Regression as an Ill-Posed Inverse Problem

Spectral Filters, Regularization, and Runtime Trade-offs

Project Summary

Executive Summary

This project reframes regression as an *inverse problem* and uses the *spectrum* (singular values / eigenvalues) as the main language for understanding (i) instability under multicollinearity, (ii) bias–variance trade-offs under regularization, and (iii) computational scaling of common solvers. The central thesis is that many seemingly different estimators are best understood as *spectral filters*: they replace an unbounded inverse (OLS) by bounded, direction-dependent shrinkage, stabilizing prediction in ill-conditioned and overparameterized regimes.

Problem Framing: Why Regression Becomes Ill-Posed

In finite-dimensional linear regression,

$$y = X\beta + \varepsilon,$$

estimation amounts to solving $X\beta \approx y$. With the thin SVD $X = U\Sigma V^\top$ (rank r), OLS can be written as

$$\widehat{\beta}_{\mathrm{OLS}} = V\Sigma^{-1}U^\top y = \sum_{j=1}^{r} \frac{1}{\sigma_j}\langle u_j, y\rangle v_j.$$

When the design spectrum decays sharply, small σ_j create large amplification factors $1/\sigma_j$, yielding unstable coefficients and high variance. The same phenomenon appears in operator limits (compact operators on Hilbert spaces) where eigenvalues $\lambda_i \downarrow 0$ make the inverse unbounded; thus, ill-conditioning is not just numerical—it is structural.

Core Idea: Regularization as Spectral Filtering

The project studies three regularization mechanisms through a unified spectral lens.

(A) Tikhonov / Ridge: explicit bounded filter

Ridge solves

$$\widehat{\beta}_{\mathrm{ridge}}(\lambda) = \arg\min_{\beta} \left\{ \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \right\}, \quad (X^\top X + \lambda I)\widehat{\beta} = X^\top y.$$

In the SVD basis this becomes

$$\widehat{\beta}_{\mathrm{ridge}}(\lambda) = VD(\lambda)U^\top y, \qquad D_{jj}(\lambda) = \frac{\sigma_j}{\sigma_j^2 + \lambda},$$

i.e., ridge applies the *bounded* spectral filter

$$g_j^{\text{ridge}}(\lambda) = \frac{\sigma_j}{\sigma_j^2 + \lambda}.$$

Large-variance tail directions ($\sigma_j^2 \ll \lambda$) are aggressively damped; well-identified directions ($\sigma_j^2 \gg \lambda$) behave OLS-like.

(B) Kernel Ridge with Nyström: explicit filter on kernel spectrum

To move beyond linear features, kernel ridge regression (KRR) is considered with a Gaussian RBF kernel and Nyström approximation for scalability. The Nyström eigenspace yields approximate kernel eigenvalues κ_i and a KRR filter

$$g_i^{\text{KRR}}(\lambda) = \frac{\kappa_i}{\kappa_i + \lambda}, \quad df_{\text{KRR}}(\lambda) \approx \sum_i \frac{\kappa_i}{\kappa_i + \lambda}.$$

Conceptually, KRR is ridge in an RKHS: it uses the same filtering mechanism, now acting on the spectrum of a (regularized) kernel operator rather than $X^\top X$.

(C) Early-stopped Gradient Descent: implicit filter

Full-batch gradient descent on squared loss, initialized at $\beta^{(0)} = 0$,

$$\beta^{(t+1)} = \beta^{(t)} + \eta X^\top \big(y - X\beta^{(t)}\big),$$

induces an *implicit* spectral filter:

$$\beta^{(t)} = \sum_{j:\sigma_j>0} g_j^{\text{GD}}(t)\, \alpha_j\, v_j, \qquad g_j^{\text{GD}}(t) = \frac{1 - (1 - \eta\sigma_j^2)^t}{\sigma_j}, \quad \alpha_j = \langle u_j, y \rangle.$$

As $t \to \infty$, $g_j^{\text{GD}}(t) \to 1/\sigma_j$ and GD approaches OLS in the singular subspace; for finite t, high-σ_j directions are fit earlier, while ill-conditioned directions remain suppressed—so iteration count acts as a regularization knob.

Experimental Design

1) Synthetic spectral stress-test: Toeplitz AR(1) designs

To isolate spectral effects, the project simulates Gaussian designs with Toeplitz AR(1) covariance:

- Correlation levels: $\rho \in \{0.10, 0.50, 0.95, 0.99\}$.
- Sample sizes: $n \in \{1000, 2000, \ldots, 10000\}$ with $p = n/10$.
- Data generation: each row is $N_p(0, \Sigma(\rho))$ with $\Sigma_{j\ell} = \rho^{|j-\ell|}$.
- Spectral diagnostics: randomized truncated SVD; truncated condition number $\kappa_k(X) = \sigma_1/\sigma_k$.

- Signal alignment: coefficients constructed in the right-singular basis with $\beta_j^\star \propto j^{-1}$, placing most signal in leading singular directions.

This provides a controlled continuum from well-conditioned to severely ill-conditioned regimes.

2) Real-data high-dimensional regime: BlogFeedback (UCI) with polynomial features

A realistic ill-conditioning + heavy-tail setting is created from the BlogFeedback dataset:

- Dataset scale: 52,397 posts, 280 numerical predictors, response = comments in next 24 hours.
- Target engineering: $y = \log(1 + \text{comments})$ to stabilize heavy-tailed counts.
- Subsample and predictors: $n = 5000$ random posts, $d = 60$ selected predictors.
- Feature map: all degree-2 polynomial terms (no explicit intercept), producing

$$p = d + \frac{d(d+1)}{2} = 60 + \frac{60 \cdot 61}{2} = 1890,$$

  followed by column standardization so the penalty is isotropic in Euclidean coordinates.
- Diagnostics (spectral + leverage): a small number of singular directions captures most variance (low effective rank) and row norms show extreme leverage, both of which amplify the need for regularization.

Results and Key Findings

Spectral behavior and generalization

Across both synthetic and real designs, the same qualitative pattern emerges:

- **OLS instability is spectral.** When spectra are concentrated, $1/\sigma_j$ explodes in tail directions; variance dominates and test error degrades sharply.
- **All three methods are filters.** Ridge, Nyström-KRR, and early-stopped GD differ algorithmically but act similarly as shrinkage operators in a spectral basis.
- **U-shaped test error.** Test MSE is minimized at an intermediate regularization level: weak regularization behaves OLS-like (low bias, high variance), while strong regularization over-shrinks (high bias, low variance).
- **Real-data confirmation.** On the polynomial BlogFeedback design ($p = 1890$), cross-validation strongly prefers substantial shrinkage; regularization collapses coefficient norms and stabilizes predictions in the presence of heavy tails and high-leverage rows.

Computational trade-offs (runtime scaling)

A key practical message is that the spectral viewpoint naturally motivates scalable solvers:

- Normal equations exhibit cubic scaling in high-dimensional regimes.
- Randomized truncated SVD, Nyström features, and iterative GD behave essentially quadratically in n (with fixed subspace rank / landmark count), enabling controlled approximation of the dominant spectral directions.

- On the BlogFeedback polynomial design, truncated-SVD ridge and Nyström-RBF ridge achieve near "low-dimensional" training times (after one-off precomputations), while GD is dominated by repeated passes over data.

Skills and Methods Demonstrated

- Spectral diagnostics: SVD-based instability analysis, condition numbers, effective rank, leverage/row-norm inspection.
- Regularization theory: ridge/Tikhonov as bounded inverse; degrees-of-freedom and bias–variance interpretation in spectral coordinates.
- Scalable approximations: randomized truncated SVD; Nyström approximation for kernel ridge.
- Optimization as implicit regularization: early stopping in gradient descent interpreted as a spectral low-pass filter.
- Experimental rigor: synthetic designs with controllable spectra (Toeplitz AR(1)) and empirical validation on a real, heavy-tailed dataset with engineered high-dimensional features.

Conclusion

In modern overparameterized, collinear, and heavy-tailed settings, regression should be designed and interpreted as *filtering an inverse problem*. Making the filter explicit—via ridge, kernel methods, or early stopping—yields a principled and computationally scalable route to stable estimation, interpretable bias–variance behavior, and improved generalization.