Metric-Fair Classifier Derandomization

Jimmy Wu* Yatong Chen† Yang Liu‡

June 20, 2022

Abstract

We study the problem of *classifier derandomization* in machine learning: given a stochastic binary classifier $f : X \to [0, 1]$, sample a deterministic classifier $\hat{f} : X \to \{0, 1\}$ that approximates the output of f in aggregate over any data distribution. Recent work revealed how to efficiently derandomize a stochastic classifier with strong output approximation guarantees, but at the cost of individual fairness — that is, if f treated similar inputs similarly, $\hat{f}$ did not. In this paper, we initiate a systematic study of classifier derandomization with metric fairness guarantees. We show that the prior derandomization approach is almost maximally metric-unfair, and that a simple "random threshold" derandomization achieves optimal fairness preservation but with weaker output approximation. We then devise a derandomization procedure that provides an appealing tradeoff between these two: if f is α-metric fair according to a metric d with a locality-sensitive hash (LSH) family, then our derandomized $\hat{f}$ is, with high probability, $O(\alpha)$-metric fair and a close approximation of f. We also prove generic results applicable to all (fair and unfair) classifier derandomization procedures, including a bias-variance decomposition and reductions between various notions of metric fairness.

1 Introduction

We study the general problem of *derandomizing* stochastic classification models. Consider a typical binary classification setting defined by a feature space $X \subseteq \mathbb{R}^n$ and labels $\{0, 1\}$; we wish to devise a procedure that, given a *stochastic* or *randomized* classifier $f : X \to [0, 1]$, efficiently samples a *deterministic* classifier $\hat{f} : X \to \{0, 1\}$ from some family of functions $\mathcal{F}$, such that $\hat{f}$ preserves various qualities of f.

Stochastic classifiers arise naturally in both theory and practice. For example, they are frequently the solutions to constrained optimization problems encoding complex evaluation metrics [Nar18], group fairness [GZG+17; ABD+18], individual fairness [DHP+12; RY18; KRR18; SKR19], and robustness to adversarial attacks [PMA+19; CRK19; PER+20; BG20]. Stochastic classifiers are also the natural result of taking an ensemble of individual classifiers [Die00; GZG+17].

However, they may be undesirable for numerous reasons: a stochastic classifier is not robust to repeated attacks, since even one that is instance-wise 99% accurate will likely err after a few hundred attempts; by the same token, they violate intuitive notions of fairness since even the *same* individual may be treated differently over multiple classifications. For these reasons, Cotter, Gupta, and Narasimhan [CGN19] recently presented a procedure for derandomizing a stochastic classifier while approximately preserving the outputs of f with high probability. However, the authors observe that their construction results in similar individuals typically being given very different predictions — in other words, it does not satisfy *individual fairness* — and ask whether it is possible to obtain a family of deterministic classifiers that preserves both aggregate outputs and individual fairness.

Another motivation for studying individually fair decision making comes from the game-theoretic setting of *strategic classification*, wherein decision subjects may modify their features to obtain a desired outcome from the classifier [HMP+16; CDP15; CPP+18; DRS+18; CLP20]. A metric-fair stochastic classifier — and by extension, a metric-fair derandomization procedure — offers significant protection against such manipulations. See Appendix B for more on this topic.

*Email: jimmywu126@gmail.com
†University of California, Santa Cruz, Department of Computer Science and Engineering. Email: ychen592@ucsc.edu
‡University of California, Santa Cruz, Department of Computer Science and Engineering. Email: yangliu@ucsc.edu

1.1 Our Contributions

In this paper, we initiate a systematic study of classifier derandomization with individual fairness preservation. In line with many recent works, we formalize individual fairness as *metric fairness*, which requires the classifier to output similar predictions on close point pairs in some metric space (X, d) [DHP+12; KRR18; FSV16]. Roughly, f is metric-fair if there are constants $\alpha, \beta > 0$ such that for all $x, x' \in X$,

$$|f(x) - f(x')| \leq \alpha \cdot d(x, x') + \beta$$

A sampled deterministic classifier $\hat{f} \sim \mathcal{F}$ is metric-fair when this inequality holds in expectation.

Under this formalism, we obtain the following results:

1. We make precise the observation of [CGN19] that their derandomization procedure, based on pairwise-independent hash functions, does not preserve individual fairness. In fact, we prove that it is almost *maximally* metric-unfair regardless of how fair the original stochastic classifier was (Section 2.1).
2. We demonstrate that a very simple derandomization procedure, based on setting a single random threshold $r \sim [0, 1]$, attains near-perfect expected fairness preservation, and prove that no better fairness preservation is possible (Section 2.2). However, this procedure's output approximation has higher variance than the pairwise-independent hashing approach in general.
3. We devise a derandomization procedure that achieves nearly the best of both worlds, preserving aggregate outputs with high probability, with only modest loss of metric fairness (Section 3). In particular, when f has fairness parameters (α, β), sampling $\hat{f}$ from our family $\mathcal{F}_{\mathsf{LS}}$ yields expected fairness parameters at most $(\alpha + \frac{1}{2}, \beta + \epsilon)$. We also show a high-probability aggregate fairness guarantee: *most* deterministic classifiers in $\mathcal{F}$ assign *most* close pairs the same prediction. These guarantees hold for the class of metrics d that possess locality-sensitive hashing (LSH) schemes, which includes a wide variety of generic and data-dependent metrics.
4. We prove structural lemmas applicable to all classifier derandomization procedures: first, a bias-variance decomposition for the error of a derandomization $\hat{f}$ of f; second, a set of reductions showing that metric fairness-preserving derandomizations also preserve notions of *aggregate* and *threshold* fairness.

A practically appealing aspect of our LSH-based derandomization method is that it is completely oblivious to the original stochastic classifier, in that it requires no knowledge of how f was trained, and its fairness guarantee holds for whatever fairness parameters f happens to satisfy on each pair $(x, x') \in X^2$. The technique can therefore be applied as an independent post-processing step — for example, on the many fair stochastic classifiers detailed in recent works [RY18; KRR18]. The burden on the model designer is thus reduced to selecting an LSHable metric feature space (X, d) that is appropriate for the classification task.

1.2 Preliminaries

Given a stochastic classifier $f : X \to [0, 1]$ and distance function $d : X \times X \to [0, 1]$, we wish to design an efficiently sampleable set $\mathcal{F}$ of deterministic binary classifiers $\hat{f} : X \to \{0, 1\}$; we call $\mathcal{F}$ a *family of deterministic classifiers*, or a *derandomization of* f. Moreover, we would like $\mathcal{F}$ to have the following properties:

Output approximation: $\hat{f}$ sampled uniformly[1] from $\mathcal{F}$ simulates or approximates f in aggregate over any distribution. More precisely, define the *pointwise* bias and variance of $\hat{f}$ with respect to f on a sample $x \in X$ as

$$\text{bias}(\hat{f}, f, x) := \mathop{\mathbb{E}}_{\hat{f} \sim \mathcal{F}} \left[\hat{f}(x) \right] - f(x) \qquad \text{and} \qquad \text{variance}(\hat{f}, x) := \mathop{\text{Var}}_{\hat{f} \sim \mathcal{F}} \left(\hat{f}(x) \right)$$

Now let $\mathcal{D}$ be a distribution over X. The *aggregate* bias and variance of $\hat{f}$ with respect to f on $\mathcal{D}$ are

$$\text{bias}(\hat{f}, f, \mathcal{D}) := \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\text{bias}(\hat{f}, f, x) \right] \qquad \text{and} \qquad \text{variance}(\hat{f}, \mathcal{D}) := \mathop{\text{Var}}_{\hat{f} \sim \mathcal{F}} \left(\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\hat{f}(x) \right] \right)$$

[1] In this paper, we will always sample uniformly from families of classifiers and hash functions; thus $\hat{f} \sim \mathcal{F}$ means $\hat{f} \sim \text{Unif}(\mathcal{F})$, and $h \sim \mathcal{H}$ means $h \sim \text{Unif}(\mathcal{H})$.

We seek a family $\mathcal{F}$ for which both of these quantities are small. This is a useful notion of a good approximation of f since in practice, classifiers are typically applied *in aggregate* on some dataset or in deployment. In Section 4.4 we also point out that low bias and variance in the above sense implies that $\hat{f}$ and f are nearly indistinguishable when compared according to any binary loss functions, such as accuracy, false positive rate, etc.

Individual fairness: Similar individuals are likely to be treated similarly. We formally define this notion as *metric fairness*, which says that that the classifier should be an approximately Lipschitz-continuous function relative to a given distance metric:

Definition 1.1 ((α, β, d)-metric fairness)**.** *Let $\alpha \geq 1$[2] and $\beta \geq 0$, let $d : X^2 \to [0, 1]$ be a metric, and let $x, x' \in X$. We say a stochastic classifier $f : X \to [0, 1]$ satisfies (α, β, d)-*metric fairness *on (x, x'), or is (α, β, d)-*fair *on (x, x'), if*

$$|f(x) - f(x')| \leq \alpha \cdot d(x, x') + \beta \tag{1}$$

Similarly, a deterministic classifier family $\mathcal{F}$ is (α, β, d)-fair on (x, x') if

$$\mathop{\mathbb{E}}_{\hat{f} \sim \mathcal{F}} \left[\left| \hat{f}(x) - \hat{f}(x') \right| \right] \leq \alpha \cdot d(x, x') + \beta \tag{2}$$

*When this condition is satisfied for all $(x, x') \in X^2$, we simply say the classifier (or family) is (α, β, d)-*fair*.*

To intuit this definition, notice that when a classifier satisfies metric fairness with $\beta = 0$, the difference between its predictions on some pair of points x and x' scales in proportion to their distance. To conform to this idea of fairness, it is important that the derandomization procedures we design do not substantially increase these fairness parameters, but especially β.

The above definition of metric fairness is most closely related to those of Rothblum and Yona [RY18], whose focus is learning a "probably approximately metric-fair" model that generalizes to unseen data; and Kim, Reingold, and Rothblum [KRR18], whose focus is in-sample learning when the metric d is not fully specified. Both works take inspiration from the metric-based notion of individual fairness introduced in [DHP+12]. Crucially however, the aforementioned works provide guarantees exclusively for stochastic classifiers, and to our knowledge, this is the case for all papers to date whose focus is learning metric-fair classifiers.

In addition to this pairwise notion of metric fairness, we will also develop *aggregate* fairness guarantees for various derandomization procedures. To that end, let $X^2_{\leq \tau} := \left\{ (x, x') \in X^2 \;\middle|\; d(x, x') \leq \tau \right\}$ denote the set of point pairs within some distance $\tau \in [0, 1]$. Our aggregate fairness bounds will state that, with high probability over the sampling of $\hat{f} \sim \mathcal{F}$, most pairs $(x, x') \in X^2_{\leq \tau}$ receive the same prediction from $\hat{f}$.

2 Output Approximation Versus Fairness

We begin our study of metric-fair classifier derandomization by contrasting two approaches: first, the "pairwise-independent" derandomization of [CGN19], which achieves a low-variance approximation of the original stochastic classifier, but does not preserve metric fairness; and second, a simple "random threshold" derandomization that perfectly preserves metric fairness, at the cost of higher output variance.

2.1 Pairwise-Independent Derandomization

The construction of Cotter, Narasimhan, and Gupta [CGN19] makes use of a pairwise-independent hash function family $\mathcal{H}_{\mathsf{PI}}$, i.e. a set of functions $h_{\mathsf{PI}} : B \to [k]$ such that

$$\Pr_{h \sim \mathcal{H}_{\mathsf{PI}}} [h(b) = i, h(b') = j] = \frac{1}{k^2} \quad \forall b \neq b' \in B, \quad i, j \in [k]$$

Observe that a family that satisfies this property is also uniform, i.e. $\Pr_{h \sim \mathcal{H}_{\mathsf{PI}}}[h(b) = i] = 1/k$ for all b, i.

[2] We enforce $\alpha \geq 1$, and not merely $\alpha \geq 0$, so that the codomain of f is $[0, 1]$ rather than potentially $[0, \alpha]$ (or some other interval of length $\alpha < 1$). Requiring $\alpha \geq 1$ thus makes f a proper stochastic classifier and enables direct comparisons between different fairness parameters. This is no loss of generality since (α, β, d)-fairness for $\alpha < 1$ can also be expressed as $(1, \frac{\beta}{\alpha}, \frac{d}{\alpha})$-fairness or, with some loss of generality, $(1, \beta + \alpha, d)$-fairness.

The classifier family they propose is then[3]

$$\mathcal{F}_{\mathsf{PI}} := \left\{ \hat{f}_{h_{\mathsf{PI}}} \;\middle|\; h_{\mathsf{PI}} \in \mathcal{H}_{\mathsf{PI}} \right\}, \quad \text{where} \quad \hat{f}_{h_{\mathsf{PI}}}(x) := \mathbb{1}\left\{ f(x) \geq \frac{h_{\mathsf{PI}}(\pi(x))}{k} \right\} \tag{3}$$

where $\pi : X \to B$ is some fixed *bucketing* function that discretizes the input (since the pairwise-independent hash family has finite domain).

Let us develop some intuition for this construction. First, thinking of k as large, each $\hat{f}_{h_{\mathsf{PI}}} \in \mathcal{F}_{\mathsf{PI}}$ essentially assigns a pseudo-random threshold $\frac{h_{\mathsf{PI}}(\pi(x))}{k} \in [0,1]$ to each input x, so that $\hat{f}(x) = 1$ if and only if $f(x)$ exceeds the threshold. Since h_{PI} is a uniform hash function family, $h_{\mathsf{PI}}(\pi(x))$ is uniform over $[k]$; this endows $\mathcal{F}_{\mathsf{PI}}$ with low bias with respect to f. Using this idea and the pairwise-independence of $\mathcal{H}_{\mathsf{PI}}$, the authors show that this classifier family exhibits low bias and variance of approximation:

Theorem 2.1 (Bias and variance of pairwise-independent derandomization [CGN19] (simplified)). *Let f be a stochastic classifier, $\mathcal{D}$ a distribution over X, and $\pi : X \to B$ a bucketing function. Then $\hat{f} \sim \mathcal{F}_{\mathsf{PI}}$ satisfies*

$$\mathrm{bias}(\hat{f}_{\mathsf{PI}}, f, \mathcal{D}) \leq \frac{1}{k} \qquad \text{and} \qquad \mathrm{variance}(\hat{f}_{\mathsf{PI}}, f, \mathcal{D}) \leq \max_{b \in B} \Pr_{x \sim \mathcal{D}}[\pi(x) = b] \cdot \mathop{\mathbb{E}}_{x \sim \mathcal{D}}[f(x)(1 - f(x))] + \frac{1}{k}$$

Moreover, $\hat{f}_{\mathsf{PI}}$ can be sampled using $O(\log |B| + \log k)$ uniform random bits.

To understand this variance bound, observe that for a given data distribution $\mathcal{D}$, the bound is stronger or weaker depending on how well π disperses samples into different buckets in B. When there exists some $b \in B$ such that $\Pr_{x \sim \mathcal{D}}[\pi(x) = b] \approx 1$, $\mathrm{variance}(\hat{f}_{\mathsf{PI}}, f, \mathcal{D}) \approx \mathbb{E}_{x \sim \mathcal{D}}[f(x)(1 - f(x))]$ essentially tracks the stochasticity of f. At the other extreme when $\Pr_{x \sim \mathcal{D}}[\pi(x) = b] = 1/|B|$ for all $b \in B$, $\mathrm{variance}(\hat{f}_{\mathsf{PI}}, f, \mathcal{D}) \approx 1/|B|$.

As the authors pointed out (but did not formalize), $\hat{f}_{\mathsf{PI}}$ does not preserve pairwise fairness in general. We make this observation precise by showing that it is always possible to design a dataset, of any desired size, such that the pairwise-independent derandomization treats *every* pair of points unfairly for nearly any $\beta < 1/2$.

Proposition 2.2 (Unfairness of pairwise-independent derandomization). *For every $N \geq 2$, $\alpha \geq 1$, $\beta < \frac{1}{2} - \frac{1}{2k}$, and metric $d : \mathbb{R}^n \times \mathbb{R}^n \to [0,1]$, there exist a set $X \subset \mathbb{R}^n$ of size N and stochastic classifier $f : X \to [0,1]$ such that the following hold:*

1. *f is nontrivial and $(1, 0, d)$-fair.*
2. *$\mathcal{F}_{\mathsf{PI}}$ violates (α, β, d)-metric fairness for* every *pair $(x, x') \in X^2, x \neq x'$.*

If k is not too small, this says that derandomizing using pairwise-independent hashing is almost maximally unfair, as a uniform random binary function $\hat{g} : X \to \{0,1\}$ satisfies $\mathbb{E}[|\hat{g}(x) - \hat{g}(x')|] = 1/2$, and therefore achieves $\beta = 1/2$.

Proof sketch of Proposition 2.2. Consider any $\alpha \geq 1$, $\beta \in \left(0, \frac{1}{2} - \frac{1}{2k}\right)$, and $N \geq 2$. We choose X to be some set of N points on a sufficiently small sphere about the origin, and let f be a classifier that maps half of the points in X to $\frac{1+\epsilon}{2}$ and the other half to $\frac{1-\epsilon}{2}$. When $\epsilon > 0$ is sufficiently small, it can be shown that f is $(1, 0, d)$-fair over X. However, $\mathcal{F}_{\mathsf{PI}}$ is not (α, β, d)-fair on any point pair $(x, x') \in X^2$. The reason is that since f is almost maximally stochastic (i.e. $f(x) \approx 1/2$ for all x), and $\mathcal{H}_{\mathsf{PI}}$ is pairwise-independent, the binary outputs $\hat{f}(x)$ and $\hat{f}(x')$ are about as likely to be the same as they are likely to be different. Hence $\mathbb{E}_{\hat{f} \sim \mathcal{F}_{\mathsf{PI}}}[|\hat{f}(x) - \hat{f}(x')|] \approx 1/2$, violating (α, β, d)-metric fairness. See Appendix A.1 for the full proof. □

2.2 Random Threshold Classifier

It turns out that there is a near-trivial derandomization that achieves optimal preservation of metric fairness, namely the following *random threshold* classifier family:

$$\mathcal{F}_{\mathsf{RT}} := \{\hat{f}_r \mid r \in [0,1]\}, \text{ where } \hat{f}_r := \mathbb{1}\{f(x) \geq r\} \tag{4}$$

Formally we make the following observation, whose proof is in Appendix A.2.

[3] For the sake of clearer exposition, we simplify the deterministic classifier used in [CGN19], which is actually $\hat{f}_{h_{\mathsf{PI}}}(x) := \mathbb{1}\{f(x) \geq \frac{2h_{\mathsf{PI}}(x)-1}{2k}\}$; this does not change Theorem 2.1 or Proposition 2.2 beyond a $1/2k$ additive difference in the bias, variance, and β.

Proposition 2.3 (Random threshold derandomization guarantees). *Let f be an (α, β, d)-fair stochastic classifier and $\mathcal{D}$ a distribution over X. Then the deterministic classifier family $\mathcal{F}_{\mathsf{RT}}$ is also (α, β, d)-fair. Moreover,*

$$\text{bias}(\hat{f}_{\mathsf{RT}}, f, \mathcal{D}) = 0 \qquad \text{and} \qquad \text{variance}(\hat{f}_{\mathsf{RT}}, f, \mathcal{D}) \leq \mathop{\mathbb{E}}_{x \sim \mathcal{D}}[f(x)(1 - f(x))]$$

Note that while this derandomization preserves the original fairness parameters perfectly, its variance can be substantially higher than that of $\mathcal{F}_{\mathsf{PI}}$ depending on the choice of bucketing function π in Equation (3).

One subtlety here is that $\mathcal{F}_{\mathsf{RT}}$ is an infinite set, and is therefore not sampleable in practice. For the more realistic scenario in which the threshold r is a number of some fixed precision $\epsilon > 0$, the statements in Proposition 2.3 hold up to additive error ϵ, and $\hat{f}_{\mathsf{RT}}$ can be sampled using $O(\log(1/\epsilon))$ uniform random bits. In this case $\mathcal{F}_{\mathsf{RT}}$ is $(\alpha, \beta + \epsilon, d)$-fair, and as we can show, this is in fact necessary:

Proposition 2.4 ($(\alpha, 0, d)$-metric fairness is impossible for finite deterministic families). *Let $d : X \times X \to [0, 1]$ be a metric over a convex set $X \subseteq \mathbb{R}^n$, and let $\mathcal{F}$ be a finite family of deterministic classifiers, at least one of which is nontrivial. Then for every $\alpha \geq 1$ and $\beta < 1/|\mathcal{F}|$, $\mathcal{F}$ is not (α, β, d)-fair.*

Proof sketch. Since $\mathcal{F}$ contains a nontrivial classifier $\hat{f}$, we can pick sufficiently close points around a discontinuity of $\hat{f}$ and show that in expectation, $\mathcal{F}$ fails to achieve roughly $(\alpha, 1/|\mathcal{F}|, d)$-fairness on this point pair. See Appendix A.3 for details. □

The main consequence is that there is an irreducible amount of additive unfairness $\beta > 0$ that cannot be avoided when constructing a fair deterministic classifier family. Indeed, the derandomization $\mathcal{F}$ we present in Section 3 has $|\mathcal{F}| \geq 1/\beta$, thus avoiding the impossible regime indicated by Proposition 2.4.

3 Fair Derandomization via Locality-Sensitive Hashing

In this section, we construct a deterministic classifier family that combines much of the appeal of both the pairwise-independent derandomization (low output variance) and the random threshold derandomization (strong fairness preservation). This new approach utilizes two types of hashing: first, a pairwise-independent hash family $\mathcal{H}_{\mathsf{PI}}$ as before; and second, a locality-sensitive hash family:[4]

Definition 3.1 (Locality-sensitive hash (LSH) family). *Let X be a set of hashable items, B a set of buckets, and $d : X^2 \to [0, 1]$ a metric distance function. We say a set $\mathcal{H}_{\mathsf{LS}}$ of functions $h : X \to B$ is a* locality-sensitive family of hash functions *for d if for all $x, x' \in X$,*

$$\Pr_{h \sim \mathcal{H}_{\mathsf{LS}}}[h(x) \neq h(x')] = d(x, x')$$

Locality-sensitive hashing is a well-studied technique, and LSH families have been constructed for many standard distances and similarities, such as L_1 [IM98], L_2 [AI06], cosine [Cha02], Jaccard [Bro97], various data-dependent metrics [JKG08; AIN+14; AR15], and more.

Our derandomization works as follows: suppose $f : X \to [0, 1]$ is a stochastic classifier, $\mathcal{H}_{\mathsf{LS}}$ is a family of locality-sensitive hash functions $h_{\mathsf{LS}} : X \to B$, and $\mathcal{H}_{\mathsf{PI}}$ is a family of pairwise-independent hash functions $h_{\mathsf{PI}} : B \to [k]$ for some positive integer k. Our family of deterministic classifiers is then

$$\mathcal{F}_{\mathsf{LS}} := \left\{ \hat{f}_{h_{\mathsf{LS}}, h_{\mathsf{PI}}} \;\middle|\; h_{\mathsf{LS}} \in \mathcal{H}_{\mathsf{LS}}, h_{\mathsf{PI}} \in \mathcal{H}_{\mathsf{PI}} \right\}, \quad \text{where} \quad \hat{f}_{h_{\mathsf{LS}}, h_{\mathsf{PI}}}(x) := \mathbb{1}\left\{ f(x) \geq \frac{h_{\mathsf{PI}}(h_{\mathsf{LS}}(x))}{k} \right\}. \tag{5}$$

Let us develop some intuition for this construction. First, thinking of k as large, each $\hat{f} \in \mathcal{F}_{\mathsf{LS}}$ essentially assigns a pseudo-random threshold $\frac{h_{\mathsf{PI}}(h_{\mathsf{LS}}(x))}{k} \in [0, 1]$ to each input x, so that $\hat{f}(x) = 1$ if and only if $f(x)$ exceeds the threshold. Since the outer hash function h_{PI} is pairwise-independent, and therefore also uniform, $h_{\mathsf{PI}}(h_{\mathsf{LS}}(\cdot))$ is uniform over $[k]$. This endows $\mathcal{F}_{\mathsf{LS}}$ with low bias and variance with respect to f, as we explain in Section 3.1.

Second, the composition of two different hash functions gives us our fairness guarantee: h_{LS} maps close point pairs x, x' to the same bucket, then h_{PI} disperses pairs that were not hashed together — most of which are distant. This separation of point pairs by distance is precisely what enables good preservation of metric fairness, as we prove in Section 3.2.

[4] We use the definition of LSH as coined by Charikar [Cha02]. See [IM98] for an alternative gap-based definition in the same spirit.

3.1 Approximation of Outputs

We show the following bounds on the bias and variance of our derandomization. The proof is deferred to Appendix A.4.

Theorem 3.2 (Bias and variance of derandomized classifier)**.** *Let f be a stochastic classifier, $\hat{f} \sim \mathcal{F}_{\mathsf{LS}}$, and $\mathcal{D}$ a distribution over X. Then*

$$\text{bias}(\hat{f}, f, \mathcal{D}) \le \frac{1}{k} \quad \text{and} \quad \text{variance}(\hat{f}, f, \mathcal{D}) \le \mathop{\mathbb{E}}_{h_{\mathsf{LS}} \sim \mathcal{H}_{\mathsf{LS}}} \left[\max_{b \in B} \Pr_{x \sim \mathcal{D}} [h_{\mathsf{LS}}(x) = b] \right] \cdot \mathop{\mathbb{E}}_{x \sim \mathcal{D}} [f(x)(1 - f(x))] + \frac{1}{k}.$$

The above variance bound is similar in form to that of the pairwise-independent derandomization (Theorem 2.1), but with added randomization over the sampling of locality-sensitive hash function: when most choices of h_{LS} distribute points $x \sim \mathcal{D}$ into buckets relatively evenly, the bound is as small as $O(1/|B|)$; when most hashes are collisions, the bound may be as large as $\mathbb{E}_{x \sim \mathcal{D}}[f(x)(1 - f(x))]$, essentially tracking the stochasticity of f.

3.2 Preservation of Metric Fairness

We can now show that our derandomization procedure approximately preserves metric fairness, both in the sense of expected fairness for any pair of points (the usual convention in the metric fairness literature), as well as in aggregate over all point pairs.

Theorem 3.3 (Locality-sensitive derandomization preserves metric fairness)**.** *Let f be an (α, β, d)-fair stochastic classifier, where d is a metric with an LSH family $\mathcal{H}_{\mathsf{LS}}$ with $k \ge 2/\epsilon$ buckets. Then $\mathcal{F}_{\mathsf{LS}}$ is a deterministic classifier family satisfying the following:*

- *(Pairwise fairness) Consider any $x, x' \in X$, and assume without loss of generality that $f(x) \le f(x')$. Then*

$$\mathop{\mathbb{E}}_{\hat{f} \sim \mathcal{F}_{\mathsf{LS}}} \left[\left| \hat{f}(x) - \hat{f}(x') \right| \right] \le [\alpha + 2f(x)(1 - f(x'))] \cdot d(x, x') + \beta + \epsilon$$

- *(Aggregate fairness) For any distance threshold $\tau \in [0, 1]$, with probability at least $1 - \delta$ over the sampling of $\hat{f}$,*

$$\Pr_{(x,x') \sim X^2_{\le \tau}} \left[\hat{f}(x) \ne \hat{f}(x') \right] \le \left(1 + \frac{1}{\sqrt{\delta}} \right) ([\alpha + 2f(x)(1 - f(x'))] \cdot \tau + \beta + \epsilon).$$

The above fairness guarantees can be simplified by noticing that since $f(x) \le f(x')$ w.l.o.g., $f(x)(1 - f(x')) \le 1/4$; this yields the following worst-case bounds over f and (x, x'):

Corollary 3.4 (Worst-case fairness)**.** *When f is (α, β, d)-fair, $\mathcal{F}_{\mathsf{LS}}$ satisfies the following:*

- *(Pairwise fairness) $\left(\alpha + \frac{1}{2}, \beta + \epsilon, d\right)$-metric fairness on any $(x, x') \in X^2$, i.e.*

$$\mathop{\mathbb{E}}_{\hat{f} \sim \mathcal{F}_{\mathsf{LS}}} \left[\left| \hat{f}(x) - \hat{f}(x') \right| \right] \le \left(\alpha + \frac{1}{2} \right) \cdot d(x, x') + \beta + \epsilon.$$

- *(Aggregate fairness) For any distance threshold $\tau \in [0, 1]$, with probability at least $1 - \delta$ over the sampling of $\hat{f}$,*

$$\Pr_{(x,x') \sim X^2_{\le \tau}} \left[\hat{f}(x) \ne \hat{f}(x') \right] \le \left(1 + \frac{1}{\sqrt{\delta}} \right) \left(\alpha\tau + \frac{\tau}{2} + \beta + \epsilon \right).$$

In expectation and with high probability, therefore, the generated deterministic classifier approximates the fairness guarantee of the original classifier to within a small constant factor when there exists an LSH family $\mathcal{H}$ for d. To get a better sense what kind of guarantees this gives us, consider the following example:

Example 3.5. *Let f be a $(1, 0, d)$-fair stochastic classifier, and suppose we derandomize it to some $\hat{f} \sim \mathcal{F}_{\mathsf{LS}}$, choosing $k = 500$. Then by Corollary 3.4,*

- *(Pairwise fairness)* $\hat{f}$ *is* $(3/2, \epsilon, d)$*-metric fair.*
- *(Aggregate fairness) With probability at least* $1 - \delta = 3/4$ *(over the sampling of* $\hat{f}$*), at least* 76% *of point pairs within distance* $\tau = 1/20$ *receive identical predictions.*

We present a sketch of the proof of Theorem 3.3; see Appendix A.5 for the complete proof.

Proof sketch of Theorem 3.3. Consider any $x, x' \in X$. Since $\hat{f}$ is binary and $\mathcal{H}_{\mathsf{LS}}$ is locality-sensitive,

$$\begin{aligned}\mathop{\mathbb{E}}_{\hat{f}\sim\mathcal{F}_{\mathsf{LS}}}\left[\left|\hat{f}(x) - \hat{f}(x')\right|\right] &= \Pr_{\substack{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}\\ h_{\mathsf{PI}}\sim\mathcal{H}_{\mathsf{PI}}}}\left[\hat{f}(x) \neq \hat{f}(x')\right] \\ &= \Pr_{h_{\mathsf{LS}},h_{\mathsf{PI}}}\left[\hat{f}(x) \neq \hat{f}(x') \;\middle|\; h_{\mathsf{LS}}(x) = h_{\mathsf{LS}}(x')\right] \cdot (1 - d(x,x')) \\ &\quad + \Pr_{h_{\mathsf{LS}},h_{\mathsf{PI}}}\left[\hat{f}(x) \neq \hat{f}(x') \;\middle|\; h_{\mathsf{LS}}(x) \neq h_{\mathsf{LS}}(x')\right] \cdot d(x,x')\end{aligned}$$

From here, the proof is a systematic analysis of conditional probabilities. To give some intuition, notice that the event $[\hat{f}(x) \neq \hat{f}(x') \mid h_{\mathsf{LS}}(x) = h_{\mathsf{LS}}(x')]$ occurs precisely when $\frac{h_{\mathsf{PI}}(h_{\mathsf{LS}}(x))}{k}$ falls between $f(x)$ and $f(x')$; by the uniformity of $\mathcal{H}_{\mathsf{PI}}$, the probability of this is roughly $|f(x) - f(x')| \leq \alpha \cdot d(x,x') + \beta$. This is one of several cases that use the uniformity and symmetry properties of the composed hash function $h_{\mathsf{PI}}(h_{\mathsf{LS}}(\cdot))$ to express $|\hat{f}(x) - \hat{f}(x')|$ in terms of $|f(x) - f(x')|$. In some cases this is not possible, resulting in an additive $2f(x)(1 - f(x'))$ loss in α. ☐

3.3 Sample Complexity

Since the LSH-based derandomization procedure involves sampling two hash functions $\mathcal{H}_{\mathsf{PI}}$ and $\mathcal{H}_{\mathsf{LS}}$, it samples $\hat{f}$ using $O(\log|B| + \log k + S_d(X,B))$ random bits, where $O(\log|B| + \log k)$ is the number of bits used to sample a pairwise-independent hash function [Rub12], and $S_d(X,B)$ is the number of random bits required to sample a locality-sensitive hash function for metric d with domain X and range B. When the metric is the Euclidean distance, for example, $O(\dim X)$ random bits suffice [Ras19].

4 Structural Lemmas for Fair Classifier Derandomization

In this section, we present generic results applicable to all classifier derandomization procedures, as well as unify different definitions of fairness used in this paper and others.

4.1 Bias-Variance Decomposition

Up to this point, a "stochastic" classifier has signified any function f from X to $[0,1]$; in this sense, it does not necessarily contain any randomness of its own. However, when it comes time to perform a binary decision on some input x, $f(x)$ is typically interpreted as the probability of outputting 1, i.e. we use the (truly random) binary function $\mathbb{1}_f(x) \sim \mathrm{Bern}(f(x))$.

By how much does this prediction typically differ from that of some pre-sampled deterministic classifier $\hat{f}$? We show that this error can be decomposed into the bias of $\hat{f}$ and the variance of both $\hat{f}$ and f:

Lemma 4.1 (Bias-variance decomposition)**.** *Let* $f : X \to [0,1]$ *be a stochastic classifier and* $\mathcal{F}$ *a deterministic classifier family. Then for any* $x \in X$*,*

$$\mathop{\mathbb{E}}_{f,\hat{f}}\left[\left|\hat{f}(x) - \mathbb{1}_f(x)\right|\right] \leq \left|\mathrm{bias}(\hat{f}, \mathbb{1}_f, x)\right| + 2\left(\mathop{\mathrm{Var}}_{f}(\mathbb{1}_f(x)) + \mathop{\mathrm{Var}}_{\hat{f}\sim\mathcal{F}}\left(\hat{f}(x)\right)\right)^{2/3}$$

We defer the proof to Appendix A.6. For now, let us interpret this decomposition and see how it applies to the derandomization approaches laid out in previous sections. Recall that for all three derandomizations — $\mathcal{F}_{\mathsf{PI}}$, $\mathcal{F}_{\mathsf{RT}}$, and $\mathcal{F}_{\mathsf{LS}}$ — the bias was either zero or could be made arbitrarily small. As for the variance, we see two types: the first, $\mathrm{Var}_f(\mathbb{1}_f(x))$, is equal to $f(x)(1 - f(x))$, i.e. the variance of a Bernoulli with parameter $f(x)$; it therefore quantifies the inherent stochasticity of the given classifier f, over which we have no control. In contrast, the second variance arises from sampling the deterministic classifier $\hat{f}$, which depends greatly on the procedure being used. Thus a comparison of the expected error of these approaches boils down to this latter variance, for which

the pairwise-independent and locality-sensitive hashing approaches compare favorably against the simple random threshold.

4.2 Metric Fairness and Threshold Fairness

Friedler, Scheidegger, and Venkatasubramanian [FSV16] propose an alternative threshold-based notion of individual fairness that implements the mantra that "similar individuals should receive similar treatment," but only extends this constraint to pairs of inputs within a certain distance of interest:

Definition 4.2 ((σ, τ, d)-threshold fairness)**.** *Fix some constants $\sigma, \tau \in (0, 1)$. We say a stochastic classifier f is (σ, τ, d)-*threshold fair *if for all $x, x' \in X$ such that $d(x, x') \leq \sigma$, we have $|f(x) - f(x')| \leq \tau$. We say a deterministic classifier family $\mathcal{F}$ is (σ, τ, d)-threshold fair if for all $x, x' \in X$ such that $d(x, x') \leq \sigma$, we have $\mathbb{E}_{\hat{f} \sim \mathcal{F}}[|\hat{f}(x) - \hat{f}(x')|] \leq \tau$.*

Neither metric fairness nor threshold fairness fully subsumes the other. However, we can still show the following *algorithmic* reduction: if we wish to derandomize a stochastic classifier while preserving threshold fairness, then it suffices to use any procedure that preserves metric fairness. For example, suppose we have a derandomization procedure that worsens the input classifier's fairness parameters α and β to at most $a \cdot \alpha$ and $b \cdot \beta$, respectively, for some small constants $a, b \geq 1$. We should also expect this procedure to preserve threshold fairness, within certain parameters related to a, b. This is what we prove in the following lemma, but for more general fairness preservation functions:

Lemma 4.3 (Metric-fair derandomization preserves threshold fairness)**.** *Suppose we have a procedure that, given an (α, β, d)-metric fair stochastic classifier f, samples a deterministic classifier $\hat{f}$ from an $(A(\alpha), B(\beta), d)$-metric fair family $\mathcal{F}$, for some functions $A, B : \mathbb{R} \to \mathbb{R}$. Then this same procedure also derandomizes any (σ, τ, d)-threshold fair stochastic classifier to a deterministic classifier from a $(\sigma, A(0) \cdot \sigma + B(\tau), d)$-threshold fair family.*

Applying this to the random threshold and locality-sensitive derandomization procedures yields the following:

Corollary 4.4 (Threshold fairness-preserving derandomizations)**.** *Let f be a (σ, τ, d)-threshold fair stochastic classifier. Then*

- *The family $\mathcal{F}_{\mathsf{RT}}$ is (σ, τ, d)-threshold fair.*
- *If d is LSHable, the family $\mathcal{F}_{\mathsf{LS}}$, for a choice of $k \geq 4/\sigma$, is $(\sigma, \sigma + \tau, d)$-threshold fair.*

The proofs are deferred to Appendix A.7.

4.3 Pairwise Fairness and Aggregate Fairness

Throughout most of this paper (and in most of the individual fairness literature), we have been focused on pairwise notion of fairness, such as metric fairness (Definition 1.1) and threshold fairness (Definition 4.2). One shortcoming of these definitions is that even if a classifier satisfies them for any particular pair of points (x, x'), they do not hold simultaneously for all input pairs; thus once we sample a specific deterministic classifier $\hat{f}$, it may be unfair for many pairs. Fortunately, as we now show, these pairwise statements imply high-probability aggregate fairness guarantees: if $\mathcal{F}$ is a metric-fair family, then *most* deterministic classifiers in $\mathcal{F}$ assign *most* close pairs the same prediction.

To that end, for all distances $\tau \in [0, 1]$, let $X^2_{\leq \tau} := \left\{(x, x') \in X^2 \;\middle|\; d(x, x') \leq \tau\right\}$ denote the set of point pairs within distance τ. Then we can bound the fraction of τ-close pairs that receive different predictions:

Lemma 4.5 (Pairwise fairness implies aggregate fairness)**.** *Let $\mathcal{F}$ be an (α, β, d)-fair deterministic classifier family. Then for any distance threshold $\tau \in [0, 1]$, with probability at least $1 - \delta$ over the sampling of $\hat{f} \sim \mathcal{F}$,*

$$\Pr_{(x,x') \sim X^2_{\leq \tau}} \left[\hat{f}(x) \neq \hat{f}(x')\right] \leq \left(1 + \frac{1}{\sqrt{\delta}}\right)(\alpha\tau + \beta).$$

The proof is deferred to Appendix A.8.

4.4 Output Approximation and Loss Approximation

In this paper, we have analyzed the output approximation qualities of various derandomization techniques using the definitions of bias and variance in Section 1.2, which say that the output of $\hat{f}$ should resemble that of f, either on a single point x or in aggregate over some distribution $\mathcal{D}$.

An alternative set of definitions of bias and variance, put forth in [CGN19], instead measures how well $\hat{f}$ preserves the *loss* of f according to one or more binary loss functions ℓ. This property, which we might call *loss approximation*, is useful since in practice, classifiers are typically compared based on criteria such as accuracy, false positive rate, etc. evaluated on a dataset — and these are essentially binary loss functions averaged over a data distribution.

Concretely, let $\ell : \{0,1\} \times \{0,1\} \to \{0,1\}$ be a loss function and let $(x, y) \in X \times \{0,1\}$ be an instance with its corresponding label. The loss on this instance incurred by a (stochastic or deterministic) classifier f is defined as

$$L(f, x, y) := f(x)\ell(1, y) + (1 - f(x))\ell(0, y)$$

The (pointwise) bias and variance of $\hat{f}$ under this loss are then

$$\text{bias}(\hat{f}, f, x, y, \ell) := \left| \mathop{\mathbb{E}}_{\hat{f} \sim \mathcal{F}} \left[L(\hat{f}, x, y) \right] - L(f, x, y) \right| \qquad \text{and} \qquad \text{variance}(\hat{f}, x, y, \ell) := \mathop{\text{Var}}_{\hat{f} \sim \mathcal{F}} \left(L(\hat{f}, x, y) \right)$$

We observe that these are closely related to the simpler definitions given in Section 1.2:

Lemma 4.6. *For any $\ell : \{0,1\} \times \{0,1\} \to \{0,1\}$, $x \in X$, and $y \in \{0,1\}$,*

$$\text{bias}(\hat{f}, f, x, y, \ell) \leq \left| \text{bias}(\hat{f}, f, x) \right| \qquad \text{and} \qquad \text{variance}(\hat{f}, x, y, \ell) \leq \text{variance}(\hat{f}, x)$$

Thus even when the goal is to compute a derandomization that simulates the performance of f on one or more binary loss functions, it essentially suffices to use a derandomization that merely simulates the raw output of f itself. See Appendix A.9 for the proof of this lemma.

5 Discussion

We offer some brief notes regarding practical considerations for our derandomization framework.

A framework for derandomization Our results give machine learning practitioners a time- and space-efficient way to remove randomness — with the inherent brittleness, security vulnerabilities, and other issues that stochasticity entails — from their deployed models while approximately preserving fairness constraints enforced during training. Notably, our derandomization procedure has the useful quality of being *oblivious* to f, its training process, and even its actual fairness parameters α and β. It can therefore be applied as an independent post-processing step — for example, on the stochastic classifiers generated by the algorithms of [RY18], [KRR18], and others. The burden on the model designer is thus reduced to selecting a metric feature space (X, d) that is both appropriate for the classification task and for which an LSH family exists.

This simplification comes with inherent constraints: it was shown in [Cha02] that only metrics (or similarities ϕ whose complement d is a metric) can have LSH schemes, though not all of them do. On the positive side, recent work has shown that various non-LSHable similarities can be approximated by LSHable similarities with some provable distortion bound [CKP+19].

Separation of feature sets Throughout this paper, we have assumed that the inner hash function h_{LS} and classifiers f and $\hat{f}$ all share the same domain X; however, this is in no way necessary. In fact, from a fairness perspective, it is often prudent to distinguish between the features used for ensuring fairness and those used purely for inference, i.e. we may have

$$f : X \to [0,1], \ \hat{f} : X \to \{0,1\}, \text{ and } h_{\mathsf{LS}} : Z \to B$$

The feature set Z should be chosen, in tandem with an appropriate LSHable metric $d : Z \to [0,1]$, so as to measure similarity or difference between inputs on the basis of attributes that should be treated equitably; on the other

hand, the feature set X can be designed primarily to maximize predictive accuracy, and need not have any overlap with Z. The fairness guarantees of Theorem 3.3 and Corollary 3.4 then hold with respect to the metric space (Z, d) rather than (X, d).

Future work: guarantees for protected attributes This paper has focused on classifier derandomization with individual fairness guarantees, but it is also worthwhile to investigate the effect of derandomization from a group fairness perspective — for example, if it is possible to design an LSHable metric such that the derandomization preserves notions of fairness with respect to a protected attribute.

Acknowledgement This work is partially supported by the National Science Foundation (NSF) under grants IIS-2007951, IIS-2143895, IIS-2040800 (FAI program in collaboration with Amazon), and CCF-2023495.

References

[ABD+18] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. "A reductions approach to fair classification". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 60–69.

[AI06] Alexandr Andoni and Piotr Indyk. "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions". In: *2006 47th annual IEEE symposium on foundations of computer science (FOCS'06)*. IEEE. 2006, pp. 459–468.

[AIN+14] Alexandr Andoni, Piotr Indyk, Huy L Nguyen, and Ilya Razenshteyn. "Beyond locality-sensitive hashing". In: *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*. SIAM. 2014, pp. 1018–1028.

[AR15] Alexandr Andoni and Ilya Razenshteyn. "Optimal data-dependent hashing for approximate near neighbors". In: *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. 2015, pp. 793–801.

[BG20] Mark Braverman and Sumegha Garg. "The Role of Randomness and Noise in Strategic Classification". In: *1st Symposium on Foundations of Responsible Computing (FORC 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik. 2020.

[Bro97] Andrei Z Broder. "On the resemblance and containment of documents". In: *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*. IEEE. 1997, pp. 21–29.

[BS11] Michael Brückner and Tobias Scheffer. "Stackelberg games for adversarial prediction problems". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2011, pp. 547–555.

[CDP15] Yang Cai, Constantinos Daskalakis, and Christos Papadimitriou. "Optimum statistical estimation with strategic data sources". In: *Conference on Learning Theory*. PMLR. 2015, pp. 280–296.

[CGN19] Andrew Cotter, Maya Gupta, and Harikrishna Narasimhan. "On making stochastic classifiers deterministic". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2019).

[Cha02] Moses S Charikar. "Similarity estimation techniques from rounding algorithms". In: *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*. 2002, pp. 380–388.

[CKP+19] Flavio Chierichetti, Ravi Kumar, Alessandro Panconesi, and Erisa Terolli. "On the distortion of locality sensitive hashing". In: *SIAM Journal on Computing* 48.2 (2019), pp. 350–372.

[CLP20] Yiling Chen, Yang Liu, and Chara Podimata. "Learning Strategy-Aware Linear Classifiers". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 15265–15276.

[CPP+18] Yiling Chen, Chara Podimata, Ariel D Procaccia, and Nisarg Shah. "Strategyproof linear regression in high dimensions". In: *Proceedings of the 2018 ACM Conference on Economics and Computation*. 2018, pp. 9–26.

[CRK19] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. "Certified adversarial robustness via randomized smoothing". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 1310–1320.

[CWL21] Yatong Chen, Jialu Wang, and Yang Liu. "Linear Classifiers that Encourage Constructive Adaptation". In: *arXiv preprint arXiv:2011.00355* (2021).

[DHP+12] 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*. 2012, pp. 214–226.

[Die00] Thomas G Dietterich. "Ensemble methods in machine learning". In: *International workshop on multiple classifier systems*. Springer. 2000, pp. 1–15.

[DRS+18] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. "Strategic classification from revealed preferences". In: *Proceedings of the 2018 ACM Conference on Economics and Computation*. 2018, pp. 55–70.

[FSV16] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. "On the (im)possibility of fairness". In: *arXiv preprint arXiv:1609.07236* (2016).

[GZG+17] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna Gummadi, and Adrian Weller. "On Fairness, Diversity, and Randomness in Algorithmic Decision Making". In: *4th Workshop on Fairness, Accountability, and Transparency in Machine Learning*. 2017.

[HMP+16] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. "Strategic classification". In: *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*. 2016, pp. 111–122.

[IM98] Piotr Indyk and Rajeev Motwani. "Approximate nearest neighbors: towards removing the curse of dimensionality". In: *Proceedings of the thirtieth annual ACM symposium on Theory of computing*. 1998, pp. 604–613.

[JKG08] Prateek Jain, Brian Kulis, and Kristen Grauman. "Fast image search for learned metrics". In: *2008 IEEE Conference on computer vision and pattern recognition*. IEEE. 2008, pp. 1–8.

[KRR18] Michael P Kim, Omer Reingold, and Guy N Rothblum. "Fairness through computationally-bounded awareness". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 2018, pp. 4847–4857.

[Nar18] Harikrishna Narasimhan. "Learning with complex loss functions and constraints". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2018, pp. 1646–1654.

[PER+20] Rafael Pinot, Raphael Ettedgui, Geovani Rizk, Yann Chevaleyre, and Jamal Atif. "Randomization matters How to defend against strong adversarial attacks". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 7717–7727.

[PMA+19] Rafael Pinot, Laurent Meunier, Alexandre Araujo, Hisashi Kashima, Florian Yger, Cedric Gouy-Pailler, and Jamal Atif. "Theoretical evidence for adversarial robustness through randomization". In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 11838–11848.

[Ras19] Cyrus Rashtchian. *Lecture 09: LSH*. URL: http://madscience.ucsd.edu/notes/lec9.pdf. 2019.

[Rub12] Ronitt Rubinfeld. *MIT 6.842*. URL: https://people.csail.mit.edu/ronitt/COURSE/S12/handouts/lec5.pdf. 2012.

[RY18] Guy Rothblum and Gal Yona. "Probably approximately metric-fair learning". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 5680–5688.

[SKR19] Saeed Sharifi-Malvajerdi, Michael Kearns, and Aaron Roth. "Average individual fairness: Algorithms, generalization and experiments". In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 8242–8251.

A Omitted Proofs

A.1 Unfairness of Pairwise-Independent Derandomization

Proof of Proposition 2.2. For any $\delta > 0$, let $\mathbb{S}_\delta := \{x \in \mathbb{R}^n \mid d(x, \mathbf{0}) = \delta\}$ be the sphere of radius δ around the origin. Consider any $\alpha \geq 1$ and $\beta \in \left(0, \frac{1}{2} - \frac{1}{2k}\right)$, and choose X to be some subset of $\mathbb{S}_\delta$ of size $|X| = N$ in which the closest two points are positioned at distance ϵ from one another, where

$$0 < \epsilon := \min_{x,x' \in X} d(x, x') < \frac{1}{2} - \frac{1}{2k} - \beta.$$

Now let f be a classifier that maps half of the points in X to $\frac{1+\epsilon}{2}$, and the other half to $\frac{1-\epsilon}{2}$. f is $(1, 0, d)$-fair over X, since for any $x, x' \in X$,

$$|f(x) - f(x')| \leq \left|\frac{1+\epsilon}{2} - \frac{1-\epsilon}{2}\right| = \epsilon \leq d(x, x')$$

However, $\mathcal{F}_{\mathsf{PI}}$ is not (α, β, d)-fair on any point pair. To see this, consider any $x \neq x' \in X$; we show that for $\hat{f} \sim \mathcal{F}_{\mathsf{PI}}$, $|\hat{f}(x) - \hat{f}(x')|$ is typically large relative to $d(x, x')$:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\hat{f}\sim\mathcal{F}_{\mathsf{PI}}}\left[\left|\hat{f}(x) - \hat{f}(x')\right|\right] &= \Pr_{\hat{f}\sim\mathcal{F}_{\mathsf{PI}}}\left[\hat{f}(x) \neq \hat{f}(x')\right] && (\hat{f} \in \{0,1\}) \\
&= \Pr_{\hat{f}\sim\mathcal{F}_{\mathsf{PI}}}\left[\hat{f}(x) = 1, \hat{f}(x') = 0\right] + \Pr_{\hat{f}\sim\mathcal{F}_{\mathsf{PI}}}\left[\hat{f}(x) = 0, \hat{f}(x') = 1\right] \\
&= \Pr_{h\sim\mathcal{H}_{\mathsf{PI}}}\left[f(x) \geq \frac{h(x)}{k}, f(x') < \frac{h(x')}{k}\right] + \Pr_{h\sim\mathcal{H}_{\mathsf{PI}}}\left[f(x) < \frac{h(x)}{k}, f(x') \geq \frac{h(x')}{k}\right] \\
&\geq \Pr_{h\sim\mathcal{H}_{\mathsf{PI}}}\left[\frac{1-\epsilon}{2} \geq \frac{h(x)}{k}, \frac{1+\epsilon}{2} < \frac{h(x')}{k}\right] + \Pr_{h\sim\mathcal{H}_{\mathsf{PI}}}\left[\frac{1+\epsilon}{2} < \frac{h(x)}{k}, \frac{1-\epsilon}{2} \geq \frac{h(x')}{k}\right] \\
&= \Pr_{h\sim\mathcal{H}_{\mathsf{PI}}}\left[\frac{h(x)}{k} \leq \frac{1-\epsilon}{2}\right] \cdot \Pr_{h\sim\mathcal{H}_{\mathsf{PI}}}\left[\frac{h(x')}{k} > \frac{1+\epsilon}{2}\right] \\
&\quad + \Pr_{h\sim\mathcal{H}_{\mathsf{PI}}}\left[\frac{h(x)}{k} > \frac{1+\epsilon}{2}\right] \cdot \Pr_{h\sim\mathcal{H}_{\mathsf{PI}}}\left[\frac{h(x')}{k} \leq \frac{1-\epsilon}{2}\right] && \text{(by pairwise independence)} \\
&\geq \left(\frac{1-\epsilon}{2} - \frac{1}{k}\right)\left(1 - \frac{1+\epsilon}{2} - \frac{1}{k}\right) + \left(1 - \frac{1+\epsilon}{2} - \frac{1}{k}\right)\left(\frac{1-\epsilon}{2} - \frac{1}{k}\right) && \text{(by (7))} \\
&= \frac{1}{2}\left(1 - 2\epsilon + \epsilon^2\right) - \frac{1-\epsilon}{2k} + \frac{1}{k^2} \\
&\geq \frac{1}{2} - \epsilon - \frac{1}{2k}
\end{aligned}$$

The distance between any two points in $\mathbb{S}_\delta$, and therefore X, is at most 2δ; hence for a choice of $\delta \in \left(0, \frac{1/2-\beta-\epsilon-1/2k}{2\alpha}\right)$ (which is possible since $\beta < \frac{1}{2} - \frac{1}{2k}$ and $\epsilon < \frac{1}{2} - \frac{1}{2k} - \beta$), we have

$$\mathop{\mathbb{E}}_{h\sim\mathcal{H}}\left[\left|\hat{f}(x) - \hat{f}(x')\right|\right] \geq \frac{1}{2} - \epsilon - \frac{1}{2k} = 2\alpha \cdot \frac{1/2 - \beta - \epsilon - 1/2k}{2\alpha} + \beta > \alpha \cdot 2\delta + \beta \geq \alpha \cdot d(x, x') + \beta$$

which is a violation of (α, β, d)-metric fairness (Eq. (2)) and applies to all pairs $x, x' \in X$. ∎

A.2 Random Threshold Derandomization Guarantees

Proof of Proposition 2.3. Let f be an (α, β, d)-fair classifier, and consider any $x, x' \in X$. We have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\hat{f}_r\sim\mathcal{F}_{\mathsf{RT}}}\left[\left|\hat{f}_r(x) - \hat{f}_r(x')\right|\right] &= \Pr_{\hat{f}_r\sim\mathcal{F}_{\mathsf{RT}}}\left[\hat{f}_r(x) \neq \hat{f}_r(x')\right] && (\hat{f} \in \{0,1\}) \\
&= \Pr_{\hat{f}_r\sim\mathcal{F}_{\mathsf{RT}}}\left[\hat{f}_r(x) = 0, \hat{f}_r(x') = 1\right] + \Pr_{\hat{f}_r\sim\mathcal{F}_{\mathsf{RT}}}\left[\hat{f}_r(x) = 1, \hat{f}_r(x') = 0\right]
\end{aligned}$$

$$
\begin{aligned}
&= \Pr_{r\sim[0,1]}[f(x) < r \le f(x')] + \Pr_{r\sim[0,1]}[f(x') < r \le f(x)] \\
&= |f(x) - f(x')| \\
&\le \alpha \cdot d(x, x') + \beta && (f \text{ is } (\alpha, \beta, d)\text{-fair})
\end{aligned}
$$

which shows that $\mathcal{F}_{\mathsf{RT}}$ is also (α, β, d)-fair. To compute the bias, note that for any $x \in X$,

$$
\mathop{\mathbb{E}}_{\hat{f}_r\sim\mathcal{F}_{\mathsf{RT}}}\left[\hat{f}_r(x)\right] = \Pr_{r\sim[0,1]}[f(x) \ge r] = f(x) \tag{6}
$$

which implies $\text{bias}(\hat{f}_r, f, x) = 0$ for all x and hence $\text{bias}(\hat{f}, f, \mathcal{D})$ for all $\mathcal{D}$. Finally for the variance, we have

$$
\begin{aligned}
\text{variance}(\hat{f}_r, \mathcal{D}) &:= \mathop{\text{Var}}_{\hat{f}_r\sim\mathcal{F}_{\mathsf{RT}}}\left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\hat{f}_r(x)]\right) \\
&= \mathop{\mathbb{E}}_{r\sim[0,1]}\left[\left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\hat{f}_r(x)\right]\right)^2\right] - \left(\mathop{\mathbb{E}}_{r\sim[0,1]}\left[\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\hat{f}_r(x)\right]\right]\right)^2 \\
&= \mathop{\mathbb{E}}_{r\sim[0,1]}\left[\left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\hat{f}_r(x)\right]\right)^2\right] - \left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{r\sim[0,1]}\left[\hat{f}_r(x)\right]\right]\right)^2 \\
&= \mathop{\mathbb{E}}_{r\sim[0,1]}\left[\mathop{\mathbb{E}}_{x,x'\sim\mathcal{D}}\left[\hat{f}_r(x)\hat{f}_r(x')\right]\right] - \mathop{\mathbb{E}}_{x,x'\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{r\sim[0,1]}\left[\hat{f}_r(x)\right]\mathop{\mathbb{E}}_{r\sim[0,1]}\left[\hat{f}_r(x')\right]\right] \\
&= \mathop{\mathbb{E}}_{x,x'\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{r\sim[0,1]}\left[\hat{f}_r(x)\hat{f}_r(x')\right] - \mathop{\mathbb{E}}_{r\sim[0,1]}\left[\hat{f}_r(x)\right]\mathop{\mathbb{E}}_{r\sim[0,1]}\left[\hat{f}_r(x')\right]\right] \\
&= \mathop{\mathbb{E}}_{x,x'\sim\mathcal{D}}\left[\mathop{\text{Cov}}_{r\sim[0,1]}\left(\hat{f}_r(x), \hat{f}_r(x')\right)\right] \\
&\le \mathop{\mathbb{E}}_{x,x'\sim\mathcal{D}}\left[\sqrt{\mathop{\text{Var}}_{r\sim[0,1]}\left(\hat{f}_r(x)\right)\mathop{\text{Var}}_{r\sim[0,1]}\left(\hat{f}_r(x')\right)}\right] && \text{(Cauchy-Schwarz inequality)} \\
&= \left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\sqrt{\mathop{\text{Var}}_{r\sim[0,1]}\left(\hat{f}_r(x)\right)}\right]\right)^2 \\
&\le \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\mathop{\text{Var}}_{r\sim[0,1]}\left(\hat{f}_r(x)\right)\right] && \text{(Jensen's inequality)} \\
&= \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{r\sim[0,1]}\left[\hat{f}_r(x)\right]\left(1 - \mathop{\mathbb{E}}_{r\sim[0,1]}\left[\hat{f}_r(x)\right]\right)\right] \\
&= \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[f(x)(1 - f(x))] && \text{(Eq. (6))}
\end{aligned}
$$

as required. □

A.3 Perfect Deterministic Fairness is Impossible for Finite Families

Proof of Proposition 2.4. Consider any $\alpha \ge 1$ and $\beta \in (0, 1/|\mathcal{F}|)$; it suffices to exhibit a pair of points $x, x' \in X$ such that

$$
\mathop{\mathbb{E}}_{\hat{f}\sim\mathcal{F}}\left[\left|\hat{f}(x) - \hat{f}(x')\right|\right] > \alpha \cdot d(x, x') + \beta.
$$

For any $\delta > 0$, define the *ball of radius* δ *around* x to be $\mathbb{B}_\delta(x) := \{x' \in X \mid d(x, x') \le \delta\}$. By assumption, $\mathcal{F}$ contains at least one nontrivial classifier (i.e. one function that is not identically 1 or 0); let $\hat{f}$ be one such classifier. Since $X \subseteq \mathbb{R}^n$ is convex and d is a metric, $\hat{f}$ must be discontinuous at some point $x \in X$, meaning that for all $\delta > 0$, there exists $x' \in \mathbb{B}_\delta(x)$ such that $\hat{f}(x) = 1 - \hat{f}(x')$. Choose any $\delta^* \in \left(0, \frac{1/|\mathcal{F}| - \beta}{\alpha}\right)$, and consider some $x^* \in \mathbb{B}_{\delta^*}(x)$.

We have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\hat{f}\sim\mathcal{F}}\left[\left|\hat{f}(x)-\hat{f}(x^*)\right|\right] &\geq \frac{1}{|\mathcal{F}|} && \text{(at least one function in } \mathcal{F} \text{ is discontinuous at } x) \\
&= \alpha\left(\frac{1/|\mathcal{F}|-\beta}{\alpha}\right)+\beta \\
&> \alpha\cdot\delta^*+\beta && (\delta^* < \tfrac{1/|\mathcal{F}|-\beta}{\alpha}) \\
&\geq \alpha\cdot d(x,x^*)+\beta && (x^*\in\mathbb{B}_{\delta^*}(x))
\end{aligned}$$

which shows that $\mathcal{F}$ is not (α,β,d)-fair. ☐

A.4 Output Approximation of Locality-Sensitive Derandomization

Proof of Theorem 3.2. We will repeatedly use the following fact: by the uniformity of $\mathcal{H}_{\mathsf{PI}}$, for all $0\leq a<b\leq 1$ and $x\in X$ we have

$$\Pr_{\substack{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}\\ h_{\mathsf{PI}}\sim\mathcal{H}_{\mathsf{PI}}}}\left[a\leq\frac{h_{\mathsf{PI}}(h_{\mathsf{LS}}(x))}{k}\leq b\right]\in\left(b-a-\frac{1}{k},b-a+\frac{1}{k}\right) \tag{7}$$

Thus for all $x\in X$,

$$\mathop{\mathbb{E}}_{\hat{f}\sim\mathcal{F}_{\mathsf{LS}}}\left[\hat{f}(x)\right]=\Pr_{\hat{f}\sim\mathcal{F}_{\mathsf{LS}}}\left[\hat{f}(x)=1\right]=\Pr_{\substack{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}\\ h_{\mathsf{PI}}\sim\mathcal{H}_{\mathsf{PI}}}}\left[f(x)\geq\frac{h_{\mathsf{PI}}(h_{\mathsf{LS}}(x))}{k}\right]\in\left(f(x)-\frac{1}{k},f(x)+\frac{1}{k}\right)$$

which implies $\mathrm{bias}(\hat{f},f,x)\leq\frac{1}{k}$ for all $x\in X$ and hence $\mathrm{bias}(\hat{f},f,\mathcal{D})\leq\frac{1}{k}$ for all $\mathcal{D}$.

Now we bound the variance. Define the *bucketed* stochastic classifier

$$g(x)=\frac{1}{k}\sum_{i=1}^{k}\mathbb{1}\left\{f(x)\geq\frac{i}{k}\right\}$$

In other words, $g(x)$ is the smallest multiple of $1/k$ greater than $f(x)$. Note that $|g(x)-f(x)|\leq\frac{1}{k}$ for all x. Additionally, define the *deterministic* classifier family $\mathcal{G}_{\mathsf{LS}}$ from g just as $\mathcal{F}_{\mathsf{LS}}$ was defined from f in Eq. (5), i.e.

$$\mathcal{G}_{\mathsf{LS}}:=\left\{\hat{g}_{h_{\mathsf{LS}},h_{\mathsf{PI}}}\mid h_{\mathsf{LS}}\in\mathcal{H}_{\mathsf{LS}},h_{\mathsf{PI}}\in\mathcal{H}_{\mathsf{PI}}\right\},\quad\text{where}\quad\hat{g}_{h_{\mathsf{LS}},h_{\mathsf{PI}}}(x):=\mathbb{1}\left\{g(x)\geq\frac{h_{\mathsf{PI}}(h_{\mathsf{LS}}(x))}{k}\right\}. \tag{8}$$

It essentially suffices to analyze $\hat{g}$ instead of $\hat{f}$, since in the end, we simply incur an additional bias or variance of $\frac{1}{k}$. To begin, observe that for any distribution $\mathcal{D}$ over X,

$$\begin{aligned}
\mathrm{variance}(\hat{f},f,\mathcal{D}) &= \mathrm{variance}(\hat{g},g,\mathcal{D}) \\
&:= \mathop{\mathrm{Var}}_{\hat{g}\sim\mathcal{G}_{\mathsf{LS}}}\left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\hat{g}(x)]\right) \\
&= \mathop{\mathbb{E}}_{\substack{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}\\ h_{\mathsf{PI}}\sim\mathcal{H}_{\mathsf{PI}}}}\left[\left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\hat{g}(x)]\right)^2\right]-\left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\hat{g}(x)]\right)^2 \\
&= \mathop{\mathbb{E}}_{\substack{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}\\ h_{\mathsf{PI}}\sim\mathcal{H}_{\mathsf{PI}}}}\left[\left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\hat{g}(x)]\right)^2\right]-\left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[g(x)]\right)^2
\end{aligned}$$

To evaluate the first term, note that for any $x,x'\in X$,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\substack{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}\\ h_{\mathsf{PI}}\sim\mathcal{H}_{\mathsf{PI}}}}[\hat{g}(x)\hat{g}(x')] \\
&= \mathop{\mathbb{E}}_{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}}\left[\mathop{\mathbb{E}}_{h_{\mathsf{PI}}\sim\mathcal{H}_{\mathsf{PI}}}[\mathbb{1}\{h_{\mathsf{LS}}(x)=h_{\mathsf{LS}}(x')\}\hat{g}(x)\hat{g}(x')]+\mathop{\mathbb{E}}_{h_{\mathsf{PI}}\sim\mathcal{H}_{\mathsf{PI}}}[\mathbb{1}\{h_{\mathsf{LS}}(x)\neq h_{\mathsf{LS}}(x')\}\hat{g}(x)\hat{g}(x')]\right] \\
&= \mathop{\mathbb{E}}_{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}}\left[\mathop{\mathbb{E}}_{h_{\mathsf{PI}}\sim\mathcal{H}_{\mathsf{PI}}}[\mathbb{1}\{h_{\mathsf{LS}}(x)=h_{\mathsf{LS}}(x')\}\hat{g}(x)\hat{g}(x')]+\mathbb{1}\{h_{\mathsf{LS}}(x)\neq h_{\mathsf{LS}}(x')\}g(x)g(x')\right] \qquad \text{(pairwise independence)}
\end{aligned}$$

Thus the first term of the variance is

$$\begin{aligned}
\mathop{\mathbb{E}}_{\substack{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}\\ h_{\mathsf{PI}}\sim\mathcal{H}_{\mathsf{PI}}}}\left[\left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\hat{g}(x)]\right)^2\right] &= \mathop{\mathbb{E}}_{\substack{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}\\ h_{\mathsf{PI}}\sim\mathcal{H}_{\mathsf{PI}}}}\left[\mathop{\mathbb{E}}_{x,x'\sim\mathcal{D}}[\hat{g}(x)\hat{g}(x')]\right] \\
&= \mathop{\mathbb{E}}_{x,x'\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{\substack{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}\\ h_{\mathsf{PI}}\sim\mathcal{H}_{\mathsf{PI}}}}[\hat{g}(x)\hat{g}(x')]\right] \\
&= \mathop{\mathbb{E}}_{x,x'\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}}\left[\mathop{\mathbb{E}}_{h_{\mathsf{PI}}\sim\mathcal{H}_{\mathsf{PI}}}[\mathbb{1}\{h_{\mathsf{LS}}(x)=h_{\mathsf{LS}}(x')\}\hat{g}(x)\hat{g}(x')]+\mathbb{1}\{h_{\mathsf{LS}}(x)\neq h_{\mathsf{LS}}(x')\}g(x)g(x')\right]\right]
\end{aligned}$$

Next consider the second term:

$$\left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[g(x)]\right)^2 = \mathop{\mathbb{E}}_{x,x'\sim\mathcal{D}}[g(x)g(x')]$$

Putting these together, we have

$$\begin{aligned}
&\text{variance}(\hat{f},f,\mathcal{D}) \\
&= \mathop{\mathbb{E}}_{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}}\left[\mathop{\mathbb{E}}_{h_{\mathsf{PI}}\sim\mathcal{H}_{\mathsf{PI}}}\left[\mathop{\mathbb{E}}_{x,x'\sim\mathcal{D}}[\mathbb{1}\{h_{\mathsf{LS}}(x)=h_{\mathsf{LS}}(x')\}\hat{g}(x)\hat{g}(x')]\right]-\mathop{\mathbb{E}}_{x,x'\sim\mathcal{D}}[\mathbb{1}\{h_{\mathsf{LS}}(x)=h_{\mathsf{LS}}(x')\}g(x)g(x')]\right] \\
&= \mathop{\mathbb{E}}_{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}}\left[\mathop{\mathbb{E}}_{x,x'\sim\mathcal{D}}\left[\mathbb{1}\{h_{\mathsf{LS}}(x)=h_{\mathsf{LS}}(x')\}\cdot\left(\mathop{\mathbb{E}}_{h_{\mathsf{PI}}}[\hat{g}(x)\hat{g}(x')]-g(x)g(x')\right)\right]\right] \\
&= \mathop{\mathbb{E}}_{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}}\left[\mathop{\mathbb{E}}_{x,x'\sim\mathcal{D}}\left[\mathbb{1}\{h_{\mathsf{LS}}(x)=h_{\mathsf{LS}}(x')\}\cdot\left(\mathop{\mathbb{E}}_{h_{\mathsf{PI}}}[\hat{g}(x)\hat{g}(x')]-\mathop{\mathbb{E}}_{h_{\mathsf{PI}}}[\hat{g}(x)]\mathop{\mathbb{E}}_{h_{\mathsf{PI}}}[\hat{g}(x')]\right)\right]\right] \\
&= \mathop{\mathbb{E}}_{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}}\left[\mathop{\mathbb{E}}_{x,x'\sim\mathcal{D}}\left[\mathbb{1}\{h_{\mathsf{LS}}(x)=h_{\mathsf{LS}}(x')\}\cdot\mathop{\text{Cov}}_{h_{\mathsf{PI}}}(\hat{g}(x),\hat{g}(x'))\right]\right] \\
&\leq \mathop{\mathbb{E}}_{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}}\left[\mathop{\mathbb{E}}_{x,x'\sim\mathcal{D}}\left[\mathbb{1}\{h_{\mathsf{LS}}(x)=h_{\mathsf{LS}}(x')\}\cdot\sqrt{\mathop{\text{Var}}_{h_{\mathsf{PI}}}(\hat{g}(x))\mathop{\text{Var}}_{h_{\mathsf{PI}}}(\hat{g}(x'))}\right]\right] && \text{(Cauchy-Schwarz inequality)} \\
&= \mathop{\mathbb{E}}_{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}}\left[\sum_{b\in B}\left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\mathbb{1}\{h_{\mathsf{LS}}(x)=b\}\cdot\sqrt{\mathop{\text{Var}}_{h_{\mathsf{PI}}}(\hat{g}(x))}\right]\right)^2\right] \\
&= \mathop{\mathbb{E}}_{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}}\left[\sum_{b\in B}\left(\Pr_{x\sim\mathcal{D}}[h_{\mathsf{LS}}(x)=b]\cdot\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\sqrt{\mathop{\text{Var}}_{h_{\mathsf{PI}}}(\hat{g}(x))}\;\middle|\;h_{\mathsf{LS}}(x)=b\right]\right)^2\right] \\
&\leq \mathop{\mathbb{E}}_{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}}\left[\sum_{b\in B}\left(\Pr_{x\sim\mathcal{D}}[h_{\mathsf{LS}}(x)=b]\right)^2\cdot\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\mathop{\text{Var}}_{h_{\mathsf{PI}}}(\hat{g}(x))\;\middle|\;h_{\mathsf{LS}}(x)=b\right]\right] && \text{(Jensen's inequality)} \\
&= \mathop{\mathbb{E}}_{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}}\left[\sum_{b\in B}\left(\Pr_{x\sim\mathcal{D}}[h_{\mathsf{LS}}(x)=b]\right)^2\cdot\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[g(x)(1-g(x))\mid h_{\mathsf{LS}}(x)=b]\right] \\
&\leq \mathop{\mathbb{E}}_{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}}\left[\left(\max_{b\in B}\Pr_{x\sim\mathcal{D}}[h_{\mathsf{LS}}(x)=b]\right)\sum_{b\in B}\Pr_{x\sim\mathcal{D}}[h_{\mathsf{LS}}(x)=b]\cdot\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[g(x)(1-g(x))\mid h_{\mathsf{LS}}(x)=b]\right] \\
&= \mathop{\mathbb{E}}_{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}}\left[\max_{b\in B}\Pr_{x\sim\mathcal{D}}[h_{\mathsf{LS}}(x)=b]\right]\cdot\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[g(x)(1-g(x))] \\
&\leq \mathop{\mathbb{E}}_{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}}\left[\max_{b\in B}\Pr_{x\sim\mathcal{D}}[h_{\mathsf{LS}}(x)=b]\right]\cdot\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[f(x)(1-f(x))+\frac{1}{k}\right] && \text{(bias}(f,g,x)\leq\tfrac{1}{k}\text{ for all }x\text{)}
\end{aligned}$$

□

A.5 Fairness of LSH-Based Derandomization

Proof of Theorem 3.3. We first prove pairwise metric fairness. Consider any $x, x' \in X$, and assume without loss of generality that $f(x) \leq f(x')$. We have

$$\mathop{\mathbb{E}}_{\hat{f}\sim\mathcal{F}_{\mathsf{LS}}}\left[\left|\hat{f}(x)-\hat{f}(x')\right|\right]$$

$$
\begin{aligned}
&= \Pr_{\substack{h_{\mathsf{LS}}\sim\mathcal{H}_{\mathsf{LS}}\\ h_{\mathsf{PI}}\sim\mathcal{H}_{\mathsf{PI}}}}\left[\hat{f}(x)\neq\hat{f}(x')\right] && (\hat{f}\in\{0,1\})\\
&= \underbrace{\Pr_{\substack{h_{\mathsf{LS}}\\ h_{\mathsf{PI}}}}\left[\hat{f}(x)\neq\hat{f}(x') \;\middle|\; h_{\mathsf{LS}}(x)=h_{\mathsf{LS}}(x')\right]}_{p_1}\cdot\Pr_{h_{\mathsf{LS}}}[h_{\mathsf{LS}}(x)=h_{\mathsf{LS}}(x')]\\
&\quad+\underbrace{\Pr_{\substack{h_{\mathsf{LS}}\\ h_{\mathsf{PI}}}}\left[\hat{f}(x)\neq\hat{f}(x') \;\middle|\; h_{\mathsf{LS}}(x)\neq h_{\mathsf{LS}}(x')\right]}_{p_2}\cdot\Pr_{h_{\mathsf{LS}}}[h_{\mathsf{LS}}(x)\neq h_{\mathsf{LS}}(x')] && (9)
\end{aligned}
$$

We evaluate p_1 and p_2 separately. First, noting that a pairwise-independent hash family is also uniform, we have

$$
\begin{aligned}
&\Pr_{h_{\mathsf{LS}},h_{\mathsf{PI}}}\left[\hat{f}(x)=0,\hat{f}(x')=1 \;\middle|\; h_{\mathsf{LS}}(x)=h_{\mathsf{LS}}(x')\right]\\
&= \Pr_{h_{\mathsf{LS}},h_{\mathsf{PI}}}\left[f(x)<\frac{h_{\mathsf{PI}}(h_{\mathsf{LS}}(x))}{k},f(x')\geq\frac{h_{\mathsf{PI}}(h_{\mathsf{LS}}(x'))}{k} \;\middle|\; h_{\mathsf{LS}}(x)=h_{\mathsf{LS}}(x')\right]\\
&= \Pr_{h_{\mathsf{LS}},h_{\mathsf{PI}}}\left[f(x)<\frac{h_{\mathsf{PI}}(h_{\mathsf{LS}}(x))}{k}\leq f(x') \;\middle|\; h_{\mathsf{LS}}(x)=h_{\mathsf{LS}}(x')\right]\\
&= \Pr_{h_{\mathsf{LS}},h_{\mathsf{PI}}}\left[f(x)<\frac{h_{\mathsf{PI}}(h_{\mathsf{LS}}(x))}{k}\leq f(x')\right] && (h_{\mathsf{PI}}\text{ is uniform})
\end{aligned}
$$

By symmetry, $\Pr_{h_{\mathsf{LS}},h_{\mathsf{PI}}}[\hat{f}(x)=1,\hat{f}(x')=0 \mid h_{\mathsf{LS}}(x)=h_{\mathsf{LS}}(x')]=\Pr_{h_{\mathsf{LS}},h_{\mathsf{PI}}}[f(x)\geq\frac{h_{\mathsf{PI}}(h_{\mathsf{LS}}(x))}{k}>f(x')]$; but this equals zero, since $f(x)\leq f(x')$. Thus

$$
\begin{aligned}
p_1 &= \Pr_{h_{\mathsf{LS}},h_{\mathsf{PI}}}\left[\hat{f}(x)=1,\hat{f}(x')=0 \;\middle|\; h_{\mathsf{LS}}(x)=h_{\mathsf{LS}}(x')\right]+\Pr_{h_{\mathsf{LS}},h_{\mathsf{PI}}}\left[\hat{f}(x)=0,\hat{f}(x')=1 \;\middle|\; h_{\mathsf{LS}}(x)=h_{\mathsf{LS}}(x')\right]\\
&= \Pr_{h_{\mathsf{LS}},h_{\mathsf{PI}}}\left[f(x)<\frac{h_{\mathsf{PI}}(h_{\mathsf{LS}}(x))}{k}\leq f(x')\right]\\
&= |f(x)-f(x')|\pm\frac{2}{k} && (\text{by Eq. (7)})
\end{aligned}
$$

Next, to compute p_2, we have

$$
\begin{aligned}
&\Pr_{h_{\mathsf{LS}},h_{\mathsf{PI}}}\left[\hat{f}(x)=1,\hat{f}(x')=0 \;\middle|\; h_{\mathsf{LS}}(x)\neq h_{\mathsf{LS}}(x')\right]\\
&= \Pr_{h_{\mathsf{LS}},h_{\mathsf{PI}}}\left[f(x)\geq\frac{h_{\mathsf{PI}}(h_{\mathsf{LS}}(x))}{k},f(x')<\frac{h_{\mathsf{PI}}(h_{\mathsf{LS}}(x'))}{k} \;\middle|\; h_{\mathsf{LS}}(x)\neq h_{\mathsf{LS}}(x')\right]\\
&= \Pr_{h_{\mathsf{LS}},h_{\mathsf{PI}}}\left[f(x)\geq\frac{h_{\mathsf{PI}}(h_{\mathsf{LS}}(x))}{k}, \;\middle|\; h_{\mathsf{LS}}(x)\neq h_{\mathsf{LS}}(x')\right]\cdot\Pr_{h_{\mathsf{LS}},h_{\mathsf{PI}}}\left[f(x')<\frac{h_{\mathsf{PI}}(h_{\mathsf{LS}}(x'))}{k} \;\middle|\; h_{\mathsf{LS}}(x)\neq h_{\mathsf{LS}}(x')\right]\\
&\qquad (h_{\mathsf{PI}}\text{ is pairwise independent})\\
&= f(x)(1-f(x'))\pm\frac{1}{k} && (h_{\mathsf{PI}}\text{ is uniform})
\end{aligned}
$$

and by symmetry, $\Pr_{h_{\mathsf{LS}},h_{\mathsf{PI}}}[\hat{f}(x)=0,\hat{f}(x')=1 \mid h_{\mathsf{LS}}(x)\neq h_{\mathsf{LS}}(x')]=(1-f(x))f(x')\pm\frac{1}{k}$. Thus

$$
\begin{aligned}
p_2 &= \Pr_{h_{\mathsf{LS}},h_{\mathsf{PI}}}\left[\hat{f}(x)=1,\hat{f}(x')=0 \;\middle|\; h_{\mathsf{LS}}(x)\neq h_{\mathsf{LS}}(x')\right]+\Pr_{h_{\mathsf{LS}},h_{\mathsf{PI}}}\left[\hat{f}(x)=0,\hat{f}(x')=1 \;\middle|\; h_{\mathsf{LS}}(x)\neq h_{\mathsf{LS}}(x')\right]\\
&= f(x)-2f(x')f(x)+f(x')\pm\frac{2}{k}
\end{aligned}
$$

Substituting p_1 and p_2 back into Eq. (9) yields

$$\begin{aligned}
\mathop{\mathbb{E}}_{h_{\mathsf{LS}},h_{\mathsf{PI}}}\left[\left|\hat{f}(x)-\hat{f}(x')\right|\right] &= p_1\cdot\Pr_{h_{\mathsf{LS}}}[h_{\mathsf{LS}}(x)=h_{\mathsf{LS}}(x')]+p_2\cdot\Pr_{h_{\mathsf{LS}}}[h_{\mathsf{LS}}(x)\neq h_{\mathsf{LS}}(x')]\\
&= |f(x)-f(x')|\cdot(1-d(x,x'))+(f(x)-2f(x')f(x)+f(x'))\cdot d(x,x')\pm\frac{2}{k} && (h_{\mathsf{LS}}\text{ is LSH})\\
&= |f(x)-f(x')|+2f(x)(1-f(x'))\cdot d(x,x')\pm\frac{2}{k} && (10)\\
&\le \alpha\cdot d(x,x')+\beta+2f(x)(1-f(x'))\cdot d(x,x')+\frac{2}{k} && (f\text{ is }(\alpha,\beta,d)\text{-fair})\\
&\le [\alpha+2f(x)(1-f(x'))]\cdot d(x,x')+\beta+\epsilon && (k\ge 2/\epsilon)
\end{aligned}$$

which proves the pairwise fairness bound. The aggregate fairness bound then follows from Lemma 4.5.

□

A.6 Bias-Variance Decomposition

Proof of Lemma 4.1. For any $c>0$, we have

$$\begin{aligned}
\left|\hat{f}(x)-\mathbb{1}_f(x)\right| &\le \left|\mathop{\mathbb{E}}_{f,\hat{f}}\left[\hat{f}(x)-\mathbb{1}_f(x)\right]\right|+\left|\hat{f}(x)-\mathbb{1}_f(x)-\mathop{\mathbb{E}}_{f,\hat{f}}\left[\hat{f}(x)-\mathbb{1}_f(x)\right]\right|\\
&\le \left|\mathop{\mathbb{E}}_{f,\hat{f}}\left[\hat{f}(x)-\mathbb{1}_f(x)\right]\right|+c\cdot\mathop{\mathrm{Var}}_{f,\hat{f}}\left(\hat{f}(x)-\mathbb{1}_f(x)-\mathop{\mathbb{E}}_{f,\hat{f}}\left[\hat{f}(x)-\mathbb{1}_f(x)\right]\right)\\
&\qquad\qquad (\text{by Chebyshev's inequality, w.p. } 1-1/c^2)\\
&\le \left|\mathop{\mathbb{E}}_{f,\hat{f}}\left[\hat{f}(x)-\mathbb{1}_f(x)\right]\right|+c\cdot\mathop{\mathrm{Var}}_{\hat{f}}\left(\hat{f}(x)-\mathop{\mathbb{E}}_{\hat{f}}\left[\hat{f}(x)\right]\right)+c\cdot\mathop{\mathrm{Var}}_{f}\left(\mathbb{1}_f(x)-\mathop{\mathbb{E}}_{f}[\mathbb{1}_f(x)]\right)\\
&\qquad\qquad (\hat{f}(x)-\mathbb{E}_{\hat{f}}[\hat{f}(x)]\text{ and }\mathbb{1}_f(x)-\mathbb{E}_f[\mathbb{1}_f(x)]\text{ have mean zero})\\
&\le \left|\mathop{\mathbb{E}}_{f,\hat{f}}\left[\hat{f}(x)-\mathbb{1}_f(x)\right]\right|+c\cdot\mathop{\mathrm{Var}}_{\hat{f}}\left(\hat{f}(x)\right)+c\cdot\mathop{\mathrm{Var}}_{f}\left(\mathbb{1}_f(x)\right)
\end{aligned}$$

The above calculation fails with probability at most $1/c^2$, in which case the left-hand side still obeys the simple bound $|\hat{f}(x)-\mathbb{1}_f(x)|\le 1$. Thus taking expectations of both sides, we have

$$\mathop{\mathbb{E}}_{f,\hat{f}}\left[\left|\hat{f}(x)-\mathbb{1}_f(x)\right|\right]\le\left|\mathop{\mathbb{E}}_{f,\hat{f}}\left[\hat{f}(x)-\mathbb{1}_f(x)\right]\right|+c\cdot\mathop{\mathrm{Var}}_{\hat{f}}\left(\hat{f}(x)\right)+c\cdot\mathop{\mathrm{Var}}_{f}\left(\mathbb{1}_f(x)\right)+\frac{1}{c^2}$$

with probability 1 for any $c>0$. A choice of $c=(\mathrm{Var}_{\hat{f}\sim\mathcal{F}}(\hat{f}(x))+\mathrm{Var}_f(\mathbb{1}_f(x)))^{-1/3}$ yields the result. □

A.7 Metric-Fair Derandomization Preserves Threshold Fairness

Proof of Lemma 4.3. First, fix some $\sigma\in(0,1)$ and let $X^2_{\le\sigma}:=\left\{(x,x')\in X^2 \;\middle|\; d(x,x')\le\sigma\right\}$. Observe the following translations between metric and threshold fairness on this set:

1. If f is (σ,τ,d)-threshold fair, then for any $(x,x')\in X^2_{\le\sigma}$,
$$|f(x)-f(x')|\le\tau=0\cdot d(x,x')+\tau$$
So, f is also $(0,\tau,d)$-metric fair on such pairs (x,x').
2. If f is (α,β,d)-metric fair on all $(x,x')\in X^2_{\le\sigma}$, then for such pairs,
$$|f(x)-f(x')|\le\alpha\cdot d(x,x')+\beta\le\alpha\sigma+\beta$$
So, f is also $(\sigma,\alpha\sigma+\beta,d)$-threshold fair.

Now suppose we run our derandomization procedure on a (σ,τ,d)-threshold fair stochastic classifier f. Let $\mathcal{F}$ be the deterministic classifier family from which we sample our output. Then f is $(0,\tau,d)$-metric fair over $X^2_{\le\sigma}$ (by observation 1 above), $\mathcal{F}$ is then $(A(0),B(\tau),d)$-metric fair over $X^2_{\le\sigma}$ (by the fairness preservation guarantee), and $\mathcal{F}$ is also $(\sigma,A(0)\cdot\sigma+B(\tau),d)$-threshold fair (by observation 2). □

Proof of Corollary 4.4. If f is (σ, τ, d)-threshold fair, then $\mathcal{F}_{\mathsf{LS}}$ is (σ, τ', d)-threshold fair, where

$$\begin{aligned}
\tau' &= A(0) \cdot \sigma + B(\tau) && \text{(Lemma 4.3)} \\
&= \frac{1}{2} \cdot \sigma + \tau + \frac{2}{k} && \text{(Corollary 3.4)} \\
&= \sigma + \tau && \text{(choice of } k \geq 4/\sigma\text{)}
\end{aligned}$$

□

A.8 Pairwise Fairness Implies Aggregate Fairness

Proof of Lemma 4.5. For all distances $\xi \in [0,1]$, let $X^2_\xi := \left\{(x,x') \in X^2 \;\middle|\; d(x,x') = \xi\right\}$ denote the set of point pairs at distance exactly ξ. Then, for any given $\hat{f} \in \mathcal{F}$, let

$$\rho_\xi(\hat{f}) := \Pr_{(x,x') \sim X^2_\xi}\left[\hat{f}(x) \neq \hat{f}(x')\right] \qquad \text{and} \qquad \rho_{\leq\tau}(\hat{f}) := \Pr_{(x,x') \sim X^2_{\leq\tau}}\left[\hat{f}(x) \neq \hat{f}(x')\right]$$

denote the fraction of pairs at distance ξ and within τ, respectively, to which $\hat{f}$ assigns different outputs. Treating $\rho_\xi(\hat{f})$ as a random variable of $\hat{f}$, we have

$$\mathop{\mathbb{E}}_{\hat{f}\sim\mathcal{F}}\left[\rho_\xi(\hat{f})\right] = \mathop{\mathbb{E}}_{\hat{f}\sim\mathcal{F}}\left[\Pr_{\substack{(x,x')\\ \sim X^2_\xi}}\left[\hat{f}(x) \neq \hat{f}(x')\right]\right] = \mathop{\mathbb{E}}_{\hat{f}\sim\mathcal{F}}\left[\mathop{\mathbb{E}}_{\substack{(x,x')\\ \sim X^2_\xi}}\left[\left|\hat{f}(x) - \hat{f}(x')\right|\right]\right] = \mathop{\mathbb{E}}_{\substack{(x,x')\\ \sim X^2_\xi}}\left[\mathop{\mathbb{E}}_{\hat{f}\sim\mathcal{F}}\left[\left|\hat{f}(x) - \hat{f}(x')\right|\right]\right] \tag{11}$$

Thus the fraction of separated pairs within distance τ is

$$\begin{aligned}
\mathop{\mathbb{E}}_{\hat{f}\sim\mathcal{F}}\left[\rho_{\leq\tau}(\hat{f})\right] &:= \mathop{\mathbb{E}}_{\hat{f}\sim\mathcal{F}}\left[\Pr_{(x,x')\sim X^2_{\leq\tau}}\left[\hat{f}(x) \neq \hat{f}(x')\right]\right] \\
&= \int_0^\tau \mathop{\mathbb{E}}_{\hat{f}\sim\mathcal{F}}\left[\Pr_{(x,x')\sim X^2_{\leq\tau}}\left[\hat{f}(x) \neq \hat{f}(x') \;\middle|\; d(x,x') = \xi\right] \cdot \Pr_{(x,x')\sim X^2_{\leq\tau}}\left[d(x,x') = \xi\right]\, d\xi\right] \\
&= \int_0^\tau \mathop{\mathbb{E}}_{\hat{f}\sim\mathcal{F}}\left[\Pr_{(x,x')\sim X^2_\xi}\left[\hat{f}(x) \neq \hat{f}(x')\right]\right] \cdot \Pr_{(x,x')\sim X^2_{\leq\tau}}\left[d(x,x') = \xi\right]\, d\xi \\
&= \int_0^\tau \mathop{\mathbb{E}}_{(x,x')\sim X^2_\xi}\left[\mathop{\mathbb{E}}_{\hat{f}\sim\mathcal{F}}\left[\left|\hat{f}(x) - \hat{f}(x')\right|\right]\right] \cdot \Pr_{(x,x')\sim X^2_{\leq\tau}}\left[d(x,x') = \xi\right]\, d\xi && \text{(by Eq. (11))} \quad (12) \\
&\leq \int_0^\tau (\alpha\xi + \beta) \Pr_{(x,x')\sim X^2_{\leq\tau}}\left[d(x,x') = \xi\right]\, d\xi && \text{(by } (\alpha,\beta,d)\text{-fairness)} \quad (13) \\
&\leq (\alpha\tau + \beta) \int_0^\tau \Pr_{(x,x')\sim X^2_{\leq\tau}}\left[d(x,x') = \xi\right]\, d\xi \\
&= \alpha\tau + \beta && (14)
\end{aligned}$$

Since $\rho_{\leq\tau} \in [0,1]$, $\mathrm{Var}(\rho_{\leq\tau}) = \mathbb{E}[\rho^2_{\leq\tau}] - \mathbb{E}[\rho_{\leq\tau}]^2 \leq \mathbb{E}[\rho_{\leq\tau}]$. Thus applying Chebyshev's inequality to Eq. (14) yields

$$\Pr_{\hat{f}\sim\mathcal{F}}\left[\rho > \left(1 + \frac{1}{\sqrt{\delta}}\right)(\alpha\tau + \beta)\right] \leq \Pr_{\hat{f}\sim\mathcal{F}}\left[\rho > \left(1 + \frac{1}{\sqrt{\delta}}\right)\mathop{\mathbb{E}}_{\hat{f}\sim\mathcal{F}}[\rho]\right] \leq \delta$$

which proves the claim. □

A.9 Output Approximation and Loss Approximation

Proof of Lemma 4.6. For any $x \in X$ and $y \in \{0,1\}$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\hat{f}\sim\mathcal{F}}\left[L(\hat{f},x,y)\right] &= \mathop{\mathbb{E}}_{\hat{f}\sim\mathcal{F}}\left[\ell(\hat{f}(x),y)\right] && (\hat{f}(x)\in\{0,1\})\\
&= \mathop{\mathbb{E}}_{\hat{f}}\left[\ell(\hat{f}(x),y) \;\middle|\; \hat{f}(x)=1\right]\cdot\Pr_{\hat{f}}\left[\hat{f}(x)=1\right] + \mathop{\mathbb{E}}_{\hat{f}}\left[\ell(\hat{f}(x),y) \;\middle|\; \hat{f}(x)=0\right]\cdot\Pr_{\hat{f}}\left[\hat{f}(x)=0\right]\\
&= \ell(1,y)\cdot\mathop{\mathbb{E}}_{\hat{f}}\left[\hat{f}(x)\right] + \ell(0,y)\cdot\left(1-\mathop{\mathbb{E}}_{\hat{f}}\left[\hat{f}(x)\right]\right)\\
&= \ell(1,y)f(x) + \ell(0,y)\left(1-f(x)\right) \pm \mathrm{bias}(\hat{f},f,x)\\
&= f(x)\ell(1,y) + (1-f(x))\ell(0,y) \pm \mathrm{bias}(\hat{f},f,x)
\end{aligned}$$

which proves the first inequality concerning the bias. For the variance, notice that since ℓ is binary, either $\mathrm{Var}_{\hat{f}}\left(\ell(\hat{f}(x),y)\right) = \mathrm{Var}_{\hat{f}}\left(\hat{f}(x)\right)$ or $\mathrm{Var}_{\hat{f}}\left(\ell(\hat{f}(x),y)\right) = 0$. ∎

B Manipulation Deterrence in Strategic Classification

Fair derandomization procedures carry implications for the *strategic classification* problem, a popular framework for modeling the behavior of self-interested agents subject to classification decisions [HMP+16; CDP15; CPP+18; DRS+18; CLP20]. Formally, strategic classification is a Stackelberg game, or a sequential game between two players:

1. First, a *decision maker* or *model designer* publishes a classifier. Traditionally, this means a stochastic classifier $f : X \to [0,1]$, but in our setting, the model designer may publish a family of deterministic classifiers $\mathcal{F}$, and promises to select a single classifier from $\mathcal{F}$ uniformly at random.
2. Next, a *strategic agent* or *decision subject*, who is associated with some feature vector $x \in X$, decides either to present their true features x, or to change or *manipulate* their features to some $x' \in X$ to obtain the favorable outcome $\hat{f}(x') = 1$ with higher probability. However, the agent incurs a cost $c(x,x') \geq 0$ for altering their features.

Given a (stochastic or deterministic) classifier $f : X \to [0,1]$ and cost function $c : X^2 \to [0,1]$, the *utility* of an agent with original features x who changes to x' is defined as

$$U_f(x,x') := f(x') - c(x,x')$$

and the utility-maximizing move $\Delta_f(x) := \arg\max_{x'\in X} U_f(x,x')$ is called the *best response* of x under f and c.

In the following proposition, we observe a general connection between metric fairness and strategic manipulation; namely that the more fair a classifier is with respect to a metric cost function, the less incentive agents have to manipulate their features. The reason is intuitive: if a classifier is a smooth function, then an agent x cannot expect their outcome to change much by moving to some nearby point x'.

Proposition B.1 (Metric fairness implies reduced manipulation incentive)**.** *Let c be a metric cost function and let f be a (α,β,c)-metric fair classifier. Then the maximum utility gained by manipulating x to x' is*

$$U_f(x,x') - U_f(x,x) \leq (\alpha-1)\cdot c(x,x') + \beta.$$

If f is a deterministic classifier drawn from a family $\mathcal{F}$, then this holds in expectation over the sampling of f.

Proof of Proposition B.1. Under a classifier f, an individual with original features $x \in X$ who changes to $x' \in X$ derives utility

$$\begin{aligned}
U_f(x,x') &= f(x') - c(x,x')\\
&\leq f(x) + |f(x') - f(x)| - c(x,x')\\
&\leq f(x) + \alpha\cdot c(x,x') + \beta - c(x,x') && (f \text{ is } (\alpha,\beta,c)\text{-fair})\\
&= f(x) + (\alpha-1)\cdot c(x,x') + \beta\\
&= U_f(x,x) + (\alpha-1)\cdot c(x,x') + \beta
\end{aligned}$$

which proves the claim for stochastic classifiers. The proof for a deterministic family $\mathcal{F}$ results from taking an expectation $\mathbb{E}_{f\sim\mathcal{F}}[\cdot]$ on both sides. □

Braverman and Garg [BG20] already observed this fact for a stochastic classifier with $\alpha = 1$ and $\beta = 0$, in which case there is no incentive to manipulate. Note that by Proposition 2.4, deterministic families cannot achieve such small fairness parameters; hence the upper bound of Proposition B.1 cannot rule out *some* incentive to manipulate. Nevertheless, it presents a nontrivial worst-case guarantee since, for a classifier without any fairness constraints, there may be individuals near the decision boundary who can flip their decision from, for example, $f(x) = 0$ to $f(x') = 1$ at near-zero cost, thereby gaining utility $U(x, x') - U(x, x) \approx 1$ through manipulation.

Cost functions studied in the strategic classification literature include the L_2 [HMP+16; BS11] and Mahalanobis [CWL21] distances, both of which are metrics with known LSH families [AI06; JKG08]. Therefore, stochastic classifiers trained to be fair with respect to these costs automatically reduce incentives to manipulate features, and if such classifiers are derandomized using fairness-preserving methods, this quality is probably approximately preserved.