REDUCING THE ADAPTATION GAP THROUGH REACHABLE FISHER GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Parameter-efficient fine-tuning (PEFT) determines not only how many parameters are trained but which local directions a model can move, so similar adapters can affect subgroup losses differently. The relevant curvature matrices are infeasible to form at adapter scale, so practice uses scalar summaries, such as the Fisher trace. What can it reveal, and what does it discard? We answer through the *reachable Fisher*: each subgroup's full-model Fisher pulled back through the adapter Jacobian, under likelihood losses the Gauss–Newton curvature the adapter can reach, whose trace is computable from score-gradient norms without forming the matrix. Under matched subgroup gradients, if the reachable-Fisher difference is positive definite on model-changing directions by a margin exceeding the Hessian–Fisher defect, every sufficiently small nonzero update along them increases the signed gap. The restricted operator norm of the Hessian difference instead determines worst-case quadratic change; a matrix-free Frobenius discrepancy upper-bounds its reachable-Fisher component, with defect and estimation errors kept separate. A trace cannot certify definiteness, nor do matched traces control magnitude: equality rules out a positive-definite difference yet permits the largest operator mismatch that trace allows. LoRA's factorization also precludes unrestricted definiteness in raw coordinates. Across 306 single-seed models, higher trace accompanies greater subgroup difficulty in 75.7% of 1,218 eligible evaluations, and trace matching during training narrows the best–worst subgroup gap in all 30 dataset–encoder–adapter combinations. Yet a held-out audit exposes what it leaves uncontrolled: the observed-label operator discrepancy falls in 23 of these 30, the primary unbiased squared-Frobenius statistic in only 16. The trace is therefore a scalable diagnostic and training heuristic, not a certificate of local gap behavior or matrix alignment.

1 INTRODUCTION

An adapter chooses more than how many parameters to train: it chooses the local directions in which the model can move. Two adapters with similar parameter counts and aggregate performance can therefore affect *subgroups*, predefined groups of inputs such as skin-tone, age, or race groups, differently. This effect is inherently geometric. First-order differences depend on how subgroup losses slope along the directions available to the adapter; once those slopes match, their local behavior depends on how the losses curve along the same directions. Recent work accordingly uses curvature-related quantities in subgroup-aware training and auditing (Tran et al., 2022; Wang et al., 2023; Sharma et al., 2026; Dai et al., 2025; Roy et al., 2026).

We study this geometry through the *reachable Fisher*. Let an adapter be a smooth map T from p trainable parameters ϕ to full model parameters $\theta = T(\phi)$, with Jacobian $J = \partial T/\partial \phi$. For subgroup a with full-model Fisher F_a, we define $F_a^{\mathrm{R}} = J^\top F_a J$. This pullback to the trainable coordinates retains the directions the adapter can express and excludes those it cannot reach (Amari, 1998; Martens, 2020). The exclusion matters: summaries computed from the full-model Fisher can create or hide an apparent reachable difference (Theorem 3). Under our likelihood assumptions the Fisher is the positive-semidefinite Gauss–Newton component of the loss curvature; the Hessian in adapter coordinates differs from F_a^{R} by a *Hessian–Fisher defect* collecting a model-curvature residual and a reparameterization term in the second derivatives of T, which we retain explicitly in every guarantee. The pullback does not remove coordinate dependence, so we compare traces only within a fixed model and parameterization.

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.

At realistic adapter scale, forming the $p \times p$ reachable Fisher is infeasible, whereas its trace is computable from squared adapter-score norms without constructing the matrix. The trace measures total reachable curvature but discards its directional distribution. We therefore distinguish the normalized trace gap from the Frobenius discrepancy $\|F_a^{\mathrm{R}} - F_b^{\mathrm{R}}\|_F$, which measures full matrix mismatch. This distinction matters empirically: trace matching reduces the normalized trace gap in 20 of 30 audited cells, yet in 9 of those 20 the primary unbiased estimator of the squared Frobenius discrepancy increases. The scalar moved as intended even when the matrix-level estimate moved oppositely. We therefore ask: *what does reachable-Fisher geometry determine about subgroup-loss changes under small PEFT updates, and what can it not determine?* The answer separates direction from magnitude: when must every sufficiently small admissible update move a signed subgroup gap in the same direction, and how large can the worst-case local change be? These questions require different spectral information. Definiteness determines whether a quadratic curvature contribution has the same sign in every admissible direction, whereas an operator norm determines its largest absolute magnitude. A trace cannot certify definiteness or generally control the operator norm.

Factorized adapters add a complication. For LoRA (Hu et al., 2021), T adds a scaled rank-r update BA to the frozen weights, but (B, A) and $(BG, G^{-1}A)$ produce the same update for any invertible $r \times r$ matrix G. This symmetry guarantees nonzero directions that leave the model unchanged. More generally, all directions that are inactive to first order form the Jacobian kernel, $\ker J = \{\delta : J\delta = 0\}$. Directional statements must therefore be made on its orthogonal complement, the *horizontal subspace* $\mathcal{H}_\phi = (\ker J)^\perp$, where every nonzero direction changes the model to first order. For subgroups a and b with losses L_a and L_b, we study the signed surrogate-loss gap $L_a - L_b$. Its local change contains a linear slope term and a quadratic curvature term. When subgroup slopes match on the reachable directions, the linear term vanishes. If $F_a^{\mathrm{R}} - F_b^{\mathrm{R}}$ is positive definite on $\mathcal{H}_\phi$ with a margin larger than the Hessian–Fisher defect, then every sufficiently small nonzero horizontal additive update increases the signed gap; when $L_a - L_b \geq 0$, it also widens the absolute gap. By contrast, the restricted operator norm of the Hessian difference determines the worst-case quadratic change; the reachable-Fisher operator norm and its Frobenius upper bound control it only up to the defect. Trace equality rules out the sufficient positive-definite regime without generally controlling magnitude.

This geometric question complements methods that optimize worst-group risk or impose outcome constraints (Sagawa et al., 2019; Hardt et al., 2016; Agarwal et al., 2018). Our theory motivates auditing trace matching at the matrix level rather than treating it as a guarantee of improved subgroup outcomes; empirically, we evaluate the trace as a diagnostic and training heuristic. Our contributions:

- **We identify the relevant space for PEFT geometry.** We characterize the LoRA Jacobian kernel exactly, prove that its dimension is at least r^2 when $r \leq \min(d_{\mathrm{in}}, d_{\mathrm{out}})$, extend the obstruction to DoRA (Liu et al., 2024), and show that definiteness must be evaluated on $\mathcal{H}_\phi$ (Section 3.1).
- **We separate direction from magnitude.** We give a defect-aware sufficient condition under which every sufficiently small horizontal update increases a signed subgroup gap (Theorem 1, Corollary 1) and characterize the worst-case quadratic change exactly through an operator norm (Theorem 2). Appendix J verifies these identities on a small nonlinear network and illustrates how a non-negligible defect can prevent the reachable-Fisher certificate from applying.
- **We establish what common summaries miss.** Full-model summaries and spectra computed separately for each group can both fail (Theorem 3), and trace matching has precise limitations (Theorem 4). We also derive a matrix-free Gram identity that computes the Frobenius discrepancy between the two groups' sampled reachable Fishers without forming either matrix (Proposition 2).
- **We test the resulting diagnostic and intervention.** Across 306 single-seed models and 1,218 eligible model–split–subgroup-axis evaluations, higher reachable-Fisher trace accompanies greater subgroup difficulty in 75.7% of cases. Trace matching during training narrows the best–worst subgroup gap in all 30 paired task–encoder–adapter cells (median relative reduction 17.2%) and the surrogate-loss gap in 27. In the held-out audit, the observed-label operator discrepancy decreases in 23 of these 30, whereas the primary unbiased squared-Frobenius statistic decreases in only 16.

Takeaways. Within a fixed model and parameterization, the reachable-Fisher trace is a scalable diagnostic and matrix-free training target, but not a certificate of matrix alignment, local gap direction, or improved subgroup performance. The Gram identity makes what it misses auditable, suggesting a practical workflow: use the trace as a heuristic, then verify subgroup outcomes and matrix discrepancies. Our guarantees are local, concern signed surrogate-loss gaps under horizontal additive updates, and do not extend to rate-parity criteria such as equalized odds (Appendix H).

2 Related work

PEFT and subgroup performance. Adapters differ in rank, scaling, magnitude–direction split, initialization and basis (Hu et al., 2021; Kalajdzievski, 2023; Liu et al., 2024; Meng et al., 2024; Kopiczko et al., 2024), and adapter-space flatness need not imply full-space flatness (Li et al., 2024). Benchmarks find large method–task variation in utility and fairness (Dutt et al., 2023; Jin et al., 2024), no consistent ordering between LoRA and full fine-tuning in subgroup utility (Ding et al., 2024), and fairness-aware PEFT methods that select components, reweight groups or constrain representations (Dutt et al., 2024; Sukumaran et al., 2024; Kamalaruban et al.; Zhou et al., 2026).

Curvature objects and their estimators. Four matrices are often called curvature: the Hessian of the loss; the generalized Gauss–Newton matrix, which drops the Hessian's term involving second derivatives of the network; the *expected* Fisher, which averages score outer products over labels drawn from the model; and the *observed-label (empirical)* Fisher, which uses the dataset labels instead. They coincide only under specific conditions (Martens, 2020), and the empirical Fisher is generally neither the Fisher nor an unbiased estimate of it (Kunstner et al., 2019). Matrix-free products, Lanczos methods and structured approximations make curvature tractable at scale (Pearlmutter, 1994; Ghorbani et al., 2019; Yao et al., 2020; Martens & Grosse, 2015; Dangel et al., 2019); sharpness is not reparameterization invariant (Dinh et al., 2017), which matters when one predictor has many adapter coordinates. Fisher-guided PEFT uses Fisher information to choose *where* to adapt (Sung et al., 2021; Feng et al., 2026); we ask how subgroups act on a subspace once it is chosen.

Curvature and subgroups. Pruning enlarges both accuracy gaps and gaps in gradient norm or top Hessian eigenvalue (Tran et al., 2022); CUMA matches input-space curvature across groups (Wang et al., 2023); CurvFed, FairSAM and FairMerging use curvature statistics inside training or merging (Sharma et al., 2026; Dai et al., 2025; Liu et al.); FLARE uses a label-dependent Fisher penalty to discover latent groups (Roy et al., 2026). These methods use curvature *inside* a procedure and validate it empirically. We instead ask what reachable curvature *provably* controls, distinguish the empirical training proxy from the expected Fisher, and show that the trace these methods rely on mixes a mean-gradient term with a dispersion term. Appendix D gives an extended review.

3 Reachable Fisher geometry of the subgroup gap

We now make precise, and prove, the introduction's claim that different spectral properties of the reachable Fisher answer different questions about a subgroup gap, while a trace answers neither. The argument has four steps: the adapter fixes the reachable directions (Section 3.1); curvature along them raises two questions, direction and magnitude (Section 3.2); common summaries lose information both need (Section 3.3); and all of it concerns surrogate losses (Section 3.4). Section 4 then derives their computable counterparts, which Sections 5 and 6 evaluate. Proofs are in Appendix B.

3.1 Setup and the master decomposition

Adapter geometry. Frozen weights $\theta_0 \in \mathbb{R}^P$ are adapted through a smooth map $T : \mathbb{R}^p \to \mathbb{R}^P$, $p \le P$, so $\theta = T(\phi)$; full fine-tuning has $T(\phi) = \phi$ (so $J = I$), and for LoRA each adapted weight is $W = W_0 + s\,BA$ with $B \in \mathbb{R}^{d_{\text{out}} \times r}$, $A \in \mathbb{R}^{r \times d_{\text{in}}}$, a fixed nonzero scale s ($s = 2$ in our runs; Appendix E), and $\phi = (A, B)$. The Jacobian $J(\phi) = \partial T / \partial \phi$ maps adapter perturbations to first-order changes of the full model, so locally the adapter moves θ only within $\operatorname{range} J$.

Subgroup losses. Let $\mathcal{A}$ be a finite set of subgroups with distributions $\mathcal{D}_a$ over $z = (x, y)$, with input marginal $\mathcal{D}_a^X$, and write $L_a(\phi) = \mathbb{E}_{\mathcal{D}_a}[\ell(T(\phi); z)]$, $g_a = \nabla_\phi L_a$ and $H_a = \nabla_\phi^2 L_a$. The *signed gap* is $\Delta_{ab} = L_a - L_b$; exchanging a and b reverses every directional conclusion below.

Reachable Fisher. For $\ell = -\log p_\theta(y \mid x)$ let F_a be the full-model Fisher of group a. Pulling it back through J gives the *reachable Fisher*:

$$F_a = \mathbb{E}_{x \sim \mathcal{D}_a^X} \mathbb{E}_{y \sim p_\theta(\cdot|x)} \big[\nabla_\theta \log p_\theta \, \nabla_\theta \log p_\theta^\top \big], \qquad F_a^{\mathrm{R}}(\phi) = J(\phi)^\top F_a(T(\phi))\, J(\phi). \tag{1}$$

$F_a \in \mathbb{R}^{P \times P}$ is never formed; $F_a^{\mathrm{R}} \in \mathbb{R}^{p \times p}$ is the object of interest.

Assumption A2 (likelihood form). $p_\theta(y \mid x)$ is an exponential family whose natural parameter is the network output, with ℓ its negative log-likelihood, so the generalized Gauss–Newton matrix equals

F_a (Martens, 2020). A2 covers softmax cross-entropy and fixed-variance Gaussian regression but does *not* give $H_a = F_a^{\mathrm{R}}$; the smoothness assumption **A1** needed for the expansion follows below.

The horizontal subspace. The sign certificate below requires positive definiteness, which raw factorized coordinates cannot provide: they contain directions that do not change the model to first order (*gauge* directions). For rank-r LoRA with $s \neq 0$ and $r \leq \min(d_{\mathrm{in}}, d_{\mathrm{out}})$, Proposition 4 gives, with $r_A = \operatorname{rank} A$ and $r_B = \operatorname{rank} B$,

$$\dim \ker J = d_{\mathrm{out}}(r - r_A) + d_{\mathrm{in}}(r - r_B) + r_A r_B \;\geq\; r^2 \quad \text{at every } \phi,$$

and the lower bound extends to DoRA (Corollary 3). Hence $\lambda_{\min}(F_a^{\mathrm{R}} - F_b^{\mathrm{R}}) \leq 0$ everywhere. We therefore restrict every directional statement to the *horizontal subspace* $\mathcal{H}_\phi = (\ker J(\phi))^\perp$, with $\lambda_{\min}^{\mathcal{H}}(S) = \min\{v^\top S v : v \in \mathcal{H}_\phi, \|v\| = 1\}$ and $\|S\|_2^{\mathcal{H}} = \max\{|v^\top S v| : v \in \mathcal{H}_\phi, \|v\| = 1\}$, and assume $d_{\mathcal{H}} := \dim \mathcal{H}_\phi \geq 1$. Unless noted, all eigenvalues and norms in this section are these restricted ones.

Caveat: a first-order restriction. Every result concerns straight additive updates $\delta \in \mathcal{H}_\phi$. At the LoRA origin $A = B = 0$ the subspace is $\{0\}$ and the statements are vacuous, and at the standard $B = 0$ initialization two updates differing by an element of $\ker J$ can move the gap with opposite signs at second order (Remark 2); the results do not transfer to an unprojected optimizer.

Assumption A1 (smoothness). Each L_a is C^3 (three times continuously differentiable) on the ball $B(\phi, \varepsilon_0)$ of adapter settings within distance ε_0 of ϕ, with third derivatives bounded by M (Appendix A.1 explains the role of each). It is what bounds the remainder below.

Proposition 1 (Exact gap expansion)**.** *Under **A1**, for $\|\delta\| \leq \varepsilon \leq \varepsilon_0$,*

$$\Delta_{ab}(\phi+\delta) - \Delta_{ab}(\phi) = \underbrace{(g_a - g_b)^\top \delta}_{\text{slope}} + \underbrace{\tfrac{1}{2}\delta^\top C_{ab}\, \delta}_{\text{curvature}} + R_{ab}(\delta), \quad C_{ab} := H_a - H_b, \;\; |R_{ab}| \leq \tfrac{M}{3}\|\delta\|^3. \tag{2}$$

Under **A2**, $H_a = F_a^{\mathrm{R}} + \Xi_a$, where the *Hessian–Fisher defect* Ξ_a collects a reparameterization term (second derivatives of T) and a model-curvature residual, so the curvature matrix splits as

$$C_{ab} = \underbrace{(F_a^{\mathrm{R}} - F_b^{\mathrm{R}})}_{\text{reachable Fisher}} + \underbrace{(\Xi_a - \Xi_b)}_{\text{defect}}, \qquad \eta_{ab}^{\mathcal{H}} = \|\Xi_a - \Xi_b\|_2^{\mathcal{H}} \leq \eta_{ab} = \|\Xi_a - \Xi_b\|_2. \tag{3}$$

For LoRA-type adapters, the reparameterization part is bounded by $|s|\,\|\nabla_W \ell_a - \nabla_W \ell_b\|_F$ (Proposition 3), where $\nabla_W \ell_a$ is the gradient at the deployed weight $W = W_0 + sBA$ and $|s|$ is the adapter scale fixed above, so this half of the defect is small exactly when the two subgroup gradients agree there. The model-curvature residual is not bounded, so η_{ab} must be assumed small or measured. This expansion leads to two different questions: when must the curvature term have the same sign for every admissible direction, and how large can it be in the worst case?

3.2 Two questions: direction and magnitude

At a point where the subgroup slopes match ($g_a = g_b$), the problem becomes spectral. If C_{ab} is positive in every horizontal direction, every sufficiently small update increases the signed gap; if it is indefinite, the chosen direction decides the sign, so the restricted minimum eigenvalue answers the directional question. If instead we ask only how large the quadratic change can be, the relevant quantity is the largest absolute eigenvalue, the restricted operator norm. Table 1 summarizes the answers. Both describe the curvature term, which governs the gap only under slope matching: $g_a = g_b$ makes the linear term vanish for every update, not just one. This is stronger than the one-step constraint $(g_a - g_b)^\top \delta = 0$, which removes the linear term only along the chosen δ: a slope can be removed by a single constraint, whereas no nonzero update can eliminate a positive-definite quadratic form. We first state the directional result for the exact Hessian difference.

Theorem 1 (Forced increase on the horizontal subspace)**.** *Write $\rho_{\min} := \lambda_{\min}^{\mathcal{H}}(C_{ab})$ for the smallest eigenvalue of the curvature difference restricted to $\mathcal{H}_\phi$. Assume **A1**, $g_a = g_b$, and $\rho_{\min} > 0$, that is, C_{ab} is positive definite* on *$\mathcal{H}_\phi$. Then for every nonzero $\delta \in \mathcal{H}_\phi$ with $\|\delta\| \leq \min(\varepsilon_0, 3\rho_{\min}/4M)$, where ε_0, M are from **A1** and the second bound is $+\infty$ when $M = 0$,*

$$\Delta_{ab}(\phi + \delta) - \Delta_{ab}(\phi) \;\geq\; \tfrac{1}{4}\rho_{\min}\|\delta\|^2 \;>\; 0.$$

The signed gap therefore strictly increases for every admissible horizontal additive update. If in addition $\Delta_{ab}(\phi) \geq 0$, the absolute gap $|\Delta_{ab}|$ strictly widens.

Table 1: What controls direction and magnitude after slope matching. All quantities are restricted to $\mathcal{H}_\phi$; without slope matching, the linear term generally dominates small updates.

Question	Exact governing quantity	Reachable-Fisher certificate
Curvature term positive in every direction?	$\lambda^{\mathcal{H}}_{\min}(C_{ab}) > 0$	$\lambda^{\mathcal{H}}_{\min}(F^{\mathrm{R}}_a - F^{\mathrm{R}}_b) > \eta^{\mathcal{H}}_{ab}$
Worst-case curvature change?	$\frac{1}{2}\|C_{ab}\|^{\mathcal{H}}_2 \varepsilon^2$	at most $\frac{1}{2}\big(\|F^{\mathrm{R}}_a - F^{\mathrm{R}}_b\|^{\mathcal{H}}_2 + \eta^{\mathcal{H}}_{ab}\big)\varepsilon^2$

The margin $\rho_{\min}$ controls both halves of the statement: it sets how far the conclusion extends, through the step bound $3\rho_{\min}/4M$, and how fast the gap grows, through $\frac{1}{4}\rho_{\min}\|\delta\|^2$. A larger margin buys a larger admissible step and a faster guaranteed increase; the step bound is where the cubic remainder of (2) still costs less than half the quadratic term (Appendix J reports a measured value). Each assumption is necessary. Slope matching removes the first-order channel; the horizontal restriction removes parameterization-null directions; and the sign of the initial gap is needed only to turn signed-gap growth into absolute-gap widening. Remark 4 and Proposition 4 give counterexamples when any assumption is dropped. The exact Hessian difference is generally unavailable; Corollary 1 replaces it with the reachable-Fisher difference at the price of the defect tolerance.

Corollary 1 (Reachable-Fisher sufficient condition)**.** *Write $\eta^{\mathcal{H}}_{ab} = \|\Xi_a - \Xi_b\|^{\mathcal{H}}_2 \le \eta_{ab}$. If $\lambda^{\mathcal{H}}_{\min}(F^{\mathrm{R}}_a - F^{\mathrm{R}}_b) > \eta^{\mathcal{H}}_{ab}$ then $\lambda^{\mathcal{H}}_{\min}(C_{ab}) > 0$; if* additionally *$g_a = g_b$, then Theorem 1 applies with margin $\rho_{\min} \ge \lambda^{\mathcal{H}}_{\min}(F^{\mathrm{R}}_a - F^{\mathrm{R}}_b) - \eta^{\mathcal{H}}_{ab}$. By Proposition 4 the restriction to $\mathcal{H}_\phi$ is necessary: the unrestricted condition $\lambda_{\min}(F^{\mathrm{R}}_a - F^{\mathrm{R}}_b) > \eta_{ab}$ is unsatisfiable for factorized adapters.*

Takeaway. Reachable-Fisher definiteness certifies the sign only when its restricted minimum eigenvalue exceeds the Hessian–Fisher defect; it says nothing about magnitude.

Theorem 2 (Worst-case growth of the two channels)**.** *Over $\delta \in \mathcal{H}_\phi$ with $\|\delta\| \le \varepsilon$, let $\mathcal{Q}_{ab}(\varepsilon) = \max \frac{1}{2}|\delta^\top C_{ab}\delta|$ be the worst-case second-order term and $\mathcal{G}_{ab}(\varepsilon) = \max |\Delta_{ab}(\phi+\delta) - \Delta_{ab}(\phi)|$ the worst-case gap change. Then*

$$\mathcal{Q}_{ab}(\varepsilon) = \tfrac{1}{2}\|C_{ab}\|^{\mathcal{H}}_2 \varepsilon^2, \qquad \mathcal{G}_{ab}(\varepsilon) \le \|g_a - g_b\|\varepsilon + \tfrac{1}{2}\|C_{ab}\|^{\mathcal{H}}_2 \varepsilon^2 + \tfrac{M}{3}\varepsilon^3,$$

the first attained at a dominant eigenvector of C_{ab} restricted to $\mathcal{H}_\phi$. If $g_a = g_b$ the bound is tight to third order: $|\mathcal{G}_{ab}(\varepsilon) - \frac{1}{2}\|C_{ab}\|^{\mathcal{H}}_2 \varepsilon^2| \le \frac{M}{3}\varepsilon^3$.

Takeaway. By (3), $\|C_{ab}\|^{\mathcal{H}}_2$ and $\|F^{\mathrm{R}}_a - F^{\mathrm{R}}_b\|^{\mathcal{H}}_2$ differ by at most $\eta^{\mathcal{H}}_{ab}$, so matching reachable Fishers does less than matching Hessians: it leaves a second-order gap change of up to $\frac{1}{2}\eta_{ab}\varepsilon^2$ (Theorem 6; exact Hessian matching is treated in Theorem 7).

3.3 What practical summaries can reveal

Two distinct failures separate a practical statistic from the quantities above: it may be computed in the wrong space, or it may scalarize away the directional information that survives. We take them in turn, then give a matrix-level remedy. We write $S \succeq 0$ for positive semidefinite S. A *Jacobian-blind* statistic is computed from the full-model Fishers F_a, F_b without J, such as a norm of $F_a - F_b$ or a spectral summary computed separately for each group. Such a statistic cannot determine reachable geometry because it does not know how the adapter Jacobian is oriented: a large full-model difference may lie entirely outside $\mathrm{range}\, J$, while two full-model Fishers with identical spectra may interact differently with it.

Theorem 3 (Full-model summaries do not determine reachable geometry)**.** (i) *For every $c > 0$ there exist $F_a, F_b \succeq 0$ and a Jacobian J with $\|F_a - F_b\|_2 = c$ yet $F^{\mathrm{R}}_a = F^{\mathrm{R}}_b$.* (ii) *There exist $F_a, F_b \succeq 0$ with identical spectra such that $F^{\mathrm{R}}_a - F^{\mathrm{R}}_b$ is positive definite on $\mathcal{H}_\phi$.*

A full-model audit can thus report a large disparity of which nothing survives the projection, or miss a reachable difference that meets the premise of Corollary 1; without J, no such statistic identifies the reachable projection.

Moving to the correct space is not enough. A trace records the sum of eigenvalues, not how curvature is distributed across directions, so two reachable Fishers can have equal traces and very different geometry.

Theorem 4 (Limits of trace matching). *For $F_a^{\mathrm{R}}, F_b^{\mathrm{R}} \succeq 0$ of size $p \geq 2$:* (i) Necessary condition. $F_a^{\mathrm{R}} = F_b^{\mathrm{R}}$ *implies* $\operatorname{tr} F_a^{\mathrm{R}} = \operatorname{tr} F_b^{\mathrm{R}}$*; the converse fails.* (ii) General failure. *For any $\tau > 0$,* $\sup\{\|F_a^{\mathrm{R}} - F_b^{\mathrm{R}}\|_2 : \operatorname{tr} F_a^{\mathrm{R}} = \operatorname{tr} F_b^{\mathrm{R}} = \tau\} = \tau$*, attained by an indefinite difference: equal traces permit the largest spectral mismatch the trace budget allows.* (iii) Structured positive result. *If $F_a^{\mathrm{R}} - F_b^{\mathrm{R}} \succeq 0$ (Loewner ordering), then $\|F_a^{\mathrm{R}} - F_b^{\mathrm{R}}\|_2 \leq \operatorname{tr} F_a^{\mathrm{R}} - \operatorname{tr} F_b^{\mathrm{R}}$, so equal traces force equal matrices.*

Corollary 2 (A small normalized trace gap excludes the sufficient regime). *If $\lambda_{\min}^{\mathcal{H}}(F_a^{\mathrm{R}} - F_b^{\mathrm{R}}) > 0$ then $\operatorname{tr} F_a^{\mathrm{R}} - \operatorname{tr} F_b^{\mathrm{R}} \geq d_{\mathcal{H}}\, \lambda_{\min}^{\mathcal{H}}(F_a^{\mathrm{R}} - F_b^{\mathrm{R}})$. Hence $|\operatorname{tr} F_a^{\mathrm{R}} - \operatorname{tr} F_b^{\mathrm{R}}|/d_{\mathcal{H}} \leq \eta_{ab}^{\mathcal{H}}$ excludes the hypothesis of Corollary 1 in both orientations.*

Trace therefore has two asymmetric uses. A sufficiently small normalized trace gap rules out the sufficient positive-definiteness regime; without Loewner ordering it cannot certify matrix matching or bound an indefinite mismatch. The observed-label trace (of the empirical reachable Fisher $\widehat{F}_a^{\mathrm{R,obs}}$, the average of $\nabla_\phi \ell_i \nabla_\phi \ell_i^\top$ over group a's examples, with mean gradient $\hat{g}_a$ and gradient covariance $\widehat{\Sigma}_a$) also mixes the channels: $\operatorname{tr} \widehat{F}_a^{\mathrm{R,obs}} = \|\hat{g}_a\|^2 + \operatorname{tr} \widehat{\Sigma}_a$, a mean-gradient term that is related to but does not determine $g_a - g_b$, plus within-group dispersion (Lemma 2). A matrix discrepancy avoids the second failure, and can be evaluated without forming a $p \times p$ matrix.

Proposition 2 (Matrix-free Gram identity). *For groups with n_a and n_b examples, any factors $G_a \in \mathbb{R}^{m_a \times p}$, $G_b \in \mathbb{R}^{m_b \times p}$ (one or more rows per example) and $\widehat{S}_c = \frac{1}{n_c} G_c^\top G_c$,*

$$\left\|\widehat{S}_a - \widehat{S}_b\right\|_F^2 = \frac{\|G_a G_a^\top\|_F^2}{n_a^2} - \frac{2\|G_a G_b^\top\|_F^2}{n_a n_b} + \frac{\|G_b G_b^\top\|_F^2}{n_b^2}, \tag{4}$$

computable in $O\big((m_a^2 + m_a m_b + m_b^2)p\big)$ time with no $p \times p$ matrix formed.

Per-example gradients supply such factors for both empirical reachable Fishers (Appendix C), which is what lets Section 6 audit trained models at the matrix level. The Frobenius discrepancy upper-bounds the operator norm, and so bounds the curvature channel up to defect and estimator error (Corollary 4); like any norm, it does not certify definiteness.

3.4 Scope: surrogate-loss gaps versus predictive-performance gaps

What do these loss-geometric results imply for predictive-performance gaps? Less than a direct translation: surrogate-loss differences do not determine balanced accuracy, mean absolute error, or rate-based criteria. The result below gives the bridge available under classification calibration and near-Bayes convergence. For *binary* classification let R_a be group-a 0–1 risk, $R_a^\star$ its Bayes risk, ℓ classification-calibrated (driving its excess risk to zero drives the excess 0–1 risk to zero) with calibration function ψ (which converts excess surrogate risk into excess 0–1 risk), and $\epsilon_a = L_a - L_a^\star$, where $L_a^\star$ is the group-a Bayes surrogate risk over all measurable predictors, not the minimum over the PEFT class (Bartlett et al., 2006).

Theorem 5 (Near-Bayes convergence, not a floor). *For every ϕ, with $\gamma_{ab} = |R_a^\star - R_b^\star|$,*

$$\big|\, |R_a - R_b| - \gamma_{ab} \,\big| \;\leq\; \max\{\psi^{-1}(\epsilon_a), \psi^{-1}(\epsilon_b)\},$$

and hence

$$|R_a - R_b| \;\geq\; \max\big\{0,\; \gamma_{ab} - \max_c \psi^{-1}(\epsilon_c)\big\}.$$

Consequently $|R_a - R_b| \to \gamma_{ab}$ along any sequence of predictors whose excess surrogate risks both tend to zero.

Near Bayes optimality the risk gap approaches the Bayes-risk difference, which is a limit, not a floor (Remark 8). The result extends to balanced error and multi-group maxima (Remark 9), not to multiclass balanced accuracy or MAE; Appendix A.4 gives the full scope.

4 From theory to computable statistics

Section 3 identifies matrix quantities governing local subgroup-gap changes, but a dense $p \times p$ reachable Fisher is infeasible at scale: with $p = 1.6$M, one requires about 10 TB. We compute

three matrix-free statistics from per-example gradients: a model-expected trace for diagnosis, an observed-label trace for training, and an observed-label Frobenius discrepancy for the held-out audit; the workflow is to match cheaply, then verify at the matrix level. The experiments evaluate these statistics, not the theorems' conditions: they do not establish slope matching, horizontal updates, or a small Hessian–Fisher defect. Theory addresses surrogate-loss gaps while applications prioritize predictive performance, so Sections 5 and 6 report both.

4.1 Post-training diagnostic: does trace rank subgroup difficulty?

The diagnostic tests whether subgroups to which the fitted model is more locally sensitive are also harder for that model. For each input we compute the model-expected reachable-Fisher trace contribution with respect to ϕ,

$$t_i = \begin{cases} \sum_{c=1}^{C} q_\theta(c \mid x_i) \|\nabla_\phi \log q_\theta(c \mid x_i)\|_2^2, & \text{categorical,} \\ \sigma^{-2} \|\nabla_\phi \mu_\theta(x_i)\|_2^2, & \text{Gaussian regression,} \end{cases} \tag{5}$$

where $q_\theta(c \mid x_i)$ is the predicted probability of class c among C classes and $\mu_\theta(x_i)$ the predicted mean, with σ^2 estimated from held-out residuals; we average within subgroups to obtain $\widehat{t}_a = \operatorname{tr} \widehat{F}_a^{\mathrm{R,exp}}$, the trace of a sample estimate of F_a^{R} in (1). For each fixed model, data split and subgroup axis (the attribute defining the groups, e.g. skin tone), we correlate these traces with subgroup difficulties across the groups on that axis:

$$r = \operatorname{corr}_a(\widehat{t}_a, d_a), \qquad d_a = 1 - \mathrm{BAcc}_a \text{ (classif.)}, \quad d_a = \mathrm{MAE}_a \text{ (regr.)}, \tag{6}$$

with balanced accuracy (BAcc) and mean absolute error (MAE) computed per group on the same evaluation split as $\widehat{t}_a$. This is an *aligned* Pearson correlation: $r > 0$ means higher trace accompanies greater difficulty. Because a correlation across two groups is necessarily ± 1, it is reported only for axes with at least three groups. Correlations are averaged after Fisher's z transformation. Correlation measures alignment; the coefficient of variation (CV) of $\{\widehat{t}_a\}$ separately measures how unevenly trace is spread across groups; it is comparable only within a fixed model and parameterization (Lemma 1).

4.2 Training intervention: can trace matching narrow subgroup gaps?

The intervention tests whether equalizing the subgroup traces during training narrows their performance gap. A group-balanced sampler draws equally from the constrained groups, reducing sampling noise for rare groups. Let $u_i = \nabla_\phi \ell_i$ be the per-example observed-label gradient, via vectorized automatic differentiation, and let $\mathcal{A}_B$ denote the groups in batch B. For each group, $c_a = \frac{1}{n_a} \sum_{i \in a} \|u_i\|_2^2 = \operatorname{tr} \widehat{F}_a^{\mathrm{R,obs}}$. We compare c_a with the detached mean $c_{\mathrm{ref}} = \mathrm{sg}\big[|\mathcal{A}_B|^{-1} \sum_{a \in \mathcal{A}_B} c_a\big]$ through $e_a = (c_a - c_{\mathrm{ref}}) / \max\{c_{\mathrm{ref}}, \epsilon_{\mathrm{den}}\}$, which normalizes traces that vary by orders of magnitude and avoids division by zero. The penalty $R(\phi, \mu) = \sum_a \mu_a \rho_\kappa(e_a)$ uses the Huber loss with $\kappa = 1$, quadratic near matching and linear for outlying estimates. We optimize

$$\mathcal{L}_{\mathrm{task}}(\phi) + \omega R(\phi, \mu). \tag{7}$$

The weights μ follow a bounded exponentiated-gradient update on moving averages of e_a, giving more weight to groups that remain above the reference. With $g_{\mathrm{task}} = \nabla_\phi \mathcal{L}_{\mathrm{task}}$ and $g_{\mathrm{pen}} = \nabla_\phi R$, we set $\omega = \min\{1, \beta \|g_{\mathrm{task}}\| / \|g_{\mathrm{pen}}\|\}$ with $\beta = \frac{1}{2}$. This caps the penalty-gradient norm at half the task-gradient norm and guarantees $g_{\mathrm{task}}^\top (g_{\mathrm{task}} + \omega g_{\mathrm{pen}}) \geq \frac{1}{2} \|g_{\mathrm{task}}\|_2^2$, so before AdamW preconditioning the penalty may rotate but cannot reverse the task gradient (Appendix C reports uncapped behavior). Retaining the task loss prevents the solution in which R is reduced by making the predictor insensitive.

What the intervention does not guarantee. The penalty acts on the trace, which leaves three gaps between it and the theory. It does not enforce slope matching, so the linear term of (2) is untouched, and $c_a = \|\hat{g}_a\|^2 + \operatorname{tr} \widehat{\Sigma}_a$ mixes a mean-gradient term with within-group dispersion (Lemma 2). It acts on the observed-label empirical Fisher rather than the model-expected Fisher of the theory. And a small trace gap bounds a matrix discrepancy only under Loewner ordering (Theorem 4(iii)), which we do not measure. Its effect therefore has to be checked at the matrix level: after training we measure, for the baseline and trace-matched models on the same held-out group-balanced batches, the trace gap and the Frobenius discrepancy of Proposition 2, which bounds the curvature channel up to defect and estimator error (Corollary 4); the audit protocol is in Appendix C.4.

4.3 DATASETS AND MODEL GRID

Table 2 lists the four public datasets and the model grid. The 17 pretrained image encoders span supervised, self-supervised, convolutional, hierarchical and vision–language pretraining, each trained with full fine-tuning, LoRA, DoRA, rsLoRA, PiSSA and VeRA (PiSSA and VeRA cannot attach to ResNet-50), giving 306 models; training details are in Appendix E.

Table 2: Diagnostic datasets and model grid. A subgroup axis is the attribute defining the groups; group counts in parentheses. Full details in Table 5.

Dataset	Task	Subgroup axes (groups)	Models
Fitzpatrick17k (Groh et al., 2021)	3-class diagnosis	skin tone (7), binary tone (3)	17 encoders
UTKFace (Zhang et al., 2017)	age regression	race (5)	17 encoders
FHIBE face (Xiang et al., 2025)	age regression	ancestry (6)	17 encoders
Diabetes, 130 hospitals (Strack et al., 2014)	30-day readmission	race (6)	FT-Transformer (Gorishniy et al., 2021)

4.4 STATISTICAL UNITS AND UNCERTAINTY

This subsection defines the 306-model diagnostic sweep; Section 6 describes the 30 paired intervention cells separately. A diagnostic *cell* is one trained model evaluated on one split and one subgroup axis (the "train" split is an internal slice of the training partition never used for gradient updates, so all splits are out-of-fit), for example a ViT–LoRA model on the UTKFace test split across race groups; each eligible cell contributes one correlation (6). Excluding two-group axes leaves 1,218 cells. The *positive-sign rate* is the fraction of cells with $r > 0$. Cells are not independent (they share encoders, datasets and splits), so cell shares are descriptive and the 17 encoders are the repeated unit for image-task uncertainty (2,000-replicate bootstrap). The performance gap of a model on an axis is $\max_a d_a - \min_a d_a$. The sweep uses one training seed.

5 RESULTS

Reachable Fisher tracks subgroup difficulty, heterogeneously. Across 1,218 cells, r is positive in 75.7% and has a descriptive Fisher-z mean of 0.637. Averaging encoders equally (Table 3), alignment is strongest for UTKFace (0.930), intermediate for FHIBE face (0.601), positive but architecture-sensitive for Fitzpatrick17k (0.320), where one encoder is strongly negative (Figure 2, Appendix F), and weak for the single tabular architecture (0.117), which has no encoder-level interval. Replacing task difficulty with the surrogate loss, the quantity the theorems concern, strengthens the pooled association to 85.5% positive over 406 test-split cells, though not uniformly: Fitzpatrick17k nearly doubles to 0.611 while FHIBE face falls to 0.411. Method-specific and normalization checks are in Appendix F. The sweep therefore supports the ranking premise behind the penalty but does not test a theorem; by Lemma 1, traces are comparable only within a fixed model and parameterization.

Table 3: Subgroup Fisher–difficulty alignment. Image-task means weight encoders equally, with encoder-bootstrap 95% intervals; sign rates are descriptive. Surrogate-loss results use the test split only, which is why their cell counts are smaller. The pooled row is a cell-level Fisher-z mean and is not comparable with the encoder-cluster means.

	task-metric difficulty			surrogate-loss difficulty		
Setting	cells	positive sign	encoder-cluster mean r [95% CI]	cells	positive sign	mean r
Diabetes FT-Transformer	18	61.1%	0.117 [not estimable]	6	66.7%	+0.284
Fitzpatrick17k	600	59.3%	0.320 [0.184, 0.424]	200	81.5%	+0.611
UTKFace	300	98.3%	0.930 [0.901, 0.950]	100	98.0%	+0.886
FHIBE face	300	86.7%	0.601 [0.511, 0.679]	100	82.0%	+0.411
All cells (pooled)	1,218	75.7%	0.637	406	85.5%	+0.668

6 DOES REACHABLE-FISHER TRACE MATCHING NARROW SUBGROUP GAPS?

Arms and baseline. We compare the trace-matching objective of Section 4.2 with *group-balanced task-loss training*: the same objective without the penalty, using the same sampler, schedule, seed,

and initialization. The sampler draws equal counts from every constrained group and the task loss is the unweighted mean over the batch, so both arms minimize the group-balanced risk (uniform p_a); pooled utility is nevertheless reported over all test examples, so it is a population quantity. Because the two arms share the sampler and the loss, their difference isolates the curvature penalty rather than the effect of oversampling.

Protocol. Three image tasks (Fitzpatrick17k, UTKFace, and the face age-regression task of FHIBE) are crossed with ConvNeXt, ViT and SigLIP 2 encoders and LoRA, DoRA and PiSSA adapters; a tabular task (diabetes readmission) is crossed with the same adapters on an FT-Transformer. This gives 30 cells, each trained once per arm. Batch size is tuned per cell as a hyperparameter (64–1024); both arms of a cell share it. The retained checkpoint is selected by the pooled validation metric, never by a subgroup gap (Appendix C). The primary outcome is the test-split best–worst subgroup gap G in balanced accuracy or MAE, summarized as $R_{\mathrm{gap}} = 100\,(G_{\mathrm{P}} - G_{\mathrm{H}})/G_{\mathrm{P}}$ (P plain, H Huber-penalized trace matching; positive favours trace matching). The secondary outcome is pooled utility over all test examples. EOD and AOD are in Appendix G.

Results. Cells share tasks, encoders and adapters and are single-seed, so all statistics are descriptive.

- *Gap.* Trace matching narrows the gap in all 30 cells, with median $R_{\mathrm{gap}} = 17.2\%$ (task-blocked bootstrap $[11,\ 29]\%$). The effect is present in every task and adapter: medians are 25.7% (Fitzpatrick17k), 14.3% (UTKFace), 11.1% (FHIBE face) and 23.0% (diabetes), and 17.2%, 17.3% and 15.5% for LoRA, DoRA and PiSSA.
- *Utility.* 23 cells narrow the gap without a loss of pooled utility and 7 at a small cost (Figure 3); median changes are $+0.5$ pp balanced accuracy and -0.09 years MAE. The largest classification loss is -2.4 pp (Fitzpatrick17k/ViT/PiSSA), where the gap falls from 0.260 to 0.139.
- *Surrogate-loss gap.* The gap in the training loss, the quantity the theorems concern, narrows in 27 of 30 cells (median -22.8%, bootstrap $[-32.2, -16.1]\%$) and moves with the performance gap (Spearman $\rho = +0.57$).
- *Matrix audit.* On identical held-out group-balanced batches (Appendix C), trace matching lowers the normalized trace gap in 20 of 30 cells (median -12.0%) and the observed-label empirical operator discrepancy, which Theorem 4(ii) says the trace cannot bound, in 23 (median -10.2%, bootstrap $[-18.5, -3.5]\%$). The primary unbiased squared-Frobenius statistic falls in only 16 (median -4.3%, interval spanning zero): of the 20 cells whose trace gap fell, it rose in 9. Using the plug-in norms, the median operator-to-Frobenius ratio is 0.72, so that bound is loose by about a third. Per-cell values are in Table 13.
- *Exploratory association.* Cells whose expected-Fisher CV fell more also narrowed the gap more (Spearman $\rho = +0.23$ on relative changes; the CV fell in 25 of 30 cells).

Reading the results against the theory. Theorem 1 motivates the penalty: under slope matching, a reachable-Fisher difference definite above the defect forces the signed gap up, so shrinking it is the lever measurable at scale. A trace penalty reaches that condition only under Loewner ordering (Theorem 4 (iii)), which we do not measure, so these outcomes are compatible with the theory without being predicted by it. Appendix G.1 compares adapter and full-model penalty scope.

7 Discussion and conclusion

Subgroup behaviour under PEFT depends on the curvature visible through the adapter, and the sign and magnitude of a gap change need different matrix information. The theory is sharp but conditional by construction: it needs slope matching, horizontal updates, and a small defect. On a small LoRA network where every matrix is formed exactly, the expansion and the worst-case term behave as proved while the reachable-Fisher certificate fails, the defect being large (Appendix J); the distance between the theory's objects and the computable ones is measured here, not assumed (Appendix A.4). Empirically, the trace orders subgroup difficulty within a model and, as a training signal, narrows gaps modestly and unevenly; we recommend it within a fixed parameterization, alongside direct outcome measurements. **Scope and outlook.** Closing that distance is the natural next step: estimating the defect directly, moving beyond the in-distribution regime, where Theorem 5 leaves loss geometry least room, and training against the audited discrepancy rather than its trace.

AI USE STATEMENT

We used generative AI tools for language editing, LaTeX debugging, and consistency checks against experimental outputs. All AI-assisted work was reviewed and verified by the authors, who take full responsibility for the final content.

ETHICS STATEMENT

This is a secondary analysis of existing benchmark datasets and stored model outputs; it collects no new human-subject data. The attributes labeled skin tone, race, sex, and gender are dataset-specific measurements or annotations and should not be treated as interchangeable biological categories. Subgroup estimates for small cells can be unstable, and a diagnostic association must not be interpreted as evidence that a model is safe or equitable in deployment. Any release must follow the original datasets' licenses and governance requirements. Models evaluated here are research artifacts and are not validated for clinical use.

REPRODUCIBILITY STATEMENT

All the code used in this study is made available through this Anonymous GitHub repository.[1]

Section 3 states the main formal results and their assumptions; Appendix A gives the extended statements and scope, and Appendix B provides complete proofs and machine-precision checks of the constructed examples and algebraic identities. Section 4 specifies the model grid, metrics and statistical unit; Appendix G lists every paired run and documents how the reported results are regenerated; Appendix C specifies the training objective and matrix-audit procedure; and Appendix J presents the exact small-network verification. Appendix E records checkpoint identifiers, splits, estimator details, hyperparameters, seeds, and exclusions. The supplemental repository contains the experiment runner, saved outputs, analysis materials, and the materials needed to regenerate the reported tables and figures.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pp. 60–69. PMLR, 2018.

Anissa Alloula, Charles Jones, Ben Glocker, and Bartłomiej W Papież. Subgroups matter for robust bias mitigation. *arXiv preprint arXiv:2505.21363*, 2025.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Mahesh Bhosale, Abdul Wasi, Shantam Srivastava, Shifa Latif, Tianyu Luan, Mingchen Gao, David Doermann, and Xuan Gong. Fairllava: Fairness-aware parameter-efficient fine-tuning for large vision-language assistants. *arXiv preprint arXiv:2603.26008*, 2026.

[1] https://anonymous.4open.science/r/Optimization-Fisher-With-Low-Ranked-Adaptation-In-Deep-Learning-Models-461E

Alceu Bissoto, Trung-Dung Hoang, Tim Flühmann, Susu Sun, Christian F Baumgartner, and Lisa M Koch. Subgroup performance analysis in hidden stratifications. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 594–603. Springer, 2025.

Aleksandar Botev, Hippolyt Ritter, and David Barber. Practical gauss-newton optimisation for deep learning. In *International Conference on Machine Learning*, pp. 557–565. PMLR, 2017.

Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 347–356. IEEE, 2021.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, Jim Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, et al. Meta clip 2: A worldwide scaling recipe. *Advances in Neural Information Processing Systems*, 38:48009–48036, 2026.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International conference on machine learning*, pp. 2189–2200. PMLR, 2021.

Yucong Dai, Jie Ji, Xiaolong Ma, and Yongkai Wu. Fairsam: Fair classification on corrupted image data through sharpness-aware minimization. *arXiv preprint arXiv:2503.22934*, 2025.

Felix Dangel, Frederik Kunstner, and Philipp Hennig. Backpack: Packing more into backprop. *arXiv preprint arXiv:1912.10985*, 2019.

Zhoujie Ding, Ken Ziyu Liu, Pura Peetathawatchai, Berivan Isik, and Sanmi Koyejo. On fairness of low-rank adaptation of large models. *arXiv preprint arXiv:2405.17512*, 2024.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028. PMLR, 2017.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Raman Dutt, Linus Ericsson, Pedro Sanchez, Sotirios A Tsaftaris, and Timothy Hospedales. Parameter-efficient fine-tuning for medical image analysis: The missed opportunity. *arXiv preprint arXiv:2305.08252*, 2023.

Raman Dutt, Ondrej Bohdal, Sotirios Tsaftaris, and Timothy Hospedales. Fairtune: Optimizing parameter efficient fine tuning for fairness in medical image analysis. In *International Conference on Learning Representations*, volume 2024, pp. 50254–50277, 2024.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024.

Zhi-Quan Feng, Ying-Jia Lin, and Hung-Yu Kao. Learning in the fisher subspace: A guided initialization for lora fine-tuning. *arXiv preprint arXiv:2605.01046*, 2026.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pp. 2232–2241. PMLR, 2019.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in neural information processing systems*, 34:18932–18943, 2021.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1820–1828. IEEE, 2021.

Dezheng Han and Shuaishuai Guo. Filora: Parameter-efficient fine-tuning with fisher information-guided low-rank adaptation. *IEEE Signal Processing Letters*, 2026.

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

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International conference on machine learning*, pp. 1929–1938. PMLR, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International conference on machine learning*, pp. 1939–1948. PMLR, 2018.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pp. 289–305. Springer, 2024.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of bayesian neural nets via local linearization. In *International conference on artificial intelligence and statistics*, pp. 703–711. PMLR, 2021.

Cheongjae Jang, Sungyoon Lee, Frank Park, and Yung-Kyun Noh. A reparametrization-invariant sharpness measure based on information geometry. *Advances in neural information processing systems*, 35:27893–27905, 2022.

Ruinan Jin, Zikang Xu, Yuan Zhong, Qingsong Yao, Qi Dou, S Kevin Zhou, and Xiaoxiao Li. Fairmedfm: fairness benchmarking for medical imaging foundation models. *Advances in Neural Information Processing Systems*, 37:111318–111357, 2024.

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with lora. *arXiv preprint arXiv:2312.03732*, 2023.

Parameswaran Kamalaruban, Mark Anderson, Stuart Burrell, Maeve Madigan, Piotr Skalski, and David Sutton. Fairness-aware low-rank representation fine-tuning. In *Trustworthy AI for Good (AI4GOOD) Workshop@ ICML 2026*.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pp. 2564–2572. PMLR, 2018.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254, 2019.

Minyoung Kim, Da Li, Shell X Hu, and Timothy Hospedales. Fisher sam: Information geometry and sharpness aware minimisation. In *International Conference on Machine Learning*, pp. 11148–11161. PMLR, 2022.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. Pmlr, 2017.

Dawid Kopiczko, Tijmen Blankevoort, and Yuki Asano. Vera: Vector-based random matrix adaptation. In *International Conference on Learning Representations*, volume 2024, pp. 6815–6835, 2024.

Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical fisher approximation for natural gradient descent. *Advances in neural information processing systems*, 32, 2019.

Tyler LaBonte, John C Hill, Xinchen Zhang, Vidya Muthukumar, and Abhishek Kumar. The group robustness is in the details: revisiting finetuning under spurious correlations. *Advances in Neural Information Processing Systems*, 37:121598–121629, 2024.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. *Advances in neural information processing systems*, 2, 1989.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Tao Li, Zhengbao He, Yujun Li, Yasheng Wang, Lifeng Shang, and Xiaolin Huang. Flat-lora: Low-rank adaptation over a flat loss landscape. *arXiv preprint arXiv:2409.14396*, 2024.

Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. In *The 22nd international conference on artificial intelligence and statistics*, pp. 888–896. PMLR, 2019.

Bing Liu, Xinrui Shan, Boyu Zhang, Qiankun Zhang, Bin Yuan, Jing Wang, and Xianjun Deng. Fairmerging: Rethinking model merging through the lens of fairness. In *Forty-third International Conference on Machine Learning*.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International conference on machine learning*, pp. 6781–6792. PMLR, 2021a.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9992–10002. Ieee, 2021b.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11966–11976. IEEE, 2022.

Yingtao Luo, Zhixun Li, Qiang Liu, and Jun Zhu. Fairness without demographics through learning graph of gradients. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 918–926, 2025.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pp. 2408–2417. PMLR, 2015.

James Martens et al. Deep learning via hessian-free optimization. In *Icml*, volume 27, pp. 735–742, 2010.

Michael S Matena and Colin Raffel. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37: 121038–121072, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Juneyoung Park, Seongbae Lee, Han-Sang Lee, Kyuho Lee, Minjae Kim, Seungheon Hyeon, Kiduk Kwon, Seongwan Kim, and Jaeho Lee. Fora: Fisher-orthogonal rank adaptation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2605.29317*, 2026.

Barak A Pearlmutter. Fast exact multiplication by the hessian. *Neural computation*, 6(1):147–160, 1994.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *International conference on learning representations*, 2018.

Shaily Roy, Harshit Sharma, Daniel A Adler, Srijan Sen, Tanzeem Choudhury, and Asif Salekin. Ethical fairness in ubiquitous health sensing without known attributes. *arXiv preprint arXiv:2603.13373*, 2026.

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. *arXiv preprint arXiv:1911.08731*, 2019.

Harshit Sharma, Shaily Roy, and Asif Salekin. Curvfed: Curvature-aligned federated learning for fairness without demographics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 10(2):1–43, 2026.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Sidak Pal Singh and Dan Alistarh. Woodfisher: Efficient second-order approximation for neural network compression. *Advances in Neural Information Processing Systems*, 33:18098–18109, 2020.

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.

Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014(1):781670, 2014.

Rohan Sukumaran, Aarash Feizi, Adriana Romero-Sorian, and Golnoosh Farnadi. Fairlora: Unpacking bias mitigation in vision models with fairness-driven low-rank adaptation. *arXiv preprint arXiv:2410.17358*, 2024.

Yi-Lin Sung, Varun Nair, and Colin A Raffel. Training neural networks with fixed sparse masks. *Advances in neural information processing systems*, 34:24193–24205, 2021.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Cuong Tran, Ferdinando Fioretto, Jung-Eun Kim, and Rakshit Naidu. Pruning has a disparate impact on model accuracy. *Advances in neural information processing systems*, 35:17652–17664, 2022.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pp. 459–479. Springer, 2022.

Haotao Wang, Junyuan Hong, Jiayu Zhou, and Zhangyang Wang. How robust is your fairness? evaluating and sustaining fairness under unseen distribution shifts. *Transactions on machine learning research*, 2023:https–openreview, 2023.

Xuanqian Wang, Jing Li, Ivor W Tsang, and Yew-Soon Ong. Towards harmless rawlsian fairness regardless of demographic prior. *Advances in Neural Information Processing Systems*, 37:80908–80935, 2024.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16133–16142. Ieee, 2023.

Alice Xiang, Jerone TA Andrews, Rebecca L Bourke, William Thong, Julienne M LaChance, Tiffany Georgievski, Apostolos Modas, Aida Rahmattalabbi, Yunhao Ba, Shruti Nagpal, et al. Fair human-centric image dataset for ethical ai benchmarking. *Nature*, 648(8092):97–108, 2025.

Kang Xue, Ming Dong, Xinhui Tu, and Tingting He. Fish-tuning: Enhancing peft methods with fisher information. *arXiv preprint arXiv:2504.04050*, 2025.

Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE international conference on big data (Big data)*, pp. 581–590. IEEE, 2020.

Zhewei Yao, Amir Gholami, Sheng Shen, Mustafa Mustafa, Kurt Keutzer, and Michael Mahoney. Adahessian: An adaptive second order optimizer for machine learning. In *proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 10665–10673, 2021.

Chaobin Zeng and Shusheng Zhao. Fi-lora: Dynamic and lightweight low-rank adaptation guided by fisher information. *IEEE Signal Processing Letters*, 2026.

Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. In *International Conference on Learning Representations*, volume 2024, pp. 5078–5123, 2024.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 4352–4360. IEEE, 2017.

Songqi Zhou, Zeyuan Liu, and Benben Jiang. Fairnet: Dynamic fairness correction without performance loss via contrastive conditional lora. *Advances in Neural Information Processing Systems*, 38:90081–90114, 2026.

Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. Medfair: Benchmarking fairness for medical imaging. *arXiv preprint arXiv:2210.01725*, 2022.

A EXTENDED THEORY

This appendix collects the discussion that accompanies Section 3: what the restriction to $\mathcal{H}_\phi$ means, why it is forced, the remarks that follow each result, and a full statement of scope. Proofs follow in Appendix B.

A.1 READING ASSUMPTION A1

Assumption **A1** makes the second-order analysis of Section 3 precise. Each of its three parts has one job.

The ball $B(\phi, \varepsilon_0)$. This is the set $\{\phi' \in \mathbb{R}^p : \|\phi' - \phi\| \le \varepsilon_0\}$ of adapter settings within Euclidean distance ε_0 of the current point ϕ. All results are local: they concern updates δ with $\|\delta\| \le \varepsilon \le \varepsilon_0$, so that $\phi + \delta$ stays inside the ball where the smoothness and the bound M hold. The value of ε_0 need not be known; it only has to exist.

C^3 smoothness. Each subgroup loss L_a is three times continuously differentiable on the ball, so its gradient g_a, Hessian H_a and third derivative exist and vary continuously. This is what licenses a Taylor expansion to second order with an explicit third-order remainder. It holds for smooth activations and losses (for example $\tanh$, GELU and softmax cross-entropy); for piecewise-linear activations such as ReLU it holds only away from their kinks, so for such networks **A1** is an idealization.

The constant M. M bounds the third derivative on the ball: $|D^3 L_a(\phi')[\delta, \delta, \delta]| \le M\|\delta\|^3$ for all ϕ' in the ball and all δ. It measures how fast the curvature itself changes. By Taylor's theorem with Lagrange remainder, each L_a differs from its quadratic model by at most $\frac{M}{6}\|\delta\|^3$, so the gap $\Delta_{ab} = L_a - L_b$ differs from its slope-plus-curvature model by at most $\frac{M}{3}\|\delta\|^3$ (Proposition 1). A small M means the quadratic model stays accurate over larger steps. M also sets the step size in Theorem 1: for $\|\delta\| \le 3\rho_{\min}/(4M)$ the cubic remainder is too small to overturn the quadratic increase. M is not estimated in our experiments; Appendix J checks on a small network that the remainder indeed scales as $\|\delta\|^3$.

A.2 THE HORIZONTAL RESTRICTION

$\mathcal{H}_\phi = (\ker J(\phi))^\perp$ is a Euclidean choice of representative for the first-order quotient by the *entire* first-order kernel $\ker J(\phi)$, which contains the $\mathrm{GL}(r)$ gauge tangent and can be strictly larger at rank-deficient points (Proposition 4(ii)).

What this restriction does and does not mean. $\mathcal{H}_\phi$ is a first-order object. Every curvature statement of Section 3 concerns *straight additive updates* $\delta \in \mathcal{H}_\phi$, and we do *not* identify such a δ with an equivalence class of updates: Remark 2 exhibits a LoRA point at the standard $B = 0$ initialization where two raw updates differing by an element of $\ker J$, and hence representing the same first-order predictor change, produce gap changes of *opposite sign* at second order. A conclusion of the form "$\mathcal{S}_\varepsilon = \{0\}$" therefore means *no nonzero horizontal additive update*, not "no nonzero quotient class". Building a genuine second-order quotient would require a regular full-rank stratum and a smooth local horizontal section, machinery we do not develop and which in any case does not cover the $B = 0$ initialization our experiments use.

Proposition 4 is the reason every curvature statement in Section 3 is restricted to $\mathcal{H}_\phi$. Read on the full coordinate space $\mathbb{R}^p$, the hypotheses of Theorem 1 and Corollary 1 are *empty* for exactly the adapters this paper is about: the $\mathrm{GL}(r)$ symmetry $BA = (Be^{tX})(e^{-tX}A)$ guarantees a null direction, so no reachable Fisher difference is ever positive definite and no C_{ab} is either once slopes match. Restricting to horizontal additive updates removes those directions and restores satisfiability: on $\mathcal{H}_\phi$ the definiteness conditions can hold, and we verify this numerically in the supplement. The obstruction is not a technicality about degenerate points: whenever $r \le \min(d_{\text{in}}, d_{\text{out}})$, which holds for every adapter we train, (8) gives it at every ϕ. The repair, however, is correspondingly modest: it buys a well-posed class of updates to quantify over, not a quotient manifold on which to do second-order geometry (Remark 2).

A.3 Remarks on the main results

Corollary 1 is the only route from reachable Fisher geometry to the forced-increase conclusion, and it is conditional on three things: slope matching, the defect being smaller than the reachable gap *on* $\mathcal{H}_\phi$, and the restriction to $\mathcal{H}_\phi$ itself, without which Proposition 4 makes it vacuous. We do not claim $F_a^{\mathrm{R}} - F_b^{\mathrm{R}}$ is the *only* lever on C_{ab}: by Proposition 3 the defect $\Xi_a - \Xi_b$ is another, and a method that changed it would change C_{ab} too. We claim it is the lever that is measurable at scale and that the estimator of Section C.4 acts on.

Minimizing the worst-case *second-order* term is thus exactly minimizing $\|C_{ab}\|_2^{\mathcal{H}}$, which by Proposition 3 agrees with minimizing $\|F_a^{\mathrm{R}} - F_b^{\mathrm{R}}\|_2^{\mathcal{H}}$ to within $\eta_{ab}^{\mathcal{H}}$. This is an additive approximation, not an equivalence, since the defect varies with ϕ and can partly cancel the Fisher term.

Part (ii) offers one concrete explanation for why adapter-robustness audits conflict (Ding et al., 2024; Sukumaran et al., 2024; Li et al., 2024): two studies measuring full-model sharpness on models whose reachable geometries differ as in (ii) record identical curvature. We do not claim this is the cause of any particular disagreement in that literature.

Three cautions. The extremal difference $\mathrm{diag}(\tau, -\tau, 0, \dots)$ is *indefinite*, so Theorem 1 does not apply to it and no forced-increase reading is available there. $\mathcal{Q}_{ab}$ is governed by $\|C_{ab}\|_2^{\mathcal{H}}$, not by $\|F_a^{\mathrm{R}} - F_b^{\mathrm{R}}\|_2$, which is why the consequence carries $\eta_{ab}^{\mathcal{H}}$. And $\mathcal{Q}_{ab}$ is only the quadratic channel: the displayed bound is on the gap itself only because slope matching removes the linear term. Combined with the decomposition $\operatorname{tr} \widehat{F}_a^{\mathrm{R,obs}} = \|\hat{g}_a\|_2^2 + \operatorname{tr} \widehat{\Sigma}_a$ (Lemma 2), which shows a trace penalty mixes a mean-gradient term (related to, but not determining, the slope channel $g_a - g_b$) with a within-group dispersion term, this is a concrete deficiency of the objective implemented in Section C.

Three terms separate the computed penalty from the gap, and none is cosmetic. The *slope* term cannot be dropped: for a fixed-variance Gaussian model at a point where $F_a^{\mathrm{R}} = F_b^{\mathrm{R}} = H_a = H_b$ and $\widehat{\mathcal{P}} = \zeta_c = \eta_{ab} = M = 0$ but $g_a - g_b = 1$, the exact worst-case gap change is ε, while a bound omitting the slope term would give 0. The *defect* $\eta_{ab}^{\mathcal{H}}$ is bounded only in part (Proposition 3). And the *estimator* terms ζ_c are not sampling noise alone: the observed-label empirical Fisher is in general neither the model-expected Fisher nor an unbiased estimate of it (Kunstner et al., 2019), so ζ_c carries a systematic component that does not vanish with batch size. This is why the Frobenius discrepancy of Section C.4 is used as a matrix-level audit and not presented as a gap guarantee.

A.4 Scope: what is and is not established

Established, unconditionally. The projection identity and invariance classification (Lemma 1); the gauge redundancy of factorized adapters (Proposition 4); the trace extremum and the ordered-case control by the trace gap (Theorem 4) and the decomposition (Lemma 2); the Gram identity (Proposition 2); the matrix constructions of Theorem 3; and, for rate-parity criteria outside our scope, Theorem 8 in Appendix H.

Established, conditionally. For additive updates $\delta \in \mathcal{H}_\phi$, under slope matching $g_a = g_b$ and a C_{ab} positive definite on $\mathcal{H}_\phi$, every such small nonzero update increases the signed gap (Theorem 1) and the admissible set is $\{0\}$ (Theorem 7(ii)); these become statements about the reachable Fisher only through Corollary 1, which needs $\lambda_{\min}^{\mathcal{H}}(F_a^{\mathrm{R}} - F_b^{\mathrm{R}}) > \eta_{ab}^{\mathcal{H}}$. Worst-case second-order growth is $\frac{1}{2}\|C_{ab}\|_2^{\mathcal{H}}\varepsilon^2$ (Theorem 2), and the Frobenius discrepancy bounds that channel up to $\zeta_a + \zeta_b + \eta_{ab}^{\mathcal{H}}$, not the gap (Corollary 4).

Not established. That the *definiteness and forced-increase* claims have content on raw LoRA/DoRA coordinates; by Proposition 4 and Corollary 3 they do not. (Theorems 2 and 6 remain meaningful there; it is specifically positive definiteness that is unattainable.) That $\mathcal{H}_\phi$ is a second-order quotient, or that a horizontal result transfers to other representatives of the same first-order class, or to an unprojected optimizer (Remark 2). That $F_a^{\mathrm{R}} - F_b^{\mathrm{R}}$ is the *only* lever on C_{ab}; the defect is another and we bound only its reparameterization half (Proposition 3). That $\eta_{ab}^{\mathcal{H}}$ or ζ_c is small for deep networks, or that either was measured in our runs. That reducing $\widehat{\mathcal{P}}$ reduces the *loss* gap, let alone a *performance* gap. That γ_{ab} is a floor (Remark 8). That the defect bound covers DoRA, one of the three adapters in Section 6. That the trace penalty narrows gaps reliably: Section 6 finds a modest, heterogeneous effect that is not statistically established. That any scalar norm of $F_a^{\mathrm{R}} - F_b^{\mathrm{R}}$ (trace, Frobenius or spectral) certifies the definiteness condition of Corollary 1. That Theorem 5 covers

multiclass balanced accuracy or MAE; it is stated for binary 0–1 risk and, by Remark 9, binary balanced error.

The claim we defend. *For straight additive updates constrained to a horizontal complement of the local Jacobian kernel, and once the slope channel is controlled: if the reachable-Fisher difference is positive definite on that complement with restricted minimum eigenvalue above the Hessian–Fisher defect, every sufficiently small nonzero such update increases the signed subgroup loss gap. Separately, the worst-case size of the curvature channel is an operator norm; trace equality cannot control it, and the Frobenius discrepancy bounds it, at a factor-m cost and up to an estimator gap we do not bound. Neither scalar certifies the definiteness condition. Whether the curvature channel matters for a measured performance gap is not established here, and our own measurements do not show that it does.*

Open. Finite-sample error in $\widehat{F}_a^{\mathrm{R}}$ for small groups; whether a matching penalty run to convergence reaches a matched point; non-smooth adapters; bounding $\mathcal{E}_a - \mathcal{E}_b$; and distribution shift, where the relevant object is the reachable curvature under the target rather than the source distribution.

B Additional results and proofs

Statements that appear in Section 3 are restated here with their original numbers; results that appear only in this appendix are stated in full where they are proved. Where a hypothesis is load-bearing we give the counterexample that shows it cannot be dropped. All constructions, the positive ones and the counterexamples alike, are checked to machine precision by `verify_theory.py`, included in the supplementary material at the root of the submission archive, which is written so that it fails if any statement is strengthened back to a form we had to retract; running `python verify_theory.py` prints one line per check and a final verdict.

B.1 The gap expansion and the Hessian–Fisher defect

Proposition 1 (Exact gap expansion)**.** *Under **A1**, for $\|\delta\| \le \varepsilon \le \varepsilon_0$,*

$$\Delta_{ab}(\phi+\delta) - \Delta_{ab}(\phi) = \underbrace{(g_a - g_b)^\top \delta}_{\text{slope}} + \underbrace{\tfrac{1}{2}\delta^\top C_{ab}\,\delta}_{\text{curvature}} + R_{ab}(\delta), \quad C_{ab} := H_a - H_b, \;\; |R_{ab}| \le \tfrac{M}{3}\|\delta\|^3. \tag{2}$$

Proof. Apply Taylor's theorem with Lagrange remainder to L_a and to L_b separately at ϕ: for $c \in \{a, b\}$, $L_c(\phi+\delta) = L_c(\phi) + g_c^\top \delta + \tfrac{1}{2}\delta^\top H_c \delta + r_c$ with $|r_c| \le \tfrac{M}{6}\|\delta\|^3$ by **A1**. Subtracting the two expansions and setting $R_{ab} = r_a - r_b$ gives (2), with $|R_{ab}| \le \tfrac{M}{6}\|\delta\|^3 + \tfrac{M}{6}\|\delta\|^3 = \tfrac{M}{3}\|\delta\|^3$. □

Proposition 3 (Hessian–Fisher defect)**.** *Under **A2**, $H_a = F_a^{\mathrm{R}} + \Xi_a$ with*

$$\Xi_a = \underbrace{\sum\nolimits_k [\nabla_\theta \ell_a]_k \nabla_\phi^2 T_k}_{\mathcal{T}_a \text{ (reparameterization)}} + \underbrace{J^\top(\nabla_\theta^2 \ell_a - F_a)J}_{\mathcal{E}_a \text{ (model curvature)}}, \qquad C_{ab} = (F_a^{\mathrm{R}} - F_b^{\mathrm{R}}) + (\Xi_a - \Xi_b).$$

Write $\eta_{ab} = \|\Xi_a - \Xi_b\|_2$. For LoRA, rsLoRA and PiSSA, $T(\phi) = \theta_0 + s\,BA$ with fixed scale s ($s = \alpha/r$ for LoRA and PiSSA, $s = \alpha/\sqrt{r}$ for rsLoRA); let $\nabla_W \ell_a$ denote the gradient with respect to the deployed *weight $W = W_0 + sBA$, i.e. before s is absorbed. Then, for a single adapted block,*

$$\|\mathcal{T}_a - \mathcal{T}_b\|_2 \;\le\; |s|\,\|\nabla_W \ell_a - \nabla_W \ell_b\|_F,$$

and with several blocks the bound holds with the maximum of the right-hand side over blocks.

We bound only the reparameterization half of the defect. The model-curvature residual $\mathcal{E}_a - \mathcal{E}_b$ is *not* bounded here: it vanishes for models linear in θ and is small near interpolation, but for a deep network we treat η_{ab} as a quantity to be assumed small or measured, not one we have shown to be small. For DoRA the magnitude–direction split contributes a further term of the same type which we also do not bound. This is the principal gap between the geometry we can compute, $F_a^{\mathrm{R}} - F_b^{\mathrm{R}}$, and the geometry that governs the expansion, C_{ab}.

Proof. Differentiating $L_a(\phi) = \ell_a(T(\phi))$ twice by the chain rule gives

$$H_a = J^\top \big(\nabla^2_\theta \ell_a\big) J + \sum\nolimits_{k=1}^{P} \big[\nabla_\theta \ell_a\big]_k \nabla^2_\phi T_k.$$

Adding and subtracting $J^\top F_a J = F_a^{\mathrm{R}}$ inside the first term gives $H_a = F_a^{\mathrm{R}} + \mathcal{T}_a + \mathcal{E}_a$ with $\mathcal{T}_a, \mathcal{E}_a$ as stated, where **A2** is used only to identify F_a with the generalized Gauss–Newton matrix so that $\mathcal{E}_a = J^\top(\nabla^2_\theta \ell_a - F_a)J$ is exactly the model-curvature residual. Subtracting the identity for b gives the expression for C_{ab}.

For the LoRA-family bound, fix one adapted block with $T(\phi) = \theta_0 + s\,BA$, $W = W_0 + sBA$. The map is bilinear in $\phi = (A, B)$, and its second differential in direction $(\dot A, \dot B)$ is $2s\,\dot B \dot A$, so

$$\mathcal{T}_a[(\dot A, \dot B), (\dot A, \dot B)] = 2s\,\langle \nabla_W \ell_a, \dot B \dot A \rangle,$$

with $\nabla_W \ell_a$ the gradient with respect to the deployed weight. For $\|\dot A\|_F^2 + \|\dot B\|_F^2 = 1$, submultiplicativity and $2\|\dot B\|_F\|\dot A\|_F \le \|\dot A\|_F^2 + \|\dot B\|_F^2 = 1$ give $\|\dot B \dot A\|_F \le \frac{1}{2}$, so Cauchy–Schwarz yields $|(\mathcal{T}_a - \mathcal{T}_b)[(\dot A, \dot B), (\dot A, \dot B)]| \le |s|\,\|\nabla_W \ell_a - \nabla_W \ell_b\|_F$; since $\mathcal{T}_a - \mathcal{T}_b$ is symmetric this is the stated operator-norm bound. The constant is $|s|$ for every member of the family; only the value of s differs (α/r for LoRA and PiSSA, which differ only in initialization, $\alpha/\sqrt{r}$ for rsLoRA). If instead the gradient is taken with respect to the unscaled product $\Delta W = BA$, then $\nabla_{\Delta W}\ell = s\nabla_W \ell$ and the bound reads $\|\nabla_{\Delta W}\ell_a - \nabla_{\Delta W}\ell_b\|_F$ with constant 1. With several blocks, $\mathcal{T}_a - \mathcal{T}_b$ is block diagonal across blocks (each block's T depends only on its own factors), so its operator norm is the maximum of the blockwise bounds. □

Remark 1 (What the bound does and does not cover). The bound controls $\mathcal{T}_a - \mathcal{T}_b$ only. The residual difference $\mathcal{E}_a - \mathcal{E}_b$ is not bounded: it vanishes when the network is linear in θ, and is small near interpolation because $\nabla^2_\theta \ell_a - F_a$ carries the factor $\nabla_f \ell_a$, but for a deep network at a general point we have no bound. Consequently η_{ab} is a quantity to be assumed small or measured, not one shown to be small, and every statement converting between C_{ab} and $F_a^{\mathrm{R}} - F_b^{\mathrm{R}}$ carries it explicitly. For DoRA the magnitude–direction reparameterization contributes a further term of the same type, which we also do not bound; this matters because DoRA is one of the three adapters in Section 6, whereas the bound does cover the LoRA and PiSSA runs.

Proposition 4 (Gauge redundancy of factorized adapters)**.** *Let $T(\phi) = \theta_0 + s\,BA$ with $\phi = (A, B)$, $B \in \mathbb{R}^{d_{\mathrm{out}} \times r}$, $A \in \mathbb{R}^{r \times d_{\mathrm{in}}}$ and fixed scale $s \neq 0$ (the LoRA, rsLoRA and PiSSA parameterization), and assume $1 \le r \le \min(d_{\mathrm{in}}, d_{\mathrm{out}})$. Write $r_A = \operatorname{rank} A$, $r_B = \operatorname{rank} B$.*

(i) Gauge tangent. *For every $X \in \mathbb{R}^{r\times r}$ the curve $A(t) = e^{-tX}A$, $B(t) = Be^{tX}$ satisfies $B(t)A(t) = BA$ for all t, and its tangent $v = (-XA, BX)$ lies in* $\ker J(\phi)$. *The gauge tangent space $\{(-XA, BX)\}$ has dimension exactly $r^2 - (r - r_A)(r - r_B)$.*

(ii) Kernel. *Exactly,*

$$\dim\ker J(\phi) = d_{\mathrm{out}}(r - r_A) + d_{\mathrm{in}}(r - r_B) + r_A r_B \;\ge\; r^2, \tag{8}$$

and $\dim\ker J(\phi)$ *exceeds the gauge tangent dimension by exactly $(r - r_A)(d_{\mathrm{out}} - r_B) + (r - r_B)(d_{\mathrm{in}} - r_A)$. The kernel* can *be strictly larger than the gauge tangent at rank-deficient points, but need not be: for $r = d_{\mathrm{in}} = 1$, $d_{\mathrm{out}} = 2$, $A = [1]$, $B = 0$ both are one-dimensional.*

(iii) Consequences. *For all groups a, b and every ϕ there is a nonzero $v \in \ker J(\phi)$, so*

$$(F_a^{\mathrm{R}} - F_b^{\mathrm{R}})\,v = 0 \qquad \textit{and so} \qquad \lambda_{\min}(F_a^{\mathrm{R}} - F_b^{\mathrm{R}}) \le 0,$$

and if in addition $g_a = g_b$ then some nonzero v has $v^\top C_{ab} v = 0$, so $C_{ab} \not\succ 0$. (At the origin $A = B = 0$ the gauge tangent is zero; there any nonzero pure-A direction $v = (U, 0)$ serves, since $B\,(A + tU) = 0$ along the line.)

Without $r \le \min(d_{\mathrm{in}}, d_{\mathrm{out}})$ the lower bound r^2 in (8) fails: $d_{\mathrm{in}} = d_{\mathrm{out}} = 1$, $r = 2$ gives $\dim\ker J = 3$ *at generic points.*

Proof. Since $\frac{d}{dt}e^{tX} = Xe^{tX}$, the curve $A(t) = e^{-tX}A$, $B(t) = Be^{tX}$ gives $B(t)A(t) = Be^{tX}e^{-tX}A = BA$ for every t, so $T(\phi(t)) = T(\phi)$ identically. Its tangent at $t = 0$ is $v = (\dot A, \dot B) = (-XA, BX)$, and by the product rule

$$J(\phi)v = s\big(\dot B A + B\dot A\big) = s\big(BXA - BXA\big) = 0,$$

so $v \in \ker J(\phi)$. The map $X \mapsto (-XA, BX)$ is linear with kernel $\{X : XA = 0,\ BX = 0\}$, so the gauge orbit contributes $r^2 - \dim\{X : XA = 0, BX = 0\}$ dimensions; if A has full row rank then $XA = 0$ forces $X = 0$, and symmetrically for B of full column rank, so in that case the gauge orbit alone already has dimension r^2.

Gauge tangent dimension. The map $X \mapsto (-XA, BX)$ is linear, with kernel $\{X : XA = 0,\ BX = 0\}$. The condition $XA = 0$ says every row $x^\top$ of X satisfies $x^\top A = 0$, i.e. $x \in (\operatorname{col} A)^\perp \subseteq \mathbb{R}^r$, a space of dimension $r - r_A$; the condition $BX = 0$ says every column of X lies in $\ker B \subseteq \mathbb{R}^r$, of dimension $r - r_B$. Together they say $X \in \ker B \otimes (\operatorname{col} A)^\perp$, i.e. $X = \sum_i u_i w_i^\top$ with $u_i \in \ker B$ and $w_i \in (\operatorname{col} A)^\perp$, a space of dimension $(r - r_A)(r - r_B)$. The gauge tangent therefore has dimension $r^2 - (r - r_A)(r - r_B)$.

Kernel dimension. Since $J(\dot A, \dot B) = s(\dot B A + B\dot A)$ and $s \neq 0$,

$$\operatorname{range} J = \{\dot B A\} + \{B\dot A\} = \big(\mathbb{R}^{d_{\text{out}}} \otimes \operatorname{row} A\big) + \big(\operatorname{col} B \otimes \mathbb{R}^{d_{\text{in}}}\big),$$

with summands of dimensions $d_{\text{out}} r_A$ and $d_{\text{in}} r_B$ and intersection exactly $\operatorname{col} B \otimes \operatorname{row} A$, of dimension $r_A r_B$ (for subspaces $U \subseteq \mathbb{R}^m$, $V \subseteq \mathbb{R}^n$, $(\mathbb{R}^m \otimes V) \cap (U \otimes \mathbb{R}^n) = U \otimes V$). Hence $\operatorname{rank} J = d_{\text{out}} r_A + d_{\text{in}} r_B - r_A r_B$ exactly and, with $p = r(d_{\text{in}} + d_{\text{out}})$,

$$\dim \ker J = p - \operatorname{rank} J = d_{\text{out}}(r - r_A) + d_{\text{in}}(r - r_B) + r_A r_B.$$

Using the assumption $d_{\text{out}}, d_{\text{in}} \geq r$ this is at least $r(r - r_A) + r(r - r_B) + r_A r_B = r^2 + (r - r_A)(r - r_B) \geq r^2$, giving (8). Subtracting the gauge dimension,

$$\dim \ker J - \big(r^2 - (r - r_A)(r - r_B)\big) = (r - r_A)(d_{\text{out}} - r_B) + (r - r_B)(d_{\text{in}} - r_A),$$

which is zero at full rank $r_A = r_B = r$, and zero also for $r = d_{\text{in}} = 1$, $d_{\text{out}} = 2$, $A = [1]$, $B = 0$ (first term 0, second term $(1 - 0)(1 - 1) = 0$), but equals 1 at the point of Remark 2 ($d_{\text{in}} = 2$). A rank-deficient point therefore *may* have a kernel strictly larger than the gauge tangent; it need not. Without $r \leq \min(d_{\text{in}}, d_{\text{out}})$ the final inequality fails, e.g. $d_{\text{in}} = d_{\text{out}} = 1$, $r = 2$, $r_A = r_B = 1$ gives $\dim \ker J = 3 < 4$.

For the consequences, pick a nonzero $v \in \ker J(\phi)$ lying on a curve $\phi(t)$ along which T is constant: away from the origin $r^2 - (r - r_A)(r - r_B) \geq 1$, so some X gives a nonzero gauge tangent on the curve $(e^{-tX}A, Be^{tX})$; at the origin $A = B = 0$ take the straight line $\phi(t) = (tU, 0)$ with $U \neq 0$, along which $BA \equiv 0$ and $\phi''(0) = 0$. In either case $F_a^{\mathrm{R}} v = J^\top F_a J v = 0$ for every group, hence $(F_a^{\mathrm{R}} - F_b^{\mathrm{R}})v = 0$ with $v \neq 0$: zero is an eigenvalue of the symmetric matrix $F_a^{\mathrm{R}} - F_b^{\mathrm{R}}$, so $\lambda_{\min}(F_a^{\mathrm{R}} - F_b^{\mathrm{R}}) \leq 0$. Finally, L_a and L_b are constant along $\phi(t)$, so $\Delta_{ab}(\phi(t))$ is constant and

$$0 = \frac{d^2}{dt^2} \Delta_{ab}(\phi(t))\Big|_{t=0} = v^\top C_{ab} v + (g_a - g_b)^\top \phi''(0).$$

With $g_a = g_b$ the second term vanishes (at the origin it vanishes regardless, since $\phi''(0) = 0$), leaving $v^\top C_{ab} v = 0$ for a nonzero v, so C_{ab} is not positive definite. □

Corollary 3 (The symmetry survives any map through BA)**.** *Let $\phi = (A, B, m)$, where m collects any further trainable parameters (for DoRA, the magnitude vector). If $T(\phi) = G(BA, m)$ for any differentiable G, then T is constant along the curve of Proposition 4(i) with m held fixed, and $v = (-XA, BX, 0) \in \ker J(\phi)$, and likewise along the pure-A line at the origin. This covers LoRA, rsLoRA and PiSSA ($G(W, m) = \theta_0 + sW$, no m) and also DoRA, whose adapted weight $m(W_0 + BA)/\|W_0 + BA\|_c$ depends on (A, B) only through BA wherever the normalization is defined. Since the (A, B) block of J is $D_1 G \cdot J_{BA}$, where J_{BA} is the Jacobian of $(A, B) \mapsto BA$, $\ker J \supseteq \ker J_{BA} \times \{0\}$: under the same rank assumption, part (i), the lower bound $\dim \ker J \geq r^2$ of part (ii), and part (iii) hold verbatim; the exact kernel formula becomes a lower bound.*

Proof. If $T(\phi) = G(BA, m)$ then, with m fixed along the curve, $T(\phi(t)) = G(B(t)A(t), m) = G(BA, m) = T(\phi)$ for all t, so $J(\phi)v = \frac{d}{dt} T(\phi(t))|_{t=0} = 0$; the same holds along the pure-A line $(tU, 0)$ at the origin. By the chain rule the (A, B) block of J is $D_1 G(BA, m)\, J_{BA}$ with J_{BA} the Jacobian of $(A, B) \mapsto BA$, so $\ker J \supseteq \ker J_{BA} \times \{0\}$, and $\dim \ker J_{BA}$ is given exactly by (8) (the case $G =$ identity, $s = 1$); hence $\dim \ker J \geq r^2$ under the rank assumption. Part (iii) used only the existence of a nonzero v on a curve of constant T, which we have. For DoRA, $W = m\,(W_0 + BA)/\|W_0 + BA\|_c$ is a function of BA alone at fixed m and W_0, and m is not moved by the curve, so the hypothesis holds wherever the column norm is nonzero. □

Remark 2 (The horizontal space is first-order only). Take $T(a_1, a_2, b) = b\,(a_1, a_2)$ and $\Delta(w_1, w_2) = \frac{1}{2}w_1^2 - w_2$, so $\Delta_{ab}(a_1, a_2, b) = \frac{1}{2}b^2a_1^2 - ba_2$. At $(1, 0, 0)$ the Jacobian is $J = \left(\begin{smallmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{smallmatrix}\right)$, so $\ker J = \mathrm{span}\{e_{a_1}, e_{a_2}\}$ and $\mathcal{H}_\phi = \mathrm{span}\{e_b\}$; the gradient vanishes and $C_{ab} = \left(\begin{smallmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 1 \end{smallmatrix}\right)$, so $\lambda^{\mathcal{H}}_{\min}(C_{ab}) = e_b^\top C_{ab} e_b = 1 > 0$ and Theorem 1 applies. Indeed $\Delta_{ab}(\phi + (0, 0, t)) - \Delta_{ab}(\phi) = \frac{1}{2}t^2 > 0$. However $\Delta_{ab}(\phi + (0, t, t)) - \Delta_{ab}(\phi) = \frac{1}{2}t^2 - t^2 = -\frac{1}{2}t^2 < 0$, and $(0, t, t) - (0, 0, t) = (0, t, 0) \in \ker J$. Two updates agreeing to first order in predictor space therefore move the gap in opposite directions at second order, so $\mathcal{H}_\phi$ cannot be read as a second-order quotient and no conclusion transfers from a horizontal representative to an arbitrary one. The point is the standard $B = 0$ initialization, not an exotic configuration. Here $\dim \ker J = 2$ while the gauge tangent is one-dimensional ($r = 1$, $r_A = 1$, $r_B = 0$, $d_{\text{in}} = 2$, $d_{\text{out}} = 1$, so the excess of Proposition 4(ii) is $0 + 1 = 1$): the kernel is strictly larger than the gauge tangent at this point.

Remark 3 (Why the restriction is not optional). Proposition 4 shows the unrestricted hypotheses "$F_a^{\mathrm{R}} - F_b^{\mathrm{R}} \succ 0$" and "$C_{ab} \succ 0$ with $g_a = g_b$" are *never* satisfied for LoRA, rsLoRA or PiSSA at any ϕ with A or B of full rank. Restricting to $\mathcal{H}_\phi = (\ker J)^\perp$ excises exactly the directions responsible and leaves the conditions satisfiable: in the supplement we sample random (F_a, F_b) at a random (A, B) and find $\lambda_{\min}(F_a^{\mathrm{R}} - F_b^{\mathrm{R}}) \le 0$ in every draw, while $\lambda^{\mathcal{H}}_{\min}(F_a^{\mathrm{R}} - F_b^{\mathrm{R}}) > 0$ occurs in roughly half of them. Since $\ker J$ is precisely the set of directions that leave the predictor unchanged to first order, $\mathcal{H}_\phi$ is a Euclidean choice of representative for that first-order quotient, and quantifying over $\delta \in \mathcal{H}_\phi$ is the weakest restriction that makes the hypotheses attainable. It is not more than that: Remark 2 shows the identification fails at second order.

B.2 Necessity and sufficiency

Theorem 1 (Forced increase on the horizontal subspace). *Write $\rho_{\min} := \lambda^{\mathcal{H}}_{\min}(C_{ab})$ for the smallest eigenvalue of the curvature difference restricted to $\mathcal{H}_\phi$. Assume **A1**, $g_a = g_b$, and $\rho_{\min} > 0$, that is, C_{ab} is positive definite on $\mathcal{H}_\phi$. Then for every nonzero $\delta \in \mathcal{H}_\phi$ with $\|\delta\| \le \min(\varepsilon_0, 3\rho_{\min}/4M)$, where ε_0, M are from **A1** and the second bound is $+\infty$ when $M = 0$,*

$$\Delta_{ab}(\phi + \delta) - \Delta_{ab}(\phi) \;\ge\; \tfrac{1}{4}\rho_{\min}\|\delta\|^2 \;>\; 0.$$

The signed gap therefore strictly increases for every admissible horizontal additive update. If in addition $\Delta_{ab}(\phi) \ge 0$, the absolute gap $|\Delta_{ab}|$ strictly widens.

Proof. By Proposition 1 with $g_a = g_b$ the slope term vanishes, so for $\delta \in \mathcal{H}_\phi$,

$$\Delta_{ab}(\phi + \delta) - \Delta_{ab}(\phi) = \tfrac{1}{2}\delta^\top C_{ab}\delta + R_{ab}(\delta) \;\ge\; \tfrac{1}{2}\rho_{\min}\|\delta\|^2 - \tfrac{M}{3}\|\delta\|^3,$$

using $\delta^\top C_{ab}\delta \ge \lambda^{\mathcal{H}}_{\min}(C_{ab})\|\delta\|^2$ for $\delta \in \mathcal{H}_\phi$ and $|R_{ab}| \le \frac{M}{3}\|\delta\|^3$. Since $\|\delta\| \le \varepsilon \le 3\rho_{\min}/4M$,

$$\tfrac{M}{3}\|\delta\|^3 \le \tfrac{M}{3}\varepsilon\|\delta\|^2 \le \tfrac{M}{3} \cdot \tfrac{3\rho_{\min}}{4M}\|\delta\|^2 = \tfrac{\rho_{\min}}{4}\|\delta\|^2,$$

so the increment is at least $(\frac{1}{2} - \frac{1}{4})\rho_{\min}\|\delta\|^2 = \frac{1}{4}\rho_{\min}\|\delta\|^2 > 0$ for $\delta \ne 0$. (When $M = 0$ the remainder vanishes identically, the constraint $\|\delta\| \le 3\rho_{\min}/4M$ is read as vacuous, and the increment is at least $\frac{1}{2}\rho_{\min}\|\delta\|^2$.) This is an increase of the *signed* gap $\Delta_{ab} = L_a - L_b$. If $\Delta_{ab}(\phi) \ge 0$ then $|\Delta_{ab}(\phi + \delta)| = \Delta_{ab}(\phi + \delta) > \Delta_{ab}(\phi) = |\Delta_{ab}(\phi)|$, so the absolute gap widens; without that orientation an increase in Δ_{ab} may reduce $|\Delta_{ab}|$. □

Remark 4 (Both hypotheses are necessary). *Slope matching.* Let the exact gap along a ray be $\Delta_{ab}(\phi + tu) = t + \frac{1}{2}t^2$ for a unit u, so $g_a - g_b$ has component 1 along u, $C_{ab} = 1 \succ 0$ and $M = 0$. Taking $t = -0.1$ gives $\Delta_{ab}(\phi + tu) - \Delta_{ab}(\phi) = -0.095 < 0$: a nonzero update that strictly *decreases* the gap. Hence the conclusion, and with it Theorem 7(ii), fails without $g_a = g_b$. *Orientation.* If $\Delta_{ab}(\phi) = -1$ and the signed gap increases to -0.9, the absolute gap has narrowed. "Widening" therefore requires $\Delta_{ab}(\phi) \ge 0$, which is a labelling convention on the pair and not an additional assumption on the model.

Corollary 1 (Reachable-Fisher sufficient condition). *Write $\eta^{\mathcal{H}}_{ab} = \|\Xi_a - \Xi_b\|^{\mathcal{H}}_2 \le \eta_{ab}$. If $\lambda^{\mathcal{H}}_{\min}(F_a^{\mathrm{R}} - F_b^{\mathrm{R}}) > \eta^{\mathcal{H}}_{ab}$ then $\lambda^{\mathcal{H}}_{\min}(C_{ab}) > 0$; if* additionally *$g_a = g_b$, then Theorem 1 applies with margin $\rho_{\min} \ge \lambda^{\mathcal{H}}_{\min}(F_a^{\mathrm{R}} - F_b^{\mathrm{R}}) - \eta^{\mathcal{H}}_{ab}$. By Proposition 4 the restriction to $\mathcal{H}_\phi$ is necessary: the unrestricted condition $\lambda_{\min}(F_a^{\mathrm{R}} - F_b^{\mathrm{R}}) > \eta_{ab}$ is unsatisfiable for factorized adapters.*

Proof. By Proposition 3, $C_{ab} = (F_a^{\mathrm{R}} - F_b^{\mathrm{R}}) + (\Xi_a - \Xi_b)$ with both terms symmetric. Compressing to $\mathcal{H}_\phi$ and applying Weyl's inequality to the restrictions,

$$\lambda_{\min}^{\mathcal{H}}(C_{ab}) \;\geq\; \lambda_{\min}^{\mathcal{H}}(F_a^{\mathrm{R}} - F_b^{\mathrm{R}}) + \lambda_{\min}^{\mathcal{H}}(\Xi_a - \Xi_b) \;\geq\; \lambda_{\min}^{\mathcal{H}}(F_a^{\mathrm{R}} - F_b^{\mathrm{R}}) - \eta_{ab}^{\mathcal{H}},$$

positive under the stated hypothesis. Theorem 1 additionally requires $g_a = g_b$, which is why it is stated as a separate condition here. That the restriction to $\mathcal{H}_\phi$ cannot be removed is Proposition 4. □

Remark 5 ($F_a^{\mathrm{R}} - F_b^{\mathrm{R}}$ is not the only lever). Since $C_{ab} = (F_a^{\mathrm{R}} - F_b^{\mathrm{R}}) + (\Xi_a - \Xi_b)$, a procedure that altered the defect would alter C_{ab} without touching the reachable Fishers, and Proposition 3 explicitly retains that term. We therefore do not claim $\|F_a^{\mathrm{R}} - F_b^{\mathrm{R}}\|$ is the only available intervention, only that it is the one our estimator acts on and the one measurable at foundation-model scale.

Theorem 2 (Worst-case growth of the two channels)**.** *Over $\delta \in \mathcal{H}_\phi$ with $\|\delta\| \leq \varepsilon$, let $\mathcal{Q}_{ab}(\varepsilon) = \max \frac{1}{2}|\delta^\top C_{ab}\delta|$ be the worst-case second-order term and $\mathcal{G}_{ab}(\varepsilon) = \max |\Delta_{ab}(\phi+\delta) - \Delta_{ab}(\phi)|$ the worst-case gap change. Then*

$$\mathcal{Q}_{ab}(\varepsilon) = \tfrac{1}{2}\|C_{ab}\|_2^{\mathcal{H}}\varepsilon^2, \qquad \mathcal{G}_{ab}(\varepsilon) \leq \|g_a - g_b\|\varepsilon + \tfrac{1}{2}\|C_{ab}\|_2^{\mathcal{H}}\varepsilon^2 + \tfrac{M}{3}\varepsilon^3,$$

the first attained at a dominant eigenvector of C_{ab} restricted to $\mathcal{H}_\phi$. If $g_a = g_b$ the bound is tight to third order: $|\mathcal{G}_{ab}(\varepsilon) - \frac{1}{2}\|C_{ab}\|_2^{\mathcal{H}}\varepsilon^2| \leq \frac{M}{3}\varepsilon^3$.

Proof. Let Π be the orthogonal projector onto $\mathcal{H}_\phi$ and $S = \Pi C_{ab}\Pi$ the compression, which is symmetric. Since $\delta \mapsto \delta^\top C_{ab}\delta$ is homogeneous of degree two and agrees with $\delta^\top S\delta$ on $\mathcal{H}_\phi$, its maximum over $\{\delta \in \mathcal{H}_\phi : \|\delta\| \leq \varepsilon\}$ is $\lambda_{\max}(S|_{\mathcal{H}})\varepsilon^2$, attained at ε times a corresponding eigenvector. Applying this to $\pm C_{ab}$ and using $\|C_{ab}\|_2^{\mathcal{H}} = \max\{\lambda_{\max}(S|_{\mathcal{H}}), -\lambda_{\min}(S|_{\mathcal{H}})\}$ gives $\mathcal{Q}_{ab}(\varepsilon) = \frac{1}{2}\|C_{ab}\|_2^{\mathcal{H}}\varepsilon^2$ with attainment. For $\mathcal{G}_{ab}$, apply the triangle inequality to (2): $|(g_a - g_b)^\top\delta| \leq \|g_a - g_b\|\varepsilon$ by Cauchy–Schwarz, the quadratic term is at most $\frac{1}{2}\|C_{ab}\|_2^{\mathcal{H}}\varepsilon^2$ on $\mathcal{H}_\phi$, and $|R_{ab}| \leq \frac{M}{3}\varepsilon^3$. When $g_a = g_b$ the slope term is absent and the same two-sided argument gives $|\mathcal{G}_{ab}(\varepsilon) - \frac{1}{2}\|C_{ab}\|_2^{\mathcal{H}}\varepsilon^2| \leq \frac{M}{3}\varepsilon^3$. □

Theorem 6 (Fisher matching controls the second-order channel up to the defect)**.** *If $F_a^{\mathrm{R}} = F_b^{\mathrm{R}}$ then for all $\|\delta\| \leq \varepsilon \leq \varepsilon_0$,*

$$\left|\Delta_{ab}(\phi+\delta) - \Delta_{ab}(\phi) - (g_a - g_b)^\top\delta\right| \;\leq\; \tfrac{1}{2}\eta_{ab}\varepsilon^2 + \tfrac{M}{3}\varepsilon^3.$$

In particular any $\delta \perp (g_a - g_b)$ holds the gap fixed to $O(\eta_{ab}\varepsilon^2 + \varepsilon^3)$.

Proof. By Proposition 3, $F_a^{\mathrm{R}} = F_b^{\mathrm{R}}$ gives $C_{ab} = \Xi_a - \Xi_b$, so $|\frac{1}{2}\delta^\top C_{ab}\delta| \leq \frac{1}{2}\eta_{ab}\|\delta\|^2$. Substituting into (2) and bounding $|R_{ab}| \leq \frac{M}{3}\varepsilon^3$ gives the display; taking $\delta \perp (g_a - g_b)$ removes the slope term. □

Theorem 7 (Two feasibility results)**.** *Define the second-order model $\widehat{\Delta}_{ab}(\delta) = (g_a - g_b)^\top\delta + \frac{1}{2}\delta^\top C_{ab}\delta$. Since $\widehat{\Delta}_{ba} = -\widehat{\Delta}_{ab}$, fix an* orientation*: let $\mathcal{O}$ contain, for each unordered pair $\{a,b\}$, exactly one ordered pair (a,b) with $\Delta_{ab}(\phi) \geq 0$ (ties broken arbitrarily), so that a is the currently worse-off group. Let*

$$\mathcal{S}_\varepsilon := \left\{\delta \in \mathcal{H}_\phi :\; \|\delta\| \leq \varepsilon,\; \widehat{\Delta}_{ab}(\delta) \leq 0 \;\; \forall (a,b) \in \mathcal{O}\right\},$$

the horizontal updates under which no oriented gap increases in the truncated quadratic model*. Fix pooling weights $\pi_a \geq 0$ with $\sum_a \pi_a = 1$, let $\bar{g} = \sum_a \pi_a g_a$ be the gradient of the pooled loss $\sum_a \pi_a L_a$, and call $\delta \in \mathcal{H}_\phi$ a* pooled descent direction *if $\bar{g}^\top\delta < 0$. (Quantifying over both orders of every pair would instead force $\widehat{\Delta}_{ab}(\delta) = 0$.)* (i) *If $C_{ab} = 0$ for all pairs, $\mathcal{S}_\varepsilon$ is the intersection of a ball with a closed convex polyhedral cone in $\mathcal{H}_\phi$, and it contains a pooled descent direction iff $-P_{\mathcal{H}}\bar{g} \notin \mathrm{cone}\{P_{\mathcal{H}}(g_a - g_b) : (a,b) \in \mathcal{O}\}$, which is a single LP.* (ii) *If $g_a = g_b$ for all pairs and $\lambda_{\min}^{\mathcal{H}}(C_{ab}) > 0$ for some $(a,b) \in \mathcal{O}$, then $\mathcal{S}_\varepsilon = \{0\}$ for every $\varepsilon > 0$; the same holds for the exact gap for ε as in Theorem 1.*

Stating $\mathcal{S}_\varepsilon$ through the truncated model is what makes (i) exact; phrased on the true loss it would inherit the cubic remainder. The slope hypothesis in (ii) cannot be dropped: the example after Theorem 1 has a definite C_{ab} and a nonzero gap-decreasing update. And by Proposition 4, "$\mathcal{S}_\varepsilon = \{0\}$" must be read on $\mathcal{H}_\phi$: raw parameters can always move along a gauge orbit without changing the predictor or any gap. What *exact Hessian* matching $C_{ab} = 0$ buys is therefore conditional but real: it converts the truncated quadratic feasibility question into a linear one. Reachable-Fisher matching $F_a^{\mathrm{R}} = F_b^{\mathrm{R}}$ does not by itself do this: it leaves $C_{ab} = \Xi_a - \Xi_b$, so the quadratic term is bounded by $\frac{1}{2}\eta_{ab}\varepsilon^2$ (Theorem 6) but not removed, and the exact constraint is not convexified. Our method matches empirical reachable Fishers, not Hessians.

Proof. (i) With $C_{ab} = 0$ for all pairs the quadratic term of the model $\widehat{\Delta}_{ab}$ vanishes identically (no remainder is involved, since $\mathcal{S}_\varepsilon$ is defined through the truncated model), and each constraint reduces to $(g_a - g_b)^\top \delta \le 0$, $(a,b) \in \mathcal{O}$, for $\delta \in \mathcal{H}_\phi$; a finite intersection of homogeneous halfspaces with a subspace is a closed convex polyhedral cone. Let A have rows $(P_{\mathcal{H}}(g_a - g_b))^\top$, $(a,b) \in \mathcal{O}$, and $c = -P_{\mathcal{H}}\bar{g}$. By Farkas' lemma exactly one holds: there is δ with $A\delta \le 0$ and $c^\top \delta > 0$; or there is $y \ge 0$ with $A^\top y = c$, i.e. $c \in \mathrm{cone}\{P_{\mathcal{H}}(g_a - g_b) : (a,b) \in \mathcal{O}\}$. Because the constraint set is a cone, feasibility of the strict inequality is unaffected by the norm bound after rescaling.

(ii) With $g_a = g_b$ the model reduces to $\widehat{\Delta}_{ab}(\delta) = \frac{1}{2}\delta^\top C_{ab}\delta \ge \frac{1}{2}\lambda_{\min}^{\mathcal{H}}(C_{ab})\|\delta\|^2 > 0$ for every nonzero $\delta \in \mathcal{H}_\phi$, so no such δ satisfies the constraint and $\mathcal{S}_\varepsilon = \{0\}$ for every $\varepsilon > 0$; no smallness condition is needed because the model carries no remainder. For the exact gap, Theorem 1 gives the same conclusion on the stated ball. The slope hypothesis is necessary: Remark 4 exhibits a definite C_{ab} together with a nonzero gap-decreasing update. The restriction to $\mathcal{H}_\phi$ is likewise necessary by Proposition 4, and $\mathcal{S}_\varepsilon = \{0\}$ must be read as "no nonzero *horizontal additive* update": a step along a gauge orbit changes neither predictor nor gap, and by Remark 2 a non-horizontal representative of the same first-order class may behave differently at second order. □

B.3 Reachable versus full-model geometry

Throughout, $\mathcal{R}_\phi := \mathrm{range}\, J(\phi) \subseteq \mathbb{R}^P$ is the reachable subspace and P_ϕ the orthogonal projector onto it.

Lemma 1 (Projection and invariance)**.** *$F_a^{\mathrm{R}} - F_b^{\mathrm{R}} = J^\top P_\phi (F_a - F_b) P_\phi J$: the subgroup geometry enters only through the compression of $F_a - F_b$ onto $\mathcal{R}_\phi$. Under a reparameterization $\phi = S(\psi)$ with invertible Jacobian K, $F_a^{\mathrm{R}} \mapsto K^\top F_a^{\mathrm{R}} K$; hence $\mathrm{tr}\, F_a^{\mathrm{R}}$ and $\lambda_{\max}(F_a^{\mathrm{R}})$ are not invariant, the across-group coefficient of variation $\mathrm{CV}\{\mathrm{tr}\, F_a^{\mathrm{R}}\}_a$ is invariant under isotropic $K = cI$ but* not *under general K, and the inertia of $F_a^{\mathrm{R}} - F_b^{\mathrm{R}}$ (its numbers of positive, negative and zero eigenvalues) is invariant, as are the generalized eigenvalues of $(F_a^{\mathrm{R}}, F_b^{\mathrm{R}})$ whenever that pencil is regular (i.e. $\det(F_a^{\mathrm{R}} - \lambda F_b^{\mathrm{R}}) \not\equiv 0$); by Proposition 4 the pencil is singular on $\ker J$, so regularity should be read on $\mathcal{H}_\phi$.*

Proof. Every column of J lies in $\mathcal{R}_\phi$, hence $P_\phi J = J$ and $J^\top(F_a - F_b)J = (P_\phi J)^\top (F_a - F_b)(P_\phi J) = J^\top P_\phi (F_a - F_b) P_\phi J$.

For the reparameterization claims, the chain rule applied to (1) gives $F_a^{\mathrm{R},(\psi)} = K^\top F_a^{\mathrm{R},(\phi)} K$. Taking $K = cI$ scales every $\mathrm{tr}\, F_a^{\mathrm{R}}$ by c^2 without changing the predictor, so the trace and $\lambda_{\max}$ are not invariant; note that this particular K leaves the across-group coefficient of variation $\mathrm{CV}\{\mathrm{tr}\, F_a^{\mathrm{R}}\}_a$ unchanged, since it rescales all group traces by the same factor. A non-isotropic K is required to move the CV, and suffices: with $p = 2$, $F_a^{\mathrm{R}} = \mathrm{diag}(1,0)$ and $F_b^{\mathrm{R}} = \mathrm{diag}(0,1)$ the traces are $(1,1)$ and $\mathrm{CV} = 0$, whereas under $K = \mathrm{diag}(1,2)$ they become $(1,4)$ with $\mathrm{CV} = 0.6$. Hence the CV is invariant under isotropic rescaling but not under general reparameterization.

The difference transforms as $F_a^{\mathrm{R},(\psi)} - F_b^{\mathrm{R},(\psi)} = K^\top (F_a^{\mathrm{R},(\phi)} - F_b^{\mathrm{R},(\phi)}) K$, a congruence by an invertible matrix, so its inertia is preserved by Sylvester's law of inertia. Finally, if $F_a^{\mathrm{R},(\phi)} v = \lambda F_b^{\mathrm{R},(\phi)} v$ then $w = K^{-1} v$ satisfies $F_a^{\mathrm{R},(\psi)} w = \lambda F_b^{\mathrm{R},(\psi)} w$, so the generalized eigenvalues of the pencil are invariant *provided the pencil is regular*. By Proposition 4 both F_a^{R} and F_b^{R} annihilate $\ker J$, so $\det(F_a^{\mathrm{R}} - \lambda F_b^{\mathrm{R}}) \equiv 0$ on $\mathbb{R}^p$ and the pencil is singular there; the statement should be read for the compressions to $\mathcal{H}_\phi$, where regularity is generic. □

Theorem 3 (Full-model summaries do not determine reachable geometry)**.** (i) *For every $c > 0$ there exist $F_a, F_b \succeq 0$ and a Jacobian J with $\|F_a - F_b\|_2 = c$ yet $F_a^{\mathrm{R}} = F_b^{\mathrm{R}}$.* (ii) *There exist $F_a, F_b \succeq 0$ with identical spectra such that $F_a^{\mathrm{R}} - F_b^{\mathrm{R}}$ is positive definite on $\mathcal{H}_\phi$.*

Proof. (i) Take $P = 2$, $p = 1$, $J = e_1$ and $F_a = \mathrm{diag}(1, c)$, $F_b = \mathrm{diag}(1, 0)$. Both are positive semidefinite, $\|F_a - F_b\|_2 = \|\,\mathrm{diag}(0, c)\|_2 = c$, and $F_a^{\mathrm{R}} = e_1^\top F_a e_1 = 1 = e_1^\top F_b e_1 = F_b^{\mathrm{R}}$; the disparity lies entirely in $\mathcal{R}^\perp$, consistent with Lemma 1. The consequence for the gap is exactly Theorem 6, i.e. the second-order channel is bounded by $\frac{1}{2}\eta_{ab}\varepsilon^2$, not that the gap is fixed, since the slope term is unconstrained by this construction.

(ii) Take $P = 2, p = 1, J = e_1$ and

$$F_a = \mathrm{diag}(2, 1), \qquad F_b = \mathrm{diag}(1, 2) = QF_aQ^\top, \qquad Q = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Q is orthogonal, so F_a and F_b are orthogonally similar and share the spectrum $\{1, 2\}$; every function of the eigenvalues alone therefore agrees. Nevertheless $F_a^{\mathrm{R}} = 2$, $F_b^{\mathrm{R}} = 1$ and $F_a^{\mathrm{R}} - F_b^{\mathrm{R}} = 1 > 0$. Here $p = 1$ and J is injective, so $\ker J = \{0\}$, $\mathcal{H}_\phi = \mathbb{R}$, and no gauge restriction is active; under the remaining hypotheses of Theorem 1 (slope matching and, via Corollary 1, $\eta_{ab}^{\mathcal{H}} < 1$), the forced-increase conclusion applies, while no summary depending only on the individual spectra of F_a and F_b distinguishes the two groups (the joint quantity $\|F_a - F_b\|_2 = 1$ does). For a factorized adapter the same construction is read on $\mathcal{H}_\phi$. Padding both matrices with a common block embeds the construction in any $P > p$. □

B.4 Trace insufficiency and a matrix-discrepancy alternative

Lemma 2 (Trace decomposition)**.** *With $\widehat{F}_a^{\mathrm{R,obs}} = \frac{1}{n_a}\sum_{i\in a} \nabla_\phi \ell_i \nabla_\phi \ell_i^\top$, mean $\hat{g}_a$ and within-group covariance $\widehat{\Sigma}_a$ (both with the $1/n_a$ convention),* $\mathrm{tr}\,\widehat{F}_a^{\mathrm{R,obs}} = \frac{1}{n_a}\sum_{i\in a}\|\nabla_\phi \ell_i\|_2^2 = \|\hat{g}_a\|_2^2 + \mathrm{tr}\,\widehat{\Sigma}_a$.

Proof. Write $v_i = \nabla_\phi \ell_i$. Since $\mathrm{tr}(v_i v_i^\top) = \|v_i\|_2^2$, linearity of the trace gives the first equality. The parallel-axis identity $\frac{1}{n_a}\sum_{i\in a} v_i v_i^\top = \hat{g}_a \hat{g}_a^\top + \widehat{\Sigma}_a$, with $\widehat{\Sigma}_a$ using the $1/n_a$ convention, followed by taking traces gives the second. □

Theorem 4 (Limits of trace matching)**.** *For $F_a^{\mathrm{R}}, F_b^{\mathrm{R}} \succeq 0$ of size $p \geq 2$:* (i) Necessary condition. *$F_a^{\mathrm{R}} = F_b^{\mathrm{R}}$ implies $\mathrm{tr}\,F_a^{\mathrm{R}} = \mathrm{tr}\,F_b^{\mathrm{R}}$; the converse fails.* (ii) General failure. *For any $\tau > 0$, $\sup\{\|F_a^{\mathrm{R}} - F_b^{\mathrm{R}}\|_2 : \mathrm{tr}\,F_a^{\mathrm{R}} = \mathrm{tr}\,F_b^{\mathrm{R}} = \tau\} = \tau$, attained by an indefinite difference: equal traces permit the largest spectral mismatch the trace budget allows.* (iii) Structured positive result. *If $F_a^{\mathrm{R}} - F_b^{\mathrm{R}} \succeq 0$ (Loewner ordering), then $\|F_a^{\mathrm{R}} - F_b^{\mathrm{R}}\|_2 \leq \mathrm{tr}\,F_a^{\mathrm{R}} - \mathrm{tr}\,F_b^{\mathrm{R}}$, so equal traces force equal matrices.*

Proof. (i) The implication is immediate, and the converse fails by the construction in (ii), whose two matrices have equal traces but differ. Quantitatively, $|\,\mathrm{tr}(F_a^{\mathrm{R}} - F_b^{\mathrm{R}})| = |\langle I, F_a^{\mathrm{R}} - F_b^{\mathrm{R}}\rangle_F| \leq \|I\|_F \|F_a^{\mathrm{R}} - F_b^{\mathrm{R}}\|_F = \sqrt{p}\,\|F_a^{\mathrm{R}} - F_b^{\mathrm{R}}\|_F$ by Cauchy–Schwarz.

(ii) *Upper bound.* Let $A, B \succeq 0$ with $\mathrm{tr}\,A = \mathrm{tr}\,B = \tau$. Weyl's inequality gives $\lambda_{\max}(A - B) \leq \lambda_{\max}(A) - \lambda_{\min}(B) \leq \lambda_{\max}(A) \leq \mathrm{tr}\,A = \tau$, using $\lambda_{\min}(B) \geq 0$ and that the eigenvalues of $A \succeq 0$ are nonnegative so the largest is at most their sum. Symmetrically $\lambda_{\max}(B - A) \leq \tau$, and since $A - B$ is symmetric, $\|A - B\|_2 = \max\{\lambda_{\max}(A - B), \lambda_{\max}(B - A)\} \leq \tau$.

Attainment. For $F_a^{\mathrm{R}} = \mathrm{diag}(\tau, 0, \ldots, 0)$ and $F_b^{\mathrm{R}} = \mathrm{diag}(0, \tau, 0, \ldots, 0)$ both traces equal τ and $F_a^{\mathrm{R}} - F_b^{\mathrm{R}} = \mathrm{diag}(\tau, -\tau, 0, \ldots, 0)$ has spectral norm τ.

(iii) If $C := F_a^{\mathrm{R}} - F_b^{\mathrm{R}} \succeq 0$, its eigenvalues are nonnegative, so $\|C\|_2 = \lambda_{\max}(C) \leq \sum_i \lambda_i(C) = \mathrm{tr}\,C$. If $\mathrm{tr}\,F_a^{\mathrm{R}} = \mathrm{tr}\,F_b^{\mathrm{R}}$ as well, then $\mathrm{tr}\,C = 0$ forces every eigenvalue to vanish, so $C = 0$. □

Corollary 2 (A small normalized trace gap excludes the sufficient regime)**.** *If $\lambda_{\min}^{\mathcal{H}}(F_a^{\mathrm{R}} - F_b^{\mathrm{R}}) > 0$ then $\mathrm{tr}\,F_a^{\mathrm{R}} - \mathrm{tr}\,F_b^{\mathrm{R}} \geq d_{\mathcal{H}}\,\lambda_{\min}^{\mathcal{H}}(F_a^{\mathrm{R}} - F_b^{\mathrm{R}})$. Hence $|\,\mathrm{tr}\,F_a^{\mathrm{R}} - \mathrm{tr}\,F_b^{\mathrm{R}}|/d_{\mathcal{H}} \leq \eta_{ab}^{\mathcal{H}}$ excludes the hypothesis of Corollary 1 in both orientations.*

Proof. Write $C = F_a^{\mathrm{R}} - F_b^{\mathrm{R}}$. Since $F_c^{\mathrm{R}} v = J^\top F_c J v = 0$ for $v \in \ker J$, and F_c^{R} is symmetric, we have $F_c^{\mathrm{R}} = P_{\mathcal{H}} F_c^{\mathrm{R}} P_{\mathcal{H}}$ and C vanishes on $\ker J$. If $\lambda_{\min}^{\mathcal{H}}(C) > 0$, then $C \succeq 0$ on all of $\mathbb{R}^p$, and taking the trace in an orthonormal basis of $\mathcal{H}_\phi$ extended by one of $\ker J$ gives $\operatorname{tr} C = \sum_{i \le d_{\mathcal{H}}} u_i^\top C u_i \ge d_{\mathcal{H}}\, \lambda_{\min}^{\mathcal{H}}(C)$, i.e. $\lambda_{\min}^{\mathcal{H}}(C) \le \operatorname{tr} C / d_{\mathcal{H}}$ (this holds for any symmetric C vanishing on $\ker J$, since the minimum of the Rayleigh quotient on $\mathcal{H}_\phi$ is at most its average over an orthonormal basis). Hence if $|\operatorname{tr} C|/d_{\mathcal{H}} \le \eta_{ab}^{\mathcal{H}}$ then $\lambda_{\min}^{\mathcal{H}}(C) \le \eta_{ab}^{\mathcal{H}}$, violating the hypothesis of Corollary 1; the same argument applies to $-C$ for the reverse orientation. Note that this excludes only the *sufficient* condition: C_{ab} may still be positive definite on $\mathcal{H}_\phi$ through the defect. □

Remark 6 (Trace equality and the magnitude of the gap change). Write $C = F_a^{\mathrm{R}} - F_b^{\mathrm{R}}$. Theorem 4 concerns C, whereas Theorem 2 gives $\mathcal{Q}_{ab}(\varepsilon) = \frac{1}{2}\|C_{ab}\|_2^{\mathcal{H}} \varepsilon^2$, *not* a formula in $\|F_a^{\mathrm{R}} - F_b^{\mathrm{R}}\|_2$; the two differ by the defect. For reachable Fishers the restricted and full spectral norms of C coincide: since $C = P_{\mathcal{H}} C P_{\mathcal{H}}$, for any unit w, $|w^\top C w| = |(P_{\mathcal{H}} w)^\top C (P_{\mathcal{H}} w)| \le \|C\|_2^{\mathcal{H}} \|P_{\mathcal{H}} w\|^2 \le \|C\|_2^{\mathcal{H}}$, so $\tau_{\mathcal{H}} = \|C\|_2^{\mathcal{H}} = \|C\|_2$, which is at most the common trace τ of Theorem 4(ii). The horizontal restriction matters for C_{ab} and the defect, which do not vanish on $\ker J$, not for the reachable Fishers themselves. Since $\big|\, \|C_{ab}\|_2^{\mathcal{H}} - \tau_{\mathcal{H}} \,\big| \le \eta_{ab}^{\mathcal{H}}$, under slope matching Theorem 2 gives $|\mathcal{G}_{ab}(\varepsilon) - \frac{1}{2}\tau_{\mathcal{H}}\varepsilon^2| \le \frac{1}{2}\eta_{ab}^{\mathcal{H}}\varepsilon^2 + \frac{M}{3}\varepsilon^3$: the *center* of this bound is $\frac{1}{2}\tau_{\mathcal{H}}\varepsilon^2$, and it is an exact value only if in addition $\eta_{ab}^{\mathcal{H}} = 0$ and $M = 0$ (an exactly quadratic gap). Equating the two is invalid: with $F_a^{\mathrm{R}} - F_b^{\mathrm{R}} = \operatorname{diag}(\tau, -\tau)$ and $\Xi_a - \Xi_b = \operatorname{diag}(0, 3\tau)$ one gets $C_{ab} = \operatorname{diag}(\tau, 2\tau) \succ 0$, so the definiteness premise holds, yet $\mathcal{Q}_{ab}(\varepsilon) = \tau\varepsilon^2 \ne \frac{1}{2}\tau\varepsilon^2$. Note also that the extremal difference $\operatorname{diag}(\tau, -\tau, 0, \dots)$ is itself indefinite, so Theorem 1 does not apply to it and no forced-increase reading attaches to the extremal construction.

Proposition 2 (Matrix-free Gram identity)**.** *For groups with n_a and n_b examples, any factors $G_a \in \mathbb{R}^{m_a \times p}$, $G_b \in \mathbb{R}^{m_b \times p}$ (one or more rows per example) and $\widehat{S}_c = \frac{1}{n_c} G_c^\top G_c$,*

$$\big\|\widehat{S}_a - \widehat{S}_b\big\|_F^2 = \frac{\|G_a G_a^\top\|_F^2}{n_a^2} - \frac{2\|G_a G_b^\top\|_F^2}{n_a n_b} + \frac{\|G_b G_b^\top\|_F^2}{n_b^2}, \tag{4}$$

computable in $O\big((m_a^2 + m_a m_b + m_b^2)p\big)$ time with no $p \times p$ matrix formed.

Proof. Expand $\|\widehat{S}_a - \widehat{S}_b\|_F^2 = \operatorname{tr}(\widehat{S}_a^2) - 2\operatorname{tr}(\widehat{S}_a \widehat{S}_b) + \operatorname{tr}(\widehat{S}_b^2)$. By cyclicity of the trace,

$$\operatorname{tr}(\widehat{S}_a \widehat{S}_b) = \frac{1}{n_a n_b} \operatorname{tr}\big(G_a^\top G_a G_b^\top G_b\big) = \frac{1}{n_a n_b} \operatorname{tr}\big(G_a G_b^\top (G_a G_b^\top)^\top\big) = \frac{\|G_a G_b^\top\|_F^2}{n_a n_b},$$

and the same computation with $b = a$ gives $\operatorname{tr}(\widehat{S}_a^2) = \|G_a G_a^\top\|_F^2 / n_a^2$. Nothing uses $m_a = n_a$: the normalization $1/n_a$ and the row count m_a enter separately.

Two factors are used in this paper. *Observed*: rows $\nabla_\phi \ell_i$ at the observed labels ($m_a = n_a$) give the observed-label empirical reachable Fisher $\widehat{F}_a^{\mathrm{R,obs}}$. *Expected*: with $q_\theta(c \mid x)$ the predicted probability of class c among C classes, rows $\sqrt{q_\theta(c \mid x_i)}\, \nabla_\phi \log q_\theta(c \mid x_i)$ over all classes ($m_a = n_a C$), or $\sigma^{-1} \nabla_\phi \mu_\theta(x_i)$ for Gaussian regression with predicted mean μ_θ and variance σ^2 ($m_a = n_a$), give the sample estimate $\widehat{F}_a^{\mathrm{R,exp}}$ of the model-expected reachable Fisher (1). For the expected factor, $\frac{1}{n_a} G_a^\top G_a = \frac{1}{n_a} \sum_i \sum_c q_\theta(c \mid x_i) \nabla_\phi \log q_\theta(c \mid x_i) \nabla_\phi \log q_\theta(c \mid x_i)^\top$, which is the sample average over inputs of the per-input expected score outer product, i.e. $\widehat{F}_a^{\mathrm{R,exp}}$; for Gaussian regression the score is $\sigma^{-2}(y - \mu)\nabla_\phi \mu$ and its expectation over y gives $\sigma^{-2} \nabla_\phi \mu \nabla_\phi \mu^\top$. All three Gram blocks are required. Forming $G_a G_a^\top$, $G_a G_b^\top$ and $G_b G_b^\top$ costs $O\big((m_a^2 + m_a m_b + m_b^2)p\big)$ and memory $O\big(m_a^2 + m_a m_b + m_b^2 + (m_a{+}m_b)p\big)$; no $p \times p$ object appears. The trace $\operatorname{tr} \widehat{S}_a = \frac{1}{n_a}\|G_a\|_F^2$ costs $O(m_a p)$. For $m_a \asymp m_b \asymp m$ the discrepancy is a factor $\Theta(m)$ more expensive; without balance the self-Gram terms dominate when one group is much larger. Both are linear in p. □

Corollary 4 (The Frobenius discrepancy bounds the second-order channel)**.** *Let $\widehat{\mathcal{P}}(\phi) = \sum_{a<b} \|\widehat{F}_a^{\mathrm{R,obs}} - \widehat{F}_b^{\mathrm{R,obs}}\|_F^2$ be computed by (4) from the* observed-label *per-example gradients actually used in training, and let $\zeta_c = \|\widehat{F}_c^{\mathrm{R,obs}} - F_c^{\mathrm{R}}\|_2$ measure its departure from the model-expected reachable Fisher of (1). Then for $\delta \in \mathcal{H}_\phi$,*

$$\mathcal{G}_{ab}(\varepsilon) \;\le\; \|g_a - g_b\|\,\varepsilon \;+\; \tfrac{1}{2}\big(\sqrt{\widehat{\mathcal{P}}(\phi)} + \zeta_a + \zeta_b + \eta_{ab}^{\mathcal{H}}\big)\varepsilon^2 \;+\; \tfrac{M}{3}\varepsilon^3.$$

Driving $\widehat{\mathcal{P}} \to 0$ therefore bounds the second-order channel only up to $\zeta_a + \zeta_b + \eta_{ab}^{\mathcal{H}}$, and does not *bound the gap, because the $O(\varepsilon)$ slope term remains and generally dominates.*

Proof. Each summand of $\widehat{\mathcal{P}}$ is nonnegative, so $\|\widehat{F}_a^{\mathrm{R,obs}} - \widehat{F}_b^{\mathrm{R,obs}}\|_2 \le \|\widehat{F}_a^{\mathrm{R,obs}} - \widehat{F}_b^{\mathrm{R,obs}}\|_F \le \sqrt{\widehat{\mathcal{P}}(\phi)}$. The training statistic is the observed-label empirical matrix, whereas Proposition 3 is stated for the model-expected F_c^{R}; two applications of the triangle inequality give

$$\|F_a^{\mathrm{R}} - F_b^{\mathrm{R}}\|_2 \;\le\; \sqrt{\widehat{\mathcal{P}}(\phi)} + \zeta_a + \zeta_b.$$

Combining with $\|C_{ab}\|_2^{\mathcal{H}} \le \|F_a^{\mathrm{R}} - F_b^{\mathrm{R}}\|_2 + \eta_{ab}^{\mathcal{H}}$ and substituting into the $\mathcal{G}_{ab}$ bound of Theorem 2, which retains the slope term, gives the display. The ζ_c are not controlled by the penalty and do not vanish with batch size in general: the observed-label empirical Fisher is neither the expected Fisher nor an unbiased estimator of it (Kunstner et al., 2019), so ζ_c contains a systematic component as well as sampling error. □

Remark 7 (The slope term cannot be omitted). Take the fixed-variance Gaussian model $p_\phi(y) = \mathcal{N}(\phi, 1)$ covered by **A2**, evaluated at $\phi = 0$, with the group *distributions* $Y_a \equiv -1$ and $Y_b \in \{-1, +1\}$ equiprobably. This specifies the laws, not only their means, so that every term but the slope is exactly zero. The observed-label gradient is $\nabla_\phi \ell = \phi - y$, so at $\phi = 0$ the observed-label empirical Fishers are $\mathbb{E}[y^2] = 1$ for both groups, matching the model Fisher $F_a^{\mathrm{R}} = F_b^{\mathrm{R}} = 1$; hence $\widehat{\mathcal{P}} = \zeta_a = \zeta_b = 0$. Also $H_a = H_b = 1$ so $\eta_{ab} = 0$, and $M = 0$. But the mean gradients are $g_a = 1$ and $g_b = 0$, so $g_a - g_b = 1$, the gap change is exactly δ, and $\mathcal{G}_{ab}(\varepsilon) = \varepsilon$. A bound omitting the slope term would assert $\mathcal{G}_{ab}(\varepsilon) \le 0$. This is why Corollary 4 is a bound on the second-order channel only, and why the Frobenius discrepancy of Section C.4 is not presented as a gap guarantee.

B.5 Performance gaps

Theorem 5 (Near-Bayes convergence, not a floor)**.** *For every ϕ, with $\gamma_{ab} = |R_a^\star - R_b^\star|$,*

$$\big|\, |R_a - R_b| - \gamma_{ab} \,\big| \;\le\; \max\{\psi^{-1}(\epsilon_a), \psi^{-1}(\epsilon_b)\},$$

and hence

$$|R_a - R_b| \;\ge\; \max\big\{0,\; \gamma_{ab} - \max_c \psi^{-1}(\epsilon_c)\big\}.$$

Consequently $|R_a - R_b| \to \gamma_{ab}$ along any sequence of predictors whose excess surrogate risks both tend to zero.

Proof. Decompose $R_a - R_b = (R_a^\star - R_b^\star) + (R_a - R_a^\star) - (R_b - R_b^\star)$. By classification calibration, $0 \le R_c - R_c^\star \le \psi^{-1}(\epsilon_c)$ (Bartlett et al., 2006), so the difference of the excess terms lies in $[-\psi^{-1}(\epsilon_b), \psi^{-1}(\epsilon_a)]$ and has absolute value at most $\max_c \psi^{-1}(\epsilon_c)$. The reverse triangle inequality applied to $|R_a - R_b|$ and $\gamma_{ab} = |R_a^\star - R_b^\star|$ gives the two-sided display, and rearranging with $|R_a - R_b| \ge 0$ gives the lower bound. If $\epsilon_a, \epsilon_b \to 0$ along a sequence then $\psi^{-1}(\epsilon_c) \to 0$, since ψ is continuous, convex and vanishes only at 0, so $|R_a - R_b| \to \gamma_{ab}$. □

The lower bound is $\max\{0, \gamma_{ab} - \max_c \psi^{-1}(\epsilon_c)\}$ and is vacuous whenever some group carries excess risk at least γ_{ab}; Remark 8 shows this is not an artifact of the bound, since gaps strictly below γ_{ab} are attainable by degrading the better-served group. We therefore make no claim that γ_{ab} lower-bounds the achievable gap.

Remark 8 (γ_{ab} is not a lower bound). A predictor with $R_a^\star = 0.2$, $R_b^\star = 0.1$ and $R_a = R_b = 0.2$ achieves gap $0 < \gamma_{ab} = 0.1$, consistent with the inequality because group b then carries excess risk 0.1. Gaps below γ_{ab} are therefore attainable, but only by degrading the better-served group, that is, by levelling down. What Theorem 5 rules out is closing the gap *while* both groups approach their Bayes risk.

Remark 9 (Balanced error and several groups). For binary labels, balanced error $\frac{1}{2}(\mathrm{FNR}_a + \mathrm{FPR}_a)$ is the 0–1 risk of group a under the reweighted distribution $\widetilde{\mathcal{D}}_a$ that keeps the class-conditional laws of $\mathcal{D}_a$ and sets $\Pr(Y{=}1) = \frac{1}{2}$. Applying Theorem 5 to $\widetilde{\mathcal{D}}_a$ and $\widetilde{\mathcal{D}}_b$, with L_c the class-balanced surrogate risk and $R_c^\star$ the balanced Bayes risk, gives the same conclusion for balanced error; the proof is unchanged. For a maximum gap over several groups, apply the theorem to each pair:

$\max_{a,b} |R_a - R_b|$ converges to $\max_{a,b} \gamma_{ab}$ when every excess risk vanishes. Multiclass balanced accuracy would require a multiclass calibration inequality, and MAE a separate result for conditional-median prediction; we prove neither.

Remark 10 (Consistency with the measurements). For binary classification Theorem 5 and Remark 9 suggest one mechanism for Table 6: where adaptation leaves the model close to its per-group Bayes risk, the gap is close to the Bayes-risk difference regardless of curvature dispersion. The reported gap ratios, however, use multiclass balanced accuracy (Fitzpatrick17k) and MAE (UTKFace, FHIBE face), which the theorem does not cover, so for those entries the mechanism is an analogy, not an explanation, and for UTKFace we offer none. We did not measure excess risks or Bayes risks, and alternative explanations (estimation noise, the single seed, coordinate dependence of the CV ratios, and the 15 non-independent method–task pairs behind $\rho = -0.17$) are not excluded. The theory does identify where matching could show an effect: where excess risk is large relative to the Bayes-risk difference, i.e. under-fitted or shifted subgroups.

C Training objective

Section 3 gives sufficient conditions and diagnostic quantities; it does not show that a successful adapter must equalize any trace or matrix discrepancy. Theorem 2 shows that, for additive updates in the horizontal subspace $\mathcal{H}_\phi$ of Section 3.1, the worst-case *second-order* term of the subgroup gap over a trust region of radius ε is exactly $\frac{1}{2}\|C_{ab}\|_2^{\mathcal{H}}\varepsilon^2$, which agrees with $\frac{1}{2}\|F_a^{\mathrm{R}} - F_b^{\mathrm{R}}\|_2^{\mathcal{H}}\varepsilon^2$ to within the defect $\eta_{ab}^{\mathcal{H}}$. A natural target is therefore a *matrix* discrepancy between reachable Fishers, for that channel only, since the slope term $\|g_a - g_b\|\varepsilon$ dominates it for small updates and is not addressed by any curvature statistic. This section builds a trainable objective around that target. We first describe the per-example quantity the implementation computes (the squared norm of the loss gradient with respect to the parameters actually being trained, whose subgroup mean is exactly a reachable-Fisher trace), then, in Section C.4, the Frobenius discrepancy, a computable upper bound on the operator norm that appears in the magnitude bound, which Proposition 2 makes computable without forming any $p \times p$ matrix and which we use to audit the trained models. Figure 1 summarizes the pipeline.

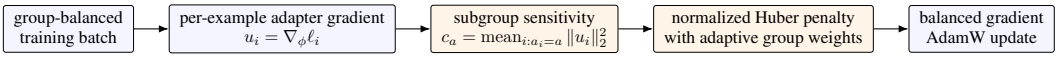


Figure 1: Training pipeline. After training, the pairwise Gram discrepancy of Equation 16 audits both arms (Section C.4). The task loss uses every row in the batch. The curvature penalty uses observed-label gradients with respect to the trainable adapter parameters and task head, aggregates their squared norms by subgroup, and combines the resulting penalty gradient with the task gradient.

C.1 Model, groups, and trainable parameters

Let $\mathcal{D} = \{(x_i, y_i, a_i)\}_{i=1}^N$ contain an input x_i, target y_i, and recorded subgroup label $a_i \in \{1, \dots, K\}$. For each experiment, these labels define the K subgroups whose sensitivities are matched during training.

The paired runs of Section 6 adapt the backbone with LoRA, DoRA or PiSSA (Hu et al., 2021; Liu et al., 2024; Meng et al., 2024). LoRA and PiSSA use $W = W_0 + s\,BA$, with PiSSA initializing (A, B) from the leading singular components of W_0. For DoRA the adapted weight is

$$W = m \frac{W_0 + BA}{\|W_0 + BA\|_c}, \qquad B \in \mathbb{R}^{d_{\mathrm{out}} \times r}, \quad A \in \mathbb{R}^{r \times d_{\mathrm{in}}}, \tag{9}$$

where W_0 is frozen, A and B form the low-rank update, and m is a trainable column-wise magnitude. The paired runs of Section 6 use rank $r = 8$, scale $\alpha = 16$ (so $s = \alpha/r$), and zero adapter dropout for all three adapters, with the task head trained in full; the diagnostic sweep uses the same LoRA-family settings, with VeRA at rank 256 (Appendix E). We use ϕ to denote the complete set of trainable parameters, consisting of the adapter parameters and the task head. All gradients used by the proposed method are calculated with respect to ϕ.

C.2 SUBGROUP SENSITIVITY FROM PER-EXAMPLE GRADIENTS

Let $\ell_i(\phi) = \ell(f_{\theta,\phi}(x_i), y_i)$ be the unreduced observed-label loss, where θ denotes the frozen parameters. For each example, the implementation computes

$$u_i = \nabla_\phi \ell_i(\phi), \qquad t_i = \|u_i\|_2^2. \tag{10}$$

For subgroup a, the corresponding observed-label empirical reachable Fisher (in the notation of Section 3; it is $p \times p$, not the $P \times P$ full-model F_a) would be

$$\widehat{F}_a^{\mathrm{R,obs}} = \frac{1}{n_a} \sum_{i \in \mathcal{B}: a_i = a} u_i u_i^\top. \tag{11}$$

Its trace measures total gradient sensitivity across all trainable directions and gives one scalar comparable across subgroups, computable exactly without constructing the full matrix. It is a *necessary* matching condition (equal matrices have equal traces), but by Theorem 4 it is not sufficient, and its insufficiency is extremal: among positive semidefinite matrices of equal trace τ, the spectral discrepancy can be as large as τ. Lemma 2 adds that the empirical trace decomposes as $\operatorname{tr} \widehat{F}_a^{\mathrm{R,obs}} = \|\hat{g}_a\|_2^2 + \operatorname{tr} \widehat{\Sigma}_a$, so a trace penalty equalizes a *sum* of a mean-gradient term and a within-group dispersion term, and the two can trade off while the penalty reports success. The mean-gradient term is related to, but does not determine, the slope channel $g_a - g_b$ of Proposition 1: equal norms $\|\hat{g}_a\| = \|\hat{g}_b\|$ do not imply $\hat{g}_a = \hat{g}_b$. We therefore treat the trace statistic below as the base estimator and the training signal, and Section C.4 complements it with a matrix discrepancy that controls the magnitude of the curvature channel and audits what the trace penalty achieved. By linearity of the trace, $\operatorname{tr}(u_i u_i^\top) = \|u_i\|_2^2$, so the average of the scalars t_i is exactly $\operatorname{tr} \widehat{F}_a^{\mathrm{R,obs}}$. In contrast, explicitly storing $\widehat{F}_a^{\mathrm{R,obs}}$ costs $O(p^2)$ for $p = \dim(\phi)$. The ConvNeXt–DoRA runs contain approximately 1.58 million trainable parameters; one dense FP32 subgroup matrix would therefore contain about 2.5×10^{12} entries and require roughly 10 TB. Our matrix-free computation follows the same principle as per-example-gradient and matrix–vector-product curvature methods, but it does not require a randomized trace estimator because the trace of each outer product has the closed form above (Pearlmutter, 1994; Dangel et al., 2019; Hutchinson, 1989). For the examples from subgroup a in the current batch $\mathcal{B}$, we compute

$$c_a(\phi) = \frac{1}{n_a} \sum_{i \in \mathcal{B}: a_i = a} t_i, \qquad n_a = |\{i \in \mathcal{B} : a_i = a\}|. \tag{12}$$

Thus, c_a measures how strongly the current examples from subgroup a act on the parameters being adapted. A larger value indicates a stronger local response in the trainable parameter space. The proposed objective uses this signal to bring subgroup sensitivities closer together during training.

C.3 NORMALIZING AND PENALIZING SUBGROUP DIFFERENCES

The objective is motivated by Theorem 1 rather than by the diagnostic sweep: when slopes match and the reachable-Fisher difference is positive definite on the horizontal subspace with restricted minimum eigenvalue above the Hessian–Fisher defect, every small nonzero horizontal update increases the signed loss gap, so reducing the discrepancy is one available intervention, and the one that is measurable at scale. A trace penalty reaches that condition only when the two groups' reachable Fishers are Loewner ordered (Theorem 4(iii)), which we do not measure; applied to every pair, as here, it is a heuristic suggested by that special case rather than an intervention the theory guarantees. It is not the only one: the defect is another lever (Remark 5), and none of the three conditions (slope matching, a small defect, the horizontal-step restriction) is enforced by the procedure below. The sweep of Section 5 supplies a consistency check on the ranking premise, but the motivation does not rest on it. During training, we use the observed-label statistic c_a because it is computed directly from the loss gradients used to update the model; Section C keeps it distinct from the expected-Fisher geometry used in evaluation.

Raw trace values vary by orders of magnitude across datasets, parameterizations, and training epochs, and can be much larger than the task loss. We therefore compare each subgroup with a detached within-batch reference. In the default method, the reference is the unweighted mean over the constrained subgroups present in the batch:

$$c_{\mathrm{ref}} = \mathrm{sg}\left[\frac{1}{|\mathcal{A}_B|} \sum_{a \in \mathcal{A}_B} c_a\right], \qquad e_a = \frac{c_a - c_{\mathrm{ref}}}{\max\{c_{\mathrm{ref}}, \epsilon_{\mathrm{den}}\}}, \tag{13}$$

where $\mathcal{A}_B$ is the set of constrained groups present in the batch, sg denotes stop-gradient, and $\epsilon_{\text{den}} = 10^{-8}$ is a fixed floor that keeps the ratio defined when every constrained group has zero statistic (the same floor is used in Equations 17 and 18). The quantity e_a is the normalized version of c_a. It measures the difference between subgroup a and the batch reference as a proportion of that reference. For example, $e_a = 0.2$ means that c_a is 20% above the subgroup mean, while $e_a = -0.2$ means that it is 20% below the mean. This normalization places all subgroup differences on a common scale, even when the raw trace values vary across datasets or training stages. The reference is detached so that it provides a fixed target during the current update.

We apply a Huber penalty to the normalized deviations,

$$\rho_\kappa(e) = \begin{cases} \frac{1}{2}e^2, & |e| \le \kappa, \\ \kappa(|e| - \frac{1}{2}\kappa), & |e| > \kappa, \end{cases} \qquad \kappa = 1. \tag{14}$$

Here ρ_κ denotes the Huber loss and κ is its transition threshold (Huber, 1992) (we avoid δ, which denotes an adapter update in Section 3). The implementation uses $\kappa = 1$. Since e_a is a relative deviation, this threshold corresponds to an absolute difference equal to the reference value. The loss is quadratic when $|e_a| \le 1$, so small subgroup differences are corrected smoothly. It becomes linear when $|e_a| > 1$, which limits the effect of an unusually large batch estimate. In this way, the penalty remains sensitive near the matching point without allowing one extreme subgroup value to dominate the update. With subgroup weights μ_a, the regularizer is

$$R(\phi, \mu) = \sum_{a \in \mathcal{A}_B} \mu_a \rho_\kappa(e_a). \tag{15}$$

The default reference is the subgroup mean. The ablations replace it with the minimum subgroup trace or replace Huber with linear, absolute, Euclidean-norm, or one-sided hinge penalties. These alternatives are specified as variants of the same pipeline rather than as separate methods.

C.4 A MATRIX DISCREPANCY FOR AUDITING TRAINED MODELS

Theorem 4 shows that trace equality cannot control the operator norm that sets the worst-case curvature channel; Corollary 4 shows that the Frobenius discrepancy does bound it, up to the estimator and defect terms. It is one sufficient, computable matrix norm, not the only one, and like any scalar norm it does not certify the definiteness condition of Corollary 1. Let $G_a \in \mathbb{R}^{n_a \times p}$ collect the observed-label per-example gradients u_i of the constrained rows of subgroup a in the current batch, so $\widehat{F}_a^{\mathrm{R,obs}} = \frac{1}{n_a} G_a^\top G_a$. By Proposition 2 the pairwise Frobenius discrepancy is available without forming a $p \times p$ matrix:

$$d_{ab}^2 = \left\| \widehat{F}_a^{\mathrm{R,obs}} - \widehat{F}_b^{\mathrm{R,obs}} \right\|_F^2 = \frac{\|G_a G_a^\top\|_F^2}{n_a^2} - \frac{2\|G_a G_b^\top\|_F^2}{n_a n_b} + \frac{\|G_b G_b^\top\|_F^2}{n_b^2}. \tag{16}$$

Only the $n \times n$ Gram matrices $G_a G_b^\top$ appear. For a balanced batch with n rows per group and $|\mathcal{A}_B|$ constrained groups, all pairs cost $O(|\mathcal{A}_B|^2 n^2 p)$ time and $O(|\mathcal{A}_B|^2 n^2 + |\mathcal{A}_B| np)$ memory, against $O(|\mathcal{A}_B| np)$ for the trace. As an operation count on the statistic alone this is a factor $\Theta(n)$ more work; with at most 64 rows per group the Gram blocks are at most 64×64, so the overhead is bounded by the per-group batch size on an operation that already materializes G_a. No step of (16) depends on p beyond the single matrix product, which is why the discrepancy is affordable at the 1.58M trainable parameters of the ConvNeXt–DoRA runs, where a dense $\widehat{F}_a^{\mathrm{R,obs}}$ would need roughly 10 TB.

That count is not the wall-clock cost, because it omits what the two penalties share. Both gradients are a sum of one Hessian-vector product per example and differ only in the vector: $\sum_a (\partial R/\partial c_a)(2/n_a) \sum_{i \in a} H_i u_{ai}$ for the trace, and $\sum_i H_i v_i$ with $v = (W + W^\top)G, W = \partial R/\partial K$, for the Frobenius form. Each product can be formed and discarded one example at a time, so neither keeps more than one second-order graph alive, and the Frobenius penalty pays extra only for the detached G and v and for the small Gram matrix. Table 4 measures single training steps from a stored checkpoint at the training batch composition. The shared per-example term dominates: the Frobenius penalty costs $1.00\times$ the trace penalty on ConvNeXt–DoRA and $0.93\times$ on ViT–LoRA, with identical peak memory, while both cost 23–372$\times$ a step with no penalty. We therefore use the trace during training for scalability and simplicity rather than for speed, and reserve the Frobenius discrepancy for the post-hoc audit, whose statistic alone costs 1–2% of a penalized step.

Table 4: Measured cost of one training step, median of 20 timed steps after 3 warm-up steps on UTK-Face at the training batch composition (25 examples per subgroup $\times$ 5 groups). The optimizer update is excluded; it is identical in every row. NVIDIA RTX 5090, PyTorch 2.14.0+cu130, CUDA 13.0.

	ConvNeXt–DoRA ($p = 1.58$M)			ViT–LoRA ($p = 1.34$M)		
Step	s/step	$\times$trace	peak MB	s/step	$\times$trace	peak MB
no penalty	0.478	—	14,656	0.289	—	6,539
trace penalty	177.63	1.00	14,668	6.562	1.00	6,549
Frobenius penalty	176.98	1.00	14,668	6.093	0.93	6,549
trace statistic only	2.953	0.02	2,568	1.556	0.24	2,195
Frobenius statistic only	2.692	0.02	5,832	1.436	0.22	5,248

Normalization and robustification carry over unchanged from the trace penalty. Writing $s_{ab} = \mathrm{sg}[\frac{1}{2}(\mathrm{tr}\, F_a^{\mathrm{R}} + \mathrm{tr}\, F_b^{\mathrm{R}})]$ for a detached scale reference (which by Lemma 1 is the same reparameterization-sensitive quantity the trace penalty already used, so no new coordinate dependence is introduced), the normalized discrepancy and the regularizer are

$$\tilde{e}_{ab} = \frac{d_{ab}}{\max\{s_{ab}, \epsilon_{\mathrm{den}}\}}, \qquad R_{\mathrm{F}}(\phi, \mu) = \sum_{a<b} \mu_{ab}\, \rho_\kappa(\tilde{e}_{ab}), \tag{17}$$

with ρ_κ the Huber loss of Equation 14 and μ_{ab} pair weights updated by the same bounded exponentiated rule as Equation 18, with ν_{ab} formed from the moving average of $\tilde{e}_{ab}$ instead of the subgroup gap. The discrepancy $d_{ab} \geq 0$ is zero iff the two empirical reachable Fishers agree, and unlike a trace difference it cannot hide directional differences through cancellation of eigenvalues of opposite sign. It is *not* one-sided: minimizing it may lower one group's matrix, raise the other's, or move both, so matching can still be achieved by increasing a subgroup's curvature; the task loss is what discourages the degenerate versions of this. Corollary 4 then gives the property the trace objective lacks, with all three gaps between penalty and gap made explicit: writing $\widehat{\mathcal{P}}(\phi) = \sum_{a<b} d_{ab}^2$ for the quantity actually computed from observed-label gradients,

$$\mathcal{G}_{ab}(\varepsilon) \leq \|g_a - g_b\|\varepsilon + \tfrac{1}{2}\big(\sqrt{\widehat{\mathcal{P}}} + \zeta_a + \zeta_b + \eta_{ab}^{\mathcal{H}}\big)\varepsilon^2 + \tfrac{M}{3}\varepsilon^3$$

for every pair and every horizontal update direction. Driving the penalty down bounds the second-order channel *up to* $\zeta_a + \zeta_b + \eta_{ab}^{\mathcal{H}}$; it does *not* bound the gap, because the $O(\varepsilon)$ slope term survives and generally dominates. The ζ_c arise because this section computes an observed-label empirical matrix while the theory is stated for the model-expected reachable Fisher, and they carry a systematic component that larger batches do not remove (Kunstner et al., 2019). This is why the penalty is combined with the task loss and the balanced step below rather than offered as a standalone guarantee.

Three practical caveats follow from the theory rather than from tuning. Zeroth, by Proposition 4 and Corollary 3 the factorized adapter, including DoRA, carries a $\mathrm{GL}(r)$ gauge symmetry. At a fixed point, each reachable Fisher annihilates the at least r^2 gauge directions of $\ker J$ per adapted weight (for $r \leq \min(d_{\mathrm{in}}, d_{\mathrm{out}})$, as in all our runs). The scalar statistics are nevertheless *not* constant along a gauge orbit, because J itself changes along it. In the scalar case $T(a, b) = ab$ with full-model Fisher f_c, $F_c^{\mathrm{R}} = f_c\left(\begin{smallmatrix} b^2 & ab \\ ab & a^2 \end{smallmatrix}\right)$ and $\mathrm{tr}\, F_c^{\mathrm{R}} = f_c(a^2 + b^2)$: the gauge change $(a, b) \mapsto (ta, b/t)$ preserves the predictor ab but changes the trace. Traces, Frobenius discrepancies and the penalty can therefore move along function-preserving directions, and a comparison of two arms of the same adapter family can partly reflect different factor scalings. The theorems are unaffected (they are stated in Euclidean adapter coordinates), and the definiteness conditions of Section 3 hold only for horizontal additive updates, whereas the optimizer below is not projected onto $\mathcal{H}_\phi$; by Remark 2 a non-horizontal step can move the gap the other way at second order, particularly at the rank-deficient $B = 0$ initialization. The audit below removes the $\mathrm{GL}(r)$ freedom by evaluating each checkpoint at its *balanced* representative. First, d_{ab} is a p-dimensional matrix discrepancy estimated from $n_a + n_b$ gradients, so it is rank-deficient whenever $n_a + n_b < p$; the estimate is then of the discrepancy restricted to the span of the sampled gradients, and balanced batches matter more for (16) than for the trace. Second, Lemma 1 implies d_{ab} is not invariant to general reparameterization, nor even to isotropic rescaling: under $K = cI$ it scales as $d_{ab} \mapsto c^2 d_{ab}$, and only the normalized $\tilde{e}_{ab} = d_{ab}/s_{ab}$ of Equation 17 is invariant there. It may therefore be compared across subgroups within a fixed adapter but not across adapter families; the inertia of $F_a^{\mathrm{R}} - F_b^{\mathrm{R}}$ is a coordinate-invariant alternative

that we do not report here. A training arm built on R_{F} should use the identical sampler, initialization, optimizer, schedule and update budget as the trace arm (Section C.7), so that their difference isolates the statistic.

Audit protocol. The audit measures what the trace penalty did to the matrix it cannot control. For each of the 30 primary cells we load the final `plain` and trace-matching checkpoints and, in evaluation mode, draw the *same* 20 group-balanced held-out batches of $n = 64$ examples per subgroup for both.

Gauge canonicalization. Before measuring, every factorized layer is replaced by its balanced representative: with the rank-r SVD $BA = U\Sigma V^\top$, set $B \leftarrow U\Sigma^{1/2}$ and $A \leftarrow \Sigma^{1/2}V^\top$, so that $B^\top B = AA^\top$; any DoRA magnitude is left unchanged. This preserves the predictor, and balanced factorizations of a full-rank product differ only by an orthogonal $G \in \mathrm{O}(r)$, which acts on the adapter coordinates as an orthogonal map; traces and Frobenius norms of the reachable Fishers are invariant under it. The audit statistics are therefore invariant to the $\mathrm{GL}(r)$ gauge, and `validate_exact.py` checks this on a small network (Appendix J). We also report the unbalanced values.

Statistics. Let u_{ai} be the observed-label per-example gradient of example i of group a with respect to the trained parameters, K their Gram matrix, and $s_{ab} = \frac{1}{2}(\operatorname{tr} \widehat{F}_a^{\mathrm{R,obs}} + \operatorname{tr} \widehat{F}_b^{\mathrm{R,obs}})$. For every subgroup pair we compute the trace gap, the plug-in Frobenius discrepancy d_{ab}^2 of (16), and the unbiased estimator of the squared discrepancy

$$\widehat{d}^2_{ab,\mathrm{U}} = \frac{\sum_{i\neq j}(u_{ai}^\top u_{aj})^2}{n_a(n_a-1)} + \frac{\sum_{i\neq j}(u_{bi}^\top u_{bj})^2}{n_b(n_b-1)} - \frac{2\sum_{i,j}(u_{ai}^\top u_{bj})^2}{n_a n_b},$$

whose expectation, for independent draws, is the squared Frobenius distance between the two groups' population observed-label reachable Fishers (the plug-in value adds a positive per-example noise term that can differ between arms). It can be negative in finite samples; we report the normalized *squared* value $\widehat{d}^2_{ab,\mathrm{U}}/s_{ab}^2$ directly, without clipping or taking a square root. The nonzero eigenvalues of $\widehat{F}_a^{\mathrm{R,obs}} - \widehat{F}_b^{\mathrm{R,obs}} = G^\top D G$, with $D = \operatorname{diag}(n_a^{-1}\mathbf{1}, -n_b^{-1}\mathbf{1})$, equal those of the small matrix $K^{1/2}DK^{1/2}$, so the operator norm on the span of the sampled gradients costs one $(n_a{+}n_b)$-dimensional eigendecomposition. From the same quantities we obtain the surrogate-loss gap from the per-example losses.

Analysis. All quantities are compared between the two arms of a cell only. The primary audit statistic, fixed before measurement, is $\widehat{d}^2_{ab,\mathrm{U}}/s_{ab}^2$ averaged over pairs; the normalized trace gap is the manipulation check, the operator-to-Frobenius ratio measures how tight the Frobenius bound is, and the surrogate-loss gap connects the audit to the loss channel of Proposition 1. Cells are compared pairwise by win counts, median paired changes with a cell bootstrap, and sign tests; across cells we report Spearman correlations of the change in trace gap with the change in Frobenius discrepancy, and of the change in Frobenius discrepancy with the changes in surrogate-loss and performance gaps. All are read descriptively, like the rest of Section 6. `measure_frobenius.py` implements the audit and checks every identity above against dense matrices.

C.5 Adaptive subgroup weights

A uniform penalty would apply the same strength to every subgroup throughout training. We instead adjust the subgroup weights so that groups with consistently high sensitivity receive greater emphasis. This follows the general idea of group-specific weighting used in constrained fairness and group-robust optimization (Agarwal et al., 2018; Sagawa et al., 2019).

Batch estimates of c_a can vary from one step to the next. We therefore maintain an exponential moving average $\bar{c}_a$ for each subgroup. The current batch is used to update the model parameters, while the moving averages determine how the subgroup weights change. This allows the weights to respond to persistent differences rather than to a single batch.

Let $\bar{c}_{\mathrm{ref}}$ be the mean reference computed from the moving averages, matching the reference used by the penalty. The relative gap and exponentiated update are

$$\nu_a = \operatorname{clip}\left(\frac{\bar{c}_a - \bar{c}_{\mathrm{ref}}}{\max\{\bar{c}_{\mathrm{ref}}, \epsilon_{\mathrm{den}}\}}, -\nu_{\max}, \nu_{\max}\right), \qquad \log \mu \leftarrow \Pi_\tau(\log \mu + \eta_\mu \nu). \tag{18}$$

The relative gap ν_a (we avoid γ, which denotes a Bayes-risk difference in Section 3) compares the smoothed sensitivity of subgroup a with the smoothed subgroup mean. If a subgroup remains above the mean, ν_a is positive and its weight increases. If it remains below the mean, its weight decreases. Therefore, the penalty does not treat all subgroup deviations at a fixed rate; it places more emphasis on groups whose high sensitivity persists over training.

The update is performed in log-weight space, following exponentiated or mirror-ascent weighting (Beck & Teboulle, 2003). The operator Π_τ centers the log-weights and clamps them to $[-\tau, \tau]$, and $\mu = \mathrm{softmax}(\log \mu)$ converts them into non-negative weights that sum to one. These two operations prevent the weighting from becoming concentrated on only one subgroup.

We use moving-average coefficient 0.9, an epoch-level log-weight rate of 0.3, gap bound $\nu_{\max} = 1$, and trust-region bound $\tau = 3$. The per-step value of η_μ is derived from the number of batches in an epoch, making the rate comparable across datasets and batch sizes.

C.6 Why the task and mitigation terms are combined

For one training step, the intended update combines the pooled task loss with the curvature regularizer:

$$\mathcal{L}_{\mathrm{step}}(\phi) = \mathcal{L}_{\mathrm{task}}(\phi) + \omega R(\phi, \mu). \tag{19}$$

Neither term is sufficient alone. The pooled task loss is required to learn the predictor, but it constrains only average fit: two parameter settings can have similar pooled loss while having very different subgroup gradient norms, and groups that generate larger gradients can dominate its update even in a count-balanced batch. This is precisely the degree of freedom targeted by R. Conversely, R alone says nothing about fitting y. It is minimized whenever the subgroup statistics match, including degenerate solutions that are uniformly insensitive and inaccurate; with a symmetric mean reference the trace form can also match by raising below-mean traces. The Frobenius form of Equation 16 shares this property (it can be reduced by raising either group's matrix) and differs from the trace form only in that equal traces no longer imply a zero penalty. The task loss therefore anchors predictive utility, while the mitigation term selects among useful descent directions by discouraging unequal subgroup sensitivity.

The two gradients are computed separately. Differentiating $t_i = \|\nabla_\phi \ell_i\|^2$ introduces Hessian–vector products, and its gradient can be much larger than the ordinary task gradient even when the penalty value is small. In an uncapped development run, the penalty gradient was about 15 times the task-gradient norm; clipping only the summed gradient then effectively removed the task direction and balanced accuracy remained at chance (0.500). We therefore set

$$\omega = \begin{cases} \min\left(1, \beta \dfrac{\|g_{\mathrm{task}}\|_2}{\|g_{\mathrm{pen}}\|_2}\right), & g_{\mathrm{pen}} \neq 0, \\ 1, & g_{\mathrm{pen}} = 0, \end{cases} \qquad g = g_{\mathrm{task}} + \omega g_{\mathrm{pen}}, \qquad \beta = 0.5. \tag{20}$$

(We write ω rather than s to avoid a clash with the adapter scale of Proposition 3.) This caps the curvature-gradient norm at one half of the task-gradient norm while preserving the relative subgroup weights. It also gives a direct first-order safeguard. For $\beta < 1$,

$$g_{\mathrm{task}}^\top g \geq (1 - \beta)\|g_{\mathrm{task}}\|_2^2 > 0, \tag{21}$$

unless the task gradient is zero. Hence the negative combined gradient cannot reverse the task descent direction before AdamW preconditioning; the mitigation term can rotate that direction, but not overwhelm it. The combined gradient is subsequently clipped to norm 1 before the AdamW update. This mechanism explains why the two terms can operate together; whether they improve held-out utility or fairness remains an empirical question rather than a consequence of the bound.

The image models require chunked computation. The implementation first computes the trace values without retaining a second-order graph, uses them to obtain c_a and the derivative of R with respect to each c_a, and then recomputes the weighted trace in small chunks. Because c_a is a mean, each example receives the detached weight $(\partial R/\partial c_a)/n_a$. Summing the chunk gradients therefore gives the same penalty gradient as a full-batch computation; chunking changes peak memory, not the objective. In the reported chunked runs, every constrained row in the batch contributes to the statistic. The task loss also uses every row, including rows outside the constraint set.

C.7 Balanced batches and comparison arms

Ordinary shuffled batches can contain very few examples from a rare group, making c_a far noisier for that group than for common groups. The group-balanced sampler draws the same number of examples from every constrained subgroup, with replacement when necessary. If unconstrained categories exist, an additional slice from those categories is included for the task loss. Because this sampler changes how often examples are seen, an epoch is a fixed number of balanced draws rather than one pass through the dataset.

The protocol specifies training arms sharing a sampler, schedule, seed, and model initialization. The `plain` arm uses only the pooled task loss. The `curvature` arm uses the trace statistic of Equation 12. Holding the sampler fixed across arms isolates the effect of the subgroup statistic from the effect of oversampling. The pairwise discrepancy of Equation 16 is used as the matrix audit of these arms; it could also replace the trace statistic in a `frobenius` training arm, which would isolate the effect of penalizing a matrix discrepancy rather than a trace (Theorem 4, Corollary 4).

What this submission reports. Section 6 compares the `plain` and `curvature` (Huber trace) arms on a 30-cell grid and audits both with the Frobenius discrepancy. Theorem 5 suggests that any matching intervention should move measured gaps most where excess risk is large relative to the Bayes-risk difference (under-fitted or shifted subgroups) rather than in the in-distribution settings evaluated here.

Paired runs use AdamW with learning rate 5×10^{-4}, no weight decay, gradient clipping at 1, and seed 12345. They are scheduled for at most 60 epochs with patience 10, and the retained checkpoint is selected by the pooled validation metric rather than by a fairness gap. Batch size is tuned per cell as a hyperparameter, over values between 64 and 1024; the selected value is 64–512 for 29 of the 30 cells and 1024 for UTKFace ViT–DoRA, and every selected value is reported in Table 10 of Appendix G.

C.8 Post-training evaluation

Post-training evaluation is necessary because the training objective establishes only an optimization-space intervention. It serves four separate checks. First, the matrix audit of Section C.4 tests on held-out batches whether the quantity acted on during training (the trace) and the matrix discrepancy it cannot control actually became more similar across groups. Second, task performance, both its mean and its worst subgroup, tests the failure mode in which traces match because useful learning collapsed. Third, direct outcome metrics such as subgroup performance range, equal-opportunity difference, and average-odds difference test fairness in prediction space; these cannot be inferred from R. Fourth, comparison with the `plain` arm, which shares the balanced sampler, determines whether a change is attributable to curvature matching rather than to oversampling alone. Section 6 reports these checks for the `plain` and `curvature` arms (the outcome metrics partly in Appendix G); a `frobenius` training arm is a natural extension.

The training penalty uses the observed-label trace because that is the quantity implemented by the differentiable loss-gradient path. Evaluation also computes the expected-Fisher trace of Equation 5, as an out-of-objective check: agreement would show that the intervention transfers from the empirical proxy to the model-distribution geometry studied in the diagnostic sweep, whereas disagreement reveals proxy mismatch rather than hiding it behind a shared estimator. A smaller curvature disparity is treated as evidence that the optimization-space target changed, not as proof that prediction-level performance gaps improved.

D Extended related work

D.1 Parameter-efficient adaptation and fair fine-tuning

PEFT changes more than the number of trainable parameters: it changes the coordinates and directions through which a pretrained model can move. LoRA restricts each weight update to a low-rank product (Hu et al., 2021); rsLoRA changes its rank-dependent scaling, DoRA separates magnitude from direction, PiSSA initializes from leading singular components, and VeRA learns scaling vectors over

shared random bases (Kalajdzievski, 2023; Liu et al., 2024; Meng et al., 2024; Kopiczko et al., 2024). Expressivity results make the role of rank explicit (Zeng & Lee, 2024), while Flat-LoRA shows that a solution that is flat in adapter coordinates can remain sharp in the full parameter space (Li et al., 2024).

Broad evaluations show why this question should not be answered from one adapter or backbone. A medical-imaging benchmark spanning 17 PEFT methods and six datasets found strong performance in low-data regimes but substantial method–task variation (Dutt et al., 2023). FairMedFM similarly reports heterogeneous utility–fairness trade-offs across 20 foundation models, 17 medical datasets, and several adaptation protocols (Jin et al., 2024). Direct comparisons of LoRA and full fine-tuning find no consistent ordering in subgroup utility or calibration (Ding et al., 2024), and full-model fine-tuning studies show that worst-group accuracy can depend sharply on balancing strategy and backbone scale (LaBonte et al., 2024). Fairness-aware PEFT methods respond by selecting trainable components, regularizing group-conditioned losses, activating conditional low-rank updates, or removing sensitive information from the learned representation (Dutt et al., 2024; Sukumaran et al., 2024; Kamalaruban et al.; Zhou et al., 2026; Bhosale et al., 2026). Our method is complementary: it regularizes subgroup sensitivity in the adapter space and evaluates outcome-level fairness separately.

D.2 The Hessian, generalized Gauss–Newton matrix, and Fisher information

Several matrices called "curvature" appear in deep learning, but they answer different questions. The Hessian $H = \nabla^2_\theta L$ is the exact local second derivative of the chosen training objective. It can be indefinite in a nonconvex network and therefore records both positive and negative curvature. If the loss is a convex function of the model output, the generalized Gauss–Newton (GGN) matrix pulls the output-space loss Hessian back through the network Jacobian. The GGN is positive semidefinite and omits the residual term containing second derivatives of the network itself (Botev et al., 2017; Martens, 2020). Thus, the Hessian and GGN coincide for a model linear in its parameters, and can be close when the omitted model-curvature term is small, but they are not generally identical.

For a conditional probabilistic model, the expected Fisher is the covariance of score gradients when the label is sampled from the model. It is positive semidefinite and defines the Riemannian metric used by natural-gradient descent (Amari, 1998; Martens, 2020). Under standard regularity conditions, the model-expected negative log-likelihood Hessian equals the Fisher; for common exponential-family output models, the Fisher also equals the corresponding GGN (Martens, 2020). These identities explain why the Fisher can serve as a stable Hessian surrogate in likelihood models, but only for the relevant expectation and loss. The equality need not hold for a finite dataset, a misspecified model, an arbitrary loss, or a restricted parameterization.

A further distinction is crucial here. The empirical Fisher is the average outer product of gradients evaluated at the *observed* labels. Despite its name, it is generally neither the Fisher under the model distribution nor an unbiased Monte Carlo estimate of that Fisher; it also need not approximate the Hessian well (Kunstner et al., 2019). Its trace is simply the mean squared norm of per-example observed-label gradients. That makes it inexpensive, nonnegative, and directly tied to the examples driving an optimizer, but it should be interpreted as an optimization-sensitivity statistic rather than as exact second-order curvature. Our training penalty uses this empirical trace. Our post-training diagnostic instead samples or sums over model-predicted labels to estimate the expected Fisher. We keep the two quantities separate throughout.

D.3 Estimating curvature at neural-network scale

Forming a dense Hessian or Fisher for p parameters costs $O(p^2)$ storage, so most methods estimate only the structure needed by a downstream task. Pearlmutter's reverse-over-forward construction computes exact Hessian–vector products at roughly gradient cost without materializing the matrix (Pearlmutter, 1994). Hessian-free optimization combines such products with iterative linear solvers (Martens et al., 2010), while stochastic Lanczos quadrature uses products to estimate the eigenvalue density and leading spectral structure (Ghorbani et al., 2019). PyHessian packages trace, top-eigenvalue, and spectral-density estimation for modern networks (Yao et al., 2020). These matrix-free methods preserve more global structure than a diagonal estimate, but require repeated products.

Structured approximations trade some fidelity for cheaper storage and inversion. K-FAC approximates layerwise Fisher blocks by Kronecker products (Martens & Grosse, 2015); related block-diagonal Gauss–Newton methods propagate curvature through the network (Botev et al., 2017). BackPACK extends backpropagation to recover per-example gradients, diagonal Hessian or GGN terms, and matrix-factor representations (Dangel et al., 2019). At the simplest end, diagonal Fisher, diagonal Hessian, and their traces retain only coordinatewise or aggregate sensitivity. Hutchinson's randomized estimator recovers a trace or diagonal from matrix–vector products (Hutchinson, 1989); AdaHessian combines such diagonal estimation with spatial and temporal averaging inside an optimizer (Yao et al., 2021); and low-rank or Lanczos approximations retain dominant eigendirections. The choice is task dependent: natural-gradient updates require an inverse or linear solve, sharpness studies emphasize extreme eigenvalues or neighborhoods, and pruning or importance scoring often needs only diagonal entries.

Our estimator occupies the inexpensive end of this spectrum. For each example, it obtains the gradient only with respect to the trainable DoRA and head parameters, then sums its squared entries. Averaging these scalars within a subgroup gives the trace of a subgroup empirical-Fisher matrix in the reachable adaptation coordinates. We do not form, invert, or eigendecompose a matrix. Differentiating the trace penalty during training does require second-order automatic differentiation, which motivates the chunked accumulation described in Section C.

D.4 What Fisher and Hessian estimates have been used for

The oldest use is optimization. Newton and Hessian-free methods use loss curvature, whereas natural gradient preconditions the gradient by the Fisher metric and is approximately invariant to smooth reparameterization (Amari, 1998; Martens et al., 2010; Martens, 2020). K-FAC makes a structured natural-gradient approximation practical for neural networks (Martens & Grosse, 2015). Fisher SAM similarly defines the adversarial neighborhood of sharpness-aware minimization with information geometry rather than a Euclidean norm (Foret et al., 2020; Kim et al., 2022).

Curvature also supports diagnosis and generalization studies. Work on flat minima relates local sensitivity to generalization (Hochreiter & Schmidhuber, 1997; Keskar et al., 2016), and visualization and spectral studies examine Hessian slices, outliers, and bulk eigenvalue structure (Li et al., 2018; Ghorbani et al., 2019). However, Euclidean sharpness can be changed by a function-preserving reparameterization (Dinh et al., 2017); information-geometric measures have therefore been proposed to obtain reparameterization-invariant notions of sensitivity (Liang et al., 2019; Jang et al., 2022). This caveat is especially relevant to PEFT, where the same predictor can be represented in different adapter coordinates.

A third line of work treats curvature as parameter importance. Optimal Brain Damage used diagonal second derivatives for pruning (LeCun et al., 1989); modern methods such as WoodFisher approximate an inverse empirical Fisher to predict the effect of deleting weights (Singh & Alistarh, 2020). Fisher-weighted penalties protect parameters important to old tasks in elastic weight consolidation (Kirkpatrick et al., 2017), and Fisher-weighted averaging supports model merging (Matena & Raffel, 2022). Hessian inverses also underlie influence functions that approximate how training points affect predictions (Koh & Liang, 2017). In Bayesian deep learning, the Hessian, GGN, or Fisher supplies the local precision in Laplace approximations (Ritter et al., 2018; Immer et al., 2021). These uses are related by a common local-sensitivity calculation, but they do not make the underlying matrices interchangeable.

Fisher-guided PEFT brings these ideas into restricted trainable subspaces. FISH Mask and FISH-Tuning select trainable parameters or modules (Sung et al., 2021; Xue et al., 2025); Learning in the Fisher Subspace and FiLoRA select or initialize low-rank directions (Feng et al., 2026; Han & Guo, 2026); FI-LoRA allocates rank using importance, and FoRA selects adapted layers (Zeng & Zhao, 2026; Park et al., 2026). These methods use Fisher information to decide *where* adaptation should occur. Our question is different: once a trainable subspace has been chosen, do observed subgroups act on that subspace with comparable strength?

D.5 Subgroup robustness with incomplete demographics

When group annotations are available, group DRO optimizes worst-group risk, while equal opportunity and equalized odds compare conditional error rates; constrained reductions turn such criteria into trainable objectives (Sagawa et al., 2019; Hardt et al., 2016; Agarwal et al., 2018). Subgroup fairness, multicalibration, and multiaccuracy extend evaluation beyond a small set of coarse, predeclared groups (Kearns et al., 2018; Hébert-Johnson et al., 2018; Kim et al., 2019). The choice of coarse, fine-grained, intersectional, or noisy group definitions can itself change which mitigation method appears effective (Alloula et al., 2025). This is particularly important in medical imaging, where conclusions also vary with the dataset, backbone, fairness metric, and model-selection rule (Zong et al., 2022; Jin et al., 2024).

Without complete demographic annotations, prior work replaces named groups with distributional, error-based, or representation-based surrogates. Distributionally robust and adversarially reweighted objectives protect high-loss subsets without observing their identities (Hashimoto et al., 2018; Lahoti et al., 2020). GEORGE and EIIL infer latent subclasses or environments, whereas Just Train Twice uses errors from an initial model to upweight likely minority examples (Sohoni et al., 2020; Creager et al., 2021; Liu et al., 2021a). More recent methods minimize training-loss variance without demographic priors or construct soft neighborhoods from observed-label gradients (Wang et al., 2024; Luo et al., 2025). In medical imaging, feature-based subgroup discovery can expose performance gaps that are larger than those found from available metadata (Bissoto et al., 2025). Our method does not infer latent groups: it assumes recorded subgroup labels for both aggregation and training. The expected-Fisher contribution is computed per input without a subgroup label, but forming a subgroup statistic still requires the labels at the aggregation step.

D.6 Curvature and subgroup behavior

Several fairness studies nevertheless use Hessian or Fisher-based quantities as subgroup signals. Pruning can enlarge both accuracy gaps and gaps in gradient norm or the top Hessian eigenvalue (Tran et al., 2022). CUMA matches distributions of input-space loss curvature across observed groups under shift (Wang et al., 2023). CurvFed regularizes the top eigenvalue of a parameter-space Fisher surrogate without demographics in federated training (Sharma et al., 2026); FairSAM modifies sharpness-aware training to balance robustness across groups (Dai et al., 2025); and FairMerging uses subgroup gradient and curvature statistics to control disparities introduced by weight-space model merging (Liu et al.). Among fairness-without-demographics methods, the recent FLARE preprint is closest in motivation: it combines embeddings, cross-entropy loss, and a label-dependent Fisher penalty to discover latent behavioral clusters, then trains cluster-specific models (Roy et al., 2026). All of these methods use geometry inside an optimization, aggregation, or subgroup-discovery procedure.

Our contribution differs from all of these in what it establishes rather than in which statistic it computes. Prior work uses curvature *inside* an optimization, aggregation, or subgroup-discovery procedure and validates it empirically. We instead ask what curvature in the reachable subspace *provably* controls. Section 3 shows that, once the slope channel is controlled, a reachable-Fisher difference that is positive definite on the horizontal subspace, above the Hessian–Fisher defect, forces the signed gap upward under every small horizontal update; that full-model statistics that ignore the adapter Jacobian can err in both directions; and that trace equality cannot control the operator norm that sets the worst-case curvature channel. This offers one possible explanation for why adapter-robustness findings conflict (Ding et al., 2024; Sukumaran et al., 2024; Li et al., 2024), though we do not claim it accounts for any particular disagreement in that literature. Unlike CUMA, our perturbation is in parameter rather than input space. Unlike CurvFed and FLARE, we distinguish the empirical training proxy from the expected-Fisher geometry used for evaluation, and we prove that the trace both methods rely on conflates a mean-gradient term with a dispersion term. Unlike work that uses curvature to accelerate optimization, prune parameters, or approximate a posterior, our target is a matrix discrepancy across known subgroups, with a matrix-free estimator and a guarantee attached to it. Finally, we state a limit: the performance gap converges to a Bayes-risk difference rather than being bounded below by it. This lets the premise be tested without treating geometric alignment as a fairness guarantee.

E REPRODUCIBILITY DETAILS

Table 5: Evaluation settings. Split sizes are the stored fine-tuning train/validation/test partitions; diabetes additionally uses 42,741 earlier encounters for pretraining before applying FFT or PEFT. "Axes" lists the criteria of subgroups used in curvature-difficulty correlations.

Dataset	task	output	train/val/test	axes (groups)	encoders
Fitzpatrick17k	image classification	3 classes	11,592/2,484/2,484	tone (7), binary tone (3)	17
UTKFace	image regression	age	13,796/2,956/2,957	race (5)	17
FHIBE face	image regression	age	5,556/1,192/1,210	ancestry (6)	17
Diabetes 130H	tabular classification	binary	28,495/20,353/10,177	race (6)	FT-T

E.1 TRAINING

Every model receives a fresh prediction head. Full fine-tuning updates all backbone and head parameters; PEFT updates the head and adapter only. LoRA-family rank is 8 with scale 16 and no adapter dropout; VeRA rank is 256. We use AdamW for at most 20 epochs, two warm-up epochs followed by cosine decay, gradient clipping at 1, and early stopping with patience 5. Image learning rates are 10^{-4} for full fine-tuning, 5×10^{-4} for LoRA-family methods, and 5×10^{-3} for VeRA; the tabular full-fine-tuning rate is 5×10^{-4}. Each training partition contains an internal 15% held-out slice that is never used for gradient updates, so its reported "train" metrics are out-of-fit.

E.2 IMAGE ROSTER

The image roster comprises ViT, DeiT, CrossViT, RoPE-ViT, MaxViT, Swin, ResNet-50, ConvNeXt, ConvNeXt V2, SigLIP 2, MetaCLIP 2, DINOv2, DINOv3, I-JEPA, MAE, BEiT, and EVA-02. It spans supervised, self-supervised, convolutional, hierarchical, and vision–language pretraining families (Dosovitskiy, 2020; Touvron et al., 2021; Chen et al., 2021; Heo et al., 2024; Tu et al., 2022; Liu et al., 2021b; He et al., 2016; Liu et al., 2022; Woo et al., 2023; Tschannen et al., 2025; Chuang et al., 2026; Oquab et al., 2023; Siméoni et al., 2025; Assran et al., 2023; He et al., 2022; Bao et al., 2021; Fang et al., 2024). Each compatible block is trained with full fine-tuning, LoRA, DoRA, rsLoRA, PiSSA, and VeRA. PiSSA and VeRA cannot attach to the convolution-only ResNet implementation, producing 100 models per image dataset rather than 102. With six models for the tabular architecture, the completed grid contains $3 \times 100 + 6 = 306$ models.

E.3 DATA PREPARATION AND LEAKAGE CONTROLS

Fitzpatrick17k removes 17 rows explicitly marked as wrongly labeled, deduplicates byte-identical files by hash, and stratifies a 70/15/15 split on apparent skin tone crossed with the three-way target. The source does not provide lesion or patient identifiers; near-duplicate crops may therefore cross splits. FHIBE face keeps primary-subject rows only (7,958 crops from 1,651 subjects; the secondary rows carry another person's demographics) and splits 70/15/15 grouped by `subject_id`, stratified on ancestry crossed with the subject's median age binned at <25/25–34/35–49/50+. No subject appears in more than one split, which matters because subjects contribute several photographs; the 15% out-of-fit slice within Train is itself subject-grouped. UTKFace uses a 70/15/15 split stratified by the race–gender cross-product. The diabetes cohort is ordered by encounter identifier: 42% pretraining, 28% fine-tuning train, 20% validation, and 10% test. Both tabular models use identical rows. Image training partitions reserve 15% of the nominal training split for out-of-fit reporting; the same principle is used by the tabular pipeline. Split seed, run seed, and bootstrap seed are all 12,345 but are represented separately in code.

E.4 EXACT MODEL ROSTER

The image checkpoint identifiers are: `google/vit-base-patch16-224`, `facebook/deit-base-patch16-224`, `crossvit_18_dagger_408.in1k`, `vit_base_patch16_rope_reg1_gap_256.sbb_in1k`, `maxvit_base_tf_384.in1k`, `microsoft/swin-base-patch4-window7-224`, `microsoft/resnet-50`, `facebook/convnext-b`

ase-224, facebook/convnextv2-base-22k-224, google/siglip2-base-patch16-224, facebook/metaclip-2-worldwide-b16, facebook/dinov2-base, facebook/dinov3-vitb16-pretrain-lvd1689m, facebook/ijepa_vith14_1k, facebook/vit-mae-base, microsoft/beit-base-patch16-224-pt22k, and eva02_base_patch14_448.mim_in22k_ft_in22k_in1k. FT-Transformer uses numeric feature tokens, categorical embeddings, and Transformer blocks. Tabular preprocessing is fitted on the pretraining partition and then frozen to prevent leakage.

E.5 Fisher computation

Per-sample gradients are computed with functional automatic differentiation and vectorization, with a per-sample loop fallback for unsupported operators. For multiclass outputs, the implementation visits all classes and weights squared score gradients by the predicted class probabilities. For binary outputs it evaluates the Bernoulli expectation. For regression, σ^2 is estimated from held-out residuals and is never replaced with the observed target inside the score. A single unshuffled pass stores per-example traces; subgroup means and bootstrap intervals are obtained by masking this vector. Thus no subgroup is given a different Fisher pass.

F Additional empirical results

Cross-method dispersion does not track gaps. Within each backbone and axis we form the ratio of PEFT to full-fine-tuning Fisher CV, and likewise for the gap, over the 16 encoders that support every method. These ratios compare different parameterizations and are descriptive (Lemma 1). Ten of 15 method–task CV ratios are below one and 11 of 15 gap ratios are, with Spearman $\rho = -0.30$ across 15 non-independent pairs (Table 6): a lower cross-method CV does not accompany a smaller gap.

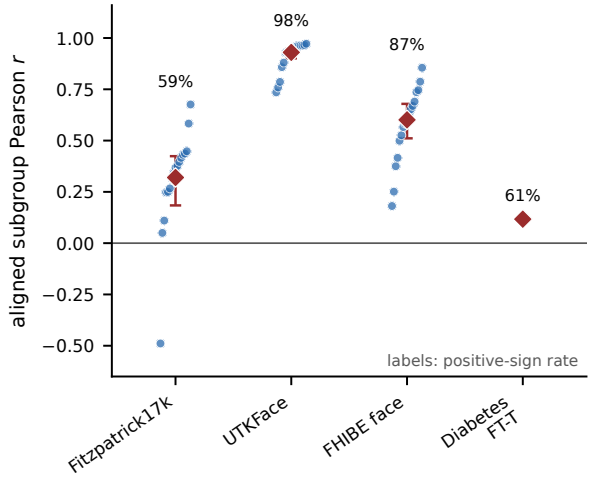


Figure 2: Encoder-level alignment by task. Blue points are encoder-level Fisher-z means, red diamonds are equal-encoder task means, and error bars bootstrap encoders; percentages are descriptive positive-sign rates over cells. Summary values are in Table 3 and method-conditioned means in Table 8.

F.1 Robustness checks

Three design checks delimit the result. First, method-stratified alignment remains high on UTKFace and positive for every method on both dermatology tasks, so the headline is not produced by full fine-tuning alone. Second, replacing $\widehat{t}_a$ with $\widehat{t}_a/\widehat{t}_{\text{pooled}}$ leaves every within-split correlation exactly unchanged; this verifies that normalization cannot manufacture the sign. Third, removing two-level sex and gender axes prevents the algebraic $r = \pm 1$ artifact. The remaining axes contain three to

Table 6: Balanced 16-encoder comparison to full fine-tuning. Fitzpatrick17k entries are balanced-accuracy changes in percentage points (higher is better); UTKFace and FHIBE face are MAE changes in years (lower is better). Ratios are geometric means across the three image tasks; values below one denote a reduction. Both ratios compare different parameterizations and are descriptive.

MMethod	params (%)	ΔFitz BAcc	ΔUTK MAE	ΔFHIBE MAE	Fisher-CV ratio	gap ratio
LoRA	1.57	+2.81	−0.34	−1.33	1.077	0.823
DoRA	1.66	+3.21	−0.31	−1.38	1.095	0.902
rsLoRA	1.57	+1.46	−0.28	−0.79	1.123	0.901
PiSSA	1.55	+1.01	+0.00	−0.88	0.956	0.946
VeRA	0.10	−0.19	−0.04	−1.23	0.603	0.745

fourteen groups. These checks do not address run-to-run training variation; multi-seed replication remains necessary.

F.2 Per-task PEFT ratios

Table 7 expands the geometric means in Table 6. The roster has 16 encoders because PiSSA and VeRA cannot attach to the convolution-only ResNet-50, which is therefore excluded from every cross-method comparison. For method m, task t and encoder e, let $V_{m,t,e}$ be the Fisher CV and $G_{m,t,e} = \max_a d_a - \min_a d_a$ the test-split performance gap. The entries are $\bar{V}_{m,t} = \mathrm{median}_e\ V_{m,t,e}/V_{\mathrm{FT},t,e}$ and $\bar{G}_{m,t} = \mathrm{median}_e\ G_{m,t,e}/G_{\mathrm{FT},t,e}$, where a ratio is defined only when its full-fine-tuning denominator is positive, and Table 6 reports the geometric mean over the three tasks. The entries are medians over the 16 encoders; encoders are weighted equally, and none is dropped. Fitzpatrick17k has two subgroup axes (skin tone and binary tone); for that task we average the two values of $G_{m,t,e}$, and likewise of $V_{m,t,e}$, before forming the ratio. The other tasks have a single axis. Every full-fine-tuning denominator $G_{\mathrm{FT},t,e}$ was strictly positive, so no encoder-level ratio was dropped for a zero denominator. These ratios compare different parameterizations and are descriptive (Lemma 1). Fisher CV is reduced on Fitzpatrick17k and UTKFace but increased on FHIBE face for LoRA, DoRA and rsLoRA. Direct subgroup performance gaps do not follow the same pattern: DoRA increases the Fitzpatrick gap by 19.5%, and PiSSA increases the UTKFace gap by 1.9% despite reducing Fisher dispersion.

Table 7: Per-task ratios relative to full fine-tuning on the balanced 16-encoder roster.

Quantity	task	LoRA	DoRA	rsLoRA	PiSSA	VeRA
Fisher CV	Fitzpatrick17k	0.953	0.845	1.009	0.907	0.913
	UTKFace	0.933	0.927	0.915	0.950	0.537
	FHIBE face	1.405	1.676	1.536	1.016	0.446
Performance gap	Fitzpatrick17k	0.943	1.195	1.031	0.977	1.013
	UTKFace	0.856	0.938	0.915	1.019	0.929
	FHIBE face	0.690	0.656	0.777	0.851	0.440

F.3 Method-conditioned difficulty alignment

Table 8 reports method-conditioned alignment. The tabular setting has only three correlations per entry, one per split, so its values are descriptive and should not be used to rank methods.

Table 8: Fisher-z mean aligned Pearson r by setting and method.

Setting	full FT	LoRA	DoRA	rsLoRA	PiSSA	VeRA
Fitzpatrick17k	0.337	0.217	0.422	0.321	0.381	0.330
UTKFace	0.897	0.929	0.942	0.932	0.934	0.945
FHIBE face	0.400	0.700	0.626	0.587	0.552	0.655
Diabetes FT-T	0.149	0.130	−0.039	0.163	0.470	−0.206

G TRACE-MATCHING EXPERIMENT: DETAILS

Design and analysis.

- *Cells.* Each cell is one task, encoder and adapter, trained once with group-balanced task-loss training (P) and once with the Huber-penalized trace-matching objective (H).
- *Primary analysis.* All 30 cells, each at the batch size selected for it as a hyperparameter (Section C.7). Both arms of a pair always share that batch size, which is what makes the pair a controlled comparison; the arms are never allowed to differ in it.
- *Batch size* Batch size is tuned per cell as a hyperparameter, over values between 64 and 1024; the selected value is 64–512 for 29 of the 30 cells and 1024 for UTKFace ViT–DoRA, and every selected value is reported in Table 10.
- *Sign convention.* Throughout, $R_{\text{gap}} = 100\,(G_{\text{P}} - G_{\text{H}})/G_{\text{P}}$, where G is the test-split best–worst gap; positive favours trace matching.
- *Uncertainty.* The 30 cells share tasks, encoders and adapters, so every interval and test below is descriptive.
  - Cell bootstrap of the median (10,000 replicates, seed 0): $[12,\ 21]\%$.
  - Task-blocked bootstrap (resample tasks, then cells within tasks): $[11,\ 29]\%$.
  - Two-sided sign test $p = 1.86e-09$; Wilcoxon test on $\log(G_{\text{H}}/G_{\text{P}})$ $p = 1.86e-09$.
- *Utility quadrants.*
  - Gap narrows, utility not worse: 23.
  - Gap narrows, utility slightly worse: 7.
  - Gap not narrower, utility not worse: 0.
  - Worst utility changes: -2.4 pp balanced accuracy (Fitzpatrick17k/ViT/PiSSA) and $++0.26$ years MAE (FHIBE face/SigLIP 2/LoRA).
- *Equalized rates (secondary, classification only).* EOD falls in 7 of the 12 classification cells and AOD in 6 of 12. Appendix H explains why the theory makes no prediction for these criteria. `mitigation/primary_analysis/summary.csv`, the same file that produces Section 6 .
- *Mechanism correlation.* The Fisher-CV correlation is reported descriptively, without a p-value: $\rho = +0.23$ between the relative changes in CV and gap, and the CV decreased in 25 of 30 cells. The per-cell values, plain and trace matching, are in Table 14 and in the supplementary file `mechanism_cv.csv`, recomputed from the stored checkpoints by `mitigation/mechanism_cv.py`.
- *Matrix audit.* Table 13 gives the per-cell change in the normalized trace gap, unbiased squared Frobenius discrepancy, operator norm and surrogate-loss gap, from `mitigation/audit/fisher_audit.csv` and `loss_gaps.csv`.

Selection of the penalty form. Before fixing the primary intervention, we conducted a fixed-batch development sweep at batch size 256 across all datasets. We compared the mean-referenced Huber penalty in Equation (15) with an ℓ_1 penalty, a squared ℓ_2 penalty, a one-sided squared-hinge penalty, a variant using the minimum subgroup trace rather than the mean as the reference, and a raw-deviation variant. The best individual alternative was dataset dependent: ℓ_1 produced the smallest best–worst task-performance gap on some datasets, whereas squared ℓ_2 performed best on others. Huber was the only penalty form that consistently reduced the gap relative to the plain arm across all datasets in this sweep. We therefore fixed Huber for the primary paired experiment. This choice provides a robust compromise between the two behaviors: with $\kappa = 1$, the Huber loss is quadratic for $|e_a| \leq 1$, like a squared ℓ_2 penalty near the matching point, and linear for $|e_a| > 1$, like an ℓ_1 penalty for large deviations. The fixed-batch sweep was used to select the penalty form; the subsequently reported primary comparison fixes Huber and selects batch size separately for each dataset–encoder–adapter cell as described in Section C.7.

Batch size is a per-cell hyperparameter rather than a fixed constant, so it is reported in Table 10; both arms of a pair always share it, which is what makes the pair a controlled comparison. Figures 6–9 draw the same 30 cells, one figure per dataset, each panel one encoder $\times$ PEFT cell.

Table 9: Primary analysis by task and adapter, each cell at its selected batch size. Image rows pool the three encoders; each diabetes row is its single FT-Transformer cell, so its counts are out of one. Gap wins count cells with $R_{\text{gap}} > 0$. Plain and trace gaps are means over the cells in the row. "Gap and utility" counts cells that narrow the gap without a loss of pooled utility. Δ utility is the mean change in balanced accuracy (percentage points, higher is better) or MAE (years, lower is better).

Task	PEFT	gap wins	plain gap	trace gap	median R_{gap} (%)	gap and utility	Δ utility
Fitzpatrick17k	LoRA	**3/3**	0.247	0.146	+36	1/3	+0.1
Fitzpatrick17k	DoRA	**3/3**	0.238	0.201	+15	2/3	−0.0
Fitzpatrick17k	PiSSA	**3/3**	0.219	0.153	+32	2/3	+0.5
UTKFace (MAE, yr)	LoRA	**3/3**	1.57	1.39	+14	3/3	−0.26
UTKFace (MAE, yr)	DoRA	**3/3**	1.52	1.28	+15	2/3	−0.22
UTKFace (MAE, yr)	PiSSA	**3/3**	1.87	1.59	+12	3/3	−0.23
FHIBE face (MAE, yr)	LoRA	**3/3**	3.79	3.41	+11	2/3	+0.03
FHIBE face (MAE, yr)	DoRA	**3/3**	4.02	2.48	+33	3/3	−1.24
FHIBE face (MAE, yr)	PiSSA	**3/3**	5.19	4.53	+7	2/3	−0.73
Diabetes FT-T	LoRA	**1/1**	0.043	0.036	+18	1/1	+0.7
Diabetes FT-T	DoRA	**1/1**	0.057	0.037	+35	1/1	+0.6
Diabetes FT-T	PiSSA	**1/1**	0.075	0.058	+23	1/1	+0.9

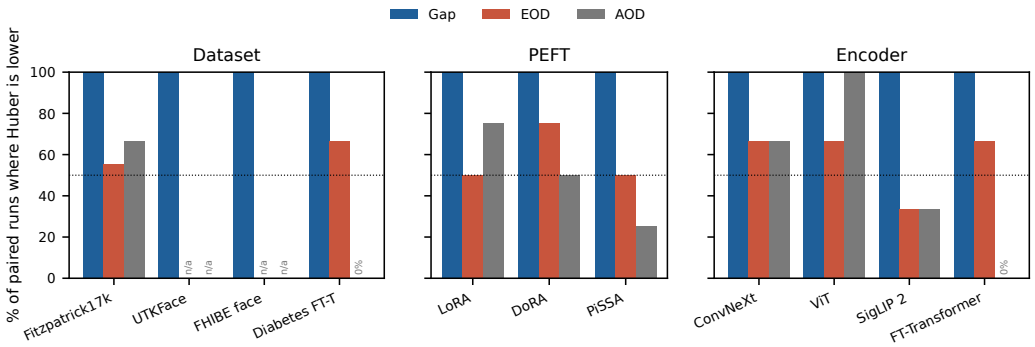


Figure 3: Share of the 30 cells, each at its selected batch size, in which trace matching lowers the gap, EOD or AOD relative to the plain arm, by dataset, adapter and encoder. EOD and AOD are undefined for the regression tasks (n/a); 0% marks a measured value that never favoured trace matching. EOD and AOD count the 12 classification cells only. The dotted line marks 50%.

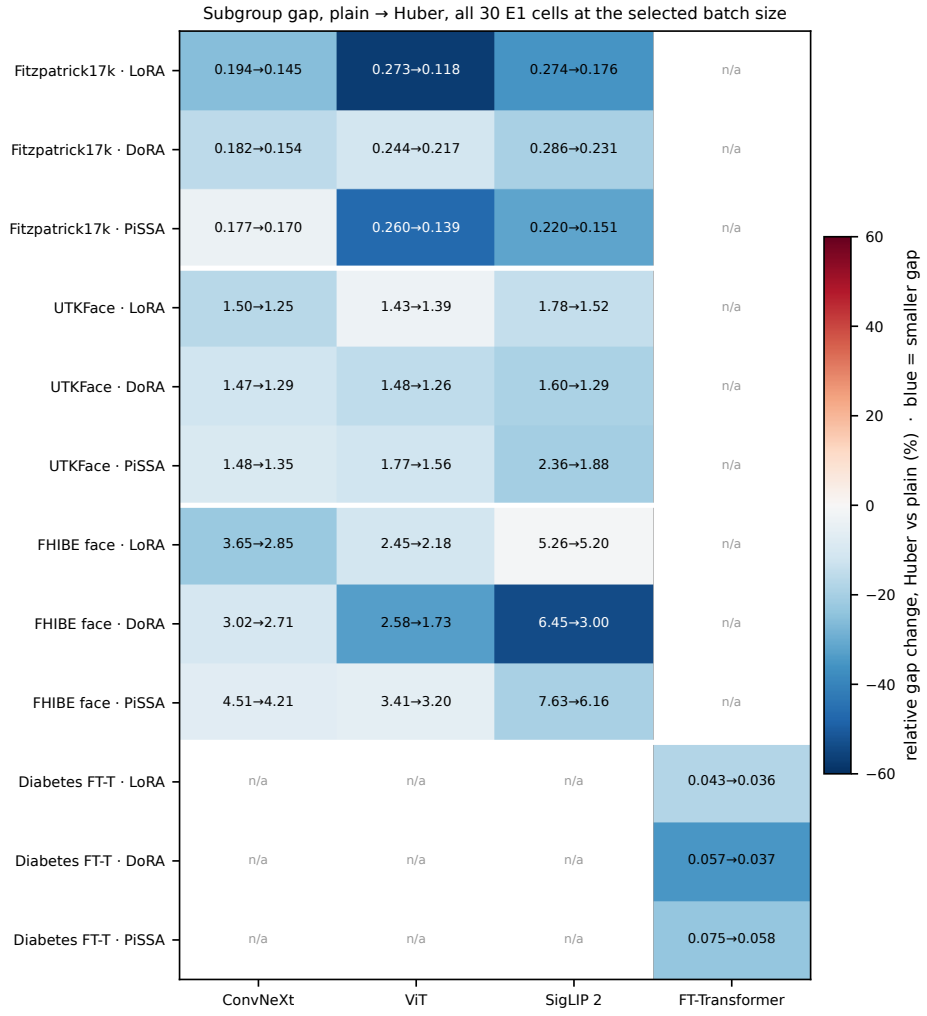


Figure 4: Best–worst gap, plain→trace matching, for all 30 cells, each at its selected batch size. Colour is R_{gap}; blue means trace matching narrows the gap. The image encoders do not apply to diabetes and the FT-Transformer does not apply to the image tasks (n/a).

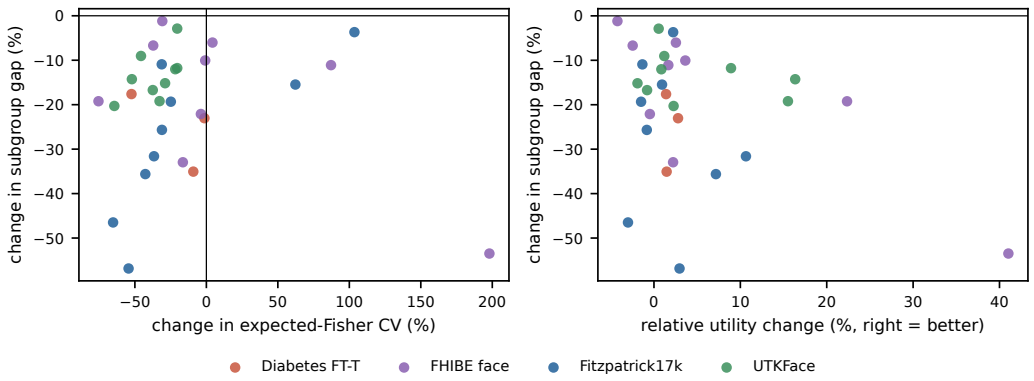


Figure 5: Primary cells, from the experiment pipeline. The axes use a *change* convention, $\Delta G = 100(G_{\mathrm{H}} - G_{\mathrm{P}})/G_{\mathrm{P}} = -R_{\mathrm{gap}}$ and likewise for the CV, so negative values are reductions. Left: ΔCV against ΔG over all 30 cells (Spearman $\rho = +0.23$; no p-value is reported because the cells share tasks, encoders and adapters). Right: relative change in pooled utility (right is better) against ΔG. The per-cell CV values are in Table 14.

Table 10: Primary Eval 1 results for all 30 cells. P and H denote plain and Huber-curvature training. BS is the selected batch size shared by the two arms of each pair. Bracketed values are percentage reductions, so positive values favor Huber. Balanced accuracy is reported for classification and MAE for regression; regression EOD is undefined (—).

Dataset	Encoder	PEFT	BS	Worst P→H	Pooled P→H	Gap P→H [drop]	EOD P→H [drop]
Fitzpatrick17k	ConvNeXt	DoRA	256	$0.603 \to 0.594$	$0.708 \to 0.716$	$0.182 \to 0.154$ [15.5%]	$0.182 \to 0.096$ [47.4%]
Fitzpatrick17k	ConvNeXt	LoRA	256	$0.598 \to 0.617$	$0.733 \to 0.725$	$0.194 \to 0.145$ [25.7%]	$0.153 \to 0.116$ [24.1%]
Fitzpatrick17k	ConvNeXt	PiSSA	256	$0.557 \to 0.613$	$0.716 \to 0.721$	$0.177 \to 0.170$ [3.7%]	$0.064 \to 0.145$ [−127.7%]
Fitzpatrick17k	ViT	DoRA	128	$0.566 \to 0.570$	$0.747 \to 0.737$	$0.244 \to 0.217$ [10.9%]	$0.108 \to 0.072$ [34.0%]
Fitzpatrick17k	ViT	LoRA	128	$0.496 \to 0.644$	$0.738 \to 0.727$	$0.273 \to 0.118$ [56.9%]	$0.063 \to 0.072$ [−13.8%]
Fitzpatrick17k	ViT	PiSSA	128	$0.534 \to 0.598$	$0.740 \to 0.715$	$0.260 \to 0.139$ [46.5%]	$0.145 \to 0.059$ [59.6%]
Fitzpatrick17k	SigLIP 2	DoRA	256	$0.479 \to 0.543$	$0.739 \to 0.741$	$0.286 \to 0.231$ [19.3%]	$0.050 \to 0.078$ [−56.8%]
Fitzpatrick17k	SigLIP 2	LoRA	256	$0.524 \to 0.660$	$0.739 \to 0.762$	$0.274 \to 0.176$ [35.6%]	$0.153 \to 0.138$ [9.6%]
Fitzpatrick17k	SigLIP 2	PiSSA	512	$0.448 \to 0.580$	$0.639 \to 0.673$	$0.220 \to 0.151$ [31.6%]	$0.070 \to 0.160$ [−127.2%]
UTKFace	ConvNeXt	DoRA	64	$4.46 \to 4.37$	$4.06 \to 4.01$	$1.47 \to 1.29$ [12.0%]	—
UTKFace	ConvNeXt	LoRA	128	$4.66 \to 4.56$	$4.20 \to 4.17$	$1.50 \to 1.25$ [16.7%]	—
UTKFace	ConvNeXt	PiSSA	64	$4.46 \to 4.40$	$4.06 \to 3.98$	$1.48 \to 1.35$ [9.0%]	—
UTKFace	ViT	DoRA	1024	$4.63 \to 4.65$	$4.06 \to 4.15$	$1.48 \to 1.26$ [15.2%]	—
UTKFace	ViT	LoRA	64	$4.36 \to 4.34$	$3.86 \to 3.84$	$1.43 \to 1.39$ [2.9%]	—
UTKFace	ViT	PiSSA	256	$5.09 \to 4.65$	$4.55 \to 4.13$	$1.77 \to 1.56$ [11.8%]	—
UTKFace	SigLIP 2	DoRA	128	$4.94 \to 4.19$	$4.47 \to 3.77$	$1.60 \to 1.29$ [19.2%]	—
UTKFace	SigLIP 2	LoRA	128	$4.97 \to 4.29$	$4.49 \to 3.77$	$1.78 \to 1.52$ [14.3%]	—
UTKFace	SigLIP 2	PiSSA	128	$5.68 \to 5.32$	$4.97 \to 4.78$	$2.36 \to 1.88$ [20.3%]	—
FHIBE Face	ConvNeXt	DoRA	128	$7.54 \to 7.32$	$5.16 \to 4.94$	$3.02 \to 2.71$ [10.1%]	—
FHIBE Face	ConvNeXt	LoRA	128	$7.90 \to 7.50$	$5.15 \to 5.06$	$3.65 \to 2.85$ [22.1%]	—
FHIBE Face	ConvNeXt	PiSSA	256	$8.98 \to 8.88$	$5.47 \to 5.55$	$4.51 \to 4.21$ [6.7%]	—
FHIBE Face	ViT	DoRA	256	$6.95 \to 6.27$	$4.84 \to 4.83$	$2.58 \to 1.73$ [32.9%]	—
FHIBE Face	ViT	LoRA	128	$6.87 \to 6.68$	$4.86 \to 4.78$	$2.45 \to 2.18$ [11.1%]	—
FHIBE Face	ViT	PiSSA	128	$7.64 \to 7.77$	$5.36 \to 5.05$	$3.41 \to 3.20$ [6.0%]	—
FHIBE Face	SigLIP 2	DoRA	64	$13.96 \to 7.74$	$8.94 \to 5.46$	$6.45 \to 3.00$ [53.5%]	—
FHIBE Face	SigLIP 2	LoRA	256	$10.03 \to 10.00$	$5.82 \to 6.08$	$5.26 \to 5.20$ [1.2%]	—
FHIBE Face	SigLIP 2	PiSSA	128	$14.74 \to 11.56$	$8.78 \to 6.81$	$7.63 \to 6.16$ [19.2%]	—
Diabetes FT-T	FT-T	DoRA	128	$0.591 \to 0.615$	$0.633 \to 0.639$	$0.057 \to 0.037$ [35.1%]	$0.151 \to 0.135$ [10.3%]
Diabetes FT-T	FT-T	LoRA	256	$0.610 \to 0.621$	$0.633 \to 0.640$	$0.043 \to 0.036$ [17.6%]	$0.127 \to 0.159$ [−25.1%]
Diabetes FT-T	FT-T	PiSSA	128	$0.565 \to 0.597$	$0.629 \to 0.638$	$0.075 \to 0.058$ [23.0%]	$0.198 \to 0.161$ [18.8%]

G.1 Cost of adapter versus full-model penalty scope

The penalty differentiates each per-example loss with respect to the adapter and head only (*adapter* scope). The alternative differentiates with respect to every parameter, including the frozen backbone (*full* scope).

- *Setup.* LoRA, DoRA and PiSSA on SigLIP 2 (375M parameters) and ViT (87M parameters), on FHIBE face and UTKFace, with the Huber penalty and identical hyperparameters, so scope is the only difference within a row.
- *Time.* Full scope is slower in all 12 pairs, by 41–62%. The increase is smallest for DoRA (41–43%) and largest for PiSSA (59–62%), and it is stable across the two datasets, which differ in size by roughly a factor of two.
- *Memory.* Full scope raises peak allocation from 6.4–6.9 GB to 13.5–14.2 GB for LoRA and PiSSA, roughly doubling it, and from 12.7–13.0 GB to 20.6–21.1 GB for DoRA. The penalty must retain activations for the frozen backbone under full scope, so the increase is in the backbone's activation memory rather than in parameter state; it is therefore independent of dataset size.
- *Utilization.* Full scope also runs at higher GPU utilization (94.5–97.7% against 89.9–95.6%), so the extra time is additional work rather than a stall.
- *Scope of this comparison.* These runs measure cost only. They do not test whether full scope changes subgroup gaps or task quality, so the choice of adapter scope in the main experiments rests on cost together with the identifiability argument of Lemma 1, not on a measured quality comparison.

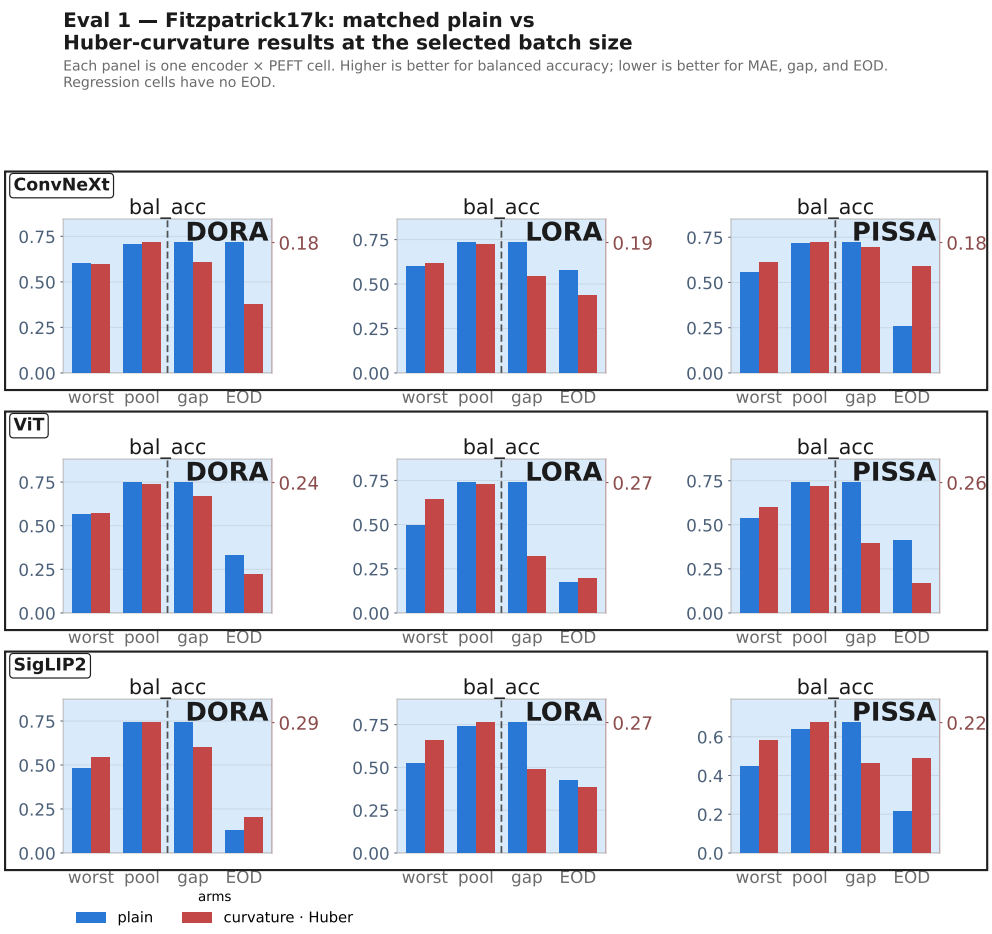


Figure 6: Fitzpatrick17k primary cells: matched plain versus Huber-curvature results at the selected batch size. Rows are the three encoders, columns the three PEFT methods. Higher is better for balanced accuracy; lower is better for gap and EOD.

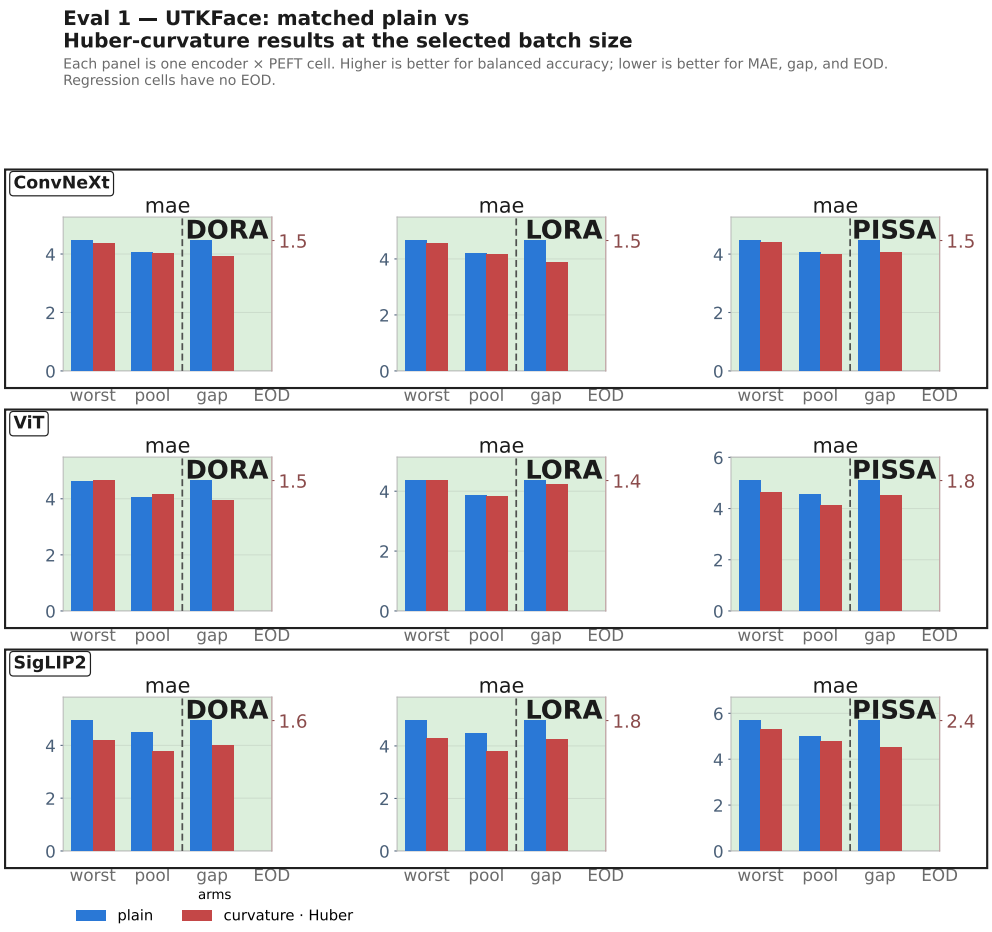


Figure 7: UTKFace primary cells: matched plain versus Huber-curvature results at the selected batch size. Rows are the three encoders, columns the three PEFT methods. Higher is better for balanced accuracy; lower is better for MAE and gap. UTKFace is a regression task, so EOD is undefined.

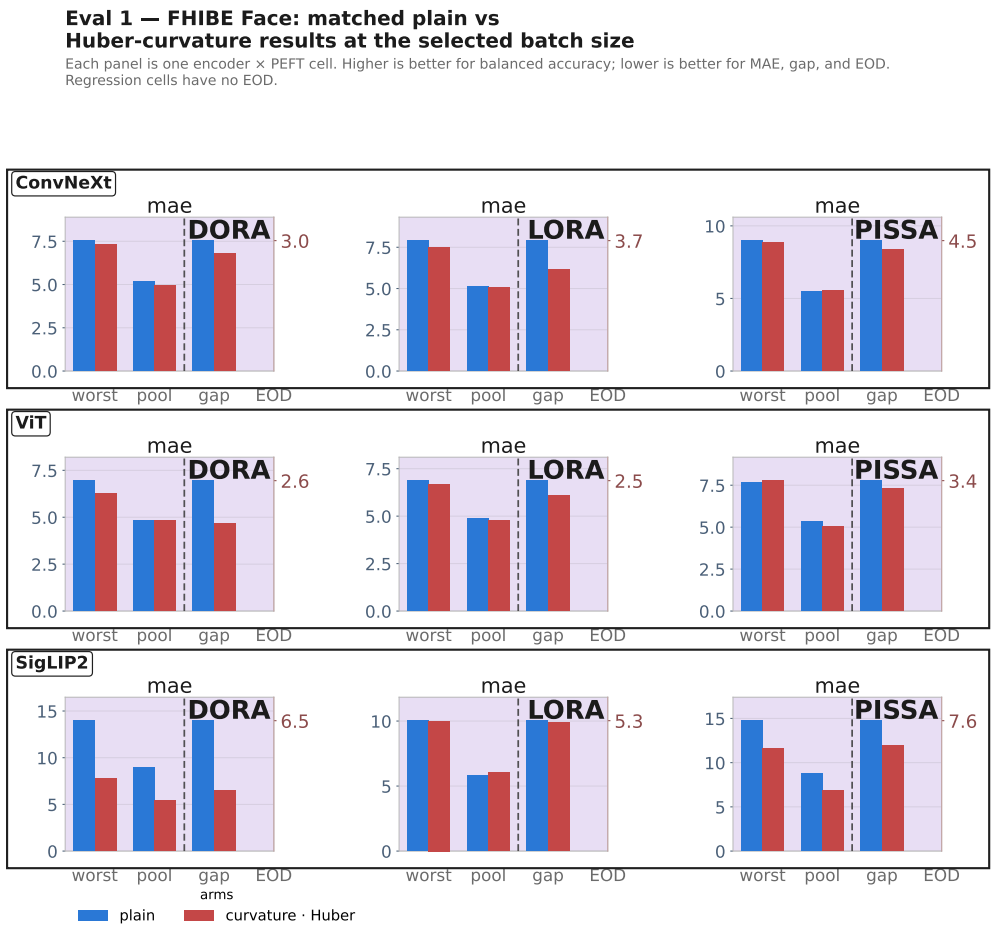


Figure 8: FHIBE Face primary cells: matched plain versus Huber-curvature results at the selected batch size. Rows are the three encoders, columns the three PEFT methods. Higher is better for balanced accuracy; lower is better for MAE and gap. FHIBE Face is a regression task, so EOD is undefined.

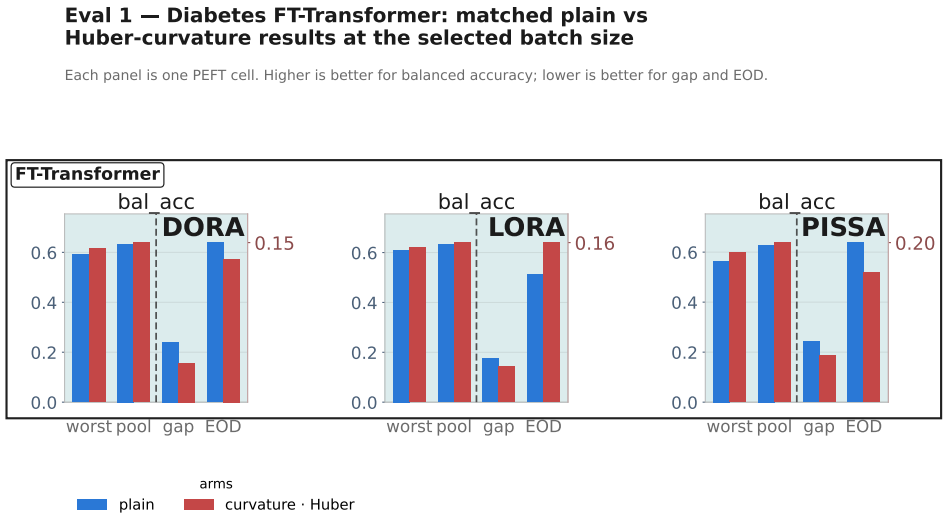


Figure 9: Diabetes FT-Transformer primary cells: matched plain versus Huber-curvature results at the selected batch size, one panel per PEFT method. Higher is better for balanced accuracy; lower is better for gap and EOD.

Table 11: Measured cost of adapter versus full penalty scope on a primary evaluation device NVIDIA RTX 5090. Each entry is the mean ± standard deviation over repeated timed runs. Train time is seconds per run; Δt is the increase from adapter to full scope. Peak memory is the maximum allocated during the run, and is independent of the dataset because the batch composition is fixed, so the two datasets repeat the same value. GPU utilization is the mean over the run. This table reports cost only; it makes no quality comparison between the two scopes.

			train time (s)			peak mem (GB)		GPU util (%)
Encoder	Dataset	PEFT	adapter	full	Δt	adapter	full	adapter / full
ViT	FHIBE face	LoRA	63.30 ± 0.21	101.30 ± 0.34	+60%	6.42	13.55	92.2 / 95.1
		DoRA	87.41 ± 0.32	123.74 ± 0.58	+42%	12.69	20.55	90.7 / 94.8
		PiSSA	62.55 ± 0.44	100.67 ± 0.32	+61%	6.39	13.53	92.2 / 95.0
	UTKFace	LoRA	155.35 ± 0.49	248.97 ± 0.37	+60%	6.42	13.55	95.6 / 97.3
		DoRA	212.89 ± 0.52	304.80 ± 0.57	+43%	12.69	20.55	92.8 / 96.2
		PiSSA	152.04 ± 0.50	246.57 ± 0.32	+62%	6.39	13.53	95.4 / 97.7
SigLIP 2	FHIBE face	LoRA	68.78 ± 0.49	108.37 ± 0.66	+58%	6.88	14.16	92.6 / 95.0
		DoRA	93.30 ± 0.60	131.24 ± 0.49	+41%	12.97	21.13	89.9 / 94.5
		PiSSA	67.41 ± 0.43	107.33 ± 0.50	+59%	6.84	14.14	92.7 / 95.2
	UTKFace	LoRA	168.15 ± 0.64	266.36 ± 0.28	+58%	6.88	14.16	95.4 / 97.6
		DoRA	227.19 ± 0.26	323.24 ± 0.33	+42%	12.97	21.13	93.3 / 96.4
		PiSSA	164.25 ± 0.44	263.99 ± 0.27	+61%	6.84	14.14	95.1 / 97.6

Table 12: Measured cost of adapter versus full penalty scope, on a second evaluation device (NVIDIA GB10). Each entry is the mean ± standard deviation over the three training epochs. Train time is seconds per epoch; Δt is the increase from adapter to full scope. Peak memory is the maximum allocated during the run, and is independent of the dataset because the batch composition is fixed, so the two datasets repeat the same value. GPU utilization is the mean over the run. This table reports cost only; it makes no quality comparison between the two scopes.

			train time (s)			peak mem (GB)		GPU util (%)
Encoder	Dataset	PEFT	adapter	full	Δt	adapter	full	adapter / full
ViT	FHIBE face	LoRA	377.83 ± 0.30	594.81 ± 0.68	$+57\%$	6.42	13.55	93.9 / 94.3
		DoRA	542.34 ± 0.87	757.80 ± 0.76	$+40\%$	12.69	20.55	94.1 / 94.3
		PiSSA	344.20 ± 0.72	590.89 ± 0.75	$+72\%$	6.39	13.53	93.6 / 94.5
	UTKFace	LoRA	940.16 ± 0.64	1477.05 ± 0.49	$+57\%$	6.42	13.55	94.6 / 94.9
		DoRA	1350.09 ± 0.69	1883.82 ± 1.25	$+40\%$	12.69	20.55	94.5 / 94.6
		PiSSA	853.01 ± 0.52	1462.13 ± 0.70	$+71\%$	6.39	13.53	94.3 / 94.8
SigLIP 2	FHIBE face	LoRA	421.83 ± 0.46	657.00 ± 0.88	$+56\%$	6.88	14.16	93.8 / 94.6
		DoRA	589.28 ± 0.72	818.11 ± 0.93	$+39\%$	12.97	21.13	94.0 / 94.4
		PiSSA	387.44 ± 0.77	652.53 ± 0.93	$+68\%$	6.84	14.14	93.7 / 94.4
	UTKFace	LoRA	1048.41 ± 0.57	1656.39 ± 0.29	$+58\%$	6.88	14.16	94.5 / 94.9
		DoRA	1456.91 ± 0.37	2038.55 ± 0.60	$+40\%$	12.97	21.13	94.5 / 94.7
		PiSSA	961.78 ± 0.64	1630.53 ± 0.69	$+70\%$	6.84	14.14	94.3 / 94.8

Table 13: Post-training audit, per cell. Each entry is the change from the plain arm to trace matching, in percent, so negative favours trace matching; bold marks a reduction. Trace gap, Frobenius discrepancy and operator norm are normalised by s_{ab} and averaged over subgroup pairs and held-out batches; the loss gap is the best–worst subgroup gap in the training loss.

Task	Encoder	Adapter	trace gap	Frobenius	operator norm	loss gap
Fitzpatrick17k	ConvNeXt	DoRA	**-25**	+98	+1	**-49**
Fitzpatrick17k	ConvNeXt	LoRA	**-31**	**-84**	**-4**	**-40**
Fitzpatrick17k	ConvNeXt	PiSSA	+20	**-41**	+10	+19
Fitzpatrick17k	SigLIP 2	DoRA	+4	**-18**	**-3**	**-33**
Fitzpatrick17k	SigLIP 2	LoRA	+20	+66	**-4**	**-16**
Fitzpatrick17k	SigLIP 2	PiSSA	**-17**	**-25**	+19	**-32**
Fitzpatrick17k	ViT	DoRA	**-27**	+36	**-14**	**-51**
Fitzpatrick17k	ViT	LoRA	**-9**	+102	**-18**	**-33**
Fitzpatrick17k	ViT	PiSSA	+3	+75	**-2**	**-51**
UTKFace	ConvNeXt	DoRA	**-9**	**-25**	**-8**	**-4**
UTKFace	ConvNeXt	LoRA	**-12**	**-48**	**-6**	**-22**
UTKFace	ConvNeXt	PiSSA	**-24**	**-41**	**-14**	**-28**
UTKFace	SigLIP 2	DoRA	**-20**	+204	**-12**	**-53**
UTKFace	SigLIP 2	LoRA	**-9**	+103	**-7**	**-37**
UTKFace	SigLIP 2	PiSSA	**-20**	**-4**	**-23**	**-24**
UTKFace	ViT	DoRA	+3	**-20**	+2	+8
UTKFace	ViT	LoRA	**-20**	**-8**	**-19**	**-3**
UTKFace	ViT	PiSSA	+3	+26	+7	**-20**
FHIBE face	ConvNeXt	DoRA	**-13**	+17	**-27**	**-13**
FHIBE face	ConvNeXt	LoRA	**-14**	+7	**-14**	**-31**
FHIBE face	ConvNeXt	PiSSA	+0	**-14**	**-21**	**-8**
FHIBE face	SigLIP 2	DoRA	+42	**-38**	**-2**	**-63**
FHIBE face	SigLIP 2	LoRA	+16	+26	+23	+8
FHIBE face	SigLIP 2	PiSSA	**-12**	**-25**	**-55**	**-41**
FHIBE face	ViT	DoRA	**-21**	**-10**	**-20**	**-30**
FHIBE face	ViT	LoRA	+17	+19	+11	**-16**
FHIBE face	ViT	PiSSA	**-8**	+3	**-23**	**-15**
Diabetes FT-T	FT-T	DoRA	**-38**	**-5**	**-26**	**-18**
Diabetes FT-T	FT-T	LoRA	**-46**	**-5**	**-28**	**-20**
Diabetes FT-T	FT-T	PiSSA	**-33**	+40	**-31**	**-16**

Table 14: Expected-Fisher trace coefficient of variation across the constrained subgroups, per cell, recomputed from the stored checkpoints on the test split by `mitigation/mechanism_cv.py`. ΔCV is the change from the plain arm to trace matching in percent, so negative means trace matching made the subgroup traces more even; bold marks a reduction. These are the per-cell values behind Figure 5.

Task	Encoder	Adapter	CV plain	CV matched	ΔCV
Fitzpatrick17k	ConvNeXt	DoRA	0.095	0.154	+62
Fitzpatrick17k	ConvNeXt	LoRA	0.206	0.142	**-31**
Fitzpatrick17k	ConvNeXt	PiSSA	0.161	0.327	+104
Fitzpatrick17k	SigLIP 2	DoRA	0.164	0.123	**-25**
Fitzpatrick17k	SigLIP 2	LoRA	0.165	0.095	**-43**
Fitzpatrick17k	SigLIP 2	PiSSA	0.139	0.088	**-37**
Fitzpatrick17k	ViT	DoRA	0.315	0.217	**-31**
Fitzpatrick17k	ViT	LoRA	0.414	0.189	**-54**

Task	Encoder	Adapter	CV plain	CV matched	ΔCV
Fitzpatrick17k	ViT	PiSSA	0.358	0.125	**-65**
UTKFace	ConvNeXt	DoRA	0.157	0.123	**-22**
UTKFace	ConvNeXt	LoRA	0.213	0.133	**-37**
UTKFace	ConvNeXt	PiSSA	0.197	0.107	**-46**
UTKFace	SigLIP 2	DoRA	0.232	0.156	**-33**
UTKFace	SigLIP 2	LoRA	0.189	0.091	**-52**
UTKFace	SigLIP 2	PiSSA	0.352	0.125	**-64**
UTKFace	ViT	DoRA	0.138	0.098	**-29**
UTKFace	ViT	LoRA	0.110	0.088	**-20**
UTKFace	ViT	PiSSA	0.168	0.134	**-20**
FHIBE face	ConvNeXt	DoRA	0.224	0.222	**-1**
FHIBE face	ConvNeXt	LoRA	0.258	0.248	**-4**
FHIBE face	ConvNeXt	PiSSA	0.247	0.155	**-37**
FHIBE face	SigLIP 2	DoRA	0.115	0.342	+198
FHIBE face	SigLIP 2	LoRA	0.278	0.192	**-31**
FHIBE face	SigLIP 2	PiSSA	0.740	0.181	**-75**
FHIBE face	ViT	DoRA	0.208	0.174	**-16**
FHIBE face	ViT	LoRA	0.082	0.153	+87
FHIBE face	ViT	PiSSA	0.155	0.161	+4
Diabetes FT-T	FT-T	DoRA	0.107	0.097	**-9**
Diabetes FT-T	FT-T	LoRA	0.107	0.051	**-52**
Diabetes FT-T	FT-T	PiSSA	0.102	0.101	**-1**

H EQUALIZED-RATE CRITERIA

Our claims concern the per-group performance gap. Rate-parity criteria condition on the label and are a different object. We define them here because they appear as a secondary outcome in Section 6, and we give a construction showing why the curvature analysis should not be expected to bear on them.

Definitions. Equal-opportunity difference (EOD) is the between-group difference in true-positive rates; average-odds difference (AOD) averages the true-positive and false-positive rate differences (Hardt et al., 2016). Both condition on the label, so they apply only to the classification settings, Fitzpatrick17k and Diabetes 130H. We do not threshold the regression targets, and no rate metric is computed for UTKFace or FHIBE face. Two conventions make the classification case concrete. Fitzpatrick17k's three-way target is binarized one-vs-rest, each class in turn positive, and the three values are macro-averaged; diabetes readmission is already binary. Every axis has more than two groups (skin tone seven, binary tone three, diabetes race six), so we compute per-group rates and report the largest pairwise difference $\max_{g,g'}|r_g - r_{g'}|$, which is nonnegative and smaller is better.

Why no association should be expected. Per-group task difficulty and between-group rate parity are related but non-equivalent, and the non-equivalence is exact.

Theorem 8 (Loss geometry does not determine EOD)**.** *There exist two subgroup distributions and a one-parameter model $q_\phi(x_1) = \sigma(\phi)$, $q_\phi(x_2) = 1 - \sigma(\phi)$ such that for* every ϕ*,*

$$L_a(\phi) = L_b(\phi), \quad g_a(\phi) = g_b(\phi), \quad H_a(\phi) = H_b(\phi), \quad F_a^{\mathrm{R}}(\phi) = F_b^{\mathrm{R}}(\phi), \quad \eta_{ab}(\phi) = 0,$$

so the entire loss geometry is identical and the gap vanishes identically, yet at the Bayes-optimal ϕ the rates satisfy $\mathrm{EOD} = 0.4$ *and* $\mathrm{AOD} \approx 0.244$*.*

Proof. Binary labels; two input atoms x_1, x_2 with $\eta(x_1) = \Pr(Y{=}1 \mid x_1) = 0.9$ and $\eta(x_2) = 0.1$. Group a places mass $(\frac{1}{2}, \frac{1}{2})$ on (x_1, x_2); group b places mass $(\frac{1}{10}, \frac{9}{10})$. Take the one-parameter model $q_\phi(x_1) = \sigma(\phi)$, $q_\phi(x_2) = 1 - \sigma(\phi)$, where $q_\phi(x)$ is the predicted probability of $Y = 1$ and σ is the logistic function.

The two atoms contribute the identical function of ϕ. Writing $s = \sigma(\phi)$, the expected negative log-likelihood at x_1 is $-[0.9 \log s + 0.1 \log(1 - s)]$, and at x_2, where the model predicts $1 - s$ and

the truth is $\eta = 0.1$, it is $-[0.1\log(1-s) + 0.9\log s]$, which is the same expression. Hence for *any* mixture weights the group loss equals that common function, so $L_a(\phi) = L_b(\phi)$ for all ϕ, and consequently $g_a = g_b$, $H_a = H_b$ and $\Delta_{ab} \equiv 0$ with all derivatives zero.

The Fisher also coincides. At x_1 the score is $\partial_\phi \log s = 1 - s$ for $y = 1$ and $\partial_\phi \log(1-s) = -s$ for $y = 0$, giving Fisher contribution $s(1-s)^2 + (1-s)s^2 = s(1-s)$. At x_2 the predicted probability is $1-s$, the scores are $-s$ and $1-s$ respectively, and the contribution is $(1-s)s^2 + s(1-s)^2 = s(1-s)$. Both atoms contribute $s(1-s)$, so $F_a = F_b = s(1-s)$ for any mixture; with $p = 1$ and $J = 1$, $F_a^{\mathrm{R}} = F_b^{\mathrm{R}}$ and $\eta_{ab} = |H_a - F_a^{\mathrm{R}} - (H_b - F_b^{\mathrm{R}})| = 0$.

Rates differ at the Bayes point. The model is well specified: at $\sigma(\phi) = 0.9$ it reproduces η exactly, so this is the global minimizer of both group losses. Thresholding at $\frac{1}{2}$ predicts $\hat{y} = 1$ exactly on x_1. Group a: $\Pr(Y{=}1) = \frac{1}{2}(0.9) + \frac{1}{2}(0.1) = 0.5$ and $\Pr(\hat{y}{=}1, Y{=}1) = \frac{1}{2}(0.9) = 0.45$, so $\mathrm{TPR}_a = 0.9$; $\Pr(Y{=}0) = 0.5$ and $\Pr(\hat{y}{=}1, Y{=}0) = 0.05$, so $\mathrm{FPR}_a = 0.1$. Group b: $\Pr(Y{=}1) = \frac{1}{10}(0.9) + \frac{9}{10}(0.1) = 0.18$ and $\Pr(\hat{y}{=}1, Y{=}1) = 0.09$, so $\mathrm{TPR}_b = 0.5$; $\Pr(Y{=}0) = 0.82$ and $\Pr(\hat{y}{=}1, Y{=}0) = 0.01$, so $\mathrm{FPR}_b = 0.01/0.82 \approx 0.0122$. Therefore $\mathrm{EOD} = |0.9 - 0.5| = 0.4$ and $\mathrm{AOD} = \frac{1}{2}(0.4 + |0.1 - 0.0122|) \approx 0.244$. □

Unlike the results of Section 3, Theorem 8 needs no slope, defect or smallness hypothesis: the loss geometry is identical in every order and for every ϕ, and the rate disparity is nevertheless 0.4. The mechanism is that TPR conditions on $Y = 1$ and so depends on the group base rate, and in this construction changing the mixture weights changes the conditional-rate denominators but not the loss geometry, because both input atoms contribute the identical loss and Fisher for every ϕ. In particular, the equal-mixture pair (both groups with mass $(\frac{1}{2}, \frac{1}{2})$) has exactly the same loss geometry as the pair above but $\mathrm{EOD} = 0$: the geometry does not determine EOD.

I Interpretive boundaries and failure modes

Each boundary below is now a consequence of a stated result rather than a caution, and we give the governing result in each case.

Coordinate dependence. Raw gradient norms depend on the chosen trainable coordinates, so changing the adapter family, rank, or parameter scaling can change the trace without a corresponding change in predictive behavior. Lemma 1 makes this exact: under $\phi = S(\psi)$ with Jacobian K, $F_a^{\mathrm{R}} \mapsto K^\top F_a^{\mathrm{R}} K$, so taking $K = cI$ rescales every trace by c^2 while leaving the predictor untouched. The across-group Fisher CV is unaffected by that isotropic case, since all group traces move together, but it is not invariant in general: with $F_a^{\mathrm{R}} = \mathrm{diag}(1, 0)$, $F_b^{\mathrm{R}} = \mathrm{diag}(0, 1)$ and $K = \mathrm{diag}(1, 2)$ the CV moves from 0 to 0.6. We therefore compare subgroup traces only within a fixed model and parameterization, and note that the inertia of $F_a^{\mathrm{R}} - F_b^{\mathrm{R}}$ (its numbers of positive, negative and zero eigenvalues) and the generalized eigenvalues of the pencil $(F_a^{\mathrm{R}}, F_b^{\mathrm{R}})$ are the coordinate-free alternatives.

The trace is a summary, not the governing quantity. Theorem 2 shows the second-order term is governed by $\|C_{ab}\|_2^{\mathcal{H}}$, which agrees with $\|F_a^{\mathrm{R}} - F_b^{\mathrm{R}}\|_2^{\mathcal{H}}$ only up to the defect $\eta_{ab}^{\mathcal{H}}$; and Theorem 4 shows that matching traces leaves it as large as the common trace: among positive semidefinite matrices of equal trace τ, the spectral discrepancy can be exactly τ. Lemma 2 adds that the empirical trace mixes $\|\hat{g}_a\|_2^2$ with $\mathrm{tr}\,\widehat{\Sigma}_a$, so a trace penalty can succeed by trading one term against the other; equalizing $\|\hat{g}_a\|^2$ does not imply $g_a = g_b$. Any conclusion drawn from trace or CV agreement alone is therefore weaker than it appears; Section C.4 gives a matrix discrepancy that does bound the curvature channel's magnitude, though, like any scalar norm, not its sign.

Expected Fisher is not observed-label empirical Fisher. The training penalty uses $\nabla\ell_i \nabla\ell_i^\top$ at the observed label. The post-training diagnostic in Equation 5 instead averages score gradients over the model's predictive distribution. These matrices can behave differently (Kunstner et al., 2019); we report them separately and do not use the expected-Fisher interpretation to describe the quantity optimized during training. Note that Assumption **A2** of Section 3 identifies the Gauss–Newton matrix with the *expected* Fisher only, which is why the defect Ξ_a is carried explicitly rather than discarded.

Difficulty is not fairness. Equation 6 tests whether subgroup ordering agrees with one task metric; EOD and AOD condition on the target and compare error rates, and Theorem 8 shows that the loss-geometric quantities studied here do not determine them. Calibration, demographic parity, worst-group risk and intersectional harms impose further, separate requirements that no scalar local sensitivity can certify.

A matched geometry is not a small gap. As excess surrogate risks vanish the performance gap tends to $\gamma_{ab} = |R_a^\star - R_b^\star|$ (Theorem 5), so exact geometric matching does not drive it to zero. Note that γ_{ab} is *not* a lower bound: smaller gaps are attainable by degrading the better-served group (Remark 8). A reader should therefore not treat a reduced discrepancy as evidence of improved outcomes, nor an unchanged gap as evidence that the geometry did not move; the two are only loosely coupled.

The curvature results are restricted to horizontal additive updates. By Proposition 4 and Corollary 3, $\ker J(\phi)$ contains the tangent space of the $\mathrm{GL}(r)$ orbit $BA = (Be^{tX})(e^{-tX}A)$ and, for $r \le \min(d_{\mathrm{in}}, d_{\mathrm{out}})$, has dimension at least r^2 everywhere, so $F_a^{\mathrm{R}} - F_b^{\mathrm{R}}$ always has a null direction and is never positive definite on raw (A, B) coordinates. Every definiteness hypothesis in Section 3 is therefore stated for $\delta \in \mathcal{H}_\phi$, and $\mathcal{S}_\varepsilon = \{0\}$ means "no nonzero horizontal additive update". This is a first-order restriction and nothing more: Remark 2 gives a LoRA point at the standard $B = 0$ initialization where two updates differing by a kernel element produce second-order gap changes of opposite sign. Nothing in Section C projects onto $\mathcal{H}_\phi$. At a fixed point the reachable Fishers annihilate the gauge directions, but the trace and Frobenius statistics are not constant along a gauge orbit (Appendix C.4), so the penalty can also move the factors along function-preserving directions; this is not the same as verifying the hypotheses, and the optimizer's raw steps are not covered by the theorems at all.

The curvature results assume slope matching. Theorems 1 and 7(ii), the forced-increase reading of Theorem 3(ii), and the gap consequence of trace equality (Remark 6) all require $g_a = g_b$. Without it a first-order term dominates and the conclusions fail outright (Remark 4). Nothing in Section C enforces slope matching exactly, so the theory describes a channel the objective addresses, not a property the trained models are shown to have.

J EXACT VERIFICATION ON A SMALL ADAPTED NETWORK

The deep-model experiments cannot form Hessians, so they cannot check the theorems' quantities directly. Here we do so on a network small enough that every matrix is formed exactly, to show that the results are operational rather than only formal. The script `validate_exact.py` regenerates every number.

Setup. Inputs $x \in \mathbb{R}^4$ pass through $\tanh(Wx)$ with a rank-2 LoRA weight $W = W_0 + BA$ ($W_0 \in \mathbb{R}^{6\times 4}$ frozen), then a frozen linear head with three classes and softmax cross-entropy; the trainable parameters are $\phi = (A, B)$, $p = 20$. Two subgroups of 400 inputs each have shifted input means, and labels come from a fixed teacher, so the model is misspecified. The evaluation point is reached by 300 steps of full-batch gradient descent on the pooled loss from a full-rank initialization. All computations are in float64 with exact Hessians, model-expected reachable Fishers and adapter Jacobians. The data are synthetic and generated by the script.

Reading. Table 15 shows the local theory working as stated on an actual nonlinear adapted network. The second-order expansion has a cubic remainder, Theorem 2's worst-case term is attained, and once the slope term is removed the sign of the curvature term predicts the sign of the gap change. On a subspace where C_{ab} is positive definite, every sampled step increases the gap, as Theorem 1 requires (the theorem is stated for slope matching on all of $\mathcal{H}_\phi$; here the slope term vanishes on the subspace itself, and the same proof applies). Two results are cautionary and consistent with the paper's qualifications. First, the reachable-Fisher certificate of Corollary 1 fails on the very subspace where C_{ab} is definite, and not marginally: there $\lambda_{\min}(F_a^{\mathrm{R}} - F_b^{\mathrm{R}}) = -0.100$ is negative, so no bound on the defect could restore the certificate, and the definiteness of C_{ab} on S_+ is carried by the defect itself. On $\mathcal{H}_\phi$ the defect is of comparable size to the Fisher difference ($\eta_{ab}^{\mathcal{H}}$ is 0.66 of it). The defect

Table 15: Exact checks on the small network (`validate_exact.py`). S is the part of $\mathcal{H}_\phi$ orthogonal to $g_a - g_b$, on which the slope term vanishes for every update (the one-step constraint), and S_+ the span of the positive eigenvectors of C_{ab} restricted to S.

Check	Result
Gauge (Prop. 4)	$\dim\ker J = 4 = r^2$; $\|F_c^{\mathrm{R}} v\| \le 10^{-15}$ for $v \in \ker J$; the unrestricted $\lambda_{\min}(F_a^{\mathrm{R}} - F_b^{\mathrm{R}}) = -4.45 < 0$.
Expansion (Prop. 1)	over 8 random horizontal directions, the remainder after the slope and curvature terms scales as $\varepsilon^{3.05}$ (median log–log slope).
Worst case (Thm. 2)	along the dominant restricted eigenvector (eigenvalue -3.13, so $\|C_{ab}\|_2^{\mathcal{H}} = 3.13$), the second-order change matches $\mathcal{Q}_{ab}(\varepsilon) = \frac{1}{2}\|C_{ab}\|_2^{\mathcal{H}}\varepsilon^2$ in magnitude to within 34.6%, 5.8%, 2.2%, 1.2% and 0.6% at $\varepsilon = 0.2, 0.1, 0.05, 0.025, 0.0125$.
Sign on S	C_{ab} restricted to S is indefinite (eigenvalues in $[-2.01, 1.27]$); the sign of the actual gap change matched $\mathrm{sign}(\delta^\top C_{ab}\delta)$ in 216 of 216 random directions at $\varepsilon = 0.01$.
Forced increase on S_+ (Thm. 1)	$\dim S_+ = 6$, $\rho_{\min} = 0.0105$; all 1,200 sampled steps increased the gap, by at least $2.79 \times \frac{1}{4}\rho_{\min}\|\delta\|^2$.
Certificate (Cor. 1)	on S_+, $\lambda_{\min}(F_a^{\mathrm{R}} - F_b^{\mathrm{R}}) = -0.100$ against a defect of 0.171: the reachable-Fisher certificate fails although C_{ab} is positive definite there; on $\mathcal{H}_\phi$, $\eta_{ab}^{\mathcal{H}}/\|F_a^{\mathrm{R}} - F_b^{\mathrm{R}}\|_2^{\mathcal{H}} = 0.66$.
Defect control	with logits linear in the trainable parameters, $\max\|H_c - F_c^{\mathrm{R}}| = 2 \times 10^{-16}$.
Audit identities	Gram-identity Frobenius, trace, and the operator norm from $K^{1/2}DK^{1/2}$ match the dense matrices to relative error $\le 3 \times 10^{-15}$.
Balanced representative	a random $\mathrm{GL}(2)$ gauge change leaves predictions unchanged (2×10^{-15}) but moves the trace gap from -9.38 to -8.16 and the Frobenius discrepancy from 11.42 to 12.85; after balancing, both copies give -10.50 and 12.34.

term in the guarantees is not a formality. Second, trace and Frobenius statistics change along a function-preserving gauge orbit and are made invariant by the balanced representative used in the audit (Appendix C.4). This is one small network and one evaluation point; it shows the statements are checkable and behave as proved, not how large the defect is in the deep models.