AnchorFuse: Stable Model Fusion for Small Validation Groups

Anonymous authors
Paper under double-blind review

Abstract

Combining pretrained predictors can raise average performance while leaving the weakest group behind. With small validation groups, choosing mixture weights becomes an estimation problem. AnchorFuse forms a probability mixture of frozen predictors by measuring leave-one-model-out contributions at a shared anchor, weighting them by group error structure, and applying one entropy-regularized update. We decompose deletion contributions into an anchor-gradient term and a nonnegative curvature remainder. A shift-invariant sensitivity bound separates score estimation error from the response of the update at fixed hyperparameters. Across four benchmarks, AnchorFuse improves worst-group utility over the strongest evaluated minimax control by 1.50–2.80 points with the full estimation set. When the estimation set is reduced to a nominal one eighth while the selection set is retained, its margins over this control reach 5.95 points on COMPAS and 6.65 points on Clinical. A 27-setting coefficient study, independently retuned component removals, and repeated data partitions further characterize the complete procedure.

1 Introduction

Uniform averaging improves overall utility over the validation-selected single model on all four tasks in our experiments, yet changes worst-group utility by only $+0.13$, -0.48, $+0.17$, and $+0.90$ points on Adult, COMPAS, Clinical, and CelebA. This uneven allocation of ensemble benefit is consistent with recent evidence on group-specific task difficulty and predictive diversity (Schweighofer et al., 2025; Claucich et al., 2025).

Reweighting existing models is inexpensive, and the worst-group log-likelihood is concave in probability-mixture weights. Yet tractable optimization can be difficult to estimate: a small, noisy validation sample may determine which group is active at the optimum. Group-robust learning has long distinguished empirical worst-group fit from generalization, and recent work studies the effects of group scarcity and pooling (Sagawa et al., 2020; Awasthi et al., 2024; Cousins et al., 2024). We study this estimation problem with pretrained predictors and a fixed pool.

AnchorFuse deletes one member at a time from a shared anchor ensemble and evaluates the remaining mixture on the same group examples (Figure 1). Residual correlation, predictive diversity, and calibration error modulate these deletion effects; they are estimated statistics, not trainable parameters. A standard entropic update converts the scores into one global mixture. Group labels are used for estimation and evaluation; inference uses the same weights for every example. Here *stability* concerns the update's response to score perturbations; priority estimation error enters the analysis separately. Our contributions are:

- **A fusion rule for small groups.** Group-weighted deletion scores at a common anchor yield global mixture weights in closed form (Section 3).
- **An analysis of these scores.** A gradient–curvature decomposition explains finite deletion, and a shift-invariant bound separates contribution error, priority error, and update sensitivity (Section 4).
- **An evaluation under estimation scarcity.** Four benchmarks, minimax and constant-priority controls, coefficient sweeps, retuned ablations, and repeated partitions assess the complete procedure (Section 5).

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.

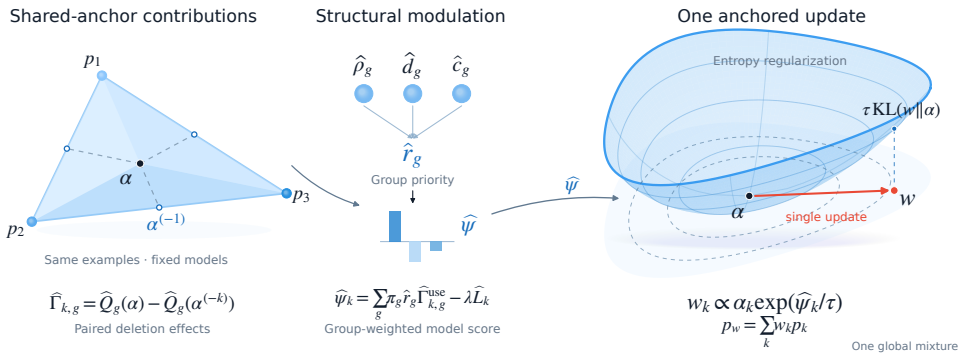

Figure 1: Overview of AnchorFuse with three models. Paired deletions at the shared anchor estimate each model's contribution to each group, and structural priorities combine these contributions into model scores. The blue surface shows the KL penalty; the red arrow marks the single entropy-regularized update to a global mixture.

2 Group-aware fusion with a fixed model pool

Let $p_1, \ldots, p_K$ be fixed binary probabilistic predictors, $K \geq 2$, and $G \in \mathcal{G}$ a group label. A mixture is $p_w(y \mid x) = \sum_k w_k p_k(y \mid x)$ with $w \in \Delta_K$. For log-likelihood, base probabilities are clipped to $[\epsilon, 1-\epsilon]$, $0 < \epsilon < 1/2$. Let $U_g(w)$ denote group utility (accuracy or AUROC), and define

$$Q_g(w) = \mathbb{E}[\log p_w(Y \mid X) \mid G = g], \qquad \widehat{Q}_g(w) = \frac{1}{n_g} \sum_{i:G_i=g} \log p_w(y_i \mid x_i). \tag{1}$$

The pool is fixed before estimation; group labels are used for fitting and evaluation, but not inference.

Utility and benefit. Worst-group utility is $U_{\min}(w) = \min_g U_g(w)$; group benefit relative to fixed p_{ref} is $T_g(w) = U_g(w) - U_g(p_{\text{ref}})$. These define different objectives, related by

$$\min_g T_g(w) \leq U_{\min}(w) - U_{\min}(p_{\text{ref}}) \leq \max_g T_g(w). \tag{2}$$

Adding $\min_g T_g$ to each reference utility proves the lower bound; evaluating at a worst reference group proves the upper bound. Since the worst group can change, benefit claims require the group-wise vector.

Convex benchmark. Each Q_g is concave because log-likelihood is concave in a positive affine mixture. Hence, for fixed constants c_g, maximizing $\min_g[Q_g(w) - c_g]$ is a convex optimization problem with a saddle-point formulation by Sion's theorem (Sion, 1958). Setting $c_g = 0$ targets worst-group log-likelihood; setting $c_g = Q_g(p_{\text{ref}})$ targets log-likelihood benefit and can change the optimizer. Accuracy and AUROC remain evaluation metrics. This benchmark is restricted to the convex hull of the fixed pool; function-space group DRO permits a broader predictor class (Thudi & Maddison, 2025).

Why the small-group regime is difficult. For independent within-group observations at fixed w, $\mathrm{Var}[\widehat{Q}_g(w)] = \mathrm{Var}[\log p_w(Y \mid X) \mid G = g]/n_g$, so smaller groups have larger sampling variance at fixed within-group variance. Selecting weights from these estimates also makes the solution depend on their errors. Our scarcity experiments therefore vary only the estimation sample, holding the model pool and test set fixed within each run; the hyperparameter-selection data remain separate and unchanged.

3 ANCHORFUSE: STRUCTURAL WEIGHTING OF ANCHORED CONTRIBUTIONS

AnchorFuse uses a full-support anchor $\alpha \in \Delta_K$, uniform in the primary experiments. Deleting model k gives $\alpha_k^{(-k)} = 0$ and $\alpha_j^{(-k)} = \alpha_j/(1-\alpha_k)$ for $j \neq k$. The deletion mixture $(p_\alpha - \alpha_k p_k)/(1-\alpha_k)$ is available from cached predictions without refitting.

3.1 ESTIMATE CONTRIBUTIONS AT A SHARED ANCHOR

For each model k and each scored group g with $n_g > 0$, we use the paired contrast

$$\widehat{\Gamma}_{k,g} = \widehat{Q}_g(\alpha) - \widehat{Q}_g(\alpha^{(-k)}) = \frac{1}{n_g} \sum_{i:G_i=g} \log \frac{p_\alpha(y_i \mid x_i)}{p_{\alpha^{(-k)}}(y_i \mid x_i)}. \tag{3}$$

A positive value means deletion lowers the group's anchor log-likelihood. The contrast is paired: for independent within-group observations, $A = \log p_\alpha(Y \mid X)$, and $B = \log p_{\alpha^{(-k)}}(Y \mid X)$, its variance is $[\mathrm{Var}(A \mid g) + \mathrm{Var}(B \mid g) - 2\,\mathrm{Cov}(A, B \mid g)]/n_g$. Positive covariance reduces variance relative to independent means.

3.2 MODULATE CONTRIBUTIONS BY GROUP STRUCTURE

With signed residuals $e_k = p_k(1 \mid X) - Y$, we compute three group statistics and a raw priority:

$$\widehat{\rho}_g = \frac{2}{K(K-1)} \sum_{i<j} \widehat{\mathrm{Corr}}(e_i, e_j \mid g), \qquad \widehat{d}_g = \widehat{\mathbb{E}}[\mathrm{Var}_k(p_k(1 \mid X)) \mid g],$$

$$\widehat{c}_g = \widehat{\mathrm{ECE}}_g(p_\alpha), \qquad \widehat{r}_g^{\mathrm{raw}} = \mathrm{sigmoid}(\eta_\rho \widetilde{\rho}_g - \eta_d \widetilde{d}_g + \eta_c \widetilde{c}_g). \tag{4}$$

The statistics are residual correlation, predictive diversity, and anchor calibration error, with ten equal-width ECE bins (Guo et al., 2017); all three coefficients default to one. Let $\mathcal{E}$ contain the groups with at least ten observations and computable residual statistics; a constant outcome label alone does not make a group ineligible. Each statistic is centered across $\mathcal{E}$ and divided by $\max\{s, 10^{-6}\}$, where s is its population standard deviation across those groups; tildes denote the standardized values. For group weights $\pi_g \geq 0$ summing to one (prevalence by default), eligible groups receive $\widehat{r}_g^{\mathrm{raw}}$, and all other groups receive $\bar{r} = \sum_{g\in\mathcal{E}} \pi_g \widehat{r}_g^{\mathrm{raw}} / \sum_{g\in\mathcal{E}} \pi_g$. If this denominator is zero, a fallback priority must be supplied separately.

The rule prioritizes correlated errors and poor calibration, and discounts predictive diversity. For centered residuals with common variance σ_g^2 and correlation ρ_g, uniform averaging reduces squared error relative to a single member by $\sigma_g^2(K-1)(1-\rho_g)/K$. Thus, at fixed variance, higher correlation leaves more error after averaging. Priorities identify groups needing attention; deletion contributions identify their supporting members.

With two eligible groups and $\Delta = u_1 - u_2$, standardization gives $\widetilde{u}_1 = \Delta/\max\{|\Delta|, 2\times 10^{-6}\}$ and $\widetilde{u}_2 = -\widetilde{u}_1$. Away from the floor, these encode ordering rather than difference magnitude: sampling-induced reversals can change priorities substantially. Proposition 3 retains this error explicitly; the floor ensures continuity near ties.

We write $\widehat{r}_g \in [0,1]$ for the final priority and $\widehat{\Gamma}^{\mathrm{use}}$ for the contribution array that enters the update. Without contribution pooling, $\widehat{\Gamma}^{\mathrm{use}} = \widehat{\Gamma}$; a pooled array is treated as a separate estimator input, with correction $\Delta_{k,g}^{\mathrm{pool}} = \widehat{\Gamma}_{k,g}^{\mathrm{use}} - \widehat{\Gamma}_{k,g}$.

3.3 MAKE ONE ENTROPY-REGULARIZED UPDATE

Let $\lambda \geq 0$, and let $\widehat{L}_k$ be model k's marginal log loss on the estimation set. We set

$$\widehat{\psi}_k = \sum_g \pi_g \widehat{r}_g \widehat{\Gamma}_{k,g}^{\mathrm{use}} - \lambda \widehat{L}_k, \qquad w_k(\widehat{\psi}) = \frac{\alpha_k \exp(\widehat{\psi}_k/\tau)}{\sum_j \alpha_j \exp(\widehat{\psi}_j/\tau)}, \quad \tau > 0. \tag{5}$$

Equation (5) is the standard entropic proximal map (**?**): it uniquely maximizes $\langle w, \widehat{\psi}\rangle - \tau\,\mathrm{KL}(w\|\alpha)$ and approaches the anchor as $\tau \to \infty$. Scores are computed once. For N cached rows, the cost

is $O(NK^2 + K|\mathcal{G}|)$ time and $O(NK + K|\mathcal{G}| + K^2)$ memory with sequential group processing, before additional pooling. For H settings with fixed inputs, weights cost $O(HK)$ and selection costs $O(HN_{\mathrm{hp}}K)$, plus $O(HN_{\mathrm{hp}} \log N_{\mathrm{hp}})$ for AUROC sorting. Inference adds $O(K)$ work beyond model forward passes.

Algorithm 1: AnchorFuse for a fixed probability-mixture pool
Input: cached predictions and outcome/group labels on D_{stat}; $\alpha, \pi, \tau, \lambda$; numerical and structural settings; fallback priority and any pooling inputs.

1. Evaluate p_α and the renormalized deletion mixtures $p_{\alpha^{(-k)}}$.
2. Compute paired contributions $\widehat{\Gamma}_{k,g}$ using Eq. (3).
3. Compute $\widehat{r}_g$ from Eq. (4) and the specified fallback; apply any pooling to obtain $\widehat{\Gamma}^{\mathrm{use}}$.
4. Form $\widehat{\psi}$ and evaluate Eq. (5) with a numerically stable softmax.

Output: $p_{\mathrm{AF}} = \sum_k w_k(\widehat{\psi}) p_k$.

Priority and scale. For $S = \sum_g \pi_g \widehat{r}_g > 0$, Eq. (5) is equivalent to using $q_g = \pi_g \widehat{r}_g / S$, temperature τ/S, and loss coefficient λ/S. Modulation changes both allocation and contribution scale. To separate them, define for each run s

$$S^{(s)} = \sum_g \pi_g \widehat{r}_g^{(s)}, \qquad r_{g,\mathrm{const}}^{(s)} = S^{(s)}, \qquad \sum_g \pi_g r_{g,\mathrm{const}}^{(s)} = S^{(s)}, \tag{6}$$

preserving mass because $\sum_g \pi_g = 1$. Hold $\widehat{\Gamma}^{\mathrm{use}}, \widehat{L}, \alpha, \pi, \tau, \lambda$, samples, pooling, and fallback outputs fixed and recompute weights without retuning. The arithmetic mean generally changes mass for nonuniform π; $r_g = 1$ changes allocation and mass. Equal mass need not imply equal $\|w - \alpha\|_1$. Constant shifts $Q_g \mapsto Q_g - c_g$ cancel in deletion contrasts, leaving AnchorFuse unchanged.

A worked example. Suppose two predictors assign observed-label probabilities $(0.9, 0.6)$ on group A and $(0.4, 0.8)$ on group B, whose prevalences are $\pi = (0.8, 0.2)$. With raw contributions, a uniform anchor, $\lambda = 0$, and $\tau = 0.2$, Eq. (5) gives $w_1 = \mathrm{sigmoid}[\sum_g \pi_g r_g \log(p_{1,g}/p_{2,g})/0.2]$. Specified priorities $(0.2, 0.8)$ give $w_1 = 0.443$, compared with $w_1 = 0.574$ under constant priorities $r_g = 0.32$ with the same total mass. Group B's mixture probability rises from 0.570 to 0.623, while group A's falls from 0.772 to 0.733. This illustrates allocation at fixed mass; estimating useful priorities from small samples remains the central problem.

4 What the deletion scores estimate

Let $\Gamma_{k,g} = Q_g(\alpha) - Q_g(\alpha^{(-k)})$ denote the population contribution. Expectations condition on the fixed pool; $\partial_k Q_g$ uses its extension to positive mixture weights.

Proposition 1 (Gradient term and deletion curvature)**.** *Let $a_k = \alpha_k/(1-\alpha_k)$ and $d_k = a_k\{p_k(Y \mid X)/p_\alpha(Y \mid X) - 1\}$. Then $d_k < 1$ and*

$$\begin{aligned} \Gamma_{k,g} &= a_k\{\partial_k Q_g(\alpha) - 1\} + R_{k,g}, \\ R_{k,g} &= \mathbb{E}[-\log(1-d_k) - d_k \mid g] \geq 0. \end{aligned} \tag{7}$$

If $|d_k| \leq b < 1$ almost surely within group g, then

$$\frac{\mathbb{E}[d_k^2 \mid g]}{2(1+b)^2} \leq R_{k,g} \leq \frac{\mathbb{E}[d_k^2 \mid g]}{2(1-b)^2}. \tag{8}$$

For a uniform anchor, $a_k = 1/(K-1)$ for every member.

Proof. The deleted probability is $p_\alpha(1-d_k) > 0$, and $\partial_k Q_g(\alpha) = \mathbb{E}[p_k/p_\alpha \mid g]$. Taylor's identity $-\log(1-d) - d = d^2 \int_0^1 (1-t)/(1-td)^2\, dt$ gives the decomposition and nonnegativity, and $1 - b \leq 1 - td \leq 1 + b$ gives both bounds. □

The gradient term measures local change; the remainder records curvature along a complete deletion. Empirically, $\widehat{G}_{k,g} = a_k \widehat{\mathbb{E}}[p_k/p_\alpha - 1 \mid g]$ gives $\widehat{\Gamma}_{k,g} = \widehat{G}_{k,g} + \widehat{R}_{k,g}$. A first-order approximation keeps $\widehat{G}$ alone, whereas the no-LOO ablation discards both terms (Section 5.4).

Common scores cancel. For the unpooled contribution, the anchor term $\sum_g \pi_g \widehat{r}_g \widehat{Q}_g(\alpha)$ is identical for every member and cancels in the normalization. Thus

$$w_k \propto \alpha_k \exp\left[-\frac{\sum_g \pi_g \widehat{r}_g \widehat{Q}_g(\alpha^{(-k)}) + \lambda \widehat{L}_k}{\tau}\right]. \tag{9}$$

The anchor defines the deletions; relative scores determine allocation. Pooling adds $\sum_g \pi_g \widehat{r}_g \Delta^{\text{pool}}_{k,g}/\tau$ to the exponent. Any score error shared by all members cancels, motivating the following bound.

Proposition 2 (Shift-invariant sensitivity)**.** *At fixed α and $\tau > 0$, let $\operatorname{osc}(v) = \max_k v_k - \min_k v_k$. Then*

$$\|w(\psi) - w(\psi')\|_1 \le \min\left\{2, \frac{\operatorname{osc}(\psi - \psi')}{2\tau}\right\} \le \frac{\|\psi - \psi'\|_\infty}{\tau}. \tag{10}$$

In particular, $\|w(\psi) - \alpha\|_1 \le \min\{2, \operatorname{osc}(\psi)/(2\tau)\}$.

Proof. The Jacobian maps v to $\big(w_k(v_k - \langle w, v\rangle)/\tau\big)_k$. By Cauchy–Schwarz, $\sum_k w_k |v_k - \langle w, v\rangle| \le \sqrt{\operatorname{Var}_w(v)} \le \|v\|_\infty$. Subtracting the midpoint of the perturbation range and integrating along the segment between the two score vectors gives the oscillation bound. The simplex diameter gives the cap of two, and $w(0) = \alpha$ gives the anchor bound. □

Increasing τ reduces sensitivity and moves weights toward the anchor. The next result separates sampling error in the priorities and contributions.

Proposition 3 (Contribution and structural error)**.** *Suppose $\widehat{\Gamma}^{\text{use}} = \widehat{\Gamma}$. Fix $\pi, \lambda, \alpha, \tau$ and comparison priorities $r_g \in [0,1]$. Let $L_k = \mathbb{E}[-\log p_k(Y \mid X)]$, $\psi_k = \sum_g \pi_g r_g \Gamma_{k,g} - \lambda L_k$, and $C = \log((1-\epsilon)/\epsilon)$. Then*

$$\|w(\widehat{\psi}) - w(\psi)\|_1 \le \frac{1}{\tau}\left[\sum_g \pi_g \left(\max_k |\widehat{\Gamma}_{k,g} - \Gamma_{k,g}| + C|\widehat{r}_g - r_g|\right) + \lambda \max_k |\widehat{L}_k - L_k|\right]. \tag{11}$$

Proof. Expand $\widehat{r}\widehat{\Gamma} - r\Gamma = \widehat{r}(\widehat{\Gamma} - \Gamma) + (\widehat{r} - r)\Gamma$, use $|\Gamma| \le C$ and $0 \le \widehat{r} \le 1$, and apply Proposition 2. The bound holds even when the two estimated terms are dependent. □

For conditionally i.i.d. within-group examples, Hoeffding's inequality gives, with probability at least $1 - \delta$ and simultaneously for all k and g,

$$|\widehat{\Gamma}_{k,g} - \Gamma_{k,g}| \le C\sqrt{\frac{2\log(2K|\mathcal{G}|/\delta)}{n_g}}. \tag{12}$$

For pooled contributions, the triangle inequality adds $\tau^{-1}\sum_g \pi_g P_g$ to the right-hand side of Eq. (11), where $P_g = \max_k |\Delta^{\text{pool}}_{k,g}|$; this term bounds an estimator change, not the population pooling bias alone. Two caveats apply. Equation (12) assumes independent group samples and does not directly cover the outcome-stratified minimum-count sampling used in our scarcity experiments. Moreover, although the standardization floor ensures continuity near ties at fixed eligibility, a small denominator can still amplify noise. The comparison is made at fixed hyperparameters and does not assert minimax optimality.

From weights to predictions. Write $f_w(x) = p_w(1 \mid x)$ and $\eta = \|w - w'\|_1/2$. Since $\sum_k (w_k - w'_k) = 0$, centering each base probability at $1/2$ gives $|f_w - f_{w'}| \le \eta$. Group accuracy therefore changes by at most the group probability of the margin event $|f_w(X) - 1/2| \le \eta$. When both classes occur, group AUROC changes by at most the probability of $|f_w(X^+) - f_w(X^-)| \le 2\eta$ for independent positive–negative pairs. Outside these margins, predictions and rankings cannot change; the bounds do not specify the direction of any utility change.

Table 1: Shared settings for the primary fixed-pool comparisons. Percentages refer to all records; D_{meta} is not used by probability-mixture methods.

Setting	Value
Data roles	Train/test: 60/20%; $D_{\text{stat}}/D_{\text{hp}}/D_{\text{meta}}$: 6/6/8%
Anchor and coefficients	$\alpha_k = 1/K$; $(\eta_\rho, \eta_d, \eta_c) = (1, 1, 1)$
Numerical settings	$\epsilon = 10^{-3}$; ten equal-width ECE bins; $n_{\min} = 10$; SD floor 10^{-6}
Fusion search	$\tau \in \{0.1, 0.5, 1, 5\}$; $\lambda \in \{0, 0.1, 0.5, 1\}$
Selection	Maximum worst-group utility on D_{hp}; ties broken by overall utility
Estimation fractions	$\rho \in \{1, 1/2, 1/4, 1/8\}$ of D_{stat}; D_{hp} and test data fixed

5 Experiments

5.1 Experimental setting

Tasks and data roles. We study four binary prediction tasks: Adult income prediction (Becker & Kohavi, 1996) ($N = 48{,}842$, two sex groups), COMPAS two-year general recidivism (Angwin et al., 2016) (a source cohort of 6,150 records, two focal race groups), MIMIC-III in-hospital mortality (Johnson et al., 2016) ($N = 35{,}364$, eight ethnicity–sex groups; hereafter Clinical), and CelebA Blond_Hair prediction (Liu et al., 2015) ($N = 202{,}599$, four label–sex groups). The COMPAS source cohort contains 3,175 African-American and 2,103 Caucasian records; records from other racial groups are retained for training. These are source-cohort counts, not test-set sizes. Clinical uses the first 48 hours of each adult patient's first ICU stay and requires a stay of at least 48 hours; its groups cross White, Black, Hispanic, and Asian categories with sex. Its positive rate is 13.4%, and its smallest group has 636 records. CelebA uses an identity-disjoint custom split; its source dataset contains 1,387 Blond Male images (0.685%).

The non-vision pools contain MLPs with hidden widths (64), $(128, 64)$, and $(256, 128, 64)$, a LightGBM model with 200 estimators, and an ℓ_2-regularized logistic regression ($K = 5$). Neural and logistic models use Adam with learning rate 10^{-3} for at most 50 epochs, with early stopping (patience five) on an internal 10% carve-out of the training set. CelebA uses four ImageNet-initialized ResNet-18 models with 224×224 inputs and distinct optimization schedules and augmentations. Utility is accuracy on every task except Clinical, where it is AUROC. Overall utility is computed on the designated test population, and worst-group utility is the within-run minimum over groups. We report means over ten paired seeds on a 0–100 scale, together with the run standard deviation (SD).

Table 1 lists the common settings; D_{stat} is the estimation set and D_{hp} the selection set. Primary fixed-pool methods share cached predictions and estimation draws, and best-single selection uses the same criterion. Scarcity draws are nested and stratified by outcome and group: each stratum with original count $n_{y,g} \geq 2$ retains $\max\{2, \lfloor \rho n_{y,g} \rfloor\}$ observations. The realized budget fraction for mixture fitting and selection is $B = (N_{\text{stat,used}} + N_{\text{hp}})/(N_{\text{stat,full}} + N_{\text{hp}})$. With equal nominal full sizes, $B = (\rho + 1)/2$ before rounding and minimum retention, i.e., 56.25% at $\rho = 1/8$. The selection set remains complete throughout.

Comparisons and selection. We instantiate the group-DRO objective (Sagawa et al., 2020) on the fixed pool: MMW maximizes $\min_g \widehat{Q}_g(w)$ over the probability-mixture simplex. We also define three variants as controls. MMW-ER adds the regularizer $-\tau\,\mathrm{KL}(w\|\alpha)$; MMW-ER-S additionally shrinks each group score toward the pooled score, $\widetilde{Q}_g = [n_g \widehat{Q}_g + m\widehat{Q}_{\text{pool}}]/(n_g + m)$ with $m \in \{5, 10, 20\}$; and MMW-B centers each group score on the shared best-single reference. MMW uses exponentiated ascent with smoothing 0.05, step sizes $\{0.01, 0.05, 0.1, 0.5\}$, and at most 2,000 steps, stopping when the ℓ_1 change between iterates falls below 10^{-5}; this is a stopping rule, not a certificate of optimality for the original minimax objective. MMW-ER searches four regularization temperatures. FairMA-EO adapts fair model averaging (Fermanian et al., 2025) by minimizing marginal cross-entropy plus an equalized-odds penalty with coefficient $\gamma \in \{0.01, 0.1, 1, 5, 10\}$. Best-single selection and uniform averaging complete the comparison. AnchorFuse uses prevalence weights and the grid in Table 1. Primary fixed-pool methods have the same data access, although their search-grid sizes differ.

Table 2: Performance with the full estimation set (mean ± run SD over ten paired seeds). All/Worst denote overall/worst-group utility. Bold marks the highest mean in each column. MMW variants are defined in Section 5.1; † denotes an equalized-odds adaptation.

	Adult		COMPAS	
Method	All	Worst	All	Worst
Best single	84.48 ± 0.27	76.22 ± 0.61	67.15 ± 0.81	61.40 ± 1.33
Uniform	85.66 ± 0.18	76.35 ± 0.49	68.38 ± 0.59	60.92 ± 1.11
MMW	84.81 ± 0.40	77.92 ± 0.83	67.49 ± 0.76	62.95 ± 1.54
MMW-ER	85.02 ± 0.31	78.01 ± 0.55	67.78 ± 0.64	63.02 ± 1.10
MMW-B	85.10 ± 0.36	80.35 ± 0.62	68.05 ± 0.68	65.70 ± 1.05
FairMA-EO† (Fermanian et al., 2025)	84.92 ± 0.36	76.80 ± 0.58	67.28 ± 0.75	61.85 ± 1.18
AnchorFuse	$\mathbf{85.80} \pm 0.22$	$\mathbf{82.15} \pm 0.38$	$\mathbf{68.65} \pm 0.55$	$\mathbf{67.20} \pm 0.85$

	Clinical		CelebA	
Method	All	Worst	All	Worst
Best single	81.35 ± 0.40	72.88 ± 1.06	87.12 ± 0.35	78.15 ± 0.82
Uniform	$\mathbf{83.60} \pm 0.31$	73.05 ± 0.91	88.40 ± 0.29	79.05 ± 0.79
MMW	82.38 ± 0.45	75.16 ± 1.23	88.02 ± 0.37	80.55 ± 0.89
MMW-ER	82.60 ± 0.41	75.42 ± 0.92	88.06 ± 0.34	80.72 ± 0.71
MMW-B	82.42 ± 0.44	76.82 ± 0.98	87.95 ± 0.38	82.58 ± 0.70
FairMA-EO† (Fermanian et al., 2025)	82.15 ± 0.45	73.80 ± 0.95	87.85 ± 0.36	79.15 ± 0.80
AnchorFuse	83.25 ± 0.35	$\mathbf{79.62} \pm 0.70$	$\mathbf{88.52} \pm 0.27$	$\mathbf{84.88} \pm 0.52$

5.2 Overall and worst-group utility

With the full estimation set, MMW-B is the strongest evaluated minimax control in worst-group utility (Table 2). AnchorFuse improves over it by 1.80, 1.50, 2.80, and 2.30 points on Adult, COMPAS, Clinical, and CelebA, and over MMW-ER by 4.14, 4.18, 4.20, and 4.16 points. Its overall utility also exceeds that of both controls on every task. Relative to uniform averaging, overall utility changes by $+0.14$, $+0.27$, -0.35, and $+0.12$ points; on Clinical, the 0.35-point overall reduction accompanies a 6.57-point worst-group increase. These comparisons evaluate the group minimum, whose relation to groupwise benefit is given by Eq. (2).

Across five independent outer partitions with three training seeds each, AnchorFuse achieves worst-group utility of 82.03 ± 0.70, 67.05 ± 1.22, and 79.48 ± 1.05 on Adult, COMPAS, and Clinical, exceeding MMW-ER by 4.07, 4.17, and 4.28 points; the SDs summarize the combined variation across all 15 runs.

5.3 Performance under estimation scarcity

At the nominal fraction $1/8$, AnchorFuse exceeds MMW-ER by 4.88, 5.50, 5.90, and 5.00 points on the four tasks and MMW-ER-S by 4.60, 5.05, 5.00, and 4.65 points (Figure 2). On COMPAS and Clinical, its margins over MMW-B are 5.95 and 6.65 points (Table 3); intermediate MMW-B measurements are not available.

AnchorFuse also degrades more gracefully as the estimation set shrinks. Between full data and $1/8$, its worst-group utility drops by 1.75 points on COMPAS and 1.77 on Clinical, compared with 3.07 and 3.47 for MMW-ER (43.0% and 49.0% smaller losses) and 6.20 and 5.62 for MMW-B. Log-likelihood fitting remains well defined in single-class groups. All clinical positive strata initially contain at least two observations, and at $1/8$ none of the 80 group–run cells retains fewer than two positives ($0/80$); empirical AUROC additionally requires negatives.

5.4 Coefficient choice and component comparisons

Constant-priority comparisons. The constant-priority procedure exceeds $r_g = 1$ by 0.74 points on COMPAS and 0.85 on Clinical, and AnchorFuse exceeds the constant variant by 3.73 and 3.70 points (Table 3). These procedure differences do not isolate allocation: that requires the fixed-input intervention in Eq. (6), without retuning. The table's run SDs are not paired confidence intervals.

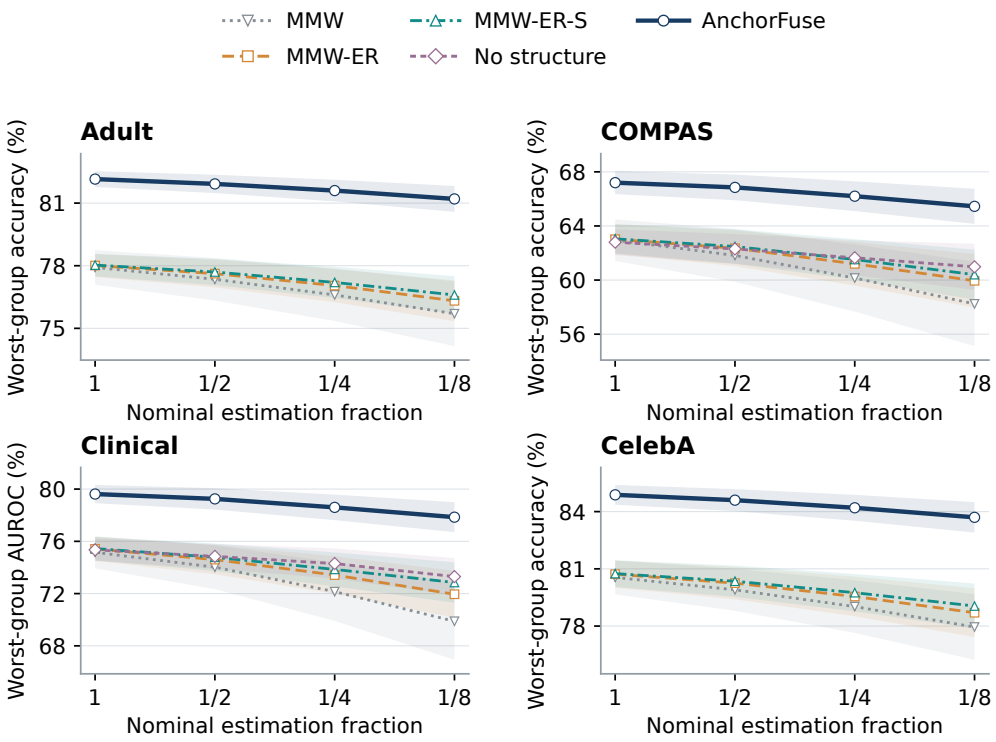

Figure 2: Worst-group test utility across nominal sampling fractions of D_{stat}, with one run SD per curve. Minimum-count retention can change the realized fractions; selection and test data remain fixed. Each curve is measured at the four fractions. Only full-size and $1/8$ MMW-B results are available, reported in Tables 2 and 3.

Table 3: Procedure comparisons at nominal estimation fraction $1/8$ (worst-group utility; mean $\pm$ run SD). The constant variant uses the full rule's mass $S = \sum_g \pi_g \widehat{r}_g$. Each row is a complete procedure, not a paired estimate of the fixed-input intervention in Eq. (6).

Method	COMPAS	Clinical
MMW	58.24 ± 3.12	69.88 ± 2.95
MMW-B	59.50 ± 2.40	71.20 ± 2.15
MMW-ER	59.95 ± 1.98	71.95 ± 1.80
MMW-ER-S	60.40 ± 1.85	72.85 ± 1.55
AnchorFuse, $r_g = 1$	60.98 ± 1.70	73.30 ± 1.42
AnchorFuse, constant $r_g = S$	61.72 ± 1.55	74.15 ± 1.30
AnchorFuse	$\mathbf{65.45} \pm 1.30$	$\mathbf{77.85} \pm 1.15$

Sensitivity to the three coefficients. Table 4 compares the default $(1, 1, 1)$ with selection over $\{0.5, 1, 2\}^3$ under the final structural configuration. Selecting the coefficients on D_{hp} improves worst-group utility by 0.14, 0.11, and 0.08 points on Adult, COMPAS, and Clinical. A separate sweep at fixed (τ, λ) gives ranges of at most 0.42 points on Adult and 0.71 on COMPAS. These small mean gains support the shared default; the fixed-parameter sweeps quantify sensitivity within the evaluated grid.

Component removals. The complete rule outperforms every independently retuned removal in Table 5. Removing structure causes the largest drop; among individual statistics, removing correlation has the largest effect. Retuning makes these comparisons assessments of each complete

Table 4: Structural-coefficient selection with the full estimation set (worst-group utility; mean ± run SD). Selected coefficients are chosen on D_{hp} from the 27 settings in $\{0.5, 1, 2\}^3$.

Coefficients	Adult	COMPAS	Clinical
Fixed $(1, 1, 1)$	82.15 ± 0.38	67.20 ± 0.85	79.62 ± 0.70
Selected on D_{hp}	82.29 ± 0.35	67.31 ± 0.82	79.70 ± 0.68
Difference	$+0.14$	$+0.11$	$+0.08$

procedure. The no-LOO variant reduces to loss-based entropy weighting and thus removes both the gradient and curvature terms of Section 4; the no-entropy variant selects the single maximum-score model.

Table 5: Worst-group utility under component removal with the full estimation set (mean ± run SD). Each variant is independently retuned on D_{hp} under a common search budget. No structure sets $r_g = 1$; no LOO (leave-one-model-out) contribution sets $\Gamma_{k,g} = 0$; no entropy penalty selects a maximum-score simplex vertex.

Variant	Adult	COMPAS	Clinical
No structure	77.85 ± 0.55	62.80 ± 0.95	75.35 ± 0.85
No correlation	79.05 ± 0.58	64.20 ± 1.00	76.60 ± 0.90
No diversity	79.86 ± 0.52	65.10 ± 0.95	77.35 ± 0.85
No calibration	81.34 ± 0.46	66.45 ± 0.90	78.40 ± 0.78
No LOO contribution	78.42 ± 0.62	63.55 ± 1.10	76.10 ± 0.95
No entropy penalty	80.61 ± 1.38	65.35 ± 1.60	77.60 ± 1.45
AnchorFuse	82.15 ± 0.38	67.20 ± 0.85	79.62 ± 0.70

6 Related work and scope

Ensembles and post hoc intervention. Deep ensembles support uncertainty estimation (Lakshminarayanan et al., 2017; Ovadia et al., 2019). Fair model averaging trades accuracy against fairness (Fermanian et al., 2025); Hardt et al. (2016) target equalized odds, and FRAPPÉ learns input-dependent corrections (Tifrea et al., 2024). AnchorFuse uses one global mixture over a fixed pool.

Robust objectives and group scarcity. Group DRO studies worst-group generalization (Sagawa et al., 2020); MixMax characterizes optimal data mixtures for function-space group DRO (Thudi & Maddison, 2025). Sparse-group and pooling analyses address complementary estimation questions (Awasthi et al., 2024; Cousins et al., 2024). Here predictors remain frozen.

Scope. The guarantees concern weight sensitivity at fixed hyperparameters, with priority error explicit in Proposition 3; they exclude data-dependent changes in τ or the pool. Experiments use four or five models, the stated solvers and grids, and complete selection data. Run SDs combine each protocol's sources of variability. Conditional update stability and empirical prediction performance are distinct conclusions.

7 Conclusion

AnchorFuse combines anchored deletion contributions and structural priorities in one entropy-regularized update. A gradient–curvature decomposition explains finite deletion, and a sensitivity bound controls the response to score error. With selection data retained, AnchorFuse improves worst-group prediction over the tested minimax baselines across the evaluated tasks and estimation budgets.

Reproducibility Statement

The main text specifies the fusion update, its core derivations, and the reported experimental settings; Appendix D gives additional implementation details. The accompanying package contains the tabulated results, figure-generation scripts, and a NumPy reference implementation of the raw contributions, residual statistics, and score-to-weight map, in which priority and contribution pooling are explicit inputs. Synthetic tests verify the mathematical identities. Cache-based diagnostic tools support matched-priority comparisons and minimax objective checks; they have not been run on the benchmark caches. Benchmark training code and per-run prediction caches are not included.

Ethics Statement

The benchmarks include sensitive attributes and de-identified clinical records. We use them to study how a fixed ensemble distributes predictive performance across groups; our results do not constitute approval for clinical or criminal-justice deployment. Recidivism labels and demographic categories have measurement limitations. Improved worst-group prediction does not imply equalized odds, demographic parity, or an individual no-harm guarantee. Access to the clinical data and any release of derived artifacts must comply with the applicable data-use terms.

AI use statement

Generative AI assisted with manuscript revision, literature and consistency checks, mathematical derivations and proof organization, and the preparation of reference code and figures. The AI assistant generated no benchmark measurements. The authors are responsible for verifying the final manuscript, code, and reported results.

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. ProPublica, 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Pranjal Awasthi, Satyen Kale, and Ankit Pensia. Semi-supervised group DRO: Combating sparsity with unlabeled data. In *Proceedings of the 35th International Conference on Algorithmic Learning Theory*, volume 237 of *Proceedings of Machine Learning Research*, pp. 125–160, 2024.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. URL https://archive.ics.uci.edu/dataset/2/adult.

Estanislao Claucich, Sara Hooker, Diego H. Milone, Enzo Ferrante, and Rodrigo Echeveste. Fairness of deep ensembles: On the interplay between per-group task difficulty and under-representation. *arXiv preprint arXiv:2501.14551*, 2025. URL https://arxiv.org/abs/2501.14551.

Cyrus Cousins, I. Elizabeth Kumar, and Suresh Venkatasubramanian. To pool or not to pool: Analyzing the regularizing effects of group-fair training on shared models. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4573–4581, 2024.

Jean-David Fermanian, Dominique Guégan, and Xuwen Liu. Fair learning by model averaging. *Risk and Decision Analysis*, 11(1–2):20–49, 2025. doi: 10.1177/15697371251321734.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 2016. doi: 10.1038/sdata.2016.35.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015. URL https://openaccess.thecvf.com/content_iccv_2015/html/Liu_Deep_Learning_Face_ICCV_2015_paper.html.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

Kajetan Schweighofer, Adrián Arnaiz-Rodríguez, Sepp Hochreiter, and Nuria M. Oliver. The disparate benefits of deep ensembles. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 53663–53702. PMLR, 2025. URL https://proceedings.mlr.press/v267/schweighofer25a.html.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Anvith Thudi and Chris J. Maddison. MixMax: Distributional robustness in function space via optimal data mixtures. In *International Conference on Learning Representations*, 2025.

Alexandru Tifrea, Preethi Lahoti, Ben Packer, Yoni Halpern, Ahmad Beirami, and Flavien Prost. FRAPPé: A group fairness framework for post-processing everything. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48321–48343. PMLR, 2024.

A Proofs and scope of the analysis

A.1 Convex probability mixtures

Proposition 4 (Convex benchmark, different reference points)**.** *For fixed constants c_g, maximizing $\min_g[Q_g(w) - c_g]$ on Δ_K is a convex optimization problem, and*

$$\max_{w\in\Delta_K}\min_{q\in\Delta_{|\mathcal{G}|}}\sum_g q_g[Q_g(w)-c_g] = \min_{q\in\Delta_{|\mathcal{G}|}}\max_{w\in\Delta_K}\sum_g q_g[Q_g(w)-c_g]. \tag{13}$$

Proof. For any fixed example (x, y), the map $w \mapsto p_w(y \mid x)$ is affine and positive, and the logarithm is concave, so $w \mapsto \log p_w(y \mid x)$ is concave. Taking an expectation conditional on $G = g$ preserves concavity. If f_g is concave for each g, the hypograph of $\min_g f_g$ is the intersection of their convex hypographs, so the minimum is also concave; this proves the optimization claim. The function $\sum_g q_g[Q_g(w) - c_g]$ is continuous, concave in w, and linear in q on the product of two compact convex simplices, so Sion's theorem gives Eq. (13). □

Subtracting group-specific constants changes which groups are active in the minimum, so it can change both the least-favorable group mixture and the optimal model weights.

Squared-loss analogue. For scalar predictions, define the group residual vector $e = (f_1(X) - Y, \ldots, f_K(X) - Y)$ and $M_g = \mathbb{E}[ee^\top \mid G = g] \succeq 0$. Negative squared-error utility is $-w^\top M_g w$. Relative to a fixed model with group risk $m_{\mathrm{ref},g}$, the benefit is $m_{\mathrm{ref},g} - w^\top M_g w$, which is concave in w and has the same saddle-point structure. Under the special second-moment model $M_g = \sigma_g^2[(1-\rho_g)I + \rho_g \mathbf{1}\mathbf{1}^\top]$, uniform averaging has risk $\sigma_g^2[1 + (K-1)\rho_g]/K$; subtracting this from a single member's risk gives the expression in Section 3.2. Interpreting ρ_g as a Pearson error correlation additionally requires the corresponding centered, equal-variance assumptions.

What homogeneous group moments imply. If $M_g = M$ for all g, then the benefit disparity

$$\max_g T_g(w) - \min_g T_g(w) = \max_g m_{\mathrm{ref},g} - \min_g m_{\mathrm{ref},g}$$

is independent of w. This does *not* imply that an optimum, or AnchorFuse, must remain at the anchor: a nonuniform weighting may reduce the common risk of every group. For arbitrary moments,

$$\max_g T_g(w) - \min_g T_g(w) \le \mathrm{osc}(m_{\mathrm{ref}}) + \max_{g,h}\|M_g - M_h\|_{\mathrm{op}},$$

since $\|w\|_2 \le 1$. This bounds the squared-loss benefit disparity; it is not a guarantee on accuracy or AUROC disparity.

A.2 Deletion scores and local improvement

Let $z_k = p_k(Y \mid X) - p_\alpha(Y \mid X)$. The deleted mixture satisfies

$$p_{\alpha^{(-k)}}(Y \mid X) = p_\alpha(Y \mid X) - a_k z_k = p_\alpha(Y \mid X)(1 - d_k).$$

Both mixtures are strictly positive, so $d_k < 1$. Because the clipped probabilities bound the derivative, differentiation under the expectation gives $\partial_k Q_g(\alpha) = \mathbb{E}[p_k(Y \mid X)/p_\alpha(Y \mid X) \mid g]$, and consequently $\mathbb{E}[d_k \mid g] = a_k\{\partial_k Q_g(\alpha) - 1\}$. Adding and subtracting d_k inside $\mathbb{E}[-\log(1-d_k) \mid g]$ yields the identity in Eq. (7), and the scalar inequality $-\log(1-d) \ge d$ for $d < 1$ proves nonnegativity. More precisely, Taylor's integral remainder gives

$$-\log(1-d) - d = d^2\int_0^1 \frac{1-t}{(1-td)^2}\,dt.$$

When $|d| \le b < 1$, the denominator lies between $(1-b)^2$ and $(1+b)^2$, and $\int_0^1 (1-t)\,dt = 1/2$, which proves Eq. (8). The condition $|d_k| \le b < 1$ is needed only for these relative-change bounds, not for the identity. Clipping alone supplies the unconditional bounds

$$\frac{a_k^2}{2(1-\epsilon)^2}\mathbb{E}[z_k^2 \mid g] \le R_{k,g} \le \frac{a_k^2}{2\epsilon^2}\mathbb{E}[z_k^2 \mid g],$$

because $p_\alpha - t a_k z_k = (1-t)p_\alpha + t p_{\alpha^{(-k)}}$ lies in $[\epsilon, 1-\epsilon]$ for $0 \le t \le 1$. For $\alpha_k = 1/K$, we have $a_k = 1/(K-1)$.

A.3 Closed form and the sensitivity bound

The Lagrangian of $\langle w, \psi \rangle - \tau \sum_k w_k \log(w_k/\alpha_k)$ under $\sum_k w_k = 1$ has the stationarity condition $\psi_k - \tau[\log(w_k/\alpha_k) + 1] + \xi = 0$, and normalization gives Eq. (5). Because the objective is strictly concave on the simplex, this solution is unique, and it has full support.

For fixed α and τ, differentiation gives

$$[Dw(\psi)v]_k = \tau^{-1} w_k \left(v_k - \sum_j w_j v_j \right).$$

View v_k as a random variable with probability mass w_k. Its mean absolute deviation is bounded by its standard deviation, which is at most $\|v\|_\infty$ for values in $[-\|v\|_\infty, \|v\|_\infty]$. Hence $\|Dw(\psi)v\|_1 \le \|v\|_\infty/\tau$, and integration along the segment between the two score vectors gives the ℓ_∞ bound. The softmax is invariant to adding a common constant, so subtracting the midpoint of the range of $\psi - \psi'$ first gives the oscillation bound in Proposition 2; the simplex diameter of two supplies the cap in Eq. (10). In particular, with $c = (\max_k \psi_k + \min_k \psi_k)/2$ and $w(0) = \alpha$,

$$\|w(\psi) - \alpha\|_1 = \|w(\psi - c\mathbf{1}) - w(0)\|_1 \le \frac{\|\psi - c\mathbf{1}\|_\infty}{\tau} = \frac{\operatorname{osc}(\psi)}{2\tau}.$$

These bounds hold at a fixed temperature. They do not account for instability from a data-dependent change of τ or from selecting a different model pool.

A.4 Error decomposition and concentration

For Proposition 3, write

$$\widehat{\psi}_k - \psi_k = \sum_g \pi_g \left[\widehat{r}_g (\widehat{\Gamma}_{k,g} - \Gamma_{k,g}) + (\widehat{r}_g - r_g)\Gamma_{k,g} \right] - \lambda(\widehat{L}_k - L_k).$$

Both the anchor and deleted-mixture probabilities lie in $[\epsilon, 1-\epsilon]$, so each individual log-ratio, and hence its expectation, lies in $[-C, C]$. Applying the triangle inequality, $0 \le \widehat{r}_g \le 1$, and the maximum over k, followed by Proposition 2, gives Eq. (11). This deterministic statement requires no independence between $\widehat{r}_g$ and $\widehat{\Gamma}_{k,g}$. It assumes that π and λ are the same on both sides; error from estimating prevalences or tuning a different penalty is not included.

Conditional on a model pool that is independent of the estimation sample, and on a group with n_g independent examples, the log-ratio variables lie in an interval of length $2C$. Hoeffding's inequality gives

$$\Pr\{|\widehat{\Gamma}_{k,g} - \Gamma_{k,g}| > t\} \le 2\exp[-n_g t^2/(2C^2)].$$

Setting $t = C\sqrt{2\log(2K|\mathcal{G}|/\delta)/n_g}$ and taking a union bound over the $K|\mathcal{G}|$ pairs gives Eq. (12); independence across models or across these pairs is unnecessary. If examples are dependent (for example, repeated stays of one patient), this concentration statement requires an appropriate dependence argument. The result covers the raw empirical contrast in Eq. (3); contribution pooling changes the estimator and requires a separate error analysis (Appendix A.8).

A.5 Decision and ranking stability

Let $f_w(x) = p_w(1 \mid x)$ and $\eta = \|w - w'\|_1/2$. Since $\sum_k (w_k - w'_k) = 0$,

$$|f_w(x) - f_{w'}(x)| = \left| \sum_k (w_k - w'_k)\{p_k(1 \mid x) - 1/2\} \right| \le \eta.$$

For a fixed threshold $1/2$, a predicted label can change only if $|f_w(X) - 1/2| \le \eta$. Thus, for each group,

$$|\operatorname{Acc}_g(w) - \operatorname{Acc}_g(w')| \le \Pr\{|f_w(X) - 1/2| \le \eta \mid G = g\}.$$

If both classes occur in group g, take independent $X^+ \sim P(X \mid Y = 1, g)$ and $X^- \sim P(X \mid Y = 0, g)$. With the usual half credit for ties, the AUROC contribution of a pair can change only if its original score gap has magnitude at most 2η. Consequently,

$$|\operatorname{AUROC}_g(w) - \operatorname{AUROC}_g(w')| \le \Pr\{|f_w(X^+) - f_w(X^-)| \le 2\eta\}.$$

For log-likelihood evaluated on clipped mixtures, the same argument gives $|Q_g(w) - Q_g(w')| \le \eta/\epsilon$. These are sensitivity bounds; the threshold and pair margins determine how informative they are for a particular group.

A.6 Structural scale and two-group priorities

Let $S = \sum_g \pi_g \widehat{r}_g > 0$ and $q_g = \pi_g \widehat{r}_g / S$. Dividing the score and the temperature in Eq. (5) by S yields the equivalent parameters $(q, \lambda/S, \tau/S)$. The constant-priority intervention in Eq. (6) sets every $r_g = S$, so $\sum_g \pi_g r_g = S$ under both allocations. Its score is $S \sum_g \pi_g \widehat{\Gamma}^{\mathrm{use}}_{k,g} - \lambda \widehat{L}_k$, evaluated with the original temperature and anchor, while contributions, losses, group weights, regularization parameters, samples, pooling, and fallback outputs are held fixed. By contrast, setting $r_g = |\mathcal{G}|^{-1} \sum_h \widehat{r}_h$ preserves the mass only when this arithmetic mean equals S. Table 3 reports complete-procedure performance, not this fixed-input paired effect. If externally supplied priorities give $S = 0$, the original update remains defined, but the normalized reparameterization does not apply.

With two eligible groups, write $\Delta = u_1 - u_2$ and $\delta = 10^{-6}$. The standardized values are

$$\widetilde{u}_1 = \frac{\Delta}{\max\{|\Delta|, 2\delta\}}, \qquad \widetilde{u}_2 = -\widetilde{u}_1.$$

They equal ± 1 only when $|\Delta| \ge 2\delta$ and vary continuously through zero inside that interval. Away from the floor, unit structural coefficients and three nonconstant components produce sigmoid arguments in $\{-3, -1, 1, 3\}$. Changing the coefficients can still alter the sigmoid's saturation and the relative influence of disagreeing components. For a fixed eligibility set, the floor removes the zero-variance discontinuity but does not eliminate structural estimation error.

A.7 Relation between utility and benefit

Write $a_g = U_g(p_{\mathrm{ref}})$ and $t_g = U_g(w) - a_g$. Adding $\min_g t_g$ to every a_g proves the lower bound in Eq. (2), and evaluating at a minimizer of a_g proves the upper bound. With $\Delta_{\min} = U_{\min}(w) - U_{\min}(p_{\mathrm{ref}})$, $T_{\min} = \min_g t_g$, and benefit disparity $\mathrm{BD} = \max_g t_g - \min_g t_g$, this reads

$$T_{\min} \le \Delta_{\min} \le T_{\min} + \mathrm{BD}.$$

The same inequalities hold for means over the same runs. For accuracy on exhaustive, disjoint groups, the overall improvement also equals the prevalence-weighted average of the group improvements. Pooled AUROC has no corresponding averaging identity.

A.8 Contribution pooling and its target

Define $\Delta^{\mathrm{pool}}_{k,g} = \widehat{\Gamma}^{\mathrm{use}}_{k,g} - \widehat{\Gamma}_{k,g}$. The operational score differs from the raw score by $\sum_g \pi_g \widehat{r}_g \Delta^{\mathrm{pool}}_{k,g}$. Applying Proposition 2 and the triangle inequality adds at most

$$\frac{1}{\tau} \sum_g \pi_g \widehat{r}_g \max_k |\Delta^{\mathrm{pool}}_{k,g}| \le \frac{1}{\tau} \sum_g \pi_g P_g$$

to Eq. (11). This bounds an observed change in the estimator; it is not automatically a population bias. Common-anchor cancellation remains valid for the raw score, whereas the pooled version retains a model-dependent correction:

$$w_k \propto \alpha_k \exp\left\{ \frac{-\sum_g \pi_g \widehat{r}_g \widehat{Q}_g(\alpha^{(-k)}) + \sum_g \pi_g \widehat{r}_g \Delta^{\mathrm{pool}}_{k,g} - \lambda \widehat{L}_k}{\tau} \right\}.$$

For a convex pooling estimator, if used, write $\widehat{\Gamma}^{\mathrm{use}}_{k,g} = (1 - \beta_g)\widehat{\Gamma}_{k,g} + \beta_g \widehat{\Gamma}_{k,P}$ with $0 \le \beta_g \le 1$, where $\widehat{\Gamma}_{k,P}$ is the pooled empirical contrast. If $\Gamma_{k,P}$ denotes the corresponding population contrast, then

$$\begin{aligned} |\widehat{\Gamma}^{\mathrm{use}}_{k,g} - \Gamma_{k,g}| \le{}& (1 - \beta_g)|\widehat{\Gamma}_{k,g} - \Gamma_{k,g}| \\ &+ \beta_g |\widehat{\Gamma}_{k,P} - \Gamma_{k,P}| + \beta_g |\Gamma_{k,P} - \Gamma_{k,g}|. \end{aligned}$$

The three terms separate group sampling error, pooled sampling error, and population-target shift. The inequality holds pointwise, including for a data-dependent β_g.

A.9 A WORKED ALLOCATION EXAMPLE

Consider two groups with prevalences $(0.8, 0.2)$ and two predictors whose observed-label probabilities are $(p_{1,A}, p_{2,A}) = (0.9, 0.6)$ and $(p_{1,B}, p_{2,B}) = (0.4, 0.8)$, constant within each group for this illustration. For a uniform anchor, $\lambda = 0$, and $\tau = 0.2$,

$$\Gamma_{1,g} - \Gamma_{2,g} = \log(p_{1,g}/p_{2,g}), \qquad w_1 = \text{sigmoid}\left[\frac{\sum_g \pi_g r_g \log(p_{1,g}/p_{2,g})}{0.2}\right].$$

We specify priorities $(0.2, 0.8)$ to isolate allocation, giving $S = 0.32$; these priorities are inputs to the example rather than estimates from its constant predictions.

Priority	w_1	$p_w(Y \mid X, G = A)$	$p_w(Y \mid X, G = B)$
Constant $r_g = 0.32$	0.574	0.772	0.570
Group priorities $(0.2, 0.8)$	0.443	0.733	0.623

Worst-group log-likelihood rises from -0.561 to -0.473, while overall log-likelihood changes from -0.319 to -0.343. Because all four mixture probabilities exceed $1/2$, this example does not demonstrate an accuracy gain. It illustrates the allocation effect of relative priorities at a fixed total score scale; whether the structural estimator identifies useful priorities is an empirical question.

B Complete utility summaries

The primary summaries contain ten paired runs. Appendix D specifies the clinical retention rule and the post-retention event counts. Overall utility is accuracy on Adult, COMPAS, and CelebA and pooled AUROC on Clinical. Worst-group utility is the minimum over groupwise values within each run, averaged over runs. Standard deviations quantify run variability and are not confidence intervals. The complete scarcity values below are the numerical source for Figure 2.

B.1 Scarcity sweeps

Table 6: Worst-group utility at each nominal sampling fraction of D_{stat} (mean $\pm$ run SD over ten paired runs).

Dataset	Method	1	1/2	1/4	1/8
Adult	MMW	77.92 ± 0.83	77.35 ± 1.02	76.60 ± 1.25	75.70 ± 1.55
	MMW-ER	78.01 ± 0.55	77.62 ± 0.66	77.05 ± 0.80	76.32 ± 0.98
	MMW-ER-S	78.03 ± 0.54	77.70 ± 0.63	77.20 ± 0.75	76.60 ± 0.90
	AnchorFuse	82.15 ± 0.38	81.92 ± 0.44	81.60 ± 0.52	81.20 ± 0.62
COMPAS	MMW	62.95 ± 1.54	61.82 ± 1.95	60.15 ± 2.48	58.24 ± 3.12
	MMW-ER	63.02 ± 1.10	62.35 ± 1.32	61.20 ± 1.61	59.95 ± 1.98
	MMW-ER-S	63.05 ± 1.08	62.48 ± 1.28	61.52 ± 1.50	60.40 ± 1.85
	AnchorFuse, no structure	62.80 ± 0.95	62.30 ± 1.10	61.65 ± 1.30	60.98 ± 1.70
	AnchorFuse	67.20 ± 0.85	66.85 ± 0.95	66.20 ± 1.10	65.45 ± 1.30
Clinical	MMW	75.16 ± 1.23	74.02 ± 1.68	72.15 ± 2.24	69.88 ± 2.95
	MMW-ER	75.42 ± 0.92	74.60 ± 1.12	73.42 ± 1.44	71.95 ± 1.80
	MMW-ER-S	75.45 ± 0.90	74.78 ± 1.05	73.85 ± 1.30	72.85 ± 1.55
	AnchorFuse, no structure	75.35 ± 0.85	74.85 ± 0.95	74.30 ± 1.15	73.30 ± 1.42
	AnchorFuse	79.62 ± 0.70	79.25 ± 0.82	78.60 ± 0.98	77.85 ± 1.15
CelebA	MMW	80.55 ± 0.89	79.90 ± 1.10	79.02 ± 1.38	77.95 ± 1.72
	MMW-ER	80.72 ± 0.71	80.25 ± 0.85	79.55 ± 1.04	78.70 ± 1.28
	MMW-ER-S	80.75 ± 0.70	80.35 ± 0.82	79.75 ± 0.98	79.05 ± 1.18
	AnchorFuse	84.88 ± 0.52	84.60 ± 0.58	84.20 ± 0.68	83.70 ± 0.80

B.2 Secondary retraining and parity references

The primary fixed-pool comparisons are in Table 2. For context, Table 7 adds the post-processing of Hardt et al. (2016), which targets equalized odds, and retraining variants of GroupDRO (Sagawa et al., 2020) and MixMax (Thudi & Maddison, 2025), which use a different training budget. These rows are supplementary references rather than solutions to the fixed-pool problem.

Table 7: Additional retraining and parity references (mean $\pm$ run SD over ten seeds). These methods have different objectives or training budgets from the primary fixed-pool methods.

Dataset	Method	Overall utility	Worst-group utility
Adult	Hardt post-proc. (Hardt et al., 2016)	83.25 ± 0.45	77.08 ± 0.39
	GroupDRO (Sagawa et al., 2020)	83.69 ± 0.42	77.55 ± 0.46
	MixMax (Thudi & Maddison, 2025)	83.95 ± 0.44	77.72 ± 0.50
COMPAS	Hardt post-proc. (Hardt et al., 2016)	65.92 ± 0.91	62.38 ± 0.85
	GroupDRO (Sagawa et al., 2020)	66.38 ± 0.96	62.72 ± 1.02
	MixMax (Thudi & Maddison, 2025)	66.60 ± 0.92	62.90 ± 1.05
Clinical	Hardt post-proc. (Hardt et al., 2016)	79.92 ± 0.70	74.08 ± 0.79
	GroupDRO (Sagawa et al., 2020)	80.85 ± 0.52	74.83 ± 0.88
	MixMax (Thudi & Maddison, 2025)	81.05 ± 0.50	75.02 ± 0.92
CelebA	GroupDRO (Sagawa et al., 2020)	86.45 ± 0.48	80.22 ± 0.64
	MixMax (Thudi & Maddison, 2025)	86.72 ± 0.45	80.51 ± 0.68

C Additional empirical comparisons

C.1 Repeated outer partitions

Table 8: Worst-group utility over five independently generated partitions with three training seeds each. Dispersions are standard deviations across the 15 runs, including partition variation; they are not cluster-level standard errors.

Dataset	Uniform	MMW	MMW-ER	AnchorFuse
Adult	76.31 ± 0.72	77.85 ± 1.02	77.96 ± 0.84	82.03 ± 0.70
COMPAS	60.75 ± 1.48	62.80 ± 1.85	62.88 ± 1.46	67.05 ± 1.22
Clinical	72.89 ± 1.21	74.95 ± 1.60	75.20 ± 1.27	79.48 ± 1.05

AnchorFuse exceeds MMW-ER by 4.07, 4.17, and 4.28 points on Adult, COMPAS, and Clinical, with runs paired within partition and seed. Because each partition changes the training, estimation, selection, and test populations together, this comparison complements the fixed-test-set scarcity sweep; it does not isolate estimation-set sampling variance.

C.2 Run variability and computation

Run variability. All performance comparisons report mean $\pm$ run SD. The repeated-partition table summarizes 15 runs arranged as five partitions with three seeds each. Its dispersion includes both partition and training variation, and the partition remains the unit for assessing sensitivity to the data split.

Computational comparison. AnchorFuse forms one score vector and evaluates one normalized exponential update, whereas the iterative minimax controls repeatedly evaluate their objectives and update the simplex weights. Appendix D.6 separates the costs of prediction caching, structural estimation, weight computation, and hyperparameter evaluation. Because the minimax objectives are concave, the distinction concerns computation and estimation rather than escape from suboptimal local optima.

D Experimental protocol and implementation details

D.1 Cohorts and base predictors

Adult predicts whether income exceeds \$50,000. COMPAS predicts general recidivism within two years in a 6,150-record source cohort, whose two focal groups contain 3,175 African-American and 2,103 Caucasian records. Records from other racial groups are retained for training, and these counts do not describe the test-set population. Clinical predicts in-hospital mortality from features in the first 48 hours of each adult patient's first ICU stay; eligibility requires a stay of at least 48 hours. Its eight groups cross White, Black, Hispanic, and Asian categories with sex. The clinical cohort has 35,364 records, a 13.4% positive rate, and 636 records in its smallest group. CelebA predicts Blond_Hair; the Blond Male group has 1,387 images, approximately 0.685% of 202,599. Because CelebA uses an identity-disjoint custom partition, its results refer to that partition rather than to the standard benchmark split.

The non-vision pool contains MLPs with hidden layers (64), $(128, 64)$, and $(256, 128, 64)$, a LightGBM model with 200 estimators, and an ℓ_2-regularized logistic regression. The neural and logistic models use Adam with learning rate 10^{-3} for at most 50 epochs, with early stopping (patience five) on an internal 10% carve-out of the training set. Categorical inputs are one-hot encoded, and continuous inputs are normalized with training-set statistics. CelebA uses ImageNet-initialized ResNet-18 models with 224×224 inputs, ImageNet normalization, and distinct optimization schedules and augmentations. The fine-tuning search considered learning rates in $\{10^{-5}, 3 \times 10^{-5}, 10^{-4}\}$ and weight decay in $\{10^{-4}, 10^{-3}\}$.

D.2 Numerical conventions and fallback

Probabilities are clipped at $\epsilon = 10^{-3}$ for log-loss calculations. The structural settings are ten equal-width ECE bins, a sample-size threshold $n_{\min} = 10$, and the cross-group standard-deviation denominator $\max\{s, 10^{-6}\}$. Let $\mathcal{E}$ contain the groups that meet the sample-size threshold and for which the residual statistics used by the score are computable. Eligibility does not require outcome variation: a label-homogeneous CelebA group remains eligible whenever its residual statistics can be evaluated. Each statistic is standardized as

$$\widetilde{u}_g = \frac{u_g - \bar{u}_{\mathcal{E}}}{\max\{s_{u,\mathcal{E}}, 10^{-6}\}},$$

where the mean and population standard deviation are taken over $\mathcal{E}$. The update in Eq. (5) is then determined by a final priority array and a contribution array.

The common priority $\bar{r}$ assigned to ineligible groups satisfies the pooled relation $\bar{r} = \sum_g \pi_g r_g$ over the final priorities. When the eligible groups have positive total mass, $\sum_{g\in\mathcal{E}} \pi_g > 0$, this relation reduces to

$$\bar{r} = \frac{\sum_{g\in\mathcal{E}} \pi_g r_g^{\mathrm{raw}}}{\sum_{g\in\mathcal{E}} \pi_g}.$$

Outside that domain, a fallback priority is an additional input to the update. Likewise, the raw contribution in Eq. (3) and any pooled contribution $\widehat{\Gamma}^{\mathrm{use}}$ are separate inputs, and their difference is accounted for in Appendix A.8. Label-homogeneous groups retain a well-defined log-loss deletion contrast; label constancy alone does not make that contrast unavailable.

D.3 Estimation draws and group counts

The outer split is 60/20/20, and the validation partition is divided into 8%, 6%, and 6% of all records for D_{meta}, D_{stat}, and D_{hp}. Scarcity draws are nested and stratified by outcome and group, and ρ denotes a nominal sampling fraction. A stratum containing at least two original observations retains $\max\{2, \lfloor \rho n_{y,g} \rfloor\}$ of them, with $\rho \in \{1, 1/2, 1/4, 1/8\}$. The realized sample count is the sum of the retained stratum counts and need not equal $\rho|D_{\mathrm{stat}}|$. Minimum-count enforcement affects the outcome composition within a group, so these draws differ from i.i.d. sampling from that group's population distribution. The realized budget for the two roles of mixture fitting and selection is $(N_{\mathrm{stat,used}} + N_{\mathrm{hp}})/(N_{\mathrm{stat,full}} + N_{\mathrm{hp}})$. Because the two sets have equal nominal full sizes, this budget is 56.25% at $\rho = 1/8$ before rounding and minimum retention. D_{meta} is not used by probability-mixture methods and is excluded from this budget.

All clinical positive strata in the full estimation set contain at least two observations. At nominal fraction $1/8$, all 80 group–run estimation cells retain at least two positives; the number with fewer than two is $0/80$ (0%). Empirical AUROC additionally requires negative observations.

D.4 Selection and baseline solvers

AnchorFuse searches the 16 (τ, λ) combinations in Table 1. MMW uses exponentiated ascent on a smoothed minimum with smoothing temperature 0.05 and step sizes $\{0.01, 0.05, 0.1, 0.5\}$ selected on D_{hp}. It stops after 2,000 steps or when $\|w^{(t+1)} - w^{(t)}\|_1 < 10^{-5}$. Iterate change is a stopping criterion, not a certificate of global optimality. MMW-ER searches four regularization temperatures, and MMW-ER-S additionally searches $m \in \{5, 10, 20\}$. MMW-B uses the shared best-single reference. FairMA-EO minimizes marginal cross-entropy plus an equalized-odds penalty with coefficient $\gamma \in \{0.01, 0.1, 1, 5, 10\}$.

The coefficient-sensitivity experiment evaluates $\{0.5, 1, 2\}^3$ with the numerical settings in Appendix D.2, and Table 4 reports the means and run SDs. Fixed-(τ, λ) sweeps under the same configuration give ranges of at most 0.42 points on Adult and 0.71 on COMPAS. Selection on D_{hp} and the fixed-parameter range are separate summaries: a small gain from selection does not by itself bound the worst grid configuration.

D.5 Statistical reporting

The primary tables report ten paired runs, and the repeated-partition results comprise five outer partitions with three training seeds each. All displayed dispersions are run SDs. Runs within a partition share data, and method comparisons are paired by partition and training seed.

D.6 Computational account

The prediction cache uses $O(NK)$ memory and the contribution table $O(K|\mathcal{G}|)$. Retaining every group correlation matrix adds $O(K^2|\mathcal{G}|)$; processing groups sequentially reduces this additional workspace to $O(K^2)$. These matrices contain estimated statistics, not trainable parameters. Simple pooling of already computed contrasts does not change the leading cost; an alternative pooling routine has its own cost.

For H choices of (τ, λ) with fixed priorities and contribution arrays, cached score components give all weight vectors in $O(HK)$ time after aggregation. Evaluating them on N_{hp} held-out rows costs $O(HN_{\mathrm{hp}}K)$; standard AUROC evaluation additionally sorts scores, at a cost of at most $O(HN_{\mathrm{hp}} \log N_{\mathrm{hp}})$ over the disjoint groups and the pooled population. An equal-mass ECE implementation would also require sorting, whereas the equal-width implementation used here does not. At inference, combining the K model outputs adds $O(K)$ work beyond their forward passes.

D.7 Reference implementation

The reference implementation computes deletion mixtures explicitly for verification; the cached shortcut in Section 3 gives the $O(NK)$ deletion cost. The accompanying NumPy code separates empirical contributions, structural statistics, optional priority construction, and the score-to-weight map, and it accepts explicit structural coefficients, a standardization floor, and externally supplied priorities and contribution arrays. Its synthetic tests check deletion recomputation, estimator identities, the score-to-weight update and its invariances, the separation of these inputs, and the scale reparameterization. This mathematical reference is separate from benchmark model training.