Multicalibrated Partitions for Importance Weights

Parikshit Gopalan PGOPALAN@VMWARE.COM
Omer Reingold REINGOLD@STANFORD.EDU
Vatsal Sharan VSHARAN@USC.EDU
Udi Wieder UWIEDER@VMWARE.COM

Editors: Sanjoy Dasgupta and Nika Haghtalab

Abstract

The ratio between the probability that two distributions assign to points in the domain are called importance weights or density ratios and they play a fundamental role in machine learning and information theory. However, there are strong lower bounds known for point-wise accurate estimation of density ratios, and most theoretical guarantees require strong assumptions about the distributions. We motivate the problem of seeking accuracy guarantees for the distribution of importance weights conditioned on sub-populations belonging to a family $\mathcal{C}$ of subsets of the domain. We formulate *sandwiching bounds* for sets: upper and lower bounds on the expected importance weight conditioned on a set; as a notion of set-wise accuracy for importance weights. We argue that they capture intuitive expectations about importance weights, and are not subject to the strong lower bounds for point-wise guarantees. We introduce the notion of multi-calibrated partitions for a class $\mathcal{C}$, inspired by recent work on multi-calibration in supervised learning (Hébert-Johnson et al., 2018) and show that the importance weights resulting from such partitions do satisfy sandwiching bounds. In contrast, we show that importance weights returned by popular algorithms in the literature may violate the sandwiching bounds. We present an efficient algorithm for constructing multi-calibrated partitions, given a weak agnostic learner for the class $\mathcal{C}$.

Keywords: Importance weighting, agnostic learning, algorithmic fairness, maximum entropy

1. Introduction

Given two distributions P and R over a domain $\mathcal{X}$, the ratio $w^*(x) = R(x)/P(x)$ defines a function $w : \mathcal{X} \to \mathbb{R}$ known as the density-ratio or importance weights of R relative to P. We consider the problem of estimating this function given access to random samples from P and R.

This problem has been studied by several communities (often under different monikers) and has a wide variety of applications. In machine learning, it arises both in unsupervised problems such as anomaly detection and supervised settings such as domain adaptation. For anomaly detection in the inlier-based (Hido et al., 2008; Sugiyama et al., 2012c) or semi-supervised model (Chandola et al., 2009), P represents normal points or inliers, and R represents our observations. The goal is to find points x where $P(x)$ is small relative to $R(x)$; essentially the points where $w^*(x)$ is high (Schölkopf et al., 2001; Hodge and Austin, 2004; Hido et al., 2008; Smola et al., 2009). In domain adaptation, 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 (Shimodaira, 2000; Zadrozny, 2004; Sugiyama and Müller, 2005; Bickel et al., 2007; Cortes et al., 2010; Ben-David and Urner, 2012). In information theory, importance weight estimation (aka

Radon-Nikodym derivative estimation), is applied to estimate various divergences such as the KL divergence and Renyi divergences between the distributions P and R (Nguyen et al., 2007, 2010; Yamada et al., 2013; Wang et al., 2005, 2009). These divergence measures themselves find several applications, such as two-sample tests for distinguishing between two distributions (Wornowizki and Fried, 2016) and for independence testing (Suzuki et al., 2008). 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 (Rosenbaum and Rubin, 1983). For further details and applications, we refer the reader to the book Sugiyama et al. (2012c).

As a counter-weight to these applications, there are strong information-theoretic lower bounds known. The problem of estimating importance weights from samples is known to be hard for arbitrary P and R, as are applications like domain adaptation and divergence estimation (Ben-David and Urner, 2012; Batu et al., 2013). These lower bounds stem from sample complexity results on testing distributions, where given samples from both P and R we must decide whether $P = R$ or if they are far apart in statistical distance (Batu et al., 2013; Valiant, 2011). These results imply that computing any reasonable point-wise approximation to w^* requires sample complexity exponential in the dimension of $\mathcal{X}$.

There are several algorithms known for density-ratio estimation in the literature based on moment matching, logistic regression, the Kullback-Liebler information estimation procedure (KLIEP) and more (Sugiyama et al., 2012c). Many of these estimate importance weights by learning some distribution Q whose importance weight function w relative to P belongs to a certain parametric family. They choose the parameters to minimize some divergence between the distribution R and the model Q. For appropriate choices of divergence, this paradigm captures several popular approaches to density-ratio estimation including all those mentioned above (Sugiyama et al., 2012a). If the true importance weight function w^* happens to lie in the parametric family, the algorithm will converge to the right answer. In the realistic non-realizable setting where w^* is not in the family, the algorithm finds the *best approximation* from the family, with quality measured by the chosen divergence. However, it is unclear how good this best approximation is, and the known guarantees are not always meaningful in some of the aforementioned applications.

Consider for example a researcher analyzing data about a disease outbreak in a county, where each point represents demographic and medical information about an individual. They model the data collected from patients as a distribution R, the general population in the county as another distribution P, and build a model w of the importance weights. Their goal is understanding the vulnerability of certain sub-populations that are represented by conjunctions of attributes.

1. Let C be a sub-population such that $R(C)/P(C) > 10$, so that the prevalence within C is 10 times higher than was expected based on the prior. The researcher would like a random sample of datapoints drawn from R conditioned on being in the set C to be assigned large weight by w, ideally at least 10. If not, w might not alert them to the increased prevalence within C.

2. Let C' be a sub-population such that a random sample of datapoints drawn from R conditioned on being in C' is assigned an average weight of 10 by w. Does this mean that the true importance weights are large in expectation for R conditioned on C'? Or might having large weights for C' be a false alarm?

In analogy to proof systems (Arora and Barak, 2006, Chapter 8), these conditions ask for **completeness** and **soundness** of the importance weights w respectively. Completeness requires that if a set C

is important under R, then it receives large weights w on average under R. Soundness requires that if the average weight under R assigned to a set C' is large, this indicates that the set is important under R. Requiring such guarantees is natural from a group fairness perspective, especially in applications like anomaly detection. On one hand, (with the right formulation) set-wise guarantees do not imply strong point-wise guarantees, hence known lower bounds do not apply. On the other hand, we will show that (perhaps surprisingly) popular algorithms in the literature cannot give such guarantees. At a high level, these algorithms find the model which minimizes expected loss, and might not necessarily capture the behavior on sub-populations accurately.

Our Contributions We summarize the main contributions of this paper briefly.

- We propose requiring set-wise accuracy guarantees for a class $\mathcal{C}$ of sets that include sub-populations we care about. We formulate Sandwiching bounds as a notion of set-wise accuracy for importance weights, and show that they capture the **completeness** and **soundness** for importance weights.
- We introduce the notion of **multi-calibrated partitions** for a class $\mathcal{C}$, inspired by recent work on multi-calibration in supervised learning (Hébert-Johnson et al., 2018). We show that the importance weights resulting from such partitions do satisfy sandwiching bounds.
- We present an efficient algorithm for constructing such multi-calibrated partitions, given a **weak agnostic learner** for the class $\mathcal{C}$, which adapts the Boosting by branching programs algorithm (Mansour and McAllester, 2002) to importance weight estimation.
- We show that importance weights returned by popular algorithms such as log-linear KLIEP (Sugiyama et al., 2008, 2012c) (equivalently MaxEnt (Jaynes, 1957; Della Pietra et al., 1997; Dudik et al., 2007)) may violate the Sandwiching bounds, by constructing explicit examples.

Notation. 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 $\mathbf{x}$ according to distribution P. For $A \subseteq \mathcal{X}$, let $P(A) = \Pr_{\mathbf{x}\sim P}[\mathbf{x} \in A]$ and $P|_A$ denote P conditioned on A. To every distribution $Q(x)$, one can associate an importance weight function relative to P by $w(x) = Q(x)/P(x)$, we denote this by $Q = w \cdot P$. For any $A \subseteq \mathcal{X}$ observe that

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_A}[w(\mathbf{x})] = \sum_{x\in A} \frac{P(x)}{P(A)} \frac{Q(x)}{P(x)} = \frac{Q(A)}{P(A)}. \tag{1}$$

In our setting, we have a *target* distribution R and a *prior* distribution P, where $R = w^* \cdot P$. Our aim is to find importance weights w such that $Q = w \cdot P$ is a good model for R. We will consider a family of $\mathcal{C} = \{C \subseteq \mathcal{X}\}$ of subsets which include the sub-populations for which we desire guarantees. The collection can be infinite, and contain overlapping subsets, for instance taking $\mathcal{C}$ to be all decision trees or neural nets of a given size lets us capture sub-populations which are conjunctions of attributes. For our algorithmic results, we assume an efficient algorithm to learn the class of indicator functions of $C \in \mathcal{C}$, formally that $\mathcal{C}$ is *weakly agnostically learnable* (see Definition 13).

2. Setwise guarantees and sandwiching bounds

We now rigorously formulate a notion called *sandwiching bounds* which formally capture the completeness and soundness requirements. We desire guarantees for a collection of sets $\mathcal{C} = \{C \subseteq \mathcal{X}\}$ that include the sub-populations of interest. While our definitions make sense for arbitrary $\mathcal{C}$, our algorithmic results require the indicators of the sets to be efficiently weakly agnostically learnable. For a distribution R and a set C, let $R|_C$ denote the distribution R conditioned on C. Ideally[1] we would like the following **strict Sandwiching bounds** to hold for every $C \in \mathcal{C}$:

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

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 do not have access to w^* explicitly). When $P|_C$ and $R|_C$ are identical, the upper and lower bounds in Equation (2) are equal. But in general, they could be far apart.

Let us see why sandwiching bounds indeed capture the aforementioned requirements. For the lower bound, observe that by Equation (1), $\mathbb{E}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})] = R(C)/P(C)$ hence this inequality captures condition (1). The upper bound can be interpreted as saying that we want our weights to be conservative, they should not exaggerate the prevalence of anomalies within a set. 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 (2) implies the following outer inequality for w^* (independent of the model w):

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

This is a consequence of convexity, as is apparent from the following restatement of Equation (2), which says that for sandwiching, we want the weights w to have the right correlation with the true weights w^* under $P|_C$.

Lemma 1 *Equation (2) is equivalent to*

$$\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})]\right)^2 \leq \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w(\mathbf{x})w^*(\mathbf{x})] \leq \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}\left[w^*(\mathbf{x})^2\right]. \tag{4}$$

Proof We can rewrite the expectation under $R|_C$ in Equation (2) in terms of expectation under $P|_C$ as follows

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w(\mathbf{x})] = \frac{\sum_{x\in C} p(x)w(x)w^*(x)}{\sum_{x\in C} p(x)w^*(x)} = \frac{\mathbb{E}_{\mathbf{x}\sim P|_C}[w(\mathbf{x})w^*(\mathbf{x})]}{\mathbb{E}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})]} \tag{5}$$

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w^*(\mathbf{x})] = \frac{\sum_{x\in C} p(x)w^*(x)^2}{\sum_{x\in C} p(x)w^*(x)} = \frac{\mathbb{E}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})^2]}{\mathbb{E}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})]}. \tag{6}$$

Plugging these into Equation (2), we can rewrite those inequalities as

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

which is equivalent to Equation (4). ■

1. In practice we allow a bit of slack, see Theorem 5.

3. Multi-calibrated partitions and sandwiching bounds

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 $P|_S$. In this section, we assume that P and R are both supported on the entire domain $\mathcal{X}$, so that the distributions $P|_S$ and $R|_S$ are well-defined for any non-empty subset S. We will relax this requirement in Section 3.1

Definition 2 *Given a prior distribution P and a target distribution R supported on $\mathcal{X}$, and a partition $\mathcal{S}$ of $\mathcal{X}$, the $(P, R, \mathcal{S})$-reweighted distribution Q over $\mathcal{X}$ is given by*

$$Q(x) = R(S)P(x|S) \text{ for } S \in \mathcal{S} \text{ s.t. } x \in S. \tag{7}$$

Equivalently, $Q = w \cdot P$ where $w(x) = R(S)/P(S)$ for all $x \in S \in \mathcal{S}$.

The equivalence follows by observing that $R(S)P(x|S) = R(S)P(x)/P(S) = w(x)P(x)$.

Intuitively, the goal of multi-calibration is to find a partition $\mathcal{S}$, hopefully of small size, whose reweighting will be sufficient to get accuracy for a large family of tests $\mathcal{C}$. We formalize this below.

Definition 3 *(α-multi-calibration) Let $\alpha > 0$. 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*

$$\Big|Q(C \cap S) - R(C \cap S)\Big| \leq \alpha R(S). \tag{8}$$

The multi-calibration condition has the following equivalent formulations (proved in Appendix A):

Lemma 4 *Equation (8) is equivalent to either of the following equations*

$$\Big|P(C|S) - R(C|S)\Big| \leq \alpha, \tag{9}$$

$$\left|w(S) - \mathop{\mathbb{E}}_{\mathbf{x} \sim P|_{C \cap S}}[w^*(x)]\right| \leq \frac{\alpha R(S)}{P(C \cap S)}. \tag{10}$$

Equation (9) reveals that the definition of multi-calibration is symmetric in P and R. Equation (10) connects our definition to multi-calibration in the supervised setting from Hébert-Johnson et al. (2018). To see this, consider the importance weights w^* of R as the ground truth, and w as our prediction which is fixed for each $S \in \mathcal{S}$. For each S and $C \in \mathcal{C}$, we compare our prediction to the conditional expectation of the ground truth, analogous to the classical notion of calibration for a predictor (Dawid, 1982).

Multi-calibration implies sandwiching bounds. For the weight function w, and $k \geq 1$ define

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

Observe $\|w\|_1 = 1$ and $\|w\|_k$ increases with k. The following result is proved in Appendix A:

Theorem 5 *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*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})] - 2\alpha\frac{\|w\|_2^2}{R(C)} \leq \mathop{\mathbb{E}}_{\mathbf{x}\sim R|_c}[w(\mathbf{x})] \leq \mathop{\mathbb{E}}_{\mathbf{x}\sim R|_C}[w^*(\mathbf{x})] + 3\alpha\frac{\|w\|_2^2}{R(C)} \tag{11}$$

Comparing Theorem 5 to the strict sandwiching requirement stated in Equation (2), we see that the theorem allows for a slack of roughly $\alpha \|w\|_2^2 / R(C)$. Some such slack is unavoidable given our model where we only have access to random samples from both P and R. One can view α above as the accuracy parameter which decides the sample complexity. With more samples, α can be made smaller, hence we approach strict sandwiching, which can be thought of as the limit with infinitely many samples. In contrast, in Section 4, we give negative examples showing previous algorithms achieving α-multi-accuracy do not satisfy Sandwiching bounds even for $\alpha = 0$. In particular, no bound analogous to Equation (11) holds for those algorithms, even as the number of samples goes to infinity.

We justify why the terms $R(C)$ and $\|w\|_2$ appear in the slack. With access only to samples we cannot hope to give guarantees for sets C which are very small, for the same reason that guarantees for singleton point sets are not possible. In particular, with only a finite number of samples, we cannot tell whether $R(C)$ is 0, or just very small, similarly with $P(C)$. Hence, we can only expect meaningful guarantees for sets where $R(C)$ is reasonably large, say $1/\alpha$.

Similarly, some bound on the norm of the importance weights seems unavoidable. Imagine that we see a set S in our partition with 0.1-fraction of the samples from R, but no samples from P. It is hard to tell with finitely many samples whether $P(S) = 0$ so the importance weight here tends to infinity or whether $P(S)$ is non-zero but small, say, $1/10\alpha$, so the importance weight is just very large. The simple solution would be to assume a hard bound $\max w^*(x) \leq B$ for some constant B. We only require a bound on the l_2 norm which is weaker, because our weights always satisfy $\|w\|_2^2 \leq \|w^*\|_2^2 \leq \max_x w^*(x)$.

3.1. A relaxed notion of multi-calibration

We now relax the notion of multi-calibration to handle the case where R and P might not have identical supports. We do this in a *robust* manner, that lets us handle the case where one of P and R assigns non-zero but tiny probability to some set T. For such T, enforcing the closeness of $P(C|T)$ and $R(C|T)$ based on random samples will be very expensive in terms of sample complexity. Such T corresponds to a region where the importance weight $w(T) = R(T)/P(T)$ is either very high or very low. We will relax our defintion to allow regions in the partition with $w(T) \leq \beta$ or $w(T') \geq 1/\beta$ for some parameter β without having to determine exactly what those weights are (which in itself would require many samples). This motivates the following definition.

Definition 6 *Let $\alpha \geq 0, \beta \geq 0$, 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*

$$P(T_0) \leq \min(\beta, R(T_0)),\ R(T_1) \leq \min(\beta, P(T_1)), \tag{12}$$

$$\forall C \in \mathcal{C}, i \in [m],\ \left|Q(C \cap S_i) - R(C \cap S_i)\right| \leq \alpha R(S_i). \tag{13}$$

(α, β)-multi-calibration permits two *exceptional* subsets T_0 and T_1 that do not satisfy Equation (8), 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. The advantage of allowing for $\mathcal{T}$ is that for every $S \in \mathcal{S} \setminus \mathcal{T}$, we *can ensure* that $w(S) = R(S)/P(S)$ lies in the range $[\beta, 1/\beta]$, since sets violating this bound may be absorbed in $\mathcal{T}$. This lets us enforce Equation (13) with sample complexity that depends on $1/\beta$.

When $\beta = 0$ we recover the notion of α-multi-calibration. 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.

Lemma 7 *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 $R^h|_{T_1} = P|_{T_1}$. If the partition $\mathcal{S}$ is (α, β)-multi-calibrated for $(P, R, \mathcal{C})$, then $\mathrm{d_{TV}}(P^h, P) \leq \beta$, $\mathrm{d_{TV}}(R^h, R) \leq \beta$, and the partition $\mathcal{S}$ is α-multi-calibrated for $(P^h, R^h, \mathcal{C})$.*

We show a sandwiching bound for (α, β)-multi-calibration. The error terms depend on β and $\|w\|_4^2$ in comparison to Theorem 5. The proof appears in Appendix B.

Theorem 8 *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*

$$\ell(\alpha, \beta, w) = \alpha \|w\|_2^2 + \sqrt{\beta}\, \|w\|_4^2 . \tag{14}$$

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

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim P|_C}[w^*(\mathbf{x})] - \frac{2\ell(\alpha, \beta, w)}{R(C)} - \frac{2(\alpha + 2\beta)}{P(C)} \leq \mathop{\mathbb{E}}_{\mathbf{x} \sim R|_c}[w(\mathbf{x})] \leq \mathop{\mathbb{E}}_{\mathbf{x} \sim R|_C}[w^*(\mathbf{x})] + \frac{3\ell(\alpha, \beta, w)}{R(C)}. \tag{15}$$

4. Multi-accuracy alone does not guarantee sandwiching

We consider two well-studied methods for density ratio estimation; log-linear KLIEP (Sugiyama et al., 2008, 2012c) and MaxEnt (Jaynes, 1957; Della Pietra et al., 1997; Kazama and Tsujii, 2003; Dudik et al., 2007), which are essentially duals of each other. Our motivation for considering these algorithms is two-fold. Firstly, they can be interpreted as giving a set-wise guarantee we call multi-accuracy, in analogy with the corresponding notion in supervised learning (Hébert-Johnson et al., 2018; Kim et al., 2019), which is weaker than multi-calibration. Secondly, these algorithms are known to out-perform other density-ratio estimation algorithms in the non-realizable setting (Kanamori et al., 2010). To state these algorithms, we define two families of distributions.

Definition 9 *For $\alpha \in [0, 1]$, the distribution Q is α-multi-accurate in expectation (α-multiAE) for $(R, \mathcal{C})$ if for every $C \in \mathcal{C}$, it holds that $|Q(C) - R(C)| \leq \alpha$. Define $K^\alpha = K^\alpha(R, \mathcal{C})$ to be the set of all α-multi-accurate distributions for $(R, \mathcal{C})$.*

α-multi-calibration for the partition $\mathcal{S}$ implies multi-accuracy in expectation for the re-weighted distribution Q; this follows by summing Equation (8) over all $S \in \mathcal{S}$. For every $\mathcal{C}$ and $\alpha \geq 0$, the set K^α is a convex set, since it is given by linear constraints, and it is non-empty since $R \in K^\alpha$. Our second family of distributions are Gibbs distributions.

Definition 10 *A Gibbs distribution is a distribution of the form*

$$Q(x) = P(x) \exp\left(\sum_{c \in \mathcal{C}} \lambda_c c(x) - \lambda_0\right). \tag{16}$$

Let $\mathcal{G} = \mathcal{G}(P, \mathcal{C})$ denote the set of all Gibbs distributions.

Note that the free parameters are $\lambda_{\mathcal{C}} = \{\lambda_c\}_{c \in \mathcal{C}}$, from these, we set the normalization constant λ_0 to ensure that $\mathbb{E}_P[w(\mathbf{x})] = 1$, so that Q is a distribution. We now describe log-linear KLIEP and MaxEnt, both of which find a multi-accurate Gibbs distribution.

1. **Log-linear KLIEP** (Sugiyama et al., 2008; Nguyen et al., 2007; Sugiyama et al., 2012c) : Find the Gibbs distribution $Q \in \mathcal{G}$ that minimizes $D(R\|Q)$. The goal in Sugiyama et al. (2008) is to find a good density-ratio estimate. Essentially this algorithm is proposed in the work of Nguyen et al. (2007) for estimating KL divergence.

2. **MaxEnt** (Jaynes, 1957; Della Pietra et al., 1997; Kazama and Tsujii, 2003; Dudik et al., 2004, 2007) : Find the distribution $Q^\alpha \in K^\alpha$ that minimizes $D(Q\|P)$.

We have not found the equivalence of these algorithms noted explicitly in the literature, but it follows from known results on convex duality (Della Pietra et al., 1997; Kazama and Tsujii, 2003; Dudik et al., 2007).

Lemma 11 *(Dudik et al., 2007, Theorem 2) For $Q \in \mathcal{G}(P, \mathcal{C})$ as in Equation (16), let $\ell_1(Q) = \sum_{c \in \mathcal{C}} |\lambda_c|$. $Q^\alpha \in K^\alpha \cap \mathcal{G}$ is the optimal solution to each of the following programs:*

$$\min_{Q \in K^\alpha} D(Q\|P), \tag{17}$$

$$\min_{Q \in \mathcal{G}} D(R\|Q) + \alpha \ell_1(Q). \tag{18}$$

The first program is the one solved by MaxEnt. The second is an ℓ_1-regularized version of the program considered by KLIEP. We derive the exact KLIEP program by setting $\alpha = 0$, in the MaxEnt literature this was analyzed by Della Pietra et al. (1997). Our main result in this section is that these algorithms do not guarantee sandwiching bounds. The proof is in Appendix C.

Theorem 12 *Fix any constant $B > 1$, and $\alpha \geq 0$. There exist distributions P, R on $\{0,1\}^n$, a collections of sets $\mathcal{C}$ and $C \in \mathcal{C}$ such that if $Q^\alpha = w \cdot P$ is the solution to Program (17) then*

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

There exist distributions P_1, R_1 on $\{0,1\}^n$, a collections of sets $\mathcal{C}_1$ and $C_1 \in \mathcal{C}_1$ such that if $Q_1^\alpha = w_1 \cdot P_1$ is the solution to Program (17), then writing $R_1 = w_1^* \cdot P_1$,

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim R_1|_C}[w_1(\mathbf{x})] > B \mathop{\mathbb{E}}_{\mathbf{x} \sim R_1|_C}[w_1^*(\mathbf{x})].$$

Let us give some intuition for why multi-accuracy by itself cannot guarantee sandwiching whereas multi-calibration does. We remind the reader of Lemma 1, which rephrases the sandwiching conditions as saying that the correlation between $w(x)$ and $w^*(x)$ under $P|_C$ behaves as expected. Multi-accuracy can be rephrased as saying that

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

Thus while it guarantees that $w(x)$ and $w^*(x)$ have similar expectations under $P|_C$, it does not give guarantees on their correlation. In contrast, the calibration property of multi-calibration (captured by Equation (10)) guarantees that conditioned on each value $w(x) = w(S)$, the expected value of $w^*(x)$ under $P|_C$ is close to $w(S)$. This ensures that $w(x)$ and $w^*(x)$ are better correlated.

5. Algorithm for 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 inspired by Boosting via Branching Programs from Mansour and McAllester (2002). We first define weak agnostic learning (Ben-David et al., 2001; Kalai et al., 2008). Given a collection of sets $\mathcal{C} \subseteq 2^{\mathcal{X}}$, we associate $C \in \mathcal{C}$ with its indicator function $c : \mathcal{X} \to \{0, 1\}$. With this view, we can also regard the set $\mathcal{C}$ as a hypothesis class with binary-valued functions $c : \mathcal{X} \to \{0, 1\}$.

Definition 13 *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 a set $C \in \mathcal{C}$ with corresponding indicator function $c : \mathcal{X} \to \{0, 1\}$ such that $\Pr_{\mathcal{D}}[c(\mathbf{x}) = \mathbf{y}] \geq (1 + \alpha)/2$, then the learner will return an indicator function $c' : \mathcal{X} \to \{0, 1\}$ corresponding to some $C' \in \mathcal{C}$ such that $\Pr_{\mathcal{D}}[c'(\mathbf{x}) = \mathbf{y}] \geq (1 + \alpha')/2$ for some $0 < \alpha' \leq \alpha$.*

We allow $\alpha - \alpha'$ to depend on L, typically it decreases with L. In the above definition, for simplicity we have defined the learner to be a proper learner (it returns a hypothesis within $\mathcal{C}$), 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.

Informally, the weak agnostic learning assumption says that if there is a hypothesis in $\mathcal{C}$ that labels the data reasonably well (say with 0-1 loss of 0.7), then we can efficiently find one that has a non-trivial advantage over random guessing (say with 0-1 loss of 0.51). Weak agnostic learning was introduced in (Ben-David et al., 2001). It captures a common modeling assumption in practice, and is a well-studied notion in the computational learning literature Kalai et al. (2008); Kanade and Kalai (2009); Feldman (2009). The assumption has also been used in previous works in the multi-calibration literature (Hébert-Johnson et al., 2018; Jung et al., 2021).

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 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 $\mathcal{T}^0, \mathcal{T}^1$ 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 T_1. Doing the merging at the end simplifies the analysis, but 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 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.

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 $S_1 = S \cap \bar{C}$.

We will analyze the Split operation using $D\left(Q_t \| P\right)$ as the potential function.

Lemma 14 *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 can write the LHS as

$$\begin{aligned}
&R(S \cap C) \log\left(\frac{R(S \cap C)}{P(S \cap C)}\right) + R(S \cap \bar{C}) \log\left(\frac{R(S \cap \bar{C})}{P(S \cap \bar{C})}\right) - R(S) \log\left(\frac{R(S)}{P(S)}\right) \\
&= R(S \cap C) \log\left(\frac{R(S \cap C)P(S)}{R(S)P(S \cap C)}\right) + R(S \cap \bar{C}) \log\left(\frac{R(S \cap \bar{C})P(S)}{R(S)P(S \cap \bar{C})}\right) \\
&= R(S)\left(R(C|S) \log\left(\frac{R(C|S)}{P(C|S)}\right) + R(\bar{C}|S) \log\left(\frac{R(\bar{C}|S)}{P(\bar{C}|S)}\right)\right). \qquad (19)
\end{aligned}$$

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 applying Pinsker's inequality (Cover and Thomas, 2006) gives the desired bound:

$$R(C|S) \log\left(\frac{R(C|S)}{P(C|S)}\right) + R(\bar{C}|S) \log\left(\frac{R(\bar{C}|S)}{P(\bar{C}|S)}\right) \geq \left|R(C|S) - P(C|S)\right|^2 \geq 4\alpha'^2. \qquad (20)$$

■

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

Lemma 15 *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

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

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.

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

We use these in Algorithm 3 which computes a multi-calibrated partition.

To sketch the overall analysis briefly, we have shown that $D(Q_t \| P)$ increases during a Split, and decreases during a Merge. But there are m Split operations between any two Merge operations, so overall $D(Q_t \| P)$ increases, but can never exceed $D(R \| P)$. Details appear in appendix D.

Theorem 16 *Given an (α, α', L) weak agnostic learning algorithm for $\mathcal{C}$, Algorithm 3 returns an (α, β)-multi-calibrated partition of size $m = O(\log(1/\beta)/(\alpha'^2\beta))$ that can be represented by a $\mathcal{C}$-branching program where each node is labelled by $c \in \mathcal{C}$. The algorithm performs $T = \widetilde{O}(D(R\|P)/(\beta^2\alpha'^4)$ Split and Merge operations. 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.*

6. Other Related Work

In Section 1, we discussed some of the diverse applications of importance weights or density-ratio estimation which span many communities. We refer the reader to the book (Sugiyama et al., 2012b) for a more comprehensive overview of the related work, especially in the context of the machine learning literature. In addition to density ratio estimation, relevant results appear in the literature under the subject of divergence estimation (Nguyen et al., 2007, 2010; Wang et al., 2005, 2009), learning max-Entropy distributions (Jaynes, 1957; Dudik et al., 2007) and domain adaptation (Cortes et al., 2010; Ben-David and Urner, 2012; Redko et al., 2019). Kernel based approaches for estimating importance weights have also been proposed, starting with Kernel Mean Matching (KMM) introduced in Huang et al. (2007). If the underlying kernel is universal, then under the limit of infinite data KMM provably recovers the true importance weights (Cortes et al., 2008; Huang et al., 2007). In the limit of finite data however, kernel based approaches can be interpreted as generalizations of moment-matching methods such as (Qin, 1998) which seek to match some moments of the data (Sugiyama et al., 2012a). In the context of our work, such based approaches guarantee multi-accuracy in expectation with respect to the moments or the feature space more generally.

Calibration has been well-studied in the statistics literature, in the context of forecasting (Dawid, 1982). It was introduced in the algorithmic fairness literature by Kleinberg et al. (2017). The

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

Inputs: $\alpha, \beta > 0$, distributions P, R, 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 $\text{Merge}(\delta)$.
2. If the weak agnostic leaner finds $S \in \mathcal{S}^t$, $C \in \mathcal{C}$ such that

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

   2.1. Run $\text{Split}(S, C)$ 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 $\mathcal{S} = \mathcal{S}^t \cup \{T_0\} \cup \{T_1\}$.

notion of multi-calibration as a multi-group fairness notion in supervised learning was introduced in Hébert-Johnson et al. (2018) (see also Kearns et al. (2018)) and has subsequently generated significant interest (Kim et al., 2019; Jung et al., 2021; Barda et al., 2020; Dwork et al., 2021). But all previous work to our knowledge has been in the supervised setting. Our work appears to be the first to introduce notions of group-fairness in unsupervised learning. Our algorithm for computing a multi-calibrated partition is inspired by the Boosting by Branching programs work of Mansour and McAllester (2002), which in turn builds on Kearns and Mansour (1999). The work of Kalai et al. (2008) showed that the Mansour-McAllester algorithm can be viewed as an agnostic boosting algorithm, improving on the results of Ben-David et al. (2001) who introduced the notion of agnostic boosting.

7. Conclusion

In this work, we have put forth a theoretical framework for reasoning about the accuracy of importance weights for sub-populations of the dataset. We presented an algorithm that provably gives much

stronger guarantees in this regard than previously known algorithms in the literature. The next step is to implement this algorithm and test its guarantees for real-world datasets. The implications of these stronger guarantees for the numerous applications of importance weights should also be explored: among these we consider divergence estimation and anomaly detection to be particularly promising.

References

Sanjeev Arora and Boaz Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, 2006.

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

Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing closeness of discrete distributions. *J. ACM*, 60(1):4:1–4:25, 2013. doi: 10.1145/2432622.2432626.

Shai Ben-David and Ruth Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *Algorithmic Learning Theory - 23rd International Conference, ALT*, 2012.

Shai Ben-David, Philip M. Long, and Yishay Mansour. Agnostic boosting. In *14th Annual Conference on Computational Learning Theory, COLT*, 2001.

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.

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

Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 38–53. Springer, 2008.

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

Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006. ISBN 978-0-471-24195-9.

A. P. Dawid. Objective probability forecasts. *University College London, Dept. of Statistical Science. Research Report 14*, 1982.

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.

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

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.

Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Outcome indistinguishability. In *ACM Symposium on Theory of Computing (STOC'21)*, 2021.

Vitaly Feldman. Distribution-specific agnostic boosting. *arXiv preprint arXiv:0909.2927*, 2009.

Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, 2018.

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.

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

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.

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

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.

Adam Tauman Kalai, Yishay Mansour, and Elad Verbin. On agnostic boosting and parity learning. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing, Victoria, British Columbia, Canada, May 17-20, 2008*, pages 629–638. ACM, 2008.

Varun Kanade and Adam Kalai. Potential-based agnostic boosting. *Advances in Neural Information Processing Systems*, 22:880–888, 2009.

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.

Jun'ichi Kazama and Jun'ichi Tsujii. Evaluation and extension of maximum entropy models with inequality constraints. In *EMNLP '03: Proceedings of the 2003 conference on Empirical Methods in Natural Language Processing*, pages 137–144, 01 2003.

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

Michael J. Kearns and Yishay Mansour. On the boosting ability of top-down decision tree learning algorithms. *J. Comput. Syst. Sci.*, 58(1):109–128, 1999.

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.

Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS*, 2017.

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

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.

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.

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

I. Redko, E. Morvant, M. Habrard, A.and Sebban, and Y Bennani. *Advances in Domain Adaptation Theory*. Elsevier, 2019.

Paul Rosenbaum and Donald Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 04 1983. doi: 10.1093/biomet/70.1.41.

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.

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.

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

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.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Von Bunau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 2008.

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, 2012a.

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

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012c. ISBN 978-0-521-19017-6.

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.

Paul Valiant. Testing symmetric properties of distributions. *SIAM J. Comput.*, 40(6):1927–1968, 2011. doi: 10.1137/080734066.

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.

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.

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

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.

Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the Twenty-First International Conference on Machine Learning*, page 114, 2004.

Appendix A. Proofs from Section 3

Proof [Proof of Lemma 4]

The first equivalence holds since we have $Q(C \cap S) = R(S)P(C|S)$, while $R(C \cap S) = R(S)R(C|S)$. We substitute these in Equation (8) and divide by $R(S)$ to derive Equation (9).

Applying Equation (1) 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 the definition of Q we have

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

Hence the LHS of Equation (8) can be written as

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

We derive Equation (10) by diving both sides of Equation (8) by $P(C \cap S)$. ■

Next, we prove Theorem 5. Using the formulation in Equation (4), 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 Lemma 17 and 18.

Lemma 17 *We have*

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

Proof [Proof of Lemma 17] 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 (10).

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

Hence using Equation (10) we have

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

■

Lemma 18 *We have*

$$\left(\mathop{\mathbb{E}}_{x\sim P|_C}[w^*(\mathbf{x})]\right)^2 - 2\alpha\frac{R(C)}{P(C)} \le \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}[w(\mathbf{S})^2] \le \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})^2] + 2\alpha\frac{\|w\|_2^2}{P(C)} \tag{22}$$

Proof [Proof of Lemma 18] We start with the lower bound. By Equation (10) 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.

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})^2] &= \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_{\mathbf{S}\cap C}}[w^*(\mathbf{x})^2] \ge \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_{\mathbf{S}\cap C}}[w^*(\mathbf{x})]\right)^2 \\
&\ge \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}\left[\left(w(S) - \frac{\alpha R(S)}{P(S\cap C)}\right)^2\right] \\
&\ge \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}\left[w(S)^2 - 2w(S)\frac{\alpha R(S)}{P(S\cap C)}\right].
\end{aligned} \tag{23}$$

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 (23) 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. ■

We now put these together to prove Theorem 5.

Proof [Proof of Theorem 5] We claim the following inequalities hold

$$\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(x)]\right)^2 - \alpha\frac{\|w\|_2^2 + R(C)}{P(C)} \le \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w(\mathbf{x})w^*(\mathbf{x})] \le \mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(x)^2] + 3\alpha\frac{\|w\|_2^2}{P(C)}. \tag{24}$$

These are an immediate consequence of Lemma 17 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 18 which gives a sandwiching bound for $\mathbb{E}_{\mathbf{S}\sim P|_C}[w(\mathbf{S})^2]$.

Equation (24) equivalent to Equation (11). To see this, we use the following equalities from Equation (5) and (6):

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

We plug these into Equation (24) and divide throughout by $\mathbb{E}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})] = R(C)/P(C)$ to derive Equation (11) and complete the proof. ■

Appendix B. Sandwiching for (α, β)-multi-calibration

In this section, we prove Theorem 8 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 $\mathcal{S} = \{S_1, \ldots, S_m, T_0, T_1\}$ 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.

We first formalize our claim that we can assume weights for $S \in \mathcal{S} \setminus \mathcal{T}$ may be assumed to be bounded, by showing that sets whose weights are outside this range have small probability under one of the distributions.

Lemma 19 *Let $T \subseteq \mathcal{X}$ and $c \ge 1$ be such that $w(T) = R(T)/P(T) \ge c/\beta$. Then $P(T) \le \beta/c$.*

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

We provide the proof of Lemma 7

Proof [Proof of Lemma 7] The statistical distance bounds hold since P^h and P only differ on T_0 and $P^h(T_0) = P(T_0) \le \beta$. We verify that the partition is multi-calibrated by showing that $|P(C|S) - R(C|S)| \le \alpha$ for every state $S \in \mathcal{S}$. For any $i \in [m]$, we have

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

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 (13). 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 20 *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 $(P, R, \mathcal{C})$.*

Lemma 21 *For all $C \in \mathcal{C}$, we have*

$$\sum_{i\in[m]} \frac{P(S_i \cap C)}{P(C)} \left(\frac{R(S_i \cap C)^2}{P(S_i \cap C)^2} - \frac{R(S_i)R(S_i \cap C)}{P(S_i)P(S_i \cap C)} \right) \geq -\frac{3\alpha \|w\|_2^2}{P(C)}, \tag{25}$$

$$\sum_{i\in[m]} \frac{P(S_i \cap C)}{P(C)} \left(\frac{R(S_i)R(S_i \cap C)}{P(S_i)P(S_i \cap C)} - \frac{R(S_i)^2}{P(S_i)^2} \right) \geq -\frac{\alpha \|w\|_2^2}{P(C)}. \tag{26}$$

Proof Using the multi-calibration condition (Equation (9)), we have

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

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

Subtracting the two bounds and averaging over S_is we get

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

which proves Equation (25).

We now prove (26). By the multi-calibration condition,

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

Hence

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

■

Next we consider the T_0 term and show the following bounds

Lemma 22 *For all $C \in \mathcal{C}$, we have*

$$\frac{P(T_0 \cap C)}{P(C)} \left(\frac{R(T_0 \cap C)^2}{P(T_0 \cap C)^2} - \frac{R(T_0) R(T_0 \cap C)}{P(T_0) P(T_0 \cap C)} \right) \geq -\frac{\sqrt{\beta}\, \|w\|_4^2}{P(C)}, \tag{27}$$

$$\frac{P(T_0 \cap C)}{P(C)} \left(\frac{R(T_0) R(T_0 \cap C)}{P(T_0) P(T_0 \cap C)} - \frac{R(T_0)^2}{P(T_0)^2} \right) \geq -\frac{\sqrt{\beta}\, \|w\|_4^2}{P(C)}. \tag{28}$$

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

$$\frac{P(T_0 \cap C)}{P(C)} \left(\frac{R(T_0) R(T_0 \cap C)}{P(T_0) P(T_0 \cap C)} - \frac{R(T_0)^2}{P(T_0)^2} \right) \geq -\frac{1}{P(C)} \frac{R(T_0)^2}{P(T_0)}. \tag{29}$$

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 (29) completes the proof. ■

Finally for the set T_1 we show the following.

Lemma 23 *For all $C \in \mathcal{C}$, we have*

$$\frac{P(T_1 \cap C)}{P(C)} \left(\frac{R(T_1 \cap C)^2}{P(T_1 \cap C)^2} - \frac{R(T_1) R(T_1 \cap C)}{P(T_1) P(T_1 \cap C)} \right) \geq -\frac{\beta}{P(C)}, \tag{30}$$

$$\frac{P(T_1 \cap C)}{P(C)} \left(\frac{R(T_1 \cap C) R(T_1)}{P(T_1 \cap C) P(T_1)} - \frac{R(T_1)^2}{P(T_1)^2} \right) \geq -\frac{\beta}{P(C)}. \tag{31}$$

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 (30). Equation (31) is shown similarly. ■

Lemma 24 *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

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})^2 - w(\mathbf{x})w^*(\mathbf{x})] \geq -\frac{3}{P(C)}(\alpha\|w\|_2^2 + \sqrt{\beta}\,\|w\|_4^2). \tag{32}$$

We have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w^*(\mathbf{x})^2] &= \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_{\mathbf{S}\cap C}}[w^*(\mathbf{x})^2] \geq \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_{\mathbf{S}\cap C}}[w^*(\mathbf{x})]\right)^2 \\
&= \sum_{S\in\mathcal{S}}\frac{P(S\cap C)}{P(C)}\frac{R(S\cap C)^2}{P(S\cap C)^2} \\
&= \sum_{i\in[m]}\frac{P(S_i\cap C)}{P(C)}\frac{R(S_i\cap C)^2}{P(S_i\cap C)^2} + \sum_{j\in\{0,1\}}\frac{P(T_j\cap C)}{P(C)}\frac{R(T_j\cap C)^2}{P(T_j\cap C)^2}.
\end{aligned} \tag{33}$$

On the other hand,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_C}[w(x)w^*(\mathbf{x})] &= \mathop{\mathbb{E}}_{\mathbf{S}\sim P|_C}\left[w(\mathbf{S})\mathop{\mathbb{E}}_{\mathbf{x}\sim P|_{\mathbf{S}\cap C}}[w^*(\mathbf{x})]\right] = \sum_{S\in\mathcal{S}}\frac{P(S\cap C)}{P(C)}\frac{R(S)}{P(S)}\frac{R(S\cap C)}{P(S\cap C)} \\
&= \sum_{i\in[m]}\frac{P(S_i\cap C)}{P(C)}\frac{R(S_i)R(S_i\cap C)}{P(S_i)P(S_i\cap C)} + \sum_{j\in\{0,1\}}\frac{P(T_j\cap C)}{P(C)}\frac{R(T_j)R(T_j\cap C)}{P(T_j)P(T_j\cap C)}.
\end{aligned} \tag{34}$$

We subtract the Equation (34) from (33). We then apply the lower bounds from Equation (25) to bound the contribution from the S_is, Equation (28) for T_0 and Equation (31) 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 (32).

To derive the claim from this, we use the following equalities from Equation (5) and (6):

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

Plugging these into Equation (32) 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 25 *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 (34)

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

Recall the bounds from Equations (26), (28) and (31) which state

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

Adding these bounds, we get

$$\sum_{S\in\mathcal{S}}\frac{P(S\cap C)}{P(C)}\frac{R(S)R(S\cap C)}{P(S)P(S\cap C)} \geq \sum_{S\in\mathcal{S}}\frac{P(S\cap C)}{P(C)}\frac{R(S)^2}{P(S)^2} - \frac{2}{P(C)}(\alpha\|w\|_2^2 + \sqrt{\beta}\|w\|_4^2). \quad (35)$$

We also have

$$\sum_{S\in\mathcal{S}}\frac{P(S\cap C)}{P(C)}\frac{R(S)^2}{P(S)^2} \geq \left(\sum_{S\in\mathcal{S}}\frac{P(S\cap C)}{P(C)}\frac{R(S)}{P(S)}\right)^2. \quad (36)$$

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 20 showing that (α, β)-multi-calibration implies $(\alpha + 2\beta)$-multi-accuracy. Plugging this into Equation (36) gives

$$\sum_{S \in \mathcal{S}} \frac{P(S \cap C)}{P(C)} \frac{R(S)^2}{P(S)^2} \geq \left(\frac{R(C) - \alpha - 2\beta}{P(C)} \right)^2 \geq \left(\frac{R(C)}{P(C)} \right)^2 - 2(\alpha + 2\beta) \frac{R(C)}{P(C)^2}. \tag{37}$$

Putting Equations (35) and (36) together with Equation (34) gives

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

Using Equations (5) and (6) and diving both sides by $R(C)/P(C)$ gives

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

■

Appendix C. Proof of Theorem 12: Gap instances

In this section we will show that the importance weights w^α found by the solution to the program in (17) (the problem solved by MaxEnt, which is equivalent to the problem solved by KLIEP) 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^α 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 26 *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 any $\alpha \geq 0$ returns a distribution Q^α with importance weights w^α such that*

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

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

We can compute the following importance weights

1. $w^\alpha(11) = (5/8)^2/(1/2)^2 = 25/16$ whereas $w^*(11) = 1$.
2. $w^\alpha(10) = (5/8 \cdot 3/8)/(1/2)^2 = 15/16$, 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^α assigns these points weights less than 1, and instead assigns a high weight to 11. Thus an algorithm that was labelling points with w^α 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

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

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

For the case $\alpha > 0$, first note that the maximum entropy distribution is still a product distribution since conditioning reduces entropy. Secondly, as we increase α the bias of the individual coordinates in Q moves towards $1/2$, but this only makes the gap larger (since $\mathbb{E}_{\mathbf{x}\sim R|_C}[w^\alpha(\mathbf{x})] = 1$ if $Q^\alpha = P$). ■

Intuitively, in the example above, while Q^α 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 27 *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 any $\alpha \geq 0$ returns a distribution Q^α with importance weights w^α such that*

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

Proof We now consider the k-wise tensor of the instances constructed in Lemma 26. 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 $Q_k^\alpha = (Q^\alpha)^k$ which is just the product distribution on $\{0,1\}^{2k}$ with $\Pr[x_i = 1] = 5/8$ for every coordinate.

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

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

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

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

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

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 12.

Theorem 28 *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 any $\alpha \geq 0$ returns a distribution Q^α with importance weights w^α such that*

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

Proof We consider the case $\alpha = 0$, the general case follows as in the proof of Lemma 26 by a suitable choice of parameters. 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^α is the product distribution on x_0 and x_1 where $\Pr[x_0 = 1] = 11/16$ and $\Pr[x_1 = 1] = 1/2$. 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:

$$w^*(01) = 3/4, w^*(11) = 5/4$$
$$w^\alpha(01) = 5/8, w^\alpha(11) = 11/8.$$

Hence we have the conditional expectations

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

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

Appendix D. Analysis of Algorithm 3

Theorem 29 *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 19, 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, proving Theorem 16

Proof [Proof of Theorem 16] 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 14, 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 $c \in \mathcal{C}$.