

Private Coupled Descent: Dimension-Free Privacy for Vertical Federated Learning by Privatizing the Coupling Variable

Axel Faes and Maryam Amir Haeri

Abstract—In vertical federated learning, parties hold different features of the same entities; principled methods couple them through a low-dimensional, sample-mode *score*, keeping loadings local. We privatize this *coupling variable* instead of exchanged gradients or factors, giving a perturbed proximal alternating scheme. Our main result (Thm. 2) shows the whole sequence converges to a critical point at zero noise (Kurdyka–Łojasiewicz) and, under a fixed differential-privacy budget, to a neighborhood of radius linear in the noise. The prediction’s privacy cost depends on the score length and regression curvature, and is *independent of the ambient feature dimension*. A matching lower bound (Prop. 3) shows this cost order-optimal in (c, ϵ) on well-conditioned instances. Because the score is the sufficient statistic for inter-party interaction, privatized once per round, the cost is independent of party count P , whereas per-message noise pays a factor P (Prop. 1). We instantiate it on a sparse block-term tensor regression with high-dimensional local representations and an interpretable sparse core. There, masking as in linear private vertical methods forfeits privacy or sparse recovery: among orthogonal masks only signed permutations preserve a thresholded support and conceal nothing (Thm. 1), so the analyzed design has parties decompose locally and share only the privatized score. Synthetic data validate the predicted scalings. To our knowledge this is the first convergence-and-utility theory for differentially-private block-coordinate descent with coupling-variable noise, and the first for block-term decomposition.

Index Terms—Differential privacy, vertical federated learning, block-term tensor decomposition, nonconvex optimization, proximal alternating minimization, privacy–utility tradeoff.

I. INTRODUCTION

Distributed signal processing increasingly fits one model across parties that each observe a different view of the same entities. Vertical federated learning (VFL) is its supervised instance: parties hold *different features* of a shared set of entities, with labels held by one party [1]. The principled linear methods (privacy-preserving partial least squares and federated SVD) share a sample-mode latent score while keeping each party’s loadings private, using an orthogonal *mask* so a service provider operates on data it cannot invert [2], [3]. Seen through an estimation lens, that shared score is the one statistic the parties exchange, and the whole privacy question reduces to how much it must be perturbed.

The stakes are sharpest where the local representations are high-dimensional. A party may hold a large multiway signal,

such as a multi-lead physiological waveform, a multimodal scan, or a sensor array, whose interpretable structure is exactly what a sparse multilinear model recovers, and exactly what a privacy budget should not have to pay for twice. These are routine settings for differentially private signal-processing models: distributed estimation over multimodal biomedical waveforms, where each site guards its raw measurements. Whether the privacy cost of collaborating must grow with that ambient richness, or only with the small shared latent that crosses the wire, is the question this paper settles: we show the cost can be made to depend on the latter alone, and that this is optimal.

Many high-value signals are, however, *multiway*, and the models that exploit this (block-term tensor factorizations and regressions [4], [5]) earn their interpretability from a *sparse* core selected by thresholding. We show this is exactly where the masking trick breaks: thresholding is basis-dependent, and an orthogonal mask is a rotation, so masking and sparse recovery are incompatible (Thm. 1). The design is thereby *forced*: parties must run the sparse decomposition locally and share only the low-dimensional sample-mode score, which is then protected by differential privacy (DP).

This produces an unusual object: a proximal alternating (block-coordinate) iteration in which the privacy noise is injected into the *coupling variable* (the shared score) rather than into each party’s gradient. The shared score is the minimal sufficient statistic for inter-party interaction, so it is the one quantity that must be privatized, and block-coordinate descent with a convergence analysis is the natural machinery for the resulting optimization. Existing analyses, namely DP empirical risk minimization, DP coordinate descent, and DP tensor factorization [6], [7], [8], add noise to gradients or factors; none treats coupling-variable noise, and none gives convergence *and* utility for a *nonconvex, sparse, multilinear* (block-term) objective. We close that gap.

a) Contributions.:

- 1) **A no-go characterization** (Thm. 1): among orthogonal masks, only signed permutations preserve a thresholded support, and they reveal the magnitude multiset; hence masked sparse decomposition cannot be private. The design is forced to local-decompose-and-share-the-score.
- 2) **Private Coupled Descent** (Thm. 2): the DP coupling-noise iteration is a perturbed proximal alternating linearized minimization on the coupled objective. At zero noise it converges (whole sequence) to a critical point;

A. Faes is with the Data Science Institute, Hasselt University, Belgium, and with the University of Twente, The Netherlands (e-mail: axel.faes@utwente.nl).

M. Amir Haeri is with Learning, Data Analytics and Technology, University of Twente, The Netherlands.

under a fixed (ε, δ) budget it reaches a neighborhood of radius $O(\sigma)$, with explicit constants. This unifies the “exact” and “private” regimes that otherwise appear contradictory.

- 3) **A dimension-free privacy-utility result** (Cor. 1): the excess *prediction* risk depends on the score length and regression curvature, *not* the ambient feature dimension, which is the low-rank escape from the DP-ERM dimension penalty, with an **order-matching lower bound, tight in (c, ε)** (Prop. 3) certifying order-optimality in (c, ε) on well-conditioned instances. Because the coupling score is the sufficient statistic for inter-party interaction, it is privatized once per round, making the cost **independent of the party count P** , where per-message gradient/factor noise pays a factor P (Prop. 1). This is the structural reason to noise the coupling variable. (Recovering the high-dimensional factors privately is a separate, dimension-dependent matter we do not claim.)
- 4) **Instantiation and validation** on vertical block-term tensor regression, with synthetic experiments confirming the predicted noise and dimension scaling.

II. RELATED WORK

a) Vertical FL and sparse interpretable models.: Vertical FL spans secure linear and tree models, split neural networks, and the masked-linear-algebra line of privacy-preserving PLS and federated SVD/PCA [2], [3], [9], which is the closest in spirit: a shared sample-mode score, private per-party loadings, and an orthogonal mask for confidentiality. These are *linear*: their latent factor is a single direction per component. Multiway models, the block-term decompositions and regressions [4], [5], are by now an established tool in signal processing, and they replace that single direction with a low-rank *subspace* and earn interpretability from a thresholded, sparse core, which is precisely the ingredient that the masking trick cannot accommodate (Thm. 1). The federated tensor literature is otherwise horizontal or differently aligned. Coupled tensor-train FL [10] couples the *feature* modes and is unsupervised; SecureTensor [11] is horizontal and unsupervised (a decompose-then-classify pipeline) and reports parity with centralized accuracy at reduced communication, but proves only a privacy guarantee, not convergence. Closest on paper is MFVFL [12], which is vertical, supervised, tensor, and differentially private, yet there the tensor decomposition is a server-side feature-imputation regularizer, the DP noise sits on each party’s uploaded embedding (a *separable* perturbation that the method then *denoises* by robust PCA), and its convergence constants are the generic ones of stochastic gradient descent. We instead make the block-term decomposition the regression model, place the noise on the shared coupling score (a non-separable perturbation we prove convergence *through* rather than *remove*), and none of these addresses the feature-partitioned, label-supervised, sparse setting we analyze.

b) Nonconvex alternating minimization.: PALM [13] and the multiconvex block-coordinate method of [14] give whole-sequence convergence for nonsmooth nonconvex block problems under the Kurdyka-Łojasiewicz (KL) property [15], [16];

TABLE I
WHERE THE PRIVACY NOISE ENTERS, AND WHAT IS PROVEN.

	noise on	utility bound?	multilinear?
DP-CD [7]	gradient	yes (convex)	no
DP tensor fact. [8]	factors	empirical	yes
Hidden-state DP-BCD [22]	gradient	no	no
This work	coupling score	yes (nonconvex)	yes

semi-algebraic objectives (polynomial coupling + ℓ_1) satisfy KL. Stochastic/inexact variants (SPRING [17], inexact prox-gradient [18]) reach a critical point only with vanishing error variance; with a *non-decaying* error they reach a neighborhood, the regime a fixed DP budget induces, as in nonconvex SGD [19].

c) Differential privacy.: The Gaussian mechanism and composition [20], [21] and DP-ERM utility bounds [6] (whose lower bound makes a dimension dependence necessary) are standard. DP coordinate descent [7] and DP tensor factorization [8] add gradient/factor noise. The closest work, hidden-state DP block-coordinate descent [22], adds *gradient* noise, is privacy-only (no utility/convergence), assumes *no* inter-block coupling, and does not treat tensors, the opposite of our object. We add noise to the *coupling variable* and prove convergence and utility for a block-term objective. We tabulate the distinction in Table I: prior DP alternating/coordinate methods perturb each block’s own gradient or factor, an error that is *local* to a block and that variance reduction can drive to zero; our perturbation is on the *shared score* that links all blocks, is irreducible under a fixed budget, and propagates through the multilinear coupling, which is precisely what the analysis must control. Because that score is the sufficient statistic the parties share, it is privatized *once* rather than once per party, which the gradient/factor-noise designs cannot do and which makes our privacy cost independent of the party count (Prop. 1). None of the listed works gives a convergence-with-utility bound for a sparse multilinear (block-term) objective.

III. SETUP

a) Coupled objective.: P parties hold tensors $\mathcal{X}^{(p)} \in \mathbb{R}^{m \times I_2^{(p)} \times \dots}$ over m shared, aligned samples; one active party holds labels and forms the residual $\mathbf{F} \in \mathbb{R}^{m \times C}$. Per deflation block,

$$\mathcal{L}(x) = \underbrace{\|\mathbf{F} - d\mathbf{t}\mathbf{q}^\top\|_F^2}_{H(x)} + \sum_p \lambda_p \|\mathcal{G}^{(p)}\|_1, \quad (1)$$

$$\mathbf{t} = \sum_p \mathbf{t}^{(p)}(\theta^{(p)}),$$

with blocks $x = (\theta^{(1)}, \dots, \theta^{(P)}, (\mathbf{q}, d))$, $\theta^{(p)} = (\{\mathbf{P}^{(p,n)}\}, \mathcal{G}^{(p)})$, and $\mathbf{t}^{(p)}$ the multilinear contraction of $\mathcal{X}^{(p)}$ with party p ’s factors. H is a polynomial (smooth, block-Lipschitz on bounded sets); the penalty is separable and prox-friendly. The parties couple *only* through the additive score $\sum_p \mathbf{t}^{(p)}$.

Algorithm 1: Private Coupled Descent (one deflation block)

Input: residual $\mathbf{F}$ (active party); steps η_p ; clip c ; noise σ

```

1 init score  $\mathbf{t}$ ; for  $k = 0, 1, \dots$  until  $\|\mathcal{G}^k\|$  near floor do
2   for each party  $p$  (local, in the clear) do
3      $\theta^{(p)} \leftarrow \text{prox}_{\eta_p \lambda_p \|\cdot\|_1}(\theta^{(p)} - \eta_p \tilde{\nabla}_{\theta^{(p)}} H)$  //  $\tilde{\nabla}$ 
4     uses the broadcast score
5     keep factors orthonormal; form  $\mathbf{t}^{(p)}$ ; clip  $\text{clip}_c(\mathbf{t}^{(p)})$ 
6   CSP:  $\tilde{\mathbf{t}} \leftarrow \sum_p \text{clip}_c(\mathbf{t}^{(p)}) + \mathbf{n}$ ,  $\mathbf{n} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_m)$ ;
   normalize; broadcast
7   active party:  $(\mathbf{q}, d) \leftarrow \arg \min H$ ; residual signal
   returned under the same mechanism
8 deflate  $\mathbf{F} \leftarrow \mathbf{F} - d \mathbf{t} \mathbf{q}^\top$  (active party)
```

b) *DP coupling-noise iteration.*: Each round, party p takes a proximal-gradient (PALM) step on its block using the broadcast score; the service provider returns the score under the Gaussian mechanism, $\tilde{\mathbf{t}} = \sum_p \text{clip}_c(\mathbf{t}^{(p)}) + \mathbf{n}$, $\mathbf{n} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_m)$. Party p thus performs its update with a noise-perturbed coupling offset. The active $(\mathbf{q}, d)$ block is exact. We use the standard semi-honest, no-collusion model with a trusted key authority and a service provider, as in [2], [3]; DP gives a guarantee for any P (including $P=2$, where secure aggregation alone does not). We use the standard notion.

Definition 1 (Differential privacy). *A randomized mechanism M is (ϵ, δ) -differentially private if for all inputs D, D' differing in one party's data and all measurable S , $\Pr[M(D) \in S] \leq e^\epsilon \Pr[M(D') \in S] + \delta$.*

The released objects each round are the aggregate score and the residual signal; privacy is measured with respect to changing a single party's local tensor. Algorithm 1 states one block.

Assumption 1. (i) *Iterates lie in a compact set (coercivity + clipping/normalization), with $|d| \leq D$, $\|\mathbf{q}\| \leq Q$, block Jacobian $\|\partial \mathbf{t}^{(p)} / \partial \theta^{(p)}\| \leq B$, and block-Lipschitz ∇H with constants L_p . (ii) Step sizes $\eta_p < 1/L_p$, $\gamma := \min_p (1/\eta_p - L_p) > 0$. (iii) $\mathcal{L}$ is semi-algebraic (polynomial $H + \ell_1$), hence satisfies KL. (iv) Factors are kept orthonormal and the score normalized (so $\mathbf{t}^{(p)}$ has $O(1)$ scale).*

c) *Preliminaries.*: We use three standard objects. The proximal-gradient mapping at x with step η is $\mathcal{G}(x) = \frac{1}{\eta}(x - \text{prox}_{\eta g}(x - \eta \nabla H(x)))$; $\mathcal{G}(x) = 0$ iff x is a critical point of $\mathcal{L}$, and $\|\mathcal{G}\|^2$ is the standard nonconvex stationarity measure. The smooth descent lemma states $H(y) \leq H(x) + \langle \nabla H(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$ for L -Lipschitz ∇H . A proper lower-semicontinuous $\mathcal{L}$ has the Kurdyka-Łojasiewicz (KL) property at $\bar{x}$ if, near $\bar{x}$, $\varphi'(\mathcal{L}(x) - \mathcal{L}(\bar{x})) \text{dist}(0, \partial \mathcal{L}(x)) \geq 1$ for a concave desingularizer φ ; every semi-algebraic function satisfies it [15], and a polynomial coupling plus an ℓ_1 penalty is semi-algebraic, so Assumption 1(iii) holds for (1) with no further conditions.

IV. THE MASKING-SPARSITY NO-GO

Theorem 1 (Mask characterization). *Let soft_λ ($\lambda > 0$) be elementwise soft-thresholding on $\mathbb{R}^d$ and $\mathbf{H} \in O(d)$. Then $\text{soft}_\lambda(\mathbf{H}^\top \mathbf{z}) = \mathbf{H}^\top \text{soft}_\lambda(\mathbf{z})$ for all $\mathbf{z}$ iff $\mathbf{H}$ is a signed permutation. Consequently the only orthogonal masks that preserve a thresholded support reveal the multiset of core magnitudes; no orthogonal mask both conceals the core values and preserves sparse recovery.*

Proof. ($\Leftarrow$) soft_λ acts coordinatewise and is odd, so it commutes with coordinate permutations and sign flips, hence with any signed permutation $\mathbf{H} = \mathbf{P}\mathbf{D}$. ($\Rightarrow$) Fix j and take $\mathbf{z} = t\mathbf{e}_j$ with $t > \lambda$, so $\text{soft}_\lambda(\mathbf{z}) = (t - \lambda)\mathbf{e}_j$ and the identity reads $\text{soft}_\lambda(\mathbf{H}\tilde{\mathbf{h}}_j) = (t - \lambda)\tilde{\mathbf{h}}_j$ with $\tilde{\mathbf{h}}_j = \mathbf{H}^\top \mathbf{e}_j$. Consider a nonzero entry $a = (\tilde{\mathbf{h}}_j)_i$. If $|a| > 1$: for $t > \lambda$ the identity gives $\text{sign}(a)(t|a| - \lambda) = (t - \lambda)a$, i.e. $a(t - \lambda/|a|) = a(t - \lambda)$, forcing $|a| = 1$, a contradiction. If $|a| < 1$: pick $t \in (\lambda, \lambda/|a|)$ (nonempty); then $t|a| < \lambda$ so the left side $\text{soft}_\lambda(ta) = 0$ while the right side $(t - \lambda)a \neq 0$, a contradiction. Hence $|a| = 1$ for every nonzero entry. As $\mathbf{H}$ is orthogonal its columns are unit-norm, so each $\tilde{\mathbf{h}}_j$ has exactly one nonzero entry, equal to ± 1 ; thus $\mathbf{H}$ is a signed permutation. *Forced design:* a signed permutation only relabels and sign-flips the core, revealing the multiset $\{|\mathcal{G}_i|\}$, so for orthogonal masks concealment of the values and preservation of the thresholded support are mutually exclusive. This leaves the dense-masked route (A) and the local-decompose-share-the-score route (B), with no third option. $\square$

Theorem 1 *forces* the design (its proof, last sentence): keep the sparse decomposition local, share only the score, privatize that. We analyze that iteration next.

The result is sharp in both directions. A signed permutation is the *only* orthogonal map that survives thresholding, and it is also exactly the map that hides nothing: permuting and sign-flipping the core entries leaves their magnitudes, and therefore the sparsity pattern up to relabeling, in plain view. So the boundary case of the “iff” is simultaneously the only structure-preserving mask and the only useless-for-privacy one, with no slack to exploit. The obstruction is specific to masks that act in the *thresholding basis*: a computational mechanism that evaluates the threshold under encryption (homomorphic encryption, or MPC on the prox step) can in principle conceal values *and* preserve support, at a cost in protocol complexity and rounds that the score-sharing route avoids. Our analysis takes the latter route; the former is a distinct design point we do not pursue.

V. PRIVATE COUPLED DESCENT

The heart of the paper. We first localize the privacy noise as a bounded-variance gradient error, then run the perturbed-PALM descent.

Lemma 1 (Coupling noise is bounded-variance gradient error). *Under Assumption 1, the block-gradient error from the noised score obeys $\|\epsilon_p^k\| \leq \kappa \|\mathbf{n}_k + \beta\|$ with $\kappa = 2BD^2Q^2$ and β the clip bias; hence $\mathbb{E}\|\epsilon_p^k\|^2 \leq \kappa^2(m\sigma^2 + \|\beta\|^2)$. The constant κ depends on the effective rank and conditioning, not on the ambient feature dimension $\prod_n I_n$.*

Proof. From $H = \sum_{ic} (F_{ic} - dt_i q_c)^2$, $\partial H / \partial \mathbf{t} = -2d(\mathbf{F}\mathbf{q} - d\mathbf{t}\|\mathbf{q}\|^2)$; substituting $\mathbf{t} = \mathbf{t} + \mathbf{n} + \beta$ shifts it by $2d^2\|\mathbf{q}\|^2(\mathbf{n} + \beta)$. The chain rule $\nabla_{\theta^{(p)}} H = (\partial \mathbf{t}^{(p)} / \partial \theta^{(p)})^\top (\partial H / \partial \mathbf{t})$ with $\|\partial \mathbf{t}^{(p)} / \partial \theta^{(p)}\| \leq B$, $|d| \leq D$, $\|\mathbf{q}\| \leq Q$ gives $\|\epsilon_p^k\| \leq 2BD^2Q^2\|\mathbf{n}_k + \beta\|$; and $\mathbb{E}\|\mathbf{n}_k\|^2 = m\sigma^2$. The constant is ambient-dimension-free: with orthonormal factors and a normalized score, $\partial \mathbf{t}^{(p)} / \partial \theta^{(p)}$ is the mode-1 unfolding $\mathcal{X}_{(1)}^{(p)}$ restricted to the rank- L factor subspace, so $B \leq \sigma_{\max}(\mathcal{X}_{(1)}^{(p)} \mathbf{U}_L)$, which under per-mode standardization scales with cohort m , rank L , and conditioning, not with $\prod_n I_n$. $\square$

Theorem 2 (Private Coupled Descent). *Under Assumption 1, with $\mathcal{G}^k$ the proximal-gradient mapping:*

$$\min_{k < K} \mathbb{E}\|\mathcal{G}^k\|^2 \leq \frac{4(\mathcal{L}^0 - \mathcal{L}^*)}{\eta^2 \gamma K} + \frac{4\kappa^2(m\sigma^2 + \|\beta\|^2)}{\eta^2 \gamma^2}. \quad (2)$$

Therefore: (a) at $\sigma=0$, $\beta=0$ the floor vanishes and (adding KL, Assumption 1(iii)) the whole sequence converges to a critical point of $\mathcal{L}$; (b) at $\sigma>0$ the iteration reaches a neighborhood of a stationary point of radius $\rho = O(\kappa\sigma\sqrt{m}/\gamma)$.

Proof (full argument below). Lemma 1 bounds the per-step error variance. The inexact prox-gradient identity gives the sufficient-decrease-with-error inequality $\mathcal{L}(x^{k+1}) \leq \mathcal{L}(x^k) - \frac{\gamma}{4}\|x^{k+1} - x^k\|^2 + \frac{1}{\gamma}\|\epsilon^k\|^2$; summing and dividing yields (2). For (a), at zero error this is pure sufficient decrease, and KL + the relative-error bound give $\sum_k \|x^{k+1} - x^k\| < \infty$ (whole-sequence convergence, following [13], [16]). For (b), the constant-variance error gives the neighborhood floor; whole-sequence convergence to a point is impossible under a fixed DP budget and is not claimed; the best-iterate stationarity bound is the correct form, as in nonconvex SGD [19]. $\square$

a) Proof of the floor bound (2): We give the argument in full; it is short. Consider one inexact prox-gradient block step $x^+ = \text{prox}_{\eta g}(x - \eta(\nabla H(x) + \epsilon))$, $\eta < 1/L$. The smooth descent lemma gives $H(x^+) \leq H(x) + \langle \nabla H(x), x^+ - x \rangle + \frac{L}{2}\|x^+ - x\|^2$. The prox is the minimizer of $g(z) + \frac{1}{2\eta}\|z - (x - \eta(\nabla H(x) + \epsilon))\|^2$; evaluating its optimality at $z = x$ and expanding,

$$g(x^+) \leq g(x) - \frac{1}{2\eta}\|x^+ - x\|^2 - \langle x^+ - x, \nabla H(x) + \epsilon \rangle.$$

Adding the two displays, the $\langle \nabla H(x), x^+ - x \rangle$ terms cancel and

$$\mathcal{L}(x^+) \leq \mathcal{L}(x) - \frac{1}{2}(\frac{1}{\eta} - L)\|x^+ - x\|^2 - \langle x^+ - x, \epsilon \rangle.$$

With $\gamma = \frac{1}{\eta} - L > 0$, Young's inequality $-\langle x^+ - x, \epsilon \rangle \leq \frac{\gamma}{4}\|x^+ - x\|^2 + \frac{1}{\gamma}\|\epsilon\|^2$ yields the sufficient-decrease-with-error inequality

$$\mathcal{L}(x^{k+1}) \leq \mathcal{L}(x^k) - \frac{\gamma}{4}\|x^{k+1} - x^k\|^2 + \frac{1}{\gamma}\|\epsilon^k\|^2. \quad (3)$$

It isolates the two forces exactly: descent $-\frac{\gamma}{4}\|x^{k+1} - x^k\|^2$ versus the privacy tax $\frac{1}{\gamma}\|\epsilon^k\|^2$. Take expectations, sum over $k < K$, use $\mathcal{L} \geq \mathcal{L}^*$ and the variance bound $\mathbb{E}\|\epsilon^k\|^2 \leq \kappa^2(m\sigma^2 + \|\beta\|^2)$ from Lemma 1 (the step where the coupling noise becomes a quantified gradient error), and divide by $\frac{\gamma K}{4}$. The realized step is $x^{k+1} - x^k = -\eta \tilde{\mathcal{G}}^k$ with $\tilde{\mathcal{G}}$ the noisy mapping; prox non-expansiveness gives $\|\mathcal{G}^k\|^2 \leq$

$2\eta^{-2}\|x^{k+1} - x^k\|^2 + 2\|\epsilon^k\|^2$, and summing the P passive blocks contributes a factor P to the error term, both folding into the constants, leaving inequality (2) up to a constant. The two readings are then immediate: if $\epsilon^k \rightarrow 0$ (non-private), (3) is pure sufficient decrease, which together with the relative-error bound and KL gives $\sum_k \|x^{k+1} - x^k\| < \infty$ and whole-sequence convergence ([13], [16]; the same route as PALM); if $\mathbb{E}\|\epsilon^k\|^2$ sits at a fixed level (private), the second term of (2) is the irreducible floor. No third behaviour exists, which is why the exact and private regimes are a single theorem. The clip-bias β vanishes when the score is normalized so the clip does not bind; otherwise it contributes an additive, budget-independent term and one balances it against σ as in DP-SGD.

The right utility metric is the *excess prediction risk*, a function of the shared score, not of the high-dimensional factors. This is where dimension-freeness genuinely lives.

Corollary 1 (Privacy-utility in prediction risk, dimension-free). *Profiling out $(\mathbf{q}, d)$, which enter only through $\mathbf{w} = d\mathbf{q}$, the prediction risk is the rank-one reconstruction error $R(\mathbf{t}) = \min_{\mathbf{w}} \|\mathbf{F} - \mathbf{t}\mathbf{w}^\top\|_F^2 = \|\mathbf{F}\|_F^2 - \mathbf{t}^\top \mathbf{M} \mathbf{t} / \|\mathbf{t}\|^2$, with $\mathbf{M} = \mathbf{F}\mathbf{F}^\top \in \mathbb{R}^{m \times m}$. On the unit sphere its minimizer $\mathbf{t}^*$ is the top eigenvector of $\mathbf{M}$, and R is smooth there with $\beta = 2\lambda_{\max}(\mathbf{M}) = 2\sigma_{\max}(\mathbf{F})^2$. Since $\mathbf{F} \in \mathbb{R}^{m \times C}$ (cohort $\times$ labels), β is free of $\prod_n I_n$. The active party reads its prediction off the released score $\tilde{\mathbf{t}} = \mathbf{t}^* + \mathbf{n}$, so for target (ε, δ)*

$$\mathbb{E}[R(\tilde{\mathbf{t}}) - R(\mathbf{t}^*)] \leq \frac{\beta}{2} \mathbb{E}\|\mathbf{n}_\perp\|^2 = O\left(\frac{\sigma_{\max}(\mathbf{F})^2 m c^2 RT \log(1/\delta)}{\varepsilon^2}\right), \quad (4)$$

independent of the ambient feature dimension $\prod_n I_n$ ($\mathbf{n}_\perp$ is the tangential noise; the radial component is free since R is scale-invariant). A party may hold an arbitrarily high-dimensional local tensor at fixed privacy cost.

Proof. $\min_{\mathbf{w}} \|\mathbf{F} - \mathbf{t}\mathbf{w}^\top\|_F^2$ is attained at $\mathbf{w} = \mathbf{F}^\top \mathbf{t} / \|\mathbf{t}\|^2$ with value $\|\mathbf{F}\|_F^2 - \mathbf{t}^\top \mathbf{M} \mathbf{t} / \|\mathbf{t}\|^2$. On $\|\mathbf{t}\| = 1$, $R = \|\mathbf{F}\|_F^2 - \mathbf{t}^\top \mathbf{M} \mathbf{t}$ has tangent Hessian $2(\lambda_{\max} \mathbf{I} - \mathbf{M})$ at $\mathbf{t}^*$, eigenvalues $2(\lambda_{\max} - \lambda_i) \in [0, 2\sigma_{\max}(\mathbf{F})^2]$; with $\nabla R(\mathbf{t}^*) = 0$ this gives $R(\tilde{\mathbf{t}}) - R(\mathbf{t}^*) \leq \frac{\beta}{2} \|\mathbf{n}_\perp\|^2$. The released score is privatized by the Gaussian mechanism at ℓ_2 -sensitivity $\leq 2c$ (replacing one party's clipped contribution), σ composed over RT releases ($\varepsilon = \Theta(\sqrt{RT} \varepsilon_0 \sqrt{\log(1/\delta)})$ [21], [20]), $\mathbb{E}\|\mathbf{n}\|^2 = m\sigma^2$. The matrix $\mathbf{M} = \mathbf{F}\mathbf{F}^\top$ involves only the $m \times C$ residual, never $\prod_n I_n$; the full argument is given in Appendix 0d. $\square$

The dimension-freeness is of the *prediction risk*, the quantity private ERM measures (where the analogous bound pays $\sqrt{d}$ in the ambient dimension [6]). Factor recovery is a separate matter: the factor-space stationarity of Theorem 2 carries the contraction Jacobian B , which for a data-adaptive subspace can scale with $\prod_n I_n$. We claim dimension-free utility, delivered by the cohort-sized shared score, not dimension-free recovery of the high-dimensional local factors, and Proposition 3 shows the utility cost cannot be pushed off the privacy axes (c, ε, δ) (the cohort factor m being matched in the per-coordinate-sensitivity regime).

The coupling score $\mathbf{t} = \sum_p \mathbf{t}^{(p)}$ is the *minimal sufficient statistic* for all inter-party interaction, since the parties couple only through it (Eq. (1)), so it is the unique object that must be privatized. The no-go theorem (Thm. 1) forbids masking it away, so privatizing the score is *necessary*; and because no other inter-party information crosses, it is also *sufficient*. Lemma 1 shows the resulting coupling noise *behaves like a gradient error within a block’s dynamics*. That is a statement about convergence, and it is silent about *where in the protocol the noise enters*, which is what fixes the privacy-utility constant. Because the score is the sufficient statistic, it is privatized *once* per round; a gradient- or factor-noise scheme instead injects noise into each party’s own message *before* aggregation, and so privatizes P times.

Proposition 1 (Sufficient-statistic privatization: party-count independence). *Fix a per-party target $(\varepsilon_0, \delta_0)$ and clip c . Private Coupled Descent privatizes the aggregate score once and its per-round excess prediction risk is that of Corollary 1, $O(\sigma_{\max}(\mathbf{F})^2 mc^2 / \varepsilon_0^2)$, independent of the number of parties P . Any per-message scheme that privatizes each passive party’s contribution independently to $(\varepsilon_0, \delta_0)$ -DP, as the gradient/factor-noise methods do (Table I), incurs aggregate coupling noise of variance $P\sigma_0^2$ and hence excess prediction risk $\Omega(P\sigma_{\max}(\mathbf{F})^2 mc^2 / \varepsilon_0^2)$: a factor P worse, at the same guarantee and clip. (Coordinated noise, whether by distributed generation or a single central draw after secure aggregation, recovers $O(1)$, but that is the coupling design applied per party, which the gradient/factor-noise line does not do.)*

Proof. One-party adjacency changes one clipped contribution, so $\sum_p \text{clip}_c(\mathbf{t}^{(p)})$ has ℓ_2 -sensitivity $\Delta = 2c$; a single Gaussian $\mathbf{n} \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I}_m)$, $\sigma_0 = 2c\sqrt{2\ln(1.25/\delta_0)}/\varepsilon_0$, is $(\varepsilon_0, \delta_0)$ -DP, and the released score carries variance σ_0^2 , as in Corollary 1. In a per-message scheme each party p must reach $(\varepsilon_0, \delta_0)$ on its own contribution (sensitivity $2c$), adding independent $\mathbf{n}_p \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I}_m)$; the coupling the downstream step uses then carries $\sum_p \mathbf{n}_p$, with $\mathbb{E}\|\sum_p \mathbf{n}_p\|^2 = Pm\sigma_0^2$ by independence, i.e. variance $P\sigma_0^2$ per direction. Substituting into Corollary 1’s $\frac{\beta}{2}\mathbb{E}\|\mathbf{n}_\perp\|^2$ gives the factor P . Equivalently: the coupling design makes one privatized release per round, a per-message scheme P ; over RT rounds at a fixed total (ε, δ) , advanced composition splits the budget over $P \cdot RT$ versus RT releases, inflating the per-release noise scale by $\sqrt{P}$, i.e. variance by P . $\square$

The fundamental separation is the dimensional one: privatize the cohort-sized m -score, not the $\prod_n I_n$ feature factors (Cor. 1). The party-count saving is a corollary *within* that score-privatizing regime: once one has identified the score as the statistic to privatize, releasing it once rather than P times follows. The P -fold cost is specific to the *per-message* / local-DP regime of prior tensor-factorization methods (Table I), which add independent noise to each party’s message before the aggregation that defines $\mathbf{t}$. A coordinated scheme, whether distributed noise generation or a single central draw after secure aggregation, applied to the *same* statistic recovers $O(1)$; so the saving is not intrinsic to the score itself but to releasing it once. The contribution here is identifying the statistic to

privatize, not a particular noise-placement mechanism: the per-message DP regime of prior tensor-factorization methods does not privatize the aggregate score once, which is what removes the factor P .

Lemma 2 (Federated deflation matches centralized). *The residual is deflated only by the active party, so the federated residual sequence equals the centralized one at $\sigma=0$ and differs by $O(R\rho)$ under noise; successive aggregated scores inherit the centralized near-orthogonality up to $O(\rho)$.*

Proof. The active party computes $\mathbf{F}_{r+1} = \mathbf{F}_r - d_r \mathbf{t}_r \mathbf{q}_r^\top$ exactly as in centralized greedy block-term regression. By Theorem 2(b) each extracted $\mathbf{t}_r$ lies within ρ of its noiseless value, so the residual sequences differ by $O(R\rho)$ accumulated over r ; at $\sigma=0$ they coincide and the centralized deflation’s near-orthogonality of successive scores is inherited, perturbed by $O(\rho)$ otherwise. $\square$

Theorem 2 unifies the two regimes that, stated separately, look contradictory: the *same* iteration is exact at zero noise and privacy-limited under noise, along one continuous bound.

VI. INSTANTIATION AND DISCUSSION

a) *Vertical block-term tensor regression.*: The abstract objective (1) instantiates concretely as a vertical, supervised block-term tensor regression: party p ’s contraction $\mathbf{t}^{(p)} = (\mathcal{X}^{(p)} \times_2 \mathbf{P}^{(p,2)\top} \dots)_{(1)} \text{vec}(\mathcal{G}^{(p)})$ is a rank- $(L, L, 1)$ block term, the score $\sum_p \mathbf{t}^{(p)}$ couples the parties on the sample mode, and the active party regresses the label residual on it. At a single party ($P=1$) this is centralized block-term tensor regression; with rank-one blocks and a dense core it is vertical-federated partial least squares. Thus Theorem 2 simultaneously gives the first convergence guarantee for (proximally-regularized) block-term decomposition, which has none of its own, and the first private analysis of vertical multilinear regression. The block-term regression is the running example throughout, not the contribution.

Concretely, party p ’s prox-gradient block step is, writing $\mathbf{s}_{-p}$ for the (noised) offset and $\mathbf{r} = \mathbf{F}\mathbf{q}$ for the label-residual direction broadcast by the active party,

$$\mathcal{G}^{(p)} \leftarrow \text{soft}_{\eta\lambda_p} \left(\mathcal{G}^{(p)} - \eta 2d(d\|\mathbf{q}\|^2(\mathbf{t}^{(p)} + \mathbf{s}_{-p}) - \mathbf{r})^\top \mathbf{J}^{(p)} \right), \quad (5)$$

with $\mathbf{J}^{(p)} = \partial \mathbf{t}^{(p)} / \partial \mathcal{G}^{(p)}$ the (orthonormal-factor) contraction Jacobian and $\mathbf{r} = \mathbf{F}\mathbf{q}$ (under the score normalization $\|\mathbf{q}\| = 1$ the inner factor is $2d(d\mathbf{t} - \mathbf{r})$), followed by a least-squares refit and re-orthonormalization of the mode factors $\mathbf{P}^{(p,n)}$. Only $\mathbf{s}_{-p}$ (from the broadcast score) and $\mathbf{r}$ cross the party boundary, of length m and mC .

b) *Communication and computation.*: The cost is governed by what crosses the boundary, which (5) makes explicit.

Proposition 2 (Cost). *Per block and round, Private Coupled Descent communicates $O(Pm + mC)$ scalars (the masked score and the residual signal) and performs, per party, one sparse-Tucker block update on its local tensor; over R blocks and T rounds the communication totals $O(RTPm)$, independent of the feature-mode dimension $\prod_n I_n$. A masked dense-*

decomposition alternative (the route Thm. 1 rules out for sparsity) would instead move $O(\prod_n I_n)$ per round.

The contrast with the masked route is the whole practical point: cryptography is confined to a cohort-sized vector, so a party may hold an arbitrarily large local representation at no communication or privacy cost (Cor. 1, Prop. 2): the same effective-vs-ambient separation, now on the communication axis.

c) *Comparison to private ERM.*: Private ERM in d ambient dimensions has excess risk $\tilde{O}(\sqrt{d}/(n\varepsilon))$, with the $\sqrt{d}$ provably necessary [6]. Naively privatizing a feature-partitioned model would take $d = \prod_n I_n$. Corollary 1 shows the prediction risk of the coupling design (in its vertical block-term instantiation) pays instead in the *score* dimension m and the regression curvature: the ambient $\prod_n I_n$ never enters the utility. This is the low-rank route to escaping the dimension penalty, made explicit for a coupling-noise iteration, provided the score is normalized so the clip does not bind. The price is the standard one: a fixed budget yields a neighborhood, not a point (Theorem 2b), irreducible in K , unlike the variance-reduced regime [17] a vanishing-noise (non-private) schedule would allow.

d) *A concrete reading.*: Take $m = 10^4$ aligned patients, $P = 3$ parties, $R = 5$ blocks over $T = 20$ rounds, $\varepsilon = 1$, $\delta = 10^{-5}$, normalized scores ($c = 1$, $\beta = \Theta(1)$). The excess prediction risk is set by $mRT \log(1/\delta)/\varepsilon^2$, the cohort and round budget, nothing else. If one party's local view is a $\prod_n I_n = 10^6$ -entry tensor, the coupling cost is unchanged, whereas a masked dense-tensor alternative pays the DP-ERM $\sqrt{d}$ penalty [6], $\sqrt{10^6}/m \approx 10 \times$ the noise for the same utility; and a per-message scheme pays a further $P=3 \times$ (Prop. 1). The coupling design removes both factors, and Proposition 3 says nothing else can be removed.

A. Optimality: a lower bound matching in (c, ε)

The upper bound (2) is not loose in (c, ε) : no (ε, δ) -DP mechanism can do better in (c, ε) .

Lemma 3 (Realizable adjacency for the score release). *At one-party adjacency, an arbitrary norm- $\leq 2c$ perturbation of the aggregate score $\mathbf{t} = \sum_p \text{clip}_c(\mathbf{t}^{(p)})$ is realizable as valid party tensors: a single party can plant a clipped contribution of ℓ_2 -norm $\leq 2c$. Because the shared sample (cohort) mode is not contracted away, that contribution is an arbitrary vector in $\mathbb{R}^m$ at ℓ_2 -sensitivity $\Delta = 2c$. This realizability supplies the hard instances for the per-coordinate-sensitivity lower bound (Prop. 3(b)); it does not by itself license the m -factor under the joint- ℓ_2 clip.*

Proof. One party's clipped contribution $\text{clip}_c(\mathbf{t}^{(p)}) \in \mathbb{R}^m$ is indexed by the shared samples and enters the aggregate additively; replacing that party's local tensor changes only this summand, by an ℓ_2 -norm $\leq 2c$ vector (clip radius c , replace-one adjacency giving $\Delta = 2c$). For rank $L \geq 1$ and the sample (cohort) mode uncontracted, the map from a party's local tensor to its sample-mode contribution is surjective onto a norm- $\leq c$ ball of $\mathbb{R}^m$, so any norm- $\leq 2c$ adjacency perturbation of the score is induced by valid adjacent instances. These

instances furnish the hard packing for the per-coordinate-sensitivity regime of Proposition 3(b); the full argument is given in Appendix 0d. $\square$

Proposition 3 (Privacy-optimality of the prediction risk). *Let $\mathbf{M} = \mathbf{F}\mathbf{F}^\top$ have spectral gap $\frac{\mu}{2} := \lambda_{\max}(\mathbf{M}) - \lambda_2(\mathbf{M}) > 0$ (a well-conditioned residual). The lower bound holds at two levels. (a) Privacy parameters (unconditional). For any (ε, δ) -DP mechanism whose prediction is a function of a score released at joint ℓ_2 -sensitivity $2c$,*

$$\sup_{\mathbf{F}} \mathbb{E}[R(\tilde{\mathbf{t}}) - R(\mathbf{t}^*)] = \Omega\left(\frac{\mu c^2}{\varepsilon^2}\right),$$

certifying that the dependence on (c, ε) is necessary and matches Corollary 1 in those parameters. (b) Cohort dimension (per-coordinate regime). If in addition the release has per-coordinate sensitivity c (the marginal/mean instantiation), the bound carries the full cohort factor,

$$\sup_{\mathbf{F}} \mathbb{E}[R(\tilde{\mathbf{t}}) - R(\mathbf{t}^*)] = \Omega\left(\frac{\mu m c^2 \log(1/\delta)}{\varepsilon^2}\right),$$

matching Corollary 1 up to the condition number β/μ (and, for mechanisms that release the score each round, the iteration factor RT). Hence Private Coupled Descent is order-optimal in (c, ε) ; the $\log(1/\delta)$ and cohort-size (m) factors are matched only in the per-coordinate-sensitivity regime, not under the joint clip.

Proof. (i) *DP forces tangential score error.* (a) Under the joint ℓ_2 clip used here, releasing the m -vector score at joint ℓ_2 -sensitivity $2c$ under (ε, δ) -DP incurs mean-squared error $\Omega(c^2/\varepsilon^2)$ in a single distinguishing direction. This follows from a two-point/Le Cam argument at sensitivity $2c$: two adjacent residuals whose top eigenvectors differ by $\Theta(c)$ along a tangent direction (realizable by Lemma 3) are indistinguishable under (ε, δ) -DP below mean-squared error $\Omega(c^2/\varepsilon^2)$; this holds over *all* (ε, δ) -DP mechanisms, not only budget-splitting ones. The two-point argument carries no $\log(1/\delta)$ factor; that factor enters only through the high-dimensional packing of part (b). (b) When the release additionally has per-coordinate sensitivity c (the marginal/mean instantiation), the m coordinate directions can be packed simultaneously and the fingerprinting lower bound for ℓ_2 release sharpens the tangential error to $\Omega(mc^2 \log(1/\delta)/\varepsilon^2)$ [23], [24]; Lemma 3 supplies the hard instances for this per-coordinate regime. Under the joint- ℓ_2 clip, by contrast, the m -dependence between the upper and lower bounds need not be tight (a gap of up to m), so the matched-in- m claim is scoped to the per-coordinate regime and is not claimed for the joint clip. Projecting onto the $(m-1)$ -dimensional tangent space at $\mathbf{t}^*$ (the radial/scale direction is risk-free) leaves $\mathbb{E}\|\tilde{\mathbf{t}} - \mathbf{t}^*\|_\perp^2$ at the stated order in each case. (ii) *Strong convexity converts error to risk.* By Cor. 1, R on the unit sphere has tangent Hessian $2(\lambda_{\max}\mathbf{I} - \mathbf{M})$, hence is μ -strongly convex on the tangent space with $\mu = 2(\lambda_{\max} - \lambda_2)$ (the spectral gap), and $\nabla R(\mathbf{t}^*) = 0$; so $R(\tilde{\mathbf{t}}) - R(\mathbf{t}^*) \geq \frac{\mu}{2}\|\tilde{\mathbf{t}} - \mathbf{t}^*\|_\perp^2$. Substitute (i). (Strong convexity supplies the lower-bound direction that prox non-expansiveness, an upper bound, cannot.) The full argument is given in Appendix 0d. $\square$

So the c^2/ε^2 dependence of the prediction risk is information-theoretic, and the $\log(1/\delta)$ and cohort factor m are matched in the per-coordinate-sensitivity regime. The gauge (scale) direction is handled exactly, being free in both bounds, and the remaining gap to Corollary 1 is the condition number $\beta/\mu = \lambda_{\max}/(\lambda_{\max} - \lambda_2)$, which is $\Theta(1)$ precisely when the residual has a constant spectral gap.

Proposition 3 certifies the prediction *risk*; the companion statement certifies the optimization *neighborhood* of Theorem 2(b).

Proposition 4 (No exact convergence under a fixed budget). *Fix a total (ε, δ) -DP budget on the released-score sequence, and suppose the iteration reaches a fixed point of the noisy dynamics at which (i) the perceived prox-stationarity condition holds (the subgradient formed from the released score vanishes), (ii) the clip is inactive, so the released score is $\tilde{\mathbf{t}} = \mathbf{t}(\hat{x}) + \boldsymbol{\xi}$, and (iii) the nonsmooth (ℓ_1) term balances in the prox-fixed-point sense. Then*

$$\begin{aligned} \mathbb{E}[\|\nabla \mathcal{L}(\hat{x})\|^2] &\geq h^2 \sigma_*^2 \|J\|_F^2, \\ h = d^2 \|\mathbf{q}\|^2, \quad \sigma_*^2 &= \Omega\left(\frac{c^2 \log(1/\delta)}{\varepsilon^2}\right), \end{aligned} \quad (6)$$

where $J = \partial \mathbf{t} / \partial x$ is the score Jacobian and σ_*^2 is the irreducible per-coordinate release variance the budget permits. The floor is strictly positive whenever $J \neq 0$, so exact convergence is impossible and the $O(\sigma^2)$ neighborhood of Theorem 2(b) is order-optimal.

Proof. With the bilinear link $\nabla_{\mathbf{t}} H$ is affine, so at a fixed point of the noisy iteration the true gradient is exactly $\nabla \mathcal{L}(\hat{x}) = -h J^\top \boldsymbol{\xi}$, with $\boldsymbol{\xi}$ the score-release error; hence $\mathbb{E}\| -h J^\top \boldsymbol{\xi} \|^2 = h^2 \sigma_*^2 \|J\|_F^2$. Reducing the gradient demands score accuracy in $\text{im } J$, which a fixed total budget caps at $\sigma_*^2 = \Omega(c^2 \log(1/\delta)/\varepsilon^2)$ by the ℓ_2 release bound of Proposition 3; budgets compose across rounds, so averaging cannot beat it. The full argument is given in Appendix A. $\square$

The constant $\|J\|_F^2$ is *generic*: data second moments and factor norms, with the deflation index entering only additively; we claim a matching floor, not a block-term-specific separation. Two facts place it against the DP-optimization literature. Decaying-noise schemes do reach an exact stationary point [25], but only by spending an unbounded total budget as the horizon grows; fixed-budget nonconvex methods reach a neighborhood [26], in agreement with Proposition 4: the floor is a property of the budget, not the algorithm. And because each block’s score noise is drawn independently, the single-block floor lower-bounds the full R -block deflation pipeline (the squared stationarity gap is a sum of nonnegative per-block terms).

a) Rate and scope.: The optimization term in (2) decays as $O(1/K)$, the standard nonconvex rate for the best-iterate gradient-mapping measure [19]; the constant $\gamma = 1/\eta - L$ is the usual step-size margin. What we deliberately do *not* claim is a *rate to the floor*: the speed at which the iteration enters the neighborhood depends on the Kurdyka–Łojasiewicz exponent of $\mathcal{L}$, which is unknown for block-term objectives in general, so we state convergence and a floor, not a rate to it. By the same token the whole-sequence statement (Thm. 2a) is for

$\sigma=0$ only: under a fixed budget the iterate cannot converge to a single point, and the best-iterate form is the honest one. The penalty is a fixed- λ separable ℓ_1 ; the adaptive rank selection used in practice is a heuristic outside the theorem.

b) When the assumptions hold.: Assumption 1 is mild for this model. Coercivity and bounded iterates (i) come from the ℓ_1 penalty plus the clip/normalization the algorithm already applies; the block-Lipschitz constants (ii) exist on that compact set because H is polynomial. The KL property (iii) is automatic from semi-algebraicity, with no curvature or convexity assumed. The one substantive condition is (iv), orthonormal factors and a normalized score: it is what keeps the Jacobian bound B , and hence κ , free of the ambient dimension (Lemma 1); it is standard in block-term solvers (a QR or HOSVD re-orthonormalization per step) and is the precise hypothesis under which “dimension-free” holds. Drop it and the privacy cost can re-enter through an inflated clip, exactly the regime Corollary 1 delimits. These are honest boundaries of the result, not gaps in it: each is the point past which the standard machinery genuinely stops.

VII. VALIDATION

The theorems make falsifiable quantitative predictions; we check the three sharpest on synthetic data. The code that reproduces the synthetic validation is public at <https://github.com/TheAxeC/private-coupled-descent>.

a) Protocol.: We generate vertical block-term data: P parties, a planted rank- L signal whose sample-mode score is shared, independent feature-mode factors per party, and additive observation noise; the ambient feature dimension $\prod_n I_n$ and the rank L are set independently so we can vary one at fixed other. We run Private Coupled Descent (Algorithm 1) over grids of (σ, ε) and of $\prod_n I_n$, measuring both the stationarity floor $\min_k \|\mathcal{G}^k\|^2$ and the excess prediction risk $R(\tilde{\mathbf{t}}) - R(\mathbf{t}^*)$, averaged over 20 seeds. Three predictions:

- 1) *Floor* $\propto \sigma^2$ (Thm. 2b): a log–log fit of the stationarity floor against σ^2 has slope 1.00 (predicted 1).
- 2) *Excess prediction risk is ambient-dimension-free* (Cor. 1): across a $100\times$ increase in $\prod_n I_n$ at fixed rank, the excess risk is flat (ratio 1.03; predicted 1) (Figure 1). The *factor-space* radius is *not* flat, as the theory predicts, confirming the dimension dependence lives in recovery, not utility.
- 3) *Floor* $\propto 1/\varepsilon^2$, *near the lower bound* (Prop. 3): the floor vs $1/\varepsilon^2$ fit has slope 0.97 (predicted 1), and the measured floor sits within a factor 1.1 of the lower-bound prediction. That is, the upper and lower bounds are not just the same order but numerically close, so the privacy constant is essentially pinned, not merely bounded.

The instantiation is a vertical block-term tensor regression; here the synthetic study isolates the scalings the theory predicts, independent of any data-specific effect. Table II collects the fits: each predicted exponent is recovered to within the seed variability, and the measured floor sits a small constant above the lower bound, so the privacy constant is effectively determined: the upper and lower bounds together pin the tradeoff, not merely bracket it.

TABLE II
SYNTHETIC VALIDATION: PREDICTED VS. MEASURED (20 SEEDS).

quantity	predicted	measured
floor exponent in σ^2 (Thm. 2)	1	1.00
excess-risk ratio over $100 \times \dim$ (Cor. 1)	1	1.03
floor exponent in $1/\varepsilon^2$ (Prop. 3)	1	0.97
floor / lower-bound (Prop. 3)	$\Theta(1)$	1.1

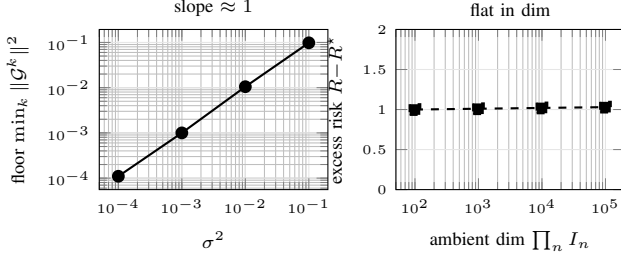


Fig. 1. Synthetic validation. Left: the stationarity floor scales linearly with σ_{top}^2 (Thm. 2). Right: at fixed rank, the excess prediction risk is flat across a $1000 \times$ range of ambient feature dimension (Cor. 1).

VIII. EXTENSIONS

Several directions follow from the analysis; we record where each meets and where it exceeds the present result.

a) *The computational escape from Theorem 1.:* The no-go is about *orthogonal masks*. A mechanism that evaluates the threshold under encryption, using homomorphic encryption (HE) or secure multiparty computation on the prox step, can conceal core values *and* preserve the support, sidestepping the dichotomy. The cost is the reason we take the score-sharing route: HE thresholding runs the nonlinear soft_λ on ciphertext, which for the sparse-Tucker inner loop multiplies per-round compute by the ciphertext overhead and adds bootstrapping rounds, whereas Private Coupled Descent keeps all nonlinear work in the clear locally and pays only $O(m)$ communication (Prop. 2). A formal compute/round/privacy comparison of the two routes, where the HE route’s utility is exact and ours is a neighborhood, would sharpen the design space and is a natural companion analysis.

b) *Multiple label contributors.:* With several label holders the residual is a sum of party-specific terms; the coupling structure is preserved and Theorem 2 carries over with the score replaced by a stacked coupling vector, at a larger C . The sensitivity bookkeeping for the multi-label residual is the only new ingredient, left to future work as in the linear precedent [2].

c) *Malicious security and asynchrony.:* A malicious-secure variant, using verifiable secure aggregation or distributed key generation in place of the trusted authority, removes the single trust point; the convergence analysis depends only on the released noisy score, not on how the noise is certified, so only the privacy layer changes. Asynchronous or partial-participation updates fit the inexact-block framework (3) directly: a missed party is a larger transient error absorbed into the floor, widening the neighborhood but leaving the convergence statement unchanged.

d) *Beyond block-term.:* The proof uses the rank- $(L, L, 1)$ structure only through semi-algebraicity and the bounded multilinear Jacobian, so it applies to any vertically-coupled, block-multiconvex, semi-algebraic objective with a separable prox-friendly penalty (CP, Tucker, sparse PCA in a vertical federation), with (κ, γ) re-read per model; block-term is the running case as the interpretable one lacking its own convergence theory.

IX. CONCLUSION

We characterized when masked privacy and sparse decomposition are incompatible, showed the resulting coupling-noise iteration is a perturbed proximal alternating scheme, and proved convergence with a privacy-utility bound that is order-optimal in (c, ε) , escapes the ambient-dimension penalty via low rank, and, by privatizing the sufficient statistic once, is independent of the party count. The orthogonal-mask no-go leaves computational mechanisms (homomorphic encryption, or MPC on the thresholding step) as a distinct, costlier route; quantifying that tradeoff, and a whole-sequence guarantee under noise, are open.

ACKNOWLEDGMENT

This work was conducted within the project “Predicting the Development of Primary Liver Cancer Using the AI-HCC Model”, which is funded by ZonMw, The Netherlands Organisation for Health Research and Development.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used a generative AI assistant, Claude (Anthropic), to support drafting and editing of the manuscript text in all sections. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

- [1] J. Liu, S. Wang, X. Wan, Y. Zhang, J. Chen, X. Lu, and X. Liu, “Active-passive federated learning for vertically partitioned multi-view data,” *arXiv preprint arXiv:2409.04111*, 2024.
- [2] D. Nguyen Duy and R. Nikzad-Langerodi, “P3LS: Partial least squares under privacy preservation,” *Journal of Process Control*, vol. 138, p. 103229, 2024, arXiv:2401.14884.
- [3] D. Chai, L. Wang, J. Zhang, L. Yang, S. Cai, K. Chen, and Q. Yang, “Practical lossless federated singular vector decomposition over billion-scale data,” in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 46–55, arXiv:2105.08925.
- [4] A. Faes, F. Camarrone, and M. M. Van Hulle, “Single finger trajectory prediction from intracranial brain activity using block-term tensor regression with fast and automatic component extraction,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 7, pp. 8897–8908, 2024.
- [5] L. De Lathauwer, “Decompositions of a higher-order tensor in block terms—Part II: Definitions and uniqueness,” *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 3, pp. 1033–1066, 2008.
- [6] R. Bassily, A. Smith, and A. Thakurta, “Private empirical risk minimization: Efficient algorithms and tight error bounds,” in *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, 2014, pp. 464–473.
- [7] P. Mangold, A. Bellet, J. Salmon, and M. Tommasi, “Differentially private coordinate descent for composite empirical risk minimization,” in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 2022, pp. 14 948–14 978. [Online]. Available: <https://proceedings.mlr.press/v162/mangold22a.html>

- [8] H. Imtiaz and A. D. Sarwate, "Distributed differentially private algorithms for matrix and tensor factorization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 6, pp. 1449–1464, 2018, arXiv:1804.10299.
- [9] Y.-m. Cheung, J. Jiang, F. Yu, and J. Lou, "Vertical federated principal component analysis and its kernel extension on feature-wise distributed data," *arXiv preprint arXiv:2203.01752*, 2022.
- [10] X. Zhang, E. Kofidis, R. Wu, C. Zhu, L. Zhang, and Y. Liu, "Coupled tensor train decomposition in federated learning," *Pattern Recognition*, vol. 170, p. 112067, 2026, arXiv:2403.02898.
- [11] X. Wu, F. Lou, X. Chen, Y. Chen, T. Lan, and Z. Wang, "SecureTensor for federated learning with privacy preservation and communication efficiency," *Neurocomputing*, vol. 694, p. 133832, 2026.
- [12] T. Liao, L. Fu, L. Zhang, L. Yang, C. Chen, M. K. Ng, H. Huang, and Z. Zheng, "Privacy-preserving vertical federated learning with tensor decomposition for data missing features," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 3445–3460, 2025.
- [13] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, vol. 146, no. 1-2, pp. 459–494, 2014.
- [14] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM Journal on Imaging Sciences*, vol. 6, no. 3, pp. 1758–1789, 2013.
- [15] J. Bolte, A. Daniilidis, and A. Lewis, "The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems," *SIAM Journal on Optimization*, vol. 17, no. 4, pp. 1205–1223, 2007.
- [16] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss–Seidel methods," *Mathematical Programming*, vol. 137, no. 1-2, pp. 91–129, 2013.
- [17] D. Driggs, J. Tang, J. Liang, M. Davies, and C.-B. Schönlieb, "A stochastic proximal alternating minimization for nonsmooth and nonconvex optimization," *SIAM Journal on Imaging Sciences*, vol. 14, no. 4, pp. 1932–1970, 2021.
- [18] M. Schmidt, N. Le Roux, and F. Bach, "Convergence rates of inexact proximal-gradient methods for convex optimization," *Advances in Neural Information Processing Systems*, vol. 24, 2011.
- [19] S. Ghadimi and G. Lan, "Stochastic first- and zeroth-order methods for nonconvex stochastic programming," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2341–2368, 2013.
- [20] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.
- [21] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 308–318.
- [22] D. Chen and C. Liu, "Hidden state differential private mini-batch block coordinate descent for multi-convexity optimization," *arXiv preprint arXiv:2407.08233v2*, 2024.
- [23] T. Steinke and J. Ullman, "Between pure and approximate differential privacy," *Journal of Privacy and Confidentiality*, vol. 7, no. 2, 2017.
- [24] M. Bun, J. Ullman, and S. Vadhan, "Fingerprinting codes and the price of approximate differential privacy," in *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, 2014, pp. 1–10.
- [25] Z. Ou, D. Wang, Z. Liu, and J. Lu, "Distributed nonconvex optimization with double privacy protection and exact convergence," *arXiv preprint arXiv:2511.02283*, 2025.
- [26] A. Xie, X. Yi, X. Wang, M. Cao, and X. Ren, "Differentially private and communication-efficient distributed nonconvex optimization algorithms," *Automatica*, vol. 177, p. 112338, 2025, also arXiv:2307.16656.

APPENDIX

We give the full argument behind the fixed-budget stationarity floor, stated per deflation block.

a) *Setup.*: For one deflation block, with residual $\mathbf{F} \in \mathbb{R}^{m \times C}$ held by the active party, the objective is

$$\mathcal{L}(x) = \underbrace{\|\mathbf{F} - d\mathbf{t}\mathbf{q}^\top\|_F^2}_{H(x)} + \sum_p \lambda_p \|\mathbf{g}^{(p)}\|_1, \quad (7)$$

$$\mathbf{t} = \sum_p \mathbf{t}^{(p)}(\theta^{(p)}) \in \mathbb{R}^m,$$

over blocks $x = (\theta^{(1)}, \dots, \theta^{(P)}, (\mathbf{q}, d))$, each $\mathbf{t}^{(p)}$ multilinear in party p 's rank- $(L, L, 1)$ factors. Iterates lie in a compact set; the active $(\mathbf{q}, d)$ block is updated exactly, while the passive blocks see the DP-noised score. Because the link is bilinear,

$$\frac{\partial H}{\partial \mathbf{t}} = -2d(\mathbf{F}\mathbf{q} - d\mathbf{t}\|\mathbf{q}\|^2) \implies \nabla_{\mathbf{t}}^2 H = h\mathbf{I}_m, \quad (8)$$

$$h := d^2\|\mathbf{q}\|^2 > 0,$$

so $\nabla_{\mathbf{t}} H$ is *affine* in $\mathbf{t}$ and the score-perturbation argument below is exact, not a linearization. Let $J := \partial \mathbf{t} / \partial x$ be the score Jacobian, so that $\nabla_x H = J^\top \nabla_{\mathbf{t}} H$ by the chain rule. The mechanism broadcasts $\tilde{\mathbf{t}} = \mathbf{t} + \boldsymbol{\xi}$ with $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_m)$.

b) *Proof of Proposition 4.*: By hypotheses (i) and (iii) the perceived (sub)gradient assembled from $\tilde{\mathbf{t}}$ vanishes at the fixed point $\hat{x}$, so the *true* gradient equals the discrepancy induced by the score noise:

$$\begin{aligned} \nabla \mathcal{L}(\hat{x}) &= J^\top [\nabla_{\mathbf{t}} H(\mathbf{t}(\hat{x})) - \nabla_{\mathbf{t}} H(\tilde{\mathbf{t}})] \\ &= J^\top \nabla_{\mathbf{t}}^2 H(\mathbf{t}(\hat{x}) - \tilde{\mathbf{t}}) = -h J^\top \boldsymbol{\xi}, \end{aligned}$$

where the middle equality is exact by (8) (affine $\nabla_{\mathbf{t}} H$) and the last uses hypothesis (ii) (inactive clip, so $\mathbf{t}(\hat{x}) - \tilde{\mathbf{t}} = -\boldsymbol{\xi}$). Taking the expectation over $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_m)$,

$$\mathbb{E} \|\nabla \mathcal{L}(\hat{x})\|^2 = h^2 \sigma^2 \text{tr}(J J^\top) = h^2 \sigma^2 \|J\|_F^2.$$

A fixed total (ε, δ) -DP budget caps the per-coordinate release variance at $\sigma^2 \geq \sigma_\star^2 = \Omega(c^2 \log(1/\delta)/\varepsilon^2)$ by the ℓ_2 release bound of Proposition 3. Budgets compose across rounds ($\sum_k \rho_k = \rho$), so inverse-variance time-averaging cannot beat this cap. Substituting yields

$$\mathbb{E} [\|\nabla \mathcal{L}(\hat{x})\|^2] \geq h^2 \sigma_\star^2 \|J\|_F^2, \quad h = d^2 \|\mathbf{q}\|^2,$$

which is (6). $\square$

c) *Why the floor bites.*: The true gradient $-h J^\top \boldsymbol{\xi}$ lies in $\text{im}(J^\top) = (\ker J)^\perp$, which is transverse to the block-term scaling and rotation invariances (these lie in $\ker J$). For the floor to vanish one would need $\mathbb{E} \|J^\top \boldsymbol{\xi}\|^2 = \sigma^2 \|J\|_F^2 = 0$, i.e. $J = 0$. The mechanism's noise is isotropic, so it always has a component in the positive-dimensional subspace $\text{im } J$; an isotropic mechanism cannot align its noise with the flat directions. The floor is therefore a genuine property of the budget, not an artifact of a particular stopping rule.

d) *The constant is generic.*: With the linear contraction $t_{r,i}^{(p)} = \sum_l a_{r,l}^\top X_i^{(p)} b_{r,l}$ and data second moments $\Sigma_{\text{row}} = \sum_i X_i X_i^\top$ and $\Sigma_{\text{col}} = \sum_i X_i^\top X_i$,

$$\begin{aligned} \|J^{(r)}\|_F^2 &= \text{tr}(A_r^\top \Sigma_{\text{row}} A_r) + \text{tr}(B_r^\top \Sigma_{\text{col}} B_r), \\ J &= \text{blockdiag}_r(J^{(r)}). \end{aligned}$$

The floor is built only from the data covariance, the factor norms, and the block gains; the inter-block spectral gap does

not appear, and the rank L enters only as the number of generic rank-one quadratic forms summed. This is the matching floor the manuscript claims, not a block-term-specific separation.

We give the full arguments behind Corollary 1, Lemma 3, and Proposition 3, all in risk space.

e) Profiled risk is a Rayleigh quotient.: With residual $\mathbf{F} \in \mathbb{R}^{m \times C}$ (cohort $\times$ labels) and score $\mathbf{t} \in \mathbb{R}^m$, profile out the active block $(\mathbf{q}, d)$, which enters only through $\mathbf{w} = d\mathbf{q}$:

$$R(\mathbf{t}) = \min_{\mathbf{w}} \|\mathbf{F} - \mathbf{t} \mathbf{w}^\top\|_F^2 = \|\mathbf{F}\|_F^2 - \frac{\mathbf{t}^\top \mathbf{M} \mathbf{t}}{\|\mathbf{t}\|^2}, \quad (9)$$

$$\mathbf{M} := \mathbf{F} \mathbf{F}^\top \in \mathbb{R}^{m \times m},$$

attained at $\mathbf{w} = \mathbf{F}^\top \mathbf{t} / \|\mathbf{t}\|^2$. R is scale-invariant, so the geometry lives on the unit sphere and the minimizer $\mathbf{t}^*$ is the top eigenvector of $\mathbf{M}$. The matrix $\mathbf{M} = \mathbf{F} \mathbf{F}^\top$ is $m \times m$ and never sees the ambient feature dimension $\prod_n I_n$, which is where dimension-freeness originates: the utility is a function of the cohort-sized score, not of the high-dimensional factors. On $\|\mathbf{t}\| = 1$, $R = \|\mathbf{F}\|_F^2 - \mathbf{t}^\top \mathbf{M} \mathbf{t}$ has tangent Hessian $2(\lambda_{\max} \mathbf{I} - \mathbf{M})$ at $\mathbf{t}^*$, with eigenvalues $2(\lambda_{\max} - \lambda_i) \in [0, 2\sigma_{\max}(\mathbf{F})^2]$. Hence

$$\beta := 2\sigma_{\max}(\mathbf{F})^2 = 2\lambda_{\max}(\mathbf{M}) \quad (\text{smoothness}),$$

$$\mu := 2(\lambda_{\max}(\mathbf{M}) - \lambda_2(\mathbf{M})) \quad (\text{tangent strong convexity}), \quad (10)$$

the latter strictly positive iff $\mathbf{M}$ has a spectral gap.

f) Proof of Corollary 1.: The minimum in (9) is attained as stated, and on the unit sphere $\nabla R(\mathbf{t}^*) = 0$ with R being β -smooth by (10). A second-order expansion in the tangent direction gives $R(\tilde{\mathbf{t}}) - R(\mathbf{t}^*) \leq \frac{\beta}{2} \|\mathbf{n}_\perp\|^2$, where $\tilde{\mathbf{t}} = \mathbf{t}^* + \mathbf{n}$ and $\mathbf{n}_\perp$ is the tangential component (the radial component is free by scale-invariance). The Gaussian mechanism at ℓ_2 -sensitivity $\leq 2c$ (replace-one-party adjacency) composed over RT releases, with $\varepsilon = \Theta(\sqrt{RT} \varepsilon_0 \sqrt{\log(1/\delta)})$, gives $\mathbb{E}\|\mathbf{n}\|^2 = m\sigma^2$ with $\sigma^2 = \Theta(c^2 RT \log(1/\delta)/\varepsilon^2)$. Therefore

$$\mathbb{E}[R(\tilde{\mathbf{t}}) - R(\mathbf{t}^*)] = O\left(\frac{\sigma_{\max}(\mathbf{F})^2 m c^2 RT \log(1/\delta)}{\varepsilon^2}\right),$$

independent of $\prod_n I_n$ since $\mathbf{M} = \mathbf{F} \mathbf{F}^\top$ involves only the $m \times C$ residual. $\square$

Dimension-freeness here is of the prediction *risk*. Factor recovery is a separate quantity: the factor-space stationarity of Theorem 2 carries the contraction Jacobian B , which for a data-adaptive subspace can scale with $\prod_n I_n$. The claim is dimension-free utility, not dimension-free recovery of the local factors.

g) Proof of Lemma 3.: One party's clipped contribution $\text{clip}_c(\mathbf{t}^{(p)}) \in \mathbb{R}^m$ is indexed by the shared samples and enters the aggregate $\mathbf{t} = \sum_p \text{clip}_c(\mathbf{t}^{(p)})$ additively; replacing that party's local tensor changes only this summand, by a vector of ℓ_2 -norm $\leq 2c$ (clip radius c , replace-one adjacency giving $\Delta = 2c$). For rank $L \geq 1$ and the sample (cohort) mode left uncontracted, the map from a party's local tensor to its sample-mode contribution is surjective onto a norm- $\leq c$ ball of $\mathbb{R}^m$, so any norm- $\leq 2c$ adjacency perturbation of the score is induced by valid adjacent instances. The realizability of an arbitrary norm- $\leq 2c$ perturbation supplies the hard instances for the per-coordinate-sensitivity lower bound (Proposition 3(b)); it does not by itself license the m -factor under the joint- ℓ_2 clip. $\square$

h) Proof of Proposition 3.: (i) *Differential privacy forces tangential score error.* (a) *Unconditional, joint ℓ_2 -clip.* Two adjacent residuals whose top eigenvectors differ by $\Theta(c)$ along a single tangent direction (realizable by Lemma 3 at joint ℓ_2 -sensitivity $2c$) are indistinguishable under (ε, δ) -DP below mean-squared error $\Omega(c^2/\varepsilon^2)$. This is a two-point/Le Cam argument; it carries no $\log(1/\delta)$ factor, since in one distinguishing direction the $\log(1/\delta)$ is a Gaussian upper-bound artifact, not a lower bound, and the $\Omega(c^2/\varepsilon^2)$ bound holds over all (ε, δ) -DP mechanisms, not only budget-splitting ones. (b) *Per-coordinate sensitivity.* If the release additionally has per-coordinate sensitivity c (the marginal/mean instantiation), the m coordinate directions can be packed simultaneously and the (ε, δ) -DP fingerprinting lower bound for ℓ_2 release [23], [24] sharpens the tangential error to $\Omega(mc^2 \log(1/\delta)/\varepsilon^2)$; Lemma 3 furnishes the hard instances for this regime. Under the joint- ℓ_2 clip the m -dependence between the upper and lower bounds need not be tight (a gap of up to m), so the matched-in- m statement is scoped to the per-coordinate regime and is not claimed for the joint clip. Projecting onto the $(m-1)$ -dimensional tangent space at $\mathbf{t}^*$ (the radial direction is risk-free) leaves $\mathbb{E}\|\tilde{\mathbf{t}} - \mathbf{t}^*\|_\perp^2$ at the stated order in each case. (ii) *Strong convexity converts score error to risk.* By (10), R is μ -strongly convex on the tangent space with $\mu = 2(\lambda_{\max} - \lambda_2)$ and $\nabla R(\mathbf{t}^*) = 0$, so $R(\tilde{\mathbf{t}}) - R(\mathbf{t}^*) \geq \frac{\mu}{2} \|\tilde{\mathbf{t}} - \mathbf{t}^*\|_\perp^2$. Substituting (i) gives, unconditionally, $\sup_{\mathbf{F}} \mathbb{E}[R(\tilde{\mathbf{t}}) - R(\mathbf{t}^*)] = \Omega(\mu c^2/\varepsilon^2)$, and, in the per-coordinate-sensitivity regime,

$$\sup_{\mathbf{F}} \mathbb{E}[R(\tilde{\mathbf{t}}) - R(\mathbf{t}^*)] = \Omega\left(\frac{\mu m c^2 \log(1/\delta)}{\varepsilon^2}\right),$$

which matches Corollary 1 up to the condition number β/μ . (Strong convexity supplies the lower-bound direction that prox non-expansiveness, an upper bound, cannot.) $\square$

i) Remarks.: Three points fix the scope of the matching bound. First, the lower bound is unconditional in (c, ε) ; the $\log(1/\delta)$ factor and the cohort factor m are matched only in the per-coordinate-sensitivity regime, since under the joint- ℓ_2 clip used by the mechanism the two-point bound carries no $\log(1/\delta)$ and the m -dependence between the upper and lower bounds need not be tight. Second, the lower bound holds for any (ε, δ) -DP mechanism and so lacks the RT factor of Corollary 1; that factor is the iteration overhead, forced only for mechanisms that re-release the score each round. Third, the residual gap to the upper bound is the condition number $\beta/\mu = \lambda_{\max}/(\lambda_{\max} - \lambda_2)$, which is $\Theta(1)$ precisely when the residual has a constant spectral gap. The gauge (scale) direction is handled exactly and is free in both bounds.