Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

A Flexible Fair Learning Framework via Group-Aware Surrogate Loss Reweighting

Wen Xu[†], Elham Dolatabadi[‡, ◇]
[†] University of Toronto
[‡] York University
[◇] Vector Institute

Abstract

This paper presents a new algorithmic fairness framework called $\boldsymbol{\alpha}$-$\boldsymbol{\beta}$ Fair Machine Learning ($\boldsymbol{\alpha}$-$\boldsymbol{\beta}$ FML), designed to optimize fairness levels across sociodemographic attributes. Our framework employs a new family of surrogate loss functions, paired with loss reweighting techniques, allowing precise control over fairness-accuracy trade-offs through tunable hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. To efficiently solve the learning objective, we propose Parallel Stochastic Gradient Descent with Surrogate Loss (P-SGD-S) and establish convergence guarantees for both convex and nonconvex loss functions. Experimental results demonstrate that our framework improves overall accuracy while reducing fairness violations, offering a smooth trade-off between standard empirical risk minimization and strict minimax fairness. Results across multiple datasets confirm its adaptability, ensuring fairness improvements without excessive performance degradation.

Keywords: Fairness, machine learning, surrogate loss, convergence analysis.

1. Introduction

Machine learning (ML) models are increasingly deployed in high-stakes domains such as healthcare, finance, and criminal justice, where disparities across protected groups can translate into unequal real-world outcomes [1]. A central challenge in fair ML is that fairness is inherently context dependent: Improving performance for underserved groups often trades off against average accuracy, and no single fairness notion is universally appropriate [1, 2]. This creates a need for learning frameworks that can flexibly navigate fairness-accuracy trade-offs rather than enforcing a single rigid operating point.

Existing approaches to fair ML prediction span pre-processing, in-processing, and post-processing. Pre-processing methods modify the training data before learning [3, 4]; post-processing methods adjust the outputs of an already trained predictor [5–8]; and in-processing methods incorporate fairness directly into the optimization procedure, often through fairness constraints, regularization, or worst-group objectives [9–12]. While effective, many of these methods optimize either fixed fairness criteria or extreme minimax objectives, offering limited flexibility in adapting to application-dependent fairness requirements. The closest work to ours is q-fair federated learning (q-FFL) [13], which also adopts a surrogate-loss perspective, but uses a uniform fairness parameter across groups and does not provide convergence analysis.

In this paper, we propose $\boldsymbol{\alpha}$-$\boldsymbol{\beta}$ Fair Machine Learning ($\boldsymbol{\alpha}$-$\boldsymbol{\beta}$ FML), a model-agnostic in-processing framework based on group-aware surrogate loss reweighting. Inspired by α-fair resource allocation in communication networks [14] and fairness-aware learning in federated settings [13], our formulation introduces group weights $\boldsymbol{\alpha}$ and group-specific surrogate parameters $\boldsymbol{\beta}$, enabling controlled interpolation between standard empirical risk minimization and increasingly fairness-oriented objectives, including minimax-style behavior in the limit. To solve the resulting problem, we develop Parallel Stochastic Gradient Descent with Surrogate Loss (P-SGD-S) and establish convergence guarantees for both convex and nonconvex losses. Experiments on protected-feature classification tasks show that our framework

achieves a favorable fairness–accuracy trade-off and can be tuned smoothly across operating points.

The main contributions of this work are as follows:

- We propose $\boldsymbol{\alpha}$-$\boldsymbol{\beta}$ FML, a model-agnostic fair learning framework that uses group-aware surrogate loss reweighting to flexibly control fairness across protected groups or imbalanced labels.
- We develop P-SGD-S for optimizing the proposed objective and provide convergence guarantees for both convex and nonconvex settings.
- We empirically show on multiple classification tasks that our framework improves the fairness-accuracy trade-off relative to standard ERM-style training and minimax baselines, while allowing smooth control through hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

2. Problem Setup and α-β Fair Objective

We consider a supervised binary classification problem with feature vector $Z = (A, X) \in \mathcal{Z}$, label $Y \in \{0, 1\}$, and predictor $\hat{Y} = h_{\boldsymbol{w}}(Z)$ parameterized by $\boldsymbol{w} \in \mathcal{W} \subseteq \mathbb{R}^d$. Here, $A \in \{0, 1\}$ denotes a binary protected attribute and X denotes the remaining features. Let $\mathcal{S}_a$ be the index set of training samples with $A = a$, and let $S_a = |\mathcal{S}_a|$ for $a \in \{0, 1\}$. We use a nonnegative loss $\ell(\boldsymbol{w}, (\boldsymbol{z}, y))$, and denote $\ell_j(\boldsymbol{w}) = \ell(\boldsymbol{w}, (\boldsymbol{z}_j, y_j))$.

To measure group disparity, we use equality of accuracy (EA) violation,

$$\epsilon_{\text{EA}} = \Big| P(\hat{Y} = Y \mid A = 0) - P(\hat{Y} = Y \mid A = 1) \Big|, \tag{2.1}$$

where smaller ϵ_{EA} indicates more balanced accuracy across groups. Other fairness metrics, including demographic parity and equality of opportunity, are reported in the appendix.

A standard baseline is empirical risk minimization (ERM):

$$\min_{\boldsymbol{w} \in \mathcal{W}} \frac{1}{N} \sum_{j=1}^{N} \ell_j(\boldsymbol{w}). \tag{2.2}$$

To obtain a tunable fairness–accuracy trade-off, we define the β-fair surrogate loss

$$f_\beta(\boldsymbol{w}, (\boldsymbol{z}, y)) = \frac{(1 + \ell(\boldsymbol{w}, (\boldsymbol{z}, y)))^{1+\beta}}{1 + \beta}, \qquad \beta \geq 0. \tag{2.3}$$

When $\beta > 0$, larger losses receive disproportionately higher penalties.

For group-specific parameters $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and weights $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$ satisfying $\alpha_0, \alpha_1 \geq 0$ and $\alpha_0 + \alpha_1 = 1$, define

$$F_{\beta_a}(\boldsymbol{w}) = \frac{1}{S_a} \sum_{i \in \mathcal{S}_a} f_{\beta_a}(\boldsymbol{w}, (\boldsymbol{z}_i, y_i)), \qquad a \in \{0, 1\}, \tag{2.4}$$

and optimize the group-aware objective

$$\min_{\boldsymbol{w} \in \mathcal{W}} L_{(\boldsymbol{\alpha}, \boldsymbol{\beta})}(\boldsymbol{w}) := \alpha_0 F_{\beta_0}(\boldsymbol{w}) + \alpha_1 F_{\beta_1}(\boldsymbol{w}). \tag{2.5}$$

Here, $\boldsymbol{\alpha}$ controls the explicit relative importance of the groups, while $\boldsymbol{\beta}$ controls how strongly high-loss samples are emphasized within each group.

This formulation recovers several important special cases. Setting $\alpha_a = S_a/N$ and $\beta_0 = \beta_1 = 0$ reduces equation 2.5 to ERM in equation 2.2. If $\beta_0 = \beta_1 = q > 0$ and the offset 1 in equation 2.3 is removed, the objective matches the surrogate form of q-FFL [13] for two groups. As $\beta_0, \beta_1 \to \infty$ with $\alpha_0, \alpha_1 > 0$, the objective increasingly emphasizes the worst losses, approaching minimax-style fairness [15]. Besides. our formulation can be further extended to include multiple protected features and categorical features beyond binary.

Algorithm 1: Parallel Stochastic Gradient Descent with Surrogate Loss (P-SGD-S)

Input: learning rate $\gamma_i^{(t)}$ for $i \in \mathcal{I}$, hyperparameters α_i and β_i for $i \in \mathcal{I}$, and training rounds T.
Output: $\{\boldsymbol{w}^{(t)}\}_{t=0}^{T}$.
1: Initialize $\boldsymbol{w}^{(0)} \in \mathcal{W}$.
2: **for** each round $t = 0, \ldots, T-1$ **do**
3: **for** each $i \in \mathcal{I}$ **do**
4: Sample $\xi_i^{(t)} \subseteq \mathcal{S}_i$ uniformly at random.
5: Obtain stochastic gradient $\tilde{\nabla} F_{\beta_i}(\boldsymbol{w}^{(t)})$ via equation 3.1.
6: Update the model via equation 3.2.
7: **end for**
8: Aggregate the model via equation 3.3.
9: **end for**

3. Algorithm Design and Convergence Analysis

To solve equation 2.5, we use Parallel Stochastic Gradient Descent with Surrogate Loss (P-SGD-S). At round t, for each group $a \in \{0, 1\}$, we sample a mini-batch $\xi_a^{(t)} \subseteq \mathcal{S}_a$ uniformly at random and compute

$$\tilde{\nabla} F_{\beta_a}(\boldsymbol{w}^{(t)}; \xi_a^{(t)}) = \frac{1}{|\xi_a^{(t)}|} \sum_{j \in \xi_a^{(t)}} (1 + \ell_j(\boldsymbol{w}^{(t)}))^{\beta_a} \nabla \ell_j(\boldsymbol{w}^{(t)}). \tag{3.1}$$

We then perform a group-wise update

$$\boldsymbol{w}_a^{(t+1)} = \boldsymbol{w}^{(t)} - \eta\, \tilde{\nabla} F_{\beta_a}(\boldsymbol{w}^{(t)}; \xi_a^{(t)}), \tag{3.2}$$

followed by aggregation

$$\boldsymbol{w}^{(t+1)} = \alpha_0 \boldsymbol{w}_0^{(t+1)} + \alpha_1 \boldsymbol{w}_1^{(t+1)}. \tag{3.3}$$

Equivalently,

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta \sum_{a=0}^{1} \alpha_a \tilde{\nabla} F_{\beta_a}(\boldsymbol{w}^{(t)}; \xi_a^{(t)}). \tag{3.4}$$

Thus, the two group-specific updates can be computed in parallel before aggregation. The pseudo-code of P-SGD-S is shown in Algorithm 1.

We analyze P-SGD-S under standard assumptions: $L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}$ is L_0-smooth, each $\tilde{\nabla} F_{\beta_a}$ is an unbiased estimator of ∇F_{β_a}, and the stochastic gradient variance is bounded by σ^2.

Theorem 1. *Let $\{\boldsymbol{w}^{(t)}\}_{t=0}^{T-1}$ be generated by P-SGD-S with stepsize $\eta \le 1/L_0$.*

Convex case. *If each ℓ_j is convex and $\hat{\boldsymbol{w}} = \frac{1}{T}\sum_{t=0}^{T-1} \boldsymbol{w}^{(t)}$, then*

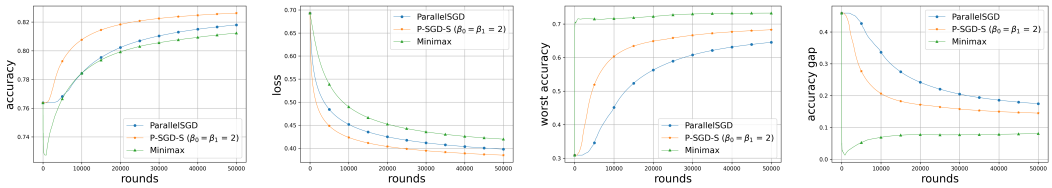
$$\mathbb{E}\big[L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\hat{\boldsymbol{w}})\big] - L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^*) \le \frac{\|\boldsymbol{w}^{(0)} - \boldsymbol{w}^*\|^2}{2\eta T} + \eta\sigma^2, \tag{3.5}$$

where $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w} \in \mathcal{W}} L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w})$.

Nonconvex case. *Let $L^* = \inf_{\boldsymbol{w} \in \mathcal{W}} L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w})$. Then*

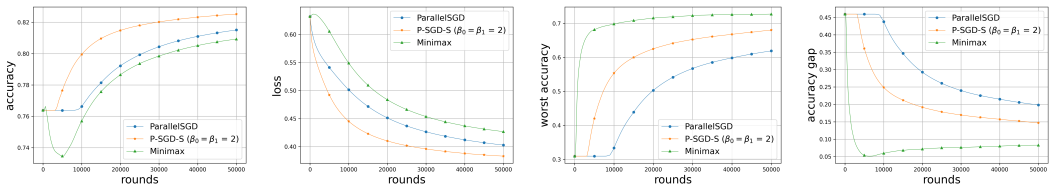
$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\Big[\|\nabla L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)})\|^2\Big] \le \frac{2\big(L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(0)}) - L^*\big)}{\eta T} + L_0\eta\sigma^2. \tag{3.6}$$

Choosing $\eta = \Theta(T^{-1/2})$ yields the standard $O(T^{-1/2})$ convergence rate in both settings. Proofs are deferred to the Appendix A.



(a) Overall test accuracy *(b)* Overall test loss *(c)* Worst Accuracy *(d)* EA violation

Figure 1. Comparison of accuracy, loss, worst accuracy, and ϵ_{EA} for convex loss on Adult. Parallel SGD operates without explicit fairness constraints, while Minimax enforces an extreme fairness constraint by prioritizing the worst-performing group.



(a) Overall test accuracy *(b)* Overall test loss *(c)* Worst Accuracy *(d)* EA violation

Figure 2. Comparison of accuracy, loss, worst accuracy, and ϵ_{EA} for nonconvex loss on Adult. Parallel SGD operates without explicit fairness constraints, while Minimax enforces an extreme fairness constraint by prioritizing the worst-performing group.

4. **Experiments**

We use the Adult dataset [16], focusing on education-level disparities in economic opportunity, and the COMPAS dataset [17], examining sex-based disparities in criminal risk assessment. We consider Doctorate (*group* 0) and non-Doctorate (*group* 1) as protected groups for Adult while we consider females (*group* 0) and males (*group* 1) as protected groups for COMPAS. Both datasets reflect real-world sociodemographic inequities. Our experiments cover both convex and non-convex classification tasks, solving for the corresponding loss functions. For both datasets, models are trained for $50,000$ iterations.

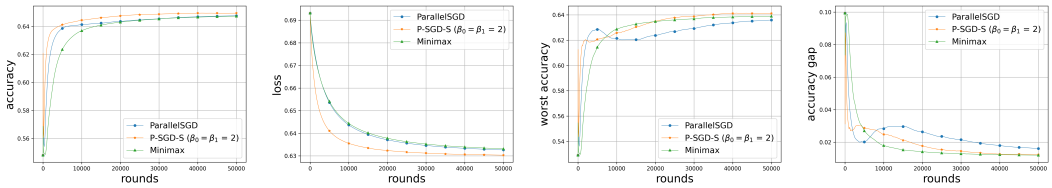
We consider two classes of benchmarks as follows:

- Surrogate-Min: solving the minimization optimization problem in equation 2.5 using our proposed P-SGD-S as outlined in Algorithm 1;
- Minimax: solving the minimax optimization problem proposed in [15], i.e., minimizing the loss for the worst mixture of all group distributions,

$$\min_{\boldsymbol{w}\in\mathcal{W}} \max_{\boldsymbol{\lambda}\in\Lambda} \sum_{i\in\mathcal{I}} \lambda_i F_{\beta_i}(w), \tag{4.1}$$

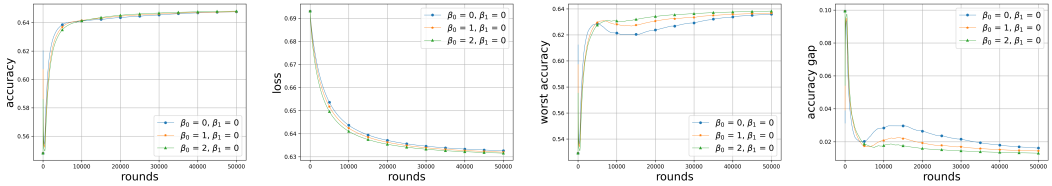
where $\boldsymbol{\lambda} \in \Lambda \subseteq \Delta_{|\mathcal{I}|-1}$ and $\beta_i = 0, \forall i \in \mathcal{I}$, using a stochastic gradient descent ascent (SGDA) algorithm.

We choose specific values for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in the Surrogate-Min class of algorithms. The original ERM formulation is a special case of the Surrogate-Min class, where $\alpha_0 = S_0/N$, $\alpha_1 = S_1/N$, and $\beta_0 = \beta_1 = 0$. For clarity of comparison, we call the corresponding algorithm Parallel SGD. For the Minimax type of algorithms, by solving the minimax optimization problem in equation 4.1, the resulting model tends to have uniform accuracy among different groups, i.e., imposing the most extreme min-max fairness on accuracy. We evaluate model performance using the following four metrics on the test set: average test accuracy, the worst group accuracy (the lowest accuracy between the two protected groups), and the EA violation. More results on other fairness notions [5, 9] are deferred to Appendix B.



(a) Overall test accuracy *(b)* Overall test loss *(c)* Worst Accuracy *(d)* EA violation

Figure 3. Comparison of accuracy, loss, worst accuracy, and ϵ_{EA} for convex loss on COMPAS. PARALLEL SGD operates without explicit fairness constraints, while MINIMAX enforces an extreme fairness constraint by prioritizing the worst-performing group in group fairness.



(a) Overall test accuracy *(b)* Overall test loss *(c)* Worst Accuracy *(d)* EA violation

Figure 4. Impact of varying $\boldsymbol{\beta}$ values on training dynamics for convex loss on COMPAS. With $\beta_1 = 0$, increasing β_0 prioritizes the minority group, improving worst-group accuracy and reducing ϵ_{EA}.

4.1. **Balanced Fairness with Improved Accuracy**

To evaluate our framework against benchmark methods, we set $\beta_1 = \beta_0 = 2$. The results for the convex loss case for Adult dataset are presented in Figure 1, while those for the COMPAS dataset are shown in Figure 3. For the nonconvex loss case, the corresponding results are provided in Figure 2. Our approach demonstrates superior average accuracy compared to both the PARALLEL SGD and MINIMAX (Figure 1a and Figure 3a), indicating that our fairness-aware utility function is being properly utilized. More importantly, our method improves worst-group accuracy compared to PARALLEL SGD while remaining slightly lower than MINIMAX (Figure 1c and Figure 3c). This aligns with our expectations, as MINIMAX explicitly prioritizes worst-case group performance at the cost of overall accuracy, making it a more extreme approach. In contrast, our method achieves a more balanced trade-off, enhancing fairness without imposing overly restrictive constraints on model performance. A similar trend is observed for EA violation, where our approach effectively reduces disparities compared to PARALLEL SGD but does not enforce as strict a correction as MINIMAX (Figure 1d and Figure 3d).

4.2. **Flexibility in Fairness Optimization**

We demonstrate the adaptability of our framework by varying $\boldsymbol{\beta}$ values on training dynamics. The results are shown in Figure 4 for COMPAS. Specifically, we fix $\beta_1 = 0$ and vary β_0, effectively shifting focus toward the minority group. For instance, in the COMPAS dataset, where Females are the minority, adjusting β_0 redistributes optimization focus toward this group. As expected, increasing β_0 enhances our fairness metric by improving worst-group accuracy (Figures 4c) while also reducing EA violation (Figures 4d).

These results demonstrate that our approach is highly adaptable and can be tuned to align with different fairness definitions and objectives. By simply adjusting a single parameter β_0, our method offers a smooth fairness-performance trade-off, spanning from PARALLEL SGD

(no explicit fairness constraints) to MINIMAX fairness (maximizing worst-group performance at the cost of overall accuracy).

5. Conclusion

Our reweighting techniques for both convex and nonconvex settings effectively address sociodemographic disparities in classification tasks while maintaining strong predictive performance. By tuning fairness parameters, our method provides a flexible trade-off between no fairness constraints and strict fairness enforcement, allowing practitioners to adapt fairness interventions to specific contexts without imposing overly rigid constraints. Future work will extend this framework to intersectional fairness analysis, more complex multimodal datasets, and integration with state-of-the-art deep learning models, including transformers, to further enhance its applicability in real-world decision-making systems.

References

[1] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and machine learning: limitations and opportunities*. fairmlbook.org, 2019.

[2] A. D. Selbst, D. Boyd, S. A. Friedler, et al. "Fairness and abstraction in sociotechnical systems". In: *Proceedings of FAccT*. 2019, pp. 59–68.

[3] N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi, et al. "The case for process fairness in learning: Feature selection for fair decision making". In: *NIPS Symposium on Machine Learning and the Law*. 2016.

[4] F. Calmon, D. Wei, B. Vinzamuri, et al. "Optimized pre-processing for discrimination prevention". In: *Proceedings of NeurIPS*. 2017.

[5] M. Hardt, E. Price, and N. Srebro. "Equality of opportunity in supervised learning". In: *Proceedings of NeurIPS*. 2016.

[6] B. Fish, J. Kun, and Á. D. Lelkes. "A confidence-based approach for balancing fairness and accuracy". In: *Proceedings of SDM*. 2016.

[7] C. Dwork, N. Immorlica, A. T. Kalai, et al. "Decoupled classifiers for group-fair and efficient machine learning". In: *Proceedings of FAccT*. 2018.

[8] W. Alghamdi, H. Hsu, H. Jeong, et al. "Beyond Adult and COMPAS: Fair multi-class prediction via information projection". In: *Proceedings of NeurIPS*. 2022.

[9] M. B. Zafar, I. Valera, M. G. Rodriguez, et al. "Fairness constraints: Mechanisms for fair classification". In: *Proceedings of AISTATS*. 2017.

[10] S. Baharlouei, M. Nouiehed, A. Beirami, et al. "Rényi fair inference". In: *Proceedings of ICLR*. 2020.

[11] M. Lohaus, M. Perrot, and U. Von Luxburg. "Too relaxed to be fair". In: *Proceedings of ICML*. 2020.

[12] W. Yao, Z. Zhou, Z. Li, et al. "Understanding fairness surrogate functions in algorithmic fairness". In: *Transactions on Machine Learning Research* (2024).

[13] T. Li, M. Sanjabi, A. Beirami, et al. "Fair resource allocation in federated learning". In: *Proceedings of ICLR*. 2020.

[14] J. Mo and J. Walrand. "Fair end-to-end window-based congestion control". In: *IEEE/ACM Transactions on Networking* 8.05 (2000), pp. 556–567.

[15] T. Hashimoto, M. Srivastava, H. Namkoong, et al. "Fairness without demographics in repeated loss minimization". In: *Proceedings of ICML*. 2018.

[16] B. Becker and R. Kohavi. *Adult*. UCI Machine Learning Repository. 1996.

[17] J. Larson, S. Mattu, L. Kirchner, et al. *How we analyzed the COMPAS recidivism algorithm*. URL: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. 2016.

In the following appendices, we provide proof details and additional experiments.

Appendix A. Proof of Theorem 1

Proof. We first prove the result of the convex case. From the L_0-smoothness of the surrogate loss, we have

$$L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t+1)}) - L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)}) \leq \nabla L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)})^\top(\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^{(t)}) + \frac{L_0}{2}\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^{(t)}\|^2 \tag{A.1}$$

$$= -\gamma \nabla L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)})^\top(\sum_{i\in\mathcal{I}} \alpha_i \tilde{\nabla} F_{\beta_i}(\boldsymbol{w}^{(t)}; \xi_i^{(t)})) + \frac{L_0\gamma^2}{2}\|\sum_{i\in\mathcal{I}} \alpha_i \tilde{\nabla} F_{\beta_i}(\boldsymbol{w}^{(t)}; \xi_i^{(t)})\|^2, \tag{A.2}$$

where we substitute in the update rule in equation A.2. Taking expectations on both sides, we have

$$\begin{aligned}
&\mathbb{E}[L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t+1)})] - L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)}) \\
&\leq -\gamma\|\nabla L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)})\|^2 + \frac{L_0\gamma^2}{2}\mathbb{E}[\|\sum_{i\in\mathcal{I}} \alpha_i \tilde{\nabla} F_{\beta_i}(\boldsymbol{w}^{(t)}; \xi_i^{(t)}) - \nabla L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)}) + \nabla L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)})\|^2] && \text{(A.3)} \\
&= (-\gamma + \frac{L_0\gamma^2}{2})\|\nabla L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)})\|^2 + \frac{L_0\gamma^2}{2}\mathbb{E}[\|\sum_{i\in\mathcal{I}} \alpha_i \tilde{\nabla} F_{\beta_i}(\boldsymbol{w}^{(t)}; \xi_i^{(t)}) - \nabla L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)})\|^2] && \text{(A.4)} \\
&\leq -\frac{\gamma}{2}\|\nabla L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)})\|^2 + \frac{L_0\gamma^2\sigma^2}{2} && \text{(A.5)} \\
&\leq -\frac{\gamma}{2}\|\nabla L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)})\|^2 + \frac{\gamma\sigma^2}{2}, && \text{(A.6)}
\end{aligned}$$

where equation A.3 is by the unbiased stochastic gradient assumption, equation A.5 is by assuming $\gamma \leq 1/L_0$ and the bounded variance assumption, and equation A.6 is by assuming $\gamma \leq 1/L_0$ again. From the update rule, we have

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^*\|^2] \\
&= \mathbb{E}[\|\boldsymbol{w}^{(t)} - \gamma(\sum_{i\in\mathcal{I}} \alpha_i \tilde{\nabla} F_{\beta_i}(\boldsymbol{w}^{(t)}; \xi_i^{(t)})) - \boldsymbol{w}^*\|^2] && \text{(A.7)} \\
&= A^{(t)} - 2\gamma\nabla L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)})^\top(\boldsymbol{w}^{(t)} - \boldsymbol{w}^*) + \gamma^2\mathbb{E}[\|\sum_{i\in\mathcal{I}} \alpha_i \tilde{\nabla} F_{\beta_i}(\boldsymbol{w}^{(t)}; \xi_i^{(t)})\|^2] && \text{(A.8)} \\
&\leq A^{(t)} + 2\gamma(L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^*) - L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)})) + \gamma^2\mathbb{E}[\|\sum_{i\in\mathcal{I}} \alpha_i \tilde{\nabla} F_{\beta_i}(\boldsymbol{w}^{(t)}; \xi_i^{(t)})\|^2] && \text{(A.9)} \\
&= A^{(t)} + 2\gamma(L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^*) - L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)})) + \gamma^2\mathbb{E}[\|\sum_{i\in\mathcal{I}} \alpha_i \tilde{\nabla} F_{\beta_i}(\boldsymbol{w}^{(t)}; \xi_i^{(t)}) - \nabla L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)}) + \nabla L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)})\|^2] && \text{(A.10)} \\
&= A^{(t)} + 2\gamma(L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^*) - L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)})) + \gamma^2(\sigma^2 + \|\nabla L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)})\|^2) && \text{(A.11)} \\
&\leq A^{(t)} + 2\gamma(L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^*) - L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)})) + \gamma^2(2\sigma^2 + \frac{2}{\gamma}(L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)}) - \mathbb{E}[L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t+1)})])) && \text{(A.12)} \\
&= A^{(t)} + 2\gamma(L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^*) - \mathbb{E}[L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t+1)})]) + 2\gamma^2\sigma^2, && \text{(A.13)}
\end{aligned}$$

where equation A.8 is by expanding the square and defining $A^{(t)} = \|\boldsymbol{w}^{(t)} - \boldsymbol{w}^*\|^2$, equation A.9 is by the convexity of $L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}$ and equation A.12 is by equation A.6. Rearranging terms in equation A.13, we have

$$\mathbb{E}[L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t+1)})] - L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^*) \leq \frac{1}{2\gamma}(A^{(t)} - \mathbb{E}[A^{(t+1)}]) + \gamma\sigma^2. \tag{A.14}$$

Summing equation A.14 over t from 0 to $T-1$, taking total expectation, and dividing both sides by T, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t+1)})] - L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^*) &\leq \frac{1}{2\gamma T}(\|\boldsymbol{w}^{(0)} - \boldsymbol{w}^*\|^2 - \mathbb{E}[\|\boldsymbol{w}^{(T)} - \boldsymbol{w}^*\|^2]) + \gamma\sigma^2 && \text{(A.15)} \\
&\leq \frac{1}{2\gamma T}(\|\boldsymbol{w}^{(0)} - \boldsymbol{w}^*\|^2) + \gamma\sigma^2. && \text{(A.16)}
\end{aligned}$$

As $L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}$ is a convex function, we have

$$\mathbb{E}[L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\frac{1}{T}\sum_{t=0}^{T-1}\boldsymbol{w}^{(t)})] - L(\boldsymbol{w}^*) \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t+1)})] - L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^*), \tag{A.17}$$

which completes the proof for the convex case.

Table 1. Comparison on different fairness metrics for convex and nonconvex loss on Adult.

	Convex Loss			Nonconvex Loss		
	$\epsilon_{\text{EA}} \downarrow$	$\epsilon_{\text{DP}} \downarrow$	$\epsilon_{\text{EO}} \downarrow$	$\epsilon_{\text{EA}} \downarrow$	$\epsilon_{\text{DP}} \downarrow$	$\epsilon_{\text{EO}} \downarrow$
SURROGATE-MIN						
$\beta_0 = \beta_1 = 0$	0.1747	**0.2846**	**0.1043**	0.1983	**0.3526**	**0.1564**
$\beta_0 = \beta_1 = 1$	0.1589	0.3147	0.1309	0.1712	0.3922	0.1953
$\beta_0 = \beta_1 = 2$	0.1452	0.3451	0.1596	0.1466	0.4248	0.2295
MINIMAX	**0.0809**	0.4827	0.3216	**0.0829**	0.4949	0.3179

Now we derive the result for the non-convex case. We derive the bound using the same proof in the convex case up to equation A.5.

$$\mathbb{E}[L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t+1)})] - L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)}) \leq -\frac{\gamma}{2}\|\nabla L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)})\|^2 + \frac{L_0\gamma^2\sigma^2}{2}. \tag{A.18}$$

Summing equation A.18 over t from 0 to $T-1$, taking total expectation, and dividing both sides by T, we have

$$\mathbb{E}[L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(T)})] - L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(0)}) \leq -\frac{\gamma}{2T}\sum_{t=0}^{T-1}\|\nabla L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)})\|^2 + \frac{L_0\gamma^2\sigma^2}{2}. \tag{A.19}$$

Rearranging terms, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(t)})\|^2 \leq \frac{2}{\gamma T}(L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(0)}) - \mathbb{E}[L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(T)})]) + L_0\gamma\sigma^2 \leq \frac{2}{\gamma T}(L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w}^{(0)}) - L^*) + L_0\gamma\sigma^2, \tag{A.20}$$

where $L^* = \min_{\boldsymbol{w}\in\mathcal{W}} L_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{w})$. □

Appendix B. Additional Experiments

In this section, we provide additional experimental details for Adult and COMPAS.

Adult. The Adult dataset contains 48,842 samples, split into 32,561 training and 16,281 test examples for a salary prediction task in the feature-partitioned setting. We use education level as the protected feature, defining Doctorate (*group* 0) and non-Doctorate (*group* 1) as the two groups. The training set contains 413 and 32,148 samples from the two groups, and the test set contains 181 and 16,100 samples, respectively. We use only categorical features. Due to the strong imbalance, we set $\boldsymbol{\alpha}$ according to the training-group proportions, i.e., $\alpha_0 = 0.0127$ and $\alpha_1 = 0.9873$. We assign β_0 and β_1 to the Doctorate and non-Doctorate groups, respectively. For the convex case of SURROGATE-MIN, we train logistic regression with cross-entropy loss; for the nonconvex case, we train a neural network with one hidden layer of 10 neurons and ReLU activation.

COMPAS. The COMPAS dataset contains 6,172 samples with 14 features. We use the same 2/3-1/3 train-test split as for Adult and take sex as the protected feature, with Females (*group* 0) and Males (*group* 1) as the two groups. The training set contains 782 and 3,332 samples, and the test set contains 393 and 1,665 samples, respectively. As in Adult, we use only categorical features. We set $\boldsymbol{\alpha}$ according to the training-group proportions, i.e., $\alpha_0 = 0.1901$ and $\alpha_1 = 0.8099$. We assign β_0 and β_1 to the Female and Male groups, respectively. For the convex case of SURROGATE-MIN, we train logistic regression with cross-entropy loss; for the nonconvex case, we train a multi-layer neural network.

We further consider other popular fairness notions. The Demographic Parity (DP) [9] violation is defined as

$$\epsilon_{\text{DP}} = |P(\hat{Y} = 1|A = 0) - P(\hat{Y} = 1|A = 1)|, \tag{B.1}$$

and the Equality of Opportunity (EO) [5] violation ϵ_{EO} is defined as

$$\epsilon_{\text{EO}} = |P(\hat{Y} = 1|Y = 1, A = 0) - P(\hat{Y} = 1|Y = 1, A = 1)|. \tag{B.2}$$

Based on the definitions of ϵ_{EA}, ϵ_{DP}, ϵ_{EO}, it is clear that a lower value for each metric means a greater degree of fairness imposed by the model.

The results of different fairness metrics for both convex and nonconvex loss of various $\boldsymbol{\beta}$ under minimization optimization are shown in Table 1. It is evident that for both convex and nonconvex loss, with higher $\boldsymbol{\beta}$, the learned model tends to have a smaller EA violation, yet larger DP violation and EO violation in Table 1. This observation matches with the argument in [1] that those fairness metrics can be contradictory. In the extreme case of MINIMAX, the EA violation is very low, meaning the accuracy of each group is almost uniform. Our SURROGATE MIN class of algorithms with different values of $\boldsymbol{\beta}$ can obtain different levels of fairness with respect to those fairness metrics between the vanilla ERM and MINIMAX.