Accuracy vs. Accuracy: Computational Tradeoffs Between Classification Rates and Utility

Noga Amit[*] MIT

Omer Reingold[*] Stanford

Guy N. Rothblum[*] Apple

Abstract

We revisit the foundations of fairness and its interplay with utility and efficiency in supervised learning settings where training labels are richer than binary outcomes, such as risk estimates (probabilities), individual types, or rankings. We introduce new notions of accurate classification rates for subgroups in the population, defined by comparing the induced positive-classification rates to those under the Bayes-optimal rich predictor. Our main contributions are computational impossibility results: we show that simultaneously achieving these rate-accuracy guarantees and natural desiderata such as calibration or loss minimization is, in some cases, computationally infeasible, even when training examples are labeled by the Bayes-optimal rich predictor. Unlike prior impossibility results in this area, these desiderata are simultaneously satisfied by the Bayes-optimal predictor, and each can be achieved efficiently in isolation.

1 Introduction

Many deployed ML systems follow a predict-then-decide pipeline. A model outputs a prediction about an individual that is richer than a Boolean outcome—for example, a risk estimate (a probability), a distribution over types, or a ranking—and a downstream decision rule maps this prediction to a binary action (approve/deny, admit/reject, intervene/not intervene). In such settings, it is natural to evaluate not only predictive quality, but also the binary decisions induced by the prediction, including their aggregate behavior on subpopulations.

A simple lending example illustrates a gap between predictor-level and decision-level guarantees. Suppose a predictor outputs repayment probabilities and the deployed decision rule approves loans above a fixed cutoff. A predictor can be well-calibrated and yet induce approval rates that differ substantially from those induced by the Bayes-optimal predictor: calibration is an on-average guarantee about predictions, but approval rates depend on how many individuals end up on each side of the cutoff. This motivates asking when guarantees about rich predictions translate into guarantees about the downstream decisions they induce.

Why consider rich predictions at all, rather than learning the binary decision directly? In many applications, the prediction is meant to be reusable across downstream choices: the same learned predictor may be used for multiple downstream tasks, decision rules may change over time (due to constraints, business goals, or regulation), and audits may target both predictive properties (e.g., calibration) and decision-level outcomes (e.g., positive-decision rates). This modularity motivates treating the rich prediction itself as the object to be learned: it preserves information that can support future decisions and audits, whereas a binary decision commits to a single downstream use.

Setting. Individuals have features x drawn from a distribution $\mathcal{D}_{\mathcal{X}}$ over $\mathcal{X}$. We let $\mathcal{C}$ denote a collection of (possibly intersecting) subpopulations, which specifies the groups on which we require

[*]Authors listed alphabetically.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

guarantees. Throughout, our multigroup notions are required to hold for every $S \in \mathcal{C}$. Nature specifies a Bayes-optimal rich predictor $R^* : \mathcal{X} \to \mathcal{Y}$ (where $\mathcal{Y}$ may represent probabilities, rankings, or distributions over a finite set of abstract categories, which we call types). When the rich label is a single probability for a binary outcome, $R^*(x)$ is simply $p^*(x) = \Pr[y = 1 \mid x]$. The learner receives samples labeled by $R^*(x)$ and outputs a (rich) predictor $\tilde{R} : \mathcal{X} \to \mathcal{Y}$.

We recall two standard multigroup notions from the algorithmic fairness literature. A predictor $\tilde{R}$ is *multi-accurate* if, for every group $S \in \mathcal{C}$, its average predicted type distribution is close coordinate-wise to the average type distribution specified by R^*. Informally, for every $S \in \mathcal{C}$ and type t, the averages of $R^*_t(x)$ and $\tilde{R}_t(x)$ over S are close. *Multi-calibration* [Hébert-Johnson et al., 2018] is stronger: it requires such agreement to hold even after conditioning on the prediction made by $\tilde{R}$.

A *decision rule* is a randomized map $\rho : \mathcal{Y} \to \{0, 1\}$ that converts a rich prediction into a binary decision, whereas an *action function* is a randomized map $h : \mathcal{X} \to \{0, 1\}$ that makes a binary decision directly from features (action functions are also commonly called *binary classifiers*, and sometimes *policies*). Given a loss function ℓ, we denote the Bayes-optimal loss-minimizing action function by h^*_ℓ. Namely, $h^*_\ell(x)$ denotes the action minimizing expected loss under $R^*(x)$.

1.1 Our contributions

We introduce two accuracy notions that assess a learned rich predictor via the subpopulation-level positive-decision/classification rates it induces. Along the way, we relate our new notions to standard multigroup guarantees such as multi-accuracy and multi-calibration. Highlights include:

Decision-accuracy vs. calibration. Let $\rho : \mathcal{Y} \to \{0, 1\}$ be a decision rule that (randomly) maps rich predictions to binary decisions. We formalize *decision-accuracy* by requiring that for each group, the positive-decision rate induced by applying ρ to a learned predictor $\tilde{R}$ is close to the rate induced by applying ρ to the Bayes-optimal rich predictor R^*. That is, for every $S \in \mathcal{C}$,

$$\left|\mathbb{E}[\rho(\tilde{R}(x)) \mid x \in S] - \mathbb{E}[\rho(R^*(x)) \mid x \in S]\right| \leq \varepsilon. \tag{1}$$

For general decision rules, we prove a computational separation: there are settings where no efficient learner can output a predictor that is both decision-accurate and calibrated, even when training examples are labeled by $R^*(x)$. We complement this with a tight positive result: for affine decision rules (defined next in Section 3.1), calibration (and even weaker conditions) implies decision-accuracy, yielding a characterization of the decision rules that admit decision-accuracy.

Classification-accuracy vs. loss minimization. Let $h : \mathcal{X} \to \{0, 1\}$ be an action function that (randomly) maps individuals to binary classifications. Motivated by the above separation for predictors, we define *classification-accuracy* by requiring that, for each group, the positive-classification rate of h matches the rate of the Bayes-optimal loss-minimizing action function h^*_ℓ (which is induced by R^* and a loss function ℓ). That is, for every $S \in \mathcal{C}$,

$$|\mathbb{E}[h(x) \mid x \in S] - \mathbb{E}[h^*_\ell(x) \mid x \in S]| \leq \varepsilon. \tag{2}$$

We show a strong negative result: for non-trivial loss functions, it can be computationally infeasible to learn an action function that is classification-accurate and also has loss competitive with the minimal benchmark class of constant functions. In other words, to obtain classification accuracy one may incur expected loss strictly larger than the constant functions (classify everyone as 0 or everyone as 1). As above, this infeasibility result holds even when training examples are labeled by $R^*(x)$.

Positive results: one desideratum at a time. Although the above pairs of guarantees can be incompatible, we give efficient algorithms achieving each desideratum on its own. We show how multi-accuracy can be used to obtain decision-accuracy and classification-accuracy, and leverage strong results from prior works on multi-calibration and omniprediction [Gopalan et al., 2022] to show how to learn action functions with competitive loss.

The notions and their associated objects are summarized in Table 1. The key message is that the desiderata we consider are Bayes-consistent and each is efficiently achievable on its own, yet certain combinations are computationally incompatible. In particular, our negative results apply even in settings where the Bayes-optimal predictor (and the Bayes-optimal action function) have succinct descriptions; a super-polynomial-time learner can recover R^* from a polynomial-size training set and satisfy all desiderata, but no efficient learner can.

Table 1: New accuracy notions: Multi-Accuracy-on-Decision (MAD, Equation (1)) and Multi-Accuracy-on-Classification (MAC, Equation (2)).

Objects	Accuracy notion
predictor $\tilde{R} : \mathcal{X} \to \mathcal{Y}$, decision rule $\rho : \mathcal{Y} \to \{0, 1\}$ action function $h : \mathcal{X} \to \{0, 1\}$	**MAD:** On every group, the positive-decision rate induced by applying ρ to $\tilde{R}$ matches that induced by applying ρ to R^*. **MAC:** On every group, the positive-classification rate of h matches that of the loss-minimizing action function.

Implications for fairness. Although our desiderata are Bayes-consistent accuracy guarantees, they are naturally motivated by fairness concerns: subpopulations can be harmed by both false positives and false negatives, and these harms are reflected in both decision rates and loss-based objectives [Liu et al., 2018]. Our positive results show that rich labels enable strong subgroup-level protections such as calibration, loss minimization, and our new rate-accuracy notions, each achievable efficiently on its own. Our negative results show that, even with rich labels, certain pairs of these protections can be computationally incompatible, forcing a policy maker to choose which protections to provide.

2 Related work

Prior incompatibility results. Classic impossibility results in algorithmic fairness, most notably those of Kleinberg et al. [2017] and Chouldechova [2017], identify inherent statistical tensions between calibration and error-rate parity constraints across demographic groups; these tensions persist even given complete knowledge of the Bayes-optimal conditional probabilities. The importance of these separations lies in exposing conceptual contradictions among the notions. Recent work has also studied fairness–accuracy and fairness–fairness tradeoffs through Pareto-frontier formulations [Chai et al., 2026]. In contrast, our separations are computational in nature, and illustrate incompatibilities between different notions that are all grounded in, and in line with, accuracy.

Multi-accuracy and multi-calibration for predictions. The literature studies several notions of calibration in the multi-type (or multi-class) setting (see Gopalan et al. [2024a] for a full discussion). We highlight two extremes among these: *Canonical* (or *distribution) calibration* [Kull and Flach, 2015, Widmann et al., 2019] conditions on every possible predicted type distribution, and yields a robust accuracy guarantee, but the relevant event space grows exponentially with the number of types, leading to exponential sample complexity even for a single group. In contrast, *class-wise* (or *coordinate-wise) calibration* [Kull et al., 2019] conditions separately on each coordinate of the predicted distribution (a polynomial-size family of events) and can be obtained efficiently. In Appendices C.1 and C.2 we include formal definitions of the variants we study and the relevant learning algorithms for achieving them.

Other variants of multi-class calibration. Canonical calibration requires predicted label probabilities to match the true label distribution among instances with the same prediction. As discussed in the previous paragraph, this is often infeasible in multi-class settings due to exponential sample complexity in the number of labels. To address this, Zhao et al. [2021] propose *decision calibration*, which ensures that predicted and true distributions are indistinguishable to decision-makers with a bounded number of actions. Their recalibration algorithm runs with sample complexity polynomial in both the number of actions and labels. Gopalan et al. [2024a] strengthen this by introducing *projected smooth calibration*, which guarantees robustness for all binary classification tasks of the form "is the label in a subset $T \subseteq \mathcal{T}$?" and achieves it efficiently in both time and samples. Our work also supports downstream binary tasks, whereas our emphasis is not on efficient recalibration per se, but rather on characterizing which decision or classification guarantees can be achieved efficiently.

Uncertainty decomposition via higher-order calibration. The recent work of Ahdritz et al. [2025] introduces higher-order calibration for predictors that output distributions over label distributions. They aim to decompose uncertainty into *epistemic* and *aleatoric* components via higher-order calibration, while relying on access to multiple independent labels per instance (as Nature itself is probabilistic). This enables a rigorous separation between model-induced uncertainty and inherent randomness. With the tension between the two kinds of uncertainty in mind, one can understand

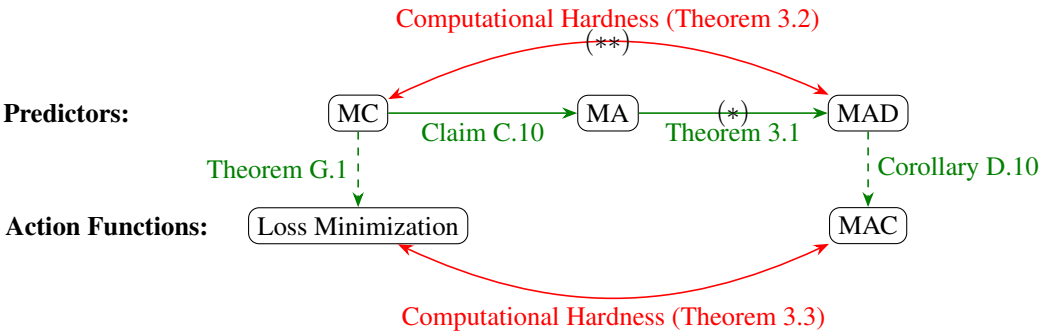


Figure 1: Positive and negative results. MC: Multi-Calibration, MA: Multi-Accuracy, MAD: Multi-Accuracy-on-Decision, MAC: Multi-Accuracy-on-Classification. The top row contains properties defined over a predictor and the bottom row contains properties defined over the action function obtained by applying a decision rule. A green arrow from X to Y means that satisfying X implies satisfying Y. A green dashed arrow from X to Y means that applying a decision rule to a predictor satisfying X yields an action function satisfying Y. A red arrow between X and Y means that achieving both simultaneously is computationally hard. The implication marked $(*)$ holds for affine decision rules, and the impossibility marked $(**)$ holds for rules that are far from affine. The dashed implication from MAD to MAC additionally assumes that Nature is deterministic.

our work as embedding uncertainty into the notion of MAC, which penalizes *epistemic* uncertainty—when the model fails to separate different types in the data—while accepting *aleatoric* uncertainty as intrinsic.[2] This contrasts with fairness notions like balance or equalized odds, which penalize even purely aleatoric deviations. In our framework, aleatoric uncertainty defines the baseline—we do not ask the model to outperform the inherent randomness of Nature, only to capture structure when it exists. In this way, our approach captures the epistemic/aleatoric distinction implicitly, by requiring consistency with the Bayes-optimal predictor, and without relying on multiple labels per instance.

Evidence-based rankings. Dwork et al. [2019] study the problem of learning rankings from binary outcomes alone, without access to individual probabilities. Their framework defines a ranking as a real-valued function over the domain $\mathcal{X}$, consistent with our formulation up to the fact that their range is continuous (i.e., interval $[0, 1]$), whereas we focus on coarse, discrete rankings (i.e., a rank is a number in $\{1, \ldots, k\}$). Additionally, our definitions are extended to allow probabilistic outputs, i.e., distributions over ranks. In their setup, the quality of a learned ranking is evaluated indirectly: rather than comparing it to a ground-truth ranking, as in our case, they assess the ranking through the best predictor it can induce. They propose *evidence-consistency* and *domination-compatibility*, where the former is stronger than the latter. Since their assumptions are more restrictive (due to limited supervision), the corresponding guarantees are also weaker: the notion we consider of coordinate-wise-multi-accuracy (which assumes access to individual ranks) implies evidence-consistency.

Relation to linearizing statistics. Multi-accuracy can be read as keeping the group averages of $\tilde{R}$ correct: for each group $S \in \mathcal{C}$ and type $t \in \mathcal{T}$, the average of $\tilde{R}_t(x)$ over S matches that of $R^*_t(x)$. For affine ρ, the groupwise decision rates depend only on these averages, so multi-accuracy implies decision-accuracy (Theorem 3.1). This mirrors the recipe in Gopalan et al. [2024b]: predict the right statistics and keep them calibrated on groups to obtain near-optimal actions.

3 Overview of results and techniques

We provide a more detailed overview of our results and their proofs, focusing on the negative results.

[2]For instance, if all applicants in a subgroup have Bayes-optimal rich label assigning probability $1/2$ to being tier-1 and probability $1/2$ to being tier-3, then MAC may tolerate rejecting them under a threshold-based decision rule. But if the subgroup instead consists of a mixture of applicants whose Bayes-optimal rich labels are pure tier-1 and pure tier-3, and the model fails to distinguish them, then MAC flags this as a failure of learning.

3.1 Decision-accuracy vs. calibration

Multi-accurate decisions. Recall that in the multi-type setting, a predictor's output is a distribution over a finite type space $\mathcal{T}$, and $\mathcal{Y}$ denotes the set of such distributions. A *decision rule* $\rho : \mathcal{Y} \to \{0, 1\}$ maps a prediction (a distribution over $\mathcal{T}$) to a binary decision. In deployments, ρ is typically designed with the Bayes-optimal rich predictor R^* in mind, but is applied to a learned predictor $\tilde{R}$.

Recall the formalization of multi-accuracy-on-decision (MAD) from Equation (1): MAD compares the decision rate induced by $\tilde{R}$ to the decision rate induced by the Bayes-optimal rich predictor R^*. Thus, to define or audit MAD, one needs access to the rich labels $R^*(x)$, not merely to binary outcomes. For example, suppose a loan rule approves applicants whose repayment probability is at least 0.95, and in a group S half of the observed loans are repaid. This observation is consistent both with the case where every applicant has repayment probability 0.5, in which case the Bayes-optimal rule rejects everyone, and with the case where half the applicants have repayment probability 1 and half have probability 0, in which case the Bayes-optimal rule approves half the group. These two cases induce different MAD benchmarks, even though they may be indistinguishable from binary outcomes alone. This motivates our focus on rich-label settings, where training examples reveal each individual's rich label $R^*(x)$.

Decision-accuracy (with respect to some decision rule) and calibrated predictions are both desirable properties for a predictor, and enforcing them for subgroups can address certain fairness concerns. Can we simultaneously achieve both of these natural notions? We give a tight characterization of the decision rules for which this is and is not possible: close-to-affine rules allow coexistence, while far-from-affine rules lead to computational hardness.

Affine decision rules: multi-accuracy implies decision-accuracy. We say that a decision rule is *affine* if, on input a probability distribution over types, it operates by sampling a type from the distribution and applying a fixed function that (randomly) maps each type to a decision in $\{0, 1\}$.[3]

Example. Suppose a predictor $\tilde{R}$ outputs, for each applicant, a distribution over three future student types: tier-1, tier-2, and tier-3. Consider an affine decision rule that samples a type from this distribution and then admits with probability 1 for tier-1, 0.5 for tier-2, and 0 for tier-3; equivalently, the admission probability is the predicted mass on tier-1 plus half the predicted mass on tier-2. Suppose that on some subgroup S, the Bayes-optimal predictor R^* has average type distribution $(0.4, 0.2, 0.4)$ (for example, it assigns $(0.5, 0.2, 0.3)$ for half of the applicants and $(0.3, 0.2, 0.5)$ for the other half). The admission rate induced by R^* on S is $0.4 + 0.5 \cdot 0.2 = 0.5$. But any predictor $\tilde{R}$ that is multi-accurate has essentially the same average type distribution on S, namely $(0.4, 0.2, 0.4)$, as multi-accuracy requires. Therefore, under the same affine rule, $\tilde{R}$ induces essentially the same admission rate on S, namely about 50%.

We show that this intuition holds generally: for affine decision rules, multi-accurate predictors (and so do multi-calibrated predictors) induce multi-accurate decisions (MAD).

Theorem 3.1 (Informal, see Theorem D.4)**.** *Let $\rho : \mathcal{Y} \to \{0, 1\}$ be an affine decision rule. If a predictor $\tilde{R}$ is multi-accurate w.r.t. a collection of groups $\mathcal{C}$, then it is also decision-accurate for every group in $\mathcal{C}$.*

We note that the error in the decision-accuracy guarantee depends on the error parameter in the multi-accuracy guarantee. Moreover, it suffices that the decision rule be *close* to affine; the resulting closeness parameter also contributes to the overall decision-accuracy error. This means that our characterization is *tight*: Theorem 3.1 shows that rules that are close to affine still allow coexistence, whereas Theorem 3.2 presented next will show hardness for rules that are far from affine.

Non-affine rules: calibration and decision-accuracy are in conflict. In many settings, affine decision rules are too restrictive. For example, a lender may wish to approve only applicants who are nearly certain to repay the loan under the predicted rich label, rather than applicants who have, say, only a 90% chance of repayment; this kind of "near-certainty" requirement is inherently non-affine in

[3]This formulation is equivalent to the standard definition of affineness: applying the rule to a convex combination of two distributions over types yields the same convex combination of the outcomes obtained by applying the rule to each distribution separately. We refer the reader to Definition C.13 for the full formal definition of being affine and close to affine.

the predicted distribution. Can we obtain predictors that are both calibrated and decision-accurate for such non-affine decision rules?

We prove a general negative result: for decision rules that are far from affine (informally, they do not approximately preserve convex combinations), there are data distributions for which it is computationally infeasible to learn a predictor that is simultaneously calibrated and decision-accurate. Importantly, the conflict is neither conceptual nor a consequence of simultaneously enforcing many group constraints: the Bayes-optimal rich predictor R^* is both calibrated and decision-accurate, but cannot be learned efficiently (nor can any other predictor satisfying both guarantees). At the same time, each guarantee is natural and efficiently achievable in isolation; the hardness only arises when we require both at once.

Example. Let us consider our running example of lending. Suppose half of the applicants belong to a subpopulation $\mathcal{X}_1$ with true repayment probability 0.99, and the other half belong to a subpopulation $\mathcal{X}_2$ with true repayment probability 0.90, but the two are computationally hidden from the learner's features. Then an efficient calibrated predictor $\tilde{R}$ may be forced to average over the hidden partition and assign all applicants repayment probability 0.945. If the bank approves only applicants with predicted repayment probability above the threshold 0.95 (this rule is far from affine), then the Bayes-optimal predictor R^* approves all of $\mathcal{X}_1$ and none of $\mathcal{X}_2$ (overall approval rate 50%), while $\tilde{R}$ rejects everyone (overall approval rate 0%).[4]

The calibration notion we consider is coordinate-wise, a very relaxed requirement; proving impossibility even under this weak notion strengthens the result. Moreover, recall that we work in the rich-label model: the learner has sample access to pairs $(x, R^*(x))$, where $R^*(x) \in \mathcal{Y}$ is a distribution over types. This further strengthens the lower bound by showing that the obstruction is not insufficient supervision from binary outcomes, but computational hardness.

Theorem 3.2 (Informal, see Theorem E.1)**.** *Assume the existence of a computationally hidden subpopulation. Let $\rho : \mathcal{Y} \to \{0, 1\}$ be a decision rule that is Lipschitz and far from affine. There exists a data distribution for which no polynomial-time learner can output a predictor that is both coordinate-wise calibrated and decision-accurate w.r.t. ρ (except with negligible probability).*

The assumption says that there is a subpopulation of the required constant density whose membership is not inferable from the learner's observed features. We formalize it via computational indistinguishability: no efficient procedure can reliably tell whether an individual belongs to the subpopulation or to its complement. We give two formal realizations. The first is information-theoretic, modeling latent attributes, missing covariates, or privacy-constrained access to features: labels may be efficiently generated from the full feature set, while the learner's observed features make the relevant partition inaccessible. The second is cryptographic, based on one-way functions, a lightweight and well-established assumption. See Appendix C.4 for the formal definitions and proofs.

Proof overview for Theorem 3.2. Given a non-affine decision rule ρ, we construct a distribution for which no efficient learner can output a predictor that is both calibrated and decision-accurate, even for a single subpopulation. The idea behind the proof is to construct a setting where any efficiently learned calibrated predictor must smooth out distinctions between inputs, even when their true predictions are different. To do so, we consider the following setup. Since ρ violates affineness, there exist two vectors (representing distributions over labels) y and y' such that

$$\frac{1}{2}\rho(y) + \frac{1}{2}\rho(y') \neq \rho\left(\frac{1}{2}y + \frac{1}{2}y'\right). \tag{3}$$

For simplicity, we use the constant $1/2$ in this argument, though the same reasoning applies to any convex combination. We construct the domain as $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$, where each subset constitutes half of $\mathcal{X}$ and the two are computationally indistinguishable, relying on the hidden-subpopulation assumption. The Bayes-optimal predictor R^* assigns the distribution y on $\mathcal{X}_1$ and y' on $\mathcal{X}_2$. Informally, any efficiently learned calibrated predictor $\tilde{R}$ cannot reliably distinguish $\mathcal{X}_1$ from $\mathcal{X}_2$; calibration then forces it to behave like the "squashed" predictor that outputs the mixture $\frac{1}{2}y + \frac{1}{2}y'$ rather than the group-dependent labels y and y', on any $x \in \mathcal{X}$.

[4] Strictly speaking, this threshold rule is not Lipschitz, as required by Theorem 3.2. However, this is easily resolved by smoothing the rule: we can replace it with a randomized Lipschitz rule whose approval probability agrees with the threshold rule outside the interval $[0.947, 0.953]$ and works linearly within this interval. This leaves the approval rates in the example unchanged.

To evaluate decision-accuracy, we compare the expected value of ρ applied to $\tilde{R}$ and to R^*. For R^*, this expectation equals $\frac{1}{2}\rho(y) + \frac{1}{2}\rho(y')$, whereas for $\tilde{R}$ it equals $\rho\left(\frac{1}{2}y + \frac{1}{2}y'\right)$. By Equation (3), these values differ, and thus $\tilde{R}$ is not decision-accurate with respect to ρ.

The formal proof of the general case is the most technically involved part of this work, as it introduces several technical challenges. First, $\tilde{R}$ is only *approximately* calibrated, and may behave arbitrarily on small subsets. This interacts subtly with the weak calibration requirement we assume, under which the level sets[5] of $\tilde{R}$ do not form a partition, complicating the identification of a large subset where $\tilde{R}$ fails to be decision-accurate. Second, since we aim to refute *approximate* decision-accuracy, we must ensure that ρ is bounded away from any affine function. To control this, we impose a Lipschitz condition on ρ, and argue that without such regularity, the resulting lower bound would be vacuous. Finally, since $\tilde{R}$ must be discretized to be learned efficiently, care is required to ensure the argument is robust to any choice of discretization. □

Non-affine rules arise naturally from loss minimization. Given a loss function, the loss-minimizing decision rule takes the positive action (e.g., approve/admit/intervene) precisely when its expected loss is smaller than that of the negative action (e.g., deny/reject/not intervene). In a medical setting, for example, a treatment with severe side effects may be prescribed only when the predicted distribution places overwhelming mass on types corresponding to very high medical risk. The loss function trades off the harm of prescribing an unnecessary or dangerous treatment against the harm of failing to treat a patient who is in fact at high risk, and the induced rule prescribes treatment exactly when the expected benefit outweighs the expected side-effect cost. Such thresholding of expected loss differences is inherently far from affine.

In fact, loss-minimizing decision rules are far from affine for every non-trivial[6] loss function (see Lemma E.8 and Remark G.5). Combined with the negative result, this implies that learning a calibrated predictor that is decision-accurate with respect to the loss-minimizing rule may be computationally infeasible.[7] This leads to the next question: if applying a loss-minimizing rule to a learned predictor may be problematic, can we instead learn an action function directly while preserving classification-accuracy?

3.2 Classification-accuracy vs. loss minimization

Multi-accurate classifications. Binary decisions need not be produced by applying a decision rule to a learned predictor. Accordingly, we study general *action functions* $h : \mathcal{X} \to \{0, 1\}$, which map features directly to a binary action. We work in a loss-minimization framework: a loss function ℓ assigns a loss $\ell(t, a)$ to taking action $a \in \{0, 1\}$ when the individual's type is $t \in \mathcal{T}$.

Given an individual x and an action a, the relevant quantity to consider is the expected loss under the individual's rich label $R^*(x)$, namely $\mathbb{E}_{t\sim R^*(x)}[\ell(t, a)]$. If we had access to R^*, we could choose, for each individual x separately, the action $a \in \{0, 1\}$ that minimizes $\mathbb{E}_{t\sim R^*(x)}[\ell(t, a)]$, yielding the Bayes-optimal loss-minimizing action function h^*_ℓ. For a subpopulation S, this benchmark induces a positive-classification rate on S. Recalling Equation (2), we say that an action function is *multi-accurate-on-classification* (MAC) for a collection of groups $\mathcal{C}$ if it matches these groupwise positive-classification rates for every $S \in \mathcal{C}$. Note that h^*_ℓ both minimizes loss and (trivially) satisfies MAC for any collection $\mathcal{C}$; thus the negative result of Theorem 3.3 presented next is purely computational, as was the previous negative result.

Loss minimization and classification-accuracy are in conflict. A second natural objective when learning action functions is to (approximately) minimize expected loss relative to a benchmark class. In contrast to our multigroup rate notions, loss minimization is a global objective. Our second negative result shows that, for essentially any non-trivial loss function (see Footnote 6), loss minimization can

[5] This term is from Hébert-Johnson et al. [2018]. In our relaxed *coordinate-wise* multi-calibration setting, a "level set" groups individuals whose predicted probability for a fixed type is approximately the same. These sets can overlap across different types, since each individual is assigned a full distribution over types.

[6] We only require that the loss is not degenerate: there exists a type for which one action is significantly better than the other, but no single action is optimal for all types (otherwise the problem is trivial, since a constant action minimizes loss).

[7] Technically, the loss-minimizing decision rule is not Lipschitz, so the impossibility result cannot apply directly. However, in Claim E.9, we show that it extends for these rules, for any non-trivial loss function.

be incompatible with classification-accuracy. Indeed, the incompatibility already appears when the benchmark class for loss minimization consists only of the two constant action functions.

Example. Let us consider our running example of preventive medicine. Suppose 75% of patients belong to a subpopulation $\mathcal{X}_1$ for whom treatment is harmful, with loss 1 for "treat" and loss 0 for "do not treat", while the remaining 25% belong to a subpopulation $\mathcal{X}_2$ for whom treatment is beneficial, with loss 0 for "treat" and loss 0.2 for "do not treat". The Bayes-optimal action function treats exactly $\mathcal{X}_2$, so the treatment rate is 25%. But if $\mathcal{X}_1$ and $\mathcal{X}_2$ are computationally hidden from the learner's features, then any efficiently learnable action function that also treats about 25% of patients cannot reliably pick out $\mathcal{X}_2$, and instead treats an essentially random 25% of the population. Its expected loss is then

$$0.75 \cdot 0.25 \cdot 1 + 0.25 \cdot 0.75 \cdot 0.2 = 0.225,$$

which is worse than the constant action function "treat nobody", whose loss is only $0.25 \cdot 0.2 = 0.05$.

The following theorem formalizes this phenomenon: for every non-trivial loss function, classification-accuracy can force any efficient action function to have loss worse than a constant benchmark.

Theorem 3.3 (Informal, see Theorem F.1)**.** *Assume the existence of a computationally hidden subpopulation. For any non-trivial loss function ℓ (as above), there exists a data distribution such that any polynomial-time action function that satisfies classification-accuracy must incur expected loss strictly larger than one of the constant action functions (classify everyone as 0 or everyone as 1).*

Proof overview for Theorem 3.3. For any non-trivial loss function, we construct a Nature under which any efficiently computable action function that is classification-accurate fails to minimize loss. We prove this with respect to the most restricted hypothesis class, which contains only the two constant functions that either accept or reject all individuals. The proof idea is as follows. The non-triviality of the loss implies the existence of two labels, or types, t and t', such that *rejecting* type t results in a significantly lower loss compared to accepting it, with a difference of at least $\alpha > 0$, while *accepting* type t' is preferable in terms of loss, though the difference may be smaller. Here, "reject" and "accept" correspond to the binary actions 0 and 1, but any other binary action can work.

We define the domain to contain an indistinguishable subset $\mathcal{X}_1 \subseteq \mathcal{X}$ of size $3/4$ of $\mathcal{X}$, relying on the hidden-subpopulation assumption. Let R^* assign type t to each $x \in \mathcal{X}_1$ and type t' to the remaining individuals. Consequently, the loss-optimal action function rejects everyone in $\mathcal{X}_1$ and accepts the rest, resulting in an expected acceptance rate of $1/4$ over $\mathcal{X}$. Thus, any action function h that is classification-accurate must accept roughly $1/4$ of the domain.

By indistinguishability, the set of accepted individuals must intersect $\mathcal{X}_1$ in approximately a $3/4$ fraction: if the efficiently computable action function h accepted a noticeably different fraction from $\mathcal{X}_1$ than from its complement, then h itself would distinguish $\mathcal{X}_1$ from $\mathcal{X} \setminus \mathcal{X}_1$. This implies that the expected loss of h conditioned on the set $\{h(x) = 1\}$ matches the expected loss of accepting the entire domain. Let L_1 denote this quantity and L_0 denote the analogous quantity for rejecting everyone. Then, the expected loss of h is given by $3L_0/4 + L_1/4$, and the difference between this loss and that of the function that rejects everyone is $(L_1 - L_0)/4$. On the other hand, by the non-triviality of the loss, $L_1 - L_0$ is lower-bounded by $\alpha/2$, allowing us to lower-bound the loss-minimization error of h by at least $\alpha/8$. Hence no polynomial-time action function can satisfy classification-accuracy and simultaneously match (or beat) the best constant action. □

3.3 Positive results: one desideratum at a time

While our negative results show that certain pairs of natural desiderata can be computationally incompatible, we also give positive results showing that each desideratum is efficiently achievable on its own. Our positive results build on predictors satisfying multi-accuracy or multi-calibration. We summarize these results here, emphasizing the case of loss minimization, which is the most subtle.

Decision and classification accuracy. First, as already proved in Theorem 3.1, MAD can be achieved by multi-accuracy. Moreover, in the natural special case when Nature's rich predictor R^* is deterministic, multi-accuracy suffices to obtain both MAD and MAC. Concretely, randomly instantiating a multi-accurate predictor (sampling a type for each individual from its predicted distribution) yields MAD, and applying the loss-minimizing decision rule yields MAC. After random instantiation, the rule maps the label to $\{0, 1\}$. MAD is achieved with any such mapping, while MAC

is achieved when this mapping is the loss-minimizing action for the realized label (i.e., it chooses the loss-minimizing action given the sampled label). Unlike direct loss minimization, in the MAC path the loss-minimizing rule is applied only *after* random instantiation. Figure 2 visualizes the two paths.

Loss minimization. Finally, we study learning an action function whose expected loss is competitive with a benchmark class. Building on the theory of omnipredictors [Gopalan et al., 2022], we prove that applying a loss-minimizing decision rule to a predictor that is multi-calibrated for a collection of sets $\mathcal{C}$ yields an action function competitive with a benchmark class closely related to $\mathcal{C}$.

Theorem 3.4 (Informal, see Theorem G.1)**.** *Let $\mathcal{H}$ be a hypothesis class over $\mathcal{X}$, and let $\mathcal{C}$ be a collection of subsets induced by $\mathcal{H}$, i.e., $\{S_h : h \in \mathcal{H}\} \subseteq \mathcal{C}$ where $S_h = \{x \in \mathcal{X} : h(x) = 1\}$. For every loss function ℓ, if $\tilde{R}$ satisfies canonical $\mathcal{C}$-multi-calibration, then applying the loss-minimizing decision rule to $\tilde{R}$ yields an action function whose expected loss is competitive with every $h \in \mathcal{H}$.*

A full comparison to the setting of Gopalan et al. [2022] can be found in Appendix G; in what follows, we focus on the results that go beyond prior work. Since the calibration guarantee used in Theorem 3.4 is strong and (as discussed above) may be expensive to obtain, we also study relaxed settings in which weaker calibration guarantees, and thus more efficient learning, suffice for loss minimization. In Appendix G.1.1, we focus on type spaces corresponding to success probabilities (as in the loan repayment example) and losses that are linear in the predicted probability; in this case, it suffices to multi-calibrate the predicted expectations, reducing the problem to one-dimensional prediction. In Appendix G.1.2, we allow general type spaces but fix a single loss function. Finally, Appendix G.1.3 gives a reduction from loss minimization to auditing multi-calibration; while this can still be computationally demanding, it avoids the exponential sample complexity inherent in canonical multi-calibration.

3.4 Limitations and implications

Limitations. First, our results are theoretical: the paper proves that the incompatibilities we identify can occur, but does not show that they arise in typical deployed systems or measure their frequency empirically. Second, the lower bounds rely on the existence of a hidden subpopulation, whose membership is inaccessible to efficient learners. While we give both information-theoretic and cryptographic realizations of this assumption, the results should be interpreted as identifying a computational obstruction rather than as claiming that the obstruction appears in every deployment. Moreover, our positive results inherit the computational and sample-complexity requirements of multi-accuracy and multi-calibration.

Implications. Figure 2, which builds on Figure 1, highlights the central choice exposed by our results. Given a multi-calibrated predictor $\tilde{R}$, one can either apply the loss-minimizing decision rule (a non-affine rule) to directly obtain a loss-minimizing action function, or use a decision rule that first randomly instantiates $\tilde{R}$ (an affine rule) to achieve MAD or MAC. Thus, calibration plays a central and somewhat counter-intuitive role. On the one hand, multi-calibrated predictors provide a versatile starting point from which we can efficiently obtain strong guarantees—including decision-accuracy, classification-accuracy, and loss competitiveness—depending on how we post-process the predictions. On the other hand, our separations show that these guarantees can be computationally incompatible when demanded simultaneously.

The resolution is that calibration is not itself the end goal; different downstream goals require using the predictor in different ways at decision time. Since both approaches operate on the same predictor $\tilde{R}$, the choice can be deferred to deployment: one can train a multi-calibrated predictor and later decide whether to prioritize MAC or loss competitiveness.

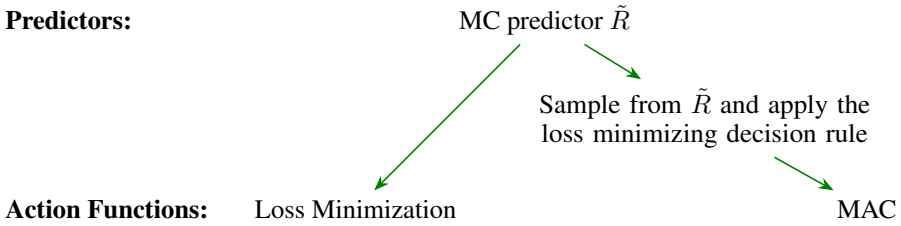


Figure 2: Choice in how to use a predictor. While a single predictor may enable both loss minimization and MAC, these goals cannot always be achieved simultaneously. This leads to two distinct paths: applying the loss-minimizing decision rule to the full distribution $\tilde{R}$ to directly minimize loss (left), or first sampling from $\tilde{R}$ and then applying the decision rule to achieve MAC (right). The MAC guarantee for the right-hand path assumes that Nature is deterministic.

Acknowledgments and Disclosure of Funding

We thank Shafi Goldwasser for her collaboration in early stages of this project, which motivated our work. We also thank Parikshit Gopalan for many useful discussions, and Charlotte Gates Peale for helpful feedback on an earlier draft. Omer Reingold is supported by the Simons Foundation Collaboration on the Theory of Algorithmic Fairness, the Sloan Foundation Grant 2020-13941 and the Simons Foundation investigators award 689988. The authors declare no competing interests.

References

Gustaf Ahdritz, Aravind Gollakota, Parikshit Gopalan, Charlotte Peale, and Udi Wieder. Provable uncertainty decomposition via higher-order calibration. In *International Conference on Learning Representations*, 2025. To appear.

Junyi Chai, Shenyu Lu, and Xiaoqian Wang. Rethinking pareto frontier: On the optimal trade-offs in fair classification. In *The Fourteenth International Conference on Learning Representations*, 2026.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

Philip Dawid. On individual risk. *Synthese*, 194(9):3445–3474, 2017. doi: 10.1007/s11229-015-0953-4.

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, 2019. doi: 10.1109/FOCS.2019.00016.

Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Outcome indistinguishability. In Samir Khuller and Virginia Vassilevska Williams, editors, *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pages 1095–1108. ACM, 2021. doi: 10.1145/3406325.3451064. URL `https://doi.org/10.1145/3406325.3451064`.

Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Beyond bernoulli: Generating random outcomes that cannot be distinguished from nature. In Sanjoy Dasgupta and Nika Haghtalab, editors, *International Conference on Algorithmic Learning Theory, 29 March - 1 April 2022, Paris, France*, volume 167 of *Proceedings of Machine Learning Research*, pages 342–380. PMLR, 2022. URL `https://proceedings.mlr.press/v167/dwork22a.html`.

Oded Goldreich, Shafi Goldwasser, and Silvio Micali. How to construct random functions. *J. ACM*, 33(4):792–807, August 1986. ISSN 0004-5411. doi: 10.1145/6490.6503. URL `https://doi.org/10.1145/6490.6503`.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In Mark Braverman, editor, *13th Innovations in Theoretical Computer Science*

Conference (ITCS 2022), volume 215 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 79:1–79:21, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-217-4. doi: 10.4230/LIPIcs.ITCS.2022.79. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2022.79.

Parikshit Gopalan, Lunjia Hu, and Guy N. Rothblum. On computationally efficient multi-class calibration. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 1983–2026. PMLR, 30 Jun–03 Jul 2024a.

Parikshit Gopalan, Princewill Okoroafor, Prasad Raghavendra, Abhishek Shetty, and Mihir Singhal. Omnipredictors for regression and the approximate rank of convex functions. In *Annual Conference on Computational Learning Theory*, 2024b. URL https://api.semanticscholar.org/CorpusID:267301310.

Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1944–1953. PMLR, 2018. URL http://proceedings.mlr.press/v80/hebert-johnson18a.html.

Johan Håstad, Russell Impagliazzo, Leonid A. Levin, and Michael Luby. A pseudorandom generator from any one-way function. *SIAM J. Comput.*, 28(4):1364–1396, March 1999. ISSN 0097-5397. doi: 10.1137/S0097539793244708. URL https://doi.org/10.1137/S0097539793244708.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 43:1–43:23, Dagstuhl, Germany, 2017. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-029-3. doi: 10.4230/LIPIcs.ITCS.2017.43. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2017.43.

Meelis Kull and Peter Flach. Novel decompositions of proper scoring rules for classification: score adjustment as precursor to calibration. In *Proceedings of the 2015th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I*, ECMLPKDD'15, pages 68–85, Gewerbestrasse 11 CH-6330, Cham (ZG), CHE, 2015. Springer. ISBN 9783319235271. doi: 10.1007/978-3-319-23528-8_5. URL https://doi.org/10.1007/978-3-319-23528-8_5.

Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: obtaining well-calibrated multiclass probabilities with dirichlet calibration. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3150–3158. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/liu18c.html.

Luca Trevisan, Madhur Tulsiani, and Salil P. Vadhan. Regularity, boosting, and efficiently simulating every high-entropy distribution. In *Proceedings of the 24th Annual IEEE Conference on Computational Complexity*, pages 126–136. IEEE Computer Society, 2009.

David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests in multi-class classification: A unifying framework. In *Neural Information Processing Systems*, 2019. URL https://api.semanticscholar.org/CorpusID:202770409.

Shengjia Zhao, Michael Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 22313–22324. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/bbc92a647199b832ec90d7cf57074e9e-Paper.pdf.

A Appendix roadmap

The remainder of the paper is organized as follows. In Appendix B we further discuss the motivation for learning with rich labels and contrast it with the standard Boolean setting. Appendix C provides the full definitions and preliminaries. In Appendix D we show how multi-accuracy guarantees for rich predictors translate into decision-accuracy and classification-accuracy guarantees for the induced binary decisions. In Appendix E we show that calibration can contradict decision-accuracy. In Appendix F we show that loss minimization can contradict classification-accuracy. Finally, in Appendix G we show how multi-calibration can be translated into loss minimization via post-processing.

B Further discussion on rich labels

Background: Boolean labels. Much of the algorithmic fairness literature focuses on scenarios where the labels correspond to Boolean outcomes. The training data comes from a distribution $\mathcal{D}$ on pairs (x, y^*) where x is a feature vector from a space $\mathcal{X}$ and $y^* \in \{0, 1\}$ is a label. For example, x could correspond to a patient's medical history, symptoms, and test results, and y^* could correspond to whether they have experienced significant cardiovascular disease in the subsequent decade. Alternatively, as in the running example given in Section 1, x could represent the information of a loan applicant, such as income, credit history, and other financial factors, and y^* corresponds to whether the loan was repaid. The purpose of the learning model in such scenarios is to decide on an appropriate action. For example, deciding on a medical intervention such as prescribing cholesterol medication or on whether to approve a loan application. A natural first step in such decisions is to estimate the probability of the outcome conditioned on the features. For example, "the patient has an 8% chance of having a heart attack in the next decade" or "there is a 30% chance that the applicant will default on the loan". We refer to this task as *prediction*. Decisions are often based on risk predictions: given a predicted level of risk, a particular decision may or may not be warranted.

In this model of learning, the true probability $p^*(x) = \Pr[y^* = 1 \mid x]$ is never provided to the learning algorithm (here, p^* is the Bayes-optimal predictor). Indeed, p^* is often considered completely out of reach, and in fact the meaning of individual probabilities has deep philosophical implications and has been debated in Statistics for generations [Dawid, 2017], with more recent computational perspectives [Dwork et al., 2021]. Given these challenges, a typical approach is to approximate p^* by an efficient predictor $\tilde{p}$ and to execute some decision rule on $\tilde{p}$ that we would consider appropriate if applied on p^*. For example, we could rank all the patients based on $\tilde{p}$ (indicating probability of future heart attack) and invite those with high values to a consultation with their primary care team on making lifestyle changes. Similarly, we can accept loan applications whenever $\tilde{p}$ is larger than 0.95, indicating a tolerable risk for the bank.

Let us revisit the example from Section 3, that motivated the study of Decision-Accuracy. For a set of individuals S, multi-accuracy-on-decision requires that members of S see the same distribution of decisions when the decision rule is applied to $\tilde{p}$ as they would when it is applied to p^*. While this is a very natural notion of accuracy and of fairness, it is often way beyond the reach of a learning algorithm. Consider for example a sample of the outcomes, where half of the members of S repaid the loan and half did not. If there are no efficiently learnable patterns that correlate with loan repayment in S, then it could be that every member of S has p^* being half, or perhaps half of the values are one and half are zeros, or perhaps two thirds of the values are $3/4$ and the rest are zeros. Each of these cases could induce a very different distribution on the decisions to members of S, say, under the decision rule that gives a positive decision to individuals whose probability of repaying the loan is at least 95%.

Our setting: rich labels. The above example highlights a fundamental challenge in achieving strong notions of accurate decision rates: we aim to base decisions on risk scores, yet we never observe the true probabilities, not even during training. When comparing two historical loan applicants, for instance, we cannot determine the correct ranking between them based solely on observed outcomes (e.g., repaid vs. defaulted). Even if Alice repaid her loan and Bob did not, it is possible that Bob had a higher true probability of repayment under the ground truth distribution p^*, but was simply unlucky, while Alice was fortunate despite a lower p^* value. In other words, p^* encodes uncertainty

that we cannot directly access; we only observe one realization of it per individual through historical outcomes.

In this work, we study the task of learning given access to richer labels, that reflect this underlying uncertainty, and our goal is to use this richer data to support more accurate decision and classification rates. One natural approach is to include the true probabilities p^* for individuals in the training data. An alternative is to include rankings of individuals. If these rankings are derived from probabilities, the two approaches are effectively equivalent. However, rankings can also encode broader considerations beyond simple outcome probabilities. For example, in college admissions, we might assess applicants not only by their probability of graduation but also by their expected contribution to their cohort or to society more broadly. Even more generally, we can think of each individual as having a "type" that supports multiple kinds of rankings, depending on the institution's goals, such as different prioritization of student applicants by a university.

Where could rich labels come from? A natural source is historical data annotated by one or more domain experts. Of course, even expert judgments can carry biases, and the very act of defining which "types" individuals are categorized into may have fairness implications. Still, experts may possess information unavailable to algorithms—for example, the nuanced understanding a physician might have about a patient's condition. Additionally, the expertise used to generate rich labels can include knowledge of fairness considerations specific to the task or domain. In such cases, these labels can be understood not as representations of an objective ground truth, but rather as a way to encode human values into the learning process.

C Preliminaries

We model Nature as a joint distribution over individuals and labels, denoted by $\mathcal{D}$. Individuals are drawn from a discrete domain $\mathcal{X}$; throughout, we assume that each $x \in \mathcal{X}$ can be represented as a Boolean string $x \in \{0,1\}^n$, which encodes the features of the individual. For simplicity, we write $n = \log|\mathcal{X}|$ to mean $n = \lceil \log|\mathcal{X}| \rceil$. We use $x \sim \mathcal{D}_{\mathcal{X}}$ to denote a sample from Nature's marginal distribution over individuals.

The labels are drawn from an abstract outcome space $\mathcal{T}$, which is assumed to be finite, i.e., $k := |\mathcal{T}| < \infty$. We refer to elements of $\mathcal{T}$ as *types*. Distributions over $\mathcal{T}$ are represented using stochastic vectors, denoted by $\mathcal{Y} \in \Delta(\mathcal{T})$, where $\Delta(\mathcal{T})$ is the set of all discrete probability distributions over $\mathcal{T}$. The following definition formalizes $\mathcal{Y}$.

Definition C.1 (Stochastic Vectors)**.** *Let $\mathcal{T}$ be a finite outcome space of size $k \in \mathbb{N}$. The set of k-dimensional stochastic vectors is defined as*

$$\mathcal{Y} := \left\{ y \in [0,1]^k \text{ such that } \sum_{i=1}^{k} y_i = 1 \right\}.$$

The probability of sampling $t \in \mathcal{T}$ from $y \in \mathcal{Y}$ is given by its t^{th} coordinate, y_t. We denote this by $t \sim y$. If $\mathcal{T}$ admits a total order, we assume that this order determines the coordinate indexing.

We define $R^* : \mathcal{X} \to \mathcal{Y}$ as the predictor induced by Nature. If Nature is deterministic, then R^* outputs a unit vector in $\mathcal{Y}$, corresponding to a one-hot encoding. We denote the t^{th} unit vector by e_t. Overall, we express Nature's joint distribution as

$$\mathcal{D} = \mathcal{D}_{\mathcal{X}} \times R^*(\mathcal{D}_{\mathcal{X}}).$$

The primary objects of study in this work are *probabilistic predictors*,

$$\tilde{R} : \mathcal{X} \to \mathcal{Y},$$

which are intended to approximate R^*.

For a predictor R, we use $R(x) \in \mathcal{Y}$ to denote its estimated distribution for an individual x. A random label sampled according to this distribution is denoted by $t(x) \sim R(x)$. We assume that each sample from R is drawn independently. The predictor R is said to be deterministic if, for every $x \in \mathcal{X}$, it outputs only unit vectors.

Throughout this work, we evaluate notions of group fairness, which are parameterized by a collection of sets $\mathcal{C} \subseteq \{0,1\}^{|\mathcal{X}|}$. Each subset $S \in \mathcal{C}$ is referred to as a *subpopulation* or a *group*.

Probabilistic predictors for a deterministic Nature. The study of probabilistic predictors is well-motivated even when Nature is entirely deterministic. Consider the following example: a subpopulation $S \in \mathcal{C}$ consists of two disjoint groups, $S = S_1 \cup S_2$, where $|S_1| = |S_2| = \frac{1}{2}|S|$, such that $S \in \mathcal{C}$ but S_1 and S_2 are computationally indistinguishable from each other (in particular, $S_1, S_2 \notin \mathcal{C}$). Suppose that the ground truth predictor R^* assigns label t to all individuals in S_1 and label t' to all individuals in S_2.

A *deterministic* predictor $\tilde{R}$, which outputs only unit vectors, cannot achieve better accuracy or calibration than simply selecting a (pseudo-)random subset of S of half its size to assign label t, while assigning label t' to the remaining individuals. However, if this predictor were used repeatedly—for instance, by multiple universities in their admission processes—the same individuals would consistently receive the same label, potentially leading to systematic unfairness.

In contrast, a *probabilistic* predictor $\tilde{R}$ could assign label t or t' to each individual with equal probability (i.e., $1/2$). As a result, each application of $\tilde{R}$ would yield different labels for the same individuals, ensuring that unfair outcomes do not persist across multiple usages.

C.1 Multi-Accuracy

Multi-Accuracy for (deterministic) predictors [Hébert-Johnson et al., 2018, Trevisan et al., 2009], requires a predictor to be accurate across multiple subpopulations of the domain. The goal is to ensure that a predictor's expected output is unbiased with respect to Nature, not just globally but also on structured subpopulations. Multi-accuracy is parameterized by a class of subpopulations $\mathcal{C}$, which controls the strength of the guarantee: the richer the class, the stronger the requirement. The following definition extends this notion to the probabilistic setting and requires approximate accuracy on each label in $\mathcal{T}$ individually. We therefore refer to it as *coordinate-wise* multi-accuracy.

Definition C.2 (Coordinate-Wise-Multi-Accuracy for Probabilistic Predictors)**.** *Let $\mathcal{C} \subseteq \{0,1\}^{|\mathcal{X}|}$ be a collection of sets. For any error parameter $\alpha \geq 0$, we say that $\tilde{R}$ satisfies $(\mathcal{C}, \alpha)$-coordinate-wise-multi-accuracy w.r.t. $\mathcal{D}$ if for all $S \in \mathcal{C}$ and all $t \in \mathcal{T}$,*

$$\left| \Pr_{\substack{x \sim \mathcal{D}_{\mathcal{X}} \\ t^*(x) \sim R^*(x)}} [t^*(x) = t \wedge x \in S] - \Pr_{\substack{x \sim \mathcal{D}_{\mathcal{X}} \\ \tilde{t}(x) \sim \tilde{R}(x)}} \left[\tilde{t}(x) = t \wedge x \in S\right] \right| \leq \alpha.$$

For the special case where $\mathcal{T}$ is an ordered set, we introduce an alternative (and stronger) definition called threshold-multi-accuracy. We denote $\mathcal{T} = [k]$ and interpret it as a set of ranks. Note that we refer to 1 as the best rank and to k as the worst. To distinguish between sampling a rank and a type, we use the notation $r(x) \sim R(x)$ instead of $t(x) \sim R(x)$. When Nature is deterministic, we often denote it by r^* and assume that it outputs a rank directly rather than a unit vector. In Appendix H.1.1, we provide a detailed discussion on the relationship between different definitions for this special case.

Definition C.3 (Threshold-Multi-Accuracy for Probabilistic Rankings)**.** *Let $\mathcal{C} \subseteq \{0,1\}^{|\mathcal{X}|}$ be a collection of sets. For any error parameter $\alpha \geq 0$, we say that $\tilde{R}$ satisfies $(\mathcal{C}, \alpha)$-threshold-multi-accuracy w.r.t. $\mathcal{D}$ if for every $S \in \mathcal{C}$ and every $\tau \in [k]$,*

$$\left| \Pr_{\substack{x \sim \mathcal{D}_{\mathcal{X}} \\ r^*(x) \sim R^*(x)}} [r^*(x) > \tau \wedge x \in S] - \Pr_{\substack{x \sim \mathcal{D}_{\mathcal{X}} \\ \tilde{r}(x) \sim \tilde{R}(x)}} [\tilde{r}(x) > \tau \wedge x \in S] \right| \leq \alpha.$$

Remark C.4. *In Definition C.3, we define multi-accuracy with respect to threshold events of the form $\{r(x) > \tau\}$. However, both this threshold condition and the coordinate-wise condition in Definition C.2 can be generalized to arbitrary subsets of labels, without increasing the complexity of learning. This can be useful in cases a loss function is known in advance. See Remark D.5 for further discussion of the implications for error bounds and runtime.*

In Appendix H.1.1 we discuss the relationship between threshold-multi-accuracy and coordinate-wise-multi-accuracy. In particular, Claim H.4 and Claim H.5 prove that the threshold definition is strictly stronger.

Next, we show that accuracy on a set implies accuracy on any large enough subset of it. The claim applies to both multi-accuracy definitions presented above; however, for concreteness, we

state it in terms of coordinate-wise-multi-accuracy for types (Definition C.2). Its proof appears in Appendix H.1.

Claim C.5. *Let S, S' be two sets such that $S' \subseteq S$. If $\tilde{R}$ is α-coordinate-wise-accurate w.r.t. $\mathcal{D}$ on S, then it is α'-coordinate-wise-accurate w.r.t. $\mathcal{D}$ on S', where $\alpha' = \alpha + \Pr_{x\sim\mathcal{D}_\mathcal{X}}[x \in S \setminus S']$.*

Finally, we show that these notions can be learned efficiently. Hébert-Johnson et al. [2018] demonstrated a connection between learning multi-accurate and multi-calibrated predictors and weak agnostic learning over the collection $\mathcal{C}$. In particular, this suggests that structural properties of $\mathcal{C}$ can be exploited, if present. Following their approach, we show that efficient weak agnostic learning over $\mathcal{C}$ implies efficient learning of multi-accurate and multi-calibrated (k-dimensional) predictors on $\mathcal{C}$.

Theorem C.6 (Theorem H.7, simplified)**.** *If there is a weak agnostic learner for $\mathcal{C}$ that runs in time T, then there is an algorithm for learning $(\mathcal{C}, \alpha)$-coordinate-wise-multi-accurate probabilistic predictor $\tilde{R}$ that runs in time $\tilde{O}(k \cdot T/\alpha^6)$ and uses $\tilde{O}(1/\alpha^2)$ samples.*

As noted in Theorem H.7, these complexity bounds also apply to learning threshold-multi-accuracy probabilistic rankings. We refer the reader to Appendix H.1.2 for full details.

C.2 Multi-Calibration

Next, we define Multi-Calibration, a natural strengthening of Multi-Accuracy. While calibration remains meaningful even when allowing arbitrary predictions $\tilde{R}(x) \in \mathcal{Y}$, computational feasibility in our constructions necessitates maintaining a certain level of discretization. Formally, we introduce the following definition of λ-discretization of the $[0,1]$ interval.

Definition C.7 (λ-discretization)**.** *Let $\lambda > 0$. The* λ-discretization *of $[0,1]$, denoted by $\Lambda[0,1] = \left\{\frac{\lambda}{2}, \frac{3\lambda}{2}, \ldots, 1 - \frac{\lambda}{2}\right\}$, is the set of $1/\lambda$ evenly spaced real values over $[0,1]$. For $v \in \Lambda[0,1]$, let*

$$\lambda(v) = [v - \lambda/2, v + \lambda/2)$$

be the λ-interval centered around v (except for the final interval, which is $[1-\lambda, 1]$).

To define the values on which we condition to measure approximate calibration, akin to the "level sets" in Hébert-Johnson et al. [2018], we consider two approaches. The first approach conditions on all possible λ-discretized outputs of the predictor, and is therefore referred to as *full*-multi-calibration.

Definition C.8 (Full-Multi-Calibration for Probabilistic Predictors)**.** *Let $\mathcal{C} \subseteq \{0,1\}^{|\mathcal{X}|}$ be a collection of sets and $\lambda \in (0,1)$ a discretization parameter. For any set $S \in \mathcal{C}$ and any λ-discretized vector $R \in (\Lambda[0,1])^k$, we define S_R w.r.t. $\tilde{R} : \mathcal{X} \to \mathcal{Y}$ as:*

$$S_R = \left\{ x \in S \text{ such that } \forall t \in \mathcal{T},\ \Pr_{\tilde{t}(x)\sim\tilde{R}(x)}\left[\tilde{t}(x) = t\right] \in \lambda\left(R_t\right) \right\}.$$

For any error parameter $\alpha \geq 0$, we say that $\tilde{R}$ satisfies $(\mathcal{C}, \alpha, \lambda)$-full-multi-calibration w.r.t. $\mathcal{D}$ if for all $S \in \mathcal{C}$, $R \in (\Lambda[0,1])^k$ and $t \in \mathcal{T}$,

$$\left| \Pr_{\substack{x\sim\mathcal{D}_\mathcal{X} \\ t^*(x)\sim R^*(x)}}\left[t^*(x) = t \wedge x \in S_R\right] - \Pr_{\substack{x\sim\mathcal{D}_\mathcal{X} \\ \tilde{t}(x)\sim\tilde{R}(x)}}\left[\tilde{t}(x) = t \wedge x \in S_R\right] \right| \leq \alpha.$$

The number of sets to consider is exponential in k, i.e., $\left|(S_R)_{R\in(\Lambda[0,1])^k}\right| = \left(\frac{1}{\lambda}\right)^k$, which for some settings of the parameters could be computationally infeasible. For this reason, we deploy relaxed notions of multi-calibration when they are sufficient. We propose a relaxed definition where approximate calibration is measured after sampling a label from the probabilistic predictor, hence the number of sets to consider is $\left|(S_{t,\beta})_{t\in\mathcal{T},\beta\in\Lambda[0,1]}\right| \leq k/\lambda$, which admits an efficient learning procedure as described in Theorem H.8.

Beyond serving as a natural relaxation of full-multi-calibration, that allows for efficient computation, this notion is used in our first negative result (Theorem E.1). Notably, it strengthens the impossibility result by demonstrating it with respect to this relaxed notion rather than full-multi-calibration.

Definition C.9 (Coordinate-Wise-Multi-Calibration for Probabilistic Predictors)**.** *Let $\mathcal{C} \subseteq \{0,1\}^{|\mathcal{X}|}$ be a collection of sets and $\lambda \in (0,1)$ a discretization parameter. For any set $S \in \mathcal{C}$, label $t \in \mathcal{T}$, and discretized probability $\beta \in \Lambda[0,1]$, define $S_{t,\beta}$ w.r.t. $\tilde{R} : \mathcal{X} \to \mathcal{Y}$ as:*

$$S_{t,\beta} := \left\{ x \in S \text{ such that } \Pr_{\tilde{t}(x)\sim\tilde{R}(x)}[\tilde{t}(x) = t] \in \lambda(\beta) \right\}.$$

For any error parameter $\alpha \geq 0$, we say that $\tilde{R}$ satisfies $(\mathcal{C}, \alpha, \lambda)$-coordinate-wise-multi-calibration w.r.t. $\mathcal{D}$ if for all $S \in \mathcal{C}$, $t \in \mathcal{T}$ and $\beta \in \Lambda[0,1]$,

$$\left| \Pr_{\substack{x\sim\mathcal{D}_{\mathcal{X}} \\ t^*(x)\sim R^*(x)}} [t^*(x) = t \wedge x \in S_{t,\beta}] - \Pr_{\substack{x\sim\mathcal{D}_{\mathcal{X}} \\ \tilde{t}(x)\sim\tilde{R}(x)}} \left[\tilde{t}(x) = t \wedge x \in S_{t,\beta}\right] \right| \leq \alpha.$$

The following claim establishes that coordinate-wise-multi-calibration (and, in particular, full multi-calibration) is a stronger requirement than coordinate-wise-multi-accuracy. The proof applies to both the coordinate-wise and threshold-based definitions and is presented in the coordinate-wise setting for concreteness.

Claim C.10 (Coordinate-Wise-Multi-Calibration Implies Coordinate-Wise-Multi-Accuracy)**.** *In the setting of Definition C.9, if $\tilde{R}$ is $(\mathcal{C}, \alpha, \lambda)$-coordinate-wise-multi-calibrated, then $\tilde{R}$ is also $(\mathcal{C}, \alpha/\lambda)$-coordinate-wise-multi-accurate.*

To illustrate the motivation for Multi-Calibration, consider again the example where $S = S_1 \cup S_2$, with S_1, S_2 indistinguishable and Nature defined such that $R^*(x) = t$ for all $x \in S_1$ and $R^*(x) = t'$ for all $x \in S_2$. Compare two predictors: (1) $\tilde{R}_1$: assigns t' to all $x \in S_1$ and t to all $x \in S_2$. (2) $\tilde{R}_2$: for all $x \in S$,

$$\tilde{R}_2(x) = \begin{cases} t \text{ w.p. } 1/2, \\ t' \text{ w.p. } 1/2. \end{cases}$$

The predictor $\tilde{R}_2$ satisfies perfect coordinate-wise-multi-calibration on S (that is, with error $\alpha = 0$), and even perfect full-multi-calibration (and thus perfect coordinate-wise-multi-accuracy), whereas $\tilde{R}_1$ has a coordinate-wise-multi-calibration error of $1/2$ on S, despite being perfectly coordinate-wise-multi-accurate. This example illustrates that while coordinate-wise-multi-accuracy ensures unbiased predictions over subpopulations, only multi-calibration prevents systematic misallocation of probabilities, ensuring that predicted distributions align with the true conditional distributions.

We conclude this section by showing that coordinate-wise-multi-calibration can be learned efficiently. As in Theorem C.6, we formulate this in terms of weak agnostic learning over $\mathcal{C}$.

Theorem C.11 (Theorem H.8, simplified)**.** *If there is a weak agnostic learner for $\mathcal{C}$ that runs in time T, then there is an algorithm for learning $(\mathcal{C}, \alpha, \lambda)$-coordinate-wise-multi-calibrated probabilistic predictor $\tilde{R}$ that runs in time $\tilde{O}(k \cdot T/(\lambda\alpha^6))$ and uses $\tilde{O}(1/\alpha^4)$ samples.*

C.3 Decision Rules

The following definitions and claims consider a randomized predicate $f : \mathcal{Y} \to \{0,1\}$, which we later refer to as a *decision rule*.

Definition C.12 (Lipschitz)**.** *Let $M \in \mathbb{R}^+$ be a positive constant. A randomized predicate $f : \mathcal{Y} \to \{0,1\}$ is called M-*Lipschitz *if $\forall y, y' \in \mathcal{Y}$,*

$$\left| \mathop{\mathbb{E}}_{f}[f(y) - f(y')] \right| \leq M \cdot ||y - y'||_\infty.$$

Definition C.13 (ε-Close to Affine)**.** *Let $\varepsilon \in [0,1]$. A randomized predicate $f : \mathcal{Y} \to \{0,1\}$ is called* ε-close to affine *if for any $y, y' \in \mathcal{Y}$ and any $\gamma \in [0,1]$,*

$$\left| \mathop{\mathbb{E}}_{f}[f(\gamma y + (1-\gamma)y')] - \left(\gamma \mathop{\mathbb{E}}_{f}[f(y)] + (1-\gamma) \mathop{\mathbb{E}}_{f}[f(y')] \right) \right| \leq \varepsilon.$$

A predicate is affine if it is 0-close to affine, and it is ε-far from affine if it is not ε-close to affine. The following claim shows that the definition with two points (Definition C.13) suffices to ensure that the inequality holds with any number of points, with an increase of at most a factor of 2. The proof of the claim and its corollary can be found in Appendix H.2.

Claim C.14. *Let $m \in \mathbb{N}$, $\gamma_1, \ldots, \gamma_m \in [0,1]$ s.t. $\sum_{i=1}^{m} \gamma_i = 1$, and $y_1, \ldots, y_m \in \mathcal{Y}$. Let $\varepsilon \in [0,1]$ and assume that $f : \mathcal{Y} \to \{0,1\}$ is ε-close to affine. Then,*

$$\left| \mathop{\mathbb{E}}_{f} \left[f\left(\sum_{i=1}^{m} \gamma_i y_i \right) \right] - \sum_{i=1}^{m} \gamma_i \mathop{\mathbb{E}}_{f} [f(y_i)] \right| \leq 2\varepsilon.$$

Corollary C.15. *If $f : \mathcal{Y} \to \{0,1\}$ is ε-close to affine, then there exists an affine randomized predicate $f' : \mathcal{Y} \to \{0,1\}$ such that $\forall y \in \mathcal{Y}$,*

$$|\mathop{\mathbb{E}}_{f}[f(y)] - \mathop{\mathbb{E}}_{f'}[f'(y)]| \leq 2\varepsilon.$$

The following claim shows that for a Lipschitz predicate, any two points that violate its affinity must be sufficiently separated. Its proof can be found in Appendix H.2 as well.

Claim C.16. *Let $\varepsilon \in [0,1]$ and $M \in \mathbb{R}^+$. Let $f : \mathcal{Y} \to \{0,1\}$ be a randomized predicate such that:*

- *f is M-Lipschitz;*
- *f is ε-far from affine: that is, there exists $y, y' \in \mathcal{Y}$ and $\gamma \in (0,1)$ such that*

$$\left| \mathop{\mathbb{E}}_{f} \left[f\left(\gamma y + (1-\gamma) y' \right) \right] - \left(\gamma \mathop{\mathbb{E}}_{f} [f(y)] + (1-\gamma) \mathop{\mathbb{E}}_{f} [f(y')] \right) \right| > \varepsilon.$$

Then, $||y - y'||_\infty \geq \varepsilon / M$.

C.4 Indistinguishability

The impossibility results (Theorems E.1 and F.1) rely on the existence of an indistinguishable subset with respect to the uniform distribution. We present two methods for constructing such a subset: one based on cryptographic assumptions, specifically the existence of one-way functions, and another that is information-theoretic, applicable when no limitation is placed on the complexity of Nature. We begin with a formal definition of indistinguishability.

Definition C.17 (Indistinguishable Subset)**.** *Let $\mathcal{X}$ be a domain and let $\mathcal{D}_\mathcal{X}$ be a distribution. For a security parameter $\kappa \in \mathbb{N}$, we say that $\emptyset \subsetneq \mathcal{X}_1 \subsetneq \mathcal{X}$ is a* κ-indistinguishable subset *if for any $x \sim \mathcal{D}_\mathcal{X}$, membership in $\mathcal{X}_1$ is not efficiently decidable (w.r.t. κ). That is, for every probabilistic $\mathsf{poly}(\kappa)$-time algorithm $\mathcal{A}$ and all sufficiently large κ,*

$$\left| \Pr_{\mathcal{A}, x \sim \mathcal{D}_\mathcal{X}} [\mathcal{A}(x) = 1 \mid x \in \mathcal{X}_1] - \Pr_{\mathcal{A}, x \sim \mathcal{D}_\mathcal{X}} [\mathcal{A}(x) = 1 \mid x \in \mathcal{X} \setminus \mathcal{X}_1] \right| = \mathsf{neg}(\kappa),$$

where $\mathsf{neg}(\cdot)$ denotes a negligible function, i.e., a function $\mathsf{neg} : \mathbb{N} \to [0,1]$ such that for every polynomial $p(\cdot)$, there exists $N \in \mathbb{N}$ for which $\mathsf{neg}(\kappa) < 1/p(\kappa)$ for all $\kappa \geq N$.

Note that the probabilities are well-defined (i.e., we do not condition on an event of probability zero), since the definition requires $\mathcal{X}_1$ to be a non-empty proper subset. Indeed, both the empty set and the entire universe $\mathcal{X}$ are never indistinguishable.

Cryptographic approach. In what follows, we show how to use cryptographic assumptions to construct an indistinguishable subset of a given size. In particular, we assume the existence of one-way functions (OWFs), often regarded as the "minimal" assumption in cryptography. The construction of the subset is computationally efficient; however, under the assumption that OWFs exist, determining membership in the subset is computationally infeasible. The proof is deferred to Appendix H.3.

Lemma C.18. *Let $\mathcal{X}$ be a domain, and let $\mathcal{U}_\mathcal{X}$ denote the uniform distribution over $\mathcal{X}$. Assume one-way functions exist w.r.t. security parameter $\kappa \geq \log |\mathcal{X}|$. For any constant $\gamma \in (0,1)$, there exists a subset $\mathcal{X}_1 \subset \mathcal{X}$ such that*

- $|\Pr_{x\sim\mathcal{U}_\mathcal{X}}[x \in \mathcal{X}_1] - \gamma| = \mathsf{neg}(\kappa)$*;*
- $\mathcal{X}_1$ *is* κ*-indistinguishable w.r.t.* $\mathcal{U}_\mathcal{X}$*;*
- *There exists a secret key such that, given the key, membership in* $\mathcal{X}_1$ *is efficiently computable.*

Information-theoretic approach. The alternative approach we present is information-theoretic. If we do not impose any computational constraints on the way we model Nature, then determining membership in a subset defined by this circuit can be computationally infeasible. In particular, if Nature is sufficiently "complex" to resemble a random function, it may inherently contain an indistinguishable subset.

We emphasize that this perspective does not require Nature to be a complex function that runs in superpolynomial time; rather, it reflects that, given our *partial view* of the feature set defining an individual, Nature appears indistinguishable from a random function. It remains entirely possible that, with access to the *full* set of features, Nature could still be computed efficiently.

In this setting as well, for any constant $\gamma \in (0,1)$, we can assume the existence of a Nature that admits an indistinguishable subset of size γ, up to a negligible rounding error (arising from the chosen binary representation of individuals).

The following claim applies in both settings, when considering the uniform distribution over the domain $\mathcal{X}$. Its proof is deferred to Appendix H.3 as well.

Claim C.19. *Let* $\mathcal{X}$ *be a domain, and let* $\mathcal{U}_\mathcal{X}$ *denote the uniform distribution over* $\mathcal{X}$*. Assume* $\mathcal{X}$ *contains a* κ*-indistinguishable subset* $\mathcal{X}_1$ *as per Definition C.17 w.r.t. a security parameter* $\kappa \geq \log|\mathcal{X}|$*. For any* $\mathsf{poly}(\kappa)$*-time algorithm* $\mathcal{A} : \mathcal{X} \to \{0,1\}$*, if* $\Pr[\mathcal{A}(x)=1]$ *is non-negligible in* κ*, then*

$$\left| \Pr_{\substack{x\sim\mathcal{U}_\mathcal{X} \\ \mathcal{A}\ \textit{randomness}}}[x \in \mathcal{X}_1 | \mathcal{A}(x)=1] - \Pr_{x\sim\mathcal{U}_\mathcal{X}}[x \in \mathcal{X}_1] \right| = \mathsf{neg}(\kappa).$$

Notice that an immediate corollary of the claim shows that any efficiently-recognizable subset $\tilde{\mathcal{X}} \subseteq \mathcal{X}$ of non-negligible size intersects an indistinguishable subset $\mathcal{X}_1 \subsetneq \mathcal{X}$ in approximately the same fraction as $\mathcal{X}$ does. This follows by interpreting $\mathcal{A}$ as a (deterministic) procedure that decides membership in a set $\tilde{\mathcal{X}}$.

D Translating accuracy of predictions to accuracy of classifications

We now turn to the study of decision rules, also known as post-processing functions. A typical approach to decision-making is to approximate R^* by an efficient predictor $\tilde{R}$ and then apply a decision rule to $\tilde{R}$ in the same manner that one would apply it to R^*. For a collection of sets $\mathcal{C}$, we define *multi-accuracy-on-decision* to mean that for any set $S \in \mathcal{C}$, members of S receive the same distribution of decisions when the decision rule is applied to $\tilde{R}$ as they would when it is applied to R^*. Formally, a decision rule $\rho : \mathcal{Y} \to \{0,1\}$ takes as input a distribution over labels and outputs either 0 or 1, potentially in a randomized manner. Any randomized predicate, as defined above, can serve as a (randomized) decision rule.

Definition D.1 (Multi-Accuracy-on-Decision)**.** *For a collection of sets* $\mathcal{C} \subseteq \{0,1\}^{|\mathcal{X}|}$ *and* $\varepsilon \geq 0$*, a probabilistic predictor* $\tilde{R} : \mathcal{X} \to \mathcal{Y}$ *satisfies* $(\mathcal{C}, \varepsilon)$*-multi-accuracy-on-decision w.r.t.* $\mathcal{D}$ *and* ρ *if for all* $S \in \mathcal{C}$*,*

$$\left| \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X},\rho}\left[\rho\left(\tilde{R}(x)\right) \mid x \in S\right] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X},\rho}\left[\rho\left(R^*(x)\right) \mid x \in S\right] \right| \leq \varepsilon.$$

Characterization of affine decision rules. Before stating the result, we present an alternative characterization of affine decision rules: those that first sample a label and then apply a function that (randomly) maps the label to a decision in $\{0,1\}$. This characterization will be useful in the proof of Theorem D.4.

Definition D.2. *A randomized decision rule $\rho : \mathcal{Y} \to \{0,1\}$ is called an* Instantiate-Then-Act *rule if there exists a randomized function $g : \mathcal{T} \to \{0,1\}$ such that ρ works as follows: On input $y \in \mathcal{Y}$, it samples $t \sim y$ and outputs $g(t)$.*

Lemma D.3. *Instantiate-Then-Act (ITA) rules are equivalent to Affine rules (Definition C.13).*

Proof. First, we show that affineness implies ITA. Let $f : \mathcal{Y} \to \{0,1\}$ be an affine rule. By Claim C.14, for any $z_1, \ldots, z_k \in \mathcal{Y}$ and any $\gamma_1, \ldots, \gamma_k \in [0,1]$,

$$\mathop{\mathbb{E}}_{f}\left[f\left(\sum_{i=1}^{k}\gamma_i z_i\right)\right] = \sum_{i=1}^{k}\gamma_i \mathop{\mathbb{E}}_{f}\left[f(z_i)\right].$$

We define $g : \mathcal{T} \to \{0,1\}$ as follows: for any $t \in \mathcal{T}$, $g(t)$ is distributed like $f(e_t)$ where e_t is the t^{th} unit vector. Notice that, defining $\rho : \mathcal{Y} \to \{0,1\}$ such that on input y it samples $t \sim y$ and outputs $g(t)$, we get that

$$\Pr_{\rho}[\rho(y)=1] = \mathop{\mathbb{E}}_{t\sim y}\left[g(t)\right] = \sum_{t\in\mathcal{T}}\Pr_{y}[t=t']\,\mathbb{E}\left[g(t)\right] = \sum_{t\in\mathcal{T}} y_t\,\mathbb{E}\left[g(t)\right] = \sum_{t\in\mathcal{T}} y_t\,\mathbb{E}\left[f(e_t)\right]$$

and on the other hand,

$$\Pr_{f}\left[f(y)=1\right] = \mathbb{E}\left[f\left(\sum_{t\in\mathcal{T}} y_t e_t\right)\right] = \sum_{t\in\mathcal{T}} y_t\,\mathbb{E}[f(e_t)].$$

Hence, f and ρ are identically distributed. Next, we show that ITA implies affineness. Let $g : \mathcal{T} \to \{0,1\}$ and let ρ be as in Definition D.2. We define $f : \mathcal{Y} \to \{0,1\}$ so that, on input $y \in \mathcal{Y}$, the random variable $f(y)$ follows a $\mathrm{Ber}\left(\sum_{t\in\mathcal{T}} y_t\,\mathbb{E}[g(t)]\right)$ distribution. It follows directly that f and ρ are identically distributed. To see that ρ is affine, let $y, y' \in \mathcal{Y}$ and $\gamma \in [0,1]$. Notice that

$$\begin{aligned}
\mathbb{E}\left[f\left(\gamma y + (1-\gamma)y'\right)\right] &= \sum_{t\in\mathcal{T}}\left(\gamma y_t + (1-\gamma)y'_t\right)\mathbb{E}[g(t)] \\
&= \gamma\sum_{t\in\mathcal{T}} y_t\,\mathbb{E}[g(t)] + (1-\gamma)\sum_{t\in\mathcal{T}} y'_t\,\mathbb{E}[g(t)] \\
&= \gamma\,\mathbb{E}\left[f(y)\right] + (1-\gamma)\,\mathbb{E}\left[f(y')\right],
\end{aligned}$$

and the lemma follows. □

The following theorem shows that a coordinate-wise-multi-accurate (coordinate-wise-MA) predictor, as defined in Definition C.2, implies multi-accuracy-on-decision (MAD) with respect to any decision rule that is close to affine.

Theorem D.4 (coordinate-wise-MA implies MAD)**.** *Fix $\mathcal{C} \subseteq \{0,1\}^{|\mathcal{X}|}$ and let $\delta := \min_{S\in\mathcal{C}}\Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[x \in S]$. Let $\rho : \mathcal{Y} \to \{0,1\}$ be a decision rule that is ε-close to affine. If a predictor $\tilde{R}$ is $(\mathcal{C},\alpha)$-coordinate-wise-multi-accurate w.r.t. $\mathcal{D}$, then it is $(\mathcal{C},\varepsilon')$-multi-accurate-on-decision w.r.t. $\mathcal{D}$ and ρ, where $\varepsilon' \le k\alpha/\delta + 2\varepsilon$.*

Furthermore, if $\mathcal{T}$ is an ordered set, $\tilde{R}$ is $(\mathcal{C},\alpha)$-threshold-multi-accurate and ρ is deterministic and monotone on unit vectors, then $\varepsilon' \le \alpha/\delta + 2\varepsilon$.

Proof. By Corollary C.15, there exists an affine decision rule ρ' such that $\forall y \in \mathcal{Y}$,

$$\left|\mathop{\mathbb{E}}_{\rho'}[\rho'(y)] - \mathop{\mathbb{E}}_{\rho}[\rho(y)]\right| \le 2\varepsilon.$$

Since ρ' is affine, Lemma D.3 implies that it can also be viewed as an Instantiate-Then-Act (ITA) rule. That is, there exists a randomized function $g : \mathcal{T} \to \{0,1\}$ such that ρ' works as follows: On

input $y \in \mathcal{Y}$, it samples $t \sim y$ and outputs $g(t)$. This implies that, for any $S \in \mathcal{C}$,

$$\left| \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}, \rho'} \left[\rho' \left(\tilde{R}(x) \right) \mid x \in S \right] - \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}, \rho'} \left[\rho' \left(R^*(x) \right) \mid x \in S \right] \right| =$$

$$\left| \mathop{\mathbb{E}}_{x, \tilde{t}(x), g} \left[g \left(\tilde{t}(x) \right) \mid x \in S \right] - \mathop{\mathbb{E}}_{x, t^*(x), g} \left[g \left(t^*(x) \right) \mid x \in S \right] \right| =$$

$$\left| \sum_{t \in \mathcal{T}} \left(\Pr_{x, \tilde{t}(x)} [\tilde{t}(x) = t \mid x \in S] - \Pr_{x, t^*(x)} [t^*(x) = t \mid x \in S] \right) \mathop{\mathbb{E}}_{g} [g(t)] \right| \leq$$

$$\sum_{t \in \mathcal{T}} \left| \Pr_{x, \tilde{t}(x)} [\tilde{t}(x) = t \mid x \in S] - \Pr_{x, t^*(x)} [t^*(x) = t \mid x \in S] \right| \leq \frac{k\alpha}{\delta}.$$

This means that

$$\left| \mathbb{E} \left[\rho \left(\tilde{R}(x) \right) \mid x \in S \right] - \mathbb{E} \left[\rho \left(R^*(x) \right) \mid x \in S \right] \right| \leq \frac{k\alpha}{\delta} + 2\varepsilon.$$

For the furthermore part, without loss of generality let us assume that $\mathcal{T} = [k]$. By the assumption, ρ is deterministic and monotone on unit vectors. Notice that ρ' as defined in Corollary C.15 coincides with ρ on unit vectors. On the other hand, expressing ρ' as an ITA rule, ρ' also coincides with g on unit vectors, meaning that for all $\tau \in [k]$,

$$g(\tau) = \rho(e_\tau).$$

This implies that g is both deterministic and monotone on unit vectors, i.e.,

$$\forall \tau \in [k], \ \mathop{\mathbb{E}}_{g} g(\tau) \in \{0, 1\} \text{ and } \forall \tau_1 < \tau_2, \ g(\tau_1) \geq g(\tau_2). \tag{4}$$

Therefore, g can be interpreted as a threshold function, meaning there exists $\bar{\tau} \in [k]$ such that[8]

$$g(\tau) = \begin{cases} 1 & \text{if } \tau \leq \bar{\tau}, \\ 0 & \text{otherwise.} \end{cases}$$

Hence, in this case, MAD translates to

$$\left| \mathop{\mathbb{E}}_{x, \tilde{r}(x), g} \left[g \left(\tilde{r}(x) \right) \mid x \in S \right] - \mathop{\mathbb{E}}_{x, r^*(x), g} \left[g \left(r^*(x) \right) \mid x \in S \right] \right| =$$

$$\left| \Pr_{x, \tilde{r}(x)} [\tilde{r}(x) \leq \bar{\tau} \mid x \in S] - \Pr_{x, r^*(x)} [r^*(x) \leq \bar{\tau} \mid x \in S] \right|,$$

which removes the k multiplicative blowup, if $\tilde{R}$ is threshold-multi-accurate according to Definition C.3. Importantly, as noted in Appendix C.1, threshold-multi-accuracy can be achieved with the same sample and time complexity as coordinate-wise multi-accuracy, despite being a mathematically stronger notion. □

Remark D.5. *As mentioned in Remark C.4, multi-accuracy can alternatively be required with respect to a subset or a collection of subsets of labels. Referring back to Equation* (4)*, such definitions correspond to deterministic functions g (i.e., functions that output either 0 or 1 for each label). As in the threshold-based case, this eliminates the multiplicative factor of k in the error bound.*

Next, we define Multi-Accuracy for Classification. To this end, we first formally define loss minimization for an action function. An action function is a randomized mapping $h : \mathcal{X} \to \{0, 1\}$ that assigns individuals to either 0 or 1, which can be interpreted as "accepted" or "rejected." An action function *may* be constructed via a decision rule composed with a predictor, but it can also be defined directly. More generally, decision rules or action functions may have more than two possible actions, but we restrict attention to the binary case for simplicity.

Definition D.6 (Loss Minimization)**.** *Let $\ell : \mathcal{T} \times \{0, 1\} \to [0, 1]$ be a loss function, and let $\mathcal{H} \subseteq \{\mathcal{X} \to \{0, 1\}\}$ be a class of action functions. For $\varepsilon \geq 0$, an action function h is said to be $(\mathcal{H}, \ell, \varepsilon)$-loss-minimizing if*

$$\forall h' \in \mathcal{H}, \quad \mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D}_{\mathcal{X}} \\ h \text{ randomness} \\ t^*(x) \sim R^*(x)}} \left[\ell \left(t^*(x), h(x) \right) \right] \leq \mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D}_{\mathcal{X}} \\ h' \text{ randomness} \\ t^*(x) \sim R^*(x)}} \left[\ell(t^*(x), h'(x)) \right] + \varepsilon.$$

[8] This does not cover the case where $\rho' \equiv 0$, but in that case, the argument follows trivially.

We denote the expected loss of h over $\mathcal{D}$ by $\mathsf{exp\text{-}loss}_{\mathcal{D}}(h)$. Any loss function induces a loss-minimizing decision rule, which we formally define as follows:

Definition D.7 (Loss-Minimizing Decision Rule)**.** *Let $\ell : \mathcal{T} \times \{0,1\} \to [0,1]$ be a loss function. The* loss-minimizing decision rule $\rho^*_\ell : \mathcal{Y} \to \{0,1\}$ *with respect to ℓ is defined as*

$$\rho^*_\ell(y) = \begin{cases} 1 & \text{if } \mathbb{E}_{t\sim y}\left[\ell(t,1) - \ell(t,0)\right] < 0, \\ 0 & \text{otherwise.} \end{cases}$$

Equivalently, we can express this as

$$\rho^*_\ell(y) = \arg\min_{a\in\{0,1\}} \mathbb{E}_{t\sim y}\left[\ell(t,a)\right],$$

where ties—cases in which the expected losses for acceptance and rejection are equal—are broken in favor of rejection. Notice that this decision rule is deterministic, and independent of $\mathcal{D}$. We additionally define the loss-minimizing action function, which is obtained by applying the loss-minimizing decision rule to Nature's predictor. This action function achieves the minimum possible loss compared to any other action function. In fact, this motivates our definition of the loss-minimizing decision rule: it makes optimal decisions under the assumption that Nature's predictor provides the true underlying probabilities.

Definition D.8 (Loss-Minimizing Action Function)**.** *Let $\rho^*_\ell : \mathcal{Y} \to \{0,1\}$ be as in Definition D.7. The* loss-minimizing action function $h^*_\ell : \mathcal{X} \to \{0,1\}$ *w.r.t. ℓ is defined as*

$$h^*_\ell = \rho^*_\ell \circ R^*.$$

Note that h^*_ℓ is deterministic as well. The following fairness criterion measures the accuracy of representation across different subpopulations *after classification*. We say that an action function h is Accurate-on-Classification if the fraction of individuals accepted according to h is close to the fraction accepted by h^*_ℓ. We formally define this measure in the general multigroup setting.

Definition D.9 (Multi-Accuracy-on-Classification)**.** *Let $\ell : \mathcal{T} \times \{0,1\} \to [0,1]$ be a loss function, and let $\mathcal{C} \subseteq \{0,1\}^{|\mathcal{X}|}$ be a collection of sets. We say that $h : \mathcal{X} \to \{0,1\}$ is $(\mathcal{C}, \ell, \varepsilon)$-multi-accurate-on-classification w.r.t. $\mathcal{D}$ if for all $S \in \mathcal{C}$,*

$$\left| \underset{\substack{x\sim\mathcal{D}_{\mathcal{X}} \\ h \text{ randomness}}}{\mathbb{E}} \left[h(x) \mid x \in S\right] - \underset{x\sim\mathcal{D}_{\mathcal{X}}}{\mathbb{E}} \left[h^*_\ell(x) \mid x \in S\right] \right| \leq \varepsilon.$$

We remark that one could define Multi-Accuracy-on-Classification (MAC) with respect to an arbitrary decision rule ρ^* instead of ρ^*_ℓ, by requiring that the fraction of individuals accepted according to h is close to the fraction accepted by $\rho^* \circ R^*$. Achieving MAC under this alternative definition can be done using the same argument presented next in Corollary D.10.

We now show that MAC action functions can be obtained efficiently. As a corollary of Theorem D.4, we define an affine decision rule and show that composing it with any multi-accurate predictor yields a MAC action function. In particular, this implies that a single predictor can simultaneously satisfy multi-accuracy-on-decision (MAD) with respect to a decision rule, and satisfy MAC when composed with that rule. As shown in the following corollary, the rule we define can be viewed as an affine analogue of the loss-minimizing decision rule.

Notice that this corollary holds only when Nature is deterministic, i.e., R^* outputs unit vectors for any $x \in \mathcal{X}$. See Remark D.13 for more details on this point.

Corollary D.10 (coordinate-wise-MA implies MAC)**.** *Let ℓ be a loss function and assume R^* is deterministic. In the setting of Theorem D.4, there exists a decision rule ρ such that for any $(\mathcal{C}, \alpha)$-coordinate-wise-multi-accurate predictor $\tilde{R}$, the action function $h = \rho \circ \tilde{R}$ is $(\mathcal{C}, \ell, \varepsilon)$-multi-accurate-on-classification for $\varepsilon \leq k\alpha/\delta$.*

Proof. We define the function $g : \mathcal{T} \to \{0,1\}$ as follows:

$$\forall t \in \mathcal{T},\ g(t) = \rho^*_\ell(e_t) = \arg\min_{a\in\{0,1\}}\{\ell(t,a)\},$$

where e_t is the t^{th} unit vector. Let $\rho : \mathcal{Y} \to \{0,1\}$ be defined as an ITA rule (Definition D.2) w.r.t. g, that is: on input $y \in \mathcal{Y}$, it samples $t \sim y$ and outputs $g(t)$. Since ρ is affine, then Theorem D.4 implies that $\tilde{R}$ is $(\mathcal{C}, \varepsilon)$-MAD w.r.t. $\mathcal{D}$ and ρ, where $\varepsilon \leq k\alpha/\delta$. That is, for any $S \in \mathcal{C}$,

$$\left| \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}, \rho} \left[\rho\left(\tilde{R}(x)\right) \mid x \in S \right] - \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}, \rho} \left[\rho\left(R^*(x)\right) \mid x \in S \right] \right| \leq \varepsilon.$$

On the other hand, since Nature is deterministic, it holds for any $x \in S$ that

$$\rho\left(R^*(x)\right) = \rho^*_\ell\left(R^*(x)\right) = h^*_\ell(x).$$

Putting it together,

$$\left| \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}, \rho} \left[h(x) \mid x \in S \right] - \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}, \rho} \left[h^*_\ell(x) \mid x \in S \right] \right| \leq \varepsilon.$$

Once again, if the definition of multi-accuracy implies ε-MAD w.r.t. $\mathcal{D}$ and ρ with $\varepsilon \leq \alpha/\delta$, then h achieves $(\mathcal{C}, \ell, \varepsilon)$-MAC with $\varepsilon \leq \alpha/\delta$ as well. □

To conclude this section, we show that MAD can be obtained efficiently for *any* decision rule, not necessarily affine, under the same assumption that Nature is deterministic (see Remark D.13 for further discussion). Compared with Theorem D.4, this implies the following: in the deterministic setting, achieving MAD for an arbitrary decision rule reduces to the task of learning a multi-accurate predictor—a task we have already shown can be performed efficiently. In contrast, in the probabilistic setting, the uncertainty can be overcome only for affine rules.

To formally state this argument, we define a standard primitive known as *random instantiation*.

Definition D.11 (Random Instantiation)**.** *Let $\mathcal{E} \subseteq \mathcal{Y}$ denote the set of unit vectors over $\mathcal{T}$. Let $R : \mathcal{X} \to \mathcal{Y}$ be a probabilistic predictor. A* random instantiation *of R is a function $r : \mathcal{X} \to \mathcal{E}$ defined by independently sampling $r(x) \sim R(x)$ for each $x \in \mathcal{X}$, and returning the corresponding unit vector. We denote this by $r \sim R$.*

The proof is straightforward and follows the same logic presented in Theorem D.4. Notice that Remark D.5 applies to Lemma D.12 as well: the multiplicative factor of k in the error bound can be avoided by adopting a different variant of multi-accuracy, such as threshold-multi-accuracy or multi-accuracy with respect to a collection of label subsets.

Lemma D.12 (Coordinate-Wise Multi-Accuracy Implies MAD for Any Rule)**.** *Assume R^* is deterministic. In the setting of Theorem D.4, for any decision rule ρ and any random instantiation $\tilde{r} \sim \tilde{R}$ of a $(\mathcal{C}, \alpha)$-coordinate-wise-multi-accurate predictor $\tilde{R}$, the predictor $\tilde{r}$ is $(\mathcal{C}, \varepsilon)$-multi-accurate-on-decision with respect to $\mathcal{D}$ and ρ, where $\varepsilon \leq k\alpha/\delta$.*

Proof. Let $\tilde{r} \sim \tilde{R}$ be a random instantiation of a $(\mathcal{C}, \alpha)$-coordinate-wise-multi-accurate predictor $\tilde{R}$, and recall that $e_t \in \mathcal{E}$ denotes the t^{th} unit vector. Then, for any $S \in \mathcal{C}$ and any $t \in \mathcal{T}$, we can write

$$\left| \Pr_{x \sim \mathcal{D}_{\mathcal{X}}} \left[R^*(x) = e_t \wedge x \in S \right] - \Pr_{x \sim \mathcal{D}_{\mathcal{X}}} \left[\tilde{r}(x) = e_t \wedge x \in S \right] \right| \leq \alpha,$$

since R^* is deterministic. Hence, for any decision rule ρ,

$$\begin{aligned}
&\left| \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}, \rho} \left[\rho\left(\tilde{r}(x)\right) \mid x \in S \right] - \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}, \rho} \left[\rho\left(R^*(x)\right) \mid x \in S \right] \right| = \\
&\left| \sum_{t \in \mathcal{T}} \left(\Pr_x[\tilde{r}(x) = e_t \mid x \in S] - \Pr_x[R^*(x) = e_t \mid x \in S] \right) \mathop{\mathbb{E}}_{\rho}[\rho(e_t)] \right| \leq \\
&\sum_{t \in \mathcal{T}} \left| \Pr[\tilde{r}(x) = e_t \mid x \in S] - \Pr[R^*(x) = e_t \mid x \in S] \right| \leq \frac{k\alpha}{\delta}.
\end{aligned}$$

This implies that the MAD error of $\tilde{r}$ is at most $k\alpha/\delta$, and the lemma follows. □

Remark D.13 (Deterministic Nature)**.** *As discussed in the introduction, we consider a rich set of labels that capture the uncertainty of Nature, such as values of the Bayes-optimal predictor, rather than binary outcomes. With such labels, it is appropriate to model Nature as deterministic. A randomized Nature with such rich labels poses a greater challenge than the standard agnostic learning model, and thus we cannot hope to achieve the strong fairness guarantees sought in this work unless Nature is assumed to be deterministic.*

E When accuracy of predictions contradicts accuracy of decisions

In this section, we present our first impossibility result. The following theorem establishes that if a decision rule is far from affine, then there exists a setting of Nature in which it is computationally infeasible to efficiently learn a predictor that is both calibrated and accurate-on-decision, even for a single subpopulation.

Notably, the calibration requirement used here follows the relaxed definition in Definition C.9, thereby strengthening the result by relying on a weaker assumption. Moreover, in contrast to simplified approaches, our proof holds for any discretization parameter λ and avoids trivial bounds on the error parameters. See Remarks E.5 and E.6 for further discussion.

Then, in Appendix E.1, we show that the loss-minimizing decision rule is far from affine, and that this implies that it may be computationally infeasible to learn a calibrated predictor that is accurate-on-decision with respect to the loss-minimizing decision rule.

Theorem E.1 (Calibration and Accuracy-on-Decision Requires Affineness)**.** *Assume one-way functions exist w.r.t. a security parameter $\kappa \geq \max\left\{\log|\mathcal{X}|, O((M/\varepsilon)^2)\right\}$. Let $\varepsilon \in (0,1)$ and $M \in \mathbb{R}^+$. Let $\lambda \in (0,1)$ be a discretization parameter satisfying $\lambda \leq \varepsilon/16M$ and let $\alpha \geq 0$ be an error parameter such that $\alpha = O(\lambda^2\varepsilon/kM)$.*

If $\rho : \mathcal{Y} \to \{0,1\}$ is M-Lipschitz and ε-far from affine, then there exists $\mathcal{D}$ such that there is no $\mathsf{poly}(\kappa)$-time learner that can output with non-negligible probability a predictor $\tilde{R}$ which is both (α,λ)-coordinate-wise-calibrated w.r.t. $\mathcal{D}$ and $\varepsilon/2$-accurate-on-decision w.r.t. $\mathcal{D}$ and ρ.

For convenience, we instantiate the assumption that a computationally hidden subpopulation exists by assuming one-way functions. However, as discussed in Appendix C.4, the assumption of one-way functions can be removed if no restriction is placed on the complexity of Nature.

Proof. Let $\rho : \mathcal{Y} \to \{0,1\}$ be a randomized decision rule which is M-Lipschitz and ε-far from affine (see Definitions C.12 and C.13). This means that there exists $y, y' \in \mathcal{Y}$ and $\gamma \in (0,1)$ such that

$$\left|\mathbb{E}\left[\rho\left(\gamma y + (1-\gamma)y'\right)\right] - \left(\gamma\,\mathbb{E}\left[\rho(y)\right] + (1-\gamma)\,\mathbb{E}\left[\rho(y')\right]\right)\right| > \varepsilon.$$

Let $\mathcal{U}_\mathcal{X}$ denote the uniform distribution over $\mathcal{X}$. First, we perturb γ by adding uniformly random noise $\nu \sim \text{Uniform}\left[-\frac{M}{\varepsilon\kappa}, \frac{M}{\varepsilon\kappa}\right]$. Let

$$\gamma_\nu = \gamma + \nu \tag{5}$$

denote the perturbed version of γ. This "randomization" helps mitigate discretization issues when analyzing how R^* behaves on subsets defined by $\tilde{R}$. A more detailed discussion appears in Remark E.5.

Under the assumption that one-way functions exist, Lemma C.18 implies the existence of a subset $\mathcal{X}_1 \subseteq \mathcal{X}$ of size $|\mathcal{X}_1| = \gamma'|\mathcal{X}|$ for $|\gamma' - \gamma_\nu| = \mathsf{neg}(\kappa)$, such that $\mathcal{X}_1$ is κ-indistinguishable w.r.t. $\mathcal{U}_\mathcal{X}$ (see Definition C.17) and Nature is efficiently computable. Taking $\mathcal{X}_2 := \mathcal{X} \setminus \mathcal{X}_1$,

$$\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2.$$

We define Nature's predictor w.r.t. $\mathcal{U}_\mathcal{X}$ as follows:

$$R^*|_{\mathcal{X}_1} \equiv y \text{ and } R^*|_{\mathcal{X}_2} \equiv y',$$

and define $\mathcal{D} = (\mathcal{U}_\mathcal{X}, R^*(\mathcal{U}_\mathcal{X}))$. Towards contradiction, assume there exists a $\mathsf{poly}(\kappa)$-time learner $\mathcal{L}$, with access to labeled samples from $\mathcal{D}$, that outputs with non-negligible probability a probabilistic predictor which is both (α,λ)-coordinate-wise-calibrated w.r.t. $\mathcal{D}$ (see Definition C.9) and $\varepsilon/2$-accurate-on-decision w.r.t. $\mathcal{D}$ and ρ (see Definition D.1). Take $\tilde{R} : \mathcal{X} \to \mathcal{Y}$ to denote this predictor. Let $(\mathcal{X}_{t,\beta})_{t\in\mathcal{T},\beta\in\Lambda[0,1]}$ be the "level sets" as per the definition of coordinate-wise-calibration:

$$\mathcal{X}_{t,\beta} := \left\{ x \in \mathcal{X} \text{ such that } \Pr_{\tilde{t}(x)\sim\tilde{R}(x)}[\tilde{t}(x) = t] \in \lambda(\beta) \right\}.$$

We stress that these sets are not necessarily disjoint, however, for a fixed t, they form a partition: $\mathcal{X} = \mathcal{X}_{t,\frac{\lambda}{2}} \cup \cdots \cup \mathcal{X}_{t,1-\frac{\lambda}{2}}$. First, we argue that calibration implies that for any label t, only one level set can be large.

Lemma E.2. *For any $t \in \mathcal{T}$, with all but a negligible probability, there exists a unique $\beta_t^* \in \Lambda[0,1]$ for which $\left|\mathcal{X}_{t,\beta_t^*}\right| > \frac{2\alpha}{\lambda}|\mathcal{X}|$.*

Proof. Let $t \in \mathcal{T}$ and $\beta \in \Lambda[0,1]$ such that $|\mathcal{X}_{t,\beta}| > \frac{2\alpha}{\lambda}$. By (α,λ)-coordinate-wise-calibration of $\tilde{R}$,

$$\left| \Pr_{\substack{x\sim\mathcal{U}_\mathcal{X} \\ t^*(x)\sim R^*(x)}} [t^*(x) = t \wedge x \in \mathcal{X}_{t,\beta}] - \Pr_{\substack{x\sim\mathcal{U}_\mathcal{X} \\ \tilde{t}(x)\sim \tilde{R}(x)}} \left[\tilde{t}(x) = t \wedge x \in \mathcal{X}_{t,\beta}\right] \right| \leq \alpha,$$

which implies that

$$\left| \Pr\left[t^*(x) = t \mid x \in \mathcal{X}_{t,\beta}\right] - \Pr\left[\tilde{t}(x) = t \mid x \in \mathcal{X}_{t,\beta}\right] \right| \leq \frac{\alpha}{\Pr\left[x \in \mathcal{X}_{t,\beta}\right]} < \frac{\lambda}{2}. \tag{6}$$

We observe that the first term can be expressed in terms of y and y'. Taking y_t to denote the t^{th} coordinate of y,

$$y_t = \Pr\left[t^*(x) = t \mid x \in \mathcal{X}_{t,\beta} \cap \mathcal{X}_1\right] \text{ and } y'_t = \Pr\left[t^*(x) = t \mid x \in \mathcal{X}_{t,\beta} \cap \mathcal{X}_2\right],$$

hence,

$$\Pr\left[t^*(x) = t \mid x \in \mathcal{X}_{t,\beta}\right] = y_t \cdot \frac{\Pr\left[\mathcal{X}_{t,\beta} \cap \mathcal{X}_1\right]}{\Pr\left[x \in \mathcal{X}_{t,\beta}\right]} + y'_t \cdot \frac{\Pr\left[\mathcal{X}_{t,\beta} \cap \mathcal{X}_2\right]}{\Pr\left[x \in \mathcal{X}_{t,\beta}\right]}. \tag{7}$$

Let us denote

$$\gamma_{t,\beta} := \frac{\Pr\left[\mathcal{X}_{t,\beta} \cap \mathcal{X}_1\right]}{\Pr\left[x \in \mathcal{X}_{t,\beta}\right]},$$

so we can rewrite the right-hand side of Equation (7) as $\gamma_{t,\beta} \cdot y_t + (1 - \gamma_{t,\beta}) \cdot y'_t$.

Recall that the level sets are defined with respect to *discretized* probabilities. Even if, for a fixed β, the probabilities

$$\gamma' \cdot y_t + (1 - \gamma') \cdot y'_t$$

and

$$\gamma_{t,\beta} \cdot y_t + (1 - \gamma_{t,\beta}) \cdot y'_t$$

are very close, they may still be mapped to different λ-intervals by the discretization (see Definition C.7). To prevent this, we introduced random noise to γ and defined R^* with respect to γ_ν (see Remark E.5 for a detailed discussion). Consequently, γ' and $(\gamma_{t,\beta})_{\beta\in\Lambda[0,1]}$ are also random variables with respect to the noise ν.

Our goal is to show that $(\gamma_{t,\beta} \cdot y_t + (1 - \gamma_{t,\beta}) \cdot y'_t)$ is independent of β–that is, although each β has a different $\gamma_{t,\beta}$, all these convex combinations are mapped to the same discretized value β_t^* to which $(\gamma' \cdot y_t + (1 - \gamma') \cdot y'_t)$ is mapped. To this end, define

$$\gamma_{t,\max} = \underset{\gamma_{t,\beta},\, \beta\in\Lambda[0,1]}{\arg\max} \left|(\gamma' \cdot y_t + (1 - \gamma') \cdot y'_t) - (\gamma_{t,\beta} \cdot y_t + (1 - \gamma_{t,\beta}) \cdot y'_t)\right|,$$

and let $\beta_t^* \in \Lambda[0,1]$ be such that

$$(\gamma' \cdot y_t + (1 - \gamma') \cdot y'_t) \in \lambda(\beta_t^*).$$

We seek to bound the probability that the convex combination with respect to $\gamma_{t,\max}$ is not mapped to β_t^*.

Claim E.3. $\Pr_\nu\left[(\gamma_{t,max} \cdot y_t + (1 - \gamma_{t,max}) \cdot y'_t) \notin \lambda(\beta_t^*)\right] = \mathsf{neg}(\kappa)$.

Proof. First, observe that for all β,

$$|(\gamma' \cdot y_t + (1 - \gamma') \cdot y'_t) - (\gamma_{t,\beta} \cdot y_t + (1 - \gamma_{t,\beta}) \cdot y'_t)| \leq |\gamma_{t,\max} - \gamma'| = \mathsf{neg}(\kappa),$$

where the last inequality follows from Claim C.19. Similarly,

$$|(\gamma_\nu \cdot y_t + (1 - \gamma_\nu) \cdot y'_t) - (\gamma' \cdot y_t + (1 - \gamma') \cdot y'_t)| \leq |\gamma' - \gamma_\nu| = \mathsf{neg}(\kappa).$$

Thus, combining both inequalities,

$$|(\gamma_\nu \cdot y_t + (1 - \gamma_\nu) \cdot y'_t) - (\gamma_{t,\beta} \cdot y_t + (1 - \gamma_{t,\beta}) \cdot y'_t)| = \mathsf{neg}(\kappa).$$

This implies that the "bad" event occurs only if $(\gamma_\nu \cdot y_t + (1-\gamma_\nu) \cdot y'_t)$ is negligibly close to one of the values in $\Lambda[0,1] = \left\{\frac{\lambda}{2}, \frac{3\lambda}{2}, \ldots, 1 - \frac{\lambda}{2}\right\}$. Applying Claim C.16, we get

$$|\gamma y + (1-\gamma)y' - \gamma_\nu y + (1-\gamma_\nu)y'| = |(\gamma - \gamma_\nu)y + (\gamma_\nu - \gamma)y'| = |\nu| \cdot ||y - y'|| \geq |\nu| \cdot \frac{\varepsilon}{M}.$$

Since $\nu \sim \text{Uniform}\left[-\frac{M}{\varepsilon\kappa}, \frac{M}{\varepsilon\kappa}\right]$, the probability of the event occurring is at most $\frac{\mathsf{neg}(\kappa)}{2/\kappa} = \mathsf{neg}(\kappa)$. The claim follows. □

Returning to Equation (6), we established that

$$\left|(\gamma_{t,\beta} \cdot y_t + (1-\gamma_{t,\beta}) \cdot y'_t) - \Pr\left[\tilde{t}(x) = t \mid x \in \mathcal{X}_{t,\beta}\right]\right| < \frac{\lambda}{2}.$$

With overwhelming probability, this implies that $\Pr\left[\tilde{t}(x) = t \mid x \in \mathcal{X}_{t,\beta}\right]$ remains within $\frac{\lambda}{2}$ of β_t^* for any β, ensuring that $\Pr\left[\tilde{t}(x) = t \mid x \in \mathcal{X}_{t,\beta}\right] \in \lambda(\beta_t^*)$. Consequently, $\mathcal{X}_{t,\beta}$ is independent of β, completing the proof of the lemma. □

Applying Lemma E.2, we conclude that for each $t \in \mathcal{T}$, there exists a unique set $\mathcal{X}_{t,\beta_t^*}$ such that $\left|\mathcal{X}_{t,\beta_t^*}\right| > \frac{2\alpha}{\lambda}|\mathcal{X}|$. Since the level sets for each t are disjoint, it follows that

$$\left|\mathcal{X}_{t,\beta_t^*}\right| \geq \left(1 - \frac{2\alpha}{\lambda^2}\right)|\mathcal{X}|.$$

Defining

$$\tilde{\mathcal{X}} = \bigcap_{t \in \mathcal{T}} \mathcal{X}_{t,\beta_t^*},$$

we obtain

$$\left|\tilde{\mathcal{X}}\right| \geq \left(1 - \frac{2k\alpha}{\lambda^2}\right)|\mathcal{X}|. \tag{8}$$

Now, let

$$\tilde{\gamma} := \frac{\left|\mathcal{X}_1 \cap \tilde{\mathcal{X}}\right|}{\left|\tilde{\mathcal{X}}\right|} \text{ and } \tilde{y} := \mathop{\mathbb{E}}_{x \sim \mathcal{D}_\mathcal{X}}\left[\tilde{R}(x) \mid x \in \tilde{\mathcal{X}}\right],$$

where the expectation is taken coordinate-wise (i.e., $\tilde{y} \in \mathcal{Y}$). Following the same decomposition as in Equation (7), we have

$$\Pr\left[t^*(x) = t \wedge x \in \tilde{\mathcal{X}}\right] = \tilde{\gamma} \cdot y_t + (1-\tilde{\gamma}) \cdot y'_t.$$

By Claim C.5, $\tilde{R}$ is α'-accurate w.r.t. $\mathcal{D}$ on $\tilde{\mathcal{X}}$, where $\alpha' = \alpha + \frac{2k\alpha}{\lambda^2} \leq \frac{3k\alpha}{\lambda^2}$. That is,

$$\|\tilde{y} - (\tilde{\gamma} \cdot y + (1-\tilde{\gamma}) \cdot y')\|_\infty \leq \frac{3k\alpha}{\lambda^2} \cdot \frac{1}{\Pr[x \in \tilde{\mathcal{X}}]} \leq \frac{3k\alpha}{\lambda^2 - 2k\alpha}.$$

By the definition of $\tilde{\mathcal{X}}$, for any $x \in \tilde{\mathcal{X}}$,

$$\left\|\tilde{y} - \tilde{R}(x)\right\|_\infty \leq \lambda.$$

Thus, for any $x \in \tilde{\mathcal{X}}$,

$$\left\|\tilde{R}(x) - (\tilde{\gamma} \cdot y + (1-\tilde{\gamma}) \cdot y')\right\|_\infty \leq \frac{3k\alpha}{\lambda^2 - 2k\alpha} + \lambda.$$

Since ρ is M-Lipschitz, it follows that for any $x \in \tilde{\mathcal{X}}$,

$$\left|\mathop{\mathbb{E}}_{\rho}\left[\rho\left(\tilde{R}(x)\right)\right] - \mathop{\mathbb{E}}_{\rho}\left[\rho\left(\tilde{\gamma} \cdot y + (1-\tilde{\gamma}) \cdot y'\right)\right]\right| \leq \frac{3Mk\alpha}{\lambda^2 - 2k\alpha} + \lambda M. \tag{9}$$

In particular, taking the expectation over all $x \in \tilde{\mathcal{X}}$ gives:

$$\left|\mathop{\mathbb{E}}_{x,\rho}\left[\rho\left(\tilde{R}(x)\right)\right] - \mathop{\mathbb{E}}_{\rho}\left[\rho\left(\tilde{\gamma} \cdot y + (1-\tilde{\gamma}) \cdot y'\right)\right]\right| \leq \frac{3Mk\alpha}{\lambda^2 - 2k\alpha} + \lambda M. \tag{10}$$

Recall the assumption that ρ is ε-far from affine:

$$|\mathbb{E}\left[\rho\left(\gamma \cdot y + (1-\gamma)\cdot y'\right)\right] - \left(\gamma \cdot \mathbb{E}\left[\rho(y)\right] + (1-\gamma)\cdot \mathbb{E}\left[\rho(y')\right]\right)| > \varepsilon. \tag{11}$$

The following claim shows that this also applies to $\tilde{\mathcal{X}}$.

Claim E.4. $|\mathbb{E}\left[\rho\left(\tilde{\gamma} \cdot y + (1-\tilde{\gamma})\cdot y'\right)\right] - \left(\tilde{\gamma} \cdot \mathbb{E}\left[\rho(y)\right] + (1-\tilde{\gamma})\cdot \mathbb{E}\left[\rho(y')\right]\right)| \geq \varepsilon - \frac{2M^2}{\varepsilon\kappa}$.

Proof. Since the ℓ_∞ norm of stochastic vectors is at most 1,

$$\left\|(\gamma \cdot y + (1-\gamma)\cdot y') - (\tilde{\gamma} \cdot y + (1-\tilde{\gamma})\cdot y')\right\|_\infty \leq |\gamma - \tilde{\gamma}| \cdot \|y - y'\|_\infty \leq |\gamma - \tilde{\gamma}|.$$

Since ρ is M-Lipschitz, it follows that

$$|\mathbb{E}\left[\rho\left(\gamma \cdot y + (1-\gamma)\cdot y'\right)\right] - \mathbb{E}\left[\rho\left(\tilde{\gamma} \cdot y + (1-\tilde{\gamma})\cdot y'\right)\right]| \leq M|\gamma - \tilde{\gamma}|.$$

Additionally, since $|\,\mathbb{E}\left[\rho(y) - \rho(y')\right]\,| \leq 1$,

$$|\left(\gamma \cdot \mathbb{E}\left[\rho(y)\right] + (1-\gamma)\cdot \mathbb{E}\left[\rho(y')\right]\right) - \left(\tilde{\gamma} \cdot \mathbb{E}\left[\rho(y)\right] + (1-\tilde{\gamma})\cdot \mathbb{E}\left[\rho(y')\right]\right)| \leq |\gamma - \tilde{\gamma}|.$$

Combining these bounds with the assumption in Equation (11) gives

$$|\mathbb{E}\left[\rho\left(\tilde{\gamma} \cdot y + (1-\tilde{\gamma})\cdot y'\right)\right] - \left(\tilde{\gamma} \cdot \mathbb{E}\left[\rho(y)\right] + (1-\tilde{\gamma})\cdot \mathbb{E}\left[\rho(y')\right]\right)| \geq \varepsilon - (M+1)|\gamma - \tilde{\gamma}|.$$

By Claim C.19, we have $|\gamma' - \gamma_\nu| = \mathsf{neg}(\kappa)$ and $|\tilde{\gamma} - \gamma'| = \mathsf{neg}(\kappa)$. Since $|\gamma - \gamma_\nu| = |\nu| \leq \frac{M}{\varepsilon\kappa}$, it follows that

$$|\gamma - \tilde{\gamma}| \leq |\gamma - \gamma_\nu| + |\gamma_\nu - \gamma'| + |\gamma' - \tilde{\gamma}| \leq \mathsf{neg}(\kappa) + \mathsf{neg}(\kappa) + \frac{M}{\varepsilon\kappa}, \tag{12}$$

and the claim follows. □

Returning to Equation (10), the claim implies that

$$\left|\mathop{\mathbb{E}}_{x,\rho}\left[\rho\left(\tilde{R}(x)\right)\right] - \left(\tilde{\gamma} \cdot \mathop{\mathbb{E}}_{\rho}\left[\rho(y)\right] + (1-\tilde{\gamma})\cdot \mathop{\mathbb{E}}_{\rho}\left[\rho(y')\right]\right)\right| \geq \varepsilon - \frac{2M^2}{\varepsilon\kappa} - \frac{3Mk\alpha}{\lambda^2 - 2k\alpha} - \lambda M > \frac{3\varepsilon}{4},$$

by taking $\kappa \geq \frac{30M^2}{\varepsilon^2}$, $\alpha \leq \frac{\lambda^2\varepsilon}{60Mk}$, and $\lambda \leq \frac{\varepsilon}{16M}$.

Thus, we have established a lower bound on the accuracy-on-decision (AD) error of $\tilde{R}$ with respect to $\mathcal{D}$ and ρ on $\tilde{\mathcal{X}}$. To translate this into a lower bound over all of $\mathcal{X}$, we recall from Equation (8) that

$$\eta := \Pr_{x\sim\mathcal{U}_\mathcal{X}}[x \notin \tilde{\mathcal{X}}] \leq \frac{2k\alpha}{\lambda^2}.$$

By our choice of parameters, $\eta \leq \varepsilon/16$, and the lower bound above is at least $3\varepsilon/4$. Since both $\rho(\tilde{R}(x))$ and $\rho(R^*(x))$ lie in $[0,1]$, the contribution of $\mathcal{X}\backslash\tilde{\mathcal{X}}$ to the AD gap can change the expectation by at most η. Therefore, the AD error over $\mathcal{X}$ is at least

$$(1-\eta)\cdot\frac{3\varepsilon}{4} - \eta > \frac{\varepsilon}{2},$$

contradicting the assumption that $\tilde{R}$ is $\varepsilon/2$-AD with respect to $\mathcal{D}$ and ρ. □

The following remarks highlight two key technical choices in the proof strategy. They clarify how the argument remains valid for any choice of discretization parameter, and why we adopt a Lipschitz assumption over weaker continuity conditions.

Remark E.5 (Avoiding the randomization of γ)**.** *Getting back to Lemma E.2, our goal was to find a large subset $\tilde{\mathcal{X}}$ where $\tilde{R}$ is close (in ℓ_∞ norm) to the expected value of R^* on $\mathcal{X}$. To achieve this, we prove that for any label t, at most one of the $1/\lambda$ level sets of t can be large, and we define $\tilde{\mathcal{X}}$ as the intersection of these large level sets. However, the t^{th} coordinate of the expected value of R^* over each level set $\mathcal{X}_{t,\beta}$ is not exactly $(\gamma' \cdot y_t + (1-\gamma')\cdot y'_t)$, but rather $(\gamma_{t,\beta} \cdot y_t + (1-\gamma_{t,\beta})\cdot y'_t)$, where $\gamma_{t,\beta}$ is negligibly close to γ'. Nevertheless, they may still fall into different λ-intervals.*

To address this, we introduce a small random perturbation to γ, which propagates to γ' and $\gamma_{t,\beta}$, ensuring that the issue arises with only negligible probability, as shown in Claim E.3.

An alternative approach is to define λ accordingly to prevent this issue. By assuming a uniform bound on the negligible term controlling this difference, we can set λ so that all values remain within the same λ-interval. This bound can be achieved, for instance, by restricting the learner that outputs $\tilde{R}$ to run within a fixed polynomial time. This approach would eliminate the need for the randomization introduced in Equation (5) *and significantly reduce the technical overhead in Lemma E.2.*

However, any method that merely proves the existence of a suitable λ, even without additional assumptions, is weaker than our approach, which guarantees that the proof holds for any λ.

Remark E.6 (Relaxing the Lipschitz condition)**.** *The theorem applies to Lipschitz rules that are far from affine. One may ask if requiring continuity is enough. Indeed, since $\mathcal{Y}$ is a compact set, continuity on $\mathcal{Y}$ implies uniform continuity. Let us revisit, for example, Equation* (9)*, and denote the two vectors there by z and z'. Using uniform continuity, for any desired bound μ, we know that there exists $\delta > 0$ such that if $||y - y'|| < \delta$, then $\mathbb{E}[\rho(z)] - \mathbb{E}[\rho(z')] < \mu$. However, there is no guarantee of how small δ should be. In particular, it can be inverse exponential in ε, and lead to trivial bounds on α and λ, such that the result only holds in the regime where, e.g., they are both smaller than $2^{-\varepsilon}$. To obtain meaningful bounds, we introduce the Lipschitzness requirement.*

E.1 Loss-minimizing decision rule is far from affine

Returning to Definition D.7, we now prove that the loss-minimizing decision rule ρ^*_ℓ is far from affine for any non-trivial loss function. In Remark G.5, we will show how this statement can also be derived from the arguments presented earlier.

The class of loss functions we consider as non-trivial includes all loss functions that do not trivialize the tasks of loss minimization or being accurate-on-classification by making them either too easy or impossible.[9]

Definition E.7 (Family of Non-Trivial Loss Functions $\mathcal{L}_\alpha$)**.** *The family of non-trivial loss functions $\mathcal{L}_\alpha \subseteq \{\mathcal{T} \times \{0,1\} \to [0,1]\}$ is defined such that any $\ell \in \mathcal{L}_\alpha$ satisfies the following two conditions:*

1. *there exists $t \in \mathcal{T}$ such that $|\ell(t,1) - \ell(t,0)| \geq \alpha$.*
2. *there exist $t', t'' \in \mathcal{T}$ such that $\ell(t',1) - \ell(t',0) \geq 0$ and $\ell(t'',1) - \ell(t'',0) < 0$.*

Lemma E.8 (Loss-Minimizing Decision Rule is Far from Affine)**.** *The decision rule ρ^*_ℓ is $1/2$-far from affine for any non-trivial $\ell \in \mathcal{L}_\alpha$.*

Proof. Since ℓ is non-trivial (Definition E.7), there exists $t, t' \in \mathcal{T}$ such that $\ell(t,1) - \ell(t,0) \geq 0$ and $\ell(t',1) - \ell(t',0) < 0$ (these are t', t'' from the second condition). Consider the unit vectors e_t and $e_{t'}$, where each vector has a 1 at coordinate t (or t', respectively) and 0 elsewhere. By definition, the loss-minimizing decision rule satisfies

$$\rho^*_\ell(e_t) = 0 \text{ and } \rho^*_\ell(e_{t'}) = 1.$$

Now, consider their convex combination:

$$v = \frac{1}{2}e_t + \frac{1}{2}e_{t'}.$$

On one hand, the convex combination of their decision rule values is

$$\frac{1}{2}\rho^*_\ell(e_t) + \frac{1}{2}\rho^*_\ell(e_{t'}) = \frac{1}{2}.$$

On the other hand, $\rho^*_\ell(v) \in \{0,1\}$ by its definition — it is a deterministic decision rule that outputs a bit for each fixed vector. This means that

$$\left|\rho^*_\ell(v) - \left(\frac{1}{2}\rho^*_\ell(e_t) + \frac{1}{2}\rho^*_\ell(e_{t'})\right)\right| = \frac{1}{2},$$

so ρ^*_ℓ is at least $1/2$-far from affine. □

[9] To see why the second condition in Definition E.7 ensures that the task of being multi-accurate-on-classification is not trivialized, recall that the loss-minimizing decision rule (Definition D.7) breaks ties in favor of rejection. Hence, if ℓ violates the second condition, then the loss-minimizing action function is constant: either $h^*_\ell \equiv 0$ or $h^*_\ell \equiv 1$.

Although Lemma E.8 establishes that the loss-minimizing decision rule ρ_ℓ^* is far from affine, it is not Lipschitz, and therefore Theorem E.1 does not apply to it directly. The following claim extends the impossibility to ρ_ℓ^* by applying Theorem E.1 to a Lipschitz smoothing of this rule and showing that the resulting lower bound transfers to the original, non-Lipschitz rule.

Claim E.9. *For any non-trivial loss $\ell \in \mathcal{L}_\alpha$, the conclusion of Theorem E.1 extends to ρ_ℓ^* and gives accuracy-on-decision error of $1/8$.*

Proof sketch. First suppose that the expected loss difference is strictly positive at some point and strictly negative at another. By linearity, we may choose a smoothing parameter η satisfying $\eta \leq \alpha/8$ where α is the non-triviality parameter, and $y_0, y_s \in \mathcal{Y}$ such that

$$\mathop{\mathbb{E}}_{t\sim y_0}[\ell(t,1) - \ell(t,0)] = -2\eta, \qquad \mathop{\mathbb{E}}_{t\sim y_s}[\ell(t,1) - \ell(t,0)] = 16\eta,$$

or with both signs reversed. Let $\bar{\rho}_{\ell,\eta}$ be the symmetric smoothing of ρ_ℓ^*: its acceptance probability is 1 below $-\eta$, 0 above η, and linear in between, that is,

$$\Pr[\bar{\rho}_{\ell,\eta}(y) = 1] = \begin{cases} 1 & \text{if } \mathbb{E}_{t\sim y}[\ell(t,1) - \ell(t,0)] \leq -\eta, \\ \frac{\eta - \mathbb{E}_{t\sim y}[\ell(t,1)-\ell(t,0)]}{2\eta} & \text{if } |\mathbb{E}_{t\sim y}[\ell(t,1) - \ell(t,0)]| < \eta, \\ 0 & \text{if } \mathbb{E}_{t\sim y}[\ell(t,1) - \ell(t,0)] \geq \eta. \end{cases}$$

This decision rule is $k/(2\eta)$-Lipschitz (see Definition C.12). For the linear combination

$$z := \frac{3}{4}y_0 + \frac{1}{4}y_s,$$

the loss difference equals $5\eta/2$ (or $-5\eta/2$), and the acceptance probabilities of $\bar{\rho}_{\ell,\eta}$ at (y_0, y_s, z) are respectively $(1, 0, 0)$ or $(0, 1, 1)$. Thus, $\bar{\rho}_{\ell,\eta}$ is $1/2$-far from affine (or any other constant strictly smaller than $3/4$). Note that comparing to Theorem E.1, y_0 and y_s are analogous to y and y' respectively, and γ is analogous to $3/4$.

Apply Theorem E.1 with $\varepsilon = 1/2$. The constructed R^* takes only the values y_0, y_s on which $\bar{\rho}_{\ell,\eta}$ and ρ_ℓ^* agree. Moreover, the bound preceding Equation (9), together with $\tilde{\gamma}$ being close to $3/4$ by Equation (12), implies that, for every discretization parameter and calibration error satisfying the explicit hypotheses of Theorem E.1, the expected loss difference at $\tilde{R}(x)$ remains outside the smoothing interval $[-\eta, \eta]$ for every $x \in \tilde{\mathcal{X}}$. Hence, the two rules also agree on $\tilde{R}(x)$ throughout $\tilde{\mathcal{X}}$. Taking $\Pr[x \notin \tilde{\mathcal{X}}] \leq 1/32$, the $1/4$ lower bound for $\bar{\rho}_{\ell,\eta}$ yields an accuracy-on-decision error greater than $1/4 - 1/32 > 1/8$ for ρ_ℓ^*.

Finally, if the only available non-negative point y_0 is a tie, the symmetric smoothing disagrees with ρ_ℓ^* at y_0. In this case, use the one-sided smoothing whose acceptance probability is 1 below $-\eta$, 0 from 0 onward, and linear on $[-\eta, 0]$. Choose y_s with loss difference at most -8η. Then the loss difference at $\frac{3}{4}y_0 + \frac{1}{4}y_s$ is at most -2η, the acceptance probabilities are $(0, 1, 1)$, and the same argument applies. □

F When loss minimization contradicts accuracy of classifications

As we have argued, achieving accuracy-on-decision may sometimes be impossible, implying that accuracy-on-classification might not be attainable when relying on predictors. However, this does not preclude the possibility of classifying individuals directly—classification methods need not operate by applying a decision rule to a risk predictor. In this section, we show that loss minimization may not align with accuracy-on-classification, even for general action functions. Specifically, for any non-trivial loss function, we construct a setting in which any (possibly randomized) action function that satisfies accuracy-on-classification necessarily fails to minimize the expected loss. This holds even when the hypothesis class is restricted to just the two constant functions that either accept or reject each individual with probability 1.

The following theorem establishes the central argument of this section: achieving both optimal loss minimization and accurate representation in classification—even for a single subpopulation—may be computationally infeasible. However, these objectives can be attained independently. In Corollary D.10, we show that a multi-accurate predictor enables MAC, while in Theorem G.1, we

demonstrate that loss minimization is possible given a multi-calibrated predictor. This confirms that while loss minimization and MAC can each be achieved separately—and even with the same predictor—they may not necessarily be achieved by applying the same decision rule to that predictor.

Theorem F.1 (Loss Minimization Contradicts Accuracy-on-Classification)**.** *Assume one-way functions exist w.r.t. a security parameter $\kappa \geq \log|\mathcal{X}|$. Let $\varepsilon, \alpha \in (0,1)$, and let $\mathcal{H} = \{h_0, h_1\}$ where $h_0 \equiv 0$ and $h_1 \equiv 1$. For any $\ell \in \mathcal{L}_\alpha$, there exists a Nature $\mathcal{D}$ such that for any $\mathsf{poly}(\kappa)$-time computable $h : \mathcal{X} \to \{0,1\}$, if h is (ℓ, ε)-accurate-on-classification then h incurs loss-minimization error at least $\Omega(\alpha) - O(\varepsilon)$ with respect to ℓ and $\mathcal{H}$.*

As discussed in Appendix C.4 and before Theorem E.1, the assumption of one-way functions is a modeling choice and can be removed if no restriction is placed on the complexity of Nature.

Proof. Recall from Definition E.7 that $\mathcal{L}_\alpha$ denotes the family of non-trivial loss functions. Let $\ell \in \mathcal{L}_\alpha$, and let $\mathcal{U}_\mathcal{X}$ be the uniform distribution over $\mathcal{X}$. Define two constants $\varepsilon_{\mathsf{AC}}, \varepsilon_\ell \in (0,1)$ to be specified later. Since ℓ is non-trivial, there exist types $t, t', t'' \in \mathcal{T}$ such that

$$|\ell(t,1) - \ell(t,0)| \geq \alpha, \quad \ell(t',1) - \ell(t',0) \geq 0, \quad \ell(t'',1) - \ell(t'',0) < 0.$$

Depending on the sign of $\ell(t,1) - \ell(t,0)$, we define $\beta = \ell(t',1) - \ell(t',0)$ (when the sign is negative) or $\beta = \ell(t'',0) - \ell(t'',1)$ (when the sign is positive), so that $\beta \geq 0$ or $\beta > 0$, respectively. The comparison between α and β induces four analogous cases. We focus on the case where

$$\ell(t,1) - \ell(t,0) \geq \alpha, \quad \ell(t',0) - \ell(t',1) = \beta > 0,^{10} \quad \alpha \geq \beta.$$

In this case, rejecting an individual of type t results in a significantly lower loss than accepting them, while for type t', acceptance yields lower loss than rejection, though the gap may be smaller. The remaining cases follow analogously by interchanging the roles of t and t' or t'', and yield the same lower bound.

We define Nature according to the pair of labels t and t'. Under the assumption that one-way functions exist, Lemma C.18 implies the existence of a subset $\mathcal{X}_1 \subseteq \mathcal{X}$ of size

$$|\mathcal{X}_1| = \left(\frac{3}{4} + \mathsf{neg}(\kappa)\right)|\mathcal{X}|,$$

such that $\mathcal{X}_1$ is κ-indistinguishable w.r.t. $\mathcal{U}_\mathcal{X}$ (see Definition C.17) and Nature is efficiently computable.

Defining $\mathcal{X}_2 := \mathcal{X} \setminus \mathcal{X}_1$, we write

$$\mathcal{X} = \mathcal{X}_1 \uplus \mathcal{X}_2.$$

We specify Nature's deterministic prediction as follows:

$$R^*|_{\mathcal{X}_1} = t, \quad R^*|_{\mathcal{X}_2} = t',$$

and take $\mathcal{D} = \mathcal{U}_\mathcal{X} \times R^*(\mathcal{U}_\mathcal{X})$.[11]

Let h be a $\mathsf{poly}(\kappa)$-time computable randomized action function that is $(\ell, \varepsilon_{\mathsf{AC}})$-accurate-on-classification. By our definition of Nature,

$$h^*_\ell(x) = \begin{cases} 0, & x \in \mathcal{X}_1, \\ 1, & x \in \mathcal{X}_2. \end{cases}$$

Therefore,

$$\mathbb{E}\left[h^*_\ell(x)\right] = \Pr[x \in \mathcal{X}_2] = \frac{1}{4} + \mathsf{neg}(\kappa).$$

By $(\ell, \varepsilon_{\mathsf{AC}})$-accuracy-on-classification (see Definition D.9) and since $\mathbb{E}\left[h(x)\right] = \Pr[h(x) = 1]$, it follows that

$$\Pr_{x,h}[h(x) = 1] \geq \frac{1}{4} - \varepsilon_{\mathsf{AC}} + \mathsf{neg}(\kappa). \tag{13}$$

[10] Here we rename t'' as t' for notational simplicity.

[11] So far, we defined a deterministic Nature such that it outputs unit vectors $e_t, e_{t'} \in \mathcal{Y}$, rather than labels $t, t' \in \mathcal{T}$. For notational convenience, we now define it to output a label in this case, and denote it by R^*.

In particular, this implies that $\Pr[h(x) = 1]$ is non-negligible. Since h is $\mathsf{poly}(\kappa)$-time computable and $\mathcal{X}_1$ is κ-indistinguishable, Claim C.19 applies, yielding

$$\Pr\left[x \in \mathcal{X}_1 | h(x) = 1\right] = \frac{3}{4} + \mathsf{neg}(\kappa), \tag{14}$$

and, analogously,

$$\Pr\left[x \in \mathcal{X}_2 | h(x) = 1\right] = \frac{1}{4} + \mathsf{neg}(\kappa).$$

Finally, we proceed to compute the expected loss of the loss minimizer h_0 and of h. Notice that $h_0 \in \mathcal{H}$, and its expected loss is

$$\mathop{\mathbb{E}}_{x}\left[\ell\left(R^*(x), h_0(x)\right)\right] = \mathbb{E}\left[\ell\left(R^*(x), 0\right)\right] = \frac{3}{4} \cdot \ell(t, 0) + \frac{1}{4} \cdot \ell(t', 0) := L_0.$$

Similarly, the expected loss of h_1 is

$$\mathop{\mathbb{E}}_{x}\left[\ell\left(R^*(x), h_1(x)\right)\right] = \mathbb{E}\left[\ell\left(R^*(x), 1\right)\right] = \frac{3}{4} \cdot \ell(t, 1) + \frac{1}{4} \cdot \ell(t', 1) := L_1.$$

This implies that

$$L_1 - L_0 = \frac{3}{4}\left(\ell(t, 1) - \ell(t, 0)\right) - \frac{1}{4}\left(\ell(t', 0) - \ell(t', 1)\right) \geq \frac{3}{4}\alpha - \frac{1}{4}\beta \geq \frac{\alpha}{2}, \tag{15}$$

where the last inequality follows from the assumption that $\alpha \geq \beta$.

Next, we compute the expected loss of h over $\mathcal{U}_{\mathcal{X}}$:

$$\begin{aligned}
\mathop{\mathbb{E}}_{x,h}\left[\ell\left(R^*(x), h(x)\right)\right] &= \mathbb{E}\left[\ell\left(R^*(x), 0\right) | h(x) = 0\right] \cdot \Pr[h(x) = 0] \\
&\quad + \mathbb{E}\left[\ell\left(R^*(x), 1\right) | h(x) = 1\right] \cdot \Pr[h(x) = 1] \\
&\geq \mathbb{E}\left[\ell\left(R^*(x), 0\right) | h(x) = 0\right] \cdot \left(\frac{3}{4} - \varepsilon_{\mathsf{AC}}\right) \\
&\quad + \mathbb{E}\left[\ell\left(R^*(x), 1\right) | h(x) = 1\right] \cdot \left(\frac{1}{4} - \varepsilon_{\mathsf{AC}}\right) + \mathsf{neg}(\kappa),
\end{aligned} \tag{16}$$

where the last inequality follows from Equation (13).

We argue that this quantity is close to $\frac{3}{4}L_0 + \frac{1}{4}L_1$. First, observe that

$$\begin{aligned}
\mathbb{E}\left[\ell\left(R^*(x), 1\right) | h(x) = 1\right] &= \ell\left(t, 1\right) \cdot \Pr\left[x \in \mathcal{X}_1 \mid h(x) = 1\right] \\
&\quad + \ell\left(t', 1\right) \cdot \Pr\left[x \in \mathcal{X}_2 \mid h(x) = 1\right] \\
&= \frac{3}{4} \cdot \ell\left(t, 1\right) + \frac{1}{4} \cdot \ell\left(t', 1\right) + \mathsf{neg}(\kappa) = L_1 + \mathsf{neg}(\kappa),
\end{aligned}$$

where the second-to-last equality follows from Equation (14). Similarly, we obtain

$$\mathbb{E}\left[\ell\left(R^*(x), 0\right) | h(x) = 0\right] = \frac{3}{4} \cdot \ell\left(t, 0\right) + \frac{1}{4} \cdot \ell\left(t', 0\right) + \mathsf{neg}(\kappa) = L_0 + \mathsf{neg}(\kappa).$$

Substituting into Equation (16), we conclude that

$$\mathbb{E}\left[\ell\left(R^*(x), h(x)\right)\right] \geq \left(\frac{3}{4} - \varepsilon_{\mathsf{AC}}\right) \cdot L_0 + \left(\frac{1}{4} - \varepsilon_{\mathsf{AC}}\right) \cdot L_1 + \mathsf{neg}(\kappa).$$

By $(\mathcal{H}, \ell, \varepsilon_\ell)$-loss-minimization of h (see Definition D.6), we have

$$\begin{aligned}
\varepsilon_\ell &\geq \mathop{\mathbb{E}}_{x,h}\left[\ell\left(R^*(x), h(x)\right)\right] - \mathop{\mathbb{E}}_{x}\left[\ell\left(R^*(x), h_0(x)\right)\right] \\
&\geq \left(\frac{3}{4} - \varepsilon_{\mathsf{AC}}\right) \cdot L_0 + \left(\frac{1}{4} - \varepsilon_{\mathsf{AC}}\right) \cdot L_1 + \mathsf{neg}(\kappa) - L_0 \\
&= \frac{1}{4}\left(L_1 - L_0\right) - \varepsilon_{\mathsf{AC}}\left(L_1 + L_0\right) + \mathsf{neg}(\kappa) \\
&\geq \frac{\alpha}{8} - \varepsilon_{\mathsf{AC}}\left(L_1 + L_0\right) + \mathsf{neg}(\kappa),
\end{aligned}$$

where the last inequality follows from Equation (15). Since $L_1 + L_0 \leq 2 - \frac{3}{4}\alpha$, we conclude that

$$\varepsilon_\ell \geq \frac{\alpha}{8} - \varepsilon_{\mathsf{AC}}\left(2 - \frac{3}{4}\alpha\right) + \mathsf{neg}(\kappa).$$

The theorem follows by setting $\varepsilon = \varepsilon_{\mathsf{AC}}$, but the analysis allows for a more general tradeoff between $\varepsilon_{\mathsf{AC}}$ and ε_ℓ. For instance, requiring the accuracy-on-classification error to be smaller than $\frac{\alpha}{80}$ implies that $(\mathcal{H}, \ell, \frac{\alpha}{10})$-loss-minimization is impossible. □

G Translating accuracy of predictions to loss minimization

Recall the definitions of the loss-minimizing decision rule (Definition D.7) and full-multi-calibration (Definition C.8). The following theorem shows that composing a multi-calibrated predictor with the loss-minimizing decision rule results in a loss-minimizing action function. The hypothesis class with respect to which loss minimization is measured corresponds to the class of subpopulations for which the predictor is multi-calibrated. Notably, this implies that the loss function need not be known during learning; one can first learn a multi-calibrated predictor and later post-process it using the loss-minimizing decision rule to obtain an action function that minimizes any given loss.

This guarantee aligns with the notion of *omnipredictors* introduced by Gopalan et al. [2022], which are predictors that admit post-processing to minimize any loss function within a given class. Specifically, Theorem 8.1 in Gopalan et al. [2022] shows that a multi-calibrated predictor is an omnipredictor. Our proof follows their high-level approach and proof technique, with a few adaptations to fit our setting. First, the class of loss functions we consider, as well as the assumptions imposed on them, differ from those in Gopalan et al. [2022]. Second, in Lemma G.4 we extend the argument by showing that composing the original predictor with a loss-minimizing decision rule yields an action function that minimizes loss, whereas the original paper establishes this property for a different primitive. Moreover, in Appendix G.1, we show how to relax the multi-calibration requirement. We note that the notion of multi-calibration used in the following proof is identical to that of Gopalan et al. [2022], though their presentation is phrased in terms of covariance.

Theorem G.1 (Multi-Calibration for Loss Minimization)**.** *Let $\mathcal{C} \subseteq \{0,1\}^{|\mathcal{X}|}$ be a collection of sets, and let $\mathcal{H} = \mathcal{H}_\mathcal{C} \subseteq \{\mathcal{X} \to \{0,1\}\}$ be the hypothesis class consisting of all functions that compute membership in sets from $\mathcal{C}$. Let $\lambda \in (0,1)$ be a discretization parameter, and let $\ell : \mathcal{T} \times \{0,1\} \to [0,1]$ be a loss function. If $\tilde{R}$ is $(\mathcal{C}, \alpha, \lambda)$-full-multi-calibrated w.r.t. $\mathcal{D}$, then $\rho^*_\ell \circ \tilde{R}$ is $(\mathcal{H}, \ell, 3k(\alpha + \lambda))$-loss-minimizing.*

Proof. Throughout the proof, it will be convenient to use the equivalence between two sampling processes: the standard process where we first sample an element from $\mathcal{X}$ and then a label, and the alternative of first sampling a label and then an element. Given a distribution $\mathcal{D}_\mathcal{X}$ over $\mathcal{X}$, each element $x \in \mathcal{X}$ is associated with a stochastic vector $R^*(x) \in \mathcal{Y}$, which defines a distribution over the set of types $\mathcal{T}$. The *marginal label distribution* $\mathcal{D}_\mathcal{T}$ is given by

$$\mathcal{D}_\mathcal{T}(t) = \sum_{x \in \mathcal{X}} \mathcal{D}_\mathcal{X}(x) R^*(x)_t,$$

where $R^*(x)_t$ is the t^{th} coordinate of $R^*(x)$. The *distribution over elements given a label* $\mathcal{D}_{\mathcal{X}|\mathcal{T}}$ is

$$\mathcal{D}_{\mathcal{X}|\mathcal{T}}(x \mid t) = \frac{\mathcal{D}_\mathcal{X}(x) R^*(x)_t}{\mathcal{D}_\mathcal{T}(t)}.$$

Thus, sampling (x,t) by first drawing $x \sim \mathcal{D}_\mathcal{X}$ and then $t \sim R^*(x)$ is equivalent to first drawing $t \sim \mathcal{D}_\mathcal{T}$ and then $x \sim \mathcal{D}_{\mathcal{X}|\mathcal{T}}(\cdot \mid t)$. For clarity, we now fix the notation used throughout the proof.

- In what follows, we define a partition $T = (T_v)_{v \in (\Lambda[0,1])^k}$ of $\mathcal{X}$. For sampling a subset T_v from the partition T according to its mass under $\mathcal{D}_\mathcal{X}$, that is, $\sum_{x \in T_v} \mathcal{D}_\mathcal{X}(x)$, we use $T_v \sim T$.
- We use $x \sim T_v$ to denote sampling $x \sim \mathcal{D}_\mathcal{X}$ conditioned on the event that $x \in T_v$.
- We use $t \sim \mathcal{D}_\mathcal{T}|_{T_v}$ to denote sampling from the marginal label distribution restricted to the set T_v.

- We use $x \sim \mathcal{D}_{\mathcal{X}|\mathcal{T}=t,T_v}$ for sampling an element from a set T_v according to the distribution over elements given a label t.
- For any $h \in \mathcal{H}$, we let $\hat{h} : T \to \{0,1\}$ denote a random variable following a $\text{Ber}\left(\mathbb{E}_{h,x\sim T_v}[h(x)]\right)$ distribution for any $T_v \in T$.
- We restate full multi-calibration in the covariance formulation, following Gopalan et al. [2022]. If $\tilde{R}$ is $(\mathcal{C}, \alpha, \lambda)$-full-multi-calibrated with respect to $\mathcal{D}$, as defined in Definition C.8, then there exists a partition $T = (T_v)_{v\in(\Lambda[0,1])^k}$ of $\mathcal{X}$ such that for any $h \in \mathcal{H}$, $t \in \mathcal{T}$,
$$\mathop{\mathbb{E}}_{T_v\sim T}\left[\Pr_{t'\sim\mathcal{D}_{\mathcal{T}|T_v}}[t'=t]\cdot\left|\mathop{\mathbb{E}}_{h,x\sim T_v}[h(x)] - \mathop{\mathbb{E}}_{h,x\sim\mathcal{D}_{\mathcal{X}|\mathcal{T}=t,T_v}}[h(x)]\right|\right] \le \alpha+\lambda. \tag{17}$$
- We define the predictor R^T as follows:
$$\forall x \in T_v, \quad R^{T_v}(x) := \mathop{\mathbb{E}}_{x\sim T_v}[R^*(x)],$$
where the expectation is taken coordinate-wise, so that R^T outputs a vector. That is, R^T is constant on each T_v and equals the coordinate-wise expectation of R^* over that part.
- We define its corresponding loss-minimizing action function as $h_\ell^T := \rho_\ell^* \circ R^T$.
- Following the terminology set in Gopalan et al. [2022], we refer to R^T as the *canonical predictor* and to h_ℓ^T as the *canonical action function*.
- We denote $\tilde{h} := \rho_\ell^* \circ \tilde{R}$.

The proof consists of two steps. First, we show that the canonical action function minimizes the loss. We begin by proving that its loss is at most that of any function that remains constant on each partition set (such as the canonical predictor). We then extend this result to all functions in the hypothesis class. Second, we demonstrate that its expected loss is close to that of $\tilde{h}$.

The following claim relates the expected losses of h and $\hat{h}$ over T_v and a fixed label to their expected values.

Claim G.2. *Let $t \in \mathcal{T}$ and $T_v \in T$.*
$$\mathop{\mathbb{E}}_{\hat{h},x\sim\mathcal{D}_{\mathcal{X}|\mathcal{T}=t,T_v}}\left[\ell\left(t,\hat{h}(x)\right)\right] - \mathop{\mathbb{E}}_{h,x\sim\mathcal{D}_{\mathcal{X}|\mathcal{T}=t,T_v}}[\ell(t,h(x))] \le \left|\mathop{\mathbb{E}}_{h,x\sim T_v}[h(x)] - \mathop{\mathbb{E}}_{h,x\sim\mathcal{D}_{\mathcal{X}|\mathcal{T}=t,T_v}}[h(x)]\right|.$$

Proof. The first term on the left-hand side can be expressed as
$$\begin{aligned}\mathop{\mathbb{E}}_{\hat{h},x\sim\mathcal{D}_{\mathcal{X}|\mathcal{T}=t,T_v}}\left[\ell\left(t,\hat{h}(x)\right)\right] &= \Pr_{\hat{h},x\sim\mathcal{D}_{\mathcal{X}|\mathcal{T}=t,T_v}}\left[\hat{h}(x)=1\right]\cdot\ell(t,1) + \Pr_{\hat{h},x\sim\mathcal{D}_{\mathcal{X}|\mathcal{T}=t,T_v}}\left[\hat{h}(x)=0\right]\cdot\ell(t,0)\\ &= \mathop{\mathbb{E}}_{h,x\sim T_v}[h(x)]\cdot\ell(t,1) + \left(1-\mathop{\mathbb{E}}_{h,x\sim T_v}[h(x)]\right)\cdot\ell(t,0),\end{aligned}$$
since $\hat{h}$ is fixed on T_v. The second term on the left-hand side can be expressed as
$$\mathop{\mathbb{E}}_{h,x\sim\mathcal{D}_{\mathcal{X}|\mathcal{T}=t,T_v}}[\ell(t,h(x))] = \mathop{\mathbb{E}}_{h,x\sim\mathcal{D}_{\mathcal{X}|\mathcal{T}=t,T_v}}[h(x)]\cdot\ell(t,1) + \left(1-\mathop{\mathbb{E}}_{h,x\sim\mathcal{D}_{\mathcal{X}|\mathcal{T}=t,T_v}}[h(x)]\right)\cdot\ell(t,0).$$
Hence, their difference is bounded by
$$\begin{aligned}&\mathop{\mathbb{E}}_{h,x\sim T_v}[h(x)]\cdot\ell(t,1) + \left(1-\mathop{\mathbb{E}}_{h,x\sim T_v}[h(x)]\right)\cdot\ell(t,0)\\ &\qquad - \mathop{\mathbb{E}}_{h,x\sim\mathcal{D}_{\mathcal{X}|\mathcal{T}=t,T_v}}[h(x)]\cdot\ell(t,1) - \left(1-\mathop{\mathbb{E}}_{h,x\sim\mathcal{D}_{\mathcal{X}|\mathcal{T}=t,T_v}}[h(x)]\right)\cdot\ell(t,0)\\ &\qquad = (\ell(t,1)-\ell(t,0))\cdot\left(\mathop{\mathbb{E}}_{h,x\sim T_v}[h(x)] - \mathop{\mathbb{E}}_{h,x\sim\mathcal{D}_{\mathcal{X}|\mathcal{T}=t,T_v}}[h(x)]\right)\\ &\qquad \le \left|\mathop{\mathbb{E}}_{h,x\sim T_v}[h(x)] - \mathop{\mathbb{E}}_{h,x\sim\mathcal{D}_{\mathcal{X}|\mathcal{T}=t,T_v}}[h(x)]\right|.\end{aligned}$$
□

We use the claim to bound the difference between the expected losses of $\hat{h}$ and h. For any $h \in \mathcal{H}$, we write this difference as

$$\begin{aligned}
&\mathsf{exp\text{-}loss}_{\mathcal{D}}(\hat{h}) - \mathsf{exp\text{-}loss}_{\mathcal{D}}(h) \leq \\
&\mathop{\mathbb{E}}_{T_v \sim T} \mathop{\mathbb{E}}_{t \sim \mathcal{D}_{\mathcal{T}}|_{T_v}} \left| \mathop{\mathbb{E}}_{\hat{h}, x \sim \mathcal{D}_{\mathcal{X}|\mathcal{T}=t,T_v}} \left[\ell\left(t, \hat{h}(x)\right) \right] - \mathop{\mathbb{E}}_{h, x \sim \mathcal{D}_{\mathcal{X}|\mathcal{T}=t,T_v}} \left[\ell\left(t, h(x)\right) \right] \right| \leq \\
&\mathop{\mathbb{E}}_{T_v \sim T} \mathop{\mathbb{E}}_{t \sim \mathcal{D}_{\mathcal{T}}|_{T_v}} \left| \mathop{\mathbb{E}}_{h, x \sim T_v} [h(x)] - \mathop{\mathbb{E}}_{h, x \sim \mathcal{D}_{\mathcal{X}|\mathcal{T}=t,T_v}} [h(x)] \right| = \\
&\mathop{\mathbb{E}}_{T_v \sim T} \sum_{t \in \mathcal{T}} \Pr_{t' \sim \mathcal{D}_{\mathcal{T}}|_{T_v}} [t' = t] \left| \mathop{\mathbb{E}}_{h, x \sim T_v} [h(x)] - \mathop{\mathbb{E}}_{h, x \sim \mathcal{D}_{\mathcal{X}|\mathcal{T}=t,T_v}} [h(x)] \right| = \\
&\sum_{t \in \mathcal{T}} \mathop{\mathbb{E}}_{T_v \sim T} \left[\Pr_{t' \sim \mathcal{D}_{\mathcal{T}}|_{T_v}} [t' = t] \left| \mathop{\mathbb{E}}_{h, x \sim T_v} [h(x)] - \mathop{\mathbb{E}}_{h, x \sim \mathcal{D}_{\mathcal{X}|\mathcal{T}=t,T_v}} [h(x)] \right| \right] \leq k(\alpha + \lambda), \qquad (18)
\end{aligned}$$

where the last inequality follows from the $(\mathcal{C}, \alpha, \lambda)$-full-multi-calibration of $\tilde{R}$, as stated in Equation (17).

Next, recall the definitions of the canonical predictor R^T and the canonical action function h_ℓ^T. The following claim achieves our first goal, that is, proving that the canonical action function minimizes the loss.

Claim G.3 (Corollary 6.2 in Gopalan et al. [2022])**.** $\forall h : T \to \{0,1\}$, $\mathsf{exp\text{-}loss}_{\mathcal{D}}\left(h_\ell^T\right) \leq \mathsf{exp\text{-}loss}_{\mathcal{D}}\left(h\right)$.

Proof. First, notice that

$$\mathsf{exp\text{-}loss}_{\mathcal{D}}\left(h_\ell^T\right) = \mathop{\mathbb{E}}_{T_v \sim T} \mathop{\mathbb{E}}_{x \sim T_v} \mathop{\mathbb{E}}_{t \sim R^{T_v}(x)} \left[\ell(t, h_\ell^T(x)) \right].$$

A useful fact throughout is that sampling $t \sim R^{T_v}$ gives the same marginal distribution over labels as first sampling $x \sim D_X$ conditioned on $x \in T_v$ and then sampling $t \sim R^*(x)$. This implies that

$$\rho_\ell^*\left(R^{T_v}(x)\right) = \mathop{\arg\min}_{a \in \{0,1\}} \mathop{\mathbb{E}}_{t \sim R^{T_v}(x)} [\ell(t,a)] = \mathop{\arg\min}_{a \in \{0,1\}} \mathop{\mathbb{E}}_{t \sim R^*|_{T_v}(x)} [\ell(t,a)],$$

hence,

$$\mathop{\mathbb{E}}_{t \sim R^{T_v}(x)} \left[\ell(t, h_\ell^T(x)) \right] \leq \min \left\{ \mathop{\mathbb{E}}_{t \sim R^*|_{T_v}(x)} [\ell(t,0)], \mathop{\mathbb{E}}_{t \sim R^*|_{T_v}(x)} [\ell(t,1)] \right\}.$$

Next, let $h : T \to \{0,1\}$. Once again,

$$\mathsf{exp\text{-}loss}_{\mathcal{D}}\left(h\right) = \mathop{\mathbb{E}}_{T_v \sim T} \mathop{\mathbb{E}}_{x \sim T_v} \mathop{\mathbb{E}}_{h, t \sim R^*|_{T_v}(x)} \left[\ell(t, h(x)) \right].$$

Focusing on the last term,

$$\begin{aligned}
\mathop{\mathbb{E}}_{h, t \sim R^*|_{T_v}(x)} [\ell(t, h(x))] &= \mathop{\mathbb{E}}_{h}[h] \cdot \mathop{\mathbb{E}}_{t \sim R^*|_{T_v}(x)} [\ell(t,1)] + \left(1 - \mathop{\mathbb{E}}_{h}[h]\right) \cdot \mathop{\mathbb{E}}_{t \sim R^*|_{T_v}(x)} [\ell(t,0)] \\
&\geq \min \left\{ \mathop{\mathbb{E}}_{t \sim R^*|_{T_v}(x)} [\ell(t,0)], \mathop{\mathbb{E}}_{t \sim R^*|_{T_v}(x)} [\ell(t,1)] \right\} \\
&\geq \mathop{\mathbb{E}}_{t \sim R^{T_v}(x)} \left[\ell(t, h_\ell^T(x)) \right],
\end{aligned}$$

and the claim follows. □

Recall that $\mathsf{exp\text{-}loss}_{\mathcal{D}}(\hat{h}) - \mathsf{exp\text{-}loss}_{\mathcal{D}}(h) \leq k(\alpha + \lambda)$ by Equation (18). Combining it with the claim yields that h_ℓ^T is $(\mathcal{H}, \ell, k(\alpha + \lambda))$-loss-minimizing. The following lemma completes the proof, by showing that the difference between $\mathsf{exp\text{-}loss}_{\mathcal{D}}\left(h_\ell^T\right)$ and $\mathsf{exp\text{-}loss}_{\mathcal{D}}\left(\tilde{h}\right)$ is small.

Lemma G.4. $\mathsf{exp\text{-}loss}_{\mathcal{D}}\left(\tilde{h}\right) - \mathsf{exp\text{-}loss}_{\mathcal{D}}\left(h_\ell^T\right) \leq 2k(\alpha + \lambda)$.

Proof. Let us denote

$$L_1 := \mathop{\mathbb{E}}_{\substack{x\sim T_v \\ t\sim R^*|_{T_v}(x)}} [\ell(t,1)]\,,$$

and, analogously,

$$L_0 := \mathop{\mathbb{E}}_{\substack{x\sim T_v \\ t\sim R^*|_{T_v}(x)}} [\ell(t,0)]\,.$$

This means that for any R,

$$\mathsf{exp\text{-}loss}_{\mathcal{D}}\left((\rho_\ell^*(R)\right) = \mathop{\mathbb{E}}_{x\sim T_v}[\rho_\ell^*(R(x))]\cdot L_1 + \left(1 - \mathop{\mathbb{E}}_{x\sim T_v}[\rho_\ell^*(R(x))]\right)\cdot L_0,$$

since ρ_ℓ^* is not randomized. Hence,

$$\begin{aligned}
\mathsf{exp\text{-}loss}_{\mathcal{D}}(\tilde{h}) - \mathsf{exp\text{-}loss}_{\mathcal{D}}(h_\ell^T) &= \mathop{\mathbb{E}}_{x\sim T_v}\left[\rho_\ell^*(\tilde{R}(x))\right]\cdot L_1 + \left(1 - \mathop{\mathbb{E}}_{x\sim T_v}\left[\rho_\ell^*(\tilde{R}(x))\right]\right)\cdot L_0 \\
&\quad - \mathop{\mathbb{E}}_{x\sim T_v}\left[\rho_\ell^*(R^{T_v}(x))\right]\cdot L_1 + \left(1 - \mathop{\mathbb{E}}_{x\sim T_v}\left[\rho_\ell^*(R^{T_v}(x))\right]\right)\cdot L_0 \\
&= (L_1 - L_0)\left(\mathbb{E}\left[\rho_\ell^*(\tilde{R}(x))\right] - \mathbb{E}\left[\rho_\ell^*(R^{T_v}(x))\right]\right) \\
&\leq |L_1 - L_0|\cdot\left|\mathbb{E}\left[\rho_\ell^*(\tilde{R}(x))\right] - \mathbb{E}\left[\rho_\ell^*(R^{T_v}(x))\right]\right|. \qquad (19)
\end{aligned}$$

We proceed by considering two cases. If $|L_1 - L_0| \leq 2k(\alpha+\lambda)$, the claim follows immediately. Otherwise, $|L_1 - L_0| > 2k(\alpha+\lambda)$. Let us assume, without loss of generality, that $L_0 - L_1 > 2k(\alpha+\lambda)$. First, notice that

$$\begin{aligned}
&\left|\mathop{\mathbb{E}}_{\substack{x\sim T_v \\ t^*(x)\sim R^*(x)}}[\ell(t^*(x),0) - \ell(t^*(x),1)] - \mathop{\mathbb{E}}_{\substack{x\sim T_v \\ \tilde{t}(x)\sim\tilde{R}(x)}}\left[\ell(\tilde{t}(x),0) - \ell(\tilde{t}(x),1)\right]\right| = \qquad (20) \\
&\left|\mathop{\mathbb{E}}_{x\sim T_v}\left[\sum_{t\in\mathcal{T}}\left(\Pr[t^*(x)=t] - \Pr[\tilde{t}(x)=t]\right)\cdot(\ell(t,0)-\ell(t,1))\right]\right| \leq \\
&\mathop{\mathbb{E}}_{x\sim T_v}\sum_{t\in\mathcal{T}}\left|\Pr[t^*(x)=t] - \Pr[\tilde{t}(x)=t]\right|\cdot|\ell(t,0)-\ell(t,1)| \leq 2k\alpha,
\end{aligned}$$

by $(\mathcal{C},\alpha,\lambda)$-full-multi-calibration of $\tilde{R}$. Notice that the first term in Equation (20) is exactly $L_0 - L_1$, hence we conclude that

$$\mu := \mathop{\mathbb{E}}_{\substack{x\sim T_v \\ \tilde{t}(x)\sim\tilde{R}(x)}}\left[\ell(\tilde{t}(x),0) - \ell(\tilde{t}(x),1)\right] \geq L_0 - L_1 - 2k\alpha \geq 2k\lambda.$$

Next, we argue that this leads to $\rho_\ell^*(\tilde{R}) = 1$. Specifically, observe that for any $x, x' \in T_v$, we have $||\tilde{R}(x) - \tilde{R}(x')||_\infty \leq \lambda$ by full-multi-calibration. Consequently, this implies that

$$\forall x\sim T_v, \left|\mathop{\mathbb{E}}_{\tilde{t}(x)\sim\tilde{R}(x)}\left[\ell(\tilde{t}(x),0) - \ell(\tilde{t}(x),1)\right] - \mu\right| \leq k\lambda,$$

Namely, for each $x\sim T_v$, the expected difference between the loss when rejecting and when accepting cannot deviate by more than $k\lambda$ from its expectation over all of T_v. Since $\mu \geq 2k\lambda$, it follows that

$$\mathop{\mathbb{E}}_{\tilde{r}(x)\sim\tilde{R}(x)}[\ell(\tilde{r}(x),0) - \ell(\tilde{r}(x),1)] > 0$$

for any $x\sim T_v$. Consequently, we obtain

$$\mathop{\mathbb{E}}_{x\sim T_v}\left[\rho_\ell^*(\tilde{R}(x))\right] = 1.$$

On the other hand, by definition of R^{T_v}, sampling $t\sim R^{T_v}$ gives the same marginal distribution as first sampling $x\sim D_X$ conditioned on $x\in T_v$ and then sampling $t\sim R^*(x)$. Since $L_0 > L_1$, it follows that

$$\mathop{\mathbb{E}}_{x\sim T_v}\left[\rho_\ell^*(R^{T_v}(x))\right] = 1.$$

Returning to Equation (19), the case where $|L_1 - L_0| > 2k(\alpha + \lambda)$ ensures that

$$\mathbb{E}\left[\rho_\ell^*(\tilde{R}(x))\right] = \mathbb{E}\left[\rho_\ell^*(R^{T_v}(x))\right],$$

which completes the proof. □

Recalling that h_ℓ^T is $(\mathcal{H}, \ell, k(\alpha + \lambda))$-loss-minimizing, we conclude that $\tilde{h}$ is $(\mathcal{H}, \ell, 3k(\alpha + \lambda))$-loss-minimizing, and the theorem follows. □

Remark G.5. *Returning to Lemma E.8, which shows that the loss-minimizing decision rule is far from affine, we observe that this claim also follows as a corollary of the arguments just presented. Suppose, for the sake of contradiction, that ρ_ℓ^* were close to affine. Then, for any calibrated predictor $\tilde{R}$, the action function $h = \rho_\ell^*(\tilde{R})$ would satisfy MAC by Corollary D.10. By Theorem G.1, h would also be loss-minimizing. However, this contradicts Theorem F.1, which shows that an action function cannot simultaneously satisfy MAC and minimize loss.*

G.1 Relaxation of the multi-calibration requirement

As discussed in Sections 2 and 3, full-multi-calibration (Definition C.8) is often infeasible due to exponential sample complexity in k, the number of labels. This section is devoted to relaxing the full-multi-calibration requirement of Theorem G.1 in order to achieve efficient sample complexity. We identify three settings in which this relaxation is possible: (i) when the label space is $\mathcal{T} = \Lambda[0, 1]$; (ii) when the loss function is known in advance during training of the predictor $\tilde{R}$; and (iii) when a *family* of loss functions is known during training. In the first two settings, the training time can be reduced to that of weak agnostic learning over the hypothesis class.

G.1.1 Probabilities and linear losses

We first consider the case where the label space is $\mathcal{T} = \Lambda[0, 1]$, and the predictor $\tilde{R}$ maps each input x to a distribution over $\Lambda[0, 1]$. In this setting, we argue that all natural loss functions $\ell : \Lambda[0, 1] \times \{0, 1\} \to [0, 1]$ are linear in their first argument. This corresponds to a modeling where the loss depends only on the realized outcome, rather than the probability assigned to it. For instance, in the context of loan approvals, the incurred loss depends solely on whether the borrower repays the loan; once the outcome is revealed, the specific predicted probability (e.g., 0.3 vs. 0.8) becomes irrelevant.

We show that under linear loss functions, probabilistic predictors offer no advantage over deterministic ones that output the expected prediction. In particular, the predictor that maps x to the expected value under $\tilde{R}(x)$ incurs the same loss as $\tilde{R}$ itself.

We define the *expectation-predictor*, which replaces a randomized prediction with its mean. Formally, for a predictor $\tilde{R}$ and input x, we set

$$\tilde{q}(x) = \sum_{\beta \in \Lambda[0,1]} \beta \cdot \tilde{R}(x)_\beta, \tag{21}$$

and define

$$\underset{\tilde{R}}{\mathbb{E}}(x) = (\tilde{q}(x), 1 - \tilde{q}(x)). \tag{22}$$

We refer to $\mathbb{E}_{\tilde{R}}$ as the *expectation-predictor*. This transformation maps a k-dimensional predictor to a 2-dimensional one, consistent with the binary label space $\{0, 1\}$. The expectation is taken over the distribution induced by $\tilde{R}$ with x fixed. Although we could equivalently define the expectation-predictor as a scalar-valued function outputting $\tilde{q}(x)$, we retain the stochastic vector form for consistency with our prior notation. Analogously, we define

$$q^*(x) = \sum_{\beta \in \Lambda[0,1]} \beta \cdot R^*(x)_\beta,$$

and define Nature's expectation-predictor $\mathbb{E}_{R^*}$ in the same way. We measure Multi-Calibration with respect to $\mathcal{D} = \mathcal{D}_\mathcal{X} \times \mathbb{E}_{R^*}(\mathcal{D}_\mathcal{X})$, using the formulation in Definition C.8 with $k = 2$, consistent with Gopalan et al. [2022]. Note that this is equivalent to training $\tilde{q}$ to be multi-calibrated with respect to $\mathcal{D} = \mathcal{D}_\mathcal{X} \times q^*(\mathcal{D}_\mathcal{X})$, and we later show how to achieve this efficiently using samples from R^*. We now state the main consequence of this reduction:

Corollary G.6. *Let $\mathcal{C}$, $\mathcal{H}$, and λ be as in Theorem G.1, with $\mathcal{T} = \Lambda[0,1]$. Let $\ell : \Lambda[0,1] \times \{0,1\} \to [0,1]$ be a loss function that is linear in its first argument. If the expectation-predictor $\mathbb{E}_{\tilde{R}}$ is $(\mathcal{C}, \alpha, \lambda)$-full-multi-calibrated with respect to $\mathcal{D}$, then the action function $\rho^*_\ell \circ \tilde{R}$ is $(\mathcal{H}, \ell, 6(\alpha+\lambda))$-loss-minimizing.*

Proof. The idea is to leverage the linearity of ℓ, which is natural when $\mathcal{T} = \Lambda[0,1]$, to reduce it to a "binary" loss function $\ell' : \{0,1\} \times \{0,1\} \to [0,1]$. We define ℓ' to coincide with ℓ on binary inputs: for all $b, a \in \{0,1\}$, let $\ell'(b,a) = \ell(b,a)$.[12]

We now apply Theorem G.1 to the expectation-predictor $\mathbb{E}_{\tilde{R}}$, with respect to ℓ' and $\mathcal{H}$. By the theorem, the induced action function

$$\tilde{f} = \rho^*_{\ell'} \circ \mathop{\mathbb{E}}_{\tilde{R}}$$

is $(\mathcal{H}, \ell', 6(\alpha+\lambda))$-loss-minimizing, where $\rho^*_{\ell'}$ denotes the optimal decision rule for $k = 2$ (as in Definition D.7).

Our goal is to show that

$$\tilde{h} = \rho^*_\ell \circ \tilde{R}$$

is $(\mathcal{H}, \ell, 6(\alpha+\lambda))$-loss-minimizing. To this end, we first establish that for any $x \in \mathcal{X}$, any predictor $R : \mathcal{X} \to \mathcal{Y}$ over $\mathcal{T} = \Lambda[0,1]$ (with corresponding q and expectation-predictor $\mathbb{E}_R$ as defined in Equations (21) and (22)), and any action $a \in \{0,1\}$, the following holds:

$$\mathop{\mathbb{E}}_{b \sim \mathbb{E}_R(x)} [\ell'(b,a)] = \mathop{\mathbb{E}}_{\beta \sim R(x)} [\ell(\beta, a)]\,. \tag{23}$$

Indeed, the left-hand side of Equation (23) expands to

$$q(x) \cdot \ell'(1,a) + (1 - q(x)) \cdot \ell'(0,a) = q(x) \cdot (\ell'(1,a) - \ell'(0,a)) + \ell'(0,a).$$

The right-hand side is

$$\sum_{\beta \in \Lambda[0,1]} R(x)_\beta \cdot \ell(\beta, a).$$

Since ℓ is linear in its first argument, we have

$$\ell(\beta, a) = \beta \cdot (\ell(1,a) - \ell(0,a)) + \ell(0,a),$$

and thus

$$\begin{aligned}\mathop{\mathbb{E}}_{\beta \sim R(x)} [\ell(\beta, a)] &= \left(\sum_{\beta \in \Lambda[0,1]} \beta \cdot R(x)_\beta \right) \cdot (\ell(1,a) - \ell(0,a)) + \ell(0,a) \\ &= q(x) \cdot (\ell(1,a) - \ell(0,a)) + \ell(0,a).\end{aligned}$$

By the definition of ℓ', the two sides are equal.

Equation (23) implies that for any $a \in \{0,1\}$,

$$\mathop{\mathbb{E}}_{b \sim \mathbb{E}_{\tilde{R}}(x)} [\ell'(b,a)] = \mathop{\mathbb{E}}_{\beta \sim \tilde{R}(x)} [\ell(\beta, a)]\,,$$

and, in particular,

$$\mathop{\arg\min}_{a \in \{0,1\}} \mathop{\mathbb{E}}_{b \sim \mathbb{E}_{\tilde{R}}(x)} [\ell'(b,a)] = \mathop{\arg\min}_{a \in \{0,1\}} \mathop{\mathbb{E}}_{\beta \sim \tilde{R}(x)} [\ell(\beta, a)]\,,$$

which implies that $\tilde{f} = \tilde{h}$. Applying Equation (23) again yields

$$\mathop{\mathbb{E}}_{b \sim \mathbb{E}_{R^*}(x)} \left[\ell'(b, \tilde{f}(x))\right] = \mathop{\mathbb{E}}_{\beta \sim R^*(x)} \left[\ell(\beta, \tilde{f}(x))\right],$$

so

$$\mathop{\mathbb{E}}_{b \sim \mathbb{E}_{R^*}(x)} \left[\ell'(b, \tilde{f}(x))\right] = \mathop{\mathbb{E}}_{\beta \sim R^*(x)} \left[\ell(\beta, \tilde{h}(x))\right]. \tag{24}$$

[12] Strictly speaking, ℓ is not defined on inputs like $\ell(0,a)$ and $\ell(1,a)$, since $0, 1 \notin \Lambda[0,1]$. We slightly abuse notation by interpreting $\ell'(b,a) = \ell(b,a)$ as shorthand for $\ell'(0,a) = \ell(\lambda/2, a)$ and $\ell'(1,a) = \ell(1-\lambda/2, a)$, where $\lambda/2$ and $1-\lambda/2$ are the smallest and largest elements in $\Lambda[0,1]$, respectively.

Similarly, for any $h \in \mathcal{H}$, Equation (23) implies

$$\mathop{\mathbb{E}}_{b\sim\mathbb{E}_{R^*}(x)}\left[\ell'(b,h(x))\right] = \mathop{\mathbb{E}}_{\beta\sim R^*(x)}\left[\ell(\beta,h(x))\right]. \tag{25}$$

Since $\tilde{f}$ is $(\mathcal{H},\ell',6(\alpha+\lambda))$-loss-minimizing, it holds for all $h\in\mathcal{H}$ that

$$\mathop{\mathbb{E}}_{b\sim\mathbb{E}_{R^*}(x)}\left[\ell'(b,\tilde{f}(x))\right] \le \mathop{\mathbb{E}}_{b\sim\mathbb{E}_{R^*}(x)}\left[\ell'(b,h(x))\right] + 6(\alpha+\lambda).$$

Substituting from Equations (24) and (25), we conclude that for all $h\in\mathcal{H}$,

$$\mathop{\mathbb{E}}_{\beta\sim R^*(x)}\left[\ell(\beta,\tilde{h}(x))\right] \le \mathop{\mathbb{E}}_{\beta\sim R^*(x)}\left[\ell(\beta,h(x))\right] + 6(\alpha+\lambda),$$

which completes the proof. □

Efficient training of $\tilde{q}$. We consider the standard one-dimensional multi-calibration setting, consistent with Definition C.8 instantiated with $k=1$.[13] We reduce samples from R^* to samples from a one-dimensional predictor q^*: for each $x\sim\mathcal{D}_{\mathcal{X}}$, we sample $\beta^*(x)\sim R^*(x)$ and then draw $y^*(x)\sim\text{Ber}(\beta^*(x))$. By the law of total probability,

$$\begin{aligned}\Pr\left[y^*(x)=1\right] &= \sum_{\beta\in\Lambda[0,1]}\Pr\left[\beta^*(x)=\beta\right]\cdot\Pr\left[y^*(x)=1 \mid \beta^*(x)=\beta\right]\\ &= \sum_{\beta\in\Lambda[0,1]} R^*(x)_\beta\cdot\beta = q^*(x).\end{aligned}$$

To efficiently train $\tilde{q}$, we apply the algorithm of Theorem 3.10 in Hébert-Johnson et al. [2018]. This algorithm achieves sample complexity

$$O\left(\frac{\log(|\mathcal{C}|/\alpha\lambda\delta)}{\alpha^4\lambda^{3/2}}\right),$$

where δ bounds the failure probability of the algorithm. As for the training time, Theorem 4.2 in Hébert-Johnson et al. [2018] shows that the task of learning a multi-calibrated predictor reduces to weak agnostic learning over $\mathcal{C}$, following the same framework described at the end of Appendix H.1.2.

G.1.2 Types and a single loss

Next, we present an alternative way to relax the multi-calibration assumption in a different setting. Here, the loss function is known in advance, before training the predictor. On the other hand, this approach applies to the more general case of types, where we make no assumptions on the outcome space. The relation between Corollary G.6 and Corollary G.7 can be viewed as a tradeoff between the generality of the outcome space and the prior knowledge of the loss function: the former supports arbitrary losses but requires the outcome space to be a discrete set of probabilities, whereas the latter accommodates arbitrary outcome spaces at the cost of assuming that the loss function is fixed during training.

We define the *loss-weighted predictor* with respect to a loss function ℓ, which replaces randomized predictions with scores that reflect the expected loss. Formally, for a loss function $\ell:\mathcal{T}\times\{0,1\}\to[0,1]$, a predictor R, and an input $x\in\mathcal{X}$, we define:

$$p_{R,\ell}(x) := \frac{1-\sum_{t\in\mathcal{T}}\left(\ell(t,1)-\ell(t,0)\right)\cdot R(x)_t}{2}. \tag{26}$$

This implies that $\forall x\in X$,

$$\rho^*_\ell(R(x))=1 \iff \mathop{\mathbb{E}}_{t(x)\sim R(x)}\left[\ell(t(x),1)-\ell(t(x),0)\right]<0 \iff p_{R,\ell}(x)>\frac{1}{2}.$$

We denote

$$\tilde{p}(x) := p_{\tilde{R},\ell}(x), \text{ and } p^*(x) := p_{R^*,\ell}(x).$$

[13] Formally, Definition C.8 defines a stochastic vector-valued predictor, but the same definition applies when the predictor outputs a scalar.

We refer to $\tilde{p}$ as the *loss-weighted predictor* and measure (one-dimensional) multi-calibration of $\tilde{p}$ w.r.t. $\mathcal{D} = \mathcal{D}_{\mathcal{X}} \times p^*(\mathcal{D}_{\mathcal{X}})$. Note that $\tilde{p}(x), p^*(x) \in [0,1]$ for all x by construction. The following corollary shows that multi-calibration of $\tilde{p}$ suffices to guarantee that $\rho_\ell^* \circ \tilde{R}$ is loss-minimizing. This result can be viewed as a simplified special case of Theorem G.1. For simplicity, we trivially assume that $\mathcal{X} \in \mathcal{C}$ and that the discretization is aligned with the threshold $1/2$.

Corollary G.7. *Let $\mathcal{C}$, $\mathcal{H}$, and λ be as in Theorem G.1, and let $\ell : \mathcal{T} \times \{0,1\} \to [0,1]$ be a loss function. If the loss-weighted predictor $\tilde{p}$ is $(\mathcal{C}, \alpha, \lambda)$-full-multi-calibrated with respect to $\mathcal{D}$, then the action function $\rho_\ell^* \circ \tilde{R}$ is $(\mathcal{H}, \ell, 4\alpha/\lambda)$-loss-minimizing.*

Proof. Let $h \in \mathcal{H}$ and write $\tilde{h}_\ell := \rho_\ell^* \circ \tilde{R}$. For any $h' : \mathcal{X} \to \{0,1\}$, define

$$\Delta_{h'}(x) := \mathop{\mathbb{E}}_{t^*(x) \sim R^*(x)} \left[\ell(t^*(x), h'(x))\right].$$

By the definition of p^* and $\tilde{p}$,

$$\mathop{\mathbb{E}}_{t^*(x) \sim R^*(x)} \left[\ell(t^*(x), 1) - \ell(t^*(x), 0)\right] = 1 - 2p^*(x),$$

and similarly the loss difference according to $\tilde{p}$ is $1 - 2\tilde{p}(x)$. Therefore,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} \left[\Delta_{\tilde{h}_\ell}(x) - \Delta_h(x)\right] \\
&= \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} \left[(1 - 2p^*(x)) \cdot \left(\tilde{h}_\ell(x) - h(x)\right)\right] \\
&= \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} \left[(1 - 2\tilde{p}(x)) \cdot \left(\tilde{h}_\ell(x) - h(x)\right)\right] + 2 \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} \left[(\tilde{p}(x) - p^*(x)) \cdot \left(\tilde{h}_\ell(x) - h(x)\right)\right].
\end{aligned}$$

The first term is non-positive, since $\tilde{h}_\ell(x) = 1$ iff $\tilde{p}(x) > 1/2$, and hence $\tilde{h}_\ell(x)$ minimizes the loss according to $\tilde{p}(x)$. Thus,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} \left[\Delta_{\tilde{h}_\ell}(x) - \Delta_h(x)\right] \\
&\leq 2 \left| \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} \left[(\tilde{p}(x) - p^*(x))\tilde{h}_\ell(x)\right] \right| + 2 \left| \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} \left[(\tilde{p}(x) - p^*(x))h(x)\right] \right|.
\end{aligned}$$

The second term is at most $2\alpha/\lambda$, since $\{x : h(x) = 1\} \in \mathcal{C}$ and summing the full-multi-calibration guarantee over all λ-level sets of $\tilde{p}$ contributes at most α/λ. The first term is also at most $2\alpha/\lambda$, since $\{x : \tilde{h}_\ell(x) = 1\} = \{x : \tilde{p}(x) > 1/2\}$ is a union of λ-level sets of $\tilde{p}$ inside $\mathcal{X}$, and $\mathcal{X} \in \mathcal{C}$. Altogether,

$$\mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D}_{\mathcal{X}} \\ t^*(x) \sim R^*(x)}} \left[\ell\left(t^*(x), \tilde{h}_\ell(x)\right)\right] \leq \mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D}_{\mathcal{X}} \\ t^*(x) \sim R^*(x)}} \left[\ell\left(t^*(x), h(x)\right)\right] + \frac{4\alpha}{\lambda}.$$

Since this holds for every $h \in \mathcal{H}$, $\tilde{h}_\ell$ is $(\mathcal{H}, \ell, 4\alpha/\lambda)$-loss-minimizing. □

Efficient training of $\tilde{p}$. As in Appendix G.1.1, we consider the standard one-dimensional multi-calibration setting and apply the algorithm of Hébert-Johnson et al. [2018] to efficiently train $\tilde{p}$. To do so, we reduce samples from R^* to samples from p^*: for each $x \sim \mathcal{D}_{\mathcal{X}}$, we first sample $t^*(x) \sim R^*(x)$, and then define $y^*(x) \sim \text{Ber}\left(\frac{1 - (\ell(t^*(x),1) - \ell(t^*(x),0))}{2}\right)$. By the law of total probability,

$$\begin{aligned}
\Pr_{y^*(x)}\left[y^*(x) = 1\right] &= \sum_{t \in \mathcal{T}} \Pr_{t^*(x) \sim R^*(x)}\left[t^*(x) = t\right] \cdot \frac{1 - (\ell(t,1) - \ell(t,0))}{2} \\
&= \frac{1 - \mathbb{E}_{t^*(x) \sim R^*(x)}\left[\ell(t^*(x), 1) - \ell(t^*(x), 0)\right]}{2} = p^*(x).
\end{aligned}$$

G.1.3 Types and a family of losses

Corollaries G.6 and G.7 can be viewed as reductions from the task of loss minimization to that of training a one-dimensional predictor to be multi-calibrated. Generalizing Corollary G.7 to a *family* of loss functions requires a stronger guarantee: the learned predictor must be multi-calibrated not only with respect to a single loss, but simultaneously for every loss in a given family $\mathcal{L}$.

To that end, we aim to satisfy multi-calibration with respect to all convex combinations induced by $\mathcal{L}$, rather than a fixed weighting. We reduce this problem to identifying, in each round, a pair (ℓ, h) with $\ell \in \mathcal{L}$ and $h \in \mathcal{H}$ such that the expected loss of $\rho^*_\ell \circ \tilde{R}$ exceeds that of h. We refer to the procedure that searches for such violating pairs as an *auditor*. In our setting, the auditor is realized by evaluating a family of distinguishers — each parameterized by (ℓ, h) — that attempt to distinguish Nature's outcomes from those generated by $\tilde{R}$. A distinguisher that succeeds reveals a violation, and the corresponding pair is used to update the predictor. As in previous settings, query complexity remains low as long as each update affects a non-negligible portion of the domain. However, the total runtime is dominated by the auditing step: namely, the search over $\mathcal{L} \times \mathcal{H}$. Unlike the single-loss case, this setting does not reduce to weak agnostic learning over $\mathcal{C}$.

Corollary G.8. *Let $\mathcal{C}$ and $\mathcal{H}$ be as in Theorem G.1, and let $\mathcal{L} \subseteq \{\mathcal{T} \times \{0,1\} \to [0,1]\}$ be a family of loss functions. Given $\tilde{O}\left(\log k \cdot (\log|\mathcal{C}| + \log|\mathcal{L}|)/\varepsilon^4\right)$ samples, one can train a predictor $\tilde{R}$ such that, for every $\ell \in \mathcal{L}$, the composed rule $\rho^*_\ell \circ \tilde{R}$ is $(\mathcal{H}, \ell, \varepsilon)$-loss-minimizing.*

Proof. In Appendix H.1.2, we present *Outcome Indistinguishability* (OI) [Dwork et al., 2021] and discuss an algorithm for training an Outcome Indistinguishable predictor. At a high level, the idea behind OI is to define a family of computationally bounded distinguishers $\mathcal{A}$, such that a target objective (in our case, loss minimization) reduces to the distinguishers' inability to tell apart Nature's outcomes from those generated based on the predictions. A probabilistic predictor $\tilde{R} : \mathcal{X} \to \mathcal{Y}$ is said to be $(\mathcal{A}, \varepsilon)$-OI with respect to $\mathcal{D}$ if for any $A \in \mathcal{A}$,

$$\left| \Pr_{(x,t^*(x))\sim\mathcal{D}} \left[A\left(x, t^*(x); \tilde{R}(x)\right) = 1 \right] - \Pr_{(x,\tilde{t}(x))\sim\mathcal{D}_{\mathcal{X}} \times \tilde{R}(x)} \left[A\left(x, \tilde{t}(x); \tilde{R}(x)\right) = 1 \right] \right| \leq \varepsilon.$$

This variant is called Sample-Access-OI. Theorem H.6 (from Dwork et al. [2021]) shows that Sample-Access-OI can be achieved using

$$\tilde{O}\left(\frac{\log k \cdot \log|\mathcal{A}|}{\varepsilon^4}\right)$$

samples. We refer the reader to Appendix H.1.2 for a detailed exposition.

We define a family of distinguishers such that if $\tilde{R}$ is OI with respect to them, then $\rho^*_\ell \circ \tilde{R}$ is $(\mathcal{H}, \ell, \varepsilon)$-loss-minimizing. Each distinguisher is parameterized by a loss function $\ell \in \mathcal{L}$ and a hypothesis $h \in \mathcal{H}$, and aims to detect whether h outperforms $\rho^*_\ell \circ \tilde{R}$ in terms of loss. Formally, the randomized Boolean distinguisher $A_{\ell,h}$ operates as follows:

$$A_{\ell,h}\left(x, t(x), \tilde{R}(x)\right) = \text{Compute } p(x) := \ell(t(x), \rho^*_\ell(\tilde{R}(x))) - \ell(t(x), h(x))$$
$$\text{Output } 1 \text{ with probability } \frac{p(x)+1}{2}.$$

While the distinguishers could have been defined to output $p(x)$ directly in $[-1, 1]$, we instead normalize and sample a bit, to be consistent with the formulation of Dwork et al. [2021].

We define

$$\tilde{p}(x) := \ell(\tilde{t}(x), \rho^*_\ell(\tilde{R}(x))) - \ell(\tilde{t}(x), h(x)) \quad \text{and} \quad p^*(x) := \ell(t^*(x), \rho^*_\ell(\tilde{R}(x))) - \ell(t^*(x), h(x)).$$

By Theorem H.6, given

$$\tilde{O}\left(\frac{\log k \cdot \log|\mathcal{A}|}{(\varepsilon')^4}\right)$$

samples from R^*, we can learn a predictor $\tilde{R}$ such that for all $A_{\ell,h} \in \mathcal{A}$,

$$\left| \Pr_{(x,t^*(x))\sim\mathcal{D}} \left[A_{\ell,h}(x, t^*(x); \tilde{R}(x)) = 1 \right] - \Pr_{(x,\tilde{t}(x))\sim\mathcal{D}_{\mathcal{X}} \times \tilde{R}(x)} \left[A_{\ell,h}(x, \tilde{t}(x); \tilde{R}(x)) = 1 \right] \right| \leq \varepsilon'.$$

Equivalently,

$$\left|\mathbb{E}_{x,t^*(x)}\left[A_{\ell,h}(x,t^*(x);\tilde{R}(x))\right]-\mathbb{E}_{x,\tilde{t}(x)}\left[A_{\ell,h}(x,\tilde{t}(x);\tilde{R}(x))\right]\right|\leq\varepsilon'. \tag{27}$$

Let us first focus on the right term. By the definition of ρ^*_ℓ (see Definition D.7),

$$\mathbb{E}_{\tilde{t}(x)\sim\tilde{R}(x)}\left[\ell\left(\tilde{t}(x),\rho^*_\ell(\tilde{R}(x))\right)\right]\leq\mathbb{E}_{\tilde{t}(x)\sim\tilde{R}(x)}\left[\ell\left(\tilde{t}(x),h(x)\right)\right],$$

so

$$\mathbb{E}_{\tilde{t}(x)\sim\tilde{R}(x)}\left[\ell\left(\tilde{t}(x),\rho^*_\ell(\tilde{R}(x))\right)-\ell\left(\tilde{t}(x),h(x)\right)\right]\leq 0.$$

Recalling the definition of $\tilde{p}(x)$, we get that

$$\forall x\in\mathcal{X},\ \mathbb{E}_{\tilde{t}}\left[\tilde{p}(x)\right]\leq 0\implies\frac{\mathbb{E}_{\tilde{t}}\left[\tilde{p}(x)\right]+1}{2}\leq\frac{1}{2},$$

and, overall,

$$\mathbb{E}_{x,\tilde{t}(x),A_{\ell,h}}\left[A_{\ell,h}\left(x,\tilde{t}(x);\tilde{R}(x)\right)\right]=\frac{\mathbb{E}_{\tilde{t}}\left[\tilde{p}(x)\right]+1}{2}\leq\frac{1}{2}.$$

Returning to Equation (27), this implies that

$$\mathbb{E}_{x,t^*(x),A_{\ell,h}}\left[A_{\ell,h}\left(x,t^*(x);\tilde{R}(x)\right)\right]\leq\frac{1}{2}+\varepsilon',$$

that is,

$$\mathbb{E}_{x,t^*(x)}\left[\frac{p^*(x)+1}{2}\right]\leq\frac{1}{2}+\varepsilon',$$

so $\mathbb{E}_{x,t^*(x)}[p^*(x)]\leq 2\varepsilon'$. By the definition of p^*, this means

$$\mathbb{E}_{x,t^*(x)}\left[\ell(t^*(x),\rho^*_\ell(\tilde{R}(x)))-\ell(t^*(x),h(x))\right]\leq 2\varepsilon',$$

hence,

$$\mathbb{E}_{x,t^*(x)}\left[\ell(t^*(x),\rho^*_\ell(\tilde{R}(x)))\right]\leq\mathbb{E}_{x,t^*(x)}\left[\ell(t^*(x),h(x))\right]+2\varepsilon'.$$

Setting $\varepsilon'=\varepsilon/2$ concludes the proof. □

H Deferred proofs

This section provides the rather standard proofs deferred from Appendix C.

H.1 Multi-accuracy and multi-calibration

We start with the proof of Claim C.5.

Claim H.1. *Let S, S' be two sets such that $S'\subseteq S$. If $\tilde{R}$ is α-coordinate-wise-accurate w.r.t. $\mathcal{D}$ on S, then it is α'-coordinate-wise-accurate w.r.t. $\mathcal{D}$ on S', where $\alpha'=\alpha+\Pr_{x\sim\mathcal{D}_\mathcal{X}}\left[x\in S\setminus S'\right]$.*

Proof. For any $t\in\mathcal{T}$,

$$\begin{aligned}&\Pr\left[t^*(x)=t\wedge x\in S'\right]-\Pr\left[\tilde{t}(x)=t\wedge x\in S'\right]\\&=\Pr\left[t^*(x)=t\wedge x\in S\right]-\Pr\left[\tilde{t}(x)=t\wedge x\in S\right]\\&-\Pr\left[t^*(x)=t\wedge x\in S\setminus S'\right]+\Pr\left[\tilde{t}(x)=t\wedge x\in S\setminus S'\right],\end{aligned}$$

where the probabilities are taken over $x\sim\mathcal{D}_\mathcal{X}, t^*(x)\sim R^*(x)$ and $x\sim\mathcal{D}_\mathcal{X},\tilde{t}(x)\sim\tilde{R}(x)$, respectively. By assumption, the difference is bounded by $\Pr_{x\sim\mathcal{D}_\mathcal{X}}\left[x\in S\setminus S'\right]+\alpha$. □

Next, we prove Claim C.10.

Claim H.2 (Coordinate-Wise-Multi-Calibration Implies Coordinate-Wise-Multi-Accuracy)**.** *In the setting of Definition C.9, if $\tilde{R}$ is $(\mathcal{C}, \alpha, \lambda)$-coordinate-wise-multi-calibrated, then $\tilde{R}$ is also $(\mathcal{C}, \alpha/\lambda)$-coordinate-wise-multi-accurate.*

Proof. Let $S \in \mathcal{C}$ and $t \in \mathcal{T}$. By the definition of $S_{t,\beta}$, we have $S \subseteq \bigcup_{\beta \in \Lambda[0,1]} S_{t,\beta}$, so

$$\begin{aligned}&\left|\Pr[t^*(x) = t \wedge x \in S] - \Pr[\tilde{t}(x) = t \wedge x \in S]\right| \\ &\leq \sum_{\beta \in \Lambda[0,1]} \left|\Pr[t^*(x) = t \wedge x \in S_{t,\beta}] - \Pr[\tilde{t}(x) = t \wedge x \in S_{t,\beta}]\right| \leq \frac{\alpha}{\lambda}.\end{aligned}$$

□

H.1.1 Notions of multi-accurate rankings

In this section, we examine the relationship between two definitions of Multi-Accuracy for Probabilistic Rankings, previously introduced in Appendix C.1. The first and stronger definition, Definition C.3, evaluates accuracy with respect to *thresholds*. Specifically, for every set S in the collection $\mathcal{C}$ and every rank τ, the probabilities assigned by R^* and $\tilde{R}$ to an individual from S being above threshold τ are similar.

The second definition, which is also natural, enforces *equality* on ranks, namely, coordinate-wise-equality. This alternative is motivated by decision rules that operate on specific segments—for example, a particular bank or university might target a specific segment of the population. The following definition restates Definition C.2, with a modified notation suited to the case of rankings, where $\mathcal{T} = [k]$.

Definition H.3 (Coordinate-Wise-Multi-Accuracy for Probabilistic Rankings)**.** *In the setting of Definition C.3, a probabilistic ranking $\tilde{R} : \mathcal{X} \to \mathcal{Y}$ satisfies $(\mathcal{C}, \alpha)$-coordinate-wise-multi-accuracy w.r.t. $\mathcal{D}$ if for every $S \in \mathcal{C}$ and every $\tau \in [k]$,*

$$\left| \Pr_{\substack{x \sim \mathcal{D}_{\mathcal{X}} \\ r^*(x) \sim R^*(x)}} [r^*(x) = \tau \wedge x \in S] - \Pr_{\substack{x \sim \mathcal{D}_{\mathcal{X}} \\ \tilde{r}(x) \sim \tilde{R}(x)}} [\tilde{r}(x) = \tau \wedge x \in S] \right| \leq \alpha.$$

The following claim shows that this definition is strictly weaker than the threshold definition.

Claim H.4. *$(\mathcal{C}, \alpha)$-threshold-multi-accuracy implies $(\mathcal{C}, 2\alpha)$-coordinate-wise-multi-accuracy.*

Proof. Let $S \in \mathcal{C}$ and $\tau \in [k]$. If $\tau = k$, then $(\mathcal{C}, \alpha)$-coordinate-wise-multi-accuracy follows directly, since $r(x) = k \iff r(x) > k - 1$. Otherwise, by the assumption,

$$|\Pr[r^*(x) > \tau \wedge x \in S] - \Pr[\tilde{r}(x) > \tau \wedge x \in S]| \leq \alpha$$

and

$$|\Pr[r^*(x) > \tau - 1 \wedge x \in S] - \Pr[\tilde{r}(x) > \tau - 1 \wedge x \in S]| \leq \alpha.$$

Hence,

$$\begin{aligned}&|\Pr[r^*(x) = \tau \wedge x \in S] - \Pr[\tilde{r}(x) = \tau \wedge x \in S]| = \\ &|\Pr[r^*(x) > \tau - 1 \wedge x \in S] - \Pr[\tilde{r}(x) > \tau - 1 \wedge x \in S] - \\ &(\Pr[r^*(x) > \tau \wedge x \in S] - \Pr[\tilde{r}(x) > \tau \wedge x \in S])| \leq 2\alpha\end{aligned}$$

follows from triangle inequality.

On the other hand, note that an error of 2α is tight. Let $k \geq 3$, $S \in \mathcal{C}$ and let $\tilde{R}$ be such that

$$\begin{aligned}\Pr[\tilde{r}(x) = k \wedge x \in S] &= \Pr[r^*(x) = k \wedge x \in S] + \alpha, \\ \Pr[\tilde{r}(x) = k - 1 \wedge x \in S] &= \Pr[r^*(x) = k - 1 \wedge x \in S] - 2\alpha, \\ \Pr[\tilde{r}(x) = k - 2 \wedge x \in S] &= \Pr[r^*(x) = k - 2 \wedge x \in S] + \alpha, \\ \forall \tau \leq k - 3,\ \Pr[\tilde{r}(x) = \tau \wedge x \in S] &= \Pr[r^*(x) = \tau \wedge x \in S],\end{aligned}$$

where all probabilities are taken over sampling an element $x \sim \mathcal{D}_{\mathcal{X}}$ and a rank from either $\tilde{R}(x)$ or $R^*(x)$. This ranking $\tilde{R}$ is α-threshold-multi-accurate on S, yet its coordinate-wise-multi-accuracy error is 2α. □

On the other hand, the following claim shows that Definition H.3 implies Definition C.3 with multiplicative k blowup in the error parameter:

Claim H.5. *$(\mathcal{C}, \alpha)$-coordinate-wise-multi-accuracy implies $(\mathcal{C}, \Theta(k \cdot \alpha))$-threshold-multi-accuracy.*

Proof. First, we show that $(\mathcal{C}, \alpha)$-coordinate-wise-multi-accuracy implies $(\mathcal{C}, O(k \cdot \alpha))$-threshold-multi-accuracy. Let $S \in \mathcal{C}$ and $\tau \in [k-1]$. By assumption,

$$\begin{aligned}
&|\Pr[r^*(x) > \tau \wedge x \in S] - \Pr[\tilde{r}(x) > \tau \wedge x \in S]| = \\
&|\Pr[r^*(x) \in \{\tau+1, \ldots, k\} \wedge x \in S] - \Pr[\tilde{r}(x) \in \{\tau+1, \ldots, k\} \wedge x \in S]| = \\
&\left| \sum_{i=\tau+1}^{k} (\Pr[r^*(x) = i \wedge x \in S] - \Pr[\tilde{r}(x) = i \wedge x \in S]) \right| \leq (k-1) \cdot \alpha.
\end{aligned}$$

Next, we demonstrate that this bound is tight by presenting an example where a ranking is $(\mathcal{C}, \alpha)$-coordinate-wise-multi-accurate but incurs a threshold-multi-accuracy error of $(k-1) \cdot \alpha$. Consider the following example. Let $k \in \mathbb{N}_{\mathsf{even}}$ and let R^* be a deterministic Nature that is uniform over the ranks $\forall S \in \mathcal{C}$, namely,

$$\forall \tau \in [k],\ \Pr[r^*(x) = \tau \wedge x \in S] = \frac{1}{k}.$$

Consider the following deterministic ranking $\tilde{R}$: For $S \in \mathcal{C}$,

$$\tau \in \{1, \ldots, k/2\},\ \Pr[\tilde{r}(x) = \tau \wedge x \in S] = \frac{1}{k} + \alpha,$$

and

$$\tau \in \{k/2+1, \ldots, k\},\ \Pr[\tilde{r}(x) = \tau \wedge x \in S] = \frac{1}{k} - \alpha.$$

By definition, it follows immediately that $\tilde{R}$ is $(\mathcal{C}, \alpha)$-coordinate-wise-multi-accurate. On the other hand, consider $\tau = \frac{k}{2} - 1$ in Definition C.3:

$$\left| \Pr\left[r^*(x) > \frac{k}{2} - 1 \wedge x \in S \right] - \Pr\left[\tilde{r}(x) > \frac{k}{2} - 1 \wedge x \in S \right] \right| = (k/2) \cdot \alpha.$$

This means that $\tilde{R}$ is $(\mathcal{C}, \Omega(k \cdot \alpha))$-threshold-multi-accurate, and the claim follows. □

H.1.2 Learning multi-accurate and multi-calibrated predictors

We use the framework of *Outcome Indistinguishability*, introduced by Dwork et al. [2021] in the context of binary prediction tasks. It articulates the goals of prediction using the language of computational indistinguishability: a predictor is outcome-indistinguishable (OI) if no computationally bounded observer can distinguish Nature's outcomes from those generated based on the predictions. This framework ensures that predictive models yield outcomes that cannot be efficiently refuted using real-life observations.

Dwork et al. also provide algorithms for learning OI predictors. The intuition behind their learning approach is as follows: if there exists a computationally bounded distinguisher A that distinguishes between Nature R^* and the learned predictor $\tilde{R}$, use A to update $\tilde{R}$; otherwise, R^* and $\tilde{R}$ are indistinguishable, and the process terminates. This strategy can be interpreted from different perspectives—as a form of boosting or gradient descent—and is closely related to the approach used by Hébert-Johnson et al. [2018] to learn multi-calibrated predictors.

In a subsequent work ([Dwork et al., 2022]), Dwork et al. extend the concept of outcome indistinguishability beyond binary (Bernoulli) outcomes to more complex, discrete, or continuous domains. They provide learning algorithms for constructing OI predictors, which we use next to achieve multi-accurate and multi-calibrated probabilistic predictors.

Formally, let $\mathcal{A}$ be a class of distinguishers and ε an indistinguishability parameter. A probabilistic predictor $\tilde{R} : \mathcal{X} \to \mathcal{Y}$ is said to be $(\mathcal{A}, \varepsilon)$-*no-access*-OI with respect to $\mathcal{D}$ if for any $A \in \mathcal{A}$,

$$\left| \Pr_{\substack{(x, t^*(x)) \sim \mathcal{D} \\ A\text{'s randomness}}} [A(x, t^*(x)) = 1] - \Pr_{\substack{(x, \tilde{t}(x)) \sim \mathcal{D}_{\mathcal{X}} \times \tilde{R}(x) \\ A\text{'s randomness}}} \left[A\left(x, \tilde{t}(x)\right) = 1\right] \right| \leq \varepsilon,$$

and *sample-access*-OI with respect to $\mathcal{D}$ if for any $A \in \mathcal{A}$,

$$\left| \Pr_{\substack{(x,t^*(x))\sim\mathcal{D} \\ A\text{'s randomness}}} \left[A\left(x, t^*(x); \tilde{R}(x)\right) = 1 \right] - \Pr_{\substack{(x,\tilde{t}(x))\sim\mathcal{D}_{\mathcal{X}}\times\tilde{R}(x) \\ A\text{'s randomness}}} \left[A\left(x, \tilde{t}(x); \tilde{R}(x)\right) = 1 \right] \right| \leq \varepsilon.$$

Dwork et al. present other, stronger variants of outcome indistinguishability, but we focus exclusively on these two. Moreover, we comment that our representation of distributions as stochastic vectors, which explicitly encode the probability mass function, coincides with the representation used by Dwork et al. for discrete random variables over a finite domain $|\mathcal{T}| = k < \infty$.

Theorem H.6 (Learning OI Predictors, Theorem 20 in Dwork et al. [2022])**.** *There exists an algorithm for learning an $(\mathcal{A}, \varepsilon)$-outcome-indistinguishable probabilistic predictor $\tilde{R} : \mathcal{X} \to \mathcal{Y}$ with respect to $\mathcal{D}$. The sample complexity is bounded by*

$$\tilde{O}\left(\frac{\log k \cdot \log |\mathcal{A}|}{\varepsilon^4}\right)$$

for the case of Sample-Access-OI, and by $O\left(\frac{\log |\mathcal{A}|}{\varepsilon^2}\right)$ for the case of No-Access-OI. Assuming that the distinguishers in $\mathcal{A}$ can be evaluated in time $\mathit{time}_{\mathcal{A}}$, the running time is

$$\tilde{O}\left(\frac{k \cdot \mathit{time}_{\mathcal{A}} \cdot \log |\mathcal{A}|}{\varepsilon^6} + \frac{|\mathcal{A}| \cdot \mathit{time}_{\mathcal{A}} \cdot \log k}{\varepsilon^4}\right).$$

With overwhelming probability over its samples, the algorithm produces a probabilistic predictor $\tilde{R}$ that can be evaluated in time

$$\tilde{O}\left(\frac{k \cdot \mathit{time}_{\mathcal{A}} \cdot \log |\mathcal{A}|}{\varepsilon^2}\right)$$

for the case of Sample-Access-OI, and by $O\left(k \cdot \mathit{time}_{\mathcal{A}} \cdot \log |\mathcal{A}|\right)$ for the case of No-Access-OI.

Recall Definition C.2, where $k = |\mathcal{T}|$. Using Theorem H.6, we derive our result for learning coordinate-wise-multi-accurate probabilistic predictors:

Theorem H.7 (Learning Coordinate-Wise-Multi-Accurate Predictors)**.** *Let $\mathcal{C} \subseteq \{0,1\}^{|\mathcal{X}|}$ be a collection of sets such that the set membership evaluation time for any $S \in \mathcal{C}$ is T. There exists an algorithm for learning a $(\mathcal{C}, \alpha)$-coordinate-wise-multi-accurate probabilistic predictor $\tilde{R} : \mathcal{X} \to \mathcal{Y}$ with respect to $\mathcal{D}$. The sample complexity is bounded by*

$$O\left(\frac{\log |\mathcal{C}| + \log k}{\alpha^2}\right).$$

The running time is

$$\tilde{O}\left(\frac{k \cdot T \cdot |\mathcal{C}|}{\alpha^6}\right).$$

With overwhelming probability over its samples, the algorithm produces a probabilistic predictor $\tilde{R}$ that can be evaluated in time $O\left(k \log k \cdot T \cdot \log |\mathcal{C}|\right)$.

Proof. We define the family of no-access-distinguishers $\mathcal{A} = (A_{S,t})_{S\in\mathcal{C}, t\in\mathcal{T}}$ as follows:

$$A_{S,t}(x, y) = \begin{cases} 1 & \text{if } y = t \text{ and } x \in S, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\tilde{R} : \mathcal{X} \to \mathcal{Y}$ be an $(\mathcal{A}, \alpha)$-no-access-OI predictor with respect to $\mathcal{D}$, guaranteed by Theorem H.6. This implies that for any $A_{S,t} \in \mathcal{A}$,

$$\left| \Pr_{(x,t^*(x))\sim\mathcal{D}} \left[A_{S,t}\left(x, t^*(x)\right) = 1 \right] - \Pr_{\left(x,\tilde{t}(x)\right)\sim\mathcal{D}_{\mathcal{X}}\times\tilde{R}(x)} \left[A_{S,t}\left(x, \tilde{t}(x)\right) = 1 \right] \right| \leq \alpha.$$

Expanding the definition of $A_{S,t}$, this ensures that for any $S \in \mathcal{C}$ and any $t \in \mathcal{T}$,

$$\left| \Pr_{(x,t^*(x))\sim\mathcal{D}} \left[t^*(x) = t \wedge x \in S \right] - \Pr_{\left(x,\tilde{t}(x)\right)\sim\mathcal{D}_{\mathcal{X}}\times\tilde{R}(x)} \left[\tilde{t}(x) = t \wedge x \in S \right] \right| \leq \alpha.$$

Thus, $\tilde{R}$ satisfies $(\mathcal{C}, \alpha)$-coordinate-wise-multi-accuracy, which completes the proof. The claimed complexity bounds follow from Theorem H.6. □

To learn threshold-multi-accurate rankings directly, as per Definition C.3, which is stronger than Definition C.2 in case of rankings, we can alternatively define the distinguishers

$$A_{S,\tau}(x,y) = \begin{cases} 1 & \text{if } y > \tau \text{ and } x \in S, \\ 0 & \text{otherwise,} \end{cases}$$

for any $\tau \in [k]$, and achieve threshold-multi-accuracy without any increase in complexity.

Next, we address the more complex task of learning multi-calibrated predictors. The following theorem shows that coordinate-wise-multi-calibration, as defined in Definition C.9, can be achieved efficiently.

Theorem H.8 (Learning Multi-Calibrated Predictors)**.** *Let $\mathcal{C} \subseteq \{0,1\}^{|\mathcal{X}|}$ be a collection of sets such that the set membership evaluation time for any $S \in \mathcal{C}$ is T, and let $\lambda \in (0,1)$ be a discretization parameter. There exists an algorithm for learning a $(\mathcal{C}, \alpha, \lambda)$-coordinate-wise-multi-calibrated probabilistic predictor $\tilde{R} : \mathcal{X} \to \mathcal{Y}$ with respect to $\mathcal{D}$. The sample complexity is bounded by*

$$\tilde{O}\left(\frac{\log^2 k \cdot \left(\log|\mathcal{C}| + \log\left(\frac{1}{\lambda}\right)\right)}{\alpha^4}\right).$$

The running time is

$$\tilde{O}\left(\frac{k \cdot T \cdot \left(|\mathcal{C}| \cdot \log(k) + \log\frac{1}{\lambda}\right)}{\lambda \cdot \alpha^6}\right).$$

With overwhelming probability over its samples, the algorithm produces a probabilistic predictor $\tilde{R}$ that can be evaluated in time

$$\tilde{O}\left(\frac{k \cdot T \cdot \log\left(\log k + \log|\mathcal{C}| + \log\frac{1}{\lambda}\right)}{\alpha^2}\right).$$

Proof. We define the family of sample-access-distinguishers $\mathcal{A} = (A_{S,t,\beta})_{S\in\mathcal{C},t\in\mathcal{T},\beta\in\Lambda[0,1]}$ with respect to the discretization λ as follows:

$$A_{S,t,\beta}(x,y;\tilde{R}(x)) = \begin{cases} 1 & \text{if } y = t \text{ and } x \in S \text{ and } \tilde{R}(x)_t \in \lambda(\beta), \\ 0 & \text{otherwise,} \end{cases}$$

where $\tilde{R}(x)_t$ denotes the t^{th} coordinate of the stochastic vector $\tilde{R}(x)$. Notice that evaluating $A_{S,t,\beta}$ can be done in time T, since $\Lambda[0,1]$ is known in advance.

Let $\tilde{R} : \mathcal{X} \to \mathcal{Y}$ be $(\mathcal{A}, \alpha)$-sample-access-OI predictor with respect to $\mathcal{D}$, guaranteed by Theorem H.6. This implies that for any $A_{S,t,\beta} \in \mathcal{A}$,

$$\left|\Pr_{(x,t^*(x))\sim\mathcal{D}}\left[A_{S,t,\beta}\left(x,t^*(x);\tilde{R}(x)\right)=1\right] - \Pr_{\left(x,\tilde{t}(x)\right)\sim\mathcal{D}_{\mathcal{X}}\times\tilde{R}(x)}\left[A_{S,t,\beta}\left(x,\tilde{t}(x);\tilde{R}(x)\right)=1\right]\right| \le \alpha.$$

Recall the definition of $S_{t,\beta}$ w.r.t. $\tilde{R}$:

$$S_{t,\beta} := \left\{x \in S \text{ such that } \Pr_{\tilde{t}(x)\sim\tilde{R}(x)}[\tilde{t}(x)=t] \in \lambda(\beta)\right\}.$$

By definition of $A_{S,t,\beta}$, this ensures that for any $S_{t,\beta}$ and any $t \in \mathcal{T}$,

$$\left|\Pr_{(x,t^*(x))\sim\mathcal{D}}[t^*(x)=t \wedge x \in S_{t,\beta}] - \Pr_{\left(x,\tilde{t}(x)\right)\sim\mathcal{D}_{\mathcal{X}}\times\tilde{R}(x)}\left[\tilde{t}(x)=t \wedge x \in S_{t,\beta}\right]\right| \le \alpha.$$

Thus, $\tilde{R}$ satisfies $(\mathcal{C}, \alpha, \lambda)$-coordinate-wise-multi-calibration w.r.t. $\mathcal{D}$.

Notice that the number of sets to consider is $\left|(S_{t,\beta})_{t\in\mathcal{T},\beta\in\Lambda[0,1]}\right| = k/\lambda$, therefore $|\mathcal{A}| = \frac{k\cdot|\mathcal{C}|}{\lambda}$ and the claimed complexity bounds follow from Theorem H.6. □

Replacing the linear dependence on $|\mathcal{C}|$ via weak agnostic learning. The running time bounds in Theorems H.7 and H.8 scale linearly with $|\mathcal{C}|$ due to the need to evaluate every distinguisher $A \in \mathcal{A}$ in each iteration, where each set $S \in \mathcal{C}$ corresponds to a distinguisher. To better understand this complexity, we revisit Theorem 20 in Dwork et al. [2022], which decomposes the learning time into three components: sample generation, searching, and reweighting. The most computationally expensive step is searching, where, in each iteration, we must identify a distinguisher $A \in \mathcal{A}$ that differentiates between R^* and the current predictor $\tilde{R}$. This is done via exhaustive search: iterating over all distinguishers in $\mathcal{A}$ and computing their distinguishing advantage. To estimate the distinguishing advantage, each distinguisher requires $O\left((\log|\mathcal{A}| + \log N)/\varepsilon^2\right)$ samples per iteration, where N is the iteration complexity. Since in the case of a discrete label distribution we have $N \leq O(\log k/\varepsilon^2)$, we obtain the complexity

$$\tilde{O}\left(\frac{|\mathcal{A}| \cdot \text{time}_{\mathcal{A}} \cdot \log k}{\varepsilon^4}\right),$$

as presented in Theorem H.6, where $\text{time}_{\mathcal{A}}$ is an upper bound on the time required to evaluate a single distinguisher.

To eliminate the need for exhaustive search over $\mathcal{A}$, we follow a standard approach and replace this step with a black-box auditing procedure. Given access to the current predictor $\tilde{R}$ and a sample from $\mathcal{D}$, the auditor returns a set $S \in \mathcal{C}$ and a label $t \in \mathcal{T}$ such that the predictor violates the calibration or accuracy condition on (S, t) by at least $\varepsilon = \alpha$, if such a violation exists. If no such pair is found, the predictor is guaranteed to be $(\mathcal{C}, \alpha)$-multi-accurate. A similar approach applies to multi-calibration, where the auditor additionally returns a probability $\beta \in \Lambda[0, 1]$. In this framework, the overall time complexity is determined by the number of correction steps and the running time of the auditor.

In the case of coordinate-wise multi-accuracy and multi-calibration, the auditing step can be implemented via a weak agnostic learner. This follows from Theorem 4.2 in Hébert-Johnson et al. [2018], which shows how to achieve multi-calibration using weak agnostic learning, and thus, multi-accuracy as well. Although their setting is one-dimensional, our framework operates coordinate-wise rather than globally on the entire predictor. Consequently, the same argument applies, albeit with a multiplicative factor of k in the running time. Thus, under the assumption that the class $\mathcal{C}$ admits an efficient weak agnostic learner, the total running time of the learning algorithm can be expressed in terms of the time complexity of this learner, effectively removing the linear dependence on $|\mathcal{C}|$. Theorems C.6 and C.11 are stated in terms of weak agnostic learning. We refer the reader to Gopalan et al. [2024a] for a detailed discussion on how weak agnostic learning can be leveraged to implement efficient auditing procedures for multi-calibration and multi-accuracy.

H.2 Decision rules

We start with the proof of Claim C.14.

Claim H.9. *Let $m \in \mathbb{N}$, $\gamma_1, \ldots, \gamma_m \in [0, 1]$ s.t. $\sum_{i=1}^m \gamma_i = 1$, and $y_1, \ldots, y_m \in \mathcal{Y}$. Let $\varepsilon \in [0, 1]$ and assume that $f : \mathcal{Y} \to \{0, 1\}$ is ε-close to affine. Then,*

$$\left| \mathop{\mathbb{E}}_{f}\left[f\left(\sum_{i=1}^m \gamma_i y_i \right) \right] - \sum_{i=1}^m \gamma_i \mathop{\mathbb{E}}_{f}\left[f(y_i) \right] \right| \leq 2\varepsilon.$$

Proof. First, let us rewrite the second sum:

$$\sum_{i=1}^m \gamma_i \,\mathbb{E}[f(y_i)] = \gamma_m \,\mathbb{E}[f(y_m)] + \sum_{i=1}^{m-1} \gamma_i \,\mathbb{E}[f(y_i)] = \gamma_m \,\mathbb{E}[f(y_m)] + \left(\sum_{i=1}^{m-1} \gamma_i \right) \cdot \sum_{i=1}^{m-1} \frac{\gamma_i}{\sum_{i=1}^{m-1} \gamma_i} \,\mathbb{E}[f(y_i)]$$

$$= \gamma_m \,\mathbb{E}[f(y_m)] + (1 - \gamma_m) \cdot \sum_{i=1}^{m-1} \frac{\gamma_i}{1 - \gamma_m} \,\mathbb{E}[f(y_i)].$$

Since $(\gamma_i / 1 - \gamma_m)_{i \in [m-1]}$ sums to 1, the expression $\sum_{i=1}^{m-1} (\gamma_i / 1 - \gamma_m) \cdot y_i$ forms a convex combination of vectors in $\mathcal{Y}$. As $\mathcal{Y}$ is a convex set, this combination remains within $\mathcal{Y}$. Thus, the assumption from Definition C.13 applies, yielding that

$$\left| \sum_{i=1}^{m-1} \frac{\gamma_i}{1 - \gamma_m} \,\mathbb{E}[f(y_i)] - \mathbb{E}\left[f\left(\sum_{i=1}^{m-1} \frac{\gamma_i}{1 - \gamma_m} \cdot y_i \right) \right] \right| \leq \varepsilon.$$

This means that

$$\left|\mathbb{E}\left[f\left(\sum_{i=1}^{m}\gamma_i y_i\right)\right]-\sum_{i=1}^{m}\gamma_i\,\mathbb{E}[f(y_i)]\right|=$$
$$\left|\mathbb{E}\left[f\left(\sum_{i=1}^{m}\gamma_i y_i\right)\right]-\gamma_m\,\mathbb{E}[f(y_m)]-(1-\gamma_m)\cdot\sum_{i=1}^{m-1}\frac{\gamma_i}{1-\gamma_m}\,\mathbb{E}[f(y_i)]\right|\leq$$
$$(1-\gamma_m)\cdot\varepsilon+\left|\mathbb{E}\left[f\left(\sum_{i=1}^{m}\gamma_i y_i\right)\right]-\gamma_m\,\mathbb{E}[f(y_m)]-(1-\gamma_m)\cdot\mathbb{E}\left[f\left(\sum_{i=1}^{m-1}\frac{\gamma_i}{1-\gamma_m}\cdot y_i\right)\right]\right|. \tag{28}$$

Following the same arguments, we rewrite the first sum:

$$f\left(\sum_{i=1}^{m}\gamma_i y_i\right)=f\left(\gamma_m\cdot y_m+(1-\gamma_m)\cdot\sum_{i=1}^{m-1}\frac{\gamma_i}{1-\gamma_m}\cdot y_i\right),$$

and by the assumption,

$$\left|\mathbb{E}\left[f\left(\gamma_m\cdot y_m+(1-\gamma_m)\cdot\sum_{i=1}^{m-1}\frac{\gamma_i}{1-\gamma_m}\cdot y_i\right)\right]\right.$$
$$\left.-\gamma_m\cdot\mathbb{E}[f(y_m)]-(1-\gamma_m)\cdot\mathbb{E}\left[f\left(\sum_{i=1}^{m-1}\frac{\gamma_i}{1-\gamma_m}\cdot y_i\right)\right]\right|\leq\varepsilon.$$

Plugging in into Equation (28),

$$\left|\mathbb{E}\left[f\left(\sum_{i=1}^{m}\gamma_i y_i\right)\right]-\sum_{i=1}^{m}\gamma_i\,\mathbb{E}[f(y_i)]\right|$$
$$\leq(1-\gamma_m)\cdot\varepsilon+\varepsilon+\left|\gamma_m\cdot\mathbb{E}[f(y_m)]+(1-\gamma_m)\cdot\mathbb{E}\left[f\left(\sum_{i=1}^{m-1}\frac{\gamma_i}{1-\gamma_m}\cdot y_i\right)\right]\right.$$
$$\left.-\gamma_m\,\mathbb{E}[f(y_m)]-(1-\gamma_m)\cdot\mathbb{E}\left[f\left(\sum_{i=1}^{m-1}\frac{\gamma_i}{1-\gamma_m}\cdot y_i\right)\right]\right|$$
$$=(2-\gamma_m)\cdot\varepsilon\leq 2\varepsilon.$$

□

Next, we derive Corollary C.15.

Corollary H.10. *If $f:\mathcal{Y}\to\{0,1\}$ is ε-close to affine, then there exists an affine randomized predicate $f':\mathcal{Y}\to\{0,1\}$ such that $\forall y\in\mathcal{Y}$,*

$$|\mathop{\mathbb{E}}_{f}[f(y)]-\mathop{\mathbb{E}}_{f'}[f'(y)]|\leq 2\varepsilon.$$

Proof. Recall that e_i denotes the i^{th} unit vector in k dimensions. Any vector y can be expressed as $y=\sum_{i\in[k]}y_i\cdot e_i$, where y_i is its i^{th} coordinate. We define f' as follows:

$$f'(y)\sim\text{Ber}\left(\sum_{i\in[k]}y_i\cdot\mathbb{E}[f(e_i)]\right).$$

Clearly, f' is affine, and the corollary follows from Claim C.14. □

Finally, we provide the proof of Claim C.16.

Claim H.11. *Let $\varepsilon\in[0,1]$ and $M\in\mathbb{R}^+$. Let $f:\mathcal{Y}\to\{0,1\}$ be a randomized predicate such that:*

- *f is M-Lipschitz;*

- *f is ε-far from affine: that is, there exists $y, y' \in \mathcal{Y}$ and $\gamma \in (0,1)$ such that*

$$\left| \mathop{\mathbb{E}}_{f} \left[f\left(\gamma y + (1-\gamma) y'\right) \right] - \left(\gamma \mathop{\mathbb{E}}_{f} \left[f(y) \right] + (1-\gamma) \mathop{\mathbb{E}}_{f} \left[f(y') \right] \right) \right| > \varepsilon.$$

Then, $||y - y'||_\infty \geq \varepsilon / M$.

Proof. First, since f is M-Lipschitz,
$|\mathbb{E}\left[f\left(\gamma y + (1-\gamma)y'\right)\right] - \mathbb{E}\left[f(y)\right]| \leq M||\gamma y + (1-\gamma)y' - y|| = M(1-\gamma)||y - y'|| \leq M||y - y'||$,
and following the same argument,

$$|\mathbb{E}\left[f\left(\gamma y + (1-\gamma)y'\right)\right] - \mathbb{E}\left[f(y')\right]| \leq M||y - y'||.$$

This implies, by the assumption,

$$\begin{aligned}
\varepsilon &< |\mathbb{E}\left[f\left(\gamma y + (1-\gamma)y'\right)\right] - \left(\gamma \mathbb{E}\left[f(y)\right] + (1-\gamma)\mathbb{E}\left[f(y')\right]\right)| \\
&= |\gamma \left(\mathbb{E}\left[f\left(\gamma y + (1-\gamma)y'\right)\right] - \mathbb{E}\left[f(y)\right]\right) + (1-\gamma)\left(\mathbb{E}\left[f\left(\gamma y + (1-\gamma)y'\right)\right] - \mathbb{E}\left[f(y')\right]\right)| \\
&\leq \gamma M||y - y'|| + (1-\gamma)M||y - y'|| = M||y - y'||,
\end{aligned}$$

and the claim follows. □

H.3 Indistinguishability

First, we provide the proof of Lemma C.18.

Lemma H.12. *Let $\mathcal{X}$ be a domain, and let $\mathcal{U}_\mathcal{X}$ denote the uniform distribution over $\mathcal{X}$. Assume one-way functions exist w.r.t. security parameter $\kappa \geq \log|\mathcal{X}|$. For any constant $\gamma \in (0,1)$, there exists a subset $\mathcal{X}_1 \subset \mathcal{X}$ such that*

- *$|\Pr_{x \sim \mathcal{U}_\mathcal{X}}\left[x \in \mathcal{X}_1\right] - \gamma| = \mathsf{neg}(\kappa)$;*
- *$\mathcal{X}_1$ is κ-indistinguishable w.r.t. $\mathcal{U}_\mathcal{X}$;*
- *There exists a secret key such that, given the key, membership in $\mathcal{X}_1$ is efficiently computable.*

Proof. It is well-known in cryptography that the existence of one-way functions implies the existence of pseudorandom functions (PRFs) [Håstad et al., 1999, Goldreich et al., 1986]. In fact, we will show that weak-PRFs are sufficient for the construction, though the existence of OWFs remains a necessary assumption in both cases.

Recall that $\kappa \geq \log|\mathcal{X}| = n$, where n is the length of the Boolean string representing any $x \in \mathcal{X}$. Let $F_s : \{0,1\}^n \to \{0,1\}^\ell$ denote the $\mathsf{poly}(\kappa)$-size circuit computing F_s, where $\ell = \lceil \log(1/\gamma) + \kappa \rceil$ and $s \in \{0,1\}^\kappa$ denotes the secret key, which is hard-wired to the circuit. No efficient computation can access s or learn anything about it.

F_s defines membership in the subset $\mathcal{X}_1$ in the following manner: On input $x \sim \mathcal{U}_\mathcal{X}$, we interpret the output $F_s(x)$ as a real number in $[0,1]$. We define $x \in \mathcal{X}_1$ if and only if $F_s(x) \leq \gamma$.[14] By our choice of ℓ, the precision error of this length-ℓ binary string is at most $2^{-\ell}$, therefore

$$\left| \Pr_{x \sim \mathcal{U}_\mathcal{X}}\left[x \in \mathcal{X}_1\right] - \gamma \right| = \mathsf{neg}(\kappa).$$

The κ-indistinguishability of $\mathcal{X}_1$ follows from the pseudorandomness of F_s. Consider a probabilistic $\mathsf{poly}(\kappa)$-time algorithm $\mathcal{A}$ such that

$$\left| \Pr_{\mathcal{A}, x \sim \mathcal{U}_\mathcal{X}}\left[\mathcal{A}(x) = 1 \mid x \in \mathcal{X}_1\right] - \Pr_{\mathcal{A}, x \sim \mathcal{U}_\mathcal{X}}\left[\mathcal{A}(x) = 1 \mid x \in \mathcal{X} \setminus \mathcal{X}_1\right] \right| = \varepsilon(\kappa), \tag{29}$$

where $\varepsilon(\cdot)$ is a non-negligible function in κ. Recall that, since γ is a constant, $\Pr[x \in \mathcal{X}_1] \in (0,1)$, hence Equation (29) is well-defined. We reduce $\mathcal{A}$ to an adversary $\mathcal{B}$ for the pseudorandom function F_s. Conceptually, the adversary $\mathcal{B}$ has oracle access to an unknown function H, which can either be the PRF or a truly random function. The goal of $\mathcal{B}$ is to distinguish between the two, using its oracle access to $\mathcal{A}$.

[14] For convenience, we write $F_s(x) \leq \gamma$ to indicate that the real value in $[0,1]$ represented by the binary string $F_s(x)$ is at most γ.

- **Input:** oracle access to $\mathcal{A} : \{0,1\}^n \to \{0,1\}$, $H : \{0,1\}^n \to \{0,1\}^\ell$.
- $\mathcal{B}$ samples inputs $x_1, \ldots, x_N \sim \mathcal{U}_\mathcal{X}$ for $N = O\left(1/\varepsilon^2(\kappa)\right)$. Let $S = \{x_i\}_{i\in[N]}$.
- $\mathcal{B}$ makes N oracle calls to H and defines
$$S_1 = \{x_i \in S : H(x_i) \leq \gamma\}, \qquad S_0 = S \setminus S_1.$$
If either S_0 or S_1 is empty, $\mathcal{B}$ returns 0.
- $\mathcal{B}$ makes N oracle calls to $\mathcal{A}$ using fresh randomness, and computes
$$\hat{p}_1 := \frac{1}{|S_1|}\sum_{x_i\in S_1}\mathcal{A}(x_i), \qquad \hat{p}_0 := \frac{1}{|S_0|}\sum_{x_i\in S_0}\mathcal{A}(x_i).$$
$\mathcal{B}$ returns 1 if $|\hat{p}_1 - \hat{p}_0| > \varepsilon(\kappa)/2$, and 0 otherwise.

By construction, $\mathcal{B}$ is a probabilistic algorithm that runs in $\mathsf{poly}(\kappa)$ time. Moreover, observe that $\mathcal{B}$ queries its oracle H only on uniformly random inputs, which allows us to prove weak-PRF security as commented above. To demonstrate that $\mathcal{B}$ breaks the security of F_s, we proceed by showing that
$$\left|\Pr_{s,\mathcal{B},\mathcal{A}}\left[\mathcal{B}^{\mathcal{A},F_s}(1^\kappa) = 1\right] - \Pr_{G_\kappa,\mathcal{B},\mathcal{A}}\left[\mathcal{B}^{\mathcal{A},G_\kappa}(1^\kappa) = 1\right]\right| \geq \frac{1}{3},$$
where $G_\kappa : \{0,1\}^n \to \{0,1\}^\ell$ is a truly random function, namely, it is uniformly distributed over the set of all functions mapping $n(\kappa)$-bit-long strings to $\ell(\kappa)$-bit-long strings. Notice that, when $H = F_s$,
$$\begin{aligned}&|\Pr[\mathcal{A}(x_i) = 1 | x_i \in S_1] - \Pr[\mathcal{A}(x_i) = 1 | x_i \in S \setminus S_1]| = \\ &|\Pr[\mathcal{A}(x_i) = 1 | F_s(x_i) \leq \gamma] - \Pr[\mathcal{A}(x_i) = 1 | F_s(x_i) > \gamma]|.\end{aligned}$$
Hence, by a Chernoff bound, applying Equation (29) to the samples $x_i \sim \mathcal{U}_\mathcal{X}$ implies that, for a sufficiently large constant in the choice $N = O(1/\varepsilon^2(\kappa))$, with probability at least $2/3$,
$$|\hat{p}_1 - \hat{p}_0| > \frac{\varepsilon(\kappa)}{2}.$$
Therefore,
$$\Pr_{s,\mathcal{B},\mathcal{A}}\left[\mathcal{B}^{\mathcal{A},F_s}(1^\kappa) = 1\right] \geq \frac{2}{3}.$$
On the other hand, when $H = G_\kappa$, the event $G_\kappa(x) \leq \gamma$ is independent of x and of the output of $\mathcal{A}(x)$. Therefore,
$$\Pr[\mathcal{A}(x) = 1 \mid G_\kappa(x) \leq \gamma] = \Pr[\mathcal{A}(x) = 1 \mid G_\kappa(x) > \gamma].$$
Again by a Chernoff bound, with probability at least $2/3$,
$$|\hat{p}_1 - \hat{p}_0| \leq \frac{\varepsilon(\kappa)}{2}.$$
Thus,
$$\Pr_{G_\kappa,\mathcal{B},\mathcal{A}}\left[\mathcal{B}^{\mathcal{A},G_\kappa}(1^\kappa) = 1\right] \leq \frac{1}{3}.$$
All in all, the distinguishing advantage of $\mathcal{B}$ is at least $1/3$, which contradicts the assumed security of F_s.

□

Next, we restate Claim C.19 together with its proof.

Claim H.13. *Let $\mathcal{X}$ be a domain, and let $\mathcal{U}_\mathcal{X}$ denote the uniform distribution over $\mathcal{X}$. Assume $\mathcal{X}$ contains a κ-indistinguishable subset $\mathcal{X}_1$ as per Definition C.17 w.r.t. a security parameter $\kappa \geq \log|\mathcal{X}|$. For any $\mathsf{poly}(\kappa)$-time algorithm $\mathcal{A} : \mathcal{X} \to \{0,1\}$, if $\Pr[\mathcal{A}(x) = 1]$ is non-negligible in κ, then*
$$\left|\Pr_{\substack{x\sim\mathcal{U}_\mathcal{X}\\ \mathcal{A}\ randomness}}[x \in \mathcal{X}_1 | \mathcal{A}(x) = 1] - \Pr_{x\sim\mathcal{U}_\mathcal{X}}[x \in \mathcal{X}_1]\right| = \mathsf{neg}(\kappa).$$

Proof. Let $\mathcal{A} : \mathcal{X} \to \{0,1\}$ be a $\mathsf{poly}(\kappa)$-time algorithm and let $\mathcal{X}_2 := \mathcal{X} \setminus \mathcal{X}_1$. Let us denote

$$\tilde{\gamma} := \Pr_{\substack{x \sim \mathcal{U}_{\mathcal{X}} \\ \mathcal{A}\ randomness}} [x \in \mathcal{X}_1 \mid \mathcal{A}(x) = 1] \text{ and } \gamma' := \Pr_{x \sim \mathcal{U}_{\mathcal{X}}} [x \in \mathcal{X}_1] .$$

The advantage of $\mathcal{A}$ in distinguishing membership in $\mathcal{X}_1$ is:

$$\begin{aligned}
&|\Pr[\mathcal{A}(x) = 1 \mid x \in \mathcal{X}_1] - \Pr[\mathcal{A}(x) = 1 \mid x \in \mathcal{X}_2]| = \\
&\Pr[\mathcal{A}(x) = 1] \cdot \left| \frac{\Pr[x \in \mathcal{X}_1 \mid \mathcal{A}(x) = 1]}{\Pr[x \in \mathcal{X}_1]} - \frac{\Pr[x \in \mathcal{X}_2 \mid \mathcal{A}(x) = 1]}{\Pr[x \in \mathcal{X}_2]} \right| = \\
&\Pr[\mathcal{A}(x) = 1] \cdot \left| \frac{\tilde{\gamma}}{\gamma'} - \frac{1 - \tilde{\gamma}}{1 - \gamma'} \right| \geq 4 \Pr[\mathcal{A}(x) = 1] \cdot |\tilde{\gamma} - \gamma'| ,
\end{aligned}$$

where the last inequality follows from the fact that for all $\gamma' \in [0,1]$, we have $\gamma'(1 - \gamma') \leq \frac{1}{4}$. Recall that, by assumption, $\Pr[\mathcal{A}(x) = 1]$ is non-negligible in κ. Since the advantage of any efficient algorithm in recognizing membership in $\mathcal{X}_1$ is negligible, we conclude that $|\tilde{\gamma} - \gamma'| = \mathsf{neg}(\kappa)$. $\square$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction state the main contributions: new rate-based notions, computational impossibility results, and positive results showing that each desideratum can be achieved in isolation. The assumptions and scope are stated in the theorem overviews and formalized in the appendices.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The paper discusses the main limitations: the negative results rely on hidden subpopulations, the work is theoretical rather than empirical, and the implications for fairness are conceptual and do not by themselves prescribe a deployment procedure. We include an explicit discussion of these in the last section of the main body.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes].

   Justification: The paper is theoretical. The assumptions for the main results are stated in the theorem statements and fully formalized in the appendices. Complete proofs are provided for the positive and negative results, with proof overviews in the main text.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A].

   Justification: The paper does not include experiments; its contributions are theorem-level definitions, positive results, and computational lower bounds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A].

   Justification: The paper does not introduce experiments, datasets, or code artifacts.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A].

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A].

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A].

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work is theoretical and does not involve human subjects, datasets, released models, or deployment. The paper discusses fairness-motivated decision-making settings and does not introduce methods that enable misuse.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: The paper discusses implications for fairness in rich-label decision-making settings. The results show that natural subgroup-level protections—including calibration, loss minimization, and accurate decision/classification rates—may be computationally

incompatible in some settings, even though each is individually achievable and all are satisfied by the Bayes-optimal benchmark. This highlights a potential need to prioritize among fairness-relevant desiderata when designing downstream decision systems. The paper also discusses that rich labels may themselves encode human judgments or values, which introduces additional fairness considerations.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The paper does not release data, models, or systems with misuse risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A].

    Justification: The paper does not use existing datasets, code, models, or other external assets beyond cited prior work.

    Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A].

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing or human-subject research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not involve human-subject research.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: The core method development and theoretical results do not involve LLMs as a methodological component. Any use of LLMs was limited to editing and formatting and did not affect the scientific content.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.