HappyMap: A Generalized Multicalibration Method

Zhun Deng
Department of Computer Science, Columbia University, New York, NY, USA

Cynthia Dwork
Department of Computer Science, Harvard University, Cambridge, MA, USA

Linjun Zhang
Department of Statistics, Rutgers University, Piscataway, NJ, USA

Abstract

Multicalibration is a powerful and evolving concept originating in the field of algorithmic fairness. For a predictor f that estimates the outcome y given covariates x, and for a function class $\mathcal{C}$, multi-calibration requires that the predictor $f(x)$ and outcome y are indistinguishable under the class of auditors in $\mathcal{C}$. Fairness is captured by incorporating demographic subgroups into the class of functions $\mathcal{C}$. Recent work has shown that, by enriching the class $\mathcal{C}$ to incorporate appropriate propensity re-weighting functions, multi-calibration also yields target-independent learning, wherein a model trained on a source domain performs well on unseen, future, target domains (approximately) captured by the re-weightings.

Formally, multicalibration with respect to $\mathcal{C}$ bounds $\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f(x),x)\cdot(f(x)-y)]\right|$ for all $c\in\mathcal{C}$. In this work, we view the term $(f(x)-y)$ as just one specific mapping, and explore the power of an enriched class of mappings. We propose *s-Happy Multicalibration*, a generalization of multi-calibration, which yields a wide range of new applications, including a new fairness notion for uncertainty quantification, a novel technique for conformal prediction under covariate shift, and a different approach to analyzing missing data, while also yielding a unified understanding of several existing seemingly disparate algorithmic fairness notions and target-independent learning approaches.

We give a single *HappyMap* meta-algorithm that captures all these results, together with a sufficiency condition for its success.

2012 ACM Subject Classification Theory of computation → Design and analysis of algorithms; Theory of computation → Theory and algorithms for application domains

Keywords and phrases algorithmic fairness, target-independent learning, transfer learning

Digital Object Identifier 10.4230/LIPIcs.ITCS.2023.41

Funding This work was supported in part by Alfred P. Sloan Foundation G-2020-13941, Simons Foundation Collaborative Research Grant 733782, and NSF DMS-2015378.

Acknowledgements We thank all the reviewers for their comments and suggestions. We are indebted to Aaron Roth for his insightful and valuable feedback, which greatly improved the paper.

1 Introduction

Prediction algorithms score individuals or individual instances, assigning to each a score in $[0,1]$ typically interpreted as a probability, for example, the probability that it will rain tomorrow. The predictions are calibrated if, for all $v\in[0,1]$, among those instances assigned the value v, a v fraction have a positive outcome. Calibration has been viewed as the *sine qua non* of prediction for decades [7].

The requirement that a predictor be simultaneously calibrated on each of two or more disjoint groups, meaning, it is calibrated on each group when viewed in isolation, was first proposed as a *fairness* condition by Kleinberg, Mullainathan, and Raghavan [28].

Inspired by [28] and in an attempt to bridge the gap between *individual fairness* [13], which demands that similar individuals be treated similarly but requires task-specific measures of similarity, and *group fairness*, which can be specious [13], Hébert-Johnson, Kim, Reingold, and Rothblum proposed *multicalibration*, which requires calibration on a (possibly large) pre-specified collection of *arbitrarily intersecting* sets that can be identified within a specified class of computations [21]. A related, independent, work of Kearns, Neel, Roth, and Wu considered the analogous setting but for Boolean-valued classifiers. They argued that including intersecting groups can prevent fairness gerrymandering, and developed multi-parity [25].

The area has blossomed in theory and in application. For example, multicalibration has been used for fair ranking [14], and for providing an indistinguishability-based interpretation of individual probabilibilities, i.e., probabilities for non-repeatable events [15]. Multicalibration has also been shown to yield *omniprediction*, meaning that for every "nice" loss function ℓ, the scores assigned by the multicalibrated predictor can be post-processed, with no additional training, to be competitive, on the training distribution, with the best predictor in $\mathcal{C}$ [18].

▶ **Definition 1** (Multicalibration [21] as presented in [27]). *Let $\mathcal{C} \subseteq \{[0,1] \times \mathcal{X} \to \mathbb{R}\}$ be a collection of functions. For a given distribution supported on $\mathcal{X} \times \mathcal{Y}$, a predictor $f : \mathcal{X} \mapsto [0,1]$ is $(\mathcal{C}, \alpha)$-multicalibrated over $\mathcal{D}$ if $\forall c \in \mathcal{C}$:*

$$\left| \mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f(x), x) \cdot (f(x) - y)] \right| \leq \alpha. \tag{1}$$

Fairness is captured by incorporating demographic subgroups into the class of functions $\mathcal{C}$. By enriching the class $\mathcal{C}$ to incorporate appropriate propensity re-weighting functions, multicalibration also yields target-independent learning, wherein a model trained on a source domain performs well on unseen, future, target domains (approximately) captured by the re-weightings. [27].

In this work, we view the term $(f(x) - y)$ in Equation (1) as just one specific mapping $s(f, y) : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$, and explore the power of an enriched class of mappings. To this end, we propose *HappyMap*, a generalization of multicalibration, which yields a wide range of new applications, including a new algorithm for fair uncertainty quantification, a novel technique for conformal prediction under distributional (a.k.a. covariate) shift, and a different approach to analyzing missing data, while also yielding a unified understanding of several existing seemingly disparate algorithmic fairness notions and target-independent learning approaches.

We give a single *HappyMap* meta-algorithm that captures all these results, together with a sufficiency condition for its success. Roughly speaking, the requirement is that the mapping have an anti-derivative satisfying a smoothness-like assumption (see Section 3). We say such a mapping is *happy*. Loosely speaking, the anti-derivative serves as a potential function, which yields an upper bound on the number of iterations of our algorithm.

Summary of Contributions.

1. We propose *HappyMap*, a generalization of multicalibration, by enriching the class of mappings (alternatives for the term $(f(x) - y)$ in Equation 2), as discussed above, and provide a *HappyMap* meta-algorithm having comparable running time and sample complexity to the other multicalibration algorithms in the literature, and give sufficient conditions for its success (Section 3).
2. We demonstrate the flexibility of *HappyMap* by first applying it to obtain generalized versions of *fair uncertainty quantification* [32] (Section 4).
3. Furthermore, we apply *HappyMap* to problems in *target-independent* learning that lie beyond the statistical estimation problems considered in [27], obtaining target-independent statistical inference and uncertainty quantification. Our approach also yields a fruitful new perspective on analyzing missing data, giving new solutions to this problem (Section 5).

A High-Level Perspective. The seminal work of Hébert-Johnson et al. [21] has long tantalized us with the suggestion of a new paradigm for machine learning. We coin the term *micro-learning* to describe gradient descent, a learning paradigm paradigm based on loss minimization, consisting of a sequence of local model updates. In contrast, the multiaccuracy and multicalibration algorithms of [21] produce predictors via interactions (through weak agnostic learning) with auditors who find large problem areas without the intercession of a loss function, and make correspondingly large model updates, suggested to us what we call *macro-learning*. This perspective has a compelling transparency story, *"the learning algorithm finds large errors and fixes them"*, which is particularly appealing when the auditors are interrogating the treatment of demographic subgroups. The fact that macro-learning can be used as post-processing [21] tells us that the two paradigms can work in concert. The concept of s-Happy Multicalibration proposed in this paper advances the broad vision of macro-learning.

2 Preliminaries

2.1 Notation

For $d \in \mathbb{N}+$, any convex set $S \subseteq \mathbb{R}^d$ and vector $v \in \mathbb{R}^d$, we use $\Pi_S(v)$ to denote the the projection of v on S under Euclidean distance. We sometimes use the probability notion $\mathbb{P}$ also for density function, for instance $\mathbb{P}(x)$ means the density function evaluated at x for continuous random vector x. Let us denote $X \in \mathcal{X}$ as the feature vector (typically $\mathcal{X} = \mathbb{R}^d$), $Y \in \mathcal{Y}$ (typically $\mathcal{Y} = \mathbb{R}$ for regression problems and $\mathcal{Y} = \{0,1\}$ for classification problems) as the response that we are trying to predict. For a joint distribution of $(x, y) \in \mathcal{X} \times \mathcal{Y}$, let us denote the marginal distribution of x and y as $\mathcal{D}^{\mathcal{X}}$ and $\mathcal{D}^{\mathcal{Y}}$ respectively. For two positive sequences $\{a_k\}$ and $\{b_k\}$, we write $a_k = \mathcal{O}(b_k)$ (or $a_n \lesssim b_n$), and $a_k = o(b_k)$, if $\lim_{k\to\infty}(a_k/b_k) < \infty$ and $\lim_{k\to\infty}(a_k/b_k) = 0$ respectively. $\tilde{\mathcal{O}}(\cdot)$ denotes the term, neglecting the logarithmic factors. We also write $a_k = \Theta(b_k)$ (or $a_k \asymp b_k$) if $a_k \lesssim b_k$ and $b_k \lesssim a_k$. We use $\mathcal{O}_p(\cdot)$ to denote stochastic boundedness: a random variable $Z_k = \mathcal{O}_p(a_k)$ for some real sequence a_k if $\forall \epsilon > 0$, there exists $M, N > 0$ such that if $k > N$, $\mathbb{P}(|Z_k/a_k| > M) \leq \epsilon$. For two numbers $a < b$, we use the notation $U(a, b)$ to denote the uniform distribution on $[a, b]$. For $\mu \in \mathbb{R}$ and $\sigma > 0$, we use $N(\mu, \sigma^2)$ to denote a normal distribution with mean μ and variance σ^2.

3 s-Happy Multicalibration

We now formally state our generalization of multicalibration.

▶ **Definition 2** (s-Happy Multicalibration)**.** *Let $\mathcal{C} \subseteq \{\mathbb{R} \times \mathcal{X} \to \mathbb{R}\}$ be a collection of functions. For a given distribution supported on $\mathcal{X} \times \mathcal{Y}$ and a mapping $s : \mathbb{R} \times \mathcal{X} \mapsto \mathbb{R}$, a predictor $f : \mathcal{X} \mapsto \mathbb{R}$ is $(\mathcal{C}, \alpha)$-s-happily multicalibrated over $\mathcal{D}$ and s if for all $c \in \mathcal{C}$:*

$$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f(x), x) \cdot s(f(x), y)]\right| \leq \alpha. \tag{2}$$

When $\mathcal{D}$ and s are clear from the context, we will also simply say f is $(\mathcal{C}, \alpha)$-happily multicalibrated. Sometimes we will constrain the range of f to a certain convex set O, for example, $[0, 1]$; in this case we say $f : \mathcal{X} \mapsto O$ is (C, α)-s-happily multicalibrated when it satisfies Eq. (2). When $c(f(x), x)$ is independent of f and only depends on x, we will simply write $c(x)$.

s-Happy Multicalibration (Equation (2)) captures several multi-group fairness notions in the literature. For example, if we let $c(f(x), x)$ be be independent of f, and take $s(f(x), y) = f(x) - y$, we recover multi-accuracy [26, 21]; if we take $c(f(x), x) = \tilde{c}(x) \cdot w(f(x))$ for some functions $\tilde{c}$ and w and $s(f(x), y) = f(x) - y$, we recover the low-degree multicalibration notion in [19]; if we take $s(f(x), y)$ to be a non-negative loss function and $c(x)$ to be indicator functions of different groups, we recover the minimax group fairness in [9].

The *HappyMap* Meta-Algorithm. We now describe the *HappyMap* meta-algorithm and prove its key properties. For simplicity, we describe the *population* version of our algorithm, where we are allowed to access $\mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f(x), x)s(f(x), y)]$. Of course, in practice, we will need to estimate these quantities from training data, and so we describe the sample version of Algorithm 1 in Appendix B.1 and derive the corresponding required sample complexity. In keeping with the multi-group fairness literature [21, 26], we can either use a fresh sample per iteration or, when samples are limited, apply techniques from adaptive data analysis [10, 11, 12] to re-use the same samples in each iteration, as suggested in [21].

We consider the general case and aim to return a $f : \mathcal{X} \mapsto O$, for a given convex set O, which is $(\mathcal{C}, \alpha)$-happily multicalibrated. Without loss of generality, let us assume the class $\mathcal{C}$ is symmetric in the sense that c and $-c$ are both in $\mathcal{C}$ (we can always augment $\tilde{\mathcal{C}}$ by including $-c$ for $c \in \tilde{\mathcal{C}}$, so this is without loss of generality). The algorithm is invoked with an initial predictor f_0, which can be trivial ($f_0(x) = c$ for all $x \in \mathcal{X}$) or can be an artisanal predictor imbued with extensive domain knowledge.

Algorithm 1 HappyMap.

Input: Tolerance $\alpha > 0$, step size $\eta > 0$, bound $T \in \mathbb{N}_+$ on number of iterations, initial predictor $f_0(\cdot)$, distribution $\mathcal{D}$, convex set $O \subseteq \mathbb{R}$, mapping $s : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$

Set $t = 0$

while $t < T$ *and* $\exists c_t \in \mathcal{C} : \mathbb{E}_{(x,y)\sim\mathcal{D}}[c_t(f_t(x), x)s(f_t(x), y)] > \alpha$ **do**

  Let c_t be an arbitrary element of $\mathcal{C}$ satisfying the condition in the while statement

  $\forall x \in \mathcal{X},\ f_{t+1}(x) = \Pi_O\left[f_t(x) - \eta \cdot c(f(x), x)\right]$

  $t = t + 1$

end

Output: $f_t(\cdot)$

As in previous work [21, 32, 27, 25], we do not address the complexity of the weak learner whose job it is to search for functions $c_t \in \mathcal{C}$ satisfying the condition of the while loop, or to report that none exists. The problem is at least as hard as agnostic learning [21, 25]. See [5] for discussion of this issue and its implications for fairness. Henceforth, our unit of computational complexity will be an invocation of the weak learner, i.e., an iteration of the while loop.

Analysis. Theorem 3 below summarizes the theoretical guarantees for Algorithm 1. As is common in the literature, the proof of termination relies on a potential function argument (see, e.g., [4, 21]). The key new ingredient is the notion of a *happy* mapping, which provides the conditions under which we can find the necessary potential function.

We will require some assumptions, which we state and discuss before stating the theorem.

Three Assumptions. (a) For all $c \in \mathcal{C}, x \in \mathcal{X}$, $\mathbb{E}_{x\sim\mathcal{D}^{\mathcal{X}}}[c^2(f(x), x)] \leq B$, where $\mathcal{D}^{\mathcal{X}}$ is the marginal distribution of $\mathcal{D}$ on x; (b) There exists a potential function $\mathcal{L} : \mathbb{R} \times \mathcal{Y} \mapsto \mathbb{R}$, constants $C^l_{\mathcal{L}}, C^u_{\mathcal{L}}$, positive constant $\kappa_{\mathcal{L}} > 0$, such that (1) for all $f, \tilde{f} : \mathcal{X} \mapsto \mathbb{R}$,

$C^l_{\mathcal{L}} \leq \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(f(x),y)]$, (2) $C^l_{\mathcal{U}} \geq \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(f_0(x),y)]$ for the initialization f_0, and (3) $\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(f(x),y)-\mathcal{L}(\tilde{f}(x),y)] \geq \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x)-\tilde{f}(x))s(f(x),y)] - \kappa_{\mathcal{L}}\mathbb{E}_{x\sim\mathcal{D}^{\mathcal{X}}}[(f(x)-\tilde{f}(x))^2]$; (c) For all $y \in \mathcal{Y}$ and $v \in \mathbb{R}$, $\mathcal{L}(\Pi_O[v],y) \leq \mathcal{L}(v,y)$.

Assumption (a) is routine and Assumption (c) says that the potential function decreases upon projection with respect to its first coordinate. One concrete example is the case in which $O = \mathcal{Y} = [0,1]$ and $\mathcal{L}(v,y) = |v-y|^2$.

We now turn to Assumption (b), focusing on *how to construct the potential function* $\mathcal{L}$. First note that the assumption is closely related to *smoothness*, a widely used concept in optimization. We use the following fact.

Fact. If $\mathcal{L}(\cdot,\cdot) : \mathbb{R}\times\mathcal{X} \mapsto \mathbb{R}$ is $2\kappa_{\mathcal{L}}$-smooth with respect to its first coordinate, i.e. $\mathcal{L}(\cdot,\cdot)$ is continuously differentiable with respect to its first coordinate, and the corresponding partial derivative is $2\kappa_{\mathcal{L}}$-Lipchitz, then, for all $x \in \mathcal{X}, y \in \mathcal{Y}$ and $f, \tilde{f} : \mathcal{X} \mapsto \mathbb{R}$,

$$\mathcal{L}(f(x),y) - \mathcal{L}(\tilde{f}(x),y) \geq \partial_u \mathcal{L}(u,y)|_{u=f(x)}(f(x)-\tilde{f}(x)) - \kappa_{\mathcal{L}}(f(x)-\tilde{f}(x))^2. \quad (3)$$

Thus, if $s(f(x),y)$ is differentiable with respect to its first coordinate and $|\partial_u s(u,y)| \leq \kappa_{\mathcal{L}}$, we can take the potential function to be the anti-derivative of s, specifically, $\mathcal{L}(f(x),y) = \int_g^{f(x)} s(u,y)du$ for any $g \in \mathbb{R}$, as long as we can ensure $\int_g^{f(x)} s(u,y)du \geq C^l_{\mathcal{L}}$ for all $f(x)$ and y. Then, taking expectation over both sides of this equation and (3), we can satisfy (b). In fact, even when the function $s(u,y)$ is not differentiable everywhere with respect to u, we can still follow the general idea above to construct $\mathcal{L}$, and assumption (b) can still be satisfied with respect to the expectation of $\mathcal{L}$. For example, in Section 5.2, when $s(u,y) = \mathbb{1}\{u \leq y\} - (1-\delta)$, taking $\mathcal{L}(f(x),y) = (1-\delta)\cdot f(x) - \min(f(x)-y,0)$ will satisfy the desired condition.

▶ **Theorem 3.** *Under Assumptions (a)-(c) above, and if $\mathcal{C}$ is symmetric*[1]*, then Algorithm 1 with a suitably chosen $\eta = \mathcal{O}(\alpha/(\kappa_{\mathcal{L}}B))$ converges in $T = \mathcal{O}((C^u_{\mathcal{L}} - C^l_{\mathcal{L}})\kappa_{\mathcal{L}}B/\alpha^2)$ iterations and outputs a function f satisfying*

$$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f(x),x)s(f(x),y)]\right| \leq \alpha.$$

Proof of Theorem 3. First, since the class $\mathcal{C}$ is symmetric, that means as long as we can prove for all $c \in \mathcal{C}$,

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f_t(x),x)s(f_t(x),y)] \leq \alpha,$$

for the output $f_t(\cdot)$ of Algorithm 1, we will also have for all $c \in \mathcal{C}$,

$$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f_t(x),x)s(f_t(x),y)]\right| \leq \alpha.$$

By our assumption, there exists a potential function $\mathcal{L}$, a constant $C_{\mathcal{L}}$, and a non-negative constant $\kappa_{\mathcal{L}}$, such that for any $x \in \mathcal{X}$,

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(f_t(x),y) - \mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(f_{t+1}(x),y) \geq \mathbb{E}_{(x,y)\sim\mathcal{D}}(f_t(x)-f_{t+1}(x))s(f_t(x),y) - \mathbb{E}_{(x,y)\sim\mathcal{D}}\kappa_{\mathcal{L}}(f_t(x)-f_{t+1}(x))^2.$$

[1] Recall that symmetry is without loss of generality.

As a result, we have

$$\begin{aligned}&\mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(f_t(x),y)-\mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(f_{t+1}(x),y)\\ \geq&\mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(f_t(x),y)-\mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}\big(f_t(x)-\eta c_t(f_t(x),x),y\big)\\ \geq&\mathbb{E}_{(x,y)\sim\mathcal{D}}\eta c_t(f_t(x),x)\cdot s(f_t(x),y)-\kappa_{\mathcal{L}}\mathbb{E}_{(x,y)\sim\mathcal{D}}(\eta c_t(f_t(x),x))^2.\end{aligned}$$

The first inequality is because of (c) in our assumption and the second inequality is because of (b) in our assumption.

Given $\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[c^2(f(x),x)]\leq B$, if there exists $c_t\in\mathcal{C}$, $\mathbb{E}_{(x,y)\sim\mathcal{D}}[c_t(f_t(x),x)s(f_t(x),y)]>\alpha$

$$\begin{aligned}&\mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(f_t(x),y)-\mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(f_{t+1}(x),y)\\ \geq&\eta\mathbb{E}_{(x,y)\sim\mathcal{D}}c_t(f_t(x),x)\cdot s(f_t(x),y)-\kappa_{\mathcal{L}}\mathbb{E}_{(x,y)\sim\mathcal{D}}(\eta c_t(f_t(x),x))^2\geq\eta\alpha-\kappa_{\mathcal{L}}\eta^2B.\end{aligned}$$

Take $\eta=\alpha/(2\kappa_{\mathcal{L}}B)$, we have

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(f_t(x),y)-\mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(f_{t+1}(x),y)\geq\frac{\alpha^2}{2\kappa_{\mathcal{L}}B}.$$

Since $C^u_{\mathcal{L}}\geq\mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(f_0(x),y)$ and $\mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(f(x),y\geq C^l_{\mathcal{L}}$ for all f by assumption, each update will result in a progress at least $\frac{\alpha^2}{2\kappa_{\mathcal{L}}B}$ for each iteration if it happens, we know there are at most $2\kappa_{\mathcal{L}}B(C^u_{\mathcal{L}}-C^l_{\mathcal{L}})/\alpha^2$ updates. ◀

In the following sections, we will always assume $\mathcal{C}$ is symmetric. Also, for simplicity, we assume $\mathcal{C}$ is closed with respect to L_2-norm.

4 Application: Algorithmic Fairness in Prediction Intervals

Conformal prediction is a popular approach to uncertainty quantification in prediction models. Continuing in this vein, Romano *et al.* proposed a new group fairness criterion, *equalized coverage* [32], in which the goal is to construct a prediction interval $\mathcal{I}(x)$ that covers y with comparable probability across all protected groups of interest. More precisely, given a collection of disjoint protected demographic groups $\mathcal{A}\subset 2^{\mathcal{X}}$, the set-valued function $\mathcal{I}:\mathcal{X}\to\mathbb{R}$ provides *equalized coverage* if

$$\mathbb{P}(y\in\mathcal{I}(x)\mid x\in A)\geq 1-\delta,\ \text{ for all } A\in\mathcal{A}. \tag{4}$$

In this section, we focus for simplicity on *one-sided prediction intervals*; by applying our results twice, or to a non-conformity score, we achieve two-sided intervals (see Corollary 5 and Remark 6 below). Moreover, as usual in the multicalibration literature, our results hold even for the case of arbitrarily intersecting population subgroups. A somewhat different version of the intersectional case was also studied in [20, 24]; later, we will explain how HappyMap can be used for this as well.

In the one-sided version of (4), we let $\mathcal{I}(x)=(l_\delta(x),\infty)$ be a one-sided $(1-\delta)$-prediction interval, and study the following coverage criterion:

$$\mathbb{P}(x\in A)\cdot|\mathbb{P}(y\geq l_\delta(x)\mid x\in A)-(1-\delta)|\leq\alpha,\ \text{ for all } A\in\mathcal{A}, \tag{5}$$

which can then be rewritten as

$$|\mathbb{E}[\mathbb{1}(x\in A)\cdot(\mathbb{1}(y\geq l_\delta(x))-(1-\delta))]|\leq\alpha,\ \text{ for all } A\in\mathcal{A}.$$

A few remarks about the problem definition are in order. First, there is a trivial deterministic solution to the original equalized coverage problem in Equation (4), whether or not the groups are disjoint: just set $\mathcal{I}(x) = \mathbb{R}$.[2] Similarly, while our (one- or two-sided) version in Equation (5) rules this out, it admits the trivial randomized solution in which we take $l_\delta(x) = -\infty$ with probability $1-\delta$, and $l_\delta(x) = \infty$ with probability δ. One approach to ruling out both trivial solutions is to require a stronger condition, *multivalidity*, proposed in [20, 24]. Adopting our notation, instead of asking for a small $|\mathbb{P}(y \geq l_\delta(x) \mid x \in A) - (1-\delta)|$ as in Eq. (6), *multivalidity* requires $|\mathbb{P}(y \geq l_\delta(x) \mid x \in A, l_\delta(x)) - (1-\delta)|$ to be small. Analogous to the relationship between multicalibration and multi-accuracy, multivalidity is stronger than our requirement in Equation (6). As discussed below, we can use HappyMap for this as well. In Theorem 9, we will further provide a different argument that our algorithm is doing something nontrivial, even when we do not enforce multivalidity.

In the following, we generalize Eq. (5) to obtain the *Intersectional Equalized Coverage* requirement:

$$\sup_{c \in \mathcal{C}} \left| \mathbb{E}[c(x)(\mathbb{1}\{y \geq l_\delta(x)\} - (1-\delta))] \right| \leq \alpha, \tag{6}$$

where $\mathcal{C}$ denotes an arbitrary pre-specified collection of functions (including indicator functions of pre-specified sub-populations, and also more general continuous functions, which is typical in the multicalibration literature). For example, $\mathcal{C}$ might be the functions that can be computed by decision trees of a fixed depth. Applying HappyMap with $s(l, y) = (1-\delta) - \mathbb{1}\{l \leq y\}$, yields the following result.

▶ **Theorem 4.** *Suppose that $y \mid x$ is a continuous random variable* [3]*, the conditional density of y given x is upper bounded by K_p, and $|\mathbb{E}[y]| < C$ for some universal constant $C > 0$. In addition, suppose that $\mathbb{E}_{x \sim \mathcal{D}^x}[c^2(x)] \leq B$ for all $c \in \mathcal{C}$. Then for a suitably chosen $\eta = \mathcal{O}(\alpha/(K_p B))$, using the potential function $\mathcal{L}(l, y) = (1-\delta) \cdot l - \min(l - y, 0)$, HappyMap (Algorithm 1) converges in $T = \mathcal{O}(K_p B^2/\alpha^2)$ steps, and outputs a function $l_\delta(\cdot)$ satisfying*

$$\sup_{c \in \mathcal{C}} \left| \mathbb{E}[c(x)(\mathbb{1}\{y \geq l_\delta(x)\} - (1-\delta))] \right| \leq \alpha.$$

Proof of Theorem 4. In order to apply Theorem 3, it is sufficient to verify that the potential function $\mathcal{L}(l, y) = (1-\delta)l - \min(l - y, 0)$ satisfies the smooth-like condition, that is

$$\mathbb{E}[\mathcal{L}(l, y) - \mathcal{L}(l', y)] \geq \mathbb{E}[(l(x) - l'(x))s(f(x), y)] - \kappa_L \mathbb{E}[(l(x) - l'(x))^2].$$

In particular, we have

$$\begin{aligned}
\mathcal{L}(l, y) - \mathcal{L}(l', y) =& (1-\delta)(l - l') + \min(l' - y, 0) - \min(l - y, 0) \\
=& (1-\delta)(l - l') + (l' - l)\mathbb{1}\{l - y, l' - y < 0\} - (l - y)\mathbb{1}\{1' - y > 0 > l - y\} \\
&+ (l' - y)\mathbb{1}\{1' - y < 0 < l - y\} \\
=& (1-\delta)(l - l') + (l' - l)\mathbb{1}\{l - y < 0\} - (l' - y)\mathbb{1}\{1' - y > 0 > l - y\} \\
&+ (l' - y)\mathbb{1}\{1' - y < 0 < l - y\} \\
=& (l - l')((1-\delta) - \mathbb{1}\{l - y < 0\}) + (l' - y)(\mathbb{1}\{1' < y < l\} - \mathbb{1}\{1' > y > l\}).
\end{aligned}$$

[2] This does not mean that previous algorithms are trivial!
[3] Our analysis can be directly extended to the case where y is stored in finite precision and α is taken to be larger than the precision error

We have

$$|(l'-y)(\mathbb{1}\{1'<y<l\}-\mathbb{1}\{1'>y>l\})| \le |l-l'|\cdot(\mathbb{1}\{1'<y<l\}+\mathbb{1}\{1'>y>l\})).$$

The above inequality is due to the fact that if a $|l-l'|$ perturbation can change the sign of $l'-y$, then $|l'-y|<\eta|c(x)|$.

Assuming the conditional density of y given x is upper bounded by K_p, we then have

$$\mathbb{E}[|(l'-y)(\mathbb{1}\{\ell'<y<l\}+\mathbb{1}\{\ell'>y>l\})|] \le L\eta^2\mathbb{E}[c(x)^2] \le K_p|l-l'|^2.$$

In addition, we verify that $\mathbb{E}[\mathcal{L}(l,y)]$ has a uniform lower bound. We discuss in cases. If $l-y<0$ (i.e., $l<y$), we have $\mathcal{L}(l,y)=(1-\delta)\cdot l-\min(l-y,0)=(1-\delta)\cdot l-l+y=y-\delta\cdot l>(1-\delta)\cdot y$; if $l-y>0$ (i.e., $l>y$), we have $\mathcal{L}(l,y)=(1-\delta)\cdot l-\min(l-y,0)=(1-\delta)\cdot l>(1-\delta)\cdot y$.

Since $|\mathbb{E}[y]|<C$ for some universal constant $C>0$ by assumption, we have

$$\mathbb{E}[\mathcal{L}(l,y)] \ge -(1-\delta)C. \qquad \blacktriangleleft$$

The following Corollary is immediate from Theorem 4.

▶ **Corollary 5.** *If we apply the algorithm above (i.e. the HappyMap specialized to our current setting) to two different cutoff values $\delta/2$ and $(1-\delta/2)$ and obtain $l_{\delta/2}(x)$ and $l_{1-\delta/2}(x)$ respectively, we obtain the two-sided prediction interval $\mathcal{I}(x)=[l_{\delta/2}(x), l_{1-\delta/2}(x)]$ such that* $\sup_{c\in\mathcal{C}}\mathbb{E}[c(x)(\mathbb{1}\{y\in\mathcal{I}(x)\}-(1-\delta))]\le 2\alpha$.

▶ Remark 6. In the conformal prediction literature, there is another commonly used way to construct two-sided prediction intervals based on non-conformity scores. More specifically, the non-conformity score $m(x,y)$ is a metric that measures how the response y fails to "conform" to a prediction $h(x)$, where h is an arbitrarily fixed prediction function. For example, a popular choice of the non-conformity score for regression tasks is $m(x,y)=|y-h(x)|$. More choices of the non-conformity scores can be found in [35, 1]. Given a non-conformity score $m(x,y)$, the set $\{y: m(x,y)\le f(x)\}$ will naturally yield a two-sided interval for y. To this end, applying our method directly to $(x,\tilde{y})$ with $\tilde{y}:=m(x,y)$ will then produce a valid $(1-\delta)$ prediction interval.

Recall that multivalidity [20, 24] bounds $|\mathbb{P}(y\ge l_\delta(x)\mid x\in A, l_\delta(x))-(1-\delta)|$" for all A and value of $l_\delta(x)$. Analogous to the relationship between multicalibration and multi-accuracy, multivalidity is generally stronger than the requirement in Equation (6), resulting in potentially much longer prediction intervals (a more detailed discussion is deferred to Section E.3). By considering c of the form $c(l_\delta(x),x)=\mathbb{1}\{l_\delta(x)\in I\}$ for $I\in\mathcal{I}$, where $\mathcal{I}$ is the collection of small bins $\mathcal{I}=\{[-C,-C+\lambda],[-C+\lambda,-C+2\lambda],...,[C-2\lambda,C-\lambda],[C-\lambda,C]\}$ for some discretization level λ and C being the upper bound of $|y|$, s-Happy Multicalibration recovers multivalidity, and applying Algorithm 1 yields a new algorithm that achieves multivalidity.

The extension of (4) to intersecting sets is also considered in [17]. They define the set collection $\mathcal{A}$ *ex post facto*, after the training data have been collected, to be all sufficiently large subsets of the training set, and then enumerate over all these (exponentially many) sets. In contrast, as is typical in the multicalibration literature, we name the sets *a priori* and rely on weak learning, which can be more efficient than exhaustive search when the collection $\mathcal{A}$ has special structure, even if the collection is infinite.

HappyMap produces a prediction interval that is fair with respect to a large collection of groups. In the following, we present a result analyzing the utility, i.e., the width of the constructed prediction interval, when HappyMap is applied as *post-processing* of a high-quality, but not necessarily group-fair, initial conformal map l_0. We are agnostic regarding the source of l_0: it may be given to us or we may obtain it by splitting the sample and building a high-quality but fairness unawareness conformal mapping using any standard method.

To facilitate the analysis, we first introduce the following two definitions on quantiles. These definitions are standard and follow the literature of quantile estimation, see [6, 36] and the reference therein.

▶ **Definition 7.** *For a distribution Q with supp $Q \subset [-C, C]$ for a universal constant $C > 0$, let us denote the τ-quantile of Q by $F_\tau(Q) := \{t \in \mathbb{R} : Q((-\infty, t]) \geq \tau$ and $Q([t, \infty)]) \geq 1-\tau\}$. In case of $F_\tau(Q)$ being an interval, we write $t_{\min}(Q) = \min F_\tau(Q)$ and $t_{\max}(Q) = \max F_\tau(Q)$. Q is said to have a τ-quantile of type $q \in (1, \infty)$ if there exist constants $\alpha_Q > 0$ and $b_Q > 0$ such that for all $s \in [0, \alpha_Q]$,*

$$Q(t_{\min} - s, t_{\min}) \geq b_Q s^{q-1}$$
$$Q(t_{\max}, t_{\max} + s) \geq b_Q s^{q-1}.$$

Since we are not interested in a single distribution Q on $\mathbb{R}$ but in distributions P on $\mathcal{X} \times \mathbb{R}$, the following definition extends the previous definition to such P.

▶ **Definition 8.** *Let $p \in (0, \infty]$, $q \in (1, \infty)$ and P be a distribution on $\mathcal{X} \times \mathbb{R}$. P is said to have a τ-quantile of p-average type q, if there exists a set $\Omega_{\mathcal{X}} \subset \mathcal{X}$ such that $\mathbb{P}(x \in \Omega_{\mathcal{X}}) = 1$, supp$(P(\cdot|x)) \subset [-C, C]$ for a universal constant $C > 0$, $P(\cdot|x)$ has a τ-quantile of type q for all any $x \in \Omega_{\mathcal{X}}$, and there exist $\alpha_{P(\cdot|x)}$ and $b_{P(\cdot|x)}$ as defined in Definition 7, such that $\mathbb{E}[|b^{-p}_{P(\cdot|x)}(x) \cdot \alpha^{(1-q)p}_{P(\cdot|x)}(x)|] < \infty$.*

We then have the following result showing that applying HappyMap comes at nearly no cost in the width of the prediction interval: if the input l_0 is close to the optimal quantile function, then the HappyMap algorithm can preserve this approximation.

▶ **Theorem 9.** *Suppose the conditions in Theorem 4 hold, and P, the distribution on $\mathcal{X} \times \mathcal{Y}$, has a δ-quantile of p-average type q. Assume $pq/(p+1) \geq 2$. Let $l^*_\delta(x) \in F_\delta(P(\cdot \mid x))$ be any element in the δ-quantile of $Y \mid X = x$. If an input l_0 satisfies $\mathbb{E}_{x\sim P_X}(l_0(x) - l^*_\delta(x))^2 \leq \beta$, then there exists a universal constant C, such that the output of the HappyMap Algorithm 1, l_δ, satisfies*

$$\mathbb{E}(l_\delta(x) - l^*_\delta(x))^2 \leq C\beta.$$

Proof of Theorem 9. Recall that we use the potential function $\mathcal{L}(l, y) = (1-\delta)l - \min(l-y, 0)$ when we apply HappyMap to this current problem. According to the derivations in the Proof of Theorem 4. We have that

$$\begin{aligned}&|\mathbb{E}[\mathcal{L}(l, y)] - \mathbb{E}[\mathcal{L}(l', y)]|\\ &\leq |\mathbb{E}[(l - l')((1-\delta) - \mathbb{1}\{l - y < 0\})]| + |\mathbb{E}[(l' - y)(\mathbb{1}\{1' < y < l\} - \mathbb{1}\{1' > y > l\})]|\\ &\leq |\mathbb{E}[(l - l')((1-\delta) - \mathbb{1}\{l - y < 0\})]| + K_p \cdot \mathbb{E}[|l - l'|^2].\end{aligned}$$

Since $l^*_\delta(x) = \arg\min_l \mathbb{E}[\mathcal{L}(l, y) \mid x]$, and $\mathbb{P}(l^*_\delta(x) < y \mid X = x) = 1 - \delta$, we have

$$\mathbb{E}[\mathcal{L}(l_0, y)] - \mathbb{E}[\mathcal{L}(l^*_\delta, y)] \leq K_p \cdot \mathbb{E}|l_0(x) - l^*_\delta(x)|^2 \leq K_p\beta.$$

Moreover, by the Proof of Theorem 4, we have $\mathbb{E}[\mathcal{L}(l, y)] \leq \mathbb{E}[\mathcal{L}(l_0, y)]$, implying

$$\mathbb{E}[\mathcal{L}(l, y)] - \mathbb{E}[\mathcal{L}(l^*_\delta, y)] \leq K_p\beta.$$

Then by Theorem 2.7 in [36], and use the fact that the function L_2 norm is dominated by the function $L_{pq/(p+1)}$ norm when $pq/(p+1) \geq 2$, we have

$$\mathbb{E}[|l(x) - l^*_\delta(x)|^2] \leq C' \cdot \mathbb{E}[\mathcal{L}(l, y)] - \mathbb{E}[\mathcal{L}(l^*_\delta, y)] \leq C' K_p \beta. \qquad ◀$$

The above theorem indicates that when the conditional distribution $Y \mid X = x$ is unimodal and symmetric[4], if we have a good initialization function f_0 with approximation error $\beta = o(1)$, then the prediction interval constructed as in either Corollary 5 or Remark 6 would converge to the width of the optimal $(1 - \delta)$ prediction interval $[l^*_{\delta/2}(x), l^*_{1-\delta/2}(x)]$ where we know the distribution $Y \mid X = x$ exactly. We note that $\beta = o(1)$ can be potentially achieved by first splitting the data into two halves: l_0 is trained on the first half, and processed by HappyMap on the second half. Such data-splitting is common in the conformal prediction literature [29, 33, 30].

5 Application: Target-independent Learning

Kim et al. [27] demonstrated that, with an appropriate collection $\mathcal{C}$ of propensity reweighting functions, training a $(\mathcal{C}, \alpha)$-multicalibrated predictor on source data $\mathcal{D}_{so}$ yields efficient estimation of statistics of interest on data drawn from previously unseen target distributions $\mathcal{D}_{ta}$ [27]. In this section, we use HappyMap to extend these results to problems in target-independent learning that lie beyond the statistical estimation problems considered in [27]: target-independent prediction and uncertainty quantification. Our approach also yields a fruitful new perspective on analyzing missing data, giving new solutions to this problem.

Let us use $\mathcal{Z} = \{so, ta\}$ to indicate sampling in the source and target distributions, respectively. As in [27], we consider a joint distribution $\mathcal{D}_{joint}$ over (x, y, z) triples. The source, respectively, target, populations $\mathcal{D}_{so}$ and $\mathcal{D}_{ta}$ can be viewed as the joint distribution on (x, y), conditioning on $z = ta$ or $z = so$, respectively. We similarly use the notation $D^{\mathcal{X}}_z$ and $D^{\mathcal{Y}}_z$ to denote the marginal distributions obtained by conditioning on $z \in \mathcal{Z}$. As in [27], we assume certain functional relationship between x and y is the same on the distributions $\mathcal{D}_{so}$ and $\mathcal{D}_{ta}$, which is known as "ignorability" in the causal inference literature and "covariate shift" in machine learning. Formally, we have:

▶ **Assumption 10** (Covariate shift assumption). *For a triplet (x, y, z) drawn from $\mathcal{D}_{joint}$,*

$$\mathbb{P}_{(x,y,z)\sim\mathcal{D}_{joint}}(x, y, z) = \mathbb{P}(x)\mathbb{P}(y|x)\mathbb{P}(z|x).$$

Based on Assumption 10, we have

$$\mathbb{P}_{(x,y)\sim\mathcal{D}_{so}}(y|x) = \mathbb{P}_{(x,y)\sim\mathcal{D}_{ta}}(y|x).$$

By convention and without loss of generality, throughout this section, we assume uniform prior over $\mathcal{Z} = \{so, ta\}$, i.e. $\mathbb{P}(z = ta) = \mathbb{P}(z = so)$.

A common tool for performing covariate shift studies is *Propensity score reweighting*. The propensity score is defined to be $e(x) = \mathbb{P}(z = so|x)$, and the propensity score ratio $\frac{1-e(x)}{e(x)} = \frac{\mathbb{P}(z=ta|x)}{\mathbb{P}(z=so|x)}$ can be used to convert an expectation over $\mathcal{D}^{\mathcal{X}}_{so}$ to an expectation over $\mathcal{D}^{\mathcal{X}}_{ta}$. However, without observing samples from $\mathcal{D}_{ta}$ at training time, we cannot estimate the propensity score ratio. Kim et al. [27] proposed multicalibrating with respect to the class $\mathcal{C} = \{\frac{1-\sigma(x)}{\sigma(x)} : \sigma \in \Sigma\}$, for a family of functions Σ that (it is hoped) captures the propensity score ratios of interest. They showed that multicalibration with respect to this class (in fact,

[4] Similar assumptions have been used in conformal prediction literature, see, e.g. [29, 30].

even multi-accuracy with respect to this class) ensures that the resulting predictor provides estimation accuracy competitive with the best propensity reweighting function in the class $\mathcal{C}$, a notion they call *universal adaptability*. In the realizable, case, when the propensity ratio for the unseen target domain is in the class $\mathcal{C}$, f is guaranteed to yield a good estimate for the statistic of interest on the target domain.

5.1 Universally Adaptive Predictors under ℓ_2 Loss

As a warm-up exercise, which does not require the full power of HappyMap, we consider universal adaptivity of *predictors* $f : \mathcal{X} \to [0, 1]$ under ℓ_2 loss. This is more complex than statistical estimation under covariate shift, and it requires a more complicated class of functions $c(f(x), x)$.

Formally, our goal is to obtain a prediction f with a small estimation error

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

for $\mathcal{Y} = [0, 1]$. We note that this quantity is commonly used as a measure of the quality of a prediction function. By Assumption 10 (covariate shift), $\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[y|x] = \mathbb{E}_{(x,y)\sim\mathcal{D}_{ta}}[y|x]$), yielding

$$\begin{aligned}\mathbb{E}_{(x,y)\sim\mathcal{D}_{ta}}(f(x) - \mathbb{E}_{(x,y)\sim\mathcal{D}_{ta}}[y|x])^2 &= \mathbb{E}_{(x,y)\sim\mathcal{D}_{ta}}(f(x) - \mathbb{E}_{(x,y)\sim\mathcal{D}_{ta}}[y|x])(f(x) - y) \\ &= \mathbb{E}_{(x,y)\sim\mathcal{D}_{ta}}(f(x) - \mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[y|x])(f(x) - y).\end{aligned}$$

Using the propensity score ratio, we have

$$\begin{aligned}&\mathbb{E}_{(x,y)\sim\mathcal{D}_{ta}}(f(x) - \mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[y|x])(f(x) - y) = \\ &\qquad \mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}\left(\frac{1 - e(x)}{e(x)}(f(x) - \mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[y|x])\right)(f(x) - y).\end{aligned}$$

At training time, we cannot see the samples from $\mathcal{D}_{ta}$, so we cannot estimate $e(\cdot)$ from samples as in the classical reweighting approach. However, following [27], one can use a class $\{\frac{1-\sigma(x)}{\sigma(x)} : \sigma \in \Sigma\}$ to represent the propensity score ratios of interest, and try to find an $f : \mathbb{R} \mapsto [0, 1]$ with a small error:

$$\sup_{\sigma\in\Sigma} \mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}\left(\frac{1 - \sigma(x)}{\sigma(x)}(f(x) - \mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[y|x])\right) \cdot (f(x) - y).$$

This almost gives us the form we need in Eq. (2); it remains only to deal with $\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[y|x]$. This is accomplished by introducing a class $\mathcal{P} = \{p : \mathcal{X} \mapsto [0, 1]\}$ containing, it is hoped, a good approximation of $\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[y|x]$, and trying to find an $f : \mathbb{R} \mapsto [0, 1]$ such that

$$\left|\sup_{\sigma\in\Sigma, p\in\mathcal{P}} \mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}\left(\frac{1 - \sigma(x)}{\sigma(x)}(f(x) - p(x))\right) \cdot (f(x) - y)\right| \le \alpha, \tag{7}$$

for some small value α. In other words, we take $\mathcal{C} = \{c(f(x), x) = \pm\frac{1-\sigma(x)}{\sigma(x)}(f(x) - p(x)) : \sigma \in \Sigma, p \in \mathcal{P}\}$, where $\sigma, p \in (0, 1)$. We define the approximation error by

$$\beta_1(p) = \sqrt{\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}(p(x) - \mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[y|x])^4}$$

and

$$\beta_2(\sigma) = \sqrt{\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}(\frac{1 - \sigma(x)}{\sigma(x)} - \frac{1 - e(x)}{e(x)})^4}.$$

Although we do not use the full power of HappyMap, since we have stayed with the original mapping $(f(x) - y)$, we can nonetheless apply Theorem 3 to obtain:

▶ **Theorem 11.** *Assume $\sigma(\cdot) \in (c_1, c_2)$, where $0 < c_1 < c_2 < 1$. Then, we have $|c(f(x), x)| \leq 2(1 - c_1)/c_1 := B$. Let $\beta = \inf_{\sigma \in \Sigma, p \in \mathcal{P}} \sqrt{2B^2\beta_1(p) + 2\beta_2(\sigma)}$. Suppose we run Algorithm 1 with a suitably chosen $\eta = \mathcal{O}(\alpha/B)$, then the algorithm converges in $T = \mathcal{O}(2B/\alpha^2)$ iterations, using the potential function $\mathcal{L}(f(x), y) = 1/2(f(x) - y)^2$ and $O = [0, 1]$, which results in*

$$\Big|\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[\frac{1 - e(x)}{e(x)}(f_t(x) - \mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[y|x])(f_t(x) - y)]\Big| \leq \alpha + \beta,$$

for the output $f_t(\cdot)$ of Algorithm 1.

The proof of Theorem 11 is deferred to Section D.1.

5.2 Universally Adaptive Conformal Prediction

In this section, we apply HappyMap to achieve universally adaptive conformal prediction, wherein we train a conformal prediction model on source data while ensuring validity on unseen target distributions, in the covariate shift setting.

Recall that the goal of standard conformal prediction is to obtain a $(1 - \delta)$ prediction interval for y. In this section, for simplicity of presentation, we consider the one-sided interval $(l(x), \infty)$ and focus on the case in which the response y is continuous[5] Letting $\mathcal{D}_{ta}$ denote the unseen target distribution, our goal is to construct $l(\cdot)$ such that $\mathbb{P}_{(x,y)\sim\mathcal{D}_{ta}}(l(x) \leq y)$ is close to the desired level $(1 - \delta)$:

$$|\mathbb{P}_{(x,y)\sim\mathcal{D}_{ta}}(l(x) \leq y) - (1 - \delta)| \leq \alpha.$$

Note that the above inequality implies $\mathbb{P}_{(x,y)\sim\mathcal{D}_{ta}}(l(x) \leq y) \geq 1 - \delta - \alpha$, which is the more standard requirement used in the conformal prediction literature. By Assumption 10, one can rewrite this probability:

$$\mathbb{P}_{(x,y)\sim\mathcal{D}_{ta}}(l(x) \leq y) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}\Big(\frac{1 - e(x)}{e(x)}\mathbb{1}\{l(x) \leq y\}\Big).$$

Now, in the same spirit as in the previous section (specifically, Eq. (7)), we seek $l(\cdot)$ such that

$$\Big|\sup_{\sigma\in\Sigma}\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[\frac{1 - \sigma(x)}{\sigma(x)} \cdot (\mathbb{1}\{l(x) \leq y\} - (1 - \delta))]\Big| \leq \alpha, \tag{8}$$

for unseen $\mathcal{D}_{ta}$ and for some $\alpha > 0$. To this end, we apply HappyMap (Algorithm 1) with $\mathcal{C} = \{c(x) = \frac{1-\sigma(x)}{\sigma(x)} : \sigma \in \Sigma\}$ and the mapping $s(l, y) = \mathbb{1}\{l \leq y\} - (1 - \delta)$.

▶ **Theorem 12.** *Suppose that $y \mid x$ is a continuous random variable, the conditional density of y given x is upper bounded by K_p, and $|\mathbb{E}[y]| < C$ for some universal constant $C > 0$. Assume $\mathbb{E}_x[c^2(x)] \leq B$, $\forall c \in \mathcal{C}$, and consider the HappyMap meta-algorithm (Algorithm 1) with $s(l, y) = \mathbb{1}\{l \leq y\} - (1 - \delta)$. Then for any target distribution whose marginal density function of x, $p_{ta}(x)$, and let β satisfy $\inf_{c\in\mathcal{C}} \mathbb{E}[(c(x) - \frac{p_{ta}(x)}{p_{so}(x)})^2] \leq \beta^2$, there exists a choice of $\eta = \mathcal{O}(\alpha/(K_pB))$ such that Algorithm 1 terminates in $T = \mathcal{O}(LB/\alpha^2)$, using the potential function $\mathcal{L}(l, y) = (1 - \delta) \cdot l - \min(l - y, 0)$. The resulting function l_δ satisfies*

$$|\mathbb{P}_{(x,y)\sim\mathcal{D}_{ta}}(l_\delta(x) \leq y) - (1 - \delta)| \leq \alpha + \beta.$$

---

[5] As in Section 4, we can handle the two-sided case with two applications of our technique (Corollary 5).

The proof Theorem 12 follows a combination of the universal adaptivity argument and the analysis of Theorem 4, and is deferred to Section D.2.

▶ Remark 13. To our knowledge, Tibshirani et al. [37] were the first to consider conformal prediction under covariate shift. Their method relies on exact knowledge of $(1-e(x))/e(x)$. In contrast, we achieve a valid asymptotic coverage guarantee without exact knowledge of $(1-e(x))/e(x)$, provided something close to this score is in Σ.

Finally, we note that we can extend our results to universally adaptive equalized coverage and universal multivalidity if we enrich $\mathcal{C}$. For example, by including in $\mathcal{C}$ both $\{c(x) = \frac{1-\sigma(x)}{\sigma(x)} : \sigma \in \Sigma\}$ and the sub-population indicator functions, we can achieve equalized coverage under covariate shift.

5.3 Prediction with Missing Data

s-Happy Multicalibration can also be applied to the study of missing data. Suppose we consider learning tasks that predict response $y \in [0,1]$ based on features $x \in \mathcal{X}$. Let X denote the complete data matrix, where the i-th row x_i is the features of the i-th observation and the response vector is $Y = (y_1, ..., y_n)^\top$. Suppose the data are i.i.d., with $(x_i, y_i) \overset{i.i.d.}{\sim} \mathcal{D}$. In the missing data problem, some entries of X are missing, so we define $x_i^{(1)}$ to be the components of x_i that are observed for observation i, and $x_i^{(0)}$ the components of x_i that are missing. In addition, we let R_i be the indicator of complete cases, that is, $R_i = 1$ iff x_i is fully observed (i.e. $x_i^{(1)} = x_i$).

There are three major missing data mechanisms: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). We say a dataset is MCAR if the distribution of missingness indicators R is independent of X. For MAR, R depends on X only through its observed components, i.e., $R_i \perp\!\!\!\perp x_i^{(0)} | x_i^{(1)}$. For MNAR, the distribution of R could depend on the missing components of X. Our method, described next, applies to all three mechanisms.

Suppose our goal is to learn a predictor function f such that the test mean squared error (MSE) $\mathbb{E}_{(x,y)\sim\mathcal{D}}[(y - f(x))^2]$ is small. There are two principal methods in the literature: weighting and imputation. The weighting approach [31] minimizes the following loss function that is evaluated on the complete data,

$$\arg\min_f \sum_{i=1}^n \mathbb{1}\{R_i = 1\} w(x_i)(y_i - f(x_i))^2, \tag{9}$$

where $w(x_i)$ approximates $\frac{\mathbb{P}(x_i|R_i=1)}{\mathbb{P}(x_i)}$.

We will adapt the weighting approach to our framework, and use $\mathcal{C}$ as a class of functions where some $c(f(x), x) \in \mathcal{C}$ approximates $\frac{\mathbb{P}(x|R_i=1)}{\mathbb{P}(x)}(\mathbb{E}[y|x] - f(x))$. Since $\mathbb{E}_{(x,y)\sim\mathcal{D}}[(y - f(x))^2]$ can be equivalently written as $\mathbb{E}_{(x,y)\sim\mathcal{D}|R=1}[\frac{\mathbb{P}(x|R_i=1)}{\mathbb{P}(x)}(\mathbb{E}[y|x] - f(x)) \cdot (y - f(x))]$. Instead of minimizing the loss function as in (9), we aim to find an f such that $\sup_{c\in\mathcal{C}} \mathbb{E}_{(x,y)\sim\mathcal{D}|R=1}[c(x)(y - f(x))] \le \alpha$, where $\mathcal{C}$ is a class of functions that approximates $\frac{\mathbb{P}(x|R_i=1)}{\mathbb{P}(x)}(\mathbb{E}[y|x] - f(x))$. Applying HappyMap with such a function class $\mathcal{C}$ and $s(f, y) = f - y$, we obtain the following result.

▶ **Theorem 14.** *Suppose* $\inf_{c\in\mathcal{C}} \sqrt{\mathbb{E}[(c(f(x), x) - \frac{\mathbb{P}(x|R_i=1)}{\mathbb{P}(x)}(\mathbb{E}[y|x] - f(x))]^2} \le \beta$, *and*

$$\sup_{c\in\mathcal{C}} \mathbb{E}[c^2(f(x), x)] \le B.$$

Consider the HappyMap meta-algorithm (Algorithm 1) with $s(f, y) = f - y$. Then there exists a suitable choice of $\eta = \mathcal{O}(\alpha/B)$ such that Algorithm 1 terminates in $T = \mathcal{O}(LB/\alpha^2)$, using the potential function $\mathcal{L}(f(x), y) = 1/2(f(x) - y)^2$ and $O = [0, 1]$. The resulting function l_δ satisfies

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[(y - f(x))^2] \le \alpha + \beta.$$

The proof of Theorem 14 is deferred to Section D.3.

▶ Remark 15. The analysis above considered learning a predictor with small mean squared error in the missing data setting. The same idea can be used to other settings when there is missing data, including constructing prediction intervals when part of the observed data are missing, and enforcing multi-group fairness notions (that are studied in Section 4) with missing data.

6 Discussion and Conclusion

In this paper, we propose s-Happy Multicalibration, a generalization of multi-calibration, unifying many existing algorithmic fairness notions and target-independent learning approaches, and yielding a wide range of new applications, including a new fairness notion for uncertainty quantification, a novel technique for conformal prediction under covariate shift, and a different approach to analyzing missing data. At a higher level, we advance the field of macro-learning.

An interesting future direction is to extend s-Happy Multicalibration to (low-dimensional) representation functions. As argued recently in the deep learning community, pre-training or representation learning is a powerful tool in improving final prediction accuracy[6]. The hope is that a suitable extension of s-Happy Multicalibration for representation learning will further facilitate target-independent learning by using an extended HappyMap that post-processes the representation learned through pre-training. In addition, our current target-independent learning requires the covariate shift assumption (that is, the conditional distribution $y \mid x$ is invariant across different environments). An interesting question is to study how to use s-Happy Multicalibration for target-independent learning when there are some other types of invariance, e.g., as in invariant risk minimization (IRM) [2, 8].

The power of the "multi-X"/macro-learning framework, including s-Happy Multicalibration, stems from choosing a rich collection $\mathcal{C}$. However, a rich $\mathcal{C}$ means a harder task for the weak agnostic learner, and increased sample complexity. To fully realize the potential of macro-learning it would be helpful to understand, even from the perspective of heuristics, the trade-off between the power of $\mathcal{C}$ and the computational complexity of the weak learner.

References

1 Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint*, 2021. arXiv:2107.07511.

2 Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint*, 2019. arXiv:1907.02893.

3 Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Practical adversarial multivalid conformal prediction. *arXiv preprint*, 2022. arXiv:2206.01067.

[6] For example, in Table 5 of https://arxiv.org/pdf/2204.02937.pdf, pre-training improves accuracy increases from 60% to 90+%.

4 Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

5 Maya Burhanpurkar, Zhun Deng, Cynthia Dwork, and Linjun Zhang. Scaffolding sets. *arXiv preprint*, 2021. `arXiv:2111.03135`.

6 Andreas Christmann and Ingo Steinwart. How svms can estimate quantiles and the median. *Advances in neural information processing systems*, 20, 2007.

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

8 Zhun Deng, Frances Ding, Cynthia Dwork, Rachel Hong, Giovanni Parmigiani, Prasad Patil, and Pragya Sur. Representation via representations: Domain generalization via adversarially learned invariant representations. *arXiv preprint*, 2020. `arXiv:2006.11478`.

9 Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 66–76, 2021.

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

11 Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

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

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

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

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

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

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

18 Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. *arXiv preprint*, 2021. `arXiv:2109.05389`.

19 Parikshit Gopalan, Michael P Kim, Mihir Singhal, and Shengjia Zhao. Low-degree multicalibration. *arXiv preprint*, 2022. `arXiv:2203.01255`.

20 Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M Pai, and Aaron Roth. Online multivalid learning: Means, moments, and prediction intervals. *arXiv preprint*, 2021. `arXiv:2101.01739`.

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

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

23 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.

24 Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. *arXiv preprint*, 2022. `arXiv:2209.15145`.

25 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. PMLR, 2018.

26 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.

27 Michael P Kim, Christoph Kern, Shafi Goldwasser, Frauke Kreuter, and Omer Reingold. Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences*, 119(4):e2108097119, 2022.

28 Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint*, 2016. `arXiv:1609.05807`.

29 Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96, 2014.

30 Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2021.

31 Lingling Li, Changyu Shen, Xiaochun Li, and James M Robins. On weighting approaches for missing data. *Statistical methods in medical research*, 22(1):14–30, 2013.

32 Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel J Candès. With malice towards none: Assessing uncertainty via equalized coverage. *arXiv preprint*, 2019. `arXiv:1908.05428`.

33 Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32:3543–3553, 2019.

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

35 Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

36 Ingo Steinwart and Andreas Christmann. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225, 2011.

37 Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

A Additional Literature Review

Hébert-Johnson et al. introduced the method for obtaining efficient multi-calibrated predictors in the batch setting [22], multicalibration in the online setting has a long history in statistics (not with this name, however; see, for example, [16, 34] and the references therein). In parallel with [22], [25] introduced notions of multi-group fairness, including statistical parity subgroup fairness and false positive subgroup fairness. [23] extends the notion of multicalibration to higher moments in the batch setting, and [20] provides an online solution. In particular, taking $s(f(x), y) = f(x)^k - y^k$ recovers the raw moment version of moment multicalibration; they also consider calibrating high-order central moments using a novel and different technique. Moreover, multicalibration has also been applied in the batch setting to solve several problems of the same flavor: fair ranking [14]; *ommiprediction*, which is roughly learning a predictor that, for a given class of loss functions, can be post-processed to minimize loss for any function in the class [18]; and providing an alternative to propensity scoring for the purposes of generalization to future populations [27]. We summarize those previous concepts in the following table and show their relationship with our new notion.

Table 1 Summary of the relationship with prevoius concepts.

Relationship between previous concpets and the s-Happy Multicalibration		
Concept name	choice of $c(f(x), x)$	choice of $s(f(x), y)$
Multiaccuracy [26]	$c(x)$	$f(x) - y$
Multicalibration [21]	$c(f(x), x)$	$f(x) - y$
Low-degree Multicalibration [19]	$\tilde{c}(x)w(f(x))$ for some function $\tilde{c}$	$f(x) - y$
Minimax Group Fairness [9]	$\mathbb{1}\{x \in G\}$ for demographic group G	non-negative loss
Omni-predictor [18] (more in Appendix E.1)	$\mathbb{1}\{f(x) \in I_v\}$ and $c(x)\mathbb{1}\{f(x) \in I_v\}$	$f(x) - y$

B Omitted Details for Section 3

B.1 Details for the Sample Version Algorithm

In this section, we will provide the corresponding sample version algorithm for Algorithm 1 and its formal theoretical guarantee. Before that, similarly as in [26], let us introduce an **informal** but handy concept. For a dataset D, we use $\mathbb{E}_{(x,y)\sim D}$ to denote the empirical expectation over D. Throughout this section, we assume $\sup_{c\in\mathcal{C}} \|c\|_\infty \leq B_0$ for some universal constant $B_0 > 0$.

▶ **Definition 16** (Dimension of a function class). *We use $d(\mathcal{C})$ to denote the dimension of an agnostically learnable class $\mathcal{C}$, such that if the sample size $m \geq C_1 \frac{d(\mathcal{C})+\log(1/\delta)}{\alpha^2}$ for some universal constant $C_1 > 0$, then the random samples S_m from $\mathcal{D}$ guarantee uniform convergence over $\mathcal{C}$ with error at most α with failure probability at most δ, that is, for any fixed f and fixed s with $\|s\|_\infty \leq C_2$ for some universal constant $C_2 > 0$:*

$$\sup_{c(f(x),x)\in\mathcal{C}} \left| \mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f(x),x)s(f(x),y)] - \mathbb{E}_{(x,y)\sim S_m}[c(f(x),x)s(f(x),y)] \right| \leq \alpha.$$

Examples of such upper bounds for this dimension include metric entropy for $\mathcal{C}$.

Given the above preparations, let us provide the following sample version algorithm.

Algorithm 2 HappyMap.

Input: Step size $\eta > 0$, bound $T \in \mathbb{N}_+$ on number of iterations, initial predictor $f_0(\cdot)$, distribution $\mathcal{D}$, convex set $O \subseteq \mathbb{R}$, mapping $s : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$, T validation datasets $D_0, \cdots, D_T$, each one with m samples
Set $t = 0$
while $t < T$ *and* $\exists c_t \in \mathcal{C} : \mathbb{E}_{(x,y)\sim D_t}[c(f_t(x), x)s(f_t(x), y)] > 3\alpha/4$ **do**
  Let c_t be an arbitrary element of $\mathcal{C}$ satisfying the condition in the while statement
  $\forall x \in \mathcal{X}$, $f_{t+1}(x) = \Pi_O\big[f_t(x) - \eta \cdot c(f(x), x)\big]$
  $t = t + 1$
end
Output: $f_t(\cdot)$

▶ **Theorem 17.** *Under our assumptions, if $\mathcal{C}$ is symmetric, suppose we run Algorithm 2 with a suitably chosen $\eta = \mathcal{O}(\alpha/(\kappa_{\mathcal{L}} B))$ and $m = \Omega(T \cdot \frac{d(\mathcal{C})+\log(1/\delta)}{\alpha^2})$, then with probability at least $1-\delta$, the algorithm converges in $T = \mathcal{O}((C^u_{\mathcal{L}} - C^l_{\mathcal{L}})\kappa_{\mathcal{L}} B/\alpha^2)$, which results in*

$$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f(x),x)s(f(x),y)]\right| \leq \alpha$$

for the final output f of Algorithm 2.

Proof. We can take a suitably chosen $m = \Omega(T \cdot \frac{d(\mathcal{C})+\log(1/\delta)}{\alpha^2})$, such that for all $t \in [T]$,

$$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f_t(x),x)s(f_t(x),y)] - \mathbb{E}_{(x,y)\sim\mathcal{D}_t}[c(f_t(x),x)s(f_t(x),y)]\right| \leq \alpha/4.$$

Thus, whenever Algorithm 2 updates, we know

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f_t(x),x)s(f_t(x),y)] \geq \alpha/2.$$

Thus, the progress for the underlying potential function is at least $\frac{\alpha^2}{8\kappa_{\mathcal{L}} B}$. Following similar proof of Algorithm 1, as long as T satisfying $(C^U_{\mathcal{L}} - C^l_{\mathcal{L}})/\frac{\alpha^2}{8\kappa_{\mathcal{L}} B} < T$, we know Algorithm 2 provides a solution f such that

$$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f(x),x)s(f(x),y)]\right| \leq \alpha. \quad ◀$$

▶ Remark 18. The sample version results for all the theorems in Section 4,5 can naturally follow the above theorem. We will not reiterate that in our paper.

C Omitted Details for Section 4

C.1 Proof of Remark 5

Proof. Since $\mathbb{1}\{y \in \mathcal{I}(x)\} = \mathbb{1}\{y \leq l_{\delta/2}(x)\} - \mathbb{1}\{y \leq l_{1-\delta/2}(x)\}$ and $(1-\delta) = (1-\delta/2)-\delta/2$, we have

$$\mathbb{1}\{y \in \mathcal{I}(x)\} - (1-\delta) = [\mathbb{1}\{y \leq l_{\delta/2}(x)\} - \delta/2] - [\mathbb{1}\{y \leq l_{1-\delta/2}(x)\} - (1-\delta/2)].$$

Therefore, $\sup_{c\in\mathcal{C}} \mathbb{E}[c(x)(\mathbb{1}\{y \in \mathcal{I}(x)\} - (1-\delta))] \leq \sup_{c\in\mathcal{C}} \mathbb{E}[c(x)(\mathbb{1}\{y \leq l_{\delta/2}(x)\} - \delta/2] + \sup_{c\in\mathcal{C}} \mathbb{E}[c(x)(\mathbb{1}\{y \leq l_{1-\delta/2}(x)\} - (1-\delta/2))] \leq 2\alpha$ ◀

D Omitted Details for Section 5

D.1 Proof of Theorem 11

Proof. $\mathcal{C} = \{c(f(x),x) = \pm\frac{1-\sigma(x)}{\sigma(x)}(f(x)-p(x)) : \sigma \in \Sigma, p \in \mathcal{P}\}$, where $\sigma, p \in (0,1)$ and$\mathcal{L}(f(x),y) = 1/2(f(x)-y)^2$ and $O = [0,1]$. Then, we directly apply the proof of Theorem 3, and we can obtain for all $\sigma \in \Sigma$ and $p \in \mathcal{P}$,

$$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}[\frac{1-\sigma(x)}{\sigma(x)}(f_t(x)-p(x))(f_t(x)-y)]\right| \leq \alpha.$$

To obtain the final guarantee, for any $\hat{f}$, let us denote $\delta_{\sigma,\hat{f},p} = \phi^*_{\hat{f}}(x) - \phi_{\sigma,\hat{f},p}(x)$, where $\phi^*_{\hat{f}}$ is the underlying truth $\frac{1-e(x)}{e(x)}(\hat{f}(x) - \mathbb{E}(y|x))$ and $\phi_{\sigma,\hat{f},p}(x) = \frac{1-\sigma(x)}{\sigma(x)}(\hat{f}(x)-p(x))$.

$$\begin{aligned}\mathbb{E}_{(x,y)\sim\mathcal{D}_{ta}}[\big(\hat{f}(x)-\mathbb{E}(y|x)\big)^2] &= \mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[\frac{1-e(x)}{e(x)}(\hat{f}(x)-\mathbb{E}(y|x))(\hat{f}(x)-y)]\\ &= \mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}(\phi_{\sigma,\hat{f},p}(x)(\hat{f}(x)-y))+\mathbb{E}_s((\phi^*_{\hat{f}}(x)-\phi_{\sigma,\hat{f},p}(x))(\hat{f}(x)-y))\\ &\le \alpha+\sqrt{\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}(\hat{f}-y)^2\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}\delta^2_{\sigma,\hat{f},p}}.\end{aligned}$$

We should notice that we can choose suitable $\sigma\in\Sigma, p\in\mathcal{P}$ and the above bound can be written as:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_{ta}}[\Big(\hat{f}(x)-\mathbb{E}(y|x)\Big)^2]\le\alpha+\min_{\sigma\in\Sigma,p\in\mathcal{P}}\sqrt{\mathbb{E}_s(\hat{f}-y)^2\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}\delta^2_{\sigma,\hat{f},p}}$$

Now, the only task is to bound $\min_{\sigma\in\Sigma,p\in\mathcal{P}}\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}\delta^2_{\sigma,\hat{f},p}$. Let us denote the bias

$$\beta_{st}=\min_{\sigma\in\Sigma}\sqrt{E_{(x,y)\sim\mathcal{D}_{so}}(\frac{1-\sigma(x)}{\sigma(x)}-\frac{1-e(x)}{e(x)})^4}.$$

Thus, we have

$$\min_{\sigma\in\Sigma,p\in\mathcal{P}}\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}\delta^2_{\sigma,\hat{f},p}\le 2\beta_{st}\sqrt{\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}(\hat{f}-\mathbb{E}(y|x))^4}+$$
$$2\min_{p\in\mathcal{P}}\sqrt{\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[\frac{1-\bar{\sigma}(x)}{\bar{\sigma}(x)}]^4}\sqrt{\mathbb{E}_s[p(x)-\mathbb{E}(y|x)]^4}$$

where $\bar{\sigma}$ is the one reach the min of β_{st}. So, combining with the fact that $f,p,y\in[0,1]$, we can obtain the final result. ◀

D.2 Proof of Theorem 12

Proof. Following Theorem 3 and proof of Theorem 4 (which shows that the function $s(l,y)=\mathbb{1}\{l\le y\}-(1-\delta)$ satisfies the conditions (b)-(c) of Theorem 3), we have

$$\Big|\sup_{c\in\mathcal{C}}\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[c(x)\cdot(\mathbb{1}\{l(x)\le y\}-(1-\delta))]\Big|\le\alpha.$$

Since $\mathbb{P}_{(x,y)\sim\mathcal{D}_{ta}}(l(x)\le y)=\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}\Big(\frac{1-e(x)}{e(x)}\mathbb{1}\{l(x)\le y\}\Big)$, we then have

$$\begin{aligned}&|\mathbb{P}_{(x,y)\sim\mathcal{D}_{ta}}(l(x)\le y)-(1-\delta)|\\ &\quad\le\alpha+|\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[(c(x)-\frac{1-e(x)}{e(x)})\cdot(\mathbb{1}\{l(x)\le y\}-(1-\delta))]|\\ &\quad\le\alpha+\sqrt{\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[(c(x)-\frac{1-e(x)}{e(x)})^2]\cdot\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[(\mathbb{1}\{l(x)\le y\}-(1-\delta))^2]}\\ &\quad\le\alpha+\beta,\end{aligned}$$

where the last inequality uses the fact that $|\mathbb{1}\{l(x)\le y\}-(1-\delta)|<1$. ◀

D.3 Proof of Theorem 14

Proof. Following Theorem 3 and the fact that the potential function $\mathcal{L}(f(x),y)=1/2(f(x)-y)^2$ satisfies the conditions (b)-(c) of Theorem 3), we have

$$\Big|\sup_{c\in\mathcal{C}}\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[c(x)\cdot(y-f(x))]\Big|\le\alpha.$$

Since $\mathbb{E}_{(x,y)\sim\mathcal{D}}[(y-f(x))^2]$ can be equivalently written as $\mathbb{E}_{(x,y)\sim\mathcal{D}|R=1}[\frac{\mathbb{P}(x|R_i=1)}{\mathbb{P}(x)}(\mathbb{E}[y|x]-f(x))\cdot(y-f(x))]$, we then have

$$\begin{aligned}&\mathbb{E}_{(x,y)\sim\mathcal{D}}[(y-f(x))^2]\\&\qquad\le\alpha+|\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[(c(f(x),x)-\frac{\mathbb{P}(x\mid R_i=1)}{\mathbb{P}(x)}(\mathbb{E}[y|x]-f(x)))\cdot(y-f(x))]|\\&\qquad\le\alpha+\sqrt{\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[(c(x)-\frac{\mathbb{P}(x\mid R_i=1)}{\mathbb{P}(x)}(\mathbb{E}[y|x]-f(x)))^2]\cdot\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[(y-f(x))^2]}\\&\qquad\le\alpha+\beta,\end{aligned}$$

where the last inequality uses the fact that after projection to $O=[0,1]$, $|f(x)-y|<1$. ◀

E Relation to the existing literature

E.1 Relation to Omnipredictors

In a recent work, [18] proposed to use a notion of multi-calibration to create an $(\mathcal{L},\mathcal{C})$ omnipredictor – a predictor that could be used to optimize any loss in a family $\mathcal{L}$, that is, the output of such a predictor can be post-processed to produce a small loss compared with any hypothesis from a given function class $\mathcal{G}$.

We first state the definition of multi-calibration in [18].

▶ **Definition 19** ([18]). *Let $\mathcal{D}$ be a distribution on $\mathcal{X}\times\{0,1\}$. The partition $\mathcal{S}=\{S_1,...,S_m\}$ of $\mathcal{X}$ is α-multicalibrated for $\mathcal{G}$, $\mathcal{D}$ if for every $i\in[m]$ and $c\in\mathcal{G}$, the conditional distribution $\mathcal{D}_i=\mathcal{D}\mid x\in S_i$ satisfies*

$$|\mathrm{Cov}_{\mathcal{D}_i}(c(x),y)|\le\alpha. \tag{10}$$

While this notion is seemingly different from the Definition 2. [18] showed that when $c(x)$ are Boolean functions, Definition 19 is equivalent to mean-multicalibration [23]. Our next theorem further extends this relationship to real-valued function c, and shows that our generalized notion in Equation (2) in fact implies (10).

▶ **Theorem 20.** *Consider a function class $\mathcal{G}$. Denote $\Lambda=\{\lambda,3\lambda,5\lambda,...,1-\lambda]\}$and $I_v=[v-\lambda,v+\lambda]$. Take $\mathcal{I}_\Lambda=\{I_v\}_{v\in\Lambda}$ and we let $\mathcal{C}$ include $\mathbb{1}\{f(x)\in I_v\}$ and $c(x)\mathbb{1}\{f(x)\in I_v\}$ for $c\in\mathcal{G}$. Assuming $|c(x)|<L$ and $\mathbb{P}(x\in I_v)\ge\gamma$.*

If a predictor f satisfies α-multi-calibration with respect to $\mathcal{C}$ in Equation (2) with $s(f(x),y)=f(x)-y$, then the set partition $\mathcal{S}_\Lambda$ defined by this predictor f satisfies $(\frac{(L+1)\alpha}{\gamma}+\lambda)$-multi-calibration with respect to $\mathcal{G}$ in Equation (10).

Proof. Recall that $\mathcal{G}$ include $\mathbb{1}\{f(x)\in I_v\}$ and $c(x)\mathbb{1}\{f(x)\in I_v\}$. Let us define $S_v=\{x:f(x)\in I_v\}$. Then (2) implies

$$\begin{aligned}|\mathbb{E}[\mathbb{1}\{x\in S_v\}\cdot(f(x)-y)]|&<\alpha,\\|\mathbb{E}[\mathbb{1}\{x\in S_v\}c(x)\cdot(f(x)-y)]|&<\alpha.\end{aligned}$$

Then

$$\begin{aligned}&|\mathbb{E}[\mathbb{1}\{x\in S_v\}c(x)\cdot(y-\mathbb{E}[y\mid x\in S_v])]|\le\alpha+|\mathbb{E}[\mathbb{1}\{x\in S_v\}c(x)\cdot(f(x)-\mathbb{E}[y\mid x\in S_v])]|\\&\le\alpha+|\mathbb{E}[\mathbb{1}\{x\in S_v\}c(x)\cdot(\mathbb{E}[f(x)-y\mid x\in S_v])]|+|\mathbb{E}[\mathbb{1}\{x\in S_v\}c(x)\cdot(f(x)-\mathbb{E}[f(x)\mid x\in S_v])]|\end{aligned}$$

For the second term, we have

$$|\mathbb{E}[\mathbb{1}\{x \in S_v\}c(x) \cdot (\mathbb{E}[f(x) - y \mid x \in S_v])]| \le |\mathbb{E}[\mathbb{1}\{x \in S_v\}|c(x)| \cdot \frac{\alpha}{\mathbb{P}(x \in S_v)}]| \le \alpha L.$$

For the third term, since $|f(x) - \mathbb{E}[f(x) \mid x \in S_v]| < \lambda$ for any $x \in S_v$, we have

$$|\mathbb{E}[\mathbb{1}\{x \in S_v\}c(x) \cdot (f(x) - \mathbb{E}[f(x) \mid x \in S_v])]| \le L\lambda\mathbb{P}(x \in S_v).$$

Combining all the pieces, we have

$$|\mathbb{E}[\mathbb{1}\{x \in S_v\}c(x) \cdot (y - \mathbb{E}[y \mid x \in S_v])]| \le (L+1)\alpha + L\lambda\mathbb{P}(x \in S_v),$$

implying

$$\begin{aligned} |\mathrm{Cov}(c(x), y \mid x \in S_v)| &= |\mathbb{E}[c(x)(y - \mathbb{E}[y \mid x \in S_v]) \mid x \in S_v]| \\ &= \frac{|\mathbb{E}[c(x)(y - \mathbb{E}[y \mid x \in S_v]) \cdot \mathbb{1}\{x \in S_v\}]|}{\mathbb{P}(x \in S_v)} \\ &\le \frac{(L+1)\alpha}{\gamma} + \lambda. \end{aligned}$$ ◀

E.2 Generalized Statistical Parity Subgroup Fairness

Kearns et al. [25] consider multi-parity (also called statistical parity subgroup fairness), in which, given a pre-specified collection of possibly intersecting demographic subgroups, the demographics of those assigned a positive outcome are the same as the demographics of the population as whole. In this section, we show that the constraints of statistical parity and equality of false positive (and false negative) rates can be expressed by appropriate choice of mappings in HappyMap. The multi-group versions of these guarantees can be achieved trivially via a constant function. In this section, in contrast to [25], we do not give guarantees for our algorithms. We state our observations in anticipation of extensions of HappyMap to constrained optimization.

Following [25], we aim to find randomized classifiers that satisfy this fairness notion. We consider binary classification, where $\mathcal{Y} = \{0, 1\}$. We let $\mathcal{U}(x) = \mathbb{1}\{U < f(x)\}$ for uniform random variable $U \sim U(0, 1)$, where $f(x) \in [0, 1]$ denotes a fitted probability function such that $\mathcal{U}(x) \sim Bern(f(x))$. Suppose there is a collection of demographic group indicator functions $\mathscr{G} = \{g : \mathcal{X} \to \{0, 1\}$, where $g(x) = 1$ indicates that an individual with features x is in group g. The multi-parity requirement says that, for all $g \in \mathscr{G}$,

$$|\mathbb{P}(g(x) = 1) \cdot (\mathbb{P}(\mathcal{U}(x) = 1) - \mathbb{P}(\mathcal{U}(x) = 1 \mid g(x) = 1))| \le \alpha. \tag{11}$$

This inequality can be rewritten as

$$|\mathbb{E}[(\mathbb{1}\{g(x) = 1\} - \mathbb{P}(g(x) = 1)) \cdot \mathbb{1}\{U < f(x)\}]| \le \alpha.$$

This notion only considers fairness auditors that are indicator functions where the range of g is $\{0, 1\}$. In the following, we generalize multiparity to allow real-valued g's:

$$\sup_{g \in \mathcal{G}} \Big|\mathbb{E}[(g(x) - \mathbb{E}[g(x)]) \cdot \mathbb{1}\{U < f(x)\}]\Big| \le \alpha, \tag{12}$$

where $\mathcal{G} \subset [0, 1]^{\mathcal{X}}$ denotes a collection of auditors with possibly continuous ranges. As suggested by [26], this allows the inclusion of more flexible statistical tests.

By taking $c(f(x), x) = g(x) - \mathbb{E}[g(x)]$, and $s(f, y) = \mathbb{1}\{U < f\}$, an application of HappyMap to this setting (with the slight modification that the **while** condition should check $\exists g \in \mathcal{G} : \mathbb{E}_{U,(x,y)\sim\mathcal{D}}[(g(x) - \mathbb{E}[g(x)]) \cdot (\mathbb{1}\{f_t(x) > U\})] > \alpha$, where the expectation is taken over the randomness of both the data and U) yields the following guarantee.

▶ **Theorem 21.** *Suppose that $\mathbb{E}_{x\sim\mathcal{D}^{\mathcal{X}}}[c^2(x)] \le B$ for all $c \in \mathcal{C}$. Taking a suitably chosen $\eta = \mathcal{O}(\alpha/B)$, then HappyMap (Algorithm 1), when run with $T = \mathcal{O}(B/\alpha^2)$ and the potential function $\mathcal{L}(f(x), y) = (1 - \delta) \cdot f(x) - \mathbb{E}_U[\min(f(x) - U, 0)]$ outputs a function $l(\cdot)$ satisfying*

$$\sup_{g\in\mathcal{G}} \mathbb{E}[(g(x) - \mathbb{E}[g(x)]) \cdot \mathbb{1}\{U < f(x)\}] \le \alpha.$$

Proof of Theorem 21. The proof of Theorem 21 directly follows the proof of Theorem 4, by replacing the y there by U. The same derivation still apply to this case, and we omit the details here. ◀

▶ Remark 22. We can similarly generalize the false positive subgroup fairness notion considered in [25], that is, $|\mathbb{P}(g(x) = 1, y = 0) \cdot (\mathbb{P}(\mathcal{U}(x) = 1) - \mathbb{P}(\mathcal{U}(x) = 1 \mid g(x) = 1))| \le \alpha$, to the case where we consider all groups that can be identified by certain classifiers: $\sup_{g\in\mathcal{G}} \Big|\mathbb{E}[(g(x) - \mathbb{E}[g(x) \mid y = 0]) \cdot \mathbb{1}\{U < f(x)\} \mid y = 0]\Big| \le \alpha$. Applying HappyMap to the distribution $\mathcal{D} \mid y = 0$ yields a classifier that satisfies this constraint.

E.3 Comparison between marginal coverage and calibrated coverage

In this section, we compare the notion in Eq. (6) and multivalidity, in the setting where $\mathcal{C} = \{1\}$. In this special case, we call them marginal coverage and calibrated coverage respectively. We first introduce some basics and definitions as below.

Suppose y is a continuous random variable. For $\delta > 0$, we aim to construct the $(1 - \delta)$ prediction interval $C_n(x) = (l(x), u(x))$ based on n i.i.d. samples $(x_i, y_i) \sim P_{x,y}$, the joint distribution of (x, y). Following the set-up in [20, 3], the prediction interval $C_n(x)$ is constructed based on a non-conformity score $m(x, y)$, eg. $|y - h(x)|$ for an arbitrary prediction function h, and a conformity threshold function $q(x)$. Then $C_n(x) = \{y : m(x, y) < q(x)\}$.

The marginal coverage guarantee asks that $C_n(x)$ satisfies

$$\mathbb{P}(y \in C_n(x)) = 1 - \delta, \tag{13}$$

for all $P_{x,y}$, where the probability is taken over all training samples and $(x, y) \sim P_{x,y}$.

The calibrated coverage guarantee asks that $C_n(x)$ satisfy

$$\mathbb{P}(y \in C_n(x) \mid q(x) = q) = 1 - \delta, \tag{14}$$

for all $P_{x,y}$ and almost all q. We then have the following theorem.

▶ **Theorem 23.** *Suppose that an interval $C_n(x)$ constructed based on n samples has $1 - \delta$ calibrated coverage in the sense of condition (14). Then for any $P_{x,y}$, with probability 1,*

$$|C_n(x)| = \infty,$$

at almost all points q that $q(x) = q$ is non-atom of P_q, where P_q denotes the marginal distribution of $q(x)$.

Here, a point q is a non-atom for P_q if q is in the support of P_q and if $\mathbb{P}(B(q, \epsilon)) \to 0$ as $\epsilon \to 0$, where $B(q, \epsilon)$ is the Euclidean ball centred at q with radius ϵ.

We note that the key part of this Theorem is the "finite-sample" and view $C_n(\cdot)$ as an interval produced by applying a certain method to the training samples. When a certain method produces an interval from n samples and needs to satisfy (14) for a large class of distributions, Theorem 23 suggests such an interval has to have infinite length. In contrast, the standard conformal prediction methods, such as the split conformal [35], produces a finite-length interval that satisfies Equation (13) for all $P_{x,y}$.

▶ Remark 24. In [3], some regularity conditions on $P_{x,y}$ are imposed, making the class of distributions smaller and therefore Theorem 23 does not apply. This suggests that some regularity conditions are necessary for calibrated coverage, but not for marginal coverage. Furthermore, under certain additional regularity conditions, calibrated coverage would produce short intervals and rule out long ones. For example, let us assume $y \in [-1, 1]$. Taking the nonconformity score $m(x, y) = |y|$, we have that all intervals that satisfy (14) must have length less than 2. This is due to the fact that $q(x)$ has to be smaller than 1, otherwise conditioning on $q(x) = q$ for $q \geq 1$ would make the left hand side of (14) equal to 1. In contrast, the intervals that satisfy (13) might have infinite length. For example, taking $C_n(x) = \mathbb{R}$ for $(1 - \delta) \times 100\%$ samples of x, and $C_n(x) = 0$ otherwise. Such a $C_n(x)$ satisfies (13), but not the stronger requirement (14).

▶ Remark 25. We would also like to point out that Theorem 23 only holds for all non-atomic points in P_q. It suggests that discretization of $q(x)$ is necessary to achieve calibrated coverage (or multivalid) under weak assumptions.

Proof of Theorem 23. The key idea of this proof follows [29] and [17]. Let us prove the theorem for any given distribution P_q. For ease of presentation, we omit the subscript q and write the distribution as P. We first fix arbitrary $\epsilon, B > 0$.

For a non-atom point x_0, there exists a $\delta_0 > 0$ such that $\mathbb{P}_P(q(x) \in B(q(x_0), \delta_0) < \epsilon_n$ where $\epsilon_n = 2\{1 - (1 - \epsilon^2/8)^{1/n}\}$. We also let $B_0 = \frac{B}{2(1-\delta)}$.

We then define a new distribution Q, by letting

$$\mathbb{P}_Q(A) = \mathbb{P}_P(A \cap S^c) + \mathbb{P}_U(A \cap S),$$

where $S = \{(x, y) : q(x) \in B(q(x_0), \delta_0)\}$ and U is uniform in $\{(x, y) : q(x) \in B(q(x_0), \delta_0), |y| \leq B_0\}$ and has total mass $\mathbb{P}_P(S)$.

By definition, we can verify $TV(P, Q) \leq \epsilon_n$, and therefore $TV(P^n, Q^n) \leq \epsilon$.

Since (14) also holds for Q, we then have

$$\int_{C_n(x)} dQ(y \mid q(x)) \geq 1 - \delta,$$

for all x such that $q(x) \in B(q(x_0), \delta_0)$. For such x's, we have $|dQ(y \mid q(x))| \leq \frac{1}{2B_0}$, implying $|C_n(x)| \geq 2B_0(1 - \delta) = B$. Therefore, for such x's

$$\mathbb{P}_Q(|C_n(x)| > B) = 1,$$

and it follows that

$$\mathbb{P}_P(|C_n(x)| > B) \geq \mathbb{P}_Q(|C_n(x)| > B) - \epsilon = 1 - \epsilon.$$

Since ϵ and B are arbitrary, we complete the proof. ◀