Journal of Machine Learning Research 21 (2020) 1-16 Submitted 10/19; Revised 6/20; Published 7/20

The Optimal Ridge Penalty for Real-world High-dimensional Data Can Be Zero or Negative due to the Implicit Ridge Regularization

Dmitry Kobak DMITRY.KOBAK@UNI-TUEBINGEN.DE
Institute for Ophthalmic Research
University of Tübingen
Otfried-Müller-Straße 25, 72076 Tübingen

Jonathan Lomond JONATHAN.LOMOND@GMAIL.COM
Toronto, Canada

Benoit Sanchez BEN.SAN3@GMAIL.COM
Paris, France

Editor: Ambuj Tewari

Abstract

A conventional wisdom in statistical learning is that large models require strong regularization to prevent overfitting. Here we show that this rule can be violated by linear regression in the underdetermined $n \ll p$ situation under realistic conditions. Using simulations and real-life high-dimensional datasets, we demonstrate that an explicit positive ridge penalty can fail to provide any improvement over the minimum-norm least squares estimator. Moreover, the optimal value of ridge penalty in this situation can be negative. This happens when the high-variance directions in the predictor space can predict the response variable, which is often the case in the real-world high-dimensional data. In this regime, low-variance directions provide an implicit ridge regularization and can make any further positive ridge penalty detrimental. We prove that augmenting any linear model with random covariates and using minimum-norm estimator is asymptotically equivalent to adding the ridge penalty. We use a spiked covariance model as an analytically tractable example and prove that the optimal ridge penalty in this case is negative when $n \ll p$.

Keywords: High-dimensional, ridge regression, regularization

1. Introduction

In recent years, there has been increasing interest in prediction problems in which the sample size n is much smaller than the dimensionality of the data p. This situation is known as $n \ll p$ and often arises in computational chemistry and biology, e.g. in chemometrics, brain imaging, or genomics (Hastie et al., 2009). The standard approach to such problems is "to bet on sparsity" (Hastie et al., 2015) and to use linear models with regularization performing feature selection, such as the lasso (Tibshirani, 1996), the elastic net (Zou and Hastie, 2005), or the Dantzig selector (Candes and Tao, 2007).

In this paper we study ordinary least squares (OLS) linear regression with loss function

$$\mathcal{L} = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2, \tag{1}$$

where $\mathbf{X}$ is a $n \times p$ matrix of predictors and $\mathbf{y}$ is a $n \times 1$ matrix of responses. Assuming $n > p$ and full-rank $\mathbf{X}$, the unique solution minimizing this loss function is given by

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \tag{2}$$

This estimator is unbiased and has small variance when $n \gg p$. As p grows for a fixed n, $\mathbf{X}^\top\mathbf{X}$ becomes poorly conditioned, increasing the variance and leading to overfitting. The expected error can be decreased by shrinkage as provided e.g. by the ridge estimator (Hoerl and Kennard, 1970), a special case of Tikhonov regularization (Tikhonov, 1963),

$$\hat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}, \tag{3}$$

which minimizes the loss function with an added ℓ_2 penalty

$$\mathcal{L}_\lambda = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2. \tag{4}$$

The closer p is to n, the stronger the overfitting and the more important it is to use regularization. It seems intuitive that when p becomes larger than n, regularization becomes indispensable and small values of $\lambda \approx 0$ would yield hopeless overfitting. A popular textbook (James et al., 2013), for example, claims that *"though it is possible to perfectly fit the training data in the high-dimensional setting, the resulting linear model will perform extremely poorly on an independent test set, and therefore does not constitute a useful model."* Here we show that this intuition is incomplete.

Specifically, we empirically demonstrate and mathematically prove the following:

(i) when $n \ll p$, the $\lambda \to 0$ limit, corresponding to the minimum-norm OLS solution, can have good generalization performance;

(ii) explicit ridge regularization with $\lambda > 0$ can fail to provide any further improvement;

(iii) moreover, the optimal value of λ in this regime can be *negative*;

(iv) this happens when the response variable is predicted by the high-variance directions while the low-variance directions together with the minimum-norm requirement effectively perform shrinkage and provide implicit ridge regularization.

Our results provide a simple counter-example to the common understanding that large models with little regularization do not generalize well. This has been pointed out as a puzzling property of deep neural networks (Zhang et al., 2017), and has been subject to a very active ongoing research since then, performed independently from our work (the first version of this manuscript was released as a preprint in May 2018). Several groups reported that very different statistical models can display $\backslash/\backslash$-shaped (*double descent*) risk curves as a function of model complexity, extending the classical U-shaped risk curves and having small or even the smallest risk in the overparametrized regime (Advani et al., 2020; Spigler et al., 2019; Belkin et al., 2019a). The same phenomenon was later demonstrated for modern deep learning architectures (Nakkiran et al., 2020a). In the context of linear or kernel methods, the high-dimensional regime when the model is rich enough to fit any training data with zero loss, has been called *ridgeless* regression or *interpolation* (Liang and Rakhlin, 2020; Hastie et al., 2019). The fact that such interpolating estimators can have low risk has been called *benign overfitting* (Bartlett et al., 2020; Chinot and Lerasle, 2020) and *harmless interpolation* (Muthukumar et al., 2020).

Our finding (i) is in line with this body of parallel literature. Findings (ii) and (iii) have not, to the best of our knowledge, been described anywhere else. Existing studies of high-dimensional ridge regression found that, under some generic assumptions, the ridge risk at some $\lambda > 0$ always dominates the minimum-norm OLS risk (Dobriban and Wager, 2018; Hastie et al., 2019). Our results highlight that the optimal value of ridge penalty can be zero or even negative, suggesting that real-world $n \ll p$ datasets can have very different statistical structure compared to the common theoretical models (Dobriban and Wager, 2018). Finding (ii) has been observed for kernel methods (Liang and Rakhlin, 2020) and for random features regression (Mei and Montanari, 2019); our results demonstrate that (ii) can happen in a simpler situation of ridge regression with Gaussian features. We are not aware of any existing work reporting that the optimal ridge penalty can be *negative*, as

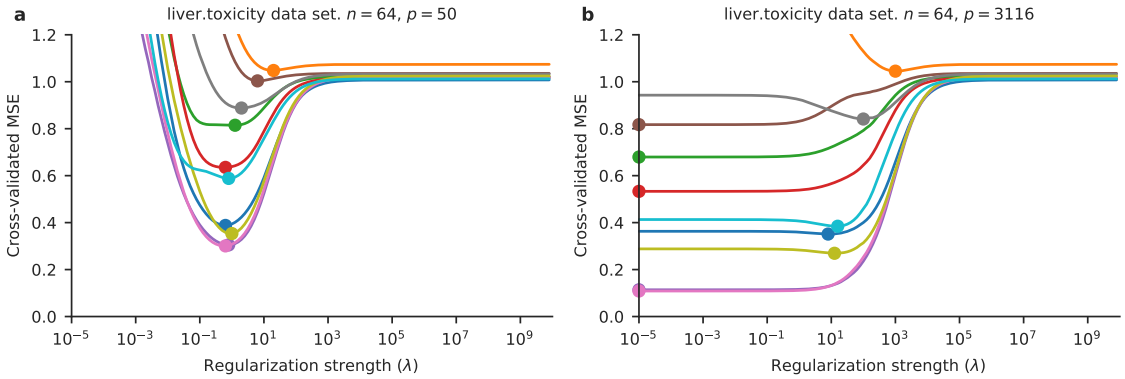

Figure 1: Cross-validation estimate of ridge regression performance for the `liver.toxicity` dataset. **a.** Using $p = 50$ randomly chosen predictors. **b.** Using all $p = 3116$ predictors. Lines correspond to 10 dependent variables. Dots show minimum values.

per our finding (iii). Finally, finding (iv) is related to the results of Bibas et al. (2019) and Bartlett et al. (2020); the connection between the minimum-norm OLS and the ridge estimators was also studied by Dereziński et al. (2019).

The code in Python can be found at `http://github.com/dkobak/high-dim-ridge`.

2. Results

2.1 A case study of ridge regression in high dimensions

We used the `liver.toxicity` dataset (Bushel et al., 2007) from the R package `mixOmics` (Rohart et al., 2017) as a motivational example to demonstrate the phenomenon. This dataset contains microarray expression levels of $p = 3116$ genes and 10 clinical chemistry measurements in liver tissue of $n = 64$ rats. We centered and standardized all the variables before the analysis.

We used `glmnet` library (Friedman et al., 2010) to predict each chemical measurement from the gene expression data using ridge regression. `Glmnet` performed 10-fold cross-validation (CV) for various values of regularization parameter λ. We ran CV separately for each of the 10 dependent variables. When we used $p = 50$ random predictors, there was a clear minimum of mean squared error (MSE) for some $\lambda_{\text{opt}} > 0$, and smaller values of λ yielded much higher MSE, i.e. led to overfitting (Figure 1a). This is in agreement with Hoerl and Kennard (1970) who proved that when $n < p$, the optimal penalty λ_{opt} is always larger than zero. The CV curves had a similar shape when $p \gtrsim n$, e.g. $p = 75$.

However, when we used all $p \gg n$ predictors, the curves changed dramatically (Figure 1b). For five dependent variables out of ten, the lowest MSE corresponded to the smallest value of λ that we tried. Four other dependent variables had a minimum in the middle of the λ range, but the limiting MSE value at $\lambda \to 0$ was close to the minimal one. This is counter-intuitive: despite having more predictors than samples, tiny values of $\lambda \approx 0$ provide optimal or near-optimal estimator.

We observed the same effect in various other genomics datasets with $n \ll p$ (Kobak et al., 2018). We believe it is a general phenomenon and not a peculiarity of this particular dataset.

2.2 Minimum-norm OLS estimator

When $n < p$, the limiting value of the ridge estimator at $\lambda \to 0$ is the minimum-norm OLS estimator. This can be shown using a thin singular value decomposition (SVD) of the predictor matrix $\mathbf{X} =$

$\mathbf{USV}^\top$ (with $\mathbf{S}$ square and all its diagonal values non-zero):

$$\hat{\boldsymbol{\beta}}_0 = \lim_{\lambda\to 0}\hat{\boldsymbol{\beta}}_\lambda = \lim_{\lambda\to 0}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} = \lim_{\lambda\to 0}\mathbf{V}\frac{\mathbf{S}}{\mathbf{S}^2+\lambda}\mathbf{U}^\top\mathbf{y} = \mathbf{V}\mathbf{S}^{-1}\mathbf{U}^\top\mathbf{y} = \mathbf{X}^\dagger\mathbf{y}, \tag{5}$$

where $\mathbf{X}^\dagger = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}$ denotes pseudo-inverse of $\mathbf{X}$ and operations on the diagonal matrix $\mathbf{S}$ are assumed to be element-wise and applied only to the diagonal.

The estimator $\hat{\boldsymbol{\beta}}_0$ gives one possible solution to the OLS problem and, as any other solution, it provides a perfect fit on the training set:

$$\|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_0\|^2 = \|\mathbf{y}-\mathbf{X}\mathbf{X}^\dagger\mathbf{y}\|^2 = \|\mathbf{y}-\mathbf{y}\|^2 = 0. \tag{6}$$

The $\hat{\boldsymbol{\beta}}_0$ solution is the one with the minimum ℓ_2 norm:

$$\hat{\boldsymbol{\beta}}_0 = \arg\min\Big\{\|\boldsymbol{\beta}\|^2 \;\Big|\; \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 = 0\Big\}. \tag{7}$$

Indeed, any other solution can be written as a sum of $\hat{\boldsymbol{\beta}}_0$ and a vector from the $(p-n)$-dimensional subspace orthogonal to the column space of $\mathbf{V}$. Any such vector yields a valid OLS solution but increases its norm compared to $\hat{\boldsymbol{\beta}}_0$ alone.

This allows us to rephrase the observations made in the previous section as follows: when $n \ll p$, the minimum-norm OLS estimator can have lower risk (expected squared error) than any ridge estimator with $\lambda > 0$.

2.3 Simulation using spiked covariance model

We qualitatively replicated this empirically observed phenomenon with a simple model where all p predictors are positively correlated to each other and all have the same effect on the response variable.

Let $\mathbf{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ be a p-dimensional vector of predictors with covariance matrix $\boldsymbol{\Sigma}$ having all diagonal values equal to $1+\rho$ and all non-diagonal values equal to ρ. This is known as *spiked covariance model*: $\boldsymbol{\Sigma} = \mathbf{I} + \rho\mathbf{1}\mathbf{1}^\top$ deviates from the spherical covariance $\mathbf{I}$ in only one dimension. Let the response variable be $y = \mathbf{x}^\top\boldsymbol{\beta} + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and $\boldsymbol{\beta} = (b, b\ldots b)^\top$ has all identical elements. We select $b = \sigma\sqrt{\alpha/(p+p^2\rho)}$ in order to achieve signal-to-noise ratio $\mathrm{Var}[\mathbf{x}^\top\boldsymbol{\beta}]/\mathrm{Var}[\varepsilon] = \mathrm{Var}[\mathbf{x}^\top\boldsymbol{\beta}] = \alpha$. In all simulations we fix $\sigma^2 = 1$, $\rho = 0.1$ and $\alpha = 10$.

Using this model with different values of p, we generated many ($N_{\mathrm{rep}} = 100$) training sets $(\mathbf{X}, \mathbf{y})$ with $n = 64$ each, as in the `liver.toxicity` dataset analyzed above. Using each training set, we computed $\hat{\boldsymbol{\beta}}_\lambda = \mathbf{V}\frac{\mathbf{S}}{\mathbf{S}^2+\lambda}\mathbf{U}^\top\mathbf{y}$ for various values of λ and then found MSE (risk) of $\hat{\boldsymbol{\beta}}_\lambda$ using the formula

$$R(\hat{\boldsymbol{\beta}}_\lambda) = \mathbb{E}_{\mathbf{x},\varepsilon}\big[\big((\mathbf{x}^\top\boldsymbol{\beta}+\varepsilon) - \mathbf{x}^\top\hat{\boldsymbol{\beta}}_\lambda\big)^2\big] = (\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})^\top\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}) + \sigma^2. \tag{8}$$

We normalized the MSE by $\mathrm{Var}[y] = \boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} + \sigma^2 = (\alpha+1)\sigma^2$. Then we averaged normalized MSEs across N_{rep} training sets to get an estimate of the expected normalized MSE. The results for $p \in \{50, 75, 150, 1000\}$ (Figure 2a–d) match well to what we previously observed in real data (Figure 1): when $n > p$ or $n \lesssim p$, the MSE had a clear minimum for some positive value of λ. But when $n \ll p$, the minimum MSE was achieved by the $\lambda = 0$ minimum-norm OLS estimator.

Figure 2e shows the expected normalized MSE of the OLS and the minimum-norm OLS estimators for $p \in [10, 1000]$. The true signal-to-noise ratio was always $\alpha = 10$, so the best attainable normalized MSE was always $1/(10+1) \approx 0.09$. With $p = 10$, OLS yielded a near-optimal performance. As p increased, OLS began to overfit and each additional predictor increased the MSE. Near $p \approx n = 64$ the expected MSE became very large, but as p increased even further, the MSE of the minimum-norm OLS quickly decreased again.

The risk of the optimal ridge estimator was close to the oracle risk for all dimensionalities (Figure 2e, dashed line), and did not show any divergence at $p = n$. However, as $p > n$ grew, the

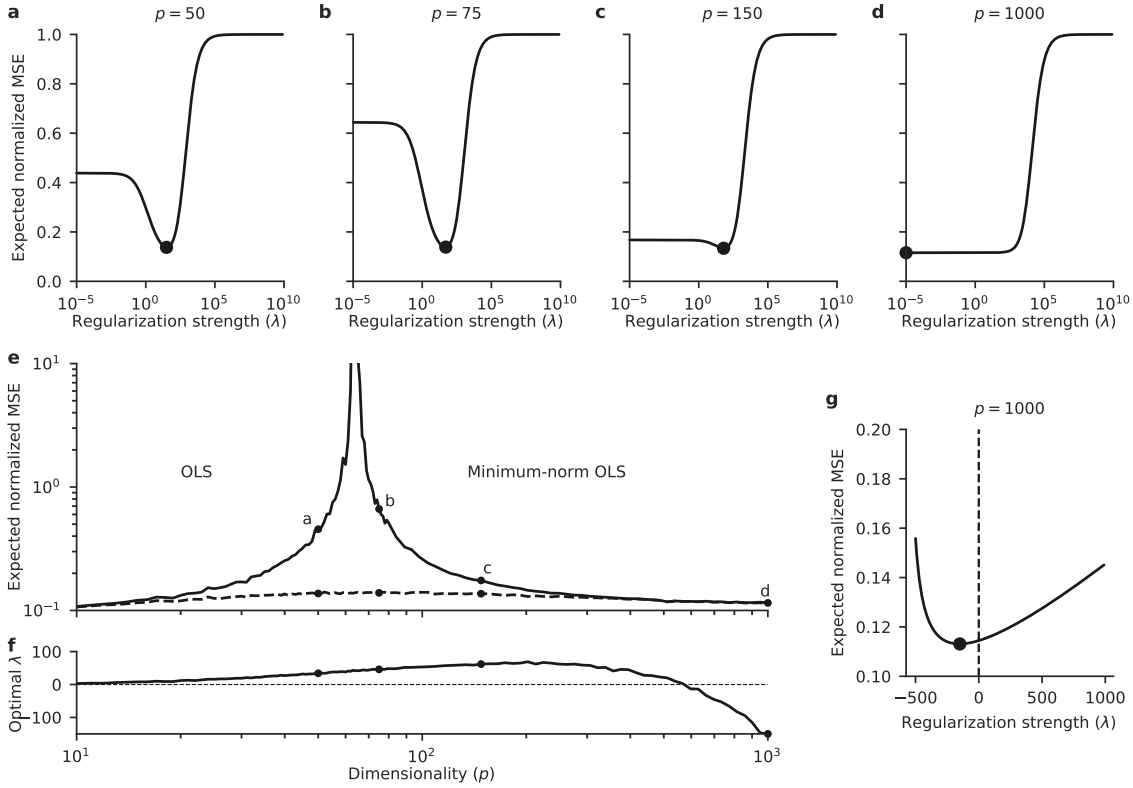

Figure 2: a–d. Expected normalized MSE of ridge estimators using a model with correlated predictors. On all subplots $n = 64$. Subplots correspond to the number of predictors p taking values 50, 75, 150, and 1000. Dots mark the points with minimum risk. **e.** Expected normalized MSE of OLS (for $n < p$) and minimum-norm OLS (for $p > n$) estimators using the same model with $p \in [10, 1000]$. Dots mark the dimensionalities corresponding to subplots (a–d). Dashed line: the expected normalized MSE of the optimal ridge estimator. **f.** The values of λ minimizing the expected risk. For $p \gtrsim 600$, the optimal value of ridge penalty was negative: $\lambda_{\text{opt}} < 0$. **g.** Expected normalized MSE of ridge estimators for $p = 1000$ including negative values of λ. The minimum was attained at $\lambda_{\text{opt}} = -150$.

gain compared to the minimum-norm OLS estimator became smaller and smaller and in the $p \gg n$ regime eventually disappeared. Moreover, for sufficiently large values of p, the optimal regularization value λ_{opt} became negative (Figure 2f). We found it to be the case for $p \gtrsim 600$. In sufficiently large dimensionalities, the expected risk as a function of λ had a minimum not at zero (Figure 2d), but at some negative value of λ (Figure 2g). For $p = 1000$, the lowest risk was achieved at $\lambda_{\text{opt}} = -150$ (Figure 2g).

To investigate this further, we found the optimal regularization value λ_{opt} for different sample sizes $n \in [10, 100]$ and different dimensionalities $p \in [20, 1000]$ (Figure 3). For the spherical covariance matrix ($\rho = 0$), λ_{opt} did not depend on the sample sizes and grew linearly with dimensionality (Figure 3a), in agreement with the analytical formula $\lambda_{\text{opt}} = p\sigma^2/\|\boldsymbol{\beta}\|^2 = p/\alpha$ (Nakkiran et al., 2020b). But in our model with $\rho = 0.1$, for any given sample size, the optimal value λ_{opt} in sufficiently high dimensionality was negative. The smallest dimensionality necessary for this to happen grew nonlinearly with the sample size (Figure 3b).

This result might appear to contradict the literature; for example, Dobriban and Wager (2018) and later Hastie et al. (2019) studied high-dimensional asymptotics of ridge regression performance

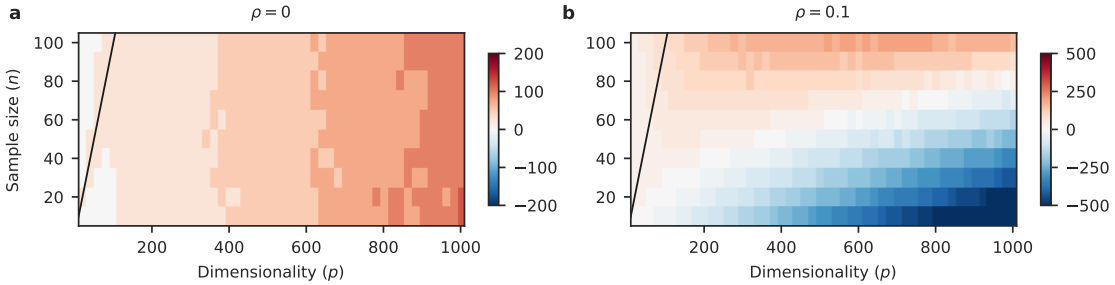

Figure 3: a. The optimal regularization parameter λ_{opt} as a function of sample size (n) and dimensionality (p) in the model with uncorrelated predictors ($\rho = 0$). In this case $\lambda_{\text{opt}} = p\sigma^2/\|\boldsymbol{\beta}\| = p/\alpha$. Black line corresponds to $n = p$. **b.** The optimal regularization parameter λ_{opt} in the model with correlated predictors ($\rho = 0.1$).

for $p, n \to \infty$ while $p/n = \gamma$ and proved, among other things, that the optimal λ is always positive. Their results hold for an arbitrary covariance matrix $\boldsymbol{\Sigma}$ when the elements of $\boldsymbol{\beta}$ are random with mean zero. The key property of our simulation is that $\boldsymbol{\beta}$ is not random and does not point in a random direction; instead, it is aligned with the first principal component (PC1) of $\boldsymbol{\Sigma}$.

While such a perfect alignment can never hold exactly in real-world data, it is plausible that $\boldsymbol{\beta}$ often points in a direction of sufficiently high predictor variance. Indeed, principal component regression (PCR) that discards all low-variance PCs and only uses high-variance PCs for prediction is known to work well for many real-world $n \ll p$ datasets (Hastie et al., 2009). In the next section we show that the low-variance PCs can provide an implicit ridge regularization.

2.4 Implicit ridge regularization provided by random low-variance predictors

Here we prove that augmenting a model with randomly generated low-variance predictors is asymptotically equivalent to the ridge shrinkage.

Theorem 1 *Let $\hat{\boldsymbol{\beta}}_\lambda$ be a ridge estimator of $\boldsymbol{\beta} \in \mathbb{R}^p$ in a linear model $y = \mathbf{x}^\top\boldsymbol{\beta} + \varepsilon$, given some training data $(\mathbf{X}, \mathbf{y})$ and some value of λ. We construct a new estimator $\hat{\boldsymbol{\beta}}_q$ by augmenting $\mathbf{X}$ with q columns $\mathbf{X}_q$ with i.i.d. elements, randomly generated with mean 0 and variance λ/q, fitting the model with minimum-norm OLS, and taking only the first p elements. Then*

$$\hat{\boldsymbol{\beta}}_q \xrightarrow[q\to\infty]{\text{a.s.}} \hat{\boldsymbol{\beta}}_\lambda.$$

In addition, for any given $\mathbf{x}$, let $\hat{y}_\lambda = \mathbf{x}^\top\hat{\boldsymbol{\beta}}_\lambda$ be the response predicted by the ridge estimator, and $\hat{y}_{\text{augm}}$ be the response predicted by the augmented model including the additional q parameters using $\mathbf{x}$ extended with q random elements (as above). Then:

$$\hat{y}_{\text{augm}} \xrightarrow[q\to\infty]{\text{a.s.}} \hat{y}_\lambda.$$

Proof Let us write $\mathbf{X}_{\text{augm}} = \begin{bmatrix}\mathbf{X} & \mathbf{X}_q\end{bmatrix}$. The minimum-norm OLS estimator can be written as

$$\hat{\boldsymbol{\beta}}_{\text{augm}} = \mathbf{X}_{\text{augm}}^\dagger\mathbf{y} = \mathbf{X}_{\text{augm}}^\top(\mathbf{X}_{\text{augm}}\mathbf{X}_{\text{augm}}^\top)^{-1}\mathbf{y}. \tag{9}$$

By the strong law of large numbers,

$$\mathbf{X}_{\text{augm}}\mathbf{X}_{\text{augm}}^\top = \mathbf{X}\mathbf{X}^\top + \mathbf{X}_q\mathbf{X}_q^\top \to \mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_n. \tag{10}$$

The first p components of $\hat{\boldsymbol{\beta}}_{\text{augm}}$ are

$$\hat{\boldsymbol{\beta}}_q = \mathbf{X}^\top(\mathbf{X}_{\text{augm}}\mathbf{X}_{\text{augm}}^\top)^{-1}\mathbf{y} \to \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_n)^{-1}\mathbf{y}. \tag{11}$$

Note that $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)\mathbf{X}^\top = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_n)$. Multiplying this equality by $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}$ on the left and $(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_n)^{-1}$ on the right, we obtain the following standard identity:

$$\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_n)^{-1} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top. \tag{12}$$

Finally:

$$\hat{\boldsymbol{\beta}}_q \to (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{y} = \hat{\boldsymbol{\beta}}_\lambda. \tag{13}$$

To prove the second statement of the Theorem, let us write $\mathbf{x}_{\text{augm}} = \begin{bmatrix}\mathbf{x}\\ \mathbf{x}_q\end{bmatrix}$. The predicted value using the augmented model is:

$$\begin{aligned}
\hat{y}_{\text{augm}} = \mathbf{x}_{\text{augm}}^\top\hat{\boldsymbol{\beta}}_{\text{augm}} &= \mathbf{x}_{\text{augm}}^\top\mathbf{X}_{\text{augm}}^\top(\mathbf{X}_{\text{augm}}\mathbf{X}_{\text{augm}}^\top)^{-1}\mathbf{y} && (14)\\
&= \begin{bmatrix}\mathbf{x}\\ \mathbf{x}_q\end{bmatrix}^\top \begin{bmatrix}\mathbf{X} & \mathbf{X}_q\end{bmatrix}^\top(\mathbf{X}\mathbf{X}^\top + \mathbf{X}_q\mathbf{X}_q^\top)^{-1}\mathbf{y} && (15)\\
&= \mathbf{x}^\top\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \mathbf{X}_q\mathbf{X}_q^\top)^{-1}\mathbf{y} + \mathbf{x}_q^\top\mathbf{X}_q^\top(\mathbf{X}\mathbf{X}^\top + \mathbf{X}_q\mathbf{X}_q^\top)^{-1}\mathbf{y} && (16)\\
&\to \mathbf{x}^\top\hat{\boldsymbol{\beta}}_\lambda + 0_{1\times n}(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_n)^{-1}\mathbf{y} && (17)\\
&= \mathbf{x}^\top\hat{\boldsymbol{\beta}}_\lambda = \hat{y}_\lambda. && (18)
\end{aligned}$$

■

Note that the Theorem requires the random predictors to be independent from each other, but does *not* require them to be independent from the existing predictors or from the response variable.

From the first statement of the Theorem it follows that the expected MSE of the truncated estimator $\hat{\boldsymbol{\beta}}_q$ converges to the expected MSE of the ridge estimator $\hat{\boldsymbol{\beta}}_\lambda$. From the second statement it follows that the expected MSE of the augmented estimator on the augmented data also converges to the expected MSE of the ridge estimator.

We extended the simulation from Section 2.3 to confirm this experimentally. We considered the same toy model as above with $n = 64$ and $p = 50$. Figure 4a (identical to Figure 2a) shows the expected MSE of ridge estimators for different values of λ. The optimal λ in this case happened to be $\lambda_{\text{opt}} = 31$. Figure 4b demonstrates that extending the model with $q \to \infty$ random predictors with variances λ_{opt}/q, using the minimum-norm OLS estimator, and truncating it at p dimensions is asymptotically equivalent to the ridge estimator with λ_{opt}. As the total number of predictors $p + q$ approached n, MSE of the extended model increased. When $p + q$ became larger than n, minimum-norm shrinkage kicked in and MSE started to decrease. As q grew even further, MSE approached the limiting value. In this case, $q \approx 200$ already got very close to the limiting performance.

As demonstrated in the proof, it is not necessary to truncate the minimum-norm estimator. The dashed line in Figure 4b shows the expected MSE of the full $(p+q)$-dimensional vector of regression coefficients. It converges slightly slower but to the same asymptotic value.

What if one does not know the value of λ_{opt} and uses random predictors with some fixed arbitrary variance to augment the model? Figure 4c shows what happens when variance is set to 1. In this case the MSE curve has a minimum at a particular q_{opt} value. This means that adding random predictors with some fixed small variance could in principle be used as an arguably bizarre but viable regularization strategy similar to ridge regression, and cross-validation could be employed to select the optimal number of random predictors.

If using random predictors as a regularization tool, one would truncate $\hat{\boldsymbol{\beta}}_{\text{augm}}$ at p dimensions (solid line in Figures 4c). The MSE values of non-truncated $\hat{\boldsymbol{\beta}}_{\text{augm}}$ (dashed line) are interesting

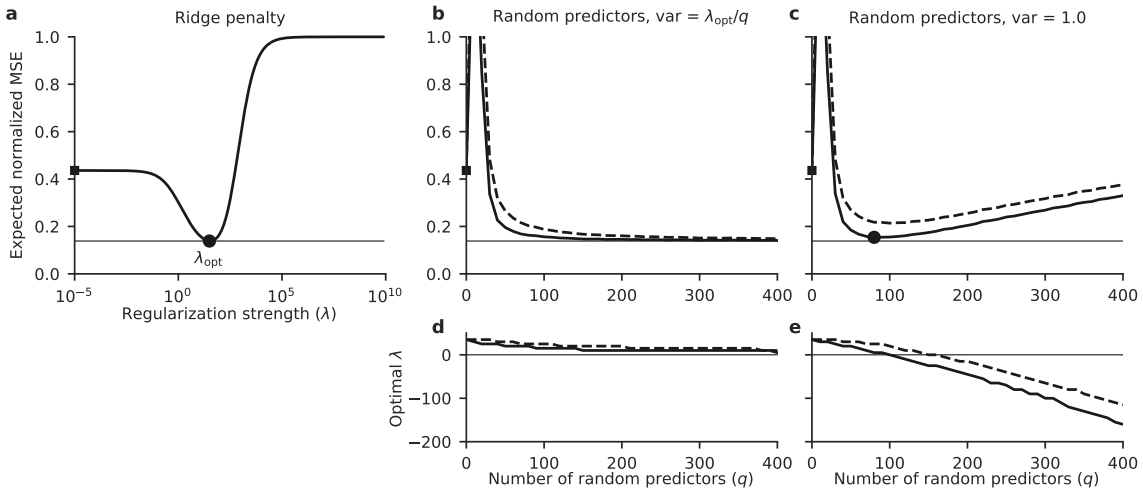

Figure 4: a. Expected MSE as a function of ridge penalty in the toy model with $p = 50$ weakly correlated predictors that are all weakly correlated with the response ($n = 64$). This is the same plot as in Figure 2a. The dot denotes minimal risk and the square denotes the MSE of the OLS estimator ($\lambda = 0$). The horizontal line shows the optimal risk corresponding to λ_{opt}. **b.** Augmenting the model with up to $q = 400$ random predictors with variance λ_{opt}/q. Solid line corresponds to $\hat{\boldsymbol{\beta}}_q$ (i.e. $\hat{\boldsymbol{\beta}}_{\text{augm}}$ truncated to p predictors); dashed line corresponds to the full $\hat{\boldsymbol{\beta}}_{\text{augm}}$. **c.** Augmenting the model with up to $q = 400$ random predictors with variance equal to 1. **d.** The optimal ridge penalty λ_{opt} in the model augmented with random predictors with adaptive variance, as in panel (b). **e.** The optimal ridge penalty λ_{opt} in the model augemented with random predictors with variance 1, as in panel (c).

because this corresponds to real-life $n \ll p$ situations such as in the `liver.toxicity` dataset discussed above. Our interpretation is that a small subset of high-variance PCs is actually predicting the dependent variable, while the large pool of low-variance PCs acts as an implicit regularizer.

In the simulations shown in Figure 4c, the parameter q controls regularization strength and there is some optimal value q_{opt} yielding minimum expected risk. If $q < q_{\text{opt}}$, this regularization is too weak and some additional ridge shrinkage with $\lambda > 0$ could be beneficial. But if $q > q_{\text{opt}}$, then the regularization is too strong and no additional ridge penalty can improve the expected risk. In this situation the expected MSE as a function of $\log(\lambda)$ will be monotonically increasing on the real line, in agreement with what we saw in Figure 2d and Figure 1b. Moreover, in this regime the expected MSE as a function of λ has a minimum at a negative value $\lambda_{\text{opt}} < 0$, as we saw in Figure 2f.

We used ridge estimators on the augmented model to demonstrate this directly. Figure 4e shows the optimal ridge penalty value λ_{opt} for each q. It crosses zero around the same value of q that yields the minimum risk with $\lambda = 0$ (Figure 4c). For larger values of q, the optimal ridge penalty λ_{opt} is negative. This shows that negative λ_{opt} is due to the over-shrinkage provided by the implicit ridge regularization arising from low-variance random predictors. It is due to *implicit over-regularization*.

2.5 Mathematical analysis for the spiked covariance model

It would be interesting to derive some sufficient conditions on $(\boldsymbol{\Sigma}, \boldsymbol{\beta}, \sigma^2, n, p)$ that would lead to $\lambda_{\text{opt}} \le 0$. One possible approach is to compute the derivative of $\mathbb{E}_{(\mathbf{X},\mathbf{y})} R(\hat{\boldsymbol{\beta}}_\lambda)$ with respect to λ at $\lambda \to 0^+$. If the derivative is positive, then $\lambda_{\text{opt}} \le 0$.

The derivative of the risk (Equation 8) can be computed as follows:

$$\frac{\partial}{\partial \lambda} R(\hat{\boldsymbol{\beta}}_\lambda) = \frac{\partial}{\partial \lambda} (\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})^\top \boldsymbol{\Sigma} (\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}) = 2(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})^\top \boldsymbol{\Sigma} \frac{\partial \hat{\boldsymbol{\beta}}_\lambda}{\partial \lambda}. \tag{19}$$

Using the standard identity $d\mathbf{A}^{-1} = -\mathbf{A}^{-1}(d\mathbf{A})\mathbf{A}^{-1}$, we get that

$$\frac{\partial \hat{\boldsymbol{\beta}}_\lambda}{\partial \lambda} = -(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-2}\mathbf{X}^\top\mathbf{y}. \tag{20}$$

Plugging this into the derivative of the risk and setting $\lambda = 0$, we obtain

$$\frac{\partial}{\partial \lambda} R(\hat{\boldsymbol{\beta}}_\lambda)\Big|_{\lambda=0} = 2\boldsymbol{\beta}^\top\boldsymbol{\Sigma}(\mathbf{X}^\top\mathbf{X})^{\dagger 2}\mathbf{X}^\top\mathbf{y} - 2\mathbf{y}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^\dagger\boldsymbol{\Sigma}(\mathbf{X}^\top\mathbf{X})^{\dagger 2}\mathbf{X}^\top\mathbf{y}, \tag{21}$$

where we denote $(\mathbf{X}^\top\mathbf{X})^{\dagger k} = \mathbf{V}\mathbf{S}^{-2k}\mathbf{V}^\top$. Remembering that $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta}$ and taking the expectation, we get

$$\begin{aligned}\frac{\partial}{\partial \lambda} \mathbb{E}_{(\mathbf{X},\mathbf{y})} R(\hat{\boldsymbol{\beta}}_\lambda)\Big|_{\lambda=0} = {}& 2\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\mathbb{E}_{\mathbf{X}}(\mathbf{X}^\top\mathbf{X})^\dagger\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top\mathbb{E}_{\mathbf{X}}(\mathbf{X}^\top\mathbf{X})^{\dagger 0}\boldsymbol{\Sigma}(\mathbf{X}^\top\mathbf{X})^\dagger\boldsymbol{\beta} \\ & - 2\sigma^2\mathbb{E}_{\mathbf{X}}\operatorname{Tr}\Big[(\mathbf{X}^\top\mathbf{X})^{\dagger 0}\boldsymbol{\Sigma}(\mathbf{X}^\top\mathbf{X})^{\dagger 2}\Big],\end{aligned} \tag{22}$$

where we used that $\mathbb{E}_{\boldsymbol{\eta}}[\mathbf{a}^\top\boldsymbol{\eta}] = 0$ and $\mathbb{E}_{\boldsymbol{\eta}}[\boldsymbol{\eta}^\top\mathbf{A}\boldsymbol{\eta}] = \sigma^2 \operatorname{Tr}[\mathbf{A}]$ for any vector $\mathbf{a}$ and any matrix $\mathbf{A}$ that are independent of $\boldsymbol{\eta}$.

We now apply this to the spiked covariance model studied above. For convenience, we write $\boldsymbol{\Sigma} = \mathbf{I} + c\boldsymbol{\beta}^\top\boldsymbol{\beta}$. Plugging this in, and denoting

$$P_k = \mathbb{E}_{\mathbf{X}}\Big[\boldsymbol{\beta}^\top(\mathbf{X}^\top\mathbf{X})^{\dagger k}\boldsymbol{\beta}\Big] = \mathbb{E}_{(\mathbf{V},\mathbf{S})}\Big[\boldsymbol{\beta}^\top\mathbf{V}\mathbf{S}^{-2k}\mathbf{V}^\top\boldsymbol{\beta}\Big], \tag{23}$$

we obtain

$$\frac{\partial}{\partial \lambda} \mathbb{E}_{(\mathbf{X},\mathbf{y})} R(\hat{\boldsymbol{\beta}}_\lambda)\Big|_{\lambda=0} = 2c\|\boldsymbol{\beta}\|^2 P_1 - 2cP_0P_1 - 2\sigma^2\mathbb{E}_{\mathbf{X}}\operatorname{Tr}(\mathbf{S}^{-4}) - 2c\sigma^2 P_2. \tag{24}$$

For the spherical covariance matrix, $c = 0$, and hence the derivative is always negative, in agreement with the fact that $\lambda_{\text{opt}} > 0$ for all $\boldsymbol{\beta}$, n, and p (Nakkiran et al., 2020b). When $c > 0$, the derivative can be positive or negative, depending on which term dominates.

We are interested in understanding the $p \gg n$ behaviour. In simulations shown above (Figures 2–4), we had $\|\boldsymbol{\beta}\|^2 = \alpha\sigma^2/(1+\rho p) = \mathcal{O}(1/p)$ and $c = \rho p/\|\boldsymbol{\beta}\|^2 = \mathcal{O}(p^2)$. For $p \gg n$ and $0 < \rho \ll 1$, all n singular values of $\mathbf{X}$ are close to $\sqrt{p}$. This makes contribution of the third term, which is the largest when $n \approx p$ due to near-zero singular values in $\mathbf{X}$, asymptotically negligible because it behaves as $\mathcal{O}(1/p^2)$. The $\boldsymbol{\beta}$ aligns with the leading singular vector in $\mathbf{V}$ and is approximately orthogonal to the others, meaning that $P_k = \|\boldsymbol{\beta}\|^2\mathcal{O}(1/p^k) = \mathcal{O}(1/p^{k+1})$. Putting everything together, we see that the first, the second, and the fourth term, all behave as $\mathcal{O}(1/p)$.

The fourth term is roughly α/ρ times smaller than the first two. In our simulations $\alpha/\rho = 100$, making the fourth term asymptotically negligible. The first two terms have identical asymptotic behaviour, however the first one is always larger because $P_0 < \|\boldsymbol{\beta}\|^2$. This makes the overall sum asymptotically positive, proving that $\lambda_{\text{opt}} \le 0$ in the $p \to \infty$ limit.

We numerically computed the derivative using Equation 24 and averaging over $N_{\text{rep}} = 100$ random training set matrices $\mathbf{X}$ to approximate the expectation values (Figure 5). This confirmed that the derivative was negative and hence $\lambda_{\text{opt}} \le 0$ for $p \gtrsim 600$, in agreement with Figure 2f.

The above results were obtained in the $p \to \infty$ limit, while holding n constant. We hypothesize that for any given value of c in the spiked covariance model, there is some n/p ratio for which $\lambda_{\text{opt}} = 0$ when both $n, p \to \infty$. This remains a question for future work. [Two manuscripts investigating this question in a more general setting (Richards et al., 2020; Wu and Xu, 2020) appeared while our paper was in press.]

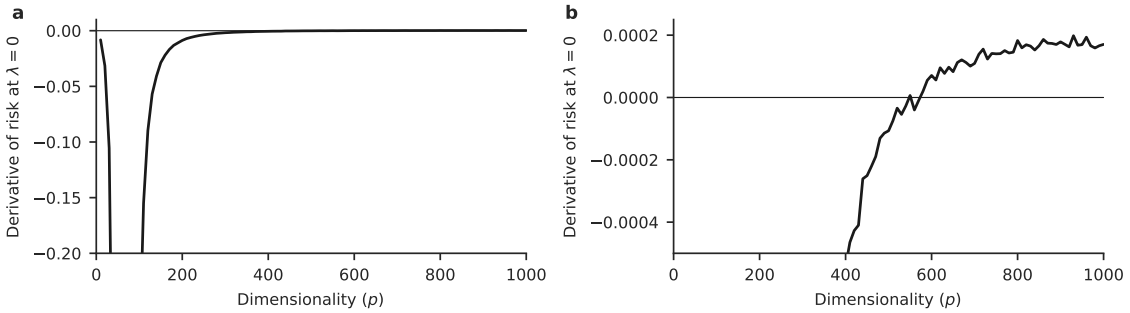

Figure 5: a. The derivative of the expected risk as a function of ridge penalty λ at $\lambda = 0$, in the model with p weakly correlated predictors (Eq. 24). Sample size $n = 64$. **b.** Zoom-in into panel (a). The derivative becomes positive for $p \gtrsim 600$, implying that $\lambda_{\text{opt}} \leq 0$.

2.6 Implicit over-regularization using random Fourier features on MNIST

For our final example, we used the setup from Nakkiran et al. (2020a,b), and asked whether the same phenomenon ($\lambda_{\text{opt}} < 0$) can be observed using random Fourier features on MNIST.

We normalized all pixel intensity values to lie between -1 and 1, and transformed the $28 \times 28 = 784$ pixel features into 2000 random Fourier features by drawing a random matrix $\mathbf{W} \in \mathbb{R}^{784\times 1000}$ with all elements i.i.d. from $\mathcal{N}(0, \sigma = 0.1)$, computing $\exp(-i\mathbf{XW})$, and taking its real and imaginary parts as separate features. This procedure approximates kernel regression with the Gaussian kernel, and standard deviation of the $\mathbf{W}$ elements corresponds to the standard deviation of the kernel (Rahimi and Recht, 2008). We used $n = 64$ randomly selected images as a training set, and used the MNIST test set with 10 000 images to compute the risk. We used the digit value (from 0 to 9) as the response variable y, with squared error loss function. The model included the intercept which was not penalized. To estimate the expected risk, we averaged the risks over $N_{\text{rep}} = 100$ random draws of training sets.

We found that the expected risk was minimized at $\lambda_{\text{opt}} \approx -80$, when the expectation was computed across all 80/100 training sets that had the smallest singular value $s^2_{\min} > 100$ (Figure 6). For any given training set, the risk diverged at $\lambda = s_{\min}$, and the smallest singular value that we observed across 100 draws was $s^2_{\min} = 40$. The average risk across 20 samples with $s^2_{\min} < 100$ had multiple diverging peaks for $\lambda \in [-100, 0]$ (Figure 6b, dashed line).

The derivative of risk with respect to λ at $\lambda = 0$ that we computed in the previous section can be formally understood as the derivative at $\lambda \to 0^+$. Negative derivative implies that $\lambda = 0$ yields better expected risk than any positive value. However, if the generative process allows singular values of $\mathbf{X}$ to become arbitrarily small, then $\lambda < 0$ can possibly yield diverging *expected* risk. That said, for any given training set, the risk will not diverge for $\lambda \in (-s^2_{\min}, \infty)$ and the minimal *conditional* risk (conditioned on the training set) can be attained at $\lambda_{\text{opt}} < 0$. Indeed, in our MNIST example, the average across all 100 training sets was monotonically decreasing until $\lambda \approx -35$ (Figure 6b).

3. Discussion

Summary and related work

We have demonstrated that the minimum-norm OLS interpolating estimator tends to work well in the $n \ll p$ situation and that a positive ridge penalty can fail to provide any further improvement. This is because the large pool of low-variance predictors (or principal components of predictors), together with the minimum-norm requirement, can perform sufficient shrinkage on its own. This phenomenon goes against the conventional wisdom (see Introduction) but is in line with the large

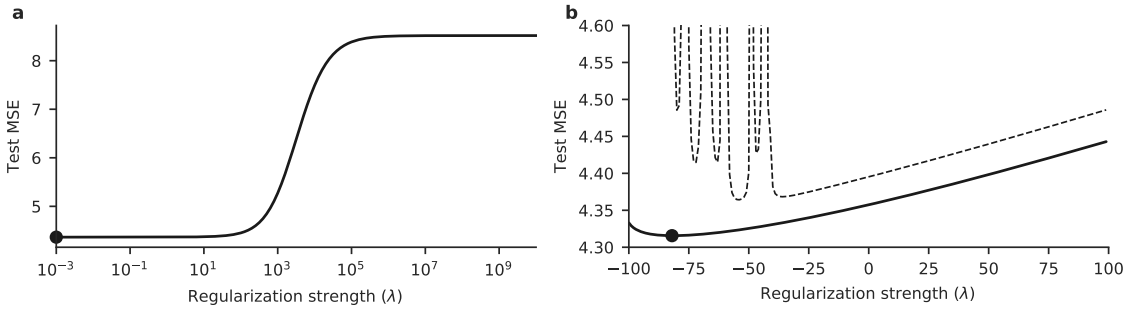

Figure 6: a. Expected risk of ridge regression on MNIST data using random Fourier features as predictors and digit value as the response. Sample size $n = 64$, number of Fourier features $p = 2000$. When $\lambda \in \mathbb{R}^+$, the risk is minimized at $\lambda = 0$. **b.** When λ is allowed to take negative values, the risk is minimized at $\lambda \approx -80$, across all training sets with $s^2_{\min} > 100$ (solid line; the average over 80/100 cases). Training sets with $s^2_{\min} < 100$ had diverging risk around $\lambda = -s^2_{\min}$ (dashed line; the average over 20/100 cases).

body of ongoing research kindled by Zhang et al. (2017) and mostly done in parallel to our work (Advani et al., 2020; Spigler et al., 2019; Belkin et al., 2018a,b, 2019a,b,c; Nakkiran, 2019; Nakkiran et al., 2020a,b; Liang and Rakhlin, 2020; Hastie et al., 2019; Bartlett et al., 2020; Chinot and Lerasle, 2020; Muthukumar et al., 2020; Mei and Montanari, 2019; Bibas et al., 2019; Dereziński et al., 2019; Negrea et al., 2019). See Introduction for more context.

We stress that the minimum-norm OLS estimator $\hat{\boldsymbol{\beta}}_0 = \mathbf{X}^\dagger \mathbf{y}$ is not an exotic concept. It is given by exactly the same formula as the standard OLS estimator when the latter is written in terms of the pseudoinverse of the design matrix: $\hat{\boldsymbol{\beta}}_{\text{OLS}} = \mathbf{X}^\dagger \mathbf{y}$. When dealing with an under-determined problem, statistical software will often output the minimum-norm OLS estimator by default.

That positive ridge penalty can fail to improve the estimator risk has been observed for kernel regression (Liang and Rakhlin, 2020) and for random features linear regression (Mei and Montanari, 2019). Our results show that this can also happen in a simpler situation of ridge regression with Gaussian features. Our contribution is to use the spiked covariance model to demonstrate and analyze this phenomenon. Moreover, we showed that the optimal ridge penalty in this situation can be negative.

In their seminal paper on ridge regression, Hoerl and Kennard (1970) proved that there always exists some $\lambda_{\text{opt}} > 0$ that yields a lower MSE than $\lambda = 0$. However, their proof was based on the assumption that $\mathbf{X}^\top \mathbf{X}$ is full rank, i.e. $n > p$. When the predictor covariance $\boldsymbol{\Sigma}$ is spherical, λ_{opt} is also always positive, for any n and p (Nakkiran et al., 2020a). Similarly, Dobriban and Wager (2018) and later Hastie et al. (2019) proved that $\lambda_{\text{opt}} > 0$ for any $\boldsymbol{\Sigma}$ in the asymptotic $p, n \to \infty$ case while $p/n = \gamma$, based on the assumption that $\boldsymbol{\beta}$ is randomly oriented. Here we argue that real-world $n \ll p$ problems can demonstrate qualitatively different behaviour with $\lambda_{\text{opt}} \le 0$. This happens when $\boldsymbol{\Sigma}$ is not spherical and $\boldsymbol{\beta}$ is pointing in its high-variance direction. This interpretation is related to the findings of Bibas et al. (2019) and Bartlett et al. (2020).

Augmenting the samples vs. augmenting the predictors

It is well-known that ridge estimator can be obtained as an OLS estimator on the augmented data:

$$\mathcal{L}_\lambda = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 = \left\| \begin{bmatrix} \mathbf{y} \\ \mathbf{0}_{p\times 1} \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I}_{p\times p} \end{bmatrix} \boldsymbol{\beta} \right\|^2. \tag{25}$$

While for this standard trick, both $\mathbf{X}$ and $\mathbf{y}$ are augmented with p additional *rows*, in this manuscript we considered augmenting $\mathbf{X}$ alone with q additional *columns*.

At the same time, from the above formula and from the proof of Theorem 1, we can see that if $\mathbf{y}$ is augmented with q additional zeros and $\mathbf{X}$ is augmented with q additional rows with all elements having zero mean and variance λ/q, then the resulting estimator will converge to $\hat{\boldsymbol{\beta}}_\lambda$ when $q \to \infty$. This means that augmenting $\mathbf{X}$ with q random samples and using OLS is very similar to augmenting it with q random predictors and using minimum-norm OLS.

More generally, it is known that corrupting $\mathbf{X}$ with noise in various ways [e.g. additive noise (Bishop, 1995) or multiplicative noise (Srivastava et al., 2014)] can be equivalent to adding the ridge penalty. Augmenting $\mathbf{X}$ with random predictors can also be seen as a way to corrupt $\mathbf{X}$ with noise.

Minimum-norm estimators in other statistical methods

Several statistical learning methods use optimization problems similar to the minimum-norm OLS:

$$\min\|\boldsymbol{\beta}\|_2 \text{ subject to } \mathbf{y} = \mathbf{X}\boldsymbol{\beta}. \tag{26}$$

One is the linear support vector machine classifier for linearly separable data, known to be *maximum margin* classifier (here $y_i \in \{-1, 1\}$) (Vapnik, 1996):

$$\min\|\boldsymbol{\beta}\|_2 \text{ subject to } y_i(\boldsymbol{\beta}^\top \mathbf{x}_i + \beta_0) \geq 1 \text{ for all } i. \tag{27}$$

Another is basis pursuit (Chen et al., 2001):

$$\min\|\boldsymbol{\beta}\|_1 \text{ subject to } \mathbf{y} = \mathbf{X}\boldsymbol{\beta}. \tag{28}$$

Both of them are more well-known and more widely applied in *soft* versions where the constraint is relaxed to hold only approximately. In case of support vector classifiers, this corresponds to the soft-margin version applicable to non-separable datasets. In case of basis pursuit, this corresponds to basis pursuit denoising (Chen et al., 2001), which is equivalent to lasso (Tibshirani, 1996). The Dantzig selector (Candes and Tao, 2007) also minimizes $\|\boldsymbol{\beta}\|_1$ subject to $\mathbf{y} \approx \mathbf{X}\boldsymbol{\beta}$, but uses ℓ_∞-norm approximation instead of the ℓ_2-norm. In contrast, our manuscript considers the case where constraint $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ is satisfied exactly.

In the classification literature, it has for a long time been a common understanding that the maximum margin linear classifier is a good choice for linearly separable problems (which is the case when $n < p$). When using the hinge loss as in Equation 27, maximum margin is equivalent to minimum norm, so from this point of view good performance of the minimum-norm OLS estimator is not unreasonable. However, when using quadratic loss as we do in this manuscript, minimum norm (for a binary y) is *not* equivalent to maximum margin; and for a continuous y the concept of margin does not apply at all. Still, the intuition remains the same: minimum norm requirement performs implicit regularization.

Minimum-norm estimator with kernel trick

Minimum-norm OLS estimator can be easily kernelized. Indeed, if $\mathbf{x}_{\text{test}}$ is some test point, then

$$\hat{y}_{\text{test}} = \mathbf{x}_{\text{test}}^\top \hat{\boldsymbol{\beta}}_0 = \mathbf{x}_{\text{test}}^\top \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{y} = \mathbf{k}^\top \mathbf{K}^{-1}\mathbf{y}, \tag{29}$$

where $\mathbf{K} = \mathbf{X}\mathbf{X}^\top$ is a $n \times n$ matrix of scalar products between all training points and $\mathbf{k} = \mathbf{X}\mathbf{x}_{\text{test}}$ is a vector of scalar products between all training points and the test point. The *kernel trick* consists of replacing all scalar products with arbitrary kernel functions. As an example, Gaussian kernel corresponds to the effective dimensionality $p = \infty$ and so trivially $n \ll p$ for any n. How exactly our results extend to such $p = \infty$ situations is an interesting question beyond the scope of this paper. It has been shown that Gaussian kernel can achieve impressive accuracy on MNIST and CIFAR10 data without any explicit regularization (Zhang et al., 2017; Belkin et al., 2018b; Liang and Rakhlin, 2020) and that positive ridge regularization decreases the performance (Liang and Rakhlin, 2020).

Minimum-norm estimator via gradient descent

In the $n < p$ situation, if gradient descent is initialized at $\boldsymbol{\beta} = 0$ then it will converge to the minimum-norm OLS solution (Zhang et al., 2017; Wilson et al., 2017) [see also Soudry et al. (2018) and Poggio et al. (2017) for the case of logistic loss]. Indeed, each update step is proportional to $\nabla_{\boldsymbol{\beta}} \mathcal{L} = \mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ and so lies in the row space of $\mathbf{X}$, meaning that the final solution also has to lie in the row space of $\mathbf{X}$ and hence must be equal to $\hat{\boldsymbol{\beta}}_0 = \mathbf{X}^\dagger \mathbf{y} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{y}$. If initial value of $\boldsymbol{\beta}$ is not exactly 0 but sufficiently close, then the gradient descent limit might be close enough to $\hat{\boldsymbol{\beta}}_0$ to work well.

Zhang et al. (2017) hypothesized that this property of gradient descent can shed some light on the remarkable generalization capabilities of deep neural networks. They are routinely trained with the number of model parameters p greatly exceeding n, meaning that such a network can be capable of perfectly fitting any training data; nevertheless, test-set performance can be very high. Moreover, increasing network size p can improve test-set performance even after p is large enough to ensure zero training error (Neyshabur, 2017; Nakkiran et al., 2020a), which is qualitatively similar to what we observed here. It has also been shown that in the $p \gg n$ regime, the ridge (or early stopping) regularization does not noticeably improve the generalization performance (Nakkiran et al., 2020b).

Our work focused on *why* the minimum-norm OLS estimator performs well. We confirmed its generalization ability and clarified the situations in which it can arise. Our results do not explain the case of highly nonlinear under-determined models such as deep neural networks, but perhaps can provide an inspiration for future work in that direction.

Acknowledgements

This paper arose from the online discussion at `https://stats.stackexchange.com/questions/328630` in February 2018; JL and BS answered the question asked by DK. We thank all other participants of that discussion, in particular `@DikranMarsupial` and `@guy` for pointing out several important analogies. We thank Ryan Tibshirani for a very helpful discussion and Philipp Berens for comments and support. We thank anonymous reviewers for suggestions that strongly improved the paper. DK was financially supported by the German Excellence Strategy (EXC 2064; 390727645), the Federal Ministry of Education and Research (FKZ 01GQ1601) and the National Institute of Mental Health of the National Institutes of Health under Award Number U19MH114830. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

M. S. Advani, A. M. Saxe, and H. Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 2020.

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

M. Belkin, D. J. Hsu, and P. Mitra. Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. In *Advances in Neural Information Processing Systems*, pages 2300–2311, 2018a.

M. Belkin, S. Ma, and S. Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, 2018b.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32): 15849–15854, 2019a.

M. Belkin, D. Hsu, and J. Xu. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019b.

M. Belkin, A. Rakhlin, and A. B. Tsybakov. Does data interpolation contradict statistical optimality? In *International Conference on Artificial Intelligence and Statistics*, 2019c.

K. Bibas, Y. Fogel, and M. Feder. A new look at an old problem: A universal learning approach to linear regression. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 2304–2308. IEEE, 2019.

C. M. Bishop. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7 (1):108–116, 1995.

P. R. Bushel, R. D. Wolfinger, and G. Gibson. Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. *BMC Systems Biology*, 1(1):15, 2007.

E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

G. Chinot and M. Lerasle. Benign overfitting in the large deviation regime. *arXiv preprint arXiv:2003.05838*, 2020.

M. Dereziński, F. Liang, and M. W. Mahoney. Exact expressions for double descent and implicit regularization via surrogate random design. *arXiv preprint arXiv:1912.04533*, 2019.

E. Dobriban and S. Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: the Lasso and Generalizations*. CRC press, 2015.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning*, volume 112. Springer, 2013.

D. Kobak, Y. Bernaerts, M. A. Weis, F. Scala, A. Tolias, and P. Berens. Sparse reduced-rank regression for exploratory visualization of paired multivariate datasets. *bioRxiv*, page 302208, 2018.

T. Liang and A. Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48(3):1329–1347, 2020.

S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

V. Muthukumar, K. Vodrahalli, V. Subramanian, and A. Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 2020.

P. Nakkiran. More data can hurt for linear regression: Sample-wise double descent. *arXiv preprint arXiv:1912.07242*, 2019.

P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020a.

P. Nakkiran, P. Venkat, S. Kakade, and T. Ma. Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*, 2020b.

J. Negrea, G. K. Dziugaite, and D. M. Roy. In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. *arXiv preprint arXiv:1912.04265*, 2019.

B. Neyshabur. *Implicit Regularization in Deep Learning*. PhD thesis, Toyota Technological Institute at Chicago, 2017.

T. Poggio, K. Kawaguchi, Q. Liao, B. Miranda, L. Rosasco, X. Boix, J. Hidary, and H. Mhaskar. Theory of deep learning III: explaining the non-overfitting puzzle. *arXiv preprint arXiv:1801.00173*, 2017.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2008.

D. Richards, J. Mourtada, and L. Rosasco. Asymptotics of ridge (less) regression under general source condition. *arXiv preprint arXiv:2006.06386*, 2020.

F. Rohart, B. Gautier, A. Singh, and K.-A. Le Cao. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Computational Biology*, 13(11):e1005752, 2017.

D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

S. Spigler, M. Geiger, S. d'Ascoli, L. Sagun, G. Biroli, and M. Wyart. A jamming transition from under-to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 52(47):474001, 2019.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15 (1):1929–1958, 2014.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

A. N. Tikhonov. On the solution of ill-posed problems and the method of regularization. *Dokl. Akad. Nauk SSSR*, 151(3):501–504, 1963.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1996.

A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, pages 4151–4161, 2017.

D. Wu and J. Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *arXiv preprint arXiv:2006.05800*, 2020.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.