Part 0: Introduction: a brand new theory for modern neural networks

- Framing the problem. Modern neural networks defies classical understanding. For example, against classical statistical learning: tradeoff between bias and variance. Overparametrization (quantitatively defined under VC dim for example) should lead to overfitting and worse performances. Modern networks employ large numbers of parameters, often exceeding training data size by several orders of magnitude. Still, empirical studies proves the contrary. Cite AlexNet and, above all, Zhang, Bengio et Al (2017) providing a nice experiment on this subject (overparametrized nets performing well on natural data, overfit on random labels. Somehow, in the landscape of all possible network configuration with zero training loss, the network with gradien descent chooses the solution that's able to generalize well). All of this a first hint that gen. error may depend on data structure too. Then Belkin et Al (2019) with empirical double descent (explaining the "inductive bias of smoothness") to harmonize together old and new perspective.

- On the same level, empirical neural scaling laws Kaplan @ OpenAI (2020). The microscopic architecture seems irrelevant, only macroscopic variable (num. parameters, compute time) seems to matter. Interesting "statphys" perspective, btw. A power law implies "bigger is better", which again is paradoxical under "old" classical statistical learning theory.

Part 1: Random feature models and kernel machines (~ paper 1)

- Consider (Belkin 2018: To Understand Deep Learning We Need to Understand Kernel Learning): this kinda provides a justification for us shifting from neural nets to kernels or random feature models. Here, explain lazy regime for NNs and, most importantly, Neural Tangent Kernel to defend this switch even further (Jacot 2020 e le note di Biroli)
- Definition of random feature models (starting from seminal paper Rahimi Recht and Montecarlo sampling, hence connecting to kernels). Random feature models (RFM) are also interesting as shallow neural networks. Learning task definition under RFM (definition, ...). Characterizing the learning problem by the spectral property of the integral operator $Tf = \int dx p(x) \varphi(x;w)f(x)$.
- Introduction to deterministic equivalent, highlighting self-averaging property (quenched disorder over random weights). Deterministic formula as in Paper 1 and kernel limit under $p\to\infty$.
- Scaling law derivation under source and capacity conditions (assuming exact power-law spectral decay!)

Part 2: How data affects spectral structure (~ Paper 2)

- Focus on kernel ridge regression. How does the structure of data influence the spectral properties of the learning problem?

- Theorems and propositions in Paper 2 for dot-product kernels. Spectral gap and continuous spectrum differentiation. Generalization error depends also on data structure

- Da qui in poi la parte di ricerca
