# Lecture notes on the inductive biases of high-dimensional ridge regression

Jacob A. Zavatone-Veth*

October 6, 2024

**Abstract**

These are notes for lectures given as part of Harvard Applied Math 226 on 30 September and 2 October 2024. They are not intended to be entirely complete or entirely rigorous.

## Contents



*Society of Fellows and Center for Brain Science
Harvard University
Cambridge, MA 02138
jzavatoneveth@fas.harvard.edu

# 1 Introduction to ridge regression

In these lecture notes, we will consider the generalization error of ridge regression in high dimensions. First, some introductory comments are in order. Namely, what is ridge regression, and why should we care about it?

## 1.1 Interpolation

We will present three perspectives on why one should care about ridge regression. The first perspective is that of simple linear interpolation. Suppose one has a system of $p$ equations in $d$ variables, written in matrix form as

$$\mathbf{X}\mathbf{w} = \mathbf{y} \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{p\times d}$ is the design matrix, $\mathbf{y} \in \mathbb{R}^p$ is the target, and $\mathbf{w} \in \mathbb{R}^d$ is the vector of trainable parameters. For simplicity, assume that $\mathbf{X}$ is full rank.

Ridge regression unifies the standard solutions to this problem as limiting cases of the same estimator. Let

$$\hat{R}(\mathbf{w}) = \frac{1}{p}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \tag{2}$$

be the mean-squared error of a candidate solution to this linear system, and consider the ridge estimator

$$\mathbf{w} = \arg\min_{\mathbf{w}} \hat{R}(\mathbf{w}) + \lambda\|\mathbf{w}\|^2. \tag{3}$$

One can easily see that

$$\mathbf{w} = \frac{1}{p}\left(\frac{1}{p}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y}. \tag{4}$$

In the overdetermined regime $p > d$, $\mathbf{X}^\top\mathbf{X}$ is invertible, and we can directly take the $\lambda \downarrow 0$ limit to obtain the ordinary least-squares solution

$$\mathbf{w} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}. \tag{5}$$

In the underdetermined regime $p < d$ we can apply the push-through identity to write

$$\mathbf{w} = \frac{1}{p}\mathbf{X}^\top\left(\frac{1}{p}\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}\right)^{-1}\mathbf{y}, \tag{6}$$

whence as $\mathbf{X}\mathbf{X}^\top$ is invertible in this regime we can pass to the limit $\lambda \downarrow 0$ to obtain the minimum-norm solution

$$\mathbf{w} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{y}. \tag{7}$$

## 1.2 Dynamics

Consider the gradient flow

$$\frac{d}{dt}\mathbf{w} = -\nabla_{\mathbf{w}}\hat{R}(\mathbf{w}) = -\frac{1}{p}\mathbf{X}^\top\mathbf{X}\mathbf{w} + \frac{1}{p}\mathbf{X}^\top\mathbf{y} \tag{8}$$

as would result from doing gradient descent on the least-squares cost introduced above. If one starts from $\mathbf{w} = \mathbf{0}$—as is true in the NTK setting—then it is not hard to show that the fixed point of these dynamics will be the minimum-norm interpolant. This holds also in discrete time [1].

## 1.3 Bayesian inference

A final perspective comes from Bayesian inference. Suppose that we assume a likelihood

$$\mathbf{y} \mid \mathbf{X}, \mathbf{w} = \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}) \tag{9}$$

and a prior

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2\mathbf{I}). \tag{10}$$

Then, one can show with a bit of algebra that the posterior mean—and thus the minimum mean-squared error (MMSE) estimator is the ridge estimator with parameter

$$\lambda = \frac{\sigma^2}{\sigma_0^2}. \tag{11}$$

# 2 Generalization

We now turn to the question of generalization, *i.e.*, the error of prediction for unseen inputs.

## 2.1 Assumptions on data

To make sense of this question, we must either make some assumption on the data forming our regressor's world, or commit to the pessimism of worst-case analysis (that is, to see what the maximum possible error would be for an adversarially-chosen input).

As a simple but non-trivial model, we will assume that the world is Gaussian (we will later return to the question of how strong this assumption actually is in high dimensions). That is, we assume covariates are drawn as

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \tag{12}$$

for a positive-definite symmetric covariance matrix $\mathbf{\Sigma}$ with bounded spectral norm. From these covariates we want to predict targets given by a fixed linear projection of $\mathbf{x}$ plus independent noise:

$$y \,|\, \mathbf{x} \sim \mathcal{N}(\langle \mathbf{w}_*, \mathbf{x} \rangle, \eta^2) \tag{13}$$

where $\mathbf{w}_* \in \mathbb{R}^d$ is a fixed vector. In other words, we can write

$$y \overset{d}{=} \langle \mathbf{w}_*, \mathbf{x} \rangle + \epsilon \tag{14}$$

where the noise $\epsilon \sim \mathcal{N}(0, \eta^2)$ is independent of $\mathbf{x}$.

We assume that we have access to a training dataset of $p$ i.i.d. samples from this distribution. Using this dataset, we compute the ridge estimate

$$\mathbf{w} = \frac{1}{p}(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} \tag{15}$$

where we collect the covariates into the design matrix $\mathbf{X} \in \mathbb{R}^{p \times d}$ and the vector $\mathbf{y} \in \mathbb{R}^p$, and define the empirical covariance matrix

$$\hat{\mathbf{\Sigma}} = \frac{1}{p} \mathbf{X}^\top \mathbf{X}. \tag{16}$$

Collecting the noise samples into a vector $\boldsymbol{\epsilon} \in \mathbb{R}^p$, we can expand the estimate into a "signal" term depending on $\mathbf{w}_*$ and a "noise" term depending on $\boldsymbol{\epsilon}$:

$$\mathbf{w} = (\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-1} \hat{\mathbf{\Sigma}} \mathbf{w}_* + \frac{1}{p}(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \boldsymbol{\epsilon}. \tag{17}$$

It is useful to re-write this in terms of the difference between the true signal and the estimate:

$$\mathbf{w}_* - \mathbf{w} = \lambda (\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-1} \mathbf{w}_* - \frac{1}{p}(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \boldsymbol{\epsilon} \tag{18}$$

We can now define the generalization error: draw a test point $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, and predict the corresponding target $y$ using $\langle \mathbf{w}, \mathbf{x} \rangle$. We measure the error of this prediction using the mean-squared error averaged over the choice of test point *and* the training noise, *i.e.*, we define

$$R(\mathbf{w}) = \mathbb{E}_{\boldsymbol{\epsilon}} \mathbb{E}_{(\mathbf{x},y)} (\langle \mathbf{w}, \mathbf{x} \rangle - y)^2 = \mathbb{E}_{\boldsymbol{\epsilon}} (\mathbf{w}_* - \mathbf{w})^\top \mathbf{\Sigma} (\mathbf{w}_* - \mathbf{w}) + \eta^2. \tag{19}$$

Substituting in the expression for $\mathbf{w}$ and evaluating the $\boldsymbol{\epsilon}$ average, one finds that

$$R = \lambda^2 \mathbf{w}_*^\top (\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-1} \mathbf{\Sigma} (\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-1} \mathbf{w}_* + \frac{\eta^2}{p} \operatorname{tr}[(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-2} \hat{\mathbf{\Sigma}} \mathbf{\Sigma}] + \eta^2 \tag{20}$$

You will show this on the homework.

We can further simplify this by noting that

$$(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-1} \mathbf{\Sigma} (\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-1} = -\frac{\partial}{\partial J} (\hat{\mathbf{\Sigma}} + J \mathbf{\Sigma} + \lambda \mathbf{I})^{-1} \Big|_{J=0} \tag{21}$$

and

$$(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-2} \hat{\mathbf{\Sigma}} = -\frac{\partial}{\partial \lambda} (\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-1} \hat{\mathbf{\Sigma}} \tag{22}$$

$$= \frac{\partial}{\partial \lambda} [\lambda (\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-1}] \tag{23}$$

which allows us to write $R$ in terms of the resolvent $(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-1}$ and its shifted counterpart $(\hat{\mathbf{\Sigma}} + J\mathbf{\Sigma} + \lambda \mathbf{I})^{-1}$.

## 2.2 The form of the ridge estimator and the classical statistics limit

At this point a few comments are in order. First, the matrix

$$(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-1} \hat{\mathbf{\Sigma}} \tag{24}$$

can be interpreted as a sort of low-pass filter: directions corresponding to eigenvalues of $\hat{\mathbf{\Sigma}}$ much larger than $\lambda$ are unchanged, while those corresponding to small eigenvalues are discarded.

Second, consider the classical statistics limit of $p \to \infty$ for fixed $d$. There, $\hat{\mathbf{\Sigma}} \to \mathbf{\Sigma}$ by the strong law of large numbers, so

$$\lim_{p \to \infty} R = \lambda^2 \mathbf{w}_*^\top (\mathbf{\Sigma} + \lambda \mathbf{I})^{-2} \mathbf{\Sigma} \mathbf{w}_* + \eta^2. \tag{25}$$

Further taking the ridgeless limit gives

$$\lim_{\lambda \downarrow 0} \lim_{p \to \infty} R = \eta^2, \tag{26}$$

reflecting the fact that in this limit ridge regression recovers the true signal $\mathbf{w}_*$.

# 3 High-dimensional asymptotics

In the previous section, we derived a formula for the generalization error where all randomness is packaged in terms of the resolvent of the empirical covariance matrix, $(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-1}$, or slight modifications thereof. We know these resolvents behave simply in the limit $p \to \infty$ with $d$ fixed. Our task is now to study them in the richer **high-dimensional** limit

$$d, p \to \infty \quad \text{with} \quad d/p \to q \in (0, \infty). \tag{27}$$

There are many ways to derive the limiting behavior of the resolvents (and thus the generalization error), which we reviewed in [2]. Here, we leverage an important set of results from random matrix theory known as **strong deterministic equivalence**. We will use these results without proof, with the goal of highlighting their existence and utility.

## 3.1 An overview of deterministic equivalence

The key players in this story are the matrix-valued resolvent of the sample covariance:

$$\hat{\mathbf{G}}(\lambda) = (\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-1} \tag{28}$$

and its tracial counterpart

$$\hat{g}(\lambda) = \frac{1}{d} \operatorname{tr}[\hat{\mathbf{G}}(\lambda)], \tag{29}$$

along with the corresponding quantities for the population covariance,

$$\mathbf{G}(\lambda) = (\mathbf{\Sigma} + \lambda \mathbf{I})^{-1} \tag{30}$$

and

$$g(\lambda) = \frac{1}{d} \operatorname{tr}[\mathbf{G}(\lambda)]. \tag{31}$$

To define the high-dimensional limit, we assume that we have a sequence of population covariance matrices $\mathbf{\Sigma}$, indexed by the dimension $d$, such that $\lim_{d\to\infty} g(\lambda)$ is well-defined. We will not be too precise here.

A (perhaps *the*) foundational result of modern random matrix theory is the **Marchenko–Pastur theorem**, which can be stated as a *weak* deterministic equivalent for $\hat{g}(\lambda)$: In the high-dimensional limit, one has

$$\lim_{\substack{d,p\to\infty \\ d/p\to q}} \hat{g}(\lambda) = \lim_{d\to\infty} g(\kappa) \tag{32}$$

almost surely, where $\kappa$ is the unique positive solution to

$$\kappa = \frac{\lambda}{1 - q\frac{1}{d}\operatorname{tr}[(\mathbf{\Sigma} + \kappa \mathbf{I})^{-1}\mathbf{\Sigma}]} = \frac{\lambda}{1 - q[1 - \kappa g(\kappa)]}. \tag{33}$$

That is, in the high-dimensional limit the traced empirical resolvent at a given ridge $\lambda$ is equivalent to the traced population resolvent at a sample-size-dependent *renormalized* ridge $\kappa$. From this deterministic equivalent for the resolvent one can infer the limiting distribution of eigenvalues of $\hat{\mathbf{\Sigma}}$; see [2] for details.

Strong deterministic equivalence is a wide-ranging extension of this idea to the matrix-valued resolvent. Essentially, it says that the same equivalence holds for the matrix-valued resolvent so long as one promises to query the matrix in certain ways. We adopt a somewhat weak but useful definition of strong deterministic equivalence following Bach [3]: For two sequences of (possibly random) matrices $\mathbf{A}$ and $\mathbf{B}$ indexed by their dimension $d$, we say that

$$\mathbf{A} \sim \mathbf{B} \tag{34}$$

if

$$\lim_{d\to\infty} \frac{\operatorname{tr}(\mathbf{AM})}{\operatorname{tr}(\mathbf{BM})} = 1 \tag{35}$$

for all sequences $\mathbf{M}$ of test matrices with bounded spectral norm, where the limit is in probability.

With this definition, one has

$$\hat{\mathbf{G}}(\lambda) \sim \frac{\kappa}{\lambda}\mathbf{G}(\kappa), \tag{36}$$

where $\kappa$ is given as above. As in the weak deterministic equivalent, this has a natural interpretation as the random fluctuations in the empirical covariance renormalizing the ridge [2].

## 3.2 The asymptotic generalization error

With these deterministic equivalents in hand, we show in the sequel that

$$R \sim \frac{\kappa^2}{1-\gamma}\mathbf{w}_*^\top(\mathbf{\Sigma}+\kappa\mathbf{I})^{-2}\mathbf{\Sigma}\mathbf{w}_* + \eta^2\frac{\gamma}{1-\gamma} + \eta^2 \tag{37}$$

where, as before, $\kappa$ is the unique positive solution to

$$\kappa = \frac{\lambda}{1 - q\frac{1}{d}\operatorname{tr}[(\mathbf{\Sigma}+\kappa\mathbf{I})^{-1}\mathbf{\Sigma}]} \tag{38}$$

and

$$\gamma = 1 - \frac{1}{\partial\kappa/\partial\lambda} = q\frac{1}{d}\operatorname{tr}[(\mathbf{\Sigma}+\kappa\mathbf{I})^{-2}\mathbf{\Sigma}^2]. \tag{39}$$

This derivation roughly follows Atanasov et al. [2], though our approach to the signal term here is slightly different.

## 3.3 Deterministic equivalent for the signal term

Here, we describe how to derive a deterministic equivalent for the signal term using a shifted resolvent. Our goal is to show that

$$\lambda^2(\hat{\mathbf{\Sigma}}+\lambda\mathbf{I})^{-1}\mathbf{\Sigma}(\hat{\mathbf{\Sigma}}+\lambda\mathbf{I})^{-1} \sim \frac{\kappa^2}{1-\gamma}(\mathbf{\Sigma}+\kappa\mathbf{I})^{-2}\mathbf{\Sigma}. \tag{40}$$

One approach, as in Atanasov et al. [2], is to study the shifted resolvent $(\hat{\mathbf{\Sigma}} + J\mathbf{\Sigma} + \lambda\mathbf{I})^{-1}$ directly. Another approach[1] is to observe that we can also write it in terms of the resolvent of a sample covariance matrix with re-shaped population covariance:

$$(\hat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-1}\mathbf{\Sigma}(\hat{\mathbf{\Sigma}} + \lambda\mathbf{I})^{-1} = -\frac{1}{\lambda}\frac{\partial}{\partial J}(\hat{\mathbf{\Sigma}} + \lambda J\mathbf{\Sigma} + \lambda\mathbf{I})^{-1}\Big|_{J=0} \tag{41}$$

$$= -\frac{1}{\lambda}\frac{\partial}{\partial J}(\mathbf{I} + J\mathbf{\Sigma})^{-1/2}(\hat{\mathbf{\Sigma}}_J + \lambda\mathbf{I})^{-1}(\mathbf{I} + J\mathbf{\Sigma})^{-1/2}\Big|_{J=0}, \tag{42}$$

where the matrix

$$\hat{\mathbf{\Sigma}}_J = (\mathbf{I} + J\mathbf{\Sigma})^{-1/2}\hat{\mathbf{\Sigma}}(\mathbf{I} + J\mathbf{\Sigma})^{-1/2} \tag{43}$$

is the sample covariance for data with a modified population covariance

$$\mathbf{\Sigma}_J = (\mathbf{I} + J\mathbf{\Sigma})^{-1/2}\mathbf{\Sigma}(\mathbf{I} + J\mathbf{\Sigma})^{-1/2}. \tag{44}$$

From this, we can use the result for Wishart matrices stated above to obtain

$$(\hat{\mathbf{\Sigma}}_J + \lambda\mathbf{I})^{-1} \sim \frac{\kappa_J}{\lambda}(\mathbf{\Sigma}_J + \kappa_J\mathbf{I})^{-1} \tag{45}$$

where $\kappa_J$ is the unique positive solution to

$$\kappa_J = \frac{\lambda}{1 - q\frac{1}{d}\operatorname{tr}[(\mathbf{\Sigma}_J + \kappa_J\mathbf{I})^{-1}\mathbf{\Sigma}_J]}. \tag{46}$$

Expanding out the form of $\mathbf{\Sigma}_J$, we find that

$$(\mathbf{I} + J\mathbf{\Sigma})^{-1/2}(\hat{\mathbf{\Sigma}}_J + \lambda\mathbf{I})^{-1}(\mathbf{I} + J\mathbf{\Sigma})^{-1/2} \sim \frac{\kappa_J}{\lambda}(\mathbf{\Sigma} + \kappa_J J\mathbf{\Sigma} + \kappa_J\mathbf{I})^{-1} \tag{47}$$

where the self-consistent equation becomes

$$\kappa_J = \frac{\lambda}{1 - q\frac{1}{d}\operatorname{tr}[(\mathbf{\Sigma} + \kappa_J J\mathbf{\Sigma} + \kappa_J\mathbf{I})^{-1}\mathbf{\Sigma}]}. \tag{48}$$

We now must evaluate the derivatives. First, we observe that $\kappa_{J=0} = \kappa$ for $\kappa$ without the subscript defined as before. Next, we compute

$$-\frac{\partial}{\partial J}\kappa_J(\mathbf{\Sigma} + \kappa_J J\mathbf{\Sigma} + \kappa_J\mathbf{I})^{-1}\Big|_{J=0} = \kappa^2(\mathbf{\Sigma} + \kappa\mathbf{I})^{-2}\mathbf{\Sigma} - \frac{\partial\kappa_J}{\partial J}\Big|_{J=0}(\mathbf{\Sigma} + \kappa\mathbf{I})^{-2}\mathbf{\Sigma}. \tag{49}$$

Finally, implicitly differentiating the self-consistent equation and using the definition of $\kappa$, we have

$$\frac{\partial\kappa_J}{\partial J}\Big|_{J=0} = -\frac{\kappa^2}{\lambda}\left[\kappa\gamma + q\frac{1}{d}\operatorname{tr}[(\mathbf{\Sigma} + \kappa\mathbf{I})^{-2}\mathbf{\Sigma}]\frac{\partial\kappa_J}{\partial J}\Big|_{J=0}\right] \tag{50}$$

[1] This approach is suggested but not worked out in Cengiz Pehlevan's notes for the 2024 Analytical Connectionism Summer School.

where we let

$$\gamma = q\frac{1}{d}\operatorname{tr}[(\boldsymbol{\Sigma}+\kappa\mathbf{I})^{-2}\boldsymbol{\Sigma}^2] \tag{51}$$

as elsewhere. Solving for the derivative, this gives

$$\begin{aligned}\left.\frac{\partial \kappa_J}{\partial J}\right|_{J=0} &= -\left[1+\frac{\kappa^2}{\lambda}q\frac{1}{d}\operatorname{tr}[(\boldsymbol{\Sigma}+\kappa\mathbf{I})^{-2}\boldsymbol{\Sigma}]\right]^{-1}\frac{\kappa^2}{\lambda}\kappa\gamma &(52)\\ &= -\left[1+\frac{\kappa}{\lambda}\left(1-\frac{\lambda}{\kappa}\right)-\frac{\kappa}{\lambda}\gamma\right]^{-1}\frac{\kappa}{\lambda}\kappa^2\gamma &(53)\\ &= -\frac{\kappa^2\gamma}{1-\gamma}, &(54)\end{aligned}$$

using the identity

$$(\boldsymbol{\Sigma}+\kappa\mathbf{I})^{-1} = \frac{1}{\kappa}\left[\mathbf{I}-(\boldsymbol{\Sigma}+\kappa\mathbf{I})^{-1}\boldsymbol{\Sigma}\right] \tag{55}$$

and the fact that

$$1-\frac{\lambda}{\kappa} = q\frac{1}{d}\operatorname{tr}[(\boldsymbol{\Sigma}+\kappa\mathbf{I})^{-1}\boldsymbol{\Sigma}]. \tag{56}$$

Substituting this in, we have that

$$-\frac{\partial}{\partial J}\kappa_J(\boldsymbol{\Sigma}+\kappa_J J\boldsymbol{\Sigma}+\kappa_J\mathbf{I})^{-1}\Big|_{J=0} = \frac{\kappa^2}{1-\gamma}(\boldsymbol{\Sigma}+\kappa\mathbf{I})^{-2}\boldsymbol{\Sigma}. \tag{57}$$

Recalling our initial objective and collecting results, we have shown that

$$\lambda^2(\hat{\boldsymbol{\Sigma}}+\lambda\mathbf{I})^{-1}\boldsymbol{\Sigma}(\hat{\boldsymbol{\Sigma}}+\lambda\mathbf{I})^{-1} \sim \frac{\kappa^2}{1-\gamma}(\boldsymbol{\Sigma}+\kappa\mathbf{I})^{-2}\boldsymbol{\Sigma}. \tag{58}$$

This gives the desired deterministic equivalent for the signal term.

## 3.4 Deterministic equivalent for the noise term

We now turn to the (simpler) noise term. The desired result is that

$$\frac{1}{p}\operatorname{tr}[(\hat{\boldsymbol{\Sigma}}+\lambda\mathbf{I})^{-2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}] \sim \frac{\gamma}{1-\gamma} \tag{59}$$

for $\kappa$ and $\gamma$ as defined before. Our starting point is the fact that, as noted before,

$$\begin{aligned}(\hat{\boldsymbol{\Sigma}}+\lambda\mathbf{I})^{-2}\hat{\boldsymbol{\Sigma}} &= -\frac{\partial}{\partial\lambda}(\hat{\boldsymbol{\Sigma}}+\lambda\mathbf{I})^{-1}\hat{\boldsymbol{\Sigma}} &(60)\\ &= \frac{\partial}{\partial\lambda}[\lambda(\hat{\boldsymbol{\Sigma}}+\lambda\mathbf{I})^{-1}]. &(61)\end{aligned}$$

We can now immediately use the fact that

$$(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-1} \sim \frac{\kappa}{\lambda}(\mathbf{\Sigma} + \kappa \mathbf{I})^{-1} \tag{62}$$

to obtain

$$\frac{\partial}{\partial \lambda}[\lambda(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-1}] \sim \frac{\partial}{\partial \lambda}[\kappa(\mathbf{\Sigma} + \kappa \mathbf{I})^{-1}] \tag{63}$$

$$= \frac{\partial \kappa}{\partial \lambda}(\mathbf{\Sigma} + \kappa \mathbf{I})^{-2}\mathbf{\Sigma} \tag{64}$$

from which we find that

$$\frac{1}{p}\operatorname{tr}[(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-2}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}] \sim q\frac{1}{d}\operatorname{tr}[(\mathbf{\Sigma} + \kappa \mathbf{I})^{-2}\mathbf{\Sigma}^2]\frac{\partial \kappa}{\partial \lambda} \tag{65}$$

$$= \gamma \frac{\partial \kappa}{\partial \lambda} \tag{66}$$

using the definition of $\gamma$. What remains is to show the relationship between $\partial\kappa/\partial\lambda$ and $\gamma$. Implicitly differentiating the self-consistent equation and using the definition of $\kappa$, we have

$$\frac{\partial \kappa}{\partial \lambda} = \frac{1}{1 - q\frac{1}{d}\operatorname{tr}[(\mathbf{\Sigma} + \kappa \mathbf{I})^{-1}\mathbf{\Sigma}]} - \frac{\lambda}{\{1 - q\frac{1}{d}\operatorname{tr}[(\mathbf{\Sigma} + \kappa \mathbf{I})^{-1}\mathbf{\Sigma}]\}^2} q\frac{1}{d}\operatorname{tr}[(\mathbf{\Sigma} + \kappa \mathbf{I})^{-2}\mathbf{\Sigma}]\frac{\partial \kappa}{\partial \lambda} \tag{67}$$

$$= \frac{\kappa}{\lambda} - \frac{\kappa}{\lambda}\left[1 - \frac{\lambda}{\kappa} - \gamma\right]\frac{\partial \kappa}{\partial \lambda}. \tag{68}$$

Solving for $\partial\kappa/\partial\lambda$ and simplifying, we find that

$$\frac{\partial \kappa}{\partial \lambda} = \frac{1}{1 - \gamma}. \tag{69}$$

Therefore, we at last obtain the desired result:

$$\frac{1}{p}\operatorname{tr}[(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I})^{-2}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}] \sim \gamma \frac{\partial \kappa}{\partial \lambda} \tag{70}$$

$$= \frac{\gamma}{1 - \gamma}. \tag{71}$$

# 4 Gaussian universality

Now that we have this result, let's step back and think a bit about what we have accomplished. Under the assumption that our data was Gaussian, we figured out how to derive a deterministic asymptotic for the generalization error that captures the effect of randomness in the training data in the high-dimensional limit.

Before trying to extract some phenomenology from this result, it's important to ask whether these insights will actually apply to any interesting, realistic settings. The first objection one might have is that we have assumed that the covariates are Gaussian. Fortunately that is not an issue in the ridge regression context thanks to the phenomenon of Gaussian universality: one can show that so long as the data distribution is reasonably well-behaved the asymptotic generalization error will be equal to that for Gaussian data with matched mean and covariance. This includes, for instance, kernel regression settings like the NTK where the covariates live in strictly infinite dimension, as seen empirically by Canatar et al. [4]. Rigorous study of these universality results is an ongoing area of research, see for instance Misiakiewicz and Saeed [5] for some recent developments.

# 5 Phenomenology

With these results in hand—and some faith in their generality—let us try to extract qualitative insights.

## 5.1 Implicit regularization

First, we consider the significance of the fact that the ridge $\lambda$ is renormalized to $\kappa$. Recall that $\kappa$ is the unique positive solution to

$$\kappa = \frac{\lambda}{1 - q\frac{1}{d}\operatorname{tr}[(\mathbf{\Sigma} + \kappa \mathbf{I})^{-1}\mathbf{\Sigma}]} \tag{72}$$

as $d \to \infty$. It is useful to re-write this in terms of the limiting distribution of the eigenvalues $\sigma$ of $\Sigma$, *i.e.*, using the fact that

$$\lim_{d\to\infty} \frac{1}{d}\operatorname{tr}[(\mathbf{\Sigma} + \kappa \mathbf{I})^{-1}\mathbf{\Sigma}] = \mathbb{E}_\sigma\left[\frac{\sigma}{\sigma + \kappa}\right], \tag{73}$$

we have

$$\kappa = \frac{\lambda}{1 - q\mathbb{E}_\sigma\left[\frac{\sigma}{\sigma+\kappa}\right]}. \tag{74}$$

Below we prove the following facts about $\kappa$:

1. We have $\kappa \geq \lambda$, and $\lim_{q\downarrow 0} \kappa = \lambda$.
2. Increasing overparameterization increases $\kappa$, *i.e.*,

$$\frac{\partial \kappa}{\partial q} > 0. \tag{75}$$

   Indeed, as $q \to \infty$ we have $\kappa \to \infty$, corresponding to the fact that in this limit ridge regression gives the zero predictor with risk $R \sim \mathbf{w}_*^\top \mathbf{\Sigma} \mathbf{w}_*$.

3. In the ridgeless limit, the behavior of $\kappa$ depends on the overparameterization ratio $q$. In the underparameterized regime $q < 1$, $\kappa \downarrow 0$ as $\lambda \downarrow 0$, while in the overparmeterized regime $q > 1$, $\kappa$ tends to the unique positive solution of the equation

$$1 = q\mathbb{E}_\sigma \left[ \frac{\sigma}{\sigma + \kappa} \right]. \tag{76}$$

Therefore, the renormalized ridge can remain strictly positive even as the explicit ridge tends to zero.

4. Structure in the eigenvalue distribution decreases the renormalized ridge. In particular, let $\bar{\kappa}$ be the unique positive solution to

$$\bar{\kappa} = \frac{\lambda}{1 - q\bar{\sigma}/(\bar{\sigma} + \bar{\kappa})}, \tag{77}$$

with $\bar{\sigma} = \mathbb{E}_\sigma[\sigma]$, which is the self-consistent equation for an isotropic covariance matrix with matching mean eigenvalue. Then, one can show that

$$\kappa \leq \bar{\kappa}. \tag{78}$$

As the equation for $\bar{\kappa}$ is quadratic, it can be solved explicitly, giving

$$\bar{\kappa} = \frac{\lambda + (q-1)\bar{\sigma} + \sqrt{[\lambda + (q-1)\bar{\sigma}]^2 + 4\bar{\sigma}\lambda}}{2}. \tag{79}$$

We can easily work out that

$$\lim_{\lambda \downarrow 0} \bar{\kappa} = \begin{cases} 0 & q < 1 \\ (q-1)\bar{\sigma} & q > 1, \end{cases} \tag{80}$$

which agrees with the solution to the limiting equation derived above in this case.

We now prove these claims. For simplicity, we will assume that the support of the eigenvalue distribution is strictly bounded away from zero. That is, we assume that there are constants $0 < c \leq C$ such that $c \leq \sigma \leq C$. Moreover, we assume that $q \neq 1$.

It is first useful to record properties of the function

$$f(\sigma, \kappa) = \frac{\sigma}{\sigma + \kappa}, \tag{81}$$

viewed as a map from $[c, C] \times [0, \infty) \to [0, 1]$. First, we obviously have

$$0 \leq f(\sigma, \kappa) \leq 1, \tag{82}$$

with $f(\sigma, \kappa = 1) = 1$ for all $\sigma$. Next, we have

$$\frac{\partial f}{\partial \kappa} = -\frac{\sigma}{(\sigma + \kappa)^2} < 0 \tag{83}$$

for all $\sigma$, $\kappa$, while

$$\frac{\partial f}{\partial \sigma} = \frac{\kappa}{(\sigma+\kappa)^2} \geq 0 \tag{84}$$

with equality iff $\kappa = 0$. Finally, we have

$$\frac{\partial^2 f}{\partial \kappa^2} = \frac{2\sigma}{(\sigma+\kappa)^3} > 0, \tag{85}$$

while

$$\frac{\partial^2 f}{\partial \sigma^2} = -\frac{2\kappa}{(\sigma+\kappa)^3} \leq 0 \tag{86}$$

with equality iff $\kappa = 0$. In words, for any fixed $\sigma$, $f(\sigma, \kappa)$ is a strictly decreasing, strictly convex function of $\kappa$, while for any fixed $\kappa$, $f$ is a increasing concave function of $\sigma$, with both inequalities being strict if $\kappa > 0$. If $\kappa = 0$, then $f(\sigma, \kappa = 0) = 1$ for all $\sigma$. Moreover, all of these derivatives are bounded in magnitude.

Define

$$m(\kappa) = \mathbb{E}_\sigma \left[ \frac{\sigma}{\sigma+\kappa} \right]. \tag{87}$$

In light of the above, $m$ is a strictly decreasing, strictly convex function from $[0, \infty)$ to $[0, 1]$ with $m(0) = 1$. This makes it straightforward to prove the uniqueness of $\kappa$ and the claim that $\kappa \geq \lambda$. Recall that we defined $\kappa$ as the solution to

$$\kappa = \frac{\lambda}{1 - q m(\kappa)}. \tag{88}$$

As $0 \leq m(\kappa) \leq 1$, the denominator of the right-hand-side of this is equation is bounded from below by one, which shows that any solution must have $\kappa \geq \lambda$. Indeed, for any finite $\lambda$ one sees that $\kappa = 0$ is not a solution as $\lambda/(1-q) \neq 0$. The left-hand-side of this equation is obviously strictly increasing in $\kappa$, while the right-hand-side is strictly decreasing, whence the solution is unique.

We now turn our attention to $\partial\kappa/\partial q$. Implicitly differentiating the self-consistent equation, we have

$$\frac{\partial \kappa}{\partial q} = \frac{\lambda}{[1 - q m(\kappa)]^2} \left( m(\kappa) + q m'(\kappa) \frac{\partial \kappa}{\partial q} \right) \tag{89}$$

or

$$\frac{\partial \kappa}{\partial q} = \left( 1 - \frac{\lambda}{[1 - q m(\kappa)]^2} q m'(\kappa) \right)^{-1} \frac{\lambda}{[1 - q m(\kappa)]^2} m(\kappa) \tag{90}$$

where we remind the reader that $m'(\kappa) < 0$. From this, we conclude that $\partial\kappa/\partial q > 0$.

Now we consider the limit $\lambda \downarrow 0$. As $m(\kappa) \leq 1$, $1 - q m(\kappa)$ has no zeros if $q < 1$. Therefore, if $q < 1$ we have $\lim_{\lambda \downarrow 0} \kappa = 0$. In contrast, if $q > 1$ then $1 - q m(\kappa)$ has a unique positive root at $m(\kappa) = 1/q$. That this root provides the correct limit in the $q > 1$ regime follows from the fact that

$\lambda/(1-q)$ is negative, so if $\kappa$ vanished as some power of $\lambda$ one would arrive at a contradiction with the requirement that it is non-negative.

We now finally use the fact that $f(\sigma, \kappa)$ is concave in $\sigma$ to study the effect of structure in the eigenvalue distribution. This argument follows [3, 6]. For any $\kappa > 0$, Jensen's inequality implies that

$$m(\kappa) = \mathbb{E}_\sigma \left[ \frac{\sigma}{\sigma + \kappa} \right] \leq \frac{\bar{\sigma}}{\bar{\sigma} + \kappa} \tag{91}$$

where

$$\bar{\sigma} = \mathbb{E}_\sigma[\sigma] \tag{92}$$

with equality if and only if the distribution is a point mass. Under our assumptions, we of course have that $c \leq \bar{\sigma} \leq C$, so this bound is non-vacuous. As a result, we have that

$$\frac{1}{1 - qm(\kappa)} \leq \frac{1}{1 - q\bar{\sigma}/(\bar{\sigma} + \kappa)} \tag{93}$$

pointwise in $\kappa$, with equality at $\kappa = 0$. Both of these functions are decreasing in $\kappa$. Therefore, if we let $\bar{\kappa}$ be the unique positive solution to

$$\bar{\kappa} = \frac{\lambda}{1 - q\bar{\sigma}/(\bar{\sigma} + \bar{\kappa})}, \tag{94}$$

and let $\kappa$ solve $\kappa = \lambda/[1 - qm(\kappa)]$ as usual, we must have

$$\kappa \leq \bar{\kappa}. \tag{95}$$

As the equation for $\bar{\kappa}$ is quadratic, it can be solved explicitly, giving

$$\bar{\kappa} = \frac{\lambda + (q-1)\bar{\sigma} + \sqrt{[\lambda + (q-1)\bar{\sigma}]^2 + 4\bar{\sigma}\lambda}}{2}. \tag{96}$$

We can easily work out that

$$\lim_{\lambda \downarrow 0} \bar{\kappa} = \begin{cases} 0 & q < 1 \\ (q-1)\bar{\sigma} & q > 1, \end{cases} \tag{97}$$

which agrees with the solution to the limiting equation derived above in this case.

## 5.2 Double-descent

The change in the behavior at $q = 1$ suggests that something interesting might be going on there in terms of the generalization error. In fact there is: the generalization error diverges.

This behavior is easy to illustrate in the isotropic setting $\mathbf{\Sigma} = \mathbf{I}$, where the generalization error in the limit $\lambda \downarrow 0$ can be written down explicitly. From before, we know that in this case we have

$$\lim_{\lambda \downarrow 0} \kappa = \begin{cases} 0 & q < 1 \\ q - 1 & q > 1. \end{cases} \tag{98}$$

From this, we find that

$$\gamma = \begin{cases} q & q < 1 \\ \dfrac{1}{q} & q > 1 \end{cases} \tag{99}$$

whence

$$R \sim \begin{cases} \dfrac{q}{1-q}\eta^2 + \eta^2 & q < 1 \\ 1 - \dfrac{1}{q} + \dfrac{1}{q-1}\eta^2 + \eta^2 & q > 1. \end{cases} \tag{100}$$

This result was first derived by Krogh and Hertz [7] in 1992, using rather different methods!

This divergence can be related to the spectrum of the empirical covariance. As the isotropic case is interesting in its own right, we give a detailed study of its properties. This gives us more insight into the emergence of double-descent, and it turns out that we will encounter a few interesting phenomena along the way. In this case, we can average over isotropically-distributed $\mathbf{w}_*$ (with $\mathbb{E}[\mathbf{w}_* \mathbf{w}_*^\top] = \mathbf{I}/d$) without loss of generality as the data are rotation-invariant. Then, one finds *even at finite* $d, p$ that

$$\mathbb{E}_{\mathbf{w}_*}[R] = \mathbb{E}_\sigma \left[ \frac{\lambda^2}{(\sigma + \lambda)^2} \right] + \eta^2 q \mathbb{E}_\sigma \left[ \frac{\sigma}{(\sigma + \lambda)^2} \right] + \eta^2 \tag{101}$$

where expectation is taken with respect to the distribution of eigenvalues of $\hat{\mathbf{\Sigma}}$.

It is easy to see that the signal term cannot generate divergences, as the function

$$\frac{\lambda^2}{(\sigma + \lambda)^2} \tag{102}$$

is bounded from above by 1 for any $\sigma, \lambda \geq 0$. Moreover, one can check by differentiation that this function is in fact increasing in $\lambda$ for any $\sigma > 0$, while its derivative vanishes if $\sigma = 0$. In contrast, considering the noise term, the function

$$\frac{\sigma}{(\sigma + \lambda)^2} \tag{103}$$

is decreasing in $\lambda$ for any $\sigma > 0$, and its derivative vanishes if $\sigma = 0$. We have the pointwise bound

$$\frac{\sigma}{(\sigma + \lambda)^2} \leq \frac{1}{\lambda}, \tag{104}$$

which becomes vacuous as $\lambda \downarrow 0$. As an aside, we can see that

$$\frac{\partial}{\partial \lambda} \mathbb{E}_{\mathbf{w}_*}[R] = 2(\lambda - \eta^2 q) \mathbb{E}_\sigma \left[ \frac{\sigma^2}{(\sigma + \lambda)^2} \right] \tag{105}$$

whence the optimal ridge is

$$\lambda_* = \eta^2 q. \tag{106}$$

From this, we can see that only the noise term can generate possible divergences as $\lambda \downarrow 0$. Our task is therefore to study how divergences appear. By the Marchenko-Pastur theorem, the spectral density of $\hat{\mathbf{\Sigma}}$ tends in the limit $d, p \to \infty$ with $d/p \to q$ to

$$\frac{q-1}{q} \delta(\sigma) \mathbf{1}_{q>1} + \frac{\sqrt{(\sigma_+ - \sigma)(\sigma - \sigma_-)}}{2\pi q \sigma} \mathbf{1}_{\sigma \in [\sigma_-, \sigma_+]}, \tag{107}$$

where

$$\sigma_\pm = (1 \pm \sqrt{q})^2. \tag{108}$$

It is easy to see that the eigenvalues precisely equal to zero do not contribute to the noise term, and therefore may be neglected. As $q \to 1$ there is an accumulation of eigenvalues near zero, which generates the divergence as $\lambda \downarrow 0$. One way to see this is to suppose that we first took $q \to 1$ for fixed $\lambda$ and then took $\lambda \downarrow 0$. In this case the density of eigenvalues tends to

$$\frac{1}{2\pi} \sqrt{\frac{4-\sigma}{\sigma}} \mathbf{1}_{\sigma \in [0,4]} \tag{109}$$

and we find that

$$\mathbb{E}_\sigma \left[ \frac{\sigma}{(\sigma + \lambda)^2} \right] = \frac{2 + \lambda - \sqrt{(4+\lambda)\lambda}}{2\sqrt{(4+\lambda)\lambda}} \tag{110}$$

As $\lambda \downarrow 0$, this diverges as $1/\sqrt{\lambda}$. To complete the picture at $q = 1$, we have a similar explicit formula for the signal term:

$$\mathbb{E}_\sigma \left[ \frac{\lambda^2}{(\sigma + \lambda)^2} \right] = \sqrt{\frac{\lambda}{4+\lambda}}. \tag{111}$$

This leads to a simple tradeoff between the signal and noise terms at $q = 1$ as a function of $\lambda$: the noise term blows up as $1/\sqrt{\lambda}$ as $\lambda \downarrow 0$ and decays as $1/\lambda^2$ as $\lambda \to \infty$, while the signal term decays as $\sqrt{\lambda}$ as $\lambda \downarrow 0$ and tends to 1 as $\lambda \to \infty$.

## 5.3 Spectral bias

We now aim to get a more detailed understanding of the inductive bias of ridge regression. We will show, follwing Canatar et al. [4], that ridge regression has a **spectral bias**: learning is faster along eigendirections of $\mathbf{\Sigma}$ with large eigenvalue.

Focusing on the signal term in the generalization error (and making for notational convenience the assumption that $\mathbf{\Sigma}$ has discrete spectrum; this either follows by working at large but finite $d$ or can be relaxed at the cost of introducing some additional notation as in [1]), we can write

$$R \sim \sum_k \sigma_k w_{*,k}^2 R_k \tag{112}$$

for

$$R_k = \frac{1}{1-\gamma} \frac{\kappa^2}{(\sigma_k + \kappa)^2} \tag{113}$$

where we work in a basis where $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots)$.

The sign of the derivative of $R_k$ with respect to $q$ is not immediately obvious. It is easier to consider the change in the ratio of two different mode errors. One can easily compute

$$\frac{\partial}{\partial q} \log \frac{R_k}{R_{k'}} = \frac{\partial}{\partial q} \log \frac{(\sigma_{k'} + \kappa)^2}{(\sigma_k + \kappa)^2} \tag{114}$$

$$= 2 \left( \frac{1}{\sigma_{k'} + \kappa} - \frac{1}{\sigma_k + \kappa} \right) \frac{\partial \kappa}{\partial q} \tag{115}$$

$$= 2 \frac{\sigma_k - \sigma_{k'}}{(\sigma_k + \kappa)(\sigma_{k'} + \kappa)} \frac{\partial \kappa}{\partial q} \tag{116}$$

But, we recall from before that

$$\frac{\partial \kappa}{\partial q} > 0, \tag{117}$$

so

$$\frac{\partial}{\partial q} \log \frac{R_k}{R_{k'}} > 0 \tag{118}$$

whenever $\sigma_k > \sigma_{k'}$. In other words,

$$\frac{\partial}{\partial q} \log R_k > \frac{\partial}{\partial q} \log R_{k'} \tag{119}$$

whenever $\sigma_k > \sigma_{k'}$. As increasing the number of training examples decreases the overparameterization ratio $q$, this means that modes with larger eigenvalue are learned faster in the sense of logarithmic derivatives.

We can obtain a stronger result by noting that as $q \to \infty$ we have

$$\lim_{q \to \infty} \frac{R_k}{R_{k'}} = \lim_{q \to \infty} \frac{(\sigma_{k'} + \kappa)^2}{(\sigma_k + \kappa)^2} = 1, \tag{120}$$

so

$$\log \frac{R_k}{R_{k'}} \tag{121}$$

is an increasing function that tends to zero as $q \to \infty$. Solving backward from that initial condition, we conclude that

$$\log \frac{R_k}{R_{k'}} < 0 \tag{122}$$

whenever $\sigma_k > \sigma_{k'}$, or, in other words,

$$R_k < R_{k'}. \tag{123}$$

This proves that the mode errors decrease with increasing eigenvalue magnitude.

## 5.4 Scaling laws

Power-law decays in covariance spectra are ubiquitous in natural data. How does ridge regression perform when faced with such a task? Moreover, one of the salient findings in recent studies of large language models is the observation of **neural scaling laws**: most simply, the generalization error decays as a power law in the number of training examples. An important problem is therefore to study the factors that determine such scaling laws in simple models. For more details, see [2].

With these motivations in mind, suppose that

$$\sigma_k = k^{-\alpha} \tag{124}$$

for some $\alpha \geq 0$, and

$$\sigma_k w_{*,k}^2 = k^{-(1+2\alpha r)} \tag{125}$$

for some $r \geq 0$. Here, $\alpha$ is known as the **capacity** exponent, while $r$ is the **source** exponent.

A few remarks are in order. First, if $\alpha > 1$ then the covariance matrix is trace class in the limit $d \to \infty$ as $\mathrm{tr}(\mathbf{\Sigma}) = \sum_{k=1}^{d} k^{-\alpha}$. Moreover, we have

$$\|\mathbf{w}_*\|_2^2 = \sum_{k=1}^{d} k^{-1-(2r-1)\alpha}, \tag{126}$$

so the norm of $\mathbf{w}_*$ will diverge as $d \to \infty$ unless $r > 1/2$. Finally, the exponent $r$ measures how much power remains above mode $k$ in the sense of the norm $\mathbf{w}^\top \mathbf{\Sigma} \mathbf{w}$:

$$\sum_{k'>k} \sigma_k w_{*,k}^2 \sim \int_k^\infty \frac{dk'}{(k')^{1+2\alpha r}} = \frac{1}{2\alpha r} k^{-2\alpha r} \tag{127}$$

One can show that under these assumptions the generalization error of ridge regression will be well-approximated by a power law with an exponent determined by $\alpha$, $r$, and the ridge $\lambda$. Here we will consider only the ridgeless limit in the overparameterized regime for $\alpha > 1$ for zero noise ($\eta = 0$); see [2] for a more general analysis. In this case we will take an ordered limit $d \to \infty$ then $p \to \infty$ in our asymptotic; it can be rigorously justified that the result is actually correct [5].

Under these conditions, we will argue that

$$R \sim C p^{-2\alpha \min\{r,1\}}, \tag{128}$$

for some constant $C$. This argument follows [2].

Using our previous arguments about $\kappa$ in the overparameterized regime, we have (with $\eta = 0$)

$$R \sim \frac{\kappa^2}{1-\gamma} \sum_{k=1}^{d} \frac{k^{-(1+2\alpha r)}}{(k^{-\alpha} + \kappa)^2} \tag{129}$$

where $\kappa$ solves

$$\frac{1}{p} \sum_{k=1}^{d} \frac{1}{1 + k^\alpha \kappa} = 1 \tag{130}$$

and

$$\gamma = \frac{1}{p}\sum_{k=1}^{d}\frac{1}{(1+k^{\alpha}\kappa)^2}. \tag{131}$$

Our first task is to analyze the behavior of $\kappa$. In the limit $d \to \infty$ we can approximate the (convergent!) sum over integer $k$ by an integral:

$$\sum_{k=1}^{\infty}\frac{1}{1+k^{\alpha}\kappa} \sim \int_1^{\infty}\frac{dk}{1+k^{\alpha}\kappa} \tag{132}$$

$$= \kappa^{-1/\alpha}\int_{\kappa^{1/\alpha}}^{\infty}\frac{du}{1+u^{\alpha}} \tag{133}$$

where we have put $u = k\kappa^{1/\alpha}$. We now note that the function

$$\int_{\kappa^{1/\alpha}}^{\infty}\frac{du}{1+u^{\alpha}} \tag{134}$$

is decreasing in $\kappa$, with

$$\int_0^{\infty}\frac{du}{1+u^{\alpha}} = \frac{\pi}{\alpha}\csc\frac{\pi}{\alpha} > 0. \tag{135}$$

From this, we have a self-consistent approximate solution

$$\kappa \sim Ap^{-\alpha} \tag{136}$$

as $p \to \infty$ for $A = \frac{\pi}{\alpha}\csc\frac{\pi}{\alpha}$. From this, we find that $\gamma$ tends to a constant:

$$\gamma \sim \frac{1}{p}\int_1^{\infty}\frac{dk}{(1+k^{\alpha}\kappa)^2} \tag{137}$$

$$= \frac{1}{\kappa^{1/\alpha}p}\int_{\kappa^{1/\alpha}}^{\infty}\frac{du}{(1+u^{\alpha})^2} \tag{138}$$

$$\sim \frac{1}{A}\int_{A^{1/\alpha}p^{-1}}^{\infty}\frac{du}{(1+u^{\alpha})^2} \tag{139}$$

$$\sim \frac{1}{A}\frac{\alpha-1}{\alpha}\frac{\pi}{\alpha}\csc\frac{\pi}{\alpha} \tag{140}$$

$$= 1 - \frac{1}{\alpha}, \tag{141}$$

where as before we let $u = \kappa^{1/\alpha}k$. Finally, we consider the signal term:

$$\kappa^2\sum_{k=1}^{\infty}\frac{k^{-(1+2\alpha r)}}{(k^{-\alpha}+\kappa)^2} \sim \kappa^2\int_1^{\infty}\frac{k^{-(1+2\alpha r)}}{(k^{-\alpha}+\kappa)^2}dk \tag{142}$$

$$= \kappa^{2r}\int_{\kappa^{1/\alpha}}^{\infty}\frac{u^{-(1+2\alpha r)}}{(u^{-\alpha}+1)^2}du \tag{143}$$

$$\sim A^{2r}p^{-2\alpha r}\int_{A^{1/\alpha}p^{-1}}^{\infty}\frac{u^{-(1+2\alpha r)}}{(u^{-\alpha}+1)^2}du \tag{144}$$

$$= A^{2r}p^{-2\alpha r}\int_{A^{1/\alpha}p^{-1}}^{\infty}\frac{u^{-1+2\alpha(1-r)}}{(1+u^{\alpha})^2}du. \tag{145}$$

We can see from the second-to-last expression that the integral is clearly convergent as $u \to \infty$, but the behavior near the lower limit requires some examination. Indeed, we see from the last expression that if $r > 1$ it becomes divergent as $u \downarrow 0$. If $r < 1$ the integral tends to a constant and we have

$$\kappa^2 \sum_{k=1}^{\infty} \frac{k^{-(1+2\alpha r)}}{(k^{-\alpha} + \kappa)^2} \sim B p^{-2\alpha r} \tag{146}$$

for a constant $B = A^{2r} \int_0^\infty \frac{u^{-1+2\alpha(1-r)}}{(1+u^\alpha)^2} du$. To extract the dominant behavior with $r > 1$ we integrate by parts, noting that contribution of the $u \to \infty$ boundary term vanishes, and neglect sub-dominant corrections:

$$A^{2r} p^{-2\alpha r} \int_{A^{1/\alpha} p^{-1}}^{\infty} \frac{u^{-1+2\alpha(1-r)}}{(1+u^\alpha)^2} du \sim C p^{-2\alpha} + \frac{A^{2r} p^{-2\alpha r}}{(1-r)} \int_{A^{1/\alpha} p^{-1}}^{\infty} \frac{u^{2\alpha(1-r)} u^{\alpha-1}}{(1+u^\alpha)^3} du. \tag{147}$$

Repeating this process, we see that we will have

$$C_1 p^{-2\alpha} + C_2 p^{-3\alpha} + C_3 p^{-4\alpha} + \dots \tag{148}$$

as the $j$-th iteration yields a term scaling as $p^{-2\alpha r - (j-1)\alpha - 2\alpha(1-r)} = p^{-(j+1)\alpha}$. Thus, the first term dominates as $p \to \infty$, and the signal scales as $p^{-2\alpha}$.

Combining this with the result we found before for $r < 1$ and using the fact that $\gamma$ tends to a constant, we conclude that

$$R \sim C p^{-2\alpha \min\{r,1\}}, \tag{149}$$

for a constant $C$ that we do not bother to write down.

# 6 Applications to neuroscience

The results developed above have broader applications to neuroscience and machine learning, because they can provide some answer to the following question: Given a set of features, what functions are easy to linearly decode, in the sense that the weights can be learned from few examples? The idea of spectral bias provides one answer to this question: we know that target functions that are well-aligned to eigendirections of $\mathbf{\Sigma}$ with large eigenvalue can be learned efficiently. To measure this notion of alignment between a task and the features from which one wants to learn it, Canatar et al. [4] proposed the cumulative power

$$C_k = \frac{\sum_{k' \leq k} \sigma_k w_{*,k}^2}{\sum_{k'} \sigma_k w_{*,k}^2}. \tag{150}$$

Under the source-capacity assumptions, we have

$$\sum_{k' \leq k} \sigma_k w_{*,k}^2 \sim \int_1^k \frac{dk'}{(k')^{1+2\alpha r}} \tag{151}$$

$$= -\frac{(k')^{-2\alpha r}}{2\alpha r}\Bigg|_{k'=1}^{k'=k} \tag{152}$$

$$= \frac{1 - k^{-2\alpha r}}{2\alpha r}, \tag{153}$$

from which we see that

$$C_k \sim 1 - k^{-2\alpha r}. \tag{154}$$

This illustrates the fact that, for fixed $\alpha$, tasks with larger $r$ are easier.

# References


[1] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949 – 986, 2022. doi: 10.1214/21-AOS2133. URL https://doi.org/10.1214/21-AOS2133.

[2] Alexander B Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024.

[3] Francis Bach. High-dimensional analysis of double descent for linear regression with random projections. *SIAM Journal on Mathematics of Data Science*, 6(1):26–50, 2024. doi: 10.1137/23M1558781. URL https://doi.org/10.1137/23M1558781.

[4] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1):2914, 2021.

[5] Theodor Misiakiewicz and Basil Saeed. A non-asymptotic theory of kernel ridge regression: deterministic equivalents, test error, and GCV estimator. *arXiv preprint arXiv:2403.08938*, 2024.

[6] Jacob A Zavatone-Veth and Cengiz Pehlevan. Learning curves for deep structured Gaussian feature models. In *Advances in Neural Information Processing Systems*, 2023.

[7] Anders Krogh and John A Hertz. Generalization in a linear perceptron in the presence of noise. *Journal of Physics A: Mathematical and General*, 25(5):1135, 1992.