Multicalibrated Partitions for Importance Weights

Parikshit Gopalan* VMware Research
Omer Reingold† Stanford University
Vatsal Sharan‡ MIT
Udi Wieder§ VMware Research

Abstract

The ratio between the probability that two distributions R and P give to points x are known as importance weights or propensity scores and play a fundamental role in many different fields, most notably, statistics and machine learning. Among its applications, importance weights are central to domain adaptation, anomaly detection, and estimations of various divergences such as the KL divergence and Renyi divergences between R and P, which in turn have numerous applications. We consider the common setting where R and P are only given through samples from each distribution. The vast literature on estimating importance weights is either heuristic, or makes strong assumptions about R and P or on the importance weights themselves. Indeed, relying on cryptographic assumptions, we show the impossibility of efficiently computing pointwise accurate importance weights.

In this paper, we explore a computational perspective to the estimation of importance weights, which factors in the limitations and possibilities obtainable with bounded computational resources. We significantly strengthen previous work that use the MaxEntropy approach, that define the importance weights based on a distribution Q closest to P, that looks the same as R on every set $C \in \mathcal{C}$, where $\mathcal{C}$ may be a huge collection of sets. We show that the MaxEntropy approach may fail to assign high average scores to sets $C \in \mathcal{C}$, even when the average of ground truth weights for the set is evidently large. We similarly show that it may overestimate the average scores to sets $C \in \mathcal{C}$. We therefore formulate Sandwiching bounds as a notion of set-wise accuracy for importance weights. We study these bounds to show that they capture natural completeness and soundness requirements from the weights and are appealing from the point of view of accuracy and fairness in heterogeneous populations. We present an efficient algorithm that under standard learnability assumptions computes weights which satisfy these bounds.

Our techniques rely on a new notion of multicalibrated partitions of the domain of the distributions. While being a relatively small collection of disjoint sets, such partitions reflect, in a well-defined sense, the complexity of the much larger collection of arbitrarily intersecting sets in $\mathcal{C}$. Stratifying the domain based on multicalibrated partitions implies our computational objectives for importance weights and appear to be useful objects in their own right.

*Email: pgopalan@vmware.com
†Most of the work performed while visiting VMware Research. Research supported in part by NSF Award IIS-1908774. Email: reingold@stanford.edu
‡Most of the work performed while visiting VMware Research. Email: vsharan@mit.edu
§Email: uwieder@vmware.com

Contents

1 Introduction

Consider a scenario where we are observing samples drawn from a distribution R over some domain $\mathcal{X}$. The domain $\mathcal{X}$ might represent demographic information of people, whereas R might be the distribution of people who voted, who applied to college or were afflicted with the flu virus. We have some knowledge about the distribution, perhaps from previous experience, which is captured by a prior distribution P. While P could be a good baseline, it is not necessarily a good model for R. How can we modify P to better represent R? The statistical technique of *importance weighting*, also called density ratio estimation, provides an approach to this problem. For each $x \in \mathcal{X}$, let us define its importance weight under R to be $w^*(x) = R(x)/P(x)$. This defines a function $w^* : \mathcal{X} \to \mathbb{R}^+$.

The problem of estimating these importance weights has been studied by several communities (often under different monikers) and is central to a wide variety of applications. In machine learning, it arises both in unsupervised problems such as anomaly detection and supervised settings such as in domain adaptation. For anomaly detection, the objective is to find points which have a small probability under a prior P but high probability under the observed distribution R; essentially the points with high importance weights [SPST+01, HA04, HTK+08, SST09]. In domain adaptation (also known as covariate shift), we get labelled training data from a distribution P but are interested in good prediction accuracy under a different test distribution R for which we only observe unlabelled data. The importance weights of R under P can be used to reweigh the loss function to correct for the distributional shift [Shi00, Zad04, SM05, BBS07, CMM10].

In information theory, importance weight estimation (where it is sometimes known as Radon-Nikodym derivative estimation), is applied to estimate various divergences such as the KL divergence and Renyi divergences between the distributions P and R [NWJ07, NWJ10, YSK+13, WKV05, WKV09]. These divergence measures themselves find several applications, such as two-sample tests for distinguishing between two distributions [WF16] and for independence testing [SSSK08].

In econometrics and statistics, importance weights (under the guise of propensity scores) play a major role in the theory of causal inference from observational data. Propensity scores denote the probability of an individual being selected for a treatment—this is just a scaling of the importance weights of the distribution R of the individuals selected for the treatment, under the prior P of the entire population. These scores are widely used to correct for bias introduced by confounding variables which influence the probability of getting the treatment itself (see survey by [Aus11]). Starting with [RR83] there is a rich body of work focused on estimating propensity scores [HIR03, IR14, SKIH11].

Let us examine the role of importance weights in anomaly detection more closely. In the semi-supervised setting [CBK09], we are given samples from R and wish to identify anomalous points or regions. Our prior knowledge is encoded by a distribution P (for which we might even have a closed form). A simple approach might be to identify points that are unlikely under P as anomalies, after all, seeing an unlikely point is indeed more surprising than seeing a likely point. However this is not a good approach for defining anomalies. For example, suppose P is Gaussian, and R has far more points close to the mean than P predicts, as was the case with Mendel's experiments in genetics [Fis36]. Points in the tail of the distribution P would be considered anomalies, even though the number of such points is lower than expected, while the excess density near the mean would be missed. In other words, we do expect low-probability points to appear in our sample of R and they should only be considered anomalous if they appear in R with higher probability than in P. The importance weights of R under P do exactly that and are therefore more suitable as a measure of anomaly. In the example above, they correctly flag the excess density near the mean as being anomalous.

1.1 Computational Limitations

Given the dramatic impact of importance weights in so many research fields and applications, a natural question is how do we compute them, and what guarantees can one hope to achieve. We consider the setting where we only have access to R and P through random samples drawn from the distributions. Not only is this the most restrictive setting, it is the realistic assumption in most applications, for

example when we want to determine how the spending patterns of customers this month differ from twelve months ago.

Computing the true weights $w^*(x) = R(x)/P(x)$ for each point based on samples from R and P is impossible in general (we will shortly justify it under standard cryptographic assumptions). Much of the literature, in different communities, makes strong assumptions on R and P, or on the ratio function $w^*(x) = R(x)/P(x)$ which make learning the weights possible. It is not obvious that such assumptions are justified in the real-world applications that motivate these works. Even if the calculated weights are approximately correct on average, it is less obvious why they would be correct on various sub-populations. This raises concerns of algorithmic discrimination: imagine for example image processing software used by a security firm that identifies movement by individuals of a particular ethnicity as anomalous with unfairly high probabilities.

A cryptographic barrier: Barring any distributional assumptions, *can we non-trivially approximate the importance weights in polynomial time?* Under standard cryptographic assumptions, the answer is that *we cannot*: Let $R^0 \equiv P$ be the uniform distribution on $\{0,1\}^d$. Let R^1 be the output of a cryptographically secure pseudorandom generator [Gol00] on a uniform seed. Every x has importance weight 1 under R^0. On the other hand, with probability 1, an x sampled from R^1 will have importance weights exponential in d. If an algorithm only gets access to random samples from R, it cannot distinguish $R = R^0$ from $R = R^1$ and hence cannot accurately determine whether every observation is hugely anomalous or perfectly normal—despite the huge gap in the weights. We borrow the terminology of anomaly detection, but this observation is relevant much more broadly. We observe that in this example, the case $R^0 = P$ is as simple as it gets and that the example could be generalized to other distributions (as long as they have large enough entropy).

This example shows that for arbitrary distributions P and R, **accuracy for point-wise scores is impossible**. It suggests that for provable guarantees, we need to look for **set-wise accuracy guarantees**, and restrict our attention to a collection of *nice* sets, whose complexity we constrain so as to exclude bad examples such as the support of a pseudorandom generator. We will model this by a collection of sets $\mathcal{C} \subseteq 2^{\mathcal{X}}$, where $\mathcal{X}$ is the domain of the distributions R and P, which will define the statistical tests that our weights must pass. Membership in every set $C \in \mathcal{C}$ has to be easy to test (we will subsequently add some learnability assumptions). $\mathcal{C}$ may represent an algorithm's computational resources in the sense that the algorithm really only uses samples from a distribution Q to compute $Q(C) = \sum_{x \in C} Q(x)$ for $C \in \mathcal{C}$. This would correspond to a statistical query algorithm that queries sets in $\mathcal{C}$. In the fairness context, we would like $\mathcal{C}$ to capture the protected subgroups and any other sub-population for which we want guarantees on our weights.

1.2 Sandwiching Bounds

In order to formulate the set-wise accuracy guarantees we wish to achieve, let us consider the setting of anomaly detection. Consider a researcher analyzing health data, where each point represents a patient. R represents the distribution of data they actually see, whereas P represents a prior. $\mathcal{C}$ is a (possibly huge) collection of medical conditions that can be diagnosed from the data. A learning algorithm has assigned importance weights $w : \mathcal{X} \to \mathbb{R}$ to the points. We articulate two desirable properties of the weights.

1. Let $C \in \mathcal{C}$ be a medical condition such that $R(C)/P(C) > 10$, so that the prevalence of C in the real world is 10 times higher than was expected in the prior. The researcher would like a random R-sample from C to be assigned large weight by w, ideally at least 10. If not, w might not alert them to the increased prevalence of C.

2. Let $C' \in \mathcal{C}$ be a medical condition such that a random R-sample from C' is assigned an average weight of 10 by w. The researcher would like this to imply that C' is truly important under R in some precise sense. If not, having large weights w in expectation for C' is not a reliable signal of importance under R and might be a false alarm.

In analogy to proof systems, these conditions ask for **completeness** and **soundness** of the importance weights respectively. Completeness requires that if a set C is important under R, then it receives large weights w on average. Soundness requires that if the average weight under R assigned to a set C' is large, this indicates that the set is important.

We now rigorously formulate these intuitive requirements. For a distribution R and a set C, let $R|_C$ denote the distribution R conditioned on C. We would like the following **Sandwiching bounds** to hold for every $C \in \mathcal{C}$:

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})] \leq \mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w(\mathbf{x})] \leq \mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w^*(\mathbf{x})]. \tag{1}$$

The quantity in the middle is one that we can compute from random samples of R, given w. We want it to be sandwiched between the expectations of the ground-truth scores w^* under $P|_C$ and $R|_C$ (note that we don't have w^* explicitly).

Let us see why sandwiching bounds indeed capture the aforementioned requirements. For the lower bound, the expectation on the left can be written as

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})] = \frac{R(C)}{P(C)}.$$

Hence this inequality captures condition (1). For the upper bound, we want our weights to be conservative, they should not exaggerate the prevalence of anomalies within a set; so we require

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w(\mathbf{x})] \leq \mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w^*(\mathbf{x})].$$

This captures the soundness requirement in condition (2), if we were to replace the learned weights w by the ground truth w^*, the average weights would only increase.

Equation (1) implies the following outer inequality for w^*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})] \leq \mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w^*(\mathbf{x})]. \tag{2}$$

One can show this is a consequence of convexity. This is apparent from the following restatement of Equation (1) which is proved in Lemma 3.2. Intuitively, this says that we want the weights w to have the right correlation with the true weights w^* under $P|_C$.

$$\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})]\right)^2 \leq \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w(\mathbf{x})w^*(\mathbf{x})] \leq \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}\left[w^*(\mathbf{x})^2\right]. \tag{3}$$

Having proposed a notion of set-wise accuracy for importance weights, we now ask how one can find weights that achieve these bounds. We first examine a notion called multi-accuracy, which is implicit in the MaxEnt algorithm for learning importance weights [DPDPL97, DPS04, DPS07]. We will show that this falls short of our goal. We will then propose a stronger notion called multi-calibration, and show that this indeed gives weights that satisfy the Sandwiching bounds.

1.3 Multi-accuracy and the MaxEnt Algorithm

Given importance weights $w(x)$ that define a candidate distribution $Q(x) = w(x)P(x)$, a natural property to require is that each set $C \in \mathcal{C}$ is given similar weight by Q as under R. We say that our weights are **multi-accurate** for $\mathcal{C}$ if $R(C) \approx Q(C)$ for every set $C \in \mathcal{C}$. If the weights $w(x)$ are multi-accurate, then Q is indistinguishable from R by $\mathcal{C}$. To distinguishers that can only perform these tests, the weight function w is just as plausible as the true weight function w^*.

While the term multi-accuracy is new in this context, this notion is implicit in the MaxEntropy approach to distribution learning, especially the elegant work of [DPDPL97, DPS04, DPS07]. The set of multi-accurate distributions forms a polytope. The MaxEnt algorithm finds the distribution Q in

this polytope which minimizes the KL divergence between Q and P, using a (weak agnostic) learning algorithm for $\mathcal{C}$. Intuitively, minimizing $D(Q\|P)$ gives the multi-accurate distribution Q that most closely matches the prior P.

We can reinterpret multi-accuracy as saying the importance weights are correct on average for every set $C \in \mathcal{C}$ under $P|_C$ (rather than under $R|_C$). More precisely, multi-accuracy implies

$$\frac{Q(C)}{P(C)} = \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w(x)] \approx \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(x)] = \frac{R(C)}{P(C)} \tag{4}$$

which is why we call it multi-accuracy in analogy to [HKRR18, KGZ19]. This seems similar in form to the lower sandwiching bound. Yet perhaps surprisingly, multi-accuracy is not strong enough to guarantee sandwiching. Intuitively, while multi-accuracy guarantees that $w(x)$ and $w^*(x)$ have similar expectations under $P|_C$, it does not guarantee either of the required bounds on the expectation of $w(x)$ *under* $R|_C$. This is because multi-accuracy does not constrain the distribution of importance weights $w(x)$ within C. However Equation (3) suggests that sandwiching requires good correlation between w and w^* under $P|_C$.

Building on this intuition, we construct examples of multi-accurate weights where the desired inequalities in the sandwiching bound can be violated by an arbitrary multiplicative factor. More precisely, for any $B > 1$ we show instances where MaxEnt returns importance weights w^{ME} such that $\mathbb{E}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})] \geq B\,\mathbb{E}_{\mathbf{x}\sim R|_C}[w^{\mathrm{ME}}(\mathbf{x})]$ (whereas sandwiching requires $\mathbb{E}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})] \leq \mathbb{E}_{\mathbf{x}\sim R|_C}[w(\mathbf{x})]$). We show separate examples exhibiting a similar violation of the upper bound (the examples are separate since by Equation (2) in every case at least one of the bounds holds). Furthermore, these violations apply to the solution found by MaxEnt (rather than to a contrived distribution that happens to be multi-accurate).

1.4 Partitions and Multicalibration

We introduce the notion of multi-calibration for importance weights, adapting ideas from recent work on multi-group fairness in the supervised setting [HKRR18] to the unsupervised setting. Multi-calibration is a strengthening of Multi-accuracy. We show that multi-calibrated importance weights guarantee sandwiching bounds.

Partitions: Our first major contribution is to shift from thinking about distributions and importance weights to thinking about partitions of the domain $\mathcal{X}$. A partition $\mathcal{S}$ of $\mathcal{X}$ is a collection of disjoint subsets $\{S_i\}_{i=1}^m$ whose union is $\mathcal{X}$. To each point $x \in S$, we assign the importance weight $w(x) = w(S) = R(S)/P(S)$. This naturally defines a distribution Q where

$$Q(x) = w(x)P(x) = \frac{R(S)}{P(S)}P(x) = R(S)P(x|S).$$

This lets us think of Q as a hybrid distribution, where we first sample a state in the partition using R, and a point $x \in S$ using $P|_S$. Using the trivial partition of a single set $S_1 = \mathcal{X}$ gives $Q = P$, and at the other extreme using the partition consisting of all singletons gives $Q = R$. While it might seem counterintuitive to restrict the family of importance weights while seeking stronger guarantees, it serves to highlight the key challenge in assigning importance weights: finding regions of $\mathcal{X}$ that should receive similar importance weights under R. Once we partition space into these regions, the choice of importance weights $R(S)/P(S)$ is natural.

Multi-calibration for partitions: Now given a set of tests $\mathcal{C}$, we ask that the importance weights resulting from our partition be calibrated for every $C \in \mathcal{C}$. While calibration is usually defined for $[0, 1]$ valued variables, in our setting it means the following. Consider the set $C \cap S_i$. Every point in this set is assigned the importance weight $w(S_i) = R(S_i)/P(S_i)$. Multi-calibration requires that these weights are indeed correct on average for this set, namely that:

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_{S_i\cap C}}[w^*(\mathbf{x})] = \frac{R(C \cap S_i)}{P(C \cap S_i)} \approx \frac{R(S_i)}{P(S_i)}. \tag{5}$$

The power of this definition comes from requiring this condition hold for every $C \in \mathcal{C}$ and $S_i \in \mathcal{S}$.

Before going further, let us see why this is a desirable notion in its own right from a fairness perspective as well as for correctness in a heterogeneous population. Returning to the example of researcher analyzing health data, suppose the weights w arise from a multi-calibrated partition. Let $S_i \in \mathcal{S}$ be such that $w(S_i) = R(S_i)/P(S_i) = 10$. For any condition C, $C \cap S_i$ is the subset of C that is assigned the weight 10. By Equation (5) Multi-calibration guarantees that average of w^* over these points is close to 10, thus the weights are justified. Intuitively, this ensures that w is not just correct in expectation over C, but that condtioned on C, it is well correlated with w^*.

Our main technical contribution is showing that the importance weights associated with a multi-calibration partition do indeed satisfy the Sandwiching bounds. This raises the question of efficient computation. We first define a generalization of multi-calibration which can be computed sample efficiently, and which preserves the desirable properties of multi-calibration. We give an algorithm for computing such a multicalibrated partition, assuming access to a weak agnostic learner for the class $\mathcal{C}$. The number of states in the output partition is independent of $\mathcal{C}$. Our algorithm is inspired by the boosting algorithm of Mansour and MacAllester [MM02]. Thus under the same learnability assumptions underlying the MaxEnt algorithm, we get a stronger guarantee.

1.5 Summary of Our Contributions

We consider the problem of computing importance weights for a distribution R with respect to a prior P, given sample access to both distributions.

1. Based on standard cryptographic assumptions, we argue that point-wise accuracy for importance weights is not possible without making strong assumptions on P and R. We propose requiring set-wise accuracy guarantees for a class $\mathcal{C}$ of sets that represent statistical tests. We formulate Sandwiching bounds as a notion of set-wise accuracy for importance weights, and show that they capture natural *completeness* and *soundness* requirements.

2. We show that the notion of multi-accuracy inherent in the MaxEnt algorithm does not guarantee Sandwiching bounds, by constructing explicit examples where the bounds are violated.

3. We introduce the notion of multi-calibration for partitions, inspired by recent work in supervised learning [HKRR18]. We show that the importance weights resulting from such partitions do guarantee sandwiching bounds.

4. We define a generalization of multi-calibration that can be computed in a sample-efficient manner. We present an efficient algorithm for constructing such multi-calibrated partitions.

Outline of this Paper

In Section 2, we formally define our notions of multi-accuracy and multi-calibration for partitions. In Section 3, we show that multi-calibration gives importance weights that satisfy the sandwiching bounds. We then turn to efficient computation. The original definition of multi-calibration can require extremely high sample complexity. In Section 4 we give a relaxed definition of multi-calibration that can be achieved with a limited number of samples. In Section 5, we provide an efficient algorithm to find a partition that meets the relaxed definition of multi-calibration. We discuss more related work in Section 6. In Appendix A, we exhibit instances where the multi-accurate distributions found by MaxEnt violate the sandwiching bound, up to arbitrary multiplicative factors. We defer additional proofs to Appendix B and C.

2 Multi-accurracy and Multi-calibration

We use $[m] = \{1, \ldots, m\}$. We use capitals $(P, Q, R, \cdots)$ to denote distributions and boldface $\mathbf{x}, \mathbf{y}, \cdots$ to denote random variables. We use $\mathbf{x} \sim P$ to denote sampling the variable $\mathbf{x}$ according to distribution P,

and $\mathbf{x} \sim P|_A$ to denote sampling from P conditioned on an event $A \subseteq \mathcal{X}$, where $P(x|A) = P(x)/P(A)$ for each $x \in A$. For two distributions P, Q over $\mathcal{X}$ let $\mathrm{d_{TV}}(P,Q) = \sum_{x\in\mathcal{X}} |P(x) - Q(x)|$. To every distribution $Q(x)$, one can associate a function $w(x) = Q(x)/P(x)$, which we call the importance weights of Q relative to P. For any $A \subseteq \mathcal{X}$ observe that

$$Q(A) = \sum_{x\in A} Q(x) = \sum_{x\in A} w(x)P(x) = P(A) \sum_{x\in A} w(x)P(x|A) = P(A) \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_A}[w(x)]. \tag{6}$$

In our setting, we have a *target* distribution R and a *prior* distribution P, and denote $w^*(x) = R(x)/P(x)$. Our aim is to find importance weights $w(x)$ such that $Q(x) = w(x)P(x)$ is close to R. We will consider a family of sets $\mathcal{C}$ which could be thought of for instance as decision trees or neural nets of a given depth. The indicator functions of sets $C \in \mathcal{C}$ can be viewed as statistical tests. We will assume there is an efficient algorithm to compute these functions.

2.1 Multi-accuracy and MaxEnt

In this section, we formally define the notion of multiaccuracy, and prove that it does not guarantee the Sandwiching bounds.

Definition 2.1. *Let $\alpha > 0$, let $\mathcal{C} \subseteq 2^{\mathcal{X}}$ be a collection of sets. An importance weight function $w : \mathcal{X} \to \mathbb{R}$ is α-multi-accurate in expectation (α-multiAE) for $(P, R, \mathcal{C})$ if for every $C \in \mathcal{C}$, the distribution $Q(x) = w(x)P(x)$ satisfies*

$$|Q(C) - R(C)| \leq \alpha. \tag{7}$$

As such, the definition of multi-accuracy only requires indistinguishability from R for tests in $\mathcal{C}$. We might also want Q to be close to P under some divergence, this is equivalent to minimizing a regularizer term in the weights [DPS07]. The MaxEnt algorithm of Dudik, Phillips and Schapire [DPS04, DPS07] minimizes $D(Q\|P)$ which amounts to using $w\log(w)$ as the regularizer. In the case where P is uniform, this minimization is equivalent to maximizing the entropy of Q, hence the name MaxEnt. Let w^{ME} denote the α-multi-accurate importance weight function so that the corresponding distribution Q^{ME} minimizes $D(Q\|P)$. Then w^{ME} is the optimal solution to a convex optimization problem that has an efficient algorithm [DPS04]. The following theorem, proved in Appendix A, shows that the MaxEnt algorithm does not guarantee Sandwiching bounds.

Theorem 2.2. *For any constant $B > 1$, there exist distributions P, R on $\{0,1\}^n$, a collections of sets $\mathcal{C}$ and $C \in \mathcal{C}$ such that the MaxEnt algorithm run on $(P, R, \mathcal{C})$ with $\alpha = 0$ returns a distribution Q^{ME} with importance weights w^{ME} such that*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})] > B \mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w^{\mathrm{ME}}(\mathbf{x})].$$

Similarly, for any constant $B > 1$, there also exist distributions P, R on $\{0,1\}^n$, such that the MaxEnt algorithm run on $(P, R, \mathcal{C})$ with $\alpha = 0$ finds importance weights w^{ME} such that

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w^{\mathrm{ME}}(\mathbf{x})] > B \mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w^*(\mathbf{x})].$$

2.2 Partitions and α-multi-calibration

A collection of disjoint subsets $\mathcal{S} = \{S_i\}_{i=1}^m$ such that $\cup_i S_i = \mathcal{X}$ is called a partition of $\mathcal{X}$ of size m. For each $x \in \mathcal{X}$, there exists a unique $S \in \mathcal{S}$ containing it. The family of distributions Q we consider are obtained by fixing a partition $\mathcal{S}$ of $\mathcal{X}$ and then reweighing each $S \in \mathcal{S}$ so that its weight matches R. Within S, we retain the marginal distribution $P|_S$. We define this formally below:

Definition 2.3. *Given a prior distribution P and a target distribution R over $\mathcal{X}$, and a partition $\mathcal{S}$ of $\mathcal{X}$, the $(P, R, \mathcal{S})$-reweighted distribution Q over $\mathcal{X}$ is given by*

$$Q(x) = R(S)P(x|S) \text{ for } S \in \mathcal{S} \text{ s.t. } x \in S. \tag{8}$$

Equivalently, the importance weights of Q relative to P are given by

$$w(x) = \frac{R(S)}{P(S)} \text{ for } S \in \mathcal{S} \text{ s.t. } x \in S. \tag{9}$$

Since w is constant within each $S \in \mathcal{S}$, we can view it as a weight function $w : \mathcal{S} \to \mathbb{R}$. When P, R and $\mathcal{S}$ are clear from context, we simply refer to Q as the reweighted distribution. We have the following expressions for any $A \subseteq \mathcal{X}$ and $S \in \mathcal{S}$,

$$Q(A \cap S) = \sum_{x \in A \cap S} R(S)P(x|S) = R(S)P(A|S), \tag{10}$$

$$Q(A) = \sum_{S \in \mathcal{S}} Q(A \cap S) = \sum_{S \in \mathcal{S}} R(S)P(A|S). \tag{11}$$

One can sample from Q, given random samples from R and P. We first select $S \in \mathcal{S}$ by drawing a single sample from R. We then sample $\mathbf{x}_i \sim P$ until $\mathbf{x}_i \in S$, and return that as our sample from Q. Observe that one can interpolate between P and R by making the partition finer. If the partition only consists of the single set $\mathcal{X}$, then $Q = P$. If $\mathcal{X}$ is discrete and we take $\mathcal{S}$ to be the collection of singleton sets, then $Q = R$.

Our goal in choosing the partition $\mathcal{S}$ will be to ensure that the reweighted distribution should be accurate for a set of statistical tests $\mathcal{C}$, while keeping the number of states small. Note that the class $\mathcal{C}$ of tests could be large, possibly inifinite (say all halfspaces or neural nets). Our hope is that reweighting a small-sized partition $\mathcal{S}$ will be sufficient to get accuracy for a large family of tests. We formalize this with the notion of α-multi-calibration.

Definition 2.4. *(α-multi-calibration) Let $\alpha > 0$, let $\mathcal{C} \subseteq 2^{\mathcal{X}}$ be a collection of sets. A partition $\mathcal{S}$ of $\mathcal{X}$ is α-multi-calibrated for $(P, R, \mathcal{C})$ if for every $C \in \mathcal{C}$ and $S \in \mathcal{S}$, the $(P, R, \mathcal{S})$-reweighted distribution Q satisfies*

$$\Big|Q(C \cap S) - R(C \cap S)\Big| \leq \alpha R(S). \tag{12}$$

Equivalently, $\mathcal{S}$ is α-multi-calibrated for $(P, R, \mathcal{C})$ if for every $C \in \mathcal{C}$ and $S \in \mathcal{S}$, it holds that

$$\Big|P(C|S) - R(C|S)\Big| \leq \alpha. \tag{13}$$

Let us show that the two formulations are indeed equivalent. By Equation (10) we have $Q(C \cap S) = R(S)P(C|S)$. Since $R(C \cap S) = R(S)R(C|S)$, we substitute these in Equation (12) and divide by $R(S)$ to derive Equation (13). Some observations about the definition:

- **α-multi-calibration implies α-multi-accuracy**. Multi-calibration for $\mathcal{S}$ implies multi-accuracy in expectation for the distribution Q; this follows easily by summing Equation (12) over all $S \in \mathcal{S}$. Note that we have not used an explicit regularizer to ensure that Q is close to P. This condition is instead enforced by keeping the number of states m small, and only allowing the natural weights $w(S) = R(S)/P(S)$.
- **Multi-calibration is symmetric in the distributions.** If the partition $\mathcal{S}$ is α-multi-calibrated for $(P, R, \mathcal{C})$, it is also α-multi-calibrated for $(R, P, \mathcal{C})$. This is clear from Equation (13) which is symmetric in the two distributions. But note that the $(P, R, \mathcal{S})$ reweighted distribution Q and the $(R, P, \mathcal{S})$- reweighted distribution Q' that Equation (12) refers to are different.

The following technical lemma will be used in proving Sandwiching. We think of the importance weights w of Q and w^* of R relative to P as random variables under the distribution P and compare their conditional expectations for each $C \in \mathcal{C}$. Readers familiar with the definitions of multi-accuracy and multi-calibration to the corresponding notions in the supervised setting from [HKRR18] will notice the similarity. The proof is in Appendix B.

Lemma 2.5. *The weight function $w : \mathcal{X} \to \mathbb{R}$ is α-multiAE for $(P, R, \mathcal{C})$ iff for every $C \in \mathcal{C}$,*

$$\left| \mathop{\mathbb{E}}_{\mathbf{x} \sim P|_C}[w(x)] - \mathop{\mathbb{E}}_{\mathbf{x} \sim P|_C}[w^*(x)] \right| \leq \frac{\alpha}{P(C)}. \tag{14}$$

The partition $\mathcal{S}$ is α-multi-calibrated for $(P, R, \mathcal{C})$ iff for every $C \in \mathcal{C}$ and $S \in \mathcal{S}$,

$$\left| w(S) - \mathop{\mathbb{E}}_{\mathbf{x} \sim P|_{C \cap S}}[w^*(x)] \right| \leq \frac{\alpha R(S)}{P(C \cap S)}. \tag{15}$$

3 Multi-calibration implies Sandwiching Bounds

In this section, we assume that weight function w comes from a partition $\mathcal{S}$ that is α-multi-calibration for $(P, R, \mathcal{S})$. For the weight function w, and $k \geq 1$ define the quantity

$$\|w\|_k = \left(\mathop{\mathbb{E}}_{\mathbf{x} \sim P}[w(\mathbf{x})^k] \right)^{1/k} = \left(\sum_{S \in \mathcal{S}} \frac{R(S)^k}{P(S)^{k-1}} \right)^{1/k}.$$

Observe $\|w\|_1 = 1$ and $\|w\|_k$ increases with k. The main result of this Section is the following:

Theorem 3.1. *If the partition $\mathcal{S}$ is α-multi-calibrated for $(P, R, \mathcal{C})$ and $w : \mathcal{X} \to \mathbb{R}$ is the corresponding importance weight function, then*

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim P|_C}[w^*(\mathbf{x})] - 2\alpha \frac{\|w\|_2^2}{R(C)} \leq \mathop{\mathbb{E}}_{\mathbf{x} \sim R|_c}[w(\mathbf{x})] \leq \mathop{\mathbb{E}}_{\mathbf{x} \sim R|_C}[w^*(\mathbf{x})] + 3\alpha \frac{\|w\|_2^2}{R(C)} \tag{16}$$

As a first step, we prove the equivalence of the two formulations of the sandwiching bounds that were presented in the introduction.

Lemma 3.2. *Equations (1) and (3) are equivalent.*

Proof. One can rewrite the expectations under $R|_C$ in Equation (1) in terms of expectations under $P|_C$ as follows

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim R|_C}[w(\mathbf{x})] = \frac{\sum_{x \in C} p(x) w(x) w^*(x)}{\sum_{x \in C} p(x) w^*(x)} = \frac{\mathbb{E}_{\mathbf{x} \sim P|_C}[w(\mathbf{x}) w^*(\mathbf{x})]}{\mathbb{E}_{\mathbf{x} \sim P|_C}[w^*(x)]} \tag{17}$$

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim R|_C}[w^*(\mathbf{x})] = \frac{\sum_{x \in C} p(x) w^*(x)^2}{\sum_{x \in C} p(x) w^*(x)} = \frac{\mathbb{E}_{\mathbf{x} \sim P|_C}[w^*(\mathbf{x})^2]}{\mathbb{E}_{\mathbf{x} \sim P|_C}[w^*(x)]}. \tag{18}$$

Plugging these into Equation (1), we can rewrite those inequalities as

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim P|_C}[w^*(\mathbf{x})] \leq \frac{\mathbb{E}_{\mathbf{x} \sim P|_C}[w(\mathbf{x}) w^*(\mathbf{x})]}{\mathbb{E}_{\mathbf{x} \sim P|_C}[w^*(x)]} \leq \frac{\mathbb{E}_{\mathbf{x} \sim P|_C}[w^*(\mathbf{x})^2]}{\mathbb{E}_{\mathbf{x} \sim P|_C}[w^*(x)]}$$

which is equivalent to Equation (3). ∎

We now proceed with the proof. Using the formulation in Equation (3), we will analyze $\mathbb{E}_{\mathbf{x} \sim P|_C}[w(\mathbf{x}) w^*(\mathbf{x})]$. The key steps are the next two technical lemmas whose proofs are in the appendix.

Lemma 3.3. *We have*

$$\left| \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w(\mathbf{x})w^*(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}[w(\mathbf{S})^2] \right| \le \frac{\alpha \|w\|_2^2}{P(C)}. \tag{19}$$

Lemma 3.4. *We have*

$$\left(\mathop{\mathbb{E}}_{x\sim P|_C}[w^*(\mathbf{x})] \right)^2 - 2\alpha \frac{R(C)}{P(C)} \le \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}[w(\mathbf{S})^2] \le \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})^2] + 2\alpha \frac{\|w\|_2^2}{P(C)} \tag{20}$$

We now put these together to prove Theorem 3.1.

Proof of Theorem 3.1. We claim the following inequalities hold

$$\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(x)] \right)^2 - \alpha \frac{\|w\|_2^2 + R(C)}{P(C)} \le \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w(\mathbf{x})w^*(\mathbf{x})] \le \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(x)^2] + 3\alpha \frac{\|w\|_2^2}{P(C)}. \tag{21}$$

These are an immediate consequence of Lemma 3.3 showing that $\mathbb{E}_{\mathbf{x}\sim P|_C}[w(\mathbf{x})w^*(\mathbf{x})]$ and $\mathbb{E}_{\mathbf{S}\sim P|_C}[w(\mathbf{S})^2]$ are close, and Lemma 3.4 which gives a sandwiching bound for $\mathbb{E}_{\mathbf{S}\sim P|_C}[w(\mathbf{S})^2]$.

Equation (21) equivalent to Equation (16). To see this, we use the following equalities from Equation (17) and (18):

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w(\mathbf{x})w^*(\mathbf{x})] = \mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w(\mathbf{x})] \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})],$$
$$\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})^2] = \mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w^*(\mathbf{x})] \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})].$$

We plug these into Equation (21) and divide throughout by $\mathbb{E}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})] = R(C)/P(C)$ to derive Equation (16) and complete the proof. □

4 (α, β)-multi-calibration

Multi-calibration requires the closeness of $R(C|S)$ and $P(C|S)$. However it might be the case that one of P and R assign very little probability to some set S. Since in our model, we only get random samples from P and R, enforcing the condition required for multi-calibration might be very expensive in terms of sample complexity, which might depend polynomially on $\max_x(R(x)/P(x), P(x)/R(x))$. This motivates a relaxation that we call (α, β)-multi-calibration.

Definition 4.1. *Let $\alpha \ge 0, \beta \ge 0$, let $\mathcal{C} \subseteq 2^{\mathcal{X}}$ be a collection of sets. The partition $\mathcal{S} = \{S_1, \ldots, S_m, T_0, T_1\}$ is (α, β)-multi-calibrated for $\mathcal{C}$ under P, R if (1) $P(T_0) \le \min(\beta, R(T_0))$, (2) $R(T_1) \le \min(\beta, P(T_1))$, (3) for every $C \in \mathcal{C}$ and $i \in [m]$,*

$$\left| Q(C \cap S_i) - R(C \cap S_i) \right| \le \alpha R(S_i). \tag{22}$$

(α, β)-multi-calibration permits two *exceptional* subsets T_0 and T_1 that do not satisfy Equation (12), but these subsets must have small measure under P and R respectively. We will use $\mathcal{T} = \{T_0, T_1\}$ to denote the exceptional sets. Intuitively, we think of T_0 as a region that has small measure under P, although $R(T_0)$ could be large. It is hard to ensure that Equation (22) is met for T_0 since samples from P seldom lie in it. Similarly, we allow a region T_1 where R allocates low probability, but $P(T_1)$ could be large. The advantage of allowing for $\mathcal{T}$ is that for every $S \in \mathcal{S} \setminus \mathcal{T}$, we *can assume* that $w(S) = R(S)/P(S)$ is bounded by $O(1/\beta)$. This intuition is formalized in the following lemma.

Lemma 4.2. *Let $T \subseteq \mathcal{X}$ and $c \ge 1$ be such that $R(T)/P(T) \ge c/\beta$. Then $P(T) \le \beta/c$.*

Proof. Since $R(T)/P(T) \ge c/\beta$, we have $P(T) \le \beta R(T)/c \le \beta/c$. □

When $\beta = 0$ we recover the notion of α-multi-calibration. Analogously when $\beta > 0$, we show that the distributions R and P are β-close in statistical distance to distributions R^h and P^h respectively that are indeed α-multicalibrated.

Definition 4.3. *Define the distribution P^h which is identical to P on $\mathcal{S} \setminus \{T_0\}$. Let $P^h(T_0) = P(T_0)$, and $P^h|_{T_0} = R|_{T_0}$. Similarly, define R^h to be identical to R on $\mathcal{S} \setminus \{T_1\}$. Let $R^h(T_1) = R(T_1)$, and $R^h|_{T_1} = P|_{T_1}$.*

Lemma 4.4. *If the partition $\mathcal{S}$ is (α, β)-multi-calibrated for $(P, R, \mathcal{C})$, then*

- $\mathrm{d_{TV}}(P^h, P) \leq \beta$, $\mathrm{d_{TV}}(R^h, R) \leq \beta$.
- *The partition $\mathcal{S}$ is α-multi-calibrated for $(P^h, R^h, \mathcal{C})$.*

Proof. The statistical distance bounds hold since P^h and P only differ on T_0 and $P^h(T_0) = P(T_0) \leq \beta$. We verify that the partition is multi-calibrated by showing that $|P(C|S) - R(C|S)| \leq \alpha$ for every state $S \in \mathcal{S}$. For any $i \in [m]$, we have

$$\left|R^h(C|S_i) - P^h(C|S_i)\right| = \left|R(C|S_i) - P(C|S_i)\right| \leq \alpha.$$

where the equality holds since since P^h and P (and R^h and R) are identical on the states S_i for $i \in [m]$ and the inequality is from Equation (22). The conditional distributions $P^h|_{T_0}$ and $R^h|_{T_0}$ are identical since they both equal $R|_{T_0}$ by construction. Hence $R^h(C|T_0) = P^h(C|T_0)$ for all $C \in \mathcal{C}$, so the condition holds. A similar argument holds for T_1. □

A corollary is that (α, β)-multi-calibration implies $(\alpha + 2\beta)$-multi-accuracy.

Lemma 4.5. *If $\mathcal{S}$ is (α, β)-multi-calibrated for $(P, R, \mathcal{C})$, then the $(P, R, \mathcal{S})$-reweighted distribution Q is $\gamma = (\alpha + 2\beta)$-multi-accurate for $(P, R, \mathcal{C})$.*

Finally, we can show a sandwiching bound for (α, β)-multi-calibration. The error terms now also depend on β and $\|w\|_4^2$ in comparison to Theorem 3.1.

Theorem 4.6. *Assume the partition $\mathcal{S}$ is (α, β)-multi-calibrated for $(P, R, \mathcal{C})$ and $w : \mathcal{X} \to \mathbb{R}$ is the corresponding importance weight function. Let*

$$\ell(\alpha, \beta, w) = \alpha \|w\|_2^2 + \sqrt{\beta}\, \|w\|_4^2 . \tag{23}$$

Then for every $C \in \mathcal{C}$,

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim P|_C}[w^*(\mathbf{x})] - \frac{2\ell(\alpha, \beta, w)}{R(C)} - \frac{2(\alpha + 2\beta)}{P(C)} \leq \mathop{\mathbb{E}}_{\mathbf{x} \sim R|_c}[w(\mathbf{x})] \leq \mathop{\mathbb{E}}_{\mathbf{x} \sim R|_C}[w^*(\mathbf{x})] + \frac{3\ell(\alpha, \beta, w)}{R(C)}. \tag{24}$$

The proof appears in Appendix C.

5 Algorithm for (α, β)-multi-calibration

In this section, we give an efficient algorithm that computes a multicalibrated partition, given access to a weak agnostic learner for $\mathcal{C}$. The algorithm is reminiscent of the algorithm for Boosting via Branching Programs due to [MM02]. We first define weak agnostic learning. Given a collection of sets $\mathcal{C} \subseteq 2^{\mathcal{X}}$, we can associate every set C with its indicator function $c : \mathcal{X} \to \{0, 1\}$.

Definition 5.1. *A (α, α', L)-weak agnostic learning algorithm for a class $\mathcal{C}$ is given L samples from a distribution $\mathcal{D} = (\mathbf{x}, \mathbf{y})$ where $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \{0, 1\}$. If there exists $c \in \mathcal{C}$ such that $\Pr_{\mathcal{D}}[c(\mathbf{x}) = \mathbf{y}] \geq (1+\alpha)/2$, then the learner will return $c' \in \mathcal{C}'$ such that $\Pr_{\mathcal{D}}[c'(\mathbf{x}) = \mathbf{y}] \geq (1+\alpha')/2$ for some $0 < \alpha' \leq \alpha$. The class $\mathcal{C}$ is said to be (α, α', L)-weakly agnostically learnable if such a leaner exists.*

We allow $\alpha - \alpha'$ to depend on L, typically it decreases with L. For simplicity, we will assume that $\mathcal{C}' = \mathcal{C}$, and do not allow for probability of error. Given two distributions P and R, a weak learner for $\mathcal{C}$ can be used to find $C \in \mathcal{C}$ such that $|R(C) - P(C)|$ is large, a view that we will use hereafter.

Lemma 5.2. *Let $\mathcal{A}$ be an (α, α', L)-weak agnostic learner for $\mathcal{C}$. Given distributions P and R, if there exists $C \in \mathcal{C}$ so that $|R(C) - P(C)| \geq \alpha$, given $O(L)$ samples from each of R and P, $\mathcal{A}$ can be used to find $C' \in \mathcal{C}$ such that $|R(C') - P(C')| \geq \alpha'$.*

Proof. Assume that $R(C) > P(C)$. Define a distribution $\mathcal{D}$ where we output $(\mathbf{x} \sim P, 0)$ or $(\mathbf{x} \sim R, 1)$ each with probability $1/2$. For any $c \in \mathcal{C}$ we have

$$\Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[c(\mathbf{x}) = y] = \frac{1}{2}\left(\Pr_{\mathbf{x}\sim P}[c(x) = 0] + \Pr_{\mathbf{x}\sim R}[c(\mathbf{x}) = 1]\right) = \frac{1 - P(C) + R(C)}{2} = \frac{1}{2} + \frac{R(C) - P(C)}{2}.$$

If there exists $C \in \mathcal{C}$ such that $R(C) - P(C) \geq \alpha$, $\mathcal{A}$ run on $\mathcal{D}$ will return $C' \in \mathcal{C}$ such that $R(C') - P(C') \geq \alpha'$. To generate L samples from $\mathcal{D}$, it suffices to have $O(L)$ samples from both R and P. □

Our algorithm for achieving (α, β)-multi-calibration uses the weak learners to find distinguishing sets. We next describe our algorithm.

5.1 Algorithm for Multi-Calibration

Given sample access to distributions P, R, we will construct a multicalibrated partition by starting from the trivial partition and iteratively modifying it till we achieve multi-calibration. We assume access to a weak agnostic learner for the class $\mathcal{C}$.

We use $(\mathcal{S}^t, \mathcal{T}_0, \mathcal{T}_1)$ to denote the t^{th} partition, and Q^t to denote the corresponding reweighted distribution. The partition consists of three groups of sets:

- **Large weights:** $\mathcal{T}_0$ consisting of sets T such that $R(T)/P(T) \geq 2/\beta$.
- **Small weights:** $\mathcal{T}_1$ consisting of sets T such that $R(T)/P(T) \leq \beta/2$.
- **Medium weights:** $\mathcal{S}^t$ will consists of sets S such that $R(S)/P(S) \in [\beta/2, 2/\beta]$.

The collections $\mathcal{T}^0, \mathcal{T}^1$ both start empty and grow monotonically. Once set T is added to either, that set is not modified. All sets in $\mathcal{T}_0$ will eventually be merged into a single set T_0 such that $P(T_0) \leq \beta$, while the sets in $\mathcal{T}_1$ will be merged into a single set $\mathcal{T}_1$. Doing the merging at the end simplifies the analysis, but for intuition it is fine to think of each as a single state that keeps growing.

Our algorithm will mostly focus on the medium sets in $\mathcal{S}^t$, although occasionally sets will be added to $\mathcal{T}_0$ or $\mathcal{T}_1$, hence we use the superscript t to account for how it changes over iterations. The algorithm combines two basic operations.

- Split: This operation takes $S \in \mathcal{S}_t$ where $R(S), P(S)$ are sufficiently large, and $C \in \mathcal{C}$ such that $|P(C|S) - R(C|S)| > \alpha'$, and split S into two states, $C \cap S$ and $\bar{C} \cap S$. We find the pair S, C by running the weak agnostic learner to distinguish the distributions $P|_S$ and $R|_S$. The new sets are classified as small, medium or large.
- Merge: This operation is applied to $\mathcal{S}^t$ when the number states in it goes beyond a certain bound. It merges those states in $\mathcal{S}^t$ with similar importance weights into a single state, and halves the number of states.

We first state and analyze each of these operations. We view both operations as updating the current partition.

Note that the states S_0 and S_1 might end up in any of $\mathcal{T}_0, \mathcal{T}_1$ or $\mathcal{S}^{t+1}$ depending on the value of $R(S_i)/P(S_i)$. We will analyze the Split operation using $D\left(Q_t \| P\right)$ as the potential function.

Algorithm 1 Split(S, C)

Input:

- $S \in \mathcal{S}_t$ s.t. $R(S) \geq \beta/4m$, $P(S) \geq \beta/4m$.
- $C \in \mathcal{C}$ s.t. $|R(C|S) - P(C|S)| > \alpha'$.

Replace S with the two states $S_0 = S \cap C$ and $S_1 = S \cap \bar{C}$.

Lemma 5.3. *We have* $D\left(Q_{t+1} \| P\right) - D\left(Q_t \| P\right) \geq 4R(S)\alpha'^2$.

Proof. Since Q_{t+1} differs from Q_t by splitting S into $S \cap C$ and $S \cap \bar{C}$, we have

$$\begin{aligned} D\left(Q_{t+1} \| P\right) - D\left(Q_t \| P\right) &= R(S \cap C) \log\left(\frac{R(S \cap C)}{P(S \cap C)}\right) + R(S \cap \bar{C}) \log\left(\frac{R(S \cap \bar{C})}{P(S \cap \bar{C}C)}\right) - R(S) \log\left(\frac{R(S)}{P(S)}\right) \\ &= R(S \cap C) \log\left(\frac{R(S \cap C)P(S)}{R(S)P(S \cap C)}\right) + R(S \cap \bar{C}) \log\left(\frac{R(S \cap \bar{C})P(S)}{R(S)P(S \cap \bar{C})}\right) \\ &= R(S)\left(R(C|S) \log\left(\frac{R(C|S)}{P(C|S)}\right) + R(\bar{C}|S) \log\left(\frac{R(\bar{C}|S)}{P(\bar{C}|S)}\right)\right). \end{aligned} \tag{25}$$

The expression in braces is the KL divergence between two Bernoulli random variables that are 1 with probability $R(C|S)$ and $P(C|S)$ respectively. Hence we can apply Pinsker's inequality to get

$$R(C|S) \log\left(\frac{R(C|S)}{P(C|S)}\right) + R(\bar{C}|S) \log\left(\frac{R(\bar{C}|S)}{P(\bar{C}|S)}\right) \geq \left|R(C|S) - P(C|S)\right|^2 \geq 4\alpha'^2. \tag{26}$$

Plugging this into Equation (25) gives the desired bound. ∎

We now describe the merge operation. For parameters δ to be specified later, we divide the interval $[\beta/2, 2/\beta]$ into $O(\log(1/\beta)/\delta)$ geometric scales of $\exp(\delta)$. Merge ensures that the number of states is at most m.

Algorithm 2 Merge(δ)

Input: parameter δ.

1. Let $m = \lceil \frac{1}{\delta} \log\left(\frac{4}{\beta^2}\right) \rceil$.
2. For each $i \in \{1, \ldots, m\}$:
   Form a new state S_i by merging all states $S' \in \mathcal{S}_j$ such that
   $$\frac{R(S')}{P(S')} \in \left(\frac{e^{(i-1)\delta}\beta}{2}, \frac{e^{i\delta}\beta}{2}\right].$$
3. Let $\mathcal{S}^{t+1} = \{S_i\}_{i=1}^m$, discarding any empty states.

Unlike Split, Merge can reduce the KL divergence, but we can bound the loss.

Lemma 5.4. *We have* $D\left(Q_t \| P\right) - D\left(Q_{t+1} \| P\right) \leq \delta$.

Proof. Let $S'_1, \ldots, S'_\ell \in \mathcal{S}_t$ denote the states that are merged to form $S_i \in \mathcal{S}_{t+1}$. For each $k \in [\ell]$,

$$\frac{R(S'_k)/P(S'_k)}{R(S_i)/P(S_i)} \leq e^{\delta}.$$

We use this to bound the decrease in potential from S_i as

$$\begin{aligned}\sum_{k=1}^{\ell} R(S'_k)\log\left(\frac{R(S'_k)}{P(S'_k)}\right) - R(S_i)\log\left(\frac{R(S_i)}{P(S_i)}\right) &= \sum_{k=1}^{\ell} R(S'_k)\left(\log\left(\frac{R(S'_k)}{P(S'_k)}\right) - \log\left(\frac{R(S_i)}{P(S_i)}\right)\right)\\ &= \sum_{k=1}^{\ell} R(S'_k)\log\left(\frac{R(S'_k)/P(S'_k)}{R(S_i)/P(S_i)}\right)\\ &\leq \sum_{k=1}^{\ell} R(S'_k)\delta = R(S_i)\delta.\end{aligned}$$

The claim follows by summing over all $S_i \in \mathcal{S}_{t+1}$. □

We now state and analyze our main algorithm.

Algorithm 3 Multi-Calibrate$(P, R, \mathcal{C}, \alpha, \beta)$

Inputs:

- Parameters $\alpha, \beta > 0$.
- Distributions P, R.
- A class $\mathcal{C}$ that is (α, α', L)-weakly agnostically learnable.

Output: A partition that is (α, β)-multicalibrated for $\mathcal{C}$ under P, R.

Let $\mathcal{S}^1 = \{\mathcal{X}\}, \mathcal{T}_0^1 = \mathcal{T}_1^1 = \{\}$.
Let $\delta = \beta\alpha'^2/2$, and $m = \lceil \frac{1}{\delta}\log\left(\frac{4}{\beta^2}\right)\rceil$.
For $t \geq 1$

1. If $|\mathcal{S}_t| \geq 2m$, then run Merge(δ).
2. If the weak agnostic leaner finds $S \in \mathcal{S}^t$, $C \in \mathcal{C}$ such that

$$R(S) \geq \beta/4m, P(S) \geq \beta/4m, \ \Big|R(C|S) - P(C|S)\Big| \geq \alpha'.$$

   2.1. Run Split(S, C) and obtain S_0, S_1
   2.2. If $P(S_0) < \beta/4m$ and $P(S_0) < R(S_0)$, place S_0 in $\mathcal{T}_0$. Else, if $R(S_0) < \beta/4m$ place S_0 in $\mathcal{T}_1$.
   2.3. Repeat previous step for S_1
   2.4. Repeat the loop

   If the weak learner fails, exit the loop.

Post-Processing:

1. Move all $S \in \mathcal{S}^t$ such that $P(S) < \beta/4m$ and $P(S) \leq R(S)$ from $\mathcal{S}^t$ to T_0. Move all remaining $S \in \mathcal{S}^t$ such that $R(S) < \beta/4m$ from $\mathcal{S}^t$ to T_1.
2. Merge all $T \in \mathcal{T}_0$ into a single state T_0. Merge all $T \in \mathcal{T}_1$ into a single state T_1.
3. Return the partition $\mathcal{S} = \mathcal{S}^t \cup \{T_0\} \cup \{T_1\}$.

Theorem 5.5. *Algorithm 3 returns a partition $\mathcal{S}$ that is (α, β)-multi-calibrated for $\mathcal{C}$ under P, R.*

Proof. We first prove that $P(T_0) \leq \beta$ and $P(T_0) \leq R(T_0)$. We can write $T_0 = \cup_i T'_i \cup_j S'_j$ where the sets T'_i were added to T_0 during the loop, when they were created during a Split operation, and the sets S'_j were moved from $\mathcal{S}^t$ in the post-processing step. Then $R(T'_j)/P(T'_j) \geq 2/\beta$ for all j, hence $R(\cup_j T'_j)/P(\cup_j T'_j) \geq 2/\beta$. But by Lemma 4.2, this implies that $P(\cup_j T'_j) \leq \beta/2$. The sets S'_j are added to T_0 because $P(S'_j) \leq \beta/4m$. Since there are at most $2m$ such sets (else we would have run Merge), we have $P(\cup_j S'_j) \leq 2m\beta/4m \leq \beta/2$. Overall

$$P(T_0) \leq P(\cup_i T'_i) + P(\cup_j S'_j) \leq \beta/2 + \beta/2 = \beta.$$

Further, for every set T merged into T_0 it holds that $P(T) \leq R(T)$ and therefore $P(T_0) \leq R(T_0)$. A similar argument shows that $R(T_1) \leq \beta$ and $R(T_1) < P(T_1)$.

We need to show that every set $S \in \mathcal{S}^t$ satisfies $\|R(C|S) - P(C|S)\| \leq \alpha$ for all $C \in \mathcal{C}$. Note that S satisfies $R(S) \geq \beta/4m$ and $P(S) \geq \beta/4m$, else it would have been removed from $\mathcal{S}^t$ in the post-processing step. Hence, if it violates this condition, the weak agnostic learner would find a $C' \in \mathcal{C}$ such that $\|R(C'|S) - P(C'|S)\| \geq \alpha'$, so we would not exit the loop at the t^{th} iteration.

This shows that the partition $\mathcal{S}$ is (α, β)-multi-calibrated. □

Next we analyze the running time and sample complexity.

Theorem 5.6. *Given an (α, α', L) weak agnostic learning algorithm for $\mathcal{C}$, Algorithm 3 returns an (α, β)-multi-calibrated partition within T iterations where $T = \widetilde{O}(KL(R, P)/(\beta^2\alpha'^4))$. It makes $O(T)$ calls to the weak agnostic learner, where each call requires $\widetilde{O}(L/(\beta^2\alpha'^2))$ samples from each of R and P.*

Proof. Each iteration but the last involves one call to either Split or Merge. We bound the number of calls to Merge, denoted ℓ. Assume the merge operations happen in interations $t^1 < t^2 \cdots < t^\ell$). Every Split operation increases the number of states by 1, whereas Merge reduces it from $2m$ to a number is the range $\{1, \ldots, m\}$. Hence $2m \leq t^{k+1} - t^k \geq m$. Each Split operation acts on a set S where $R(S) \geq \beta/4m$, and by Lemma 5.3, it increase the KL divergence by $4R(S)\alpha'^2$. The Merge operation decreases it by $\delta = \alpha'^2\beta/2$. Hence we have

$$D\left(Q_{t^{k+1}} \| P\right) - D\left(Q_{t^k} \| P\right) \geq m\frac{\beta}{4m}4\alpha'^2 - \delta = \delta.$$

Thus the KL divergence between successive Merge operations increases by δ. We start with the trivial partition, so $Q^1 = P$. Since $\mathcal{S}^T$ is partition, if Q^T denotes the corresponding reweighted distribution, then $D\left(Q^T \| P\right) \leq D\left(R \| P\right)$. Hence

$$\ell\delta \leq D\left(Q^T \| P\right) - D\left(Q^1 \| P\right) \leq D\left(R \| P\right)$$

hence $\ell \leq D\left(R \| P\right)/\delta$. The total number of iterations is bounded by

$$T \leq (2m+1)\ell = O(\log(1/\beta)D\left(R \| P\right)/\delta^2) = \widetilde{O}(D\left(R \| P\right)/(\beta^2\alpha'^4).$$

For one Split iteration, we might make $O(m)$ calls to the weak learner, one per state to find the pair S, C on which to run Split. However, once we fail to find a good C for S, we do not need to try S again until the state is modified, which cannot happen before the next Merge iteration. This shows that there are at most $4m$ calls to the agnostic learner between two merge operations, $2m$ successful ones and $2m$ unsuccessful ones. Hence the number of calls to the learner is bounded by $4m\ell = O(T)$.

Finally, we address the sample complexity. We need to run the learner on the distributions $P|_S$ and $R|_S$ where $R(S), P(S) \geq \beta/4m$. If the sample complexity of the agnostic learner is L then $O(Lm/\beta) = \widetilde{O}(L/(\alpha'\beta)^2)$ samples from each of P and R will suffice to ensure that we have sufficiently many samples from $P|_S$ and $R|_S$ respectively. □

Finally we note that the partition we compute can be represented by a $\mathcal{C}$-branching program where each node is labelled by $c \in \mathcal{C}$.

6 Other Related Work

In Section 1, we discussed some of the diverse applications of importance weights which span many communities. We refer the reader to the book [SSK12a] for a more comprehensive overview of the related work, especially in the context of the machine learning literature. Here, we provide a brief overview of some of the important techniques for computing importance weights.

Owing to extensive applications, there has been considerable interest in showing theoretical guarantees for various approaches for importance weight estimation. An influential work in this regard is [NWJ10] which shows a variational characterization of KL divergence as an optimization problem over some hypothesis class $\mathcal{H}$, the solution to which also yields the importance weights. The optimization problem is convex if $\mathcal{H}$ is convex, and they establish guarantees on the rate of convergence of the estimated importance weights to the true weights. However, their analysis crucially assumes that the true importance weights lie in the hypothesis class $\mathcal{H}$—which is a strong assumption in most applications. There are also other works which show guarantees on the recovered importance weights if the true importance weights belong to simple hypothesis classes such as linear functions [KSS10, KHS09]. Under practical scenarios where the true importance weights are too complex to be modelled by any simple hypothesis class $\mathcal{H}$, these results do not provide any guarantees for the predicted importance weights to be accurate or calibrated.

To allow more expressive hypothesis classes, kernel based approaches for estimating importance weights have also been proposed, starting with Kernel Mean Matching (KMM) introduced in [HGB+07]. If the underlying kernel is universal, then under the limit of infinite data KMM provably recovers the true importance weights [CMRR08, HGB+07]. In the limit of finite data however, kernel based approaches can be interpreted as generalizations of moment-matching methods such as [Qin98] which seek to match some moments of the data [SSK12b]. In the context of our work, this implies that the kernel based approaches guarantee indistinguishability with respect to the matched moments (or more generally, the feature space induced by the kernel), which is equivalent to multi-accuracy in expectation with respect to the moments (or the feature space more generally). Also, as we discussed before, a widely studied body of work which can also be interpreted as getting multi-accuracy guarantees is the concept of maximum entropy for distribution estimation [Jay57, PDS04, DPS04, DPS07]

References

[Aus11] Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

[BBS07] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, pages 81–88, 2007.

[CBK09] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[CMM10] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Advances in neural information processing systems*, pages 442–450, 2010.

[CMRR08] Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *International conference on algorithmic learning theory*, pages 38–53. Springer, 2008.

[DPDPL97] Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. Inducing features of random fields. *IEEE transactions on pattern analysis and machine intelligence*, 19(4):380–393, 1997.

[DPS04] Miroslav Dudik, Steven J Phillips, and Robert E Schapire. Performance guarantees for regularized maximum entropy density estimation. In *International Conference on Computational Learning Theory*, pages 472–486. Springer, 2004.

[DPS07] Miroslav Dudik, Steven J Phillips, and Robert E Schapire. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *Journal of Machine Learning Research*, 8(Jun):1217–1260, 2007.

[Fis36] Ronald A Fisher. Has mendel's work been rediscovered? *Annals of science*, 1(2):115–137, 1936.

[Gol00] Oded Goldreich. Candidate one-way functions based on expander graphs. *IACR Cryptol. ePrint Arch.*, 2000:63, 2000.

[HA04] Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85–126, 2004.

[HGB+07] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608, 2007.

[HIR03] Keisuke Hirano, Guido W. Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

[HKRR18] Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1944–1953. PMLR, 2018.

[HTK+08] Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, and Takafumi Kanamori. Inlier-based outlier detection via direct density ratio estimation. In *2008 Eighth IEEE International Conference on Data Mining*, pages 223–232. IEEE, 2008.

[IR14] Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 01 2014.

[Jay57] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[KGZ19] Michael P. Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box postprocessing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

[KHS09] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection. In *Advances in neural information processing systems*, pages 809–816, 2009.

[KSS10] Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. Theoretical analysis of density ratio estimation. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 93(4):787–798, 2010.

[MM02] Yishay Mansour and David McAllester. Boosting using branching programs. *Journal of Computer and System Sciences*, 64(1):103–112, 2002.

[NWJ07] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Nonparametric estimation of the likelihood ratio and divergence functionals. In *2007 IEEE International Symposium on Information Theory*, pages 2016–2020. IEEE, 2007.

[NWJ10] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[PDS04] Steven J Phillips, Miroslav Dudik, and Robert E Schapire. A maximum entropy approach to species distribution modeling. In *Proceedings of the twenty-first international conference on Machine learning*, page 83, 2004.

[Qin98] Jing Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998.

[RR83] Paul Rosenbaum and Donald Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 04 1983.

[Shi00] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[SKIH11] Elizabeth Stuart, Gary King, Kosuke Imai, and Daniel Ho. Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42, 06 2011.

[SM05] Masashi Sugiyama and Klaus-Robert Müller. Input-dependent estimation of generalization error under covariate shift. *Statistics and Decisions-International Journal Stochastic Methods and Models*, 23(4):249–280, 2005.

[SPST+01] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[SSK12a] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[SSK12b] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, 2012.

[SSSK08] Taiji Suzuki, Masashi Sugiyama, Jun Sese, and Takafumi Kanamori. Approximating mutual information by maximum likelihood density ratio estimation. In *New challenges for feature selection in data mining and knowledge discovery*, pages 5–20, 2008.

[SST09] Alex Smola, Le Song, and Choon Hui Teo. Relative novelty detection. In *Artificial Intelligence and Statistics*, pages 536–543, 2009.

[WF16] Max Wornowizki and Roland Fried. Two-sample homogeneity tests based on divergence measures. *Computational Statistics*, 31(1):291–313, 2016.

[WKV05] Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, 51(9):3064–3074, 2005.

[WKV09] Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405, 2009.

[YSK+13] Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural computation*, 25(5):1324–1370, 2013.

[Zad04] Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, page 114, 2004.

A Gap Instances for MaxEnt

In this section we will show that the importance weights w^{ME} found by MaxEnt need not satisfy the sandwiching bounds. Indeed, for either direction of the sandwiching bound, we will show instances where the inequality is off by an arbitrarily large constant factor. Thus while one would like $\mathbb{E}_{P|_C}[w^*(x)] \leq \mathbb{E}_{R|_C}[w(x)]$, we will exhibit P, R and $\mathcal{C}$ such that the importance weights w^{ME} found by MaxEnt are such that the ratio $\mathbb{E}_{P|_C}[w^*(x)]/\,\mathbb{E}_{R|_C}[w(x)]$ is arbitrarily large, and similarly for the upper bound. Both our counterexamples work by starting with a small example on $\{0,1\}^2$ that shows some small constant gap and then tensoring to amplify the gap.

Lemma A.1. *There exist distribution P, R on $\{0,1\}^2$, a collections of sets $\mathcal{C}$ and $C \in \mathcal{C}$ such that the MaxEnt algorithm run on $(P, R, \mathcal{C})$ with $\alpha = 0$ returns a distribution Q^{ME} with importance weights w^{ME} such that*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})] > \mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w^{\mathrm{ME}}(\mathbf{x})].$$

Proof. Let P be the uniform distribution on $\{0,1\}^2$. Let R be the distribution where

$$R(00) = 0, R(01) = R(10) = 3/8, R(11) = 1/4.$$

We denote the two coordinates x_0, x_1, and let $\mathcal{C}$ consist of all subcubes of dimension 1. Hence $\mathcal{C} = \{x : x_i = a\}_{i\in\{0,1\},a\in\{0,1\}}$.

The distribution Q^{ME} for $\alpha = 0$ is the product distribution which matches the marginal distributions on each coordinate: $Q^{\mathrm{ME}}(x_0 = 1) = Q^{\mathrm{ME}}(x_1 = 1) = 5/8$ and the coordinates are independent. The multi-accuracy constraints $Q^{\mathrm{ME}}(x_0 = 1) = R(x_0 = 1)$ and $Q^{\mathrm{ME}}(x_1 = 1) = R(x_1 = 1)$ are clearly satisfied, and Q^{ME} is the maximum entropy distribution satisfying these constraints.

We can compute the following importance weights

1. $w^{\mathrm{ME}}(11) = (5/8)^2/(1/4)^2 = 25/16$ whereas $w^*(11) = 1$.
2. $w^{\mathrm{ME}}(10) = (5/8 \cdot 3/8)/(1/4)^2 = 15/16$, whereas $w^*(10) = (3/8)/(1/4) = 3/2$; ditto for 01.

For intuition as to why this is a gap example, note that this shows that while w^* assigns high weights to 01 and 10, w^{ME} assigns these points weights less than 1, and instead assigns a high weight to 11. Thus an algorithm that was labelling points with w^{ME} exceeding 1 as anomalies would report 11 as the sole anomaly, and miss both 01 and 10.

We consider the set $C = \{10, 11\} = \{x : x_0 = 1\}$. Note that $P|_C$ is uniform on $x_1 \in \{0,1\}$, whereas $R|_C(x_1 = 1) = 2/5$. Then it follows that

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w^{\mathrm{ME}}(\mathbf{x})] = 3/5 \cdot 15/16 + 2/5 \cdot 25/16 = 19/16$$
$$\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})] = 1/2 \cdot 1 + 1/2 \cdot 3/2 = 5/4$$

hence $\mathbb{E}_{\mathbf{x}\sim P|_C} w^*(\mathbf{x}) > \mathbb{E}_{\mathbf{x}\sim R|_C} w^{\mathrm{ME}}(\mathbf{x})$. ∎

Intuitively, in the example above, while Q^{ME} assigns the right weight of $5/8$ to the set C, within C the distribution of weight is misaligned with R, leading to low expected weight under $R|_C$. We now tensor this example to amplify the gap.

Theorem A.2. *For any constant $B > 1$, there exist distributions P, R on $\{0,1\}^n$, a collections of sets $\mathcal{C}$ and $C \in \mathcal{C}$ such that the MaxEnt algorithm run on $(P, R, \mathcal{C})$ with $\alpha = 0$ returns a distribution Q^{ME} with importance weights w^{ME} such that*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})] > B \mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w^{\mathrm{ME}}(\mathbf{x})].$$

Proof. We now consider the k-wise tensor of the instances constructed in Lemma A.1. The domain is $\{0,1\}^{2k}$ where the coordinates are denoted $x_0, \ldots, x_{2k-1}$. We consider the pair of distributions $P_k = (P)^k$ which is uniform on $2k$ bits, $R_k = (R)^k$ which the product of k independent copies of R on the pairs $\{x_{2i}x_{2i+1}\}_{i=1}^{k-1}$. Let $\mathcal{C}_k$ consist of all subcubes of dimension k where we restrict one co-ordinate out of x_{2i}, x_{2i+1} for $i \in \{0, \ldots, k-1\}$. One can verify that MaxEnt returns $Q_k^{\mathrm{ME}} = (Q^{\mathrm{ME}})^k$ which is just the product distribution on $\{0,1\}^{2k}$ with $\Pr[x_i = 1] = 5/8$ for every coordinate.

Let w_k^{ME} and $w_k^*(x)$ denote the importance weights of Q^{ME} and R_k with respect to P_k. A key observations is that importance weights tensor: for any $x \in \{0,1\}^{2k}$,

$$w_k^*(x) = \frac{R_k(x)}{P_k(x)} = \prod_{i=0}^{k-1} \frac{R(x_{2i}x_{2i+1})}{P(x_{2i}x_{2i+1})} = \prod_{i=0}^{k-1} w^*(x_{2i}x_{2i+1})$$

$$w_k^{\mathrm{ME}}(x) = \frac{Q_k^{\mathrm{ME}}(x)}{P_k(x)} = \prod_{i=0}^{k-1} \frac{Q^{\mathrm{ME}}(x_{2i}x_{2i+1})}{P(x_{2i}x_{2i+1})} = \prod_{i=0}^{k-1} w^{\mathrm{ME}}(x_{2i}x_{2i+1}).$$

We consider the set $C = \{x : x_{2i} = 1, i \in \{0, \ldots, k-1\}\}$. The key property of this set is that the conditional distributions $P_k|_C = (P|_C)^k$ and $R_k|_C = (R|_C)^k$ are also product distributions of the conditional distributions. Hence we have

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim R_k|_C}[w_k^{\mathrm{ME}}(\mathbf{x})] = \mathop{\mathbb{E}}_{\mathbf{x}\sim R_k|_C}\left[\prod_{i=0}^{k-1} w^{\mathrm{ME}}(\mathbf{x}_{2i}\mathbf{x}_{2i+1})\right] = \prod_{i=1}^{k-1} \mathop{\mathbb{E}}_{\mathbf{x}_{2i}\mathbf{x}_{2i+1}\sim R|_C}[w^{\mathrm{ME}}(\mathbf{x}_{2i}\mathbf{x}_{2i+1})] = (19/16)^k$$

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim P_k|_C}[w_k^*(\mathbf{x})] = \mathop{\mathbb{E}}_{\mathbf{x}\sim P_k|_C}\left[\prod_{i=0}^{k-1} w^*(\mathbf{x}_{2i}\mathbf{x}_{2i+1})\right] = \prod_{i=1}^{k-1} \mathop{\mathbb{E}}_{\mathbf{x}_{2i}\mathbf{x}_{2i+1}\sim P|_C}[w^*(\mathbf{x}_{2i}\mathbf{x}_{2i+1})] = (5/4)^k.$$

Now take k sufficiently large so that $(5/4)^k > B(19/16)^k$. ∎

We now construct a gap example for the other direction of the sandwiching bounds, where $\mathbb{E}_{R|_C}[w(x)] > \mathbb{E}_{R_C}[w^*(x)]$. Again we start with a small constant gap and amplify it by tensoring. We will only describe the construction for achieving the small constant gap, the tensoring step is identical to Theorem A.2.

Theorem A.3. *For any constant $B > 1$, there exist distributions P, R on $\{0,1\}^n$, a collections of sets $\mathcal{C}$ and $C \in \mathcal{C}$ such that the MaxEnt algorithm run on $(P, R, \mathcal{C})$ with $\alpha = 0$ returns a distribution Q^{ME} with importance weights w^{ME} such that*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w^{\mathrm{ME}}(\mathbf{x})] > B \mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w^*(\mathbf{x})].$$

Proof. As before let P be uniform on $\{0,1\}^2$. Consider the distribution R given by

$$R(00) = 2/16, R(10) = 6/16, R(01) = 3/16, R(11) = 5/16.$$

As before we let $\mathcal{C}$ consist of all subcubes of dimension 1. The distribution Q^{ME} is the product distribution on x_0 and x_1 where $\Pr[x_0 = 1] = 1/2$ and $\Pr[X_1 = 1] = 11/16$. We will use the set $C = \{01, 11\}$, so that $R|_C(01) = 3/8, R|_C(11) = 5/8$.

We compute the importance weights within C as follows:

$$w^*(01) = 3/4, w^*(11) = 5/4$$
$$w^{\mathrm{ME}}(01) = 5/8, w^{\mathrm{ME}}(11) = 11/8.$$

Hence we have the conditional expectations

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w^*(\mathbf{x})] = 3/8 \cdot 3/4 + 5/8 \cdot 5/4 = 34/32.$$
$$\mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w^{\mathrm{ME}}(\mathbf{x})] = 3/8 \cdot 5/8 + 5/8 \cdot 11/8 = 35/32$$

hence $\mathbb{E}_{\mathbf{x}\sim R|_C}[w^{\mathrm{ME}}(\mathbf{x})] > \mathbb{E}_{\mathbf{x}\sim R|_C}[w^*(\mathbf{x})]$. We can amplify this gap by tensoring. □

B Additional Proofs

Proof of Lemma 2.5. Applying Equation (6) to distributions Q and R for the set C, we get

$$Q(C) = P(C) \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w(x)], \quad R(C) = P(C) \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(x)].$$

Hence we get

$$\Big|Q(C) - R(C)\Big| = P(C)\left|w(S) - \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_{C\cap S}}[w^*(x)]\right|$$

and hence dividing both sides of Equation (7) by $P(C)$ gives Equation (14).

Applying Equation (6) to R with $A = C \cap S$, we get

$$R(C \cap S) = P(C \cap S) \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_{C\cap S}}[w^*(x)]$$

whereas by Equation (10),

$$Q(C \cap S) = R(S)P(C|S) = \frac{R(S)P(C \cap S)}{P(S)} = w(S)P(C \cap S).$$

Hence the LHS of Equation (12) can be written as

$$\Big|R(C \cap S) - Q(C \cap S)\Big| = P(C \cap S)\left|w(S) - \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_{C\cap S}}[w^*(x)]\right|$$

We derive Equation (15) by diving both sides of Equation (12) by $P(C \cap S)$. □

Proof of Lemma 3.3. We sample from the distribution $P|_C$ in two steps:

1. We first sample $\mathbf{S} \in \mathcal{S}$ according to the marginal distribution induced by $P|_C$ where $\Pr[\mathbf{S} = S_i] = P(C \cap S_i)/P(C)$.
2. We then sample $\mathbf{x} \in \mathbf{S}$ according to $P|_{C\cap\mathbf{S}}$ so that $\Pr[\mathbf{x} = x] = P(x)/P(C \cap \mathbf{S})$.

This allows us to use the fact that $w(x) = w(\mathbf{S})$ remains constant within each set of the partition, and that multi-calibration implies that $\mathbb{E}_{P|_{\mathbf{S}\cap C}}[w^*(x)]$ is close to $w(\mathbf{S})$ by Equation (15).

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w(\mathbf{x})w^*(\mathbf{x})] = \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_{\mathbf{S}\cap S}}[w(\mathbf{x})w^*(\mathbf{x})] = \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C} w(\mathbf{S}) \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_{\mathbf{S}\cap C}}[w^*(\mathbf{x})].$$

Hence using Equation (15) we have

$$\begin{aligned}\left|\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w(\mathbf{x})w^*(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}[w(\mathbf{S})^2]\right| &= \left|\mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}\left[w(\mathbf{S})\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_{\mathbf{S}\cap C}}[w^*(\mathbf{x})] - w(\mathbf{S})^2\right]\right| \\ &\le \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}\left[w(\mathbf{S})\left|\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_{\mathbf{S}\cap C}}[w^*(\mathbf{x})] - w(\mathbf{S})\right|\right] \\ &\le \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}\left[w(\mathbf{S})\frac{\alpha R(\mathbf{S})}{P(\mathbf{S}\cap C)}\right] \\ &= \sum_{S\in\mathcal{S}}\frac{P(S\cap C)}{P(C)}\frac{R(S)}{P(S)}\frac{\alpha R(S)}{P(S\cap C)} \\ &= \sum_{S\in\mathcal{S}}\frac{\alpha R(S)^2}{P(S)P(C)} = \frac{\alpha\|w\|^2}{P(C)}.\end{aligned}$$

□

Proof of Lemma 3.4. We start with the lower bound. By Equation (15) we have

$$\mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}[w(\mathbf{S})] \ge \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}\left[\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_{\mathbf{S}\cap C}}[w^*(\mathbf{x})] - \frac{\alpha R(S)}{P(C\cap S)}\right] = \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})] - \frac{\alpha}{P(C)}.$$

Using this bound and the convexity of x^2,

$$\mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}[w(\mathbf{S})^2] \ge \left(\mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}[w(\mathbf{S})]\right)^2 \ge \left(\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})] - \frac{\alpha}{P(C)}\right)^2 \ge \left(\mathop{\mathbb{E}}_{x\sim P|_C}[w^*(\mathbf{x})]\right)^2 - 2\alpha\frac{R(C)}{P(C)}.$$

We now show the upper bound.

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})^2] &= \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_{\mathbf{S}\cap C}}[w^*(\mathbf{x})^2] \ge \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_{\mathbf{S}\cap C}}[w^*(\mathbf{x})]\right)^2 \\ &\ge \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}\left[\left(w(S) - \frac{\alpha R(S)}{P(S\cap C)}\right)^2\right] \\ &\ge \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}\left[w(S)^2 - 2w(S)\frac{\alpha R(S)}{P(S\cap C)}\right]. \qquad (27)\end{aligned}$$

We have

$$\mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}\left[w(S)\frac{\alpha R(S)}{P(S\cap C)}\right] = \sum_{S\in\mathcal{S}}\frac{P(S\cap C)}{P(C)}\frac{R(S)}{P(S)}\frac{\alpha R(S)}{P(S\cap C)} = \sum_{S\in\mathcal{S}}\alpha\frac{R(S)^2}{P(S)P(C)} = \frac{\alpha\|w\|^2}{P(C)}.$$

Plugging this into Equation (27) gives

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})^2] \ge \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}[w(S)^2] - 2\frac{\alpha\|w\|^2}{P(C)}$$

which gives the desired upper bound. □

Proof of Lemma 4.5. Fix any $C \in \mathcal{C}$. Since $\mathcal{S}$ is α-multi-calibrated for $(P^h, R^h, \mathcal{C})$ and α-multi-calibration implies α-multi-accuracy, $|P^h(C) - R^h(C)| \le \alpha$. Also from Lemma 4.4,

$$\left|R(C) - R^h(C)\right| \le \mathrm{d_{TV}}(R, R^h) \le \beta, \quad \left|P(C) - P^h(C)\right| \le \mathrm{d_{TV}}(P, P^h) \le \beta.$$

Now using the triangle inequality,

$$\left|R(C) - P(C)\right| \le \left|R^h(C) - P^h(C)\right| + \left|R(C) - R^h(C)\right| + \left|P^h(C) - P(C)\right| \le \alpha + 2\beta.$$

□

C Sandwiching for (α, β)-multi-calibration

In this section, we prove Theorem 4.6 which asserts that for

$$\ell(\alpha, \beta, w) = \alpha \|w\|_2^2 + \sqrt{\beta}\, \|w\|_4^2$$

the following bounds hold for every $C \in \mathcal{C}$,

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim P|_C}[w^*(\mathbf{x})] - \frac{2\ell(\alpha,\beta,w)}{R(C)} - \frac{2(\alpha+2\beta)}{P(C)} \le \mathop{\mathbb{E}}_{\mathbf{x} \sim R|_c}[w(\mathbf{x})] \le \mathop{\mathbb{E}}_{\mathbf{x} \sim R|_C}[w^*(\mathbf{x})] + \frac{3\ell(\alpha,\beta,w)}{R(C)}.$$

Throughout this section, we assume that $\mathcal{S} = \{S_1, \dots, S_m, T_0, T_1\}$ is (α, β)-multi-calibration for $(P, R, \mathcal{C})$. We will use $S \in \mathcal{S}$ to denote a generic set in the partition, which could one of the S_is or T_js. We will now prove a sequence of technical lemmas that will be used to prove our bounds.

Lemma C.1. *For all $C \in \mathcal{C}$, we have*

$$\sum_{i \in [m]} \frac{P(S_i \cap C)}{P(C)} \left(\frac{R(S_i \cap C)^2}{P(S_i \cap C)^2} - \frac{R(S_i) R(S_i \cap C)}{P(S_i) P(S_i \cap C)} \right) \ge -\frac{3\alpha \|w\|_2^2}{P(C)}, \tag{28}$$

$$\sum_{i \in [m]} \frac{P(S_i \cap C)}{P(C)} \left(\frac{R(S_i) R(S_i \cap C)}{P(S_i) P(S_i \cap C)} - \frac{R(S_i)^2}{P(S_i)^2} \right) \ge -\frac{\alpha \|w\|_2^2}{P(C)}. \tag{29}$$

Proof. Using the multi-calibration condition (Equation (13)), we have

$$\frac{R(S_i \cap C)^2}{P(S_i \cap C)^2} \ge \left(\frac{R(S_i)}{P(S_i)} - \alpha \frac{R(S_i)}{P(S_i \cap C)} \right)^2 \ge \frac{R(S_i)^2}{P(S_i)^2} - 2\alpha \frac{R(S_i)^2}{P(S_i) P(S_i \cap C)}$$

$$\frac{R(S_i) R(S_i \cap C)}{P(S_i) P(S_i \cap C)} \le \frac{R(S_i)}{P(S_i)} \left(\frac{R(S_i)}{P(S_i)} + \alpha \frac{R(S_i)}{P(S_i \cap C)} \right) = \frac{R(S_i)^2}{P(S_i)^2} + \alpha \frac{R(S_i)^2}{P(S_i) P(S_i \cap C)}.$$

Subtracting the two bounds and averaging over S_is we get

$$\begin{aligned}
\sum_{i \in [m]} \frac{P(S_i \cap C)}{P(C)} \left(\frac{R(S_i \cap C)^2}{P(S_i \cap C)^2} - \frac{R(S_i) R(S_i \cap C)}{P(S_i) P(S_i \cap C)} \right) &\ge -3\alpha \sum_{i \in [m]} \frac{P(S_i \cap C)}{P(C)} \frac{R(S_i)^2}{P(S_i) P(S_i \cap C)} \\
&= -\frac{3\alpha}{P(C)} \sum_{i \in [m]} \frac{R(S_i)^2}{P(S_i)} \\
&\ge -\frac{3\alpha}{P(C)} \|w\|_2^2
\end{aligned}$$

which proves Equation (28).

We now prove (29). By the multi-calibration condition,

$$\begin{aligned}
\sum_{i \in [m]} \frac{P(S_i \cap C)}{P(C)} \frac{R(S_i) R(S_i \cap C)}{P(S_i) P(S_i \cap C)} &\ge \sum_{i \in [m]} \frac{P(S_i \cap C)}{P(C)} \frac{R(S_i)}{P(S_i)} \left(\frac{R(S_i)}{P(S_i)} - \alpha \frac{R(S_i)}{P(S_i \cap C)} \right) \\
&= \sum_{i \in [m]} \frac{P(S_i \cap C)}{P(C)} \frac{R(S_i)^2}{P(S_i)^2} - \sum_{i \in [m]} \frac{R(S_i)^2}{P(C) P(S_i)}.
\end{aligned}$$

Hence

$$\sum_{i \in [m]} \frac{P(S_i \cap C)}{P(C)} \left(\frac{R(S_i) R(S_i \cap C)}{P(S_i) P(S_i \cap C)} - \frac{R(S_i)^2}{P(S_i)^2} \right) \ge -\frac{\alpha \|w\|_2^2}{P(C)}.$$

□

Next we consider the T_0 term and show the following bounds

Lemma C.2. *For all $C \in \mathcal{C}$, we have*

$$\frac{P(T_0 \cap C)}{P(C)} \left(\frac{R(T_0 \cap C)^2}{P(T_0 \cap C)^2} - \frac{R(T_0)R(T_0 \cap C)}{P(T_0)P(T_0 \cap C)} \right) \geq -\frac{\sqrt{\beta}\,\|w\|_4^2}{P(C)}, \tag{30}$$

$$\frac{P(T_0 \cap C)}{P(C)} \left(\frac{R(T_0)R(T_0 \cap C)}{P(T_0)P(T_0 \cap C)} - \frac{R(T_0)^2}{P(T_0)^2} \right) \geq -\frac{\sqrt{\beta}\,\|w\|_4^2}{P(C)}. \tag{31}$$

Proof. If we have

$$\frac{R(T_0 \cap C)}{P(T_0 \cap C)} \geq \frac{R(T_0)}{P(T_0)}$$

then clearly both the LHSes are non-negative, hence both bounds hold. Assume this is not the case, then we have

$$\begin{aligned}
\frac{P(T_0 \cap C)}{P(C)} \left(\frac{R(T_0 \cap C)^2}{P(T_0 \cap C)^2} - \frac{R(T_0)R(T_0 \cap C)}{P(T_0)P(T_0 \cap C)} \right) &\geq -\frac{P(T_0 \cap C)}{P(C)} \frac{R(T_0)R(T_0 \cap C)}{P(T_0)P(T_0 \cap C)} \\
&\geq -\frac{P(T_0 \cap C)}{P(C)} \frac{R(T_0)^2}{P(T_0)^2} \\
&\geq -\frac{1}{P(C)} \frac{R(T_0)^2}{P(T_0)}
\end{aligned}$$

and similarly

$$\frac{P(T_0 \cap C)}{P(C)} \left(\frac{R(T_0)R(T_0 \cap C)}{P(T_0)P(T_0 \cap C)} - \frac{R(T_0)^2}{P(T_0)^2} \right) \geq -\frac{1}{P(C)} \frac{R(T_0)^2}{P(T_0)}. \tag{32}$$

We can bound this as

$$\frac{R(T_0)^2}{P(T_0)} = P(T_0) \frac{R(T_0)^2}{P(T_0)^2} = \left(P(T_0) \cdot P(T_0) w(T_0)^4 \right)^{1/2} \leq \sqrt{\beta}\,\|w\|_4^2$$

where we use $P(T_0) \leq \beta$ and $P(T_0)w(T_0)^4 \leq \|w\|_4^4$. Plugging this into Equation (32) completes the proof. ☐

Finally for the set T_1 we show the following.

Lemma C.3. *For all $C \in \mathcal{C}$, we have*

$$\frac{P(T_1 \cap C)}{P(C)} \left(\frac{R(T_1 \cap C)^2}{P(T_1 \cap C)^2} - \frac{R(T_1)R(T_1 \cap C)}{P(T_1)P(T_1 \cap C)} \right) \geq -\frac{\beta}{P(C)}, \tag{33}$$

$$\frac{P(T_1 \cap C)}{P(C)} \left(\frac{R(T_1 \cap C)R(T_1)}{P(T_1 \cap C)P(T_1)} - \frac{R(T_1)^2}{P(T_1)^2} \right) \geq -\frac{\beta}{P(C)}. \tag{34}$$

Proof. If

$$\frac{R(T_1 \cap C)}{P(T_1 \cap C)} \geq \frac{R(T_1)}{P(T_1)}$$

then both LHSs are non-negative, so the bound holds. Else,

$$\frac{R(T_1 \cap C)}{P(T_1 \cap C)} \leq \frac{R(T_1)}{P(T_1)} \leq 1$$

where the inequality is by second by the definition of T_1. So we have the lower bound

$$\frac{P(T_1 \cap C)}{P(C)} \left(\frac{R(T_1 \cap C)^2}{P(T_1 \cap C)^2} - \frac{R(T_1)R(T_1 \cap C)}{P(T_1)P(T_1 \cap C)} \right) \geq -\frac{P(T_1 \cap C)}{P(C)} \frac{R(T_1)R(T_1 \cap C)}{P(T_1)P(T_1 \cap C)} \geq -\frac{\beta}{P(C)}$$

since $P(T_1 \cap C) \leq \beta$, and the other two ratios are at most 1. This proves Equation (33). Equation (34) is shown similarly. ☐

Lemma C.4. *For all $C \in \mathcal{C}$, we have*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w^*(\mathbf{x})] + \frac{3}{R(C)}(\alpha\|w\|_2^2 + \sqrt{\beta}\,\|w\|_4^2) \geq \mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w(\mathbf{x})|].$$

Proof. We first show the bound

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})^2 - w(\mathbf{x})w^*(\mathbf{x})] \geq -\frac{3}{P(C)}(\alpha\|w\|_2^2 + \sqrt{\beta}\,\|w\|_4^2). \tag{35}$$

We have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})^2] &= \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_{\mathbf{S}\cap C}}[w^*(\mathbf{x})^2] \geq \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_{\mathbf{S}\cap C}}[w^*(\mathbf{x})]\right)^2 \\
&= \sum_{S\in\mathcal{S}} \frac{P(S\cap C)}{P(C)}\frac{R(S\cap C)^2}{P(S\cap C)^2} \\
&= \sum_{i\in[m]} \frac{P(S_i\cap C)}{P(C)}\frac{R(S_i\cap C)^2}{P(S_i\cap C)^2} + \sum_{j\in\{0,1\}} \frac{P(T_j\cap C)}{P(C)}\frac{R(T_j\cap C)^2}{P(T_j\cap C)^2}.
\end{aligned} \tag{36}$$

On the other hand,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w(x)w^*(\mathbf{x})] &= \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}\left[w(\mathbf{S})\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_{\mathbf{S}\cap C}}[w^*(\mathbf{x})]\right] = \sum_{S\in\mathcal{S}} \frac{P(S\cap C)}{P(C)}\frac{R(S)}{P(S)}\frac{R(S\cap C)}{P(S\cap C)} \\
&= \sum_{i\in[m]} \frac{P(S_i\cap C)}{P(C)}\frac{R(S_i)R(S_i\cap C)}{P(S_i)P(S_i\cap C)} + \sum_{j\in\{0,1\}} \frac{P(T_j\cap C)}{P(C)}\frac{R(T_j)R(T_j\cap C)}{P(T_j)P(T_j\cap C)}.
\end{aligned} \tag{37}$$

We subtract the Equation (37) from (36). We then apply the lower bounds from Equation (28) to bound the contribution from the S_is, Equation (31) for T_0 and Equation (34) for T_1 to get

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})^2 - w(\mathbf{x})w^*(\mathbf{x})] &\geq -\frac{1}{P(C)}(3\alpha\,\|w\|_2^2 + \sqrt{\beta}\,\|w\|_4^2 + \beta) \\
&\geq -\frac{3}{P(C)}(\alpha\,\|w\|_2^2 + \sqrt{\beta}\,\|w\|_4^2)
\end{aligned}$$

which proves the bound claimed in Equation (35).

To derive the claim from this, we use the following equalities from Equation (17) and (18):

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w(\mathbf{x})w^*(\mathbf{x})] &= \mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w(\mathbf{x})]\frac{R(C)}{P(C)}, \\
\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})^2] &= \mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w^*(\mathbf{x})]\frac{R(C)}{P(C)}.
\end{aligned}$$

Plugging these into Equation (35) gives

$$\frac{R(C)}{P(C)}\mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w^*(x) - w(x)] \geq -\frac{3}{P(C)}(\alpha\,\|w\|_2^2 + \sqrt{\beta}\,\|w\|_4^2)$$

which gives the claimed bound upon rearranging. ∎

Lemma C.5. *For all $C \in \mathcal{C}$, we have*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w(\mathbf{x})] + \frac{2}{R(C)}(\alpha\|w\|_2^2 + \sqrt{\beta}\,\|w\|_4^2) + \frac{2(\alpha+2\beta)}{P(C)} \geq \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})|].$$

Proof. Recall that by Equation (37)

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w(x)w^*(\mathbf{x})] = \sum_{i\in[m]} \frac{P(S_i\cap C)}{P(C)}\frac{R(S_i)R(S_i\cap C)}{P(S_i)P(S_i\cap C)} + \sum_{j\in\{0,1\}} \frac{P(T_j\cap C)}{P(C)}\frac{R(T_j)R(T_j\cap C)}{P(T_j)P(T_j\cap C)}.$$

Recall the bounds from Equations (29), (31) and (34) which state

$$\begin{aligned}
\sum_{i\in[m]} \frac{P(S_i\cap C)}{P(C)}\left(\frac{R(S_i)R(S_i\cap C)}{P(S_i)P(S_i\cap C)} - \frac{R(S_i)^2}{P(S_i)^2}\right) &\geq -\frac{\alpha\left\|w\right\|_2^2}{P(C)}\\
\frac{P(T_0\cap C)}{P(C)}\left(\frac{R(T_0)R(T_0\cap C)}{P(T_0)P(T_0\cap C)} - \frac{R(T_0)^2}{P(T_0)^2}\right) &\geq -\frac{\sqrt{\beta}\left\|w\right\|_4^2}{P(C)}\\
\frac{P(T_1\cap C)}{P(C)}\left(\frac{R(T_1\cap C)R(T_1)}{P(T_1\cap C)P(T_1)} - \frac{R(T_1)^2}{P(T_1)^2}\right) &\geq -\frac{\beta}{P(C)}.
\end{aligned}$$

Adding these bounds, we get

$$\sum_{S\in\mathcal{S}} \frac{P(S\cap C)}{P(C)}\frac{R(S)R(S\cap C)}{P(S)P(S\cap C)} \geq \sum_{S\in\mathcal{S}} \frac{P(S\cap C)}{P(C)}\frac{R(S)^2}{P(S)^2} - \frac{2}{P(C)}(\alpha\left\|w\right\|_2^2 + \sqrt{\beta}\left\|w\right\|_4^2). \tag{38}$$

We also have

$$\sum_{S\in\mathcal{S}} \frac{P(S\cap C)}{P(C)}\frac{R(S)^2}{P(S)^2} \geq \left(\sum_{S\in\mathcal{S}} \frac{P(S\cap C)}{P(C)}\frac{R(S)}{P(S)}\right)^2. \tag{39}$$

But note that

$$\sum_{S\in\mathcal{S}} P(S\cap C)\frac{R(S)}{P(S)} = \sum_{S\in\mathcal{S}} R(S)P(C|S) = Q(C) \geq R(C) - \alpha - 2\beta$$

by Lemma 4.5 showing that (α,β)-multi-calibration implies $(\alpha+2\beta)$-multi-accuracy. Plugging this into Equation (39) gives

$$\sum_{S\in\mathcal{S}} \frac{P(S\cap C)}{P(C)}\frac{R(S)^2}{P(S)^2} \geq \left(\frac{R(C)-\alpha-2\beta}{P(C)}\right)^2 \geq \left(\frac{R(C)}{P(C)}\right)^2 - 2(\alpha+2\beta)\frac{R(C)}{P(C)^2}. \tag{40}$$

Putting Equations (38) and (39) together with Equation (37) gives

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w(\mathbf{x})w^*(\mathbf{x})] \geq \left(\mathop{\mathbb{E}}_{\mathbf{S}\sim P|C}[w^*(x)]\right)^2 - \frac{2}{P(C)}(\alpha\left\|w\right\|_2^2 + \sqrt{\beta}\left\|w\right\|_4^2) - 2(\alpha+2\beta)\frac{R(C)}{P(C)^2}.$$

Using Equations (17) and (18) and diving both sides by $R(C)/P(C)$ gives

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w(\mathbf{x})] \geq \mathop{\mathbb{E}}_{\mathbf{x}\sim P|C}[w^*(\mathbf{x})] - \frac{2}{R(C)}(\alpha\left\|w\right\|_2^2 + \sqrt{\beta}\left\|w\right\|_4^2) - \frac{2(\alpha+2\beta)}{P(C)}.$$

□