WHY ARE BIASES NOT RARE IN AI SYSTEMS? THEORETICAL IMPOSSIBILITY AND TRADE-OFFS

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we establish a set of theoretical impossibility results, termed the *No-Free-Fairness* theorems, that identify three fundamental sources of disparity in learning systems. First, we show that when a task exhibits irreducible cost on a subgroup, any decision rule must trade off overall performance with disparity, yielding an inherent fairness–cost frontier. Second, we prove that even in ideal, noise-free settings where a perfectly fair and accurate solution exists, finite-sample learning alone induces nontrivial subgroup disparity, ruling out distribution-free fairness guarantees. More seriously, enforcing strict relative fairness creates a statistical bottleneck: achieving low cost may require exponentially many samples. Third, we show that limitations of the model class can independently induce disparity: if the model cannot represent accurate solutions for a subgroup, fairness remains unattainable regardless of data or training procedure. Overall, these results demonstrate that unfairness is not solely a consequence of biased data or imperfect algorithm, but arises from the intrinsic structure of decision problems, the constraints of finite data, and the expressivity of models. Our framework applies broadly beyond standard supervised learning, and suggests that achieving fairness requires explicit trade-offs and should be treated as a core design consideration.

1 INTRODUCTION

As machine learning systems are increasingly deployed in high-stakes domains, from credit underwriting to generative AI, concerns regarding algorithmic fairness have moved to the forefront of the field (Kearns et al., 2018; Buolamwini & Gebru, 2018; Sheng et al., 2021; D'Amour et al., 2022). While the community has proposed numerous fairness definitions and mitigation strategies (Dwork et al., 2012; Hardt et al., 2016; Liu et al., 2018; Hashimoto et al., 2018; Zhang, 2024; Cheng et al., 2025), a growing literature establishes trade-offs between fairness criteria and predictive performance (Kleinberg et al., 2017; Chouldechova, 2017; Pleiss et al., 2017). Most existing impossibility results, however, operate at the level of *metric incompatibility*: they show that under differing base rates, certain fairness notions cannot be satisfied simultaneously. This leaves open three questions:

Q1. *Does disparity emerge inevitably from the structural properties of statistical learning itself?*
Q2. *Can disparity always be resolved through improved data curation or algorithms?*
Q3. *Is there unavoidable trade-off in settings which admit perfectly fair and accurate solutions?*

Prior work provides partial answers to Q1 and Q2. Using absolute disparity, Cummings et al. (2019) show that one cannot simultaneously achieve exact fairness and non-trivial accuracy, under differential privacy constraint. Pinzón et al. (2022) further establish the existence of tasks for which no predictor is both fair and accurate, even without privacy constraints. However, these results are largely confined to classification settings and rely on existential constructions. As such, **it remains unclear whether these limitations extend to more general learning problems, and whether they reflect isolated pathological cases or more pervasive phenomena.**

In this work, we study fairness through the *risk ratio*, rather than the more commonly used absolute disparity (Cummings et al., 2019; Pinzón et al., 2022). While absolute disparity is analytically convenient and admits standard convergence guarantees (Woodworth et al., 2017; Agarwal et al., 2018; Laakom et al., 2025), it is inherently scale-dependent and can obscure disproportionate harm in low-risk regimes. In contrast, the risk ratio captures *relative* differences in outcomes, is invariant

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

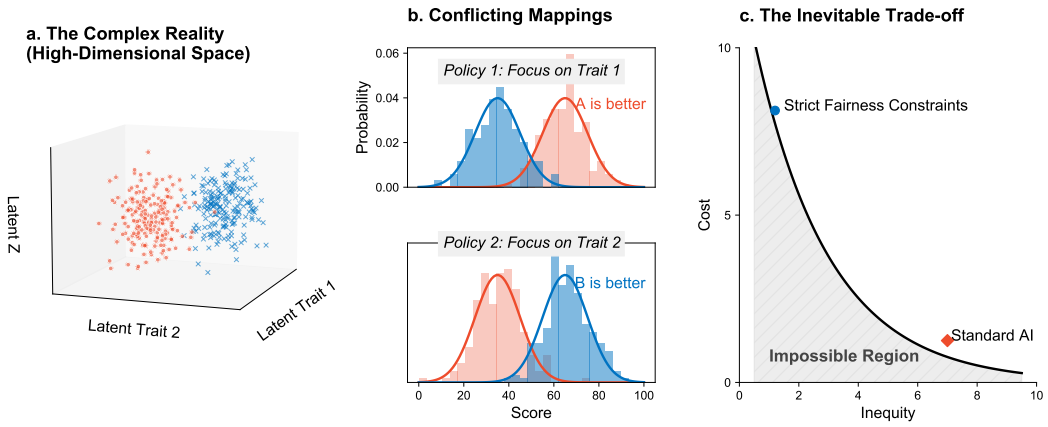


Figure 1: Inherent Fairness–Cost Trade-off. (a) Individuals (or agents) are represented as points in a high-dimensional latent space capturing task-relevant attributes (traits). (b) Each task, together with its associated decision rule or policy, induces a mapping from this latent space to lower-dimensional outcomes. We illustrate that for some valid tasks, this mapping necessarily collapses information unevenly across subpopulations, leading to differential costs. (c) As a consequence, achievable systems lie on a fairness–cost Pareto frontier, where regions corresponding to simultaneously perfect fairness and minimal cost are theoretically unattainable.

to the overall scale of risk, aligns with ratio-based regulatory standards, and remains meaningful under class imbalance and rare-event settings (see Appendix D).

We establish a unified set of **No-Free-Fairness** results that identify fundamental limits to achieving fairness in learning systems. Moving beyond metric incompatibility, we show that disparity can arise from the basic mechanics of learning itself, and that such effects occur under broad conditions. Our contributions isolate three distinct sources of unavoidable disparity:

- *No-Free-Fairness from Data (Task-Inherent Limits).* We show that when a problem is inherently uncertain for some subgroup, no decision rule can simultaneously achieve both high overall performance and equal outcomes across groups. In such settings, subgroup costs are intrinsically coupled: improving performance for one group can affect another when task difficulty is heterogeneous. In other words, when some cases are fundamentally harder than others, perfect fairness is impossible. As a result, all decision rules are constrained by a fundamental *fairness–cost trade-off*, forming a Pareto frontier as depicted in Figure 1.
- *No-Free-Fairness from Learning (Algorithmic and Statistical Limits).* Even in ideal settings which admit perfectly fair and accurate decision rules, we show that learning from finite data introduces unavoidable disparities. In particular, small or rare groups are harder to learn accurately, which can lead to uneven performance. Moreover, enforcing strict fairness can require dramatically (exponentially many) more data, creating a fundamental *tension between fairness and data efficiency*. Strictly enforcing near-perfect fairness can lead to non-vanishing overall risk, even with unlimited data.
- *No-Free-Fairness from Model Classes (Structural Limits).* We show that the choice of model itself can create fairness limitations. If a model lacks the flexibility to capture patterns specific to a subgroup, then unequal performance can persist even with unlimited, high-quality data. In this case, the limitation comes from how the model represents the problem, not from the data or the learning process.

The first and third results hold true for any non-negative cost function. Therefore, those results apply well to a wide range of contexts, e.g., regression, ranking, gererative tasks. Moreover, "uncertainty" is ubiquitous in practice (Northcutt et al., 2021; Jiang et al., 2018; Song et al., 2023; Koh et al., 2021). Taken together, our results show that fairness is not solely a consequence of biased data or imperfect algorithms. Rather, it is jointly determined by task structure, statistical estimation, and model expressivity. In particular, we provide strong evidence that (Q1) unfairness can emerge from intrinsic properties of learning, that (Q2) universal elimination of disparity through improved

data or algorithmic design alone is unattainable in general, and that (Q3) there is a fundamental tension between fairness and data efficiency, even in ideal settings where perfectly fair and accurate solutions exist. Achieving fairness therefore requires explicit trade-offs.

2 IMPOSSIBILITY RESULTS

This section presents our main impossibility results, touching on three axes: data distribution, algorithm, and hypothesis class. To facilitate detailed discussions, we need the following concepts.

2.1 PRELIMINARIES

Let $\mathcal{X}$ be a measurable space of individuals (or entities, states), and let $\mathcal{Z}$ denote a space of outcomes capturing all stochastic elements of the environment. A (possibly randomized) decision rule (or policy or hypothesis or model) is a measurable mapping $h : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y}$ can be a special case or subset of $\mathcal{Z}$.

Let $(x, Z) \sim P$ be a joint distribution over $\mathcal{X} \times \mathcal{Z}$. We define a non-negative cost function $\ell(h(x), Z)$ to measure an undesirable property of the rule h (e.g., loss, error, latency, or resource waste) for a specific instance (x, Z), and the population *risk* $a(h) = \mathbb{E}[\ell(h(x), Z)]$. For any measurable subset $\mathcal{X}_b \subseteq \mathcal{X}$ with $p_b = \mathbb{P}(x \in \mathcal{X}_b) \in (0, 1)$, define $a_b(h) = \mathbb{E}[\ell(h(x), Z) \mid x \in \mathcal{X}_b]$ and $a_{\bar{b}}(h) = \mathbb{E}[\ell(h(x), Z) \mid x \notin \mathcal{X}_b]$.

A *learner* (or decision-maker or learning algorithm) is a (possibly randomized) mapping $\mathcal{A}$ that takes as input a dataset $S \sim P^n$ and outputs a decision rule $h_S = \mathcal{A}(S)$. The *learning task* is to produce a rule with small population risk $a(h_S)$.

Instantiation of ℓ for common fairness notions. Many standard fairness metrics can be expressed as group-conditional risks $a_b(h)$ via appropriate choices of ℓ:

- *Error parity.* For binary classification with $Z = y \in \{0, 1\}$ and $h(x) \in \{0, 1\}$, choosing $\ell(h(x), Z) = \mathbf{I}\{h(x) \neq Z\}$ yields $a_b(h) = \mathbb{P}(h(x) \neq y \mid x \in \mathcal{X}_b)$, where $\mathbf{I}$ is the indicator function. So disparities correspond to differences (or ratios) in misclassification rates across groups.
- *Equalized odds.* Group-specific error rates are obtained by masking the loss: false negative rate (FNR) via $\ell(h(x), Z) = \mathbf{I}\{Z = 1\}\mathbf{I}\{h(x) = 0\}$, giving $a_b(h) = \mathbb{P}(h(x) = 0 \mid Z = 1, x \in \mathcal{X}_b)$; false positive rate (FPR) via $\ell(h(x), Z) = \mathbf{I}\{Z = 0\}\mathbf{I}\{h(x) = 1\}$, giving $a_b(h) = \mathbb{P}(h(x) = 1 \mid Z = 0, x \in \mathcal{X}_b)$.
- *Demographic parity.* This notion depends only on predictions and can be written with $\ell(h(x), Z) = \mathbf{I}\{h(x) = 1\}$ (independent of Z), yielding $a_b(h) = \mathbb{P}(h(x) = 1 \mid x \in \mathcal{X}_b)$.
- *Cost-based fairness.* In more general decision settings, let Z encode outcomes and define ℓ as an application-specific cost. Then $a_b(h)$ measures expected cost within group b, and fairness corresponds to controlling disparities in these costs.

These instantiations show that a broad class of group fairness criteria can be unified as disparities in conditional risks under suitable choices of ℓ.

Fairness in this work. We study fairness through the *risk ratio* $\varepsilon = a_{\bar{b}}(h)/a_b(h)$, rather than the more commonly used absolute disparity $\Delta(h) = |a_{\bar{b}}(h) - a_b(h)|$ (Woodworth et al., 2017; Agarwal et al., 2018; Laakom et al., 2025; Cummings et al., 2019; Pinzón et al., 2022). While the latter is analytically convenient and admits standard polynomial convergence guarantees, it is inherently scale-dependent and can obscure disproportionate harm in low-risk regimes. In particular, for fixed absolute gap $\Delta(h)$, the ratio satisfies $\varepsilon = 1 + \Delta(h)/a_b(h)$ (when $a_{\bar{b}}(h) \geq a_b(h)$), which becomes large as $a_b(h) \to 0$. In contrast, ε directly captures multiplicative disparity, is invariant to global rescaling of risks, aligns with ratio-based regulatory criteria, and remains meaningful under class imbalance and rare-event settings. We provide a detailed justification for this choice in Appendix D.

2.2 NO-FREE-FAIRNESS TASK

We first investigate the structural properties of a learning task. In practice, a sampling procedure can induce a data distribution P and hence determines the hardness of a learning task. The following theorem reveals that the task itself can inherently induce equity, whose proof is in Appendix A.

Theorem 2.1 (No-Free-Fairness Task)**.** *Given a non-negative cost ℓ, consider any distribution P over a data space $\mathcal{X} \times \mathcal{Z}$ and a measurable region $\mathcal{X}_b \subseteq \mathcal{X}$ with measure $p_b = \mathbb{P}(X \in \mathcal{X}_b) \in (0, 1)$. Let $a_b^* = \inf_h a_b(h)$, where the infimum is taken over all measurable decision rules h, and let $h^* \in \arg\min_h a(h)$ be a Bayes-optimal rule.*

1. *If $a_b^* > 0$, then any measurable decision rule h with $\varepsilon = \frac{a_{\bar{b}}(h)}{a_b(h)}$ satisfies*

$$a(h) \geq \big[p_b + (1 - p_b)\varepsilon\big] a_b^*. \tag{1}$$

2 (Price of Parity). *Any measurable rule h_{fair} with exact parity, $a_b(h_{\text{fair}}) = a_{\bar{b}}(h_{\text{fair}})$, satisfies*

$$a(h_{\text{fair}}) - a(h^*) \geq \max\left\{p_b\big[a_{\bar{b}}(h^*) - a_b(h^*)\big], (1 - p_b)\big[a_b(h^*) - a_{\bar{b}}(h^*)\big]\right\}. \tag{2}$$

Hence, exact parity incurs a strictly positive excess cost relative to h^* whenever $a_b(h^*) \neq a_{\bar{b}}(h^*)$.

This theorem applies to any decision problem in which a measurable rule h maps inputs to actions and incurs a non-negative cost. The condition $a_b^* > 0$ means that group b has a strictly positive irreducible cost: no measurable decision rule can make its conditional cost arbitrarily close to zero. This may arise from inherent uncertainty, overlapping outcome distributions, or stochastic responses in applications such as medical decision-making, resource allocation, and other decision problems under uncertainty.

Despite having a simple proof, Theorem 2.1 has several important implications:

- *Structural Lower Bound on Performance.* Theorem 2.1 shows that the overall cost is constrained jointly by the irreducible group cost a_b^*, the group mass p_b, and the disparity ratio ε. In particular, setting $\varepsilon = 1$ in equation 1 gives $a(h) \geq a_b^*$. Thus, no decision rule can attain exact parity with an overall cost below the smallest achievable cost for group b, even when this infimum is not attained by any measurable rule. Unlike impossibility results based on incompatibility among some fairness criteria (Kleinberg et al., 2017; Chouldechova, 2017; Pleiss et al., 2017), this limitation follows directly from the task.
- *A Symmetric Price of Parity.* equation 2 quantifies the performance loss required to remove disparity under a Bayes-optimal rule, regardless of which group has the larger Bayes cost. If $a_b(h^*) > a_{\bar{b}}(h^*)$, exact parity requires an excess cost of at least $(1 - p_b)[a_b(h^*) - a_{\bar{b}}(h^*)]$. Since the Bayes cost on group b cannot be reduced, parity must be attained by raising the cost on its complement. Conversely, if $a_{\bar{b}}(h^*) > a_b(h^*)$, the excess cost is at least $p_b[a_{\bar{b}}(h^*) - a_b(h^*)]$, and the cost on group b must be raised. Thus, whenever the Bayes-optimal group costs differ, exact parity necessarily entails leveling down on the group with the smaller Bayes cost, consistent with empirical observations of this phenomenon (Mittelstadt et al., 2023; Petersen et al., 2023).
- *Disparity Can Be Bayes Optimal.* The theorem shows that unequal group costs need not result from model misspecification, optimization failure, or the use of a suboptimal decision rule. The Bayes-optimal rule itself may satisfy $a_b(h^*) \neq a_{\bar{b}}(h^*)$ because the two groups have different irreducible costs. In this case, enforcing exact parity moves the decision rule away from Bayes optimality and may increase aggregate cost. Fairness and Bayes optimality are simultaneously attainable only if some Bayes-optimal rule is fair.
- *Task Dependence and Lack of Uniform Guarantees.* The bounds depend on a_b^* and the Bayes-optimal group costs $a_b(h^*)$ and $a_{\bar{b}}(h^*)$, all of which are properties of the underlying distribution rather than of a particular learning algorithm. Even when the input marginal P_X remains fixed, changes in the conditional distribution of Z given X can alter these quantities and, consequently, the attainable fairness–cost trade-off. A rule that attains a favorable trade-off in one environment may therefore fail to do so under another, reflecting the sensitivity of fairness guarantees to the underlying task (Koh et al., 2021).

Together, Theorem 2.1 identifies two related structural limitations. A positive irreducible cost a_b^* imposes a nonzero lower bound on the overall cost of parity, while unequal Bayes-optimal group costs imply a strictly positive excess cost relative to Bayes-optimal performance. These limitations are independent of the model, learning algorithm, or optimization procedure and are instead determined by the intrinsic difficulty and relative probability mass of the groups.

Remark 2.1 (On the practicality of the condition $a_b^* > 0$)**.** *The condition $a_b^* > 0$ corresponds to the standard notion of a* non-realizable *setting in learning theory: there is no decision rule that achieves*

zero population cost under P. Far from being restrictive, this is the typical regime in practice. In real-world problems, noise, ambiguity, unobserved confounders, and intrinsic stochasticity in the outcome mechanism (e.g., $P_{Y|X}$ in supervised learning) imply that even the optimal rule incurs nonzero cost on some subset $\mathcal{X}_b$, i.e., $a_b^* > 0$ (Zhang et al., 2021; Arora et al., 2019). Moreover, many application domains are inherently noisy or ill-defined—such as medical diagnosis, human annotation, or natural language understanding—where multiple plausible outcomes or measurement noise make perfect decisions unattainable (Northcutt et al., 2021; Jiang et al., 2018; Song et al., 2023; Petersen et al., 2023; Mittelstadt et al., 2023). Distribution shift (Koh et al., 2021) further reinforces this point: even if a problem appears realizable under one distribution, it may be non-realizable when the environment changes.

Therefore, $a_b^* > 0$ should be viewed as a mild and ubiquitous property of realistic decision problems. Importantly, it is precisely this irreducible cost that makes fairness nontrivial: if zero cost were achievable everywhere, disparities could be eliminated without trade-offs. Theorem 2.1 thus characterizes the practically relevant regime in which fairness–cost trade-offs are unavoidable.

2.3 No-Free-Fairness Algorithm

The previous result shows the scenarios, where the learning tasks may not be learnable. However, for some tractable tasks (where P is *realizable*, i.e., $a(h^*) = 0$), one may hope to find some fair solutions, by careful modeling and data curation. Unfortunately, it may not always be the case as revealed by the following result for binary classification, whose proof appears in Appendix B.

Theorem 2.2 (No-Free-Fairness Algorithm)**.** *Consider any measurable input space $\mathcal{X}$ which admits a non-atomic distribution, 0-1 loss ℓ, any real number $p_b \in (0,1)$, any $c \in (0,1]$. Denote $\mathcal{A}$ as a (possibly randomized) learner that maps each dataset S to a decision rule $\mathcal{A}(S)$, and $\varepsilon_S = \frac{a_{\bar{b}}(\mathcal{A}(S))}{\max\{a_b(\mathcal{A}(S)),c\}}$ as the cost ratio. Let $\mathcal{P}_{real}(p_b)$ be the class of all realizable probability distributions over $\mathcal{X} \times \{0,1\}$ that contain a subset $\mathcal{X}_b$ with marginal probability p_b. Then*

$$\inf_{\mathcal{A}} \sup_{P \in \mathcal{P}_{real}(p_b)} \mathbb{E}_{S\sim P^n}[\varepsilon_S] \geq \frac{1}{4}; \qquad \inf_{\mathcal{A}} \sup_{P \in \mathcal{P}_{real}(p_b)} \frac{\mathbb{E}_{S\sim P^n}[a(\mathcal{A}(S))]}{\mathbb{E}_{S\sim P^n}[\varepsilon_S]} \geq c(1-p_b). \tag{3}$$

Theorem 2.2 establishes minimax lower bounds on the (smoothed) cost ratio ε_S and its relation to overall cost. Since ε_S is asymmetric and may not directly exhibit deviation from parity, the following corollary makes the fairness–cost limitation explicit through the symmetric quantity $D_c(h_S) - 1$.

Corollary 2.1. *Denote $D_c(h) = \max\left\{\frac{\max\{a_{\bar{b}}(h),c\}}{\max\{a_b(h),c\}}, \frac{\max\{a_b(h),c\}}{\max\{a_{\bar{b}}(h),c\}}\right\}$, given the setting in Theorem 2.2. For any $c \in (0, 1/4)$,*

$$\inf_{\mathcal{A}} \sup_{P \in \mathcal{P}_{\mathrm{real}}(p_b)} \mathbb{E}_S\Big[[D_c(h_S) - 1] + \frac{1}{p_b} a(h_S)\Big] \geq \frac{1}{4} - c. \tag{4}$$

Various implications follow from Theorem 2.2 and Corollary 2.1:

- *Minimax Limits in the Realizable Regime.* Theorem 2.2 gives distribution-free minimax lower bounds even when the data-generating distribution is realizable and admits a rule with zero cost. Thus, no learner can uniformly control the directional cost ratio and overall cost over all distributions in $\mathcal{P}_{\mathrm{real}}(p_b)$. Unlike Theorem 2.1, where irreducible group cost creates a pointwise limitation, this result identifies a finite-sample limitation even when a perfectly accurate rule exists.
- *Explicit Fairness–Cost Trade-off.* Corollary 2.1 uses $D_c(h_S) - 1$, which is invariant to exchanging b and $\bar{b}$, to measure deviation from parity. Inequality (4) shows that no learner can uniformly make both the disparity and the overall cost small. In particular, decreasing $a(h_S)$ forces $D_c(h_S)$ to go away from 1 in the worst case, meaning more disparity.
- *Limitations of the "Clean Data" Narrative.* A common explanation attributes unfairness mainly to biased, noisy, or inconsistent data (Bommasani et al., 2021; Trewin et al., 2019; Buyl & De Bie, 2024; Calmon et al., 2017; Caton & Haas, 2024). Our results show that this explanation is incomplete: their lower bounds hold over realizable distributions for which a perfectly accurate rule exists. Hence, clean and consistent data alone do not guarantee that a finite-sample learner can attain both low overall cost and low disparity. The difficulty

arises because different rules may agree on the observed data while behaving differently across groups, a phenomenon related to underspecification (D'Amour et al., 2022).

- *Fairness as a No-Free-Lunch Phenomenon.* These results provide a fairness analogue of the No Free Lunch principle (Wolpert & Macready, 1997). Without additional assumptions on the distribution, target rule, or group structure, no learner can guarantee uniformly low overall cost and low disparity from finite data. Such guarantees therefore require suitable inductive biases, structural assumptions, representative sampling, or explicit fairness objectives. Together with Theorem 2.1, the results distinguish two sources of fairness limitations: finite-sample uncertainty when a perfectly fair and accurate rule exists, and irreducible group cost when it does not.

Remark 2.2 (Practical relevance of threshold c)**.** *We use c to define ε_S not only to avoid degeneracy when $a_b(h_S)$ of a model h_S is small, but also to encode a* minimum level of statistical relevance*. When $a_b(h_S) < c$, we have $\varepsilon_S = a_{\bar{b}}(h_S)/c$ which is strictly less than the true disparity $a_{\bar{b}}(h_S)/a_b(h_S)$; it is possible that $a_b(h_S)$ is negligible while $a_{\bar{b}}(h_S)$ remains large, in which case the disparity is substantial and ε_S remains informative.*

The practical relevance issue arises specifically in the regime where both $a_b(h_S)$ *and* $a_{\bar{b}}(h_S)$ *are small (e.g., below c). In this case, the ratio $a_{\bar{b}}(h_S)/a_b(h_S)$ can be dominated by negligible fluctuations, making large relative disparities practically uninformative. When $a_b(h_S) < c$, the metric effectively transitions to $\varepsilon_S \approx a_{\bar{b}}(h_S)/c$, emphasizing* absolute cost control *rather than relative disparity. Thus, c serves as a* relevance threshold*: above c, ε_S reflects relative fairness; below c, it deprioritizes ratio-based comparisons when both costs are already negligible.*

This relevance issue is not unique to ratio-based metrics. Difference-based measures such as $|a_{\bar{b}}(h_S) - a_b(h_S)|$ are inherently scale-insensitive*: a gap of the same magnitude is treated equally regardless of whether costs are large or near zero. Consequently, such metrics may overemphasize negligible absolute gaps in low-cost regimes, while ratio-based metrics without a floor may overstate disparities. From this perspective, the floor c provides a principled interpolation by enforcing a minimum scale below which relative comparisons are no longer considered meaningful.*

2.3.1 Fairness vs. Convergence Rates: A Statistical Tension

Theorem 2.2 introduces a regularization floor $c \in (0, 1]$ in the disparity metric $\varepsilon_S = a_{\bar{b}}(h_S)/\max\{a_b(h_S), c\}$, where $h_S = \mathcal{A}(S)$. If c is fixed, then there exist realizable distributions for which $\mathbb{E}[\varepsilon_S] = \Omega(1)$ while simultaneously $\mathbb{E}[a(h_S)] = \Omega(1)$. This shows that *distribution-free statistical consistency is incompatible with enforcing a relative fairness*. On the other hand, allowing $c = c(n)$ to depend on n reveals a fundamental statistical tension, as shown below.

Corollary 2.2. $\inf_{\mathcal{A}} \sup_{P \in \mathcal{P}_{\mathrm{real}}(p_b)} \mathbb{E}_{S\sim P^n}[a(\mathcal{A}(S))] \geq \frac{1-p_b}{4} c(n)$*, for every n and $c(n) \in (0, 1]$.*

This result implies that any distribution-free guarantee on the expected risk must satisfy $\sup_P \mathbb{E}[a(h_S)] = \Omega(c(n))$. Consequently, enforcing a relatively strict and slowly decaying floor (e.g., $c(n) = 1/\log n$) yields correspondingly slow convergence rates. In particular, achieving statistical consistency (i.e., $\mathbb{E}[a(h_S)] \to 0$) necessarily requires $c(n) \to 0$.

A Statistical Tension. The above analysis shows that $c(n)$ governs a fundamental trade-off. Bounding the maximum possible disparity requires a non-vanishing $c(n)$, but this induces a non-vanishing worst-case cost floor that strictly precludes statistical consistency. Conversely, achieving vanishing cost necessitates $c(n) \to 0$, under which ε_S increasingly reduces to raw disparity $\frac{a_{\bar{b}}(h_S)}{a_b(h_S)}$. Because the worst-case upper bound of this ratio scales as $1/c(n)$, allowing $c(n) \to 0$ leaves the disparity loosely constrained and mathematically permits severe unfairness. This reveals an inherent tension between distribution-free fairness guarantees and classical consistency.

A Bottleneck of Sample Complexity. Corollary 2.2 shows that $\sup_P \mathbb{E}_S[a(h_S)] = \Omega(c(n))$, revealing a fundamental statistical bottleneck specific to *relative* fairness. Enforcing a nontrivial lower bound $c(n)$ on subgroup costs (to control ratios) directly limits how fast the overall cost can decay. Concretely, the choice of $c(n)$ determines preference on sample complexity or equity:

- *Prioritizing fairness leads to exponential sample complexity.* If $c(n)$ decays slowly (e.g., $c(n) = 1/\log n$), then subgroup costs may not easily become too small, which tightly controls the disparity ε_S. However, achieving $\mathbb{E}_S[a(h_S)] \leq \nu$ now requires $c(n) \lesssim \nu$,

implying $n \geq e^{\Omega(1/\nu)}$. In this regime, the interaction between rare subgroups and strict ratio constraints converts the usual sublinear sample complexity into an exponential one.

- *Prioritizing convergence leads to vacuous fairness.* If $c(n)$ decays rapidly (e.g., $c(n) = 1/n$ or $1/n^2$), standard polynomial convergence rates are recovered. However, the worst-case disparity scales as $\varepsilon_S = \Omega(1/c(n))$, so $c(n) = \mathcal{O}(1/n^2)$ allows $\varepsilon_S = \Omega(n^2)$. In this case, fairness guarantees can become vacuous, as arbitrarily large disparities are permitted.

The Relevance Threshold. This dichotomy reveals a pathology of ratio-based fairness metrics in large-sample regimes. As $n \to \infty$, absolute costs $a_b(h_S)$ and $a_{\bar{b}}(h_S)$ both tend to 0. If they decay at slightly different rates (e.g., $1/n$ vs. $1/n^2$), their difference vanishes while their ratio ε_S diverges. The floor $c(n)$ acts as a *relevance threshold*: when $a_b(h_S) \gtrsim c(n)$, the ratio ε_S enforces meaningful relative comparisons; when $a_b(h_S) \ll c(n)$, the ratio is effectively disregarded. This formalizes the intuition that once costs are negligible across all groups, amplifying vanishing differences is not meaningful. Consequently, the choice of $c(n)$ encodes a principled transition between prioritizing strict equity and prioritizing overall utility.

2.4 No-Free-Fairness with Limited Model Classes

We have seen that disparity can arise from the task itself or from the learning procedure. We now examine a third factor: the role of model capacity and inductive bias. Similar with Theorem 2.1, we can easily show the following result (see Appendix C).

Theorem 2.3 (Limited Architecture)**.** *Given a non-negative cost ℓ, consider any distribution P over a data space $\mathcal{X} \times \mathcal{Z}$, and a nonempty class $\mathcal{H}$ of measurable decision rules. Denote $a^*_{b,\mathcal{H}} = \inf_{h\in\mathcal{H}} a_b(h)$ and $a^*_{\bar{b},\mathcal{H}} = \inf_{h\in\mathcal{H}} a_{\bar{b}}(h)$ for any given group $\mathcal{X}_b \subseteq \mathcal{X}$ with index b. If $a^*_{b,\mathcal{H}} > 0$ for some b with $p_b = \mathbb{P}(\mathcal{X}_b) \in (0,1)$, then any decision rule $h \in \mathcal{H}$ with $\varepsilon = \frac{a_{\bar{b}}(h)}{a_b(h)}$ satisfies $a(h) \geq \big[p_b + (1-p_b)\varepsilon\big] a^*_{b,\mathcal{H}}$. Furthermore, when $h^*_{\mathcal{H}} \in \arg\min_{h\in\mathcal{H}} a(h)$ exists, any decision rule $h_{\text{fair}} \in \mathcal{H}$ with exact parity, $a_b(h_{\text{fair}}) = a_{\bar{b}}(h_{\text{fair}})$, satisfies*

$$a(h_{\text{fair}}) - a(h^*_{\mathcal{H}}) \geq \max\left\{0, \max\left\{a^*_{b,\mathcal{H}}, a^*_{\bar{b},\mathcal{H}}\right\} - a(h^*_{\mathcal{H}})\right\}. \tag{5}$$

Theorem 2.3 identifies restrictions on the decision class as a source of fairness limitations. A restricted or misspecified class $\mathcal{H}$ may have $a^*_{b,\mathcal{H}} > 0$, meaning that no rule in $\mathcal{H}$ can achieve arbitrarily small cost on group b. In particular, exact parity implies $a(h) \geq a^*_{b,\mathcal{H}}$, while equation 5 gives a strictly positive price of parity whenever $\max\{a^*_{b,\mathcal{H}}, a^*_{\bar{b},\mathcal{H}}\} > a(h^*_{\mathcal{H}})$. The difference $a^*_{b,\mathcal{H}} - a^*_b \geq 0$, where a^*_b is the infimum over all measurable rules, further quantifies the additional group cost induced by restricting decisions to $\mathcal{H}$. This class-induced gap may reflect limited features, restrictive architectures, or unsuitable inductive biases. Moreover, because shared parameters can couple decisions across $\mathcal{X}_b$ and its complement, the groupwise infima $a^*_{b,\mathcal{H}}$ and $a^*_{\bar{b},\mathcal{H}}$ may be attained by different rules and need not be simultaneously achievable.

Together with Theorems 2.2 and 2.1, this result distinguishes three sources of fairness limitations: finite-sample uncertainty, intrinsic task difficulty, and restrictions on the decision class. These sources require different remedies: additional data can reduce estimation error, while richer model classes can reduce approximation error, but neither can eliminate irreducible group costs. For modern models, a positive gap may arise from limitations in the representation, architecture. Although the theorem does not imply that architecture causes every observed disparity, it motivates interventions that improve representations, allocate capacity more effectively, or enlarge the model class rather than relying exclusively on additional data or post-processing.

Connection to Modern Models. Empirical studies have documented demographic biases, representational harms, and disparities in model behavior across social groups in large language models (Bolukbasi et al., 2016; Sheng et al., 2021; Abid et al., 2021; Gehman et al., 2020; Bommasani et al., 2021; Gallegos et al., 2024). Theorem 2.3 suggests one possible structural explanation: the deployed model class, representation, or adaptation mechanism may impose a positive group-specific approximation floor. The theorem does not establish that observed disparities in a particular model are caused by architecture, since the quantities $a^*_{b,\mathcal{H}}$ and $a^*_{\bar{b},\mathcal{H}}$ also depend on the task. It nevertheless motivates some ways to reduce disparities, e.g., modifying representations, increasing relevant capacity, introducing subgroup-sensitive adaptation, or changing the model class.

3 COMPARISON WITH PRIOR IMPOSSIBILITY RESULTS.

The fundamental tension between predictive accuracy and algorithmic fairness has been extensively studied, yet existing theories isolate different (and often orthogonal) sources of this tension. Our framework unifies and sharpens these perspectives by identifying three distinct mechanisms: *task-inherent* (Theorem 2.1), *algorithm-inherent* (Theorem 2.2), and *architecture-inherent* (Theorem 2.3). Below, we contrast these mechanisms with prior work along several key dimensions.

Task-Inherent Bias, Algorithmic Cost, and Statistical Bottlenecks. Cummings et al. (2019) show that a good classifier cannot simultaneously achieve both differential privacy and exact fairness (equal opportunity), revealing a fundamental fairness–cost trade-off. Pinzón et al. (2022) further establish the existence of tasks for which no predictor is both accurate and exactly fair. In these works, the tension arises because fairness constraints force the learner to deviate from minimizing $a(h)$, leading to gaps of the form $a(h_{\text{fair}}) - a(h^*)$ that depend on the chosen constraint.

In contrast, our results show that a nontrivial trade-off exists *even before imposing any constraint*. In Theorems 2.1 and 2.3, the lower bound $a(h) \geq \varepsilon\, a_b^\dagger \big[1 - p_b \mathbf{I}(\varepsilon > 1)\big]$, where $a_b^\dagger \in \{a_b(h^*), a_b^*\}$, is intrinsic to the problem. Here $a_b(h^*)$ captures irreducible (task-inherent) cost, while $a_b^* = \min_{g \in \mathcal{H}} a_b(g)$ captures capacity-induced cost. Thus, the fairness–cost frontier is fundamentally constrained by *statistical or representational heterogeneity*, rather than by externally imposed constraints. More importantly, while Pinzón et al. (2022) establish task limitations through carefully constructed worst-case tasks, our Theorem 2.1 shows that such limitations arise naturally in realistic settings, as discussed in Remark 2.1.

Beyond these structural limits, our analysis also reveals a distinct *statistical bottleneck* for relative fairness. Corollary 2.2 shows that enforcing a nontrivial fairness floor $c(n)$ directly constrains the achievable convergence rate. In particular, maintaining strict relative fairness (slowly decaying $c(n)$) can force *exponential sample complexity*, while recovering classical rates (rapidly decaying $c(n)$) renders fairness guarantees vacuous. This phenomenon has no analogue in prior analyses.

Metric Incompatibility vs. Intrinsic Disparity. Foundational impossibility results (Kleinberg et al., 2017; Chouldechova, 2017; Pleiss et al., 2017) show that multiple fairness criteria (e.g., calibration and equalized odds) cannot be simultaneously satisfied when base rates differ: $P(Y = 1 \mid \mathcal{X}_b) \neq P(Y = 1 \mid \mathcal{X}_{\bar{b}})$. Many fairness metrics often rely on concepts of independence, separation, and sufficiency, which cannot hold *simultaneously* in non-trivial settings (Berk et al., 2021). These are *multi-metric incompatibility* results: impossibility arises from conflicting definitions.

Our framework instead considers a *single disparity measure*, $\varepsilon = a_{\bar{b}}(h)/a_b(h)$, and shows that even this minimal notion cannot be optimized jointly with accuracy. Hence, disparity persists *without invoking incompatible constraints*. Moreover, unlike Dwork et al. (2012), which requires a task-specific similarity metric to define individual fairness, our results operate purely at the level of expected risks, showing that even coarse group-level parity is fundamentally constrained. This demonstrates that unfairness is not merely a definitional artifact, but a consequence of heterogeneous error landscapes.

Beyond Classification: Loss-Agnostic Fairness Limits. A further distinction is that much of the existing theory is inherently tied to classification, where fairness notions are defined via confusion matrix quantities such as FPR and FNR (Hardt et al., 2016; Kleinberg et al., 2017; Chouldechova, 2017; Pleiss et al., 2017). These formulations do not easily extend to regression, ranking, or generative modeling.

In contrast, our results in Theorems 2.1 and 2.3 are expressed in terms of a general non-negative cost ℓ and subgroup risks $a_b(h)$, and therefore apply uniformly across learning settings. This shows that the fairness–accuracy tension is not a peculiarity of classification metrics, but a general property of risk minimization under heterogeneous populations. Consequently, our framework extends the scope of fairness impossibility results to a broader class of modern learning problems, including regression and generative modeling, where prior theory provides limited guidance.

Distributional Assumptions vs. Minimax Finite-Sample Limits. Many prior results rely on specific distributional conditions (e.g., differing base rates or noise levels). In contrast, Theorem 2.2 establishes a *distribution-free minimax bound* for $\mathbb{E}_S[\varepsilon_S]$. This holds even in realizable settings where $a_b(h^*) = a_{\bar{b}}(h^*) = 0$. Thus, unfairness can arise purely from finite-sample variability, with-

out noise or structural bias. Together with Theorem 2.1, this yields a unified picture: *disparity is unavoidable both in finite-sample regimes and in asymptotic regimes with irreducible error.*

Rates and Sample Complexity under Fairness Constraints. Beyond existence results (Cummings et al., 2019; Pinzón et al., 2022), our framework yields quantitative implications for statistical rates. Corollary 2.2 shows that any distribution-free guarantee must satisfy $\sup_P \mathbb{E}_{S\sim P^n}[a(h_S)] = \Omega(c(n))$, thereby constraining how the expected risk can decay as a function of the sample size n. In particular, achieving statistical consistency (i.e., $\mathbb{E}[a(h_S)] \to 0$ as $n \to \infty$) requires $c(n) \to 0$, which effectively relaxes the fairness constraint and permits larger disparity. Conversely, maintaining a non-vanishing, scale-stable notion of fairness (constant c) induces a non-vanishing worst-case error floor.

This result seems orthogonal to some prior studies (Agarwal et al., 2018; Donini et al., 2018; Woodworth et al., 2017; Cummings et al., 2019) which show polynomial convergence rates for the expected risk. The main reason for this difference comes from the use of absolute disparity such as $|a_b(h) - a(h)|$ in (Agarwal et al., 2018) or $|a_b(h) - a_{\bar{b}}(h)|$ in (Donini et al., 2018; Woodworth et al., 2017; Cummings et al., 2019). As discussed before, while absolute disparity is convenient for analysis, it is inherently scale-dependent and can obscure disproportionate harm in low-risk regimes. Therefore, despite providing various insights, prior guarantees with polynomial rates may not reflect the true cost when imposing fairness constraints in practice.

In contrast, the use of risk ratio in our framework can better align with ratio-based regulatory standards. Our results show that fairness constraints fundamentally limit the achievable *statistical rates* and sample complexity in a distribution-free setting. The limitation arises even with perfect empirical risk minimization, and is driven by statistical variability and subgroup structure.

Architecture and Representation vs. Data-Centric Views. Data-centric perspectives (Bender et al., 2021; Bommasani et al., 2021) attribute bias primarily to skewed or noisy datasets. Theorem 2.3 shows that even with ideal data, the hypothesis class can induce disparity. This connects to approximation theory: restricting $\mathcal{H}$ induces non-uniform approximation error across the population, which translates into fairness gaps. Hence, disparity can persist even under perfect data curation, unless the model class is sufficiently expressive. This complements empirical findings on representation bias and subgroup performance gaps, and suggests that *architectural capacity is a crucial determinant of fairness*.

Implications for Mitigation Strategies. Existing mitigation approaches, such as post-hoc thresholding (Hardt et al., 2016) or data repair (Feldman et al., 2015), operate within a fixed model class $\mathcal{H}$ and dataset. Our results show that such methods may not overcome structural limits imposed by $a_b^\dagger > 0$ or finite-sample bounds. Therefore, meaningful mitigation may require *changing the problem itself*: enriching features, increasing subgroup coverage, or expanding $\mathcal{H}$ to reduce a_b^*. This reframes fairness as a systems-level design problem rather than a purely algorithmic one.

4 Conclusion

We establish a unified set of *No-Free-Fairness* results that characterize fundamental limits of achieving fairness in learning systems along three axes. These results move beyond metric incompatibility and existential constructions, demonstrating that disparity can arise from general and non-pathological properties of tasks, learners, and models. In this sense, disparity is not merely a consequence of flawed data or misaligned optimization, but a structural feature of statistical learning that must be explicitly managed.

Our analysis focuses on relative fairness measured by the risk ratio and primarily studies population risks and worst-case guarantees, which may not capture all practically relevant notions of fairness or distributional structure encountered in real systems. While our results extend beyond standard classification settings, they do not fully address dynamic, strategic, or feedback-driven environments where decisions influence future data. Finally, although we argue that the identified phenomena arise under broad conditions, quantifying their exact severity in specific real-world applications remains an open empirical question. These limitations point to important directions for future work.

AI USE STATEMENT

During the preparation of this manuscript, we used generative AI tools to assist with language refinement, improve readability and exposition, strengthen the discussion of related work, enhance the presentation of mathematical proofs, and explore possible interpretations of our theoretical findings. All substantive ideas, analyses, proofs, and conclusions were developed and validated by the authors. We carefully reviewed and verified all AI-assisted outputs and remain fully responsible for the accuracy and integrity of the final manuscript.

ETHICS STATEMENT

This work studies fundamental limits of fairness in learning systems. Our results suggest that unfairness cannot always be eliminated through better optimization, larger datasets, or post-hoc correction alone, because disparities may arise from intrinsic uncertainty, finite-sample effects, or limitations of the model class itself. By formalizing these constraints, our work may help practitioners and policymakers develop more realistic expectations regarding the capabilities and limitations of AI systems deployed in high-stakes settings.

A potential positive impact of this work is to encourage more transparent evaluation of fairness trade-offs and more principled design of learning systems, particularly for underrepresented or difficult subpopulations. Our framework may also motivate future research on adaptive data collection, robust model design, and application-specific notions of equity.

At the same time, these impossibility results could be misinterpreted as suggesting that fairness efforts are futile. This is not the intent of our work. Rather, our results emphasize that fairness requires explicit assumptions, design choices, and trade-offs, and cannot generally be guaranteed in a distribution-free manner. Understanding these limitations is important for developing more reliable and accountable AI systems.

REFERENCES

Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. In *AAAI/ACM Conference on AI, Ethics, and Society*, pp. 298–306, 2021.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pp. 60–69. PMLR, 2018.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning (ICML)*, pp. 322–332, 2019.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on Fairness, Accountability, and Transparency*, pp. 610–623, 2021.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29, pp. 4349–4357, 2016.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang,

Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pp. 77–91. PMLR, 2018.

Maarten Buyl and Tijl De Bie. Inherent limitations of ai fairness. *Communications of the ACM*, 67 (2):48–55, 2024.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. *Advances in Neural Information Processing Systems*, 30, 2017.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56 (7), 2024. URL https://doi.org/10.1145/3616865.

Xiaoqing Cheng, Ruizhe Chen, Hongying Zan, Yuxiang Jia, and Min Peng. Biasfilter: An inference-time debiasing framework for large language models. In *Findings of the Association for Computational Linguistics: EMNLP*, 2025.

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

Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In *The 27th Conference on User Modeling, Adaptation and Personalization*, pp. 309–315, 2019.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

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

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP*, 2020.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1929–1938. PMLR, 2018.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1):65–75, 2025.

Wonje Jeung, Dongjae Jeon, Ashkan Yousefpour, and Jonghyun Choi. Large language models still exhibit bias in long text. In *Findings of the Association for Computational Linguistics: ACL*, 2025.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pp. 2304–2313, 2018.

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

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS)*, pp. 43–1, 2017.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard L Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Firas Laakom, Haobo Chen, Jürgen Schmidhuber, and Yuheng Bu. Fairness overfitting in machine learning: An information-theoretic perspective. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 32078–32115, 13–19 Jul 2025. URL `https://proceedings.mlr.press/v267/laakom25a.html`.

Chenxi Liu, Jianing Zheng, Yushu Liu, Xi Wang, Yuting Zhang, Qiang Fu, Wenwen Yu, Ting Yu, Wang Jiang, Dan Wang, et al. Potential to perpetuate social biases in health care by chinese large language models: a model evaluation study. *International Journal for Equity in Health*, 24(1): 206, 2025.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pp. 3150–3158. PMLR, 2018.

Brent Mittelstadt, Sandra Wachter, and Chris Russell. The unfairness of fair machine learning: Leveling down and strict egalitarianism by default. *Mich. Tech. L. Rev.*, 30:1, 2023.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

Judea Pearl. Interpretation and identification of causal mediation. *Psychological methods*, 19(4): 459, 2014.

Eike Petersen, Sune Holm, Melanie Ganz, and Aasa Feragen. The path toward equal performance in medical machine learning. *Patterns*, 4(7), 2023.

Carlos Pinzón, Catuscia Palamidessi, Pablo Piantanida, and Frank Valencia. On the impossibility of non-trivial accuracy in presence of fairness constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7993–8000, 2022.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in Neural Information Processing Systems*, 30, 2017.

Harishwar Reddy, Madhusudan Srinivasan, and Upulee Kanewala. Metamorphic testing for fairness evaluation in large language models: Identifying intersectional bias in llama and gpt. *arXiv preprint arXiv:2504.07982*, 2025.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. Societal biases in language generation: Progress and challenges. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

Hwanjun Song, Minseok Kim, Dongmin Park, Yong Shin, and Jae-Gil Lee. Learning with noisy labels: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, 2023.

Shari Trewin, Sara Basson, Michael Muller, Stacy Branham, Jutta Treviranus, Daniel Gruen, Daniel Hebert, Natalia Lyckowski, and Erich Manser. Considerations for ai fairness for people with disabilities. *AI Matters*, 5(3):4063, December 2019. URL `https://doi.org/10.1145/3362077.3362086`.

David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.

Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pp. 1920–1953. PMLR, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Wenbin Zhang. Ai fairness in practice: Paradigm, challenges, and prospects. *Ai Magazine*, 45(3): 386–395, 2024.

A Proof for No-free-fairness Task

Proof of Theorem 2.1. First, we observe that $a_b(h) \geq a_b^* > 0$. Therefore, $\varepsilon = \frac{a_{\bar{b}}(h)}{a_b(h)}$ is well-defined. Observe that

$$\begin{aligned} a(h) &= p_b a_b(h) + (1-p_b) a_{\bar{b}}(h) && (6)\\ &= p_b a_b(h) + (1-p_b)\varepsilon a_b(h) && (7)\\ &= \big[p_b + (1-p_b)\varepsilon\big] a_b(h) && (8)\\ &\geq \big[p_b + (1-p_b)\varepsilon\big] a_b^*. && (9) \end{aligned}$$

This proves inequality (1).

We next consider the Bayes-optimal rule. Note that h^* minimizes both $a_b(h)$ and $a_{\bar{b}}(h)$ over all measurable decision rules.

Enforcing exact parity gives

$$\begin{aligned} a(h_{\text{fair}}) &= p_b a_b(h_{\text{fair}}) + (1-p_b) a_{\bar{b}}(h_{\text{fair}}) && (10)\\ &= p_b a_b(h_{\text{fair}}) + (1-p_b) a_b(h_{\text{fair}}) && (11)\\ &= a_b(h_{\text{fair}}). && (12) \end{aligned}$$

Since h^* has the smallest risk for group b, we have $a_b(h_{\text{fair}}) \geq a_b(h^*)$. Hence,

$$\begin{aligned} a(h_{\text{fair}}) - a(h^*) &\geq a_b(h^*) - a(h^*) && (13)\\ &= a_b(h^*) - \big[p_b a_b(h^*) + (1-p_b) a_{\bar{b}}(h^*)\big] && (14)\\ &= (1-p_b)\big[a_b(h^*) - a_{\bar{b}}(h^*)\big]. && (15) \end{aligned}$$

Similarly, exact parity implies $a(h_{\text{fair}}) = a_{\bar{b}}(h_{\text{fair}})$. Since h^* has the smallest risk for group $\bar{b}$, we have $a_{\bar{b}}(h_{\text{fair}}) \geq a_{\bar{b}}(h^*)$. Therefore,

$$\begin{aligned} a(h_{\text{fair}}) - a(h^*) &\geq a_{\bar{b}}(h^*) - a(h^*) && (16)\\ &= a_{\bar{b}}(h^*) - \big[p_b a_b(h^*) + (1-p_b) a_{\bar{b}}(h^*)\big] && (17)\\ &= p_b\big[a_{\bar{b}}(h^*) - a_b(h^*)\big]. && (18) \end{aligned}$$

Taking the maximum of (15) and (18) leads to (2), completing the proof. □

B Proof for No-free-fairness algorithm

Theorem 2.2 is a corollary of the following result.

Theorem B.1. *Consider any measurable input space $\mathcal{X}$ which admits a non-atomic distribution, any (possibly randomized) learning algorithm $\mathcal{A}$ that maps each dataset S to $h_S : \mathcal{X} \to \{0,1\}$, 0-1 loss, any real number $p_b \in (0,1)$, any $c \in (0,1]$ that may depend on n and p_b.*

Then there exist a realizable distribution P over $\mathcal{X} \times \{0,1\}$ and a measurable subset $\mathcal{X}_b \subseteq \mathcal{X}$ with $\mathbb{P}(\mathcal{X}_b) = p_b$ such that:

$$\begin{aligned} \mathbb{E}_S[\varepsilon_S] &\geq 1/4, && (19)\\ \mathbb{E}_S[a(h_S)] &\geq c(1-p_b)\mathbb{E}_S[\varepsilon_S], && (20) \end{aligned}$$

where $\varepsilon_S = a_{\bar{b}}(h_S)/\max\{a_b(h_S), c\}$.

Proof. Choose any integer $m \geq 2n$ where n is the number of samples in S. The proof contains three main steps.

Step 1: Partition of $\mathcal{X}$ and define P. By the assumption of $\mathcal{X}$, there exist a partition $\mathcal{X} = \mathcal{X}_b \cup \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_m$ of $\mathcal{X}$ into disjoint measurable sets, and a marginal distribution P_X over $\mathcal{X}$ so that

$$P_X(\mathcal{X}_b) = p_b, \qquad P_X(\mathcal{X}_i) = \frac{1-p_b}{m}, \quad i \in [m]. \tag{21}$$

Draw $\theta = (\theta_1, \ldots, \theta_m) \sim \mathrm{Unif}(\{0,1\}^m)$. Define $P_{Y|X}$ by choosing

$$Y = \begin{cases} 0 & \text{if } x \in \mathcal{X}_b, \\ \theta_i & \text{if } x \in \mathcal{X}_i. \end{cases} \tag{22}$$

Let $P = P_X \times P_{Y|X}$. This distribution is realizable by $h^*(x) = 0$ on $\mathcal{X}_b$ and $h^*(x) = \theta_i$ on $\mathcal{X}_i$, hence $a(h^*) = 0$.

Step 2: Lower bound on $a_{\bar{b}}(h_S)$ and ε_S. Let $I_S = \{i \in [m] : \mathcal{X}_i \cap S \neq \emptyset\}$ and $U_S = [m] \setminus I_S$. Then $|I_S| \leq n$, hence $|U_S| \geq m - n \geq m/2$.

For any $i \in U_S$, the variable θ_i is independent of S, hence

$$\mathbb{E}_\theta[\mathbf{I}\{h_S(x) \neq Y\} \mid x \in \mathcal{X}_i, S] = \tfrac{1}{2}. \tag{23}$$

Therefore,

$$\begin{aligned} \mathbb{E}_{S,\theta}[a_{\bar{b}}(h_S)] &= \mathbb{E}_S\left[\sum_{i=1}^m \frac{1}{m}\mathbb{E}_\theta[\mathbf{I}\{h_S(x) \neq Y\} \mid x \in \mathcal{X}_i, S]\right] && (24) \\ &\geq \mathbb{E}_S\left[\sum_{i \in U_S} \frac{1}{m} \cdot \frac{1}{2}\right] && (25) \\ &= \mathbb{E}_S\left[\frac{|U_S|}{2m}\right] && (26) \\ &\geq \frac{1}{4}. && (27) \end{aligned}$$

Since the expected value of the risk over the random choice of θ satisfies the lower bound, there must exist at least one specific realization of θ (and thus a specific realizable distribution P) for which the following holds:

$$\mathbb{E}_S[a_{\bar{b}}(h_S)] \geq 1/4. \tag{28}$$

Since $a_b(h_S) \leq 1$ and $c \leq 1$, we have $\max\{a_b(h_S), c\} \leq 1$, which implies $\varepsilon_S \geq a_{\bar{b}}(h_S)$. As a result, we have

$$\mathbb{E}_S[\varepsilon_S] \geq \mathbb{E}_S[a_{\bar{b}}(h_S)] \geq 1/4. \tag{29}$$

Step 3: Bounding $a(h_S)$. Note that

$$a(h_S) = p_b a_b(h_S) + (1 - p_b) a_{\bar{b}}(h_S) \geq (1 - p_b) a_{\bar{b}}(h_S). \tag{30}$$

We next consider the following cases.

Case 1: $a_b(h_S) \geq c$. Observe that $\varepsilon_S = \frac{a_{\bar{b}}(h_S)}{a_b(h_S)} \leq \frac{a_{\bar{b}}(h_S)}{c}$. So $a_{\bar{b}}(h_S) \geq c\varepsilon_S$.

Case 2: $a_b(h_S) < c$. Then $\varepsilon_S = \frac{a_{\bar{b}}(h_S)}{c}$, and hence $a_{\bar{b}}(h_S) = c\varepsilon_S$.

Both cases show that $a_{\bar{b}}(h_S) \geq c\varepsilon_S$. Combining this with (30), we obtain $a(h_S) \geq c(1 - p_b)\varepsilon_S$. This holds for every θ. Taking expectation over S yields

$$\mathbb{E}_S[a(h_S)] \geq c(1 - p_b)\mathbb{E}_S[\varepsilon_S], \tag{31}$$

completing the proof. □

Proof of Corollary 2.2. For any (possibly randomized) learning algorithm $\mathcal{A}$, Theorem B.1 shows the existence of a realizable distribution P satisfying $\mathbb{E}_S[a(h_S)] \geq c(n)(1 - p_b)\mathbb{E}_S[\varepsilon_S]$ and $\mathbb{E}_S[\varepsilon_S] \geq 1/4$. Hence, $\mathbb{E}_S[a(h_S)] \geq c(n)(1 - p_b)/4$. This suggests that

$$\sup_{P \in \mathcal{P}_{\mathrm{real}}(p_b)} \mathbb{E}_S[a(h_S)] \geq c(n)(1 - p_b)/4.$$

Taking $\inf_{\mathcal{A}}$ for both sides completes the proof. □

Proof of Corollary 2.1. For any (possibly randomized) learning algorithm $\mathcal{A}$, using the same distribution P as constructed in the proof of Theorem B.1, we obtain (28) as

$$\mathbb{E}_S[a_{\bar{b}}(h_S)] \geq \frac{1}{4}.$$

Furthermore, $a(h_S) = p_b a_b(h_S) + (1-p_b)a_{\bar{b}}(h_S) \geq p_b a_b(h_S)$. This implies $\mathbb{E}_S[a_b(h_S)] \leq \frac{1}{p_b}\mathbb{E}_S[a(h_S)]$. Therefore,

$$\mathbb{E}_S[|a_{\bar{b}}(h_S) - a_b(h_S)|] \geq \frac{1}{4} - \frac{1}{p_b}\mathbb{E}_S[a(h_S)].$$

Finally, since $D_c(h_S) - 1 \geq |a_{\bar{b}}(h_S) - a_b(h_S)| - c$ as shown in Lemma B.2 for any $c \in (0, 1/4)$, we have $\mathbb{E}_S[D_c(h_S) - 1] \geq \frac{1}{4} - c - \frac{1}{p_b}\mathbb{E}_S[a(h_S)]$. In other words,

$$\mathbb{E}_S[D_c(h_S) - 1] + \frac{1}{p_b}\mathbb{E}_S[a(h_S)] \geq \frac{1}{4} - c.$$

Since $P \in \mathcal{P}_{\mathrm{real}}(p_b)$, it follows that

$$\sup_{P \in \mathcal{P}_{\mathrm{real}}(p_b)} \left[\mathbb{E}_S[D_c(h_S) - 1] + \frac{1}{p_b}\mathbb{E}_S[a(h_S)]\right] \geq \frac{1}{4} - c.$$

Since this holds for every possibly randomized learning algorithm $\mathcal{A}$, taking the infimum over $\mathcal{A}$ proves (4). □

Lemma B.2. *$D_c - 1 \geq |x - y| - c$, for any $x, y \in [0,1]$ and $c \in (0,1]$, where $D_c = \max\{\frac{\max\{x,c\}}{\max\{y,c\}}, \frac{\max\{y,c\}}{\max\{x,c\}}\}$.*

Proof. Let $u = \max\{x, c\}$ and $v = \max\{y, c\}$. Then $u, v \in [c, 1]$, so

$$D_c - 1 = \max\{\frac{u}{v}, \frac{v}{u}\} - 1 = \frac{|u - v|}{\min\{u, v\}} \geq |u - v|.$$

We now lower bound $|u - v|$ by cases:

If $x, y \geq c$, then $u = x$ and $v = y$, hence $D_c - 1 \geq |x - y| \geq |x - y| - c$.

If $x, y < c$, then $u = v = c$, so $D_c - 1 = 0$. Since $|x - y| \leq c$, we get $0 \geq |x - y| - c$.

If $x < c \leq y$, then $u = c$ and $v = y$, so $D_c - 1 \geq y - c \geq y - x - c = |x - y| - c$, where we used $x \geq 0$. The case $y < c \leq x$ is symmetric.

Therefore, in all cases, we have $D_c - 1 \geq |x - y| - c$. This proves the lemma. □

C Proof for No-free-fairness Architecture

Proof of Theorem 2.3. For every $h \in \mathcal{H}$, we have $a_b(h) \geq a^*_{b,\mathcal{H}} > 0$, so $\varepsilon = \frac{a_{\bar{b}}(h)}{a_b(h)}$ is well-defined. Observe that

$$\begin{aligned} a(h) &= p_b a_b(h) + (1-p_b)a_{\bar{b}}(h) && (32)\\ &= p_b a_b(h) + (1-p_b)\varepsilon a_b(h) && (33)\\ &= \big[p_b + (1-p_b)\varepsilon\big] a_b(h) && (34)\\ &\geq \big[p_b + (1-p_b)\varepsilon\big] a^*_{b,\mathcal{H}}. && (35) \end{aligned}$$

This proves the first result.

If $\varepsilon \leq 1$, then $p_b + (1-p_b)\varepsilon = \varepsilon + p_b(1-\varepsilon) \geq \varepsilon$. If $\varepsilon > 1$, then $p_b + (1-p_b)\varepsilon \geq (1-p_b)\varepsilon$. Combining these two cases gives $a(h) \geq \varepsilon a^*_{b,\mathcal{H}}\big[1 - p_b \mathbf{I}\{\varepsilon > 1\}\big]$.

We next prove inequality (5). Let $r = a_b(h_{\text{fair}}) = a_{\bar{b}}(h_{\text{fair}})$. Exact parity gives

$$\begin{aligned} a(h_{\text{fair}}) &= p_b a_b(h_{\text{fair}}) + (1-p_b) a_{\bar{b}}(h_{\text{fair}}) && (36)\\ &= p_b r + (1-p_b) r && (37)\\ &= r. && (38) \end{aligned}$$

Since $h_{\text{fair}} \in \mathcal{H}$, we have $r \geq a^*_{b,\mathcal{H}}$ and $r \geq a^*_{\bar{b},\mathcal{H}}$. Therefore,

$$a(h_{\text{fair}}) - a(h^*_{\mathcal{H}}) \geq \max\left\{a^*_{b,\mathcal{H}}, a^*_{\bar{b},\mathcal{H}}\right\} - a(h^*_{\mathcal{H}}). \tag{39}$$

Moreover, since $h^*_{\mathcal{H}}$ minimizes the overall cost over $\mathcal{H}$, we have

$$a(h_{\text{fair}}) - a(h^*_{\mathcal{H}}) \geq 0. \tag{40}$$

Taking the maximum of (39) and (40) proves (5). This concludes the proof. □

D JUSTIFICATION FOR THE RISK RATIO AS A FAIRNESS METRIC

Throughout this work, we formalize relative fairness via the *risk ratio* $\mathrm{RR}(h) = a_{\bar{b}}(h)/a_b(h)$ rather than the mathematically friendlier absolute difference $\Delta(h) = |a_{\bar{b}}(h) - a_b(h)|$. Although $\Delta(h)$ admits standard polynomial convergence rates in realizable PAC regimes (Woodworth et al., 2017; Agarwal et al., 2018; Laakom et al., 2025), it is intrinsically *scale-dependent* and can be statistically and operationally misleading. In contrast, ratio-based notions encode *proportional harm*, which is the quantity of interest in many high-stakes settings. Below we articulate the theoretical and practical imperatives for adopting $\mathrm{RR}(h)$, thereby motivating why the statistical bottlenecks identified in our main results are fundamental.

- **Scale sensitivity in low-risk regimes.** Absolute difference ignores the baseline magnitude of risk. For instance, let's fix the same $\Delta(h) = 0.01$. If $a_b(h) = 0.51$ and $a_{\bar{b}}(h) = 0.50$, then $\mathrm{RR}(h) \approx 1.02$, indicating negligible disparity. In contrast, if $a_b(h) = 0.02$ and $a_{\bar{b}}(h) = 0.01$, then $\mathrm{RR}(h) = 2$, i.e., a 100% multiplicative increase in risk. More generally, for $a_{\bar{b}}(h) \geq a_b(h)$ we have $\mathrm{RR}(h) = 1 + \Delta(h)/a_b(h)$, which diverges as $a_b(h) \to 0$. Hence, absolute difference $\Delta(h)$ systematically under-penalizes disparities in low-incidence regimes.
- **Sensitivity under multiplicative improvements.** Suppose model updates induce multiplicative improvements $a_b(h') = \gamma_b a_b(h)$ and $a_{\bar{b}}(h') = \gamma_{\bar{b}} a_{\bar{b}}(h)$ with $\gamma_b, \gamma_{\bar{b}} \in (0,1)$. Then $\mathrm{RR}(h')/\mathrm{RR}(h) = \gamma_{\bar{b}}/\gamma_b$. Even when both groups improve, relative disparity worsens whenever $\gamma_{\bar{b}} > \gamma_b$. In contrast, $\Delta(h') = \gamma_{\bar{b}} a_{\bar{b}}(h) - \gamma_b a_b(h)$ may shrink purely due to global error reduction, masking a strict widening of proportional inequity.
- **Compatibility with ratio-based constraints in regulation.** Many regulatory standards are explicitly ratio-based. For instance, the U.S. Equal Employment Opportunity Commission's (EEOC) "Four-Fifths Rule" requires the ratio of selection rates to be at least 0.8. Such constraints are naturally expressed as lower bounds on a ratio. Optimizing $\Delta(h)$ provides no guarantee of satisfying constraints of the form $\mathrm{RR}(h) \leq \tau$ or $\mathrm{RR}(h) \geq \tau$, whereas directly controlling $\mathrm{RR}(h)$ yields immediate compliance guarantees (cf. (Feldman et al., 2015; Hardt et al., 2016)).
- **Alignment with epidemiological and risk-analysis standards.** The ratio $\mathrm{RR}(h)$ coincides with *relative risk*, the canonical measure in epidemiology for quantifying disproportionate harm. This ensures that fairness assessments in ML are directly interpretable within established scientific frameworks (cf. (Pearl, 2014)).
- **Scale invariance and cross-domain interpretability.** The quantity $\Delta(h)$ is tied to the absolute scale of the task and lacks a canonical interpretation across domains. In contrast, $\mathrm{RR}(h)$ is dimensionless: $\mathrm{RR}(h) = 1.5$ uniformly encodes a 50% excess risk regardless of whether $a_b(h)$ is 10^{-3} or 0.3. This invariance is essential for transferring fairness thresholds across heterogeneous applications.
- **Statistical behavior under class imbalance.** When $p_b \ll 1$, accurate estimation of $a_b(h)$ requires $n \gg 1/p_b$, and errors in estimating $a_b(h)$ can dominate $\Delta(h)$. For the ratio, the relevant quantity is multiplicative error, e.g., $|\widehat{a}_{\bar{b}}(h)/\widehat{a}_b(h) - a_{\bar{b}}(h)/a_b(h)|$, which directly reflects proportional uncertainty. Our lower bounds show that controlling such quantities necessarily incurs sample complexity scaling with $1/\min\{a_b(h), a_{\bar{b}}(h)\}$, revealing an intrinsic dependence on rare-event regimes.

- **Incompatibility of absolute control with proportional guarantees.** Without assumptions on $a_b(h)$, there is no distribution-free mapping from a bound $\Delta(h) \leq \nu$ to a meaningful bound on $\mathrm{RR}(h)$. Indeed, $\mathrm{RR}(h) \leq 1 + \nu/a_b(h)$, which becomes vacuous as $a_b(h) \to 0$. This means prior results on absolute disparity do not directly apply to ratio-based fairness. Therefore, guaranteeing proportional fairness necessitates direct control of ratio-type quantities, which in turn induces the sharper statistical trade-offs established in our main results.

In summary, absolute disparity is analytically convenient but misaligned with proportional notions of harm. In contrast, $\mathrm{RR}(h)$ is scale-invariant, regulation-compatible, and statistically faithful to rare-event regimes. Consequently, the stringent sample complexity requirements and trade-offs identified in our results are unavoidable: they arise from the intrinsic difficulty of estimating and controlling multiplicative disparities.