Overcoming the Resolution Limit: Significance-Aware Regularization for Intersectional Fairness

Antonio Ferrara **André Panisson** **Francesco Cozzi**
Alan Perotti **Francesco Bonchi**
Intesa Sanpaolo AI Research, Turin, Italy
{name.surname}@intesasanpaolo.com

Abstract

Intersectional fairness requires evaluating models across fine-grained subgroups. However, as group definitions become more granular, the sample size per subgroup shrinks, causing fairness estimates to become increasingly unstable. Existing in-processing methods are *certainty-blind*: they penalize observed disparities without distinguishing statistically certified unfairness from sampling noise. This leads to *fairness overfitting*, where models sacrifice utility to "fix" spurious disparities in sparse regimes. To address this, we introduce the *significant-excess* metric family, a continuous, uncertainty-aware fairness score that measures only the portion of a subgroup's disparity exceeding what sampling variation can explain at a specific confidence level. Building on this metric, we propose SAFER (Significance-Aware Fairness Regularizer), a differentiable in-processing method that gates fairness penalties using a soft-count likelihood-ratio evidence statistic. SAFER satisfies a *provable safety property*: when every subgroup disparity lies well below the gate threshold, the unconstrained optimum is nearly stationary for SAFER. At that optimum, the back-propagated contribution of its significance-gated component is exponentially small, and the additional fairness contribution comes from a small ungated term with a fixed weight. Mechanism ablations across three benchmark datasets and purpose-built stress variants show that removing any component of SAFER (threshold smoothing, the significance gate, or a dense-tail term) produces the failure the theory predicts on at least one stress variant. In sparse intersectional settings, SAFER achieves a better fairness–utility Pareto frontier than the baseline methods, and it preserves utility when certified unfairness is absent.

1 Introduction

Intersectional fairness [8, 16] requires evaluating models across conjunctions of protected attributes, but as subgroups become more granular, their sample sizes inevitably shrink. In these sparse regimes, fairness estimates become noise-dominated, and intersectional fairness reaches a *resolution limit*: the point at which the statistical uncertainty exceeds the disparity being measured [38, 14, 13]. Indeed, even widely adopted heuristics such as the four-fifths rule lack statistical grounding and can both miss and fabricate disparities [35]. Ignoring estimation uncertainty leads to *fairness overfitting*, where the model memorizes group noise in the training set rather than genuine bias [28]. Existing in-processing methods are *certainty-blind*: they penalize observed disparities without distinguishing certified unfairness from sampling noise, causing *fairness overfitting* in the sparse intersectional regime. Conversely, audit tools such as [13] incorporate uncertainty awareness, but only as a post-hoc binary accept/reject decision rather than a continuous metric or a differentiable training signal. If fairness regularization is to behave sensibly in sparse intersectional regimes, the optimization criterion itself must be able to distinguish well-supported disparities from noisy ones.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Whether gating on significance helps depends on how many training subgroups lie above the gate threshold. If none do, the gated fairness contributions are suppressed. If some do, the gate selects which disparities receive greater weight. If all do, the gate is effectively inactive.

In this context, this paper makes three main contributions:

- **Uncertainty-aware fairness measurement (Section 3).** We introduce the *significant-excess* metric family, which measures only the portion of each subgroup's disparity that exceeds what sampling variation can explain at a given confidence level. The construction applies to any parity criterion expressible as a continuously differentiable function of multinomial cell probabilities, including Statistical Parity [11], Disparate Impact [12], Equalized Odds and Equal Opportunity [17], and yields a continuous, uncertainty-aware fairness score that is zero when no violation is statistically certified.
- **Significance-gated optimization (Section 4).** We propose **SAFER** (Significance-Aware Fairness Regularizer), a differentiable in-processing method that gates fairness penalties using a soft-count likelihood-ratio evidence statistic. The gate attenuates subgroup penalties according to their evidence of disparity; statistical certification comes from the held-out audit. SAFER satisfies a provable *safety property*: when every disparity lies well below the gate threshold, the optimum of unconstrained empirical risk minimization (ERM) is nearly stationary for SAFER. At that optimum, the back-propagated contribution of its significance-gated component is exponentially small, and the additional fairness contribution comes from a small ungated term with a fixed weight.
- **Regime characterization and empirical validation (Section 5 and Appendix F.3).** We identify three operating regimes (safety, selective gating, gate always open), defined by which training subgroups cross the gate's threshold, and test each through mechanism ablations on purpose-built stress variants and a controlled benchmark.

The claims are regime-conditional: SAFER gives a better Pareto frontier than the baselines where certified disparities exist and preserves the utility of the unconstrained model where they do not, and removing any component of SAFER produces the failure the theory predicts on at least one stress variant.

2 Related Work

Intersectional and subgroup fairness. Equality of Opportunity [17] and related parity criteria are the standard building blocks for group fairness, but models can satisfy coarse constraints yet violate fairness on structured intersections [25]. Differential fairness treats intersectionality as a first-class object [23]; worst-case comparisons formalize the combinatorial structure of multi-attribute subgroups [15]; and multicalibration [19] provides simultaneous calibration guarantees over rich subgroup families. Molina and Loiseau [32] bound intersectional violations via marginal constraints; Himmelreich et al. [20] formalize the tension between group granularity and statistical power. Our work targets the same setting but a different bottleneck: once many intersectional groups are audited simultaneously, their disparities become heterogeneous in statistical reliability, and this reliability must enter both evaluation and optimization.

Uncertainty-aware fairness. SAFT [13] is the work closest to ours: it tests whether each subgroup's confidence interval excludes parity, applying multiplicity corrections across the subgroup family. This makes subgroup size and estimator variance part of the fairness decision; certainty-blind metrics [22] ignore both. We build on SAFT in two directions. First, we replace binary per-subgroup flags with a continuous *significant-excess* score that preserves severity information beyond the accept/reject decision, yielding a single metric for ranking models (Section 3). Second, this statistical logic motivates a differentiable evidence gate that modulates subgroup penalties during training. SAFER keeps SAFT's intervals for the audit and trains with a soft-count likelihood-ratio statistic, because a plug-in Wald standard error is unstable on small soft counts and a posterior-sampling test is too expensive inside a training loop (Section 4.2). More broadly, a growing body of work grounds fairness evaluation in statistical inference: Cherian and Candes [7] provide finite-sample valid tests for auditing; Tang et al. [37] replace Wald statistics with empirical-likelihood tests; and Lo et al. [30] advocate classical hypothesis testing for AI fairness. On the evaluation side, Barrainkua et al. [3] show that fairness conclusions are unstable under resampling, and Rosenblatt and Witter [36] benchmark how uncertainty can reverse fairness rankings. Kuzucu et al. [27] treat predictive uncertainty itself as

a fairness criterion, whereas we use statistical uncertainty of the fairness *estimate* to gate optimization. Besse et al. [5] were among the first to propose confidence intervals for testing disparate impact, which replaced point estimates with interval-based decisions.

Fairness-aware optimization. A large literature mitigates unfairness during or after training. Post-processing methods adjust a fixed classifier's thresholds [17], but cannot reshape the learned representation and become impractical when the number of intersectional subgroups is combinatorially large. Among in-processing approaches, Exponentiated Gradient [1] solves a sequence of weighted learning problems, Adversarial Debiasing [40] and Prejudice Remover [24] embed fairness pressure through an adversary or a dependence penalty, and differentiable regularizers such as Bechavod and Ligett [4] and FAIRRET [6] penalize group discrepancy via smooth surrogates. Recent intersectional-specific optimizers include Lee et al. [29], who regularize over multitype attributes, and Němeček et al. [33], who enforce exact intersectional constraints via mixed-integer programming. Both remain certainty-blind. Distributionally robust optimization (DRO) [18] and minimax group fairness [9] minimize worst-case subgroup loss, while FairDRO [34] casts group fairness as classwise robust optimization. All these methods are *certainty-blind*: once a subgroup disparity is observed, it contributes to the objective regardless of whether it is statistically distinguishable from sampling noise. SAFER differs by using a likelihood-ratio evidence gate to modulate disparity penalties during training, while retaining the held-out audit for statistical certification.

3 Measuring Intersectional Fairness

We introduce the *significant-excess metric* family to address the *resolution limit* identified in Section 1. To this end, these metrics are designed to answer two questions simultaneously: does a subgroup exhibit a non-zero disparity compared to the rest of the population, and is that disparity statistically significant enough to be distinguished from sampling noise?

Setup. Let $\mathfrak{S}$ be a family of protected subgroups, including intersectional groups formed by conjunctions of sensitive attributes. For each subgroup $S \in \mathfrak{S}$, we compare the behavior of a model on S and its complement $\bar{S}$. Both are sets of individuals: S collects those with the subgroup's attribute values and $\bar{S}$ the rest of the population. We write $\phi(S)$ for a generic fairness functional expressible as a continuously differentiable function of multinomial cell probabilities. We write $\phi_n(S)$ for its empirical estimate on a test sample of size n. For each subgroup disparity $\phi_n(S)$ we require a confidence interval of level $(1-\alpha)$, expressible by a lower and upper bound $[\ell_\alpha(S),\, u_\alpha(S)]$, which quantifies the uncertainty of the estimation. We assume $\ell_\alpha(S) \le \phi_n(S) \le u_\alpha(S)$.

Definition 1 (Significant excess)**.** For a subgroup S and significance level α, we define the *significant excess* of fairness criterion ϕ as

$$\mathcal{E}_\alpha^\phi(S) \;=\; \begin{cases} \max\big(0,\; \ell_\alpha(S)\big) & \text{if } \phi_n(S) \ge 0, \\ \max\big(0,\; -u_\alpha(S)\big) & \text{if } \phi_n(S) < 0. \end{cases} \tag{1}$$

The significant excess is the distance from zero to the interval. When the interval contains zero, the disparity is not statistically certified and the contribution is exactly zero. When the interval excludes zero, $\mathcal{E}_\alpha^\phi(S)$ measures the portion of the observed disparity that remains after subtracting what sampling noise can plausibly explain.

Basic properties. The significant-excess score satisfies:

1. **Zero iff no certified unfairness under the interval rule.** $\mathcal{E}_\alpha^\phi(S) = 0$ if and only if the interval for $\phi(S)$ contains zero.
2. **Scale preservation.** $\mathcal{E}_\alpha^\phi(S)$ is expressed in the same units as ϕ.
3. **Uncertainty awareness.** The score is nonincreasing when the interval expands by inclusion.
4. **Continuity.** The score is continuous in the interval endpoints.

Finally, we refer to the *Mean Significant Measure Excess* as the mean aggregation over the nonempty audited family of subgroups with defined estimates and intervals:

$$\overline{\mathcal{E}}_\alpha^\phi(f) = \frac{1}{|\mathfrak{S}|} \sum_{S \in \mathfrak{S}} \mathcal{E}_\alpha^\phi(S). \tag{2}$$

Let us consider two fairness metrics as an example, the Statistical Parity (SP) gap and the Equal Opportunity (EO) gap:

$$SP(S) = P(\hat{Y} = 1 \mid S) - P(\hat{Y} = 1 \mid \bar{S}), \tag{3}$$

$$EO(S) = P(\hat{Y} = 1 \mid Y = 1, S) - P(\hat{Y} = 1 \mid Y = 1, \bar{S}). \tag{4}$$

For SP, the relevant cells are predicted positive/negative outcomes conditioned on S and $\bar{S}$. For EO, the same construction is applied after conditioning on the positive-label population $Y = 1$. Both can be written as continuously differentiable functions of multinomial cell probabilities. Given a confidence level α, we can define the *Mean Significant Parity Excess* (MSPE) by applying Equations 1 and 2 with $\phi(S) = SP(S)$, and the *Mean Significant Equal-Opportunity Excess* (MSEE) with $\phi(S) = EO(S)$.

$$MSPE_\alpha(f) \;=\; \overline{\mathcal{E}}_\alpha^{SP}(f), \qquad MSEE_\alpha(f) \;=\; \overline{\mathcal{E}}_\alpha^{EO}(f). \tag{5}$$

We obtain the confidence intervals $[\ell_\alpha(S),\, u_\alpha(S)]$ with the SAFT framework [13], because its test adapts to both large and small subgroups.

4 SAFER: A Differentiable Surrogate for Significant-Excess Metrics

$MSEE_\alpha$ (Section 3) measures only the statistically certified portion of each subgroup's disparity. It is not directly trainable, because it depends on hard predictions, which are piecewise constant in θ, and on an absolute value and a positive part, which are not differentiable at zero. SAFER replaces these with differentiable approximations and is built around two ideas. First, a *soft-count likelihood-ratio test* (LRT) provides a differentiable significance statistic that requires no explicit variance estimation. This matters in the sparse intersectional cells, where Wald-type statistics become unreliable (§4.2). Second, a significance gate built from this statistic yields a provable *safety property*: when every subgroup disparity lies well below the gate threshold, the ERM optimum is nearly stationary for SAFER. At that optimum, the back-propagated contribution of the significance-gated component is exponentially small, and the additional fairness contribution is proportional to the weight of a small ungated term (§4.3). Proofs are in Appendix J.

4.1 Threshold-Smoothed Soft Counts

Let $f_\theta : \mathcal{X} \to [0, 1]$ be the model score. For each subgroup $S \in \mathfrak{S}$ write Π_S for the conditioning population: $\Pi_S = S$ for Statistical Parity, and $\Pi_S = (S,\, Y{=}1)$ for Equal Opportunity, matching Section 3. Define three population positive rates, relative to the hard binary predictions, the raw predicted probabilities and the τ smoothed sigmoid:

$$p_S^{\mathrm{hard}}(\theta) = \mathbb{E}\big[\mathbf{1}\{f_\theta(X) > \tfrac{1}{2}\} \;\big|\; \Pi_S\big]\,, \tag{6}$$

$$p_S^{\mathrm{raw}}(\theta) = \mathbb{E}[f_\theta(X) \mid \Pi_S]\,, \tag{7}$$

$$p_S^{\tau}(\theta) = \mathbb{E}\big[\sigma_\tau\big(f_\theta(X) - \tfrac{1}{2}\big) \;\big|\; \Pi_S\big]\,, \qquad \sigma_\tau(u) = \sigma(u/\tau), \tag{8}$$

where σ is the logistic sigmoid. The corresponding disparity is $\phi_S^\bullet = p_S^\bullet - p_{\bar{S}}^\bullet$ for each variant $\bullet \in \{\mathrm{hard}, \mathrm{raw}, \tau\}$. For example in this notation, ϕ_S^{hard} is exactly the population functional $\phi(S)$ of Section 3. Furthermore, we use hats (e.g., $\widehat{\phi}_S^\tau$) to denote the empirical analogues computed from sample soft counts. The decision threshold $1/2$ is the Bayes rule when the scores estimate class probabilities and both error types cost the same. It is a property of the deployed classifier and is fixed before training. The construction applies unchanged to any other fixed threshold t by replacing $f_\theta(x) - \frac{1}{2}$ with $f_\theta(x) - t$; Appendix F.1 reports SAFER retrained and audited at thresholds from 0.1 to 0.9.

In particular, replacing $\mathbf{1}\{f_\theta(x) > \frac{1}{2}\}$ with $\sigma_\tau(f_\theta(x) - \frac{1}{2})$ yields a differentiable approximation whose quality is controlled linearly by τ:

Theorem 2. *If, for some $\delta > 0$, the conditional score distributions on Π_S and $\Pi_{\bar{S}}$ admit densities bounded by L on $\{u : |u - \frac{1}{2}| \le \delta\}$, then $|\phi_S^\tau(\theta) - \phi_S^{\mathrm{hard}}(\theta)| \le 4L\tau \log 2 + 2e^{-\delta/\tau}$.*

A natural alternative, using the raw probability $f_\theta(x)$ directly (p_S^{raw}), does not work:

Proposition 3 (Raw soft counts are not hard-decision consistent)**.** *For every $\eta > 0$, there exist continuous bounded-density score distributions such that $|\phi_S^{\mathrm{raw}}(\theta) - \phi_S^{\mathrm{hard}}(\theta)| \geq 1 - \eta$, and no constant rescaling of ϕ_S^{raw} matches ϕ_S^{hard} uniformly.*

Two groups sitting just across the $\frac{1}{2}$ boundary from opposite sides produce a hard gap near 1 but a raw-probability gap near 0; two groups both well above the boundary produce the reverse. No single correction can fix both regimes, so raw soft counts can rank models differently from the hard-decision audit. Threshold smoothing also localizes gradients to examples near the decision boundary, where predictions can flip (Theorem 7 in Appendix J).

4.2 A Differentiable LRT for Subgroup Disparities

Given threshold-smoothed soft counts, we need a differentiable evidence statistic to select subgroup disparities for training. An approach would be the Bernoulli-form Wald statistic $|Z_S| = |\widehat{\phi}_S^{\tau}|/\widehat{SE}_S^{\phi}$, where $\widehat{SE}_S^{\phi}$ is the two-proportion standard-error proxy of property (ii), but this has two problems in the sparse intersectional regime. First, $\widehat{SE}_S^{\phi}$ is numerically unstable when soft counts are small. The second problem matters more for optimization: the Wald loss can decrease by *inflating the standard-error estimate* rather than reducing the disparity, since both numerator and denominator depend on θ.

SAFER instead uses the Bernoulli-form two-sample deviance statistic computed on threshold-smoothed soft counts. For subgroup S define the soft counts $\tilde{n}_{1,S}(\theta) = \sum_{i \in \Pi_S} \sigma_\tau\big(f_\theta(x_i) - \frac{1}{2}\big)$ and $\tilde{n}_{0,S}(\theta) = |\Pi_S| - \tilde{n}_{1,S}(\theta)$. Let $\tilde{n}_S = \tilde{n}_{1,S} + \tilde{n}_{0,S}$ and define analogous quantities for $\bar{S}$. Write $\hat{p}_S = \tilde{n}_{1,S}/\tilde{n}_S$, $\hat{p}_{\bar{S}} = \tilde{n}_{1,\bar{S}}/\tilde{n}_{\bar{S}}$, and $\hat{p} = (\tilde{n}_{1,S} + \tilde{n}_{1,\bar{S}})/(\tilde{n}_S + \tilde{n}_{\bar{S}})$ for the pooled null estimate. The likelihood-ratio statistic for the soft-parity null $H_0 \colon p_S^{\tau} = p_{\bar{S}}^{\tau}$ is

$$\Lambda_S(\theta) = -2\left[\tilde{n}_{1,S} \log \frac{\hat{p}}{\hat{p}_S} + \tilde{n}_{0,S} \log \frac{1-\hat{p}}{1-\hat{p}_S} + \tilde{n}_{1,\bar{S}} \log \frac{\hat{p}}{\hat{p}_{\bar{S}}} + \tilde{n}_{0,\bar{S}} \log \frac{1-\hat{p}}{1-\hat{p}_{\bar{S}}}\right]. \tag{9}$$

For SP, these are soft counts of predicted positive/negative across S and $\bar{S}$; for EO, the same construction restricts to the positive-label population.

Three properties make the soft-count LRT a suitable significance statistic for intersectional fairness training:

- **(i) No variance estimation.** Λ_S is a closed-form function of the soft counts: it never estimates or differentiates through $\widehat{SE}_S^{\phi}$. This removes the gradient path through which the Wald loss can be lowered by inflating the standard error.
- **(ii) Boundary stability.** In sparse cells the Bernoulli-form variance proxy $(\widehat{SE}_S^{\phi})^2 = \hat{p}_S(1 - \hat{p}_S)/\tilde{n}_S + \hat{p}_{\bar{S}}(1 - \hat{p}_{\bar{S}})/\tilde{n}_{\bar{S}}$ is unstable: with only a few effective observations, threshold-smoothed rates can approach zero or one, the variance estimate collapses toward zero, and $|Z_S|$ grows without bound, so a few tiny, saturated subgroups can dominate the objective. Λ_S is computed from the Bernoulli log-likelihood itself and contains no division by a plug-in variance estimate that can vanish, so it avoids this boundary singularity.
- **(iii) Asymptotic equivalence.** For interior rates and vanishing disparities, a second-order expansion gives $\Lambda_S = Z_S^2\,(1 + o_p(1))$, so the two statistics use the same reference cutoff z_α. For a fixed predictor evaluated on independent observations under soft parity, $\Lambda_S \Rightarrow c\chi_1^2$, $0 \leq c \leq 1$, and the Bernoulli reference cutoff is asymptotically conservative (Appendix J.6).

During training the gate is an optimization device; statistical certification is provided by the held-out hard-count audit of Section 3.

LRT-based threshold scale. To express the significance threshold in the raw-disparity units of $MSEE_\alpha$ when the disparity is nonzero, we define an effective standard-error proxy from the empirical disparity and the deviance:

$$\widehat{SE}_{\mathrm{eff},S}(\theta) = \frac{|\widehat{\phi}_S^{\tau}(\theta)|}{\sqrt{\Lambda_S(\theta)}}\,. \tag{10}$$

This conversion puts the dimensionless evidence statistic on the disparity scale, so that the training threshold carries the units of ϕ. Under the local conditions of property (iii), $\widehat{SE}_{\mathrm{eff},S} = \widehat{SE}^{\phi}_{S}(1 + o_p(1))$, recovering the Bernoulli-form normalization. Before numerical clamping, the threshold therefore scales with $\sqrt{1/n_S + 1/n_{\bar{S}}}$ for SP and $\sqrt{1/m_S + 1/m_{\bar{S}}}$ for EO. It thus shrinks as the inverse square root of the sample size when both comparison groups grow in proportion. The implementation regularizes the ratio and clamps it for numerical stability.

The trained objective. SAFER penalizes excess disparity above the LRT-based training threshold, with the threshold detached from the gradient computation:

$$\mathcal{L}_{\mathrm{dir}}(\theta) = \frac{1}{|\mathfrak{S}|} \sum_{S \in \mathfrak{S}} \mathrm{softplus}_{\tau_s}\left(\sqrt{\widehat{\phi}^{\tau}_{S}(\theta)^2 + \varepsilon^2} \;-\; z_\alpha \;\mathrm{sg}\Big[\widehat{SE}_{\mathrm{eff},S}(\theta)\Big] \right), \tag{11}$$

where $\mathrm{softplus}_\tau(u) = \tau \log(1 + e^{u/\tau})$ and $\mathrm{sg}[\cdot]$ denotes stop-gradient (the enclosed expression is treated as a constant during back-propagation). Detachment ensures that the optimizer can only reduce the loss by shrinking $\widehat{\phi}^{\tau}_{S}$, not by inflating the threshold. The threshold still updates between optimizer steps as θ changes, so it remains adaptive: it does not provide a gradient path. Within one update the gated term therefore gives the optimizer no first-order incentive to raise the threshold; across updates the threshold follows the predictions, which every term of the objective changes. Appendix J.9 gives a scalar example and a diagnostic over training runs, in which the mean threshold shows no upward drift. A canonical version of Eq. (11), in which the surrogate and the metric share the population Wald threshold $B_S = z_\alpha SE^{\phi}_{S}$, approximates $MSEE^{\mathrm{pop}}_{\alpha}$ (Eq. (A.8)) pointwise with error at most $\varepsilon + \tau_s \log 2 + 4L\tau \log 2 + 2e^{-\delta/\tau}$ (Theorem 8 in Appendix J).

4.3 Safety Property and Operating Regimes

The softplus gate in $\mathcal{L}_{\mathrm{dir}}$ separates subgroups above the gate threshold from those below it. Let

$$u_S(\theta) = \sqrt{\widehat{\phi}^{\tau}_{S}(\theta)^2 + \varepsilon^2} \;-\; z_\alpha \;\mathrm{sg}\Big[\widehat{SE}_{\mathrm{eff},S}(\theta)\Big] .$$

Since the stop-gradient (sg) freezes the threshold for differentiation, $\nabla_\theta u_S = \nabla_\theta \sqrt{\widehat{\phi}^{\tau}_{S}(\theta)^2 + \varepsilon^2}$. Throughout, ∇_θ of an expression containing sg denotes the back-propagated fairness contribution with the detached quantity held fixed; Proposition 4 bounds it. The chain rule and $\mathrm{softplus}'_{\tau_s}(u) = \sigma(u/\tau_s)$ give

$$\big\|\nabla_\theta \,\mathrm{softplus}_{\tau_s}(u_S)\big\| \le \sigma(u_S/\tau_s)\, \big\|\nabla_\theta u_S\big\|. \tag{12}$$

When $u_S < 0$ (disparity below the gate threshold), the prefactor $\sigma(u_S/\tau_s)$ decays exponentially as e^{u_S/τ_s}, suppressing the fairness gradient on that subgroup.

Dense tail. Under strong regularization, active disparities may stay close to the hinge boundary where the softplus gate attenuates the gradient substantially: by Eq. (12) the gradient is multiplied by $\sigma(u_S/\tau_s)$, which equals $1/2$ at the gate boundary and is already below 0.01 at $u_S = -0.05$ for $\tau_s = 10^{-2}$, so under strong regularization the optimization can stall before an active disparity is removed. To prevent stalling, SAFER adds a small ungated term:

$$\mathcal{L}_{\mathrm{tail}}(\theta) = \frac{w_{\mathrm{tail}}}{|\mathfrak{S}|} \sum_{S \in \mathfrak{S}} \sqrt{\Lambda_S(\theta)}\,, \tag{13}$$

where $w_{\mathrm{tail}} \ge 0$. This term carries no gate multiplier, so it provides a gradient even when the softplus is saturated. Since $\sqrt{\Lambda_S}$ locally expresses the disparity in units of the Bernoulli-form normalization, the tail weights each subgroup by its statistical support (Appendix J.7). Because the term is ungated, the safety bound below includes an additional term proportional to w_{tail}, while the gated part of the gradient keeps its exponential bound; with $w_{\mathrm{tail}} = 0$ the objective is gated only. We use the fixed value $w_{\mathrm{tail}} = 0.003$ in all runs of the full method. When active subgroups dominate, the tail is negligible relative to $\mathcal{L}_{\mathrm{dir}}$.

Proposition 4 (Safety at the unconstrained optimum)**.** *If $u_S(\theta^\star_{\mathrm{ERM}}) \le -\epsilon^\star$ for every $S \in \mathfrak{S}$, and the score and statistic gradients are bounded near $\theta^\star_{\mathrm{ERM}}$, then*

$$\big\|\nabla_\theta \mathcal{L}_{\mathrm{dir}}(\theta^\star_{\mathrm{ERM}})\big\| = O\Big(e^{-\epsilon^\star/\tau_s}\Big)\,, \tag{14}$$

$$\big\|\nabla_\theta (\mathcal{L}_{\mathrm{dir}} + \mathcal{L}_{\mathrm{tail}})(\theta^\star_{\mathrm{ERM}})\big\| \le O\Big(e^{-\epsilon^\star/\tau_s}\Big) + \frac{w_{\mathrm{tail}}}{|\mathfrak{S}|} \sum_{S \in \mathfrak{S}} \big\|\nabla_\theta \sqrt{\Lambda_S(\theta^\star_{\mathrm{ERM}})}\big\|. \tag{15}$$

When every subgroup lies below the gate boundary by a margin $\epsilon^\star$ large relative to τ_s, the ERM optimum is nearly stationary for SAFER: the gated fairness contribution there is exponentially small, and the dense-tail residual (Eq. (13)) enters the full objective multiplied by $\lambda w_{\mathrm{tail}}$.

Three operating regimes. The gate threshold determines which subgroups contribute to the fairness gradient. Write $\mathcal{A}(\theta) = \{S \in \mathfrak{S} : u_S(\theta) > 0\}$ for the *active set* of training subgroups: a subgroup is active when $u_S > 0$, which, up to the smoothing constant ε and the numerical clamps, means $T_S = \sqrt{\Lambda_S} > z_\alpha$, with $z_\alpha = 1.96$ at $\alpha = 0.05$. At a given checkpoint θ, the active set places a dataset in one of three regimes:

- **(i) Safety** ($\mathcal{A}(\theta) = \varnothing$): when all subgroups at the ERM point lie well below the boundary relative to τ_s, the gated fairness contribution is exponentially small, and the unconstrained optimum is nearly stationary for SAFER up to the dense-tail residual.
- **(ii) Selective gating** ($0 < |\mathcal{A}(\theta)| < |\mathfrak{S}|$, active and inactive subgroups coexist): the gate concentrates model capacity on the disparities for which the data provide statistical support.
- **(iii) Gate always open** ($\mathcal{A}(\theta) = \mathfrak{S}$, all subgroups above the boundary): when all lie well above it relative to τ_s, the gate is effectively inactive, and SAFER approaches a certainty-blind penalty.

The regime is read off the active set at a checkpoint; when we characterize a dataset we use the checkpoint of the unconstrained model, and the regime can change during training. Whether a subgroup is active depends on its disparity, on its outcome rates, and on the counts on both sides of the comparison (Appendix J.8), so the boundary between regimes is set by the evidence in each subgroup. Appendix F.3 states the hypothesis these regimes suggest and tests it on a controlled benchmark.

Full objective. For fairness criterion ϕ, SAFER minimizes

$$\mathcal{L}_{\mathrm{SAFER}}(\theta) = \mathcal{L}_{\mathrm{pred}}(\theta) \; + \; \lambda\big[\mathcal{L}_{\mathrm{dir}}(\theta) + \mathcal{L}_{\mathrm{tail}}(\theta)\big], \tag{16}$$

where $\mathcal{L}_{\mathrm{pred}}$ is binary cross-entropy and $\lambda \geq 0$ controls the fairness–utility tradeoff.

Extensions. The construction extends to any fairness functional differentiable in multinomial cell probabilities: substituting SP gives MSPE directly, Equalized Odds adds an FPR-parity hinge, and Disparate Impact uses the log-DI metric with the same SP parity null (Appendix I).

5 Experiments

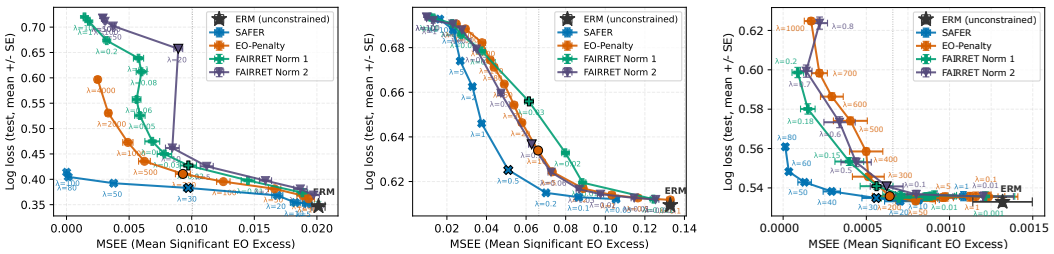
The experiments serve two purposes. First, we compare SAFER on the Pareto frontier against the strongest available in-processing fairness baselines on three datasets spanning sample sizes and disparity regimes. Second, we use constructed *stress variants* of the datasets (low training budgets, deeper intersection levels and small validation budgets) to probe the three mechanisms predicted by Section 4. Each variant targets a setting in which one SAFER component should be needed, so removing that component in an ablation should have a measurable effect. The full codebase is available at `https://github.com/Ambress92/Safer`.

Datasets and Stress Variants. Public benchmarks with multiple protected attributes and non-trivial intersectional structures are scarce. The main figures use the three most established ones (Adult Income [26], COMPAS [2] and German Credit [21]). Table 1 summarizes their stress variants, each of which targets one of the mechanisms of Section 4. The table also reports ACS Coverage [10] and Taiwan Credit [39], which add a public-coverage task with larger samples and a credit-risk task with many intersections. Unless a train-test split or intersection level is stated explicitly, dataset names below refer to the default variants in Table 1. The default COMPAS variant uses the more severe 20/80 train-test split and is reserved for mechanism ablations; the headline COMPAS comparison instead uses a balanced 50/50 split for a reasonable training budget.

Adult probes the threshold-smoothing mechanism (almost every intersection has few samples). COMPAS probes the dense-tail mechanism under strong regularization (small training set, moderate disparity). German probes the safety regime (almost no SP/EO violation is statistically certifiable at $\alpha = 0.05$ under ERM in our observed splits). ACS Coverage has larger effective subgroup sizes, while Taiwan Credit tests whether the comparison extends to a larger credit-risk task with many audited intersections but a relatively small positive-label budget. Appendix A details preprocessing, protected attributes, subgroup construction, and split-level audit statistics.

Table 1: Datasets, stress variants, and the mechanism each variant probes. $|\mathfrak{S}|$ is the number of audited intersectional subgroups at the indicated level; $\bar{n}_S$ is the mean training-set subgroup size; $\bar{m}_S$ is the mean positive-label subgroup size relevant to EO. All values are approximate averages over splits.

Dataset	Level	Split	n_{train}	$\lvert\mathfrak{S}\rvert$	$\bar{n}_S$ / $\bar{m}_S$	Targeted mechanism
Adult Income	4	10/90	4,522	228	19.9 / 4.9	threshold smoothing: high intersectionality, low budget
COMPAS	3	20/80	993	35	28.4 / 14.2	dense tail: small training set, strong regularization
German Credit	3	50/50	500	32	15.6 / 10.9	significance gate: safety regime
ACS Coverage	3	50/50	25,000	71	352.1 / 103.5	higher-sample ACS benchmark
Taiwan Credit	4	50/50	15,000	285	52.6 / 11.6	high-intersectionality credit risk



(a) Adult Income (b) COMPAS (c) German Credit

Figure 1: Held-out log-loss versus MSEE Pareto frontiers under Equal Opportunity. The vertical axis is the mean test log-loss over runs across splits and seeds and the horizontal axis is MSEE on hard test predictions; lower is better on both. Markers are the values of the fairness weight λ along each method's sweep, annotated and connected in increasing order, with error bars of ± 1 standard error in both coordinates; the star is ERM ($\lambda = 0$). The methods are SAFER, EO-Penalty, FAIRRET Norm 1 and FAIRRET Norm 2. Adult Income and German Credit use the default variants in Table 1; COMPAS uses the balanced 50/50 variant. The dashed vertical line marks $0.5\times$ the ERM MSEE, which is the eligibility threshold used in Table 2; outlined markers indicate the selected qualifying points reported in that table.

Protocol and Metrics Each method and λ is evaluated over at least 10 stratified shuffle splits, with at least 3 random initialization seeds per split (Appendix B gives the counts per experiment). All headline neural methods share the same MLP backbone, prediction-loss schedule, and held-out validation-only model selection. We report $MSEE_\alpha$ at $\alpha = 0.05$ on hard test predictions (Eq. (5)) against held-out log-loss, with standard error (SE) bars over the runs across splits and seeds. MSEE and MSPE use unadjusted per-subgroup intervals; reported violation counts use Benjamini–Hochberg (BH)-corrected subgroup p-values. Hyperparameter grids, BCE-guarded checkpointing, and clustered-SE checks are in Appendix B.

5.1 Pareto Comparison Against Baselines

We compare SAFER against ERM, EO-Penalty and two FAIRRET [6] regularizers. EO-Penalty is the certainty-blind squared-disparity regularizer ($\mathcal{L}_{\text{pred}} + \frac{\lambda}{|\mathfrak{S}|}\sum_S(\widehat{\phi}_S^{\text{raw}})^2$) and is the closest direct comparator to the SAFER significant-excess hinge. FAIRRET Norm 1 and FAIRRET Norm 2 are ratio-normalized differentiable fairness regularizers. EO-Penalty and both FAIRRET variants operate on raw soft probabilities rather than threshold-smoothed predictions, which makes them direct empirical tests of the inconsistency claim of Proposition 3. We also report thresholded summaries for Exponentiated Gradient [1], Prejudice Remover [24], and Adversarial Debiasing [40]. Complete per-dataset details are deferred to Appendix D, because these methods do not all share the matched neural in-training backbone used in the headline Pareto plots.

Table 2 summarizes the thresholded fairness–utility comparison across all five datasets and all baselines. For each dataset and method, the entry is the best mean test log loss among the swept hyperparameters whose test MSEE is at least 50% below the ERM MSEE on the same dataset. A $*$ indicates that no swept hyperparameter achieved that fairness threshold, so the method did not produce a qualifying tradeoff under this criterion. Appendix D reports the complete per-dataset tables, including the lowest-MSEE point found for each ineligible $*$ case.

Table 2: Thresholded fairness–utility summary across datasets and baselines. Each entry is the best mean test log loss among swept hyperparameters with MSEE at least 50% below ERM MSEE; $*$ means no qualifying hyperparameter was found.

	Adult Income	COMPAS	German Credit	Taiwan	ACS
SAFER	$\mathbf{0.3829} \pm 3e{-4}$	$\mathbf{0.6252} \pm 7e{-4}$	$\mathbf{0.5349} \pm 1e{-3}$	$\mathbf{0.4479} \pm 4e{-4}$	$\mathbf{0.5198} \pm 3e{-4}$
EO-Penalty	$0.4106 \pm 9e{-4}$	$0.6339 \pm 5e{-4}$	$0.5358 \pm 1e{-3}$	$0.4527 \pm 6e{-4}$	$0.5223 \pm 4e{-4}$
FAIRRET Norm 1	$0.4272 \pm 1e{-3}$	$0.6559 \pm 9e{-4}$	$0.5409 \pm 2e{-3}$	$0.4668 \pm 1e{-3}$	$0.5240 \pm 4e{-4}$
FAIRRET Norm 2	$0.4618 \pm 3e{-3}$	$0.6366 \pm 7e{-4}$	$0.5408 \pm 2e{-3}$	$0.4624 \pm 1e{-3}$	$0.5314 \pm 4e{-4}$
AdvDebiasing	*	*	*	*	*
ExponGrad	*	*	*	$0.6256 \pm 3.5e{-2}$	$0.6535 \pm 1.3e{-2}$
PrejudiceRemover	*	*	*	$0.6216 \pm 1.1e{-2}$	*

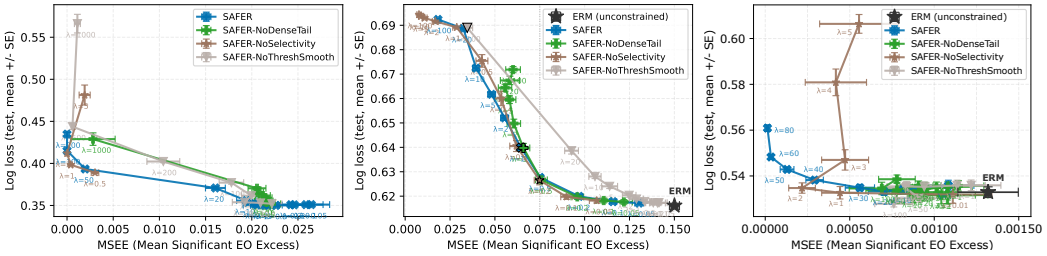


Figure 2: Single-knob ablations of SAFER on the default Adult Income, COMPAS, and German Credit variants from Table 1. The panels show Adult Income (left), COMPAS (center) and German Credit (right). Axes, markers and error bars are as in Figure 1. The arms are SAFER and its three single-knob removals: NoDenseTail ($w_{\text{tail}} = 0$), NoSelectivity (gate threshold set to zero) and NoThreshSmooth (raw probabilities instead of threshold-smoothed counts).

Figure 1 shows the EO Pareto frontiers for Adult, COMPAS, and German. Across panels, SAFER preserves near-ERM log-loss while crossing the threshold and reaches lower MSEE than EO-Penalty or FAIRRET without comparable log-loss collapse. Both observations are consistent with the alignment argument of Theorem 2: optimizing a threshold-smoothed surrogate that approximates the hard-decision metric with an error linear in τ is more effective at reducing that metric than optimizing a raw-soft-count surrogate, especially in regions where the soft-count and hard-decision quantities can disagree by an order-one amount (Proposition 3). SAFER's frontier matches or improves on that of a baseline that optimizes the Wald statistic directly (Appendix F.2). The per-dataset threshold summaries, including the minimum-MSEE points for methods marked ineligible in Table 2, are reported in Appendix D.

The corresponding SP Pareto plots are in Appendix H. Under SP the effective per-subgroup sample sizes n_S are substantially larger than EO's m_S, so more subgroups are certified already at ERM and the gate contributes less. SAFER nevertheless reaches lower log-loss at matched MSPE than the raw-probability penalty on all three datasets, and the difference is largest on Adult. We attribute this to the threshold-smoothing effect of Theorem 2: with the gate close to open, the hinge approaches a smoothed-disparity penalty, and SAFER's remaining advantage comes from targeting hard-decision disparities.

5.2 Ablation study

Section 4 predicted three mechanisms: threshold smoothing (Theorems 2, 7), the gate threshold (Proposition 4), and the dense tail (Eq. (12)). Each picks out one stress variant in which it is predicted to be empirically necessary. We test all three with single-knob ablations against the SAFER backbone (NoThreshSmooth, NoSelectivity, NoDenseTail); details are in Appendix G.

Threshold smoothing. On Adult and COMPAS, NoThreshSmooth is Pareto-dominated across the mid-MSEE range; at matched $MSEE \approx 10^{-3}$ on Adult, removing threshold smoothing costs roughly 0.14 nats of log-loss. This is the empirical counterpart of Theorems 2 and 7: raw probabilities target a soft-count quantity that is biased relative to the hard-decision metric and whose gradient is not concentrated where decisions can flip.

Dense-tail term. On COMPAS, NoDenseTail plateaus at $MSEE \approx 0.06$ from $\lambda \geq 5$ while SAFER reaches $MSEE \approx 0.018$ at $\lambda = 100$ (about ten standard errors apart). On Adult, NoDenseTail similarly fails to reach the low-MSEE region of SAFER's frontier: it remains near ERM MSEE (≈ 0.02) for almost the entire λ sweep and moves toward smaller MSEE only at the largest tested λ, where it is Pareto-dominated by SAFER (Figure 2, left). The mechanism is Eq. (12): as the optimizer drives disparities toward the gate threshold, $\sigma(u_S/\tau_s)$ drops below $\frac{1}{2}$ and the hinge becomes a vanishing fraction of the gradient; the tail's gradient carries no $\sigma(u_S/\tau_s)$ factor and lets optimization continue. On German Credit, NoDenseTail tracks SAFER, because the significance gate already keeps the hinge inactive.

Gate threshold. On German Credit, almost no EO violation is statistically certified at $\alpha = 0.05$ in our sweeps; SAFER preserves ERM-level log-loss (≈ 0.53) up to $\lambda = 20$, while NoSelectivity climbs to log-loss ≈ 0.69 by $\lambda = 20$. This is the failure mode that Proposition 4 predicts when the optimizer can act on threshold-smoothed but uncertified signal.

Cross-Regime Summary. Each removal produces the failure that Section 4 predicts. Without threshold smoothing, a Pareto gap opens on Adult and COMPAS, as Theorems 2 and 7 predict. Without the dense tail, the frontier stops short of low MSEE on the same two datasets, because of the hinge suppression in Eq. (12). Without the gate threshold, the log-loss rises on German Credit, as Proposition 4 predicts. NoSelectivity has no effect on COMPAS, and on Adult it tracks SAFER up to $\lambda = 1$ and loses utility at larger λ (Appendix G.4). The ablation map is in Appendix G.5.

6 Conclusion

Intersectional fairness auditing encounters a resolution limit: as subgroup definitions grow finer, the statistical uncertainty of each disparity estimate comes to exceed the disparity itself. We showed that ignoring this limit, as existing in-processing methods do, leads to fairness overfitting, where the optimizer spends model capacity fitting sampling noise rather than genuine disparities.

To address this, we introduced the *significant-excess* metric family, which continuously measures only the portion of each subgroup's disparity that exceeds what sampling variation can explain at a given confidence level, and *SAFER*, a differentiable regularizer that gates fairness penalties on a soft-count likelihood-ratio evidence statistic. SAFER satisfies a provable safety property: when every subgroup disparity lies well below the gate threshold, the ERM optimum is nearly stationary for SAFER. At that optimum, the gated fairness contribution is exponentially small, and the contribution of the small ungated term is proportional to its weight.

Experiments across several benchmarks and purpose-built stress variants support a regime-conditional claim: SAFER achieves a better Pareto frontier than the baselines where certified disparities exist and preserves the utility of the unconstrained model where they do not, and removing any component of SAFER produces the failure the theory predicts on at least one stress variant.

Limitations and future work. A zero significant excess means that no excess disparity is certifiable at the chosen confidence level and sample size; a real disparity in a small subgroup can remain uncertified, and SAFER then withholds the gradient that would reduce it (Appendix E). The current evaluation is limited to tabular benchmarks with at most a few hundred intersectional subgroups. Scaling to high-cardinality attribute spaces and to settings where subgroup membership is uncertain or partially observed remains open. A natural extension is to include multiplicity corrections in the training-time significance test, so that SAFER's gate also controls the family-wise error rate.

Acknowledgments and Disclosure of Funding

This work is conducted as part of the Horizon Europe project PRE-ACT (Prediction of Radiotherapy side effects using explainable AI for patient communication and treatment modification). It is supported by the European Commission through the Horizon Europe Program (Grant Agreement number 101057746), by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22 00058, and by the UK government (Innovate UK) under application number 10061955. The authors are employed by Intesa Sanpaolo. The authors declare no other competing interests.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, pages 60–69, 2018.

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

[3] Ainhize Barrainkua, Paula Gordaliza, Jose A. Lozano, and Novi Quadrianto. Uncertainty matters: Stable conclusions under unstable assessment of fairness results. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1198–1206. PMLR, 2024.

[4] Yahav Bechavod and Katrina Ligett. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.

[5] Philippe Besse, Eustasio del Barrio, Paula Gordaliza, and Jean-Michel Loubes. Confidence intervals for testing disparate impact in fair learning. *arXiv preprint arXiv:1807.06362*, 2018.

[6] Maarten Buyl, Marybeth Defrance, and Tijl De Bie. fairret: a framework for differentiable fairness regularization terms. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

[7] John J. Cherian and Emmanuel J. Candès. Statistical inference for fairness auditing. *Journal of Machine Learning Research*, 25(149):1–49, 2024.

[8] Kimberle Crenshaw. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43(6):1241–1299, 1991.

[9] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 66–76, 2021.

[10] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery.

[12] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

[13] Antonio Ferrara, Francesco Cozzi, Alan Perotti, André Panisson, and Francesco Bonchi. Size-adaptive hypothesis testing for fairness. In *Advances in Neural Information Processing Systems (NeurIPS) 2025*, 2025.

[14] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness. In *2020 IEEE 36th international conference on data engineering (ICDE)*, pages 1918–1921. IEEE, 2020.

[15] Avijit Ghosh, Lea Genuit, and Mary Reagan. Characterizing intersectional group fairness with worst-case comparisons. In *Proceedings of the 2nd Workshop on Diversity in Artificial Intelligence (AIDBEI)*, volume 142 of *PMLR*, pages 22–34, 2021.

[16] Usman Gohar and Lu Cheng. A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. In Edith Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6619–6627. International Joint Conferences on Artificial Intelligence Organization, 2023. Survey Track.

[17] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, pages 3315–3323, 2016.

[18] Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, pages 1929–1938, 2018.

[19] Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (Computationally-Identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, pages 1939–1948, 2018.

[20] Johannes Himmelreich, Arbie Hsu, Ellen Veomett, and Kristian Lum. The intersectionality problem for algorithmic fairness. In Miriam Rateike, Awa Dieng, Jamelle Watson-Daniels, Ferdinando Fioretto, and Golnoosh Farnadi, editors, *Proceedings of the Algorithmic Fairness Through the Lens of Metrics and Evaluation*, volume 279 of *Proceedings of Machine Learning Research*, pages 68–95. PMLR, 2025.

[21] Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994.

[22] Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, 1(2):1–52, 2024.

[23] Rashidul Islam, Kamrun Naher Keya, Shimei Pan, Anand D. Sarwate, and James R. Foulds. Differential fairness: An intersectional framework for fair AI. *Entropy*, 25(4):660, 2023.

[24] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, volume 7524 of *Lecture Notes in Computer Science*, pages 35–50. Springer, 2012.

[25] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, pages 2564–2572, 2018.

[26] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 202–207, 1996.

[27] Selim Kuzucu, Jiaee Cheong, Hatice Gunes, and Sinan Kalkan. Uncertainty as a fairness measure. *J. Artif. Int. Res.*, 81, 2025.

[28] Firas Laakom, Haobo Chen, Jürgen Schmidhuber, and Yuheng Bu. Fairness overfitting in machine learning: An information-theoretic perspective. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 32078–32115. PMLR, 2025.

[29] Ho Ming Lee, Katrien Antonio, Benjamin Avanzi, Lorenzo Marchi, and Rui Zhou. Machine learning with multitype protected attributes: Intersectional fairness through regularisation. *arXiv preprint arXiv:2509.08163*, 2025.

[30] Victor Lo, Sayan Datta, and Youssouf Salami. Bringing practical statistical science to AI and predictive model fairness testing. *AI and Ethics*, 5(3):2149–2164, 2025.

[31] Yan Luo, Muhammad Osama Khan, Yu Tian, Min Shi, Zehao Dou, Tobias Elze, Yi Fang, and Mengyu Wang. FairVision: Equitable deep learning for eye disease screening via fair identity scaling. *arXiv preprint arXiv:2310.02492*, 2023.

[32] Mathieu Molina and Patrick Loiseau. Bounding and approximating intersectional fairness through marginal fairness. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[33] Jiří Němeček, Mark Kozdoba, Illia Kryvoviaz, Tomáš Pevný, and Jakub Mareček. Intersectional fairness via mixed-integer optimization. *arXiv preprint arXiv:2601.19595*, 2026.

[34] Taeeon Park, Sangwon Jung, Sanghyuk Chun, and Taesup Moon. FairDRO: Group fairness regularization via classwise robust optimization. *Neural Networks*, 182:106891, 2025.

[35] Manish Raghavan and Pauline T Kim. Limitations of the "four-fifths rule" and statistical parity tests for measuring fairness. *Geo. L. Tech. Rev.*, 8:93, 2024.

[36] Lucas Rosenblatt and R. Teal Witter. Fairlyuncertain: A comprehensive benchmark of uncertainty in algorithmic fairness. *arXiv preprint arXiv:2410.02005*, 2024.

[37] Jie Tang, Chuanlong Xie, Xianli Zeng, and Lixing Zhu. Empirical likelihood-based fairness auditing: Distribution-free certification and flagging. *arXiv preprint arXiv:2601.20269*, 2026.

[38] Ke Yang, Joshua R. Loftus, and Julia Stoyanovich. Causal intersectionality and fair ranking. In *Symposium on Foundations of Responsible Computing*, 2021.

[39] I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2):2473–2480, 2009.

[40] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 335–340, 2018.

A Datasets and stress variants

All experiments use the same preprocessing code as the released scripts. Protected attributes are removed from the feature matrix and kept only for subgroup construction. Subgroup masks include all conjunctions up to the specified intersection level, with groups of size less than two omitted in the corresponding split.

Adult Income. We use the OpenML version of the UCI Adult Income dataset [26]. The label is income $>$ \$50K. The protected attributes are race, sex, age binned into $\{< 25, 25\text{–}45, 45\text{–}65, > 65\}$, and native country binarized as US/non-US. Non-protected numeric and categorical features are retained, with categorical features one-hot encoded. The stress variant in the main paper uses intersection level 4 and a 10/90 train/test split. After dropping rows with missing values, the processed data contain 45,222 examples, 55 non-protected features, and a positive-label rate of 24.8%. Over 10 stratified splits used for a dataset audit, this dataset variant has 4,522 training examples, 227.7 ± 4.0 training-set subgroup masks and 256.2 ± 2.6 test-set subgroup masks. This is a budget of only 19.9 ± 0.4 training examples per audited subgroup, and only 4.9 ± 0.1 positive-label examples per subgroup for EO. For the ERM model fit, the test-set EO audit has 220.7 ± 2.6 EO-defined subgroups and certifies 61.4 ± 8.8 significant EO violations at $\alpha = 0.05$ after BH correction.

COMPAS. We use the OpenML COMPAS recidivism dataset [2]. The label is two-year recidivism. The protected attributes are race (African-American/Caucasian in the OpenML version), sex, and age category (< 25, 25–45, > 45). The processed data contain 4,966 examples, 5 non-protected features, and a positive-label rate of 50.0%. The headline Pareto comparison uses intersection level 3 and a 50/50 train/test split. Over ten stratified splits this gives 2,483 training examples and exactly 35 audited subgroup masks in both train and test, a budget of 70.9 training examples and 35.5 positive-label examples per audited subgroup. We run an ERM test-set EO audit with 10 splits that certifies 27.3 ± 2.0 significant EO violations at $\alpha = 0.05$ after BH correction. The mechanism ablation uses the more severe 20/80 split: 993 training examples, the same 35 subgroup masks, and only 28.4 training examples and 14.2 positives per audited subgroup. Under this stress split, ERM certifies 29.3 ± 2.2 significant EO violations, making it the high-regularization stress test for the dense-tail component.

German Credit. We use the OpenML version of the Statlog German Credit dataset [21]. The label is good credit. The protected attributes are sex derived from `personal_status`, age binned into $\{< 30, 30\text{–}50, > 50\}$, and foreign-worker status. The processed data contain 1,000 examples, 53

non-protected features, and a positive-label rate of 70.0%. The stress variant uses intersection level 3 and a 50/50 train/test split. Over ten stratified splits this gives 500 training examples, 32.1 ± 1.4 training-set subgroup masks, and 31.1 ± 0.3 test-set subgroup masks, a budget of 15.6 ± 0.7 training examples and 10.9 ± 0.5 positive-label examples per audited subgroup. For the ERM model fit on these same splits, the test-set EO audit has 31.1 ± 0.3 EO-defined subgroups and certifies only 0.3 ± 0.9 significant EO violations at $\alpha = 0.05$ after BH correction, with MSEE 0.0007 ± 0.0018. This is therefore a safety-regime test rather than a fairness-improvement test.

ACS Coverage. We use the ACS Public Coverage task from the Folktables benchmark [10], which predicts public health insurance coverage from American Community Survey records. The protected attributes are race, sex, and age. The reported variant uses intersection level 3 and a 50/50 train/test split. The data contains 50,000 examples, so each split has 25,000 training examples. Across the dataset audit there are exactly 71 training-set and test-set subgroup masks, all EO-defined. This corresponds to a budget of 352.1 training examples and 103.5 positive-label examples per audited subgroup ($\approx 29.4\%$ positives). Because the subgroup budget is much larger than in Adult, COMPAS, or German, this variant serves as a higher-sample ACS coverage benchmark rather than a sparse-cell stress test.

Taiwan Credit. We use the UCI Default of Credit Card Clients dataset [39], which predicts credit card default. The protected attributes are sex, education, marriage, and age. The headline variant uses intersection level 4 and a 50/50 train/test split. The processed data contain 30,000 examples, so each split has 15,000 training examples. Over ten stratified splits this gives 285.4 ± 2.8 training-set subgroup masks and 285.9 ± 2.3 test-set subgroup masks. This corresponds to a budget of 52.6 ± 0.5 training examples and 11.6 ± 0.1 positive-label examples per audited subgroup; 245.8 subgroups are EO-defined on average. This variant therefore provides a high-intersectionality credit-risk benchmark with a larger total training set than Adult but still a relatively small positive-label budget per audited subgroup.

B Training details and hyperparameters

Model and optimizer. The headline neural methods use the same MLP classifier: hidden layers $(64, 32)$, ReLU activations, dropout 0.2, Adam optimizer ($\mathrm{lr} = 10^{-3}$, weight decay 5×10^{-4}), batch size 256, gradient clipping at norm 1, and at most 200 epochs. Prediction loss is binary cross-entropy. Fairness statistics are computed from full-training-set soft counts each epoch rather than mini-batches, because subgroup-level LRT and significant-excess quantities are unstable on mini-batch counts.

Splits. The main Adult L4 10/90 and COMPAS L3 50/50 Pareto comparisons use 100 stratified shuffle splits and 3 random seeds per split. The mechanism ablations use 10 splits $\times$ 3 seeds on Adult L4 10/90, 100 splits $\times$ 3 seeds on COMPAS L3 20/80, and 30 splits $\times$ 3 seeds on German L3 50/50. Within the training portion of every split, 10% is held out for model selection. The held-out test partition is never used for hyperparameter or checkpoint selection.

Checkpointing. SAFER and its single-knob ablations use the joint BCE-guarded training configuration. The selector first restricts attention to checkpoints whose validation BCE is within a relative tolerance of the best validation BCE, and then selects the checkpoint with the smallest guarded score, using the full-training-set fairness loss for the fairness term. This prevents high λ runs from selecting collapsed models while still allowing fairness improvements when their validation-BCE cost is small. ERM selects by validation BCE alone. The headline EO-Penalty and FAIRRET runs use their standard joint-training selectors unless otherwise stated.

SAFER hyperparameters. The main SAFER configuration is the unweighted subgroup average (`use_size_weight=false`) with $\alpha = 0.05$, threshold smoothing temperature $\tau = 0.05$, softplus temperature $\tau_s = 0.01$, absolute-value smoothing $\varepsilon = 10^{-3}$, detached LRT-based $\widehat{SE}_{\mathrm{eff}}$, and LRT dense-tail weight $w_{\mathrm{tail}} = 0.003$. The NoThreshSmooth ablation sets τ to null; NoSelectivity sets the hinge threshold to zero; NoDenseTail sets $w_{\mathrm{tail}} = 0$.

Baselines. ERM is the unconstrained MLP. EO-Penalty is BCE plus an unweighted squared EO-disparity penalty. FAIRRET Norm 1 and FAIRRET Norm 2 use the ratio-normalized differentiable fairness regularizers of FAIRRET [6], with the EO variant conditioning on positive labels. We also swept ExponGrad [1] with TPR-parity constraints, Prejudice Remover, and adversarial debiasing

Table A.1: Regularization grids used in the Results section. Main Pareto comparisons use 100 splits $\times$ 3 seeds; Adult ablations use 10 splits $\times$ 3 seeds.

Experiment	Method	Grid
Adult L4 10/90 main	SAFER	$0.05, 0.1, 0.2, 0.3, 0.5, 1, 2, 5, 10, 20, 30, 50, 80, 100$
Adult L4 10/90 main	EO-Penalty	$10, 20, 50, 100, 200, 500, 1000, 2000, 4000$
Adult L4 10/90 main	FAIRRET Norm 2	$0.01, 0.1, 0.5, 1, 2, 5, 10, 20, 50, 100, 500, 1000$
Adult L4 10/90 ablation	SAFER / NoThresh / NoTail	$5, 10, 20, 50, 100, 200, 500, 1000$
Adult L4 10/90 ablation	NoSelectivity	$0.5, 1, 2, 5, 10, 20, 50, 100$
COMPAS L3 50/50 main	SAFER	$0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 100$
COMPAS L3 50/50 main	EO-Penalty	$0.1, 0.5, 1, 2, 5, 10, 20, 30, 50, 80, 100, 200, 500, 1000$
COMPAS L3 50/50 main	FAIRRET Norm 1	$0.001, 0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 1, 10$
COMPAS L3 50/50 main	FAIRRET Norm 2	$0.005, 0.01, 0.02, 0.03, 0.06, 0.1, 0.15, 0.2, 0.3, 0.5, 1, 10, 100$
COMPAS L3 20/80 ablation	SAFER / NoThresh / NoTail	$0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 100$
COMPAS L3 20/80 ablation	NoSelectivity	$0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 1, 2, 5, 10, 100$
German L3 50/50 ablation	SAFER / NoThresh / NoTail	$1, 2, 5, 10, 20, 50, 100$
German L3 50/50 ablation	NoSelectivity	$0.01, 1, 2, 3, 4, 5, 10, 20, 50, 100$

Table A.2: Estimated CPU wall-clock time for the full sweeps on the 10-core, 32 GB Apple M1 Max machine with eight joblib workers. SAFER-ablation variants (NoThreshSmooth, NoSelectivity, and NoDenseTail) have the same resource profile as SAFER up to small constant-factor changes.

Experiment	Method	λ jobs	Split $\times$ seed runs per job	Estimated wall time
Adult L4 10/90 main	SAFER	14	100×3	3–6 h
Adult L4 10/90 main	EO-Penalty	9	100×3	1–1.5 h
Adult L4 10/90 main	FAIRRET Norm 1 / Norm 2	12–13	100×3	20–45 min
COMPAS L3 50/50 main	SAFER	9	100×3	20–40 min
COMPAS L3 50/50 main	EO-Penalty	14	100×3	15–30 min
COMPAS L3 50/50 main	FAIRRET Norm 1 / Norm 2	9–13	100×3	5–10 min
German L3 50/50 main	SAFER	7	100×3	20–30 min
German L3 50/50 main	EO-Penalty	14	100×3	10–20 min
German L3 50/50 main	FAIRRET Norm 1 / Norm 2	9–13	100×3	under 10 min
Adult L4 10/90 ablation	SAFER-family arm	8	30×3	35–55 min per arm
COMPAS L3 20/80 ablation	SAFER-family arm	10–11	100×3	20–40 min per arm
German L3 50/50 ablation	SAFER-family arm	7–10	30×3	5–15 min per arm

where compatible. These additional baselines are included in the thresholded summary, but they do not all share the same neural in-training backbone as the headline Pareto comparisons.

Uncertainty reporting. Figures report means and ± 1 standard error over all combinations of split and seed. Because three seeds share the same data split, these observations are not fully independent. We therefore treat the plotted SE as descriptive and verified that the method ordering is unchanged when seeds are averaged within split before computing SE.

C Computational resources

All reported neural sweeps were run on CPU workers. The machine used for the final local runs was an Apple MacBook Pro with an Apple M1 Max processor (10 CPU cores: 8 performance and 2 efficiency cores) and 32 GB of unified memory, running macOS/Darwin on arm64. GPU execution was disabled in the experiment scripts. Independent λ jobs were parallelized across up to 8 CPU worker processes. Each worker trains one combination of method and λ over its assigned sequence of splits and seeds. Peak memory was dominated by the Python/PyTorch worker processes and the dataset matrices; on this 32 GB machine the reported CPU sweeps ran without memory pressure when running all eight processes.

Table A.2 gives wall-clock estimates for reproducing the full neural sweeps. These estimates exclude plotting and table aggregation, which take seconds to minutes, and should be interpreted as order-of-magnitude wall-clock requirements for this CPU workstation rather than hardware-independent constants.

D Per-dataset thresholded baseline summaries

The following tables expand Table 2 with per-dataset details. For eligible rows, we report the best mean test log loss among swept hyperparameters whose MSEE is at least 50% below the ERM MSEE on that dataset. For ineligible rows, corresponding to $*$ entries in the main summary table, we report the minimum-MSEE point found by the sweep.

Table A.3: Adult Income: Best mean test log loss with MSEE at least 50% below ERM MSEE. (ERM MSEE=0.0201, threshold=0.0101; ineligible rows show the minimum-MSEE point).

Method	Parameter	Eligible	Log loss	AUC	MSEE
SAFER	$\lambda = 30$	yes	$\mathbf{0.3829} \pm 0.0003$	$\mathbf{0.8818} \pm 0.0003$	0.0097 ± 0.0004
EO-Penalty	$\lambda = 200$	yes	0.4106 ± 0.0009	0.8624 ± 0.0006	0.0093 ± 0.0006
FAIRRET Norm 1	$\lambda = 0.02$	yes	0.4272 ± 0.0012	0.8561 ± 0.0007	0.0097 ± 0.0006
FAIRRET Norm 2	$\lambda = 10$	yes	0.4618 ± 0.0030	0.8456 ± 0.0022	0.0085 ± 0.0005
AdvDebiasing	$\alpha_{\text{adv}} = 0.5$	no	0.4770 ± 0.0066	0.8208 ± 0.0034	0.0157 ± 0.0009
ExponGrad	$\epsilon_{\text{EG}} = 0.05$	no	0.5936 ± 0.0238	0.7373 ± 0.0066	0.0335 ± 0.0018
PrejudiceRemover	$\lambda = 5$	no	1.1022 ± 0.1035	0.4367 ± 0.0835	0.0165 ± 0.0038

Table A.4: COMPAS: Best mean test log loss with MSEE at least 50% below ERM (ERM MSEE=0.1292, threshold=0.0646; ineligible rows show the minimum-MSEE point).

Method	Parameter	Eligible	Log loss	AUC	MSEE
SAFER	$\lambda = 0.5$	yes	$\mathbf{0.6252} \pm 0.0007$	0.7148 ± 0.0009	0.0510 ± 0.0013
EO-Penalty	$\lambda = 10$	yes	0.6339 ± 0.0005	$\mathbf{0.7171} \pm 0.0009$	0.0661 ± 0.0015
FAIRRET Norm 1	$\lambda = 0.03$	yes	0.6559 ± 0.0009	0.6975 ± 0.0018	0.0614 ± 0.0021
FAIRRET Norm 2	$\lambda = 0.1$	yes	0.6366 ± 0.0007	0.7117 ± 0.0010	0.0628 ± 0.0015
AdvDebiasing	$\alpha_{\text{adv}} = 0.5$	no	0.7381 ± 0.0449	0.6522 ± 0.0170	0.1029 ± 0.0095
ExponGrad	$\epsilon_{\text{EG}} = 0.05$	no	0.7211 ± 0.0061	0.5291 ± 0.0096	0.1259 ± 0.0045
PrejudiceRemover	$\lambda = 0.1$	no	0.6595 ± 0.0037	0.7066 ± 0.0023	0.0939 ± 0.0019

Table A.5: German Credit: Best mean test log loss with MSEE at least 50% below ERM (ERM MSEE=0.0013, threshold=0.0007; ineligible rows show the minimum-MSEE point).

Method	Parameter	Eligible	Log loss	AUC	MSEE
SAFER	$\lambda = 30$	yes	$\mathbf{0.5349} \pm 0.0014$	$\mathbf{0.7659} \pm 0.0017$	0.0006 ± 0.0001
EO-Penalty	$\lambda = 200$	yes	0.5358 ± 0.0013	0.7554 ± 0.0017	0.0006 ± 0.0001
FAIRRET Norm 1	$\lambda = 0.12$	yes	0.5409 ± 0.0016	0.7536 ± 0.0019	0.0006 ± 0.0001
FAIRRET Norm 2	$\lambda = 0.4$	yes	0.5408 ± 0.0016	0.7544 ± 0.0018	0.0006 ± 0.0001
AdvDebiasing	$\alpha_{\text{adv}} = 0.1$	no	1.3272 ± 0.0141	0.7247 ± 0.0024	0.0027 ± 0.0003
ExponGrad	$\epsilon_{\text{EG}} = 0.1$	no	0.5664 ± 0.0033	0.7371 ± 0.0027	0.0018 ± 0.0002
PrejudiceRemover	$\lambda = 5$	no	0.8915 ± 0.0110	0.6380 ± 0.0098	0.0104 ± 0.0009

Table A.6: Taiwan: Best mean test log loss with MSEE at least 50% below ERM (ERM MSEE=0.0060, threshold=0.0030; ineligible rows show the minimum-MSEE point).

Method	Parameter	Eligible	Log loss	AUC	MSEE
SAFER	$\lambda = 2$	yes	$\mathbf{0.4479} \pm 0.0004$	$\mathbf{0.7552} \pm 0.0008$	0.0021 ± 0.0002
EO-Penalty	$\lambda = 5$	yes	0.4527 ± 0.0006	0.7531 ± 0.0010	0.0025 ± 0.0002
FAIRRET Norm 1	$\lambda = 0.002$	yes	0.4668 ± 0.0010	0.7442 ± 0.0014	0.0014 ± 0.0002
FAIRRET Norm 2	$\lambda = 0.05$	yes	0.4624 ± 0.0012	0.7455 ± 0.0015	0.0021 ± 0.0002
AdvDebiasing	$\alpha_{\text{adv}} = 5$	no	0.4569 ± 0.0009	0.7546 ± 0.0013	0.0032 ± 0.0002
ExponGrad	$\epsilon_{\text{EG}} = 0.1$	yes	0.6256 ± 0.0347	0.6096 ± 0.0101	0.0027 ± 0.0006
PrejudiceRemover	$\lambda = 0.01$	yes	0.6216 ± 0.0107	0.6093 ± 0.0103	0.0017 ± 0.0002

E Subgroups below the certification boundary

A zero significant excess for a subgroup states only that the available sample cannot certify an excess at the chosen confidence level. The level α sets the trade between reacting to sampling noise and

Table A.7: ACS: Best mean test log loss with MSEE at least 50% below ERM (ERM MSEE=0.0904, threshold=0.0452; ineligible rows show the minimum-MSEE point).

Method	Parameter	Eligible	Log loss	AUC	MSEE
SAFER	$\lambda = 0.3$	yes	$\mathbf{0.5198} \pm 0.0003$	$\mathbf{0.7472} \pm 0.0006$	0.0236 ± 0.0009
EO-Penalty	$\lambda = 10$	yes	0.5223 ± 0.0004	0.7427 ± 0.0007	0.0383 ± 0.0009
fairret-norm1	$\lambda = 0.005$	yes	0.5240 ± 0.0004	0.7401 ± 0.0006	0.0357 ± 0.0011
fairret-norm2	$\lambda = 0.05$	yes	0.5314 ± 0.0004	0.7320 ± 0.0007	0.0218 ± 0.0009
AdvDebiasing	$\alpha_{\text{adv}} = 0.5$	no	0.5299 ± 0.0111	0.7450 ± 0.0030	0.0605 ± 0.0025
ExponGrad	$\epsilon_{\text{EG}} = 0.005$	yes	0.6535 ± 0.0134	0.5623 ± 0.0078	0.0257 ± 0.0029
PrejudiceRemover	$\lambda = 10$	no	0.6196 ± 0.0019	0.6401 ± 0.0018	0.0464 ± 0.0013

leaving low-power disparities uncorrected; a larger α protects smaller groups at a utility cost. The size sweep of Appendix F.3 makes the dependence concrete: the same planted gap of -0.15 is certified in about 9% of runs at $m_S = 5$ and in about 95% at $m_S = 120$. We therefore recommend reporting MSEE together with raw subgroup disparities, confidence intervals, subgroup counts and a power or minimum-detectable-effect analysis, and treating an uncertified subgroup as unresolved; possible responses are collecting more data, pooling groups coarsely or hierarchically, or policy constraints that do not rely on statistical significance alone. Where harm is suspected in a subgroup too small to certify, the appropriate evidence is of a different kind: the structure of the decision procedure, the features it uses, the causal path from the protected attribute to the outcome, and documented practice. A complete audit combines statistical outcome testing with this process evidence, and the certification boundary marks where the hand-over occurs. Medical-imaging benchmarks whose small subgroups carry genuine disparities, such as Harvard-FairVision30k [31], are a natural next test of SAFER outside the tabular setting.

F Additional experiments

F.1 Decision-threshold sensitivity

The threshold t does not enter the prediction loss, which trains probabilistic scores; it defines the deployed classifier and therefore both its accuracy and its hard-decision fairness audit. We use $t = 0.5$ because, when scores estimate class probabilities and false positives and false negatives cost the same, it is the Bayes-optimal decision threshold. Since t enters SAFER's objective, it is fixed before training: by a deployment requirement such as a cost matrix or a regulatory decision rule, or, when no policy value exists, on validation data with a preliminary ERM model, and then held fixed while all methods are trained and compared. To test the dependence on this default we retrained SAFER for every $t \in \{0.1, \ldots, 0.9\}$ on the three datasets of the main Pareto comparison, five seeds per threshold, and audited each model at its own threshold. Table A.8 reports the test log-loss, the paired difference $\Delta\text{MSEE} = \text{MSEE}_{\text{SAFER}} - \text{MSEE}_{\text{ERM}}$ against the ERM audit at the same threshold (negative values are a fairness improvement), and the accuracy. Log-loss is computed from probabilistic scores, so for a fixed model it does not depend on t; its variation across columns measures how t changes the retrained model through the fairness term. Log-loss is stable across t on Adult and German and more sensitive on COMPAS. The fairness benefit persists across operationally meaningful thresholds: Adult has non-positive ΔMSEE throughout, COMPAS improves for t between 0.2 and 0.9, and the changes on German are negligible because little certified disparity is available to reduce. Extreme thresholds produce nearly constant classifiers with trivially low MSEE (accuracy 0.500 on COMPAS for $t \leq 0.3$, 0.700 on German for $t \leq 0.2$), so those columns are not useful operating points. This is a sensitivity analysis; t is never tuned on the test set.

F.2 A baseline that optimizes the Wald statistic

Wald-Stat penalizes the standardized Wald statistic directly:

$$\mathcal{L}_{\text{Wald-Stat}} = \mathcal{L}_{\text{pred}} + \frac{\lambda}{|\mathfrak{S}|} \sum_{S \in \mathfrak{S}} \left[\text{softplus}_{\tau_s} \left(|Z_S| - z_\alpha \right)^2 + w_{\text{tail}} |Z_S| \right], \qquad Z_S = \frac{\widehat{\phi}_S^{\tau}}{\widehat{SE}_S^{\phi}}, \quad \text{(A.1)}$$

where $\widehat{\phi}_S^{\tau}$ is the disparity computed from the same threshold-smoothed counts as SAFER and $\widehat{SE}_S^{\phi}$ its two-proportion plug-in standard error (Section 4.2). Both stay in the computation graph, so

Table A.8: SAFER retrained and audited at decision thresholds t from 0.1 to 0.9: test log-loss, paired ΔMSEE against ERM at the same threshold, and accuracy (mean $\pm$ standard deviation over five seeds).

	t	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Adult	log-loss	$.336 \pm .002$	$.337 \pm .002$	$.337 \pm .003$	$.336 \pm .002$	$.336 \pm .002$	$.336 \pm .002$	$.337 \pm .002$	$.338 \pm .003$	$.337 \pm .003$
	ΔMSEE	$-.008 \pm .009$	$-.013 \pm .005$	$-.001 \pm .009$	$-.007 \pm .006$	$-.009 \pm .006$	$-.008 \pm .003$	$-.014 \pm .004$	$-.010 \pm .005$	$-.008 \pm .005$
	accuracy	$.696 \pm .004$	$.778 \pm .003$	$.822 \pm .003$	$.839 \pm .002$	$.843 \pm .002$	$.837 \pm .002$	$.817 \pm .002$	$.785 \pm .004$	$.762 \pm .002$
COMPAS	log-loss	$.615 \pm .008$	$.615 \pm .007$	$.623 \pm .006$	$.633 \pm .005$	$.640 \pm .008$	$.643 \pm .007$	$.628 \pm .005$	$.620 \pm .005$	$.615 \pm .005$
	ΔMSEE	$.000 \pm .000$	$-.005 \pm .004$	$-.089 \pm .019$	$-.131 \pm .021$	$-.084 \pm .018$	$-.069 \pm .018$	$-.078 \pm .012$	$-.051 \pm .015$	$-.011 \pm .005$
	accuracy	$.500 \pm .000$	$.500 \pm .000$	$.500 \pm .000$	$.535 \pm .040$	$.631 \pm .016$	$.531 \pm .024$	$.510 \pm .006$	$.505 \pm .003$	$.502 \pm .001$
German	log-loss	$.531 \pm .035$	$.531 \pm .035$	$.532 \pm .035$	$.532 \pm .035$	$.531 \pm .035$	$.531 \pm .035$	$.531 \pm .035$	$.532 \pm .035$	$.531 \pm .035$
	ΔMSEE	$.0000 \pm .0000$	$.0000 \pm .0000$	$.0000 \pm .0000$	$-.0005 \pm .0013$	$+.0002 \pm .0014$	$-.0005 \pm .0007$	$-.0014 \pm .0018$	$-.0041 \pm .0078$	$-.0006 \pm .0027$
	accuracy	$.700 \pm .000$	$.700 \pm .000$	$.702 \pm .004$	$.722 \pm .018$	$.731 \pm .019$	$.737 \pm .014$	$.673 \pm .109$	$.590 \pm .162$	$.478 \pm .103$

gradients flow through the statistic, including the route that lowers the penalty by changing the estimated standard error. Wald-Stat uses the same critical value, threshold smoothing, temperatures, dense-tail weight, architecture, optimizer and checkpointing protocol as SAFER; it tests whether optimizing Wald significance directly yields a stronger fairness–utility frontier than SAFER's LRT-based construction. Because the two objectives live on different scales, we sweep λ independently for each and match operating points by mean test log-loss after examining both sweeps; Table A.9 summarizes the matched pairs. Entries are SAFER/Wald-Stat means $\pm$ standard deviations over independent runs over splits and seeds, 10 for Adult and COMPAS and 5 for German, rounded to three decimals; several differences are small relative to this variability, so we read the frontier rather than individual rows. For the following table, the ERM reference values in the caption come from these same runs, a lighter protocol than the 100 splits with three seeds of the main sweeps, so they differ from the ERM values of the main tables.

Table A.9: SAFER versus Wald-Stat at matched test log-loss. Each row pairs a SAFER and a Wald-Stat fairness weight $(\lambda_{\mathrm{S}}, \lambda_{\mathrm{W}})$ with similar mean test log-loss; entries are SAFER/Wald-Stat, mean $\pm$ standard deviation over runs. ERM log-loss/MSEE: Adult $0.334/0.041$, COMPAS $0.613 \pm 0.009/0.127 \pm 0.027$, German $0.531 \pm 0.035/0.006 \pm 0.005$.

	$(\lambda_{\mathrm{S}}, \lambda_{\mathrm{W}})$	log-loss S/W	MSEE S/W
Adult	$(1, 0.0001)$	$.337 \pm .003/.338 \pm .004$	$.028 \pm .013/.040 \pm .013$
	$(5, 0.015)$	$.344 \pm .004/.344 \pm .005$	$.028 \pm .015/.035 \pm .013$
	$(15, 0.5)$	$.372 \pm .005/.372 \pm .005$	$.025 \pm .011/.025 \pm .014$
	$(30, 2)$	$.387 \pm .003/.391 \pm .003$	$.005 \pm .010/.002 \pm .007$
	$(50, 3)$	$.397 \pm .003/.397 \pm .004$	$.0001 \pm .0002/.0003 \pm .0009$
COMPAS	$(0.1, 0.001)$	$.615 \pm .009/.614 \pm .009$	$.086 \pm .026/.087 \pm .025$
	$(0.5, 0.005)$	$.625 \pm .010/.621 \pm .008$	$.058 \pm .013/.067 \pm .017$
	$(1, 0.015)$	$.647 \pm .012/.648 \pm .007$	$.041 \pm .015/.052 \pm .019$
	$(3, 0.05)$	$.671 \pm .008/.669 \pm .006$	$.036 \pm .026/.046 \pm .030$
	$(20, 1)$	$.690 \pm .003/.690 \pm .003$	$.040 \pm .048/.038 \pm .050$
German	$(0.25, 0.0005)$	$.531 \pm .035/.531 \pm .035$	$.006 \pm .005/.006 \pm .005$
	$(1, 0.01)$	$.532 \pm .035/.531 \pm .035$	$.007 \pm .005/.008 \pm .006$
	$(15, 0.02)$	$.532 \pm .032/.532 \pm .035$	$.006 \pm .006/.007 \pm .006$
	$(20, 0.1)$	$.533 \pm .032/.533 \pm .034$	$.004 \pm .004/.005 \pm .005$
	$(40, 3)$	$.540 \pm .033/.541 \pm .030$	$.0005 \pm .0007/.0015 \pm .0021$

Diagnostics on Adult, stratified by the positive-label count m_S, explain the gap. Across the five matched points, the median of $|Z_S|/\sqrt{\Lambda_S}$ is 1.00 for $m_S \geq 30$. For $m_S < 10$, saturation drives the plug-in standard error toward zero and the ratio's 90th percentile is 37 to 43, a roughly fortyfold inflation in the most extreme cases. Among about $1{,}491$ pairs of subgroup and seed per point, Wald-Stat flags 41 to 57 disparities that the LRT does not, and never the reverse. These saturated groups are 2.7% of the total but absorb 99% of Wald-Stat's penalty and 86% of its final-epoch gradient over the smallest subgroups ($m_S < 30$), against 43% for SAFER. At the near-ERM Adult point, the residual certified excess of the groups with $m_S > 300$ is 0.0029 for SAFER and 0.0214 for Wald-Stat, so SAFER preserves capacity for the better-supported groups. SAFER also has lower mean MSEE near ERM on Adult and at the intermediate COMPAS and German points, with gaps that are modest relative to the variability across splits and seeds; under strong regularization both methods converge near zero MSEE at a greater utility cost.

F.3 A controlled benchmark for the selective-gating regime

The operating regimes of Section 4.3 suggest a hypothesis for the gate threshold: when both active and inactive subgroups are present, the gate gives SAFER its largest advantage over NoSelectivity, because it withholds penalties from unsupported disparities while keeping those with support. The benchmark below tests it by separating statistical evidence from disparity magnitude. It has one large reference cell and 34 probe cells. The generator is calibrated so that its population decision rule has a true-positive rate of 0.85 on the reference cell and 0.70 on every unfair probe, a planted reference gap of -0.15; null probes have zero gap. Only the positive-label count m_S of the unfair probes varies, and the reference count is fixed and large. Group identity is hidden from the predictor and used only by the fairness loss and the audit, which keep the S-versus-$\bar{S}$ comparison of the main text. We summarize each evidence stratum by the certification statistic $\sqrt{\Lambda_S}$ computed on the hard predictions of the final ERM model on the training split, the same statistic as in the size sweep below (Table A.10). The null and strong-evidence stratum means lie on opposite sides of the gate in every one of the 10 runs, so the benchmark is in the selective-gating regime; cell by cell the picture is noisier, with at least one null cell certified in 6 runs and one strong-evidence cell uncertified in 1 run.

Table A.10: Evidence strata of the probe cells for the final ERM model: cells per stratum, planted gap, mean positive-label count m_S in the training split, mean certification statistic $\sqrt{\Lambda_S}$ on hard training predictions, and fraction of cells certified ($\sqrt{\Lambda_S} > 1.96$); per-run stratum means, reported as mean $\pm$ standard deviation over 10 runs.

probe stratum	cells	planted gap	m_S	mean $\sqrt{\Lambda_S}$	certified fraction
null	10	0	105	$.90 \pm .16$	$.07 \pm .07$
weak evidence	12	-0.15	6	$1.11 \pm .25$	$.11 \pm .07$
boundary	8	-0.15	34	$1.77 \pm .29$	$.45 \pm .13$
strong evidence	4	-0.15	204	$4.72 \pm .50$	$.98 \pm .08$

Generator and protocol. Every cell draws labels $y \sim \mathrm{Bernoulli}(0.5)$. One signal feature is distributed as $\mathcal{N}(\Phi^{-1}(\mathrm{TPR}_{\mathrm{cell}}), 1)$ for positives and $\mathcal{N}(-\Phi^{-1}(0.85), 1)$ for negatives, where $\mathrm{TPR}_{\mathrm{cell}}$ is the planted true-positive rate of the cell and Φ is the standard normal distribution function; eight pure-noise $\mathcal{N}(0, 1)$ features are appended. The reference cell has $45{,}000$ examples and planted rate 0.85, which pins the decision threshold of a group-blind classifier so that each cell recovers its planted rate and the complement of every probe has a true-positive rate close to 0.85. A probe cell with design size m has $4m$ examples, so about $0.9m$ positives fall in the training split; the design sizes are $10, 10, 20, 20, 40, 40, 120, 120, 400, 400$ for the ten null probes (planted rate 0.85), $5, 5, 6, 6, 7, 7, 8, 8, 5, 6, 7, 8$ for the twelve weak-evidence probes, $30, 30, 35, 35, 40, 40, 45, 45$ for the eight boundary probes and $150, 150, 300, 300$ for the four strong-evidence probes (planted rate 0.70). Each of the 10 runs draws a fresh dataset of $54{,}832$ examples, a fresh split and a fresh initialization; within a run all arms share the initialization and the dropout stream, so comparisons are paired. Half of the data is held out as test; of the rest, 90% trains the model and 10% is set aside for validation and not used. The model is an MLP with hidden layers $(64, 32)$ and dropout 0.2, trained full-batch for 120 epochs with Adam (learning rate 10^{-3}, weight decay 5×10^{-4}); the final model is evaluated, without early stopping or checkpoint selection, so that λ traces the whole fairness–utility frontier. SAFER uses the values of the main experiments ($\alpha = 0.05$, $\tau = 0.05$, $\tau_s = 0.01$, $\varepsilon = 10^{-3}$, $w_{\mathrm{tail}} = 0.003$); NoSelectivity removes only the gate and EO-Penalty adds λ times the squared EO gap to the prediction loss. The λ grid is chosen separately for each arm, from near-ERM to a near-constant predictor. Every cell is scored on its training and its test members with hard predictions at 0.5; a cell is skipped on a split where it or its complement has fewer than two positives, which affects only the weak-evidence stratum (10 to 12 of its 12 cells scored per run). Stratum values are cell averages, reported as mean $\pm$ standard deviation over the 10 runs. The generator, the harness and the per-run results are released with the code.

We compare SAFER with NoSelectivity (NS), which removes only the gate, and with EO-Penalty (EO), over independent λ sweeps. Table A.12 reports the complete sweeps: for every arm and λ, the test log-loss and the mean residual certified excess in each evidence stratum. Points whose mean test log-loss reaches 0.5 approach the log-loss 0.693 of a constant predictor, and their near-zero excess comes from degenerate predictions. From the remaining part of the frontiers we chose three approximately matched mean test-log-loss levels, reported in Table A.11 with the fairness weight and

test log-loss of each arm and the mean residual certified excess in each evidence stratum, in percentage points, as means $\pm$ standard deviations over 10 runs. The methods overlap at the near-ERM level A. At B and C, SAFER has the lowest mean excess in the boundary and strong-evidence strata; in original units, SAFER's C values in these two strata are 0.0079 and 0.0227, against 0.0184 and 0.0465 for NS and 0.0108 and 0.0429 for EO. SAFER's null-stratum mean at C is higher and the standard deviations overlap, so the evidence is directional, with no statistically significant pairwise differences at matched utility. The full sweeps show the same pattern: EO's strong-stratum excess stays near 4 percentage points from $\lambda = 50$ to $\lambda = 600$ while its log-loss rises from 0.47 to 0.63, and NS becomes unstable above $\lambda \approx 1$, with large run-to-run variation of the log-loss from $\lambda = 1.25$ on. The strong-evidence gain is an indirect effect of the whole objective: the gate is already active for those probes, and withholding unsupported gradients preserves optimization capacity for well-supported ones. Table A.13 reports the raw gap $|\phi_S|$ on training and on test members for the non-degenerate points. On the weak-evidence stratum, relative to ERM (training gap 0.185, test gap 0.200), EO at $\lambda = 50$ lowers the training gap by 0.090 but the test gap by only 0.037, whereas SAFER at $\lambda = 15$, at similar log-loss, lowers them by 0.098 and 0.071: the certainty-blind penalty fits training-set noise in the sparse cells. An independent size sweep (Table A.14) locates the transition for this planted gap: the certified fraction rises from 0.09 ± 0.09 at design size $m = 5$ to 0.52 ± 0.15 at $m = 30$, where the mean statistic equals 1.96, and to 0.95 ± 0.08 at $m = 120$; the value $m \approx 30$ is specific to this gap and reference size. On the real datasets, Adult and COMPAS were built as candidates for this regime; NoSelectivity tracks SAFER across the sweep on COMPAS and up to $\lambda = 1$ on Adult, where it then loses utility (Appendix G.4).

Table A.11: Matched-utility comparison on the controlled benchmark: fairness weight λ, test log-loss, and mean residual certified excess per evidence stratum in percentage points (mean $\pm$ standard deviation over 10 runs; lower is better).

set	arm	λ	test log-loss	null	weak	boundary	strong
–	ERM	0	$.356 \pm .004$	$.89 \pm 1.03$	3.49 ± 2.69	$2.01 \pm .74$	7.42 ± 1.30
A	SAFER	2	$.362 \pm .003$	$1.09 \pm .77$	3.35 ± 2.13	$1.51 \pm .56$	6.35 ± 1.16
	NS	.5	$.361 \pm .004$	$1.09 \pm .76$	3.63 ± 1.90	$1.54 \pm .63$	6.58 ± 1.01
	EO	5	$.367 \pm .004$	$1.07 \pm .72$	3.14 ± 2.16	$1.54 \pm .56$	6.47 ± 1.07
B	SAFER	5	$.381 \pm .006$	1.14 ± 1.06	2.88 ± 1.88	$1.25 \pm .73$	4.90 ± 1.28
	NS	1	$.379 \pm .008$	$1.23 \pm .92$	3.30 ± 2.48	$1.49 \pm .81$	5.63 ± 1.05
	EO	10	$.381 \pm .006$	$1.17 \pm .86$	3.25 ± 2.24	$1.44 \pm .54$	5.98 ± 1.08
C	SAFER	15	$.457 \pm .031$	$1.41 \pm .50$	2.51 ± 1.62	$.79 \pm .76$	2.27 ± 1.29
	NS	1.25	$.467 \pm .066$	$1.06 \pm .84$	3.32 ± 2.24	1.84 ± 1.26	4.65 ± 1.29
	EO	50	$.466 \pm .011$	$1.04 \pm .91$	2.61 ± 1.73	$1.08 \pm .83$	$4.29 \pm .96$

G Ablation study details

The arm definitions and detailed numbers for the mechanism ablations of Section 5.2 are given below.

G.1 Arm definitions

The four arms are evaluated on the three stress variants of Table 1 under the protocols of Section 5:

- **SAFER**: full method, all three components active.
- **NoThreshSmooth**: τ disabled (raw probabilities feed the soft counts).
- **NoSelectivity**: the threshold term $z_\alpha \widehat{SE}_{\mathrm{eff},S}$ is removed from the softplus argument, leaving $\sqrt{\widehat{\phi}_S^{\tau\,2} + \varepsilon^2}$ alone.
- **NoDenseTail**: $w_{\mathrm{tail}} = 0$.

Apart from the toggled knob, every other element of the loss, optimizer, and protocol is identical across arms.

Table A.12: Complete λ sweeps on the controlled benchmark: test log-loss and mean residual certified excess per evidence stratum in percentage points (mean $\pm$ standard deviation over 10 runs; lower is better). NS is NoSelectivity and EO is EO-Penalty. Rows marked † have mean test log-loss at or above 0.5: these models approach the constant predictor (log-loss 0.693), and their small excess values come from degenerate predictions rather than from fairer ones. The matched levels of Table A.11 are drawn from the unmarked rows.

arm	λ	test log-loss	null	weak	boundary	strong
ERM	0	$.356 \pm .004$	$.89 \pm 1.03$	3.49 ± 2.69	$2.01 \pm .74$	7.42 ± 1.30
SAFER	0.5	$.357 \pm .003$	$1.08 \pm .94$	4.02 ± 2.43	$1.91 \pm .84$	6.95 ± 1.21
SAFER	1	$.358 \pm .003$	$1.13 \pm .79$	3.55 ± 2.27	$1.76 \pm .62$	6.77 ± 1.38
SAFER	1.5	$.360 \pm .003$	$1.19 \pm .80$	3.16 ± 2.12	$1.63 \pm .49$	6.74 ± 1.29
SAFER	2	$.362 \pm .003$	$1.09 \pm .77$	3.35 ± 2.13	$1.51 \pm .56$	6.35 ± 1.16
SAFER	2.5	$.365 \pm .004$	$1.02 \pm .73$	3.28 ± 2.07	$1.32 \pm .65$	6.15 ± 1.22
SAFER	5	$.381 \pm .006$	1.14 ± 1.06	2.88 ± 1.88	$1.25 \pm .73$	4.90 ± 1.28
SAFER	10	$.423 \pm .018$	$1.13 \pm .67$	2.53 ± 1.53	$.71 \pm .59$	2.90 ± 1.25
SAFER	15	$.457 \pm .031$	$1.41 \pm .50$	2.51 ± 1.62	$.79 \pm .76$	2.27 ± 1.29
SAFER	20	$.487 \pm .037$	1.36 ± 1.19	2.87 ± 1.98	$.63 \pm .81$	2.01 ± 1.58
SAFER	30	$.539 \pm .018^{\dagger}$	$.75 \pm 1.04$	3.13 ± 3.91	$.89 \pm .86$	1.63 ± 1.24
SAFER	40	$.581 \pm .021^{\dagger}$	$.04 \pm .07$	2.68 ± 3.33	$.50 \pm .75$	$.31 \pm .29$
SAFER	50	$.620 \pm .024^{\dagger}$	$.54 \pm .96$	3.77 ± 2.72	$.53 \pm .73$	$.43 \pm .74$
SAFER	70	$.665 \pm .013^{\dagger}$	$.50 \pm .82$	3.33 ± 3.61	$.42 \pm .77$	$.14 \pm .24$
SAFER	100	$.684 \pm .004^{\dagger}$	$.90 \pm 1.02$	2.62 ± 2.36	$.24 \pm .39$	$.20 \pm .23$
SAFER	150	$.690 \pm .002^{\dagger}$	$.20 \pm .32$	$.30 \pm .52$	$.12 \pm .22$	$.02 \pm .06$
SAFER	200	$.692 \pm .001^{\dagger}$	$.03 \pm .10$	$.04 \pm .12$	$.01 \pm .03$	$.00 \pm .00$
SAFER	300	$.694 \pm .002^{\dagger}$	$.00 \pm .00$	$.00 \pm .00$	$.00 \pm .00$	$.00 \pm .00$
SAFER	500	$.696 \pm .003^{\dagger}$	$.00 \pm .00$	$.00 \pm .00$	$.00 \pm .00$	$.00 \pm .00$
NS	0.25	$.357 \pm .004$	$1.09 \pm .88$	4.15 ± 2.69	$1.83 \pm .64$	6.80 ± 1.10
NS	0.5	$.361 \pm .004$	$1.09 \pm .76$	3.63 ± 1.90	$1.54 \pm .63$	6.58 ± 1.01
NS	0.75	$.366 \pm .005$	$1.05 \pm .68$	3.41 ± 2.30	$1.38 \pm .78$	6.15 ± 1.26
NS	1	$.379 \pm .008$	$1.23 \pm .92$	3.30 ± 2.48	$1.49 \pm .81$	5.63 ± 1.05
NS	1.25	$.467 \pm .066$	$1.06 \pm .84$	3.32 ± 2.24	1.84 ± 1.26	4.65 ± 1.29
NS	1.5	$.528 \pm .073^{\dagger}$	$.57 \pm .74$	4.96 ± 4.86	1.73 ± 1.21	3.18 ± 1.93
NS	1.75	$.562 \pm .077^{\dagger}$	$.58 \pm .58$	4.60 ± 4.17	$.82 \pm .94$	1.10 ± 1.51
NS	2	$.586 \pm .071^{\dagger}$	$.86 \pm .90$	2.71 ± 2.86	$.55 \pm .43$	$.95 \pm 1.24$
NS	2.25	$.608 \pm .060^{\dagger}$	$.48 \pm .62$	1.90 ± 2.28	$.40 \pm .47$	$.59 \pm .78$
NS	2.5	$.627 \pm .051^{\dagger}$	$.36 \pm .53$	1.32 ± 1.70	$.30 \pm .38$	$.43 \pm .51$
NS	2.75	$.643 \pm .043^{\dagger}$	$.30 \pm .41$	1.15 ± 1.74	$.13 \pm .15$	$.23 \pm .38$
NS	3	$.653 \pm .039^{\dagger}$	$.37 \pm .46$	$.62 \pm .85$	$.18 \pm .24$	$.21 \pm .28$
NS	3.5	$.671 \pm .018^{\dagger}$	$.14 \pm .26$	$.24 \pm .55$	$.04 \pm .07$	$.09 \pm .22$
NS	3.75	$.677 \pm .009^{\dagger}$	$.14 \pm .34$	$.17 \pm .45$	$.02 \pm .05$	$.07 \pm .20$
NS	4	$.681 \pm .006^{\dagger}$	$.09 \pm .26$	$.13 \pm .39$	$.01 \pm .02$	$.03 \pm .08$
NS	4.5	$.685 \pm .004^{\dagger}$	$.07 \pm .21$	$.08 \pm .23$	$.00 \pm .00$	$.00 \pm .00$
NS	5	$.688 \pm .003^{\dagger}$	$.03 \pm .11$	$.04 \pm .12$	$.02 \pm .05$	$.00 \pm .00$
NS	5.4	$.689 \pm .002^{\dagger}$	$.02 \pm .06$	$.02 \pm .07$	$.01 \pm .04$	$.00 \pm .00$
EO	5	$.367 \pm .004$	$1.07 \pm .72$	3.14 ± 2.16	$1.54 \pm .56$	6.47 ± 1.07
EO	10	$.381 \pm .006$	$1.17 \pm .86$	3.25 ± 2.24	$1.44 \pm .54$	5.98 ± 1.08
EO	20	$.409 \pm .009$	1.13 ± 1.01	2.87 ± 2.09	$1.36 \pm .78$	$5.06 \pm .91$
EO	50	$.466 \pm .011$	$1.04 \pm .91$	2.61 ± 1.73	$1.08 \pm .83$	$4.29 \pm .96$
EO	60	$.479 \pm .010$	$1.03 \pm .91$	2.57 ± 1.73	$1.04 \pm .76$	4.20 ± 1.03
EO	200	$.561 \pm .011^{\dagger}$	$1.20 \pm .77$	3.27 ± 1.91	$1.13 \pm .93$	4.37 ± 1.23
EO	600	$.633 \pm .011^{\dagger}$	$.61 \pm .76$	2.56 ± 1.78	$1.23 \pm .79$	3.89 ± 1.94
EO	1000	$.660 \pm .008^{\dagger}$	$.54 \pm .54$	3.83 ± 2.85	$.93 \pm .76$	3.22 ± 2.31
EO	1500	$.674 \pm .006^{\dagger}$	$.68 \pm .56$	3.20 ± 1.59	$.77 \pm .86$	2.20 ± 1.72
EO	2000	$.681 \pm .004^{\dagger}$	$.49 \pm .67$	3.05 ± 3.95	$.61 \pm .66$	1.74 ± 1.42
EO	2500	$.684 \pm .003^{\dagger}$	$.42 \pm .76$	3.13 ± 3.28	$.55 \pm .90$	1.41 ± 1.08
EO	5000	$.690 \pm .002^{\dagger}$	$.41 \pm .67$	1.77 ± 2.13	$.50 \pm .65$	$.35 \pm .52$
EO	7500	$.691 \pm .001^{\dagger}$	$.39 \pm .71$	2.03 ± 2.85	$.31 \pm .46$	$.12 \pm .22$
EO	10000	$.692 \pm .001^{\dagger}$	$.27 \pm .52$	1.27 ± 1.66	$.42 \pm .75$	$.06 \pm .13$

Table A.13: Raw EO gap $|\phi_S|$ on test and on training members, mean over the cells of each evidence stratum, for the non-degenerate points of the sweeps (test log-loss below 0.5; mean $\pm$ standard deviation over 10 runs). Train and test values come from the same model. A training gap that falls much faster than the test gap on the sparse strata (weak, boundary) indicates fitting of training-set noise.

			test $\lvert\phi_S\rvert$				train $\lvert\phi_S\rvert$			
arm	λ	test log-loss	null	weak	boundary	strong	null	weak	boundary	strong
ERM	0	$.356 \pm .004$	$.048 \pm .015$	$.200 \pm .027$	$.126 \pm .027$	$.137 \pm .013$	$.054 \pm .009$	$.185 \pm .031$	$.124 \pm .022$	$.134 \pm .017$
SAFER	0.5	$.357 \pm .003$	$.050 \pm .015$	$.201 \pm .028$	$.124 \pm .030$	$.131 \pm .012$	$.050 \pm .011$	$.175 \pm .026$	$.111 \pm .021$	$.125 \pm .015$
SAFER	1	$.358 \pm .003$	$.050 \pm .014$	$.187 \pm .027$	$.122 \pm .029$	$.129 \pm .014$	$.046 \pm .011$	$.170 \pm .027$	$.100 \pm .021$	$.116 \pm .015$
SAFER	1.5	$.360 \pm .003$	$.049 \pm .014$	$.182 \pm .025$	$.118 \pm .028$	$.128 \pm .013$	$.045 \pm .009$	$.169 \pm .027$	$.095 \pm .019$	$.110 \pm .014$
SAFER	2	$.362 \pm .003$	$.047 \pm .013$	$.185 \pm .024$	$.117 \pm .027$	$.124 \pm .012$	$.043 \pm .009$	$.162 \pm .026$	$.092 \pm .016$	$.106 \pm .012$
SAFER	2.5	$.365 \pm .004$	$.045 \pm .011$	$.182 \pm .025$	$.115 \pm .026$	$.121 \pm .013$	$.040 \pm .008$	$.155 \pm .020$	$.087 \pm .013$	$.103 \pm .012$
SAFER	5	$.381 \pm .006$	$.041 \pm .009$	$.170 \pm .023$	$.105 \pm .033$	$.105 \pm .015$	$.038 \pm .009$	$.139 \pm .018$	$.071 \pm .007$	$.085 \pm .014$
SAFER	10	$.423 \pm .018$	$.033 \pm .007$	$.143 \pm .027$	$.084 \pm .026$	$.077 \pm .018$	$.031 \pm .007$	$.099 \pm .016$	$.054 \pm .010$	$.053 \pm .008$
SAFER	15	$.457 \pm .031$	$.035 \pm .013$	$.129 \pm .031$	$.080 \pm .034$	$.068 \pm .022$	$.027 \pm .006$	$.087 \pm .025$	$.044 \pm .015$	$.036 \pm .007$
SAFER	20	$.487 \pm .037$	$.041 \pm .023$	$.129 \pm .047$	$.076 \pm .038$	$.065 \pm .027$	$.031 \pm .014$	$.095 \pm .035$	$.041 \pm .016$	$.034 \pm .011$
NS	0.25	$.357 \pm .004$	$.050 \pm .015$	$.194 \pm .033$	$.123 \pm .026$	$.129 \pm .011$	$.046 \pm .014$	$.147 \pm .027$	$.111 \pm .025$	$.125 \pm .014$
NS	0.5	$.361 \pm .004$	$.048 \pm .014$	$.191 \pm .031$	$.119 \pm .028$	$.127 \pm .010$	$.040 \pm .010$	$.128 \pm .021$	$.102 \pm .024$	$.117 \pm .014$
NS	0.75	$.366 \pm .005$	$.044 \pm .010$	$.183 \pm .028$	$.115 \pm .028$	$.121 \pm .013$	$.038 \pm .009$	$.121 \pm .019$	$.094 \pm .024$	$.113 \pm .015$
NS	1	$.379 \pm .008$	$.044 \pm .011$	$.171 \pm .027$	$.113 \pm .030$	$.114 \pm .011$	$.035 \pm .011$	$.106 \pm .014$	$.080 \pm .015$	$.105 \pm .018$
NS	1.25	$.467 \pm .066$	$.053 \pm .018$	$.172 \pm .039$	$.120 \pm .041$	$.106 \pm .017$	$.037 \pm .013$	$.094 \pm .017$	$.080 \pm .023$	$.102 \pm .029$
EO	5	$.367 \pm .004$	$.048 \pm .013$	$.182 \pm .027$	$.119 \pm .030$	$.125 \pm .011$	$.045 \pm .012$	$.127 \pm .021$	$.099 \pm .018$	$.118 \pm .016$
EO	10	$.381 \pm .006$	$.046 \pm .011$	$.182 \pm .026$	$.114 \pm .027$	$.119 \pm .012$	$.042 \pm .009$	$.108 \pm .015$	$.090 \pm .015$	$.109 \pm .017$
EO	20	$.409 \pm .009$	$.043 \pm .011$	$.171 \pm .021$	$.108 \pm .029$	$.108 \pm .010$	$.039 \pm .008$	$.102 \pm .010$	$.074 \pm .012$	$.099 \pm .018$
EO	50	$.466 \pm .011$	$.040 \pm .011$	$.163 \pm .024$	$.099 \pm .030$	$.098 \pm .010$	$.036 \pm .009$	$.095 \pm .013$	$.064 \pm .012$	$.088 \pm .022$
EO	60	$.479 \pm .010$	$.041 \pm .011$	$.157 \pm .020$	$.099 \pm .030$	$.097 \pm .011$	$.036 \pm .011$	$.096 \pm .014$	$.064 \pm .012$	$.087 \pm .021$

Table A.14: Size sweep: ERM models on a variant of the benchmark with one reference cell ($n = 150{,}000$, true-positive rate 0.85) and six probe cells per design size m, all with planted gap -0.15. For each m, the mean certification statistic $\sqrt{\Lambda_S}$ over the probe cells and runs, computed on hard training predictions of the final ERM model, and the fraction of probe cells certified ($\sqrt{\Lambda_S} > 1.96$) as mean $\pm$ standard deviation over 10 runs. At $m = 5$, two of the 60 cells had fewer than two positives and were skipped.

m	5	10	15	20	30	50	80	120	150
mean $\sqrt{\Lambda_S}$	1.03	1.29	1.55	1.49	1.96	2.15	2.99	3.57	4.11
certified fraction	$.09 \pm .09$	$.23 \pm .16$	$.33 \pm .18$	$.27 \pm .18$	$.52 \pm .15$	$.58 \pm .24$	$.85 \pm .18$	$.95 \pm .08$	$.97 \pm .07$
cells (all runs)	58	60	60	60	60	60	60	60	60

G.2 Per-mechanism detailed numbers

Threshold smoothing. On Adult L4 10/90 (10 splits$\times$3 seeds), reaching $MSEE \approx 10^{-3}$ requires $\lambda = 1000$ at log-loss ≈ 0.57 for NoThreshSmooth, versus $\lambda = 200$ at log-loss ≈ 0.43 for SAFER. At matched fairness, the difference is roughly 0.14 nats. On COMPAS L3 2080, the gap is 0.01–0.02 nats at matched MSEE through the mid-range (Figure 2, center).

Dense tail. On COMPAS L3 2080, at $\lambda \leq 2$ the four arms are within the $\pm$SE bands of one another. From $\lambda = 5$ onward, NoDenseTail plateaus at $MSEE \approx 0.06$; the gap to SAFER widens monotonically and reaches roughly ten standard errors at $\lambda = 100$ (SAFER 0.018 ± 0.001 versus NoDenseTail 0.060 ± 0.004). On Adult L4 1090 the truncation has a different scale but the predicted form: NoDenseTail stays close to ERM MSEE (≈ 0.02) for almost every λ, and only at the largest tested λ does it move toward smaller MSEE, where it is Pareto-dominated by SAFER. On German L3 50/50 NoDenseTail tracks SAFER, because the significance gate already keeps the hinge inactive (Proposition 4).

Gate threshold (NoSelectivity on German). SAFER preserves ERM-level log-loss (≈ 0.53) up to $\lambda = 20$ and degrades only mildly (0.59 at $\lambda = 100$). NoSelectivity climbs from log-loss 0.53 at $\lambda = 1$ through 0.60 at $\lambda = 5$ and reaches 0.69 by $\lambda = 20$; the error bars on its MSEE become very large in this range.

G.3 NoThreshSmooth on German: no departure from ERM

NoThreshSmooth on German exhibits a different and benign failure mode from NoSelectivity: the raw soft-count gradient is too small to move the model on this small dataset, so the optimizer never leaves ERM regardless of λ. This is consistent with Proposition 3: when subgroup distributions are tightly clustered well above (or below) the decision boundary, raw probabilities and hard decisions disagree but the threshold-smoothing step that would expose the hard-decision gap is absent. This safe behavior is therefore a side effect of the small gradient and is not guaranteed by the method.

G.4 Selective-gating: what did not appear

The hypothesis for the selective-gating regime, stated in Appendix F.3, is that the gate threshold gives SAFER an advantage over an ungated alternative when active and inactive subgroups coexist. The Adult L4 1090 and COMPAS L3 2080 results in Figure 2 (left and center) do not exhibit this behavior: NoSelectivity tracks SAFER up to $\lambda = 1$ on Adult, where it then loses utility, and across the sweep on COMPAS. We interpret this as evidence that the gate's empirical role is asymmetric on these two datasets: it suppresses harm in the safety regime, while the evidence for its selective-regime benefit is directional and comes from the controlled benchmark of Appendix F.3, where the evidence strata are isolated.

G.5 Cross-regime ablation map

Table A.15: Empirical effect of removing each component on each stress variant. "≈ 0" indicates indistinguishable from full SAFER within $\pm$SE. "$\Downarrow\Downarrow$" indicates Pareto-dominated by SAFER across the regularization sweep. "$\downarrow$" indicates a Pareto gap that opens at high λ. "harm" indicates that the ablation degrades utility without any certified-fairness benefit. "n/a" indicates the ablation was not run rigorously at the matched protocol.

Ablation	Adult L4 1090	COMPAS L3 2080	German L3 5050
NoThreshSmooth (Theorem 2)	$\Downarrow\Downarrow$	$\Downarrow\Downarrow$	≈ 0 (no traction)
NoDenseTail (Eq. (12))	$\downarrow$ at high λ	$\downarrow$ at high λ	≈ 0
NoSelectivity (Proposition 4)	≈ 0, harm at high λ	≈ 0	harm

The Eq. (12) mechanism predicts that NoDenseTail under-reaches SAFER at high λ on certified-violation datasets, and we observe this on both Adult L4 1090 and COMPAS L3 2080. The shape of the truncation differs: COMPAS shows a clean plateau at $MSEE \approx 0.06$ from $\lambda = 5$ onward, while Adult shows NoDenseTail stuck near ERM MSEE for almost the entire λ sweep, escaping only at the largest tested λ and remaining Pareto-dominated by SAFER (Figure 2, left). Both shapes are consistent with the same hinge-suppression argument; a more controlled comparison of where each dataset's subgroup statistics sit relative to the gate boundary would tighten the quantitative match and is left to future work.

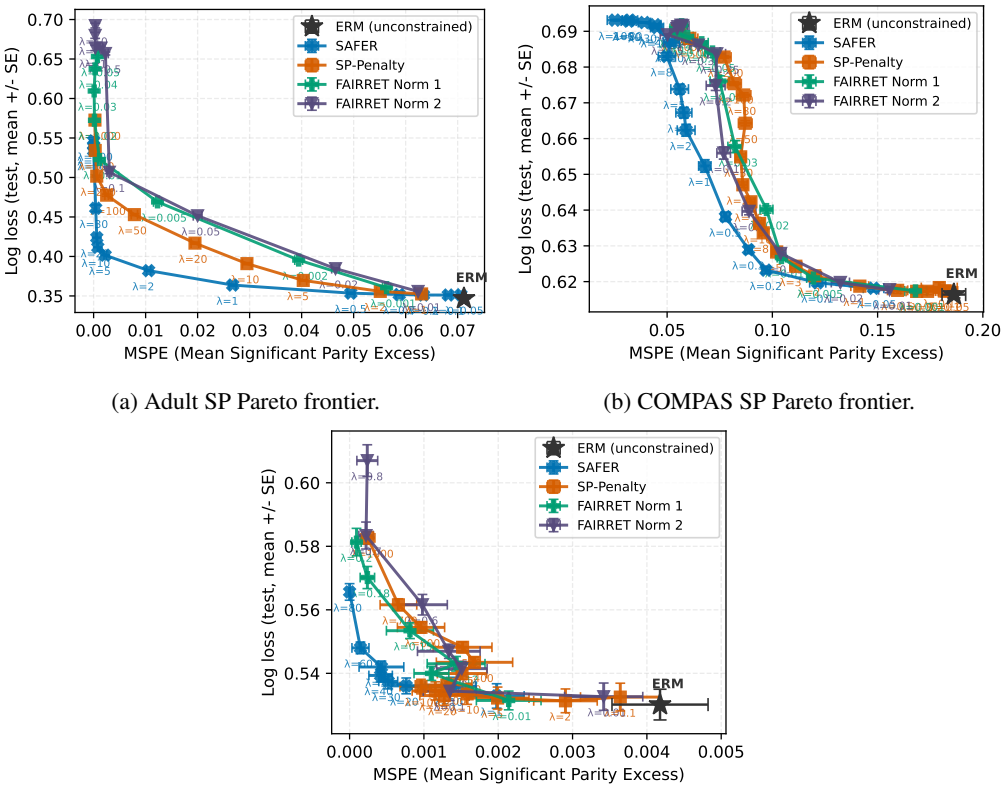
H Statistical Parity results

The main text focuses on EO because its positive-label conditioning makes the sparse-cell issue most severe. SP has larger effective subgroup sizes because it conditions on all examples in a subgroup. Figure A.1 shows the Adult, COMPAS and German SP Pareto frontiers. The penalty baseline, SP-Penalty, is the certainty-blind squared-disparity regularizer applied to the SP gap. Results are consistent with the EO interpretation in Section 5.

I Other instantiations

I.1 Equalized Odds

Section 4 shows that the SAFER framework extends beyond the single-component SP/EO case to multi-component criteria, by summing componentwise hinges. We demonstrate this concretely for



(a) Adult SP Pareto frontier.

(b) COMPAS SP Pareto frontier.

(c) German Credit SP Pareto frontier.

Figure A.1: Pareto frontiers for Adult, COMPAS and German Credit under Statistical Parity.

Equalized Odds [17], which requires simultaneous parity of the true-positive rate (TPR) and the false-positive rate (FPR) across subgroups:

$$EO(S) = P(\hat{Y}{=}1 \mid Y{=}1, S) - P(\hat{Y}{=}1 \mid Y{=}1, \bar{S}), \tag{A.2}$$

$$FP(S) = P(\hat{Y}{=}1 \mid Y{=}0, S) - P(\hat{Y}{=}1 \mid Y{=}0, \bar{S}). \tag{A.3}$$

The TPR component is the Equal Opportunity criterion already treated in the main paper; FPR parity is the complementary condition on negative-label examples.

FPR soft counts. Section 4 defines TPR soft counts by restricting to positive-label examples. The FPR component requires the analogous restriction to negative-label examples:

$$\tilde{n}^{(-)}_{1,S} = \sum_{i:\, y_i=0} \hat{y}_i \, \mathbf{1}_S(i), \qquad \tilde{n}^{(-)}_{0,S} = \sum_{i:\, y_i=0} (1-\hat{y}_i) \, \mathbf{1}_S(i). \tag{A.4}$$

These induce a differentiable FPR estimate $\hat{f}_S = \tilde{n}^{(-)}_{1,S} / \left(\tilde{n}^{(-)}_{1,S} + \tilde{n}^{(-)}_{0,S}\right)$ and, via the identical two-sample proportion construction, a differentiable FPR disparity $\widehat{FP}_S(\theta) = \hat{f}_S - \hat{f}_{\bar{S}}$ and LRT statistic $T^{\mathrm{FPR}}_S(\theta) = \sqrt{\Lambda^{\mathrm{FPR}}_S}$ where Λ^{FPR}_S is the log-likelihood ratio for the null of equal FPR between S and $\bar{S}$.

Combined SAFER objective. The Equalized Odds variant penalizes both components independently:

$$\mathcal{L}^{\mathrm{EqOdds}}_{\mathrm{SAFER}}(\theta) = \mathcal{L}_{\mathrm{pred}}(\theta) + \frac{\lambda}{|\mathfrak{S}|} \sum_{S \in \mathfrak{S}} \left[\mathcal{L}^{\mathrm{TPR}}_{\mathrm{dir},S}(\theta) + \mathcal{L}^{\mathrm{FPR}}_{\mathrm{dir},S}(\theta) \right]. \tag{A.5}$$

Each component uses the same direct significant-excess form as Eq. (11), but with soft counts restricted to the corresponding label stratum. The significance gate and the safety argument apply

componentwise: a subgroup whose TPR and FPR components both lie below the gate threshold has only the exponentially small below-threshold gradient described by Eq. (12), plus any dense-tail residual.

Effective sample sizes. For EqOdds, the binding constraint is the smaller of the two label-conditioned counts:

$$m_S^{\mathrm{EqOdds}} = \min\big(m_S^{(+)},\, m_S^{(-)}\big), \quad m_S^{(+)} = |\{i \in S : y_i = 1\}|, \quad m_S^{(-)} = |\{i \in S : y_i = 0\}|.$$

Whenever $\hat{\rho} \ll 1$ (the typical regime for sparse intersectional subgroups), the positive-label count $m_S^{(+)} \approx n_S \hat{\rho}_S$ is the binding term; when negative labels are rarer, the FPR component is binding instead.

Metric. Specializing Definition 1 to the two components yields the combined certified-excess metric:

$$\mathrm{MSE}_\alpha^{\mathrm{EqOdds}}(\theta) = \frac{1}{|\mathfrak{S}|} \sum_{S \in \mathfrak{S}} \big[\mathcal{E}_\alpha^{\mathrm{TPR}}(S) + \mathcal{E}_\alpha^{\mathrm{FPR}}(S)\big], \tag{A.6}$$

where each component is the significant excess of Definition 1 applied to the respective label stratum. This metric is zero if and only if no subgroup has a certified violation in either the TPR or FPR component, inheriting property 1 of the significant-excess family.

I.2 Disparate Impact instantiation

Disparate Impact (DI) [12] measures the ratio of positive-prediction rates across subgroups:

$$DI(S) = \frac{P(\hat{Y}{=}1 \mid S)}{P(\hat{Y}{=}1 \mid \bar{S})}.$$

We use DI in its *parity* formulation, which tests whether $DI(S) = 1$, rather than the thresholded 0.8-rule used in legal practice. The parity null $DI(S) = 1$ is equivalent to $P(\hat{Y}{=}1 \mid S) = P(\hat{Y}{=}1 \mid \bar{S})$, which is exactly the null of Statistical Parity. Consequently, test statistic and the two-sample LRT statistic for DI parity are identical to the one already defined for SP. Therefore, the SAFER training objective for DI is the same as the SP variant of Eq. (16).

Although the training objective is shared with SP, the significant-excess metric can be expressed in log-ratio units, by defining the log-DI functional $\phi^{DI}(S) = \log DI(S) = \log p_S - \log p_{\bar{S}}$. The significant-excess is derived by following Definition 1 given a confidence interval $[\ell_\alpha(S),\, u_\alpha(S)]$ from a statistical test on $\phi^{DI}(S)$.

J Proofs for Section 4

Throughout this appendix, we write $\sigma_\tau(u) = \sigma(u/\tau)$, where σ is the logistic sigmoid, and $\mathrm{softplus}_\tau(u) = \tau \log(1 + e^{u/\tau})$.

J.1 Basic sigmoid bounds

Lemma 5 (Sigmoid–step difference)**.** *Let σ be the logistic sigmoid and $\sigma_\tau(u) = \sigma(u/\tau)$. For every $u \in \mathbb{R}$ and $\tau > 0$,*

$$\big|\sigma_\tau(u) - \mathbf{1}\{u > 0\}\big| = \sigma(-|u|/\tau).$$

Proof. If $u > 0$, the difference is $1 - \sigma(u/\tau) = \sigma(-u/\tau)$. If $u < 0$, it is $\sigma(u/\tau) = \sigma(-|u|/\tau)$. At $u = 0$, both sides are $1/2$. □

Lemma 6 (Integral and derivative bounds)**.** *For every $\tau > 0$,*

$$\int_{-\infty}^{\infty} \sigma(-|v|/\tau)\, dv = 2\tau \log 2, \qquad \sigma'(z) \le e^{-|z|} \quad \forall z \in \mathbb{R}.$$

Proof. The integral follows by the change of variables $w = v/\tau$:

$$\int_{-\infty}^{\infty} \sigma(-|v|/\tau)\, dv = 2\tau \int_0^{\infty} \frac{e^{-w}}{1+e^{-w}}\, dw = 2\tau \log 2.$$

For the derivative bound, use $\sigma'(z) = \sigma(z)(1-\sigma(z))$. If $z \geq 0$, then $1 - \sigma(z) = \sigma(-z) \leq e^{-z}$. Symmetry gives the case $z < 0$. □

J.2 Proof of Theorem 2

Let F_S be the conditional distribution of $f_\theta(X)$ on the conditioning population Π_S, and define

$$p_S^\tau = \mathbb{E}_{F_S}\left[\sigma_\tau(f_\theta(X) - \tfrac{1}{2})\right], \qquad p_S^{\mathrm{hard}} = \mathbb{E}_{F_S}\left[\mathbf{1}\{f_\theta(X) > \tfrac{1}{2}\}\right].$$

By Lemma 5,

$$|p_S^\tau - p_S^{\mathrm{hard}}| \leq \int \sigma(-|u - \tfrac{1}{2}|/\tau)\, dF_S(u).$$

Let $N = \{u : |u - \frac{1}{2}| \leq \delta\}$ be the interval on which, by assumption, F_S has a density bounded by L, so that $dF_S(u) \leq L\, du$ on N. We split the integral at N. On N, substituting $v = u - \frac{1}{2}$, extending the range of integration to the real line and using Lemma 6,

$$\int_N \sigma(-|u - \tfrac{1}{2}|/\tau)\, dF_S(u) \leq L \int_{-\infty}^{\infty} \sigma(-|v|/\tau)\, dv = 2L\tau \log 2.$$

Outside N we have $|u - \frac{1}{2}| > \delta$, so $\sigma(-|u - \frac{1}{2}|/\tau) \leq \sigma(-\delta/\tau) \leq e^{-\delta/\tau}$, and the mass of F_S outside N is at most one; hence

$$\int_{N^c} \sigma(-|u - \tfrac{1}{2}|/\tau)\, dF_S(u) \leq e^{-\delta/\tau}.$$

Adding the two parts,

$$|p_S^\tau - p_S^{\mathrm{hard}}| \leq 2L\tau \log 2 + e^{-\delta/\tau}.$$

The same bound holds for $\bar{S}$. Since $\phi_S^\bullet = p_S^\bullet - p_{\bar{S}}^\bullet$, the triangle inequality gives:

$$|\phi_S^\tau - \phi_S^{\mathrm{hard}}| \leq |p_S^\tau - p_S^{\mathrm{hard}}| + |p_{\bar{S}}^\tau - p_{\bar{S}}^{\mathrm{hard}}| \leq 4L\tau \log 2 + 2e^{-\delta/\tau}.$$

The remainder is also linear in τ: since $e^{-x} \leq 1/(ex)$ for $x > 0$, we have $2e^{-\delta/\tau} \leq 2\tau/(e\delta)$, so the whole bound is at most $(4L \log 2 + 2/(e\delta))\, \tau$. Outside N the score distribution is arbitrary; in particular it may have atoms at 0 and 1. □

J.3 Proof of Proposition 3

Fix $\eta > 0$. Choose $\kappa < \eta/2$ and let F_S concentrate smoothly in a small interval around $1/2 + \kappa$, while $F_{\bar{S}}$ concentrates smoothly in a small interval around $1/2 - \kappa$. Then

$$\phi_S^{\mathrm{raw}} \approx 2\kappa, \qquad \phi_S^{\mathrm{hard}} \approx 1,$$

so

$$|\phi_S^{\mathrm{raw}} - \phi_S^{\mathrm{hard}}| \geq 1 - \eta$$

for sufficiently narrow intervals. The distributions may be chosen with continuous densities bounded by a finite constant.

To see that no constant rescaling works uniformly, compare this construction with a second one in which both groups are well above the threshold, with mean scores differing by $\Delta > 0$. Then $\phi_S^{\mathrm{raw}} \approx \Delta$ while $\phi_S^{\mathrm{hard}} = 0$. Matching the first construction as $\kappa \to 0$ requires a rescaling constant of order $1/(2\kappa)$, which then makes the rescaled raw disparity in the second construction diverge instead of matching zero. Hence a single constant cannot align raw soft counts with hard-decision disparities across both regimes. □

J.4 Gradient localization at the decision boundary

The following result, deferred from Section 4.1, shows that threshold smoothing localizes the fairness gradient to examples near the decision boundary $\{x : f_\theta(x) = \frac{1}{2}\}$.

Theorem 7 (Gradient localization)**.** *For SP and EO,*

$$\left|\frac{\partial \phi_S^\tau}{\partial f_\theta(x)}\right| \le \frac{c_S(x)}{\tau}\,\exp\!\left(-\frac{|f_\theta(x)-\frac{1}{2}|}{\tau}\right), \tag{A.7}$$

for a bounded subgroup-membership weight $c_S(x)$. As $\tau \to 0^+$, the gradient concentrates on $\{x : f_\theta(x) = \frac{1}{2}\}$.

Proof. For $x \in \Pi_S$,

$$\frac{\partial p_S^\tau}{\partial f_\theta(x)} = \frac{1}{\tau}\sigma'\big((f_\theta(x)-\tfrac{1}{2})/\tau\big)\, c_S^+(x),$$

where $c_S^+(x)$ is the appropriate conditioning-population weight. The corresponding complement term has the same form with a possibly different weight. Therefore, for a bounded subgroup-membership weight $c_S(x)$,

$$\left|\frac{\partial \phi_S^\tau}{\partial f_\theta(x)}\right| \le \frac{c_S(x)}{\tau}\sigma'\big((f_\theta(x)-\tfrac{1}{2})/\tau\big) \le \frac{c_S(x)}{\tau}e^{-|f_\theta(x)-1/2|/\tau},$$

using Lemma 6. Finally, $\tau^{-1}\sigma'((u-1/2)/\tau)$ is an approximate identity that converges weakly to a Dirac mass at $u = 1/2$, yielding the distributional limit. □

J.5 Canonical surrogate approximation

The main text (Eq. (11)) uses an LRT-based threshold. Here we state and prove a metric-alignment bound for a canonical version in which the surrogate and the metric share the population Wald threshold B_S. The bound shows that the surrogate *shape* tracks $MSEE_\alpha^{\mathrm{pop}}$, defined below. This population comparison does not bound the discrepancy between the implemented LRT-based training loss and the SAFT audit metric.

Let $B_S(\theta) = z_\alpha\, SE_S^\phi(\theta)$, where SE_S^ϕ is evaluated with population cell probabilities. Define

$$MSEE_\alpha^{\mathrm{pop}}(\theta) = \frac{1}{|\mathfrak{S}|}\sum_{S\in\mathfrak{S}}\big[|\phi_S^{\mathrm{hard}}(\theta)| - B_S(\theta)\big]_+, \tag{A.8}$$

$$\mathcal{L}_{\mathrm{can}}(\theta) = \frac{1}{|\mathfrak{S}|}\sum_{S\in\mathfrak{S}}\mathrm{softplus}_{\tau_s}\left(\sqrt{\phi_S^\tau(\theta)^2+\varepsilon^2} - B_S(\theta)\right). \tag{A.9}$$

Theorem 8 (Canonical surrogate approximation)**.** *Under Theorem 2's density assumption, with the same L and δ, for every $\varepsilon, \tau_s, \tau > 0$,*

$$\big|\mathcal{L}_{\mathrm{can}}(\theta) - MSEE_\alpha^{\mathrm{pop}}(\theta)\big| \le \varepsilon + \tau_s\log 2 + 4L\tau\log 2 + 2e^{-\delta/\tau}. \tag{A.10}$$

Proof. Fix a subgroup S, write $B_S = B_S(\theta)$, and define

$$h_S = \mathrm{softplus}_{\tau_s}\left(\sqrt{(\phi_S^\tau)^2+\varepsilon^2} - B_S\right), \qquad m_S = \big[|\phi_S^{\mathrm{hard}}| - B_S\big]_+.$$

The elementary inequalities

$$\big|\sqrt{x^2+\varepsilon^2} - |x|\big| \le \varepsilon, \qquad 0 \le \mathrm{softplus}_{\tau_s}(u) - [u]_+ \le \tau_s\log 2,$$

and the Lipschitz property of $x \mapsto [x]_+$ imply

$$|h_S - m_S| \le \varepsilon + \tau_s\log 2 + \big||\phi_S^\tau| - |\phi_S^{\mathrm{hard}}|\big|.$$

Since $\big||a|-|b|\big| \le |a-b|$, Theorem 2 gives

$$|h_S - m_S| \le \varepsilon + \tau_s\log 2 + 4L\tau\log 2 + 2e^{-\delta/\tau}.$$

Averaging this bound over $S \in \mathfrak{S}$ proves Eq. (A.10). □

J.6 Asymptotic equivalence of the soft-count LRT and the Wald statistic

Write $Q_i = \sigma_\tau(f_\theta(x_i) - \frac{1}{2}) \in (0,1)$ for the soft count of observation i, so that $\hat{p}_S$ and $\hat{p}_{\bar{S}}$ are the means of the Q_i over Π_S and $\Pi_{\bar{S}}$, $\hat{p}$ is the pooled mean, and $\tilde{n}_S = |\Pi_S|$, $\tilde{n}_{\bar{S}} = |\Pi_{\bar{S}}|$. With $D(a \,\|\, b) = a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b}$ the Bernoulli divergence, Eq. (9) reads

$$\Lambda_S = 2\big[\tilde{n}_S \, D(\hat{p}_S \,\|\, \hat{p}) + \tilde{n}_{\bar{S}} \, D(\hat{p}_{\bar{S}} \,\|\, \hat{p})\big].$$

Set $\Delta = \hat{p}_S - \hat{p}_{\bar{S}}$ and $N = \tilde{n}_S + \tilde{n}_{\bar{S}}$, so that $\hat{p}_S - \hat{p} = (\tilde{n}_{\bar{S}}/N)\,\Delta$ and $\hat{p}_{\bar{S}} - \hat{p} = -(\tilde{n}_S/N)\,\Delta$. Since $D(a \,\|\, b) = \frac{(a-b)^2}{2b(1-b)} + O(|a-b|^3)$ for b bounded away from 0 and 1,

$$\Lambda_S = \frac{\Delta^2}{\hat{p}(1-\hat{p})\,(1/\tilde{n}_S + 1/\tilde{n}_{\bar{S}})}\,\big(1 + O(|\Delta|)\big).$$

The plug-in Wald variance of property (ii) satisfies $(\widehat{SE}_S^{\,\phi})^2 = \hat{p}(1-\hat{p})\,(1/\tilde{n}_S + 1/\tilde{n}_{\bar{S}})\,(1 + O(|\Delta|))$, because $\hat{p}_S(1-\hat{p}_S)$ and $\hat{p}_{\bar{S}}(1-\hat{p}_{\bar{S}})$ differ from $\hat{p}(1-\hat{p})$ by $O(|\Delta|)$. Under the null and under local alternatives, where disparities are of the order of their standard error, $\Delta \to 0$ in probability; this is the range in which the gate operates. There $\Lambda_S = Z_S^2\,(1 + o_p(1))$, and for nonzero Δ Eq. (10) gives $\widehat{SE}_{\mathrm{eff},S} = \widehat{SE}_S^{\,\phi}\,(1 + o_p(1))$.

Null law of the soft-count statistic. Fix $\tau > 0$ and the predictor independently of the sample. Assume independent observations within and across groups, identically distributed within each group, with a common soft mean $p \in (0,1)$ under the null and $\tilde{n}_S/N \to \rho \in (0,1)$. For fractional counts the Bernoulli form overstates the variance: in each group $\mathrm{Var}(Q_i) = p(1-p) - \mathbb{E}[Q_i(1-Q_i)]$, and the central limit theorem applied to Δ in the expansion above gives $\Lambda_S \Rightarrow c\,\chi_1^2$ with

$$0 \le c = \frac{(1-\rho)\,v_S + \rho\,v_{\bar{S}}}{p(1-p)} \le 1, \qquad v_S = \mathrm{Var}(Q_i \mid i \in \Pi_S), \quad v_{\bar{S}} = \mathrm{Var}(Q_i \mid i \in \Pi_{\bar{S}}).$$

The sampling standard error of the soft mean contrast is $\sqrt{v_S/\tilde{n}_S + v_{\bar{S}}/\tilde{n}_{\bar{S}}}$; the Bernoulli-form denominator is a proxy for it. For $c > 0$, the reference gate opens under soft parity with asymptotic probability $P(\chi_1^2 > z_\alpha^2/c) \le \alpha$. The same factor c scales Z_S^2, preserving the local equivalence of the two statistics. For $c > 0$, division by a consistent estimate of c would give asymptotic nominal size under these same fixed-predictor, soft-null assumptions; SAFER retains the Bernoulli reference cutoff.

For exact hard indicators $\mathbf{1}\{f_\theta(X) > 1/2\}$, the corresponding deviance instead has Wilks' χ_1^2 null limit when the common hard positive rate is interior. To relate this to soft counts as $\tau \downarrow 0$, assume that neither conditional score distribution has mass at $1/2$ and that the limiting hard rates are interior. Bounded convergence then gives $\mathrm{Var}(Q_i)/[(\mathbb{E}Q_i)(1-\mathbb{E}Q_i)] \to 1$ in each group, and hence $c \to 1$ along models satisfying soft parity.

At fixed τ, equal hard positive rates need not give equal soft means, so the null of the gate is soft parity, and certification of hard-decision parity is the task of the held-out hard-count audit of Section 3. The fixed-predictor result above applies when the statistic is evaluated on data independent of the fit. During training the sample is also used to fit f_θ, and there the gate is an optimization device that uses the audit's reference cutoff.

J.7 Gradient identities for Section 4.3

Recall from Section 4.3 that $u_S(\theta) = \sqrt{\widehat{\phi}_S^{\tau}(\theta)^2 + \varepsilon^2} - z_\alpha \,\mathrm{sg}[\widehat{SE}_{\mathrm{eff},S}(\theta)]$, where $\mathrm{sg}[\cdot]$ denotes stop-gradient. Since the stop-gradient makes the threshold a constant for differentiation, $\nabla_\theta u_S = \nabla_\theta \sqrt{\widehat{\phi}_S^{\tau}(\theta)^2 + \varepsilon^2}$. By the chain rule and $\mathrm{softplus}'_{\tau_s}(u) = \sigma(u/\tau_s)$,

$$\partial_\theta \,\mathrm{softplus}_{\tau_s}(u_S(\theta)) = \sigma(u_S/\tau_s)\,\partial_\theta u_S(\theta).$$

Taking norms gives Eq. (12). If $u_S \le 0$, then $\sigma(u_S/\tau_s) \le 1/2$; if $u_S < 0$, then $\sigma(u_S/\tau_s) \le \exp(u_S/\tau_s)$.

The dense-tail contribution for subgroup S is $w_{\mathrm{tail}}\sqrt{\Lambda_S(\theta)}$. Differentiating gives $w_{\mathrm{tail}}\partial_\theta\sqrt{\Lambda_S(\theta)}$. This expression contains no factor $\sigma(u_S/\tau_s)$, because the LRT statistic is computed from the soft counts directly and does not pass through the significant-excess softplus.

Size of the suppression. With $\tau_s = 10^{-2}$, the multiplier $\sigma(u_S/\tau_s)$ equals $1/2$ at the gate boundary $u_S = 0$, 0.0067 at $u_S = -0.05$, and 0.9933 at $u_S = 0.05$. The suppression below the boundary is the intended safety mechanism, but it also acts on active subgroups that strong regularization pushes back toward the boundary, so the high-regularization frontier can stall; the dense tail supplies a weak gradient path that does not contain this multiplier.

Why $\sqrt{\Lambda_S}$. Three candidates have the required descent direction. An ungated raw-disparity term $|\widehat{\phi}^{\tau}_S|$ would be certainty-blind and would penalize noisy disparities regardless of their statistical support. The statistic Λ_S itself would place the tail on a squared evidence scale and produce much steeper precision-dependent gradients. The intermediate choice $\sqrt{\Lambda_S}$ remains on the same standardized evidence scale as the gate and satisfies $\sqrt{\Lambda_S} = (|\widehat{\phi}^{\tau}_S|/\widehat{SE}^{\phi}_S)(1 + o_p(1))$ near the gate boundary (property (iii) in Section 4.2), so it retains the uncertainty weighting without introducing a new estimator or threshold; its strength is governed by the single weight w_{tail}.

Effect on the safety guarantee. Because the tail is ungated, Eq. (15) carries an additive residual proportional to w_{tail}, whose contribution to the complete objective scales as $\lambda w_{\mathrm{tail}}$; only the gated component retains the exponentially small gradient guarantee below the gate boundary. Setting $w_{\mathrm{tail}} = 0$ removes the residual and recovers the gated-only objective. All full-SAFER runs use the fixed value $w_{\mathrm{tail}} = 0.003$; the NoDenseTail ablation sets it to zero. The proposition bounds the back-propagated fairness contribution at one point. How far training moves from the ERM optimum depends also on the curvature of the prediction loss there: for a scalar quadratic prediction loss with curvature κ, a constant residual gradient of size r shifts the minimizer by r/κ. The hypothesis $u_S \le -\epsilon^\star$ is a margin condition: an empty active set alone gives only the multiplier bound $\sigma(u_S/\tau_s) \le 1/2$, and the exponential bound needs $\epsilon^\star$ large relative to τ_s.

J.8 Operating regimes: the gate boundary

The regimes describe the state of the gate at a checkpoint. The implemented boundary is $u_S = 0$; in the regular unclamped case with nonzero disparity the active side $u_S > 0$ is $T_S = \sqrt{\Lambda_S} > z_\alpha$. Softplus makes the transition between the gate states continuous through the multiplier $\sigma(u_S/\tau_s)$ of Eq. (12), whose values at $u_S = \pm 0.05$ are given above and give a concrete meaning to "well below" and "well above" the boundary. There is no universal sample-size boundary. For EO the local approximation

$$T_S \approx \frac{|\phi_S|}{\sqrt{p_S(1-p_S)/m_S + p_{\bar{S}}(1-p_{\bar{S}})/m_{\bar{S}}}}$$

shows that certification depends jointly on the disparity magnitude, the outcome rates, and the positive-label counts m_S, $m_{\bar{S}}$ on both sides; for SP the total subgroup counts n_S, $n_{\bar{S}}$ take their place.

J.9 Stop-gradient and the threshold across updates

Write $d_S(\theta) = \sqrt{\widehat{\phi}^{\tau}_S(\theta)^2 + \varepsilon^2}$ for the smoothed disparity and $b_S(\theta) = z_\alpha \widehat{SE}_{\mathrm{eff},S}(\theta)$ for the threshold. At iteration k, SAFER differentiates $\ell_S(\theta) = \mathrm{softplus}_{\tau_s}\big(d_S(\theta) - \mathrm{sg}[b_S(\theta)]\big)$, so that

$$\nabla_\theta \ell_S(\theta_k) = \sigma\left(\frac{d_{S,k} - b_{S,k}}{\tau_s}\right) \nabla_\theta d_S(\theta_k).$$

The threshold $b_{S,k}$ is held fixed during that backward pass only. After the parameters are updated, $b_{S,k+1} = b_S(\theta_{k+1})$ is recomputed from the new predictions and can increase or decrease.

A scalar example. Let $d(\theta) = \theta$, $b(\theta) = 2\theta$ and $\theta_k = 1$, and absorb the positive softplus factor and the learning rate into an effective step of 0.1; the example sets the gate factor aside and shows only the direction of the two gradient routes. With stop-gradient, the direct term differentiates only d, giving $\theta_{k+1} = 0.9$; on the next forward pass the threshold is recomputed and changes from 2 to 1.8. Without stop-gradient, the derivative of the excess is $d'(\theta) - b'(\theta) = -1$, so descent instead gives $\theta_{k+1} = 1.1$: the disparity increases from 1 to 1.1, but the threshold rises faster, from 2 to 2.2, and the excess decreases. Detachment removes this route to a lower penalty.

Threshold drift in practice. We logged the unweighted mean of b_S over the subgroups eligible for the EO loss (at least two positive-label examples in both S and $\bar{S}$: 156 subgroups on Adult and 35 on COMPAS) together with the complete fairness loss. Table A.16 reports means $\pm$ standard deviations over 10 Adult and 5 COMPAS runs; "final" is epoch 199 on Adult and each seed's early-stopped epoch on COMPAS. Near initialization the scores are approximately constant, so disparities and the fairness loss are small, and both grow as the prediction loss learns predictive structure. The mean threshold moves from 0.281 to 0.269 on Adult and from 0.128 to 0.110 on COMPAS. These are averages over subgroups, and individual thresholds can rise; the guarantee itself concerns one update, in which the gated term supplies no gradient through b_S.

Table A.16: Mean gate threshold b_S over eligible subgroups and complete fairness loss at the start and end of training (mean $\pm$ standard deviation over runs).

dataset (runs)	mean b_S: initial / final	fairness loss: initial / final
Adult (10)	$.281 \pm .032$ / $.269 \pm .014$	$.0015 \pm .0007$ / $.0037 \pm .0011$
COMPAS (5)	$.128 \pm .021$ / $.110 \pm .003$	$.0034 \pm .0045$ / $.0205 \pm .0023$

J.10 Proof of Proposition 4

At $\theta^\star_{\mathrm{ERM}}$, assume $u_S(\theta^\star_{\mathrm{ERM}}) \le -\epsilon^\star$ for all S. By Eq. (12),

$$\left\|\partial_\theta \operatorname{softplus}_{\tau_s}(u_S(\theta^\star_{\mathrm{ERM}}))\right\| \le e^{-\epsilon^\star/\tau_s}\left\|\partial_\theta u_S(\theta^\star_{\mathrm{ERM}})\right\|.$$

The bounded-gradient regularity condition in Proposition 4 allows the finite average of $\|\partial_\theta u_S\|$ to be absorbed into the constant. Summing over subgroups gives Eq. (14). Adding the dense-tail gradients and applying the triangle inequality gives Eq. (15). □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction are directly supported by the theoretical results and experimental evaluation presented in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in the conclusion, including scope of evaluation and open directions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theorems and propositions are numbered with assumptions stated explicitly; complete proofs are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper and appendix provide full details on datasets, preprocessing, model architecture, optimizer, hyperparameter grids, split protocols, and evaluation metrics.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All datasets are publicly available benchmarks. Code is publicly available at `https://github.com/Ambress92/Safer`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The experimental section and appendix specify architecture, optimizer, hyperparameter grids, and data splits.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All results report ± 1 standard error over multiple splits and random seeds, with the source of variability stated.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Yes, all necessary details are reported in the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics. It uses only publicly available datasets and addresses algorithmic fairness.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper is directly concerned with positive societal impact through improving fairness in machine learning. Limitations that may affect real-world deployment are discussed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release pretrained models or scraped datasets that carry misuse risk.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets used are publicly available and cited with their original publications.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce new datasets or pretrained models as released assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human subjects research.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core methodology does not involve LLMs.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.