Mathematics of Deep Learning
Lecture 6: Random design analysis of ridge regression

Bruno Loureiro
Département d'Informatique, École Normale Supérieure - PSL & CNRS, France

21/02/2025

Get in touch at: bruno.loureiro@di.ens.fr

1 Setting

Consider a supervised regression setting with training data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R} : i \in [n]\}$. In this lecture, we will be interested in the analysis of ridge regressor $f(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_\lambda) = \langle \hat{\boldsymbol{\theta}}_\lambda, \boldsymbol{x} \rangle$ with:

$$\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y}) := \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle)^2 + \lambda ||\boldsymbol{\theta}||_2^2 \tag{1.1}$$

$$= \left(\boldsymbol{X}^\top \boldsymbol{X} + n\lambda \boldsymbol{I}_d \right)^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{1.2}$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and $\boldsymbol{y} \in \mathbb{R}^n$ denote the covariate matrix and response vector, obtained by stacking $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ row-wise. Since the ridge predictor is a linear function, it can only express linear dependences on the data. Therefore, it is natural to assume that data has been independently drawn from a linear model:

$$y_i = \langle \boldsymbol{\theta}_\star, \boldsymbol{x}_i \rangle + \varepsilon_i. \tag{1.3}$$

In particular, we will be interested in analysing the so-called *random design setting*.

Assumption 1. Throughout this lecture, we assume that:

- Gaussian covariates, i.e. $\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2} \boldsymbol{z}_i$ with $\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ and $\boldsymbol{\Sigma} \succ 0$ a positive-definite matrix.
- The label noise are drawn independently from $\boldsymbol{x}_i$, are zero-mean and have finite variance $\mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty$.
- The target weights have finite norm $||\boldsymbol{\theta}_\star||_2^2 < \infty$.

Note that the above implicitly define a joint distribution $p(\boldsymbol{x}, y)$. In particular, the Bayes risk is given by the noise variance $R_\star = \sigma^2$, and corresponds to the Bayes predictor $f_\star(\boldsymbol{x}) = \langle \boldsymbol{\theta}_\star, \boldsymbol{x} \rangle$.

Our goal is to derive a precise characterisation of the excess risk:

$$R(\hat{\boldsymbol{\theta}}_\lambda) - \sigma^2 = \mathbb{E}_{\boldsymbol{x}} \left[\left(f_\star(\boldsymbol{x}) - f(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_\lambda) \right)^2 \right] \tag{1.4}$$

$$= \langle \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}_\star, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}_\star) \rangle \tag{1.5}$$

$$:= ||\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}_\star||_{\boldsymbol{\Sigma}}^2 \tag{1.6}$$

where in the last equality we simply recognised the definition of the *Mahalanobis norm.* In other words: the excess risk is a mean-squared error weighted by the most relevant directions in the data (i.e. with largest eigenvalues).

Note that the excess risk above is a random quantity (since $\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X},\boldsymbol{y})$ is a function of the training data, which is random). To simplify the analysis, we will rather consider the average of the expected risk with respect to the label noise:

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[R(\hat{\boldsymbol{\theta}}_\lambda)\right]-\sigma^2=\mathbb{E}_{\boldsymbol{\varepsilon}}\left[||\hat{\boldsymbol{\theta}}_\lambda-\boldsymbol{\theta}_\star||^2_{\boldsymbol{\Sigma}}\right] \tag{1.7}$$

With more work, one can show concentration of $R(\hat{\boldsymbol{\theta}}_\lambda)$ over $\boldsymbol{\varepsilon}$ under an additional sub-Gaussian assumption.

⚠ $\mathbb{E}_{\boldsymbol{\varepsilon}}\left[R(\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X},\boldsymbol{y}))\right]$ is still a random quantity since it depends on the random data $\boldsymbol{X},\boldsymbol{y}$.

2 Bias-variance decomposition

In regression problems with additive noise $y_i=f_\star(\boldsymbol{x}_i)+\varepsilon_i$, it is common to write the expected excess risk in terms of a (squared) bias and variance decomposition with respect to the label noise:

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[R(\hat{\boldsymbol{\theta}}_\lambda)\right]-\sigma^2=B(f_\star,\boldsymbol{X},\lambda)+V(\boldsymbol{X},\lambda) \tag{2.1}$$

where:

$$B(f_\star,\boldsymbol{X},\lambda)=\mathbb{E}_{\boldsymbol{x}}\left[(f_\star(\boldsymbol{x})-\mathbb{E}_{\boldsymbol{\varepsilon}}[f(\boldsymbol{x};\hat{\boldsymbol{\theta}})])^2\right] \tag{2.2}$$

$$V(\boldsymbol{X},\lambda)=\mathrm{Var}_{\boldsymbol{\varepsilon}}(f(\boldsymbol{x};\hat{\boldsymbol{\theta}}))=\mathbb{E}_{\boldsymbol{x},\boldsymbol{\varepsilon}}\left[(f(\boldsymbol{x};\hat{\boldsymbol{\theta}})-\mathbb{E}_{\boldsymbol{\varepsilon}}[f(\boldsymbol{x};\hat{\boldsymbol{\theta}})])^2\right] \tag{2.3}$$

In particular, note that the variance is independent of the target function $f_\star$.

⚠ Note that the (squared) bias and variance are defined with respect to the training data label noise $\boldsymbol{\varepsilon}\in\mathbb{R}^n$. In particular, they are still random functions of $\boldsymbol{X}$.

Explicit expressions for the bias and variance can be worked out from the definition. But in our case it is simpler to note that we can write:

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X},\boldsymbol{y})&=\left(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\theta}_\star+\boldsymbol{\varepsilon}) &&(2.4)\\
&=\left(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\theta}_\star+\left(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon} &&(2.5)\\
&\overset{(a)}{=}\left(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d\right)^{-1}(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d-n\lambda\boldsymbol{I}_d)\boldsymbol{\theta}_\star+\left(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon} &&(2.6)\\
&=\boldsymbol{\theta}_\star-n\lambda\left(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{\theta}_\star+\left(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon} &&(2.7)
\end{aligned}$$

where in (a) we added and subtracted $n\lambda\boldsymbol{I}_d$. Therefore, we have:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\varepsilon}}\left[R(\hat{\boldsymbol{\theta}}_\lambda)\right]-\sigma^2&=\mathbb{E}_{\boldsymbol{\varepsilon}}\left[||\hat{\boldsymbol{\theta}}_\lambda-\boldsymbol{\theta}_\star||^2_{\boldsymbol{\Sigma}}\right] &&(2.8)\\
&=(n\lambda)^2\langle\boldsymbol{\theta}_\star,\left(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{\theta}_\star\rangle\\
&\quad+\sigma^2\,\mathrm{Tr}\left\{\boldsymbol{X}^\top\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d\right)^{-1}\right\} &&(2.9)
\end{aligned}$$

where we used that which allow us to identify:

$$B(\boldsymbol{\theta}_\star,\boldsymbol{X},\lambda)=(n\lambda)^2\langle\boldsymbol{\theta}_\star,\left(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{\theta}_\star\rangle \tag{2.10}$$

$$V(\boldsymbol{X},\lambda)=\sigma^2\,\mathrm{Tr}\left\{\boldsymbol{X}^\top\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d\right)^{-2}\boldsymbol{\Sigma}\right\} \tag{2.11}$$

where in the last expression we used the fact that the matrices inside the trace commute. It is also common to see the expressions above written in terms of the data empirical covariance matrix:

$$\hat{\boldsymbol{\Sigma}}_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top = \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} \tag{2.12}$$

which reads:

$$B(\boldsymbol{\theta}_\star, \boldsymbol{X}, \lambda) = \lambda^2 \langle \boldsymbol{\theta}_\star, \left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{\Sigma} \left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{\theta}_\star \rangle \tag{2.13}$$

$$V(\boldsymbol{X}, \lambda, \sigma^2) = \frac{\sigma^2}{n} \operatorname{Tr}\left\{\hat{\boldsymbol{\Sigma}}_n \left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-2} \boldsymbol{\Sigma}\right\} \tag{2.14}$$

Characterising the behaviour of the quantities above becomes, at this point, a random matrix theory problem. Before looking at the general result, let's do a warm up.

2.1 Warm-up: ordinary least-squares

We now consider the ordinary least-squares case $\lambda = 0$, which is equivalent to solving a system of n equations with d unknowns:

$$\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\theta} \overset{!}{=} \boldsymbol{X}^\top \boldsymbol{y} \tag{2.15}$$

For $n \geq d$, since $\boldsymbol{X}^\top \boldsymbol{X} \in \mathbb{R}^{n\times n}$ is invertible with high-probability, the solution to this problem is unique. For $n < d$, $\boldsymbol{X}^\top \boldsymbol{X}$ is rank defficient, i.e. we have a linear system with more unknowns than variables, and many solutions exist. These solutions can be explicitly written as:

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{v}}(\boldsymbol{X}, \boldsymbol{y}) = \boldsymbol{X}^{+} \boldsymbol{y} + \boldsymbol{v} \tag{2.16}$$

where $\boldsymbol{X}^{+}$ is the Moore-Penrose inverse of $\boldsymbol{X}$ and $\boldsymbol{v}$ is any vector in the kernel of $\boldsymbol{X}^\top \boldsymbol{X}$. One particular solution is the *least-norm solution* for which $\boldsymbol{v} = 0$, and corresponds to the $\lambda \to 0^{+}$ limit of the ridge estimator $\hat{\boldsymbol{\theta}}_\lambda$.

Interestingly, the bias and variance of the ordinary least-squares estimator can be easily computed when $n > d + 1$. Since $\hat{\boldsymbol{\Sigma}}_n$ is almost surely invertible in this case, we have:

$$B(\boldsymbol{\theta}_\star, \boldsymbol{X}, \lambda = 0^{+}) = 0 \tag{2.17}$$

$$V(\boldsymbol{X}, \lambda = 0^{+}, \sigma^2) = \frac{\sigma^2}{n} \operatorname{Tr}\left\{\hat{\boldsymbol{\Sigma}}_n^{-1} \boldsymbol{\Sigma}\right\} \tag{2.18}$$

i.e. the excess risk is fully given by the variance. Writing $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}$ for $\boldsymbol{Z}$ a Gaussian i.i.d. matrix with zero mean and unit variance, we have:

$$V(\boldsymbol{X}, \lambda = 0^{+}, \sigma^2) = \sigma^2 \operatorname{Tr}\left\{(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\right\} \tag{2.19}$$

Curiously, this is independent of the data covariance matrix. The random matrix $(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}$ is an inverse Wishart matrix with n degrees-of-freedom, a well-studied random matrix ensemble. In particular, its mean is equal to:

$$\mathbb{E}[(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}] = \frac{1}{n - d - 1} \boldsymbol{I}_d \tag{2.20}$$

implying that:

$$\mathbb{E}[V(\boldsymbol{X}, \lambda = 0^{+}, \sigma^2)] = \frac{\sigma^2 d}{n - d - 1} \tag{2.21}$$

This result in expectation can be turned into a high-probability bound for the excess risk when $n \to \infty$.[1] Curiously, this expression is fairly close to the excess risk under fixed design setting $\mathbb{E}[R] - \sigma^2 = {}^{\sigma^2 d}/_{n}$, and imply that under the well-specified case the risk converges to zero at a $O(n^{-1})$ rate as $n \to \infty$. An alternative but asymptotic way to get this result is to recognise that this is exactly the $z \to 0$ limit of the Stieltjes transform of the Marchenko-Pastur distribution.

3 High-dimensional asymptotics

We now leverage the random matrix theory results discussed in the previous lecture to provide a sharp characterisation of the bias variances in the high-dimensional proportional asymptotics $n, d \to \infty$ with ${}^{d}/_{n} \to \gamma = \Theta(1)$. Recall that in the previous lecture we have shown that:

Theorem 1. Let $\hat{\boldsymbol{\Sigma}}_n = {}^{1}/_{n}\boldsymbol{X}^\top \boldsymbol{X} \in \mathbb{R}^{d\times d}$ with $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}$, where $\boldsymbol{Z}$ is a sub-Gaussian matrix with zero mean and unit variance and $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ is a positive-definite matrix with eigenvalues $\operatorname{spec}(\boldsymbol{\Sigma}) = \{\lambda_k : k \in [d]\} \subset \mathbb{R}_+$ and bounded operator norm $||\boldsymbol{\Sigma}||_{\text{op}} < C$. Assume that the empirical measure of eigenvalues $\hat{\mu}_n = {}^{1}/_{d}\sum_{i\in[d]} \delta_{\lambda_i}$ converges (weakly) to a probability distribution μ on $\mathbb{R}_+$ with compact support as $d \to \infty$. Then, for any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d\times d}$ with bounded operator norm, the following asymptotic equivalents hold in the proportional limit where $d \to \infty$ with ${}^{d}/_{n} \to \gamma > 0$:

$$\operatorname{Tr}\{\boldsymbol{A}(\hat{\boldsymbol{\Sigma}}_n - z\boldsymbol{I}_d)^{-1}\} \asymp -\frac{1}{z\tilde{s}(z)} \operatorname{Tr}\left\{\boldsymbol{A}\left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}(z)}\boldsymbol{I}_d\right)^{-1}\right\} \tag{3.1}$$

$$\begin{aligned}\operatorname{Tr}\{\boldsymbol{A}(\hat{\boldsymbol{\Sigma}}_n - z\boldsymbol{I}_d)^{-1}\boldsymbol{B}(\hat{\boldsymbol{\Sigma}}_n - z\boldsymbol{I}_d)^{-1}\} &\asymp \frac{1}{z^2\tilde{s}(z)^2} \operatorname{Tr}\left\{\boldsymbol{A}\left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}(z)}\boldsymbol{I}_d\right)^{-1} \boldsymbol{B}\left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}(z)}\boldsymbol{I}_d\right)^{-1}\right\} \\ &\quad + \frac{1}{z^2\tilde{s}(z)^2} \frac{\operatorname{Tr}\left\{\boldsymbol{A}\left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}(z)}\boldsymbol{I}_d\right)^{-2}\boldsymbol{\Sigma}\right\} \cdot \operatorname{Tr}\left\{\boldsymbol{B}\left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}(z)}\boldsymbol{I}_d\right)^{-2}\boldsymbol{\Sigma}\right\}}{n - \operatorname{Tr}\left\{\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + {}^{1}/_{\tilde{s}(z)}\boldsymbol{I}_d)^{-2}\right\}}\end{aligned} \tag{3.2}$$

where $\tilde{s}(z)$ is the unique solution of the following self-consistent equation:

$$\frac{1}{\tilde{s}(z)} + z = \frac{\gamma}{d} \operatorname{Tr}\left\{\boldsymbol{\Sigma}(\tilde{s}(z)\boldsymbol{\Sigma} + \boldsymbol{I}_d)^{-1}\right\} \tag{3.3}$$

Remark 1. Note that in eq. (3.1) and (3.2) we used the *asymptotic equivalent notation* "$\asymp$". We say $a_n \asymp b_n$ as $n \to \infty$ if $a_n = \Theta(b_n)$ or equivalently $a_n = O(b_n)$ and $b_n = O(a_n)$ see **??** for a detailed discussion. In particular, this implies that $\lim_{n\to\infty} {}^{a_n}/_{b_n} \to 1$. When employing this notation in our context, we are always referring to the proportional asymptotical limit, and when dealing with random quantities the convergence will be almost surely or in probability. When both sides of $\asymp$ are of the same order in n, this implies convergence (a.s. or in probability) of the normalised quantities, e.g. ${}^{a_n}/_{n} \to {}^{b_n}/_{n}$ if $a_n, b_n = \Theta(n)$. Therefore, the main convenience of this notation is to speak of asymptotic limits without having to care for the normalisation of the quantities involved.

Now rewriting the bias and variance in the form of eqs. (3.2) and (3.2):

$$B(\boldsymbol{\theta}_\star, \boldsymbol{X}, \lambda) = \lambda^2 \operatorname{Tr}\left\{\boldsymbol{\theta}_\star\boldsymbol{\theta}_\star^\top \left(\hat{\boldsymbol{\Sigma}}_n + \lambda\boldsymbol{I}_d\right)^{-1} \boldsymbol{\Sigma}\left(\hat{\boldsymbol{\Sigma}}_n + \lambda\boldsymbol{I}_d\right)^{-1}\right\} \tag{3.4}$$

$$V(\boldsymbol{X}, \lambda, \sigma^2) = \frac{\sigma^2}{n} \operatorname{Tr}\left\{\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\Sigma}}_n + \lambda\boldsymbol{I}_d\right)^{-1}\right\} - \frac{\lambda\sigma^2}{n} \operatorname{Tr}\left\{\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\Sigma}}_n + \lambda\boldsymbol{I}_d\right)^{-2}\right\} \tag{3.5}$$

[1] For instance, by showing that $\mathbb{E}\operatorname{Tr}\left[(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\right]^2$ is vanishing with $n \to \infty$.

We can readily apply theorem 1. Let's start with the bias. Using eq. (3.2) with $\boldsymbol{A} = \boldsymbol{\theta}_\star \boldsymbol{\theta}_\star^\top$ and $\boldsymbol{B} = \boldsymbol{\Sigma}$ evaluated at $z = -\lambda$ give us:

$$
\begin{aligned}
B(\boldsymbol{\theta}_\star, \boldsymbol{X}, \lambda) &\asymp \frac{1}{\tilde{s}(-\lambda)^2} \operatorname{Tr}\left\{ \boldsymbol{\theta}_\star \boldsymbol{\theta}_\star \left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}(-\lambda)} \boldsymbol{I}_d \right)^{-2} \boldsymbol{\Sigma} \right\} \\
&\quad + \frac{1}{\tilde{s}(-\lambda)^2} \frac{\operatorname{Tr}\left\{ \boldsymbol{\theta}_\star \boldsymbol{\theta}_\star \left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}(-\lambda)} \boldsymbol{I}_d \right)^{-2} \boldsymbol{\Sigma} \right\} \cdot \operatorname{Tr}\left\{ \boldsymbol{\Sigma}^2 \left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}(-\lambda)} \boldsymbol{I}_d \right)^{-2} \right\}}{n - \operatorname{Tr}\left\{ \boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + {}^{1}/_{\tilde{s}(-\lambda)} \boldsymbol{I}_d)^{-2} \right\}} \\
&= \frac{1}{\tilde{s}(-\lambda)^2} \boldsymbol{\theta}_\star^\top \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}(-\lambda)} \boldsymbol{I}_d \right)^{-2} \boldsymbol{\theta}_\star \left[1 + \frac{\operatorname{Tr}\left\{ \boldsymbol{\Sigma}^2 \left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}(-\lambda)} \boldsymbol{I}_d \right)^{-2} \right\}}{n - \operatorname{Tr}\left\{ \boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + {}^{1}/_{\tilde{s}(-\lambda)} \boldsymbol{I}_d)^{-2} \right\}} \right]
\end{aligned}
\tag{3.6}
$$

It will be convenient to define $\kappa(\lambda) := {}^{1}/_{\tilde{s}(-\lambda)}$ and rewrite the brackets$1 + \frac{x}{1-x} = \frac{1}{1-x}$:

$$
B(\boldsymbol{\theta}_\star, \boldsymbol{X}, \lambda) \asymp \frac{\kappa(\lambda)^2 \langle \boldsymbol{\theta}_\star, \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma} + \kappa(\lambda) \boldsymbol{I}_d \right)^{-2} \boldsymbol{\theta}_\star \rangle}{1 - \frac{1}{n} \operatorname{Tr}\left\{ \boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \kappa(\lambda) \boldsymbol{I}_d)^{-2} \right\}}
\tag{3.7}
$$

The variance is composed of two terms. For the first, we use eq. (3.1) with $\boldsymbol{A} = \boldsymbol{\Sigma}$:

$$
\frac{\sigma^2}{n} \operatorname{Tr}\left\{ \boldsymbol{\Sigma} \left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d \right)^{-1} \right\} \asymp \frac{\sigma^2 \kappa(\lambda)}{n\lambda} \operatorname{Tr}\left\{ \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma} + \kappa(\lambda) \boldsymbol{I}_d \right)^{-1} \right\}
\tag{3.8}
$$

while for the second we use eq. (3.2) with $\boldsymbol{A} = \boldsymbol{\Sigma}$ and $\boldsymbol{B} = \boldsymbol{I}_d$:

$$
\begin{aligned}
\frac{\lambda \sigma^2}{n} \operatorname{Tr}\left\{ \boldsymbol{\Sigma} \left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d \right)^{-2} \right\} &\asymp \frac{\sigma^2 \kappa(\lambda)^2}{n\lambda} \operatorname{Tr}\left\{ \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma} + \kappa(\lambda) \boldsymbol{I}_d \right)^{-2} \right\} \left[1 + \frac{\operatorname{Tr}\left\{ \boldsymbol{\Sigma}^2 \left(\boldsymbol{\Sigma} + \kappa(\lambda) \boldsymbol{I}_d \right)^{-2} \right\}}{n - \operatorname{Tr}\left\{ \boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \kappa(\lambda) \boldsymbol{I}_d)^{-2} \right\}} \right] \\
&= \frac{\sigma^2 \kappa(\lambda)^2}{n\lambda} \frac{\operatorname{Tr}\left\{ \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma} + \kappa(\lambda) \boldsymbol{I}_d \right)^{-2} \right\}}{1 - \frac{1}{n} \operatorname{Tr}\left\{ \boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \kappa(\lambda) \boldsymbol{I}_d)^{-2} \right\}}
\end{aligned}
\tag{3.9}
$$

Putting together:

$$
V(\boldsymbol{X}, \lambda, \sigma^2) \asymp \frac{\sigma^2 \kappa(\lambda)}{n\lambda} \left[\operatorname{Tr}\left\{ \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma} + \kappa(\lambda) \boldsymbol{I}_d \right)^{-1} \right\} - \kappa(\lambda) \frac{\operatorname{Tr}\left\{ \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma} + \kappa(\lambda) \boldsymbol{I}_d \right)^{-2} \right\}}{1 - \frac{1}{n} \operatorname{Tr}\left\{ \boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \kappa(\lambda) \boldsymbol{I}_d)^{-2} \right\}} \right]
\tag{3.10}
$$

Noting that we can write:

$$
\begin{aligned}
\kappa \operatorname{Tr}\left\{ \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma} + \kappa \boldsymbol{I}_d \right)^{-2} \right\} &= \operatorname{Tr}\left\{ (\boldsymbol{\Sigma} + \kappa \boldsymbol{I}_d - \boldsymbol{\Sigma}) \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma} + \kappa \boldsymbol{I}_d \right)^{-2} \right\} \\
&= \operatorname{Tr}\left\{ \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma} + \kappa \boldsymbol{I}_d \right)^{-1} \right\} - \operatorname{Tr}\left\{ \boldsymbol{\Sigma}^2 \left(\boldsymbol{\Sigma} + \kappa \boldsymbol{I}_d \right)^{-2} \right\}
\end{aligned}
\tag{3.11}
$$

We can equate the denominator and simplify the first term:

$$
\begin{aligned}
V(\boldsymbol{X}, \lambda, \sigma^2) &\asymp \frac{\sigma^2 \kappa(\lambda)}{n\lambda} \operatorname{Tr}\left\{ \boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \kappa(\lambda) \boldsymbol{I}_d)^{-2} \right\} \frac{1 - \frac{1}{n} \operatorname{Tr}\left\{ \boldsymbol{\Sigma} (\boldsymbol{\Sigma} + \kappa(\lambda) \boldsymbol{I}_d)^{-1} \right\}}{1 - \frac{1}{n} \operatorname{Tr}\left\{ \boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \kappa(\lambda) \boldsymbol{I}_d)^{-2} \right\}} \\
&\overset{(a)}{=} \sigma^2 \frac{\operatorname{Tr}\left\{ \boldsymbol{\Sigma}^2 \left(\boldsymbol{\Sigma} + \kappa(\lambda) \boldsymbol{I}_d \right)^{-2} \right\}}{n - \operatorname{Tr}\left\{ \boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \kappa(\lambda) \boldsymbol{I}_d)^{-2} \right\}}
\end{aligned}
\tag{3.12}
$$

where in (a) we used the self-consistent equation eq. (3.3) to rewrite ${}^{1}/_{n} \operatorname{Tr}\left\{ \boldsymbol{\Sigma} (\boldsymbol{\Sigma} + \kappa \boldsymbol{\Sigma})^{-1} \right\} = 1 - {}^{\lambda}/_{\kappa}$. Summarising, this leads to the following result:

Proposition 1 (Asymptotic risk of ridge regression)**.** Under Assumption 1, the asymptotic excess risk of the ridge regressor eq. (1.1) converges, in the proportional limit $n, d \to \infty$ with $d/n \to \gamma > 0$ is given by:

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[R(\hat{\boldsymbol{\theta}}_\lambda)\right] - \sigma^2 \xrightarrow{a.s.} \mathcal{B}(\boldsymbol{\theta}_\star, \boldsymbol{\Sigma}, \lambda, \gamma) + \mathcal{V}(\boldsymbol{\Sigma}, \lambda, \sigma^2, \gamma), \text{ as } n, d \to \infty \tag{3.13}$$

where the asymptotic bias $\mathcal{B}$ and variance $\mathcal{V}$ are given by:

$$\begin{aligned}\mathcal{B}(\boldsymbol{\theta}_\star, \boldsymbol{\Sigma}, \lambda, \gamma) &= \frac{\kappa(\lambda)^2 \langle \boldsymbol{\theta}_\star, \boldsymbol{\Sigma}\left(\boldsymbol{\Sigma} + \kappa(\lambda)\boldsymbol{I}_d\right)^{-2} \boldsymbol{\theta}_\star \rangle}{1 - \frac{1}{n}\operatorname{Tr}\left\{\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \kappa(\lambda)\boldsymbol{I}_d)^{-2}\right\}} \\ \mathcal{V}(\boldsymbol{\Sigma}, \lambda, \sigma^2, \gamma) &= \sigma^2 \frac{\operatorname{Tr}\left\{\boldsymbol{\Sigma}^2\left(\boldsymbol{\Sigma} + \kappa(\lambda)\boldsymbol{I}_d\right)^{-2}\right\}}{n - \operatorname{Tr}\left\{\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \kappa(\lambda)\boldsymbol{I}_d)^{-2}\right\}}\end{aligned} \tag{3.14}$$

where $\kappa(\lambda) \geq 0$ is the unique solution of the following self-consistent equation:

$$1 - \frac{\lambda}{\kappa} = \frac{1}{n}\operatorname{Tr}\left\{\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \kappa\boldsymbol{I}_d)^{-1}\right\} \tag{3.15}$$

Remark 2. A few comments on this result are in order.

- Note that without loss of generality we can assume $\boldsymbol{\Sigma}$ to be a diagonal matrix $\boldsymbol{\Sigma} = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$. Therefore, a sufficient condition for different terms in proposition 1 to be well defined in the asymptotic limit is that the empirical spectral measure of $\boldsymbol{\Sigma}$ admits a limit:

  $$\frac{1}{d}\sum_{\lambda \in \operatorname{spec}(\boldsymbol{\Sigma})} \delta_\lambda \xrightarrow{\text{weakly}} \mu \tag{3.16}$$

  in which case all the traces can be written in terms of expectations with respect to μ, for example:

  $$\frac{1}{n}\operatorname{Tr}\{\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \kappa\boldsymbol{I}_d)^{-1}\} \to \gamma \int \mu(\mathrm{d}t)\frac{t}{t+\kappa}, \text{ as } n, d \xrightarrow{d/n \to \gamma} \infty. \tag{3.17}$$

- Nevertheless, we deliberately chose write the equations above in terms of traces since, despite being derived in the proportional regime, the universality of these formulas is remarkable. For instance, Cheng and Montanari (2024) have derived multiplicative, non-asymptotic rates for the limit above under fairly generic assumptions on the covariates. These formulas hold even in the $d \to \infty$ case where $\boldsymbol{\theta}_\star$ can be seen as an element of a Hilbert space and $\boldsymbol{\Sigma}$ a covariance operator, as it was first noted by Bordelon et al. (2020); Cui et al. (2021) and proven in (Misiakiewicz and Saeed, 2024) under some conditions on the tail of the covariance spectrum.

3.1 Degrees-of-freedom interpretation

Proposition 1 involve the following two quantities, known in the literature as the *degrees of freedom* of the matrix $\boldsymbol{\Sigma}$:

$$\operatorname{df}_1(\kappa) \coloneqq \operatorname{Tr}\{\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma} + \kappa\boldsymbol{I}_d\right)^{-1}\} = \sum_{j=1}^{d}\frac{\lambda_j}{\kappa + \lambda_j} \tag{3.18}$$

$$\operatorname{df}_2(\kappa) \coloneqq \operatorname{Tr}\{\boldsymbol{\Sigma}^2\left(\boldsymbol{\Sigma} + \kappa\boldsymbol{I}_d\right)^{-2}\} = \sum_{j=1}^{d}\frac{\lambda_j^2}{(\kappa + \lambda_j)^2} \tag{3.19}$$

The degrees-of-freedom is a widespread notion in the signal processing and kernel literature, where it is often used as a notion of *effective dimension* when comparing kernel operators defined on infinie dimensional Hilbert spaces, see for example (Zhang, 2005; Caponnetto and De Vito, 2007).

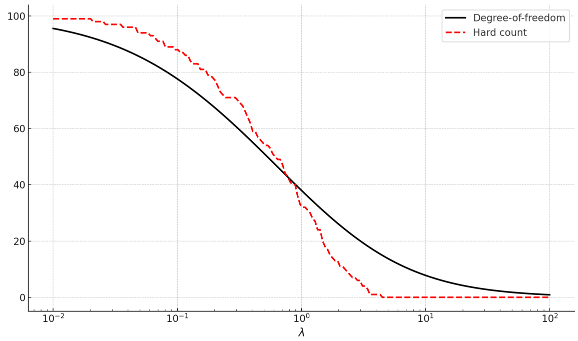

Figure 1: Degrees-of-freedom $\mathrm{df}_1(\lambda)$ and hard count $\phi(\lambda)$ as a function of $\lambda \geq 0$ for $\lambda_i \sim \exp(i)$ independently, $i \in [n]$

Note that $\mathrm{df}_1, \mathrm{df}_2$ are strictly decreasing functions of $\kappa \geq 0$, and since $\boldsymbol{\Sigma} \succeq 0$ we have:

$$0 \leq \mathrm{df}_2(\kappa) \leq \mathrm{df}_1(\kappa) \leq \mathrm{rank}(\boldsymbol{\Sigma}) \tag{3.20}$$

with equality on the right for $\kappa = 0$. The degrees-of-freedom $\mathrm{df}_1(\kappa), \mathrm{df}_2(\kappa)$ can be seen as a "soft count" of how many eigenvalues are larger than the parameter κ, since eigenvalues $\lambda_j \gg \kappa$ contribute to the sum, while eigenvalues $\lambda_j \ll \kappa$ are shrank. To make this relationship more quantitative, consider the *hard count* of how many eigenvalues of $\boldsymbol{\Sigma}$ are larger than a certain value κ:

$$\phi(\kappa) \coloneqq \sum_{j=1}^{d} \mathbf{1}_{\lambda_j \geq \kappa} = \#\{k : \lambda_k \geq \kappa\}, \tag{3.21}$$

Note $1 - \phi(\kappa)$ is the *cumulative distribution function* (c.d.f.) of the empirical spectral distribution $\hat{\mu}_{\boldsymbol{\Sigma}}$. This can also be written as an integral over $\hat{\mu}_{\boldsymbol{\Sigma}}$:

$$\phi(\kappa) = d \int_{\kappa}^{\infty} \hat{\mu}_{\boldsymbol{\Sigma}}(\mathrm{d}\lambda) = d \int_{\mathbb{R}} \mathbf{1}_{\lambda \geq \kappa}(\lambda)\ \hat{\mu}_{\boldsymbol{\Sigma}}(\mathrm{d}\lambda) \tag{3.22}$$

to be compared with:

$$\mathrm{df}_1(\kappa) = d \int_{\mathbb{R}} \frac{\lambda}{\lambda + \kappa} \hat{\mu}_{\boldsymbol{\Sigma}}(\mathrm{d}\lambda) \tag{3.23}$$

3.2 Equivalent denoising problem

The asymptotic formulas in proposition 1 have an intuitive interpretation in terms of an effectively denoising problem. To see this, consider the problem of retrieving $\boldsymbol{\theta}_\star \in \mathbb{R}^d$ from the following noisy observation:

$$\boldsymbol{u} = \boldsymbol{\Sigma}^{1/2} \boldsymbol{\theta}_\star + \frac{\tau}{\sqrt{n}} \boldsymbol{z}, \qquad \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d). \tag{3.24}$$

where $\tau > 0$ is the noise standard deviation. Then, the following regularised estimator:

$$\hat{\boldsymbol{\theta}}_{\text{den.}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\mathrm{argmin}} \left\{ ||\boldsymbol{u} - \boldsymbol{\Sigma}^{1/2} \boldsymbol{\theta}||_2^2 + \kappa_\star ||\boldsymbol{\theta}||_2^2 \right\} \tag{3.25}$$

is statistically equivalent to the one in proposition 1 in the proportional high-dimensional limit, as long as we identify:

$$\tau^2 = \sigma^2 + \mathbb{E}_{\boldsymbol{z}}[||\hat{\boldsymbol{\theta}}_{\text{den.}} - \boldsymbol{\theta}_\star||_{\boldsymbol{\Sigma}}^2] \tag{3.26}$$

To see this, note that the quadratic problem in eq. (3.25) has explicit solution:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\text{den.}} &= (\boldsymbol{\Sigma}+\kappa_\star \boldsymbol{I}_d)^{-1}\boldsymbol{\Sigma}^{1/2}\boldsymbol{u} \\ &= (\boldsymbol{\Sigma}+\kappa_\star \boldsymbol{I}_d)^{-1}\boldsymbol{\Sigma}^{1/2}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}_\star + \frac{\tau}{\sqrt{n}}\boldsymbol{z}\right) && (3.27) \\ &= \boldsymbol{\theta}_\star - \kappa_\star(\boldsymbol{\Sigma}+\kappa_\star \boldsymbol{I}_d)^{-1}\boldsymbol{\theta}_\star + \frac{\tau}{\sqrt{n}}(\boldsymbol{\Sigma}+\kappa_\star \boldsymbol{I}_d)^{-1}\boldsymbol{\Sigma}^{1/2}\boldsymbol{z} && (3.28)\end{aligned}$$

Hence, we have:

$$\mathbb{E}_{\boldsymbol{z}}\left[||\hat{\boldsymbol{\theta}}_{\text{den.}}-\boldsymbol{\theta}_\star||^2_{\boldsymbol{\Sigma}}\right] = \kappa_\star^2\langle\boldsymbol{\theta}_\star, \boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\kappa_\star \boldsymbol{I}_d)^{-2}\boldsymbol{\theta}_\star\rangle + \frac{\tau^2}{n}\operatorname{Tr}\left\{\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\kappa_\star \boldsymbol{I}_d)^{-2}\right\} \tag{3.29}$$

It remains to find τ^2. For that, we insert this in eq. (3.26) and solve for τ^2 to yield:

$$\tau^2 = \frac{\sigma^2 - \kappa_\star^2\langle\boldsymbol{\theta}_\star, \boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\kappa_\star \boldsymbol{I}_d)^{-2}\boldsymbol{\theta}_\star\rangle}{1-\frac{1}{n}\operatorname{Tr}\left\{\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\kappa_\star \boldsymbol{I}_d)^{-2}\right\}} \tag{3.30}$$

Inserting this back into eq. (3.29) give us:

$$\mathbb{E}_{\boldsymbol{z}}\left[||\hat{\boldsymbol{\theta}}_{\text{den.}}-\boldsymbol{\theta}_\star||^2_{\boldsymbol{\Sigma}}\right] = \frac{\kappa_\star^2\langle\boldsymbol{\theta}_\star, \boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}+\kappa_\star \boldsymbol{I}_d\right)^{-2}\boldsymbol{\theta}_\star\rangle}{1-\frac{1}{n}\operatorname{Tr}\left\{\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\kappa_\star \boldsymbol{I}_d)^{-2}\right\}} + \sigma^2\frac{\operatorname{Tr}\left\{\boldsymbol{\Sigma}^2\left(\boldsymbol{\Sigma}+\kappa_\star \boldsymbol{I}_d\right)^{-2}\right\}}{n-\operatorname{Tr}\left\{\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\kappa_\star \boldsymbol{I}_d)^{-2}\right\}} \tag{3.31}$$

which is precisely the expression for the asymptotic excess risk from proposition 1.

Remark 3. Two comments are in order:

- The equivalent denoising problem gives a nice interpretation of the different quantities involved in proposition 1. For instance, $\kappa_\star(\lambda)$ appears exactly in the same role as λ in section 1, and therefore plays the role of an effective, self-induced ℓ_2-regularisation.
- In fact, the self-induced regularisation is always larger than the original ridge regularisation. This can be seen from studying the behaviour of the self-consistent eq. (3.15) for both $\gamma > 1$ and $\gamma < 1$:
  - For $\gamma < 1$ ($d < n$), we have $\mathrm{df}_1(\kappa(\lambda)) \leq d < n$, and therefore the self-consistent eq. (3.15) implies:

    $$0 \leq 1 - \frac{\lambda}{\kappa} \leq \gamma. \tag{3.32}$$

    Since $\lambda \mapsto \kappa_\star(\lambda)$ is non-decreasing, the solution must satisfy:

    $$\kappa_\star(\lambda) \in \left[\lambda, \frac{\lambda}{1-\gamma}\right], \qquad . \tag{3.33}$$

    with in particular $\kappa(0) = 0$.
  - For $\gamma > 1$ ($d > n$), the self-consistent eq. (3.15) has a solution $\kappa_\star(0) > 0$ defined by the implicit equation:

    $$\mathrm{df}_1(\kappa(0)) = n \tag{3.34}$$

  In other words, the effective regularisation is always larger or equal the original regularisation: $\kappa_\star \geq \lambda$. Since $\kappa \mapsto \mathrm{df}_1(\kappa)$ is a convex map, by Jensen's inequality:

  $$\mathrm{df}_1(\kappa(\lambda)) \leq \frac{\operatorname{Tr}\boldsymbol{\Sigma}}{\kappa(\lambda) + {}^1\!/\!_d\operatorname{Tr}\boldsymbol{\Sigma}} \leq \frac{\operatorname{Tr}\boldsymbol{\Sigma}}{\kappa(\lambda)} \tag{3.35}$$

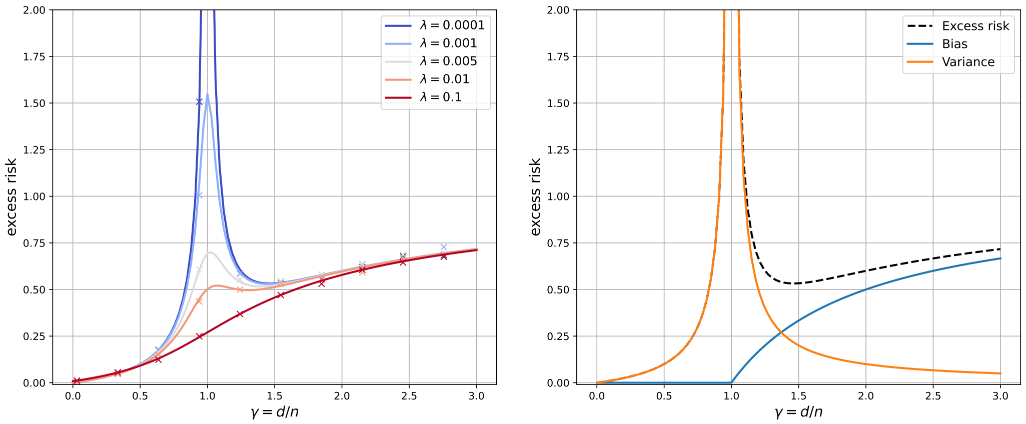

Figure 2: Excess risk of ridge regression as a function of $\gamma = {}^{d}/_{n}$ for $\sigma^2 = 0.1$, $||\boldsymbol{\theta}_\star||_2^2 = 1$ and isotropic covariates $\boldsymbol{\Sigma} = \boldsymbol{I}_d$. (**Left**) Increasing values of λ. (**Right**) Bias-variance decomposition of the excess risk for the ridge interpolator $\lambda = 0^+$ (a.k.a. ordinary least-squares estimator).

which from eq. (3.15) implies that:

$$0 \leq 1 - \frac{\lambda}{\kappa(\lambda)} \leq \frac{\operatorname{Tr} \boldsymbol{\Sigma}}{\kappa(\lambda)} \tag{3.36}$$

and again, since $\lambda \mapsto \kappa(\lambda)$ is non-decreasing:

$$\kappa_\star(\lambda) \in \left[\lambda, \lambda + \frac{\operatorname{Tr} \boldsymbol{\Sigma}}{n}\right] \tag{3.37}$$

Therefore, in both cases we have an effective regularisation larger than the original ridge regularisation: $\kappa_\star(\lambda) \geq \lambda$.

- Similarly, ${}^{\tau}/_{\sqrt{n}}$ plays a similar role to the noise level σ. It is interesting to note that in the effective denoising problem, the effective noise level is itself a function of the excess risk, and is a decreasing function of the number of samples n.

The denoising problem in eq. (3.24) and (3.25) is also known in the statistical literature as a *sequence model*, see for example (Tsybakov, 2008).

3.3 Case study: isotropic covariates

Let's explore our result on possibly the simplest setting, the case of isotropic covariates $\boldsymbol{\Sigma} = \boldsymbol{I}_d$. In this case, the self-consistent eq. (3.15) is simply a quadratic equation:

$$1 - \frac{\lambda}{\kappa} = \frac{\gamma}{1+\kappa} \tag{3.38}$$

This admits two solutions:

$$\kappa_\pm(\lambda) = \frac{1}{2}\left(\lambda - 1 + \gamma \pm \sqrt{(1-\gamma-\lambda)^2 + 4\lambda}\right) \tag{3.39}$$

of which only the positive branch $\kappa_\star(\lambda) := \kappa_+(\lambda)$ is positive for $\lambda \geq 0$. Further, the bias and variance simplify to:

$$\begin{aligned}\mathcal{B}(\boldsymbol{\theta}_\star, \lambda, \gamma) &= \frac{\kappa_\star(\lambda)^2}{(1+\kappa_\star(\lambda))^2 - \gamma}||\boldsymbol{\theta}_\star||_2^2 \\ \mathcal{V}(\lambda, \sigma^2, \gamma) &= \frac{\sigma^2\gamma}{(1+\kappa_\star(\lambda))^2 - \gamma}\end{aligned} \tag{3.40}$$

Figure 2 (left) illustrates the asymptotic risk as a function of $\gamma = {}^{d}/_{n}$ for different values on regularisation λ. Note that for small values of λ, the excess risk becomes a non-monotonic function of γ, with a divergence developing at $\gamma = 1$ as $\lambda \to 0^{+}$. On fig. 2 (right), we plot the bias and variance contribution to the excess risk in this limit, which shows that this divergence is mainly driven by the variance. This behaviour can be understood from the explicit solution eq. (3.39). Note that:

$$\lim_{\lambda \to 0^{+}} \kappa_{\star}(\lambda) = \frac{1}{2}(\gamma - 1 + |\gamma - 1|) = \begin{cases} \gamma - 1 & \gamma > 1 \\ 0 & \gamma \leq 1 \end{cases} \tag{3.41}$$

and therefore:

$$\lim_{\lambda \to 0^{+}} \mathcal{B}(\boldsymbol{\theta}_{\star}, \lambda, \gamma) = \begin{cases} 0 & \gamma \leq 1 \\ 1 - \frac{1}{\gamma} & \gamma > 1 \end{cases} \tag{3.42}$$

$$\lim_{\lambda \to 0^{+}} \mathcal{V}(\lambda, \sigma^{2}, \gamma) = \begin{cases} \frac{\gamma\sigma^{2}}{1-\gamma} & \gamma < 1 \\ \frac{\sigma^{2}}{\gamma - 1} & \gamma > 1 \end{cases} \tag{3.43}$$

with a divergence going as $\mathcal{V} \asymp O({}^{1}/_{|1-\gamma|})$ at around $\gamma = 1$.

Remark 4. A few comments are in order.

- Note that the $\gamma < 1$ solution is incredibly close to the non-asymptotic expression we found directly by looking at the ordinary least-squares estimator in section 2.1 for $n > d + 1$, with perfect agreement at the limit.

- However, the exact asymptotic formula also give us the behaviour of the least-square solution in the $\gamma \geq 1$ regime. The first curious observation is that at $\gamma = 1$ the variance blows up as $O(|\gamma - 1|^{-1})$, and indeed this is precisely the case where the expected value of the inverse Wishart distribution ceases to exist.

- Consistently to our general discussion in remark 3, for $\gamma > 1$ we have $\kappa_{\star}(\lambda = 0^{+}) = \gamma - 1 > 0$. Recall that from the equivalent denoising problem in eq. (3.25), $\kappa_{\star}$ plays the role of the ridge regularisation. This means that we have a non-zero, *self-induced regularisation* in the region $\gamma > 1$. The larger γ, the stronger is this regularisation.

- In the regime $\gamma > 1$, the bias is also non-zero. This is intuitive since we are choosing one (the minimum norm) among all the existing zero loss solutions in this regime. Curiously, the variance also decreases for $\gamma > 1$.

- Although the singularity at $\gamma = 1$ resembles the double descent phenomenon observed in neural networks, the minimum of the risk is achieved at the $\gamma < 1$ region, meaning that it is not beneficial to take $d > n$. In other words, the minimum norm solution overfits.

3.4 Case study: the double descent phenomenon

The isotropic case captures the non-monotonic behaviour of the excess risk around the interpolation threshold, but different from neural networks the least-norm solution in this case still overfits in the "overparametrised" regime.

To discuss a model that captures the benign overfitting in neural networks, we need to consider a richer, anisotropic model. Note that one of the limitations of the isotropic case is that the number of parameters in the model $f(\boldsymbol{x}; \boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \boldsymbol{x} \rangle$ is the same as the dimensionality of the covariates. Since the covariates are isotropic, increasing the number of parameters is akin to increasing the dimensionality of the covariate space, effectively decreasing the signal-to-noise ratio or sample complexity ${}^{n}/_{d}$ of the problem. This is different from neural networks with more than one-layer, e.g. $f(\boldsymbol{x}; \theta) = \langle \boldsymbol{a}, \sigma(\boldsymbol{W}\boldsymbol{x}) \rangle$

with $\boldsymbol{W} \in \mathbb{R}^{p\times d}$, where we can increase the number of parameters by increasing the width p without changing the input dimension d.

We now introduce a model that seeks to mimic the behaviour of neural networks. Labels are generated from an isotropic latent Gaussian variable:

$$y_i = \langle \boldsymbol{\beta}_\star, \boldsymbol{z}_i \rangle + \xi_i, \qquad \boldsymbol{z}_i \sim \mathcal{N}(0, \boldsymbol{I}_d), \qquad \xi_i \sim \mathcal{N}(0, \tau^2) \tag{3.44}$$

However, the statistician does not observe the latent covariates, but rather a noisy projection:

$$\boldsymbol{x}_i = \boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{u}_i \tag{3.45}$$

where $\boldsymbol{W} \in \mathbb{R}^{p\times d}$ is a fixed matrix and $\boldsymbol{u}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ i.i.d. In other words, the statistician performs regression on the training data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^{p+1} : i = 1, \dots, n\}$. This model, known under the umbrella of *hidden manifold model* (Goldt et al., 2020) or *latent space model* (Hastie et al., 2022), models the well-known *manifold hypothesis* that high-dimensional data depends on a few "relevant features" lying on a lower-dimensional manifold.

This model is a particular case of the one introduced in assumption 1. To see this, note that $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ are jointly Gaussian variables:

$$\begin{bmatrix} y_i \\ \boldsymbol{x}_i \\ \boldsymbol{z}_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ \boldsymbol{0}_p \\ \boldsymbol{0}_d \end{bmatrix}, \begin{bmatrix} \tau^2 + ||\boldsymbol{\beta}||_2^2 & (\boldsymbol{W}\boldsymbol{\beta}_\star)^\top & \boldsymbol{\beta}_\star^\top \\ \boldsymbol{W}\boldsymbol{\beta}_\star & \boldsymbol{W}\boldsymbol{W}^\top + \boldsymbol{I}_p & \boldsymbol{W}^\top \\ \boldsymbol{\beta}_\star & \boldsymbol{W} & \boldsymbol{I}_d \end{bmatrix} \right), \qquad \text{i.i.d.} \tag{3.46}$$

Therefore, by Gaussian conditioning we have:

$$y_i | \boldsymbol{x}_i \sim \mathcal{N}\left((\boldsymbol{W}^\top\boldsymbol{W} + \boldsymbol{I}_p)^{-1}\boldsymbol{W}\boldsymbol{\beta}_\star, \boldsymbol{x}_i, \tau^2 + \langle \boldsymbol{\beta}_\star, (\boldsymbol{W}^\top\boldsymbol{W} + \boldsymbol{I}_d)^{-1}\boldsymbol{\beta}_\star \rangle \right) \tag{3.47}$$

In other words, this model is statistically equivalent to:

$$y_i = \langle \boldsymbol{\theta}_\star, \boldsymbol{x}_i \rangle + \varepsilon_i \tag{3.48}$$

with:

$$\boldsymbol{\theta}_\star = (\boldsymbol{W}\boldsymbol{W}^\top + \boldsymbol{I}_p)^{-1}\boldsymbol{W}\boldsymbol{\beta}_\star \tag{3.49}$$

and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is an effective Gaussian noise with variance $\sigma^2 = \tau^2 + \langle \boldsymbol{\beta}_\star, (\boldsymbol{W}^\top\boldsymbol{W} + \boldsymbol{I}_d)^{-1}\boldsymbol{\beta}_\star \rangle$.

Remark 5. From the perspective of the latent model, the effective noise accounts for both for the label noise τ^2 but also for the model misspecification (i.e. the fact that we are fitting in a p dimensional space instead of the d dimensional space the signal lives). In particular, when $p = 0$ we have $\sigma^2 = \tau^2 + ||\boldsymbol{\beta}_\star||_2^2$. Note that beyond the anisotropy, a key difference of the model above is that the target weights in eq. (3.49) are correlated with the top right eigenvectors of $\boldsymbol{W}$.

For concreteness, let's look at a simple particular case:

- We assume that $n, p, d \to \infty$ at constant rates $\gamma = {}^{p}/_{n}$ and $\alpha = {}^{n}/_{d}$.
- We assume $\boldsymbol{\beta}_\star \in \mathbb{S}^{d-1}$, i.e. $||\boldsymbol{\beta}_\star||_2 = 1$.
- We assume that $\boldsymbol{W} \in \mathbb{R}^{p\times d}$ is given by:

$$\boldsymbol{W} = \begin{cases} \begin{bmatrix} \sqrt{{}^{p}/_{d}}\boldsymbol{I}_d \\ \boldsymbol{0}_{(p-d)\times d} \end{bmatrix} & \text{if } p \geq d \\ \begin{bmatrix} \boldsymbol{I}_p & \boldsymbol{0}_{p\times(d-p)} \end{bmatrix} & \text{if } p < d \end{cases} \tag{3.50}$$

Note this implies:

$$\boldsymbol{W}^\top\boldsymbol{W} = \begin{cases} {}^{p}/_{d}\boldsymbol{I}_d & \text{if } p \geq d \\ \begin{bmatrix} \boldsymbol{I}_p & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} & \text{if } p < d \end{cases}, \qquad \boldsymbol{W}\boldsymbol{W}^\top = \begin{cases} \begin{bmatrix} {}^{p}/_{d}\boldsymbol{I}_d & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} & \text{if } p \geq d \\ \boldsymbol{I}_p & \text{if } p < d \end{cases} \tag{3.51}$$

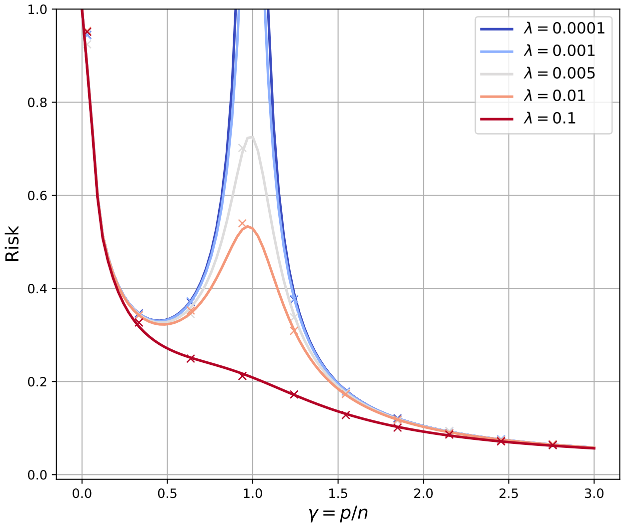

Figure 3: Risk of ridge regression as a function of $\gamma = {}^{p}/_{n}$ for the latent space model defined in Section 3.4 for $\tau^2 = 0$, $\alpha = {}^{n}/_{d} = 10$. Solid curves show the theoretical result, obtained from solving the self-consistent eq. (3.54), and crosses are finite size simulations with $d = 100$.

Remark 6. For the third assumption, any full-rank matrix with fixed Frobenius norm $||\boldsymbol{W}||_{\mathrm{F}}^2 = p$ would be equally good. Equation (3.50) is the simplest such example.

Under these assumptions, we have the following simplification:

$$\boldsymbol{\theta}_\star = (\boldsymbol{W}\boldsymbol{W}^\top + \boldsymbol{I}_p)^{-1}\boldsymbol{W}\boldsymbol{\beta}_\star = \boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W} + \boldsymbol{I}_d)^{-1}\boldsymbol{\beta}_\star = \frac{\boldsymbol{W}\boldsymbol{\beta}_\star}{1+\alpha\gamma}$$

$$\sigma^2 = \begin{cases} \tau^2 + \frac{1}{1+\alpha\gamma} & \text{if } p \geq d \\ \tau^2 + 1 - {}^{\alpha\gamma}/_2 & \text{if } p < d \end{cases} \tag{3.52}$$

Further, we can also simplify the expression of the degrees-of-freedom:

$$\mathrm{df}_a(\kappa) = \mathrm{Tr}\left\{\boldsymbol{\Sigma}^a(\boldsymbol{\Sigma} + \kappa\boldsymbol{I}_p)^{-a}\right\} = \begin{cases} d\left(\frac{\alpha\gamma+1}{\alpha\gamma+1+\kappa}\right)^a + \frac{p-d}{(1+\kappa)^a} & \text{if } p \geq d \\ p\left(\frac{\alpha\gamma+1}{\alpha\gamma+1+\kappa}\right)^a & \text{if } p < d \end{cases}, \qquad a = 1,2 \tag{3.53}$$

Therefore, in the proportional high-dimensional limit, the self-consistent equation eq. (3.15) reads:

$$1 - \frac{\lambda}{\kappa} = \min(\gamma, {}^{1}/_{\alpha})\frac{\alpha\gamma+1}{\alpha\gamma+\kappa+1} + \left(\gamma - \frac{1}{\alpha}\right)_+ \frac{1}{1+\kappa} \tag{3.54}$$

As before, this is a quadratic equation that can be solved explicitly. However, differently from the isotropic case the expressions are cumbersome, so we refrain from writing them here, and focus instead on the discussion of the interpolator $\lambda = 0^+$.

For $\gamma < 1$ ($p < n$), $\kappa(0) = 0$ is a solution of eq. (3.54), and we have $\mathcal{B} = 0$. Then, the risk is completely dominated by the variance:

$$R = \sigma^2 + \mathcal{V} = \sigma^2\left(1 + \frac{\gamma}{1-\gamma}\right) = \begin{cases} \left(\tau^2 + \frac{1}{1+\alpha\gamma}\right)\frac{1}{1-\gamma} & \text{if } p \geq d \\ \left(\tau^2 + 1 - \frac{\alpha\gamma}{2}\right)\frac{1}{1-\gamma} & \text{if } p < d \end{cases} \tag{3.55}$$

Note that as $\gamma \to 1^+$, the variance (and hence the risk) diverge as $(1-\gamma)^{-1}$, just as in the isotropic case studied in section 3.3. Assuming that $\alpha > 1$, for $\gamma > 1$ the expressions for the bias and the variance read:

$$\begin{aligned}\mathcal{B}(\gamma,\alpha,\tau^2) &= \kappa_\star(0)^2\frac{A}{1-B}\\ \mathcal{V}(\gamma,\alpha,\tau^2) &= \left(\tau^2+\frac{1}{\alpha\gamma+1}\right)\frac{B}{1-B}.\end{aligned} \tag{3.56}$$

where we have defined:

$$A(\gamma,\alpha,\tau^2) = \langle\boldsymbol{\theta}_\star, \boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}+\kappa(\lambda)\boldsymbol{I}_d\right)^{-2}\boldsymbol{\theta}_\star\rangle = \frac{\alpha\gamma}{(\alpha\gamma+1)(\alpha\gamma+\kappa_\star(0)+1)^2} \tag{3.57}$$

$$B(\gamma,\alpha,\tau^2) = \frac{1}{n}\mathrm{df}_2 = \frac{1}{\alpha}\left(\frac{\alpha\gamma+1}{\alpha\gamma+1+\kappa_\star(0)}\right)^2 + \left(\gamma-\frac{1}{\alpha}\right)\frac{1}{(1+\kappa_\star(0))^2} \tag{3.58}$$

and $\kappa_\star(0) > 0$ is given by:

$$\kappa_\star(0) = \gamma - 1 - \frac{\alpha\gamma}{2} + \frac{1}{2}\sqrt{(4+\alpha^2)\gamma^2 - 4\gamma} \tag{3.59}$$

Figure 3 shows the risk of ridge regression for the latent variable model for different values of the regularisation λ. For $\lambda \approx 0^+$ (interpolator), we can clearly see the divergence at $\gamma \to 1$ discussed above, also known as *double descent* or *interpolation peak*. Differently from the isotropic case in section 3.3, for $\gamma > 1$, the risk is a decreasing function of γ, meaning that overparametrisation does not hurt generalisation. Moreover, the minimal risk is achieved at large parametrisation $\gamma \to \infty$, when the predictor perfectly interpolates the training data. This phenomenon is known as *benign overfitting*, and is at odds with the classical statistical intuition that interpolating the training data always hurts generalisation.

Remark 7 (Historical note)**.** Both the interpolation peak (Opper et al., 1990; Krogh and Hertz, 1991) and observation that neural networks continue to improve their performance as the number of neurons is increased are quite old (Geman et al., 1992), see (Loog et al., 2020) for a detailed historical discussion. These results were mostly forgotten, and were independently rediscovered in the recent development of machine learning theory driven by the deep learning boom (Zhang et al., 2021). The term "double descent" was coined by Belkin et al. (2019), motivated by different empirical works that observed an interpolation peak in the context of neural networks (Nakkiran et al., 2021; Spigler et al., 2019).

4 To go further

4.1 Random Features Model

The double descent curve discussed in Section 3.4 is not a particular feature of the latent variables model, and manifests in different problems of interest in machine learning. One important example is the *random features model*, where the predictor is given by:

$$f_\theta(\boldsymbol{x}) = \langle\boldsymbol{a}, \sigma(\boldsymbol{W}\boldsymbol{x})\rangle \tag{4.1}$$

where $\boldsymbol{W} \in \mathbb{R}^{p\times d}$ is a full-rank matrix, typically taken to be random, and the weights $\boldsymbol{\theta} \in \mathbb{R}^p$ are trained by ridge regression:

$$\begin{aligned}\hat{\boldsymbol{a}}_\lambda(\boldsymbol{X},\boldsymbol{y}) &= \underset{\boldsymbol{a}\in\mathbb{R}^p}{\operatorname{argmin}}\frac{1}{2n}\sum_{i=1}^{n}(y_i - \langle\boldsymbol{a},\sigma(\boldsymbol{W}\boldsymbol{x})\rangle)^2 + \frac{\lambda}{2}||\boldsymbol{a}||_2^2\\ &= \frac{1}{n}\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}+\lambda\boldsymbol{I}_p\right)^{-1}\boldsymbol{\Phi}^\top\boldsymbol{y}\end{aligned} \tag{4.2}$$

where we have defined the features matrix $\boldsymbol{\Phi} = \sigma(\boldsymbol{X}\boldsymbol{W}^\top) \in \mathbb{R}^{p\times d}$. Note that this model is equivalent to a two-layer neural network of width p with frozen first layer weights. Indeed, it has been widely studied in the literature as a proxy model for neural networks in the lazy regime.[2]

The challenge of studying this model is that the features matrix $\boldsymbol{\Phi}$ is not a Gaussian matrix, as it was the case for the latent variable model discussed in section 3.4. Nevertheless, quite remarkably Theorem 1 can still be applied to characterise the asymptotic properties of the random features model in the proportional regime, thanks to a phenomenon known as *Gaussian universality*, and which was first discussed in this context by Mei and Montanari (2022); Gerace et al. (2020). We now give a brief intuitive discussion, referring the reader interested in the details to the original literature.

The ridge operator in eq. (4.2):

$$\boldsymbol{y} \in \mathbb{R}^n \mapsto \frac{1}{n}\left(\frac{1}{n}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \lambda\boldsymbol{I}_p\right)^{-1}\boldsymbol{\Phi}^\top\boldsymbol{y} \tag{4.3}$$

projects the response onto the column-space of $\text{Image}(\boldsymbol{\Phi}^\top) \subset \mathbb{R}^p$, which is a linear subspace of the feature space. To see this mathematically, denote by $\boldsymbol{\Phi} = \sum_{j=1}^r \lambda_j \boldsymbol{u}_j\boldsymbol{v}_j^\top$ the singular-value decomposition of the features $\boldsymbol{\Phi}$ with $r := \text{rank}(\boldsymbol{\Phi}) \leq \min(n, p)$. Then, we can re-write eq. (4.2) as:

$$\hat{\boldsymbol{a}}_\lambda(\Phi, \boldsymbol{y}) = \sum_{j=1}^r \frac{\lambda_j}{\lambda_j^2 + n\lambda}\langle \boldsymbol{u}_j, \boldsymbol{y}\rangle \boldsymbol{v}_j \tag{4.4}$$

Therefore, assuming that $y_i = f_\star(\boldsymbol{x}) + \varepsilon_i$ for some target function $f_\star : \mathbb{R}^d \to \mathbb{R}$, the predictor $f(\boldsymbol{x}; \hat{\boldsymbol{a}}_\lambda) = \langle \hat{\boldsymbol{a}}_\lambda, \boldsymbol{\varphi}(\boldsymbol{x})\rangle$ can learn at best a linear component of the target function $f_\star$ in the space spanned by the features $\boldsymbol{\varphi}(\boldsymbol{x})$. For instance, in the vanilla ridge case $\boldsymbol{\varphi}(\boldsymbol{x}) = \boldsymbol{x}$ this would imply that only a linear component of the target can be learned: $f_\star(\boldsymbol{x}) = \langle \boldsymbol{\beta}_\star, \boldsymbol{x}\rangle + f_\star^{>1}(\boldsymbol{x})$, with the non-linear component $f_\star^{>1}$ effectively behaving as part of the label noise ε_i when projected on $\hat{\boldsymbol{a}}_\lambda$. A non-linear feature map $\boldsymbol{\varphi}(\boldsymbol{x})$ therefore allows, in principle, to learn higher order, non-linear components.

To make this discussion more concrete, it is useful to decompose the target function in an orthonormal basis with respect to the distribution of the covariates. Since we assume $\boldsymbol{x}_i \sim \mathcal{N}(0, {}^1/_d\boldsymbol{I}_d)$, this is given by the Hermite polynomials:

$$f_\star(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^d} c_{\boldsymbol{\alpha}} h_{\boldsymbol{\alpha}}(\boldsymbol{x}) \tag{4.5}$$

where $h_{\boldsymbol{\alpha}}(\boldsymbol{x})$ are the Hermite tensors, which form an orthonormal basis of $L^2(\mathbb{R}^d, \gamma_d)$ — where we denote γ_d the Gaussian p.d.f. in dimension d. This basis induces an orthogonal decomposition of $L^2(\mathbb{R}^d, \gamma_d) = \bigoplus_{\ell\geq 1} V_\kappa$, where V_κ is the linear space spanned by polynomials of degree $\ell = |\boldsymbol{\alpha}|$. The coefficients $c_{\boldsymbol{\alpha}}$ quantify how much of the total energy of the target $||f_\star||_{\gamma_d}^2 = \sum_{\boldsymbol{\alpha}} c_{\boldsymbol{\alpha}}^2$ lies in each subspace.

Assuming the features $\boldsymbol{\Phi}$ are full-rank $r = \min(n, p)$[3], since the ridge predictor in eq. (4.4) spans a linear subspace of dimension r, a naive power counting suggests that to learn the component of the target in subspace V_ℓ requires $r = O(d^\ell)$, with the minimum between the number of samples n and the width p being the bottleneck for approximating V_ℓ. Therefore, in a polynomial scaling regime $n, p = \Theta(d^\ell)$, we can learn at best a degree ℓ polynomial approximation of the target function $f_\star$. In particular, under the proportional asymptotics discussed in Section 3, it is enough to consider a linear target function $f_\star(\boldsymbol{x}) = \langle \boldsymbol{\beta}_\star, \boldsymbol{x}\rangle$.

⚠ It is important to keep in mind the discussion in this section is specific to ridge regression.

[2] Although, as we have seen in Lecture 4, random features are only one component of the kernel in the lazy regime, the other being the NTK.

[3] For instance, for $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, {}^1/_d\boldsymbol{I}_d)$ and $\boldsymbol{w}_j \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, $\boldsymbol{\Phi} = \sigma(\boldsymbol{X}\boldsymbol{W}^\top)$ will be a full-rank matrix with high-probability with respect to $\boldsymbol{X}, \boldsymbol{W}$.

An important consequence of this discussion is that in the high-dimensional limit a random features map sees the target function at a limited resolution. Considering the expansion of the feature map in the Hermite basis:

$$\varphi_j(\boldsymbol{x}) = \sigma\left(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle\right) = \sum_{\ell\geq 0} b_\ell h_\ell(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle) \tag{4.6}$$

Its first and second moments are given by:

$$\mathbb{E}_{\boldsymbol{x}}[\sigma\left(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle\right)] = b_0 \tag{4.7}$$

$$\mathbb{E}_{\boldsymbol{x}}[\sigma\left(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle\right)\sigma\left(\langle \boldsymbol{w}_{0,k}, \boldsymbol{x}\rangle\right)] = \sum_{\ell\geq 0} b_\ell^2 \left(\frac{\langle \boldsymbol{w}_j, \boldsymbol{w}_{0,k}\rangle}{d}\right)^\ell \tag{4.8}$$

In particular, note that if $\boldsymbol{w}_j \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, with high-probability ${}^1\!/\!{}_d\langle \boldsymbol{w}_j, \boldsymbol{w}_k\rangle = O(d^{-1/2})$ for $j \neq k$ and ${}^1\!/\!{}_d||\boldsymbol{w}_j||^2 = 1$, meaning that to leading order in d, the features population covariance $\boldsymbol{\Sigma} = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{\varphi}(\boldsymbol{x})\boldsymbol{\varphi}(\boldsymbol{x})^\top]$ is given by:[4]

$$\boldsymbol{\Sigma} = b_0^2 \boldsymbol{1}_p\boldsymbol{1}_p^\top + b_1^2 \frac{\boldsymbol{W}_0\boldsymbol{W}_0^\top}{d} + b_\star^2 \boldsymbol{I}_p + o_{\mathbb{P},d}(1) \tag{4.9}$$

where we have defined:

$$b_\star^2 = \sum_{\ell\geq 2} b_\ell^2 = \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\sigma(z)^2\right] - b_0^2 - b_1^2 \tag{4.10}$$

This implies that under the proportional high-dimensional limit, the features $\boldsymbol{\varphi}(\boldsymbol{x}) = \sigma(\boldsymbol{W}_0\boldsymbol{x})$ have the same first and second moments as the following Gaussian covariates:

$$\boldsymbol{g} = b_0\boldsymbol{1}_p + b_1\boldsymbol{W}_0\boldsymbol{x} + b_\star\boldsymbol{u}, \qquad \boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p) \tag{4.11}$$

This is exactly the latent variable model we studied in section 3.4! This suggests that in the proportional high-dimensional limit, we can trade the study of the original non-linear random features model in eq. (4.2) for the study of an equivalent Gaussian covariate model. This is an instance of a more general universality phenomenon, known as a Gaussian equivalence. We refer the reader for the original works for a full discussion of Gaussian universality in this context (Mei and Montanari, 2022; Gerace et al., 2020; Goldt et al., 2022; Hu and Lu, 2022; Montanari and Saeed, 2022)

4.2 Benign overfitting

Whether a predictor can benignantly overfit the data will crucially depends on the geometry of the covariates. Intuitively, benign overfitting means that the predictor is able to fit the signal in the data (so it can generalise) while also fitting the noise (so it can interpolate). This is only possible when the signal is strong, and lies in a sufficiently small subspace of the covariate space, ensuring that there is enough "room" left to accommodate the noise.

Note that this is precisely the case in the latent variable model discussed above: the signal $\boldsymbol{\beta}_\star$ is in $\mathbb{R}^d$ while the predictor and the covariates are in $\boldsymbol{x}, \boldsymbol{\theta} \in \mathbb{R}^p$. For $p \geq d$, the covariate covariance $\boldsymbol{\Sigma}$ has a block structure, with the directions corresponding to the d dimensional signal space being $O({}^p\!/\!{}_d)$, while the remaining $p - d$ directions being $O(1)$. Hence, when the system is overparametrised $p \gg d$, the covariance has a small but strong signal block, with a weak but large orthogonal block.

General conditions for benign overfitting in the context of ridge regression were first studied by Bartlett et al. (2020); Tsigler and Bartlett (2023). Here, we will closely follow an argument from Misiakiewicz and Montanari (2024) which is based on the formulas from proposition 1.

[4] Note this is normalised such that $\operatorname{Tr}\boldsymbol{\Sigma} = \Theta(p)$

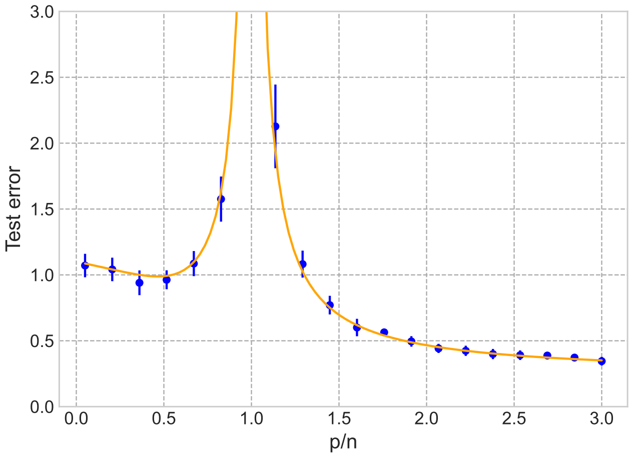

Figure 4: Test error of the random features ridge regressor eq. (4.2) as function of $\gamma = {}^{p}\!/\!_{n}$ at fixed ${}^{n}\!/\!_{d} = 1.5$ and $\lambda = 0^{+}$. The solid line denote the theoretical result obtained from proposition 1 under the Gaussian equivalent covariance eq. (4.9), and points denote finite-size simulations with $d = 500$.

⚠ Technically, the discussion that follows requires a non-asymptotic control of the risk, which goes beyond our proportional asymptotic result proposition 1. Nevertheless, as previously discussed, one can derive non-asymptotic multiplicative rates for the deterministic equivalents in theorem 1. We refer the interested reader to (Cheng and Montanari, 2024; Misiakiewicz and Saeed, 2024; Defilippis et al., 2024).

From eq. (3.20) and the self-consistent equation for $\kappa_\star$, we know that for any $\lambda \geq 0$:

$$\operatorname{Tr} \boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \kappa_\star \boldsymbol{I}_d)^{-2} \leq \operatorname{Tr} \boldsymbol{\Sigma} (\boldsymbol{\Sigma} + \kappa_\star \boldsymbol{I}_p)^{-1} = n \left(1 - \frac{\lambda}{\kappa_\star(\lambda)} \right) \leq n \tag{4.12}$$

We now assume that actually we have a tighter control of $\mathrm{d}f_2(\kappa_\star)$:

$$\operatorname{Tr} \boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \kappa_\star \boldsymbol{I}_d)^{-2} \leq n \left(1 - \frac{1}{c_\star} \right) \tag{4.13}$$

for a constant $c_\star \in (1, \infty)$ which is problem dependent. This implies an immediate upper-bound on the bias and variance:

$$\begin{aligned} \mathcal{B}(\boldsymbol{\theta}_\star, \boldsymbol{\Sigma}, \lambda, \gamma) &= \frac{\kappa(\lambda)^2 \langle \boldsymbol{\theta}_\star, \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma} + \kappa_\star \boldsymbol{I}_d \right)^{-2} \boldsymbol{\theta}_\star \rangle}{1 - \frac{1}{n} \operatorname{Tr} \{ \boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \kappa_\star \boldsymbol{I}_d)^{-2} \}} \leq c_\star \kappa_\star^2 \langle \boldsymbol{\theta}_\star, \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma} + \kappa_\star \boldsymbol{I}_d \right)^{-2} \boldsymbol{\theta}_\star \rangle \\ \mathcal{V}(\boldsymbol{\Sigma}, \lambda, \sigma^2, \gamma) &= \sigma^2 \frac{\operatorname{Tr} \left\{ \boldsymbol{\Sigma}^2 \left(\boldsymbol{\Sigma} + \kappa_\star \boldsymbol{I}_d \right)^{-2} \right\}}{n - \operatorname{Tr} \{ \boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \kappa_\star \boldsymbol{I}_d)^{-2} \}} \leq \frac{c_\star \sigma^2}{n} \operatorname{Tr} \boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \kappa_\star \boldsymbol{I}_d)^{-2} \end{aligned} \tag{4.14}$$

reducing the problem to the study of $\d{\mathrm{f}}_2(\kappa_\star)$ and the quadratic form in the bias. The key idea to control these terms is to split the target and the covariance into a low- and a high-frequency part:

$$\boldsymbol{\Sigma} = \sum_{\ell=1}^{k_\star} \lambda_\ell \boldsymbol{v}_\ell \boldsymbol{v}_\ell^\top + \sum_{\ell=k_\star+1}^{d} \lambda_\ell \boldsymbol{v}_\ell \boldsymbol{v}_\ell^\top := \boldsymbol{\Sigma}_{\leq k_\star} + \boldsymbol{\Sigma}_{> k_\star} \tag{4.15}$$

$$\boldsymbol{\theta}_\star = \sum_{\ell=1}^{k_\star} \langle \boldsymbol{\theta}_\star, \boldsymbol{v}_\ell \rangle \boldsymbol{v}_\ell + \sum_{\ell=k_\star+1}^{d} \langle \boldsymbol{\theta}_\star, \boldsymbol{v}_\ell \rangle \boldsymbol{v}_\ell := \boldsymbol{\theta}_{\star, \leq k_\star} + \boldsymbol{\theta}_{\star, > k_\star} \tag{4.16}$$

where, motivated by the discussion in section 3.1 we take the cut-off to be:

$$k_\star = \max\{k : \lambda_k \geq \kappa_\star\} \tag{4.17}$$

Using this decomposition allow us to further upper-bound the bias and variance. Starting with the variance, we have:

$$\begin{aligned}
\operatorname{Tr} \boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\kappa_\star \boldsymbol{I}_d)^{-2} &= \sum_{\ell=1}^{k_\star} \frac{\lambda_\ell^2}{(\lambda_\ell+\kappa_\star)^2} + \sum_{\ell=k_\star+1}^{d} \frac{\lambda_\ell^2}{(\lambda_\ell+\kappa_\star)^2} \\
&\leq \sum_{\ell=1}^{k_\star} \frac{\lambda_\ell^2}{\lambda_\ell^2} + \sum_{\ell=k_\star+1}^{d} \frac{\lambda_\ell^2}{\kappa_\star^2} \qquad (4.18) \\
&\leq k_\star + \sum_{\ell=k_\star+1}^{d} \frac{\lambda_\ell^2}{\lambda_{k_\star+1}^2} \qquad (4.19)
\end{aligned}$$

where in the last inequality we have used that by construction $\lambda_{k+1} < \kappa_\star$ to upper-bound the second term. While this is a perfectly good bound, to bring it closer to the result derived by Bartlett et al. (2020), we can use the self-consistent equations to further rewrite the second term. Indeed, we can bound:

$$\begin{aligned}
n \geq \operatorname{Tr} \boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\kappa_\star \boldsymbol{I}_d)^{-1} &= \sum_{\ell=1}^{k_\star} \frac{\lambda_\ell}{\lambda_\ell+\kappa_\star} + \sum_{\ell=k_\star+1}^{d} \frac{\lambda_\ell}{\lambda_\ell+\kappa_\star} \\
&\geq \sum_{\ell=1}^{k_\star} \frac{\lambda_\ell}{2\lambda_\ell} + \sum_{\ell=k_\star+1}^{d} \frac{\lambda_\ell}{2\kappa_\star} \\
&= \frac{k_\star}{2} + \sum_{\ell=k_\star+1}^{d} \frac{\lambda_\ell}{2\lambda_{k_\star+1}} \qquad (4.20)
\end{aligned}$$

In particular, since $k_\star \geq 1$ we have $n \geq \sum_{\ell>k_\star} {}^{\lambda_\ell}/_{2\lambda_{k_\star+1}}$ and so the variance term can be bounded by:

$$\begin{aligned}
\mathcal{V} &\leq \frac{c_\star \sigma^2}{n}\left[k_\star + \sum_{\ell=k_\star+1}^{d} \frac{\lambda_\ell^2}{\lambda_{k_\star+1}^2}\right] \\
&\leq c_\star \sigma^2 \left[\frac{k_\star}{n} + \frac{\sigma_{k_\star^2}}{\sigma_{k_\star+1}^2} \frac{\left(\sum_{\ell>k_\star} \lambda_\ell\right)^2}{\sum_{\ell>k_\star} \lambda_\ell^2} n\right] \qquad (4.21)
\end{aligned}$$

The ratio:

$$r(k) = \frac{\left(\sum_{\ell>k} \lambda_\ell\right)^2}{\sum_{\ell>k} \lambda_\ell^2} \qquad (4.22)$$

is a classical quantity is Physics and Random Matrix Theory, where $r(1)$ is *spectrum participation ratio*. It measures how "localised" is the spectrum of a matrix. Indeed, if all eigenvalues are equal, $r(1) = 1$, while if only a few eigenvalues dominate the spectrum, we have $r(1) \ll 1$. Here, we define it with respect to the tail of the spectrum.

Similarly, we can bound the bias:

$$\begin{aligned}
\mathcal{B} &\leq c_\star \sum_{\ell=1}^{d} \frac{\kappa_\star^2 \lambda_\ell}{(\lambda_\ell+\kappa_\star)^2} \langle \boldsymbol{\theta}_\star, \boldsymbol{v}_\ell \rangle^2 \qquad (4.23) \\
&\leq c_\star \left[\sum_{\ell=1}^{k_\star} \frac{\kappa_\star^2}{\lambda_k} \langle \boldsymbol{\theta}_\star, \boldsymbol{v}_\ell \rangle^2 + \sum_{\ell=k_\star+1}^{d} \lambda_\ell \langle \boldsymbol{\theta}_\star, \boldsymbol{v}_\ell \rangle^2\right] \qquad (4.24) \\
&\leq c_\star \left[\lambda_{k_\star}^2 ||\boldsymbol{\theta}_{\star,\leq k_\star}||_{\boldsymbol{\Sigma}^{-1}}^2 + ||\boldsymbol{\theta}_{\star,>k_\star}||_{\boldsymbol{\Sigma}}^2\right] \qquad (4.25)
\end{aligned}$$

Summarising, we have the following upper-bounds:

$$\mathcal{V} \leq c_\star \sigma^2 \left[\frac{k_\star}{n} + \frac{\sigma_{k_\star^2}}{\sigma^2_{k_\star+1}} \frac{\left(\sum_{\ell > k_\star} \lambda_\ell\right)^2}{\sum_{\ell > k_\star} \lambda_\ell^2} n \right] \tag{4.26}$$

$$\mathcal{B} \leq c_\star \left[\lambda_{k_\star}^2 ||\boldsymbol{\theta}_{\star, \leq k_\star}||^2_{\boldsymbol{\Sigma}^{-1}} + ||\boldsymbol{\theta}_{\star, > k_\star}||^2_{\boldsymbol{\Sigma}} \right] \tag{4.27}$$

$$\frac{k_\star}{2} + \sum_{\ell = k_\star + 1} \frac{\lambda_\ell}{2\lambda_{k_\star+1}} \leq n \tag{4.28}$$

with $c_\star \in (1, \infty)$. Note this is valid for all $\lambda \geq 0$. We can also obtain a lower-bound for $k_\star$ by assuming $\lambda \leq {}^{\kappa_\star}/_2$, which can always be satisfied by taking λ small enough, since $\kappa_\star(\lambda) \geq \lambda \geq 0$. In this case:

$$\operatorname{Tr} \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \kappa_\star \boldsymbol{I}_d)^{-1} = n \left(1 - \frac{\lambda}{\kappa_\star} \right) \geq n \tag{4.29}$$

Hence:

$$\begin{aligned} n \leq \operatorname{Tr} \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \kappa_\star \boldsymbol{I}_d)^{-1} &= \sum_{\ell=1}^{k_\star} \frac{\lambda_\ell}{\lambda_\ell + \kappa_\star} + \sum_{\ell = k_\star + 1}^{d} \frac{\lambda_\ell}{\lambda_\ell + \kappa_\star} \\ &\leq \sum_{\ell=1}^{k_\star} \frac{2\lambda_\ell}{\lambda_\ell} + \sum_{\ell = k_\star + 1}^{d} \frac{2\lambda_\ell}{\kappa_\star} \\ &= 2k_\star + 2 \sum_{\ell = k_\star + 1} \frac{\lambda_\ell}{\lambda_{k_\star + 1}} \end{aligned} \tag{4.30}$$

This lower-bound implies that $k_\star \to \infty$ as $n \to \infty$. Together, this provides everything we need to characterise when overfitting is benign.

Assuming that $\lambda_k \to 0$ as $k \to \infty$, we have that:

$$\mathcal{B} \to 0 \qquad \text{as } n \to \infty \tag{4.31}$$

provided that $||\boldsymbol{\theta}_\star||_{\boldsymbol{\Sigma}^{-1}} < \infty$. A sufficient condition is that $\boldsymbol{\theta}_\star$ only with finitely many directions of the covariance $\boldsymbol{\Sigma}$. What about the variance? Assuming that ${}^{\lambda_{k_\star}}/_{\lambda_{k_\star+1}} < \infty$ as $k_\star \to \infty$, in order for the variance to vanish we need that:

$$\frac{k_\star}{n} \to 0, \qquad \frac{n}{r(k_\star)} \to 0. \tag{4.32}$$

References

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

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

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR, 2020.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

Chen Cheng and Andrea Montanari. Dimension free ridge regression. *The Annals of Statistics*, 52(6): 2879–2912, 2024.

Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *Advances in Neural Information Processing Systems*, 34:10131–10143, 2021.

Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic equivalents and scaling laws for random feature regression. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 104630–104693. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/bd18189308a4c45c7d71ca83acf3deaa-Paper-Conference.pdf.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pages 3452–3462. PMLR, 2020.

Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10 (4):041044, 2020.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pages 426–471. PMLR, 2022.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

Hong Hu and Yue M. Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory, in press*, 2022.

Anders Krogh and John Hertz. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, 1991.

Marco Loog, Tom Viering, Alexander Mey, Jesse H Krijthe, and David MJ Tax. A brief prehistory of double descent. *Proceedings of the National Academy of Sciences*, 117(20):10625–10626, 2020.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75 (4):667–766, 2022.

Theodor Misiakiewicz and Andrea Montanari. Six lectures on linearized neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(10):104006, 2024.

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.

Andrea Montanari and Basil N Saeed. Universality of empirical risk minimization. In *Conference on Learning Theory*, pages 4310–4312. PMLR, 2022.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Manfred Opper, W Kinzel, J Kleinz, and R Nehl. On the ability of the optimal perceptron to generalise. *Journal of Physics A: Mathematical and General*, 23(11):L581, 1990.

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, oct 2019. doi: 10.1088/1751-8121/ab4c8b. URL https://dx.doi.org/10.1088/1751-8121/ab4c8b.

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023.

A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2008. ISBN 9780387790527. URL https://books.google.fr/books?id=mwB8rUBsbqoC.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Tong Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural computation*, 17(9):2077–2098, 2005.