Learning features with two-layer neural networks, one step at a time

Bruno Loureiro
@ CSD, DI-ENS & CNRS

brloureiro@gmail.com

DIMACS Workshop on Modeling Randomness in Neural Network Training
06.06.2024

Learning features with two-layer neural networks, one step at a time

Bruno Loureiro
@ CSD, DI-ENS & CNRS

brloureiro@gmail.com

DIMACS Workshop on Modeling Randomness in Neural Network Training
06.06.2024

DIMACS Workshop on Modeling Randomness in Neural Network Training

June 5-7, 2024 at Rutgers University

About | Participants | Schedule

The DIMACS Workshop on Modeling Randomness in Neural Network Training: Mathematical, Statistical, and Numerical Guarantees will be held at the DIMACS Center at Rutgers University from **June 5-7, 2024**. The central question of this workshop is: *what can random matrix theory tell us about neural networks, modern machine learning, and AI?*

One goal of the workshop will be to create bridges between the different mathematical and computational communities by bringing together researchers with a diverse set of perspectives on neural networks. Topics of interest include:

- understanding matrix-valued random processes that arise during NN training,
- modeling/measuring uncertainty and designing estimators for training processes,
- applications to these designs within optimization algorithms.

How Two-Layer Neural Networks Learn, One (Giant) Step at a Time

Yatin Dandi[1,3], Florent Krzakala[1], Bruno Loureiro[2], Luca Pesce[1], and Ludovic Stephan[1]

arXiv: 2305.18270

Asymptotics of feature learning in two-layer networks after one gradient-step

Hugo Cui[1], Luca Pesce[2], Yatin Dandi[2,1], Florent Krzakala[2], Yue M. Lu[3],
Lenka Zdeborová[1], and Bruno Loureiro[4]

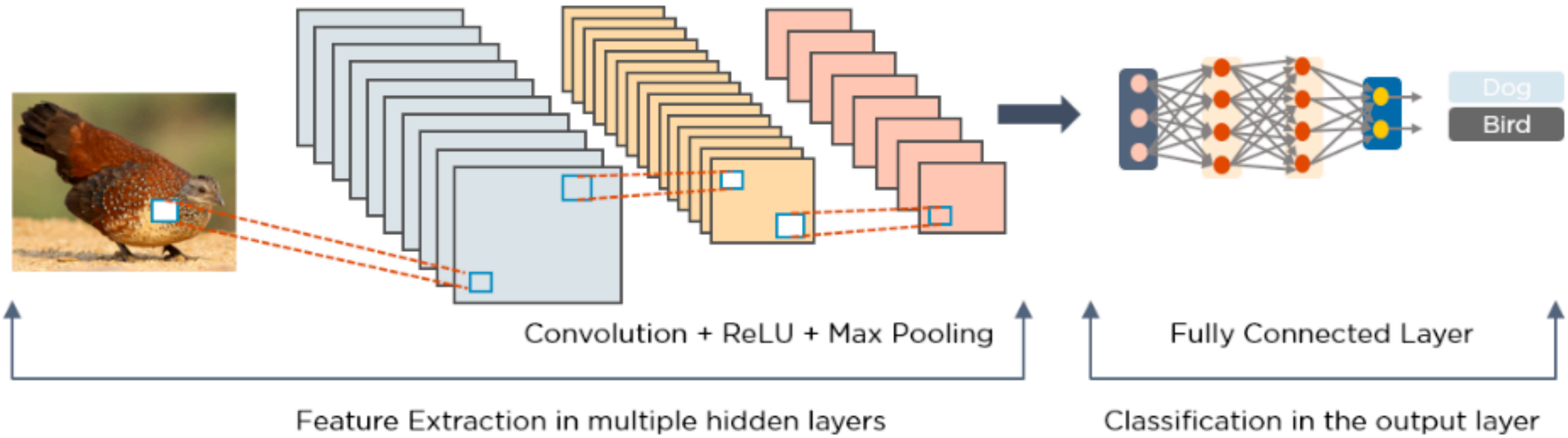
arXiv: 2402.04980
(ICML 2024)

Feature Learning after One Gradient Descent Step: A Random Matrix Theory Perspective

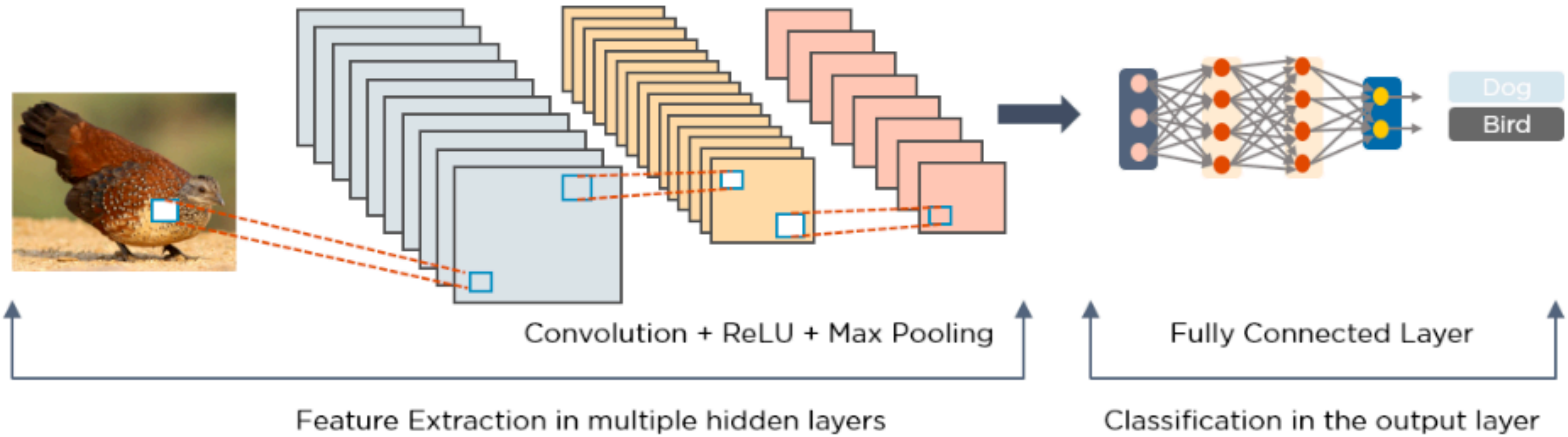
Yatin Dandi[1], Luca Pesce[2], Hugo Cui[1], Florent Krzakala[2], Yue M. Lu[3], and Bruno Loureiro[4]

arXiv: 2406.XXXX

Neural networks are good because they adapt and "learn features" from the data

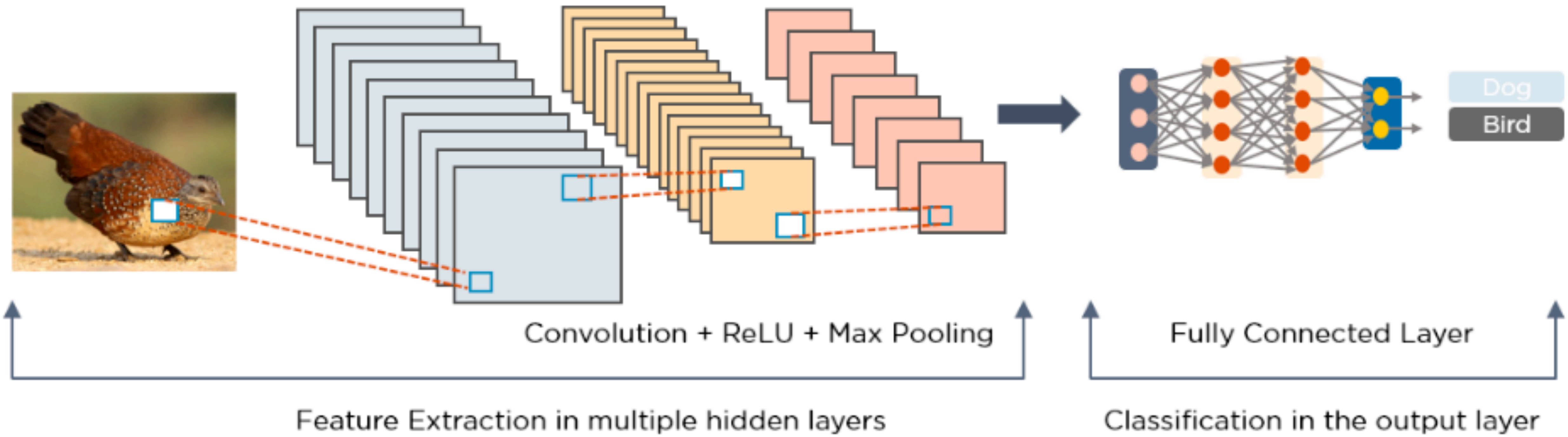


Neural networks are good because they adapt and "learn features" from the data



🤔 But what this exactly means?

Neural networks are good because they adapt and "learn features" from the data



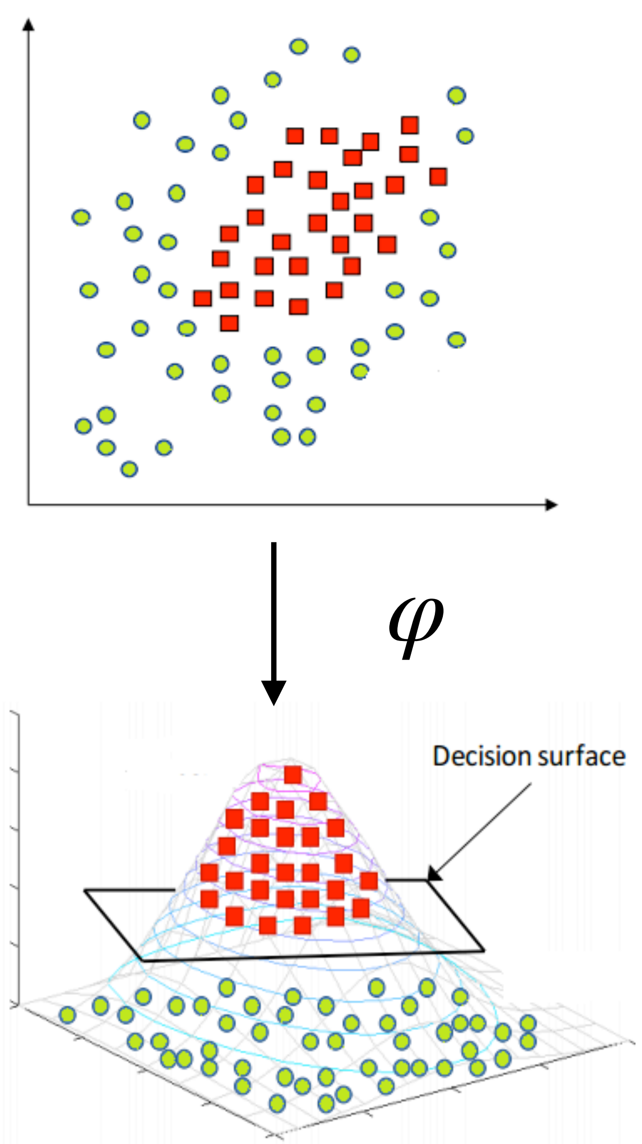
🤔 But what this exactly means?

Goal: make sense of this in a simple setting

Today's menu

Initialization

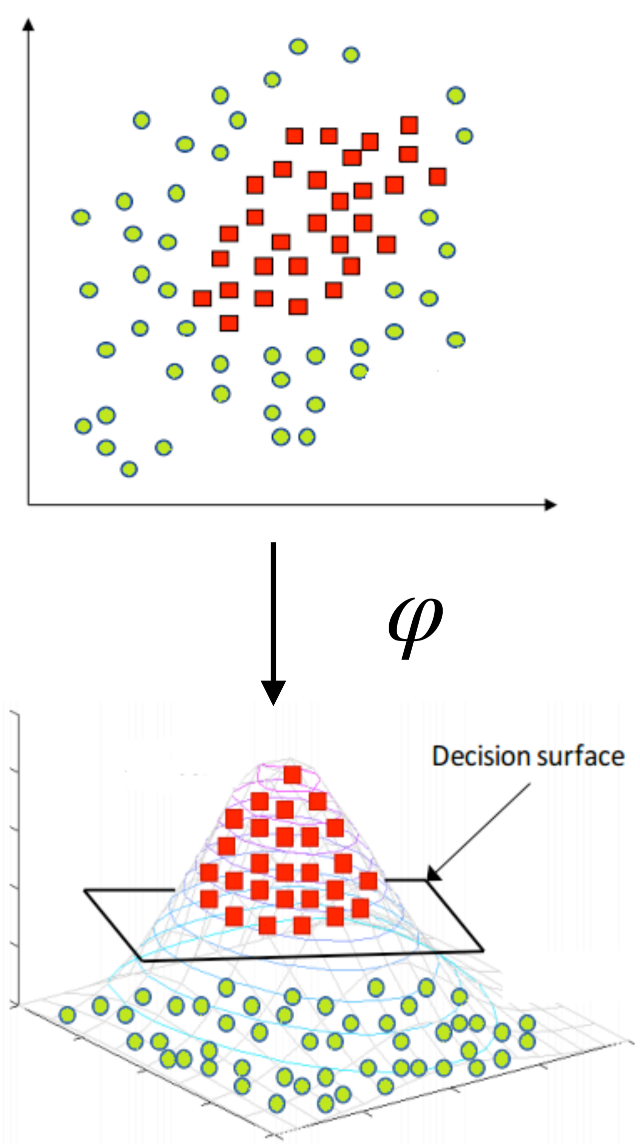
Random features and kernels



<u>Today's menu</u>

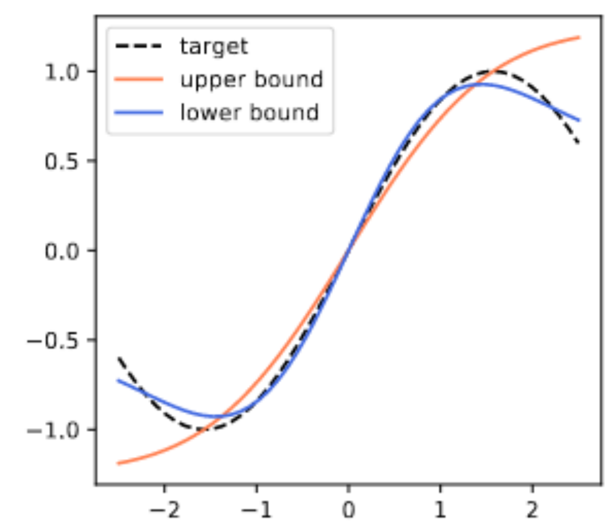
Initialization

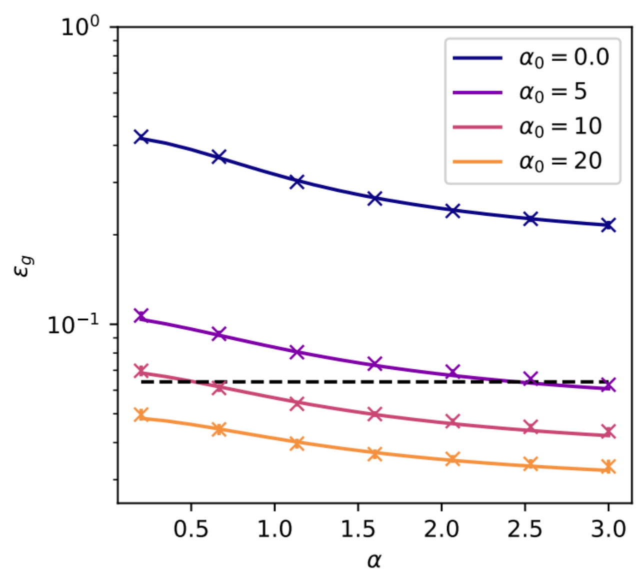
Random features and kernels



One step

Exact asymptotics for one GD step

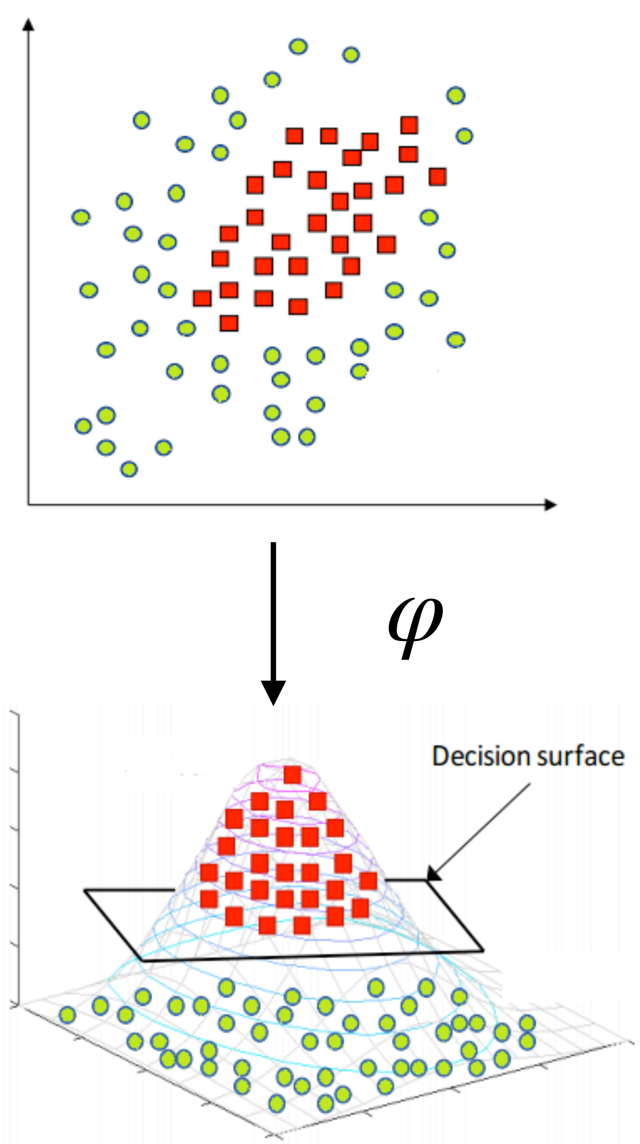




Today's menu

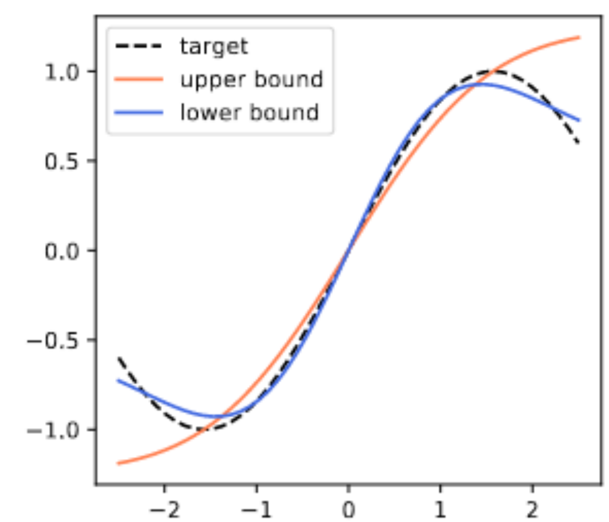
Initialization

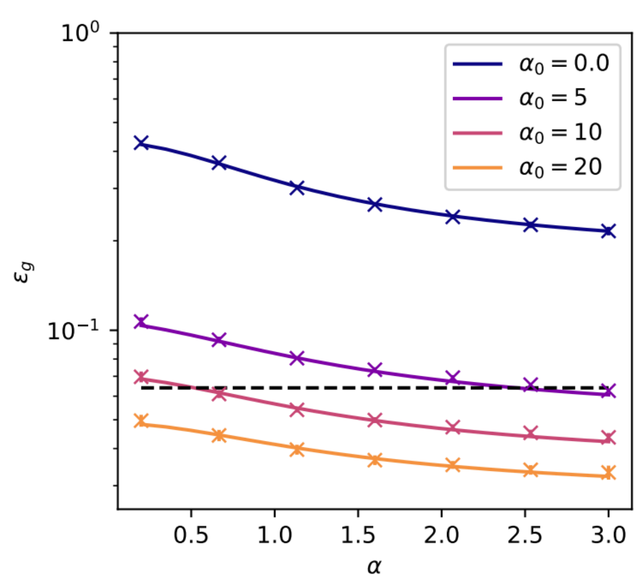
Random features and kernels



One step

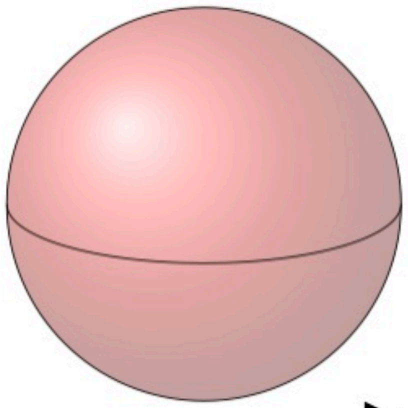
Exact asymptotics for one GD step





Few steps

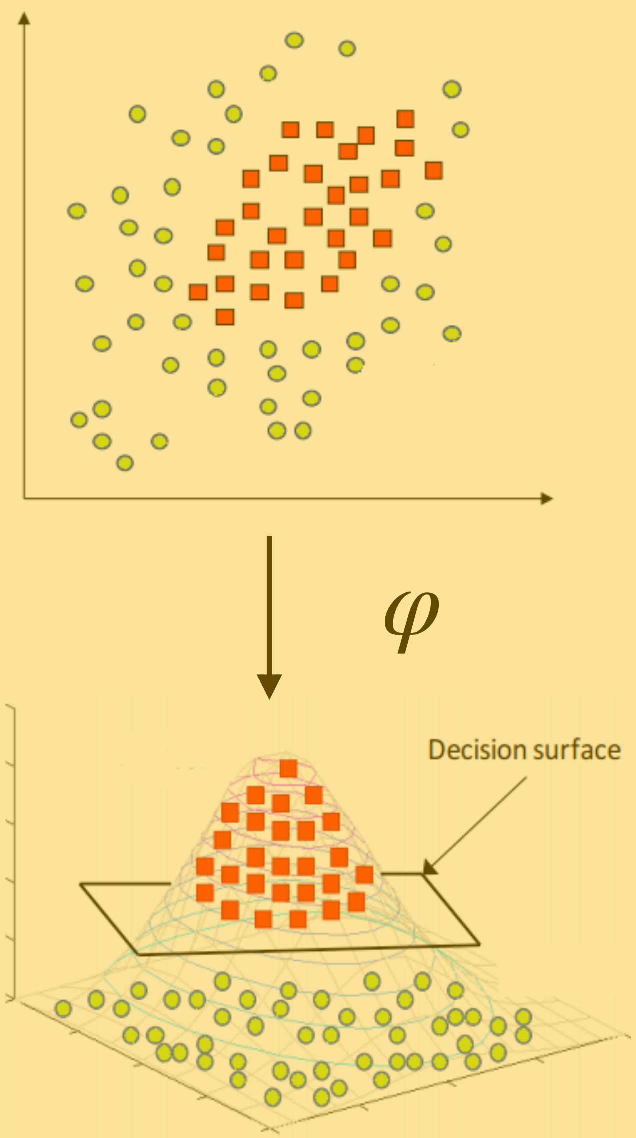
Learning staircase functions



Today's menu

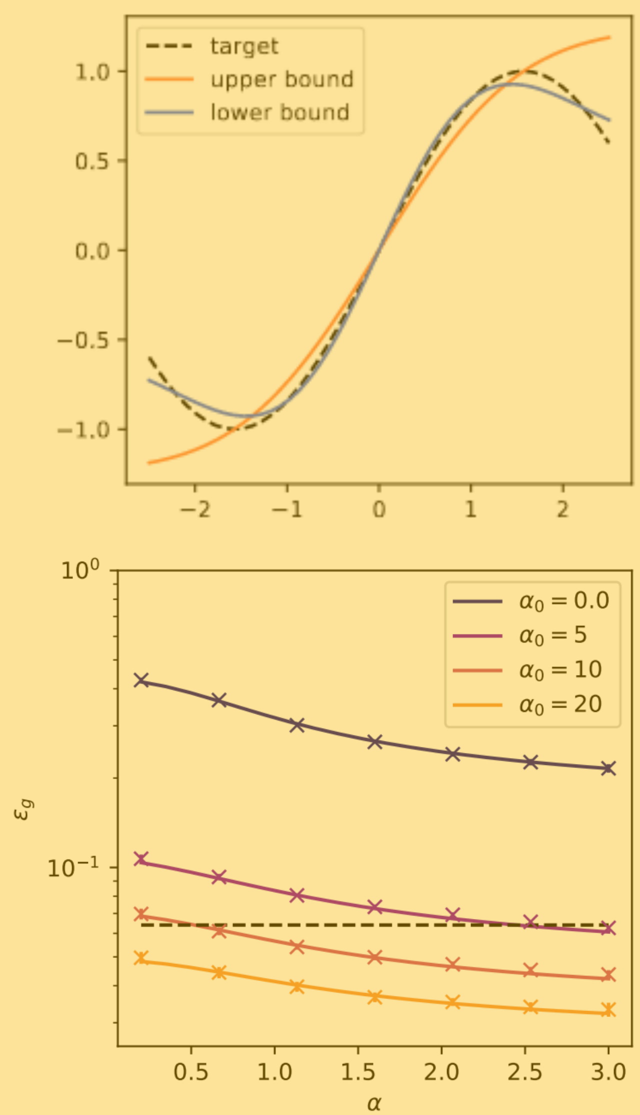
Initialization

Random features and kernels



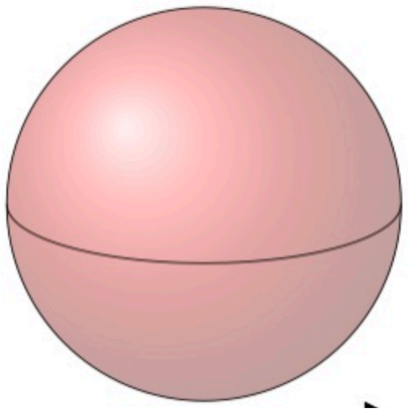
One step

Exact asymptotics for one GD step



Few steps

Learning staircase functions



Setting

Let $(x_i, y_i)_{i\in[n]} \in \mathbb{R}^{d+1}$ be the training data. We assume:

$$y_i = f_\star(x_i) + z_i$$

$$x_i \sim \mathcal{N}(0, I_d/d) \qquad\qquad z_i \sim \mathcal{N}(0, \Delta)$$

Setting

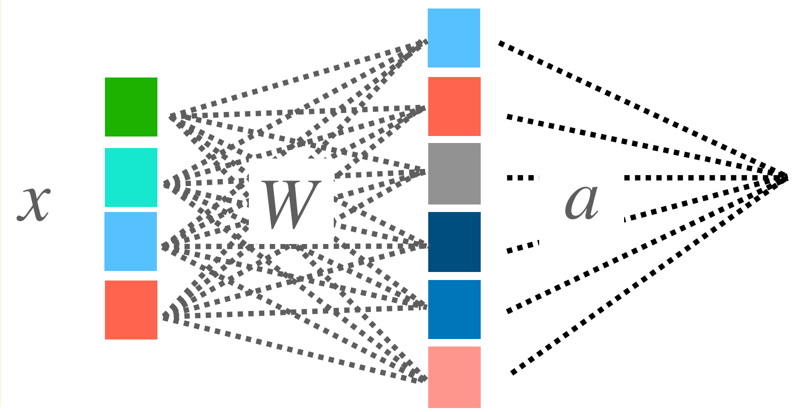
Let $(x_i, y_i)_{i\in[n]} \in \mathbb{R}^{d+1}$ be the training data. We assume:

$$y_i = f_\star(x_i) + z_i$$

$$x_i \sim \mathcal{N}(0, I_d/d) \qquad\qquad z_i \sim \mathcal{N}(0, \Delta)$$

We are interested in the performance of 2 layer NNs:

$$f(x; a, W) = \frac{1}{\sqrt{p}} \sum_{k=1}^{p} a_k \sigma\left(\langle w_k, x\rangle\right)$$



Setting

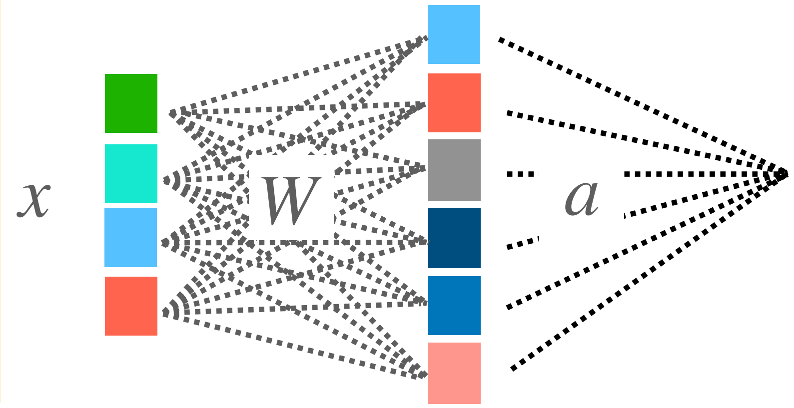
Let $(x_i, y_i)_{i\in[n]} \in \mathbb{R}^{d+1}$ be the training data. We assume:

$$y_i = f_\star(x_i) + z_i$$

$$x_i \sim \mathcal{N}(0, I_d/d) \qquad z_i \sim \mathcal{N}(0, \Delta)$$

We are interested in the performance of 2 layer NNs:

$$f(x; a, W) = \frac{1}{\sqrt{p}} \sum_{k=1}^{p} a_k \sigma\left(\langle w_k, x\rangle\right)$$



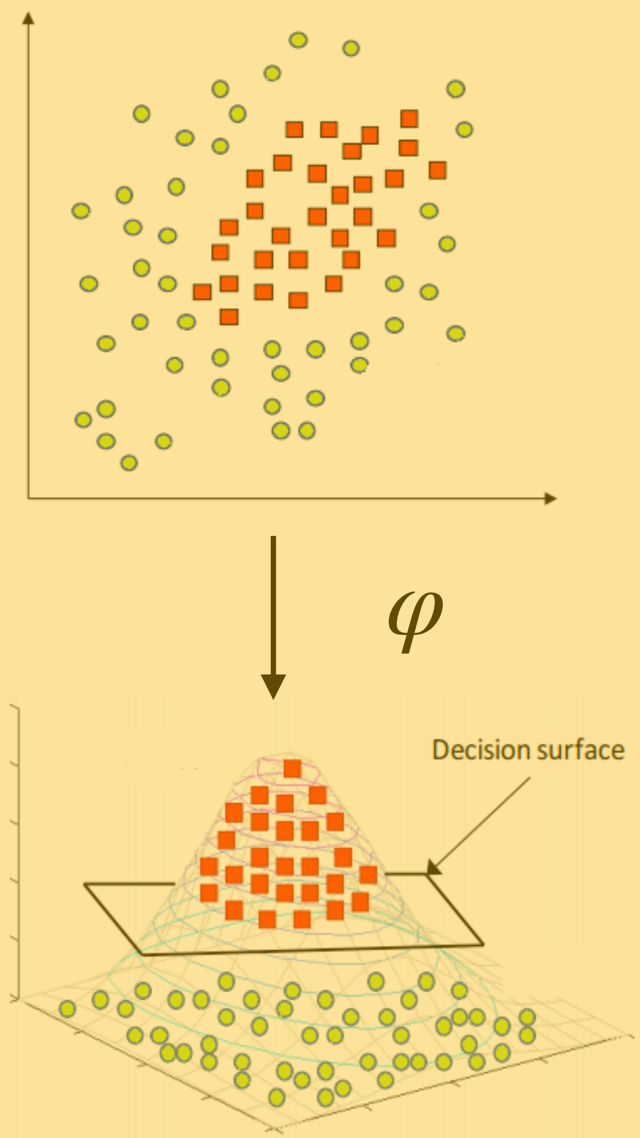
When trained over ERM:

$$\min_{a,W} \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i; a, W))^2 + \lambda r(a, W)$$

Today's menu

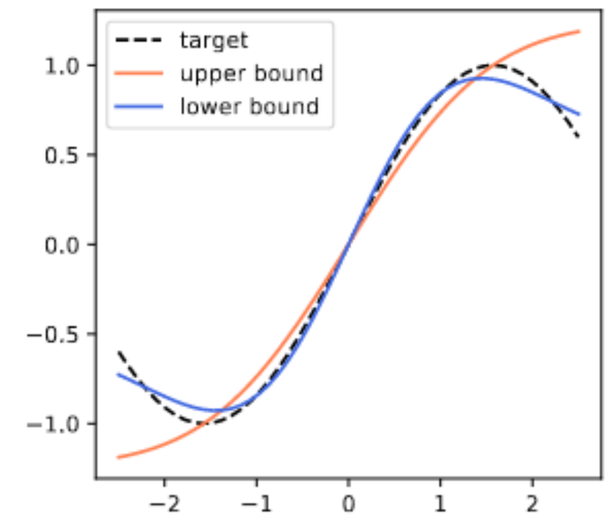
Initialization

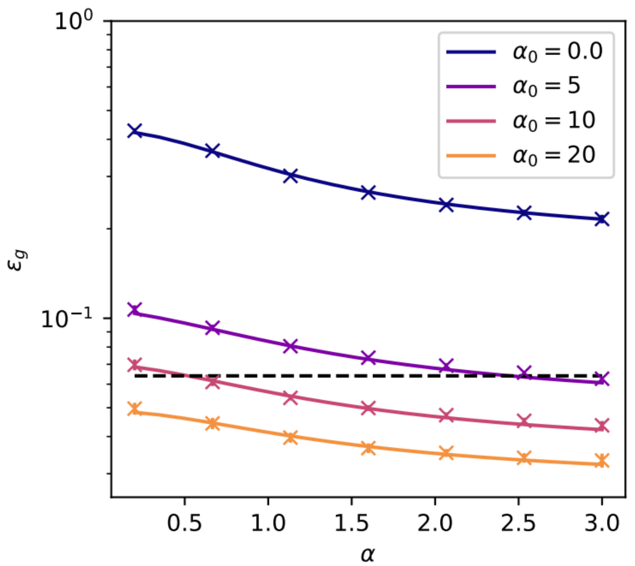
Random features and kernels



One step

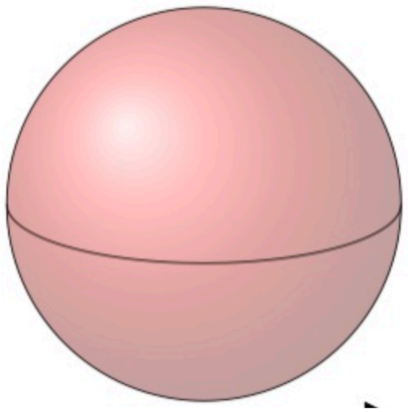
Exact asymptotics for one GD step





Few steps

Learning staircase functions



Initialisation

[Jacot, Gabriel, Hongler '18; Chizat, Bach '19; Neal '94; Lee et al. '19]

Start by looking at fixed W_0:

$$\hat{a}_\lambda(X, y) = \underset{a}{\text{argmin}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle a, \sigma(W^0 x_i) \rangle)^2 + \lambda ||a||_2^2$$

Initialisation

[Jacot, Gabriel, Hongler '18; Chizat, Bach '19;
Neal '94; Lee et al. '19]

Start by looking at fixed W_0:

$$\hat{a}_\lambda(X, y) = \underset{a}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle a, \sigma(W^0 x_i) \rangle)^2 + \lambda ||a||_2^2$$

a.k.a. as Random Features Model, which approximates a kernel method:

$$K_{\mathrm{RF}}(x, x') = \mathbb{E}_{w_0} \left[\sigma\left(\langle w^0, x \rangle\right) \sigma\left(\langle w^0, x' \rangle\right) \right] \approx \frac{1}{p} \sum_{k=1}^{p} \sigma\left(\langle w_k^0, x \rangle\right) \sigma\left(\langle w_k^0, x' \rangle\right)$$

[Retch, Raimi 2007]

Initialisation

[Jacot, Gabriel, Hongler '18; Chizat, Bach '19; Neal '94; Lee et al. '19]

Start by looking at fixed W_0:

$$\hat{a}_\lambda(X, y) = \underset{a}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle a, \sigma(W^0 x_i) \rangle)^2 + \lambda ||a||_2^2$$

a.k.a. as Random Features Model, which approximates a kernel method:

$$K_{\mathrm{RF}}(x, x') = \mathbb{E}_{w_0} \left[\sigma\left(\langle w^0, x \rangle\right) \sigma\left(\langle w^0, x' \rangle\right) \right] \approx \frac{1}{p} \sum_{k=1}^{p} \sigma\left(\langle w_k^0, x \rangle\right) \sigma\left(\langle w_k^0, x' \rangle\right)$$

[Retch, Raimi 2007]

🤔 What can we learn with that?

Mei, Montanari '19; Ghorbani, Mei, Misiakiewicz, Montanari '19, '20, '21; Gerace, **BL**, Krzakala, Mézard, Zdeborová '20; Goldt, **BL**, Reeves, Krzakala, Mézard, Zdeborová '21 Dhiffalah & Lu '20; Hu & Lu '20; Liang, Sur '20; Jacot, Simsek, Spadaro, Hongler, Gabriel '20; **BL**, Gerbelot, Refinetti, Sicuro, Krzakala '22; Mei, Misiakiewicz, Montanari '22; Fan, Wang 2020; Schröder, Cui, Dmitriev, **BL** '23, 24; Defilippis, **BL**, Misiakiewicz ;24

Limitations of RF

Theorem [Mei, Misiakiewicz, Montanari '22, informal]:

For isotropic data (e.g. $x \sim \mathrm{Unif}(\mathbb{S}^{d-1})$), with $n, p = \Theta(d^\kappa)$ one can learn at best a polynomial approximation of degree κ of the target $f_\star(x)$

$$\mathbb{E}\,||f_\star(x) - f(x; \hat{a}_\lambda, W^0)||_2^2 = ||P_{\leq\kappa} f_\star||_{L_2}^2 + o_d(1)$$

Limitations of RF

Theorem [Mei, Misiakiewicz, Montanari '22, informal]:

For isotropic data (e.g. $x \sim \mathrm{Unif}(\mathbb{S}^{d-1})$), with $n, p = \Theta(d^\kappa)$ one can learn at best a polynomial approximation of degree κ of the target $f_\star(x)$

$$\mathbb{E}\,||f_\star(x) - f(x; \hat{a}_\lambda, W^0)||_2^2 = ||P_{\leq\kappa} f_\star||_{L_2}^2 + o_d(1)$$

In particular, for $n, p = \Theta(d)$, can learn at best a linear approximation of $f_\star$

$$f_\star(x) = \langle \theta_\star, x \rangle + f_{NL}(x)$$

Limitations of RF

Theorem [Mei, Misiakiewicz, Montanari '22, informal]:

For isotropic data (e.g. $x \sim \text{Unif}(\mathbb{S}^{d-1})$), with $n, p = \Theta(d^\kappa)$ one can learn at best a polynomial approximation of degree κ of the target $f_\star(x)$

$$\mathbb{E}\,||f_\star(x) - f(x; \hat{a}_\lambda, W^0)||_2^2 = ||P_{\leq\kappa} f_\star||_{L_2}^2 + o_d(1)$$

In particular, for $n, p = \Theta(d)$, can learn at best a linear approximation of $f_\star$

$$f_\star(x) = \langle \theta_\star, x \rangle + f_{NL}(x)$$

Intuition:

$$\sigma(\langle w^0, x \rangle) = \mu_0 + \mu_1 \langle w^0, x \rangle + \sum_{\alpha \geq 2} \frac{\mu_\alpha}{\alpha!} \text{He}_\alpha(\langle w^0, x \rangle)$$

$$\mu_\alpha = \mathbb{E}[\text{He}_\alpha(z)\sigma(z)]$$

Limitations of RF

Theorem [Mei, Misiakiewicz, Montanari '22, informal]:

For isotropic data (e.g. $x \sim \mathrm{Unif}(\mathbb{S}^{d-1})$), with $n, p = \Theta(d^\kappa)$ one can learn at best a polynomial approximation of degree κ of the target $f_\star(x)$

$$\mathbb{E}\,||f_\star(x) - f(x; \hat{a}_\lambda, W^0)||_2^2 = ||P_{\leq\kappa} f_\star||_{L_2}^2 + o_d(1)$$

In particular, for $n, p = \Theta(d)$, can learn at best a linear approximation of $f_\star$

$$f_\star(x) = \langle \theta_\star, x\rangle + f_{NL}(x)$$

Intuition:

$$\sigma(\langle w^0, x\rangle) = \mu_0 + \mu_1\langle w^0, x\rangle + \sum_{\alpha\geq 2} \frac{\mu_\alpha}{\alpha!} \mathrm{He}_\alpha(\langle w^0, x\rangle)$$

$= \Theta(d^{-\alpha/2})$

$$\mu_\alpha = \mathbb{E}[\mathrm{He}_\alpha(z)\sigma(z)]$$

Limitations of RF

Theorem [Mei, Misiakiewicz, Montanari '22, informal]:

For isotropic data (e.g. $x \sim \mathrm{Unif}(\mathbb{S}^{d-1})$), with $n, p = \Theta(d^\kappa)$ one can learn at best a polynomial approximation of degree κ of the target $f_\star(x)$

$$\mathbb{E}\,||f_\star(x) - f(x; \hat{a}_\lambda, W^0)||_2^2 = ||P_{\leq\kappa} f_\star||_{L_2}^2 + o_d(1)$$

In particular, for $n, p = \Theta(d)$, can learn at best a linear approximation of $f_\star$

$$f_\star(x) = \langle \theta_\star, x \rangle + f_{NL}(x)$$

<u>Intuition</u>:

$$\sigma(\langle w^0, x \rangle) = \mu_0 + \mu_1 \langle w^0, x \rangle + \sum_{\alpha \geq 2} \frac{\mu_\alpha}{\alpha!} \underbrace{\mathrm{He}_\alpha(\langle w^0, x \rangle)}_{= \Theta(d^{-\alpha/2})}$$

$$\approx \mu_0 + \mu_1 \langle w, x \rangle + \mu_\star \xi$$

$$\mu_\alpha = \mathbb{E}[\mathrm{He}_\alpha(z)\sigma(z)] \qquad \mu_\star = \sqrt{\mathbb{E}[\sigma(z)^2] - \mu_0^2 - \mu_1^2}$$

Gaussian equivalence

Consider the following two ERM problems:

$$\hat{a}_\lambda(X, y) = \underset{a}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle a, \sigma(W^0 x_i) \rangle)^2 + \lambda ||a||_2^2$$

$$\hat{a}_\lambda^G(X, y) = \underset{a}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle a, \mu_0 1 + \mu_1 W^0 x_i + \mu_\star z_i \rangle)^2 + \lambda ||a||_2^2$$

Then, in the limit $d \to \infty$ with $n, p = \Theta(d)$:

Gaussian equivalence principle (GEP)

[Goldt et al. '19; Mei & Montanari '19; Hu& Lu '20]

$$|R(\hat{a}_\lambda) - R(\hat{a}_\lambda^G)| \to 0$$

Definitions:

Consider the unique fixed point of the following system of equations

$$\begin{cases} \hat{V}_s = \frac{\alpha}{\gamma}\kappa_1^2 \mathbb{E}_{\xi,y}\left[\mathcal{Z}\left(y,\omega_0\right)\frac{\partial_\omega \eta(y,\omega_1)}{V}\right], \\ \hat{q}_s = \frac{\alpha}{\gamma}\kappa_1^2 \mathbb{E}_{\xi,y}\left[\mathcal{Z}\left(y,\omega_0\right)\frac{\left(\eta(y,\omega_1)-\omega_1\right)^2}{V^2}\right], \\ \hat{m}_s = \frac{\alpha}{\gamma}\kappa_1 \mathbb{E}_{\xi,y}\left[\partial_\omega \mathcal{Z}\left(y,\omega_0\right)\frac{\left(\eta(y,\omega_1)-\omega_1\right)}{V}\right], \\ \hat{V}_w = \alpha\kappa_\star^2 \mathbb{E}_{\xi,y}\left[\mathcal{Z}\left(y,\omega_0\right)\frac{\partial_\omega \eta(y,\omega_1)}{V}\right], \\ \hat{q}_w = \alpha\kappa_\star^2 \mathbb{E}_{\xi,y}\left[\mathcal{Z}\left(y,\omega_0\right)\frac{\left(\eta(y,\omega_1)-\omega_1\right)^2}{V^2}\right], \end{cases}$$

$$\begin{cases} V_s = \frac{1}{\hat{V}_s}\left(1 - z\ g_\mu(-z)\right), \\ q_s = \frac{\hat{m}_s^2+\hat{q}_s}{\hat{V}_s}\left[1 - 2z g_\mu(-z) + z^2 g_\mu'(-z)\right] \\ \qquad -\frac{\hat{q}_w}{(\lambda+\hat{V}_w)\hat{V}_s}\left[-z g_\mu(-z) + z^2 g_\mu'(-z)\right], \\ m_s = \frac{\hat{m}_s}{\hat{V}_s}\left(1 - z\ g_\mu(-z)\right), \\ V_w = \frac{\gamma}{\lambda+\hat{V}_w}\left[\frac{1}{\gamma} - 1 + z g_\mu(-z)\right], \\ q_w = \gamma\frac{\hat{q}_w}{(\lambda+\hat{V}_w)^2}\left[\frac{1}{\gamma} - 1 + z^2 g_\mu'(-z)\right], \\ \qquad +\frac{\hat{m}_s^2+\hat{q}_s}{(\lambda+\hat{V}_w)\hat{V}_s}\left[-z g_\mu(-z) + z^2 g_\mu'(-z)\right], \end{cases}$$

$$\begin{cases} \eta(y,\omega) = \underset{x\in\mathbb{R}}{\operatorname{argmin}}\left[\frac{(x-\omega)^2}{2V} + \ell(y,x)\right] \\ \mathcal{Z}(y,\omega) = \int \frac{dx}{\sqrt{2\pi V^0}} e^{-\frac{1}{2V^0}(x-\omega)^2}\delta\left(y - f^0(x)\right) \end{cases}$$

where $V = \kappa_1^2 V_s + \kappa_\star^2 V_w$, $V^0 = \rho - \frac{M^2}{Q}$, $Q = \kappa_1^2 q_s + \kappa_\star^2 q_w$, $M = \kappa_1 m_s$, $\omega_0 = M/\sqrt{Q}\xi$, $\omega_1 = \sqrt{Q}\xi$

and g_μ is the Stieltjes transform of $W_0 W_0^T \mu_0 = \mathbb{E}\left[\sigma(z)\right], \mu_1 \equiv \mathbb{E}\left[z\sigma(z)\right], \mu_\star \equiv \mathbb{E}\left[\sigma(z)^2\right] - \mu_0^2 - \mu_1^2$, and $z \sim \mathcal{N}(0,1)$

In the high-dimensional limit:

$$R(\hat{a}_\lambda) = \mathbb{E}_{\lambda,\nu}\left[(f^0(\nu) - \hat{f}(\lambda))^2\right]$$

with $(\nu,\lambda) \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\rho & M^\star \\ M^\star & Q^\star\end{pmatrix}\right)$

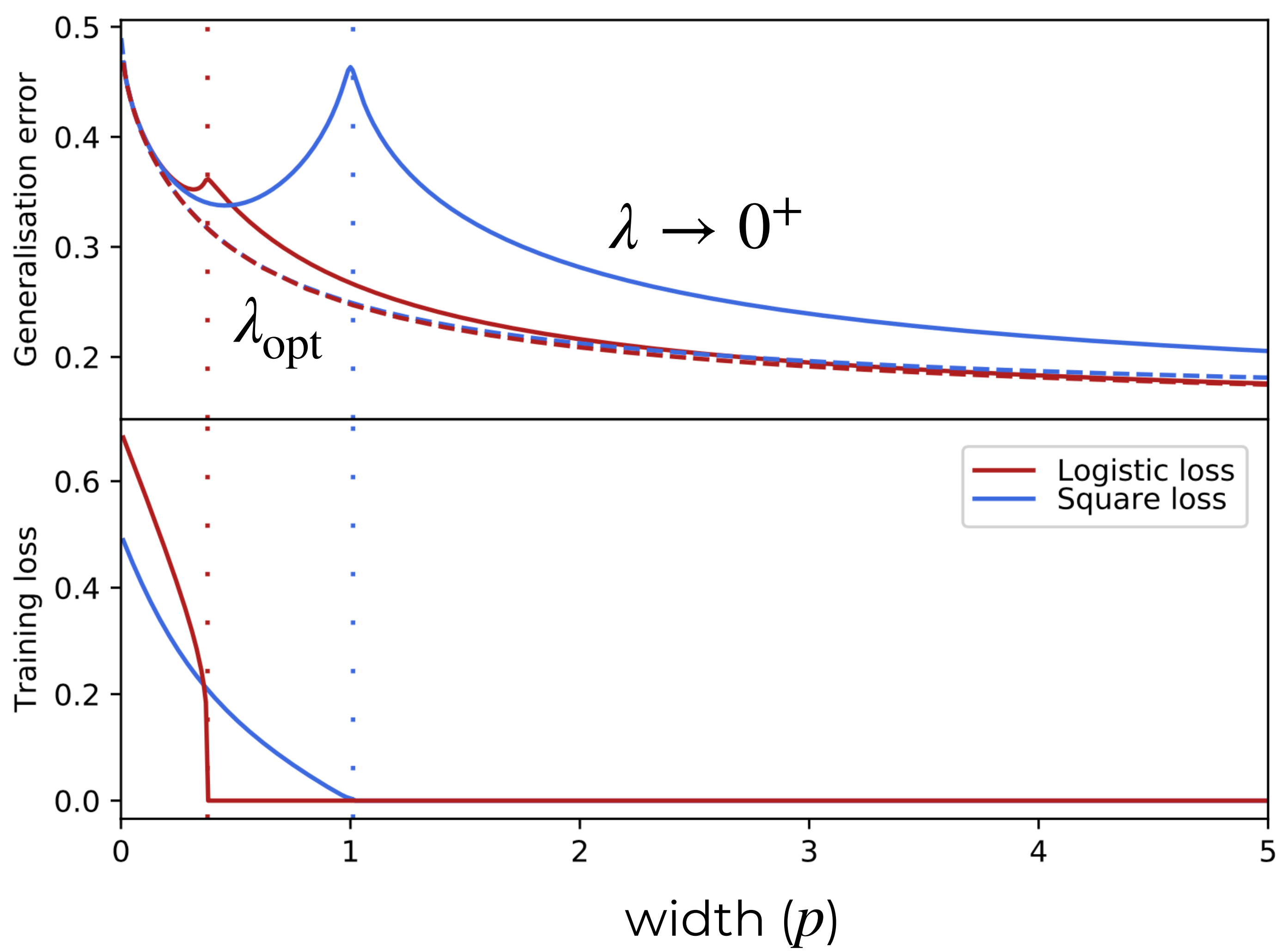
$$\hat{R}_n(\hat{a}_\lambda) = \frac{\lambda}{2\alpha}q_w^\star + \mathbb{E}_{\xi,y}\left[\mathcal{Z}\left(y,\omega_0^\star\right)\ell\left(y,\eta(y,\omega_1^\star)\right)\right]$$

with $\omega_0^\star = M^\star/\sqrt{Q^\star}\xi$, $\omega_1^\star = \sqrt{Q^\star}\xi$

Gaussian equivalence

Gaussian equivalence principle (GEP)
[Goldt et al. '19; Mei & Montanari '19; Hu& Lu '20]

$$|R(\hat{a}_\lambda) - R(\hat{a}^G_\lambda)| \to 0$$

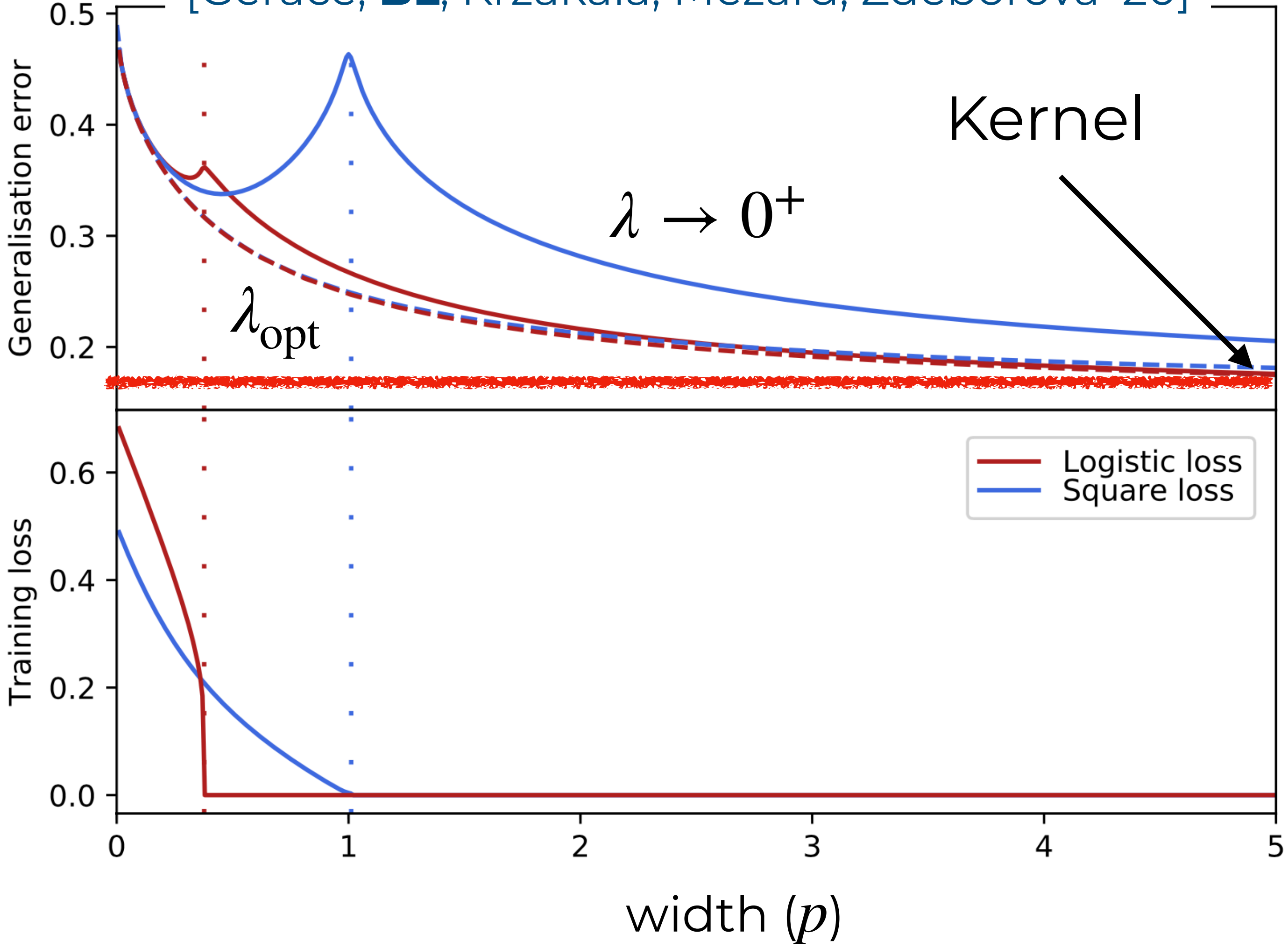


Gaussian equivalence

Gaussian equivalence principle (GEP)
[Goldt et al. '19; Mei & Montanari '19; Hu& Lu '20]

$$| R(\hat{a}_\lambda) - R(\hat{a}^G_\lambda) | \to 0$$

[Gerace, **BL**, Krzakala, Mézard, Zdeborová '20]



Gaussian universality

↕

Equivalence to a linear model

↕

Limited expressivity

Partial Summary

Kernels/RF are able to learn “anything”, but they need “a lot” of data.

Partial Summary

Kernels/RF are able to learn "anything", but they need "a lot" of data.

In particular, with $n, p = \Theta(d)$, only learn linear functions.

Partial Summary

Kernels/RF are able to learn "anything", but they need "a lot" of data.

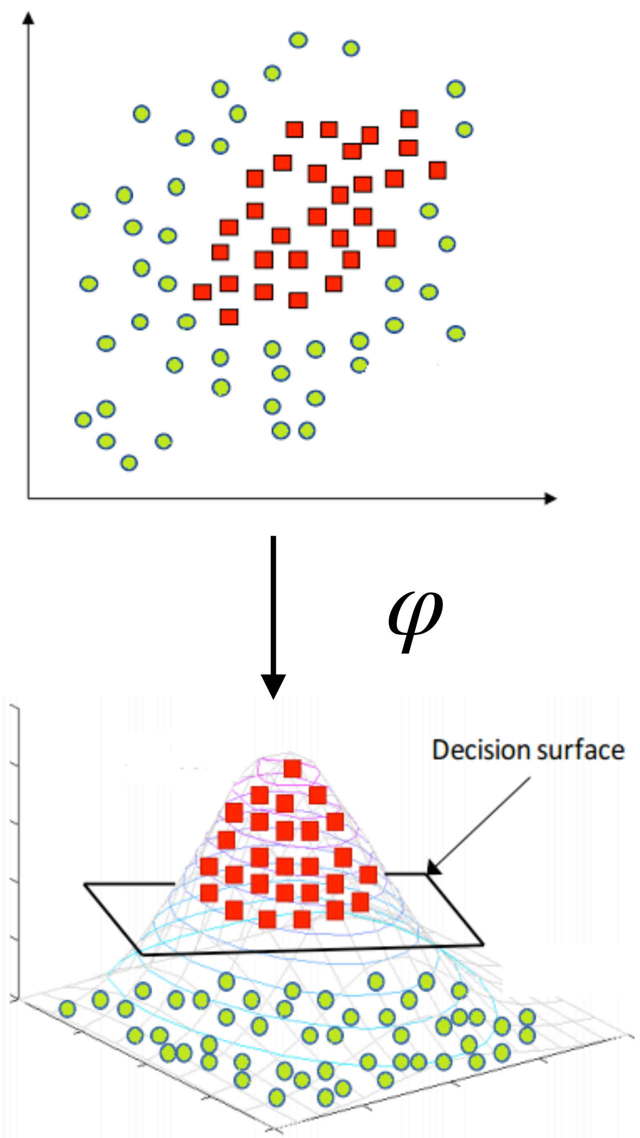
In particular, with $n, p = \Theta(d)$, only learn linear functions.

To do better, need to learn features.

Today's menu

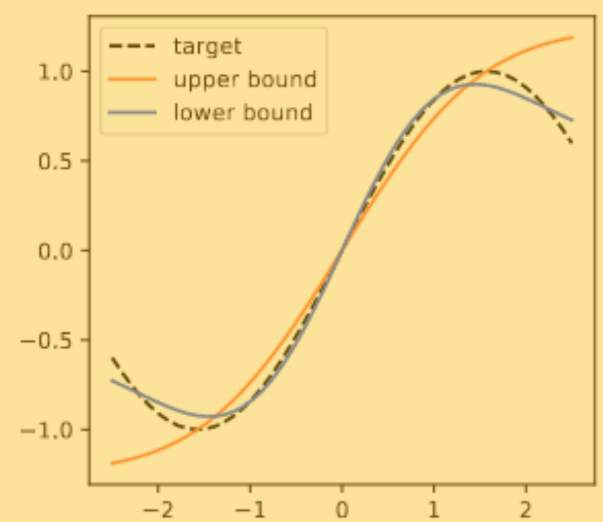
Initialization

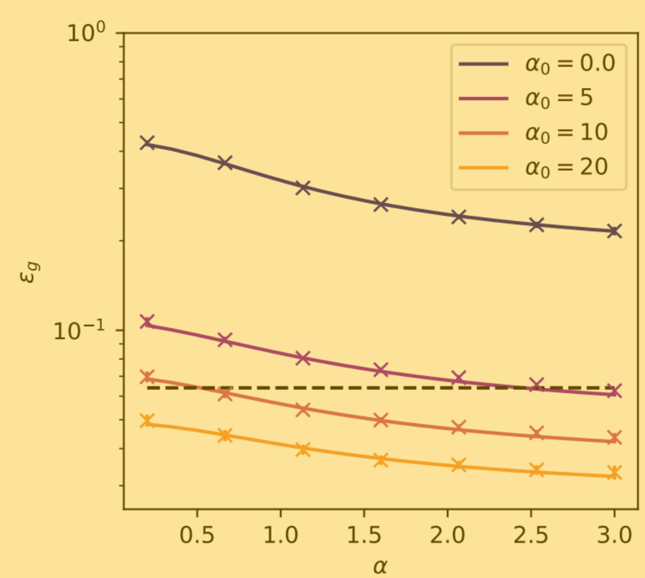
Random features and kernels



One step

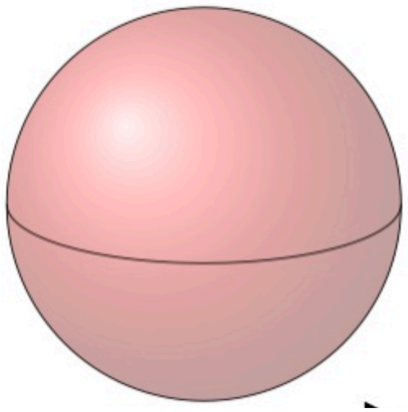
Exact asymptotics for one GD step





Few steps

Learning staircase functions



One step of GD

Consider one step of GD from initialisation a^0, W^0 with fresh batch $(x_i, y_i)_{i \in [n_0]}$

$$W^1 = W^0 - \frac{\eta}{2n} \sum_{i=1}^{n_B} \nabla_w (y_i - f(x_i; a^0, W^0))^2$$

One step of GD

Consider one step of GD from initialisation a^0, W^0 with fresh batch $(x_i, y_i)_{i \in [n_0]}$

$$W^1 = W^0 - \frac{\eta}{2n} \sum_{i=1}^{n_B} \nabla_w (y_i - f(x_i; a^0, W^0))^2$$

🤔 Can we learn more than $f_\star(x) = \langle \theta_\star, x \rangle$?

One step of GD

Consider one step of GD from initialisation a^0, W^0 with fresh batch $(x_i, y_i)_{i \in [n_0]}$

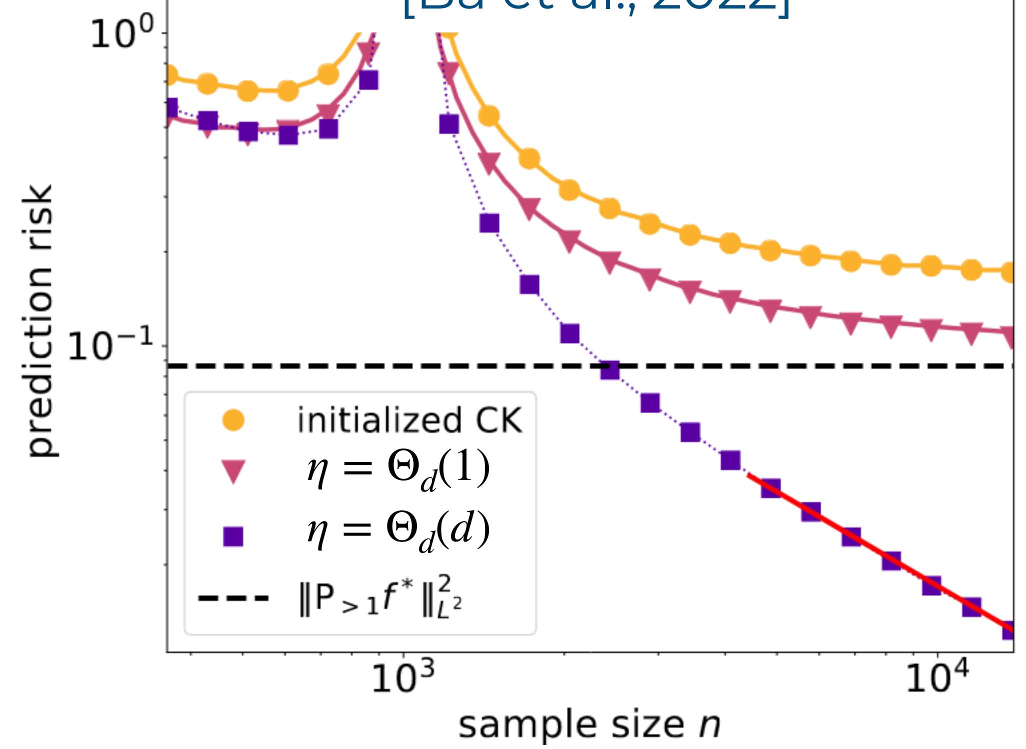
$$W^1 = W^0 - \frac{\eta}{2n} \sum_{i=1}^{n_B} \nabla_w (y_i - f(x_i; a^0, W^0))^2$$

🤔 Can we learn more than $f_\star(x) = \langle \theta_\star, x \rangle$?

- For $n, p = \Theta(d)$ and $\eta = \Theta(1)$, no! GEP still valid.
- $\eta = \Theta_d(d)$ sufficient to learn more.

Can we characterise what?

[Ba et al., 2022]



What you learn in one-step of SGD?

Consider a multi-index model, $\sqrt{p}a^0 \sim \mathrm{Unif}([-1,+1])$, η large enough.

$$f_\star(x) = g(\langle w_1^\star, x\rangle, \cdots, \langle w_r^\star, x\rangle)$$

$$g : \mathbb{R}^r \to \mathbb{R} \quad w_k^\star \in \mathbb{S}^{d-1}(\sqrt{d})$$

$$\frac{\langle w_i^1, w_k^\star\rangle}{||w_i^1|| \cdot ||w_k^\star||} \overset{d\to\infty}{>} 0$$

What you learn in one-step of SGD?

Consider a multi-index model, $\sqrt{p}a^0 \sim \mathrm{Unif}([-1,+1])$, η large enough.

$$f_\star(x) = g(\langle w_1^\star, x\rangle, \cdots, \langle w_r^\star, x\rangle)$$
$$g : \mathbb{R}^r \to \mathbb{R} \quad w_k^\star \in \mathbb{S}^{d-1}(\sqrt{d})$$

$$\frac{\langle w_i^1, w_k^\star\rangle}{||w_i^1|| \cdot ||w_k^\star||} \overset{d\to\infty}{>} 0$$

Key idea: Hermite tensor decomposition

$$g(z_1, \ldots, z_r) = \mu_0 + \sum_i \mu_i^{(1)} z_i + \sum_{ij} \mu_{ij}^{(2)} h_2(z_i) h_2(z_j) + \ldots$$

Hardness ≈ large leap

What you learn in one-step of SGD?

Consider a multi-index model, $\sqrt{p}a^0 \sim \mathrm{Unif}([-1, +1])$, η large enough.

$$f_\star(x) = g(\langle w_1^\star, x\rangle, \cdots, \langle w_r^\star, x\rangle)$$
$$g : \mathbb{R}^r \to \mathbb{R} \quad w_k^\star \in \mathbb{S}^{d-1}(\sqrt{d})$$

$$\frac{\langle w_i^1, w_k^\star\rangle}{||w_i^1|| \cdot ||w_k^\star||} \overset{d\to\infty}{>} 0$$

Key idea: Hermite tensor decomposition

$$g(z_1, \ldots, z_r) = \mu_0 + \sum_i \mu_i^{(1)} z_i + \sum_{ij} \mu_{ij}^{(2)} h_2(z_i) h_2(z_j) + \ldots$$

Hardness ≈ large leap

Examples:

$$g(z) = z_1 + z_1 z_2 + z_1 z_2 z_3 \qquad \ell = 1$$

$$g(z) = \mathrm{He}_k(z_1) \qquad \ell = k$$

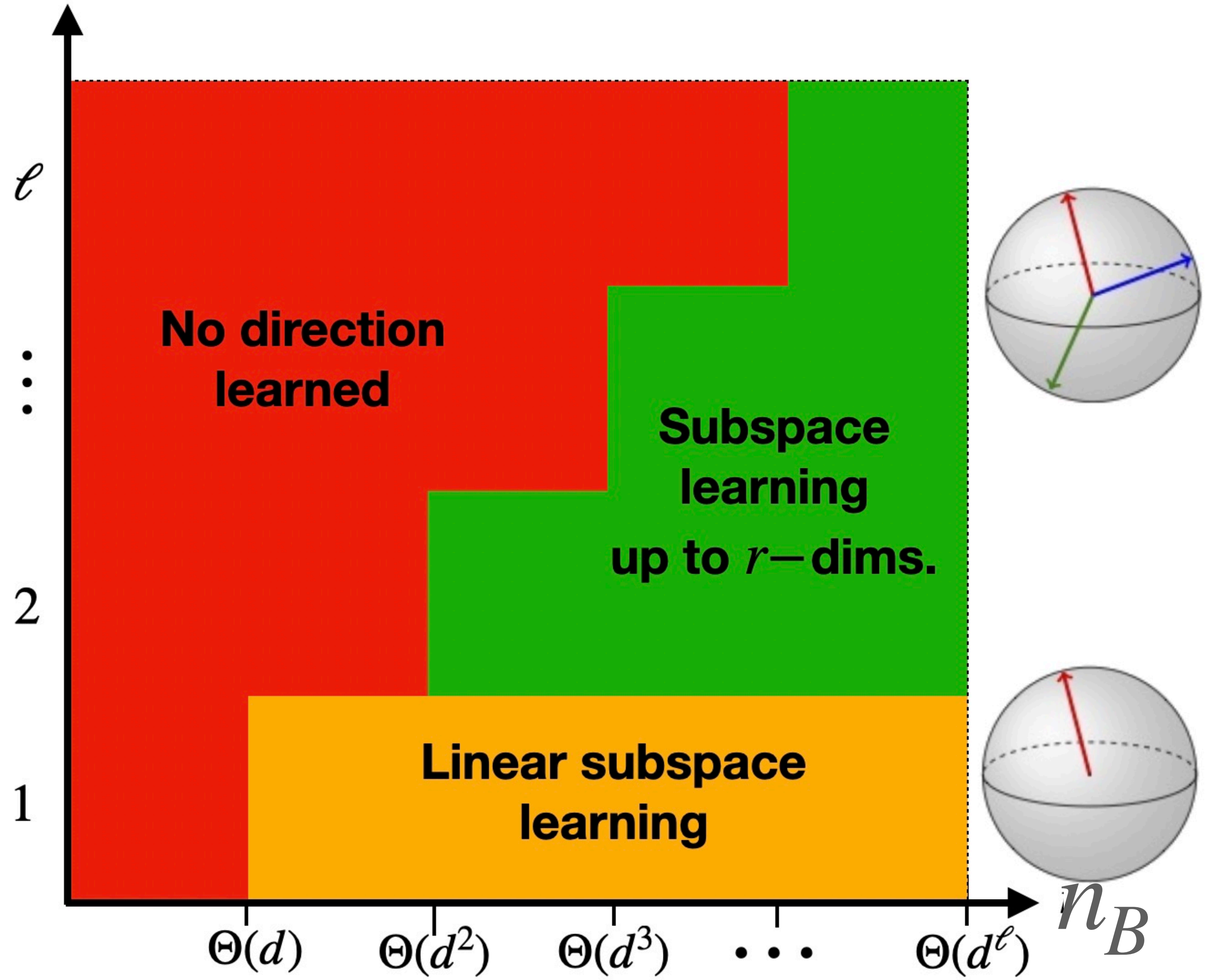
$$g(z) = z_1 z_2 z_3 z_4 \qquad \ell = 4$$

What you learn in one-step of SGD?

Consider a multi-index model, $\sqrt{p}a^0 \sim \mathrm{Unif}([-1, +1])$, η large enough.

$$f_\star(x) = g(\langle w_1^\star, x\rangle, \cdots, \langle w_r^\star, x\rangle)$$
$$g : \mathbb{R}^r \to \mathbb{R} \quad w_k^\star \in \mathbb{S}^{d-1}(\sqrt{d})$$

$$\frac{\langle w_i^1, w_k^\star\rangle}{||w_i^1|| \cdot ||w_k^\star||} \overset{d\to\infty}{>} 0$$



What you learn in one-step of SGD?

Theorem 1. *Let ℓ be the leap index of $f^\star$ equation 1, and assume that $n = \mathcal{O}(d^{\ell-\delta})$ for some $\delta > 0$. Then, with probability at least $1 - cpe^{-c(\delta)\log(d)^2}$, there exists a universal constant c such that for any $i \in [p]$,*

$$\frac{\langle w_i^{t=1}, w_k^\star \rangle}{||w_i^{t=1}|| \cdot ||w_k^\star||} \leq c\,\frac{\mathrm{polylog}(d)}{d^{(1\wedge\delta)/2}}. \tag{7}$$

In other words, for *every* neuron i, only a vanishing fraction of the weight $\boldsymbol{w}_i^1$ lies in the target subspace $V^\star$. In particular, if $\delta > 1$, this large gradient step does not improve over the initial random feature weights.

On the other hand, when $n = \Omega(d^\ell)$, we are able to characterize exactly what is being learned:

Theorem 2. *Assume that the ℓ-th Hermite coefficient μ_ℓ of σ is nonzero, and set the learning rate $\eta = pd^{\frac{\ell-1}{2}}$. Then, with probability at least $1 - ce^{-c\log(d)^2}$, there exists a random variable X independent of d with positive expectation such that*

$$\frac{\langle w_i^{t=1}, w_k^\star \rangle}{||w_i^{t=1}|| \cdot ||w_k^\star||} \geq X_i, \tag{8}$$

where $X_1, \ldots, X_p$ are i.i.d copies of X. Further, let $\boldsymbol{u}_1^\star, \ldots, \boldsymbol{u}_{r_\ell}^\star$ be the higher-order singular vectors of $C_\ell^\star$, and define $V_\ell^\star = \mathrm{span}(\boldsymbol{u}_1^\star, \ldots, \boldsymbol{u}_{r_\ell}^\star)$. Then, the projections $\boldsymbol{\pi}_i^1$ asymptotically belong to $V_\ell^\star$, in the sense that there exists a constant c such that

$$\|(I - \Pi_{V_\ell^\star})\boldsymbol{\pi}_i^1\| \leq c\,\frac{\mathrm{polylog}(d)}{\sqrt{d}}, \tag{9}$$

and they span the space $V_\ell^\star$.

[Dandi, Krzakala, **BL**, Pesce, Stephan '23]

[Damian, Lee, Soltanolkotabi '22] implies the positive part of (ii) for $n = O(d^2)$
[Ba, Erdogdu, Suzuki, Wang, Wu, Yang '22] proved a rank-one property for single index teacher for $n = O(d)$ in (i)

Partial Summary

With a single gradient step and

$$n, p, \eta = \Theta(d)$$

can learn at best a non-linear function of one direction

$$f_\star(x) = g(\langle \theta_\star, x \rangle)$$

Partial Summary

With a single gradient step and

$$n, p, \eta = \Theta(d)$$

can learn at best a non-linear function of one direction

$$f_\star(x) = g(\langle \theta_\star, x \rangle)$$

🤔 Can we get sharp asymptotics for the error?

Mapping to a sRF model

After a single gradient step with $n, p, \gamma = \Theta(d)$:

$$W^1 = W^0 - \frac{\eta}{2n} \sum_{i=1}^{n_B} \nabla_w (g(\langle \theta_\star, x_i \rangle) - f(x_i; a^0, W^0))^2$$

Mapping to a sRF model

After a single gradient step with $n, p, \gamma = \Theta(d)$:

$$W^1 = W^0 - \frac{\eta}{2n} \sum_{i=1}^{n_B} \nabla_w (g(\langle \theta_\star, x_i \rangle) - f(x_i; a^0, W^0))^2$$

We can decompose:

$$W^1 = W^0 + \breve{u}\breve{v} + \Delta$$

[Ba et al., '22]

$$\breve{u} = \eta \mu_1 a^0 \in \mathbb{R}^p \qquad \breve{v} = \frac{1}{n_B} \sum_{i=1}^{n_B} \breve{\sigma}(W^0 x_i) g(\langle \theta_\star, x_i \rangle) x_i \in \mathbb{R}^d$$

$$\breve{\sigma}(z) = \sigma(z) - \mu_1$$

$$\mu_1 = \mathbb{E}[\sigma(z) z]$$

Taking $a^0 = 1_p$, after some massage...

Mapping to a sRF model

After a single gradient step with $n, p, \gamma = \Theta(d)$:

$$W^1 = W^0 - \frac{\eta}{2n} \sum_{i=1}^{n_B} \nabla_w (g(\langle \theta_\star, x_i \rangle) - f(x_i; a^0, W^0))^2$$

We can decompose:

$$W^1 = W + ruv$$

$$w_k \in \mathbb{S}^{d-1}(\sqrt{c})$$
$$u \in \mathbb{S}^{d-1}(\sqrt{p})$$
$$v \in \mathbb{S}^{d-1}$$

$$r = \frac{\eta}{d}\frac{p}{d}\mu_1 \sqrt{\frac{d}{n_B}\mu_2^\star + \mu_1^{\star 2}} \qquad c = 1 + \frac{\eta^2 d}{n_B p^2}\mu_1^2 \breve{\mu}_1^2 \mu_2^\star \qquad \langle v, \theta_\star \rangle = \frac{\mu_1^\star}{\sqrt{\frac{d}{n_B}\mu_2^\star + \mu_1^{\star 2}}}$$

$$\mu_1 = \mathbb{E}[\sigma(z)z] \qquad \mu_2 = \mathbb{E}[\sigma(z)^2] \qquad \breve{\mu}_1^2 = \mathbb{E}[(\sigma(z)z - \mu_1)^2]$$

"Spiked Random Features"

[Cui, Pesce, Dandi, Krzakala, Lu, Zdeborová, **BL** '24]

Conditional GEP

Recall that for the standard RF model

Gaussian Equivalence Theorem (GET)

$$\sigma\left(\langle w^0, x\rangle\right) \approx \mu_0 + \mu_1 \langle w^0, x\rangle + \mu_\star \xi$$

[Goldt et al. 19; Mei, Montanari '19; Hu & Lu '20]

Conditional GEP

Recall that for the standard RF model

Gaussian Equivalence Theorem (GET)

$$\sigma\left(\langle w^0, x\rangle\right) \approx \mu_0 + \mu_1\langle w^0, x\rangle + \mu_\star \xi$$

[Goldt et al. 19; Mei, Montanari '19; Hu & Lu '20]

We can show that for a sRF model with $a^0 = 1_p$:

cGET [Dandi, Krzakala, **BL**, Pesce, Stephan '23]

$$\sigma\left(\langle w^1, x\rangle\right) \approx \mu_0(\langle v, x\rangle) + \mu_1(\kappa)\langle w^0, x^\perp\rangle + \mu_\star(\kappa)\xi$$

$$\kappa = \langle v, x\rangle \qquad x = \kappa\theta_\star + x^\perp$$

Conditional GEP

Recall that for the standard RF model

Gaussian Equivalence Theorem (GET)

$$\sigma\left(\langle w^0, x\rangle\right) \approx \mu_0 + \mu_1 \langle w^0, x\rangle + \mu_\star \xi$$

[Goldt et al. 19;
Mei, Montanari '19;
Hu & Lu '20]

We can show that for a sRF model with $a^0 = 1_p$:

cGET [Dandi, Krzakala, **BL**, Pesce, Stephan '23]

$$\sigma\left(\langle w^1, x\rangle\right) \approx \mu_0(\langle v, x\rangle) + \mu_1(\kappa)\langle w^0, x^\perp\rangle + \mu_\star(\kappa)\xi$$

$$\kappa = \langle v, x\rangle \qquad x = \kappa\theta_\star + x^\perp$$

Examples: $\sigma(z) = \mathrm{sign}$

$$\mu_0(\kappa) = \mathrm{erf}\left(\frac{\kappa}{\sqrt{2}}\right) \quad \mu_1(\kappa) = \sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}\kappa^2}$$

$$\mu_2(\kappa) = 1 - \mu_0(\kappa)^2 - \mu_1(\kappa)^2$$

Main result

Together, this allow us to characterise the risk:

$$R(\hat{a}_\lambda) = \mathbb{E}[(g(\langle \theta_\star, x \rangle) - \langle \hat{a}_\lambda, \sigma(W^1 x_i))^2]$$

Where:

$$\hat{a}_\lambda(X, y) = \underset{a}{\text{argmin}} \frac{1}{2n} \sum_{i=1}^{n} (g(\langle \theta_\star, x_i \rangle - \langle a, \sigma(W^1 x_i))^2 + \lambda ||a||_2^2$$

Main result

Together, this allow us to characterise the risk:

$$R(\hat{a}_\lambda) = \mathbb{E}[(g(\langle \theta_\star, x \rangle) - \langle \hat{a}_\lambda, \sigma(W^1 x_i))^2]$$

Where:

$$\hat{a}_\lambda(X, y) = \underset{a}{\text{argmin}} \frac{1}{2n} \sum_{i=1}^{n} (g(\langle \theta_\star, x_i \rangle - \langle a, \sigma(W^1 x_i))^2 + \lambda ||a||_2^2$$

More precisely, for $a^0 = 1_p$ in the limit $d \to \infty$ with $n, p, \eta = \Theta(d)$:

$$R(\hat{a}_\lambda) = \mathbb{E}_{\kappa,z}\left[\left(g\left(\gamma\kappa + \sqrt{1-\gamma^2}z\right) - \mu_0(\kappa)m - \mu_1(\kappa)\kappa\zeta - \frac{\mu_1(\kappa)\psi}{\sqrt{\rho}}z\right)^2 + \mu_1(\kappa)^2 q_1 + \mu_2(\kappa)^2 q_2 - \frac{\mu_1(\kappa)^2\psi^2}{\rho}\right]$$

$$m = \frac{1^\top \hat{a}_\lambda}{\sqrt{p}} \qquad q_1 = \frac{\langle W^\top \hat{a}_\lambda, \Pi^\perp W^\top \hat{a}_\lambda \rangle}{p} \qquad q_2 = \frac{||\hat{a}_\lambda||_2^2}{p} \qquad \zeta = \frac{\langle \hat{a}_\lambda, Wv \rangle}{\sqrt{dp}}$$

Exact asymptotics ($a^0 = 1_p$)

$$\begin{cases} V_1 = \int \frac{d\nu(\varrho,\tau,\pi)\varrho}{\lambda + \hat{V}_1\varrho + \hat{V}_2} \\ V_2 = \int \frac{d\nu(\varrho,\tau,\pi)}{\lambda + \hat{V}_1\varrho + \hat{V}_2} \\ m = \frac{\mathbb{E}_{\kappa,y}\left[\frac{\mu_0(\kappa)(\sigma_\star(\kappa,y) - \mu_1(\kappa)\kappa\zeta)}{1+V(\kappa)}\right]}{\mathbb{E}_\kappa\left[\frac{\mu_0(\kappa)^2}{1+V(\kappa)}\right]} \\ \zeta = \hat{\zeta}\sqrt{\beta}\int d\nu(\varrho,\tau,\pi)\varrho\tau^2 \frac{1}{\lambda + \hat{V}_1\varrho + \hat{V}_2} + \beta^{\frac{3}{2}}\hat{\zeta}\hat{V}_1 \frac{I(\hat{V}_1,\hat{V}_2)^2}{1 - \beta\hat{V}_1 I(\hat{V}_1,\hat{V}_2)} \\ \psi = \hat{\psi}\sqrt{\beta}\int \frac{d\nu(\varrho,\tau,\pi)\varrho\pi^2}{\lambda + \hat{V}_1\varrho + \hat{V}_2} \end{cases}$$

$$\begin{cases} q_1 = \int d\nu(\varrho,\tau,\pi)\varrho \frac{\left(\hat{q}_1\varrho + \hat{q}_2 + \hat{\zeta}^2\varrho\tau^2 + \hat{\psi}^2\varrho\pi^2\right)}{\left(\lambda + \hat{V}_1\varrho + \hat{V}_2\right)^2} - \beta\hat{\zeta}^2 \frac{I(\hat{V}_1,\hat{V}_2)^2}{\left(1 - \beta\hat{V}_1 I(\hat{V}_1,\hat{V}_2)\right)^2} \\ \qquad -\hat{\zeta}^2 \frac{\int \frac{\tau^2\varrho^2 d\nu(\varrho,\tau,\pi)}{(\lambda + \hat{V}_1\varrho + \hat{V}_2)^2}\left[\left(1 - \beta\hat{V}_1 I(\hat{V}_1,\hat{V}_2)\right)^2 - 1\right]}{\left(1 - \beta\hat{V}_1 I(\hat{V}_1,\hat{V}_2)\right)^2} \\ q_2 = \int \frac{\left(\hat{q}_1\varrho + \hat{q}_2 + \hat{\zeta}^2\varrho\tau^2 + \hat{\psi}^2\varrho\pi^2\right)d\nu(\varrho,\tau,\pi)}{\left(\lambda + \hat{V}_1\varrho + \hat{V}_2\right)^2} \\ \qquad -\hat{\zeta}^2 \int \frac{\tau^2\varrho d\nu(\varrho,\tau,\pi)}{(\lambda + \hat{V}_1\varrho + \hat{V}_2)^2}\left[1 - \frac{1}{\left(1 - \beta\hat{V}_1 I(\hat{V}_1,\hat{V}_2)\right)^2}\right] \end{cases}$$

$$\begin{cases} \hat{V}_1 = \frac{\alpha}{\beta}\mathbb{E}_\kappa \frac{\rho\mu_1(\kappa)^2}{1+V(\kappa)} \\ \hat{V}_2 = \frac{\alpha}{\beta}\mathbb{E}_\kappa \frac{\rho\mu_2(\kappa)^2}{1+V(\kappa)} \\ \hat{\zeta} = \frac{\alpha}{\sqrt{\beta}}\mathbb{E}_{\kappa,y}\kappa\mu_1(\kappa)\frac{b(\kappa,y)}{1+V(\kappa)} \\ \hat{\psi} = \frac{\alpha}{\sqrt{\beta}}\mathbb{E}_{\kappa,y}\frac{y\mu_1(\kappa)b(\kappa,y) + \psi\mu_1(\kappa)^2}{1+V(\kappa)} \end{cases}$$

$$\begin{cases} \hat{q}_1 = \frac{\alpha}{\beta}\mathbb{E}_{\kappa,y}\mu_1(\kappa)^2 \frac{b(\kappa,y)^2 + \rho q(\kappa) - \mu_1(\kappa)^2\psi^2}{\left(1+V(\kappa)\right)^2} \\ \hat{q}_2 = \frac{\alpha}{\beta}\mathbb{E}_{\kappa,y}\mu_2(\kappa)^2 \frac{b(\kappa,y)^2 + \rho q(\kappa) - \mu_1(\kappa)^2\psi^2}{\left(1+V(\kappa)\right)^2} \end{cases}$$

$$\alpha_0 = n_B/d \qquad \beta = p/d$$
$$\alpha = n/d \qquad \tilde{\eta} = \eta/d$$
$$\kappa = \langle v, x\rangle \qquad \rho = 1 - \gamma^2$$
$$\gamma = \langle v, \theta_\star\rangle$$

$$W = \sum_{i=1}^{\min(p,d)} \lambda_i e_i f_i^\top \qquad \Pi^\perp = I_d - vv^\top$$

$$\nu(\varrho,\tau,\pi) = \frac{1}{p}\sum_{i=1}^{\min(p,d)} \delta\left(\lambda_i - \varrho\right)\delta\left(f_i^\top v - \tau\right)\delta\left(f_i^\top \Pi^\perp \vec{\theta} - \pi\right)$$

Partial Summary

Single step of GD can be approximated by a spiked RF model

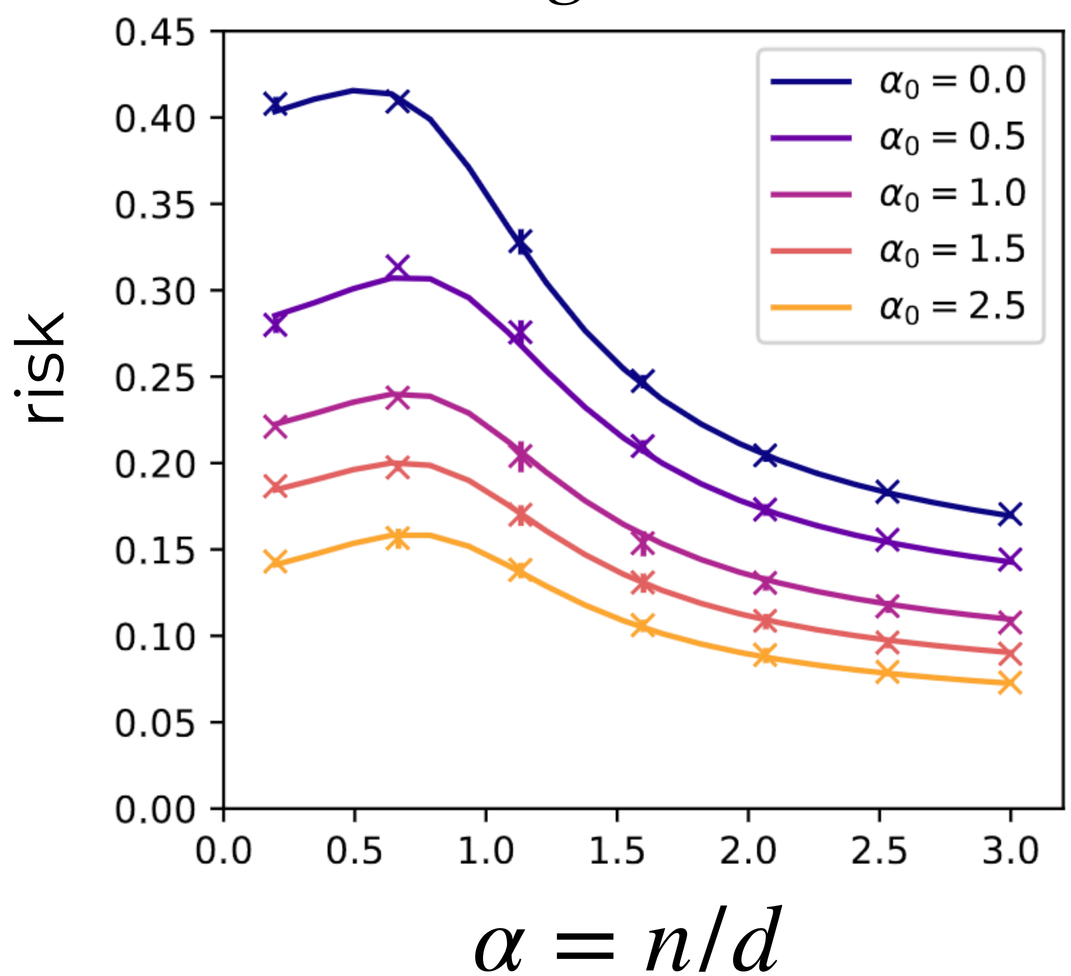
Conditional GET allow us to handle non-linearity.

Can derive a sharp asymptotic description of the error.

Batch size

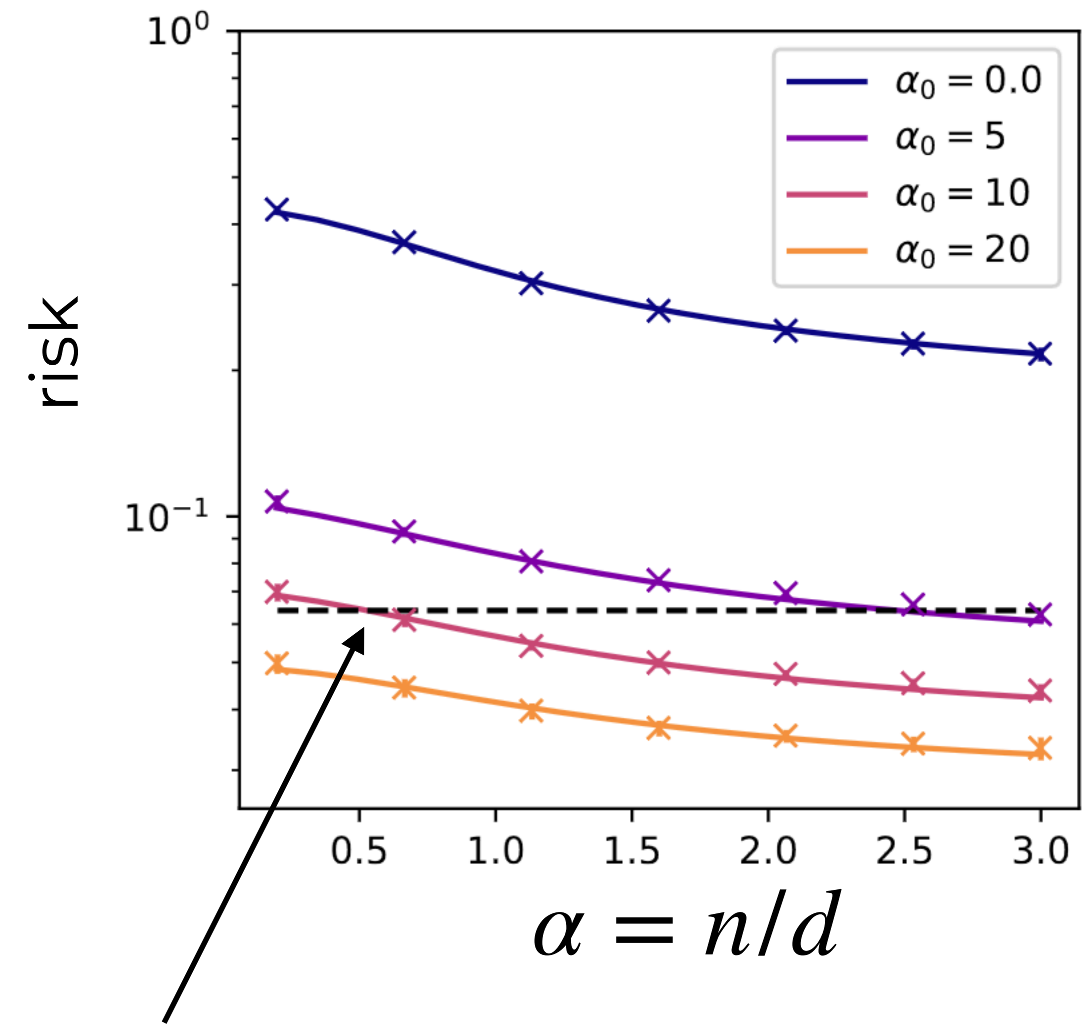
$\tilde{\eta} = 1 \qquad \lambda = 10^{-2}$

$\sigma = g = \tanh$



$\tilde{\eta} = 3 \qquad \lambda = 0.1$

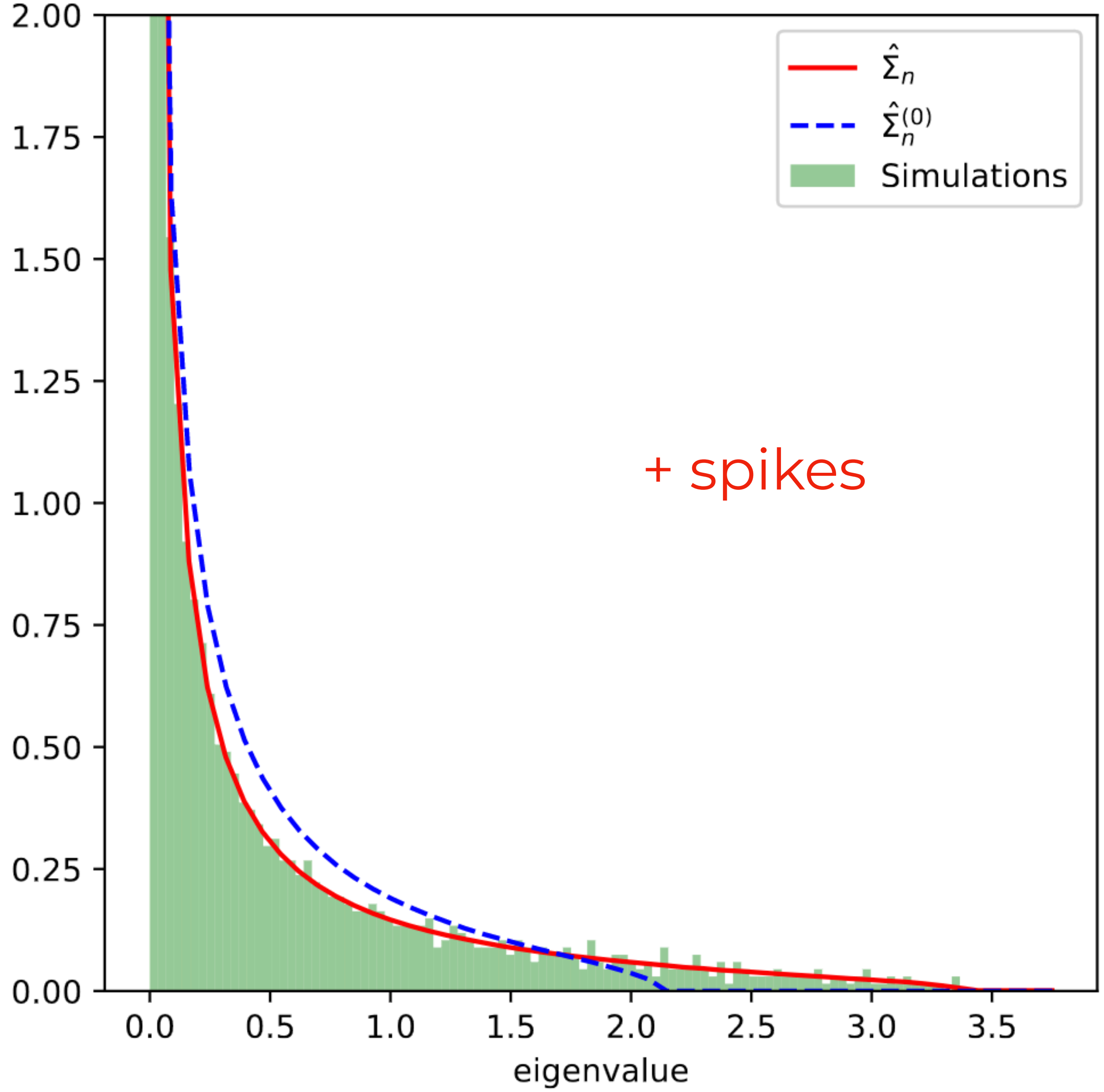
$\sigma = \tanh \qquad g = \text{sign}$



Best linear predictor

$||P_{\kappa \leq 1} f_\star||^2$

Spectral properties



+ spikes

Risk bounds

Recall that. Noting that this is monotonic in $\alpha_0 = n_B/d$:

$$R(\hat{a}_\lambda) = \mathbb{E}_{\kappa,z}\left[\left(g\left(\gamma\kappa + \sqrt{1-\gamma^2}z\right) - \mu_0(\kappa)m - \mu_1(\kappa)\kappa\zeta - \frac{\mu_1(\kappa)\psi}{\sqrt{\rho}}z\right)^2 + \mu_1(\kappa)^2 q_1 + \mu_2(\kappa)^2 q_2 - \frac{\mu_1(\kappa)^2\psi^2}{\rho}\right]$$

Risk bounds

Recall that. Noting that this is monotonic in $\alpha_0 = n_B/d$:

$$\inf_{\lambda \geq 0} R(\alpha, \lambda, \tilde{\eta}, \beta) \leq \inf_{b_1} \mathbb{E}[(g(\kappa) - b_1 \mu_0(\kappa))^2]$$

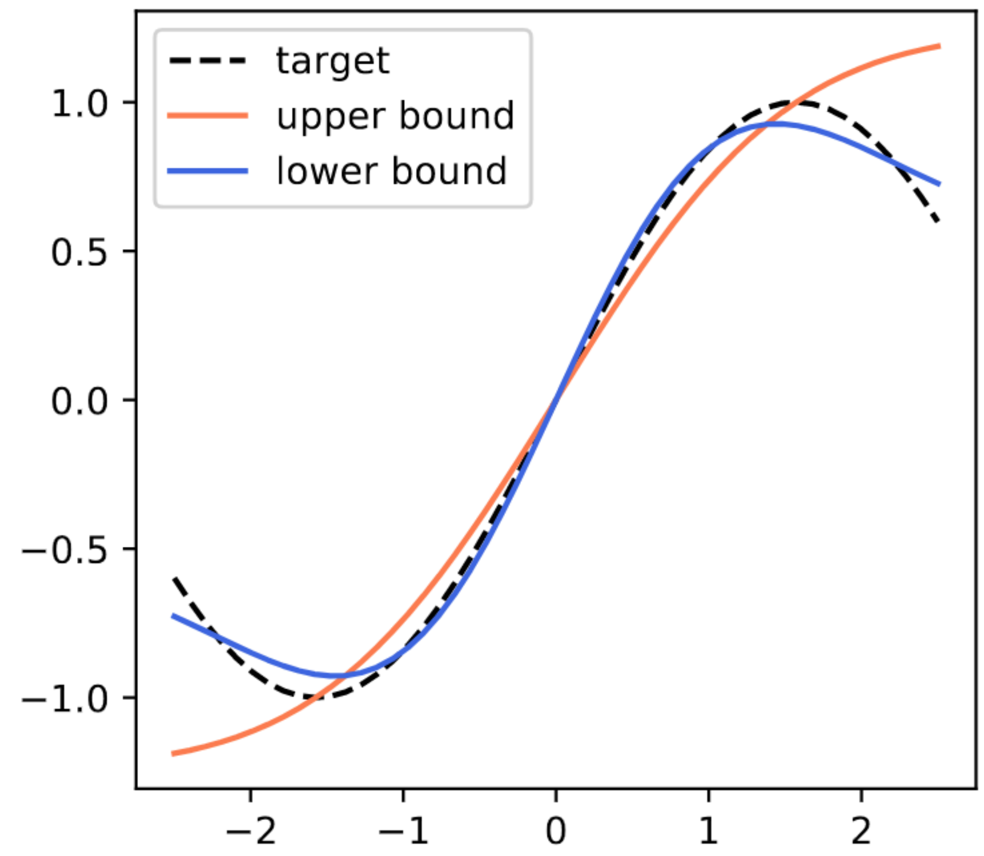
$$\inf_{\lambda \geq 0} R(\alpha, \lambda, \tilde{\eta}, \beta) \geq \inf_{b_1, b_2} \mathbb{E}[(g(\kappa) - b_1 \mu_0(\kappa) - b_2 \mu_1(\kappa)\kappa)^2]$$

$c = \gamma = 1$

$r = 0.9$

$g = \sin$

$\sigma = \tanh$



Risk bounds

Recall that. Noting that this is monotonic in $\alpha_0 = n_B/d$:

$$\inf_{\lambda\geq 0} R(\alpha,\lambda,\tilde{\eta},\beta) \leq \inf_{b_1} \mathbb{E}[(g(\kappa) - b_1\mu_0(\kappa))^2]$$

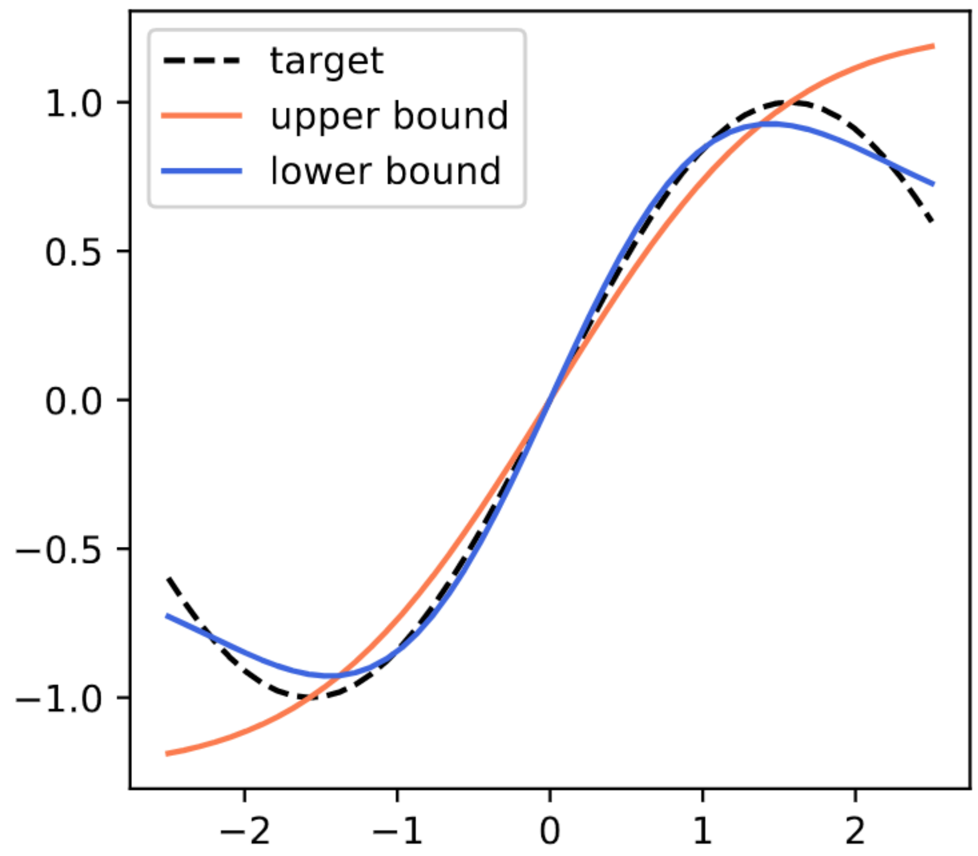
$$\inf_{\lambda\geq 0} R(\alpha,\lambda,\tilde{\eta},\beta) \geq \inf_{b_1,b_2} \mathbb{E}[(g(\kappa) - b_1\mu_0(\kappa) - b_2\mu_1(\kappa)\kappa)^2]$$

$c = \gamma = 1$

$r = 0.9$

$g = \sin$

$\sigma = \tanh$



n.b.:

1. $L_2(\mathcal{N})$ distance between g and $\mathrm{span}(\mu_0, \mu_1')$

2. Can make tighter by optimising over $\tilde{\eta}$

A note on initialisation

So far, assumed $a^0 = 1_p$. But can be generalised to finite support $a^0 \in V$.

$$\sigma(W^1 x) \asymp \begin{bmatrix} \mu_0(u_1\kappa) \\ \vdots \\ \mu_0(u_p\kappa) \end{bmatrix} + \begin{bmatrix} \mu_1(u_1\kappa) \\ \vdots \\ \mu_1(u_p\kappa) \end{bmatrix} \odot Wx + \begin{bmatrix} \mu_2(u_1\kappa) \\ \vdots \\ \mu_2(u_p\kappa) \end{bmatrix} \odot \xi$$

$$u \in V^p \qquad \xi \sim \mathcal{N}(0, I_p)$$

A note on initialisation

So far, assumed $a^0 = 1_p$. But can be generalised to finite support $a^0 \in V$.

$$\sigma(W^1 x) \asymp \begin{bmatrix} \mu_0(u_1\kappa) \\ \vdots \\ \mu_0(u_p\kappa) \end{bmatrix} + \begin{bmatrix} \mu_1(u_1\kappa) \\ \vdots \\ \mu_1(u_p\kappa) \end{bmatrix} \odot Wx + \begin{bmatrix} \mu_2(u_1\kappa) \\ \vdots \\ \mu_2(u_p\kappa) \end{bmatrix} \odot \xi$$

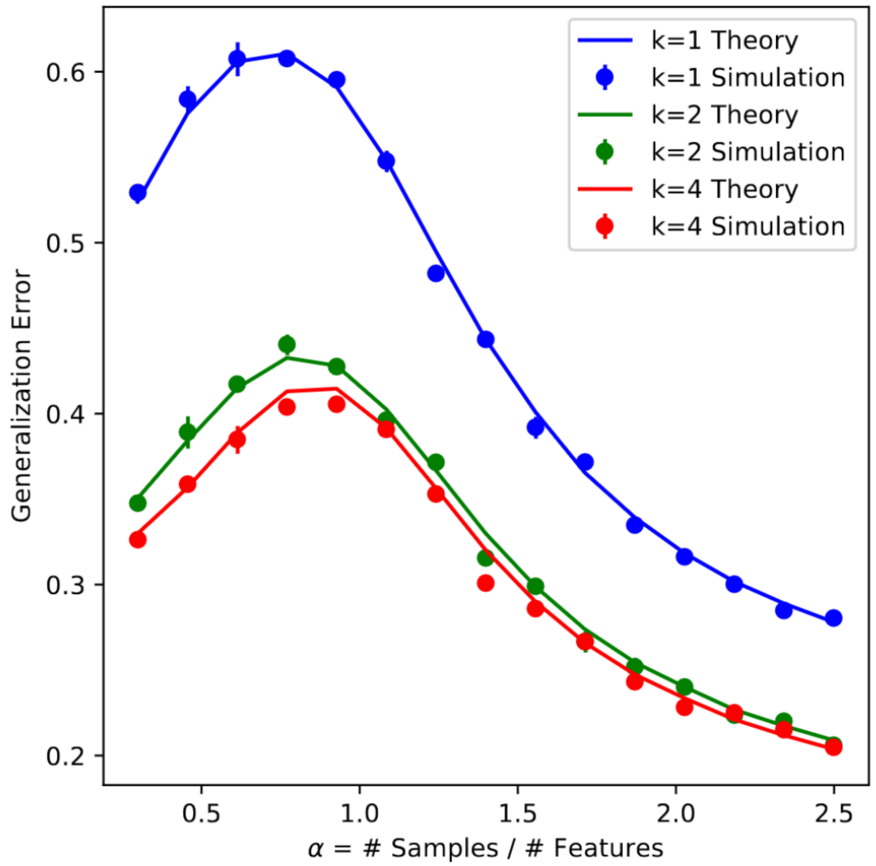
$$u \in V^p \qquad \xi \sim \mathcal{N}(0, I_p)$$

This now spans a richer functional basis:

$$\{\mu_0(\omega\cdot), \mu_1'(\omega\cdot)\}_{\omega\in V}$$

For instance, in the limit $\lambda, \alpha_0, \tilde{\eta} \to \infty$:

$$\sigma(W^1 x)_k \asymp \mu_0(u_k\kappa)$$



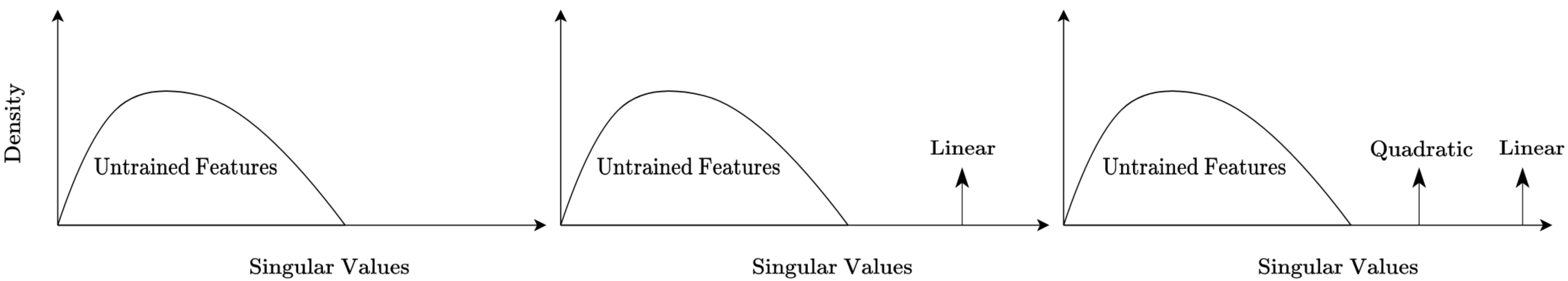
Single neuron with random weights.

Complementary regime

A Theory of Non-Linear Feature Learning with One Gradient Step in Two-Layer Neural Networks

Behrad Moniri*† Donghwan Lee*‡ Hamed Hassani† Edgar Dobriban§

$\eta = 0$ $\eta = \Theta(d^s)$ $\frac{1}{2} < s < 1$



Main ideas

SGD step ⟶ sRF model ⟶ cGET

$$\varphi_i = \sigma(W_1 x_i) \quad \approx \quad \sigma(\tilde{W} x_i + \langle v, x_i \rangle u^\top) \quad \approx \quad \mu_0(\kappa_i u) + \mu_1(\kappa_i u)\tilde{W} x_i^\perp + \mu_\star(\kappa_i u)\xi_i$$

Main ideas

SGD step $\longrightarrow$ sRF model $\longrightarrow$ cGET

$$\varphi_i = \sigma(W_1 x_i) \quad \approx \quad \sigma(\tilde{W}x_i + \langle v, x_i \rangle u^\top) \quad \approx \quad \mu_0(\kappa_i u) + \mu_1(\kappa_i u)\tilde{W}x_i^\perp + \mu_\star(\kappa_i u)\xi_i$$

2 stages of deterministic equivalent: over X and $\tilde{W}$
(leave-one-out + Burkholder)

Main challenges:

- For $u_j \in \{\zeta_1, \ldots, \zeta_k\}$, with prob. $\pi_j = p_j/p$, need to handle k spikes separately.
- For bulk, need deterministic equivalent for block-structured Wishart matrices

$$M = (C_e \odot \tilde{W}\tilde{W}^\top + D_e)^{-1}$$

$$\sum_{j=1}^{k} p_j = p$$

$$C_e = \begin{bmatrix} C_{11}1_{p_1\times p_1} & C_{12}1_{p_1\times p_2} & \cdots & C_{1k}1_{p_1\times p_k} \\ C_{21}1_{p_2\times p_1} & C_{22}1_{p_2\times p_2} & \cdots & C_{2k}1_{p_2\times p_k} \\ \vdots & \vdots & \ddots & \vdots \end{bmatrix} \in \mathbb{R}^{k\times k}$$

$$D_e = \begin{bmatrix} D_{11}I_{p_1\times p_1} & 0 & \cdots & 0 \\ 0 & D_{22}I_{p_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \end{bmatrix} \in \mathbb{R}^{k\times k}$$

Conclusion

In proportional asymptotics,
kernels can learn at best a linear approximation

With one gradient step, 2LNN learn
do better than kernels along
one (and only one) direction

We can provide a sharp asymptotic description
on what is learned

Conclusion

In proportional asymptotics,
kernels can learn at best a linear approximation

With one gradient step, 2LNN learn
do better than kernels along
one (and only one) direction

We can provide a sharp asymptotic description
on what is learned

Multiple steps, same batch,
continuous weights

Collaborators in these works

L. Zdeborová
(EPFL)

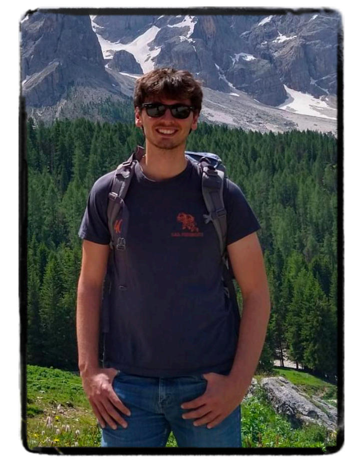
F. Krzakala
(EPFL)

L. Stephane
(EPFL)

C. Gerbelot
(Courant)

G. Reeves
(Duke U.)

H. Cui
(EPFL)

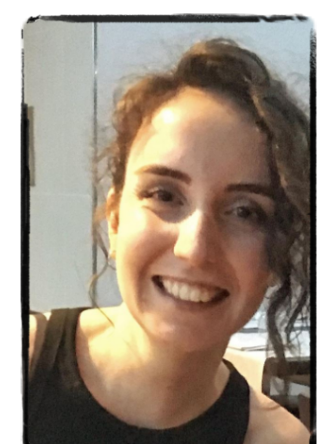


L Pesce
(EPFL)

Y. Dandi
(EPFL)

Y.M. Lu
(Harvard U.)

S. Goldt
(SISSA)



F. Gerace
(SISSA)

M. Mézard
(Bocconi U.)

Thank you!