Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Simultaneous Norm Optimality for Quadratic Regularization under Gain and Conditioning Constraints

Bich Van Nguyen

nbvan@vnu.edu.vn

VNU University of Engineering and Technology

Ngoc Anh Khong

VNU University of Engineering and Technology

Research Article

Keywords:

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11445957/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Simultaneous Norm Optimality for Quadratic Regularization under Gain and Conditioning Constraints

Nguyen Bich Van[1*] and Khong Ngoc Anh[1]

[1]Institute for Artificial Intelligence, VNU University of Engineering and Technology, Hanoi, Vietnam.

*Corresponding author(s). E-mail(s): nbvan@vnu.edu.vn;
Contributing authors: 22022549@vnu.edu.vn;

Abstract

Quadratic regularization stabilizes least-squares problems, but bounds on conditioning and coefficient amplification also constrain the attainable fit. For a fixed nonzero real design matrix, we construct an explicit penalty that minimizes every unitarily invariant norm of the fitted-target distortion under a positive penalty floor, a Hessian condition-number ceiling, and a target-to-coefficient amplification ceiling. The optimization is over all admissible symmetric penalties, including noncommuting ones. A sharp inverse-Hessian diagonal bound, attained simultaneously in all singular directions, and weak majorization establish the common optimum. We characterize the fitted-target and coefficient maps of all Frobenius minimizers and obtain uniqueness of the penalty at full column rank. For an individual target, a noncommuting penalty can improve upon this norm-optimal rule. We prove an exact convex reduction of that problem to the positive singular subspace, with a lift preserving the objective and all constraints. The explicit solution is evaluated by a singular value decomposition and scalar maxima; its relation to modified Tikhonov flooring is made precise. Independent conic computations, prescribed-spectrum checks, and comparisons with ridge and modified Tikhonov regularization examine the formulas and their numerical implications. Six fixed-feature learning benchmarks illustrate reduced training-fit distortion, while validation-selected predictive outcomes remain mixed.

Keywords: Quadratic regularization, numerical linear algebra, unitarily invariant norms, weak majorization, condition-number constraints, rank reduction

1 Introduction

Quadratic regularization is a standard tool in numerical linear algebra for stabilizing least-squares problems and controlling the sensitivity of their solutions. Its numerical effect depends on both the spectrum of the regularized Hessian and the amplification of the target-to-coefficient map. Ridge regression uses a scalar multiple of the identity as the penalty, whereas modified Tikhonov methods adjust different singular directions separately [1–3]. This raises a matrix approximation question: under simultaneous bounds on conditioning and coefficient amplification, which admissible quadratic penalty changes the least-squares fitted-target map the least?

We study this question for a fixed design matrix and a positive lower bound on the penalty. The admissible penalties are arbitrary real symmetric matrices satisfying two further conditions: an upper bound on the condition number of the regularized Hessian and an upper bound on the norm of the target-to-coefficient map. In particular, a competing penalty need not commute with the Gram matrix. The objective is the difference between the regularized fitted-target map and the orthogonal projection onto the range of the design.

Our main contribution is an explicit solution theory for this joint feasible class, covering simultaneous norm optimality, identification of all Frobenius-optimal estimators, and exact target-dependent dimension reduction. It consists of the following three results.

(i) *One optimizer for an entire family of norms.* We construct a penalty Q_*, determined only by the design and the three budgets, that minimizes every unitarily invariant fitted-target distortion norm over all admissible symmetric penalties (Theorem 3.4). A sharp bound on each inverse-Hessian diagonal entry is attained simultaneously; weak majorization then proves the common optimality. Thus the choice of norm does not require a new optimizer, and commutation with the

Gram matrix is a property of the constructed solution, not a restriction on its competitors.

(ii) *Identification through equality.* Every Frobenius minimizer induces the same fitted-target map and the same coefficient map. The equality argument excludes coupling between the positive singular subspace and the null coefficient space. In particular, the penalty itself is unique at full column rank (Proposition 3.5). This identifies the estimators associated with all minimizing penalties, beyond determining the optimal objective value.

(iii) *Exact reduction for individual targets.* The common norm optimizer need not be optimal for a specified target. We give a noncommuting improvement and prove that the target-dependent problem reduces exactly to the positive singular subspace. Every reduced optimizer has a feasible lift preserving the objective and all three constraints (Theorem 3.7); no commutation assumption is imposed on its active block.

The operator-norm specialization gives three necessary and sufficient fit-feasibility conditions (Theorem 3.1 and Corollary 3.2). Diagonal target second moments yield the average-fit consequence in Corollary 3.6. These are consequences of the same attainable inverse bound.

The distinction from earlier work concerns these explicit solutions and structural properties. The fixed-target and squared Frobenius problems fit the general ConMOM model of Li et al. [4], and the Frobenius objective can also be represented in the moment-fitting code accompanying Livne et al. [5, 6]. We make these connections explicit in Section 2 and after Theorem 3.7. Our contributions are the common closed-form optimizer, its equality characterization, and the exact reduction and lift for the specified joint constraints.

One application is output-layer training in an extreme learning machine, where the nonlinear hidden features are fixed before solving a least-squares problem [7]. In this setting, sensitivity of the inverse map and error in its numerical evaluation are separate questions [8, 9]. Our results apply conditionally on the realized features. Numerical comparisons with ridge and modified Tikhonov regularization illustrate the training-fit conclusions; they do not give a generalization guarantee.

This article extends the ICICPE 2026 report [10], which concerns the unregularized map $H^{+}T$, its perturbation bound, and pseudoinverse solvers. The joint constrained optimization problem and the results on simultaneous norm optimality, equality cases, second moments, and target-dependent rank reduction were not contained in that report. The extreme learning machine (ELM) setting, compact SVD, and inverse-iteration family are common background. The numerical records below are separate from the conference success-count table and heatmap, neither of which is reproduced or pooled with the present experiments.

The paper is organized as follows. Section 2 reviews the relevant literature and fixes notation. Section 3 contains the main results and their proofs. Section 4 gives numerical examples and benchmark comparisons. We conclude with the scope of the results and some further questions. The appendices contain scalar feasibility conditions, inverse-iteration identities, residual estimates, and details of the numerical protocols.

2 Preliminaries and related work

2.1 Relation to earlier optimization problems

Spectral regularization and parameter selection are established topics. Ridge regression replaces reciprocal singular values by bounded filters [1]. OCReP chooses a ridge parameter by conditioning the regularized pseudoinverse [11], whereas our condition bound applies to the Hessian. Numerical-rank effects in pseudoinverse learning have

been studied directly [12]. A scalar diagonal shift enforcing a condition ceiling is standard reconditioning algebra [13]; neither that shift nor a new ridge estimator is claimed here.

Modified Tikhonov regularization is a close antecedent. Fuhry and Reichel [2, Theorem 2.1] give a matrix-nearness characterization of a spectral floor, and Noschese and Reichel [3] develop related condition and nearness constructions. These results already allow general symmetric perturbations and yield commuting solutions. Won et al. [14] study condition-constrained covariance estimation and use a scalar spectral sandwich. Tanaka and Nakata [15, Theorems 1–2] reduce condition-constrained invariant-norm matrix nearness to spectral optimization, using standard pinching. Their objective measures distance to a prescribed matrix; ours measures the design-weighted inverse map $P-\Phi(A+Q)^{-1}\Phi^T$ under simultaneous floor, gain, and condition constraints. For the latter objective, we obtain a common optimizer from a bound on the inverse Hessian that is attained in every singular direction.

Li et al. [4, problem (1.2)] consider a more general condition-number-constrained least-squares matrix problem, with an arbitrary linear operator and an additional closed convex constraint set. After an inverse-Hessian change of variables, both our fixed-target problem and our squared Frobenius distortion problem are special cases of that model; the identification is given after Theorem 3.7. Their Lemma 2.2 gives spectral clipping for the nearest-matrix subproblem, and their Theorem 3.1 establishes convergence of a partial inexact alternating direction method. The present work gives an explicit common optimizer across unitarily invariant distortion norms for a particular joint constraint set, characterizes its Frobenius equality case, and proves an exact compression and lift for the fixed-target problem.

Livne et al. [5] study conditioned positive definite estimation by regularized Cholesky optimization (RCO). Their publicly available CPD implementation [6]

includes least-squares moment fitting, numerical continuation, and Newton–CG optimization. There is a direct overlap at the objective level: with the variable $X = (A+Q)^{-1}$, the squared Frobenius distortion $\left\|P - \Phi X \Phi^T\right\|_{\mathrm{F}}^2$ is a single-term moment-fitting objective, apart from a constant normalization. This objective alone therefore does not distinguish our problem. In the CPD implementation of RCO, the default regularizer for an $m \times m$ positive definite matrix M is

$$\psi(M) = \log \operatorname{tr}(M) - m^{-1} \log \det(M).$$

Continuation selects a trade-off using the residual and the measured condition number. This scale-invariant regularizer does not by itself impose our inverse-floor and gain constraints. CPD also includes a separate Tanaka–Nakata routine for condition-bounded matrix nearness; that routine must be distinguished from RCO. Our results concern the explicit simultaneous norm optimum and the equality and lifting properties for the particular joint feasible set, rather than a new general framework for conditioned positive definite estimation.

Earlier biased-estimation theory also optimizes variance subject to bias-gradient bounds, with closed forms in commuting cases and semidefinite formulations more generally [16]. Our amplification ceiling bounds a worst-case operator norm, not total variance; the positive penalty floor and Hessian condition ceiling are additional constraints. The target-specific formulation below uses standard inverse-order, spectral-sandwich, and norm-LMI tools [14, 17]. Its particular exact rank reduction is the useful structural result.

These relations locate the present results within existing matrix optimization. Model inclusion and objective overlap do not by themselves supply the common optimizer in Theorem 3.4, the coefficient-map equality characterization in Proposition 3.5, or the constraint-preserving lift in Theorem 3.7. We establish these properties directly

for the joint feasible class below, using standard spectral and convex-analytic tools. Approximation and generalization for random features form a separate subject; see, for example, [18].

2.2 Notation and the ridge objective

Throughout the paper, all matrices are real. We use $\|\cdot\|_2$ and $\|\cdot\|_{\mathrm{F}}$ for the operator and Frobenius norms, respectively, and $\preceq$ for the Loewner order on symmetric matrices. The symbol I_m denotes the identity matrix of order m. Let $H \in \mathbb{R}^{N\times L}$ be nonzero and let $T \in \mathbb{R}^{N\times c}$ contain training targets. Write $s_H = \|H\|_{\mathrm{F}}$ and $\Phi = H/s_H$; all held-out feature rows use the same training-derived scale. Set $P = \Phi\Phi^{+}$. For $\lambda > 0$, define

$$J_\lambda(B) = \tfrac{1}{2}\,\|\Phi B - T\|_{\mathrm{F}}^2 + \tfrac{\lambda}{2}\,\|B\|_{\mathrm{F}}^2\,, \tag{1}$$

$$G_\lambda = \Phi^T\Phi + \lambda I_L, \quad C = \Phi^T T, \quad B_\lambda = G_\lambda^{-1}C = R_\lambda T, \quad R_\lambda = G_\lambda^{-1}\Phi^T. \tag{2}$$

The original-coordinate coefficients are $\beta_\lambda = B_\lambda/s_H$ and the corresponding ridge parameter for H is $s_H^2\lambda$. Thus predictions are unchanged by this bookkeeping normalization. A gain bound for B must be divided by s_H to obtain a bound for β; coefficient gains should not be compared across unreported feature units. In particular, $\kappa_2(cH) = \kappa_2(H)$ whereas $\|(cH)^{+}\|_2 = \|H^{+}\|_2/|c|$ for $c \neq 0$: conditioning alone does not determine absolute target amplification. For inconsistent targets, the unregularized fitted value is $HH^{+}T = P_H T$, the orthogonal projection onto range(H), rather than T itself.

Let $\Phi = U_r\Sigma_r V_r^T$ be the compact singular value decomposition (SVD), with $r = \mathrm{rank}(\Phi) \geq 1$ and $\sigma_1 \geq \cdots \geq \sigma_r > 0$. Define

$$a = \lambda_{\min}(\Phi^T\Phi), \qquad b = \sigma_1^2, \qquad a_+ = \sigma_r^2. \tag{3}$$

If $r < L$, including a wide full-row-rank matrix, then $a = 0$ whereas $a_+ > 0$. The Hessian sees null coefficient directions; the target-to-coefficient operator does not. Classical diagonalization gives

$$\kappa_2(G_\lambda) = \frac{b+\lambda}{a+\lambda}, \quad R_\lambda = V_r \operatorname{diag}\left(\frac{\sigma_i}{\sigma_i^2+\lambda}\right) U_r^T, \quad g_\lambda = \|R_\lambda\|_2 = \max_i \frac{\sigma_i}{\sigma_i^2+\lambda}. \tag{4}$$

At fixed Φ and λ, the sharp bound

$$\|B_\lambda(T+\Delta T) - B_\lambda(T)\|_{\mathrm{F}} \le g_\lambda \, \|\Delta T\|_{\mathrm{F}} \tag{5}$$

holds, with equality for a target perturbation in a maximizing left-singular direction. This is not a result about redrawing features, reselecting λ, or the correctness of a classification label. Computational error $\widehat{B} - B_\lambda$ is also distinct from regularization bias relative to $\Phi^+ T$.

3 Optimal quadratic regularization under joint constraints

In this section, we determine an optimal quadratic penalty under the three prescribed constraints. We first treat the operator norm, then prove simultaneous optimality for all unitarily invariant norms and characterize the Frobenius equality case. We then treat the error for a fixed target and describe how to evaluate the resulting spectral rule.

For a fixed $\lambda_0 > 0$, let $Q = Q^T \succeq \lambda_0 I_L$ and set

$$A = \Phi^T \Phi, \quad G_Q = A + Q, \quad R_Q = G_Q^{-1} \Phi^T, \quad S_Q = \Phi R_Q. \tag{6}$$

The objective is $J_Q(B) = \frac{1}{2}\|\Phi B - T\|_{\mathrm{F}}^2 + \frac{1}{2}\operatorname{tr}(B^T Q B)$ and its minimizer is $B_Q = R_Q T$. Fix $K_* > 1$ and $\Gamma > 0$, and define

$$\mathcal{Q} = \{Q = Q^T \succeq \lambda_0 I_L : \kappa_2(G_Q) \le K_*,\ \|R_Q\|_2 \le \Gamma\}. \tag{7}$$

When a horizon is imposed, use the K_* of Lemma A.1 and initialize the inverse iteration with $I_L/\lambda_{\max}(G_Q)$. The same eigenvalue argument makes this horizon exactly equivalent to the corresponding condition ceiling.

Complete V_r to an orthogonal matrix V and extend the singular values by $s_i = \sigma_i$ for $i \le r$ and $s_i = 0$ for $i > r$. Define

$$c_i = \max\{s_i^2 + \lambda_0,\ s_i/\Gamma\}, \qquad C_* = \max_i c_i = \max\{b + \lambda_0, \sqrt{b}/\Gamma\}, \tag{8}$$

$$d_i^* = \max\{c_i, C_*/K_*\}, \qquad G_* = V \operatorname{diag}(d_i^*) V^T, \quad Q_* = G_* - A. \tag{9}$$

The matrices in (9) are independent of the choice of basis in each repeated-singular-value subspace, since the corresponding values d_i^* coincide. The same observation applies to the null space.

We begin with the operator-norm optimum.

Theorem 3.1 (Sharp uniform fit boundary over all quadratic penalties) *The matrix Q_* belongs to $\mathcal{Q}$ and*

$$\min_{Q \in \mathcal{Q}} \|P - S_Q\|_2 = \Theta_* := \max_{1 \le i \le r} \left(1 - \frac{\sigma_i^2}{d_i^*}\right) \tag{10}$$

$$= \max\left\{\frac{\lambda_0}{a_+ + \lambda_0},\ 1 - \Gamma\sqrt{a_+},\ 1 - \frac{K_* a_+}{C_*}\right\}. \tag{11}$$

Thus a penalty in $\mathcal{Q}$ with uniform fit distortion at most θ exists if and only if $\Theta_* \le \theta$. The lower bound holds without assuming that Q and A commute. The term "minimax" refers to worst-case unit-norm training-target distortion, not statistical risk or out-of-sample prediction.

Proof First we verify that Q_* is feasible. By construction, $d_i^* \geq s_i^2 + \lambda_0$ and $d_i^* \geq s_i/\Gamma$; also $C_*/K_* \leq d_i^* \leq C_*$ and $\max_i d_i^* = C_*$. Consequently $Q_* \succeq \lambda_0 I_L$, $\kappa_2(G_*) \leq K_*$, and $\left\|G_*^{-1}\Phi^T\right\|_2 = \max_{i\leq r} s_i/d_i^* \leq \Gamma$. Its distortion is the right-hand side of (10).

For the converse, take any $Q \in \mathcal{Q}$ and put $X = G_Q^{-1}$ and $x_i = v_i^T X v_i > 0$. Inversion reverses the positive-definite order, so

$$0 \prec X \preceq (A + \lambda_0 I_L)^{-1}, \qquad x_i \leq \frac{1}{s_i^2 + \lambda_0}. \tag{12}$$

For $i \leq r$, the gain condition gives $s_i x_i \leq s_i \|X v_i\|_2 = \left\|X\Phi^T u_i\right\|_2 \leq \Gamma$. For $i > r$ only (12) is needed. Hence $x_i \leq 1/c_i$ for every i. Choose j with $c_j = C_*$. The Rayleigh principle and the condition constraint give

$$\lambda_{\min}(X) \leq x_j \leq C_*^{-1}, \qquad \lambda_{\max}(X) \leq K_* \lambda_{\min}(X) \leq K_*/C_*.$$

Combining these inequalities yields $x_i \leq \min\{1/c_i, K_*/C_*\} = 1/d_i^*$. Furthermore, (12) implies $0 \preceq S_Q \preceq P$. Therefore, for every positive singular direction,

$$\left\|P - S_Q\right\|_2 \geq u_i^T (P - S_Q) u_i = 1 - s_i^2 x_i \geq 1 - s_i^2/d_i^*.$$

Taking the maximum over i proves (10). To obtain the second expression, observe that

$$1 - \frac{s^2}{\max\{s^2 + \lambda_0, s/\Gamma, C_*/K_*\}} = \max\left\{\frac{\lambda_0}{s^2 + \lambda_0}, 1 - \Gamma s, 1 - \frac{K_* s^2}{C_*}\right\}.$$

Each of the three functions decreases with $s > 0$; their maximum over the positive spectrum is attained at $s = \sigma_r$. This proves (11) and the feasibility equivalence. □

Corollary 3.2 (Three necessary and sufficient budget tests) *For $0 < \theta < 1$, uniform feasibility over $\mathcal{Q}$ is equivalent to*

$$\lambda_0 \leq \frac{\theta a_+}{1 - \theta}, \qquad \Gamma \geq \frac{1 - \theta}{\sqrt{a_+}}, \qquad K_* \geq \frac{(1 - \theta) C_*}{a_+}. \tag{13}$$

Proof By Theorem 3.1, feasibility is equivalent to requiring each term in (11) to be at most θ. Rearranging these three inequalities gives (13), since all denominators are positive. □

The gain obstruction persists even if the condition ceiling is relaxed. The condition obstruction uses the smallest *positive* squared singular value, while null coefficient directions are regularized separately. This explains why the scalar-ridge obstruction need not be fundamental.

Proposition 3.3 (Least spectral Hessian and comparison with ridge) *Among the members of $\mathcal{Q}$ diagonal in the chosen singular basis, d^* is the componentwise least Hessian spectrum. For every such member with eigenvalues d_i and every target matrix T,*

$$\|(P - S_*)T\|_{\mathrm{F}} \le \left\|(P - S_Q)T\right\|_{\mathrm{F}}, \tag{14}$$

$$\left\|B_* - \Phi^+ T\right\|_{\mathrm{F}} \le \left\|B_Q - \Phi^+ T\right\|_{\mathrm{F}}. \tag{15}$$

In particular these inequalities hold against every feasible scalar ridge parameter $\lambda \ge \lambda_{\min}$, where the explicit floor is given by (31) in Appendix A. A strict improvement occurs if some $i \le r$ has $d_i > d_i^*$ and $(U_r^T T)_{i,:} \neq 0$.

Proof Feasibility gives $d_i \ge c_i$, whence $\max_j d_j \ge C_*$. The condition constraint implies $d_i \ge \max_j d_j / K_* \ge C_*/K_*$, so $d_i \ge d_i^*$. The squared quantities in (14) and (15) are, respectively,

$$\sum_{i \le r} \left(1 - \frac{s_i^2}{d_i}\right)^2 \left\|(U_r^T T)_{i,:}\right\|_2^2, \qquad \sum_{i \le r} \left(\frac{1}{s_i} - \frac{s_i}{d_i}\right)^2 \left\|(U_r^T T)_{i,:}\right\|_2^2.$$

Both scalar factors are positive and strictly increasing in $d_i \ge s_i^2 + \lambda_0$. This proves the inequalities and the stated sufficient condition for strictness. A feasible scalar ridge is a member of this class with $d_i = s_i^2 + \lambda$. □

A strict separation example. Let $\Phi = \mathrm{diag}(\sqrt{0.9}, \sqrt{0.1})$, $K_* = 2$, $\Gamma = 10$, and $\lambda_0 = 10^{-3}$. Then $C_* = 0.901$, $d^* = (0.901, 0.4505)$, and $\Theta_* = 1 - 0.1/0.4505 \approx 0.7780244$. Every feasible scalar ridge has $\lambda \ge 0.7$ and uniform distortion at least $0.7/(0.1+0.7) = 0.875$. Thus $\theta = 0.8$ is feasible for matrix regularization and infeasible

for scalar ridge. This is a training-fit comparison under fixed numerical budgets; it implies no universal test-risk advantage.

3.1 Simultaneous norm optimality and equality cases

We now use the inverse bound from the preceding proof to compare the whole distortion spectrum. Put

$$e_i^* = 1 - \sigma_i^2/d_i^*, \quad i = 1, \dots, r, \qquad E_Q = P - S_Q.$$

We write e_i for the ith standard coordinate vector (distinct from the scalar e_i^*). For vectors sorted in decreasing order, $x \prec_w y$ means that every partial sum of x is bounded by the corresponding partial sum of y. This is weak majorization, not coordinatewise eigenvalue ordering or a Loewner inequality.

Theorem 3.4 (A common optimizer for every unitarily invariant norm) *Let e^* be the decreasing rearrangement of $(e_1^*, \dots, e_r^*)$ padded with $N - r$ zeros. For every $Q \in \mathcal{Q}$,*

$$e^* \prec_w \lambda(E_Q). \tag{16}$$

Consequently the same Q_* minimizes $\|P - S_Q\|_{\mathrm{ui}}$ over $\mathcal{Q}$ for every unitarily invariant norm on $\mathbb{R}^{N\times N}$. In particular,

$$\min_{Q\in\mathcal{Q}} \|P - S_Q\|_{S_p} = \left(\sum_{i=1}^{r} (e_i^*)^p\right)^{1/p}, \quad 1 \le p < \infty, \tag{17}$$

$$\min_{Q\in\mathcal{Q}} \|P - S_Q\|_{(k)} = \sum_{j=1}^{\min(k,r)} e_{[j]}^*, \quad 1 \le k \le N, \tag{18}$$

where S_p denotes the Schatten norm, (k) the Ky Fan norm, and $e_{[j]}^*$ the decreasing rearrangement. The case $p = \infty$ is Theorem 3.1.

Proof The proof of Theorem 3.1 gives $x_i = v_i^T G_Q^{-1} v_i \le 1/d_i^*$ for every feasible penalty. Moreover $E_Q \succeq 0$ and $E_Q = P E_Q P$. Its compression $F = U_r^T E_Q U_r$ has diagonal entries

$f_i = 1 - \sigma_i^2 x_i \geq e_i^* > 0$. For $k \leq r$, let I_k index the k largest entries of e^*. The Ky Fan variational principle gives

$$\sum_{j=1}^{k} \lambda_j(F) = \max_{W^T W = I_k} \operatorname{tr}(W^T F W) \geq \sum_{i \in I_k} f_i \geq \sum_{j=1}^{k} e^*_{[j]}.$$

For $k > r$, zero padding gives the same inequality. Since E_Q is positive semidefinite, its eigenvalues are its singular values. The Ky Fan dominance theorem therefore yields the inequality for every unitarily invariant norm [19]. At Q_*, $F = \operatorname{diag}(e_i^*)$, so all inequalities attain equality simultaneously. The stated specializations follow from the definitions of the norms. □

The finite-p and Ky Fan optima depend in general on the whole positive spectrum. The operator-norm optimum is the special case that reduces to the two endpoints in (11).

The next proposition describes all Frobenius minimizers.

Proposition 3.5 (Frobenius equality and identification of the predictor) *If $Q \in \mathcal{Q}$ minimizes $\|P - S_Q\|_{\mathrm{F}}$, then $S_Q = S_*$ and $R_Q = R_*$. Every such Q commutes with A. If $r = L$, the unique minimizing penalty is Q_*. If $r < L$, only the penalty on the null coefficient space can remain undetermined by this objective. No corresponding uniqueness claim is made for the operator norm alone.*

Proof Let Q be a Frobenius minimizer and let $F = U_r^T E_Q U_r$ as in the preceding proof. Then

$$\|F\|_{\mathrm{F}}^2 = \sum_i f_i^2 + \sum_{i \neq j} F_{ij}^2 \geq \sum_i (e_i^*)^2.$$

Equality forces $F = \operatorname{diag}(e_i^*)$ and $x_i = 1/d_i^*$ for every $i \leq r$. Set $X = G_Q^{-1}$. At least one defining bound is active for each d_i^*. If $d_i^* = s_i^2 + \lambda_0$, then the positive semidefinite matrix $(A + \lambda_0 I)^{-1} - X$ has zero quadratic form on v_i, so $X v_i = x_i v_i$. If $d_i^* = s_i/\Gamma$, the inequalities $x_i \leq \|X v_i\|_2 \leq \Gamma/s_i = x_i$ give the same conclusion. If $d_i^* = C_*/K_*$, use $X \preceq (K_*/C_*) I$ and the identical zero-quadratic-form argument. Thus $X v_i = v_i/d_i^*$ for every positive direction.

Symmetry eliminates coupling to the null space, and

$$X\Phi^T = V_r \operatorname{diag}(s_i/d_i^*)U_r^T = R_*.$$

The positive block of X is fixed and diagonal, while its null block commutes with the zero block of A. Inverting preserves this block structure. If there is no null block, $X = G_*^{-1}$ and $Q = Q_*$. □

Strict convexity of the squared Frobenius norm already gives uniqueness of the fitted map S_Q over the convex inverse feasible set. Proposition 3.5 additionally identifies the coefficient map R_Q: the active inverse bounds exclude coupling between the positive singular subspace and the null coefficient space.

Corollary 3.6 (Average fit for diagonal target second moments) *Fix Φ and the budgets. Suppose T is random with finite second moment and*

$$U_r^T \mathbb{E}[TT^T]U_r = \operatorname{diag}(w_1, \ldots, w_r), \qquad w_i \geq 0.$$

Among penalties selected before observing T,

$$\min_{Q \in \mathcal{Q}} \mathbb{E}\left\|(P - S_Q)T\right\|_{\mathrm{F}}^2 = \sum_i w_i (e_i^*)^2, \tag{19}$$

attained by Q_*. Isotropic second moments give a multiple of the squared Frobenius optimum.

Proof Since $E_Q = U_r F U_r^T$, $\mathbb{E}\left\|E_Q T\right\|_{\mathrm{F}}^2 = \sum_i w_i \left\|Fe_i\right\|_2^2 \geq \sum_i w_i f_i^2 \geq \sum_i w_i (e_i^*)^2$. The diagonal F at Q_* attains the bound. □

The second moment $\mathbb{E}[TT^T]$ need not be a covariance matrix. The requirement that Q be chosen before observing T is essential to the stated optimization problem. The corollary concerns average training fit and does not assert out-of-sample risk optimality. For off-diagonal second moments, the optimizer may differ, as the following example shows.

3.2 Target-specific optimization without a commutation assumption

We first show that the commutation restriction in Proposition 3.3 cannot be omitted. Retain the parameters of the preceding two-dimensional example and set $T = (1, 1)^T$. Consider

$$X = \begin{pmatrix} 1.09 & 0.25 \\ 0.25 & 1.90 \end{pmatrix}, \qquad Q = X^{-1} - A.$$

This is feasible: $Q \succeq 0.001 I_2$, $\kappa_2(X) \approx 1.93410 < 2$, and $\left\|X\Phi^T\right\|_2 \approx 1.08945 < 10$. The first condition also follows directly from $X \preceq \operatorname{diag}(1/0.901, 1/0.101)$: both diagonal entries of the difference are positive and its determinant is positive. Nevertheless, $\left\|(I - \Phi X \Phi^T)T\right\|_2^2 = 0.543361$, below the value 0.6053232 for Q_*. The uniform minimax theorem and this target-specific improvement are compatible.

Theorem 3.7 (An exact rank-reduced convex formulation) *Write $Z = U_r^T T$ and $\Sigma = \Sigma_r$. If $Z \neq 0$, the minimum of $\left\|(P - S_Q)T\right\|_{\mathrm{F}}^2$ over $Q \in \mathcal{Q}$ equals the minimum of*

$$\begin{aligned} \min_{X_r = X_r^T,\ t \geq 0} \quad & \|(I_r - \Sigma X_r \Sigma) Z\|_{\mathrm{F}}^2 \\ \text{subject to} \quad & X_r \preceq (\Sigma^2 + \lambda_0 I_r)^{-1}, \quad t I_r \preceq X_r \preceq K_* t I_r, \\ & \begin{pmatrix} \Gamma I_r & X_r \Sigma \\ \Sigma X_r & \Gamma I_r \end{pmatrix} \succeq 0. \end{aligned} \tag{20}$$

Every optimizer has $X_r \succ 0$ and $t > 0$. Completing V_r by V_0, the lift

$$X = V_r X_r V_r^T + t V_0 V_0^T, \qquad Q = X^{-1} - A \tag{21}$$

is an optimal feasible penalty. The null term is omitted when $r = L$. If $Z = 0$, all feasible penalties have zero loss and Q_* can be used.

Proof We prove the equality of the optimal values by compression and lifting. If $X = G_Q^{-1} \succ 0$, the full-space constraints are equivalent to

$$X \preceq (A + \lambda_0 I_L)^{-1}, \quad tI_L \preceq X \preceq K_* tI_L, \quad \left\| X\Phi^T \right\|_2 \leq \Gamma$$

for some $t > 0$. Inverse order gives the first equivalence; the spectral sandwich gives the condition bound; the Schur complement represents the gain as a norm LMI. These are standard convex formulations [14, 17]; the full-space least-squares problem is a specialization of [4, problem (1.2)].

Compress X to $X_r = V_r^T X V_r$. Principal compression preserves the order and sandwich constraints, and $\|X_r\Sigma\|_2 \leq \|XV_r\Sigma\|_2 = \left\|X\Phi^T\right\|_2$. The objective depends only on this compression and equals the displayed squared norm. Thus every full feasible penalty supplies a feasible reduced point with the same objective.

Conversely, at a reduced feasible point with $t > 0$, the lift (21) is positive definite and has spectrum in $[t, K_* t]$. The active floor holds by assumption. Also $t \leq e_1^T X_r e_1 \leq 1/(b + \lambda_0) < 1/\lambda_0$, so the lifted null block satisfies the floor constraint. The lifted gain equals $\|X_r\Sigma\|_2 \leq \Gamma$, and its objective is unchanged. Therefore this lift yields a full feasible penalty.

It remains to prove that an optimal reduced solution admits this lift. The reduced feasible set is nonempty, closed, and bounded, so the continuous convex objective attains a minimum. A singular feasible X_r forces $t = 0$ and $X_r = 0$. For small $\epsilon > 0$, $(X_r, t) = (\epsilon I_r, \epsilon)$ is feasible, and the derivative of the objective at zero is $-2\|\Sigma Z\|_{\mathrm{F}}^2 < 0$. Thus no optimizer is singular, proving attainability of the inverse lift and equality of the optimal values. □

The formulation has $r(r+1)/2 + 1$ scalar variables and a largest displayed LMI of order $2r$, instead of the full-space gain LMI of order $L + N$. Computing the SVD and Z still depends on N and L. The lift proves the existence of an optimizer without active–null coupling; it does not restrict the active block to commute with Σ^2, nor identify every full-space optimizer. The squared-norm objective has a standard semidefinite epigraph if a linear-objective SDP is required. This is an exact specialization and dimension reduction, not a new general technique for representing condition

constraints. Since the optimizing penalty depends on T, fixed-Q gain bounds do not automatically bound the effect of reoptimizing after target perturbation.

Relation to the ConMOM model. The relation to [4, problem (1.2)] can be made explicit. With $X = G_Q^{-1}$, define the linear map and the closed convex set

$$\mathcal{L}_T(X) = \Phi X \Phi^T T, \qquad C_T = PT, \tag{22}$$

$$\mathcal{D}_\Phi = \{X = X^T : 0 \preceq X \preceq (A + \lambda_0 I_L)^{-1},\ \left\|X\Phi^T\right\|_2 \leq \Gamma\}. \tag{23}$$

The fixed-target objective is $\|\mathcal{L}_T(X) - C_T\|_{\mathrm{F}}^2$. Thus, apart from the immaterial factor $1/2$, it has the form of their objective, with $\mathcal{D} = \mathcal{D}_\Phi$ and condition bound K_*. For squared Frobenius distortion, take instead $\mathcal{L}(X) = \Phi X \Phi^T$ and $C = P$. More generally, for a fixed target second moment $M = \mathbb{E}[TT^T] \succeq 0$, expected squared fit has this form with $\mathcal{L}_M(X) = \Phi X \Phi^T M^{1/2}$ and $C_M = PM^{1/2}$. Corollary 3.6 gives an explicit optimizer under its diagonal-moment hypothesis.

To state the correspondence at the boundary, we use the closed condition cone

$$\Omega_{K_*} = \{X \succeq 0 : \lambda_{\max}(X) \leq K_* \lambda_{\min}(X)\} = \{X : \exists t \geq 0,\ tI_L \preceq X \preceq K_* t I_L\}.$$

Its only singular member is zero; we do not assign a ratio condition number to that matrix. If $PT \neq 0$, the proof of Theorem 3.7 excludes zero from the optimum, so the closed formulation and the positive-definite inverse problem have the same optimal value. If $PT = 0$, both values are zero and Q_* is a feasible penalty. The Frobenius case follows by taking $T = I_N$, for which $PT = P \neq 0$.

This inclusion of models does not by itself supply a formula for the optimizer. In particular, [4, Lemma 2.1] compares eigenvector alignments for fixed spectra under unitarily invariant norms; it does not select one spectrum solving our design-weighted problem for every such norm. The additional assertions here are the attainable inverse

bound, the common optimizer and equality case, and the exact rank reduction with a feasible lift. Applying the iterative method of [4] would also require handling projection onto $\mathcal{D}_\Phi$ and verifying its algorithmic assumptions; we do not infer an inexpensive projection or a performance advantage from the model inclusion.

3.3 Evaluation and computational cost of the spectral solution

The optimizer can be applied without forming a dense inverse Hessian, a full orthogonal completion of V_r, or the $N \times N$ fitted-target map. Algorithm 1 evaluates the formulas of Theorems 3.1 and 3.4.

Algorithm 1. Jointly constrained spectral regularization

Input: nonzero $\Phi \in \mathbb{R}^{N\times L}$, targets $T \in \mathbb{R}^{N\times c}$, and budgets $\lambda_0 > 0$, $K_* > 1$, $\Gamma > 0$.

1. Compute $\Phi = U_r \Sigma_r V_r^T$, with positive singular values σ_i and $b = \sigma_1^2$.
2. Set $C_* = \max\{b + \lambda_0, \sqrt{b}/\Gamma\}$ and

$$d_i^* = \max\{\sigma_i^2 + \lambda_0, \sigma_i/\Gamma, C_*/K_*\}, \quad i = 1, \dots, r.$$

3. Compute $Z = U_r^T T$ and return $B_* = V_r \operatorname{diag}(\sigma_i/d_i^*)Z$. If needed, compute the fitted targets as $U_r \operatorname{diag}(\sigma_i^2/d_i^*)Z$.
4. Report the gain $\max_i \sigma_i/d_i^*$ and distortion $\max_i(1 - \sigma_i^2/d_i^*)$. For the full Hessian condition number, include the null-direction eigenvalue $d_0 = \max\{\lambda_0, C_*/K_*\}$ when $r < L$.

Optional penalty: with d_0 as above, $Q_* = d_0 I_L + V_r \operatorname{diag}(d_i^* - \sigma_i^2 - d_0)V_r^T$.

For a dense design, a standard economy-size SVD costs $O(NL\min\{N,L\})$ arithmetic operations. Given its positive singular factors, the budget-dependent scalar work is $O(r)$ and coefficient evaluation costs $O((N+L)rc)$; these costs exclude explicitly forming the optional dense penalty. The same factors and Z can be reused for different budgets and scalar offsets. Thus evaluating the common norm optimizer requires no iterative matrix optimization after the decomposition. This statement concerns the explicit norm optimizer, not the target-specific conic solve in Theorem 3.7. Computing small singular directions and deciding numerical rank remain separate floating-point issues; the exact-rank theorem does not certify a cutoff chosen in finite precision.

For the closed-form choice, one SVD and scalar maxima suffice. To allow validation beyond the least-distorting point, the path $G_*(\tau) = G_* + \tau I_L$, $\tau \geq 0$, preserves all numerical constraints: the condition number is nonincreasing, and each gain $s_i/(d_i^*+\tau)$ with $s_i > 0$ strictly decreases. The condition number is constant when G_* is a scalar multiple of I_L. Validation over a predeclared grid containing 0 is possible; it need not select $\tau = 0$. The exact distortion at each τ is $\max_{i\leq r}(1 - s_i^2/(d_i^* + \tau))$. Direct coefficients and positive-direction hyperpower filters are

$$B_* = V_r \operatorname{diag}\left(\frac{\sigma_i}{d_i^*}\right) U_r^T T, \qquad Y_k \Phi^T = V_r \operatorname{diag}\left(\frac{\sigma_i}{d_i^*}[1 - (1 - d_i^*/C_*)^{p^k}]\right) U_r^T.$$

The gain is at most Γ at every exact-arithmetic iterate. Stationarity bounds in Theorem B.2 extend by replacing $\lambda\widehat{B}$ with $Q_*\widehat{B}$ and $a + \lambda$ with $\min_i d_i^*$.

The theoretical proofs are independent of numerical rank decisions. In computation, singular values below a specified relative cutoff are treated explicitly as a diagnostic subspace, as detailed below. When $\lambda_0 \downarrow 0$ and the gain constraint is inactive, (9) reduces to the familiar floor $d_i = \max\{s_i^2, b/K_*\}$ [2, 3]. The joint-constraint theorem includes this established special case.

4 Numerical evaluation

4.1 Direct checks of the optimization statements

We first test the matrix problem independently of prediction. Three prescribed spectra are used: $(0.8, 0.19, 0.01)^{1/2}$ in three dimensions; $(0.7, 0.1, 0.1, 0.05, 0.05)^{1/2}$ in five dimensions; and the first spectrum with three null coefficient directions. For $\lambda_0 \in \{10^{-4}, 0.02, 0.2\}$, 25 logarithmically spaced values of $K_* \in [1.1, 1000]$, and 25 values of $\Gamma \in [0.2, 100]$, there are 5,625 budget combinations. All evaluated floors, gains, condition bounds, and the two forms of Θ_* agree within the recorded tolerances (relative tolerance 10^{-10} for the condition and gain ceilings, and absolute tolerance 10^{-12} for the identity). Figure 1 displays one slice and the scalar-ridge gap. A further 400 randomly generated feasible, generally noncommuting competitors satisfy every checked Ky Fan partial-sum inequality and the diagonal-second-moment inequality within 10^{-10}. These computations support consistency; the proofs establish the results.

We separately solve the reduced feasible problem using Clarabel and SCS through CVXPY. Each solver minimizes target loss, spectral distortion, and squared Frobenius distortion on four cases, giving 24 solves. The first two cases use the two-dimensional separation example with $Z = (1, 1)^T$ or $Z = I_2$. The gain-active case uses squared spectrum $(0.8, 0.19, 0.01)$, $Z = (1, 1, 1)^T$, $K_* = 20$, $\Gamma = 0.5$, and $\lambda_0 = 0.001$. The wide case has $r = 3$, $L = 7$, the same spectrum, $Z_{1,:} = (1, 0)$, $Z_{2,:} = (0.5, 1)$, $Z_{3,:} = (1, 0.3)$, $K_* = 4$, $\Gamma = 3$, and $\lambda_0 = 0.01$. A separate full-space solve checks equality of the wide-case target value with the reduced formulation.

Table 1 compares the closed-form rule with the independent conic solutions.

The two solvers agree with the spectral and Frobenius formulas to the displayed precision. For the separation target, the numerical target optimum is about 0.525083, below 0.605323 for Q_*; for the wide target, the values are about 0.927552 and 0.989919. The wide full-space solve gives 0.9275516381. The most negative evaluated constraint

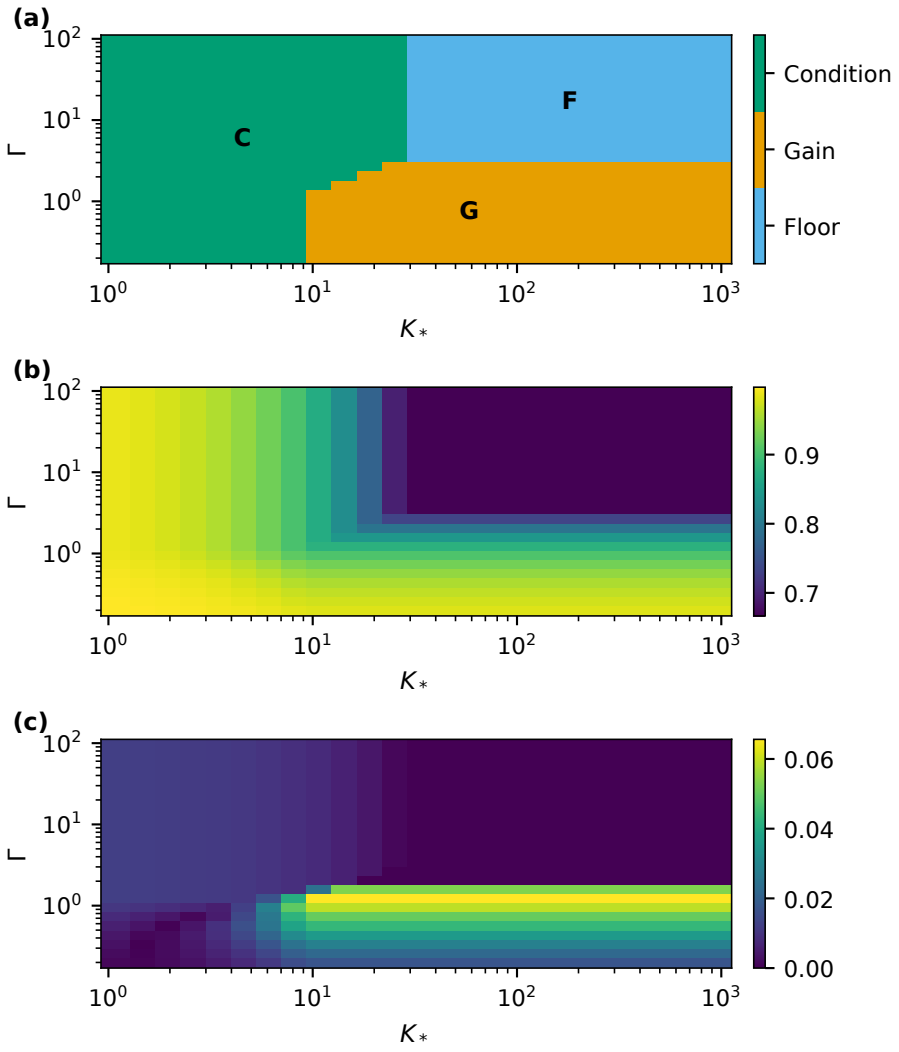


Fig. 1 Budget sweep for squared spectrum $(0.8, 0.19, 0.01)$ and $\lambda_0 = 0.02$. (a) Term attaining the sharp maximum (ties assigned in formula order): F, floor; G, gain; C, condition. (b) Optimal uniform distortion. (c) Least-feasible ridge distortion minus the optimum

slack over the 24 reduced solves is approximately -5.37×10^{-9} and the largest available SCS gap is 1.31×10^{-9}. These are ordinary solver diagnostics, not rigorous optimality enclosures. The tests do not establish scalability for large rank.

Table 1 Independent conic solves of the reduced feasible set. Target and Frobenius objectives are squared losses; spectral is the operator norm. Numerical agreement is a consistency check, not an interval certificate.

Case	Objective	Closed-form rule	Clarabel	SCS
separation-target	target	0.605323	0.525083	0.525083
separation-target	spectral	0.778024	0.778024	0.778024
separation-target	frobenius	0.605323	0.605323	0.605323
separation-isotropic	target	0.605323	0.605323	0.605323
separation-isotropic	spectral	0.778024	0.778024	0.778024
separation-isotropic	frobenius	0.605323	0.605323	0.605323
gain-target	target	1.819683	1.819683	1.819683
gain-target	spectral	0.950000	0.950000	0.950000
gain-target	frobenius	1.819683	1.819683	1.819683
wide-target	target	0.989919	0.927552	0.927552
wide-target	spectral	0.950617	0.950617	0.950617
wide-target	frobenius	0.907636	0.907636	0.907636

4.2 A calibrated modified Tikhonov comparator

To compare with the closest spectral-floor construction [2, 3], apply modified Tikhonov flooring to $A+\lambda_0 I$ and calibrate its threshold to the same budgets. Put $b_0 = b+\lambda_0$ and

$$\mu_0^2 = \max\left\{ b_0/K_*, \max_{i:\ s_i/\Gamma > s_i^2+\lambda_0} s_i/\Gamma \right\}, \qquad d_i^{\mathrm{MT}} = \max\{s_i^2+\lambda_0, \mu_0^2\}, \tag{24}$$

where an empty inner maximum is zero. This threshold is sufficient for conditioning because all eigenvalues lie between μ_0^2 and $\max(b_0, \mu_0^2)$ whenever the floor acts; if it does not act, their minimum already exceeds b_0/K_*. It also lifts every direction that would violate the gain ceiling. Thus it is a feasible spectral comparator and Proposition 3.3 applies. This is a budget-calibrated adaptation of the published filter, not a replication of its original discrepancy-based parameter selection. It isolates the cost of applying one common spectral floor when gain constraints are direction dependent.

Table 2 Benchmark sizes before partitioning and candidate hidden widths. No sample cap is applied to these bundled data sets.

Data set	Observations	Inputs	Task	Widths
Iris	150	4	Classification	64, 128
Wine	178	13	Classification	64, 128
Breast Cancer	569	30	Classification	64, 128
Digits	1,797	64	Classification	64, 128
Diabetes	442	10	Regression	64, 128
RAND HIE	20,190	9	Regression	128, 256, 512

4.3 Matched validation on six benchmarks

Table 2 lists the benchmarks. Preprocessing, seeded splits, grouped splitting for RAND HIE, feature generation, and target scaling are specified in Appendix C. The comparison uses identical matrices, validation losses, test splits, and tie-breaking rules for three paths:

$$d_i^{\mathrm{ridge}}(\tau) = s_i^2 + \lambda_{\min} + \tau, \quad d_i^{\mathrm{MT}}(\tau) = \max\{s_i^2 + \lambda_0, \mu_0^2 + \tau\}, \quad d_i^{\mathrm{matrix}}(\tau) = d_i^* + \tau.$$

All use twelve predeclared offset evaluations $\tau \in \{0\} \cup \{10^{-10}, \ldots, 10^{0}\}$; modified Tikhonov evaluations can coincide before the threshold becomes active. Every path remains feasible. Validation selects the offset and representation without inspecting test targets. These offset grids differ from the supplementary clamped ridge grid and are not pooled with it.

The default budgets are $K_* = 10^4$, $\Gamma = 100$, and $\lambda_0 = 10^{-12}$; the gain-tight regime changes only Γ to 10. There are 390 feature matrices, 780 paired design–regime cases, 2,340 estimator–configuration records, and 360 selected-model records. The evaluated gain and condition constraints hold with relative tolerance 10^{-9}. At zero offset the envelope never has greater retained-target distortion than either comparator, and Table 3 quantifies strict improvements. The gain term changes at least one envelope

Table 3 Zero-offset reduction in relative retained-target distortion, comparator minus the matrix envelope. Positive values favor the envelope. All comparisons use identical designs and budgets; strict means greater than 10^{-12}.

Regime	Comparator	Strict reduction	Median	Maximum
Default	Ridge	364/390	0.014871	0.056373
Default	Modified Tikhonov	347/390	0.008877	0.077136
Gain-tight	Ridge	390/390	0.018921	0.037802
Gain-tight	Modified Tikhonov	390/390	0.044068	0.123541

eigenvalue relative to the condition-only envelope in 354/390 default cases and all 390 gain-tight cases.

For these empirical matrices define $I_\eta = \{i : \sigma_i > \eta\sigma_1\}$, $P_\eta = \sum_{i\in I_\eta} u_i u_i^T$, and use the reported diagnostic

$$\psi_{\eta,Q}(T) = \frac{\|P_\eta(P - S_Q)T\|_{\mathrm{F}}}{\|P_\eta T\|_{\mathrm{F}}}, \qquad \eta = 10^{-12}, \tag{25}$$

with zero assigned when the denominator vanishes. The projector and numerator use the same retained subspace. This compares retained training-target energy, not exact-rank uniform feasibility. An additional controlled audit varies the cutoff between the designed exact spectrum, 10^{-14}, 10^{-12}, and 10^{-10} (Appendix D).

Table 4 reports the validation-selected predictive results for the three regularization paths.

Predictive outcomes are mixed. The matrix rule does not consistently beat ridge or modified Tikhonov on held-out scores, and its mean Diabetes RMSE exceeds ridge in both regimes. Standard deviations describe ten overlapping seeded split-and-feature runs; no statistical significance or population-level superiority is asserted. The conic target optimizer is used only in the small direct optimization checks, not in these benchmark predictions.

Table 4 Matched validation over twelve offset evaluations per design. Entries are mean ± sample standard deviation over ten seeds. Accuracy is in percent (larger is better); RMSE is in original units (smaller is better).

Regime	Data	Metric	Ridge	Modified Tikhonov	Matrix envelope
Default	Iris	Accuracy	95.33 ± 4.77	96.00 ± 3.78	95.33 ± 4.77
Default	Wine	Accuracy	97.78 ± 1.76	97.50 ± 1.58	97.50 ± 1.58
Default	Breast Cancer	Accuracy	97.19 ± 1.15	97.28 ± 0.97	97.19 ± 1.15
Default	Digits	Accuracy	95.17 ± 0.85	94.64 ± 1.08	95.17 ± 0.85
Default	Diabetes	RMSE	55.335 ± 3.832	55.082 ± 3.337	55.450 ± 3.669
Default	RAND HIE	RMSE	4.301 ± 0.335	4.303 ± 0.333	4.302 ± 0.335
Gain-tight	Iris	Accuracy	95.00 ± 5.03	94.67 ± 3.58	94.67 ± 4.77
Gain-tight	Wine	Accuracy	98.06 ± 2.29	98.33 ± 1.94	97.78 ± 2.19
Gain-tight	Breast Cancer	Accuracy	97.63 ± 0.83	97.54 ± 1.00	97.89 ± 0.74
Gain-tight	Digits	Accuracy	94.14 ± 1.08	92.67 ± 1.48	94.17 ± 1.21
Gain-tight	Diabetes	RMSE	54.262 ± 3.162	54.551 ± 3.037	54.635 ± 3.039
Gain-tight	RAND HIE	RMSE	4.294 ± 0.332	4.295 ± 0.331	4.294 ± 0.332

5 Concluding remarks

We have determined a single explicit penalty that solves an entire family of distortion-norm minimization problems under simultaneous floor, gain, and condition constraints. Its optimality holds against every admissible symmetric penalty, including noncommuting competitors. The mathematical mechanism is a sharp inverse-Hessian diagonal bound attained in every singular direction, followed by weak majorization. The equality analysis identifies both the fitted-target and coefficient maps of every Frobenius minimizer and gives penalty uniqueness at full column rank.

For individual targets, noncommuting penalties can improve the fit. The exact compression and feasible lift characterize this second optimization problem without retaining irrelevant null-space variables. Together, the common norm optimizer, the equality characterization, and the target-dependent reduction form the structural contribution of this work within established condition-constrained matrix optimization.

The computational contribution is the direct spectral evaluation of the common optimizer and the exact reduction of the target-specific problem to rank-sized matrix variables. The empirical evidence supports the structural statements without a universal prediction claim. The calibrated modified Tikhonov comparison directly tests the extra directional gain constraint; independent conic solves test the optimization formulas. Exact-rank and retained-subspace feasibility are reported separately. Scalar residual diagnostics concern the chosen objective, while repeated timing includes reusing coefficients already obtained during validation. Dense direct solvers remain strong references in this CPU setting.

Several limits remain. The target-dependent convex formulation has no general large-rank complexity advantage beyond the exact dimension reduction. Diagonal target second moments are a substantive restriction for the average-fit corollary. Fixed-penalty gain bounds do not cover revalidation, input perturbations, or feature resampling. Numerical residuals and conic gaps are nominal double-precision diagnostics, not certified enclosures. The small classification benchmarks and two regression tasks do not establish broad predictive superiority; grouped covariate splits do not replace unavailable participant identifiers. Further work can address validated rounding, structured large-rank optimization, and statistically justified budget selection.

A Scalar feasibility and compatible training fit

A.1 An iteration horizon is a condition-number ceiling

Fix $K > 1$, $\Gamma > 0$, an integer polynomial order $p \geq 2$, an integer update horizon $M \geq 0$, an inverse-residual tolerance $0 < \rho < 1$, and $\lambda_0 > 0$. Consider

$$Y_0 = \frac{I_L}{b + \lambda}, \quad E_k = I_L - G_\lambda Y_k, \quad Y_{k+1} = Y_k \sum_{j=0}^{p-1} E_k^j. \tag{26}$$

Order two is Newton–Schulz. The initialization ensures nonnegative scalar error factors, not optimal minimax scaling. The three operational specifications are

$$\kappa_2(G_\lambda) \le K, \qquad g_\lambda \le \Gamma, \qquad \|E_M\|_2 \le \rho. \tag{27}$$

Lemma A.1 (Reduction of the horizon constraint) *For* (26) *in exact arithmetic,*

$$\|E_M\|_2 = q^{p^M}, \qquad q = \frac{b-a}{b+\lambda} = 1 - \frac{1}{\kappa_2(G_\lambda)}. \tag{28}$$

Consequently, the last inequality of (27) *is equivalent to*

$$\kappa_2(G_\lambda) \le K_M := \frac{1}{1-\rho^{1/p^M}}. \tag{29}$$

The first and third specifications together are therefore exactly the single ceiling $K_* := \min\{K, K_M\}$.

Proof The finite geometric identity gives $G_\lambda Y_{k+1} = (I_L - E_k)\sum_{j=0}^{p-1} E_k^j = I_L - E_k^p$, so $E_{k+1} = E_k^p$. Since Y_0 is a scalar matrix, all iterates commute with G_λ. Its eigenvalues lie in $[a+\lambda, b+\lambda]$; those of E_0 lie in $[0, (b-a)/(b+\lambda)]$. The upper endpoint is attained, which proves the first identity. The second follows from (4). Rearranging $q \le \rho^{1/p^M}$ proves (29). If $a = b$, then $E_0 = 0$ and $\kappa_2(G_\lambda) = 1$, so the same statement holds. □

Remark A.2 (Dependent specifications) The horizon is a solver-dependent interpretation of conditioning, not an independent source of stability. For $p = 2$, $M = 20$, and $\rho = 10^{-8}$, $K_M \approx 56\,924$. Thus $K = 10^4$ already enforces the default horizon in the experiments. Smaller horizons can become restrictive. The statement concerns matrix-polynomial updates, not wall-clock time, equal work across orders, or attained floating-point accuracy.

Theorem A.3 (Least admissible regularization) *Set*

$$\lambda_C = \left[\frac{b - K_* a}{K_* - 1}\right]_+, \qquad \lambda_\Gamma = \max\left\{0, \max_{1\le i\le r}\left(\frac{\sigma_i}{\Gamma} - \sigma_i^2\right)\right\}, \tag{30}$$

$$\lambda_{\min} = \max\{\lambda_0, \lambda_C, \lambda_\Gamma\}. \tag{31}$$

Then (27) *and* $\lambda \geq \lambda_0$ *hold if and only if* $\lambda \geq \lambda_{\min}$. *In particular,* (31) *is an exact minimum, not just a sufficient regularization rule.*

Proof Lemma A.1 reduces the two condition specifications to $(b+\lambda)/(a+\lambda) \leq K_*$. Since $K_* > 1$, this is equivalent to $\lambda \geq (b - K_* a)/(K_* - 1)$. For each positive singular value, $\sigma_i/(\sigma_i^2 + \lambda) \leq \Gamma$ is equivalent to $\lambda \geq \sigma_i/\Gamma - \sigma_i^2$. Intersecting all scalar upper intervals with $[\lambda_0, \infty)$ gives exactly (31), proving necessity, sufficiency, and minimality. □

The theorem combines a classical ridge shift, ridge gains, and the explicit iteration identity. Its operational content is the feasible parameter set and the distinction between independent and dependent requirements. It does not establish statistical optimality of the minimum parameter.

For numerical evaluation, writing $\ell = \log(\rho)/p^M$ gives

$$K_M = -\frac{1}{\operatorname{expm1}(\ell)}, \quad \lambda_K = \left[\frac{b - Ka}{K - 1}\right]_+, \quad \lambda_M = \left[(b-a)\operatorname{expm1}(-\ell) - a\right]_+ . \tag{32}$$

Use $\lambda_C = \lambda_K$ if $K \leq K_M$ and $\lambda_C = \lambda_M$ otherwise. This equivalent evaluation avoids subtracting nearly equal numbers in $K_M - 1$ when K_M is close to one. The independent four-term expression $\max(\lambda_0, \lambda_K, \lambda_M, \lambda_\Gamma)$ is retained as a software consistency check.

A.2 Uniform stability–fit feasibility

Let $P = \Phi\Phi^+$ and $S_\lambda = \Phi R_\lambda$. The unregularized least-squares fitted targets are PT, while ridge fits $S_\lambda T$. A fit-distortion requirement constrains their difference, not generalization error.

Theorem A.4 (Uniform feasibility interval) *For* $0 < \theta < 1$,

$$\|P - S_\lambda\|_2 = \frac{\lambda}{a_+ + \lambda}. \tag{33}$$

Therefore (27), $\lambda \geq \lambda_0$, *and* $\|P - S_\lambda\|_2 \leq \theta$ *are simultaneously feasible if and only if*

$$\lambda_{\min} \leq \lambda_{\max}^{\text{fit}} := \frac{\theta a_+}{1-\theta}. \tag{34}$$

Every admissible parameter lies in $[\lambda_{\min}, \lambda_{\max}^{\text{fit}}]$. *On this interval the smallest parameter minimizes training residual and coefficient bias relative to* $\Phi^+ T$ *for every fixed* T. *Both minima are unique when* $PT \neq 0$*; when* $PT = 0$, *every feasible parameter is a minimizer.*

Proof On range(Φ), the eigenvalues of $P - S_\lambda$ are $\lambda/(\sigma_i^2 + \lambda)$, while on its orthogonal complement they are zero. The largest is (33). Rearrangement gives the upper endpoint and Theorem A.3 gives the lower endpoint. Write $T = U_r Z + T_\perp$, $U_r^T T_\perp = 0$. Orthogonality gives

$$\|\Phi B_\lambda - T\|_{\text{F}}^2 = \|T_\perp\|_{\text{F}}^2 + \sum_i \left(\frac{\lambda}{\sigma_i^2 + \lambda}\right)^2 \|Z_{i,:}\|_2^2,$$

$$\left\|B_\lambda - \Phi^+ T\right\|_{\text{F}}^2 = \sum_i \left(\frac{\lambda}{\sigma_i(\sigma_i^2 + \lambda)}\right)^2 \|Z_{i,:}\|_2^2.$$

All terms are nondecreasing in $\lambda > 0$. If $PT \neq 0$, at least one weight $\|Z_{i,:}\|_2^2$ is positive, so both quantities are strictly increasing and their minima occur uniquely at the smallest feasible parameter. If $PT = 0$, all these weights vanish and both quantities are constant on the feasible interval. □

For example, $\Phi = \text{diag}(1, 10^{-6})/\sqrt{1 + 10^{-12}}$ with $K = 10^4$ requires approximately $\lambda \geq 10^{-4}$. Requiring $\theta = 0.1$ imposes $\lambda \leq 1.11 \times 10^{-13}$. No scalar ridge parameter can meet both requirements. Infeasibility is a property of the specifications, not a failure of the optimizer.

A.3 Target-specific fit and a less conservative interval

A uniform requirement protects every target direction, including directions containing negligible target energy. When the actual fitted target is known, an explicit scalar test can be less conservative.

Proposition A.5 (Target-specific feasibility) *Assume $PT \neq 0$ and define*

$$\psi_T(\lambda)^2 := \frac{\|(P - S_\lambda)T\|_{\mathrm{F}}^2}{\|PT\|_{\mathrm{F}}^2} = \frac{\sum_i [\lambda/(\sigma_i^2 + \lambda)]^2 \left\|(U_r^T T)_{i,:}\right\|_2^2}{\sum_i \left\|(U_r^T T)_{i,:}\right\|_2^2}. \tag{35}$$

The function ψ_T is continuous and strictly increasing on $(0, \infty)$, with limits 0 and 1. For every $0 < \theta < 1$ there is a unique $\lambda_{\theta,T} > 0$ with $\psi_T(\lambda_{\theta,T}) = \theta$. Joint numerical and target-fit feasibility holds exactly when

$$\lambda_{\min} \leq \lambda_{\theta,T}. \qquad \textit{Moreover,} \quad \lambda_{\theta,T} \geq \lambda_{\max}^{\mathrm{fit}}. \tag{36}$$

The upper endpoint can be computed by bracketing and bisection. If $PT = 0$, both fitted targets are zero and the unnormalized fit-distortion requirement imposes no upper bound.

Proof Expanding $(P - S_\lambda)T$ in U_r gives (35). Each summand is continuous and strictly increasing whenever its target weight is positive. At least one such weight is positive because $PT \neq 0$, proving strict monotonicity and the limits. The intermediate value theorem gives the unique root and the interval test. Furthermore, $\psi_T(\lambda) \leq \|P - S_\lambda\|_2$, so the uniform upper endpoint is always admissible for the target-specific condition. The case $PT = 0$ follows from $U_r^T T = 0$. □

This result is a target-energy refinement of classical spectral filtering, not a new loss function. It permits a direct feasibility audit rather than assuming that conditioning and preservation of the fitted targets must be compatible. The predictive benchmark below selects parameters without an upper fit constraint; a separate synthetic experiment tests both fit intervals.

B All-iterate bounds and residual-controlled computation

B.1 Monotone filters along the inverse iteration

Theorem B.1 (Amplification during iteration) *For (26), let* $B_k = Y_k\Phi^T T$ *and* $e_i = 1 - (\sigma_i^2+\lambda)/(b+\lambda)$. *Then*

$$Y_k\Phi^T = V_r \operatorname{diag}\left(\frac{\sigma_i}{\sigma_i^2+\lambda}(1-e_i^{p^k})\right) U_r^T. \tag{37}$$

If $\lambda \geq \lambda_{\min}$, *every exact-arithmetic iterate satisfies*

$$\left\|Y_k\Phi^T\right\|_2 \leq g_\lambda \leq \Gamma, \qquad \|B_k - B_\lambda\|_{\mathrm{F}} \leq q^{p^k}\,\|B_\lambda\|_{\mathrm{F}}\,. \tag{38}$$

More precisely,

$$\|B_k - B_\lambda\|_{\mathrm{F}}^2 = \sum_i e_i^{2p^k} \frac{\sigma_i^2}{(\sigma_i^2+\lambda)^2} \left\|(U_r^T T)_{i,:}\right\|_2^2. \tag{39}$$

Proof From $G_\lambda Y_k = I_L - E_0^{p^k}$, write $Y_k = G_\lambda^{-1}(I_L - E_0^{p^k})$ and diagonalize on the singular bases. This gives (37). Since $0 \leq e_i \leq q < 1$, the factor $1 - e_i^{p^k}$ lies in $[0,1]$, proving the gain bound. Subtract the ridge filter, use orthonormality for (39), and bound every e_i by q. □

The all-iterate result depends on the nonnegative scalar error factors. Alternative initialization, another high-order polynomial, or arbitrary roundoff perturbations need not preserve it. Our implementations use orders two and three, symmetrize the approximate inverse, and recompute residuals; they are not assumed to satisfy the exact filter identity bit for bit.

B.2 A posteriori coefficient and prediction error

Theorem B.2 (Stationarity residual bounds) *For any approximate coefficient matrix* $\widehat{B}$, *define its exact original-design residual*

$$D = \Phi^T(T - \Phi\widehat{B}) - \lambda\widehat{B}. \tag{40}$$

Then

$$\left\|\widehat{B} - B_\lambda\right\|_{\mathrm{F}} \le \frac{\|D\|_{\mathrm{F}}}{a+\lambda}, \tag{41}$$

$$J_\lambda(\widehat{B}) - J_\lambda(B_\lambda) = \tfrac{1}{2}\operatorname{tr}(D^T G_\lambda^{-1} D) \le \frac{\|D\|_{\mathrm{F}}^2}{2(a+\lambda)}. \tag{42}$$

For a normalized held-out feature vector $h \in \mathbb{R}^L$, the maximum absolute score error is at most

$$\left\|h^T(\widehat{B} - B_\lambda)\right\|_\infty \le \delta(h) := \frac{\|h\|_2 \, \|D\|_{\mathrm{F}}}{a+\lambda}. \tag{43}$$

If the largest entry of the exact product $h^T\widehat{B}$ is unique and exceeds the second largest by more than $2\delta(h)$, its class equals that of the exact ridge predictor. For a floating-point score vector $\widehat{z}$ with a validated bound $\left\|\widehat{z} - h^T\widehat{B}\right\|_\infty \le \varepsilon_{\mathrm{score}}$, a margin greater than $2(\delta(h) + \varepsilon_{\mathrm{score}})$ suffices.

Proof The normal equations imply $\widehat{B} - B_\lambda = -G_\lambda^{-1} D$ and $\left\|G_\lambda^{-1}\right\|_2 = (a+\lambda)^{-1}$, giving (41). Expanding the quadratic objective about its minimizer cancels the linear term and yields one half of the squared G_λ-norm of the coefficient error. Substituting $-G_\lambda^{-1} D$ proves (42). Apply Cauchy–Schwarz to each score coordinate for (43). A top score can decrease by at most $\delta(h)$ while a competitor can increase by at most $\delta(h)$; the stated margin therefore preserves the top class. □

The coefficient and objective inequalities are consequences of positive-definite quadratic optimization. Their use here is as solver-independent stopping diagnostics. Agreement with the exact ridge classifier is not correctness relative to the unknown label, and is not an adversarial-input or clinical reliability claim.

Corollary B.3 (Combined target and solver error) *Keep Φ and $\lambda \ge \lambda_{\min}$ fixed. If an approximate solve for $T + \Delta T$ has exact residual norm at most η, then*

$$\left\|\widehat{B}_\Delta - B_\lambda(T)\right\|_{\mathrm{F}} \le \Gamma \, \|\Delta T\|_{\mathrm{F}} + \frac{\eta}{a+\lambda}. \tag{44}$$

Proof Insert $B_\lambda(T + \Delta T)$, use (5) and Theorem B.2, and apply the triangle inequality. □

In floating point, an evaluated residual $\widehat{D}$ is not the exact D. Given validated bounds $\left\|D-\widehat{D}\right\|_{\mathrm{F}} \le \varepsilon_D$ and $0 \le \underline{a} \le a$, with outward-rounded norm and arithmetic evaluation, a rigorous computable counterpart is

$$\left\|\widehat{B}-B_\lambda\right\|_{\mathrm{F}} \le \frac{\left\|\widehat{D}\right\|_{\mathrm{F}}+\varepsilon_D}{\underline{a}+\lambda}. \tag{45}$$

Our experiments use double precision, an SVD point estimate of a, and $\varepsilon_D = 0$. The reported values are therefore *nominal residual diagnostics*, not outward-rounded interval certificates. Recomputing (40) with the original feature matrix avoids using only the residual of an already rounded normal-equation matrix, but it does not by itself eliminate rounding error.

B.3 Spectrum-free feasibility and calibration cost

Proposition B.4 (A spectrum-free sufficient rule) *For $\|\Phi\|_{\mathrm{F}} = 1$, a sufficient regularization floor is*

$$\overline{\lambda}_{\min} = \max\left\{\lambda_0, \frac{1}{K_*-1}, \frac{1}{4\Gamma^2}\right\}. \tag{46}$$

Using $Y_0 = I_L/(1+\lambda)$ then satisfies the specified inverse horizon in exact arithmetic. In general, (46) is sufficient rather than minimal.

Proof Unit Frobenius norm gives $b \le 1$ and $a \ge 0$, hence $\kappa_2(G_\lambda) \le (1+\lambda)/\lambda$. Also, $s/\Gamma - s^2 \le 1/(4\Gamma^2)$ for every $s \ge 0$. For the stated initialization, the initial error eigenvalues lie in $[0, 1/(1+\lambda)]$. Thus its residual after M updates is at most $(1/(1+\lambda))^{p^M}$. The term $1/(K_*-1)$ enforces both the curvature ceiling and this worst-case horizon. Together these observations prove the sufficient rule. □

The spectrum-free rule uses only normalization and budgets. Validation can be performed by Cholesky solves without an SVD. Exact spectral calibration usually

permits less regularization, whereas the conservative rule removes the decomposition cost; neither option is universally preferable. General certified spectral enclosures can interpolate between these extremes. Uncertified singular-value estimates remain diagnostics, not guaranteed enclosures.

C Data and supplementary scalar protocol

C.1 Numerical feasibility and validation

For a base grid $\mathcal{L}$, a lower floor f, and an optional upper fit limit u, define

$$\mathcal{L}(f, u) = \{\max(\ell, f) : \ell \in \mathcal{L},\ \max(\ell, f) \leq u\}, \tag{47}$$

with duplicate parameters removed. An empty set is reported as candidate-grid infeasibility. This does not alone imply that the continuous feasible interval is empty: the grid may omit a feasible endpoint. The main prediction experiments use $u = \infty$; the fit experiments investigate finite u. All condition-, gain-, horizon-, and joint-constraint ablations use the same base grid, validation loss, data splits, and candidate hidden matrices. Differences in parameter tuning are therefore not mistaken for evidence about one constraint.

Algorithm 2. Feasibility-constrained output training

Input: training/validation features and targets, numerical budgets, candidate parameters, coefficient tolerance ϵ_B.

1. Normalize all feature splits with the training Frobenius norm. Obtain either the exact spectral floor (31) or the sufficient floor (46).
2. Select a feasible parameter by validation mean squared error (MSE); select the feature configuration by the same validation rule. Do not inspect test targets.
3. Freeze the chosen features and λ. Solve the ridge objective using a direct solver or the stated hyperpower iteration.
4. Recompute (40). Accept an iterative solve when $\|D\|_{\mathrm{F}} / (a + \lambda) \leq \epsilon_B \|T\|_{\mathrm{F}}$, or use a validated version when available. Use the valid lower bound $a = 0$ in a spectrum-free branch without spectral information. Record the update limit and nonfinite failures explicitly.
5. Return coefficients, selected parameter, diagnostics, and optional margin-based predictor-agreement checks. Any fallback uses the same objective and has separately reported cost.

Calibration by SVD is part of the exact-floor implementation, not a hidden free operation. Dense formation of $\Phi^T \Phi$ costs $O(NL^2)$ and each inverse-polynomial matrix product costs $O(L^3)$. For a small number of target columns, direct solves or conjugate gradients can be cheaper than an inverse approximation. We compare these possibilities at the same chosen λ, and separately measure complete fixed-representation training pipelines.

C.2 Synthetic matrices and formula audits

The controlled solver experiment uses $(N, L) = (240, 80), (96, 96), (80, 160)$. Independent Gaussian QR factors define singular bases. With $m = \min(N, L)$, the well-conditioned spectrum decreases linearly from 1 to 10^{-1}; the moderate and severe spectra decrease geometrically from 1 to 10^{-4} and 10^{-12}, respectively. For the rank-deficient and clustered spectra, the first $m - \lfloor m/10 \rfloor$ values decrease linearly from 1 to 10^{-1}, and the remaining $\lfloor m/10 \rfloor$ values are set to zero or 10^{-14}, respectively. Matrices are normalized to unit Frobenius norm. Ten seeds per shape–spectrum pair give 150 designs; orders two and three give 300 same-parameter ridge solves. Targets have three Gaussian columns. These records were regenerated from the accompanying baseline script, not transcribed from conference tables.

A separate audit checks the collapsed and uncollapsed floors, the condition–horizon identity, active-constraint minimality, and residual bounds on 200 small random designs with orders two to four, horizons zero to nine, and varied budgets. Another 300 cases compare uniform and target-specific fit feasibility: 50 designs, Gaussian or leading-five-direction targets, and $\theta \in \{0.1, 0.5, 0.9\}$. The latter experiment defines a numerical positive singular subspace using cutoff $10^{-12}\sigma_1$; its displayed projectors use that same subspace. This convention is distinct from asserting an exact rank in floating point. Numerical checks supplement, and do not replace, the proofs.

C.3 Classification and regression data

Table 2 lists four classification benchmarks and two regression benchmarks. Bundled scikit-learn data [20–22] and the statsmodels RAND HIE subset [23] are used through the recorded package versions. Digits consists of 8×8 images and is not MNIST. RAND HIE predicts the count of outpatient visits as a squared-error regression benchmark, not a count-distribution model or a clinical/causal application.

Seeds 0–9 define training/validation/test partitions. Classification uses stratified 60/20/20 partitions; Diabetes uses random partitions of the same proportions. RAND HIE contains 2,760 distinct covariate rows among 20,190 observations. Its split groups identical covariate rows and assigns groups, not rows, in 60/20/20 proportions. Thus no identical covariate vector crosses splits, but participant identifiers are unavailable; we do not claim patient-independent evaluation. Actual row counts are retained in the split manifest.

Input standardization is fitted only on training samples. Regression targets are centered and scaled using training mean and population standard deviation; root mean squared error (RMSE) is converted back to original target units. Hidden weights are Gaussian with variance $1/d$, biases are uniform on $[-1, 1]$, and activations are tanh, ReLU, and sigmoid. Widths are 64 and 128 for the five smaller data sets, and 128, 256, and 512 for RAND HIE. Identical preactivations are used across activation choices for a given seed and width. Classification uses one-hot targets, with no additional output intercept. Each estimator selects its parameter and width/activation by validation MSE; test scores never enter selection and no train–validation refit is performed.

The design gives 390 realized feature matrices, 3,510 estimator–configuration records, and 540 selected-model records. Only ten selected test scores per estimator and data set enter predictive summaries. Standard deviations describe the seeded split-and-feature protocol; overlapping splits are not treated as independent populations and no universal significance claim is made.

C.4 Baselines, constraints, and perturbations

The shared ridge grid is $\mathcal{L} = \{10^{-10}, 10^{-9}, \ldots, 10^{0}\}$. Five common-validation variants use no additional constraint (Ridge-CV), a condition floor (Condition-CV), a gain floor (Gain-CV), a horizon floor (Horizon-CV), or the joint floor (Joint-CV). Spectrum-free-CV uses (46) and Cholesky-based validation; SVD is used

only for its ex-post audit, not to choose its parameter. A pseudoinverse baseline retains singular values above $10^{-12}\sigma_1$; TSVD-CV selects a relative cutoff from $\{10^{-12}, 10^{-6}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. The OCReP-rule baseline uses $\sigma_1\sigma_r$ on the same numerical positive spectrum. In wide or rank-deficient designs this is an explicitly defined extension of that scalar rule, not a full replication of OCReP.

The default specifications are

$$K = 10^4, \quad \Gamma = 100, \quad p = 2, \quad M = 20, \quad \rho = 10^{-8}, \quad \lambda_0 = 10^{-12}, \quad \epsilon_B = 10^{-7}. \tag{48}$$

They are operational choices, not statistically optimal constants. The order-two floor also suffices for order three, and both orders solve the same selected ridge objective. The inverse-residual specification and the absolute coefficient stopping target are distinct.

For every realized feature matrix, matched solves use Cholesky, augmented QR via LAPACK `gelsy`, conjugate gradients (relative tolerance 10^{-11}, at most $10L$ steps per right-hand side), and hyperpower orders two and three (at most 20 updates). Residuals are evaluated with the original design. The independently computed SVD ridge solution supplies the reference, not a rank-truncated unregularized pseudoinverse. Non-accepted runs are preserved and no fallback is hidden.

Fixed-design regression perturbations use relative target magnitudes $10^{-4}, 10^{-2}, 10^{-1}$ in Gaussian and maximizing singular directions, for pseudoinverse, Ridge-CV, and Joint-CV. Features and selected parameters remain fixed. The experiment records gain and noisy-test RMSE; it does not represent revalidation after noisy observations. A separate synthetic perturbation test provides the same operator check at four magnitudes. The purpose is to evaluate (5), not claim that random noise usually realizes the worst direction.

C.5 Implementation, timing, and reproducibility

The scalar tables use the recorded reproduction of 29 September 2026, with Python 3.12.14, NumPy 2.3.5, SciPy 1.17.0, scikit-learn 1.8.0, pandas 2.2.3, and statsmodels 0.14.6. New matrix, conic, cutoff, and timing records use Python 3.12.14 and the same numerical package versions. Conic computations additionally use CVXPY 1.9.3, Clarabel 0.11.1, and SCS 3.3.1. Exact versions, CPU model, thread settings, and source-data hashes are in the manifests. All timed BLAS computations use one thread.

The timing table and figure are regenerated exclusively from one new controlled run. Each fixed-representation pipeline includes standardization, feature construction, normalization, calibration, parameter validation, and a final solve or reuse of the already selected coefficients. Loading, split construction, and outer architecture search are excluded. Five policies are compared: exact SVD calibration followed by Cholesky or hyperpower-2, spectrum-free calibration followed by Cholesky, and a reuse variant of each calibration policy. Reuse returns the coefficients already evaluated during validation. One warm-up and five measurements for each of six designs give 150 records, with rotated execution order. The conservative spectrum-free floor may select a different parameter. Historical timing aggregates are not mixed into this run.

Online Resource 1 includes complete build scripts, numerical records, split indices, data hashes, and environment manifests. The scalar and new matrix data directories are separate. Every displayed table and figure can be rebuilt from the supplied records without an unavailable manuscript template. Rerunning the experiments requires the listed dependencies and publicly available bundled data.

Table 5 Fit feasibility on the designed exact spectrum and on the retained spectrum ($\eta = 10^{-12}$). Each row has five spectra and ten seeds. Exact-spectrum calculations use the prescribed singular values and bases, not a numerical rank estimate.

Target	θ	Exact uniform	Exact target	Retained uniform	Retained target
Gaussian	0.1	20/50	20/50	30/50	30/50
Gaussian	0.5	20/50	30/50	30/50	30/50
Gaussian	0.9	20/50	46/50	30/50	47/50
Leading	0.1	20/50	50/50	30/50	50/50
Leading	0.5	20/50	50/50	30/50	50/50
Leading	0.9	20/50	50/50	30/50	50/50

D Supporting scalar numerical results

D.1 Numerical feasibility and target-specific fit

All 200 randomized formula checks and 48 explicit boundary cases pass. The largest computed discrepancy in $q = 1 - 1/\kappa_2(G_\lambda)$ is 2.70×10^{-16}; the reduced and four-term floors agree numerically. Boundary cases include isotropic, rank-one, wide, and severely ill-conditioned spectra and horizon zero; invalid budgets are rejected. In the 390 expanded realized feature matrices, the pre-validation floor is driven by the condition constraint in 305 cases, the gain constraint in 59 cases, and the positive numerical floor in 26 cases. The horizon never independently determines the default floor, as Remark A.2 requires. Validation chooses the exact joint floor in 118 of these configurations.

The exact designed-spectrum audit accepts 120/300 cases under the uniform criterion and 246/300 under the target-specific criterion. With the relative cutoff 10^{-12}, the corresponding counts are 180/300 and 257/300. Table 5 reports both, and Table 6 shows the cutoff sensitivity. These are different subspace questions; the retained-subspace counts must not be interpreted as exact-rank feasibility. Figure 2 makes the distinction explicit.

Table 6 Sensitivity of feasibility to the relative spectral cutoff. The designed exact spectrum retains all prescribed positive singular values, including the 10^{-14} cluster.

Cutoff	Uniform	Target-specific
Exact design	120/300	246/300
10^{-14}	180/300	256/300
10^{-12}	180/300	257/300
10^{-10}	180/300	259/300

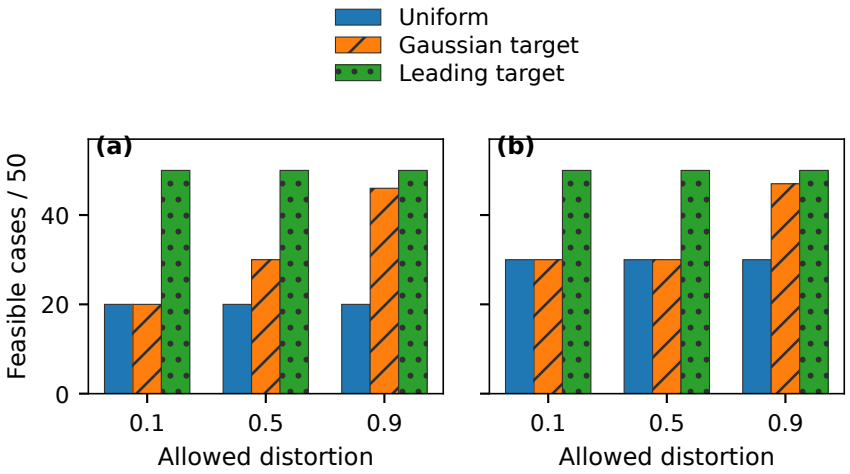


Fig. 2 Feasible cases among 50 controlled designs for each target type and fit tolerance. The operator condition protects every target; the target-specific condition uses projected training-target energy. (a) The prescribed exact spectrum. (b) The retained subspace at relative cutoff 10^{-12}

The target-specific improvement is not automatic predictive improvement. It says that protecting every weak feature direction can be stricter than protecting the target energy actually present. A target-specific feasible interval should not be interpreted as preserving unseen targets or guaranteeing generalization. The main prediction comparisons do not impose this upper constraint.

Table 7 Same-parameter ridge solves across 390 feature configurations. Step counts are not equal computational work; CG reports the maximum across target columns. Classification disagreement is measured against the SVD ridge predictor, not ground truth.

Solver	Accepted	Median steps	Max. relative error	Disagreements
Cholesky	390/390	–	5.44×10^{-13}	0
Augmented QR	390/390	–	2.65×10^{-14}	0
CG	390/390	48	3.98×10^{-9}	0
Hyperpower-2	390/390	14	2.25×10^{-8}	0
Hyperpower-3	390/390	9	3.70×10^{-8}	0

D.2 Matched ridge solves and nominal diagnostics

All 300 regenerated synthetic hyperpower solves meet the stopping target, with maximum relative error 4.94×10^{-9} against the same-parameter SVD ridge reference. All 1950 expanded matched-objective solves, including 780 hyperpower solves, also meet the implemented nominal coefficient target. Table 7 reports errors separately from stopping flags. There are no observed classification-label disagreements with the exact-ridge reference; all such predictions pass the nominal margin test.

Finite-precision diagnostic bounds are not strict verified enclosures. The largest error-to-nominal-bound ratio among all expanded solves is 2.5838, for an augmented-QR result with absolute error 1.10×10^{-12}, relative error 1.30×10^{-14}, and nominal bound 4.28×10^{-13}. At this scale, residual evaluation and reference error cannot be ignored. The maximum ratio among the hyperpower runs is 0.9999996. The full records preserve these discrepancies rather than converting diagnostic values into machine-verified guarantees.

Figure 3 compares measured coefficient errors with the nominal residual bounds.

A solve that reaches the residual target approximates its chosen regularized objective. It does not recover the unregularized pseudoinverse to the same accuracy. Similarly, successful margin diagnostics indicate agreement with the reference predictor, not correctness of all test labels.

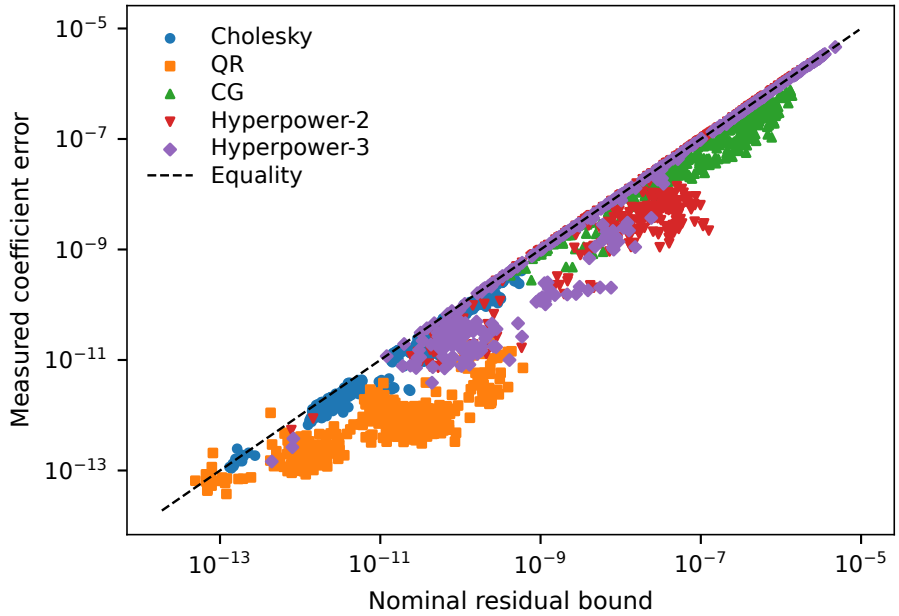


Fig. 3 Measured coefficient errors against same-parameter SVD ridge references versus nominal stationarity-residual bounds. These are floating-point diagnostics, not interval-validated enclosures; errors near roundoff include reference-solution error

Table 8 Test accuracy (%, mean ± sample standard deviation over ten seeded partitions and feature draws). Every validation-based variant uses the same selection protocol; no test score enters model selection.

Estimator	Iris	Wine	Breast Cancer	Digits
Pseudoinverse	89.33 ± 5.40	93.33 ± 2.99	96.93 ± 1.19	94.94 ± 0.83
TSVD + validation	95.00 ± 4.78	97.50 ± 2.43	97.11 ± 0.72	94.94 ± 0.83
Ridge + validation	95.00 ± 4.51	98.33 ± 1.43	97.19 ± 1.00	95.17 ± 0.85
OCReP spectral rule	96.00 ± 3.78	97.78 ± 2.19	97.11 ± 1.02	92.97 ± 0.96
Condition + validation	95.33 ± 4.77	97.50 ± 1.58	97.19 ± 1.00	95.17 ± 0.85
Gain + validation	95.33 ± 4.77	98.06 ± 1.34	97.19 ± 1.00	95.17 ± 0.85
Horizon + validation	95.33 ± 4.77	98.33 ± 1.43	97.19 ± 1.00	95.17 ± 0.85
Joint + validation	95.33 ± 4.77	97.50 ± 1.58	97.19 ± 1.00	95.17 ± 0.85
Spectrum-free + validation	95.33 ± 4.77	97.50 ± 1.58	97.19 ± 1.00	95.17 ± 0.85

D.3 Common-validation ablations and predictive performance

Tables 8 and 9 report held-out classification and regression scores under the common validation protocol.

Table 9 Regression in original target units: RMSE (mean ± sample standard deviation) and mean test R^2 over ten seeded partitions. RAND HIE uses identical-covariate-group splits and squared loss; this is not a clinical or count-distribution claim.

Estimator	Diabetes RMSE	R^2	RAND HIE RMSE	R^2
Pseudoinverse	59.973 ± 3.850	0.339	4.382 ± 0.309	0.032
TSVD + validation	54.623 ± 2.802	0.453	4.303 ± 0.336	0.067
Ridge + validation	55.153 ± 3.167	0.442	4.299 ± 0.334	0.069
OCReP spectral rule	54.881 ± 3.424	0.447	4.302 ± 0.334	0.068
Condition + validation	55.153 ± 3.167	0.442	4.300 ± 0.335	0.069
Gain + validation	55.039 ± 3.220	0.445	4.302 ± 0.335	0.068
Horizon + validation	55.068 ± 3.203	0.444	4.299 ± 0.334	0.069
Joint + validation	55.153 ± 3.167	0.442	4.300 ± 0.335	0.069
Spectrum-free + validation	55.153 ± 3.167	0.442	4.299 ± 0.334	0.069

Joint-CV has no uniform predictive advantage. It matches Ridge-CV on Breast Cancer, Digits, and Diabetes, has a small descriptive increase on Iris, and is less accurate on Wine; RAND HIE RMSE values are close. Neither these differences nor any table entry is declared statistically significant. RAND HIE mean R^2 is low (approximately 0.07), so the medium-size experiment demonstrates numerical behavior under the specified squared-error model rather than strong predictive fit.

Condition-CV and Joint-CV select the same representation and parameter in all 60 default selected-model cases. This is an important negative ablation result: the gain constraint does not improve their selected predictions in this default protocol. Nevertheless, Ridge-CV meets the gain and condition ceilings in only 50/60 selected cases, versus 60/60 for Joint-CV. Validation alone and numerical feasibility are therefore distinguishable even when their predictive scores are similar.

To test binding regimes explicitly, a diagnostic follow-up fixes width 128 and sigmoid features on Iris, Digits, Diabetes, and RAND HIE. With ten seeds, five common-validation policies, and three regimes (default, $\Gamma = 10$, or $M = 8$), it contains 600 rows. This follow-up is designed to probe the constraints, not to tune final predictive settings using test results. In the gain-tight regime, Condition-CV meets the gain ceiling in only 1/40 cases; Joint-CV meets it in 40/40. In the horizon-tight

Table 10 Constraint fulfilment in the fixed-width sigmoid stress diagnostic. Gain-tight uses $\Gamma = 10$; horizon-tight uses $M = 8$; all other budgets remain at their defaults. Every method uses the same validation-grid rule.

Regime	Method	Condition	Gain	Horizon
Default	Ridge-CV	17/40	17/40	17/40
Default	Condition-CV	40/40	40/40	40/40
Default	Gain-CV	16/40	40/40	40/40
Default	Horizon-CV	17/40	17/40	40/40
Default	Joint-CV	40/40	40/40	40/40
Gain-tight	Ridge-CV	17/40	1/40	17/40
Gain-tight	Condition-CV	40/40	1/40	40/40
Gain-tight	Gain-CV	40/40	40/40	40/40
Gain-tight	Horizon-CV	17/40	1/40	40/40
Gain-tight	Joint-CV	40/40	40/40	40/40
Horizon-tight	Ridge-CV	17/40	17/40	0/40
Horizon-tight	Condition-CV	40/40	40/40	0/40
Horizon-tight	Gain-CV	16/40	40/40	0/40
Horizon-tight	Horizon-CV	40/40	40/40	40/40
Horizon-tight	Joint-CV	40/40	40/40	40/40

regime, Condition-CV meets the stricter horizon in 0/40 cases, versus 40/40 for Joint-CV. These are different requested operating points; they do not overturn the exact condition–horizon equivalence.

Table 10 records constraint fulfilment in the three stress regimes.

All constraint variants use validation on the same base-grid rule. This avoids confounding a single assigned parameter with a searched grid. The horizon-only variant is weaker than the condition-only variant at the default settings, as Lemma A.1 predicts. Equal predictions from different variants are not evidence that the constraints are universally equivalent; they can result from validation selecting the same interior parameter.

Figure 4 compares the coefficient gains of paired validation-only and jointly constrained fits.

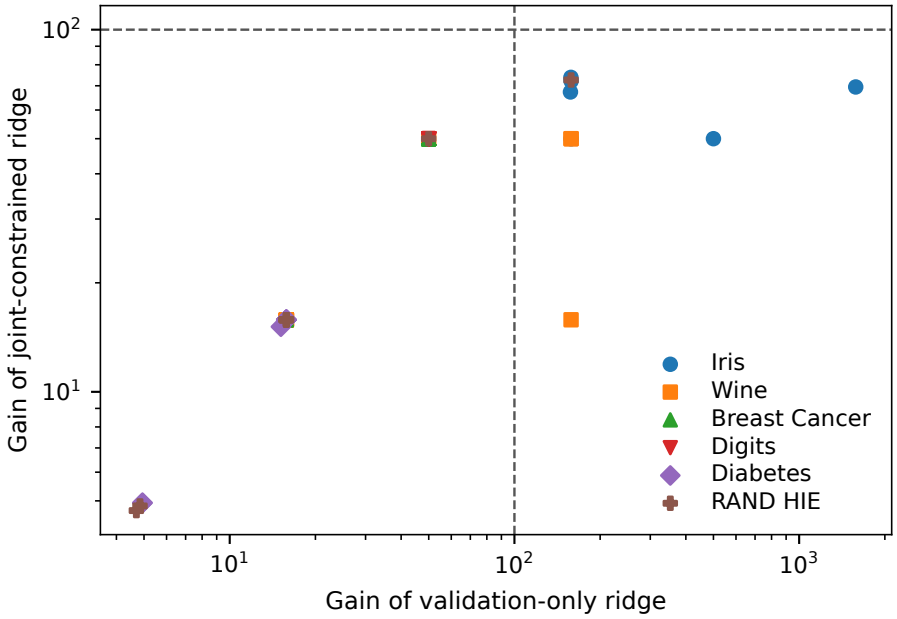


Fig. 4 Target-to-coefficient gains for the 60 selected models, paired by data set and seed, under ordinary ridge validation and joint constraints. Dashed reference lines mark the prescribed gain ceiling of 100. Similar predictive scores need not imply identical numerical sensitivity

D.4 Noise and computation–accuracy trade-offs

The expanded regression perturbation study contains 2,700 records. Maximizing directions attain the exact fixed-parameter operator gain with maximum relative discrepancy 2.89×10^{-15}. Gaussian-direction gains are smaller; their maximum observed-to-operator-gain ratio is 0.521. These calculations evaluate a linear response at a fixed parameter, not a robustness claim for a reselected model. The corresponding noisy RMSE records are supplied without assuming monotone or universal improvements in generalization. Figure 5 shows the observed gains for Gaussian and maximizing perturbation directions.

Table 11 compares complete pipeline times for the five training policies.

In this run, the ratio of exact-calibration to spectrum-free median time is 1.88–2.92 for the resolve policies and 2.19–3.64 for reuse. These are policy comparisons: the conservative floor can change the selected parameter. Reusing the selected coefficients

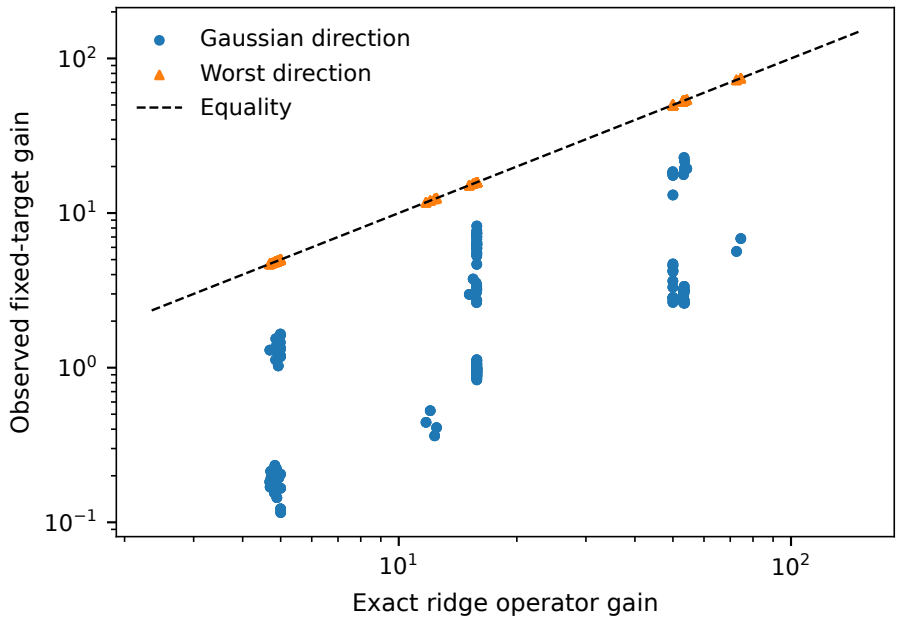


Fig. 5 Observed target-noise gain for joint-constrained regression fits compared with the exact fixed-parameter ridge gain. Maximizing directions attain the operator norm up to numerical error, while Gaussian directions generally do not. No parameter re-selection follows perturbation

Table 11 Median complete fixed-representation pipeline times (ms), five repetitions, one controlled run. Resolve policies recompute the selected coefficients; reuse policies return the coefficients retained during validation. Quartiles and component times are in Online Resource 1.

Data	Width	Exact/Chol.	Exact/HP2	Free/Chol.	Exact/reuse	Free/reuse
Digits	128	13.4	22.8	7.1	12.6	5.8
Digits	256	36.3	80.0	16.5	32.1	12.4
Digits	512	147.9	430.3	55.8	135.5	39.4
RAND HIE	128	133.1	144.8	49.5	124.1	38.7
RAND HIE	256	318.1	413.3	120.9	295.3	96.5
RAND HIE	512	938.5	1247.7	320.9	877.4	240.8

avoids an unnecessary final solve; it does not remove the preceding validation cost. There is no hardware-independent speed claim. The exact hyperpower pipeline is slower than exact Cholesky in all six designs in this run.

Figure 6 displays these median pipeline times across the six fixed designs.

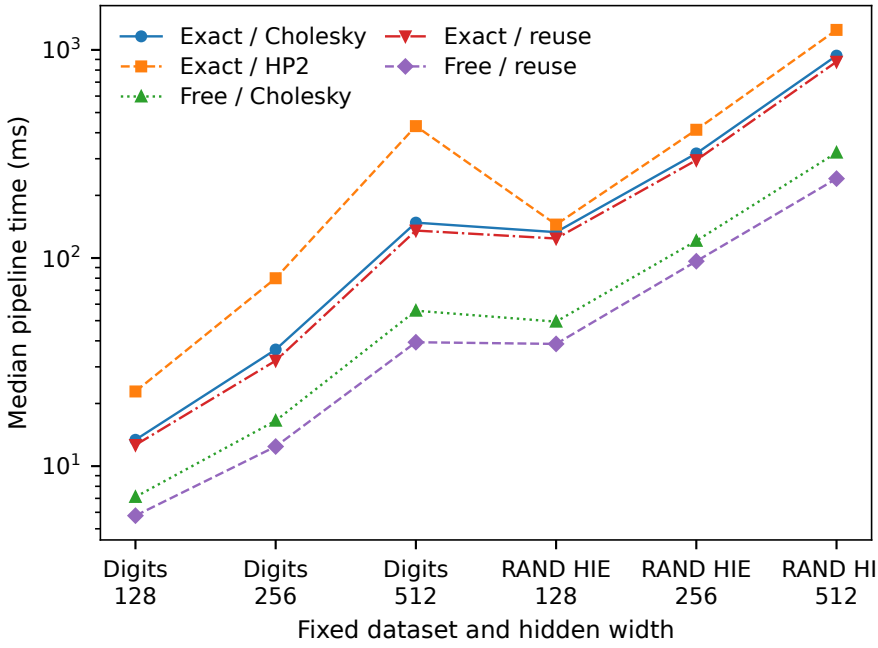


Fig. 6 Median complete fixed-representation pipeline time over five repetitions. Preprocessing, calibration, parameter search, and final solution are included; the outer architecture search is not. A spectrum-free floor changes the admissible parameters, so these curves compare training policies, not equal-objective solver kernels

Statements and Declarations

Funding

This research is partially supported by the Institute of Mathematics, Vietnam Academy of Science and Technology, Hanoi, under grant number CSCL01.01/25-26.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Author contributions

Nguyen Bich Van developed the mathematical foundations and provided the proofs. Khong Ngoc Anh handled the data and implemented the code for the numerical experiments.

Data availability

The numerical records, split indices, source-data hashes, and environment manifests supporting this study are provided in Online Resource 1. The underlying benchmark data are available through the scikit-learn and statsmodels dataset loaders cited in the text. The scripts obtain these bundled datasets from the installed packages; the archive contains the recorded outputs and provenance rather than separate copies of every raw dataset.

Code availability

The experiment and artifact-generation scripts, dependency list, and reproduction instructions are provided in Online Resource 1.

Supplementary information

Online Resource 1 (ESM_1.zip): Experiment code, recorded numerical outputs, data provenance, environment manifests, and instructions for reproducing the computational results.

References

[1] Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. Technometrics **12**, 55–67 (1970). https://doi.org/10.1080/00401706.1970.10488634

[2] Fuhry, M., Reichel, L.: A new Tikhonov regularization method. Numerical Algorithms **59**, 433–445 (2012). https://doi.org/10.1007/s11075-011-9498-x

[3] Noschese, S., Reichel, L.: Some matrix nearness problems suggested by Tikhonov regularization. Linear Algebra and its Applications **502**, 366–386 (2016). https://doi.org/10.1016/j.laa.2015.04.008

[4] Li, J.-f., Li, W., Vong, S.-W.: Efficient algorithms for solving condition number-constrained matrix minimization problems. Linear Algebra and its Applications **607**, 190–230 (2020). https://doi.org/10.1016/j.laa.2020.08.007

[5] Livne, O.E., Castellano, K.E., McCaffrey, D.F.: Numerical algorithm for estimating a conditioned symmetric positive definite matrix under constraints. Numerical Linear Algebra with Applications **31**, e2559 (2024). https://doi.org/10.1002/nla.2559

[6] Livne, O.E.: CPD: Algorithms for Conditioned Symmetric Positive Definite Matrix Estimation Under Constraints. Educational Testing Service, source code (2022). Version at commit `75de3027bc7d81cadfe947ae3661cc22c66763f4`. https://github.com/EducationalTestingService/cpd.

[7] Huang, G.-B., Zhu, Q.-Y., Siew, C.-K.: Extreme learning machine: theory and applications. Neurocomputing **70**, 489–501 (2006). https://doi.org/10.1016/j.neucom.2005.12.126

[8] Ben-Israel, A., Greville, T.N.E.: Generalized Inverses: Theory and Applications, 2nd edn. Springer, New York (2003). https://doi.org/10.1007/b97366

[9] Higham, N.J.: Accuracy and Stability of Numerical Algorithms, 2nd edn. SIAM, Philadelphia (2002). https://doi.org/10.1137/1.9780898718027

[10] Nguyen, B.V., Khong, N.A.: Spectral Stability of Pseudoinverse-Based Extreme Learning Machine. Proceedings of the 10th International Conference on Interdisciplinary Research on Computer Science, Psychology and Education (ICICPE 2026), vol. 10, no. 1, pp. 84–86, Chiang Mai, Thailand, 19–21 August 2026. ISSN 2508-7436 (2026). Public preprint: arXiv:2607.08581v1. https://doi.org/10.48550/arXiv.2607.08581.

[11] Cancelliere, R., Gai, M., Gallinari, P., Rubini, L.: OCReP: An optimally conditioned regularization for pseudoinversion based neural training. Neural Networks **71**, 76–87 (2015). https://doi.org/10.1016/j.neunet.2015.07.015

[12] Cancelliere, R., Deluca, R., Gai, M., Gallinari, P., Rubini, L.: An analysis of numerical issues in neural training by pseudoinversion. Computational and Applied Mathematics **36**, 599–609 (2017). https://doi.org/10.1007/s40314-015-0246-z

[13] Tabeart, J.M., Dance, S.L., Lawless, A.S., Nichols, N.K., Waller, J.A.: Improving the condition number of estimated covariance matrices. Tellus A: Dynamic Meteorology and Oceanography **72**, 1696646 (2020). https://doi.org/10.1080/16000870.2019.1696646

[14] Won, J.-H., Lim, J., Kim, S.-J., Rajaratnam, B.: Condition-number-regularized covariance estimation. Journal of the Royal Statistical Society, Series B **75**, 427–450 (2013). https://doi.org/10.1111/j.1467-9868.2012.01049.x

[15] Tanaka, M., Nakata, K.: Positive definite matrix approximation with condition number constraint. Optimization Letters **8**, 939–947 (2014). https://doi.org/10.1007/s11590-013-0632-7

[16] Eldar, Y.C.: Minimum variance in biased estimation: bounds and asymptotically optimal estimators. IEEE Transactions on Signal Processing **52**, 1915–1930 (2004). https://doi.org/10.1109/TSP.2004.828929

[17] Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2004). https://web.stanford.edu/~boyd/cvxbook/

[18] Rudi, A., Rosasco, L.: Generalization properties of learning with random features. Advances in Neural Information Processing Systems **30** (2017). https://proceedi

ngs.neurips.cc/paper_files/paper/2017/hash/61b1fb3f59e28c67f3925f3c79be81a1-Abstract.html

[19] Bhatia, R.: Matrix Analysis. Springer, New York (1997). https://doi.org/10.1007/978-1-4612-0653-8

[20] Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011). https://jmlr.org/papers/v12/pedregosa11a.html

[21] Scikit-learn developers: Toy datasets. Documentation v1.8; data loaded using scikit-learn 1.8.0. https://scikit-learn.org/1.8/datasets/toy_dataset.html.

[22] Scikit-learn developers: Load diabetes, regression dataset. Documentation v1.8; raw features obtained with `scaled=False`. https://scikit-learn.org/1.8/modules/generated/sklearn.datasets.load_diabetes.html.

[23] Statsmodels developers: RAND Health Insurance Experiment Data. Documentation v0.14.4; data loaded using statsmodels 0.14.6. https://www.statsmodels.org/v0.14.4/datasets/generated/randhie.html .

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.zip