RAWLS–SHAPLEY FAIRNESS: A NASH-BARGAINING APPROACH TO EQUITABLE MACHINE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose *Rawls–Shapley* (RS) learning, a machine-learning fairness objective grounded in Yaari's (1981) formal result that applying Rawls' Difference Principle through Shapley's utility-transfer framework uniquely recovers the Nash bargaining solution. Translating this result into a training objective, we define each group's utility gain and aggregate these gains through a Nash-product formulation. This gives rise to the Rawls–Shapley objective, whose optimization automatically places greater emphasis on groups with lower utility gains, thereby incorporating Rawlsian priority into the learning process in a principled and differentiable manner. We instantiate RS in two settings: **RS-Known**, which uses observed group labels, and **RS-Proxy**, which trains a supervised proxy predictor on group labels to derive soft group memberships (a "limited demographics" regime, distinct from the fully label-free setting). We evaluate both variants on Adult income and COMPAS recidivism benchmarks alongside ERM, Adversarially Reweighted Learning (ARL), Logistic Regression, and XGBoost, reporting accuracy, AUC, and equalized-odds gap across protected groups. Empirically, RS-Proxy achieves the lowest equalized-odds gap among the RS and non-parity neural objectives on Adult (0.145), while EqOdds-Reg achieves the best parity across both datasets; The two RSvariants achieve broadly similar worst-group accuracy. On COMPAS, where group labels are noisy and less predictable, all methods converge to similar performance, consistent with our graceful-degradation analysis.

1 INTRODUCTION

Machine learning systems deployed in consequential domains such as credit scoring, criminal risk assessment, and medical triage can exhibit systematic performance disparities across demographic groups Barocas & Selbst (2016); Buolamwini & Gebru (2018); Angwin et al. (2016). A large body of fairness research addresses these disparities through statistical parity criteria, including demographic parity Zafar et al. (2017) and equalized odds Hardt et al. (2016). Although such constraints provide clear notions of parity, reducing disparities does not necessarily imply improving the outcomes of the worst-performing groups, parity may also be attained by reducing performance for groups that already perform well.

A complementary line of work emphasizes worst-group performance. Motivated by Rawls' Difference Principle Rawls (1971), max-min formulations seek to maximize $\min_s U_s(\theta)$ and thereby prioritize the least advantaged group. Related ideas appear in distributionally robust optimization (DRO) Hashimoto et al. (2018) and Adversarially Reweighted Learning (ARL) Lahoti et al. (2020), which place greater emphasis on high-loss subpopulations. Strict max-min optimization, however, can be highly concentrated: when one group uniquely attains the minimum utility, the local objective is governed primarily by that group, while improvements to other groups receive no marginal value until the identity of the worst group changes.

We consider an alternative based on Nash welfare, motivated by Yaari's connection between Rawlsian fairness, Shapley bargaining, and the Nash solution Yaari (1981). Let $G_s(\theta) = U_s(\theta) - d_s$ denote the differentiable utility gain of group s relative to disagreement level d_s. We optimize

$$\mathcal{L}_{\mathrm{RS}}^{\varepsilon}(\theta) = -\sum_{s=1}^{K} \log\big(G_s(\theta) + \varepsilon\big), \tag{1}$$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

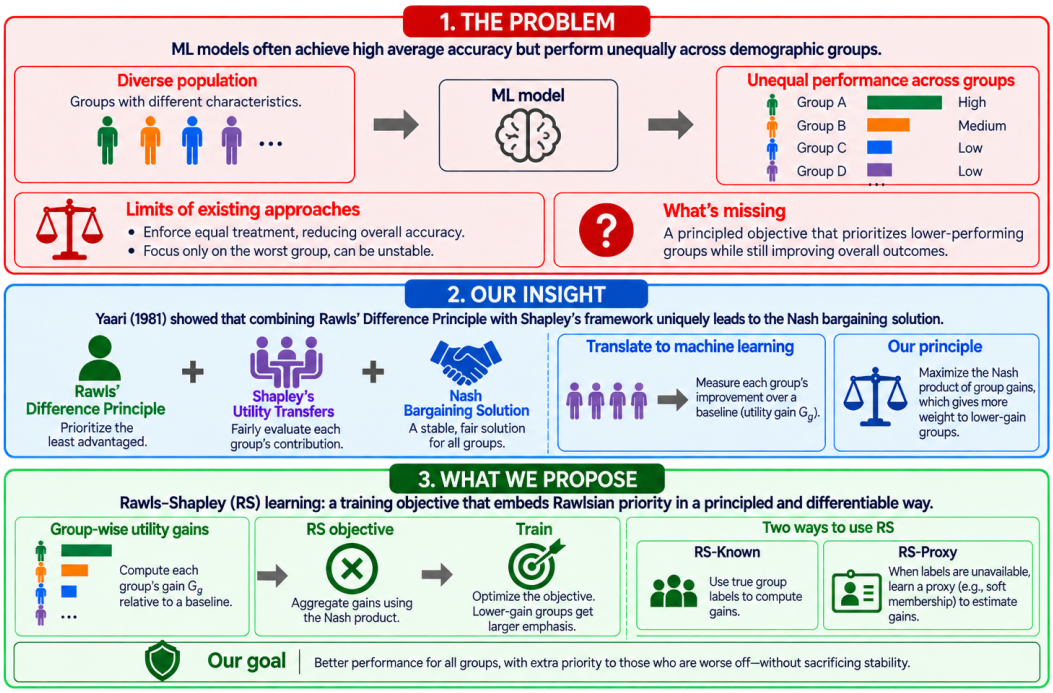


Figure 1: Overview of the proposed Rawls–Shapley fairness framework.

which induces a marginal coefficient proportional to $1/(G_s + \varepsilon)$ for each group. Consequently, lower-gain groups receive greater emphasis while all groups continue to influence the optimization. In our experiments, $d_s = 0$, so $G_s = U_s \in [0, 1]$. Under the convex utility-profile assumptions developed later, the exact Nash-welfare solution admits the supporting-hyperplane and equal-weight-gain characterization associated with Yaari's result (Figure 1).

Our contributions are:

1. We formulate group fairness directly in utility space through a Nash-welfare objective that gives larger marginal coefficients to lower-gain groups without reducing to strict max-min optimization.
2. We consider two demographic-access regimes: RS-Known uses observed group labels during training, whereas RS-Proxy uses soft memberships from a supervised proxy model when demographic information is limited.
3. We characterize the RS solution in convex utility-profile space, derive its inverse-gain gradient weighting, and establish a Nash-product certificate for worst-group utility gain.
4. We compare nine methods on two real-world datasets using predictive performance, group-wise AUC, and equalized-odds metrics.

2 Rawls–Shapley Fairness

2.1 From Distributive Justice to Nash Welfare

Let F be a feasible choice set and $u_s : F \to \mathbb{R}$ the utility of party $s \in \{1, \ldots, K\}$. Each $x \in F$ induces a utility profile $u(x) = (u_1(x), \ldots, u_K(x))$, with feasible profile set $U = \{u(x) : x \in F\} \subseteq \mathbb{R}^K$. Yaari's Principle of Non-Polarisation assumes that U is convex Yaari (1981). Rawls' Difference Principle instead emphasizes the least advantaged party, leading to the max-min rule $x^{MM} \in \arg\max_{x \in F} \min_s u_s(x)$ Rawls (1971).

Yaari combines this Rawlsian perspective with Shapley's utility-transfer framework Shapley (1969). Let $d = (d_1, \ldots, d_K)$ denote a disagreement profile and define the utility gain $G_s(x) = u_s(x) - d_s$.

A feasible choice x^* is Rawls–Shapley relative to d if there exists $\boldsymbol{\lambda}^* \in \mathrm{int}(\Delta^{K-1})$ such that

$$\sum_{s=1}^{K} \lambda_s^* G_s(x^*) \geq \sum_{s=1}^{K} \lambda_s^* G_s(x), \qquad \forall x \in F_d, \tag{2}$$

$$\lambda_1^* G_1(x^*) = \cdots = \lambda_K^* G_K(x^*), \tag{3}$$

where $F_d = \{x \in F : u_s(x) \geq d_s,\ \forall s\}$. The first condition selects a Pareto-efficient gain profile through a supporting hyperplane, while the second equalizes weighted gains.

Theorem 1 (Yaari 1981)**.** *Let $\mathcal{G} = \{(u_1(x) - d_1, \ldots, u_K(x) - d_K) : x \in F_d\}$ be nonempty, compact, and convex, and suppose it contains a strictly positive profile. Then x^* is Rawls–Shapley relative to d if and only if*

$$x^* \in \arg \max_{\substack{x \in F \\ u_s(x) \geq d_s,\ \forall s}} \prod_{s=1}^{K} \big(u_s(x) - d_s\big). \tag{4}$$

The maximizing strictly positive gain profile is unique, although multiple feasible choices may induce the same profile.

Thus, Rawls–Shapley fairness admits a Nash-welfare representation over utility gains, providing the basis for our learning objective.

2.2 From Nash Welfare to Group Utility

We consider binary classification with $\mathcal{D} = \{(x_i, y_i, s_i)\}_{i=1}^n$, where $y_i \in \{0, 1\}$ and $s_i \in \{1, \ldots, K\}$ denotes group membership. A classifier $h_\theta : \mathcal{X} \to \mathbb{R}$ produces $p_i = \sigma(h_\theta(x_i))$. We define the differentiable soft-correctness score $c_i(\theta) = y_i p_i + (1-y_i)(1-p_i)$, which equals $1 - |p_i - y_i|$ for binary labels. The corresponding group utility is

$$U_s(\theta) = \frac{1}{|\mathcal{D}_s|} \sum_{i \in \mathcal{D}_s} c_i(\theta), \qquad \mathcal{D}_s = \{i : s_i = s\}. \tag{5}$$

Hence $U_s \in [0, 1]$, with larger values indicating better group-level performance, while remaining differentiable through the classifier.

The utility scale is consequential. In particular, choosing $U_s = \exp(-L_s)$ for group loss L_s gives $-\sum_s \log U_s = \sum_s L_s$, reducing Nash welfare to ordinary group-balanced loss minimization. We therefore define welfare directly in terms of soft group utility.

2.3 RS Objective and Implicit Welfare Weighting

Translating Yaari's gain formulation to the hypothesis space, let $G_s(\theta) = U_s(\theta) - d_s$. The exact Nash-welfare objective is $\mathcal{L}_{\mathrm{RS}}^0(\theta) = -\sum_{s=1}^K \log G_s(\theta)$ for $G_s(\theta) > 0$. For numerical stability, we optimize

$$\mathcal{L}_{\mathrm{RS}}^{\varepsilon}(\theta) = -\sum_{s=1}^{K} \log\big(G_s(\theta) + \varepsilon\big), \tag{6}$$

where $\varepsilon > 0$ is a numerical stabilizer. Minimizing equation 6 maximizes $\prod_s (G_s + \varepsilon)$, with the exact Nash product recovered at $\varepsilon = 0$. In our experiments, $d_s = 0$, so $G_s = U_s$.

Differentiating equation 6 yields

$$\nabla_\theta \mathcal{L}_{\mathrm{RS}}^{\varepsilon} = -\sum_{s=1}^{K} \frac{1}{G_s(\theta) + \varepsilon} \nabla_\theta U_s(\theta), \tag{7}$$

since d_s is constant. Thus, each group receives an implicit marginal coefficient proportional to $1/(G_s + \varepsilon)$: lower-gain groups receive greater emphasis, while all groups continue to contribute to the update. This distinguishes RS from strict max-min optimization, which assigns marginal importance only to the currently worst-off group. RS instead implements a smooth Nash-welfare tradeoff that preserves priority toward lower-utility groups without discarding improvements elsewhere.

Relation to existing fairness objectives. Distributionally robust optimization (DRO) emphasizes worst-case subpopulations Hashimoto et al. (2018), while Adversarially Reweighted Learning (ARL) identifies and up-weights high-loss regions Lahoti et al. (2020). In contrast, RS follows directly from the Rawls–Shapley Nash-welfare characterization and operates on differentiable group utility gains rather than directly reweighting losses.

3 Theoretical Properties

We establish three properties of the RS objective. First, under compactness and convexity of the feasible utility-gain set, the exact RS objective has a unique Nash-welfare solution in utility-profile space that satisfies the Rawls–Shapley supporting-hyperplane and equal-weighted-gain conditions. Second, the Nash product yields a lower certificate on worst-group utility gain, while remaining distinct from strict max-min fairness. Third, we clarify the weaker guarantees available in neural-network parameter space.

These distinctions are important because Yaari's characterization assumes a convex feasible utility set, whereas neural-network parameterizations generally induce nonconvex optimization problems. The global characterization applies in utility-profile space under the stated compactness and convexity assumptions. Neural-network training, however, is generally nonconvex; therefore, convergence to a stationary parameter vector does not guarantee recovery of the globally optimal Rawls–Shapley utility profile.

3.1 Utility-Profile Characterization

Let $\mathcal{G} = \{(G_1(h), \ldots, G_K(h)) : h \in \mathcal{H}\} \subset \mathbb{R}^K_{++}$ denote the feasible set of group utility-gain profiles. For $\varepsilon = 0$, minimizing the RS objective is equivalent to maximizing $\sum_{s=1}^K \log g_s$, or equivalently the Nash product $\prod_{s=1}^K g_s$.

Theorem 2. *(See Appendix D.1)[RS utility-profile characterization] Assume that $\mathcal{G}$ is nonempty, compact, and convex. Then*

$$g^* = \arg\max_{g \in \mathcal{G}} \sum_{s=1}^K \log g_s \tag{8}$$

is unique. Define $\lambda_s^* = (1/g_s^*)/\sum_{j=1}^K (1/g_j^*)$. Then $\boldsymbol{\lambda}^* \in \Delta^{K-1}$ and

$$\sum_{s=1}^K \lambda_s^* g_s^* \geq \sum_{s=1}^K \lambda_s^* g_s \quad \forall g \in \mathcal{G}, \qquad \lambda_s^* g_s^* = \lambda_t^* g_t^* \quad \forall s, t. \tag{9}$$

Thus, the exact RS solution is unique in utility-profile space and satisfies the supporting-hyperplane and equal-weighted-gain conditions underlying the Rawls–Shapley characterization. The classifier or parameter vector producing that profile need not be unique. Theorem 2 applies to the exact objective ($\varepsilon = 0$); $\varepsilon > 0$ in implementation serves only as numerical stabilization.

3.2 Worst-Group Certificate and Max-Min Relation

Although Nash welfare is not a strict max-min criterion, its product structure provides a pointwise certificate for the worst-group gain.

Proposition 1. *(See in Appendix D.2)[Worst-group gain certificate] If $0 < G_s(\theta) \leq 1$ for all s, then*

$$\min_s G_s(\theta) \geq \exp\big(-\mathcal{L}_{\mathrm{RS}}^0(\theta)\big). \tag{10}$$

For the stabilized objective, if $0 \leq G_s(\theta) \leq 1$, then

$$\min_s G_s(\theta) \geq \frac{\exp\big(-\mathcal{L}_{\mathrm{RS}}^\varepsilon(\theta)\big)}{(1+\varepsilon)^{K-1}} - \varepsilon. \tag{11}$$

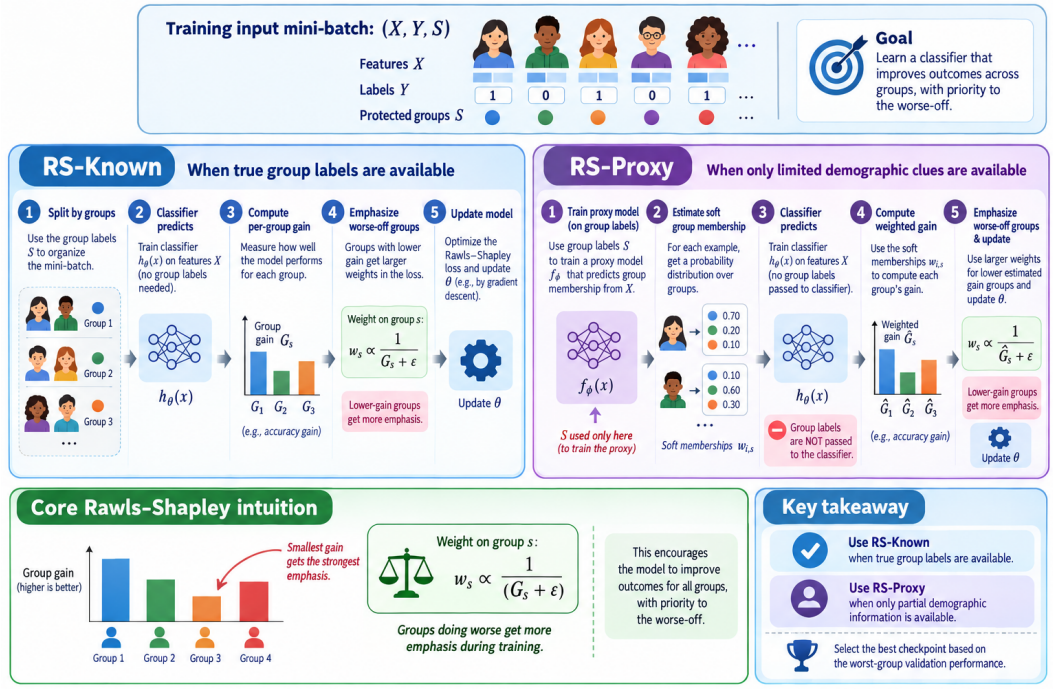


Figure 2: Training framework for RS-Known and RS-Proxy.

The certificate can be loose: if every group has gain $a \in (0, 1)$, the exact bound gives $a \geq a^K$. It therefore does not imply a constant-factor approximation to max-min fairness or monotonic improvement of the worst group during optimization.

More fundamentally, RS and max-min encode different preferences. For $A = (0.5, 0.5)$ and $B = (0.4, 0.9)$, max-min selects A because $\min_s A_s = 0.5 > 0.4 = \min_s B_s$, whereas Nash welfare selects B because $0.5 \cdot 0.5 = 0.25 < 0.36 = 0.4 \cdot 0.9$. RS therefore prioritizes lower-gain groups through inverse-gain marginal weighting without reducing to a pure max-min rule.

3.3 Parameter-Space Implications

Equation equation 7 shows that RS assigns group s a marginal coefficient proportional to $1/(G_s+\varepsilon)$. At a neural-network stationary point θ^*, $\sum_{s=1}^{K}(G_s(\theta^*) + \varepsilon)^{-1}\nabla_\theta G_s(\theta^*) = 0$. This is only a first-order condition in parameter space: it does not imply global optimality, uniqueness, Pareto efficiency, or the supporting-hyperplane condition in Theorem 2. Those conclusions require the convex utility-profile assumptions stated above.

4 Algorithmic Instantiations

We consider two training regimes that differ in access to demographic attributes. Figure 2 summarizes both variants.

4.1 RS-Known: Observed Group Membership

When group labels s_i are available during training, group utilities are computed directly using equation 5. We use group-stratified minibatches with $|\mathcal{B}_s| > 0$ for every group, ensuring that each mini-batch utility U_s, gain $G_s = U_s - d_s$, and the corresponding RS objective are well defined. Model parameters are then updated by minimizing $\mathcal{L}_{\mathrm{RS}}^{\varepsilon}$. The detailed algorithm is given in Appendix section A.2.

4.2 RS-PROXY: LIMITED DEMOGRAPHICS

When protected attributes are not provided directly to the classifier, we consider a *limited-demographics* regime in which group labels are used only to train a proxy model. Let $f_\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^K$ predict group membership, with soft assignments $w_{i,s} = \mathrm{softmax}(f_\phi(x_i, y_i))_s$. The classifier itself never receives s_i as an input. This setting is therefore distinct from fully label-free fairness, which would require unsupervised group discovery.

Using the soft-correctness score $c_i(\theta)$, proxy-weighted group utility is estimated as

$$\widehat{U}_s(\theta, \phi) = \frac{\sum_{i=1}^n w_{i,s} c_i(\theta)}{\sum_{i=1}^n w_{i,s} + \varepsilon_{\mathrm{den}}}, \tag{12}$$

where $\varepsilon_{\mathrm{den}} > 0$ stabilizes the denominator. With $\widehat{G}_s = \widehat{U}_s - d_s$, RS-Proxy minimizes

$$\mathcal{L}^{\varepsilon}_{\mathrm{RS-P}}(\theta, \phi) = -\sum_{s=1}^{K} \log\big(\widehat{G}_s(\theta, \phi) + \varepsilon\big). \tag{13}$$

As in RS-Known, we use $d_s = 0$ in the experiments. RS-proxy algorithm is given in Appendix section A.1.

Graceful degradation. If the proxy becomes uninformative, then $w_{i,s} \approx 1/K$ and $\widehat{U}_s \approx \overline{U}$, where $\overline{U}(\theta) = n^{-1} \sum_i c_i(\theta)$. Consequently,

$$\mathcal{L}^{\varepsilon}_{\mathrm{RS-P}} \approx -K \log\big(\overline{U}(\theta) + \varepsilon\big). \tag{14}$$

Thus, as recoverable group structure vanishes, the group-adaptive component of RS-Proxy disappears and the objective approaches maximization of aggregate soft correctness. This limiting objective is not identical to cross-entropy ERM.

5 EXPERIMENTS

5.1 DATASETS AND PRE-PROCESSING

We evaluate on Adult Income and COMPAS Recidivism, two widely used fairness benchmarks. **Adult** Kohavi (1996) contains $n = 48{,}841$ instances with income above \$50K as the target. We form four intersectional groups from sex (Male/Female) and race (White/Non-White): Female/Non-White, Female/White, Male/Non-White, and Male/White. **COMPAS** Angwin et al. (2016) contains approximately 6,172 instances after standard ProPublica filtering, with two-year recidivism as the target; groups are defined by sex and race (Black/Non-Black), again yielding four intersectional groups. COMPAS outcomes are known to contain substantial measurement noise Eckhouse (2017).

Protected attributes are excluded from classifier inputs. They are used only to construct groups for RS-Known and evaluation, and to train the proxy model in RS-Proxy. Features are standardized using training-set statistics. We use 70%/15%/15% train/validation/test splits stratified by group and report means over five independent runs.

5.2 BASELINES AND IMPLEMENTATION

We compare nine methods: standard MLP empirical risk minimization (ERM); Adversarially Reweighted Learning (ARL) Lahoti et al. (2020); RS-Known and RS-Proxy; an equalized-odds-regularized MLP (EqOdds-Reg); ℓ_2-regularized logistic regression (LR) and its inverse group-frequency reweighted variant (LR+RW); and XGBoost Chen & Guestrin (2016) with and without inverse group-frequency reweighting (XGB+RW). LR uses $C = 1$, and XGBoost uses 200 estimators with maximum depth 4.

EqOdds-Reg augments cross-entropy with

$$\mathcal{L}_{\mathrm{EO}} = L_{\mathrm{CE}} + \frac{\mu}{K} \sum_{s=1}^{K} \Big(|\widehat{\mathrm{TPR}}_s - \widehat{\mathrm{TPR}}| + |\widehat{\mathrm{FPR}}_s - \widehat{\mathrm{FPR}}| \Big), \tag{15}$$

Table 1: Adult Income results (mean over 5 runs). ↑ higher is better; ↓ lower (less disparity) is better. **Bold**: best value per column among all methods. Underline: best among neural (MLP) methods.

Method	Acc-Avg↑	Acc-Min↑	AUC-Avg↑	AUC-Min↑	Δ_{EO} ↓	EqOpp↓
Baseline (ERM)	0.8531	0.8142	0.9069	0.8826	0.1685	0.1201
ARL	0.8520	0.8121	0.9072	0.8822	0.1608	0.1123
RS-Known	0.8522	0.8121	0.8975	0.8727	0.1979	0.1486
RS-Proxy	0.8522	0.8127	0.8943	0.8707	0.1448	0.1052
EqOdds-Reg	0.8495	0.8122	0.8980	0.8802	**0.1037**	**0.0796**
Logistic Reg.	0.8490	0.8079	0.9034	0.8771	0.1664	0.1173
LR+RW	0.8487	0.8082	0.9028	0.8760	0.1887	0.1407
XGBoost	**0.8694**	**0.8326**	**0.9240**	**0.9029**	0.1499	0.1134
XGB+RW	0.8688	0.8316	0.9228	0.9008	0.1644	0.1269

Table 2: COMPAS Recidivism results (mean over 5 runs). Same notation as Table 1.

Method	Acc-Avg↑	Acc-Min↑	AUC-Avg↑	AUC-Min↑	Δ_{EO} ↓	EqOpp↓
Baseline (ERM)	0.6857	0.6545	0.7426	0.6612	0.3248	0.1945
ARL	0.6799	0.6605	0.7368	0.6524	0.3618	0.2318
RS-Known	0.6819	0.6471	0.7356	0.6572	0.3218	0.2043
RS-Proxy	0.6784	0.6514	0.7307	0.6440	0.3325	0.2227
EqOdds-Reg	0.6670	0.6284	0.7271	0.6527	**0.2740**	0.1922
Logistic Reg.	0.6819	0.6484	0.7344	0.6369	0.3973	0.2687
LR+RW	0.6799	0.6499	0.7359	0.6486	0.3339	0.2172
XGBoost	**0.6909**	**0.6658**	**0.7374**	**0.6586**	0.3294	**0.1907**
XGB+RW	0.6855	0.6604	0.7343	0.6567	0.3147	0.1803

with $\mu = 1$, using sigmoid probabilities as soft TPR/FPR indicators.

All MLP-based methods use the same $64 \to 32$ architecture with ReLU activations, Adam optimization, learning rate 10^{-3}, batch size 256, and 50 epochs. The checkpoint with the highest worst-group validation accuracy is retained.

5.3 Evaluation Metrics

We report overall, macro-averaged group, and minimum-group accuracy (**Acc-Avg**, **Acc-Mac**, **Acc-Min**) and AUC (**AUC-Avg**, **AUC-Mac**, **AUC-Min**). Fairness is measured by the equalized-odds gap $\Delta_{\mathrm{EO}} = \max_s |\widehat{\mathrm{TPR}}_s - \widehat{\mathrm{TPR}}| + \max_s |\widehat{\mathrm{FPR}}_s - \widehat{\mathrm{FPR}}|$ and the equal-opportunity gap (**EqOpp-Gap**), given by the TPR term alone. Lower gap values indicate smaller group disparities.

5.4 Results

Tables 1 and 2 summarise the main results averaged over 5 runs; Table 3 gives per-group AUC on the last run.

Adult Income (Table 1). Among neural methods, ARL achieves the highest overall AUC (0.907), while RS-Known and RS-Proxy obtain 0.898 and 0.894, respectively, while RS-Proxy attains the lowest equalized-odds gap among all neural approaches (0.145), closely followed by EqOdds-Reg (0.104). Notably, EqOdds-Reg achieves the best Δ_{EO} overall (0.104) and the best EqOpp-Gap (0.080), illustrating that directly penalising TPR/FPR disparity is the most targeted strategy for parity. Worst-group accuracy is similar across the neural methods, ranging from 0.8121 to 0.8142.

Reweighted variants (LR+RW, XGB+RW) do not consistently improve upon their unweighted counterparts.

COMPAS Recidivism (Table 2). Results on COMPAS are notably tighter across all methods, reflecting the dataset's noisy labels and lower group identifiability. Among neural methods, ARL achieves the best Acc-Min (0.661), while RS-Known achieves the best Δ_{EO} among the RS variants (0.322). EqOdds-Reg again achieves the best Δ_{EO} overall (0.274), but at the cost of Acc-Min (0.628), the lowest of all methods. RS-Proxy's performance closely tracks ERM on COMPAS, consistent with the graceful-degradation prediction: when groups are less computationally identifiable, the proxy weights approach uniform, and the RS-Proxy objective converges to standard ERM.

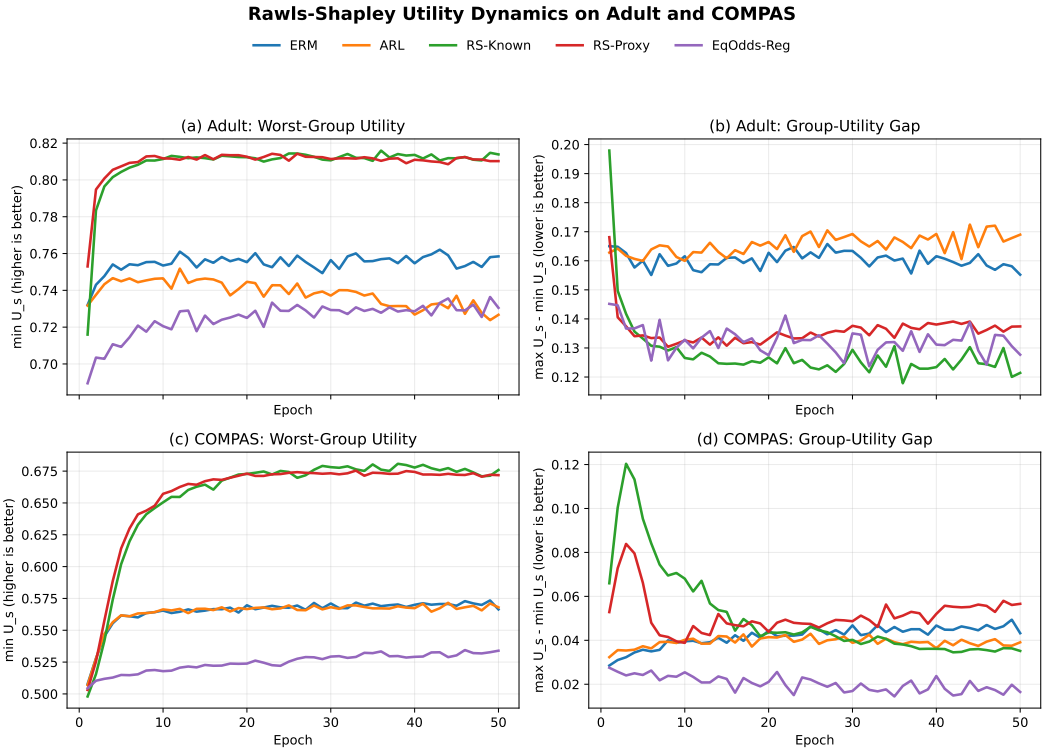


Figure 3: Rawls–Shapley Utility Dynamics on Adult and COMPAS.

Per-group AUC (Table 3). On Adult, AUC is highest for Female/Non-White and Female/White groups across most methods, while Male/White shows the lowest AUC with an artefact of class imbalance within that group. ARL achieves the best Female/Non-White AUC (0.929), while XGBoost dominates all groups in absolute terms. RS-Known and RS-Proxy show more balanced AUC profiles across groups than the ERM baseline, consistent with the Nash-product objective distributing gradient more evenly.

Figure 3 presents the utility-aligned training dynamics on the Adult and COMPAS datasets. Across both datasets, RS-Known and RS-Proxy substantially improve worst-group utility compared with ERM, ARL, and EqOdds-Reg, demonstrating the ability of the proposed objective to prioritize worse-off groups. On Adult, RS-Known additionally achieves a smaller group-utility gap than the competing methods, indicating a more balanced distribution of utility across protected groups. On COMPAS, RS-Known and RS-Proxy similarly improve worst-group utility, although the utility-gap advantage is less pronounced, with EqOdds-Reg achieving the smallest gap. Overall, these results show that Rawls–Shapley learning consistently strengthens worst-group utility across datasets while its effect on utility parity varies with the underlying data and group structure.

Utility–Parity Trade-off. A recurring theme across both datasets is the trade-off between *utility* (AUC-Min, Acc-Min) and *parity* (Δ_{EO}). RS-Known and ARL improve utility for worst-off groups

Table 3: Per-group AUC on Adult Income and COMPAS Recidivism (last run). For Adult, groups are F/NW = Female/Non-White, F/W = Female/White, M/NW = Male/Non-White, and M/W = Male/White. For COMPAS, groups are F/NB = Female/Non-Black, F/B = Female/Black, M/NB = Male/Non-Black, and M/B = Male/Black. **Bold** indicates the best result in each column.

	Adult Income				**COMPAS Recidivism**			
Method	**F/NW**	**F/W**	**M/NW**	**M/W**	**F/NB**	**F/B**	**M/NB**	**M/B**
Baseline (ERM)	0.9151	0.9212	0.9083	0.8797	0.7611	0.7612	0.7006	0.7091
ARL	**0.9295**	0.9267	0.9056	0.8792	0.7683	0.7586	0.6984	**0.7127**
RS-Known	0.8946	0.9202	0.8893	0.8691	0.7263	0.7520	**0.7011**	0.6964
RS-Proxy	0.8941	0.9142	0.8873	0.8674	0.7473	0.7609	0.6939	0.7062
EqOdds-Reg	0.8948	0.9172	0.8970	0.8755	0.7714	0.7698	0.6547	0.7071
Logistic Reg.	0.9074	0.9210	0.8964	0.8723	0.7135	**0.7731**	0.6829	0.7103
LR+RW	0.9091	0.9213	0.8969	0.8713	0.7371	0.7573	0.6914	0.7024
XGBoost	0.9255	**0.9382**	0.9131	**0.9004**	0.7729	0.7725	0.6856	0.6923
XGB+RW	0.9202	**0.9382**	**0.9146**	0.8983	**0.7791**	0.7652	0.6899	0.6816

by targeting the Nash product or adversarially weighted loss, respectively; EqOdds-Reg directly penalises statistical gaps. The two objectives are not equivalent: reducing Δ_{EO} via EqOdds-Reg can reduce Acc-Min (most clearly on COMPAS, where Acc-Min drops from 0.654 to 0.628), while the RS objectives improve individual-group performance without explicitly targeting parity. This confirms that max-min and parity fairness are complementary rather than substitutable.

6 Conclusion

We presented Rawls–Shapley (RS) learning, grounded in Yaari's Yaari (1981) connection between Rawls' Difference Principle, Shapley's utility-transfer framework, and the Nash bargaining solution. The adaptive weighting does not imply that every parameter update improves the worst-off group, since the update also depends on the magnitudes and directions of the group gradients.

Empirically, RS-Proxy achieves the best equalized-odds gap among neural methods on Adult, while EqOdds-Reg achieves the best parity across both datasets. On COMPAS, the graceful-degradation property of RS-Proxy holds. It closely matches ERM when groups are not computationally identifiable. XGBoost without fairness constraints achieves the highest utility on both datasets, reinforcing that fairness objectives involve a real utility–parity trade-off that method designers must make explicit.

AI USE STATEMENT

Large language models, i.e., ChatGPT & Claude, were used to refine the manuscript's language, improve clarity, and enhance readability. All AI-assisted content was critically reviewed, edited, and approved by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

Our work adheres to the ICLR Code of Ethics and established principles of research integrity and reproducibility. This study does not involve human or animal subjects and uses publicly available benchmark datasets in accordance with their applicable licensing and usage requirements. The proposed framework does not collect or process personally identifiable information and does not introduce additional privacy, security, or legal compliance concerns beyond those associated with the benchmark datasets. We are not aware of any conflicts of interest or foreseeable harmful societal impacts arising directly from this work. We acknowledge that fairness methods may have different implications across datasets, populations, and deployment contexts; therefore, the proposed approach should be carefully evaluated before application in high-stakes settings.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility and transparency of our results. Upon publication, we will release the complete codebase, pretrained models, and detailed instructions for reproducing the experiments. The dataset preprocessing procedures and evaluation metrics are documented in the main paper and appendix, with additional implementation details provided in the supplementary materials. To further support reproducibility, we use fixed random seeds and specify the relevant hardware and software environments used in our experiments.

REFERENCES

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016. Accessed: 2024-01-01.

Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104: 671–732, 2016.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pp. 77–91. PMLR, 2018.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

Laurel Eckhouse. Big data may be reinforcing racial bias in the criminal justice system. The Washington Post, 2017. Accessed: 2024-01-01.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, pp. 3315–3323, 2016.

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*, pp. 1929–1938. PMLR, 2018.

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*, pp. 202–207, 1996.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. Fairness without demographics through adversarially reweighted learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 728–740, 2020.

John Rawls. *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA, 1971.

Lloyd S. Shapley. Utility comparison and the theory of games. In *La Décision: Agrégation et Dynamique des Ordres de Préférence*, pp. 251–263. Editions du CNRS, Paris, 1969.

Menahem E. Yaari. Rawls, Edgeworth, Shapley, Nash: Theories of distributive justice re-examined. *Journal of Economic Theory*, 24(1):1–39, 1981.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1171–1180, 2017.

A ALGORITHMS FOR RS-KNOWN AND RS-PROXY

A.1 ALGORITHMS FOR RS-PROXY

Algorithm 1: RS-Proxy Training (Limited Demographics)

Require: Data $\{(x_i, y_i, s_i)\}$, learner h_θ, proxy predictor f_ϕ (linear), K, η, T, B, $\{d_s\}_{s=1}^K$, ε_{den}, ε
Require: **Note:** s_i are used only to train f_ϕ, never passed to h_θ
Ensure: Best checkpoint θ^*

Initialise θ, ϕ
for $t = 1$ **to** T **do**
  for each mini-batch $\mathcal{B}$ **do**
    Compute $p_i = \sigma(h_\theta(x_i))$, $c_i = y_i p_i + (1 - y_i)(1 - p_i)$
    [Proxy predictor step which uses group labels s_i]
    $\mathcal{L}_\phi \leftarrow \mathrm{CE}(f_\phi(x_i, y_i),\ s_i)$
    $\phi \leftarrow \phi - \eta \nabla_\phi \mathcal{L}_\phi$
    [Learner step where group labels NOT used]
    $w_{i,s} \leftarrow \mathrm{softmax}(f_\phi(x_i, y_i))_s$ (detached from ϕ graph)
    for each group s **do**
      $\widehat{U}_s \leftarrow \big(\sum_i w_{i,s} c_i\big) / \big(\sum_i w_{i,s} + \varepsilon_{\text{den}}\big)$
      $\widehat{G}_s \leftarrow \widehat{U}_s - d_s$
    end for
    $\mathcal{L}^{\varepsilon}_{\text{RS-P}} \leftarrow -\sum_s \log(\widehat{G}_s + \varepsilon)$
    $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}^{\varepsilon}_{\text{RS-P}}$
  end for
  Save best θ by worst-group accuracy on validation set
end for
return θ^*

A.2 ALGORITHMS FOR RS-KNOWN

Algorithm 2: RS-Known Training

Require: Data $\{(x_i, y_i, s_i)\}$, learner h_θ, groups $\mathcal{S}$, lr η, epochs T, group-stratified batch size B, disagreement points $\{d_s\}_{s=1}^K$, ε
Ensure: Best checkpoint θ^*

Initialise θ
for $t = 1$ **to** T **do**
  for each mini-batch $\mathcal{B} \subseteq \mathcal{D}$ **do**
    Compute probabilities: $p_i = \sigma(h_\theta(x_i))$ for $(x_i, y_i, s_i) \in \mathcal{B}$
    Compute element-wise soft-correct: $c_i = y_i p_i + (1 - y_i)(1 - p_i)$
    for each group $s \in \mathcal{S}$ **do**
      $U_s \leftarrow \frac{1}{|\mathcal{B}_s|} \sum_{i:\, s_i = s} c_i$ *(soft accuracy)*
      $G_s \leftarrow U_s - d_s$
    end for
    $\mathcal{L}^{\varepsilon}_{\text{RS}} \leftarrow -\sum_{s \in \mathcal{S}} \log(G_s + \varepsilon)$
    $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}^{\varepsilon}_{\text{RS}}$
  end for
  Save best θ by worst-group accuracy on validation set
end for
return θ^*

B THEORY COMPONENTS

B.1 SENSITIVITY TO ε

The parameter $\varepsilon > 0$ is introduced primarily for numerical stability and controls the degree of variation in the inverse-gain coefficients $\frac{1}{G_s(\theta)+\varepsilon}$. As $\varepsilon \to 0$, the stabilized objective approaches the exact log-Nash objective,

$$\mathcal{L}_{\mathrm{RS}}^{\varepsilon}(\theta) \longrightarrow -\sum_{s=1}^{K} \log G_s(\theta),$$

provided that every group gain is strictly positive. The corresponding coefficients approach $1/G_s(\theta)$. Hence lower-gain groups receive larger marginal weight, although the weights concentrate strongly on a single group only when its gain is much smaller than the others.

For large ε, a Taylor expansion gives

$$\begin{aligned}\mathcal{L}_{\mathrm{RS}}^{\varepsilon}(\theta) &= -\sum_{s=1}^{K} \log\big(G_s(\theta)+\varepsilon\big) \\ &= -K\log\varepsilon - \frac{1}{\varepsilon}\sum_{s=1}^{K} G_s(\theta) + O\big(\varepsilon^{-2}\big).\end{aligned}$$

Therefore, up to constants and positive rescaling, minimizing the objective approaches maximizing the sum of the group gains. It does not generally become cross-entropy ERM. In all experiments, we use $\varepsilon = 10^{-8}$, which is negligible relative to the observed utility values and serves only to prevent numerical singularities.

B.2 UTILITY-BASED VS. LOSS-BASED OBJECTIVES

A central design choice in this work is using soft accuracy U_s rather than cross-entropy loss L_s as the utility measure. As argued in Section 2, using $\exp(-L_s)$ as a proxy utility collapses the RS objective to group-balanced ERM. Using soft accuracy preserves the nonlinear inverse-utility weighting of the log-Nash objective. Although

$$-\log\prod_{s=1}^{K} U_s = -\sum_{s=1}^{K}\log U_s$$

is additively separable at the utility level, it differs from a linear sum of group utilities because the marginal coefficient of group s is $1/U_s$. Thus, its distinguishing feature is adaptive inverse-utility weighting, not cross-group derivative interaction.

B.3 IMPLICIT WEIGHT DYNAMICS

During RS-Known training, define the normalized implicit coefficient of group s at epoch t by

$$\lambda_s^{\varepsilon}(t) = \frac{1/(G_s(t)+\varepsilon)}{\sum_{j=1}^{K} 1/(G_j(t)+\varepsilon)}.$$

By construction,

$$\lambda_s^{\varepsilon}(t)\big(G_s(t)+\varepsilon\big) = \frac{1}{\sum_{j=1}^{K} 1/(G_j(t)+\varepsilon)}$$

for every group s and at every epoch t. This equality is an algebraic consequence of the definition of $\lambda_s^{\varepsilon}(t)$; it is not a convergence criterion.

It shows only that lower stabilized gains receive larger normalized coefficients. For $\varepsilon = 0$, weights of this form support the global Nash-welfare solution under the assumptions of Theorem 2. In the neural-network implementation, the identity alone does not establish global optimality or Yaari's supporting-hyperplane condition.

C Results Comparison

Comparison with group-balanced ERM. Group-balanced ERM minimizes $\frac{1}{K}\sum_s L_s$ and therefore assigns equal coefficients to the group-loss terms. In contrast, the RS objective differentiates group utility gains and assigns the coefficient $1/(G_s+\varepsilon)$ to $\nabla_\theta G_s$. Thus, lower-gain groups receive larger scalar coefficients. For the exact objective, the global solution under the assumptions of Theorem 2 satisfies condition equation 3; no corresponding global conclusion is claimed for a neural-network stationary point.

Comparison with max-min. The max-min objective $\min_\theta \max_s L_s$ concentrates the entire gradient on a single worst group at each step, potentially causing oscillation. $\mathcal{L}_{\mathrm{RS}}$ smoothly distributes gradient across all groups with the largest scalar coefficient assigned to the lowest-gain group, which is more stable in practice.

D Proofs and Explanations of the Theorems

D.1 Proof of Theorem 2

Let $\Phi(g)=\sum_{s=1}^K \log g_s$. Since $\mathcal{G}\subset\mathbb{R}^K_{++}$ is compact and Φ is continuous on $\mathbb{R}^K_{++}$, a maximizer exists. Since Φ is strictly concave and $\mathcal{G}$ is convex, the maximizing utility profile is unique.

Let g^* be the unique maximizer. The first-order optimality condition for maximizing a differentiable concave function over a convex set gives

$$\nabla\Phi(g^*)^\top(g-g^*)\le 0 \qquad \text{for all } g\in\mathcal{G}.$$

Since $\nabla\Phi(g^*)=\left(\frac{1}{g_1^*},\dots,\frac{1}{g_K^*}\right)$, we obtain

$$\sum_{s=1}^K \frac{1}{g_s^*}(g_s-g_s^*)\le 0.$$

Equivalently,

$$\sum_{s=1}^K \frac{1}{g_s^*}g_s \le \sum_{s=1}^K \frac{1}{g_s^*}g_s^* \qquad \text{for all } g\in\mathcal{G}.$$

Define

$$\lambda_s^*=\frac{1/g_s^*}{\sum_{j=1}^K 1/g_j^*}.$$

Because $g_s^*>0$ for every s, we have $\lambda_s^*>0$ and $\sum_s\lambda_s^*=1$. Hence $\lambda^*\in\Delta^{K-1}$. Normalizing the previous inequality gives

$$\sum_{s=1}^K \lambda_s^* g_s \le \sum_{s=1}^K \lambda_s^* g_s^* \qquad \text{for all } g\in\mathcal{G},$$

which proves equation 2.

Finally,

$$\lambda_s^* g_s^* = \frac{(1/g_s^*)g_s^*}{\sum_{j=1}^K 1/g_j^*} = \frac{1}{\sum_{j=1}^K 1/g_j^*},$$

which is independent of s. Therefore,

$$\lambda_s^* g_s^* = \lambda_t^* g_t^* \qquad \text{for all } s,t.$$

This proves equation 9.

D.2 Proof of Proposition 1

Let

$$m = \min_s G_s(\theta).$$

For the exact objective,

$$\exp\bigl(-\mathcal{L}_{\mathrm{RS}}^{0}(\theta)\bigr) = \prod_{s=1}^{K} G_s(\theta).$$

Since $0 < G_s(\theta) \le 1$ for all s, the product is at most the smallest factor:

$$\prod_{s=1}^{K} G_s(\theta) \le m.$$

Therefore,

$$\min_s G_s(\theta) = m \ge \exp\bigl(-\mathcal{L}_{\mathrm{RS}}^{0}(\theta)\bigr).$$

For the stabilized objective,

$$\exp\bigl(-\mathcal{L}_{\mathrm{RS}}^{\varepsilon}(\theta)\bigr) = \prod_{s=1}^{K} \bigl(G_s(\theta) + \varepsilon\bigr).$$

Since $m = \min_s G_s(\theta)$ and $G_s(\theta) \le 1$, the smallest shifted factor is $m + \varepsilon$, and every other shifted factor is at most $1 + \varepsilon$. Hence

$$\prod_{s=1}^{K} \bigl(G_s(\theta) + \varepsilon\bigr) \le (m + \varepsilon)(1 + \varepsilon)^{K-1}.$$

Rearranging gives

$$m \ge \frac{\exp\bigl(-\mathcal{L}_{\mathrm{RS}}^{\varepsilon}(\theta)\bigr)}{(1 + \varepsilon)^{K-1}} - \varepsilon.$$

This proves the stabilized bound.