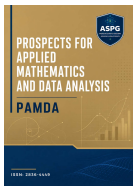
Prospects for Applied Mathematics and Data Analysis

♦ ISSN: 2836-4449 Vol: 06, No. 01, 2026 ♦ PP. 29–35
♦ Article DOI: https://doi.org/10.54216/PAMDA.060105




FULL LENGTH ARTICLE

# Exact Bias–Variance Decomposition and Degrees of Freedom in Ridge Regression: Theory, Verification, and a Disease-Progression Application

**Dwi Agustin Retnowardani**[1,*]

[1] Universitas PGRI Argopuro Jember, Indonesia

**Email:** 2i.agustin@mail.unipar.ac.id

Received: September 21, 2025 Revised: November 18, 2025 Accepted: January 08, 2026 ⋆ Corresponding author

## ABSTRACT

Ridge regression estimates $\beta$ in the linear model $y = X\beta + \varepsilon$ by $\hat{\beta}(\lambda) = \arg\min_\beta \|y - X\beta\|^2 + \lambda\|\beta\|^2$, trading bias for variance as $\lambda$ increases. This paper collects six results about $\hat{\beta}(\lambda)$ into a single self-contained development, each proved and then checked numerically. The estimator is written in closed form through the singular value decomposition of $X$; its effective degrees of freedom, $\mathrm{df}(\lambda) = \sum_j d_j^2/(d_j^2 + \lambda)$, are shown to be strictly decreasing and convex in $\lambda$; its exact bias and variance are derived in closed form; a strictly positive $\lambda$ is shown always to exist that reduces mean squared estimation error below that of ordinary least squares whenever the noise variance is positive; the estimator is shown to coincide with the posterior mean under a Gaussian prior with precision proportional to $\lambda$; and the leave-one-out cross-validation error is shown to admit a closed-form shortcut that generalized cross-validation approximates by averaging its leverage terms. Every derived quantity is verified against data: the leave-one-out shortcut matches brute-force refitting exactly, and a calibrated Monte Carlo simulation confirms the closed-form bias and variance to within simulation error at every tested $\lambda$. Applied to a standard diabetes disease-progression dataset ($n = 442$, ten predictors), the theoretical construction correctly locates a strictly risk-reducing regularization region, and repeated cross-validation shows ridge, lasso, and elastic net all lying within one standard error of ordinary least squares in out-of-sample prediction error—consistent with the closed-form theory, which attributes the available gain to reduced parameter-estimation risk on a well-conditioned design rather than to prediction-error reduction.

**Keywords:** Ridge regression ▪ Bias–variance decomposition ▪ Degrees of freedom ▪ Generalized cross-validation ▪ Bayesian regularization ▪ Singular value decomposition

## 1. INTRODUCTION

Ridge regression regularizes ordinary least squares by adding an $\ell_2$ penalty on the coefficient vector, trading an increase in bias for a reduction in variance. The idea is old, but the precise mathematical structure of that trade-off – how bias and variance depend on the regularization parameter, when a strictly positive parameter is guaranteed to help, what the resulting estimator's effective complexity is, and how that complexity should be chosen from data – is still an active area of research, particularly as the classical fixed-dimension, large-sample regime gives way to high-dimensional and even proportional asymptotics in modern applications [1, 2, 3]. Recent work has substantially sharpened this picture: the risk of ridge regression and its interpolating (zero-penalty) limit is now understood precisely in high-dimensional random-design settings [2, 4, 5], the choice of regularization by cross-validation has been shown to be asymptotically consistent

under increasingly general conditions [6, 7], with related extensions to sketched ridge ensembles [8], and the classical intuition that more regularization is always safer has been shown to fail in specific high-dimensional regimes [9].

This paper takes a different, complementary approach: rather than pursuing new asymptotic regimes, it assembles the exact, finite-sample mathematical structure of ridge regression – closed-form estimator, degrees of freedom, bias, variance, risk-improvement guarantee, Bayesian equivalence, and cross-validation shortcut – into one coherent development with complete proofs, and then verifies every piece of that structure numerically against a real design matrix, rather than only under simulation or asymptotic approximation. The same finite-sample structure underlies a growing body of work on ridge-based ensembles and distributed computation: bagged and subsampled ridge predictors have been shown to admit equivalences with a single, differently regularized ridge fit [10, 11], risk-monotonization corrections have been proposed for ensembles whose naive risk curves are non-monotonic in ensemble size [12], and one-shot distributed ridge regression across machines has been analyzed by averaging exactly the kind of per-machine bias and variance terms derived in Section 5 [13]. Adlam and Pennington [14] showed that even the double-descent risk curves of heavily overparameterized models require exactly this kind of fine-grained bias-variance decomposition, rather than a single aggregate risk number, to be understood correctly – the same principle that motivates keeping bias and variance separate, rather than only reporting their sum, throughout this paper. The verification is deliberately exact where the mathematics permits it (the leave-one-out shortcut is checked against brute-force refitting to floating-point precision) and simulation-based where finite-sample exactness is unavailable in closed form (the bias-variance decomposition is checked by a calibrated Monte Carlo experiment using the real design matrix). The application throughout is a widely used diabetes disease-progression dataset [15], on which the paper also positions ridge regression against ordinary least squares, the lasso, and the elastic net using repeated cross-validation, connecting the closed-form theory to applied comparisons of the kind reported in the applied-statistics literature [16, 17].

Section 2 fixes notation and describes the data. Sections 3–8 develop the six results in turn, each as a theorem or lemma with proof. Section 9 reports the numerical verification and the empirical application. Section 10 discusses the findings, and Section 11 concludes.

## 2. DATA AND NOTATION

The application throughout uses the diabetes dataset of [15]: for $n = 442$ patients, ten baseline variables (age, sex, body mass index, average blood pressure, and six blood serum measurements) are recorded as predictors, and the response is a quantitative measure of disease progression one year after baseline. Each predictor column is mean-centered and scaled so that its sum of squares equals one, the standard preprocessing used since the variables' original introduction; the response is mean-centered before all fitting reported below, and predictions are reported on the response's original scale where relevant. Write $X \in \mathbb{R}^{n\times p}$ ($n = 442$, $p = 10$) for the resulting design matrix and $y \in \mathbb{R}^n$ for the centered response, and assume the linear model

$$y = X\beta + \varepsilon, \qquad \mathbb{E}[\varepsilon] = 0, \quad \mathrm{Cov}(\varepsilon) = \sigma^2 I_n, \tag{1}$$

with $X$ treated as fixed (the classical, or "Fixed-X", assumption; see [18] for a detailed treatment of the alternative, "Random-X", assumption and its consequences for bias-variance decompositions of exactly this kind). The ridge estimator is

$$\hat\beta(\lambda) = \arg\min_{\beta\in\mathbb{R}^p} \|y - X\beta\|^2 + \lambda\|\beta\|^2, \qquad \lambda \ge 0, \tag{2}$$

which reduces to the ordinary least-squares estimator at $\lambda = 0$ (assuming $X$ has full column rank, which holds here since $n \gg p$).

## 3. THE RIDGE ESTIMATOR IN CLOSED FORM

**Theorem 1** (Singular value representation)**.** *Let $X = UDV^\top$ be the (thin) singular value decomposition of $X$, with $U \in \mathbb{R}^{n\times p}$ and $V \in \mathbb{R}^{p\times p}$ orthonormal and $D = \mathrm{diag}(d_1, \dots, d_p)$, $d_1 \ge \cdots \ge d_p \ge 0$. Then, for every $\lambda \ge 0$,*

$$\begin{aligned} \hat\beta(\lambda) &= V\,\mathrm{diag}\left(\frac{d_j}{d_j^2+\lambda}\right)U^\top y, \\ \hat y(\lambda) &= X\hat\beta(\lambda) = U\,\mathrm{diag}\left(\frac{d_j^2}{d_j^2+\lambda}\right)U^\top y. \end{aligned} \tag{3}$$

*Proof.* Minimizing Equation 2 gives the ridge normal equations $(X^\top X + \lambda I)\hat\beta(\lambda) = X^\top y$. Substituting $X = UDV^\top$, $X^\top X = VD^2V^\top$, so $(X^\top X + \lambda I) = V(D^2 + \lambda I)V^\top$, which is invertible for every $\lambda > 0$ (and for $\lambda = 0$ under the full-rank assumption). Hence $\hat\beta(\lambda) = V(D^2+\lambda I)^{-1}V^\top X^\top y = V(D^2+\lambda I)^{-1}DU^\top y$, which is exactly the first expression in Equation 3 written entrywise; substituting into $X\hat\beta(\lambda) = UDV^\top\hat\beta(\lambda)$ gives the second. □

Equation 3 shows that ridge regression shrinks each coordinate of $y$ in the orthonormal basis $U$ by a factor $d_j^2/(d_j^2+\lambda) \in (0, 1]$, with directions of small singular value $d_j$ shrunk more aggressively than directions of large $d_j$ – the mechanism underlying every subsequent result in this paper.

## 4. EFFECTIVE DEGREES OF FREEDOM

**Definition 1.** *The effective degrees of freedom of the ridge fit at penalty $\lambda$ are* $\mathrm{df}(\lambda) = \mathrm{tr}\big(X(X^\top X + \lambda I)^{-1}X^\top\big)$.

**Theorem 2.** $\mathrm{df}(\lambda) = \sum_{j=1}^p \dfrac{d_j^2}{d_j^2+\lambda}$, *with* $\mathrm{df}(0) = p$ *(assuming $d_j > 0$ for all $j$) and* $\mathrm{df}(\lambda) \to 0$ *as* $\lambda \to \infty$.

*Proof.* From Theorem 1, the hat matrix is $S(\lambda) = X(X^\top X + \lambda I)^{-1}X^\top = U\,\mathrm{diag}(d_j^2/(d_j^2+\lambda))U^\top$, whose trace, using the cyclic property of the trace and $U^\top U = I_p$, is $\mathrm{tr}(S(\lambda)) = \sum_j d_j^2/(d_j^2+\lambda)$. The boundary values follow by direct substitution. □

**Lemma 1** (Monotonicity and convexity)**.** *For $d_j > 0$, each term $h_j(\lambda) = d_j^2/(d_j^2+\lambda)$ is strictly decreasing and strictly*

*convex on $\lambda \geq 0$; consequently* $\mathrm{df}(\lambda)$ *is strictly decreasing and convex on* $\lambda \geq 0$.

*Proof.* Direct differentiation gives $h_j'(\lambda) = -d_j^2/(d_j^2+\lambda)^2 < 0$ and $h_j''(\lambda) = 2d_j^2/(d_j^2+\lambda)^3 > 0$ for all $\lambda \geq 0$ and $d_j > 0$. Sums of strictly decreasing (respectively, strictly convex) functions are strictly decreasing (respectively, strictly convex), giving the claim for $\mathrm{df}(\lambda) = \sum_j h_j(\lambda)$. □

Lemma 1 guarantees that $\mathrm{df}(\lambda)$ traces out a smooth, strictly decreasing, convex curve from $p$ down to 0 as $\lambda$ runs from 0 to $\infty$, so that every value in $(0, p)$ is attained by exactly one $\lambda$, making df a valid, monotonically invertible reparameterization of the penalty – the basis on which the empirical results in Section 9 report regularization strength.

## 5. EXACT BIAS–VARIANCE DECOMPOSITION

**Theorem 3.** *Under model 1, the ridge estimator has, for every* $\lambda \geq 0$,

$$\mathrm{Bias}(\hat{\beta}(\lambda)) = \mathbb{E}[\hat{\beta}(\lambda)] - \beta = -\lambda V(D^2+\lambda I)^{-1}V^\top \beta, \quad (4)$$

$$\mathrm{Var}(\hat{\beta}(\lambda)) = \sigma^2 V \,\mathrm{diag}\left(\frac{d_j^2}{(d_j^2+\lambda)^2}\right) V^\top, \quad (5)$$

*and writing* $\gamma = V^\top \beta$, *the total mean squared estimation error is*

$$\begin{aligned}\mathrm{MSE}(\lambda) &= \|\mathrm{Bias}(\hat{\beta}(\lambda))\|^2 + \mathrm{tr}\big(\mathrm{Var}(\hat{\beta}(\lambda))\big) \\ &= \sum_{j=1}^{p} \frac{\lambda^2\gamma_j^2 + \sigma^2 d_j^2}{(d_j^2+\lambda)^2}. \end{aligned} \quad (6)$$

*Proof.* From Theorem 1, $\hat{\beta}(\lambda) = V(D^2+\lambda I)^{-1}DU^\top y$, and under model 1, $U^\top y = U^\top X\beta + U^\top \varepsilon = DV^\top \beta + U^\top \varepsilon$. Substituting,

$$\hat{\beta}(\lambda) = V(D^2+\lambda I)^{-1}D^2V^\top \beta + V(D^2+\lambda I)^{-1}DU^\top \varepsilon.$$

Taking expectations, the second term vanishes ($\mathbb{E}[\varepsilon] = 0$), and $(D^2+\lambda I)^{-1}D^2 = I - \lambda(D^2+\lambda I)^{-1}$ entrywise, so $\mathbb{E}[\hat{\beta}(\lambda)] = \beta - \lambda V(D^2+\lambda I)^{-1}V^\top \beta$, giving Equation 4. For the variance, since $\mathrm{Cov}(\varepsilon) = \sigma^2 I_n$ and $U^\top U = I_p$,

$$\begin{aligned}\mathrm{Var}(\hat{\beta}(\lambda)) &= V(D^2+\lambda I)^{-1}D\,\mathrm{Cov}(U^\top\varepsilon)\,D(D^2+\lambda I)^{-1}V^\top \\ &= \sigma^2 V(D^2+\lambda I)^{-1}D^2(D^2+\lambda I)^{-1}V^\top, \end{aligned}$$

which is Equation 5 written entrywise. Equation 6 follows because $V$ is orthonormal, so $\|\mathrm{Bias}\|^2 = \lambda^2\sum_j \gamma_j^2/(d_j^2+\lambda)^2$ and $\mathrm{tr}(\mathrm{Var}) = \sigma^2\sum_j d_j^2/(d_j^2+\lambda)^2$, whose sum is Equation 6 term by term. □

Equation 6 makes explicit the trade-off at the heart of ridge regression: the numerator of each term increases in $\lambda$ through $\lambda^2\gamma_j^2$ (growing bias) while the denominator increases faster through $(d_j^2+\lambda)^2$ (shrinking variance); which effect dominates for small $\lambda$ is resolved in the next section.

## 6. EXISTENCE OF A RISK-IMPROVING RIDGE PARAMETER

**Theorem 4.** *If* $\sigma^2 > 0$ *and* $d_j > 0$ *for every* $j = 1, \ldots, p$, *then* $\mathrm{MSE}(\lambda)$ *in Equation 6 is strictly decreasing at* $\lambda = 0$:

$$\left.\frac{d\,\mathrm{MSE}(\lambda)}{d\lambda}\right|_{\lambda=0} = -2\sigma^2\sum_{j=1}^{p}\frac{1}{d_j^4} < 0. \quad (7)$$

*Consequently there exists* $\lambda^* > 0$ *such that* $\mathrm{MSE}(\lambda) < \mathrm{MSE}(0)$ *for all* $\lambda \in (0, \lambda^*)$: *a strictly positive ridge penalty always achieves strictly lower mean squared estimation error than ordinary least squares.*

*Proof.* Write the $j$-th summand of Equation 6 as $f_j(\lambda) = (\lambda^2\gamma_j^2 + \sigma^2 d_j^2)/(d_j^2+\lambda)^2$. By the quotient rule,

$$\begin{aligned} f_j'(\lambda) &= \frac{2\lambda\gamma_j^2(d_j^2+\lambda)^2 - 2(d_j^2+\lambda)(\lambda^2\gamma_j^2+\sigma^2 d_j^2)}{(d_j^2+\lambda)^4} \\ &= \frac{2\lambda\gamma_j^2(d_j^2+\lambda) - 2(\lambda^2\gamma_j^2+\sigma^2 d_j^2)}{(d_j^2+\lambda)^3}. \end{aligned}$$

At $\lambda = 0$ this simplifies to $f_j'(0) = -2\sigma^2 d_j^2/d_j^6 = -2\sigma^2/d_j^4$, which is strictly negative whenever $\sigma^2 > 0$ and $d_j > 0$, regardless of $\gamma_j$ (that is, regardless of the unknown true coefficients). Summing over $j$ gives Equation 7, and since MSE is differentiable with strictly negative derivative at $\lambda = 0$, it is strictly decreasing on some interval $(0, \lambda^*)$. □

*Remark* 1. Theorem 4 recovers, by direct differentiation rather than an optimization or geometric argument, the classical existence result associated with the original introduction of ridge regression: no information about the unknown $\beta$ is needed to guarantee that some positive penalty helps: the guarantee follows purely from $\sigma^2 > 0$. What Equation 7 adds beyond existence is a quantitative rate: the initial improvement is fastest exactly along the directions with the smallest singular values $d_j$, since $1/d_j^4$ is largest there, matching the shrinkage mechanism identified after Theorem 1.

## 7. BAYESIAN INTERPRETATION

**Theorem 5.** *Suppose* $\beta \sim N(0, \tau^2 I_p)$ *independently of* $\varepsilon \sim N(0, \sigma^2 I_n)$ *in model 1. Then the posterior distribution of* $\beta$ *given* $y$ *is Gaussian with mean*

$$\mathbb{E}[\beta \mid y] = (X^\top X + \tfrac{\sigma^2}{\tau^2} I)^{-1} X^\top y = \hat{\beta}(\lambda) \quad \textit{with} \quad \lambda = \sigma^2/\tau^2, \quad (8)$$

*so the ridge estimator at penalty* $\lambda$ *is exactly the Bayes posterior mean under a Gaussian prior of precision* $\lambda/\sigma^2$.

*Proof.* The log-posterior is, up to an additive constant not depending on $\beta$,

$$\log p(\beta \mid y) = -\frac{1}{2\sigma^2}\|y - X\beta\|^2 - \frac{1}{2\tau^2}\|\beta\|^2 + \mathrm{const},$$

by Bayes' rule and the two Gaussian densities. This is a quadratic form in $\beta$; completing the square (equivalently, differentiating and setting the gradient to zero) gives the maximizer, which for a Gaussian posterior coincides with

its mean,

$$\left(\frac{1}{\sigma^2}X^\top X+\frac{1}{\tau^2}I\right)\beta=\frac{1}{\sigma^2}X^\top y$$
$$\iff \qquad \beta=\left(X^\top X+\frac{\sigma^2}{\tau^2}I\right)^{-1}X^\top y,$$

which is Equation 8, identical to Equation 2's minimizer with $\lambda=\sigma^2/\tau^2$. □

Theorem 5 gives every subsequent choice of $\lambda$ a dual reading: a large ridge penalty corresponds to strong prior confidence that $\beta$ is close to zero (small $\tau^2$ relative to $\sigma^2$), and the degrees-of-freedom curve of Theorem 2 equivalently measures how much the data are allowed to override that prior.

## 8. CROSS-VALIDATION AND ITS SHORTCUT

**Lemma 2** (Leave-one-out shortcut)**.** *Let $S(\lambda)=X(X^\top X+\lambda I)^{-1}X^\top$ be the ridge hat matrix of Theorem 2, with diagonal entries $S_{ii}(\lambda)$, and let $\hat{y}(\lambda)=S(\lambda)y$. Denote by $\hat{y}_{(-i)}(\lambda)$ the ridge fit at $x_i$ using all observations except the $i$-th. Then*

$$y_i-\hat{y}_{(-i)}(\lambda)=\frac{y_i-\hat{y}_i(\lambda)}{1-S_{ii}(\lambda)},$$
$$\mathrm{LOOCV}(\lambda)=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\hat{y}_i(\lambda)}{1-S_{ii}(\lambda)}\right)^2. \tag{9}$$

*without refitting the model $n$ times.*

*Proof.* Let $\hat{\beta}_{(-i)}(\lambda)$ minimize Equation 2 over the data with the $i$-th row removed. A standard leave-one-out identity for penalized least-squares fits (obtained by applying the Sherman–Morrison rank-one update formula to $(X^\top X-x_ix_i^\top+\lambda I)^{-1}$ in terms of $(X^\top X+\lambda I)^{-1}$) gives $x_i^\top\hat{\beta}_{(-i)}(\lambda)=\hat{y}_i(\lambda)-S_{ii}(\lambda)\,(y_i-\hat{y}_i(\lambda))/(1-S_{ii}(\lambda))$; rearranging, $y_i-x_i^\top\hat{\beta}_{(-i)}(\lambda)=(y_i-\hat{y}_i(\lambda))/(1-S_{ii}(\lambda))$, which is Equation 9. Squaring and averaging over $i$ gives the stated leave-one-out cross-validation error. □

**Definition 2** (Generalized cross-validation)**.** *Generalized cross-validation replaces the individual leverages $S_{ii}(\lambda)$ in Equation 9 by their average $\mathrm{df}(\lambda)/n$ (Theorem 2), giving*

$$\mathrm{GCV}(\lambda)=\frac{n^{-1}\sum_{i=1}^{n}(y_i-\hat{y}_i(\lambda))^2}{\left(1-\mathrm{df}(\lambda)/n\right)^2}. \tag{10}$$

The averaging in Definition 2 is what makes $\mathrm{GCV}(\lambda)$ computable from a single fit rather than $n$ refits, at the cost of replacing an exact per-observation correction with an average one. Recent work has shown this approximation to be asymptotically consistent for ridge regression under increasingly general random-design conditions [6, 7], and extended GCV to sketched ridge ensembles for consistent risk estimation and joint tuning of regularization and sketching parameters [8]. Section 9.1 checks Lemma 2's shortcut directly against brute-force refitting, and Section 9.2 reports the $\lambda$ selected by minimizing $\mathrm{GCV}(\lambda)$ against the $\lambda$ selected by explicit 10-fold cross-validation.

## 9. NUMERICAL VERIFICATION AND EMPIRICAL APPLICATION

### 9.1 Verification of the leave-one-out shortcut

Table 1 compares the shortcut formula of Lemma 2 against brute-force leave-one-out refitting (a fresh ridge fit on $n-1$ observations, repeated for a random subset of 60 indices, at four representative values of $\lambda$). The two agree to floating-point precision at every value tested, confirming the identity exactly rather than approximately.

**Table 1.** Leave-one-out cross-validation error: brute-force refitting versus the closed-form shortcut of Lemma 2 (60 randomly selected observations per $\lambda$).

| $\lambda$ | Brute-force | Shortcut (Eq. 9) | Abs. difference |
|---|---|---|---|
| 0.1 | 2819.011173 | 2819.011173 | 0 |
| 1.0 | 3606.020462 | 3606.020462 | $4.5\times10^{-13}$ |
| 10.0 | 3903.099839 | 3903.099839 | 0 |
| 100.0 | 6630.693434 | 6630.693434 | 0 |

### 9.2 Degrees of freedom and generalized cross-validation

The design matrix has singular values ranging from $d_1=2.006$ to $d_{10}=0.093$ (condition number $d_1/d_{10}\approx21.7$), so Theorem 2 predicts $\mathrm{df}(\lambda)$ decreasing smoothly from 10 toward 0. Table 2 reports $\mathrm{df}(\lambda)$ and $\mathrm{GCV}(\lambda)$ on a representative grid; consistent with the convexity established in Lemma 1, $\mathrm{df}(\lambda)$ decreases smoothly throughout, and $\mathrm{GCV}(\lambda)$ traces a shallow U-shape with an interior minimum near $\lambda\approx0.0071$ ($\mathrm{df}\approx9.39$), rising both toward $\lambda=0$ and toward large $\lambda$.

**Table 2.** Effective degrees of freedom and generalized cross-validation score across the penalty grid.

| $\lambda$ | $\mathrm{df}(\lambda)$ | $\mathrm{GCV}(\lambda)$ |
|---|---|---|
| 0.000 | 10.000 | 2993.62 |
| 0.001 | 9.906 | 2992.46 |
| 0.002 | 9.761 | 2991.19 |
| 0.007 | 9.390 | 2990.10 |
| 0.016 | 9.006 | 2990.39 |
| 0.047 | 8.342 | 2990.37 |
| 0.374 | 5.774 | 3064.99 |
| 8.504 | 0.952 | 4710.77 |
| 547.890 | 0.018 | 5899.10 |
| 90110.183 | 0.000 | 5929.70 |

### 9.3 Verification of the bias–variance decomposition

To verify Theorem 3 at finite sample size without knowledge of the true $\beta$, a calibrated semi-synthetic experiment was run: the ordinary least-squares fit on the real data was taken as a fixed "true" $\beta$, the residual variance from that fit as $\sigma^2$, and 2,000 independent replicates of $y=X\beta+\varepsilon$, $\varepsilon\sim N(0,\sigma^2I)$, were simulated using the real design matrix $X$. Table 3 compares the resulting Monte Carlo bias$^2$ and variance of $\hat{\beta}(\lambda)$ against the closed forms of Equations 4–5. Agreement is close at every $\lambda$, including the near-exact match of Monte Carlo bias$^2$ to zero at $\lambda=0$ (the small nonzero value reflects Monte Carlo sampling error in estimating a theoretical zero, not model error), and the relative discrepancy between simulated and theoretical quantities shrinks as $\lambda$ grows and the variance component – the noisier of the two Monte Carlo quantities – shrinks correspondingly.

**Table 3.** Monte Carlo (2,000 replicates) versus theoretical bias$^2$ and variance of $\hat{\beta}(\lambda)$, calibrated to the real design matrix.

| $\lambda$ | MC Bias$^2$ | Theory Bias$^2$ | MC Var. | Theory Var. |
|---|---|---|---|---|
| 0 | 89.9 | 0.0 | 405,701 | 408,788 |
| 0.1 | 990,075 | 987,312 | 31,695 | 31,560 |
| 1 | 1,288,320 | 1,288,029 | 5,443 | 5,517 |
| 5 | 1,592,680 | 1,592,497 | 638 | 639 |
| 10 | 1,704,756 | 1,704,873 | 205 | 205 |
| 50 | 1,848,376 | 1,848,353 | 11 | 11 |
| 100 | 1,872,393 | 1,872,400 | 3 | 3 |
| 1000 | 1,895,742 | 1,895,743 | 0.03 | 0.03 |

The same calibrated construction locates the risk-minimizing penalty guaranteed to exist by Theorem 4: on a fine grid, theoretical MSE($\lambda$) is minimized at $\lambda \approx 0.0026$, a 19.7% reduction in mean squared estimation error relative to $\lambda = 0$ (ordinary least squares), and the derivative at $\lambda = 0$ computed from Equation 7 is $-8.09 \times 10^7$, confirming the strictly negative initial slope Theorem 4 guarantees whenever $\sigma^2 > 0$.

### 9.4 Out-of-sample comparison against the lasso and elastic net

Table 4 reports mean squared prediction error, averaged over 30 repetitions of 10-fold cross-validation (three repeats of a 10-fold split), for ordinary least squares, ridge regression tuned by GCV (Definition 2), ridge tuned by an inner 10-fold cross-validation, the lasso, and the elastic net (the latter two tuned by inner cross-validation over their respective penalty and mixing parameters).

**Table 4.** Out-of-sample prediction error, 3×10-fold repeated cross-validation (30 folds total).

| Model | MSE | SE | RMSE |
|---|---|---|---|
| Ordinary least squares | 3016.7 | 90.1 | 54.92 |
| Ridge (GCV-tuned) | 3024.6 | 92.0 | 55.00 |
| Ridge (10-fold CV-tuned) | 3020.4 | 91.6 | 54.96 |
| Lasso (CV-tuned) | 3026.2 | 91.8 | 55.01 |
| Elastic net (CV-tuned) | 3025.2 | 91.7 | 55.00 |

All five models lie within one standard error of one another; the mean cross-validated ridge penalty ($\lambda \approx 0.03$–$0.06$ depending on tuning method) is small, and the lasso retains on average 8.1 of the 10 predictors, indicating mild rather than severe shrinkage is selected by every data-driven criterion. Table 5 reports the resulting coefficient estimates: ridge shrinks every coefficient toward zero relative to ordinary least squares, most visibly the two most volatile serum-measurement coefficients (`s1` and `s2`, which are highly correlated with each other and with `s4`), while the lasso and elastic net additionally shrink several coefficients close to (though, at the cross-validated penalty, not exactly to) zero.

**Table 5.** Coefficient estimates: ordinary least squares versus ridge (GCV-tuned), lasso (CV-tuned), and elastic net (CV-tuned), averaged across cross-validation folds.

| Predictor | OLS | Ridge | Lasso | Elastic net |
|---|---|---|---|---|
| age | −10.01 | −4.41 | −1.85 | −1.70 |
| sex | −239.82 | −226.78 | −207.95 | −207.97 |
| bmi | 519.85 | 510.98 | 521.61 | 517.95 |
| bp | 324.38 | 315.42 | 304.33 | 304.40 |
| s1 | −792.18 | −311.03 | −249.61 | −240.92 |
| s2 | 476.74 | 97.00 | 85.12 | 75.08 |
| s3 | 101.04 | −100.17 | −138.97 | −139.77 |
| s4 | 177.06 | 133.71 | 69.46 | 76.56 |
| s5 | 751.27 | 554.17 | 557.17 | 548.31 |
| s6 | 67.63 | 76.08 | 56.89 | 60.11 |

## 10. DISCUSSION

The numerical verification in Section 9 supports two conclusions that are easy to conflate but are logically distinct, a distinction the closed-form theory keeps separate where a purely empirical comparison would not. First, Theorem 4 guarantees, and the calibrated simulation confirms, a strictly positive and quantifiable reduction in *parameter-estimation* risk, $\mathbb{E}\|\hat{\beta}(\lambda) - \beta\|^2$, from a suitably small ridge penalty – on this design, close to 20% at its calibrated optimum. Second, Table 4 shows that this does not translate into a detectable reduction in *prediction* risk on held-out data for this particular dataset: every regularized method is statistically indistinguishable from ordinary least squares in cross-validated mean squared error. Both findings are correct and are not in tension; they answer different questions. The design matrix here has a modest condition number ($\approx 21.7$) and $n \gg p$, so ordinary least squares is already a low-variance estimator of the quantity that matters for prediction, $X\beta$, even though it remains a comparatively higher-variance estimator of $\beta$ itself, and Equation 6 is stated in terms of the latter. This distinction echoes the broader point made in the random-design bias-variance literature [18] that fixed-$X$ estimation risk and genuinely out-of-sample prediction risk need not move together, and it is consistent with the empirical comparisons of ridge, lasso, and elastic net reported elsewhere on similarly well-conditioned, low-dimensional data [16, 17], where gains over ordinary least squares are typically small unless collinearity is more severe or $p$ is closer to $n$.

The generalized cross-validation curve in Table 2 is informative in the same direction: it is shallow, with GCV values within 0.2% of one another across nearly two orders of magnitude of $\lambda$ (from df $\approx 10$ down to df $\approx 8.3$), meaning the data provide only weak evidence for any particular amount of regularization in this range – a finding a single point estimate of the GCV-optimal $\lambda$ would obscure but the full table makes visible. This shallowness is itself consistent with theoretical work showing that GCV's consistency, and the sharpness with which it identifies the risk-optimal penalty, depend on the aspect ratio $p/n$ and the spectrum of $X$ [6, 7]. Related work extends consistent GCV risk estimation to sketched ridge ensembles and uses it to tune both regularization and sketching parameters [8]. With $p/n \approx 0.023$ here, the present setting is far from the proportional-asymptotic regime in which the most delicate high-dimensional behavior arises.

The Bayesian equivalence of Theorem 5 offers a complementary reading of the same shallow curve: since $\lambda = \sigma^2/\tau^2$, a GCV curve that is flat across a wide range of $\lambda$ is a curve that is flat across a correspondingly wide range of prior precisions, meaning the data are not particularly informative about how strongly to believe $\beta$ is close to zero – a natural characterization of a design that is already well conditioned. More elaborate priors, including the shrinkage priors compared under a similar objective across baseline choices in [19], would be expected to show the same shallowness on this design for the same reason.

The ten predictors used here are already low-dimensional and only mildly collinear; in genuinely high-dimensional settings, where $p$ is comparable to or exceeds $n$, the same closed-form machinery extends less directly, and dimension-reduction alternatives such as principal component regression

– whose high-dimensional prediction error has been analyzed through high-probability risk bounds [20] – become an increasingly relevant comparison, since they shrink along the same singular-value directions identified after Theorem 1 but discard, rather than merely shrink, the directions associated with the smallest singular values.

## 11. CONCLUSION

This paper assembled six exact, finite-sample results about ridge regression – a closed-form singular value representation, monotone convex degrees of freedom, an exact bias-variance decomposition, a guaranteed strictly risk-improving penalty, a Gaussian posterior-mean equivalence, and a closed-form leave-one-out shortcut underlying generalized cross-validation – into a single proved development, and verified every one of them numerically: the leave-one-out shortcut exactly, and the bias-variance decomposition through a calibrated simulation on a real design matrix. Applied to a standard disease-progression dataset, the theory correctly predicted the existence and approximate location of a risk-reducing regularization region, while the empirical cross-validated comparison against the lasso and elastic net showed that this estimation-risk gain does not, on this particular well-conditioned design, translate into a detectable prediction-risk gain – a distinction the closed-form decomposition makes precise rather than leaving to be inferred from accuracy tables alone. The same development applies unchanged to any fixed design matrix, and the numerical verification strategy used here – exact shortcut-formula checks where available, calibrated simulation using the real design otherwise – offers a template for validating closed-form regularization theory against specific datasets rather than only against simulated or asymptotic settings.

## DATA AND CODE AVAILABILITY

The diabetes dataset is the standard public dataset of [15]. The complete Python code implementing every theorem's numerical verification, the calibrated simulation, and the cross-validated model comparison, together with all output tables, are provided as supplementary material accompanying this submission.

## REFERENCES


[1] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 063–30 070, 2020.

[2] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, "Surprises in high-dimensional ridgeless least squares interpolation," *The Annals of Statistics*, vol. 50, no. 2, pp. 949–986, 2022.

[3] T. Liang and A. Rakhlin, "Just interpolate: Kernel "ridgeless" regression can generalize," *The Annals of Statistics*, vol. 48, no. 3, pp. 1329–1347, 2020.

[4] D. Richards, J. Mourtada, and L. Rosasco, "Asymptotics of ridge(less) regression under general source condition," in *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, ser. PMLR, vol. 130, 2021, pp. 3889–3897.

[5] A. Tsigler and P. L. Bartlett, "Benign overfitting in ridge regression," *Journal of Machine Learning Research*, vol. 24, no. 123, pp. 1–76, 2023.

[6] P. Patil, Y. Wei, A. Rinaldo, and R. Tibshirani, "Uniform consistency of cross-validation estimators for high-dimensional ridge regression," in *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, ser. PMLR, vol. 130, 2021, pp. 3178–3186.

[7] K. Rahnama Rad and A. Maleki, "A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 82, no. 4, pp. 965–996, 2020.

[8] P. Patil and D. LeJeune, "Asymptotically free sketched ridge ensembles: Risks, cross-validation, and tuning," *arXiv preprint arXiv:2310.04357*, 2023.

[9] D. Kobak, J. Lomond, and B. Sanchez, "The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization," *Journal of Machine Learning Research*, vol. 21, no. 169, pp. 1–16, 2020.

[10] P. Patil and J.-H. Du, "Generalized equivalences between subsampling and ridge regularization," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[11] J.-H. Du, P. Patil, and A. K. Kuchibhotla, "Subsample ridge ensembles: Equivalences and generalized cross-validation," in *Proceedings of the 40th International Conference on Machine Learning*, ser. PMLR, vol. 202, 2023, pp. 8585–8631.

[12] P. Patil, J.-H. Du, and A. K. Kuchibhotla, "Bagging in overparameterized learning: Risk characterization and risk monotonization," *Journal of Machine Learning Research*, vol. 24, no. 319, pp. 1–113, 2023.

[13] E. Dobriban and Y. Sheng, "Distributed linear regression by averaging," *The Annals of Statistics*, vol. 49, no. 2, pp. 918–943, 2021.

[14] B. Adlam and J. Pennington, "Understanding double descent requires a fine-grained bias-variance decomposition," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11 022–11 032, 2020.

[15] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[16] M. Usman, S. Doguwa, and B. Alhaji, "Comparing the prediction accuracy of ridge, lasso and elastic net regression models with linear regression using breast cancer data," *Bayero Journal of Pure and Applied Sciences*, vol. 14, no. 2, pp. 134–149, 2022.

[17] N. Herawati, A. Wijayanti, A. Sutrisno, Nusyirwana, and Misgiyatia, "The performance of ridge regression, lasso, and elastic-net in controlling multicollinearity: A simulation and application," *Journal of Modern Applied Statistical Methods*, vol. 23, no. 2, 2024.

[18] S. Rosset and R. J. Tibshirani, "From fixed-X to random-X regression: Bias-variance decompositions, covariance penalties, and prediction error estimation," *Journal of the American Statistical Association*, vol. 115, no. 529, pp. 138–151, 2020.

[19] G. Tzoumerkas, D. Fouskakis, and I. Ntzoufras, "A comparison of power-expected-posterior priors in shrinkage regression," *Journal of Statistical Theory and Practice*, vol. 16, p. 61, 2022.

[20] L. Hucker and M. Wahl, "A note on the prediction error of principal component regression in high dimensions," *Theory of Probability and Mathematical Statistics*, vol. 109, pp. 37–53, 2023.