Regression as an Ill-Posed Inverse Problem: Spectral Filters, Regularization, and Runtime Trade-offs

1 Spectral and Operator-Theoretic View of Regression

Regression can be viewed through linear operators and their spectra. In the finite-dimensional linear model

$$y = X\beta + \varepsilon, \qquad X \in \mathbb{R}^{n\times p},$$

the design matrix X acts as a linear operator mapping coefficients β to fitted values. Estimation amounts to approximately solving the inverse problem $X\beta \approx y$.

In the infinite-dimensional limit (e.g. nonparametric regression or function estimation), X is replaced by a compact, self-adjoint operator T on a Hilbert space H (e.g. $L^2(\mathcal{X})$). By the spectral theorem, T admits an eigen-expansion

$$Tf = \sum_{i\in I} \lambda_i \langle f, \phi_i\rangle\, \phi_i, \qquad \lambda_i \geq 0,\ \lambda_i \downarrow 0,$$

on an orthonormal basis $\{\phi_i\}$ of $\overline{\mathrm{Range}(T)}$.[1] The decay $\lambda_i \longrightarrow 0$ is structurally elegant but creates instability for naive inversion: solving $Tf = y$ requires division by small eigenvalues, the operator analogue of ill-conditioned least squares.[2]

Project goal and working hypothesis. Using spectral decompositions, we study this instability and show how Tikhonov (ridge) regularization replaces unbounded division by eigenvalues with bounded *spectral filters*. This makes explicit how regularization trades bias for variance and stabilizes the estimator [3, 5, 9]. Our working hypothesis is that, in designs with highly concentrated spectra (such as Gaussian AR(1) Toeplitz covariances and the polynomially expanded BlogFeedback design), explicit Tikhonov regularisation, Nyström-approximated kernel ridge, and early-stopped gradient descent all act as *spectral filters* that primarily attenuate small-singular-value directions. We expect this common spectral mechanism to (i) produce broadly similar bias–variance trade-offs and U-shaped test-error curves as the regularisation parameter (or iteration count) is varied, and (ii) explain the observed stability and runtime differences relative to unregularised least squares

Why this perspective? The finite-dimensional SVD and the spectral theorem for compact operators are two versions of the same idea. In finite dimensions, SVD clarifies multicollinearity and ill-conditioning; in infinite dimensions, the same spectral picture governs kernel methods, covariance operators, and integral operators [4, 10]. This operator viewpoint, common in learning theory and inverse problems, makes the effect of regularization on the spectrum—and hence on the bias–variance trade-off, stability, and generalization—transparent.

Regression as an Inverse Problem on a Hilbert Space

In linear regression we observe (X, y) and seek β such that

$$X\beta \approx y.$$

[1] See, for example, standard functional analysis notes such as Salamon, *Functional Analysis* [8].

[2] For the connection between inverse problems and regression, see, e.g., the MIT notes on inverse problems and the monographs by Engl, Hanke, and Neubauer, and by Kirsch [3, 6, 7].

When $X^\top X$ is invertible and well-conditioned, the normal equations

$$X^\top X \beta = X^\top y$$

yield the (unique) OLS solution $\hat{\beta} = (X^\top X)^{-1} X^\top y$.

In an infinite-dimensional setting, let $T : H \longrightarrow H$ be compact, self-adjoint, and positive. Writing

$$y = \sum_i \langle y, \phi_i \rangle \phi_i$$

in the eigenbasis of T, the formal solution of $Tf = y$ is

$$f = \sum_{i:\,\lambda_i > 0} \frac{1}{\lambda_i} \langle y, \phi_i \rangle \, \phi_i.$$

If y has non-negligible components in directions where λ_i is tiny, the factors $1/\lambda_i$ explode. The induced inverse $T^\dagger$ is unbounded and the solution map $y \mapsto f$ is not continuous, the classical Hadamard notion of instability in inverse problems.[3]

Thus, both in finite and infinite dimensions, decaying spectra create ill-posedness: solving $Tf = y$ or $X^\top X \beta = X^\top y$ involves large amplification along poorly identified directions [3].

Spectral Decomposition and Instability of OLS

Let $X = U \Sigma V^\top$ be the (thin) SVD with singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$ and rank r. Whenever the least-squares solution is identifiable, we can write

$$\hat{\beta}_{\text{OLS}} \;=\; V \Sigma^{-1} U^\top y \;=\; \sum_{j=1}^{r} \frac{1}{\sigma_j} \, (u_j^\top y) \, v_j,$$

where u_j and v_j are the singular vectors. Each data component $u_j^\top y$ is scaled by $1/\sigma_j$: small σ_j produce large coefficients and high variance.

In the operator setting, formally solving $Tf = y$ gives

$$f = \sum_{i:\,\lambda_i > 0} \lambda_i^{-1} \langle y, \phi_i \rangle \phi_i,$$

so noise in directions with small λ_i is similarly amplified. The spectral decomposition therefore provides a unified explanation for why OLS becomes unstable under ill-conditioning and why compact operators lead to ill-posed inverse problems [3, 6].

[3] See, e.g., Kirsch, *An Introduction to the Mathematical Theory of Inverse Problems* [6].

Tikhonov Regularization as a Spectral Filter

Tikhonov (ridge) regularization stabilizes this inversion by penalizing the coefficients. In finite dimensions it solves

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \left\{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\},$$

with normal equations

$$(X^\top X + \lambda I)\,\hat{\beta}_{\text{ridge}} = X^\top y.$$

Using the SVD of X gives

$$\hat{\beta}_{\text{ridge}} = V D(\lambda) U^\top y, \qquad D_{jj}(\lambda) = \frac{\sigma_j}{\sigma_j^2 + \lambda}.$$

Thus ridge applies the spectral filter

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

to each singular direction. For $\sigma_j^2 \gg \lambda$ this approximates $1/\sigma_j$ (OLS-like behavior), while for $\sigma_j^2 \ll \lambda$ it behaves like σ_j/λ and remains bounded. Directions with $\sigma_j = 0$ are discarded. This is the classical Tikhonov regularization scheme in inverse problems and ridge regression in statistics [3, 4, 5, 9].

In the operator setting, if $T\phi_i = \lambda_i \phi_i$ then the regularized solution of $(T + \lambda I) f_\lambda = y$ is

$$f_\lambda = \sum_i \frac{1}{\lambda_i + \lambda} \langle y, \phi_i \rangle\, \phi_i,$$

so Tikhonov replaces $1/\lambda_i$ by the bounded filter $1/(\lambda_i + \lambda)$ with operator norm at most $1/\lambda$. In both finite- and infinite-dimensional cases, regularization replaces an unbounded inverse by a bounded spectral filter. Detailed per-direction bias, variance, and degrees-of-freedom formulas for these filters are collected in Appendix A [3, 4].

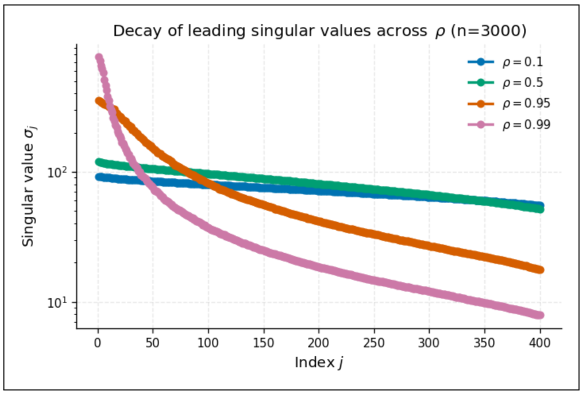


Figure 1: Decay of leading singular values across AR(1) correlation levels.

Spectral Simulation Design: Toeplitz AR(1) Designs

To connect the operator-theoretic picture with finite-sample behavior, we consider random designs whose population covariance has Toeplitz AR(1) structure, a standard testbed in spectral analyses of regression and inverse problems [3, 4]. For

$$\rho \in \{0.10, 0.50, 0.95, 0.99\}, \quad n \in \{1000, 2000, \ldots, 10000\},$$

we generate $X \in \mathbb{R}^{n \times p}$ with $p = n/10$ by simulating n independent Gaussian AR(1) series of length p:

$$X_{i1} \sim N(0,1), \qquad X_{ij} = \rho X_{i,j-1} + \eta_{ij}, \quad \eta_{ij} \sim N(0, 1 - \rho^2),\ j = 2, \ldots, p.$$

This yields

$$\mathrm{Cov}(X_{ij}, X_{i\ell}) = \rho^{|j-\ell|},$$

so the population covariance $\Sigma(\rho)$ is Toeplitz with entries $\Sigma(\rho)_{j\ell} = \rho^{|j-\ell|}$. The rows of X are i.i.d. $N_p(0, \Sigma(\rho))$ and

$$\frac{1}{n} X^\top X \to \Sigma(\rho) \quad \text{a.s. as } n \to \infty.$$

We column-standardize X so that each column has unit empirical variance, preserving the correlation (and hence spectral) structure up to sampling noise. For each (ρ, n) we compute a randomized truncated SVD

$$X \approx U_k \Sigma_k V_k^\top, \qquad \Sigma_k = \mathrm{diag}(\sigma_1, \ldots, \sigma_k),$$

and study the truncated condition number

$$\kappa_k(X) = \frac{\sigma_1}{\sigma_k}.$$

As $\rho \uparrow 1$, the spectrum becomes sharply concentrated: a few leading singular values dominate and the tail collapses, so $\kappa_k(X)$ grows rapidly.

Figure 1 illustrates this effect: higher correlations produce faster spectral decay and more severe ill-conditioning.

Data-generating Process in the Right-singular Basis

To align with the spectral analysis, we construct the true coefficient vector $\beta^\star \in \mathbb{R}^p$ in the right-singular basis. With

$$X \approx U_r \Sigma_r V_r^\top,$$

we set

$$\beta^\star = \sum_{j=1}^{r} \beta_j^\star v_j, \qquad \beta_j^\star \propto j^{-\alpha},$$

and use $\alpha = 1$ in the reported runs. The coefficients decay like $1/j$, so $\beta^\star$ is "smooth" relative to the spectrum and most signal lies in leading singular directions. Writing $\beta_V^\star = (\beta_1^\star, \ldots, \beta_r^\star)^\top$ and normalizing $\|\beta_V^\star\|_2 = 1$, we set $\beta^\star = V_r^\top \beta_V^\star$ to keep $\|\beta^\star\|_2$ comparable across (ρ, n).

We then generate responses from

$$y = X\beta^\star + \varepsilon, \qquad \varepsilon \sim N_n(0, \sigma_\varepsilon^2 I_n),$$

and an independent test set $(X_{\text{test}}, y_{\text{test}})$ with the same Toeplitz structure and noise variance. The same $\beta^\star$ is used for all methods (ridge, kernel ridge, gradient descent).

Linear Ridge, Kernel Ridge, and Gradient Descent

Linear OLS vs Ridge in the Spectral Basis

Given $X \approx U_k \Sigma_k V_k^\top$, the OLS estimator (in the identifiable subspace) is

$$\hat{\beta}_{\text{OLS}} = V_k \Sigma_k^{-1} U_k^\top y = \sum_{j=1}^{k} \frac{1}{\sigma_j} \langle y, u_j \rangle v_j,$$

with spectral filter $g_j^{\text{OLS}} = 1/\sigma_j$.

Ridge with penalty $\lambda > 0$ uses

$$\hat{\beta}_{\text{ridge}}(\lambda) = V_k D(\lambda) U_k^\top y, \qquad D_{jj}(\lambda) = \frac{\sigma_j}{\sigma_j^2 + \lambda},$$

so

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

We compute $\lambda \mapsto \text{Bias}^2(\lambda)$, $\lambda \mapsto \text{Var}(\lambda)$, and

$$R_{\text{test}}(\lambda) = \frac{1}{n_{\text{test}}} \left\| X_{\text{test}} \hat{\beta}_{\text{ridge}}(\lambda) - y_{\text{test}} \right\|_2^2,$$

on a logarithmic grid of λ scaled by σ_1^2. As ρ increases and the tail of the spectrum collapses, OLS variance and test MSE grow sharply, while ridge maintains bounded filters and admits an optimal $\lambda > 0$, consistent with classical ridge theory [4, 5].

Figure 2 compares $1/\sigma_j$ and $g_j^{\text{ridge}}(\lambda)$ for $\rho = 0.99$. The OLS filter diverges in the tail as $\sigma_j \to 0$, while the ridge filter is uniformly bounded and heavily damps the most ill-conditioned directions.

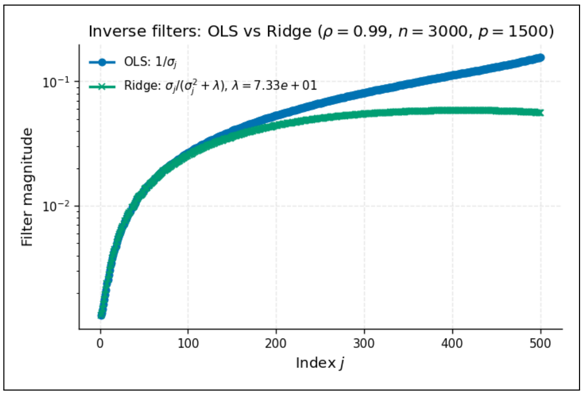


Figure 2: Inverse filters $1/\sigma_j$ vs. $g_j^{\text{ridge}}(\lambda)$ in a highly correlated design.

Kernel Ridge and Nyström Spectral Approximation

To move beyond linear features, we also consider kernel ridge regression with a Gaussian RBF kernel

$$k(x, x') = \exp\left(- \frac{\|x - x'\|_2^2}{2\ell^2} \right),$$

with lengthscale ℓ chosen by the median-distance heuristic on a subsample. The $n \times n$ Gram matrix $K_{ij} = k(x_i, x_j)$ is approximated via Nyström using m landmarks: we form $W \in \mathbb{R}^{m \times m}$ and $C \in \mathbb{R}^{n \times m}$, compute $W = Q \Lambda Q^\top$, and set

$$Z = C\, Q \Lambda^{-1/2} Q^\top, \qquad K_{\text{nys}} := Z Z^\top.$$

An economical SVD $Z = U_Z \Sigma_Z V_Z^\top$ yields approximate kernel eigenvalues $\kappa_i = s_i^2$ [11].

Kernel ridge in the RKHS of k has solution

$$\hat{f}_\lambda(\cdot) = \sum_{i=1}^{n} \alpha_i(\lambda) k(x_i, \cdot), \qquad \alpha(\lambda) = (K + \lambda I_n)^{-1} y,$$

and in the Nyström eigenspace applies the filter

$$g_i^{\mathrm{KRR}}(\lambda) = \frac{\kappa_i}{\kappa_i + \lambda}.$$

The effective degrees of freedom are approximated by

$$\mathrm{df}_{\mathrm{KRR}}(\lambda) \approx \sum_i \frac{\kappa_i}{\kappa_i + \lambda},$$

and we compute both $\mathrm{df}_{\mathrm{KRR}}(\lambda)$ and the corresponding test MSE

$$R_{\mathrm{test}}^{\mathrm{KRR}}(\lambda) = \frac{1}{n_{\mathrm{test}}} \left\| \hat{y}_{\lambda,\mathrm{test}}^{\mathrm{KRR}} - y_{\mathrm{test}} \right\|_2^2.$$

Qualitatively, kernel ridge implements the same spectral mechanism as linear ridge, but now on the spectrum of the kernel operator [4, 10].

Implicit Spectral Regularization: Ridge vs. Early-Stopped Gradient Descent

Ridge is an *explicit* spectral filter. Many algorithms instead employ iterative methods such as gradient descent, which induce *implicit* spectral regularization, a phenomenon widely studied in modern learning theory [4, 10].

Consider the squared loss

$$L(\beta) = \frac{1}{2} \|y - X\beta\|_2^2,$$

and gradient descent with step size $\eta > 0$ and initialization $\beta^{(0)} = 0$:

$$\beta^{(t+1)} = \beta^{(t)} + \eta X^\top (y - X\beta^{(t)}), \qquad t = 0, 1, 2, \dots.$$

Using the SVD $X = U\Sigma V^\top$ and writing $\alpha_j = u_j^\top y$, one obtains

$$\beta^{(t)} = \sum_{j:\,\sigma_j > 0} g_j^{\mathrm{GD}}(t)\, \alpha_j\, v_j, \qquad g_j^{\mathrm{GD}}(t) = \frac{1 - (1 - \eta\sigma_j^2)^t}{\sigma_j},$$

with η scaled as $\eta = \texttt{gd\_step\_scale}/\sigma_1^2$ and $\texttt{gd\_step\_scale} < 2$ for stability.

Two facts summarize the behavior:

- As $t \to \infty$, $(1 - \eta\sigma_j^2)^t \to 0$ and $g_j^{\mathrm{GD}}(t) \to 1/\sigma_j$, so gradient descent converges to the OLS solution in the singular subspace.
- For finite t, directions with larger σ_j^2 are fit faster since $(1 - \eta\sigma_j^2)^t \approx e^{-\eta\sigma_j^2 t}$ decays more rapidly. Early-stopped gradient descent therefore acts as a low-pass filter ordered by σ_j^2: well-identified directions are fitted first, ill-conditioned ones remain suppressed.

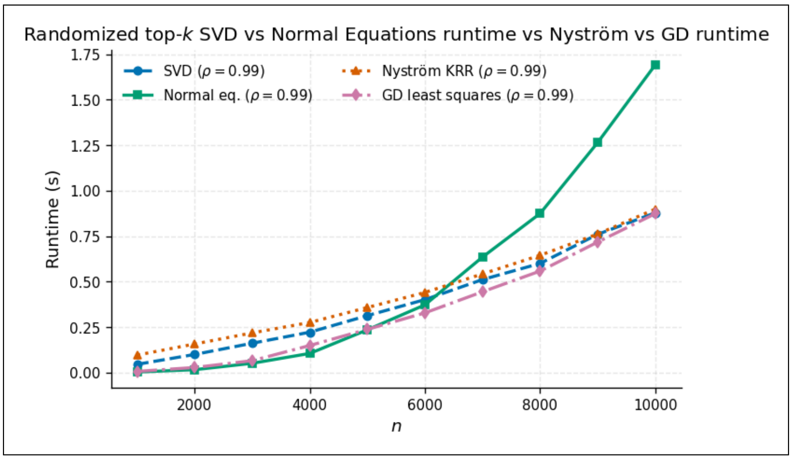


Figure 3: Scaling of runtime for normal equations, randomized SVD, Nyström KRR, and gradient descent as n grows with $p = n/2$.

The fitted values after t steps are

$$\hat{y}^{(t)} = X\beta^{(t)} = \sum_j \left(1 - (1 - \eta\sigma_j^2)^t\right) \alpha_j \, u_j,$$

with smoothing factors

$$s_j^{\mathrm{GD}}(t) = 1 - (1 - \eta\sigma_j^2)^t,$$

and effective degrees of freedom

$$\mathrm{df}_{\mathrm{GD}}(t) = \sum_j s_j^{\mathrm{GD}}(t) = \sum_j \left[1 - (1 - \eta\sigma_j^2)^t\right].$$

Empirically, $\mathrm{df}_{\mathrm{GD}}(t)$ increases from 0 to $\mathrm{rank}(X)$ as t grows; test error first decreases (bias reduction), then increases (variance from small-σ_j directions), mirroring ridge.

Comparing filters,

$$g_j^{\mathrm{ridge}}(\lambda) = \frac{\sigma_j}{\sigma_j^2 + \lambda},$$

$$g_j^{\mathrm{GD}}(t) = \frac{1 - (1 - \eta\sigma_j^2)^t}{\sigma_j},$$

reveals their similarity. For moderate t,

$$1 - (1 - \eta\sigma_j^2)^t \approx 1 - e^{-\eta\sigma_j^2 t},$$

so $g_j^{\mathrm{GD}}(t)$ behaves qualitatively like a ridge filter with effective penalty $\lambda \approx 1/(\eta t)$. Both families damp the same small-σ_j directions; they differ mainly in how the filter is parametrized (continuous λ vs. discrete t, η).

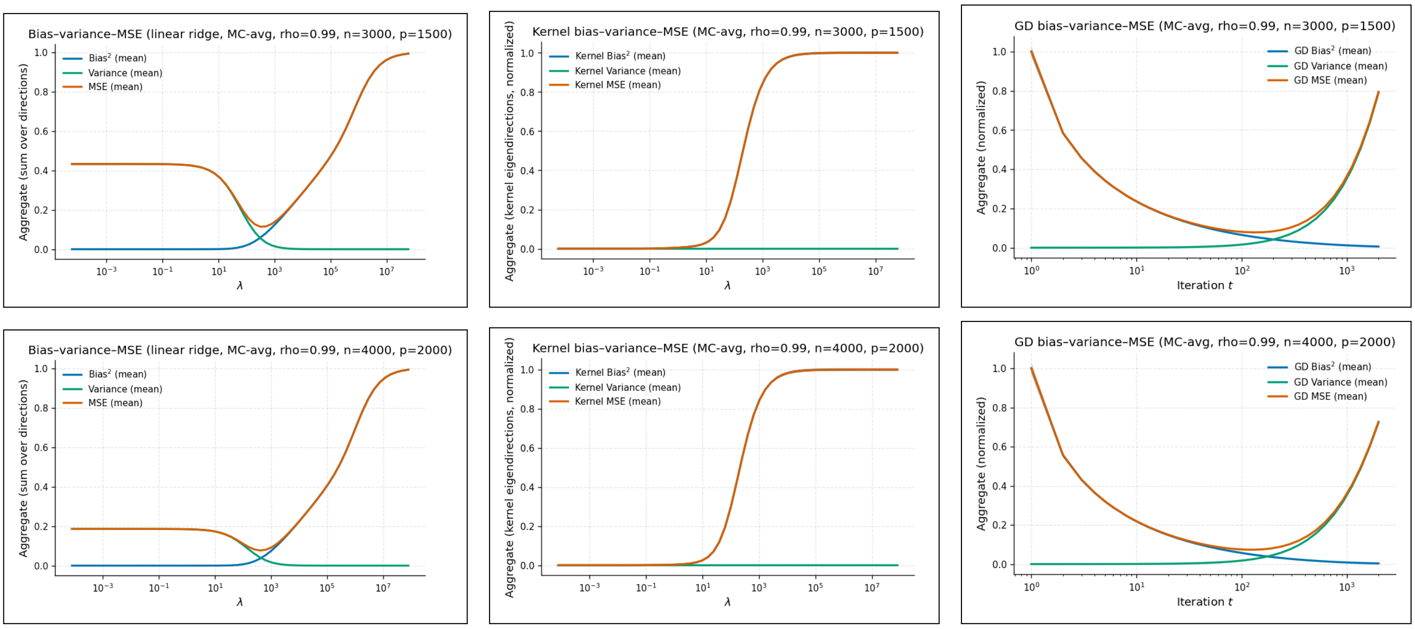


Figure 4: Spectral bias–variance trade-offs for ridge (left), kernel ridge with Nyström approximation (middle), and early-stopped gradient descent (right) in a highly correlated AR(1) design ($\rho = 0.99$).

Runtime Comparison

Figure 3 reports Monte Carlo–averaged runtimes as n increases, with $p = n/2$, SVD rank $k_{\text{top}} = 400$ and $m = 500$ Nyström landmarks. The randomized top-k SVD curve rises roughly quadratically in n (on the log–linear scale) because each pass costs $O(npq)$ with fixed $q \approx k_{\text{top}}$. The normal–equations solver scales as $O(np^2 + p^3) = O(n^3)$ and bends upward much more steeply as n approaches 10^4.

Nyström KRR builds an $n \times m$ feature matrix and performs an SVD of Z plus $O(m^3)$ operations; with fixed m this is $O(n^2)$ but with a smaller constant than full SVD or normal equations [11]. Gradient descent shares the same $O(n^2)$ scaling in our implementation, dominated by matrix–vector products needed to form y and y_{test}. Overall, normal equations are clearly cubic, whereas randomized SVD, Nyström KRR, and GD behave essentially quadratically.

Spectral Views of Regularization in the Simulations

Figure 4 summarizes the spectral behavior of the three regularization schemes in the ill-conditioned AR(1) design ($\rho = 0.99$). In the ridge panel, the filter $g_j^{\text{ridge}}(\lambda) = \sigma_j/(\sigma_j^2 + \lambda)$ yields the usual pattern: small λ produces low bias and high variance, large λ produces high bias and low variance, and the prediction MSE is minimized at an intermediate λ [3, 4].

The kernel ridge panel shows the same mechanism in the RKHS: filters $\kappa_i/(\kappa_i + \lambda)$ damp tail eigenvalues of the kernel operator, so only a small number of dominant eigendirections contribute significantly to the fit. The GD panel reflects the implicit filter $g_j^{\text{GD}}(t)$: for small t, the estimator is heavily shrunk toward zero; for large t, it approaches OLS in the singular basis and inherits its variance explosion in small-σ_j directions. In all three cases, the optimal regularization level lies strictly between the extremes of “no fit” and “unstable OLS”.

2 Data and Polynomial Feature Construction

We base our empirical study on the *BlogFeedback* dataset from the UCI Machine Learning Repository [1, 2]. The original dataset contains 52,397 blog posts, each described by 280 numerical attributes, and a response variable given by the number of comments received in the subsequent 24 hours. The raw response is highly right–skewed and heavy–tailed, so we apply the variance–stabilising transformation

$$y = \log\bigl(1 + \text{comments}\bigr),$$

and use y as the target in all subsequent regression analyses, following standard practice in regression modeling for heavy-tailed count data [4].

For the high–dimensional experiments we draw a random subsample of $n = 5{,}000$ posts and select $d = 60$ input variables from the 280 available predictors, yielding a working design matrix $X_{\text{sub}} \in \mathbb{R}^{n\times d}$. To introduce controlled nonlinearity and collinearity we expand these d base predictors using all degree–2 polynomial terms (without an explicit intercept). Writing $x = (x_1, \ldots, x_d)^\top$, we define

$$\phi(x) = \bigl(x_1, \ldots, x_d,\ x_1^2, \ldots, x_d^2,\ x_1x_2, \ldots, x_{d-1}x_d\bigr) \in \mathbb{R}^p,$$

and set $X_{\text{poly}} = \phi(X_{\text{sub}}) \in \mathbb{R}^{n\times p}$. For $d = 60$ the resulting dimension is

$$p = d + \frac{d(d+1)}{2} = 60 + \frac{60\cdot 61}{2} = 1{,}890,$$

so the model operates in an overparameterised regime with p of the same order as n. Prior to fitting any models, each column of X_{poly} is standardised to have zero mean and unit variance so that the ridge penalty acts isotropically in Euclidean coordinates [4].

2.1 Exploratory analysis of the BlogFeedback design

We first summarise the behaviour of the response and the selected predictors before introducing regularisation.

Target distribution. The raw response y_{raw} (number of comments in the next 24 hours) is extremely heavy–tailed and zero–inflated. On the full sample we observe

$$\text{min} = 0, \quad \text{max} = 1424, \quad \text{mean} \approx 6.8, \quad \text{sd} \approx 37.7,$$

with the median at 0 and the 99th percentile already above 150 comments. Figure 5a shows that most posts receive no or very few comments, while a small fraction receive hundreds. Applying the transformation $y = \log(1 + y_{\text{raw}})$ (Figure 5b) compresses the right tail and yields a more regular target: on the $n = 5{,}000$ subsample, min = 0, max $\approx$ 6.7, mean $\approx$ 0.67, and sd $\approx$ 1.14, with the upper quartile around $\log(1+3) \approx 1.10$. The mass at $y = 0$ is still visible, but the bulk of the distribution is now concentrated on a moderate scale, which explains why the transformed target is more amenable to least–squares methods [4].

At the covariate level, the $d = 60$ selected predictors are mostly nonnegative, heavily right–skewed, and only moderately correlated with the transformed response (absolute correlations up to about 0.44). Their

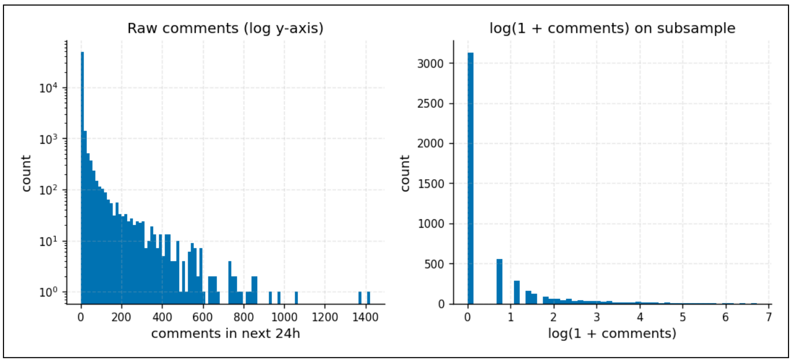


Figure 5: Distribution of the response before and after the $\log(1 + \cdot)$ transformation. The raw counts exhibit a long, sparse right tail; the transformed target is still skewed but much more compressed, preventing extreme observations from dominating the loss.

distributions and correlation structure indicate substantial but noisy signal and motivate the use of nonlinear feature maps. Detailed marginal summaries, the list of top y–correlated predictors, and correlation heatmaps are reported in Appendix B.

Effective rank and row norms of the polynomial design. On the standardised polynomial design $X_{\text{poly}} \in \mathbb{R}^{n\times p}$ we compute a truncated singular value decomposition. Approximately $k \approx 67$ singular directions capture 90% of the total variance, and $k \approx 260$ explain 99%. Hence, although $p = 1{,}890$ columns are present, the effective rank is much smaller; most directions in feature space carry negligible variance and are therefore susceptible to noise amplification.

Row norms of the standardised polynomial design are also highly variable: the mean ℓ_2 norm is around 18.4, but the maximum exceeds 2,000, with a standard deviation of about 36.9. These extreme rows correspond to posts where several base predictors jointly take large values, so that their polynomial combinations explode. Such high–leverage points further motivate the use of regularised estimators [4].

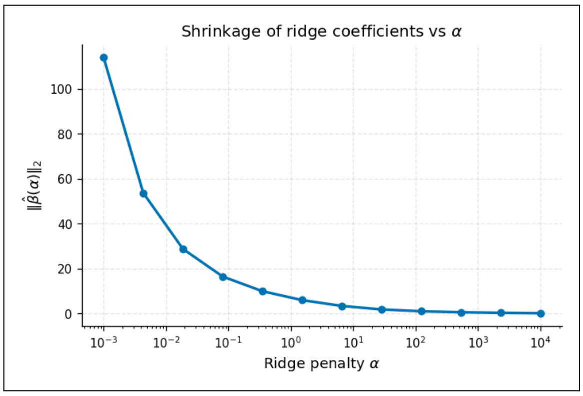


Figure 6: Euclidean norm of the ridge coefficient vector $\|\hat{\beta}(\alpha)\|_2$ along the regularisation path.

2.2 Ridge regression and spectral regularisation on BlogFeedback

We now fit ridge regression on the standardised polynomial design $X \in \mathbb{R}^{n\times p}$, $p = 1890$, with response $y_i = \log(1 + \text{comments}_i)$:

$$y = X\beta + \varepsilon, \qquad \hat{\beta}_\alpha = \arg\min_{\beta}\left\{\tfrac{1}{2}\|y - X\beta\|_2^2 + \tfrac{\alpha}{2}\|\beta\|_2^2\right\}.$$

Ridge path and shrinkage. Figure 7 shows 5-fold cross–validated training and validation MSE as a function of the ridge penalty α. For very small penalties ($\alpha \approx 10^{-3}$) the training error is extremely low but the validation error is huge, reflecting severe overfitting of the noisy, heavy–tailed target. As α increases, the validation curve drops rapidly: shrinking unstable directions reduces variance much faster than it increases bias. Beyond $\alpha \gtrsim 10^{-1}$ the curve is almost flat, and the CV–optimal penalty is at the upper end of the grid, $\alpha^\star \approx 10^4$, corresponding to a strongly regularised solution [4].

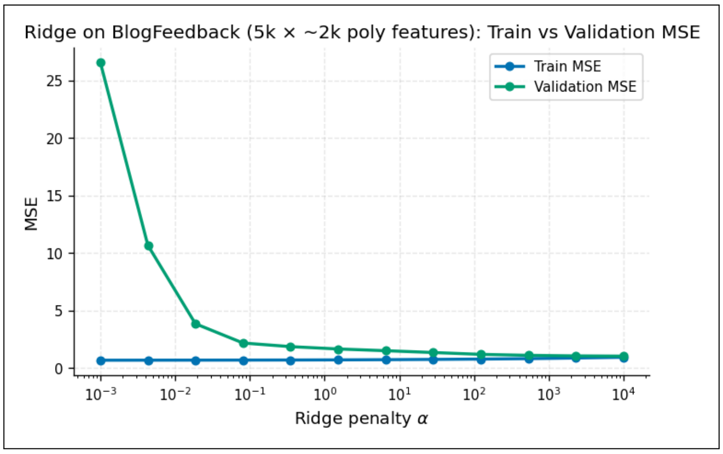


Figure 7: Training and validation MSE for ridge regression on the polynomial BlogFeedback design as a function of the penalty α (log scale). The dramatic drop in validation error for small α reflects the suppression of high–variance directions; the near–flat tail shows that once these are controlled, additional shrinkage has little effect.

The coefficient norm path in Figure 6 confirms this behaviour. The Euclidean norm $\|\hat{\beta}(\alpha)\|_2$ decreases by roughly two orders of magnitude as α moves from 10^{-3} to 10^4: for tiny α the ridge solution nearly inverts the ill–conditioned design and yields very large coefficients; as α grows, all directions are aggressively shrunk and the effective complexity of the model collapses.

Effect of α in data space. Scatter plots of $\hat{y}$ versus y for three representative penalties (Figure 9) illustrate the induced bias–variance trade–off. For a very small penalty ($\alpha = 10^{-3}$) the cloud is highly dispersed around the identity line, especially for larger y: the model chases rare high–comment posts and badly overfits. For a moderate penalty ($\alpha = 10^2$) the cloud tightens and the spread around the identity line shrinks, but the fitted line has slope < 1, so large responses are systematically underestimated. At the CV–optimal penalty ($\alpha^\star = 10^4$) predictions are strongly shrunk toward the global mean: variance is very low, but large responses are heavily biased downwards, a classic high–bias/low–variance regime [4].

This behaviour is summarised more compactly in Figure 8, which plots the regression lines from $\hat{y}$ onto y for OLS and several ridge penalties. Relative to the identity line, OLS has the steepest slope and thus the largest response to variation in y, but also the highest variance. As α increases, the slope of the ridge line decreases and the intercept increases, pulling all predictions toward the mean and visibly flattening the line; the motion of these lines encodes the bias–variance trade–off in prediction space.

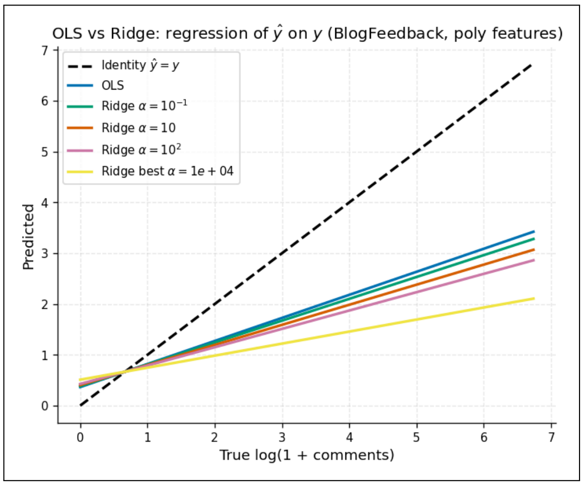
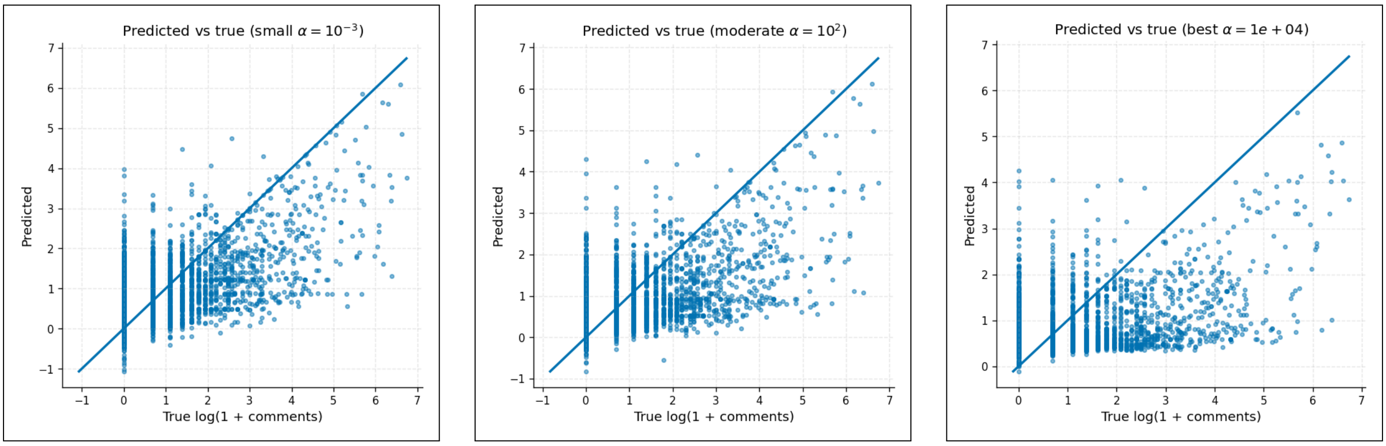


Figure 8: Regression of $\hat{y}$ on y for OLS and several ridge penalties. The dashed line is the identity $\hat{y} = y$. Increasing α moves the ridge lines from a high–variance, steep response (close to OLS) toward flatter, more biased fits concentrated near the global mean.

Spectral viewpoint. Let $X_{\text{std}} = U_k \Sigma_k V_k^\top$ denote the truncated SVD of the standardised polynomial design using the top $k = 500$ singular directions. In this basis, OLS applies the inverse filter $1/\sigma_j$ to each singular value σ_j, which explodes along small singular values and amplifies noise. Ridge instead uses the



(a) Small penalty $\alpha = 10^{-3}$ (b) Moderate penalty $\alpha = 10^{2}$ (c) CV–optimal penalty $\alpha = 10^{4}$

Figure 9: Predicted versus true $\log(1 + \text{comments})$ for three values of the ridge penalty α; the 45° line is shown for reference. As α increases, the point cloud contracts toward the line but also flattens, reflecting lower variance and higher bias.

spectral filter

$$g_j(\alpha) = \frac{\sigma_j}{\sigma_j^2 + \alpha},$$

which behaves like $1/\sigma_j$ when $\sigma_j^2 \gg \alpha$ but damps directions with $\sigma_j^2 \ll \alpha$.

Figure 10 displays $g_j(\alpha)$ for several penalties together with the OLS filter $1/\sigma_j$. For moderate penalties, the leading singular directions (large σ_j) are almost unchanged, while the trailing ~ 200 directions are strongly shrunk. Under the CV–optimal penalty, all 500 directions are heavily damped; the ridge filter curve lies well below the OLS curve for every j, explaining the strong shrinkage seen in coefficient norms and prediction plots.

Bias–variance curves for SVD, Nyström, and GD. We next compare explicit spectral regularisation with an implicit one by estimating bias^2, variance, and MSE via Monte Carlo over random train/test splits.

For truncated–SVD ridge (Figure 11a) the normalised bias^2, variance, and MSE all drop sharply as α increases from 10^{-3} to roughly 10^{-2}: damping the most ill–conditioned directions immediately reduces test error. Beyond this point, all three curves are essentially flat and close to zero, indicating that once the unstable tail of the spectrum is controlled, further increases in α only have a second–order effect on performance.

For Nyström–RBF features with $m = 500$ landmarks (Figure 11b), the variance component is small across the entire range of α, while the bias term dominates the MSE. The MSE curve exhibits a shallow minimum at intermediate penalties: very small α retains slightly more variance, whereas very large α oversmooths and misses signal in the kernel feature space. The slow movement of these curves reflects the strong regularising effect of the RBF mapping itself [4, 11].

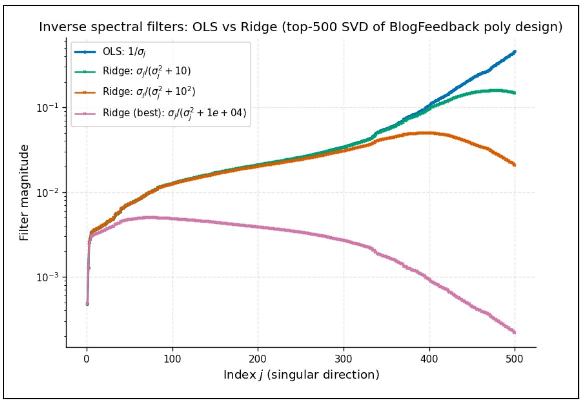
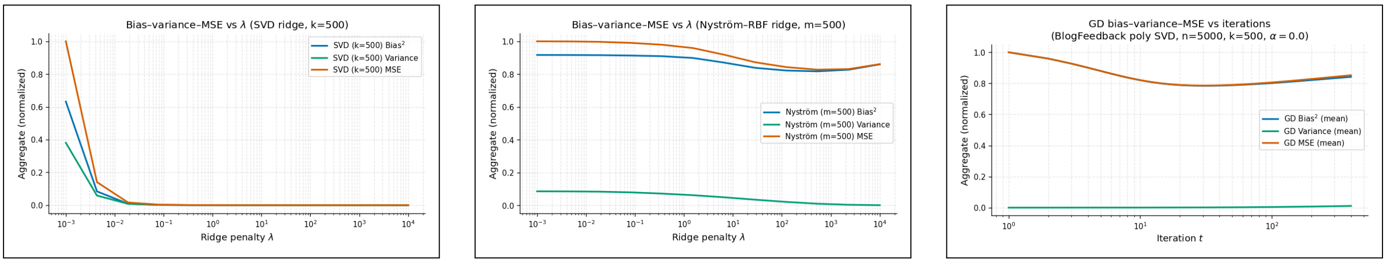


Figure 10: Inverse spectral filters for OLS and ridge regression applied to the top 500 singular directions of the standardised polynomial design. Moving from OLS to large α, the filter curves tilt downward and flatten, showing how ridge progressively suppresses low–variance directions.



(a) Normalised bias^2, variance, and MSE for ridge in the top-500 SVD subspace as a function of λ.

(b) Normalised bias^2, variance, and MSE for Nyström–RBF ridge with $m = 500$ features as a function of λ.

(c) Normalised bias^2, variance, and MSE of GD in the $k = 500$ SVD subspace as a function of t.

Figure 11: Normalised bias^2, variance, and MSE for (a) ridge in the SVD subspace, (b) Nyström–RBF ridge, and (c) gradient descent, showing the U-shaped test error as a function of the regularisation parameter (λ or iteration t) in each spectral setting.

Finally, gradient descent run in the $k = 500$ SVD subspace with $\alpha = 0$ (Figure 11c) exhibits implicit regularisation through early stopping. In the first few dozen iterations the bias and MSE curves fall rapidly as the algorithm fits the dominant singular directions. After this “sweet spot” the curves flatten, and for larger t the MSE slowly rises as the algorithm begins to fit noise in smaller singular directions. The U–shaped MSE trajectory as a function of t mirrors the U–shape in α for explicit ridge: iteration count acts as an implicit regularisation parameter [4].

3 Conclusion

This project has treated regression explicitly as an inverse problem on a Hilbert space and used the singular–value spectrum as the primary lens for understanding stability, generalisation, and computational cost. In both the synthetic Toeplitz AR(1) designs and the polynomially expanded BlogFeedback design, the same phenomenon appears: the spectrum is highly concentrated, the effective rank is much smaller than the ambient dimension, and naive inversion amplifies noise in small–variance directions. The OLS inverse $1/\sigma_j$ is therefore intrinsically unstable in these regimes, and the heavy tails and high–leverage points in the data make this instability visible as exploding variance and poor test performance.

Against this backdrop, all three regularisation schemes studied here can be understood as spectral filters that replace unbounded inversion by bounded, direction–dependent shrinkage. Linear ridge applies $g_j(\lambda) = \sigma_j/(\sigma_j^2 + \lambda)$ in the SVD basis; Nyström kernel ridge applies the analogous filter $\kappa_i/(\kappa_i + \lambda)$ to approximate kernel eigenvalues; and early–stopped gradient descent realises an implicit filter $[1 - (1 - \eta\sigma_j^2)^t]/\sigma_j$ whose effective penalty is controlled by the iteration count. Despite their very different algorithmic implementations, the bias–variance curves and degrees–of–freedom trajectories for these methods exhibit the same qualitative pattern: strong regularisation heavily shrinks all directions (high bias, low variance), weak regularisation reverts to an OLS–like inverse (low bias, high variance), and test MSE is minimised at an intermediate point where only the ill–conditioned spectral tail is damped.

The runtime experiments further highlight that once one adopts a spectral point of view, there is no reason to privilege cubic–time normal equations over more scalable spectral or iterative schemes. Randomised truncated SVD, Nyström features, and gradient descent all achieve essentially quadratic scaling in n while providing fine control over which spectral directions are retained or suppressed. Taken together, the results suggest a practical message: in modern overparameterised, collinear, and heavy–tailed settings, regression procedures should be designed and interpreted as *spectral filters*. Making the filter explicit—whether through ridge, kernel methods, or early stopping—not only stabilises estimation and improves prediction, but also offers a clear, operator–theoretic language for comparing algorithms and reasoning about their behaviour beyond black–box performance curves.

4 Limitations

While the experiments provide a coherent spectral narrative, the study has several limitations that constrain the scope of the conclusions.

Designs and datasets. On the synthetic side, the analysis focuses on Gaussian AR(1) Toeplitz covariances with a small grid of correlation parameters ρ and a particular choice of p/n scaling. This is a very structured and well-behaved ensemble: eigenvectors are close to a fixed Fourier-like basis, and the eigenvalues follow a smooth decay pattern. Real-world designs frequently exhibit block structure, heteroskedasticity, heavy tails, or adversarial collinearity that are not captured by this AR(1) family. On the empirical side, the BlogFeedback dataset is only one example, with a single choice of subsampling scheme, predictor subset, and degree-2 polynomial map. Thus, while the results are consistent with the inverse-problem perspective, they should be interpreted as illustrative rather than representative of all high-dimensional regression problems.

Modeling assumptions and evaluation metrics. The theoretical and simulation analyses rely heavily on homoscedastic Gaussian noise, a fixed "smooth" signal aligned with the leading singular vectors, and squared-error loss as the primary performance metric. These assumptions simplify the spectral derivations but underplay phenomena such as outliers, heavy-tailed errors, covariate shift, or signals that live partly in low-variance directions. Moreover, the study emphasises MSE and effective degrees of freedom; it does not investigate robustness, calibration, prediction intervals, or task-specific metrics (e.g., ranking quality or tail prediction for rare, high-comment posts). As a consequence, the practical implications for non-Gaussian, misspecified, or distributionally shifted settings remain largely unexplored.

Algorithmic scope and hyperparameter handling. Algorithmically, the project concentrates on ridge-type quadratic penalties, Nyström-approximated kernel ridge, and full-batch gradient descent with a simple step-size schedule. Other important regularisation mechanisms— ℓ_1 penalties and sparsity, elastic nets, early stopping in stochastic gradient methods, or modern deep architectures—are not examined within the same spectral framework. In addition, hyperparameters (e.g., the RBF lengthscale, Nyström landmark count, SVD truncation level, and GD step size) are chosen via fixed heuristics or coarse grids rather than a systematic search or nested cross-validation. This limits the ability to claim hyperparameter-robust conclusions and leaves open how sensitive the spectral patterns are to these design choices.

5 Extensions and Future Work

Sharper operator-theoretic and high-dimensional analysis. On the infinite-dimensional side, the present work uses the spectral theorem and Tikhonov regularisation mainly at a conceptual level. A natural extension is to impose standard source conditions on the true regression function (e.g. $f^\star$ lying in a range space of powers of the operator T) and to derive nonasymptotic or asymptotic convergence rates for Tikhonov under various eigenvalue decay regimes. In parallel, on the finite-sample side one could move from fixed-p/n AR(1) designs to high-dimensional asymptotics with $p, n \to \infty$ and $p/n \to \gamma \in (0, \infty)$, leveraging random matrix theory to obtain analytic predictions for test MSE, effective degrees of freedom, and the onset of double-descent behaviour. This would connect the empirical curves in the project to a more rigorous operator-theoretic and random-matrix framework.

Implicit regularisation and a broader catalogue of spectral filters. The project illustrates explicit regularisation via ridge and implicit regularisation via early-stopped gradient descent, but many optimisation procedures (heavy-ball and Nesterov momentum, conjugate gradients, coordinate descent, SGD) can also be understood as inducing specific spectral filters $g_j(\cdot)$ as a function of iteration. Extending the analysis to these methods, deriving their induced filters in the SVD or Nyström eigenspaces, and comparing their bias–variance and runtime profiles to ridge on AR(1) and BlogFeedback designs would yield a more complete picture of *algorithm-dependent* regularisation. In parallel, one could compare ridge to other explicit spectral schemes (truncated SVD, elastic net, or tailored filter families) to build a small "catalogue" of regularisers whose behaviour is interpretable directly from the spectrum.

Robustness, feature design, and broader benchmarks. The BlogFeedback analysis reveals heavy tails in both the response and the row norms of the polynomial design, suggesting that squared loss with pure ℓ_2 penalties may not be robust to high-leverage points. A natural extension is to investigate robust losses (Huber, Tukey) or leverage-trimming combined with ridge, and to characterise their impact in the spectral domain. On the feature side, one could systematically vary the feature map—degrees of polynomial expansion, spline or wavelet bases, random Fourier features versus Nyström for the same kernel—and study how each choice reshapes the spectrum and interacts with the chosen filter. Replicating the full spectral analysis on additional real datasets with different correlation and tail behaviour (e.g., image or text features) would test how broadly the "spectral filter" viewpoint extends beyond the specific BlogFeedback and AR(1) settings considered here.

References

[1] Krisztián Buza. Blogfeedback data set. UCI Machine Learning Repository, 2014. Available at the UCI Machine Learning Repository.

[2] Krisztián Buza. Feedback prediction for blogs. In *Discovery Challenge, ECML PKDD 2014*, 2014. Source of the BlogFeedback data set used in our experiments.

[3] Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*, volume 375 of *Mathematics and Its Applications*. Kluwer Academic Publishers, Dordrecht, 1996.

[4] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2 edition, 2009.

[5] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[6] Andreas Kirsch. *An Introduction to the Mathematical Theory of Inverse Problems*, volume 120 of *Applied Mathematical Sciences*. Springer, New York, 2 edition, 2011.

[7] MIT OpenCourseWare. Inverse problems. Course notes, 2010. Lecture notes for 18.325, Topics in Applied Mathematics.

[8] Dietmar A. Salamon. Functional analysis. Lecture notes, 2019. Available from the author's webpage; accessed 2024.

[9] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-Posed Problems*. Winston and Sons, Washington, DC, 1977.

[10] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[11] Christopher K. I. Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13 (NIPS)*, pages 682–688, 2001.

Appendix A: Bias–Variance Trade-off and Effective Degrees of Freedom

Express $\beta^\star = \sum_j \beta_j^\star v_j$ in the right-singular basis and suppose $\varepsilon \sim N(0, \sigma_\varepsilon^2 I_n)$. A standard calculation shows that for each j,

$$\mathrm{Bias}_j(\lambda) = -\frac{\lambda}{\sigma_j^2 + \lambda}\beta_j^\star, \qquad \mathrm{Var}_j(\lambda) = \sigma_\varepsilon^2 \frac{\sigma_j^2}{(\sigma_j^2 + \lambda)^2},$$

and hence

$$\mathrm{MSE}_j(\lambda) = \mathrm{Bias}_j(\lambda)^2 + \mathrm{Var}_j(\lambda).$$

Aggregating over j yields total squared bias, variance, and MSE:

$$\mathrm{Bias}^2(\lambda) = \sum_j \mathrm{Bias}_j(\lambda)^2, \quad \mathrm{Var}(\lambda) = \sum_j \mathrm{Var}_j(\lambda), \quad \mathrm{MSE}(\lambda) = \sum_j \mathrm{MSE}_j(\lambda).$$

Ridge also defines a linear smoother on y with smoothing factors $\sigma_j^2/(\sigma_j^2 + \lambda)$ along each u_j, and effective degrees of freedom

$$\mathrm{df}_{\mathrm{ridge}}(\lambda) = \sum_j \frac{\sigma_j^2}{\sigma_j^2 + \lambda}.$$

As λ increases, $\mathrm{df}_{\mathrm{ridge}}(\lambda)$ decreases and the estimator moves from low-bias/high-variance (near OLS) to high-bias/low-variance [4].

For gradient descent with step size η and initialization $\beta^{(0)} = 0$, the tth iterate has spectral filter

$$g_j^{\mathrm{GD}}(t) = \frac{1 - (1 - \eta\sigma_j^2)^t}{\sigma_j},$$

and the fitted values are

$$\hat{y}^{(t)} = \sum_j s_j^{\mathrm{GD}}(t)\, \alpha_j\, u_j, \qquad s_j^{\mathrm{GD}}(t) = 1 - (1 - \eta\sigma_j^2)^t.$$

The effective degrees of freedom for GD are

$$\mathrm{df}_{\mathrm{GD}}(t) = \sum_j \left[1 - (1 - \eta\sigma_j^2)^t\right],$$

which increase from 0 to $\mathrm{rank}(X)$ as t grows. As $t \to \infty$, $s_j^{\mathrm{GD}}(t) \to 1$ and GD converges to the OLS solution in the singular subspace [3, 4].

Appendix B: Additional BlogFeedback Exploratory Analysis and Runtimes

Behaviour of the base predictors. For the $d = 60$ selected predictors we compute basic marginal summaries and their Pearson correlation with the transformed target y (Table 1). The ten most y–correlated features have absolute correlations in the range 0.19–0.44: the signal is clearly present but far from deterministic.

Table 1: Top base predictors ranked by $|\mathrm{corr}(x_j, y)|$ on the $n = 5000$ subsample. “local idx” denotes the index within the chosen 60-dimensional subset, and “orig col” the corresponding column in the original 280-dimensional design.

local idx	orig col	mean	sd	frac zero	skew	corr_y
50	3	346.15	446.72	0.04	2.06	0.44
33	10	14.25	28.85	0.04	4.70	0.44
35	6	28.38	38.70	0.04	2.88	0.44
22	16	42.11	52.73	0.05	2.06	0.43
40	5	15.45	32.57	0.04	5.00	0.43
28	18	290.64	377.91	0.05	2.10	0.40
59	13	260.47	322.19	0.04	1.86	0.40
26	23	255.98	320.92	0.04	1.88	0.40
4	19	22.25	61.98	0.31	5.31	0.20
19	22	-229.26	268.40	0.04	-1.52	-0.19

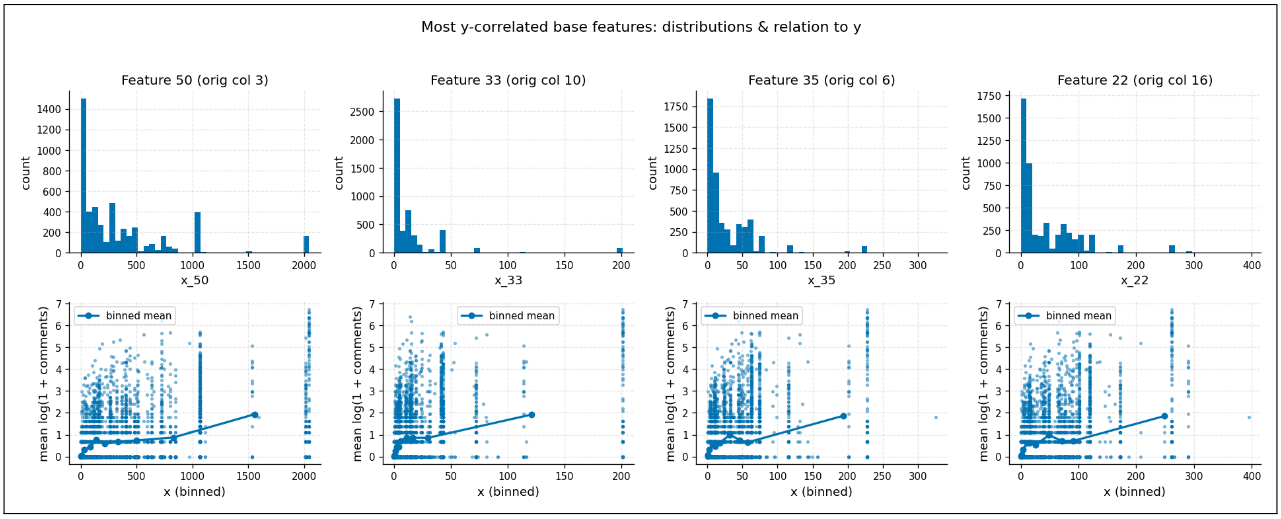


Figure 12: Top row: empirical distributions of the four most y-correlated base features, all of which are nonnegative and heavily right–skewed. Bottom row: binned conditional means of $\log(1 + \text{comments})$; the upward trends with large scatter explain why individual features are only moderately predictive and motivate richer feature maps.

Most of these covariates are nonnegative and strongly right–skewed with a nontrivial fraction of zeros (typically 3–7%, and up to 31% for one feature), followed by a long tail of large values. The top panels of Figure 12 show this pattern. The bottom panels display binned conditional means of y given each predictor: the average $\log(1 + \text{comments})$ increases with the predictor, but the vertical scatter within each bin remains large. This indicates that the important covariates are informative but noisy and that nonlinear and interaction effects are plausible, motivating the polynomial expansion [4].

Correlation structure and multicollinearity. To assess linear dependence, we inspect the sample correlation matrix of the first 20 standardized base features (Figure 13) and boxplots of the ten most y–correlated predictors (Figure 14). Most off–diagonal entries lie between roughly -0.2 and 0.6, with a few stronger positive or negative relationships. Thus the original 60-dimensional design is moderately, but not pathologically, collinear. After polynomial expansion, these dependencies are amplified: squares and pairwise products of already correlated variables generate many nearly redundant columns, which foreshadows a spectrum with many very small singular values and an ill–conditioned least–squares problem [4].

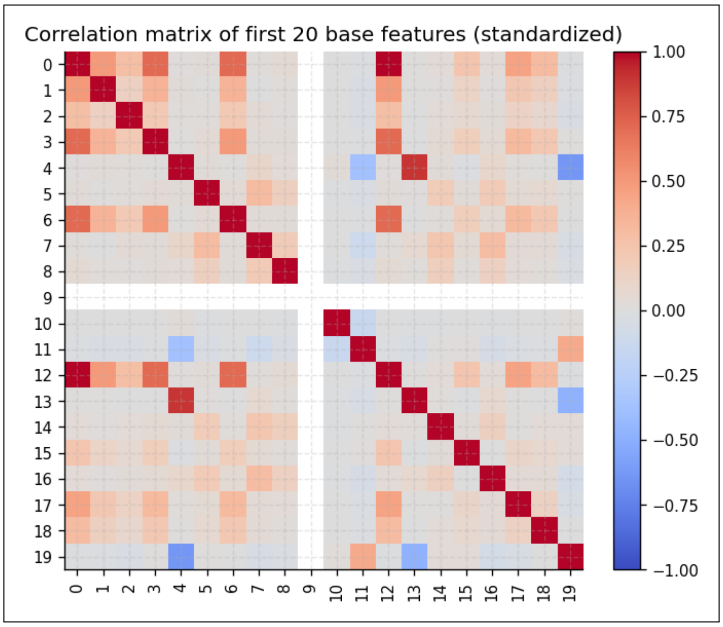


Figure 13: Sample correlation matrix of the first 20 standardized base predictors in X_{sub}. Blocks of moderate positive correlation and a few strong negative pairs show that the design already has substantial dependence before polynomial expansion.

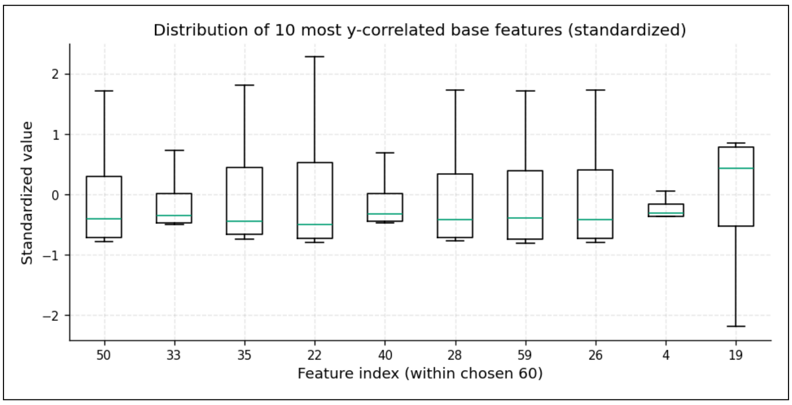


Figure 14: Boxplots of the standardized values of the ten most y-correlated base predictors. The heterogeneous spreads and occasional outliers indicate heterogeneous scales and the presence of high–leverage observations even after standardisation.

Runtime comparison (BlogFeedback). Table 2 reports average training times (over three runs) for four solvers on the polynomial design. For truncated–SVD ridge and Nyström ridge, the one–off cost of computing the global SVD and Nyström features is amortised and not included.

Method	Description	Avg. runtime (s)
Normal equations (poly)	Solve $(X^\top X + \alpha I)\beta = X^\top y$ in $\mathbb{R}^{1890}$	0.050
Truncated SVD ridge (poly, $k = 500$)	Closed form in rank-k SVD subspace	0.001
Gradient descent (poly, $k = 500$)	GD in SVD subspace (200 iterations)	0.109
Nyström–RBF ridge ($m = 500$)	Normal equations in 500-D Nyström space	0.002

Table 2: Average training time of the four ridge solvers on the polynomial BlogFeedback design (excluding one–off SVD and Nyström precomputation). Truncated SVD and Nyström achieve speeds comparable to low–dimensional regression, while GD is dominated by repeated passes over the data.