CROSSROADS OF AI & SOCIETY
JULY 15–16 2026 · PARIS
ARXIV:2601.13448

# Fairness-informed Pareto Optimization: an Efficient Bilevel Framework

**SOFIANE TANJI**[1] · **SAMUEL VAITER**[2] · **YASSINE LAGUEL**[3] [1] UCLouvain · [2] CNRS, LJAD · [3] Université Côte d'Azur

MOTIVATION

## Fair learning as multi-objective optimization

Data-driven decisions in hiring, credit, or health care can perform unevenly across sensitive groups $a \in \{1, \dots, S\}$. Each group's performance is measured by its own empirical loss

$$F_a(w) = \frac{1}{n_a} \sum_{i:a_i=a} \ell(f(w, x_i), y_i),$$

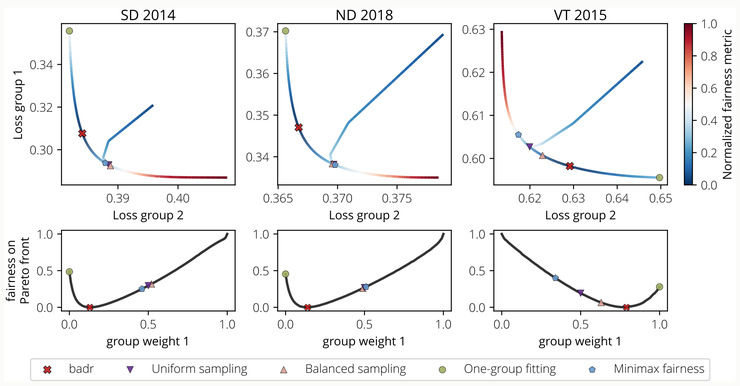
and fairness becomes a multi-objective problem:

$$\min_{w \in \mathbb{R}^d} \; [F_1(w), \dots, F_S(w)].$$

A model is **Pareto efficient** if no group's loss can be reduced without increasing another's. In-processing methods based on fairness constraints or penalties do not guarantee this property.

EXISTING ISSUE

## Penalization leaves the Pareto front




→ models trained with an individual-fairness penalty of increasing weight trace a path that departs from the Pareto front of group losses: penalized models are dominated.

SCALARIZATION

## The weighted-sum method

Given group weights $\lambda \in \Delta_S$, the weighted-sum method replaces the multi-objective problem with

$$w_\lambda = \arg\min_{w \in \mathbb{R}^d} \sum_{a=1}^{S} \lambda_a F_a(w).$$

For strongly convex losses, every $\lambda$ yields a Pareto-efficient model, and $\lambda \mapsto w_\lambda$ spans the whole front (Kuhn & Tucker, 1951).

Existing Pareto-efficient approaches correspond to fixed choices of $\lambda$:

- ERM: group proportions $\lambda_a = n_a/n$
- Balanced sampling: $\lambda_a \propto 1/n_a$
- Minimax fairness / DRO: adversarial worst-case $\lambda$

The effect of a fixed scalarization on a given fairness metric is neither controlled nor predictable.

THE IDEA

**Treat the group weights $\lambda$ as variables, and learn the weights whose Pareto-efficient model minimizes a chosen fairness metric.**

FORMULATION

## BADR: Bilevel ADaptive Rescalarization

For any differentiable (or smoothable) unfairness metric $\mathrm{Fair} : \mathbb{R}^d \to \mathbb{R}$, we select the point of the Pareto front minimizing it:

$$\min_{\lambda \in \Delta_S} \; \mathrm{Fair}(w_\lambda)$$

$$\text{s.t.} \;\; w_\lambda \in \arg\min_{w \in \mathbb{R}^d} \sum_{a=1}^{S} \lambda_a F_a(w).$$

ALGORITHM

### badr-sgd

1 **input** $w_0, v_0, \lambda_0$; **steps** $\tau, \rho, \gamma$; $C_\gamma$
2 **for** $t = 0$ **to** $T-1$ **do**
3 $\quad w_{t+1} \leftarrow w_t - \tau \tilde{\nabla} F(w_t)^\top \lambda_t$
4 $\quad v_{t+1} \leftarrow v_t - \rho\Big(\tilde{\nabla}\,\mathrm{Fair}(w_t) + \sum_{a=1}^{S} \lambda_{t,a} \tilde{\nabla}^2 F_a(w_t)\, v_t\Big)$
5 $\quad \lambda_{t+1} \leftarrow \mathrm{Proj}_{\Delta_S}\Big(\lambda_t - \gamma\, \mathrm{Clip}_{C_\gamma}\Big(\tilde{\nabla} F(w_t)\, v_t\Big)\Big)$
6 **return** $w_T, \lambda_T$

Exact gradients give the deterministic variant `badr-gd`, where **no clipping is needed**: clipping only controls the stochastic error of the hypergradient estimate.

ASSUMPTIONS

## Weaker assumptions

We ask for the standard assumptions in bilevel literature:

- Fair is Lipschitz with Lipschitz gradient on $\mathbb{R}^d$.
- Each $F_a$ is $\mu$-strongly convex with Lipschitz gradient and Lipschitz Hessian; hence $w_\lambda$ is unique.
- Stochastic oracles $\tilde{\nabla} F_a$, $\tilde{\nabla}^2 F_a$, $\tilde{\nabla}\,\mathrm{Fair}$ are unbiased with bounded variance $\mathcal{V}$.

but we **weaken** the usual smoothness requirement. The cross derivative $\nabla^2_{w\lambda} \mathcal{F}(w, \lambda) = [\nabla F_1(w) \cdots \nabla F_S(w)]$ is Lipschitz, and we bound it *only* on the solution set $\{(w_\lambda, \lambda') : \lambda, \lambda' \in \Delta_S\}$, not on all of $\mathbb{R}^d \times \Delta_S$.

This is essential here: since $\nabla F_a$ is unbounded on $\mathbb{R}^d$, the lower objective is **not globally smooth in** $\lambda$, so existing single-loop analyses (e.g. SOBA) do not apply.

GUARANTEES

## Both levels converge

THEOREM · CONVERGENCE OF BADR-SGD

With $\tau, \rho \propto 1/L$ and $\gamma$ small enough, for all $T \geqslant 1$,

$$\frac{1}{T} \sum_{t<T} \mathbb{E}\|\overline{D}_{\lambda,t}\| + \tilde{\alpha}^{\frac{1}{2}} \mathbb{E}\|w_t - w_{\lambda_t}\| \leqslant \mathcal{S}\left[\frac{8\Delta_\varphi}{\gamma T} + \frac{C_0}{T} + M^\top \mathcal{V} + r\, C_\gamma^2\right]$$

with $\varphi(\lambda) = \mathrm{Fair}(w_\lambda)$ the implicit objective, $\overline{D}_{\lambda,t}$ its projected gradient, $\mathcal{V}$ the oracle variances, $\mathcal{S}[u] = u/C_\gamma + \sqrt{u}$. Mini-batches and horizon $\propto \varepsilon^{-2}$ give an $\varepsilon$-stationary point at cost $\mathcal{O}(\varepsilon^{-4})$, matching the stochastic bilevel lower bound.

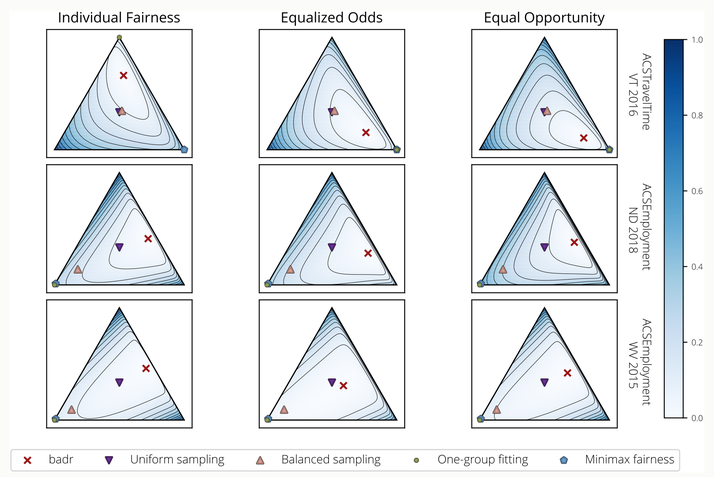
The proof rests on a new Lyapunov function

$$\psi_t = \mathbb{E}\Big[\varphi(\lambda_t) + \alpha\|w_t - w_{\lambda_t}\|^2 + \beta\|v_t - v_{\lambda_t}\|^2 + \delta\|w_t - w_{\lambda_t}\|^2\|v_t\|^2\Big].$$

As $\psi_t$ dominates both the upper-level gradient and the lower-level error, its decrease certifies **both levels at once**: stationarity of $\varphi$ and optimality of $w_t$.

EXPERIMENT 1

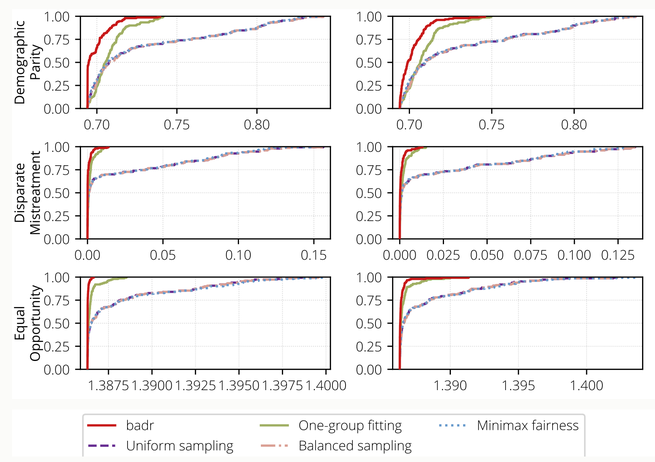
## Fixed scalarization misses the optimum




→ Normalized unfairness over the simplex of group weights (3 groups; 3 datasets × 3 metrics): scalarization baselines miss the minimizer, which badr attains.

EXPERIMENT 2

## Fairer on both train and test




→ ECDFs over 255 datasets: badr concentrates at low unfairness on both train and test.

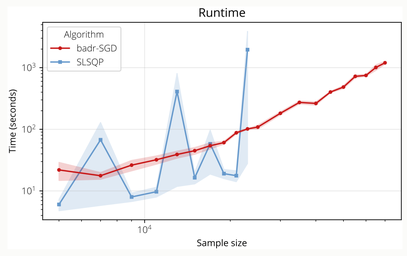
HEADLINE

**Lowest test unfairness on every metric, at no accuracy cost**

SOFTWARE

## `badr`: Pareto-fair models in one line

```
import badr as bdr
dset = bdr.datasets.fetch_adult()
model = bdr.models.LogisticRegression()
metric = bdr.metrics.IndividualFairness()
bdr.Badr(dset, model, metric).run()
```




→ badr-sgd scales where two-loop solvers (SLSQP) go out of memory at 25 k samples.

**PAPER**
arxiv.org/abs/2601.13448

**CODE**
github.com/…/badr