STABILIZED SPECTRAL PROXIMAL UPDATES FOR DISTRIBUTED LOW-RANK REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

In many applications, several responses are predicted from common predictors while the observations are stored on computers that cannot pool them. Estimating a low-rank coefficient matrix from such data remains a difficult task when local samples have fewer observations than predictors. Corrected local fitting problems can be unbounded, and nuclear-norm penalties shrink strong effects. In this article, we propose a stabilized spectral proximal method that exchanges model estimates and gradients. A majorizing metric computed from one scalar per worker, with the penalty on the matching transformed coefficient, makes every round an exact singular-value update: a nuclear-norm stage followed by a nonconvex refinement that leaves strong singular values unshrunk. Theoretical results show that the computed iterates satisfy finite-iteration prediction bounds, including designs with fewer combined observations than predictors, and that observed singular-value gaps certify the selected rank. Under signal and curvature conditions, the iterates recover the identifiable rank and converge to the least-squares fit of that rank. We also characterize when gradient exchange transmits fewer scalars than sufficient statistics. Simulations show exact rank recovery, accuracy close to centralized estimation with the same nonconvex penalty, and fewer transmitted scalars when predictors greatly outnumber responses.

1 INTRODUCTION

Predicting q responses from p common predictors can require many coefficients. The OES10 employment data contain 298 predictors and 16 responses for 403 observations (Tsoumakas et al., 2011; Spyromitros-Xioufis et al., 2016), requiring 4,768 unrestricted coefficients. Low-rank regression uses shared linear combinations of predictors (Anderson, 1951; Bunea et al., 2011); rank is the smallest number of combinations needed. When observations are distributed across computers, called workers, pooling them can be infeasible or prohibited.

Distributed methods average local estimates (Zhang et al., 2013) or combine pooled gradients with the curvature of one worker's data (Shamir et al., 2014; Jordan et al., 2019; Wang et al., 2017). When that worker has fewer observations than predictors, the gradient-corrected objective is linear along locally undetermined directions and can decrease without bound, even under a rank-one restriction. A proximal term restores a well-defined subproblem (Shamir et al., 2014; Hendrikx et al., 2020; Fan et al., 2023). With a spectral penalty and a general local Gram matrix, these subproblems do not reduce to ordinary spectral thresholding. These analyses concern convex penalties and do not establish spectral rank recovery. Gradient-based federated methods avoid local inverses (Yuan et al., 2021; Bao et al., 2022), but their convex nuclear-norm penalty shrinks strong singular values.

Nuclear-norm penalties reduce regression dimension (Yuan et al., 2007; Negahban & Wainwright, 2011) and estimate partially or indirectly observed matrices (Candès & Recht, 2009; Koltchinskii et al., 2011; Cai et al., 2010; Mazumder et al., 2010; Zhou & Li, 2014). Smoothly clipped absolute deviation (SCAD) and the minimax concave penalty (MCP) reduce shrinkage bias (Fan & Li, 2001; Zou & Li, 2008; Zhang, 2010; Loh & Wainwright, 2015; Fan et al., 2018). Low-rank extensions recover rank under sufficient signal strength (Gui et al., 2016; Park et al., 2024; 2026), and rank-constrained projected gradient methods converge under restricted curvature (Jain et al., 2010; 2014). Spectral optimization (Lu et al., 2015; Yao et al., 2019; Mai, 2025) and adaptive nuclear-norm

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.

regression (Chen et al., 2013) treat pooled data. Variable-metric proximal algorithms (Combettes & Vũ, 2014) do not generally permit ordinary spectral thresholding under an unmatched metric.

Distributed optimization methods also use consensus (Boyd et al., 2011) or network communication (Maros & Scutari, 2022), and recent federated studies treat general objectives and penalties (Zhang et al., 2026; Zhou et al., 2026) and gains from additional workers and local updates (Huang et al., 2026). For distributed low-rank regression, Chen et al. (2022a) assume common predictor covariances and use a full-rank master design, and Yang et al. (2023) penalize a surrogate loss by the nuclear norm for quantile regression with matrix predictors and a scalar response. Chen et al. (2022b) study quantile loss. Other methods distribute predictors (Huang & Tsay, 2024), use communication networks (Wai et al., 2017; Qiao & Chen, 2024), or reduce penalization bias (Choi & Park, 2025). To the best of our knowledge, no existing analysis establishes prediction bounds and rank recovery for the estimates computed by a distributed algorithm with a nonconvex spectral penalty when every local sample has fewer observations than predictors and predictor second-moment matrices differ across workers.

In this article, we propose a stabilized spectral proximal method for a common low-rank coefficient matrix, allowing unequal sample sizes and predictor distributions. A coordinating worker, called the master, adds to its local curvature a multiple of the identity computed from one scalar per worker. The resulting quadratic surrogate bounds the pooled loss from above. Placing the penalty on the matching transformed coefficient makes each proximal step an exact singular-value map. A convex stage initializes the estimate. Each SCAD refinement decreases its penalized objective and leaves strong singular values unshrunk. Each round exchanges one $p \times q$ matrix in each direction.

First, we bound the prediction error of the computed iterates after finitely many rounds, including designs in which the combined sample has fewer observations than predictors. Under independent Gaussian noise and a bounded curvature ratio on the observed predictor space, the first-stage statistical bound has the minimax rank and dimension dependence. For the second stage, a finite-iteration comparison separates selected-rank approximation error, noise, and the error of stopping, without assuming that the selected rank is correct.

Second, we establish descent, full-sequence convergence, and rank recovery for the computed updates without a full-rank local or pooled design. From the singular values it computes, the master can verify the rank of an update and whether its nonzero singular values are unshrunk. Under signal and curvature conditions of the beta-min form used for folded-concave spectral penalties on pooled data (Park et al., 2024; 2026), the updates keep the true rank and converge to the least-squares fit of that rank. Under singular pooled designs, they recover the rank identifiable from the data.

Third, we compare gradient exchange with uploading sufficient statistics, counting all tuning candidates. Gradient exchange transmits fewer scalars when the total number of rounds is below a threshold of order p/q, and with prescribed tuning under locally singular designs, prediction error and the communication ratio both decrease to zero. Exact gradients reveal the sufficient statistics after $\lceil p/q \rceil + 1$ queries, and this number is necessary over positive-semidefinite Gram matrices. Bit-complexity bounds (Acharya et al., 2019; Li et al., 2023) use a different measure. In the two primary simulations, the method recovers the rank in every replication, achieves accuracy close to centralized SCAD, and is more accurate than distributed nuclear-norm methods, one-shot averaging, and distributed reduced-rank regression. When predictors greatly outnumber responses, it also transmits fewer scalars than a sufficient-statistic upload. Under singular pooled designs, its second stage reduces the Stage I prediction error in every replication.

The remainder of the article is organized as follows. Section 2 defines the model, Section 3 the estimator, Section 4 the guarantees, Section 5 the simulations and two public-data analyses, and Section 6 the implications and limitations. Proofs and additional results are in the appendices.

2 Distributed multivariate regression

There are $m \geq 2$ workers. Worker $s \in \{1, \ldots, m\}$ stores $n_s \geq 1$ observations of p predictors in $X_s \in \mathbb{R}^{n_s \times p}$ and q responses in $Y_s \in \mathbb{R}^{n_s \times q}$. We study $Y_s = X_s B^\star + E_s$, with common low-rank coefficient matrix $B^\star \in \mathbb{R}^{p \times q}$ and error matrix E_s. Predictor distributions may differ across workers.

Let $\|\cdot\|_2$, $\|\cdot\|_F$, and $\|\cdot\|_{\mathrm{op}}$ denote the Euclidean, Frobenius, and operator norms. $\mathrm{tr}(A)$ and $\ker(A)$ denote trace and null space. The matrix inner product is $\langle A, B \rangle = \mathrm{tr}(A^\top B)$, and $\sigma_j(A)$ lists singular

values in decreasing order. We use I_p for the identity matrix, $\lambda_{\min}$ and $\lambda_{\max}$ for extremal eigenvalues, and $A \preceq B$ when $B - A$ is positive semidefinite. For $A \succeq 0$, $A^{1/2}$ is its positive-semidefinite square root. For scalars, $u \wedge v = \min\{u, v\}$ and $u_+ = \max\{u, 0\}$; $\lceil u \rceil$ is the smallest integer at least u. Positive numerical constants denoted by c or C may vary between occurrences; $\lesssim$, $\gtrsim$, and $O(\cdot)$ suppress numerical constant factors.

Let $N = \sum_s n_s$ and $w_s = n_s/N$. We define

$$\mathcal{L}(B) = \sum_{s=1}^{m} w_s \mathcal{L}_s(B), \qquad \mathcal{L}_s(B) = \frac{1}{2n_s} \|Y_s - X_s B\|_F^2, \tag{1}$$

where $B \in \mathbb{R}^{p \times q}$. We define $\widehat{\Sigma}_s = X_s^\top X_s/n_s$, $\widehat{Z}_s = X_s^\top Y_s/n_s$, $\widehat{\Sigma} = \sum_s w_s \widehat{\Sigma}_s$, and $\widehat{Z} = \sum_s w_s \widehat{Z}_s$. The local gradient is $G_s(B) = \widehat{\Sigma}_s B - \widehat{Z}_s$, and the pooled gradient is $G(B) = \sum_s w_s G_s(B) = \widehat{\Sigma} B - \widehat{Z}$. Let X and Y denote the matrices obtained by stacking the worker rows. For a positive-semidefinite matrix K, we define $\|A\|_K^2 = \mathrm{tr}(A^\top K A)$.

The pooled loss in (1) depends on the data through $\widehat{\Sigma}$ and $\widehat{Z}$. Uploading $\widehat{\Sigma}_s$ and $\widehat{Z}_s$ costs $S_{\mathrm{stat}} = p(p+1)/2 + pq$ scalars per remote worker and permits all subsequent fits without further communication. Sample sizes are registered before fitting and are not charged to either protocol. Our synchronous parameter-server method uses exact gradients and one setup scalar. Worker 1 is the master; it keeps its local observations and never forms a pooled $p \times p$ Gram matrix.

3 Stabilized spectral proximal estimation

We first obtain a convex initial estimate by nuclear-norm regularization, then use SCAD to reduce shrinkage of strong effects. Both steps use pooled gradients in the same stabilized coordinate.

3.1 Stabilized curvature

We stabilize the master curvature using one scalar from each worker. Each worker sends $\ell_s = \lambda_{\max}(\widehat{\Sigma}_s)$, and the master computes $\bar{\ell} = \sum_s w_s \ell_s$. If $\bar{\ell} = 0$, every X_s and $\widehat{Z}_s$ equals zero, and the zero matrix minimizes the pooled loss. For $\bar{\ell} > 0$, let $\tau > 0$ be the stabilization level. We define

$$H = \widehat{\Sigma}_1 + \gamma I_p, \qquad \gamma = (1+\tau)\bar{\ell}, \qquad C = H^{1/2} B. \tag{2}$$

Because $\widehat{\Sigma} \preceq \bar{\ell} I_p$, $H - \widehat{\Sigma} \succeq \tau \bar{\ell} I_p$. Thus H is positive definite even when $n_1 < p$. For an iterate B_t, we define the quadratic surrogate

$$Q_t(B) = \mathcal{L}(B_t) + \langle G(B_t), B - B_t \rangle + \frac{1}{2} \left\| H^{1/2}(B - B_t) \right\|_F^2 = \mathcal{L}(B) + \frac{1}{2} \|B - B_t\|_{H - \widehat{\Sigma}}^2 \tag{3}$$

Equation (3) is exact: the surrogate agrees with the pooled loss to first order at B_t and adds positive curvature in every predictor direction.

If two predictor columns coincide at the master but differ elsewhere, the local data leave their coefficient difference undetermined, and the gradient-corrected local objective can be unbounded below along such a direction. Proximal terms also stabilize local subproblems in Shamir et al. (2014) and Fan et al. (2023). Here the positive gap $H - \widehat{\Sigma}$ makes the surrogate an upper bound for the pooled loss. A general positive metric does not permit ordinary singular-value thresholding, and Appendix S1.2 gives both calculations. To obtain an exact spectral update, we place the penalty on $C = H^{1/2} B$. Appendix S28 gives exact updates, descent bounds, and one-step rank guarantees for the isotropic metric γI_p and other positive-definite metrics that strictly majorize the pooled curvature. The local term adapts the penalty scale to the master's predictor variation without changing the per-round communication.

3.2 Step 1: Convex initial estimation

We obtain the initial estimate from a convex problem in the transformed coordinate. Let $f(C) = \mathcal{L}(H^{-1/2} C)$ and $M = H^{-1/2} \widehat{\Sigma} H^{-1/2}$. Then $0 \preceq M \preceq (1+\tau)^{-1} I_p$. Let $\lambda_1, \lambda_2 \geq 0$ denote the

penalty levels, and let T_1, T_2 be nonnegative integers counting the rounds in Steps 1 and 2. The Stage I target is

$$\widehat{C} \in \arg\min_{C}\{F_1(C) := f(C) + \lambda_1 \|C\|_*\}, \tag{4}$$

where $\|C\|_*$ is the nuclear norm. At $\lambda_1 = 0$, $\widehat{C}$ denotes the least-squares minimizer of smallest Frobenius norm. The transformation in (2) specifies the surrogate metric and measures singular values in units of local predictive variation, because $\|C\|_F^2 = \|X_1 B\|_F^2 / n_1 + \gamma \|B\|_F^2$. Rank is unchanged because H is nonsingular.

The transformed gradient is $\nabla f(C) = H^{-1/2} G(H^{-1/2} C)$. In this coordinate, minimizing the surrogate plus the nuclear penalty gives $\mathrm{SVT}_{\lambda_1}\{C_t - \nabla f(C_t)\}$, where $C_t = H^{1/2} B_t$ and SVT_λ applies singular-value soft thresholding at level λ.

Stage I uses accelerated proximal-gradient updates with unit step (Beck & Teboulle, 2009). Let D_t be the Nesterov extrapolated point. The master broadcasts $H^{-1/2} D_t$, aggregates local gradients, and forms $Z_t = D_t - \nabla f(D_t)$. Algorithm 1 specifies the extrapolation coefficient α_t and update $C_{t+1} = \mathrm{SVT}_{\lambda_1}(Z_t)$. The initial estimate for Stage II is the computed iterate C_{T_1}, which need not equal $\widehat{C}$.

3.3 Step 2: SCAD refinement

We refine C_{T_1} with SCAD to reduce shrinkage of strong effects. In Stage II, $D_t = C_t$, and the master computes the same gradient update without extrapolation. We minimize the surrogate plus a folded-concave spectral penalty through the scalar proximal map applied to the singular values of Z_t. For the SCAD penalty with $a > 2$, this map is

$$\Theta_{\lambda,a}(z) = \begin{cases} 0, & 0 \le z \le \lambda, \\ z - \lambda, & \lambda < z \le 2\lambda, \\ \{(a-1)z - a\lambda\}/(a-2), & 2\lambda < z \le a\lambda, \\ z, & z > a\lambda. \end{cases} \tag{5}$$

The map sets singular values at most λ_2 to zero and leaves those above $a\lambda_2$ unchanged. Soft thresholding subtracts λ_1 from every kept singular value. Stage II thus reduces the shrinkage of strong effects. The minimax concave penalty map is given in Appendix S1.1.

In Algorithm 1, $Z_t = U_t \mathrm{diag}(d_t) V_t^\top$ is a thin singular value decomposition and d_t lists the singular values. The operator diag forms a diagonal matrix, and Θ acts entrywise. After T_1 Stage I rounds and T_2 Stage II rounds, the final estimator is $\widehat{B} = H^{-1/2} C_{T_1+T_2}$.

Algorithm 1 Stabilized spectral proximal estimation

1: Given $T_1, T_2, \lambda_1, \lambda_2, \tau$ and $a > 2$, set $C_0 = D_0 = 0$ and $\alpha_0 = 1$.
2: Collect ℓ_s and construct H from (2).
3: **for** $t = 0, \ldots, T_1 - 1$ **do**
4: Aggregate $G_s(H^{-1/2} D_t)$ with weights w_s and set $Z_t = D_t - H^{-1/2} G(H^{-1/2} D_t)$.
5: Set $C_{t+1} = \mathrm{SVT}_{\lambda_1}(Z_t)$, $\alpha_{t+1} = \{1 + (1 + 4\alpha_t^2)^{1/2}\}/2$.
6: Set $D_{t+1} = C_{t+1} + (\alpha_t - 1)(C_{t+1} - C_t)/\alpha_{t+1}$.
7: **end for**
8: Reset $D_{T_1} = C_{T_1}$.
9: **for** $t = T_1, \ldots, T_1 + T_2 - 1$ **do**
10: Aggregate $G_s(H^{-1/2} C_t)$ with weights w_s and set $Z_t = C_t - H^{-1/2} G(H^{-1/2} C_t)$.
11: If $Z_t = U_t \mathrm{diag}(d_t) V_t^\top$, set $C_{t+1} = U_t \mathrm{diag}\{\Theta_{\lambda_2,a}(d_t)\} V_t^\top$.
12: **end for**
13: Return $\widehat{B} = H^{-1/2} C_{T_1+T_2}$.

A thin singular value decomposition of X_1 applies $H^{1/2}$ and $H^{-1/2}$ with $O(n_1 p + pq)$ memory when $n_1 < p$. A truncated decomposition (Halko et al., 2011) remains exact when the omitted singular values are certified at or below the zero threshold of the penalty by Proposition 10. Appendices S17 and S33 treat certified spectral-radius bounds and metric error.

The multistage method of Bao et al. (2022) uses a convex norm penalty under restricted curvature. Our exact SCAD updates permit singular local Gram matrices.

4 THEORETICAL RESULTS

4.1 PREDICTION AND POPULATION GUARANTEES

We separate statistical prediction error from the error of stopping after finitely many updates.

Suppose $Y_s = X_s B^\star + E_s$, where E_s is the error matrix. Let $r = \operatorname{rank}(B^\star)$ and let the E_s be stacked into E. Define $C^\star = H^{1/2} B^\star$ and

$$\widetilde{W} = H^{-1/2} \sum_{s=1}^{m} w_s X_s^\top E_s / n_s, \qquad \varepsilon = \left\| \widetilde{W} \right\|_{\mathrm{op}}, \qquad \kappa = \lambda_{\min}(M).$$

Proposition 3 links the matched penalty to pooled prediction, with two-sided norm bounds when $\kappa > 0$. Descent and stationarity hold for every choice of master (Proposition 30).

Let $\widehat{B}_1 = H^{-1/2} \widehat{C}$. The following bound uses curvature only on the observed predictor space.

Let $d = \operatorname{rank}(X)$, let Π_X be the orthogonal projection onto the row space of X, and let κ_+ be the smallest positive eigenvalue of M. The standing condition $\bar{\ell} > 0$ gives $d \geq 1$ and $\kappa_+ > 0$. Let $C_X^\star = \Pi_X C^\star$ and $r_X = \operatorname{rank}(X B^\star)$. Lemma 3 proves that $\operatorname{rank}(C_X^\star) = r_X$ and that every iterate of Algorithm 1 remains in this predictor subspace.

Theorem 1. *If $\lambda_1 \geq \varepsilon$, then*

$$\left\| \widehat{C} - C_X^\star \right\|_F \leq \frac{2\sqrt{2 r_X}\, \lambda_1}{\kappa_+}, \qquad \frac{\left\| X(\widehat{B}_1 - B^\star) \right\|_F}{\sqrt{N}} \leq \frac{2\sqrt{2 r_X}\, \lambda_1}{\sqrt{\kappa_+}}.$$

For the zero initialization and every Stage I iterate with $1 \leq T \leq T_1$,

$$\left\| C_T - \widehat{C} \right\|_F \leq \frac{2 \left\| \widehat{C} \right\|_F}{\sqrt{\kappa_+}(T+1)}.$$

Theorem 1 controls prediction even when $N < p$. The identifiable coefficient target $\Pi_X B^\star$ gives the same fitted responses as $B^\star$, whereas rank-one coefficients with arbitrarily different components in $\ker(X)$ can be observationally equivalent by Proposition 35.

Theorem 5 extends this bound to approximately low-rank coefficients with coefficient one on the approximation error. For $1 \leq T \leq T_1$ and $\lambda_1 \geq \varepsilon$, Proposition 23 adds at most $2 \left\| C_0 - \widehat{C} \right\|_F / (T+1)$ to the statistical prediction bound, and Appendix S38.3 bounds prediction for new observations.

Let $x_{si}^\top$ and $e_{si}^\top$ denote rows of X_s and E_s. For a scalar random variable Z, the sub-Gaussian Orlicz norm is $\|Z\|_{\psi_2} = \inf\{b > 0 : \mathbb{E} \exp(Z^2/b^2) \leq 2\}$, with conditional expectation given the designs for noise; conditional bounds hold almost surely. Let $L, \mu, K_x, \sigma > 0$ and let $\Sigma_s^0 = \mathbb{E}(x_{si} x_{si}^\top)$. The following condition permits different predictor second moments across workers.

Condition 1. *The vectors $\{x_{si}\}_{s,i}$ are independent, with*

$$\lambda_{\max}(\Sigma_s^0) \leq L, \qquad \lambda_{\min}\Big(\sum_s w_s \Sigma_s^0 \Big) \geq \mu > 0, \qquad \left\| u^\top x_{si} \right\|_{\psi_2} \leq K_x (u^\top \Sigma_s^0 u)^{1/2}, \quad u \in \mathbb{R}^p.$$

Conditional on $\{X_s\}_s$, the vectors $\{e_{si}\}_{s,i}$ are independent, $\mathbb{E}(e_{si} \mid \{X_s\}_s) = 0$, and $\left\| v^\top e_{si} \right\|_{\psi_2} \leq \sigma \|v\|_2$ for $v \in \mathbb{R}^q$.

Condition 1 permits nonzero predictor means and unequal sample sizes. Let $n_{\min} = \min_s n_s$, $\Sigma^0 = \sum_s w_s \Sigma_s^0$, and $\chi = \max_s \left\| (\Sigma^0)^{-1/2} \Sigma_s^0 (\Sigma^0)^{-1/2} \right\|_{\mathrm{op}}$. This relative second-moment bound satisfies $1 \leq \chi \leq L/\mu$ and equals one for common matrices. Sub-Gaussian concentration (Vershynin, 2018) gives the following guarantees.

Corollary 1. *Under Condition 1, let* $0 < \delta < 1$. *If* $N \gtrsim K_x^4 \chi\{p + \log(1/\delta)\}$, *then, with probability at least* $1 - \delta$,

$$\lambda_{\min}(\widehat{\Sigma}) \geq \frac{\mu}{2}, \qquad \kappa \geq \frac{\mu/2}{\gamma + \ell_1}, \qquad \varepsilon \lesssim \sigma \sqrt{\frac{p + q + \log(1/\delta)}{(1+\tau)N}}.$$

The same event has $\max_s \ell_s \lesssim L K_x^2 \{1 + p/n_{\min} + \log(m/\delta)/n_{\min}\}$.

Corollary 1 controls pooled curvature without requiring any $n_s \geq p$. With λ_1 at the score scale, Theorem 6 gives prediction error of order $\sigma\{r(p + q + \log(1/\delta))/(\kappa(1+\tau)N)\}^{1/2}$, which has the rank and dimension dependence of low-rank regression bounds (Negahban & Wainwright, 2011; Chen et al., 2013), with $\kappa(1+\tau)$ as the curvature factor. Appendices S13 and S26 treat worker-specific noise and Gaussian effective dimension.

Under independent Gaussian noise, Proposition 5 gives the fixed-design minimax prediction lower bound $c\sigma^2 r(d + q - r)/N$ for $1 \leq r \leq d \wedge q$ (Bunea et al., 2011; Tsybakov, 2009). Conditional on the designs, Proposition 60 gives $\varepsilon \leq c\sigma\{\|M\|_{\mathrm{op}}(d + q + t)/N\}^{1/2}$ with probability at least $1 - 2e^{-t}$ for $t > 0$. With λ_1 at this scale, Theorem 1 matches the rank and dimension dependence of the lower bound when $\|M\|_{\mathrm{op}}/\kappa_+$ is bounded and $\log(2/\delta) \leq d + q$. The upper bound holds with high probability, whereas the lower bound concerns expected loss.

4.2 Convergence and rank recovery of computed updates

For $z \geq 0$, the scalar SCAD penalty is $P_\lambda(z) = \int_0^z [\lambda \wedge \{(a\lambda - u)_+/(a-1)\}]\, du$. Let $P_2(C) = \sum_j P_{\lambda_2}\{\sigma_j(C)\}$ and $F_2(C) = f(C) + P_2(C)$, with limiting subdifferential ∂F_2. Generalized singular-value thresholding gives the exact matrix proximal map (Lu et al., 2015). The following theorem establishes descent of the computed Stage II updates, and its convergence statements concern an infinite continuation of the same update.

Theorem 2. *For every Stage II iteration,*

$$F_2(C_t) - F_2(C_{t+1}) \geq \frac{\tau}{2(1+\tau)} \|C_{t+1} - C_t\|_F^2 .$$

Hence $\sum_t \|C_{t+1} - C_t\|_F^2 < \infty$, *and every cluster point of an infinite Stage II sequence is a limiting stationary point of* F_2.

For a set of matrices $\mathcal{S}$, define $\mathrm{dist}(0, \mathcal{S}) = \inf_{U \in \mathcal{S}} \|U\|_F$.

Corollary 2. *For every integer* $K \geq 1$,

$$\min_{T_1 \leq t < T_1 + K} \mathrm{dist}^2\{0, \partial F_2(C_{t+1})\} \leq \frac{2(1+\tau)}{\tau K}\{F_2(C_{T_1}) - \inf_C F_2(C)\}.$$

The complete exact Stage II sequence converges to a stationary point of F_2.

Corollary 2 bounds the stationarity error at order $K^{-1/2}$. Boundedness of the exact updates and semi-algebraic convergence give the full-sequence result without pooled invertibility (Attouch et al., 2013).

Stationarity alone does not control prediction error. We next compare the computed estimate with the best fit of the same rank.

Let $t \geq T_1$ and $1 \leq k \leq d \wedge q$. Let the singular-value sequence be extended by zero beyond $p \wedge q$. For $\sigma_k(Z_t) > 0$, define $\vartheta_t = \sigma_{k+1}(Z_t)/\sigma_k(Z_t)$.

Theorem 3. *Suppose* $\sigma_k(Z_t) > a\lambda_2$, $\sigma_{k+1}(Z_t) \leq \lambda_2$, *and* $\vartheta_t < \kappa_+$. *Then, for every* $B^\star \in \mathbb{R}^{p \times q}$,

$$\frac{\left\|X(H^{-1/2}C_{t+1} - B^\star)\right\|_F^2}{N} \leq \inf_{\mathrm{rank}(XB) \leq k} \frac{\|X(B - B^\star)\|_F^2}{N} + \frac{\{\sqrt{2k}\varepsilon + \|C_{t+1} - C_t\|_F\}^2}{\kappa_+ - \vartheta_t}.$$

Theorem 3 separates approximation, noise, and finite-iteration error without assuming the true rank. Its numerical evaluation needs a lower bound on κ_+. Theorem 15 certifies all later ranks and convergence to the least-squares fit of the selected rank, and Corollary 35 gives population conditions for this limit.

For the rank-recovery results that follow, suppose $1 \le r < p \wedge q$. Let

$$R_t = (I_p - M)(C_t - C^\star) + \widetilde{W}, \qquad \eta_t = \|R_t\|_{\mathrm{op}}.$$

The update matrix satisfies $Z_t = C^\star + R_t$, and the following theorem applies singular-value perturbation bounds (Stewart & Sun, 1990; Bhatia, 1997) to convert this identity into a rank statement.

Theorem 4. *Suppose $\eta_t < \lambda_2$ and $\sigma_r(C^\star) > a\lambda_2 + \eta_t$. Then the computed SCAD update has*

$$\mathrm{rank}(C_{t+1}) = r, \qquad \|C_{t+1} - C^\star\|_F \le 2\sqrt{2r}\eta_t.$$

Theorem 4 proves rank recovery for the computed update. The observable inequalities $\sigma_r(Z_t) > a\lambda_2$ and $\sigma_{r+1}(Z_t) \le \lambda_2$ certify that its first r singular values are kept and all others equal zero. Taking $t = T_1 + T_2 - 1$ applies the result to the returned estimator, and Corollary 4 gives the MCP counterpart.

Corollary 3 links the convex initialization to the first rank-correct update. Under its additional radius and separation conditions, Theorem 7 keeps the rank and unshrunk singular values at every subsequent update, with error bounded by a geometrically decreasing term and a statistical term of order $\sqrt{r}\varepsilon/\kappa$.

Under Condition 1 and the sample-size bound of Corollary 1, Corollaries 11 and 12 give sufficient signal, tuning, and Stage I duration conditions for these rank guarantees from $C_0 = 0$ with probability at least $1 - \delta$. Under singular pooled designs, Proposition 61 and Corollary 34 give population conditions for recovering the identifiable rank throughout Stage II.

4.3 Communication and information from gradients

Gradient exchange reduces the transmitted scalar count only before a dimension-dependent number of rounds. Let $T = T_1 + T_2$.

Proposition 1. *Counting one setup scalar, a $p \times q$ broadcast, and a $p \times q$ return per round, Algorithm 1 transmits*

$$S_{\mathrm{grad}} = 1 + 2Tpq$$

scalars per remote worker. It uses fewer scalars than sufficient-statistic aggregation if and only if

$$T < \frac{1}{2} + \frac{p+1}{4q} - \frac{1}{2pq}.$$

Because sufficient statistics can be reused across tuning candidates, Proposition 33 applies the same comparison to the total number of rounds over all candidates.

Corollary 36 gives singular local designs where prediction error and the one-fit communication ratio both decrease to zero under prescribed tuning.

A zero query and $\lceil p/q \rceil$ further queries recover worker sufficient statistics (Proposition 7). This count is necessary over all positive-semidefinite Gram matrices (Theorem 22), although not necessarily for a prescribed deficient worker rank.

5 Numerical studies

We compare the proposed SCAD estimator with Stage I, centralized nuclear and SCAD estimators, one-shot averaging, consensus alternating direction method of multipliers (ADMM), distributed reduced-rank regression (D-RRR), and Federated Dual Averaging (FedDualAvg) (Chen et al., 2022a; Yuan et al., 2021). The centralized estimators fit one-stage penalties to the pooled data. We report test response mean squared error (MSE) and, in simulations, relative Frobenius error $\left\|\widehat{B} - B^\star\right\|_F / \|B^\star\|_F$. Simulations use 100 independent replications. In the two primary simulations, the proposed tuning fractions and round counts are fixed before the replications, and comparator

penalties use independent validation observations. Communication totals cover one fit and exclude tuning for every method. Appendix S40 gives all designs, tuning grids, stopping rules, and paired summaries.

5.1 Locally singular designs and communication accounting

We first compare prediction, rank recovery, and communication in two synthetic settings. The proposed estimator recovers rank in every replication and achieves accuracy close to centralized SCAD in Table 1. In the first setting, worker covariances differ in correlation and scale, every local Gram matrix is singular, and $N = 480 > p$. The second uses a larger p/q to study communication savings. Stage I and Stage II use 30 and 20 rounds in the first setting and 15 and 5 in the second.

Table 1: Synthetic results over 100 replications. Numbers in parentheses represent standard deviations; percentages in brackets are exact-rank frequencies.

Method	Test MSE	Relative error	Rank [%]	Rounds	Scalars
Locally singular designs: $m = 8$, $n_s = 60$, $p = 120$, $q = 12$, $r = 3$					
Proposed	1.0890 (.0093)	.1278 (.0049)	3.00 [100]	50.0	144,001
Stage I	1.1458 (.0109)	.1624 (.0055)	11.94 [0]	30.0	86,401
FedDualAvg	1.1509 (.0118)	.1659 (.0062)	11.94 [0]	50.0	144,000
Central nuclear	1.1494 (.0115)	.1650 (.0061)	11.96 [0]	0	8,700
Central SCAD	1.0884 (.0103)	.1273 (.0053)	3.13 [92]	0	8,700
One-shot	2.8591 (.0863)	.5768 (.0132)	12.00 [0]	1.0	1,440
ADMM	1.1508 (.0116)	.1659 (.0062)	11.96 [0]	129.8	373,709
Communication setting: $m = 5$, $n_s = 1200$, $p = 300$, $q = 3$, $r = 2$					
Proposed	1.0384 (.0135)	.0675 (.0020)	2.00 [100]	20.0	36,001
Stage I	1.0489 (.0137)	.0768 (.0020)	3.00 [0]	15.0	27,001
FedDualAvg	1.0446 (.0138)	.0731 (.0024)	2.17 [83]	20.0	36,000
Central nuclear	1.0483 (.0137)	.0765 (.0021)	3.00 [0]	0	46,050
Central SCAD	1.0386 (.0136)	.0677 (.0020)	2.15 [85]	0	46,050
One-shot	1.0665 (.0142)	.0908 (.0023)	3.00 [0]	1.0	900
ADMM	1.0485 (.0137)	.0767 (.0021)	3.00 [0]	58.1	104,616
D-RRR	1.2228 (.0270)	.1682 (.0086)	2.00 [100]	.3	612

In the locally singular setting, the proposed method has lower prediction and coefficient errors than Stage I, FedDualAvg, one-shot averaging, ADMM, and centralized nuclear regularization in all 100 replications. D-RRR is undefined in this setting because its master least-squares step requires a full-rank local design. In the communication setting, the proposed method and FedDualAvg use 20 rounds and respectively transmit 36,001 and 36,000 scalars, which are strictly below the sufficient-statistic count of 46,050. The paired proposed-minus-FedDualAvg test MSE is -0.00615 with standard error 0.00023, and the proposed method has lower error in all 100 replications.

Across both designs, every Stage II update has the true rank in all 100 replications. Every returned estimate satisfies the finite-iteration fitting conditions. We compare its loss with the best fit of its selected rank in Table 5. Appendix S40.2 gives the complete rank diagnostics.

5.2 Prediction under singular pooled designs

We next study prediction when the pooled design is singular. Four settings have $N = 80 < p = 120$, and three separate the full coefficient rank four from the identifiable rank three. In all settings, the method selects the identifiable rank in every replication, and its second stage reduces the Stage I prediction error in every replication. Its mean prediction error is also below that of the one-stage centralized nuclear and SCAD fits, as reported in Table 10.

With a singular population covariance in the second setting, Stage II reduces mean prediction loss from 2.9875 to 1.8173 and test MSE from 1.4319 to 1.2996. Appendix S40.8 gives comparisons under smaller positive eigenvalues and weaker signals.

5.3 Public data and additional diagnostics

We finally compare prediction on two public multi-target datasets, SCM20d and OES10.

SCM20d from the Mulan collection (Tsoumakas et al., 2011; Spyromitros-Xioufis et al., 2016) has 61 predictors and 16 responses. We combine its published partitions and use 100 splits with 8,000 training and 966 test observations. Training data are divided among ten workers, each reserving ten percent for tuning. All methods share splits and partitions. D-RRR follows equation (11) of Chen et al. (2022a), with adaptive exponent 10.

Table 2: SCM20d results over 100 paired splits. Numbers in parentheses represent standard deviations.

Method	Test MSE	Rank	Rounds
Proposed	.41340 (.01786)	8.72 (.62)	300.0
Stage I	.41503 (.01809)	10.11 (.31)	100.0
Central nuclear	.41661 (.01820)	9.84 (.37)	0
Central SCAD	.41369 (.01788)	9.99 (.10)	0
One-shot	.41632 (.01822)	16.00 (.00)	1.0
D-RRR	.42193 (.02569)	9.26 (2.22)	1.7

The proposed estimator has the lowest mean error in Table 2. Against centralized SCAD, the mean paired difference is -0.000292 with standard error 0.000033, and the proposed error is lower in 81 splits. It is lower in at least 97 splits against each other method, as reported in Supplementary Table 6.

We use OES10 to study prediction with locally small samples: it has 403 observations, 298 predictors, and 16 responses (Tsoumakas et al., 2011; Spyromitros-Xioufis et al., 2016). Each worker holds 25 training observations, and hence no local design has full column rank and D-RRR is undefined. Across 100 splits with training-only preprocessing and tuning, mean test error is .18062 for the proposed method, versus .18330, .18308, and .20111 for centralized nuclear regularization, centralized SCAD, and one-shot averaging. The paired win frequencies are 69%, 73%, and 87%, respectively, in Table 8.

In the evaluated setting, all fifteen choices of metric, master, and stabilization level in Table 7 select rank three, and their mean test errors differ by less than 0.11%. In a single-host socket experiment with $p = 1500$ and $q = 2$ (Table 9), gradient exchange is faster on average and sends fewer scalars than sufficient-statistic aggregation, with relative coefficient difference 7.1×10^{-15}.

6 Discussion

Stabilization makes distributed low-rank fitting with a nonconvex spectral penalty a sequence of closed-form updates that need no full-rank local design. SCAD refinement from the convex stage removes the shrinkage of strong singular values, and the master certifies the selected rank from singular values it computes. Under Theorem 15, the limit is the reduced-rank least-squares fit of the selected rank.

Singular pooled designs identify only a coefficient component, and our guarantees concern that component. Gradient exchange transmits fewer scalars than sufficient-statistic aggregation when the total number of rounds, including tuning candidates, satisfies Proposition 33. When this condition fails, the same estimator can be computed from uploaded sufficient statistics. Exact gradients reveal sufficient statistics after enough queries (Proposition 7), and hence the method gives no formal privacy guarantee.

The analysis concerns synchronous row-partitioned quadratic regression without coefficient sparsity and excludes peer-to-peer consensus, delays, and online updates. Extending Appendix S5 to asynchronous protocols requires controlling gradient errors from delays while preserving descent and rank recovery.

AI USE STATEMENT

In this work, we used generative AI tools (large language model assistants) to run and check numerical summaries. We also used these tools for literature search and editing the manuscript for readability. We did not use generative AI tools to generate data sets; the synthetic data are produced by the released code, and the public data are the published Mulan files. The authors reviewed all AI-assisted work: the code was tested with the unit tests and verification scripts in the supplementary code, and every reported number was recomputed from the raw outputs. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Appendix S40 specifies every simulation and data-analysis setting, including tuning, preprocessing, and evaluation measures. The appendices give complete proofs. The supplementary code records the seeds and contains the estimator, unit tests for the metric and proximal maps, experiment drivers, session records, and machine-readable summaries for every reported result, including the Stage II path diagnostics.

ETHICS STATEMENT

The methodological and synthetic studies involve no human participants. The public SCM20d and OES10 data are used under their existing public distribution for regression benchmarking. The study introduces no personal data, intervention, or deployment decision.

REFERENCES

Jayadev Acharya, Chris De Sa, Dylan Foster, and Karthik Sridharan. Distributed learning with sublinear communication. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 40–50, 2019.

Theodore W. Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, 22(3):327–351, 1951. doi: 10.1214/aoms/1177729580.

Hedy Attouch, Jérôme Bolte, and Benar F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Mathematical Programming*, 137(1–2):91–129, 2013. doi: 10.1007/s10107-011-0484-9.

Yajie Bao, Michael Crawshaw, Shan Luo, and Mingrui Liu. Fast composite optimization and statistical recovery in federated learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1508–1536. PMLR, 2022.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. doi: 10.1137/080716542.

Rajendra Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer, New York, 1997. doi: 10.1007/978-1-4612-0653-8.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. doi: 10.1561/2200000016.

Florentina Bunea, Yiyuan She, and Marten H Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2):1282–1309, 2011.

Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010. doi: 10.1137/080738970.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009. doi: 10.1007/s10208-009-9045-5.

Canyi Chen, Wangli Xu, and Liping Zhu. Distributed estimation in heterogeneous reduced rank regression: With application to order determination in sufficient dimension reduction. *Journal of Multivariate Analysis*, 190:104991, 2022a. doi: 10.1016/j.jmva.2022.104991.

Kun Chen, Hongbo Dong, and Kung-Sik Chan. Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100(4):901–920, 2013. doi: 10.1093/biomet/ast036.

Xi Chen, Weidong Liu, and Xiaojun Mao. Robust reduced rank regression in a distributed setting. *Science China Mathematics*, 65:1707–1730, 2022b. doi: 10.1007/s11425-020-1785-0.

Hongshin Choi and Seyoung Park. Distributed reduced-rank regression for large-scale data. In *Proceedings of The Korean Data Analysis Society*, pp. 89, November 2025.

Patrick L. Combettes and B. C. Vũ. Variable metric forward–backward splitting with applications to monotone inclusions in duality. *Optimization*, 63(9):1289–1318, 2014. doi: 10.1080/02331934.2012.733883.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. doi: 10.1198/016214501753382273.

Jianqing Fan, Han Liu, Qiang Sun, and Tong Zhang. I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *The Annals of Statistics*, 46(2):814–841, 2018.

Jianqing Fan, Yongyi Guo, and Kaizheng Wang. Communication-efficient accurate statistical estimation. *Journal of the American Statistical Association*, 118(542):1000–1010, 2023. doi: 10.1080/01621459.2021.1969238.

Huan Gui, Jiawei Han, and Quanquan Gu. Towards faster rates and oracle property for low-rank matrix estimation. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2300–2309. PMLR, 2016.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011. doi: 10.1137/090771806.

Hadrien Hendrikx, Lin Xiao, Sébastien Bubeck, Francis Bach, and Laurent Massoulié. Statistically preconditioned accelerated gradient method for distributed optimization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4203–4227. PMLR, 2020.

Kun Huang, Shi Pu, and Karl Henrik Johansson. Achieving linear speedup for composite federated learning. arXiv preprint arXiv:2602.03357, 2026. Version 2.

Shuo-Chieh Huang and Ruey S. Tsay. Scalable high-dimensional multivariate linear regression for feature-distributed data. *Journal of Machine Learning Research*, 25(205):1–59, 2024.

Prateek Jain, Raghu Meka, and Inderjit S. Dhillon. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, volume 23, pp. 937–945, 2010.

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional M-estimation. In *Advances in Neural Information Processing Systems*, volume 27, pp. 685–693, 2014.

Michael I. Jordan, Jason D. Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019. doi: 10.1080/01621459.2018.1429274.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

Yi Li, Honghao Lin, and David Woodruff. ℓ_p-regression in the arbitrary partition model of communication. In *Proceedings of the Thirty-Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 4902–4928, 2023.

Po-Ling Loh and Martin J. Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16:559–616, 2015.

Canyi Lu, Changbo Zhu, Chunyan Xu, Shuicheng Yan, and Zhouchen Lin. Generalized singular value thresholding. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 1805–1811, 2015. doi: 10.1609/aaai.v29i1.9464.

The Tien Mai. Robust reduced rank regression under heavy-tailed noise and missing data via non-convex penalization. *arXiv preprint arXiv:2512.24450*, 2025. doi: 10.48550/arXiv.2512.24450.

Marie Maros and Gesualdo Scutari. Acceleration in distributed sparse regression. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36832–36844, 2022. doi: 10.52202/068431-2669.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010.

Sahand N. Negahban and Martin J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011. doi: 10.1214/10-AOS850.

Seyoung Park, Eun Ryung Lee, and Hongyu Zhao. Low-rank regression models for multiple binary responses and their applications to cancer cell-line encyclopedia data. *Journal of the American Statistical Association*, 119(545):202–216, 2024. doi: 10.1080/01621459.2022.2105704.

Seyoung Park, Eun Ryung Lee, Hyunjin Kim, and Hongyu Zhao. Transfer learning under large-scale low-rank regression models. *Journal of the American Statistical Association*, 121(554):1154–1166, 2026. doi: 10.1080/01621459.2025.2555057.

Nan Qiao and Canyi Chen. Fast and robust low-rank learning over networks: A decentralized matrix quantile regression approach. *Journal of Computational and Graphical Statistics*, 33(4): 1214–1223, 2024. doi: 10.1080/10618600.2024.2353640.

Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate Newton-type method. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1000–1008. PMLR, 2014.

Eleftherios Spyromitros-Xioufis, Grigorios Tsoumakas, William Groves, and Ioannis Vlahavas. Multi-target regression via input space expansion: Treating targets as inputs. *Machine Learning*, 104(1):55–98, 2016. doi: 10.1007/s10994-016-5546-z.

G. W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Computer Science and Scientific Computing. Academic Press, Boston, 1990. ISBN 0126702306.

Grigorios Tsoumakas, Eleftherios Spyromitros-Xioufis, Jozef Vilcek, and Ioannis Vlahavas. MULAN: A Java library for multi-label learning. *Journal of Machine Learning Research*, 12(71):2411–2414, 2011.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009. doi: 10.1007/b13794.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, 2018. doi: 10.1017/9781108231596.

Hoi-To Wai, Anna Scaglione, Jean Lafond, and Eric Moulines. Fast and privacy preserving distributed low-rank regression. In *2017 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 4451–4455. IEEE, 2017. doi: 10.1109/ICASSP.2017.7952998.

Jialei Wang, Mladen Kolar, Nathan Srebro, and Tong Zhang. Efficient distributed learning with sparsity. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3636–3645. PMLR, 2017.

Yaohong Yang, Lei Wang, Jiamin Liu, Rui Li, and Heng Lian. Communication-efficient estimation of quantile matrix regression for massive datasets. *Computational Statistics & Data Analysis*, 187: 107812, 2023. doi: 10.1016/j.csda.2023.107812.

Quanming Yao, James T. Kwok, Taifeng Wang, and Tie-Yan Liu. Large-scale low-rank matrix learning with nonconvex regularizers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2628–2643, 2019. doi: 10.1109/TPAMI.2018.2858249.

Honglin Yuan, Manzil Zaheer, and Sashank Reddi. Federated composite optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12253–12266. PMLR, 2021.

Ming Yuan, Ali Ekici, Zhaosong Lu, and Renato Monteiro. Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):329–346, 2007. doi: 10.1111/j.1467-9868.2007.00591.x.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010. doi: 10.1214/09-AOS729.

Jiaojiao Zhang, Jiang Hu, and Mikael Johansson. Non-convex composite federated learning with heterogeneous data. *Automatica*, 183:112695, 2026. doi: 10.1016/j.automatica.2025.112695.

Yuchen Zhang, John C Duchi, and Martin J Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14:3321–3363, 2013.

Hua Zhou and Lexin Li. Regularized matrix regression. *Journal of the Royal Statistical Society: Series B*, 76(2):463–483, 2014. doi: 10.1111/rssb.12031.

Yuan Zhou, Jiachen Zhong, Xinli Shi, Guanghui Wen, and Xinghuo Yu. FedCanon: Non-convex composite federated learning with efficient proximal operation on heterogeneous data. *IEEE Transactions on Signal Processing*, 74:215–229, 2026. doi: 10.1109/TSP.2025.3642025.

Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1509–1533, 2008. doi: 10.1214/009053607000000802.

S1 PROOFS

The metric identities give a common scale for prediction error and the computed spectral updates. We use them to prove the Stage I bounds, Stage II convergence, and rank guarantees. The communication results follow from the exact-gradient protocol.

S1.1 METRIC IDENTITIES AND SPECTRAL PROXIMAL MAPS

The stabilized quadratic surrogate is strictly convex even when the local design is singular. In the transformed coordinate, its penalized update reduces to scalar minimizations of the singular values.

Lemma 1. *The matrices in* (2) *satisfy*

$$H - \widehat{\Sigma} \succeq \tau\bar{\ell} I_p, \qquad 0 \preceq M \preceq (1+\tau)^{-1} I_p.$$

Equation (3) *holds exactly.*

Lemma 1 proves that the stabilization controls all predictor directions, including the null space of the master design.

Proof. The proof establishes the curvature bounds by aggregating the local spectral radii, and it verifies the surrogate identity by expanding the quadratic loss.

Step 1: Bounding the metric matrices. For every s, $\widehat{\Sigma}_s \preceq \ell_s I_p$, which gives $\widehat{\Sigma} \preceq \bar{\ell} I_p$. Hence

$$H - \widehat{\Sigma} = \widehat{\Sigma}_1 + (1+\tau)\bar{\ell} I_p - \widehat{\Sigma} \succeq \tau\bar{\ell} I_p.$$

Because $H \succeq (1+\tau)\bar{\ell} I_p$,

$$M = H^{-1/2}\widehat{\Sigma}H^{-1/2} \preceq \bar{\ell} H^{-1} \preceq (1+\tau)^{-1} I_p.$$

Step 2: Verifying the surrogate identity. Let $\Delta = B - B_t$. Expansion of the quadratic loss gives

$$\mathcal{L}(B) = \mathcal{L}(B_t) + \langle G(B_t), \Delta \rangle + \tfrac{1}{2}\left\langle \widehat{\Sigma}\Delta, \Delta \right\rangle.$$

Substitution into Q_t proves (3). This completes the proof. □

The spectral proximal map reduces the update to scalar minimizations over the singular values. Each round can thus be computed without an inner optimization problem. For $\lambda \geq 0$ and $b > 1$, the scalar MCP penalty is $P_\lambda(z) = \int_0^z (\lambda - u/b)_+ \, du$ for $z \geq 0$. We use θ_λ for the scalar proximal map, denoted by $\theta_{\lambda,b}$ for MCP.

Lemma 2. *Let P_λ be SCAD with $a > 2$ or MCP with $b > 1$. If $Z = U\mathrm{diag}(d)V^\top$, then*

$$U\mathrm{diag}\{\theta_\lambda(d)\}V^\top \in \arg\min_C \left\{ \tfrac{1}{2}\|C - Z\|_F^2 + \sum_j P_\lambda\{\sigma_j(C)\} \right\},$$

where $\theta_\lambda = \Theta_{\lambda,a}$ for SCAD. For MCP,

$$\theta_{\lambda,b}(z) = \begin{cases} 0, & 0 \leq z \leq \lambda, \\ b(z-\lambda)/(b-1), & \lambda < z \leq b\lambda, \\ z, & z > b\lambda. \end{cases}$$

Lemma 2 establishes that Algorithm 1 computes a global proximal minimizer at each round.

Proof. The main difficulty is minimizing the penalized quadratic objective over matrices, which we resolve by decoupling the problem into scalar minimizations using von Neumann's trace inequality.

Step 1: Lower bound via trace inequality. Let $C = \widetilde{U}\mathrm{diag}(c)\widetilde{V}^\top$ with $c_j \geq 0$. Von Neumann's trace inequality gives

$$\|C - Z\|_F^2 \geq \sum_j (c_j - d_j)^2,$$

with equality when the singular vectors are aligned.

Step 2: Scalar minimizations. For $d \geq 0$, the scalar candidate solves

$$\min_{c \geq 0} \ \tfrac{1}{2}(c-d)^2 + P_\lambda(c).$$

Lemmas 16 and 17 give the unique scalar minimizers, including (5). Both maps are nondecreasing, hence the minimizing singular values keep their order. Aligned singular vectors attain the matrix lower bound. This completes the proof. □

S1.2 Local curvature and spectral updates

A singular local Gram matrix causes two distinct difficulties. A rank constraint alone need not bound the corrected local surrogate. A positive metric does not make ordinary singular-value thresholding valid for an unmatched penalty.

For the first point, let $v \in \ker(X_1)$ satisfy $v^\top \widehat{\Sigma} v > 0$, take a unit vector $u \in \mathbb{R}^q$, and let $Y = Xvu^\top$. At the zero iterate, the corrected local quadratic with curvature $\widehat{\Sigma}_1$ is unbounded along the rank-one matrices $B = tvu^\top$, because

$$\left\langle G(0), tvu^\top \right\rangle + \tfrac{1}{2}\left\| \widehat{\Sigma}_1^{1/2} tvu^\top \right\|_F^2 = -t\, v^\top \widehat{\Sigma} v \longrightarrow -\infty.$$

Such a direction exists whenever the master misses a predictor direction represented by another worker. A bounded SCAD or MCP penalty cannot prevent this divergence. The positive quadratic term in (3) does prevent it.

For the same design and $B_0 = 0$, the corrected local quadratic and the stabilized surrogate satisfy

$$\left\langle G(0), tvu^\top \right\rangle + \tfrac{1}{2}\left\| \widehat{\Sigma}_1^{1/2} tvu^\top \right\|_F^2 = -t\, v^\top \widehat{\Sigma} v,$$
$$\left\langle G(0), tvu^\top \right\rangle + \tfrac{1}{2}\left\| H^{1/2} tvu^\top \right\|_F^2 = -t\, v^\top \widehat{\Sigma} v + \tfrac{1}{2}\gamma t^2 \left\| v \right\|_2^2 .$$

The stabilized expression is minimized at $t = v^\top \widehat{\Sigma} v / (\gamma \left\| v \right\|_2^2)$. Because $\ker(X) \subseteq \ker(X_1)$, the direction exists exactly when the pooled design has larger rank. This occurs in Section 5 and OES10, where each worker holds at most 25 observations on 298 predictors.

For the second point, it suffices to consider one response and two predictors. Let $K = \mathrm{diag}(k_1, k_2)$ with $k_1, k_2 > 0$ and $k_1 \neq k_2$, let z have two nonzero coordinates, and let $0 < \lambda < \left\| Kz \right\|_2$. The unique minimizer of $\frac{1}{2}(c-z)^\top K(c-z) + \lambda \left\| c \right\|_2$ is nonzero. Ordinary singular-value thresholding would keep its direction parallel to z. A minimizer of the form $c = \alpha z$ with $\alpha > 0$ would satisfy

$$(\alpha - 1)Kz + \lambda z / \left\| z \right\|_2 = 0,$$

which is impossible because Kz and z are not parallel. Thus even the one-column nuclear-norm proximal problem is not solved by ordinary spectral shrinkage under a general predictor metric. Our penalty is placed on $C = H^{1/2}B$ itself. The transformed proximal problem has the Euclidean geometry required by Lemma 2.

S1.3 PROOF OF THE PREDICTION BOUND ON THE OBSERVED PREDICTOR SPACE

A singular pooled design does not identify every coefficient direction. We show that the zero-initialized algorithm remains on the observed predictor space and prove Theorem 1 there. Appendix S38 gives the conditional score bound, population sufficient conditions, and rank consequences.

For a matrix A, let $\mathrm{range}(A)$ denote its column space. The projection Π_X acts on predictor coordinates. It is distinct from the projection P_X onto the column space of X used in Appendix S34. All assertions below keep the metric and exact updates in Algorithm 1.

Lemma 3. *For $\lambda_1 > 0$, the Stage I target is unique. With the specified minimum-norm choice at $\lambda_1 = 0$,*

$$\Pi_X H = H\Pi_X, \qquad \mathrm{range}(M) = \mathrm{range}(X^\top), \qquad \Pi_X \widehat{C} = \widehat{C},$$
$$\Pi_X C_t = C_t, \qquad \Pi_X D_t = D_t, \qquad H^{-1/2} C_X^\star = \Pi_X B^\star, \qquad \mathrm{rank}(C_X^\star) = r_X.$$

Lemma 3 identifies the target of the computed path without imposing a sparsity condition or modifying the algorithm. Uniqueness for positive nuclear regularization holds even when the pooled Gram matrix is singular.

Proof. The proof establishes null-space preservation under the stabilized metric and tracks the support of the exact iterates.

Step 1: Commutation and null space. For $v \in \ker(X)$, $X_s v = 0$ for every worker, hence

$$\widehat{\Sigma}_s v = 0, \qquad Hv = \gamma v.$$

By symmetry, H preserves $\ker(X)$ and its orthogonal complement. Thus its powers commute with Π_X, and

$$\ker(M) = \ker(X), \qquad A := H^{-1/2}\widehat{Z} \in \mathrm{range}(X^\top).$$

Step 2: Objective support and uniqueness. The range identity holds columnwise. For any C,

$$f(C) = f(\Pi_X C), \qquad C^\top C = (\Pi_X C)^\top (\Pi_X C) + \{(I_p - \Pi_X)C\}^\top \{(I_p - \Pi_X)C\}.$$

Eigenvalue monotonicity gives $\|\Pi_X C\|_* \le \|C\|_*$. If $(I_p - \Pi_X)C \neq 0$, the positive semidefinite summand has positive trace, hence at least one eigenvalue and the nuclear norm increase strictly. Every minimizer for $\lambda_1 > 0$ thus lies in $\mathrm{range}(X^\top)$, where f is κ_+-strongly convex. At $\lambda_1 = 0$, the unique solution on this space is the least-squares solution of smallest Frobenius norm.

Step 3: Iterate support. If $\Pi_X D = D$, then $(I_p - M)D + A$ has the same support. Soft thresholding, SCAD, and MCP map zero to zero and keep this support, as does linear extrapolation. Induction from $C_0 = D_0 = 0$ proves the iterate assertions.

Step 4: Rank of the coefficient projection. Commutation gives $H^{-1/2} C_X^\star = \Pi_X B^\star$. Because X is injective and $H^{1/2}$ invertible on the row space of X,

$$\mathrm{rank}(C_X^\star) = \mathrm{rank}(\Pi_X B^\star) = \mathrm{rank}(X\Pi_X B^\star) = \mathrm{rank}(XB^\star) = r_X.$$

This completes the proof. □

The optimality equation gives the prediction bound under $\lambda_1 \ge \varepsilon$.

Proof of Theorem 1. The proof uses restricted strong convexity on the observed predictor space and the subgradient inequality to bound statistical error. An accelerated objective bound then controls finite-iteration error.

Step 1: Restricted subgradient bound. Let $\Delta = \widehat{C} - C_X^\star$. Lemma 3 gives $\Pi_X \Delta = \Delta$, and $MC_X^\star = MC^\star$ gives $\nabla f(C_X^\star) = -\widetilde{W}$. Target optimality yields a nuclear-norm subgradient V such that

$$M\Delta - \widetilde{W} + \lambda_1 V = 0.$$

Let $\mathcal{T}$ be the tangent space at $C_X^\star$. Let $\Delta_\mathcal{T}$ and $\Delta_{\mathcal{T}^\perp}$ be the orthogonal projections of Δ onto $\mathcal{T}$ and its orthogonal complement. The subgradient inequality and Lemmas 13 and 12 give

$$\begin{aligned}\langle M\Delta, \Delta\rangle &\le \left\langle \widetilde{W}, \Delta \right\rangle + \lambda_1\{\|C_X^\star\|_* - \left\|\widehat{C}\right\|_*\} \\ &\le (\lambda_1+\varepsilon)\left\|\Delta_\mathcal{T}\right\|_* + (\varepsilon-\lambda_1)\left\|\Delta_{\mathcal{T}^\perp}\right\|_* \le 2\sqrt{2r_X}\lambda_1\left\|\Delta\right\|_F.\end{aligned}$$

Step 2: Prediction error from restricted strong convexity. On the support of Δ,

$$\kappa_+\left\|\Delta\right\|_F^2 \le \langle M\Delta,\Delta\rangle = \frac{\left\|X(\widehat{B}_1 - B^\star)\right\|_F^2}{N}.$$

Cancel $\|\Delta\|_F$ for the coefficient bound and use $\|\Delta\|_F \le \langle M\Delta,\Delta\rangle^{1/2}/\sqrt{\kappa_+}$ for the prediction bound. If $r_X = 0$, the inequality gives $\Delta = 0$.

Step 3: Finite-iteration bound. The objective bound in Theorem 6 requires no pooled invertibility. Restricted strong convexity and Lemma 3 give

$$\frac{\kappa_+}{2}\left\|C_T - \widehat{C}\right\|_F^2 \le F_1(C_T) - F_1(\widehat{C}) \le \frac{2\left\|\widehat{C}\right\|_F^2}{(T+1)^2}.$$

Taking square roots proves the finite-iteration assertion. This completes the proof. □

S1.4 Proof of Theorem 5

The comparator in the following theorem can have any coefficient rank, and only the rank of its fitted response matrix enters the estimation term.

Theorem 5. *If $\lambda_1 \ge \varepsilon$, then, for every $B^\star \in \mathbb{R}^{p\times q}$,*

$$\frac{\left\|X(\widehat{B}_1 - B^\star)\right\|_F^2}{N} \le \inf_{B\in\mathbb{R}^{p\times q}}\left\{\frac{\|X(B-B^\star)\|_F^2}{N} + \frac{8\lambda_1^2\mathrm{rank}(XB)}{\kappa_+}\right\}.$$

The prediction comparison uses curvature only for differences on the observed predictor space. The subgradient inequality keeps the approximation term unchanged.

Proof. The prediction comparison uses curvature only for differences on the observed predictor space, while the subgradient inequality bounds the noise and penalty terms.

Step 1: Subgradient bound and projection. Let $B \in \mathbb{R}^{p\times q}$ and let $A = \Pi_X H^{1/2} B$, $k = \mathrm{rank}(XB)$, and $D = \widehat{C} - A$. Lemma 3 gives

$$\mathrm{rank}(A) = k, \qquad \Pi_X D = D, \qquad XH^{-1/2}A = XB.$$

Optimality gives $V \in \partial\left\|\widehat{C}\right\|_*$ with $M(\widehat{C} - C_X^\star) - \widetilde{W} + \lambda_1 V = 0$. Let $\mathcal{T}$ be the tangent space generated by the singular vectors of A, with $\mathcal{T} = \{0\}$ if $k = 0$. Let $D_\mathcal{T}$ and $D_{\mathcal{T}^\perp}$ be the corresponding projections. The subgradient inequality and Lemmas 12 and 13 give

$$\begin{aligned}\left\langle M(\widehat{C} - C_X^\star), D\right\rangle &\le \left\langle \widetilde{W}, D\right\rangle + \lambda_1\{\|A\|_* - \left\|\widehat{C}\right\|_*\} \\ &\le (\varepsilon+\lambda_1)\left\|D_\mathcal{T}\right\|_* + (\varepsilon-\lambda_1)\left\|D_{\mathcal{T}^\perp}\right\|_* \\ &\le 2\lambda_1\sqrt{2k}\left\|D\right\|_F \le 2\lambda_1\sqrt{2k/\kappa_+}\left\|M^{1/2}D\right\|_F.\end{aligned}$$

Step 2: Metric expansion and bound. Let $z = \left\|M^{1/2}D\right\|_F$. Expanding the squared norms yields

$$\begin{aligned}&\left\|M^{1/2}(\widehat{C} - C_X^\star)\right\|_F^2 - \left\|M^{1/2}(A - C_X^\star)\right\|_F^2 \\ &\quad= 2\left\langle M(\widehat{C} - C_X^\star), D\right\rangle - z^2 \le 4\lambda_1\sqrt{2k/\kappa_+}\,z - z^2 \le \frac{8k\lambda_1^2}{\kappa_+}.\end{aligned}$$

By the definition of M, the squared norms are the two prediction losses. Take the infimum over B. This completes the proof. □

S1.5 Proof of Theorem 3

The proof compares the exact spectral update with every lower-rank approximation of the mean response. The normal-component inequality in Lemma 6 controls the nonconvex rank constraint; it is the local-concavity inequality of Barber & Ha (2018).

Proof. The proof combines the normal-component inequality for the nonconvex rank constraint with the affine gradient expansion to isolate the optimization and statistical errors.

Step 1: Geometry of the spectral update. Let $\mathcal{P}_k(Z)$ denote a best rank-at-most-k approximation of Z in Frobenius norm. Let $A = C_{t+1}$, $\Delta = A - C_t$, and $N_t = Z_t - A$. The two separation conditions and (5) imply

$$A = \mathcal{P}_k(Z_t), \qquad \operatorname{rank}(A) = k, \qquad \sigma_k(A) = \sigma_k(Z_t), \qquad \|N_t\|_{\mathrm{op}} = \sigma_{k+1}(Z_t).$$

The positive separation between the selected and removed singular values makes this approximation unique. Lemma 3 gives $\Pi_X A = A$ and $\Pi_X C_t = C_t$, and $\Pi_X \widetilde{W} = \widetilde{W}$ because the columns of every $X_s^\top E_s$ lie in $\operatorname{range}(X^\top)$, which $H^{-1/2}$ preserves by that lemma.

Step 2: Restricted strong convexity and gradient expansion. For any B with $\operatorname{rank}(XB) \le k$, let

$$C_B = \Pi_X H^{1/2} B, \qquad D = A - C_B, \qquad V_t = (I_p - M)\Delta.$$

Because H commutes with Π_X and $XH^{-1/2}$ is injective on the row space of X,

$$XH^{-1/2} C_B = XB, \qquad \operatorname{rank}(C_B) = \operatorname{rank}(XB) \le k,$$
$$\operatorname{rank}(D) \le 2k, \qquad \langle MD, D\rangle \ge \kappa_+ \|D\|_F^2 .$$

The gradient is affine. Hence

$$\nabla f(A) = \nabla f(C_t) + M\Delta = -N_t - V_t, \qquad M(A - C_X^\star) = \widetilde{W} - N_t - V_t.$$

Also $\|V_t\|_F \le \|\Delta\|_F$, because $0 \preceq M \preceq I_p$.

Step 3: Prediction error bound. Lemma 6, applied with $Y = C_B$, yields

$$-2\langle N_t, D\rangle \le \vartheta_t \|D\|_F^2 .$$

This inequality also holds when $k = p \wedge q$, for then $N_t = 0$. Expanding the difference of the two prediction errors gives

$$\begin{aligned}
&\frac{\left\|X(H^{-1/2}A - B^\star)\right\|_F^2}{N} - \frac{\|X(B - B^\star)\|_F^2}{N} \\
&\quad = 2\langle M(A - C_X^\star), D\rangle - \langle MD, D\rangle \\
&\quad \le 2\left\langle \widetilde{W} - V_t, D\right\rangle - (\kappa_+ - \vartheta_t)\|D\|_F^2 \\
&\quad \le 2\{\sqrt{2k}\varepsilon + \|\Delta\|_F\}\|D\|_F - (\kappa_+ - \vartheta_t)\|D\|_F^2 \\
&\quad \le \frac{\{\sqrt{2k}\varepsilon + \|\Delta\|_F\}^2}{\kappa_+ - \vartheta_t}.
\end{aligned}$$

The penultimate inequality uses $|\left\langle \widetilde{W}, D\right\rangle| \le \varepsilon \|D\|_* \le \sqrt{2k}\varepsilon \|D\|_F$; the last inequality maximizes a scalar quadratic. Taking the infimum over B proves the result. The argument is deterministic for every exact update satisfying the conditions. This completes the proof. □

S1.6 Boundedness of exact Stage II updates

A folded-concave penalty is bounded and cannot make the objective coercive in the pooled null space. The exact spectral update nevertheless controls that component of the iterate. This gives boundedness of the computed sequence without a full-rank design or a restriction on its Stage II starting point.

Lemma 4. *For exact SCAD or MCP updates from any finite C_{T_1},*

$$\|(I_p - \Pi_X)C_{t+1}\|_F \le \|(I_p - \Pi_X)C_t\|_F , \qquad t \ge T_1, \qquad \sup_{t \ge T_1} \|C_t\|_F < \infty.$$

Lemma 4 controls the actual sequence, not every sublevel set. Its proof uses exact spectral thresholding. Arbitrary inexact updates need not preserve the null-space bound, which is why the separate inexact convergence result keeps its curvature condition.

Proof. The proof relies on exact spectral thresholding to control the null-space component. The gradient step leaves this component unchanged, and thresholding cannot increase its norm.

Step 1: Null-space norm bound. Let $Z_t = U_t \mathrm{diag}(d_t) V_t^\top$. For either penalty, its scalar map θ satisfies $0 \le \theta(z) \le z$ for $z \ge 0$. Define

$$A_t = \sum_{j: d_{t,j} > 0} \frac{\theta(d_{t,j})}{d_{t,j}} v_{t,j} v_{t,j}^\top,$$

where $v_{t,j}$ is column j of V_t. Then $0 \preceq A_t \preceq I_q$ and $C_{t+1} = Z_t A_t$. Because $\nabla f(C_t) = MC_t - H^{-1/2}\widehat{Z}$ has columns in the row space of X,

$$(I_p - \Pi_X) Z_t = (I_p - \Pi_X) C_t, \qquad \left\| (I_p - \Pi_X) C_{t+1} \right\|_F \le \left\| (I_p - \Pi_X) C_t \right\|_F .$$

Step 2: Objective coercivity on the observed space. For $A = H^{-1/2}\widehat{Z}$ and $D = \Pi_X C$, the quadratic identity and Young's inequality give

$$f(C) = f(D) \ge \frac{\kappa_+}{2} \|D\|_F^2 - \|A\|_F \|D\|_F \ge \frac{\kappa_+}{4} \|D\|_F^2 - \frac{\|A\|_F^2}{\kappa_+}.$$

The loss constant $\|Y\|_F^2/(2N)$ and both penalties are nonnegative. Exact proximal minimization gives the descent bound of Theorem 2 for SCAD and MCP.

Step 3: Boundedness of the iterate sequence. Descent bounds the objective by its value at C_{T_1}, hence $\Pi_X C_t$ is bounded. The null-space contraction bounds the orthogonal component, hence C_t is bounded. This completes the proof. □

S1.7 Coefficient and objective bounds for Stage I

The following theorem establishes coefficient error under positive pooled curvature and objective convergence without pooled invertibility.

Theorem 6. *If $\kappa > 0$ and $\lambda_1 \ge \varepsilon$, then*

$$\left\| \widehat{C} - C^\star \right\|_F \le \frac{2\sqrt{2r}\lambda_1}{\kappa}, \qquad \frac{\left\| X(\widehat{B}_1 - B^\star) \right\|_F}{\sqrt{N}} \le \frac{2\sqrt{2r}\lambda_1}{\sqrt{\kappa}}.$$

If $\kappa > 0$, the Stage I iterates from any C_0 satisfy, for every integer $1 \le T \le T_1$,

$$F_1(C_T) - F_1(\widehat{C}) \le \frac{2\left\| C_0 - \widehat{C} \right\|_F^2}{(T+1)^2}, \qquad \left\| C_T - \widehat{C} \right\|_F \le \frac{2\left\| C_0 - \widehat{C} \right\|_F}{\sqrt{\kappa}(T+1)}.$$

The objective inequality remains valid without $\kappa > 0$ or $\lambda_1 \ge \varepsilon$.

Proof. The proof first establishes the coefficient error under positive pooled curvature using the subgradient inequality, and then converts the standard accelerated objective bound into distance from the centralized target.

Step 1: Coefficient and prediction bounds. Let $\Delta = \widehat{C} - C^\star$ and let $\mathcal{T}$ be the tangent space at $C^\star$, with orthogonal projections $\Delta_{\mathcal{T}}$ and $\Delta_{\mathcal{T}^\perp}$. Optimality in (4) gives $V \in \partial \left\| \widehat{C} \right\|_*$ with $M\Delta - \widetilde{W} + \lambda_1 V = 0$. The subgradient inequality, Lemmas 13 and 12, and $\lambda_1 \ge \varepsilon$ give

$$\begin{aligned} \langle M\Delta, \Delta \rangle &\le \left\langle \widetilde{W}, \Delta \right\rangle + \lambda_1 \{ \|C^\star\|_* - \left\| \widehat{C} \right\|_* \} \\ &\le (\lambda_1 + \varepsilon) \|\Delta_{\mathcal{T}}\|_* + (\varepsilon - \lambda_1) \|\Delta_{\mathcal{T}^\perp}\|_* \le 2\sqrt{2r}\lambda_1 \|\Delta\|_F . \end{aligned}$$

Because $M \succeq \kappa I_p$, this proves the coefficient bound. Substituting $\|\Delta\|_F \le \langle M\Delta, \Delta\rangle^{1/2}/\sqrt{\kappa}$ into the same inequality proves the prediction bound, using $\langle M\Delta, \Delta\rangle = \left\|X(\widehat{B}_1 - B^\star)\right\|_F^2/N$.

Step 2: Objective and iterate convergence. By Lemma 1, ∇f has Lipschitz constant at most one. Theorem 4.4 of Beck & Teboulle (2009), with step one, gives

$$F_1(C_T) - F_1(\widehat{C}) \le 2\left\|C_0 - \widehat{C}\right\|_F^2/(T+1)^2.$$

This holds without $\kappa > 0$ or $\lambda_1 \ge \varepsilon$. For $\kappa > 0$, strong convexity gives

$$\tfrac{1}{2}\kappa\left\|C_T - \widehat{C}\right\|_F^2 \le F_1(C_T) - F_1(\widehat{C}).$$

This completes the proof. □

S1.8 Proof of Corollary 1

The argument uses concentration of squared projections of independent sub-Gaussian rows and a conditional matrix deviation bound for the noise term.

Lemma 5. *Under Condition 1, without its lower second-moment bound, for every $t > 0$, the following two bounds hold with probability at least $1 - 4e^{-t}$:*

$$\left\|\widehat{\Sigma} - \textstyle\sum_s w_s \Sigma_s^0\right\|_{\mathrm{op}} \le CK_x^2 L\left\{\sqrt{\frac{p+t}{N}} + \frac{p+t}{N}\right\},$$

$$\max_s \left\|\widehat{\Sigma}_s - \Sigma_s^0\right\|_{\mathrm{op}} \le CK_x^2 L\left\{\sqrt{\frac{p+t+\log m}{n_{\min}}} + \frac{p+t+\log m}{n_{\min}}\right\},$$

Conditional on the designs, with probability at least $1 - 2e^{-t}$,

$$\left\|\widetilde{W}\right\|_{\mathrm{op}} \le C\sigma\sqrt{\|M\|_{\mathrm{op}}}\sqrt{\frac{p+q+t}{N}}.$$

Lemma 5 controls the empirical quantities in the deterministic theory without imposing an empirical Gram matrix assumption.

Proof. The proof combines the concentration of sub-exponential empirical quadratic forms over a covering net with a conditional sub-Gaussian bound for the score matrix.

Step 1: Gram matrix concentration. For fixed u in the Euclidean unit sphere $\mathbb{S}^{p-1}$, the centered variables $(u^\top x_{si})^2 - \mathbb{E}(u^\top x_{si})^2$ have sub-exponential norm at most $CK_x^2 L$. Bernstein's inequality gives

$$\Pr\left(\left|u^\top\left\{\widehat{\Sigma} - \sum_s w_s\Sigma_s^0\right\}u\right| > z\right) \le 2\exp\left[-cN\min\left\{\frac{z^2}{K_x^4L^2}, \frac{z}{K_x^2L}\right\}\right].$$

A $1/4$-net has at most 9^p points and bounds a symmetric operator norm by twice the largest absolute quadratic form (Vershynin, 2018, Chapters 2 and 4). A union bound with $z = CK_x^2L\{(p+t)/N\}^{1/2} + CK_x^2L(p+t)/N$ proves the first inequality; summing the worker bounds proves the second.

Step 2: Score matrix deviation. Conditional on the designs, for $u \in \mathbb{S}^{p-1}$ and $v \in \mathbb{S}^{q-1}$,

$$u^\top\widetilde{W}v = \frac{1}{N}\sum_{s,i} u^\top H^{-1/2}x_{si}\, e_{si}^\top v$$

is sub-Gaussian with squared scale at most

$$\frac{C\sigma^2}{N^2}\sum_{s,i}(u^\top H^{-1/2}x_{si})^2 = \frac{C\sigma^2}{N}u^\top Mu \le \frac{C\sigma^2\,\|M\|_{\mathrm{op}}}{N}.$$

A union bound over $1/4$-nets of $\mathbb{S}^{p-1}$ and $\mathbb{S}^{q-1}$ proves the last inequality. This completes the proof. □

Proof of Corollary 1. The proof controls the minimum pooled eigenvalue and the maximum local eigenvalue simultaneously to bound the transformed curvature from below before applying the score deviation inequality.

Step 1: Curvature bounds. Let $t = \log(6/\delta)$. Proposition 18 and the sample-size premise give $\widehat{\Sigma} \succeq \Sigma^0/2$, hence

$$\lambda_{\min}(\widehat{\Sigma}) \geq \mu/2$$

with failure probability at most $2e^{-t}$. The local bound in Lemma 5 gives

$$\max_s \ell_s \lesssim LK_x^2\{1 + p/n_{\min} + \log(m/\delta)/n_{\min}\}$$

Here $K_x \geq 1$, because the ψ_2 norm of $u^\top x_{si}$ is at least its root second moment, and $\log(6m/\delta) \lesssim \log(m/\delta)$ because $m \geq 2$. Combining the two bounds gives

$$\kappa = \lambda_{\min}(H^{-1/2}\widehat{\Sigma}H^{-1/2}) \geq \frac{\lambda_{\min}(\widehat{\Sigma})}{\lambda_{\max}(H)} \geq \frac{\mu/2}{\gamma + \ell_1}.$$

Step 2: Score bound and intersection. The conditional bound in Lemma 5 gives

$$\left\|\widetilde{W}\right\|_{\text{op}} \lesssim \sigma\sqrt{\|M\|_{\text{op}}}\sqrt{\{p + q + \log(1/\delta)\}/N}.$$

By Lemma 1, $\|M\|_{\text{op}} \leq (1+\tau)^{-1}$. The two Gram matrix bounds and the conditional score bound each fail with probability at most $2e^{-t}$. Their union bound is $6e^{-t} = \delta$. This completes the proof. □

S1.9 Proof of Theorem 2

The proof uses only the exact proximal property in Lemma 2 and the curvature bound in Lemma 1.

Proof. The proof combines the exact proximal optimality bound with the curvature inequality to establish monotonic descent and square summability of the update steps.

Step 1: Proximal optimality and descent bound. Let $\Delta_t = C_{t+1} - C_t$. Proximal optimality relative to C_t gives

$$\langle \nabla f(C_t), \Delta_t \rangle + \tfrac{1}{2}\|\Delta_t\|_F^2 + \sum_j P_{\lambda_2}\{\sigma_j(C_{t+1})\} - \sum_j P_{\lambda_2}\{\sigma_j(C_t)\} \leq 0.$$

Because $\|M\|_{\text{op}} \leq (1+\tau)^{-1}$,

$$f(C_{t+1}) - f(C_t) \leq \langle \nabla f(C_t), \Delta_t \rangle + \frac{1}{2(1+\tau)}\|\Delta_t\|_F^2.$$

Adding these inequalities proves

$$F_2(C_t) - F_2(C_{t+1}) \geq \frac{\tau}{2(1+\tau)}\|\Delta_t\|_F^2.$$

Step 2: Stationarity of cluster points. The objective is bounded below, which proves square summability. Necessary optimality for the proximal problem gives

$$0 \in \Delta_t + \nabla f(C_t) + \partial F_{2,\text{pen}}(C_{t+1}),$$

where $F_{2,\text{pen}}$ is the SCAD spectral penalty and ∂ is the limiting subdifferential. Along any convergent subsequence, $\Delta_t \to 0$; continuity of ∇f and closedness of the limiting subdifferential give $0 \in \nabla f(\bar{C}) + \partial F_{2,\text{pen}}(\bar{C})$. This completes the proof. □

S1.10 Proof of Theorem 4 and Corollary 3

The proof identifies the exact singular-value ranges used by the computed proximal map.

Proof of Theorem 4. The proof tracks the singular values of the surrogate variable relative to the transformed true coefficient matrix. The separation condition identifies which singular values the SCAD update leaves unchanged.

Step 1: Surrogate perturbation bound. The regression model gives

$$\nabla f(C_t) = M(C_t - C^\star) - \widetilde{W}, \qquad Z_t = C^\star + (I_p - M)(C_t - C^\star) + \widetilde{W} = C^\star + R_t.$$

Weyl's inequality gives

$$\sigma_{r+1}(Z_t) \le \eta_t < \lambda_2, \qquad \sigma_r(Z_t) \ge \sigma_r(C^\star) - \eta_t > a\lambda_2.$$

Step 2: Exact thresholding mapping. Equation (5) sets every singular value after r equal to zero and leaves the first r unchanged. Hence C_{t+1} is the best rank-r approximation of Z_t. Its error satisfies

$$\|C_{t+1} - C^\star\|_{\mathrm{op}} \le \|C_{t+1} - Z_t\|_{\mathrm{op}} + \|R_t\|_{\mathrm{op}} \le 2\eta_t.$$

The difference has rank at most $2r$, which gives the Frobenius bound. The observable statement follows directly from (5). This completes the proof. □

Define

$$\rho_{T_1} = (1-\kappa)\left\{\frac{2\left\|C_0 - \widehat{C}\right\|_F}{\sqrt{\kappa}(T_1+1)} + \frac{2\sqrt{2r}\lambda_1}{\kappa}\right\} + \varepsilon.$$

The following corollary gives a sufficient Stage I duration for entering the rank-identification region.

Corollary 3. *Suppose $\kappa > 0$, $\lambda_1 \ge \varepsilon$, $T_1 \ge 1$, and $T_2 \ge 1$. The first Stage II update obeys Theorem 4 whenever*

$$\rho_{T_1} < \lambda_2, \qquad \sigma_r(C^\star) > a\lambda_2 + \rho_{T_1}.$$

Proof of Corollary 3. The proof bounds the iterate distance at the end of Stage I to verify rank recovery in the first Stage II update.

Step 1: Iterate distance bound. Because the eigenvalues of M lie in $[\kappa, 1]$,

$$\eta_{T_1} \le (1-\kappa)\|C_{T_1} - C^\star\|_F + \varepsilon.$$

The triangle inequality and Theorem 6 give

$$\|C_{T_1} - C^\star\|_F \le \frac{2\left\|C_0 - \widehat{C}\right\|_F}{\sqrt{\kappa}(T_1+1)} + \frac{2\sqrt{2r}\lambda_1}{\kappa}.$$

Step 2: Sufficient conditions. The stated conditions imply those of Theorem 4. This completes the proof. □

S1.11 Proof of Proposition 1

The proposition compares scalar counts for one fitted value of the tuning parameters.

Proof. The proof counts the scalar transmissions for exact gradients and directly compares them to the sufficient-statistic dimensions.

Step 1: Protocol accounting. Setup sends ℓ_s; sample sizes are registered for both protocols. Each round sends and returns pq entries, giving $S_{\mathrm{grad}} = 1 + 2Tpq$. Direct transmission sends $p(p+1)/2$ Gram matrix entries and pq cross-product entries.

Step 2: Comparison inequality. Thus

$$1 + 2Tpq < p(p+1)/2 + pq$$

is equivalent to the stated inequality. This completes the proof. □

S2 RANK RECOVERY THROUGH THE RETURNED ITERATE

The first Stage II update can recover rank without ensuring that later updates remain in the same region. We prove a sufficient condition for all subsequent updates. The argument uses the curvature of the set of low-rank matrices to control the discarded normal component. This removes the rank-dependent contraction condition in Proposition 15.

The direct perturbation bound in Proposition 15 uses $A_r = 2\sqrt{2r}(1-\kappa)$. If all worker Gram matrices equal a positive-definite matrix $\widehat{\Sigma}$, simultaneous diagonalization gives $\kappa = \lambda_{\min}(\widehat{\Sigma})/\{\lambda_{\min}(\widehat{\Sigma}) + (1+\tau)\lambda_{\max}(\widehat{\Sigma})\} \le (2+\tau)^{-1}$. Hence $A_r > 1$ for every $r \ge 1$, even in this case of identical worker Gram matrices. The following local bound controls persistent rank recovery without requiring $A_r < 1$.

S2.1 LOCAL GEOMETRY OF THE SPECTRAL UPDATE

For a matrix Z, let $\mathcal{P}_r(Z)$ be a best approximation of rank at most r in Frobenius norm. The following inequality is the low-rank local-concavity bound of Barber & Ha (2018), Lemma 5.1 and Appendix C.1. We give the short approximation argument to specify the form needed here.

Let $A = \mathcal{P}_r(Z)$, $N = Z - A$, and $s = \sigma_r(A)$.

Lemma 6. *If $s > 0$, then for every matrix Y of rank at most r,*

$$\langle N, Y - A\rangle \le \frac{\|N\|_{\mathrm{op}}}{2s}\|Y - A\|_F^2 .$$

Lemma 6 controls the normal component discarded by the spectral update. Its constant depends on the smallest kept singular value, not on the number of kept singular values.

Proof. The proof compares squared distances from a scaled perturbation to the best rank-r approximation and to any matrix satisfying the same rank bound. Expanding the squares gives the inner-product bound.

Step 1: Singular value decomposition and optimality. For $N = 0$ the bound holds directly. For $N \ne 0$, let $n = \|N\|_{\mathrm{op}}$. A singular value decomposition gives $A^\top N = AN^\top = 0$. Because $\|(s/n)N\|_{\mathrm{op}} = s$, A is a best rank-r approximation of $A + (s/n)N$, including ties at s.

Step 2: Distance expansion. Hence

$$\|(s/n)N\|_F^2 \le \|A + (s/n)N - Y\|_F^2 = \|A - Y\|_F^2 + \|(s/n)N\|_F^2 - 2(s/n)\langle N, Y - A\rangle .$$

Rearrangement proves the inequality. This completes the proof. □

S2.2 PERSISTENCE AND ERROR CONTRACTION

The next theorem controls rank and estimation error throughout the Stage II sequence.

Let $e_t = \|C_t - C^\star\|_F$ and $s_\star = \sigma_r(C^\star)$. For a radius $R > 0$, let $\bar\eta = (1-\kappa)R + \varepsilon$ and $q_\kappa = (1-\kappa)/(1-\kappa/2)$. We consider the Stage II iteration of Algorithm 1 continued with the same penalty and metric.

Theorem 7. *Suppose $\kappa > 0$, $t_0 \ge T_1$, $e_{t_0} \le R$, and*

$$R \ge \frac{2\sqrt{2r}\varepsilon}{\kappa}, \qquad \bar\eta < \lambda_2, \qquad s_\star > a\lambda_2 + \bar\eta, \qquad \bar\eta \le \frac{\kappa s_\star}{1+\kappa}.$$

Then, for every $k \ge 1$, the kept singular values are unshrunk and

$$\mathrm{rank}(C_{t_0+k}) = r, \qquad e_{t_0+k} \le q_\kappa^k e_{t_0} + \frac{2\sqrt{2r}\varepsilon}{\kappa}(1 - q_\kappa^k).$$

Theorem 7 bounds every subsequent iterate by a geometrically decreasing initial error and a statistical error term. It permits every $\kappa > 0$, with stronger separation and a larger error bound when pooled curvature is weak. The condition $e_{t_0} \le R$ is an intermediate deterministic premise. Corollary 12 derives it from the prescribed zero initialization and population assumptions.

Proof. The proof combines the exact thresholding mapping with the local concavity bound of the low-rank projection to obtain an error recurrence with a contraction factor and an additive noise term.

Step 1: Rank and singular value bounds. Suppose $e_t \le R$. Then $\eta_t \le \bar{\eta}$. Theorem 4 and (5) give

$$A := C_{t+1} = \mathcal{P}_r(Z_t), \qquad \operatorname{rank}(A) = r, \qquad \sigma_r(A) \ge s_\star - \bar{\eta}.$$

Let $N_t = Z_t - A$ and $E = A - C^\star$. Weyl's inequality gives $\|N_t\|_{\mathrm{op}} = \sigma_{r+1}(Z_t) \le \bar{\eta}$.

Step 2: Error contraction inequality. Because $E = R_t - N_t$, Lemma 6 with $Y = C^\star$ yields

$$\left(1 - \frac{\|N_t\|_{\mathrm{op}}}{2\sigma_r(A)}\right) \|E\|_F^2 \le \langle R_t, E\rangle \le \{(1-\kappa)e_t + \sqrt{2r}\varepsilon\} \|E\|_F .$$

The last step uses $\operatorname{rank}(E) \le 2r$, which gives $\|E\|_* \le \sqrt{2r}\,\|E\|_F$. The local signal condition gives

$$\frac{\|N_t\|_{\mathrm{op}}}{2\sigma_r(A)} \le \frac{\bar{\eta}}{2(s_\star - \bar{\eta})} \le \frac{\kappa}{2}.$$

Step 3: Geometric recurrence. If $E = 0$ the desired bound is immediate. Division in the nonzero case gives

$$e_{t+1} \le q_\kappa e_t + \frac{\sqrt{2r}\varepsilon}{1 - \kappa/2}.$$

Because $0 < \kappa < 1$, $0 < q_\kappa < 1$. The radius condition implies

$$q_\kappa R + \frac{\sqrt{2r}\varepsilon}{1 - \kappa/2} \le R.$$

Induction preserves the radius and rank r without shrinkage. Summing the geometric recurrence proves the error bound without restricting $\operatorname{rank}(C_{t_0})$. This completes the proof. □

S3 Prediction of the computed Stage II estimates

Rank recovery alone does not identify the least-squares fit of that rank. We show that a computed Stage II update approaches this fit when the discarded singular component is small relative to the positive pooled curvature. The argument applies on the observed predictor space and does not require an invertible pooled Gram matrix. The limiting statistical error bound has no inverse-curvature factor; the signal and iteration conditions still depend on curvature.

S3.1 A finite-iteration comparison with reduced-rank regression

Let P_X be the orthogonal projection onto the column space of X, and let $\xi_X = \|P_X E\|_{\mathrm{op}}/\sqrt{N}$. This noise quantity is used only in the analysis. For an exact Stage II update, let $A = C_{t+1}$, $N_t = Z_t - A$, and $k = \operatorname{rank}(A)$. When $1 \le k \le d \wedge q$, define

$$\nu_t = \frac{\|N_t\|_{\mathrm{op}}}{\sigma_k(A)}, \qquad b_t = 1 - \frac{\nu_t}{\kappa_+}, \qquad h_t = \frac{\|C_t - A\|_F}{\sqrt{\kappa_+}}.$$

The next result compares the actual update with every rank-k fit. A stationarity guarantee alone does not give this comparison.

Theorem 8. *Let $t \ge T_1$, $1 \le k \le d \wedge q$, and $b_t > 0$. If $\sigma_k(Z_t) > a\lambda_2$ and $\sigma_{k+1}(Z_t) \le \lambda_2$, then*

$$0 \le f(A) - \min_{\substack{\Pi_X D = D \\ \operatorname{rank}(D) \le k}} f(D) \le \frac{h_t^2}{2b_t},$$

$$\frac{\left\|X(H^{-1/2}A - B^\star)\right\|_F^2}{N} \le \inf_{\substack{B \in \mathbb{R}^{p\times q} \\ \operatorname{rank}(XB) \le k}} \left\{ \frac{\|X(B - B^\star)\|_F^2}{N} + \frac{\{\xi_X\sqrt{k + \operatorname{rank}(XB)} + h_t\}^2}{b_t} \right\}.$$

Theorem 8 separates approximation, noise, and incomplete optimization. The coefficient of the approximation error is one. This deterministic comparison does not require knowledge of the true rank. The singular values and step difference are available at the master, but κ_+ is not determined by the setup scalars alone. A numerical certificate needs an additional lower curvature bound; computing it from pooled sufficient statistics must be included in the communication cost. Corollary 35 instead gives sufficient conditions in population parameters.

Proof. The proof expands the objective differences using the affine gradient on the observed predictor space and bounds the normal component of the exact spectral update.

Step 1: Objective lower bound. The singular-value conditions and (5) imply $A = \mathcal{P}_k(Z_t)$. Lemma 3 gives $\Pi_X A = A$ and $\Pi_X C_t = C_t$. For D in the stated constraint set, let $u = \left\|M^{1/2}(D - A)\right\|_F$. On the observed predictor space,

$$\|D - A\|_F \le u/\sqrt{\kappa_+}, \qquad \nabla f(A) = (I_p - M)(C_t - A) - N_t.$$

By Lemma 6 and $\|I_p - M\|_{\mathrm{op}} \le 1$, quadratic expansion gives

$$\begin{aligned} f(D) - f(A) &= \langle (I_p - M)(C_t - A) - N_t, D - A\rangle + \tfrac{1}{2}u^2 \\ &\ge \tfrac{1}{2}u^2 - h_t u - \tfrac{1}{2}\nu_t \|D - A\|_F^2 \ge \tfrac{1}{2}b_t u^2 - h_t u. \end{aligned}$$

The rank constraint is closed, and f is coercive on the observed predictor space because $\kappa_+ > 0$. Its minimum over that constraint set exists. When $k = d \wedge q$, the discarded component is zero and the same inequality holds. Minimizing the last quadratic over $u \ge 0$ proves the first bound.

Step 2: Geometry of the score projection. For the noise term, let $\dagger$ denote the Moore–Penrose inverse and define $M^{\dagger/2} = (M^\dagger)^{1/2}$. Let $Q = N^{-1/2}XH^{-1/2}M^{\dagger/2}$. A singular value decomposition of $XH^{-1/2}$ gives

$$Q^\top Q = \Pi_X, \qquad QQ^\top = P_X, \qquad M^{\dagger/2}\widetilde{W} = Q^\top E/\sqrt{N}, \qquad \left\|M^{\dagger/2}\widetilde{W}\right\|_{\mathrm{op}} = \xi_X.$$

Step 3: Prediction error comparison. For any B in the comparison, let $D = \Pi_X H^{1/2}B$ and $j = \mathrm{rank}(XB)$. The commutation and injectivity argument in the proof of Lemma 3 gives $\mathrm{rank}(D) = j$ and $XH^{-1/2}D = XB$. Also,

$$\mathrm{rank}\{M^{1/2}(A - D)\} \le k + j, \qquad |\left\langle \widetilde{W}, A - D\right\rangle| \le \xi_X\sqrt{k + j}\, u.$$

Expanding $Y = XB^\star + E$ gives

$$\begin{aligned} &\frac{\left\|X(H^{-1/2}A - B^\star)\right\|_F^2 - \|X(B - B^\star)\|_F^2}{N} \\ &= 2\{f(A) - f(D)\} + 2\left\langle \widetilde{W}, A - D\right\rangle \\ &\le -b_t u^2 + 2\{h_t + \xi_X\sqrt{k + j}\}u \le \frac{\{h_t + \xi_X\sqrt{k + j}\}^2}{b_t}. \end{aligned}$$

Taking the infimum over B proves the prediction inequality. This completes the proof. □

The projected noise admits a bound that is independent of the positive singular values of X. Conditional on X, assume independent centered error rows satisfying the noise bound in Condition 1.

Proposition 2. *For $0 < \delta < 1$, with conditional probability at least $1 - \delta$,*

$$\xi_X \le C\sigma\sqrt{\frac{d + q + \log(2/\delta)}{N}}.$$

Proposition 2 controls the noise in the fitted response space. The same bound can be used in Theorem 8 simultaneously at every update satisfying its deterministic premises, including a data-dependent stopping iteration. It does not assert that these premises hold at a selected tuning value.

Proof. The proof applies a standard covering argument to the projection of the sub-Gaussian error matrix onto the column space of the predictors.

Step 1: Conditional deviation and covering net. Choose an orthonormal basis $U \in \mathbb{R}^{N\times d}$ for the column space of X. Then $\|P_X E\|_{\mathrm{op}} = \|U^\top E\|_{\mathrm{op}}$. For unit vectors $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^q$, conditional independence gives

$$\|u^\top U^\top E v\|_{\psi_2}^2 \le C\sigma^2 \sum_{i=1}^{N} \{(Uu)_i\}^2 = C\sigma^2.$$

Apply Lemmas 21 and 23 to $1/4$-nets in the two spheres. The sub-Gaussian tail bound and a union bound give

$$\Pr\{\|U^\top E\|_{\mathrm{op}} > 2z \mid X\} \le 2\exp\{(d+q)\log 9 - cz^2/\sigma^2\}.$$

Step 2: Probability bound. Taking $z = C\sigma\sqrt{d+q+\log(2/\delta)}$ and dividing by $\sqrt{N}$ proves the assertion. The bound concerns only X, E, which proves the simultaneous interpretation. This completes the proof. □

S3.2 Convergence to the rank-constrained least-squares fit

We next strengthen the separation condition in the persistent-rank result. For $1 \le r_X < d \wedge q$, let $e_{X,t} = \|C_t - C_X^\star\|_F$, $s_X = \sigma_{r_X}(C_X^\star)$, $\bar\eta_X = (1-\kappa_+)R + \varepsilon$, and $q_+ = (1-\kappa_+)/(1-\kappa_+/2)$. The result concerns the unchanged exact SCAD sequence, not an additional rank-constrained refit.

Theorem 9. *Let $t_0 \ge T_1$, $R > 0$, and $1 \le r_X < d \wedge q$. Suppose*

$$e_{X,t_0} \le R, \qquad R \ge \frac{2\sqrt{2r_X}\varepsilon}{\kappa_+}, \qquad \bar\eta_X < \lambda_2, \qquad s_X > a\lambda_2 + \bar\eta_X, \qquad \bar\eta_X \le \frac{\kappa_+ s_X}{2+\kappa_+}.$$

There is a unique C^{o} minimizing $f(C)$ over $\Pi_X C = C$, $\mathrm{rank}(C) \le r_X$. For every integer $j \ge 1$,

$$\mathrm{rank}(C_{t_0+j}) = \mathrm{rank}(C^{\mathrm{o}}) = r_X, \qquad \|C_{t_0+j} - C^{\mathrm{o}}\|_F \le q_+^j \|C_{t_0} - C^{\mathrm{o}}\|_F \le 2Rq_+^j,$$

$$\frac{\|X(H^{-1/2}C_{t_0+j} - B^\star)\|_F}{\sqrt{N}} \le 2\sqrt{r_X}\,\xi_X + 2\sqrt{\|M\|_{\mathrm{op}}}Rq_+^j.$$

Theorem 9 identifies the statistical target of the computed Stage II sequence under stronger signal separation. Its limit is the ordinary reduced-rank least-squares fit on the observed predictor space. Uniqueness does not extend to unobserved coefficient components, and no global minimization of F_2 is asserted. Proposition 2 gives prediction error of order $\sigma\sqrt{r_X(d+q)/N}$ apart from confidence and iteration terms. This is the usual low-rank prediction order, not a new minimax rate. The entry and separation conditions, and the geometric factor q_+, still deteriorate when κ_+ is small.

Proof. The proof shows that exact Stage II iterates converge to the unique rank-constrained least-squares fit on the observed predictor space by verifying strict separation and positive curvature.

Step 1: Iterate boundedness and fixed point. The stronger last premise implies the premises of Corollary 33. Hence, for $t \ge t_0$,

$$e_{X,t} \le R, \qquad A = C_{t+1} = \mathcal{P}_{r_X}(Z_t), \qquad \sigma_{r_X}(A) \ge s_X - \bar\eta_X > 0,$$

$$\|N_t\|_{\mathrm{op}} \le \bar\eta_X, \qquad \frac{\|N_t\|_{\mathrm{op}}}{\sigma_{r_X}(A)} \le \frac{\bar\eta_X}{s_X - \bar\eta_X} \le \frac{\kappa_+}{2}.$$

Corollary 2 gives a limit C^{o}. The uniform lower singular-value bound gives $\mathrm{rank}(C^{\mathrm{o}}) = r_X$. With $Z^{\mathrm{o}} = C^{\mathrm{o}} - \nabla f(C^{\mathrm{o}})$ and $N^{\mathrm{o}} = Z^{\mathrm{o}} - C^{\mathrm{o}}$, continuity gives

$$C^{\mathrm{o}} = \mathcal{P}_{r_X}(Z^{\mathrm{o}}), \qquad \frac{\|N^{\mathrm{o}}\|_{\mathrm{op}}}{\sigma_{r_X}(C^{\mathrm{o}})} \le \kappa_+/2.$$

Step 2: Global optimality and uniqueness. The singular-value gap is strict because the first r_X singular values exceed $a\lambda_2$ and the next is at most $\bar\eta_X < \lambda_2$. At this fixed point the proof of Theorem 8, with $h_t = 0$ and $b_t \ge 1/2$, gives

$$f(D) - f(C^{\mathrm{o}}) \ge \tfrac{1}{4}\left\|M^{1/2}(D - C^{\mathrm{o}})\right\|_F^2$$

for every $\Pi_X D = D$ of rank at most r_X. Positive curvature on that space proves global minimality and uniqueness. Taking $B = B^\star$ in its prediction inequality also gives

$$\frac{\left\|X(H^{-1/2}C^{\mathrm{o}} - B^\star)\right\|_F}{\sqrt{N}} \le 2\sqrt{r_X}\,\xi_X.$$

Step 3: Geometric contraction. For the geometric bound, let $\Delta = A - C^{\mathrm{o}}$. The quadratic gradient identity gives

$$Z_t - Z^{\mathrm{o}} = (I_p - M)(C_t - C^{\mathrm{o}}), \qquad \Delta = (I_p - M)(C_t - C^{\mathrm{o}}) - (N_t - N^{\mathrm{o}}).$$

Apply Lemma 6 first at A with comparison C^{o} and then at C^{o} with comparison A. All differences belong to the observed predictor space. Thus

$$\|\Delta\|_F^2 \le (1 - \kappa_+)\,\|C_t - C^{\mathrm{o}}\|_F\,\|\Delta\|_F + \tfrac{1}{2}\kappa_+\,\|\Delta\|_F^2\,.$$

Division when $\Delta \neq 0$ proves the one-step contraction; the zero case satisfies it directly. Iteration gives the first error bound. Both C_{t_0} and C^{o} are within R of $C_X^\star$, and hence their distance is at most $2R$. Finally,

$$\frac{\left\|XH^{-1/2}(C_{t_0+j} - C^{\mathrm{o}})\right\|_F}{\sqrt{N}} \le \sqrt{\|M\|_{\mathrm{op}}}\,\|C_{t_0+j} - C^{\mathrm{o}}\|_F\,.$$

The triangle inequality proves the prediction bound. This completes the proof. □

The projection inequality in Lemma 6 is established in Barber & Ha (2018). Nonconvex spectral penalties and rank recovery are studied in Gui et al. (2016). The conclusion here specifies when the unchanged distributed sequence reaches the unpenalized reduced-rank fit and gives a finite-iteration prediction comparison on a possibly singular observed predictor space. It does not establish that the sufficient signal conditions describe the weakest possible regime.

S4 Additional theoretical results

The stabilized penalty dominates the pooled prediction norm and is comparable to it when the pooled design has full column rank. A singular design still permits prediction bounds, although some coefficient directions are unidentifiable. The following results also compare prediction with a minimax bound and control finite-iteration error.

S4.1 Prediction geometry of the matched penalty

The proof of Proposition 3 uses an order relation for matrix square roots. The following lemma records the required relation before it is applied to the metric penalty.

Lemma 7. *Let $A_1, A_2 \in \mathbb{R}^{p\times p}$ be positive semidefinite and let $A_1 \preceq A_2$. Then, for every $B \in \mathbb{R}^{p\times q}$,*

$$\left\|A_1^{1/2}B\right\|_F \le \left\|A_2^{1/2}B\right\|_F, \qquad \left\|A_1^{1/2}B\right\|_* \le \left\|A_2^{1/2}B\right\|_*.$$

Lemma 7 shows that Loewner comparison of predictor metrics gives the corresponding comparison of prediction-scale Frobenius and nuclear norms.

Proof. The proof applies operator monotonicity of the matrix square root to transfer Loewner order to the spectral penalty norms.

Step 1: Frobenius norm bound. The Frobenius relation follows from

$$\left\|A_j^{1/2}B\right\|_F^2 = \mathrm{tr}(B^\top A_j B), \qquad j = 1, 2,$$

and $B^\top A_1 B \preceq B^\top A_2 B$.

Step 2: Nuclear norm bound. For the nuclear norm,

$$\left\|A_j^{1/2}B\right\|_* = \mathrm{tr}\{(B^\top A_j B)^{1/2}\}.$$

The matrix square-root map is operator monotone on the positive semidefinite cone. Hence

$$(B^\top A_1 B)^{1/2} \preceq (B^\top A_2 B)^{1/2}.$$

Taking traces proves the second relation. This completes the proof. □

The following proposition compares the matched penalty with pooled prediction in both nuclear and Frobenius norm.

Proposition 3. *For every $B \in \mathbb{R}^{p\times q}$,*

$$\sqrt{1+\tau}\left\|\widehat{\Sigma}^{1/2}B\right\|_* \le \left\|H^{1/2}B\right\|_*.$$

If $\kappa > 0$, then

$$\left\|H^{1/2}B\right\|_* \le \kappa^{-1/2}\left\|\widehat{\Sigma}^{1/2}B\right\|_*.$$

The same inequalities hold with the Frobenius norm in place of the nuclear norm.

Proof of Proposition 3. The proof establishes the required matrix inequalities by congruence before applying the Frobenius and nuclear norm comparisons.

Step 1: Metric lower bounds. Lemma 1 gives

$$H^{-1/2}\widehat{\Sigma}H^{-1/2} \preceq (1+\tau)^{-1}I_p.$$

Congruence by $H^{1/2}$ yields

$$(1+\tau)\widehat{\Sigma} \preceq H.$$

Lemma 7, applied to $(1+\tau)\widehat{\Sigma}$ and H, gives both lower inequalities in the proposition.

Step 2: Metric upper bounds. If $\kappa > 0$, then

$$\kappa I_p \preceq H^{-1/2}\widehat{\Sigma}H^{-1/2}.$$

The same congruence gives $\kappa H \preceq \widehat{\Sigma}$. Applying Lemma 7 to κH and $\widehat{\Sigma}$ proves both upper inequalities. This completes the proof. □

The metric also determines the correct dual norm for the pooled score. Let $\mathcal{R}_H(B) = \left\|H^{1/2}B\right\|_*$. This relation explains why the tuning scale in the theory and code is $\left\|H^{-1/2}G(0)\right\|_{\mathrm{op}}$.

Proposition 4. *The dual norm satisfies*

$$\mathcal{R}_H^*(A) := \sup_{\mathcal{R}_H(B)\le 1}\langle A, B\rangle = \left\|H^{-1/2}A\right\|_{\mathrm{op}}.$$

The zero matrix is a Stage I minimizer if and only if

$$\lambda_1 \ge \left\|H^{-1/2}\widehat{Z}\right\|_{\mathrm{op}}.$$

Proposition 4 connects the transformed penalty, its score bound, and the zero threshold used to construct the tuning grid.

Proof. The proof uses a change of coordinates in the trace inner product to identify the dual unit ball and applies the exact subdifferential of the nuclear norm at zero.

Step 1: Dual norm equivalence. Let $C = H^{1/2}B$. Then

$$\langle A, B\rangle = \left\langle H^{-1/2}A, C\right\rangle.$$

Nuclear and operator norms are dual, hence

$$\sup_{\|C\|_*\le 1}\left\langle H^{-1/2}A, C\right\rangle = \left\|H^{-1/2}A\right\|_{\mathrm{op}}.$$

This proves the first assertion.

Step 2: Subdifferential at zero. The optimality condition for $C = 0$ in (4) is

$$0 \in -H^{-1/2}\widehat{Z} + \lambda_1 \partial \|0\|_* .$$

Because

$$\partial \|0\|_* = \{V \in \mathbb{R}^{p\times q} : \|V\|_{\mathrm{op}} \le 1\},$$

the inclusion holds exactly under the stated inequality. This completes the proof. □

S4.2 Prediction when the pooled design is singular

Theorem 10 treats the component of the coefficient matrix that is identifiable through X. Its proof uses neither invertibility nor a restricted eigenvalue.

For $C \in \mathbb{R}^{p\times q}$, let $B_C = H^{-1/2}C$. The following theorem bounds prediction without a lower curvature condition.

Theorem 10. *Suppose $\lambda_1 \ge \varepsilon$, $\delta \ge 0$, and $F_1(C) \le F_1(C^\star) + \delta$. Then*

$$\frac{\|X(B_C - B^\star)\|_F^2}{N} \le 4\lambda_1 \|C^\star\|_* + 2\delta.$$

For a Stage I iterate with $1 \le T \le T_1$, one may let

$$\delta = \frac{2\left\|C_0 - \widehat{C}\right\|_F^2}{(T+1)^2}.$$

Proof of Theorem 10. The proof bounds the prediction error directly from the objective inequality without using strong convexity.

Step 1: Metric and score bounds. Let $\Delta = C - C^\star$. The regression model gives

$$f(C) - f(C^\star) = \frac{1}{2}\langle M\Delta, \Delta\rangle - \left\langle \widetilde{W}, \Delta \right\rangle.$$

The premise $F_1(C) \le F_1(C^\star) + \delta$ implies

$$\frac{1}{2}\langle M\Delta, \Delta\rangle \le \left\langle \widetilde{W}, \Delta \right\rangle + \lambda_1\{\|C^\star\|_* - \|C\|_*\} + \delta.$$

By operator-nuclear duality and $\lambda_1 \ge \varepsilon$,

$$\left\langle \widetilde{W}, \Delta \right\rangle \le \lambda_1 \|\Delta\|_* \le \lambda_1\{\|C\|_* + \|C^\star\|_*\}.$$

Step 2: Prediction error isolation. It follows that

$$\frac{1}{2}\langle M\Delta, \Delta\rangle \le 2\lambda_1 \|C^\star\|_* + \delta.$$

The definition of M gives

$$\langle M\Delta, \Delta\rangle = \frac{1}{N}\left\|XH^{-1/2}\Delta\right\|_F^2 = \frac{1}{N}\|X(B_C - B^\star)\|_F^2 .$$

This proves the first assertion. The accelerated objective bound in Theorem 6 gives the stated value of δ for C_T. This completes the proof. □

Small singular values contribute an approximation term to the coefficient-error bound. For each $1 \le k \le p \wedge q$, let $C_k^\star$ be a best rank-k approximation to $C^\star$ and let $a_k = \|C^\star - C_k^\star\|_*$.

Theorem 11. *Suppose $\kappa > 0$ and $\lambda_1 \ge \varepsilon$. Then, for every $1 \le k \le p \wedge q$,*

$$\left\|\widehat{C} - C^\star\right\|_F \le \frac{2\sqrt{2k}\lambda_1}{\kappa} + \sqrt{\frac{2\lambda_1 a_k}{\kappa}}.$$

Theorem 11 extends the Stage I error bound to approximately low-rank coefficients. The first term is the estimation error for the leading rank-k component, and the second term is the approximation error.

Proof. The proof uses nuclear-norm decomposability at the low-rank approximation and the subgradient inequality to separate estimation error from the approximation remainder.

Step 1: Metric and score bounds. Let $\Delta = \widehat{C} - C^\star$ and let $\mathcal{T}_k$ be the tangent space at $C_k^\star$, with orthogonal projections $\Delta_{\mathcal{T}_k}$ and $\Delta_{\mathcal{T}_k^\perp}$. Nuclear-norm decomposability and the triangle inequality give

$$\|C^\star\|_* - \|C^\star + \Delta\|_* \le \|\Delta_{\mathcal{T}_k}\|_* - \left\|\Delta_{\mathcal{T}_k^\perp}\right\|_* + 2a_k.$$

The optimality equation in the proof of Theorem 6, $\lambda_1 \ge \varepsilon$, and $\operatorname{rank}(\Delta_{\mathcal{T}_k}) \le 2k$ imply

$$\kappa \|\Delta\|_F^2 \le 2\sqrt{2k}\lambda_1 \|\Delta\|_F + 2\lambda_1 a_k.$$

Step 2: Estimation error isolation. For nonnegative x, u, v, $x^2 \le ux + v$ implies $x \le u + \sqrt{v}$. Apply this relation with $x = \|\Delta\|_F$, $u = 2\sqrt{2k}\lambda_1/\kappa$, and $v = 2\lambda_1 a_k/\kappa$. This completes the proof. □

The metric scale converts transformed coefficient bounds to the original coefficient matrix. It also converts the rank signal condition to the singular values of $B^\star$.

Lemma 8. *For every $A \in \mathbb{R}^{p\times q}$ and every $j \le p \wedge q$,*

$$\sqrt{\gamma}\,\sigma_j(A) \le \sigma_j(H^{1/2}A) \le \sqrt{\gamma + \ell_1}\,\sigma_j(A).$$

In particular,

$$\left\|H^{-1/2}A\right\|_F \le \gamma^{-1/2}\|A\|_F.$$

Lemma 8 expresses statistical conclusions in the original coefficient scale. Local rank deficiency does not appear because $\gamma > 0$ controls every predictor direction.

Proof. The proof relies on the min-max characterization of singular values and the bounds on the modified metric.

Step 1: Rayleigh quotient inequalities. The metric satisfies

$$\gamma I_p \preceq H \preceq (\gamma + \ell_1) I_p.$$

For every vector $v \in \mathbb{R}^q$,

$$\gamma \|Av\|_2^2 \le \left\|H^{1/2}Av\right\|_2^2 \le (\gamma + \ell_1)\|Av\|_2^2.$$

Step 2: Singular value bounds. The min-max characterization of singular values gives the first two inequalities. The last relation follows from $H^{-1} \preceq \gamma^{-1} I_p$ and a trace calculation. This completes the proof. □

Combining Lemma 8 with Theorem 6 shows directly that, if $\kappa > 0$ and $\lambda_1 \ge \varepsilon$, then

$$\left\|\widehat{B}_1 - B^\star\right\|_F \le \frac{2\sqrt{2r}\lambda_1}{\kappa\sqrt{\gamma}}.$$

When Theorem 11 is applied at approximation rank k, replace r by k in the preceding bound and add $\{2\lambda_1 a_k/(\kappa\gamma)\}^{1/2}$. These bounds follow by applying the lemma to $A = \widehat{C} - C^\star$ and using $\widehat{B}_1 - B^\star = H^{-1/2}(\widehat{C} - C^\star)$. They show how stabilization changes coefficient scaling. A larger γ reduces the conversion factor, while it may also reduce κ and change the transformed signal.

S4.3 MINIMAX PREDICTION BENCHMARK

The proof of Proposition 5 reduces multivariate regression to Gaussian matrix denoising on the column space of the design. The next lemma gives the required finite packing.

Lemma 9. *For every pair of integers $u, v \geq 1$, there is a set $\mathcal{V} \subset \{0,1\}^{u \times v}$ such that*

$$\log |\mathcal{V}| \geq c_0 uv, \qquad \|V - V'\|_F^2 \geq c_1 uv$$

for distinct $V, V' \in \mathcal{V}$, where $c_0, c_1 > 0$ are absolute constants.

Lemma 9 gives a parameter set with separation proportional to its number of free entries. It is used with Fano's inequality in the proof of Proposition 5.

Proof. The proof constructs the packing set using the Varshamov–Gilbert bound on binary matrices.

Step 1: Set construction. Identify binary $u \times v$ matrices with vectors of length uv. For $uv \geq 8$, the Varshamov–Gilbert construction gives at least $\exp(c_0 uv)$ vectors separated by Hamming distance $c_1 uv$. For $uv < 8$, use the two constant matrices and decrease c_0, c_1.

Step 2: Distance metric. Squared Frobenius distance equals Hamming distance. This completes the proof. □

The following proposition establishes the fixed-design minimax prediction benchmark under Gaussian noise.

Proposition 5. *Let E have independent $N(0, \sigma^2)$ entries, and let $1 \leq r \leq d \wedge q$. For an absolute constant $c > 0$,*

$$\inf_{\widetilde{B}} \sup_{\operatorname{rank}(B) \leq r} \mathbb{E}_B \frac{\left\| X(\widetilde{B} - B) \right\|_F^2}{N} \geq c\sigma^2 \frac{r(d + q - r)}{N}.$$

Proof of Proposition 5. The proof reduces multivariate regression to Gaussian matrix denoising on the column space of the design and uses Fano's inequality with a binary packing set.

Step 1: Reduction to Gaussian denoising. Let $Q \in \mathbb{R}^{N \times d}$ have orthonormal columns spanning the column space of X. Let

$$A = Q^\top X B, \qquad W = Q^\top Y = A + Q^\top E.$$

The entries of $Q^\top E$ are independent $N(0, \sigma^2)$. Because $Q^\top X$ has row rank d, every $A \in \mathbb{R}^{d \times q}$ equals $Q^\top X B$ for some B with $\operatorname{rank}(B) \leq \operatorname{rank}(A)$. Also,

$$\left\| X(\widetilde{B} - B) \right\|_F = \left\| Q^\top X(\widetilde{B} - B) \right\|_F.$$

The component $(I_N - QQ^\top)Y$ is independent of W and its distribution does not depend on A. It suffices to bound the risk in

$$W = A + Z, \qquad Z_{ij} \overset{\text{ind}}{\sim} N(0, \sigma^2).$$

For $A = te_1 e_1^\top$ with coordinate vectors in $\mathbb{R}^d$ and $\mathbb{R}^q$, the prior $t \sim N(0, v^2)$ has Bayes risk $\sigma^2 v^2/(\sigma^2 + v^2) \to \sigma^2$ as $v \to \infty$. This gives the required order when rq or rd is bounded by an absolute constant.

Step 2: Fano bound for the reduced problem. For larger rq, apply Lemma 9 with $u = r, v = q$. For $c_2 > 0$, let

$$A_V = c_2 \sigma \begin{pmatrix} V \\ 0 \end{pmatrix}, \qquad V \in \mathcal{V}.$$

Each A_V has rank at most r. Distinct parameters have squared distance at least

$$c_1 c_2^2 \sigma^2 rq.$$

The Kullback–Leibler divergence from the distribution at zero is

$$\frac{\|A_V\|_F^2}{2\sigma^2} \leq \frac{c_2^2 rq}{2}.$$

Choose c_2 to bound this by $\log|\mathcal{V}|/8$. For rq large enough to absorb the $\log 2$ term, Fano's inequality (Tsybakov, 2009) gives

$$\inf_{\widetilde{A}} \sup_{\operatorname{rank}(A)\le r} \mathbb{E}_A \left\|\widetilde{A} - A\right\|_F^2 \ge c_3\sigma^2 rq.$$

Using the first r columns and the scalar bound for bounded rd gives

$$\inf_{\widetilde{A}} \sup_{\operatorname{rank}(A)\le r} \mathbb{E}_A \left\|\widetilde{A} - A\right\|_F^2 \ge c_3\sigma^2 rd.$$

The larger bound is at least

$$\frac{c_3}{2}\sigma^2 r(d+q) \ge \frac{c_3}{2}\sigma^2 r(d+q-r).$$

Divide by N and use the regression reduction. This completes the proof. □

S4.4 Finite-iteration stationarity and full-sequence convergence

We first bound a subgradient of the Stage II objective by the step between two computed iterates. This relative-error relation and Theorem 2 give the finite stationarity rate.

Lemma 10. *For every Stage II iteration,*

$$\operatorname{dist}\{0, \partial F_2(C_{t+1})\} \le \|C_{t+1} - C_t\|_F .$$

Lemma 10 bounds a stationarity measure by the computed step between consecutive iterates. No oracle point or unobserved selection event enters the bound.

Proof. The proof bounds a stationarity measure directly using the exact proximal optimality condition and the Lipschitz continuity of the gradient.

Step 1: Proximal optimality equation. Let $\Delta_t = C_{t+1} - C_t$. Optimality of the exact proximal update gives a matrix $V_{t+1} \in \partial P_2(C_{t+1})$ such that

$$0 = \Delta_t + \nabla f(C_t) + V_{t+1}.$$

Hence

$$\nabla f(C_{t+1}) + V_{t+1} = (M - I_p)\Delta_t.$$

Step 2: Subdifferential distance bound. Lemma 1 gives $0 \preceq M \preceq (1+\tau)^{-1} I_p$ and hence $\|I_p - M\|_{\mathrm{op}} \le 1$. It follows that

$$\|\nabla f(C_{t+1}) + V_{t+1}\|_F \le \|\Delta_t\|_F .$$

Taking the distance to the limiting subdifferential proves the lemma. This completes the proof. □

Proof of Corollary 2. The proof bounds stationarity error by the average squared norm of the iterate differences and uses the Kurdyka–Łojasiewicz property for full-sequence convergence.

Step 1: Finite-iteration stationarity bound. Summing Theorem 2 over $t = T_1, \dots, T_1 + K - 1$ gives

$$\frac{\tau}{2(1+\tau)} \sum_{t=T_1}^{T_1+K-1} \|C_{t+1} - C_t\|_F^2 \le F_2(C_{T_1}) - \inf_C F_2(C).$$

Lemma 10 yields

$$\min_{T_1 \le t < T_1+K} \operatorname{dist}^2\{0, \partial F_2(C_{t+1})\} \le \frac{1}{K} \sum_{t=T_1}^{T_1+K-1} \|C_{t+1} - C_t\|_F^2 ,$$

which proves the finite-iteration bound.

Step 2: Full-sequence convergence. Lemma 4 gives boundedness without $\kappa > 0$. The objective is continuous, proper, and semi-algebraic by Corollary 18. Theorem 2 and Lemma 10 verify H1–H2 of Attouch et al. (2013, Theorem 2.9); continuity along a convergent subsequence verifies H3. That theorem gives finite length and convergence to one stationary point. This completes the proof. □

The stationarity bound can be written directly in terms of the proximal residual produced by the code. This gives a stopping rule that does not require evaluation of a spectral subdifferential.

Let prox_{P_2} denote the exact spectral proximal map in Lemma 2. Define the proximal residual by

$$\mathcal{G}_2(C) = C - \mathrm{prox}_{P_2}\{C - \nabla f(C)\}.$$

Proposition 6. *Then $\mathcal{G}_2(C_t) = C_t - C_{t+1}$ and, for every integer $K \geq 1$,*

$$\min_{T_1 \leq t < T_1 + K} \|\mathcal{G}_2(C_t)\|_F^2 \leq \frac{2(1+\tau)}{\tau K}\{F_2(C_{T_1}) - \inf_C F_2(C)\}.$$

Proposition 6 identifies the exact quantity that can be monitored after each communication round. A small iterate difference is a valid stationarity diagnostic under the fixed metric.

Proof. The proof relies on the definition of the update and the descent bound to control the minimum proximal residual.

Step 1: Iterate identities. The first identity is the definition of the Stage II update. Summing the descent inequality exactly as in the proof of Corollary 2 gives

$$\sum_{t=T_1}^{T_1+K-1} \|\mathcal{G}_2(C_t)\|_F^2 \leq \frac{2(1+\tau)}{\tau}\{F_2(C_{T_1}) - \inf_C F_2(C)\}.$$

Step 2: Minimum over iterations. The minimum is bounded by the average. This completes the proof. □

S4.5 MCP OBJECTIVE AND RANK IDENTIFICATION

The MCP proximal map has the same zero region as SCAD and leaves singular values above $b\lambda_2$ unchanged. These two facts are sufficient for the corresponding rank result.

The following corollary extends descent, convergence, and rank recovery to the MCP update.

Corollary 4. *For MCP with $b > 1$, Theorem 2 and Corollary 2 hold for the MCP objective. If $1 \leq r < p \wedge q$ and*

$$\eta_t < \lambda_2, \qquad \sigma_r(C^\star) > b\lambda_2 + \eta_t,$$

then $\mathrm{rank}(C_{t+1}) = r$ and $\|C_{t+1} - C^\star\|_F \leq 2\sqrt{2r}\eta_t$.

Proof of Corollary 4. The proof verifies that the exact MCP proximal update preserves the same singular-value separation properties used for the SCAD penalty.

Step 1: Stationarity. By Lemma 2, the MCP update is an exact proximal minimizer. The proofs of Theorem 2 and Lemma 10 use only this property and the Lipschitz bound for ∇f, hence the finite-round bound in Corollary 2 applies. Lemma 4 and Corollary 18 are stated for MCP as well, hence the full-sequence convergence in Corollary 2 also holds.

Step 2: Signal separation and rank identification. Under the rank premises, Weyl's inequality gives

$$\sigma_{r+1}(Z_t) \leq \eta_t < \lambda_2, \qquad \sigma_r(Z_t) \geq \sigma_r(C^\star) - \eta_t > b\lambda_2.$$

The MCP map keeps the first r singular values and sets the others to zero. The error argument in Theorem 4 gives

$$\mathrm{rank}(C_{t+1}) = r, \qquad \|C_{t+1} - C^\star\|_F \leq 2\sqrt{2r}\eta_t.$$

This completes the proof. □

The signal condition can be stated in the scale of the original coefficient matrix. Lemma 8 gives

$$\sigma_r(C^\star) = \sigma_r(H^{1/2}B^\star) \geq \sqrt{\gamma}\,\sigma_r(B^\star).$$

Hence, if $\eta_t < \lambda_2$ and

$$\sqrt{\gamma}\,\sigma_r(B^\star) > a\lambda_2 + \eta_t,$$

then the SCAD update satisfies the conclusions of Theorem 4. Replacing a by b gives the MCP conclusion. This sufficient beta-min statement is stronger than the transformed condition when the lower singular-value bound in Lemma 8 is not tight.

S4.6 Exact recovery of worker sufficient statistics

We make the query matrices in Proposition 7 explicit. Let I_k contain the indices $(k-1)q+1$ through kq, truncated at p, and let $E_k \in \mathbb{R}^{p\times q}$ place the corresponding coordinate vectors in its first $|I_k|$ columns.

Let $K = \lceil p/q \rceil$ and set the remaining columns of each E_k to zero. The following proposition establishes exact recovery of the worker summaries.

Proposition 7. *The values of $G_s(0)$ and $G_s(E_k)$ for $k = 1, \dots, K$ determine $\widehat{Z}_s$ and $\widehat{\Sigma}_s$ exactly.*

Proof of Proposition 7. The proof extracts the worker sufficient statistics by querying the gradient at zero and at the partitioned coordinate blocks.

Step 1: Recovery of the cross-product matrix. The worker gradient is $G_s(B) = \widehat{\Sigma}_s B - \widehat{Z}_s$. The zero query gives

$$\widehat{Z}_s = -G_s(0).$$

Step 2: Recovery of the Gram matrix. For $k = 1, \dots, K$,

$$G_s(E_k) - G_s(0) = \widehat{\Sigma}_s E_k.$$

The nonzero columns are those of $\widehat{\Sigma}_s$ indexed by I_k. Because $I_1, \dots, I_K$ partition $\{1, \dots, p\}$, the differences recover the Gram matrix. This completes the proof. □

The preceding construction uses one zero query and $\lceil p/q \rceil$ coordinate-block queries. It is an identification statement and does not improve the scalar comparison in Proposition 1.

Proposition 8. *For every worker s and every B,*

$$\mathcal{L}_s(B) = \frac{\|Y_s\|_F^2}{2n_s} - \left\langle \widehat{Z}_s, B \right\rangle + \frac{1}{2} \left\langle \widehat{\Sigma}_s B, B \right\rangle.$$

Thus $\widehat{\Sigma}_s$ and $\widehat{Z}_s$ determine the local objective up to an additive constant.

Proposition 8 states why sufficient-statistic aggregation is the exact batch benchmark. The coefficient-dependent quadratic objective is available after these matrices are released.

Proof. The proof expands the local least-squares objective directly in terms of the sufficient statistics.

Step 1: Objective expansion. Expansion gives

$$\begin{aligned} \|Y_s - X_s B\|_F^2 &= \|Y_s\|_F^2 - 2\mathrm{tr}(Y_s^\top X_s B) + \mathrm{tr}(B^\top X_s^\top X_s B) \\ &= \|Y_s\|_F^2 - 2n_s \left\langle \widehat{Z}_s, B \right\rangle + n_s \left\langle \widehat{\Sigma}_s B, B \right\rangle. \end{aligned}$$

Step 2: Constant and quadratic terms. Division by $2n_s$ proves the identity. This completes the proof. □

S5 Perturbed-gradient updates

The main manuscript uses exact gradients. This section quantifies deterministic perturbations without assigning them a privacy, delay, or quantization mechanism. Let $\Xi_t \in \mathbb{R}^{p\times q}$ be an error in the transformed gradient and define

$$C_{t+1} \in \arg\min_C \left[\langle \nabla f(C_t) + \Xi_t, C - C_t \rangle + \frac{1}{2} \|C - C_t\|_F^2 + P_2(C) \right].$$

Let

$$c_\tau = \frac{\tau}{2(1+\tau)}, \qquad \Delta_t = C_{t+1} - C_t.$$

The following theorem gives descent up to the perturbation energy and a corresponding stationarity bound.

Theorem 12. *For every perturbed update,*

$$F_2(C_t) - F_2(C_{t+1}) \geq \frac{c_\tau}{2} \|\Delta_t\|_F^2 - \frac{1}{2c_\tau} \|\Xi_t\|_F^2 ,$$

and

$$\operatorname{dist}\{0, \partial F_2(C_{t+1})\} \leq \|\Delta_t\|_F + \|\Xi_t\|_F .$$

If $\sum_t \|\Xi_t\|_F^2 < \infty$*, then* $\sum_t \|\Delta_t\|_F^2 < \infty$*, and every cluster point is stationary.*

Theorem 12 shows which part of the exact-gradient result is stable under numerical or communication error. Square-summable transformed errors preserve asymptotic stationarity, while persistent errors produce an explicit residual term.

Proof. The proof uses Young's inequality to bound the perturbation term in the descent inequality and proximal optimality to bound stationarity error.

Step 1: Proximal optimality and smooth descent. Proximal optimality relative to C_t gives

$$\langle \nabla f(C_t) + \Xi_t, \Delta_t \rangle + \frac{1}{2} \|\Delta_t\|_F^2 + P_2(C_{t+1}) - P_2(C_t) \leq 0.$$

The smooth part satisfies

$$f(C_{t+1}) - f(C_t) \leq \langle \nabla f(C_t), \Delta_t \rangle + \frac{1}{2(1+\tau)} \|\Delta_t\|_F^2 .$$

Step 2: Descent inequality. Adding the two inequalities yields

$$F_2(C_t) - F_2(C_{t+1}) \geq c_\tau \|\Delta_t\|_F^2 + \langle \Xi_t, \Delta_t \rangle .$$

Young's inequality gives

$$\langle \Xi_t, \Delta_t \rangle \geq -\frac{c_\tau}{2} \|\Delta_t\|_F^2 - \frac{1}{2c_\tau} \|\Xi_t\|_F^2 ,$$

Step 3: Stationarity bound. There is $V_{t+1} \in \partial P_2(C_{t+1})$ such that

$$0 = \Delta_t + \nabla f(C_t) + \Xi_t + V_{t+1}.$$

It follows that

$$\begin{aligned} \|\nabla f(C_{t+1}) + V_{t+1}\|_F &= \|(M - I_p)\Delta_t - \Xi_t\|_F \\ &\leq \|\Delta_t\|_F + \|\Xi_t\|_F . \end{aligned}$$

Step 4: Stationarity of cluster points. Summing the descent bound and using the lower bound for F_2 gives

$$\frac{c_\tau}{2} \sum_t \|\Delta_t\|_F^2 \leq F_2(C_{T_1}) - \inf_C F_2(C) + \frac{1}{2c_\tau} \sum_t \|\Xi_t\|_F^2 .$$

Square summability gives $\Delta_t \to 0$ and $\Xi_t \to 0$. Along a convergent subsequence, the optimality relation and closedness of the limiting subdifferential give stationarity. This completes the proof. □

The next corollary gives a finite bound under nonzero perturbations. It separates optimization error from accumulated gradient error.

For $K \geq 1$, let

$$\mathcal{E}_K = \sum_{t=T_1}^{T_1+K-1} \|\Xi_t\|_F^2 .$$

Let $C_\tau = 2 + 8(1+\tau)^2/\tau^2$.

Corollary 5. *For every integer* $K \geq 1$*,*

$$\min_{T_1 \leq t < T_1+K} \operatorname{dist}^2\{0, \partial F_2(C_{t+1})\} \leq \frac{8(1+\tau)}{\tau K}\{F_2(C_{T_1}) - \inf_C F_2(C)\} + \frac{C_\tau \mathcal{E}_K}{K},$$

Corollary 5 shows that the average transformed error energy determines the finite perturbation term. If $\mathcal{E}_K = O(1)$, the same stationarity rate holds for the perturbed updates.

Proof. The proof bounds the sum of squared iterate differences, converts it to a bound for squared stationarity errors, and bounds the minimum by their average.

Step 1: Sum of squared iterate differences. Theorem 12 gives

$$\sum_{t=T_1}^{T_1+K-1} \|\Delta_t\|_F^2 \le \frac{2}{c_\tau}\{F_2(C_{T_1}) - \inf_C F_2(C)\} + \frac{1}{c_\tau^2}\mathcal{E}_K.$$

Step 2: Stationarity error bound. The stationarity inequality and $(x+y)^2 \le 2x^2 + 2y^2$ give

$$\sum_{t=T_1}^{T_1+K-1} \mathrm{dist}^2\{0, \partial F_2(C_{t+1})\} \le 2\sum_t \|\Delta_t\|_F^2 + 2\mathcal{E}_K.$$

Substitute $c_\tau = \tau/\{2(1+\tau)\}$ and divide by K. This completes the proof. □

Gradient perturbations also enter the rank condition through one operator-norm term. This gives a direct tolerance for approximate aggregation.

Let the perturbed update use

$$Z_t^\Xi = C_t - \nabla f(C_t) - \Xi_t.$$

Proposition 9. *Let* $1 \le r < p \wedge q$. *If*

$$\eta_t + \|\Xi_t\|_{\mathrm{op}} < \lambda_2, \qquad \sigma_r(C^\star) > a\lambda_2 + \eta_t + \|\Xi_t\|_{\mathrm{op}},$$

then the SCAD update has rank r *and*

$$\|C_{t+1} - C^\star\|_F \le 2\sqrt{2r}\{\eta_t + \|\Xi_t\|_{\mathrm{op}}\}.$$

Proposition 9 quantifies the margin required beyond the exact-gradient certificate. The bound concerns the transformed aggregation error received by the master.

Proof. The proof applies Weyl's inequality to the perturbed target matrix using the operator norm of the perturbation.

Step 1: Proximal target error. The update matrix satisfies

$$Z_t^\Xi = C^\star + R_t - \Xi_t.$$

Hence

$$\left\|Z_t^\Xi - C^\star\right\|_{\mathrm{op}} \le \eta_t + \|\Xi_t\|_{\mathrm{op}}.$$

Step 2: Rank identification. Weyl's inequality and the SCAD threshold regions give rank r. The proof of Theorem 4, with R_t replaced by $R_t - \Xi_t$, gives the stated Frobenius bound. This completes the proof. □

S6 Exact thresholding with a truncated singular value decomposition

The full singular value decomposition is unnecessary when all singular values above the zero threshold have been computed and the remaining spectral norm is certified. This section states the exact condition; it does not treat an uncertified approximate decomposition.

Let $Z \in \mathbb{R}^{p\times q}$ have singular value decomposition $Z = U\mathrm{diag}(d)V^\top$. For $1 \le k < p \wedge q$, let $U_k\mathrm{diag}(d_1,\ldots,d_k)V_k^\top$ contain its leading k singular components. The following proposition applies to soft thresholding, SCAD, and MCP.

Proposition 10. *Let $\theta : [0,\infty) \to [0,\infty)$ satisfy $\theta(z) = 0$ for $0 \le z \le \lambda$. If $d_{k+1} \le \lambda$, then*

$$U\mathrm{diag}\{\theta(d)\}V^\top = U_k\mathrm{diag}\{\theta(d_1),\dots,\theta(d_k)\}V_k^\top.$$

Proposition 10 gives an exact stopping certificate for a partial decomposition. The next singular value must be bounded by the relevant penalty threshold.

Proof. The proof relies on the fact that the thresholding function sends singular values at or below the threshold to zero, isolating the computed leading components.

Step 1: Thresholding property. For $j > k$, $d_j \le d_{k+1} \le \lambda$, hence $\theta(d_j) = 0$.

Step 2: Map truncation. The full spectral map contains only the first k terms, which proves the identity. This completes the proof. □

The same statement can be written using a residual norm produced by a Krylov method. This form is useful because it avoids computing d_{k+1} explicitly.

Corollary 6. *Let U_k and V_k have orthonormal columns, and let $d_1 \ge \cdots \ge d_k \ge 0$. Suppose*

$$Z = U_k\mathrm{diag}(d_1,\dots,d_k)V_k^\top + R_k^{\mathrm{svd}}, \qquad U_k^\top R_k^{\mathrm{svd}} = 0, \qquad R_k^{\mathrm{svd}}V_k = 0, \qquad \left\|R_k^{\mathrm{svd}}\right\|_{\mathrm{op}} \le \lambda.$$

Then the thresholded matrix formed from $(U_k, \mathrm{diag}(d_1,\dots,d_k), V_k)$ equals the full spectral proximal update.

Corollary 6 separates numerical approximation from certification. An approximate decomposition is exact for the estimator only after its orthogonal remainder is bounded by the threshold.

Proof. The proof uses the orthogonality conditions to obtain a block-diagonal representation, with all singular values of the remainder at or below the threshold.

Step 1: Block diagonalization. The orthogonality relations give a block-diagonal representation whose singular values are those in $\mathrm{diag}(d_1,\dots,d_k)$ and R_k^{svd}.

Step 2: Remainder thresholding. Every singular value of the remainder is at most λ, and the spectral map sends the remainder to zero. This completes the proof. □

The threshold certificate also determines the rank directly.

Proposition 11. *Under all conditions of Corollary 6, suppose, for the SCAD map,*

$$d_k > a\lambda.$$

Then the computed proximal update has rank k, keeps $d_1,\dots,d_k$, and equals the full update. For MCP, replace a by b.

Proposition 11 combines computational termination and rank certification in one check. It concerns the singular values of the computed update matrix and does not assume an oracle selection rule.

Proof. The proof combines the separation of the leading singular values with the exact truncation threshold to certify the rank of the computed update and its equality to the full update.

Step 1: Leading components. The first k singular values exceed the upper knot and remain unchanged.

Step 2: Truncated components. Corollary 6 maps the remaining values to zero. This completes the proof. □

S7 Computational form of the stabilized metric

This section derives the matrix actions used by the implementation. The formulas show that local rank deficiency reduces the size of the master decomposition and never requires a dense inverse.

Let

$$\frac{X_1}{\sqrt{n_1}} = U_1 S_1 V_1^\top, \qquad r_1 = \mathrm{rank}(X_1),$$

be a thin singular value decomposition with positive diagonal entries in S_1. The following proposition gives every real power of H required by the method.

Proposition 12. *For $\alpha \in \mathbb{R}$ and $A \in \mathbb{R}^{p\times q}$,*

$$H^\alpha A = \gamma^\alpha A + V_1 \left\{ (S_1^2 + \gamma I_{r_1})^\alpha - \gamma^\alpha I_{r_1} \right\} V_1^\top A.$$

Proposition 12 gives the square-root and inverse-square-root operations used in every reported proposed fit with the local metric. Their storage depends on $r_1 p$, not p^2.

Proof. The proof uses the thin singular value decomposition of the scaled design to explicitly apply functional calculus to the block structure of the metric.

Step 1: Subspace decomposition. The master Gram matrix is

$$\widehat{\Sigma}_1 = V_1 S_1^2 V_1^\top.$$

Complete V_1 to an orthogonal matrix $[V_1, V_{1\perp}]$. Then

$$H = V_1(S_1^2 + \gamma I_{r_1})V_1^\top + \gamma V_{1\perp} V_{1\perp}^\top.$$

Step 2: Functional calculus. Functional calculus gives

$$H^\alpha = V_1(S_1^2 + \gamma I_{r_1})^\alpha V_1^\top + \gamma^\alpha V_{1\perp} V_{1\perp}^\top.$$

Use $V_{1\perp} V_{1\perp}^\top = I_p - V_1 V_1^\top$ and multiply by A. This completes the proof. □

The next proposition records arithmetic and storage costs under dense linear algebra. It distinguishes worker gradient evaluation, master metric actions, and spectral thresholding.

Let $r_1 = \operatorname{rank}(X_1)$ and let k be the block size used at a threshold-certified update.

Proposition 13. *One round has worker cost $O(n_s pq)$ at worker s, master aggregation cost $O(mpq)$, master metric-action cost $O(pq + pr_1 q)$, and full spectral cost $O\{pq(p \wedge q)\}$. A block Krylov computation using J matrix–block passes has matrix–block multiplication cost $O(Jpqk)$, with separate costs for orthogonalization and the projected decomposition. The dense implementation uses master storage $O(pr_1 + pq)$ apart from the local observations when worker gradients are accumulated sequentially.*

Proposition 13 identifies the operations that can dominate in different regimes. The in-process experiments cache each worker covariance and cross-product. This has setup cost $O(n_s p^2 + n_s pq)$, per-round worker cost $O(p^2 q)$, and worker storage $O(p^2 + pq)$. Their current master reduction can hold m gradient matrices temporarily and use $O(mpq)$ memory; sequential accumulation gives the bound in the proposition. These choices change local computation and memory but not the communicated matrices. The socket experiment uses the rowwise gradient calculation stated in the proposition.

Proof. The proof counts the floating-point operations for each linear algebra primitive based on the dimensions of the local data and the shared coefficient iterate.

Step 1: Gradient evaluation and aggregation. The products $X_s B$ and $X_s^\top (X_s B - Y_s)$ cost $O(n_s pq)$ operations; gradient aggregation costs $O(mpq)$.

Step 2: Master matrix actions. Proposition 12 uses scaling and multiplication by $V_1^\top, V_1$, costing $O(pq + pr_1 q)$. Dense singular value decomposition costs $O\{pq(p \wedge q)\}$. In J block Krylov passes of size k, multiplication by $Z, Z^\top$ costs $O(Jpqk)$; orthogonalization and the projected decomposition have separate implementation-dependent costs.

Step 3: Storage. Storing V_1 and a transformed coefficient costs $O(pr_1 + pq)$. This completes the proof. □

The transformed smooth objective has an exact quadratic representation. This identity is useful for checking gradients and objective values without reference to an iterative surrogate.

Lemma 11. *Up to a constant independent of C,*

$$f(C) = \frac{1}{2}\langle MC, C\rangle - \left\langle H^{-1/2}\widehat{Z}, C\right\rangle, \qquad \nabla f(C) = MC - H^{-1/2}\widehat{Z}.$$

Lemma 11 proves that Stage I is accelerated proximal gradient and Stage II is ordinary proximal gradient in the transformed coordinate. The method's distinct statistical object is the spectral penalty defined in that same coordinate.

Proof. The proof applies the symmetry of the metric inverse square root to transform the pooled quadratic objective.

Step 1: Metric substitution. Substitute $B = H^{-1/2}C$ into the quadratic expansion

$$\mathcal{L}(B) = \frac{1}{2}\left\langle \widehat{\Sigma}B, B\right\rangle - \left\langle \widehat{Z}, B\right\rangle + \text{const.}$$

Step 2: Symmetry relations. Symmetry of $H^{-1/2}$ gives

$$\left\langle \widehat{\Sigma}H^{-1/2}C, H^{-1/2}C\right\rangle = \langle MC, C\rangle, \quad \left\langle \widehat{Z}, H^{-1/2}C\right\rangle = \left\langle H^{-1/2}\widehat{Z}, C\right\rangle.$$

The gradient follows by differentiating the quadratic objective in C. This completes the proof. □

S8 Low-rank geometry used in the proofs

This section gives the tangent-space and perturbation facts used by the statistical bounds. The results are stated for the transformed coefficient because that is the parameter penalized by the method.

Let

$$C^\star = U^\star D^\star (V^\star)^\top$$

be a compact singular value decomposition of rank r. Define

$$\mathcal{T} = \{U^\star A^\top + G(V^\star)^\top : A \in \mathbb{R}^{q\times r},\ G \in \mathbb{R}^{p\times r}\}.$$

The next lemma gives the two orthogonal projections and their ranks. The next two lemmas use only this decomposition, hence they apply verbatim to any matrix of rank $r_0 \geq 0$ in place of $C^\star$, with r_0 in place of r and $\mathcal{T} = \{0\}$ when $r_0 = 0$; Theorems 1 and 5 apply them at $C^\star_X$ and at $\Pi_X H^{1/2}B$.

Lemma 12. *For every $A \in \mathbb{R}^{p\times q}$, with $P_U = U^\star(U^\star)^\top$ and $P_V = V^\star(V^\star)^\top$,*

$$\mathcal{P}_{\mathcal{T}}(A) = P_U A + AP_V - P_U AP_V, \quad \mathcal{P}_{\mathcal{T}^\perp}(A) = (I_p - P_U)A(I_q - P_V).$$

Every matrix in $\mathcal{T}$ has rank at most $2r$.

Lemma 12 isolates changes in the true row or column singular spaces. This is the effective low-dimensional component in the Stage I error bound.

Proof. The proof decomposes an arbitrary matrix into orthogonal components aligned with the true singular spaces.

Step 1: Orthogonal decomposition. Decompose A into the four blocks induced by P_U and P_V:

$$A = P_U AP_V + P_U A(I_q - P_V) + (I_p - P_U)AP_V + (I_p - P_U)A(I_q - P_V).$$

Step 2: Projection identification. The first three blocks belong to $\mathcal{T}$ and the last to $\mathcal{T}^\perp$, giving the projection. Each element of $\mathcal{T}$ is a sum of two matrices of rank at most r, hence has rank at most $2r$. This completes the proof. □

The nuclear norm is decomposable across the true singular spaces. The following inequality is the precise form used in the basic inequality.

Lemma 13. *For every $A \in \mathbb{R}^{p\times q}$,*

$$\|C^\star + A\|_* - \|C^\star\|_* \geq \|\mathcal{P}_{\mathcal{T}^\perp}(A)\|_* - \|\mathcal{P}_{\mathcal{T}}(A)\|_* .$$

Lemma 13 separates error in the true singular spaces from the remaining component. This distinction gives the rank-dependent terms in the Stage I bound.

Proof. The proof isolates the complementary tangent space component using the block-diagonal structure of the nuclear norm.

Step 1: Orthogonal blocks. Let $A_{\mathcal{T}} = \mathcal{P}_{\mathcal{T}}(A)$ and $A_{\mathcal{T}^\perp} = \mathcal{P}_{\mathcal{T}^\perp}(A)$. The matrices $C^\star$ and $A_{\mathcal{T}^\perp}$ have orthogonal row and column spaces, hence

$$\|C^\star + A_{\mathcal{T}^\perp}\|_* = \|C^\star\|_* + \|A_{\mathcal{T}^\perp}\|_* .$$

Step 2: Triangle inequality. The triangle inequality gives

$$\begin{aligned}\|C^\star + A\|_* &= \|C^\star + A_{\mathcal{T}^\perp} + A_{\mathcal{T}}\|_* \\ &\geq \|C^\star + A_{\mathcal{T}^\perp}\|_* - \|A_{\mathcal{T}}\|_* .\end{aligned}$$

Substitution proves the lemma. This completes the proof. □

The next lemma bounds the Stage I error outside the true tangent space by its component within that space. Its penalty condition is stronger than that of Theorem 6.

Let $\Delta = \widehat{C} - C^\star$.

Lemma 14. *If $\lambda_1 > 0$ and $\lambda_1 \geq 2\varepsilon$, then*

$$\|\mathcal{P}_{\mathcal{T}^\perp}(\Delta)\|_* \leq 3\,\|\mathcal{P}_{\mathcal{T}}(\Delta)\|_* , \qquad \|\Delta\|_* \leq 4\sqrt{2r}\,\|\Delta\|_F .$$

Lemma 14 follows from the penalized optimization problem. The nuclear-norm bound depends on the signal rank and does not require the convex estimate to have that rank.

Proof. The proof applies the basic inequality and decomposability to bound the component outside the tangent space.

Step 1: Basic inequality. The basic inequality and nonnegativity of the quadratic term give

$$0 \leq \left\langle \widetilde{W}, \Delta \right\rangle + \lambda_1\{\|C^\star\|_* - \|C^\star + \Delta\|_*\}.$$

By duality, the score term is at most

$$\frac{\lambda_1}{2}\left\{\|\mathcal{P}_{\mathcal{T}}(\Delta)\|_* + \|\mathcal{P}_{\mathcal{T}^\perp}(\Delta)\|_*\right\}.$$

Step 2: Cone condition. Lemma 13 then gives

$$\frac{1}{2}\,\|\mathcal{P}_{\mathcal{T}^\perp}(\Delta)\|_* \leq \frac{3}{2}\,\|\mathcal{P}_{\mathcal{T}}(\Delta)\|_* .$$

This proves the cone relation.

Step 3: Nuclear norm bound. Lemma 12 and the rank-norm inequality give

$$\begin{aligned}\|\Delta\|_* &\leq 4\,\|\mathcal{P}_{\mathcal{T}}(\Delta)\|_* \leq 4\sqrt{2r}\,\|\mathcal{P}_{\mathcal{T}}(\Delta)\|_F \\ &\leq 4\sqrt{2r}\,\|\Delta\|_F .\end{aligned}$$

This completes the proof. □

The rank proof uses singular-value perturbation and a rank bound for differences. We state both facts in one lemma.

Lemma 15. *For matrices A and B of the same dimensions,*

$$|\sigma_j(A) - \sigma_j(B)| \le \|A - B\|_{\mathrm{op}}$$

for every j. If $\mathrm{rank}(A) \le r$ and $\mathrm{rank}(B) \le r$, then

$$\|A - B\|_F \le \sqrt{2r}\, \|A - B\|_{\mathrm{op}}.$$

Lemma 15 converts operator control of the update matrix into rank and Frobenius conclusions. These are the two steps used in Theorem 4.

Proof. The proof combines Weyl's inequality with a restricted rank bound.

Step 1: Weyl inequality. The first bound is Weyl's inequality (Stewart & Sun, 1990; Bhatia, 1997).

Step 2: Restricted rank bound. Because $\mathrm{rank}(A - B) \le 2r$, summing squared singular values gives $\|A - B\|_F^2 \le 2r\, \|A - B\|_{\mathrm{op}}^2$. This completes the proof. □

S9 Detailed scalar proximal calculations

This section verifies the scalar maps used in Algorithm 1. The calculations also show uniqueness at and between the threshold knots and justify the rank regions in Theorem 4.

For $c \ge 0$, the SCAD penalty is

$$P_{\lambda,a}(c) = \begin{cases} \lambda c, & 0 \le c \le \lambda, \\ \{-c^2 + 2a\lambda c - \lambda^2\}/\{2(a-1)\}, & \lambda < c \le a\lambda, \\ (a+1)\lambda^2/2, & c > a\lambda. \end{cases}$$

The next lemma derives its proximal minimizer.

Lemma 16. *If $a > 2$ and $z \ge 0$, the unique minimizer of*

$$\min_{c \ge 0} \left\{ \frac{1}{2}(c - z)^2 + P_{\lambda,a}(c) \right\}$$

is $\Theta_{\lambda,a}(z)$ in (5)*.*

Lemma 16 verifies the exact threshold used in the implementation. In particular, the upper SCAD region has no shrinkage.

Proof. The proof solves the scalar proximal optimization explicitly by finding the zeroes of the derivative on each piece of the SCAD penalty.

Step 1: First and second intervals. On $0 \le c \le \lambda$, the objective derivative is

$$c - z + \lambda.$$

The minimizer is zero for $0 \le z \le \lambda$ and $c = z - \lambda$ for $\lambda < z < 2\lambda$.

Step 2: Third interval. On $\lambda < c \le a\lambda$, the derivative is

$$c - z + \frac{a\lambda - c}{a - 1} = \frac{a - 2}{a - 1} c - z + \frac{a\lambda}{a - 1}.$$

Its zero is

$$c = \frac{(a - 1)z - a\lambda}{a - 2}.$$

This value lies in $(\lambda, a\lambda]$ for $2\lambda < z \le a\lambda$.

Step 3: Fourth interval and global continuity. For $c > a\lambda$, the derivative is $c - z$, giving $c = z$ when $z > a\lambda$. Because $a > 2$, the derivative is continuous and strictly increasing across the three intervals. The formulas agree at the knots, proving uniqueness and (5). This completes the proof. □

For MCP with $b > 1$, let

$$P^{\mathrm{MCP}}_{\lambda,b}(c) = \begin{cases} \lambda c - c^2/(2b), & 0 \le c \le b\lambda, \\ b\lambda^2/2, & c > b\lambda. \end{cases}$$

The next lemma gives the scalar map used in Lemma 2.

Lemma 17. *If $b > 1$ and $z \ge 0$, the unique minimizer of $\min_{c\ge 0}\{\frac{1}{2}(c-z)^2 + P^{\mathrm{MCP}}_{\lambda,b}(c)\}$ is*

$$\theta_{\lambda,b}(z) = \begin{cases} 0, & 0 \le z \le \lambda, \\ b(z-\lambda)/(b-1), & \lambda < z \le b\lambda, \\ z, & z > b\lambda. \end{cases}$$

Lemma 17 shows that MCP has the same zero threshold as SCAD and a different upper knot. This is the only change required by Corollary 4.

Proof. The proof solves the scalar proximal optimization explicitly for the two pieces of the MCP penalty.

Step 1: First interval. On $0 \le c \le b\lambda$, the derivative is

$$c - z + \lambda - c/b = \frac{b-1}{b}c - z + \lambda.$$

Its zero is $b(z-\lambda)/(b-1)$ for $\lambda < z \le b\lambda$. Zero is optimal for $z \le \lambda$.

Step 2: Second interval and continuity. Above $b\lambda$, the derivative is $c - z$, giving $c = z$. Because $b > 1$, the derivative is continuous and strictly increasing across the knot, proving uniqueness. This completes the proof. □

The scalar maps preserve the order of singular values. The SCAD slopes on its four intervals are

$$0, \qquad 1, \qquad \frac{a-1}{a-2}, \qquad 1,$$

and the values agree at the interval boundaries. The MCP slopes are

$$0, \qquad \frac{b}{b-1}, \qquad 1,$$

with agreement at the boundaries. Every slope is nonnegative, which proves that componentwise minimization preserves the ordering required of singular values.

S10 Population quantities and empirical curvature

Population second moments and conditional noise tails determine sufficient conditions for coefficient and rank recovery. We derive the required empirical Gram matrix bounds from Condition 1.

Let

$$\Sigma^0 = \sum_{s=1}^{m} w_s \Sigma^0_s.$$

The following lemma gives a direct deterministic conversion from Gram matrix deviations to the transformed curvature parameter.

Lemma 18. *If $\lambda_{\min}(\Sigma^0) \ge \mu$ and $\left\|\widehat{\Sigma} - \Sigma^0\right\|_{\mathrm{op}} \le u < \mu$, then*

$$\kappa \ge \frac{\mu - u}{\gamma + \ell_1}.$$

Lemma 18 isolates the only pooled Gram matrix deviation needed for coefficient recovery. No worker Gram matrix has to be invertible.

Proof. The proof bounds the minimum eigenvalue using Weyl's inequality and applies the definition of the metric.

Step 1: Unscaled minimum eigenvalue. Weyl's inequality gives

$$\lambda_{\min}(\widehat{\Sigma}) \geq \lambda_{\min}(\Sigma^0) - \left\|\widehat{\Sigma} - \Sigma^0\right\|_{\mathrm{op}} \geq \mu - u.$$

Step 2: Metric scaling. For every nonzero $v \in \mathbb{R}^p$,

$$\frac{v^\top \widehat{\Sigma} v}{v^\top H v} \geq \frac{\mu - u}{\lambda_{\max}(H)} \geq \frac{\mu - u}{\gamma + \ell_1}.$$

Taking the infimum gives the result. This completes the proof. □

The stabilization level depends on local upper eigenvalues. The next corollary combines Lemma 5 with the definition of γ.

Corollary 7. *Under the design assumptions of Condition 1, with independence required only within workers and without its lower second-moment bound, let $0 < \delta < 1/2$. With probability at least $1 - 2\delta$,*

$$\gamma \lesssim (1+\tau) L K_x^2 \left\{1 + \frac{p}{n_{\min}} + \frac{\log(m/\delta)}{n_{\min}}\right\}.$$

Corollary 7 quantifies the price of locally small samples. When $n_{\min} < p$, the metric remains positive definite, but its scale can grow with $p/n_{\min}$.

Proof. The proof applies the local covariance concentration bound to every worker simultaneously.

Step 1: Local concentration. The worker-wise argument in the proof of Lemma 5, applied with $t = \log(1/\delta)$, gives the stated upper bound for $\max_s \ell_s$.

Step 2: Aggregation bound. Because

$$\bar{\ell} = \sum_s w_s \ell_s \leq \max_s \ell_s$$

and $\gamma = (1+\tau)\bar{\ell}$, the result follows. This completes the proof. □

The following corollary combines the empirical curvature and score bounds into an explicit high-probability Stage I statement in the original coefficient scale.

Corollary 8. *Under Condition 1, let $0 < \delta < 1$. Suppose the sample-size premise of Corollary 1 holds and choose*

$$\lambda_1 = C\sigma\sqrt{\frac{p + q + \log(1/\delta)}{(1+\tau)N}}$$

with C sufficiently large. Then, with probability at least $1 - \delta$,

$$\left\|\widehat{B}_1 - B^\star\right\|_F \lesssim \frac{\sigma\sqrt{r\{p + q + \log(1/\delta)\}}}{\kappa\sqrt{(1+\tau)N\gamma}}.$$

Corollary 8 makes the statistical and metric factors explicit. The nominal dimension term is the usual low-rank complexity $r(p+q)$.

Proof. The proof combines the deterministic coefficient error bound with the score concentration condition.

Step 1: Score threshold. Corollary 1 gives $\lambda_1 \geq 2\varepsilon$ after increasing C.

Step 2: Coefficient conversion. Apply Theorem 6 and Lemma 8. This completes the proof. □

The noise bound in Lemma 5 is conditional on all designs. The next lemma records the variance proxy that explains the factor $\|M\|_{\mathrm{op}}$.

Lemma 19. *Under the conditional noise assumptions of Condition 1, conditional on the designs, for unit vectors $u \in \mathbb{R}^p$ and $v \in \mathbb{R}^q$,*

$$\left\| u^\top \widetilde{W} v \right\|_{\psi_2}^2 \le \frac{C\sigma^2}{N} u^\top M u.$$

Lemma 19 shows that the same metric that majorizes the pooled curvature also normalizes the score. This is why the score scale contains $(1+\tau)^{-1/2}$.

Proof. The proof verifies the sub-Gaussian norm of the score in arbitrary directions conditional on the designs.

Step 1: Summand bounds. Conditional on the designs,

$$u^\top \widetilde{W} v = \frac{1}{N} \sum_{s=1}^{m} \sum_{i=1}^{n_s} u^\top H^{-1/2} x_{si}\, e_{si}^\top v.$$

The summands are independent, centered, and sub-Gaussian with squared norm bounded by

$$C\sigma^2 \{u^\top H^{-1/2} x_{si}\}^2 / N^2.$$

Step 2: Variance proxy summation. The squared norm of their sum is at most the sum of these bounds. Thus

$$\begin{aligned} \left\| u^\top \widetilde{W} v \right\|_{\psi_2}^2 &\le \frac{C\sigma^2}{N^2} \sum_{s,i} \{u^\top H^{-1/2} x_{si}\}^2 \\ &= \frac{C\sigma^2}{N} u^\top H^{-1/2} \widehat{\Sigma} H^{-1/2} u. \end{aligned}$$

This completes the proof. □

S11 Further rank-certification consequences

The main rank theorem concerns one computed update. This section gives two consequences that help interpret repeated Stage II updates. The first gives a deterministic error recurrence whenever the rank conditions continue to hold.

Let

$$e_t = \|C_t - C^\star\|_F, \qquad A_r = 2\sqrt{2r}(1-\kappa), \qquad b_r = 2\sqrt{2r}\varepsilon.$$

The following proposition does not assume that rank identification persists. It states the recurrence on any consecutive range where the theorem's premises can be verified.

Proposition 14. *For integers $t_0 \ge T_1$ and $K \ge 1$, suppose $\eta_t < \lambda_2$ and $\sigma_r(C^\star) > a\lambda_2 + \eta_t$ at iterations $t = t_0, \dots, t_0 + K - 1$. Then*

$$e_{t_0+k} \le A_r^k e_{t_0} + b_r \sum_{j=0}^{k-1} A_r^j, \qquad k = 1, \dots, K.$$

Proposition 14 gives a coarse sufficient error recurrence. Its factor A_r is not a necessary contraction threshold; Theorem 7 uses the local projection bound to obtain contraction for every $\kappa > 0$ under stronger signal separation.

Proof. The one-step rank bound and the operator norm of $I_p - M$ give the error recurrence.

Step 1: Recursive bound. Theorem 4 and the definition of R_t give

$$\begin{aligned} e_{t+1} &\le 2\sqrt{2r}\eta_t \le 2\sqrt{2r}\{\|I_p - M\|_{\mathrm{op}}\, e_t + \varepsilon\} \\ &= A_r e_t + b_r, \end{aligned}$$

where $\|I_p - M\|_{\mathrm{op}} = 1 - \kappa$.

Step 2: Iteration. Iteration of this inequality proves the result. This completes the proof. □

The next proposition gives a sufficient invariant region. Its contraction condition is stated explicitly and is not hidden in a general rank claim.

Proposition 15. *Suppose $A_r < 1$. Let $R > 0$ satisfy*

$$(1-\kappa)R + \varepsilon < \lambda_2, \quad \sigma_r(C^\star) > a\lambda_2 + (1-\kappa)R + \varepsilon, \quad A_r R + b_r \le R.$$

If $t_0 \ge T_1$ and $e_{t_0} \le R$, then every later SCAD iterate has rank r and

$$e_{t_0+k} \le A_r^k R + \frac{b_r(1-A_r^k)}{1-A_r}, \qquad k \ge 1.$$

Proposition 15 identifies a sufficient setting for persistent rank recovery. The condition $A_r < 1$ is not required by Theorem 7. For example, if every empirical Gram matrix equals ℓI_p with $\ell > 0$, then $\kappa = (2+\tau)^{-1} < 1/2$ and $A_r > 1$ for every $r \ge 1$. The refined result can still apply after sufficient convex iterations and with sufficient signal separation.

Proof. The proof verifies that the initial error bound implies the separation conditions and is preserved by the update.

Step 1: Base case conditions. Suppose $e_t \le R$. Then

$$\eta_t \le (1-\kappa)R + \varepsilon.$$

The first two premises give the two conditions in Theorem 4. Hence $\operatorname{rank}(C_{t+1}) = r$ and

$$e_{t+1} \le A_r R + b_r \le R.$$

Step 2: Induction and error bound. Induction preserves rank r and $e_t \le R$. Summing the recurrence in Proposition 14 proves the error bound. This completes the proof. □

The observable singular-value gap can be made stable under a numerical perturbation of the update matrix. This gives a margin for rank certification with approximate singular values.

Lemma 20. *Suppose $\zeta \ge 0$ and a matrix $\widetilde{Z}_t$ satisfies $\left\|\widetilde{Z}_t - Z_t\right\|_{\mathrm{op}} \le \zeta$. If*

$$\sigma_r(\widetilde{Z}_t) > a\lambda_2 + \zeta, \qquad \sigma_{r+1}(\widetilde{Z}_t) \le \lambda_2 - \zeta,$$

then the exact SCAD update from Z_t has rank r and keeps its first r singular values.

Lemma 20 states a deterministic tolerance for an approximate spectral diagnostic. It does not identify the true rank unless the population separation conditions also hold.

Proof. The proof bounds the exact singular values using the approximate ones via Weyl's inequality.

Step 1: Singular value bounds. Weyl's inequality gives

$$\sigma_r(Z_t) \ge \sigma_r(\widetilde{Z}_t) - \zeta > a\lambda_2$$

and

$$\sigma_{r+1}(Z_t) \le \sigma_{r+1}(\widetilde{Z}_t) + \zeta \le \lambda_2.$$

Step 2: Exact update. By (5), the update leaves the first r singular values of Z_t unchanged and sets the remaining singular values to zero. This completes the proof. □

S12 Alignment of estimands, algorithms, and reported computations

Local rowwise gradients and gradients computed from cached cross-products are identical. The transformed and coefficient objectives have the same values at $C = H^{1/2}B$.

The estimator returned by Algorithm 1 is

$$\widehat{B} = H^{-1/2} C_{T_1+T_2}.$$

The code constructs the same metric with

$$\gamma = (1+\tau)\sum_s w_s \lambda_{\max}(\widehat{\Sigma}_s),$$

applies Proposition 12, and evaluates worker gradients at the transmitted coefficient matrix. The next proposition confirms that caching worker cross-products for local computation does not change the gradient oracle.

Proposition 16. *For every worker s and every B,*

$$\frac{1}{n_s} X_s^\top (X_s B - Y_s) = \widehat{\Sigma}_s B - \widehat{Z}_s.$$

For $w_s = n_s/N$,

$$\sum_s w_s \frac{1}{n_s} X_s^\top (X_s B - Y_s) = \frac{1}{N} X^\top (XB - Y).$$

Proposition 16 shows that local caching and unequal worker sample sizes change neither the algorithm nor its estimand.

Proof. The proof expands the local products and sums them with weights proportional to sample sizes.

Step 1: Local caching. Distribute $X_s^\top$ over the difference:

$$\frac{1}{n_s} X_s^\top (X_s B - Y_s) = \frac{1}{n_s} X_s^\top X_s B - \frac{1}{n_s} X_s^\top Y_s.$$

This proves the worker identity.

Step 2: Pooled identity. Because $w_s/n_s = 1/N$,

$$\sum_s w_s \frac{1}{n_s} X_s^\top (X_s B - Y_s) = \frac{1}{N} \sum_s X_s^\top (X_s B - Y_s).$$

Stacking the worker rows proves the pooled identity. This completes the proof. □

The Stage II objective check in the implementation is justified by Theorem 2. The following proposition states the exact objective in the coefficient coordinate.

Proposition 17. *For $C = H^{1/2} B$,*

$$F_2(C) = \mathcal{L}(B) + \sum_j P_{\lambda_2}\{\sigma_j(H^{1/2}B)\}, \quad \operatorname{rank}(H^{1/2}B) = \operatorname{rank}(B).$$

Hence objective descent and rank in transformed coordinates equal those of the implemented coefficient-scale criterion.

Proposition 17 connects the transformed proof, the criterion evaluated by the numerical implementation, and the reported coefficient rank.

Proof. The change of coordinates gives the pooled objective, and invertibility of the metric preserves rank.

Step 1: Objective identity. The objective identity follows from $f(C) = \mathcal{L}(H^{-1/2}C)$ and $C = H^{1/2}B$.

Step 2: Rank identity. The matrix $H^{1/2}$ is nonsingular because $H \succeq \gamma I_p$ and $\gamma > 0$. Left multiplication by a nonsingular matrix preserves rank. This completes the proof. □

S13 COMPLETE CONCENTRATION DERIVATIONS

This section derives the empirical bounds used in Corollary 1. It separates scalar concentration, sphere approximation, pooled and worker Gram matrix control, and the rectangular score bound. The argument permits nonzero predictor means, nonidentical second-moment matrices, and unequal sample sizes.

We first record a covering fact for Euclidean spheres. It is used for both symmetric quadratic forms and rectangular bilinear forms.

Lemma 21. *For $d \geq 1$ and $0 < \epsilon < 1$, the unit sphere $\mathbb{S}^{d-1}$ has an ϵ-net $\mathcal{N}_\epsilon$ satisfying*

$$|\mathcal{N}_\epsilon| \leq (1 + 2/\epsilon)^d.$$

Lemma 21 reduces operator-norm control to finitely many scalar deviations. Its logarithmic size is proportional to the matrix dimension.

Proof. The proof bounds the size of a maximal separated set using a volume comparison.

Step 1: Net construction. Choose a maximal set $\mathcal{N}_\epsilon \subset \mathbb{S}^{d-1}$ with pairwise distances greater than ϵ. Maximality makes it an ϵ-net.

Step 2: Volume bound. The Euclidean balls of radius $\epsilon/2$ centered at its points are disjoint and lie in the ball of radius $1 + \epsilon/2$. Comparing volumes gives

$$|\mathcal{N}_\epsilon|(\epsilon/2)^d \leq (1 + \epsilon/2)^d.$$

Rearranging proves the bound. This completes the proof. □

Gram matrix concentration requires control over all directions. The following inequality obtains the operator norm from quadratic forms on a finite net.

Lemma 22. *Let $A \in \mathbb{R}^{d\times d}$ be symmetric and let $\mathcal{N}_\epsilon$ be an ϵ-net of $\mathbb{S}^{d-1}$ with $0 < \epsilon < 1/2$. Then*

$$\|A\|_{\mathrm{op}} \leq (1 - 2\epsilon)^{-1} \max_{u\in\mathcal{N}_\epsilon} |u^\top A u|.$$

Lemma 22 is the deterministic step behind Gram matrix concentration. We use $\epsilon = 1/4$, which gives the factor two.

Proof. The proof bounds the continuous maximum by a discrete maximum over the net plus a continuity error.

Step 1: Net approximation. Choose a unit vector x such that $|x^\top A x| = \|A\|_{\mathrm{op}}$. There is $u \in \mathcal{N}_\epsilon$ with $\|x - u\|_2 \leq \epsilon$. Then

$$|x^\top A x - u^\top A u| \leq |(x - u)^\top A x| + |u^\top A(x - u)| \leq 2\epsilon \|A\|_{\mathrm{op}}.$$

Step 2: Discrete maximum. Hence

$$\|A\|_{\mathrm{op}} \leq \max_{u\in\mathcal{N}_\epsilon} |u^\top A u| + 2\epsilon \|A\|_{\mathrm{op}}.$$

Rearranging proves the result. This completes the proof. □

Rectangular matrices require two sphere nets. The following lemma gives the corresponding deterministic inequality.

Lemma 23. *Let $A \in \mathbb{R}^{p\times q}$ and let $\mathcal{N}_p$ and $\mathcal{N}_q$ be ϵ-nets of $\mathbb{S}^{p-1}$ and $\mathbb{S}^{q-1}$ with $0 < \epsilon < 1/2$. Then*

$$\|A\|_{\mathrm{op}} \leq (1 - 2\epsilon)^{-1} \max_{u\in\mathcal{N}_p, v\in\mathcal{N}_q} |u^\top A v|.$$

Lemma 23 is used for the transformed score. It gives dimension $p + q$ because the two net sizes multiply.

Proof. The proof bounds the bilinear form maximum using independent approximations for the left and right singular vectors.

Step 1: Net approximation. Choose unit vectors x, y with $x^\top Ay = \|A\|_{\mathrm{op}}$ after changing a sign when needed. Select $u \in \mathcal{N}_p$ and $v \in \mathcal{N}_q$ with $\|x-u\|_2 \le \epsilon$ and $\|y-v\|_2 \le \epsilon$. Then

$$|x^\top Ay - u^\top Av| \le |(x-u)^\top Ay| + |u^\top A(y-v)| \le 2\epsilon \|A\|_{\mathrm{op}}.$$

Step 2: Discrete maximum. Rearranging as in Lemma 22 proves the result. This completes the proof. □

For a scalar random variable Z, let $\|Z\|_{\psi_1} = \inf\{b > 0 : \mathbb{E}\exp(|Z|/b) \le 2\}$ denote the sub-exponential Orlicz norm. We next record the scalar deviation used for Gram matrices. The sub-Gaussian premise in Condition 1 implies a uniform sub-exponential bound for centered squares.

Lemma 24. *Let Z be sub-Gaussian with $\|Z\|_{\psi_2} \le K(\mathbb{E}Z^2)^{1/2}$. Then*

$$\left\|Z^2 - \mathbb{E}Z^2\right\|_{\psi_1} \le CK^2\mathbb{E}Z^2.$$

Lemma 24 converts the design assumption into the Bernstein scale for one quadratic form.

Proof. The proof applies the Orlicz product inequality to bound the sub-exponential norm of the square.

Step 1: Uncentered square. The standard Orlicz product inequality gives

$$\left\|Z^2\right\|_{\psi_1} \le C\|Z\|_{\psi_2}^2.$$

Step 2: Centering. Centering changes the sub-exponential norm by at most a factor two. Hence

$$\left\|Z^2 - \mathbb{E}Z^2\right\|_{\psi_1} \le C\|Z\|_{\psi_2}^2 \le CK^2\mathbb{E}Z^2.$$

This completes the proof. □

For a lower eigenvalue bound, absolute Gram matrix error can impose an unnecessary dependence on the condition number of Σ^0. Relative error depends on χ, defined before Corollary 1. This distinction is useful even when every worker second-moment matrix is singular.

Proposition 18. *Under the design assumptions of Condition 1, for every $t > 0$, with probability at least $1 - 2e^{-t}$,*

$$\left\|(\Sigma^0)^{-1/2}(\widehat{\Sigma} - \Sigma^0)(\Sigma^0)^{-1/2}\right\|_{\mathrm{op}} \le CK_x^2\left\{\sqrt{\frac{\chi(p+t)}{N}} + \frac{\chi(p+t)}{N}\right\}.$$

Proposition 18 gives $\Sigma^0/2 \preceq \widehat{\Sigma} \preceq 3\Sigma^0/2$ when $N \ge CK_x^4\chi(p+t)$. No comparison between a realized local Gram matrix and the pooled Gram matrix is assumed. The same result applies on the support of a singular population second-moment matrix, as used in Proposition 61.

Proof. The proof applies Bernstein's inequality for nonidentically distributed variables to a net over the sphere.

Step 1: Normalized variables. Let $z_{si} = (\Sigma^0)^{-1/2}x_{si}$ and $V_s = (\Sigma^0)^{-1/2}\Sigma_s^0(\Sigma^0)^{-1/2}$. Then

$$V_s \succeq 0, \qquad \sum_s w_s V_s = I_p, \qquad \max_s \|V_s\|_{\mathrm{op}} = \chi.$$

For a fixed unit vector u, let $v_s = u^\top V_s u$ and $\xi_{si} = (u^\top z_{si})^2 - v_s$. Lemma 24 gives

$$\|\xi_{si}\|_{\psi_1} \le CK_x^2 v_s, \qquad \sum_s n_s v_s = N, \qquad \sum_s n_s v_s^2 \le \chi N.$$

Step 2: Bernstein bound. Bernstein's inequality for independent, nonidentically distributed subexponential variables (Vershynin, 2018, Theorem 2.8.1) gives

$$\Pr\Big\{\Big|\frac{1}{N}\sum_{s,i}\xi_{si}\Big| > z\Big\} \le 2\exp\Big[-cN\min\Big\{\frac{z^2}{K_x^4\chi}, \frac{z}{K_x^2\chi}\Big\}\Big].$$

Step 3: Uniform bound. A $1/4$-net has at most 9^p points. Lemma 22 and a union bound give an exponent at most $p\log 9 - cN\min\{z^2/(K_x^4\chi), z/(K_x^2\chi)\}$. Taking

$$z = CK_x^2\left\{\sqrt{\frac{\chi(p+t)}{N}} + \frac{\chi(p+t)}{N}\right\}$$

proves the assertion after increasing C. This completes the proof. □

The dependence on χp is necessary for the empirical Gram matrix to be positive definite uniformly over the stated designs. To see this, take two workers, an integer $1 \le n_1 < N$, $w_1 = n_1/N = 1/\chi$, $\Sigma_1^0 = \chi I_p$, and $\Sigma_2^0 = 0$. Independent Gaussian rows at the first worker and zero rows at the second satisfy the design conditions with $\Sigma^0 = I_p$ and an absolute K_x. The pooled empirical Gram matrix has rank at most $n_1 = N/\chi$, hence it is singular if $N < \chi p$. This example gives a necessary sample size for equality of the empirical and population predictor spaces. Replacing I_p by an orthogonal projection gives the same dependence on the dimension of its range.

The pooled Gram matrix averages outer products of independent but nonidentically distributed rows. The next proposition gives an absolute-error bound in terms of the largest population eigenvalue.

Proposition 19. *Under the design assumptions of Condition 1, without its lower second-moment bound, for every $t > 0$, with probability at least $1 - 2e^{-t}$,*

$$\Big\|\widehat{\Sigma} - \Sigma^0\Big\|_{\mathrm{op}} \le CK_x^2 L\left\{\sqrt{\frac{p+t}{N}} + \frac{p+t}{N}\right\}.$$

Proposition 19 shows that heterogeneous predictor second moments do not change the pooled sample size in the operator bound when their matrices have a common upper eigenvalue bound.

Proof. The proof applies Bernstein's inequality to a net using the uniform bound on local eigenvalues.

Step 1: Summand bounds. Let $u \in \mathbb{S}^{p-1}$ and let

$$Z_{si} = (u^\top x_{si})^2 - u^\top \Sigma_s^0 u.$$

Lemma 24 and $u^\top \Sigma_s^0 u \le L$ give

$$\|Z_{si}\|_{\psi_1} \le CK_x^2 L.$$

Step 2: Bernstein bound. Because

$$u^\top(\widehat{\Sigma} - \Sigma^0)u = \frac{1}{N}\sum_{s=1}^{m}\sum_{i=1}^{n_s} Z_{si},$$

Bernstein's inequality gives

$$\Pr\Big\{|u^\top(\widehat{\Sigma} - \Sigma^0)u| > z\Big\} \le 2\exp\left[-cN\min\left\{\frac{z^2}{K_x^4L^2}, \frac{z}{K_x^2L}\right\}\right].$$

Step 3: Uniform bound. By Lemma 21, a $1/4$-net $\mathcal{N}$ has $|\mathcal{N}| \le 9^p$. A union bound and Lemma 22 give

$$\Pr\{\Big\|\widehat{\Sigma} - \Sigma^0\Big\|_{\mathrm{op}} > 2z\} \le 2\exp\left[p\log 9 - cN\min\left\{\frac{z^2}{K_x^4L^2}, \frac{z}{K_x^2L}\right\}\right].$$

Choose

$$z = CK_x^2 L\left\{\sqrt{\frac{p+t}{N}} + \frac{p+t}{N}\right\}$$

For sufficiently large C, the exponent is at most $-t$; absorb the norm factor two into C. This completes the proof. □

The worker-wise bound follows from the same argument and a union bound. We state it separately because it controls the setup metric.

Proposition 20. *Under the design assumptions of Condition 1, with independence required only within workers and without its lower second-moment bound, for every $t > 0$, with probability at least $1 - 2e^{-t}$,*

$$\max_{1\le s\le m} \left\|\widehat{\Sigma}_s - \Sigma_s^0\right\|_{\mathrm{op}} \le CK_x^2 L \left\{ \sqrt{\frac{p + t + \log m}{n_{\min}}} + \frac{p + t + \log m}{n_{\min}} \right\}.$$

Proposition 20 permits $n_{\min} < p$. In that case the bound grows, but the metric remains positive definite because its smallest eigenvalue is bounded below by γ.

Proof. The proof applies the pooled concentration argument to each worker and unions over workers.

Step 1: Local bound. Apply Proposition 19 to worker s, replacing N, Σ^0 by n_s, Σ_s^0. For $z > 0$,

$$\Pr\left\{ \left\|\widehat{\Sigma}_s - \Sigma_s^0\right\|_{\mathrm{op}} > z \right\} \le 2\exp\left[p\log 9 - cn_s \min\left\{ \frac{z^2}{K_x^4 L^2}, \frac{z}{K_x^2 L} \right\} \right].$$

Step 2: Union over workers. Sum this probability over s, use $n_s \ge n_{\min}$, and choose z with $p + t + \log m$ in place of $p + t$. This completes the proof. □

We now derive the conditional score bound. This argument permits arbitrary dependence among the q entries of one error vector because Condition 1 controls every linear contrast.

Proposition 21. *Under the conditional noise assumptions of Condition 1, conditional on the designs, for every $t > 0$, with probability at least $1 - 2e^{-t}$,*

$$\left\|\widetilde{W}\right\|_{\mathrm{op}} \le C\sigma \sqrt{\|M\|_{\mathrm{op}}} \sqrt{\frac{p + q + t}{N}}.$$

Proposition 21 yields the low-rank dimension $p + q$ and not the parameter count pq. The metric upper bound then produces the factor $(1 + \tau)^{-1/2}$.

Proof. The proof bounds the transformed score using the conditional variance proxy and a union bound over two nets.

Step 1: Pointwise bound. If $\|M\|_{\mathrm{op}} = 0$, then $XH^{-1/2} = 0$ and $\widetilde{W} = 0$. For $\|M\|_{\mathrm{op}} > 0$ and unit $u \in \mathbb{R}^p, v \in \mathbb{R}^q$, Lemma 19 gives

$$\left\|u^\top \widetilde{W} v\right\|_{\psi_2}^2 \le \frac{C\sigma^2 \|M\|_{\mathrm{op}}}{N}.$$

Thus

$$\Pr\{|u^\top \widetilde{W} v| > z \mid X_1, \dots, X_m\} \le 2\exp\left\{ -\frac{cNz^2}{\sigma^2 \|M\|_{\mathrm{op}}} \right\}.$$

Step 2: Union bound. For $1/4$-nets $\mathcal{N}_p, \mathcal{N}_q$ with at most 9^{p+q} pairs, a union bound gives

$$\Pr\left\{ \max_{u\in\mathcal{N}_p, v\in\mathcal{N}_q} |u^\top \widetilde{W} v| > z \mid X_1, \dots, X_m \right\} \le 2\exp\left\{ (p+q)\log 9 - \frac{cNz^2}{\sigma^2 \|M\|_{\mathrm{op}}} \right\}.$$

Step 3: Continuous norm. Choose

$$z = C\sigma \sqrt{\|M\|_{\mathrm{op}}} \sqrt{\frac{p + q + t}{N}}$$

and apply Lemma 23. This completes the proof. □

The common noise scale in Condition 1 can be replaced by worker-specific scales. This extension is useful when local response variability differs across partitions. For scales $\sigma_s \geq 0$, define

$$M_\sigma = H^{-1/2}\left(\sum_s w_s \sigma_s^2 \widehat{\Sigma}_s\right) H^{-1/2}.$$

Theorem 13. *Conditional on the designs, suppose the vectors $\{e_{si}\}_{s,i}$ are independent, $\mathbb{E}(e_{si} \mid \{X_s\}_s) = 0$, and*

$$\left\|v^\top e_{si}\right\|_{\psi_2} \leq \sigma_s \|v\|_2, \qquad v \in \mathbb{R}^q.$$

Then, for every $t > 0$, conditional on the designs, with probability at least $1 - 2e^{-t}$,

$$\left\|\widetilde{W}\right\|_{\mathrm{op}} \leq C\sqrt{\|M_\sigma\|_{\mathrm{op}}}\sqrt{\frac{p+q+t}{N}}.$$

Theorem 13 shows that heterogeneous response noise enters through a weighted sum of the worker Gram matrices. The simpler bound with $\max_s \sigma_s$ follows from Lemma 1.

Proof. The proof generalizes the conditional score bound to distinct noise scales.

Step 1: Pointwise bound. For unit u, v, conditional independence gives

$$\begin{aligned}\left\|u^\top \widetilde{W} v\right\|_{\psi_2}^2 &\leq \frac{C}{N^2}\sum_{s,i}\sigma_s^2\{u^\top H^{-1/2}x_{si}\}^2 \\ &= \frac{C}{N}u^\top M_\sigma u \leq \frac{C\|M_\sigma\|_{\mathrm{op}}}{N}.\end{aligned}$$

Step 2: Continuous norm. Repeat the two-net argument in Proposition 21. This completes the proof. □

The preceding bounds prove Lemma 5 with explicit probability allocation. We give the final assembly to make the dependency clear.

Corollary 9. *Under Condition 1, without its lower second-moment bound, for every $t > 0$, with probability at least $1 - 6e^{-t}$, Propositions 19, 20, and 21 hold simultaneously.*

Corollary 9 controls the pooled Gram matrix, worker Gram matrices, and transformed score on one event under the population condition. These simultaneous bounds justify using the estimated metric in the prediction and rank analysis.

Proof. The proof sums the failure probabilities of the three structural bounds.

Step 1: Individual bounds. Each Gram matrix bound has failure probability at most $2e^{-t}$. Integrating the conditional score bound gives the same probability.

Step 2: Union bound. A union bound proves the result. This completes the proof. □

S14 Finite Stage I accuracy and communication rounds

The main theorem separates statistical error from optimization error. This section gives explicit round counts for coefficient accuracy, prediction accuracy, and entry into the rank-identification region. These results explain when a communication advantage can coexist with a prescribed statistical target.

Let D_{init} denote the initial distance to the convex target. When $\kappa > 0$, let R_{stat} denote the coefficient-error bound:

$$D_{\mathrm{init}} = \left\|C_0 - \widehat{C}\right\|_F, \qquad R_{\mathrm{stat}} = \frac{2\sqrt{2r}\lambda_1}{\kappa}.$$

The first proposition gives the complete Stage I error decomposition.

Proposition 22. *Suppose $\kappa > 0$, $\lambda_1 \geq \varepsilon$, and $1 \leq T \leq T_1$. Then*

$$\|C_T - C^\star\|_F \leq \frac{2D_{\text{init}}}{\sqrt{\kappa}(T+1)} + R_{\text{stat}},$$

$$\left\|H^{-1/2}C_T - B^\star\right\|_F \leq \frac{1}{\sqrt{\gamma}}\left\{\frac{2D_{\text{init}}}{\sqrt{\kappa}(T+1)} + R_{\text{stat}}\right\}.$$

Proposition 22 gives the optimization and statistical terms in the same transformed norm. The first decreases with communication rounds and the second does not.

Proof. The proof bounds the optimization error and statistical error in the transformed scale.

Step 1: Triangle inequality. The triangle inequality gives

$$\|C_T - C^\star\|_F \leq \left\|C_T - \widehat{C}\right\|_F + \left\|\widehat{C} - C^\star\right\|_F.$$

Step 2: Component bounds. Apply the two bounds in Theorem 6. Lemma 8 gives the coefficient-scale inequality. This completes the proof. □

The zero initialization admits a bound that depends only on the transformed signal and statistical radius. If $\kappa > 0$, $\lambda_1 \geq \varepsilon$, and $C_0 = 0$, then

$$\left\|\widehat{C}\right\|_F \leq \|C^\star\|_F + \left\|\widehat{C} - C^\star\right\|_F,$$

and Theorem 6 gives

$$D_{\text{init}} \leq \|C^\star\|_F + R_{\text{stat}}.$$

This relation removes the unknown Stage I target from a sufficient round bound. The transformed signal norm remains, as expected for a global distance guarantee.

Proposition 22 also gives the round count needed to make optimization error no larger than a prescribed tolerance. If $\kappa > 0$, $\lambda_1 \geq \varepsilon$, $1 \leq T \leq T_1$, $u > 0$, and

$$T + 1 \geq \frac{2D_{\text{init}}}{\sqrt{\kappa}\, u},$$

then

$$\|C_T - C^\star\|_F \leq R_{\text{stat}} + u.$$

The bound states the communication cost of reducing optimization error below the statistical scale. Taking $u = R_{\text{stat}}$ makes the two terms equal.

Prediction accuracy does not require conversion to coefficient distance. The quadratic loss controls the prediction difference between an iterate and the convex target directly.

Proposition 23. *Suppose $\lambda_1 \geq \varepsilon$ and $1 \leq T \leq T_1$. Then*

$$\frac{\left\|X(H^{-1/2}C_T - B^\star)\right\|_F}{\sqrt{N}} \leq \frac{2D_{\text{init}}}{T+1} + \frac{2\sqrt{2r_X}\lambda_1}{\sqrt{\kappa_+}}.$$

Proposition 23 applies to a singular pooled design. Its optimization term has no inverse-curvature factor, although the initial distance and statistical term still depend on the design. For an approximately low-rank response mean, the square root of the right-hand side in Theorem 5 can replace the statistical term.

Proof. The proof converts the accelerated objective bound into prediction error using the quadratic pooled loss.

Step 1: Convexity lower bound. Optimality, the subgradient inequality, and quadratic expansion give, for every C,

$$F_1(C) - F_1(\widehat{C}) \geq \tfrac{1}{2}\left\|M^{1/2}(C - \widehat{C})\right\|_F^2.$$

Step 2: Optimization bound. The accelerated objective bound in Theorem 6 gives

$$\frac{\left\|XH^{-1/2}(C_T - \widehat{C})\right\|_F}{\sqrt{N}} = \left\|M^{1/2}(C_T - \widehat{C})\right\|_F \leq \frac{2D_{\text{init}}}{T+1}.$$

Step 3: Triangle inequality. The triangle inequality and Theorem 1 give the bound; Theorem 5 gives its approximation version. This completes the proof. □

The next result solves the Stage I entry inequality in Corollary 3. It separates the statistical margin from the optimization rounds. Define

$$d_{\text{entry}} = \min\{\lambda_2, \sigma_r(C^\star) - a\lambda_2\} - (1-\kappa)R_{\text{stat}} - \varepsilon.$$

Proposition 24. *Suppose $\kappa > 0$, $\lambda_1 \geq \varepsilon$, $T_1 \geq 1$, and $T_2 \geq 1$. If $d_{\text{entry}} > 0$ and*

$$T_1 + 1 > \frac{2(1-\kappa)D_{\text{init}}}{\sqrt{\kappa}\, d_{\text{entry}}},$$

then the first Stage II update has rank r and satisfies Theorem 4.

Proposition 24 separates statistical error from the optimization rounds. This sufficient certificate requires a positive entry margin. Additional convex rounds reduce the optimization term but not the statistical error term.

Proof. The proof verifies the two premises of the entry corollary using the round condition.

Step 1: Round condition. The round condition gives

$$(1-\kappa)\frac{2D_{\text{init}}}{\sqrt{\kappa}(T_1+1)} < d_{\text{entry}}.$$

Step 2: Margin verification. By the definition of d_{entry},

$$\rho_{T_1} < \min\{\lambda_2, \sigma_r(C^\star) - a\lambda_2\}.$$

These are the two premises in Corollary 3. This completes the proof. □

The communication comparison can now be combined with the entry bound. This result states a sufficient condition and does not claim that every statistically valid setting has lower communication. Let

$$T_{\max} = \frac{1}{2} + \frac{p+1}{4q} - \frac{1}{2pq}.$$

Corollary 10. *Suppose $\kappa > 0$, $\lambda_1 \geq \varepsilon$, $d_{\text{entry}} > 0$, and $T_2 \geq 1$. Assume that*

$$\frac{2(1-\kappa)D_{\text{init}}}{\sqrt{\kappa}\, d_{\text{entry}}} + T_2 + 1 < T_{\max}.$$

Then there exists an integer $T_1 \geq 1$ for which Algorithm 1 has a rank-identified first Stage II update and transmits fewer scalars than sufficient-statistic aggregation.

Corollary 10 shows that large p/q alone is not enough. The statistical margin and Stage I conditioning must also permit entry before the scalar threshold is reached.

Proof. The proof constructs a feasible round assignment from the strict communication inequality.

Step 1: Round assignment. Let

$$L_{\text{entry}} = \frac{2(1-\kappa)D_{\text{init}}}{\sqrt{\kappa}\, d_{\text{entry}}}$$

and choose $T_1 = \lfloor L_{\text{entry}} \rfloor + 1$, where $\lfloor u \rfloor$ is the largest integer at most u. Then $T_1 + 1 > L_{\text{entry}}$ and $T_1 + T_2 \leq L_{\text{entry}} + T_2 + 1 < T_{\max}$.

Step 2: Conclusions. Proposition 24 gives rank identification, and Proposition 1 gives the communication comparison. This completes the proof. □

S14.1 HIGH-PROBABILITY RANK RECOVERY

Corollary 3 is stated through empirical curvature and score quantities. The following result gives a population sufficient condition for exact rank recovery after a finite number of Stage I rounds. For $0 < \delta < 1$ and a sufficiently large absolute constant C, define

$$L_\delta = CLK_x^2 \left\{ 1 + \frac{p}{n_{\min}} + \frac{\log(m/\delta)}{n_{\min}} \right\}, \qquad \varepsilon_\delta = C\sigma \sqrt{\frac{p+q+\log(1/\delta)}{(1+\tau)N}},$$

$$\kappa_\delta = \frac{\mu}{2(2+\tau)L_\delta}, \qquad R_\delta = \frac{2\sqrt{2r}\lambda_1}{\kappa_\delta}, \qquad D_\delta = \sqrt{(2+\tau)L_\delta}\,\|B^\star\|_F + R_\delta,$$

and

$$\rho_\delta(T) = \frac{2D_\delta}{\sqrt{\kappa_\delta}(T+1)} + R_\delta + \varepsilon_\delta.$$

Corollary 11. *Under Condition 1, suppose the sample-size premise of Corollary 1 holds and Algorithm 1 uses $C_0 = 0$, $T_1 = T \geq 1$, $T_2 \geq 1$, and $\lambda_1 \geq \varepsilon_\delta$. If*

$$\rho_\delta(T) < \lambda_2, \qquad \sqrt{\frac{(1+\tau)\mu}{2}}\,\sigma_r(B^\star) > a\lambda_2 + \rho_\delta(T),$$

then, with probability at least $1-\delta$,

$$\operatorname{rank}(C_{T+1}) = r, \qquad \|C_{T+1} - C^\star\|_F \leq 2\sqrt{2r}\rho_\delta(T).$$

Corollary 11 separates the population signal condition from the finite optimization error. The singular-value condition has the same beta-min form used for rank recovery with folded-concave spectral penalties (Park et al., 2024; 2026).

Proof. On the event in Corollary 1, increasing C if needed gives

$$\max_s \ell_s \leq L_\delta, \qquad \varepsilon \leq \varepsilon_\delta, \qquad \lambda_{\min}(\widehat{\Sigma}) \geq \mu/2.$$

Hence

$$\gamma + \ell_1 \leq (2+\tau)L_\delta, \qquad \kappa \geq \kappa_\delta, \qquad \|C^\star\|_F \leq \sqrt{(2+\tau)L_\delta}\,\|B^\star\|_F.$$

Theorem 6 and $C_0 = 0$ imply

$$\left\| C_0 - \widehat{C} \right\|_F \leq \|C^\star\|_F + \frac{2\sqrt{2r}\lambda_1}{\kappa} \leq D_\delta.$$

Because $0 \preceq M \preceq (1+\tau)^{-1} I_p$,

$$\eta_T \leq (1-\kappa)\,\|C_T - C^\star\|_F + \varepsilon \leq \rho_\delta(T).$$

Also, $\widehat{\Sigma} \preceq \bar{\ell} I_p$ and $\lambda_{\min}(\widehat{\Sigma}) \geq \mu/2$ give $\gamma \geq (1+\tau)\mu/2$. Thus

$$\sigma_r(C^\star) \geq \sqrt{\gamma}\,\sigma_r(B^\star) \geq \sqrt{\frac{(1+\tau)\mu}{2}}\,\sigma_r(B^\star).$$

The two premises of Theorem 4 follow, and that theorem gives the conclusion. This completes the proof. □

The preceding corollary concerns the first Stage II update. The following corollary gives population conditions for the returned iterate after any positive number of Stage II rounds. With the constants defined above, let

$$R_{\delta,T} = \frac{2D_\delta}{\sqrt{\kappa_\delta}(T+1)} + R_\delta, \qquad \eta_{\delta,T} = (1-\kappa_\delta)R_{\delta,T} + \varepsilon_\delta, \qquad s_\delta = \sqrt{\frac{(1+\tau)\mu}{2}}\,\sigma_r(B^\star),$$

and $q_\delta = (1-\kappa_\delta)/(1-\kappa_\delta/2)$.

Corollary 12. *Under Condition 1 and the sample-size premise of Corollary 1, let $C_0 = 0$, $T_1 = T \geq 1$, and $\lambda_1 \geq \varepsilon_\delta$. Suppose*

$$\eta_{\delta,T} < \lambda_2, \qquad s_\delta > a\lambda_2 + \eta_{\delta,T}, \qquad \eta_{\delta,T} \leq \frac{\kappa_\delta s_\delta}{1+\kappa_\delta}.$$

With probability at least $1-\delta$, simultaneously for all $k \geq 1$,

$$\operatorname{rank}(C_{T+k}) = r, \qquad \|C_{T+k} - C^\star\|_F \leq q_\delta^k R_{\delta,T} + \frac{2\sqrt{2r}\varepsilon_\delta}{\kappa_\delta}(1-q_\delta^k).$$

Corollary 12 connects the data model, the zero initialization, and the final SCAD estimate without assuming rank recovery along the computed sequence. Its signal condition is sufficient and need not hold for tuning parameters chosen by prediction validation.

Proof. The joint event in the proof of Corollary 11 gives

$$\kappa \geq \kappa_\delta > 0, \qquad \varepsilon \leq \varepsilon_\delta, \qquad \sigma_r(C^\star) \geq s_\delta, \qquad e_T \leq R_{\delta,T}.$$

Also, $\lambda_1 \geq \varepsilon_\delta$ gives $R_{\delta,T} \geq R_\delta \geq 2\sqrt{2r}\varepsilon_\delta/\kappa_\delta$. For $e_t \leq R_{\delta,T}$, the update error is at most $\eta_{\delta,T}$, and the first two signal inequalities give rank r without shrinkage. The proof of Theorem 7, with $\|I_p - M\|_{\mathrm{op}} \leq 1-\kappa_\delta$ and the third signal inequality, gives

$$e_{t+1} \leq \frac{(1-\kappa_\delta)e_t + \sqrt{2r}\varepsilon_\delta}{1-\kappa_\delta/2}.$$

The radius is at least the fixed point of this recurrence. Induction and summation prove both conclusions on the same event for all k. This completes the proof. □

The objective error has a faster iteration exponent than the coefficient distance. The objective inequality in Theorem 6 implies that, for $v > 0$ and an integer $1 \leq T \leq T_1$, the condition

$$T + 1 \geq \frac{\sqrt{2}D_{\mathrm{init}}}{\sqrt{v}},$$

gives

$$F_1(C_T) - F_1(\widehat{C}) \leq v.$$

This is the T^{-2} accelerated objective rate. Coefficient distance has the slower T^{-1} conversion because strong convexity is used after the objective bound.

S15 Rank selection and shrinkage separation

The signal condition in Theorem 4 guarantees both rank recovery and absence of shrinkage on the kept singular values. Rank recovery alone requires a weaker separation. This section states the distinction precisely.

Proposition 25. *Suppose $\eta_t < \lambda_2$. Then the SCAD update has rank at most r. If $\sigma_r(C^\star) > \lambda_2 + \eta_t$, then it has rank exactly r and*

$$\|C_{t+1} - C^\star\|_F \leq \sqrt{2r}(\lambda_2 + \eta_t).$$

Proposition 25 separates exact rank from retention without shrinkage. The stronger upper-knot condition in Theorem 4 removes the λ_2 term from the error bound.

Proof. Weyl's inequality gives

$$\sigma_{r+1}(Z_t) \leq \eta_t < \lambda_2.$$

Thresholding gives $\operatorname{rank}(C_{t+1}) \leq r$. Under the signal condition,

$$\sigma_r(Z_t) \geq \sigma_r(C^\star) - \eta_t > \lambda_2,$$

hence $\mathrm{rank}(C_{t+1}) = r$. The SCAD map satisfies

$$0 \le z - \Theta_{\lambda_2,a}(z) \le \lambda_2.$$

It follows that

$$\|C_{t+1} - Z_t\|_{\mathrm{op}} \le \lambda_2.$$

Hence

$$\|C_{t+1} - C^\star\|_{\mathrm{op}} \le \lambda_2 + \eta_t.$$

Because both matrices have rank r, Lemma 15 gives the Frobenius bound. This completes the proof. □

The same weak separation applies to MCP because its zero threshold is also λ_2.

Corollary 13. *Proposition 25 holds for MCP with the same conditions and conclusion.*

Corollary 13 confirms that the rank-only condition depends on the lower knot, while no-shrinkage depends on the penalty-specific upper knot.

Proof. The MCP map is zero on $[0, \lambda_2]$, positive above λ_2, and satisfies $0 \le z - \theta_{\lambda_2,b}(z) \le \lambda_2$. Weyl's inequality gives

$$\sigma_{r+1}(Z_t) \le \eta_t < \lambda_2,$$

which proves $\mathrm{rank}(C_{t+1}) \le r$. Under the signal condition,

$$\sigma_r(Z_t) \ge \sigma_r(C^\star) - \eta_t > \lambda_2,$$

hence the rank equals r. The scalar bound and the MCP spectral map give

$$\|C_{t+1} - C^\star\|_{\mathrm{op}} \le \lambda_2 + \eta_t.$$

Both matrices have rank r. Lemma 15 proves the Frobenius bound. This completes the proof. □

We next compare soft thresholding and SCAD at the same update matrix. This isolates spectral shrinkage from optimization and communication effects.

Proposition 26. *Suppose*

$$\sigma_r(Z) > a\lambda, \qquad \sigma_{r+1}(Z) \le \lambda.$$

Let C_{scad} and C_{soft} be the SCAD and soft-thresholded updates at Z with threshold λ. Then

$$C_{\mathrm{scad}} = \mathcal{P}_r(Z), \qquad \|C_{\mathrm{scad}} - C_{\mathrm{soft}}\|_F = \sqrt{r}\lambda, \qquad \|C_{\mathrm{scad}} - C_{\mathrm{soft}}\|_* = r\lambda,$$

where $\mathcal{P}_r$ is the best rank-r approximation defined in Appendix S2.

Proposition 26 quantifies the shrinkage removed by Stage II after the upper-knot separation occurs. The comparison holds at a common update matrix and does not attribute the difference to a change in data or gradients.

Proof. Let

$$Z = U\mathrm{diag}(d)V^\top.$$

The SCAD map leaves $d_1, \dots, d_r$ unchanged and maps the remaining singular values to zero. Thus

$$C_{\mathrm{scad}} = U_r\mathrm{diag}(d_1, \dots, d_r)V_r^\top = \mathcal{P}_r(Z).$$

Soft thresholding gives

$$C_{\mathrm{soft}} = U_r\mathrm{diag}(d_1 - \lambda, \dots, d_r - \lambda)V_r^\top.$$

The difference has r singular values equal to λ. Its Frobenius and nuclear norms are $\sqrt{r}\lambda$ and $r\lambda$. This completes the proof. □

The observable rank of a spectral proximal update is determined by the lower threshold. This gives a direct count without reference to a true parameter. Let $\mathrm{prox}_{P_\lambda}(Z)$ denote the matrix update in Lemma 2, and let $\#A$ denote the number of elements of a finite set A.

Lemma 25. *For SCAD or MCP,*

$$\operatorname{rank}\{\operatorname{prox}_{P_\lambda}(Z)\} = \#\{j : \sigma_j(Z) > \lambda\}.$$

Lemma 25 states the exact rank computed by the implementation. Equality at a threshold is mapped to zero under the convention in Algorithm 1.

Proof. Both scalar maps equal zero on $[0, \lambda]$ and are strictly positive on (λ, ∞). Lemma 2 applies the map to each singular value. This completes the proof. □

The no-false-positive condition can be checked using any operator-norm bound on the update error.

Corollary 14. *If*

$$\|Z_t - C^\star\|_{\mathrm{op}} < \lambda_2,$$

then the SCAD and MCP updates have rank at most r.

Corollary 14 excludes false positive singular values through a bound on the total update error. A beta-min condition is needed only to prevent false negatives.

Proof. Weyl's inequality gives

$$\sigma_{r+1}(Z_t) \le \|Z_t - C^\star\|_{\mathrm{op}} < \lambda_2.$$

Lemma 25 and the threshold convention prove the result. This completes the proof. □

S16 A GENERAL SPECTRAL THRESHOLD RESULT

The proofs for SCAD and MCP extend to other spectral penalties through the scalar proximal map. We first verify the scalar ordering property of Lu et al. (2015). We then separate rank identification, which needs two threshold regions, from descent and stationarity.

Let $p_\lambda : [0, \infty) \to [0, \infty)$ be lower semicontinuous and let

$$\theta_\lambda(z) \in \arg\min_{c \ge 0} \left\{ \frac{1}{2}(c - z)^2 + p_\lambda(c) \right\}$$

be any selected minimizer. Such a minimizer exists because the scalar objective is lower semicontinuous and coercive. Define

$$P_\lambda(C) = \sum_j p_\lambda\{\sigma_j(C)\}.$$

The first proposition gives the exact matrix update.

Proposition 27. *If $Z = U\operatorname{diag}(d)V^\top$, then*

$$U\operatorname{diag}\{\theta_\lambda(d)\}V^\top \in \arg\min_C \left\{ \frac{1}{2}\|C - Z\|_F^2 + P_\lambda(C) \right\}.$$

Proposition 27 needs no separate monotonicity assumption. Scalar proximal optimality ensures that the selected minimizers preserve singular-value order.

Proof. For $z_1 > z_2 \ge 0$, let $c_i = \theta_\lambda(z_i)$. Comparing each scalar minimizer with the other and adding gives

$$\tfrac{1}{2}(c_1 - z_1)^2 + \tfrac{1}{2}(c_2 - z_2)^2 \le \tfrac{1}{2}(c_2 - z_1)^2 + \tfrac{1}{2}(c_1 - z_2)^2, \qquad (c_1 - c_2)(z_1 - z_2) \ge 0.$$

Thus every selected map is nondecreasing. Let $C = \widetilde{U}\operatorname{diag}(c)\widetilde{V}^\top$. Von Neumann's trace inequality gives

$$\langle C, Z \rangle \le \sum_j c_j d_j,$$

with equality when the singular vectors are aligned. Hence

$$\frac{1}{2}\|C - Z\|_F^2 + P_\lambda(C) \geq \sum_j \left[\frac{1}{2}(c_j - d_j)^2 + p_\lambda(c_j)\right].$$

The coordinatewise minimum is attained at $c_j = \theta_\lambda(d_j)$. Monotonicity preserves the nonincreasing order of the singular values. Aligning the singular vectors attains equality. This completes the proof. □

The next theorem gives rank recovery whenever the scalar map has a zero region and an identity region. Let $0 < \lambda_{\mathrm{lo}} \leq \lambda_{\mathrm{hi}}$ denote their two thresholds.

Theorem 14. *If*

$$\theta_\lambda(z) = 0 \quad \textit{for } 0 \leq z \leq \lambda_{\mathrm{lo}}, \qquad \theta_\lambda(z) = z \quad \textit{for } z > \lambda_{\mathrm{hi}},$$

and

$$\eta_t < \lambda_{\mathrm{lo}}, \qquad \sigma_r(C^\star) > \lambda_{\mathrm{hi}} + \eta_t,$$

then the computed update has rank r and

$$\|C_{t+1} - C^\star\|_F \leq 2\sqrt{2r}\eta_t.$$

Theorem 14 shows that rank identification depends on the two scalar threshold regions. SCAD has $(\lambda_{\mathrm{lo}}, \lambda_{\mathrm{hi}}) = (\lambda, a\lambda)$ and MCP has $(\lambda, b\lambda)$.

Proof. Weyl's inequality gives

$$\sigma_{r+1}(Z_t) \leq \eta_t < \lambda_{\mathrm{lo}} \quad \sigma_r(Z_t) \geq \sigma_r(C^\star) - \eta_t > \lambda_{\mathrm{hi}}.$$

The map keeps the first r singular values and sets the others to zero, giving a best rank-r approximation of Z_t. The error bound follows as in the proof of Theorem 4. This completes the proof. □

The descent and stationarity arguments also extend to this penalty when the matrix proximal update is exact. Let

$$F_\lambda(C) = f(C) + P_\lambda(C).$$

Proposition 28. *Every exact proximal update in Proposition 27 satisfies*

$$F_\lambda(C_t) - F_\lambda(C_{t+1}) \geq \frac{\tau}{2(1+\tau)}\|C_{t+1} - C_t\|_F^2.$$

Every cluster point is stationary.

Proposition 28 does not require continuity of the selected proximal map. It concerns cluster points, not convergence of the whole sequence. Rank identification still depends on the threshold regions in Theorem 14.

Proof. Exact proximal optimality gives

$$\langle \nabla f(C_t), \Delta_t \rangle + \frac{1}{2}\|\Delta_t\|_F^2 + P_\lambda(C_{t+1}) - P_\lambda(C_t) \leq 0.$$

Smoothness of f with constant $(1+\tau)^{-1}$ gives descent. The objective is bounded below, hence $\sum_t \|C_{t+1} - C_t\|_F^2 < \infty$. If $C_{t_j} \to \bar{C}$, then $C_{t_j+1} \to \bar{C}$. Let $Z_j = C_{t_j} - \nabla f(C_{t_j})$ and $\bar{Z} = \bar{C} - \nabla f(\bar{C})$. Continuity of ∇f gives $Z_j \to \bar{Z}$. For every matrix D, exact proximal minimality and lower semicontinuity of P_λ give

$$\begin{aligned}\tfrac{1}{2}\left\|\bar{C} - \bar{Z}\right\|_F^2 + P_\lambda(\bar{C}) &\leq \liminf_j \left\{\tfrac{1}{2}\left\|C_{t_j+1} - Z_j\right\|_F^2 + P_\lambda(C_{t_j+1})\right\} \\ &\leq \tfrac{1}{2}\left\|D - \bar{Z}\right\|_F^2 + P_\lambda(D).\end{aligned}$$

Thus $\bar{C}$ is a proximal minimizer at $\bar{Z}$. Fermat's rule gives

$$0 \in \nabla f(\bar{C}) + \partial P_\lambda(\bar{C}).$$

This completes the proof. □

S17 CERTIFIED CURVATURE BOUNDS AND MASTER CHOICE

The setup does not require exact local largest eigenvalues. Certified upper bounds are sufficient. This section also states how the metric changes with the selected master.

Suppose worker s returns $L_s \geq \lambda_{\max}(\widehat{\Sigma}_s)$ and let

$$\bar{L} = \sum_s w_s L_s, \qquad H_L = \widehat{\Sigma}_1 + (1+\tau)\bar{L} I_p.$$

The next proposition gives the same majorization property as Lemma 1.

Proposition 29. *The metric H_L satisfies*

$$H_L - \widehat{\Sigma} \succeq \tau \bar{L} I_p, \qquad 0 \preceq H_L^{-1/2} \widehat{\Sigma} H_L^{-1/2} \preceq (1+\tau)^{-1} I_p.$$

Proposition 29 permits certified spectral upper bounds at setup. A power-iteration Rayleigh quotient alone is a lower estimate and does not certify majorization. Corollary 29 converts a certified relative error bound to an upper bound.

Proof. The worker bounds give

$$\widehat{\Sigma} = \sum_s w_s \widehat{\Sigma}_s \preceq \sum_s w_s L_s I_p = \bar{L} I_p.$$

Hence

$$H_L - \widehat{\Sigma} = \widehat{\Sigma}_1 + (1+\tau)\bar{L} I_p - \widehat{\Sigma} \succeq \tau \bar{L} I_p.$$

Also $H_L \succeq (1+\tau)\bar{L} I_p$, which gives the transformed upper bound by congruence. This completes the proof. □

Any worker can serve as master. The deterministic theory depends on the corresponding metric through κ and the transformed signal.

For master j, define

$$H_j = \widehat{\Sigma}_j + (1+\tau)\bar{\ell} I_p.$$

Proposition 30. *For every master j, Lemma 1, Proposition 3, Theorem 2, Corollary 2, and Proposition 6 hold with H_j and the corresponding transformed objective.*

Proposition 30 shows that validity does not depend on selecting a full-rank or representative master. Statistical constants and finite-round behavior can still vary across masters.

Proof. For every j, $\widehat{\Sigma}_j \succeq 0$, $\widehat{\Sigma} \preceq \bar{\ell} I_p$, and $H_j \succeq (1+\tau)\bar{\ell} I_p$, as required in Lemma 1. Also $H_j v = \gamma v$ for $v \in \ker(X)$, permitting Lemma 4. The listed results apply in the transformed coordinate. This completes the proof. □

Changing the master changes the penalty scale as well as the metric. The following comparison bounds that change in scale before any comparison of prediction errors.

Lemma 26. *For masters j and k and every B,*

$$\left\| H_j^{1/2} B \right\|_* \leq \sqrt{\frac{\gamma + \ell_j}{\gamma}} \left\| H_k^{1/2} B \right\|_*,$$

and the same relation holds with the Frobenius norm.

Lemma 26 gives a deterministic sensitivity bound. It can be loose when local Gram matrix eigenspaces are aligned, which motivates an empirical comparison across master choices.

Proof. The metric bounds

$$\gamma I_p \preceq H_k, \qquad H_j \preceq (\gamma + \ell_j) I_p$$

imply

$$H_j \preceq \frac{\gamma + \ell_j}{\gamma} H_k.$$

Apply Lemma 7. This completes the proof. □

Local curvature can improve the transformed condition number when master and pooled Gram matrices are aligned. For $A \succ 0$, let $\mathrm{cond}(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$. The next proposition gives an exact calculation in a proportional case.

Proposition 31. *Suppose $\widehat{\Sigma}_1 = c\widehat{\Sigma}$ for $c \geq 0$ and $\widehat{\Sigma} \succ 0$. Then*

$$\mathrm{cond}(M) = \mathrm{cond}(\widehat{\Sigma}) \frac{\gamma + c\lambda_{\min}(\widehat{\Sigma})}{\gamma + c\lambda_{\max}(\widehat{\Sigma})}.$$

For $c > 0$ and nonconstant eigenvalues, this is strictly smaller than $\mathrm{cond}(\widehat{\Sigma})$.

Proposition 31 gives one setting in which the master curvature improves optimization geometry. No condition-number dominance is claimed for arbitrary heterogeneous Gram matrices.

Proof. The two Gram matrices commute and have common eigenvectors. If λ_j is an eigenvalue of $\widehat{\Sigma}$, then the corresponding eigenvalue of M is $\frac{\lambda_j}{\gamma + c\lambda_j}$. This function is increasing in λ_j. Hence

$$\mathrm{cond}(M) = \frac{\lambda_{\max}/(\gamma + c\lambda_{\max})}{\lambda_{\min}/(\gamma + c\lambda_{\min})},$$

which is the stated formula. The ratio multiplying $\mathrm{cond}(\widehat{\Sigma})$ is below one under the strict premises. This completes the proof. □

Without alignment of the Gram matrices, the local term reduces every generalized Rayleigh quotient relative to the isotropic metric, but it need not improve their ratio.

Proposition 32. *Let $H_0 = \gamma I_p$. For every nonzero v,*

$$\frac{v^\top \widehat{\Sigma} v}{v^\top H v} \leq \frac{v^\top \widehat{\Sigma} v}{v^\top H_0 v}.$$

This ordering alone gives no deterministic ordering between $\mathrm{cond}(M)$ and $\mathrm{cond}(H_0^{-1/2} \widehat{\Sigma} H_0^{-1/2})$.

Proposition 32 explains why the isotropic ablation is necessary. The local metric has a prediction interpretation and can improve aligned designs, but arbitrary heterogeneity does not guarantee faster conditioning.

Proof. Because $H = H_0 + \widehat{\Sigma}_1 \succeq H_0$, $v^\top H v \geq v^\top H_0 v$. The numerators are equal and nonnegative, which proves the pointwise relation.

For the condition numbers, let $p = 2$ and $\widehat{\Sigma} = \mathrm{diag}(4, 1)$. If $\widehat{\Sigma}_1 = \widehat{\Sigma}$, then

$$\mathrm{cond}(M) = 4\frac{\gamma + 1}{\gamma + 4} < 4 = \mathrm{cond}(H_0^{-1/2} \widehat{\Sigma} H_0^{-1/2}).$$

If $\widehat{\Sigma}_1 = \mathrm{diag}(0, 1)$, then

$$\mathrm{cond}(M) = 4\frac{\gamma + 1}{\gamma} > 4 = \mathrm{cond}(H_0^{-1/2} \widehat{\Sigma} H_0^{-1/2}).$$

Both master Gram matrices are positive semidefinite. Thus neither condition number uniformly bounds the other. This completes the proof. □

A well-conditioned pooled Gram matrix alone does not bound κ away from zero. To see this, let $p = 2$ and let $k \geq 2$ be an integer. Consider two workers with $w_1 = 1/k$ and $w_2 = 1 - 1/k$. Let

$$\widehat{\Sigma}_1 = \mathrm{diag}(k, 0), \qquad \widehat{\Sigma}_2 = \mathrm{diag}\{0, k/(k-1)\}.$$

Then $\widehat{\Sigma} = I_2$, $\bar{\ell} = 2$, $\gamma = 2(1+\tau)$, and

$$H = \mathrm{diag}(k + \gamma, \gamma), \qquad M = \mathrm{diag}\{(k+\gamma)^{-1}, \gamma^{-1}\}.$$

Thus $\kappa = (k+\gamma)^{-1} \to 0$, although the pooled Gram matrix has condition number one. Pooled conditioning alone does not give a uniform positive lower bound for the curvature in Theorem 6. The example establishes this curvature dependence; it gives no lower bound for estimation risk.

The stabilization factor τ controls descent margin and metric scale in opposite directions. As τ increases, the smoothness upper bound $(1+\tau)^{-1}$ decreases and the descent coefficient $\tau/\{2(1+\tau)\}$ increases. At the same time, $\gamma = (1+\tau)\bar{\ell}$ and H increase in Loewner order. Thus τ changes both the optimization step and the penalized prediction metric, and its empirical effect must be examined.

S18 End-to-end communication accounting

Proposition 1 counts one fit at fixed tuning values. Tuning changes the comparison because sufficient statistics can be reused across all candidate estimators. We count the communication for fitting all candidates and for distributed validation separately.

Suppose J gradient-based candidate fits use $T_1^{(j)} + T_2^{(j)}$ rounds and reuse the same setup scalar. Let

$$T_{\mathrm{total}} = \sum_{j=1}^{J} \{T_1^{(j)} + T_2^{(j)}\}.$$

The following proposition compares the communication used to fit all tuning candidates.

Proposition 33. *The gradient protocol transmits*

$$S_{\mathrm{grad,tune}} = 1 + 2pqT_{\mathrm{total}}$$

scalars per remote worker. Sufficient-statistic aggregation transmits

$$S_{\mathrm{stat,tune}} = p(p+1)/2 + pq$$

scalars per remote worker and reuses them for every candidate. The gradient protocol is smaller if and only if

$$T_{\mathrm{total}} < \frac{1}{2} + \frac{p+1}{4q} - \frac{1}{2pq}.$$

Proposition 33 applies the break-even inequality to the rounds used to fit all tuning candidates. The main numerical table reports one fit at the selected tuning values.

Proof. Setup costs one scalar and each gradient round costs $2pq$ entries. After the sufficient-statistic upload, the master can fit all candidates without further worker communication. Comparing the counts gives the bound for T_{total}. This completes the proof. □

The total network count multiplies both per-worker quantities by the number of remote workers. The break-even inequality is unchanged.

Corollary 15. *With one master and $m-1$ remote workers,*

$$S_{\mathrm{grad,network}} = (m-1)(1 + 2pqT), \quad S_{\mathrm{stat,network}} = (m-1)\{p(p+1)/2 + pq\}.$$

Their ordering is the same as in Proposition 1.

Corollary 15 confirms that unequal sample sizes do not alter scalar dimensions. They affect the aggregation weights and computational loads, not the per-worker matrix sizes.

Proof. Every remote worker sends objects of the stated dimensions. Sum the per-worker counts over $m-1$ workers. Multiplication by the same positive number preserves the comparison. This completes the proof. □

Distributed validation requires the final candidate at each worker and one returned local loss. This cost should be stated in a full system account.

Proposition 34. *If the final $p \times q$ coefficient matrix for each of J candidates is broadcast for distributed validation and every remote worker returns one local validation loss, the tuning count becomes $1 + 2pqT_{\text{total}} + J(pq+1)$ scalars per remote worker.*

Proposition 34 isolates the cost of distributed validation. If every candidate is already available at each worker, its Jpq broadcast term can be omitted. The result does not include raw validation transfer or model-specific preprocessing communication.

Proof. For each candidate, the master sends pq coefficient entries and each worker returns one scalar loss. Add $J(pq+1)$ to the count in Proposition 33. This completes the proof. □

The sufficient-statistic protocol also requires the response sum of squares only when the full objective value is needed. One scalar $\|Y_s\|_F^2 / n_s$ from each worker determines this coefficient-independent term. It does not affect the matrix-order comparison, and the optimizer needs only $\widehat{\Sigma}_s$ and $\widehat{Z}_s$.

S19 Fixed-design identifiability and coefficient lower bounds

Theorem 10 and Proposition 5 distinguish prediction from coefficient estimation. This section proves that the distinction is unavoidable when the pooled design is singular.

Proposition 35. *If $\operatorname{rank}(X) < p$ and $q \geq 1$, then for every $R > 0$ there are rank-one matrices B_0, B_1 such that*

$$XB_0 = XB_1, \qquad \|B_1 - B_0\|_F > R.$$

Proposition 35 shows that low rank alone cannot identify coefficient directions in the pooled null space. A coefficient theorem for $N < p$ needs additional predictor structure beyond rank.

Proof. Choose a nonzero $h \in \ker(X)$ and a unit vector $v \in \mathbb{R}^q$. Let

$$B_0 = hv^\top, \qquad B_1 = (1+t)hv^\top.$$

Then $XB_1 = 0 = XB_0$. Both matrices have rank one, and $\|B_1 - B_0\|_F = t\,\|h\|_2$. Choose $t > R/\|h\|_2$. This completes the proof. □

Under a full-rank design, the prediction lower bound gives a coefficient lower bound. The design upper eigenvalue determines the conversion.

Corollary 16. *Suppose $\widehat{\Sigma} \preceq L_X I_p$, $\operatorname{rank}(X) = p$, and $1 \leq r \leq p \wedge q$. Under the Gaussian model in Proposition 5,*

$$\inf_{\widetilde{B}} \sup_{\operatorname{rank}(B) \leq r} \mathbb{E}_B \left\| \widetilde{B} - B \right\|_F^2 \geq \frac{c\sigma^2 r(p+q-r)}{NL_X}.$$

Corollary 16 gives the fixed-design benchmark for Frobenius coefficient loss. The Stage I upper bound has the same low-rank dimension under bounded metric and curvature factors.

Proof. For every matrix A,

$$\frac{1}{N}\|XA\|_F^2 = \left\langle \widehat{\Sigma}A, A \right\rangle \leq L_X \|A\|_F^2.$$

Hence coefficient loss is at least prediction loss divided by L_X. Apply Proposition 5 with $d = p$. This completes the proof. □

The rank of the prediction matrix can be smaller than the coefficient rank when the design is singular. The next lemma records the exact relation.

Lemma 27. *For every B,*

$$\operatorname{rank}(XB) \le \operatorname{rank}(B) \wedge \operatorname{rank}(X).$$

If X has full column rank, then $\operatorname{rank}(XB) = \operatorname{rank}(B)$.

Lemma 27 explains why the minimax prediction dimension uses $d = \operatorname{rank}(X)$. Pooled injectivity preserves the coefficient rank in prediction for every B.

Proof. The rank inequality for a matrix product gives the upper bound. If X has full column rank, it has a left inverse X^+ satisfying $X^+X = I_p$. Hence

$$\operatorname{rank}(B) = \operatorname{rank}(X^+XB) \le \operatorname{rank}(XB) \le \operatorname{rank}(B).$$

This completes the proof. □

S20 Deterministic model approximation

The deterministic theory extends to a nonzero approximation residual. This section states the exact replacement in the score and does not impose a stochastic model on the residual.

Suppose

$$Y = XB^\star + A + E,$$

where $A \in \mathbb{R}^{N\times q}$ is fixed conditional on X. Partition A by worker rows and define

$$\widetilde{W}_A = H^{-1/2}\frac{X^\top(A+E)}{N}, \qquad \varepsilon_A = \left\|\widetilde{W}_A\right\|_{\mathrm{op}}.$$

The following proposition gives the exact extension.

Proposition 36. *Theorems 6, 10, and 4 remain valid after replacing $\widetilde{W}$ and ε by $\widetilde{W}_A$ and ε_A.*

Proposition 36 shows that deterministic model error enters through its transformed score. A residual orthogonal to the design does not affect the estimator.

Proof. The gradient at $C^\star$ becomes

$$\nabla f(C^\star) = -H^{-1/2}X^\top(A+E)/N = -\widetilde{W}_A.$$

Every proof of the three cited theorems uses the noise only through this gradient and its operator norm. The update identity becomes

$$Z_t = C^\star + (I_p - M)(C_t - C^\star) + \widetilde{W}_A.$$

Substitution of these two identities into the three cited proofs gives the stated bounds and rank conclusion. This completes the proof. □

The approximation and stochastic noise contributions can be separated directly. The operator-norm triangle inequality gives

$$\varepsilon_A \le \left\|H^{-1/2}X^\top A/N\right\|_{\mathrm{op}} + \varepsilon.$$

This bound separates a deterministic bias term from the stochastic score used in Corollary 1.

If the approximation residual is orthogonal to the pooled predictor space, then $X^\top A = 0$ gives $\widetilde{W}_A = \widetilde{W}$. The pooled gradient is unchanged at every B, which leaves every proposed iterate unchanged from the model without A. Thus only the projection of model error onto the predictor space affects the method.

S21 FIXED POINTS AND STRICT RANK CERTIFICATES

Corollary 2 gives convergence of the complete exact Stage II sequence, including singular pooled designs. A strict singular-value gap at its limit makes the computed rank eventually constant.

Let

$$\mathcal{T}_2(C) = \mathrm{prox}_{P_2}\{C - \nabla f(C)\}.$$

The SCAD and MCP scalar threshold maps are continuous, and the corresponding spectral maps are continuous.

Proposition 37. *If $C_t \to \bar{C}$ along the complete Stage II sequence, then*

$$\bar{C} = \mathcal{T}_2(\bar{C}), \qquad 0 \in \partial F_2(\bar{C}).$$

Proposition 37 identifies the limiting object as a fixed point of the exact update used by the code.

Proof. By Theorem 2, $\|C_{t+1} - C_t\|_F \to 0$, hence $C_{t+1} \to \bar{C}$. Continuity of ∇f and the spectral threshold map gives $\bar{C} = \mathcal{T}_2(\bar{C})$. Proximal optimality yields $0 \in \nabla f(\bar{C}) + \partial P_2(\bar{C})$. This completes the proof. □

If the limiting update matrix has a strict gap around the lower threshold, the selected rank is eventually constant.

For a candidate limit $\bar{C}$, define

$$\bar{Z} = \bar{C} - \nabla f(\bar{C}).$$

Proposition 38. *Suppose $C_t \to \bar{C}$. If, for some $1 \le k < p \wedge q$ and $\zeta > 0$,*

$$\sigma_k(\bar{Z}) \ge \lambda_2 + \zeta, \qquad \sigma_{k+1}(\bar{Z}) \le \lambda_2 - \zeta,$$

then $\mathrm{rank}(C_t) = k$ for all sufficiently large t.

Proposition 38 gives an asymptotic rank conclusion from a strict observable gap. It does not identify k with the true rank without the signal conditions in Theorem 4.

Proof. Continuity of $C \mapsto C - \nabla f(C)$ gives $Z_t \to \bar{Z}$. Hence $\|Z_t - \bar{Z}\|_{\mathrm{op}} < \zeta/2$ for all sufficiently large t. Weyl's inequality gives

$$\sigma_k(Z_t) > \lambda_2, \qquad \sigma_{k+1}(Z_t) < \lambda_2.$$

Lemma 25 gives rank k for the next update. Reindexing proves the statement for C_t. This completes the proof. □

The stronger upper-knot gap gives eventual absence of shrinkage on the kept singular values.

Corollary 17. *Under Proposition 38, if*

$$\sigma_k(\bar{Z}) \ge a\lambda_2 + \zeta,$$

then every sufficiently late SCAD update keeps its first k singular values unchanged. For MCP, replace a by b.

Corollary 17 distinguishes stable rank from stable removal of shrinkage at the limit.

Proof. For all sufficiently large t, Weyl's inequality gives $\sigma_k(Z_t) > a\lambda_2$ for SCAD. The upper branch of the scalar map is the identity. The MCP proof is the same. This completes the proof. □

S22 FINITE-ITERATION RANK AND CONVERGENCE CERTIFICATES

The preceding prediction bound concerns a single computed update. This section gives a condition at a finite iteration that also controls every later update with the same penalty and metric. The condition uses a certified lower bound on observed-space curvature, not a distance from the unknown coefficient matrix.

Let $t \geq T_1$ and $1 \leq k \leq d \wedge q$. Let $0 < \underline{\kappa} \leq \kappa_+$ be a certified bound and let $\omega = 1 - \underline{\kappa}$. For $R > 0$ with $\sigma_k(Z_t) > \omega R$, define

$$\vartheta_R = \frac{\sigma_{k+1}(Z_t) + \omega R}{\sigma_k(Z_t) - \omega R}, \qquad \rho_R = \frac{\omega}{1 - \vartheta_R}.$$

The following theorem concerns the exact SCAD sequence continued after iteration t.

Theorem 15. *Suppose*

$$\sigma_{k+1}(Z_t) + \omega R < \lambda_2, \qquad \sigma_k(Z_t) - \omega R > a\lambda_2, \qquad \vartheta_R < \underline{\kappa},$$
$$\|C_{t+1} - C_t\|_F \leq (1 - \rho_R)R.$$

There is a unique fixed point $\overline{C}$ in $\{C : \Pi_X C = C,\ \|C - C_t\|_F \leq R\}$, and, for every $j \geq 0$,

$$\left\|C_{t+j} - \overline{C}\right\|_F \leq \frac{\rho_R^j}{1 - \rho_R} \|C_{t+1} - C_t\|_F .$$

Every update C_{t+j} with $j \geq 1$ has rank k and its nonzero singular values are unchanged by the SCAD map.

Theorem 15 guarantees the future rank of the computed sequence. The limit is the unique least-squares fit of rank at most k on the observed predictor space, but this conclusion does not identify k with a population rank. Unlike Theorem 7, it does not use $B^\star$, ε, or an initial error relative to the signal. Evaluating this certificate requires a valid curvature bound in addition to the singular values and update difference.

Proof. Let $\mathcal{B} = \{C : \Pi_X C = C,\ \|C - C_t\|_F \leq R\}$ and define $Z(C) = C - \nabla f(C)$ and $\mathcal{T}(C) = \mathrm{prox}_{P_2}\{Z(C)\}$. For C, D in the observed predictor space,

$$Z(C) - Z(D) = (I_p - M)(C - D), \qquad \|Z(C) - Z(D)\|_F \leq \omega \|C - D\|_F .$$

The second inequality in Lemma 1 and $\underline{\kappa} \leq \kappa_+ \leq (1 + \tau)^{-1} < 1$ give $0 < \omega < 1$. For $C \in \mathcal{B}$, singular-value perturbation gives

$$\sigma_k\{Z(C)\} \geq \sigma_k(Z_t) - \omega R > a\lambda_2, \qquad \sigma_{k+1}\{Z(C)\} \leq \sigma_{k+1}(Z_t) + \omega R < \lambda_2.$$

Hence $\mathcal{T}(C) = \mathcal{P}_k\{Z(C)\}$ and $\mathrm{rank}\{\mathcal{T}(C)\} = k$. Let $A = \mathcal{T}(C)$, $B = \mathcal{T}(D)$, $N_C = Z(C) - A$, and $N_D = Z(D) - B$. Lemma 6, first at A and then at B, gives

$$\begin{aligned}
\langle Z(C) - Z(D), A - B\rangle &= \|A - B\|_F^2 + \langle N_C - N_D, A - B\rangle \\
&\geq \left\{1 - \frac{\|N_C\|_{\mathrm{op}}}{2\sigma_k(A)} - \frac{\|N_D\|_{\mathrm{op}}}{2\sigma_k(B)}\right\} \|A - B\|_F^2 \\
&\geq (1 - \vartheta_R) \|A - B\|_F^2 .
\end{aligned}$$

For $k = p \wedge q$ both normal components are zero, and the same bound holds. Cauchy–Schwarz implies

$$\|\mathcal{T}(C) - \mathcal{T}(D)\|_F \leq \frac{\|Z(C) - Z(D)\|_F}{1 - \vartheta_R} \leq \rho_R \|C - D\|_F .$$

The premise $\vartheta_R < \underline{\kappa}$ gives $0 < \rho_R < 1$. The row-space identity in Lemma 3 also gives $\Pi_X \mathcal{T}(C) = \mathcal{T}(C)$. For $C \in \mathcal{B}$,

$$\begin{aligned}
\|\mathcal{T}(C) - C_t\|_F &\leq \|\mathcal{T}(C) - \mathcal{T}(C_t)\|_F + \|\mathcal{T}(C_t) - C_t\|_F \\
&\leq \rho_R R + \|C_{t+1} - C_t\|_F \leq R.
\end{aligned}$$

Thus $\mathcal{T}$ maps the closed ball into itself and is a contraction there. The contraction mapping theorem gives a unique fixed point $\overline{C} \in \mathcal{B}$ and convergence of the sequence beginning at C_t. Repeated contraction gives

$$\|C_{t+j+1} - C_{t+j}\|_F \le \rho_R^j \|C_{t+1} - C_t\|_F .$$

Summing this inequality from j to infinity proves the stated error bound. The singular-value inequalities hold throughout $\mathcal{B}$, and hence every subsequent update has rank k and is unshrunk. At the limit the update difference is zero and its singular-value ratio is less than $\underline{\kappa} \le \kappa_+$. The quadratic comparison in the proof of Theorem 8 then proves that $\overline{C}$ is the unique global least-squares fit of rank at most k on the observed predictor space. This completes the proof. □

The certificate can hold when all local designs and the pooled design are singular. The next construction also shows that a finite Stage I initialization can enter the certified region.

Proposition 39. *For every $d_0 \ge 1$, $p > 2d_0$, $q \ge 1$, $1 \le k \le 2d_0 \wedge q$, $\tau > 0$, and $\lambda_2 > 0$, there are two workers with $n_1 = n_2 = d_0$ and a rank-k signal for which Theorem 15 holds after finitely many Stage I updates from zero. Each local Gram matrix and the pooled Gram matrix are singular. The conditions continue to hold on an open set of response perturbations.*

Proposition 39 shows that local and pooled singularity do not make the certificate empty. It does not give the probability that the certificate holds in the reported experiments.

Proof. Choose orthogonal projections P_1, P_2 of rank d_0 with $P_1 P_2 = 0$ and let $P = P_1 + P_2$. There are $d_0 \times p$ matrices X_s with

$$X_s^\top X_s / d_0 = 2P_s, \qquad \widehat{\Sigma} = P, \qquad \gamma = 2(1+\tau), \qquad H = 2P_1 + \gamma I_p.$$

For example, multiply any orthonormal row basis for the range of P_s by $\sqrt{2d_0}$. The observed predictor space is the range of P, and

$$M = (2+\gamma)^{-1} P_1 + \gamma^{-1} P_2, \qquad \kappa_+ = (2+\gamma)^{-1} > 0.$$

Choose $C^\star$ with $PC^\star = C^\star$, rank k, and $s = \sigma_k(C^\star) > a\lambda_2$. Let $B^\star = H^{-1/2} C^\star$ and $Y = XB^\star$. Then $\widetilde{W} = 0$, $\nabla f(C^\star) = 0$, and the exact Stage II map fixes $C^\star$.

Let $\underline{\kappa} = \kappa_+$ and $\omega = 1 - \kappa_+$. Choose $R > 0$ such that

$$R < \min\left\{ \frac{\lambda_2}{\omega}, \frac{s - a\lambda_2}{\omega}, \frac{\kappa_+ s}{\omega(1+\kappa_+)} \right\}.$$

At the center $C^\star$, the upper singular-value remainder is zero. The three strict inequalities in Theorem 15 follow from this choice, and the update difference is zero, strictly less than $(1-\rho_R)R$. Singular values, the SCAD map in (5), and the affine gradient are continuous. Thus the same four conditions hold for every starting matrix in a sufficiently small observed-space neighborhood of $C^\star$.

In the noiseless model, Theorem 1 gives

$$\|C_{T_1} - C^\star\|_F \le \frac{2\sqrt{2k}\lambda_1}{\kappa_+} + \frac{2\left\|\widehat{C}\right\|_F}{\sqrt{\kappa_+}(T_1+1)}.$$

Choose $\lambda_1 > 0$ sufficiently small and then a finite T_1 sufficiently large that this upper bound is within that neighborhood. This proves entry from zero without assuming the desired initial error as a model condition.

With these choices fixed, every one of the finitely many Stage I operations is continuous in Y: gradients are affine, extrapolation is linear, and singular-value soft thresholding is continuous. The first Stage II map and its singular values are also continuous. The strict inequalities at iteration T_1 thus persist for all responses in a sufficiently small open neighborhood of $XB^\star$. This neighborhood contains nonzero noise matrices. Because $p > 2d_0$, the two local matrices and their pooled matrix are singular throughout. This completes the proof. □

A deterministic use of Theorem 15 requires a certified bound $\underline{\kappa}$. Computing it from pooled sufficient statistics incurs their upload cost. The population bound $k_{\delta,0}$ in Proposition 61 is a different option when the stated population constants are available; its certificate is then valid on the probability event established in that proposition. An unverified fitted eigenvalue approximation cannot replace either bound. With finite-precision calculations, the curvature bound and spectral inequalities must also account for numerical error; the numerical-rank statement in Proposition 56 alone does not verify these inequalities. A curvature-verification procedure using additional gradient queries must count those queries in Proposition 33.

S23 SEMI-ALGEBRAIC CONVERGENCE DETAILS

This section verifies the structural premise used for full-sequence convergence in Corollary 2. It records why the matrix SCAD and MCP objectives satisfy the Kurdyka–Łojasiewicz property.

The graph of the scalar SCAD penalty is a finite union of sets defined by polynomial equalities and inequalities. The same statement holds for MCP. Hence both scalar penalties are semi-algebraic. Here $P_{\lambda,a}$ and $P^{\mathrm{MCP}}_{\lambda,b}$ denote the scalar SCAD and MCP penalties with shape parameters a and b, respectively.

Lemma 28. *The functions*

$$C \mapsto \sum_j P_{\lambda,a}\{\sigma_j(C)\}, \qquad C \mapsto \sum_j P^{\mathrm{MCP}}_{\lambda,b}\{\sigma_j(C)\}$$

are continuous and semi-algebraic.

Lemma 28 gives the Kurdyka–Łojasiewicz property used by the convergence theorem of Attouch et al. (2013).

Proof. Continuity follows from continuity of the singular values and scalar penalties. For semi-algebraicity, let U, V, d satisfy

$$U^\top U = I, \qquad V^\top V = I, \qquad C = U\mathrm{diag}(d)V^\top, \qquad d_1 \geq \cdots \geq d_{p\wedge q} \geq 0.$$

These constraints are polynomial. Each penalty has finitely many polynomial pieces, hence adding $\sum_j P_\lambda(d_j) \leq t$ defines a semi-algebraic set. Its projection onto (C, t) is the spectral penalty epigraph, which is semi-algebraic by the Tarski–Seidenberg theorem. This completes the proof. □

The complete objective is also semi-algebraic because its smooth part is quadratic.

Corollary 18. *The SCAD and MCP objectives in transformed coordinates are proper, continuous, and semi-algebraic.*

Corollary 18 establishes the required objective structure without placing it in a statistical condition.

Proof. Add the quadratic polynomial in Lemma 11 to the continuous semi-algebraic penalty in Lemma 28. Finite sums preserve both properties. This completes the proof. □

Strong pooled curvature makes the Stage II sublevel set bounded.

Lemma 29. *If $\kappa > 0$, then every set $\{C : F_2(C) \leq h\}$ is bounded.*

Lemma 29 controls every sublevel set under pooled invertibility. For the exact computed sequence, Lemma 4 gives boundedness even when these unrestricted sublevel sets are unbounded.

Proof. Lemma 11 and completion of the square give

$$f(C) \geq \frac{\kappa}{4}\|C\|_F^2 - K$$

for a finite constant K depending on M and $H^{-1/2}\widehat{Z}$. The SCAD penalty is nonnegative, hence

$$F_2(C) \geq \frac{\kappa}{4}\|C\|_F^2 - K.$$

Every finite sublevel set is bounded. This completes the proof. □

The sufficient-decrease and relative-error properties have already been proved. The next proposition states the resulting finite-length conclusion explicitly.

Proposition 40. *The exact SCAD and MCP Stage II sequences satisfy*

$$\sum_{t=T_1}^{\infty} \|C_{t+1} - C_t\|_F < \infty.$$

Proposition 40 is stronger than square summability. It implies that the complete sequence is Cauchy.

Proof. Theorem 2 and Lemma 10 give sufficient decrease and relative error. Lemma 4 gives boundedness for both penalties. Corollary 18 gives continuity and the Kurdyka–Łojasiewicz property. Theorem 2.9 of Attouch et al. (2013) gives finite length. This completes the proof. □

S24 Further relations among the statistical targets

The proposed convex target penalizes $H^{1/2}B$, while centralized coefficient nuclear-norm regression penalizes B and adaptive prediction nuclear-norm regression penalizes $\widehat{\Sigma}^{1/2}B$. This section compares the three deterministic criteria without claiming that their minimizers coincide.

Define

$$\mathcal{R}_{\mathrm{coef}}(B) = \|B\|_*, \qquad \mathcal{R}_{\mathrm{pred}}(B) = \left\|\widehat{\Sigma}^{1/2}B\right\|_*, \qquad \mathcal{R}_H(B) = \left\|H^{1/2}B\right\|_*.$$

The metric bounds and Lemma 7 give, for every B,

$$\sqrt{\gamma}\,\mathcal{R}_{\mathrm{coef}}(B) \le \mathcal{R}_H(B) \le \sqrt{\gamma+\ell_1}\,\mathcal{R}_{\mathrm{coef}}(B).$$

Thus the proposed penalty remains a norm equivalent to the coefficient nuclear norm, with directional scaling determined by the master Gram matrix.

The relation to the pooled prediction penalty is sharper when κ is controlled. If $\kappa > 0$, Proposition 3 gives

$$\sqrt{1+\tau}\,\mathcal{R}_{\mathrm{pred}}(B) \le \mathcal{R}_H(B) \le \kappa^{-1/2}\mathcal{R}_{\mathrm{pred}}(B).$$

This comparison shows that the metric penalty approximates prediction-space regularization when κ is bounded away from zero.

The transformed objective is invariant to orthogonal rotations of the predictor coordinates when the data and master are rotated in the same way.

Proposition 41. *Let $Q \in \mathbb{R}^{p\times p}$ be orthogonal and replace X_s by X_sQ and B by $Q^\top B$. Then the loss, metric penalty, transformed iterates, fitted responses, and scalar communication count are unchanged after the corresponding coordinate rotation.*

Proposition 41 confirms that the method does not depend on an arbitrary orthogonal basis for the predictor space.

Proof. The local Gram matrix becomes $Q^\top\widehat{\Sigma}_sQ$, preserving its largest eigenvalue. The metric and its square root become $Q^\top HQ$ and $Q^\top H^{1/2}Q$. Thus

$$(Q^\top HQ)^{1/2}Q^\top B = Q^\top H^{1/2}B,$$

Orthogonality preserves singular values. Gradients and proximal updates transform by $Q^\top$; fitted responses satisfy $X_sQQ^\top B = X_sB$. Matrix dimensions remain unchanged. This completes the proof. □

General rescaling is not an invariance because it changes the chosen local prediction metric. Standardization must hence be defined from training observations, as done in the public-data analysis.

Lemma 30. *Let $R \in \mathbb{R}^{q\times q}$ be orthogonal and replace Y_s by Y_sR and B by BR. Then the proposed estimates are multiplied on the right by R, while ranks, objective values, and communication counts are unchanged.*

Lemma 30 confirms that rank estimation is invariant to an orthogonal basis change among the responses.

Proof. The loss satisfies

$$\|Y_s R - X_s B R\|_F = \|Y_s - X_s B\|_F .$$

The transformed coefficient becomes $H^{1/2}BR$, which has the same singular values as $H^{1/2}B$. Gradients and proximal updates rotate on the right. Ranks and matrix dimensions are unchanged. This completes the proof. □

S25 Summary of theoretical scope

The results distinguish prediction, coefficient estimation, stationarity, and rank recovery. When $\widehat{\Sigma}$ is singular, Theorem 10 gives a curvature-free prediction bound, while Theorem 1 gives a fast bound on the observed predictor subspace. Proposition 35 excludes uniform recovery of null-space coefficients. Under full pooled rank, Theorem 6 gives coefficient and prediction rates. Proposition 5 gives the prediction benchmark in both rank regimes. Stage II has a finite stationarity bound without a rank premise. Rank identification then depends on the lower threshold, while removal of shrinkage depends on the upper threshold.

The communication statements have the same separation. Proposition 1 concerns one fit at fixed tuning values. Proposition 33 concerns all candidate fits. Sufficient-statistic aggregation reconstructs the batch quadratic objective up to a coefficient-independent constant, and Proposition 7 shows that unrestricted exact gradients also reveal those statistics after finitely many designed queries. The proposed protocol is justified only by its finite scalar and master-storage bounds in the stated range.

The empirical comparison remains distinct from these theorems. Proposition 31 proves improvement under covariance alignment, while Proposition 32 gives no universal condition-number ordering. In 100 replications of the locally singular design, the worker-1 local metric at $\tau = .10$ has mean transformed condition number 14.32, compared with 15.57 for the isotropic metric. Its paired test MSE is lower by $.000482$ with standard error $.000071$. The mean test MSE across the eight master choices ranges from 1.09009 to 1.09047. These comparisons are specific to the studied design and do not give a universal ordering.

S26 Gaussian effective-dimension bounds

The sub-Gaussian score bound in Corollary 1 uses the upper dimensions p and q. Under Gaussian errors, the transformed covariance gives a sharper effective-dimension expression. This section derives that bound conditional on the complete design.

We first state a Gaussian matrix inequality. It applies to rectangular matrices with covariance on both sides.

Lemma 31. *Let $G \in \mathbb{R}^{p\times q}$ have independent $N(0,1)$ entries, and let $A \in \mathbb{R}^{p\times p}$ and $B \in \mathbb{R}^{q\times q}$ be fixed. For every $t > 0$, with probability at least $1 - 2e^{-t^2/2}$,*

$$\|AGB\|_{\mathrm{op}} \le \|A\|_F \|B\|_{\mathrm{op}} + \|A\|_{\mathrm{op}} \|B\|_F + t \|A\|_{\mathrm{op}} \|B\|_{\mathrm{op}} .$$

Lemma 31 replaces the two nominal dimensions by Frobenius-to-operator norm ratios. The squares of these ratios are the effective ranks of the left and right covariance matrices.

Proof. Let g_0, $g \in \mathbb{R}^p$, and $h \in \mathbb{R}^q$ be independent standard Gaussian variables, independent of G. For unit u, v, let $a = A^\top u$, $b = Bv$, and

$$X_{u,v} = a^\top G b + g_0 \|a\|_2 \|b\|_2 , \qquad Y_{u,v} = \|b\|_2 g^\top a + \|a\|_2 h^\top b.$$

Their pointwise variances agree. For a second index with vectors a', b', the covariance difference is

$$\mathrm{Cov}(X_{u,v}, X_{u',v'}) - \mathrm{Cov}(Y_{u,v}, Y_{u',v'}) = \{\|a\|_2 \|a'\|_2 - a^\top a'\}\{\|b\|_2 \|b'\|_2 - b^\top b'\} \ge 0.$$

Slepian's inequality and conditional Jensen's inequality in g_0 give

$$\mathbb{E}\,\|AGB\|_{\mathrm{op}} \le \mathbb{E}\sup_{u,v} X_{u,v} \le \mathbb{E}\sup_{u,v} Y_{u,v} \le \|B\|_{\mathrm{op}}\,\mathbb{E}\,\|Ag\|_2 + \|A\|_{\mathrm{op}}\,\mathbb{E}\,\left\|B^\top h\right\|_2 .$$

The comparison extends from finite nets to the spheres by continuity. Because $\mathbb{E}\,\|Ag\|_2 \le \|A\|_F$ and $\mathbb{E}\,\left\|B^\top h\right\|_2 \le \|B\|_F$,

$$\mathbb{E}\,\|AGB\|_{\mathrm{op}} \le \|A\|_F\,\|B\|_{\mathrm{op}} + \|A\|_{\mathrm{op}}\,\|B\|_F .$$

For concentration, the map

$$G \mapsto \|AGB\|_{\mathrm{op}}$$

has Frobenius Lipschitz constant $\|A\|_{\mathrm{op}}\,\|B\|_{\mathrm{op}}$ because

$$|\,\|AG_1B\|_{\mathrm{op}} - \|AG_2B\|_{\mathrm{op}}\,| \le \|A(G_1-G_2)B\|_{\mathrm{op}} \le \|A\|_{\mathrm{op}}\,\|B\|_{\mathrm{op}}\,\|G_1-G_2\|_F .$$

Gaussian concentration (Vershynin, 2018), with the exact tail constant in Boucheron et al. (2013, Section 5.4), gives the tail bound. This completes the proof. □

We apply the lemma first to independent spherical response noise. Conditional on the design, the transformed score has an exact Gaussian representation.

Let $G \in \mathbb{R}^{p\times q}$ have independent $N(0,1)$ entries. The symbol $\overset{d}{=}$ denotes equality in distribution.

Theorem 16. *Suppose the entries of E are independent $N(0,\sigma^2)$ conditional on X. Then*

$$\widetilde{W} \overset{d}{=} \frac{\sigma}{\sqrt{N}} M^{1/2} G,$$

conditional on X. For every $t>0$, with probability at least $1-2e^{-t^2/2}$,

$$\varepsilon \le \frac{\sigma}{\sqrt{N}}\left\{\sqrt{\mathrm{tr}(M)} + \sqrt{\|M\|_{\mathrm{op}}}(\sqrt{q}+t)\right\}.$$

Theorem 16 replaces p by the effective dimension $\mathrm{tr}(M)/\|M\|_{\mathrm{op}}$. It can be substantially smaller when the pooled prediction geometry is concentrated in a low-dimensional predictor subspace.

Proof. For each response column e_j,

$$H^{-1/2}X^\top e_j/N \sim N\left(0, \frac{\sigma^2}{N} H^{-1/2}\widehat{\Sigma}H^{-1/2}\right) = N(0,\sigma^2 M/N).$$

The columns are independent, which proves the distributional identity. Apply Lemma 31 with

$$A = M^{1/2}, \qquad B = I_q.$$

Then

$$\|A\|_F = \sqrt{\mathrm{tr}(M)}, \quad \|A\|_{\mathrm{op}} = \sqrt{\|M\|_{\mathrm{op}}}, \quad \|B\|_F = \sqrt{q}, \quad \|B\|_{\mathrm{op}} = 1.$$

Multiplication by $\sigma/\sqrt{N}$ proves the result. This completes the proof. □

The response coordinates may be correlated. The next theorem permits a general Gaussian response covariance and identifies its effective rank.

Theorem 17. *Suppose the error rows are independent $N(0,\Omega)$ conditional on X, where $\Omega \succeq 0$. Then*

$$\widetilde{W} \overset{d}{=} \frac{1}{\sqrt{N}} M^{1/2} G \Omega^{1/2},$$

and, for every $t>0$, with probability at least $1-2e^{-t^2/2}$,

$$\varepsilon \le \frac{1}{\sqrt{N}}\{\sqrt{\mathrm{tr}(M)\,\|\Omega\|_{\mathrm{op}}} + \sqrt{\|M\|_{\mathrm{op}}\,\mathrm{tr}(\Omega)} + t\sqrt{\|M\|_{\mathrm{op}}\,\|\Omega\|_{\mathrm{op}}}\}.$$

Theorem 17 allows arbitrary dependence among responses. Its two complexity terms distinguish predictor and response effective dimensions.

Proof. Let vec denote columnwise vectorization and $\otimes$ denote the Kronecker product. Then

$$\operatorname{vec}(\widetilde{W}) \sim N\left(0, \frac{1}{N}\Omega \otimes M\right).$$

This is the vectorized distribution of $N^{-1/2}M^{1/2}G\Omega^{1/2}$. Apply Lemma 31 with $A = M^{1/2}$ and $B = \Omega^{1/2}$. The identities

$$\left\|M^{1/2}\right\|_F = \sqrt{\operatorname{tr}(M)}, \qquad \left\|M^{1/2}\right\|_{\mathrm{op}} = \sqrt{\|M\|_{\mathrm{op}}}$$

and their analogues for Ω give the result. This completes the proof. □

When $M \neq 0$ and $\Omega \neq 0$, define the effective ranks

$$r_M = \frac{\operatorname{tr}(M)}{\|M\|_{\mathrm{op}}}, \qquad r_\Omega = \frac{\operatorname{tr}(\Omega)}{\|\Omega\|_{\mathrm{op}}}.$$

If either matrix equals zero, $\widetilde{W} = 0$ and the score bounds are immediate. The next corollary treats the nonzero case.

Corollary 19. *Under Theorem 17, suppose $M \neq 0$ and $\Omega \neq 0$. For every $t > 0$,*

$$\varepsilon \leq \sqrt{\frac{\|M\|_{\mathrm{op}} \|\Omega\|_{\mathrm{op}}}{N}}\{\sqrt{r_M} + \sqrt{r_\Omega} + t\}$$

with probability at least $1 - 2e^{-t^2/2}$.

Corollary 19 gives the intrinsic dimension of the Gaussian score. The nominal bound follows from $r_M \leq p$ and $r_\Omega \leq q$.

Proof. Factor $\{\|M\|_{\mathrm{op}} \|\Omega\|_{\mathrm{op}} / N\}^{1/2}$ from the three terms in Theorem 17. This completes the proof. □

The Gaussian score bound gives a corresponding Stage I prediction rate. We state the result conditionally because the design may be fixed or random.

Corollary 20. *Suppose $\kappa > 0$, $\Omega \neq 0$, and the model in Theorem 17 holds. For $t > 0$, choose*

$$\lambda_1 = 2\sqrt{\frac{\|M\|_{\mathrm{op}} \|\Omega\|_{\mathrm{op}}}{N}}\{\sqrt{r_M} + \sqrt{r_\Omega} + t\}.$$

Then, with probability at least $1 - 2e^{-t^2/2}$,

$$\frac{\left\|X(\widehat{B}_1 - B^\star)\right\|_F}{\sqrt{N}} \leq \frac{4\sqrt{2r}}{\sqrt{\kappa}}\sqrt{\frac{\|M\|_{\mathrm{op}} \|\Omega\|_{\mathrm{op}}}{N}}\{\sqrt{r_M} + \sqrt{r_\Omega} + t\}.$$

Corollary 20 shows that the prediction bound depends on the effective predictor and response dimensions through the transformed score.

Proof. Corollary 19 gives $\lambda_1 \geq 2\varepsilon$ on the stated event. Apply the prediction conclusion of Theorem 6. This completes the proof. □

The expected score has the same two effective-dimension terms without a tail parameter.

Proposition 42. *Under Theorem 17, conditional on X,*

$$\mathbb{E}(\varepsilon \mid X) \leq \frac{1}{\sqrt{N}}\left\{\sqrt{\operatorname{tr}(M) \|\Omega\|_{\mathrm{op}}} + \sqrt{\|M\|_{\mathrm{op}} \operatorname{tr}(\Omega)}\right\}.$$

Proposition 42 gives the average stochastic scale of the metric dual norm.

Proof. Use the distributional identity in Theorem 17 and the expectation part of Lemma 31. This completes the proof. □

S27 APPROXIMATELY LOW-RANK COEFFICIENT CLASSES

Theorem 11 is an oracle inequality indexed by an approximation rank. This section specializes it to coefficients with decaying singular values.

Let $0 < \nu < 1$ and suppose

$$\sum_{j=1}^{p\wedge q} \sigma_j(C^\star)^\nu \le R_\nu.$$

The following lemma bounds the nuclear-norm tail after the first k singular values.

Lemma 32. *For every $k \ge 1$,*

$$\sum_{j>k} \sigma_j(C^\star) \le \frac{\nu}{1-\nu} R_\nu^{1/\nu} k^{1-1/\nu}.$$

Lemma 32 converts approximate low rank into the approximation term in Theorem 11.

Proof. Because the singular values are nonincreasing,

$$j\sigma_j(C^\star)^\nu \le \sum_{\ell=1}^{j} \sigma_\ell(C^\star)^\nu \le R_\nu.$$

Hence

$$\sigma_j(C^\star) \le R_\nu^{1/\nu} j^{-1/\nu}.$$

The integral bound gives

$$\sum_{j>k} j^{-1/\nu} \le \int_k^\infty x^{-1/\nu}\, dx = \frac{\nu}{1-\nu} k^{1-1/\nu}.$$

This completes the proof. □

The approximation rank balances estimation error for the kept singular components against the error from omitting the remaining components.

Corollary 21. *Under Theorem 11 and the premise of Lemma 32,*

$$\left\|\widehat{C} - C^\star\right\|_F \le \inf_{1\le k\le p\wedge q} \Big[\frac{2\sqrt{2k}\lambda_1}{\kappa} + \sqrt{2}\Big\{\frac{\nu\lambda_1 R_\nu^{1/\nu}}{(1-\nu)\kappa} k^{1-1/\nu}\Big\}^{1/2}\Big].$$

Corollary 21 lets the approximation rank balance estimation and singular-value tail error over a Schatten class. For an exactly rank-k coefficient, Theorem 11 has $a_k = 0$ and gives the exact-rank bound directly.

Proof. For each k, the best rank-k approximation has nuclear tail

$$a_k = \sum_{j>k} \sigma_j(C^\star).$$

Apply Lemma 32 in Theorem 11 and minimize over k. This completes the proof. □

The slow prediction result also gives a direct bound for a nuclear-norm bounded class. It does not require selecting an approximation rank.

Proposition 43. *If $\|C^\star\|_* \le R_*$ and $\lambda_1 \ge \varepsilon$, then*

$$\frac{\left\|X(\widehat{B}_1 - B^\star)\right\|_F^2}{N} \le 4\lambda_1 R_*.$$

Proposition 43 is the singular-design benchmark for a bounded nuclear-norm target. Its rate is slower than the exact-rank fast rate because it uses no curvature conversion.

Proof. Apply Theorem 10 with $C = \widehat{C}$ and $\delta = 0$. This completes the proof. □

The finite Stage I iterate has the same bound with an explicit optimization term.

Corollary 22. *Under Proposition 43, for every integer* $1 \le T \le T_1$,

$$\frac{\left\|X(H^{-1/2}C_T - B^\star)\right\|_F^2}{N} \le 4\lambda_1 R_* + \frac{4D_{\mathrm{init}}^2}{(T+1)^2}.$$

Corollary 22 gives the prediction cost of stopping Stage I after a finite number of rounds without pooled invertibility.

Proof. Theorem 10 permits

$$\delta = \frac{2D_{\mathrm{init}}^2}{(T+1)^2}.$$

Substitution proves the result. This completes the proof. □

S28 General positive-definite majorizers

The proposed metric is one computable member of a broader deterministic class. This section states the exact conditions under which the transformed spectral update and its guarantees remain valid. The result can be used if a tighter certified majorizer becomes available without releasing the pooled Gram matrix.

Let $\widetilde{H} \succ 0$ and define

$$\widetilde{C} = \widetilde{H}^{1/2}B, \qquad \widetilde{M} = \widetilde{H}^{-1/2}\widehat{\Sigma}\widetilde{H}^{-1/2}, \qquad L_{\widetilde{H}} = \left\|\widetilde{M}\right\|_{\mathrm{op}}, \qquad \kappa_{\widetilde{H}} = \lambda_{\min}(\widetilde{M}).$$

Assume $L_{\widetilde{H}} < 1$. Let P be a proper lower-semicontinuous spectral penalty whose unit-step proximal problem has a minimizer for every matrix. Its effective domain is $\operatorname{dom} P = \{C : P(C) < \infty\}$. Define $\widetilde{f}(C) = \mathcal{L}(\widetilde{H}^{-1/2}C)$ and $\widetilde{F}_2(C) = \widetilde{f}(C) + P(C)$. The following proposition gives the surrogate identity and exact update.

Proposition 44. *For every* B_t,

$$\mathcal{L}(B_t) + \langle G(B_t), B - B_t\rangle + \frac{1}{2}\left\|\widetilde{H}^{1/2}(B - B_t)\right\|_F^2 = \mathcal{L}(B) + \frac{1}{2}\left\|B - B_t\right\|_{\widetilde{H}-\widehat{\Sigma}}^2$$

The penalty applied to $\widetilde{H}^{1/2}B$ *has the exact proximal update*

$$\widetilde{C}_{t+1} \in \operatorname{prox}_P\{\widetilde{C}_t - \nabla\widetilde{f}(\widetilde{C}_t)\},$$

where prox_P *denotes the set of exact spectral proximal minimizers.*

Proposition 44 identifies the algebraic role of the matched construction. Using the same positive-definite matrix in the proximal metric and the penalty coordinate turns the update into an ordinary spectral proximal map. In the remaining results of this section, C_t denotes the iterates in the $\widetilde{H}$ coordinate.

Proof. Expansion of the quadratic loss gives

$$\mathcal{L}(B) = \mathcal{L}(B_t) + \langle G(B_t), B - B_t\rangle + \frac{1}{2}\left\|B - B_t\right\|_{\widehat{\Sigma}}^2.$$

Subtracting gives the surrogate remainder. Under $C = \widetilde{H}^{1/2}B$, the proximal quadratic is $\|C - C_t\|_F^2/2$ and the linear coefficient is $\nabla\widetilde{f}(C_t)$. Completing the square gives the update. This completes the proof. □

The descent coefficient depends only on the transformed smoothness bound.

Theorem 18. *If $C_t \in \operatorname{dom} P$, every selected exact Stage II update under $\widetilde{H}$ satisfies*

$$\widetilde{F}_2(C_t) - \widetilde{F}_2(C_{t+1}) \geq \frac{1 - L_{\widetilde{H}}}{2} \|C_{t+1} - C_t\|_F^2 .$$

Further,

$$\operatorname{dist}\{0, \partial \widetilde{F}_2(C_{t+1})\} \leq \|C_{t+1} - C_t\|_F .$$

Theorem 18 gives a positive decrease coefficient under strict majorization. The proposed metric gives $L_{\widetilde{H}} \leq (1+\tau)^{-1}$ using only setup scalars.

Proof. The gradient of $\widetilde{f}$ is affine with Hessian $\widetilde{M}$, hence its Lipschitz constant is $L_{\widetilde{H}}$. Let $\Delta_t = C_{t+1} - C_t$. Proximal optimality gives

$$\left\langle \nabla \widetilde{f}(C_t), \Delta_t \right\rangle + \frac{1}{2} \|\Delta_t\|_F^2 + P(C_{t+1}) - P(C_t) \leq 0.$$

Smoothness gives

$$\widetilde{f}(C_{t+1}) - \widetilde{f}(C_t) \leq \left\langle \nabla \widetilde{f}(C_t), \Delta_t \right\rangle + \frac{L_{\widetilde{H}}}{2} \|\Delta_t\|_F^2 .$$

Adding proves descent. The proximal optimality inclusion gives a subgradient at C_{t+1} whose norm satisfies

$$\left\| \nabla \widetilde{f}(C_{t+1}) - \nabla \widetilde{f}(C_t) - \Delta_t \right\|_F = \left\| (\widetilde{M} - I_p)\Delta_t \right\|_F \leq \|\Delta_t\|_F .$$

The last inequality uses $0 \preceq \widetilde{M} \preceq L_{\widetilde{H}} I_p$ and $L_{\widetilde{H}} < 1$. This completes the proof. □

The Stage I fast-rate theorem has the same form under the general metric. Let $\widehat{C}_{\widetilde{H}}$ minimize $\widetilde{f}(C) + \lambda_1 \|C\|_*$ and let $\widehat{B}_{\widetilde{H}} = \widetilde{H}^{-1/2} \widehat{C}_{\widetilde{H}}$.

Within this section, let

$$\widetilde{W} = \widetilde{H}^{-1/2} X^\top E / N, \qquad \widetilde{\varepsilon} = \left\| \widetilde{W} \right\|_{\mathrm{op}} .$$

Theorem 19. *If $\kappa_{\widetilde{H}} > 0$ and $\lambda_1 \geq \widetilde{\varepsilon}$, then*

$$\left\| \widehat{C}_{\widetilde{H}} - \widetilde{H}^{1/2} B^\star \right\|_F \leq \frac{2\sqrt{2r}\lambda_1}{\kappa_{\widetilde{H}}}, \qquad \frac{\left\| X(\widehat{B}_{\widetilde{H}} - B^\star) \right\|_F}{\sqrt{N}} \leq \frac{2\sqrt{2r}\lambda_1}{\sqrt{\kappa_{\widetilde{H}}}}.$$

Theorem 19 shows that the statistical proof depends on the metric through a dual score and a transformed curvature. The proposed H is distinguished by its communication-feasible certificate and thin master representation.

Proof. Let $C^\star_{\widetilde{H}} = \widetilde{H}^{1/2} B^\star$ and $\Delta = \widehat{C}_{\widetilde{H}} - C^\star_{\widetilde{H}}$. Let $\mathcal{T}$ be the tangent space at $C^\star_{\widetilde{H}}$, with orthogonal projections $\Delta_{\mathcal{T}}$ and $\Delta_{\mathcal{T}^\perp}$. The optimality equation and nuclear-norm decomposability give

$$\left\langle \widetilde{M}\Delta, \Delta \right\rangle \leq (\lambda_1 + \widetilde{\varepsilon}) \|\Delta_{\mathcal{T}}\|_* + (\widetilde{\varepsilon} - \lambda_1) \|\Delta_{\mathcal{T}^\perp}\|_* \leq 2\sqrt{2r}\lambda_1 \|\Delta\|_F .$$

Apply the coefficient and prediction argument of Theorem 6 with M and κ replaced by $\widetilde{M}$ and $\kappa_{\widetilde{H}}$. The prediction identity is $\left\langle \widetilde{M}\Delta, \Delta \right\rangle = \left\| X\widetilde{H}^{-1/2}\Delta \right\|_F^2 / N$. This completes the proof. □

The rank theorem also has a direct general-metric form. Let $\widetilde{Z}_t = C_t - \nabla \widetilde{f}(C_t)$.

Let

$$\widetilde{R}_t = (I_p - \widetilde{M})(C_t - \widetilde{H}^{1/2} B^\star) + \widetilde{W}, \qquad \widetilde{\eta}_t = \left\| \widetilde{R}_t \right\|_{\mathrm{op}} .$$

Proposition 45. *Suppose $1 \leq r < p \wedge q$. For the SCAD update under $\widetilde{H}$, if*

$$\widetilde{\eta}_t < \lambda_2, \qquad \sigma_r(\widetilde{H}^{1/2} B^\star) > a\lambda_2 + \widetilde{\eta}_t,$$

then $\operatorname{rank}(C_{t+1}) = r$ *and* $\left\| C_{t+1} - \widetilde{H}^{1/2} B^\star \right\|_F \leq 2\sqrt{2r}\widetilde{\eta}_t$.

Proposition 45 confirms that the rank proof is a property of the computed transformed update and not of the particular scalar upper bound used to construct H.

Proof. The update identity is

$$\widetilde{Z}_t = \widetilde{H}^{1/2} B^\star + \widetilde{R}_t.$$

Apply Weyl's inequality and the SCAD threshold regions as in Theorem 4. This completes the proof. □

S29 Stage I regularization paths

The code selects Stage I tuning values from a grid below the zero threshold in Proposition 4. This section gives deterministic monotonicity and stability properties for the convex targets.

For $\lambda \geq 0$, let

$$\widehat{C}(\lambda) \in \arg\min_C \{ f(C) + \lambda \|C\|_* \}$$

denote the unique minimizer for $\lambda > 0$ and the minimizer of smallest Frobenius norm for $\lambda = 0$. Lemma 3 proves existence, uniqueness with this convention, and support on the row space of X. The first proposition gives monotonicity of the penalty and loss along this target family.

Proposition 46. *If $0 \leq \lambda' < \lambda$, then*

$$\left\| \widehat{C}(\lambda) \right\|_* \leq \left\| \widehat{C}(\lambda') \right\|_*, \qquad f\{\widehat{C}(\lambda')\} \leq f\{\widehat{C}(\lambda)\}.$$

Proposition 46 states the usual bias-complexity ordering in the metric prediction coordinate. A larger penalty cannot increase the nuclear norm or decrease the unpenalized loss.

Proof. Optimality at the two tuning values gives

$$f\{\widehat{C}(\lambda)\} + \lambda \left\| \widehat{C}(\lambda) \right\|_* \leq f\{\widehat{C}(\lambda')\} + \lambda \left\| \widehat{C}(\lambda') \right\|_*,$$
$$f\{\widehat{C}(\lambda')\} + \lambda' \left\| \widehat{C}(\lambda') \right\|_* \leq f\{\widehat{C}(\lambda)\} + \lambda' \left\| \widehat{C}(\lambda) \right\|_*.$$

Adding and rearranging gives

$$(\lambda - \lambda')\{ \left\| \widehat{C}(\lambda) \right\|_* - \left\| \widehat{C}(\lambda') \right\|_* \} \leq 0.$$

Substitute this into the second optimality inequality for the loss bound. This completes the proof. □

The target remains stable along the tuning path when the pooled Gram matrix is singular. Only curvature on the observed predictor space enters the bound. Let $d_0 = d \wedge q$, where $d = \operatorname{rank}(X)$ and κ_+ is defined before Theorem 1.

Proposition 47. *For $\lambda, \lambda' \geq 0$,*

$$\left\| \widehat{C}(\lambda) - \widehat{C}(\lambda') \right\|_F \leq \frac{\sqrt{d_0}}{\kappa_+} |\lambda - \lambda'|.$$

Proposition 47 gives a deterministic resolution bound for the Stage I tuning grid. It does not select the grid or replace validation.

Proof. Let $C = \widehat{C}(\lambda)$ and $C' = \widehat{C}(\lambda')$. There are $V \in \partial \|C\|_*$ and $V' \in \partial \|C'\|_*$ such that

$$\nabla f(C) + \lambda V = 0, \qquad \nabla f(C') + \lambda' V' = 0.$$

Subtract the equations and take the inner product with $C - C'$:

$$\langle M(C - C'), C - C' \rangle + \lambda \langle V - V', C - C' \rangle + (\lambda - \lambda') \langle V', C - C' \rangle = 0.$$

Monotonicity makes the middle term nonnegative. By Lemma 3, $\Pi_X(C - C') = C - C'$. Also $\|\Pi_X V'\|_{\mathrm{op}} \leq 1$ and $\operatorname{rank}(\Pi_X V') \leq d_0$, hence $\|\Pi_X V'\|_F \leq \sqrt{d_0}$. Thus

$$\kappa_+ \|C - C'\|_F^2 \leq |\lambda - \lambda'| \sqrt{d_0} \|C - C'\|_F.$$

Cancel the positive distance; for $C = C'$ the bound holds directly. This completes the proof. □

The zero threshold separates tuning values that admit the zero target from those that require a nonzero target. Define

$$\lambda_{\max} = \left\| H^{-1/2} \widehat{Z} \right\|_{\mathrm{op}}.$$

Corollary 23. *The zero matrix is a minimizer if and only if $\lambda \geq \lambda_{\max}$. It is the unique minimizer if also $\lambda > 0$ or $\kappa > 0$. If $0 \leq \lambda < \lambda_{\max}$, every minimizer is nonzero.*

Corollary 23 includes the threshold itself whenever it is positive. At $\lambda = \lambda_{\max} = 0$, all minimizers satisfy $XC = 0$, and the target of smallest Frobenius norm is zero.

Proof. Let $Z_0 = H^{-1/2} \widehat{Z}$. The zero matrix is a minimizer if and only if

$$0 \in -Z_0 + \lambda \partial \left\| 0 \right\|_*,$$

which is equivalent to $\left\| Z_0 \right\|_{\mathrm{op}} \leq \lambda$. If $\lambda > \lambda_{\max}$ and $C \neq 0$, nuclear-norm duality gives

$$\begin{aligned} f(C) + \lambda \left\| C \right\|_* - f(0) &= \frac{1}{2} \langle MC, C \rangle - \langle Z_0, C \rangle + \lambda \left\| C \right\|_* \\ &\geq \frac{1}{2} \langle MC, C \rangle + (\lambda - \lambda_{\max}) \left\| C \right\|_* > 0. \end{aligned}$$

At $\lambda = \lambda_{\max} > 0$, Lemma 3 gives uniqueness. If $\lambda = \lambda_{\max} = 0$, then $Z_0 = 0$ and the objective difference is $\langle MC, C \rangle / 2$. Its minimizers have columns in $\ker(M) = \ker(X)$, giving uniqueness for $\kappa > 0$. For $\lambda < \lambda_{\max}$, zero fails the optimality condition. This completes the proof. □

The same optimality identity gives a sharper prediction bound than converting the coefficient bound through the largest eigenvalue of M.

Corollary 24. *Under Proposition 47,*

$$\frac{\left\| X H^{-1/2} \{ \widehat{C}(\lambda) - \widehat{C}(\lambda') \} \right\|_F}{\sqrt{N}} \leq \sqrt{\frac{d_0}{\kappa_+}} |\lambda - \lambda'|.$$

Corollary 24 bounds discretization of the convex tuning path on the observed design. Its constant depends on $d \wedge q$ and κ_+, not on invertibility in unobserved predictor directions.

Proof. For $D = \widehat{C}(\lambda) - \widehat{C}(\lambda')$, monotonicity in the proof of Proposition 47 gives

$$\frac{1}{N} \left\| X H^{-1/2} D \right\|_F^2 = \langle MD, D \rangle \leq |\lambda - \lambda'| \sqrt{d_0} \left\| D \right\|_F.$$

Apply Proposition 47 and take square roots. This completes the proof. □

S30 POOLED AND LOCAL RANK REGIMES

The local and pooled design ranks play different roles. For every worker s, the matrices X_s and $X_s^\top X_s$ have the same null space, which gives

$$\operatorname{rank}(\widehat{\Sigma}_s) = \operatorname{rank}(X_s) \leq n_s \wedge p.$$

In particular, $n_s < p$ implies that $\widehat{\Sigma}_s$ is singular. Local singularity in the first simulation is thus deterministic and not a low-probability event.

Pooling can restore full column rank even when every worker Gram matrix is singular.

Proposition 48. *The pooled Gram matrix satisfies*

$$\ker(\widehat{\Sigma}) = \bigcap_{s : w_s > 0} \ker(X_s).$$

Hence $\widehat{\Sigma} \succ 0$ if and only if the worker row spaces jointly span $\mathbb{R}^p$.

Proposition 48 is the exact algebraic distinction used by Theorem 6. No individual worker needs a full-rank design.

Proof. For every $v \in \mathbb{R}^p$,

$$v^\top \widehat{\Sigma} v = \sum_s w_s \frac{\|X_s v\|_2^2}{n_s}.$$

The sum is zero exactly when $X_s v = 0$ for every worker with positive weight. This proves the null-space identity and the positive-definiteness conclusion. This completes the proof. □

The stabilizing metric is full rank regardless of both local and pooled ranks. If $\bar{\ell} > 0$ and $\tau > 0$, then $\widehat{\Sigma}_1 \succeq 0$ and $\gamma = (1+\tau)\bar{\ell}$ give

$$\lambda_{\min}(H) \geq (1+\tau)\bar{\ell} > 0.$$

This relation proves existence of every transformed update in Algorithm 1, including when $N < p$.

The pooled curvature parameter equals a generalized eigenvalue. This gives an alternative expression useful for diagnostics.

Lemma 33.

$$\kappa = \inf_{v \neq 0} \frac{v^\top \widehat{\Sigma} v}{v^\top H v}.$$

Lemma 33 shows that κ measures pooled prediction curvature relative to the local metric, not relative to the Euclidean coefficient scale.

Proof. Let $u = H^{1/2} v$ in the Rayleigh quotient of $M = H^{-1/2} \widehat{\Sigma} H^{-1/2}$. This gives

$$\frac{u^\top M u}{u^\top u} = \frac{v^\top \widehat{\Sigma} v}{v^\top H v}.$$

Take the infimum. This completes the proof. □

S31 Additional communication and storage relations

The scalar communication count does not measure master memory. This section records the two resources separately for the gradient and sufficient-statistic protocols.

Proposition 49. *When $n_1 < p$, the proposed method stores the thin master basis and coefficient matrices using $O(n_1 p + pq)$ numbers beyond the master observations. Direct storage of the pooled sufficient statistics uses $O(p^2 + pq)$ numbers.*

Proposition 49 gives a separate reason to use a gradient interface when p is large. It is a storage comparison and does not alter the scalar communication inequality.

Proof. The thin right singular basis of X_1 has at most $n_1 p$ entries; coefficients and gradients have pq. Proposition 12 avoids dense metric storage. Direct storage requires p^2 Gram matrix entries and pq cross-product entries. This completes the proof. □

Symmetry reduces transmission but does not remove dense master storage after reconstruction. A symmetric $p \times p$ Gram matrix has $p(p+1)/2$ distinct entries. An ordinary full-matrix representation uses p^2 array locations, while packed symmetric storage uses $p(p+1)/2$ and has the same quadratic order. The gradient protocol can stream worker returns. The master adds $w_s G_s(B)$ to one accumulator and discards the local matrix. The pooled gradient then requires $O(pq)$ memory beyond the current iterate and metric representation. The number of workers affects latency and total traffic but not this streaming memory.

S32 Additional proof of the Stage I accelerated bound

Theorem 6 cites the accelerated proximal gradient result of Beck & Teboulle (2009). This section verifies the premises of that result for the exact transformed objective.

For the Stage I sequence, let

$$a_t = \alpha_t.$$

Define the quadratic upper model

$$Q(C; D) = f(D) + \langle \nabla f(D), C - D \rangle + \frac{1}{2} \|C - D\|_F^2 + \lambda_1 \|C\|_* .$$

The next lemma verifies majorization and exact minimization.

Lemma 34. *For all C, D,*

$$F_1(C) \le Q(C; D),$$

and the Stage I update satisfies

$$C_{t+1} \in \arg\min_C Q(C; D_t).$$

Lemma 34 proves that Algorithm 1 uses a valid unit step and solves each convex proximal subproblem exactly.

Proof. The Hessian of f is $M \preceq (1+\tau)^{-1} I_p \preceq I_p$. Hence

$$f(C) \le f(D) + \langle \nabla f(D), C - D \rangle + \frac{1}{2} \|C - D\|_F^2 .$$

Adding the nuclear penalty proves majorization. Completing the square in Q gives

$$\arg\min_C Q(C; D_t) = \mathrm{SVT}_{\lambda_1}\{D_t - \nabla f(D_t)\},$$

which is the Stage I update. This completes the proof. □

The acceleration parameters satisfy the identity required by the estimate-sequence proof.

Lemma 35. *The recursion*

$$a_{t+1} = \{1 + (1 + 4a_t^2)^{1/2}\}/2$$

satisfies

$$a_{t+1}^2 - a_{t+1} = a_t^2, \qquad a_t \ge (t+2)/2.$$

Lemma 35 gives the telescoping coefficient and its quadratic growth.

Proof. The first identity follows by squaring $2a_{t+1} - 1 = (1 + 4a_t^2)^{1/2}$. For the second, $a_0 = 1 = 2/2$. If $a_t \ge (t+2)/2$, then

$$a_{t+1} \ge \frac{1 + 2a_t}{2} \ge \frac{t+3}{2}.$$

Induction proves the result. This completes the proof. □

Lemma 35 verifies the acceleration recursion used in Algorithm 1. Lemma 34 verifies the unit-step quadratic majorization. These are the two algorithm-specific premises of Beck & Teboulle (2009, Theorem 4.4), which gives the objective bound in Theorem 6. Strong convexity with parameter κ then gives the stated distance bound.

S33 STABILITY UNDER METRIC APPROXIMATION

The setup metric may be formed from certified approximations to the worker spectral radii. This section quantifies the resulting changes in the penalty, transformed score, curvature, and statistical bound. The results concern the same estimator and do not introduce a second optimization procedure.

Let $H_0, H_1 \succ 0$. For $h = 0, 1$, define

$$M_h = H_h^{-1/2}\widehat{\Sigma}H_h^{-1/2}, \qquad \kappa_h = \lambda_{\min}(M_h), \qquad L_h = \lambda_{\max}(M_h), \qquad \varepsilon_h = \left\|H_h^{-1/2}X^\top E/N\right\|_{\mathrm{op}}.$$

The following proposition gives all deterministic comparisons needed below.

Proposition 50. *Suppose* $0 \le \delta < 1$ *and*

$$(1-\delta)H_0 \preceq H_1 \preceq (1+\delta)H_0.$$

Then, for every $B \in \mathbb{R}^{p\times q}$ *and* $1 \le j \le p \wedge q$,

$$\sqrt{1-\delta}\,\sigma_j(H_0^{1/2}B) \le \sigma_j(H_1^{1/2}B) \le \sqrt{1+\delta}\,\sigma_j(H_0^{1/2}B),$$
$$\sqrt{1-\delta}\left\|H_0^{1/2}B\right\|_* \le \left\|H_1^{1/2}B\right\|_* \le \sqrt{1+\delta}\left\|H_0^{1/2}B\right\|_*,$$

and the same norm inequalities hold for the Frobenius norm. Further,

$$\frac{\varepsilon_0}{\sqrt{1+\delta}} \le \varepsilon_1 \le \frac{\varepsilon_0}{\sqrt{1-\delta}}, \qquad \frac{\kappa_0}{1+\delta} \le \kappa_1 \le \frac{\kappa_0}{1-\delta}, \qquad \frac{L_0}{1+\delta} \le L_1 \le \frac{L_0}{1-\delta}.$$

Proposition 50 shows that a relative Loewner error changes every statistical and optimization factor by an explicit multiplicative amount. It also controls each transformed signal singular value, which is the scale used by the rank theorem.

Proof. Congruence by B gives

$$(1-\delta)B^\top H_0 B \preceq B^\top H_1 B \preceq (1+\delta)B^\top H_0 B.$$

The eigenvalues of $B^\top H_h B$ are the squared singular values of $H_h^{1/2}B$. Monotonicity and summation give the singular-value and norm bounds. Inverse Loewner order gives

$$\frac{1}{1+\delta}H_0^{-1} \preceq H_1^{-1} \preceq \frac{1}{1-\delta}H_0^{-1}.$$

For $S = X^\top E/N$,

$$\left\|H_h^{-1/2}S\right\|_{\mathrm{op}}^2 = \lambda_{\max}(S^\top H_h^{-1}S), \qquad h = 0, 1.$$

Congruence by S gives the score bound. For curvature,

$$\kappa_h = \inf_{v\neq 0}\frac{v^\top\widehat{\Sigma}v}{v^\top H_h v}, \qquad L_h = \sup_{v\neq 0}\frac{v^\top\widehat{\Sigma}v}{v^\top H_h v}, \qquad h = 0, 1.$$

Substitute the denominator bounds. This completes the proof. □

S34 STABILITY OF THE CONVEX TARGET TO RESPONSE PERTURBATIONS

The Stage I target is a deterministic function of the pooled predictor-response cross-product. This section bounds its change when responses or transmitted cross-products are perturbed. The design and setup metric remain fixed throughout.

For a response matrix $Y \in \mathbb{R}^{N\times q}$ and $\lambda \ge 0$, define

$$A_Y = H^{-1/2}X^\top Y/N$$

and let

$$C_\lambda(Y) \in \arg\min_C \left\{\frac{1}{2}\langle MC, C\rangle - \langle A_Y, C\rangle + \lambda\|C\|_*\right\}.$$

When $\kappa > 0$, this minimizer is unique. The following theorem gives its deterministic modulus of continuity.

For $Y, Y' \in \mathbb{R}^{N\times q}$, let $D = C_\lambda(Y) - C_\lambda(Y')$ and $\Delta_A = A_Y - A_{Y'}$.

Theorem 20. *If $\kappa > 0$, then*

$$\left\|M^{1/2}D\right\|_F \le \left\|M^{-1/2}\Delta_A\right\|_F, \qquad \|D\|_F \le \frac{\|\Delta_A\|_F}{\kappa}.$$

Theorem 20 shows that the convex target is nonexpansive in pooled prediction scale after the response perturbation is measured in the corresponding dual scale. The second inequality converts this relation to the transformed coefficient coordinate.

Proof. There are nuclear-norm subgradients V and V' such that

$$MC_\lambda(Y) - A_Y + \lambda V = 0, \qquad MC_\lambda(Y') - A_{Y'} + \lambda V' = 0.$$

Subtract and take the inner product with D:

$$\left\|M^{1/2}D\right\|_F^2 = \langle \Delta_A, D\rangle - \lambda \langle V - V', D\rangle .$$

Monotonicity of the nuclear-norm subdifferential gives $\langle V - V', D\rangle \ge 0$. Hence

$$\left\|M^{1/2}D\right\|_F^2 \le \left\langle M^{-1/2}\Delta_A, M^{1/2}D\right\rangle \le \left\|M^{-1/2}\Delta_A\right\|_F \left\|M^{1/2}D\right\|_F .$$

Cancellation gives the first bound. Cauchy–Schwarz also gives

$$\kappa \|D\|_F^2 \le \|\Delta_A\|_F \|D\|_F ,$$

This completes the proof. □

The fitted responses remain stable even when the pooled design is singular. Let P_X be the orthogonal projection onto the column space of X. For full column rank, $P_X = X(X^\top X)^{-1}X^\top$. A Stage I minimizer exists for every $\lambda \ge 0$: the objective is coercive when $\lambda > 0$, and it is a least-squares objective when $\lambda = 0$.

Corollary 25. *For any two Stage I minimizers at the same $\lambda \ge 0$,*

$$\frac{\left\|XH^{-1/2}\{C_\lambda(Y) - C_\lambda(Y')\}\right\|_F}{\sqrt{N}} \le \frac{\|P_X(Y - Y')\|_F}{\sqrt{N}} \le \frac{\|Y - Y'\|_F}{\sqrt{N}}.$$

Corollary 25 proves a unit Lipschitz bound for the fitted training responses, including singular pooled designs. Taking $Y' = Y$ shows that these responses are unique even when the coefficient minimizer is not. The bound holds for every penalty level and worker partition.

Proof. Let $D = C_\lambda(Y) - C_\lambda(Y')$ and $\mathsf{X} = XH^{-1/2}$. Subtracting the optimality equations, monotonicity and $P_X\mathsf{X}D = \mathsf{X}D$ give

$$\begin{aligned}\|\mathsf{X}D\|_F^2 &= N\langle MD, D\rangle \le N\langle A_Y - A_{Y'}, D\rangle \\ &= \langle P_X(Y - Y'), \mathsf{X}D\rangle \le \|P_X(Y - Y')\|_F \|\mathsf{X}D\|_F .\end{aligned}$$

Cancel $\|\mathsf{X}D\|_F$ when positive and use $\|P_X\|_{\mathrm{op}} \le 1$. The zero case holds directly.

For $\kappa > 0$, let $\Delta_A = A_Y - A_{Y'}$. Then

$$M = \mathsf{X}^\top\mathsf{X}/N, \qquad \Delta_A = \mathsf{X}^\top(Y - Y')/N.$$

Because X and X have the same column space,

$$\left\|M^{-1/2}\Delta_A\right\|_F^2 = \frac{1}{N}\mathrm{tr}\left[(Y - Y')^\top\mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top(Y - Y')\right] = \frac{1}{N}\|P_X(Y - Y')\|_F^2 .$$

Thus the bound equals that of Theorem 20, with left side $\left\|M^{1/2}D\right\|_F$. This completes the proof. □

The coefficient change depends on pooled curvature and on the smallest eigenvalue of the setup metric.

Corollary 26. *Under Theorem 20,*

$$\left\|H^{-1/2}\{C_\lambda(Y) - C_\lambda(Y')\}\right\|_F \le \frac{\sqrt{\|M\|_{\mathrm{op}}}}{\kappa\sqrt{\gamma N}} \|Y - Y'\|_F .$$

Corollary 26 makes the identifiability cost explicit. Prediction remains contractive, while coefficient stability deteriorates when κ approaches zero.

Proof. The operator norm of $H^{-1/2}X^\top/N$ is $\sqrt{\|M\|_{\mathrm{op}}/N}$. Hence

$$\|\Delta_A\|_F \le \sqrt{\frac{\|M\|_{\mathrm{op}}}{N}} \|Y - Y'\|_F .$$

Apply the second inequality in Theorem 20 and $\left\|H^{-1/2}\right\|_{\mathrm{op}} \le \gamma^{-1/2}$. This completes the proof. □

The same calculation controls predictions on a second fixed design. Let $X_0 \in \mathbb{R}^{n_0 \times p}$.

Proposition 51. *Under Theorem 20,*

$$\frac{\left\|X_0 H^{-1/2}\{C_\lambda(Y) - C_\lambda(Y')\}\right\|_F}{\sqrt{n_0}} \le \frac{\|X_0\|_{\mathrm{op}} \sqrt{\|M\|_{\mathrm{op}}}}{\kappa\sqrt{n_0\gamma N}} \|Y - Y'\|_F .$$

Proposition 51 separates training curvature from the scale of a new design. It gives a deterministic bound for changes in fitted values on the response-prediction scale used in the numerical studies.

Proof. For any matrix D,

$$\left\|X_0 H^{-1/2} D\right\|_F \le \|X_0\|_{\mathrm{op}} \left\|H^{-1/2}\right\|_{\mathrm{op}} \|D\|_F .$$

Use the bound for $\|D\|_F$ in the proof of Corollary 26 and divide by $\sqrt{n_0}$. This completes the proof. □

Distributed response perturbations enter through their pooled Euclidean energy. Let $D_s \in \mathbb{R}^{n_s \times q}$ perturb the response matrix at worker s, and let D_{all} be their row concatenation.

Proposition 52. *Under Corollary 25, replacing every Y_s by $Y_s + D_s$ changes the fitted training responses by at most*

$$\frac{\|P_X D_{\mathrm{all}}\|_F}{\sqrt{N}} \le \left\{\frac{1}{N}\sum_{s=1}^{m} \|D_s\|_F^2\right\}^{1/2}.$$

If only worker s is perturbed, the upper bound is $\|D_s\|_F/\sqrt{N}$.

Proposition 52 shows that unequal worker sizes need no additional stability factor. A worker affects the target through the observations it contributes to the pooled loss.

Proof. Apply Corollary 25 with $Y - Y' = D_{\mathrm{all}}$. Row concatenation gives $\|D_{\mathrm{all}}\|_F^2 = \sum_s \|D_s\|_F^2$. This completes the proof. □

Errors in transmitted cross-products admit a related bound. Suppose the pooled cross-product $\widehat{Z}$ is replaced by $\widehat{Z} + \Delta_Z$, while H and $\widehat{\Sigma}$ remain fixed.

Let $\widehat{C}$ and $\widehat{C}^\Delta$ be the two Stage I targets.

Proposition 53. *If $\kappa > 0$, then*

$$\frac{\left\|XH^{-1/2}(\widehat{C}^\Delta - \widehat{C})\right\|_F}{\sqrt{N}} \le \left\|M^{-1/2}H^{-1/2}\Delta_Z\right\|_F, \quad \left\|H^{-1/2}(\widehat{C}^\Delta - \widehat{C})\right\|_F \le \frac{\|\Delta_Z\|_F}{\kappa\gamma}.$$

Proposition 53 gives a deterministic tolerance for finite-precision aggregation of predictor-response summaries. It concerns the convex target and does not assign a privacy interpretation to the perturbation.

Proof. Because $\kappa > 0$, $D_Y = NX(X^\top X)^{-1}\Delta_Z$ is well defined and $X^\top D_Y/N = \Delta_Z$. Apply Theorem 20 with $\Delta_A = H^{-1/2}\Delta_Z$. For the second bound,

$$\left\|\widehat{C}^{\Delta} - \widehat{C}\right\|_F \le \frac{\left\|H^{-1/2}\Delta_Z\right\|_F}{\kappa} \le \frac{\|\Delta_Z\|_F}{\kappa\sqrt{\gamma}}.$$

Multiplication by $H^{-1/2}$ gives the stated result. This completes the proof. □

S35 Statistical consequences of metric approximation

A metric approximation changes both the transformed score and the curvature. The following corollary bounds the resulting coefficient and prediction errors. Let $\widehat{C}_1$ minimize the Stage I objective in the coordinate $C = H_1^{1/2}B$ and let $C_1^\star = H_1^{1/2}B^\star$.

Corollary 27. *Under Proposition 50, suppose*

$$L_0 < 1 - \delta, \qquad \kappa_0 > 0, \qquad \lambda_1 \ge \frac{\varepsilon_0}{\sqrt{1-\delta}}.$$

Then the unit-step Stage I iterations under H_1 are valid and

$$\left\|\widehat{C}_1 - C_1^\star\right\|_F \le \frac{2\sqrt{2r}(1+\delta)\lambda_1}{\kappa_0}, \quad \frac{\left\|X(H_1^{-1/2}\widehat{C}_1 - B^\star)\right\|_F}{\sqrt{N}} \le 2\sqrt{\frac{2r(1+\delta)}{\kappa_0}}\lambda_1.$$

Corollary 27 gives a direct tolerance for approximate metric construction. The premise preserves strict majorization, while the conclusion separates score and curvature changes.

Proof. Proposition 50 gives

$$L_1 \le \frac{L_0}{1-\delta} < 1, \qquad \varepsilon_1 \le \frac{\varepsilon_0}{\sqrt{1-\delta}}, \qquad \kappa_1 \ge \frac{\kappa_0}{1+\delta}.$$

The penalty premise gives $\lambda_1 \ge \varepsilon_1$. Apply Theorem 19 with H_1. Substitution of the lower bound for κ_1 proves both assertions. This completes the proof. □

We now specialize the comparison to the metric used by Algorithm 1. Suppose each worker returns a certified bound L_s satisfying

$$\ell_s \le L_s \le (1+\rho)\ell_s, \qquad \rho \ge 0.$$

Let $\bar{L} = \sum_s w_s L_s$ and

$$H_L = \widehat{\Sigma}_1 + (1+\tau)\bar{L}I_p, \qquad M_L = H_L^{-1/2}\widehat{\Sigma}H_L^{-1/2}.$$

Proposition 54. *The exact and certified metrics satisfy*

$$H \preceq H_L \preceq (1+\rho)H.$$

If $\varepsilon_L = \left\|H_L^{-1/2}X^\top E/N\right\|_{\mathrm{op}}$ and $\kappa_L = \lambda_{\min}(M_L)$, then

$$\varepsilon_L \le \varepsilon, \qquad \frac{\kappa}{1+\rho} \le \kappa_L \le \kappa, \qquad \|M_L\|_{\mathrm{op}} \le \|M\|_{\mathrm{op}}.$$

Further, for every B and every $j \le p \wedge q$,

$$\sigma_j(H^{1/2}B) \le \sigma_j(H_L^{1/2}B) \le \sqrt{1+\rho}\,\sigma_j(H^{1/2}B).$$

Proposition 54 proves that certified overestimation cannot invalidate descent or enlarge the transformed score. Its statistical cost is at most the stated curvature factor.

Proof. The worker inequalities give

$$0 \le \bar{L} - \bar{\ell} \le \rho\bar{\ell}.$$

Hence $H_L \succeq H$ and

$$H_L - H = (1+\tau)(\bar{L} - \bar{\ell})I_p \preceq \rho(1+\tau)\bar{\ell}I_p \preceq \rho H.$$

This proves the metric comparison. Inverse Loewner order gives $\varepsilon_L \le \varepsilon$. The generalized Rayleigh quotients give

$$\frac{1}{1+\rho}\frac{v^\top \widehat{\Sigma} v}{v^\top H v} \le \frac{v^\top \widehat{\Sigma} v}{v^\top H_L v} \le \frac{v^\top \widehat{\Sigma} v}{v^\top H v}.$$

Taking infima and suprema proves the curvature claims. Congruence by B and eigenvalue monotonicity prove the signal inequalities. This completes the proof. □

The corresponding Stage I guarantee does not require a larger penalty level because the certified metric reduces the dual score. Let $\widehat{C}_L$ minimize $\mathcal{L}(H_L^{-1/2}C) + \lambda_1 \|C\|_*$.

Corollary 28. *If $\kappa > 0$ and $\lambda_1 \ge \varepsilon$, then the Stage I target under H_L satisfies*

$$\left\|\widehat{C}_L - H_L^{1/2}B^\star\right\|_F \le \frac{2\sqrt{2r}(1+\rho)\lambda_1}{\kappa}, \quad \frac{\left\|X(H_L^{-1/2}\widehat{C}_L - B^\star)\right\|_F}{\sqrt{N}} \le 2\sqrt{\frac{2r(1+\rho)}{\kappa}}\lambda_1.$$

Corollary 28 shows that a relative setup tolerance has a transparent effect on the statistical constant and preserves the same tuning order.

Proof. Proposition 54 gives $\varepsilon_L \le \varepsilon$ and $\kappa_L \ge \kappa/(1+\rho)$. Proposition 29 verifies strict majorization. Apply Theorem 19. This completes the proof. □

A certified lower approximation can be converted to the preceding upper-bound setting by deterministic scaling.

Corollary 29. *Suppose $0 \le \rho < 1$ and*

$$(1-\rho)\ell_s \le \underline{\ell}_s \le \ell_s.$$

Then $L_s = \underline{\ell}_s/(1-\rho)$ satisfies

$$\ell_s \le L_s \le \{1 + \rho/(1-\rho)\}\ell_s.$$

Hence Proposition 54 and Corollary 28 apply with ρ replaced by $\rho/(1-\rho)$.

Corollary 29 gives a safe conversion when a local eigensolver gives a relative lower approximation and a certified tolerance.

Proof. Divide the assumed inequalities by $1-\rho$. The resulting upper factor is $(1-\rho)^{-1} = 1 + \rho/(1-\rho)$. Apply the cited results. This completes the proof. □

The maximal Stage I penalty also changes predictably under certified overestimation.

Let

$$\lambda_{\max,L} = \left\|H_L^{-1/2}\widehat{Z}\right\|_{\mathrm{op}}, \qquad \lambda_{\max} = \left\|H^{-1/2}\widehat{Z}\right\|_{\mathrm{op}}.$$

Proposition 55. *Then*

$$\frac{\lambda_{\max}}{\sqrt{1+\rho}} \le \lambda_{\max,L} \le \lambda_{\max}.$$

Proposition 55 quantifies the effect of approximate setup on the tuning grid. Its bounds give the corresponding rescaling of a grid defined relative to the upper value.

Proof. Apply inverse Loewner order to $H \preceq H_L \preceq (1+\rho)H$ and use

$$\left\|K^{-1/2}\widehat{Z}\right\|_{\mathrm{op}}^2 = \lambda_{\max}(\widehat{Z}^\top K^{-1}\widehat{Z}), \qquad K \in \{H, H_L\}.$$

This completes the proof. □

S36 INEXACT SPECTRAL PROXIMAL UPDATES

Algorithm 1 uses an exact spectral map in the mathematical analysis. A numerical singular value decomposition may stop with a nonzero optimization tolerance. The next results give verifiable objective and stationarity bounds for such an update.

For a Stage II iterate C_t, define the proximal model

$$\mathcal{Q}_t(C) = f(C_t) + \langle \nabla f(C_t), C - C_t \rangle + \frac{1}{2}\|C - C_t\|_F^2 + P_2(C).$$

Let $c = \tau/(1+\tau)$. Let $u_t, v_t \geq 0$ measure objective and first-order errors in this model.

Theorem 21. *Suppose C_{t+1} satisfies*

$$\mathcal{Q}_t(C_{t+1}) \leq \mathcal{Q}_t(C_t) + u_t, \qquad \operatorname{dist}\{0, \partial \mathcal{Q}_t(C_{t+1})\} \leq v_t.$$

Then

$$F_2(C_t) - F_2(C_{t+1}) \geq \frac{c}{2}\|C_{t+1} - C_t\|_F^2 - u_t, \quad \operatorname{dist}\{0, \partial F_2(C_{t+1})\} \leq v_t + \|C_{t+1} - C_t\|_F.$$

Theorem 21 separates the error of the spectral subproblem from the progress of the outer method. Exact thresholding has $u_t = v_t = 0$ and recovers the descent and stationarity structure used in the main results.

Proof. Let $D_t = C_{t+1} - C_t$. Because f is quadratic with Hessian M,

$$f(C_{t+1}) = f(C_t) + \langle \nabla f(C_t), D_t \rangle + \frac{1}{2}\langle M D_t, D_t \rangle.$$

Hence

$$F_2(C_{t+1}) = \mathcal{Q}_t(C_{t+1}) - \frac{1}{2}\langle (I_p - M) D_t, D_t \rangle.$$

Lemma 1 gives

$$I_p - M \succeq \frac{\tau}{1+\tau} I_p = c I_p.$$

Because $\mathcal{Q}_t(C_t) = F_2(C_t)$, the objective premise yields

$$F_2(C_{t+1}) \leq F_2(C_t) + u_t - \frac{c}{2}\|D_t\|_F^2.$$

Choose $R_{t+1} \in \partial \mathcal{Q}_t(C_{t+1})$ with $\|R_{t+1}\|_F \leq v_t$ and $V_{t+1} \in \partial P_2(C_{t+1})$ satisfying

$$R_{t+1} = \nabla f(C_t) + D_t + V_{t+1}.$$

The corresponding subgradient of F_2 equals

$$\nabla f(C_{t+1}) + V_{t+1} = R_{t+1} - (I_p - M) D_t.$$

The triangle inequality and $\|I_p - M\|_{\mathrm{op}} \leq 1$ give the stationarity bound. This completes the proof. □

The finite-iteration consequence depends only on accumulated objective and first-order tolerances. Define

$$U_K = \sum_{t=T_1}^{T_1+K-1} u_t, \qquad V_K = \sum_{t=T_1}^{T_1+K-1} v_t^2.$$

Corollary 30. *For $K \geq 1$, suppose the two premises of Theorem 21 hold for $t = T_1, \ldots, T_1 + K - 1$. Then*

$$\sum_{t=T_1}^{T_1+K-1} \|C_{t+1} - C_t\|_F^2 \leq \frac{2}{c}\{F_2(C_{T_1}) - \inf_C F_2(C) + U_K\},$$

and

$$\min_{T_1 \leq t < T_1+K} \operatorname{dist}^2\{0, \partial F_2(C_{t+1})\} \leq \frac{4}{cK}\{F_2(C_{T_1}) - \inf_C F_2(C) + U_K\} + \frac{2V_K}{K}.$$

Corollary 30 gives a finite-iteration stationarity bound for approximate spectral computations. The stationarity bound has the exact $K^{-1/2}$ order when accumulated tolerances are bounded. If $U_K/K + V_K/K$ tends to zero, the best-iterate stationarity residual converges to zero.

Proof. Sum Theorem 21 and use the lower bound for F_2. Its stationarity bound and $(x+y)^2 \le 2x^2 + 2y^2$ give

$$\sum_t \operatorname{dist}^2\{0, \partial F_2(C_{t+1})\} \le 2\sum_t \|C_{t+1} - C_t\|_F^2 + 2V_K.$$

Substitute the summed descent bound and bound the minimum by the average over K steps. This completes the proof. □

Summable tolerances preserve stationary cluster points. Full-sequence convergence under inexact updates would need a stronger error condition and is not asserted.

Corollary 31. *Suppose $\kappa > 0$ and the two premises of Theorem 21 hold for every $t \ge T_1$, with*

$$\sum_{t=T_1}^{\infty} u_t < \infty, \qquad \sum_{t=T_1}^{\infty} v_t^2 < \infty.$$

Then the inexact sequence is bounded, $\|C_{t+1} - C_t\|_F \to 0$, and every cluster point is stationary for F_2.

Corollary 31 states the asymptotic conclusion supported by the two numerical residuals. It does not replace the stronger exact-update convergence statement in Corollary 2.

Proof. Corollary 30 gives square summability of the iterate differences. Descent gives

$$F_2(C_t) \le F_2(C_{T_1}) + \sum_{j=T_1}^{t-1} u_j.$$

By $\kappa > 0$ and Lemma 29, the sequence is bounded. Square summability gives $C_{t+1} - C_t \to 0$ and $v_t \to 0$. Let

$$A_t = F_2(C_t) + \sum_{j=t}^{\infty} u_j.$$

Descent makes A_t nonincreasing and bounded below. Its tail tends to zero, hence $F_2(C_t)$ converges. If $C_{t_j} \to \bar{C}$, then $C_{t_j+1} \to \bar{C}$, and the first-order premise gives $V_{t_j+1} \in \partial P_2(C_{t_j+1})$ with

$$\nabla f(C_{t_j+1}) + V_{t_j+1} \to 0.$$

Continuity of F_2 and closedness of its limiting subdifferential give $0 \in \partial F_2(\bar{C})$. This completes the proof. □

The objective residual also gives a direct acceptance rule. Under Theorem 21,

$$u_t < \frac{c}{2}\|C_{t+1} - C_t\|_F^2$$

implies $F_2(C_{t+1}) < F_2(C_t)$. This rule uses quantities from the current round and makes no claim about an uncomputed rank-constrained solution.

An approximate returned matrix can still give a certified numerical rank. Let Z have singular values $d_1 \ge \cdots \ge d_{p\wedge q}$, and let $\mathcal{T}_{\lambda,a}(Z)$ denote the exact SCAD spectral map.

Proposition 56. *Let $1 \le r < p \wedge q$. Suppose*

$$d_r > a\lambda, \qquad d_{r+1} \le \lambda, \qquad \left\|\widetilde{C} - \mathcal{T}_{\lambda,a}(Z)\right\|_{\mathrm{op}} \le \zeta.$$

If $d_r > 2\zeta$, then $\widetilde{C}$ has exactly r singular values larger than ζ.

Proposition 56 distinguishes exact algebraic rank from a rank reported at numerical tolerance. A strict threshold gap makes the reported rank stable to a certified operator error.

Proof. The exact SCAD map has rank r, keeps its first r singular values, and sets all later singular values equal to zero. Weyl's inequality gives

$$\sigma_r(\widetilde{C}) \geq d_r - \zeta > \zeta, \qquad \sigma_{r+1}(\widetilde{C}) \leq \zeta.$$

Monotonicity of the singular values proves the result. This completes the proof. □

S37 Information in exact gradient queries

Proposition 7 gives a sufficient set of exact queries for recovering a worker Gram matrix. The next theorem first treats gradient returns alone, with no spectral-radius information. Proposition 57 then includes the setup scalar transmitted by Algorithm 1. Both lower bounds concern the full positive semidefinite class, not a subclass with a prescribed deficient rank.

Fix one worker and let

$$G(B) = \Sigma B - Z,$$

where $\Sigma \succeq 0$ is symmetric and $Z \in \mathbb{R}^{p\times q}$. Suppose $G(0)$ has already revealed Z. A query may depend on every earlier query and returned gradient. Let $\mathrm{col}(B)$ denote the column space of a matrix B.

Theorem 22. *Let $B_1, \ldots, B_K \in \mathbb{R}^{p\times q}$ be adaptive exact-gradient queries. If*

$$\dim \mathrm{span}\{\mathrm{col}(B_1), \ldots, \mathrm{col}(B_K)\} < p,$$

then the returned gradients do not identify Σ over the positive semidefinite cone. Hence identification of every $\Sigma \succeq 0$ requires $Kq \geq p$.

Theorem 22 shows that adaptivity cannot identify an arbitrary Gram matrix before the queried columns span all p predictor directions.

Proof. Let $\mathcal{S}$ be the span in the theorem. If $\dim(\mathcal{S}) < p$, choose a unit vector $v \in \mathcal{S}^{\perp}$ and any $a > 0$. Define

$$\Sigma' = \Sigma + avv^{\top}.$$

Then $\Sigma' \succeq 0$, $\Sigma' \neq \Sigma$, and

$$\Sigma' B_k - \Sigma B_k = avv^{\top} B_k = 0, \qquad k = 1, \ldots, K.$$

Thus the returned gradients agree. For adaptive queries, induction gives identical queries and gradients through round K, leaving Σ unidentified.

Because $\dim(\mathcal{S}) \leq Kq$, identification requires $Kq \geq p$. This completes the proof. □

The setup scalar restricts the admissible Gram matrices. The following construction keeps that scalar fixed and proves the same lower bound on the number of queries when $p \geq 2$.

Proposition 57. *Let $p \geq 2$ and let $\ell > 0$ be known. After $G(0)$, identification of every $\Sigma \succeq 0$ satisfying $\lambda_{\max}(\Sigma) = \ell$ requires at least $\lceil p/q \rceil$ adaptive exact matrix queries.*

Proposition 57 matches the upper bound for recovering the Gram matrix even when the spectral radius is known. It is a worst-case statement. For $p = 1$, the known radius already identifies the Gram matrix, and hence the restriction $p \geq 2$ is necessary.

Proof. Let Z be fixed and query $\Sigma = \ell I_p$. If $Kq < p$, choose a unit v orthogonal to all queried columns and $0 < a < \ell$. Let

$$\Sigma' = \ell I_p - avv^{\top}.$$

Then $\Sigma' \succ 0$, $\Sigma' \neq \Sigma$, and $p \geq 2$ gives

$$\lambda_{\max}(\Sigma') = \ell = \lambda_{\max}(\Sigma), \qquad (\Sigma' - \Sigma)B_k = 0 \quad \text{for } k = 1, \ldots, K.$$

The spectral radius and the gradient at zero agree. Induction gives identical adaptive transcripts, proving the lower bound $\lceil p/q \rceil$, attained by coordinate-block queries. Both Gram matrices are realizable with at least p worker rows; the argument does not extend to a prescribed singular-design class. This completes the proof. □

The coordinate-block construction in Proposition 7 attains the lower bound.

Corollary 32. *Without a known spectral radius, the smallest number of exact $p \times q$ gradient queries after the zero query that identifies every Gram matrix over the full positive semidefinite cone is $K_\star = \lceil p/q \rceil$.*

Corollary 32 gives the exact query count for Gram matrix recovery. Proposition 57 gives the same worst-case count for a known positive radius when $p \geq 2$. If $p = 1$ or the known radius equals zero, that radius already determines the Gram matrix. It concerns recovery through full gradient matrices and does not state a lower bound for every possible compressed communication protocol.

Proof. Theorem 22 gives $K \geq \lceil p/q \rceil$. Proposition 7 gives equality by querying coordinate blocks. This completes the proof. □

Counting only returned matrices already makes adaptive recovery more expensive than direct symmetric Gram matrix transmission.

Proposition 58. *The zero return and $K_\star$ nonzero gradient returns contain*

$$S_{\text{return}} = (K_\star + 1)pq \geq p^2 + pq$$

scalars. Direct transmission of Σ and Z contains

$$S_{\text{direct}} = p(p+1)/2 + pq$$

distinct scalars. For $p > 1$,

$$S_{\text{return}} - S_{\text{direct}} \geq p(p-1)/2 > 0.$$

Proposition 58 confirms that repeated gradient queries are not an efficient way to reconstruct the sufficient statistics. The proposed method reduces scalar communication when it satisfies the finite break-even inequality in Proposition 1.

Proof. Each returned gradient has pq entries. Because $K_\star q \geq p$,

$$(K_\star + 1)pq \geq p^2 + pq.$$

Subtract the direct count and simplify. This completes the proof. □

The same construction gives an explicit unresolved direction of the Gram matrix before the lower bound is reached. If $Kq < p$, the matrices Σ and $\Sigma' = \Sigma + avv^\top$ from the proof of Theorem 22 agree on all K adaptive query returns. Their difference has rank one and operator norm a, where a can be any prescribed positive value. This states the information left unresolved at a finite horizon. It is not a privacy guarantee because later queries can recover the missing direction.

The complete transcript determines the Gram matrix only on the span of the queried columns. The next proposition characterizes the remaining symmetric directions.

Let

$$Q = [B_1 \ \cdots \ B_K], \qquad \mathcal{S} = \text{col}(Q), \qquad d = \dim(\mathcal{S}).$$

Proposition 59. *Two symmetric matrices Σ and Σ' give the same exact-gradient returns at all queries if and only if*

$$(\Sigma - \Sigma')Q = 0.$$

If $P_{\mathcal{S}}$ is the orthogonal projection onto $\mathcal{S}$, every compatible symmetric difference $A = \Sigma' - \Sigma$ satisfies

$$A = (I_p - P_{\mathcal{S}})A(I_p - P_{\mathcal{S}}).$$

The vector space of such differences has dimension $\frac{(p-d)(p-d+1)}{2}$.

Proposition 59 quantifies the information missing from a finite transcript. Each newly queried predictor direction reduces the unresolved symmetric subspace, and complete recovery occurs exactly at $d = p$.

Proof. For fixed Z, equality of returned gradients is equivalent to

$$\Sigma B_k = \Sigma' B_k, \qquad k = 1, \dots, K.$$

Concatenation proves the first equivalence. Because $AQ = 0$, $AP_{\mathcal{S}} = 0$; symmetry gives $P_{\mathcal{S}} A = 0$. Expanding

$$A = \{P_{\mathcal{S}} + I_p - P_{\mathcal{S}}\} A \{P_{\mathcal{S}} + I_p - P_{\mathcal{S}}\}$$

leaves an arbitrary symmetric matrix on $\mathcal{S}^{\perp}$, with $(p-d)(p-d+1)/2$ free entries. This completes the proof. □

S38 Prediction and rank on a singular population support

Theorem 1 bounds prediction on the observed predictor space. This section controls its score and curvature and gives sufficient conditions for rank recovery along the same computed sequence. The population support may be a strict subspace of the predictor space; no recovery is asserted outside that support. The subspace identity and the main prediction proof are in Appendix S1.3.

For the next result, the design is fixed, and $d = \operatorname{rank}(X)$ is its rank. Conditional on it, assume that the error rows are independent, centered, and satisfy the noise bound in Condition 1. No distributional condition on the predictors is used.

Proposition 60. *For every $t > 0$, conditionally on X, with probability at least $1 - 2e^{-t}$,*

$$\varepsilon \le C\sigma \sqrt{\frac{\|M\|_{\mathrm{op}} (d + q + t)}{N}}.$$

Proposition 60 replaces the predictor dimension by the observed design rank under the same sub-Gaussian noise condition. The Gaussian effective-rank bound in Appendix S26 can be sharper under its stronger noise model.

Proof. Let $U \in \mathbb{R}^{p \times d}$ be an orthonormal basis for the row space of X. The columns of every $X_s^{\top} E_s$ lie in that row space, which $H^{-1/2}$ preserves by Lemma 3, hence

$$\widetilde{W} = U U^{\top} \widetilde{W}, \qquad \left\| \widetilde{W} \right\|_{\mathrm{op}} = \left\| U^{\top} \widetilde{W} \right\|_{\mathrm{op}}.$$

For unit $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^q$, conditional independence and the noise bound give

$$\left\| u^{\top} U^{\top} \widetilde{W} v \right\|_{\psi_2}^2 \le \frac{C\sigma^2}{N} u^{\top} U^{\top} M U u \le \frac{C\sigma^2 \|M\|_{\mathrm{op}}}{N}.$$

Lemmas 21 and 23 reduce the bound to at most 9^{d+q} pairs in $1/4$-nets. A union bound for the sub-Gaussian tails proves the result. This completes the proof. □

Combining the score bound with Theorem 1 gives a statistical prediction guarantee under a fixed, possibly singular design. Its tuning value depends on the fixed design but not on the responses.

Let $0 < \delta < 1$ and choose

$$\lambda_1 = C\sigma \sqrt{\frac{\|M\|_{\mathrm{op}} \{d + q + \log(2/\delta)\}}{N}}.$$

Proposition 60 and Theorem 1 give, for a sufficiently large absolute C, with conditional probability at least $1 - \delta$,

$$\frac{\left\| X(\widehat{B}_1 - B^{\star}) \right\|_F^2}{N} \le C' \sigma^2 \frac{\|M\|_{\mathrm{op}}}{\kappa_+} \frac{r_X \{d + q + \log(2/\delta)\}}{N}.$$

This bound separates low-rank dimension from conditioning on the observed predictor space. The dimension term agrees with Proposition 5 up to a constant when the effective condition number

$\|M\|_{\mathrm{op}}/\kappa_+$ is bounded and the confidence term is not larger than $d+q$. This is a high-probability upper bound; no expected-risk conclusion is inferred from a single confidence level.

To obtain the bound, let $t=\log(2/\delta)$ in Proposition 60 and substitute the resulting value of λ_1 into Theorem 1.

The empirical curvature in this bound can also be controlled from a singular population second-moment matrix. Let $\Sigma^0=\sum_s w_s\Sigma_s^0$. Keep the independence, upper second-moment, and sub-Gaussian conditions of Condition 1, but replace its full-rank lower bound by the following population specification. The matrix Σ^0 has rank $d_\Sigma\geq 1$, and its smallest positive eigenvalue is at least $\mu_+>0$. Let $\mathcal{S}=\mathrm{range}(\Sigma^0)$ and let $\Pi_{\mathcal{S}}$ be its orthogonal projection. In this section, define $\chi=\max_s\left\|(\Sigma^0)^{\dagger/2}\Sigma_s^0(\Sigma^0)^{\dagger/2}\right\|_{\mathrm{op}}$, where $\dagger$ denotes the Moore–Penrose inverse. Thus $1\leq\chi\leq L/\mu_+$, and the definition agrees with the full-rank case. For $0<\delta<1$, let

$$L_{\delta,0}=CLK_x^2\left\{1+\frac{d_\Sigma+\log(m/\delta)}{n_{\min}}\right\},\qquad k_{\delta,0}=\frac{\mu_+}{2(2+\tau)L_{\delta,0}},$$

$$\lambda_{\delta,0}=C\sigma\sqrt{\frac{d_\Sigma+q+\log(1/\delta)}{(1+\tau)N}},\qquad r_\Sigma=\mathrm{rank}(\Pi_{\mathcal{S}}B^\star).$$

Here C is a sufficiently large absolute constant.

Proposition 61. *If $N\gtrsim K_x^4\chi\{d_\Sigma+\log(1/\delta)\}$ and $\lambda_1=\lambda_{\delta,0}$, then, with probability at least $1-\delta$,*

$$\Pi_X=\Pi_{\mathcal{S}},\qquad d=d_\Sigma,\qquad \kappa_+\geq k_{\delta,0},\qquad \varepsilon\leq\lambda_1,\qquad \frac{\left\|X(\widehat{B}_1-B^\star)\right\|_F}{\sqrt{N}}\leq\frac{2\sqrt{2r_\Sigma}\lambda_1}{\sqrt{k_{\delta,0}}}.$$

Proposition 61 replaces positive definiteness in all p predictor directions by positive curvature only on the population support. This is a population assumption, not a condition imposed on the realized sample. Individual worker second-moment matrices may have different, deficient supports. No claim is made for coefficient directions outside $\mathcal{S}$.

Proof. For $v\in\ker(\Sigma^0)$, positivity and $w_s>0$ imply

$$0=v^\top\Sigma^0v=\sum_s w_sv^\top\Sigma_s^0v,\qquad \mathbb{E}(v^\top x_{si})^2=0.$$

Apply this to a finite basis of $\ker(\Sigma^0)$ to obtain $x_{si}\in\mathcal{S}$ almost surely. For an orthonormal basis U_0 of $\mathcal{S}$, let $z_{si}=U_0^\top x_{si}$. These independent sub-Gaussian rows have local second moments bounded by L and pooled second moment at least $\mu_+I_{d_\Sigma}$. Apply Propositions 18 and 20 in dimension d_Σ, each with failure probability $\delta/3$. The sample-size premise bounds the relative Gram error by $1/2$, giving

$$\lambda_{\min}(U_0^\top\widehat{\Sigma}U_0)\geq\mu_+/2,\qquad \max_s\ell_s\leq L_{\delta,0}$$

The heterogeneity factor is χ because $U_0(U_0^\top\Sigma^0U_0)^{-1/2}U_0^\top=(\Sigma^0)^{\dagger/2}$. The positive lower bound on $\mathcal{S}$ gives $\Pi_X=\Pi_{\mathcal{S}}$ and $d=d_\Sigma$. Also,

$$\lambda_{\max}(H)\leq(2+\tau)L_{\delta,0},\qquad \kappa_+\geq\frac{\mu_+}{2(2+\tau)L_{\delta,0}}.$$

Proposition 60 and $\|M\|_{\mathrm{op}}\leq(1+\tau)^{-1}$ give $\varepsilon\leq\lambda_1$ with conditional failure probability $\delta/3$. Constants absorb the logarithmic factors because $d_\Sigma,q\geq 1$. A union bound, Theorem 1, and $r_X=r_\Sigma$ complete the argument. This completes the proof. □

The same subspace reduction applies to persistent rank recovery. Its target is the rank of the fitted mean matrix, which need not equal the original coefficient rank. For $1\leq r_X<d\wedge q$ and $R>0$, let $e_{X,t}=\|C_t-C_X^\star\|_F$, $s_X=\sigma_{r_X}(C_X^\star)$, $q_+=(1-\kappa_+)/(1-\kappa_+/2)$, and $\bar{\eta}_X=(1-\kappa_+)R+\varepsilon$.

Corollary 33. *Let $1\leq r_X<d\wedge q$, $t_0\geq T_1$, and $R>0$. Suppose*

$$e_{X,t_0}\leq R,\qquad R\geq\frac{2\sqrt{2r_X}\varepsilon}{\kappa_+},\qquad \bar{\eta}_X<\lambda_2,\qquad s_X>a\lambda_2+\bar{\eta}_X,\qquad \bar{\eta}_X\leq\frac{\kappa_+s_X}{1+\kappa_+}.$$

Then, for every integer $k \geq 1$,

$$\operatorname{rank}(C_{t_0+k}) = r_X, \qquad e_{X,t_0+k} \leq q_+^k e_{X,t_0} + \frac{2\sqrt{2r_X}\varepsilon}{\kappa_+}(1-q_+^k).$$

Corollary 33 concerns the unchanged computed SCAD sequence. It is a deterministic implication, not an assumed random event. Theorem 1 gives its initial error bound, while Proposition 61 gives population control of the score and curvature. Additional signal separation is still needed; the corollary does not assert uniform recovery of $\operatorname{rank}(B^\star)$ under a singular design.

Proof. Choose U as in Proposition 60 and let

$$\check{C}_t = U^\top C_t, \qquad \check{C}^\star = U^\top C_X^\star, \qquad \check{M} = U^\top M U, \qquad \check{W} = U^\top \widetilde{W}.$$

By Lemma 3, multiplication by U preserves the Frobenius norms, nonzero singular values, and ranks. The reduced update satisfies

$$\check{Z}_t = \check{C}^\star + (I_d - \check{M})(\check{C}_t - \check{C}^\star) + \check{W}, \qquad \lambda_{\min}(\check{M}) = \kappa_+, \qquad \left\|\check{W}\right\|_{\mathrm{op}} = \varepsilon.$$

The SCAD update and singular-value gaps are unchanged. The proof of Theorem 7, using Lemma 6, gives

$$e_{X,t+1} \leq \frac{(1-\kappa_+)e_{X,t} + \sqrt{2r_X}\varepsilon}{1-\kappa_+/2}$$

for $e_{X,t} \leq R$. The lower bound on R makes this radius invariant. Induction and summation prove the assertions. This completes the proof. □

The preceding result can be stated using only population parameters, the tuning values, and the prescribed number of Stage I iterations. This removes the need to assume an empirical entry error. Under the population model of Proposition 61, let $1 \leq r_\Sigma < d_\Sigma \wedge q$, let $B_0 = \Pi_{\mathcal{S}} B^\star$, and define

$$A_\delta = \frac{2\sqrt{2r_\Sigma}\lambda_{\delta,0}}{k_{\delta,0}}, \qquad B_\delta = \sqrt{(2+\tau)L_{\delta,0}}\,\|B_0\|_F, \qquad s_\delta = \sqrt{(1+\tau)\mu_+/2}\,\sigma_{r_\Sigma}(B_0),$$

$$E_\delta(T) = A_\delta + \frac{2(B_\delta + A_\delta)}{\sqrt{k_{\delta,0}}(T+1)}, \qquad q_\delta = \frac{1-k_{\delta,0}}{1-k_{\delta,0}/2}, \qquad \eta_\delta = (1-k_{\delta,0})R + \lambda_{\delta,0}.$$

All these quantities are deterministic for a prescribed $R > 0$ and confidence level $0 < \delta < 1$.

Corollary 34. *Under the model and sample-size condition of Proposition 61, let $\lambda_1 = \lambda_{\delta,0}$ and $T_1 \geq 1$. Suppose*

$$R \geq E_\delta(T_1), \qquad \eta_\delta < \lambda_2, \qquad s_\delta > a\lambda_2 + \eta_\delta, \qquad \eta_\delta \leq \frac{k_{\delta,0}s_\delta}{1+k_{\delta,0}}.$$

With probability at least $1-\delta$, for every integer $k \geq 1$,

$$\operatorname{rank}(C_{T_1+k}) = r_\Sigma, \qquad e_{X,T_1+k} \leq q_\delta^k E_\delta(T_1) + A_\delta(1-q_\delta^k).$$

Corollary 34 connects the population model to every computed Stage II iterate after the prescribed initialization. The signal and iteration conditions are sufficient and can be conservative. The recovered rank is that of B_0, not that of an arbitrary coefficient component outside the population support.

Proof. The joint event in Proposition 61 gives

$$\Pi_X = \Pi_{\mathcal{S}}, \quad r_X = r_\Sigma, \quad \kappa_+ \geq k_{\delta,0}, \quad \varepsilon \leq \lambda_{\delta,0}, \quad \lambda_{\max}(H) \leq (2+\tau)L_{\delta,0}.$$

Also, $\bar{\ell} \geq \lambda_{\max}(\widehat{\Sigma}) \geq \mu_+/2$, hence $H \succeq (1+\tau)\mu_+ I_p/2$. Because $C_X^\star = H^{1/2}B_0$, singular-value monotonicity gives

$$\|C_X^\star\|_F \leq B_\delta, \qquad s_X \geq s_\delta.$$

Theorem 1 gives $\left\|\widehat{C} - C_X^\star\right\|_F \leq A_\delta$ and $\left\|\widehat{C}\right\|_F \leq B_\delta + A_\delta$. Its finite-iteration bound implies

$$e_{X,T_1} \leq E_\delta(T_1) \leq R, \qquad R \geq A_\delta \geq \frac{2\sqrt{2r_X}\varepsilon}{\kappa_+}.$$

For $e_{X,t} \le R$, the perturbation norm is at most η_δ. The gaps give rank r_X and the projection-curvature bound in Corollary 33. Monotonicity of $u/(1+u)$ transfers the last premise to κ_+, s_X. Applying that proof with curvature $k_{\delta,0}$ gives

$$e_{X,t+1} \le \frac{(1-k_{\delta,0})e_{X,t} + \sqrt{2r_\Sigma}\lambda_{\delta,0}}{1-k_{\delta,0}/2} = q_\delta e_{X,t} + (1-q_\delta)A_\delta.$$

Because $0 < k_{\delta,0} \le \kappa_+ < 1$, $0 < q_\delta < 1$. Also $E_\delta(T_1) \ge A_\delta$. Induction preserves $e_{X,t} \le R$, and summation gives the bound on the same event for all iterations. This completes the proof. □

S38.1 Population conditions for the Stage II least-squares limit

Theorem 9 has a deterministic entry premise. The following corollary derives that premise from the original zero initialization. Use the population model and constants $L_{\alpha,0}, k_{\alpha,0}, \lambda_{\alpha,0}, E_\alpha, s_\alpha, q_\alpha, \eta_\alpha$ defined above at confidence level $\alpha = \delta/2$. The tuning values, T_1, and R are prescribed quantities.

Corollary 35. *Suppose* $1 \le r_\Sigma < d_\Sigma \wedge q$, $N \gtrsim K_x^4 \chi\{d_\Sigma + \log(2/\delta)\}$, $\lambda_1 = \lambda_{\delta/2,0}$, *and* $T_1 \ge 1$. *If*

$$R \ge E_{\delta/2}(T_1), \qquad \eta_{\delta/2} < \lambda_2, \qquad s_{\delta/2} > a\lambda_2 + \eta_{\delta/2}, \qquad \eta_{\delta/2} \le \frac{k_{\delta/2,0}s_{\delta/2}}{2+k_{\delta/2,0}},$$

then, with probability at least $1-\delta$, *the conclusions of Theorem 9 hold with* $t_0 = T_1$ *and* $r_X = r_\Sigma$. *Simultaneously, for every integer* $j \ge 1$,

$$\frac{\left\|X(H^{-1/2}C_{T_1+j} - B^\star)\right\|_F}{\sqrt{N}} \le C\sigma\sqrt{\frac{r_\Sigma\{d_\Sigma + q + \log(2/\delta)\}}{N}} + \frac{2R}{\sqrt{1+\tau}}q_{\delta/2}^j.$$

Corollary 35 gives a statistical guarantee for the returned estimate using population signal, design, noise, tuning, and iteration conditions. Worker second-moment matrices may be different and singular. The result does not require a condition imposed on the fitted sample, and it does not justify arbitrary validation-selected tuning values.

Proof. Apply Proposition 61 with failure probability $\delta/2$. The proof of Corollary 34 gives on that event

$$\Pi_X = \Pi_{\mathcal{S}}, \quad d = d_\Sigma, \quad r_X = r_\Sigma, \quad \kappa_+ \ge k_{\delta/2,0}, \quad \varepsilon \le \lambda_{\delta/2,0}, \quad s_X \ge s_{\delta/2},$$

$$e_{X,T_1} \le E_{\delta/2}(T_1) \le R, \qquad R \ge \frac{2\sqrt{2r_X}\varepsilon}{\kappa_+}, \qquad \bar\eta_X \le \eta_{\delta/2}.$$

The function $u/(2+u)$ is increasing for $u > 0$. The last premise gives

$$\bar\eta_X \le \eta_{\delta/2} \le \frac{k_{\delta/2,0}s_{\delta/2}}{2+k_{\delta/2,0}} \le \frac{\kappa_+ s_X}{2+\kappa_+}.$$

The other separation premises transfer in the same direction. Theorem 9 applies. Also, $q_+ \le q_{\delta/2}$ because $(1-u)/(1-u/2)$ decreases on $(0,1)$. Apply Proposition 2 with conditional failure probability $\delta/2$, integrate over the designs, and use a union bound. On the resulting event $d = d_\Sigma$ and $\|M\|_{\mathrm{op}} \le (1+\tau)^{-1}$. Substitution into Theorem 9 proves the inequality for every j on the same event. This completes the proof. □

The conditions have nonzero examples. Consider fixed dimensions and a fixed number of workers, with $n_{\min} \to \infty$, fixed positive curvature on $\mathcal{S}$, and $\sigma_{r_\Sigma}(B_0) > 0$. Then $k_{\delta/2,0}$ is bounded away from zero, $\lambda_{\delta/2,0} \to 0$, and $s_{\delta/2}$ is bounded away from zero. Choose T_1 such that $E_{\delta/2}(T_1) \le 2A_{\delta/2}$ and let $R = 2A_{\delta/2}$. For positive noise scale, $\eta_{\delta/2} \to 0$ and $\lambda_2 = 2\eta_{\delta/2}$ satisfies the three separation inequalities for sufficiently large N. The worker distributions can have different deficient supports whose span is $\mathcal{S}$. This example shows that the population conditions are attainable. It does not give uniformly weak signal requirements or ensure communication savings before the round threshold.

S38.2 A POPULATION REGIME WITH STATISTICAL ACCURACY AND LOWER COMMUNICATION

The preceding population conditions can hold while every local design is singular and the complete two-stage fit transmits fewer scalars than sufficient-statistic aggregation. Let q, $1 \le r_0 < q$, $0 < \delta < 1$, $a > 2$, $\tau > 0$, and positive population bounds $\mu, L, K_x, \sigma, b, B$ be fixed. Let p increase through the even integers, with $m = p$ and $n_s = p/2$ for every worker. Use the constants of Proposition 61 with $d_\Sigma = p$ and $\mu_+ = \mu$. Prescribe

$$R = \frac{4\sqrt{2r_0}\lambda_{\delta/2,0}}{k_{\delta/2,0}}, \qquad T_1 = \lceil K\sqrt{p} \rceil, \qquad T_2 = \lceil \log p \rceil, \qquad \lambda_1 = \lambda_{\delta/2,0}, \qquad \lambda_2 = 2\eta_{\delta/2}.$$

These tuning values use the stated population bounds and the rank upper bound r_0, not the unknown rank or singular vectors of $B^\star$.

Corollary 36. *Under Condition 1, suppose*

$$1 \le r = \operatorname{rank}(B^\star) \le r_0, \qquad \sigma_r(B^\star) \ge b, \qquad \|B^\star\|_F \le B.$$

There are $K, C, p_0 > 0$, depending only on the fixed bounds, such that for $p \ge p_0$, with probability at least $1 - \delta$,

$$\operatorname{rank}(C_{T_1+j}) = r \quad (j \ge 1), \qquad \frac{\left\|X(H^{-1/2}C_{T_1+T_2} - B^\star)\right\|_F^2}{N} \le \frac{C}{p}.$$

Every local design is singular. The two-stage scalar count per remote worker satisfies

$$\frac{1 + 2(T_1 + T_2)pq}{p(p+1)/2 + pq} \le \frac{C}{\sqrt{p}}.$$

Corollary 36 is an explicit consequence of the statistical and communication bounds, not a new communication lower bound. It shows that their sufficient regimes intersect with a growing number of workers and prescribed tuning. The comparison concerns one complete two-stage fit and does not establish an advantage for an arbitrary tuning search. Corollary 10 gives the earlier comparison for the first rank-correct update; the present result also controls the returned prediction and every later rank.

Proof. Here $N = p^2/2$, $n_{\min} = p/2$, and $\chi \le L/\mu$. The population constants satisfy, for all sufficiently large p,

$$0 < c \le k_{\delta/2,0} < 1, \qquad c \le L_{\delta/2,0} \le C, \qquad \frac{c}{\sqrt{p}} \le \lambda_{\delta/2,0} \le \frac{C}{\sqrt{p}}.$$

The lower bound on $\lambda_{\delta/2,0}$ uses the fixed positive noise bound σ. The sample-size condition in Corollary 35 holds because $N = p^2/2$ and χ is bounded. Also $\mathcal{S} = \mathbb{R}^p$, $B_0 = B^\star$, and

$$A_{\delta/2} \le R/2, \qquad B_{\delta/2} \le CB, \qquad s_{\delta/2} \ge \sqrt{(1+\tau)\mu/2}\, b > 0.$$

Thus, uniformly over the stated coefficient class,

$$E_{\delta/2}(T_1) \le R/2 + \frac{C}{K\sqrt{p}} \le R$$

for a sufficiently large fixed K, because $R \ge c/\sqrt{p}$. Also $\eta_{\delta/2} > 0$ and $\eta_{\delta/2} \le C/\sqrt{p}$. The choice $\lambda_2 = 2\eta_{\delta/2}$ gives

$$\eta_{\delta/2} < \lambda_2, \qquad a\lambda_2 + \eta_{\delta/2} < s_{\delta/2}, \qquad \eta_{\delta/2} \le \frac{k_{\delta/2,0} s_{\delta/2}}{2 + k_{\delta/2,0}}$$

for all sufficiently large p. Corollary 35 proves the rank assertion and, using $R \le C/\sqrt{p}$,

$$\frac{\left\|X(H^{-1/2}C_{T_1+T_2} - B^\star)\right\|_F}{\sqrt{N}} \le C\sqrt{\frac{p + q + \log(2/\delta)}{p^2}} + \frac{2R}{\sqrt{1+\tau}} q_{\delta/2}^{T_2} \le \frac{C}{\sqrt{p}}.$$

Squaring proves the prediction bound. The deterministic inequality $\operatorname{rank}(X_s) \le n_s = p/2 < p$ proves local singularity. Finally, $T_1 + T_2 \le C\sqrt{p}$ and Proposition 1 give

$$\frac{1 + 2(T_1 + T_2)pq}{p(p+1)/2 + pq} \le \frac{2}{p^2} + \frac{4q(T_1 + T_2)}{p} \le \frac{C}{\sqrt{p}}.$$

The same ratio holds for the total network counts because both per-worker totals are multiplied by $m - 1$. This completes the proof. □

The population class includes unequal and singular worker second moments. For example, let P project onto the first $p/2$ coordinates, take half the worker covariances equal to $2P$ and half equal to $2(I_p - P)$, and use independent centered Gaussian rows. Then $\Sigma^0 = I_p$, every local population covariance is singular, and the sub-Gaussian and curvature bounds are uniform in p. Independent centered Gaussian error rows satisfy the stated noise condition. The coefficient class is nonempty whenever $B \ge b\sqrt{r}$ for at least one $1 \le r \le r_0$.

S38.3 Prediction for new observations

The population second-moment bounds also control prediction for a new observation. Let x_{new} be independent of the training sample and have the mixture of worker predictor distributions with weights w_s. Its second-moment matrix is Σ^0. For a training-sample estimator $\widetilde{B}$, the conditional mean squared error in predicting ${B^\star}^\top x_{\mathrm{new}}$ is $\left\|(\Sigma^0)^{1/2}(\widetilde{B} - B^\star)\right\|_F^2$. Coefficient components outside $\mathcal{S}$ do not contribute to this loss.

Under the model and sample-size condition of Proposition 61, let $\lambda_1 = \lambda_{\delta,0}$. Its proof gives, with probability at least $1 - \delta$,

$$\tfrac{1}{2}\Sigma^0 \preceq \widehat{\Sigma} \preceq \tfrac{3}{2}\Sigma^0, \qquad \Pi_X = \Pi_{\mathcal{S}}, \qquad \kappa_+ \ge k_{\delta,0}, \qquad \varepsilon \le \lambda_{\delta,0}.$$

The second-moment matrices equal zero on $\mathcal{S}^\perp$, hence these inequalities hold on $\mathbb{R}^p$. The lower bound controls population prediction loss by twice its empirical counterpart. The upper bound and Theorem 5 then give

$$\left\|(\Sigma^0)^{1/2}(\widehat{B}_1 - B^\star)\right\|_F^2 \le \inf_{B \in \mathbb{R}^{p\times q}} \left\{ 3\left\|(\Sigma^0)^{1/2}(B - B^\star)\right\|_F^2 + \frac{16\lambda_{\delta,0}^2 \operatorname{rank}(\Pi_{\mathcal{S}} B)}{k_{\delta,0}} \right\}.$$

Here $\operatorname{rank}(XB) = \operatorname{rank}(\Pi_{\mathcal{S}} B)$ because X is injective on $\mathcal{S}$. The bound compares the convex target with every approximation to the population response mean. The sample-size condition ensures that the observed design represents every population predictor direction that contributes to prediction.

Under the conditions of Corollary 34, the same comparison applies to every computed Stage II iterate. On the event in its proof, $C_X^\star = H^{1/2}\Pi_{\mathcal{S}} B^\star$ and $M \preceq (1+\tau)^{-1} I_p$, giving

$$\left\|(\Sigma^0)^{1/2}(H^{-1/2}C_t - B^\star)\right\|_F^2 \le 2\left\|M^{1/2}(C_t - C_X^\star)\right\|_F^2 \le \frac{2e_{X,t}^2}{1+\tau}.$$

Substituting the error bound in that corollary proves, with probability at least $1 - \delta$, simultaneously for every integer $k \ge 1$,

$$\left\|(\Sigma^0)^{1/2}(H^{-1/2}C_{T_1+k} - B^\star)\right\|_F \le \sqrt{\frac{2}{1+\tau}}\left\{q_\delta^k E_\delta(T_1) + A_\delta(1 - q_\delta^k)\right\}.$$

This controls the returned estimator under the stated signal and iteration conditions. It does not imply that each Stage II update reduces test error.

S39 Consequences for numerical interpretation

The results in this supplement determine what each reported numerical quantity can support. Exact rank frequency in the two designs of Table 1 evaluates the conclusion of Theorem 4. Correct-rank frequency in Table 10 compares the returned rank with r_X, the target of Corollary 33. Relative

Frobenius error evaluates coefficient recovery and hence concerns settings with pooled identifiability. Test response mean squared error evaluates prediction and remains meaningful without coefficient identifiability. Appendix S38.3 controls the corresponding error in the conditional mean. For a new response independent of the training sample, conditional mean-zero noise makes expected response MSE equal to this mean prediction loss divided by q, plus the average response noise variance. The communication column evaluates Proposition 1 for one fit at the selected tuning values.

The public SCM20d analysis has no known true coefficient rank. Its selected-rank column is a descriptive model-complexity measure, not a rank-recovery probability. Its paired test errors compare predictive performance under identical splits. Proposition 33 shows why these tuned public-data fits are not used to claim an end-to-end communication advantage.

Proposition 31 gives a positive aligned-covariance result, while Proposition 32 excludes a universal optimization advantage over an isotropic majorizer. The metric, master, and stabilization ablations quantify these choices in the locally singular design. SCM20d and OES10 provide two public-data evaluations, and the socket experiment compares communication time across separate processes on one host. Generality beyond these multi-target benchmarks and timing on a physical multi-node system remain unmeasured.

S40 Numerical and implementation details

This appendix records the experimental settings, tuning rules, metrics, and implementation checks. All reported summaries use 100 replications. The proposed estimator uses $\tau = 0.10$, unless the stabilization value is varied explicitly, and each SCAD fit uses $a = 3.7$.

For the public datasets, paired differences compare methods on the same training and test split. Their standard errors describe variation over the reported random splits of the fixed dataset. They do not quantify variation over newly sampled datasets. Selected ranks describe fitted model complexity because the true coefficient rank is unknown.

S40.1 Synthetic experiments

For each replication, we generate $B^\star = UDV^\top$. The matrices U and V are the first r columns of the thin QR factors of independent standard Gaussian matrices of dimensions $p \times r$ and $q \times r$. The diagonal entries of D are $(6, 5, 4)$ in the locally singular setting and $(4, 3)$ in the communication setting. In the first setting, let $b_s = 2(s-1)/7 - 1$. Worker s has autoregressive correlation $\rho_s = 0.30 + 0.25 b_s$ and predictor scale $d_{sj} = \exp[\{-0.20 + 0.40(j-1)/119\} b_s]$; its population covariance has entries $d_{sj} d_{sk} \rho_s^{|j-k|}$. In the second setting, every worker has covariance entries $0.2^{|j-k|}$. Independent Gaussian errors have variance one. Conditional on $B^\star$, these designs and errors satisfy Condition 1. Independent validation and test samples contain 1,600 and 4,000 observations, with equal numbers from every worker distribution.

The locally singular setting fixes $\lambda_1/\lambda_{\max} = 0.04$, $\lambda_2/\lambda_1 = 16$, $T_1 = 30$, and $T_2 = 20$. The communication setting fixes these values at 0.02, 12, 15, and 5. Here $\lambda_{\max} = \left\| H^{-1/2} G(0) \right\|_{\mathrm{op}}$. Centralized and one-shot penalties use independent validation samples. The centralized SCAD iterations use the exact proximal map of the step-size-scaled penalty. D-RRR uses adaptive exponent 10 and is evaluated only when the master design has full column rank. Its communication total counts one $p \times q$ broadcast and one $p \times q$ gradient return per selected distributed update; its master-local initializer requires no remote transmission. ADMM uses augmented Lagrangian penalty parameter $\rho = 1$ and the centralized nuclear-norm tuning value. FedDualAvg uses all workers, complete local gradients, and one or five local steps. Validation selects the nuclear penalty, client learning rate, server learning rate, and local step count. The two learning rates use the 7×7 grid in Yuan et al. (2021); the rounds equal those of the proposed fit.

Relative Frobenius error is $\left\| \widehat{B} - B^\star \right\|_F / \left\| B^\star \right\|_F$. Test response mean squared error averages all test observations and responses. Rank uses singular values above 10^{-5}. Table 3 gives the complete paired comparisons for test MSE and relative Frobenius error.

Table 3: Paired proposed-minus-comparator differences over 100 replications. SE denotes standard error.

Setting	Comparator	Δ MSE (SE)	Lower MSE (%)	Δ error (SE)	Lower error (%)
Locally singular	Stage I	−.05676 (.00072)	100	−.03463 (.00040)	100
	FedDualAvg	−.06186 (.00085)	100	−.03818 (.00049)	100
	Central nuclear	−.06037 (.00082)	100	−.03720 (.00047)	100
	Central SCAD	.00058 (.00035)	44	.00052 (.00028)	38
	One-shot	−1.77005 (.00875)	100	−.44903 (.00140)	100
	ADMM	−.06180 (.00084)	100	−.03813 (.00048)	100
Communication	Stage I	−.01048 (.00024)	100	−.00931 (.00010)	100
	FedDualAvg	−.00615 (.00023)	100	−.00562 (.00016)	100
	Central nuclear	−.00992 (.00023)	100	−.00900 (.00010)	100
	Central SCAD	−.00019 (.00007)	61	−.00019 (.00005)	64
	One-shot	−.02806 (.00039)	100	−.02332 (.00019)	100
	ADMM	−.01010 (.00023)	100	−.00918 (.00010)	100
	D-RRR	−.18442 (.00224)	100	−.10067 (.00087)	100

S40.2 Rank persistence along the computed sequence

We examine whether the rank selected by the first SCAD update is kept through the returned estimate. The two synthetic designs, tuning parameters, and stopping rules are those specified above. For each of the 100 replications, we record all 20 Stage II updates in the locally singular setting and all five in the communication setting. Initial and final errors in Table 4 are $\|C_{T_1} - C^\star\|_F$ and $\|C_{T_1+T_2} - C^\star\|_F$. Rank retention and absence of shrinkage are counted once per replication and require the corresponding property at every Stage II update.

Table 4: Rank persistence and transformed coefficient error. Numbers in parentheses represent standard deviations.

Setting	Stage II rounds	Rank kept (%)	Unshrunk (%)	Initial error	Final error
Locally singular	20	100	100	4.0567 (0.1356)	3.1788 (0.1216)
Communication	5	100	100	0.7366 (0.0191)	0.6462 (0.0186)

Table 4 shows that the first SCAD update recovers the true rank and every subsequent update keeps it in both designs. The observed two-threshold gap holds at every update, and the SCAD map (5) leaves all kept singular values unchanged. The 95% Wilson lower bound for the probability of maintaining rank throughout Stage II is 0.9630 in each design. After the first Stage II update, the mean transformed coefficient error decreases from 3.6814 to 3.1788 in the locally singular setting and from 0.6665 to 0.6462 in the communication setting. These reductions show that the later iterations improve estimation within the selected rank. The final test error is lower than that of Stage I in all 100 replications in both designs, as reported in Table 3.

The observed spectral gap certifies the computed rank and lack of shrinkage, while the known simulation coefficient determines whether the rank is correct. Thus the path results demonstrate persistent rank recovery without asserting that the population sufficient conditions hold for the chosen tuning values.

S40.3 Fitting error of the returned Stage II estimate

We examine whether the returned estimate is close to the best least-squares fit of its selected rank. For each of the 100 replications in the two designs of Table 1, we evaluate the singular-value conditions and fitting bound in Theorem 8. The comparison rank is the rank selected by the estimator, without using the true rank. Table 5 reports the loss difference from the global rank-constrained least-squares fit and its upper bound. Numbers in parentheses represent standard deviations.

Table 5: Finite-iteration comparison with rank-constrained least squares.

Setting	Loss gap	Upper bound
Locally singular	.017287 (.003322)	.157449 (.163940)
Communication	.000679 (.000031)	.001896 (.000074)

The finite-iteration conditions hold for every returned estimate in both designs, and every observed loss gap is below its bound. The ratio of the actual gap to the bound is at most .1875 in the locally singular design and .3784 in the communication design. These comparisons quantify the fitting accuracy of the returned estimate in addition to its rank recovery. Training sufficient statistics are used to evaluate curvature and the least-squares comparison. These calculations are separate from the distributed fit and are not included in its communication count.

In the communication setting, the future-rank condition in Theorem 15 holds in all 100 replications. This conclusion concerns that setting. The prediction and rank results under singular pooled designs in Table 10 do not establish that the bound in Theorem 8 applies. All condition evaluations use numerical singular values and eigenvalues and are numerical diagnostics, without certified rounding-error bounds.

S40.4 SCM20D DATA

The SCM20d analysis uses the 8,966 observations distributed with the Mulan multi-target regression collection. Each replication samples 8,000 training observations and uses the remaining 966 for testing. Centering and scaling parameters are estimated from the training observations. The training sample is randomly divided into ten equal worker blocks. Within each block, 10 percent is reserved for tuning. The proposed grid uses $\lambda_1/\lambda_{\max} \in \{0.005, 0.01, 0.02, 0.04, 0.07, 0.10, 0.15\}$ and $\lambda_2/\lambda_1 \in \{2, 4, 6, 8, 10, 12, 16\}$. The Stage I value minimizes validation error after at most 100 rounds, and the Stage II ratio minimizes validation error after at most 200 exact SCAD updates from that fit. The selected values are refitted on all 8,000 training observations using 100 Stage I rounds and 200 Stage II rounds. Centralized and one-shot methods use the same tuning observations. D-RRR uses five-fold training-only tuning within each worker. Test observations are not used in preprocessing or selection. Selected rank counts singular values above 10^{-4}.

The paired differences in Table 6 compare the methods on the same 100 splits.

Table 6: Paired SCM20d test MSE comparisons over 100 splits.

Comparator	Proposed minus comparator	Standard error	Proposed lower (%)
Stage I	−.001628	.000072	98
Central nuclear	−.003206	.000108	100
Central SCAD	−.000292	.000033	81
One-shot	−.002915	.000111	100
D-RRR	−.008527	.001649	97

S40.5 METRIC, MASTER, AND STABILIZATION ABLATION

The ablation uses a separate set of 100 independent replications of the locally singular synthetic design and the same 30 Stage I and 20 Stage II rounds. Each variant fixes $\lambda_1/\lambda_{\max} = .04$ and $\lambda_2/\lambda_1 = 16$, where $\lambda_{\max}$ is computed in the metric used by that variant. We compare $H = \widehat{\Sigma}_s + (1+\tau)\bar{\ell} I_p$ across all eight choices of master and compare it with $H = (1+\tau)\bar{\ell} I_p$ for $\tau \in \{.02, .10, .50, 1\}$. The transformed condition number is the ratio of the largest and smallest eigenvalues of $H^{-1/2}\widehat{\Sigma}H^{-1/2}$.

Table 7 reports the complete results. At $\tau = .10$, worker 1 gives lower test error and relative Frobenius error than the isotropic metric in 71 and 72 replications. The local metric has a smaller transformed condition number for every master in this design. At each stabilization value, the local metric from worker 1 gives lower mean prediction and coefficient errors than the isotropic metric.

Table 7: Metric, master, and stabilization results over 100 replications. Parentheses contain standard errors. Every fit selects rank three.

Metric and master	τ	Test MSE	Relative error	Condition number
Local, worker 1	.02	1.09040 (.00097)	.12817 (.00058)	14.28
Local, worker 1	.10	1.09044 (.00097)	.12821 (.00058)	14.32
Local, worker 1	.50	1.09044 (.00097)	.12818 (.00058)	14.49
Local, worker 1	1.00	1.09080 (.00096)	.12848 (.00059)	14.65
Isotropic	.02	1.09083 (.00096)	.12852 (.00058)	15.57
Isotropic	.10	1.09092 (.00096)	.12860 (.00058)	15.57
Isotropic	.50	1.09096 (.00097)	.12862 (.00058)	15.57
Isotropic	1.00	1.09122 (.00096)	.12883 (.00059)	15.57
Local, worker 2	.10	1.09047 (.00096)	.12819 (.00058)	14.08
Local, worker 3	.10	1.09036 (.00096)	.12812 (.00058)	13.76
Local, worker 4	.10	1.09039 (.00096)	.12813 (.00058)	13.48
Local, worker 5	.10	1.09031 (.00095)	.12805 (.00057)	13.06
Local, worker 6	.10	1.09022 (.00096)	.12801 (.00058)	12.70
Local, worker 7	.10	1.09020 (.00096)	.12797 (.00058)	12.15
Local, worker 8	.10	1.09009 (.00096)	.12789 (.00058)	11.79

S40.6 OES10 DATA

OES10 is distributed by the same Mulan collection as SCM20d. It contains 403 observations, 298 predictors, and 16 responses. Each replication samples 200 training observations, assigns 25 observations to each of eight workers, and uses the remaining 203 observations for testing. Predictor and response means and standard deviations are estimated from the training sample. Each worker reserves five observations for tuning. The proposed and centralized grids use $\lambda/\lambda_{\max} \in \{.002, .005, .01, .02, .04, .07, .10, .15, .22\}$, with each grid normalized by the corresponding method's zero threshold, and the proposed Stage II ratios are $\{2, 4, 6, 8, 10, 12, 16\}$. The selected Stage I fit uses 100 rounds, and the selected Stage II ratio is applied for the one update used during validation. One-shot averaging uses the same tuning observations. D-RRR is undefined because each master design has fewer rows than columns. Test observations are not used in preprocessing or tuning. Selected rank counts singular values above 10^{-4}.

Table 8 shows lower mean error for the proposed estimator than the three external comparators. Its Stage I initialization attains nearly the same error, with a mean difference below 5×10^{-5}, and hence the gain over the external comparators is already present after the convex stage.

Table 8: OES10 results over 100 paired splits. Parentheses contain standard deviations for method summaries and standard errors for paired differences.

Method	Test MSE	Rank	Proposed minus method	Proposed lower (%)
Proposed	.18062 (.08603)	14.13 (2.22)	—	—
Stage I	.18058 (.08588)	14.20 (2.15)	.00004 (.00006)	59
Central nuclear	.18330 (.08690)	13.22 (1.81)	−.00268 (.00068)	69
Central SCAD	.18308 (.08634)	13.16 (1.95)	−.00246 (.00059)	73
One-shot	.20111 (.10180)	16.00 (.00)	−.02049 (.00205)	87

S40.7 SOCKET TIMING

The timing study uses one master R process and four worker R processes connected by parallel socket connections on one physical host. The host has a 16-core Apple M4 Max processor and 128 GB memory and runs macOS 26.2. Every process holds its row block before timing begins; process creation and initial data placement are excluded. The setting has $n_s = 200$, $p = 1500$, $q = 2$, and rank two. The proposed fit uses eight Stage I rounds and two Stage II rounds. Both protocols include the worker spectral-radius calculation. The gradient protocol then times ten synchronous gradient exchanges. The competing protocol times each local covariance and cross-product calculation, transmission of the packed sufficient statistics, master aggregation, and the same

ten local spectral updates. The scalar total in Table 9 includes the spectral-radius scalar for both protocols. Proposition 1 omits this scalar for sufficient-statistic aggregation because the master can recover it from the transmitted covariance. Protocol order alternates across replications, and the numerical libraries use one thread per process.

Table 9 reports the timing results. The gradient protocol is faster in all 100 paired replications, and the geometric mean of the sufficient-statistic to gradient time ratio is 11.02. The maximum relative difference between the two computed coefficient matrices is 7.07×10^{-15}.

Table 9: Single-host socket timing over 100 replications. Parentheses contain standard errors.

Protocol	Mean seconds	Median seconds	Scalars per remote worker
Gradient	.17690 (.00753)	.1580	60,001
Sufficient statistics	1.92111 (.05888)	1.8100	1,128,751

S40.8 Prediction and rank under singular pooled designs

A singular pooled design identifies only part of the coefficient matrix. We examine whether the computed estimator estimates this component and reduces prediction error after the convex stage. We use four workers with 20 observations each, $p = 120$, $q = 12$, and independent standard Gaussian response errors. Each of the four settings uses 100 independent replications. Coefficient factors and population support bases are fixed before generating the observations.

In the first setting, each worker has independent $N(0, I_{120})$ predictor rows. The pooled design has rank 80, and the coefficient matrix has rank three with singular values $(6, 5, 4)$. In the second setting, the population covariance has rank 40. We fix an orthonormal basis for its support, assign ten distinct basis directions to each worker, and use independent Gaussian coordinates with variance four in those directions. The pooled population covariance has 40 positive eigenvalues equal to one. The coefficient component on its support has rank three and singular values $(6, 5, 4)$. We add a coefficient component of unit Frobenius norm in an orthogonal predictor direction and an orthogonal response direction. The full coefficient matrix has rank four, while the fitted mean has rank three. The added component contributes zero to the response mean under each worker population.

The third setting keeps the same supports and coefficient matrix but uses 40 positive pooled eigenvalues equally spaced from 0.1 to one. Each worker variance is four times the corresponding pooled eigenvalue. The fourth setting keeps the second covariance and reduces the three identified coefficient singular values to $(3, 2.5, 2)$. These two settings assess sensitivity to weaker curvature and weaker signal while keeping the same population support and rank.

Independent validation and test samples have 1,600 and 4,000 observations, with equal numbers from the worker populations. The Stage I penalty uses fractions $0.015, 0.025, 0.04, 0.07, 0.10, 0.15, 0.22$ of $\left\|H^{-1/2}G(0)\right\|_{\mathrm{op}}$. Validation first selects the Stage I estimate after 100 iterations. Starting from this estimate, validation selects the returned Stage II estimate over $\lambda_2/\lambda_1 \in \{2, 4, 6, 8, 10, 12, 16\}$. Stage II has at most 20 iterations and relative coefficient-change tolerance 10^{-7}. The centralized nuclear and SCAD comparators use the same seven penalty fractions of $\left\|\widehat{Z}\right\|_{\mathrm{op}}$, at most 500 iterations, and tolerance 10^{-6}. All four methods use the same validation and test observations. Rank counts singular values above 10^{-5}.

Prediction loss is $\left\|X(\widehat{B} - B^\star)\right\|_F^2 / N$, as in Theorem 1. Test MSE averages squared response errors over the independent test observations and responses. Correct-rank frequency compares the returned rank with $r_X = \mathrm{rank}(XB^\star) = 3$. In the three singular-population settings, this target differs from the full coefficient rank four.

Table 10: Prediction and identifiable-rank selection under singular pooled designs. Numbers in parentheses represent standard deviations.

Method	Prediction loss	Test MSE	Rank	Correct rank (%)
Full population covariance; $d = 80$, $r_X = 3$				
Proposed	3.1880 (0.2692)	3.5198 (0.2674)	3.00 (0.00)	100
Stage I	5.0117 (0.5202)	3.6661 (0.2637)	11.85 (0.36)	0
Central nuclear	5.1986 (0.6253)	3.6782 (0.2625)	11.83 (0.40)	0
Central SCAD	3.7230 (0.4046)	3.6007 (0.2830)	7.86 (1.90)	3
Singular population covariance; $d = 40$, $r_X = 3$				
Proposed	1.8173 (0.2133)	1.2996 (0.0624)	3.00 (0.00)	100
Stage I	2.9875 (0.3031)	1.4319 (0.0753)	11.51 (0.75)	0
Central nuclear	3.0154 (0.3252)	1.4374 (0.0769)	11.58 (0.67)	0
Central SCAD	1.8709 (0.2137)	1.3133 (0.0630)	4.00 (1.02)	40
Smaller positive eigenvalues; $d = 40$, $r_X = 3$				
Proposed	1.7250 (0.1835)	1.2682 (0.0413)	3.00 (0.00)	100
Stage I	2.6346 (0.2809)	1.3671 (0.0516)	10.50 (1.05)	0
Central nuclear	2.6643 (0.3068)	1.3721 (0.0525)	10.53 (1.10)	0
Central SCAD	1.8336 (0.2074)	1.3073 (0.0554)	3.34 (0.62)	74
Smaller signal; $d = 40$, $r_X = 3$				
Proposed	1.7442 (0.2043)	1.2408 (0.0406)	3.00 (0.00)	100
Stage I	2.3717 (0.2691)	1.3098 (0.0483)	10.04 (1.19)	0
Central nuclear	2.3929 (0.2653)	1.3119 (0.0486)	10.22 (1.04)	0
Central SCAD	2.0211 (0.2569)	1.3025 (0.0610)	6.24 (1.70)	6

Table 10 shows lower mean prediction loss and test MSE for the proposed method than for all three comparators in each setting. Relative to Stage I, the paired mean prediction-loss reductions are 1.8237, 1.1702, 0.9096, and 0.6275, with Monte Carlo standard errors 0.0481, 0.0221, 0.0253, and 0.0199, respectively. Both prediction loss and test MSE decrease in all 100 pairs per setting. The second stage selects the identifiable rank three in every replication, including the settings with smaller positive eigenvalues and smaller signals.

The 8,000 recorded Stage II updates remain in the observed predictor space: the largest value of $\|(I_p - \Pi_X)C_t\|_F$ is 7.08×10^{-13}. This agrees with the subspace identity in Lemma 3. The rank frequencies in Table 10 concern the returned estimate. They do not assert rank retention at every preceding update. Validation selects the penalties for prediction; these empirical results do not verify the score condition in Theorem 1 or the sufficient inequalities in Corollary 34.

S40.9 Implementation checks

The code checks $H^{-1/2}H^{1/2}A = A$, an upper spectral bound for M, and the transformed update identity. It compares the scalar SCAD and MCP proximal maps with dense one-dimensional grid searches at fixed test values and checks each recorded Stage II objective sequence within numerical tolerance. Cached worker covariances and cross-products accelerate local loss and gradient evaluation in the experiments; they are not transmitted to the master by the proposed protocol. The thin singular value decomposition implementation applies $H^{1/2}$ and $H^{-1/2}$ without constructing a dense $p \times p$ master matrix when $n_1 < p$.

Supplementary references

Rina Foygel Barber and Wooseok Ha. Gradient descent with non-convex constraints: Local concavity determines convergence. *Information and Inference: A Journal of the IMA*, 7(4):755–806, 2018. doi: 10.1093/imaiai/iay002.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. doi: 10.1093/acprof:oso/9780199535255.001.0001.