Aaron Roth

Uncertain: Modern Topics in Uncertainty Estimation

INCOMPLETE WORKING DRAFT

Contents

Part I

Foundations

1

Basic Setting and Definitions

CONTENTS

We will consider 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. Depending on the setting, the label domain might be real valued ($\mathcal{Y} = \mathbb{R}$)—the *regression* setting, binary valued ($\mathcal{Y} = \{0, 1\}$)—the *binary classification* setting, or consist of some larger finite unordered set—the *multiclass classification* setting. Sometimes we will consider the regression setting in which the label domain is rescaled to the unit interval $\mathcal{Y} = [0, 1]$.

We will sometimes assume the existence of a distribution $\mathcal{D} \in \Delta\mathcal{Z}$. Given such a distribution, we will write $\mathcal{D}_{\mathcal{X}}$ to denote the marginal distribution over features: $\mathcal{D}_{\mathcal{X}} \in \Delta\mathcal{X}$ induced by $\mathcal{D}$. We will write $\mathcal{D}_{\mathcal{Y}}(x) \in \Delta\mathcal{Y}$ to denote the conditional distribution over labels induced by $\mathcal{D}$ when we condition on a particular feature vector x. $\mathcal{D}_{\mathcal{Y}}(x)$ captures all of the information about the label that is contained by the feature vector x, and is frequently the object that we are trying to approximate with our models and uncertainty quantification. A model is just some function $f : \mathcal{X} \to [0, 1]$, and our (typically unatainable goal) is to find a model f^* that has the property that for all $x \in \mathcal{X}$, $f^*(x) = \mathbb{E}_{y \sim \mathcal{D}_{\mathcal{Y}}(x)}[y]$ is the *conditional label expectation given* x.

Suppose we try and solve this problem and come up with some model f. How can we evaluate whether f is any good? If we are in a regression setting and our goal is 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 1 (Squared Error) *The squared error (also known as* Brier score*) of a predictor f on a distribution $\mathcal{D}$ is:*

$$B(f, \mathcal{D}) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x) - y)^2]$$

We will sometimes elide the distribution $\mathcal{D}$ when it is clear from context.

The Batch Setting

In the *batch* setting, we are given a *batch* or *sample* of n datapoints D sampled i.i.d. from $\mathcal{D}$, which we will write as $D \in \mathcal{Z}^n$. We will want algorithms that use D to learn something useful about $\mathcal{D}$.

We will sometimes treat a sample D as if it is a distribution: sampling from it, taking expectations over it, etc. When we do this, we are identifying D with the discrete distribution that places weight $1/n$ on each example $(x, y) \in D$. For example, we can compute the squared error of a predictor over a sample D which evaluates to:

$$B(f, D) = \frac{1}{n} \sum_{(x,y) \in D} (f(x) - y)^2$$

The Sequential Setting

In the sequential setting, data is revealed to the algorithm one example at a time, and the algorithm must make predictions before learning the label of each point. We will not always assume that the data is drawn from a distribution — often we will assume nothing about the data generation process at all, which might even be adversarial. In such cases our goals will pertain to the empirical performance of the predictions. At a high level the setting proceeds as follows, in rounds $t \in \{1, \ldots, T\}$.

1. The *adversary* chooses a (distribution over) feature vectors $x_t \in \mathcal{X}$ and labels $y_t \in \mathcal{Y}$. The realized feature vector x_t is shown to the *learner*, but not the label.
2. The learner makes some prediction p_t.
3. Finally the learner observes the realized label y_t.

Here the prediction p_t could be anything — it could try to predict the label itself, or the label mean (in the case in which $\mathcal{Y} \subset \mathbb{R}$), or it could be a prediction *set*. We'll be more specific about our goals as we proceed.

An interaction for T rounds generates a *transcript* π, which just encodes the sequence of examples and predictions across the T rounds: $\pi = \{(x_t, p_t, y_t)\}_{t=1}^{T}$.

We might assume that $(x_t, y_t) \sim \mathcal{D}$ are drawn i.i.d. from some unknown distribution $\mathcal{D}$ — more generally that the examples are drawn from an *exchangeable* distribution, which just means that their probability is permutation invariant. But more frequently we will assume that the sequence of examples is arbitrary, and will evaluate the performance of algorithms by considering empirical measures of accuracy as evaluated on the transcript π, in the worst case over adversaries.

2

A Simple Goal: Marginal Estimation

CONTENTS

We introduce the problem of learning a model that is faithful to the distribution in some formal sense with a goal that is extremely weak — too weak, on its face — in part as a straw man that will focus our attention on how we can meaningfully ask for stronger guarantees. Our initial aim will only be to find a model that matches the mean of a distribution. But we will also see that marginal guarantees like these are widely used for estimating other properties of a distribution — especially quantiles. We will talk about this at length when we get to *conformal prediction*, and we will think about ways in which we can strengthen those guarantees as well.

2.1 Means

Recall that in a regression setting in which $\mathcal{Y} \subseteq [0,1]$, our goal is to learn a model f^* such that $f^*(x) = \mathbb{E}_{y \sim \mathcal{D}(x)}[y]$ — i.e. that correctly captures the conditional label mean for each $x \in \mathcal{X}$. Of course its not clear how to do this (or even to test if we have succeeded), but we can begin with a minimal sanity check: *marginal mean consistency*:

Definition 2 *A model $f : \mathcal{X} \to [0,1]$ has marginal mean consistency error α if:*

$$\left| \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}}[f(x)] - \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}}[y] \right| = \alpha$$

If $\alpha = 0$ we'll just say that f satisfies marginal mean consistency.

This minimal sanity check is an example of a *marginal* guarantee because

it depends on f only through an unconditional expectation $\mathbb{E}[f(x)]$, rather than constraining the behavior of f conditional on any property of x. In other words, its just an average over all inputs to f. f^* satisfies marginal mean consistency, so if our model f does not, this means that our model f must not be f^*. Of course, failure to satisfy marginal mean consistency is easy to fix: Let:

$$\Delta = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[f(x)] \quad \text{and} \quad \hat{f}(x) = f(x) + \Delta.$$

It is easy to see that $\hat{f}$ satisfies marginal mean consistency:

Lemma 2.1.1 *$\hat{f}(x) = f(x) + \Delta$ satisfies marginal mean consistency.*

Proof 1

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\hat{f}(x)] &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[f(x)] + \Delta \\
&= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[f(x)] + \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[f(x)] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y]
\end{aligned}$$

as desired.

What is less obvious is that $\hat{f}$ is more accurate than f — as measured by its squared error.

Lemma 2.1.2 *Fix any distribution $\mathcal{D}$, let $f : \mathcal{X} \to [0,1]$ be any model, let $\Delta = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y] - \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[f(x)]$, and let $\hat{f}(x) = f(x) + \Delta$. Then over the distribution $\mathcal{D}$:*

$$B(\hat{f}, \mathcal{D}) = B(f, \mathcal{D}) - \Delta^2$$

Proof 2 *We can directly compute:*

$$\begin{aligned}
B(f,\mathcal{D}) - B(\hat{f},\mathcal{D}) &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[(f(x)-y)^2 - (\hat{f}(x)-y)^2\right] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[f(x)^2 - 2f(x)y + y^2 - \hat{f}(x)^2 + 2\hat{f}(x)y - y^2\right] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[f(x)^2 - 2f(x)y - (f(x)+\Delta)^2 + 2(f(x)+\Delta)y\right] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[f(x)^2 - 2f(x)y - f(x)^2 - 2\Delta f(x) - \Delta^2 + 2f(x)y + 2\Delta y\right] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[-2\Delta f(x) - \Delta^2 + 2\Delta y\right] \\
&= 2\Delta \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y - f(x)] - \Delta^2 \\
&= 2\Delta^2 - \Delta^2 \\
&= \Delta^2
\end{aligned}$$

So not only is it easy to fix a model that does not satisfy marginal mean consistency, it is always in our interest to do so if we care about accuracy: the fix is strictly accuracy improving (as measured by squared error).

2.2 Quantiles

Rather than asking for a model that matches the mean of a distribution marginally, we can ask for a model that matches a target quantile of a distribution marginally. For simplicity, we will assume that all marginal label distributions $\mathcal{D}(x)$ are continuous.

Definition 3 *Fix any $0 \leq q \leq 1$. τ is a q-quantile of a label distribution if:*

$$\Pr_y[y \leq \tau] = q$$

We also write $Q(\tau) = q$.

Once again, our goal might be to produce a model $f : \mathcal{X} \to [0, 1]$ that on each input x outputs a value $f(x)$ that is an q-quantile of the conditional label distribution $\mathcal{D}(x)$. This will generally be impossible, but we can define marginal quantile consistency as a simple sanity check analogue of marginal mean conssitency.

Definition 4 *A model $f : \mathcal{X} \to [0, 1]$ has marginal quantile consistency error α with respect to a target quantile q if:*

$$\left| \Pr_{(x,y)\sim\mathcal{D}}[y \leq f(x)] - q \right| = \alpha$$

If $\alpha = 0$ we'll say that f satisfies marginal quantile consistency for target quantile q.

Just as the Brier score will play a key role in our analysis of models that aim to match distributional means, *pinball loss* will play a key role in our analysis of models that aim to match distributional quantiles.

Definition 5 *The* pinball *loss function for target quantile q is:*

$$L_q(\tau, y) = \begin{cases} (y - \tau)q & y > \tau \\ (\tau - y)(1 - q) & y \leq \tau \end{cases}$$

Given a data distribution $\mathcal{D}$ and a function $f : \mathcal{X} \to [0, 1]$, write:

$$PB_q(f, \mathcal{D}) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[L_q(f(x), y)]$$

We will sometimes elide the distribution $\mathcal{D}$ when it is clear from context.

Observe that for $q = 1/2$, this is simply (a scaling of) the absolute value difference function: $L_{1/2}(\tau, y) = \frac{1}{2}|\tau - y|$. Just as the constant that minimizes the Brier score on a distribution is its mean, the constant that minimizes the pinball loss for a target quantile q is a q-quantile:

Lemma 2.2.1 *For any continuous distribution over y and any $0 \leq q \leq 1$*

$$\tau_q = \arg\min_{\tau \in [0,1]} \mathbb{E}_y[L_q(\tau, y)]$$

is a q-quantile.

Proof 3 *Since the distribution is continuous, this is a continuous convex function and takes its minimum at a point that has a (sub)derivative equal to 0. Thus We can calculate a (sub)derivative of the function:*

$$\begin{aligned} \frac{d\,\mathbb{E}_y[L_q(\tau, y)]}{d\tau} &= \mathbb{E}_y[(1-q)\mathbb{1}[y \leq \tau] - q\mathbb{1}[y > \tau]] \\ &= \mathbb{E}_y[\mathbb{1}[y \leq \tau] - q] \\ &= \Pr_y[y \leq \tau] - q \end{aligned}$$

Thus by inspection we see that there is a sub-derivative that takes value 0 for exactly values τ_q for which $\Pr_y[y \leq \tau_q] = q$ — i.e. q-quantiles of the distribution.

It will also be useful for us later on to have an analogue of Lemma 2.1.2 for pinball loss — i.e. a lemma that says that if we start with a model f that is far from satisfying marginal quantile consistency, and then apply a shift so that it does, that we make quantifiable progress in terms of reducing the expected pinball loss for the function.

Suppose that $f : \mathcal{X} \to [0,1]$ has marginal quantile consistency error α. Let $\Delta \in \mathbb{R}$ be such that: $\Pr_{(x,y)\sim\mathcal{D}}[y \leq f(x)+\Delta] = q$. Such a value Δ is guaranteed to exist since we have assumed that the conditional label distributions $\mathcal{D}(x)$ are continuous, and so $\Pr_{(x,y)\sim\mathcal{D}}[y \leq f(x)+\Delta]$ is a continuous monotonically increasing function taking values in the full range $[0,1]$. Let $\hat{f}(x) = f(x) + \Delta$. By construction, $\hat{f}$ satisfies marginal quantile consistency with respect to target quantile q. It also has improved pinball loss. But in order to claim that the pinball loss has improved by an amount that we can bound away from 0, we will need to assume that the conditional label distributions has bounded probability density — or equivalently that its cumulative distribution function is Lipschitz-continuous.

Definition 6 *A conditional label distribution $\mathcal{D}(x)$ is ρ-Lipschitz continuous (or just ρ-Lipschitz) if for all $0 \leq \tau \leq \tau' \leq 1$:*

$$\Pr_{y\sim\mathcal{D}(x)}[y \leq \tau'] - \Pr_{y\sim\mathcal{D}(x)}[y \leq \tau] \leq \rho(\tau' - \tau)$$

A distribution over labelled examples $\mathcal{D}$ is ρ-Lipschitz if for each $x \in \mathcal{X}$, $\mathcal{D}(x)$ is ρ-Lipschitz.

The above definition is actually somewhat stronger than we need right now — we don't need the Lipschitz condition simultaneously for each conditional label distribution $\mathcal{D}(x)$, but only marginally over the whole distribution — but this stronger condition will be useful for us later on.

Lemma 2.2.2 *Fix any distribution over labeled examples $\mathcal{D}$ that is ρ-Lipschitz. Fix any model $f : \mathcal{X} \to [0,1]$ that has marginal consistency error α with respect to target quantile q, and let $\hat{f}(x) = f(x) + \Delta$ with Δ chosen such that $\hat{f}$ satisfies marginal quantile consistency for quantile q. Then:*

$$PB_q(\hat{f}) \leq PB_q(f) - \frac{\alpha^2}{2\rho}$$

and

$$PB_q(f) \leq PB_q(\hat{f}) + |\Delta|\alpha - \frac{\alpha^2}{2\rho}$$

Proof 4 *As in the proof of Lemma 2.2.1, we can compute:*

$$\begin{aligned}
\frac{dPB_q(f(x)+\tau)}{d\tau} &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\left[\frac{d\,\mathbb{E}_{y\sim\mathcal{D}(x)}[L_q(f(x)+\tau, y)]}{d\tau}\right] \\
&= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\left[\Pr_{y\sim\mathcal{D}(x)}[y \leq f(x)+\tau] - q\right] \\
&= \Pr_{(x,y)\sim\mathcal{D}}[y \leq f(x)+\tau] - q
\end{aligned}$$

We can now compute:

$$\begin{aligned}
PB_q(\hat{f}(x)) - PB_q(f(x)) &= PB_q(f(x)+\Delta) - PB_q(f(x)) \\
&= \int_0^\Delta \frac{dPB_q(f(x)+\tau)}{d\tau} d\tau \\
&= \int_0^\Delta \left(\Pr_{(x,y)\sim\mathcal{D}}[y \leq f(x)+\tau] - q\right) d\tau \\
&= \begin{cases} \int_0^\Delta \Pr_{(x,y)\sim\mathcal{D}}[y \leq f(x)+\tau]d\tau - |\Delta|q & \Delta \geq 0 \\ \int_0^\Delta \Pr_{(x,y)\sim\mathcal{D}}[y \leq f(x)+\tau]d\tau + |\Delta|q & \Delta < 0 \end{cases}
\end{aligned}$$

$\Pr_{(x,y)\sim\mathcal{D}}[y \leq f(x)+\tau]$ is a non-negative function that is increasing in τ, and so if $\Delta < 0$ (i.e. if initially $f(x)$ is over*-predicting the q'th quantile), then we have that $\int_0^\Delta \Pr_{(x,y)\sim\mathcal{D}}[y \leq f(x)+\tau]d\tau$ evaluates to the negative of the area under the CDF of the distribution between $f(x)+\Delta$ and $f(x)$.*

Similarly the integral takes positive value if $\Delta > 0$ and corresponds to the area under the CDF between $f(x)$ and $f(x)+\Delta$. First we consider the case in which $\Delta > 0$. We need to bound $\int_0^\Delta \Pr_{(x,y)\sim\mathcal{D}}[y \leq f(x)+\tau]d\tau$. Here we will use the Lipschitz condition to upper bound the maximum possible area under the CDF. The worst case is that the CDF of the label distribution increases as

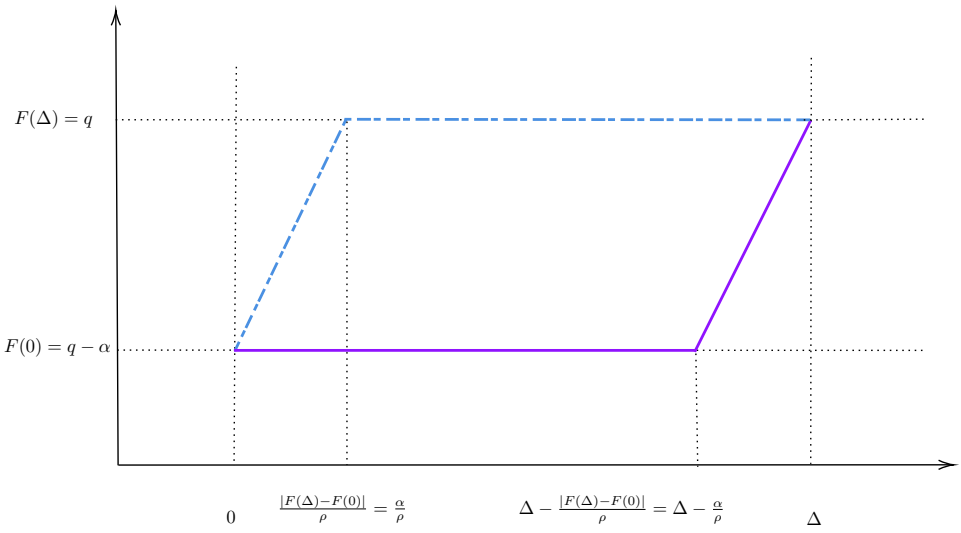

FIGURE 2.1
Upper and lower bounding the local area under the curve when $\Delta > 0$. Here $F(\tau) = \Pr_{(x,y)\sim\mathcal{D}}[y \leq f(x) + \tau]$

quickly as possible at a linear rate from $q - \alpha$ to q between $\tau = 0$ and $\tau = \alpha/\rho$, and then maintains a constant value at q from $\tau = \alpha/\rho$ to $\tau = \Delta$ (See Figure 2.1). Calculating the area under this worst case curve, we have:

$$\begin{aligned}\int_0^{\Delta} \Pr_{(x,y)\sim\mathcal{D}}[y \leq f(x) + \tau]d\tau &\leq (\Delta - \frac{\alpha}{\rho}) \cdot q + \frac{\alpha}{\rho}(q - \alpha) + \frac{\alpha}{\rho} \cdot \frac{\alpha}{2} \\ &= q\Delta - \frac{q\alpha}{\rho} + \frac{q\alpha}{\rho} - \frac{\alpha^2}{\rho} + \frac{\alpha^2}{2\rho} \\ &= q\Delta - \frac{\alpha^2}{2\rho}\end{aligned}$$

Combining with above, we have that:

$$PB_q(\hat{f}) - PB_q(f) \leq q|\Delta| - \frac{\alpha^2}{2\rho} - q|\Delta| = -\frac{\alpha^2}{2\rho}$$

Next we can lower bound the area under the CDF. Again by the Lipschitz condition, the smallest area under the CDF that respects the Lipschitz condition arises if the CDF remains constant taking value $q - \alpha$ from $\tau = 0$ to $\tau = \Delta - \frac{\alpha}{\rho}$ before increasing at a linear rate to q from $\tau = \Delta - \frac{\alpha}{\rho}$ to $\tau = \Delta$.

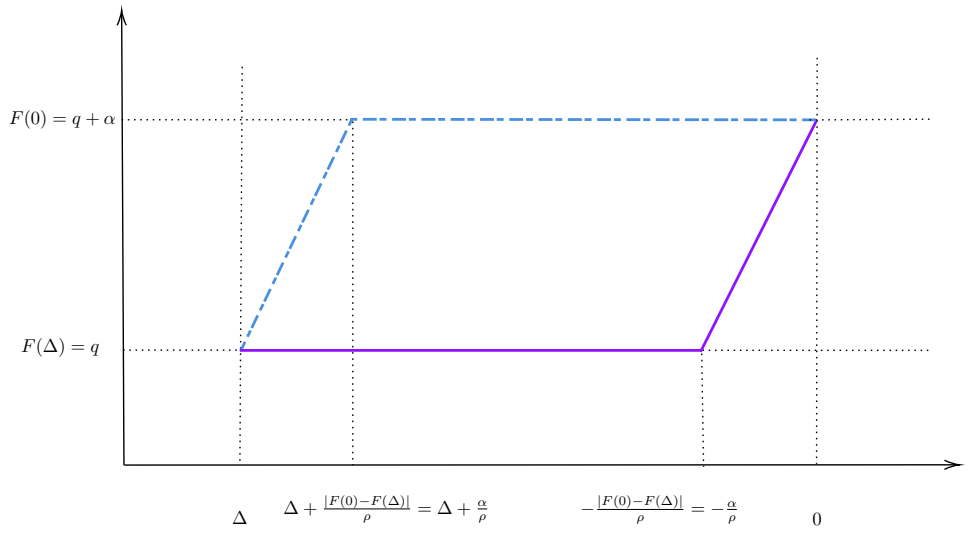

FIGURE 2.2
Upper and lower bounding the local area under the curve when $\Delta < 0$. Here $F(\tau) = \Pr_{(x,y)\sim\mathcal{D}}[y \leq f(x) + \tau]$

See figure 2.1. In this case the area is:

$$\begin{aligned}\int_0^{\Delta} \Pr_{(x,y)\sim\mathcal{D}}[y \leq f(x) + \tau]d\tau &\geq \Delta(q - \alpha) + \frac{\alpha^2}{2\rho} \\ &= \Delta q - \Delta\alpha + \frac{\alpha^2}{2\rho}\end{aligned}$$

Combining with the above we have that:

$$PB_q(\hat{f}) - PB_q(f) \geq |\Delta|q - |\Delta|\alpha + \frac{\alpha^2}{2\rho} - |\Delta|q = -|\Delta|\alpha + \frac{\alpha^2}{2\rho}$$

In the remaining case in which $\Delta < 0$, our worst cases are reversed (we need to maximize the area under the curve to lower bound the integral and minimize the area under the curve to upper bound the integral). Once again, the CDF that minimizes the area under the curve subject to the Lipschitz constraint behaves as follows (See figure 2.2): The CDF remains constant at q between $\tau = \Delta$ and $\tau = -\alpha/\rho$, before increasing as quickly as possible at a linear rate up to value $q + \alpha$ between $\tau = -\alpha/\rho$ and $\tau = 0$. In this case we have that:

$$\begin{aligned}\int_0^{\Delta} \Pr_{(x,y)\sim\mathcal{D}}[y \leq f(x) + \tau]d\tau &\leq -\left(|\Delta|q + \frac{\alpha^2}{2\rho}\right) \\ &= -q|\Delta| - \frac{\alpha^2}{2\rho}\end{aligned}$$

Again combining with above we get that:

$$PB_q(\hat{f}) - PB_q(f(x)) \leq -q|\Delta| - \frac{\alpha^2}{2\rho} + q|\Delta| = -\frac{\alpha^2}{2\rho}$$

Finally, the CDF that maximizes the area under the curve subject to the Lipschitz constraint increases at a linear rate from $\tau = \Delta$ to $\tau = \Delta + \alpha/\rho$ from value q to value $q + \alpha$, and then remains constant at $q + \alpha$ from $\tau = \Delta + \frac{\alpha}{\rho}$ to $\tau = 0$. Computing the area under this curve, we get:

$$\begin{aligned}\int_0^{\Delta} \Pr_{(x,y)\sim\mathcal{D}}[y \leq f(x) + \tau]d\tau &\geq -\left(\frac{\alpha^2}{2\rho} + |\Delta|q + \alpha(|\Delta| - \frac{\alpha}{\rho})\right) \\ &= \frac{\alpha^2}{2\rho} - |\Delta|q - |\Delta|\alpha\end{aligned}$$

Together with the above we have that:

$$PB_q(\hat{f}) - PB_q(f(x)) \geq \frac{\alpha^2}{2\rho} - |\Delta|q - |\Delta|\alpha + q|\Delta| = \frac{\alpha^2}{2\rho} - |\Delta|\alpha$$

which completes the proof of the lemma.

So once again, if a model fails to satisfy marginal quantile consistency, it is easy to fix the model, and once again, doing so is accuracy improving — this time as measured via Pinball loss.

We'll make one more observation: we proved Lemma 2.2.2 under the assumption that the underlying distribution $\mathcal{D}$ was ρ-Lipschitz. But eventually, when we want to apply similar arguments to algorithms run on *data* sampled from some underlying distribution, we will face the problem that the empirical distribution over a finite dataset is discrete, and hence cannot be Lipschitz at fine enough resolutions. But we observe that Lemma 2.2.2 actually only speaks to updates Δ that are applied to fix marginal consistency error of scale α — and if the underlying distribution is ρ-Lipschitz, we must have that $\Delta \geq \frac{\alpha}{\rho}$. So we really only require that the underlying distribution is Lipschitz at scales larger than $\frac{\alpha}{\rho}$. Here we define the condition we need — Lipschitzness only at large enough scales:

Definition 7 *Fix $\rho, r > 0$. A conditional label distribution $\mathcal{D}(x)$ is (ρ, r)-Lipschitz continuous (or just (ρ, r)-Lipschitz) if for all $0 \leq \tau \leq \tau' \leq 1$ such that $\tau' - \tau \geq r$:*

$$\Pr_{y\sim\mathcal{D}(x)}[y \leq \tau'] - \Pr_{y\sim\mathcal{D}(x)}[y \leq \tau] \leq \rho(\tau' - \tau)$$

A distribution over labelled examples $\mathcal{D}$ is (ρ, r)-Lipschitz if for each $x \in \mathcal{X}$, $\mathcal{D}(x)$ is (ρ, r)-Lipschitz.

Using the insight that our proof of Lemma 2.2.2 really only used the condition of $(\rho, \frac{\alpha}{\rho})$-Lipschitz continuity, we have the following lemma:

Lemma 2.2.3 *Fix any $\rho, \alpha > 0$. Fix any distribution over labeled examples $\mathcal{D}$ that is $(\rho, \frac{\alpha}{\rho})$-Lipschitz. Fix any model $f : \mathcal{X} \to [0,1]$ that has marginal consistency error α with respect to target quantile q, and let $\hat{f}(x) = f(x) + \Delta$ with Δ chosen such that $\hat{f}$ satisfies marginal quantile consistency for quantile q. Then:*

$$PB_q(\hat{f}) \leq PB_q(f) - \frac{\alpha^2}{2\rho}$$

and

$$PB_q(f) \leq PB_q(\hat{f}) + |\Delta|\alpha - \frac{\alpha^2}{2\rho}$$

2.2.1 Generalizing From Data

Thus far we have been acting as if we have direct access to the data distribution $\mathcal{D}$, and in particular, given a fixed model f can compute the quantity Δ such that $\hat{f}(x) = f(x) + \Delta$ satisfies our marginal consistency desideratum with respect to either means or quantiles. But of course we generally will not have direct access to $\mathcal{D}$, and will instead have only a *sample* $D \sim \mathcal{D}^n$ of n points drawn i.i.d. from $\mathcal{D}$. What we will generally do (now, and for more complex algorithms in later chapters) is run our algorithms on the *empirical distribution* over our sample D, and then prove that the guarantees that our algorithms have on D carry over (with small loss) to $\mathcal{D}$.

Theorem 1 *Fix any model f and distribution $\mathcal{D}$, and let $D \sim \mathcal{D}^n$ consist of n samples drawn i.i.d. from $\mathcal{D}$. Let Δ be such that $\hat{f}(x) = f(x) + \Delta$ satisfies marginal mean consistency on D. Then with probability $1 - \delta$ over the draw of D, $\hat{f}$ has marginal mean consistency error at most α on $\mathcal{D}$, for:*

$$\alpha \leq \sqrt{\frac{2\log(2/\delta)}{n}}$$

Proof 5 *This is an application of Hoeffding's inequality (Theorem 46) which we quote here in its first use:*

> *Let $X_1, \ldots, X_n$ be independent random variables bounded such that for each i, $a_i \leq X_i \leq b_i$. Let $S_n = \sum_{i=1}^n X_i$ denote their sum. Then for all $t > 0$:*
>
> $$\Pr\left[|S_n - \mathbb{E}[S_n]| \geq t\right] \leq 2\exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

In our case, we have that $\Delta = \frac{1}{n}\sum_{(x,y)\in D}(y - f(x))$, and each term $\frac{1}{n}(y - f(x))$ is bounded such that:

$$-\frac{1}{n} \leq \frac{1}{n}(y - f(x)) \leq \frac{1}{n}$$

We also have that $\mathbb{E}_D \Delta = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y - f(x)]$. *Thus we can apply Hoeffding's inequality to conclude that:*

$$\Pr[|\Delta - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y - f(x)]| \geq t] \leq 2\exp\left(\frac{-nt^2}{2}\right)$$

Setting the right hand side to be at most δ *and solving for* t, *we find that:*

$$\Pr\left[|\Delta - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y - f(x)]| \geq \sqrt{\frac{2\log(2/\delta)}{n}}\right] \leq \delta$$

Finally, recall that by definition, $\hat{f}$ *has marginal mean consistency error:*

$$\begin{aligned}\left|\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\hat{f}(x) - y]\right| &= \left|\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f(x) + \Delta - y]\right| \\ &\leq \left|\Delta - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y - f(x)]\right| \\ &\leq \sqrt{\frac{2\log(2/\delta)}{n}}\end{aligned}$$

where the last inequality holds with probability $1 - \delta$, *as established by Hoeffding's inequality.*

Theorem 2 *Fix any model* f *and distribution* $\mathcal{D}$, *and let* $D \sim \mathcal{D}^n$ *consist of* n *samples drawn i.i.d. from* $\mathcal{D}$. *Let* Δ *be such that* $\hat{f}(x) = f(x) + \Delta$ *has quantile mean consistency error* α' *with respect to some target quantile* q *on* D. *Then with probability* $1 - \delta$ *over the draw of* D, $\hat{f}$ *has marginal quantile consistency error at most* α *with respect to target quantile* q *on* $\mathcal{D}$, *for:*

$$\alpha \leq \alpha' + \sqrt{\frac{\log(2/\delta)}{2n}}$$

Proof 6 *This is an application of the DKW (Dvoretzky–Kiefer–Wolfowitz) inequality (Theorem 49) which we quote here in its first use:*

> *Let* $\mathcal{D} \in \mathcal{Z}^n$ *be any distribution and let* $D \sim \mathcal{D}^n$ *consist of* n *points sampled i.i.d. from* D. *Let* $F(c) = \Pr_{(x,y)\sim\mathcal{D}}[y \leq c]$ *denote the CDF of the label distribution induced by* $\mathcal{D}$, *and let* $\hat{F}_D(c) = \frac{1}{n}\sum_{(x,y)\in D} \mathbb{1}[y \leq c]$ *denote the CDF of the empirical label distribution induced by* D. *Then for every* $t > 0$*:*
>
> $$\Pr\left[\sup_{c\in\mathbb{R}} \left|F(c) - \hat{F}_D(C)\right| \geq t\right] \leq 2\exp\left(-2nt^2\right)$$

Consider the distribution $\mathcal{D}'$ *which is derived from* $\mathcal{D}$ *by replacing the label* y *of each example* (x, y) *with the label* $y' = y - f(x)$. *We apply the DKW inequality to this distribution. By definition,* Δ *is chosen such that*

$$\left|\Pr_{(x,y)\sim D}[y \leq f(x) + \Delta] - q\right| = \alpha'$$

rearranging, this is:

$$|\hat{F}_D(\Delta) - q| \leq \alpha'$$

Applying the DKW inequality with $t = \sqrt{\frac{\log(2/\delta)}{2n}}$*, we have that with probability* $1 - \delta$*:*

$$\left|\hat{F}_D(\Delta) - F(\Delta)\right| \leq \sqrt{\frac{\log(2/\delta)}{2n}}$$

And so $|F(\Delta) - q| \leq \alpha' + \sqrt{\frac{\log(2/\delta)}{2n}}$ *Expanding out the definition of* $F(\Delta)$ *we have that*

$$\begin{aligned}
\alpha' + \sqrt{\frac{\log(2/\delta)}{2n}} &\geq \left| \Pr_{(x,y)\sim\mathcal{D}}[y - f(x) \leq \Delta] - q \right| \\
&= \left| \Pr_{(x,y)\sim\mathcal{D}}[y \leq f(x) + \Delta] - q \right| \\
&= \left| \Pr_{(x,y)\sim\mathcal{D}}[y \leq \hat{f}(x)] - q \right|
\end{aligned}$$

The result is that we can simply proceed as if our sample is our underlying distribution when we aim for marginal consistency — and our marginal consistency error on the underlying distribution is guaranteed to be larger than our empirical marginal consistency error by at most ϵ with probability $1 - \delta$, whenever $n \geq \Omega\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$.

2.3 Sequential Prediction

What about when we are in a sequential prediction setting, and there is no distribution? Even when examples are selected by an adversary, we can still talk about marginal mean and quantile consistency (and all of the other distributional measures that we will introduce in later chapters). We will always evaluate these guarantees ex-post, over the empirical distribution over the transcript.

Definition 8 *Fix a transcript* $\pi = \{(x_t, p_t, y_t)\}_{t=1}^T$ *consisting of examples* $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$ *and predictions* $p_t \in [0, 1]$*. The transcript satisfies marginal mean consistency with error* α *if:*

$$\left| \frac{1}{T}\sum_{t=1}^{T} p_t - \frac{1}{T}\sum_{t=1}^{T} y_t \right| = \alpha$$

It satisfies marginal quantile consistency with respect to a target quantile q *and error* α *if:*

$$\left| \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}[y_t \leq p_t] - q \right| = \alpha$$

Our goal in sequential prediction settings will generally be to derive algorithms that guarantee that against *any adversary*, they generate a transcript π that with high probability (or with certainty) satisfies some notion of statistical consistency. In general, solving these problems in the adversarial sequential setting is only more difficult than solving them in the batch, distributional setting, in a formal sense: If we have algorithms that promise consistency guarantees in the sequential setting, then by running them on data that is in fact drawn i.i.d. from some distribution, we can also obtain the same guarantees in the batch setting. The reverse is not true — the sequential setting is generally strictly harder.

Here we give a warm-up version of this style of theorem, just for mean marginal consistency. In this case, it is easy to more directly get the same kinds of out of sample guarantees — but more later we will see more sophisticated versions of this kind of online-to-offline reduction. It will be an application of Hoeffding's inequality, which we state as Theorem 46.

Theorem 3 *Suppose we have an algorithm $\mathcal{A}$ that when run against any adversary for T rounds generates a transcript π that satisfies marginal mean consistency with error at most α. Suppose we have some model $f : \mathcal{X} \to [0, 1]$ and a data distribution $\mathcal{D}$, and consider the following procedure to simulate an adversary. At each round t we:*

1. *Sample $(\hat{x}_t, \hat{y}) \sim \mathcal{D}$*
2. *Feed algorithm $\mathcal{A}$ the sample $(x_t, y_t) = (\hat{x}, \hat{y}_t - f(\hat{x}))$*

This results in some transcript $\pi = \{(x_t, p_t, y_t)\}_{t=1}^T$. Let $\Delta = \frac{1}{T}\sum_{t=1}^T p_t$ and let $\hat{f}(x) = f(x) + \Delta$. Then for any $\delta > 0$, with probability $1 - \delta$, $\hat{f}$ satisfies marginal mean consistency with error α' for:

$$\alpha' \leq \alpha + \sqrt{\frac{2\log(2/\delta)}{T}}$$

Proof 7 *Since π is promised to satisfy marginal mean consistency with error at most α, we know that:*

$$\left|\frac{1}{T}\sum_{t=1}^T p_t - \frac{1}{T}\sum_{t=1}^T y_t\right| \leq \alpha$$

Let:

$$\bar{\Delta} = \frac{1}{T}\sum_{t=1}^T (\hat{y}_t - f(\hat{x}_t))$$

Plugging in the definitions of Δ and y_t we have that:

$$\left|\Delta - \bar{\Delta}\right| \leq \alpha$$

Note also that since $(\hat{x}_t, \hat{y}_t)$ are sampled i.i.d. from $\mathcal{D}$, we have that:

$$\mathbb{E}[\bar{\Delta}] = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y - f(x)]$$

We can now apply Hoeffding's inequality (Theorem 46) to the quantity $\bar{\Delta} = \frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - f(\hat{x}_t))$. Each term in the sum is bounded between $-1/T \leq \frac{1}{T}(\hat{y}_t - f(\hat{x}_t)) \leq 1/T$ and so we have for any $\epsilon > 0$:

$$\Pr\left[\left|\bar{\Delta} - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y - f(x)]\right| \geq \epsilon\right] \leq 2\exp\left(\frac{-2T\epsilon^2}{4}\right)$$

The right hand side is at most δ when we have:

$$\epsilon \geq \sqrt{\frac{2\log(2/\delta)}{T}}$$

We therefore have that with probability $1 - \delta$:

$$\begin{aligned}
\left|\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\hat{f}(x) - y]\right| &= \left|\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f(x) + \Delta - y]\right| \\
&\leq \left|\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f(x) + \bar{\Delta} - y]\right| + \alpha \\
&= \left|\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[f(x) + \mathbb{E}[\bar{\Delta}] + (\bar{\Delta} - \mathbb{E}[\bar{\Delta}]) - y\right]\right| + \alpha \\
&= \left|\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[(\bar{\Delta} - \mathbb{E}[\bar{\Delta}])\right]\right| + \alpha \\
&\leq \alpha + \sqrt{\frac{2\log(2/\delta)}{T}}
\end{aligned}$$

as desired.

Next, we'll see a simple algorithm that can guarantee marginal mean consistency with error on the order of $O(1/T)$ on *any* sequence of length T — i.e. without assuming that the data points come from a distribution. The algorithm will be silly on its face as a prediction algorithm — always predicting that today's outcome will be equal to yesterday's outcome. Its excellent performance (as measured by marginal mean consistency) tells us something about the weakness of marginal guarantees.

Algorithm 1 Online-Marginal-Mean-Predictor

Let $y_0 = 0$
for $t = 1$ to T **do**
  Observe x_t (and ignore it!)
  Predict $p_t = y_{t-1}$
  Observe y_t.

If we imagine using this algorithm to predict weather, then what it does is the following: If it rained yesterday, it predicts a 100% chance of rain today. Otherwise it predicts a 0% chance of rain. And yet:

Theorem 4 *For any sequence of examples of length T, $\{(x_t, y_t)\}_{t=1}^T$ Online-Marginal-Mean-Predictor (Algorithm 1) produces a transcript that satisfies marginal mean consistency with error α for $\alpha \leq 1/T$.*

Proof 8 *Using the fact that $p_t = y_{t-1}$ (and $y_0 = 0$) we compute:*

$$\begin{aligned}\left|\frac{1}{T}\sum_{t=1}^{T} p_t - \frac{1}{T}\sum_{t=1}^{T} y_t\right| &= \left|\frac{1}{T}\sum_{t=1}^{T} y_{t-1} - \frac{1}{T}\sum_{t=1}^{T} y_t\right| \\ &= \frac{1}{T}\left|y_0 - y_T\right| \\ &\leq \frac{1}{T}\end{aligned}$$

Now lets do the same for quantiles. First we argue that obtaining marginal quantile consistency in the online setting is sufficient to obtain marginal quantile consistency on a distribution, and then show a simple deterministic algorithm for obtaining marginal quantile consistency in the online adversarial setting. There will be one major difference, which is that to convert an online sequence of predictions to an offline quantile predictor, we cannot simply average the predicted quantiles as we did with predicted means (because the relationship between the numeric value of quantiles and their inverse CDF value is not linear). Instead, we will *randomize* over the sequence of predictions, which will result in an offline randomized quantile predictor.

Theorem 5 *Suppose we have an algorithm $\mathcal{A}$ that when run against any adversary for T rounds generates a transcript π that satisfies marginal quantile consistency with error at most α for some target quantile q. Suppose we have some model $f : \mathcal{X} \to [0, 1]$ and a data distribution $\mathcal{D}$, and consider the following procedure to simulate an adversary. At each round t we:*

1. *Sample $(\hat{x}_t, \hat{y}_t) \sim \mathcal{D}$*
2. *Feed algorithm $\mathcal{A}$ the sample $(x_t, y_t) = (\hat{x}, \hat{y}_t - f(\hat{x}))$*

This results in some transcript $\pi = \{(x_t, p_t, y_t)\}_{t=1}^T$. Let Δ be the random variable that takes value in $\{p_1, \ldots, p_T\}$ uniformly at random (i.e. $\Delta = p_1$ with probability $1/T$, $\Delta = p_2$ with probability $1/T$, etc.) Let $\hat{f}(x)$ be the randomized predictor defined has $\hat{f}(x) + \Delta$. Then for any $\delta > 0$, with probability $1 - \delta$, $\hat{f}$ satisfies marginal quantile consistency with error α' with respect to target quantile q for:

$$\alpha' \leq \alpha + \sqrt{\frac{2\ln(2/\delta)}{T}}$$

In other words:

$$\left| \Pr_{(x,y)\sim\mathcal{D},\Delta}[y \leq f(x) + \Delta] - q \right| \leq \alpha'$$

Proof 9 *Since π is promised to satisfy marginal quantile consistency w.r.t. quantile q with error at most α, we know that:*

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}[y_t \leq p_t] - q \right| \leq \alpha$$

Plugging in the definition of y_t we have that:

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}[\hat{y}_t - f(\hat{x}_t) \leq p_t] - q \right| \leq \alpha$$

Let $\mathcal{D}'$ be the label distribution induced by outputting the label $y - f(x)$ for $(x, y) \sim \mathcal{D}$ and let F denote its CDF: $F(x) = \Pr_{y\sim\mathcal{D}'}[y \leq x]$. We want to be able to say that $\frac{1}{T}\sum_{t=1}^{T} F(p_t) \approx q$, but we have a problem: the indicators $\mathbb{1}[y_t \leq p_t]$ are not independent random variables even though the y_t are, since each p_t is potentially chosen as a function of all previous labels $y_1, \ldots, y_{t-1}$. Hence we cannot apply Hoeffding's inequality. But all is not lost! We will need Azuma's inequality (Theorem 48) which we quote here before its first use:

> *Let $X_1, \ldots, X_n$ be random variables (not necessarily independent) bounded such that for each i, $|X_i| \leq c_i$. Let $X_{<i}$ denote the prefix $X_1, X_2, \ldots, X_{i-1}$. Then for all $t > 0$:*
>
> $$\Pr\left[\left| \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}[X_i | X_{<i}] \right| \geq t \right] \leq 2\exp\left(\frac{-t^2}{2\sum_{i=1}^{n} c_i^2} \right)$$

Recall that for a sequential prediction algorithm, p_t can be chosen as a function of past examples — but must be independent of the current example y_t. Hence we do have that $\mathbb{E}_{y_t}[\mathbb{1}[y_t \leq p_t | y_{<t}] = F(p_t)$. For us, the random variables are $1/T(\mathbb{1}[y_t \leq p_t])$ which are bounded by $c_t = 1/T$. Thus we can apply Azuma's inequality with $t = \sqrt{\frac{2\ln(2/\delta)}{T}}$ to conclude that:

$$\Pr\left[\left| \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}[y_t \leq p_t] - \frac{1}{T} \sum_{t=1}^{T} F(p_t) \right| \geq \sqrt{\frac{2\ln(2/\delta)}{T}} \right] \leq \delta$$

Combining this with our guarantee of marginal quantile consistency, with probability $1 - \delta$ we have that:

$$\left| \frac{1}{T} \sum_{t=1}^{T} F(p_t) - q \right| \leq \alpha + \sqrt{\frac{2\ln(2/\delta)}{T}}$$

Finally we can compute:

$$\begin{aligned}\left|\Pr_{(x,y)\sim\mathcal{D},\Delta}[y \le f(x)+\Delta] - q\right| &= \left|\frac{1}{T}\sum_{t=1}^{T}\Pr_{(x,y)\sim\mathcal{D}}[y \le f(x)+p_t] - q\right| \\ &= \left|\frac{1}{T}\sum_{t=1}^{T}F(p_t) - q\right| \\ &\le \alpha + \sqrt{\frac{2\ln(2/\delta)}{T}}\end{aligned}$$

where the last inequality holds with probability $1-\delta$ *over the draws of* $\{(x_t,y_t)\}_{t=1}^{T}$.

Next, we give our algorithm for making predictions that satisfy online marginal quantile consistency for any target quantile q against any adversarially chosen sequence of examples. The algorithm takes as input a "learning rate" parameter η, and can be viewed and analyzed as online gradient descent on the pinball loss. But the specific form of the resulting update also lends itself to a very simple analysis showing that its quantile error tends to 0 at a rate of $1/T$, just as our algorithm for obtaining marginal mean consistency does.

Algorithm 2 Online-Marginal-Quantile-Predictor(q,η)

Let $p_1 = 0$
for $t = 1$ to T **do**
  Observe x_t (and ignore it!)
  Predict p_t
  Observe y_t.
  Let $p_{t+1} = p_t + \eta(q - \mathbb{1}[y_t \le p_t])$

Theorem 6 *For any sequence of examples of length* T*, any target quantile* $q \in [0,1]$ *and any update parameter* $\eta > 0$*, Online-Marginal-Quantile-Predictor*(q,η) *(Algorithm 2) produces a transcript that satisfies marginal quantile consistency with error* α *for* $\alpha \le \frac{1+\eta}{\eta T}$

Proof 10 *Examining the update rule* $p_{t+1} = p_t + \eta(q - \mathbb{1}[y_t \le p_t])$ *and solving for* $\mathbb{1}[y_t \le p_t])$*, we see:*

$$\mathbb{1}[y_t \le p_t] = q - \frac{p_{t+1} - p_t}{\eta}$$

So, we can compute:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}[y_t \le p_t] &= q - \frac{1}{\eta T}\sum_{t=1}^{T}(p_{t+1} - p_t) \\ &= q - \frac{p_{T+1} - p_1}{\eta T} \\ &= q - \frac{p_{T+1}}{\eta T}\end{aligned}$$

Next observe that for all t, $|p_t - p_{t+1}| \le \eta$, and since $y_t, q \in [0,1]$, if $p_t \ge 1$, it must be that $\mathbb{1}[y_t \le p_t] = 1$ then $p_{t+1} < p_t$ and similarly if $p_t \le 0$ then $p_{t+1} > p_t$. Hence we must have for all t that:

$$-\eta \le p_t \le 1 + \eta$$

So we have:

$$\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}[y_t \le p_t] - q\right| = \left|\frac{p_{T+1}}{\eta T}\right| \le \frac{1+\eta}{\eta T}$$

Remark 2.3.1 *In fact, there is an even simpler algorithm that can guarantee marginal quantile consistency against an adversary, with error tending to 0 at a rate of $1/T$. For a q fraction of rounds, predict $p_t = 1$, and for a $1-q$ fraction of rounds predict $p_t = 0$. Because we know that $y_t \in [0,1]$, we have that on the q fraction of rounds for which $p_t = 1$, $\mathbb{1}[y_t \le p_t] = 1$ and for the remaining $1 - q$ fraction of rounds, $\mathbb{1}[y_t \le p_t] = 0$. Hence we can satisfy marginal quantile consistency in an entirely data independent way, which should make us suspicious of marginal guarantees and make us ask for something stronger.*

References and Further Reading

Lemma 2.2.2 (bounding the change in pinball loss as a function of the change in predicted quantile under a Lipschitz condition on the distribution) is adapted from Jung et al. [2022]. Algorithm 2 and its analysis are adapted from Gibbs and Candes [2021], who derive it in the context of conformal prediction (which we will see later in Chapter 7).

3

Calibration

CONTENTS

The marginal guarantees we saw in Chapter 2 were easy to obtain, but extremely weak. In this chapter we'll see one way to go beyond marginal guarantees, by making *calibrated* predictions. Calibration on its own is also quite weak, but not as weak as a marginal guarantee, and should be thought of as one step up in terms of a "sanity check" intended to falsify whether we have learned the true conditional label distribution.

3.1 Introduction to Calibration

In this section we'll focus on regression problems in which the label domain is real valued: $\mathcal{Y} \subset [0,1]$. A natural special case is when we are predicting binary outcomes: $\mathcal{Y} = \{0,1\}$, but everything we say holds also for the general real valued case.

In such a setting, we often want to solve the *regression* problem: that is, to find a model $f : \mathcal{X} \to [0,1]$ that has the property that for all $x \in \mathcal{X}$, $f^*(x) = \mathbb{E}_{y \sim \mathcal{D}_{\mathcal{Y}}(x)}[y]$ is the *conditional label expectation given* x. Of course, we don't generally expect to actually find this function (for a variety of reasons), but that's going to be our goal.

Suppose we try and solve this problem and come up with some model f. How can we evaluate whether f is any good? If our goal is 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 9 (Squared Error) *The squared error (also known as* Brier score*) of a predictor f is:*

$$B(f) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x) - y)^2]$$

On the other hand, if we want our predictions $f(x)$ to have the same probabilistic semantics as $f^*(x)$ — namely that they be a prediction about the *expected value of* y, then we might want that $f(x)$ be calibrated. Calibration asks that the predictions of f be correct conditional on its own predictions: Roughly that $\mathbb{E}_{(x,y)\sim\mathcal{D}}[y|f(x) = v] = v$ for all v. So that the conditioning event makes sense, we will restrict attention so functions f that have a range of finite cardinality, and study *average* calibration error. Let $R(f) = \{f(x) : x \in \mathcal{X}\}$ denote the range of f, and let $m = |R(f)|$ denote the cardinality of f's range. We will assume $m < \infty$.

Definition 10 (Average Calibration Error) *The average calibration error of a predictor f on a distribution $\mathcal{D}$ is:*

$$K_1(f, \mathcal{D}) = \sum_{v \in R(f)} \Pr_{(x,y)\sim\mathcal{D}}[f(x) = v] \left| v - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y|f(x) = v] \right|$$

The average squared calibration error is:

$$K_2(f, \mathcal{D}) = \sum_{v \in R(f)} \Pr_{(x,y)\sim\mathcal{D}}[f(x) = v] \left(v - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y|f(x) = v] \right)^2$$

Finally, we can define a notion of maximum calibration error. Just as with our average notions, we weight by the probability mass of the levelsets to avoid needing to measure quantities over sets with tiny mass:

$$K_\infty(f, \mathcal{D}) = \max_{v \in R(f)} \Pr_{(x,y)\sim\mathcal{D}}[f(x) = v] \left| v - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y|f(x) = v] \right|$$

When the distribution $\mathcal{D}$ is clear from context we will sometimes elide it and simply write $K_1(f), K_2(f), K_\infty(f)$, etc.

Sometimes it will be more convenient to work with one of these quantities over another, but they are closely related to one another:

Lemma 3.1.1 *For any predictor $f : \mathcal{X} \to [0, 1]$,*

$$K_2(f) \leq K_1(f) \leq \sqrt{K_2(f)}$$
$$K_\infty(f) \leq K_1(f) \leq mK_\infty(f)$$

Proof 11 *$K_2(f) \leq K_1(f)$ follows from the fact that since v and y are bounded in $[0, 1]$, term by term $\left(v - \mathbb{E}_{(x,y)\sim\mathcal{D}}[y|f(x) = v]\right)^2 \leq$*

$\left|v - \mathbb{E}_{(x,y)\sim\mathcal{D}}[y|f(x)=v]\right|$. $K_1(f) \leq \sqrt{K_2(f)}$ *follows from the Cauchy-Schwarz inequality:*

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

$K_\infty(f) \leq K_1(f)$ *follows from the fact that a sum of non-negative terms upper bounds a maximum over the terms, and* $K_1(f) \leq mK_\infty(f)$ *follows from the fact that* $K_1(f)$ *is a sum of* m *terms each of which is upper bounded by* $K_\infty(f)$.

Unlike squared error, which we may never be able to drive to zero (because of inherent unpredictability), we can in principle drive calibration error to zero: observe that $K_2(f^*) = 0$.

In fact, f^* also minimizes the squared error over the set of all functions because $f^*(x)$ minimizes squared error point-wise per prediction x:

Lemma 3.1.2 *Fix any distribution on labels* $\mathcal{D}_\mathcal{Y}$. *Let* $v^* = \mathbb{E}_{\mathcal{D}_\mathcal{Y}}[y]$ *denote the true label expectation, and let* $\hat{v} = v^* + \Delta$ *for some* $\Delta \neq 0$. *Then:*

$$\mathop{\mathbb{E}}_{y\sim\mathcal{D}_\mathcal{Y}}[(\hat{v}-y)^2 - (v^*-y)^2] = \Delta^2$$

Proof 12

$$\begin{aligned}
\mathop{\mathbb{E}}_{y\sim\mathcal{D}_\mathcal{Y}}\left[(\hat{v}-y)^2 - (v^*-y)^2\right] &= \mathop{\mathbb{E}}_{y\sim\mathcal{D}_\mathcal{Y}}\left[\hat{v}^2 - 2\hat{v}y - (v^*)^2 + 2v^*y\right] \\
&= \mathop{\mathbb{E}}_{y\sim\mathcal{D}_\mathcal{Y}}\left[(v^*+\Delta)^2 - 2\hat{v}y - (v^*)^2 + 2v^*y\right] \\
&= \mathop{\mathbb{E}}_{y\sim\mathcal{D}_\mathcal{Y}}\left[2v^*\Delta + \Delta^2 - 2(v^*+\Delta)y + 2v^*y\right] \\
&= \mathop{\mathbb{E}}_{y\sim\mathcal{D}_\mathcal{Y}}\left[2v^*\Delta + \Delta^2 - 2\Delta y\right] \\
&= 2v^*\Delta + \Delta^2 - 2v^*\Delta \\
&= \Delta^2
\end{aligned}$$

3.2 Calibrating a Model f

Suppose we are given a model f with large average calibration error $K_2(f)$. Can we fix it? And will fixing it come at the cost of accuracy (say, as measured by squared error $B(f)$)? The answers are "Yes", and "No" respectively! :-) There is a simple algorithm that takes as input an arbitrary model f and outputs a modified model $\hat{f}$ such that:

1. $\hat{f}$ has as low average calibration error as we like: For any α, we can produce $\hat{f}$ such that $K_2(\hat{f}) \leq \alpha$

2. $\hat{f}$ has strictly lower squared error than f if f was not already calibrated.
3. The range of $\hat{f}$ is only smaller than the range of f: $|R(\hat{f})| \leq |R(f)|$

The basic idea will be to take some intermediate model f_t and then "patch" it if it is not already calibrated, to produce a better model f_{t+1}. We will focus on the simplest possible "patch", and form our calibrated model by simply stringing them together.

Definition 11 (Value Patch) *Given a model f and a pair of values $v, v' \in [0,1]$, we say that the* value patch *applied to f with pair (v, v') is the function:*

$$h(x, f; v \to v') = \begin{cases} v' & f(x) = v \\ f(x) & otherwise \end{cases}$$

Algorithm 3 Calibrate$(f, \alpha, \mathcal{D})$

Let $f_0 = f$ and $t = 0$.
while $K_2(f_t, \mathcal{D}) \geq \alpha$ **do**
Let:

$$v_t \in \arg\max_{v \in R(f_t)} \Pr_{(x,y)\sim\mathcal{D}}[f_t(x) = v] \left(v - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y | f_t(x) = v] \right)^2$$

$$v'_t = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y | f_t(x) = v_t]$$

Let $f_{t+1} = h(x; f_t, v_t \to v'_t)$ and $t = t + 1$.
Output f_t.

We can now analyze the algorithm.

Theorem 7 *After T rounds, where $T \leq \frac{m}{\alpha}$, Algorithm 3 outputs a model f_T such that $K_2(f_T) \leq \alpha$ and $B(f_T) \leq B(f)$.*

Proof 13 *Observe that at each round before the algorithm halts, since $K_2(f_t) \geq \alpha$ we must have that:*

$$\Delta_t \equiv \Pr_{(x,y)\sim\mathcal{D}}[f_t(x) = v_t] \left(v_t - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y | f(x) = v_t] \right)^2 \geq \frac{\alpha}{m}$$

Rearranging, we also have that:

$$(v_t - v'_t)^2 = \frac{\Delta_t}{\Pr_{(x,y)\sim\mathcal{D}}[f_t(x) = v_t]}$$

Let $\mathcal{D}(v_t) = \mathcal{D}|(f_t(x) = v_t)$ be the distribution that results from conditioning on the event that $f_t(x) = v_t$ and let $\mathcal{D}(\bar{v}_t) = \mathcal{D}|(f_t(x) \neq v_t)$ be the distribution that results from conditioning on the event that $f_t(x) \neq v_t$. We have from Lemma 3.1.2 that:

$$\begin{aligned}
& B(f_{t+1}, \mathcal{D}) - B(f_t, \mathcal{D}) \\
= \; & \Pr[f_t(x) = v_t](B(f_{t+1}, \mathcal{D}(v_t)) - B(f_t, \mathcal{D}(v_t)) + \Pr[f_t(x) \neq v_t](B(f_{t+1}, \mathcal{D}(\bar{v}_t)) - B(f_t, \mathcal{D}(\bar{v}_t)) \\
= \; & \Pr[f_t(x) = v_t](B(f_{t+1}, \mathcal{D}(v_t)) - B(f_t, \mathcal{D}(v_t)) \\
= \; & \Pr[f_t(x) = v_t](v_t - v_t')^2 \\
= \; & \Delta_t \\
\geq \; & \frac{\alpha}{m}
\end{aligned}$$

Here the second to last inequality follows from Lemma 3.1.2. Since for any model $f : \mathcal{X} \to [0,1]$, $B(f, \mathcal{D}) \leq 1$ and for any model f_T $B(f_T, \mathcal{D}) \geq 0$, the algorithm must halt after at most $T \leq \frac{m}{\alpha}$ many rounds. Since each iteration decreases squared error, it must be that $B(f_T, \mathcal{D}) \leq B(f, \mathcal{D})$.

In fact, this argument is wasteful, although its form will be useful for us later when we investigate stronger forms of calibration. However for simple calibration, there is a simple one-shot algorithm that obtains perfect calibration and decreases squared error by exactly the amount of the calibration error of the original model.

Algorithm 4 One-Shot-Calibrate($f, \mathcal{D}$)

For each $v \in R(f)$ let $c(v) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y|f(x) = v])$

Output the model $\hat{f}$ defined as $\hat{f}(x) = c(f(x))$.

Theorem 8 *For any function f, One-Shot-Calibrate($f, \mathcal{D}$) (Algorithm 4 outputs a model $\hat{f}$ such that $K_2(\hat{f}) = 0$ and $B(\hat{f}) = B(f) - K_2(f)$.*

Proof 14 *Consider any level set of $\hat{f}$: $S(v) = \{x : \hat{f}(x) = v\}$. By definition, for all $x \in S(v)$, we must have $f(x) = v'$ such that $c(v') = v$ — i.e. such that $c(f(x)) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y|f(x) = v']) = v$. Let $P(v) = \{v' : c(v') = v\}$. We have that*

$$\mathop{\mathbb{E}}_{(x,y)}[y|x \in S(v)] = \frac{\sum_{v' \in P(v)} \Pr[f(x) = v']c(v')}{\sum_{v' \in P(v)} \Pr[f(x) = v']} = v$$

Hence:

$$K_2(\hat{f}) = \sum_{v \in R(\hat{f})} \Pr[\hat{f}(x) = v] \left(v - \mathbb{E}[y|x \in S(v)]\right)^2 = 0$$

Next, observe that we can decompose the squared error of both f and $\hat{f}$

according to the level sets of f, which form a partition for $\mathcal{X}$:

$$\begin{aligned} B(f,\mathcal{D}) - B(\hat{f},\mathcal{D}) &= \mathbb{E}[(f(x)-y)^2] - \mathbb{E}[(\hat{f}(x)-y)^2] \\ &= \sum_{v \in R(f)} \Pr[f(x)=v]\, \mathbb{E}[(f(x)-y)^2 - (c(f(x))-y)^2 | f(x)=v] \\ &= \sum_{v \in R(f)} \Pr[f(x)=v]\, \mathbb{E}[(v-y)^2 - (c(v)-y)^2 | f(x)=v] \\ &= \sum_{v \in R(f)} \Pr[f(x)=v](v-c(v))^2 \end{aligned}$$

where the last equality follows from Lemma 3.1.2. But:

$$\sum_{v \in R(f)} \Pr[f(x)=v](v-c(v))^2 = \sum_{v \in R(f)} \Pr[f(x)=v](v-\mathbb{E}[y|f(x)=v])^2 = K_2(f)$$

which completes the proof.

Thus we see that mis-calibrated models can always be improved: they can be efficiently updated to have *no* calibration error, and in performing this simple update, their squared error is improved by an amount equal to their initial calibration error. This also shows that squared error can be decomposed into two terms: calibration error, and the remainder (which is sometimes called *refinement* error), and that the part corresponding to calibration error can always be removed.

We will eventually be interested in calibrating predictors using a finite sample of data from a distribution (rather than giving our algorithm the ability to directly and exactly compute expectations on the distribution), which will require proving *generalization* theorems. But we will defer this to Chapter 4, when we will prove such theorems for more demanding notions of calibration.

3.3 Quantile Calibration

We can similarly define *quantile calibration* for a target quantile q, which asks that a model f produce quantiles $f(x)$ that satisfy marginal quantile consistency not just overall, but conditional on the value of $f(x)$.

Definition 12 (Average Quantile Calibration Error) *The average quantile calibration error with respect to a target quantile q of a predictor f is:*

$$Q_1(f) = \sum_{v \in R(f)} \Pr_{(x,y)\sim\mathcal{D}}[f(x)=v] \left| q - \Pr_{(x,y)\sim\mathcal{D}}[y \le v | f(x)=v] \right|$$

The average squared quantile calibration error is:

$$Q_2(f) = \sum_{v \in R(f)} \Pr_{(x,y)\sim\mathcal{D}}[f(x) = v] \left(q - \Pr_{(x,y)\sim\mathcal{D}}[y \leq v | f(x) = v] \right)^2$$

Finally, we can define a notion of maximum quantile calibration error. Just as with our average notions, we weight by the probability mass of the levelsets to avoid needing to measure quantities over sets with tiny mass:

$$Q_\infty(f) = \max_{v \in R(f)} \Pr_{(x,y)\sim\mathcal{D}}[f(x) = v] \left| q - \Pr_{(x,y)\sim\mathcal{D}}[y \leq v | f(x) = v] \right|$$

The relationship between these different measures of quantile calibration error is the same as it is for the corresponding measures of (mean) calibration error: we restate their relationship here without proof, which is identical to the case of mean calibration.

Lemma 3.3.1 *For any predictor $f : \mathcal{X} \to [0, 1]$,*

$$Q_2(f) \leq Q_1(f) \leq \sqrt{Q_2(f)}$$
$$Q_\infty(f) \leq Q_1(f) \leq m Q_\infty(f)$$

We now give an analogue to our one-shot mean calibrator. There is also an analogous iterative version — that we will build on when we study multigroup guarantees in Chapter 4 — but as with mean calibration, it has no advantages in this setting.

Algorithm 5 One-Shot-Quantile-Calibrate$(f, q, \mathcal{D})$

For each $v \in R(f)$ let

$$c(v) = \arg\min_{v'} |q - \Pr[y \leq v' | f(x) = v]|$$

Output the model $\hat{f}$ defined as $\hat{f}(x) = c(f(x))$.

Theorem 9 *For any function f, any target quantile value $q \in [0, 1]$, and any ρ-Lipschitz distribution $\mathcal{D}$, One-Shot-Quantile-Calibrate$(f, \mathcal{D})$ (Algorithm 5 outputs a model $\hat{f}$ such that $Q_2(\hat{f}) = 0$ and $PB_q(\hat{f}) \leq PB_q(f) - \frac{1}{2\rho} Q_2(f)$.*

Proof 15 *Consider any level set of $\hat{f}$: $S(v) = \{x : \hat{f}(x) = v\}$. By definition, for all $x \in S(v)$, we must have $f(x) = v'$ such that $c(v') = v$ — i.e. such that $c(f(x))$ satisfies $\Pr_{(x,y)\sim\mathcal{D}}[y \leq c(f(x)) | f(x) = v'] = q$. Let $P(v) = \{v' : c(v') = v\}$. We have that*

$$\Pr_{(x,y)}[y \leq v | x \in S(v)] = \frac{\sum_{v' \in P(v)} \Pr[f(x) = v'] \Pr[y \leq v | f(x) = v']}{\sum_{v' \in P(v)} \Pr[f(x) = v']} = q$$

Hence:

$$Q_2(\hat{f}) = \sum_{v \in R(\hat{f})} \Pr[\hat{f}(x) = v] \left(q - \Pr[y \leq v | x \in S(v)]\right)^2 = 0$$

Next, observe that we can decompose the pinball error of both f and $\hat{f}$ according to the level sets of f, which form a partition for $\mathcal{X}$:

$$\begin{aligned} PB_q(f, \mathcal{D}) - PB_q(\hat{f}, \mathcal{D}) &= \mathbb{E}[L_q(f(x), y)] - \mathbb{E}[L_q(\hat{f}(x), y)] \\ &= \sum_{v \in R(f)} \Pr[f(x) = v] \, \mathbb{E}[L_q(f(x), y) - L_q(\hat{f}(x), y) | f(x) = v] \\ &\geq \sum_{v \in R(f)} \Pr[f(x) = v] \frac{\left(\Pr[y \leq f(x) | f(x) = v] - q\right)^2}{2\rho} \\ &= \frac{1}{2\rho} Q_2(f) \end{aligned}$$

where the second to last inequality follows from Lemma 2.2.2.

3.4 Sequential Prediction

We now return to the sequential prediction setting, this time to solve a more challenging problem than simply obtaining marginal mean or quantile consistency: our goal will be to obtain empirically calibrated predictions p_t in the worst case over sequences of observations and outcomes. (x_t, y_t). We will assume that our prediction algorithm makes predictions in the discrete grid $p_t \in [1/m] = \{0, 1/m, 2/m, \ldots, 1\}$. We begin by defining empirical analogues of our calibration scores K and Q:

Definition 13 (Average Mean and Quantile Calibration Error) *Fix any transcript $\pi = \{(p_1, x_1, y_1), \ldots, (p_T, x_T, y_T)\}$ of length T. For each $p \in [1/m]$ let $n(\pi, p) = \sum_{t=1}^{T} \mathbb{1}[p_t = p]$ be the number of times that the prediction $p_t = p$ is made over the T rounds of the transcript.*

The average squared (mean) calibration error on this transcript is:

$$K_2(\pi) = \sum_{p \in [1/m]} \frac{n(\pi, p)}{T} \left(\frac{\sum_{t=1}^{T} \mathbb{1}[p_t = p](y_t - p_t)}{n(\pi, p)} \right)^2 = \frac{1}{T} \sum_{p \in [1/m]} \left(\frac{\sum_{t=1}^{T} \mathbb{1}[p_t = p](y_t - p_t)}{\sqrt{n(\pi, p)}} \right)^2$$

It will be convenient for us to be able to refer to the un-normalized *calibration error:*

$$\hat{K}_2(\pi) = \sum_{p \in [1/m]} \left(\frac{\sum_{t=1}^{T} \mathbb{1}[p_t = p](y_t - p_t)}{\sqrt{n(\pi, p)}} \right)^2$$

Observe that $K_2(\pi) = \frac{1}{T}\hat{K}_2(\pi)$.

For a target quantile $q \in [0,1]$*, the average squared quantile calibration error on this transcript is:*

$$Q_2(\pi) = \sum_{p \in [1/m]} \frac{n(\pi,p)}{T}\left(\sum_{t=1}^{T} \frac{\mathbb{1}[p_t = p](q - \mathbb{1}[y_t \leq p_t])}{n(\pi,p)}\right)^2 = \frac{1}{T}\sum_{p \in [1/m]}\left(\sum_{t=1}^{T} \frac{\mathbb{1}[p_t = p](q - \mathbb{1}[y_t \leq p_t])}{\sqrt{n(\pi,p)}}\right)^2$$

Similarly, we define the unnormalized qunatile calibration error:

$$\hat{Q}_2(\pi) = \sum_{p \in [1/m]}\left(\sum_{t=1}^{T} \frac{\mathbb{1}[p_t = p](q - \mathbb{1}[y_t \leq p_t])}{\sqrt{n(\pi,p)}}\right)^2$$

Here any term in the sum in which $n(\pi,p) = 0$ *evaluates to* 0 *by convention.*

We will derive algorithms that (based on their performance on a transcript of length $t-1$ so far, and possibly on the next context x_t) decide on their prediction p_t at round t. After they make their prediction, they learn the true label y_t, and the transcript extends by one round. We write $\pi^{<t} = \{(p_1, x_1, y_1), \ldots, (p_{t-1}, x_{t-1}, y_{t-1})\}$ to denote a transcript corresponding to rounds $1, \ldots, t-1$, and given a record of the prediction, context, and outcome at round t (p_t, x_t, y_t) write the transcript that is extended by one round as $\pi^{\leq t} = \pi^{<t+1} = \pi^{<t} \circ (p_t, x_t, y_t)$. Similarly given a transcript π of length T we will write $\pi^{\leq t}$ to denote the prefix of this transcript of length t.

3.4.1 Sequential (Mean) Calibration

In deriving an algorithm for guaranteeing sequential mean calibration, it will be helpful for us to understand how the average squared calibration score increases from round to round, given the prediction of the algorithm p_t and the outcome y_t. It will be useful for us to develop some notation.

Definition 14 *Fixing a transcript* π *of length* T*, for any* $s \leq T$ *and* $p \in [1/m]$ *define the quantity:*

$$V_s^p(\pi) = \sum_{t=1}^{s} \frac{\mathbb{1}[p_t = p](y_t - p_t)}{\sqrt{n(\pi^{\leq s}, p)}}$$

If $n(\pi,p) = 0$ *then by convention we define* $V_s^p(\pi) = 0$.

We observe that for all p, s, π:

$$|V_s^p(\pi)| \leq \sqrt{n(\pi^{\leq s}, p)}$$

Our goal is to understand how the calibration error increases from round to round as a function of the transcript — and once we understand it, give an algorithm that guarantees that the increase is small. The next lemma bounds the increase in calibration error between rounds s and $s+1$ as a function of the transcript up through round $s+1$.

Lemma 3.4.1 *Fix any partial transcript $\pi^{\leq s}$ and any triple $(p_{s+1}, x_{s+1}, y_{s+1})$ of potential outcomes for the next round. Let $\pi^{\leq s+1} = \pi^{\leq s} \circ (p_{s+1}, x_{s+1}, y_{s+1})$ be the corresponding continuation of the transcript. Define:*

$$\Delta_{s+1}(p_{s+1}, y_{s+1}) = \hat{K}_2(\pi^{\leq s+1}) - \hat{K}_2(\pi^{\leq s})$$

to be the increase in the (unnormalized) squared calibration error that results from the transcript continuation. Then we have that:

$$\Delta_{s+1}(p_{s+1}, y_{s+1}) \leq \left(\frac{2V_s^{p_{s+1}}}{\sqrt{n(\pi^{\leq s}, p_{s+1})}} \cdot (y_{s+1} - p_{s+1}) + \frac{1}{n(\pi^{\leq s}, p_{s+1})} \right)$$

Proof 16 *Since the terms in the squared mean calibration error corresponding to predictions $p \neq p_{s+1}$ do not change, We can compute:*

$$\begin{aligned}
\Delta_{s+1}(p_{s+1}, y_{s+1}) &= \hat{K}_2(\pi^{\leq s+1}) - \hat{K}_2(\pi^{\leq s}) \\
&= \left(\left(\frac{\sum_{t=1}^{s+1} \mathbb{1}[p_t = p_{s+1}](y_t - p_t)}{\sqrt{n(\pi^{\leq s+1}, p_{s+1})}} \right)^2 - \left(\frac{\sum_{t=1}^{s} \mathbb{1}[p_t = p_{s+1}](y_t - p_t)}{\sqrt{n(\pi^{\leq s}, p_{s+1})}} \right)^2 \right) \\
&\leq \left(\left(\frac{\sum_{t=1}^{s+1} \mathbb{1}[p_t = p_{s+1}](y_t - p_t)}{\sqrt{n(\pi^{\leq s}, p_{s+1})}} \right)^2 - \left(\frac{\sum_{t=1}^{s} \mathbb{1}[p_t = p_{s+1}](y_t - p_t)}{\sqrt{n(\pi^{\leq s}, p_{s+1})}} \right)^2 \right) \\
&= \left(\left(V_s^{p_{s+1}}(\pi^{\leq s}) + \frac{y_{s+1} - p_{s+1}}{\sqrt{n(\pi^{\leq s}, p_{s+1})}} \right)^2 - V_s^{p_{s+1}}(\pi^{\leq s})^2 \right) \\
&= \left(2V_s^{p_{s+1}}(\pi^{\leq s}) \left(\frac{y_{s+1} - p_{s+1}}{\sqrt{n(\pi^{\leq s}, p_{s+1})}} \right) + \frac{(y_{s+1} - p_{s+1})^2}{n(\pi^{\leq s}, p_{s+1})} \right) \\
&\leq \left(\frac{2V_s^{p_{s+1}}}{\sqrt{n(\pi^{\leq s}, p_{s+1})}} \cdot (y_{s+1} - p_{s+1}) + \frac{1}{n(\pi^{\leq s}, p_{s+1})} \right)
\end{aligned}$$

Next, our plan is to show that for *every* transcript $\pi^{\leq s}$ there is a distribution over subsequent predictions p_{s+1} such that for *every* possible realization of y_{s+1}, $\mathbb{E}_{p_{s+1}}[\Delta_{s+1}(p_{s+1}, y_{s+1})]$ is small. If we can show this, then the algorithm that consists of playing this randomized strategy at each round will have small expected calibration loss, which we can conclude simply by summing the terms $\Delta_s(p_s, y_s)$ from $s = 1$ to T.

Towards this end, define:

$$\Delta^1_{s+1}(p_{s+1}, y_{s+1}) = \frac{2V_s^{p_{s+1}}}{\sqrt{n(\pi^{\leq s}, p_{s+1})}} \cdot (y_{s+1} - p_{s+1})$$

With this notation, Lemma 3.4.1 states that:

$$\Delta_{s+1}(p_{s+1}, y_{s+1}) \leq \Delta^1_{s+1}(p_{s+1}, y_{s+1}) + \frac{1}{n(\pi^{\leq s}, p_{s+1})}.$$

Here the term $\frac{1}{n(\pi^{\leq s}, p_{s+1})}$ evaluates to 0 if $n(\pi^{\leq s}, p_{s+1}) = 0$.

We next establish a randomized prediction strategy that makes the first term $\mathbb{E}_{p_{s+1}}[\Delta^1_{s+1}(p_{s+1}, y_{s+1})]$ small in expectation.

Lemma 3.4.2 *Fix any partial transcript $\pi^{\leq s}$. Consider the distribution over p_{s+1} that we can sample from as follows:*

1. *If $V^1_s(\pi^{\leq s}) \geq 0$: Predict $p_{s+1} = 1$ with probability 1*
2. *If $V^0_s(\pi^{\leq s}) \leq 0$: Predict $p_{s+1} = 0$ with probability 1.*
3. *Otherwise: Find a $p \in \{0, \frac{1}{m}, \ldots, \frac{m-1}{m}\}$ such that $V^p_s(\pi^{\leq s}) \geq 0$ and $V^{p+1/m}_s(\pi^{\leq s}) \leq 0$. Compute $q \in [0,1]$ such that:*

$$q \cdot \frac{V^p_s(\pi^{\leq s})}{\sqrt{n(\pi^{\leq s}, p)}} + (1-q) \cdot \frac{V^{p+1/m}_s(\pi^{\leq s})}{\sqrt{n(\pi^{\leq s}, p + \frac{1}{m})}} = 0$$

Predict $p_{s+1} = p$ with probability q and predict $p_{s+1} = p + \frac{1}{m}$ with probability $1 - q$.

This distribution has the property that for every $y_{s+1} \in [0,1]$:

$$\mathop{\mathbb{E}}_{p_{s+1}}[\Delta^1_{s+1}(p_{s+1}, y_{s+1})] \leq \frac{2}{m}$$

Proof 17 *We bound $\mathbb{E}_{p_{s+1}}[\Delta^1_{s+1}(p_{s+1}, y_{s+1})]$ separately in each of the three cases.*

Case 1:

In this case, $V^1_s(\pi^{\leq s}) \geq 0$ and $p_{s+1} = 1$. Note that since $y_{s+1} \in [0,1]$, we must have that $(y_{s+1} - p_{s+1}) \leq 0$ and so for all $y_{s+1} \in [0,1]$:

$$\Delta^1_{s+1}(p_{s+1}, y_{s+1}) = \frac{2V^{p_{s+1}}_s}{\sqrt{n(\pi^{\leq s}, p_{s+1})}} \cdot (y_{s+1} - p_{s+1}) \leq 0$$

Case 2:

In this case, $V^0_s(\pi^{\leq s}) \leq 0$ and $p_{s+1} = 0$. Note that since $y_{s+1} \in [0,1]$, we must have that $(y_{s+1} - p_{s+1}) \geq 0$ and so for all $y_{s+1} \in [0,1]$:

$$\Delta^1_{s+1}(p_{s+1}, y_{s+1}) = \frac{2V^{p_{s+1}}_s}{\sqrt{n(\pi^{\leq s}, p_{s+1})}} \cdot (y_{s+1} - p_{s+1}) \leq 0$$

Case 3:

First we observe that in this case, $V^0_s(\pi^{\leq s}) \geq 0$ and $V^1_s(\pi^{\leq s}) \leq 0$. Hence there must exist some adjacent pair $p, p + 1/m \in [1/m]$ such that $V^p_s(\pi^{\leq s}) \geq 0$ and

$V_s^{p+1/m}(\pi^{\leq s}) \leq 0$, *so the algorithm is well defined. Recall that* $q \in [0,1]$ *is such that* $q \cdot \frac{V_s^p(\pi^{\leq s})}{\sqrt{n(\pi^{\leq s},p)}} + (1-q) \cdot \frac{V_s^{p+1/m}(\pi^{\leq s})}{\sqrt{n(\pi^{\leq s},p+\frac{1}{m})}} = 0$. *We can compute:*

$$\begin{aligned}
& \mathop{\mathbb{E}}_{p_{s+1}}[\Delta_{s+1}^1(p_{s+1}, y_{s+1})] \\
\leq & \left(q \cdot \frac{2V_s^p(\pi^{\leq s})}{\sqrt{n(\pi^{\leq s},p)}} \cdot (y_{s+1} - p) + (1-q)\frac{2V_s^{p+1/m}(\pi^{\leq s})}{\sqrt{n(\pi^{\leq s},p+1/m)}} \cdot \left(y_{s+1} - p - \frac{1}{m} \right) \right) \\
= & \left(-\frac{1}{m} \cdot (1-q)\frac{2V_s^{p+1/m}(\pi^{\leq s})}{\sqrt{n(\pi^{\leq s},p+1/m)}} \right) \\
\leq & \frac{2}{m}
\end{aligned}$$

Here the last inequality follows from the fact that for all $p \in [1/m]$, $|V_s^p(\pi^{\leq s})| \leq \sqrt{n(\pi^{\leq s},p)}$.

Applying the prediction strategy defined in 3.4.2 repeatedly gives us an algorithm (Algorithm 6) for making sequential predictions that are calibrated against arbitrary sequences of outcomes.

Algorithm 6 Online-Calibrated-Predictor(m)

for $t = 1$ to T **do**
  Observe x_t (and ignore it!)
  if $V_{t-1}^1(\pi^{<t}) \geq 0$ **then**
    Predict $p_t = 1$.
  else if $V_{t-1}^0(\pi^{<t}) \leq 0$ **then**
    Predict $p_t = 0$.
  else
    Select $p \in \{0, \frac{1}{m}, \ldots, \frac{m-1}{m}\}$ such that such that $V_{t-1}^p(\pi^{<t}) \geq 0$ and $V_{t-1}^{p+1/m}(\pi^{<t}) \leq 0$.
    Compute $q \in [0,1]$ such that:

$$q \cdot \frac{V_{t-1}^p(\pi^{<t})}{\sqrt{n(\pi^{<t},p)}} + (1-q) \cdot \frac{V_{t-1}^{p+1/m}(\pi^{<t})}{\sqrt{n(\pi^{<t},p+\frac{1}{m})}} = 0$$

    Predict $p_t = p$ with probability q and predict $p_t = p + \frac{1}{m}$ with probability $1-q$.
  Observe y_t
  Let $\pi^{<t+1} = \pi^{<t} \circ (x_t, p_t, y_t)$

Theorem 10 *Against any adaptive adversary, Online-Calibrated-Predictor (Algorithm 6) invoked with the range* $[1/m]$ *induces a distribution over length*

T transcripts π such that:

$$\mathop{\mathbb{E}}_{\pi}[K_2(\pi)] \leq \frac{2}{m} + \frac{m+1}{T} \cdot (\log(T/m) + 1)$$

In particular, if we choose discretization parameter $m = \sqrt{\frac{2T}{\log T}}$ then we have:

$$\mathop{\mathbb{E}}_{\pi}[K_2(\pi)] \leq O\left(\sqrt{\frac{\log T}{T}}\right)$$

Proof 18 *Fix any length T transcript $\pi = \{(x_1, p_1, y_1), \ldots, (x_T, p_T, y_T)\}$. Since $\hat{K}_2(\pi^{\leq 0}) = 0$ we have that the telescoping sum:*

$$\sum_{t=1}^{T} \Delta_t(p_t, y_t) = \sum_{t=1}^{T} \hat{K}_2(\pi^{\leq t}) - \hat{K}_2(\pi^{\leq t-1}) = \hat{K}_2(\pi)$$

From Lemma 3.4.1 we can write this as:

$$\begin{aligned}
\hat{K}_2(\pi) &= \sum_{t=1}^{T} \Delta_t(p_t, y_t) \\
&\leq \sum_{t=1}^{T} \left(\Delta_t^1(p_t, y_t) + \frac{1}{n(\pi^{\leq t-1}, p_t)} \right) \\
&\leq \sum_{t=1}^{T} \Delta_t^1(p_t, y_t) + \max_{\tilde{\pi}} \sum_{t=1}^{T} \frac{1}{n(\tilde{\pi}^{\leq t-1}, \tilde{p}_t)}
\end{aligned}$$

where in the last step, we take the maximum over all length t transcripts $\tilde{p}i = \{(\tilde{x}_1, \tilde{p}_1, \tilde{y}_1), \ldots, (\tilde{x}_T, \tilde{p}_T, \tilde{y}_T)\}$

We now take the expectation of both sides (over the randomness of the algorithm's predictions p_t) and apply Lemma 3.4.2:

$$\begin{aligned}
\mathbb{E}[\hat{K}_2(\pi)] &\leq \sum_{t=1}^{T} \mathop{\mathbb{E}}_{p_t, y_t}[\Delta_t^1(p_t, y_t) | \pi^{<t}] + \max_{\tilde{\pi}} \sum_{t=1}^{T} \frac{1}{n(\tilde{\pi}^{\leq t-1}, \tilde{p}_t)} \\
&\leq \frac{2T}{m} + \max_{\tilde{\pi}} \sum_{t=1}^{T} \frac{1}{n(\tilde{\pi}^{\leq t-1}, \tilde{p}_t)}
\end{aligned}$$

It remains to bound $\max_{\tilde{\pi}} \sum_{t=1}^{T} \frac{1}{n(\tilde{\pi}^{\leq t-1}, \tilde{p}_t)}$ To do this, we observe that whenever $\tilde{p}_t = p$, then we must have that $n(\tilde{\pi}^{\leq t}, p) = n(\tilde{\pi}^{\leq t-1}, p) + 1$. Hence for

any transcript $\tilde{\pi}$ we can write:

$$\begin{aligned}
\sum_{t=1}^{T} \frac{1}{n(\tilde{\pi}^{\leq t-1}, \tilde{p}_t)} &= \sum_{p \in [1/m]} \sum_{t:\tilde{p}_t = p} \frac{1}{n(\tilde{\pi}^{\leq t-1}, p)} \\
&= \sum_{p \in [1/m]} \sum_{k=1}^{n(\tilde{\pi},p)-1} \frac{1}{k} \\
&\leq (m+1) \sum_{k=1}^{T/m} \frac{1}{k} \\
&= (m+1) \cdot H_{T/m} \\
&\leq (m+1) \cdot (\log(T/m) + 1)
\end{aligned}$$

Here H_k denotes the k'th Harmonic number.

Combining these bounds we find that:

$$\mathbb{E}[K_2(\pi)] = \mathbb{E}\left[\frac{\hat{K}_2(\pi)}{T}\right] \leq \frac{2}{m} + \frac{m+1}{T} \cdot (\log(T/m) + 1)$$

Add high probability bound, online to offline reduction

3.4.2 Sequential Quantile Calibration

We can derive an algorithm for making sequential predictions that are quantile calibrated in an analogous way. The derivation is almost identical to our derivation of sequential mean calibration: so we will sketch it, pointing out those parts that differ. As in our batch quantile algorithm, we will need to now assume that the adversary picks continuous *distributions* of labels at each round rather than allowing her to choose deterministically, since obtaining quantile calibration is not in general possible against point mass distributions. Moreover our quantitative bound will require assuming that the adversary's label distributions are ρ-Lipschitz and will depend on ρ.

Definition 15 *Fix a target quantile $q \in [0,1]$. Fixing a transcript π of length T, for any $s \leq T$ and $p \in [1/m]$ define the quantity:*

$$W_s^p(\pi, q) = \sum_{t=1}^{s} \frac{\mathbb{1}[p_t = p](q - \mathbb{1}[y_t \leq p_t])}{\sqrt{n(\pi^{\leq s}, p)}}$$

If $n(\pi, p) = 0$ then by convention we define $W_s^p(\pi, q) = 0$. When q is clear from context we elide it and just write $W_s^p(\pi)$

We observe that for all p, s, π:

$$|W_s^p(\pi, q)| \leq \sqrt{n(\pi^{\leq s}, p)}$$

Lemma 3.4.3 *Fix any $q \in [0,1]$ and any partial transcript $\pi^{\leq s}$ and any triple $(p_{s+1}, x_{s+1}, y_{s+1})$ of potential outcomes for the next round. Let $\pi^{\leq s+1} = \pi^{\leq s} \circ (p_{s+1}, x_{s+1}, y_{s+1})$ be the corresponding continuation of the transcript. Redefine:*

$$\Delta_{s+1}(p_{s+1}, y_{s+1}) = \hat{Q}_2(\pi^{\leq s+1}) - \hat{Q}_2(\pi^{\leq s})$$

to be the increase in the (unnormalized) squared quantile calibration error that results from the transcript continuation. Then we have that:

$$\Delta_{s+1}(p_{s+1}, y_{s+1}) \leq \left(\frac{2W_s^{p_{s+1}}}{\sqrt{n(\pi^{\leq s}, p_{s+1})}} \cdot (q - \mathbb{1}[y_{s+1} \leq p_{s+1}]) + \frac{1}{n(\pi^{\leq s}, p_{s+1})} \right)$$

The proof is essentially identical to that of Lemma 3.4.1 — we include it here for completeness.

Proof 19 *Since the terms in the squared quantile calibration error corresponding to predictions $p \neq p_{s+1}$ do not change, We can compute:*

$$\begin{aligned}
& \Delta_{s+1}(p_{s+1}, y_{s+1}) \\
= \; & \hat{Q}_2(\pi^{\leq s+1}) - \hat{Q}_2(\pi^{\leq s}) \\
= \; & \left(\left(\frac{\sum_{t=1}^{s+1} \mathbb{1}[p_t = p_{s+1}](q - \mathbb{1}[y_t \leq p_t])}{\sqrt{n(\pi^{\leq s+1}, p_{s+1})}} \right)^2 - \left(\frac{\sum_{t=1}^{s} \mathbb{1}[p_t = p_{s+1}](q - \mathbb{1}[y_t \leq p_t])}{\sqrt{n(\pi^{\leq s}, p_{s+1})}} \right)^2 \right) \\
\leq \; & \left(\left(\frac{\sum_{t=1}^{s+1} \mathbb{1}[p_t = p_{s+1}](q - \mathbb{1}[y_t \leq p_t])}{\sqrt{n(\pi^{\leq s}, p_{s+1})}} \right)^2 - \left(\frac{\sum_{t=1}^{s} \mathbb{1}[p_t = p_{s+1}](q - \mathbb{1}[y_t \leq p_t])}{\sqrt{n(\pi^{\leq s}, p_{s+1})}} \right)^2 \right) \\
= \; & \left(\left(W_s^{p_{s+1}}(\pi^{\leq s}) + \frac{(q - \mathbb{1}[y_{s+1} \leq p_{s+1}])}{\sqrt{n(\pi^{\leq s}, p_{s+1})}} \right)^2 - W_s^{p_{s+1}}(\pi^{\leq s})^2 \right) \\
= \; & \left(2W_s^{p_{s+1}}(\pi^{\leq s}) \left(\frac{(q - \mathbb{1}[y_{s+1} \leq p_{s+1}])}{\sqrt{n(\pi^{\leq s}, p_{s+1})}} \right) + \frac{(q - \mathbb{1}[y_{s+1} \leq p_{s+1}])^2}{n(\pi^{\leq s}, p_{s+1})} \right) \\
\leq \; & \left(\frac{2W_s^{p_{s+1}}}{\sqrt{n(\pi^{\leq s}, p_{s+1})}} \cdot (q - \mathbb{1}[y_{s+1} \leq p_{s+1}]) + \frac{1}{n(\pi^{\leq s}, p_{s+1})} \right)
\end{aligned}$$

Next, our plan is to show that for *every* transcript $\pi^{\leq s}$ there is a distribution over subsequent predictions p_{s+1} such that for *every* possible ρ-Lipschitz distribution over y_{s+1}, $\mathbb{E}_{p_{s+1}, y_{s+1}}[\Delta_{s+1}(p_{s+1}, y_{s+1})]$ is small. Note that here we are deviating from our derivation of mean calibration algorithms, in that we are requiring that the label y_{s+1} be drawn from a ρ-Lipschitz distribution, and we are taking the expectation over y_{s+1} as well as p_{s+1}. If we can show this, then the algorithm that consists of playing this randomized strategy at each round will have small expected quantile calibration loss, which we can conclude simply by summing the terms $\Delta_s(p_s, y_s)$ from $s = 1$ to T.

Towards this end, define:

$$\Delta^1_{s+1}(p_{s+1}, y_{s+1}) = \frac{2W_s^{p_{s+1}}}{\sqrt{n(\pi^{\leq s}, p_{s+1})}} \cdot (q - \mathbb{1}[y_{s+1} \leq p_{s+1}])$$

With this notation, Lemma 3.4.3 states that:

$$\Delta_{s+1}(p_{s+1}, y_{s+1}) \leq \Delta^1_{s+1}(p_{s+1}, y_{s+1}) + \frac{1}{n(\pi^{\leq s}, p_{s+1})}.$$

Here the term $\frac{1}{n(\pi^{\leq s}, p_{s+1})}$ evaluates to 0 if $n(\pi^{\leq s}, p_{s+1}) = 0$.

We next establish a randomized prediction strategy that makes the first term $\mathbb{E}_{p_{s+1}, y_{s+1}}[\Delta^1_{s+1}(p_{s+1}, y_{s+1})]$ small in expectation for any ρ-Lipchitz distribution over y_{s+1}.

Lemma 3.4.4 *Fix any partial transcript $\pi^{\leq s}$. Consider the distribution over p_{s+1} that we can sample from as follows:*

1. *If $W_s^1(\pi^{\leq s}) \geq 0$: Predict $p_{s+1} = 1$ with probability 1*
2. *If $W_s^0(\pi^{\leq s}) \leq 0$: Predict $p_{s+1} = 0$ with probability 1.*
3. *Otherwise: Find a $p \in \{0, \frac{1}{m}, \ldots, \frac{m-1}{m}\}$ such that $W_s^p(\pi^{\leq s}) \geq 0$ and $W_s^{p+1/m}(\pi^{\leq s}) \leq 0$. Compute $b \in [0,1]$ such that:*

$$b \cdot \frac{W_s^p(\pi^{\leq s})}{\sqrt{n(\pi^{\leq s}, p)}} + (1-b) \cdot \frac{W_s^{p+1/m}(\pi^{\leq s})}{\sqrt{n(\pi^{\leq s}, p + \frac{1}{m})}} = 0$$

Predict $p_{s+1} = p$ with probability b and predict $p_{s+1} = p + \frac{1}{m}$ with probability $1 - b$.

This distribution has the property that for every ρ-Lipschitz distribution over $y_{s+1} \in [0,1]$:

$$\mathop{\mathbb{E}}_{p_{s+1}, y_{s+1}}[\Delta^1_{s+1}(p_{s+1}, y_{s+1})] \leq \frac{2\rho}{m}$$

Proof 20 *We bound $\mathbb{E}_{p_{s+1}, y_{s+1}}[\Delta^1_{s+1}(p_{s+1}, y_{s+1})]$ separately in each of the three cases.*

Case 1:

In this case, $W_s^1(\pi^{\leq s}) \geq 0$ and $p_{s+1} = 1$. Note that since $q, y_{s+1} \in [0,1]$, we must have that $(q - \mathbb{1}[y_{s+1} \leq p_{s+1}]) \leq 0$ and so for all $y_{s+1} \in [0,1]$:

$$\Delta^1_{s+1}(p_{s+1}, y_{s+1}) = \frac{2W_s^{p_{s+1}}}{\sqrt{n(\pi^{\leq s}, p_{s+1})}} \cdot (q - \mathbb{1}[y_{s+1} \leq p_{s+1}]) \leq 0$$

Case 2:

In this case, $W_s^0(\pi^{\leq s}) \leq 0$ *and* $p_{s+1} = 0$*. Note that since* $q, y_{s+1} \in [0,1]$*, we must have that if* $y_{s+1} > 0$ *(which occurs with probability 1 if it is drawn from a continuous distribution),* $(q - \mathbb{1}[y_{s+1} \leq p_{s+1}]) \geq 0$ *and so for all* $q, y_{s+1} \in (0,1]$*:*

$$\Delta^1_{s+1}(p_{s+1}, y_{s+1}) = \frac{2W_s^{p_{s+1}}}{\sqrt{n(\pi^{\leq s}, p_{s+1})}} \cdot (q - \mathbb{1}[y_{s+1} \leq p_{s+1}]) \leq 0$$

Case 3:

First we observe that in this case, $W_s^0(\pi^{\leq s}) \geq 0$ *and* $W_s^1(\pi^{\leq s}) \leq 0$*. Hence there must exist some adjacent pair* $p, p+1/m \in [1/m]$ *such that* $W_s^p(\pi^{\leq s}) \geq 0$ *and* $W_s^{p+1/m}(\pi^{\leq s}) \leq 0$*, so the algorithm is well defined. Recall that* $b \in [0,1]$ *is such that* $b \cdot \frac{W_s^p(\pi^{\leq s})}{\sqrt{n(\pi^{\leq s}, p)}} + (1-b) \cdot \frac{W_s^{p+1/m}(\pi^{\leq s})}{\sqrt{n(\pi^{\leq s}, p+\frac{1}{m})}} = 0$*. We can compute:*

$$\begin{aligned}
& \mathop{\mathbb{E}}_{p_{s+1}, y_{s+1}}[\Delta^1_{s+1}(p_{s+1}, y_{s+1})] \\
\leq\ & \left(b \cdot \frac{2W_s^p(\pi^{\leq s})}{\sqrt{n(\pi^{\leq s}, p)}} \cdot (q - \Pr[y_{s+1} \leq p]) + (1-b) \frac{2W_s^{p+1/m}(\pi^{\leq s})}{\sqrt{n(\pi^{\leq s}, p+1/m)}} \cdot \left(q - \Pr[y_{s+1} \leq p + \frac{1}{m}] \right) \right) \\
\leq\ & \left(b \cdot \frac{2W_s^p(\pi^{\leq s})}{\sqrt{n(\pi^{\leq s}, p)}} \cdot (q - \Pr[y_{s+1} \leq p]) + (1-b) \frac{2W_s^{p+1/m}(\pi^{\leq s})}{\sqrt{n(\pi^{\leq s}, p+1/m)}} \cdot \left(q - \Pr[y_{s+1} \leq p] - \frac{\rho}{m} \right) \right) \\
=\ & \left(-\frac{\rho}{m} \cdot (1-b) \frac{2W_s^{p+1/m}(\pi^{\leq s})}{\sqrt{n(\pi^{\leq s}, p+1/m)}} \right) \\
\leq\ & \frac{2\rho}{m}
\end{aligned}$$

Here the second inequality follows from the fact that the distribution over y_{s+1} *is assumed to be* ρ*-Lipschitz, and hence for all* p*:*

$$\Pr_{y_{s+1}}\left[y_{s+1} \leq p + \frac{1}{m} \right] \leq \Pr_{y_{s+1}}[y_{s+1} \leq p] + \frac{\rho}{m}$$

The last inequality follows from the fact that for all $p \in [1/m]$*,* $|W_s^p(\pi^{\leq s})| \leq \sqrt{n(\pi^{\leq s}, p)}$*.*

Applying the prediction strategy defined in 3.4.4 repeatedly gives us an algorithm (Algorithm 7) for making sequential predictions that are calibrated against arbitrary sequences of outcomes.

Algorithm 7 Online-Quantile-Calibrated-Predictor(q, m)

for $t = 1$ to T **do**
  Observe x_t (and ignore it!)
  if $W_s^1(\pi^{<t}, q) \geq 0$ **then**
    Predict $p_t = 1$.
  else if $W_s^0(\pi^{<t}, q) \leq 0$ **then**
    Predict $p_t = 0$.
  else
    Select $p \in \{0, \frac{1}{m}, \ldots, \frac{m-1}{m}\}$ such that such that $W_s^p(\pi^{\leq s}, q) \geq 0$ and $W_s^{p+1/m}(\pi^{\leq s}, q) \leq 0$.
    Compute $b \in [0, 1]$ such that:

$$b \cdot \frac{W_s^p(\pi^{\leq s}, q)}{\sqrt{n(\pi^{\leq s}, p)}} + (1 - b) \cdot \frac{W_s^{p+1/m}(\pi^{\leq s}, q)}{\sqrt{n(\pi^{\leq s}, p + \frac{1}{m})}} = 0$$

    Predict $p_{s+1} = p$ with probability b and predict $p_{s+1} = p + \frac{1}{m}$ with probability $1 - b$.
  Observe y_t
  Let $\pi^{<t+1} = \pi^{<t} \circ (x_t, p_t, y_t)$

Theorem 11 *Against any adaptive adversary that chooses a ρ-Lipschitz distribution over y_t at each round t, Online-Quantile-Calibrated-Predictor (Algorithm 7) invoked with quantile $q \in [0, 1]$ and the range $[1/m]$ induces a distribution over length T transcripts π such that:*

$$\mathop{\mathbb{E}}_{\pi}[Q_2(\pi)] \leq \frac{2\rho}{m} + \frac{m+1}{T} \cdot (\log(T/m) + 1)$$

In particular, if we choose discretization parameter $m = \sqrt{\frac{2\rho T}{\log T}}$ then we have:

$$\mathop{\mathbb{E}}_{\pi}[Q_2(\pi)] \leq O\left(\sqrt{\frac{\rho \log T}{T}}\right)$$

Proof 21 *Fix any length T transcript $\pi = \{(x_1, p_1, y_1), \ldots, (x_T, p_T, y_T)\}$. Since $\hat{Q}_2(\pi^{\leq 0}) = 0$ we have that the telescoping sum:*

$$\sum_{t=1}^{T} \Delta_t(p_t, y_t) = \sum_{t=1}^{T} \hat{Q}_2(\pi^{\leq t}) - \hat{Q}_2(\pi^{\leq t-1}) = \hat{Q}_2(\pi)$$

From Lemma 3.4.3 we can write this as:

$$\begin{aligned}\hat{Q}_2(\pi) &= \sum_{t=1}^{T} \Delta_t(p_t, y_t) \\ &\leq \sum_{t=1}^{T} \left(\Delta_t^1(p_t, y_t) + \frac{1}{n(\pi^{\leq t-1}, p_t)} \right) \\ &\leq \sum_{t=1}^{T} \Delta_t^1(p_t, y_t) + \max_{\tilde{\pi}} \sum_{t=1}^{T} \frac{1}{n(\tilde{\pi}^{\leq t-1}, \tilde{p}_t)}\end{aligned}$$

where in the last step, we take the maximum over all length t transcripts $\tilde{p}i = \{(\tilde{x}_1, \tilde{p}_1, \tilde{y}_1), \ldots, (\tilde{x}_T, \tilde{p}_T, \tilde{y}_T)\}$

We now take the expectation of both sides (over the randomness of the algorithm's predictions p_t) and apply Lemma 3.4.4:

$$\begin{aligned}\mathbb{E}[\hat{Q}_2(\pi)] &\leq \sum_{t=1}^{T} \mathop{\mathbb{E}}_{p_t, y_t} [\Delta_t^1(p_t, y_t) | \pi^{<t}] + \max_{\tilde{\pi}} \sum_{t=1}^{T} \frac{1}{n(\tilde{\pi}^{\leq t-1}, \tilde{p}_t)} \\ &\leq \frac{2\rho T}{m} + \max_{\tilde{\pi}} \sum_{t=1}^{T} \frac{1}{n(\tilde{\pi}^{\leq t-1}, \tilde{p}_t)}\end{aligned}$$

It remains to bound $\max_{\tilde{\pi}} \sum_{t=1}^{T} \frac{1}{n(\tilde{\pi}^{\leq t-1}, \tilde{p}_t)}$ To do this, we observe that whenever $\tilde{p}_t = p$, then we must have that $n(\tilde{\pi}^{\leq t}, p) = n(\tilde{\pi}^{\leq t-1}, p) + 1$. Hence for any transcript $\tilde{p}i$ we can write:

$$\begin{aligned}\sum_{t=1}^{T} \frac{1}{n(\tilde{\pi}^{\leq t-1}, \tilde{p}_t)} &= \sum_{p \in [1/m]} \sum_{t : \tilde{p}_t = p} \frac{1}{n(\tilde{\pi}^{\leq t-1}, p)} \\ &= \sum_{p \in [1/m]} \sum_{k=1}^{n(\tilde{\pi}, p)-1} \frac{1}{k} \\ &\leq (m+1) \sum_{k=1}^{T/m} \frac{1}{k} \\ &= (m+1) \cdot H_{T/m} \\ &\leq (m+1) \cdot (\log(T/m) + 1)\end{aligned}$$

Here H_k denotes the k'th Harmonic number.

Combining these bounds we find that:

$$\mathbb{E}[Q_2(\pi)] = \mathbb{E}\left[\frac{\hat{Q}_2(\pi)}{T}\right] \leq \frac{2\rho}{m} + \frac{m+1}{T} \cdot (\log(T/m) + 1)$$

Add high probability bound, online to offline reduction

References and Further Reading

Algorithm 3 (our calibration algorithm) takes inspiration from the *multicalibration* algorithms given in Hébert-Johnson et al. [2018] (which bounds $K_\infty(f)$) and Gopalan et al. [2022b] (which bounds $K_1(f)$) (See Chapter 4 for more on multicalibration).

Calibration in the sequential setting has a long history dating back to Dawid [1982] and Dawid [1985]. The first algorithm that guaranteed worst-case calibration in the sequential setting was given by Foster and Vohra [1998] and alternative derivations were given by Foster [1999], Fudenberg and Levine [1999], Hart [2020], and others. Algorithm 6 is a variant of the algorithm given by Foster and Hart [2021] and its generalization to multicalibration given in Gupta et al. [2022]. Algorithm 7 is a variant of the online quantile multicalibration algorithm given by Bastani et al. [2022].

4

Multigroup Guarantees

CONTENTS

Marginal guarantees are easy to obtain, but very weak. We saw one way of strengthening those guarantees: calibration. But on its own calibration is also quite weak. Obtaining it in the adversarial sequential prediction setting was non-trivial, but we could obtain it in the batch setting with a simple constant predictor $\hat{f}(x) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y]$ that just predicts the mean of the marginal label distribution. Moreover, all of the techniques we've seen so far *entirely ignore the features x and depend only on the labels y*! We'll now consider a different way to strengthen marginal guarantees, first on its own, and then together with calibration. We will call these *multi-group* guarantees, and they ask for guarantees that hold conditional on the features x in various ways.

Let $\mathcal{G} \in 2^{\mathcal{X}}$ denote a collection of *groups* or *subsets* of the data domain $\mathcal{X}$. We will represent groups using their indicator functions: so $g \in \mathcal{G}$ is represented as a function $g : \mathcal{X} \to \{0, 1\}$, where $g(x) = 1$ denotes that $x \in \mathcal{X}$ is a member of group g, and $g(x) = 0$ denotes that x is not a member of $\mathcal{G}$. Given an example $x \in \mathcal{X}$, we will write $\mathcal{G}(x) = \{g \in \mathcal{G} : g(x) = 1\}$ to denote the set of groups that x is a member of. At a high level, our aim will be to obtain guarantees like mean consistency (and eventually calibration) not just

marginally, but *conditionally* on $g(x) = 1$ for every $g \in \mathcal{G}$ for some large set $\mathcal{G}$.

4.1 Group Conditional Mean Consistency

With only a finite number of samples from the distribution, we will not in general be able to provide group conditional guarantees conditional on groups that have tiny probability under our distribution, simply because we won't have seen very many points from this part of the probability space. So, the probability mass of a group will be a key parameter for us:

Definition 16 *Under a distribution $\mathcal{D}$, a group $g : \mathcal{X} \to \{0, 1\}$ has probability mass $\mu(g)$ defined as:*

$$\mu(g) = \Pr_{x \sim \mathcal{D}_{\mathcal{X}}}[g(x) = 1]$$

Definition 17 *A model $f : \mathcal{X} \to [0, 1]$ satisfies α-approximate group conditional mean consistency with respect to a set of groups $\mathcal{G} \in 2^{\mathcal{X}}$ if for every $g \in \mathcal{G}$:*

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

Notice that our requirement smoothly becomes less demanding as the measure of the group g grows smaller, allowing us to ask for stronger guarantees for groups for which we will have more data. We have parameterized things so that the scaling is at the right rate: the error within a sub-group g increases at a rate of $1/\sqrt{\mu(g)}$, which is the same rate at which the error of our best estimate of $\mathbb{E}_{(x,y)\sim\mathcal{D}}[y|g(x) = 1]$ from the data will necessarily increase.

We will now show how to update a model f that does not satisfy group conditional mean consistency to one that does, using a sequence of "patches" that are similar to how we obtained calibration. Just as in the examples we have seen thus far, these patches will be accuracy improving, and so we will quickly converge to a group conditional mean consistent model.

Definition 18 (Group Shift Patch) *Given a model f, a shift $\Delta \in \mathbb{R}$, and a group $g : \mathcal{X} \to \{0, 1\}$ we say that the* group patch *applied to f with shift Δ and group g is the function:*

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

Algorithm 8 GroupShift$(f, \alpha, \mathcal{G})$

Let $f_0 = f$ and $t = 0$.
while f_t does not satisfy α-approximate group conditional mean consistency w.r.t. $\mathcal{G}$: **do**
Let:

$$g_t \in \arg\max_{g \in \mathcal{G}} \mu(g) \left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f_t(x)|g(x)=1] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y|g(x)=1] \right)^2$$

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

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

Lemma 4.1.1 *Fix any model $f_t : \mathcal{X} \to [0,1]$ and group $g : \mathcal{X} \to \{0,1\}$. 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

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

(i.e. the update performed at Round t of Algorithm 8). Then:

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

Proof 22 *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:*

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

Theorem 12 *Given any model f, any collection of groups $\mathcal{G}$, and any $\alpha > 0$ Algorithm 8 (GroupShift) halts after $T \leq 1/\alpha$ many rounds and outputs a model f_T that satisfies α-approximate group conditional mean consistency. Moreover, if the algorithm runs for T rounds, then $B(f_T) \leq B(f) - T\alpha$.*

Proof 23 *At any round T at which the algorithm halts, by the stopping condition of the algorithm it must be that f_T satisfies α-approximate group conditional mean consistency. It remains to bound T and $B(f_T)$.*

Consider any intermediate round $t < T$ of the algorithm. We know since the algorithm has not halted that:

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

g_t realizes this maximum, so we must have:

$$\mu(g_t) \cdot \Delta_t^2 \geq \alpha$$

Thus by Lemma 4.1.1, $B(f_{t+1}) \leq B(f_t) - \alpha$. Inductively applying this claim gives $B(f_T) \leq B(f) - T\alpha$ as desired.

Since (by assumption) $y, f(x) \in [0,1]$, we have that $B(f_T) \geq 0$ and $B(f) \leq 1$. Thus we must have that $T \leq 1/\alpha$.

Unlike marginal mean consistency, group conditioned mean consistency is clearly a non-trivial promise: if $\mathcal{G} = 2^{\mathcal{X}}$, the set of all subsets, and $\alpha = 0$, then the only model satisfying α-approximate group conditional mean consistency with respect to $\mathcal{G}$ is the model encoding true conditional label distributions f^*. For smaller collections of groups $\mathcal{G}$ and larger values of α we have a necessarily weaker guarantee, but at least we have a parametric family of guarantees that allows us to interpolate between one satisfied by a (trivial) constant function, and one only satisfied by a perfect model.

4.2 Group Conditional Quantile Consistency

Our algorithm and analysis for group conditional mean consistency directly translates to quantiles when we replace the role of the Brier score in our analysis with Pinball loss. First, we can define an analogous notion for group conditional quantile consistency:

Definition 19 *A model $f : \mathcal{X} \to [0,1]$ satisfies α-approximate group conditional quantile consistency with respect to a target quantile q and set of groups $\mathcal{G} \in 2^{\mathcal{X}}$ if for every $g \in \mathcal{G}$:*

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

Our algorithm proceeds by applying the same kind of group-shift patches we used in the case of group conditional mean consistency.

Algorithm 9 QuantileGroupShift($f, \alpha, \mathcal{G}, q$)

Let $f_0 = f$ and $t = 0$.

while f_t does not satisfy α-approximate group conditional quantile consistency w.r.t. target quantile q and $\mathcal{G}$: **do**

Let:

$$g_t \in \arg\max_{g \in \mathcal{G}} \mu(g_t) \left(\Pr_{(x,y)\sim\mathcal{D}}[y \leq f_t(x) | g(x) = 1] - q \right)^2$$

$$\Delta_t = \arg\min_{\Delta} \left(\Pr_{(x,y)\sim\mathcal{D}}[y \leq f_t(x) + \Delta | g_t(x) = 1] - q \right)^2$$

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

Output f_t.

Theorem 13 *Assume that $\mathcal{D}$ is a ρ-Lipschitz continuous probability distribution. Given any model f, any collection of groups $\mathcal{G}$, any target quantile $q \in [0, 1]$ and any $\alpha > 0$ Algorithm 9 (QuantileGroupShift) halts after T rounds where:*

$$T \leq \frac{2\rho PB_q(f)}{\alpha} \leq \frac{2\rho}{\alpha}.$$

It outputs a model f_T that satisfies α-approximate group conditional quantile consistency. Moreover, if the algorithm runs for T rounds, then $PB_q(f_T) \leq PB_q(f) - T \cdot \frac{\alpha}{2\rho}$.

Proof 24 *If the algorithm halts at round T, then by definition of the halting condition it must be that f_T satisies α-approximate group conditional quantile consistency with respect to q and $\mathcal{G}$, so it remains to bound T.*

If the algorithm has not halted at round t, then by definition it must be that g_t satisfies:

$$\mu(g_t) \cdot \left(\Pr_{(x,y)\sim\mathcal{D}}[y \leq f(x) | g(x) = 1] - q \right)^2 \geq \alpha$$

Since $\mathcal{D}$ is continuous, it must be that Δ_t is such that:

$$\Pr_{(x,y)\sim\mathcal{D}}[y \leq f(x) + \Delta_t | g_t(x) = 1] = q$$

Finally by the separability of Pinball loss, we have that:

$$\begin{aligned}
& PB_q(f_t) - PB_q(f_{t+1}) \\
= \; & \Pr[g_t(x) = 1] \cdot \left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [L_q(f_t(x), y) - L_q(f_{t+1}(x), y) | g_t(x) = 1] \right) \\
\geq \; & \mu(g_t) \cdot \frac{\alpha}{2\rho\mu(g_t)} \\
= \; & \frac{\alpha}{2\rho}
\end{aligned}$$

where the inequality follows from Lemma 2.2.2 applied to the conditional distribution $\mathcal{D}|(g_t(x) = 1)$, which must also be ρ-smooth.

Applying this bound iteratively, we have that for every T, $PB_q(f_T) \leq PB_q(f) - T \cdot \frac{\alpha}{2\rho}$. Since when $f(x)$ and y are bounded in $[0, 1]$, $PB_q(f) \leq 1$ and $PB_q(f_T) \geq 0$ it must be that the total number of iterations that the algorithm runs for is bounded by:

$$T \leq \frac{2\rho PB_q(f)}{\alpha} \leq \frac{2\rho}{\alpha}$$

4.2.1 A More Direct Approach to Group Conditional Guarantees

Algorithms 8 and 9 gave us a relatively simple method for obtaining approximate group conditional mean and quantile consistency respectively. These algorithms will be a useful template for our algorithms for mean and quantile multicalibration in the next section — but it turns out that for group conditional consistency (without calibration) there is an even simpler algorithm that gives an even better guarantee. Observe that the "group shift" patches $h(x, f_t; g_t, \Delta_t)$ that Algorithms 8 and 9 apply have an extremely simple form: They add Δ_t to the output of $f_t(x)$ if $x \in \mathcal{G}$ and do nothing otherwise. Since addition is commutative, we can observe that these patches are actually order invariant! Consider any run of Algorithm 8 or 9 for T rounds, and for each group $g \in \mathcal{G}$ define the quantities:

$$\lambda_g = \sum_{t:g_t=g} \Delta_t$$

Then the final model f_T that is output can be seen to have the form:

$$f_T(x) = \hat{f}(x; \lambda) \equiv f(x) + \sum_{g \in \mathcal{G}} \lambda_g \cdot g(x)$$

So to compute a model satisfying group conditional mean consistency, we can just directly optimize over functions that have this form, which is a $|\mathcal{G}|$ dimensional convex optimization described in Algorithm 10 for group conditional mean consistency and Algorithm 11 for group conditional quantile consistency. The only difference between the two algorithms is that we minimize the Brier score for mean consistency and the pinball loss for quantile consistency.

Algorithm 10 Simple-Group-Conditional($f, \mathcal{G}$)

Let λ^* be a solution to the optimization problem:

$$\text{Minimize}_\lambda \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\left(\hat{f}(x;\lambda) - y\right)^2\right]$$

Such that:

$$\hat{f}(x;\lambda) \equiv f(x) + \sum_{g\in\mathcal{G}} \lambda_g \cdot g(x)$$

Output $\hat{f}(x;\lambda^*)$

Algorithm 11 Simple-Quantile-Group-Conditional($f, \mathcal{G}, q$)

Let λ^* be a solution to the optimization problem:

$$\text{Minimize}_\lambda \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[L_q\left(\hat{f}(x;\lambda), y\right)\right]$$

Such that:

$$\hat{f}(x;\lambda) \equiv f(x) + \sum_{g\in\mathcal{G}} \lambda_g \cdot g(x)$$

Output $\hat{f}(x;\lambda^*)$

Theorem 14 *Fix any model $f : \mathcal{X} \to [0,1]$ and class of groups $\mathcal{G}$. The model $\hat{f}(x;\lambda^*)$ output by Algorithm 10 satisfies perfect (i.e. 0-approximate) group conditional mean consistency. Moreover, if f_T is the model output by Algorithm 8, then $B(\hat{f}(\cdot;\lambda^*)) \leq B(f_T)$.*

Proof 25 *Suppose $\hat{f}(x;\lambda^*)$ does not satisfy group conditional mean consistency. Then there must be a group $g \in \mathcal{G}$ such that:*

$$\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f(x)|g(x)=1] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y|g(x)=1]\right)^2 > 0$$

Let $\Delta = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y|g(x)=1] - \mathbb{E}_{(x,y)\sim\mathcal{D}}[\hat{f}(x;\lambda^*)|g(x)=1]$ and note that $\Delta \neq 0$. In this case, by Lemma 4.1.1 the model obtained by applying the same patch as in the update rule in Algorithm 8 — i.e. $f'(x) = h(x, \hat{f}(x;\lambda^*), g, \Delta)$ is such that $B(f') < B(\hat{f}(x;\lambda^*))$. But this is a contradition to the optimality of λ^*. Let $\hat{\lambda}$ be the vector such that for all $g' \neq g$, $\hat{\lambda}_{g'} = \lambda^*_{g'}$ and such that $\hat{\lambda}_g = \lambda^*_g + \Delta$. We can write f' as $f'(x) = \hat{f}(x;\hat{\lambda})$. Since $\hat{\lambda}$ is a feasible solution to the optimization problem in Algorithm 10 — by the optimality of λ^* we must have $B(f(x;\hat{\lambda})) \geq B(f(x;\lambda^*))$.

Similarly, since f_T can be represented as $\hat{f}(x;\lambda)$ for some λ, we have $B(f_T) \geq B(f(x;\lambda^*))$.

Theorem 15 *Fix any model $f : \mathcal{X} \to [0,1]$, target quantile q, and class of groups $\mathcal{G}$. The model $\hat{f}(x;\lambda^*)$ output by Algorithm 11 satisfies perfect (i.e. 0-approximate) group conditional quantile consistency with respect to q and $\mathcal{G}$. Moreover, if f_T is the model output by Algorithm 9, then $PB_q(\hat{f}(\cdot;\lambda^*)) \leq PB_q(f_T)$.*

Proof 26 *Suppose $\hat{f}(x;\lambda^*)$ does not satisfy group conditional quantile consistency. Then there must be a group $g \in \mathcal{G}$ such that:*

$$\Delta = \arg\min_{\Delta} \left(\Pr_{(x,y)\sim\mathcal{D}}[y \leq f_t(x) + \Delta | g(x) = 1] - q \right)^2 > 0$$

In this case, by Lemma 2.2.1 applied to the distribution $\mathcal{D}|(g(x)=1)$, the model obtained by applying the same patch as in the update rule in Algorithm 9 — i.e. $f'(x) = h(x, \hat{f}(x;\lambda^*), g, \Delta)$ is such that $PB_q(f') < PB_q(\hat{f}(x;\lambda^*))$. But this is a contradition to the optimality of λ^*. Let $\hat{\lambda}$ be the vector such that for all $g' \neq g$, $\hat{\lambda}_{g'} = \lambda^*_{g'}$ and such that $\hat{\lambda}_g = \lambda^*_g + \Delta$. We can write f' as $f'(x) = \hat{f}(x;\hat{\lambda})$. Since $\hat{\lambda}$ is a feasible solution to the optimization problem in Algorithm 11 — by the optimality of λ^* we must have $PB_q(f(x;\hat{\lambda})) \geq PB_q(f(x;\lambda^*))$.

Similarly, since f_T can be represented as $\hat{f}(x;\lambda)$ for some λ, we have $PB_q(f_T) \geq PB_q(f(x;\lambda^*)$.

4.2.1.1 Generalization

What about out of sample guarantees — i.e. what if we run algorithms 10 and 11 on the empirical distributions on datasets $D \sim \mathcal{D}^n$?

Our generalization theorem will depend on the norm of the solution λ^* output by our algorithms, so it will be helpful for us to study a *regularized* version of these simple algorithms that is guaranteed to output a solution of small norm.

Definition 20 *For any vector $v \in \mathbb{R}^d$, the ℓ_1 norm is defined as:*

$$||v||_1 = \sum_{i=1}^{d} |v_i|$$

Algorithm 12 Simple-Group-Conditional-Regularized($f, \mathcal{G}, \mathcal{D}, \eta$)

Let λ^* be a solution to the optimization problem:

$$\text{Minimize}_\lambda \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\left(\hat{f}(x;\lambda) - y\right)^2\right] + \eta||\lambda||_1$$

Such that:

$$\hat{f}(x;\lambda) \equiv f(x) + \sum_{g\in\mathcal{G}} \lambda_g \cdot g(x)$$

Output $\hat{f}(x;\lambda^*)$

Algorithm 12 is identical to Algorithm 10, except that its objective function has been augmented with the regularization term $\eta||\lambda||_1$, where η is a parameter of the algorithm. The reason to add this regularization term is to guarantee that the output parameters λ^* will have small norm:

Lemma 4.2.1 *Let $f : \mathcal{X} \to [0,1]$ be any model with range $[0,1]$, let $\mathcal{G}$ be any set of groups, let $\mathcal{D}$ be any distribution over labelled example, and let $\eta > 0$. Then Simple-Group-Conditional-Regularized($f, \mathcal{G}, \mathcal{D}, \eta$) (Algorithm 12) outputs a model $\hat{f}(x, \lambda^*)$ with:*

$$||\lambda^*||_1 \leq \frac{1}{\eta}$$

Proof 27 *Suppose otherwise, and we have that $||\lambda^*||_1 > \frac{1}{\eta}$. Then since squared error is non-negative, we must have that:*

$$\begin{aligned}\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\left(\hat{f}(x;\lambda^*) - y\right)^2\right] + \eta||\lambda^*||_1 &\geq \eta||\lambda^*||_1 \\ &> 1\end{aligned}$$

On the other hand, consider the candidate solution λ_0, where $\lambda_0 = 0^{|\mathcal{G}|}$ is the all 0's vector. Since the squared error of f is bounded by 1 (since f has range in $[0,1]$), we have that:

$$\begin{aligned}\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\left(\hat{f}(x;\lambda_0) - y\right)^2\right] + \eta||\lambda_0||_1 &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\left(f(x) - y\right)^2\right] + \eta||\lambda_0||_1 \\ &\leq 1 + \eta||\lambda_0||_1 \\ &= 1\end{aligned}$$

Thus $\hat{f}(x;\lambda_0)$ has lower objective value than $\hat{f}(x;\lambda^*)$, contradicting the optimality of λ^*.

Ok — so Algorithm 12 produces solutions of small norm. How many small norm solutions are there anyhow? Obviously there are continuously many, so we need a more refined way to ask this question. To do this lets define an ϵ-net.

Definition 21 *Let $B(C,d) = \{x \in \mathbb{R}^d : ||x||_1 \leq C\}$ denote the d dimensional ℓ_1 ball of radius C. Let $N_\epsilon(C,d) \subset B(C,d)$ be some finite subset of the ball. We say that $N_\epsilon(C,d)$ is an ℓ_1 ϵ-net for $B(C,d)$ if for every $x \in B(C,d)$ there is an $x' \in N_\epsilon(C,d)$ such that $||x - x'||_1 \leq \epsilon$.*

Theorem 16 *There is a finite ℓ_1 ϵ-net for $B(C,d)$ that has cardinality:*

$$|N_\epsilon(C,d)| \leq \left(1 + \frac{2C}{\epsilon}\right)^d$$

Proof 28 *Let $N_\epsilon \subset B(C,d)$ be a maximal subset of points that are ϵ-separated — i.e. such that for all $\lambda, \lambda' \in N_\epsilon$, $||\lambda - \lambda'||_1 \geq \epsilon$, and such that no other point from $B(C,d)$ can be added to N_ϵ while maintaining this property. Observe that N_ϵ must be an ϵ-net for $B(C,d)$, since if there were any point $\lambda^* \in B(C,d)$ such that for all $\lambda \in N_\epsilon$, $||\lambda - \lambda^*||_1 > \epsilon$, then λ^* could be added to N_ϵ while preserving its ϵ-separation property, which would contradict its maximality.*

Consider the union of ℓ_1 balls of radius $\epsilon/2$ centered at each point $\lambda \in N_\epsilon$. Because of the ϵ-separation property of N_ϵ, these balls are disjoint, and so their total volume is the sum of their individual volumes: $|N_\epsilon| \cdot V^d_{\epsilon/2}$, where $V^d_{\epsilon/2}$ is the volume of a d-dimensional ℓ_1 ball of radius $\epsilon/2$. On the other hand, the union of these balls are all contained within a ball of radius $C + \epsilon/2$. Hence:

$$|N_\epsilon| \cdot V^d_{\epsilon/2} \leq V^d_{C+\epsilon/2}$$

and in particular:

$$\begin{aligned} |N_\epsilon| &\leq \frac{V^d_{C+\epsilon/2}}{V^d_{\epsilon/2}} \\ &\leq \left(\frac{C + \frac{\epsilon}{2}}{\frac{\epsilon}{2}}\right)^d \\ &= \left(\frac{2C}{\epsilon} + 1\right)^d \end{aligned}$$

What good is an ϵ-net? It is useful for two reasons. Since it is finite, we can use Hoeffding's inequality together with a union bound to argue that for every parameter vector λ' in the net, our in and out-of-sample objective values are close. And what about for parameter vectors λ that aren't in the net? We argue that they take objective value close to the objective value of the closest parameter vector λ' in the net.

Lemma 4.2.2 *Let $\lambda, \lambda' \in B(C, |\mathcal{G}|)$ be such that $||\lambda - \lambda'||_1 \leq \epsilon$. Then for all x, y, we have that:*

$$(f(x;\lambda) - y)^2 - (f(x;\lambda') - y)^2 \leq 2\epsilon C$$

Proof 29 *We can write:*

$$\begin{aligned}(f(x;\lambda)-y)^2-(f(x;\lambda')-y)^2 &= f(x;\lambda)^2-f(x;\lambda')^2+2y(f(x;\lambda')-f(x;\lambda))\\ &= (f(x;\lambda)-f(x;\lambda'))(f(x;\lambda)+f(x;\lambda')-2y)\\ &\leq ||\lambda-\lambda'||_1(||\lambda||_1+||\lambda'||_1)\\ &\leq 2\epsilon C\end{aligned}$$

We're almost done. First we argue that if $D \sim \mathcal{D}^n$, then if n is sufficiently large, the squared error of any model $f(x;\lambda)$ for $\lambda \in N_\epsilon(C, |\mathcal{G}|)$ is close under both D and $\mathcal{D}$. In fact this is true for any finite set $S \subseteq B(C, |\mathcal{G}|)$ so long as n scales with the log of $|S|$:

Lemma 4.2.3 *Fix any finite subset $S \subset B(C, |\mathcal{G}|)$ and any $\delta > 0$. Let $D \sim \mathcal{D}^n$ consist of n samples (x, y). Then with probability $1-\delta$, for every $\lambda \in S$:*

$$\left|\mathop{\mathbb{E}}_{(x,y)\sim D}[(f(x;\lambda)-y)^2]-\mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}[(f(x;\lambda)-y)^2\right| \leq (C+1)^2\sqrt{\frac{\ln\left(\frac{2|S|}{\delta}\right)}{2n}}$$

Proof 30 *First observe that since $\lambda \in B(C, |\mathcal{G}|)$, we have that for all x, $-C \leq f(x;\lambda) \leq C$. Thus for all x, y:*

$$(f(x;\lambda)-y)^2 \leq (C+1)^2$$

We can therefore apply Hoeffding's inequality to conclude that for any fixed $\lambda \in B(C, d)$:

$$\Pr_{D\sim\mathcal{D}^n}\left[\left|\mathop{\mathbb{E}}_{(x,y)\sim D}[(f(x;\lambda)-y)^2]-\mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}[(f(x;\lambda)-y)^2\right| \geq t\right] \leq 2\exp\left(\frac{-2nt^2}{(C+1)^4}\right)$$

The right hand side evaluates to δ if we take:

$$t = (C+1)^2\sqrt{\frac{\ln(2/\delta)}{2n}}$$

Replacing δ with $\delta/|S|$ and union bounding over all $\lambda \in S$ we have that with probability $1-\delta$, simultaneously for every $\lambda \in S$:

$$\left|\mathop{\mathbb{E}}_{(x,y)\sim D}[(f(x;\lambda)-y)^2]-\mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}[(f(x;\lambda)-y)^2\right| \leq (C+1)^2\sqrt{\frac{\ln\left(\frac{2|S|}{\delta}\right)}{2n}}$$

We can combine Lemma 4.2.3 (uniform convergence of squared error over points in a finite set) with Lemma 4.2.2 (squared error is Lipschitz) together with the existence of a finite ϵ-net for the ℓ_1 ball (Theorem 16 to obtain a similar claim for all of the (continuously many) vectors $\lambda \in B(C, |\mathcal{G}|)$:

Theorem 17 *Fix any $C, \delta, \epsilon > 0$. Let $D \sim \mathcal{D}^n$ consist of n samples (x, y). Then with probability $1 - \delta$, for every $\lambda \in B(C, |\mathcal{G}|)$:*

$$\left| \mathop{\mathbb{E}}_{(x,y)\sim D}[(f(x;\lambda) - y)^2] - \mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}[(f(x;\lambda) - y)^2 \right| \leq$$

$$(C+1)^2 \sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}| \ln\left(1 + \frac{2C}{\epsilon}\right)}{2n}} + 4\epsilon C$$

In particular, choosing $\epsilon = \frac{C}{\sqrt{n}}$ gives:

$$\left| \mathop{\mathbb{E}}_{(x,y)\sim D}[(f(x;\lambda) - y)^2] - \mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}[(f(x;\lambda) - y)^2 \right| \leq 2(C+1)^2 \sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}| \ln\left(1 + 2\sqrt{n}\right)}{2n}}$$

Proof 31 *Let $N_\epsilon = N_\epsilon(C, |\mathcal{G}|)$ be an ℓ_1 ϵ-net for $B(C, |\mathcal{G}|)$ of size $|N_\epsilon| \leq \left(1 + \frac{2C}{\epsilon}\right)^{|\mathcal{G}|}$ — which we know exists from Theorem 16. Applying Lemma 4.2.3 with $S = N_\epsilon$, we get that with probability $1 - \delta$, for every $\lambda \in N_\epsilon$:*

$$\left| \mathop{\mathbb{E}}_{(x,y)\sim D}[(f(x;\lambda) - y)^2] - \mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}[(f(x;\lambda) - y)^2 \right| \leq (C+1)^2 \sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}| \ln\left(1 + \frac{2C}{\epsilon}\right)}{2n}}$$

Now fix any $\lambda \in B(C, |\mathcal{G}|)$ and let $\lambda' = \arg\min_{\hat{\lambda} \in N_\epsilon} ||\hat{\lambda} - \lambda||_1$. We know from the ϵ-net property that $||\lambda - \lambda'||_1 \leq \epsilon$. Applying Lemma 4.2.2 twice we can conclude:

$$\begin{aligned}
& \left| \mathop{\mathbb{E}}_{(x,y)\sim D}[(f(x;\lambda) - y)^2] - \mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}[(f(x;\lambda) - y)^2 \right| \\
\leq & \left| \mathop{\mathbb{E}}_{(x,y)\sim D}[(f(x;\lambda') - y)^2] - \mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}[(f(x;\lambda') - y)^2 \right| + 4\epsilon C \\
\leq & (C+1)^2 \sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}| \ln\left(1 + \frac{2C}{\epsilon}\right)}{2n}} + 4\epsilon C
\end{aligned}$$

We're now ready to prove our generalization bound for Algorithm 12.

Theorem 18 *Fix any $\delta > 0$ and model $f : X \to [0, 1]$. Let $\mathcal{G}$ be any collection of groups. Let $D \sim \mathcal{D}^n$ consist of n samples (x, y). Then with probability $1 - \delta$, the model $\hat{f}(x, \lambda^*)$ output by Simple-Group-Conditional-Regularized$(f, \mathcal{G}, D, \eta)$) (Algorithm 12) satisfies α-approximate group conditional mean consistency on $\mathcal{D}$ whenever $\min_{g \in \mathcal{G}} \mu(g) \geq \alpha$ for:*

$$\alpha \leq \eta + 4\left(\frac{1}{\eta} + 1\right)^2 \sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}| \ln\left(1 + 2\sqrt{n}\right)}{2n}}$$

Choosing η to minimize this expression gives:

$$\alpha \leq O\left(\left(\frac{\ln\left(\frac{1}{\delta}\right) + |\mathcal{G}| \ln(n)}{n}\right)^{1/6}\right)$$

Proof 32 *Let*

$$\hat{\lambda} = \arg\min_{\lambda} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\left(\hat{f}(x;\lambda) - y\right)^2\right] + \eta||\lambda||_1$$

i.e. the true minimizer of regularized objective function over $\mathcal{D}$. We know from Lemma 4.2.1 that $\lambda^*, \hat{\lambda} \in B(1/\eta, |\mathcal{G}|)$. Hence from Theorem 17 and the fact that λ^* minimizes the objective function on D, we have that with probability $1-\delta$:

$$\begin{aligned}
& \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x;\lambda^*) - y)^2] + \eta||\lambda^*||_1 \\
\leq & \mathop{\mathbb{E}}_{(x,y)\sim D}[(f(x;\lambda^*) - y)^2] + \eta||\lambda^*||_1 + 2\left(\frac{1}{\eta}+1\right)^2\sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}|\ln\left(1+2\sqrt{n}\right)}{2n}} \\
\leq & \mathop{\mathbb{E}}_{(x,y)\sim D}[(f(x;\hat{\lambda}) - y)^2] + \eta||\hat{\lambda}||_1 + 2\left(\frac{1}{\eta}+1\right)^2\sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}|\ln\left(1+2\sqrt{n}\right)}{2n}} \\
\leq & \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x;\hat{\lambda}) - y)^2] + \eta||\hat{\lambda}||_1 + 4\left(\frac{1}{\eta}+1\right)^2\sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}|\ln\left(1+2\sqrt{n}\right)}{2n}}
\end{aligned}$$

Let α be the minimum value such that $f(x;\lambda^*)$ satisfies α-approximate group conditional mean consistency on $\mathcal{D}$. In other words, there exists a group g such that:

$$\mu(g)\cdot\left(\mathop{\mathbb{E}}_{\mathcal{D}}[f(x;\lambda^*) - y | g(x) = 1]\right)^2 = \alpha$$

Let $\Delta = \mathbb{E}_{\mathcal{D}}[f(x;\lambda^*) - y|g(x)=1]$, and let $h(x, f(x;\lambda^*), g, \Delta) = f(x,\lambda')$ be the result of applying a patch operation, where $\lambda'_{g'} = \lambda^*_{g'}$ for all $g' \neq g$ and $\lambda'_g = \lambda^*_g + \Delta$. By Lemma 4.1.1, we have that $B(f(x,\lambda^*),\mathcal{D}) - B(f(x,\lambda'),\mathcal{D}) > \alpha$. This will contradict the optimality of $\hat{\lambda}$ above if we have that:

$$\alpha > \eta\left(||\lambda'||_1 - ||\lambda^*||_1\right) + 4\left(\frac{1}{\eta}+1\right)^2\sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}|\ln\left(1+2\sqrt{n}\right)}{2n}}$$

To avoid the contradiction we must have that:

$$\begin{aligned}
\alpha & \leq \eta\left(||\lambda'||_1 - ||\lambda^*||_1\right) + 4\left(\frac{1}{\eta}+1\right)^2\sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}|\ln\left(1+2\sqrt{n}\right)}{2n}} \\
& \leq \eta|\Delta| + 4\left(\frac{1}{\eta}+1\right)^2\sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}|\ln\left(1+2\sqrt{n}\right)}{2n}} \\
& \leq \eta\sqrt{\frac{\alpha}{\mu(g)}} + 4\left(\frac{1}{\eta}+1\right)^2\sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}|\ln\left(1+2\sqrt{n}\right)}{2n}} \\
& \leq \eta + 4\left(\frac{1}{\eta}+1\right)^2\sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}|\ln\left(1+2\sqrt{n}\right)}{2n}}
\end{aligned}$$

where the second to last inequality follows from the fact that $|\Delta| = \sqrt{\frac{\alpha}{\mu(g)}}$ *and the last inequality follows from the assumption that* $\mu(g) \geq \alpha$.

We can carry out a similar analysis of a regularized variant of our algorithm for group conditional quantile consistency:

Algorithm 13 Simple-Quantile-Group-Conditional-Regularized($f, \mathcal{G}, q, \eta$)

Let λ^* be a solution to the optimization problem:

$$\text{Minimize}_\lambda \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[L_q \left(\hat{f}(x;\lambda), y \right) \right] + \eta ||\lambda||_1$$

Such that:

$$\hat{f}(x;\lambda) \equiv f(x) + \sum_{g \in \mathcal{G}} \lambda_g \cdot g(x)$$

Output $\hat{f}(x;\lambda^*)$

The basic strategy is the same, and so we highlight only the differences. Since Pinball loss is also bounded within $[0,1]$ when $f(x), y \in [0,1]$ we continue to have that solutions output by Algorithm 12 are norm bounded:

Lemma 4.2.4 *Let* $f : \mathcal{X} \to [0,1]$ *be any model with range* $[0,1]$*, let* $\mathcal{G}$ *be any set of groups, let* $\mathcal{D}$ *be any distribution over labelled example, and let* $\eta > 0$*. Then Simple-Quantile-Group-Conditional-Regularized*$(f, \mathcal{G}, \mathcal{D}, \eta)$ *(Algorithm 13) outputs a model* $\hat{f}(x, \lambda^*)$ *with:*

$$||\lambda^*||_1 \leq \frac{1}{\eta}$$

We get an even better Lipschitz bound on the loss function:

Lemma 4.2.5 *Let* $\lambda, \lambda' \in B(C, |\mathcal{G}|)$ *be such that* $||\lambda - \lambda'||_1 \leq \epsilon$*. Then for all* x, y*, and for all* $q \in [0,1]$ *we have that:*

$$|L_q(f(x;\lambda), y) - L_q(f(x;\lambda'), y)| \leq \epsilon$$

Similarly, since $|L_q(f(x;\lambda), y)| \leq C + 1$ (rather than $(C+1)^2$) for $\lambda \in B(C, |\mathcal{G}|)$, we get a uniform convergence bound that is improved over our version for squared loss by a factor of $(C+1)$:

Lemma 4.2.6 *Fix any* $q \in [0,1]$*, any finite subset* $S \subset B(C, |\mathcal{G}|)$ *and any* $\delta > 0$*. Let* $D \sim \mathcal{D}^n$ *consist of* n *samples* (x, y)*. Then with probability* $1 - \delta$*, for every* $\lambda \in S$*:*

$$\left| \mathop{\mathbb{E}}_{(x,y)\sim D} [L_q(f(x;\lambda), y)] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [L_q(f(x;\lambda), y)] \right| \leq (C+1) \sqrt{\frac{\ln\left(\frac{2|S|}{\delta}\right)}{2n}}$$

Combining these two improved lemmas gives a correspondingly improved uniform convergence theorem over all of $B(C, |\mathcal{G}|)$:

Theorem 19 *Fix any $q \in [0,1]$ and $C, \delta, \epsilon > 0$. Let $D \sim \mathcal{D}^n$ consist of n samples (x, y). Then with probability $1 - \delta$, for every $\lambda \in B(C, |\mathcal{G}|)$:*

$$\left| \mathop{\mathbb{E}}_{(x,y)\sim D}[L_q(f(x;\lambda), y)] - \mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}[L_q(f(x;\lambda), y)] \right| \leq$$

$$(C+1)\sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}| \ln\left(1 + \frac{2C}{\epsilon}\right)}{2n}} + 2\epsilon$$

In particular, choosing $\epsilon = \frac{C}{\sqrt{n}}$ gives:

$$\left| \mathop{\mathbb{E}}_{(x,y)\sim D}[L_q(f(x;\lambda), y)] - \mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}[L_q(f(x;\lambda), y)] \right| \leq 2(C+1)\sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}| \ln\left(1 + 2\sqrt{n}\right)}{2n}}$$

We can now obtain our generalization theorem for quantiles — but we'll need one more assumption. Recall that we have already been assuming that our label distributions have CDFs that are ρ-Lipschitz, which means that they have CDFs F such that $F(\tau) - F(\tau') \leq \rho(\tau - \tau')$. To prove our next generalization theorem, we'll also have to assume that the label distributions are not too flat — that is, that they are σ-anti-Lipschitz:

Definition 22 *A CDF F is σ anti-Lipschitz if for all $\tau \geq \tau'$, we have that: $F(\tau) - F(\tau') \geq \sigma(\tau - \tau')$. We say that a distribution $\mathcal{D} \in \Delta\mathcal{X} \times \mathcal{Y}$ is σ-anti-Lipschitz if all of its conditional label distributions $\mathcal{D}_{\mathcal{Y}}(x)$ have σ-anti-Lipschitz CDFs.*

Theorem 20 *Fix any $\delta > 0$ and model $f : X \to [0,1]$. Let $\mathcal{G}$ be any collection of groups. Let $D \sim \mathcal{D}^n$ consist of n samples (x, y) from a distribution $\mathcal{D}$ that is ρ-Lipschitz and σ-anti-Lipschitz. Then with probability $1 - \delta$, the model $\hat{f}(x, \lambda^*)$ output by Simple-Quantile-Group-Conditional-Regularized($f, \mathcal{G}, D, \eta$)) (Algorithm 13) satisfies α-approximate group conditional quantile consistency on $\mathcal{D}$ whenever $\min_{g\in\mathcal{G}} \mu(g) \geq \alpha$ for:*

$$\alpha \leq \frac{2\eta\rho}{\sigma} + 8\rho\left(\frac{1}{\eta} + 1\right)\sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}| \ln\left(1 + 2\sqrt{n}\right)}{2n}}$$

Choosing η to minimize this expression gives:

$$\alpha \leq O\left(\frac{\rho}{\sqrt{\sigma}} \cdot \left(\frac{\ln\left(\frac{1}{\delta}\right) + |\mathcal{G}| \ln(n)}{n}\right)^{1/4}\right)$$

Proof 33 *Let*

$$\hat{\lambda} = \arg\min_{\lambda} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[L_q(\hat{f}(x;\lambda), y) \right] + \eta ||\lambda||_1$$

i.e. the true minimizer of regularized objective function over $\mathcal{D}$. We know from Lemma 4.2.4 that $\lambda^*, \hat{\lambda} \in B(1/\eta, |\mathcal{G}|)$. Hence from Theorem 19 and the fact that λ^* minimizes the objective function on D, we have that with probability $1-\delta$:

$$\begin{aligned}
& \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[L_q(\hat{f}(x;\lambda^*), y)] + \eta||\lambda^*||_1 \\
\leq & \mathop{\mathbb{E}}_{(x,y)\sim D}[L_q(\hat{f}(x;\lambda^*), y)] + \eta||\lambda^*||_1 + 2\left(\frac{1}{\eta}+1\right)\sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}|\ln\left(1+2\sqrt{n}\right)}{2n}} \\
\leq & \mathop{\mathbb{E}}_{(x,y)\sim D}[L_q(\hat{f}(x;\hat{\lambda}), y)] + \eta||\hat{\lambda}||_1 + 2\left(\frac{1}{\eta}+1\right)\sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}|\ln\left(1+2\sqrt{n}\right)}{2n}} \\
\leq & \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[L_q(\hat{f}(x;\hat{\lambda}), y)]] + \eta||\hat{\lambda}||_1 + 4\left(\frac{1}{\eta}+1\right)\sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}|\ln\left(1+2\sqrt{n}\right)}{2n}}
\end{aligned}$$

Let α be the minimum value such that $f(x;\lambda^*)$ satisfies α-approximate group conditional quantile consistency on $\mathcal{D}$. In other words, there exists a group g such that:

$$\mu(g) \cdot \left(\Pr_{\mathcal{D}}[y \leq f(x;\lambda^*) - q | g(x) = 1]\right)^2 = \alpha$$

Let Δ be such that $\Pr_{\mathcal{D}}[y \leq f(x;\lambda^*) + \Delta | g(x) = 1] = q$, and let $h(x, f(x;\lambda^*), g, \Delta) = f(x, \lambda')$ be the result of applying a patch operation, where $\lambda'_{g'} = \lambda^*_{g'}$ for all $g' \neq g$ and $\lambda'_g = \lambda^*_g + \Delta$. We have that $PB_q(f(x,\lambda^*), \mathcal{D}) - PB_q(f(x,\lambda'), \mathcal{D}) > \frac{\alpha}{2\rho}$. This will contradict the optimality of $\hat{\lambda}$ above if we have that:

$$\frac{\alpha}{2\rho} > \eta\left(||\lambda'||_1 - ||\lambda^*||_1\right) + 4\left(\frac{1}{\eta}+1\right)\sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}|\ln\left(1+2\sqrt{n}\right)}{2n}}$$

To avoid the contradiction we must have that:

$$\begin{aligned}
\frac{\alpha}{2\rho} & \leq \eta\left(||\lambda'||_1 - ||\lambda^*||_1\right) + 4\left(\frac{1}{\eta}+1\right)\sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}|\ln\left(1+2\sqrt{n}\right)}{2n}} \\
& \leq \eta|\Delta| + 4\left(\frac{1}{\eta}+1\right)\sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}|\ln\left(1+2\sqrt{n}\right)}{2n}} \\
& \leq \frac{\eta}{\sigma}\sqrt{\frac{\alpha}{\mu(g)}} + 4\left(\frac{1}{\eta}+1\right)\sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}|\ln\left(1+2\sqrt{n}\right)}{2n}} \\
& \leq \frac{\eta}{\sigma} + 4\left(\frac{1}{\eta}+1\right)\sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}|\ln\left(1+2\sqrt{n}\right)}{2n}}
\end{aligned}$$

where the second to last inequality follows from the fact that $|\Delta| \leq \frac{1}{\sigma}\sqrt{\frac{\alpha}{\mu(g)}}$ *by the anti-Lipschitzness property, and the last inequality follows from the assumption that* $\mu(g) \geq \alpha$.

Solving we get:

$$\alpha \leq \frac{2\eta\rho}{\sigma} + 8\rho\left(\frac{1}{\eta}+1\right)\sqrt{\frac{\ln\left(\frac{2}{\delta}\right) + |\mathcal{G}|\ln\left(1+2\sqrt{n}\right)}{2n}}$$

4.3 Multicalibration: Group Conditional Calibration

We can go further and simultaneously ask for group conditional mean consistency and calibration. Combined, these two constraints are called multicalibration:

Definition 23 *Fix any model* $f : \mathcal{X} \to [0,1]$ *and group* $g : \mathcal{X} \to \{0,1\}$. *The average squared calibration error of* f *on* g *is:*

$$K_2(f,g,\mathcal{D}) = \sum_{v \in R(f)} \Pr_{(x,y)\sim\mathcal{D}}[f(x) = v | g(x) = 1]\left(v - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y|f(x) = v, g(x) = 1]\right)^2$$

We say that a model f *is* α*-approximately multicalibrated with respect to a collection of groups* $\mathcal{G}$ *and a distribution* $\mathcal{D}$ *if for every group* $g \in \mathcal{G}$*:*

$$K_2(f,g,\mathcal{D}) \leq \frac{\alpha}{\mu(g)}.$$

When $\mathcal{D}$ *is clear from context we just write* $K_2(f,g)$.

Just as in our previous cases, we will proceed by starting with an initial model which we will patch:

Definition 24 (Group Value Patch) *Given a model* $f : \mathcal{X} \to [0,1]$, *a group* $g : \mathcal{X} \to \{0,1\}$ *and a pair of values* $v, v' \in [0,1]$, *we say that the* group value patch *applied to* f *with pair* (v, v') *and group* g *is the function:*

$$h(x, f; v \to v', g) = \begin{cases} v' & f(x) = v \text{ and } g(x) = 1 \\ f(x) & \text{otherwise} \end{cases}$$

Algorithm 14 Multicalibrate($f, \alpha, \mathcal{G}, \mathcal{D}$) (First Attempt)

Let $f_0 = f$ and $t = 0$.
while f_t is not α-approximately multicalibrated with respect to $\mathcal{G}$: **do**
Let:

$$(v_t, g_t) \in \arg\max_{(v,g) \in R(f_t) \times \mathcal{G}} \Pr_{(x,y)\sim\mathcal{D}}[f_t(x) = v, g(x) = 1] \left(v - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y | f_t(x) = v, g(x) = 1])\right)^2$$

$$v'_t = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y | f_t(x) = v_t, g_t(x) = 1]$$

Let $f_{t+1} = h(x; f_t, v_t \to v'_t, g_t)$ and $t = t + 1$.
Output f_t.

Definition 25 *Fix a model f_t and a group $g : \mathcal{X} \to 0, 1$. For a value $v \in R(f_t)$, we write:*

$$\mu_t(v, g, \mathcal{D}) = \Pr_{(x,y)\sim\mathcal{D}}[f_t(x) = v, g(x) = 1]$$

When $\mathcal{D}$ is clear from context we just write $\mu_t(v, g)$.

Lemma 4.3.1 *Fix any intermediate round t of Algorithm 14 (Multicalibrate). We have that:*

$$B(f_t) - B(f_{t+1}) = \mu_t(v_t, g_t) \cdot (v_t - v'_t)^2$$

Proof 34 *Since by construction $f_{t+1}(x) = f_t(x)$ for every x such that either $g(x) = 0$ or $f_t(x) \neq v_t$, we have that:*

$$\begin{aligned} B(f_t) - B(f_{t+1}) &= \mu_t(v_t, g_t) \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_t(x) - y)^2 - (f_{t+1}(x) - y)^2 | g_t(x) = 1, f_t(x) = v_t] \\ &= \mu_t(v_t, g_t) \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(v_t - y)^2 - (v'_t - y)^2 | g_t(x) = 1, f_t(x) = v_t] \\ &= \mu_t(v_t, g_t) \cdot (v_t - v'_t)^2 \end{aligned}$$

Where the final equality follows from Lemma 3.1.2 and the fact that by definition $v'_t = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y | f_t(x) = v_t, g_t(x) = 1]$.

So far everything is mirroring our past derivations — but there is an important difference here that will complicate things (just a little!) in comparison to our analysis of Algorithm 3 (our calibration algorithm — without groups). The issue is that the updates for Multicalibrate can *increase* the cardinality of the range of our model: i.e. it might be that $|R(f_{t+1})| = |R(f_t)| + 1$. This is problematic for us since our updates change the function at the granularity of a group intersected with an element of $R(f_t)$ — and so as $R(f_t)$ grows, the rate of progress that we make slows down. We will still eventually get to multicalibration (since $\sum_{t=0}^{\infty} 1/(m + t)$ is a divergent series for any m), but we might need a lot of updates.

The fix is to realize that we don't need arbitrary precision to achieve α-approximate multicalibration — we only need some finite precision that depends on α. If we restrict our updates to an appropriately discretized set of m finite values, then the range of f_{t+1} can never grow above m and our problem is solved. This will also be useful to us when it comes time to argue that solving the empirical multi-calibration problem on a modestly sized dataset suffices to solve it out of sample, on the distribution from which the data was drawn.

Definition 26 *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$(v;m) = \arg\min_{v' \in [1/m]} |v - v'|$ denote the closest grid point to v in $[1/m]$. For a model $f : \mathcal{X} \to [0,1]$, let Round$(f;m)$ denote the function $f'(x) =$ Round$(f(x);m)$ that simply rounds the output of f to the nearest grid point of $[1/m]$.

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

Algorithm 15 Multicalibrate$(f, \alpha, \mathcal{G}, \mathcal{D})$

Let $m = \frac{1}{\alpha}$.
Let $f_0 = \text{Round}(f;m)$ and $t = 0$.
while f_t is not α-approximately multicalibrated with respect to $\mathcal{G}$: **do**
Let:

$$(v_t, g_t) \in \underset{(v,g) \in R(f_t) \times \mathcal{G}}{\arg\max} \Pr_{(x,y)\sim\mathcal{D}}[f_t(x) = v, g(x) = 1] \left(v - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y | f_t(x) = v, g(x) = 1])\right)^2$$

$$\tilde{v}_t = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y | f_t(x) = v_t, g_t(x) = 1] \text{ and } v'_t = \text{Round}(\tilde{v}_t; m)$$

Let $f_{t+1} = h(x; f_t, v_t \to v'_t, g_t)$ and $t = t + 1$.
Output f_t.

Lets start by proving an approximate variant of Lemma 4.3.1

Lemma 4.3.2 *Fix any intermediate round t of Algorithm 15 (Multicalibrate). We have that:*

$$B(f_t) - B(f_{t+1}) \geq \mu_t(v_t, g_t) \cdot \left((v_t - \tilde{v}_t)^2 - \frac{1}{4m^2}\right)$$

Proof 35 *Let $\tilde{f}_{t+1} = h(x; f_t, v_t \to \tilde{v}_t, g_t)$ be the hypothetical update that would have resulted had we not rounded $\tilde{v}_t$ in step t of the algorithm. This is the*

update that would have resulted from a step of Algorithm 14, and so we can apply Lemma 4.3.1 to conclude that:

$$B(f_t) - B(\tilde{f}_{t+1}) \geq \mu_t(v_t, g_t) \cdot (v_t - \tilde{v}_t)^2$$

We also have that:

$$\begin{aligned} B(f_t) - B(f_{t+1}) &= (B(f_t) - B(\tilde{f}_{t+1})) - (B(f_{t+1}) - B(\tilde{f}_{t+1})) \\ &= \mu_t(v_t, g_t) \cdot (v_t - \tilde{v}_t)^2 - (B(f_{t+1}) - B(\tilde{f}_{t+1})) \end{aligned}$$

And so it remains to upper bound $(B(f_{t+1}) - B(\tilde{f}_{t+1}))$. Let $\Delta = \tilde{v}_t - v'_t$ and note that since $v'_t = Round(\tilde{v}_t; m)$ we have that $|\Delta| \leq \frac{1}{2m}$. We can calculate:

$$\begin{aligned} B(f_{t+1}) - B(\tilde{f}_{t+1}) &= \mu_t(v_t, g_t) \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[(v'_t - y)^2 - (\tilde{v}'_t - y)^2 | g_t(x) = 1, f_t(x) = v_t\right] \\ &= \mu_t(v_t, g_t)\Delta^2 \\ &\leq \frac{\mu_t(v_t, g_t)}{4m^2} \end{aligned}$$

Here the 2nd equality follows from the fact that $\tilde{v}_t = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y | f_t(x) = v_t, g_t(x) = 1]$. Combining with the above we have:

$$\begin{aligned} B(f_t) - B(f_{t+1}) &= \mu_t(v_t, g_t) \cdot (v_t - \tilde{v}_t)^2 - (B(f_{t+1}) - B(\tilde{f}_{t+1})) \\ &\geq \mu_t(v_t, g_t) \cdot (v_t - \tilde{v}_t)^2 - \frac{\mu_t(v_t, g_t)}{4m^2} \\ &= \mu_t(v_t, g_t) \cdot \left((v_t - \tilde{v}_t)^2 - \frac{1}{4m^2}\right) \end{aligned}$$

Theorem 21 *Given any model f, any collection of groups $\mathcal{G}$, and any $1 \geq \alpha > 0$, Algorithm 15 (Multicalibrate) halts after $T < \frac{4}{\alpha^2}$ many rounds and outputs a model f_T that satisfies α-approximate multicalibration. Moreover if the algorithm runs for T rounds then $B(f_T) < B(f_0) - T\frac{\alpha^2}{4}$.*

Proof 36 *Consider any intermediate round t of the algorithm. If the algorithm has not halted, it must be because f_t is not α-approximately multicalibrated, and so we know that there exists a group $g \in \mathcal{G}$ such that $K_2(f_t, g) > \frac{\alpha}{\mu(g)}$. In other words:*

$$\sum_{v \in R(f_t)} \mu_t(g, v)\left(v - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y | f_t(x) = v, g(x) = 1]\right)^2 \geq \alpha$$

We know by construction that $|R(f_t)| \leq m+1$ and so by averaging there must exist some $v \in R(f_t)$ such that:

$$\mu_t(g, v)\left(v - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y | f_t(x) = v, g(x) = 1]\right)^2 \geq \frac{\alpha}{m+1}$$

In particular, since in the algorithm (g_t, v_t) are chosen in the algorithm to jointly maximize the left hand side of this quantity, we know that this inequality holds for the pair (g_t, v_t):

$$\mu_t(v_t, g_t) \cdot (v_t - \tilde{v}_t)^2 \geq \frac{\alpha}{m+1}$$

By Lemma 4.1.1 we have that:

$$\begin{aligned} B(f_t) - B(f_{t+1}) &\geq \mu_t(v_t, g_t) \cdot \left((v_t - \tilde{v}_t)^2 - \frac{1}{4m^2} \right) \\ &\geq \mu_t(v_t, g_t) \cdot (v_t - \tilde{v}_t)^2 - \frac{1}{4m^2} \\ &\geq \frac{\alpha}{m+1} - \frac{1}{4m^2} \\ &= \frac{\alpha^2}{\alpha+1} - \frac{\alpha^2}{4} \\ &\geq \frac{\alpha^2}{2} - \frac{\alpha^2}{4} \\ &= \frac{\alpha^2}{4} \end{aligned}$$

Iterating we therefore have that $B(f_T) \leq B(f_0) - T\frac{\alpha^2}{4}$ and since $B(f_T)$ and $B(f_0)$ are bounded in $[0,1]$, we must have that: $T \leq \frac{4}{\alpha^2}$.

Remark 4.3.1 *Theorem 21 bounds $B(f_t) - B(f_0)$. But recall that f_0 results from rounding the outputs of f to the nearest multiple of $1/m$, which might increase f's squared error by as much as $\frac{1}{m} + \frac{1}{4m^2} = \alpha + \frac{\alpha^2}{4}$ if f was very poorly calibrated at the outset. Taking this into account we can also conclude that:*

$$B(f_T) < B(f) - T\frac{\alpha^2}{4} + \alpha + \frac{\alpha^2}{4}.$$

4.4 Quantile Multicalibration

Similarly, if we let Pinball loss play the role of the Brier score in our analysis, we can derive algorithms for quantile multicalibration:

Definition 27 *Fix any model $f : \mathcal{X} \to [0,1]$, target quantile q, and group $g : \mathcal{X} \to \{0,1\}$. The average squared quantile calibration error of f on g is:*

$$Q_2(f,g) = \sum_{v \in R(f)} \Pr_{(x,y)\sim\mathcal{D}}[f(x) = v | g(x) = 1] \left(q - \Pr_{(x,y)\sim\mathcal{D}}[y \leq v | f(x) = v, g(x) = 1] \right)^2$$

We say that a model f is α-approximately quantile multicalibrated with respect to a collection of groups $\mathcal{G}$ and q if for every group $g \in \mathcal{G}$:

$$Q_2(f, g) \leq \frac{\alpha}{\mu(g)}.$$

We will use the same kind of group value patches that we used for mean multicalibration, as well as the same rounding procedure. We get the following algorithm:

Algorithm 16 QuantileMulticalibrate$(f, \alpha, q, \mathcal{G}, \rho)$

Let $m = \frac{\rho^2}{2\alpha}$.
Let $f_0 = \text{Round}(f; m)$ and $t = 0$.
while f_t is not α-approximately quantile multicalibrated with respect to $\mathcal{G}$ and q: **do**
Let:

$$(v_t, g_t) \in \underset{(v,g) \in R(f_t) \times \mathcal{G}}{\arg\max} \Pr_{(x,y)\sim\mathcal{D}}[f_t(x) = v, g(x) = 1] \left(q - \Pr_{(x,y)\sim\mathcal{D}}[y \leq v | f_t(x) = v, g(x) = 1])\right)^2$$

$$\tilde{v}_t = \arg\min_v \left| \Pr_{(x,y)\sim\mathcal{D}}[y \leq v | f_t(x) = v_t, g_t(x) = 1] - q \right| \text{ and } v'_t = \text{Round}(\tilde{v}_t; m)$$

Let $f_{t+1} = h(x; f_t, v_t \to v'_t, g_t)$ and $t = t + 1$.
Output f_t.

Lemma 4.4.1 *Fix any intermediate round t of Algorithm 16 (QuantileMulticalibrate) run with parameters α, q, and ρ. If $\mathcal{D}$ is ρ-Lipschitz, then We have that:*

$$PB_q(f_t) - PB_q(f_{t+1}) \geq \frac{\alpha^2}{2\rho^3}$$

Proof 37 *Since the algorithm has not halted at around t, it must be that f_t is not α-approximately quantile multicalibrated, and hence we know that:*

$$\Pr_{(x,y)\sim\mathcal{D}}[f_t(x) = v_t, g_t(x) = 1] \left(q - \Pr_{(x,y)\sim\mathcal{D}}[y \leq v_t | f_t(x) = v_t, g_t(x) = 1])\right)^2 \geq \frac{\alpha}{m}$$

Let $\tilde{f}_{t+1} = h(x; f_t, v_t \to \tilde{v}_t, g_t)$ be the hypothetical update that would have resulted had we not rounded $\tilde{v}_t$ in step t of the algorithm. Since $\Pr_{(x,y)\sim\mathcal{D}}[y \leq v | f_t(x) = v_t, g_t(x) = 1] = q$ we can apply Lemma 2.2.2 to the distribution

$\mathcal{D}|(f_t(x) = v_t, g_t(x) = 1)$ *(which must also be ρ-Lipschitz) to conclude that:*

$$\begin{aligned}
& PB_q(f_t) - PB_q(\tilde{f}_{t+1}) \\
= \; & \Pr[g_t(x) = 1, f_t(x) = v_t] \cdot \left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[L_q(f_t(x), y) - L_q(\tilde{f}_{t+1}(x), y) | g_t(x) = 1, f_t(x) = v_t \right] \right) \\
\geq \; & \mu(g_t, v_t) \cdot \frac{\alpha}{2\rho m \mu(g_t.v_t)} \\
= \; & \frac{\alpha}{2m\rho}
\end{aligned}$$

We also have that:

$$\begin{aligned}
PB_q(f_t) - PB_q(f_{t+1}) &= (PB_q(f_t) - PB_q(\tilde{f}_{t+1})) - (PB_q(f_{t+1}) - PB_q(\tilde{f}_{t+1})) \\
&\geq \frac{\alpha}{2m\rho} - (PB_q(f_{t+1}) - PB_q(\tilde{f}_{t+1}))
\end{aligned}$$

And so it remains to upper bound $(PB_q(f_{t+1}) - PB_q(\tilde{f}_{t+1}))$. Let $\Delta = \tilde{v}_t - v'_t$ and note that since $v'_t = Round(\tilde{v}_t; m)$ we have that $|\Delta| \leq \frac{1}{2m}$. From Lemma 2.2.2 we have that:

$$\begin{aligned}
PB_q(f_{t+1}) - PB_q(\tilde{f}_{t+1}) &= \mu(g_t, v_t) \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [L_q(v'_t, y) - L_q(\tilde{v}_t, y) | g_t(x) = 1, f_t(x) = v_t] \\
&\leq |Q(v'_t) - Q(\tilde{v}_t)| \cdot |\Delta| - \frac{(Q(v'_t) - Q(\tilde{v}_t))^2}{2\rho} \\
&\leq |Q(v'_t) - Q(\tilde{v}_t)| \cdot \frac{1}{2m} - \frac{(Q(v'_t) - Q(\tilde{v}_t))^2}{2\rho} \\
&\leq \frac{\rho}{4m^2} - \frac{\rho}{8m^2} \\
&= \frac{\rho}{8m^2}
\end{aligned}$$

where the first inequality follows from Lemma 2.2.2, the second follows from the fact that $\Delta \leq 1/2m$, and the third follows from the fact that by ρ-Lipschitzness, we must have that $|Q(v'_t) - Q(\tilde{v}_t)| \leq \frac{\rho}{2m}$.

Putting it all together we get that:

$$PB_q(f_t) - PB_q(f_{t+1}) \geq \frac{\alpha}{2m\rho} - \frac{\rho}{8m^2} = \frac{\alpha^2}{2\rho^3}$$

Here we use the fact that $m = \frac{\rho^2}{2\alpha}$.

With this progress lemma, we can state the final guarantee for Algorithm 16.

Theorem 22 *Fix any model $f : \mathcal{X} \to [0, 1]$, $\alpha > 0$, $q \in [0, 1]$, $\mathcal{G}$, and ρ. If the distribution $\mathcal{D}$ is ρ-Lipschitz, then Algorithm 16 (QuantileMulticalibrate)*

runs for T rounds and outputs a model f_T that is α-approximately quantile multicalibrated with respect to $\mathcal{G}$ and q. Moreover:

$$T \leq \frac{2\rho^3}{\alpha^2}$$

and $PB_q(f_T) \leq PB_q(f_0) - T\frac{\alpha^2}{2\rho^3}$.

Proof 38 *Lemma 4.4.1 tells us that at any intermediate round $t < T$ of the algorithm, we have that:*

$$PB_q(f_t) - PB_q(f_{t+1}) \geq \frac{\alpha^2}{2\rho^3}$$

Applying this repeatedly we have that:

$$PB_q(f_T) \leq PB_q(f_0) - T\frac{\alpha^2}{2\rho^3}$$

For labels in $[0,1]$, we have that $PB_q(f_0) \leq 1$ and $PB_q(f_T) \geq 0$. Hence we must have that $T \leq \frac{2\rho^3}{\alpha^2}$.

4.5 Out of Sample Generalization

Thus far we have presented our algorithms for multicalibration as if they have direct access to the distribution $\mathcal{D}$. In practice, they will not: We will run our algorithms on a finite sample of n points $D \in \mathcal{Z}^n$ to obtain multicalibration on the empirical distribution on the sample — but we will want our multicalibration guarantees to carry over to some other distribution. In this section, we will show that if the n points in D were sampled i.i.d. from any distribution $\mathcal{D}$, then so long as n is sufficiently large, the guarantees of multicalibration will indeed carry over to $\mathcal{D}$.

4.5.1 Mean Multicalibration

Imagine that we have a distribution $\mathcal{D} \in \Delta\mathcal{Z}$ and that we have sampled n points i.i.d. from $\mathcal{D}$ to form a dataset D: $D \sim \mathcal{D}^n$. We

Our generalization bounds follow a simple formula: We first argue that for any particular function f_t, if it is multicalibrated on D it is very likely multicalibrated on $\mathcal{D}$ as well. We then argue that for any fixed input model, Algorithm 15 can only output a model from a finite (and boundedly large set), and so we can union bound over all possible output models.

Theorem 23 *Fix any model $f_t : \mathcal{X} \to [0,1]$, any $v \in R(f_t)$, and any group $g \in \mathcal{G}$. Let $D \sim \mathcal{D}^n$ consist of n points drawn i.i.d. from $\mathcal{D}$. Then with probability $1-\delta$.*

$$\left|\mu_t(g,v,\mathcal{D})\left(v - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y|f_t(x)=v, g(x)=1]\right)^2 - \mu_t(g,v,D)\left(v - \mathop{\mathbb{E}}_{(x,y)\sim D}[y|f_t(x)=v, g(x)=1]\right)^2\right|$$

$$\leq 46\sqrt{\frac{3\mu_t(g,v,\mathcal{D})\ln(8/\delta)}{n}} + \frac{135\ln(8/\delta)}{n}$$

$$\in O\left(\sqrt{\frac{\mu_t(g,v,\mathcal{D})\ln(1/\delta)}{n}} + \frac{\ln(1/\delta)}{n}\right)$$

Proof 39 *This will be a long slog. We will beat each term into submission using Chernoff bounds in sequence, and then combine the resulting bounds.*

First we argue that with high probability, $\mu_t(g,v,D)$ and $\mu_t(g,v,\mathcal{D})$ must be close.

Lemma 4.5.1 *Fix any model $f_t : \mathcal{X} \to [0,1]$, any $v \in R(f_t)$, and any group $g \in \mathcal{G}$. Let $D \sim \mathcal{D}^n$ consist of n points drawn i.i.d. from $\mathcal{D}$. Then with probability $1-\delta$:*

$$|\mu_t(g,v,D) - \mu_t(g,v,\mathcal{D})| \leq \sqrt{\frac{3\ln(2/\delta)\mu_t(g,v,\mathcal{D})}{n}}$$

Proof 40 *We can write*

$$\mu_t(g,v,D) = \frac{1}{n}\sum_{(x,y)\in D} \mathbb{1}[g(x)=1, f_t(x)=v]$$

We have both that $0 \leq \mathbb{1}[g(x)=1, f_t(x)=v] \leq 1$ and that $\mathbb{E}_{D\sim\mathcal{D}^n}[\mu_t(g,v,D)] = \mu_t(g,v,\mathcal{D})$, and so we can apply the Chernoff bound (Theorem 47) to conclude:

$$\Pr_{D\sim\mathcal{D}^n}\left[|n\mu_t(g,v,D) - n\mu_t(g,v,\mathcal{D})| \geq \eta n\mu_t(g,v,\mathcal{D})\right] \leq 2\exp\left(-\frac{\mu_t(g,v,\mathcal{D})\eta^2}{3}\right)$$

Plugging in $\eta = \sqrt{\frac{3\ln(2/\delta)}{n\mu_t(g,v,\mathcal{D})}}$ yields:

$$\Pr_{D\sim\mathcal{D}^n}\left[|n\mu_t(g,v,D) - n\mu_t(g,v,\mathcal{D})| \geq \sqrt{3\ln(2/\delta)n\mu_t(g,v,\mathcal{D})}\right] \leq \delta$$

Dividing by n yields the theorem.

We next consider the term: $\mathbb{E}_{(x,y)\sim D}[y|f_t(x)=v, g(x)=1]$.

Lemma 4.5.2 *Fix any model $f_t : \mathcal{X} \to [0,1]$, any $v \in R(f_t)$, and any group $g \in \mathcal{G}$. Let $D \sim \mathcal{D}^n$ consist of n points drawn i.i.d. from $\mathcal{D}$. Then with probability $1-\delta$, for any $v \in R(f_t)$ such that $\mu_t(g,v,\mathcal{D}) \geq \frac{12\ln(4/\delta)}{n}$:*

$$\left| \mathop{\mathbb{E}}_{(x,y)\sim D}[y|f_t(x)=v, g(x)=1] - \mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}[y|f_t(x)=v, g(x)=1] \right| \leq 5\sqrt{\frac{3\ln(4/\delta)}{n\mu_t(g,v,\mathcal{D})}}$$

Proof 41 *We have that:*

$$\mathop{\mathbb{E}}_{(x,y)\sim D}[y|f_t(x)=v, g(x)=1] = \frac{\sum_{(x,y)\in D} y \cdot \mathbb{1}[f_t(x)=v]\mathbb{1}[g(x)=1]}{n\mu_t(g,v,D)}$$

Both the numerator and the denominator are i.i.d. sums of random variables bounded in $[0,1]$. So, we can apply the Chernoff bound (Theorem 47) to conclude that with probability $1-\delta$ we have simultaneously:

$$\left| \sum_{(x,y)\in D} y \cdot \mathbb{1}[f_t(x)=v]\mathbb{1}[g(x)=1] - n \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y \cdot \mathbb{1}[f_t(x)=v]\mathbb{1}[g(x)=1]] \right| \leq$$

$$\sqrt{3\ln(4/\delta) n \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y \cdot \mathbb{1}[f_t(x)=v]\mathbb{1}[g(x)=1]]} \leq \sqrt{3\ln(4/\delta) n\mu_t(g,v,\mathcal{D})}$$

and

$$|n\mu_t(g,v,D) - n\mu_t(g,v,\mathcal{D})| \leq \sqrt{3\ln(4/\delta)n\mu_t(g,v,\mathcal{D})}$$

Therefore we have that with probability $1-\delta$:

$$\begin{aligned}
& \mathop{\mathbb{E}}_{(x,y)\sim D}[y|f_t(x)=v, g(x)=1] \\
= \;& \frac{\sum_{(x,y)\in D} y \cdot \mathbb{1}[f_t(x)=v]\mathbb{1}[g(x)=1]}{n\mu_t(g,v,D)} \\
\leq \;& \frac{n\,\mathbb{E}_{(x,y)\sim\mathcal{D}}[y \cdot \mathbb{1}[f_t(x)=v]\mathbb{1}[g(x)=1]] + \sqrt{3\ln(4/\delta)n\mu_t(g,v,\mathcal{D})}}{n\mu_t(g,v,\mathcal{D}) - \sqrt{3\ln(4/\delta)n\mu_t(g,v,\mathcal{D})}} \\
= \;& \frac{n\,\mathbb{E}_{(x,y)\sim\mathcal{D}}[y \cdot \mathbb{1}[f_t(x)=v]\mathbb{1}[g(x)=1]] + \sqrt{3\ln(4/\delta)n\mu_t(g,v,\mathcal{D})}}{n\mu_t(g,v,\mathcal{D})\left(1 - \sqrt{\frac{3\ln(4/\delta)}{n\mu_t(g,v,\mathcal{D})}}\right)} \\
= \;& \left(\frac{1}{\left(1 - \sqrt{\frac{3\ln(4/\delta)}{n\mu_t(g,v,\mathcal{D})}}\right)}\right) \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y|f_t(x)=v, g(x)=1] + \frac{\sqrt{3\ln(4/\delta)}}{\sqrt{n\mu_t(g,v,\mathcal{D})}\left(1 - \sqrt{\frac{3\ln(4/\delta)}{n\mu_t(g,v,\mathcal{D})}}\right)} \\
\leq \;& \left(1 + 2\sqrt{\frac{3\ln(4/\delta)}{n\mu_t(g,v,\mathcal{D})}}\right)\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y|f_t(x)=v, g(x)=1] + \frac{\sqrt{3\ln(4/\delta)}}{\sqrt{n\mu_t(g,v,\mathcal{D})}}\right) \\
\leq \;& E_{(x,y)\sim\mathcal{D}}[y|f_t(x)=v, g(x)=1] + 3\sqrt{\frac{3\ln(4/\delta)}{n\mu_t(g,v,\mathcal{D})}} + 2\frac{3\ln(4/\delta)}{n\mu_t(g,v,\mathcal{D})} \\
\leq \;& E_{(x,y)\sim\mathcal{D}}[y|f_t(x)=v, g(x)=1] + 5\sqrt{\frac{3\ln(4/\delta)}{n\mu_t(g,v,\mathcal{D})}}
\end{aligned}$$

Here we have applied Lemma 4.5.1 to move between $\mu_t(g, v, D)$ and $\mu_t(g, v, \mathcal{D})$, and have relied on our assumption that $\sqrt{\frac{3\ln(4/\delta)}{n\mu_t(g,v,\mathcal{D})}} \leq \frac{1}{2}$ to apply the inequality $1/(1-x) \leq 1+2x$ for $0 \leq x \leq 1/2$. The last inequality follows because by our assumption on $\mu_t(g, v, \mathcal{D})$, $\frac{3\ln(4/\delta)}{n\mu_t(g,v,\mathcal{D})} \leq 1$ and hence $\frac{3\ln(4/\delta)}{n\mu_t(g,v,\mathcal{D})} \leq \sqrt{\frac{3\ln(4/\delta)}{n\mu_t(g,v,\mathcal{D})}}$.

We can similarly derive the inequality in the reverse direction to conclude that with probability $1-\delta$:

$$\left| \mathop{\mathbb{E}}_{(x,y)\sim D}[y|f_t(x) = v, g(x) = 1] - \mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}[y|f_t(x) = v, g(x) = 1] \right| \leq 5\sqrt{\frac{3\ln(4/\delta)}{n\mu_t(g, v, \mathcal{D})}}$$

Onwards! We now propagate our error bounds outwards:

Lemma 4.5.3 *Fix any model $f_t : \mathcal{X} \to [0, 1]$, any $v \in R(f_t)$, and any group $g \in \mathcal{G}$. Let $D \sim \mathcal{D}^n$ consist of n points drawn i.i.d. from $\mathcal{D}$. Then with probability $1-\delta$, for any $v \in R(f_t)$ such that $\mu_t(g, v, \mathcal{D}) \geq \frac{12\ln(4/\delta)}{n}$:*

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

$$45\sqrt{\frac{3\ln(4/\delta)}{n\mu_t(g, v, \mathcal{D})}}$$

Proof 42 *We compute:*

$$\begin{aligned}
& \left| \left(v - \mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}[y|f_t(x) = v, g(x) = 1] \right)^2 - \left(v - \mathop{\mathbb{E}}_{(x,y)\sim D}[y|f_t(x) = v, g(x) = 1] \right)^2 \right| \\
= & \left| 2v \left(\mathop{\mathbb{E}}_{(x,y)\sim D}[y|f_t(x) = v, g(x) = 1] - \mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}[y|f_t(x) = v, g(x) = 1] \right) + \right. \\
& \left. \left(\mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}[y|f_t(x) = v, g(x) = 1]^2 - \mathop{\mathbb{E}}_{(x,y)\sim D}[y|f_t(x) = v, g(x) = 1]^2 \right) \right| \\
\leq & 2v \left| \mathop{\mathbb{E}}_{(x,y)\sim D}[y|f_t(x) = v, g(x) = 1] - \mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}[y|f_t(x) = v, g(x) = 1] \right| + \\
& \left| \left(\mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}[y|f_t(x) = v, g(x) = 1]^2 - \mathop{\mathbb{E}}_{(x,y)\sim D}[y|f_t(x) = v, g(x) = 1]^2 \right) \right| \\
\leq & 10v\sqrt{\frac{3\ln(4/\delta)}{n\mu_t(g, v, \mathcal{D})}} + \left| \left(\mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}[y|f_t(x) = v, g(x) = 1]^2 - \mathop{\mathbb{E}}_{(x,y)\sim D}[y|f_t(x) = v, g(x) = 1]^2 \right) \right| \\
\leq & 10\sqrt{\frac{3\ln(4/\delta)}{n\mu_t(g, v, \mathcal{D})}} + 10\sqrt{\frac{3\ln(4/\delta)}{n\mu_t(g, v, \mathcal{D})}} + \left(5\sqrt{\frac{3\ln(4/\delta)}{n\mu_t(g, v, \mathcal{D})}} \right)^2 \\
\leq & 45\sqrt{\frac{3\ln(4/\delta)}{n\mu_t(g, v, \mathcal{D})}}
\end{aligned}$$

Here we have applied Lemma 4.5.2 twice. The last inequality follows because by our assumption on $\mu_t(g, v, \mathcal{D})$, $\frac{3\ln(4/\delta)}{n\mu_t(g,v,\mathcal{D})} \leq 1$ *and hence* $\frac{3\ln(4/\delta)}{n\mu_t(g,v,\mathcal{D})} \leq \sqrt{\frac{3\ln(4/\delta)}{n\mu_t(g,v,\mathcal{D})}}$.

Phew. Lets finish this. Applying Lemma 4.5.1, we have that with probability $1 - \delta/2$*:*

$$\begin{aligned} & \mu_t(g, v, D)\left(v - \mathop{\mathbb{E}}_{(x,y)\sim D}[y|f_t(x) = v, g(x) = 1]\right)^2 \\ \leq & \left(\mu_t(g, v, \mathcal{D}) + \sqrt{\frac{3\ln(2/\delta)\mu_t(g, v, \mathcal{D})}{n}}\right)\left(v - \mathop{\mathbb{E}}_{(x,y)\sim D}[y|f_t(x) = v, g(x) = 1]\right)^2 \end{aligned}$$

There are two cases to consider. The first case is when $\mu_t(g, v, \mathcal{D}) < \frac{12\ln(8/\delta)}{n}$. *In this case, since* $\left(v - \mathbb{E}_{(x,y)\sim D}[y|f_t(x) = v, g(x) = 1]\right)^2 \leq 1$ *we have:*

$$\mu_t(g, v, D)\left(v - \mathop{\mathbb{E}}_{(x,y)\sim D}[y|f_t(x) = v, g(x) = 1]\right)^2 \leq \frac{12\ln(8/\delta)}{n} + \sqrt{\frac{3\ln(2/\delta)\mu_t(g, v, \mathcal{D})}{n}}$$

In the remaining case, we can apply Lemma 4.5.3 to continue and conclude that with probability $1 - \delta$*:*

$$\begin{aligned} & \mu_t(g, v, D)\left(v - \mathop{\mathbb{E}}_{(x,y)\sim D}[y|f_t(x) = v, g(x) = 1]\right)^2 \\ \leq & \left(\mu_t(g, v, \mathcal{D}) + \sqrt{\frac{3\ln(2/\delta)\mu_t(g, v, \mathcal{D})}{n}}\right)\left(\left(v - \mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}[y|f_t(x) = v, g(x) = 1]\right)^2 + 45\sqrt{\frac{3\ln(8/\delta)}{n\mu_t(g, v, \mathcal{D})}}\right) \\ \leq & \mu_t(g, v, \mathcal{D})\left(v - \mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}[y|f_t(x) = v, g(x) = 1]\right)^2 \\ & + \sqrt{\frac{3\ln(2/\delta)\mu_t(g, v, \mathcal{D})}{n}} + 45\sqrt{\frac{3\mu_t(g, v, \mathcal{D})\ln(8/\delta)}{n}} + 45\sqrt{\frac{3\ln(2/\delta)\mu_t(g, v, \mathcal{D})}{n}}\sqrt{\frac{3\ln(8/\delta)}{n\mu_t(g, v, \mathcal{D})}} \\ \leq & \mu_t(g, v, \mathcal{D})\left(v - \mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}[y|f_t(x) = v, g(x) = 1]\right)^2 + 46\sqrt{\frac{3\mu_t(g, v, \mathcal{D})\ln(8/\delta)}{n}} + \frac{135\ln(8/\delta)}{n} \end{aligned}$$

The reverse direction follows the same way:

$$\begin{aligned} \mu_t(g, v, D)\left(v - \mathop{\mathbb{E}}_{(x,y)\sim D}[y|f_t(x) = v, g(x) = 1]\right)^2 \geq & \mu_t(g, v, \mathcal{D})\left(v - \mathop{\mathbb{E}}_{(x,y)\sim \mathcal{D}}[y|f_t(x) = v, g(x) = 1]\right)^2 \\ & -46\sqrt{\frac{3\mu_t(g, v, \mathcal{D})\ln(8/\delta)}{n}} - \frac{135\ln(8/\delta)}{n} \end{aligned}$$

which finally gives us our theorem.

Recapping where we are, we have shown that for a *single* model f_t, group g, and value v, the quantities $\mu_t(g, v, D)\left(v - \mathbb{E}_{(x,y)\sim D}[y|f_t(x) = v, g(x) = 1]\right)^2$ evaluated in-sample are close to the corresponding quantities out of sample. But we need a corresponding statement for *every* group $g \in \mathcal{G}$, every $v \in [1/m]$ and every model f that might be output by Algorithm 15. Our solution to this will simply be to *count* all possible combinations of g, v, and f, but in order to do this, we need to understand how many different distinct models might be output by Algorithm 15.

Lemma 4.5.4 *Fix any model $f : \mathcal{X} \to [0, 1]$, any finite collection of groups $\mathcal{G}$, and any $\alpha > 0$. Then there is a set of models C such that for every distribution $\mathcal{D}$ (which might be the empirical distribution over an arbitrary dataset), the model f_t output by Multicalibrate$(f, \alpha, \mathcal{G}, \mathcal{D})$ is such that $f_t \in C$, and:*

$$|C| \leq \left(\frac{|\mathcal{G}|}{\alpha^2}\right)^{\frac{4}{\alpha^2}+1}$$

Proof 43 *Given a run of Multicalibrate$(f, \alpha, \mathcal{G}, \mathcal{D})$ (Algorithm 15) for T rounds, let $\pi = \{(v_t, v'_t, g_t)\}_{t=1}^{T}$ denote the record of the quantities (v_t, v'_t, g_t) selected by the algorithm at each round t. Let $\pi^{<t} = \{(v_{t'}, v'_{t'}, g_{t'})\}_{t'=1}^{t-1}$ denote the prefix of this transcript up through round $t-1$. Observe that once we fix $\pi^{<t}$ we have also fixed the model f_t that is defined at the start of round t (independently of the distribution $\mathcal{D}$). Thus to count models that might be output by Multicalibrate$(f, \alpha, \mathcal{G}, \mathcal{D})$, it suffices to count transcripts.*

We let C denote the set of all models defined by transcripts $\pi^{<T}$ for all $T \leq \frac{4}{\alpha^2}$. Since we know from Theorem 21 that Algorithm 15 halts after at most $T \leq \frac{4}{\alpha^2}$ many rounds, the models output by Algorithm 15 must be contained in C as claimed. It remains to count the set of transcripts of length $T \leq \frac{4}{\alpha^2}$. At each round t, there are $m = 1/\alpha$ possible choices for v_t, $m = 1/\alpha$ possible choices for v'_t, and $|\mathcal{G}|$ possible choices for g_t. Hence the number of transcripts of length T is $\left(\frac{|\mathcal{G}|}{\alpha^2}\right)^T$. Thus we have:

$$|C| \leq \sum_{T=0}^{\frac{4}{\alpha^2}} \left(\frac{|\mathcal{G}|}{\alpha^2}\right)^T \leq \left(\frac{|\mathcal{G}|}{\alpha^2}\right)^{\frac{4}{\alpha^2}+1}$$

Having counted the number of models that multicalibrate might output, we can apply our union bound:

Theorem 24 *Fix any model $f : \mathcal{X} \to [0, 1]$, any finite collection of groups $\mathcal{G}$, any $\alpha > 0$ and any $\delta > 0$. Let $D \sim \mathcal{D}^n$ consist of n points drawn i.i.d. from $\mathcal{D}$. Then with probability $1-\delta$, simultaneously for every model $f_t : \mathcal{X} \to [0, 1]$ that can be output by Multicalibrate$(f, \alpha, \mathcal{G}, D)$ (Algorithm 15), any group $g \in \mathcal{G}$, and any $v \in R(f_t)$:*

$$\left|\mu_t(g, v, \mathcal{D})\left(v - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y|f_t(x) = v, g(x) = 1]\right)^2 - \mu_t(g, v, D)\left(v - \mathop{\mathbb{E}}_{(x,y)\sim D}[y|f_t(x) = v, g(x) = 1]\right)^2\right|$$

$$\leq 46\sqrt{\frac{3\mu_t(g,v,\mathcal{D})\left(\frac{4}{\alpha^2}+2\right)\ln\left(\frac{8|\mathcal{G}|}{\alpha^2\delta}\right)}{n}} + \frac{135\left(\frac{4}{\alpha^2}+2\right)\ln\left(\frac{8|\mathcal{G}|}{\alpha^2\delta}\right)}{n}$$

$$\in O\left(\frac{1}{\alpha}\sqrt{\frac{\mu_t(g,v,\mathcal{D})\ln\left(\frac{|\mathcal{G}}{\alpha\delta}\right)}{n}}\right)$$

Proof 44 *From Theorem 23 we have that for any $\delta' > 0$ and any* single *triple (f_t, g, v) we have that:*

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

$$46\sqrt{\frac{3\mu_t(g,v,\mathcal{D})\ln(8/\delta)}{n}} + \frac{135\ln(8/\delta)}{n}$$

We now count the number of triples quantified over in our theorem. Lemma 4.5.4 tells us that the number of models f_t that might be output is at most $\left(\frac{|\mathcal{G}|}{\alpha^2}\right)^{\frac{4}{\alpha^2}+1}$. The number of groups $g \in \mathcal{G}$ is $|\mathcal{G}|$, and the number of values $v \in R(f_t)$ is by construction $m = \frac{1}{\alpha}$. Hence the number of triples is at most:

$$\left(\frac{|\mathcal{G}|}{\alpha^2}\right)^{\frac{4}{\alpha^2}+1} \cdot |\mathcal{G}| \cdot \frac{1}{\alpha} \leq \left(\frac{|\mathcal{G}|}{\alpha^2}\right)^{\frac{4}{\alpha^2}+2}$$

The theorem then follows from invoking Theorem 23 with $\delta' = \frac{\delta}{\left(\frac{|\mathcal{G}|}{\alpha^2}\right)^{\frac{4}{\alpha^2}+2}}$ and then summing the failure probability δ' over all enumerated triples.

We're now ready to state our final generalization theorem:

Theorem 25 *Fix any model $f : \mathcal{X} \to [0,1]$, any finite collection of groups $\mathcal{G}$, any $\alpha > 0$ and any $\delta > 0$. Let $D \sim \mathcal{D}^n$ consist of n points drawn i.i.d. from $\mathcal{D}$. Then with probability $1-\delta$, the model $f_t : \mathcal{X} \to [0,1]$ that is output by Multicalibrate$(f, \alpha, \mathcal{G}, D)$ (Algorithm 15) is α' approximately multicalibrated with respect to $\mathcal{G}$ and $\mathcal{D}$ for:*

$$\alpha' \leq \alpha + \frac{1}{\alpha}\left(\frac{135\left(\frac{4}{\alpha^2}+2\right)\ln\left(\frac{8|\mathcal{G}|}{\alpha^2\delta}\right)}{n}\right) + 46\sqrt{\frac{3\left(\frac{4}{\alpha^2}+2\right)\ln\left(\frac{8|\mathcal{G}|}{\alpha^2\delta}\right)}{\alpha n}}$$

$$\in O\left(\alpha + \frac{\ln\left(\frac{|\mathcal{G}|}{\alpha^2\delta}\right)}{\alpha^3 n} + \sqrt{\frac{\ln\left(\frac{|\mathcal{G}|}{\alpha^2\delta}\right)}{\alpha^3 n}}\right)$$

Remark 4.5.1 *Choosing α to optimize the bound from Theorem 25, we get a model f_t that is α' approximately multicalibrated with respect to $\mathcal{G}$ and $\mathcal{D}$ for:*

$$\alpha' = \tilde{O}\left(\left(\frac{\ln\left(\frac{|\mathcal{G}|}{\delta}\right)}{n}\right)^{1/5}\right)$$

Proof 45 *We need to prove that with probability $1-\delta$, for every group $g \in \mathcal{G}$: $K_2(f_t, g, \mathcal{D}) \leq \frac{\alpha'}{\mu(g,\mathcal{D})}$.*

Expanding out the definition of $K_2(f_t, g, \mathcal{D})$ this is equivalent to proving that for every $g \in \mathcal{G}$:

$$\mu(g,\mathcal{D})K_2(f_t, g, \mathcal{D}) = \sum_{v \in R(f_t)} \mu_t(g, v, \mathcal{D})\left(v - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y|f(x) = v, g(x) = 1]\right)^2 \leq \alpha'$$

From Theorem 24 we know that with probability $1-\delta$ we have that for every *$v \in R(f_t)$ and $g \in \mathcal{G}$ we have that:*

$$\left|\mu_t(g, v, \mathcal{D})\left(v - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y|f_t(x) = v, g(x) = 1]\right)^2 - \mu_t(g, v, D)\left(v - \mathop{\mathbb{E}}_{(x,y)\sim D}[y|f_t(x) = v, g(x) = 1]\right)^2\right|$$

$$\leq 46\sqrt{\frac{3\mu_t(g, v, \mathcal{D})\left(\frac{4}{\alpha^2} + 2\right)\ln\left(\frac{8|\mathcal{G}|}{\alpha^2\delta}\right)}{n} + \frac{135\left(\frac{4}{\alpha^2} + 2\right)\ln\left(\frac{8|\mathcal{G}|}{\alpha^2\delta}\right)}{n}}$$

From Theorem 21 we know that (with probability 1), $\mu(g, D)\cdot K_2(f_t, g, D) \leq \alpha$ for every $g \in \mathcal{G}$.

Combining these bounds we have:

$$
\begin{aligned}
& \mu(g,\mathcal{D})K_2(f_t,g,\mathcal{D}) \\
= & \sum_{v\in R(f_t)} \mu_t(g,v,\mathcal{D}) \left(v - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y|f(x)=v,g(x)=1]\right)^2 \\
\leq & \sum_{v\in R(f_t)} \mu_t(g,v,D) \left(v - \mathop{\mathbb{E}}_{(x,y)\sim D}[y|f(x)=v,g(x)=1]\right)^2 + \\
& \sum_{v\in R(f_t)} \left(46\sqrt{\frac{3\mu_t(g,v,\mathcal{D})\left(\frac{4}{\alpha^2}+2\right)\ln\left(\frac{8|\mathcal{G}|}{\alpha^2\delta}\right)}{n}} + \frac{135\left(\frac{4}{\alpha^2}+2\right)\ln\left(\frac{8|\mathcal{G}|}{\alpha^2\delta}\right)}{n}\right) \\
\leq & \alpha + \sum_{v\in R(f_t)} \left(46\sqrt{\frac{3\mu_t(g,v,\mathcal{D})\left(\frac{4}{\alpha^2}+2\right)\ln\left(\frac{8|\mathcal{G}|}{\alpha^2\delta}\right)}{n}} + \frac{135\left(\frac{4}{\alpha^2}+2\right)\ln\left(\frac{8|\mathcal{G}|}{\alpha^2\delta}\right)}{n}\right) \\
\leq & \alpha + \frac{1}{\alpha}\left(\frac{135\left(\frac{4}{\alpha^2}+2\right)\ln\left(\frac{8|\mathcal{G}|}{\alpha^2\delta}\right)}{n}\right) + \sum_{v\in R(f_t)} 46\sqrt{\frac{3\mu_t(g,v,\mathcal{D})\left(\frac{4}{\alpha^2}+2\right)\ln\left(\frac{8|\mathcal{G}|}{\alpha^2\delta}\right)}{n}} \\
\leq & \alpha + \frac{1}{\alpha}\left(\frac{135\left(\frac{4}{\alpha^2}+2\right)\ln\left(\frac{8|\mathcal{G}|}{\alpha^2\delta}\right)}{n}\right) + 46\sqrt{\frac{3\mu_t(g,\mathcal{D})\left(\frac{4}{\alpha^2}+2\right)\ln\left(\frac{8|\mathcal{G}|}{\alpha^2\delta}\right)}{\alpha n}} \\
\leq & \alpha + \frac{1}{\alpha}\left(\frac{135\left(\frac{4}{\alpha^2}+2\right)\ln\left(\frac{8|\mathcal{G}|}{\alpha^2\delta}\right)}{n}\right) + 46\sqrt{\frac{3\left(\frac{4}{\alpha^2}+2\right)\ln\left(\frac{8|\mathcal{G}|}{\alpha^2\delta}\right)}{\alpha n}}
\end{aligned}
$$

Where here we have used the fact that $|R(f_t)| = \frac{1}{\alpha}$, and that because $\sqrt{\cdot}$ is a concave function, the final sum is maximized when $\mu_t(g,v,\mathcal{D}) = \alpha\mu_t(g)$ for each v.

4.5.2 Quantile Multicalibration

We can prove a very similar generalization bound for quantile multicalibration. We elide the details that are for the most part similar to the case of mean multicalibration, and state the final theorem:

Theorem 26 *Fix any model $f : \mathcal{X} \to [0,1]$, any finite collection of groups $\mathcal{G}$, any $\alpha > 0$ and any $\delta > 0$. Let $D \sim \mathcal{D}^n$ consist of n points drawn i.i.d. from a ρ-Lipschitz distribution $\mathcal{D}$. Then with probability $1-\delta$, the model $f_t : \mathcal{X} \to [0,1]$ that is output by QuantileMulticalibrate$(f,\alpha,q,\mathcal{G},D)$ (Algorithm 16) is*

α' approximately quantile multicalibrated with respect to target quantile q and $\mathcal{G}$ and $\mathcal{D}$ for:

$$\alpha' = \alpha + 42\sqrt{\frac{3\rho^2\left(\ln(\frac{4\pi^2T^2}{3\delta}) + T\ln(\frac{\rho^4|\mathcal{G}|}{\alpha^2})\right)}{2\alpha n}}$$

Remark 4.5.2 *Choosing α to optimize the bound from Theorem 25, we get a model f_t that is α' approximately multicalibrated with respect to $\mathcal{G}$ and $\mathcal{D}$ for:*

$$\alpha' = \tilde{O}\left(\left(\frac{\rho^3\ln\left(\frac{\rho^4|\mathcal{G}|}{\delta}\right)}{n}\right)^{1/5}\right)$$

4.6 Sequential Prediction

In this section we give algorithms that can promise mean and quantile multicalibration in the sequential setting, against an adversary. It will be more convenient for us to bound ℓ_∞ calibration error (K_∞) rather than the ℓ_2 calibration error that we have been working with. But to avoid trivialities, we will need a slightly modified definition that allows us to drive the calibration error to 0 while keeping m fixed. (Recall an issue with K_∞ and Q_∞ as we have defined them is that it is possible to drive them to zero at a rate of $1/m$ in a trivial manner by taking m to be very large and making sure that we do not make any particular prediction with probability greater than $1/m$.)

4.6.1 A Bucketed Calibration Definition

Recall that calibration asks informally that $\mathbb{E}_{(x,y)\sim\mathcal{D}}[y|f(x) = v] \approx v$ for all v. To avoid problems with the conditioning event, we have thus far restricted our attention to models f that have bounded range $R(f) \subseteq [1/m]$. A different solution is to allow f to have arbitrary range in $[0, 1]$, but not to condition on the event that $f(x) = v$, but rather to condition on the event that $f(x) \approx v$ for an appropriate formalization of $\approx$. We will do this through bucketing:

Definition 28 *Let m be a bucket coarseness parameter. For $i \in \{1, \ldots, m-1\}$ let $B_m(i) = [\frac{i-1}{m}, \frac{i}{m})$ and let $B_m(m) = [\frac{m-1}{m}, 1]$. Collectively, $B_m(i)$ for $i \in [m]$ form a set of m "buckets" of width $1/m$ that partition the unit interval.*

We now give a variant of our ℓ_∞ calibration definition in which informally, the conditioning event $f(x) \approx v$ is taken to mean "$f(x)$ and v are in the same bucket".

Definition 29 (Bucketed Multicalibration Error in the Distributional Setting) *Fix a predictor $f : \mathcal{X} \to [0,1]$, a collection of groups $\mathcal{G}$, and a bucket coarseness parameter m. The calibration error of f on a group g with respect to bucketing coarseness m, on distribution $\mathcal{D}$ is defined to be:*

$$C_\infty(f, m, g, \mathcal{D}) =$$

$$\max_{i \in m} \Pr_{(x,y)\sim\mathcal{D}}[f(x) \in B_m(i) | g(x) = 1] \cdot \left| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f(x) - y | f(x) \in B_m(i), g(x) = 1] \right|$$

We say that f satisfies (α, m)-multicalibration with respect to $\mathcal{G}$ on $\mathcal{D}$ if for every $g \in \mathcal{G}$:

$$C_\infty(f, m, g, \mathcal{D}) \leq \frac{\alpha}{\mu(g)}$$

This is identical to our definition of K_∞ except that the condition that $f(x) = v$ has been replaced with the condition that $f(x) \in B(i)$. We can give a corresponding definition in the sequential setting:

Definition 30 (Bucketed Multicalibration Error in the Sequential Setting) *Fix a collection of groups $\mathcal{G}$, a transcript $\pi = \{(x_1, p_1, y_1), \ldots, (x_T, p_T, y_T)\}$, and a bucket coarseness parameter m. For any $i \in [m]$ and $g \in \mathcal{G}$, let $S(\pi, g, i) = \{t : g(x_t) = 1, p_t \in B_m(i)\}$ and $S(\pi, g) = \{t : g(x_t) = 1\}$. Let $n(\pi, g, i) = |S(\pi, g, i)|$ and let $n(\pi, g) = |S(\pi, g)|$.*

The calibration error of π on a group g with respect to bucketing coarseness m is defined to be:

$$C_\infty(\pi, m, g) = \max_{i \in [m]} \frac{n(\pi, g, i)}{n(\pi, g)} \cdot \left| \frac{\sum_{t \in S(\pi,g,i)} (p_t - y_t)}{n(\pi, g, i)} \right|$$

We say that π satisfies (α, m) mutlicalibration with respect to $\mathcal{G}$ if for every $g \in \mathcal{G}$:

$$C_\infty(\pi, m, g) \leq \frac{\alpha T}{n(\pi, g)}$$

Expanding out the definitions we find that equivalently, π satisfies (α, m) multicalibration error with respect to $\mathcal{G}$ if:

$$\max_{g \in \mathcal{G}, i \in [m]} \left| \sum_{t \in S(\pi,g,i)} (p_t - y_t) \right| \leq \alpha T$$

4.6.2 Achieving Bucketed Calibration

Our goal is to design a sequential prediction algorithm that guarantees (α, m) multicalibration, with α tending to 0 with T against any possible sequence of observations. To this end, we define a surrogate loss function that replaces the

max in the definition of bucketed multicalibration with a "softmax" function based on the sums of exponentials, that is analytically better behaved but that is nevertheless a good approximation to the max function.

Definition 31 (Surrogate Loss) *For a round $s \leq T$ and a transcript π, recall that $\pi^{\leq s}$ denotes the length s prefix of π. For a group $g \in \mathcal{G}$, and a bucket $i \in [m]$ let:*

$$V_s^{g,i} = \sum_{t \in S(\pi^{\leq s}, g, i)} (y_t - p_t)$$

denote the average difference between the predictions p_t and the outcomes y_t on the subsequence of $\pi^{\leq s}$ corresponding to examples from group g and predictions in bucket i.

Fixing a parameter $\eta \in [0, \frac{1}{2}]$, define a surrogate calibration loss function at round s as:

$$L_s(\pi^{\leq s}) = \sum_{\substack{g \in \mathcal{G}, \\ i \in [m]}} \left(\exp(\eta V_s^{g,i}) + \exp(-\eta V_s^{g,i})\right).$$

When the transcript $\pi^{\leq s}$ is clear from context, we will simply write L_s.

We will leave η unspecified for now, and choose it later to optimize our bounds. Recall that what we really want to do is upper bound $\max_{G \in \mathcal{G}, i \in [n]} |V_T^{G_i}|$, which corresponds to our calibration loss. Observe that this "soft-max style" function allows us to tightly upper bound our calibration loss:

Observation 4.6.1 *For any transcript π_T, and any $\eta \in [0, \frac{1}{2}]$, we have that:*

$$\max_{g \in \mathcal{G}, i \in [m]} \left|V_T^{g,i}\right| \leq \frac{1}{\eta} \ln(L_T) \leq \max_{g \in \mathcal{G}, i \in [m]} \left|V_T^{G,i}\right| + \frac{\ln(2|\mathcal{G}|m)}{\eta}.$$

Proof 46 *For the first inequality, note that:*

$$\begin{aligned}
\max_{g \in \mathcal{G}, i \in [m]} \left|\eta V_T^{g,i}\right| &= \ln\left(\exp\left(\max_{g \in \mathcal{G}, i \in [m]} \eta \left|V_T^{g,i}\right|\right)\right) \\
&= \ln\left(\max_{g \in \mathcal{G}, i \in [m]} \exp\left(\eta \left|V_T^{g,i}\right|\right)\right) \\
&\leq \ln\left(\max_{g \in \mathcal{G}, i \in [m]} \exp\left(\eta V_T^{g,i}\right) + \exp\left(-\eta V_T^{g,i}\right)\right) \\
&\leq \ln\left(\sum_{g \in \mathcal{G}, i \in [m]} \exp\left(\eta V_T^{g,i}\right) + \exp\left(-\eta V_T^{g,i}\right)\right) \\
&= \ln(L_T)
\end{aligned}$$

Dividing by η gives the inequality. In the other direction we have that:

$$\begin{aligned}\frac{1}{\eta}\ln(L_t) &= \frac{1}{\eta}\ln\left(\sum_{g\in\mathcal{G},i\in[m]}\exp\left(\eta V_T^{g,i}\right)+\exp\left(-\eta V_T^{g,m}\right)\right)\\ &\leq \frac{1}{\eta}\ln\left(2|\mathcal{G}|m\cdot\max_{g\in\mathcal{G},i\in[m]}\exp\left(\eta\left|V_T^{g,i}\right|\right)\right)\\ &= \frac{\ln(2|\mathcal{G}|m)}{\eta}+\max_{g\in\mathcal{G},i\in[m]}\left|V_T^{g,i}\right|\end{aligned}$$

So we now feel freed to study the analytically nicer surrogate loss function.

Just as in our derivation of algorithms promising (regular) calibration guarantees against an adversary, we will be interested in bounding the *increase* in our surrogate loss function from round to round.

Definition 32 *Fix any partial transcript $\pi^{\leq s+1} = \pi^{\leq s}\circ(x_{s+1},p_{s+1},y_{s+1})$. Define:*

$$\Delta_{s+1}(\pi^{\leq s+1}) \equiv \Delta_{s+1}(\pi^{\leq s},(x_{s+1},p_{s+1},y_{s+1})) = L_{s+1}(\pi^{\leq s+1}) - L_s(\pi^{\leq s})$$

Our first step is to bound $\Delta_{s+1}(\pi^{\leq s+1})$ in terms of a quantity that is *linear* in p_{s+1} and y_{s+1}.

Lemma 4.6.1 *Fix any partial transcript $\pi^{\leq s+1} = \pi^{\leq s}\circ(x_{s+1},p_{s+1},y_{s+1})$ such that $p_{s+1}\in B(i)$. Then for any $\eta\leq 1$, we have that:*

$$\Delta_{s+1}(\pi^{\leq s+1}) \leq \eta(y_{s+1}-p_{s+1})\cdot C_s^i(x_{s+1}) + 2\eta^2 L_s(\pi^{\leq s})$$

where:

$$C_s^i(x_{s+1}) = \sum_{g\in\mathcal{G}(x_{s+1})}\exp(\eta V_s^{g,i}) - \exp(-\eta V_s^{g,i})$$

is a constant depending only on $\pi^{\leq s}$ and x_{s+1}.

Proof 47 *Observe that our surrogate loss function is a sum of terms each defined by a group $g\in\mathcal{G}$ and a bucket $i\in[m]$, and that at round $s+1$, the* change *in surrogate loss can be written as a sum over only those groups in $\mathcal{G}(x_{s+1})$ over the bucket i such that $p_{s+1}\in B(i)$, since all other terms in the*

sum cancel out. Therefore we can write:

$$\begin{aligned}
& \Delta_{s+1}(\pi^{\leq s+1}) \\
= \; & L_{s+1} - L_s \\
= \; & \sum_{g \in \mathcal{G}(x_{s+1})} \Big(\exp(\eta V_{s+1}^{g,i}) - \exp(\eta V_s^{g,i})) + \exp(-\eta V_{s+1}^{g,i}) - \exp(-\eta V_s^{g,i})\Big) \\
= \; & \sum_{g \in \mathcal{G}(x_{s+1})} \exp(\eta V_s^{g,i})(\exp(\eta(y_{s+1} - p_{s+1})) - 1) + \exp(-\eta V_s^{g,i})(\exp(-\eta(y_{s+1} - p_{s+1})) - 1) \\
\leq \; & \sum_{g \in \mathcal{G}(x_{s+1})} \exp(\eta V_s^{g,i})(\eta(y_{s+1} - p_{s+1}) + 2\eta^2) + \exp(-\eta V_s^{g,i})(-\eta(y_{s+1} - p_{s+1}) + 2\eta^2) \\
= \; & \eta(y_{s+1} - p_{s+1}) \left(\sum_{g \in \mathcal{G}(x_{s+1})} \exp(\eta V_s^{g,i}) - \sum_{g \in \mathcal{G}(x_{s+1})} \exp(-\eta V_s^{g,i}) \right) + \\
& 2\eta^2 \left(\sum_{g \in \mathcal{G}(x_{s+1})} \exp(\eta V_s^{g,i}) + \exp(-\eta V_s^{g,i}) \right) \\
\leq \; & \eta(y_{s+1} - p_{s+1}) \cdot C_s^i(x_{s+1}) + 2\eta^2 L_s(\pi^{\leq s})
\end{aligned}$$

Here the first inequality follows from the fact $\eta(y_{s+1} - p_{s+1}) \leq \eta$ *and that for* $|x| \leq 1$, $\exp(x) \leq 1 + x + x^2$.

Our goal is to find a strategy for the learner's choice of p_{s+1}, as a function of both $\pi^{\leq s}$ and x^{s+1}, that will guarantee that $\mathbb{E}_{p_{s+1}}\left[\Delta_{s+1}(\pi^{\leq s}, (x_{s+1}, p_{s+1}, y_{s+1}))\right]$ is small for *every* possible realization of y_{s+1}. We measure calibration over m buckets, but we will allow our learner to play from a larger strategy space $\left[\frac{1}{rm}\right] = \{0, \frac{1}{rm}, \ldots, \frac{rm-1}{rm}, 1\}$ for some integer $r > 1$. In the end we will see that our bounds get better with larger r, but that the algorithm we design has no dependence on r at all in its running time, so we can imagine r to be an arbitrarily large number.

Lemma 4.6.2 *Fix any transcript* $\pi^{\leq s}$ *and any* x_{s+1}. *There is a distribution on predictions* $p_{s+1} \in \left[\frac{1}{rm}\right]$ *such that for every* $y_{s+1} \in [0, 1]$:

$$\mathop{\mathbb{E}}_{p_{s+1}}\left[\Delta_{s+1}(\pi^{\leq s}, (x_{s+1}, p_{s+1}, y_{s+1}))\right] \leq L_s(\pi^{\leq s}) \cdot \left(\frac{\eta}{rm} + 2\eta^2\right)$$

The distribution can be sampled from as follows:

1. *If* $C_s^i(x_{s+1}) > 0$ *for all* i *then predict* $p_{s+1} = 1$
2. *If* $C_s^i(x_{s+1}) < 0$ *for all* i *ten predict* $p_{s+1} = 0$
3. *Otherwise, find* $i^* \in [m-1]$ *such that* $C_s^{i^*}(x_{s+1})C_s^{i^*+1}(x_{s+1}) \leq 0$ *and let* $q \in [0, 1]$ *be such that*

$$q \cdot C_s^{i^*}(x_{s+1}) + (1 - q)C_s^{i^*+1}(x_{s+1}) = 0.$$

Predict $p_{s+1} = \frac{i^*}{m} - \frac{1}{rm}$ *with probability* q *and predict* $p_{s+1} = \frac{i^*}{m}$ *with probability* $(1-q)$.

Proof 48 *From Lemma 4.6.1 we have that:*

$$\Delta_{s+1}(\pi^{\leq s+1}) \leq \eta(y_{s+1} - p_{s+1}) \cdot C_s^i(x_{s+1}) + 2\eta^2 L_s(\pi^{\leq s})$$

where $p_{s+1} \in B(i)$. *So it suffices to prove that:*

$$\mathop{\mathbb{E}}_{p_{s+1}}[\eta(y_{s+1} - p_{s+1}) \cdot C_s^{B^{-1}(p_{s+1})}(x_{s+1})] \leq \frac{\eta}{rm} L_s(\pi^{\leq s})$$

where $B^{-1}(p_{s+1}) = i$ *is the bucket such that* $p_{s+1} \in B(i)$. *We do this in cases.*

Case 1: $C_s^i(x_{s+1}) > 0$ *for all* i*:*

In this case $p_{s+1} = 1$ *and we have that* $\eta(y_{s+1} - p_{s+1}) \leq 0$ *Since* $C_s^m(x_{s+1}) > 0$, *it follows that* $\eta(y_{s+1} - p_{s+1}) \cdot C_s^m(x_{s+1}) \leq 0$.

Case 2: $C_s^i(x_{s+1}) < 0$ *for all* i*:*

In this case $p_{s+1} = 0$ *and we have that* $\eta(y_{s+1} - p_{s+1}) \geq 0$ *Since* $C_s^1(x_{s+1}) < 0$, *it follows that* $\eta(y_{s+1} - p_{s+1}) \cdot C_s^1(x_{s+1}) \leq 0$.

Case 3: Everything Else

In the remaining case, observe that since the quantities $C_s^i(x_{s+1})$ *are neither all positive or all negative, there must exist some bucket* i^* *such that* $C_s^{i^*}(x_{s+1})C_s^{i^*+1}(x_{s+1}) \leq 0$. *So sampling from the specified distribution is well defined and we can compute:*

$$\begin{aligned}
& \mathop{\mathbb{E}}_{p_{s+1}}[\eta(y_{s+1} - p_{s+1}) \cdot C_s^{B^{-1}(p_{s+1})}(x_{s+1})] \\
= \;& q \cdot \left(\eta\left(y_{s+1} - \left(\frac{i^*}{m} - \frac{1}{rm}\right)\right) \cdot C_s^{i^*}(x_{s+1})\right) + (1-q) \cdot \left(\eta\left(y_{s+1} - \frac{i^*}{m}\right) \cdot C_s^{i^*+1}(x_{s+1})\right) \\
= \;& \eta\left(y_{s+1} - \frac{i^*}{m}\right) \cdot \left(qC_s^{i^*}(x_{s+1}) + (1-q)C_s^{i^*+1}(x_{s+1})\right) + \eta q \frac{1}{rm} C_s^{i^*}(x_{s+1}) \\
= \;& \eta q \frac{1}{rm} C_s^{i^*}(x_{s+1}) \\
\leq \;& \frac{\eta}{rm}\left(\sum_{g \in \mathcal{G}(x_{s+1})} \exp(\eta V_s^{g,i}) - \exp(-\eta V_s^{g,i})\right) \\
\leq \;& \frac{\eta}{rm} L_s(\pi^{\leq s})
\end{aligned}$$

We have a concrete algorithm that implements the prediction strategy laid out in Lemma 3.4.2.

Algorithm 17 Online-Multicalibrated-Predictor($\mathcal{G}, m, r, \eta$)

for $t = 1$ to T **do**
Observe x_t and compute

$$C^i_{t-1}(x_t) = \sum_{g \in \mathcal{G}(x_t)} \exp(\eta V^{g,i}_{t-1}) - \exp(-\eta V^{g,i}_{t-1})$$

for all $i \in [m]$.
if $C^m_{t-1}(x_t) > 0$ **then**
Predict $p_t = 1$.
else if $C^1_{t-1}(x_t) < 0$ **then**
Predict $p_t = 0$.
else
Select $i^* \in [m]$ such that such that $C^{i^*}_{t-1}(x_t) \cdot C^{i^*+1}_{t-1}(x_t) \leq 0$.
Compute $q \in [0, 1]$ such that:

$$q \cdot C^{i^*}_{t-1}(x_t) + (1 - q) \cdot C^{i^*+1}_{t-1}(x_t) = 0$$

Predict $p_t = \frac{i^*}{m} - \frac{1}{rm}$ with probability q and predict $p_t = \frac{i^*}{m}$ with probability $1 - q$.
Observe y_t
Let $\pi^{<t+1} = \pi^{<t} \circ (x_t, p_t, y_t)$

Lets now analyze the expected calibration loss of Algorithm 17. We start by analyzing the expected surrogate loss:

Lemma 4.6.3 *Fix any set of groups $\mathcal{G}$, $m, r \geq 0$ and $0 \leq \eta \leq 1$. Fix any adversary, which together with Online-Multicalibrated-Predictor($\mathcal{G}, m, r, \eta$) (Algorithm 17) fixes a distribution on transcripts π. We have that:*

$$\mathop{\mathbb{E}}_{\pi}[L_T(\pi)] \leq 2|G|m \cdot \exp\left(\frac{T\eta}{rm} + 2T\eta^2\right)$$

Proof 49 *Consider the final round T. From Lemma 4.6.2, we have that for all $\pi^{<T}$, x_T, y_T:*

$$\begin{aligned}
\mathop{\mathbb{E}}_{p_T}[L_T(\pi^{\leq T})] &= L_{T-1}(\pi^{\leq T-1}) + \mathop{\mathbb{E}}_{p_T}[\Delta_T(\pi^{\leq T-1}, (x_T, p_T, y_T))] \\
&\leq L_{T-1}(\pi^{\leq T-1}) + L_{T-1}(\pi^{\leq T-1})\left(\frac{\eta}{rm} + 2\eta^2\right) \\
&= L_{T-1}\left(1 + \frac{\eta}{rm} + 2\eta^2\right) \\
&\leq L_{T-1}\exp\left(\frac{\eta}{rm} + 2\eta^2\right)
\end{aligned}$$

where the last inequality follows from $1 + x \leq \exp(x)$. Now inductively taking

the expectation with respect to $p_{T-1}, p_{T-2}, \ldots, p_1$ we get that:

$$\mathop{\mathbb{E}}_{\pi}[L_T(\pi)] \leq L_0 \exp\left(\frac{\eta}{rm} + 2\eta^2\right)^T = 2|\mathcal{G}|m \cdot \exp\left(\frac{T\eta}{rm} + 2T\eta^2\right)$$

Since $L_0 = \sum_{g\in\mathcal{G}, i\in[m]}(\exp(0) + \exp(0)) = 2|\mathcal{G}|m$.

We are now ready to state the final guarantee of Algorithm 17.

Theorem 27 *Fix any set of groups $\mathcal{G}$, $m, r \geq 0$. Let $\eta = \sqrt{\frac{\log(2|\mathcal{G}|m)}{2T}} < 1$. Fix any adversary, which together with Online-Multicalibrated-Predictor$(\mathcal{G}, m, r, \eta)$ (Algorithm 17) fixes a distribution on transcripts π. We have that π satisfies (α, m)-multicalibration error with respect to $\mathcal{G}$ where:*

$$\mathop{\mathbb{E}}_{\pi}[\alpha] \leq \frac{1}{rm} + 2\sqrt{\frac{2\ln(2|\mathcal{G}|m)}{T}}$$

In particular, if we choose $r \geq \frac{\sqrt{T}}{\epsilon m\sqrt{2\log(2|\mathcal{G}|m)}}$ then we have:

$$\mathop{\mathbb{E}}_{\pi}[\alpha] \leq (2+\epsilon)\sqrt{\frac{2\ln(2|\mathcal{G}|m)}{T}}$$

Proof 50 *Recall that (α, m)-multicalibration corresponds to the requirement that $\max_{g\in\mathcal{G}, i\in[m]} |V_T^{g,i}| \leq \alpha T$. Hence we need to show that:*

$$\mathop{\mathbb{E}}_{\pi}\left[\max_{g\in\mathcal{G}, i\in[m]} |V_T^{g,i}|\right] \leq \frac{T}{rm} + 2\sqrt{2T\ln(2|\mathcal{G}|m)}$$

We can compute:

$$\begin{aligned}
\exp\left(\eta \mathop{\mathbb{E}}_{\pi}\left[\max_{g\in\mathcal{G}, i\in[m]} |V_T^{g,i}|\right]\right) &\leq \mathop{\mathbb{E}}_{\pi}\left[\exp\left(\eta \max_{g\in\mathcal{G}, i\in[m]} |V_T^{g,i}|\right)\right] \\
&= \mathop{\mathbb{E}}_{\pi}\left[\max_{g\in\mathcal{G}, i\in[m]} \exp\left(\eta|V_T^{g,i}|\right)\right] \\
&\leq \mathop{\mathbb{E}}_{\pi}\left[\max_{g\in\mathcal{G}, i\in[m]} \left(\exp\left(\eta V_T^{g,i}\right) + \exp\left(-\eta V_T^{g,i}\right)\right)\right] \\
&\leq \mathop{\mathbb{E}}_{\pi}\left[\sum_{g\in\mathcal{G}, i\in[m]} \left(\exp\left(\eta V_T^{g,i}\right) + \exp\left(-\eta V_T^{g,i}\right)\right)\right] \\
&= \mathop{\mathbb{E}}_{\pi}\left[L_T(\pi)\right] \\
&\leq 2|G|m \cdot \exp\left(\frac{T\eta}{rm} + 2T\eta^2\right)
\end{aligned}$$

where the first inequality follows from Jensen's inequality and the convexity of

$\exp(x)$*, and the last inequality follows from Lemma 4.6.3. Taking the log of both sides and dividing by* η *gives:*

$$\mathop{\mathbb{E}}_{\pi}\left[\max_{g\in\mathcal{G},i\in[m]}|V_T^{g,i}|\right] \leq \frac{\log(2|\mathcal{G}|m)}{\eta} + \frac{T}{rm} + 2T\eta$$

Plugging in our chosen value of η *completes the proof.*

Insert high probability bound and online to offline reduction

4.6.3 Obtaining Bucketed Quantile Multicalibration

We can analogously define a "bucketed" definition of quantile multicalibration:

Definition 33 (Bucketed Multicalibration Error in the Sequential Setting) *Fix a collection of groups* $\mathcal{G}$*, a transcript* $\pi = \{(x_1, p_1, y_1), \ldots, (x_T, p_T, y_T)\}$*, and a bucket coarseness parameter* m*. The quantile calibration error of* π *on a group* g *with respect to bucketing coarseness* m *and target quantile* q *is defined to be:*

$$Q_\infty(\pi, m, g) = \max_{i\in[m]} \frac{n(\pi, g, i)}{n(\pi, g)} \cdot \left|\frac{\sum_{t\in S(\pi,g,i)}(\mathbb{1}[y_t \leq p_t] - q)}{n(\pi, g, i)}\right|$$

We say that π *satisfies* (α, m) *quantile multicalibration with respect to* $\mathcal{G}$ *if for every* $g \in \mathcal{G}$*:*

$$Q_\infty(\pi, m, g) \leq \frac{\alpha T}{n(\pi, g)}$$

Expanding out the definitions we find that equivalently, π *satisfies* (α, m) *quantile multicalibration error with respect to* $\mathcal{G}$ *if:*

$$\max_{g\in\mathcal{G},i\in[m]}\left|\sum_{t\in S(\pi,g,i)}(\mathbb{1}[y_t \leq p_t] - q)\right| \leq \alpha T$$

The derivation of an algorithm for online quantile multicalibration closely mimics our derivation for mean multicalibration, so we will state lemmas without proof when the proof is exactly analogous, and focus only on differences.

Definition 34 (Quantile Surrogate Loss) *For a round* $s \leq T$ *and a transcript* π*, recall that* $\pi^{\leq s}$ *denotes the length* s *prefix of* π*. For a group* $g \in \mathcal{G}$*, and a bucket* $i \in [m]$ *redefine:*

$$V_s^{g,i} = \sum_{t\in S(\pi^{\leq s},g,i)}(\mathbb{1}[y_t \leq p_t] - q).$$

Fixing a parameter $\eta \in [0, \frac{1}{2}]$*, continue to let the surrogate calibration loss function at round* s *as:*

$$L_s(\pi^{\leq s}) = \sum_{\substack{g\in\mathcal{G},\\ i\in[m]}}\left(\exp(\eta V_s^{g,i}) + \exp(-\eta V_s^{g,i})\right).$$

When the transcript $\pi^{\leq s}$ is clear from context, we will simply write L_s.

We can prove a direct analogue of Lemma 4.6.1 for our new quantile surrogate loss function. All we used previously about the $V_s^{g,i}$ quantities was that they were sums of terms bounded between $[-1, 1]$ which remains true in our quantile reformulation.

Lemma 4.6.4 *Fix any partial transcript $\pi^{\leq s+1} = \pi^{\leq s} \circ (x_{s+1}, p_{s+1}, y_{s+1})$ such that $p_{s+1} \in B(i)$. Then for any $\eta \leq 1$, we have that:*

$$\Delta_{s+1}(\pi^{\leq s+1}) \leq \eta(\mathbb{1}[y_{s+1} \leq p_{s+1}] - q) \cdot C_s^i(x_{s+1}) + 2\eta^2 L_s(\pi^{\leq s})$$

where:

$$C_s^i(x_{s+1}) = \sum_{g \in \mathcal{G}(x_{s+1})} \exp(\eta V_s^{g,i}) - \exp(-\eta V_s^{g,i})$$

is a constant depending only on $\pi^{\leq s}$ and x_{s+1}.

We now come to the only lemma whose statement and proof change — the bound on how much our surrogate loss changes in expectation when we play according to our multicalibration strategy (which does not change). Our bound now depends on the Lipschitz parameter of the underlying distributions played by the adversary, and holds in expectation both over the randomness of our prediction p_{s+1} and over the adversary's choice of labels y_{s+1}.

Lemma 4.6.5 *Fix any transcript $\pi^{\leq s}$ and any x_{s+1}. There is a distribution on predictions $p_{s+1} \in \left[\frac{1}{rm}\right]$ such that for every ρ-Lipschitz distribution over $y_{s+1} \in [0, 1]$:*

$$\mathop{\mathbb{E}}_{p_{s+1}, y_{s+1}} \left[\Delta_{s+1}(\pi^{\leq s}, (x_{s+1}, p_{s+1}, y_{s+1}))\right] \leq L_s(\pi^{\leq s}) \cdot \left(\frac{\eta}{\rho rm} + 2\eta^2\right)$$

The distribution can be sampled from as follows:

1. *If $C_s^i(x_{s+1}) < 0$ for all i then predict $p_{s+1} = 1$*
2. *If $C_s^i(x_{s+1}) > 0$ for all i ten predict $p_{s+1} = 0$*
3. *Otherwise, find $i^* \in [m-1]$ such that $C_s^{i^*}(x_{s+1}) C_s^{i^*+1}(x_{s+1}) \leq 0$ and let $p \in [0, 1]$ be such that*

$$p \cdot C_s^{i^*}(x_{s+1}) + (1-p) C_s^{i^*+1}(x_{s+1}) = 0.$$

Predict $p_{s+1} = \frac{i^*}{m} - \frac{1}{rm}$ with probability q and predict $p_{s+1} = \frac{i^*}{m}$ with probability $(1-q)$.

Proof 51 *From Lemma 4.6.4 we have that:*

$$\Delta_{s+1}(\pi^{\leq s+1}) \leq \eta(\mathbb{1}[y_{s+1} \leq p_{s+1}] - q) \cdot C_s^i(x_{s+1}) + 2\eta^2 L_s(\pi^{\leq s})$$

where $p_{s+1} \in B(i)$. So it suffices to prove that:

$$\mathop{\mathbb{E}}_{p_{s+1}, y_{s+1}} [\eta(\mathbb{1}[y_{s+1} \leq p_{s+1}] - q) \cdot C_s^{B^{-1}(p_{s+1})}(x_{s+1})] \leq \frac{\eta}{\rho rm} L_s(\pi^{\leq s})$$

where $B^{-1}(p_{s+1}) = i$ is the bucket such that $p_{s+1} \in B(i)$. We do this in cases.

Case 1: $C_s^i(x_{s+1}) > 0$ for all i:

In this case $p_{s+1} = 0$ and we have that $\eta(\mathbb{1}[y_{s+1} \leq p_{s+1}] - q) \leq 0$ Since $C_s^1(x_{s+1}) > 0$, it follows that $\eta(\mathbb{1}[y_{s+1} \leq p_{s+1}] - q) \cdot C_s^1(x_{s+1}) \leq 0$.

Case 2: $C_s^i(x_{s+1}) < 0$ for all i:

In this case $p_{s+1} = 1$ and we have that $\eta(\mathbb{1}[y_{s+1} \leq p_{s+1}] - q) \geq 0$ Since $C_s^m(x_{s+1}) < 0$, it follows that $\eta(\mathbb{1}[y_{s+1} \leq p_{s+1}] - q) \cdot C_s^m(x_{s+1}) \leq 0$.

Case 3: Everything Else

In the remaining case, observe that since the quantities $C_s^i(x_{s+1})$ are neither all positive or all negative, there must exist some bucket i^* such that $C_s^{i^*}(x_{s+1})C_s^{i^*+1}(x_{s+1}) \leq 0$. So sampling from the specified distribution is well defined and we can compute:

$$\begin{aligned}
& \mathop{\mathbb{E}}_{p_{s+1}, y_{s+1}}[\eta(\mathbb{1}[y_{s+1} \leq p_{s+1}] - q) \cdot C_s^{B^{-1}(p_{s+1})}(x_{s+1})] \\
= \;& p \cdot \left(\eta\left(\Pr_{y_{s+1}}\left[y_{s+1} \leq \left(\frac{i^*}{m} - \frac{1}{rm}\right)\right] - q\right) \cdot C_s^{i^*}(x_{s+1})\right) + \\
& (1-p) \cdot \left(\eta\left(\Pr_{y_{s+1}}\left[y_{s+1} \leq \frac{i^*}{m}\right] - q\right) \cdot C_s^{i^*+1}(x_{s+1})\right) \\
\leq \;& p \cdot \left(\eta\left(\Pr_{y_{s+1}}\left[y_{s+1} \leq \left(\frac{i^*}{m}\right)\right] + \frac{1}{\rho rm} - q\right) \cdot C_s^{i^*}(x_{s+1})\right) + \\
& (1-p) \cdot \left(\eta\left(\Pr_{y_{s+1}}\left[y_{s+1} \leq \frac{i^*}{m}\right] - q\right) \cdot C_s^{i^*+1}(x_{s+1})\right) \\
= \;& \eta p \frac{1}{\rho rm} C_s^{i^*}(x_{s+1}) \\
\leq \;& \frac{\eta}{\rho rm} L_s(\pi^{\leq s})
\end{aligned}$$

With our new Lemma 3.4.4 in hand, the rest follows as before:

Algorithm 18 Online-Quantile-Multicalibrated-Predictor($\mathcal{G}, m, r, \eta$)

for $t = 1$ to T **do**

Observe x_t and compute

$$C^i_{t-1}(x_t) = \sum_{g \in \mathcal{G}(x_t)} \exp(\eta V^{g,i}_{t-1}) - \exp(-\eta V^{g,i}_{t-1})$$

for all $i \in [m]$, with $V^{g,i}_{t-1}$ defined as in Definition 34.

if $C^m_{t-1}(x_t) < 0$ **then**

Predict $p_t = 1$.

else if $C^1_{t-1}(x_t) > 0$ **then**

Predict $p_t = 0$.

else

Select $i^* \in [m]$ such that such that $C^{i^*}_{t-1}(x_t) \cdot C^{i^*+1}_{t-1}(x_t) \leq 0$.

Compute $p \in [0, 1]$ such that:

$$p \cdot C^{i^*}_{t-1}(x_t) + (1-p) \cdot C^{i^*+1}_{t-1}(x_t) = 0$$

Predict $p_t = \frac{i^*}{m} - \frac{1}{rm}$ with probability p and predict $p_t = \frac{i^*}{m}$ with probability $1 - p$.

Observe y_t

Let $\pi^{<t+1} = \pi^{<t} \circ (x_t, p_t, y_t)$

We get the following final theorem:

Theorem 28 *Fix any set of groups $\mathcal{G}$, $m, r \geq 0$ and $q \in [0, 1]$. Let $\eta = \sqrt{\frac{\log(2|\mathcal{G}|m)}{2T}} < 1$. Fix any adversary who is constrained to playing ρ-Lipschitz distributions, which together with Online-Quantile-Multicalibrated-Predictor($\mathcal{G}, m, r, \eta$) (Algorithm 18) fixes a distribution on transcripts π. We have that π satisfies (α, m)-quantile-multicalibration error with respect to $\mathcal{G}$ and target quantile q where:*

$$\mathop{\mathbb{E}}_{\pi}[\alpha] \leq \frac{1}{\rho r m} + 2\sqrt{\frac{2\ln(2|\mathcal{G}|m)}{T}}$$

We can similarly prove a high probability version of this theorem:

Theorem 29 *Fix any set of groups $\mathcal{G}$, $m, r \geq 0$ and $q \in [0, 1]$. Let $\eta = \sqrt{\frac{\log(2|\mathcal{G}|m)}{2T}} < 1$. Fix $\delta > 0$ Fix any adversary who is constrained to playing ρ-Lipschitz distributions, which together with Online-Quantile-Multicalibrated-Predictor($\mathcal{G}, m, r, \eta$) (Algorithm 18) fixes a distribution on transcripts π. We have that with probability $1 - \delta$ over the randomness of π, π satisfies (α, m)-*

quantile-multicalibration error with respect to $\mathcal{G}$ and target quantile q where:

$$\alpha \leq \frac{1}{\rho r m} + 4\sqrt{\frac{2\ln\left(\frac{2|\mathcal{G}|m}{\delta}\right)}{T}}$$

References and Further Reading

Multicalibration and Group Conditional Mean Consistency (under the name "multiaccuracy") were introduced in Hébert-Johnson et al. [2018] using a slightly different definition, roughly corresponding to what we refer to our K_∞ metric. Group conditional mean consistency (multiaccuracy) was further studied in Kim et al. [2019]. Several different multicalibration algorithms have been given in the literature, including one based on analyzing the Lagrangian of a linear program Jung et al. [2021], and one based on constructing a branching program via "split" and "merge" operations Gopalan et al. [2022b], which controls an ℓ_1 variant of multicalibration closely related to our K_1 metric. The algorithm and analysis we give here (which controls multicalibration in the K_2 metric) is based on a variant of the original algorithm given by Hébert-Johnson et al. [2018] together with a rounding operation. The algorithm we give for quantile multicalibration and group conditional quantile consistency is from Jung et al. [2022]. Algorithm 10 — the algorithm for obtaining group conditional mean consistency with a one-shot minimization of squared error is due to Parikshit Gopalan. A different analysis than we give here is given in Gopalan et al. [2022a]. Its quantile analogue (Algorithm 11) was given in Jung et al. [2022]. The generalization bounds we give are new as far as we know. The online algorithm for obtaining multicalibration against an adversary is from Gupta et al. [2022]. The full proof of the generalization theorem giving out of sample guarantees for our batch quantile multicalibration algorithm can be found in Jung et al. [2022]. The online algorithm for obtaining quantile multicalibration against an adversary is an adaptation of Bastani et al. [2022] to the ℓ_∞ setting (Bastani et al. [2022] actually give a bound on an ℓ_2 variant of quantile multicalibration, which is stronger than ℓ_∞ multicalibration, but with a polynomial dependence on $|\mathcal{G}|$. Obtaining a bound on ℓ_2 mean or quantile multicalibration with a logarithmic dependence on $|\mathcal{G}|$ remains open as of this writing.)

5

Beyond Means, Quantiles, and Calibration

CONTENTS

In this chapter we will study how much "beyond" multicalibration we can go in various ways. We'll start simple: generalizing what a "group" is. Then we'll ask what other distributional properties we can multicalibrate with respect to (beyond means and quantiles). Finally we'll ask what there is beyond calibration.

5.1 Beyond Means and Quantiles

"nobreak

5.2 Beyond Calibration

6

Multicalibration for Real Valued Functions: When Does Multicalibration Imply Accuracy?

CONTENTS

In this section we think about how we can view multicalibration as a *boosting algorithm* for regression — i.e. a way to take a regression algorithm that has the capacity to predict in *slightly* non-trivial ways (better than a constant function), and produce an ensemble of regression functions that can predict optimally. Along the way, we will generalize multicalibration over a set of *groups* $\mathcal{G}$ represented by indicator functions $g : \mathcal{X} \to \{0, 1\}$ to multicalibration over a collection of arbitrary real valued functions $\mathcal{H}$, where each $h \in \mathcal{H}$ is a function $h : \mathcal{X} \to \mathbb{R}$. This generalization will also be useful for us for a number of other applications of multicalibration.

6.1 Beyond Groups

Our study of multicalibration thus far has been predicated on *groups* — i.e. subsets of the feature domain $\mathcal{X}$. We have represented groups by their indicator functions g, such that $g(x) = 1$ if x is a member of the group, and $g(x) = 0$ otherwise. Recall that what we mean by (perfect) multicalibration of a function $f : \mathcal{X} \to \mathbb{R}$ on a collection of groups $\mathcal{G}$ is that for every $g \in \mathcal{G}$ and $v \in R(f)$:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y - f(x))|f(x) = v, g(x) = 1] = 0$$

Since $g(x)$ is binary, we can equivalently re-write this multicalibration condition as the requirement that for every $g \in \mathcal{G}$ and $v \in R(f)$:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[g(x)(y - f(x))|f(x) = v] = 0$$

But although this is an equivalent condition to ask for when g is binary (i.e. a group indicator function), it now makes sense to ask for this condition even if g is an arbitrary real valued function $g : \mathcal{X} \to \mathbb{R}$. We will use this as our more general definition of multicalibration with respect to an arbitrary class of real valued functions. We will have to define approximate versions of this condition, and we will again use an ℓ_2-error variant:

Definition 35 (Multicalibration With Respect to Real Valued Functions) *Fix a distribution $\mathcal{D} \in \Delta\mathcal{Z}$ and a model $f : \mathcal{X} \to [0, 1]$. Let $\mathcal{H}$ be an arbitrary collection of real valued functions $h : \mathcal{X} \to \mathbb{R}$. We say that f is α-approximately multicalibrated with respect to $\mathcal{D}$ and $\mathcal{H}$ if for every $h \in \mathcal{H}$:*

$$K_2(f, h, \mathcal{D}) = \sum_{v\in R(f)} \Pr_{(x,y)\sim\mathcal{D}}[f(x) = v] \left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)(y - v)|f(x) = v] \right)^2 \leq \alpha$$

There is a close connection between a failure of a model f to be multicalibrated with respect to a class of functions $\mathcal{H}$ and the ability to decrease squared error of f using a simple update using a function $h \in \mathcal{H}$. We summarize the connection with two lemmas showing the connection in each direction.

First, suppose $\mathcal{H}$ contains a model h that has lower squared error than the best *constant* prediction on one of the level-sets of a calibrated model f. Then f is not multi-calibrated with respect to $\mathcal{H}$. Note that if f is calibrated, it is making the best constant prediction on each of its level sets, so the condition that h makes predictions with lower squared error than the best constant predictor on a level-set of f is the same that it makes predictions with better squared error than f on one of its level-sets.

Lemma 6.1.1 *Fix a calibrated model $f : \mathcal{X} \to \mathbb{R}$. Suppose for some $v \in R(f)$ there is an $h \in \mathcal{H}$ such that:*

$$\mathbb{E}[(f(x) - y)^2 - (h(x) - y)^2|f(x) = v] \geq \alpha$$

Then it must be that:

$$\mathbb{E}[h(x)(y - v)|f(x) = v] \geq \frac{\alpha}{2}$$

Proof 52 *We calculate:*

$$
\begin{aligned}
& \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)(y-v)|f(x)=v] \\
= & \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)y|f(x)=v] - v \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)|f(x)=v] \\
= & \frac{1}{2}\left(2 \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)y|f(x)=v] - 2v \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)|f(x)=v]\right) \\
\geq & \frac{1}{2}\left(2 \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)y|f(x)=v] - 2v \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)|f(x)=v] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h(x)-v)^2|f(x)=v]\right) \\
= & \frac{1}{2}\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[2h(x)y - h(x)^2 - v^2|f(x)=v]\right) \\
= & \frac{1}{2}\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[2h(x)y - h(x)^2 - 2vy + v^2|f(x)=v]\right) \\
= & \frac{1}{2}\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(v-y)^2 - (h(x)-y)^2|f(x)=v]\right) \\
\geq & \frac{\alpha}{2}
\end{aligned}
$$

where the 3rd to last line follows from adding and subtracting v^2 and the fact that because f is calibrated, $v\,\mathbb{E}[y|f(x)=v] = v^2$.

In the reverse direction, we show that if a model f fails to be multicalibrated with respect to a class of functions $\mathcal{H}$, then it is possible to perform a simple update on one of the level-sets of f using a function $h \in \mathcal{H}$ that witnesses a failure of multicalibration that decreases squared error on that level-set

Lemma 6.1.2 *Fix a model $f : \mathcal{X} \to \mathbb{R}$. Suppose for some $v \in R(f)$ there is an $h \in \mathcal{H}$ such that:*

$$\mathbb{E}[h(x)(y-v)|f(x)=v] \geq \alpha$$

Let $h' = v + \eta h(x)$ for $\eta = \frac{\alpha}{\mathbb{E}[h(x)^2|f(x)=v]}$. Then:

$$\mathbb{E}[(f(x)-y)^2 - (h'(x)-y)^2|f(x)=v] \geq \frac{\alpha^2}{\mathbb{E}[h(x)^2|f(x)=v]}$$

Proof 53 *We calculate:*

$$
\begin{aligned}
& \mathbb{E}[(f(x)-y)^2 - (h'(x)-y)^2|f(x)=v] \\
= & \mathbb{E}[(v-y)^2 - (v+\eta h(x)-y)^2|f(x)=v] \\
= & \mathbb{E}[v^2 - 2vy + y^2 - (v+\eta h(x))^2 + 2y(v+\eta h(x)) - y^2|f(x)=v] \\
= & \mathbb{E}[2y\eta h(x) - 2v\eta h(x) - \eta^2 h(x)^2|f(x)=v] \\
= & \mathbb{E}[2\eta h(x)(y-v) - \eta^2 h(x)^2|f(x)=v] \\
\geq & 2\eta\alpha - \eta^2\,\mathbb{E}[h(x)^2|f(x)=v] \\
= & \frac{\alpha^2}{\mathbb{E}[h(x)^2|f(x)=v]}
\end{aligned}
$$

Where the last line follows from the definition of η.

Note that there is an asymmetry in Lemma 6.1.1 and Lemma 6.1.2. Lemma 6.1.1 implies that if h has improved squared error compared to f on one of its level-sets, then h itself fails the multi-calibration condition on this levelset. On the other hand, Lemma 6.1.2 says that if h fails the multicalibration condition on some levelset v of f, then there is a function $h' = v + \eta h(x)$ that improves on the squared error of f on level-set v. We can remove this assymetry by assuming that $\mathcal{H}$ is *closed under affine transformations*

Definition 36 *A class of functions $\mathcal{H}$ is* closed under affine transformations *if for every $a, b \in \mathbb{R}$, if $h(x) \in \mathcal{H}$ then:*

$$h'(x) \equiv ah(x) + b \in \mathcal{H}$$

Most natural classes of regression functions are closed under affine transformations: linear functions, polynomials of any fixed degree d, regression trees, etc.

For classes of functions $\mathcal{H}$ that are closed under affine transformation, the relationship becomes symmetric:

Lemma 6.1.3 *Suppose $\mathcal{H}$ is closed under affine transformation. Fix a model $f : \mathcal{X} \to \mathbb{R}$ and a levelset $v \in R(f)$ of f. Then:*

1. *If f is calibrated and there exists an $h \in \mathcal{H}$ such that*

$$\mathbb{E}[(f(x) - y)^2 - (h(x) - y)^2 | f(x) = v] \geq \alpha$$

then there exists an $h' \in \mathcal{H}$ such that:

$$\mathbb{E}[h'(x)(y - v) | f(x) = v] \geq \frac{\alpha}{2}$$

2. *If there exists an $h \in \mathcal{H}$ such that:*

$$\mathbb{E}[h(x)(y - v) | f(x) = v] \geq \alpha$$

then there exists an $h' \in \mathcal{H}$ such that:

$$\mathbb{E}[(f(x) - y)^2 - (h'(x) - y)^2 | f(x) = v] \geq \frac{\alpha^2}{\mathbb{E}[h(x)^2 | f(x) = v]}$$

Proof 54 *The first part follows from Lemma 6.1.1 using $h' = h$. The second part follows from Lemma 6.1.2 using $h' = v + \eta h(x)$, where $h' \in \mathcal{H}$ by the assumption that $\mathcal{G}$ is closed under affine transformations.*

The equivalence between a failure of multicalibration with respect to $\mathcal{H}$ and the ability for some $h \in \mathcal{H}$ to improve on the squared error of f on one of f's level sets is useful for several reasons. First, it means that we can reduce the

problem of finding a model f that is multicalibrated over $\mathcal{H}$ to the standard regression problem of finding models $h \in \mathcal{H}$ that minimize squared error over subsets of the distribution $\mathcal{D}$, which is a well studied problem for which we have very good algorithms for many classes $\mathcal{H}$. The second is that it will allow us to give a simple, intuitive characterization of what properties $\mathcal{H}$ must have relative to a data distribution $\mathcal{D}$ such that multicalibration with respect to $\mathcal{H}$ implies Bayes optimal prediction with respect to $\mathcal{H}$. Importantly, since we only have to solve regression problems on subsets of the distribution $\mathcal{D}$ — for which there is a fixed Bayes optimal predictor — this will make it easy for us to enunciate conditions under which multicalibration implies accuracy; this would be more difficult if we needed to solve regression problems on distributions with different conditional label dsitributions.

6.2 Algorithmically Reducing Multicalibration to Regression

In this section we give an algorithm for computing predictors that are multicalibrated with respect to a real-valued class of functions $\mathcal{H}$. We will be interested in infinite classes $\mathcal{H}$, so we will need to think about what kind of access we have to this class of functions. What we will assume is that we have access to an algorithm $A_{\mathcal{H}}$ that given access to a distribution $\mathcal{D}$ solves the squared error regression problem on $\mathcal{D}$ over $\mathcal{H}$.

Definition 37 *$A_{\mathcal{H}}$ is a squared error regression oracle for a class of real valued functions $\mathcal{H}$ if for every $\mathcal{D} \in \Delta\mathcal{Z}$, $A_{\mathcal{H}}(\mathcal{D})$ outputs a function $h \in \mathcal{H}$ such that:*

$$h \in \arg\min_{h' \in \mathcal{H}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h'(x) - y)^2]$$

A squared error regression oracle $A_{\mathcal{H}}$ is a very natural object: for example, if $\mathcal{H}$ is the class of linear functions, then $A_{\mathcal{H}}$ is simply solves a linear regression problem (which has a solution in closed form). Polynomial squared error regression problems can also be solved in closed form. Even for model classes (e.g. regression trees and neural networks) such that the corresponding squared error regression problem is not convex, we have very good heuristics for solving the problem. So assuming that we have a squared error regression oracle for a class $\mathcal{H}$ is a very reasonable assumption. We now ask: if we have such an oracle, can we leverage it to learn a multi-calibrated predictor over $\mathcal{H}$?

Algorithm 19 RegressionMulticalibrate($f, \alpha, A_{\mathcal{H}}, \mathcal{D}, B$)

Let $m = \frac{2B}{\alpha}$.
Let $f_0 = \text{Round}(f; m)$, $\text{err}_0 = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f_0(x) - y)^2]$, $\text{err}_{-1} = \infty$ and $t = 0$.
while $(\text{err}_{t-1} - \text{err}_t) \geq \frac{\alpha}{2B}$ **do**
  for each $v \in [1/m]$ **do**
    Let $\mathcal{D}_v^{t+1} = \mathcal{D}|f_t(x) = v$.
    Let $h_v^{t+1} = A_{\mathcal{H}}(\mathcal{D}_v^{t+1})$.
  Let:

$$\tilde{f}_{t+1}(x) = \sum_{v\in[1/m]} \mathbb{1}[f_t(x) = v] \cdot h_v^{t+1}(x) \quad f_{t+1} = \text{Round}(f_{t+1}, m)$$

  Let $\text{err}_{t+1} = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f_{t+1}(x) - y)^2]$ and $t = t + 1$.
Output f_{t-1}.

Just as in our algorithm for multicalibration over groups $\mathcal{G}$ (Algorithm 15), Algorithm 19 rounds its output to the discrete range $[1/m] = \{0, \frac{1}{m}, \ldots, \frac{m-1}{m}, 1\}$. We recall that $\text{Round}(h, m)$ outputs the function:

$$\tilde{h}(x) = \min_{v\in[1/m]} |h(x) - v|$$

— i.e. the function that outputs the closest grid-point in $[1/m]$ to the function value $h_t(x)$.

Theorem 30 *Fix any distribution $\mathcal{D} \in \Delta\mathcal{Z}$, any model $f : \mathcal{X} \to [0, 1]$, any $\alpha < 1$, any class of real valued functions $\mathcal{H}$ that is closed under affine transformations, and a squared error regression oracle $A_{\mathcal{H}}$ for $\mathcal{H}$. For any bound $B > 0$ let:*

$$\mathcal{H}_B = \{h \in \mathcal{H} : h(x)^2 \leq B\}$$

be the set of functions in h with squared magnitude bounded by B. Then RegressionMulticalibrate($f, \alpha, A_{\mathcal{H}}, \mathcal{D}, B$) (Algorithm 19) halts after at most $T \leq \frac{2B}{\alpha}$ many iterations and outputs a model f_{T-1} such that f_{T-1} is α-approximately multicalibrated with respect to $\mathcal{D}$ and $\mathcal{H}_B$.

Remark 6.2.1 *Note the form of this theorem — we do not promise multicalibration at approximation parameter α for all of $\mathcal{H}$, but only for $\mathcal{H}_B$ — i.e. those functions in $\mathcal{H}$ satisfying a bound on their squared value. This is necessary, since $\mathcal{H}$ is closed under affine transformations. To see this, note that if $\mathbb{E}[h(x)(y - v)] \geq \alpha$, then it must be that $\mathbb{E}[c \cdot h(x)(y - v)] \geq c \cdot \alpha$. Since $h'(x) = ch(x)$ is also in $\mathcal{H}$ by assumption, approximate multicalibration bounds must always also be paired with a bound on the norm of the functions for which we promise those bounds.*

Remark 6.2.2 *The algorithm runs for at most $\frac{2B}{\alpha}$ iterations, and at each iteration needs to make $m + 1 = \frac{2B}{\alpha} + 1$ many calls to the squared error*

regression oracle $A_{\mathcal{H}}$. Thus to obtain α-approximate multi-calibration with respect to $\mathcal{H}_B$, it suffices to make $\frac{4B^2}{\alpha^2} + \frac{2B}{\alpha}$ many oracle calls to a regression oracle for $\mathcal{H}$.

Proof 55 *Since f_0 takes values in $[0,1]$ and $y \in [0,1]$, we have $err_0 \leq 1$, and by definition $err_T \geq 0$ for all T. By construction, if the algorithm has not halted at round t it must be that $err_t \leq err_{t-1} - \frac{\alpha}{2B}$, and so the algorithm must halt after at most $T \leq \frac{2B}{\alpha}$ many iterations to avoid a contradiction.*

It remains to show that when the algorithm halts at round T, the model f_{T-1} that it outputs is α-approximately multi-calibrated with respect to $\mathcal{D}$ and $\mathcal{H}_B$. We will show that if this is not the case, then $err_{T-1} - err_T > \frac{\alpha}{2B}$, which will be a contradiction to the halting criterion of the algorithm.

Suppose that f_{T-1} is not α-approximately multicalibrated with respect to $\mathcal{D}$ and $\mathcal{H}_B$. This means there must be some $h \in \mathcal{H}_B$ such that:

$$\sum_{v \in [1/m]} \Pr_{(x,y)\sim\mathcal{D}}[f_{T-1}(x) = v] \left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)(y-v)|f_{T-1}(x) = v] \right)^2 > \alpha$$

For each $v \in [1/m]$ define

$$\alpha_v = \Pr_{(x,y)\sim\mathcal{D}}[f_{T-1}(x) = v] \left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)(y-v)|f_{T-1}(x) = v] \right)^2$$

So we have $\sum_{v\in[1/m]} \alpha_v > \alpha$.

Applying the 2nd part of Lemma 6.1.3 we learn that for each v, there must be some $h_v \in \mathcal{H}$ such that:

$$\begin{aligned} \mathbb{E}[(f_{T-1}(x)-y)^2 - (h_v(x)-y)^2|f_{T-1}(x)=v] &> \frac{1}{\mathbb{E}[h(x)^2|f_{T-1}(x)=v]} \cdot \frac{\alpha_v}{\Pr_{(x,y)\sim\mathcal{D}}[f_{T-1}(x)=v]} \\ &\geq \frac{1}{B} \frac{\alpha_v}{\Pr_{(x,y)\sim\mathcal{D}}[f_{T-1}(x)=v]} \end{aligned}$$

where the last inequality follows from the fact that $h \in \mathcal{H}_B$ Now we can compute:

$$\begin{aligned} & \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x)-y)^2 - (\tilde{f}_T(x)-y)^2] \\ = & \sum_{v\in[1/m]} \Pr_{(x,y)\sim\mathcal{D}}[f_{T-1}(x)=v] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x)-y)^2 - (\tilde{f}_T(x)-y)^2|f_{T-1}(x)=v] \\ = & \sum_{v\in[1/m]} \Pr_{(x,y)\sim\mathcal{D}}[f_{T-1}(x)=v] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x)-y)^2 - (h_v^T(x)-y)^2|f_{T-1}(x)=v] \\ \geq & \sum_{v\in[1/m]} \Pr_{(x,y)\sim\mathcal{D}}[f_{T-1}(x)=v] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x)-y)^2 - (h_v(x)-y)^2|f_{T-1}(x)=v] \\ \geq & \sum_{v\in[1/m]} \frac{\alpha_v}{B} \\ > & \frac{\alpha}{B} \end{aligned}$$

Here the third line follows from the definition of $\tilde{f}_T$ and the fourth line follows from the fact $h_v \in \mathcal{H}$ and that h_v^T minimizes squared error on $\mathcal{D}_v^T$ amongst all $h \in \mathcal{H}$.

Finally we calculate:

$$\begin{aligned}
& err_{T-1} - err_T \\
= & \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x)-y)^2 - (f_T(x)-y)^2] \\
= & \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x)-y)^2 - (\tilde{f}_T(x)-y)^2] + \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(\tilde{f}_T(x)-y)^2 - (f_T(x)-y)^2] \\
> & \frac{\alpha}{B} + \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(\tilde{f}_T(x)-y)^2 - (f_T(x)-y)^2] \\
> & \frac{\alpha}{B} - \frac{1}{m} \\
\geq & \frac{\alpha}{2B}
\end{aligned}$$

where the last equality follows from the fact that $m \geq \frac{2B}{\alpha}$.

The 2nd inequality follows from the fact that for every pair (x, y):

$$(\tilde{f}_T(x)-y)^2 - (f_T(x)-y)^2 \geq -\frac{1}{m}$$

To see this we consider two cases. Since $y \in [0,1]$, if $\tilde{f}_T(x) > 1$ or $\tilde{f}_T(x) < 0$ then the Round operation decreases squared error and we have $(\tilde{f}_T(x)-y)^2 - (f_T(x)-y)^2 \geq 0$. In the remaining case we have $f_T(x) \in [0,1]$ and $\Delta = \tilde{f}_T(x) - f_T(x)$ is such that $|\Delta| \leq \frac{1}{2m}$. In this case we can compute:

$$\begin{aligned}
(\tilde{f}_T(x)-y)^2 - (f_T(x)-y)^2 &= (f_T(x)+\Delta-y)^2 - (f_T(x)-y)^2 \\
&= 2\Delta(f(x)-y) + \Delta^2 \\
&\geq -2|\Delta| + \Delta^2 \\
&\geq -\frac{1}{m}
\end{aligned}$$

6.3 Weak Learning, Multicalibration, and Boosting

We now turn from multicalibration to "Boosting". Our analysis of multicalibration algorithms has used squared error as a potential function — so we know that post-processing a model to make it multicalibrated does not harm accuracy (as measured by squared error). But when must multicalibration improve accuracy meaningfully? Can we find conditions on the class $\mathcal{H}$ with respect to which we are multicalibrated such that multicalibration must imply *Bayes optimality*? That is what we'll do now.

Definition 38 *Fix a distribution $\mathcal{D} \in \Delta\mathcal{Z}$ and a class of functions $\mathcal{H}$. Let $f^*(x) = \mathbb{E}_{y\sim\mathcal{D}(x)}[y]$ denote the true conditional label expectation conditional on x. We say that $\mathcal{H}$ satisfies the* weak learner *condition relative to $\mathcal{D}$ if for every $S \subset \mathcal{X}$ with $\Pr_{x\sim\mathcal{D}_\mathcal{X}}[x \in S] > 0$, if:*

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x) - y)^2 | x \in S] < \min_{c\in\mathbb{R}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(c - y)^2 | x \in S]$$

then there exists an $h \in \mathcal{H}$ such that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h(x) - y)^2 | x \in S] < \min_{c\in\mathbb{R}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(c - y)^2 | x \in S]$$

First lets pause to interpret this condition and explain why it is "weak". It is helpful to recall that $f^*(x)$ is the Bayes optimal predictor for squared error — it minimizes squared error over $\mathcal{D}$ over the set of all possible functions (we proved this in Lemma 3.1.2.) The weak learning condition requires that for every restriction of $\mathcal{D}$ to some subset $S \subset \mathcal{X}$ of its domain, if the Bayes optimal predictor performs better than a constant predictor in terms of squared error, then there must be some $h \in \mathcal{H}$ that also performs better than a constant predictor. This is a *weak* learning assumption because it might be that $f^*(x)$ performs *much* better than a constant predictor, but that the best $h \in \mathcal{H}$ performs only a little bit better than a constant predictor on S — this situation is still consistent with our assumption.

Nevertheless, we will show that the weak learning assumption is enough (together with our Algorithm 19 for multicalibration with respect to real valued functions $\mathcal{H}$) to boost the *weak learners* in $\mathcal{H}$ to a *strong learner* f — i.e. a model f that performs as well as the optimal model f^* with respect to squared error. In fact, the weak learning condition on $\mathcal{H}$ is both necessary and sufficient for multicalibration of f with respect to $\mathcal{H}$ to imply Bayes optimality of f. Our "boosting algorithm" will simply be our multicalibration algorithm!

First we define what we mean when we say that multicalibration with respect to $\mathcal{H}$ implies Bayes optimality. Note that $f^*(x)$ is multicalibrated with respect to any set of functions, so it is not enough to require that there *exist* Bayes optimal functions f that are multicalibrated with respect to $\mathcal{H}$. Instead, we have to require that *every* function that is multicalibrated with respect to $\mathcal{H}$ is Bayes optimal:

Definition 39 *Fix a distribution $\mathcal{D} \in \Delta\mathcal{Z}$. We say that multicalibration with respect to $\mathcal{H}$ implies Bayes optimality over $\mathcal{D}$ if for every $f : \mathcal{X} \to \mathbb{R}$ that is multicalibrated with respect to $\mathcal{D}$ and $\mathcal{H}$, we have:*

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x) - y)^2] = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x) - y)^2]$$

Where $f^*(x) = \mathbb{E}_{y\sim\mathcal{D}(x)}[y]$ is the function that has minimum squared error over the set of all functions.

Theorem 31 *Fix a distribution $\mathcal{D} \in \Delta\mathcal{Z}$. Let $\mathcal{H}$ be a class of functions that is*

closed under affine transformation. Multicalibration with respect to $\mathcal{H}$ implies Bayes optimality over $\mathcal{D}$ if and only if $\mathcal{H}$ satisfies the weak learner condition relative to $\mathcal{D}$.

Proof 56 *To avoid measurability issues we assume that models f have a countable range (which is true in particular whenever $\mathcal{X}$ is countable) — but this assumption can be avoided with more care.*

First we show that if $\mathcal{H}$ satisfies the weak learner condition relative to $\mathcal{D}$, then multicalibration with respect to $\mathcal{H}$ implies Bayes optimality over $\mathcal{D}$. Suppose not. Then there exists a function f that is multicalibrated with respect to $\mathcal{D}$ and $\mathcal{H}$, but is such that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2] > \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x)-y)^2]$$

By linearity of expectation we have:

$$\sum_{v\in R(f)} \Pr[f(x)=v]\cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2-(f^*(x)-y)^2|f(x)=v] > 0$$

In particular there must be some $v \in R(f)$ with $\Pr_{x\sim\mathcal{D}_\mathcal{X}}[f(x)=v] > 0$ such that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2|f(x)=v] > \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x)-y)^2|f(x)=v]$$

Let $S=\{x : f(x)=v\}$. Since f is calibrated, we know that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(v-y)^2|x\in S]] = \min_{c\in\mathbb{R}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(c-y)^2|x\in S]]$$

Thus by the weak learning assumption there must exist some $h\in\mathcal{H}$ such that:

$$\mathbb{E}[(v-y)^2-(h(x)-y)^2|x\in S] = \mathbb{E}[(f(x)-y)^2-(h(x)-y)^2|f(x)=v] > 0$$

By Lemma 6.1.3, there must therefore exist some $h'\in\mathcal{H}$ such that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h'(x)(y-v)|f(x)=v] > 0$$

implying that f is not *multicalibrated with respect to $\mathcal{D}$ and $\mathcal{H}$, a contradiction.*

In the reverse direction, we show that for any $\mathcal{H}$ that does not *satisfy the weak learning condition with respect to $\mathcal{D}$, then multicalibration with respect to $\mathcal{H}$ and $\mathcal{D}$ does not imply Bayes optimality over $\mathcal{D}$. In particular, we exhibit a function f such that f is multicalibrated with respect to $\mathcal{H}$ and $\mathcal{D}$, but such that:*

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2] > \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x)-y)^2]$$

Since $\mathcal{H}$ does not satisfy the weak learning assumption over $\mathcal{D}$, there must exist some set $S\subseteq\mathcal{X}$ with $\Pr[x\in S]>0$ such that

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x)-y)^2|x\in S] < \min_{c\in\mathbb{R}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(c-y)^2|x\in S]$$

but for every $h \in \mathcal{H}$:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h(x)-y)^2|x \in S] \geq \min_{c\in\mathbb{R}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(c-y)^2|x \in S]$$

.

Let $c(S) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y|x \in S]$. We define $f(x)$ as follows:

$$f(x) = \begin{cases} f^*(x) & x \notin S \\ c(S) & x \in S \end{cases}$$

We can calculate that:

$$\begin{aligned}
& \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2] \\
= & \Pr_{(x,y)\sim\mathcal{D}}[x \in S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(c(S)-y)^2|x \in S] + \Pr_{(x,y)\sim\mathcal{D}}[x \notin S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x)-y)^2|x \notin S] \\
> & \Pr_{(x,y)\sim\mathcal{D}}[x \in S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x)-y)^2|x \in S] + \Pr_{(x,y)\sim\mathcal{D}}[x \notin S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x)-y)^2|x \notin S] \\
= & \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x)-y)^2]
\end{aligned}$$

In other words, f is not Bayes optimal. So if we can demonstrate that f is multicalibrated with respect to $\mathcal{H}$ and $\mathcal{D}$ we are done. Suppose otherwise. Then there exists some $h \in \mathcal{H}$ and some $v \in R(f)$ such that

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)(y-v)|f(x)=v] > 0$$

By Lemma 6.1.3, there exists some $h' \in \mathcal{H}$ such that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h'(x)-y)^2|f(x)=v] < \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2|f(x)=v]$$

We first observe that it must be that $v = c(S)$. If this were not the case, by definition of f we would have that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h'(x)-y)^2|f(x)=v] < \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x)-y)^2|f(x)=v]$$

which would contradict the Bayes optimality of f^*. Having established that $v = c(S)$ we can calculate:

$$\begin{aligned}
& \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h'(x)-y)^2|f(x)=c(S)] \\
= & \Pr_{(x,y)\sim\mathcal{D}}[x \in S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h'(x)-y)^2|x \in S] + \\
& \Pr_{(x,y)\sim\mathcal{D}}[x \notin S, f(x)=c(S)] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h'(x)-y)^2|x \notin S, f(x)=c(S)] \\
\geq & \Pr_{(x,y)\sim\mathcal{D}}[x \in S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h'(x)-y)^2|x \in S] + \\
& \Pr_{(x,y)\sim\mathcal{D}}[x \notin S, f(x)=c(S)] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2|x \notin S, f(x)=c(S)]
\end{aligned}$$

where in the last inequality we have used the fact that by definition, $f(x) = f^*(x)$ for all $x \notin S$, and so is pointwise Bayes optimal for all $x \notin S$.

Hence the only way we can have $\mathbb{E}_{(x,y)\sim\mathcal{D}}[(h'(x) - y)^2 | f(x) = c(S)] < \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x) - y)^2 | f(x) = c(S)]$ is if:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h'(x) - y)^2 | x \in S] < \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(c(S) - y)^2 | x \in S]$$

But this contradicts our assumption that $\mathcal{H}$ violates the weak learning condition on S, which completes the proof.

Theorem 31 characterizes when *exact* multicalibration with respect to $\mathcal{H}$ implies *exact* Bayes optimality and vice versa. But our algorithm 19 only converges to approximate multi-calibration over a set of functions $\mathcal{H}$. What can we say about its convergence to approximate Bayes optimality when $\mathcal{H}$ satisfies the weak learning condition? To answer this question we'll need a quantitative version of our weak learning condition.

Definition 40 *Fix a distribution $\mathcal{D} \in \Delta\mathcal{Z}$ and a class of functions $\mathcal{H}$. Let $f^*(x) = \mathbb{E}_{y\sim\mathcal{D}(x)}[y]$ denote the true conditional label expectation conditional on x. We say that $\mathcal{H}$ satisfies the γ-*weak learner *condition relative to $\mathcal{D}$ if for every $S \subset \mathcal{X}$ with $\Pr_{x\sim\mathcal{D}_\mathcal{X}}[x \in S] > 0$, if:*

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x) - y)^2 | x \in S] < \min_{c\in\mathbb{R}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(c - y)^2 | x \in S] - \gamma$$

then there exists an $h \in \mathcal{H}$ such that:

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(h(x) - y)^2 | x \in S] < \min_{c\in\mathbb{R}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(c - y)^2 | x \in S] - \gamma$$

Definition 40 approaches Defininition 38 as $\gamma \to 0$. It says that when the Bayes optimal predictor improves over a constant predictor on set S *by at least some margin* γ, then there is some $h \in \mathcal{H}$ that does so as well. On the one hand, it is weaker than Definition 38 in that it does not require anything of $\mathcal{H}$ if the Bayes optimal predictor improves over a constant prediction by less than γ. On the other hand, it is stronger, in that it requires that some $h \in \mathcal{H}$ improve over a constant predictor on $\mathcal{H}$ by margin γ (rather than just infinitesimally) whenever doing so is possible.

Since the γ-weak learning condition does not make any requirements on $\mathcal{H}$ on sets for which $f^*(x)$ improves over a constant predictor by less than γ, the best we can hope to prove under this assumption is γ-approximate Bayes optimality, which is what we do next.

Theorem 32 *Fix any distribution $\mathcal{D} \in \Delta\mathcal{Z}$, any model $f : \mathcal{X} \to [0, 1]$, any $\gamma > 0$, any class of real valued functions $\mathcal{H}$ that satisfies the γ-weak learner condition relative to $\mathcal{D}$, and a squared error regression oracle $A_\mathcal{H}$ for $\mathcal{H}$. Let $\alpha = \gamma$ and $B = 1/\gamma$ (or any pair such that $\alpha/B = \gamma^2$). Then*

RegressionMulticalibrate$(f, \alpha, A_{\mathcal{H}}, \mathcal{D}, B)$ *halts after at most* $T \leq \frac{2}{\gamma^2}$ *many iterations and outputs a model* f_{T-1} *such that* f_{T-1} *is* 2γ*-approximately Bayes optimal over* $\mathcal{D}$*:*

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x) - y)^2] \leq \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f^*(x) - y)^2] + 2\gamma$$

where $f^*(x) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y]$ *is the function that minimizes squared error over* $\mathcal{D}$*.*

Proof 57 *At each round* t *before the algorithm halts, we have by construction that* $err_t \leq err_{t-1} - \frac{\alpha}{2B}$*, and since the squared error of* f_0 *is at most* 1*, and squared error is non-negative, we must have* $T \leq \frac{2B}{\alpha} = \frac{2}{\gamma^2}$*.*

Now suppose the algorithm halts at round T *and outputs* f_{T-1}*. It must be that* $err_T > err_{T-1} - \frac{\gamma^2}{2}$*. Suppose also that* f_{T-1} *is not* 2γ*-approximately Bayes optimal:*

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x) - y)^2 - (f^*(x) - y)^2] > 2\gamma$$

We can write this condition as:

$$\sum_{v\in[1/m]} \Pr[f_{T-1}(x) = v] \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x) - y)^2 - (f^*(x) - y)^2 | f_{T-1}(x) = v] > 2\gamma$$

Define the set:

$$S = \{v \in [1/m] : \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x) - y)^2 - (f^*(x) - y)^2 | f_{T-1}(x) = v] \geq \gamma\}$$

to denote the set of values v *in the range of* f_{T-1} *such that conditional on* $f_{T-1}(x) = v$*,* f_{T-1} *is at least* γ*-sub-optimal. Since we have both* $y \in [0, 1]$ *and* $f_{T-1}(x) \in [0, 1]$*, for every* v *we must have that* $\mathbb{E}[(f_{T-1}(x) - y)^2 - (f^*(x) - y)^2 | f_{T-1}(x) = v] \leq 1$*. Therefore we can bound:*

$$\begin{aligned} 2\gamma &< \sum_{v\in[1/m]} \Pr[f_{T-1}(x) = v] \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x) - y)^2 - (f^*(x) - y)^2 | f_{T-1}(x) = v] \\ &\leq \Pr_{(x,y)\sim\mathcal{D}}[x \in S] + (1 - \Pr_{(x,y)\sim\mathcal{D}}[x \in S])\gamma \end{aligned}$$

Solving we learn that:

$$\Pr_{(x,y)\sim\mathcal{D}}[x \in S] \geq \frac{2\gamma - \gamma}{(1 - \gamma)} \geq 2\gamma - \gamma = \gamma$$

Now observe that by the fact that $\mathcal{H}$ *is assumed to satisfy the* γ*-weak learning assumption with respect to* $\mathcal{D}$*, at the final round* T *of the algorithm, for every* $v \in S$ *we have that* h_v^T *satisfies:*

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x) - y)^2 - (h_v^T(x) - y)^2 | f_{T-1}(x) = v] \geq \gamma$$

Let $\tilde{err}_T = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(\tilde{f}_T(x) - y)^2]$ *Therefore we have:*

$$\begin{aligned} err_{T-1} - \tilde{err}_T &= \sum_{v\in[1/m]} \Pr_{(x,y)\sim\mathcal{D}}[f_{T-1}(x) = v] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(f_{T-1}(x) - y)^2 - (h_v^T(x) - y)^2 | f_{T-1}(x) = v] \\ &\geq \Pr_{(x,y)\sim\mathcal{D}}[f_{T-1}(x) \in S]\gamma \\ &\geq \gamma^2 \end{aligned}$$

We recall that $|\tilde{err}_T - err_T| \leq 1/m = \frac{\gamma^2}{2}$ *and so we can conclude that*

$$err_{T-1} - err_T \geq \frac{\gamma^2}{2}$$

which contradicts the fact that the algorithm halted at round T*, completing the proof.*

References and Further Reading

Kim et al. [2019] first studied *multi-accuracy* (what we call group conditional mean consistency in this book) for real valued functions, and gave a boosting like algorithm for obtaining it. Multicalibration with respect to real valued functions was first studied in Gopalan et al. [2022b] who gave an algorithm based on "split" and "merge" operations, related to boosting-by-branching-programs algorithms from the learning theory literature. Burhanpurkar et al. [2021] first ask the question what properties of a set of groups $\mathcal{G}$ are sufficient to guarantee that multicalibration with respect to $\mathcal{G}$ implies Bayes optimality — the answer they give (which is sufficient but not necessary) is that $\mathcal{G}$ contains refinements of the levelsets of the optimal predictor f^*. This can be viewed as a "strong learning" assumption in comparison to our "weak learning" assumption. The main results from this chapter, including the multicalibration algorithm that operates as a reduction to squared error regression, and the characterization that multicalibration implies Bayes Optimality if and only if $\mathcal{H}$ satisfies the weak learning condition comes from Globus-Harris et al. [2023].

Part II

Applications

7

Conformal Prediction

CONTENTS

Thus far we have restricted our attention to regression problems (in which the label domain $\mathcal{Y} = \mathbb{R}$), and have focused on estimating distributional quantities of conditional label distributions, like means and quantiles. In this chapter, we introduce a much more general framework for uncertainty quantification that reduces a very general uncertainty quantification problem to the problem of one dimensional quantile estimation. As a result, we will be able to draw on our development of powerful quantile estimation techniques to give an analogously powerful set of results for a much more general problem.

7.1 Prediction Sets and Nonconformity Scores

Suppose we have a distribution $\mathcal{D} \in \Delta\mathcal{Z}$ (although we will also consider the sequential prediction setting in which there need not be any distribution). Our goal is to be able to produce *prediction sets* as a function of observed features x that are likely to contain the corresponding label y. More specifically, we want to be able to find a function $\mathcal{T} : \mathcal{X} \to 2^{\mathcal{Y}}$ mapping unlabelled examples x to *subsets* of labels $\mathcal{T}(x)$ that have the property that the true label is contained within $\mathcal{T}(x)$ with some specified level of confidence $1 - \delta$:

$$\Pr[y \in \mathcal{T}(x)] \approx 1 - \delta$$

FIGURE 7.1
Images x about which we might have uncertainty about their labels y.

We leave unspecified for now what distribution this probability is taken over, because we will consider a spectrum of guarantees of increasing strength, mirroring our treatment of mean and quantile estimation. For example, we can ask for marginal guarantees, group conditional guarantees, calibrated guarantees, or ask for guarantees that hold empirically on adversarially chosen sequences. Prediction sets can take different forms: when we are facing a regression problem ($\mathcal{Y} = \mathbb{R}$) is is natural (but not necessary) for a prediction set to take the form of an *interval*: $\mathcal{T}(x) = [a, b]$ for some $a < b \in \mathbb{R}$. On the other hand, for a multiclass classification problem (when $\mathcal{Y}$ is some unordered discrete set), prediction sets correspond to subsets of labels — e.g. we might have $\mathcal{T}(x) = \{\text{Blueberry Muffin, Chihuahua}\}$ for x representing images from Figure 7.1.

Prediction sets are a very attractive way to quantify uncertainty: their size represents a quantitative *degree* of uncertainty. For example, if $\mathcal{T}(x)$ is a singleton, this represents certainty at the specified $1 - \delta$ level in a particular point prediction. But the contents of the set also provides insight into *where* the uncertainty lies. For example in a classification problem, there might be a high degree of uncertainty in the specific label, but a well crafted prediction set might nevertheless tell us that our uncertainty is concentrated in a region that corresponds to the same downstream action. Say, in a computer vision setting, we might be unsure of the breed of dog in front of us—so $\mathcal{T}(x)$ contains half a dozen different labels, corresponding to different breeds—but despite this uncertainty in the specifics, this prediction set gives us a high degree of confidence in what action to take—apply the breaks.

The main difficulty with thinking about producing prediction sets is that they are very high dimensional objects: In a k-label multiclass classification setting, there are 2^k different prediction sets. The main idea in *conformal pre-*

diction is to reduce these high dimensional prediction sets to one-dimensional objects using a *non-conformity score function* $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

7.1.1 Non-Conformity Scores

A "non-conformity score function" $s(x, y)$ is typically built from some *model* h for making point predictions. As a running example, lets imagine that we are in the regression setting ($\mathcal{Y} = \mathbb{R}$) and we have solved a linear regression problem to produce a model $h : \mathcal{X} \to \mathcal{Y}$ that makes point predictions. Intuitively, the "non-conformity score" $s(x, y)$ is supposed to communicate some measure by which the label y differs from the prediction of the model $h(x)$. The simplest (often too simple) non-conformity score in this setting is:

$$s(x, y) = |h(x) - y|$$

which simply measures the deviation of the label y from the point prediction $h(x)$.

Any non-conformity score function s can be used to parameterize a (now one dimensional) family of prediction sets $\mathcal{T}_s : \mathcal{X} \times \mathbb{R} \to 2^{\mathcal{Y}}$ as follows:

$$\mathcal{T}_s(x, \tau) = \{\hat{y} : s(x, \hat{y}) \leq \tau\}$$

The prediction set $\mathcal{T}(x, \tau)$ simply contains all labels $\hat{y}$ that would produce nonconformity score at most τ when paired with x: $s(x, \hat{y}) \leq \tau$. In the case of our simple regression running example, this would simply correspond to the interval centered at the point prediction $h(x)$ that has width 2τ: $\mathcal{T}_s(x, \tau) = [h(x) - \tau, h(x) + \tau]$. Although simple, a clear disadvantage of this non-conformity score is that for fixed τ, every prediction interval $\mathcal{T}(x, \tau)$ has the same width — so for methods that use a fixed value of τ — which roughly speaking are those methods that promise only marginal coverage — the prediction intervals do not give us any insight into *which examples* we have more uncertainty about compared to others.

There are many other non-conformity scores that are in wide use. For example, rather than training a regression model h that aims to predict the mean of the conditional label distribution $\mathcal{D}_{\mathcal{Y}}(x)$ (as linear regression does), we could train quantile regression models $h_{\delta/2}(x)$, $h_{1-\delta/2}(x)$ that try and predict the $\delta/2$ and $1 - \delta/2$ quantiles of the conditional label distribution $\mathcal{D}_{\mathcal{Y}}(x)$ instead. Then a natural non-conformity score would be:

$$s(x, y) = \max(h_{\delta/2}(x) - y, y - h_{1-\delta/2}(x))$$

This score starts with the candidate interval that directly arises from the quantile regression method $[h_{\delta/2}(x), h_{1-\delta/2}(x)]$, and measures how far the label y is from the interval — taking a positive value when y falls outside of the interval and a negative value when it falls inside. If the interval is correct, then the $1 - \delta$ quantile of the nonconformity score distribution will be 0 — picking

threshold $\tau = 0$ will get the target marginal coverage. But if the interval induced by the quantile regression method is not correct, then choosing different thresholds τ can systematically widen or shorten the prediction interval by τ on each end: $\mathcal{T}_s(x, \tau) = [h_{\delta/2}(x) - \tau, h_{1-\delta/2}(x) + \tau]$. This non-conformity score has the advantage that even for a fixed value of τ, the prediction intervals $\mathcal{T}_s(x, \tau)$ can have very different widths, depending on the predictions of the models $h_{\delta/2}(x)$ and $h_{1-\delta/2}(x)$.

What about for multi-class classification problems, in which $\mathcal{D}_{\mathcal{Y}}(x)$ is a discrete distribution over k possible labels? To build intuition, suppose we were given the true conditional distribution over labels given x: For each label $\hat{y} \in [k]$, $p^*_x(\hat{y}) = \Pr[y = \hat{y}|x]$. Let $\pi_{p^*_x}$ be the permutation on labels that puts them in decreasing sorted order by their underlying probability: so $p^*_x(\pi_{p^*_x}(1)) \geq p^*_x(\pi_{p^*_x}(2)) \geq \ldots \geq p^*_x(\pi_{p^*_x}(k))$. How would we find the smallest prediction set that contains the true label with probability at least $1 - \delta$? We would greedily start adding labels to our prediction set in order of their probabilities (highest probability to lowest) until the cumulative probability of the labels in our prediction set exceeded $1 - \delta$. To say this more formally, for each $t \in [k]$, let $C(t, p^*_x) = \sum_{i=1}^{t} p^*_x(\pi_{p^*_x}(i))$ denote the cumulative probability of the top t labels in likelihood sorted order. We would choose the prediction set:

$$\mathcal{T}(x) = \{\hat{y} : C(\pi^{-1}_{p^*_x}(\hat{y}), p^*_x) \leq 1 - \delta\}$$

Now suppose we have a method that gives us a score function $p_x : \mathcal{Y} \to [0, 1]$ for each example x. We might *hope* that p_x is the true probability distribution over labels, but we have no strong reason to believe that it is. For example, p_x might be the softmax outputs of the final layer of a neural network. We can nevertheless define the same quantities with respect to p_x: π_{p_x} is the permutation that places the labels in descending order according to p_x: $p_x(\pi_{p_x}(1)) \geq \ldots \geq p_x(\pi_{p_x}(k))$, and $C(t, p_x) = \sum_{i=1}^{t} p_x(\pi_{p_x}(i))$ denotes the cumulative "probability" of the top t labels according to p_x. We can then define a non-conformity score:

$$s(x, y) = C(\pi^{-1}_{p_x}(y), p_x)$$

In the event that p_x really is the true conditional label distribution conditional on x, then using this non-conformity score, the prediction sets $\mathcal{T}_s(x, \tau) = \{\hat{y} : C(\pi^{-1}_{p_x}(\hat{y}), p_x) \leq \tau\}$ are the minimum size prediction sets with coverage probability τ — and even if they are not, there exists some τ that leads to coverage with the target coverage probability.

There are plenty of other non-conformity scores that one could consider. But for the rest of this chapter, we won't worry about what the non-conformity score is — the techniques we discuss will work for *any* choice of nonconformity score.

7.2 A Weak Guarantee: Marginal Coverage in Expectation

In this section we will consider the problem of using a sample of data $D \sim \mathcal{D}^n$ (that we will call a *calibration set*) to produce prediction sets $\mathcal{T}(x)$ that obtain the following coverage guarantee on *new samples* $(x, y) \sim \mathcal{D}$ that are not contained in D.

$$1 - \delta \leq \Pr_{D \sim \mathcal{D}^n, (x,y) \sim \mathcal{D}}[y \in \mathcal{T}(x)] \leq 1 - \delta + \frac{1}{n+1}$$

This is a marginal coverage guarantee because the probability is over x as well as y, and is unconditioned. We call it a marginal guarantee in *expectation* because the probability is also taken over the calibration set D, and so could be expressed as:

$$1 - \delta \leq \mathop{\mathbb{E}}_{D \sim \mathcal{D}^n}\left[\Pr_{(x,y) \sim \mathcal{D}}[y \in \mathcal{T}(x)]\right] \leq 1 - \delta + \frac{1}{n+1}$$

This is in contrast to theorems we will see later that have high probability guarantees over the randomness of the calibration set D. Nevertheless, this guarantee is very simple to obtain, and has a very mild (inverse linear) dependence on n which makes it attractive.

Algorithm 20 SplitConformal(D, s, δ)

Let τ be the smallest value such that:

$$\sum_{i=1}^{n} \mathbb{1}[s(x_i, y_i) \leq \tau] \geq (1 - \delta)(n + 1)$$

i.e. τ is an empirical $\frac{\lceil (1-\delta)(n+1) \rceil}{n}$ quantile of D.

Output the function:

$$\mathcal{T}_D(x) = \{\hat{y} : s(x, \hat{y}) \leq \tau\}$$

The algorithm is simple, and given in Algorithm 20. Informally, it takes as input a calibration dataset D (of any size), a non-conformity score s (which must be defined independently of the calibration dataset D), and a target miscoverage rate δ. It computes a threshold τ that comes as close as possible to being an empirical $(1-\delta)$-quantile of the set of non-conformity scores induced by S on D (up to a bias correction term of roughly $\frac{n+1}{n}$), and then outputs the function $\mathcal{T}(x) \equiv \mathcal{T}_s(x, \tau)$ that uses the fixed threshold τ for every example x. As we have done previously in our discussion of quantile estimation, we will assume that the distribution on which we want to compute quantiles (which in

this case is the induced distribution on non-conformity scores) is *continuous*, which simplifies things. Recall that we can always enforce this assumption by adding arbitrarily small amounts of noise from any continuous distribution to the non-conformity scores.

Theorem 33 *Fix any distribution $\mathcal{D} \in \Delta Z$, any $0 \leq \delta \leq 1$ and any non-conformity score $s : \mathcal{Z} \to \mathbb{R}$. Assume the induced distribution on non-conformity scores $s(x, y)$ for $(x, y) \sim \mathcal{D}$ is continuous. Let $D \sim \mathcal{D}^n$ be a dataset of n points sampled i.i.d. from $\mathcal{D}$. Then for the function $\mathcal{T}_D(x)$ output by SplitConformal(D, s, δ) (Algorithm 20) we have that:*

$$1 - \delta \leq \mathop{\mathbb{E}}_{D \sim \mathcal{D}^n} \left[\Pr_{(x,y) \sim \mathcal{D}} [y \in \mathcal{T}_D(x)] \right] \leq 1 - \delta + \frac{1}{n+1}$$

In fact, the only property we will use about the distribution from which D and (x, y) are jointly drawn is that it is exchangable, which means permutation invariant — we will not need the stronger property that the points are drawn i.i.d.

Proof 58 (Proof of Theorem 33) *Because we have assumed that the non-conformity score distribution on $s(x, y)$ is continuous, with probability 1, there are no ties amongst the non-conformity scores in D — i.e. for all $i \neq j$, $s(x_i, y_i) \neq s(x_j, y_j)$. Renumber the datapoints in D in increasing order of their nonconformity scores — i.e. such that $s(x_1, y_1) < s(x_2, y_2) < \ldots < s(x_n, y_n)$. Let i^* be the unique index such that $s(x_{i^*}, y_{i^*}) = \tau$. $i^* = \lceil (1 - \delta)(n + 1) \rceil$.*

Imagine the dataset $D' = D \cup (x, y)$ containing $n + 1$ elements. Consider the event $y \in \mathcal{T}_D(x)$. This occurs exactly when $s(x, y) < \tau$, which is exactly the event that the pair (x,y) occurs before the pair (x_{i^*}, y_{i^*}) when we sort the $n + 1$ points in D' by their non-conformity scores. But since all of the points in D' are exchangable, by symmetry it must be that point (x,y) will have rank that is uniformly random in $\{1, 2, \ldots, n + 1\}$ when put in sorted order within D'. Thus the event that $y \in \mathcal{T}_D(x)$ is the event that (x, y) has rank that is less than i^*, which is:

$$\begin{aligned}
\Pr_{D,(x,y)}[y \in \mathcal{T}_D(x)] &= \frac{i^*}{n+1} \\
&= \frac{\lceil (1-\delta)(n+1) \rceil}{n+1} \\
&\geq \frac{(1-\delta)(n+1)}{n+1} \\
&= (1 - \delta)
\end{aligned}$$

We can similarly calculate:

$$\begin{aligned}\Pr_{D,(x,y)}[y \in \mathcal{T}_D(x)] &= \frac{i^*}{n+1} \\ &= \frac{\lceil (1-\delta)(n+1) \rceil}{n+1} \\ &\leq \frac{(1-\delta)(n+1)+1}{n+1} \\ &= (1-\delta) + \frac{1}{n+1}\end{aligned}$$

which completes the proof.

7.3 Dataset Conditional Bounds

The first way in which we might strengthen Theorem 33 is to give a bound that holds with high probability over the draw of $D \sim \mathcal{D}^n$ rather than only in expectation. To do this, all we need to do is find a high probability estimate for the $1-\delta$ quantile of the non-conformity score distribution, which is a problem that we already solved in Chapter 2!

Algorithm 21 HighProbabilitySplitConformal(D, s, δ, γ)

Let τ be the smallest value such that:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[s(x_i, y_i) \leq \tau] \geq (1-\delta) + \sqrt{\frac{\log(2/\gamma)}{2n}}$$

Output the function:

$$\mathcal{T}_D(x) = \{\hat{y} : s(x, \hat{y}) \leq \tau\}$$

Theorem 34 *Fix any distribution $\mathcal{D} \in \Delta Z$, any $0 \leq \delta \leq 1$ and any non-conformity score $s : \mathcal{Z} \to \mathbb{R}$. Assume the induced distribution on non-conformity scores $s(x, y)$ for $(x, y) \sim \mathcal{D}$ is continuous. Let $D \sim \mathcal{D}^n$ be a dataset of n points sampled i.i.d. from $\mathcal{D}$. Then for the function $\mathcal{T}_D(x)$ output by SplitConformal(D, s, δ, γ) (Algorithm 20) we have that with probability $1-\gamma$ over the draw of $D \sim \mathcal{D}^n$:*

$$1-\delta \leq \Pr_{(x,y)\sim\mathcal{D}}[y \in \mathcal{T}_D(x)] \leq 1-\delta+2\sqrt{\frac{\log(2/\gamma)}{2n}} + \frac{1}{n}$$

Proof 59 *By construction, τ is an empirical q-quantile for the empirical distribution of scores $s(x, y)$ over D for:*

$$(1-\delta) + \sqrt{\frac{\log(2/\gamma)}{2n}} \leq q \leq (1-\delta) + \sqrt{\frac{\log(2/\gamma)}{2n}} + \frac{1}{n}$$

From Theorem 2, we have that with probability $1-\gamma$, τ is therefore a q' quantile for the distribution of scores $s(x, y)$ over $\mathcal{D}$ for:

$$q - \sqrt{\frac{\log(2/\gamma)}{2n}} \leq q' \leq q + \sqrt{\frac{\log(2/\gamma)}{2n}}$$

Combining these two bounds, we have that with probability $1-\gamma$, τ is a q' quantile for $\mathcal{D}$ such that:

$$(1-\delta) \leq q' \leq (1-\delta) + 2\sqrt{\frac{\log(2/\gamma)}{2n}} + \frac{1}{n}$$

Since $y \in T_D(x)$ exactly when $s(x, y) \leq \tau$, we have that $\Pr[y \in T_D(x)] = q'$, which completes the proof.

7.4 Dataset and Group Conditional Bounds

We can also ask for stronger than marginal guarantees. For example, given an arbitrary collection of groups $\mathcal{G} \subseteq 2^{\mathcal{X}}$, we can ask for group conditional coverage. That is, we can ask for prediction sets $\mathcal{T}_D(x)$ that have the property that for every $g \in \mathcal{G}$:

$$\Pr_{(x,y)\sim\mathcal{D}}[y \in \mathcal{T}_D(x) | g(x) = 1] = 1 - \delta.$$

To obtain a guarantee like this, it will no longer be sufficient to parameterize $\mathcal{T}_D(x)$ with a single threshold τ: instead we will parameterize $\mathcal{T}_D^f(x)$ with a function $f : \mathcal{X} \to [0, 1]$:

$$\mathcal{T}_D^f(x) = \{\hat{y} : s(x, \hat{y}) \leq f(x)\}.$$

So: $\mathcal{T}_D(x)$ will have group conditional coverage guarantees if and only if $f(x)$ has group conditional quantile marginal consistency guarantees. We know how to do this using Algorithm 13! We can apply it here:

Algorithm 22 GroupSplitConformal$(D, s, \mathcal{G}, \delta, \gamma, \rho, \sigma, \eta)$

Let $q = 1 - \delta$.

Let λ^* be a solution to the optimization problem:

$$\text{Minimize}_\lambda \mathop{\mathbb{E}}_{(x,y)\sim D}\left[L_q\left(\hat{f}(x;\lambda), s(x,y)\right)\right] + \eta||\lambda||_1$$

Such that:

$$\hat{f}(x;\lambda) \equiv \sum_{g\in\mathcal{G}} \lambda_g \cdot g(x)$$

Output the function:

$$\mathcal{T}_D^{f(x;\lambda^*)}(x) = \{\hat{y} : s(x,\hat{y}) \leq f(x;\lambda^*)\}$$

Theorem 35 *Fix any $\gamma, \delta, \eta > 0$. Let $\mathcal{G}$ be any collection of groups. Let $D \sim \mathcal{D}^n$ consist of n samples (x, y) from a distribution $\mathcal{D}$ that is ρ-Lipschitz and σ-anti-Lipschitz. Then with probability $1 - \gamma$ over the draw of $D \sim \mathcal{D}^n$, the function $\mathcal{T}_D^f(x)$ output by GroupSplitConformal$(D, s, \mathcal{G}, \delta, \gamma, \rho, \sigma, \eta))$ (Algorithm 22) satisfies for any group g that has mass $\mu(g) \geq \alpha$:*

$$1 - \delta - \sqrt{\frac{\alpha}{\mu(g)}} \leq \Pr_{(x,y)\sim D}[y \in \mathcal{T}_D^f(x) | g(x) = 1] \leq 1 - \delta + \sqrt{\frac{\alpha}{\mu(g)}}$$

for:

$$\alpha \leq \frac{2\eta\rho}{\sigma} + 8\rho\left(\frac{1}{\eta} + 1\right)\sqrt{\frac{\ln\left(\frac{2}{\gamma}\right) + |\mathcal{G}|\ln\left(1 + 2\sqrt{n}\right)}{2n}}$$

Choosing η to minimize this expression gives:

$$\alpha \leq O\left(\frac{\rho}{\sqrt{\sigma}} \cdot \left(\frac{\ln\left(\frac{1}{\gamma}\right) + |\mathcal{G}|\ln(n)}{n}\right)^{1/4}\right)$$

Proof 60 *By Theorem 20, with probability $1-\gamma$, the function $f(x;\lambda^*)$ satisfies α-approximate group conditional marginal quantile consistency on $\mathcal{D}$ on the set of groups $g \in \mathcal{G}$ with $\mu(g) \geq \alpha$ and target quantile $q = 1 - \delta$ for:*

$$\alpha \leq \frac{2\eta\rho}{\sigma} + 8\rho\left(\frac{1}{\eta} + 1\right)\sqrt{\frac{\ln\left(\frac{2}{\gamma}\right) + |\mathcal{G}|\ln\left(1 + 2\sqrt{n}\right)}{2n}}$$

This means that for every group $g \in \mathcal{G}$ with $\mu(g) \geq \alpha$:

$$\left(\Pr[s(x,y) \leq f(x;\lambda^*) | g(x) = 1] - (1 - \delta)\right)^2 \leq \frac{\alpha}{\mu(g)}$$

or equivalently:

$$(1-\delta) - \sqrt{\frac{\alpha}{\mu(g)}} \leq \Pr[s(x,y) \leq f(x;\lambda^*)|g(x)=1] \leq (1-\delta) + \sqrt{\frac{\alpha}{\mu(g)}}$$

The result then follows the fact that $y \in \mathcal{T}_D^f(x)$ *exactly when* $s(x,y) \leq f(x;\lambda^*)$.

7.5 Multivalid Bounds

In moving to group conditional coverage, we have started defining our prediction sets $\mathcal{T}_D^f(x)$ not with a single threshold τ, but instead with a function $f(x)$ that maps each example to a potentially different threshold. In doing so, we have created the possibility that our coverage guarantees are no longer *threshold calibrated* — i.e. that conditional on the threshold $f(x)$ that we choose, our coverage guarantees may now fail to hold. In fact, without threshold calibration, it is possible to abuse the model and obtain group conditional coverage without providing any useful information about the data at all. Consider the randomized predictor which predicts the empty prediction set (and hence definitely fails to cover the label) with probability δ and predicts the full prediction set (and hence definitely covers the label) with probability $1-\delta$. This predictor has $1-\delta$ marginal coverage not just overall, but conditional on membership in each group! And yet it is defined independently of the data, and so provides no useful insight. But observe that this predictor would badly fail a threshold calibration test. Conditional on the chosen threshold, the coverage rate is either 0 or 1, in either case bounded away from the target $1-\delta$.

To correct for this, we can ask for coverage guarantees that hold conditional on both group membership and the value of the chosen threshold, which are called *multivalid coverage guarantees*. Specifically, we want that for each $g \in \mathcal{G}$ and for each $v \in R(f)$:

$$\Pr_{(x,y)\sim\mathcal{D}}[y \in \mathcal{T}_D^f(x)|g(x)=1, f(x)=v] = 1-\delta$$

Once again, we see that $\mathcal{T}_D^f$ will satisfy multivalid coverage guarantees if and only if f satisfies quantile multicalibration for target quantile $q = 1-\delta$. And once again, we know how to find such an f — use Algorithm 16!

Algorithm 23 MultivalidSplitConformal$(D, s, \mathcal{G}, \delta, \alpha, \gamma, \rho)$

Let $m = \frac{\rho^2}{2\alpha}$, $q = 1 - \delta$.
Let $f_0(x) = 0$ for all x and $t = 0$.
while f_t is not α-approximately quantile multicalibrated with respect to $\mathcal{G}$ and q: **do**
Let:

$$(v_t, g_t) \in \underset{(v,g) \in R(f_t) \times \mathcal{G}}{\arg\max} \Pr_{(x,y)\sim D}[f_t(x) = v, g(x) = 1] \left(q - \Pr_{(x,y)\sim D}[y \leq v | f_t(x) = v, g(x) = 1]) \right)^2$$

$$\tilde{v}_t = \arg\min_v \left| \Pr_{(x,y)\sim \mathcal{D}}[y \leq v | f_t(x) = v_t, g_t(x) = 1] - q \right| \text{ and } v_t' = \text{Round}(\tilde{v}_t; m)$$

Let $f_{t+1} = h(x; f_t, v_t \to v_t', g_t)$ and $t = t + 1$.
Output the function:

$$\mathcal{T}_D^{f_t}(x) = \{\hat{y} : s(x, \hat{y}) \leq f_t(x)\}$$

Theorem 36 *Fix any $\gamma, \delta, \alpha > 0$. Let $\mathcal{G}$ be any collection of groups. Let $D \sim \mathcal{D}^n$ consist of n samples (x, y) drawn from a distribution $\mathcal{D}$ that is ρ-Lipschitz. With probability $1 - \gamma$ over the draw of $D \sim \mathcal{D}^n$, the function $\mathcal{T}_D^f$ output by MultivalidSplitConformal$(D, s, \mathcal{G}, \delta, \alpha, \gamma, \rho)$ satisfies for every group $g \in \mathcal{G}$ and threshold $v \in R(f)$:*

$$1-\delta-\sqrt{\frac{\alpha'}{\Pr_{(x,y)\sim\mathcal{D}}[g(x) = 1, f(x) = v]}} \leq \Pr_{(x,y)\sim\mathcal{D}}[y \in \mathcal{T}_D^f(x) | g(x) = 1, f(x) = v] \leq$$

$$1 - \delta + \sqrt{\frac{\alpha'}{\Pr_{(x,y)\sim\mathcal{D}}[g(x) = 1, f(x) = v]}}$$

for:

$$\alpha' = \alpha + 42\sqrt{\frac{3\rho^2 \left(\ln(\frac{4\pi^2 T^2}{3\gamma}) + T \ln(\frac{\rho^4 |\mathcal{G}|}{\alpha^2}) \right)}{2\alpha n}}$$

Choosing α to optimize the bound, we get:

$$\alpha' \in \tilde{O}\left(\left(\frac{\rho^3 \ln\left(\frac{\rho^4|\mathcal{G}|}{\delta}\right)}{n} \right)^{1/5} \right)$$

Proof 61 *By Theorem 26, we have that with probability $1 - \gamma$, the final function f_t satisfies α'-approximate quantile multicalibration for target quantile*

$1-\delta$, *groups* $\mathcal{G}$, *and:*

$$\alpha' \leq \alpha + 42\sqrt{\frac{3\rho^2\left(\ln(\frac{4\pi^2T^2}{3\gamma}) + T\ln(\frac{\rho^4|\mathcal{G}|}{\alpha^2})\right)}{2\alpha n}}$$

This means that for every group $g \in \mathcal{G}$ *and* $v \in R(f_t)$*:*

$$\Pr_{(x,y)\sim\mathcal{D}}[f_t(x) = v | g(x) = 1]\left(\Pr_{(x,y)\sim\mathcal{D}}[s(x,y) \leq f_t(x) | g(x) = 1, f_t(x) = v] - (1-\delta)\right)^2$$
$$\leq Q_2(f_t, g) \leq \frac{\alpha}{\mu(g)}$$

Dividing through and taking the square root we obtain:

$$\left|\Pr_{(x,y)\sim\mathcal{D}}[s(x,y) \leq f_t(x) | g(x) = 1, f_t(x) = v] - (1-\delta)\right| \leq \sqrt{\frac{\alpha'}{\Pr_{(x,y)\sim\mathcal{D}}[g(x) = 1, f_t(x) = v]}}$$

The theorem then follows since $y \in \mathcal{T}_D^{f_t}(x)$ *exactly when* $s(x,y) \leq f_t(x)$*.*

7.6 Sequential Conformal Prediction

So far we have considered the problem of conformal prediction in the batch setting, in which we have a dataset of labelled examples D that we can use to train a model that defines a prediction set function $\mathcal{T} : \mathcal{X} \to 2^{\mathcal{Y}}$ that we can later deploy to produce prediction sets $\mathcal{T}(x)$. A major advantage of this kind of approach is that we do not need to observe labels at test time, but a major disadvantage is that we need to make strong assumptions about the test time distribution — generally that it is identical to the training distribution, and that it is distributed independently — or is at least exchangable.

In this section we apply the techniques we have developed to the sequential conformal prediction problem, which can be described as the following interaction between a learner and an adversary. In rounds $t \in \{1, \ldots, T\}$

1. The adversary chooses a feature vector $x_t \in \mathcal{X}$ and a distribution over labels $y_t \in \mathcal{Y}$.
2. The learner produces a prediction set $\mathcal{T}_{\pi^{<t}}(x_t)$.
3. The learner observes the realized label y_t.

This interaction generates a transcript $\pi = \{(x_1, \mathcal{T}_{\pi^{<1}}(x_1), y_1), \ldots, (x_T, \mathcal{T}_{\pi^{<T}}(x_T), y_T)\}$. The learner is an algorithm mapping transcript prefixes $\pi^{<t}$ and feature vectors x_t to prediction sets $\mathcal{T}_{\pi^{<t}}(x_t)$, adn the adversary is a mapping from transcript prefixes $\pi^{<t}$ to pairs of feature vectors and label distributions $\mathcal{X} \times \Delta\mathcal{Y}$.

The adversary may be arbitrary, or we may impose restrictions on the label distributions that she chooses.

The prediction sets we study will continue to be based on non-conformity score functions s — but since we no longer require exchangability, we will also allow the non-conformity score function s_t to potentially change at every round. So, for example, if our non-conformity score function is based on a model f, we can use a model f_t that is retrained on all of the examples seen so far, at each round — something that breaks the exchangability of the non-conformity scores of past and future data by introducing a dependence between the past data and the non-conformity score.

7.6.1 Sequential Marginal Coverage Guarantees

We can derive algorithms for adversarial sequential conformal prediction from our algorithms for online sequential quantile prediction. For example, we can use Algorithm 2 (which promises *marginal* quantile consistency against an adversary) to obtain a sequential conformal prediction algorithm with a corresponding marginal coverage guarantee against an adversary. To talk about coverage rates in the sequential setting, we write $\Pr_{(x_t, \mathcal{T}_t(x_t), y_t) \sim \pi}[\cdot]$ to denote the uniformly random selection of a record $(x_t, \mathcal{T}_t(x_t), y_t)$ from a transcript of T records $\pi = \{(x_t, \mathcal{T}_t(x_t), y_t)\}_{t=1}^{T}$.

Algorithm 24 Adversarial-Marginal-Conformal-Predictor(δ, η, T)

Let $q = 1 - \delta + \frac{1+\eta}{\eta T}$ and $p_1 = 0$
for $t = 1$ to T **do**
  Obtain non-conformity score s_t and observe x_t.
  Predict

$$\mathcal{T}_t(x_t) = \{\hat{y} : s_t(x_t, \hat{y}) \leq p_t\}$$

  Observe y_t.
  Let $p_{t+1} = p_t + \eta(q - \mathbb{1}[s_t(x_t, y_t) \leq p_t])$

Theorem 37 *Fix any $\delta, \eta > 0$. Paired with any adversary, Adversarial-Marginal-Conformal-Predictor(δ, η) (Algorithm 24) produces a transcript such that:*

$$(1 - \delta) \leq \Pr_{(x_t, \mathcal{T}_t(x_t), y_t) \sim \pi}[y_t \in \mathcal{T}_t(x_t)] \leq 1 - \delta + 2\frac{1+\eta}{\eta T}$$

Proof 62 *This is an application of Algorithm 2. Thus we can apply Theorem 6 to conclude that the sequence of thresholds p_t produced satisfy α-approximate marginal quantile consistency with respect to target quantile q and the sequence*

of non-conformity scores $s(x_t, y_t)$ for $\alpha \leq \frac{1+\eta}{\eta T}$. This means that:

$$q - \frac{1+\eta}{\eta T} \leq \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}[s_t(x_t, y_t) \leq p_t] \leq q + \frac{1+\eta}{\eta T}$$

Plugging in our definition of q and noting that $y_t \in \mathcal{T}_t(x_t)$ exactly when $s_t(x_t, y_t) \leq p_t$ completes the proof.

7.6.2 Sequential Multivalid Guarantees

We can similarly ask for multivalid coverage guarantees in the sequential setting — i.e. coverage guarantees that remain valid conditional on both the predicted threshold p_t and on group membership. To do this, we apply Algorithm 18, our algorithm for obtaining quantile multicalibrated predictions in the sequential setting.

Algorithm 25 Online-Multivalid-Conformal-Predictor($\mathcal{G}, m, r, \eta, \delta$)

Let $q = 1 - \delta$.
for $t = 1$ to T **do**
Obtain non-conformity score s_t and observe x_t and compute

$$C^i_{t-1}(x_t) = \sum_{g \in \mathcal{G}(x_t)} \exp(\eta V^{g,i}_{t-1}) - \exp(-\eta V^{g,i}_{t-1})$$

for all $i \in [m]$, with $V^{g,i}_{t-1}$ defined as:

$$V^{g,i}_{t-1} = \sum_{\ell \in S(\pi^{\leq t-1}, g, i)} (\mathbb{1}[s_\ell(x_\ell, y_\ell) \leq p_\ell] - q).$$

if $C^m_{t-1}(x_t) < 0$ **then**
Select $p_t = 1$.
else if $C^1_{t-1}(x_t) > 0$ **then**
Select $p_t = 0$.
else
Select $i^* \in [m]$ such that such that $C^{i^*}_{t-1}(x_t) \cdot C^{i^*+1}_{t-1}(x_t) \leq 0$.
Compute $p \in [0, 1]$ such that:

$$p \cdot C^{i^*}_{t-1}(x_t) + (1 - p) \cdot C^{i^*+1}_{t-1}(x_t) = 0$$

Select $p_t = \frac{i^*}{m} - \frac{1}{rm}$ with probability p and select $p_t = \frac{i^*}{m}$ with probability $1 - p$.
Predict:

$$\mathcal{T}_t(x_t) = \{\hat{y} : s_t(x_t, \hat{y}) \leq p_t\}$$

Observe y_t
Let $\pi^{<t+1} = \pi^{<t} \circ (x_t, p_t, y_t)$

Theorem 38 *Fix any set of groups $\mathcal{G}$, $m, r \geq 0$ and $q \in [0, 1]$. Let $\eta = \sqrt{\frac{\log(2|\mathcal{G}|m)}{2T}} < 1$. Fix $\delta > 0$ Fix any adversary who is constrained for each t to playing label distributions such that the induced distribution on non-conformity scores $s_t(x_t, y_t)$ is ρ-Lipschitz, which together with Online-Multivalid-Conformal-Predictor($\mathcal{G}, m, r, \eta, \delta$) (Algorithm 18) fixes a distribution on transcripts π. We have that with probability $1-\gamma$ over the randomness of π, for every group $g \in \mathcal{G}$ and every bucket $i \in [m]$:*

$$1-\delta-\frac{\alpha}{\mu_\pi(g,i)} \leq \Pr_{(x_t, \mathcal{T}_t(x_t), y_t) \sim \pi}[y_t \in \mathcal{T}(x_t) | p_t \in B_m(i), g(x) = 1] \leq 1-\delta+\frac{\alpha}{\mu_\pi(g,i)}$$

where

$$\mu_\pi(g,i) = \Pr_{(x_t,\mathcal{T}_t(x_t),y_t)\sim\pi}[p_t \in B_m(i), g(x_t) = 1]$$

and

$$\alpha \leq \frac{1}{\rho r m} + 4\sqrt{\frac{2\ln\left(\frac{2|\mathcal{G}|m}{\gamma}\right)}{T}}$$

Proof 63 *We can apply Theorem 29 to conclude that the sequence of thresholds p_t is (α, m)-approximately quantile multicalibrated with probability $1-\gamma$, for target quantile $q = 1-\delta$. Translating this guarantee into our notation, this means that for all buckets $i \in [m]$ and groups $g \in \mathcal{G}$:*

$$\mu_\pi(g,i) \cdot \left| q - \Pr_{(x_t,\mathcal{T}_t(x_t),y_t)\sim\pi}[\mathbb{1}[s_t(x_t,y_t) \leq p_t | p_t \in B_m(i), g(x) = 1]\right| \leq \alpha$$

The theorem then follows, since $s_t(x_t,y_t) \leq p_t$ exactly when $y \in \mathcal{T}_t(x_t)$.

References and Further Reading

See Shafer and Vovk [2008] for a classical introduction to conformal prediction and Angelopoulos and Bates [2021] for an excellent recent survey. The batch conformal prediction algorithms we present here are variants of "split conformal prediction" that use a held out calibration set — this general idea dates back to [Papadopoulos et al., 2002] and was studied in detail by Lei et al. [2018]. Romano et al. [2019] introduced quantile regression as the basis of a non-conformity score into the conformal prediction literature, and Angelopoulos et al. [2020] introduced a non-conformity score based on the soft-max output of a neural network for classification problems and demonstrated its utility on ImageNet. Romano et al. [2020] introduced the notion of group-conditional coverage as a fairness goal, and proposed a method for obtaining it for disjoint groups. Foygel Barber et al. [2020] also studied group conditional coverage guarantees for intersecting groups, and gave a conservative approach based on separately calibrating a threshold for each group, and then on a new example using the largest threshold of any group that the new example is a member of. The kind of "high probability" dataset conditional marginal guarantees we present were studied by Park et al. [2019]. Gibbs and Candes [2021] studied sequential conformal prediction with marginal coverage guarantees, and derived the algorithm we present here. Gupta et al. [2022], Bastani et al. [2022], Jung et al. [2022] introduced techniques from multicalibration into the conformal prediction literature and showed how to obtain group-wise and threshold calibrated coverage for arbitrary group structures in both the batch and sequential settings.

8

Distribution Shift

CONTENTS

Thus far we have studied prediction in two very different models:

1. In the *batch* or *distributional* setting we assume that we have sample access to a distribution $\mathcal{D}$ which we can use to train a model, that we can then deploy; it has guarantees on new data drawn from the *same* distribution.
2. In the *sequential adversarial* setting we assume data arrives sequentially and can be worst-case/generated by an adversary. But in order to make progress we assume that we learn the true label after each prediction.

But what if we want the best of both worlds — to be able to train a model on data drawn from some distribution $\mathcal{D}$, but then deploy it *without test time labels* on new data drawn from some other process, and still have guarantees about our predictions?

Of course this is impossible in general, but we can say something about it if we make assumptions about how the data distribution might shift. Suppose that we get training data from some source distribution $\mathcal{D}^s$, and then evaluate our model on a test distribution $\mathcal{D}^t$. Can we give guarantees if we assume something about how $\mathcal{D}^s$ and $\mathcal{D}^t$ relate to one another?

8.1 Likelihood Ratio Reweighting

Our goal is to learn to make predictions about labels y from examples x. If we are going to learn about the relationship between x and y on $\mathcal{D}^s$ and then

hope to do well on $\mathcal{D}^t$, then this relationship had better be similar on both distributions — in this chapter we will assume that it is the same.

Definition 41 *Two distributions $\mathcal{D}^s, \mathcal{D}^t \in \Delta\mathcal{Z}$ are said to have the same conditional label distributions if for every $x \in \mathcal{X}$, $\mathcal{D}^s_{\mathcal{Y}}(x) = \mathcal{D}^t_{\mathcal{Y}}(x)$. In other words the distributions differ only in their marginal distributions on features $\mathcal{D}^s_{\mathcal{X}}$ and $\mathcal{D}^t_{\mathcal{X}}$.*

So, two distributions that have the same conditional label distributions differ in the relative frequency with which different feature vectors x appear, but agree on how labels are distributed conditional on features — so there is some fixed "truth" that we can hope to learn.

Definition 42 (Likelihood Ratios) *For each $x \in \mathcal{X}$ let $p^s(x) = \Pr_{\mathcal{D}^s_{\mathcal{X}}}[x]$ and let $p^t(x) = \Pr_{\mathcal{D}^t_{\mathcal{X}}}[x]$ denote the probabilty mass/density that the feature distributions $\mathcal{D}^s_{\mathcal{X}}$ and $\mathcal{D}^t_{\mathcal{X}}$ respectively put on x. The $s \to t$ likelihood ratio for a point x is:*

$$w_{s\to t}(x) = \frac{p^t(x)}{p^s(x)}$$

Remark 8.1.1 *Observe that:*

$$\frac{1}{w_{s\to t}(x)} = \frac{p^s(x)}{p^t(x)} = w_{t\to s}(x)$$

$s \to t$ likelihood ratios are useful because they allow us to relate expectations taken over $\mathcal{D}^s$ to expectations taken over $\mathcal{D}^t$.

Lemma 8.1.1 *Suppose $\mathcal{D}^s$ and $\mathcal{D}^t$ have the same conditional label distribution $\mathcal{D}^s_{\mathcal{Y}}(x) = \mathcal{D}^t_{\mathcal{Y}}(x) = \mathcal{D}_{\mathcal{Y}}(x)$. Fix any $S \subseteq \mathcal{X}$. For any function $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, we have:*

$$\Pr_{(x,y)\sim\mathcal{D}^s}[x \in S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}^s}[w_{s\to t}(x)\cdot F(x,y)|x \in S] = \Pr_{(x,y)\sim\mathcal{D}^t}[x \in S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}^t}[F(x,y)|x \in S]$$

Proof 64 *For simplicity assume the distribution over $\mathcal{X}$ is discrete (otherwise repeat the derivation below with sums replaced by integrals). We have:*

$$\begin{aligned}
& \Pr_{(x,y)\sim\mathcal{D}^s}[x \in S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}^s}[w_{s\to t}(x) \cdot F(x,y)|x \in S] \\
= & \sum_{x\in S} p^s(x) \cdot w_{s\to t}(x) \cdot \mathop{\mathbb{E}}_{y\sim\mathcal{D}_{\mathcal{Y}}(x)}[F(x,y)] \\
= & \sum_{x\in S} p^s(x)\frac{p^t(x)}{p^s(x)} \cdot \mathop{\mathbb{E}}_{y\sim\mathcal{D}_{\mathcal{Y}}(x)}[F(x,y)] \\
= & \sum_{x\in S} p^t(x) \cdot \mathop{\mathbb{E}}_{y\sim\mathcal{D}_{\mathcal{Y}}(x)}[F(x,y)] \\
= & \Pr_{(x,y)\sim\mathcal{D}^t}[x \in S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}^t}[F(x,y)|x \in S]
\end{aligned}$$

Of course, even if we are explicitly given samples from $\mathcal{D}^s$ and $\mathcal{D}^t$, we will not generally know the likelihood ratios $w_{s\to t}(x)$. A common approach is to attempt to learn a function h from some class $\mathcal{H}$ that approximates them well. Since they are a function only of x, this can be done using only unlabelled examples from $\mathcal{D}^t_{\mathcal{X}}$. Suppose we attempt to approximate $w_{s\to t}(x)$ using a function h. How should we evaluate our approximation error?

Definition 43 *Suppose $\mathcal{D}^s$ and $\mathcal{D}^t$ have the same conditional label distribution. For a function $h : \mathcal{X} \to \mathbb{R}$, we write:*

$$e(h, w_{s\to t}) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}^s}[|h(x) - w_{s\to t}(x)|]$$

Similarly, for any subset $S \subseteq \mathcal{X}$ of the feature space, we write:

$$e(h, w_{s\to t}, S) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}^s}[|h(x) - w_{s\to t}(x)| \mid x \in S]$$

Remark 8.1.2 *Observe that by the law of total probability, for any collection of sets $\{S_1, \ldots, S_k\}$ that partition $\mathcal{X}$, we have that:*

$$\sum_{i=1}^{k} \Pr_{(x,y)\sim\mathcal{D}^s}[x \in S_i] e(h, w_{s\to t}, S_i) = e(h, w_{s\to t})$$

The next lemma shows that if we can estimate $w_{s\to t}$ closely in total variation distance (as measured in expectation over the source distribution $\mathcal{D}^s$), then we can closely approximate expectations over $\mathcal{D}^t$.

Lemma 8.1.2 *Suppose $\mathcal{D}^s$ and $\mathcal{D}^t$ have the same conditional label distribution. Fix any $S \subseteq \mathcal{X}$. For any function $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, and any function $h : \mathcal{X} \to \mathbb{R}$, we have:*

$$\left| \Pr_{(x,y)\sim\mathcal{D}^s}[x \in S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}^s}[h(x) \cdot F(x,y) | x \in S] - \Pr_{(x,y)\sim\mathcal{D}^t}[x \in S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}^t}[F(x,y)|x \in S] \right|$$
$$\leq \Pr_{(x,y)\sim\mathcal{D}^s}[x \in S] \cdot \max_{(x,y)\in\mathcal{Z}} |F(x,y)| \cdot e(h, w_{s\to t}, S)$$

Proof 65 *we know from Lemma 8.1.1 that we can write:*

$$\Pr_{(x,y)\sim\mathcal{D}^t}[x \in S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}^t}[F(x,y)|x \in S] = \Pr_{(x,y)\sim\mathcal{D}^s}[x \in S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}^s}[w_{s\to t}(x) \cdot F(x,y)|x \in S]$$

So, we can calculate:

$$\begin{aligned}
& \left| \Pr_{(x,y)\sim\mathcal{D}^s}[x \in S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}^s}[h(x) \cdot F(x,y)|x \in S] - \Pr_{(x,y)\sim\mathcal{D}^t}[x \in S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}^t}[F(x,y)|x \in S] \right| \\
= & \left| \Pr_{(x,y)\sim\mathcal{D}^s}[x \in S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}^s}[h(x) \cdot F(x,y)|x \in S] - \Pr_{(x,y)\sim\mathcal{D}^s}[x \in S] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}^s}[w_{s\to t}(x)F(x,y)|x \in S] \right| \\
= & \Pr_{(x,y)\sim\mathcal{D}^s}[x \in S] \left| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}^s}[F(x,y) \cdot (h(x) - w_{s\to t}(x))|x \in S] \right| \\
\leq & \Pr_{(x,y)\sim\mathcal{D}^s}[x \in S] \cdot \max_{(x,y)\in\mathcal{Z}} |F(x,y)| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}^s}[|(h(x) - w_{s\to t}(x))| \mid x \in S] \\
= & \Pr_{(x,y)\sim\mathcal{D}^s}[x \in S] \cdot \max_{(x,y)\in\mathcal{Z}} |F(x,y)| \cdot e(h, w_{s\to t}, S)
\end{aligned}$$

8.2 Multicalibration under Distribution Shift

We'll now study how multi-calibration guarantees change under distribution shift, and how the relationship between the class of functions $\mathcal{H}$ we are multicalibrated with respect to interacts with the likelihood ratios $w_{s\to t}(x)$ defined by the shift. It will be more convenient for us to work with an ℓ_1 notion of multicalibration (compared to the ℓ_2 notion we gave in Definition 35.

Definition 44 (L_1 Multicalibration For Real Valued Functions) *Fix a distribution $\mathcal{D} \in \Delta\mathcal{Z}$ and a model $f : \mathcal{X} \to [0,1]$. Let $\mathcal{H}$ be an arbitrary collection of real valued functions $h : \mathcal{X} \to \mathbb{R}$. We say that f is α-approximately L_1-multicalibrated with respect to $\mathcal{D}$ and $\mathcal{H}$ if for every $h \in \mathcal{H}$:*

$$K_1(f,h,\mathcal{D}) = \sum_{v \in R(f)} \Pr_{(x,y)\sim\mathcal{D}}[f(x)=v] \left| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)(y-v)|f(x)=v] \right| \leq \alpha$$

We say that f is α-approximately L_1-multicalibrated with respect to $\mathcal{D}$ if:

$$K_1(f,\mathcal{D}) = \sum_{v \in R(f)} \Pr_{(x,y)\sim\mathcal{D}}[f(x)=v] \left| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y-v)|f(x)=v] \right| \leq \alpha$$

Recall that we know from Lemma 3.1.1 that $K_1(f,h,\mathcal{D}) \leq \sqrt{K_2(f,h,\mathcal{D})}$. Thus, we can use algorithm 19 — which guarantees that $K_2(f,h,\mathcal{D}) \leq \alpha'$ for all $h \in \mathcal{H}$ — to obtain α-approximate L_1 multicalibration by setting $\alpha' = \alpha^2$.

Theorem 39 *Suppose $\mathcal{D}^s$ and $\mathcal{D}^t$ have the same conditional label distribution, and suppose f is α-approximately L_1-multicalibrated with respect to $\mathcal{D}^s$ and $\mathcal{H}$. Then f is also α-approximately L_1-multicalibrated with respect to $\mathcal{D}^t$ and $\mathcal{H}_{s\to t}$ where:*

$$\mathcal{H}_{s\to t} = \left\{ \frac{h(x)}{w_{s\to t}(x)} : h(x) \in \mathcal{H} \right\}$$

Proof 66 *Since f is α-approximately L_1-multicalibrated with respect to $\mathcal{D}^s$*

and $\mathcal{H}$, we have that for every $h \in \mathcal{H}$:

$$\begin{aligned}
\alpha &\geq K_1(f, h, \mathcal{D}^s) \\
&= \sum_{v \in R(f)} \Pr_{(x,y)\sim\mathcal{D}^s}[f(x) = v] \left| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}^s}[h(x)(y-v)|f(x) = v] \right| \\
&= \sum_{v \in R(f)} \left| \Pr_{(x,y)\sim\mathcal{D}^s}[f(x) = v] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}^s}[h(x)(y-v)|f(x) = v] \right| \\
&= \sum_{v \in R(f)} \left| \Pr_{(x,y)\sim\mathcal{D}^s}[f(x) = v] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}^s}\left[w_{s\to t}(x) \cdot \left(\frac{h(x)}{w_{s\to t}(x)}(y-v) \right) |f(x) = v \right] \right| \\
&= \sum_{v \in R(f)} \left| \Pr_{(x,y)\sim\mathcal{D}^t}[f(x) = v] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}^t}\left[\frac{h(x)}{w_{s\to t}(x)}(y-v)|f(x) = v \right] \right| \\
&= K_1\left(f, \frac{h}{w_{s\to t}}, \mathcal{D}^t \right)
\end{aligned}$$

Here the second to last equality follows from applying Lemma 8.1.1 to each term:

$$\Pr_{(x,y)\sim\mathcal{D}^s}[f(x) = v] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}^s}\left[w_{s\to t}(x) \cdot \left(\frac{h(x)}{w_{s\to t}(x)}(y-v) \right) |f(x) = v \right]$$

using $S = \{x : f(x) = v\}$ and $F(x, y) = \frac{h(x)}{w_{s\to t}(x)}(y - v)$.

Corollary 8.2.1 *Suppose $\mathcal{D}^s$ and $\mathcal{D}^t$ have the same conditional label distribution, and suppose f is α-approximately L_1-multicalibrated with respect to $\mathcal{D}^s$ and $\mathcal{H}$. Then if $w_{s\to t} \in \mathcal{H}$, f has at most α L_1-calibration error on $\mathcal{D}^t$:*

$$K_1(f, \mathcal{D}^t) \leq \alpha$$

Proof 67 *We apply Theorem 39. Since by assumption $w_{s\to t}(x) \in \mathcal{H}$, we can choose $h = w_{s\to t}$ and find that:*

$$\begin{aligned}
\alpha &\geq K_1\left(f, \frac{h}{w_{s\to t}}, \mathcal{D}^t \right) \\
&= \sum_{v \in R(f)} \left| \Pr_{(x,y)\sim\mathcal{D}^t}[f(x) = v] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}^t}\left[\frac{h(x)}{w_{s\to t}(x)}(y-v)|f(x) = v \right] \right| \\
&= \sum_{v \in R(f)} \Pr_{(x,y)\sim\mathcal{D}^t}[f(x) = v] \left| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}^t}[(y-v)|f(x) = v] \right| \\
&= K_1(f, \mathcal{D}^t)
\end{aligned}$$

Similarly, if we are approximately multicalibrated on $\mathcal{D}^s$ with respect to a class $\mathcal{H}$ that contains a function h that is *close* in total variation distance to $w_{s\to t}$ on $\mathcal{D}^s$, then we remain approximately calibrated on $\mathcal{D}^t$.

Lemma 8.2.1 *Suppose $\mathcal{D}^s$ and $\mathcal{D}^t$ have the same conditional label distribution, and suppose $f : \mathcal{X} \to [0,1]$ is α-approximately L_1-multicalibrated with respect to $\mathcal{D}^s$ and $\mathcal{H}$. Then f has L_1- calibration error on $\mathcal{D}^t$ at most:*

$$K_1(f, \mathcal{D}^t) \leq \alpha + \min_{h \in \mathcal{H}} e(h, w_{s \to t})$$

Proof 68 *Let $h^* = \arg\min_{h \in \mathcal{H}} e(h, w_{s \to t})$. Since f is α-approximately L_1 multicalibrated with respect to D^s and H we have:*

$$\begin{aligned}
\alpha &\geq K_1(f, h^*, D^s) \\
&= \sum_{v \in R(f)} \left| \Pr_{(x,y) \sim \mathcal{D}^s}[f(x) = v] \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}^s}[h^*(x)(y - v) | f(x) = v] \right| \\
&\geq \sum_{v \in R(f)} \left| \Pr_{(x,y) \sim \mathcal{D}^t}[f(x) = v] \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}^t}[(y - v) | f(x) = v] \right| \\
&\quad - \sum_{v \in R(f)} \Pr_{(x,y) \sim \mathcal{D}^s}[f(x) = v] \cdot e(h^*, w_{s \to t}, \{x : f(x) = v\}) \\
&= K_1(f, \mathcal{D}^t) - e(h^*, w_{s \to t})
\end{aligned}$$

where the inequality follows from Lemma 8.1.2 applied to each term:

$$\Pr_{(x,y) \sim \mathcal{D}^t}[f(x) = v] \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}^t}[(y - v) | f(x) = v]$$

choosing $S_v = \{x : f(x) = v\}$, $F(x, y) = (y - v)$, and the fact that since $y, v \in [0,1]$, $\max_y |y - v| \leq 1$. The final line follows from the observation that the collection $\{S_v\}_{v \in R(f)}$ forms a partition of $\mathcal{X}$.

8.3 Why Calibration Under Distribution Shift is Useful

On our training distribution we generally have samples of labeled data $(x, y) \sim \mathcal{D}^s$, and so we can empirically evaluate various quantities of interest. When it comes time to deploy a model, we may have unlabelled examples $x \sim \mathcal{D}^t_{\mathcal{X}}$ from the target distribution, but we may not have labelled examples. But if f is calibrated on $\mathcal{D}^t$, then there are certain things we can do with only unlabelled examples.

One very simple thing we can do is estimate the average value of the label.

Lemma 8.3.1 *Suppose f satisfies α-approximate L_1 calibration on $\mathcal{D}^t$: $K_1(f, \mathcal{D}^t) \leq \alpha$. Then:*

$$\left| \mathop{\mathbb{E}}_{x \sim \mathcal{D}^t_{\mathcal{X}}}[f(x)] - \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}^t}[y] \right| \leq \alpha$$

Proof 69 *Expanding the definition of K_1 we can write:*

$$\begin{aligned}
\alpha &\geq K_1(f, \mathcal{D}^t) \\
&= \sum_{v \in R(f)} \Pr_{x \sim \mathcal{D}^t_{\mathcal{X}}}[f(x) = v] \left| \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}^t}[(y - v) | f(x) = v] \right| \\
&= \sum_{v \in R(f)} \left| \Pr_{x \sim \mathcal{D}^t_{\mathcal{X}}}[f(x) = v] \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}^t}[y | f(x) = v] - \Pr_{x \sim \mathcal{D}^t_{\mathcal{X}}}[f(x) = v] v \right| \\
&\geq \left| \sum_{v \in R(f)} \left(\Pr_{x \sim \mathcal{D}^t_{\mathcal{X}}}[f(x) = v] \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}^t}[y | f(x) = v] - \Pr_{x \sim \mathcal{D}^t_{\mathcal{X}}}[f(x) = v] v \right) \right| \\
&= \left| \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}^t}[y] - \mathop{\mathbb{E}}_{x \sim \mathcal{D}^t_{\mathcal{X}}}[f(x)] \right|
\end{aligned}$$

If our label space is binary $\mathcal{Y} = \{0, 1\}$, then we can go beyond this, and estimate the cost of *acting on any policy depending on the predictions of f.*

Definition 45 *Fix an action space $\mathcal{A}$ and a model $f : \mathcal{X} \to [0, 1]$. A* policy *of f is any mapping $\rho : [0, 1] \to \mathcal{A}$ that chooses an action $\rho(f(x)) \in \mathcal{A}$ as a function of the prediction $f(x)$.*

We can evaluate the *cost* of a policy using a loss function:

Definition 46 *Fixing an action space $\mathcal{A}$, a loss function $\ell : \mathcal{A} \times \{0, 1\} \to \mathbb{R}$ maps action/label pairs to a real valued loss. Given a distribution $\mathcal{D}$ and a predictor $f : \mathcal{X} \to [0, 1]$, the expected cost of a policy ρ is:*

$$\ell(\rho, f, \mathcal{D}) = \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}}[\ell(\rho(f(x)), y)]$$

We can estimate the cost of any policy ρ if we have sample access to $\mathcal{D}$ — but this requires samples both of x (to compute $\rho(f(x))$) and y (to plug into the second argument of $\ell(\cdot, \cdot)$). What if we only have sample access to unlabelled examples from $\mathcal{D}_{\mathcal{X}}$? Recall that $f(x)$ *purports* to estimate $\mathbb{E}_{\mathcal{D}_{\mathcal{Y}}(x)}[y]$ — i.e. the probability that $y = 1$ conditional on x. So we can attempt to estimate $\ell(\rho, f, \mathcal{D})$ taking this as a given, using only unlabelled examples:

Definition 47 *Given an action space $\mathcal{A}$, a loss function $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$, a policy ρ, a predictor $f : \mathcal{X} \to \mathbb{R}$, and a feature distribution $\mathcal{D}_{\mathcal{X}}$, the f-*estimated *cost of ρ is:*

$$\tilde{\ell}(\rho, f, \mathcal{D}_{\mathcal{X}}) = \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}}[f(x)\ell(\rho(f(x)), 1) + (1 - f(x))\ell(\rho(f(x)), 0)]$$

Observe that for the Bayes optimal predictor — $f^*(x) = \mathbb{E}_{\mathcal{D}_{\mathcal{Y}}(x)}[y]$ — that the f^*-estimated cost of ρ: $\tilde{\ell}(\rho, f^*, \mathcal{D}_{\mathcal{X}})$ is equal to its true expected cost: $\ell(\rho, f^*, \mathcal{D})$.

We now show that the same is true if f is not Bayes optimal, but merely calibrated.

Theorem 40 *Fix an action space $\mathcal{A}$, a loss function $\ell : \mathcal{A} \times \{0,1\} \to \mathbb{R}$, a policy ρ, a distribution $\mathcal{D}$, and a predictor $f : \mathcal{X} \to \mathbb{R}$, and a distribution $\mathcal{D}$. Let:*

$$C = \max_{a \in \mathcal{A}} \left(\ell(a,0) + \ell(a,1) \right)$$

If f is α-approximately L_1-calibrated with respect to $\mathcal{D}$, then:

$$\left| \ell(\rho, f, \mathcal{D}) - \tilde{\ell}(\rho, f, \mathcal{D}_{\mathcal{X}}) \right| \leq C \cdot \alpha$$

Proof 70 *Let:*

$$k_v := \Pr_{x \sim \mathcal{D}_{\mathcal{X}}}[f(x) = v] \cdot \left| \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}}[y | f(x) = v] - v \right|$$

Since f satisfies α-approximate L_1 calibration with respect to $\mathcal{D}$, we know that:

$$K_1(f, \mathcal{D}) = \sum_{v \in R(f)} k_v \leq \alpha$$

We can now calculate:

$$\begin{aligned}
& \ell(\rho, f, \mathcal{D}) \\
= & \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}}[\ell(\rho(f(x)), y)] \\
= & \sum_{v \in R(f)} \Pr_{x \sim \mathcal{D}_{\mathcal{X}}}[f(x) = v] \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}}[\ell(\rho(f(x)), y) | f(x) = v] \\
= & \sum_{v \in R(f)} \Pr_{x \sim \mathcal{D}_{\mathcal{X}}}[f(x) = v] \left(\ell(\rho(v), 1) \Pr_{(x,y) \sim \mathcal{D}}[y = 1 | f(x) = v] + \ell(\rho(v), 0) \Pr_{(x,y) \sim \mathcal{D}}[y = 0 | f(x) = v] \right) \\
= & \sum_{a \in \mathcal{A}} \ell(a,1) \sum_{v : \rho(v) = a} \Pr_{x \sim \mathcal{D}_{\mathcal{X}}}[f(x) = v] \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}}[y | f(x) = v] + \\
& \sum_{a \in \mathcal{A}} \ell(a,0) \sum_{v : \rho(v) = a} \Pr_{x \sim \mathcal{D}_{\mathcal{X}}}[f(x) = v](1 - \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}}[y | f(x) = v]) \\
\leq & \sum_{a \in \mathcal{A}} \ell(a,1) \sum_{v : \rho(v) = a} \left(\Pr_{x \sim \mathcal{D}_{\mathcal{X}}}[f(x) = v] v + k_v \right) + \\
& \sum_{a \in \mathcal{A}} \ell(a,0) \sum_{v : \rho(v) = a} \left(\Pr_{x \sim \mathcal{D}_{\mathcal{X}}}[f(x) = v](1 - v) + k_v \right) \\
= & \sum_{v \in R(f)} \Pr_{x \sim \mathcal{D}_{\mathcal{X}}}[f(x) = v] \left(v \ell(\rho(v), 1) + (1 - v) \ell(\rho(v), 0) \right) + \sum_{v \in R(f)} k_v (\ell(\rho(v), 1) + \ell(\rho(v), 0)) \\
= & \tilde{\ell}(\rho, f, \mathcal{D}_{\mathcal{X}}) + \sum_{v \in R(f)} k_v (\ell(\rho(v), 1) + \ell(\rho(v), 0)) \\
\leq & \tilde{\ell}(\rho, f, \mathcal{D}_{\mathcal{X}}) + C \cdot \alpha
\end{aligned}$$

The other direction is identical.

A simple example of a policy and loss function arises in binary classification. Here, $\mathcal{Y} = \{0, 1\}$ and $\mathcal{A} = \{0, 1\}$: our goal is for each example x observed, to predict the true label by selecting $a \in \mathcal{A}$ such that $a = y$. The 0/1 loss function defined as $\ell^{0/1}(a, y) = \mathbb{1}[a \neq y]$ measures the frequency with which a given policy makes prediction mistakes.

For any calibrated predictor f (including the true conditional label expectation $f = f^* = \mathbb{E}_{y \sim \mathcal{D}_{\mathcal{Y}}(x)}[y]$), the following policy minimizes 0/1 loss among all policies defined as a function of f:

$$\rho^*(v) = \begin{cases} 1 & v \geq \frac{1}{2} \\ 0 & v < \frac{1}{2} \end{cases}$$

Lemma 8.3.2 *Fix any distribution $\mathcal{D}$ and any predictor $f : \mathcal{X} \to [0, 1]$ such that $K_1(f, \mathcal{D}) \leq \alpha$. Consider the policy ρ^* defined above. For any other policy $\rho : [0, 1] \to \{0, 1\}$, we have:*

$$\ell^{0,1}(\rho^*, f, \mathcal{D}) \leq \ell^{0,1}(\rho, f, \mathcal{D}) + 2\alpha$$

Proof 71 *Using Theorem 40, the fact that f satisfies α-approximate L_1-calibration, and the fact that for all y $\ell^{0/1}(0, y) + \ell^{0/1}(1, y) = 1$, we can calculate:*

$$\begin{aligned}
\ell^{0,1}(\rho, f, \mathcal{D}) &\geq \tilde{\ell}^{0,1}(\rho, f, \mathcal{D}_{\mathcal{X}}) - \alpha \\
&= \sum_{v \in R(f)} \Pr_{x \sim \mathcal{D}_{\mathcal{X}}}[f(x) = v] \left(v \cdot \ell^{0,1}(\rho(v), 1) + (1 - v)\ell^{0,1}(\rho(v), 0)\right) - \alpha \\
&\geq \sum_{v \in R(f)} \Pr_{x \sim \mathcal{D}_{\mathcal{X}}}[f(x) = v] \left(v \cdot \ell^{0,1}(\rho^*(v), 1) + (1 - v)\ell^{0,1}(\rho^*(v), 0)\right) - \alpha \\
&= \tilde{\ell}^{0,1}(\rho^*, f, \mathcal{D}_{\mathcal{X}}) - \alpha \\
&\geq \ell^{0,1}(\rho^*, f, \mathcal{D}) - 2\alpha
\end{aligned}$$

Here the first and last inequalities follow from Theorem 40. The middle inequality follows from the fact that pointwise (for each value v):

$$v \cdot \ell^{0,1}(\rho(v), 1) + (1 - v)\ell^{0,1}(\rho(v), 0) = \begin{cases} 1 - v & \rho(v) = 1 \\ v & \rho(v) = 0 \end{cases}$$

is minimized by setting $\rho(v) = 1$ when $v \geq \frac{1}{2}$ and $\rho(v) = 0$ when $v < \frac{1}{2}$, which is what the policy $\rho^*(v)$ does.

So if f is calibrated on $\mathcal{D}$, the not only is ρ^* the optimal post-processing of f to minimize classification error on $\mathcal{D}$, but we can estimate the classification error of $\rho^*(f(x))$ on $\mathcal{D}$ without the need for any labelled examples from $\mathcal{D}$ — since we can estimate $\tilde{\ell}^{0,1}(\rho, f, \mathcal{D}_{\mathcal{X}})$ using only samples from $\mathcal{D}_{\mathcal{X}}$.

Thus if we have trained a model that is multicalibrated on some distribution $\mathcal{D}^s$ with respect to some class of functions $\mathcal{H}$, then for any distribution

$\mathcal{D}^t$ such that $w_{s\to t} \in \mathcal{H}$ (or is close in statistical distance to some $h \in \mathcal{H}$), we can correctly estimate the performance of our model on $\mathcal{D}^t$ given access only to unlabelled data from $\mathcal{D}^t$, which is much more commonly available.

References and Further Reading

Multicalibration under distribution shift was studied by Kim et al. [2022], from which this chapter primarily draws — they call the phenomenon "universal adaptibility". Baek et al. [2022] suggest in a different context that predictors that are calibrated out of distribution can be used to evaluate the out of distribution performance using only unlabelled data.

9

Sufficient Statistics for Optimization

CONTENTS

In Chapter 8 we started studying the problem of choosing *actions* $a \in \mathcal{A}$ to minimize the expectation of a loss function $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ using some policy $\rho : [0,1] \to \mathcal{A}$ that we evaluate as a function of a predictor $f : \mathcal{X} \to [0,1]$. In the case of binary labels $\mathcal{Y} = \{0,1\}$, we saw that if f is calibrated on $\mathcal{D}$, then for any such policy, we can accurately estimate the loss of a policy $\rho(f(x))$ using only unlabeled data:

$$\ell(\rho, f, \mathcal{D}) \approx \tilde{\ell}(\rho, f, \mathcal{D}_{\mathcal{X}})$$

by "pretending" that $\Pr[y|f(x) = v] = v$ — so we can choose the optimal policy $\rho^*(f(x))$ under this fiction, and know that it performs as well as any other policy that is a function only of $f(x)$.

How strong is this guarantee? It depends. If $f(x) = f^*(x) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y|x]$ is the true conditional label expectation, then this guarantee means that the optimal policy ρ^* that is a function of $f(x)$ is as good as any other policy, regardless of what information about x it uses. On the other hand, if $f(x) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y]$ is simply the (calibrated) constant function, then policies $\rho(f(x))$ must also be constant functions, and so have necessarily very weak performance guarantees. In this chapter we ask under what conditions on f we can compare the performance of policies $\rho(f(x))$ to the performance of policies $h(x)$ that can depend in other ways on the features x. We assume throughout this chapter that the label space is binary: $\mathcal{Y} = \{0,1\}$.

9.1 Omnipredictors: Sufficient Statistics for Unconstrained Optimization

We recall several important definitions from Chapter 8.

Definition 48 *Fix an action space $\mathcal{A}$ and a model $f : \mathcal{X} \to [0,1]$. A* policy *of f is any mapping $\rho : [0,1] \to \mathcal{A}$ that chooses an action $\rho(f(x)) \in \mathcal{A}$ as a function of the prediction $f(x)$.*

We can evaluate the *cost* of a policy using a loss function:

Definition 49 *Fixing an action space $\mathcal{A}$, a loss function $\ell : \mathcal{A} \times \{0,1\} \to \mathbb{R}$ maps action/label pairs to a real valued loss. Given a distribution $\mathcal{D}$ and a predictor $f : \mathcal{X} \to [0,1]$, the expected cost of a policy ρ is:*

$$\ell(\rho, f, \mathcal{D}) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(\rho(f(x)), y)]$$

To compute the loss ℓ of a policy we need access to labeled examples $(x, y) \sim \mathcal{D}$. But we can estimate the loss of a policy using only unlabelled examples together with a model f if we "pretend" that $f(x)$ actually encodes the true conditional label expectation $\mathbb{E}[y|x]$:

Definition 50 *Given an action space $\mathcal{A}$, a loss function $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$, a policy ρ, a predictor $f : \mathcal{X} \to \mathbb{R}$, and a feature distribution $\mathcal{D}_{\mathcal{X}}$, the f-*estimated *cost of ρ is:*

$$\tilde{\ell}(\rho, f, \mathcal{D}_{\mathcal{X}}) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}[f(x)\ell(\rho(f(x)), 1) + (1 - f(x))\ell(\rho(f(x)), 0)]$$

Another nice property of the f-estimated loss is that we can find the policy that optimizes it without needing to know anything about the underlying distribution. Specifically, if we have a loss function ℓ in mind, we can choose the policy ρ^*_ℓ that pointwise optimizes the f-estimated cost $\tilde{\ell}$:

$$\rho^*_\ell(v) = \arg\min_{a\in\mathcal{A}} (v\ell(a, 1) + (1 - v)\ell(a, 0))$$

If f is calibrated, then the policy ρ^*_ℓ has the smallest expected loss (as measured by ℓ) of any policy that is a function of f. This statement generalizes what we proved in Lemma 8.3.2 in the special case of 0/1 loss and has essentially the same proof.

Lemma 9.1.1 *Fix any distribution $\mathcal{D}$, loss function $\ell : \mathcal{A} \times \{0,1\} \to \mathbb{R}$, and any predictor $f : \mathcal{X} \to [0,1]$ such that $K_1(f, \mathcal{D}) \leq \alpha$. Let:*

$$C = \max_{a\in\mathcal{A}} (\ell(a, 0) + \ell(a, 1))$$

Consider the policy ρ^*_ℓ defined above. For any other policy $\rho : [0,1] \to \mathcal{A}$, we have:

$$\ell(\rho^*_\ell, f, \mathcal{D}) \leq \ell(\rho, f, \mathcal{D}) + 2C\alpha$$

Proof 72 *Using Theorem 40 and the fact that f satisfies α-approximate L_1-calibration, we can calculate:*

$$\begin{aligned}
\ell(\rho, f, \mathcal{D}) &\geq \tilde{\ell}(\rho, f, \mathcal{D}_{\mathcal{X}}) - C\alpha \\
&= \sum_{v \in R(f)} \Pr_{x \sim \mathcal{D}_{\mathcal{X}}}[f(x) = v]\left(v \cdot \ell(\rho(v), 1) + (1-v)\ell(\rho(v), 0)\right) - C\alpha \\
&\geq \sum_{v \in R(f)} \Pr_{x \sim \mathcal{D}_{\mathcal{X}}}[f(x) = v]\left(v \cdot \ell(\rho^*_\ell(v), 1) + (1-v)\ell(\rho^*_\ell(v), 0)\right) - C\alpha \\
&= \tilde{\ell}(\rho^*_\ell, f, \mathcal{D}_{\mathcal{X}}) - C\alpha \\
&\geq \ell(\rho^*_\ell, f, \mathcal{D}) - 2C\alpha
\end{aligned}$$

Here the first and last inequalities follow from Theorem 40. The middle inequality follows from the fact that by definition, $\rho^*_\ell(v)$ is the minimizer of:

$$v \cdot \ell(\rho(v), 1) + (1-v)\ell(\rho(v), 0)$$

Going forward, we consider the special case of $\mathcal{A} = [0, 1]$ and aim to show that if f is multicalibrated with respect to a class of real valued functions $\mathcal{H}$, then for any convex loss function ℓ, the policy ρ^*_ℓ has optimal loss not just compared to other policies ρ of f, but compared to any $h \in \mathcal{H}$. Note that functions $h : \mathcal{X} \to [0, 1]$ are functions of x directly, rather than functions of $f(x)$, and so Lemma 9.1.1 does not imply that that $\rho^*_\ell(f(x))$ has lower loss than $h(x)$. But Lemma 9.1.1 does point in the direction of our proof strategy: we will show that if f is (approximately) multicalibrated with respect to $\mathcal{H}$ then in fact every $h \in \mathcal{H}$ is (almost) dominated by a policy $\rho(f(x))$. Thus for any loss functions satisfying the conditions of our theorem, we can do (almost) as well as any $h \in \mathcal{H}$ by playing the *optimal* policy for the f-estimated loss ρ^*_ℓ. As mentioned, our results will apply to any *convex* loss function:

Definition 51 *A loss function $\ell : [0, 1] \times \{0, 1\} \to \mathbb{R}$ is* convex *in its first argument if for all $v, v', \alpha \in [0, 1]$ and for all $y \in \{0, 1\}$:*

$$\ell(\alpha v + (1-\alpha)v', y) \leq \alpha \cdot \ell(v, y) + (1-\alpha) \cdot \ell(v', y)$$

A direct consequence of convexity that we will make use of is called Jensen's inequality:

Claim 9.1.1 (Jensen's Inequality) *Fix any loss function $\ell : [0, 1] \times \{0, 1\} \to \mathbb{R}$ that is convex in its first argument. For any $y \in \{0, 1\}$ and for any distribution $\mathcal{P} \in \Delta[0, 1]$, we have:*

$$\mathop{\mathbb{E}}_{v \sim \mathcal{P}}[\ell(v, y)] \geq \ell\left(\mathop{\mathbb{E}}_{v \sim \mathcal{P}}[v], y\right)$$

How closely we can relate the performance of a model $h(x)$ to the performance of a policy $\rho(f(x))$ will depend both on the multicalibration error that f has on the models in $\mathcal{H}$ and on how much small errors in prediction are magnified by the loss function ℓ, which we will measure by its Lipschitz constant:

Definition 52 *A loss function $\ell : [0,1] \times \{0,1\} \to \mathbb{R}$ is* L-Lipschitz *in its first argument if for all v, v' and for all $y \in \{0,1\}$:*

$$|\ell(v,y) - \ell(v',y)| \leq L \cdot |v - v'|$$

Finally we will prove a useful statement about multicalibration—if f is multicalibrated with respect to $\mathcal{H}$, then for any $h \in \mathcal{H}$, its conditional expectation (conditional on the value of f) doesn't change by very much if we additionally condition on the value of the label:

Lemma 9.1.2 *Fix any distribution $\mathcal{D}$ and class of real valued functions $\mathcal{H}$. Suppose that f is α-approximately L_1 multicalibrated with respect to $\mathcal{D}$ and $\mathcal{H}$ (as defined in Definition 44). and α-approximately L_1-calibrated. Then for any $h \in \mathcal{H}$ and $v \in R(f)$:*

$$\sum_{v \in R(f)} \Pr[f(x) = v] v(1-v) \left| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x) | f(x) = v, y = 1] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x) | f(x) = v, y = 0] \right|$$

$$\leq 2\alpha$$

Proof 73 *Let:*

$$k_v = \Pr_{(x,y)\sim\mathcal{D}}[f(x) = v] \left| v - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y | f(x) = v] \right|$$

By hypothesis, we know that:

$$\alpha \geq K_1(f, \mathcal{D}) = \sum_{v \in R(f)} k_v.$$

We can now compute:

$$\begin{aligned}
\alpha &\geq K_1(f,h,\mathcal{D}) \\
&= \sum_{v\in R(f)} \Pr_{(x,y)\sim\mathcal{D}}[f(x)=v]\left|\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)(y-v)|f(x)=v]\right| \\
&= \sum_{v\in R(f)} \Pr_{(x,y)\sim\mathcal{D}}[f(x)=v]\Big|\Pr[y=1|f(x)=v]\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)|f(x)=v,y=1](1-v) \\
&\quad -\Pr[y=0|f(x)=v]\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)|f(x)=v,y=0]v\Big| \\
&\geq \sum_{v\in R(f)} \Pr_{(x,y)\sim\mathcal{D}}[f(x)=v]\Big|v\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)|f(x)=v,y=1](1-v) \\
&\quad -(1-v)\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)|f(x)=v,y=0]v\Big| - \sum_{v\in R(f)} k_v((1-v)+v) \\
&= \sum_{v\in R(f)} \Pr_{(x,y)\sim\mathcal{D}}[f(x)=v]\Big|v(1-v)\Big(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)|f(x)=v,y=1] \\
&\quad -\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)|f(x)=v,y=0]\Big)\Big| - \sum_{v\in R(f)} k_v \\
&\geq \sum_{v\in R(f)} \Pr_{(x,y)\sim\mathcal{D}}[f(x)=v]v(1-v)\left|\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)|f(x)=v,y=1]\right. \\
&\quad \left.-\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)|f(x)=v,y=0]\right| - \alpha
\end{aligned}$$

With this lemma in hand, we can prove the main theorem of this section: that if f is multicalibrated with respect to $\mathcal{H}$, then for any convex Lipschitz loss function ℓ, the policy ρ^*_ℓ obtains loss nearly as good as the loss of the best $h \in \mathcal{H}$. Thus, once we train f, we can use it to optimize *any* such loss function ℓ and have performance guarantees relative to $\mathcal{H}$, rather than needing to solve a fresh optimization problem for each new loss function. The proof strategy is just as we have already laid out: show that the loss for any $h \in \mathcal{H}$ is comparable to the loss of some *policy* ρ of f, and therefore only higher than the loss of the *best* policy ρ^*_ℓ for ℓ.

Theorem 41 *Let $\ell : [0,1] \times \{0,1\} \to [0,1]$ be a bounded loss function that is both convex and L-Lipschitz in its first argument. Suppose that f is α-approximately L_1 calibrated with respect to a distribution $\mathcal{D}$, and also with respect to $\mathcal{D}$ and a class of real valued functions $\mathcal{H}$. That is, both that $K_1(f,\mathcal{D}) \leq \alpha$ and that for every $h \in \mathcal{H}$, $K_1(f,h,\mathcal{D}) \leq \alpha$. Let ρ^*_ℓ be the policy*

that optimizes the f-estimated loss $\tilde{\ell}$:

$$\rho^*_\ell(v) = \arg\min_{a\in\mathcal{A}} (v\ell(a,1) + (1-v)\ell(a,0))$$

Then the loss of policy ρ^*_ℓ is almost as low as the loss of any $h \in \mathcal{H}$:

$$\ell(\rho^*_\ell, f, \mathcal{D}) \leq \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(h(x),y)] + (4+4L)\alpha$$

Proof 74 *Let:*

$$k_v = \Pr_{(x,y)\sim\mathcal{D}}[f(x)=v]\left|v - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y|f(x)=v]\right|$$

By hypothesis, we know that:

$$\alpha \geq K_1(f,\mathcal{D}) = \sum_{v\in R(f)} k_v.$$

Let $H(v,1) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[h(x)|f(x)=v, y=1]$ and $H(v,0) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[h(x)|f(x)=v, y=0]$. Using Jensen's inequality and the convexity of ℓ, we can calculate:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(h(x),y)] \\
=\ &\sum_{v\in R(f)} \Pr[f(x)=v]\Big(\Pr[y=1|f(x)=v] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(h(x),1)|f(x)=v,y=1] \\
&+(\Pr[y=0|f(x)=v] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(h(x),0)|f(x)=v,y=0]\Big) \\
\geq\ &\sum_{v\in R(f)} \Pr[f(x)=v]\Big(\Pr[y=1|f(x)=v]\ell(H(v,1),1) + (\Pr[y=0|f(x)=v]\ell(H(v,0),0)\Big) \\
\geq\ &\sum_{v\in R(f)} \Pr[f(x)=v]\Big(v\ell(H(v,1),1) + (1-v)\ell(H(v,0),0)\Big) - 2\sum_{v\in R(f)} k_v \\
\geq\ &\sum_{v\in R(f)} \Pr[f(x)=v]\Big(v\ell(H(v,1),1) + (1-v)\ell(H(v,0),0)\Big) - 2\alpha
\end{aligned}$$

Now consider the policy ρ defined such that:

$$\rho(v) = \begin{cases} H(v,1) & v \geq \frac{1}{2} \\ H(v,0) & v < \frac{1}{2} \end{cases}$$

Lets compare the loss of this policy ρ with the loss of h. Continuing our derivation above we find:

$$
\begin{aligned}
& \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(h(x),y)] \\
\geq\ & \sum_{v\in R(f)} \Pr[f(x)=v]\Big(v\ell(H(v,1),1)+(1-v)\ell(H(v,0),0)\Big) - 2\alpha \\
\geq\ & \sum_{v<\frac{1}{2}} \Pr[f(x)=v]\Big(v\ell(H(v,0),1)+(1-v)\ell(H(v,0),0) - vL|H(v,0)-H(v,1)|\Big) + \\
& \sum_{v\geq\frac{1}{2}} \Pr[f(x)=v]\Big(v\ell(H(v,1),1)+(1-v)\ell(H(v,1),0) - (1-v)L|H(v,0)-H(v,1)|\Big) - 2\alpha \\
=\ & \tilde{\ell}(\rho,f,\mathcal{D}_{\mathcal{X}}) - 2\alpha - L\sum_{v\in R(f)} \Pr[f(x)=v]\min(v,1-v)|H(v,0)-H(v,1)| \\
\geq\ & \tilde{\ell}(\rho,f,\mathcal{D}_{\mathcal{X}}) - 2\alpha - 2L\sum_{v\in R(f)} \Pr[f(x)=v]v\cdot(1-v)|H(v,0)-H(v,1)| \\
\geq\ & \tilde{\ell}(\rho,f,\mathcal{D}_{\mathcal{X}}) - (2+4L)\alpha \\
\geq\ & \tilde{\ell}(\rho^*_\ell,f,\mathcal{D}_{\mathcal{X}}) - (2+4L)\alpha \\
\geq\ & \ell(\rho^*_\ell,f,\mathcal{D}) - (4+4L)\alpha
\end{aligned}
$$

Here, in the 3rd to last line, we have applied Lemma 9.1.2, which tells us that:

$$\sum_{v\in R(f)} \Pr[f(x)=v]v(1-v)\,|H(v,0)-H(v,1)| \leq 2\alpha$$

In the second to last line we have used the fact that ρ^*_ℓ is the minimizer of $\tilde{\ell}(\rho,f,\mathcal{D}_{\mathcal{X}})$ among all policies ρ. In the final line we have applied Theorem 40 to relate the f-estimated loss $\tilde{\ell}$ to the true loss ℓ, using the fact that $K_1(f,\mathcal{D}) \leq \alpha$, and that $C = \max_{a\in\mathcal{A}}(\ell(a,0)+\ell(a,1))$ is at most 2 since we have assumed that ℓ takes values in $[0,1]$.

9.2 Sufficient Statistics for Constrained Optimization

In Section 9.1 we showed that if f is multicalibrated with respect to $\mathcal{H}$, then for any (convex, Lipschitz) loss function ℓ, using the policy of f, ρ^*_ℓ that is optimal for minimizing ℓ *is almost as good as using the best* $h \in \mathcal{H}$, in terms of minimizing ℓ over the true data distribution. Note that this was an *unconstrained* optimization problem, in that there were no restrictions at all on what our policy $\rho^*(\ell)$ could look like. In this section, we consider constrained optimization problems. We continue to consider action spaces $\mathcal{A} = [0,1]$ and binary label spaces $\mathcal{Y} = \{0,1\}$.

Definition 53 *Fix a collection of real valued functions $\mathcal{H}$ of the form $h : \mathcal{X} \to [0,1]$, a collection of group indicator functions $\mathcal{G}$ of the form $g : \mathcal{X} \to \{0,1\}$, and a scalar $C \in R$. An $(\mathcal{H}, \mathcal{G}, C)$-convex minimization problem with linear constraints is defined by:*

1. *An objective function $\ell : [0,1] \times \{0,1\} \to [-C, C]$ that is convex in its first argument,*
2. *A collection of k constraints j each defined by a loss function $\ell_j : [0,1] \times \{0,1\} \to [-C, C]$ that is affine in its first argument, a group indicator function $g_j \in \mathcal{G}$, and a subset of labels $S_j \subseteq \{0,1\}$.*

Together they define the following optimization problem:

$$\arg\min_{\mathcal{P} \in \Delta\mathcal{H}} \mathop{\mathbb{E}}_{h \sim \mathcal{P}, (x,y) \sim \mathcal{D}} [\ell(h(x), y)]$$

Subject to the constraint that for each $j \in [k]$:

$$\mathop{\mathbb{E}}_{h \sim \mathcal{P}, (x,y) \sim \mathcal{D}} [\ell_j(h(x), y) | g_j(x) = 1, y \in S_j] \leq 0$$

If there is any solution $\mathcal{P}$ that satisfies all of the constraints, we say that the optimization problem is feasible. *We write $\mathcal{P}^*$ for the solution that minimizes the objective function while satisfying the constraints, and write $OPT(\mathcal{H}) = \mathbb{E}_{h \sim \mathcal{P}^*, (x,y) \sim \mathcal{D}}[\ell(h(x), y)]$ for the objective value of an optimal feasible solution.*

9.2.1 Convex Optimization

We now review some facts about convex optimization with linear constraints.

Definition 54 *Fix a $(\mathcal{H}, \mathcal{G}, C)$-convex minimization problem with linear constraints, defined by $(\ell, \{(\ell_j, g_j, S_j)\}_{j=1}^k)$. The corresponding* Lagrangian *is the function $L : \mathcal{H} \times \mathbb{R}^k_{\geq 0} \to \mathbb{R}$ defined as:*

$$L(\mathcal{P}, \lambda) = \mathop{\mathbb{E}}_{h \sim \mathcal{P}, (x,y) \sim \mathcal{D}} [\ell(h(x), y)] + \sum_{j=1}^{k} \lambda_j \mathop{\mathbb{E}}_{h \sim \mathcal{P}, (x,y) \sim \mathcal{D}} [\ell_j(h(x), y) | g_j(x) = 1, y \in S_j]$$

Definition 55 *Fix a $(\mathcal{H}, \mathcal{G}, C)$-convex minimization problem with linear constraints, and let $L : \mathcal{H} \times \mathbb{R}^k_{\geq 0} \to \mathbb{R}$ be its Lagrangian. We say that $\mathcal{P}^* \in \Delta\mathcal{H}$ and $\lambda^* \in \mathbb{R}^k$ are an optimal primal/dual pair for L if we have both that:*

1.
$$\mathcal{P}^* \in \arg\min_{P \in \Delta\mathcal{H}} L(\mathcal{P}, \lambda^*)$$

and

2.
$$\lambda^* \in \arg\max_{\lambda \in \mathbb{R}^k} L(\mathcal{P}^*, \lambda)$$

We'll state an important theorem in convex optimization here without proof:

Theorem 42 (Strong Duality and Complementary Slackness) *Fix a feasible $(\mathcal{H}, \mathcal{G}, C)$-convex minimization problem with linear constraints, defined by $(\ell, \{(\ell_j, g_j, S_j)\}_{j=1}^k)$.*

For every optimal solution $\mathcal{P}^*$, there is a corresponding vector λ^* such that $(\mathcal{P}^*, \lambda^*)$ form an optimal primal/dual pair.

Moreover every primal dual pair $(\mathcal{P}^*, \lambda^*)$, satisfies:

1. *$\mathcal{P}^*$ is a feasible, optimal solution to the optimization problem, and*
2. *For every constraint $j \in [k]$:*

$$\lambda_j^* \cdot \left(\mathop{\mathbb{E}}_{h \sim \mathcal{P}^*, (x,y) \sim \mathcal{D}} [\ell_j(h(x), y) | g_j(x) = 1, y \in S_j] \right) = 0$$

The second condition is called "Complementary Slackness"

A simple corollary of Theorem 42 is that the Lagrangian of an optimization problem takes value OPT when evaluated at an optimal primal/dual pair.

Corollary 9.2.1 *Fix a feasible $(\mathcal{H}, \mathcal{G}, C)$-convex minimization problem with linear constraints, defined by $(\ell, \{(\ell_j, g_j, S_j)\}_{j=1}^k)$. Let $L : \mathcal{H} \times \mathbb{R}^k_{\geq 0} \to \mathbb{R}$ be its corresponding Lagrangian, and let $(\mathcal{P}^*, \lambda^*)$ be an optimal primal/dual pair for L. Then:*

$$L(\mathcal{P}^*, \lambda^*) = OPT$$

.

Proof 75 *Using both parts of Theorem 42, we can compute:*

$$\begin{aligned}
& L(\mathcal{P}^*, \lambda^*) \\
= & \mathop{\mathbb{E}}_{h \sim \mathcal{P}^*, (x,y) \sim \mathcal{D}} [\ell(h(x), y)] + \sum_{j=1}^k \lambda_j^* \mathop{\mathbb{E}}_{h \sim \mathcal{P}^*, (x,y) \sim \mathcal{D}} [\ell_j(h(x), y) | g_j(x) = 1, y \in S_j] \\
= & OPT + \sum_{j=1}^k \lambda_j^* \mathop{\mathbb{E}}_{h \sim \mathcal{P}^*, (x,y) \sim \mathcal{D}} [\ell_j(h(x), y) | g_j(x) = 1, y \in S_j] \\
= & OPT
\end{aligned}$$

Here the second to last inequality follows from the fact that $\mathcal{P}^*$ is an optimal solution to the optimization problem, and the last inequality follows from complementary slackness.

9.2.2 f-estimated Optimization

The optimization problems we have defined (and their corresponding Lagrangians) are defined as expectations over both x and y — so in order to evaluate a solution $\mathcal{P}$ (or to solve for one), we need access to labelled examples. Just as we did in Section 9.1 for unconstrained optimization, given a model $f : \mathcal{X} \to \mathbb{R}$ that purports to encode $f(x) = \mathbb{E}[y|x]$, we can define an f-estimated optimization problem whose definition only involves expectations taken over features $x \sim \mathcal{D}_{\mathcal{X}}$.

Definition 56 *Fix an $(\mathcal{H}, \mathcal{G}, C)$-convex minimization problem with objective ℓ and linear constraints is defined by $\{(\ell_j, g_j, S_j)\}_{j=1}^k$. Fix a model $f : \mathcal{X} \to [0, 1]$. The corresponding f-estimated optimization problem defined as:*

$$\arg\min_{\mathcal{P} \in \Delta\mathcal{H}} \mathop{\mathbb{E}}_{h\sim\mathcal{P}, x\sim\mathcal{D}_{\mathcal{X}}} [f(x)\ell(h(x), 1) + (1 - f(x))\ell(h(x), 0)]$$

Subject to the constraints that for each $j \in [k]$ with $S = \{0, 1\}$:

$$\mathop{\mathbb{E}}_{h\sim\mathcal{P}, x\sim\mathcal{D}_{\mathcal{X}}} [f(x)\ell_j(h(x), 1) + (1 - f(x))\ell_j(h(x), 0)|g_j(x) = 1] \leq 0$$

for each $j \in [k]$ with $S = \{1\}$:

$$\mathop{\mathbb{E}}_{h\sim\mathcal{P}, x\sim\mathcal{D}_{\mathcal{X}}} [\ell_j(h(x), 1)|g_j(x) = 1] \leq 0$$

and for each $j \in [k]$ with $S = \{0\}$:

$$\mathop{\mathbb{E}}_{h\sim\mathcal{P}, x\sim\mathcal{D}_{\mathcal{X}}} [\ell_j(h(x), 0)|g_j(x) = 1] \leq 0$$

If there is any solution $\mathcal{P}$ that satisfies all of the constraints, we say that the optimization problem is feasible. *We write $\tilde{\mathcal{P}}^*$ for the solution that minimizes the objective function while satisfying the constraints, and write $\tilde{OPT} = \mathbb{E}_{h\sim\tilde{\mathcal{P}}^*, x\sim\mathcal{D}_{\mathcal{X}}}[f(x)\ell(h(x), 1) + (1 - f(x))\ell(h(x), 0)]$ for the objective value of an optimal feasible solution.*

We can similarly define the f-estimated Lagrangian:

Definition 57 *Fix an f-estimated $(\mathcal{H}, \mathcal{G}, C)$-convex minimization problem with linear constraints, defined by $(\ell, \{(\ell_j, g_j, S_j)\}_{j=1}^k)$. Partition the constraints such that $C_0 = \{j \in [k] : S_j = \{0\}\}$, $C_1 = \{j \in [k] : S_j = \{1\}\}$, and $C_{01} = \{j \in [k] : S_j = \{0, 1\}\}$.*

The corresponding f-estimated Lagrangian *is the function $\tilde{L} : \mathcal{H} \times \mathbb{R}^k_{\geq 0} \to \mathbb{R}$ defined as:*

$$\tilde{L}(\mathcal{P}, \lambda) = \mathop{\mathbb{E}}_{h\sim\mathcal{P}, x\sim\mathcal{D}_{\mathcal{X}}} [f(x)\ell(h(x), 1) + (1 - f(x))\ell(h(x), 0)]$$

$$+ \sum_{j\in C_0} \lambda_j \mathop{\mathbb{E}}_{h\sim\mathcal{P}, x\sim\mathcal{D}_{\mathcal{X}}} [\ell_j(h(x), 0)|g_j(x) = 1] + \sum_{j\in C_1} \lambda_j \mathop{\mathbb{E}}_{h\sim\mathcal{P}, x\sim\mathcal{D}_{\mathcal{X}}} [\ell_j(h(x), 1)|g_j(x) = 1]$$

$$+ \sum_{j\in C_{01}} \lambda_j \mathop{\mathbb{E}}_{h\sim\mathcal{P}, x\sim\mathcal{D}_{\mathcal{X}}} [f(x)\ell_j(h(x), 1) + (1 - f(x))\ell_j(h(x), 0)|g_j(x) = 1]$$

9.2.3 Solving Optimization Problems Without Labelled Data

Our goal is to derive a constrained optimization analogue of our results from Section 9.1, which were for *unconstrained* optimization. Namely, we would like to train a single model f using labelled data from $\mathcal{D}$, such that f is sufficient to solve a wide variety of downstream *constrained optimization* problems using only unlabelled data $x \sim \mathcal{D}_{\mathcal{X}}$ and minimal additional computation. The main idea will be to train a predictor f that is multicalibrated with respect to $\mathcal{G}$, $\mathcal{H}$, and their corresponding *product class*:

Definition 58 *Fix two classes of functions $\mathcal{G}$ and $\mathcal{H}$ mapping features to real numbers. The product class is defined as:*

$$\mathcal{G} \cdot \mathcal{H} = \{g(x) \cdot h(x) : g(x) \in \mathcal{G}, h(x) \in \mathcal{H}\}$$

We will argue that if we have such a predictor f, then if we want to solve some $(\mathcal{H}, \mathcal{G}, C)$-convex minimization problem with linear constraints, it will be sufficient to solve its corresponding f-estimated variant in which $\mathcal{H}$ has been replaced by $\mathcal{H}_{\text{all}}$, the set of *all* functions $f : \mathcal{X} \to [0, 1]$, which (as we will see) is a computationally easier task, and one that does not require a randomized solution.

Definition 59 *We write $\mathcal{H}_{all} = \{f : \mathcal{X} \to [0, 1]\}$ for the set of all real-valued functions mapping features to the unit interval.*

In particular, the optimal solution h to an f-estimated $(\mathcal{H}_{\text{all}}, \mathcal{G}, C)$-optimization problem will be a *policy* in the sense that we can write it in the form $h(x) = \rho(f(x))$, that depends only on $f(x)$.

Lemma 9.2.1 *Fix any model $f : \mathcal{X} \to [0, 1]$ and any f-estimated $(\mathcal{H}_{all}, \mathcal{G}, C)$-convex optimization problem with linear constraints. Let $h \in \mathcal{H}_{all}$ be an optimal solution to the problem. Then $h(x)$ can be written as a policy of $f(x)$: $h(x) = \rho(f(x))$ for some $\rho : [0, 1] \to [0, 1]$.*

Proof 76

Theorem 43 *Fix any feasible $(\mathcal{H}, \mathcal{G}, C)$-convex minimization problem with linear constraints, defined by $(\ell, \{(\ell_j, g_j, S_j)\}_{j=1}^k)$. Fix any $f : \mathcal{X} \to [0, 1]$ that is α-approximately L_1-calibrated and L_1-multicalibrated with respect to $\mathcal{H}$, $\mathcal{G}$, and $\mathcal{H} \cdot \mathcal{G}$. Let $\tilde{\mathcal{P}}^*$ be an optimal solution to the corresponding f-estimated $(\mathcal{H}_{all}, \mathcal{G}, C)$-optimization problem. Then we have that $\tilde{\mathcal{P}}^*$ is approximately optimal according to the original objective function and approximately satisfies the original constraints:*

$$BLAH$$

Proof 77 *First we argue about the objective value. From Theorem 42, we know that there exists a $\tilde{\lambda}^*$ such that $(\tilde{\mathcal{P}}^*, \tilde{\lambda}^*)$ form an optimal primal/dual*

pair for the corresponding f-estimated Lagrangian $\tilde{L}$. From Corollary 9.2.1 we know that $\tilde{L}(\tilde{\mathcal{P}}^*, \tilde{\lambda}^*) = \tilde{OPT}$. Similarly, let $\mathcal{P}^*$ be an optimal solution to the original $(\mathcal{H}, \mathcal{G}, C)$-optimization problem. We know from Theorem 42 that there exists a λ^* such that $(\mathcal{P}^*, \lambda^*)$ form an optimal primal/dual pair for the corresponding Lagrangian L, and from Corollary 9.2.1 that $L(\mathcal{P}^*, \lambda^*) = OPT$.

We also know from Theorem 40 since f is α-approximately calibrated with respect to $\mathcal{H}$ that:

||

Thus we can calculate:

$$\begin{aligned} \tilde{OPT} &= L(\tilde{\mathcal{P}}^*, \tilde{\lambda}^*) \\ &\leq \end{aligned}$$

Can prove this if we need it

We recall a piece of notation from our earlier chapters: $\mu(g, \mathcal{D}) = \Pr_{(x,y)\sim\mathcal{D}}[g(x) = 1]$.

Lemma 9.2.2 *Fix any distribution $\mathcal{D} \in \Delta\mathcal{Z}$, any class of group indicator functions $\mathcal{G}$ containing functions $g : \mathcal{X} \to \{0, 1\}$ and any class of real valued functions $\mathcal{H}$ containing functions $h : \mathcal{X} \to \mathbb{R}$. For each $g \in \mathcal{G}$ define the distribution $\mathcal{D}_g = \mathcal{D}|g(x) = 1$ be the conditional distribution conditional on $g(x) = 1$. Suppose a model $f : \mathcal{X} \to \mathbb{R}$ is α L_1-multicalibrated with respect to $\mathcal{D}$ and $\mathcal{G} \cdot \mathcal{H}$. Then for every $g \in \mathcal{G}$ and $h \in \mathcal{H}$:*

$$K_1(f, h, \mathcal{D}_g) \leq \frac{\alpha}{\mu(g, \mathcal{D})}$$

Proof 78 *By hypothesis we know that for every $h \in \mathcal{H}$ and $g \in \mathcal{G}$:*

$$\begin{aligned} K_1(f, g \cdot h, \mathcal{D}) &= \sum_{v \in R(f)} \Pr[f(x) = v] \, \mathbb{E}[g(x)h(x)(y - v)|f(x) = v] \\ &\leq \alpha \end{aligned}$$

Fix any $g \in \mathcal{G}$ with $\mu(g, \mathcal{D}) > 0$. We can now calculate:

$$\begin{aligned} & K_1(f, h, \mathcal{D}_g) \\ = & \sum_{v \in R(f)} \Pr[f(x) = v|g(x) = 1] \, \mathbb{E}[h(x)(y - v)|g(x) = 1, f(x) = v] \end{aligned}$$

References and Further Reading

The fact that a model f that is multicalibrated with respect to a class of functions $\mathcal{H}$ can be post-processed in such a way to be competitive with any $h \in \mathcal{H}$ as measured by any convex Lipschitz loss function was proven in Gopalan et al.

[2022b], who called such models "omnipredictors". Gopalan et al. [2022b] use a slightly different notion of calibration than we do (based on partitions of the feature space and covariance), but if a model satisfies our notion of multicalibration and is also calibrated, then it also satisfies the covariance based notion and vice versa. Gopalan et al. [2022a] give an incomparable omniprediction theorem — they show that group conditional mean consistency together with (marginal) calibration is sufficient to be competitive with any $h \in \mathcal{H}$ on any Lipschitz loss function ℓ (no longer requiring convexity of ℓ or full multicalibration) — but in general this requires group conditional mean consistency with respect to all *level sets* of functions in $\mathcal{H}$, rather than just with $\mathcal{H}$ itself.

10

Ensembling, Model Multiplicity, and the Reference Class Problem

CONTENTS

Suppose we have a prediction problem of some import: perhaps we are selling life insurance, and we want to predict the probability that particular customers will die in the next 12 months. We are in a familiar regression setting, in which we have some space of individuals $\mathcal{X}$ and would like a model $f : \mathcal{X} \to [0, 1]$, where ideally $f(\text{Bob})$ should have the semantics that "$f(\text{Bob})$ is the probability that Bob will die in the next 12 months". But what does this mean? We are predicting a probability for a single event that will occur or not — there are to be no repeated trials for which we can measure an empirical frequency. If I propose a model f_1 that purports to assign individual probabilities to people like Bob, and you propose a different model f_2, how are we to resolve which model is "better"?

10.1 Reference Classes and Model Multiplicity

Suppose we posit that there are true "individual probabilities" underlying reality — i.e. that there is in principle some number p_{Bob} that represents the probability that Bob will die in the next 12 months. This is after all the formalism that has underlied our studies so far: we have been modeling the world as if there is a distribution $\mathcal{D}$ over labelled examples (x, y), and for each individual x a conditional label distribution $\mathcal{D}_Y(x)$. We still cannot get access to these individual probabilities through data. Nevertheless, we know that the function f^* encoding the true conditional label distribution

$f^*(x) = \mathbb{E}_{y \sim \mathcal{D}(x)}[y]$ minimizes the expected Brier score:

$$f^* \in \underset{f:\mathcal{X}\to[0,1]}{\arg\min} \underset{(x,y)\sim\mathcal{D}}{\mathbb{E}} \left[(f(x) - y)^2\right]$$

Hence if we have two models such that $B(f_1) < B(f_2)$, this falsifies the hypothesis that $f_2 = f^*$ — i.e. it cannot be the case that f_2 represents the true individual probabilities, and gives us an empirical (and practical!) justification for adopting model f_1 rather than model f_2.

The "model multiplicity" problem refers to the worry that there may be multiple models f_1, f_2 that are equally accurate (such that $B(f_1) = B(f_2)$) that disagree in their predictions. In this case, accuracy gives us no basis on which to reject either model, and yet if f_1(Bob) is very different from f_2(Bob), what basis do we have to act on our predictions? Are we justified in denying Bob life insurance if it seems unprofitable according to the individual probability assigned by f_2 but seems profitable according to the individual probability assigned by f_1?

This can indeed be a problem if the models f are chosen to optimize accuracy in some fixed class. But as we will see, the situation cannot arise if the parties proposing their models are willing to update (and improve!) their models in the face of evidence that can be found in the data before them and in the competing models that are proposed! The updates needed will be of exactly the same simple "patch" form that we have studied when deriving algorithms for multicalibration and group conditional mean consistency.

10.2 Model Ensembling

Suppose we are given two models $f_1, f_2 : \mathcal{X} \to [0, 1]$. 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 60 *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. 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 using our patch operations 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 61 *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:

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

Lemma 10.2.1 *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:

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

Proof 79 *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 $\mu(U_\epsilon^\bullet(f_1, f_2))$ 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:

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

Lets 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 ($U_\epsilon^>(f_1^{t_1}, f_2^{t_2})$ or $U_\epsilon^<(f_1^{t_1}, f_2^{t_2})$) that falsifies either the assertion that f_1 encodes true conditional label expectations or the assertion that f_2 does. And not only does it falsify that at least one of f_1 or f_2 are a "correct" model — it provides a directly actionable way to improve one of the models. Recall Lemma 4.1.1, which we proved when analyzing an algorithm for guaranteeing group conditional mean consistency, and we reproduce here:

Lemma 10.2.2 *Fix any model $f_t : \mathcal{X} \to [0,1]$ and group $g : \mathcal{X} \to \{0,1\}$. 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:

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

Then:

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

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 $O(\alpha\epsilon^2)$.

We put this all together in Algorithm 26 (Reconciler).

Algorithm 26 Reconcile($f_1, f_2, \alpha, \epsilon$)

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:

$$v_*^\bullet = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y | x \in U_\epsilon^\bullet(f_1^{t_1}, f_2^{t_2})] \quad v_i^\bullet = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f_i^{t_i}(x) | x \in U_\epsilon^\bullet(f_1^{t_1}, f_2^{t_2})]$$

Let:

$$(i_t, \bullet_t) = \mathop{\arg\max}_{i\in\{1,2\}, \bullet\in\{>,<\}} \mu(U_\epsilon^\bullet(f_1^{t_1}, f_2^{t_2})) \cdot (v_*^\bullet - v_i^\bullet)^2$$

Let:

$$g_t(x) = \begin{cases} 1 & x \in U_\epsilon^\bullet(f_1^{t_1}, f_2^{t_2}) \\ 0 & \text{otherwise} \end{cases}$$

Let:

$$\tilde{\Delta}_t = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y | g_t(x) = 1] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f_i^{t_i}(x) | g_t(x) = 1]$$

$$\Delta_t = \text{Round}(\tilde{\Delta}_t; m)$$

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 44 *For any pair of models $f_1, f_2 : \mathcal{X} \to [0, 1]$ and any $\alpha, \epsilon > 0$, Algorithm 26 (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. $T \leq (B(f_1) + B(f_2)) \cdot \frac{16}{\alpha\epsilon^2}$
2. $B(f_1^{T_1}) \leq B(f_1) - T_1 \cdot \frac{\alpha\epsilon^2}{16}$ *and* $B(f_2^{T_2}) \leq B(f_2) - T_2 \cdot \frac{\alpha\epsilon^2}{16}$
3. $\mu(U_\epsilon(f_1^{T_1}, f_2^{T_2})) < \alpha$.

Proof 80 *By Lemma 10.2.1, for each round $t < T$ we must have that:*

$$\mathop{\arg\max}_{i\in\{1,2\}, \bullet\in\{>,<\}, v\in[1/2\epsilon]} \mu(S(v, \bullet)) \cdot (v_*^\bullet - v_i^\bullet)^2 \geq \frac{\alpha\epsilon^2}{8}$$

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 10.2.2, we have that:

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

.

We can now compute

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

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 10.2.2 to conclude that:*

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

Combining this with our initial calculation lets us conclude that:

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

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 or* f_1 *and* f_2 *respectively we get that:* $B(f_1^{T_1}) \leq B(f_1) - T_1 \cdot \frac{\alpha\epsilon^2}{16}$ *and* $B(f_2^{T_2}) \leq B(f_2) - T_2 \cdot \frac{\alpha\epsilon^2}{16}$. *Since Brier scores are non-negative, we conclude that* $T_1 \leq B(f_1)\frac{16}{\alpha\epsilon^2}$ *and* $T_2 \leq B(f_2)\frac{16}{\alpha\epsilon^2}$. *Thus* $T = T_1 + T_2 \leq (B(f_1) + B(f_2)) \cdot \frac{16}{\alpha\epsilon^2}$

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 probability estimates for any sufficiently large conditional probability.

Corollary 10.2.1 *Let $E \subset \mathcal{X}$ be any subset of the data space. Let f_1 and f_2 be any two models that have been output by Algorithm 26 (Reconcile) with parameters ϵ and α. Let:*

$$p_1(E) = \sum_{x \in E} \frac{\mu(x) \cdot f_1(x)}{\mu(E)} \text{ and } p_2(E) = \sum_{x \in E} \frac{\mu(x) \cdot f_2(x)}{\mu(E)}$$

be the estimates for $\mathbb{E}[y|x \in E]$ implied by models f_1 and f_2 respectively. Then:

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

Proof 81 *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:*

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

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

10.3 Sample Complexity

We have once again presented our algorithm 26 as if it has direct access to the distribution $\mathcal{D}$. Of course in general we do not have access to $\mathcal{D}$, but rather have access to some set D of n i.i.d. *samples* from $\mathcal{D}$. We will typically instead run Algorithm 26 over the *empirical distribution* over D — i.e. the distribution that puts weight $1/n$ on each datapoint $(x_i, y_i) \in D$. We will prove that with high probability over the sample of D, when Algorithm 26 is run over the empirical distribution on D, then its guarantees translate over to the distribution $\mathcal{D}$ from which D was drawn, with error parameters that go to zero with the size n of the data sample.

We begin by counting the number of potential models $f_1^{t_1}, f_2^{t_2}$ that Algorithm 26 might output.

Lemma 10.3.1 *Fix any pair of models $f_1, f_2 : \mathcal{X} \to [0, 1]$ and any $\alpha, \epsilon > 0$. Then there is a set C of pairs of models of size at most $|C| \leq$*

$(4 \cdot (m+1))^{32/\alpha\epsilon^2+1}$ *such that for any dataset distribution* $\mathcal{D}$ *on which Algorithm 26 is run, the output models* $(f_1^{t_1}, f_2^{t_2}) \in C$. *Here, as in Algorithm 26,* $m = \lceil \frac{2}{\sqrt{\alpha\epsilon}} \rceil$.

Proof 82 *Given a run of Algorithm 26 for* T *rounds, let* $\pi = \{(i_t, \bullet_t, \Delta_t)\}_{t=1}^{T}$ *denote the record of the quantities* $(i_t, \bullet_t, \Delta_t)$ *chosen at each round* t. *Let* $\pi^{<t} = \{(i_{t'}, \bullet_{t'}, \Delta_{t'})\}_{t'=1}^{t-1}$ *denote the prefix of this transcript up through round* $t-1$. *Observe that once we fix* $\pi^{<t}$ *we have also fixed the models* $f_1^{t_1}$ *and* $f_2^{t_2}$ *that are defined at the start of round* t. *To see this, assume the claim holds true at round* t. *In particular,* $\pi^{<t}$ *fixes the disagreement regions* $U_\epsilon^\bullet(f_1^{t_1}, f_2^{t_1})$ *of these two models, and therefore given the choices* $(i_t, \bullet_t, \Delta_t)$, *we have inductively defined the models present at the start of round* $t+1$.

We let C *denote the set of all pairs of models defined by transcripts* $\pi^{<T}$ *for all* $T \leq \frac{32}{\alpha\epsilon^2}$. *Since we know from Theorem 44 that Algorithm 26 halts after at most* $T \leq (B(f_1) + B(f_2)) \cdot \frac{16}{\alpha\epsilon^2} \leq \frac{32}{\alpha\epsilon^2}$ *many rounds, and hence the models output by Algorithm 26 must be contained in* C *as claimed. It remains to count the set of transcripts of length* $T \leq \frac{32}{\alpha\epsilon^2}$. *At each round* t, *there are two possible values for* i_t, *two possible values for* $\bullet_t$, *and* $m+1$ *possible choices for* Δ_t. *Hence the number of transcripts of length* T *is* $(4(m+1))^T$. *Thus we have:*

$$|C| \leq \sum_{T=0}^{\frac{32}{\alpha\epsilon^2}} (4(m+1))^T \leq (4(m+1))^{\frac{32}{\alpha\epsilon^2}+1}$$

We can now argue that if we have a sample of n datapoints D that are sampled i.i.d. from some unknown distribution $\mathcal{D}$, then if we run Algorithm 26 using the empirical distribution over D, then its guarantees hold also over $\mathcal{D}$, with error terms that tend to 0 as n grows large.

Theorem 45 *Fix any data distribution* $\mathcal{D}$ *and consider a run of Algorithm 26 over the empirical distribution over points in a dataset* $D \sim \mathcal{D}^n$ *consisting of* n *points sampled i.i.d. from* $\mathcal{D}$. *For any pair of models* $f_1, f_2 : \mathcal{X} \to [0,1]$ *and any* $\alpha, \epsilon > 0$, *Algorithm 26 (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. $T \leq \frac{16}{\alpha\epsilon^2}$
2. *For any* $\delta > 0$, *with probability at least* $1-\delta$ *over the randomness of* $D \sim \mathcal{D}^n$ *we have that:*

$$B(f_1^{T_1}) \leq B(f_1) - T_1 \cdot \frac{\alpha\epsilon^2}{16} + 2\sqrt{\frac{\left(\frac{16}{\alpha\epsilon^2}+1\right)\log\left(\frac{64\left(\lceil\frac{2}{\sqrt{\alpha\epsilon}}\rceil+1\right)}{\delta}\right)}{n}}$$

and

$$B(f_2^{T_2}) \leq B(f_2) - T_2 \cdot \frac{\alpha\epsilon^2}{16} + 2\sqrt{\frac{\left(\frac{16}{\alpha\epsilon^2}+1\right)\log\left(\frac{64\left(\lceil\frac{2}{\sqrt{\alpha\epsilon}}\rceil+1\right)}{\delta}\right)}{n}}$$

3. For any $\delta > 0$, with probability at least $1-\delta$ over the randomness of $D \sim \mathcal{D}^n$:

$$\mu(U_\epsilon(f_1^{T_1}, f_2^{T_2})) < \alpha + \sqrt{\frac{\left(\frac{32}{\alpha\epsilon^2}+1\right)\log\left(\frac{8\left(\lceil\frac{2}{\sqrt{\alpha}\epsilon}\rceil+1\right)}{\delta}\right)}{n}}.$$

Remark 10.3.1 *Theorem 45 tells us that the guarantees we proved for Algorithm 26 in Theorem 44 (when we assumed direct access to the distribution $\mathcal{D}$) continue to hold when all we have access to is a finite sample of n points from the data distribution, with additional error terms that tend to zero as n grows large. How large is large? If we want the final disagreement region to have mass at most 2α (i.e. we want the third conclusion of Theorem 45 to tell us that $\mu(U_\epsilon(f_1^{T_1}, f_2^{T_2})) < 2\alpha$), then solving for n in the error bound, we find that it suffices to have n samples for n on the order of:*

$$n \in \tilde{O}\left(\frac{\log(1/\delta)}{\alpha^3\epsilon^2}\right)$$

where the $\tilde{O}()$ notation hides logarithmic terms in $1/\alpha$ and $1/\epsilon$.

This is a remarkably small amount of data: We would need $\approx \frac{\log(1/\delta)}{\alpha\epsilon^2}$ samples just to estimate the conditional label expectation $\Pr[y=1|x \in S]$ for a conditional event S with $\mu(S) = \alpha$ up to error ϵ with probability $1-\delta$ (or for two parties with disjoint samples to agree on this conditional label expectation up to error ϵ). Theorem 45 tells us that in fact two parties can be made to agree on a $1-\alpha$ fraction of points up to error ϵ with an additional amount of data only on the order of $\tilde{O}(1/\alpha^2)$.

Proof 83 (Proof of Theorem 45) *The bound on T follows directly from Theorem 44 without modification. We focus on bounding the Brier scores and the uncertainty region for the resulting models.*

Consider any pair of models f_1, f_2. Given a finite dataset D we write $(x,y) \sim D$ to denote uniformly sampling a single datapoint from D. We start by comparing $\Pr_{(x,y)\sim D}[x \in U_\epsilon(f_1,f_2)]$ with $\Pr_{(x,y)\sim\mathcal{D}}[x \in U_\epsilon(f_1,f_2)]$. We have that:

$$\Pr_{(x,y)\sim D}[x \in U_\epsilon(f_1,f_2)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[x_i \in U_\epsilon(f_1,f_2)]$$

Since $\mathbb{1}[x_i \in U_\epsilon(f_1,f_2)] \in [0,1]$ and

$$\mathop{\mathbb{E}}_{D\sim\mathcal{D}^n}\left[\Pr_{(x,y)\sim D}[x \in U_\epsilon(f_1,f_2)]\right] = \Pr_{(x,y)\sim\mathcal{D}}[x \in U_\epsilon(f_1,f_2)]$$

we can apply Hoeffding's inequality (Theorem 46) to conclude that for every $\eta > 0$:

$$\Pr_{D\sim\mathcal{D}^n}\left[\left|\Pr_{(x,y)\sim D}[x \in U_\epsilon(f_1,f_2)] - \Pr_{(x,y)\sim\mathcal{D}}[x \in U_\epsilon(f_1,f_2)]\right| \geq \eta\right] \leq 2\exp\left(-2\eta^2 n\right)$$

Let C be the set of pairs of models guaranteed in the statement of Lemma 10.3.1. Recall that Lemma 10.3.1 guarantees us that $|C| \leq (4(m+1)^{32/\alpha\epsilon^2+1}$. We can apply the union bound to all pairs of models $(f_1, f_2) \in C$ to conclude that with probability at least $1-2|C|\exp\left(-2\eta^2 n\right)$ (over the randomness of D) we have that for every *pair $(f_1, f_2) \in C$:*

$$\left| \Pr_{(x,y)\sim D}[x \in U_\epsilon(f_1, f_2)] - \Pr_{(x,y)\sim \mathcal{D}}[x \in U_\epsilon(f_1, f_2)] \right| \leq \eta$$

Choosing

$$\eta = \sqrt{\frac{\log\left(\frac{2|C|}{\delta}\right)}{2n}}$$

we get that with probability $1-\delta$ over the draw of D, for every pair $(f_1, f_2) \in C$:

$$\begin{aligned}\left| \Pr_{(x,y)\sim D}[x \in U_\epsilon(f_1, f_2)] - \Pr_{(x,y)\sim \mathcal{D}}[x \in U_\epsilon(f_1, f_2)] \right| &\leq \sqrt{\frac{\log\left(\frac{2|C|}{\delta}\right)}{2n}} \\ &\leq \sqrt{\frac{\left(\frac{32}{\alpha\epsilon^2}+1\right)\log\left(\frac{8\left(\lceil\frac{2}{\sqrt{\alpha}\epsilon}\rceil+1\right)}{\delta}\right)}{n}}\end{aligned}$$

where the final inequality follows from plugging in our bound on $|C|$ and the definition of m.

Because we know from Theorem 44 that the models $f_1^{T_1}, f_2^{T_2}$ output by Algorithm 26 satisfy that $\Pr_{(x,y)\sim D}[x \in U_\epsilon(f_1^{T_1}, f_2^{T_2})] \leq \alpha$ we can conclude that with probability $1-\delta$:

$$\Pr_{(x,y)\sim \mathcal{D}}[x \in U_\epsilon(f_1^{T_1}, f_2^{T_2})] \leq \sqrt{\frac{\left(\frac{32}{\alpha\epsilon^2}+1\right)\log\left(\frac{8\left(\lceil\frac{2}{\sqrt{\alpha}\epsilon}\rceil+1\right)}{\delta}\right)}{n}}$$

We can bound the Brier score of the resulting models in exactly the same way. For any fixed model $f : \mathcal{X} \to [0,1]$, we can write the empirical Brier score (i.e. the Brier score as evaluated over D) as:

$$B_D(f) = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2$$

Since $(f(x_i)-y_i)^2 \in [0,1]$ and $\mathbb{E}_{D\sim\mathcal{D}^n}[B_D(f)]$, we can apply Hoeffding's inequality (Theorem 46) exactly as before to conclude that for every pair of models $(f_1^{T_1}, f_2^{T_2}) \in C$, with probability $1-\delta$:

$$\left|B_D(f_1^{T_1}) - B(f_1^{T^1})\right| \leq \sqrt{\frac{\left(\frac{16}{\alpha\epsilon^2}+1\right)\log\left(\frac{16\left(\lceil\frac{2}{\sqrt{\alpha}\epsilon}\rceil+1\right)}{\delta}\right)}{n}}$$

and with probability $1-\delta$*:*

$$\left|B_D(f_2^{T_2}) - B(f_2^{T_2})\right| \leq \sqrt{\frac{\left(\frac{16}{\alpha\epsilon^2}+1\right)\log\left(\frac{16\left(\lceil\frac{2}{\sqrt{\alpha}\epsilon}\rceil+1\right)}{\delta}\right)}{n}}$$

Observe that the same holds true for the original pair of models (f_1, f_2)*, since* $(f_1, f_2) \in C$ *(they correspond to the models output after transcripts of length 0). We further know from Theorem 44 that:* $B_D(f_1^{T_1}) \leq B_D(f_1) - T_1 \cdot \frac{\alpha\epsilon^2}{16}$ *and* $B_D(f_2^{T_2}) \leq B_D(f_2) - T_2 \cdot \frac{\alpha\epsilon^2}{16}$*.*

Instantiating these bounds for the four models $\{f_1, f_2, f_1^{T_1}, f_2^{T_2}\}$*, and setting* $\delta \leftarrow \delta/4$ *so that we can union bound over all four models, we have that with probability* $1-\delta$ *that we simultaniously have:*

$$B(f_1^{T_1}) \leq B(f_1) - T_1 \cdot \frac{\alpha\epsilon^2}{16} + 2\sqrt{\frac{\left(\frac{16}{\alpha\epsilon^2}+1\right)\log\left(\frac{64\left(\lceil\frac{2}{\sqrt{\alpha}\epsilon}\rceil+1\right)}{\delta}\right)}{n}}$$

$$B(f_2^{T_1}) \leq B(f_2) - T_2 \cdot \frac{\alpha\epsilon^2}{16} + 2\sqrt{\frac{\left(\frac{16}{\alpha\epsilon^2}+1\right)\log\left(\frac{64\left(\lceil\frac{2}{\sqrt{\alpha}\epsilon}\rceil+1\right)}{\delta}\right)}{n}}$$

References and Further Reading

The material from this Chapter is taken from Roth et al. [2022].

Bibliography

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2020.

Christina Baek, Yiding Jiang, Aditi Raghunathan, and Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. *arXiv preprint arXiv:2206.13089*, 2022.

Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Practical adversarial multivalid conformal prediction. *arXiv preprint arXiv:2206.01067*, 2022.

Maya Burhanpurkar, Zhun Deng, Cynthia Dwork, and Linjun Zhang. Scaffolding sets. *arXiv preprint arXiv:2111.03135*, 2021.

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

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

Dean P Foster. A proof of calibration via blackwell's approachability theorem. *Games and Economic Behavior*, 29(1-2):73–78, 1999.

Dean P Foster and Sergiu Hart. Forecast hedging and calibration. *Journal of Political Economy*, 129(12):3447–3490, 2021.

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*, 2020.

Drew Fudenberg and David K Levine. An easier way to calibrate. *Games and economic behavior*, 29(1-2):131–137, 1999.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34: 1660–1672, 2021.

Ira Globus-Harris, Declan Harrison, Michael Kearns, Aaron Roth, and Jessica Sorrell. Multicalibration as boosting for regression. 2023.

Parikshit Gopalan, Lunjia Hu, Michael P Kim, Omer Reingold, and Udi Wieder. Loss minimization through the lens of outcome indistinguishability. *arXiv preprint arXiv:2210.08649*, 2022a.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022b.

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

Sergiu Hart. Calibrated forecasts: The minimax proof. 2020. URL `http://www.ma.huji.ac.il/~hart/papers/calib-minmax.pdf`.

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

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

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. *arXiv preprint arXiv:2209.15145*, 2022.

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.

Michael P Kim, Christoph Kern, Shafi Goldwasser, Frauke Kreuter, and Omer Reingold. Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences*, 119(4):e2108097119, 2022.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer, 2002.

Sangdon Park, Osbert Bastani, Nikolai Matni, and Insup Lee. Pac confidence sets for deep neural networks via calibrated prediction. In *International Conference on Learning Representations*, 2019.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel Candès. With malice toward none: Assessing uncertainty via equalized coverage. *Harvard Data Science Review*, 2020.

Aaron Roth, Alexander Tolbert, and Scott Weinstein. Reconciling individual probability forecasts. *arXiv preprint arXiv:2209.01687*, 2022.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

A

Useful Probabilistic Inequalities

CONTENTS

In this appendix we collect several useful probabilistic inequalities that will be handy in our analyses.

Theorem 46 (Hoeffding's Inequality) *Let $X_1, \ldots, X_n$ be independent random variables bounded such that for each i, $a_i \leq X_i \leq b_i$. Let $S_n = \sum_{i=1}^{n} X_i$ denote their sum. Then for all $t > 0$:*

$$\Pr\left[|S_n - \mathbb{E}[S_n]| \geq t\right] \leq 2\exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

Theorem 47 (Chernoff's Bound) *Let $X_1, \ldots, X_n$ be independent random variables bounded such that for each i, $0 \leq X_i \leq 1$. Let $S_n = \sum_{i=1}^{n} X_i$ denote their sum. Then for all $\eta > 0$:*

$$\Pr\left[|S_n - \mathbb{E}[S_n]| \geq \eta\,\mathbb{E}[S_n]\right] \leq 2\exp\left(-\frac{\mathbb{E}[S_n]\eta^2}{3}\right)$$

Theorem 48 (Azuma's Inequality) *Let $X_1, \ldots, X_n$ be random variables (not necessarily independent) bounded such that for each i, $|X_i| \leq c_i$. Let $X_{<i}$ denote the prefix $X_1, X_2, \ldots, X_{i-1}$. Then for all $t > 0$:*

$$\Pr\left[\left|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}[X_i | X_{<i}]\right| \geq t\right] \leq 2\exp\left(\frac{-t^2}{2\sum_{i=1}^{n} c_i^2}\right)$$

Theorem 49 (The DKW (Dvoretzky–Kiefer–Wolfowitz) inequality) *Let $\mathcal{D} \in \mathcal{Z}^n$ be any distribution and let $D \sim \mathcal{D}^n$ consist of n points sampled i.i.d. from D. Let $F(c) = \Pr_{(x,y)\sim\mathcal{D}}[y \leq c]$ denote the CDF of the label distribution induced by $\mathcal{D}$, and let $\hat{F}_D(c) = \frac{1}{n}\sum_{(x,y)\in D} \mathbb{1}[y \leq c]$ denote the CDF of the empirical label distribution induced by D. Then for every $t > 0$:*

$$\Pr\left[\sup_{c\in\mathbb{R}} \left|F(c) - \hat{F}_D(C)\right| \geq t\right] \leq 2\exp\left(-2nt^2\right)$$