Fitting the Data Worse on Purpose Predicts Better, and How Far to Shrink Is Set by the Strength of the Signal

— shrinking the coefficients by $n/(n+\lambda)$ raises the in-sample miss monotonically while the prediction error reaches its minimum of 0.333333 at $\lambda = 50$, the best λ being dimension times noise over signal — the separator is whether the signal is stronger or weaker than noise times dimension over sample size — [Paper 744]

Yuuki YAMAGISHI (山岸 勇輝)
Yamagishi Complexity Torus Lab, Japan
September 2026

— *From thinking to syncing.* —

Abstract

The coefficients that fit the data at hand best need not predict new data best. Shrinking them on purpose can improve prediction.How far to shrink, and how much is gained, is counted by closed form and simulation.No new theorem or law is claimed. First, shrinking makes the in-sample fit worse——the miss term rises monotonically: 0.002268, 0.111111 and 0.694444 at $\lambda = 5$, 50 and 500 (Section 2). **Second, yet prediction improves——the prediction error falls from 0.500000 to 0.333333 at $\lambda = 50$ and returns to 0.708333 at 500** (Section 2). **Third, and this is the core. The best shrinkage is dimension times noise over signal——$\lambda = 50.000000$ at signal 1 and 5.000000 at signal 10** (Section 3). **Fourth, the weaker the signal, the larger the gain——the error falls to 0.666667 at signal 1 but only to 0.952381 at signal 10** (Section 3). **Fifth, simulation confirms it——over 20000 runs, 0.499461 at $\lambda = 0$ and 0.332937 at the best λ** (Section 3). **Sixth, the separator is whether the signal is stronger or weaker than noise times dimension over sample size——weak, shrink hard; strong, hardly at all** (Section 4).
Keywords: regularisation, ridge regression, coefficient shrinkage, bias and variance, signal-to-noise ratio, fit and prediction moving apart.

Scope of this paper (scope note): **No new theorem or law is claimed——ridge regression, trading bias for variance by shrinking, and the optimal λ for orthogonal regressors are all known (Hoerl and Kennard 1970, Hastie 2009).Orthogonal regressors only——for the closed form; correlated regressors are not treated.Choosing λ is not treated ——the optimum is found as if the signal strength were known; in practice it is not, and one would choose by cross-validation or similar, which is not treated.Other penalties are not treated——the lasso and variable selection are not treated.No real data are claimed——particular data sets and learners are not treated.Relation to earlier papers**: **Paper 733 showed that the error has a minimum whose location moves with the amount of data——its scope note stated that cross-validation, regularisation and early stopping would not be treated. There the minimum was made by choosing the degrees of freedom. Here, without changing them, the coefficients are shrunk and the same minimum is made with a continuous knob. Choosing degrees of freedom, and shrinking coefficients.Paper 330 is also cited——it showed that the squaring of the condition number in the normal equations is an identity; ridge's λ is what is added to that matrix's diagonal.**What is added is **showing in one table that in-sample and prediction move opposite ways**, **confirming the best λ by closed form, grid and simulation**, **showing the gain to range from 0.666667 to 0.952381 with signal strength**, and **placing the separator on signal against noise times dimension over sample size**.

1. Notation and the Question

Setting: **with orthogonal regressors, ridge regression shrinks every least-squares coefficient by $n/(n+\lambda)$. The excess prediction error is the bias introduced by shrinking plus the variance that remains.**

$$\hat{\beta}_\lambda = \frac{n}{n+\lambda}\,\hat{\beta}, \qquad \mathrm{E}(\lambda) = \Big(\frac{\lambda}{n+\lambda}\Big)^2 \|\beta\|^2 + \Big(\frac{n}{n+\lambda}\Big)^2 \frac{d\sigma^2}{n}, \qquad \lambda^* = \frac{d\sigma^2}{\|\beta\|^2}$$

Take $d = 50$, $n = 100$ and $\sigma^2 = 1$, so noise times dimension over sample size is 0.500000.

Question: **are the coefficients that fit best the ones that predict best?**

2. In-Sample and Prediction Move Opposite Ways

Computation 2.1: **at signal $\|\beta\|^2 = 1$, the shrink factor, the in-sample miss term and the prediction error were derived for each λ.**

λ	shrink factor	in-sample miss term	prediction error
0	1.000000	0.000000	0.500000
5	0.952381	0.002268	0.455782
50	0.666667	0.111111	0.333333
500	0.166667	0.694444	0.708333

> **Observation 2.2**
>
> the more one shrinks, the worse the in-sample fit.The miss term rises monotonically with λ——judged on the data at hand alone, $\lambda = 0$ is always best.

> **Observation 2.3**
>
> yet the prediction error has a minimum in between.It falls from 0.500000 to 0.333333 at $\lambda = 50$ and returns to 0.708333 at 500——fitting worse on purpose predicts better, and fitting too badly predicts worse again.

3. The Strength of the Signal Sets the Shrinkage

Computation 3.1: for two signal strengths, the best λ and the error there were derived, and checked against a grid search.

signal $\|\beta\|^2$	best λ	error at best	ratio to $\lambda = 0$
1	50.000000	0.333333	0.666667
10	5.000000	0.476190	0.952381

> **Observation 3.2**
>
> this is the core: the best shrinkage is dimension times noise over signal.Ten times the signal gives one tenth the best λ, and the grid minima fall at 50.00 and 5.00——what sets the shrinkage is not the in-sample fit but the ratio of signal to noise.

Observation 3.3

the weaker the signal, the larger the gain.The error falls to 0.666667 at signal 1 but only to 0.952381 at signal 10——the error at the optimum is noise times dimension over sample size scaled by the signal's share, so shrinking pays when the signal is weak against that quantity.

Computation 3.4: **at signal 1, adding noise to the coefficients and shrinking was repeated 20000 times.**

λ	simulation	closed form
0	0.499461	0.500000
50	0.332937	0.333333

Observation 3.5

both sit within 2 % of the closed form.That the optimum brings the error to two thirds of $\lambda = 0$ is confirmed by a separate route——it is not an artefact of the formula.

4. The Separator

Observation 4.1

the separator is whether the signal is stronger or weaker than noise times dimension over sample size.Weak, shrink hard; strong, hardly at all——"does regularisation help?" has no answer until this ratio is stated.

Observation 4.2

beside Paper 733.Paper 733 showed the minimum of error to move with the amount of data, and left regularisation aside——there the minimum was made by choosing degrees of freedom; here, by shrinking coefficients with a continuous knob.

Observation 4.3

beside Paper 330 as well.Paper 330 showed that the condition number's squaring in the normal equations is an identity, not an approximation——ridge's λ is added to that matrix's diagonal, and this paper counts it from the side of prediction error.

Observation 4.4

what is not said.Correlated regressors, the choice of λ and other penalties are not treated, and the signal strength is taken as known——what is shown is that fit and prediction move opposite ways, and the ratio that sets the shrinkage.

5. Roll Call and Closing

1. **prediction 0.500000 at $\lambda = 0$** (Computation 2.1): **the best fit.**
2. **0.333333 at $\lambda = 50$** (Computation 2.1): **worse fit, better prediction.**
3. **0.708333 at $\lambda = 500$** (Computation 2.1): **too much.**
4. **in-sample miss 0.111111** (Computation 2.1): **rising monotonically.**
5. **best 50.000000 at signal 1** (Computation 3.1): **the core.**
6. **5.000000 at signal 10** (Computation 3.1): **one tenth.**
7. **gain 0.666667** (Observation 3.3): **weak signal.**
8. **gain 0.952381** (Observation 3.3): **strong signal.**
9. **simulation 0.332937** (Computation 3.4): **on the closed form.**
10. **signal against noise times dimension over samples** (Observation 4.1): **the separator; Paper 733 and Paper 330.**

Closing: **Fitting the data worse on purpose predicts better, and how far to shrink is set by the strength of the signal.With orthogonal regressors, dimension 50 and 100 samples, shrinking raises the in-sample miss term monotonically, 0.002268, 0.111111 and 0.694444 at $\lambda = 5$, 50 and 500, while the prediction error falls from 0.500000 through 0.455782 to 0.333333 and returns to 0.708333——judged on the data at hand $\lambda = 0$ is best, and it is not best for prediction.The best λ is dimension times noise over signal: 50.000000 at signal 1 and 5.000000 at signal 10, lowering the error to 0.666667 and 0.952381 of its value——shrinking pays most when the signal is weak against noise times dimension over sample size.20000 simulated runs gave 0.499461 and 0.332937, within 2 % of the closed form.The separator is whether the signal is stronger or weaker than noise times dimension over sample size**

——“does regularisation help?” has no answer until this ratio is stated.Placed among the earlier papers——Paper 733 made the minimum by choosing degrees of freedom and left regularisation aside; this makes it by shrinking.To be honest——regressors are orthogonal, the choice of λ and other penalties are not treated, and the signal strength is taken as known.

References

1. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.
2. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*.
3. Yamagishi, Y. (2026). *That the Condition Number Squares Is an Identity, Not an Approximation* [Paper 330]. Zenodo. 10.5281/zenodo.22681423
4. Yamagishi, Y. (2026). *The Error Has a Minimum, but Where That Minimum Sits Is Not a Property of the Model: It Moves with How Much Data You Hold* [Paper 733]. Zenodo. 10.5281/zenodo.23015055

— From thinking to syncing. —

Yuuki YAMAGISHI (山岸 勇輝)

© 2026 Yamagishi Complexity Torus Lab, Japan · Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)