DIFFERENTIALLY PRIVATE WORST-CASE RATIO ANALYSIS VIA ENTROPIC SMOOTHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Discriminative dimensionality reduction seeks a subspace in which the classes remain separated, and the pairs of classes that lie closest together are the hardest to keep apart. In this paper we investigate pairwise worst-case ratio analysis (PWCRA), a criterion that protects those pairs by maximizing, over all class pairs, the smallest ratio of the between-class scatter of a pair to its own within-class scatter. The resulting problem is a non-convex, non-smooth max–min fractional program. A second difficulty is that the learned projection is a function of class means and covariances, so releasing it can reveal information about individual samples. We address both with entropic smoothing. Reversing the sign of the criterion turns it into the minimization of a pointwise maximum, and replacing that maximum by a log-sum-exp function gives a smooth surrogate whose gap to the exact maximum is additive, independent of the ambient and the target dimension, and vanishing as the temperature tends to zero. Built on majorization–minimization principles, the resulting single-loop algorithm, ESRA, assembles at every iteration a gradient with closed-form softmax weights and then takes one polar-factor update, with no inner solver. We then make the same pipeline differentially private by releasing the per-class sufficient statistics once through a Gaussian mechanism, so that the whole optimization is post-processing and the privacy budget does not grow with the number of iterations. We prove privacy, monotone descent, stationary convergence and an end-to-end utility bound, together with a stability bound for the softmax weights that sets the temperature from the noise level of the mechanism, so the tighter the budget, the warmer the smoothing at which we stop. The private variant follows the privacy budget closely, and in terms of the nearest-neighbour, nearest-mean and quadratic discriminant error rates, supported by paired significance tests, ESRA outperforms the state-of-the-art and benchmark methods.

1 INTRODUCTION

Modern acquisition technologies produce datasets whose feature dimension is very large, which raises the cost of learning, increases the risk of overfitting, and makes the fitted model hard to interpret, so dimensionality reduction has become a standard preprocessing step in supervised classification [11, 19, 24, 32–34, 45]. Linear discriminant analysis (LDA), proposed for two classes and later extended to several, remains the most widely used method of this kind [18, 36]. Its most specific weakness appears when the target dimension is smaller than the number of classes minus one, because the LDA criterion effectively averages the separations of all class pairs and is therefore dominated by the pair that is already far apart, so that the pairs which are close to each other can be projected on top of one another [20]. This is the *worst-case class separation* problem.

A more direct treatment is to optimize the worst pair explicitly. The earliest members of that family enlarge the smallest distance between class means, an idea later carried to the heteroscedastic setting [7, 41, 46, 47]. Such criteria are built on distances between classes, so two classes whose means are pushed apart can still overlap when their internal dispersion is large, and worst-case *ratio* analysis addresses this shortcoming by maximizing the smallest pairwise ratio of between-class to within-class scatter [45]. Since each pair is judged by its own within-class scatter, the criterion stays meaningful when class variances differ substantially. We refer to it throughout as pairwise worst-case ratio

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.

analysis (PWCRA), and it is the most adaptive of these treatments as well as the hardest to optimize, since it yields a non-convex, non-smooth max–min *fractional* program on the Stiefel manifold.

In [45], the proposed algorithm, worst-case discriminative dimension reduction (WDDR), recasts the non-convex fractional program as a sequence of quadratic subproblems and applies to each a semidefinite relaxation (in the sense of losing tightness), which replaces the $d \times m$ projection by a $d \times d$ matrix variable, d and m being the ambient and the target dimension; the projection is then recovered from the m leading eigenvectors of the semidefinite solution. This design introduces two drawbacks. First, the lifted variable enlarges the problem, so that each semidefinite solve costs $\mathcal{O}((d+m)^{4.5})$ operations, and the recovered projection is not guaranteed to be optimal, which may undermine class separability and the effective handling of class variances. Second, solving the semidefinite program requires manually tuning at least two hyperparameters, the stopping tolerance aside [3, 35, 45]. Appendix J reviews this literature and the compared methods.

A second difficulty has not been addressed at all for criteria of this kind. PWCRA touches the data only through class means and covariance-type statistics, and a projection fitted to such statistics can reveal information about individual samples. Differential privacy (DP) is the standard formal answer [14, 15], and private counterparts exist for principal component analysis [13, 16], empirical risk minimization [5, 12], gradient-based learning [1] and statistical-query analyses [8]. None of them carries over to a worst-case separation criterion, whose objective commits at every iteration to one single worst pair through a hard $\arg\max$, so that noise of any magnitude can flip that selection.

In this paper we first propose **ESRA**, a single-loop algorithm for PWCRA that avoids both drawbacks described above. It needs no semidefinite program and no inner solver, so its cost per iteration is far lower, and its single parameter, the temperature, is fixed in advance from a target accuracy, with no tuning. We reverse the sign of the objective and replace the resulting pointwise maximum by the log-sum-exp function at a temperature $\tau > 0$ [6, 31], whose approximation error is additive, at most τ times the logarithm of the number of class pairs and independent of both the ambient and the target dimension, so that the original problem is recovered as τ tends to zero. The fractional structure is absorbed inside the same loop by a quadratic transform [39], which we compose with a descent-lemma majorizer, and since the Frobenius norm is constant on the Stiefel manifold, every iteration reduces to one closed-form polar-factor update.

As a second contribution we propose **DP-ESRA**, which makes the same pipeline differentially private. All pairwise scatters are deterministic functions of a small collection of per-class sufficient statistics, so it suffices to privatize these statistics *once* with a Gaussian mechanism. Every later computation then uses the privatized statistics alone, and by the post-processing property of differential privacy it adds nothing to the privacy cost, however many iterations, restarts and annealing stages the optimization runs, whereas in gradient perturbation of the DP-SGD kind the privacy cost grows with every step [1, 5]. The two contributions reinforce each other, because entropic smoothing is at the same time what makes the problem easy to optimize and what renders it safe to privatize. The smoothed objective is uniformly Lipschitz in the scatter matrices, so the spectral error of the privatized statistics transfers into an end-to-end utility bound, and the softmax weights react continuously to a perturbation where the hard $\arg\max$ does not.

The main contributions of this work are given as summary:

(i) **ESRA, a single-loop solver for PWCRA.** Every iteration is a softmax-weighted gradient assembly followed by one closed-form polar-factor update, with no inner solver and no semidefinite program, so the cost per iteration stays low and nothing has to be tuned beyond the target accuracy. We prove monotone descent and stationary convergence.

(ii) **An accuracy account of the smoothing.** We show that the smoothed objective exceeds the exact one by an additive term that depends only on the temperature and the number of class pairs and vanishes with the temperature, and we turn this into a rule that fixes the temperature in advance from a target accuracy.

(iii) **DP-ESRA, private at a cost independent of the iteration count.** We bound the joint sensitivity of the per-class sufficient statistics under record replacement and release them once, so the whole optimization is post-processing and the budget does not grow with the number of iterations.

(iv) **End-to-end utility and a noise-calibrated temperature.** We bound how far the private solution's worst-case ratio can fall below the non-private optimum, and we bound how far the softmax

weights can move under a perturbation of the statistics. Balancing the two gives a temperature that is set from the noise scale of the mechanism alone, with no access to the data.

(v) **A complete empirical study** on the standard benchmarks of this literature against state-of-the-art and benchmark methods, under several classification rules, with paired significance tests, runtimes, projection plots, a privacy–utility analysis, ablations, and a comparison with a per-step gradient-perturbation solver for the same objective.

Notation. Bold uppercase and lowercase letters denote matrices and column vectors, $\mathbf{I}_m$ is the identity matrix of order m, $(\cdot)^T$ and $\mathrm{tr}(\cdot)$ are the transpose and the trace, and $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{tr}(\mathbf{A}^T\mathbf{B})$. The spectral norm of a matrix and the Euclidean norm of a vector are both written $\|\cdot\|_2$, and $\|\cdot\|_F$ and $\|\cdot\|_*$ are the Frobenius and the nuclear norm. For a symmetric $\mathbf{A}$, $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ are its extreme eigenvalues, $\mathbf{A} \succcurlyeq \mathbf{0}$ and $\mathbf{A} \succ \mathbf{0}$ denote positive semidefiniteness and definiteness, and $\Pi_{\succcurlyeq 0}(\mathbf{A})$ clips its negative eigenvalues to zero. The Stiefel manifold is $\mathcal{S} = \{\mathbf{T} \in \mathbb{R}^{d\times m} : \mathbf{T}^T\mathbf{T} = \mathbf{I}_m\}$, $[\mathrm{softmax}(\mathbf{y})]_k = e^{y_k} / \sum_l e^{y_l}$, and $f \circ g$ is the composition of two maps. Throughout, C is the number of classes, $K = C(C-1)/2$ the number of class pairs, d and m the ambient and the target dimension, n the number of samples, τ the smoothing temperature and (ε, δ) the privacy budget.

2 Problem formulation

2.1 Pairwise worst-case ratio analysis

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ contain n samples from C classes, with π_k, n_k and $\bar{\mathbf{x}}_k = \frac{1}{n_k}\sum_{\ell\in\pi_k} \mathbf{x}_\ell$ the index set, size and mean of class k. We look for a linear map $\mathbf{T} \in \mathbb{R}^{d\times m}$ with $m \ll d$ into a subspace in which the classes remain separated. For every pair (i, j) with $1 \le i < j \le C$, the between-class and within-class scatter matrices of that pair are

$$\mathbf{S}_b^{ij} = \boldsymbol{\delta}_{ij}\boldsymbol{\delta}_{ij}^T, \quad \boldsymbol{\delta}_{ij} \triangleq \bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j, \quad \mathbf{S}_w^{ij} = \boldsymbol{\Sigma}_i^w + \boldsymbol{\Sigma}_j^w, \quad \boldsymbol{\Sigma}_k^w \triangleq \sum_{\ell\in\pi_k} (\mathbf{x}_\ell - \bar{\mathbf{x}}_k)(\mathbf{x}_\ell - \bar{\mathbf{x}}_k)^T. \tag{1}$$

Here $\mathbf{S}_b^{ij}$ is an outer product and so has rank one, while $\mathbf{S}_w^{ij}$ sums two within-class scatters of rank at most $n_i - 1$ and $n_j - 1$ and is therefore singular whenever the two classes together carry fewer samples than the ambient dimension. Positive definiteness is restored in the experiments by the preprocessing of Section 5, which reduces the dimension below that sample count [33, 35], and in the private pipeline of Section 4 by a ridge $\lambda_{\mathrm{reg}}\mathbf{I}$ with a constant $\lambda_{\mathrm{reg}} > 0$ fixed in advance, which it needs in any case because a noisy second-moment matrix need not be positive semidefinite. Accordingly, $\mathbf{S}_w^{ij} \succ \mathbf{0}$ is assumed for every pair in what follows. With these definitions, the PWCRA problem seeks, over the Stiefel manifold, the projection that maximizes the smallest ratio of between-class to within-class scatter among all class pairs,

$$\max_{\mathbf{T}\in\mathcal{S}} \min_{1\le i<j\le C} r_{ij}(\mathbf{T}), \qquad r_{ij}(\mathbf{T}) \triangleq \frac{\mathrm{tr}\big(\mathbf{T}^T\mathbf{S}_b^{ij}\mathbf{T}\big)}{\mathrm{tr}\big(\mathbf{T}^T\mathbf{S}_w^{ij}\mathbf{T}\big)}. \tag{2}$$

Formulation (2) has three properties that merit discussion. First, the inner minimization renders the criterion a worst-case one, in the sense that the quality of a projection is determined by the least separated class pair, which is precisely the pair that an averaged criterion such as LDA tends to sacrifice. Second, the pair index appears in the denominator as well as in the numerator, so each pair is normalized by its own within-class scatter, which is why (2) remains meaningful when the class covariances differ substantially [45]. Third, the constraint $\mathbf{T}^T\mathbf{T} = \mathbf{I}_m$ keeps the basis orthonormal, which improves both the discriminative behaviour of the projection and its numerical stability in high dimension [18, 41, 45]. The problem is thus difficult on three counts at once, being non-convex, non-smooth and fractional, and Section 3 shows that entropic smoothing reduces it to a single loop whose every step is closed-form.

3 ESRA, A SINGLE-LOOP ALGORITHM FOR PWCRA

Since the minimum of a finite set of numbers is the negative of the maximum of their negatives, problem (2) is equivalent to

$$-\min_{\mathbf{T}\in\mathcal{S}} \max_{1\le i<j\le C} \big\{ -r_{ij}(\mathbf{T})\big\}, \qquad r_{ij}(\mathbf{T}) \triangleq \frac{\operatorname{tr}\big(\mathbf{T}^T\mathbf{S}_b^{ij}\mathbf{T}\big)}{\operatorname{tr}\big(\mathbf{T}^T\mathbf{S}_w^{ij}\mathbf{T}\big)}, \tag{3}$$

and we solve the minimization in (3), so that every surrogate built below is an upper bound on the quantity it replaces. The obstacle in (3) is the pointwise maximum, which is not differentiable wherever two pairs attain it at once, and at a worst-case solution several pairs typically do. We remove that obstacle by replacing the maximum with the log-sum-exp function at a temperature $\tau > 0$, which gives

$$\min_{\mathbf{T}\in\mathcal{S}} F_\tau(\mathbf{T}), \qquad F_\tau(\mathbf{T}) \;\triangleq\; \tau \log \sum_{1\le i<j\le C} \exp\left(-\frac{1}{\tau}\, \frac{\operatorname{tr}\big(\mathbf{T}^T\mathbf{S}_b^{ij}\mathbf{T}\big)}{\operatorname{tr}\big(\mathbf{T}^T\mathbf{S}_w^{ij}\mathbf{T}\big)} \right). \tag{4}$$

The accuracy of this replacement is controlled uniformly over the manifold, as the following lemma shows [6, 31].

Lemma 3.1 (Uniform additive approximation)**.** *For every $\mathbf{T} \in \mathcal{S}$ and every $\tau > 0$,*

$$\max_{1\le i<j\le C} \big\{ -r_{ij}(\mathbf{T})\big\} \;\le\; F_\tau(\mathbf{T}) \;\le\; \max_{1\le i<j\le C} \big\{ -r_{ij}(\mathbf{T})\big\} + \tau \log K, \tag{5}$$

with the same constant at every $\mathbf{T}$ (for the proof, see Appendix D).

The gap between the two bounds of Lemma 3.1 involves neither d, nor m, nor the scale of the data, and it vanishes as $\tau \to 0$, so that the original problem is recovered in that limit and the gap transfers directly to the quality of the solution.

Proposition 3.2 (Solution quality)**.** *Let* $\mathrm{OPT} = \max_{\mathbf{T}\in\mathcal{S}} \min_{ij} r_{ij}(\mathbf{T})$ *and let* $\widehat{\mathbf{T}}$ *be a global minimizer of* (4)*. Then* $\min_{ij} r_{ij}(\widehat{\mathbf{T}}) \ge \mathrm{OPT} - \tau \log K$*, so that* $\tau \le \rho/\log K$ *guarantees* ρ*-optimality for any target accuracy* $\rho > 0$ *fixed by the analyst (for the proof, see Appendix D).*

It remains to minimize (4) over $\mathbf{T}$, and we do so by building on majorization–minimization principles [21, 23, 43]. At the current iterate $\mathbf{T}^t$ we form an upper bound on the objective that touches it at $\mathbf{T}^t$, and we minimize that bound instead. For a criterion of the form (3) the whole construction is contained in one lemma, which takes a sign-reversed ratio directly to a function that is *affine* in $\mathbf{T}$ (for an overview of the majorization–minimization principle, see Appendix A).

Lemma 3.3 (Affine majorization of a sign-reversed ratio on $\mathcal{S}$)**.** *Let $\mathbf{A} \succcurlyeq \mathbf{0}$ and $\mathbf{B} \succ \mathbf{0}$, let $\mathbf{T}^t \in \mathcal{S}$, and set*

$$u_t = \operatorname{tr}\big((\mathbf{T}^t)^T\mathbf{A}\mathbf{T}^t\big) > 0, \quad v_t = \operatorname{tr}\big((\mathbf{T}^t)^T\mathbf{B}\mathbf{T}^t\big) > 0, \quad d_t = \frac{u_t}{v_t^2}, \quad \mathbf{G}_t = 2d_t\mathbf{B}\mathbf{T}^t - \frac{2}{v_t}\mathbf{A}\mathbf{T}^t. \tag{6}$$

Then for every $L \ge 2d_t\lambda_{\max}(\mathbf{B})$ and every $\mathbf{T} \in \mathcal{S}$,

$$-\frac{\operatorname{tr}\big(\mathbf{T}^T\mathbf{A}\mathbf{T}\big)}{\operatorname{tr}\big(\mathbf{T}^T\mathbf{B}\mathbf{T}\big)} \;\le\; c_t - \operatorname{tr}\big(\mathbf{B}_t^T\mathbf{T}\big), \qquad \mathbf{B}_t \triangleq L\,\mathbf{T}^t - \mathbf{G}_t, \tag{7}$$

where c_t does not depend on $\mathbf{T}$, with equality at $\mathbf{T} = \mathbf{T}^t$ (for the proof, see Appendix E).

We now apply Lemma 3.3 to our criterion, pair by pair, with $\mathbf{A} = \mathbf{S}_b^{ij}$ and $\mathbf{B} = \mathbf{S}_w^{ij}$, both of which carry the pair index. Writing

$$u_{ij}^t = \big\|(\mathbf{T}^t)^T\boldsymbol{\delta}_{ij}\big\|_2^2, \qquad v_{ij}^t = \operatorname{tr}\big((\mathbf{T}^t)^T\mathbf{S}_w^{ij}\mathbf{T}^t\big), \qquad d_{ij}^t = \frac{u_{ij}^t}{(v_{ij}^t)^2}, \tag{8}$$

the intermediate convex quadratic of the lemma is, for each pair,

$$-r_{ij}(\mathbf{T}) \;\le\; \hat{h}_{ij}(\mathbf{T} \mid \mathbf{T}^t) \;\triangleq\; d_{ij}^t\, \operatorname{tr}\big(\mathbf{T}^T\mathbf{S}_w^{ij}\mathbf{T}\big) - \frac{2}{v_{ij}^t}\, \operatorname{tr}\big((\mathbf{T}^t)^T\mathbf{S}_b^{ij}\mathbf{T}\big), \tag{9}$$

tight at $\mathbf{T}^t$. Because log-sum-exp is non-decreasing in each argument, substituting the K bounds (9) into (4) majorizes the smoothed objective itself,

$$F_\tau(\mathbf{T}) \;\le\; \widehat{F}_\tau(\mathbf{T} \mid \mathbf{T}^t) \;\triangleq\; \tau \log \sum_{1\le i<j\le C} \exp\big(\hat{h}_{ij}(\mathbf{T} \mid \mathbf{T}^t)/\tau\big) \qquad \text{on } \mathcal{S}, \tag{10}$$

with equality and matching gradients at $\mathbf{T}^t$. Applying the second and third steps of Lemma 3.3 to the composite (10), whose curvature is bounded by an explicit constant $\widehat{L}_\tau^{\,t}$, yields the affine majorizer the algorithm actually minimizes (for the constant and the derivation of (11), see Appendix E),

$$F_\tau(\mathbf{T}) \;\le\; c_t - \operatorname{tr}\big(\mathbf{B}_t^T\mathbf{T}\big), \qquad \mathbf{B}_t \triangleq \widehat{L}_\tau^{\,t}\,\mathbf{T}^t - \nabla F_\tau(\mathbf{T}^t). \tag{11}$$

Minimizing the right-hand side of (11) over $\mathcal{S}$ is an orthogonal Procrustes problem, and it has a closed-form answer.

Theorem 3.4 (Closed-form update)**.** *Let $\mathbf{B}_t = \mathbf{U}\mathbf{D}\mathbf{V}^T$ be a thin singular value decomposition. Then $\mathbf{T}^{t+1} = \mathbf{U}\mathbf{V}^T$ minimizes $-\operatorname{tr}(\mathbf{B}_t^T\mathbf{T})$ over $\mathcal{S}$, with optimal value $-\|\mathbf{B}_t\|_*$. When $\mathbf{B}_t$ has full column rank the minimizer is unique and equals the polar factor $\mathbf{B}_t(\mathbf{B}_t^T\mathbf{B}_t)^{-1/2}$ (for the proof, see Appendix E).*

The only quantity still to be supplied is the gradient in (11). Differentiating (10) at $\mathbf{T}^t$ by the chain rule through the log-sum-exp, and using the first-order contact with F_τ at that point, gives it in closed form,

$$\nabla F_\tau(\mathbf{T}^t) = \sum_{1\le i<j\le C} w_{ij}^t \Big[\, 2\, d_{ij}^t\, \mathbf{S}_w^{ij}\mathbf{T}^t - \tfrac{2}{v_{ij}^t}\,\mathbf{S}_b^{ij}\mathbf{T}^t \Big], \qquad w_{ij}^t = \big[\operatorname{softmax}\big(-\mathbf{r}^t/\tau\big)\big]_{ij}, \tag{12}$$

where $\mathbf{r}^t$ collects the K ratios at $\mathbf{T}^t$. The weights w_{ij}^t form a probability distribution over the class pairs, concentrating on the worst-separated pair as τ decreases and hedging over the near-worst pairs at a finite τ, which is what Section 4 shows makes the scheme robust once the statistics are privatized. It is worth mentioning that the log-sum-exp function is the entropy-regularized maximum over the simplex and these weights are its maximizers, so ESRA solves the inner maximization of the max–min problem in closed form, with no quadratic program at any iteration [6, 31].

One iteration therefore evaluates the scalars (8), forms the weights and the gradient (12), builds $\mathbf{B}_t$ by (11), and updates by Theorem 3.4, every step being closed-form. We call the resulting method **ESRA** (Entropic-Smoothing Ratio Analysis). A summary of ESRA is given in Algorithm S1 of Appendix C, which also states the annealing schedule of the temperature.

The per-iteration cost of (12) and the convergence of the resulting sequence are established in Appendix F, which shows that one iteration costs $\mathcal{O}(Cd^2m + Kdm + m^3)$ with no quadratic program, no semidefinite program and no outer loop on the ratio value, and that the objective decreases monotonically, every limit point of the iterates being stationary.

4 DP-ESRA, PRIVACY PAID ONCE AT THE LEVEL OF THE STATISTICS

4.1 THE MECHANISM

We now make the pipeline differentially private, so that the released projection reveals almost nothing about any single training record; for familiarity with the concept of differential privacy, the Gaussian mechanism and post-processing invariance, see Appendix B. Privacy requires a bound on how far one record can move the quantity we release, so for the private pipeline, and only for it, we make the standard boundedness assumption.

Assumption 1. *Every sample satisfies $\|\mathbf{x}_\ell\|_2 \le R$, enforced by the data-independent clipping $\mathbf{x} \mapsto \mathbf{x}\min\{1, R/\|\mathbf{x}\|_2\}$, and every class contains at least $n_{\min}$ samples, with C, R and $n_{\min}$ public, that is, known without access to the data. This is the standard bounded-norm assumption of the differential privacy literature, where per-example clipping to a public radius is the usual means of bounding sensitivity [1, 14].*

ESRA touches the data only through the pairwise scatters (1), and those are deterministic functions of $3C$ per-class sufficient statistics,

$$q(\mathbf{X}) = \big(n_k,\ \mathbf{s}_k,\ \mathbf{M}_k\big)_{k=1}^{C}, \qquad \mathbf{s}_k = \sum_{\ell\in\pi_k}\mathbf{x}_\ell, \qquad \mathbf{M}_k = \sum_{\ell\in\pi_k}\mathbf{x}_\ell\mathbf{x}_\ell^T, \tag{13}$$

with $\bar{\mathbf{x}}_k = \mathbf{s}_k/n_k$ and $\mathbf{\Sigma}_k^w = \mathbf{M}_k - \mathbf{s}_k\mathbf{s}_k^T/n_k$. DP-ESRA therefore privatizes $q(\mathbf{X})$ once, and every later computation is post-processing of that single release. The noise level of the release is governed by the sensitivity of q, which the following lemma bounds.

Lemma 4.1 (Joint ℓ_2-sensitivity)**.** *Under record replacement and Assumption 1, and with the matrix blocks measured in Frobenius norm, the query* (13) *satisfies*

$$\Delta_2 \;\le\; \max\Big\{\sqrt{2\,(1+R^2+R^4)},\;\; 2R\sqrt{1+R^2}\Big\}, \tag{14}$$

the first term corresponding to a record that moves from one class to another and the second to a replacement within a class (for the proof, see Appendix G).

Given a budget (ε, δ) we release $\hat{n}_k = n_k + \nu_k$, $\hat{\mathbf{s}}_k = \mathbf{s}_k + \boldsymbol{\xi}_k$ and $\widehat{\mathbf{M}}_k = \mathbf{M}_k + \mathbf{E}_k$ for $k = 1, \dots, C$, with $\nu_k \sim \mathcal{N}(0, \sigma^2)$, $\boldsymbol{\xi}_k \sim \mathcal{N}(0, \sigma^2\mathbf{I}_d)$, $\mathbf{E}_k$ a symmetric Gaussian matrix as in [16], and σ calibrated by (S12) with Δ_2 from Lemma 4.1. The scatters are then rebuilt by post-processing only,

$$\tilde{n}_k = \max\{\hat{n}_k, 1\}, \quad \hat{\bar{\mathbf{x}}}_k = \frac{\hat{\mathbf{s}}_k}{\tilde{n}_k}, \quad \widehat{\mathbf{\Sigma}}_k^w = \Pi_{\succcurlyeq 0}\big(\widehat{\mathbf{M}}_k - \tilde{n}_k \hat{\bar{\mathbf{x}}}_k \hat{\bar{\mathbf{x}}}_k^T\big), \quad \widehat{\mathbf{S}}_w^{ij} = \widehat{\mathbf{\Sigma}}_i^w + \widehat{\mathbf{\Sigma}}_j^w + \lambda_{\mathrm{reg}}\mathbf{I}_d, \tag{15}$$

with $\widehat{\mathbf{S}}_b^{ij} = \hat{\boldsymbol{\delta}}_{ij}\hat{\boldsymbol{\delta}}_{ij}^T$, $\hat{\boldsymbol{\delta}}_{ij} = \hat{\bar{\mathbf{x}}}_i - \hat{\bar{\mathbf{x}}}_j$. The clipping $\Pi_{\succcurlyeq 0}$ is needed because a noisy second-moment matrix need not be positive semidefinite, and the public ridge $\lambda_{\mathrm{reg}} > 0$ keeps the denominators away from zero; ESRA then runs unchanged on these inputs, and we call the pipeline **DP-ESRA**.

Theorem 4.2 (Privacy)**.** *Under Assumption 1, DP-ESRA is (ε, δ)-differentially private, independently of the number of iterations, of the number of annealing stages and restarts, and of the values of τ, m, λ_{reg} and all stopping tolerances (for the proof, see Appendix G).*

4.2 Utility and the noise-calibrated temperature

The utility analysis has three steps, the first of which is that the private statistics concentrate.

Lemma 4.3 (Statistical accuracy)**.** *Fix $\gamma \in (0,1)$ and suppose that $n_{\min} \gtrsim \sigma\sqrt{\log(C/\gamma)}$. With probability at least $1-\gamma$, every class k and every pair (i,j) satisfy $\|\widehat{\mathbf{\Sigma}}_k^w - \mathbf{\Sigma}_k^w\|_2 \le e_w$ and $\|\widehat{\mathbf{S}}_b^{ij} - \mathbf{S}_b^{ij}\|_2 \le e_b$, where, up to absolute constants, $e_w \lesssim \sigma(\sqrt{d} + \sqrt{\log(C/\gamma)})(1 + R + R^2/n_{\min})$ and $e_b \lesssim \frac{\sigma R}{n_{\min}}(\sqrt{d} + \sqrt{\log(C/\gamma)})$ (for the proof, see Appendix H).*

The second step turns these errors into a *uniform* perturbation of the ratios over the whole manifold.

Lemma 4.4 (Uniform ratio perturbation)**.** *Suppose that $\|\widehat{\mathbf{S}}_b^{ij} - \mathbf{S}_b^{ij}\|_2 \le e_b$ and $\|\widehat{\mathbf{S}}_w^{ij} - \mathbf{S}_w^{ij}\|_2 \le e'_w$ for every pair, with $\widehat{\mathbf{S}}_w^{ij} \succcurlyeq \lambda_{\mathrm{reg}}\mathbf{I}$. Then the private ratios $\hat{r}_{ij}$ satisfy*

$$\sup_{\mathbf{T}\in\mathcal{S}} \max_{ij} \big|\hat{r}_{ij}(\mathbf{T}) - r_{ij}(\mathbf{T})\big| \;\le\; \Xi \triangleq \frac{e_b + \kappa\, e'_w}{\lambda_{\mathrm{reg}}}, \qquad \kappa = \max_{ij} \boldsymbol{\delta}_{ij}^T \big(\mathbf{S}_w^{ij}\big)^{-1} \boldsymbol{\delta}_{ij}, \tag{16}$$

where κ upper-bounds all the ratios (for the proof, see Appendix H).

On the event of Lemma 4.3 the hypotheses of Lemma 4.4 hold with $e'_w = 2e_w + \lambda_{\mathrm{reg}}$, which absorbs the semidefinite projection and the ridge. The third step is Proposition 3.2, which transfers this closeness to the worst-case ratio attained by the private solution, and chaining the three steps gives the following end-to-end guarantee.

Theorem 4.5 (End-to-end utility)**.** *Let $\widehat{\mathbf{T}}$ be a global minimizer of the private smoothed objective built from* (15)*. With probability at least $1-\gamma$,*

$$\min_{1\le i<j\le C} r_{ij}(\widehat{\mathbf{T}}) \;\ge\; \mathrm{OPT} - 2\,\Xi - \tau \log K, \tag{17}$$

with $\Xi = \widetilde{\mathcal{O}}\big(R^2\sqrt{d}\,\sqrt{\log(1/\delta)}\,(1+\kappa)(1+1/n_{\min})/(\varepsilon\lambda_{\mathrm{reg}})\big)$, the notation $\widetilde{\mathcal{O}}$ hiding absolute constants and logarithmic factors (for the proof, see Appendix H).

The bound splits into a purely statistical term 2Ξ and a purely algorithmic term $\tau \log K$, with no multiplicative interaction. It also shows why the smoothing helps twice over. The exact criterion picks one worst pair by a hard $\arg\max$, and once the statistics are privatized any two ratios within 2Ξ of one another are indistinguishable, so that choice may be made by noise alone, whereas the softmax weights react continuously instead.

Proposition 4.6 (Stability of the softmax weights). *Let* $\mathbf{w}(\mathbf{r}) = \mathrm{softmax}(-\mathbf{r}/\tau)$ *and let* $\|\hat{\mathbf{r}} - \mathbf{r}\|_\infty \leq \Xi$. *Then* $\|\mathbf{w}(\hat{\mathbf{r}}) - \mathbf{w}(\mathbf{r})\|_1 \leq 2\Xi/\tau$, *so the rule* $\tau \geq 2\Xi/\eta$ *keeps the weight profile within total-variation distance* η *of its noise-free counterpart. Balancing the weight-instability scale* $2\Xi\kappa/\tau$ *against the smoothing gap* $\tau \log K$ *of Proposition 3.2 gives* $\tau^\star(\varepsilon) = \sqrt{2\Xi\kappa/\log K}$, *which increases as the privacy budget tightens (for the proof, see Appendix H).*

Entropic smoothing therefore also stabilizes the privatized max–min problem, and its single parameter is set from the *public* noise scale σ. In the annealing schedule of Appendix C, $\tau^\star(\varepsilon)$ replaces $\rho/\log K$ as the floor temperature, so the noisier the statistics, the warmer we stop.

5 Numerical results

The non-private experiments follow the evaluation protocol of the worst-case separation literature [35, 41, 45], and the private ones follow standard practice in differential privacy [1, 12]. Every remaining detail is recorded in Appendix K.

Datasets. We evaluate on the standard benchmarks of this literature, **COIL-20**, **CMU-PIE**, **Caltech101** and **Stable ImageNet**, which cover object recognition, face recognition and large-scale image classification [17, 22, 29, 40]. Table S2 of Appendix K gives their sizes and the feature-extraction pipeline used for Caltech101 and Stable ImageNet. We apply PCA as preprocessing and retain 98% of the energy, following [33, 35, 41, 45].

Compared methods. We compare ESRA with eight non-private methods. LDA (1936) [18, 36] maximizes the ratio of one between-class to one within-class scatter matrix, HLDA (2004) [26] extends it to unequal class covariances, WDA-eig (2020) [25] maximizes an entropically regularized Wasserstein separation, and RLDA-$\ell_{2,1}$ (2021) [33] makes LDA robust to outliers, but all four aggregate the class pairs into a single criterion, so the closest pair is left unprotected. WHMMDA (2018) [41, 42], QP-MMDA (2025) and Sparse-MMDA (2025) [35] enlarge the smallest pairwise Chernoff distance, which is a distance and not a ratio, so none of them measures a pair against its own within-class scatter, and each solves a semidefinite or a quadratic program at every iteration. WDDR/MMRA (2024) [45] optimizes PWCRA itself through the semidefinite relaxation (in the sense of losing tightness) of Section 1, with no guarantee of optimality and at least two hyperparameters to tune. Appendix J describes each method in more detail, and every competitor is run at the configuration its own paper prescribes. Since no differentially private worst-case separation method exists, the private study sets DP-ESRA against non-private ESRA and the per-step gradient-perturbation solver of Appendix P.

Protocol. Each dataset is split in half over 100 paired, seeded draws, and we report the mean error under the 1-NN, NM and QDF rules, in percent throughout. To obtain qualitative results, and to show how the proposed method and its competitors separate the classes even in the worst-case regime of the target dimension, we examine the two most severe reductions, $m = 2$ here and $m = 3$ in Appendix L, in which a projection is most constrained and a criterion protecting the average has nowhere to hide a badly separated pair, and at $m = 2$ the learned subspaces can moreover be plotted directly, as Figure 1 does. On the other hand, displaying every class of a benchmark at once would produce a heavily crowded space in which little can be read, so the reported experiments and all projection plots restrict each dataset to a few of its classes, and the released code enables any other subset to be selected and plotted. The private runs use the budgets $\varepsilon \in \{0.5, 1, 2, 4, 8, 16, \infty\}$ at $\delta = 10^{-5}$, over 50 fresh seeded splits and with no tuning on private data.

5.1 Classification results at $m = 2$

As shown in Table 1, on **Caltech101**, the harder of the two deep-feature datasets, ESRA attains a 1-NN error of 10.45 against 13.94 for the strongest competitor, WDDR/MMRA, a relative reduction of 25.0%, and under the nearest-mean rule its 10.06 improves on 14.18 by 29.1%. The two closest are precisely those that also normalize by a within-class quantity, WDDR/MMRA and RLDA-$\ell_{2,1}$, whereas the purely distance-based methods sit between 14.4 and 14.9, and LDA, which averages over pairs instead of protecting the worst one, sits at 17.18. Under QDF ESRA again returns the lowest error, 11.70 against 11.79, the two being on a par at $p = 0.737$, since a quadratic rule already models a

Table 1: Classification error in percent at $m = 2$ over 100 splits, best in bold, second best underlined.

Dataset	Caltech101			Stable ImageNet		
Method \ Classifier	1-NN	NM	QDF	1-NN	NM	QDF
LDA (1936) [18, 36]	17.18	16.82	16.87	<u>25.08</u>	<u>21.73</u>	21.74
HLDA (2004) [26]	14.53	17.62	13.86	26.94	23.58	22.51
WHMMDA (2018) [41, 42]	14.81	18.18	14.86	28.06	23.45	21.96
WDA-eig (2020) [25]	33.59	27.29	28.67	26.53	23.17	23.29
RLDA-$\ell_{2,1}$ (2021) [33]	14.50	15.77	12.56	39.07	28.62	28.77
WDDR/MMRA (2024) [45]	<u>13.94</u>	<u>14.18</u>	<u>11.79</u>	35.26	27.85	27.08
QP-MMDA (2025) [35]	14.41	17.50	14.43	27.32	22.55	<u>21.05</u>
Sparse-MMDA (2025) [35]	14.50	18.00	14.34	28.44	23.60	22.15
ESRA (proposed)	**10.45**	**10.06**	**11.70**	**18.15**	**16.37**	**15.61**

Dataset	CMU-PIE			COIL-20		
Method \ Classifier	1-NN	NM	QDF	1-NN	NM	QDF
LDA (1936) [18, 36]	4.63	11.02	4.31	<u>4.14</u>	<u>4.07</u>	<u>4.04</u>
HLDA (2004) [26]	9.54	16.24	9.48	10.68	16.30	10.61
WHMMDA (2018) [41, 42]	13.76	22.20	14.04	24.31	38.03	27.31
WDA-eig (2020) [25]	5.52	11.74	4.87	5.39	5.14	5.02
RLDA-$\ell_{2,1}$ (2021) [33]	51.26	55.15	51.50	12.72	24.90	12.29
WDDR/MMRA (2024) [45]	<u>3.16</u>	<u>7.54</u>	<u>2.77</u>	10.88	16.13	11.19
QP-MMDA (2025) [35]	16.45	24.63	16.44	27.11	42.71	31.23
Sparse-MMDA (2025) [35]	16.23	24.20	17.24	24.69	38.12	27.26
ESRA (proposed)	**0.45**	**7.10**	**0.39**	**1.16**	**1.16**	**1.29**

full covariance per class in the projected space and so narrows the room any projection has to improve on another. On **Stable ImageNet** the advantage is larger and uniform, 27.6%, 24.7% and 25.8% below the best competitor under the three rules. This dataset is informative because the ordering of the baselines is almost reversed, WDDR/MMRA and RLDA-$\ell_{2,1}$, the two best competitors on Caltech101, being here the two weakest under 1-NN, at 35.26 and 39.07, while plain LDA becomes the strongest of them. ESRA is the only method at the top on both, as one expects from a criterion that protects the least separated pair.

The two image datasets show the same behaviour with much larger margins, because the classes are close to separable once the right subspace is found. On **CMU-PIE**, ESRA reaches 0.45 under 1-NN and 0.39 under QDF against 3.16 and 2.77 for WDDR/MMRA, dividing the error by roughly seven. Under the nearest-mean rule it again holds the lower error, 7.10 to the 7.54 of WDDR/MMRA, the two being on a par at $p = 0.484$, and RLDA-$\ell_{2,1}$ is the weakest method here, above 51 under all three classifiers, which is consistent with only 42 samples per class. On **COIL-20**, ESRA gives 1.16, 1.16 and 1.29 against 4.14, 4.07 and 4.04 for LDA, a reduction of 68–72%, while the distance-based worst-case methods stay an order of magnitude behind, between 24 and 43, which is the point of Section 2.1. The paired t-tests of Appendix M confirm the advantage at the 5% level across the table, the two cells just discussed being on a par.

Qualitative analysis. On COIL-20 (Figure 1) the ESRA panel is free of overlap, whereas QP-MMDA and WHMMDA, which never measure a pair against its own within-class scatter, leave substantial mixing, in agreement with the error rates of Table 1. The projections on the three remaining datasets are shown and discussed in Appendix Q.

5.2 Privacy–utility trade-off

Table 2 reads the price of privacy directly. The trade-off is monotone on every dataset and classifier, and most of the descent happens between $\varepsilon = 1$ and $\varepsilon = 8$, so a moderate budget already buys most of the accuracy a looser one delivers. The consistency check passes, since on these same 50 seeded splits the $\varepsilon = \infty$ column reproduces non-private ESRA digit for digit, which is the empirical counterpart of Theorem 4.2. How sharply a dataset feels the budget is governed by the samples per class, exactly as the $1/n_{\min}$ term of Theorem 4.5 predicts, so the effect is strongest on CMU-PIE, which has only 42 samples per class, whereas Caltech101 stays within a factor of three of its noise-free error at $\varepsilon = 8$. The same numbers are plotted, with their dispersion across splits, in Appendix Q, and Appendix P sets DP-ESRA against a per-step gradient-perturbation solver of the same objective, which isolates the benefit of paying for privacy once.

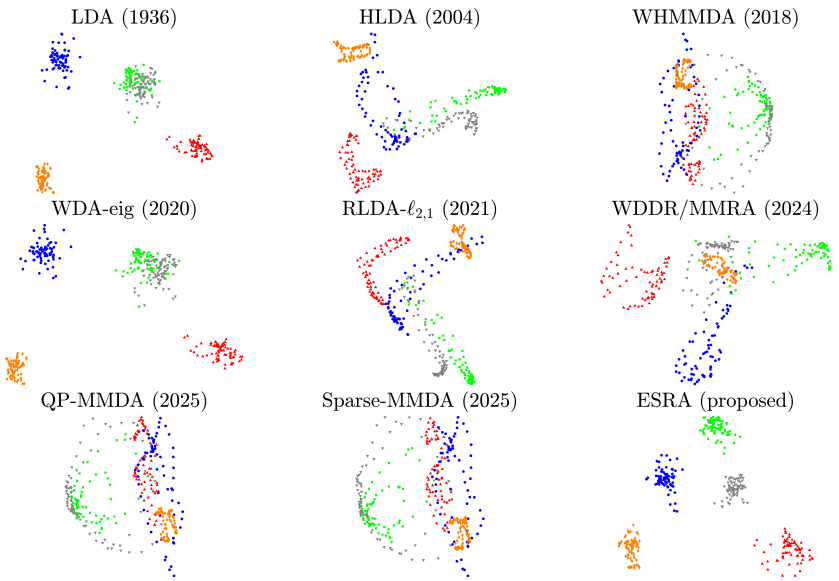

Figure 1: Two-dimensional projections on COIL-20 at $m = 2$, the methods referenced as in Table 1.

Table 2: Error of DP-ESRA in percent against the budget ε, at $\delta = 10^{-5}$ and $m = 2$, over 50 splits.

Dataset	Classifier	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	$\varepsilon = 8$	$\varepsilon = 16$	$\varepsilon = \infty$
COIL-20	1-NN	40.06	34.52	24.82	18.21	15.24	11.64	0.86
	NM	48.29	41.47	31.43	23.22	19.26	14.67	0.90
	QDF	43.02	37.18	27.39	20.14	16.27	12.34	0.94
CMU-PIE	1-NN	47.07	44.40	37.81	25.90	15.98	9.68	0.59
	NM	49.35	45.22	39.64	27.41	16.90	11.60	7.01
	QDF	45.35	41.60	35.31	23.70	14.55	8.25	0.23
Caltech101	1-NN	46.09	44.84	38.57	28.03	19.51	13.90	10.59
	NM	34.23	33.20	28.53	21.41	14.58	11.04	10.04
	QDF	35.28	34.04	29.87	22.30	15.74	12.09	11.85
Stable ImageNet	1-NN	54.24	50.35	45.03	40.40	37.87	35.41	17.92
	NM	46.56	41.78	35.79	31.58	29.98	27.88	15.59
	QDF	46.54	41.73	35.76	31.81	30.19	28.08	15.10

6 Limitations

Our problem is non-convex, and for a problem of that kind no algorithm can guarantee a global minimizer [30]. We establish monotone descent, and only objective convergence and the stationarity of the limit points can be proved.

7 Conclusion

We proposed ESRA, a single-loop algorithm for pairwise worst-case ratio analysis built on entropic smoothing, and its differentially private counterpart DP-ESRA. Reversing the sign of the criterion and replacing the resulting pointwise maximum by the log-sum-exp function turns the fractional program into a smooth minimization on the Stiefel manifold, whose every iteration is a softmax-weighted gradient assembly followed by one closed-form polar-factor update. Releasing the per-class sufficient statistics once through a Gaussian mechanism makes the whole optimization post-processing, so the privacy guarantee is independent of the number of iterations. We proved monotone descent, stationary convergence, an end-to-end utility bound in which the statistical error of the private statistics and the smoothing gap enter additively, and a stability bound that calibrates the temperature to the noise level of the mechanism. The private variant follows the privacy budget closely, and in terms of the nearest-neighbour, nearest-mean and quadratic discriminant error rates, supported by paired significance tests, ESRA outperforms the state-of-the-art and benchmark methods, with the learned projections showing the same behaviour.

AI USE STATEMENT

AI-assisted editing was used for language editing and proofreading of the main paper and supplementary materials. The authors take full responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

We release, as supplementary material submitted together with this paper, the complete MATLAB source code of the proposed methods — the non-private solver ESRA of Section 3, the private pipeline DP-ESRA of Section 4, and the per-step gradient-perturbation comparator of Appendix P — together with the scripts of our experiments and the datasets on which they were run. On the theoretical side, all assumptions are stated explicitly, the boundedness assumption that the private pipeline needs being Assumption 1 of Section 4, and complete proofs of every lemma, proposition and theorem are given in Appendices D–H. On the experimental side, the full protocol — class selection, the rule for the principal component analysis, the split construction, the seeding that makes every comparison paired, the classifiers, and the calibration of the privacy noise — is specified in Appendix K, and Appendix O reports the ablations.

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 308–318, 2016.

[2] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

[3] Prabhu Babu, Petre Stoica, and Astha Saini. Fair Principal Component Analysis (PCA): Minorization-Maximization Algorithms for Fair PCA, Fair Robust PCA and Fair Sparse PCA. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=6jTQrr3APY.

[4] Borja Balle and Yu-Xiang Wang. Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning (ICML)*, pp. 394–403, 2018.

[5] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *IEEE 55th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 464–473, 2014.

[6] Amir Beck and Marc Teboulle. Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization*, 22(2):557–580, 2012.

[7] Wei Bian and Dacheng Tao. Max-min distance analysis by using sequential SDP relaxation for dimension reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):1037–1050, 2011. doi: 10.1109/TPAMI.2010.189.

[8] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the SuLQ framework. In *Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pp. 128–138, 2005.

[9] Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023.

[10] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference (TCC)*, pp. 635–658, 2016.

[11] Xiaojun Chang, Feiping Nie, Sen Wang, Yi Yang, Xiaofang Zhou, and Chengqi Zhang. Compound rank-k projections for bilinear analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 27(7): 1502–1513, 2016. doi: 10.1109/TNNLS.2015.2441735.

[12] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

[13] Kamalika Chaudhuri, Anand D Sarwate, and Kaushik Sinha. A near-optimal algorithm for differentially-private principal components. *Journal of Machine Learning Research*, 14:2905–2943, 2013.

[14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–487, 2014.

[15] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*, pp. 265–284, 2006.

[16] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze Gauss: Optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 11–20, 2014.

[17] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 178, 2004.

[18] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2): 179–188, 1936.

[19] Jie Gui, Zhenan Sun, Shuiwang Ji, Dacheng Tao, and Tieniu Tan. Feature selection based on structured sparsity: A comprehensive study. *IEEE Transactions on Neural Networks and Learning Systems*, 28(7): 1490–1507, 2017. doi: 10.1109/TNNLS.2016.2551724.

[20] Onur C. Hamsici and Aleix M. Martinez. Bayes optimality in linear discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):647–657, 2008. doi: 10.1109/TPAMI.2007.70717.

[21] David R Hunter and Kenneth Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1): 30–37, 2004.

[22] Vitaliy Kinakh. Stable ImageNet-1K. `https://www.kaggle.com/datasets/vitaliykinakh/stable-imagenet1k`, 2022.

[23] Kenneth Lange, David R Hunter, and Ilsoon Yang. Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9(1):1–20, 2000.

[24] Xiaoping Li, Yadi Wang, and Rubén Ruiz. A survey on sparse learning models for feature selection. *IEEE Transactions on Cybernetics*, 52(3):1642–1660, 2022. doi: 10.1109/TCYB.2020.2982445.

[25] Hexuan Liu, Yunfeng Cai, You-Lin Chen, and Ping Li. Ratio trace formulation of Wasserstein discriminant analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Adv. Neural Inf. Process. Syst.*, volume 33, pp. 16821–16832. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/c37f9e1283cbd4a6edfd778fc8b1c652-Paper.pdf`.

[26] Marco Loog and Robert P. W. Duin. Linear dimensionality reduction via a heteroscedastic extension of LDA: The Chernoff criterion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6): 732–739, 2004. doi: 10.1109/TPAMI.2004.13.

[27] Albert W. Marshall and Ingram Olkin. *Inequalities: Theory of Majorization and Its Applications*. Academic Press, New York, 1979.

[28] Ilya Mironov. Rényi differential privacy. In *IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275, 2017.

[29] Sameer A. Nene, Shree K. Nayar, and Hiroshi Murase. Columbia object image library (COIL-20). Technical Report CUCS-005-96, Department of Computer Science, Columbia University, 1996.

[30] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.

[31] Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1): 127–152, 2005.

[32] Feiping Nie, Xia Dong, and Xuelong Li. Unsupervised and semisupervised projection with graph optimization. *IEEE Transactions on Neural Networks and Learning Systems*, 32(4):1547–1559, 2021. doi: 10.1109/TNNLS.2020.2984958.

[33] Feiping Nie, Zheng Wang, Rong Wang, Zhen Wang, and Xuelong Li. Towards robust discriminative projections learning via non-greedy $\ell_{2,1}$-norm minmax. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(6): 2086–2100, 2021. doi: 10.1109/TPAMI.2019.2961877.

[34] Feiping Nie, Xia Dong, Zhanxuan Hu, Rong Wang, and Xuelong Li. Discriminative projected clustering via unsupervised LDA. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):9466–9480, 2023. doi: 10.1109/TNNLS.2022.3202719.

[35] Mohammad Mahdi Omati, Prabhu Babu, Petre Stoica, and Arash Amini. A max-min approach to the worst-case class separation problem. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=EEmwBd4tfZ.

[36] C Radhakrishna Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):159–193, 1948.

[37] Meisam Razaviyayn, Mingyi Hong, and Zhi-Quan Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.

[38] Astha Saini, Petre Stoica, Prabhu Babu, and Aakash Arora. Min-max framework for majorization-minimization algorithms in signal processing applications: An overview. *Foundations and Trends® in Signal Processing*, 18(4):310–389, 2024.

[39] Kaiming Shen and Wei Yu. Fractional programming for communication systems—Part I: Power control and beamforming. *IEEE Transactions on Signal Processing*, 66(10):2616–2630, 2018.

[40] Terence Sim, Simon Baker, and Maan Bsat. The CMU pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615–1618, 2003.

[41] Bing Su, Xiaoqing Ding, Changsong Liu, and Ying Wu. Heteroscedastic max-min distance analysis. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4539–4547, 2015. doi: 10.1109/CVPR.2015.7299084.

[42] Bing Su, Xiaoqing Ding, Changsong Liu, and Ying Wu. Heteroscedastic max–min distance analysis for dimensionality reduction. *IEEE Transactions on Image Processing*, 27(8):4052–4065, 2018. doi: 10.1109/TIP.2018.2836312.

[43] Ying Sun, Prabhu Babu, and Daniel P. Palomar. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Transactions on Signal Processing*, 65(3):794–816, 2017. doi: 10.1109/TSP.2016.2601299.

[44] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[45] Zheng Wang, Feiping Nie, Canyu Zhang, Rong Wang, and Xuelong Li. Worst-case discriminative feature learning via max-min ratio analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(1):641–658, 2024. doi: 10.1109/TPAMI.2023.3323453.

[46] Yaoliang Yu, Jiayan Jiang, and Liming Zhang. Distance metric learning by minimal distance maximization. *Pattern Recognition*, 44(3):639–649, 2011. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2010.09.019. URL https://www.sciencedirect.com/science/article/pii/S0031320310004590.

[47] Yu Zhang and Dit-Yan Yeung. Worst-case linear discriminant analysis. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/4e4b5fbbbb602b6d35bea8460aa8f8e5-Paper.pdf.

SUPPLEMENTARY MATERIAL

This material is organized as follows. Appendix A reviews the majorization–minimization principle and Appendix B the differential-privacy framework, both in the form in which the paper uses them and both self-contained, so that a reader unfamiliar with either may start there. Appendix C states the pseudocode. Appendix D proves the smoothing results, Appendix E the majorization results together with the explicit gradient and aggregation formulas, and Appendix F establishes the per-iteration cost and proves the convergence theorem. Appendix G proves the privacy results and Appendix H the utility results. Appendix I compares the structure and cost of the available solvers, Appendix J reviews the related work and the compared methods and Appendix K records the complete experimental protocol. Appendices L–Q contain the remaining experiments. Every derivation is written out step by step.

A THE MAJORIZATION–MINIMIZATION PRINCIPLE

A.1 THE GENERAL FRAMEWORK

Majorization–minimization (MM) is a principle for building iterative schemes for constrained minimization problems of the form

$$\min_{\mathbf{T}\in\chi} F(\mathbf{T}), \tag{S1}$$

where F is the objective, $\mathbf{T}$ is the optimization variable and χ is the feasible set. Instead of attacking (S1) directly, one solves a sequence of simpler problems. At iteration t a surrogate function $H(\mathbf{T} \mid \mathbf{T}^t)$, called a majorizer of F at $\mathbf{T}^t$, is constructed so that

$$H(\mathbf{T} \mid \mathbf{T}^t) \;\geq\; F(\mathbf{T}), \qquad \forall\, \mathbf{T} \in \chi, \tag{S2}$$

$$H(\mathbf{T}^t \mid \mathbf{T}^t) \;=\; F(\mathbf{T}^t). \tag{S3}$$

Condition (S2) says that the surrogate is a global upper bound on F over the feasible set, and condition (S3) says that it touches F at the current iterate. The next iterate is obtained by minimizing the surrogate,

$$\mathbf{T}^{t+1} \in \operatorname*{arg\,min}_{\mathbf{T}\in\chi} H(\mathbf{T} \mid \mathbf{T}^t), \tag{S4}$$

and the procedure is repeated until the relative change of the iterate or of the objective falls below a prescribed tolerance.

The essential property of the scheme is that the objective never increases. Indeed,

$$F(\mathbf{T}^{t+1}) \overset{(a)}{\leq} H(\mathbf{T}^{t+1} \mid \mathbf{T}^t) \overset{(b)}{\leq} H(\mathbf{T}^t \mid \mathbf{T}^t) \overset{(c)}{=} F(\mathbf{T}^t), \tag{S5}$$

where (a) is the majorization condition (S2), (b) follows because $\mathbf{T}^{t+1}$ minimizes $H(\cdot \mid \mathbf{T}^t)$ over χ and $\mathbf{T}^t$ is feasible, and (c) is the tangency condition (S3). The sequence $\{F(\mathbf{T}^t)\}$ is therefore non-increasing, and it converges whenever F is bounded below on χ. A detailed treatment of the majorization–minimization principle, including systematic ways of building surrogates, can be found in [21, 23, 37, 43].

A.2 MAJORIZING A POINTWISE MAXIMUM

The objective of Section 3 is a pointwise maximum,

$$F(\mathbf{T}) = \max_{k=1,\dots,K} f_k(\mathbf{T}), \tag{S6}$$

and it can be majorized piece by piece [38]. Let $g_k(\mathbf{T} \mid \mathbf{T}^t)$ be a majorizer of f_k, so that $g_k(\cdot \mid \mathbf{T}^t) \geq f_k$ on χ and $g_k(\mathbf{T}^t \mid \mathbf{T}^t) = f_k(\mathbf{T}^t)$. Then the pointwise maximum of the individual majorizers,

$$g(\mathbf{T} \mid \mathbf{T}^t) \triangleq \max_{k=1,\dots,K} g_k(\mathbf{T} \mid \mathbf{T}^t), \tag{S7}$$

is a valid majorizer of F. The upper-bound property follows from

$$g(\mathbf{T} \mid \mathbf{T}^t) = \max_k g_k(\mathbf{T} \mid \mathbf{T}^t) \;\geq\; \max_k f_k(\mathbf{T}) = F(\mathbf{T}), \tag{S8}$$

and the tangency property from

$$g(\mathbf{T}^t \mid \mathbf{T}^t) = \max_k g_k(\mathbf{T}^t \mid \mathbf{T}^t) = \max_k f_k(\mathbf{T}^t) = F(\mathbf{T}^t). \tag{S9}$$

The construction used in the paper is (S7) with the maximum replaced by the log-sum-exp function, which is itself an upper bound on the maximum by Lemma 3.1 and is monotone in each coordinate, so both properties are preserved.

B THE DIFFERENTIAL-PRIVACY FRAMEWORK

This appendix reviews differential privacy in the form in which Section 4 uses it, following [14, 15].

B.1 THE DEFINITION AND ITS PARAMETERS

The released projection matrix is a function of the training data, and its value may depend noticeably on any single record, so that an adversary in possession of every other record could test whether a given individual is present. Differential privacy rules this out by making the released object random and by requiring its distribution to be almost unchanged when a single record is replaced.

Two datasets $\mathbf{X}$ and $\mathbf{X}'$ are *neighbouring*, written $\mathbf{X} \sim \mathbf{X}'$, when one is obtained from the other by replacing a single labelled record and leaving the remaining $n-1$ untouched. The other common convention adds or removes a record instead; we use replacement throughout, since the class sizes n_k are themselves part of what we release.

Definition B.1 ((ε, δ)-differential privacy [14, 15])**.** A randomized mechanism $\mathcal{A}$ is (ε, δ)-differentially private if, for every pair of neighbouring datasets $\mathbf{X} \sim \mathbf{X}'$ and every measurable set E of outputs,

$$\Pr[\mathcal{A}(\mathbf{X}) \in E] \;\le\; e^{\varepsilon}\, \Pr[\mathcal{A}(\mathbf{X}') \in E] + \delta. \tag{S10}$$

The two parameters are not interchangeable. The factor e^{ε} bounds the ratio by which any outcome becomes more likely when a record is swapped in, so that ε is the privacy budget, and as $\varepsilon \to 0$ the two output distributions coalesce and the release carries almost nothing about the record, whereas a large ε permits the output to depend strongly on it. The additive δ allows that bound to fail on an event of probability at most δ, and is kept well below $1/n$; were it not, the mechanism that publishes a uniformly drawn record verbatim would already satisfy (S10). What gives the definition its strength is the quantification over every measurable E, since the bound then holds for all events at once, so that no test an adversary may construct, and no quantity of side information, separates the two datasets by more than (S10) allows. The condition constrains the two output distributions and makes no assumption on the adversary's beliefs, so the guarantee is insensitive to what the adversary already knows.

B.2 SENSITIVITY AND THE GAUSSIAN MECHANISM

A mechanism is built from the amount by which the quantity to be released can move under a single replacement. For a deterministic query q with values in a Euclidean space this is its ℓ_2-*sensitivity*,

$$\Delta_2 \;\triangleq\; \max_{\mathbf{X} \sim \mathbf{X}'} \; \big\| q(\mathbf{X}) - q(\mathbf{X}') \big\|_2, \tag{S11}$$

the maximum running over all neighbouring pairs in the domain, so that Δ_2 does not depend on the dataset at hand and is a public constant, available before any data are seen. Assumption 1 enters the privacy argument here and nowhere else, since absent a bound on $\|\mathbf{x}_\ell\|_2$ a single record could displace q arbitrarily far, Δ_2 would be infinite, and no finite noise level would suffice. Lemma 4.1 evaluates (S11) for the sufficient statistics (13).

Given Δ_2, the Gaussian mechanism releases $q(\mathbf{X}) + \mathcal{N}(0, \sigma^2 \mathbf{I})$, and that release is (ε, δ)-differentially private for $\varepsilon \in (0, 1)$ whenever

$$\sigma \;\ge\; \Delta_2 \sqrt{2 \log(1.25/\delta)} \big/ \varepsilon \tag{S12}$$

[14]. The dependence displayed in (S12) is what later propagates into the utility bound of Theorem 4.5, since the noise grows linearly in the sensitivity and inversely in the budget, but only as $\sqrt{\log(1/\delta)}$,

so that δ may be made very small at little cost, and the measurements of Appendix O bear this out. Nothing in the paper rests on this particular calibration, since the analytic calibration of [4] and the concentrated-DP accountants of [10, 28] may be substituted throughout and alter only the constant in σ.

When the query returns a symmetric matrix, as the second-moment blocks of (13) do, the noise is drawn independently on and above the diagonal and then mirrored below it, which is the Analyze Gauss construction of [16]. The calibration then uses the sensitivity of the upper-triangular vectorization, and the Frobenius bound of Lemma 4.1 dominates that quantity because it counts the off-diagonal entries twice, so the calibration we apply is conservative.

B.3 Post-processing and composition

Two properties of the definition govern how a private algorithm ought to be arranged, and they pull in opposite directions.

The first is invariance under post-processing, in the sense that if $\mathcal{A}$ is (ε, δ)-differentially private and f is any map that does not access the data, then $f \circ \mathcal{A}$ is (ε, δ)-differentially private with the same parameters [14]. For deterministic f this is immediate from (S10) applied to the preimage $f^{-1}(E)$, which is itself a measurable set of outputs of $\mathcal{A}$,

$$\begin{aligned} \Pr\big[f(\mathcal{A}(\mathbf{X})) \in E\big] &= \Pr\big[\mathcal{A}(\mathbf{X}) \in f^{-1}(E)\big] \\ &\le\ e^{\varepsilon} \Pr\big[\mathcal{A}(\mathbf{X}') \in f^{-1}(E)\big] + \delta \ =\ e^{\varepsilon} \Pr\big[f(\mathcal{A}(\mathbf{X}')) \in E\big] + \delta, \end{aligned} \tag{S13}$$

and a randomized f follows by averaging over its internal randomness. No computation performed on an already private release recovers information that the release does not contain.

The second is composition, since when the data are queried T times, the t-th mechanism being $(\varepsilon_t, \delta_t)$-private, then the combined release is $\big(\sum_t \varepsilon_t, \sum_t \delta_t\big)$-private. Advanced composition improves the first sum to order $\sqrt{T \log(1/\delta')}\,\varepsilon + T\varepsilon(e^{\varepsilon} - 1)$ at the price of a slack δ', and the concentrated and Rényi accountants [1, 10, 28] sharpen the constants further [14]; under every one of them, however, the budget still grows with T.

Together these two properties account for the design of Section 4. A solver that consults the data at every step, as DP-SGD and its relatives do [1, 5], falls under composition, so that the iteration count becomes a privacy parameter and accuracy has to be traded against running the optimizer to convergence. A solver that consults the data once, and whose every subsequent step is a data-independent function of that single release, falls under post-processing instead, and its iteration count leaves the privacy account untouched. ESRA admits the second arrangement because the data enter only through the sufficient statistics (13), and Theorem 4.2 states this formally.

C Pseudocode of ESRA

Algorithm S1 states the method derived in Section 3. Every line is closed-form, in that the ratios and the softmax weights are evaluated directly, the gradient is assembled through the per-class aggregation of Appendix E.2, and the update is the polar factor of Theorem 3.4. There is no inner solver and no outer search on the value of the ratio.

Algorithm S1 ESRA, a single-loop entropic-smoothing solver for the PWCRA problem (2)

Require: $\{\boldsymbol{\delta}_{ij}\}$, $\{\boldsymbol{\Sigma}_k^w\}$, target dimension m, temperature τ, tolerance ϵ_{tol}, iteration cap $T_{\max}$.
1: Initialize $\mathbf{T}^0 \in \mathcal{S}$ and set $t \leftarrow 0$.
2: **repeat**
3: Evaluate $u_{ij}^t = \|(\mathbf{T}^t)^T \boldsymbol{\delta}_{ij}\|_2^2$, $v_{ij}^t = \mathrm{tr}((\mathbf{T}^t)^T(\boldsymbol{\Sigma}_i^w + \boldsymbol{\Sigma}_j^w)\mathbf{T}^t)$, $r_{ij}^t = u_{ij}^t / v_{ij}^t$.
4: Compute $w_{ij}^t = [\mathrm{softmax}(-\mathbf{r}^t/\tau)]_{ij}$ in shifted form and $d_{ij}^t = u_{ij}^t/(v_{ij}^t)^2$.
5: Assemble $\nabla F_\tau(\mathbf{T}^t)$ from (12) using the per-class aggregation, and $\widehat{L}_\tau^t$ from Lemma E.1.
6: Form $\mathbf{B}_t = \widehat{L}_\tau^t \mathbf{T}^t - \nabla F_\tau(\mathbf{T}^t)$, take the thin SVD $\mathbf{B}_t = \mathbf{U}\mathbf{D}\mathbf{V}^T$, set $\mathbf{T}^{t+1} = \mathbf{U}\mathbf{V}^T$, $t \leftarrow t+1$.
7: **until** $\|\mathbf{T}^t - \mathbf{T}^{t-1}\|_F / \|\mathbf{T}^{t-1}\|_F \le \epsilon_{\mathrm{tol}}$ **or** $t \ge T_{\max}$
8: **Output:** $\mathbf{T}^\star = \mathbf{T}^t$.

Annealing schedule. The temperature is annealed over stages, starting from a data-scaled τ_0 that is halved after each stage, with a warm start from the previous solution, down to the floor $\tau_{\mathrm{f}} = \rho/\log K$ dictated by Proposition 3.2. Within each stage τ is fixed, so Theorem F.1 applies to every stage, and the only change the private setting makes to this schedule is that the floor is raised, as Section 4 explains.

D PROOFS OF THE SMOOTHING RESULTS

D.1 PROOF OF LEMMA 3.1

Proof. Fix $\mathbf{T}$, write $h_{ij}(\mathbf{T}) = -r_{ij}(\mathbf{T})$ for the sign-reversed ratios and $x_{ij} = h_{ij}(\mathbf{T})$ for $1 \le i < j \le C$, and let $M = \max_{ij} x_{ij}$. Since every term of the sum is positive and one of them equals $e^{M/\tau}$, we have the elementary two-sided bound

$$e^{M/\tau} \;\le \sum_{1\le i<j\le C} e^{x_{ij}/\tau} \;\le\; K\, e^{M/\tau}, \tag{S14}$$

where the left inequality keeps only the largest term and the right inequality replaces each of the K terms by the largest one. Taking the logarithm of (S14), which is monotone, gives

$$\frac{M}{\tau} \;\le\; \log \sum_{1\le i<j\le C} e^{x_{ij}/\tau} \;\le\; \frac{M}{\tau} + \log K, \tag{S15}$$

and multiplying by $\tau > 0$ gives $M \le F_\tau(\mathbf{T}) \le M + \tau \log K$, which is the claimed sandwich. The bound holds at every $\mathbf{T}$ with the same constant $\tau \log K$, so F_τ decreases to $\max_{ij} h_{ij}$ uniformly on $\mathcal{S}$ as τ tends to zero. ■

D.2 PROOF OF PROPOSITION 3.2

Proof. Let $\mathbf{T}^\star \in \mathcal{S}$ attain $\mathrm{OPT} = \max_{\mathbf{T}\in\mathcal{S}} \min_{ij} r_{ij}(\mathbf{T})$. By the identity (3),

$$\max_{ij} h_{ij}(\mathbf{T}^\star) = -\min_{ij} r_{ij}(\mathbf{T}^\star) = -\mathrm{OPT}. \tag{S16}$$

Now let $\widehat{\mathbf{T}}$ be a global minimizer of F_τ over $\mathcal{S}$. Chaining four inequalities gives

$$\begin{aligned} -\min_{ij} r_{ij}(\widehat{\mathbf{T}}) = \max_{ij} h_{ij}(\widehat{\mathbf{T}}) \;&\overset{(a)}{\le}\; F_\tau(\widehat{\mathbf{T}}) \;\overset{(b)}{\le}\; F_\tau(\mathbf{T}^\star) \\ &\overset{(c)}{\le}\; \max_{ij} h_{ij}(\mathbf{T}^\star) + \tau \log K \;\overset{(d)}{=}\; -\mathrm{OPT} + \tau \log K, \end{aligned} \tag{S17}$$

where (a) is the left half of Lemma 3.1, (b) holds because $\widehat{\mathbf{T}}$ minimizes F_τ and $\mathbf{T}^\star$ is feasible, (c) is the right half of Lemma 3.1, and (d) is (S16). Multiplying (S17) by -1 reverses the inequality and yields $\min_{ij} r_{ij}(\widehat{\mathbf{T}}) \ge \mathrm{OPT} - \tau \log K$. Finally, requiring $\tau \log K \le \rho$ is the same as requiring $\tau \le \rho/\log K$. ■

E PROOFS OF THE MAJORIZATION RESULTS

E.1 PROOF OF LEMMA 3.3

The proof of Lemma 3.3 proceeds in three steps, and it is their composition on the Stiefel manifold that yields an affine majorizer and removes the inner optimization from the loop. The first step replaces the ratio by a convex quadratic through the quadratic transform of fractional programming [39]. The second bounds that quadratic by its first-order expansion plus a curvature term, and the third uses the constancy of the Frobenius norm on $\mathcal{S}$, which gives $\|\mathbf{T} - \mathbf{T}^t\|_F^2 = 2m - 2\,\mathrm{tr}(\mathbf{T}^T\mathbf{T}^t)$ and reduces the curvature term from a quadratic to an affine one.

Proof. The proof has three steps.

Step 1, a variational form of a scalar ratio. Let $u \geq 0$ and $v > 0$, and consider the scalar function $\varphi(y) = 2y\sqrt{u} - y^2 v$ of $y \in \mathbb{R}$. It is a concave parabola, so its maximum is found by setting $\varphi'(y) = 2\sqrt{u} - 2yv = 0$, that is $y = \sqrt{u}/v$, and the maximum value is

$$\varphi\Big(\frac{\sqrt{u}}{v}\Big) = 2\frac{\sqrt{u}}{v}\sqrt{u} - \frac{u}{v^2}v = \frac{2u}{v} - \frac{u}{v} = \frac{u}{v}. \tag{S18}$$

Hence $u/v = \max_y\{2y\sqrt{u} - y^2 v\}$, and in particular, for the specific value $y_t = \sqrt{u_t}/v_t$ built from the current iterate,

$$\frac{u}{v} \;\geq\; 2y_t\sqrt{u} - y_t^2 v. \tag{S19}$$

Step 2, linearizing the square root. Put $u(\mathbf{T}) = \operatorname{tr}(\mathbf{T}^T\mathbf{A}\mathbf{T})$ and note that

$$\sqrt{u(\mathbf{T})} = \big\|\mathbf{A}^{1/2}\mathbf{T}\big\|_F, \tag{S20}$$

which is well defined because $\mathbf{A} \succcurlyeq \mathbf{0}$. The map $\mathbf{T} \mapsto \|\mathbf{A}^{1/2}\mathbf{T}\|_F$ is a seminorm, hence convex and positively homogeneous of degree one. For a convex, positively homogeneous function f one has both the subgradient inequality $f(\mathbf{T}) \geq f(\mathbf{T}^t) + \langle \nabla f(\mathbf{T}^t), \mathbf{T} - \mathbf{T}^t\rangle$ and Euler's identity $\langle \nabla f(\mathbf{T}^t), \mathbf{T}^t\rangle = f(\mathbf{T}^t)$, and combining the two gives the tangent bound

$$f(\mathbf{T}) \;\geq\; \langle \nabla f(\mathbf{T}^t), \mathbf{T}\rangle. \tag{S21}$$

Here $\nabla f(\mathbf{T}^t) = \mathbf{A}\mathbf{T}^t/\|\mathbf{A}^{1/2}\mathbf{T}^t\|_F = \mathbf{A}\mathbf{T}^t/\sqrt{u_t}$, so (S21) reads

$$\sqrt{u(\mathbf{T})} \;\geq\; \frac{\operatorname{tr}\big((\mathbf{T}^t)^T\mathbf{A}\mathbf{T}\big)}{\sqrt{u_t}}. \tag{S22}$$

Step 3, putting the two together. Substituting (S22) into (S19) with $y_t = \sqrt{u_t}/v_t > 0$ and $v = v(\mathbf{T}) = \operatorname{tr}(\mathbf{T}^T\mathbf{B}\mathbf{T})$ gives

$$\begin{aligned}
\frac{\operatorname{tr}(\mathbf{T}^T\mathbf{A}\mathbf{T})}{\operatorname{tr}(\mathbf{T}^T\mathbf{B}\mathbf{T})} \;&\geq\; 2\,\frac{\sqrt{u_t}}{v_t}\,\sqrt{u(\mathbf{T})} - \frac{u_t}{v_t^2}\,\operatorname{tr}\big(\mathbf{T}^T\mathbf{B}\mathbf{T}\big) \\
&\geq\; 2\,\frac{\sqrt{u_t}}{v_t}\cdot\frac{\operatorname{tr}\big((\mathbf{T}^t)^T\mathbf{A}\mathbf{T}\big)}{\sqrt{u_t}} - \frac{u_t}{v_t^2}\,\operatorname{tr}\big(\mathbf{T}^T\mathbf{B}\mathbf{T}\big) \\
&=\; \frac{2}{v_t}\,\operatorname{tr}\big((\mathbf{T}^t)^T\mathbf{A}\mathbf{T}\big) - \frac{u_t}{v_t^2}\,\operatorname{tr}\big(\mathbf{T}^T\mathbf{B}\mathbf{T}\big).
\end{aligned} \tag{S23}$$

Multiplying (S23) by -1 reverses the inequality and produces exactly (7). For the tightness, the right-hand side of (7) evaluated at $\mathbf{T} = \mathbf{T}^t$ is

$$\frac{u_t}{v_t^2}\,v_t - \frac{2}{v_t}\,u_t = \frac{u_t}{v_t} - \frac{2u_t}{v_t} = -\frac{u_t}{v_t}, \tag{S24}$$

which is the value of the left-hand side at $\mathbf{T}^t$. Hence equality holds at the expansion point. ■

E.2 The gradient of the pairwise majorizer and of the smoothed objective

Applying Lemma 3.3 pairwise gives the majorizer (9),

$$\hat{h}_{ij}(\mathbf{T} \mid \mathbf{T}^t) = d_{ij}^t\,\operatorname{tr}\big(\mathbf{T}^T\mathbf{S}_w^{ij}\mathbf{T}\big) - \frac{2}{v_{ij}^t}\,\operatorname{tr}\big((\mathbf{T}^t)^T\mathbf{S}_b^{ij}\mathbf{T}\big). \tag{S25}$$

Using $\partial\operatorname{tr}(\mathbf{T}^T\mathbf{S}\mathbf{T})/\partial\mathbf{T} = 2\mathbf{S}\mathbf{T}$ for symmetric $\mathbf{S}$, and $\partial\operatorname{tr}(\mathbf{P}^T\mathbf{T})/\partial\mathbf{T} = \mathbf{P}$ for constant $\mathbf{P}$, we obtain

$$\nabla\hat{h}_{ij}(\mathbf{T} \mid \mathbf{T}^t) = 2d_{ij}^t\mathbf{S}_w^{ij}\mathbf{T} - \frac{2}{v_{ij}^t}\mathbf{S}_b^{ij}\mathbf{T}^t, \qquad \nabla^2\hat{h}_{ij} = 2d_{ij}^t\big(\mathbf{I}_m \otimes \mathbf{S}_w^{ij}\big), \tag{S26}$$

the Hessian being written in the vectorized coordinates $\operatorname{vec}(\mathbf{T})$. In particular $\mathbf{0} \preccurlyeq \nabla^2\hat{h}_{ij} \preccurlyeq 2d_{ij}^t\lambda_{\max}(\mathbf{S}_w^{ij})\mathbf{I}$, so each $\hat{h}_{ij}(\cdot \mid \mathbf{T}^t)$ is a convex quadratic.

It is useful to check that the majorizer shares its gradient with h_{ij} at the expansion point. Differentiating $h_{ij} = -u_{ij}/v_{ij}$ directly, with $\nabla u_{ij}(\mathbf{T}) = 2\mathbf{S}_b^{ij}\mathbf{T}$ and $\nabla v_{ij}(\mathbf{T}) = 2\mathbf{S}_w^{ij}\mathbf{T}$, gives

$$\nabla h_{ij}(\mathbf{T}) = -\frac{2\mathbf{S}_b^{ij}\mathbf{T}\, v_{ij} - u_{ij}\, 2\mathbf{S}_w^{ij}\mathbf{T}}{v_{ij}^2} = \frac{2u_{ij}}{v_{ij}^2}\mathbf{S}_w^{ij}\mathbf{T} - \frac{2}{v_{ij}}\mathbf{S}_b^{ij}\mathbf{T}, \tag{S27}$$

and evaluating (S27) at $\mathbf{T} = \mathbf{T}^t$ reproduces (S26) at $\mathbf{T} = \mathbf{T}^t$, since $u_{ij}^t/(v_{ij}^t)^2 = d_{ij}^t$.

It remains to combine the pairs. Write $F_\tau = \mathrm{lse}_\tau \circ \mathbf{h}$, where $\mathrm{lse}_\tau(\mathbf{x}) = \tau \log \sum_k e^{x_k/\tau}$ and $\mathbf{h}(\mathbf{T})$ collects the K values $h_{ij}(\mathbf{T})$. Differentiating $F_\tau(\mathbf{T}) = \tau \log \sum_{ij} e^{h_{ij}(\mathbf{T})/\tau}$ by the chain rule gives

$$\nabla F_\tau(\mathbf{T}) = \tau \cdot \frac{1}{\sum_{kl} e^{h_{kl}(\mathbf{T})/\tau}} \sum_{ij} e^{h_{ij}(\mathbf{T})/\tau} \cdot \frac{1}{\tau}\nabla h_{ij}(\mathbf{T}) = \sum_{ij} \frac{e^{h_{ij}(\mathbf{T})/\tau}}{\sum_{kl} e^{h_{kl}(\mathbf{T})/\tau}} \nabla h_{ij}(\mathbf{T}), \tag{S28}$$

and the coefficient of ∇h_{ij} is $[\mathrm{softmax}(\mathbf{h}(\mathbf{T})/\tau)]_{ij} = [\mathrm{softmax}(-\mathbf{r}(\mathbf{T})/\tau)]_{ij} = w_{ij}(\mathbf{T})$ because $h_{ij} = -r_{ij}$. Evaluating (S28) at $\mathbf{T} = \mathbf{T}^t$ with (S27) yields the gradient (12) used in Algorithm S1.

Derivation of the per-class aggregation. The denominator part of (12) is $\sum_{i<j} w_{ij}^t d_{ij}^t \mathbf{S}_w^{ij}\mathbf{T}^t$, which naively requires $K = C(C-1)/2$ products of a $d \times d$ matrix with a $d \times m$ matrix. Substituting $\mathbf{S}_w^{ij} = \mathbf{\Sigma}_i^w + \mathbf{\Sigma}_j^w$ and expanding,

$$\begin{aligned}
\sum_{1\le i<j\le C} w_{ij}^t d_{ij}^t\, \mathbf{S}_w^{ij}\mathbf{T}^t &= \sum_{1\le i<j\le C} w_{ij}^t d_{ij}^t \big(\mathbf{\Sigma}_i^w + \mathbf{\Sigma}_j^w\big)\mathbf{T}^t \\
&= \sum_{1\le i<j\le C} w_{ij}^t d_{ij}^t \mathbf{\Sigma}_i^w \mathbf{T}^t + \sum_{1\le i<j\le C} w_{ij}^t d_{ij}^t \mathbf{\Sigma}_j^w \mathbf{T}^t \\
&= \sum_{k=1}^{C} \Big(\sum_{j>k} w_{kj}^t d_{kj}^t + \sum_{i<k} w_{ik}^t d_{ik}^t \Big) \mathbf{\Sigma}_k^w \mathbf{T}^t = \sum_{k=1}^{C} a_k^t\, \mathbf{\Sigma}_k^w \mathbf{T}^t,
\end{aligned} \tag{S29}$$

where in the third equality we collected, for each class index k, all the terms in which $\mathbf{\Sigma}_k^w$ appears, and where

$$a_k^t = \sum_{j>k} w_{kj}^t d_{kj}^t + \sum_{i<k} w_{ik}^t d_{ik}^t = \sum_{j\neq k} w_{kj}^t d_{kj}^t, \tag{S30}$$

using the symmetry $w_{ij}^t = w_{ji}^t$ and $d_{ij}^t = d_{ji}^t$ implied by the unordered pair notation. The scalars a_k^t cost $\mathcal{O}(K)$ to form, after which only C matrix products remain. This is the source of the $\mathcal{O}(Cd^2m)$ term in the per-iteration cost.

E.3 The smoothness constant of the composite surrogate

Lemma 3.3 requires a curvature bound for the composite surrogate (10). The following lemma supplies the constant $\widehat{L}_\tau^t$ referred to in Section 3.

Lemma E.1 (Smoothness constant)**.** *With u_{ij}^t, v_{ij}^t and d_{ij}^t as in* (8)*, put*

$$\widehat{L}_\tau^t \;\triangleq\; 2 \max_{1\le i<j\le C} d_{ij}^t \lambda_{\max}\big(\mathbf{S}_w^{ij}\big) + \frac{1}{\tau} \max_{1\le i<j\le C} \widehat{G}_{ij}^2, \qquad \widehat{G}_{ij} = 2d_{ij}^t \lambda_{\max}\big(\mathbf{S}_w^{ij}\big)\sqrt{m} + \frac{2}{v_{ij}^t}\big\|\mathbf{S}_b^{ij}\mathbf{T}^t\big\|_F. \tag{S31}$$

Then $\sup_{\mathbf{T}\in\mathcal{B}} \|\nabla^2 \widehat{F}_\tau(\mathbf{T} \mid \mathbf{T}^t)\|_{\mathrm{op}} \le \widehat{L}_\tau^t$, where $\mathcal{B} = \{\mathbf{T} : \|\mathbf{T}\|_F \le \sqrt{m}\}$ is a convex set that contains $\mathcal{S}$ and $\|\cdot\|_{\mathrm{op}}$ is the operator norm.

Proof. Write $\widehat{F}_\tau = \mathrm{lse}_\tau \circ \hat{\mathbf{h}}$, where $\hat{\mathbf{h}}(\mathbf{T})$ collects the K values $\hat{h}_{ij}(\mathbf{T} \mid \mathbf{T}^t)$. For a twice differentiable composition of this type, the second directional derivative along a direction $\mathbf{\Xi}$ with $\|\mathbf{\Xi}\|_F = 1$ is

$$\nabla^2 \widehat{F}_\tau[\mathbf{\Xi},\mathbf{\Xi}] = \underbrace{\sum_{ij} w_{ij}\, \nabla^2 \hat{h}_{ij}[\mathbf{\Xi},\mathbf{\Xi}]}_{\text{(I)}} + \underbrace{\frac{1}{\tau}\Big[\sum_{ij} w_{ij} \big\langle \nabla \hat{h}_{ij}, \mathbf{\Xi}\big\rangle^2 - \Big(\sum_{ij} w_{ij} \big\langle \nabla \hat{h}_{ij}, \mathbf{\Xi}\big\rangle \Big)^2 \Big]}_{\text{(II)}}, \tag{S32}$$

where $w_{ij} = [\mathrm{softmax}(\hat{\mathbf{h}}/\tau)]_{ij}$ are non-negative and sum to one. The expression (S32) is obtained by differentiating (S28) once more, the first term collecting the second derivatives of the inner functions and the second term being the derivative of the softmax weights themselves, whose Jacobian is $\frac{1}{\tau}(\mathrm{diag}(\mathbf{w}) - \mathbf{w}\mathbf{w}^T)$.

We bound the two terms separately. For (I), each $\nabla^2 \hat{h}_{ij}$ satisfies $\mathbf{0} \preccurlyeq \nabla^2 \hat{h}_{ij} \preccurlyeq 2d_{ij}^t \lambda_{\max}(\mathbf{S}_w^{ij})\mathbf{I}$ by (S26), and a convex combination of matrices bounded above by $c_{ij}\mathbf{I}$ is bounded above by $(\max_{ij} c_{ij})\mathbf{I}$, so

$$0 \;\le\; (\mathrm{I}) \;\le\; 2 \max_{1 \le i < j \le C} d_{ij}^t \lambda_{\max}\big(\mathbf{S}_w^{ij}\big). \tag{S33}$$

For (II), the bracket is exactly the variance of the scalar random variable that takes the value $\langle \nabla \hat{h}_{ij}, \mathbf{\Xi} \rangle$ with probability w_{ij}. A variance is non-negative, and it is bounded above by the second moment, which in turn is bounded by the largest squared value, so that

$$0 \;\le\; (\mathrm{II}) \;\le\; \frac{1}{\tau} \max_{ij} \big\langle \nabla \hat{h}_{ij}, \mathbf{\Xi} \big\rangle^2 \;\le\; \frac{1}{\tau} \max_{ij} \big\| \nabla \hat{h}_{ij} \big\|_F^2, \tag{S34}$$

the last step being the Cauchy–Schwarz inequality with $\|\mathbf{\Xi}\|_F = 1$. It remains to bound $\|\nabla \hat{h}_{ij}\|_F$ on $\mathcal{B} = \{\mathbf{T} : \|\mathbf{T}\|_F \le \sqrt{m}\}$. From (S26) and the triangle inequality,

$$\begin{aligned} \big\| \nabla \hat{h}_{ij}(\mathbf{T} \mid \mathbf{T}^t) \big\|_F \;&\le\; 2d_{ij}^t \big\| \mathbf{S}_w^{ij} \mathbf{T} \big\|_F + \frac{2}{v_{ij}^t} \big\| \mathbf{S}_b^{ij} \mathbf{T}^t \big\|_F \\ &\le\; 2d_{ij}^t \lambda_{\max}\big(\mathbf{S}_w^{ij}\big)\sqrt{m} + \frac{2}{v_{ij}^t} \big\| \mathbf{S}_b^{ij} \mathbf{T}^t \big\|_F \;=\; \widehat{G}_{ij}, \end{aligned} \tag{S35}$$

where we used $\|\mathbf{S}\mathbf{T}\|_F \le \|\mathbf{S}\|_2 \|\mathbf{T}\|_F \le \lambda_{\max}(\mathbf{S})\sqrt{m}$ for $\mathbf{S} \succcurlyeq \mathbf{0}$ and $\mathbf{T} \in \mathcal{B}$. Since both (I) and (II) are non-negative, the operator norm of the Hessian is bounded by the sum of the two bounds (S33) and (S34), which is exactly $\widehat{L}_\tau^t$ of (S31). ■

E.4 Proof of Lemma 3.3, step three, the reduction to an affine bound

Proof. Let $\mathbf{T}, \mathbf{T}^t \in \mathcal{S}$. Every point of the segment $[\mathbf{T}^t, \mathbf{T}]$ has Frobenius norm at most $\sqrt{m}$ by convexity of the norm, since both endpoints have norm exactly $\sqrt{m}$ on $\mathcal{S}$; hence the segment lies in $\mathcal{B}$ and the descent lemma applies on it [30],

$$F(\mathbf{T}) \;\le\; F(\mathbf{T}^t) + \mathrm{tr}\big(\nabla F(\mathbf{T}^t)^T (\mathbf{T} - \mathbf{T}^t)\big) + \frac{L}{2} \big\| \mathbf{T} - \mathbf{T}^t \big\|_F^2. \tag{S36}$$

On $\mathcal{S}$ we have $\|\mathbf{T}\|_F^2 = \mathrm{tr}(\mathbf{T}^T\mathbf{T}) = \mathrm{tr}(\mathbf{I}_m) = m$ and likewise $\|\mathbf{T}^t\|_F^2 = m$, so

$$\big\| \mathbf{T} - \mathbf{T}^t \big\|_F^2 = \|\mathbf{T}\|_F^2 - 2\,\mathrm{tr}\big(\mathbf{T}^T\mathbf{T}^t\big) + \|\mathbf{T}^t\|_F^2 = 2m - 2\,\mathrm{tr}\big(\mathbf{T}^T\mathbf{T}^t\big). \tag{S37}$$

Substituting (S37) into (S36) and separating the terms that depend on $\mathbf{T}$ from those that do not,

$$\begin{aligned} F(\mathbf{T}) &\le F(\mathbf{T}^t) - \mathrm{tr}\big(\nabla F(\mathbf{T}^t)^T \mathbf{T}^t\big) + Lm \;+\; \mathrm{tr}\big(\nabla F(\mathbf{T}^t)^T \mathbf{T}\big) - L\,\mathrm{tr}\big((\mathbf{T}^t)^T \mathbf{T}\big) \\ &= \underbrace{\Big[F(\mathbf{T}^t) - \mathrm{tr}\big(\nabla F(\mathbf{T}^t)^T \mathbf{T}^t\big) + Lm \Big]}_{c_t} \;-\; \mathrm{tr}\Big(\big[L\mathbf{T}^t - \nabla F(\mathbf{T}^t) \big]^T \mathbf{T} \Big), \end{aligned} \tag{S38}$$

which is the affine bound (7) of Lemma 3.3 with $\mathbf{B}_t = L\mathbf{T}^t - \nabla F(\mathbf{T}^t)$. The constant c_t does not depend on $\mathbf{T}$. Finally, at $\mathbf{T} = \mathbf{T}^t$ the right-hand side of (S38) equals

$$c_t - L\,\mathrm{tr}\big((\mathbf{T}^t)^T\mathbf{T}^t\big) + \mathrm{tr}\big(\nabla F(\mathbf{T}^t)^T\mathbf{T}^t\big) = c_t - Lm + \mathrm{tr}\big(\nabla F(\mathbf{T}^t)^T\mathbf{T}^t\big) = F(\mathbf{T}^t), \tag{S39}$$

so the bound is tight at the expansion point. ■

E.5 Proof of Theorem 3.4

Proof. Minimizing $-\mathrm{tr}(\mathbf{B}_t^T\mathbf{T})$ over $\mathcal{S}$ is the same as maximizing $\mathrm{tr}(\mathbf{B}_t^T\mathbf{T})$ over $\mathcal{S}$. By the Von Neumann trace inequality [27], for any two matrices of the same size, with $\sigma_k(\cdot)$ the k-th largest singular value,

$$\mathrm{tr}\big(\mathbf{B}_t^T\mathbf{T}\big) \;\le\; \sum_{k=1}^{m} \sigma_k(\mathbf{B}_t)\,\sigma_k(\mathbf{T}). \tag{S40}$$

If $\mathbf{T} \in \mathcal{S}$ then all its singular values equal one, so (S40) becomes

$$\operatorname{tr}\big(\mathbf{B}_t^T \mathbf{T}\big) \;\le\; \sum_{k=1}^{m} \sigma_k(\mathbf{B}_t) = \|\mathbf{B}_t\|_*, \tag{S41}$$

the nuclear norm of $\mathbf{B}_t$. This bound is attained by $\mathbf{T} = \mathbf{U}\mathbf{V}^T$, because with $\mathbf{B}_t = \mathbf{U}\mathbf{D}\mathbf{V}^T$ we have

$$\operatorname{tr}\big(\mathbf{B}_t^T \mathbf{U}\mathbf{V}^T\big) = \operatorname{tr}\big(\mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{U}\mathbf{V}^T\big) = \operatorname{tr}\big(\mathbf{V}\mathbf{D}\mathbf{V}^T\big) = \operatorname{tr}\big(\mathbf{D}\mathbf{V}^T\mathbf{V}\big) = \operatorname{tr}(\mathbf{D}) = \sum_{k=1}^{m} \sigma_k(\mathbf{B}_t), \tag{S42}$$

where we used $\mathbf{U}^T\mathbf{U} = \mathbf{I}_m$, the cyclic property of the trace, and $\mathbf{V}^T\mathbf{V} = \mathbf{I}_m$. Moreover $(\mathbf{U}\mathbf{V}^T)^T\mathbf{U}\mathbf{V}^T = \mathbf{V}\mathbf{U}^T\mathbf{U}\mathbf{V}^T = \mathbf{I}_m$, so the maximizer is feasible. If $\mathbf{B}_t$ has full column rank then $\mathbf{D} \succ \mathbf{0}$, the polar decomposition $\mathbf{B}_t = (\mathbf{U}\mathbf{V}^T)(\mathbf{V}\mathbf{D}\mathbf{V}^T)$ is unique, and

$$\mathbf{B}_t\big(\mathbf{B}_t^T\mathbf{B}_t\big)^{-1/2} = \mathbf{U}\mathbf{D}\mathbf{V}^T\big(\mathbf{V}\mathbf{D}^2\mathbf{V}^T\big)^{-1/2} = \mathbf{U}\mathbf{D}\mathbf{V}^T\mathbf{V}\mathbf{D}^{-1}\mathbf{V}^T = \mathbf{U}\mathbf{V}^T, \tag{S43}$$

so the two expressions for the update coincide. If instead $\mathbf{B}_t$ is rank deficient, the minimizer is no longer unique, but $\mathbf{U}\mathbf{V}^T$ taken from any thin singular value decomposition still attains the bound (S41) and is therefore optimal, and since that is the form the algorithm uses no separate case is needed. ■

F Per-iteration cost and convergence

Two structural facts keep the cost of the gradient (12) low. First, every numerator is rank one, $\mathbf{S}_b^{ij}\mathbf{T}^t = \boldsymbol{\delta}_{ij}(\boldsymbol{\delta}_{ij}^T\mathbf{T}^t)$, so all K numerator terms together cost $\mathcal{O}(Kdm)$ and no $d \times d$ product is ever formed. Second, $\mathbf{S}_w^{ij} = \boldsymbol{\Sigma}_i^w + \boldsymbol{\Sigma}_j^w$, so the weighted denominator sum in (12) regroups by class instead of by pair,

$$\sum_{1\le i<j\le C} w_{ij}^t d_{ij}^t\, \mathbf{S}_w^{ij}\mathbf{T}^t = \sum_{k=1}^{C} a_k^t\, \boldsymbol{\Sigma}_k^w \mathbf{T}^t, \qquad a_k^t = \sum_{j\neq k} w_{kj}^t d_{kj}^t, \tag{S44}$$

because each unordered pair contributes its weight once to i and once to j, as Appendix E.2 derives. The identity (S44) replaces K matrix products by C. With the $\mathcal{O}(m^3)$ thin singular value decomposition, one iteration of Algorithm S1 therefore costs $\mathcal{O}(Cd^2m + Kdm + m^3)$, so there is no quadratic program, no semidefinite program and no outer loop on the ratio value at any point, which is the accounting Table S1 of Appendix I sets against the other solvers.

Theorem F.1 (Convergence)**.** *Fix $\tau > 0$ and let $\{\mathbf{T}^t\}$ be generated by Algorithm S1. Then* (i) *$\{F_\tau(\mathbf{T}^t)\}$ is non-increasing and convergent;* (ii) *$\|\mathbf{T}^{t+1} - \mathbf{T}^t\|_F \to 0$; and* (iii) *every accumulation point of $\{\mathbf{T}^t\}$ is a stationary point of $\min_{\mathbf{T}\in\mathcal{S}} F_\tau(\mathbf{T})$.*

Proof. Let $H_t(\mathbf{T}) = c_t - \operatorname{tr}(\mathbf{B}_t^T\mathbf{T})$ denote the linear surrogate built at iteration t, with $\mathbf{B}_t = \widehat{L}_\tau^t\mathbf{T}^t - \nabla F_\tau(\mathbf{T}^t)$.

(i) Monotone descent. On $\mathcal{S}$ we have the chain of majorizations

$$F_\tau(\mathbf{T}) \;\overset{(a)}{\le}\; \widehat{F}_\tau(\mathbf{T} \mid \mathbf{T}^t) \;\overset{(b)}{\le}\; H_t(\mathbf{T}), \tag{S45}$$

with equality throughout at $\mathbf{T} = \mathbf{T}^t$. Inequality (a) holds because $\hat{h}_{ij}(\cdot \mid \mathbf{T}^t) \ge h_{ij}$ on $\mathcal{S}$ by Lemma 3.3 and because lse_τ is non-decreasing in each coordinate, so replacing every argument by a larger one can only increase the value. Inequality (b) is the affine bound (7) of Lemma 3.3 applied to $\widehat{F}_\tau(\cdot \mid \mathbf{T}^t)$ with $L = \widehat{L}_\tau^t$, which is a valid curvature bound by Lemma E.1. Since $\mathbf{T}^{t+1}$ minimizes H_t exactly over $\mathcal{S}$ by Theorem 3.4, the generic MM argument (S5) gives

$$F_\tau(\mathbf{T}^{t+1}) \;\le\; H_t(\mathbf{T}^{t+1}) \;\le\; H_t(\mathbf{T}^t) = F_\tau(\mathbf{T}^t). \tag{S46}$$

The manifold $\mathcal{S}$ is compact and F_τ is continuous on it, so F_τ is bounded below there, and a non-increasing sequence bounded below converges.

(ii) Vanishing step length. Instead of discarding the quadratic term of the descent lemma we keep it. Repeating the derivation (S38) without the substitution (S37) in the direction of the minimizer gives the sufficient-decrease estimate

$$F_\tau(\mathbf{T}^{t+1}) \;\le\; F_\tau(\mathbf{T}^t) - \frac{\widehat{L}_\tau^{\,t}}{2}\big\|\mathbf{T}^{t+1} - \mathbf{T}^t\big\|_F^2, \tag{S47}$$

which is the standard proximal-MM inequality [9, 37]. We now argue that the constants $\widehat{L}_\tau^{\,t}$ are bounded above and below by strictly positive constants that do not depend on t. Indeed, for $\mathbf{T}^t \in \mathcal{S}$,

$$v_{ij}^t = \operatorname{tr}\big((\mathbf{T}^t)^T \mathbf{S}_w^{ij} \mathbf{T}^t\big) \;\ge\; m\,\lambda_{\min}\big(\mathbf{S}_w^{ij}\big) > 0, \qquad u_{ij}^t = \big\|(\mathbf{T}^t)^T \boldsymbol{\delta}_{ij}\big\|_2^2 \;\le\; \|\boldsymbol{\delta}_{ij}\|_2^2, \tag{S48}$$

so $d_{ij}^t = u_{ij}^t/(v_{ij}^t)^2$ lies in a fixed bounded interval, and the same is true of $\widehat{G}_{ij}$ and hence of $\widehat{L}_\tau^{\,t}$ by (S31). Writing $0 < L_{\min} \le \widehat{L}_\tau^{\,t} \le L_{\max} < \infty$ and summing (S47) over $t = 0, \dots, N-1$ telescopes to

$$\frac{L_{\min}}{2} \sum_{t=0}^{N-1} \big\|\mathbf{T}^{t+1} - \mathbf{T}^t\big\|_F^2 \;\le\; F_\tau(\mathbf{T}^0) - F_\tau(\mathbf{T}^N) \;\le\; F_\tau(\mathbf{T}^0) - \inf_{\mathcal{S}} F_\tau < \infty. \tag{S49}$$

Letting $N \to \infty$ shows that $\sum_t \|\mathbf{T}^{t+1} - \mathbf{T}^t\|_F^2 < \infty$, and therefore $\|\mathbf{T}^{t+1} - \mathbf{T}^t\|_F \to 0$.

(iii) Stationarity of the accumulation points. The surrogates have first-order contact with F_τ at the expansion point. This holds because the quadratic transform of Lemma 3.3 is a tangent-type majorizer, so $\nabla \hat{h}_{ij}(\mathbf{T}^t \mid \mathbf{T}^t) = \nabla h_{ij}(\mathbf{T}^t)$ as verified in (S27), and because both the log-sum-exp composition and the linearization of Lemma 3.3 preserve gradients at the point of tangency. Consequently

$$\nabla \widehat{F}_\tau(\mathbf{T}^t \mid \mathbf{T}^t) = \nabla F_\tau(\mathbf{T}^t), \qquad \mathbf{B}_t = \widehat{L}_\tau^{\,t}\mathbf{T}^t - \nabla F_\tau(\mathbf{T}^t). \tag{S50}$$

Since $\mathcal{S}$ is compact, the sequence $\{\mathbf{T}^t\}$ has accumulation points. Let $\mathbf{T}^{t_k} \to \bar{\mathbf{T}}$. By part (ii), $\mathbf{T}^{t_k+1} \to \bar{\mathbf{T}}$ as well. Each $\mathbf{T}^{t_k+1}$ satisfies the first-order optimality condition of $\min_{\mathbf{T}\in\mathcal{S}} H_{t_k}(\mathbf{T})$, that is, the Riemannian gradient of H_{t_k} vanishes at $\mathbf{T}^{t_k+1}$,

$$\operatorname{grad} H_{t_k}(\mathbf{T}^{t_k+1}) = \mathbf{P}_{\mathbf{T}^{t_k+1}}\big(-\mathbf{B}_{t_k}\big) = \mathbf{0}, \qquad \mathbf{P}_{\mathbf{T}}(\mathbf{G}) = \mathbf{G} - \mathbf{T}\,\mathrm{sym}\big(\mathbf{T}^T\mathbf{G}\big), \tag{S51}$$

where $\mathbf{P}_{\mathbf{T}}$ is the projection onto the tangent space of $\mathcal{S}$ at $\mathbf{T}$ and $\mathrm{sym}(\mathbf{A}) = \frac{1}{2}(\mathbf{A} + \mathbf{A}^T)$ [2]. Substituting $\mathbf{B}_{t_k}$ from (S50) and using $\mathbf{P}_{\mathbf{T}}(\mathbf{T}) = \mathbf{0}$, the condition becomes $\mathbf{P}_{\mathbf{T}^{t_k+1}}\big(\nabla F_\tau(\mathbf{T}^{t_k})\big) = \mathbf{0}$ in the limit of the two arguments coinciding. Passing to the limit along the subsequence and using the continuity of ∇F_τ and of $\mathbf{P}_{(\cdot)}(\cdot)$, we obtain

$$\operatorname{grad} F_\tau(\bar{\mathbf{T}}) = \nabla F_\tau(\bar{\mathbf{T}}) - \bar{\mathbf{T}}\,\mathrm{sym}\big(\bar{\mathbf{T}}^T \nabla F_\tau(\bar{\mathbf{T}})\big) = \mathbf{0}, \tag{S52}$$

which is exactly the KKT condition of $\min_{\mathbf{T}\in\mathcal{S}} F_\tau(\mathbf{T})$. ■

G Proofs of the privacy results

G.1 Proof of Lemma 4.1

Proof. Let $\mathbf{X}$ and $\mathbf{X}'$ differ in one labelled record, so that the pair $(\mathbf{x}, a)$ in $\mathbf{X}$ is replaced by $(\mathbf{x}', b)$ in $\mathbf{X}'$, with $\|\mathbf{x}\|_2 \le R$ and $\|\mathbf{x}'\|_2 \le R$ by Assumption 1. We compute the squared ℓ_2 change of the concatenated query (13) in the two possible cases.

Case 1, the record changes class, $a \neq b$. The counts change in two coordinates, n_a decreasing by one and n_b increasing by one, so the count block contributes $1^2 + 1^2 = 2$. The sum block loses $\mathbf{x}$ from class a and gains $\mathbf{x}'$ in class b, and since these are different blocks the two contributions add up to

$$\|-\mathbf{x}\|_2^2 + \|\mathbf{x}'\|_2^2 = \|\mathbf{x}\|_2^2 + \|\mathbf{x}'\|_2^2 \;\le\; 2R^2. \tag{S53}$$

The second-moment block likewise loses $\mathbf{x}\mathbf{x}^T$ from class a and gains $\mathbf{x}'\mathbf{x}'^T$ in class b, and for a rank-one matrix $\|\mathbf{x}\mathbf{x}^T\|_F^2 = \operatorname{tr}(\mathbf{x}\mathbf{x}^T\mathbf{x}\mathbf{x}^T) = \|\mathbf{x}\|_2^4$, so this block contributes

$$\|\mathbf{x}\mathbf{x}^T\|_F^2 + \|\mathbf{x}'\mathbf{x}'^T\|_F^2 = \|\mathbf{x}\|_2^4 + \|\mathbf{x}'\|_2^4 \;\le\; 2R^4. \tag{S54}$$

Adding the three blocks, the squared change is at most $2 + 2R^2 + 2R^4$, so

$$\big\|q(\mathbf{X}) - q(\mathbf{X}')\big\|_2 \;\le\; \sqrt{2\big(1 + R^2 + R^4\big)}. \tag{S55}$$

Case 2, the record stays in its class, $a = b$. The counts do not change at all, so the count block contributes zero. The sum block changes by $\mathbf{x}' - \mathbf{x}$, and by the triangle inequality $\|\mathbf{x}' - \mathbf{x}\|_2 \le \|\mathbf{x}'\|_2 + \|\mathbf{x}\|_2 \le 2R$, so this block contributes at most $4R^2$. The second-moment block changes by $\mathbf{x}'\mathbf{x}'^T - \mathbf{x}\mathbf{x}^T$, and again by the triangle inequality

$$\left\|\mathbf{x}'\mathbf{x}'^T - \mathbf{x}\mathbf{x}^T\right\|_F \;\le\; \|\mathbf{x}'\mathbf{x}'^T\|_F + \|\mathbf{x}\mathbf{x}^T\|_F = \|\mathbf{x}'\|_2^2 + \|\mathbf{x}\|_2^2 \;\le\; 2R^2, \tag{S56}$$

so this block contributes at most $4R^4$. Adding,

$$\left\|q(\mathbf{X}) - q(\mathbf{X}')\right\|_2 \;\le\; \sqrt{4R^2 + 4R^4} = 2R\sqrt{1 + R^2}. \tag{S57}$$

The sensitivity is the maximum over all neighbouring pairs, hence the maximum of (S55) and (S57), which is (14). ■

G.2 Proof of Theorem 4.2

Proof. The release of Section 4.1 is a single Gaussian mechanism applied to the single query q of (13). For the symmetric second-moment blocks, drawing independent Gaussian entries on and above the diagonal and mirroring them below is the Gaussian mechanism applied to the upper-triangular vectorization of $\mathbf{M}_k$; the ℓ_2-sensitivity of that vectorization is dominated by the Frobenius bound of Lemma 4.1, which counts the off-diagonal entries twice, so the calibration is conservative. This is exactly the construction of Analyze Gauss [16]. With σ chosen as in (S12), the release is therefore (ε, δ)-differentially private [4, 14].

Everything that happens afterwards is a deterministic function of the released quantities and of public parameters. In particular, the clipped counts $\tilde{n}_k$, the private means $\hat{\bar{\mathbf{x}}}_k$, the projected scatters $\widehat{\mathbf{\Sigma}}_k^w$, the ridge-regularized $\widehat{\mathbf{S}}_w^{ij}$, the temperature schedule, the smoothness constants $\widehat{L}_\tau^t$, the softmax weights and every iterate $\mathbf{T}^t$ are computed from that release without any further access to $\mathbf{X}$. By post-processing invariance [14], the composite map is (ε, δ)-differentially private with the same parameters. Since the number of iterations, the number of restarts, the annealing schedule and the values of τ, m, λ_{reg} and the tolerances enter only in this post-processing stage, the guarantee does not depend on them. ■

H Proofs of the utility results

H.1 Proof of Lemma 4.3

Proof. We first control the three noise blocks and then propagate the errors through the reconstruction (15).

Step 1, the noise events. For a symmetric Gaussian matrix with independent $\mathcal{N}(0, \sigma^2)$ entries on and above the diagonal, the standard bound for the spectral norm [44] gives, for each k,

$$\|\mathbf{E}_k\|_2 \;\le\; \sigma\big(2\sqrt{d} + c_1\sqrt{\log(C/\gamma)}\big) \tag{S58}$$

with probability at least $1 - \gamma/(3C)$; a union bound over the C classes gives the statement simultaneously for all k with probability at least $1 - \gamma/3$. Similarly, the Gaussian vector tail gives $\|\boldsymbol{\xi}_k\|_2 \le \sigma(\sqrt{d} + c_2\sqrt{\log(C/\gamma)})$ and the scalar Gaussian tail gives $|\nu_k| \le c_3\sigma\sqrt{\log(C/\gamma)}$, each simultaneously for all k with probability at least $1 - \gamma/3$. Intersecting the three events gives an event $\mathcal{E}$ of probability at least $1 - \gamma$ on which all three bounds hold. The rest of the proof is deterministic given $\mathcal{E}$.

Step 2, the counts and the means. If $n_{\min} \gtrsim \sigma\sqrt{\log(C/\gamma)}$ then $|\nu_k| \le n_k/2$, so $\tilde{n}_k = \max\{\hat{n}_k, 1\}$ satisfies $n_k/2 \le \tilde{n}_k \le 3n_k/2$. Writing

$$\hat{\bar{\mathbf{x}}}_k - \bar{\mathbf{x}}_k = \frac{\mathbf{s}_k + \boldsymbol{\xi}_k}{\tilde{n}_k} - \frac{\mathbf{s}_k}{n_k} = \frac{\boldsymbol{\xi}_k}{\tilde{n}_k} + \mathbf{s}_k\Big(\frac{1}{\tilde{n}_k} - \frac{1}{n_k}\Big) = \frac{\boldsymbol{\xi}_k}{\tilde{n}_k} - \bar{\mathbf{x}}_k\frac{\tilde{n}_k - n_k}{\tilde{n}_k}, \tag{S59}$$

and using $\|\bar{\mathbf{x}}_k\|_2 \le R$, $|\tilde{n}_k - n_k| \le |\nu_k|$ and $\tilde{n}_k \ge n_k/2 \ge n_{\min}/2$, we get

$$\left\|\hat{\bar{\mathbf{x}}}_k - \bar{\mathbf{x}}_k\right\|_2 \;\lesssim\; \frac{\|\boldsymbol{\xi}_k\|_2 + |\nu_k|R}{n_{\min}} \;\lesssim\; \frac{\sigma\big(\sqrt{d} + \sqrt{\log(C/\gamma)}\big)(1 + R)}{n_{\min}}. \tag{S60}$$

Step 3, the within-class scatters. Subtracting $\mathbf{\Sigma}_k^w = \mathbf{M}_k - n_k \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^T$ from the pre-projection estimate,

$$\big(\widehat{\mathbf{M}}_k - \tilde{n}_k \hat{\bar{\mathbf{x}}}_k \hat{\bar{\mathbf{x}}}_k^T\big) - \mathbf{\Sigma}_k^w = \mathbf{E}_k - \big(\tilde{n}_k \hat{\bar{\mathbf{x}}}_k \hat{\bar{\mathbf{x}}}_k^T - n_k \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^T\big), \tag{S61}$$

and expanding the second bracket as $\tilde{n}_k(\hat{\bar{\mathbf{x}}}_k - \bar{\mathbf{x}}_k)\hat{\bar{\mathbf{x}}}_k^T + \tilde{n}_k \bar{\mathbf{x}}_k(\hat{\bar{\mathbf{x}}}_k - \bar{\mathbf{x}}_k)^T + (\tilde{n}_k - n_k)\bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^T$, the triangle inequality together with (S60) and $\tilde{n}_k \lesssim n_k$ gives

$$\big\|\widehat{\mathbf{M}}_k - \tilde{n}_k \hat{\bar{\mathbf{x}}}_k \hat{\bar{\mathbf{x}}}_k^T - \mathbf{\Sigma}_k^w\big\|_2 \;\lesssim\; \sigma\big(\sqrt{d} + \sqrt{\log(C/\gamma)}\big)\Big(1 + R + \frac{R^2}{n_{\min}}\Big). \tag{S62}$$

The projection onto the positive semidefinite cone can at most double this error, because for any symmetric $\mathbf{A}$ and any $\mathbf{B} \succcurlyeq \mathbf{0}$,

$$\big\|\Pi_{\succcurlyeq 0}(\mathbf{A}) - \mathbf{B}\big\|_2 \le \big\|\Pi_{\succcurlyeq 0}(\mathbf{A}) - \mathbf{A}\big\|_2 + \|\mathbf{A} - \mathbf{B}\|_2 \le 2\|\mathbf{A} - \mathbf{B}\|_2, \tag{S63}$$

the last step holding because $\Pi_{\succcurlyeq 0}$ is the nearest positive semidefinite matrix in spectral norm and $\mathbf{B}$ is a feasible competitor. This gives the stated bound on e_w.

Step 4, the between-class scatters. These are rank one, $\mathbf{S}_b^{ij} = \boldsymbol{\delta}_{ij}\boldsymbol{\delta}_{ij}^T$ and $\widehat{\mathbf{S}}_b^{ij} = \hat{\boldsymbol{\delta}}_{ij}\hat{\boldsymbol{\delta}}_{ij}^T$. Using the identity $\mathbf{a}\mathbf{a}^T - \mathbf{b}\mathbf{b}^T = \mathbf{a}(\mathbf{a}-\mathbf{b})^T + (\mathbf{a}-\mathbf{b})\mathbf{b}^T$,

$$\big\|\widehat{\mathbf{S}}_b^{ij} - \mathbf{S}_b^{ij}\big\|_2 \;\le\; \big(\|\boldsymbol{\delta}_{ij}\|_2 + \|\hat{\boldsymbol{\delta}}_{ij}\|_2\big)\,\big\|\hat{\boldsymbol{\delta}}_{ij} - \boldsymbol{\delta}_{ij}\big\|_2, \tag{S64}$$

and $\|\hat{\boldsymbol{\delta}}_{ij} - \boldsymbol{\delta}_{ij}\|_2 \le \|\hat{\bar{\mathbf{x}}}_i - \bar{\mathbf{x}}_i\|_2 + \|\hat{\bar{\mathbf{x}}}_j - \bar{\mathbf{x}}_j\|_2$ is bounded by twice (S60). Since $\|\boldsymbol{\delta}_{ij}\|_2 \le 2R$, this gives the stated bound on e_b. ■

H.2 Proof of Lemma 4.4

Proof. Fix $\mathbf{T} \in \mathcal{S}$ and a pair (i, j), and abbreviate $u = \operatorname{tr}(\mathbf{T}^T \mathbf{S}_b^{ij} \mathbf{T})$, $v = \operatorname{tr}(\mathbf{T}^T \mathbf{S}_w^{ij} \mathbf{T})$, and likewise $\hat{u}, \hat{v}$ for the private quantities. For any symmetric $\mathbf{\Delta}$ and any $\mathbf{T} \in \mathcal{S}$,

$$\big|\operatorname{tr}(\mathbf{T}^T \mathbf{\Delta} \mathbf{T})\big| = \Big|\sum_{k=1}^{m} \mathbf{t}_k^T \mathbf{\Delta} \mathbf{t}_k\Big| \;\le\; \sum_{k=1}^{m} \|\mathbf{\Delta}\|_2 \|\mathbf{t}_k\|_2^2 = m\|\mathbf{\Delta}\|_2, \tag{S65}$$

where $\mathbf{t}_k$ are the columns of $\mathbf{T}$, which are unit vectors. Applying (S65) to $\mathbf{\Delta} = \widehat{\mathbf{S}}_b^{ij} - \mathbf{S}_b^{ij}$ and to $\mathbf{\Delta} = \widehat{\mathbf{S}}_w^{ij} - \mathbf{S}_w^{ij}$ gives $|\hat{u} - u| \le m e_b$ and $|\hat{v} - v| \le m e_w'$. Moreover $\hat{v} \ge m\lambda_{\mathrm{reg}}$ by assumption, and $u/v = r_{ij}(\mathbf{T}) \le \kappa$ because, for any $\mathbf{T} \in \mathcal{S}$,

$$\frac{\operatorname{tr}(\mathbf{T}^T \boldsymbol{\delta}_{ij}\boldsymbol{\delta}_{ij}^T \mathbf{T})}{\operatorname{tr}(\mathbf{T}^T \mathbf{S}_w^{ij} \mathbf{T})} = \frac{\|\mathbf{T}^T \boldsymbol{\delta}_{ij}\|_2^2}{\operatorname{tr}(\mathbf{T}^T \mathbf{S}_w^{ij} \mathbf{T})} \;\le\; \max_{\mathbf{T} \in \mathcal{S}} \frac{\|\mathbf{T}^T \boldsymbol{\delta}_{ij}\|_2^2}{\operatorname{tr}(\mathbf{T}^T \mathbf{S}_w^{ij} \mathbf{T})} \;\le\; \boldsymbol{\delta}_{ij}^T \big(\mathbf{S}_w^{ij}\big)^{-1} \boldsymbol{\delta}_{ij} \;\le\; \kappa. \tag{S66}$$

Now decompose the difference of the two ratios by adding and subtracting $u/\hat{v}$, which gives

$$\Big|\frac{\hat{u}}{\hat{v}} - \frac{u}{v}\Big| = \Big|\frac{\hat{u} - u}{\hat{v}} + u\Big(\frac{1}{\hat{v}} - \frac{1}{v}\Big)\Big| \;\le\; \frac{|\hat{u} - u|}{\hat{v}} + \frac{u}{v} \cdot \frac{|v - \hat{v}|}{\hat{v}} \;\le\; \frac{m e_b}{m\lambda_{\mathrm{reg}}} + \kappa\,\frac{m e_w'}{m\lambda_{\mathrm{reg}}} = \frac{e_b + \kappa e_w'}{\lambda_{\mathrm{reg}}}. \tag{S67}$$

The right-hand side of (S67) does not depend on $\mathbf{T}$ or on (i, j), so the bound holds uniformly, which is (16). ■

H.3 Proof of Theorem 4.5

Proof. Condition on the event of Lemma 4.3, which has probability at least $1 - \gamma$, and let Ξ be as in Lemma 4.4, where the effective within-class error is $e_w' = 2e_w + \lambda_{\mathrm{reg}}$ so as to absorb both the doubling caused by the positive semidefinite projection and the deterministic shift caused by the ridge in (15), while $\widehat{\mathbf{S}}_w^{ij} \succcurlyeq \lambda_{\mathrm{reg}}\mathbf{I}$ holds by construction. Write $\hat{r}_{ij}$ for the private ratios and $\widehat{\mathrm{OPT}} = \max_{\mathbf{T} \in \mathcal{S}} \min_{ij} \hat{r}_{ij}(\mathbf{T})$.

Step 1, the private optimum is not much worse than the true one. Let $\mathbf{T}^\star$ attain OPT. By Lemma 4.4,

$$\widehat{\mathrm{OPT}} \;\ge\; \min_{ij} \hat{r}_{ij}(\mathbf{T}^\star) \;\ge\; \min_{ij} r_{ij}(\mathbf{T}^\star) - \Xi \;=\; \mathrm{OPT} - \Xi. \tag{S68}$$

Step 2, the smoothing costs $\tau \log K$ on the private problem. Proposition 3.2, applied verbatim to the private problem, gives for the global minimizer $\widehat{\mathbf{T}}$ of the private smoothed objective

$$\min_{ij} \hat{r}_{ij}(\widehat{\mathbf{T}}) \;\geq\; \widehat{\mathrm{OPT}} - \tau \log K. \tag{S69}$$

Step 3, transfer back to the true ratios. One more application of Lemma 4.4, this time at the point $\widehat{\mathbf{T}}$, gives

$$\min_{ij} r_{ij}(\widehat{\mathbf{T}}) \;\geq\; \min_{ij} \hat{r}_{ij}(\widehat{\mathbf{T}}) - \Xi. \tag{S70}$$

Chaining (S70), (S69) and (S68) in that order,

$$\min_{ij} r_{ij}(\widehat{\mathbf{T}}) \;\geq\; \widehat{\mathrm{OPT}} - \tau \log K - \Xi \;\geq\; \mathrm{OPT} - 2\Xi - \tau \log K, \tag{S71}$$

which is (17). The two occurrences of Ξ come from the two directions of the transfer, once to move to the private problem and once to come back, and the term $\tau \log K$ is the smoothing gap; the two are added and never multiplied. The stated order of magnitude follows by substituting e_b and e_w from Lemma 4.3 into $\Xi = (e_b + \kappa e'_w)/\lambda_{\mathrm{reg}}$ and then $\sigma = \Theta\big(R^2\sqrt{\log(1/\delta)}/\varepsilon\big)$, which is the calibration (S12) with the sensitivity (14) for $R \geq 1$. ■

H.4 Proof of Proposition 4.6

Proof. Let $\mathbf{y} \mapsto \mathrm{softmax}(\mathbf{y})$ with $\mathbf{w} = \mathrm{softmax}(\mathbf{y})$. Its Jacobian is the classical expression

$$\mathbf{J}(\mathbf{y}) = \mathrm{diag}(\mathbf{w}) - \mathbf{w}\mathbf{w}^T, \tag{S72}$$

obtained by differentiating $w_k = e^{y_k}/\sum_l e^{y_l}$ with respect to y_p, which gives $\partial w_k/\partial y_p = w_k(\delta_{kp} - w_p)$. We bound the operator norm of $\mathbf{J}$ from ℓ_∞ to ℓ_1. For any $\mathbf{z}$,

$$\big\|\mathbf{J}\mathbf{z}\big\|_1 = \sum_k \big|w_k z_k - w_k\langle \mathbf{w}, \mathbf{z}\rangle\big| = \sum_k w_k\big|z_k - \langle \mathbf{w}, \mathbf{z}\rangle\big| \;\leq\; \sum_k w_k\big(|z_k| + |\langle \mathbf{w}, \mathbf{z}\rangle|\big) \;\leq\; 2\|\mathbf{z}\|_\infty, \tag{S73}$$

where we used $w_k \geq 0$, $\sum_k w_k = 1$, and $|\langle \mathbf{w}, \mathbf{z}\rangle| \leq \|\mathbf{z}\|_\infty$. The mean value theorem along the segment joining $-\mathbf{r}/\tau$ and $-\hat{\mathbf{r}}/\tau$ then gives

$$\big\|\mathbf{w}(\hat{\mathbf{r}}) - \mathbf{w}(\mathbf{r})\big\|_1 \;\leq\; \sup\big\|\mathbf{J}\big\|_{\infty\to 1} \cdot \Big\|\frac{\hat{\mathbf{r}} - \mathbf{r}}{\tau}\Big\|_\infty \;\leq\; \frac{2\|\hat{\mathbf{r}} - \mathbf{r}\|_\infty}{\tau} \;\leq\; \frac{2\Xi}{\tau}, \tag{S74}$$

which is the bound of Proposition 4.6. Setting the right-hand side equal to η gives the calibration rule $\tau \geq 2\Xi/\eta$.

For the balanced temperature, note that a perturbation of the weight profile of total-variation size $2\Xi/\tau$ moves the value of the weighted objective by at most that amount multiplied by the range of the quantities being weighted, and the ratios lie in $[0, \kappa]$ by Lemma 4.4; so the objective-scale wobble caused by weight instability is at most $2\Xi\kappa/\tau$. The smoothing gap, by Proposition 3.2, is $\tau \log K$. The first decreases in τ and the second increases in τ, so the sum is minimized where the two are equal, that is, where

$$\frac{2\Xi\kappa}{\tau} = \tau \log K \qquad\Longleftrightarrow\qquad \tau^2 = \frac{2\Xi\kappa}{\log K} \qquad\Longleftrightarrow\qquad \tau^\star = \sqrt{\frac{2\Xi\kappa}{\log K}}, \tag{S75}$$

which is the temperature of Proposition 4.6. Since Ξ grows as ε decreases, $\tau^\star$ increases as the privacy budget tightens. ■

I Complexity comparison

Table S1 sets the structure and the per-iteration cost of every solver that can be applied to the PWCRA problem beside the way each of them would have to pay for privacy, and two things follow from reading down its columns. The first is that the inner-solver column is what separates the costs, since the two-level scheme pays a semidefinite program at every iteration, whereas a closed-form inner

step leaves ESRA with a cost linear in the number of classes and cubic only in the target dimension, which is the small one. The second is that the privacy column splits the table in two, because a solver that reads the data at every step has to buy each step out of the budget, whereas releasing the statistics once puts the whole optimization on the far side of post-processing, and that follows from where the data are read, independently of how fast the solver runs. In DP-SGD-style schemes [1, 5] every additional step consumes budget, whereas DP-ESRA lets the analyst anneal the temperature and iterate to convergence at zero marginal privacy cost, the only added cost being the one-time $\mathcal{O}(Cd^2)$ mechanism and $\mathcal{O}(Cd^3)$ reconstruction of Section 4.1.

Table S1: Per-iteration complexity, algorithmic structure and privacy cost of the solvers applicable to the PWCRA problem.

Method	Structure	Inner solver	Per-iteration cost	Privacy cost
DP-SGD-PWCRA (2016) [1]	single loop	noisy gradient	$\mathcal{O}(Cd^2m)$	grows with the iterations
WDDR/MMRA (2024) [45]	two-level	SDP relaxation (in the sense of losing tightness)	$\mathcal{O}((d+m)^{4.5})$	—
ESRA (proposed)	single loop	none (closed form)	$\mathcal{O}(Cd^2m + Kdm + m^3)$	—
DP-ESRA (proposed)	single loop	none (closed form)	$\mathcal{O}(Cd^2m + Kdm + m^3)$	one-shot, iteration-free

J Related work and the compared methods

This appendix reviews the literature the paper builds on and states what each compared method computes and where it is limited. Every one of them was run at the configuration prescribed in its own paper, with its own hyperparameters, so that the comparison is fair.

LDA (1936). Linear discriminant analysis maximizes the trace ratio of one between-class scatter matrix to one within-class scatter matrix [18, 36]. Both matrices aggregate every class, so the criterion is an average over class pairs and is driven by the pairs that are already far apart. When the target dimension falls below the number of classes minus one, the two classes of the closest pair can therefore be projected on top of each other, and the between-class matrix has rank at most that same number, which caps how many informative directions the method can return at all.

HLDA (2004). Heteroscedastic LDA replaces the equal-covariance assumption by a criterion that also accounts for the differences between the class covariances [26]. It nevertheless aggregates the classes into a single criterion, so the worst-separated pair is still unprotected, and the resulting objective is handled by a generic numerical search for which no descent guarantee is available.

WHMMDA (2018). Whitened heteroscedastic max–min distance analysis enlarges the smallest pairwise Chernoff distance between classes in the whitened space [41, 42]. It maximizes a distance and not a ratio, so a pair is never measured against its own within-class scatter, and two classes that have been driven apart can still overlap when their internal dispersion is large. The subproblem it solves at every iteration is moreover a semidefinite program, so that one iteration costs $\mathcal{O}((d+m)^{4.5})$ operations.

WDA-eig (2020). Wasserstein discriminant analysis with an eigenvalue decomposition optimizes a transport-based separation measure [25]. The transport plan it rests on is entropically regularized, so the quantity actually optimized is a surrogate whose regularization weight has to be chosen, and the measure is again an aggregate over the classes, which leaves the worst-separated pair unprotected.

RLDA-$\ell_{2,1}$ (2021). Robust LDA with the $\ell_{2,1}$ norm replaces the squared Frobenius norm of the LDA criterion by a norm that limits the influence of outlying samples [33]. That robustness is bought with a regularization weight that has to be tuned, and because the criterion remains an aggregate it is sensitive to a small number of samples per class, where the robust scatter estimate itself becomes unreliable.

WDDR/MMRA (2024). Worst-case discriminative dimension reduction maximizes the smallest pairwise ratio, which is the PWCRA criterion [45]. It recasts the fractional program as a sequence of quadratic subproblems and applies to each of them a semidefinite relaxation (in the sense of losing tightness), which replaces the $d \times m$ projection by a $d \times d$ matrix variable. Once the resulting semidefinite program is solved, WDDR eigen-decomposes its solution and keeps the m leading eigenvectors, so as to return to a projection with orthonormal columns. Lifting the variable from $d \times m$ to $d \times d$ increases the dimension of the problem, so that each semidefinite solve costs $\mathcal{O}((d+m)^{4.5})$ operations, and since tightness is not guaranteed, the recovered projection may be suboptimal and may undermine the separability it is meant to protect. The procedure that solves the semidefinite program moreover has at least two hyperparameters that must be set by hand, the stopping tolerance aside.

QP-MMDA (2025) and Sparse-MMDA (2025). Both enlarge the smallest pairwise Chernoff distance under the minorization–maximization principle, the second adding a sparsity penalty on the projection [35]. They maximize a distance and not a ratio, so neither of them measures a pair against the dispersion of that same pair, both solve a quadratic program over the simplex at every iteration, and the sparse variant carries a penalty weight that has to be selected on held-out data.

K FULL EXPERIMENTAL PROTOCOL

This appendix records the protocol details that Section 5 summarizes.

Table S2: Size of each benchmark as distributed, before preprocessing.

Dataset	Samp.	Cls.	Dim.
COIL-20 [29]	1,440	20	16,384
CMU-PIE [40]	41,368	68	311,040
Caltech101 [17]	9,146	101	variable[†]
Stable ImageNet [22]	1,000	10	786,432

[†]Caltech101 images are not of a fixed size, roughly 300×200 pixels; see the dataset descriptions below.

Datasets. The four benchmarks span object recognition, face recognition and large-scale image classification, and they are chosen across a wide spectrum of feature-space sizes. **COIL-20** [29] contains $1{,}440$ grayscale images of 20 objects, 72 views per object taken at 5-degree intervals on a turntable, distributed at 128×128 resolution, so the raw pixel space is $16{,}384$-dimensional; we work from the 16×16 downsampled version. **CMU-PIE** [40] contains $41{,}368$ face images of 68 subjects under 13 poses, 43 illumination conditions and 4 expressions, captured at 640×486 resolution, so the raw pixel space is $311{,}040$-dimensional; we work from the standard 32×32 cropped faces. For **Caltech101** [17], $9{,}146$ images over 101 categories, the images are not of a fixed size, so there is no single raw dimension to quote. For both Caltech101 and **Stable ImageNet** [22] we therefore use a deep feature-extraction pipeline, with ResNet-50 as the backbone and the features read from its penultimate layer to give $2{,}048$-dimensional vectors, which are then reduced to 128 dimensions by PCA. Stable ImageNet is stored as RGB images at 512×512 resolution, so its raw pixel space is $786{,}432$-dimensional, which is impractical to work in directly. Reducing such data by PCA before the discriminant analysis is the standard preprocessing of this literature [25, 33, 35, 41, 45].

Preprocessing. We apply PCA as preprocessing and retain 98% of the energy, following [33, 35, 41, 45]. The retained dimension is therefore determined by the data, and every method receives exactly the same reduced data.

Initialization. ESRA and DP-ESRA both start from the top-m generalized eigenvectors of the summed between-class and within-class scatters, which is deterministic, costs one eigenvalue decomposition and needs no tuning. Theorem F.1 gives monotone descent from any feasible starting point, and since the problem is non-convex, the initialization also matters, which is why a deterministic, data-driven start is used.

Splits and reporting. Each dataset is split $50/50$ into a training and a test set. The split is repeated 100 times with independent random draws, seeded by the repeat index and shared across methods,

so that every comparison between two methods is paired on identical data. We report the mean classification error over the 100 repeats, and the paired two-sided t-tests of Appendix M are computed on the same 100 paired differences at the 5% level.

Classifiers. Classification in the reduced space uses three rules, the nearest neighbour (1-NN), the nearest mean (NM) and the quadratic discriminant function (QDF),

$$\mathbf{x} \in \underset{i=1,\dots,C}{\arg\min} \left\{ (\mathbf{x} - \bar{\mathbf{x}}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \bar{\mathbf{x}}_i) + \log|\boldsymbol{\Sigma}_i| \right\}, \tag{S76}$$

where $\bar{\mathbf{x}}_i$ and $\boldsymbol{\Sigma}_i$ are the mean and the covariance of class i in the projected space. The principal stress test is $m = 2$, which is the hardest setting for a dimensionality reduction method and also the one that allows the learned subspaces to be inspected visually; $m = 3$ is reported in Appendix L.

Private runs. The budget is varied over $\varepsilon \in \{0.5, 1, 2, 4, 8, 16, \infty\}$ at $\delta = 10^{-5}$, with the noise calibrated by (S12) and the sensitivity of Lemma 4.1. Each configuration is repeated over 50 seeded splits, each with its own draw of the noise, on all four datasets, at $m = 2$. The $\varepsilon = \infty$ column injects no noise and serves as a consistency check, since it must reproduce non-private ESRA exactly. No dataset-specific tuning of any kind is performed on private data, since the temperature floor is set from the public noise scale by Proposition 4.6, and the ridge λ_{reg} is a public constant.

Hardware and software. All algorithms were executed in MATLAB R2025b on a workstation carrying a 64-core AMD EPYC 7763 processor, of which 56 cores were used, and 500 GB of RAM. The runtimes of Appendix N were all measured on that machine under the same worker configuration, so that they are comparable across datasets as well as across methods.

L Results at $m = 3$

Tables S3 and S4 repeat the evaluation of Section 5.1 at projection dimension $m = 3$, over the same 100 splits and with the same nine methods.

ESRA leads across this evaluation as well. On COIL-20 and CMU-PIE, where the classes are close to separable, it holds the lowest error under 1-NN and under QDF, even though every method is already accurate there. On CMU-PIE it reaches 0.21 and 0.04 against 0.38 and 0.23 for LDA, and on COIL-20 it gives 0.32, 0.38 and 0.43 against 0.41, 0.46 and 0.47 for WDA-eig. Under the nearest-mean rule on CMU-PIE both LDA and ESRA stay below six errors in a hundred, at 4.05 and 5.96 respectively. On Stable ImageNet ESRA again leads under 1-NN, 10.47 against 11.38 at $p = 8.9 \times 10^{-6}$, and under QDF, 9.24 against 9.45, and under the nearest-mean rule it is on a par with LDA, whose 9.70 against 10.00 does not reach significance ($p = 0.067$).

On Caltech101, ESRA leads under all three classification rules at $m = 2$, where the least separated pair binds the projection. At $m = 3$ the subspace leaves room for every pair to be separated at once, so the protection of the least separated pair matters less, which is why methods that estimate lower-variance scatters reach lower errors there, WDDR/MMRA at 5.53 and 6.80 under 1-NN and NM and RLDA-$\ell_{2,1}$ at 5.11 under QDF, against 7.85, 8.62 and 11.87 for ESRA.

The privacy study is reported at $m = 2$, the dimension used for the main figures and for the privacy–utility analysis. Neither the sensitivity of Lemma 4.1 nor the mechanism of Section 4.1 depends on m, so the guarantee of Theorem 4.2 holds verbatim at any target dimension.

M Paired significance tests

Table S5 reports, for every dataset and classifier at $m = 2$ and $m = 3$, the paired two-sided t-test over the 100 shared seeded splits between ESRA and the single best-performing competitor on that dataset. The competitor is chosen by the lowest mean-of-three error, so the opponent of each test is fixed before the test is run.

A small p-value with a favourable point estimate means that the improvement is real at the 5% level, whereas a large p-value means that the two methods are statistically indistinguishable on 100 splits. At $m = 2$ the second situation arises on CMU-PIE under the nearest-mean rule ($p = 0.484$) and on

Table S3: Classification error in percent on Caltech101 and Stable ImageNet at $m = 3$, averaged over 100 splits, best in bold and second best underlined.

Dataset	Caltech101			Stable ImageNet		
Method \ Classifier	1-NN	NM	QDF	1-NN	NM	QDF
LDA (1936) [18, 36]	11.50	11.67	11.67	<u>11.38</u>	**9.70**	<u>9.45</u>
HLDA (2004) [26]	11.34	16.86	13.53	15.04	14.63	12.92
WHMMDA (2018) [41, 42]	12.45	16.99	14.95	16.21	15.11	13.22
WDA-eig (2020) [25]	25.54	23.65	24.59	11.69	10.35	10.01
RLDA-$\ell_{2,1}$ (2021) [33]	<u>6.02</u>	<u>8.10</u>	**5.11**	27.31	20.27	18.98
WDDR/MMRA (2024) [45]	**5.53**	**6.80**	<u>5.72</u>	24.78	20.44	19.90
QP-MMDA (2025) [35]	12.08	16.76	14.63	15.76	14.85	13.03
Sparse-MMDA (2025) [35]	11.95	16.79	14.24	16.16	15.10	13.19
ESRA (proposed)	7.85	8.62	11.87	**10.47**	<u>10.00</u>	**9.24**

Table S4: Classification error in percent on CMU-PIE and COIL-20 at $m = 3$, averaged over 100 splits, best in bold and second best underlined.

Dataset	CMU-PIE			COIL-20		
Method \ Classifier	1-NN	NM	QDF	1-NN	NM	QDF
LDA (1936) [18, 36]	<u>0.38</u>	**4.05**	<u>0.23</u>	<u>0.36</u>	0.50	0.51
HLDA (2004) [26]	1.50	7.17	1.28	6.69	11.95	6.73
WHMMDA (2018) [41, 42]	2.38	8.78	2.31	10.61	20.41	13.63
WDA-eig (2020) [25]	0.48	<u>5.01</u>	0.31	0.41	<u>0.46</u>	<u>0.47</u>
RLDA-$\ell_{2,1}$ (2021) [33]	31.13	31.23	19.04	6.15	11.19	4.34
WDDR/MMRA (2024) [45]	1.51	5.20	1.30	0.52	0.69	0.64
QP-MMDA (2025) [35]	2.17	8.47	1.99	12.18	24.84	15.99
Sparse-MMDA (2025) [35]	2.85	9.61	2.59	10.63	20.67	14.06
ESRA (proposed)	**0.21**	5.96	**0.04**	**0.32**	**0.38**	**0.43**

Caltech101 under QDF ($p = 0.737$), and ESRA holds the lower mean error in both, so we report the two as on a par. The three COIL-20 cells at $m = 3$ are of the same kind, ESRA again holding the lower mean error at error levels too small for 100 splits to resolve, while Stable ImageNet under the nearest-mean rule is on a par in the other direction, LDA leading by 0.30 at $p = 0.067$.

N RUNTIME COMPARISON

Table S6 reports the mean per-iteration runtime, in seconds, over the 100 splits and at both target dimensions. All timings were produced on the same machine and with the same worker configuration, so that they are comparable across datasets, across target dimensions and across methods.

All runtime comparisons were conducted under the same per-iteration examination [35]. The reason is that the nine methods do not share a stopping rule, so that a total runtime would measure the stopping rule as much as it measures the algorithm. Some of them terminate on the relative change of the objective and others on the change of the iterate, some use an absolute tolerance and others a tolerance normalized by the current value, and some simply run a fixed number of iterations without testing convergence at all. A per-iteration measurement is invariant to all of these choices, so that every algorithm is assessed under its own convergence conditions, and it is moreover the quantity that the complexity column of Table S1 predicts, so that measurement and analysis can be read against one another.

The nine methods fall into two families, and it is worth saying what separates them before reading the numbers. LDA, HLDA, WDA-eig and RLDA-$\ell_{2,1}$ are not worst-case criteria. They aggregate the class statistics into a single scatter pair and reduce to standard eigenvalue or least-squares problems, so that a low per-iteration cost is precisely what one expects of them. A worst-case criterion, by contrast, must account for the separation of every class pair under its least favourable configuration, and this leads to the minimax fractional program (2), which is an inherently more demanding object. A runtime comparison across the two families is in that sense asymmetric.

Even against these methods, the per-iteration cost of ESRA is lower than that of LDA, HLDA and WDA-eig on all four datasets and at both target dimensions, by factors of between 10 and 24, between 21 and 79, and between 2.6 and 16 respectively. RLDA-$\ell_{2,1}$, which likewise does not address the

Table S5: Paired t-tests over 100 splits, ESRA against the strongest competitor of each dataset and target dimension, with the errors in percent, the lower of the two in bold and the other underlined, a star marking significance at the 5% level and the sign giving the direction.

$m = 2$					
Dataset	Classifier	Competitor	ESRA	Comp.	p (dir.)
COIL-20	1-NN	LDA	**1.16**	4.14	1.77e-13* (+)
	NM	LDA	**1.16**	4.07	1.74e-15* (+)
	QDF	LDA	**1.29**	4.04	1.05e-13* (+)
CMU-PIE	1-NN	WDDR/MMRA	**0.45**	3.16	8.79e-05* (+)
	NM	WDDR/MMRA	**7.10**	7.54	0.484 (+)
	QDF	WDDR/MMRA	**0.39**	2.77	2.98e-05* (+)
Stable ImageNet	1-NN	LDA	**18.15**	25.08	1.45e-38* (+)
	NM	LDA	**16.37**	21.73	2.79e-32* (+)
	QDF	LDA	**15.61**	21.74	6.92e-37* (+)
Caltech101	1-NN	WDDR/MMRA	**10.45**	13.94	1.87e-32* (+)
	NM	WDDR/MMRA	**10.06**	14.18	3.33e-23* (+)
	QDF	WDDR/MMRA	**11.70**	11.79	0.737 (+)
$m = 3$					
Dataset	Classifier	Competitor	ESRA	Comp.	p (dir.)
COIL-20	1-NN	WDA-eig	**0.32**	0.41	0.117 (+)
	NM	WDA-eig	**0.38**	0.46	0.249 (+)
	QDF	WDA-eig	**0.43**	0.47	0.579 (+)
CMU-PIE	1-NN	LDA	**0.21**	0.38	0.202 (+)
	NM	LDA	5.96	**4.05**	3.62e-09* (−)
	QDF	LDA	**0.04**	0.23	0.045* (+)
Stable ImageNet	1-NN	LDA	**10.47**	11.38	8.88e-06* (+)
	NM	LDA	10.00	**9.70**	0.067 (−)
	QDF	LDA	**9.24**	9.45	0.138 (+)
Caltech101	1-NN	WDDR/MMRA	7.85	**5.53**	1.11e-27* (−)
	NM	WDDR/MMRA	8.62	**6.80**	3.07e-15* (−)
	QDF	WDDR/MMRA	11.87	**5.72**	5.24e-38* (−)

worst case, is the only one whose cost is comparable, the two never differing by more than a factor of two and a half, with ESRA ahead of it on both image datasets at either target dimension. A method that solves a minimax fractional program therefore costs no more than the averaging methods it is compared against, and less than all but one of them, which is the practical advantage of removing the inner solver altogether.

Within the worst-case family itself the differences are larger still. The two semidefinite-programming methods are the most expensive, since at $m = 2$ the per-iteration cost of WHMMDA is 30.4 s on COIL-20 and $1{,}983.8$ s on Caltech101, and that of WDDR/MMRA is 20.7 s and 798.7 s on the same two datasets, against 0.071 s and 0.045 s for ESRA, which are factors of roughly 430 and $44{,}000$ against the former and 290 and $17{,}700$ against the latter. The quadratic-programming variants are far cheaper per iteration than the semidefinite ones, yet at $m = 2$ they still cost between about 30 and 225 times what ESRA costs. The orders of magnitude are the same at $m = 3$, where WHMMDA is essentially unchanged and WDDR/MMRA rises on the two deep-feature datasets, reaching $2{,}304.2$ s per iteration on Caltech101, while ESRA stays in the tens of milliseconds throughout. The explanation is structural, since each ESRA iteration consists of a gradient assembly followed by a single thin singular value decomposition, and neither a semidefinite program nor an inner quadratic program is solved at any step, which is precisely the per-iteration accounting of Appendix I.

The overhead of the privacy layer itself, that is the one-time $\mathcal{O}(Cd^2)$ mechanism and the $\mathcal{O}(Cd^3)$ reconstruction of Section 4.1, stays below 5 ms on all four datasets and is therefore not tabulated separately. Because it is paid once for the whole optimization, it is negligible against the total cost of the iterations ESRA runs.

Table S6: Mean per-iteration runtime in seconds over the 100 splits, lowest in bold and second lowest underlined.

$m = 2$	COIL-20	CMU-PIE	Caltech101	Stable ImageNet
LDA (1936) [18, 36]	1.349	0.421	0.511	1.531
HLDA (2004) [26]	1.474	1.620	3.570	3.326
WHMMDA (2018) [41, 42]	30.43	31.27	1983.8	2250.3
WDA-eig (2020) [25]	1.128	0.077	0.585	1.190
RLDA-$\ell_{2,1}$ (2021) [33]	<u>0.073</u>	<u>0.043</u>	<u>0.109</u>	**0.077**
WDDR/MMRA (2024) [45]	20.72	14.54	798.7	1418.3
QP-MMDA (2025) [35]	3.034	2.739	3.508	3.398
Sparse-MMDA (2025) [35]	7.923	6.075	10.078	6.598
ESRA (proposed)	**0.071**	**0.030**	**0.045**	<u>0.110</u>
$m = 3$	COIL-20	CMU-PIE	Caltech101	Stable ImageNet
LDA (1936) [18, 36]	0.314	0.356	0.400	0.468
HLDA (2004) [26]	0.948	0.773	2.520	2.492
WHMMDA (2018) [41, 42]	30.18	25.43	1929.0	2194.4
WDA-eig (2020) [25]	0.132	0.084	0.456	0.306
RLDA-$\ell_{2,1}$ (2021) [33]	<u>0.023</u>	<u>0.026</u>	**0.031**	**0.017**
WDDR/MMRA (2024) [45]	20.53	14.62	2304.2	1682.8
QP-MMDA (2025) [35]	2.478	2.840	3.557	2.784
Sparse-MMDA (2025) [35]	17.263	6.598	17.353	7.678
ESRA (proposed)	**0.021**	**0.015**	<u>0.038</u>	<u>0.041</u>

O Ablations

Two ablations are reported, the first on the number of training samples and the second on the privacy parameter δ. Both use the default annealing schedule of Appendix C.

(i) Effect of the training-set size. Figure S1 varies the training fraction on CMU-PIE at $\varepsilon = 4$ with 50 splits per point, and reports all three classification rules for both the private and the non-private method, so that six curves appear, and three of their features are set out below.

The first is the behaviour the utility bound predicts. The absolute error of DP-ESRA falls monotonically in the number of training samples per class under every rule, from 55.1 at $n = 20$ to 19.9 at $n = 145$ under 1-NN, from 53.9 to 22.0 under the nearest-mean rule, and from 55.8 to 18.9 under QDF, so that a sevenfold increase in n divides the private error by between 2.4 and 3.0. This is the $1/n_{\min}$ term of the statistical error Ξ in Theorem 4.5 made visible, because the per-class means and scatters are formed from more samples and the same absolute noise perturbs them less.

The second is that the *ratio* of private to non-private error is the wrong quantity to read off the figure, since the non-private error becomes almost exact as n grows, its 1-NN error falling from 15.4 to 0.06, so that the ratio rises even while the private error keeps falling. The bound of Theorem 4.5 is additive, and it is the absolute private error that it constrains.

The third is a feature of the private curves that the non-private ones do not share. The three private curves lie almost on top of one another at every sample size, spanning 53.9 to 55.8 at $n = 20$ and 18.9 to 22.0 at $n = 145$, whereas the three non-private curves separate widely, reaching 0.06, 6.4 and 0.15 under 1-NN, NM and QDF respectively at $n = 145$. The reason is that the injected noise does not depend on the classifier, so that once it dominates it sets a common floor which hides the differences between the three decision rules, and only as n grows and that floor recedes does the choice of rule begin to matter again. The nearest-mean rule is the exception that confirms this, since it is the one rule whose non-private error settles, at about 6.4, so that it is also the rule for which the private and non-private curves come closest.

(ii) Sensitivity to δ. On Stable ImageNet we tighten δ from 10^{-4} to 10^{-6} at the two fixed budgets $\varepsilon \in \{1, 4\}$, again reporting all three classification rules. The calibration (S12) makes the prediction precise, since over that range $\sqrt{\log(1.25/\delta)}$ grows from 3.07 to 3.75, so the noise scale σ increases by 22% and the error should rise by a small, monotone amount and no more.

That is what we observe, under every rule. At $\varepsilon = 1$ the 1-NN error rises from 49.46 at $\delta = 10^{-4}$ to 50.35 at 10^{-5} and to 50.59 at 10^{-6}, the nearest-mean error from 40.99 to 41.78 to 42.26, and the

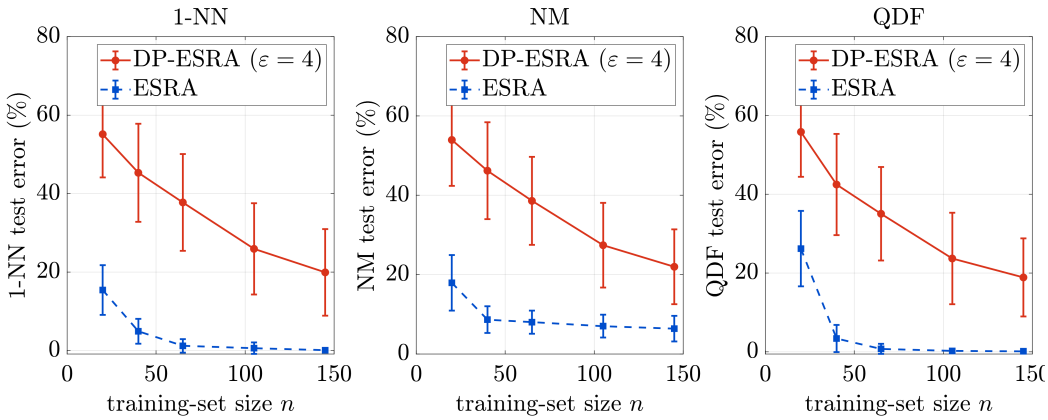

Figure S1: DP-ESRA against non-private ESRA as a function of the training-set size n on CMU-PIE, at $\varepsilon = 4$ and $\delta = 10^{-5}$, for the three classifiers.

QDF error from 40.95 to 41.73 to 42.55. At $\varepsilon = 4$ the three rise from 39.72, 31.32 and 31.51 to 40.94, 31.79 and 32.12. The largest movement anywhere in the study is 1.6, under QDF at $\varepsilon = 1$.

Set against this, moving the budget from $\varepsilon = 4$ to $\varepsilon = 1$ at a fixed δ costs about 10 under each rule, between six and nine times as much as two full orders of magnitude in δ. In this regime, therefore, the budget and not δ is the dominant lever on utility, which is the practical content of the $\sqrt{\log(1/\delta)}/\varepsilon$ shape of (S12), since δ may be set as small as a deployment requires at almost no cost in accuracy, whereas ε cannot.

P Comparison against a per-step gradient-perturbation baseline

The contrast drawn in Section 1, between a one-shot mechanism whose budget does not depend on the iteration count and gradient perturbation whose budget grows with it, is made empirical here. We implemented **DP-SGD-PWCRA**, a per-example-clipped and Poisson-subsampled Gaussian-mechanism solver for the same PWCRA objective [1], with Rényi-DP composition over the T steps and the standard subsampling amplification bound, converted to an (ε, δ) guarantee at the end [28]. So that the comparison isolates the privacy mechanism from the optimizer, its retraction step is the same polar update that ESRA uses (Theorem 3.4), and both methods are run on identical seed-shared splits with $m = 2$, $\delta = 10^{-5}$, 50 repeats and $T = 200$ steps.

Table S7 reports the three classifiers, and the comparison is best read through the *shape* of the two families of curves. DP-ESRA converts additional budget into utility while the baseline stays comparatively flat, since on Caltech101 under 1-NN, moving from $\varepsilon = 0.5$ to $\varepsilon = 16$ takes DP-ESRA from 46.09 to 13.90, a reduction of 70%, and DP-SGD-PWCRA only from 46.99 to 35.79, a reduction of 24%. On CMU-PIE the same range gives $47.07 \to 9.68$ against $49.45 \to 39.62$, and on COIL-20 $40.06 \to 11.64$ against $43.46 \to 21.13$. This is the predicted consequence of spending the budget at every step, because at a fixed ε the per-step noise scale has to grow with T, so most of the extra budget is reabsorbed by composition instead of reaching the solution. The two methods are on a par in a single cell, Caltech101 under the nearest-mean rule at $\varepsilon = 0.5$, where they read 34.23 and 34.26 against standard deviations of about 4, and that cell sits at the tightest budget of the grid.

The $\varepsilon = \infty$ column measures the two optimizers and not the two mechanisms, because the noise-free limit of DP-SGD-PWCRA is not ESRA. A per-step mechanism can weight the class pairs only through quantities it privatizes at every step, and a per-pair within-class scatter would have to be privatized anew at each of the T steps, at a further cost in budget. This is precisely the information that DP-ESRA releases once, and it explains why the noise-free error of the baseline on CMU-PIE under 1-NN is 40.08 where ESRA reaches 0.59. DP-SGD-PWCRA is moreover non-monotone on CMU-PIE, its 1-NN error standing at 39.62 at $\varepsilon = 16$ against 40.08 at $\varepsilon = \infty$, which with standard deviations of about 12 is within the noise.

Table S7: DP-ESRA against a per-step gradient-perturbation baseline on identical seed-shared splits at $m = 2$, $\delta = 10^{-5}$, 50 repeats and $T = 200$ steps, with the mean test error in percent, the lower of the two in bold and the other underlined.

Dataset	Method	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	$\varepsilon = 8$	$\varepsilon = 16$	$\varepsilon = \infty$
1-NN error								
COIL-20	DP-SGD-PWCRA (2016) [1]	43.46	42.57	41.43	38.67	30.74	21.13	15.12
	DP-ESRA (proposed)	**40.06**	**34.52**	**24.82**	**18.21**	**15.24**	**11.64**	**0.86**
CMU-PIE	DP-SGD-PWCRA (2016) [1]	49.45	48.90	48.40	47.41	44.02	39.62	40.08
	DP-ESRA (proposed)	**47.07**	**44.40**	**37.81**	**25.90**	**15.98**	**9.68**	**0.59**
Caltech101	DP-SGD-PWCRA (2016) [1]	46.99	47.47	47.75	46.94	44.27	35.79	17.07
	DP-ESRA (proposed)	**46.09**	**44.84**	**38.57**	**28.03**	**19.51**	**13.90**	**10.59**
Stable ImageNet	DP-SGD-PWCRA (2016) [1]	57.34	56.65	56.37	53.19	47.71	40.02	37.87
	DP-ESRA (proposed)	**54.24**	**50.35**	**45.03**	**40.40**	**37.87**	**35.41**	**17.92**
NM error								
COIL-20	DP-SGD-PWCRA (2016) [1]	52.02	51.27	50.49	47.40	38.58	27.56	20.47
	DP-ESRA (proposed)	**48.29**	**41.47**	**31.43**	**23.22**	**19.26**	**14.67**	**0.90**
CMU-PIE	DP-SGD-PWCRA (2016) [1]	52.88	52.50	52.44	49.07	47.52	41.58	43.71
	DP-ESRA (proposed)	**49.35**	**45.22**	**39.64**	**27.41**	**16.90**	**11.60**	**7.01**
Caltech101	DP-SGD-PWCRA (2016) [1]	34.26	34.59	34.25	34.02	32.02	26.69	23.02
	DP-ESRA (proposed)	**34.23**	**33.20**	**28.53**	**21.41**	**14.58**	**11.04**	**10.04**
Stable ImageNet	DP-SGD-PWCRA (2016) [1]	47.67	47.48	47.53	44.85	38.81	32.37	28.56
	DP-ESRA (proposed)	**46.56**	**41.78**	**35.79**	**31.58**	**29.98**	**27.88**	**15.59**
QDF error								
COIL-20	DP-SGD-PWCRA (2016) [1]	46.27	45.39	45.33	42.56	34.34	23.49	16.76
	DP-ESRA (proposed)	**43.02**	**37.18**	**27.39**	**20.14**	**16.27**	**12.34**	**0.94**
CMU-PIE	DP-SGD-PWCRA (2016) [1]	47.79	47.71	48.46	45.96	44.11	38.50	40.04
	DP-ESRA (proposed)	**45.35**	**41.60**	**35.31**	**23.70**	**14.55**	**8.25**	**0.23**
Caltech101	DP-SGD-PWCRA (2016) [1]	35.86	35.88	35.50	34.78	32.91	27.62	13.24
	DP-ESRA (proposed)	**35.28**	**34.04**	**29.87**	**22.30**	**15.74**	**12.09**	**11.85**
Stable ImageNet	DP-SGD-PWCRA (2016) [1]	47.77	47.63	47.56	44.91	38.96	32.20	28.53
	DP-ESRA (proposed)	**46.54**	**41.73**	**35.76**	**31.81**	**30.19**	**28.08**	**15.10**

Q ADDITIONAL FIGURES

This appendix collects the figures that did not fit within the nine-page budget of the main text, each of them reproduced at full text width so that the individual panels stay legible. Figures S2 and S3 complete Figure 1 of the main text with the learned two-dimensional projections on the three remaining datasets, every panel of a figure being drawn from one and the same split so that the nine methods are compared on identical data. On CMU-PIE the ESRA panel separates all classes with no residual overlap while every competitor leaves some, RLDA-$\ell_{2,1}$ even placing two classes on top of each other, as its error rates in Table 1 anticipate. On Caltech101 the ESRA panel is the cleanest of the nine, and on Stable ImageNet, the most demanding of the four for every method, the ESRA panel is again the one in which the classes are tightest and furthest apart, which is what its error rates in Table 1 record. Figures S4 and S5 plot the privacy–utility curves of Table 2, in each case the classification error against ε on a logarithmic scale for the three classifiers, with the non-private ESRA error as a dashed reference line, and the gap to that line closes fastest on Caltech101, the largest dataset, and slowest on CMU-PIE, the smallest, which is the visual counterpart of the $1/n_{\min}$ term in Theorem 4.5. Figure S6 shows how the learned projection itself changes with the privacy budget. In the last of these, the leftmost panel of each row is non-private ESRA and the remaining panels are DP-ESRA at $\varepsilon \in \{1, 4, 8, \infty\}$ on a fixed split, drawn on a shared axis scale. The subspace visibly tightens and separates as ε grows, most markedly between $\varepsilon = 1$ and $\varepsilon = 8$, and the $\varepsilon = \infty$ panel is indistinguishable from the non-private one on the left, as it must be by construction.

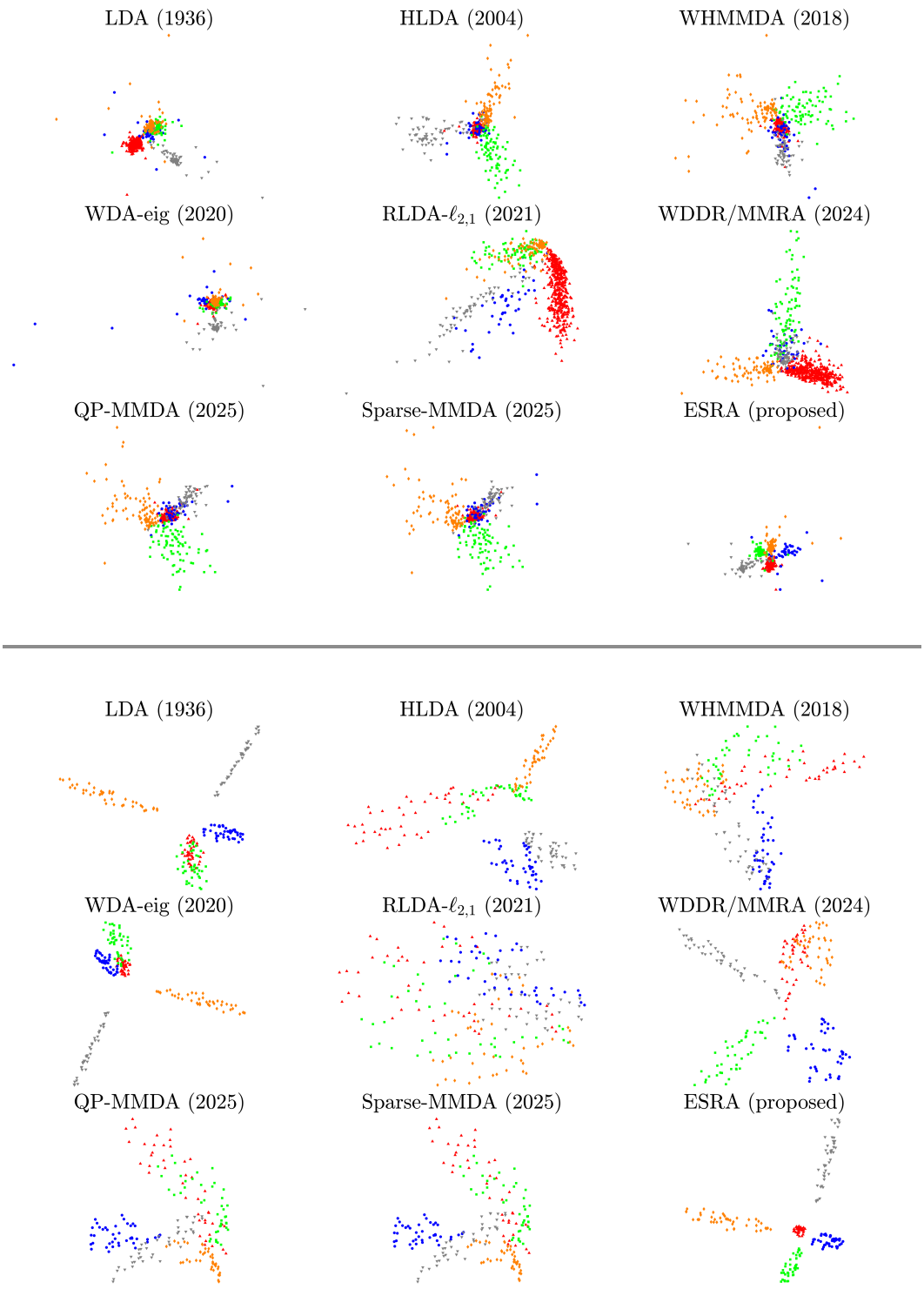

Figure S2: Two-dimensional projections at $m = 2$ on Caltech101 (top) and CMU-PIE (bottom), the methods referenced as in Table 1.

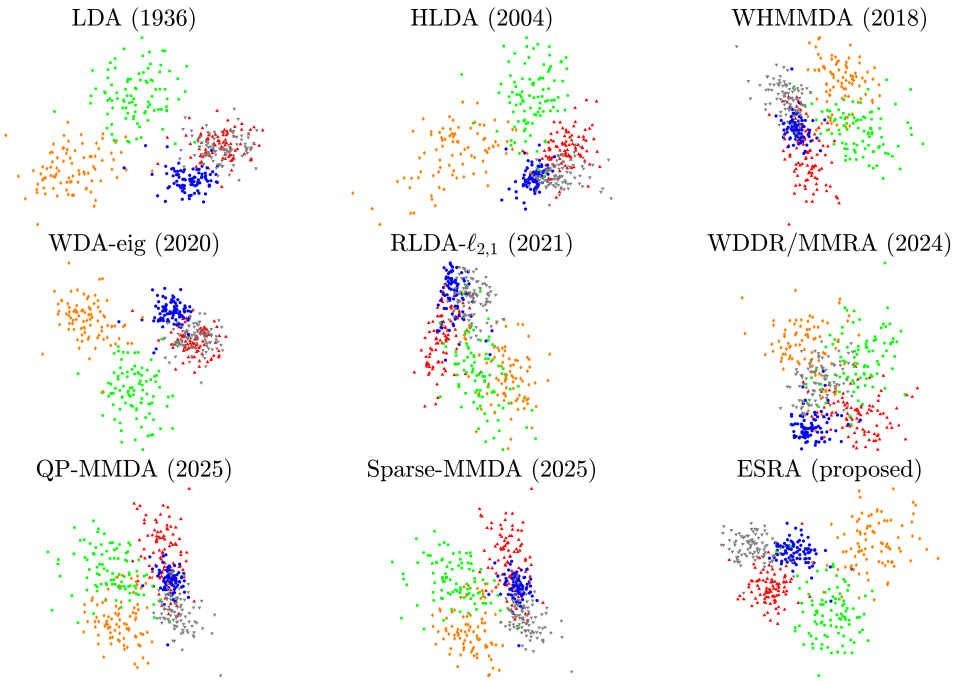

Figure S3: Two-dimensional projections at $m = 2$ on Stable ImageNet, the methods referenced as in Table 1.

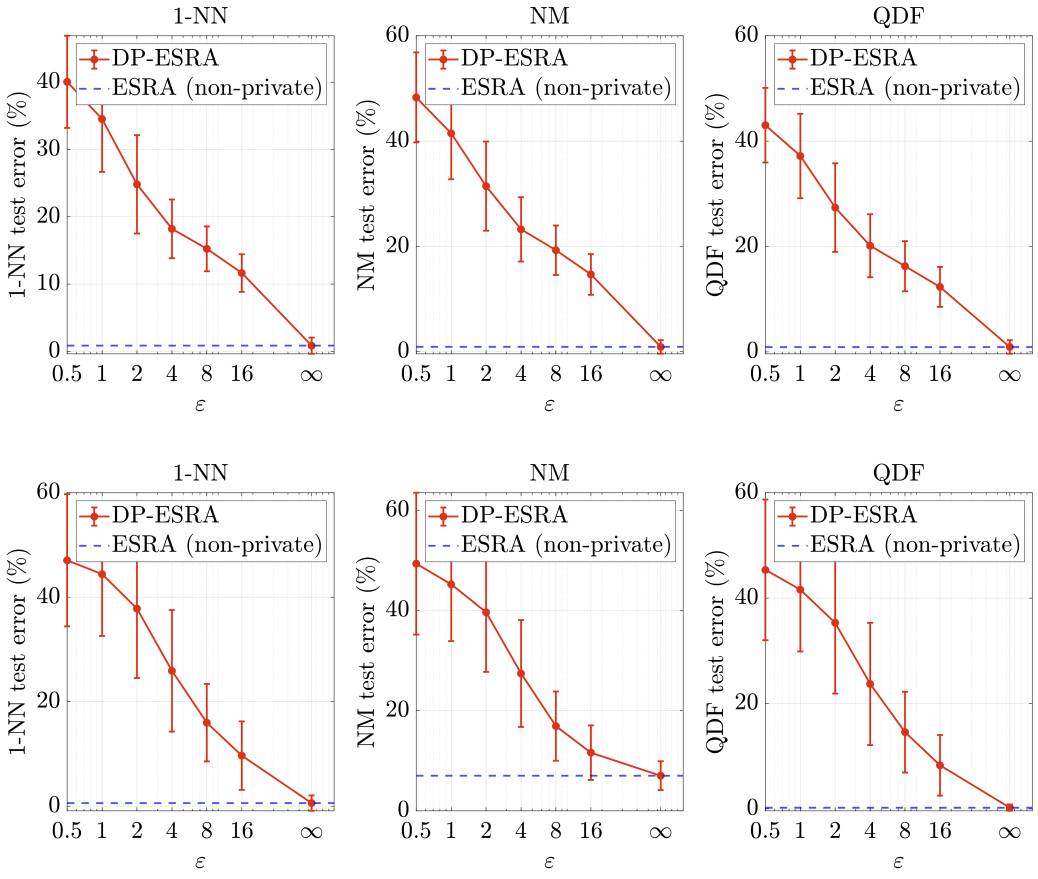

Figure S4: Privacy–utility curves of DP-ESRA on COIL-20 (top) and CMU-PIE (bottom).

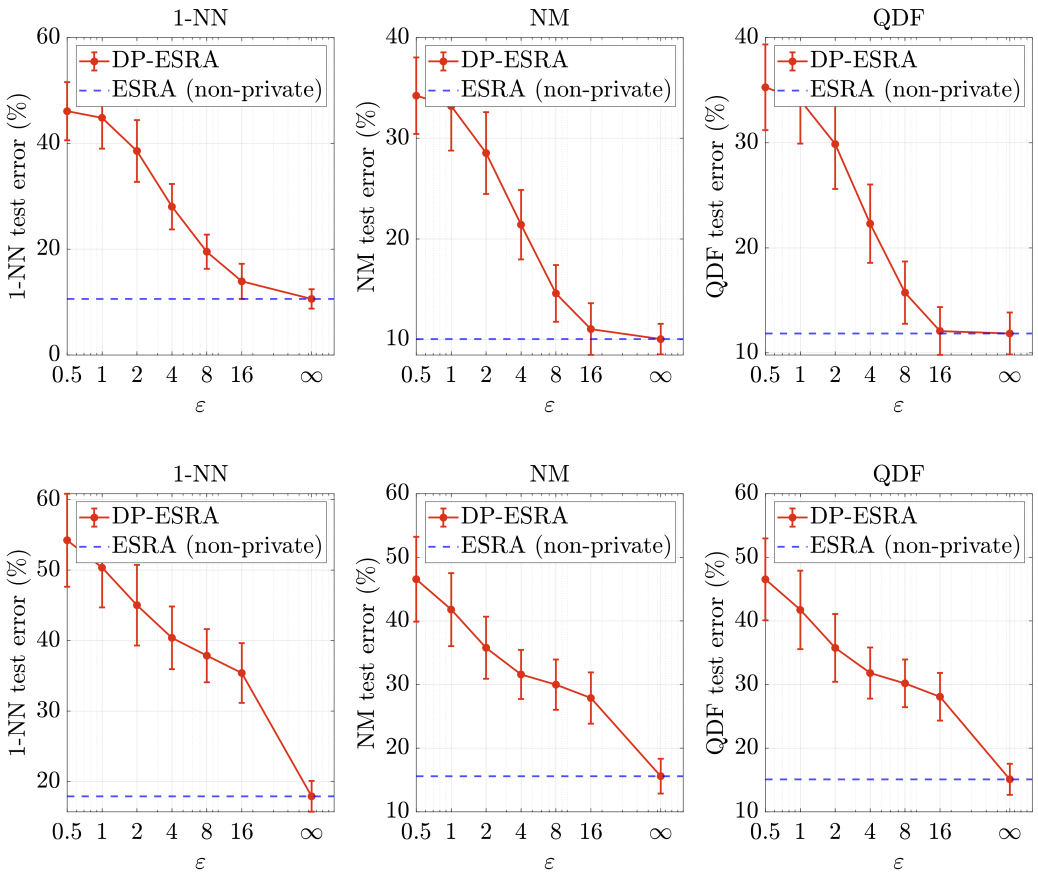

Figure S5: Privacy–utility curves of DP-ESRA on Caltech101 (top) and Stable ImageNet (bottom).

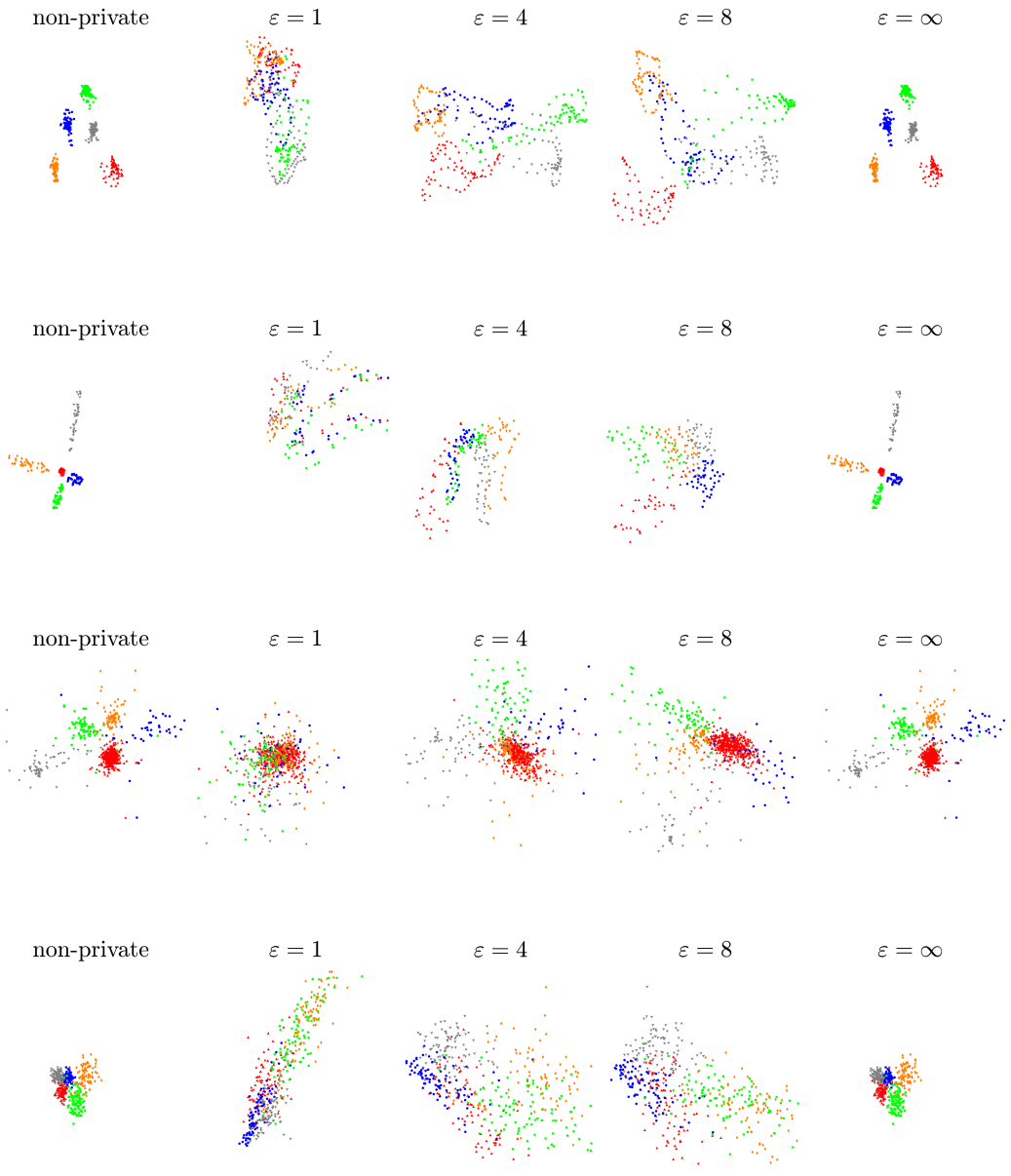

Figure S6: Effect of the budget ε on the projection at $m = 2$, on COIL-20, CMU-PIE, Caltech101 and Stable ImageNet from top to bottom, the leftmost panel of each row being non-private ESRA and the rest DP-ESRA.