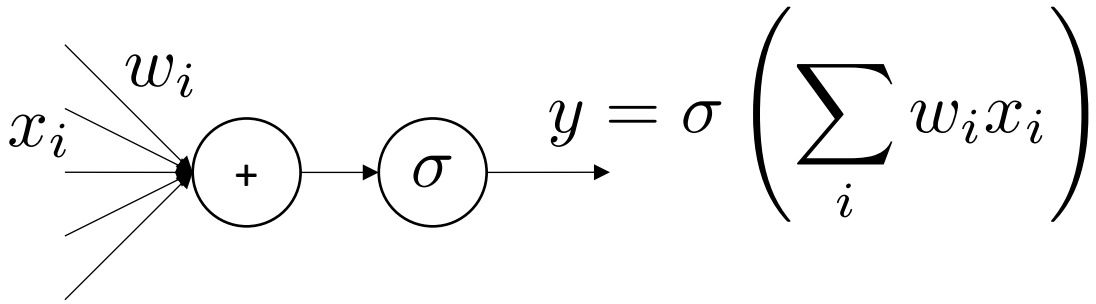
Optimal Activation Functions from the Random Feature Regression Model

Jianxin Wang & José Bento

Motivation

Neural networks are powerful function approximators.

Most networks are non-linear due to the choice of Activation Function (AF).



The selection of activation function has been given substantial attention and a good choice of activation can make substantial difference in terms of approximation power, ease of learning and implementation.

Problem

We want to find an AF that is simple and leads to good approximation in a setting where these two concepts can be formalized and quantified analytically.

Our focus is on studying the role of AF when learning functions using the Random Feature Regression model on high dimensional data, i.e.

$$d \to \infty$$

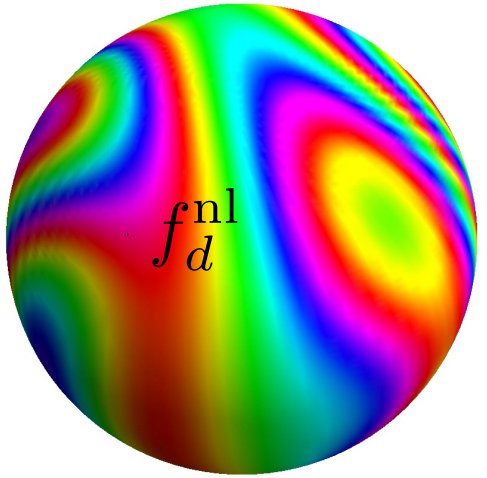
Setup

Consider a function to be learned $f_d \in \mathbb{S}^{d-1}(\sqrt{d})$, $d \to \infty$, that is the sum of a linear component and a random non-linear component:

$$f_d(x) = \beta_{d,0} + \beta_{d,1}^T x + f_d^{\mathrm{nl}}(x)\,,$$

where $\lim_{d\to\infty} \beta_{d,0}^2 = F_0^2$ and $\lim_{d\to\infty} \|\beta_{d,1}\|_2^2 = F_1^2$ and

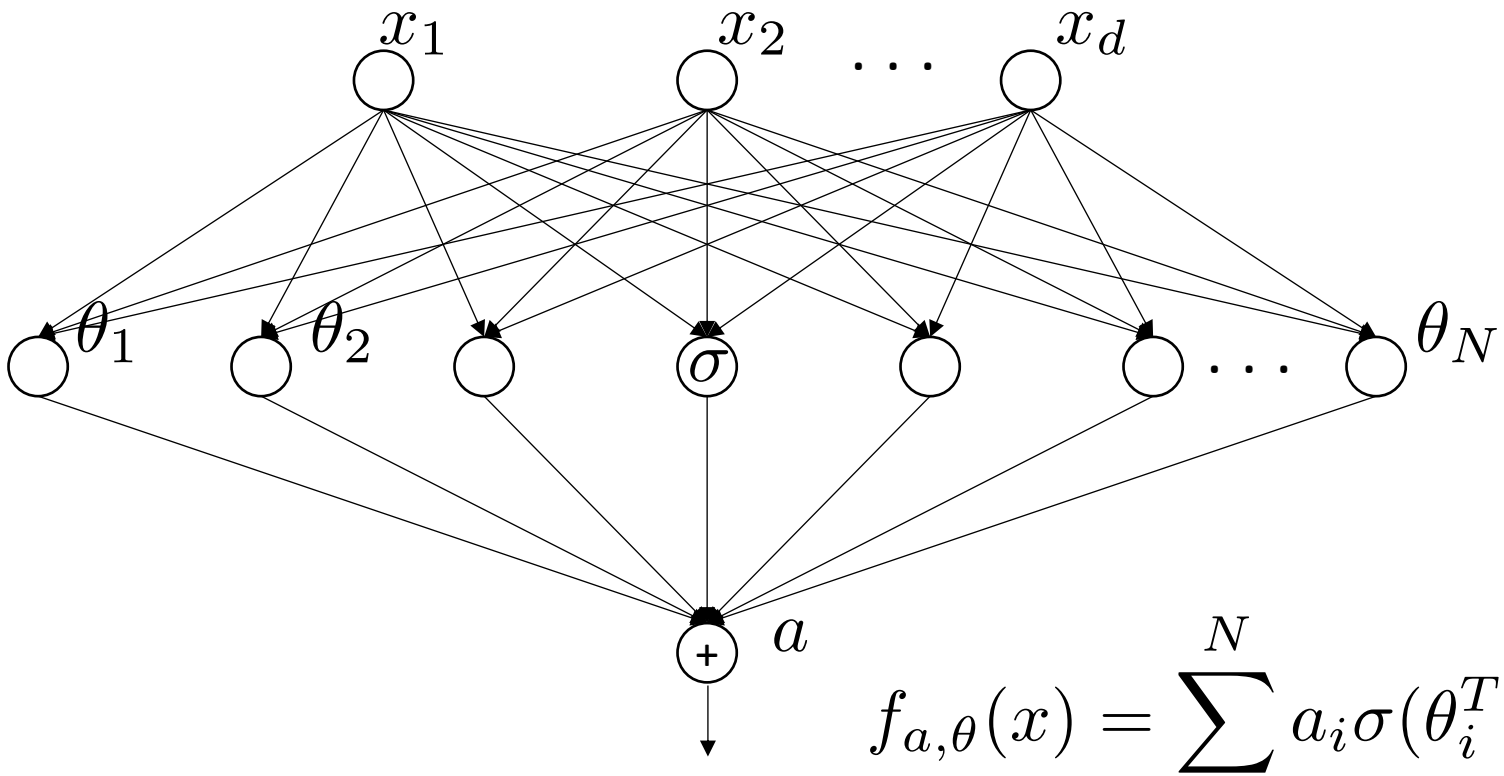
$\lim_{d\to\infty} \mathbb{E}_{f_d^{\mathrm{nl}}}\{(f_d^{\mathrm{nl}}(x))^2\} = F_\star^2$



In particular, $F_\star = 0$ means that the target function $f_d(x)$ is linear, and F_1^2 controls the magnitude of the linear component.

Setup

To learn f_d we consider the Random Feature Regression (RFR) model.
RFR is a two-layer network with random weights on the first layer.

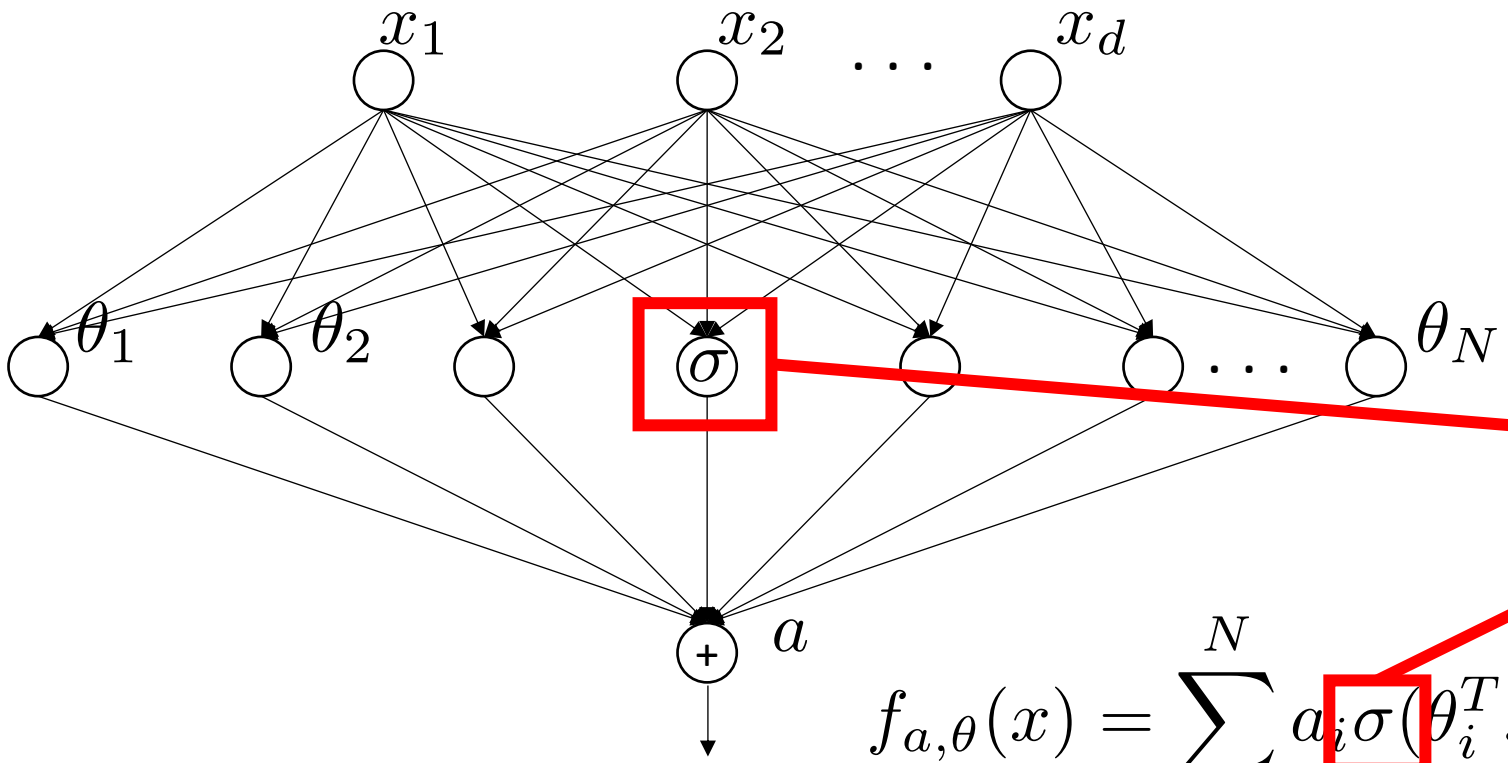


$$\{\theta_i\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} \text{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$$

$$f_{a,\theta}(x) = \sum_{i=1}^{N} a_i \sigma(\theta_i^T x/\sqrt{d}),$$

Setup

To learn f_d we consider the Random Feature Regression (RFR) model.
RFR is a two-layer network with random weights on the first layer.



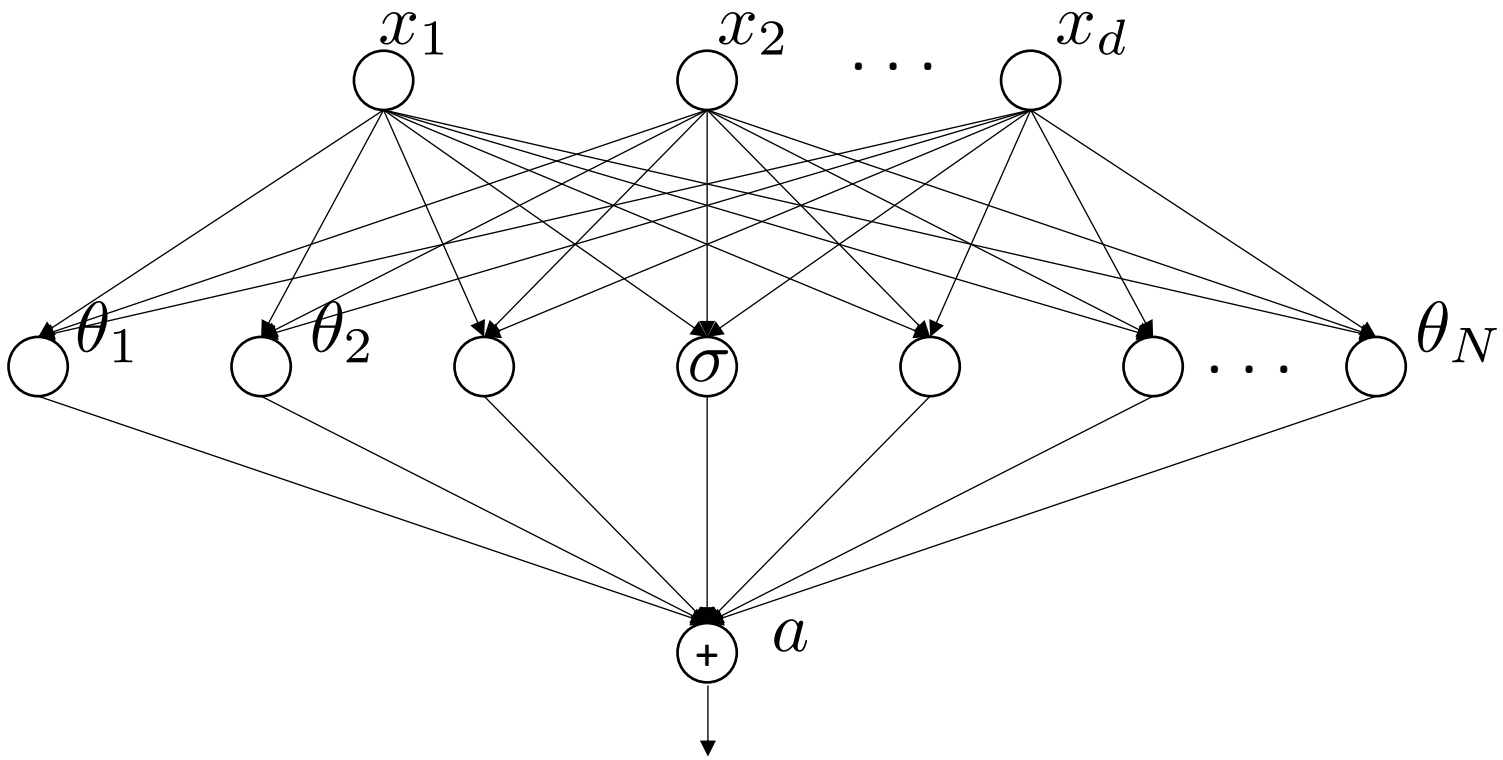
$$\{\theta_i\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} \text{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$$

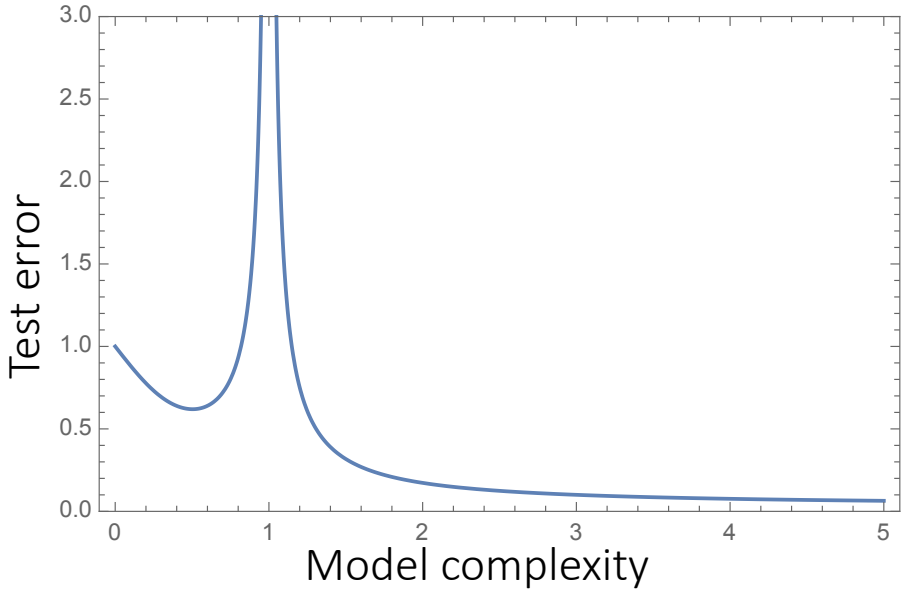
The focus of our study is this Activation Function (AF)

$$f_{a,\theta}(x) = \sum_{i=1}^{N} a_i \sigma(\theta_i^T x/\sqrt{d}),$$

Setup

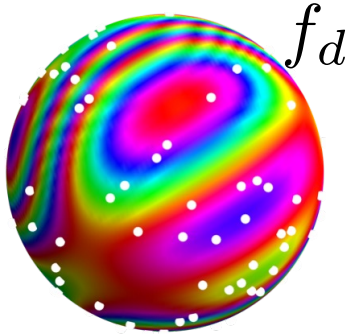
It is one of the models for which we can analytically study phenomena like the double descent curve.





Setup

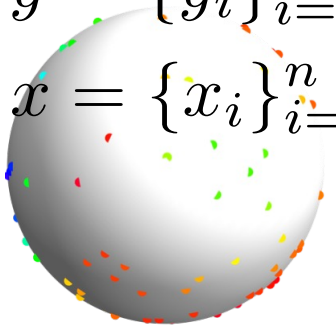
We obtain noisy samples from f_d, $y_i = f_d(x_i) + \epsilon_i$, $x_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathbb{P}_\epsilon, i = 1, \ldots, n$, that satisfy $\mathbb{E}_\epsilon(\epsilon_i) = 0$, $\mathbb{E}_\epsilon(\epsilon_i^2) = \tau^2$, $\mathbb{E}_\epsilon(\epsilon_i^4) < \infty$ and solve a regularized least square problem to learn the second layer a.



$f_d(x)$

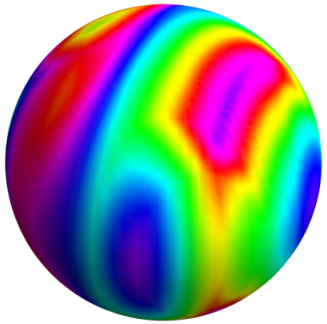
$y = \{y_i\}_{i=1}^n$

$x = \{x_i\}_{i=1}^n$



$$\min_a \frac{1}{n}\|y - f_{a,\theta}(x)\|_2^2 + \frac{N\lambda}{d}\|a\|_2^2$$

$f_{a^*,\theta}$



$$\mathcal{E} \triangleq \mathbb{E}_x((f_{a^*,\theta}(x) - f_d(x))^2)$$

Problem

What is the AF function that minimizes

$$\min_{\sigma} \ \mathcal{E}$$

that also minimizes

$$\|\sigma\|_1 \triangleq \mathbb{E}_Z(|\sigma'(Z))|) \quad \text{or} \quad \|\sigma\|_2 \triangleq \sqrt{\mathbb{E}_Z((\sigma'(Z))^2)}, \quad Z \sim \mathcal{N}(0,1)$$

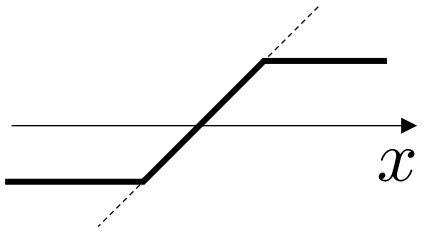
?

Main results

Theorem [Wang & Bento]

If $\|\sigma'(x)\| = \mathbb{E}(|\sigma'(Z)|), Z \sim \mathcal{N}(0,1)$,
then the optimal AF functions are of the form:

$$\sigma(x) = a + b\max\{c, \min\{d, x\}\}$$



where $c \leq d$ and the constants a, b, c and d need to be choose such that

$$\mu_0 = \mathbb{E}\{\sigma(Z)\},\ \mu_1 = \mathbb{E}\{Z\sigma(Z)\},\ \mu_2 = \mathbb{E}\{\sigma(Z)^2\}$$

Main results

Theorem [Wang & Bento]

If $\|\sigma'(x)\| = \sqrt{\mathbb{E}((\sigma'(Z))^2)}, Z \sim \mathcal{N}(0,1)$,
then the optimal AF functions are the two quadratics:

$$\sigma(x) = ax^2 + bx + c$$

$$a = \pm\frac{\mu_\star}{\sqrt{2}}, \quad b = \mu_1, \quad c = \mu_0 - a$$

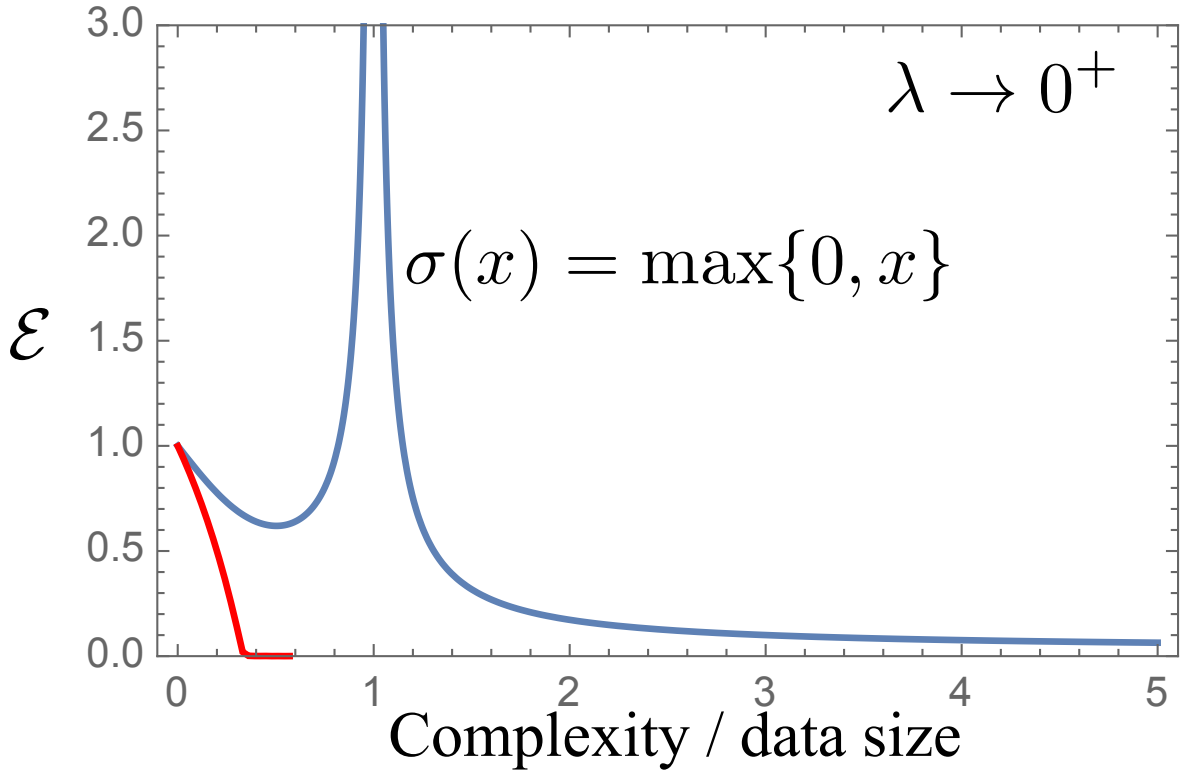
where $\mu_\star^2 = \mu_2 - \mu_0^2 - \mu_1^2$ and $\mu_0 = \mathbb{E}\{\sigma(Z)\},\ \mu_1 = \mathbb{E}\{Z\sigma(Z)\},\ \mu_2 = \mathbb{E}\{\sigma(Z)^2\}$

Remark: Note that $\mu_\star = 0$ if and only if the AF is linear.

Main results

Theorem [Wang & Bento]

Using an optimal AF (linear) can destroy the double descent curve & lead to zero test error for low-complexity models.



$F_\star = 0$

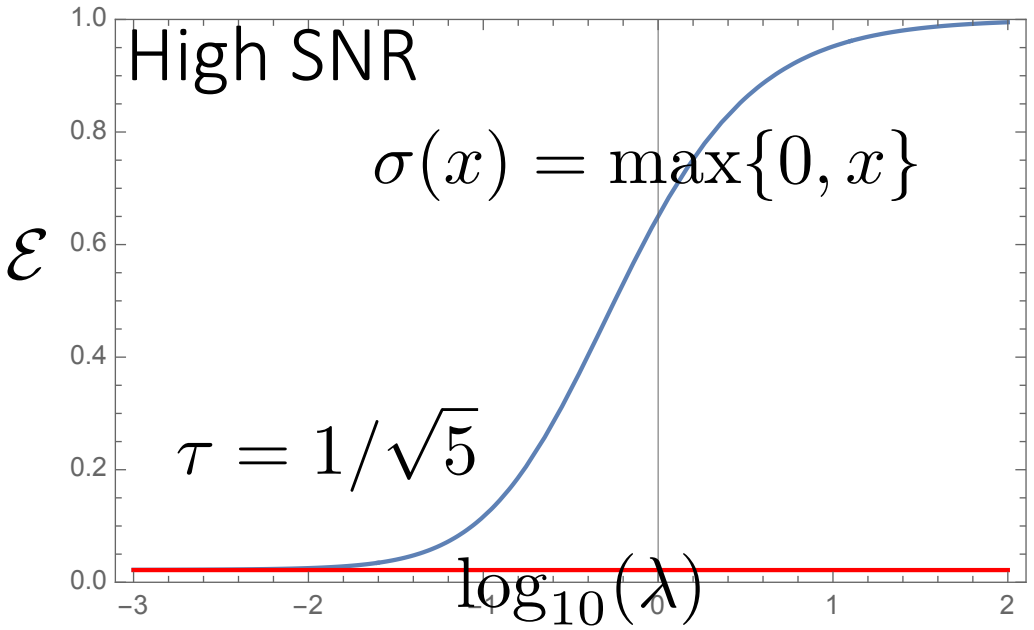
$F_1 = 1$

$\tau = 0$

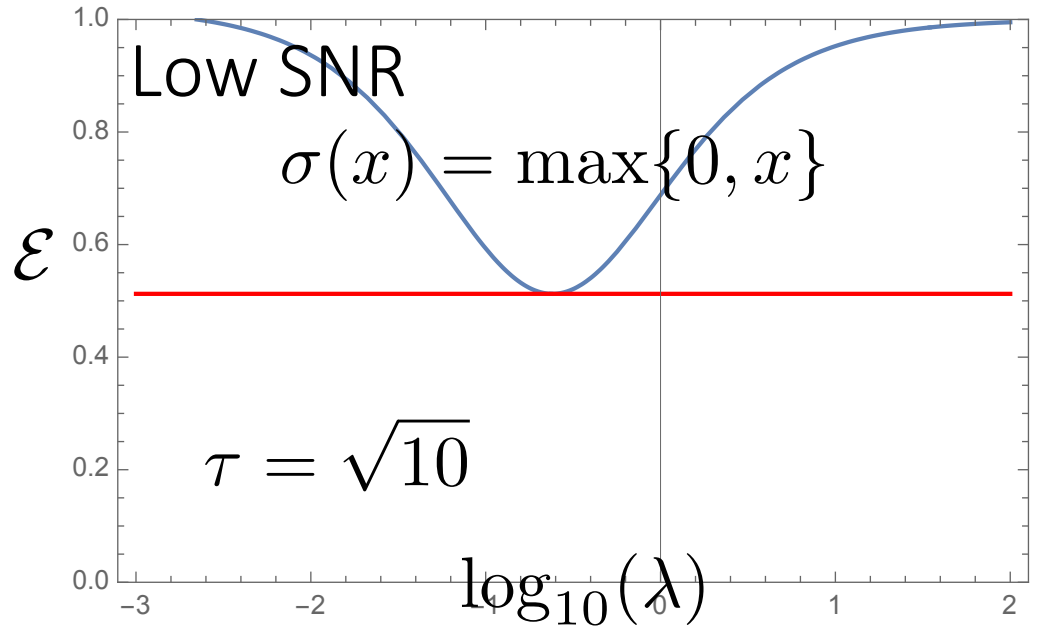
Main results

Theorem [Wang & Bento]

For overparameterized models, using an optimal AF with an arbitrary regularization is as good, or better, than using an optimal regularization with an non-optimal AF, both in low SNR and high SNR.



$$F_\star = 0$$
$$F_1 = 1$$



Thank you !

For more information check out our paper:
https://arxiv.org/abs/2206.01332