A Data-dependent Theory of Overparameterization: Phase Transition, Double Descent, and Beyond

Anonymous Authors[1]

Introduction

The big data revolution comes along with the challenging need to mine a large amount of large dimensional and complex data, with huge machine learning systems. For a machine learning system having N parameters, trained on a data set of size n, statistical analysis typically either focuses on the (statistical) population $n \to \infty$ limit, for N fixed, or the overparameterization $N \to \infty$ limit, for a given n. These two settings are technically more convenient to work with, yet less practical, as they essentially assume that one of the two dimensions is negligibly small compared to the other. More recently, by focusing on the so-called *double asymptotic* regime where n, N are both large and comparable, theoretical and empirical studies are conducted and lead to novel understanding and improved design of large-scale machine learning systems, including the popular "double descent" phenomenon [1, 2, 6].

Exploiting tools from Random Matrix Theory (RMT), here we develop a *data-dependent* theory to provide *quantitative* assessments of large-scale learning systems, as a function of their *relative complexity* N/n. This allows for a precise description of the under- to over-parameterized "phase transition" (that does not appear in the $N \to \infty$ alone analysis). The double descent phenomenon, for instance, is recognized as a direct consequence of this fundamental phase transition.

This article is based on the results in [4, 5]. We refer the interested readers to [4] for more detailed results, proofs, discussions, and numerical evaluations.

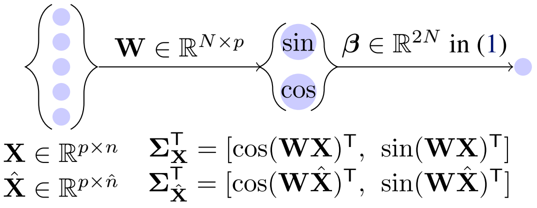
$\mathbf{W} \in \mathbb{R}^{N \times p}$
sin
cos
$\boldsymbol{\beta} \in \mathbb{R}^{2N}$ in (1)
$\mathbf{X} \in \mathbb{R}^{p \times n}$ $\quad \boldsymbol{\Sigma}_{\mathbf{X}}^{\mathsf{T}} = [\cos(\mathbf{W}\mathbf{X})^{\mathsf{T}},\ \sin(\mathbf{W}\mathbf{X})^{\mathsf{T}}]$
$\hat{\mathbf{X}} \in \mathbb{R}^{p \times \hat{n}}$ $\quad \boldsymbol{\Sigma}_{\hat{\mathbf{X}}}^{\mathsf{T}} = [\cos(\mathbf{W}\hat{\mathbf{X}})^{\mathsf{T}},\ \sin(\mathbf{W}\hat{\mathbf{X}})^{\mathsf{T}}]$

Figure 1. Illustration of RFFs regression model.

Random Fourier features as a stylized model

In this article, we consider the popular random Fourier features (RFFs) ridge regression [7], as a stylized model for the analysis of neural networks. See Fig 1 for an illustration.

The RFF ridge regressor $\boldsymbol{\beta} \in \mathbb{R}^{2N}$ is given by, for $\lambda \geq 0$,

$$\begin{aligned} \boldsymbol{\beta} \equiv \tfrac{1}{n} \boldsymbol{\Sigma}_{\mathbf{X}} \left(\tfrac{1}{n} \boldsymbol{\Sigma}_{\mathbf{X}}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{X}} + \lambda \mathbf{I}_n \right)^{-1} \mathbf{y} \cdot 1_{2N>n}, \\ + \left(\tfrac{1}{n} \boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\Sigma}_{\mathbf{X}}^{\mathsf{T}} + \lambda \mathbf{I}_{2N} \right)^{-1} \tfrac{1}{n} \boldsymbol{\Sigma}_{\mathbf{X}}\, \mathbf{y} \cdot 1_{2N<n}. \end{aligned} \tag{1}$$

The two forms of $\boldsymbol{\beta}$ in (1) are equivalent for any $\lambda > 0$ and minimize the (ridge-regularized) squared loss $\frac{1}{n} \| \mathbf{y} - \boldsymbol{\Sigma}_{\mathbf{X}}^{\mathsf{T}} \beta \|^2 + \lambda \| \boldsymbol{\beta} \|^2$ on the training set $(\mathbf{X}, \mathbf{y})$ with RFFs $\boldsymbol{\Sigma}_{\mathbf{X}}$. Our objective is to characterize the large n, p, N asymptotics of the test Mean Squared Error (MSE), E_{test}, defined as

$$E_{\text{test}} = \tfrac{1}{\hat{n}} \| \hat{\mathbf{y}} - \boldsymbol{\Sigma}_{\hat{\mathbf{X}}}^{\mathsf{T}} \beta \|^2, \tag{2}$$

with $\boldsymbol{\Sigma}_{\hat{\mathbf{X}}}^{\mathsf{T}} \in \mathbb{R}^{\hat{n} \times 2N}$ the RFFs of a test set $(\hat{\mathbf{X}}, \hat{\mathbf{y}})$ of size $\hat{n}$.

The effective kernel of large-scale RFFs

It has been established in [7] that *entry-wise* the RFF Gram matrix $\boldsymbol{\Sigma}_{\mathbf{X}}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{X}} / N$ converges to the Gaussian kernel matrix $\mathbf{K}_{\text{Gauss}} \equiv \{ \exp(-\| \mathbf{x}_i - \mathbf{x}_j \|^2 / 2) \}_{i,j=1}^n$, as $N \to \infty$. This unfolds from the fact that $[\boldsymbol{\Sigma}_{\mathbf{X}}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{X}} / N]_{ij} = \frac{1}{N} \sum_{t=1}^N [\cos(\mathbf{x}_i^{\mathsf{T}} \mathbf{w}_t) \cos(\mathbf{w}_t^{\mathsf{T}} \mathbf{x}_j) + \sin(\mathbf{x}_i^{\mathsf{T}} \mathbf{w}_t) \sin(\mathbf{w}_t^{\mathsf{T}} \mathbf{x}_j)]$, for independent $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, so that by the strong law of large numbers, for fixed n, p, $[\boldsymbol{\Sigma}_{\mathbf{X}}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{X}} / N]_{ij}$ converges to its expectation almost surely as $N \to \infty$, i.e., $[\boldsymbol{\Sigma}_{\mathbf{X}}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{X}} / N]_{ij} \xrightarrow{a.s.} \mathbf{K}_{\cos} + \mathbf{K}_{\sin}$, with $\mathbf{K}_{\cos}, \mathbf{K}_{\sin}$ defined in (4)–(5) below and $\mathbf{K}_{\cos} + \mathbf{K}_{\sin} = \mathbf{K}_{\text{Gauss}}$.

This, however, does not imply the *spectral norm* convergence $\| \boldsymbol{\Sigma}_{\mathbf{X}}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{X}} / N - \mathbf{K}_{\text{Gauss}} \| \to 0$, due to the factor n, now large, in the norm inequality $\| \mathbf{K} \|_\infty \leq \| \mathbf{K} \| \leq n \| \mathbf{K} \|_\infty$ for $\mathbf{K} \in \mathbb{R}^{n \times n}$ and $\| \mathbf{K} \|_\infty \equiv \max_{ij} | \mathbf{K}_{ij} |$. A trivial example is the case $2N < n$, where $\operatorname{rank}(\boldsymbol{\Sigma}_{\mathbf{X}}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{X}}) \leq 2N < n$ (as the sum of $2N$ rank-one matrices), while $\operatorname{rank}(\mathbf{K}_{\text{Gauss}}) = n$ for distinct data points, see [8, Theorem 2.18]. As a result, for various machine learning algorithms the performance guarantee offered by the (limiting) Gaussian kernel is less likely to agree with empirical observations in real-world large-scale problems, when $N \not\gg n$. In Fig 2 we compare the training MSEs of RFF ridge regression on MNIST

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Under review by the Workshop on the Theory of Overparameterized Machine Learning (TOPML).

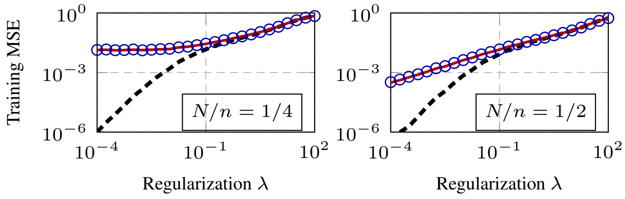

Figure 2. Training MSEs of RFF ridge regression on MNIST data, as a function of λ, for $p = 784$, $n = 1\,000$, $N = 250$ and 500. Empirical results in **blue** circles; Gaussian kernel (assuming $N \to \infty$ alone) in **black** dashed lines; and RMT predictions in **red** solid lines. Results obtained by averaging over 30 runs.

data [3] for $2N \leq n$: while the empirical results diverge from the Gaussian kernel predictions for small λ, our results *consistently* fit the empirical observations almost perfectly.

To precisely characterize the large n, p, N asymptotics of the RFF ridge regression model, we assume the follows.

Assumption 1 (High dimensional asymptotics)**.** *As $n \to \infty$, **(i)** $p/n = O(1)$, $N/n = O(1)$, $\hat{n}/n = O(1)$; **(ii)** the norms $\|\mathbf{X}\|$, $\|\hat{\mathbf{X}}\|$, $\|\mathbf{y}\|_\infty$ and $\|\hat{\mathbf{y}}\|_\infty$ are all of order $O(1)$, i.e., the data and targets are normalized with respect to n and $\hat{n}$.*

Note that the RFF regressor $\boldsymbol{\beta}$ in (1) is closely connected to

$$\mathbf{Q}(\lambda) \equiv \left(\tfrac{1}{n}\boldsymbol{\Sigma}_{\mathbf{X}}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{X}} + \lambda\mathbf{I}_n\right)^{-1} \in \mathbb{R}^{n\times n}, \quad \lambda \geq 0 \tag{3}$$

the *resolvent* of $\boldsymbol{\Sigma}_{\mathbf{X}}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{X}}/n$, the behavior of which, as discussed above, is *different* from $(\frac{N}{n}\mathbf{K}_{\mathrm{Gauss}} + \lambda\mathbf{I}_n)^{-1}$ in the large n, p, N regime and is precisely described as follows.

Theorem 1 (Asymptotic equivalent for $\mathbb{E}[\mathbf{Q}]$)**.** *Under Assumption 1, for $\mathbf{Q}$ defined in (3) and $\lambda > 0$, as $n \to \infty$,*

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\| \to 0,\ \bar{\mathbf{Q}} \equiv \left(\frac{N}{n}\left(\frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}} + \frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}}\right) + \lambda\mathbf{I}_n\right)^{-1},$$

for $\mathbf{K}_\sigma \equiv \mathbf{K}_\sigma(\mathbf{X}, \mathbf{X}) \in \mathbb{R}^{n\times n}$, $\sigma \in \{\cos, \sin\}$ and

$$[\mathbf{K}_{\cos}(\mathbf{X}, \mathbf{X}')]_{ij} = e^{-\frac{1}{2}(\|\mathbf{x}_i\|^2 + \|\mathbf{x}_j'\|^2)}\cosh(\mathbf{x}_i^{\mathsf{T}}\mathbf{x}_j'), \tag{4}$$

$$[\mathbf{K}_{\sin}(\mathbf{X}, \mathbf{X}')]_{ij} = e^{-\frac{1}{2}(\|\mathbf{x}_i\|^2 + \|\mathbf{x}_j'\|^2)}\sinh(\mathbf{x}_i^{\mathsf{T}}\mathbf{x}_j'), \tag{5}$$

with $\delta_{\cos} = \frac{1}{n}\operatorname{tr}(\mathbf{K}_{\cos}\bar{\mathbf{Q}})$ and $\delta_{\sin} = \frac{1}{n}\operatorname{tr}(\mathbf{K}_{\sin}\bar{\mathbf{Q}})$.

Taking $N/n \to \infty$, one has $\delta_{\cos}, \delta_{\sin} \to 0$ so that $\boldsymbol{\Phi} \equiv \frac{\mathbf{K}_{\cos}}{1+\delta_{\cos}} + \frac{\mathbf{K}_{\sin}}{1+\delta_{\sin}} \to \mathbf{K}_{\mathrm{Gauss}}$, in accordance with the classical large-N prediction. In this sense, the pair $(\delta_{\cos}, \delta_{\sin})$ introduced in Theorem 1 accounts for the "correction" due to the non-trivial n/N, as opposed to the $N \to \infty$ alone analysis.

Phase transition and the corresponding double descent

Both $\delta_{\cos}$ and $\delta_{\sin}$ are decreasing functions of N, as depicted in Fig 3. More importantly, Fig 3 also illustrates that $\delta_{\cos}$ and $\delta_{\sin}$ exhibit *qualitatively different* behavior, depending on the sign of $2N - n$. For not-too-small $\lambda = 1$, both $\delta_{\cos}$ and $\delta_{\sin}$ decrease *smoothly*, as N/n grows large. However, for small $\lambda \ll 1$, we see a *phase transition* behavior at the interpolation threshold $2N = n$. More precisely, $\delta_{\cos}$ and $\delta_{\sin}$ "jump" from order $O(1)$ (when $2N > n$) to much larger values of order λ^{-1} (when $2N < n$) for $\lambda = 10^{-7}$.

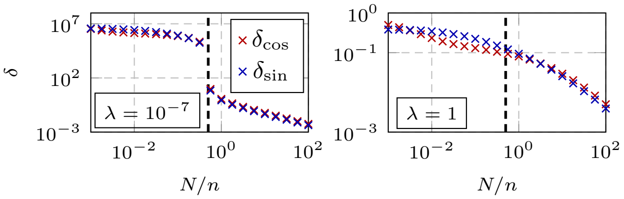

Figure 3. Behavior of $(\delta_{\cos}, \delta_{\sin})$ on MNIST data, as a function of N/n, $p = 784$, $n = 1\,000$, $\lambda = 10^{-7}$ and 1. The **black** dashed line is the interpolation threshold $2N = n$.

This phase transition can be theoretically justified by considering the *ridgeless* $\lambda \to 0$ limit in Theorem 1. Note that, for $\lambda = 0$ and $2N < n$, $\mathbf{Q}(\lambda = 0)$ in (3) is simply undefined, as it involves inverting a singular matrix $\boldsymbol{\Sigma}_{\mathbf{X}}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{X}} \in \mathbb{R}^{n\times n}$ of rank at most $2N < n$. As a consequence, we expect to see both $\mathbf{Q}$ and $\bar{\mathbf{Q}}$ scale like λ^{-1} as $\lambda \to 0$ for $2N < n$, while for $2N > n$ this is no longer the case. As a consequence, we have the following two phases:

1. *Under-parameterized* with $2N < n$, where $\mathbf{Q}$ is not well-defined ($\mathbf{Q} \sim \lambda^{-1}$) and one must consider instead the properly scaled $\lambda\delta_{\cos}, \lambda\delta_{\sin}$ and $\lambda\bar{\mathbf{Q}}$ as $\lambda \to 0$.
2. *Over-parameterized* with $2N > n$, where one can take $\lambda \to 0$ in Theorem 1 to get $\delta_{\cos}, \delta_{\sin}$ and $\bar{\mathbf{Q}}$.

From the above phase transition, we can further derive the asymptotic test MSE $\bar{E}_{\mathrm{test}}$ of the RFF ridge regression, so that under Assumption 1, $E_{\mathrm{test}} - \bar{E}_{\mathrm{test}} \xrightarrow{a.s.} 0$, as $n \to \infty$, for E_{test} defined in (2) and $\bar{E}_{\mathrm{test}}$ depending on N/n, $N/\hat{n}$, and on the data $\mathbf{X}, \hat{\mathbf{X}}$ via $\mathbf{K}_\sigma = \mathbf{K}_\sigma(\mathbf{X}, \mathbf{X})$, $\mathbf{K}_\sigma(\hat{\mathbf{X}}, \mathbf{X})$ and $\mathbf{K}_\sigma(\hat{\mathbf{X}}, \hat{\mathbf{X}})$, $\sigma \in \{\cos, \sin\}$, as in (4)–(5), as well as the targets $\mathbf{y}$ and $\hat{\mathbf{y}}$, see [4] for more details on this.

Fig 4 depicts the test MSEs for different values of regularization λ. In particular, for small $\lambda = 10^{-7}$, a double-descent behavior is observed, with a singularity at $2N = n$; while for larger $\lambda = 1$, a smoother and monotonically decreasing test error curve is observed.

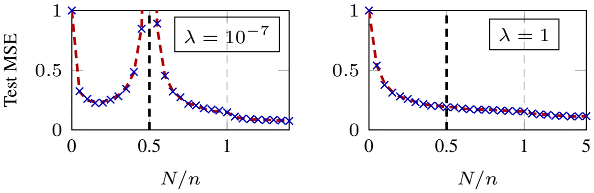

Figure 4. Empirical (**blue** crosses) and theoretical (**red** dashed lines) test MSEs of RFF regression as a function of the ratio N/n, on MNIST data, for $p = 784$, $n = 500$, $\lambda = 10^{-7}$ and 1.

References

[1] Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 2020.

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

[3] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[4] Zhenyu Liao, Romain Couillet, and Michael W Mahoney. A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. In *Advances in Neural Information Processing Systems*, volume 33, pages 13939–13950, 2020.

[5] Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

[6] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

[7] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

[8] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.