Wonders of high-dimensions:
the maths and physics of ML

Bruno Loureiro
Département d'Informatique
École Normale Supérieure & CNRS

brloureiro@gmail.com

ACDL 2023, 10-14.06.2023

Yesterday

Challenges for a "theory of ML":

- Overparametrisation can be benign
- Data structure matters
- Non-convex optimisation is hard

Many of these challenges arise due to high-dimensionality…

Statistical physics as the study of high-dimensional probability.
"More is different"

[Anderson 1972]

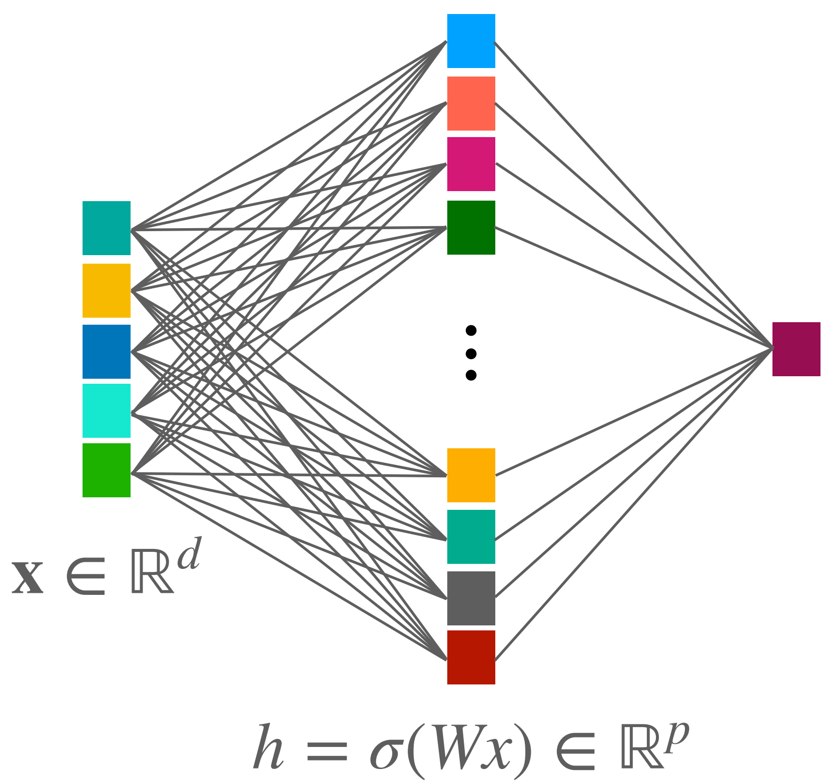
Part II

Two layer neural networks in the lazy regime

The random features model

Two-layer neural nets

We now focus our attention on two-layer neural networks:

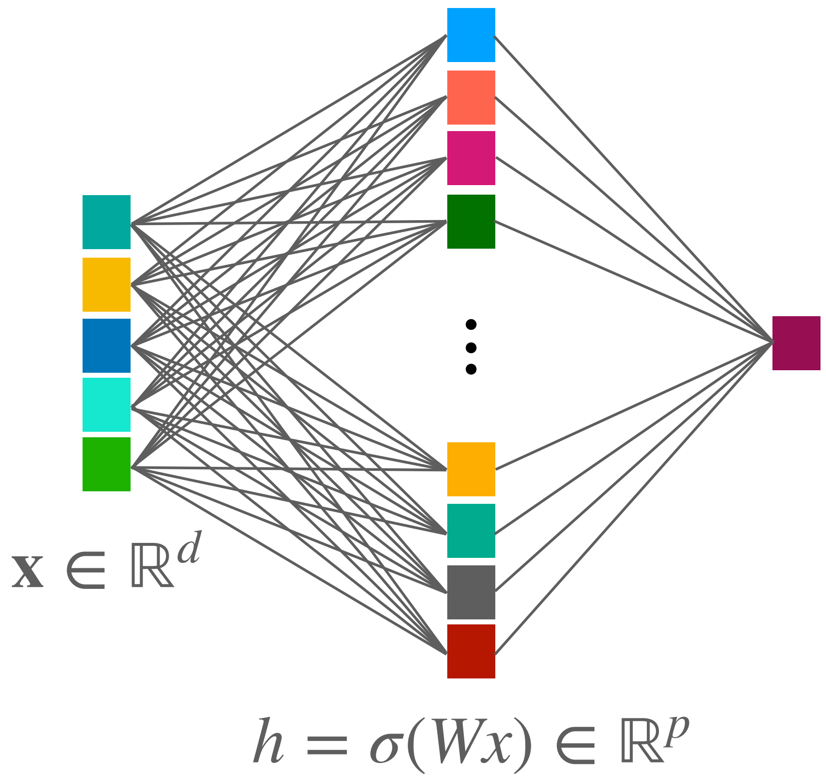


$$f(x; \Theta) = \sum_{i=1}^{p} a_i \sigma(w_i^\top x)$$

$$\Theta = (a, W) \in \mathbb{R}^p \times \mathbb{R}^{p \times d}$$

Two-layer neural nets

We now focus our attention on two-layer neural networks:



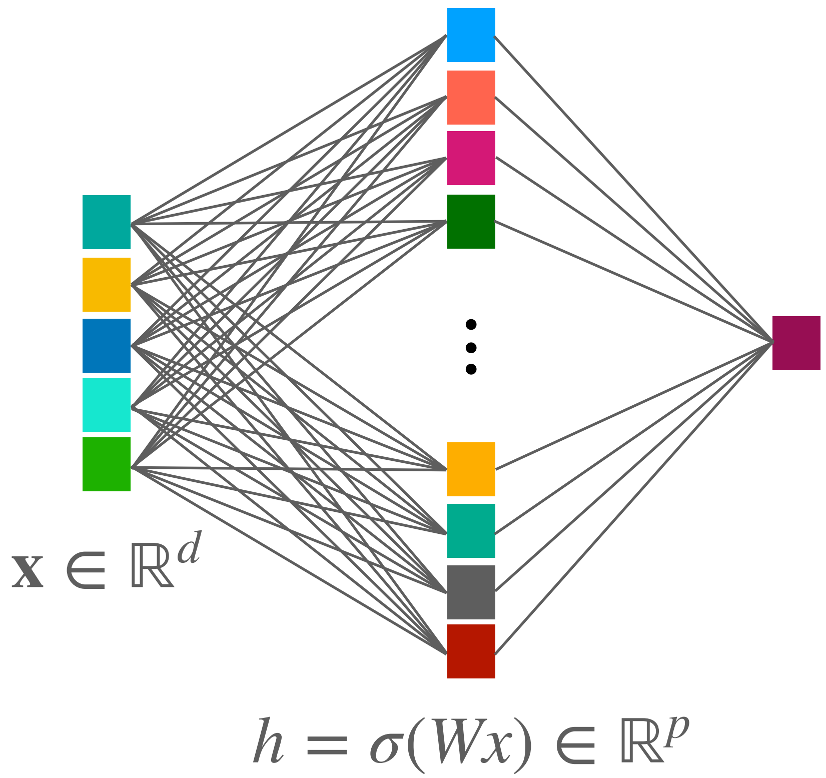
$$f(x; \Theta) = \sum_{i=1}^{p} a_i \sigma(w_i^\top x)$$

$$\Theta = (a, W) \in \mathbb{R}^p \times \mathbb{R}^{p \times d}$$

Question: What happens when $p \to \infty$?

Two-layer neural nets

We now focus our attention on two-layer neural networks:



$$f(x;\Theta) = \sum_{i=1}^{p} a_i \sigma(w_i^\top x)$$

$$\Theta = (a, W) \in \mathbb{R}^p \times \mathbb{R}^{p\times d}$$

Question: What happens when $p \to \infty$? It depends!

Lazy vs. Rich regimes.

[Jacot et al. '18; Chizat, Oyallon & Bach '19]

Lazy vs. rich regime

Let Θ_0 be a fixed set of "generic" weights.

Lazy vs. rich regime

Let Θ_0 be a fixed set of "generic" weights.

Then, on a neighbourhood of Θ_0:

$\bar{f}_{\text{lin}}(x;\Theta_0)$

$$f(x;\Theta) = f(x;\Theta_0) + \nabla_\Theta f(x;\Theta_0)^\top(\Theta - \Theta_0) + \\ + \frac{1}{2}(\Theta - \Theta_0)^\top \nabla^2_\Theta f(x;\Theta_0)(\Theta - \Theta_0) + \cdots$$

Lazy vs. rich regime

Let Θ_0 be a fixed set of "generic" weights.

Then, on a neighbourhood of Θ_0:

$$\bar{f}_{\text{lin}}(x;\Theta_0)$$

$$f(x;\Theta) = f(x;\Theta_0) + \nabla_\Theta f(x;\Theta_0)^\top(\Theta - \Theta_0) + \\ + \frac{1}{2}(\Theta - \Theta_0)^\top \nabla^2_\Theta f(x;\Theta_0)(\Theta - \Theta_0) + \cdots$$

Under one step of gradient descent, we have:

$$\Theta_1 = \Theta_0 - \eta \nabla_\Theta \hat{\mathcal{R}}_n(\Theta_0)$$

Lazy vs. rich regime

Let Θ_0 be a fixed set of "generic" weights.

Then, on a neighbourhood of Θ_0:

$$f(x;\Theta) = \underbrace{f(x;\Theta_0) + \nabla_\Theta f(x;\Theta_0)^\top(\Theta-\Theta_0)}_{\bar{f}_{\mathrm{lin}}(x;\Theta_0)} + \\ + \frac{1}{2}(\Theta-\Theta_0)^\top \nabla^2_\Theta f(x;\Theta_0)(\Theta-\Theta_0) + \cdots$$

Under one step of gradient descent, we have:

$$\Theta_1 = \Theta_0 - \eta\nabla_\Theta \hat{\mathscr{R}}_n(\Theta_0)$$

Therefore, $\bar{f}_{\mathrm{lin}}(x;\Theta_0)$ will is a good approximation if:

$$\kappa(\Theta_0) = \frac{\delta\nabla_\Theta f}{\delta\hat{\mathscr{R}}_n} \ll 1$$

Lazy vs. rich regime

For $\eta \ll 1$, the relative change in the risk is:

$$\delta \hat{\mathscr{R}}_n = \frac{|\hat{\mathscr{R}}_n(\Theta_1) - \hat{\mathscr{R}}_n(\Theta_0)|}{\hat{\mathscr{R}}_n(\Theta_0)} \underset{\eta \ll 1}{\approx} \eta \frac{||\nabla_\Theta \hat{\mathscr{R}}_n(\Theta_0)||_2}{\hat{\mathscr{R}}_n(\Theta_0)}$$

Lazy vs. rich regime

For $\eta \ll 1$, the relative change in the risk is:

$$\delta \hat{\mathscr{R}}_n = \frac{|\hat{\mathscr{R}}_n(\Theta_1) - \hat{\mathscr{R}}_n(\Theta_0)|}{\hat{\mathscr{R}}_n(\Theta_0)} \underset{\eta \ll 1}{\approx} \eta \frac{||\nabla_\Theta \hat{\mathscr{R}}_n(\Theta_0)||_2}{\hat{\mathscr{R}}_n(\Theta_0)}$$

While the relative change in the features is:

$$\delta \nabla_\Theta f \approx \eta \frac{||\nabla^2_\Theta f(\Theta_0)||}{||\nabla_\Theta f(\Theta_0)||}$$

Lazy vs. rich regime

For $\eta \ll 1$, the relative change in the risk is:

$$\delta \hat{\mathscr{R}}_n = \frac{|\hat{\mathscr{R}}_n(\Theta_1) - \hat{\mathscr{R}}_n(\Theta_0)|}{\hat{\mathscr{R}}_n(\Theta_0)} \underset{\eta \ll 1}{\approx} \eta \frac{||\nabla_\Theta \hat{\mathscr{R}}_n(\Theta_0)||_2}{\hat{\mathscr{R}}_n(\Theta_0)}$$

While the relative change in the features is:

$$\delta \nabla_\Theta f \approx \eta \frac{||\nabla^2_\Theta f(\Theta_0)||}{||\nabla_\Theta f(\Theta_0)||}$$

For the square loss, we have:

$$\kappa(\Theta_0) = ||y - f(x; \Theta_0||\frac{||\nabla^2_\Theta f(\Theta_0)||}{||\nabla_\Theta f(\Theta_0)||} \ll 1$$

Lazy vs. rich regime

Assuming that $a_{0,i} = O(1)$ and introducing a scaling:

$$f(x;\Theta) = \alpha(p)\sum_{i=1}^{p} a_i \sigma(w_i^\top x)$$

It can be shown that for $p \gg 1$: [Chizat, Oyallon & Bach '19]

$$\mathbb{E}[\kappa(\Theta_0)] \lesssim \frac{1}{\sqrt{p}} + \frac{1}{p\alpha(p)}$$

Which means $f(x;\Theta) \approx \bar{f}_{\text{lin}}(x;\Theta_0)$ if $p\alpha(p) \to \infty$ as $p \to \infty$

a.k.a. “lazy” regime

The neural tangent kernel

Under a "lazy scaling" and considering the gradient flow limit $\eta \to 0$:

$$\dot{\Theta}(t) = \frac{1}{n}\Phi^\top(y - \hat{y}_0 - \Phi(\Theta(t) - \Theta_0))$$

[Lee et al. '19]

The neural tangent kernel

Under a "lazy scaling" and considering the gradient flow limit $\eta \to 0$:

$$\dot{\Theta}(t) = \frac{1}{n}\Phi^\top(y - \hat{y}_0 - \Phi(\Theta(t) - \Theta_0))$$

This corresponds exactly to a least squares regression:

$$\min_{\Theta} \frac{1}{2n} \sum_{\nu \in [n]} (y^\nu - \Theta^\top \varphi(x^\nu))^2$$

With features:

$$\varphi(x) = \nabla_\Theta f(x; \Theta_0) = \begin{pmatrix} \sigma(W_0 x) \\ a_0 \odot \sigma'(W_0 x) \otimes x \end{pmatrix}$$

[Lee et al. '19]

The neural tangent kernel

Under a "lazy scaling" and considering the gradient flow limit $\eta \to 0$:

$$\dot{\Theta}(t) = \frac{1}{n}\Phi^\top(y - \hat{y}_0 - \Phi(\Theta(t) - \Theta_0))$$

This corresponds exactly to a least squares regression:

$$\min_{\Theta} \frac{1}{2n} \sum_{\nu \in [n]} (y^\nu - \Theta^\top \varphi(x^\nu))^2$$

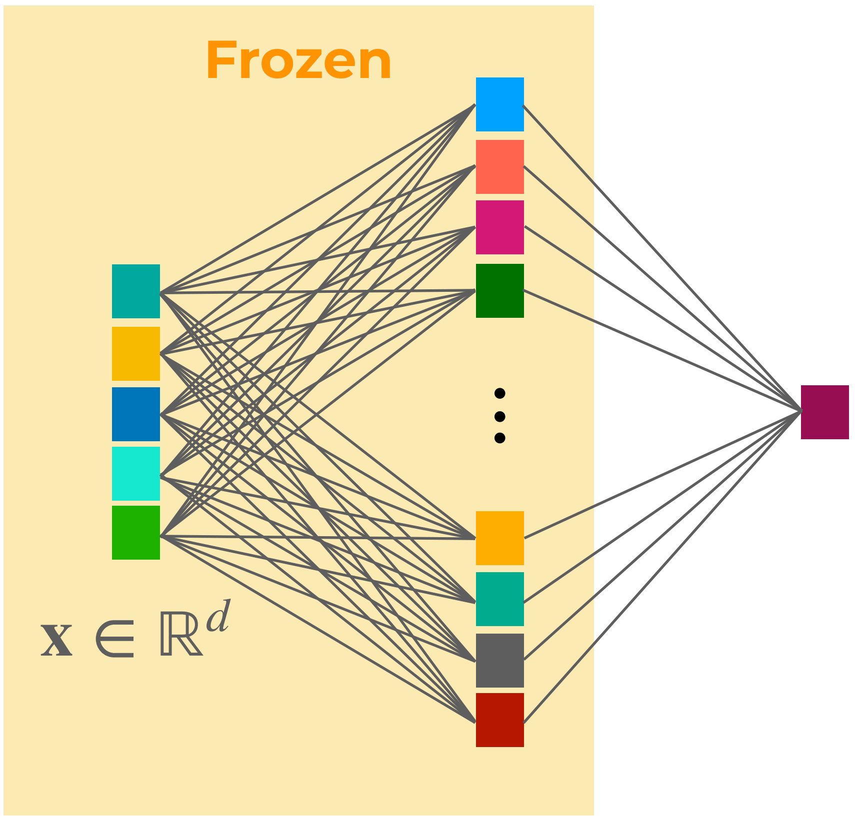
With features:

$$\varphi(x) = \nabla_\Theta f(x; \Theta_0) = \begin{pmatrix} \sigma(W_0 x) \\ a_0 \odot \sigma'(W_0 x) \otimes x \end{pmatrix} \begin{matrix} \text{Random features} \\ \text{NT features} \end{matrix}$$

[Lee et al. '19]

Random features model

[Rahimi & Recht '07]



$$W \sim p_W$$

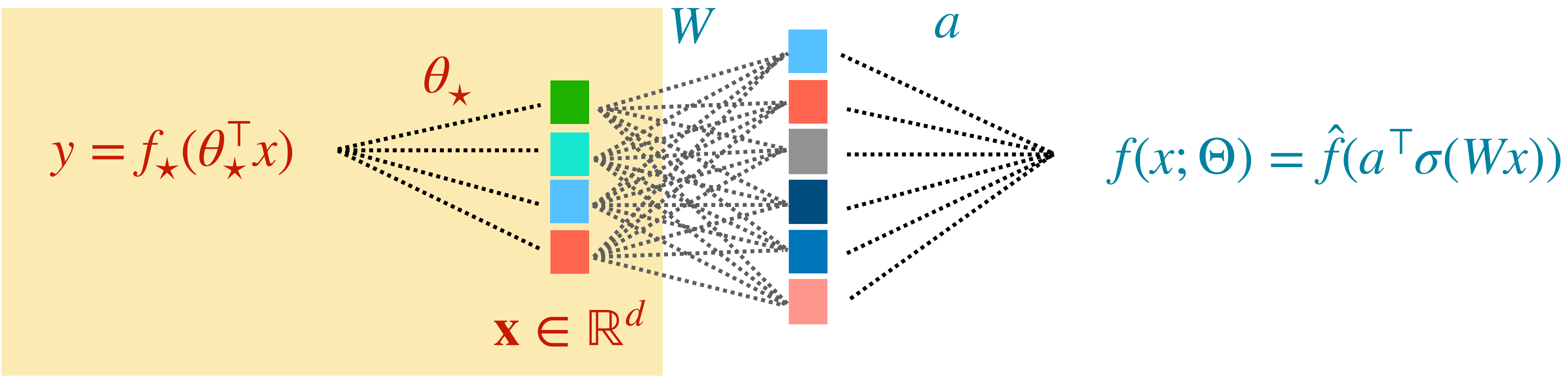
$$f(x;\Theta) = \frac{1}{\sqrt{p}} a^\top \sigma(Wx)$$

$$\min_{a \in \mathbb{R}^p} \frac{1}{n} \sum_{\nu \in [n]} \ell\left(y^\nu, \frac{1}{\sqrt{p}} a^\top \sigma(Wx^\nu)\right) + \frac{\lambda}{2} ||a||_2^2$$

Convex in $a \in \mathbb{R}^p$!

RF model: a simple set-up

[Mei & Montanari '19; Gerace et al. '20]

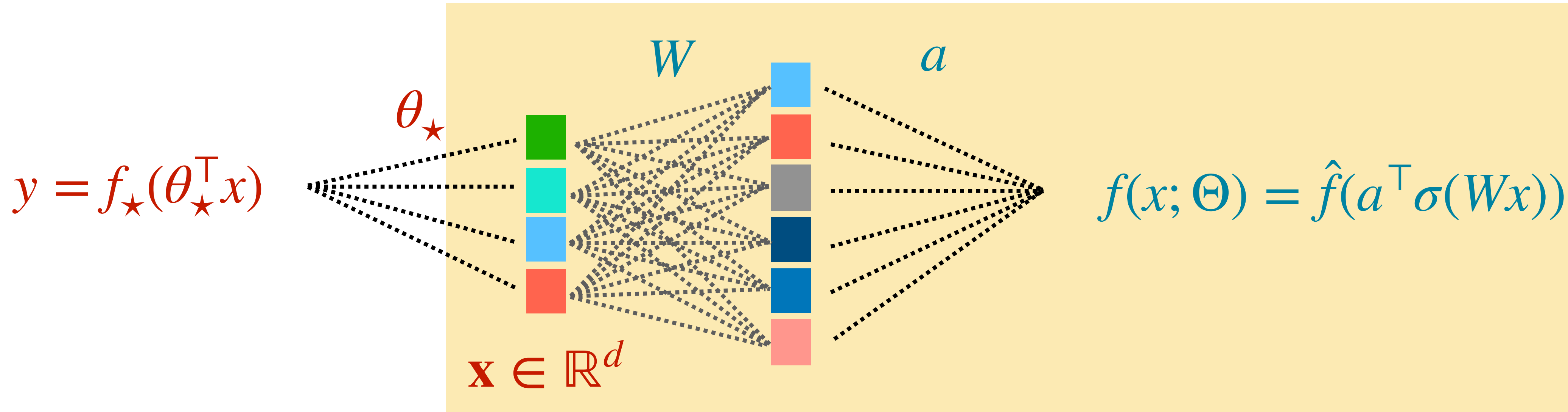


Data: $(x^\nu, y^\nu)_{\nu \in [n]} \in \mathbb{R}^d \times \mathscr{Y}$ generated as:

$$y^\nu = f_\star(\theta_\star^\top x^\nu) \qquad x^\nu \sim \mathcal{N}(0, I_d) \qquad f_\star : \mathbb{R} \to \mathscr{Y}$$

RF model: a simple set-up

[Mei & Montanari '19; Gerace et al. '20]



Data: $(x^\nu, y^\nu)_{\nu \in [n]} \in \mathbb{R}^d \times \mathcal{Y}$ generated as:

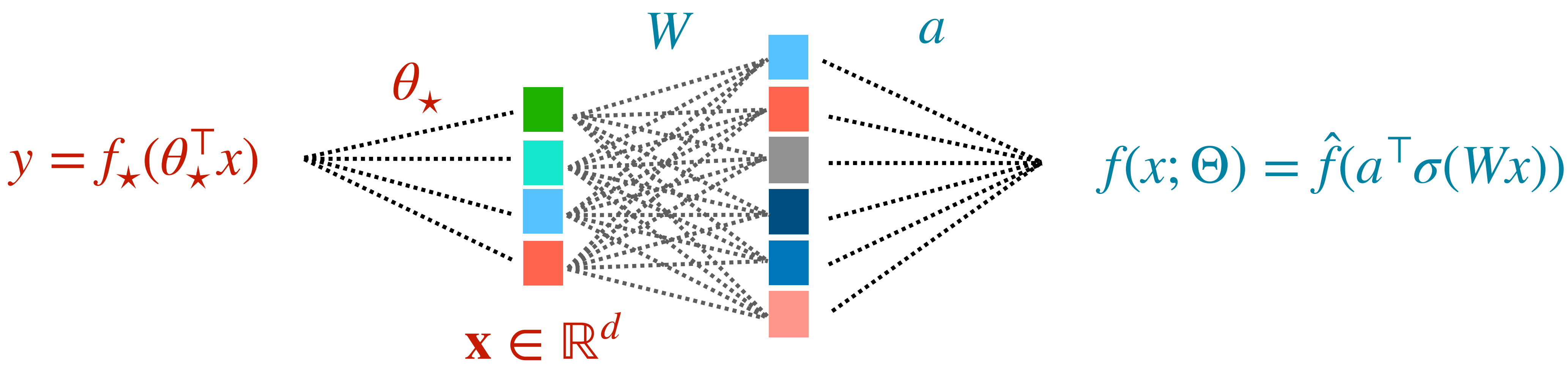
$$y^\nu = f_\star(\theta_\star^\top x^\nu) \qquad x^\nu \sim \mathcal{N}(0, I_d) \qquad f_\star : \mathbb{R} \to \mathcal{Y}$$

Hypothesis:

$$f(x;\Theta) = \hat{f}\left(a^\top \varphi(x)\right) \qquad \varphi(x) = \frac{1}{\sqrt{p}}\sigma(Wx)$$

RF model: a simple set-up

[Mei & Montanari '19; Gerace et al. '20]



Data: $(x^\nu, y^\nu)_{\nu\in[n]} \in \mathbb{R}^d \times \mathcal{Y}$ generated as:

$$y^\nu = f_\star(\theta_\star^\top x^\nu) \qquad x^\nu \sim \mathcal{N}(0, I_d) \qquad f_\star : \mathbb{R} \to \mathcal{Y}$$

Hypothesis: $$f(x;\Theta) = \hat{f}\left(a^\top \varphi(x)\right) \qquad \varphi(x) = \frac{1}{\sqrt{p}}\sigma(Wx)$$

ERM: $$\min_{a\in\mathbb{R}^p} \frac{1}{n} \sum_{\nu\in[n]} \ell\left(y^\nu, a^\top \varphi(x)\right) + \frac{\lambda}{2}||a||_2^2$$

Statistical Physics analysis

[Gerace et al. '20; Loureiro et al. '21]

As discussed in the previous lecture, introduce a Gibbs measure over the risk:

$$\mu_\beta(a) = \frac{1}{Z_\beta} e^{-\beta\left[\sum_{\nu\in[n]} \ell(y^\nu, a^\top \varphi(x)) + \frac{\lambda}{2}||a||_2^2\right]}$$

Statistical Physics analysis

[Gerace et al. '20; Loureiro et al. '21]

As discussed in the previous lecture, introduce a Gibbs measure over the risk:

$$\mu_\beta(a) = \frac{1}{Z_\beta} e^{-\beta\left[\sum_{\nu\in[n]} \ell(y^\nu, a^\top \varphi(x)) + \frac{\lambda}{2}||a||_2^2\right]}$$

$$= \frac{1}{Z_\beta} \underbrace{e^{-\frac{\beta\lambda}{2}||a||_2^2}}_{p_a(a)} \prod_{\nu\in[n]} \underbrace{e^{-\beta\ell(y^\nu, a^\top\varphi(x))}}_{p_y(y\,|\,a^\top\varphi(x))}$$

Statistical Physics analysis

[Gerace et al. '20; Loureiro et al. '21]

As discussed in the previous lecture, introduce a Gibbs measure over the risk:

$$\mu_\beta(a) = \frac{1}{Z_\beta} e^{-\beta\left[\sum_{\nu\in[n]} \ell(y^\nu, a^\top\varphi(x)) + \frac{\lambda}{2}||a||_2^2\right]}$$

$$= \frac{1}{Z_\beta} \underbrace{e^{-\frac{\beta\lambda}{2}||a||_2^2}}_{p_a(a)} \prod_{\nu\in[n]} \underbrace{e^{-\beta\ell(y^\nu, a^\top\varphi(x))}}_{p_y(y|a^\top\varphi(x))}$$

Goal: Compute $-\beta f_\beta = \lim_{d\to\infty} \frac{1}{d}\mathbb{E}[\log Z_\beta]$

When $n, d, p \to \infty$ at fixed ratio $\alpha = \frac{n}{d}$ and $\gamma = \frac{d}{p}$

Statistical Physics analysis

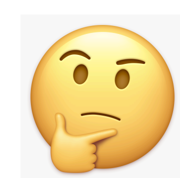
[Gerace et al. '20; Loureiro et al. '21]

As discussed in the previous lecture, introduce a Gibbs measure over the risk:

$$\mu_\beta(a) = \frac{1}{Z_\beta} e^{-\beta \left[\sum_{\nu \in [n]} \ell(y^\nu, a^\top \varphi(x)) + \frac{\lambda}{2} ||a||_2^2 \right]}$$

$$= \frac{1}{Z_\beta} \underbrace{e^{-\frac{\beta\lambda}{2}||a||_2^2}}_{p_a(a)} \prod_{\nu \in [n]} \underbrace{e^{-\beta \ell(y^\nu, a^\top \varphi(x))}}_{p_y(y | a^\top \varphi(x))}$$

Goal: Compute $\quad -\beta f_\beta = \lim_{d \to \infty} \frac{1}{d} \mathbb{E}[\log Z_\beta]$



When $n, d, p \to \infty$ at fixed ratio $\alpha = \frac{n}{d}$ and $\gamma = \frac{d}{p}$

The replica trick

Key idea: $$\log Z_\beta = \lim_{s\to 0^+} \partial_s Z_\beta^s$$ [Kac 1968]

The replica trick

Key idea: $$\log Z_\beta = \lim_{s \to 0^+} \partial_s Z_\beta^s$$ [Kac 1968]

Such that: $$-\beta f_\beta = \lim_{d \to \infty} \frac{1}{d} \lim_{s \to 0^+} \partial_s \mathbb{E}[Z_\beta^s]$$ with:

$$\mathbb{E}[Z_\beta^s] = \prod_{\nu \in [n]} \mathbb{E}_{(x^\nu, y^\nu)} \left[\int \prod_{a=1}^{s} p(\mathrm{d}a^a) p_y(y^\nu \,|\, a^{a\top} \varphi(x^\nu)) \right]$$

The replica trick

Key idea: $$\log Z_\beta = \lim_{s\to 0^+} \partial_s Z_\beta^s$$ [Kac 1968]

Such that: $$-\beta f_\beta = \lim_{d\to\infty} \frac{1}{d} \lim_{s\to 0^+} \partial_s \mathbb{E}[Z_\beta^s]$$ with:

$$\mathbb{E}[Z_\beta^s] = \prod_{\nu\in[n]} \mathbb{E}_{(x^\nu,y^\nu)} \left[\int \prod_{a=1}^{s} p(\mathrm{d}a^a) p_y(y^\nu \,|\, a^{a\top} \varphi(x^\nu)) \right]$$

3 key steps:

1. Taking the average wrt to the data
2. Taking $d \to \infty$ limit with Laplace method
3. Taking the $s \to 0^+$ limit

Interlude

Replica computation

The replica trick

Key idea: $\log Z_\beta = \lim_{s \to 0^+} \partial_s Z_\beta^s$ [Kac 1968]

Such that: $-\beta f_\beta = \lim_{d \to \infty} \frac{1}{d} \lim_{s \to 0^+} \partial_s \mathbb{E}[Z_\beta^s]$ with:

$$\mathbb{E}[Z_\beta^s] = \prod_{\nu \in [n]} \mathbb{E}_{(x^\nu, y^\nu)} \left[\int \prod_{a=1}^{s} p(\mathrm{d}a^a) p_y(y^\nu \,|\, a^{a\top} \varphi(x^\nu)) \right]$$

$$= \prod_{\nu \in [n]} \mathbb{E}_{x^\nu} \int \mathrm{d}y^\nu p_0(y^\nu \,|\, \theta_\star^\top x^\nu) \int \prod_{a=1}^{s} p(\mathrm{d}a^a) p_y(y^\nu \,|\, a^{a\top} \varphi(x^\nu))$$

The replica trick

Key idea: $$\log Z_\beta = \lim_{s\to 0^+} \partial_s Z_\beta^s$$ [Kac 1968]

Such that: $$-\beta f_\beta = \lim_{d\to\infty} \frac{1}{d} \lim_{s\to 0^+} \partial_s \mathbb{E}[Z_\beta^s]$$ with:

$$\mathbb{E}[Z_\beta^s] = \prod_{\nu\in[n]} \mathbb{E}_{(x^\nu,y^\nu)} \left[\int \prod_{a=1}^{s} p(\mathrm{d}a^a) p_y(y^\nu \,|\, a^{a\top} \varphi(x^\nu)) \right]$$

$$= \int \prod_{a=1}^{s} p(\mathrm{d}a^a) \left(\int \mathrm{d}y \; \mathbb{E}_x \left[p_0(y \,|\, \theta_\star^\top x) \prod_{a=1}^{s} p_y(y \,|\, a^{a\top} \varphi(x)) \right] \right)^n$$

The replica trick

Key idea: $$\log Z_\beta = \lim_{s\to 0^+} \partial_s Z_\beta^s$$ [Kac 1968]

Such that: $$-\beta f_\beta = \lim_{d\to\infty} \frac{1}{d} \lim_{s\to 0^+} \partial_s \mathbb{E}[Z_\beta^s]$$ with:

$$\mathbb{E}[Z_\beta^s] = \prod_{\nu\in[n]} \mathbb{E}_{(x^\nu,y^\nu)}\left[\int \prod_{a=1}^{s} p(\mathrm{d}a^a) p_y(y^\nu \,|\, a^{a\top}\varphi(x^\nu))\right]$$

$$= \int \prod_{a=1}^{s} p(\mathrm{d}a^a) \left(\int \mathrm{d}y \; \boxed{\mathbb{E}_x \left[p_0(y \,|\, \theta_\star^\top x) \prod_{a=1}^{s} p_y(y \,|\, a^{a\top}\varphi(x)) \right]} \right)^n$$

Step 1: taking the average

$$\mathbb{E}_x \left[p_0(y \mid \theta_\star^\top x) \prod_{a=1}^{s} p_y(y \mid a^{a\top} \varphi(x)) \right]$$

Step 1: taking the average

$$\mathbb{E}_x \left[p_0(y \,|\, \theta_\star^\top x) \prod_{a=1}^{s} p_y(y \,|\, a^{a\top} \varphi(x)) \right] =$$

$$= \int \mathrm{d}\nu p_\star(y \,|\, \nu) \int \prod_{a=1}^{s} \mathrm{d}\lambda^a p_y(y \,|\, \lambda^a) \mathbb{E}_x \left[\delta(\nu - \theta_\star^\top x) \prod_{a=1}^{s} \delta(\lambda^a - a^{a\top} \varphi(x)) \right]$$

Step 1: taking the average

$$\mathbb{E}_x\left[p_0(y\,|\,\theta_\star^\top x)\prod_{a=1}^{s} p_y(y\,|\,a^{a\top}\varphi(x))\right] =$$

$$= \int \mathrm{d}\nu\, p_\star(y\,|\,\nu) \int \prod_{a=1}^{s} \mathrm{d}\lambda^a p_y(y\,|\,\lambda^a)\, \mathbb{E}_x\left[\delta(\nu - \theta_\star^\top x)\prod_{a=1}^{s}\delta(\lambda^a - a^{a\top}\varphi(x))\right]$$

$p(\nu, \lambda^1, \cdots, \lambda^s)$ 🤔

Step 1: taking the average

$$\mathbb{E}_x\left[p_0(y\,|\,\theta_\star^\top x)\prod_{a=1}^{s}p_y(y\,|\,a^{a\top}\varphi(x))\right] =$$

$$= \int \mathrm{d}\nu p_\star(y\,|\,\nu)\int\prod_{a=1}^{s}\mathrm{d}\lambda^a p_y(y\,|\,\lambda^a)\mathbb{E}_x\left[\delta(\nu-\theta_\star^\top x)\prod_{a=1}^{s}\delta(\lambda^a - a^{a\top}\varphi(x))\right]$$

$p(\nu, \lambda^1, \cdots, \lambda^s)$ 🤔

Particular case: $\varphi(x) = \frac{1}{\sqrt{p}}Wx \sim \mathcal{N}(0,\Omega)$ $\quad \Omega = \frac{WW^\top}{p}$

$$p(\nu, \lambda^1, \cdots, \lambda^s) = \mathcal{N}\left(0, \begin{bmatrix} ||\theta_\star||_2^2 & \theta_\star^\top \Phi a^a \\ a^{a\top}\Phi\theta_\star & a^{a\top}\Omega a^b \end{bmatrix}\right) \qquad \Phi = \frac{W}{\sqrt{p}}$$

Step 1: taking the average

$$\mathbb{E}_x\left[p_0(y\,|\,\theta_\star^\top x)\prod_{a=1}^{s} p_y(y\,|\,a^{a\top}\varphi(x))\right] =$$

$$= \int \mathrm{d}\nu p_\star(y\,|\,\nu)\int\prod_{a=1}^{s}\mathrm{d}\lambda^a p_y(y\,|\,\lambda^a)\mathbb{E}_x\left[\delta(\nu-\theta_\star^\top x)\prod_{a=1}^{s}\delta(\lambda^a-a^{a\top}\varphi(x))\right]$$

$p(\nu,\lambda^1,\cdots,\lambda^s)$ 🤔

Particular case: $\varphi(x) = \frac{1}{\sqrt{p}}Wx \sim \mathcal{N}(0,\Omega)$ $\quad \Omega = \frac{WW^\top}{p}$

$$p(\nu,\lambda^1,\cdots,\lambda^s) = \mathcal{N}\left(0,\begin{bmatrix}\rho & m^a\\ m^a & q^{ab}\end{bmatrix}\right) \qquad \Phi = \frac{W}{\sqrt{p}}$$

Step 2: Writing as a SPP

$$\mathbb{E}[Z_\beta^s] = \int \prod_{a=1}^{s} p(\mathrm{d}a^a) \left(\int \mathrm{d}y \; \mathbb{E}_{(\nu,\lambda^a)} \left[p_\star(y \,|\, \nu) \prod_{a=1}^{s} p_y(y \,|\, \lambda^a) \right] \right)^n$$

Step 2: Writing as a SPP

$$\mathbb{E}[Z_\beta^s] = \int \prod_{a=1}^{s} p(\mathrm{d}a^a) \left(\int \mathrm{d}y \; \mathbb{E}_{(\nu,\lambda^a)} \left[p_\star(y \,|\, \nu) \prod_{a=1}^{s} p_y(y \,|\, \lambda^a) \right] \right)^n$$

Goal: Factorise this integral. Introduce:

$$1 \propto \int \prod_{a=1}^{s} \mathrm{d}m^a \delta(\sqrt{pd} \; m^a - \theta_\star \Phi^\top a^a) \int \prod_{1 \le a \le b \le s} \mathrm{d}q^{ab} \delta(p \; q^{ab} - a^{a\top} \Omega a^b)$$

Step 2: Writing as a SPP

$$\mathbb{E}[Z_\beta^s] = \int \prod_{a=1}^{s} p(\mathrm{d}a^a) \left(\int \mathrm{d}y \ \mathbb{E}_{(\nu,\lambda^a)} \left[p_\star(y\,|\,\nu) \prod_{a=1}^{s} p_y(y\,|\,\lambda^a) \right] \right)^n$$

Goal: Factorise this integral. Introduce:

$$1 \propto \int \prod_{a=1}^{s} \mathrm{d}m^a \delta(\sqrt{pd}\ m^a - \theta_\star \Phi^\top a^a) \int \prod_{1 \le a \le b \le s} \mathrm{d}q^{ab} \delta(p\ q^{ab} - a^{a\top} \Omega a^b)$$

$$= \int \prod_{a=1}^{s} \frac{\mathrm{d}m^a \mathrm{d}\hat{m}^a}{2\pi} e^{i \sum_{a=1}^{s} \hat{m}^a (\sqrt{pd} m^a - \theta_\star \Phi^\top a^a)} \times$$

$$\times \int \prod_{1 \le a \le b \le s} \frac{\mathrm{d}q^{ab} \mathrm{d}\hat{q}^{ab}}{2\pi} e^{i \sum_{1 \le a \le b \le s} \hat{q}^{ab} (p q^{ab} - a^{a\top} \Omega a^b)}$$

Step 2: Writing as a SPP

Inserting this allow us to swap integrals and decouple $p_y, p_\star$ from p_a

$$
\mathbb{E}[Z_\beta^s] = \int \prod_{a=1}^{s} \frac{\mathrm{d}m^a \mathrm{d}\hat{m}^a}{2\pi} e^{i\sqrt{pd}\sum_{a\leq b} q^{ab}\hat{q}^{ab}} \int \prod_{a\leq b} \frac{\mathrm{d}q^{ab}\mathrm{d}\hat{q}^{ab}}{2\pi} e^{ip\sum_{a\leq b} q^{ab}\hat{q}^{ab}}
$$

$$
\times \left(\int \prod_{a=1}^{s} p(\mathrm{d}a^a) e^{-i\sum_{a=1}^{s} \hat{m}^a \theta_\star^\top \Phi^\top a^a - i \sum_{a\leq b} \hat{q}^{ab} a^{a\top} \Omega a^b} \right) \times
$$

$$
\times \left(\int \mathrm{d}y \; \mathbb{E}_{(\nu,\lambda^a)} \left[p_\star(y|\nu) \prod_{a=1}^{s} p_y(y|\lambda^a) \right] \right)^n
$$

Step 2: Writing as a SPP

Inserting this allow us to swap integrals and decouple $p_y, p_\star$ from p_a

$$\mathbb{E}[Z_\beta^s] = \int \prod_{a=1}^{s} \frac{\mathrm{d}m^a \mathrm{d}\hat{m}^a}{2\pi} e^{i\sqrt{pd}\sum_{a\leq b} q^{ab}\hat{q}^{ab}} \int \prod_{a\leq b} \frac{\mathrm{d}q^{ab}\mathrm{d}\hat{q}^{ab}}{2\pi} e^{d\ \Psi^{(s)}(q^{ab},\hat{q}^{ab},m^a,\hat{m}^a)}$$

Step 2: Writing as a SPP

Inserting this allow us to swap integrals and decouple $p_y, p_\star$ from p_a

$$\mathbb{E}[Z_\beta^s] = \int \prod_{a=1}^{s} \frac{\mathrm{d}m^a \mathrm{d}\hat{m}^a}{2\pi} e^{i\sqrt{pd} \sum_{a\leq b} q^{ab}\hat{q}^{ab}} \int \prod_{a\leq b} \frac{\mathrm{d}q^{ab}\mathrm{d}\hat{q}^{ab}}{2\pi} e^{d\ \Psi^{(s)}(q^{ab},\hat{q}^{ab},m^a,\hat{m}^a)}$$

Where we defined:

$$\Psi^{(s)} = i \sum_{a,b=1}^{s} q^{ab}\hat{q}^{ab} + i\sqrt{\gamma} \sum_{a=1}^{s} m^a\hat{m}^a + \alpha \Psi_y^{(s)}(q^{ab}, m^a) + \Psi_a^{(s)}(\hat{q}^{ab}, \hat{m}^a)$$

$$\Psi_y^{(s)} = \log \int \mathrm{d}y\ \mathbb{E}_{(\nu,\lambda^a)} \left[p_\star(y|\nu) \prod_{a=1}^{s} p_y(y|\lambda^a) \right]$$

$$\Psi_a^{(s)} = \log \int \prod_{a=1}^{s} p(\mathrm{d}a^a) e^{-i \sum_{a=1}^{s} \hat{m}^a \theta_\star^\top \Phi^\top a^a - i \sum_{a\leq b}^{s} \hat{q}^{ab} a^{a\top} \Omega a^b}$$

Step 2: Writing as a SPP

This allow us to take the $d \to \infty$ exactly:

$$\mathbb{E}[Z_\beta^s] \underset{d\to\infty}{\approx} e^{d\Psi^{(s)}(q_\star^{ab},\hat{q}_\star^{ab},m_\star^a,\hat{m}_\star^a)}$$

Where $(q_\star^{ab}, \hat{q}_\star^{ab}, m_\star^a, \hat{m}_\star^a)$ are minimisers of

$$\underset{q^{ab},\hat{q}^{ab},m^a,\hat{m}^a}{\text{extr}} \Psi^{(s)}(q^{ab}, \hat{q}^{ab}, m^a, \hat{m}^a)$$

Step 2: Writing as a SPP

This allow us to take the $d \to \infty$ exactly:

$$\mathbb{E}[Z_\beta^s] \underset{d\to\infty}{\approx} e^{d\Psi^{(s)}(q_\star^{ab},\hat{q}_\star^{ab},m_\star^a,\hat{m}_\star^a)}$$

Where $(q_\star^{ab}, \hat{q}_\star^{ab}, m_\star^a, \hat{m}_\star^a)$ are minimisers of

$$\underset{q^{ab},\hat{q}^{ab},m^a,\hat{m}^a}{\mathrm{extr}} \Psi^{(s)}(q^{ab}, \hat{q}^{ab}, m^a, \hat{m}^a)$$

For general s, this is still too complicated.... But we only need to solve this for $s \to 0^+$!

Step 3: Replica symmetric ansatz

Consider the following RS ansatz:

$$m^a = m \qquad \hat{m}^a = -i\hat{m} \qquad \forall a = 1,\ldots,s$$

$$q^{aa} = r \qquad \hat{q}^{aa} = \frac{i}{2}\hat{r}$$

$$q^{ab} = r \qquad \hat{q}^{ab} = -i\hat{q} \qquad \forall a \neq b$$

$$\mathrm{Cov}(\nu, \lambda^a) = \begin{bmatrix} \rho & m & m & \cdots & m \\ m & r & q & \cdots & m \\ m & q & r & \cdots & m \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ m & q & q & \cdots & r \end{bmatrix}$$

Step 2: Writing as a saddle

This allow us to take the $d \to \infty$ exactly:

$$\mathbb{E}[Z_\beta^s] \underset{d\to\infty}{\approx} e^{d\Psi^{(s)}(q_\star^{ab},\hat{q}_\star^{ab},m_\star^a,\hat{m}_\star^a)}$$

Where $(q_\star^{ab}, \hat{q}_\star^{ab}, m_\star^a, \hat{m}_\star^a)$ are minimisers of

$$\underset{q^{ab},\hat{q}^{ab},m^a,\hat{m}^a}{\text{extr}} \Psi^{(s)}(q^{ab}, \hat{q}^{ab}, m^a, \hat{m}^a)$$

For general s, this is still too complicated.... But we only need to solve this for $s \to 0^+$!

Step 3: Replica symmetric ansatz

This allows to make the dependence on s explicit in every term

Trace terms:

$$i \sum_{a=1}^{s} m^a \hat{m}^a = sm\hat{n} \qquad i \sum_{a \leq b}^{s} q^{ab} \hat{q}^{ab} = -\frac{1}{2} r\hat{r} + \frac{s(s-1)}{2} q\hat{q}$$

Step 3: Replica symmetric ansatz

This allows to make the dependence on s explicit in every term

Trace terms:

$$i\sum_{a=1}^{s} m^a \hat{m}^a = s m \hat{n} \qquad i\sum_{a\leq b}^{s} q^{ab}\hat{q}^{ab} = -\frac{1}{2} r\hat{r} + \frac{s(s-1)}{2} q\hat{q}$$

"Prior" potential:

$$\Psi_a^{(s)} = \log \int \prod_{a=1}^{s} p(\mathrm{d}a^a) e^{-i\sum_{a=1}^{s} \hat{m}^a \theta_\star^\top \Phi^\top a^a - i\sum_{a\leq b}^{s} \hat{q}^{ab} a^{a\top} \Omega a^b}$$

Step 3: Replica symmetric ansatz

This allows to make the dependence on s explicit in every term

Trace terms:

$$i\sum_{a=1}^{s} m^a \hat{m}^a = sm\hat{n} \qquad i\sum_{a\leq b}^{s} q^{ab}\hat{q}^{ab} = -\frac{1}{2}r\hat{r} + \frac{s(s-1)}{2}q\hat{q}$$

"Prior" potential:

$$\Psi_a^{(s)} = \log \int \prod_{a=1}^{s} p(\mathrm{d}a^a) e^{-i\sum_{a=1}^{s}\hat{m}^a\theta_\star^\top\Phi^\top a^a - i\sum_{a\leq b}^{s}\hat{q}^{ab}a^{a\top}\Omega a^b}$$

$$= \log \int \prod_{a=1}^{s} p(\mathrm{d}a^a) e^{\hat{m}\theta_\star^\top\Phi^\top\sum_{a=1}^{s}a^a - \frac{\hat{r}+\hat{q}}{2}\sum_{a=1}^{s}a^{a\top}\Omega a^a + \hat{q}\sum_{a,b=1}^{s}a^{a\top}\Omega a^b}$$

Step 3: Replica symmetric ansatz

This allows to make the dependence on s explicit in every term

Trace terms:

$$i\sum_{a=1}^{s} m^a \hat{m}^a = s m \hat{n} \qquad\qquad i\sum_{a\le b}^{s} q^{ab}\hat{q}^{ab} = -\frac{1}{2} r\hat{r} + \frac{s(s-1)}{2} q\hat{q}$$

"Prior" potential:

$$\Psi_a^{(s)} = \log \int \prod_{a=1}^{s} p(\mathrm{d}a^a) e^{-i\sum_{a=1}^{s} \hat{m}^a \theta_\star^\top \Phi^\top a^a - i\sum_{a\le b}^{s} \hat{q}^{ab} a^{a\top}\Omega a^b}$$

$$= \log \int \prod_{a=1}^{s} \left(p(\mathrm{d}a^a) e^{\hat{m}\theta_\star^\top \Phi^\top a^a - \frac{\hat{r}+\hat{q}}{2} a^{a\top}\Omega a^a} \right) e^{\hat{q}\sum_{a,b=1}^{s} a^{a\top}\Omega a^b}$$

Step 3: Replica symmetric ansatz

This allows to make the dependence on s explicit in every term

Trace terms:

$$i \sum_{a=1}^{s} m^a \hat{m}^a = s m \hat{n} \qquad\qquad i \sum_{a \leq b}^{s} q^{ab} \hat{q}^{ab} = -\frac{1}{2} r \hat{r} + \frac{s(s-1)}{2} q \hat{q}$$

"Prior" potential:

$$\Psi_a^{(s)} = \log \int \prod_{a=1}^{s} p(\mathrm{d}a^a) e^{-i \sum_{a=1}^{s} \hat{m}^a \theta_\star^\top \Phi^\top a^a - i \sum_{a \leq b}^{s} \hat{q}^{ab} a^{a\top} \Omega a^b}$$

$$= \log \int \prod_{a=1}^{s} \left(p(\mathrm{d}a^a) e^{\hat{m} \theta_\star^\top \Phi^\top a^a - \frac{\hat{r} + \hat{q}}{2} a^{a\top} \Omega a^a} \right) \mathbb{E}_{\xi \sim \mathcal{N}(0, I_p)} \left[e^{\sqrt{\hat{q}} \sum_{a=1}^{s} \xi^\top \Omega^{1/2} a^a} \right]$$

Step 3: Replica symmetric ansatz

This allows to make the dependence on s explicit in every term

Trace terms:

$$i\sum_{a=1}^{s} m^a \hat{m}^a = s m \hat{n} \qquad\qquad i\sum_{a\leq b}^{s} q^{ab}\hat{q}^{ab} = -\frac{1}{2} r\hat{r} + \frac{s(s-1)}{2} q\hat{q}$$

"Prior" potential:

$$\Psi_a^{(s)} = \log \int \prod_{a=1}^{s} p(\mathrm{d}a^a) e^{-i\sum_{a=1}^{s} \hat{m}^a \theta_\star^\top \Phi^\top a^a - i\sum_{a\leq b}^{s} \hat{q}^{ab} a^{a\top}\Omega a^b}$$

$$= \log \mathbb{E}_{\xi\sim\mathcal{N}(0,I_p)} \left(\int p(\mathrm{d}a) e^{-\frac{\hat{r}+\hat{q}}{2} a^\top \Omega a + a^\top\left(\hat{m}\Phi\theta_\star + \sqrt{\hat{q}}\Omega^{1/2}\xi\right)} \right)^s$$

Step 3: Replica symmetric ansatz

This allows to make the dependence on s explicit in every term

"Likelihood" potential:

$$\Psi_y^{(s)} = \log \int \mathrm{d}y \; \mathbb{E}_{(\nu,\lambda^a)} \left[p_\star(y \,|\, \nu) \prod_{a=1}^{s} p_y(y \,|\, \lambda^a) \right]$$

Step 3: Replica symmetric ansatz

This allows to make the dependence on s explicit in every term

"Likelihood" potential:

$$\Psi_y^{(s)} = \log \int \mathrm{d}y \; \mathbb{E}_{(\nu,\lambda^a)} \left[p_\star(y\,|\,\nu) \prod_{a=1}^{s} p_y(y\,|\,\lambda^a) \right]$$

$$= \log \int \mathrm{d}y \int \mathrm{d}\nu \; p_\star(y\,|\,\nu) \int \prod_{a=1}^{s} \mathrm{d}\lambda^a p_y(y\,|\,\lambda^a) e^{-\frac{1}{2}\begin{pmatrix}\nu & \lambda^a\end{pmatrix}\begin{pmatrix}\rho & m^a \\ m^a & q^{ab}\end{pmatrix}^{-1}\begin{pmatrix}\nu \\ \lambda^a\end{pmatrix}}$$

Step 3: Replica symmetric ansatz

This allows to make the dependence on s explicit in every term

"Likelihood" potential:

$$\Psi_y^{(s)} = \log \int \mathrm{d}y \; \mathbb{E}_{(\nu,\lambda^a)} \left[p_\star(y \,|\, \nu) \prod_{a=1}^{s} p_y(y \,|\, \lambda^a) \right]$$

$$= \log \int \mathrm{d}y \int \mathrm{d}\nu \; p_\star(y \,|\, \nu) \int \prod_{a=1}^{s} \mathrm{d}\lambda^a p_y(y \,|\, \lambda^a) e^{-\frac{1}{2}(\nu \quad \lambda^a) \begin{pmatrix} \rho & m^a \\ m^a & q^{ab} \end{pmatrix}^{-1} \begin{pmatrix} \nu \\ \lambda^a \end{pmatrix}}$$

Exercise: Decouple this in s (see [Gerace et al. '21] for a solution)

Summary

Taking the limit $s \to 0^+$ and putting together, we can finally get:

$$-\beta f_\beta(\alpha, \lambda) = \underset{r,\hat{r},q,\hat{q},m,\hat{m}}{\operatorname{extr}} \Psi(r, \hat{r}, q, \hat{q}, m, \hat{m})$$

$$\Psi^{(s)} = \frac{1}{2} r\hat{r} + \frac{1}{2} q\hat{q} - \sqrt{\gamma} m\hat{m} + \alpha \Psi_y(r, q, m) + \Psi_a(\hat{r}, \hat{q}, \hat{m})$$

$$\Psi_a(\hat{r}, \hat{q}, \hat{m}) = \mathbb{E}_\xi \log \int \mathrm{d}p_a(a) e^{-\frac{\hat{r}+\hat{q}}{2} a^\top \Omega a + a^\top \left(\hat{m}\Phi\theta_\star + \sqrt{\hat{q}}\Omega^{1/2}\xi\right)}$$

$$\Psi_y(r, q, m) = \mathbb{E}_\eta \int \mathrm{d}y \, Z_0\left(y, \frac{m}{\sqrt{q}}\eta, \rho - \frac{m^2}{q}\right) \log Z_y\left(y, \sqrt{q}\eta, r - q\right)$$

$$Z_{\star/y}(y, \omega, v) = \mathbb{E}_{z \sim \mathcal{N}(\omega, v)}[p_{\star/y}(y \mid z)]$$

Comments

- Result holds for a general Gaussian Covariate model

$$y = f_\star(\theta_\star^\top u)$$

$$(u, v) \sim \mathcal{N}\left(0, \begin{bmatrix} \Psi & \Phi \\ \Phi^\top & \Omega \end{bmatrix}\right)$$

$$f(x; \Theta) = \hat{f}(a^\top v)$$

- This covers "many" feature maps $\varphi(x)$ due to universality in high-dimensions. [Goldt et al. '21; Hu & Lu '21; Saed & Montanari '22]

- Technique readily applies to other "priors" and "likelihoods".

c.f. [Zdeborová & Krzakala 2016] for a review

- Formulas can be rigorously proven using Gordon min-max inequalities or AMP methods. c.f. [Loureiro et al. '21]

Main result

Theorem (informal): for convex losses ℓ and under mild conditions on $(f_0, \hat{f})$ and $(\Psi, \Omega, \Phi, \theta_0)$, the asymptotic errors are given by:

$$\mathcal{R}(\hat{a}) \xrightarrow[d \to \infty]{} R(m_\star, q_\star) \qquad \hat{\mathcal{R}}_n(\hat{a}) \xrightarrow[d \to \infty]{} \hat{R}(v_\star, m_\star, q_\star)$$

Main result

Theorem (informal): for convex losses ℓ and under mild conditions on $(f_0, \hat{f})$ and $(\Psi, \Omega, \Phi, \theta_0)$, the asymptotic errors are given by:

$$\mathscr{R}(\hat{a}) \xrightarrow[d\to\infty]{} R(m_\star, q_\star) \qquad \hat{\mathscr{R}}_n(\hat{a}) \xrightarrow[d\to\infty]{} \hat{R}(v_\star, m_\star, q_\star)$$

Where $v_\star, m_\star, q_\star \in \mathbb{R}_+$ extremise the following potential function

$$\min_{v,q,m} \max_{\hat{v},\hat{q},\hat{m}} \frac{1}{2}(\hat{v}q - v\hat{q}) + \sqrt{\gamma} m\hat{m} + \alpha \mathbb{E}_{\eta\sim\mathcal{N}(0,1)} \left[Z_\star\left(y, \frac{m}{\sqrt{q}}\eta, \rho - \frac{m^2}{q}\right) \mathscr{M}_{v\ell(y,\cdot)}(\sqrt{q}\eta) \right]$$
$$- \frac{\hat{m}^2}{2p}(\Phi\theta_\star)^\top \left(\lambda \mathrm{I_d} + \hat{\mathrm{v}}\Omega\right)^{-1} \Phi\boldsymbol{\theta}_0 - \frac{\hat{q}}{2p}\mathrm{tr}\left(\Omega\left(\lambda\mathrm{I_d} + \hat{\mathrm{v}}\Omega\right)^{-1}\right)$$

With the following auxiliary functions:

$$Z_\star(y, \omega, v) = \mathbb{E}_{z\sim\mathcal{N}(\omega,v)}[p_\star(y|z)] \qquad \mathscr{M}_{\tau\ell(y,\cdot)}(x) = \min_z \left[\frac{1}{2\tau}(z-x)^2 + \ell(y,z) \right]$$

Main result

Theorem (informal): for convex losses ℓ and under mild conditions on $(f_0, \hat{f})$ and $(\Psi, \Omega, \Phi, \theta_0)$, the asymptotic errors are given by:

$$\mathscr{R}(\hat{a}) \xrightarrow[d\to\infty]{} R(m_\star, q_\star) \qquad \hat{\mathscr{R}}_n(\hat{a}) \xrightarrow[d\to\infty]{} \hat{R}(v_\star, m_\star, q_\star)$$

Where $v_\star, m_\star, q_\star \in \mathbb{R}_+$ extremise the following potential function

$$\min_{v,q,m} \max_{\hat{v},\hat{q},\hat{m}} \frac{1}{2}(\hat{v}q - v\hat{q}) + \sqrt{\gamma} m\hat{m} + \alpha \mathbb{E}_{\eta\sim\mathcal{N}(0,1)}\left[Z_\star\left(y, \frac{m}{\sqrt{q}}\eta, \rho - \frac{m^2}{q}\right) \mathscr{M}_{v\ell(y,\cdot)}(\sqrt{q}\eta)\right]$$
$$-\frac{\hat{m}^2}{2p}(\Phi\theta_\star)^\top\left(\lambda \mathrm{I}_\mathrm{d} + \hat{v}\Omega\right)^{-1}\Phi\boldsymbol{\theta}_0 - \frac{\hat{q}}{2p}\mathrm{tr}\left(\Omega\left(\lambda\mathrm{I}_\mathrm{d} + \hat{v}\Omega\right)^{-1}\right)$$

With the following auxiliary functions:

$$Z_\star(y, \omega, v) = \mathbb{E}_{z\sim\mathcal{N}(\omega,v)}[p_\star(y|z)] \qquad \mathscr{M}_{\tau\ell(y,\cdot)}(x) = \min_z \left[\frac{1}{2\tau}(z-x)^2 + \ell(y,z)\right]$$

Comments

- Result holds for a general Gaussian Covariate model

$$y = f_\star(\theta_\star^\top u)$$

$$f(x; \Theta) = \hat{f}(a^\top v)$$

$$(u, v) \sim \mathcal{N}\left(0, \begin{bmatrix} \Psi & \Phi \\ \Phi^\top & \Omega \end{bmatrix}\right)$$

Comments

- Result holds for a general Gaussian Covariate model

$$y = f_\star(\theta_\star^\top u)$$

$$f(x;\Theta) = \hat{f}(a^\top v)$$

$$(u, v) \sim \mathcal{N}\left(0, \begin{bmatrix} \Psi & \Phi \\ \Phi^\top & \Omega \end{bmatrix}\right)$$

[Loureiro et al. ’21]

- RF case given by:

$$\Psi = I_d \qquad \Phi = \kappa_1 W \qquad \Omega = \kappa_0 1 1^\top + \kappa_1^2 W W^\top + \kappa_\star^2 I_p$$

$$\kappa_0 = \mathbb{E}[\sigma(z)] \qquad \kappa_1 = \mathbb{E}[\sigma'(z)] \qquad \kappa_\star^2 = \mathbb{E}[\sigma(z)^2] - \kappa_1^1 - \kappa_0^2$$

[Mei & Montanari ’19; Gerace et al. ’20]

Comments

- Result holds for a general Gaussian Covariate model

$$y = f_\star(\theta_\star^\top u) \qquad (u, v) \sim \mathcal{N}\left(0, \begin{bmatrix} \Psi & \Phi \\ \Phi^\top & \Omega \end{bmatrix}\right)$$

$$f(x; \Theta) = \hat{f}(a^\top v)$$

[Loureiro et al. '21]

- RF case given by:

$$\Psi = I_d \qquad \Phi = \kappa_1 W \qquad \Omega = \kappa_0 1 1^\top + \kappa_1^2 W W^\top + \kappa_\star^2 I_p$$

$$\kappa_0 = \mathbb{E}[\sigma(z)] \qquad \kappa_1 = \mathbb{E}[\sigma'(z)] \qquad \kappa_\star^2 = \mathbb{E}[\sigma(z)^2] - \kappa_1^1 - \kappa_0^2$$

[Mei & Montanari '19; Gerace et al. '20]

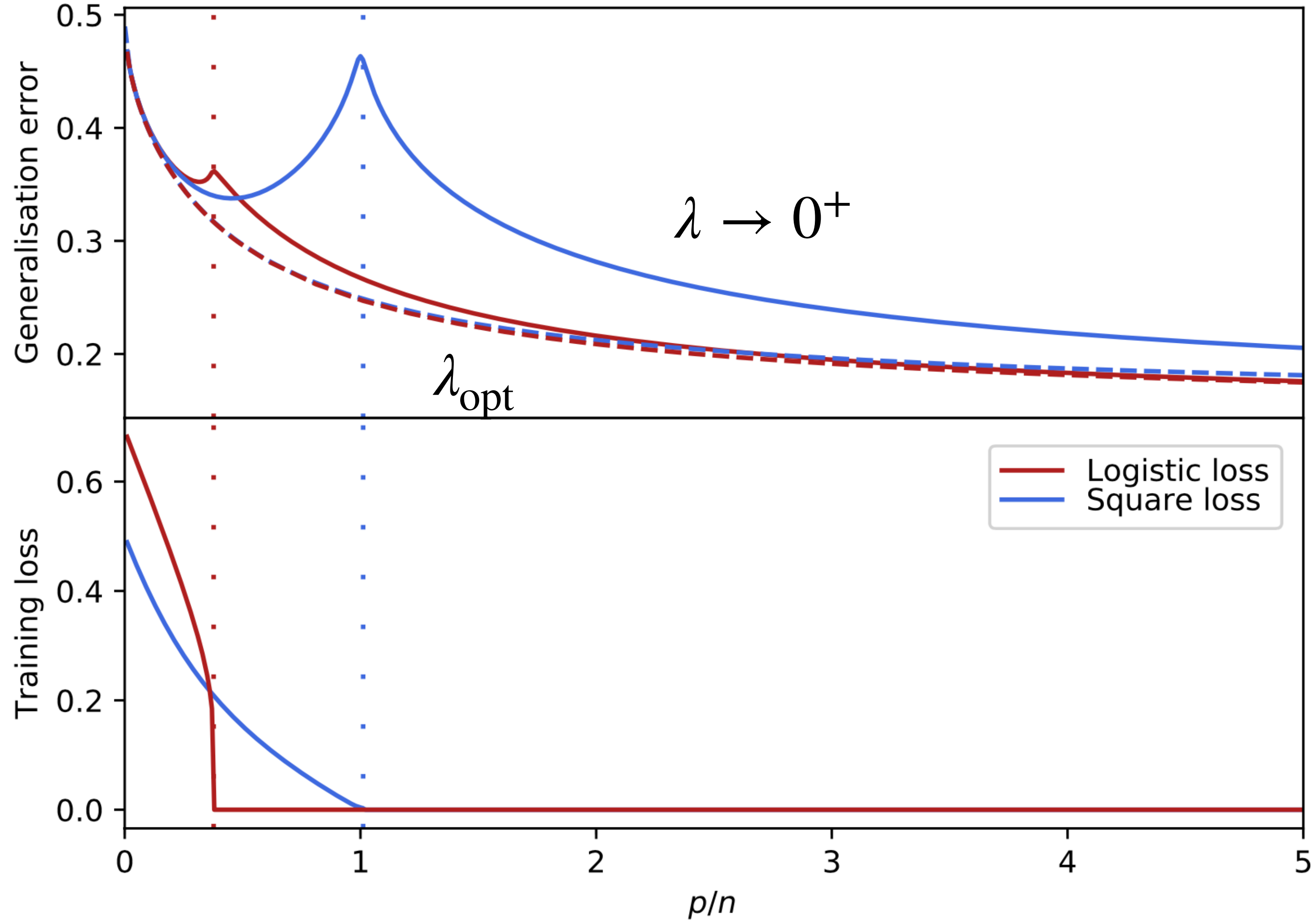
- This covers "many" feature maps $\varphi(x)$ due to universality in high-dimensions. [Goldt et al. '21; Hu & Lu '21; Saed & Montanari '22]

  Including deep case $\varphi(x) = \sigma(W_L \sigma(\cdots \sigma(W_1 x)))$ [Schröder '23]

Phenomenology

Double descent

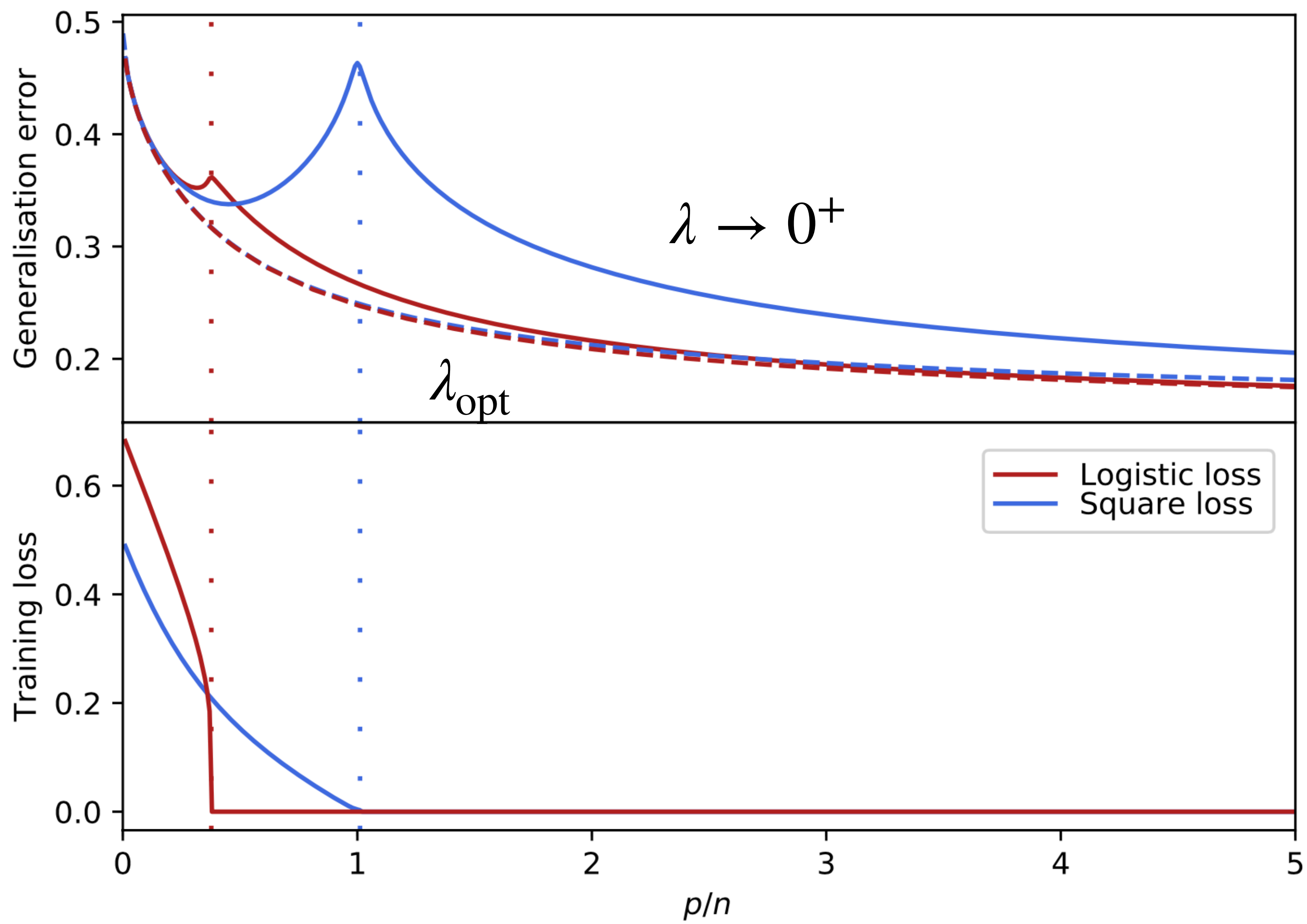
$$y^\mu = \text{sign}(w_*^\top x) \qquad \hat{y} = \text{sign}\left(w_2^\top \text{erf}\left(\text{W}_1 \text{x}\right)\right)$$

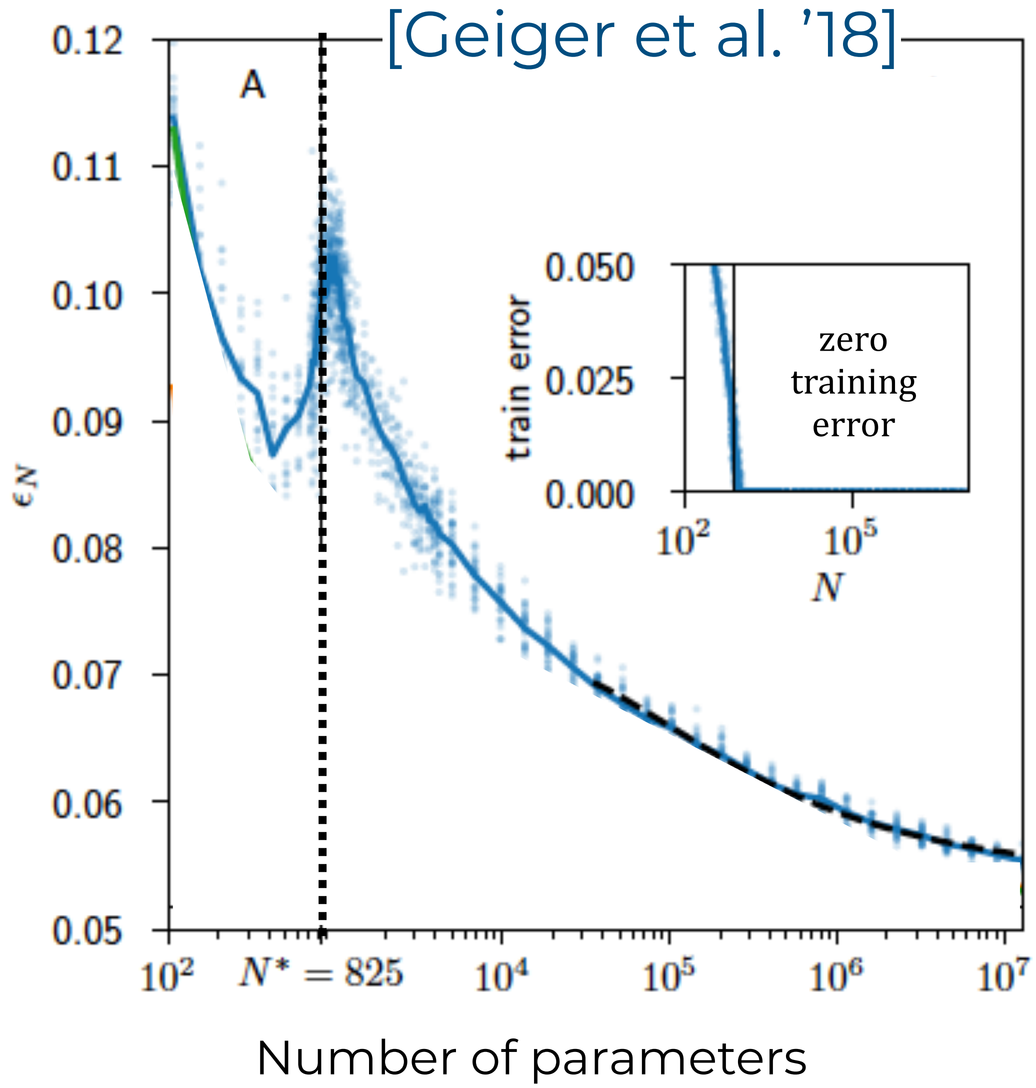


[Mei & Montanari '19; Gerace et al. '20]

Double descent

$$y^\mu = \mathrm{sign}(w_*^\top x) \qquad \hat{y} = \mathrm{sign}\left(w_2^\top \mathrm{erf}\left(W_1 x\right)\right)$$





[Mei & Montanari '19; Gerace et al. '20]

What's going on?

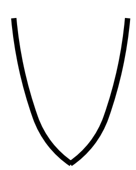
Focus on ℓ_2 loss $\lambda \to 0^+$.

$p < n$

p/n

underparametrised

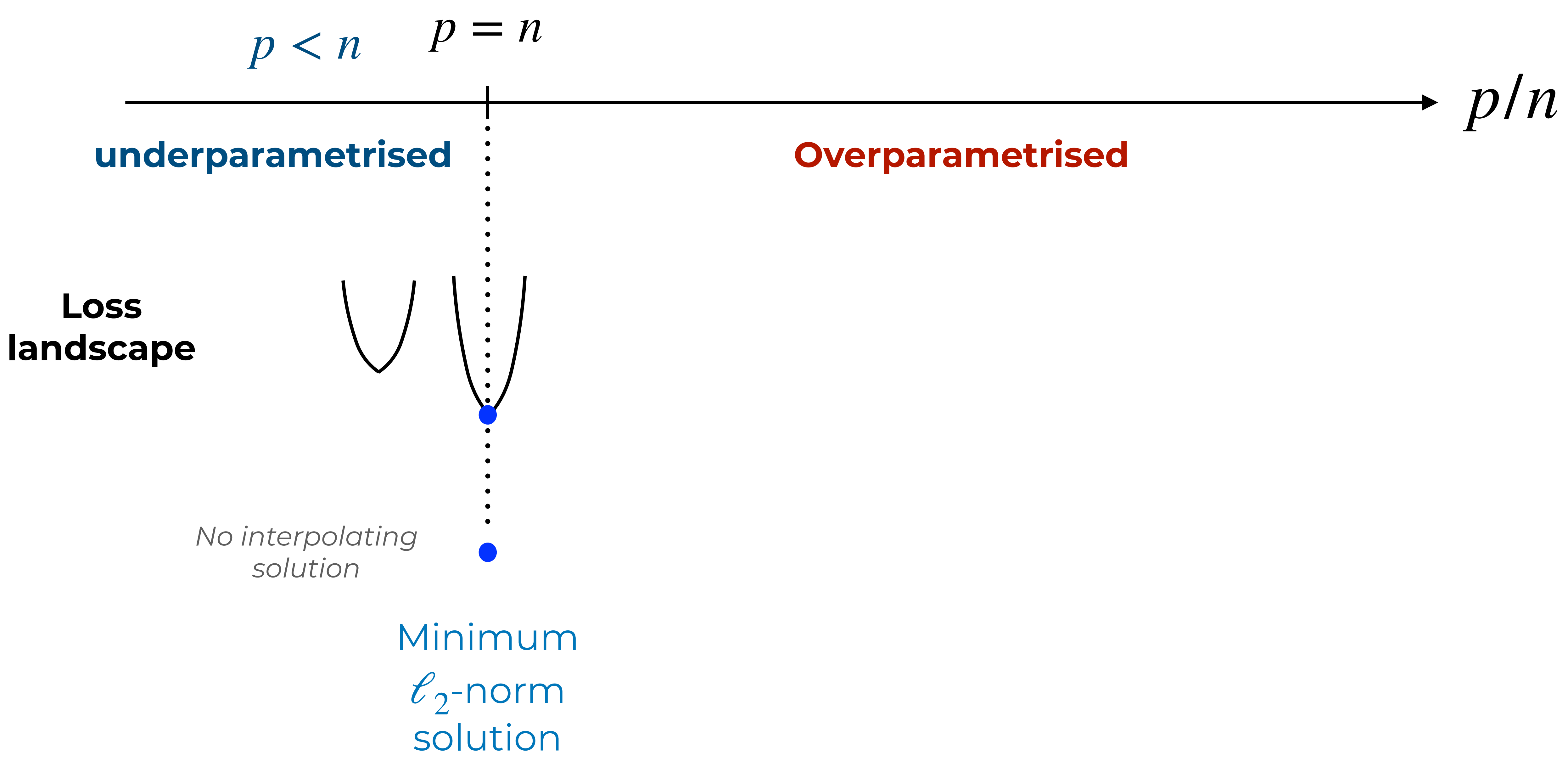
Loss landscape



No interpolating solution

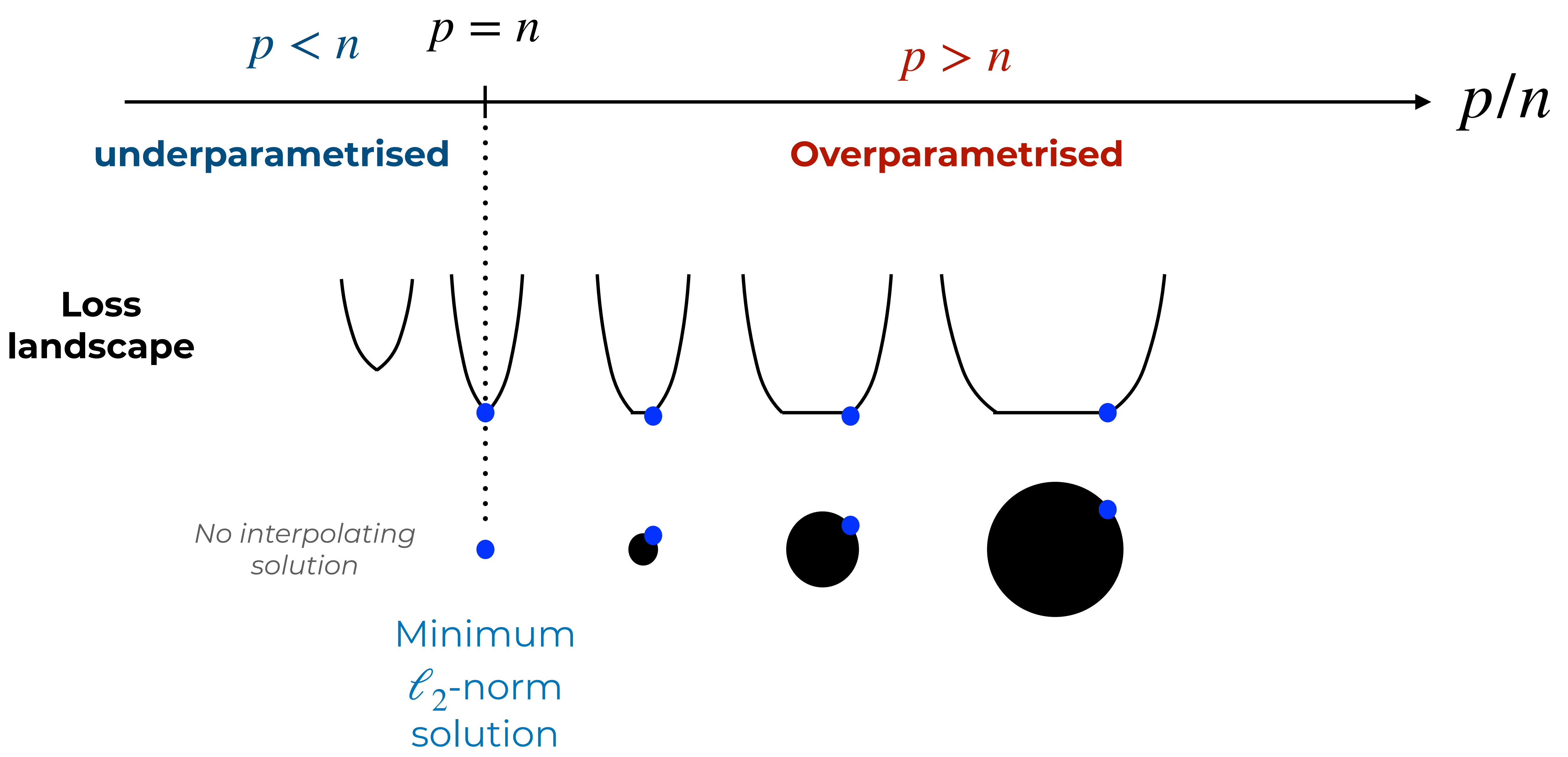
What's going on?

Focus on ℓ_2 loss $\lambda \to 0^+$.

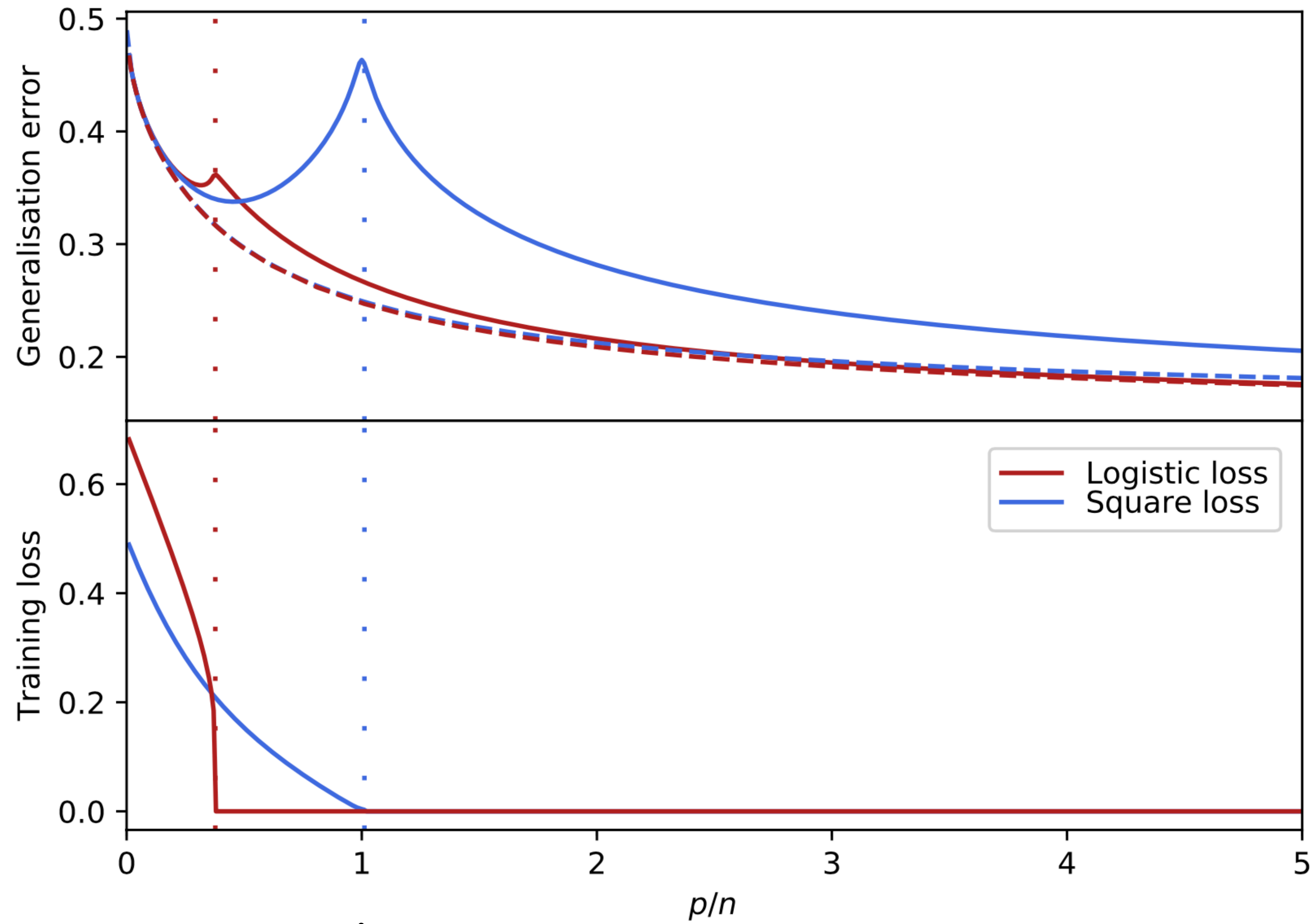


What's going on?

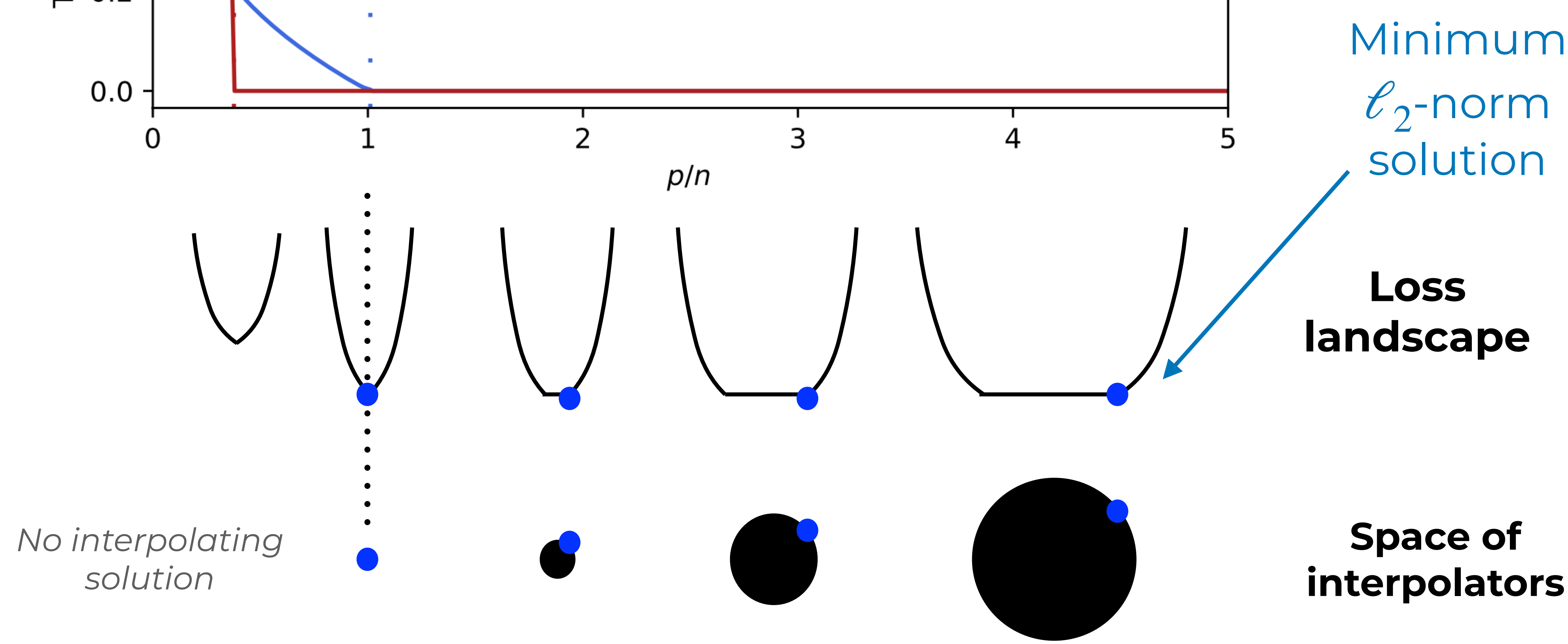
Focus on ℓ_2 loss $\lambda \to 0^+$.



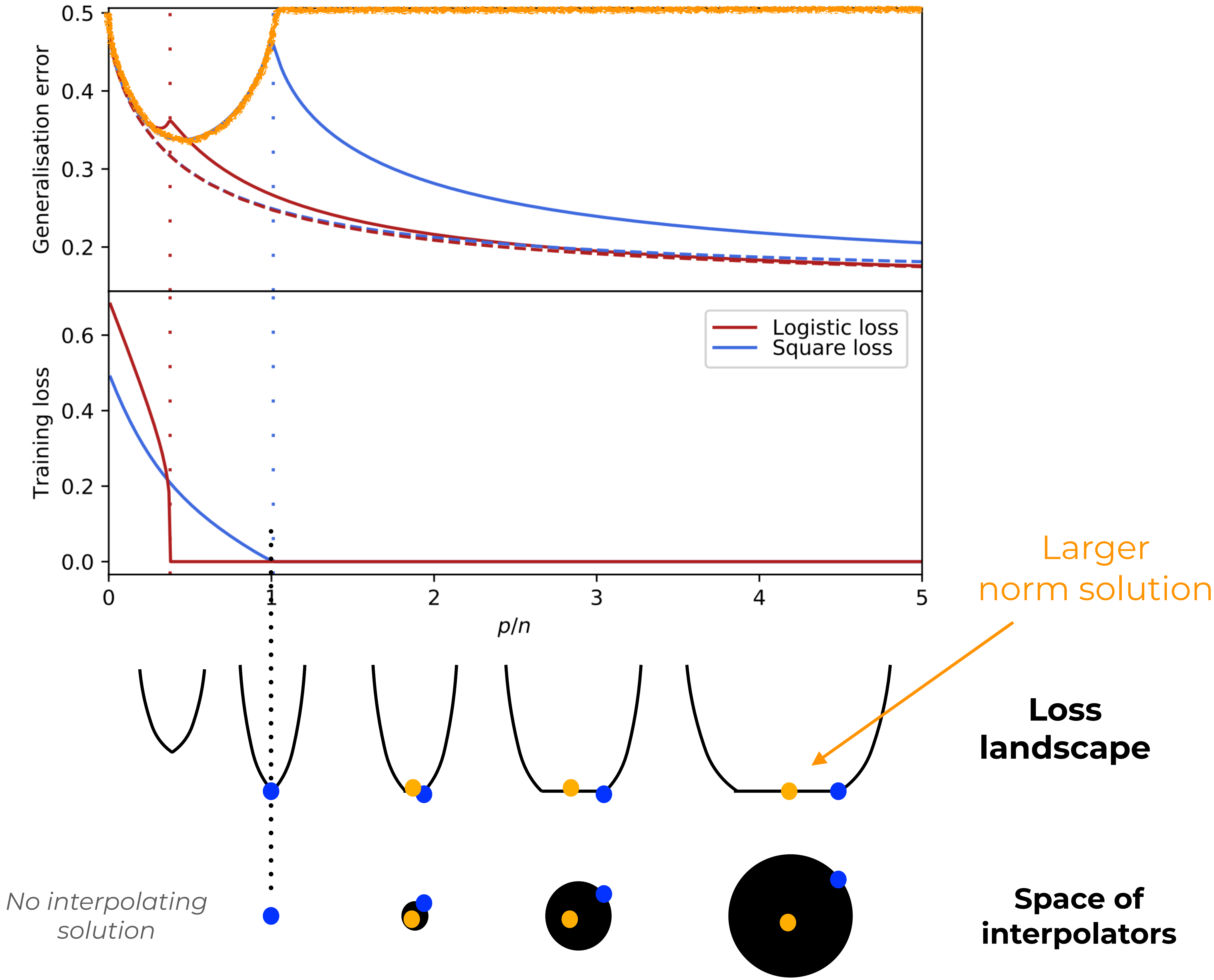
What's going on?



ℓ_2 loss
$\lambda \to 0^+$.

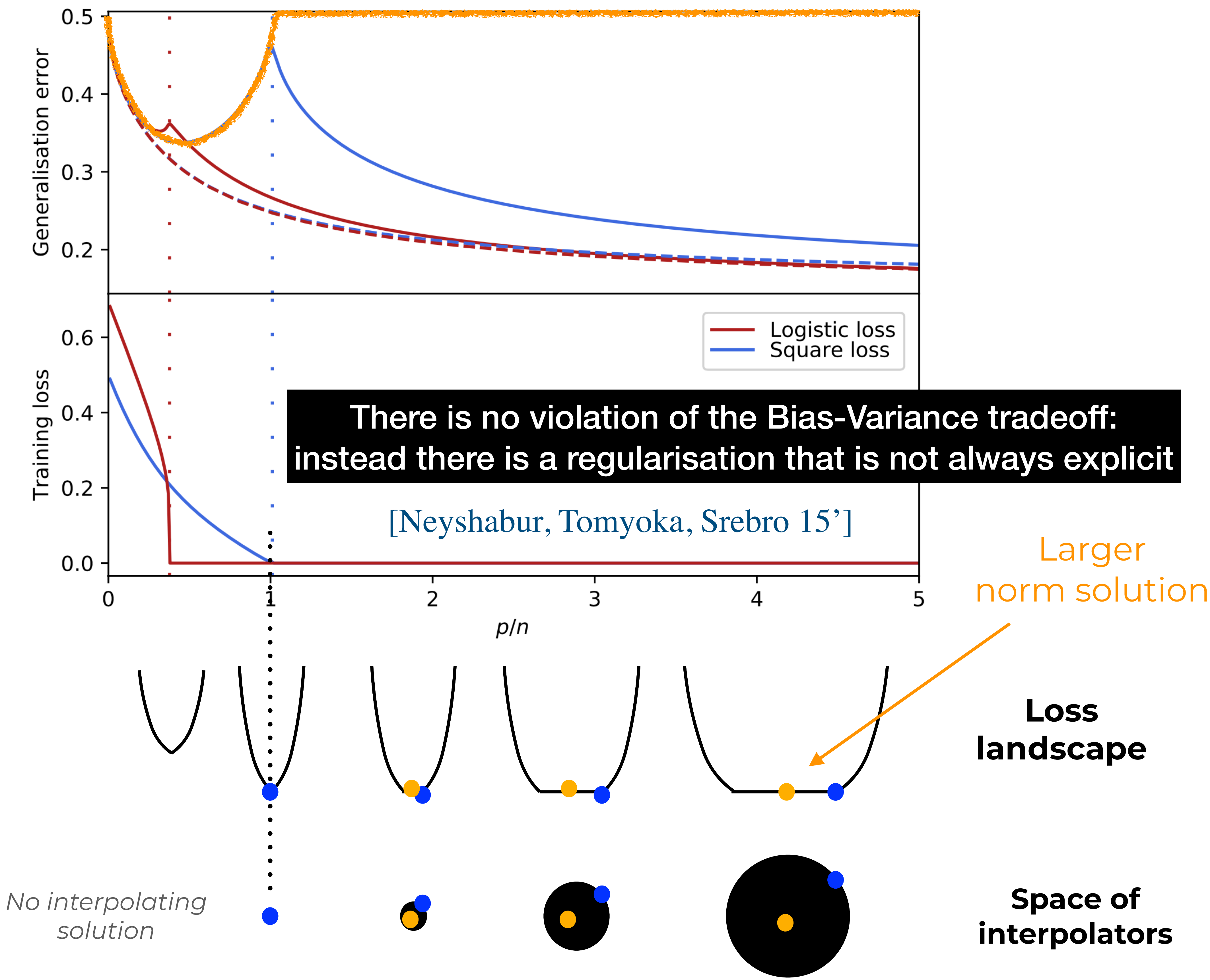


What's going on?



ℓ_2 loss
$\lambda \to 0^+$.

What's going on?



There is no violation of the Bias-Variance tradeoff: instead there is a regularisation that is not always explicit

[Neyshabur, Tomyoka, Srebro 15']

ℓ_2 loss
$\lambda \to 0^+.$

What about hinge / logistic?

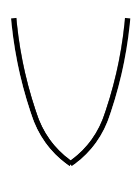
$\lambda \to 0^+$ $\qquad \ell(y,x) = \log(1 + e^{-yx})$ $\qquad \ell(y,x) = (1 - yx)_+$

$p < 2n$

p/n

underparametrised

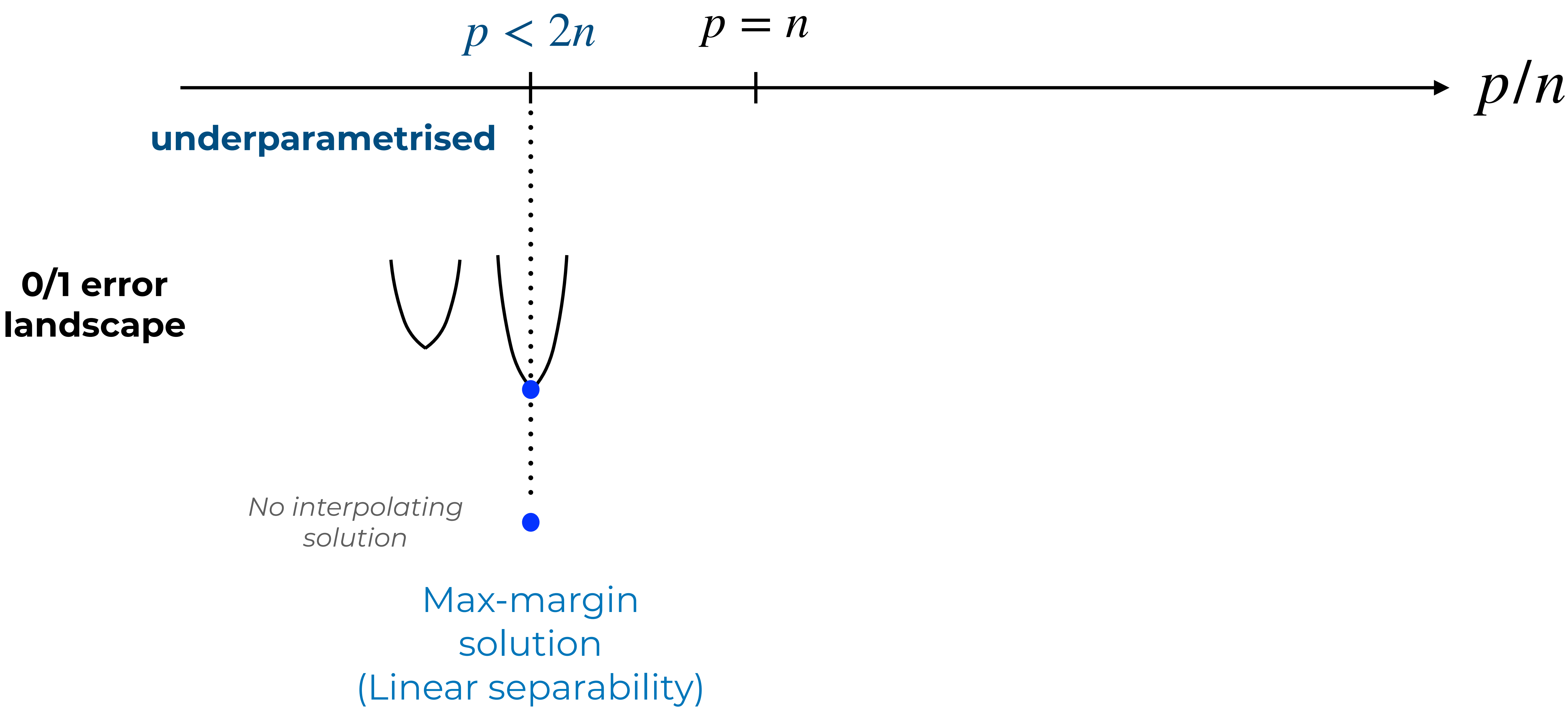
0/1 error landscape



No interpolating solution

What about hinge / logistic?

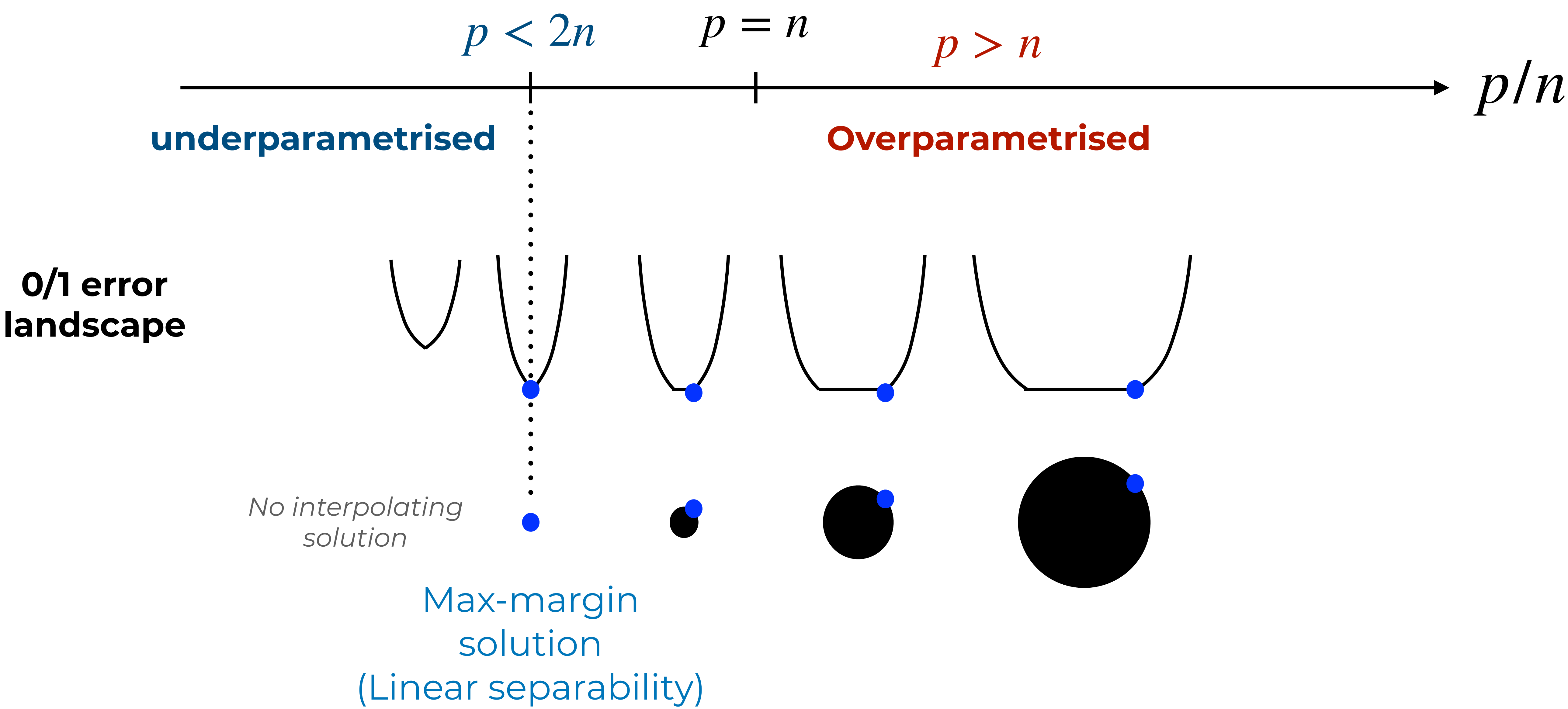
$$\lambda \to 0^+ \qquad \ell(y,x) = \log(1 + e^{-yx}) \qquad \ell(y,x) = (1 - yx)_+$$



[Rosset, Zhy, Hastie 04']

What about hinge / logistic?

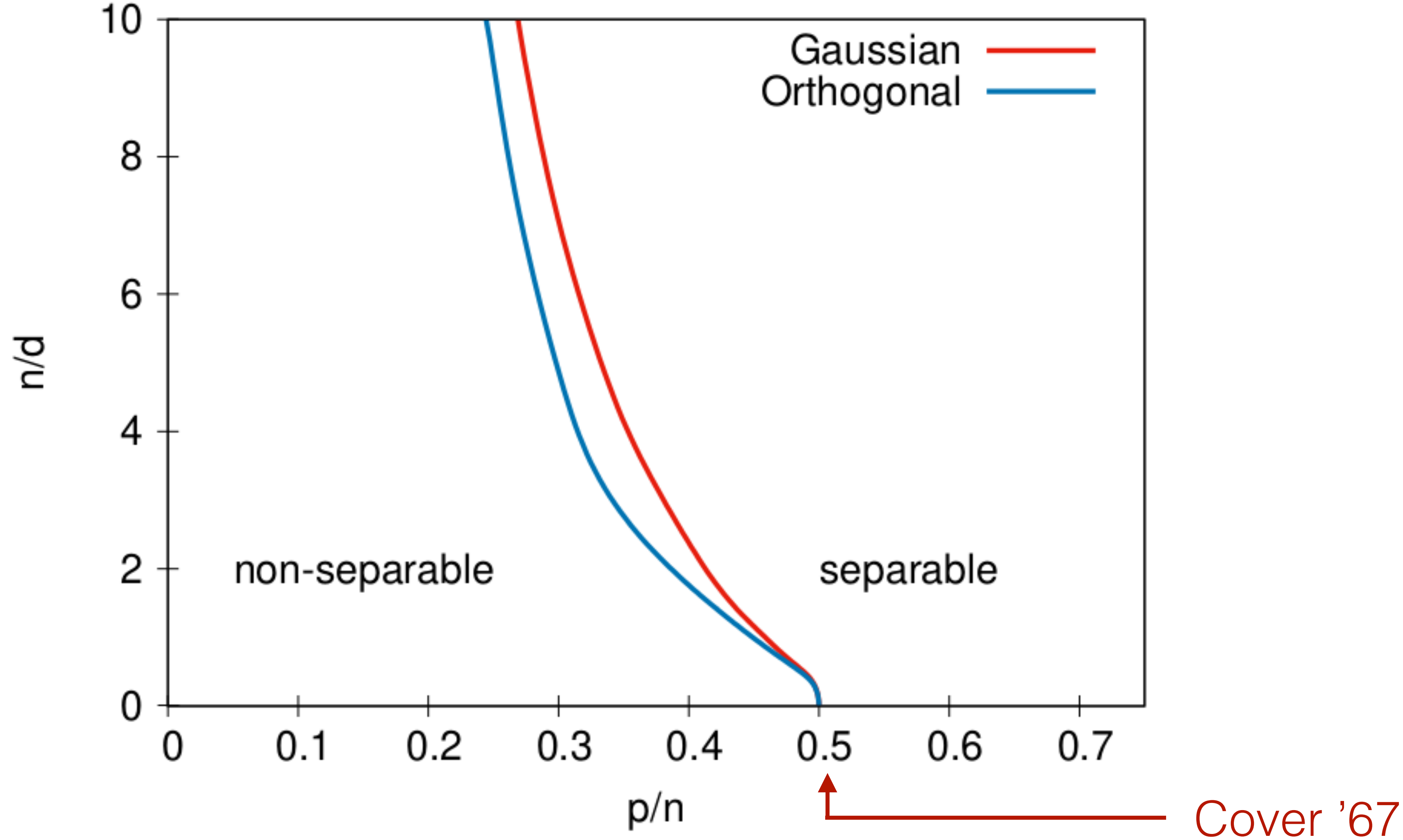
$\lambda \to 0^+$ $\quad \ell(y,x) = \log(1+e^{-yx})$ $\quad \ell(y,x) = (1-yx)_+$



[Rosset, Zhy, Hastie 04']

Linear separability

[Gerace et al. '20]



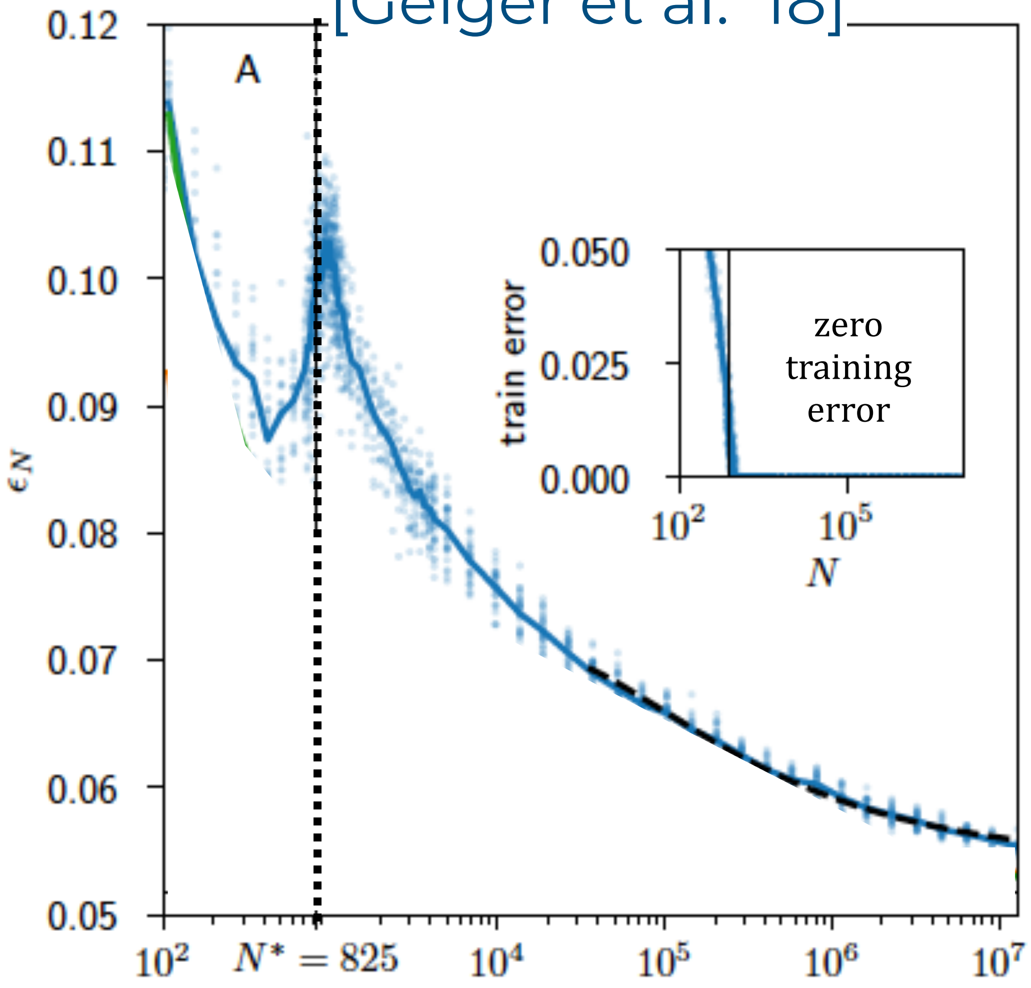
Generalise a phase transition in
[Cover '67; Gardner '87; Sur & Candes, '18]

Fluctuations and overfitting

Scaling description of generalization with number of parameters in deep learning

Mario Geiger[a,1], Arthur Jacot[b,1], Stefano Spigler[a], Franck Gabriel[b], Levent Sagun[a], Stéphane d'Ascoli[c], Giulio Biroli[c], Clément Hongler[b,2], and Matthieu Wyart[a,2]

[Geiger et al. '18]



Number of parameters

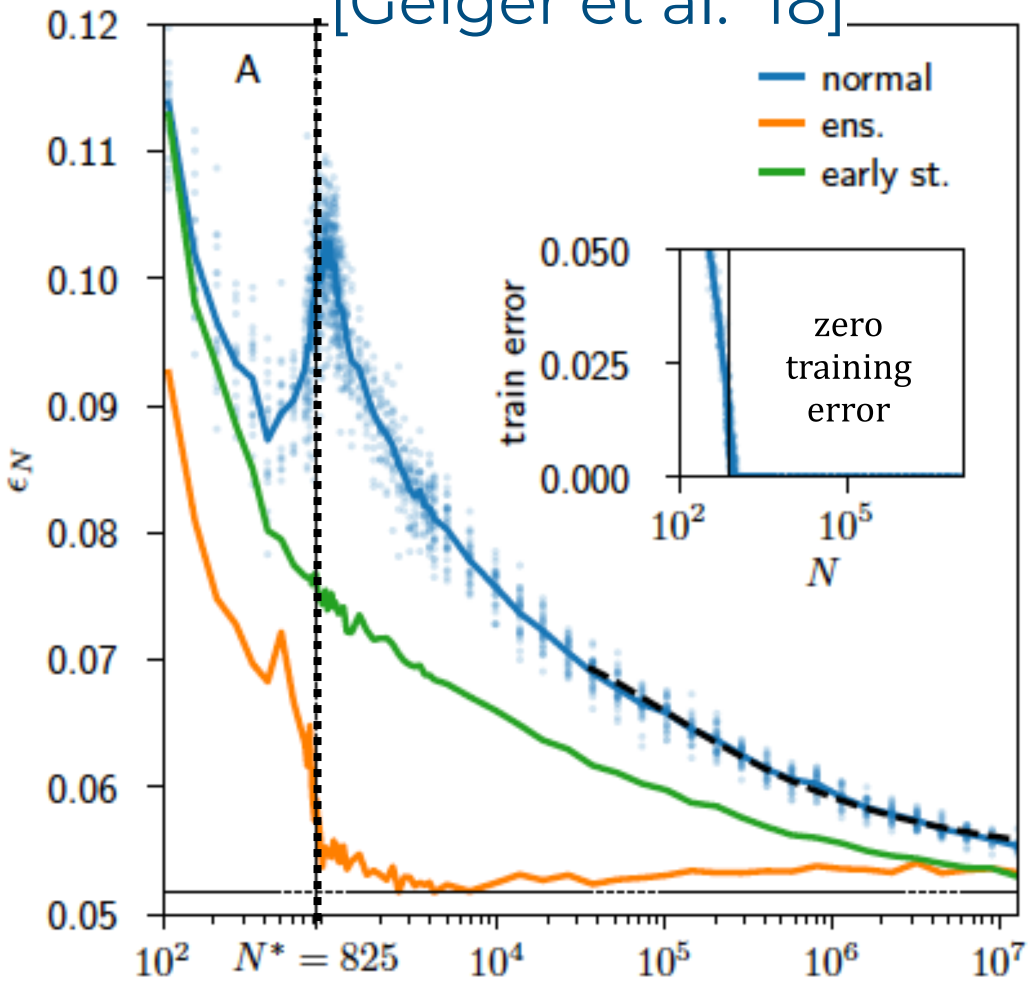
is correctly classified) [24, 25, 26, 27]. Indeed the test error (the probability of an incorrect classification for an unseen data point) has been observed to decrease as $N \to \infty$ in a slow power-law fashion [17]. In contrast, as $N \to N^*$, the test error blows up [27, 28, 17] (a phenomenon shown by the blue curve in Fig. 2). In the context of least-squares regression, the improvement of performance with N has been linked to the observed diminishing fluctuations of the DNN function after training [29], a result consistent with the notion of stronger implicit regularization with increasing N [30, 31]. This raises the question of understanding what controls these fluctuations and how they affect the test error in a classification task.

Fluctuations and overfitting

Scaling description of generalization with number of parameters in deep learning

Mario Geiger[a,1], Arthur Jacot[b,1], Stefano Spigler[a], Franck Gabriel[b], Levent Sagun[a], Stéphane d'Ascoli[c], Giulio Biroli[c], Clément Hongler[b,2], and Matthieu Wyart[a,2]

[Geiger et al. '18]

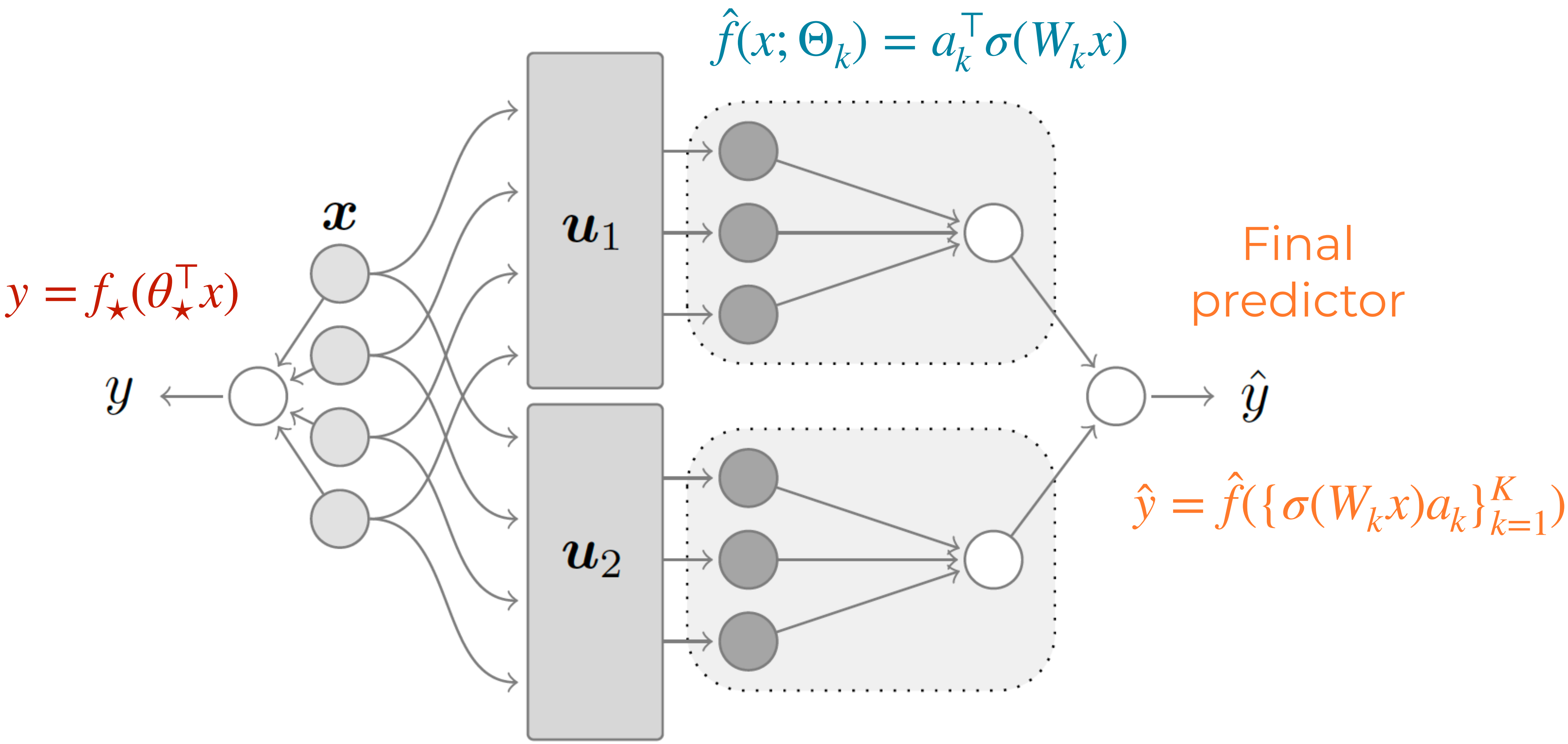


Number of parameters

is correctly classified) [24, 25, 26, 27]. Indeed the test error (the probability of an incorrect classification for an unseen data point) has been observed to decrease as $N \to \infty$ in a slow power-law fashion [17]. In contrast, as $N \to N^*$, the test error blows up [27, 28, 17] (a phenomenon shown by the blue curve in Fig. 2). In the context of least-squares regression, the improvement of performance with N has been linked to the observed diminishing fluctuations of the DNN function after training [29], a result consistent with the notion of stronger implicit regularization with increasing N [30, 31]. This raises the question of understanding what controls these fluctuations and how they affect the test error in a classification task.

Ensemble of random features

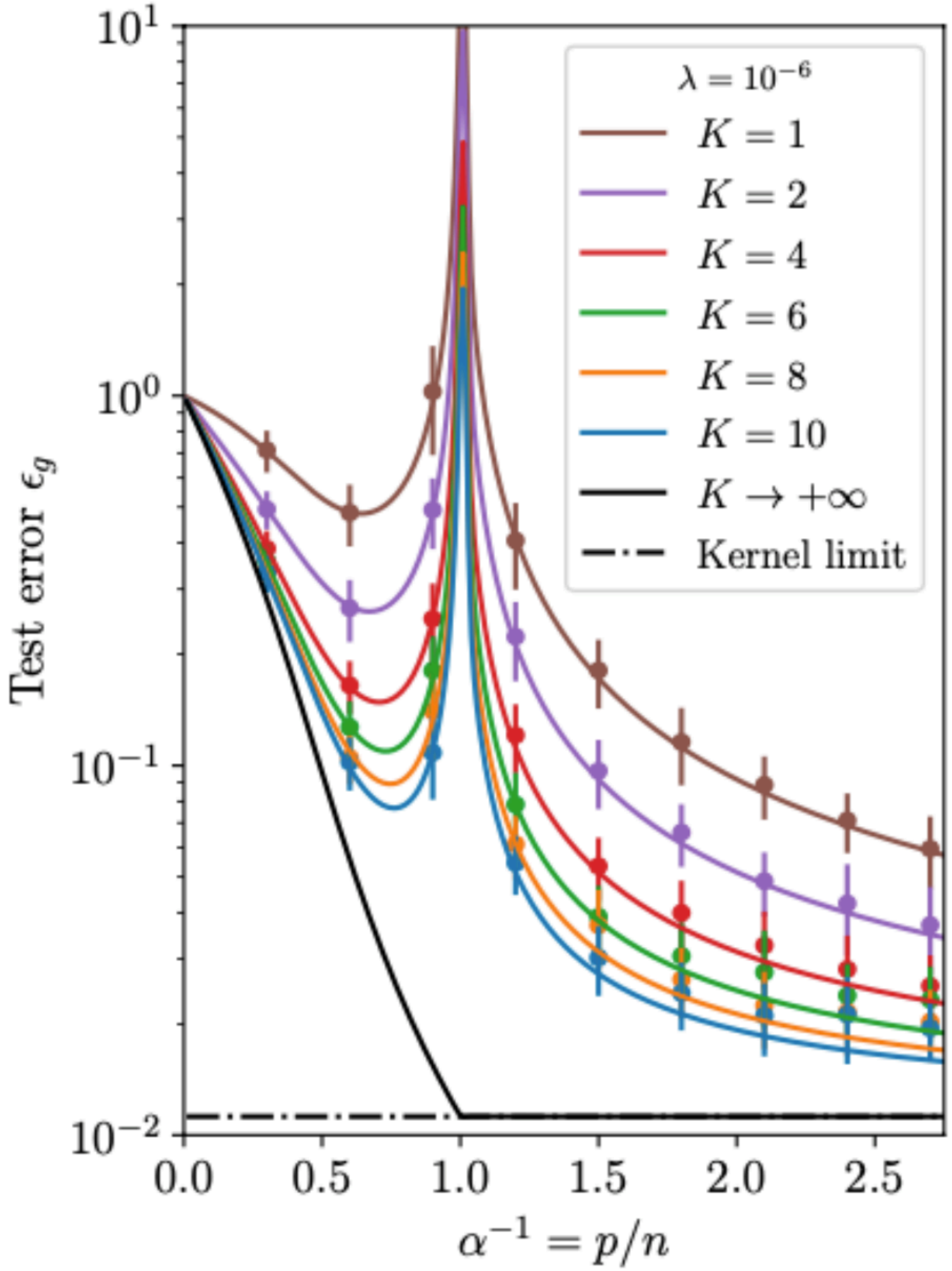
Can generalise previous discussion to an ensemble of learners:



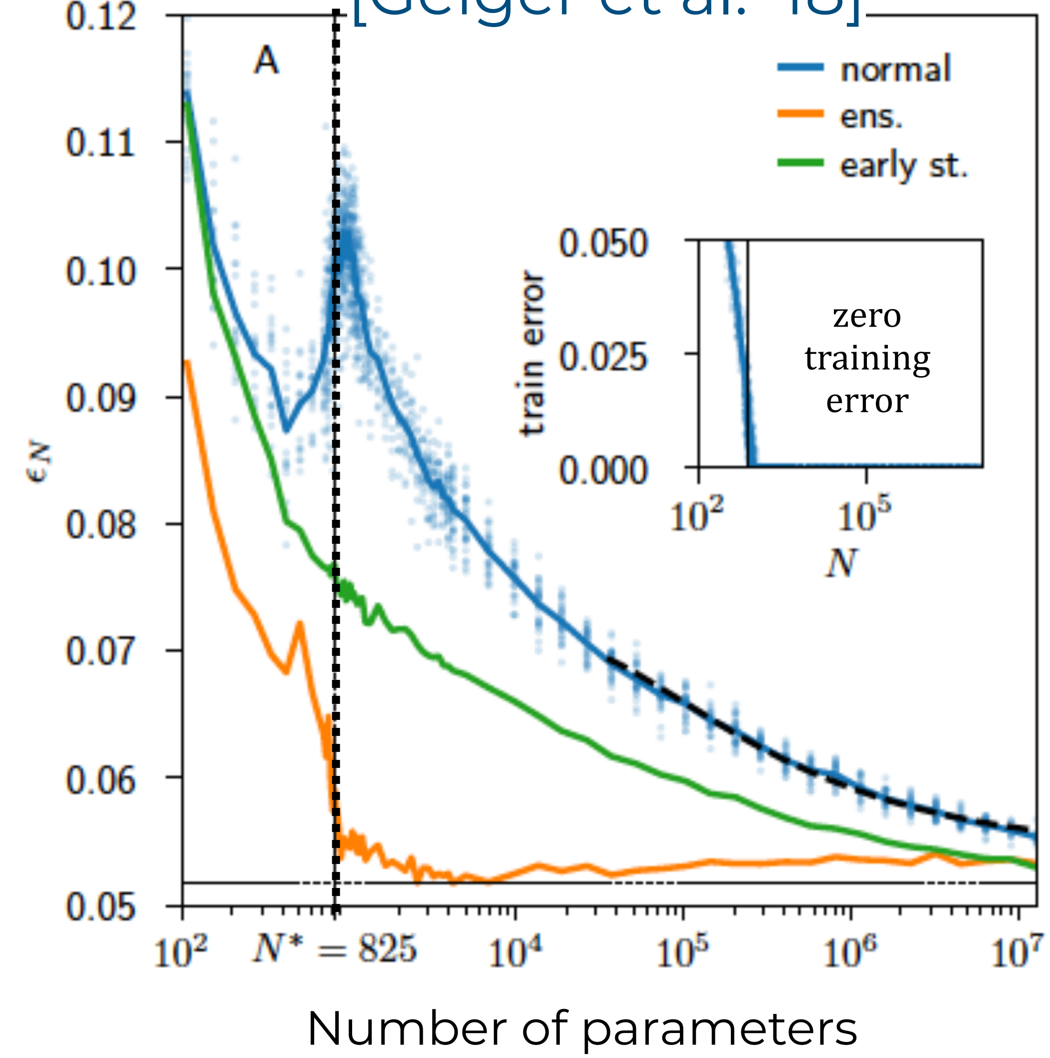
Example: $$\hat{y} = \frac{1}{K}\sum_{k=1}^{K} a_k^\top \sigma(W_k x)$$

Overfitting at interpolation

Ridge regression



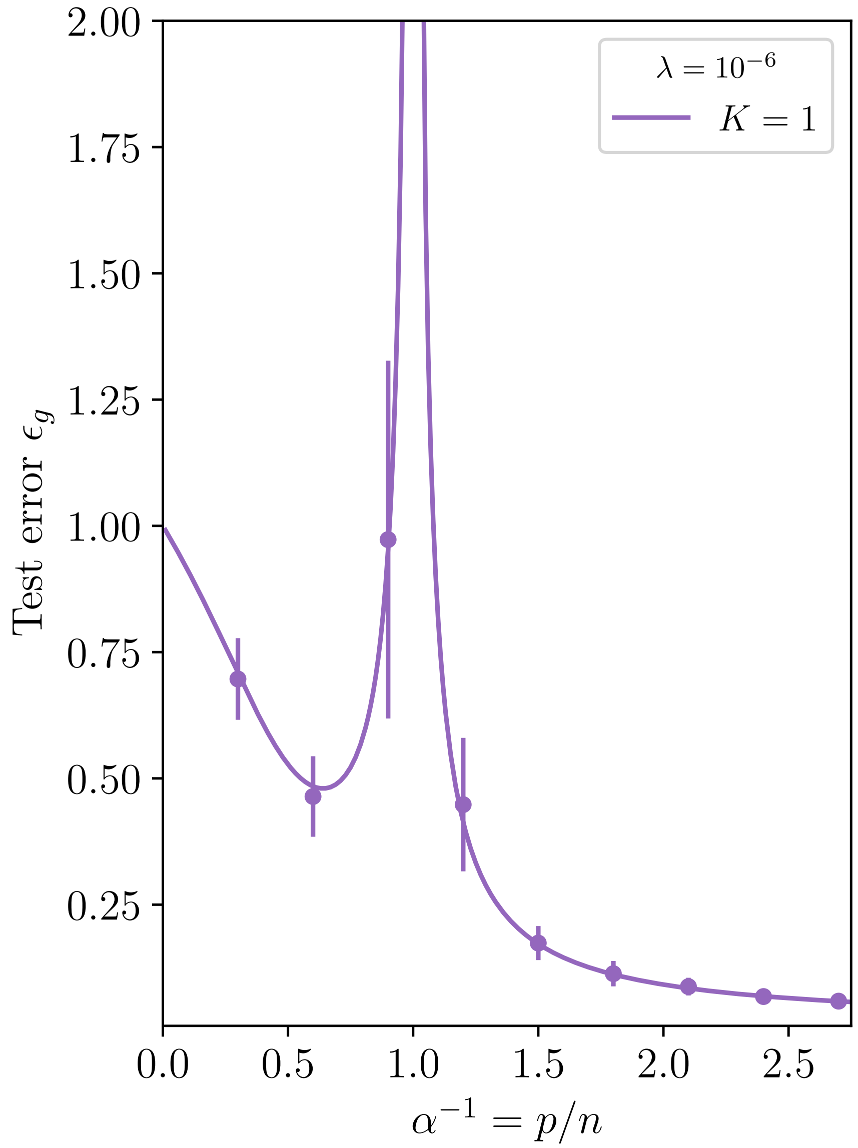
[Geiger et al. '18]



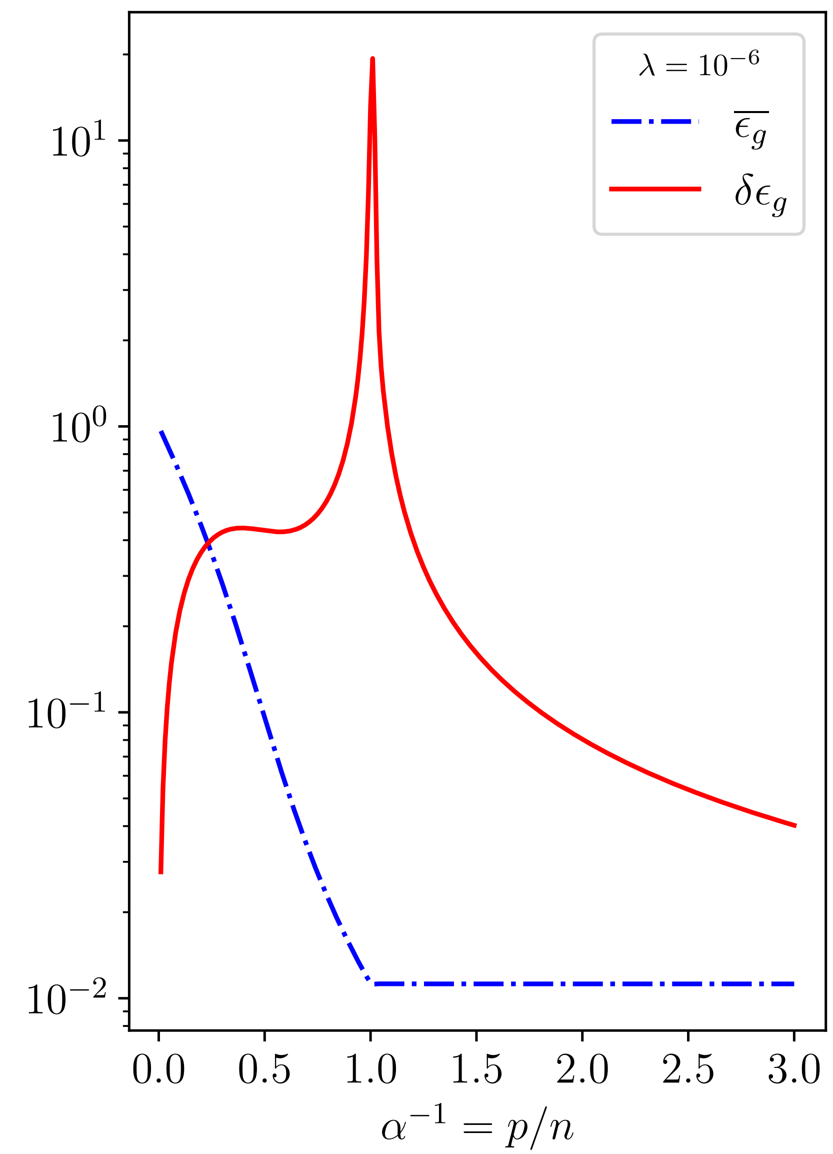
[Biroli et al '20; Loureiro et al. '22]

Bias-Variance trade-off

Ridge regression



$$\epsilon_g = \bar{\epsilon}_g + \delta\bar{\epsilon}_g$$



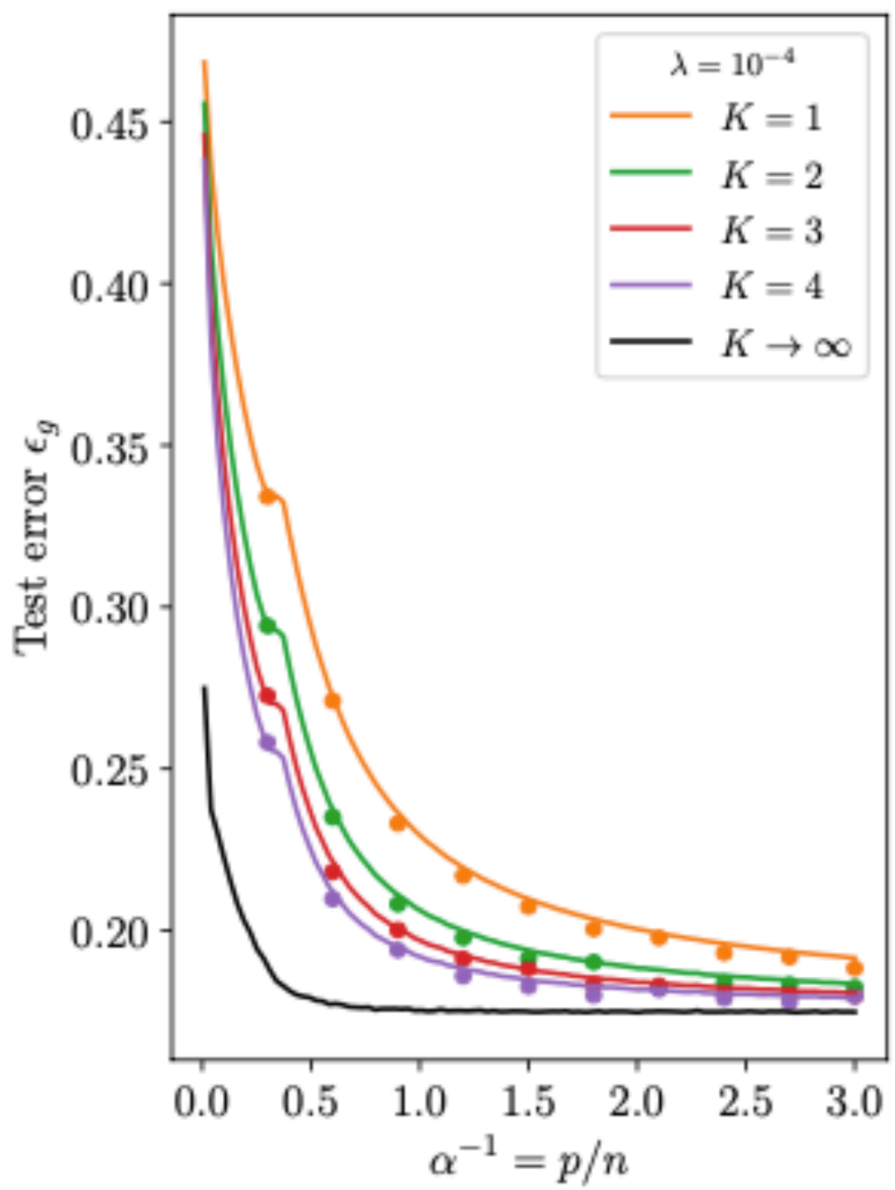
Bias (approximation error): decrease and vanishes at interpolation

Variance: overfitting of random weights W fluctuations

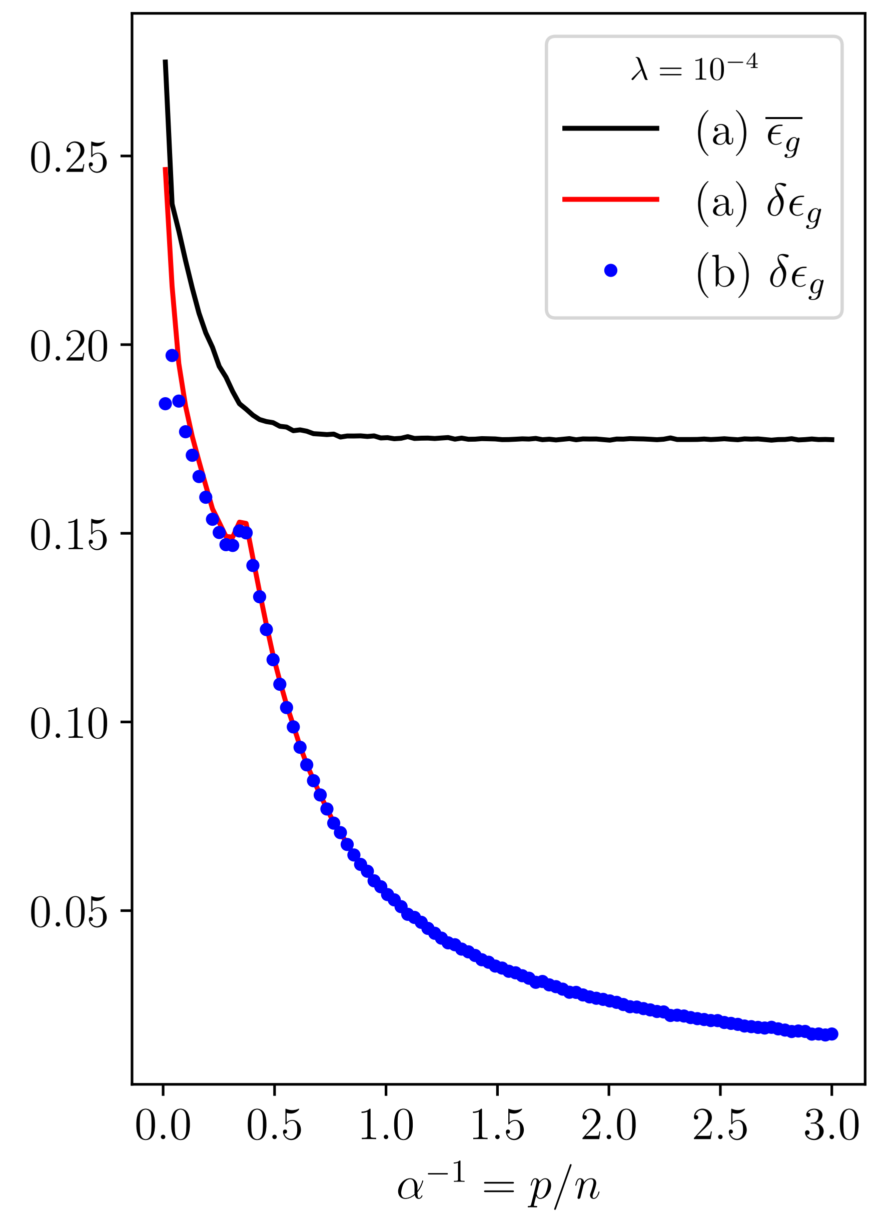
[Krogh, Vedelsby '95; Biroli et al '20; Loureiro et al. '22]

Bias-Variance trade-off

Logistic regression



$$\epsilon_g = \bar{\epsilon}_g + \delta\bar{\epsilon}_g$$



Bias (approximation error): decrease and vanishes at interpolation

Variance: overfitting of random weights W fluctuations

[Krogh, Vedelsby '95; Biroli et al '20; Loureiro et al. '22]

Limitations of RF model

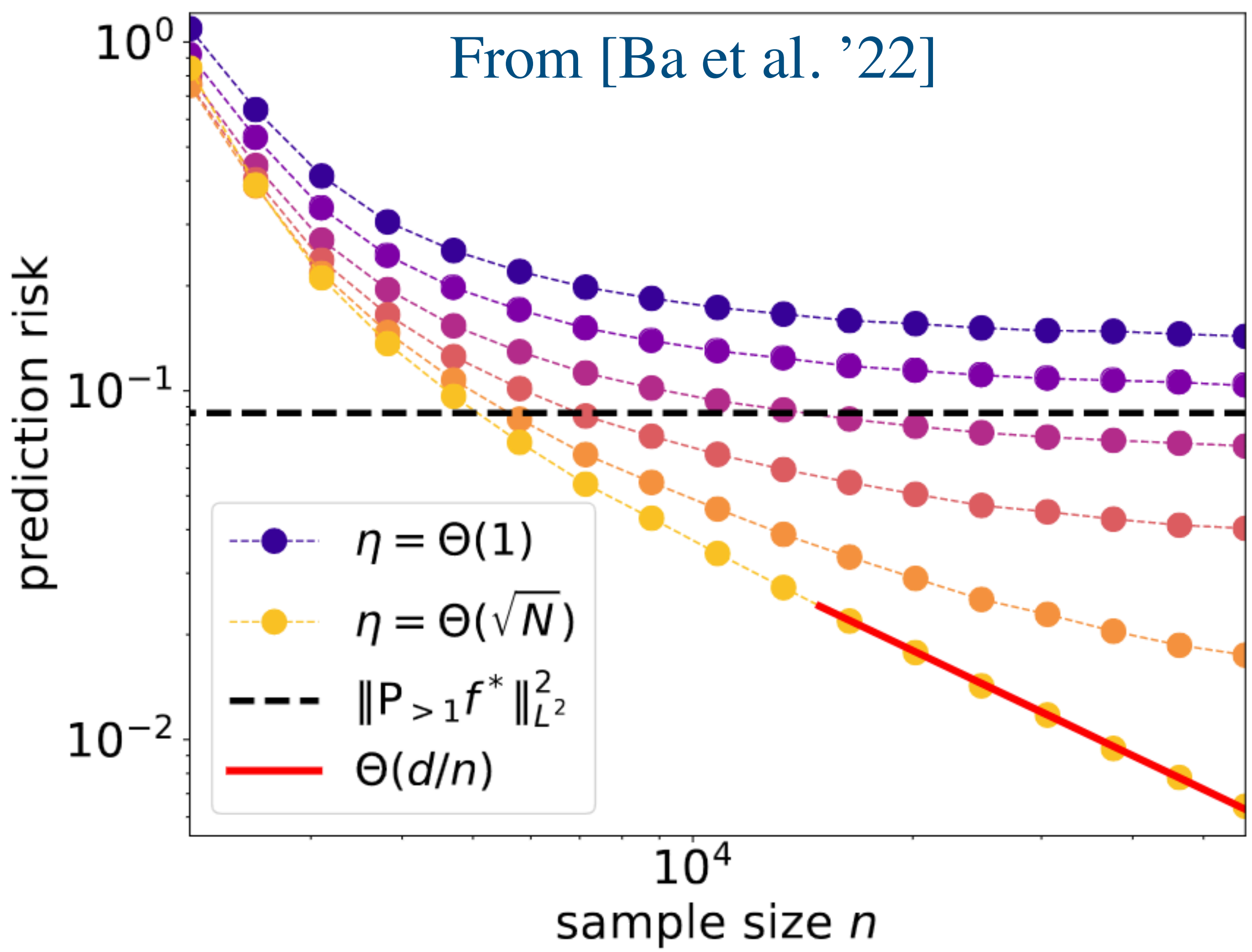
Performance of RF is bounded by kernel performance.

In this setting, for $n \propto d^{\ell}$ it can be shown that kernels can learn up to degree ℓ polynomials. [Mei, Misiakiewicz & Montanari '21]

Limitations of RF model

Performance of RF is bounded by kernel performance.

In this setting, for $n \propto d^{\ell}$ it can be shown that kernels can learn up to degree ℓ polynomials. [Mei, Misiakiewicz & Montanari '21]



Beating this requires
Learning features!

Lecture II: Summary

- ✓ “Lazy” vs “rich” regime of wide neural networks

Lecture II: Summary

- “Lazy” vs “rich” regime of wide neural networks
- Exact asymptotic results for RF model under Gaussian data

Lecture II: Summary

- “Lazy” vs “rich” regime of wide neural networks
- Exact asymptotic results for RF model under Gaussian data
- Overparametrisation might not hurt generalisation c.f. “Benign overfitting” [Bartlett et al. ’19]

Lecture II: Summary

- “Lazy” vs “rich” regime of wide neural networks
- Exact asymptotic results for RF model under Gaussian data
- Overparametrisation might not hurt generalisation c.f. “Benign overfitting” [Bartlett et al. ’19]
- Implicit bias of optimisation algorithm