PRIVACY-PRECONDITIONED REPRESENTATION LEARNING FOR DIFFERENTIALLY PRIVATE ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce *privacy preconditioning*: designing public training to improve the accuracy of a later learning stage subject to differential privacy. Rather than prepare a model only for predictive transfer, the public stage anticipates the clipping and noise required during private adaptation. We instantiate this principle in shared-representation learning: public tasks learn a low-dimensional representation, which is frozen while each new private task learns its own predictor. The public objective combines prediction loss with a penalty on either the average squared norm or the maximum norm of projected features. These norms influence the per-example gradients that private optimization clips and perturbs. For shared linear regression, we relate public prediction risk to representation recovery and bound the expected squared projected-feature norm on new tasks. Our private excess-risk bounds quantify the contributions of representation approximation, gradient clipping, and privacy noise. For the maximum penalty, a public-data coverage condition yields a bound on every private projected norm and an excess-risk guarantee using a sufficient clipping threshold. In synthetic regression tasks where the relationship between a predictive feature direction and the target varies across tasks, average-squared-norm regularization improves private mean squared error by 8.8% at a fixed privacy budget despite slightly worse non-private prediction. Repeating the experiment with fresh random seeds and newly generated tasks gives a similar improvement. Privacy preconditioning makes downstream private utility an explicit objective of public representation design, before any private records are accessed.

1 INTRODUCTION

A model trained on public data is often adapted to a new task whose examples are sensitive. Public training can supply useful features, so the private stage needs to learn only a small predictor rather than an entire model. Differential privacy limits the effect of an individual training record on the released model (Dwork et al., 2006). A common approach is differentially private stochastic gradient descent (DP-SGD) (Song et al., 2013; Bassily et al., 2014). Its widely used implementation clips each example's gradient and adds Gaussian noise to the aggregate (Abadi et al., 2016). Both operations can reduce prediction accuracy.

We introduce *privacy preconditioning*, a design principle for two-stage learning: deliberately prepare the public stage for the privacy requirements of the later stage. Public training accesses no private records and need not itself enforce differential privacy. It can nevertheless anticipate the clipping and noise that private adaptation will require. The question is therefore not only which representation transfers best without privacy, but *which public training choices lead to the most accurate adaptation under privacy constraints?* This perspective can guide other pipelines with public preparation followed by a private task; shared-representation learning provides a concrete setting in which to develop it.

An input $x \in \mathbb{R}^d$ is mapped to $Ux \in \mathbb{R}^k$, where $U \in \mathbb{R}^{k \times d}$ is shared across tasks and $k \ll d$. A task-specific parameter vector $v \in \mathbb{R}^k$ predicts $v^\top Ux$. Public tasks learn U and their own vectors; a new private task freezes U and learns only its v. For an example with target y, the squared-loss

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.

gradient satisfies

$$\|\nabla_v (y - v^\top U x)^2\|_2 = 2|y - v^\top U x| \, \|Ux\|_2.$$

This identity motivates penalizing projected feature norms during public training, while retaining prediction loss to preserve task information. Smaller gradients can permit smaller clipping thresholds and less noise at a fixed privacy budget. More generally, changing the representation can reduce the prediction-risk cost of clipping and noise, even without a smaller selected threshold.

We develop two versions of *privacy-preconditioned representation learning*. The first adds an average squared projected-feature penalty to the public prediction loss. The second uses a *maximum-norm penalty*: it penalizes the largest projected norm across all examples from all public tasks, rather than an average. This targets the largest gradients the private stage might encounter. Under a coverage condition relating private features to public examples, this public maximum bounds every private projected norm and gives a sufficient gradient bound. In both objectives, $UU^\top = I_k$ makes the rows orthonormal, preventing trivial rescaling of U and compensating rescaling of v. The penalty therefore changes the chosen subspace, not merely its coordinate scale. Model selection uses public data, and the private algorithm retains gradient clipping and valid privacy accounting.

Contributions. Our conceptual contribution is a privacy-preconditioning framework that makes later private utility a concrete design criterion for earlier public learning. We instantiate it through two projected-norm objectives and minibatch training procedures. For the maximum penalty, updates can scan all public examples or use a cached subset with an explicit bound on the gap to the true maximum.

For shared linear regression, we relate public prediction risk to recovery of the shared subspace and bound projected feature norms on new tasks. We derive private prediction-risk bounds with separate terms for approximation error of the learned representation, gradient-clipping distortion, and privacy noise. For the maximum penalty, a public coverage condition yields a uniform bound on private projected norms and an expected excess-risk guarantee. Section 5 states the results, with proofs in the appendices.

Synthetic regression experiments vary the relationship between a predictive feature direction and the target across tasks. The average-squared penalty improves private mean squared error despite slightly worse non-private prediction, and the result persists with fresh random seeds. Additional experiments on the Inner London Education Authority (ILEA) School benchmark (Bonilla et al., 2007) show that its spectral implementation instead selects a lower effective dimension. The experiments distinguish improvements in private adaptation from gains explained by dimension reduction. Other synthetic experiments show that reducing projected norms alone need not improve private utility.

2 Related Work

Private optimization and public data. Private empirical risk minimization and stochastic convex optimization characterize the statistical cost of privacy (Bassily et al., 2014; 2019). DP-SGD and Rényi differential privacy support iterative training and accounting (Abadi et al., 2016; Mironov, 2017; Wang et al., 2019; Birrell et al., 2024). Gradient clipping itself can bias optimization (Chen et al., 2020); adaptive clipping privately estimates a useful threshold during training (Andrew et al., 2021). We instead design the preceding public representation-learning stage for private adaptation and tune adaptation settings using public tasks.

Public features and pretraining can substantially improve private learning (Tramèr & Boneh, 2021; Ganesh et al., 2023). Public-data-assisted mirror descent uses the public loss to guide private optimization (Amid et al., 2022), while public or synthetic data can help recenter clipping (Nasr et al., 2023). Bu et al. (2024) analyze public pretraining and private training through Hessian-dependent loss dynamics; Thaker et al. (2024) study public representations for private transfer under distribution shift. Our focus is an explicit upstream feature-norm penalty, a subsequently frozen representation, and its effect on the private task predictor. The coverage result additionally connects a pooled public maximum to a sufficient private gradient bound.

Shared representations and private adaptation. Multi-task and meta-learning exploit common predictive structure (Maurer et al., 2016; Tripuraneni et al., 2021). FedRep alternates a shared representation with client-specific predictors and provides linear-model recovery guarantees (Collins

et al., 2021). Private model personalization has been studied under user-level privacy (Jain et al., 2021); Snedeker et al. (2025) develop sharper algorithms and guarantees in the shared linear model. We adopt the shared-model perspective but learn the representation from public tasks, then protect records within each new private task. This differs from privately learning the shared representation from participating users.

Related approaches include low-rank reparameterization and private fine-tuning of pretrained models (Yu et al., 2021; 2022; Li et al., 2022), and parameter-efficient low-rank adaptation (Hu et al., 2022). Our question is complementary: how should public training prepare the representation for private adaptation, even when its dimension is already fixed?

3 Problem Setup and Preliminaries

Tasks and prediction. Public task $j \in [m] := \{1, \dots, m\}$ has a distribution $\mathcal{D}_j$ on $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ and n_j independent examples S_j. Write $S^{\mathrm{pub}} = (S_1, \dots, S_m)$ and $V = (v_1, \dots, v_m)$. For squared loss $\ell(U, v; (x, y)) = (y - v^\top U x)^2$, define

$$\begin{aligned} L_j(U, v) &= \mathbb{E}_{\mathcal{D}_j} \ell(U, v; (x, y)), \qquad \widehat{L}_j(U, v) = \frac{1}{n_j} \sum_{i=1}^{n_j} \ell(U, v; (x_{j,i}, y_{j,i})), \\ L_{\mathrm{pub}}(U, V) &= \frac{1}{m} \sum_j L_j(U, v_j), \qquad \widehat{L}_{\mathrm{pub}}(U, V) = \frac{1}{m} \sum_j \widehat{L}_j(U, v_j). \end{aligned} \tag{1}$$

Each task has equal weight, even if sample sizes differ. Representations belong to $\mathcal{U} = \{U \in \mathbb{R}^{k \times d} : UU^\top = I_k\}$, and $P_U = U^\top U$ projects onto their row space. Matrix norms $\|\cdot\|_2$ and $\|\cdot\|_F$ denote spectral and Frobenius norms; vector norms are Euclidean.

For the recovery guarantees, tasks follow the realizable shared linear model

$$y = \langle v_j^\star, U^\star x \rangle + \epsilon, \qquad U^\star \in \mathcal{U}, \quad \mathbb{E}[\epsilon \mid x, j] = 0, \quad \mathbb{E}[\epsilon^2 \mid j] = \tau^2. \tag{2}$$

The feature second moment is $\Sigma_j = \mathbb{E}_{\mathcal{D}_j}[xx^\top]$. Transfer results draw public tasks and a new private task independently from a common task distribution $\mathcal{Q}$. Conditional on its task, the private sample contains n independent records. The synthetic experiment in Section 6 also examines a rank-five representation when the task coefficients span six directions.

Record-level privacy and adaptation. An algorithm is (ε, δ)-differentially private if $\Pr[\mathcal{A}(S) \in E] \le e^\varepsilon \Pr[\mathcal{A}(S') \in E] + \delta$ for every measurable output event E and datasets S, S' differing in one record. We use replace-one adjacency: dataset size is fixed. With U frozen, each step draws an independent uniform size-b subset B_t of the same n private records. For a positive clipping threshold C,

$$\begin{aligned} g_{t,i} &= \nabla_v \ell(U, v_t; (x_i, y_i)), \quad \bar{g}_t = \frac{1}{b} \sum_{i \in B_t} \frac{g_{t,i}}{\max\{1, \|g_{t,i}\|_2 / C\}}, \\ v_{t+1} &= v_t - \eta(\bar{g}_t + \xi_t), \qquad \xi_t \sim \mathcal{N}(0, \sigma^2 C^2 I_k / b^2). \end{aligned} \tag{3}$$

Here η is the step size and σ is the noise multiplier relative to C. The averaged-gradient sensitivity is at most $2C/b$, so its noise-to-sensitivity ratio is $\sigma/2$. The accountant uses the actual sampling, adjacency, number of steps, and target δ to select σ. A projected variant constrains v_t to a fixed ball. Both variants retain clipping on every dataset, irrespective of the distributional assumptions used for utility analysis.

4 Methodology

4.1 Preparing representations for private optimization

For squared loss, the per-example gradient is

$$\nabla_v \ell(U, v; (x, y)) = -2(y - v^\top U x) U x. \tag{4}$$

This suggests two ways to favor smaller downstream gradients while preserving predictive information. The first penalizes the average squared projected-feature norm:

$$\min_{U\in\mathcal{U},V} \widehat{J}_\lambda(U,V) := \widehat{L}_{\mathrm{pub}}(U,V) + \lambda \widehat{R}_{\mathrm{proj}}(U), \qquad \widehat{R}_{\mathrm{proj}}(U) = \frac{1}{m}\sum_j \frac{1}{n_j}\sum_i \|Ux_{j,i}\|_2^2. \quad (5)$$

The prediction loss encourages useful features; the penalty anticipates the gradients that must later be clipped and perturbed. The coefficient $\lambda \geq 0$ balances these objectives, with $\lambda = 0$ recovering unregularized shared-representation learning. We call this average-squared penalty *AvgSq*.

An average can be small even when a few projected examples have large norms. Our second approach, *Max*, therefore replaces the average penalty by the largest projected norm over the entire public sample:

$$\widehat{R}_{\max}(U) = \max_{j\in[m],\, i\in[n_j]} \|Ux_{j,i}\|_2, \qquad \widehat{J}^{\max}_\lambda(U,V) = \widehat{L}_{\mathrm{pub}}(U,V) + \lambda \widehat{R}_{\max}(U). \quad (6)$$

We minimize $\widehat{J}^{\max}_\lambda$ over the same $U \in \mathcal{U}$ and task vectors V. The maximum is unsquared and includes every record from every public task, not just the current minibatch. Section 5.3 shows when this public radius bounds the projected norms of new private records.

Let $\widehat{\Sigma}_x = m^{-1}\sum_j n_j^{-1}\sum_i x_{j,i}x_{j,i}^\top$ and $\bar{\Sigma} = m^{-1}\sum_j \Sigma_j$. Then $\widehat{R}_{\mathrm{proj}}(U) = \mathrm{Tr}(P_U\widehat{\Sigma}_x)$ and $R_{\mathrm{proj}}(U) = \mathrm{Tr}(P_U\bar{\Sigma})$ is its population counterpart for the realized public tasks. For $Q \in \mathbb{R}^{k\times k}$ satisfying $Q^\top Q = QQ^\top = I_k$, replacing (U, v) by (QU, Qv) preserves both predictions and projected norms. Together with $UU^\top = I_k$, both penalties select a subspace rather than arbitrarily rescaling features or changing coordinates within that subspace.

4.2 Training, calibration, and private adaptation

The public solver alternates minibatch task-vector and representation updates, using an orthogonal–triangular (QR) factorization to restore orthonormal rows. Its initializer adapts the moment statistic of Snedeker et al. (2025) to public data. For Max, scanning all public records supplies a subgradient of the maximum; Appendix D.3 also gives cached approximations and an alternative stochastic formulation. Appendix A gives the remaining algorithms and costs. The analysis retains public optimization error and an alternative validation-based certificate.

For a new task, freeze the representation and run (3). The synthetic calibration procedure evaluates candidate representations by running the intended clipped, noisy adaptation on held-out *public* tasks and measuring loss on their public query records. It selects (λ, C), or (λ, C, η) in the Max experiments, with zero included among the penalty values and each representation free to select its own threshold. All choices are fixed before final private evaluation. The ILEA School experiment uses a public spectral surrogate and its stated public selection protocol, not the iterative solver of (5); Appendix C.3 makes this distinction explicit.

Clipping and Gaussian noise are still applied in the private stage. Public selection does not spend privacy budget: conditional on any public output, the selected private mechanism must satisfy the same target (ε, δ) guarantee. Averaging over the public randomness preserves that guarantee (Appendix B.6). This remains true whether or not a coverage or boundedness assumption used for utility holds.

5 Theoretical Analysis

We relate public prediction risk to representation recovery and projected norms on new tasks, then derive private-risk guarantees for both penalties. Risks below are population squared losses. The reference is the true predictor's risk τ^2, not the best risk within a potentially poor learned representation.

5.1 PUBLIC RECOVERY AND TRANSFER

Let $G_V = m^{-1}\sum_j v_j^\star (v_j^\star)^\top$ describe public task diversity. If $\Sigma_j \succeq \alpha I_d$ for all tasks and $G_V \succeq \gamma I_k$, with $\alpha, \gamma > 0$, then

$$L_{\mathrm{pub}}(U,V) - \tau^2 \geq \frac{\alpha\gamma}{2}\|P_U - P_{U^\star}\|_F^2. \tag{7}$$

Feature excitation makes errors in ambient directions observable, and task diversity makes every shared predictive direction visible. Appendix B.1 proves this inequality and the following empirical consequence.

Fix a public head ball $\mathcal{V}_B = \{V : \max_j \|v_j\|_2 \leq B_v\}$ containing the true heads. Write $J_\lambda = L_{\mathrm{pub}} + \lambda R_{\mathrm{proj}}$ and $\Delta_B = \sup_{U\in\mathcal{U}, V\in\mathcal{V}_B} |J_\lambda(U,V) - \widehat{J}_\lambda(U,V)|$. If $(\widehat{U}_\lambda, \widehat{V}_\lambda)$ is within ξ_{opt} of the empirical optimum over this class, then

$$\|P_{\widehat{U}_\lambda} - P_{U^\star}\|_F^2 \leq B_{\mathrm{sub}} := \min\left\{2k, \frac{2}{\alpha\gamma}[2\Delta_B + \xi_{\mathrm{opt}} + \lambda R_{\mathrm{proj}}(U^\star)]\right\}. \tag{8}$$

The bound is expressed in terms of separate quantities representing statistical estimation error, public optimization error, and the cost of regularization. For bounded features and labels, Appendix B.2 gives an explicit finite-sample bound on Δ_B. Appendix B.2 sharpens the regularization term using covariance anisotropy, and Appendix B.2 provides a held-out certificate without ξ_{opt}. Any applicable sharper certificate can replace B_{sub}.

To transfer projected norms, let $\Sigma_{\mathrm{meta}} = \mathbb{E}_{\theta\sim\mathcal{Q}}\Sigma_\theta$ and $R_{\mathrm{meta}}(U) = \mathrm{Tr}(U\Sigma_{\mathrm{meta}}U^\top)$. If $\|x\|_2 \leq X$, then with probability at least $1-\delta_R$ over the public tasks and records,

$$\sup_{U\in\mathcal{U}} |R_{\mathrm{meta}}(U) - \widehat{R}_{\mathrm{proj}}(U)| \leq \Gamma_R, \quad \Gamma_R = \sqrt{k}X^2\left(1+\sqrt{2\log\frac{2}{\delta_R}}\right)\left(\frac{1}{\sqrt{m}} + \frac{1}{m}\sqrt{\sum_j \frac{1}{n_j}}\right). \tag{9}$$

The first term depends on the number of public tasks; the second accounts for the number of records available within each task. Uniformity allows U and λ to be learned from the same public data. In particular, averaging over a new independent task and its features bounds $\mathbb{E}_{\theta,x}\|Ux\|_2^2$ by $r_{\mathrm{pub}}(U) := \min\{X^2, \widehat{R}_{\mathrm{proj}}(U) + \Gamma_R\}$. Appendix B.3 proves this statement. It is an average over new tasks, not a worst-case bound on individual private records.

5.2 PRIVATE ADAPTATION AND PREDICTION RISK

For a task θ, let $L_\theta(U) = \inf_v L_\theta(U,v)$. The total excess risk decomposes exactly as

$$L_\theta(U,v) - \tau^2 = \underbrace{L_\theta(U) - \tau^2}_{\text{representation approximation error}} + \underbrace{L_\theta(U,v) - L_\theta(U)}_{\text{adaptation error}}. \tag{10}$$

If $\Sigma_\theta \preceq \beta I_d$, the first term is at most $\frac{\beta}{2}\|v_\theta^\star\|_2^2\|P_U - P_{U^\star}\|_F^2$. The second is quadratic in the head error, with matrix $H_U = U\Sigma_\theta U^\top$. For example, adding centered Gaussian noise of covariance $\sigma^2C^2I_k/b^2$ to an otherwise fixed update adds $\eta^2\sigma^2C^2\,\mathrm{Tr}(H_U)/b^2$ to its expected population loss. Appendix B.4.1 derives the accumulated noise cost over a full population-gradient trajectory.

The following result analyzes projected DP-SGD using minibatches drawn repeatedly from the same private training set. Assume uniformly over tasks that $\|x\|_2 \leq X$, $|y| \leq Y$, $\|v_\theta^\star\|_2 \leq B_\star$, and $\mu_0 I_d \preceq \Sigma_\theta \preceq \beta I_d$ for $\mu_0 > 0$. Project heads onto $\mathcal{B}_v = \{v : \|v\|_2 \leq B_v\}$ with $B_v \geq \beta B_\star/\mu_0$, which suffices to contain each optimal fixed-representation head. Set $B_r = Y + B_v X$.

Theorem 1 (Public-to-private risk bound)**.** *On the public recovery and transfer events above, fix $U = \widehat{U}_\lambda$ and any positive public-only threshold C. Start from a public-only $v_0 \in \mathcal{B}_v$, use $n \geq 2$ private records and independent uniform size-b subsets, and project every update in* (3) *onto $\mathcal{B}_v$. Let $0 < \eta \leq 1/(2\beta)$ and choose $\zeta > 0$ such that $q = (1+\zeta)(1-2\eta\mu_0)^2 < 1$. Define $a = 1 + 1/\zeta$,*

$S_T = \sum_{t=0}^{T-1} q^t$, *and* $\nu_{n,b} = (n-b)/(b(n-1))$. *Conditional on the public stage,*

$$\mathbb{E}[L_\theta(U, v_T) - \tau^2] \le \frac{\beta B_\star^2}{2} B_{\mathrm{sub}} + 4\beta B_v^2 q^T + \beta\eta^2 S_T \left\{ \frac{2aB_r^4 X^2(1+1/n)}{C^2} r_{\mathrm{pub}}(U) + \frac{16a(B_v^2 X^2 + Y^2)}{n} r_{\mathrm{pub}}(U) + \left(\nu_{n,b} + \frac{k\sigma^2}{b^2}\right) C^2 \right\}. \tag{11}$$

The expectation is over the new task, its records, and private optimization randomness. The multiplier σ is calibrated for this sampling scheme and horizon.

The first line bounds approximation and initialization error. Inside the braces, the C^{-2} term bounds clipping bias; the $1/n$ term accounts for estimating population gradients from n records. The last term contains minibatch gradient variability and privacy noise. The bound expresses how the clipping–noise trade-off depends on the learned representation. Appendix B.5 proves a task-specific result, and Appendix B.6 derives (11). The task-vector norm constraint is part of this guarantee; Section 7 describes how the experimental update rules differ.

5.3 Maximum-norm regularization and coverage

The maximum penalty in (6) gives a direct route to bounding private gradients when public examples cover the feature vectors of the new private task. We now state that coverage condition and the resulting private-risk guarantee. Let S_x^{pub} be the pooled public features. For a finite set Q, define $\mathrm{SymConv}(Q) = \mathrm{conv}\{a, -a : a \in Q\}$. The *coverage event* $\mathcal{E}_{\mathrm{cov}}$ requires that for each private record x_t there is a public subset $Q_t \subseteq S_x^{\mathrm{pub}}$ with $x_t \in \mathrm{SymConv}(Q_t)$. Each record may use different raw, unnormalized anchors, with no restriction on their number. Assume $\Pr(\mathcal{E}_{\mathrm{cov}}) \ge 1 - \beta_{\mathrm{cov}}$ jointly over public and private draws.

On this event, write $x_t = \sum_{a \in Q_t} c_a a$ with $\sum_a |c_a| \le 1$. The triangle inequality gives

$$\|Ux_t\|_2 \le \sum_{a \in Q_t} |c_a| \, \|Ua\|_2 \le \widehat{R}_{\max}(U) =: r_U. \tag{12}$$

This holds simultaneously for every U, including a publicly fitted representation. If $|y| \le Y$ and heads remain in $\mathcal{B}_v$, every covered training gradient is bounded by $C_U^{\mathrm{cert}} = 2r_U(Y + B_v r_U)$, at every iteration.

Theorem 2 (Coverage-based private excess risk)**.** *Suppose the bounded shared linear model holds uniformly with $\|x\|_2 \le X$, $|y| \le Y$, and $\|v_\theta^\star\|_2 \le B_\star \le B_v$, and the coverage event has probability at least $1-\beta_{\mathrm{cov}}$. Let U be any publicly learned row-orthonormal representation. Set $w_\theta^\star = (U^\star)^\top v_\theta^\star$ and $A_\theta(U) = \mathbb{E}_x\left[\left(((I - P_U)w_\theta^\star)^\top x\right)^2\right]$. Run projected updates with independent uniform size-b subsets of $n \ge 2$ private records, starting from public $v_0 \in \mathcal{B}_v$, with a positive public threshold $C_U \ge C_U^{\mathrm{cert}}$ satisfying $\mathbb{E}[C_U^2] < \infty$, a fixed public step size $\eta > 0$, and return $\bar{v}_T = T^{-1}\sum_{t=0}^{T-1} v_t$. Then*

$$\mathbb{E}[L_\theta(U, \bar{v}_T) - \tau^2] \le \mathbb{E}[A_\theta(U)] + \frac{2B_v^2}{\eta T} + \frac{\eta}{2}\left(1 + \frac{k\sigma^2}{b^2}\right)\mathbb{E}[C_U^2] + \frac{4B_v Y X + 2B_v^2 X^2}{\sqrt{n}} + \beta_{\mathrm{cov}}(Y + B_v X)^2. \tag{13}$$

The expectations are over the public stage and its randomness, the new task and its finite private sample, and the private optimization randomness. Clipping remains in the mechanism, so its differential privacy guarantee holds on all datasets, not only on $\mathcal{E}_{\mathrm{cov}}$.

The approximation term compares with the true predictor and can be bounded by public subspace recovery. At the certified threshold, the noise cost scales with $r_U^2(Y + B_v r_U)^2$, so a smaller radius lowers this cost while the prediction loss limits the opposing approximation error. Coverage removes training-gradient clipping error and the last term accounts for coverage failure. No strong-convexity or unrestricted-head containment assumption is required. Appendix D.2 proves

the theorem; Appendices D.3–D.4 give minibatch implementations and concrete coverage models, including signed-anchor and directional tasks. The Max experiments and threshold follow-up are in Appendix D.5.

The certificate is a sufficient analytical choice, not an experimental prescription. Empirical comparisons tune C on public adaptation tasks and may select $C < C_U^{\mathrm{cert}}$. Such runs remain private but need not avoid clipping; their utility is not claimed to satisfy the no-clipping bound (13).

6 Experiments

We evaluate prediction on new private tasks after all representations and adaptation settings have been selected using public data. The direct comparison is with the corresponding unregularized representation ($\lambda = 0$). We report mean squared error (MSE); the ILEA School benchmark uses normalized MSE (NMSE), dividing each school's test MSE by its test-target variance. The synthetic experiments have analytically known population risk, allowing us to calculate both the representation approximation error and private adaptation error in (10). Appendix C gives the data, grids, noise calibration, and additional comparisons.

6.1 Private-utility gains on synthetic regression tasks

In the *task-instability* experiment, one feature direction predicts the target within each task, but the strength and sign of that relationship vary across tasks. Its coefficient must therefore be learned anew during private adaptation. Inputs in $\mathbb{R}^{100}$ have bounded coordinates and additional variance in a five-dimensional subspace. Labels combine five stable predictive directions with one additional direction having a task-varying coefficient. We learn a rank-five representation for this six-dimensional predictive family, testing which information public learning should preserve for private adaptation. Appendix C.2 specifies the generator and distinguishes the feature-variance subspace from the label directions.

We use 50 public tasks, 1,000 public epochs, and joint public selection of (λ, C) with $\lambda = 0$ included. Each new task has 80 adaptation records; DP-SGD runs 100 steps with batch size 8 and fixed step size 0.01. Noise is calibrated at $\delta = 1/80^2$ for each target ε. The primary evaluation uses 24 random seeds. We repeat the full experiment on another 24 fresh seeds, generating new public task collections and private evaluation tasks with the same fixed configuration. Neither set is used for selection, and their results are not pooled.

At $\varepsilon = 2$, AvgSq lowers primary-set MSE from 1.8749 to 1.7090: an 8.8% reduction, favorable on all 24 seeds, with a 95% interval $[0.1368, 0.1965]$ for the absolute reduction. At $\varepsilon = 8$, MSE falls from 1.3668 to 1.3365, favorable on 21 of 24 seeds. The fresh-seed repetition gives closely matching

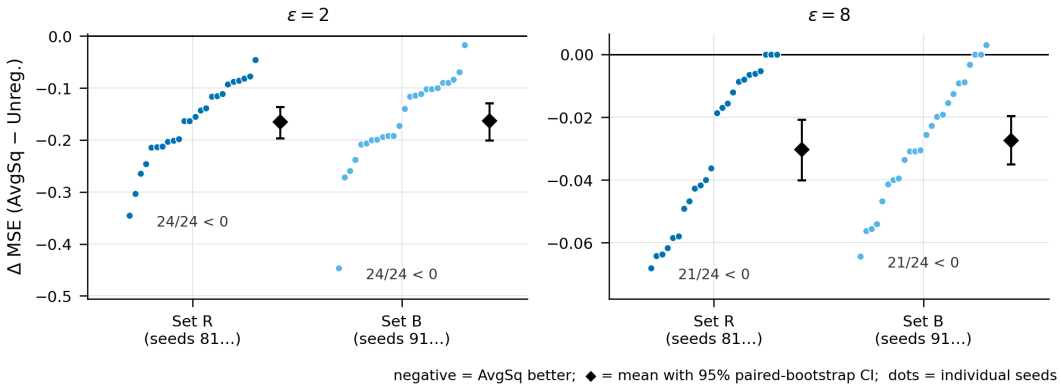

Figure 1: Task-instability experiment: AvgSq minus unregularized MSE on the same generated tasks. Each point corresponds to one public-task collection; negative values favor AvgSq. The 95% intervals for the mean difference are obtained by resampling collections while keeping the methods' results paired. The second set uses fresh seeds and task collections.

reductions of 0.1633 and 0.0274 (Figure 1). The benefit also persists at intermediate privacy budgets and at 50 and 200 private steps, with noise recalibrated for each horizon.

What improves under privacy. At $\varepsilon = 2$, the regularized representation incurs about 0.029 more approximation error but 0.195 less private adaptation error. With neither clipping nor noise, its MSE is slightly worse. Ablation experiments retain the representation, data, initialization, batches, step size, and horizon, while removing noise or both clipping and noise. Adding noise worsens each method: its MSE cost is about 0.337 for AvgSq versus 0.419 for $\lambda = 0$ in the primary set. Clipping likewise costs less under AvgSq. The private-utility gain therefore reflects smaller *accuracy costs of clipping and privacy noise*.

The advantage persists when both methods use the same threshold and noise scale. Thus the empirical benefit need not come from selecting a smaller C. An additional experiment fixes this predictive coefficient across tasks while keeping the feature distribution and the coefficient's second moment unchanged. Regularization then gives a much smaller benefit at $\varepsilon = 2$ and no benefit at $\varepsilon = 8$, suggesting that the difficulty of learning task-specific coefficients matters.

6.2 Real-data adaptation and effective dimension

ILEA School benchmark. We treat exam-score prediction at each school as a task (Bonilla et al., 2007). Five public/private assignments each contain 20 private schools, with 80 training and 20 test records per school. A public spectral surrogate selects a representation from public ridge-coefficient estimates rather than minimizing (5) directly. Each method privately learns five projected coefficients and an intercept for 200 steps with batch size 20. Public proxy tasks select its threshold and step size. The School accountant uses the fixed-size without-replacement bound of Birrell et al. (2024) at $\delta = 10^{-5}$; Appendix C.3 specifies the physical noise convention and its distinction from the Wang-based synthetic accountant.

The original regularized spectral method improves NMSE over its nominally rank-five $\lambda = 0$ counterpart by 10.9%–13.3% for $\varepsilon \in \{1, 2, 4, 8\}$, but not at $\varepsilon = 0.5$. Examining effective dimension reveals why: the regularized representation has essentially one direction with appreciable feature variance, whereas the unregularized one has nearly five. A post-hoc rank-one baseline retains the leading predictive direction of the unregularized solution and fills the other four coordinates with publicly determined covariance-null directions. It preserves the same six-parameter private interface and Gaussian perturbations and receives its own public threshold/step-size calibration.

The rank-one baseline matches or improves on the regularized method's mean NMSE (Table 1). This attributes the original ILEA School gain primarily to lower effective predictive dimension, even though the nominal number of trained and perturbed parameters is unchanged. At the strictest privacy budget, $\varepsilon = 0.5$, neither the regularized method nor the rank-one baseline beats the original $\lambda = 0$ method.

CT Slice Localization. On numerical features from computed tomography (CT) slices (Graf et al., 2011), patients define tasks. The rank-five regularized representation has modest private-MSE reductions relative to $\lambda = 0$: 0.08%–0.97% in the primary table at $\varepsilon \in \{2, 3, 5, 8\}$. This comparison holds representation dimension fixed at five and measures the additional change in private prediction associated with regularization. Privacy protects individual training slices, not whole patients. Appendix C.4 gives the protocol and results.

Table 1: ILEA School private NMSE. The post-hoc rank-one baseline reproduces the regularized spectral method's advantage over the original $\lambda = 0$ representation. Appendix C.3.2 reports the differences within each of the five assignments and their standard errors.

Privacy budget ε	Original $\lambda = 0$	Regularized spectral	Rank-one $\lambda = 0$
1	1.119	0.997	0.977
2	1.102	0.961	0.943
4	1.101	0.958	0.935
8	1.100	0.954	0.934

6.3 PROJECTED-NORM EXPERIMENTS AND THE MAXIMUM PENALTY

A separate bounded anisotropic experiment examines AvgSq when the task coefficients lie in a shared rank-five subspace. The penalty reduces expected squared projected norms on new tasks, but does not improve private MSE; public selection chooses $\lambda = 0$. An isotropic version has the same expected squared projected norm for every representation, even when the empirical public penalty decreases (Appendix C.5).

For Max, we test whether reducing the largest projected norm improves private prediction under the coverage condition of Section 5.3. We generate signed, scaled coordinate vectors so that the expected squared projected norm is constant across representations, while the maximum can vary. With thresholds and step sizes selected on public tasks rather than prescribed by C_U^{cert}, Max reduces the maximum and has slightly lower mean MSE than unregularized adaptation at privacy budgets $\varepsilon = 2$ and 8. However, variation across repeated task collections is too large to establish a private-MSE improvement. A separate 24-seed experiment with an expanded learning-rate grid gives the same conclusion. Appendix D.5 gives both comparisons and the coverage checks.

7 DISCUSSION AND LIMITATIONS

The most useful representation for private adaptation need not minimize non-private prediction risk. The task-instability experiment illustrates a favorable trade-off between representation approximation and private adaptation. On ILEA School, the spectral method instead gains through lower effective predictive dimension. These experiments identify different ways in which public representation design can improve the accuracy of later private learning.

The analysis concerns linear predictors and squared loss, with explicit boundedness and coverage assumptions for the strongest guarantees. It retains public optimization error without proving global convergence of the nonconvex public optimizer. The experiments also examine variants beyond these assumptions: the positive synthetic experiment learns five representation directions for task coefficients spanning six directions. The shared-representation methods on ILEA School and CT use the learned features Ux, but do not project the private task vector v back into a fixed norm ball after each update. This additional task-vector projection is assumed by the private-risk theorems. Extending the public representation-learning stage to nonlinear representations and testing transfer under public/private distribution shift are natural next steps.

8 CONCLUSION

Privacy preconditioning makes the accuracy of a later private stage a *concrete design consideration in public training*. The public stage need not handle private records or satisfy a privacy constraint of its own. Instead, it deliberately learns a representation that will support more accurate adaptation once clipping and privacy noise become necessary.

We develop this principle through average-squared and maximum projected-norm penalties. The theory links public prediction risk and projected feature norms to private excess risk; the coverage-based result gives a direct bound on all private projected norms. In the synthetic task-instability experiment, the average penalty improves private prediction despite slightly worse non-private fit. The ILEA School experiment instead attributes its gain to lower effective dimension, as shown by an unregularized rank-one comparison. The remaining synthetic experiments delimit when reducing projected norms is useful.

More broadly, privacy preconditioning treats public preparation and private computation as a joint design problem. The public model or intermediate output is chosen to improve the utility of the private procedure that follows, without accessing private records or consuming their privacy budget during preparation.

Reproducibility Statement

The appendices specify the algorithms, assumptions, proofs, data construction, public-only selection procedures, privacy accountants, and experimental comparisons. All numerical results are associated with the stated sampling and output conventions. The accompanying code records the configurations, split indices, and scripts for the reported comparisons.

AI Use Statement

Generative AI tools assisted with mathematical checking, experimental design and interpretation, code development, literature checks, and editing. The authors reviewed and revised the AI-assisted material and are responsible for the final statements, proofs, code, and results.

References

Martín 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*, pp. 308–318. ACM, 2016. doi: 10.1145/2976749.2978318. URL https://doi.org/10.1145/2976749.2978318.

Ehsan Amid, Arun Ganesh, Rajiv Mathews, Swaroop Ramaswamy, Shuang Song, Thomas Steinke, Vinith M. Suriyakumar, Om Thakkar, and Abhradeep Thakurta. Public data-assisted mirror descent for private model training. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 517–535. PMLR, 2022. URL https://proceedings.mlr.press/v162/amid22a.html.

Galen Andrew, Om Thakkar, Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17455–17466. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/91cff01af640a24e7f9f7a5ab407889f-Abstract.html.

Borja Balle, Gilles Barthe, Marco Gaboardi, Justin Hsu, and Tetsuya Sato. Hypothesis testing interpretations and Rényi differential privacy. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2496–2506. PMLR, 2020. URL https://proceedings.mlr.press/v108/balle20a.html.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 464–473. IEEE, 2014. doi: 10.1109/FOCS.2014.56. URL https://doi.org/10.1109/FOCS.2014.56.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, volume 32, pp. 11282–11291. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/3bd8fdb090f1f5eb66a00c84dbc5ad51-Abstract.html.

Jeremiah Birrell, Reza Ebrahimi, Rouzbeh Behnia, and Jason Pacheco. Differentially private stochastic gradient descent with fixed-size minibatches: Tighter RDP guarantees with or without replacement. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0354. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/14fef58f09f2ebe69306e0a322e3be2b-Abstract-Conference.html.

Edwin V. Bonilla, Kian M. Chai, and Christopher K. I. Williams. Multi-task gaussian process prediction. In *Advances in Neural Information Processing Systems*, volume 20, pp. 153–160. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/hash/66368270ffd51418ec58bd793f2d9b1b-Abstract.html.

Zhiqi Bu, Xinwei Zhang, Sheng Zha, Mingyi Hong, and George Karypis. Pre-training differentially private models with limited public data. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3001. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ac04e54e0a2d1927d60709019e4e7870-Abstract-Conference.html.

Xiangyi Chen, Steven Wu, and Mingyi Hong. Understanding gradient clipping in private SGD: A geometric perspective. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13773–13782. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/9ecff5455677b38d19f49ce658ef0608-Abstract.html.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2089–2099. PMLR, 2021. URL https://proceedings.mlr.press/v139/collins21a.html.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, volume 3876 of *Lecture Notes in Computer Science*, pp. 265–284. Springer, 2006. doi: 10.1007/11681878_14. URL https://doi.org/10.1007/11681878_14.

Arun Ganesh, Mahdi Haghifam, Milad Nasr, Sewoong Oh, Thomas Steinke, Om Thakkar, Abhradeep Guha Thakurta, and Lun Wang. Why is public pretraining necessary for private model training? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10611–10627. PMLR, 2023. URL https://proceedings.mlr.press/v202/ganesh23a.html.

F. Graf, H.-P. Kriegel, M. Schubert, S. Poelsterl, and A. Cavallaro. Relative location of CT slices on axial axis. UCI Machine Learning Repository, 2011. URL https://doi.org/10.24432/C5CP6G. Dataset.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Prateek Jain, John Rush, Adam Smith, Shuang Song, and Abhradeep Guha Thakurta. Differentially private model personalization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29723–29735, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f8580959e35cb0934479bb007fb241c2-Abstract.html.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=bVuP3ltATMz.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016. URL https://jmlr.org/papers/v17/15-242.html.

Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics, 1989*, volume 141 of *London Mathematical Society Lecture Note Series*, pp. 148–188. Cambridge University Press, 1989.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275. IEEE, 2017. doi: 10.1109/CSF.2017.11. URL https://doi.org/10.1109/CSF.2017.11.

Milad Nasr, Saeed Mahloujifar, Xinyu Tang, Prateek Mittal, and Amir Houmansadr. Effectively using public data in privacy preserving machine learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25718–25732. PMLR, 2023. URL https://proceedings.mlr.press/v202/nasr23a.html.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.

Conor Snedeker, Xinyu Zhou, and Raef Bassily. Private model personalization revisited. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 56018–56067. PMLR, 2025. URL https://proceedings.mlr.press/v267/snedeker25a.html.

Shuang Song, Kamalika Chaudhuri, and Anand D. Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pp. 245–248. IEEE, 2013. doi: 10.1109/GlobalSIP.2013.6736861. URL https://cseweb.ucsd.edu/~kamalika/pubs/scs13.pdf.

Pratiksha Thaker, Amrith Setlur, Zhiwei Steven Wu, and Virginia Smith. On the benefits of public representations for private transfer learning under distribution shift. In *Advances in Neural Information Processing Systems*, volume 37, pp. 27088–27120. Curran Associates, Inc., 2024. doi: 10.52202/079017-0851. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f8f01acb3df19fea82fa68b4a310058-Abstract-Conference.html.

Florian Tramèr and Dan Boneh. Differentially private learning needs better features (or much more data). In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YTWGvpFOQD-.

Nilesh Tripuraneni, Chi Jin, and Michael I. Jordan. Provable meta-learning of linear representations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10434–10443. PMLR, 2021. URL https://proceedings.mlr.press/v139/tripuraneni21a.html.

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1226–1235. PMLR, 2019. URL https://proceedings.mlr.press/v89/wang19b.html.

Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. Large scale private learning via low-rank reparametrization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12208–12218. PMLR, 2021. URL https://proceedings.mlr.press/v139/yu21f.html.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Q42f0dfjECO.

A Algorithms and Computational Details

This appendix specifies the public solver for AvgSq, the average-squared projected-feature penalty in (5), and the two private update rules. The Max penalty in (6) uses the pooled-maximum update in Appendix D.3. The ILEA School spectral implementation is specified separately in Appendix C.3.

A.1 Public initialization and updates

The moment initializer uses the cross-example statistic of Snedeker et al. (2025), now computed on public tasks without privacy noise or user-level clipping. For $n_j \geq 2$, set

$$Z_j = \frac{\sum_{a \neq b} y_{j,a} y_{j,b} x_{j,a} x_{j,b}^\top}{n_j(n_j - 1)}, \qquad Z = \frac{1}{m} \sum_j Z_j.$$

The leading k left singular vectors of Z, transposed, give U_0. This is an initializer, not a separate recovery guarantee under heterogeneous feature distributions. Task heads may start at zero or a public-only value fixed by the experimental configuration.

For a task minibatch B_j, write $\widehat{J}_j = \widehat{L}_j + \lambda |B_j|^{-1} \sum_{(x,y) \in B_j} \|Ux\|_2^2$ for its AvgSq objective. The representation gradient is

$$\nabla_U \widehat{J}_j(U, v_j; B_j) = \frac{1}{|B_j|} \sum_{(x,y) \in B_j} \left[-2(y - v_j^\top U x) v_j x^\top + 2\lambda U x x^\top \right]. \tag{14}$$

The shared update averages these gradients equally across tasks. The local head steps differentiate only the prediction loss, since the penalty does not depend on v_j.

Algorithm 1 Public shared-representation learning (AvgSq)

Require: Public tasks S_j, dimension k, penalty λ, step sizes η_U, η_v, epochs T_{meta}, local steps K_v

```
Initialize U_0 by the public moment statistic and initialize heads v_{j,0}
for t = 0, ..., T_meta − 1 do
    for each task j do
        Draw a task minibatch; take K_v prediction-loss steps in v_j with U_t fixed
    end for
    Draw fresh task minibatches and average (14) to obtain G_t
    Ũ ← U_t − η_U G_t; factor Ũ^⊤ = QR; set U_{t+1} ← Q^⊤
end for
return U_{T_meta} and the final public task heads
```

A.2 Private update rules

For a fixed public representation U, define $g_i(v) = -2(y_i - v^\top U x_i) U x_i$ and $\mathrm{clip}_C(g) = g / \max\{1, \|g\|_2 / C\}$. Both routines use the fixed projected features $U x_i$. Algorithm 2 updates the task vector v without a norm constraint; Algorithm 3 additionally projects v onto the ball $\mathcal{B}_v$ after each update. Thus task-vector projection is distinct from applying the feature map U. Every minibatch is an independent uniform size-b subset of the same private records. Initialization uses only public information; it is the mean public head or zero, as specified by the experiment.

Here $\Pi_{\mathcal{B}_v}(u) = u \min\{1, B_v / \|u\|_2\}$, with $\Pi_{\mathcal{B}_v}(0) = 0$. Projection and averaging are post-processing of private updates. The Max experiment averages $v_1, \ldots, v_T$ instead; its empirical output convention differs from the theorem's $v_0, \ldots, v_{T-1}$ average.

A.3 Complexity

Let $N = \sum_j n_j$. With $s_j = \sum_i y_{j,i} x_{j,i}$, the numerator of Z_j equals $s_j s_j^\top - \sum_i y_{j,i}^2 x_{j,i} x_{j,i}^\top$. This computes the moment in $O(Nd^2)$ time without enumerating pairs. A rank-k iterative singular-value decomposition costs $O(d^2 k)$ up to its iteration count; a full dense decomposition costs $O(d^3)$. Working memory is $O(d^2 + kd)$.

Algorithm 2 Private adaptation with a frozen representation

Require: Private records $\{(x_i, y_i)\}_{i=1}^n$, public $U, v_0, C > 0, \eta, b, T$, calibrated σ
1: **for** $t = 0, \dots, T-1$ **do**
2: Draw a uniform size-b subset $B_t \subseteq [n]$, independently of earlier draws
3: $\bar{g}_t \leftarrow b^{-1} \sum_{i \in B_t} \mathrm{clip}_C(g_i(v_t))$
4: Draw $\xi_t \sim \mathcal{N}(0, \sigma^2 C^2 I_k / b^2)$; set $v_{t+1} \leftarrow v_t - \eta(\bar{g}_t + \xi_t)$
5: **end for**
6: **return** v_T

Algorithm 3 Private adaptation with task-vector projection and output variants

Require: The inputs of Algorithm 2, a public radius $B_v > 0$, $v_0 \in \mathcal{B}_v$
1: Use the same sampling, clipped gradient, and noise as Algorithm 2
2: At every step replace its update by $v_{t+1} \leftarrow \Pi_{\mathcal{B}_v}[v_t - \eta(\bar{g}_t + \xi_t)]$
3: For Theorem 1, return v_T
4: For Theorem 2, return $T^{-1} \sum_{t=0}^{T-1} v_t$

Using size-b public batches, one AvgSq epoch costs $O(m(K_v + 1)bkd + dk^2)$, including QR, and streaming working memory is $O(kd + mk + b(d + k))$. Exact Max scans add $O(Nkd)$ per scan. Private adaptation costs $O(Tbkd)$ with on-demand projections, or $O(nkd + Tbk)$ after caching projected features in $O(nk)$ space. Projection onto the head ball adds only $O(k)$ per step. These costs exclude dataset storage.

B Public Recovery and Private Risk

This appendix proves the AvgSq recovery and transfer results and the private-risk bound of Theorem 1. We retain the notation $\widehat{J}_\lambda$ and $\widehat{R}_{\mathrm{proj}}$ from (5); AvgSq is the name of this penalty, not a different risk or objective. Probability statements distinguish sampling tasks from sampling their records. The Max proof for $\widehat{J}_\lambda^{\max}$ and its coverage assumption are in Appendix D.2.

B.1 Population recovery and empirical optimization

Fix the public task distributions in the shared model (2). Write $w_j^\star = (U^\star)^\top v_j^\star$ and $\|z\|_\Sigma^2 = z^\top \Sigma z$. Centered model noise gives

$$L_j(U, v) = \tau^2 + \|w_j^\star - U^\top v\|_{\Sigma_j}^2. \tag{15}$$

If $\Sigma_j \succ 0$, the optimal head and representation-only loss are

$$v_j^\star(U) = (U\Sigma_j U^\top)^{-1} U \Sigma_j w_j^\star, \quad L_j(U) = \tau^2 + (w_j^\star)^\top \Sigma_j w_j^\star - (U\Sigma_j w_j^\star)^\top (U\Sigma_j U^\top)^{-1} U\Sigma_j w_j^\star.$$

These follow by differentiating the quadratic in v and completing its square.

Assumption 1 (Feature excitation and task diversity)**.** *For positive α, γ, every $\Sigma_j \succeq \alpha I_d$ and $G_V = m^{-1} \sum_j v_j^\star (v_j^\star)^\top \succeq \gamma I_k$.*

Theorem 3 (Population risk controls subspace error)**.** *Under Assumption 1, every $U \in \mathcal{U}$ and V satisfy*

$$L_{\mathrm{pub}}(U, V) - \tau^2 \geq L_{\mathrm{pub}}(U) - \tau^2 \geq \frac{\alpha\gamma}{2} \|P_U - P_{U^\star}\|_F^2.$$

Proof. For any head, (15) is at least $\tau^2 + \alpha \|(I - P_U) w_j^\star\|_2^2$, since the component orthogonal to the row space cannot be fitted. Averaging this component over tasks gives

$$\frac{1}{m} \sum_j \|(I - P_U)(U^\star)^\top v_j^\star\|_2^2 = \mathrm{Tr}\big(G_V U^\star (I - P_U)(U^\star)^\top\big) \geq \gamma \,\mathrm{Tr}(P_{U^\star}(I - P_U)).$$

Both projectors have rank k, so the last trace is $\frac{1}{2}\|P_U - P_{U^\star}\|_F^2$. This proves the result for arbitrary heads and hence for their infimum. At the population minimum, the subspaces coincide and predictors differ only by an orthonormal change of latent coordinates. □

Fix $\mathcal{V}_B = \{V : \max_j \|v_j\|_2 \le B_v\}$ containing $V^\star$. Let $\Delta_B = \sup_{U\in\mathcal{U},V\in\mathcal{V}_B} |J_\lambda(U,V) - \widehat{J}_\lambda(U,V)|$, and suppose $\widehat{J}_\lambda(\widehat{U},\widehat{V}) \le \inf_{(U,V)\in\mathcal{U}\times\mathcal{V}_B} \widehat{J}_\lambda(U,V) + \xi_{\mathrm{opt}}$. For any comparator in this class, empirical approximate optimality and two deviations give

$$J_\lambda(\widehat{U},\widehat{V}) \le J_\lambda(U,V) + 2\Delta_B + \xi_{\mathrm{opt}}. \tag{16}$$

Choosing $(U^\star, V^\star)$ and dropping the nonnegative learned penalty proves (8). Alternatively, retaining it gives

$$L_{\mathrm{pub}}(\widehat{U},\widehat{V}) - \tau^2 + \lambda\left[R_{\mathrm{proj}}(\widehat{U}) - R_{\mathrm{proj}}(U^\star)\right] \le 2\Delta_B + \xi_{\mathrm{opt}}. \tag{17}$$

The public certificate requires task-head norms at most B_v; the public training routine does not impose this constraint automatically. A fitted finite candidate set can instead be assessed using the validation result below.

B.2 Explicit public concentration and refined certificates

Assume every public distribution satisfies $\|x\|_2 \le X$ and $|y| \le Y$. Define $\chi(\delta) = 1 + \sqrt{2\log(1/\delta)}$.

Lemma 1 (Weighted bounded-vector concentration)**.** *Let independent vectors Z_i in a Euclidean space satisfy $\|Z_i\| \le M$, and let deterministic weights be a_i. With probability at least $1-\delta$,*

$$\left\|\sum_i a_i(Z_i - \mathbb{E}Z_i)\right\| \le M\sqrt{\sum_i a_i^2}\,\chi(\delta).$$

The statement also applies to matrices with the Frobenius norm.

Proof. Independence and centering imply the expected squared norm is at most $M^2\sum_i a_i^2$, so Jensen bounds the expected norm by its square root. Replacing one Z_i changes the norm by at most $2M|a_i|$. The bounded-differences inequality (McDiarmid, 1989) bounds the excess above its expectation by $M\sqrt{2\log(1/\delta)\sum_i a_i^2}$. □

Proposition 1 (A bound on the public objective deviation)**.** *Conditional on the public tasks, with probability at least $1-\delta$, simultaneously for all $\lambda \ge 0$,*

$$\Delta_B \le \overline{\Delta}_B(\lambda,\delta) := \chi\!\left(\frac{\delta}{3m+1}\right)\left[\frac{(Y+B_vX)^2}{m}\sum_j \frac{1}{\sqrt{n_j}} + \frac{\lambda\sqrt{k}X^2}{m}\sqrt{\sum_j \frac{1}{n_j}}\right]. \tag{18}$$

Proof. For each task, expand its prediction loss using the moments $a_j = \mathbb{E}y^2$, $c_j = \mathbb{E}[yx]$, and $\Sigma_j = \mathbb{E}[xx^\top]$, and their empirical counterparts:

$$|L_j(U,v_j) - \widehat{L}_j(U,v_j)| \le |a_j - \widehat{a}_j| + 2B_v\|c_j - \widehat{c}_j\|_2 + B_v^2\|\Sigma_j - \widehat{\Sigma}_j\|_F.$$

The summands have bounds Y^2, YX, X^2. Apply the lemma separately to these three moments for every task. The resulting weighted average is bounded by the first term in (18). For the regularizer, $\|P_U\|_F = \sqrt{k}$, so $|\operatorname{Tr}(P_U(\widehat{\Sigma}_x - \bar{\Sigma}))| \le \sqrt{k}\|\widehat{\Sigma}_x - \bar{\Sigma}\|_F$. Apply the lemma to all public records with weights $1/(mn_j)$; their squared weights sum to $m^{-2}\sum_j 1/n_j$. A union bound over the $3m+1$ events proves the result. The event concerns moments only, so the inequality holds for all U, V, λ simultaneously. □

Anisotropy-sensitive recovery. Let $D = \|P_{\widehat{U}} - P_{U^\star}\|_F$, $c_0 = \alpha\gamma/2$, $E_\lambda = 2\Delta_B + \xi_{\mathrm{opt}}$, and $A_\Sigma = \inf_{s\in\mathbb{R}}\|\bar{\Sigma} - sI_d\|_F$. The projector difference has trace zero, hence

$$|R_{\mathrm{proj}}(\widehat{U}) - R_{\mathrm{proj}}(U^\star)| = |\operatorname{Tr}((P_{\widehat{U}} - P_{U^\star})(\bar{\Sigma} - sI_d))| \le D\|\bar{\Sigma} - sI_d\|_F.$$

Combining its infimum with (17) gives $c_0D^2 - \lambda A_\Sigma D \le E_\lambda$, and therefore

$$D \le \frac{\lambda A_\Sigma + \sqrt{\lambda^2A_\Sigma^2 + 4c_0E_\lambda}}{2c_0}. \tag{19}$$

This refinement removes an artificial population penalty cost when $\bar{\Sigma}$ is isotropic. It does not remove finite-sample or public optimization error.

Held-out validation.

Proposition 2 (Validation subspace certificate)**.** *For each realized public task reserve r_j independent validation records. Let K_{cand} candidate pairs $(U^{(c)}, V^{(c)}) \in \mathcal{U} \times \mathcal{V}_B$ be fitted without these records. With probability at least $1 - \delta_{\mathrm{val}}$, every candidate satisfies*

$$\|P_{U^{(c)}} - P_{U^\star}\|_F^2 \le \frac{2}{\alpha\gamma}\left[\widehat{L}_{\mathrm{val}}(U^{(c)}, V^{(c)}) + \epsilon_{\mathrm{val}} - \tau^2\right],$$

$$\epsilon_{\mathrm{val}} = (Y + B_v X)^2 \sqrt{\frac{\log(K_{\mathrm{cand}}/\delta_{\mathrm{val}})}{2m^2} \sum_j \frac{1}{r_j}}.$$

where $\widehat{L}_{\mathrm{val}}$ averages each task's validation loss with weight $1/m$.

Proof. Conditional on the tasks and fitted candidates, each validation loss lies in $[0, (Y + B_v X)^2]$. Weighted Hoeffding concentration (Hoeffding, 1963), with record weights $1/(mr_j)$, gives the displayed one-sided deviation for one candidate with failure $\delta_{\mathrm{val}}/K_{\mathrm{cand}}$. Union bound over candidates, then apply Theorem 3. □

A known lower bound on τ^2 can replace it. This certificate is independent of ξ_{opt}, but the validation records must remain independent of every fitted candidate being certified. Its result can be capped at the deterministic $2k$ subspace bound.

B.3 Transfer across tasks

Let public tasks be independent draws from $\mathcal{Q}$ and let records be independent conditional on their tasks. Define $\Sigma_{\mathrm{meta}} = \mathbb{E}_\theta \Sigma_\theta$ and $R_{\mathrm{meta}}(U) = \mathrm{Tr}(P_U \Sigma_{\mathrm{meta}})$.

Proposition 3 (Two-level projected-moment transfer)**.** *With bounded features,* (9) *holds with probability at least $1 - \delta_R$. On this event, for every public-only representation U and an independent new task,*

$$\mathbb{E}_\theta R_\theta(U) = R_{\mathrm{meta}}(U) \le r_{\mathrm{pub}}(U), \qquad \mathbb{E}_\theta R_\theta(U)^2 \le X^2 r_{\mathrm{pub}}(U), \quad R_\theta(U) = \mathrm{Tr}(U\Sigma_\theta U^\top).$$

Proof. Decompose $\widehat{\Sigma}_x - \Sigma_{\mathrm{meta}} = (\widehat{\Sigma}_x - \bar{\Sigma}) + (\bar{\Sigma} - \Sigma_{\mathrm{meta}})$. Conditional on the tasks, the first term is a weighted sum of record matrices and is bounded in Frobenius norm by $X^2 \chi(\delta_R/2) m^{-1} \sqrt{\sum_j 1/n_j}$. For the second, use independent task matrices Σ_j of Frobenius norm at most X^2 to obtain $X^2\chi(\delta_R/2)/\sqrt{m}$. A union bound and $\|P_U\|_F = \sqrt{k}$ yield (9). Since $0 \le R_\theta(U) \le X^2$, also $R_\theta(U)^2 \le X^2 R_\theta(U)$; the last conclusions follow. □

The event is uniform in U and thus includes public hyperparameter selection. In particular,

$$R_{\mathrm{meta}}(\widehat{U}_\lambda) - R_{\mathrm{meta}}(\widehat{U}_0) \le \widehat{R}_{\mathrm{proj}}(\widehat{U}_\lambda) - \widehat{R}_{\mathrm{proj}}(\widehat{U}_0) + 2\Gamma_R.$$

For exact empirical minimizers, optimality at two penalty levels $\lambda_1 < \lambda_2$ also gives $\widehat{R}_{\mathrm{proj}}(\widehat{U}_{\lambda_2}) \le \widehat{R}_{\mathrm{proj}}(\widehat{U}_{\lambda_1})$ by adding their two optimality inequalities. With optimization errors, the right side acquires $(\xi_1 + \xi_2)/(\lambda_2 - \lambda_1)$. Neither statement alone asserts improved private prediction.

B.4 Private approximation, clipping, and noise

For a fixed task and U, put $w^\star = (U^\star)^\top v^\star$, $H_U = U\Sigma_\theta U^\top$, $R_U = \mathrm{Tr}(H_U)$, $\mu_U = \lambda_{\min}(H_U)$ and $L_U = \lambda_{\max}(H_U)$. When $H_U \succ 0$, completing the square yields

$$v_U^\star = H_U^{-1} U \Sigma_\theta w^\star, \qquad L_\theta(U, v) - L_\theta(U) = (v - v_U^\star)^\top H_U (v - v_U^\star). \tag{20}$$

The feasible comparator $v^\circ = U w^\star$ satisfies $\|v^\circ\|_2 \le \|v^\star\|_2$, and

$$L_\theta(U) - \tau^2 \le A_\theta(U) := \mathbb{E}_x\left[\left(((I - P_U)w^\star)^\top x\right)^2\right] \le \frac{\beta}{2}\|v^\star\|_2^2 \|P_U - P_{U^\star}\|_F^2 \tag{21}$$

if $\Sigma_\theta \preceq \beta I_d$. To see the last inequality, bound by $\beta\|(I - P_U)(U^\star)^\top v^\star\|_2^2$, then use $\|(I - P_U)(U^\star)^\top\|_2^2 \le \|(I - P_U)(U^\star)^\top\|_F^2 = \frac{1}{2}\|P_U - P_{U^\star}\|_F^2$.

Assumption 2 (Bounded private model)**.** *Features satisfy $\|x\|_2 \le X$, labels satisfy $|y| \le Y$, and every private iterate lies in $\mathcal{B}_v = \{v : \|v\|_2 \le B_v\}$. The bounds are fixed independently of the private sample. In the shared model, $\|v^\star\|_2 \le B_\star$ and $|\epsilon| \le B_\epsilon$ imply the sufficient label bound $Y = B_\star X + B_\epsilon$.*

Residual and clipping bounds. For every $v \in \mathcal{B}_v$, $|y - v^\top U x| \le B_r := Y + B_v X$. Thus, with $g = -2(y - v^\top U x)Ux$,

$$\|g - \mathrm{clip}_C(g)\|_2 = (\|g\|_2 - C)_+, \quad \|g\|_2^2 \le 4B_r^2\|Ux\|_2^2, \quad \|\mathbb{E}[\mathrm{clip}_C(g) - g]\|_2 \le \frac{B_r^2 R_U}{C}.$$

The last step uses $(a-C)_+ \le a^2/(4C)$ for $a \ge 0$ and Jensen's inequality. The residual bound holds at every iterate when the task vector is projected into $\mathcal{B}_v$ after each update.

Sensitivity. Condition on the public choices and preceding private outputs. Replacing a record changes the clipped batch sum by at most $2C$, or its average by $2C/b$. Gaussian noise of standard deviation $\sigma C/b$ on the average therefore has ratio $\sigma/2$ to this sensitivity. An accountant for the actual sampling and composition establishes privacy; C cancels from the ratio, but its absolute size affects both clipping error and noise magnitude.

B.4.1 Exact Gaussian-noise accumulation in a population recursion

Proposition 4 (One-step and finite-horizon noise cost)**.** *Consider the population-gradient update without clipping or task-vector projection $v_{t+1} = v_t - 2\eta H_U(v_t - v_U^\star) - \eta\xi_t$, with independent centered $\xi_t \sim \mathcal{N}(0, s_U^2 I_k)$, $s_U = \sigma C_U/b$. For a fixed pre-noise iterate, its one-step additional expected loss is $\eta^2 s_U^2 \operatorname{Tr}(H_U)$. If $H_U \succ 0$ has eigenvalues ν_i and $0 < \eta < 1/L_U$, the noise contribution after T steps is*

$$\eta^2 s_U^2 \sum_{i=1}^{k} \nu_i \sum_{t=0}^{T-1} (1 - 2\eta\nu_i)^{2t} = \frac{\eta s_U^2}{4} \sum_{i=1}^{k} \frac{1 - (1 - 2\eta\nu_i)^{2T}}{1 - \eta\nu_i}. \tag{22}$$

Its limit is $\frac{\eta s_U^2}{4}\sum_i (1 - \eta\nu_i)^{-1}$.

Proof. With $e_t = v_t - v_U^\star$ and $A = I - 2\eta H_U$, $e_T = A^T e_0 - \eta\sum_{t=0}^{T-1} A^{T-1-t}\xi_t$. Cross terms vanish by independence and centering. The loss is $e_T^\top H_U e_T$, so the added contribution is $\eta^2 s_U^2 \sum_{t=0}^{T-1} \operatorname{Tr}(H_U A^{2t})$. Diagonalize H_U and sum the geometric series, using $1 - (1 - 2\eta\nu_i)^2 = 4\eta\nu_i(1 - \eta\nu_i)$. The one-step identity is the case with a single perturbation. Stability gives the limit. □

This calculation explains why the cost of identical parameter noise can depend on the representation. It also shows why reducing a one-step trace term does not establish the total effect of a training trajectory. Representation-dependent C_U remains in every expression. The analysis below includes clipping, task-vector projection, and minibatch updates on the same private dataset.

B.4.2 Gradient moments with independent symmetric coordinates

Proposition 5 (Gradient second moment)**.** *Let $x = Dz$, where D is deterministic and the coordinates of z are independent, symmetric, have variance one and fourth moments $\kappa_i < \infty$. Let $y = (w^\star)^\top x + \epsilon$, with independent centered noise of variance τ^2. For a fixed head define $a = w^\star - U^\top v$, $h = D^\top a$, $M = D^\top P_U D$, and $\Sigma = DD^\top$. Then*

$$\mathbb{E}\|\nabla_v \ell\|_2^2 = 4\left[(a^\top\Sigma a + \tau^2)\operatorname{Tr}(U\Sigma U^\top) + 2a^\top\Sigma P_U\Sigma a + \sum_i (\kappa_i - 3)h_i^2 M_{ii}\right].$$

Gaussian coordinates have $\kappa_i = 3$, giving the Gaussian specialization without the last term.

Proof. The squared gradient is $4(h^\top z + \epsilon)^2 z^\top M z$. Independent centered noise contributes $4\tau^2\operatorname{Tr}(M)$ and eliminates its cross term. In the fourth-moment expansion of $\mathbb{E}[(h^\top z)^2 z^\top M z]$, independence and symmetry leave only paired indices. They give $\|h\|_2^2\operatorname{Tr}(M) + 2h^\top M h$, with a diagonal correction $\sum_i (\kappa_i - 3)h_i^2 M_{ii}$. Substitution proves the formula. □

This identity includes the bounded uniform-coordinate generator, whose fourth moment is $9/5$. It does not substitute a Gaussian formula for non-Gaussian features.

B.5 Private optimization with task-vector projection and data reuse

We prove the task-specific result used by Theorem 1. Fix U and a task satisfying Assumption 2. Let $z_i = Ux_i$, K_U be a deterministic bound on $\|Ux\|_2$ (X always suffices), and set

$$\widehat{H}_U = \frac{1}{n}\sum_i z_i z_i^\top, \quad \widehat{c}_U = \frac{1}{n}\sum_i y_i z_i, \quad c_U = \mathbb{E}[yUx], \quad \widehat{R}_U = \frac{1}{n}\sum_i \|z_i\|_2^2,$$

$$e_n = 2B_v\|\widehat{H}_U - H_U\|_2 + 2\|\widehat{c}_U - c_U\|_2.$$

For the empirical clipped mean $\widehat{g}_C(v) = n^{-1}\sum_i \mathrm{clip}_C(g_i(v))$, define $d_S(v) = \widehat{g}_C(v) - \nabla L_\theta(U, v)$. The empirical unclipped gradient is $2\widehat{H}_U v - 2\widehat{c}_U$. Its difference from the population gradient is uniformly bounded by e_n. The clipping inequality above applies to the empirical sample as well, hence

$$\sup_{v\in\mathcal{B}_v}\|d_S(v)\|_2 \le \frac{B_r^2\widehat{R}_U}{C} + e_n, \qquad \sup_{v\in\mathcal{B}_v}\|d_S(v)\|_2^2 \le 2\left(\frac{B_r^4\widehat{R}_U^2}{C^2} + e_n^2\right). \tag{23}$$

Theorem 4 (Task-specific private risk). *Suppose $H_U \succ 0$, $v_U^\star \in \mathcal{B}_v$, and $v_0 \in \mathcal{B}_v$ is public-only. For Algorithm 3 returning v_T, let $0 < \eta \le 1/(2L_U)$, $a_\zeta = 1 + 1/\zeta$, and $q_U = (1+\zeta)(1-2\eta\mu_U)^2 < 1$ for some $\zeta > 0$. Set $S_T(q_U) = \sum_{t=0}^{T-1} q_U^t$ and $\nu_{n,b} = (n-b)/(b(n-1))$. Conditional on the public stage and the task distribution,*

$$\begin{aligned}\mathbb{E}[L_\theta(U, v_T) - L_\theta(U)] \le{}& L_U q_U^T\|v_0 - v_U^\star\|_2^2 \\ &+ L_U\eta^2 S_T(q_U)\left[\frac{2a_\zeta B_r^4}{C^2}\left(R_U^2 + \frac{K_U^2 R_U}{n}\right)\right. \\ &\left.+\frac{16a_\zeta(B_v^2K_U^2 + Y^2)R_U}{n} + \left(\nu_{n,b} + \frac{k\sigma^2}{b^2}\right)C^2\right].\end{aligned} \tag{24}$$

Proof. First condition on the realized dataset and the history before step t. Write $\zeta_t = \bar{g}_t - \widehat{g}_C(v_t)$. Uniform independent subset draws make this centered. For fixed vectors a_i with average $\bar{a}$, sampling without replacement gives

$$\mathbb{E}\left\|\frac{1}{b}\sum_{i\in B_t} a_i - \bar{a}\right\|_2^2 = \frac{n-b}{b(n-1)}\frac{1}{n}\sum_i \|a_i - \bar{a}\|_2^2.$$

For clipped gradients, this is at most $\nu_{n,b}C^2$. The Gaussian second moment is $k\sigma^2C^2/b^2$.

Let $h_t = v_t - v_U^\star$ and $A_U = I - 2\eta H_U$. Since $v_U^\star$ is fixed by projection, nonexpansiveness gives

$$\|h_{t+1}\|_2^2 \le \|A_U h_t - \eta d_S(v_t) - \eta\zeta_t - \eta\xi_t\|_2^2.$$

The centered independent perturbations have zero conditional cross terms. Young's inequality bounds the remaining term by $q_U\|h_t\|_2^2 + a_\zeta\eta^2\sup_{v\in\mathcal{B}_v}\|d_S(v)\|_2^2$. Iteration conditional on the fixed dataset yields

$$\mathbb{E}[\|h_T\|_2^2 \mid S] \le q_U^T\|h_0\|_2^2 + \eta^2 S_T(q_U)\left[2a_\zeta\left(\frac{B_r^4\widehat{R}_U^2}{C^2} + e_n^2\right) + \left(\nu_{n,b} + \frac{k\sigma^2}{b^2}\right)C^2\right].$$

Now average over the independent records. Because $\|z\|_2 \le K_U$,

$$\mathbb{E}\widehat{R}_U^2 \le R_U^2 + \frac{K_U^2 R_U}{n}, \qquad \mathbb{E}\|\widehat{c}_U - c_U\|_2^2 \le \frac{Y^2 R_U}{n},$$

$$\mathbb{E}\|\widehat{H}_U - H_U\|_2^2 \le \mathbb{E}\|\widehat{H}_U - H_U\|_F^2 \le \frac{K_U^2 R_U}{n}.$$

Thus $\mathbb{E}e_n^2 \le 8(B_v^2K_U^2 + Y^2)R_U/n$. Substitute these bounds and use $L_\theta(U, v_T) - L_\theta(U) \le L_U\|h_T\|_2^2$. Conditioning on the dataset before centering the minibatch error is what permits reuse; the argument does not assume fresh population examples at each step. □

At $b = n$ the subset variance vanishes, but the empirical-to-population terms remain. These terms are essential even when each iteration uses the entire private sample.

B.6 Public selection and the integrated guarantee

Lemma 2 (Public-only selection preserves privacy)**.** *Let all representation and adaptation choices be functions only of public data and randomness independent of the private dataset. If, conditional on every such choice, the private mechanism is (ε, δ)-DP for the same target parameters, the selected mechanism is also (ε, δ)-DP.*

Proof. Apply the DP inequality conditional on each public output and integrate both sides over its distribution, which is the same for neighboring private datasets. Projection and output averaging are deterministic post-processing. □

Proof of Theorem 1. The uniform lower covariance bound implies

$$\|v_U^\star\|_2 = \|(U\Sigma_\theta U^\top)^{-1} U\Sigma_\theta w_\theta^\star\|_2 \le \frac{\beta}{\mu_0}\|w_\theta^\star\|_2 \le B_v.$$

Thus the comparator containment required by the task-specific theorem holds. Also $L_U \le \beta$, $\mu_U \ge \mu_0$, $K_U \le X$, and $\|v_0 - v_U^\star\|_2^2 \le 4B_v^2$. Conditional on a public output in the transfer event, the preceding transfer proposition gives $\mathbb{E}_\theta R_U \le r_{\mathrm{pub}}(U)$ and $\mathbb{E}_\theta R_U^2 \le X^2 r_{\mathrm{pub}}(U)$. Use these in (24), bound its contraction and geometric sum by q and S_T of the main theorem, and add the approximation bound (21) with $\|P_U - P_{U^\star}\|_F^2 \le B_{\mathrm{sub}}$. The result is (11). Public recovery and transfer hold jointly with at least one minus the sum of their stated failure probabilities.

Analytical versus empirical thresholds. For fixed public quantities, the C-dependent part of (11) is $A/C^2 + BC^2$, with $A = 2aB_r^4X^2(1 + 1/n)r_{\mathrm{pub}}(U)$ and $B = \nu_{n,b} + k\sigma^2/b^2$. When $A, B > 0$, its minimizer is $(A/B)^{1/4}$. Clamping this value to any prescribed positive interval preserves privacy by the preceding lemma. This is a tuning rule for a sufficient bound, not a claim that its constants describe an empirically optimal threshold. The experiments instead use held-out public adaptation loss. This analytical clipping/noise balance is distinct from the no-clipping certificate for Max.

C Experimental Protocols and Supporting Results

We give the protocols and results for the task-instability experiment, ILEA School and its rank-one comparison, CT Slice Localization, and the bounded-anisotropic and isotropic experiments. AvgSq denotes the average-squared projected-feature penalty in (5); the regularized methods use this penalty unless a spectral variant is stated. The Max experiment uses (6) and appears in Appendix D.5. All hyperparameter selection uses public data.

C.1 Synthetic sampling, accounting, and exact risk

The two principal synthetic experiments use ambient dimension $d = 100$ and learned dimension $k = 5$. The base anisotropy profile has a_i^2 log-spaced on $[0.2, 2]$. A task draws independent $u_{\theta,i} \sim \mathrm{Unif}[0.5, 1.5]$ and generates base features $x = \mathrm{diag}(a \odot u_\theta)z$, with independent $z_i \sim \mathrm{Unif}[-\sqrt{3}, \sqrt{3}]$. Here $\odot$ is coordinatewise multiplication. Independent label noise is uniform on $[-\sqrt{3}, \sqrt{3}]$, with variance one. The task-instability experiment adds the feature spike and coefficient structure described below; the bounded-anisotropic experiment uses a uniformly distributed radius-three task head.

Public training uses 50 tasks, 64/16/20 percent training/validation/test splits within each task, 1,000 alternating epochs, five local head steps, and $\eta_U = \eta_v = 0.01$. In the joint-selection experiments, 10 held-out public calibration tasks with three adaptation repetitions select (λ, C) using query MSE after noisy adaptation on support records. The grids are $\{0, .01, .03, .1, .3, 1\}$ and $\{1, 2, 5, 10, 20, 40\}$, respectively. The unregularized comparison selects its own best C within the same public procedure. Settings are frozen before evaluation on 20 fresh tasks per outer seed.

Private adaptation uses $n = 80$ training records from a 100-record task, batch size $b = 8$, step size 0.01, and 100 updates, unless the horizon is explicitly varied. Each minibatch is an independent uniform subset of the same training set. Under replace-one adjacency, the clipped-average sensitivity is $2C/b$ and Gaussian standard deviation is $\sigma C/b$. The synthetic accountant uses the

fixed-size without-replacement bound of Wang et al. (2019) through `dp_accounting`, with noise-to-sensitivity parameter $\sigma/2$, additive Rényi composition, and the conversion of Balle et al. (2020) at $\delta = 1/80^2$. Table 2 states the attained bounds, including the small difference from a nominal budget caused by numerical calibration. Horizon comparisons recalibrate the multiplier for their own update count.

For each known synthetic task, with coefficient w_θ and second moment Σ_θ, exact population risk is

$$L_\theta(U, v) = 1 + (w_\theta - U^\top v)^\top \Sigma_\theta (w_\theta - U^\top v).$$

Let $H = U\Sigma_\theta U^\top$ and $v_U^\star = H^{-1} U \Sigma_\theta w_\theta$. The approximation error is $w_\theta^\top \Sigma_\theta w_\theta - (v_U^\star)^\top H v_U^\star$; the adaptation error is $(v - v_U^\star)^\top H (v - v_U^\star)$. This identity applies even when the true task family has dimension greater than the learned representation.

Table 2: Synthetic privacy calibration for $T = 100$, $b = 8$, and $n_{\text{train}} = 80$ (bound of Wang et al. (2019)). σ is the configured multiplier, the standard deviation of the noise added to the clipped sum divided by C; the accountant receives $\sigma/2$. The last column is the noise standard deviation on the averaged gradient at $C = 10$.

Target ε	Achieved ε	Configured σ	$\sigma C/b$
2	1.9999	7.7500	9.6875
3	3.0009	5.5508	6.9385
5	4.9993	3.7578	4.6973
8	7.9992	2.7246	3.4058

C.2 Task-instability family: generator and independent confirmation

A Gaussian 100×10 matrix is QR-factorized to obtain two row-orthonormal blocks $U_s, U_u \in \mathbb{R}^{5\times 100}$ with orthogonal row spaces. Unit vectors $\bar{v}_s, \bar{v}_u \in \mathbb{R}^5$ specify shared coefficient directions. A task generates

$$x = \operatorname{diag}(a \odot u_\theta) z + \omega U_u^\top \zeta, \qquad y = \langle v_{s,j}, U_s x \rangle + \gamma_j \langle \bar{v}_u, U_u x \rangle + \epsilon,$$

where ζ has five independent $\mathrm{Unif}[-\sqrt{3}, \sqrt{3}]$ coordinates, $v_{s,j} = 2\bar{v}_s + 0.4\rho_j$ with ρ_j uniform in the five-dimensional unit ball, and

$$\gamma_j = g(\sqrt{1-\alpha^2} + \alpha r_j), \quad r_j \sim \mathrm{Unif}[-\sqrt{3}, \sqrt{3}], \qquad g = 2\sqrt{\phi/R}.$$

Then $\mathbb{E}\gamma_j = g\sqrt{1-\alpha^2}$ and $\mathbb{E}\gamma_j^2 = g^2$. The instability parameter $\alpha \in [0, 1]$ changes the coefficient's cross-task variation while preserving this second moment and the feature law. Set $\omega^2 = (R-1)\operatorname{Tr}(U_u \Sigma_{\text{meta}}^{\text{clean}} U_u^\top)/5$, where $\Sigma_{\text{meta}}^{\text{clean}} = \frac{13}{12}\operatorname{diag}(a^2)$. This increases the *subspace-average* feature variance along U_u by the factor R. The studied regime is $\alpha = 1, R = 2, \phi = 0.5$, so $g = 1$. The parameter ϕ controls nominal relative predictive strength; it is not an exact label-variance ratio for each realized subspace and stable head. The realized ratios range from 0.37 to 0.59 at $\phi = 0.5$.

Predictive rank and exact moments. The task coefficient is $w_j = U_s^\top v_{s,j} + \gamma_j U_u^\top \bar{v}_u$. Stable-head variation spans five directions and the unstable label term adds one, so the predictive family is generically six-dimensional. The learned representation has rank five, so it cannot express all six predictive directions simultaneously. In particular, $\mathbb{E}[v_{s,j} v_{s,j}^\top] = 4\bar{v}_s \bar{v}_s^\top + (0.16/7) I_5$; we do not divide its predictive strength equally among five directions. The centered feature coordinates give

$$\Sigma_\theta = \operatorname{diag}(a_i^2 u_{\theta,i}^2) + \omega^2 U_u^\top U_u, \quad \Sigma_{\text{meta}} = \frac{13}{12}\operatorname{diag}(a^2) + \omega^2 U_u^\top U_u.$$

The support bound is $\|x\|_2 \le 1.5\sqrt{3}\|a\|_2 + \omega\sqrt{15}$. The public sample size is $n_j = 200$. Private adaptation uses the fixed features Ux, returns v_T, and does not project v onto a norm ball. The exact risk calculation above supplies approximation and adaptation errors without test-sample estimation noise.

Discovery and evaluation separation. An exploratory grid of 13 regimes used three discovery seeds. Two candidates were fixed before five validation seeds; the milder candidate $(\alpha, R, \phi) = (0.3, 2, 1)$ did not validate, while the second $(1, 2, 0.5)$ was carried forward. Neither discovery nor validation is the main evidence. Primary set R has 24 new seeds $81000003+104729i$ and replication set B has 24 disjoint seeds $91000003+104729i$, with $i = 0, \ldots, 23$. The regime, grids, and seed sets were internally frozen before their evaluation, not externally preregistered. The sets are not pooled. Tables report regularized-minus-unregularized differences; negative values favor regularization. For each outer seed, the two methods are evaluated on the same generated tasks. We calculate their MSE difference, then resample these seed-level differences 10,000 times to obtain 95% percentile-bootstrap intervals for the mean difference. The reported intervals are for individual comparisons.

Table 3: Task-instability family ($\alpha = 1$, $R = 2$, $\phi = 0.5$): exact population MSE (noise floor 1) at each representation's publicly selected (λ, C). Δ is regularized minus unregularized. λ: selected mode, with the number of seeds on which $\lambda = 0$ was selected. C and $\sigma C/b$: means over seeds (unregularized / regularized).

Set	ε	Unreg. MSE	Reg. MSE	Δ [95% interval]	$n_{<0}$	λ (mode; $\#\lambda{=}0$)	C	$\sigma C/b$
R (primary)	2	1.8749	1.7090	-0.1659 $[-0.1965, -0.1368]$	24/24	0.1; 0	3.75 / 4.13	3.63 / 4.00
R (primary)	8	1.3668	1.3365	-0.0303 $[-0.0398, -0.0209]$	21/24	0.1; 3	6.67 / 6.67	2.27 / 2.27
B (replication)	2	1.8987	1.7354	-0.1633 $[-0.2001, -0.1305]$	24/24	0.1; 0	3.38 / 4.00	3.27 / 3.88
B (replication)	8	1.3605	1.3332	-0.0274 $[-0.0351, -0.0198]$	21/24	0.1; 2	6.67 / 6.46	2.27 / 2.20
V5 (validation)	2	1.9415	1.8244	-0.1171 $[-0.2226, +0.0220]$	4/5	0.1; 0	3.80 / 4.40	3.68 / 4.26
V5 (validation)	8	1.3621	1.3247	-0.0374 $[-0.0584, -0.0155]$	4/5	0.1; 1	7.00 / 7.00	2.38 / 2.38

The primary MSE reduction is 0.1659 at $\varepsilon = 2$ and 0.0303 at $\varepsilon = 8$, with closely matching independent replication. Positive λ is selected on every seed at $\varepsilon = 2$, with modal value 0.1. Table 4 reports the budget and horizon comparisons.

Table 4: Set R privacy curve and horizons (24 seeds per row).

T	ε	σ	Unreg. MSE	Δ [95% interval]	$n_{<0}$
100	2	7.7500	1.8749	-0.1659 $[-0.1965, -0.1368]$	24
100	3	5.5508	1.5924	-0.1080 $[-0.1313, -0.0875]$	24
100	5	3.7578	1.4203	-0.0563 $[-0.0662, -0.0458]$	24
100	8	2.7246	1.3668	-0.0303 $[-0.0398, -0.0209]$	21
50	2	5.6328	1.8996	-0.1468 $[-0.1671, -0.1288]$	24
50	8	2.0093	1.4359	-0.0296 $[-0.0394, -0.0211]$	22
200	2	10.7813	1.7938	-0.1829 $[-0.2137, -0.1522]$	24
200	8	3.5840	1.3519	-0.0269 $[-0.0362, -0.0185]$	20

C.2.1 Approximation and private adaptation

At $\varepsilon = 2$, the primary set pays approximately 0.029 more approximation error but has 0.195 less adaptation error. The replication gives $+0.034$ and -0.198. With no clipping or Gaussian noise, regularization is slightly worse: $+0.0119$ in the primary set and $+0.0175$ in replication. The private improvement is therefore not simply a better ordinary predictor.

In the replication, mean per-example gradient norms fall from 5.49 to 4.05, the 99th percentile from 20.22 to 15.26, and clipping fraction from .610 to .434 at $\varepsilon = 2$. The absolute unstable-subspace contribution to projected variance falls from 1.893 to 1.559, but total projected variance falls more, from 6.746 to 3.990. Its fraction therefore rises from .281 to .393. These measurements support a change in adaptation geometry, not selective elimination of the unstable variance in relative terms.

C.2.2 Clipping and noise ablations

For each publicly chosen representation and threshold, the ablation runs retain data, initialization, step size, minibatch schedule, and horizon. CLIPPING ONLY removes Gaussian noise; NO CLIP/NO NOISE removes both. Define clipping cost as the former minus the latter and noise cost as the private loss minus clipping-only loss. Their between-method differences obey $\Delta_{\text{private}} = \Delta_{\text{no clip/no noise}} + \Delta_{\text{clipping cost}} + \Delta_{\text{noise cost}}$. This is an exact matched-run accounting identity, not

a general causal law. Different methods may select different thresholds; the next subsection also matches them.

Table 5: Between-method decomposition: regularized minus unregularized MSE. Each component gives its mean and 95% paired interval. A negative noise-cost difference means a smaller positive cost of adding noise, not a benefit from adding noise to the same method.

Set	ε	No clipping/no noise	Clipping-cost difference	Noise-cost difference
R	2	$+.0119$	$-.0961$	$-.0817$
		$[+.0067, +.0173]$	$[-.1558, -.0474]$	$[-.1339, -.0237]$
B	2	$+.0175$	$-.1294$	$-.0514$
		$[+.0072, +.0312]$	$[-.2073, -.0671]$	$[-.1081, +.0142]$
R	8	$+.0062$	$-.0176$	$-.0189$
		$[+.0016, +.0111]$	$[-.0276, -.0077]$	$[-.0316, -.0061]$
B	8	$+.0033$	$-.0132$	$-.0174$
		$[-.0009, +.0080]$	$[-.0208, -.0039]$	$[-.0279, -.0099]$

At $\varepsilon = 2$, clipping increases MSE by about .273 for the unregularized primary-set predictor and .177 for the regularized one; noise then adds .419 versus .337. Both changes increase each method's reported mean loss, but the regularized method loses less accuracy. The clipping/noise split differs across replication sets, so we do not make an exact percentage attribution part of the central claim.

C.2.3 Task stability and matched thresholds

The $\alpha = 0$ experiment preserves the feature law and $\mathbb{E}\gamma_j^2$ but makes the unstable coefficient constant across tasks. Table 6 reports the instability and variance sweep. High variance without coefficient variation gives a much smaller gain at $\varepsilon = 2$ and a slight loss at $\varepsilon = 8$. Its clipping cost increases rather than decreases under regularization. Increasing R from 2 to 4 in this task-stable setting does not reproduce the primary effect. The clipping-cost difference becomes more favorable with increasing instability, but private utility is not monotone in α.

Table 6: Task-instability and feature-variance comparisons, primary set R. Differences are regularized minus unregularized. The clean reference has the same base features but nearly shared task heads, unlike the radius-three heads of the separate bounded-anisotropic experiment.

ε	Regime (α, R)	Difference in MSE [95% interval]	Favorable seeds
2	Clean reference	-0.0173 $[-0.0216, -0.0129]$	21
2	Stable $(0, 2)$	-0.0209 $[-0.0245, -0.0174]$	24
2	Stable $(0, 4)$	-0.0213 $[-0.0250, -0.0176]$	23
2	$(0.3, 2)$	$+0.0008$ $[-0.0036, +0.0057]$	11
2	$(0.7, 2)$	-0.0501 $[-0.0644, -0.0365]$	23
2	$(1, 2)$, studied	-0.1659 $[-0.1965, -0.1368]$	24
8	Clean reference	-0.0004 $[-0.0028, +0.0025]$	11
8	Stable $(0, 2)$	$+0.0049$ $[+0.0018, +0.0083]$	8
8	Stable $(0, 4)$	$+0.0041$ $[+0.0000, +0.0084]$	12
8	$(0.3, 2)$	$+0.0105$ $[+0.0047, +0.0172]$	2
8	$(0.7, 2)$	-0.0062 $[-0.0136, +0.0014]$	12
8	$(1, 2)$, studied	-0.0303 $[-0.0398, -0.0209]$	21

The clean reference ($R = 1, \phi = 0$) is not the generator of Appendix C.5: the stable heads remain nearly shared rather than uniform in a radius-three ball. It has a small gain at $\varepsilon = 2$ and no resolved effect at $\varepsilon = 8$. The two experiments therefore compare different task-head distributions.

Publicly selected C is usually unchanged, and sometimes larger, for AvgSq. To remove threshold selection from the comparison, both representations are evaluated at each common $C \in \{1, 2, 5, 10, 20, 40\}$ with the same η and σ. Table 7 shows the direct utility differences; the advantage persists in each reported cell. The larger thresholds are stress tests, not publicly selected practical settings.

Table 7: Matched-C comparisons, replication set B. Both methods have the same threshold, step size, and noise multiplier. Negative differences favor regularization. Stars identify large-C stress tests.

ε	C	Unregularized MSE	Regularized MSE	Difference [95% interval], favorable seeds
2	1	2.203	2.137	-0.066 $[-0.078, -0.054]$, 23
2	2	1.908	1.794	-0.114 $[-0.129, -0.099]$, 24
2	5	1.914	1.710	-0.203 $[-0.219, -0.188]$, 24
2	10	2.996	2.558	-0.438 $[-0.481, -0.397]$, 24
2	20^*	7.659	6.231	-1.429 $[-1.604, -1.266]$, 24
2	40^*	26.454	21.016	-5.438 $[-6.157, -4.787]$, 24
8	1	2.144	2.108	-0.036 $[-0.046, -0.027]$, 22
8	2	1.743	1.696	-0.047 $[-0.059, -0.036]$, 22
8	5	1.356	1.326	-0.030 $[-0.037, -0.023]$, 22
8	10	1.374	1.355	-0.019 $[-0.026, -0.013]$, 21
8	20^*	1.818	1.777	-0.040 $[-0.059, -0.024]$, 21
8	40^*	3.693	3.538	-0.156 $[-0.223, -0.095]$, 21

The primary-set matched differences have the same sign throughout. At $\varepsilon = 2$ and $C = 20, 40$, the unregularized noisy trajectories still clip approximately 13.1% and 10.7% of gradients (regularized: 5.9% and 4.4%). Thus a small noise-free clipping penalty does not make the noisy recursion unclipped. The supported interpretation is a smaller incremental prediction-error cost when Gaussian noise is added to the clipped optimizer. Proposition 4 describes a related idealized process, not an exact model of these trajectories.

C.3 ILEA School: protocol, spectral objective, and privacy accounting

Each school is a regression task with 27 preprocessed covariates. Five public/private assignments each reserve 20 private schools and 100 records per private school, split 80/20. Feature transformation and target centering/scaling are fitted only on public records. The private interface is five coefficients on the learned features Ux plus an intercept; the task vector is updated without projection onto a norm ball. Comparators include the corresponding $\lambda = 0$ spectral representation, rank-five principal component analysis (PCA), full-dimensional adaptation, and intercept-only adaptation.

Public spectral surrogate. For each of 50 public schools, a ridge regression with an unpenalized intercept on standardized features yields a slope vector $\widetilde{w}_j \in \mathbb{R}^{27}$. These are the rows of W; the modal publicly cross-validated ridge level is 10. Let $\widehat{S}$ be the covariance of standardized features pooled over public records. The public solver takes the top-five eigenspace of $M_\lambda = W^\top W/m - \lambda\widehat{S}$. It exactly minimizes

$$\frac{1}{m}\sum_j \|(I - P_U)\widetilde{w}_j\|_2^2 + \lambda \operatorname{Tr}(P_U\widehat{S}) = \frac{1}{m}\sum_j \|\widetilde{w}_j\|_2^2 - \operatorname{Tr}(UM_\lambda U^\top)$$

over $U \in \mathcal{U}$, because the constant is independent of U and the trace is maximized by the leading eigenspace. This is a coefficient-reconstruction surrogate, not the empirical prediction loss in (5). In particular, it uses ridge slopes, an identity coefficient metric, and a pooled covariance penalty, rather than task-specific prediction metrics and equal-task feature second moments.

Public validation NMSE selects λ from $\{0, .01, .1, .2, .25, .3, .35, .4, .5, .6, .65, .7, .8, 1, 3, 10, 30, 100\}$; the selected values are 10 in four assignments and 3 in one. This is prediction-based representation selection, followed by separate public noisy-adaptation selection of C and η on proxy schools. The latter uses $C \in \{2^{-9}, 2^{-8}, \ldots, 2^5\}$ and $\rho = \eta C \in \{2^{-14.5+j/2} : j = 0, \ldots, 18\}$, with the production batch size, horizon, and multiplier. Each method selects its own settings. The eigensolve has no public learning-rate or training-horizon parameter.

Private mechanism and accountants. The run starts at zero, uses $T = 200$, $B = 20$ from $N = 80$ records, and clips the joint six-dimensional gradient. Every step draws an independent uniform size-20 subset without replacement. Gaussian noise with covariance $\sigma^2 C^2 I$ is added to

the clipped sum before dividing by B. Under replace-one adjacency, average sensitivity is $2C/B$, average noise standard deviation is $\sigma C/B$, and their ratio is $\sigma/2$.

School uses the fixed-size without-replacement replace-one bound of Birrell et al. (2024, Theorem 3.4), composed over 200 steps and converted at $\delta = 10^{-5}$ with Balle et al. (2020), using integer orders 2–512. That formula takes the configured σ directly. The synthetic accountant instead uses Wang et al. (2019) through `dp_accounting`, passing $\sigma/2$ as its Gaussian noise-to-sensitivity parameter. The physical meaning of the displayed σ is the same in both protocols; the applicable per-step Rényi bounds and software arguments differ.

The stored School multipliers 28.24970844 and 8.18661016 certify $\varepsilon = 2$ and 8 at orders 7 and 3. At the corresponding four-decimal reported multipliers, the independent Wang-bound calculation gives 2.23463565 and 9.45128186. These are different sufficient upper bounds for the stated mechanism, not conflicting exact privacy losses. The difference appears at the per-step Rényi bound, not in the adjacency or noise normalization. Table 8 reports the production calibration; its displayed multipliers are rounded, whereas production uses the stored precision.

Table 8: School privacy calibration for $N = 80$, $B = 20$, $T = 200$, and $\delta = 10^{-5}$, using the bound of Birrell et al. (2024, Theorem 3.4). σ is the configured multiplier, the standard deviation of the noise added to the clipped sum divided by C, and is passed to this accountant directly; the physical noise-to-sensitivity ratio is $\sigma/2$.

Target ε	σ	Achieved ε	Order
0.5	202.9493	0.4999999	19
1	66.3741	1.0000000	11
2	28.2497	2.0000000	7
4	14.5021	4.0000000	5
8	8.1866	8.0000000	3

Aggregation and original comparisons. MSE is normalized by $\max\{\mathrm{Var}_{\mathrm{test}}(y), 10^{-12}s^2_{\mathrm{pub}}\}$, with scale s_{pub} fitted publicly. Results average five adaptation repetitions within each school and 20 schools within each assignment, then summarize the five assignment values. The five assignments overlap in 3–8 private schools. We summarize their variation with one standard error, the standard deviation of the five assignment values divided by $\sqrt{5}$. Each privacy guarantee covers one fit on 80 training records; it is not a composed guarantee for releasing the whole experimental archive, and the held-out evaluation is not DP.

Table 9: ILEA School private test NMSE (mean±one standard error over five assignments). Lower is better.

ε	PROPOSED	$v^\top Ux$ ($\lambda = 0$)	PCA	FULL-w	Intercept only
0.5	1.335 ± 0.042	1.327 ± 0.046	1.243 ± 0.019	1.230 ± 0.015	$\mathbf{1.202 \pm 0.012}$
1	$\mathbf{0.997 \pm 0.016}$	1.119 ± 0.014	1.144 ± 0.007	1.151 ± 0.010	1.128 ± 0.007
2	$\mathbf{0.961 \pm 0.013}$	1.102 ± 0.012	1.134 ± 0.012	1.144 ± 0.012	1.122 ± 0.007
4	$\mathbf{0.958 \pm 0.017}$	1.101 ± 0.014	1.132 ± 0.013	1.144 ± 0.012	1.121 ± 0.009
8	$\mathbf{0.954 \pm 0.016}$	1.100 ± 0.014	1.131 ± 0.012	1.143 ± 0.011	1.120 ± 0.009

C.3.1 CLIPPING AND NOISE ABLATIONS

The ablation runs retain each production run's data, representation, initialization, learning rate, batch schedule, threshold when applicable, and horizon. Removing only noise gives CLIPPING ONLY; removing both gives NO CLIP/NO NOISE. The reported production cells are reproduced by the ablation runner. Because the saved batch schedule includes a budget index, noise-free runs use the schedule for each privacy budget, with differences agreeing within .002 across budget slices.

Table 12 separates the within-method costs from between-method advantages. At $\varepsilon = 2$, regularized clipping-only NMSE is .954 and its private NMSE is .961: adding noise worsens accuracy. Its advantage over unregularized adaptation nevertheless remains .141, compared with .146 before adding

Table 10: School assignment-level paired differences Δ_i in mean NMSE ($\lambda = 0$ minus regularized; positive favors regularization), with the mean and the nominal standard error $\mathrm{SD}/\sqrt{5}$. Rows NO CLIP/NO NOISE and CLIPPING ONLY are the ablation experiments in Appendix C.3.1.

Variant	123	2027	4099	7919	15401	Mean ± SE	Favors reg.
NO CLIP/NO NOISE	$+0.061$	$+0.039$	$+0.128$	$+0.004$	$+0.010$	$+0.048 \pm 0.022$	$5/5$
CLIPPING ONLY	$+0.148$	$+0.167$	$+0.129$	$+0.161$	$+0.126$	$+0.146 \pm 0.008$	$5/5$
DP, $\varepsilon = 0.5$	-0.060	-0.005	$+0.029$	$+0.012$	-0.016	-0.008 ± 0.015	$2/5$
DP, $\varepsilon = 1$	$+0.111$	$+0.175$	$+0.110$	$+0.125$	$+0.091$	$+0.122 \pm 0.014$	$5/5$
DP, $\varepsilon = 2$	$+0.140$	$+0.166$	$+0.133$	$+0.146$	$+0.121$	$+0.141 \pm 0.008$	$5/5$
DP, $\varepsilon = 4$	$+0.146$	$+0.163$	$+0.122$	$+0.162$	$+0.122$	$+0.143 \pm 0.009$	$5/5$
DP, $\varepsilon = 8$	$+0.148$	$+0.170$	$+0.127$	$+0.158$	$+0.127$	$+0.146 \pm 0.009$	$5/5$

noise. The regularized method receives a larger parameter-update noise scale in all five assignments, even though its selected C is smaller in three and larger in two. Since $\rho = \eta C$, the parameter-noise scale is $\sigma\rho/B$, whereas the averaged-gradient scale is $\sigma C/B$; these should not be conflated.

Table 11: Publicly selected clipping threshold and noise scales by assignment. $\rho = \eta C$ governs the parameter-space noise $\sigma\rho/B$; $\sigma C/B$ is the noise standard deviation on the averaged gradient, shown at $\varepsilon = 1$. The selected C is smaller for the regularized representation in three of five assignments and larger in two, whereas ρ is larger in all five.

	C		$\rho = \eta C$		$\sigma C/B$ at $\varepsilon = 1$	
Assignment	Reg.	$\lambda = 0$	Reg.	$\lambda = 0$	Reg.	$\lambda = 0$
123	0.0625	0.0078	0.0025	0.00088	0.207	0.026
2027	0.125	0.0625	0.0025	0.00088	0.415	0.207
4099	0.125	0.5	0.0025	0.00088	0.415	1.659
7919	0.0156	0.125	0.0035	0.00125	0.052	0.415
15401	0.0039	0.25	0.0025	0.00088	0.013	0.830

Table 12: School mechanism decomposition, mean NMSE pooled over assignments, schools and replicates. The clipping cost is CLIPPING ONLY minus NO CLIP/NO NOISE; the noise cost is DP minus CLIPPING ONLY.

Quantity	Regularized	$\lambda = 0$	Difference
No clip / no noise	0.803	0.852	$+0.048 \pm 0.022$
Clipping only	0.954	1.100	$+0.146 \pm 0.008$
Clipping cost	$+0.150 \pm 0.010$	$+0.248 \pm 0.022$	$+0.098 \pm 0.027$
Noise cost, $\varepsilon = 8$	$+0.0004$	$+0.0002$	-0.0002
Noise cost, $\varepsilon = 2$	$+0.0071$	$+0.0020$	-0.0050
Noise cost, $\varepsilon = 1$	$+0.0430$	$+0.0191$	-0.0239
Noise cost, $\varepsilon = 0.5$	$+0.3816$	$+0.2275$	-0.1540
Clipping fraction	0.955	0.925	

C.3.2 EFFECTIVE DIMENSION AND THE POST-HOC RANK-ONE COMPARISON

The nominal number of private parameters is unchanged, but the effective predictive dimension differs. The covariance $U\widehat{S}U^\top$ has exponential-entropy effective rank 1.00–1.03 for the regularized representation (REG) and 4.89–4.93 for the original $\lambda = 0$ representation. For eigenvalues normalized to proportions p_i, this rank is $\exp(-\sum_i p_i \log p_i)$; the participation ratio $1/\sum_i p_i^2$ gives the same qualitative picture. One-hot covariate groups create covariance-null directions. At the selected penalties, the regularized eigensolve uses almost only one feature-active direction, with the others null or nearly null. Its active direction is within $2.2°$–$6.8°$ of the leading unregularized predictive direction; the decomposition attributes 98%–99% of the projected-moment reduction to dropped directions rather than rotation of the retained direction.

The post-hoc RANK1-L0 baseline uses the leading predictive direction of the saved unregularized solution (the rank-one solution of the same unregularized spectral objective). It fills the remaining four rows with deterministic orthonormal covariance-null directions obtained from projected canonical vectors. It therefore keeps five projected coordinates plus an intercept, joint clipping, and noise

on all six parameters. Its threshold and step size are calibrated publicly using the same protocol, with unchanged privacy multiplier. Its comparison uses the same assignments, data splits, batch schedules, and Gaussian draws. This baseline was designed after inspecting the effective-rank finding and is identified as post hoc.

Table 13: Post-hoc ILEA School rank-one comparison. REG is the regularized spectral method. Differences are REG minus RANK1-L0 on the same assignments; positive values favor RANK1-L0. Error bars show one standard error over the five assignments.

ε	RANK1-L0 NMSE	REG minus RANK1-L0
0.5	1.389	$-.053 \pm .047$
1	.977	$+.020 \pm .010$
2	.943	$+.018 \pm .011$
4	.935	$+.023 \pm .012$
8	.934	$+.020 \pm .012$

At every $\varepsilon \geq 1$, RANK1-L0 reproduces 112–116% of the original mean advantage. The rank-one baseline improves on REG's mean NMSE, attributing the original gain primarily to lower effective predictive dimension. Neither REG nor this baseline improves on the original rank-five method at $\varepsilon = .5$. The post-hoc comparison is reported separately from the original production results.

C.4 CT SLICE LOCALIZATION

The UCI dataset contains 53,500 CT slices and 384 numerical predictors (Graf et al., 2011). Dataset metadata list 74 patients, whereas the distributed file used by the experiment has 97 patient identifiers. Tasks use those file identifiers; among 85 with at least 100 slices, 50 supply public tasks and 20 disjoint patients supply private tasks. Targets are divided by the documented range 180. The primary experiment uses 100 records per task and rank five. Public validation over five runs selects $\lambda = .001$, with $\lambda = 0$ as the direct reference. PCA and the untuned full-dimensional method are additional references.

Private adaptation learns on the fixed features Ux without task-vector projection, using $N = 80$, 100 steps, batch size 8, and $C = 1$. At $\delta = 1/80^2$, the fixed-size without-replacement replace-one RDP accountant in `dp-accounting` receives $\sigma/2$ and gives configured multipliers 7.74980396, 5.55201336, 3.75744459, and 2.72443127 at nominal budgets $\varepsilon \in \{2, 3, 5, 8\}$. Methods share patients, splits, minibatches, and Gaussian draws. Results average over 20 patients and five private repetitions, with standard errors over five public runs. Privacy is record-level for training slices, not patient-level. Table 14 reports the primary fixed-size comparison; the exact grid, seeds, and reproduction command are in the supplemental configuration and README.

Table 14: CT primary comparison at 100 records per task and $C = 1$, with mean $\pm$ standard error over five public runs. The target is normalized by its range. The last column compares directly with $\lambda = 0$; the full-dimensional method is not independently tuned.

ε	PROPOSED	$v^\top Ux$ ($\lambda = 0$)	PCA	FULL-w REF.	Rel. reduction
2	$\mathbf{0.10041 \pm 0.00039}$	0.10139 ± 0.00039	0.16183 ± 0.00442	1.78672 ± 0.00002	0.97%
3	$\mathbf{0.04193 \pm 0.00023}$	0.04220 ± 0.00023	0.06607 ± 0.00173	0.79515 ± 0.00006	0.64%
5	$\mathbf{0.01833 \pm 0.00017}$	0.01837 ± 0.00018	0.02879 ± 0.00064	0.30320 ± 0.00004	0.26%
8	$\mathbf{0.01165 \pm 0.00010}$	0.01166 ± 0.00010	0.01808 ± 0.00039	0.13657 ± 0.00003	0.08%

The direct comparison with the rank-five $\lambda = 0$ representation measures the incremental regularization gain, .08%–.97% in this table. PCA and full-dimensional adaptation provide additional reference points.

C.5 BOUNDED-ANISOTROPIC AND ISOTROPIC PROJECTED-NORM EXPERIMENTS

The base feature law in Appendix C.1 is paired with a common $U^\star \in \mathcal{U}$, independent task heads uniform in the radius-three ball, and centered unit-variance uniform label noise. Thus $\Sigma_\theta = \operatorname{diag}(a_i^2 u_{\theta,i}^2)$ and $\Sigma_{\text{meta}} = (13/12)\operatorname{diag}(a_i^2)$. This realizes the shared linear model with

bounded features $X = 23.0182$, coefficient bound $B_\star = 3$, and sufficient label bound $Y = 70.7868$. A projected comparison uses $B_v = 3$; it does not instantiate the stronger uniform containment requirement $B_v \geq \beta B_\star/\mu_0$ of Theorem 1.

Uniform standardized coordinates have fourth moment $9/5$, not the Gaussian value 3. In particular, $\mathbb{E}[\|x\|_2^4 \mid \theta] = (\sum_i s_{\theta,i}^2)^2 + (4/5)\sum_i s_{\theta,i}^4$. Proposition 5 supplies the corresponding gradient identity, with excess fourth-moment correction $-6/5$.

Joint public selection. The decisive test fixes $n_j = 200$, budgets 2 and 8, and the joint (λ, C) grids specified above. Adaptation with and without task-vector projection onto $\mathcal{B}_v$ is evaluated over five public seeds, averaging private repetitions within each seed. Ties break toward smaller λ, then smaller C. Selection chooses $\lambda = 0$ in all 20 seed/budget/algorithm cells, at $C = 5$ for budget 2 and $C = 10$ for budget 8. Table 15 shows the public selection scores and exact fresh-task components, without relying on the superseded positive-only penalty-selection rule.

Table 15: Targeted joint selection at $n_j = 200$, adaptation without task-vector projection, means over five public seeds, each penalty evaluated at its own best threshold under the joint public criterion. Calibration MSE is the public selection score; the remaining columns are exact population quantities on fresh tasks. $\lambda = 0$ minimizes the calibration score at both budgets.

ε	λ	Best C	Calib. MSE	Exact R_{meta}	Approx. err.	Adapt. err.	Clip. frac.
2	0	5	**3.0385**	4.1030	0.0969	1.8823	0.470
2	0.01	5	3.0489	4.0616	0.0982	1.8896	0.469
2	0.03	5	3.0705	3.9824	0.1027	1.9037	0.466
2	0.1	5	3.1532	3.7382	0.1328	1.9490	0.457
2	0.3	5–10	3.4052	3.2314	0.2900	2.0879	0.386
2	1	5	4.1524	2.3723	1.0418	2.0619	0.401
8	0	10	**1.8293**	4.1030	0.0969	0.7259	0.162
8	0.01	10	1.8393	4.0616	0.0982	0.7338	0.161
8	0.03	10	1.8609	3.9824	0.1027	0.7496	0.159
8	0.1	10	1.9477	3.7382	0.1328	0.8029	0.155
8	0.3	10	2.2355	3.2314	0.2900	0.9320	0.146
8	1	10	3.1572	2.3723	1.0418	1.1480	0.128

The best positive candidate, $\lambda = .01$, is worse on fresh tasks by $.0086 \pm .0006$ MSE at budget 2 and $.0093 \pm .0006$ at budget 8, in all five public seeds. Its projected population moment is lower, but both approximation and adaptation errors increase. Larger penalties lower the moment further while worsening prediction. Public selection consequently favors the unregularized representation in this experiment.

Isotropic comparison. The Gaussian feature distribution has $\Sigma_{\mathrm{meta}} = 1.3451 I_{100}$, hence $R_{\mathrm{meta}}(U) = 6.7255$ for every row-orthonormal $U \in \mathbb{R}^{5\times 100}$. In its 200-epoch geometry-only comparison, an empirical public reduction of 17.47% at $\lambda = 1$ did not transfer to the constant population quantity. The comparison illustrates the difference between the empirical public penalty and its expected value on new tasks.

Training horizon and numerical certificates. The bounded-anisotropic public stage was extended from 200 to 1,000 epochs after observing undertraining; the selected configuration and relevant validation quantities were stable by the longer horizon. The final joint-selection result above, not an earlier favorable checkpoint, is the authoritative utility comparison. Public task-diversity matrices were positive definite, with smallest eigenvalues .638–.941 across the 20 final public collections.

Worst-case certificates are not quantitatively informative at these sample sizes. The validation subspace bound before capping is approximately $2.26 \cdot 10^4$–$4.99 \cdot 10^4$, above the trivial $2k = 10$ bound; realized squared subspace errors are approximately .10–.62. The projected-moment deviation bound Γ_R is about 650–685, so r_{pub} saturates at $X^2 \approx 529.84$. These observations limit the numerical interpretation of the sufficient bounds; the empirical tuning procedure does not use them to prescribe its threshold.

D MAXIMUM-NORM REGULARIZATION: PROOFS, ALGORITHMS, AND EXPERIMENTS

For the Max objective in (6), we prove Theorem 2 and give algorithms, coverage examples, and experiments. AvgSq retains the average-squared penalty $\widehat{R}_{\mathrm{proj}}$ in (5).

D.1 COVERAGE AND THE GRADIENT BOUND

Let $S_x^{\mathrm{pub}} = \bigcup_j \{x_{j,i} : i \in [n_j]\}$ and $\mathrm{SymConv}(Q) = \mathrm{conv}\{\pm a : a \in Q\}$. The pointwise event in Section 5.3 is equivalently

$$\mathcal{E}_{\mathrm{cov}} = \{x_t \in \mathrm{SymConv}(S_x^{\mathrm{pub}}) \text{ for all } t \in [n]\}. \tag{25}$$

Every pointwise witness is a public subset, and the entire pool is admissible without a cardinality restriction. The algorithm need not construct these witnesses. The event covers all n private training records jointly with probability at least $1 - \beta_{\mathrm{cov}}$.

Uniform radius transfer. The convex-combination proof in (12) applies to every linear U. In fact,

$$\sup_{x \in \mathrm{SymConv}(S_x^{\mathrm{pub}})} \|Ux\|_2 = \max_{x \in S_x^{\mathrm{pub}}} \|Ux\|_2 = \widehat{R}_{\max}(U),$$

by the triangle inequality and inclusion of the public points in their hull. This identity holds for every U, including all publicly fitted and selected candidates. The same event covers all private records and iterates. The regularizer is computed by a maximum scan, without constructing the hull.

Threshold and unconditional privacy. With $r_U = \widehat{R}_{\max}(U)$, bounded labels and heads imply on $\mathcal{E}_{\mathrm{cov}}$

$$|y_t - v^\top U x_t| \le Y + B_v r_U, \qquad \|\nabla_v \ell(U, v; (x_t, y_t))\|_2 \le 2 r_U (Y + B_v r_U) = C_U^{\mathrm{cert}}. \tag{26}$$

Projection of the task vector onto $\mathcal{B}_v$ keeps this bound valid throughout adaptation. A public threshold $C_U \ge C_U^{\mathrm{cert}}$ therefore causes no clipping of the covered training gradients. On any dataset, however, the algorithm retains its clipping operation, so the sensitivity and accountant remain valid whether or not coverage holds. The probability β_{cov} is a utility parameter, not an extra contribution to δ.

Since Y and B_v can be loose, experiments tune C_U publicly rather than prescribe C_U^{cert}. A positive $C_U < C_U^{\mathrm{cert}}$ remains privacy-valid but need not avoid clipping.

D.2 PROOF OF THE COVERAGE-RISK THEOREM

We prove Theorem 2. Let $v^\circ = U w_\theta^\star$, so $\|v^\circ\|_2 \le B_\star \le B_v$ and $L_\theta(U, v^\circ) - \tau^2 = A_\theta(U)$. The comparator need not be the optimal head for U. This choice avoids any assumption that the unrestricted optimum belongs to the private ball.

First fix the public stage, the new task distribution, and its realized dataset. On $\mathcal{E}_{\mathrm{cov}}$, all sampled gradients are unclipped, including at data-dependent iterates. Independent uniform subset resampling therefore gives

$$\mathbb{E}[\bar{g}_t \mid S^{\mathrm{priv}}, v_t, \text{past}] = \nabla \widehat{L}_\theta(U, v_t), \qquad \mathbb{E}[\|\bar{g}_t + \xi_t\|_2^2 \mid S^{\mathrm{priv}}, v_t, \text{past}] \le C_U^2 \left(1 + \frac{k\sigma^2}{b^2}\right).$$

Euclidean projection is nonexpansive and fixes v°. Squaring the update difference, taking expectations, and using convexity of the empirical loss gives

$$2\eta \mathbb{E}[\widehat{L}_\theta(U, v_t) - \widehat{L}_\theta(U, v^\circ)] \le \mathbb{E}\|v_t - v^\circ\|_2^2 - \mathbb{E}\|v_{t+1} - v^\circ\|_2^2 + \eta^2 C_U^2 \left(1 + \frac{k\sigma^2}{b^2}\right).$$

Sum over $t = 0, \dots, T-1$, telescope, and apply Jensen's inequality to $\bar{v}_T = T^{-1} \sum_{t=0}^{T-1} v_t$. Since the ball has diameter $2B_v$, the conditional expected empirical excess is at most

$$\frac{2B_v^2}{\eta T} + \frac{\eta C_U^2}{2} \left(1 + \frac{k\sigma^2}{b^2}\right). \tag{27}$$

For population risk, write $z = Ux$, $c_\theta = \mathbb{E}[yz]$, $H_U = \mathbb{E}[zz^\top]$ and let hats denote sample means. For any $v, w \in \mathcal{B}_v$, the common y^2 term cancels when comparing their loss deviations, yielding

$$|(L_\theta - \widehat{L}_\theta)(U, v) - (L_\theta - \widehat{L}_\theta)(U, w)| \le 4B_v\|c_\theta - \widehat{c}_\theta\|_2 + 2B_v^2\|H_U - \widehat{H}_U\|_2 =: \Xi(S^{\mathrm{priv}}).$$

Conditional on U and the task, the original private records are independent and $\|Ux\|_2 \le X$. Centered sample-mean second moments and Jensen give

$$\mathbb{E}\|c_\theta - \widehat{c}_\theta\|_2 \le \frac{YX}{\sqrt{n}}, \qquad \mathbb{E}\|H_U - \widehat{H}_U\|_2 \le \sqrt{\mathbb{E}\|H_U - \widehat{H}_U\|_F^2} \le \frac{X^2}{\sqrt{n}}.$$

These expectations are taken without conditioning the records on coverage. On the good event, combine (27), Ξ, and the comparator approximation error. On the complement, the output stays in the ball, so its population excess is at most $(Y + B_vX)^2$. Multiply by the corresponding event indicators and average. All terms on the good-event upper bound are nonnegative, so dropping that indicator only enlarges the bound. The bad event contributes at most $\beta_{\mathrm{cov}}(Y + B_vX)^2$. This proves (13), including its expectation over public randomness.

A stronger support-coverage version. If an event depending only on the public data covers the entire private feature support, the same proof conditional on that event uses $\|Ux\|_2 \le r_U$ in the moment deviations. Consequently,

$$\mathbb{E}[L_\theta(U, \bar{v}_T) - \tau^2 \mid S^{\mathrm{pub}}, U, \theta] \le A_\theta(U) + \frac{2B_v^2}{\eta T} + \frac{\eta C_U^2}{2}\left(1 + \frac{k\sigma^2}{b^2}\right) + \frac{4B_vYr_U + 2B_v^2r_U^2}{\sqrt{n}}. \quad (28)$$

Finite-sample coverage alone does not justify replacing X by r_U in the population concentration terms.

Connection to public recovery. Let $\Delta_L = \sup_{U\in\mathcal{U}, V\in\mathcal{V}_B} |L_{\mathrm{pub}} - \widehat{L}_{\mathrm{pub}}|$. Approximate optimality of the Max objective over this class, with $V^\star \in \mathcal{V}_B$, gives

$$L_{\mathrm{pub}}(\widehat{U}, \widehat{V}) - \tau^2 \le 2\Delta_L + \xi_{\mathrm{opt}} + \lambda[\widehat{R}_{\max}(U^\star) - \widehat{R}_{\max}(\widehat{U})].$$

This follows by comparing with $(U^\star, V^\star)$ and applying the two prediction-loss deviations. The empirical maximum need not be replaced by a population maximum. Theorem 3 or the validation certificate controls subspace error, and (21) controls $A_\theta(U)$. Thus the prediction term and coverage assumption address the approximation and private-optimization sides of the bound, respectively.

D.3 Minibatch implementations of the pooled maximum

Enumerate the N public training features as $z_1, \ldots, z_N$, and define $f_i(U) = \|Uz_i\|_2$ and $\widehat{X} = \max_i \|z_i\|_2$. For $Uz_i \ne 0$ a subgradient is

$$\Psi_i(U) = \frac{(Uz_i)z_i^\top}{\|Uz_i\|_2}; \quad \Psi_i(U) = 0 \text{ is valid when } Uz_i = 0. \quad (29)$$

If $i^\star$ maximizes $f_i(U)$, then $\Psi_{i^\star}(U)$ is a subgradient of the entire maximum: for any W, $\max_i f_i(W) \ge f_{i^\star}(W) \ge f_{i^\star}(U) + \langle \Psi_{i^\star}(U), W - U\rangle_F$. A maximum over an ordinary random minibatch is a different objective, not an unbiased estimate of this pooled maximum.

Each scan costs $O(Nkd)$ time and can be streamed without retaining a computation graph for all records. Only the maximizing record is needed for its subgradient. Prediction-loss updates remain minibatch updates, with the original task weighting. This exact version is used in the signed-coordinate experiment.

Cached maximizers with an explicit error bound. After a full scan at U_s, keep its maximizing index i_s in a cache I. Since $|f_i(U) - f_i(U_s)| \le \widehat{X}\|U - U_s\|_2$,

$$0 \le \widehat{R}_{\max}(U) - \max_{i\in I} f_i(U) \le 2\widehat{X}\|U - U_s\|_2.$$

The full maximum can increase by at most one drift term and the saved maximizing value can decrease by at most the other. If the right side is at most $\epsilon_{\max}$, the maximizing cached gradient is an $\epsilon_{\max}$-subgradient: the subgradient inequality above holds with a subtractive $\epsilon_{\max}$ error. Refresh after a prescribed number of steps or when the drift bound exceeds the tolerance. The final reported radius comes from an exact scan.

Algorithm 4 Minibatch public learning with an exact pooled maximum

Require: Public tasks, $k, \lambda, \eta_U, \eta_v, T_{\mathrm{meta}}, K_v$
1: Initialize the representation and heads as in Algorithm 1
2: **for** $t = 0, \dots, T_{\mathrm{meta}} - 1$ **do**
3: Update each task head with K_v minibatch prediction-loss steps at fixed U_t
4: Stream the pooled features and retain $i_t \in \arg\max_i \|U_t z_i\|_2$
5: Draw fresh task batches and compute their equal-task mean prediction gradient G_L
6: $G_t \leftarrow G_L + \lambda \Psi_{i_t}(U_t)$; $\widetilde{U} \leftarrow U_t - \eta_U G_t$
7: Factor $\widetilde{U}^\top = QR$; set $U_{t+1} \leftarrow Q^\top$
8: **end for**
9: Freeze U and perform a final exact scan for its public radius r_U
10: **return** the representation, heads, and radius

An exact auxiliary-variable formulation. For nonnegative $f_i(U)$ bounded by $\widehat{X}$ on $\mathcal{U}$,

$$\max_i f_i(U) = \min_{a \in [0, \widehat{X}]} \Phi(U, a), \qquad \Phi(U, a) = a + \sum_{i=1}^{N} (f_i(U) - a)_+. \tag{30}$$

For any a, the expression is at least the largest f_i; equality holds at $a = \max_i f_i$. This is the extreme empirical-tail case of the positive-part formulation of Rockafellar & Uryasev (2000). Therefore one can minimize $\widehat{L}_{\mathrm{pub}}(U, V) + \lambda \Phi(U, a)$ with stochastic batches. For a uniform size-b_R record batch I_t, unbiased subgradient estimates are

$$\widehat{G}_R = \frac{N}{b_R} \sum_{i \in I_t} \mathbf{1}\{f_i(U) > a\} \Psi_i(U), \qquad \widehat{g}_a = 1 - \frac{N}{b_R} \sum_{i \in I_t} \mathbf{1}\{f_i(U) > a\}.$$

The zero choice at ties is a valid subgradient. At each step add $\lambda \widehat{G}_R$ to the minibatch prediction gradient, retract the representation by QR, and update $a \leftarrow \Pi_{[0, \widehat{X}]}(a - \eta_a \lambda \widehat{g}_a)$. The factor N/b_R is essential for unbiasedness. Rare exceedances can give high variance, so exact scans are simpler at the scale considered here. The scalar a can lie below the actual maximum even at a minimizer of (30); it is not a certified clipping radius. Always use the final full scan for r_U, and public adaptation trials for the experimental threshold.

D.4 Task families satisfying coverage

Signed anchors with task-dependent interior distributions. Fix nonzero $a_1, \dots, a_L \in \mathbb{R}^d$ with pairwise disjoint signed pairs $\{a_\ell, -a_\ell\}$, and $K = \mathrm{conv}\{\pm a_1, \dots, \pm a_L\}$. Suppose every task has support in K and assigns probability at least $p_\ell > 0$ to the pair $\{a_\ell, -a_\ell\}$. The remaining mass may vary by task anywhere inside K, including continuously.

Proposition 6 (Coverage from signed anchor observations)**.** *For $N = \sum_j n_j$ independent records conditional on the public tasks,*

$$\Pr[K \subseteq \mathrm{SymConv}(S_x^{\mathrm{pub}})] \geq 1 - \sum_{\ell=1}^{L} (1 - p_\ell)^N \geq 1 - L e^{-N p_{\min}}, \qquad p_{\min} = \min_\ell p_\ell.$$

Proof. For each pair, conditional independence and the uniform mass lower bound give probability at most $(1 - p_\ell)^N$ of missing both signs. Once at least one sign of every pair is observed, symmetry supplies the other and the public hull contains K. Union bound over pairs and average over the public tasks. □

This event covers entire task supports, so (28) applies. With $A = [a_1 \ \cdots \ a_L]$ of row rank d, every task has $\Sigma_\theta \succeq \sum_\ell p_\ell a_\ell a_\ell^\top \succeq p_{\min} A A^\top \succ 0$. Thus the model can also satisfy the full-dimensional excitation required for public recovery.

Directional tasks, including continuous radii. A task draws type $J \in [L]$ with probability $\pi_J > 0$, a bounded task head, and features $x = \epsilon_x A_J q_J$, where q_J is a fixed unit direction, $A_J > 0$, and ϵ_x is a uniform sign. Labels can follow the same bounded shared linear model. If any of the m public tasks has the new task's type, one matching public feature is a witness for every private point. Consequently

$$\Pr(\mathcal{E}_{\mathrm{cov}}^c) \le \sum_{\ell=1}^{L} \pi_\ell (1-\pi_\ell)^m.$$

This follows by conditioning on the new task type and counting the chance that all public task types miss it. It does not require per-task full-dimensional covariance, which is unnecessary for Theorem 2.

A continuous version has $x = \epsilon_x R A_J q_J$ with independent radii $R \in (0, 1]$ drawn from a common continuous law for each type. If each public task supplies n_{pub} records and $M_\ell \sim \mathrm{Binomial}(m, \pi_\ell)$ counts type-ℓ tasks, the largest public radius on that line covers all private records unless the maximum of the combined radial draws belongs to the private sample. By exchangeability,

$$\Pr(\mathcal{E}_{\mathrm{cov}}^c) \le \sum_{\ell} \pi_\ell \mathbb{E}\left[\frac{n}{n + n_{\mathrm{pub}} M_\ell}\right].$$

For $n_{\mathrm{pub}} = n$, the bound is $(m+1)^{-1} \sum_\ell [1 - (1-\pi_\ell)^{m+1}] \le L/(m+1)$. Indeed, $\mathbb{E}[1/(M+1)] = \int_0^1 (1 - p + pt)^m \, dt$ for $M \sim \mathrm{Binomial}(m, p)$. This is finite-dataset coverage, not full-support coverage. These directional constructions illustrate the assumption; no directional-family experiment is claimed.

D.5 Signed-coordinate experiment and threshold follow-up

Generator. Set $a_\ell = b_\ell e_\ell$, $Z = \sum_\ell b_\ell^{-2}$, choose $\Pr(J = \ell) = b_\ell^{-2}/Z$, and emit $x = \epsilon_x b_J e_J$. Every task uses this feature law with its own uniformly distributed unit-ball head. Then

$$\Sigma_\theta = I_d/Z, \qquad R_{\mathrm{meta}}(U) = k/Z, \qquad \sup_{x \in \mathrm{supp}(\mathcal{D}_{\theta,x})} \|Ux\|_2 = \max_\ell b_\ell \|U e_\ell\|_2. \tag{31}$$

The population AvgSq penalty is constant, but the maximum need not be. We use $d = 20, k = 3, m = 50, n_j = 100$, with $b_\ell = 4$ for the first five coordinates and $b_\ell = 1$ otherwise. Thus $Z = 245/16$ and $p_{\min} = 1/245$. A true row-orthonormal representation is drawn per public-task collection; independent label noise is uniform on $[-0.1, 0.1]$. The noise floor is $\tau^2 = 0.01/3$, and $X = 4, Y = 4.1, B_v = 1$ are valid. Exact risk is

$$L_\theta(U, v) = \tau^2 + (w_\theta^\star - U^\top v)^\top \Sigma_\theta (w_\theta^\star - U^\top v). \tag{32}$$

Public training and selection. All three methods use the same public solver with the appropriate penalty: 600 epochs, five local head steps, $\eta_U = \eta_v = 0.01$, and 64/16/20 training/validation/test percentages. Max scans all 3,200 public training records at each representation update. The grid is $\Lambda = \{0, .01, .03, .1, .3, 1\}$ for each penalty, with zero competing in its own selection. Ten additional public calibration tasks and three repetitions select (λ, C, η) using noisy adaptation loss with task-vector projection onto the unit ball. The common grids are $\mathcal{G}_C = \{.01, .02, .05, .1, .2, .5, 1, 2, 5\}$ and $\mathcal{G}_\eta = \{.03, .1, .3, 1, 3\}$. The original grids were fixed after public pilot adjustments, before final evaluation.

Private protocol. Each fresh task supplies 80 adaptation and 20 test records. The run uses $b = 8, T = 100$, $\delta = 1/80^2$, replace-one adjacency, and the Wang-based without-replacement accountant ($\sigma = 7.75$ and 2.7246 for $\varepsilon = 2, 8$). Heads are projected onto the unit ball, start at the projected mean public head, and the implementation averages $v_1, \ldots, v_T$. The theorem uses the different average $v_0, \ldots, v_{T-1}$. Each method receives its own publicly tuned threshold, unrestricted by C_U^{cert}. Frozen settings are evaluated on 20 fresh tasks for each of 24 independent public-task collections, excluding two pilot seeds. The 95% intervals for the mean difference use 10,000 bootstrap resamples of the 24 collection-level MSE differences, with methods evaluated on the same tasks.

Max reduces the radius substantially, but its tuned private-MSE change relative to the baseline is -0.00019 $[-0.00053, +0.00011]$ at $\varepsilon = 2$ and -0.00040 $[-0.00085, +0.00000]$ at $\varepsilon = 8$; variation

Table 16: Signed-coordinate primary experiment, means over 24 seeds. The radius is the exact public training maximum; C and η are publicly selected. All MSE values in this table use the tuned threshold, not the certificate.

ε	Method	r_U	C	η	Exact private MSE
2	Unregularized	2.403	.0213	3.00	.01066
2	AvgSq	2.269	.0213	3.00	.01027
2	Max	1.744	.0200	3.00	.01048
8	Unregularized	2.403	.0517	2.83	.00713
8	AvgSq	2.278	.0517	2.92	.00663
8	Max	1.821	.0529	2.92	.00673

across the 24 collections leaves the mean advantage unresolved. It selects $\lambda = 0$ on 7 and 6 seeds, respectively. At $\varepsilon = 2$, its threshold matches the baseline on 23 of 24 seeds. AvgSq has small favorable mean differences (-0.00039 and -0.00050, with intervals excluding zero). Since its population criterion is invariant here, that result concerns finite-sample or optimization effects of the empirical objective, not population second-moment reduction. Max and AvgSq also have similar mean MSE.

Independent learning-rate follow-up. The original search selected its largest learning rate, 3, on every seed at $\varepsilon = 2$ and almost every seed at $\varepsilon = 8$. A public-only extension added $\{4, 5, 6\}$ without retraining the representations. The expanded grid selected 4–6 on every seed at $\varepsilon = 2$ and usually 4–5, sometimes 6, at $\varepsilon = 8$, lowering mean per-seed calibration error by 1.4%–2.4% relative to the best original-grid choice. A pre-fixed decision rule therefore triggered evaluation on 24 fresh seeds, reported separately in Table 17. The Max utility conclusion is unchanged. At $\varepsilon = 2$, the smallest threshold .01 was selected on 11, 9, and 8 seeds for the baseline, AvgSq, and Max, respectively.

Table 17: Learning-rate follow-up, 24 fresh seeds. Differences are first minus second; negative MSE differences favor the first method. Intervals are 95% paired-bootstrap intervals. This is a separate evaluation from the primary experiment.

ε	Comparison	Difference in exact private MSE [interval]	Favorable seeds
2	AvgSq − baseline	$-.000592$ $[-.001085, -.000202]$	16/24
2	Max − baseline	$-.000318$ $[-.001532, +.000786]$	8/24
2	Max − AvgSq	$+.000274$ $[-.000683, +.001169]$	8/24
8	AvgSq − baseline	$-.000604$ $[-.001161, -.000154]$	17/24
8	Max − baseline	$-.000544$ $[-.001904, +.000740]$	8/24
8	Max − AvgSq	$+.000060$ $[-.001066, +.001146]$	8/24

Coverage and certificate consistency. For the 3,200 training records used by the maximum, $\beta_{\mathrm{cov}} \le 20e^{-3200/245} < 5 \cdot 10^{-5}$. The scanned maximum equaled the support radius in every primary cell. Separate runs at C_U^{cert} had zero clipping in all 144 cells, thresholds two to three orders of magnitude larger than the tuned choices, and 4–7 times higher MSE. These additional runs check the no-clipping certificate.

Max on the task-instability family. The replication set of Appendix C.2 also includes a Max arm under the protocol that returns the final task vector without projecting it onto a norm ball; features remain Ux. Relative to the baseline, its MSE differences are $-.0021$ $[-.0226, +.0196]$ at $\varepsilon = 2$ and $+.0031$ $[-.0035, +.0109]$ at $\varepsilon = 8$, with $\lambda = 0$ selected on 9 and 10 of 24 seeds. The coverage theorem is not asserted for this protocol. The mean MSE differences are small relative to variation across task collections.