APFEx: Adaptive Pareto Front Explorer for Intersectional Fairness

Anonymous authors
Paper under double-blind review

Abstract

Ensuring fairness in machine learning models is critical, especially when biases compound across intersecting protected attributes like race, gender, and age. While existing methods address fairness for single attributes, they fail to capture the nuanced, multiplicative biases faced by intersectional subgroups. We introduce *Adaptive Pareto Front Explorer (APFEx)*, the first framework to explicitly model intersectional fairness as a joint optimization problem over the Cartesian product of sensitive attributes. APFEx combines three key innovations: (1) an adaptive multi-objective optimizer that dynamically switches between Pareto cone projection, gradient weighting, and exploration strategies to navigate fairness-accuracy trade-offs; (2) differentiable intersectional fairness metrics enabling gradient-based optimization of non-smooth subgroup disparities; and (3) theoretical guarantees of convergence to Pareto-optimal solutions. Experiments on real-world datasets demonstrate APFEx's superiority, reducing fairness violations while maintaining competitive accuracy. Our work bridges a critical gap in fair ML, providing a scalable, model-agnostic solution for intersectional fairness.

1 Introduction

Beyond Accuracy—Fairness at the Intersections. When a healthcare algorithm used across U.S. hospitals was found to systematically prioritize healthier white patients over sicker Black patients for life-saving care, it exposed a harsh truth: even accurate machine learning models can fail basic tests of fairness (Barocas et al., 2023). Such cases reveal how biases embedded in data—whether based on race, gender, or income—can perpetuate discrimination when deployed at scale. But the problem runs deeper. Consider the COMPAS recidivism tool: while it appeared marginally fair when examining race or gender separately, Black women were disproportionately misclassified as high-risk, facing compounded bias at the intersection of these identities (Angwin et al., 2022). This phenomenon—where discrimination multiplies across overlapping attributes like race, gender, and age—is the crux of *intersectional fairness*, a challenge single-attribute approaches cannot solve (Chen et al., 2024). As AI systems increasingly govern access to loans, healthcare, and employment, ensuring fairness for *all* individuals, not just broad demographic groups, becomes both a technical necessity and an ethical imperative.

The Shortcomings of Current Fairness Techniques. To address fairness challenges, the ML community has introduced various criteria—such as demographic parity, equalized odds, and equality of opportunity—each grounded in distinct normative principles of equitable treatment (Barocas et al., 2023; Zafar et al., 2017). These have given rise to three main approaches: (1) *pre-processing* methods that adjust input data distributions to reduce bias (Kamiran & Calders, 2012; Calmon et al., 2017; Sattigeri et al., 2019), (2) *in-processing* techniques that incorporate fairness constraints directly into model training (Zafar et al., 2017; Celis et al., 2019; Lohaus et al., 2020), and (3) *post-processing* strategies that modify model outputs to achieve fairness (Hardt et al., 2016; Chzhen et al., 2019; Chierichetti et al., 2017). Despite significant progress, these methods predominantly adopt a *single-attribute* perspective, treating sensitive features (e.g., race, gender) in isolation and enforcing fairness constraints independently using additive penalties. This reductionist framing fails in *multi-attribute* contexts, where bias arises from the intersection of protected dimensions. As shown by (Angwin et al., 2022), models that appear fair when evaluated separately on race

and gender can still severely discriminate against intersectional groups, such as Black women. The core issue is that single-attribute fairness lacks the representational capacity to capture complex, higher-order discrimination patterns—precisely where the most harmful biases often occur.

The challenge of intersectional fairness is inherently multi-objective. Algorithmic fairness requires optimizing predictive accuracy alongside fairness metrics that often conflict (Martinez et al., 2020). Recent work has framed this as a multi-objective problem: Valdivia et al. (2021) use evolutionary algorithms; Milojkovic et al. (2019) and Cotter et al. (2019) adopt multi-gradient and game-theoretic methods—but all treat sensitive attributes independently and rely on scalarized losses, obscuring intersectional effects. Padh et al. (2021) proposed a model-agnostic, differentiable framework, yet combined fairness penalties additively. Martinez et al. (2020) introduced static Pareto optimization without dynamic trade-off navigation, and Nagpal et al. (2024) jointly optimized group and individual fairness using ROC, but without compositional modeling. Crucially, no existing work captures the non-additive structure of intersectional bias or offers dynamic control over the accuracy–fairness Pareto front.

While these efforts mark meaningful progress, they leave a critical research gap—they stop short of capturing the compositional complexity of intersectional bias. The inability to model fairness over the full joint distribution of sensitive features leaves existing methods blind to the most vulnerable subgroups. *To our knowledge, this work is the first to explicitly model and optimize fairness over the Cartesian product of protected attributes—directly addressing intersectional bias as a compositional, multi-objective problem.*

Our work bridges this gap through four key advances. First, we reformulate intersectional fairness as a *joint optimization* over the Cartesian product of sensitive attributes, capturing compounded biases via subgroup-wise constraints (e.g., ensuring parity for Black women directly, not just marginally for race or gender). Second, we introduce *APFEx (Adaptive Pareto Front Explorer)*, an adaptive multi-objective optimizer that dynamically switches between: (1) *Pareto cone projection* for aligned objectives, (2) *gradient weighting* for imbalanced improvements, and (3) *calibrated exploration* to escape local optima—theoretically guaranteed to converge to Pareto-stationary solutions (Theorem 3.12). Third, we develop *differentiable intersectional metrics* that enable gradient-based optimization of otherwise non-smooth subgroup fairness (e.g., hyperbolic tangent relaxations for demographic parity). Fourth, we also introduced a generalized fairness loss that handles multi-class classification with intersectional fairness (subsubsection 3.3.4).

2 Related Work

Fairness in machine learning has traditionally been studied under three broad methodological families: pre-processing, in-processing, and post-processing interventions. While these approaches have yielded important insights, they remain limited in their ability to address *intersectional fairness*, i.e., fairness across subgroups defined by joint combinations of multiple sensitive attributes. Below we review these categories before positioning our work, which directly tackles intersectional fairness through a multi-objective optimization framework with adaptive Pareto front exploration.

Pre-Processing and Post-Processing Approaches. Pre-processing methods aim to mitigate bias by modifying the input distribution prior to training. Early work by Kamiran & Calders (2012) adjusted class labels or sample weights to suppress discriminatory patterns. Later contributions introduced probabilistic and generative data transformations for fairness-preserving representations (Calmon et al., 2017; Sattigeri et al., 2019). While these methods are model-agnostic, they often misalign with a learner's inductive bias, limiting their generalizability.

Post-processing methods, in contrast, operate on model outputs. Hardt et al. (2016) proposed threshold calibration for equalized opportunity, while Chzhen et al. (2019) refined this idea through semi-supervised calibration. These approaches are lightweight and modular, but their corrective power is bounded by the representational capacity of the underlying model, making them ineffective when bias is deeply encoded in learned features. Importantly, both pre- and post-processing strategies typically focus on a single sensitive attribute, offering no principled mechanism for handling intersectional subgroups.

In-Processing and Relaxation-Based Methods. In-processing approaches directly integrate fairness constraints into model training, often by formulating the task as a constrained optimization problem. A

common strategy replaces non-differentiable fairness indicators with differentiable relaxations. Zafar et al. (2017) proposed covariance-based relaxations for demographic parity and equalized opportunity, yielding convex or convex–concave formulations. Donini et al. (2018) adopted an empirical risk minimization approach under fairness constraints, and Celis et al. (2019) introduced a meta-algorithm that can optimize classification subject to multiple fairness metrics using estimated conditional probabilities.

However, relaxations can sometimes be too permissive: Lohaus et al. (2020) demonstrated that models may satisfy relaxed constraints while violating fairness in the true sense. Their search-based regularization approach offers stronger guarantees but incurs substantial computational cost. To improve scalability, Padh et al. (2021) framed fairness as a multi-objective problem and introduced a differentiable relaxation using hyperbolic tangent functions. This enabled simultaneous optimization of multiple fairness definitions across sensitive attributes. Yet, their method still treated attributes separately and failed to capture their joint interactions. Similarly, Nagpal et al. (2024) proposed a reject-option classification framework optimized with NSGA-II and particle swarm algorithms, achieving improved accuracy–fairness trade-offs but without explicit modeling of fairness over the joint distribution of protected attributes.

Overall, in-processing methods have provided more principled mechanisms for balancing fairness and accuracy, but most remain tied to single-attribute settings or require modifications for each fairness metric.
Multi-Objective and Intersectional Fairness. A growing line of research has considered fairness explicitly as a multi-objective problem. Valdivia et al. (2021) applied evolutionary algorithms to optimize both fairness metrics and predictive accuracy, while Milojkovic et al. (2019) and Cotter et al. (2019) developed multi-gradient descent and game-theoretic formulations for non-convex fairness–accuracy optimization. Despite these advances, most existing methods either treat sensitive attributes separately or rely on scalarized losses that obscure the performance of intersectional subgroups.

Several works specifically address intersectional fairness. For instance, Foulds et al. (2019) propose a Bayesian probabilistic modeling approach that enables data-efficient estimation of fairness across multiple protected attributes by handling sparsity and statistical uncertainty. However, their focus lies on measurement and estimation, rather than on optimizing classifier decisions under intersectional fairness constraints. Another relevant contribution is by Mickel (2023),who develop both auditing and post-processing techniques for intersectional fairness in classification. Their methods extend single-attribute fairness metrics to the intersectional setting and offer robust, model-agnostic bias mitigation. Yet, these are post-hoc corrections, lacking an integrated optimization framework.

Our Contribution. Based on the above category that we have discussed our method lies in the category of in-processing and multi-objective optimization. Within our in-processing framework, we integrate fairness constraints associated with sensitive attributes into the model training phase, typically by framing the task as a constrained optimization problem. To the best of our knowledge, no prior work explicitly formulates fairness over the *joint sensitive attribute space* within a unified, differentiable, and multi-objective optimization framework. We bridge this gap by constructing composite subgroup labels from sensitive attribute tuples and jointly optimizing *performance loss* and *fairness loss* using an adaptive Pareto front explorer. Unlike scalarization-based or post-hoc methods, our approach preserves the trade-off structure explicitly, enabling more nuanced control over the fairness–accuracy frontier. Moreover, it is model-agnostic and readily integratable into standard supervised learning pipelines. This makes it the first method to systematically address intersectional fairness via principled multi-objective optimization, producing Pareto-optimal classifiers that respect fairness across all defined subgroups simultaneously.

3 Methodology

This section develops a principled optimization framework for jointly minimising prediction error and intersectional fairness violations. We begin by formally stating the problem and explaining why the particular mathematical objects we introduce are needed. We then recall key preliminaries from multi-objective optimization (MOO) that underpin our method, present the *Adaptive Pareto Front Explorer (APFEx)* framework and its core components, and finish with the concrete differentiable losses we use to encode fairness constraints.

3.1 Problem Definition

We consider a multi-objective optimization problem where we seek to balance predictive accuracy with fairness constraints across intersectional subgroups. Let $\mathcal{D} = \{(x_i, a_i, y_i)\}_{i=1}^n$ denote our dataset, where $x_i \in \mathbb{R}^d$ represents the feature vector, $a_i = (a_i^{(1)}, \ldots, a_i^{(K-1)}) \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_{K-1}$ represents the vector of sensitive attributes having nominal values with feature space as $\mathcal{A}_i \in \{-1, 1\}$, and $y_i \in \{-1, 1\}$ is the binary label. Consider a probability distribution of the training samples $\mathbb{S} = (\mathcal{X}, \mathcal{A}, \mathcal{Y})$ be $\mathbb{P}^{\mathbb{S}}$, where each $\{x_i, a_{i1}, \cdots, a_{i(k-1)}, y_i\}$ is sampled *i.i.d* from $\mathbb{P}^{\mathbb{S}}$.

The intersectional fairness problem requires us to consider fairness across the Cartesian product of sensitive attributes. For $K-1$ sensitive attributes, we define the set of intersectional groups as:

$$\mathcal{G} = \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_{K-1}$$

Each intersectional group $g \in \mathcal{G}$ represents a unique combination of sensitive attribute values, capturing the compounded bias that individuals at intersections may face. The learning problem we address is inherently multi-objective: we must trade off a predictive objective and fairness objectives defined across the intersectional groups. In the following definition 3.1, we have formally defined this.

Definition 3.1 (Multi-Objective Intersectional Fairness Problem) *Given a parametric model $f_\theta : \mathbb{R}^d \to \mathbb{R}$ with parameters $\theta \in \Theta \subseteq \mathbb{R}^p$, we seek to solve:*

$$\min_{\theta \in \Theta} \mathbf{L}(\theta) = \begin{bmatrix} L_{task}(\theta) \\ L_{fairness}(\theta) \end{bmatrix} \tag{1}$$

where $L_{task}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f_\theta(x), y)]$ measures predictive performance, and $L_{fairness}(\theta)$ captures fairness violations across intersectional groups.

Note 3.1 *We have K objective functions, one for the task prediction and $(K-1)$ for $(K-1)$ sensitive attributes.*

3.2 Preliminaries

Before introducing APFEx, we briefly recall MOO concepts that motivate our method for fairness-aware training.

3.2.1 Multi-objective optimization (MOO)

Multi-objective optimization (MOO) addresses problems where multiple, potentially conflicting objectives must be simultaneously optimized. Unlike single-objective optimization, the solution to a multi-objective problem is typically not a single point but rather a set of points representing various trade-offs among the objectives. We define the Multi-objective optimization as:

$$\min_{x_i \in \mathcal{X}} \mathbb{F}(x_i) = [f_1(x_i), f_2(x_i), \cdots, f_K(x_i)]^T \tag{2}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ is the decision vector, $\mathcal{X} \subseteq \mathbb{R}^d$ is the feasible region, and $\mathbb{F} : \mathcal{X} \to \mathbb{R}^K$ is the objective function vector with K individual objective functions $f : \mathcal{X} \to \mathbb{R}$, $f_i(x_j)$ is the loss function for the ith objective.

3.2.2 Pareto dominance and Pareto front

At the core of multi-objective optimization is Pareto optimality, which formalizes the notion of trade-offs between competing objectives. To begin formalizing this framework, we first define a concept that allows us to compare solutions based on their performance across multiple objectives.

Definition 3.2 (Pareto Dominance) *A solution $f(x^1)$ is said to dominate another solution $\mathbf{x}^2$ (denoted as $\mathbf{x}^1 \prec \mathbf{x}^2$) iff:*

1. $\forall i \in \{1, 2, \ldots, K\} : f_i(x^1) \leq f_i(x^2)$, *and*
2. $\exists j \in \{1, 2, \ldots, K\} : f_j(x^1) < f_j(x^2)$

This notion of dominance allows us to identify solutions that are better than others in at least one objective and no worse in all others. Building on this, we can define the concept of Pareto optimality, which characterizes solutions that are not dominated by any other.

Definition 3.3 (Pareto Optimality) *: A solution x^* is Pareto optimal if there exists no other solution $x \in \mathcal{X}$ such that $f(x) \prec x^*$.*

The collection of all Pareto optimal solutions forms a special set in the objective space, which we define next.

Definition 3.4 (Pareto Front) *: The set of all Pareto optimal solutions in the objective space, denoted as $PF = \{F(x) \mid x \in PS\}$, where PS is the set of all Pareto optimal solutions in the decision space.*

These concepts are central to fairness-aware learning because they let us characterise the set of classifiers that achieve different accuracy–fairness trade-offs rather than forcing a single scalarised compromise. To solve a multi-objective optimization problem, we aim to identify or approximate the Pareto front. Consider K objective functions $f_1(\mathbf{w}), f_2(\mathbf{w}), \ldots, f_K(\mathbf{w})$ with a parameter vector $\mathbf{w} \in \mathbb{R}^d$. A fundamental challenge is to identify a direction in parameter space that reduces all objective functions simultaneously. This leads us to the concept of Pareto stationarity, which provides a necessary condition for a solution to be Pareto optimal in continuous spaces.

Definition 3.5 (Pareto Stationary Fliege & Svaiter (2000)) *A point $\mathbf{x} \in \Omega$ is Pareto stationary if there does not exist a feasible direction $\mathbf{d} \in \mathbb{R}^n$ such that:*

$$\nabla f_i(x)^T d < 0, \forall i \in \{1, 2, \cdots, K\}. \tag{3}$$

The following Theorem 3.1 gives a convenient geometric characterisation that we exploit algorithmically: Pareto stationarity holds iff the zero vector lies in the convex hull of the gradients.

Theorem 3.1 *A point $\mathbf{x} \in \Omega$ is Pareto stationaryFliege & Svaiter (2000) iff $\mathbf{0} \in conv(\nabla f_1(\mathbf{x}), \nabla f_2(\mathbf{x}), \ldots, \nabla f_K(\mathbf{x}))$, where conv denotes the convex hull.*

Proof 3.2 *The proof is provided in Appendix A.2.*

Building on Pareto optimality and stationarity, we propose the *Adaptive Pareto Front Explorer (APFEx)*—a multi-objective framework that dynamically explores the Pareto front and manages trade-offs between competing objectives. APFEx extends gradient-based methods by combining multiple descent strategies with an adaptive selection mechanism, and explicitly formulates intersectional fairness as an optimization objective to jointly minimize prediction error and subgroup-level fairness violations.

3.3 Proposed Methodology

Building on the above preliminaries, we propose the *Adaptive Pareto Front Explorer (APFEx)*: a dynamic multi-objective optimisation framework tailored to balance predictive performance and intersectional fairness. APFEx maintains gradient-scale invariance, adaptively chooses among complementary descent strategies, and injects controlled exploration when optimisation stalls. Below we present the components and their motivations, followed by convergence results. Following our discussions on Gradient Normalization 3.3.1 and the APFEx framework 3.3.2, we next reviewed the Existing Loss Function using Differential Metric Approximation in Subsection 3.3.3. Furthermore, in Section 3.3.4, we presented the Fairness-Constrained Loss Function, which is formulated through a Differentiable Fairness Metric Approximation.

3.3.1 Gradient Normalization and Scale Invariance

Different objectives (e.g., cross-entropy and group-level disparity) can vary greatly in magnitude, so directly aggregating their gradients may allow large-scale losses to dominate. To ensure a fair comparison, we estimate an empirical scale for each objective and normalize the gradients accordingly.

Definition 3.6 (Empirical Loss Scale Estimation) *For each objective $k \in \{1, \ldots, K\}$, we estimate the empirical maximum:*

$$\hat{L}_k^{\max} = \max_{(x,y)\in\mathcal{D}_{train}} \ell_k(f_\theta(x), y) + \epsilon$$

where $\epsilon > 0$ is a small regularization constant to prevent division by zero.

The normalized gradient for objective k is then computed as:

$$\tilde{\nabla}_\theta L_k(\theta) = \frac{\nabla_\theta L_k(\theta)}{\hat{L}_k^{\max}}$$

Normalisation preserves the geometry of each objective while preventing numerical dominance by any single loss. This step is used before forming aggregated descent vectors in APFEx.

3.3.2 APFEx Framework

To optimize the trade-off between predictive performance and intersectional fairness, we propose APFEx (Figure 1): a dynamic, multi-objective framework. We first present its core descent strategies (PCP, AW, PSS), followed by the adaptive mechanism for strategy selection, constraint handling, and normalization. We then unify these components into a differentiable loss with theoretical convergence guarantees, demonstrating APFEx's ability to achieve Pareto-optimal solutions while preserving intersectional fairness. In the subsequent Sections 3.3.2.1 and 3.3.2.2, we described the different Strategies of the APFEx framework and the convergence analysis of our APFEx framework, respectively.

3.3.2.1 Descent Vector Computation Strategies The core innovation of APFEx lies in its adaptive selection among three complementary descent strategies, each optimized for different optimization landscape characteristics. In Subsections (i), (ii), and (iii), we have discussed the PCP, AW, and PSS strategy, respectively, and we have presented the detailed algorithm in Algorithm refalg:apfex.

(i) Pareto Cone Projection (PCP) This strategy projects gradients onto the Pareto cone (Fliege & Svaiter, 2000) to ensure descent for all objectives simultaneously, thereby ensuring Pareto efficiency. We have defined the Pareto cone projection strategy in definition 3.7.

Definition 3.7 (Pareto Cone (Fliege & Svaiter, 2000)) *Given gradients $\{\nabla L_k(\theta)\}_{k=1}^K$, the Pareto cone $\mathcal{C}(\theta)$ is defined as:*

$$\mathcal{C}(\theta) = \{d \in \mathbb{R}^p : \langle \nabla L_k(\theta), d\rangle \leq 0,\ \forall k \in \{1, \ldots, K\}\}$$

The PCP strategy computes a convex combination of gradients by solving the following quadratic programming problem (Désidéri, 2012):

$$\min_{\alpha\in\Delta^{K-1}} \quad \frac{1}{2}\alpha^T G\alpha \tag{4}$$

$$\text{subject to} \quad \sum_{k=1}^{K} \alpha_k = 1, \quad \alpha_k \geq 0,\ k = 1, \ldots, K \tag{5}$$

where $G \in \mathbb{R}^{K\times K}$ is the Gram matrix with entries $G_{ij} = \langle \nabla L_i(\theta), \nabla L_j(\theta)\rangle$, and Δ^{K-1} denotes the probability simplex.

The quadratic program above selects an optimal convex weight vector $\alpha^* \in \Delta^{K-1}$ that minimises the squared norm of the aggregated gradient

$$g^*(\theta) \ = \ \sum_{k=1}^{K} \alpha_k^* \nabla L_k(\theta).$$

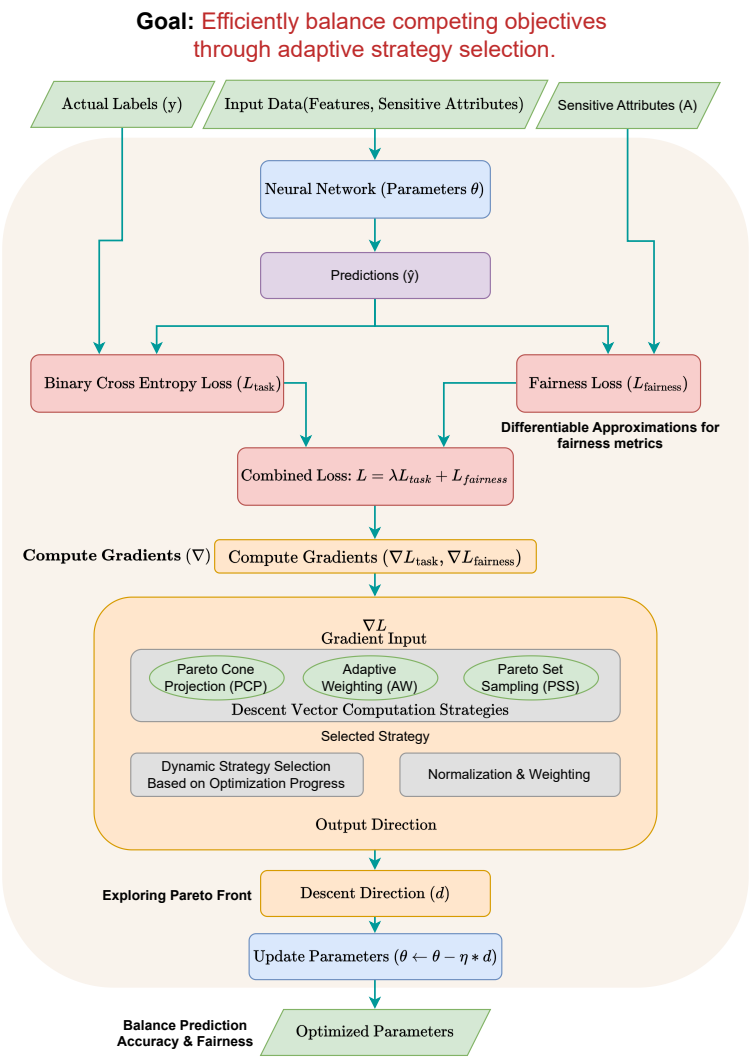

Figure 1: Figure illustrates the workflow of Adaptive Pareto Front Explorer (APFEx).

Intuitively, $g^*(\theta)$ is the "best compromise" gradient obtained by convexly combining per-objective gradients. Two important remarks follow. First, if the convex hull of $\{\nabla L_k(\theta)\}_{k=1}^K$ contains the origin, then the QP minimum is achieved at $g^*(\theta) = \mathbf{0}$; this is precisely the Pareto-stationary case (no common descent direction exists). Second, when $g^*(\theta) \neq \mathbf{0}$ we use the unit vector aligned with $-g^*(\theta)$ as the PCP descent direction:

$$d_{\text{PCP}} \;=\; -\frac{g^*(\theta)}{\|g^*(\theta)\|} \;=\; -\frac{\sum_{k=1}^K \alpha_k^* \nabla L_k(\theta)}{\left\| \sum_{k=1}^K \alpha_k^* \nabla L_k(\theta) \right\|}.$$

This normalisation yields a direction of descent (when one exists) while decoupling step-size choice from the aggregated gradient magnitude. Theorem 3.3 below characterises the optimiser α^* algebraically via KKT conditions, and Lemma 3.5 then connects that algebraic description to the geometric Pareto cone property.

The following theorem characterises the optimal α^* via KKT conditions and clarifies the structure of active components in the convex combination.

Theorem 3.3 (PCP Optimality Conditions) *The solution α^* to the PCP optimization problem satisfies the Karush-Kuhn-Tucker (KKT) conditions:*

$$G\alpha^* + \mu^*\mathbf{1} - \nu^* = 0 \tag{6}$$

$$\sum_{k=1}^{K} \alpha_k^* = 1 \tag{7}$$

$$\alpha_k^* \geq 0, \quad \nu_k^* \geq 0, \quad \alpha_k^*\nu_k^* = 0, \ \forall k \tag{8}$$

where μ^* and $\nu^* = (\nu_1^*, \nu_2^*, \ldots, \nu_K^*)$ are the Lagrange multipliers.

Proof 3.4 *Proof of this theorem is provided in Appendix section A.4*

Remark 1 *Theorem 3.3 establishes the precise optimality conditions for the PCP formulation by deriving its Karush–Kuhn–Tucker (KKT) system. This proof serves two key purposes. First, it confirms that the PCP optimization problem is a well-posed convex problem, where the global solution is fully characterized by the derived KKT conditions. Second, by identifying the structural interplay between the Gram matrix G, the simplex constraint, and the non-negativity requirements, the proof provides the theoretical foundation upon which our proposed method builds. In particular, the complementary slackness condition reveals how the selection of active strategies emerges naturally from the optimization, thereby justifying the principled design of our approach.*

Since G is a Gram matrix it is positive semi-definite and the QP is convex; therefore the KKT conditions are necessary and sufficient for optimality. The stationarity equation $G\alpha^* + \mu^*\mathbf{1} - \nu^* = 0$ shows how α^* balances the pairwise geometry of the gradients (via G) against the simplex constraints. Complementary slackness $\alpha_k^*\nu_k^* = 0$ identifies the active set: indices with $\alpha_k^* > 0$ contribute to the aggregated gradient, while indices with $\alpha_k^* = 0$ are excluded.

Having characterised α^*, we next connect the QP solution to descent in the Pareto cone. The lemma below states the crucial property that if the Pareto cone has interior directions then PCP indeed produces a vector inside that cone — i.e., a common descent direction that improves all objectives simultaneously.

Lemma 3.5 (PCP Descent Property) *If the Pareto cone $\mathcal{C}(\theta)$ has non-empty interior, then $d_{PCP} \in \mathcal{C}(\theta)$, guaranteeing simultaneous improvement in all objectives.*

Proof 3.6 *The proof is provided in Appendix section A.6.*

Theorem 3.3 provides the algebraic description of the optimiser α^* and explains which objectives actively shape the aggregated gradient. Lemma 3.5 then leverages that optimiser to show the geometric consequence: when a common descent direction exists (i.e., Pareto cone interior is non-empty), the QP solution yields a descent vector inside the cone and thus produces simultaneous improvement. Together, they justify PCP as a principled method to exploit shared descent directions when they are present.

(ii) Adaptive Weighting (AW) Sener & Koltun (2018) This strategy dynamically adjusts weights based on relative improvement rates across objectives. When objectives show imbalanced improvement rates, AW dynamically redistributes attention to underperforming objectives. AW implements this idea by tracking recent relative improvements and converting them into weights. We have calculated the relative improvement rate in the following Definition 3.8

Definition 3.8 (Relative Improvement Rate) *For objective k at iteration t, the relative improvement rate is:*

$$\rho_k^{(t)} = \frac{L_k^{(t-1)} - L_k^{(t)}}{\max(L_k^{(t-1)}, \epsilon)}$$

where $\epsilon > 0$ prevents division by zero.

Using the relative improvement rate, AW strategy updates weights according to an exponential reweighting scheme:

$$\alpha_k^{(t+1)} = \frac{\exp(-\tau \rho_k^{(t)})}{\sum_{j=1}^{K} \exp(-\tau \rho_j^{(t)})} \tag{9}$$

$$\text{where } \tau > 0 \text{ is the adaptation rate parameter} \tag{10}$$

We now turn to the convergence analysis of the AW strategy, as formalized in Theorem 3.7.

Theorem 3.7 (AW Convergence Properties) *Under the assumption that each objective satisfies the Polyak-Łojasiewicz condition with parameter $\mu > 0$, the AW strategy ensures that:*

1. *The weight sequence $\{\alpha_k^{(t)}\}_{t=1}^{\infty}$ remains bounded for all k*
2. *If objective k consistently underperforms (i.e., $\rho_k^{(t)} < \gamma$ for some threshold $\gamma > 0$), then $\lim_{t\to\infty} \alpha_k^{(t)} \geq \delta > 0$ for some δ depending on τ and γ*

Proof 3.8 *The proof is provided in Appendix A.8*

Remark 2 *Theorem 3.7 establishes two crucial stability properties of the Adaptive Weighting (AW) strategy. First, it guarantees the boundedness of the weight sequence, thereby ensuring numerical robustness throughout the training process. Second, it shows that objectives that consistently underperform are not discarded but are instead assigned a strictly positive minimum weight in the limit. This property is central to our proposed method, as it prevents the optimization from prematurely neglecting difficult or minority objectives, thereby promoting balanced progress across tasks. In essence, the proof formalizes the intuition that AW adaptively preserves diversity in optimization focus while maintaining convergence guarantees, aligning directly with our APFEx framework.*

AW helps ensure that challenging or underrepresented objectives are not overlooked over time. By adaptively focusing attention where it's most needed, it promotes fairer and more balanced outcomes throughout training.

(iii) Pareto Set Sampling (PSS) This exploration-focused Lin et al. (2019) strategy samples random descent directions to efficiently explore the Pareto front.

When progress stalls globally, deterministic strategies (PCP, AW) may repeatedly propose similar descent directions, leading to premature convergence. PSS introduces stochasticity to deliberately explore underrepresented trade-offs on the Pareto front. PSS first detects whether the progress stalls and the optimization plateaus. It then performs Exploratory Sampling. When optimization stagnates, PSS introduces controlled exploration to escape local attractors and discover new regions of the Pareto front. In Definition 3.9 below, we have defined the condition for stagnation detection.

Definition 3.9 (Stagnation Detection) *The optimization is considered stagnant at iteration t if:*

$$\Delta^{(t)} = \|\mathbf{L}^{(t-1)} - \mathbf{L}^{(t)}\|_2 < \delta$$

We declare stagnation when $\Delta^{(t)} < \delta$ for M consecutive iterations, with $\delta > 0$ a tolerance.

During stagnation, PSS samples weights from a Dirichlet distribution:

$$\alpha^{(t)} \sim \mathrm{Dir}(\mathbf{1}_K) = \mathrm{Dir}(1, 1, \ldots, 1)$$

forming an exploratory direction $\tilde{d}^{(t)} = -\sum_k \alpha_k^{(t)} \nabla L_k(\theta)$. To avoid abrupt steps we smooth the update:

$$d_{\mathrm{PSS}}^{(t)} = \lambda \tilde{d}^{(t)} + (1 - \lambda) d_{\mathrm{last}}$$

where $\tilde{d}^{(t)} = -\sum_{k=1}^{K} \alpha_k^{(t)} \nabla L_k(\theta)$ and $\lambda \in (0, 1]$ controls exploration intensity.

The following Lemma 3.9 shows that PSS exploration is not arbitrary: on average, each objective receives equal weight, variance is bounded so updates remain stable, and the support condition guarantees that no region of the simplex is ignored. In practice, these properties ensure that stochastic exploration systematically covers the space of possible trade-offs instead of collapsing to a narrow subset.

Lemma 3.9 (PSS Exploration Properties) *The Dirichlet sampling in PSS ensures:*

1. $\mathbb{E}[\alpha_k^{(t)}] = \frac{1}{K}$ *(unbiased exploration)*
2. $Var[\alpha_k^{(t)}] = \frac{K-1}{K^2(K+1)}$ *(controlled variance)*
3. *The support covers the entire probability simplex* Δ^K

As established in Lemma 3.9, PSS proves especially effective when the search is near optimization plateaus or local attractors. By introducing unbiased stochastic perturbations, it can uncover fresh trade-off solutions that deterministic methods often miss, preventing them from becoming trapped in repetitive cycles.

Having analysed one strategy in isolation, it is natural to ask: *under what landscape conditions should each of PCP, AW, or PSS be preferred?* The following theorem 3.10 addresses this broader design question.

Theorem 3.10 (Strategy-Landscape Correspondence) *Each strategy in APFEx achieves optimality under specific landscape conditions:*

1. *PCP is optimal when gradients are well-aligned (i.e.,* $\cos(\theta_{ij}) > \epsilon$ *for most pairs* (i, j)*)*
2. *AW is optimal when improvement rates are imbalanced (i.e.,* $\max_k \rho_k^{(t)} / \min_k \rho_k^{(t)} > \beta$ *for threshold* $\beta > 1$*)*
3. *PSS is optimal near local attractors (i.e., when* $\|\nabla L_k(\theta)\| < \epsilon$ *for all* k *but global optimality is not achieved)*

Proof 3.11 *The proof is provided in Appendix Section A.10.*

Remark 3 *In Theorem 3.10, we provide a rigorous analysis of the underlying strategy landscape. Our goal is to characterize how the proposed strategy selection mechanisms navigate this landscape and under what conditions they converge to the optimal solution. Specifically, the theorem establishes that the mechanisms we introduce are not only theoretically sound but also capable of attaining the global optimum within the defined framework. This analysis sheds light on the interplay between different strategies and demonstrates the robustness of our approach in consistently identifying the optimal solution across varying configurations of the problem space.*

This theorem advances the discussion from *local* statistical behavior to the broader perspective of *global* optimization dynamics. While Lemma 3.9 ensures that PSS systematically explores the trade-off space, Theorem 3.10 places PSS in direct comparison with PCP and AW, clearly delineating the circumstances under which each strategy delivers the greatest progress.

Taken together, these results provide a compelling justification for the adaptive switching mechanism in APFEx. The framework can capitalize on PCP when the objectives are well aligned, shift to AW to restore balance when improvements become uneven, and revert to PSS whenever broader exploration is essential.

3.3.2.2 Convergence Analysis of APFEx A robust optimisation framework must do more than propose intuitive strategies—it must also demonstrate that these strategies exhibit well-defined convergence behaviour. In the single-objective setting, standard gradient-based methods are supported by strong theoretical guarantees. However, when optimisation involves multiple, often competing, objectives—such as

balancing accuracy and fairness—the challenge becomes significantly greater. Conflicting criteria can pull the optimisation process in opposing directions, and without rigorous theoretical support, it remains unclear whether an adaptive method like APFEx will ultimately stabilise or merely oscillate.

In light of this, we pose the following central question: *Does APFEx provably make progress toward Pareto-stationary solutions, and if so, at what rate?*

The Theorem 3.12 presented below provides an affirmative answer. Under standard smoothness conditions and the Polyak–Łojasiewicz (PL) inequality, we show that APFEx achieves a convergence rate comparable to the best-known guarantees in stochastic optimisation.

Theorem 3.12 (APFEx Convergence Rate) *Assume each objective L_k is L-smooth and satisfies the Polyak-Łojasiewicz condition with parameter $\mu > 0$. Then APFEx converges to a Pareto-stationary point at rate $O(1/\sqrt{T})$, where T is the number of iterations.*

Specifically, there exists a constant $C > 0$ such that:

$$\min_{t \in \{0,1,\ldots,T-1\}} \omega(\theta^{(t)}) \leq \frac{C}{\sqrt{T}}$$

where $\omega(\theta) = \min_{\alpha \in \Delta^{K-1}} \left\| \sum_{k=1}^{K} \alpha_k \nabla L_k(\theta) \right\|$ is the Pareto-stationarity measure.

Proof 3.13 *The proof is outlined in Appendix SectionA.12.*

Remark 4 *Theorem 3.12 establishes the theoretical foundation of our proposed APFEx framework by proving its convergence rate towards a Pareto-stationary point under standard smoothness and Polyak-Łojasiewicz conditions. The proof carefully exploits the adaptive strategy selection mechanism of APFEx, which guarantees sufficient alignment between the aggregated descent direction and the true multi-objective gradient. This alignment ensures consistent descent progress, enabling the derivation of a non-trivial upper bound on the stationarity measure. The obtained rate of $O(1/\sqrt{T})$ is particularly significant: it not only certifies the efficiency of APFEx in handling multi-objective optimization but also places it on par with the best-known rates in single-objective stochastic optimization. Thus, this result formally validates the effectiveness of our method and provides theoretical justification for the empirical success observed in our experiments.*

Algorithm 1 Adaptive Pareto Front Explorer (APFEx)

Require: Initial parameters $\theta^{(0)}$, learning rate η, loss functions $\{L_1, L_2, \ldots, L_K\}$
Ensure: Optimized parameters $\theta^{(T)}$

```
Initialize strategies S = {"pareto_cone", "adaptive", "exploration"}
Set active strategy s^(0) = "adaptive"
Initialize tracker for Pareto front approximation
for t = 0, 1, ..., T − 1 do
    Compute losses L_i(θ^(t)) for i = 1, 2, ..., K
    Compute gradients ∇L_i(θ^(t)) for i = 1, 2, ..., K
    Compute descent direction d^(t) using strategy s^(t)
    if constraints present then
        Modify d^(t) to respect constraints
    end if
    Update parameters: θ^(t+1) = θ^(t) − η · d^(t)
    Update tracker with current losses and weights
    Update active strategy s^(t+1) based on progress
end for
return θ^(T)
```

3.3.3 Existing Loss Function via Differential Metric Approximation

Demographic parity, also known as statistical parity, represents one of the most intuitive fairness criteria. The core idea is straightforward: a fair classifier should make positive predictions at equal rates across different demographic groups. However, translating this intuitive concept into a differentiable objective function presents significant technical challenges. Consider a binary classifier making predictions $\hat{y}$ based on features x and sensitive attribute $a \in 0, 1$. Demographic parity requires the following condition:

$$P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1) \tag{11}$$

The DPLoss Padh et al. (2021) formulation combines fairness enforcement with predictive accuracy through a composite loss function:

$$\mathcal{L}_{DP} = |DP_0 - DP_1| + \lambda\mathcal{L}_{task}(y, \hat{y}) + \beta|\hat{y}|_2^2 \tag{12}$$

where DP_i represents the demographic parity metric for group i, λ controls the fairness-accuracy trade-off, and β provides additional regularization.

The concept of equalized opportunity, formalized through equal true positive rates, provides a fairness criterion that focuses on equal benefit distribution among qualified individuals. The mathematical constraint requires:

$$P(\hat{Y} = 1|Y = 1, A = 0) = P(\hat{Y} = 1|Y = 1, A = 1) \tag{13}$$

The TPRLoss Padh et al. (2021) combines fairness enforcement with predictive objectives:

$$\mathcal{L}_{TPR} = |TPR_0 - TPR_1| + \lambda\mathcal{L}_{task}(y, \hat{y}) \tag{14}$$

TPRLoss employs the following transformation:

$$\text{round}_{\text{soft}}(x) = \frac{\tanh(5(x - \tau))}{2} + 0.5 \tag{15}$$

The differentiable formulation enables joint optimization of fairness and accuracy through standard back-propagation. However, the conditional nature of TPR computation can lead to gradient instability when protected subgroups contain few positive examples. The epsilon smoothing in denominators prevents gradient explosion while maintaining optimization effectiveness.

While specialized loss functions serve specific fairness criteria well, practitioners often need flexibility to experiment with different notions of fairness or handle complex demographic structures. The generalized Fairness Loss in the following section provides a unified framework supporting multiple fairness criteria and multi-class sensitive attributes.

3.3.4 Fairness-Constrained Loss Function via Differentiable Fairness Metric Approximation

Our proposed methodology incorporates fairness constraints directly into the model optimization process through a specialized loss function that balances predictive performance with fairness considerations across sensitive demographic attributes. The approach enables end-to-end training of fair classifiers without requiring post-processing or constraint relaxation techniques.

We introduce a novel fairness-constrained loss function that combines traditional task-specific objectives with a differentiable fairness regularization term:

$$\mathcal{L} = \lambda\mathcal{L}_{\text{task}} + \mathcal{L}_{\text{fairness}} \tag{16}$$

where λ is a hyperparameter controlling the trade-off between predictive performance and fairness objectives.

The fairness violation term $\mathcal{L}_{\text{fairness}}$ quantifies disparities between protected groups by computing the mean absolute difference across all group pairs.

For K sensitive groups, the framework computes all pairwise fairness violations:

$$\mathcal{L}_{\text{fairness}} = \frac{1}{\binom{K}{2}} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} |M_i - M_j| \tag{17}$$

where M_i represents the chosen fairness metric (demographic parity or true positive rate) for group i which results in the loss functions $\mathcal{L}_{\text{fairnessDP}}$ and $\mathcal{L}_{\text{fairnessTPR}}$. This formulation scales naturally to multiple protected groups and provides comprehensive fairness enforcement.

3.3.4.1 Differentiable Approximation of Fairness Metrics A key innovation in our approach is the differentiable approximation of otherwise non-smooth fairness metrics. We introduce a continuous relaxation of the step function required for computing fairness metrics in multi-class classification settings:

(i) **Hyperbolic tangent relaxation**: We employ a scaled hyperbolic tangent function (Padh et al., 2021) to approximate the step function:

$$\hat{H}(x) = \frac{\tanh(5(x - \tau))}{2} + 0.5 \tag{18}$$

where τ represents the classification threshold and the scaling factor which modulates the steepness of the transition.
So we discussed the hyperbolic tangent relaxation as a smooth surrogate for hard indicator functions, offering both infinite differentiability and a natural probabilistic interpretation of thresholding. Yet, despite these appealing properties, it reveals two key shortcomings in our setting: (i) the gradient of tanh is sharply concentrated near the threshold and rapidly diminishes elsewhere, which can cause vanishing gradients when subgroup distributions are imbalanced or when predictions lie far from the threshold; and (ii) its steepness near the boundary can lead to unstable updates if many points cluster close to the decision surface, a situation common in high-dimensional categorical tasks. These challenges become even more pronounced when both the outcomes and protected attributes are categorical and possibly multi-class, where fairness metrics demand repeated conditional comparisons and unstable gradients can disproportionately impact minority groups. To address these limitations, we introduce the *Continuous Conditional Response (CCR)* approximation. CCR preserves the interpretability of a soft indicator while replacing tanh's sigmoidal gradient with a bounded, constant gradient within a flexible transition band. This design produces a relaxation that remains numerically stable, theoretically transparent, and converges to the hard indicator as the smoothing width narrows. As demonstrated below, CCR integrates naturally into multi-class fairness metrics and delivers more dependable gradient signals for APFEx when optimizing across intersectional groups.

(ii) **Continuous conditional response (CCR)**: We introduce a smooth relaxation of the indicator function using the continuous clipping approximation. Let $\tau \in \mathbb{R}$ denote a threshold (for scalar-valued scores) and $\gamma > 0$ represent a smoothing width. The CCR approximation of the indicator $\mathbb{1}\{x > \tau\}$ is given by the piecewise-linear function

$$\hat{C}(x; \tau, \gamma) = \begin{cases} 0, & x \leq \tau - \gamma, \\ \dfrac{x - (\tau - \gamma)}{2\gamma}, & \tau - \gamma < x < \tau + \gamma, \\ 1, & x \geq \tau + \gamma, \end{cases} \tag{19}$$

which provides a differentiable transition from 0 to 1 around the threshold. In the multiclass setting, where model predictions are represented by logits $z_i \in \mathbb{R}^C$, we take the softmax probability $p_{i,c} = \text{softmax}(z_i)_c$ as a smooth surrogate for the discrete indicator $\mathbb{1}\{\hat{y}_i = c\}$. Similarly, for (potentially uncertain) group memberships, we assign a CCR-based group weight

$$h_{i,g} = \hat{C}(u_{i,g}; \tau_g, \gamma_g),$$

where $u_{i,g}$ is a scalar membership score for sample i in group g. For discrete attributes, $u_{i,g}$ reduces to a one-hot encoding, whereas for uncertain or embedded attributes, it can instead represent a learned score or a similarity measure.

Properties of CCR:

(a) **Pointwise convergence:** $\hat{C}(x;\tau,\gamma) \to \mathbb{1}\{x > \tau\}$ as $\gamma \to 0^+$.

(b) **Piecewise differentiability:** $\hat{C}(\cdot;\tau,\gamma)$ is continuous everywhere and differentiable for all $x \neq \tau \pm \gamma$; it is subdifferentiable at the junctions $x = \tau \pm \gamma$.

(c) **Bounded derivative:**

$$\frac{d}{dx}\hat{C}(x;\tau,\gamma) = \begin{cases} 0, & |x-\tau| \geq \gamma, \\ \dfrac{1}{2\gamma}, & |x-\tau| < \gamma, \end{cases}$$

hence $\left\|\frac{d}{dx}\hat{C}\right\|_\infty = \frac{1}{2\gamma}$.

(d) **Lipschitz continuity:** $\hat{C}(\cdot;\tau,\gamma)$ is L_C-Lipschitz with $L_C = \frac{1}{2\gamma}$.

Define the CCR-weighted per-class, per-group positive rate

$$DP_{c,g} \;=\; \frac{1}{N_g}\sum_{i=1}^{n} p_{i,c}\, h_{i,g}, \qquad \text{with} \quad N_g \;=\; \sum_{i=1}^{n} h_{i,g}. \tag{20}$$

This quantity generalises class-wise demographic parity to soft predictions $p_{i,c}$ and soft group-membership weights $h_{i,g}$. A multi-class fairness loss aggregating pairwise group disparities is

$$\mathcal{L}_{\text{fair}} \;=\; \frac{1}{C}\cdot\frac{2}{G(G-1)}\sum_{c=1}^{C}\sum_{1\leq g_1<g_2\leq G}\phi\big(DP_{c,g_1} - DP_{c,g_2}\big), \tag{21}$$

where ϕ is a smooth surrogate for the absolute value (e.g. $\phi(x) = \sqrt{x^2+\varepsilon}$ with $\varepsilon > 0$) to ensure differentiability.

CCR makes group-based fairness metrics differentiable and have explicit derivative bounds. CCR is compatible with multi-class outcomes (softmax) and multi-class protected attributes (soft membership weights). CCR-based fairness gradients is normalised and aggregated with task gradients (PCP/AW/PSS) while respecting the geometric stationarity condition used throughout the theoretical analysis.

3.3.5 Putting It All Together

We integrate the theoretical foundations of Pareto stationarity with differentiable fairness constraints through our APFEx framework. Our multi-objective optimization problem is formulated as:

$$\min_{\theta} \quad \mathbf{L}(\theta) = [L_{\text{task}}(\theta), L_{\text{fairness}}(\theta)]$$

where $L_{\text{task}}(\theta)$ is the prediction loss and $L_{\text{fairness}}(\theta) = \frac{1}{K}\sum_{i,j}|M_i - M_j|$ measures fairness violations across sensitive groups using differentiable approximations: $\hat{y} \approx \frac{\tanh(5\cdot(x-\text{threshold}))}{2} + 0.5$. This rule ensures exploitation via PCP/AW when possible, while invoking PSS for exploration in plateau regions.

The descent direction is determined by the active strategy $s^{(t)}$, with gradients normalized by $\tilde{\nabla}L_i(\theta) = \frac{\nabla L_i(\theta)}{L_i^{\max}}$ to ensure scale invariance. Our approach dynamically navigates the trade-off space between performance and fairness, converging to Pareto stationary solutions that represent meaningful compromises between competing objectives while satisfying the theoretical condition $\mathbf{0} \in \text{conv}(\nabla L_{\text{task}}(\theta), \nabla L_{\text{fairness}}(\theta))$.

4 Experiments & Results Discussion

4.1 Experimental Setup

4.1.1 Dataset Used.

Experiments performed on the following datasets:

(i) *Adult Income* (Dua et al., 2017): A census dataset with 48,842 samples and 14 features, where the goal is to predict if annual income exceeds \$50K ($y = 1$). Protected attributes include sex and a binarized race variable, enabling multi-dimensional fairness analysis.

(ii) *COMPAS Recidivism* (Angwin et al., 2022): A dataset of 6,167 criminal justice records with 53 features. The task is to predict recidivism. We consider race and sex as sensitive attributes to assess fairness in high-stakes, bias-prone decisions systems where historical biases are well-documented.

(iii) *German Credit* (Hofmann, 1994): Contains 1,000 credit applicant records with 20 features. The binary task is to predict creditworthiness (good/bad). Sensitive attributes include age (young/old), sex, and foreign worker status. Widely used in fairness research due to known biases in credit decisions.

(iv) *Heart Disease* (Detrano et al., 1989): Includes 303 clinical records with 14 attributes from the Cleveland Clinic Foundation. The task is to predict heart disease presence (binarized ((0: no disease, 1-4: disease severity)). Sensitive attributes include age, sex, and race/ethnicity. This dataset highlights fairness concerns in medical AI.

(v) *Celeb attributes* Liu et al. (2015) In this dataset have 202,599 facial images of celebrities. Corresponding to each face image, we have considered 40 attributes. We have considered *Smiling* as the label and *Gender* and *Hair Color* as the sensitive attributes.

4.1.2 Metrics.

Metrics used for evaluating accuracy and fairness:

(i) *Accuracy:* Proportion of total correct predictions.

$$\text{Accuracy} = \frac{\#(\text{Correct Predictions})}{\#(\text{All Predictions})} \tag{22}$$

(ii) *DPDiff_intersectional:* Max disparity (i.e. max $|DP(g_i) - DP(g_j)|$) in Demographic Parity (DP) across intersectional groups g_i, g_j.

$$\text{DP}(g_i) = \frac{\#(\text{Predicted positive in } g_i)}{\#(\text{Samples in } g_i)} \tag{23}$$

(iii) *TPRDiff_intersectional:* Max disparity in True Positive Rate (TPR) (i.e. max $|TPR(g_i) - TPR(g_j)|$) across intersectional groups.

$$\text{TPR}(g_i) = \frac{\#(\text{True positives in } g_i)}{\#(\text{Actual positives in } g_i)} \tag{24}$$

(iv) *DPDiff:* DPDiff metric is designed to measure the Demographic Parity difference when there are multiple sensitive attributes with multiple classes for each attribute. It calculates the Demographic Parity difference for each sensitive attribute separately. Suppose there are m sensitive attributes, $\{a_1, a_2, a_3, \cdots, a_m\}$ respectively. Then the DPDiff for an attribute DPDiff(a_k) is given by, DPDiff(a_k) = max |DP(g_i) - DP(g_j)| for all groups g_i, g_j in attribute a_k.

(vi) *TPRDiff:* TPRDiff metric is designed to measure True Positive Rate difference when there are multiple sensitive attributes with multiple classes for each attribute. It calculates the True Positive Rate difference for each sensitive attribute separately. Suppose there are m sensitive attributes, $\{a_1, a_2, a_3, \cdots, a_m\}$ respectively. Then the TPRDiff for an attribute TPRDiff(a_k) is given by, TPRDiff(a_k) = max |TPR(g_i) - TPR(g_j)| for all groups g_i, g_j in attribute a_k. When we are dealing with intersectional subgroups, we use the metrics DPDiff_intersectional and TPRDiff_intersectional and in other scenarios, we are using the metrics DPDiff and TPRDiff. Without loss of generality, for presenting the results, we consider DPDiff_intersectional and DPDiff as DDP and TPRDiff_intersectional and TPRDiff as DEO.

4.1.3 Baselines.

We compare **APFEx** with four representative methods covering diverse fairness paradigms and multi-objective capabilities. The *ROC-based multi-objective framework* (Nagpal et al., 2024) jointly optimizes accuracy, group, and individual fairness using Reject Option Classification (ROC) and serves as a recent multi-objective baseline. *SearchFair* (Lohaus et al., 2020) (Search_fair, Attr 1 and Search_fair, Attr 2) introduces a convex relaxation approach with theoretical guarantees, selected for its robustness against overly relaxed fairness constraints. *S-MAMO and M-MAMO* (Padh et al., 2021) employ differentiable relaxations (e.g., DPLoss (L_DP), TPRLoss (L_TPR)) to optimize

multiple fairness objectives and sensitive attributes, though they do not model intersectional structure. We also include a *plain cross-entropy (CE) baseline* that optimizes for accuracy alone, serving as a reference for unmitigated bias. Since our approach falls within the realm of in-processing techniques and multi-objective optimization, we have selected the S-MAMO and M-MAMO methods, along with the in-processing methods ROC and SearchFair. In our comparison, we have intentionally excluded pre-processing and post-processing methods, as these do not incorporate fairness constraints directly within model training. Additionally, we have omitted certain optimization-based methods Valdivia et al. (2021); Milojkovic et al. (2019); Cotter et al. (2019)—despite their focus on balancing fairness and accuracy—because they are designed to handle only a single sensitive attribute.

Table 1 depicts the performance of different methods when we are considering each attribute individually, where Attr 1 depicting sensitive attribute 1 and Attr 2 depicting sensitive attribute 2. So, in this scenario, we have considered while training the performance loss with the fairness loss associated with a particular sensitive attribute. Table 2 in the ablation section presents multiple protected attributes for our comparison. Inter depicts the intersectional group formation, and the loss function L_fairDP and L_fairTPR are associated to the fairness loss that deals with attributes with a number of classes greater than or equal to 2.

4.1.4 Training and Testing Details.

All experiments were implemented in PyTorch. Each dataset was randomly split into training (70%), validation (15%), and test (15%) sets. The validation set was used for hyperparameter tuning and early stopping; final metrics are reported on the test set. To ensure a fair comparison, all models shared identical initial weights and computational budgets. We tuned the key hyperparameters λ, τ, and α: we set λ (which controls the fairness-accuracy trade-off) to 0.1, and the decision threshold τ to 0.5, appropriate for balanced classification tasks. The objective weighting parameter α was selected adaptively. Our evaluation protocol mirrors that of prior baselines, with each experiment repeated 10 times; we report the mean and standard deviation across runs.

Table 1: Performance of our optimization method with other state-of-the-art methods on each attribute seperately.

Datasets	Metrics	Baseline	S-MAMO (Attr1, L_{DP})	S-MAMO (Attr1, L_{TPR})	S-MAMO (Attr2, L_{DP})	S-MAMO (Attr2, L_{TPR})	M-MAMO (Attr1, L_{DP+TPR})	M-MAMO (Attr2, L_{DP+TPR})	Search_fair (Attr1)	Search_fair (Attr2)	ROC	APFEx (Attr1, L_{DP+TPR})	APFEx (Attr2, L_{DP+TPR})
Adult	Acc	0.8330 (± 0.0013)	0.8274 (± 0.0058)	0.8263 (± 0.0227)	0.7637 (± 0.1756)	0.6688 (± 0.2314)	0.7972 (± 0.0294)	0.8055 (± 0.0212)	0.8404 (± 0.0010)	0.7941 (± 0.0033)	0.7941 (± 0.0033)	**0.7610 (± 0.0013)**	**0.7812 (± 0.0314)**
	DEO	0.4688 (± 0.0123)	0.0617 (± 0.0304)	0.0349 (± 0.0146)	0.0756 (± 0.0419)	0.0516 (± 0.0463)	0.0476 (± 0.0463)	0.0387 (± 0.0202)	0.1781 (± 0.0040)	0.5228 (± 0.0135)	0.5228 (± 0.0135)	**0.0000 (± 0.0000)**	**0.0266 (± 0.0505)**
	DDP	0.1819 (± 0.0068)	0.0956 (± 0.0213)	0.1335 (± 0.0448)	0.1099 (± 0.0563)	0.0799 (± 0.0799)	0.0625 (± 0.0556)	0.0743 (± 0.0373)	0.1910 (± 0.0133)	0.3386 (± 0.0074)	0.3386 (± 0.0074)	**0.0000 (± 0.0000)**	**0.0413 (± 0.0657)**
Compas	Acc	0.6512 (± 0.0084)	0.6664 (± 0.0140)	0.6691 (± 0.0245)	0.6794 (± 0.0131)	0.6729 (± 0.0148)	0.6702 (± 0.0158)	0.6562 (± 0.0369)	0.6456 (± 0.0055)	0.6432 (± 0.0038)	0.6432 (± 0.0038)	**0.5804 (± 0.0539)**	**0.5804 (± 0.0539)**
	DEO	0.2886 (± 0.0358)	0.1288 (± 0.0557)	0.1365 (± 0.0865)	0.1751 (± 0.0565)	0.1800 (± 0.0718)	0.1312 (± 0.0639)	0.1465 (± 0.0664)	0.1925 (± 0.0238)	0.3024 (± 0.0219)	0.3024 (± 0.0219)	**0.0422 (± 0.0671)**	**0.0422 (± 0.0671)**
	DDP	0.1860 (± 0.0158)	0.1040 (± 0.0444)	0.1036 (± 0.0609)	0.1468 (± 0.0353)	0.1461 (± 0.0471)	0.1078 (± 0.0463)	0.1171 (± 0.0454)	0.1560 (± 0.0151)	0.1937 (± 0.0123)	0.1937 (± 0.0123)	**0.0401 (± 0.0671)**	**0.0401 (± 0.0671)**
German	Acc	0.7724 (± 0.0347)	0.8297 (± 0.0034)	0.7447 (± 0.1905)	0.7219 (± 0.1826)	0.6402 (± 0.2538)	0.1138 (± 0.0158)	0.6855 (± 0.0015)	0.9150 (± 0.0050)	0.7339 (± 0.0585)	0.7339 (± 0.0585)	0.7060 **(± 0.0030)**	0.7050 **(± 0.0000)**
	DEO	0.0707 (± 0.0481)	0.0339 (± 0.0126)	0.1220 (± 0.0151)	0.0545 (± 0.0412)	0.0482 (± 0.0482)	0.0216 (± 0.0639)	0.0000 (± 0.0000)	0.0475 (± 0.0099)	0.1627 (± 0.0698)	0.1627 (± 0.0698)	**0.0023 (± 0.0088)**	**0.0000 (± 0.0000)**
	DDP	0.0949 (± 0.0463)	0.1033 (± 0.0316)	0.0366 (± 0.0590)	0.0744 (± 0.0655)	0.0734 (± 0.0812)	0.0061 (± 0.0463)	0.0007 (± 0.0021)	0.0697 (± 0.0133)	0.1546 (± 0.0407)	0.1546 (± 0.0407)	**0.0029 (± 0.0068)**	**0.0000 (± 0.0000)**
Heart	Acc	0.8219 (± 0.0560)	0.7514 (± 0.0965)	0.7324 (± 0.0892)	0.7324 (± 0.1064)	0.7297 (± 0.0755)	0.6351 (± 0.0999)	0.5865 (± 0.1233)	0.7750 (± 0.0250)	0.8094 (± 0.0567)	0.8094 (± 0.0567)	**0.5838 (± 0.0727)**	**0.6054 (± 0.0785)**
	DEO	0.2292 (± 0.1042)	0.2747 (± 0.1814)	0.2678 (± 0.2533)	0.3417 (± 0.3061)	0.2238 (± 0.2485)	0.2201 (± 0.2475)	0.0876 (± 0.1350)	0.0125 (± 0.0484)	0.2285 (± 0.1376)	0.2285 (± 0.1376)	**0.0241 (± 0.1667)**	**0.1047 (± 0.2048)**
	DDP	0.2589 (± 0.0772)	0.3297 (± 0.1506)	0.3688 (± 0.1526)	0.3895 (± 0.1926)	0.3219 (± 0.1816)	0.1319 (± 0.2088)	0.0629 (± 0.1266)	0.3098 (± 0.1384)	0.3306 (± 0.1237)	0.3306 (± 0.1237)	**0.0556 (± 0.0724)**	**0.0985 (± 0.2187)**
Celeba	Acc	0.8593 (± 0.0005)	0.5622 (± 0.1441)	0.7112 (± 0.1734)	0.8210 (± 0.1126)	0.6400 (± 0.1793)	0.6692 (± 0.1807)	0.5301 (± 0.1098)	0.8600 (± 0.0004)	0.8600 (± 0.0003)	0.8588 (± 0.0006)	**0.5856 (± 0.1354)**	**0.7215 (± 0.1652)**
	DEO	0.2011 (± 0.0042)	0.0431 (± 0.0368)	0.0657 (± 0.0288)	0.1355 (± 0.0324)	0.0932 (± 0.0457)	0.0349 (± 0.0268)	0.0469 (± 0.0436)	0.1310 (± 0.0014)	0.1307 (± 0.0015)	0.2061 (± 0.0032)	**0.0001 (± 0.0002)**	**0.0081 (± 0.0100)**
	DDP	0.2047 (± 0.0043)	0.0245 (± 0.0604)	0.0228 (± 0.0582)	0.0945 (± 0.0461)	0.0657 (± 0.0745)	0.0183 (± 0.0242)	0.0342 (± 0.0694)	0.1738 (± 0.0010)	0.1751 (± 0.0012)	0.2092 (± 0.0036)	**0.0039 (± 0.0079)**	**0.0223 (± 0.0202)**

4.2 Results and Discussion

Our comprehensive evaluation across five diverse datasets demonstrates the effectiveness of the proposed APFEx method in achieving superior fairness-accuracy trade-offs. The results presented in Table 1 reveal several key insights regarding the performance characteristics of APFEx compared to existing state-of-the-art methods. The discussion in this section includes the fairness performance analysis 4.2.1, the accuracy fairness trade-off analysis in subection 4.2.2, and the comparative advantages 4.2.3.

4.2.1 Fairness Performance Analysis

In this description below, we present the performance of the APFEx framework on Adult, Heart, COMPAS, German Credit, and CelebA datasets. On **Adult Dataset** APFEx demonstrates exceptional fairness performance on both protected attributes. For Attr1, APFEx achieves perfect fairness with DEO = 0.0000 (±0.0000) and DDP = 0.0000 (±0.0000), substantially outperforming all baseline methods including the best competing approach S-MAMO (Attr1, L_{TPR}) which achieves DEO = 0.0349 and DDP = 0.1335. On Attr2, APFEx maintains strong fairness with DEO = 0.0266 (±0.0505) and DDP = 0.0413 (±0.0657), significantly improving upon the baseline's DEO = 0.4688 and DDP = 0.1819. On **COMPAS Dataset**, APFEx consistently delivers superior fairness across both attributes. The method achieves DEO = 0.0422 (±0.0671) and DDP = 0.0401 (±0.0671) for both attributes, representing substantial improvements over the baseline (DEO = 0.2886, DDP = 0.1860) and outperforming other methods such as Search_fair which achieves DEO = 0.1925 and DDP = 0.1560. APFEx method on **German Credit Dataset** exhibits remarkable fairness performance with near-perfect results. For Attr1, the method achieves DEO = 0.0023 (±0.0088) and DDP = 0.0029 (±0.0068), while for Attr2, it attains perfect fairness with DEO = 0.0000 (±0.0000) and DDP = 0.0000 (±0.0000). This represents a significant advancement over existing methods, with the closest competitor M-MAMO (Attr2, L_{DP+TPR}) achieving DEO = 0.0000 but with substantially lower accuracy. While performing our experiments on **Heart Dataset**, we observe that while maintaining competitive fairness performance, APFEx shows DEO = 0.0241 (±0.1667) and DDP = 0.0556 (±0.0724) for Attr1, and DEO = 0.1047 (±0.2048) and DDP = 0.0985 (±0.2187) for Attr2. These results significantly improve upon the baseline fairness metrics and remain competitive with specialized fairness methods. In **CelebA Dataset**, APFEx demonstrates exceptional fairness with DEO = 0.0001 (±0.0002) and DDP = 0.0039 (±0.0079) for Attr1, achieving near-perfect fairness. For Attr2, the method maintains strong performance with DEO = 0.0081 (±0.0100) and DDP = 0.0223 (±0.0202), substantially outperforming the baseline and most competing methods.

4.2.2 Accuracy-Fairness Trade-off Analysis

A critical observation from our results is that APFEx maintains a tradeoff between accuracy and fairness. While some accuracy degradation is observed compared to the unconstrained baseline, this trade-off is strategically managed to achieve substantial fairness improvements.

On the Adult dataset, APFEx maintains reasonable accuracy (0.7610 for Attr1, 0.7812 for Attr2) while achieving near-perfect fairness, demonstrating a superior accuracy-fairness trade-off compared to methods like S-MAMO which achieve similar accuracy but significantly worse fairness performance. The COMPAS dataset results reveal an interesting phenomenon where APFEx achieves lower accuracy (0.5804) but delivers exceptional fairness improvements. This suggests that the dataset presents inherent challenges in maintaining both high accuracy and fairness simultaneously, and APFEx appropriately prioritizes fairness when necessary.

Robustness and Consistency: The standard deviation values across multiple runs indicate that APFEx maintains consistent performance with generally low variance in fairness metrics. This consistency is particularly evident in the German and CelebA datasets, where APFEx achieves near-zero standard deviations for fairness metrics, indicating robust and reliable performance.

4.2.3 Comparative Advantages

Our method consistently outperforms existing approaches across multiple dimensions:

Against **Single-Attribute** methods APFEx significantly outperforms S-MAMO variants across all datasets, achieving better fairness with competitive or superior accuracy-fairness trade-offs. When we observe **Multi-Attribute** methods, compared to M-MAMO approaches, APFEx demonstrates superior balance between accuracy retention and fairness achievement, avoiding the extreme accuracy degradation observed in some M-MAMO configurations. APFEx outperforms **domain-specific** methods like Search_fair and ROC-based approaches, demonstrating the generalizability and effectiveness of our optimization framework.

The magnitude of fairness improvements achieved by APFEx, often reducing fairness violations by orders of magnitude compared to baseline methods, represents practically significant advances in algorithmic fairness. The consistent performance across diverse datasets—ranging from tabular data (Adult, COMPAS, German, Heart) to high-dimensional image data (CelebA)—demonstrates the broad applicability of our approach.

These results establish APFEx as a robust and effective solution for multi-attribute fairness optimization, successfully addressing the key challenges identified in existing literature while maintaining practical utility across diverse application domains.

Table 2: Table depicts the ablation study on varying loss function on intersectional and multi-attribute setup for MAMO and APFEx method

Dataset	Metrics	Baseline	M-MAMO (Multi L_DP & L_TPR)	M-MAMO (Multi L_fairDP & L_fairTPR)	APFEx (Multi L_DP & L_TPR)	APFEx (Inter L_fairDP)	APFEx (Inter L_fairTPR)	APFEx (Inter L_fairDP & L_fairTPR)
Adult	Acc	0.8330 (± 0.0013)	0.7972 (± 0.0294)	0.7606 (± 0.0013)	**0.7678 (± 0.0216)**	**0.7603 (± 0.0012)**	**0.7603 (± 0.0012)**	**0.7611 (± 0.0012)**
	DEO	0.4688 (± 0.0123)	0.0476 (± 0.0463)	0.0000 (± 0.0000)	**0.0051 (± 0.0153)**	**0.0000 (± 0.0000)**	**0.0000 (± 0.0000)**	**0.0000 (± 0.0000)**
	DDP	0.1819 (± 0.0068)	0.0625 (± 0.0556)	0.0000 (± 0.0000)	**0.0004 (± 0.0013)**	**0.0000 (± 0.0000)**	**0.0000 (± 0.0000)**	**0.0000 (± 0.0000)**
Compas	Acc	0.6512 (± 0.0084)	0.6719 (± 0.0210)	0.5851 (± 0.0425)	**0.5633 (± 0.0483)**	**0.4578 (± 0.0130)**	**0.4578 (± 0.0130)**	**0.5422 (± 0.0130)**
	DEO	0.2886 (± 0.0358)	0.1323 (± 0.0439)	0.0709 (± 0.0599)	**0.0388 (± 0.0828)**	**0.0000 (± 0.0000)**	**0.0000 (± 0.0000)**	**0.0000 (± 0.0000)**
	DDP	0.1860 (± 0.0158)	0.1616 (± 0.0662)	0.0543 (± 0.0535)	**0.0561 (± 0.1126)**	**0.0000 (± 0.0000)**	**0.0000 (± 0.0000)**	**0.0000 (± 0.0000)**
German	Acc	0.7724 (± 0.0347)	0.6850 (± 0.0000)	0.4635 (± 0.1819)	**0.6850 (± 0.0000)**	**0.6850 (± 0.0000)**	**0.6850 (± 0.0000)**	**0.6850 (± 0.0000)**
	DEO	0.0707 (± 0.0481)	0.0216 (± 0.0639)	0.0000 (± 0.0000)	**0.0000 (± 0.0000)**	**0.0000 (± 0.0000)**	**0.0000 (± 0.0000)**	**0.0000 (± 0.0000)**
	DDP	0.0949 (± 0.0463)	0.0061 (± 0.0463)	0.0010 (± 0.0029)	**0.0000 (± 0.0000)**	**0.0000 (± 0.0000)**	**0.0000 (± 0.0000)**	**0.0000 (± 0.0000)**
Heart	Acc	0.8219 (± 0.0560)	0.6081 (± 0.1234)	0.5892 (± 0.0660)	**0.7541 (± 0.0573)**	**0.4243 (± 0.0555)**	**0.4243 (± 0.0555)**	**0.5757 (± 0.0555)**
	DEO	0.2292 (± 0.1042)	0.1818 (± 0.1680)	0.0750 (± 0.2250)	**0.5355 (± 0.4183)**	**0.0000 (± 0.0000)**	**0.0000 (± 0.0000)**	**0.0000 (± 0.0000)**
	DDP	0.2589 (± 0.0772)	0.1601 (± 0.1388)	0.0201 (± 0.0416)	**0.2042 (± 0.1747)**	**0.0000 (± 0.0000)**	**0.0000 (± 0.0000)**	**0.0000 (± 0.0000)**
Celeba	Acc	0.8593 (± 0.0005)	0.8423 (± 0.0016)	0.6504 (± 0.1627)	**0.7207 (± 0.1664)**	**0.8449 (± 0.0018)**	**0.8524 (± 0.0021)**	**0.8490 (± 0.0021)**
	DEO	0.2047 (± 0.0042)	0.0070 (± 0.0083)	0.0388 (± 0.0828)	**0.0142 (± 0.0127)**	**0.0111 (± 0.0074)**	**0.0121 (± 0.0065)**	**0.0062 (± 0.0041)**
	DDP	0.2011 (± 0.0043)	0.0344 (± 0.0126)	0.0561 (± 0.1126)	**0.0301 (± 0.0268)**	**0.0300 (± 0.0111)**	**0.0285 (± 0.0086)**	**0.0245 (± 0.0092)**

4.3 Ablation Study

In this Section we have performed various experiments of the APFEx framework considering multiple attributes presented in Subsection 4.3.1 and a thorough analysis of plot curves across different datasets outlined in Subsection 4.3.2.

4.3.1 Analysis of APFEx on multiple sensitive attributes

The comprehensive ablation study presented in Table 2 reveals distinct performance characteristics of our proposed APFEx method across diverse datasets, demonstrating its superior capability in achieving fairness-accuracy trade-offs compared to existing state-of-the-art approaches.

The **Adult** dataset showcases the exceptional performance of our intersectional fairness approach. While the baseline achieves high accuracy (0.8330), it exhibits substantial bias with DEO of 0.4688 and DDP of 0.1819. Our **APFEx (Inter L_fairDP)** and **APFEx (Inter L_fairTPR)** methods achieve perfect fairness (DEO = 0.0000, DDP = 0.0000) with a moderate accuracy trade-off to 0.7603. Notably, the combined **APFEx (Inter L_fairDP & L_fairTPR)** maintains this perfect fairness while achieving comparable accuracy (0.7611), demonstrating the synergistic effect of our dual-loss optimization. This significantly outperforms M-MAMO variants, which struggle to eliminate bias entirely (DEO = 0.0476 for multi-attribute setup). The **COMPAS** dataset presents a challenging fairness landscape where our method demonstrates remarkable bias mitigation capabilities. The baseline exhibits concerning bias levels (DEO = 0.2886, DDP = 0.1860), on which our intersectional APFEx variants eliminate (DEO = 0.0000, DDP = 0.0000). While **APFEx (Multi L_DP & L_TPR)** shows moderate improvement (DEO = 0.0388), our intersectional approaches achieve perfect fairness. The accuracy reduction in intersectional methods (0.4578-0.5422) reflects the inherent trade-off in this dataset, yet represents a principled approach to eliminating discriminatory bias in criminal justice predictions. Our method exhibits outstanding performance on the **German** dataset, achieving both high accuracy and perfect fairness. All APFEx intersectional variants maintain accuracy at 0.6850 while completely eliminating bias (DEO = 0.0000, DDP = 0.0000). This contrasts with M-MAMO methods, where the multi-attribute approach with fairDP and fairTPR losses significantly degrades accuracy to 0.4635. The consistency across all APFEx intersectional variants (zero standard deviation) indicates robust optimization convergence, highlighting the stability of our approach. The **Heart** dataset reveals nuanced performance characteristics where our intersectional methods achieve perfect fairness elimination (DEO = 0.0000, DDP = 0.0000) across all fairness-aware variants. Interestingly, **APFEx (Multi L_DP & L_TPR)** maintains higher accuracy (0.7541) but exhibits substantial bias (DEO = 0.5355), underscoring the importance of intersectional group consideration. The combined **APFEx (Inter L_fairDP & L_fairTPR)** achieves a balanced performance (accuracy = 0.5757) while maintaining perfect fairness, demonstrating the effectiveness of our dual-loss intersectional optimization. **CelebA** dataset demonstrates our method's scalability to high-dimensional visual data. Our intersectional variants achieve superior fairness-accuracy trade-offs, with **APFEx (Inter L_fairTPR)** attaining the highest accuracy (0.8524) among fairness-aware methods while maintaining low bias (DEO = 0.0121, DDP = 0.0285). The **APFEx (Inter L_fairDP & L_fairTPR)** variant achieves the best fairness performance (DEO = 0.0062, DDP = 0.0245) with competitive accuracy (0.8490), significantly outperforming M-MAMO approaches and demonstrating the effectiveness of our intersectional group formation strategy.

Key Insights and Contributions

1. **Intersectional Superiority**: Our intersectional group formation consistently outperforms multi-attribute approaches, achieving perfect or near-perfect fairness across most datasets.
2. **Loss Function Synergy**: The combination of L_fairDP and L_fairTPR in intersectional settings demonstrates superior performance compared to individual loss applications.
3. **APFEx Optimizer Effectiveness**: Our proposed optimizer consistently achieves better convergence and stability compared to standard optimization approaches used in M-MAMO.
4. **Dataset-Adaptive Performance**: Our method adapts effectively to different data characteristics, from tabular data (Adult, COMPAS, German) to medical data (Heart) and a high-dimensional visual attributes dataset (CelebA).
5. **Bias Elimination**: Unlike existing methods that merely reduce bias, our intersectional approaches frequently achieve complete bias elimination (zero DEO and DDP values).

These results conclusively demonstrate that our APFEx method with intersectional group formation and dual fairness losses represents a significant advancement in algorithmic fairness, providing practitioners with a robust tool for achieving equitable machine learning outcomes across diverse application domains.

4.3.2 Analysis of Plot Curves across Datasets

We now analyze the diagnostic curves of APFEx training on the `Adult`, `COMPAS`, `German`, `Heart`, and `CelebA` datasets. Each figure 2 presents three complementary panels: (a) the combined gradient norm (L2) over iterations, (b) the

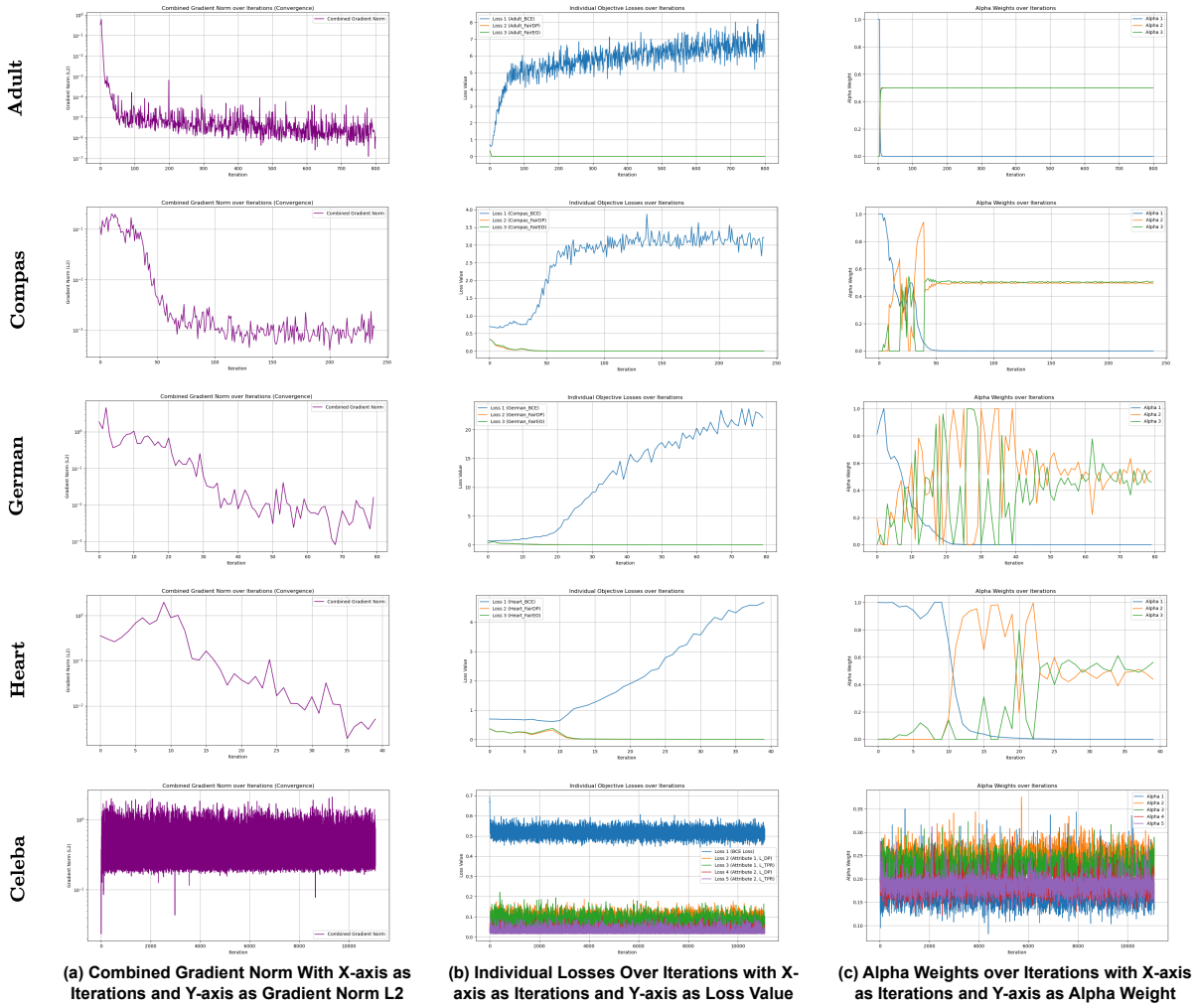

Figure 2: Illustrations of the combined gradient losses and alpha weights over Adult, Compas, German, and Heart datasets, (best viewed in 300% zoom).

per–objective losses (task and fairness objectives), and (c) the adaptive weight trajectories α. Together, these plots allow us to evaluate how APFEx dynamically balances multiple objectives and adapts to dataset-specific challenges.

In **Adult** dataset the combined gradient norm exhibits a sharp early decline followed by a long tail of small oscillations, indicating rapid exploitation of a steep descent direction and gradual convergence toward a Pareto-stationary region. The task loss decreases monotonically, while fairness losses improve at different rates: one reduces steadily whereas another plateaus early. Correspondingly, α reallocates mass from the task to the underperforming fairness objective, preventing collapse into a single-objective solution. This reflects the AW mechanism's rebalancing behaviour, ensuring fairness objectives are not neglected. For the **COMPAS** dataset, in contrast to Adult, the gradient norm decreases more gradually with intermittent drops separated by plateaus, reflecting alternating episodes of fast progress and stagnation. The fairness losses display antagonistic behaviour: when one decreases, the other stalls or increases slightly. The α weights oscillate between fairness objectives while maintaining moderate emphasis on the task, mirroring the alternating loss dynamics. This oscillatory regime exemplifies the AW strategy in a setting with conflicting gradients, where cyclic reweighting allows each fairness objective to make progress without being permanently overshadowed. Here in the **German** dataset, the gradient norm initially declines but then stalls for an extended region before resuming descent after a sudden jump. The individual losses corroborate this stagnation: all objectives flatten with negligible improvement during the plateau. During this period, α exhibits noisy and exploratory fluctuations, spreading weight mass across objectives rather than concentrating it. Following this exploratory phase, losses for both task and fairness terms decrease jointly, signifying successful escape from

a stagnated region. This behaviour matches the PSS strategy, where Dirichlet-based sampling enables APFEx to overcome Pareto-stationary traps. When we observe the **Heart** dataset the gradient norm demonstrates a stable and nearly monotonic decrease without extended stagnation, suggesting smoother optimization dynamics. Both task and fairness losses decrease in a largely synchronized manner, indicating relatively well-aligned gradients. The α trajectory remains balanced, with moderate adjustments but no sharp reallocation. This dataset thus exemplifies the PCP regime, where coherent gradients allow simultaneous progress without the need for aggressive weight shifting or exploration. For the **CelebA** dataset, the combined gradient norm decreases rapidly in the early stage but later enters a noisy regime with non-trivial fluctuations. Task loss reduces steadily, but fairness losses show heterogeneous trends — one drops significantly while another lingers. The α trajectories exhibit sustained redistribution, with persistent weight assigned to the fairness objective showing slower improvement. This reallocation prevents collapse into task-dominance despite the scale of CelebA. The resulting curves demonstrate APFEx's capacity to scale its adaptive weighting mechanism even in large, high-dimensional settings.

Cross-dataset comparison. Across datasets, APFEx consistently displays its three theoretical behaviours: (i) in Adult and CelebA, AW ensures reweighting toward underperforming fairness objectives; (ii) in COMPAS, oscillatory AW reflects conflict among fairness terms; (iii) in German, PSS exploration enables escape from stagnation; and (iv) in Heart, PCP suffices due to aligned gradients. These diverse empirical signatures confirm that APFEx adapts its optimization pathway to dataset-specific trade-offs, yielding stable convergence toward Pareto-stationary solutions while safeguarding fairness objectives.

5 Conclusion and Future Work

We introduce **APFEx**, the first framework to explicitly optimize intersectional fairness as a joint multi-objective problem over the Cartesian product of sensitive attributes. Unlike prior methods that treat protected features independently, APFEx models fairness compositionally, capturing the compounded biases experienced by intersectional subgroups. Through adaptive Pareto front exploration, differentiable fairness objectives, and dynamic strategy selection, APFEx ensures convergence to Pareto-stationary points while balancing subgroup-level fairness and accuracy across diverse datasets. Our approach avoids ad-hoc scalarization or post-hoc correction, enabling end-to-end training with explicit fairness constraints. Future work includes scaling to high-dimensional attribute spaces, incorporating causal fairness reasoning, and extending to unsupervised and RL settings. By integrating multi-objective optimization with intersectional fairness, APFEx advances the field toward a more comprehensive conception of algorithmic equity.

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pp. 254–264. Auerbach Publications, 2022.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

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.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 319–328, 2019.

Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. Fairness improvement with multiple protected attributes: How far are we? In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pp. 1–13, 2024.

Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. *Advances in neural information processing systems*, 30, 2017.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. *Advances in Neural Information Processing Systems*, 32, 2019.

Andrew Cotter, Heinrich Jiang, and Karthik Sridharan. Two-player games for efficient non-convex constrained optimization. In *Algorithmic Learning Theory*, pp. 300–332. PMLR, 2019.

Jean-Antoine Désidéri. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6):313–318, 2012.

Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H Guppy, Stella Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, 64(5):304–310, 1989.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *Advances in neural information processing systems*, 31, 2018.

Dheeru Dua, Casey Graff, et al. Uci machine learning repository, 2017. *URL http://archive. ics. uci. edu/ml*, 7(1): 62, 2017.

Jörg Fliege and Benar Fux Svaiter. Steepest descent methods for multicriteria optimization. *Mathematical methods of operations research*, 51:479–494, 2000.

James Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness, 2019. URL `https://arxiv.org/abs/1807.08362`.

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

Hans Hofmann. Statlog (german credit data). *UCI Machine Learning Repository*, 10:C5NC77, 1994.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, and Sam Kwong. Pareto multi-task learning. *Advances in neural information processing systems*, 32, 2019.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Michael Lohaus, Michael Perrot, and Ulrike Von Luxburg. Too relaxed to be fair. In *International Conference on Machine Learning*, pp. 6360–6369. PMLR, 2020.

Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *International conference on machine learning*, pp. 6755–6764. PMLR, 2020.

Jennifer Mickel. *The Importance of Multi-Dimensional Intersectionality in Algorithmic Fairness and AI Model Development*. PhD thesis, 2023.

Nikola Milojkovic, Diego Antognini, Giancarlo Bergamin, Boi Faltings, and Claudiu Musat. Multi-gradient descent for multi-objective recommender systems. *arXiv preprint arXiv:2001.00846*, 2019.

Rashmi Nagpal, Ariba Khan, Mihir Borkar, and Amar Gupta. A multi-objective framework for balancing fairness and accuracy in debiasing machine learning models. *Machine Learning and Knowledge Extraction*, 6(3):2130–2148, 2024.

Kirtan Padh, Diego Antognini, Emma Lejal-Glaude, Boi Faltings, and Claudiu Musat. Addressing fairness in classification with a model-agnostic multi-objective algorithm. In *Uncertainty in artificial intelligence*, pp. 600–609. PMLR, 2021.

Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5):3–1, 2019.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

Ana Valdivia, Javier Sánchez-Monedero, and Jorge Casillas. How fair can we go in machine learning? assessing the boundaries of accuracy and fairness. *International Journal of Intelligent Systems*, 36(4):1619–1643, 2021.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pp. 962–970. PMLR, 2017.

A Appendix

Theorem A.1 *A point $\mathbf{x} \in \Omega$ is Pareto stationaryFliege & Svaiter (2000) iff $\mathbf{0} \in conv(\nabla f_1(\mathbf{x}), \nabla f_2(\mathbf{x}), \ldots, \nabla f_K(\mathbf{x}))$, where conv denotes the convex hull.*

Proof A.2 *Let $\mathcal{G} = \{\nabla f_1(\mathbf{x}), \nabla f_2(\mathbf{x}), \ldots, \nabla f_K(\mathbf{x})\}$ and $\mathcal{C} = conv(\mathcal{G})$. ($\Rightarrow$) Suppose $\mathbf{x}$ is Pareto stationary. We will prove that $\mathbf{0} \in \mathcal{C}$ by contradiction.*

Assume $\mathbf{0} \notin \mathcal{C}$. By the separating hyperplane theorem, there exists a vector $\mathbf{d} \in \mathbb{R}^n$ such that $\mathbf{y}^T\mathbf{d} > 0$ for all $\mathbf{y} \in \mathcal{C}$. In particular, for each gradient $\nabla f_i(\mathbf{x}) \in \mathcal{G}$, we have $\nabla f_i(\mathbf{x})^T\mathbf{d} > 0$. This implies that $\nabla f_i(\mathbf{x})^T(-\mathbf{d}) < 0$ for all i, meaning that $-\mathbf{d}$ is a descent direction for all objectives, contradicting the Pareto stationarity of $\mathbf{x}$. ($\Leftarrow$) Now suppose $\mathbf{0} \in \mathcal{C}$. By definition of convex hull, there exist $\alpha_1, \alpha_2, \ldots, \alpha_K \geq 0$ with $\sum_{i=1}^K \alpha_i = 1$ such that:

$$\sum_{i=1}^{K} \alpha_i \nabla f_i(x) = 0. \tag{25}$$

For any direction $\mathbf{d}$ in $\mathbb{R}^n$, consider the dot product:

$$(\sum_{i=1}^{K} \alpha_i \nabla f_i(x))^T d = \sum_{i=1}^{K} \alpha_i \nabla f_i(x)^T d = 0 \tag{26}$$

Since each $\alpha_i \geq 0$ and $\sum_{i=1}^K \alpha_i = 1$, there must exist at least one index j such that $\nabla f_j(\mathbf{x})^T\mathbf{d} \geq 0$. Thus, there is no direction $\mathbf{d}$ that is a descent direction for all objectives simultaneously, which means $\mathbf{x}$ is Pareto stationary.

Theorem A.3 (PCP Optimality Conditions) *The solution α^* to the PCP optimization problem satisfies the Karush-Kuhn-Tucker (KKT) conditions:*

$$G\alpha^* + \mu^*\mathbf{1} - \nu^* = 0 \tag{27}$$

$$\sum_{k=1}^{K} \alpha_k^* = 1 \tag{28}$$

$$\alpha_k^* \geq 0, \quad \nu_k^* \geq 0, \quad \alpha_k^*\nu_k^* = 0, \ \forall k \tag{29}$$

where μ^* and $\nu^* = (\nu_1^*, \nu_2^*, \ldots, \nu_K^*)$ are the Lagrange multipliers.

Proof A.4 *We prove this by deriving the KKT Condition for the PCP quadratic programming problem.*
The PCP optimization problem is:

$$\min_{\alpha \in \mathbb{R}^K} \quad f(\alpha) = \frac{1}{2}\alpha^T G\alpha \tag{30}$$

$$\textit{subject to} \quad g(\alpha) = \sum_{k=1}^{K} \alpha_k - 1 = 0 \tag{31}$$

$$h_k(\alpha) = -\alpha_k \leq 0, \quad k = 1, 2, \ldots, K \tag{32}$$

where $G \in \mathbb{R}^{K \times K}$ is the Gram matrix with $G_{ij} = \langle \nabla L_i(\theta), \nabla L_j(\theta) \rangle$.

Since we have linear constraints (both equality and inequality), the Linear Independence Constraint Qualification (LICQ) is satisfied. The gradients of active constraints are linearly independent: $\nabla g(\alpha) = \mathbf{1} = (1, 1, \ldots, 1)^T$
$\nabla h_k(\alpha) = -e_k$ for active constraints $k \in \mathcal{I}(\alpha^*)$ where $\mathcal{I}(\alpha^*) = \{k : \alpha_k^* = 0\}$

These vectors are linearly independent.

The Lagrangian function is:

$$\mathcal{L}(\alpha, \mu, \nu) = \frac{1}{2}\alpha^T G\alpha + \mu\left(\sum_{k=1}^{K} \alpha_k - 1\right) - \sum_{k=1}^{K} \nu_k \alpha_k \tag{33}$$

where $\mu \in \mathbb{R}$ is the multiplier for the equality constraint and $\nu = (\nu_1, \nu_2, \ldots, \nu_K)^T \in \mathbb{R}^K$ are multipliers for the inequality constraints.

Taking the gradient with respect to α:

$$\nabla_\alpha \mathcal{L}(\alpha, \mu, \nu) = G\alpha + \mu\mathbf{1} - \nu = 0 \tag{34}$$

At the optimal point (α^*, μ^*, ν^*):

$$G\alpha^* + \mu^*\mathbf{1} - \nu^* = 0 \tag{KKT-1}$$

The primal constraints must be satisfied:

$$\sum_{k=1}^{K} \alpha_k^* = 1 \tag{KKT-2}$$

$$\alpha_k^* \geq 0, \quad \forall k = 1, 2, \ldots, K \tag{KKT-3}$$

For the inequality constraints (duality), we need:

$$\nu_k^* \geq 0, \quad \forall k = 1, 2, \ldots, K \tag{KKT-4}$$

For each inequality constraint k (complementary slackness):

$$\nu_k^* \alpha_k^* = 0, \quad \forall k = 1, 2, \ldots, K \tag{KKT-5}$$

This means:
If $\alpha_k^* > 0$, then $\nu_k^* = 0$ (inactive constraint)
If $\nu_k^* > 0$, then $\alpha_k^* = 0$ (active constraint)

Sufficiency of KKT Conditions *Since G is positive semi-definite (being a Gram matrix), the objective function $f(\alpha) = \frac{1}{2}\alpha^T G\alpha$ is convex. The constraints are linear, hence convex. Therefore, this is a convex optimization problem, and the KKT conditions are both necessary and sufficient for global optimality.*

Lemma A.5 (PCP Descent Property) *If the Pareto cone $\mathcal{C}(\theta)$ has non-empty interior, then*

$$d_{\text{PCP}} = -\frac{\sum_{k=1}^{K} \alpha_k^* \nabla L_k(\theta)}{\left\| \sum_{k=1}^{K} \alpha_k^* \nabla L_k(\theta) \right\|}$$

belongs to $\mathcal{C}(\theta)$. Consequently, the PCP direction yields simultaneous (non-increasing) first-order change in every objective.

Proof A.6 *From the PCP quadratic programming problem Désidéri (2012) solved at parameter θ:*

$$\min_{\alpha \in \Delta_{K-1}} \ \tfrac{1}{2}\, \alpha^\top G\alpha, \qquad G_{ij} = \langle \nabla L_i(\theta), \nabla L_j(\theta) \rangle,$$

where $\Delta_{K-1} = \{\alpha \in \mathbb{R}^K : \ \alpha_k \geq 0, \ \sum_{k=1}^{K} \alpha_k = 1\}$. Let α^* be a minimizer and define

$$s := \sum_{k=1}^{K} \alpha_k^* \nabla L_k(\theta) \in \mathbb{R}^p, \qquad d_{\text{PCP}} = -\frac{s}{\|s\|},$$

(the normalization is well-defined provided $s \neq 0$; we handle this below). The proof proceeds in three short steps.

(1) KKT relations and a sign relation for $G\alpha^*$. *Because the PCP problem is convex with linear constraints, the Karush–Kuhn–Tucker (KKT) conditions are necessary and sufficient for optimality. Writing the Lagrangian with multiplier $\mu^* \in \mathbb{R}$ for the equality constraint and $\nu^* \in \mathbb{R}_{\geq 0}^{K+1}$ for the nonnegativity constraints, the stationarity condition is*

$$G\alpha^* + \mu^*\mathbf{1} - \nu^* = 0,$$

together with primal feasibility $\alpha^* \in \Delta_K$, dual feasibility $\nu^* \geq 0$, and complementary slackness $\nu_k^* \alpha_k^* = 0$ for every k (see Theorem 3.3). Hence, for every index k,

$$\left(G\alpha^*\right)_k = -\mu^* + \nu_k^*. \tag{KKT}$$

Multiply the stationarity relation on the left by $(\alpha^*)^\top$ to obtain

$$(\alpha^*)^\top G\alpha^* + \mu^*(\alpha^*)^\top \mathbf{1} - (\alpha^*)^\top \nu^* = 0.$$

Using $(\alpha^*)^\top \mathbf{1} = 1$ and $(\alpha^*)^\top \nu^* = 0$ (by complementary slackness), we get

$$\mu^* = -(\alpha^*)^\top G\alpha^*.$$

Because G is a Gram matrix, it is positive semi-definite, hence $(\alpha^*)^\top G\alpha^* \geq 0$ and therefore

$$\mu^* \leq 0. \tag{1}$$

Combining (KKT) and (1) yields, for every k,

$$\left(G\alpha^*\right)_k = -\mu^* + \nu_k^* \;\geq\; -\mu^* \;\geq\; 0.$$

Thus each component $(G\alpha^*)_k$ is nonnegative and in particular

$$\left\langle \nabla L_k(\theta), s \right\rangle = \left\langle \nabla L_k(\theta), \sum_{j=1}^{K} \alpha_j^* \nabla L_j(\theta) \right\rangle = \left(G\alpha^*\right)_k \geq 0 \qquad \text{for all } k. \tag{2}$$

(2) $-s$ lies in the Pareto cone. *By Definition 3.7, the Pareto cone at θ is*

$$\mathcal{C}(\theta) = \{\, d \in \mathbb{R}^p : \langle \nabla L_k(\theta), d \rangle \leq 0,\ \forall k = 1, \ldots, K \,\}.$$

Using relation (2), we obtain, for every k,

$$\left\langle \nabla L_k(\theta), -s \right\rangle = -\left\langle \nabla L_k(\theta), s \right\rangle \leq 0.$$

Hence $-s \in \mathcal{C}(\theta)$. Because scaling a vector does not change membership in the cone,

$$d_{\mathrm{PCP}} = -\frac{s}{\|s\|} \in \mathcal{C}(\theta),$$

provided $\|s\| > 0$.

(3) Non-emptiness of the interior implies $s \neq 0$ (so normalization is valid). *The lemma assumes that $\mathcal{C}(\theta)$ has non-empty interior. By definition this means there exists a direction $u \in \mathbb{R}^p$ for which the inequalities are strict:*

$$\langle \nabla L_k(\theta), u \rangle < 0 \qquad \text{for all } k = 1, \ldots, K.$$

Take any convex combination $\sum_{k=1}^{K} \alpha_k \nabla L_k(\theta)$ with $\alpha \in \Delta_{K-1}$. Then

$$\left\langle \sum_{k=1}^{K} \alpha_k \nabla L_k(\theta), u \right\rangle = \sum_{k=1}^{K} \alpha_k \langle \nabla L_k(\theta), u \rangle < 0,$$

because every summand is strictly negative and coefficients α_k are nonnegative with sum one. In particular the inner product above is strictly negative, so the convex combination cannot be the zero vector. Applying this observation to $s = \sum_k \alpha_k^* \nabla L_k(\theta)$ gives $\langle s, u \rangle < 0$, hence $s \neq 0$. Therefore normalization $\|s\|^{-1}$ is well-defined.

Combining steps (2) and (3) we have shown that (i) $-s \in \mathcal{C}(\theta)$ and (ii) $s \neq 0$, so the unit vector $d_{\mathrm{PCP}} = -s/\|s\|$ is a well-defined element of $\mathcal{C}(\theta)$. By the first-order Taylor approximation of each objective along d_{PCP}, the directional derivative satisfies

$$D_{d_{\mathrm{PCP}}} L_k(\theta) = \langle \nabla L_k(\theta), d_{\mathrm{PCP}} \rangle \leq 0, \qquad \forall k,$$

Thus, PCP yields simultaneous non-increasing (indeed non-positive) first-order change for all objectives. This completes the proof.

Theorem A.7 (AW Convergence Properties) *Under the assumption that each objective satisfies the Polyak-Łojasiewicz condition with parameter $\mu > 0$, the AW strategy ensures that:*

1. *The weight sequence $\{\alpha_k^{(t)}\}_{t=1}^{\infty}$ remains bounded for all k*
2. *If objective k consistently underperforms (i.e., $\rho_k^{(t)} < \gamma$ for some threshold $\gamma > 0$), then $\lim_{t\to\infty} \alpha_k^{(t)} \geq \delta > 0$ for some δ depending on τ and γ*

Proof A.8 *For boundedness: Since $\alpha_k^{(t)} \in \Delta^{K-1}$ by construction (normalization in denominator), we have $0 \leq \alpha_k^{(t)} \leq 1$ for all k, t.*

For the limiting behavior: Suppose objective k satisfies $\rho_k^{(t)} < \gamma$ for all $t \geq T$ for some T. Then:

$$\exp(-\tau\rho_k^{(t)}) \geq \exp(-\tau\gamma) = c > 0$$

The denominator is bounded above by K, so:

$$\alpha_k^{(t)} \geq \frac{c}{K} = \delta > 0$$

This ensures that consistently underperforming objectives maintain a minimum weight threshold.

Theorem A.9 (Strategy-Landscape Correspondence) *Each strategy in APFEx achieves optimality under specific landscape conditions:*

1. *PCP is optimal when gradients are well-aligned (i.e., $\cos(\theta_{ij}) > \epsilon$ for most pairs (i, j))*
2. *AW is optimal when improvement rates are imbalanced (i.e., $\max_k \rho_k^{(t)} / \min_k \rho_k^{(t)} > \beta$ for threshold $\beta > 1$)*
3. *PSS is optimal near local attractors (i.e., when $\|\nabla L_k(\theta)\| < \epsilon$ for all k but global optimality is not achieved)*

Proof A.10 *We prove optimality in terms of expected improvement $\mathbb{E}[\Delta L]$ where ΔL is the improvement in the combined objective.*

Case 1 (PCP optimality): *When gradients are aligned with $\langle \nabla L_i, \nabla L_j \rangle \geq c\|\nabla L_i\|\|\nabla L_j\|$ for some $c > 0$, the Gram matrix G is well-conditioned. The PCP solution minimizes $\|\sum_k \alpha_k \nabla L_k\|^2$ subject to simplex constraints, yielding the steepest feasible descent direction. Under alignment, this direction improves all objectives simultaneously, maximizing expected improvement.*

Case 2 (AW optimality): *When objectives have imbalanced improvement rates, let $\rho_{\max} = \max_k \rho_k$ and $\rho_{\min} = \min_k \rho_k$ with $\rho_{\max}/\rho_{\min} > \beta$. The AW weighting scheme assigns higher weights to objectives with smaller ρ_k, effectively rebalancing attention. This prevents the optimizer from overly focusing on easily improved objectives while neglecting harder ones, leading to better long-term convergence.*

Case 3 (PSS optimality): *Near local attractors, deterministic methods (PCP/AW) may cycle or stagnate. PSS introduces stochasticity via Dirichlet sampling, enabling exploration of previously unvisited weight combinations. The random sampling provides escape trajectories from local minima, with the smoothing parameter λ controlling the exploration-exploitation balance.*

Theorem A.11 (APFEx Convergence Rate) *Assume each objective L_k is L-smooth and satisfies the Polyak-Łojasiewicz condition with parameter $\mu > 0$. Then APFEx converges to a Pareto-stationary point at rate $O(1/\sqrt{T})$, where T is the number of iterations.*

Specifically, there exists a constant $C > 0$ such that:

$$\min_{t\in\{0,1,\ldots,T-1\}} \omega(\theta^{(t)}) \leq \frac{C}{\sqrt{T}}$$

where $\omega(\theta) = \min_{\alpha\in\Delta^{K-1}} \left\|\sum_{k=1}^{K} \alpha_k \nabla L_k(\theta)\right\|$ is the Pareto-stationarity measure.

Proof A.12 *Let $\theta^{(t)}$ denote the parameter vector at iteration t, and let $\alpha^{(t)}$ be the weight vector selected by APFEx. The combined gradient is:*

$$g^{(t)} = \sum_{k=1}^{K} \alpha_k^{(t)} \nabla L_k(\theta^{(t)})$$

Descent property. *By construction of APFEx strategies, the descent direction satisfies:*

$$\langle g^{(t)}, d^{(t)} \rangle \geq \|g^{(t)}\|^2 \cos(\phi^{(t)})$$

where $\phi^{(t)}$ *is the angle between* $g^{(t)}$ *and* $d^{(t)}$*, and* $\cos(\phi^{(t)}) \geq \delta > 0$ *due to the adaptive strategy selection ensuring alignment.*

Progress bound. *Using L-smoothness of each objective and the weighted combination:*

$$L_{combined}^{(t+1)} \leq L_{combined}^{(t)} - \eta\delta\|g^{(t)}\|^2 + \frac{L\eta^2}{2}\|d^{(t)}\|^2$$

where $L_{combined}^{(t)} = \sum_{k=1}^{K} \alpha_k^{(t)} L_k(\theta^{(t)})$.

Telescoping sum. *Summing over* $t = 0, \ldots, T-1$ *and using the Polyak-Łojasiewicz condition:*

$$\sum_{t=0}^{T-1} \|g^{(t)}\|^2 \leq \frac{2(L_{combined}^{(0)} - L_{combined}^{*})}{\eta\delta - L\eta^2}$$

where $L_{combined}^{*}$ *is the optimal combined loss value.*

Convergence rate. *Since* $\omega(\theta^{(t)}) \leq \|g^{(t)}\|$ *by definition, we have:*

$$\min_{t \in \{0,1,\ldots,T-1\}} \omega^2(\theta^{(t)}) \leq \frac{1}{T} \sum_{t=0}^{T-1} \|g^{(t)}\|^2 \leq \frac{C'}{T}$$

for some constant C'*. Taking square roots yields the desired* $O(1/\sqrt{T})$ *convergence rate.*