Les Houches Summer school on Statistical Physics & Machine learning, July 2022

Five Lectures on Nonparametric Neural Networks

September 23, 2023

Lecturer: Andrea Montanari *Scribe: Simone Maria Giancola*

In short: a collection of Lectures from a Summer School held in France in 2022. See [MM23] for a detailed review by the lecturers.

Foreword

The following is a redaction of lectures held at the *Les Houches Summer School on Statistical Physics and Machine Learning*, in July 2022. The page of the event is here. The videos are at this link.
While I was writing them up, the instructors published an expanded version on ArXiv [MM23]. Their document is clearly of higher quality (they are also authors of many papers explored here and there). There, a full set of references in an independent section is also present.
To give a brief comment on the differences, their work is thematic, and merges different advanced sets of lectures they taught. The focus for each chapter is on topics that build up sequentially. Differently, this work is smaller and is merely a write up of their lectures from someone that saw them on YouTube. The idea is not to compare them in terms of quality since there is a clear mismatch. Despite this, maybe reading twice about the overlapping matters might be helpful for inexperienced readers like myself. This last fact motivates making the work available.

For similar reasons, this document might be vague or incorrect in some passages. Hopefully, it will be fixed with time. In particular, I am quite not satisfied with not having found some references. As my reading list gets exhausted, these will be sorted out. The main idea behind this write up is exactly getting an introduction for future readings.

The results are very recent and nice. Open problems stated by the lecturer (Prof. Montanari) were mentioned in between explanations. Some might be solved as of now.
For any remark, typo or suggestion, I am more than happy to chat[1].

Versioning This is a first version, I expect to correct mistakes and adjust it with time. I stress that there might be errors.

How to use this? I would suggest reading [MM23] to see how the lecturers envisioned the results, and if needed take this as an accompanying version.

[1] my email is `simonegiancola 09 at gmail dot com`.

Contents

Notation By $i \in [n]$ we mean $i = 1, \dots, n$. Emphasis on random variables is placed with capital letters, e.g. X. A vector is bold italic, a matrix is bold and capital. So, $\boldsymbol{x}$ is a vector, $\boldsymbol{X}$ is a random vector and $\mathbf{X}$ is a matrix. The rest is standard and understandable from context. Various objects are introduced during the arguments but are well specified.

1 Setting and Phenomenology

The setting we will consider is the classical one. Assume we are given some data $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ where $\forall i$ the pair $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ was sampled iid from a distribution $\mathbb{P}$. The task is finding a function $f : \mathbb{R}^d \to \mathbb{R}$ that achieves minimum square error. We aim to minimize:

$$\mathcal{R}(f) \equiv \mathbb{E}\left[(Y - f(\boldsymbol{X}))^2\right] \quad (\boldsymbol{X}, Y) \sim \mathbb{P} \in \mathscr{P}(\mathbb{R}^d \times \mathbb{R}),$$

which is the square loss. We will exchangeably use the term test error and risk. If $\mathbb{P}$ was available, the optimal choice would be the conditional expectation, which under mild conditions achieves the minimum test error:

$$f_*(\boldsymbol{x}) = \mathbb{E}\left[Y|\boldsymbol{X}\right] \quad \mathcal{R}(f_*) \leq \mathcal{R}(f) \quad \forall f.$$

Clearly this is somewhat a *loose* definition of a learning task. Especially, we are not specifying the sense of optimality[2] and the expectation is not available since we cannot really sample from the

[2]For all f in which space?

distribution $\mathbb{P}$ which is unknown. To overcome this issue, we will use the so called *Empirical Risk Minimization* (ERM) principle. Mathematically, for a specified loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ and space of functions $\mathcal{F}$, we hope to achieve a good performance on the optimization:

$$f_* = \arg\min_{f \in \mathcal{F}} \mathcal{R}(f) \quad \mathcal{R}(f) = \mathbb{E}\left[\ell(f(\boldsymbol{X}, Y)\right]$$

by optimizing over the *empirical risk*

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \widehat{\mathcal{R}}_n(f) \quad \widehat{\mathcal{R}}_n(f) \equiv \frac{1}{n} \sum_{i=1}^{n} \ell(f(\boldsymbol{X}_i), Y_i)$$

Remark 1.1. *Optimizing over $\mathcal{F}$ is reasonable. It provides the option to choose which space it is, and avoids the trivial case in which the empirical risk is minimized by a function that replicates the data. In some sense it reflects the power of estimation that is available.*

Remark 1.2. *Any assumption on $\mathbb{P}$ is in practice not appropriate. Yet, it helps to build an estimator $\widehat{f}$. To obtain theoretical results it is reasonable to allow for assumptions on $\mathbb{P}$.*

We focus on results about *parameter* optimization. This means that $\mathcal{F} = \{f(\cdot; \boldsymbol{\theta}) | \boldsymbol{\theta} \in \mathbb{R}^p\}$. In particular we are interested in two regimes:

- *overparametrized*, when $p \gg n$
- *non-parametric*, when $p = \infty$.

In many cases, we will implicitly assume that in an overparametrized model the train loss is zero.

Why are we interested in these parameter settings?

Neural Networks are parametric models, often overparametrized. Many older-fashioned models are of this type as well.

{ex:cub

Example 1.3 (Cubic Splines)**.** *For simplicity let $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$. Seek an optimization of the form:*

$$\widehat{f} = \arg\min \left\{ \widehat{\mathcal{R}}_n(f) + \lambda \int f''(x)\, dx \right\} \iff \widehat{f} = \arg\min \widehat{\mathcal{R}}_n(f) \quad s.t. \quad \int f''(x)\, dx \leq \rho.$$

Consider the case in which $\lambda = 0$, or equivalently $\rho \uparrow \rho^*$ such that $\widehat{\mathcal{R}}_n(f_*) = 0$. This is a perfect interpolator of the data.
If the data generating process was a linear model $y_i = \beta_* x_i + \epsilon_i$, this is not great idea. A visualization is shown in Figure 1. The result is a third order polynomial with continuous derivatives that *overfits* the model. Any prediction will be mistaken by σ units where σ is the variance of the noise ϵ (assuming the classical symmetric noise model, such as Gaussian). Indeed, regularization

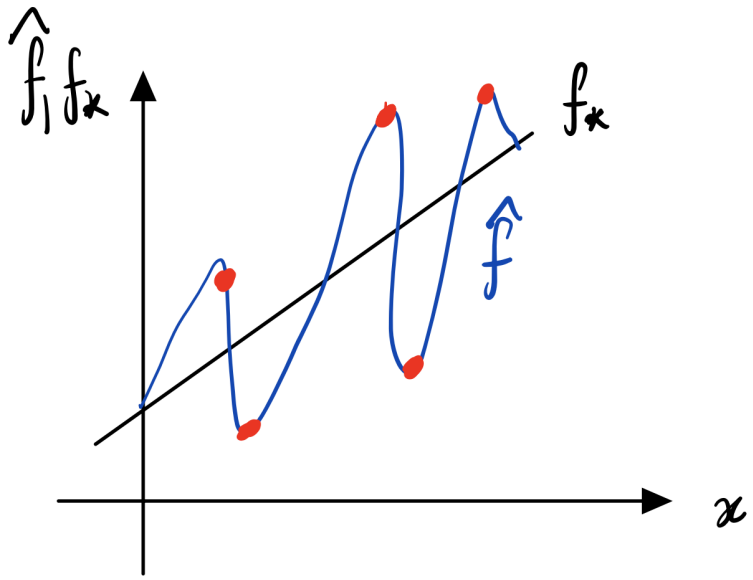

Figure 1: Spline interpolator {fig:sp

is useful in this case, and $\lambda \neq 0$ makes fitting more flexible. Recently, a suboptimality proof was given in [RZ18].

Despite this, we informally present our motivation to study perfect interpolators.

Phenomenology of Perfect Interpolators

Overparametrized models with no regularization achieve a test error that is *much greater* than the train error but *still satisfactory*.

In practical applications this is highly beneficial. At the cost of losing theoretical guarantees, applications gain in tractability, as non-regularized models are friendly. The cost of this is over-parametrizing sufficiently the model, and current technology allows for this smoothly.
The main interest is having a model specification that is well suited for optimization with Gradient Descent (GD) or Stochastic Gradient Descent (SGD). In particular, the latter converges quickly with overparametrization.

Mathematically, the simplest justification is as follows. In the parameter space $\mathbb{R}^p$ there is a manifold of perfect interpolators, expressed concisely as

$$\mathscr{M}_{ERM_0} \equiv \{\boldsymbol{\theta} : f(\boldsymbol{x}_i; \boldsymbol{\theta}) = y_i \ \forall i \in [n]\}$$

which is very *big*. Here by big, we roughly mean that for any $\boldsymbol{\theta}_0$ of initialization, there is a close point $\boldsymbol{\theta}_* \in \mathscr{M}$. A depiction in $\mathbb{R}^2$ is presented in Figure 2.

A formal justification validating this idea is available in a class of examples. We refer to these as the *linear/lazy/neural tangent* regime. In this precise group, the validation is assessed by proving

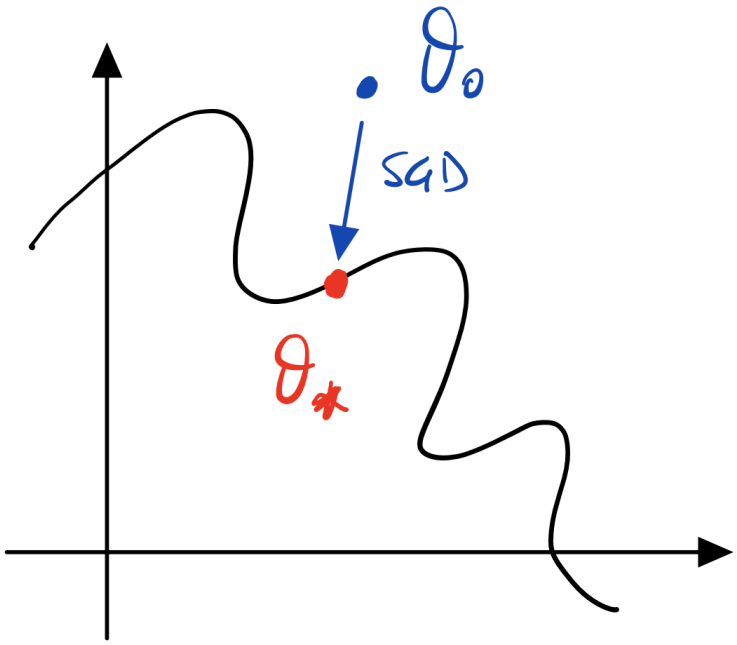

Figure 2: Convergence to the manifold $\mathscr{M}_{ERM_0}$ {fig:ma

that:

$$\mathcal{R}(f) \approx \mathcal{R}_*(f),$$

i.e. that the test error is approximately the optimal test error. The peculiar aspect is that in low-dimensional models such as the cubic spline this does not work, while in high-dimensions it does. To give more context, when referring to the optimal test error we will exchangeably mean the original object and the *excess test error*. The two differ by a constant. In mathematical terms, the risk-excess risk relationship for a noisy model with square loss is:

$$\mathcal{R}(f) = \mathbb{E}_{\boldsymbol{x}_{new}}\left[\mathbb{E}_{\boldsymbol{\epsilon}}\left[(f(\boldsymbol{x}_{new}) - y_{new})^2\right]\right] = \mathbb{E}_{\boldsymbol{x}_{new}}\left[\mathbb{E}_{\boldsymbol{\epsilon}}\left[(f(\boldsymbol{x}_{new}) - \boldsymbol{\beta}_*^\top \boldsymbol{x}_{new} - \epsilon_i)^2\right]\right] = \mathcal{R}_{excess}(f) + \sigma^2.$$

Previously, in Example 1.3, the test error was σ for any sample size n.

1.1 Linear and Lazy Regime

We aim to answer the possible question.

Closeness

Under which conditions is an initialization $\boldsymbol{\theta}_0$ close to some $\boldsymbol{\theta} \in \mathscr{M}_{ERM_0}$?

Introduce the compact notation

$$f_n(\boldsymbol{\theta}) = \begin{bmatrix} f(\boldsymbol{x}_1;\boldsymbol{\theta}) \\ \vdots \\ f(\boldsymbol{x}_n;\boldsymbol{\theta}) \end{bmatrix},$$

the question is rephrased as solving $\boldsymbol{y} = f_n(\boldsymbol{\theta})$ for $\boldsymbol{y} = [y_1, \dots, y_n]^\top$. This is an overdetermined set of linear equations. For $\boldsymbol{\theta} \approx \boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_t \equiv \boldsymbol{\theta}_0 + t(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$, a Taylor expansion gives:

$$\begin{aligned} \widetilde{\boldsymbol{y}} \equiv \boldsymbol{y} - f_n(\boldsymbol{\theta}_0) &= f_n(\boldsymbol{\theta}) - f_n(\boldsymbol{\theta}_0) \\ &\approx D\, f_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \int_0^1 \underbrace{(D\, f_n(\boldsymbol{\theta}_t) - D\, f_n(\boldsymbol{\theta}_0))(\boldsymbol{\theta} - \boldsymbol{\theta}_0)}_{=\frac{df_n(\boldsymbol{\theta}_t)}{dt}}\, dt \\ &= \boldsymbol{\Phi}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \mathcal{E}(\boldsymbol{\theta}) \end{aligned}$$

where $\boldsymbol{\Phi} \in \mathbb{R}^{n \times p}$ is the Jacobian matrix.

Remark 1.4. *The Taylor expansion has this form:*

$$g(x) = g(0) + g'(0)x + \int_0^x g'(u) - g'(0)\, du,$$

and if the function was on more than one dimension the last term would have been $\int_0^1 (\partial_t g(t \cdot \boldsymbol{x}) - \partial_t g(\mathbf{0})) \cdot \boldsymbol{x}\, dt$.

The aim is bounding $\mathcal{E}(\boldsymbol{\theta})$. To do so, we introduce a Lipschitz condition.

{ass:li

Assumption 1 (Lipschitz Condition)**.** *For a function* f_n*, it holds:*

$$L_n \equiv \sup_{\boldsymbol{\theta} \in \mathbb{R}^p} \frac{\|D\, f_n(\boldsymbol{\theta}) - D\, f_n(\boldsymbol{\theta}_0)\|_p}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2} < \infty.$$

which in particular implies that $\|\mathcal{E}(\boldsymbol{\theta})\|_2 \leq L_n \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 < \infty$.

Coming back to our problem, the solution of the equation $\widehat{\boldsymbol{y}} = \boldsymbol{\Phi}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \mathcal{E}(\boldsymbol{\theta})$ is given by the inversion formula:

$$\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \boldsymbol{\Phi}^\dagger \widehat{\boldsymbol{y}} - \underbrace{\boldsymbol{\Phi}^\dagger \mathcal{E}(\boldsymbol{\theta})}_{\equiv \delta}$$

where to invert $\boldsymbol{\Phi}$ we have used the pseudoinverse, which also ensures that $\boldsymbol{\theta}$ will be the smallest possible L^2 norm vector to $\boldsymbol{\theta}_0$. The nonlinearity in the correction error term is solved by a Fixed point equation of the map:

$$\delta = \boldsymbol{\Phi}^\dagger \mathcal{E}(\boldsymbol{\theta}_0 + \boldsymbol{\Phi}^\dagger \widehat{\boldsymbol{y}} - \delta) \equiv F(\delta).$$

If F maps balls into smaller balls, namely $F(\mathcal{B}^p(0,r)) = \mathcal{B}^p(0,r')$ with $r' < r$, then it has a fixed point with $\|\delta\|_2 < r'$. To derive it, it suffices to bound the norm:

$$r' = r'(r) = \|F(\delta)\| \leq \left\|\boldsymbol{\Phi}^\dagger\right\|_{op} L_n \left\|\boldsymbol{\Phi}^\dagger \widehat{\boldsymbol{y}} - \delta\right\|_2^2 \leq \left\|\boldsymbol{\Phi}^\dagger\right\|_{op} L_n \left(\left\|\boldsymbol{\Phi}^\dagger \widehat{\boldsymbol{y}}\right\|_2 + r\right)^2$$

where in the last passage we have used the triangular inequality. Notice also that the first norm is just the largest singular value of the matrix. We have now rephrased the problem into showing that $r'(r) < r$, where $r'(r) = a(b+r)^2$ for some $r > 0$. This is easily solved by the condition

$$L_n \left\|\boldsymbol{\Phi}^\dagger\right\|_{op} \left\|\boldsymbol{\Phi}^\dagger \widetilde{\boldsymbol{y}}\right\|_{op} \leq \frac{1}{4},$$

by which we state a Proposition just below.

{prop:c

Proposition 1.5 ([COB20])**.** *We have*

$$L_n \left\|\boldsymbol{\Phi}^\dagger\right\|_{op}^2 \|\widetilde{\boldsymbol{y}}\|_2 \leq \frac{1}{4} \iff L_n \|\widetilde{\boldsymbol{y}}\|_2 \leq \frac{(\sigma_{min}(\boldsymbol{\Phi}))^2}{4}, \tag{1}$$

{eqn:su

since the largest singular value of $\boldsymbol{\Phi}^\dagger$ *is* $\frac{1}{\sigma_{min}(\boldsymbol{\Phi})}$.

In words, we want the non-linearity of the model L_n to be small, and the Jacobian $\boldsymbol{\Phi}$ to be non-singular. If the latter case did not verify, then a linear approximation would not have been reasonable from the start.

It turns out that one can say something more in this setting. It is indeed possible to bound $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2$ and it makes sense that GD converges quickly.

{thm:du

Theorem 1.6 ([ALS19; Du+19; Zou+18; OS19b; LZB21] in the formulation of [OS19a] and [COB20])**.** *Assume that* $\lambda_0 \equiv \frac{(\sigma_{min}(\boldsymbol{\Phi}))^2}{4}$ *and that Eqn. 1 holds. Then for a gradient flow* $\boldsymbol{\theta}_t$*:*

1. *the training error converges exponentially to zero, i.e.* $\boldsymbol{\theta}_t$ *goes exponentially fast into* $\mathscr{M}_{ERM_0}$*:*

$$\widehat{\mathcal{R}}_n(\boldsymbol{\theta}_t) \leq e^{-\lambda_0 t} \widehat{\mathcal{R}}_n(\boldsymbol{\theta}_0).$$

2. *(sloppy) the generalization error is approximately that of a linear model:*

$$\mathcal{R}(f(\cdot;\boldsymbol{\theta}_t)) = \mathcal{R}(f_{lin}(\cdot;\boldsymbol{\theta}_{lin}^{(t)})) + err$$

for err *small. A formal statement is found in this review [BMR21, Thm. 5.1].*

The morale of Thm. 1.6 is that one can compute the test error of a much simpler linear model, and that there exists a ball $\mathcal{B}^p(r)$ around the initialization such that some $\boldsymbol{\theta}_* \in \mathcal{B}^p(r)$ is also in $\mathscr{M}_{ERM_0}$. An idealistic visualization of this phenomenon is Figure 3.

Remark 1.7. *Notice that nothing explicitly requires that* $\boldsymbol{\theta}_*$ *will be the closest point in norm to the start.*

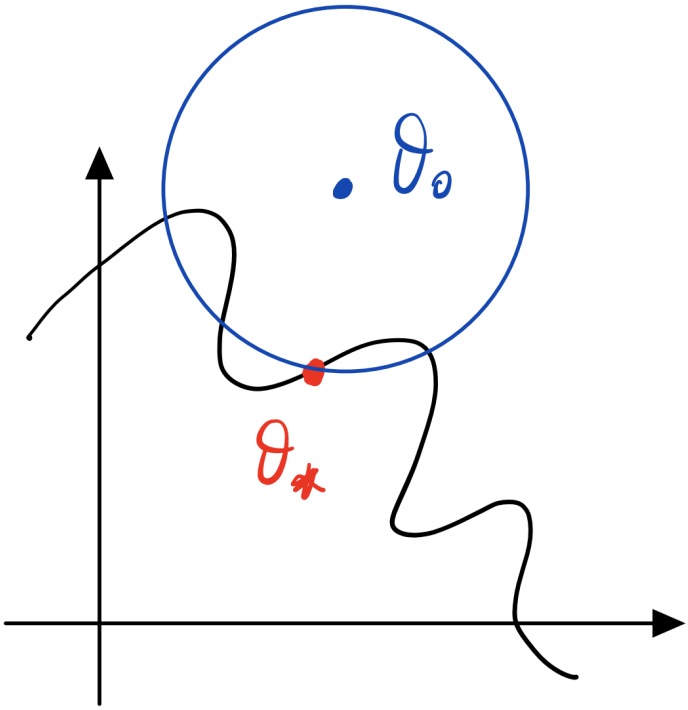

Figure 3: Closeness of $\mathscr{M}_{ERM_0}$ to starting point {fig:ma

We will try to give more details of the statement in the next paragraphs. First of all, the idea is that in this linearized (also [COB20]) regime the model will approximate the ERM procedure with a quadratic function, i.e. gradient flow on a quadratic function. This is guaranteed to converge. In some sense, the induction bias is L^2 in the coefficients.
The particular form of the linear model is:

$$f_{lin}(\boldsymbol{x};\boldsymbol{\theta}) = f(\boldsymbol{x};\boldsymbol{\theta}_0) + \langle \boldsymbol{\theta} - \boldsymbol{\theta}_0, \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x};\boldsymbol{\theta}_0)\rangle$$

which is the first order Taylor expansion in $\boldsymbol{\theta}$. The Gradient Flow (GF), is with respect to the risk for this model. Namely:

$$\widehat{\mathcal{R}}_n(f_{lin}(\boldsymbol{x};\boldsymbol{\theta})) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i;\boldsymbol{\theta}))^2 = \frac{1}{n}\left\|\widetilde{\boldsymbol{y}} - D\, f_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right\|_2^2$$

which indeed is quadratic in $\boldsymbol{\theta}$! Our parameter $\boldsymbol{\theta}_{lin}^{(t)}$ will run GF on this risk. To draw precise conclusions in generalization terms we now turn to studying a specific model. We will consider a 2-layers Neural Network, which has function:

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} a_i \cdot \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle) \quad \sigma : \mathbb{R} \to \mathbb{R}$$

where σ is an element-wise nonlinearity. Further, we set the weights of the last layer to be symmetric, namely $a_1 = \cdots = a_{\frac{N}{2}} = 1$ and $a_{\frac{N}{2}+1} = \cdots = a_N = -1$, and aim to only fit the weights $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_N$. Thus we set $\boldsymbol{\theta} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_N)$, and $\boldsymbol{\theta} \in \mathbb{R}^p, p = Nd$. The initialization of interest is spherical, i.e. $\boldsymbol{w}_{i,t=0} \sim \mathcal{U}nif(\mathbb{S}^{p-1})$.
For simplicity, we neglect the term $f_n(\boldsymbol{\theta}_0)$ which is just random and independent of data. This choice is without loss of generality, and amounts to replacing labels with values $\boldsymbol{y}' = \boldsymbol{y} - f(\mathbf{X};\boldsymbol{\theta}_0)$, or setting directly $f_n(\cdot;\boldsymbol{\theta}_0) = 0$ with a choice of weights $\boldsymbol{w}_{i,t=0}$ iid and $\boldsymbol{w}_{i+\frac{N}{2},t=0} = -\boldsymbol{w}_{i,t=0}$ for $i \in [\frac{N}{2}]$. This last technique is often named *symmetric initialization*.

Since we want to understand the linear regime generalization of this model, it is crucial to observe the gradient at initialization. In general, the gradient has form:

$$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{\sqrt{N}}\left(a_1\sigma'(\langle \boldsymbol{w}_1, \boldsymbol{x}\rangle)\boldsymbol{x}^\top, \ldots, a_n\sigma'(\langle \boldsymbol{w}_N, \boldsymbol{x}\rangle)\boldsymbol{x}^\top\right)$$

and in a linearized model, using the shortcut $\boldsymbol{b} \equiv \boldsymbol{\theta} - \boldsymbol{\theta}_0$, the function of Theorem 1.6 is

$$f_{lin}(\boldsymbol{x};\boldsymbol{b}) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} a_i \,\langle \boldsymbol{b}_i, \boldsymbol{x}\rangle\, \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle),$$

to which we add a $\frac{1}{\sqrt{d}}$ factor and absorb[3] the a_i into $\boldsymbol{b}$. Eventually, we work out a more compact expression that is:

$$f_{lin}(\boldsymbol{x};\boldsymbol{b}) = \langle \boldsymbol{b}, \varphi(\boldsymbol{x})\rangle \quad \varphi(\boldsymbol{x}) = \frac{1}{\sqrt{Nd}}\left(\sigma'(\langle \boldsymbol{w}_1, \boldsymbol{x}\rangle)\boldsymbol{x}^\top, \ldots\right),$$

[3]Notice that $a_i \in \{\pm 1\}$

for which we want to evaluate the interpolator

$$\widehat{\boldsymbol{b}} = \arg\min \left\{ \|\boldsymbol{b}\|^2 \mid \langle \boldsymbol{b}, \varphi(\boldsymbol{x}_i) \rangle = y_i \ \forall i \in [n] \right\}.$$

In the general setting, one would ask if Proposition 1.5 holds for this class of problems.

Proposition 1.8 ([OS19a]). *For $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $y_i \in O(1)$ Eqn. 1 holds w.h.p. if $Nd \geq Cn^2$*

Remark 1.9 (On the meaning of Lazy). *In [COB20] it is shown that for a large class of linear models, namely those satisfying a mild homogeneity condition, if the function is strecthed by a constant α, i.e.*

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{\alpha}{\sqrt{N}} \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle$$

then the condition becomes:

$$\alpha L_n \|\widetilde{\boldsymbol{y}}\|_2^2 \leq \frac{(\sigma_{min}(\boldsymbol{\Phi}))^2 \alpha^2}{4}.$$

In other words, despite being invariant under $\boldsymbol{\theta}$ the condition is <u>not</u> invariant under scaling of f, and for $p \gg d, Nd \geq \frac{CN^2}{\alpha}$ with α large enough the linear regime is attained. Another important requirement is that $f(\boldsymbol{x}; \boldsymbol{\theta}_0) = 0$, allowing for the condition $\widetilde{\boldsymbol{y}} = \boldsymbol{y}$, which is somehow equivalent to requiring that the function is not only affine but exactly linear. One could also interpret it as the condition that allows for lazy training, i.e. that $\widetilde{\boldsymbol{y}}$ will not depend on the scaling of α.
Additionally, α does not affect the training error $\widehat{\mathcal{R}}_n$ in the linear regime since it is reabsorbed into $\boldsymbol{b}$.

2 Three Models

We will deal with the following layers of abstraction:

1. a simple Ridge regression with random features
2. kernel regression
3. Neural Tangent Kernel regression

2.1 Linear (Ridge) Regression

As previously discussed, we want to do min-norm regression. In particular, consider:

$$\widehat{\boldsymbol{b}}_\lambda = \arg\min_{\mathbb{R}^p} \left\{ \frac{1}{n} \|\boldsymbol{y} - \mathbf{Z}\boldsymbol{b}\|_2^2 + \lambda \|\boldsymbol{b}\|_2^2 \right\} \quad \mathbf{Z} \in \mathbb{R}^{n \times p} \tag{2}$$ {eqn:ri

and its associated min norm intepolator $\widehat{\boldsymbol{b}} \equiv \widehat{\boldsymbol{b}}_{0+} = \lim_{\lambda \to 0^+} \widehat{\boldsymbol{b}}_\lambda$. Also, we ideally have a data matrix of random features $\boldsymbol{z}_i = \varphi(\boldsymbol{x}_i)$ for all $i \in [n]$.

For a Ridge regression such as the one above, the solution is explicitly found

$$\widehat{\boldsymbol{b}}_\lambda = \frac{1}{n}\mathbf{Z}^\top \left(\lambda \mathbf{I}_n + \frac{1}{n}\mathbf{Z}\mathbf{Z}^\top\right)^{-1} \boldsymbol{y} \tag{3}$$ {eqn:ri

and if we take $y_i = \boldsymbol{b}_*^\top \boldsymbol{z}_i + \epsilon_i$ with $\mathbb{E}[\epsilon_i] = 0$ and $\boldsymbol{\Sigma} \equiv \mathbb{E}\left[\boldsymbol{z}_i \boldsymbol{z}_i^\top\right]$ then the error is classically decomposed into a bias and a variance term as:

$$\mathcal{R}(\mathbf{Z};\lambda) = \left\|f - \widehat{f}\right\|_2^2 = \left\|f - \mathbb{E}\left[\widehat{f}\right]\right\|_2^2 + \left\|\widehat{f} - \mathbb{E}\left[\widehat{f}\right]\right\|_2^2.$$

We briefly sketch the argument for completeness below. The test error reads

$$\mathcal{R}(\mathbf{Z};\lambda) = \mathbb{E}_{\boldsymbol{z}_{new}}\left[\mathbb{E}_{\boldsymbol{\epsilon}}\left[\left(\left\langle \widehat{\boldsymbol{b}}_\lambda, \boldsymbol{z}_{new}\right\rangle - \langle \boldsymbol{b}_*, \boldsymbol{z}_{new}\rangle\right)^2\right]\right]$$

where we added the expectation over the noise in training vector $\boldsymbol{\epsilon}$ to simplify but one could show that it is not necessary, at the cost of a more complicated formula. As before, we inspect the excess risk. Notice also that $\widehat{\boldsymbol{b}}_\lambda$ is a function of $\mathbf{Z}$ and so is $\mathcal{R}$. Namely, both are functions of the training data.
Writing $\|\boldsymbol{v}\|_{\boldsymbol{\Sigma}} \equiv \langle \boldsymbol{v}, \boldsymbol{\Sigma}\boldsymbol{v}\rangle$, we express this as:

$$\begin{aligned}\mathcal{R}(\mathbf{Z};\lambda) = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\left\|\widehat{\boldsymbol{b}}_\lambda - \boldsymbol{b}_*\right\|_{\boldsymbol{\Sigma}}^2\right] &= \left\|\boldsymbol{b}_* - \mathbb{E}_{\boldsymbol{\epsilon}}\left[\widehat{\boldsymbol{b}}\right]\right\|_{\boldsymbol{\Sigma}}^2 + \mathbb{E}_{\boldsymbol{\epsilon}}\left[\left\|\widehat{\boldsymbol{b}}_\lambda - \mathbb{E}_{\boldsymbol{\epsilon}}\left[\widehat{\boldsymbol{b}}_\lambda\right]\right\|_{\boldsymbol{\Sigma}}^2\right] \\ &= \mathsf{Bias}(\mathbf{Z};\lambda) + \mathsf{Variance}(\mathbf{Z};\lambda).\end{aligned}$$

plugging in Eqn. 3 and with a little algebra the bias and variance expression can be found:

$$\begin{aligned}\mathsf{Bias}(\mathbf{Z};\lambda) &= \lambda^2 \langle \boldsymbol{b}_*, \mathbf{S}_\lambda \boldsymbol{\Sigma} \mathbf{S}_\lambda \boldsymbol{b}_*\rangle & \mathbf{S}_\lambda &= (\lambda \mathbf{I}_p + \widehat{\boldsymbol{\Sigma}})^{-1} \\ \mathsf{Variance}(\mathbf{Z};\lambda) &= \frac{\sigma^2}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}\mathbf{S}_\lambda\right) & \widehat{\boldsymbol{\Sigma}} &= \frac{1}{n}\mathbf{Z}\mathbf{Z}^\top = \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i \boldsymbol{z}_i\end{aligned} \tag{4}$$ {eqn:ri

where $\widehat{\boldsymbol{\Sigma}}$ is the empirical covariance.
Given the closed form expression of these random matrices, we wish to compute them with Random Matrix Theory (RMT) tools.

Remark 2.1. *Not only eigenvalues will be important. Since the bias has some vector terms, also eigenvectors will matter.*

Remark 2.2. *Subject to some conditions, we claimed in Theorem 1.6 that there exists a close interpolator in the manifold $\mathscr{M}_{ERM_0}$ and that SGD reaches it, but provided some evidence only for the first statement. The second (i.e. SGD ends up in $\mathscr{M}_{ERM_0}$) is argued starting from establishing an inequality:*

$$\frac{\mathrm{d}\widehat{\mathcal{R}}_n}{\boldsymbol{d}t} \leq -\lambda_0 \widehat{\mathcal{R}}_n,$$

and controlling the LHS by the minimum singular value of the Jacobian to continue.

End of Lecture 1

2.1.1 Sharp Characterization of proportional regime

We have convinced ourselves to look at Ridge regression (Eqns. 2-4). Also, we have derived an expression for the risk in terms of $\boldsymbol{\Sigma}$ norms with a bias-variance decomposition. We will now provide a sharp characterization in the proportional regime by [Has+20].
Without loss of generality[4], assume that the rows of the data matrix are such that $\boldsymbol{z}_i = \boldsymbol{\Sigma}\boldsymbol{u}_i$, independent and identically distributed, with standard moments and a bound on higher moments:

$$\mathbb{E}\left[\boldsymbol{u}_i\right] = 0, \quad \mathbb{E}\left[\boldsymbol{u}_i^2\right] = 1, \quad \mathbb{E}\left[\boldsymbol{u}_i^8\right] \leq C, \quad C \in \mathbb{R}.$$

Additionally, assume the variance matrix satisfies for $\lambda(\boldsymbol{\Sigma})$ an eigenvalue:

$$\lambda_{max}(\boldsymbol{\Sigma}) \leq C, \quad \frac{1}{p}\sum_{i=1}^{p} \lambda_i^{-1}(\boldsymbol{\Sigma}) \leq C, \quad C \in \mathbb{R},$$

where the last condition is verified if the minimum eigenvalue is bounded away from zero.
We state the result in the vanishing regularization regime $\lambda = 0^+$, with proportionality

$$C^{-1} \leq \frac{n}{p} \leq C, \quad C \in \mathbb{R}.$$

Let $\lambda_\star$ be the unique solution to the equation:

$$n = \mathrm{Tr}\left(\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma} + \lambda \mathbf{I}\right)^{-1}\right) \equiv F(\lambda)$$

of which we see a plot in Figure 4. Additionally, define:

$$B_n^{th} := \frac{\lambda_\star^2 \left\langle \boldsymbol{b}_*, (\boldsymbol{\Sigma} + \lambda_\star \mathbf{I})^{-2}\boldsymbol{\Sigma}\boldsymbol{b}_* \right\rangle}{1 - \frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star \mathbf{I})^{-2}\right)} \tag{5}$$ {eqn:Br

$$V_n^{th} := \frac{\sigma^2}{n}\frac{\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star \mathbf{I})^{-1}\right)}{1 - \frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star \mathbf{I})^{-2}\right)}. \tag{6}$$ {eqn:Vr

recall that the bias and variance terms admit expressions $\mathsf{Bias}(\mathbf{Z}; \lambda) = \lambda^2 \left\langle \boldsymbol{b}_*, \mathbf{S}_\lambda \boldsymbol{\Sigma} \mathbf{S}_\lambda \boldsymbol{b}_* \right\rangle$ and $\mathsf{Variance}(\mathbf{Z}; \lambda) = \frac{\sigma^2}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}\mathbf{S}_\lambda\right)$. We claim that B_n^{th}, V_n^{th} will be their predictions. In [Has+20], the following result was shown.

Theorem 2.3 ([Has+20])**.** *Let $\lambda = 0$. With the above assumptions, there exists $\overline{C} = \overline{C}(C)$ such that with high probability:* {thm:Ha

$$|\mathsf{Bias}(\mathbf{Z}; \lambda) - B_n^{th}| \leq \overline{C} n^{-\frac{1}{6}} \quad |\mathsf{Variance}(\mathbf{Z}; \lambda) - V_n^{th}| \leq \overline{C} n^{-\frac{1}{6}}.$$

For $\lambda > 0$, the exponent gives a tighter result.

[4] we can always reorient the matrix

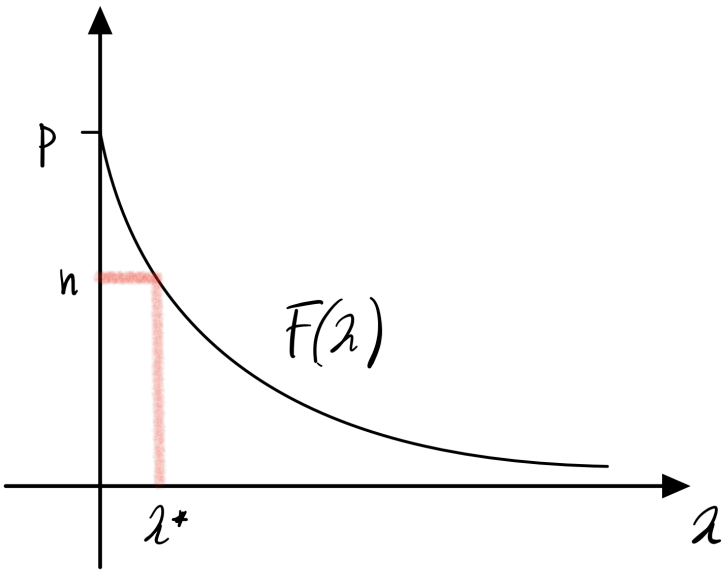

Figure 4: The $F(\lambda)$ function {fig:F

In other words, if there exists a C such that the assumptions holds, then the bias and the variance are well approximated by the theoretical predictions. The proof is a combination of RMT and anisotropic local laws, especially a non-asymptotic one proved in [KY16]. In simple words, such laws are statements about the asymptotics of $\langle \boldsymbol{v}, \mathbf{S}_\lambda \boldsymbol{v} \rangle$. Typically in RMT one studies the trace of the resolvant $\mathbf{S}_\lambda$, which is also the Steltjes transform of the distribution of eigenvalues, but this latter result is more advanced and well suited for the bias, which can be seen as a sandwiching of the a matrix in between $\boldsymbol{b}_*$ vectors.
Rather than spending time on the proof, we focus on providing a nice interpretation.
As a first remark, even though the bias is proportional to λ^2, we send $\lambda \to 0^+$. We will argue that nevertheless $\lambda_\star \neq 0$ when it is the solution of $F(\lambda)$. Indeed, as $\lambda \to 0_+$, it is also the solution of:

$$n\left(1 - \frac{\lambda}{\lambda_\star}\right) = \operatorname{Tr}\left(\mathbf{\Sigma}(\mathbf{\Sigma} + \lambda_\star \mathbf{I})^{-1}\right),$$

where the previous equation was for $\lambda = 0$. Looking at the curve in Figure 4, we can conclude that $\lambda_\star \to 0_+$ but that it is nonzero as $\lambda \to 0_+$, a fact that is guaranteed by the condition $p > n$.

The power of the result lies in the fact that we get rid of the empirical covariances $\widehat{\mathbf{\Sigma}}$, so that the matrices are deterministic. Making $\mathbf{\Sigma}$ diagonal by a change of orientation, both B_n^{th}, V_n^{th} can be written in terms of eigenvectors and thus become sums instead of matrices. Letting dimensions diverge, we eventually get the distribution of the eigenvalues.

Example 2.4. *By the above discussion:*

$$V_n^{th} = \sum_{i=1}^{p} \frac{\sigma_i^2}{\sigma_i + \lambda_\star},$$

so that for $p \to \infty$ it holds that:

$$\frac{1}{p} V_n^{th} \overset{p\to\infty}{\to} \int \frac{\sigma^2}{\sigma + \lambda_\star} \varphi(\mathrm{d}\sigma).$$

The last expression is the asymptotic distribution of eigenvalues, and it admits a nice formula if $\varphi(\mathrm{d}\sigma)$ has simple form.

We now turn to an interpretation of the formulas. Assume the denominator in Eqns. 5, 6 is such that

$$\frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star \mathbf{I})^{-2}\right) \leq \frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star \mathbf{I})^{-1}\right) \leq \frac{1}{c_0},$$

where the first bound is trivial. Notice that for $\lambda_\star = 0$ the term above is exactly 1. Then:

$$B_n^{th} \leq c_0 \lambda_\star^2 \left\langle \boldsymbol{b}_*, (\boldsymbol{\Sigma} + \lambda_\star \mathbf{I})^{-2} \boldsymbol{\Sigma} \boldsymbol{b}_* \right\rangle$$
$$V_n^{th} \leq c_0 \frac{\sigma^2}{n} \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star \mathbf{I})^{-1}\right),$$

which are basically the original bias and variance apart from constants and the difference $\lambda, \lambda_\star, \boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}}$. We also claim that they are the bias and variance for the following sequence model:

$$\boldsymbol{y}_{(s)} = \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{b}_* + \frac{\sigma}{\sqrt{n}} \boldsymbol{w} \quad \widehat{\boldsymbol{b}}_s = \arg\min_{\boldsymbol{b}} \left\{ \left\| \boldsymbol{y}_{(s)}^2 - \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{b} \right\|_2^2 + \lambda_\star \left\| \boldsymbol{b} \right\|_2^2 \right\},$$

which is a simpler Ridge regression with as design matrix the proportional covariance and $\lambda_\star$ regularization. With some manipulations, from a random $\mathbf{Z}$ and null regularization we are obtaining approximate equivalence to a deterministic well suited problem. The term *sequence* stems from the fact that setting $\boldsymbol{\Sigma}$ to be diagonal one obtains even simpler equations:

$$y_{(s),1} = \sigma_1^{\frac{1}{2}} b_1 + \frac{\sigma^2}{\sqrt{n}} w_i,$$

where σ_1 is the first eigenvalue of the matrix and σ^2 is the noise variance.

Remark 2.5. *Notice that we only get an upper bound by assumption, but wer have that the denominator is always positive by assumption. So the result is true up to constants in the interval $(0, 1 - \frac{1}{c_0})$.*

The behavior of $\lambda_\star$ From Figure 4, we would expect that as $n \to \infty$ we have $\lambda_\star \to 0^+$. Heuristically, say it is small. Take a covariance $\boldsymbol{\Sigma}$ in which the eigenvalues decay to zero:

$$\sigma_1 \geq \cdots \geq \sigma_l \downarrow 0.$$

Then, one can find a k such that:

$$\sigma_k \geq \lambda_\star \geq \sigma_{k+1},$$

with very small difference, as to assume $\lambda_\star \approx \sigma_k$. The equation fo $F(\lambda)$ becomes:

$$n = \sum_{l=1}^{\infty} \frac{\sigma_l}{\sigma_l + \lambda_\star} \approx \sum_{l=1}^{k} \underbrace{\frac{\sigma_l}{\sigma_l + \lambda_\star}}_{\approx 1} + \sum_{l \geq k} \frac{\sigma_l}{\sigma_k} \qquad \text{since } \sigma_l \gg \sigma_k$$

$$\asymp k + \frac{1}{\sigma_k} \sum_{k+1}^{\infty} \sigma_l := k + r_k^{(1)},$$

where we have defined the *effective rank*. If the eigenvalues after σ_k are either approximately σ_k or zero the ratio is the number of nonzero eigenvalues after k, thus the name.

Example 2.6. *Say $\sigma_k \asymp k^{-\alpha}$ for $\alpha > 1$ to make it summable. Then:*

$$r_k^{(1)} \asymp \frac{1}{k^{-\alpha}} \int_k^{\infty} x^{-\alpha} \mathrm{d}x \asymp k,$$

so that $n \asymp 2k \asymp k$. The effective regularization is some order of the n^{th} eigenvalue of $\mathbf{\Sigma}$, i.e. $\lambda_\star \asymp \sigma_n$.

Example 2.7. *Similarly, let $\sigma_k = k^{-1}(\log k)^{-\beta}$ where $\beta > 1$. Then:*

$$r_k^{(1)} \asymp k \log k \implies n \asymp k + k \log k \asymp k \log k \implies k \asymp \frac{n}{n \log n}.$$

We do the same procedure of splitting between larger and smaller eigenvalues for V_n^{th}. What is found is that:

$$V_n^{th} \asymp \frac{\sigma^2}{n} \mathrm{Tr}\left(\mathbf{\Sigma}^2(\mathbf{\Sigma} + \lambda_\star \mathbf{I})^{-2}\right) \asymp \frac{\sigma^2}{n} \left(k + \underbrace{\frac{1}{\sigma_k^2} \sum_{l>k} \sigma_l^2}_{r_k^{(2)}} \right) \asymp k \frac{\sigma^2}{n}$$

where in general $r_k^{(2)} \ll r_k^{(1)}$, and the last asymptotic is roughly but not always true. The variance is the same as that of Linear Regression with k parameters, and k is the effective degrees of freedom[5] in the regression problem we set. Similarly, for the bias:

$$B_n^{th} \asymp \lambda_\star^2 \left\langle \boldsymbol{b}_*, \mathbf{\Sigma}(\mathbf{\Sigma} + \lambda_\star \mathbf{I})^{-2} \boldsymbol{b}_* \right\rangle \asymp \sigma_k^2 \left[\sum_{l \leq k} \frac{\overline{b}_l^2}{\sigma_l} + \sum_{l \geq k} \overline{b}_l^2 \frac{\sigma_l}{\sigma_k^2} \right] \quad \text{in the basis of the eigenvectors of } \mathbf{\Sigma}$$

$$\asymp \sum_{l \leq k} \frac{\sigma_k^2}{\sigma_l} \overline{b}_l^2 + \sum_{l>k} \sigma_l \overline{b}_l^2,$$

[5] indeed, if the variance is identity, V_n^{th} is approximately p as $\lambda \to 0^+$, so k is the dof!

where the first term in the second asymptotics is the coefficient of $\boldsymbol{b}_*$ at the l^{th} entry in the basis of eigenvectors of $\boldsymbol{\Sigma}$, namely $\langle \boldsymbol{v}_l, \boldsymbol{b}_* \rangle$ for $\boldsymbol{v}_l$ the l^{th} eigenvector of $\boldsymbol{\Sigma}$. The second term is also interesting. We are not fitting at all the projection of $\boldsymbol{b}_*$ onto the eigenvectors after the k^{th}. It is the same as:

$$\langle \boldsymbol{b}_*, \mathbf{P}_{>k} \boldsymbol{\Sigma} \mathbf{P}_{>k} \boldsymbol{b}_* \rangle = \|\mathbf{P}_{>k} \boldsymbol{b}_*\|_{\boldsymbol{\Sigma}}^2 .$$

In general, we could always have written:

$$\widehat{\mathcal{R}}_n = \left\| \widehat{\boldsymbol{b}}_\lambda - \boldsymbol{b}_* \right\|_{\boldsymbol{\Sigma}}^2 = \left\| \mathbf{P}_{\leq k} (\widehat{\boldsymbol{b}}_\lambda - \boldsymbol{b}_*) \right\|_{\boldsymbol{\Sigma}}^2 + \left\| \mathbf{P}_{>k} (\widehat{\boldsymbol{b}}_\lambda - \boldsymbol{b}_*) \right\|_{\boldsymbol{\Sigma}}^2 ,$$

which in our case means that $\widehat{\boldsymbol{b}}_\lambda$ is zero in the second projection, and non trivial in the first projection, since $\sigma_k < \sigma_l$.

What we learn from this is that one can have both $V_n^{th}, B_n^{th} \to 0$, if for example $\frac{k}{n} \to 0$, which holds when $\boldsymbol{b}_*$ is concentrated onto the top k eigenvectors of $\boldsymbol{\Sigma}$ (i.e. second term in last equation being null nevertheless), and something more to make the first term vanish. This phenomenon is established for $\lambda = 0$, and was named *benign overfitting*, to underline that in contrast with expectations, overfitting is beneficial at high dimensions [Bar+20].

Remark 2.8. *For the rate $k \asymp \frac{n}{\log n}$ the speed is $\frac{1}{\log n}$, so it could in principle be very slow. The technique is not optimal.*

We now turn to justify further why $\lambda_\star$ should be different than zero. We start with a simple example: the *Noise in covariates model*. Let for δ small:

$$\mathbf{Z} = \mathbf{Z}_0 + \delta \mathbf{W} \quad W_{ij} \sim \mathcal{N}(0, 1).$$

A Ridge regression finds the estimator of Eqn. 3. We recall it below:

$$\widehat{\boldsymbol{b}}_\lambda = \frac{1}{n} \mathbf{Z}^\top \left(\lambda \mathbf{I}_n + \underbrace{\frac{1}{n} \mathbf{Z} \mathbf{Z}^\top}_{:= \mathbf{K}_n} \right)^{-1} \boldsymbol{y},$$

where $\frac{1}{n} \mathbf{Z} \mathbf{Z}^\top$ is a *kernel matrix*. Indeed

$$(K_n)_{ij} = \frac{1}{n} \langle \boldsymbol{z}_i, \boldsymbol{z}_j \rangle$$

$$\mathbf{K}_n = \underbrace{\frac{1}{n} \mathbf{Z}_0 \mathbf{Z}_0^\top}_{:= \mathbf{K}_{n,0}} + \frac{\delta^2}{n} \mathbf{W} \mathbf{W}^\top \qquad \text{modulo ignored some cross terms.}$$

The the second term is also expressed as:

$$\frac{\delta^2}{n} \mathbf{W} \mathbf{W}^\top = \frac{\delta^2}{n} p \mathbf{I} + \frac{\delta^2}{n} \left(\sqrt{p} \mathbf{W} \mathbf{W}^\top - \underbrace{p \mathbf{I}}_{\mathbb{E}[\mathbf{W} \mathbf{W}^\top]} \right).$$

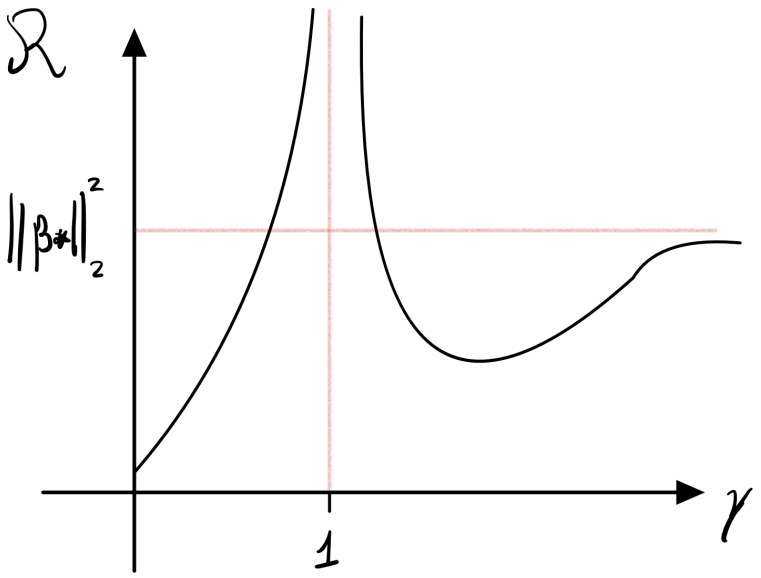

Figure 5: Risk behavior as a function of γ {fig:ri

The matrix inside the parenthesis is in $\mathbb{R}^n \times \mathbb{R}^n$, it is symmetric with zero mean and entries with norm $\pm\sqrt{p}$ being products of Gaussians, like a Wigner matrix. It is easy to prove then that the operator norm of the matrix is $\sqrt{np}$. Then we rewrite it as:

$$\frac{\delta^2}{n}\mathbf{W}\mathbf{W}^\top = \frac{\delta^2}{n}p\mathbf{I} + \delta^2\sqrt{\frac{p}{n}}\mathbf{G},$$

where $\mathbf{G}$ is a Wishart matrix. For $p \gg n$ we get that:

$$\mathbf{K}_n = \mathbf{K}_{n,0} + \frac{p}{n}\delta^2\mathbf{I} + h.o.t.,$$

and in practice $\lambda' = \lambda + \frac{p\delta^2}{n}$. The presence of noise in the covariates adds regularization to the model. It is one of the oldest examples of regularization via noise. Something almost equivalent was found in [Bis95], or recently and more in detail by [KLS20].

We can now look at specific examples, which turn into choices of $(\mathbf{\Sigma}, \boldsymbol{b}_\star)$.

Example 2.9. *Let $\mathbf{\Sigma} = \mathbf{I}; \boldsymbol{b}_\star$ be isotropic (rotationally invariant). For $\lambda = 0^+, \frac{p}{n} \to \gamma$ the behavior of the risk is as in Figure 5. We see a divergence at $\gamma = 1$ since the $\mathbf{Z}$ matrix coincides with the divergence of the condition number. As $\gamma \to \infty$ the function changes, so this is not a conceptual model for overparametrization. The number of parameters increases, but the function to learn changes. Despite the double descent phenomenon we should ignore it. This was first pointed out in [AS17].*

Example 2.10 (Latent space model)**.** *Construct the following observations:*

$$y_i = \boldsymbol{\theta}_*^\top \boldsymbol{g}_i + \xi_i; \quad z_{ij} = \boldsymbol{w}_j^\top \boldsymbol{g}_i + u_{ij},$$

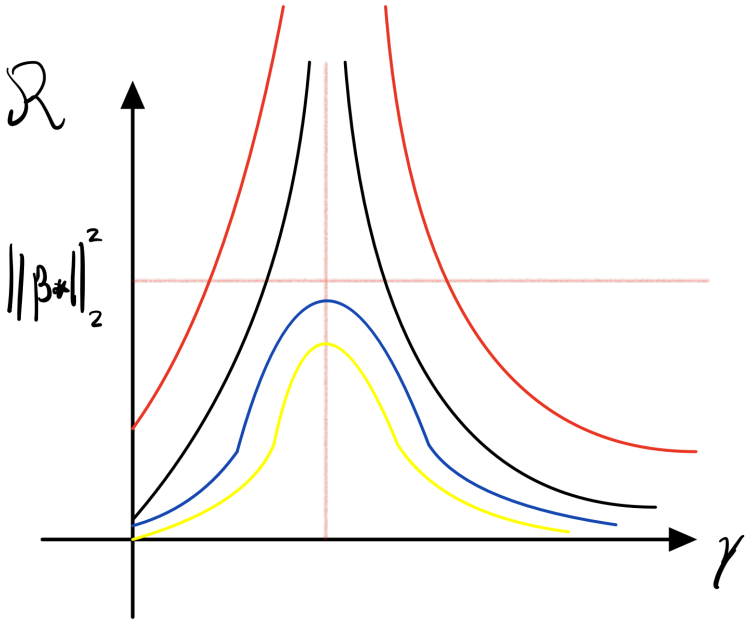

Figure 6: Risk behavior as a function of γ, different regularizations
The Risk goes to zero as $\gamma \to \infty$, differently from before, so over-parametrization induces null risk. Turning on some regularization the divergence disappears and we get the blue curve, the optimal regularizer is the yellow curve, which is monotone. In red we plot the norm of $\widehat{\boldsymbol{b}}_\lambda$, which has a qualitatively similar behavior to the risk at no regularization. {fig:mu

where we have sampled independently $\boldsymbol{g}_i \sim \mathcal{N}(0, \mathbf{I}_d), \xi_i \sim \mathcal{N}(0, \sigma_\xi^2), u_{ij} \sim \mathcal{N}(0,1)$ *for* $d \ll p$. *This is equivalent to providing* $(\boldsymbol{\Sigma}, \boldsymbol{b}_\star)$, *but underlines that there is a* d*-dimensional latent space with the true repsonse and the covariates being linear functions of the latent space. Notice that* $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{W}\mathbf{W}^\top + \mathbf{I})$ *so* $\boldsymbol{\Sigma}$ *is well specified and* $\boldsymbol{y}$ *is jointly Gaussian with* $\mathbf{Z}$. *This ensures that there exists a* $\boldsymbol{b}_\star$ *solution.*
In this setting, we can vary the overparametrization while keeping the d*-dimensional latent space fixed. This is done by projecting the covariates in different directions. The model is a conceptually good toy model.*
Assume that $\frac{p}{n} \to \gamma, \frac{d}{n} \to \frac{1}{100}$ *or any other constant. We report in Figure 6 the risk-*γ *plot.*

In the previous example, we find benign overfitting as before. The covariance $\boldsymbol{\Sigma}$ is approximately low-rank (since $d \ll p$) so that the eigenvalues decay and $\widehat{\boldsymbol{b}}_\lambda$ is aligned with the top eigenvectors.

2.2 Kernel Ridge Regression

The interest for Kernel Ridge Regression (KRR) stems from two observations:

- it is non-parametric $p \to \infty$

- is is the limit of Neural Networks in the linear regime, as we will see in the next sections.

For background, we remind the optimization problem that is sought in KRR:

$$\widehat{f}_\lambda = \arg\min \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 + \lambda \|f\|_K^2 \right\},$$

where $\|\cdot\|_K$ is the RHKS[6] norm in terms of the kernel K. By Kernel, we mean a map that is also positive semi-definite (psd). Mathematically:

$$K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, \quad (K(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j\in[n]} \succeq 0.$$

To the kernel, we associate a linear operator:

$$\mathcal{K} : L^2(\mathbb{P}) \to L^2(\mathbb{P}) \quad \mathcal{K}(g(\boldsymbol{x})) \equiv \int K(\boldsymbol{x}, \boldsymbol{x}') g(\boldsymbol{x}') \mathbb{P}(\mathrm{d}\boldsymbol{x}'),$$

where $\mathbb{P}$ is the data distribution on $\mathbb{R}^d$. We can decompose the Kernel into eigenvalues and eigenfunctions in an orthonormal set:

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{l=1}^{\infty} \lambda_l \varphi_l(\boldsymbol{x}_1) \varphi_x(\boldsymbol{x}_2) \quad \{\lambda_l, \varphi_l\}_{l=1}^{\infty}.$$

Remark 2.11. *Taking $\mathcal{K}$ a continuous function, we can bound its trace operator from above, and from below by the eigendecomposition. This makes the sum of the eigenvalues finite since*

$$C \geq \mathrm{Tr}\left(\mathcal{K}(\mathsf{id})\right) = \int K(\boldsymbol{x}, \boldsymbol{x}) \mathbb{P}(\mathrm{d}\boldsymbol{x}) = \sum_{l=1}^{\infty} \lambda_l^2,$$

and in particular we can order the eigenvalues:

$$\infty > \lambda_1^2 > \cdots > 0 \quad \lambda_l \downarrow 0.$$

The above construction allows us to define the norm of a function with respect to the kernel, which takes form:

$$\langle f, \mathcal{K} f \rangle \coloneqq \|f\|_K^2 = \sum_{l=1}^{\infty} \frac{1}{\lambda_l^2} \langle \varphi_l, f \rangle_{L^2}$$

where to define the last object we have ignored null eigenvalues.

Example 2.12. *The simplest example is for $\mathbb{P}[\mathrm{d}x] = \mathcal{U}nif(0, 2\pi)$. Say $f : [0, 2\pi] \to \mathbb{C}$ and the eigenfunctions are the Fourier period functions $\varphi_l(x) = e^{\mathrm{i}kx}$ for $k \in \mathbb{Z}$. To make the eigenvalues summable, we explicitly construct them as:*

$$\lambda_l^2 = (1 + |l|^{2s})^{-1} \quad s > \frac{1}{2}.$$

[6]Reproducing Kernel Hilbert Space

Given the eigenpairs $\{(\lambda_l, \varphi_l)\}$ *we obtain the kernel* $K(x_1, x_2) = \sum_{l=1}^{\infty} \lambda_l^2 \varphi_l(x_1)\varphi_l(x_2)$ *as before. The Kernel norm of a function will then be:*

$$\begin{aligned}\|f\|_K^2 &= \sum_{q\in\mathbb{Z}} (1 + |q|^{2s}) \langle \varphi_q, f\rangle_{L^2}^2 \\ &= \frac{1}{2\pi}\int_0^{2\pi} \left(|f(x)|^2 + |f^{(s)}(x)|^2\right) \mathrm{d}x \qquad\qquad \textit{if } s \in \mathbb{N},\end{aligned}$$

where the last passage follows by Parceval's identity, and we find a Sobolev norm.

To extend the last example, it can be shown that RHKS generalize some subsets of Sobolev spaces, and most importantly they can encode the smoothness of a function.

We will find that the optimal estimator $\widehat{f}_\lambda$ is the solution of an infinite dimensional problem with a finite dimensional description. The way we derive this is one of the many. It makes use of the Representer Theorem. Expand f in the basis of eigenvectors:

$$f = \sum_{l=1}^{\infty} a_l \varphi_l \quad \boldsymbol{\Phi} \equiv \begin{bmatrix} \varphi_1(x_1) & \varphi_2(x_1) & \cdots \\ \varphi_1(x_1) & \varphi_2(x_2) & \cdots \\ \vdots & \vdots & \vdots \\ \varphi_1(x_n) & \varphi_2(x_n) & \cdots \end{bmatrix} \in \mathbb{R}^{n\times\infty}.$$

The construction of the matrix $\boldsymbol{\Phi}$ highlights that to optimize f with KRR we are eventually doing Ridge Regression on the matrix of data $\boldsymbol{\Phi}$, which has infinite parameters. To invert the relation and find $\boldsymbol{a}$ we apply the usual formula:

$$\widehat{\boldsymbol{a}}_\lambda = \arg\min \left\{ \|\boldsymbol{y} - \boldsymbol{\Phi a}\|^2 + \lambda \left\langle \boldsymbol{a}, \mathbf{D}^{-1}\boldsymbol{a}\right\rangle \right\}, \mathbf{D} = \mathsf{diag}(\lambda_1^2, \lambda_2^2, \ldots).$$

Remark 2.13. *Notice that* $f = \boldsymbol{\Phi a}$ *and in the basis of the eigenvectors of* $\mathbf{K}$ *the matrix norm is just the diagonal of eigenvalues and* $\|f\|_K^2 = \langle \boldsymbol{a}, \mathbf{D}^{-1}\boldsymbol{a}\rangle$ *effectively. We are just rewriting the usual optimization with the new norm.*

The result of the optimization is a vector:

$$\widehat{\boldsymbol{a}}_\lambda = \mathbf{D}\boldsymbol{\Phi}^\top \left(\lambda \mathbf{I}_n + \boldsymbol{\Phi}\mathbf{D}\boldsymbol{\Phi}^\top\right)^{-1} \boldsymbol{y},$$

where the matrix $\mathbf{K}_n = \boldsymbol{\Phi}\mathbf{D}\boldsymbol{\Phi}^\top$ has entries:

$$(\mathbf{K}_n)_{ij} = \sum_{l=1}^{\infty} \lambda_l^2 \varphi_l(\boldsymbol{x}_i)\varphi_l(\boldsymbol{x}_j),$$

which is the same formula of plain Ridge Regression in all its passages. Having $\widehat{\boldsymbol{a}}_\lambda$ we eventually compute the estimator function:

$$\widehat{f}_\lambda(\boldsymbol{x}) = \sum_{l=1}^{\infty} \widehat{a}_{\lambda,l}\varphi_l(x) = K(\boldsymbol{x}, \cdot)^\top (\lambda \mathbf{I}_n + \mathbf{K}_n)^{-1}\boldsymbol{y}, \quad \mathbf{K}_n \in \mathbb{R}^{n\times n}, K(\boldsymbol{x}, \cdot) = \begin{bmatrix} K(\boldsymbol{x}, \boldsymbol{x}_1) \\ \vdots \\ K_n(\boldsymbol{x}, \boldsymbol{x}_n) \end{bmatrix}$$

End of Lecture 2

3 Neural Networks

We found that the Kernel Ridge Regression (KRR) has an explicit solutions in terms of $\mathbf{K}_n, K$. We report it below to recap:

$$\begin{aligned}\widehat{f}_\lambda &= \arg\min\left\{\sum_{i=1}^n (y_i - f(\boldsymbol{x}_i))^2 + \lambda \|f\|_K^2\right\} \\ &= K(\boldsymbol{x}, \cdot)^\top (\lambda \mathbf{I}_n + \mathbf{K}_n)^{-1} \boldsymbol{y}.\end{aligned}$$

Our results will rely heavily on this.

Open Problem #1

Redo the following for other losses ℓ.

Kernel choice

Which kernel we should choose to make contact with linearized 2-layer Neural Networks?

We briefly recap the expression for N hidden neurons which we derived before:

$$f_{lin}(\boldsymbol{x}, \boldsymbol{b}) = \langle \boldsymbol{b}, \varphi(\boldsymbol{x})\rangle \quad \varphi(\boldsymbol{x}) = \frac{1}{\sqrt{Nd}} \underbrace{\left(\sigma'(\boldsymbol{w}_1^\top \boldsymbol{x})\boldsymbol{x}^\top, \dots, \sigma'(\boldsymbol{w}_N^\top \boldsymbol{x})\boldsymbol{x}^\top\right)^\top}_{\in \mathbb{R}^{Nd}}.$$

The Ridge regression with feature map φ is equivalent to the KRR with (finite dimensional equivalent) kernel:

$$\begin{aligned}K_N(\boldsymbol{x}_1, \boldsymbol{x}_2) &= \langle \varphi(\boldsymbol{x}_1), \varphi(\boldsymbol{x}_2)\rangle \\ &= \frac{1}{Nd} \langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle \sum_{i=1}^N \sigma'(\boldsymbol{w}_i^\top \boldsymbol{x}_1)\sigma'(\boldsymbol{w}_i^\top \boldsymbol{x}_2) \\ &\overset{N\to\infty}{\to} K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{1}{d} \langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle \mathbb{E}_{\boldsymbol{w}} \left[\sigma'(\boldsymbol{w}^\top \boldsymbol{x}_1)\sigma'(\boldsymbol{w}^\top \boldsymbol{x}_2)\right],\end{aligned}$$

where φ was computed some lectures ago. If $\boldsymbol{w}_i \sim \mathcal{U}nif(\mathbb{S}^{d-1})$ the expectation is *rotationally invariant* and for $\|\boldsymbol{x}_i\|_2 = \sqrt{d}$ one gets:

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \hbar_d\left(\frac{\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle}{d}\right),$$

so the limit kernel in the linear regime is an *inner product kernel*.

We claim that the dependence on the dimension is roughly null, namely:

$$\hbar_d(q) \overset{d\to\infty}{\to} \hbar(q),$$

which is reasonable since the interpolation-benign overfitting effects are typically at $d \gg 1$. We will be a little sloppy. Notice that:

$$\begin{aligned} K(\boldsymbol{x}_1, \boldsymbol{x}_2) &\approx \frac{\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle}{d} \mathbb{E}_{\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \frac{\mathbf{I}_d}{d})} \left[\sigma'(\boldsymbol{w}^\top \boldsymbol{x}_1) \sigma'(\boldsymbol{w}^\top \boldsymbol{x}_2) \right] \\ &= \frac{\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle}{d} \mathbb{E}_{\boldsymbol{G}_1, \boldsymbol{G}_2} \left[\sigma'(\boldsymbol{G}_1) \sigma'(\boldsymbol{G}_2) \right] \qquad \boldsymbol{G}_1, \boldsymbol{G}_2 \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & \frac{\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle}{d} \\ \frac{\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle}{d} & 1 \end{bmatrix} \right) \\ &= \hbar \left(\frac{\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle}{d} \right). \end{aligned}$$

We then see that:

$$\hbar(q) = q \mathbb{E}_{\boldsymbol{G}_1, \boldsymbol{G}_2} \left[\sigma'(\boldsymbol{G}_1) \sigma'(\boldsymbol{G}_2) \right] \qquad \boldsymbol{G}_1, \boldsymbol{G}_2 \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & q \\ q & 1 \end{bmatrix} \right).$$

The only way in which d enters into the picture is in terms of the input q.

Remark 3.1. *All of the above is a heuristic procedure to arrive at an expression for the kernel. In the next lecture, we will provide details on the speed of convergence when $N \to \infty$.*

We now state a technical assumption.

{ass:ta

Assumption 2 (Taylor niceness)**.** *The map $\hbar_q$ has expansion*

$$\hbar_d(x) = \sum_{l=1}^{\infty} \frac{c_l(d)}{l!} x_l \quad c_l(d) \geq 0, \ \forall l, d.$$

Additionally, assume that $c_l(\infty) \geq 0$ for all l, which makes the activation function generic.

Example 3.2. *The first part of Ass. 2 is true if $\sigma' \in L^2$ the second if σ is the shifted* ReLu.

Remark 3.3. *Consider the L layer MLP expressed as usual by:*

$$f(\boldsymbol{x}) = \sigma \circ \mathbf{W}_1 \circ \mathbf{W}_2 \circ \cdots \circ \mathbf{W}_L(\boldsymbol{x}),$$

with standard initialization $(W_l)_{ij} \sim \mathcal{N}(0, \tau_l^2)$ the NTK kernel is also an inner product kernel. Namely:

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \hbar \left(\frac{\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle}{d} \right).$$

This should be obvious without computations. The linearization is rotation invariant. Any function of two arguments which is rotation invariant depends on the norms of the inputs and the inner product. The norms are fixed, and dependence is only one the inner product. This holds for any rotationally invariant distribution.

Remark 3.4. *The reasoning above does not apply to any architecture since the kernel changes. Consider a Convolutional Neural Network (CNN) with two layers and average pooling. We wish the kernel to be translationally invariant. Let $\boldsymbol{x} \in \mathbb{R}^d$. The CNN is invariant under* ***translations****, so we inspect:*

$$T^l(\boldsymbol{x}) = (x_l, x_{l+1}, \dots, x_{l-1}),$$

for T^l a member of the group of translations (the cyclic group). The classic example is an image translation, but we keep the $\boldsymbol{x} \in \mathbb{R}^d$ for simplicity, like a one-dimensional image. The Kernel will be:

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{1}{d} \sum_{l=0}^{d-1} \hbar \left(\frac{\langle \boldsymbol{x}_1, T^l \boldsymbol{x}_2 \rangle}{d} \right),$$

which is the average of the previous kernel over the translations of the group.

Remark 3.5. *The linearization of the MLP follows by the linearization of the case we saw at the beginning for a 2-layer net.*

3.1 More about previous problems

The results we state now hold for any infinite width kernel of 2-layer NNs and was explored in a series of works. We focus on [MMM21] for the moment. There are also extensions to CNNs. As a starting point, notice that the dependence on dimension d is inside the input, namely:

$$\hbar_d \left(\frac{\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle}{d} \right) = \hbar \left(\frac{\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle}{d} \right) \quad \|\boldsymbol{x}\| = \sqrt{d}.$$

Open Problems #2

We want to analyze the following:

1. the risk of the KRR for given kernel K
2. when is N is large enough so that the risk of the finite width kernel is *close* to the risk of the infinite width version?
3. when are NNs better than kernels?

Problems #1, #2 have been done to some extent for inner product kernels and convolutional kernels, while for #3 there are only special cases.

We now state an important Theorem. The setting is as follows. Consider an inner product kernel for which Ass. 2 holds. Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \overset{iid}{\sim} \mathcal{U}nif(\mathbb{S}^{d-1}(\sqrt{d}))$ and:

$$y_i = f_*(\boldsymbol{x}_i) + \epsilon_i \quad \mathbb{E}[\epsilon] = 0, \ \mathbb{E}\left[\epsilon^2\right] = \tau^2.$$

To make the risk $\mathbb{E}[(f_* - f)^2]$ computable, assume also that $\mathbb{E}[f_*^2] = \|f_*\|_{L^2}^2 < \infty$.

{thm:me

Theorem 3.6 (Mei, Monta, THeo [MMM21]). *With the setting above and $d \to \infty, n \to \infty$, there exists $\lambda_0 > 0$ such that for all[7] $\lambda \in [0, \lambda_0]$ the following holds: if $d^{l+\delta} \leq n \leq d^{l+1-\delta}$ for some $\delta > 0$ then*

$$\mathcal{R}(f_*, \lambda) = \mathbb{E}\left[\left(\widehat{f}_\lambda(\boldsymbol{x}_{new}) - f_*(\boldsymbol{x}_{new})\right)^2\right] = \|\mathcal{P}_{>l} f_*\|_{L^2}^2 + o_d(1)(\|f_*\|_{L^2}^2 + \tau^2)$$
$$= \min_{p \in \mathsf{poly}(l)} \mathbb{E}\left[(f_*(\boldsymbol{x}) - p(\boldsymbol{x}))^2\right] + o_d(1)(\|f_*\|_{L^2}^2 + \tau^2),$$

where $\mathcal{P}_{>l}$ is the projection to polynomials of degree greater than l.
Further, no RKHS method with inner product kernel can achieve lower risk changing λ or the training loss function and same test error with square loss.

We now turn to interpreting the result. The condition on the number of samples n is roughly translated as *data is able to fit an l degree polynomial but not marginally more than that.* Additionally, rotational invariance would imply that if one $l+1$ degree polynomial is fitted then all of them can be, leading to a contradiction. In this case, rotational invariance can be seen as a curse of dimensionality.

Remark 3.7. *For conditional kernels with f_* rotationally invariant (e.g. $\sum_{i=1}^d x_i x_{i+1}$) the same result holds with a power of d less. Namely, d^{l-1} points are needed to fit a degree l polynomial. For other groups, there is an abstract theorem [MMM21].*

A generalization for other high-dimensional isotropic distributions (e.g. Rademacher ± 1 at d dimensions) was also derived.

Open Problem #3

What happens with distributions with a latent low-dimensional structure?

The motivation for Open Problem #3 is that classes of inputs (e.g. images) often lie on a lower dimensional manifold. This idea was studied in a special case which is now briefly explained. It was first presented in [Gho+19; Gho+20; Gho+21]. Take:

$$\boldsymbol{x} = \mathbf{U}\boldsymbol{x}_1 + \mathbf{U}^\perp \boldsymbol{x}_2, \quad \mathbf{U} \in \mathbb{R}^{d \times k}, \mathbf{U}^\perp \in \mathbb{R}^{d \times (d-k)}, \ [\mathbf{U}.\mathbf{U}^\perp] \in \mathcal{O}(d),$$

where $\mathcal{O}(d)$ is the group of orthogonal matrices and $\mathbf{U}.\mathbf{U}^\perp \in \mathbb{R}^{d \times d}$ is obtained by stacking the columns of the two matrices. A visualization in $\mathbb{R}^2$ is Figure 7. Assuming further that $\boldsymbol{x}_1 \sim \mathcal{U}nif(\mathbb{S}^k_{\rho_1}), \boldsymbol{x}_2 \sim \mathcal{U}nif(\mathbb{S}^{d-k}_{\rho_2})$ with $\rho_1 > \rho_2$ the points lie on an *ellipsoid*, also depicted in Figure 7. In particular, let $k = d^\alpha$ and $\frac{\rho_1}{\rho_2} = d^\beta$. This choice of polynomial dimensions makes the previous results hold for an *effective dimension* in place of d^l.

Remark 3.8. *While this low dimensional construction may seem not intuitive, one can think of an image in the Fourier domain: low frequency components have larger variance and are collected in $\mathbf{U}$, while the others are collected in $\mathbf{U}^\perp$.*

[7] at $\lambda = 0$ we get the minimum-norm interpolator

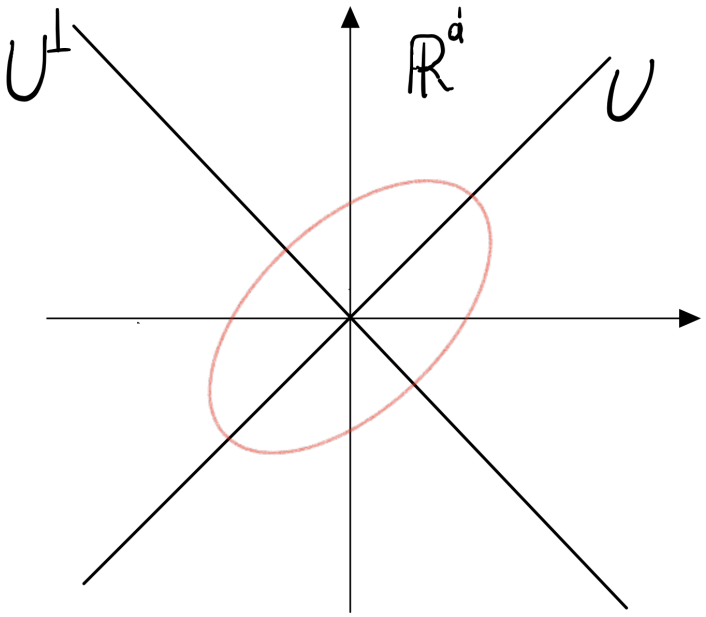

Figure 7: The ellipsoid of the synthetic example {fig:el

We come back to the interpretation of Thm. 3.6. The error can always be decomposed as:

$$\left\|\widehat{f} - f_*\right\|_{L^2}^2 = \left\|\mathcal{P}_{<l}(\widehat{f} - f_*)\right\|_{L^2}^2 + \left\|\mathcal{P}_{>l}(\widehat{f} - f_*)\right\|_{L^2}^2 .$$

From the result, we see that the first element of the RHS is null and the second is not impacted, namely $\widehat{f} \approx \mathcal{P}_{\leq l} f_*$. We then get that interpolation is optimal, precisely in the sense that $\lambda = 0$ is the best up to vanishing $o_d(1)$ corrections. The setting of Thm. 3.6 is then a second example of *benign overfitting*.

In particular if $\|\mathcal{P}_{>l} f_*\|_{L^2} \overset{l\to\infty}{\to} 0$ then the excess risk goes to zero. A depiction is Figure 8, where the *abscissa* is $\frac{\log n}{\log d} \sim l$ by the scaling of the Theorem. The risk is flat between integers and jumps when an additional degree is added. One can further think of the estimator as a decomposition:

$$\widehat{f} = f_{smooth} + f_{spiky}, \quad f_{smooth} = \mathcal{P}_{<l} f_*, f_{spiky}(\boldsymbol{x}) = f_*(\boldsymbol{x}) - \mathcal{P}_{>l} f_*(\boldsymbol{x}) \ \forall \boldsymbol{x}, \ \|f_{spiky}\|_{L^2} = o(1).$$

The above equation is understood as follows. The fitting process returns a smooth component and a spiky part that is very small in L^2 norm in high dimensions[8] (see Figure 9 blue is smooth, red is spiky).

Open Problem #4

What is the structure/geometry of these spikes? Knowing this would provide results on the risk of test points at perturbations of training points.

[8]Notably, large spikes in high dimension have small norm

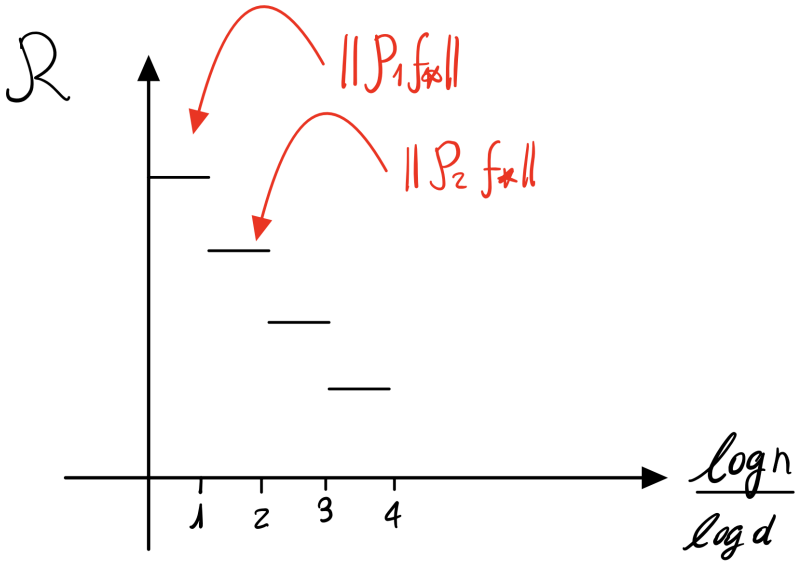

Figure 8: Staircase risk behavior {fig:st

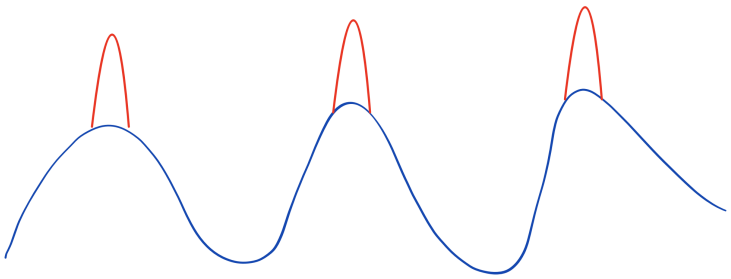

Figure 9: Estimator as a smooth(blue)+spiky (red) combination. {fig:sp

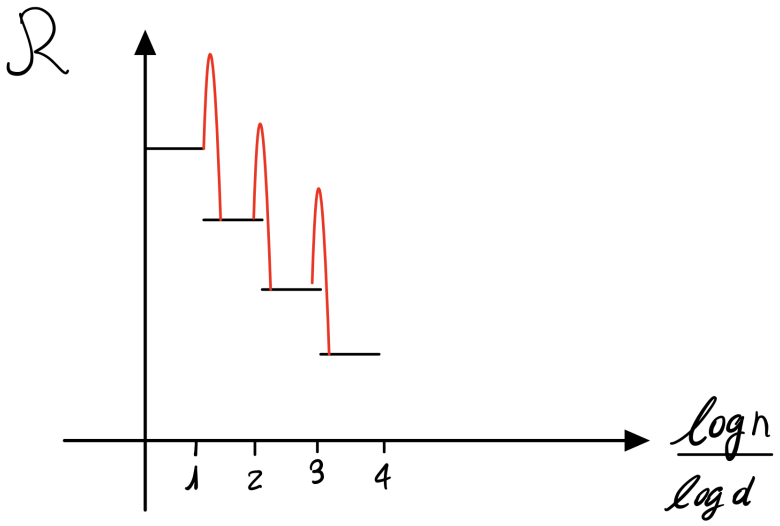

Figure 10: Staircase with bumps {fig:st

Remark 3.9 (Another perspective on Thm. 3.6)**.** *Mathematically, the risk in this setting is the same of a polynomial regression of degree l, but the function fitted in the latter case misses the spiky components.*

A question that might come to mind at this point is what happens at the boundaries of the staircase of Figure 8? This has been studied recently for the case of the sphere and the hypercube distribution of the signals [Mis22; LY23; TAP21]. What happens is that every polynomial in l term is fitted, anything that is degree $> l + 2$ is seen as noise[9]. All that matters is the component of degree 2 of the kernel and of the function. The design matrix of degree 2 polynomial behaves *roughly* as a Wishart matrix for any transition. There is a universal formula that gives the shape of these spikes, since the phenomenon is always related to this degree 2 Wishart matrix. Basically one can describe the aspect of the multiple descents in a nice way. A depiction is Figure 10.

The mathematical intuition for why this happens is worth exploring. The setting is Random Matrix Theory models with large aspect ratio or polynomial dependence between entries, which are slightly different from canonical research areas in RMT. The main aspect we will see is the structure of the empirical kernel $\mathbf{K}_n = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j\leq n}$. Since K is an inner product kernel , it diagonalizes in the basis of spherical harmonics, which we denote as:

$$\boldsymbol{Y}_l(\boldsymbol{x}) = \left[Y_{l,1}(\boldsymbol{x}), \dots, Y_{l,D(l)}(\boldsymbol{x})\right], \quad l \in \mathbb{N} \cup \{0\}, \quad \boldsymbol{x} \in \mathbb{S}^{d-1}(\sqrt{d}).$$

The functions $\{Y_{l,j}\}_{j\leq D(l)}$ are l degree homogeneous harmonic polynomials restricted to the $\sqrt{d}$ radius sphere. They can be made orthonormal, namely such that $\langle Y_{l,k}, Y_{l',k'}\rangle = \delta_{kk'}\delta_{ll'}$, so that

[9]think of a polynomial of $l = 3$ degree sampled at n^2 sample points, like a high frequency cosine in the unit interval: it appears as noise.

together they form an *orthonormal basis of square integrable functions on the sphere*. The degeneracy (# dimensions) is roughly[10] d^l, by permutation invariance it is $\frac{d^l}{l!}$. For simplicity, we will state this as $D(l) \asymp d^l$.
Obviously, rotating $\boldsymbol{x}$ the span of $\boldsymbol{Y}$ functions is the same since it is invariant. For this reason it diagonalizes an invariant kernel K in such space.
Consider a set of orthonormal functions $\{\varphi_i\}_{i\leq D}$. In general the projection onto the span $V = \mathsf{span}\{\varphi_i\}_{i\leq D}$ satisfies:

$$(\mathcal{P}_V)(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{i=1}^{\mathsf{dim}(V)} \varphi_i(\boldsymbol{x}_1)\varphi_i(\boldsymbol{x}_2)^\top.$$

So the projector onto l-degree polynomials seen as a kernel (an integral operator) reads:

$$\mathcal{P}_l(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{j=1}^{D(l)} Y_{l,j}(\boldsymbol{x}_1)Y_{l,j}(\boldsymbol{x}_2) = \boldsymbol{Y}_l^\top(\boldsymbol{x}_1)\boldsymbol{Y}_l(\boldsymbol{x}_2).$$

Now the inner product kernel diagonalizes and is invariant so we write it in terms of spherical harmonics as:

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{l=0}^{\infty} \lambda_l^2 \boldsymbol{Y}_l^\top(\boldsymbol{x}_1)\boldsymbol{Y}_l(\boldsymbol{x}_2) \succeq 0.$$

The typical size of these eigenvalues is:

$$\mathrm{Tr}\left(\mathcal{K}(\mathsf{Id})\right) = \int_{\mathbb{S}^{d-1}(\sqrt{d})} K(\boldsymbol{x}, \boldsymbol{x})\mu(\mathrm{d}\boldsymbol{x}) = \sum_{l=1}^{\infty} \lambda_l^2 D(l),$$

where $\mu(\mathrm{d}\boldsymbol{x})$ is the uniform measure over the sphere, and letting $\sigma' < \infty$ the trace is bounded. Boundedness ensures that the RHS argument is summable and we derive:

$$\lambda_l^2 \lesssim D(l)^{-1} \lesssim d^{-l}. \tag{7}$$ {eqn:as
{ex: f

Example 3.10. *Under our Assumption $\lambda_l^2 = \frac{c_l}{d^l}$, where the c_l come from Ass. 2, we have:*

$$\hbar(q) = \sum_{l=1}^{\infty} \frac{c_l}{d^l} q^l,$$

with $\hbar_d\left(\frac{\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle}{d}\right) = K(\boldsymbol{x}_1, \boldsymbol{x}_2)$.

Open Problem #4

Assume $c_l \geq 0$ for all l. Then:

$$\lambda_l^2 = \frac{c_l}{d^l}(1 + o(1)).$$

and the chain of inequalities of Eqn. 7 is tight.

[10] the number of degree l independent polynomials

Unfortunately, while the formalism of harmonic analysis on the sphere or on the hypercube is easy, the calculations get complicated for product measures and conditional kernels.
Now that we know the population kernel, what is the empirical kernel? We already know $\mathbf{K}_n = ((K(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j\leq n}$. Then:

$$\mathbf{K}_n = \sum_{m=0}^{l} \lambda_m^2 \widehat{\mathbf{Y}}_m \widehat{\mathbf{Y}}_m^\top + \sum_{m=l+1}^{\infty} \lambda_m^2 \widehat{\mathbf{Y}}_m \widehat{\mathbf{Y}}_m^\top \quad \widehat{\mathbf{Y}}_m = \begin{bmatrix} \boldsymbol{Y}_m(\boldsymbol{x}_1) \\ \vdots \\ \boldsymbol{Y}_m(\boldsymbol{x}_n) \end{bmatrix} \in \mathbb{R}^{n\times D(m)}, \quad \forall m. \tag{8}$$ {eqn:ke

The empirical kernel is composed of a low degree $\mathbf{K}_n^{\leq l}$ and a high degree part $\mathbf{K}_n^{>l}$. This is made explicit since the two have a completely different behavior. Indeed, for $m \leq l$ the matrix $\widehat{\mathbf{Y}}_m$ is *skinny* and low rank since $n \gg D(m)$, while for $m > l$ the matrix is *fat* and full rank. When $n \propto d^l$ the l^{th} term has constant aspect ratio in $\mathbf{K}_n^{\leq l}$, and it will drive the dynamics. In the skinny case when $m \leq l$ an application of the matrix Bernstein Inequality gives:

$$\frac{1}{n}\widehat{\mathbf{Y}}_m^\top \widehat{\mathbf{Y}}_m = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\boldsymbol{Y}_m(\boldsymbol{x}_i)\boldsymbol{Y}_m(\boldsymbol{x}_i)^\top}_{\in \mathbb{R}^{D(m)\times D(m)}} \overset{n\to\infty}{\approx} \mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{Y}_m(\boldsymbol{x})\boldsymbol{Y}_m(\boldsymbol{x})^\top\right], \tag{9}$$ {eqn:fi

which is expected since it is an average of many matrices. However, we can say more, since by the use of spherical harmonics we also have that:

$$\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{Y}_m(\boldsymbol{x})\boldsymbol{Y}_m(\boldsymbol{x})^\top\right] = \mathbf{I}_{D(l)},$$

since it is the matrix of inner products of spherical harmonics. Then, the matrix $\widehat{\mathbf{Y}}_m \approx \sqrt{n}\widehat{\mathbf{U}}_m$ where $\widehat{\mathbf{U}}_m$ is basically orthogonal and:

$$\mathbf{K}_n^{\leq l} \overset{n\to\infty}{\approx} \sum_{m=0}^{l} \underbrace{\frac{c_m}{m!}}_{\lambda_m} \frac{n}{d^m}\widehat{\mathbf{U}}_m\widehat{\mathbf{U}}_m^\top.$$

By this result, this matrix has eigenvalues that scale as $n, \frac{n}{d}, \frac{n}{d^2}, \ldots$, with multiplicities $1, d, d^2, \ldots$. At the $(d^l)^{th}$ term the series stops.
Instead, one can prove that the high degree part satisfies:

$$\mathbf{K}_n^{>l} \overset{n\to\infty}{\approx} \kappa\mathbf{I}_n \quad \kappa \in \mathbb{R},$$

so that the other eigenvalues are of constant order $O(1)$ with multiplicity n.

Having established the pattern of eigenvalues, we conclude that doing KRR with regularization λ is basically equivalent to doing polynomial RR with added regularization $\lambda + \kappa$ for some degree l coming from Thm. 3.6. The effective regularization is increasing as in the case of simple features.

Remark 3.11. *Recall $\hbar(q) = \sum_{l=0}^{\infty} \frac{c_l}{l!} q^l$, where the c_l are for 2-layes NNs related to the Hermite polynomials.*

{rem:KR

Remark 3.12. *KRR and Polynomial RR are essentially equivalent in the L^2 sense but the latter is missing the spiking components! The spiky part depends indeed on the high rank component.*

Eventually, we conclude that the KRR estimator:

$$\widehat{f}_\lambda(\boldsymbol{x}) = K(\boldsymbol{x}, \cdot)^\top \left(\lambda \mathbf{I} + \mathbf{K}_n\right)^{-1} \boldsymbol{y}$$

is such that:

$$\lambda \mathbf{I} + \mathbf{K}_n = (\lambda + \kappa)\mathbf{I} + \mathbf{K}_n^{\leq l},$$

for a shifted regularization polynomial regression. If $\lambda \ll 1$, since $\mathbf{K}_n^{\leq l}$ has eigenvalues $\frac{n}{d^l} \gg 1$ the regularization is useless. When $n \propto d^l$ some eigenvalues become $O(1)$ and regularization starts to be impactful.

Remark 3.13. *The equivalence holds in the L^2 sense, namely establishing it by vicinity in expected norm* $\mathbb{E}\left[\left\|\widehat{f}_{KRR} - \widehat{f}_{PRR}\right\|_{L^2}^2\right] = o(1)$.

End of Lecture 3

4 NTK and Risk analysis

We go back to min-norm regression in the Neural Tangent regime of a 2-layer Neural Network. The feature map in this case is:

$$\varphi(\boldsymbol{x}) = \frac{1}{\sqrt{Nd}}\left[\sigma'(\langle \boldsymbol{w}_1, \boldsymbol{x}\rangle)\boldsymbol{x}^\top, \dots, \sigma'(\langle \boldsymbol{w}_N, \boldsymbol{x}\rangle)\boldsymbol{x}^\top\right], \tag{10}$$

with associated linear model $f_{\mathsf{NT}}(\boldsymbol{x}; \boldsymbol{b}) = \langle \boldsymbol{b}, \varphi(\boldsymbol{x})\rangle$. The optimal parameters are found by solving the optimization problem:

$$\widehat{\boldsymbol{b}}_\lambda = \arg\min_{\boldsymbol{b}} \left\{\sum_{i=1}^n (y_i - \langle \boldsymbol{b}, \varphi(\boldsymbol{x}_i)\rangle)^2 + \lambda \|\boldsymbol{b}\|_2^2\right\}, \tag{11}$$

which is equivalent to a KRR with kernel $\mathbf{K}_{ij} = \langle \varphi(\boldsymbol{x}_i), \varphi(\boldsymbol{x}_j)\rangle$ and optimal function found via the optimization problem:

$$\widehat{f}_\lambda = \arg\min_f \left\{\sum_{i=1}^n (y_i - f(\boldsymbol{x}_i))^2 + \lambda \|f\|_{L^2}^2\right\}. \tag{12}$$

TODO probably norm is in K
Here, as the number of neurons diverges ($N \to \infty$) it holds that the empirical kernel becomes an inner product kernel $\mathcal{K}_N(\boldsymbol{x}_1, \boldsymbol{x}_2) \to \mathcal{K}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \hbar_d\left(\frac{\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle}{d}\right)$, and we studied its risk.

We now turn to study the convergence speed of this result for $\boldsymbol{x}_i \sim \mathcal{U}nif(\mathbb{S}^{d-1}(\sqrt{d}))$ and $y_i = f_*(\boldsymbol{x}_i) + \epsilon_i$ where $f_* \in L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ and $\mathbb{E}[\epsilon_i] = \tau^2$ have equal variance for all $i \in [n]$. In particular, we will compare the following risks which are always wrt a reference function f_*:

$$\mathcal{R}_{\mathsf{NT}}(\lambda) = \mathbb{E}\left[\left(f_*(\boldsymbol{x}_{new}) - \left\langle \widehat{b}_\lambda, \varphi(\boldsymbol{x}_{new})\right\rangle\right)^2\right] \tag{13}$$

$$\mathcal{R}_{\mathsf{KRR}}(\lambda) \qquad \text{infinite width kernel} \tag{14}$$

$$\mathcal{R}^{(l)}_{\mathsf{PRR}}(\lambda) \qquad \text{Risk of RR and } l \text{ degree polynomial} \tag{15}$$

where the last has roughly d^l parameters as we discussed earlier. Another previous results was that PRR is equivalent to an l-truncated KRR which in general has kernel:

$$\mathcal{K}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{m=0}^{\infty} \lambda_m^2 \boldsymbol{Y}_m(\boldsymbol{x}_1)^\top \boldsymbol{Y}_m(\boldsymbol{x}_2), \tag{16}$$

and we take the first l terms to have equivalence. Recall that the m^{th} term in this sum is a polynomial of degree m.

Assumption 3. *To state our results, we require the following:*

(A1) ***separation:*** *for some $l \geq 2$ and big $c \gg 1$ it holds that $d^l(\log d)^c \leq n \leq \frac{d^{l+1}}{(\log d)^c}$.*
In particular, for $l = 1$, it suffices to require that $\frac{d}{c_0} \leq n \leq \frac{d^2}{(\log 2)^c}$ for some c, c_0.

(A2) ***activation niceness:*** *the function σ observes a criterion that will be made precise in the next statement (Thm. 4.2).*

Remark 4.1. *Recall that the NTK term appears in the Kernel summands. Indeed:*

$$\mathcal{K}_N(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{1}{Nd} \sum_{i=1}^{N} \langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle \, \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{x}_1 \rangle) \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{x}_2 \rangle). \tag{17}$$

In this expression, we are neglecting the terms from the second layer. If we added them, there would be an additional factor often termed Random Feature (RF) Kernel. It admits an expression:

$$\frac{1}{N} \sum_{i=1}^{N} \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}_1 \rangle \, \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}_2 \rangle), \tag{18}$$

but having rank N with the NTK part having rank $Nd \gg N$, it is neglected. We are effectively only considering the first layer expansion. In Multi-Layer Perceptrons, this term would be important, since higher order interactions take place.

{thm:Mi

Theorem 4.2 ([MZ22]). *Let σ be weakly differentiable*[11] *and satisfy for some B the inequality $|\sigma'(x)| \leq B(1+|x|)^B$. Assume further that $Nd \geq n(\log Nd)^{c_0}$ for some c_0, and define the shortcut:*

$$v_*(\sigma) \equiv \sum_{k\geq 1} \langle H_{lk}, \sigma' \rangle_{L^2}^2 (\mathcal{N}(0,1)) = \min_{p \in \mathsf{poly}(l-1)} \mathbb{E}_{G\sim\mathcal{N}(0,1)} \left[(\sigma'(G) - p(G))^2 \right] > 0 \tag{19}$$

where H_{lk} are Hermite polynomials. Then:

$$\mathcal{R}_{\mathsf{NT}}(\lambda) = \mathcal{R}_{\mathsf{KRR}}(\lambda) + O\left(\tau_+^2 + \sqrt{\frac{n(\log Nd)^{c'}}{Nd}} \right), \quad \tau_+^2 = \tau^2 \mathbb{E}\left[f_*^2\right], \tag{20}$$

with $c' \in \mathbb{R}$ depending on all previous constants.

Remark 4.3. *The quickest interpretation is as follows. For $Nd \geq n(\log Nd)^{c_0}$ and $Nd \gg n$ it holds that $\mathsf{NTK} \approx \mathsf{KRR}$, and an overparametrized Neural Tangent Kernel is basically KR, which is a similar result when compared to the one of the previous lecture* [12] *of self-induced regularization in PRR. Indeed we can state that NTK is equivalent to:*

$$\mathcal{R}_{\mathsf{KRR}}^{(l)}(\lambda + v_*(\sigma)) + O\left(\tau_+^2 \sqrt{\frac{n}{Nd}(\log Nd)^c} \right) + O\left(\tau_+^2 \sqrt{\frac{n(\log n)^c}{d^{l+1}}} \right), \quad (O_p \textit{ formally}) \tag{21}$$

where $O\left(\tau_+^2\sqrt{\frac{n}{Nd}(\log Nd)^c}\right)$ is the finite-width error and the second term tells us that the polynomial-to-kernel approximation error is small for $n \ll d^{l+1}$.

Remark 4.4. *The norm of y plays a role. Recall that*

$$\mathcal{R}(\lambda) = \mathbb{E}_{\boldsymbol{\epsilon}, \boldsymbol{X}_{new}} \left[(\widehat{f} - f_*)^2 \right] = \mathsf{BIAS} + \mathsf{VARIANCE}, \tag{22}$$

and being that on the RHS we have random quantities over which we expect the asymptotic terms are wrt to the randomness, namely they should have an explicit O_p in their form, as we are not expecting wrt their distribution. Expanding the above expression we find:

$$\mathsf{BIAS} + \mathsf{VARIANCE} = \underbrace{\mathbb{E}_{\boldsymbol{X}_{new}} \left[(f_* - \mathbb{E}_{\boldsymbol{\epsilon}} \left[\widehat{f} \right])^2 \right]}_{\propto \|f_*\|_{L^2}^2} + \underbrace{\mathbb{E}_{\boldsymbol{X}_{new}} \left[(\widehat{f} - \mathbb{E}_{\boldsymbol{\epsilon}} \left[\widehat{f} \right])^2 \right]}_{\perp\!\!\!\perp f_*, \propto \tau^2} \tag{23}$$

$$= \|f_*\|_{L^2}^2 \, B^{\mathsf{normalized}} + \tau^2 V^{\mathsf{normalized}} \tag{24}$$

Open Problem #5

The potential improvements are twofold:

- Get rid of the spurious $\log$ factors in the scaling assumption
- repeat the process or find something similar for Networks with 2 hidden layers.

[11] almost everywhere differentiable with continuous derivative

[12] see Rem. 3.12 and the discussion there

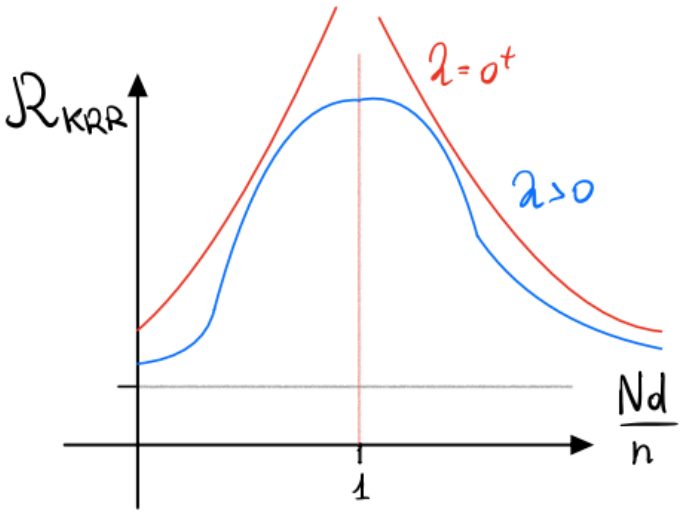

Figure 11: Kernel Ridge Regression Risk {fig:KR

Concerning the second point, we already know that with 2 hidden layers the NTK regime allows for an inner product kernel with a precise characterization, so it would suffice to describe the risk of the NTK against the risk of the KRR. In this case, the aspect ratio of the network[13] is important. Additionally, the kernel is not anymore a sum of independent terms.

In Figure 11 we can see the result of Theorem 4.2 in action. The solution f_* is not of low degree and we have a scaling for which $d^l \ll n \ll d^{l+1}$. When the plots are above the regime $(\log Nd)^c$ the regularized curve and the interpolating curve match. In particular, benign KRR overfitting transfers as a result to NTK.

4.1 Phase diagram

We now explain in detail Figure 12. The limit depends on the parameter scaling $\frac{\log n}{\log d}$, which determines the degree of the polynomial RR that is effectively performed. The lower triangular part represents the phase in which it is possible to interpolate data, the upper triangular phase does not admit this. The red phases are blocks of constant degree l, and so they identify regimes in which the risk is kept constant until the next jump. The blue phases are the result of a symmetry in behavior that was only proved for the Random Features model (see [**meiLearningInvariancesRandom2021**]). Namely, as N grows, it was shown that the RF kernel gets to the same limits of its lower triangular counterpart.

[13] i.e. the scaling of the ratio of hidden layers in the first and second layer

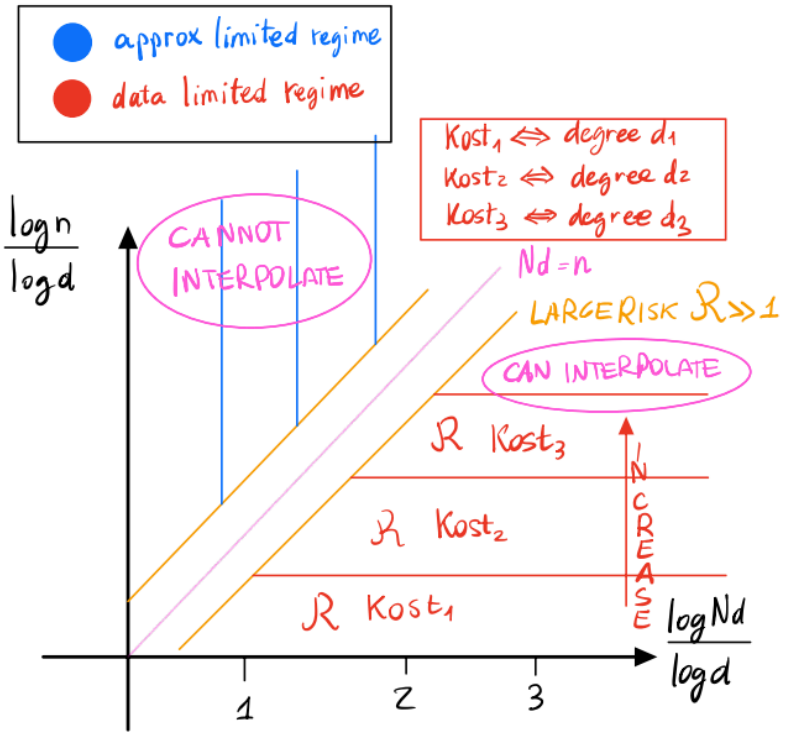

Figure 12: Phase Diagram of Kernel Regression {fig:ph

In short, the red regions represent different *data limited*[14] phases, while the blue regions are *approximation limited*[15], and we hope to operate in the data limited setting using all the parameters.

Open Problem #6

What happens above the diagonal for the Neural Tangent model?

If the distribution is known, it is possible to take the kernel directly and diagonalize it, to later do PRR on the top eigenfunctions and approximate KRR. Unfortunately, the true distribution is unknown and this is not a priori feasible for the objective we gave at the beginning.

Remark 4.5. *Any kernel is a sum of eigenfunctions. It is always possible to analyze it as an operator over the data distribution and diagonalize it. Then, a finite approximation by truncating at a regime such that $l \gg n$, where l is the degree of truncation is valid. The true problem is diagonalizing, since the operator is self-adjoint and in $\mathbb{R}^d$ where $d \gg 1$, so this is an expensive task. Even if this were feasible, the regression obtained would be still complex due to high dimensionality. Computations are not simplified.*

[14] statistical error dominates
[15] approximation error dominates

Another approach to solving this complication is as follows. Any kernel can be seen as:

$$\mathcal{K}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathbb{E}_{\boldsymbol{\vartheta}} \left[\overline{\sigma}(\boldsymbol{x}_1; \boldsymbol{\vartheta}) \overline{\sigma}(\boldsymbol{x}_2; \boldsymbol{\vartheta}) \right], \tag{25}$$

for some $\overline{\sigma}, p(\boldsymbol{\vartheta})$, where we are represting the kernel as a RF kernel for some activation function and distribution. In principle, the complexity for prediction is N^2, and one can approximate the above expression with:

$$\mathcal{K}(\boldsymbol{x}_1, \boldsymbol{x}_2) \approx \frac{1}{N} \sum_{i=1}^{N} \overline{\sigma}(\boldsymbol{x}_1; \boldsymbol{\vartheta}_i) \overline{\sigma}(\boldsymbol{x}_2; \boldsymbol{\vartheta}_i), \tag{26}$$

which is an idea by Rahimi Recht [RR07], termed **Randomized KRR Methods**.

Open Problem #7

Vaguely, the kernel works well, but how fast does a Kernel replicate ERM of NNs in RKHS?

The plot of Figure 12 however says that optimality is for $N \asymp n$, so this approach fails, but maybe the framework is valid.

Remark 4.6. *It seems that the RF model is similar to the NT model, but the RF model above is such that $Nd = pd$ at prediction time, with p controlling the generalization error. Instead, the NT model is at prediction time (matrix multiplication dominates) $Nd = p$, so p drives the generalization dynamics in both models. The error in the RF model is more complex. In some sense, at equal risk the NT model requires less complexity.*

4.2 Techniques

The rest of the lecture is focused on showcasing some techniques of the proof. In particular, we will deal with the characterization of the kernel matrix. We have that the finite width version is $\mathbf{K}_{N,n} = (\mathcal{K}_N(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j \le n}$ while the infinite width version is $\mathbf{K}_n \in \mathbb{R}^{n \times n}$.

{thm:th

Theorem 4.7. *For deterministic $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, assume that*[16]*:*

1. $\lambda_{min}(\mathbf{K}_n) \ge \frac{v(\sigma)}{2}$
2. $\mathbb{R}^{n \times d} \ni \|\mathbf{X}\|_{op} \le 2(\sqrt{n} + \sqrt{d})$*, where the bigger term is the maximum singular value*

Then w.h.p. wrt $\boldsymbol{w}_1, \dots, \boldsymbol{w}_N \sim \mathcal{U}nif(\mathbb{S}^{d-1}(\sqrt{d})$ it holds:

$$\left\| \mathbf{K}_n^{-\frac{1}{2}} \mathbf{K}_{N,n} \mathbf{K}_n^{-\frac{1}{2}} - \mathbf{I}_n \right\|_{op} \le c \sqrt{\frac{n \log Nd}{Nd}}, \tag{27}$$

where the log *is believed to be spurious.*

[16] we enforce these to have a well-defined behavior of the larged and smallest singular value. For random vectors, this is verified w.h.p.

A sensible question is reported below.

Distance Measure

Why is this projection norm the *best* way to measure the distance between $\mathbf{K}_n, \mathbf{K}_{N,n}$?

The naive distance would be $\|\mathbf{K}_{N,n} - \mathbf{K}_n\|_{op}$, the maximum eigenvalue of the difference. However, this is not nice to deal with, since the eigenvalues can get very large. As we saw previously [17], the eigenvalues of $\mathbf{K}_n$ are of order $O(n)$ with multiplicity 1 and of order $O\left(\frac{n}{d^k}\right)$ with multiplicity d^k until they become $O(1)$. The typical empirical covariance matrices do not have a bounded condition number. We immediately get a Corollary of this statement.

Corollary 4.8 (Existance of interpolator)**.** *Under the same assumptions if $Nd \geq Cn \log n$ then for all y_i there exists a $\boldsymbol{b}$ such that $f(\boldsymbol{x}_i; \boldsymbol{b}) = y_i$ for all $i \in [n]$.*

This result is understood as follows. Consider a sufficient number of parameters and the linear equations $\boldsymbol{\Phi b} = \boldsymbol{y}$ with $\mathbf{K}_{n,N} = \boldsymbol{\Phi}\boldsymbol{\Phi}^\top$. The Theorem says that the minimum eigenvalue is bounded away from zero, and one can find a full rank (invertible) matrix $\boldsymbol{\Phi}$.

Open Problem #7

Prove that $Nd = n$ is sufficient, precisely that $Nd \geq (1+\varepsilon)n$ suffices for arbitrarily small ε.

The proof of Theorem 4.7 makes use of a very important Matrix inequality which we already evoked before (see Eqn. 9).

{prop:m

Proposition 4.9 (Matrix Bernstein Inequality)**.** *Say the matrices $(\mathbf{X}_i)_{i \leq N}$ are iid and symmetric. Define the following variance and max proxyes:*

$$\sigma^2 := \left\| \sum \mathbb{E}\left[\mathbf{X}_i\right] \right\|_{op} \qquad M := \operatorname{ess\,sup} \max_{l \leq N} \|\mathbf{X}_l\|_{op}, \tag{28}$$

where ess sup *is to be intended as the essential supremum*[18]*. Then it holds:*

$$\mathbb{P}\left[\left\| \sum_{i=1}^N \mathbf{X}_i - \mathbb{E}\left[\mathbf{X}_i\right] \right\|_{op} \geq t \right] \leq 2d \exp\left\{ -\frac{t^2}{2\sigma^2 + \frac{2}{3}Mt} \right\}, \tag{29}$$

where in the exponential we recnognize an expression that is $\leq \min\left\{\frac{t^2}{\sigma^2}, \frac{t}{M}\right\}$.

The result is an asymptotic inequality:

$$\left\| \sum_{i=1}^N \mathbf{X}_i - \mathbb{E}\left[\mathbf{X}_i\right] \right\|_{op} \lesssim \sigma\sqrt{\log d} \vee M \log d \lesssim \sigma \vee M. \tag{30}$$

[17] see after Ex. 3.10 and the discussion before it for context

[18] for a rancom variable Z the essential supremum is the smallest deterministic number such that almost sure inequality holds, namely M such tht $\mathbb{P}\left[\max_l \|\mathbf{X}_l\|_{op} \leq M\right] = 1$

Remark 4.10. *The classic Bernstein inequality does not have a factor* $\log d$ *since* d*, the dimension of the matrix, is* 1 *in vectors.*

We are now ready to dive into the proof.

Proof. Of Thm. 4.7. For simplicity **drop the common lowercase** n **from now onwards**. Observe that:

$$(\mathbf{K}_N)_{ij} = \frac{1}{Nd}\sum_{l=1}^{N} \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \, \sigma'(\boldsymbol{w}_l^\top \boldsymbol{x}_1)\sigma'(\boldsymbol{w}_l^\top \boldsymbol{x}_j), \tag{31}$$

and in a slightly nicer matrix form:

$$\mathbf{K}_N = \frac{1}{Nd}\sum_{l=1}^{N} \mathbf{D}_l \mathbf{X}\mathbf{X}^\top \mathbf{D}_l \quad \mathbf{D}_l = \mathsf{diag}(\sigma'(\boldsymbol{w}_l^\top \boldsymbol{x}_i) \mid i \leq N). \tag{32}$$

Inspecting $\mathbf{K}^{-\frac{1}{2}}$, one finds:

$$\mathbf{K}^{-\frac{1}{2}}\mathbf{K}_N\mathbf{K}^{-\frac{1}{2}} = \frac{1}{N}\sum_{l=1}^{N} \frac{1}{d}\mathbf{K}^{-\frac{1}{2}}\mathbf{D}_l\mathbf{X}\mathbf{X}^\top\mathbf{D}_l\mathbf{K}^{-\frac{1}{2}} = \frac{1}{N}\sum_{l=1}^{N}\mathbf{H}_l, \tag{33}$$

where each $\mathbf{H}_l$ is iid since we are fixing $\boldsymbol{x}$ and the weights $\boldsymbol{w}_i$ are iid. We would like to bound this. In expectation, it holds:

$$\mathbb{E}\left[\mathbf{H}_l\right] = \mathbf{K}^{-\frac{1}{2}}\mathbb{E}\left[\mathbf{D}_l\mathbf{X}\mathbf{X}^\top\mathbf{D}_l\right]\mathbf{K}^{-\frac{1}{2}} = \mathbf{K}^{-\frac{1}{2}}\mathbf{K}\mathbf{K}^{-\frac{1}{2}} = \mathbf{I}_n \tag{34}$$ {eqn:H1

Remark 4.11. *For NNs with up to two layers, the product* $\mathbf{X}\mathbf{X}^\top$ *is made of independent vectors.*

Further References

A good result was found by Yizhe Zhu [ALS19]for $Nd \geq n^2$ or something like this. Also Rudy Ahlswede and Winther [AW01] found results starting from Quantum Information Theory problems. Matrix concentration has relevant statements in the work of David Gross [Gro11]. Eventually, Tropp proved this statement, and later provided a very nice introduction to RMT [Tro12b; Tro12a]. A high degree bound for the previous lecture result (the $\geq l$ part in Eqn. 8 and the surrounding discussions) was later found by Oliveira and generalized by Tropp.

In general, we would like to lose the $\log d$ factor. Being a sum of independent variables, we have a crude bound:

$$\left\|\mathbf{K}^{-\frac{1}{2}}\mathbf{K}_N\mathbf{K}^{-\frac{1}{2}} - \mathbf{I}_n\right\|_{op} \lesssim \frac{1}{N}\left[M\log n \vee \sigma\sqrt{\log n}\right], \tag{35}$$

which is an non-explicit application of Prop. 4.9. To close the asymptotic bound, we seek nice expressions of M, σ. Notice that right away:

$$\|\mathbf{H}_l\|_{op} \lesssim \frac{1}{d}\frac{1}{\sigma_{min}(\mathbf{K})}\|\mathbf{D}_l\|_{op}^2\|\mathbf{X}\|_{op}^2. \tag{36}$$

Assuming for simplicity[19] that $|\sigma'| < \infty$ and $n \geq cd$ this becomes:

$$\|\mathbf{H}_l\|_{op} \lesssim \frac{1}{d}(\sqrt{n} + \sqrt{d})^2 \lesssim nd = M. \tag{37}$$

While for the variance by the iid assumption:

$$\sigma^2 = \left\| \sum_{l=1}^{N} \mathbb{E}\left[\mathbf{H}_l^2\right] \right\|_{op} = N \left\| \mathbb{E}\left[\mathbf{H}_1^2\right] \right\|_{op}, \tag{38}$$

and in particular:

$$\mathbb{E}\left[\mathbf{H}_1^2\right] = \frac{1}{d^2}\mathbb{E}_{\boldsymbol{w}_1}\left[\mathbf{K}^{-\frac{1}{2}}\mathbf{D}_1\mathbf{X}\mathbf{X}^\top\mathbf{D}_1\mathbf{K}^{-\frac{1}{2}}\mathbf{K}^{-\frac{1}{2}}\mathbf{D}_1\mathbf{X}\mathbf{X}^\top\mathbf{D}_1\mathbf{K}^{-\frac{1}{2}}\right]. \tag{39}$$

Then, by the bounded derivative of the activation assumption and the minimum singular value of $\mathbf{K}$ being positive we directly get that $\|\mathbf{D}_1\mathbf{K}^{-1}\mathbf{D}_1\|_{op} \leq C$ in the middle.

Remark 4.12. *We would like to use the general inequality:*

$$\left\|\mathbf{A}\mathbf{M}^\top\mathbf{A}\right\|_{op} \leq \left\|\mathbf{A}\mathbf{A}^\top\right\|_{op} \|\mathbf{M}\|_{op}, \tag{40}$$

but this turns out to be slightly incorrect. The right statement is in the line of $\mathbf{A}\mathbf{M}\mathbf{A}^\top \preceq \mathbf{A}\mathbf{A}^\top$*, i.e. for any vector* $\boldsymbol{v}$ *we have* $\left\langle \boldsymbol{v}, \mathbf{A}\mathbf{M}\mathbf{A}^\top\boldsymbol{v}\right\rangle \leq \left\langle \boldsymbol{v}, \mathbf{A}\mathbf{A}^\top\boldsymbol{v}\right\rangle$*, which holds whenever* $\mathbf{M} \preceq c\mathbf{I}_n$ *for some* $c \in \mathbb{R}$ *deterministically. The proof is straightforward and requires inspecting* $\boldsymbol{v}' = \mathbf{A}^\top\boldsymbol{v}$*. Being that the inequality is a.s., one gets:*

$$\left\langle \boldsymbol{v}, \mathbb{E}\left[\mathbf{A}\mathbf{M}\mathbf{A}^\top\right]\boldsymbol{v}\right\rangle \leq \left\langle \boldsymbol{v}, \mathbb{E}\left[\mathbf{A}\mathbf{A}^\top\right]\boldsymbol{v}\right\rangle. \tag{41}$$

we reach the following asymptotic variance bound:

$$\mathbb{E}\left[\mathbf{H}_1^2\right] \lesssim \frac{1}{d^2}\mathbf{K}^{-\frac{1}{2}}\mathbb{E}_{\boldsymbol{w}_1}\left[\mathbf{D}_1\mathbf{X}\underbrace{\mathbf{X}^\top\mathbf{X}}_{\lesssim n}\mathbf{X}^\top\mathbf{D}_1\right]\mathbf{K}^{-\frac{1}{2}} \lesssim \frac{n}{d^2}\mathbf{K}^{-\frac{1}{2}}\mathbb{E}_{\boldsymbol{w}_1}\left[\mathbf{D}_1\mathbf{X}\mathbf{X}^\top\mathbf{D}_1\right]\mathbf{K}^{-\frac{1}{2}} \tag{42}$$

So far, we are only deriving crude asymptotic bound, but these expressions highlight that we can use Bernstein's inequality (Prop. 4.9) also for $\mathbf{H}_1$, which by definition and the fact that it is identity in expectation (Eqn. 34) we get:

$$\mathbb{E}\left[\mathbf{H}_1\right] \lesssim \frac{n}{d} \implies \mathbb{E}\left[\mathbf{H}_1^2\right] \lesssim \frac{n}{d^2}d \lesssim \frac{Nn}{d}. \tag{43}$$

Eventually, w.h.p. we obtain:

$$\left\|\mathbf{K}^{-\frac{1}{2}}\mathbf{K}_N\mathbf{K}^{-\frac{1}{2}} - \mathbf{I}_n\right\|_{op} \lesssim \frac{n}{Nd}\log n \vee \sqrt{\frac{n}{Nd}\log n} \lesssim \sqrt{\frac{n}{Nd}\log n}, \tag{44}$$

in the worst case. □

[19] this is not in the paper, where there is a more general statement. It should be in [MZ22].

Remark 4.13. *Recall that the inequality* $\lambda_{min}(\mathbf{K}) \leq \frac{v(\sigma)}{2}$ *was proved w.h.p. in Theorem 4.7. Namely, there exists a set* $\mathscr{K}$ *such that* $\mathbb{P}[\boldsymbol{X} \in \mathscr{K}] = 1$.

In the previous lecture, we also proved that $\frac{1}{n}\boldsymbol{Y}_m^\top \boldsymbol{Y}_m \approx \mathbf{I}_n$, the smooth part of the kernel approximation. The steps for the result are similar to those above. Instead, for the high frequency part, one must resort to the martingale version of the inequality.

End of Lecture 4

5 Limitations

Now that we saw the linear theory of 2-layer NNs, we can check its uses and connections to other results.

It is indeed possible to derive easily separation results between kernels and Neural Networks. With kernels, we mean both NT and other general forms. The simplest example is very instructive and will be detailed below.

5.1 Ridge functions Learning

{ex:ric

Example 5.1 (Ridge Function Kernel learning)**.** *Let* $\boldsymbol{x}_i \sim \mathcal{U}nif(\mathbb{S}^{d-1}(\sqrt{d}))$*, with model* $y_i = f_*(\boldsymbol{x}_i) + \epsilon_i$*, where the true function is a so-called Ridge function, i.e. a single neuron:*

$$f_*(\boldsymbol{x}) = \sigma(\langle \boldsymbol{u}_*, \boldsymbol{x} \rangle). \tag{45}$$

Recall that $\boldsymbol{x}_i$ *sampled this way all have square norm* d *and entries with variance of order constant. The risk of this model, being an inner product kernel, is available. Namely, for* $n \geq d^{l+1-\epsilon}$*, where* $\epsilon > 0$ *is arbitrarily small, it holds:*

$$\mathcal{R}(f_*; \lambda) = \|\mathcal{P}_{>l} f_*\|_{L^2}^2 + o(1), \tag{46}$$

which is true since for 2-layer NNs the condition is $Nd \gg n$*. If we decompose* f_* *into spherical harmonics, we need invariance wrt* $\boldsymbol{u}_*$ *to have a good result. We claim that for any degree there is a unique polynomial that achieves this. Namely:*

$$\forall m \leq l \exists! q_m : \|q_m\|^2 = 1,\ q_m \in V_m = \mathsf{span}(\boldsymbol{Y}_{m,j})_{j \leq D(m)},\ s.t.\ q_m(\boldsymbol{x}) = q_m(S\boldsymbol{x}) \tag{47}$$

for any rotation S *(i.e. objects such that* $S\boldsymbol{u}_* = \boldsymbol{u}_*$*). Such polynomial is obtained in uniqueness and existance as the average of spherical harmonics over the rotation. Then:*

$$q_m(\boldsymbol{x}) = P_{\mathsf{deg}=m,d}(\langle \boldsymbol{u}_*, \boldsymbol{x} \rangle), \tag{48}$$

by invariance under rotation on the spheres, with $\|\boldsymbol{u}_*\|^2 = 1$*. We get wlog that* $\langle q_m, q_{m'} \rangle = \delta_{m,m'}$*. This implies that the integral representation is a one dimensional integral:*

$$\int P_{\mathsf{deg}=m,d}(\boldsymbol{z}) P_{\mathsf{deg}=m',d}(\boldsymbol{z}) \tau_d(\boldsymbol{dz}) = \delta_{m,m'}, \tag{49}$$

{eqn:ir

where τ_d is the probability measure of $\langle \boldsymbol{u}_*, \boldsymbol{x} \rangle$ for $\boldsymbol{x}$ distributed uniformly on the sphere. Substituting Eqn. 49 into the Dirac of inner products of $q_m, q_{m'}$ we simplified the general expression for the Dirac-inner product:

$$\int q_m(\boldsymbol{x}) q_{m'}(\boldsymbol{x}) \mu(\boldsymbol{dx}), \quad \mu = \mathcal{U}nif(\mathbb{S}^{d-1}(\sqrt{d})), \tag{50}$$

which would have been an integral on the sphere. We now only care about low-dimensional projections of high-dimensional spheres, which are approximately Gaussian, so that $\tau_d \overset{d\to\infty}{\to} \gamma(d) = \mathcal{N}(0,1)$. Then if $P_{m,d} \overset{d\to\infty}{\to} P_{m,\infty}$, we have that:

$$\int P_{m,\infty}(\boldsymbol{x}) P_{m',\infty}(\boldsymbol{z}) \gamma(\boldsymbol{dz}) = \delta_{m,m'}, \tag{51}$$

which are orthonormal polynomials wrt the Gaussian measure, commonly known as Hermite Polynomials. Seen in terms of the target function, we obtain the expression:

$$f_*(\boldsymbol{x}) = \sum_{l=0}^{\infty} \widehat{\sigma}_{l,d} P_{l,d}(\langle \boldsymbol{u}_*, \boldsymbol{x} \rangle), \quad \widehat{\sigma}_{l,\infty} = \mathbb{E}\left[\sigma(G) \mathcal{H}_{l,m}(G)\right], \tag{52}$$

where the latter is seen as a projection onto Hermite polynomials. Going to the risk, we get the expression:

$$\|P_{>l} f_*\|_{L^2(\mathbb{S}^{d-1}(\sqrt{d}))} \overset{d\to\infty}{\to} \sum_{m=l+1}^{\infty} \langle \sigma, \mathcal{H}_{l,m} \rangle > 0, \tag{53}$$

assumed to be non-zero to avoid trivialities, where by trivialities we mean that σ is <u>not</u> a degree l polynomial.

Example 5.2. *Assume σ is not polynomial and for all $l \in \mathbb{N}$ the conditions $n \leq d^{l+1}$ and having an inner product kernel hold. Then any risk (NT, KRR, PRR) behaves as $\mathcal{R}(f_*, \lambda) = c_l + o(1)$. As a consequence, we need a nonpolynomial number of samples to draw from a simple Ridge function.*

Some early versions of this result appeared in [YS22]. While this is a bit disappointing since we need too many parameters, on the other hand the dependence is only on d in principle. With the right inductive bias, it should be possible to learn in $O(d)$ samples. A 1 neuron NN is able to do this task. Let the empirical risk be:

$$\widehat{\mathcal{R}}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \sigma(\boldsymbol{\theta}^\top \boldsymbol{x}_i))^2, \quad \widehat{\boldsymbol{\theta}} = \arg\min_{\mathbb{R}^d} \widehat{\mathcal{R}}_n(\boldsymbol{\theta}), \tag{54}$$

then we can state the following, which is a consequence of a more general Theorem found in [MBM17].

{prop:E

Proposition 5.3 ([MBM17]). *If $\sigma' \geq C^{-1} > 0$ and $\|\sigma'''\| < l$ with high probability Gradient Descent on the empirical risk will converge globally to $\widehat{\boldsymbol{\theta}}_n$ such that:*

$$\left\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{u}_*\right\|^2 \leq C' \sqrt{\frac{d \log n}{n}}. \tag{55}$$

In particular, GD has a unique fixed point at $\boldsymbol{u}_*$ and the gradient flow attains it if $n \gg d$ (modulo a logarithmic factor which is spurious).

Combining Example 5.1 and Proposition 5.3, we get a reasonable separation of Kernels and Neural Networks in terms of performance.

Open Problem #8

Generalize for a model $y_i = \varphi(\boldsymbol{u}_*^\top \boldsymbol{x}_i) + \epsilon_i$. Notice that for high degree $\varphi(\cdot)$ the Hermite polynomials do not generalize. However, from the Theorem conditions and proof, it appears that this could not matter much, and one just needs to inspect the population error at initialization.

Again concerning Problem #8, monotone functions $\varphi(\cdot)$ are such that an easy trick makes them learnable. Without restricting ourselves to GD, one could think of an alignment process as follows. Initialize $\widehat{\boldsymbol{\theta}}_0 = \frac{1}{n}\sum_{i=1}^n y_i \boldsymbol{x}_i$ and for $n \gg d \log d$ it will be that

$$\widehat{\boldsymbol{\theta}}_0 \approx \mathbb{E}\left[Y \boldsymbol{X}\right] = \mathbb{E}\left[\varphi(\langle u_*, \boldsymbol{X}\rangle)\boldsymbol{X}\right] = c\boldsymbol{u}_*. \tag{56}$$

where the last passage holds if $\boldsymbol{X}$ is isotropic in distribution, while the initialization is essentially a gradient. The result is already close to the target vector, a proper normalization would then just require to localize the problem again at the next step and iterate.

There are countless other separation examples, but this is interesting. An isotropic (inner product) kernel is not good with low dimensional projections of data such as images, which live on a smaller manifold.

5.2 Another Separation example and the importance of scaling

We stick to $f_*(\boldsymbol{x}) = \sigma(\langle \boldsymbol{u}_*, \boldsymbol{x}\rangle)$ and a 2 layer function $f(\boldsymbol{x}; \boldsymbol{w}) = \frac{1}{\sqrt{N}}\sum_{i=1}^N a_i \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle)$ with $a_i = \pm 1$ not trained. We already know the risk looks like Figure 13 for $n \gg d, n \ll d^C$. Theorem **TODO previous one** gives us the result at $N = 1$, while we know that for $N > \frac{n}{d}$ (modulo logarithmic factors) the NTK risk converges to $\|P_{>l} f_*\|^2$. If $Nd > Cn^2$, results in [OS19b; BMR21] guarantee that NNs Risk is approximately like the NTK risk. In lecture 3-4 we also proved that the NTK risk is approximately $\|P_{>l} f_*\|^2 = \mathcal{R}_{\mathsf{PRR}}$, while in this lecture we proved under which conditions it is strictly positive. At arbitrary number of neurons N, it could instead be that the risk behaves non monotonically, and we cannot say anything. A natural question arises.

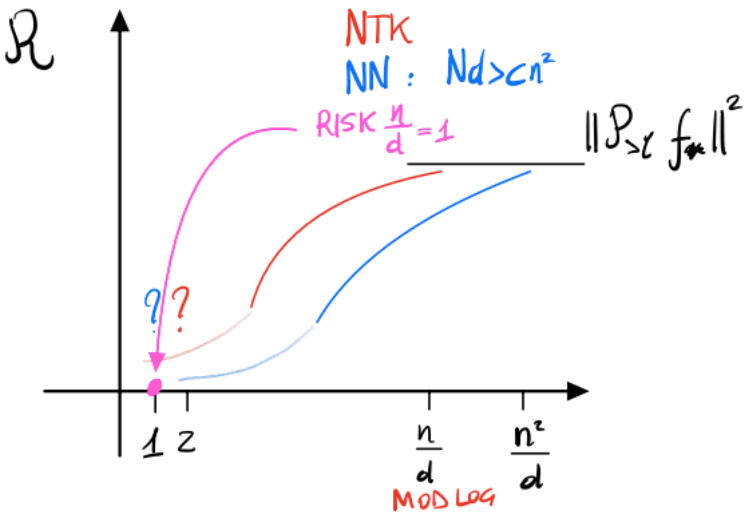

Figure 13: NTK and NN risk {fig:NT

Width

Is this a problem of width?

The quickest answer is no, since the $\frac{1}{\sqrt{N}}$ scaling is inherently wrong. The right function should be:

$$f(\boldsymbol{x};\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle), \tag{57}$$

which we will term **Mean-Field scaling**.

We now present a simple trick to get zero error. There is one initialization for which GD attains zero risk. Without training, a choice $a_i \equiv a, \boldsymbol{w}_i \equiv \boldsymbol{w}$, where $\boldsymbol{w}$ possibly is $\mathbf{0}$ at initialization is interesting. Ispecting the empirical risk:

$$\widehat{\mathcal{R}}_n(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n}(\sigma(\langle \boldsymbol{u}_*, \boldsymbol{x}_i\rangle) - \sigma(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle))^2, \tag{58}$$

we notice that the weights will move together by symmetry for the whole dynamics. Then $\widehat{\mathcal{R}}(\mathbf{W}) \overset{n,d\to\infty}{\to} 0$. One might then wonder what happens with $\boldsymbol{w}_i$ sampled randomly in the GD/SGD training dynamics. The MF scaling parired with $\boldsymbol{w}_i^{(t=0)} \sim \mathcal{N}(0, \mathbf{I}_d)$ and any $\boldsymbol{a}$ taken iid is an average over $\boldsymbol{a}, \mathbf{W}$. The function will then be:

$$f(\boldsymbol{x}) = \int_{\mathbb{R}\times\mathbb{R}^d} a\sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\widehat{\rho}_0(\mathrm{d}a, \mathrm{d}\boldsymbol{w}), \tag{59}$$

where $\widehat{\rho}_0$ is the empirical distribution over $(\boldsymbol{a}, \mathbf{W})$, namely:

$$\widehat{\rho}_0 = \frac{1}{N}\sum_{i=1}^{n} \delta_{(a_i, \boldsymbol{w}_i)}. \tag{60}$$

We recognize that at initialization $(a_i, \boldsymbol{w}_i) \sim \rho_0$ and in the Mean Field limit $\widehat{\rho}_0 \overset{N\to\infty}{\to} \rho_0$. The question now becomes if $\widehat{\rho}_t$ converges to a measure ρ_t as $N \to \infty$. We claim that the answer is yes and will provide a heuristic argument for it. For any finite N online SGD on the parameter vector $\theta_i^{(t)} = (a_i, \boldsymbol{w}_i; t) \in \mathbb{R}^{d+1}$ where $t = i \in [n]$ is such that in the limit of small learning rate $\gamma \to 0^+$ dynamics converge to the Gradient Flow on the Population Error. Namely:

$$\theta_i^{(t)} \overset{\gamma\to 0^+}{\to} -\nabla_\theta \mathcal{R}_N(\theta) = -\nabla_\theta \mathbb{E}_{(Y,\boldsymbol{X})}\left[\frac{1}{2}\left(Y - \frac{1}{N}\sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{w}_i, \boldsymbol{X}\rangle)\right)^2\right]. \tag{61}$$

In reality, we will obviously have a noisy term added to the dynamics, that accounts for the fact that we can only perform finite stepsize updates. We express this in general as:

$$\theta_i^{(t)} = -\nabla_\theta \mathcal{R}_N(\theta) + \boldsymbol{\epsilon}. \tag{62}$$

This notion is termed convergence to the *fluid limit* [Kur71]. In Statistical Physics, it was explored by Saad and Solla [SS95a; SS95b; SS95c]. In other words, consider the update of SGD at the k^{th} step, denoted as $\theta_i^{\mathsf{SGD},(k)}$ and the continuous GF limit $\theta_i^{(t)}$. Under suitable assumptions, one can establish that with high probability:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\theta_i^{\mathsf{SGD},(\lfloor \frac{t}{\gamma} \rfloor)} - \theta_i^{(t)}\right\|^2 \leq \sqrt{\gamma d}, \tag{63}$$

and for small learning rate γ the dynamics are essentially equivalent. The risk can then be expressed as:

$$\mathcal{R}_N(\theta) = \mathsf{cst} + \frac{1}{N}\sum_{i=1}^{N} V(\theta_i) + \frac{1}{2N^2}\sum_{i=1}^{N} U(\theta_i, \theta_j) \tag{64}$$

$$V(\theta_i) = -\mathbb{E}\left[f_*(\boldsymbol{x}) a_i \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle)\right] \tag{65}$$

$$U(\theta_i, \theta_j) = \mathbb{E}\left[a_i \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle) a_j \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle)\right], \tag{66}$$

where we recognize that V, U are respectively a correlation potential and a pairwise interaction term. The GF limit for finite N can be written concisely in terms of these two objects as:

$$\dot{\theta}_i^{(t)} = -\nabla_{\theta_i}\left(V(\theta_i^{(t)}) + \int U(\theta_i^{(t)}, \theta)\widehat{\rho}_t^N(\mathrm{d}\theta)\right), \tag{67}$$

with the interpretation that $\theta_i^{(t)}$ is a particle, the potential V encodes the tendency to align, and the pairwise interaction encodes repulsion among neuron particles. As before $\widehat{\rho}_t^N$ is the empirical

distribution of particles. It is intuitive to think that in the Mean Field limit $N \to \infty$, we can write the one particle/neuron potential as:

$$\psi(\theta, \varphi_t) \coloneqq V(\theta) + \int U(\theta, \theta')\varphi(\mathrm{d}\theta'), \quad \widehat{\varphi}_t^N \overset{N\to\infty}{\to} \varphi_t. \tag{68}$$

Moving to a Physics-based analysis, we might wonder what is the evolution of φ_t in the limit. The fact that neurons move locally imposes a local conservation constraint, known in literature as continuity equation:

$$\partial_t \varphi_t - \nabla_\theta \cdot J_t(\theta) = 0. \tag{69}$$

We briefly explain it as follows. $\nabla_\theta \cdot J_t(\theta)$ is a divergence term, for a current $J_t(\theta)$ that is yet to find, where the current is roughly a density of particles times a speed. Having a nice form of the potential $\psi(\theta; \varphi_t)$, we express it simply as:

$$J_t(\theta) = \varphi_t(\theta)(-\nabla_\theta \psi(\theta; \varphi_t)). \tag{70}$$

Then, the PDE reads:

$$\partial_t \varphi_t + \nabla_\theta \cdot (\varphi_t \nabla_\theta \psi(\theta; \varphi_t)), \tag{71}$$

known in literature as Mckean-Vlosov Equation, found initially by Debrushin, with many Optimal Transport applications. In particular, it brings nice interpretations as a gradient flow on probability measures with the Wasserstein metric.

Theorem 5.4. *Under regularity assumptions on σ, f_*, V, U (see below for a Remark) for $k = \lfloor \frac{t}{n} \rfloor$ we have that:*

$$\sup_{t \le T} \left| \mathcal{R}_N(\theta_i^{\mathsf{SGD},k}) - \mathcal{R}_\infty(\varphi_t) \right| \le C\sqrt{d} \left(\sqrt{\gamma} + \frac{1}{\sqrt{N}} \right) e^{-ct^3}. \tag{72}$$

Namely, for any finite time the absolute distance goes to zero for N or γ small.

Remark 5.5. *The regularity assumptions can be thought of as follows. Fix $\boldsymbol{a}$, then:*

$$V(\boldsymbol{w}) = -\mathbb{E}\left[f_*(\boldsymbol{x}) \sigma(\langle \boldsymbol{w}, \boldsymbol{x} \rangle) \right], \quad \nabla_{\boldsymbol{w}} V(\boldsymbol{w}) = -\mathbb{E}\left[f_*(\boldsymbol{x}) \sigma'(\langle \boldsymbol{w}, \boldsymbol{x} \rangle) \boldsymbol{x} \right], \tag{73}$$

is required to be bounded and independent of dimension. In our example for $\boldsymbol{X}$ Gaussian integration by parts gives:

$$\mathbb{E}\left[\varphi(\boldsymbol{u}_*^\top \boldsymbol{x}) \sigma'(\langle \boldsymbol{w}, \boldsymbol{x} \rangle) \boldsymbol{x} \right] = C < \infty. \tag{74}$$

The above Theorem is particularly useful when one takes $t \in O(1)$ so that $\mathcal{R}_\infty(\varphi_T) < \epsilon$, or $e^{-cT^3} \in O(1)$ so that one can take $N \gg d$ and $\gamma \ll \frac{1}{d}$. However, can we solve this PDE? In general this is hard, but it is feasible numerically and when there are lots of symmetries also analytically in some cases. To give an example, for a true function $f_*(\boldsymbol{x}) = \varphi(\langle \boldsymbol{u}_*, \boldsymbol{x} \rangle)$ with $\boldsymbol{x}$ isotropic (say uniform on a sphere) the dimension of the PDE in the large width limit is determined by $\theta_i = (a_i, \boldsymbol{w}_i) \in \mathbb{R}^{d+1}$ but symmetry & isotropy make ρ_t invariant under rotations of $\boldsymbol{u}_*$. At finite width, the symmetry is broken. Essentially, the PDE as $N \to \infty$ is dependent on

$$\widetilde{\theta}_i = (a_i, s, r) \in \mathbb{R}^3 \quad s = \langle \boldsymbol{w}_i, u_* \rangle, \quad r = \left\| \boldsymbol{w}_i^\perp \right\|, \tag{75}$$

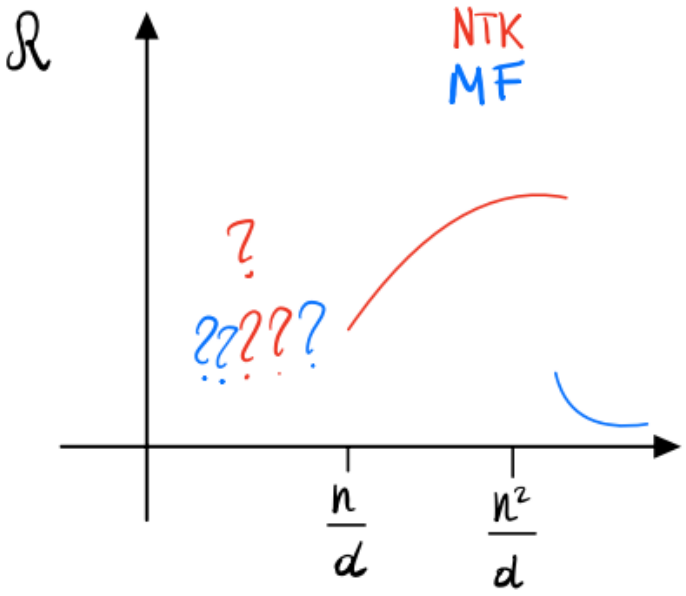

Figure 14: Mean Field NTK Risk {fig:MF

which is a considerable dimensionality reduction. Numerically, this is solvable, but it could still present analytical hurdles. For an initialization $\varphi_t(a, s, r)$, with d appearing as a parameter inside, $s \sim \mathcal{N}(0, \frac{1}{\sqrt{d}})$ and r centered at 1 with the same variance it is easy to prove the following statement.

Fact 5.6. *For all $\epsilon > 0$ there exists $T \equiv T(\epsilon, d)$ such that $\mathcal{R}_\infty(\varphi_T) \leq \epsilon$. In other words, fixing the dimension the risk can be made arbitrarily small depending on the variance $\frac{1}{d}$ induced by d.*

Open Problem #10

Prove that $T(\epsilon, \alpha^2) \equiv T(\epsilon)$ where $\alpha = \frac{1}{\sqrt{d}}$. Namely, prove that the convergence time for arbitrarily small risk is not dependent on the data dimension. This would shed light to the fact that at the $\alpha \to \infty$ limit, where the initialization is singular, the PDE methods do not apply. Numerical evidence appears to be in disaccordance with this.

What we eventually know is summarized in Figure 14. The Theorem above briefly says that $N \gg d$ matters. With one pass SGD and $n = \lfloor \frac{T}{\gamma} \rfloor$, $T \in O(1), \gamma \in O\left(\frac{1}{d}\right)$, setting a risk target ϵ, and recalling $T \equiv T(\epsilon, d)$ the required sample size is $n = cd$. In simple terms, $O(d)$ samples suffice for learning a Ridge function.

References

[ALS19] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. "A Convergence Theory for Deep Learning via Over-Parameterization". In: *Proceedings of the 36th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, May 24, 2019, pp. 242–252. URL: https://proceedings.mlr.press/v97/allen-zhu19a.html (visited on 08/05/2023).

[AS17] Madhu S. Advani and Andrew M. Saxe. *High-Dimensional Dynamics of Generalization Error in Neural Networks*. Oct. 10, 2017. arXiv: 1710.03667 [physics, q-bio, stat]. URL: http://arxiv.org/abs/1710.03667 (visited on 09/12/2023). preprint.

[AW01] R. Ahlswede and A. Winter. *Strong Converse for Identification via Quantum Channels*. Oct. 22, 2001. DOI: 10.48550/arXiv.quant-ph/0012127. arXiv: quant-ph/0012127. URL: http://arxiv.org/abs/quant-ph/0012127 (visited on 09/07/2023). preprint.

[Bar+20] Peter L. Bartlett et al. "Benign Overfitting in Linear Regression". In: *Proceedings of the National Academy of Sciences* 117.48 (Dec. 2020), pp. 30063–30070. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1907378117. arXiv: 1906.11300 [cs, math, stat]. URL: http://arxiv.org/abs/1906.11300 (visited on 08/19/2023).

[Bis95] Chris M. Bishop. "Training with Noise Is Equivalent to Tikhonov Regularization". In: *Neural Computation* 7.1 (Jan. 1995), pp. 108–116. ISSN: 0899-7667, 1530-888X. DOI: 10.1162/neco.1995.7.1.108. URL: https://direct.mit.edu/neco/article/7/1/108-116/5828 (visited on 08/19/2023).

[BMR21] Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. *Deep Learning: A Statistical Viewpoint*. Mar. 16, 2021. arXiv: 2103.09177 [cs, math, stat]. URL: http://arxiv.org/abs/2103.09177 (visited on 08/05/2023). preprint.

[COB20] Lenaic Chizat, Edouard Oyallon, and Francis Bach. *On Lazy Training in Differentiable Programming*. Jan. 7, 2020. DOI: 10.48550/arXiv.1812.07956. arXiv: 1812.07956 [cs, math]. URL: http://arxiv.org/abs/1812.07956 (visited on 11/21/2022). preprint.

[Du+19] Simon S. Du et al. *Gradient Descent Finds Global Minima of Deep Neural Networks*. May 28, 2019. arXiv: 1811.03804 [cs, math, stat]. URL: http://arxiv.org/abs/1811.03804 (visited on 08/05/2023). preprint.

[Gho+19] Behrooz Ghorbani et al. "Limitations of Lazy Training of Two-layers Neural Network". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/hash/c133fb1bb634af68c5088f3438848bfd-Abstract.html (visited on 08/19/2023).

[Gho+20] Behrooz Ghorbani et al. “When Do Neural Networks Outperform Kernel Methods?” In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 14820–14830. URL: https://proceedings.neurips.cc/paper/2020/hash/a9df2255ad642b923d95503b9a7958d8-Abstract.html (visited on 08/19/2023).

[Gho+21] Behrooz Ghorbani et al. “Linearized Two-Layers Neural Networks in High Dimension”. In: *The Annals of Statistics* 49.2 (2021), pp. 1029–1054.

[Gro11] David Gross. “Recovering Low-Rank Matrices from Few Coefficients in Any Basis”. In: *IEEE Transactions on Information Theory* 57.3 (Mar. 2011), pp. 1548–1566. ISSN: 0018-9448, 1557-9654. DOI: 10.1109/TIT.2011.2104999. arXiv: 0910.1879 [quant-ph]. URL: http://arxiv.org/abs/0910.1879 (visited on 09/07/2023).

[Has+20] Trevor Hastie et al. *Surprises in High-Dimensional Ridgeless Least Squares Interpolation*. Dec. 7, 2020. DOI: 10.48550/arXiv.1903.08560. arXiv: 1903.08560 [cs, math, stat]. URL: http://arxiv.org/abs/1903.08560 (visited on 08/19/2023). preprint.

[KLS20] Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. *Optimal Ridge Penalty for Real-World High-Dimensional Data Can Be Zero or Negative Due to the Implicit Ridge Regularization*. Apr. 9, 2020. DOI: 10.48550/arXiv.1805.10939. arXiv: 1805.10939 [math, stat]. URL: http://arxiv.org/abs/1805.10939 (visited on 08/19/2023). preprint.

[Kur71] T. G. Kurtz. “Limit Theorems for Sequences of Jump Markov Processes Approximating Ordinary Differential Processes”. In: *Journal of Applied Probability* 8.2 (June 1971), pp. 344–356. ISSN: 0021-9002, 1475-6072. DOI: 10.2307/3211904. URL: https://www.cambridge.org/core/product/identifier/S002190020003535X/type/journal_article (visited on 09/04/2023).

[KY16] Antti Knowles and Jun Yin. *Anisotropic Local Laws for Random Matrices*. Aug. 4, 2016. DOI: 10.48550/arXiv.1410.3516. arXiv: 1410.3516 [math-ph]. URL: http://arxiv.org/abs/1410.3516 (visited on 08/19/2023). preprint.

[LY23] Yue M. Lu and Horng-Tzer Yau. *An Equivalence Principle for the Spectrum of Random Inner-Product Kernel Matrices with Polynomial Scalings*. May 5, 2023. DOI: 10.48550/arXiv.2205.06308. arXiv: 2205.06308 [math, stat]. URL: http://arxiv.org/abs/2205.06308 (visited on 08/19/2023). preprint.

[LZB21] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. *On the Linearity of Large Non-Linear Models: When and Why the Tangent Kernel Is Constant*. Feb. 19, 2021. DOI: 10.48550/arXiv.2010.01092. arXiv: 2010.01092 [cs, stat]. URL: http://arxiv.org/abs/2010.01092 (visited on 08/05/2023). preprint.

[MBM17] Song Mei, Yu Bai, and Andrea Montanari. *The Landscape of Empirical Risk for Nonconvex Losses*. Jan. 14, 2017. DOI: 10.48550/arXiv.1607.06534. arXiv: 1607.06534 [stat]. URL: http://arxiv.org/abs/1607.06534 (visited on 09/07/2023). preprint.

[Mis22] Theodor Misiakiewicz. *Spectrum of Inner-Product Kernel Matrices in the Polynomial Regime and Multiple Descent Phenomenon in Kernel Ridge Regression*. Apr. 21, 2022. arXiv: 2204.10425 [math, stat]. URL: http://arxiv.org/abs/2204.10425 (visited on 08/19/2023). preprint.

[MM23] Theodor Misiakiewicz and Andrea Montanari. *Six Lectures on Linearized Neural Networks*. Aug. 25, 2023. DOI: 10.48550/arXiv.2308.13431. arXiv: 2308.13431 [cs, math, stat]. URL: http://arxiv.org/abs/2308.13431 (visited on 08/30/2023). preprint.

[MMM21] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. *Generalization Error of Random Features and Kernel Methods: Hypercontractivity and Kernel Matrix Concentration*. Jan. 26, 2021. arXiv: 2101.10588 [math, stat]. URL: http://arxiv.org/abs/2101.10588 (visited on 08/19/2023). preprint.

[MZ22] Andrea Montanari and Yiqiao Zhong. *The Interpolation Phase Transition in Neural Networks: Memorization and Generalization under Lazy Training*. June 8, 2022. DOI: 10.48550/arXiv.2007.12826. arXiv: 2007.12826 [cs, math, stat]. URL: http://arxiv.org/abs/2007.12826 (visited on 09/07/2023). preprint.

[OS19a] Samet Oymak and Mahdi Soltanolkotabi. "Overparameterized Nonlinear Learning: Gradient Descent Takes the Shortest Path?" In: *Proceedings of the 36th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, May 24, 2019, pp. 4951–4960. URL: https://proceedings.mlr.press/v97/oymak19a.html (visited on 08/05/2023).

[OS19b] Samet Oymak and Mahdi Soltanolkotabi. *Towards Moderate Overparameterization: Global Convergence Guarantees for Training Shallow Neural Networks*. Feb. 12, 2019. DOI: 10.48550/arXiv.1902.04674. arXiv: 1902.04674 [cs, math, stat]. URL: http://arxiv.org/abs/1902.04674 (visited on 08/05/2023). preprint.

[RR07] Ali Rahimi and Benjamin Recht. "Random Features for Large-Scale Kernel Machines". In: *Advances in Neural Information Processing Systems*. Vol. 20. Curran Associates, Inc., 2007. URL: https://papers.nips.cc/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html (visited on 11/20/2022).

[RZ18] Alexander Rakhlin and Xiyu Zhai. *Consistency of Interpolation with Laplace Kernels Is a High-Dimensional Phenomenon*. Dec. 28, 2018. DOI: 10.48550/arXiv.1812.11167. arXiv: 1812.11167 [cs, math, stat]. URL: http://arxiv.org/abs/1812.11167 (visited on 08/05/2023). preprint.

[SS95a] David Saad and Sara Solla. "Dynamics of On-Line Gradient Descent Learning for Multilayer Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 8. MIT Press, 1995. URL: https://proceedings.neurips.cc/paper_files/paper/1995/hash/a1519de5b5d44b31a01de013b9b51a80-Abstract.html (visited on 08/21/2023).

[SS95b] David Saad and Sara A. Solla. "Exact Solution for On-Line Learning in Multilayer Neural Networks". In: *Physical Review Letters* 74.21 (May 22, 1995), pp. 4337–4340. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.74.4337. URL: https://link.aps.org/doi/10.1103/PhysRevLett.74.4337 (visited on 08/14/2023).

[SS95c] David Saad and Sara A. Solla. "On-Line Learning in Soft Committee Machines". In: *Physical Review E* 52.4 (Oct. 1, 1995), pp. 4225–4243. ISSN: 1063-651X, 1095-3787. DOI: 10.1103/PhysRevE.52.4225. URL: https://link.aps.org/doi/10.1103/PhysRevE.52.4225 (visited on 08/21/2023).

[TAP21] Nilesh Tripuraneni, Ben Adlam, and Jeffrey Pennington. *Covariate Shift in High-Dimensional Random Feature Regression*. Nov. 16, 2021. DOI: 10.48550/arXiv.2111.08234. arXiv: 2111.08234 [cs, stat]. URL: http://arxiv.org/abs/2111.08234 (visited on 08/19/2023). preprint.

[Tro12a] Joel A. Tropp. "User-Friendly Tail Bounds for Sums of Random Matrices". In: *Foundations of Computational Mathematics* 12.4 (Aug. 1, 2012), pp. 389–434. ISSN: 1615-3383. DOI: 10.1007/s10208-011-9099-z. URL: https://doi.org/10.1007/s10208-011-9099-z (visited on 09/07/2023).

[Tro12b] Joel A. Tropp. "User-Friendly Tools for Random Matrices: An Introduction:" in: Fort Belvoir, VA: Defense Technical Information Center, Dec. 3, 2012. DOI: 10.21236/ADA576100. URL: http://www.dtic.mil/docs/citations/ADA576100 (visited on 09/07/2023).

[YS22] Gilad Yehudai and Ohad Shamir. *On the Power and Limitations of Random Features for Understanding Neural Networks*. Feb. 27, 2022. DOI: 10.48550/arXiv.1904.00687. arXiv: 1904.00687 [cs, stat]. URL: http://arxiv.org/abs/1904.00687 (visited on 09/07/2023). preprint.

[Zou+18] Difan Zou et al. *Stochastic Gradient Descent Optimizes Over-parameterized Deep ReLU Networks*. Dec. 27, 2018. DOI: 10.48550/arXiv.1811.08888. arXiv: 1811.08888 [cs, math, stat]. URL: http://arxiv.org/abs/1811.08888 (visited on 08/05/2023). preprint.