Fair Machine Unlearning: Data Removal while Mitigating Disparities

Alex Oesterling Harvard University
Jiaqi Ma UIUC
Flavio P. Calmon Harvard University
Himabindu Lakkaraju Harvard University

Abstract

The Right to be Forgotten is a core principle outlined by regulatory frameworks such as the EU's General Data Protection Regulation (GDPR). This principle allows individuals to request that their personal data be deleted from deployed machine learning models. While "forgetting" can be naively achieved by retraining on the remaining dataset, it is computationally expensive to do to so with each new request. As such, several *machine unlearning* methods have been proposed as efficient alternatives to retraining. These methods aim to approximate the predictive performance of retraining, but fail to consider how unlearning impacts other properties critical to real-world applications such as fairness. In this work, we demonstrate that most efficient unlearning methods cannot accommodate popular fairness interventions, and we propose the first *fair machine unlearning* method that can efficiently unlearn data instances from a fair objective. We derive theoretical results which demonstrate that our method can provably unlearn data and provably maintain fairness performance. Extensive experimentation with real-world datasets highlight the efficacy of our method at unlearning data instances while preserving fairness. Code is provided at `https://github.com/AI4LIFE-GROUP/fair-unlearning`.

1 INTRODUCTION

Machine learning applications, ranging from recommender systems to credit scoring, frequently involve training sophisticated models using large quantities of personal data. As a means to prevent misuse of such information, recent regulatory policies have included provisions[1] that allow data subjects to delete their information from databases and models used by organizations. These policies have given rise to a general regulatory principle known as the *right to be forgotten*. However, achieving this principle in practice is non-trivial; for instance, the naive approach of retraining a model from scratch each time a deletion request occurs is computationally prohibitive. As a result, a variety of *machine unlearning* (or simply *unlearning*) methods have been developed to effectively approximate the retraining process with reduced computational costs (Guo et al., 2020; Bourtoule et al., 2020; Izzo et al., 2021; Neel et al., 2020).

In critical real-world applications that use personal data we often seek to achieve other key properties, such as algorithmic fairness, in addition to unlearning. For example, if a social media recommender system systematically under-ranks posts from a particular demographic group, it could substantially suppress the public voices of that group (Beutel et al., 2019; Jiang et al., 2019). Similarly, if a machine learning model used for credit scoring exhibits bias, it could unjustly deny loans to certain demographic groups (Corbett-Davies and Goel, 2018). In medicine, the fair allocation of healthcare services is also critical to prevent the exacerbation of systemic racism and improve the quality of care for marginalized groups (Chen et al., 2021). Both protection against algorithmic discrimination and the right to be forgotten have been spelled out in various regulatory documents (OSTP, 2022; European Commission; CCPA, 2018), highlighting a need to achieve them simultaneously in practice.

To adhere to the above mentioned regulatory principles, practitioners must satisfy the right to be forgotten while still preventing discrimination in deployed set-

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

[1] For example, Article 17 in the European Union's General Data Protection Regulation (GDPR) (European Commission) or Section 1798.105 in the California Consumer Privacy Act (CCPA) (CCPA, 2018).

tings. However, most prior work in unlearning only considers models optimized for predictive performance, and fails to consider additional objectives. In fact, we find that most existing machine unlearning methods are not compatible with common fairness interventions. Specifically, popular unlearning methods (Guo et al., 2020; Neel et al., 2020; Izzo et al., 2021) provide theoretical analysis over training objectives that are convex and can be decomposed into sums of losses on individual training data points. On the other hand, existing fairness interventions generally have nonconvex objectives (Lowy et al., 2022) or involve pairwise or group comparisons between samples that make unlearning a datapoint nontrivial due to each instance's entangled influence on all other samples in the objective function[2] (Berk et al., 2017; Beutel et al., 2019). As such, current unlearning methods cannot be naively combined with fairness objectives, highlighting a need from real-world practitioners for a method that achieves fairness and unlearning simultaneously.

To fill in this critical gap, we formalize the problem of fair unlearning, and propose the first *fair unlearning* method that can provably unlearn over a fairness-constrained objective. Our fair unlearning approach takes a convex fairness regularizer with pairwise comparisons and unrolls these comparisons into an unlearnable form while ensuring the preservation of theoretical guarantees on unlearning. We provide theoretical results that our method can achieve a common notion of unlearning, *statistical indistinguishability* (Guo et al., 2020; Neel et al., 2020), while preserving a measure of *equality of odds* (Hardt et al., 2016), a popular group fairness metric. Furthermore, we empirically evaluate the proposed method on a variety of real-world datasets and verify its effectiveness across various data deletion request settings, such as deletion at random and deletion from minority and majority subgroups. We show that our method well approximates retraining over fair machine learning objectives in terms of both accuracy and fairness.

Our main contributions include:

- We formalize the problem of fair unlearning, i.e., how to develop models we can train fairly and unlearn from.
- We introduce the first fair unlearning method that can provably unlearn requested data points while preserving popular notions of fairness.
- We provide theoretical analysis at the intersection of fairness and unlearning that can be applied to a broad set of unlearning methods.
- We empirically verify the effectiveness of the proposed method through extensive experiments on a variety of real-world datasets and across different settings of data deletion requests.

[2]See Appendix A.4 for a detailed explanation.

2 RELATED WORK

Unlearning. The naive approach to machine unlearning is to retrain a model from scratch with each data deletion request. However, retraining is not feasible for companies with many large models or organizations with limited resources. Thus, the primary objective of machine unlearning is to provide efficient approximations to retraining. Early approaches in security and privacy attempt to achieve exact removal, where an unlearned model is identical to retraining, but are limited in model class (Cao and Yang, 2015; Ginart et al., 2019). Bourtoule et al. (2020) propose SISA, a flexible approach to exact unlearning that "shards" a dataset, dividing it and training an ensemble of models where each can be retrained separately. More recent approaches propose approximate removal, requiring the unlearned model to be "close" to the output of retraining. Some approximate removal methods focus on improving efficiency (Wu et al., 2020) and others try to preserve performance (Wu et al., 2022). While these methods apply to a large class of models, they have no formal guarantees on data removal. A second group of approximate approaches provide theoretical guarantees on the statistical indistinguishability of unlearned and retrained models. These noise-based methods leverage convex loss functions to guarantee unlearning with gradient updates (Neel et al., 2020) and Hessian methods (Guo et al., 2020; Sekhari et al.; Izzo et al., 2021). We augment this second set of approximate methods to simultaneously provide strong guarantees on data protection and preserve fairness performance while targeting a common class of models.

Fairness. There are a multitude of definitions for fairness in machine learning, such as individual fairness, multicalibration or multiaccuracy, and group fairness. Individual fairness (Dwork et al., 2012) posits that "similar individuals should be treated similarly" by a model. On the other hand, recent work has focused on multicalibration and multiaccuracy (Hébert-Johnson et al., 2018; Kearns et al., 2018; Deng et al., 2023), where predictions are required to be calibrated across subpopulations. These subpopulation definitions can be highly expressive, containing many intersectional identities from protected groups. In this work, however, we focus on the most commonly studied form of fairness, group fairness, which seeks to balance certain statistical metrics across predefined subgroups. Group fairness literature has proposed various definitions of

fairness, but the three most common definitions are Demographic Parity (Zafar et al., 2017; Feldman et al., 2015; Zliobaite, 2015; Calders et al., 2009), Equalized Odds, and Equality of Opportunity (Hardt et al., 2016). To achieve these definitions, there are generally three approaches to achieving group fairness: *preprocessing* which attempts to correct dataset imbalance to ensure fairness (Calmon et al., 2017), *in-processing* which occurs during training by modifying traditional empirical risk minimization objectives to include fairness constraints (Lowy et al., 2022; Berk et al., 2017; Agarwal et al., 2018; Martinez et al., 2020), and *postprocessing* which modifies predictions to ensure fair treatment (Alghamdi et al., 2022; Hardt et al., 2016). In this work we focus on in-processing algorithms because they simply modify an objective to account for fairness rather than requiring an additional operation before or after each unlearning request which would also have to be made unlearnable.

Intersections. Despite advancements in machine unlearning, the literature still lacks sufficient consideration of the downstream impacts of unlearning methods. While recent papers have explored the compatibility of the right to be forgotten with the right to explanation (Krishna et al., 2023), there is little work at the intersection of unlearning and fairness. In privacy literature, a thread of work has shown the incompatibility of group fairness with privacy (Esipova et al., 2022; Bagdasaryan et al., 2019; Cummings et al., 2019) but these incompatibilities arise due to privacy-specific methods, such as gradient clipping and differences in neighboring datasets. Fairness literature has studied the related problem of the influence of training data on fairness (Wang et al., 2022), but does not provide any methods for unlearning. In unlearning literature, recent empirical studies have shown that unlearning can increase disparity (Zhang et al., 2023), other works have demonstrated the incompatibility of fairness and unlearning for the SISA algorithm (Koch and Soll, 2023), and one work (Wang et al., 2023) has provided a method to achieve removal and fairness but uses a sharding and retraining algorithm over fairness-corrected graph data for GNNs. In this paper, we propose the first efficient method which achieves fairness while being provably unlearnable without requiring retraining.

3 PRELIMINARIES

Let $D = \{(x_1, y_1, s_1), \ldots, (x_n, y_n, s_n)\} \in \mathcal{D}$ be a training dataset with size n, where $\mathcal{D}$ is the set of all possible datasets. For $i = 1, 2, \ldots, n$, $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$ is a feature vector; $y_i \in \mathcal{Y} = \{0, 1\}$ is a binary label; and $s_i \in \mathcal{S} = \{a, b\}$ is a binary sensitive attribute. A learning algorithm is represented as a mapping $A : \mathcal{D} \to \mathcal{H}_\Theta$ that maps a dataset $D \in \mathcal{D}$ to a model $A(D)$ in a hypothesis space $\mathcal{H}$ parametrized by some $\theta \in \Theta$.

Unlearning. Given set of samples to be deleted $R \subseteq D$, the goal of unlearning is to efficiently find a model that resembles the model retrained on $D \setminus R$ from scratch. We define an unlearning mechanism as a mapping $M : \mathcal{H}_\Theta \times \mathcal{D} \times \mathcal{D} \to \mathcal{H}_\Theta$ that maps a learned model $A(D)$, the original dataset D, and a subset $R \subseteq D$ of data points to be deleted, to a new unlearned model $M(A(D), D, R)$. Ideally, the unlearned model $M(A(D), D, R)$ is *statistically indistinguishable* from $A(D \setminus R)$ (Guo et al., 2020; Neel et al., 2020). When both the learning algorithm A and the unlearning mechanism M are randomized, i.e., their outputs produce a probability distribution over the hypothesis $\mathcal{H}_\Theta$, the notion of statistical indistinguishability can be formalized as follows.

Definition 1 ((ϵ, δ)-Statistical Indistinguishability (Guo et al., 2020; Neel et al., 2020).)

For given $\epsilon, \delta > 0$, an unlearning mechanism M achieves (ϵ, δ)-Statistical Indistinguishability *with respect to A, if for all $\mathcal{M} \subseteq \mathcal{H}_\Theta, D \in \mathcal{D}, R \subseteq D$ and $|R| = 1$, we have*

$$\Pr(M(A(D), D, R) \in \mathcal{M}) \leq e^\epsilon \Pr(A(D \setminus R) \in \mathcal{M}) + \delta,$$
$$\Pr(A(D \setminus R) \in \mathcal{M}) \leq e^\epsilon \Pr(M(A(D), D, R) \in \mathcal{M}) + \delta.$$

Note that this definition is based on deletion of one data point ($|R| = 1$). We include results for unlearning over multiple requests ($|R| = m$) in the Appendix.

Fairness. We focus on a popular notion of group fairness, *Equalized Odds* (Hardt et al., 2016), and present results for other popular metrics such as *Demographic Parity*, *Equality of Opportunity*, and *Subgroup Accuracy* in the Appendix. Consider a data point (X, Y, S) randomly drawn from the data distribution, where $X \in \mathcal{X}, Y \in \mathcal{Y}$, and $S \in \mathcal{S}$. Note that X, Y, and S are random variables. Equalized Odds is then defined as follows.

Definition 2 (Equalized Odds (Hardt et al., 2016).) *A model $h_\theta : \mathcal{X} \to \mathcal{Y}$ satisfies* Equalized Odds *with respect to the sensitive attribute S and the outcome Y if the model prediction $h_\theta(X)$ and S are independent conditional on Y. More formally, for all $y \in \{0, 1\}$,*

$$\Pr(h_\theta(X) = 1 \mid S = a, Y = y) = \Pr(h_\theta(X) = 1 \mid S = b, Y = y).$$

In practice, when exact Equalized Odds is not achieved, we can use *Absolute Equalized Odds Difference*, the

absolute difference for both false positive rates and true positive rates between the two subgroups, as a quantitative measure for unfairness. The Absolute Equalized Odds Difference (AEOD) is defined as:

$$\text{AEOD}(\theta) := \frac{1}{2} \sum_{y \in \{0,1\}} \big| \Pr(h_\theta(X) = 1 \mid S = a, Y = y) - \Pr(h_\theta(X) = 1 \mid S = b, Y = y) \big|. \quad (1)$$

We build on prior approaches by incorporating fairness goals into the classic unlearning problem. In the next section, we propose the first method to satisfy fairness and unlearning simultaneously.

4 FAIR UNLEARNING

We provide an algorithm that can efficiently unlearn requested data points from a model trained with fair loss. We also prove theoretical guarantees on the unlearnability of our method, and show novel fairness bounds that can be applied to our method as well as prior work.

4.1 A Convex Fair Loss Function

To simultaneously achieve fairness and provable unlearning, we leverage a convex fair learning loss introduced by Berk et al. (2017). This fair loss can effectively optimize for several popular notions of group fairness, including Equalized Odds. In addition, the convexity of this loss offers mathematical convenience in the development of our provable unlearning algorithm.

We begin with a binary cross-entropy (BCE) loss for binary classification:

$$\mathcal{L}_{\text{BCE}}(\theta, D) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, x_i, y_i) + \frac{\lambda}{2} ||\theta||_2^2, \quad (2)$$

where $\ell(\theta, x_i, y_i) = -y_i \log(\langle x_i, \theta \rangle) - (1 - y_i) \log(1 - \langle x_i, \theta \rangle)$ is the logistic loss.

We then add a fairness regularizer. Let $G_a := \{i : s_i = a\}$, $G_b := \{i : s_i = b\}$ be sets of indices indicating subgroup membership for each sample in dataset D with $n_a = |G_a|$ and $n_b = |G_b|$. The fairness regularizer is defined as the following:

$$\mathcal{L}_{\text{fair}}(\theta, D) := \left(\frac{1}{n_a n_b} \sum_{i \in G_a} \sum_{j \in G_b} \mathbb{1}[y_i = y_j](\langle x_i, \theta \rangle - \langle x_j, \theta \rangle) \right)^2. \quad (3)$$

In simple terms, $\mathcal{L}_{\text{fair}}$ penalizes the pairwise difference in logits for samples with the same label between subgroups. Because $\mathcal{L}_{\text{fair}}$ considers samples that share a label regardless of label value, this term directly optimizes for Equalized Odds (Def. 2). The full fair loss becomes

$$\mathcal{L}(\theta, D) = \mathcal{L}_{\text{BCE}}(\theta, D) + \gamma \mathcal{L}_{\text{fair}}(\theta, D), \quad (4)$$

where γ is the fairness regularization parameter.

Most existing unlearning methods (Guo et al., 2020; Izzo et al., 2021; Neel et al., 2020) are specifically designed to unlearn over objective functions like Eq. (2), where the objective constitutes a sum of independent functions on each data point. However, the introduction of the fairness regularizer $\mathcal{L}_{\text{fair}}(\theta, D)$ (and most other fairness regularizers (Beutel et al., 2019)) entangles the influence of data points, rendering existing unlearning methods inapplicable. In the following section, we propose a method to address these issues and the first fair unlearning algorithm.

4.2 The Fair Unlearning Algorithm

Recall the goal of an unlearning algorithm M is to efficiently update the model $A(D)$, such that $M(A(D), D, R)$ is close to $A(D \setminus R)$. With a concrete definition of our fair loss, we can represent the fully trained model $A(D)$ by its parameters $\theta_D^* := \arg\min_\theta \mathcal{L}(\theta, D)$. Similarly, the retrained model $A(D \setminus R)$ corresponds to parameters $\theta_{D'}^* := \arg\min_\theta \mathcal{L}(\theta, D')$, where we define $D' := D \setminus R$ for simplicity.

Similar to several existing unlearning methods (Guo et al., 2020; Neel et al., 2020), our fair unlearning method consists of two steps. First we conduct an efficient update on θ_D^* to obtain updated model parameters $\theta_{D'}^-$. The goal of this part is to make the updated model parameters $\theta_{D'}^-$ close to the retrained model parameters $\theta_{D'}^*$. Then we add noise to the loss function such that both $A(D)$ and $A(D')$ are randomized, from which we can establish a statistical indistinguishability result between $M(\theta_D^*, D, R)$ and $A(D')$.

Efficient Model Update. We first introduce the efficient model update. To simplify notations, we can unroll the squared double sum in Eq. (3) into a single sum using a Cartesian product. Define $N := G_a \times G_b \times G_a \times G_b$. Then we can write $\mathcal{L}_{\text{fair}}$ as

$$\mathcal{L}_{\text{fair}}(\theta, D) = \frac{1}{|N|} \sum_{(i,j,k,l) \in N} \ell_{\text{fair}}(\theta, \{(x_f, y_f)\}_{f \in \{i,j,k,l\}}), \quad (5)$$

where

$$\ell_{\text{fair}}(\theta, \{(x_f, y_f)\}_{f \in \{i,j,k,l\}}) := \mathbb{1}[y_i = y_j](\langle x_i, \theta \rangle - \langle x_j, \theta \rangle)\mathbb{1}[y_k = y_l](\langle x_k, \theta \rangle - \langle x_l, \theta \rangle). \quad (6)$$

Without loss of generality, assume that we want to remove one single data point, the final sample (x_n, y_n, s_n),

and that this sample comes from subgroup a, i.e., $n \in G_a$. We propose an algorithm that updates θ^*_D to approximately minimize $\mathcal{L}(\theta, D')$ over the remaining dataset $D' = D \setminus R$ where $R = \{(x_n, y_n, s_n)\}$. Our algorithm takes a second-order Newton update from θ^*_D, which results in $\theta^-_{D'}$. Let $d'_a := \{i \in G_a : x_i \in D'\}$, $C_{D'} := d'_a \times G_b \times d'_a \times G_b$. We define a quantity Δ related to the residual difference between gradients over D and D' as follows:

$$\begin{aligned}\Delta := &\ell'(\theta^*_D, x_n, y_n) + \lambda\theta^*_D \\ &+ \gamma\frac{n}{|N|}\sum_{(i,j,k,l)\in N} \ell'_{\text{fair}}(\theta^*_D, \{(x_f, y_f)\}_{f\in\{i,j,k,l\}}) \\ &- \gamma\frac{n-1}{|C_{D'}|}\sum_{(i,j,k,l)\in C_{D'}} \ell'_{\text{fair}}(\theta^*_D, \{(x_f, y_f)\}_{f\in\{i,j,k,l\}}).\end{aligned} \tag{7}$$

Intuitively, the first and second term correspond to taking a gradient step in the direction of the BCE loss and ℓ_2 penalty respectively, while the third term counteracts the entangled influence the sample has on the remaining data and the fourth term takes a gradient step over the fairness penalty. By constructing Δ in this way, our unlearning algorithm becomes the following:

$$\theta^-_{D'} = \theta^*_D + H^{-1}_{\theta^*_D}\Delta, \tag{8}$$

where $H_{\theta^*_D} = \nabla^2\mathcal{L}(\theta^*_D, D')$ is the Hessian of the full loss. In Theorem 1, we achieve a bound on the gradient $||\nabla\mathcal{L}(\theta^-_{D'}; D')||_2$ and then apply this bound to perform (ϵ, δ)-unlearning.

Noisy Loss Perturbation. To achieve (ϵ, δ)-statistical indistinguishability, we add Gaussian noise to both the original training and the retraining process. Specifically, we modify the loss function by an additional term $\mathbf{b}^T\theta$, where $\mathbf{b} \sim N(0, \sigma I)^d$, resulting in the perturbed loss function $\mathcal{L}^{\mathbf{b}}(\theta, D) = \mathcal{L}(\theta, D) + \mathbf{b}^T\theta$.

Runtime Complexity. Hessian inversion takes $O(d^2n)$ time, but this can be precomputed and cached as in (Izzo et al., 2021; Guo et al., 2020). We can also rearrange terms in Eq. (7) to precompute the double sum over all elements in N which is $O(n^2)$ and then subtract terms accordingly from $N \setminus C$ which is $O(kn+k^2)$, where k is the size of our unlearning request and $k << n$. Thus, at unlearning time, our runtime is of order $O(kn + k^2)$.

Trade-off between fairness and unlearning. Although not directly related to each other, both fairness (in the form of increased regularization, γ) and unlearning (in the form of increased noise, σ), form a Pareto frontier with accuracy. Thus, for a given accuracy budget, fairness and unlearning come at a cost to one another.

4.3 Theoretical Guarantees

Theoretical Guarantee on Statistical Indistinguishability. When both the original training and retraining are randomized, the following Theorem 1 provides a guarantee that our unlearning method in Eq. (8) is (ϵ, δ)-indistinguishable from retraining for some $\epsilon, \delta > 0$.

Theorem 1 ((ϵ, δ)-unlearning.[3]) *Let θ^*_D be the output of algorithm A trained on $\mathcal{L}^{\boldsymbol{b}}$. Assume the loss ℓ is ψ-Lipschitz in its second derivative (ℓ'' is ψ-Lipschitz), and bounded in its first derivative by a constant $||\ell'(\theta, x, y)||_2 \leq g$. Assume the data is bounded such that for all $i \in n$, $||x_i||_2 \leq 1$. Then,*

$$\begin{aligned}&||\nabla\mathcal{L}(\theta^-_{D'}; D')||_2 \leq \\ &\qquad\frac{\psi}{\lambda^2(n-1)}\left(2g + ||\theta^*_D||_2\left|\left|\frac{8(n-1)}{n_a^2}\right|\right|_2\right)^2,\end{aligned} \tag{9}$$

and if $\boldsymbol{b} \sim N(0, k\epsilon'/\epsilon)^d$ with $k > 0$ and $||\nabla\mathcal{L}(\theta^-_{D'}; D')||_2 \leq \epsilon'$, then the unlearning algorithm defined by Eq. 8 is (ϵ, δ)-unlearnable with $\delta = 1.5\exp(-k^2/2)$.

For logistic loss, we know $||\ell'(\theta, x, y)||_2 \leq 1$ and that ℓ'' is 1/4-Lipschitz.

Intuitively, our goal is to achieve Definition 1 by making the likelihood of a model being output by training approximately equal to the likelihood of that same model being output by unlearning. We do this by 1) ensuring the gradient of our unlearned model is bounded, which for strongly convex loss directly means our unlearned parameter $\theta^-_{D'}$ is close to the optimal retrained parameter $\theta^*_{D'}$, and 2) adding a sufficient amount of noise to the training objective such that any parameter close to $\theta^*_{D'}$ could have been generated by A.

Theoretical Guarantee on Fairness. Finally, we show that the model output by our method, $\theta^-_{D'}$, has a bounded change in Absolute Equalized Odds Difference (Eq. (1)) in comparison to the model output by retraining, $\theta^*_{D'}$. We formalize that $\text{AEOD}(\theta^-_{D'})$ is bounded in the following theorem.

Theorem 2 (AEOD is bounded.) *Assume both $\ell_{BCE}(\theta, x, y) = \ell(\theta, x, y) + \frac{2\lambda}{n}||\theta||_2^2$ and ℓ_{fair} are L-Lipschitz. Suppose $\forall x \in \mathcal{X} \subseteq \mathbb{R}^d, ||x||_2 \leq 1$, and for all $\mathcal{Z} \subseteq \mathcal{X}$, the conditional probability $P(\mathcal{Z}|S, Y) \leq c\frac{|\mathcal{Z}|}{|\mathcal{X}|}$ for some $c \geq 1$, which controls the concentration of the probability distribution. Then, with the results from Thm. 1 we can show that $||\theta^*_{D'} - \theta^-_{D'}||_2 \leq \kappa$ for some κ*

[3]For proof and generalization to unlearning m samples, see Appendix A.1

being $O(1/n^2)$. Furthermore, we have that the absolute equalized odds difference for $\theta^-_{D'}$ is bounded by

$$AEOD(\theta^-_{D'}) \leq AEOD(\theta^*_{D'}) + c\left(1 - \frac{2\int_\mu^1 \frac{\pi^{\frac{d-1}{2}}}{\Gamma(\frac{d+1}{2})}(1-y^2)^{\frac{d-1}{2}}dy}{V}\right), \quad (10)$$

where V is the volume of the unit d-sphere, and $\mu = O(\sqrt{\kappa})$.

As n goes to infinity, κ approaches 0, the integral in Eq. (10) evaluates to $V/2$, so the second term vanishes.

Theorem 2 states that the unlearned and retrained models will be similar to each other in terms of fairness performance. By developing an unlearnable method that optimizes for fairness during training and retraining, our theory guarantees that unlearning will also be as fair. Furthermore, Thm. 2 only requires $||\theta^*_{D'} - \theta^-_{D'}||_2 \leq \kappa$ in addition to standard assumptions in unlearning literature for linear models such as Lipschitzness and bounded data. For any other method that achieves the above bound on parameter distance, Thm. 2 bounds the change in fairness of the unlearned and retrained model. Proofs for both theorems can be found in the Appendix.

5 EXPERIMENTAL RESULTS

In this section, we evaluate the effectiveness of fair unlearning in terms of both fairness and accuracy in the scenario where data points are deleted at random. Next, we consider more extreme settings for unlearning where data points are deleted disproportionately from a specific subgroup.

5.1 Experimental Setup

We evaluate the proposed fair unlearning method on three real-world datasets using a logistic regression model. However, note that previous work has shown combining an unlearnable linear layer with a DP-SGD trained neural network preserves unlearning guarantees (Guo et al., 2020). We simulate three types of data deletion requests and compare various unlearning methods with retraining. We also report the metrics of the original model trained on the full training set as a reference. For all results we repeat our experiments five times and report mean and standard deviation. Note that the effectiveness of erasing the influence of requested data points is guaranteed by our theory rather than evaluated empirically, as is common in prior unlearning literature (Guo et al., 2020; Neel et al., 2020).

Datasets. We consider three datasets commonly used in the machine learning fairness literature. 1) *COMPAS* (Angwin et al., 2016) is a classic fairness benchmark with the task of predicting a criminal defendant's likelihood of recidivism. 2) *Adult* from the UCI Machine Learning Repository (Asuncion and Newman, 2007) is a large scale dataset with the task of predicting income bracket. 3) High School Longitudinal Study (*HSLS*) is a dataset containing survey responses from high school students and parents with the task of predicting academic year (Ingels et al., 2011; Jeong et al., 2022). For all datasets, we use the individual's race as the sensitive group identity.

Data Deletion Settings. We simulate data deletions in three settings: 1) *random deletion*, where data deletion requests come randomly from the training set independent of subgroup membership; 2) *minority deletion*, where data deletion requests come randomly from the minority group of the training set; 3) *majority deletion*, where data deletion requests come randomly from the majority group of the training set.

Evaluation Metrics. We report experimental results on *Absolute Equalized Odds Difference* (AEOD) as defined in Eq. (1), with results for Demographic Parity, Equality of Opportunity, and Statistical Parity in the Appendix. In addition, we explore the tradeoffs between ϵ and δ which control different strengths of unlearning. Finally, we report runtime, test accuracy, and fairness-accuracy tradeoffs for γ in the Appendix.

Baseline Methods. We compare our method with four baselines designed for unlearning with BCE loss (Eq. (2)). In addition to brute-force retraining with BCE loss (named as **Retraining (BCE)**), we consider three unlearning methods from existing literature. The three methods are 1) **Newton** (Guo et al., 2020) a second-order approach that takes a single Newton step to approximate retraining; 2) **PRU** (Izzo et al., 2021) which uses a statistical technique called the residual method to efficiently approximate the Hessian used in the Newton step; and 3) **SISA** (Bourtoule et al., 2020) a method that splits the dataset into shards and trains a model on each shard, while unlearning by retraining each model separately to speed up the process. In addition, we compare our **Fair Unlearning** method designed for unlearning with fair loss (Eq. (4)) to brute-force retraining with fair loss (named as **Retraining (Fair)**). Finally, as references, we also evaluate models trained on the full training set with BCE loss and fair loss, respectively named as **Full Training (BCE)** and **Full Training (Fair)**. Note that the goal of all machine unlearning methods is to approximate retraining.

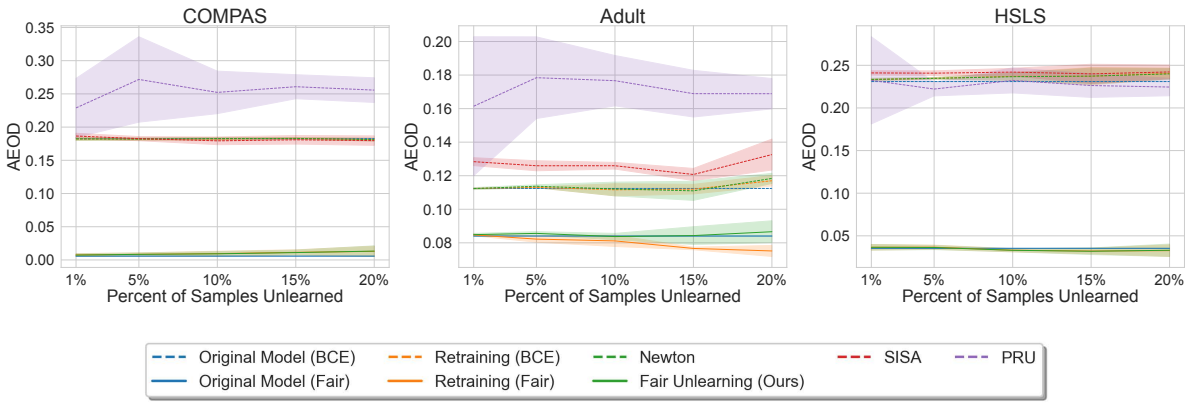


Figure 1: Absolute equalized odds difference (lower is better) for unlearning methods over random requests on COMPAS, Adult, and HSLS. Our method well-approximates fair retraining.

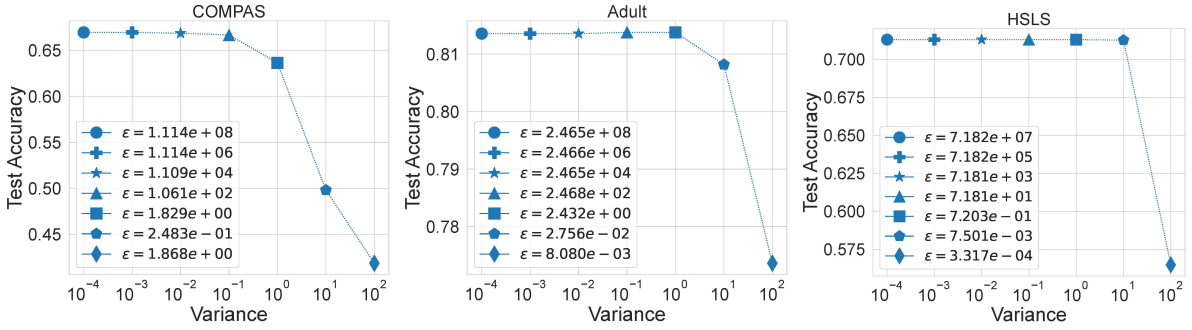


Figure 2: Unlearning performance and leakage for various levels of noise added while unlearning 100 samples according to Thm. 1. We set $\delta = 0.0001$ and report corresponding ϵ values in the legend.

Runtime Complexity. We compare the runtimes of our unlearning method and retraining over various removal sizes and report results in the Appendix. We also report precomputation times for Hessian inversion. Overall we see a 2× to 20× speedup with our method over retraining.

Unlearning Leakage. We report the unlearning leakage of our method for various levels of noise variance using the bound shown in Thm. 1. For small datasets, Guo et al. (2020) propose a tighter data dependent bound which we adapt to fair unlearning in Appendix A.2. We pick the lower value of ϵ between the two bounds for our reported results. For various magnitudes of σ, we fix $\delta = 0.0001$ and compute the accuracy of the model unlearned with our method in addition to the leakage, ϵ. Values are reported in Fig. 2. We can achieve realistic values of ϵ with reasonable amounts of noise while preserving performance.

5.2 Experimental Results on Fairness

Deletion at Random. We evaluate our method in terms of fairness when unlearning at random from our training dataset. For each of our datasets, and for both BCE and fair loss, we train a base logistic regression model on the full training dataset. Then, we randomly select 1%, 5%, 10%, 15%, and 20% of our training set to unlearn. We compare the AOED of our method to various baselines. Results are reported in Figure 1. Note that methods computed with BCE loss are shown with dashed lines whereas methods computed with fair loss are shown with solid lines.

As predicted by our theoretical results, our **Fair Unlearning** method (green) well-approximates the fairness performance of retraining with fair loss. In addition, we see that across all levels of unlearning, our method outperforms all baselines by a significant margin in AEOD for COMPAS, with similar results for Adult and HSLS. In the full statement of Thm. 1 for unlearning m samples (see Appendix A.1), we show that the unlearning bound scales with m. We similarly observe that the precision of our method decreases as we unlearn more samples. Note that **Newton** is the best approximator for **Retraining (BCE)**, oftentimes exactly covering the retraining curve, and **Fair Unlearning** similarly replicates the performance of **Retraining (Fair)**. Further, we see that certain exist-

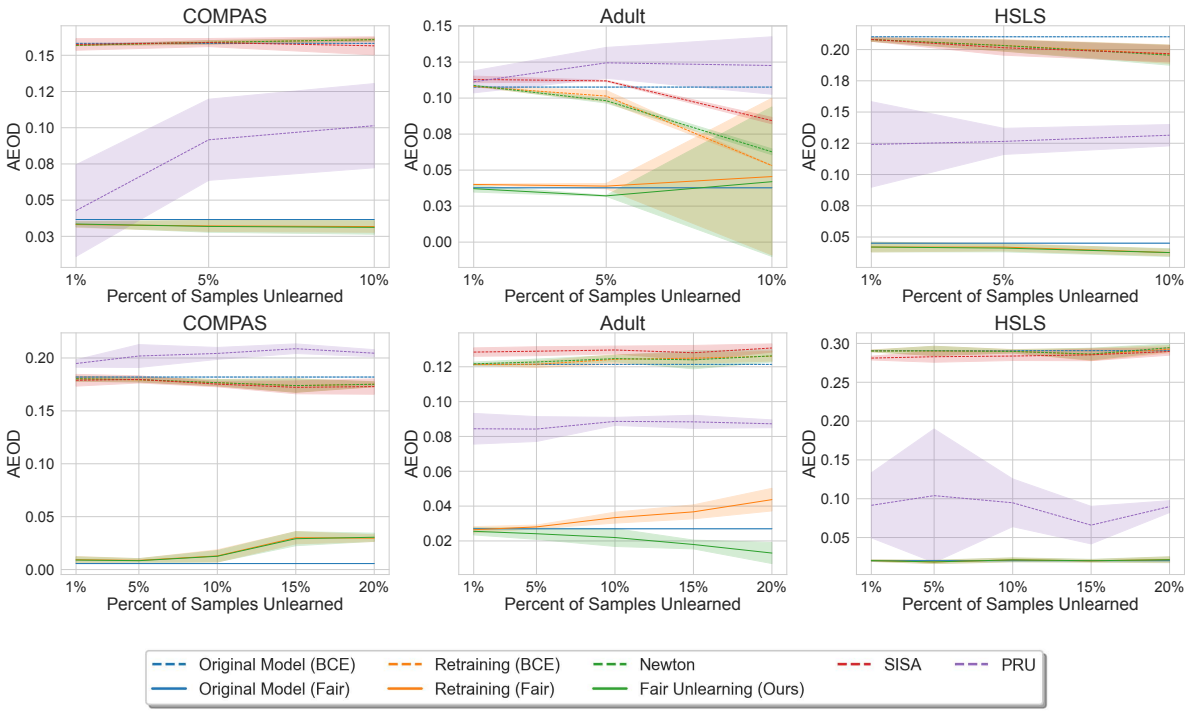


Figure 3: Absolute equalized odds difference (lower is better) for unlearning methods when unlearning from the minority (top) and majority (bottom) subgroup on COMPAS, Adult, and HSLS.

ing methods such as **PRU** and **SISA** are highly variable and can lead to greater disparity than retraining. Notably, across all datasets, our method outperforms baselines in preserving fairness.

Deletion from Subgroups. If unlearning at random preserves relative subgroup distribution in data, the worst case unlearning distribution for fairness is when unlearning requests are concentrated on specific subgroups. Subgroup deletion can also create a negative feedback loop where data deletion results in degradation in performance for that subgroup, motivating more members of said subgroup to request unlearning. To explore this, we unlearn only from minority groups and majority groups (Fig. 3). Note that when unlearning from minority groups, we only unlearn up to 10% of the dataset as the minority subgroup has (by definition) less representation in the dataset. Too many deletions from the minority group would destroy the stability of the learned models. For example, in the Adult dataset, which has 90% majority and 10% minority data, 10% deletion from the minority group is the maximum possible and already drastically reduces model performance.

We see that when unlearning from both minority and majority subgroups, our **Fair Unlearning** method approximates **Retraining (Fair)** and thus maintains the best fairness performance throughout unlearning. When unlearning from minority subgroups, for both COMPAS and HSLS, our method perfectly approximates fair retraining and largely outperforms other baselines in terms of AEOD. In Adult, note that vanilla unlearning methods improve in fairness despite unlearning from the minority subgroup. This is not unexpected as models can generalize better on subsets of a dataset, especially if the removed data are outliers, and the same can occur with fairness performance. However, we highlight that our method always achieves better AEOD than baselines. This indicates that even if removing a portion of data happens to help fairness, we can better improve fairness performance by explicitly accommodating fairness constraints. We also see that both fair retraining and unlearning have high standard deviation at 10% unlearning due to its imbalanced subgroup distribution. We do not see a similar change for baseline methods trained with BCE loss because those methods do not consider subgroups in their loss, whereas in the fair setting this error can be attributed primarily to the fair loss term which requires well-represented subgroups for accurate estimation. This behavior highlights a problem in unlearning settings that even when unlearning a small proportion of the total dataset, unlearning requests may completely destroy accurate estimates of subgroup distributions, resulting in downstream disparity when not considering fairness.

In majority subgroups, we see consistent performance

by our method across all datasets, with significant improvements in terms of AEOD. Finally, similar to the results for random unlearning, **PRU** display high error bands and atypical AEOD scores when compared to other baselines, while **Newton** remains accurate at approximating **Retraining (BCE)**.

5.3 Experimental Results on Accuracy

Finally, we validate the accuracy of **Fair Unlearning** to ensure that increased fairness does not come at a cost to performance. Trivially, AEOD can be optimized by a random guessing classifier despite its poor performance. We report test accuracy at 5% and 20% of samples unlearned across our various baselines and datasets in the supplemental material. We see that while achieving significant improvements in fairness, our method and brute-force retraining maintain performance in terms of accuracy and can even improve in some cases when unlearning at random and from subgroups.

6 CONCLUSION

In this work we show that existing methods fail to accommodate models with fairness considerations, and we resolve this issue by proposing the first fair unlearning algorithm. We prove our method is unlearnable and bound its fairness guarantees. Then, we evaluate our method on three real-world datasets and show that our method well approximates retraining with a fair loss function in terms of both accuracy and fairness. Further, we highlight worst case scenarios where unlearning requests disproportionately come from certain subgroups. Our fair unlearning method allows practitioners to achieve multiple data protection goals at the same time (fairness and unlearning) and satisfy regulatory principles. Furthermore, we hope this work sparks discussion surrounding other auxiliary goals that may need be compliant with unlearning such as robustness and interpretability, or other definitions of fairness such as multicalibration and multiaccuracy.

Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-2140743. This work is also supported in part by the NSF awards IIS-2008461, IIS-2040989, IIS-2238714, FAI-2040880, and research awards from Google, JP Morgan, Amazon, Adobe, Harvard Data Science Initiative, and the Digital, Data, and Design (D^3) Institute at Harvard. The views expressed here are those of the authors and do not reflect the official policy or position of the funding agencies.

References

European Commission. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). URL `https://eur-lex.europa.eu/eli/reg/2016/679`.

CCPA. California consumer privacy act (ccpa), 2018. URL `https://oag.ca.gov/privacy/ccpa`.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens van der Maaten. Certified Data Removal from Machine Learning Models, August 2020. URL `http://arxiv.org/abs/1911.03030`. arXiv:1911.03030 [cs, stat].

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine Unlearning, December 2020. URL `http://arxiv.org/abs/1912.03817`. arXiv:1912.03817 [cs].

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate Data Deletion from Machine Learning Models. page 10, 2021.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-Delete: Gradient-Based Methods for Machine Unlearning, July 2020.

Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2212–2220, 2019.

Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 383–390, 2019.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Richard J Chen, Tiffany Y Chen, Jana Lipkova, Judy J Wang, Drew FK Williamson, Ming Y Lu, Sharifa Sahai, and Faisal Mahmood. Algorithm fairness in ai for medicine and healthcare. *arXiv preprint arXiv:2110.00603*, 2021.

OSTP. Blueprint for an AI Bill of Rights. Technical report, The White House, Washington, DC, October 2022.

Andrew Lowy, Sina Baharlouei, Rakesh Pavan, Meisam Razaviyayn, and Ahmad Beirami. A Stochastic Optimization Framework for Fair Risk Minimization, September 2022.

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Yinzhi Cao and Junfeng Yang. Towards Making Systems Forget with Machine Unlearning. In *2015 IEEE Symposium on Security and Privacy*, pages 463–480, San Jose, CA, May 2015. IEEE. ISBN 978-1-4673-6949-7. doi: 10.1109/SP.2015.35.

Antonio Ginart, Melody Y. Guan, Gregory Valiant, and James Zou. Making AI Forget You: Data Deletion in Machine Learning, November 2019.

Yinjun Wu, Edgar Dobriban, and Susan B. Davidson. DeltaGrad: Rapid retraining of machine learning models, June 2020.

Ga Wu, Masoud Hashemi, and Christopher Srinivasa. PUMA: Performance Unchanged Model Augmentation for Training Data Removal. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8): 8675–8682, June 2022. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v36i8.20846.

Ayush Sekhari, Jayadev Acharya, Ananda Theertha Suresh, and Gautam Kamath. Remember What You Want to Forget: Algorithms for Machine Unlearning. page 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*, pages 214–226, 2012.

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

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

Zhun Deng, Cynthia Dwork, and Linjun Zhang. Happymap: A generalized multi-calibration method. *arXiv preprint arXiv:2303.04379*, 2023.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

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*, pages 259–268, 2015.

Indre Zliobaite. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, 2015.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops*, pages 13–18. IEEE, 2009.

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.

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

Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, pages 6755–6764. PMLR, 2020.

Wael Alghamdi, Hsiang Hsu, Haewon Jeong, Hao Wang, Peter Michalak, Shahab Asoodeh, and Flavio Calmon. Beyond adult and compas: Fair multi-class prediction via information projection. *Advances in Neural Information Processing Systems*, 35:38747–38760, 2022.

Satyapriya Krishna, Jiaqi Ma, and Himabindu Lakkaraju. Towards bridging the gaps between the right to explanation and the right to be forgotten. *arXiv preprint arXiv:2302.04288*, 2023.

Maria S Esipova, Atiyeh Ashari Ghomi, Yaqiao Luo, and Jesse C Cresswell. Disparate impact in differential privacy from gradient misalignment. *arXiv preprint arXiv:2206.07737*, 2022.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32, 2019.

Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In *Adjunct Publication of the*

27th Conference on User Modeling, Adaptation and Personalization, pages 309–315, 2019.

Jialu Wang, Xin Eric Wang, and Yang Liu. Understanding instance-level impact of fairness constraints. In *International Conference on Machine Learning*, pages 23114–23130. PMLR, 2022.

Dawen Zhang, Shidong Pan, Thong Hoang, Zhenchang Xing, Mark Staples, Xiwei Xu, Lina Yao, Qinghua Lu, and Liming Zhu. To be forgotten or to be fair: Unveiling fairness implications of machine unlearning methods. *arXiv preprint arXiv:2302.03350*, 2023.

Korbinian Koch and Marcus Soll. No matter how you slice it: Machine unlearning with sisa comes at the expense of minority classes. In *First IEEE Conference on Secure and Trustworthy Machine Learning*, 2023.

Cheng-Long Wang, Mengdi Huai, and Di Wang. Inductive graph unlearning. *arXiv preprint arXiv:2304.03093*, 2023.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2016.

Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

Steven J Ingels, Daniel J Pratt, Deborah R Herget, Laura J Burns, Jill A Dever, Randolph Ottem, James E Rogers, Ying Jin, and Steve Leinwand. High school longitudinal study of 2009 (hsls: 09): Base-year data file documentation. nces 2011-328. *National Center for Education Statistics*, 2011.

Haewon Jeong, Hao Wang, and Flavio P Calmon. Fairness without imputation: A decision tree approach for fair prediction with missing values. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9558–9566, 2022.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A APPENDIX

A.1 Proof of Theorem 1

We begin by introducing two lemmas.

Recall the definition of our full loss function,

$$\mathcal{L}(\theta, D) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, x_i, y_i) + \frac{\lambda}{2n}||\theta||_2^2 + \gamma \frac{1}{|N|} \sum_{i,j,k,l \in N} \ell_{\text{fair}}(\theta, \{(x_f, y_f)\}_{f\in\{i,j,k,l\}}),$$

$$\begin{aligned}\ell_{\text{fair}}(\theta, \{(x_f, y_f)\}_{f\in\{i,j,k,l\}}) &= \mathbb{1}[y_i = y_j](\langle x_i, \theta\rangle - \langle x_j, \theta\rangle)\mathbb{1}[y_k = y_l](\langle x_k, \theta\rangle - \langle x_l, \theta\rangle) \\ &= \mathbb{1}[y_i = y_j]\mathbb{1}[y_k = y_l]\left(\theta^T(x_i^T x_k - x_i^T x_l - x_j^T x_k + x_j^T x_l)\theta\right).\end{aligned}$$

For our proof of Lemma 1, we multiply by the entire objective by n:

$$\mathcal{L}(\theta, D) = \sum_{i=1}^{n} \ell(\theta, x_i, y_i) + \frac{\lambda}{2}||\theta||_2^2 + \gamma \frac{n}{|N|} \sum_{i,j,k,l \in N} \ell_{\text{fair}}(\theta, \{(x_f, y_f)\}_{f\in\{i,j,k,l\}}). \tag{11}$$

Let $\theta^-_{D'} := \theta^*_D + H^{-1}_{\theta^*_D}\Delta$ be the update step of our unlearing algorithm, where θ^*_D is our optimum when training on the entire dataset, $H^{-1}_{\theta^*_D} := (\nabla^2\mathcal{L}(\theta^*_D; D'))^{-1}$ is the inverse hessian of the loss at θ^*_D evaluated over the remaining dataset defined as $D' = D \setminus R$. Let R be our unlearning requests, and without loss of generality assume these are the last m elements of D, $\{x_{n-m}, ..., x_n\}$. Let $G_a := \{i : s_i = a\}$, $G_b := \{i : s_i = b\}$ be sets of indices indicating subgroup membership for each sample. Define $r_a := \{i \in G_a : x_i \in R\}$, $r_b := \{i \in G_b : x_i \in R\}$, and d'_a, d'_b similarly for $x_i \in D'$. Recall $n = |D|$, and let $n_a = |G_a|, n_b = |G_b|$ and similarly let $m = |R|, m_a = |r_a|, m_b = |r_b|$. Let $N = G_a \times G_b \times G_a \times G_b$, and let $C_{D'} = d'_a \times d'_b \times d'_a \times d'_b$. Then, Then, define our unlearning step Δ as the following:

$$\begin{aligned}\Delta := &\sum_{i=m}^{n} \ell'(\theta^*_D, x_i, y_i) + m\lambda\theta^*_D + \gamma\frac{n}{|N|} \sum_{(i,j,k,l)\in N} \ell'_{\text{fair}}(\theta^*_D, \{(x_f, y_f)\}_{f\in\{i,j,k,l\}}) \\ &- \gamma\frac{n-m}{|C_{D'}|} \sum_{(i,j,k,l)\in C_{D'}} \ell'_{\text{fair}}(\theta^*_D, \{(x_f, y_f)\}_{f\in\{i,j,k,l\}}).\end{aligned} \tag{12}$$

Lemma 1 *Assume the ERM loss ℓ is ψ-Lipschitz in its second derivative (ℓ'' is ψ-Lipschitz), and bounded in its first derivative by a constant $||\ell'(\theta, x, y)||_2 \le g$. Suppose the data is bounded such that for all $i \in n$, $||x_i||_2 \le 1$. Then the gradient of our unlearned model on the remaining dataset is bounded by:*

$$||\nabla\mathcal{L}(\theta^-_{D'}; D')||_2 \le \frac{\psi}{\lambda^2(n-m)}\left(2mg + ||\theta^*_D||_2\left|\left|\frac{8(m_a^2n_b^2 + m_b^2n_a^2 - m_a^2m_b^2)(n-m)}{(n_a^2n_b^2)}\right|\right|_2\right)^2. \tag{13}$$

Proof: This proof largely follows the structure of the proof of Theorem 1 in Guo et al. (2020), but differs in the addition of our fairness loss term which introduces additional terms into the bound.

Let $G(\theta) = \nabla L(\theta; D')$ be the gradient at the *remaining* dataset $D' = D \setminus R$. We want to bound the norm of the gradient at $\theta = \theta^-_{D'}$, as we know for convex loss, the optimum when retraining over the *remainder* dataset D' is zero. Keeping the norm of the unlearned weights close to zero ensures we have achieved unlearning.

We substitute the expression for the unlearned parameter and do a first-order Taylor approximation, which states there exists an $\eta \in [0, 1]$ such that the following holds:

$$\begin{aligned}G(\theta^-_{D'}) &= G(\theta^*_D + H^{-1}_{\theta^*_D}\Delta), \\ &= G(\theta^*_D) + \nabla G(\theta^*_D + \eta H^{-1}_{\theta^*_D}\Delta)H^{-1}_{\theta^*_D}\Delta.\end{aligned}$$

Recall that G is the gradient of the loss, so here ∇G is the Hessian evaluated at this η perturbed value. Let H_{θ_η} be the Hessian evaluated at the point $\theta_\eta = \theta^*_D + \eta H^{-1}_{\theta^*_D}\Delta$. Then we have

$$\begin{aligned}&= G(\theta^*_D) + H_{\theta_\eta}H^{-1}_{\theta^*_D}\Delta, \\ &= G(\theta^*_D) + \Delta + H_{\theta_\eta}H^{-1}_{\theta^*_D}\Delta - \Delta.\end{aligned} \tag{14}$$

By our choice of Δ, $G(\theta_D^*) + \Delta = 0$.

$$\begin{aligned} G(\theta_D^*) &= \nabla \mathcal{L}(\theta_D^*; D'), \\ &= \sum_{i=1}^{n-m} \ell'(\theta_D^*, x_i, y_i) + (n-m)\lambda\theta_D^* \\ &+ \gamma \frac{n-m}{|C_{D'}|} \sum_{(i,j,k,l) \in C_{D'}} \ell'_{\text{fair}}(\theta, \{(x_f, y_f)\}_{f \in \{i,j,k,l\}}). \end{aligned} \tag{15}$$

Combining (12) and (15), we see $G(\theta_D^*) + \Delta$ equals 0.

$$\begin{aligned} G(\theta_D^*) + \Delta &= \sum_{i=1}^{n} \ell'(\theta_D^*, x_i, y_i) + n\lambda\theta_D^* + \gamma \frac{n}{|N|} \sum_{i,j,k,l \in N} \ell'_{\text{fair}}(\theta, \{(x_f, y_f)\}_{f \in \{i,j,k,l\}}), \\ &= \nabla \mathcal{L}(\theta_D^*; D), \\ &= 0. \end{aligned} \tag{16}$$

We see in (16) that $G(\theta_D^*) + \Delta = \nabla \mathcal{L}(\theta_D^*; D)$. Then, $\nabla \mathcal{L}(\theta_D^*; D) = 0$ because our loss function is convex and our optimum occurs at θ_D^*.

Returning to the proof of Lemma 1, we have:

$$\begin{aligned} G(\theta_{D'}^-) &= G(\theta_D^*) + \Delta + H_{\theta_\eta} H_{\theta_D^*}^{-1} \Delta - \Delta, \\ &= 0 + H_{\theta_\eta} H_{\theta_D^*}^{-1} \Delta - \Delta, \\ &= H_{\theta_\eta} H_{\theta_D^*}^{-1} \Delta - H_{\theta_D^*} H_{\theta_D^*}^{-1} \Delta, \\ &= (H_{\theta_\eta} - H_{\theta_D^*}) H_{\theta_D^*}^{-1} \Delta. \end{aligned} \tag{17}$$

We want to bound the norm of (17).

$$\begin{aligned} ||G(\theta_{D'}^-)||_2 &= ||(H_{\theta_\eta} - H_{\theta_D^*}) H_{\theta_D^*}^{-1} \Delta||_2 \\ &\leq ||H_{\theta_\eta} - H_{\theta_D^*}||_2 ||H_{\theta_D^*}^{-1} \Delta||_2 \end{aligned}$$

First, we bound the left term. Recall we define H as the Hessian evaluated over the remainder dataset D'. Also note that ℓ''_{fair} and $\frac{\lambda}{2}||\theta||_2^2$ are constant in θ so $\ell''_{\text{fair}}(\theta_\eta, \{(x_f, y_f)\}_{f \in \{i,j,k,l\}}) - \ell''_{\text{fair}}(\theta_D^*, \{(x_f, y_f)\}_{f \in \{i,j,k,l\}}) = 0$.

Then we have

$$
\begin{aligned}
||H_{\theta_\eta} - H_{\theta^*_D}||_2 = || \sum_{i=1}^{n-m} \nabla^2 \ell(\theta_\eta, x_i, y_i) + \lambda(n-m) + \gamma \frac{n-m}{|C_{D'}|} \sum_{i,j,k,l \in C_{D'}} \ell''_{\text{fair}}(\theta_\eta, \{(x_f, y_f)\}_{f \in \{i,j,k,l\}}) \\
- \sum_{i=1}^{n-m} \nabla^2 \ell(\theta^*_D, x_i, y_i) - \lambda(n-m) - \gamma \frac{n-m}{|C_{D'}|} \sum_{i,j,k,l \in C_{D'}} \ell''_{\text{fair}}(\theta^*_D, \{(x_f, y_f)\}_{f \in \{i,j,k,l\}})||_2, \\
\leq \sum_{i=1}^{n-m} ||\nabla^2 \ell(\theta_\eta, x_i, y_i) - \nabla^2 \ell(\theta^*_D, x_i, y_i)||_2, \\
= \sum_{i=1}^{n-m} ||x_i^T (\ell''(\theta_\eta, x_i, y_i) - \ell''(\theta^*_D, x_i, y_i)) x_i||_2, \qquad (18) \\
\leq \sum_{i=1}^{n-m} ||\ell''(\theta_\eta, x_i, y_i) - \ell''(\theta^*_D, x_i, y_i)||_2 ||x_i||_2^2, \\
\leq \sum_{i=1}^{n-m} \psi ||\theta_\eta - \theta^*_D||_2 ||x_i||_2^2, \qquad (19) \\
\leq \sum_{i=1}^{n-m} \psi ||\theta_\eta - \theta^*_D||_2, \qquad (20) \\
= (n-m)\psi ||\theta_\eta - \theta^*_D||_2, \\
= (n-m)\psi ||\theta^*_D + \eta H^{-1}_{\theta^*_D} \Delta - \theta^*_D||_2, \\
= (n-m)\psi ||\eta H^{-1}_{\theta^*_D} \Delta||_2, \\
\leq (n-m)\psi ||H^{-1}_{\theta^*_D} \Delta||_2, \qquad (21)
\end{aligned}
$$

where (18) comes from chain rule, (19) comes from our assumption ℓ'' is ψ-Lipschitz, and (20) comes from our assumption that $||x_i||_2 \leq 1$.

Thus, we have:

$$
\begin{aligned}
||G(\theta^-_{D'})||_2 &\leq ||H_{\theta_\eta} - H_{\theta^*_D}||_2 ||H^{-1}_{\theta^*_D} \Delta||_2, \\
&\leq (n-m)\psi ||H^{-1}_{\theta^*_D} \Delta||_2^2 \qquad (22)
\end{aligned}
$$

We bound the quantity $||H^{-1}_{\theta^*_D} \Delta||_2$. Remember we are evaluating our Hessians and gradients over the remainder dataset, D'. With a convex ℓ and ℓ_{fair} where $\nabla^2 \ell, \nabla^2 \ell_{\text{fair}}$ are positive semi-definite and regularization of strength $\frac{n\lambda}{2}$, our loss function is $n\lambda$-strongly convex. Thus, our Hessian over the remainder dataset (of size $n-m$) is bounded:

$$
||H_\theta||_2 \geq (n-m)\lambda
$$

And so is its inverse

$$
||H^{-1}_\theta||_2 \leq \frac{1}{(n-m)\lambda} \qquad (23)
$$

Finally, we bound Δ. Recall its definition:

$$
\begin{aligned}
\Delta := \sum_{i=m}^{n} \ell'(\theta^*_D, x_i, y_i) + m\lambda\theta^*_D + \gamma \frac{n}{|N|} \sum_{(i,j,k,l) \in N} \ell'_{\text{fair}}(\theta^*_D, \{(x_f, y_f)\}_{f \in \{i,j,k,l\}}) \\
- \gamma \frac{n-m}{|C_{D'}|} \sum_{(i,j,k,l) \in C_{D'}} \ell'_{\text{fair}}(\theta^*_D, \{(x_f, y_f)\}_{f \in \{i,j,k,l\}}). \qquad (24)
\end{aligned}
$$

We bound θ_D^* to bound our second term. Recall that the gradient over the full dataset at θ_D^* equals 0.

$$0 = \nabla \mathcal{L}(\theta_D^*, D) = \sum_{i=1}^{n} \ell'(\theta_D^*, x_i, y_i) + n\lambda\theta_D^* + \gamma \frac{n}{|N|} \sum_{i,j,k,l \in N} \ell'_{\text{fair}}(\theta_D^*, \{(x_f, y_f)\}_{f \in \{i,j,k,l\}}) \tag{25}$$

We expand ℓ'_{fair} and see it is linear in θ. Denote this coefficient $F(i,j,k,l)$.

$$\begin{aligned} \ell'_{\text{fair}}(\theta_D^*, \{(x_f, y_f)\}_{f \in \{i,j,k,l\}}) &= 2 \cdot \mathbb{1}[y_i = y_j]\mathbb{1}[y_k = y_l](x_i^T x_k - x_i^T x_l - x_j^T x_k + x_j^T x_l)\theta_D^* \\ &= F(i,j,k,l)\theta_D^* \end{aligned} \tag{26}$$

We can bound θ_D^* by taking the stationary point condition (25) and solving for θ_D^*. Using the assumption the ERM loss is bounded $||\ell'(\theta, x_i, y_i)||_2 \leq g$, we see:

$$\begin{aligned} \theta_D^* &= \frac{-\sum_{i=1}^{n} \ell'(\theta_D^*, x_i, y_i)}{n\lambda + \gamma \frac{n}{|N|} \sum_{i,j,k,l \in N} F(i,j,k,l)}, \\ ||\theta_D^*||_2 &= \left|\left| \frac{\sum_{i=1}^{n} \ell'(\theta_D^*, x_i, y_i)}{n\lambda + \gamma \frac{n}{|N|} \sum_{i,j,k,l \in N} F(i,j,k,l)} \right|\right|_2, \\ ||\theta_D^*||_2 &\leq ng \left|\left| \frac{1}{n\lambda + \gamma \frac{n}{|N|} \sum_{i,j,k,l \in N} F(i,j,k,l)} \right|\right|_2, \\ ||\theta_D^*||_2 &\leq g \left|\left| \frac{1}{\lambda + \gamma \frac{1}{|N|} \sum_{i,j,k,l \in N} F(i,j,k,l)} \right|\right|_2 \end{aligned} \tag{27}$$

Now, we can bound the norm of Δ. We use the triangle inequality as well as the fact that $|C_{D'}| < |N|$, and the terms $n > m > 0$ to simplify.

$$\begin{aligned} ||\Delta||_2 &= \left|\left| \sum_{i=m}^{n} \ell'(\theta_D^*, x_i, y_i) + \theta_D^* \left[m\lambda + \gamma \left(\frac{n}{|N|} \sum_{i,j,k,l \in N} F(i,j,k,l) - \frac{n-m}{|C_{D'}|} \sum_{i,j,k,l \in C_{D'}} F(i,j,k,l) \right) \right] \right|\right|_2, \\ &\leq mg + ||\theta_D^*||_2 \left|\left| m\lambda + \gamma \left(\frac{n}{|N|} \sum_{i,j,k,l \in N} F(i,j,k,l) - \frac{n-m}{|C_{D'}|} \sum_{i,j,k,l \in C_{D'}} F(i,j,k,l) \right) \right|\right|_2, \\ &\leq mg + ||\theta_D^*||_2 \left|\left| m\lambda + \gamma \left(\frac{n}{|N|} \sum_{i,j,k,l \in N} F(i,j,k,l) - \frac{n-m}{|N|} \sum_{i,j,k,l \in C_{D'}} F(i,j,k,l) \right) \right|\right|_2, \\ &= mg + ||\theta_D^*||_2 \left|\left| m\lambda + \gamma \left(\frac{n}{|N|} \sum_{i,j,k,l \in N \setminus C_{D'}} F(i,j,k,l) + \frac{m}{|N|} \sum_{i,j,k,l \in C_{D'}} F(i,j,k,l) \right) \right|\right|_2, \\ &= gm + ||\theta_D^*||_2 \left|\left| m\lambda + \gamma \left(\frac{m}{|N|} \sum_{i,j,k,l \in N} F(i,j,k,l) + \frac{n-m}{|N|} \sum_{i,j,k,l \in N \setminus C_{D'}} F(i,j,k,l) \right) \right|\right|_2, \\ &\leq mg + ||\theta_D^*||_2 \left|\left| m\lambda + \gamma \frac{m}{|N|} \sum_{i,j,k,l \in N} F(i,j,k,l) \right|\right|_2 + ||\theta_D^*||_2 \left|\left| \frac{n-m}{|N|} \sum_{i,j,k,l \in N \setminus C_{D'}} F(i,j,k,l) \right|\right|_2 \end{aligned} \tag{28}$$

We can plug in our solution for $||\theta_D^*||_2$ from (27) and we see the second term equals g.

$$||\Delta||_2 \leq 2mg + ||\theta_D^*||_2 \left|\left| \frac{n-m}{|N|} \sum_{i,j,k,l \in N \setminus C_{D'}} F(i,j,k,l) \right|\right|_2 \tag{29}$$

Next we analyze the final term. Let, m_a, m_b be the number of removals from subgroups a and b respectively and recall that n_a, n_b is the total number of elements in subgroups a and b respectively. Recall our assumption

$||x_i||_2 \leq 1$. Note that $N \setminus C_{D'}$ is the union of two sets; first, the set of removed points from subgroup a (r_a) and their interactions with all points in subgroup b (G_b), and second the set of removed points from subgroup b (r_b) and their interactions with all points in subgroup a (G_a). Then, we can compute the cardinality $|N \setminus C_{D'}| = m_a^2 n_b^2 + m_b^2 n_a^2 - m_a^2 m_b^2$. Note that $|N| = n_a^2 n_b^2$.

$$
\begin{aligned}
&||\theta_D^*||_2 \left|\left| \frac{n-m}{|N|} \sum_{i,j,k,l \in N \setminus C_{D'}} F(i,j,k,l) \right|\right|_2 \\
&= ||\theta_D^*||_2 \left|\left| \frac{n-m}{|N|} \sum_{i,j,k,l \in N \setminus C_{D'}} 2 \cdot \mathbb{1}[y_i = y_j]\mathbb{1}[y_k = y_l](x_i^T x_k - x_i^T x_l - x_j^T x_k + x_j^T x_l) \right|\right|_2, \\
&\leq ||\theta_D^*||_2 \left|\left| \frac{8|N \setminus C_{D'}|(n-m)}{|N|} \right|\right|_2, \\
&= ||\theta_D^*||_2 \left|\left| \frac{8(m_a^2 n_b^2 + m_b^2 n_a^2 - m_a^2 m_b^2)(n-m)}{(n_a^2 n_b^2)} \right|\right|_2. \qquad (30)
\end{aligned}
$$

Combining (29) and (30),

$$
||\Delta||_2 \leq 2mg + ||\theta_D^*||_2 \left|\left| \frac{8(m_a^2 n_b^2 + m_b^2 n_a^2 - m_a^2 m_b^2)(n-m)}{(n_a^2 n_b^2)} \right|\right|_2. \qquad (31)
$$

The final form of our gradient norm $||\mathcal{L}(\theta_{D'}^-, D')||_2 = ||G(\theta_{D'}^-)||_2$ is the following.

$$
\begin{aligned}
||G(\theta_{D'}^-)||_2 &\leq (n-m)\psi ||H_{\theta_D^*}^{-1}\Delta||_2^2, \\
&\leq (n-m)\psi ||H_{\theta_D^*}^{-1}||_2^2 ||\Delta||_2^2, \\
&\leq (n-m)\psi \frac{1}{(n-m)^2\lambda^2} (2mg + ||\theta_D^*||_2 \left|\left| \frac{8(m_a^2 n_b^2 + m_b^2 n_a^2 - m_a^2 m_b^2)(n-m)}{(n_a^2 n_b^2)} \right|\right|_2)^2, \\
&= \frac{\psi}{\lambda^2(n-m)} \left(2mg + ||\theta_D^*||_2 \left|\left| \frac{8(m_a^2 n_b^2 + m_b^2 n_a^2 - m_a^2 m_b^2)(n-m)}{(n_a^2 n_b^2)} \right|\right|_2 \right)^2. \qquad \square \qquad (32)
\end{aligned}
$$

Assume the number of samples in subgroups a and b (n_a, n_b) are relatively balanced ($O(n)$), and the number of unlearned samples (m_a, m_b) are sufficiently low, such that $m_a = O(\sqrt{n_a}), m_b = O(\sqrt{n_b})$. Under these assumptions, we see that the term in the parenthesis is $O(1)$, so (32) is $O(1/n)$. When we let $m = 1$, and without loss of generality $m_a = 1, m_b = 0$, we exactly recover Thm. 1 in the main paper.

Lemma 2 (Guo et al. (2020)) *Let A be the learning algorithm that returns the unique optimum of the perturbed loss $\mathcal{L}(\theta, D) + \boldsymbol{b}^T\theta$, and let M be the an unlearning mechanism that outputs $\theta_{D'}^-$. Suppose that $||\nabla\mathcal{L}(\theta_{D'}^-, D')||_2 \leq \epsilon'$ for some $\epsilon' > 0$. Then, we have the following guarantees for M:*

1. *If $\boldsymbol{b}$ is drawn from a distribution with density $p(\boldsymbol{b}) \propto \exp -\frac{\epsilon}{\epsilon'}||\boldsymbol{b}||_2$, then M is ϵ-unlearnable for A;*
2. *If $\boldsymbol{b} \sim N(0, k\epsilon'/\epsilon)^d$ with $k > 0$, then M is (ϵ, δ)-unlearnable for A with $\delta = 1.5\exp(-k^2/2)$*

A.1.1 Proof of Theorem 1:

By combining Lemmas 1 and 2, we show that the loss gradient over the remaining dataset, evaluated at the unlearned parameter $\theta_{D'}^-$ is bounded, and we can directly plug this into Lemma 2 to achieve (δ, ϵ)-unlearning for some $\delta, \epsilon > 0$. $\square$

A.2 Data-Dependent Bound

From Eqn. (17), we are computing a bound on the norm of the following:

$$
(H_{\theta_\eta} - H_{\theta_D^*}) H_{\theta_D^*}^{-1} \Delta. \qquad (33)
$$

The Hessian for $\mathcal{L}(\theta, D')$ is

$$X^T W_\theta X + \lambda(n-m) + \gamma \frac{n-m}{|C_{D'}|} \sum_{i,j,k,l \in C_{D'}} \ell''_{\text{fair}}(\theta, \{(x_f, y_f)\}_{f \in \{i,j,k,l\}})$$

Where X is the data matrix of samples in D', and W_θ is a diagonal matrix where $W_{ii} = \ell''(\theta, x_i, y_i)$. Then, Eqn. (33) becomes

$$(X^T(W_{\theta_\eta} - W_{\theta^*_D})X)H^{-1}_{\theta^*_D}\Delta, \tag{34}$$

which by (Guo et al., 2020), Corollary 1, shows that

$$||\nabla\mathcal{L}(\theta^*_D; D')||_2 = ||(X^T(W_{\theta_\eta} - W_{\theta^*_D})X)H^{-1}_{\theta^*_D}\Delta||_2 \leq \gamma||X||_2||H^{-1}_{\theta^*_D}\Delta||_2||XH^{-1}_{\theta^*_D}\Delta||_2. \tag{35}$$

A.3 Proof of Theorem 2

We begin with a useful corollary of Lemma 1:

Corollary 1 ($||\theta^-_{D'} - \theta^*_{D'}||_2$ is bounded.) *Recall $\mathcal{L}(\theta, D')$ is $(n-m)\lambda$-strongly convex. Then, by strong convexity,*

$$\begin{aligned}
(n-m)\lambda||\theta^-_{D'} - \theta^*_{D'}||_2^2 &\leq (\nabla\mathcal{L}(\theta^-_{D'}, \mathcal{D}') - \nabla\mathcal{L}(\theta^*_{D'}, \mathcal{D}')^T(\theta^-_{D'} - \theta^*_{D'}),\\
&= (\nabla\mathcal{L}(\theta^-_{D'}, \mathcal{D}'))^T(\theta^-_{D'} - \theta^*_{D'}),\\
&\leq ||\nabla\mathcal{L}(\theta^-_{D'}, \mathcal{D}')||_2||\theta^-_{D'} - \theta^*_{D'}||_2,\\
(n-m)\lambda||\theta^-_{D'} - \theta^*_{D'}||_2 &\leq \frac{\psi}{\lambda^2(n-m)}\left(2mg + ||\theta^*_D||_2\left|\left|\frac{8(m_a^2n_b^2 + m_b^2n_a^2 - m_a^2m_b^2)(n-m)}{(n_a^2n_b^2)}\right|\right|_2\right)^2,\\
||\theta^-_{D'} - \theta^*_{D'}||_2 &\leq \frac{\psi}{\lambda^3(n-m)^2}\left(2mg + ||\theta^*_D||_2\left|\left|\frac{8(m_a^2n_b^2 + m_b^2n_a^2 - m_a^2m_b^2)(n-m)}{(n_a^2n_b^2)}\right|\right|_2\right)^2
\end{aligned} \tag{36}$$

Next, we introduce the lemmas we will use in this proof.

Lemma 3 (The interior angle between $\theta^*_{D'}$ and $\theta^-_{D'}$ is bounded.) *Let $||\theta^*_{D'} - \theta^-_{D'}||_2 \leq \kappa$. Define the angle between $\theta^*_{D'}$ and $\theta^-_{D'}$ as ϕ. Then,*

$$\begin{aligned}
\phi &\leq \arccos\left(\frac{||\theta^*_{D'}||_2^2 + ||\theta^-_{D'}||_2^2 - \kappa^2}{2||\theta^*_{D'}||_2||\theta^-_{D'}||_2}\right),\\
&\leq \arccos\left(\frac{||\theta^*_{D'}||_2^2 + (||\theta^*_{D'}||_2 - \kappa)^2 - \kappa^2}{2||\theta^*_{D'}||_2(||\theta^*_{D'}||_2 + \kappa)}\right)\\
&= \arccos\left(\frac{||\theta^*_{D'}||_2 - \kappa}{||\theta^*_{D'}||_2 + \kappa}\right)
\end{aligned} \tag{37}$$

The first inequality follows directly from applying the cosine angle formula to the triangle described by $\theta^*_{D'}$, $\theta^-_{D'}$, and $\theta^*_{D'} - \theta^-_{D'}$ and the fact that $||\theta^*_{D'} - \theta^-_{D'}||_2 \leq \kappa$. Then, we use the triangle inequality to show

$$\begin{aligned}
||\theta^-_{D'}||_2 &\leq ||\theta^*_{D'}||_2 + ||\theta^-_{D'} - \theta^*_{D'}||_2 \leq ||\theta^*_{D'}||_2 + \kappa,\\
||\theta^-_{D'}||_2 &\geq ||\theta^*_{D'}||_2 - ||\theta^*_{D'} - \theta^-_{D'}||_2 \geq ||\theta^*_{D'}||_2 - \kappa.
\end{aligned}$$

This lemma implies that as κ approaches 0, the angle ϕ approaches zero as well.

Lemma 4 (Volume of internal segment of a d-sphere) *Let S be a d-sphere with radius 1 and volume $V = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$. Then, let θ_1, θ_2 be normal vectors defining two hyperplanes such that their intersection lies tangent to the surface of S with internal angle ϕ such that the angle bisector to ϕ also bisects S. Then, the volume of the region of intersection B between the half spaces defined by the hyperplanes at ϕ within S equals*

$$\begin{aligned}
B &= V - 2\int_\mu^1 \frac{\pi^{\frac{d-1}{2}}}{\Gamma(\frac{d+1}{2})}(1-y^2)^{d-1}dy,\\
\mu &= \frac{1}{2}\sin\frac{\phi}{2}.
\end{aligned}$$

We take the full volume V and subtract off the volume of the spherical segment defined as the integral of the volume for a $d-1$-sphere over various radius values along the hyperplane to the surface of the sphere. Note that as $\phi \to 0, \mu \to 0$, and therefore $B \to 0$ as the right term approaches V.

A.3.1 Proof of Theorem 2:

Let $\theta^*_{D'}$ have a mean absolute equalized odds upper bounded by α, or

$$AEOD(\theta^*_{D'}) \leq \alpha. \tag{38}$$

This implies

$$|P(h_{\theta^*_{D'}}(X) = 1 \mid S = a, Y = 1) - P(h_{\theta^*_{D'}}(X) = 1 \mid S = b, Y = 1)| \leq \alpha, \tag{39}$$
$$|P(h_{\theta^*_{D'}}(X) = 1 \mid S = a, Y = 0) - P(h_{\theta^*_{D'}}(X) = 1 \mid S = b, Y = 0)| \leq \alpha. \tag{40}$$

Define $A_{\theta^*_{D'}} = \{x : h_{\theta^*_{D'}}(x) = 1\}, A_{\theta^-_{D'}} = \{x : h_{\theta^-_{D'}}(x) = 1\}$. Then,

$$|P(A_{\theta^*_{D'}} \mid S = a, Y = 1) - P(A_{\theta^*_{D'}} \mid S = b, Y = 1)| \leq \alpha, \tag{41}$$
$$|P(A_{\theta^*_{D'}} \mid S = a, Y = 0) - P(A_{\theta^*_{D'}} \mid S = b, Y = 0)| \leq \alpha. \tag{42}$$

Now, define the regions of agreement as $A = A_{\theta^*_{D'}} \cap A_{\theta^-_{D'}}, B = A^C_{\theta^*_{D'}} \cap A^C_{\theta^-_{D'}}$ and the regions of disagreement as $C = A^C_{\theta^*_{D'}} \cap A_{\theta^-_{D'}}, D = A_{\theta^*_{D'}} \cap A^C_{\theta^-_{D'}}$. These events are disjoint, so we can decompose our conditional probability measures in to sums over these events and express $P(A_{\theta^-_{D'}} \mid S, Y)$ in terms of $P(A_{\theta^*_{D'}} \mid S, Y)$

$$\begin{aligned} P(A_{\theta^*_{D'}} \mid S, Y) &= P(A \mid S, Y) + P(D \mid S, Y), \\ P(A_{\theta^-_{D'}} \mid S, Y) &= P(A \mid S, Y) + P(C \mid S, Y), \\ P(A_{\theta^-_{D'}} \mid S, Y) &= P(A_{\theta^*_{D'}} \mid S, Y) + P(C \mid S, Y) - P(D \mid S, Y). \end{aligned} \tag{43}$$

Then, we can bound the absolute difference:

$$\begin{aligned} &|P(A_{\theta^-_{D'}} \mid S = a, Y = 1) - P(A_{\theta^-_{D'}} \mid S = b, Y = 1)| \\ &\quad = |P(A_{\theta^*_{D'}} \mid S = a, Y = 1) - P(A_{\theta^*_{D'}} \mid S = b, Y = 1) \\ &\qquad + P(C \mid, S = a, Y = 1) - P(C \mid, S = b, Y = 1) \\ &\qquad - P(D \mid, S = a, Y = 1) + P(D \mid, S = b, Y = 1)|, \\ &\quad \leq |P(A_{\theta^*_{D'}} \mid S = a, Y = 1) - P(A_{\theta^*_{D'}} \mid S = b, Y = 1) \\ &\qquad + |P(C \mid, S = a, Y = 1) - P(C \mid, S = b, Y = 1)| \\ &\qquad + |P(D \mid, S = b, Y = 1) - P(D \mid, S = a, Y = 1)|, \\ &\quad \leq \alpha + |P(C \mid, S = a, Y = 1) - P(C \mid, S = b, Y = 1)| \\ &\qquad + |P(D \mid, S = b, Y = 1) - P(D \mid, S = a, Y = 1)|. \end{aligned} \tag{44}$$

Next, we relate the probability measures $P(\cdot \mid S, Y)$ to the size of their event spaces to establish a bound on the second and third terms in Equation 44. We assume for all events $x \subseteq \mathcal{X}$, $P(x|S, Y) \leq c\frac{|x|}{|\mathcal{X}|}$, or in other words, that our probability mass is relatively well distributed up to a constant. Then, computing $|C|$ and $|D|$ will provide upper bounds on their conditional likelihoods.

Recall our assumption that $||x_i||_2 \leq 1$. This means all of our data lies within a unit d-sphere we call S, which defines our event space $\mathcal{X}$ and has volume $V = |\mathcal{X}| = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$. We can define events $A_{\theta^*_{D'}}$ and $A_{\theta^-_{D'}}$ the halfspaces defined by $\theta^*_{D'}$ and $\theta^-_{D'}$ intersected with S. Then, $A, B, C,$ and D partition S into four disjoint volumes. We seek to bound the volume of C and D. First observe that the largest volume for $C + D$ occurs when $\theta^*_{D'}, \theta^-_{D'}$ intersect tangent to S and create a region where the angle between $\theta^*_{D'}$ and $\theta^-_{D'}$ is bisected by a hyperplane that also bisects S. This creates a region where one of C, D is empty and the other draws a large slice through S. Define

the angle between $\theta^*_{D'}$ and $\theta^-_{D'}$ as ϕ. Without loss of generality, assume $|D| = 0$. Then, by Lemmas 3 and 4, if $||\theta^*_{D'} - \theta^-_{D'}||_2 \leq d$, then ϕ is bounded and thus $|C|$ is bounded.

$$|C| = V - 2\int_{\mu}^{1} \frac{\pi^{\frac{d-1}{2}}}{\Gamma(\frac{d+1}{2})}(1-y^2)^{d-1}dy, \tag{45}$$

$$\phi \leq \arccos\left(\frac{||\theta^*_{D'}||_2 - \kappa}{||\theta^*_{D'}||_2 + \kappa}\right), \tag{46}$$

where

$$\begin{aligned}\mu &= \frac{1}{2}\sin\frac{\phi}{2},\\ &= \frac{1}{2}\sqrt{\frac{1-\cos\phi}{2}},\\ &\leq \frac{\sqrt{\kappa}}{2}.\end{aligned} \tag{47}$$

by the double angle formula for cosine.

Returning to Equation 44, we have

$$\begin{aligned}&|P(A_{\theta^-_{D'}} \mid S=a, Y=1) - P(A_{\theta^-_{D'}} \mid S=b, Y=1)|\\ &\quad\leq \alpha + |P(C\mid, S=a, Y=1) - P(C\mid, S=b, Y=1)|\\ &\quad+ |P(D\mid, S=b, Y=1) - P(D\mid, S=a, Y=1)|,\\ &\quad\leq \alpha + \max(P(C\mid, S=a, Y=1), P(C\mid, S=b, Y=1))\\ &\quad+ \max(P(D\mid, S=b, Y=1), P(D\mid, S=a, Y=1)),\\ &\quad\leq \alpha + c\frac{|C|}{|\mathcal{X}|} + c\frac{|D|}{|\mathcal{X}|},\\ &\quad= \alpha + c\frac{V - 2\int_{\mu}^{1}\frac{\pi^{\frac{d-1}{2}}}{\Gamma(\frac{d+1}{2})}(1-y^2)^{d-1}dy}{V},\\ &\quad= \alpha + c\left(1 - \frac{2\int_{\mu}^{1}\frac{\pi^{\frac{d-1}{2}}}{\Gamma(\frac{d+1}{2})}(1-y^2)^{d-1}dy}{V}\right),\end{aligned} \tag{48}$$

where

$$\mu \leq \frac{\sqrt{\kappa}}{2}.$$

By a symmetric argument, we also see

$$|P(A_{\theta^-_{D'}} \mid S=a, Y=0) - P(A_{\theta^-_{D'}} \mid S=b, Y=0)| \leq \alpha + c\left(1 - \frac{2\int_{\mu}^{1}\frac{\pi^{\frac{d-1}{2}}}{\Gamma(\frac{d+1}{2})}(1-y^2)^{d-1}dy}{V}\right). \tag{49}$$

And thus we have that the absolute mean equalized odds of $\theta^-_{D'}$ is bounded:

$$\begin{aligned}&\frac{1}{2}\big[|P(A_{\theta^-_{D'}} \mid S=a, Y=1) - P(A_{\theta^-_{D'}} \mid S=b, Y=1)|\\ &\qquad+ |P(A_{\theta^-_{D'}} \mid S=a, Y=1) - P(A_{\theta^-_{D'}} \mid S=b, Y=1)|\big],\\ &\qquad\leq \alpha + c\left(1 - \frac{2\int_{\mu}^{1}\frac{\pi^{\frac{d-1}{2}}}{\Gamma(\frac{d+1}{2})}(1-y^2)^{d-1}dy}{V}\right),\end{aligned} \tag{50}$$

where

$$h \leq \frac{\sqrt{\kappa}}{2}.$$

□

A.4 Incompatibility of existing unlearning methods.

Existing unlearning methods require a decomposition of the loss term as $\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\ell(\theta)$. For example, the most similar method to Fair Unlearning comes from Guo et al. (2020), and their unlearning step involves a second order approximation over $\Delta = \lambda\theta_D^* + \nabla\ell(\theta_D^*, x_n, y_n)$ to unlearn datapoint (x_n, y_n). Similarly to our proof of Lemma 1, adding Δ with the gradient of $\theta_{D'}^-$ over the remaining dataset $\mathcal{L}(\theta_{D'}^-, D')$ results in a cancellation that enables the rest of their proof to bound the gradient. However, in their case Δ is directly corresponds to the gradient over sample (x_n, y_n) only because of the decomposability of their loss term. This behavior makes all unlearning over decomposable loss functions the same: compute a Newton step $H^{-1}\Delta$ where H is the hessian over D' and Δ is the gradient over R. In our case our Δ is more complex because the fairness loss involves the interactions between each sample and samples in the opposite subgroup. Thus, removing one point requires removing a set of terms from the fairness loss, and thus Δ must be a more complex term to account for this rather than simply the gradient over (x_n, y_n).

A.5 Extension of fairness penalty to other definitions of group fairness.

Recall the definition of Equalized Odds:

$$\Pr(h_\theta(X) = 1 \mid S = a, Y = y) = \Pr(h_\theta(X) = 1 \mid S = b, Y = y). \tag{51}$$

In contrast, Demographic Parity does not condition on the true label, Y.

$$\Pr(h_\theta(X) = 1 \mid S = a) = \Pr(h_\theta(X) = 1 \mid S = b). \tag{52}$$

Thus, to achieve this, we remove the indicator variable in our loss function.

$$\mathcal{L}_{\text{Demo. Par.}}(\theta, D) := \left(\frac{1}{n_a n_b}\sum_{i\in G_a}\sum_{j\in G_b}(\langle x_i, \theta\rangle - \langle x_j, \theta\rangle)\right)^2. \tag{53}$$

Equality of opportunity conditions only on true positive labels, $Y = 1$.

$$\Pr(h_\theta(X) = 1 \mid S = a, Y = 1) = \Pr(h_\theta(X) = 1 \mid S = b, Y = 1). \tag{54}$$

Thus, to achieve this, we modify the indicator variable in our loss function to only consider true positives.

$$\mathcal{L}_{\text{Eq. Opp.}}(\theta, D) := \left(\frac{1}{n_a n_b}\sum_{i\in G_a}\sum_{j\in G_b}\mathbb{1}[y_i = y_j = 1](\langle x_i, \theta\rangle - \langle x_j, \theta\rangle)\right)^2. \tag{55}$$

A.6 Experiment details.

We run all of our experiments on a 32GB Tesla V100 GPU, but using tabular data and small models, these results can be replicated with weaker hardware. When training, we use $\sigma = 1$ for our objective perturbation and compute ϵ, δ bounds with $\delta = 0.0001$. We report the following hyperparameters for our final results in the main paper:

	ℓ_2 Penalty (λ)	Fairness Penalty (γ)
COMPAS	$1e-4$	1e1
Adult	$1e-4$	1e0
HSLS	$1e-2$	1e1

Table 1: Hyperparameters for experiments in paper.

A.7 Runtimes.

Below we include runtime comparisons (Table 2) for our method with naive retraining as well as precomputation times for COMPAS, Adult, and HSLS (Table 3). Our method achieves significant speedup over retraining a model.

Table 2: Runtime Comparisons for Proposed Method vs Retraining.

	COMPAS		Adult		HSLS	
Method	5% Unlearned	20% Unlearned	5% Unlearned	20% Unlearned	5% Unlearned	20% Unlearned
Retraining	3.500 ± 0.340 s	3.180 ± 0.142 s	36.796 ± 1.090 s	31.175 ± 1.040 s	14.963 ± 0.396 s	12.362 ± 0.795 s
Fair Unlearning (Ours)	0.170 ± 0.010 s	0.164 ± 0.003 s	8.800 ± 0.511 s	8.483 ± 0.938 s	6.014 ± 0.249 s	5.704 ± 0.015 s

Table 3: Precomputation Runtime Comparisons for Hessian Inversion.

	COMPAS	Adult	HSLS
Precomputation Time:	0.182 ± 0.006s	6.210 ± 0.041s	4.964 ± 0.117s

A.8 Test accuracy when unlearning.

In Tables 4, 5, and 6, we report test accuracy at various amounts of unlearning with standard deviation.

Table 4: Test accuracy when unlearning at random.

	COMPAS		Adult		HSLS	
Method	5% Unlearned	20% Unlearned	5% Unlearned	20% Unlearned	5% Unlearned	20% Unlearned
Full Training (BCE)	0.652 ± 0.000	0.652 ± 0.000	0.817 ± 0.000	0.816 ± 0.000	0.724 ± 0.000	0.724 ± 0.000
Retraining (BCE)	0.652 ± 0.001	0.654 ± 0.003	0.817 ± 0.000	0.817 ± 0.001	0.724 ± 0.002	0.724 ± 0.002
Newton (Guo et al. (2020))	0.652 ± 0.001	0.654 ± 0.003	0.817 ± 0.000	0.817 ± 0.001	0.724 ± 0.002	0.724 ± 0.002
SISA (Bourtoule et al. (2020))	0.653 ± 0.001	0.656 ± 0.004	0.817 ± 0.000	0.817 ± 0.001	0.725 ± 0.001	0.724 ± 0.003
PRU (Izzo et al. (2021))	0.688 ± 0.005	0.691 ± 0.001	0.810 ± 0.003	0.810 ± 0.001	0.726 ± 0.004	0.729 ± 0.000
Full Training (Fair)	0.647 ± 0.000	0.647 ± 0.000	0.819 ± 0.000	0.819 ± 0.000	0.713 ± 0.000	0.713 ± 0.000
Retraining (Fair)	0.646 ± 0.002	0.642 ± 0.003	0.819 ± 0.000	0.819 ± 0.001	0.713 ± 0.002	0.713 ± 0.002
Fair Unlearning (Ours)	0.646 ± 0.002	0.642 ± 0.003	0.819 ± 0.000	0.819 ± 0.001	0.713 ± 0.002	0.713 ± 0.002

Table 5: Test accuracy when unlearning from minority subgroups.

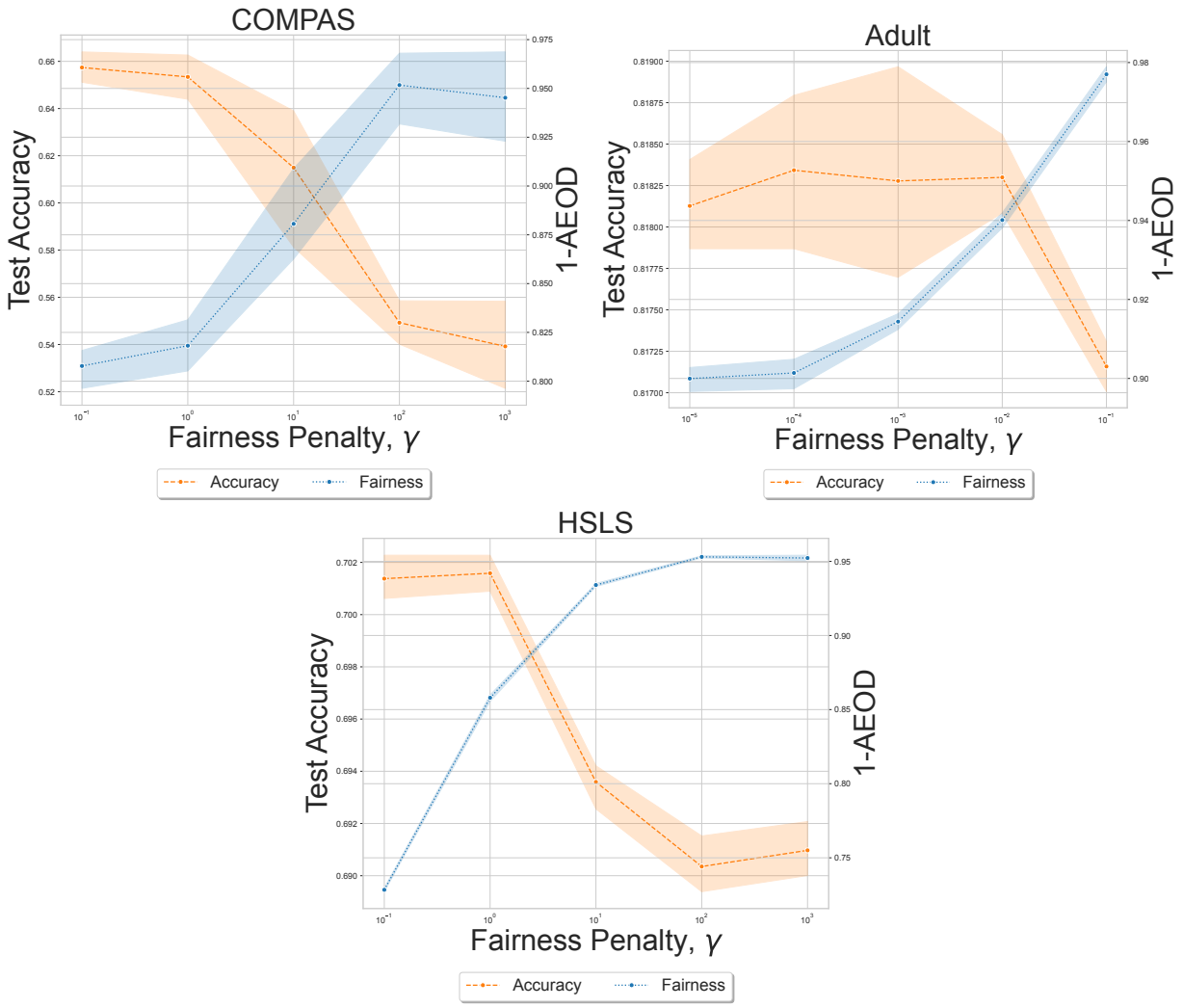
	COMPAS		Adult		HSLS	
Method	5% Unlearned	10% Unlearned	5% Unlearned	10% Unlearned	5% Unlearned	10% Unlearned
Full Training (BCE)	0.652 ± 0.000	0.652 ± 0.000	0.821 ± 0.000	0.821 ± 0.000	0.729 ± 0.000	0.729 ± 0.000
Retraining (BCE)	0.651 ± 0.002	0.652 ± 0.001	0.821 ± 0.000	0.822 ± 0.000	0.729 ± 0.001	0.730 ± 0.001
Newton (Guo et al. (2020))	0.651 ± 0.002	0.652 ± 0.001	0.821 ± 0.000	0.823 ± 0.000	0.729 ± 0.001	0.730 ± 0.002
SISA (Bourtoule et al. (2020))	0.653 ± 0.001	0.654 ± 0.002	0.822 ± 0.000	0.823 ± 0.000	0.728 ± 0.001	0.728 ± 0.001
PRU (Izzo et al. (2021))	0.661 ± 0.007	0.666 ± 0.007	0.819 ± 0.004	0.815 ± 0.035	0.720 ± 0.004	0.724 ± 0.003
Full Training (Fair)	0.653 ± 0.000	0.653 ± 0.000	0.812 ± 0.000	0.812 ± 0.000	0.723 ± 0.000	0.723 ± 0.000
Retraining (Fair)	0.653 ± 0.001	0.661 ± 0.005	0.813 ± 0.000	0.816 ± 0.001	0.713 ± 0.002	0.713 ± 0.002
Fair Unlearning (Ours)	0.652 ± 0.001	0.661 ± 0.004	0.814 ± 0.000	0.816 ± 0.002	0.724 ± 0.001	0.721 ± 0.001

Table 6: Test accuracy when unlearning from majority subgroups.

	COMPAS		Adult		HSLS	
Method	5% Unlearned	20% Unlearned	5% Unlearned	20% Unlearned	5% Unlearned	20% Unlearned
Full Training (BCE)	0.654 ± 0.000	0.654 ± 0.000	0.814 ± 0.000	0.814 ± 0.000	0.738 ± 0.000	0.738 ± 0.000
Retraining (BCE)	0.654 ± 0.001	0.652 ± 0.002	0.814 ± 0.000	0.814 ± 0.000	0.736 ± 0.000	0.736 ± 0.002
Newton (Guo et al. (2020))	0.654 ± 0.001	0.652 ± 0.002	0.814 ± 0.000	0.814 ± 0.001	0.736 ± 0.000	0.735 ± 0.002
SISA (Bourtoule et al. (2020))	0.654 ± 0.002	0.653 ± 0.004	0.814 ± 0.000	0.815 ± 0.001	0.734 ± 0.001	0.733 ± 0.002
PRU (Izzo et al. (2021))	0.696 ± 0.006	0.688 ± 0.003	0.812 ± 0.001	0.811 ± 0.002	0.723 ± 0.012	0.731 ± 0.004
Full Training (Fair)	0.666 ± 0.000	0.666 ± 0.000	0.815 ± 0.000	0.815 ± 0.000	0.725 ± 0.000	0.725 ± 0.000
Retraining (Fair)	0.663 ± 0.004	0.653 ± 0.002	0.815 ± 0.000	0.813 ± 0.000	0.724 ± 0.001	0.725 ± 0.002
Fair Unlearning (Ours)	0.663 ± 0.004	0.653 ± 0.003	0.815 ± 0.000	0.814 ± 0.001	0.724 ± 0.001	0.725 ± 0.002

A.9 Fairness-Accuracy Tradeoffs.

Below we report the impact of the fairness regularizer on the fairness and accuracy performance of the model without unlearning. We see that the fairness regularizer has a tangible impact and controls a tradeoff between the fairness and accuracy of the model. This motivates the selection of our fairness penalty, γ, as detailed in A.6 where we see a good balance between fairness and accuracy performance.



A.10 Figures for various fairness metrics.

Below we report figures for Demographic Parity, Equality of Opportunity, and Subgroup Accuracy. These results reflect the results shown in the main body of the paper, with the exception of Subgroup Accuracy which has more variable results, and sometimes worse results for fair methods. However, this is because our fairness regularizer optimizes for group fairness metrics such as Equalized Odds, Equality of Opportunity, and Demographic Parity which are metrics based on individual true and false positive rates rather than the raw test accuracy of each subgroup.

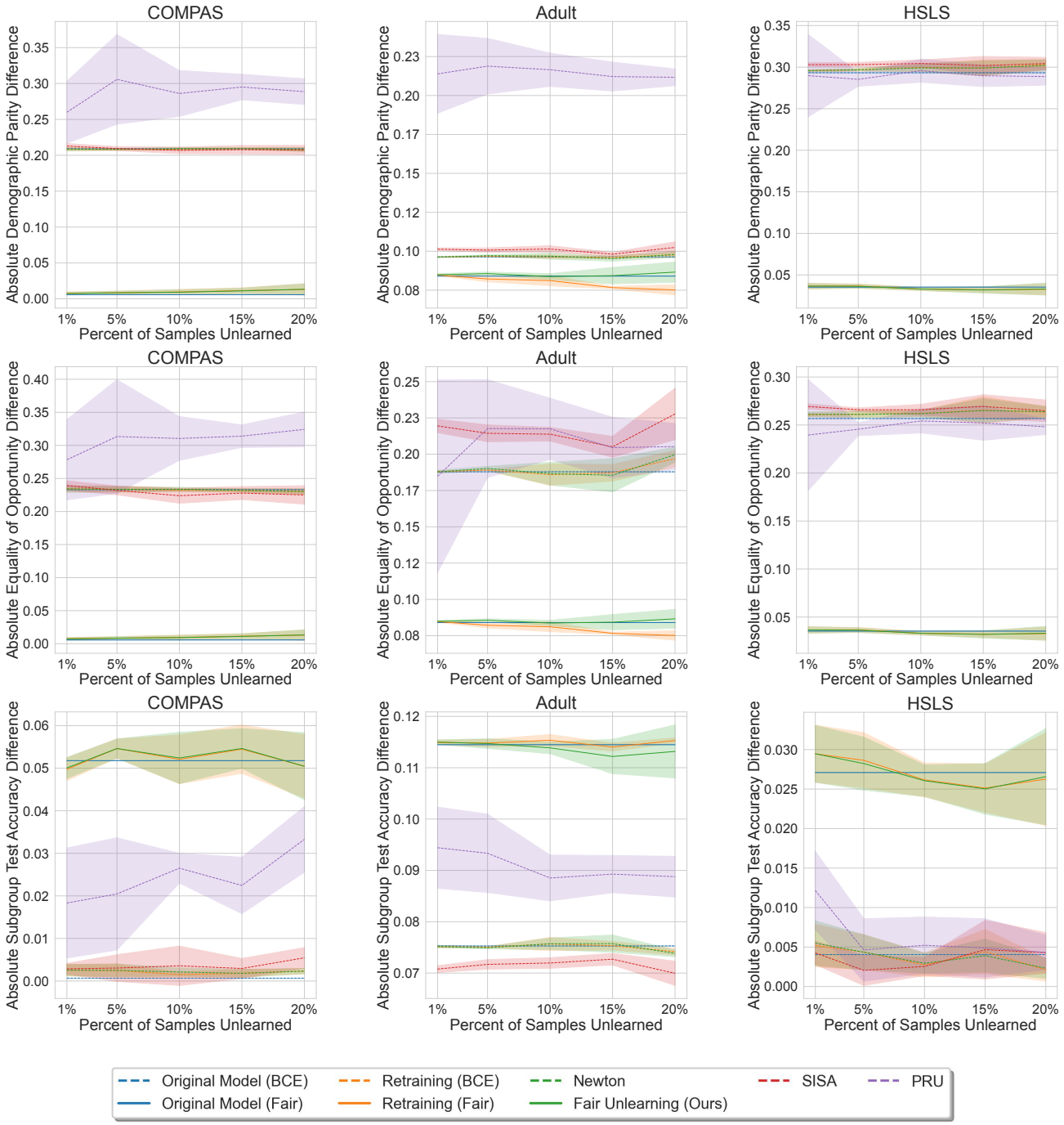


Figure 5: Absolute demographic parity (top), equality of opportunity (middle), and subgroup test accuracy (bottom) differences (lower is better) for unlearning methods over random requests on COMPAS, Adult, and HSLS.

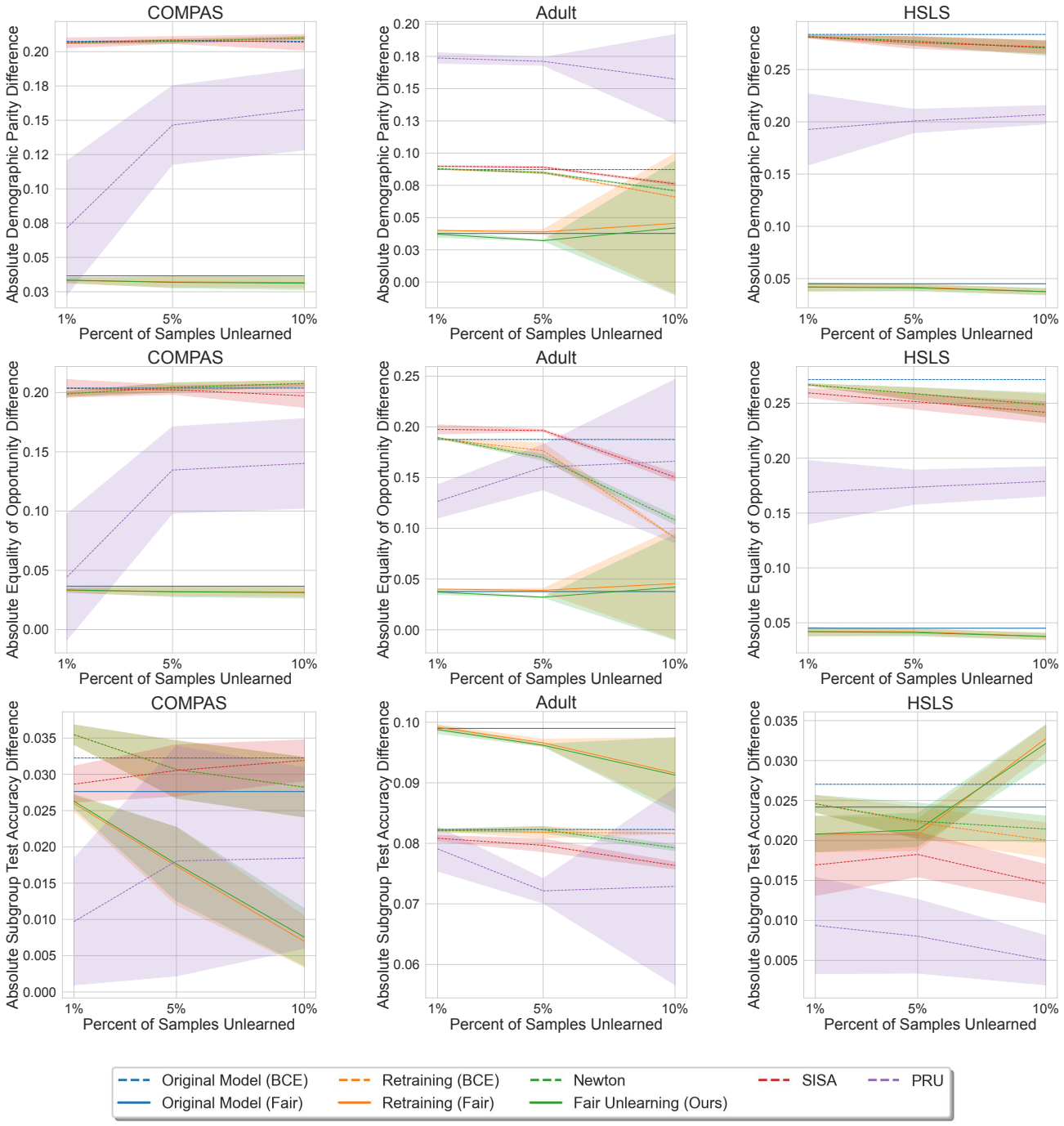


Figure 6: Absolute demographic parity (top), equality of opportunity (middle), and subgroup test accuracy (bottom) differences (lower is better) for unlearning methods when unlearning from the minority subgroup on COMPAS, Adult, and HSLS.

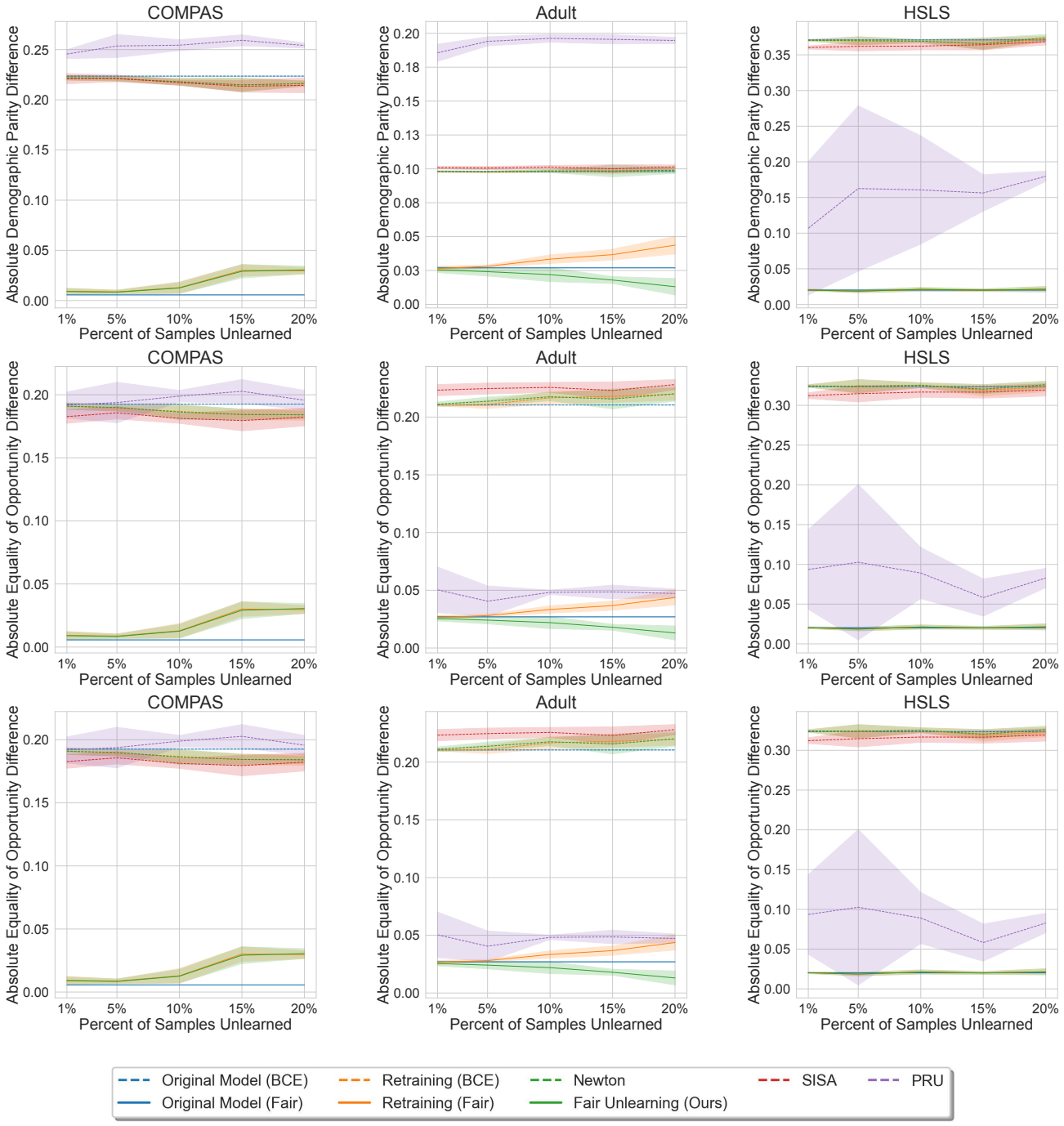


Figure 7: Absolute demographic parity (top), equality of opportunity (middle), and subgroup test accuracy (bottom) differences (lower is better) for unlearning methods when unlearning from the majority subgroup on COMPAS, Adult, and HSLS.