Generalisation error in learning with random features and the hidden manifold model

B. Loureiro
(IPhT)

F. Gerace
(IPhT)

M. Mézard
(ENS)

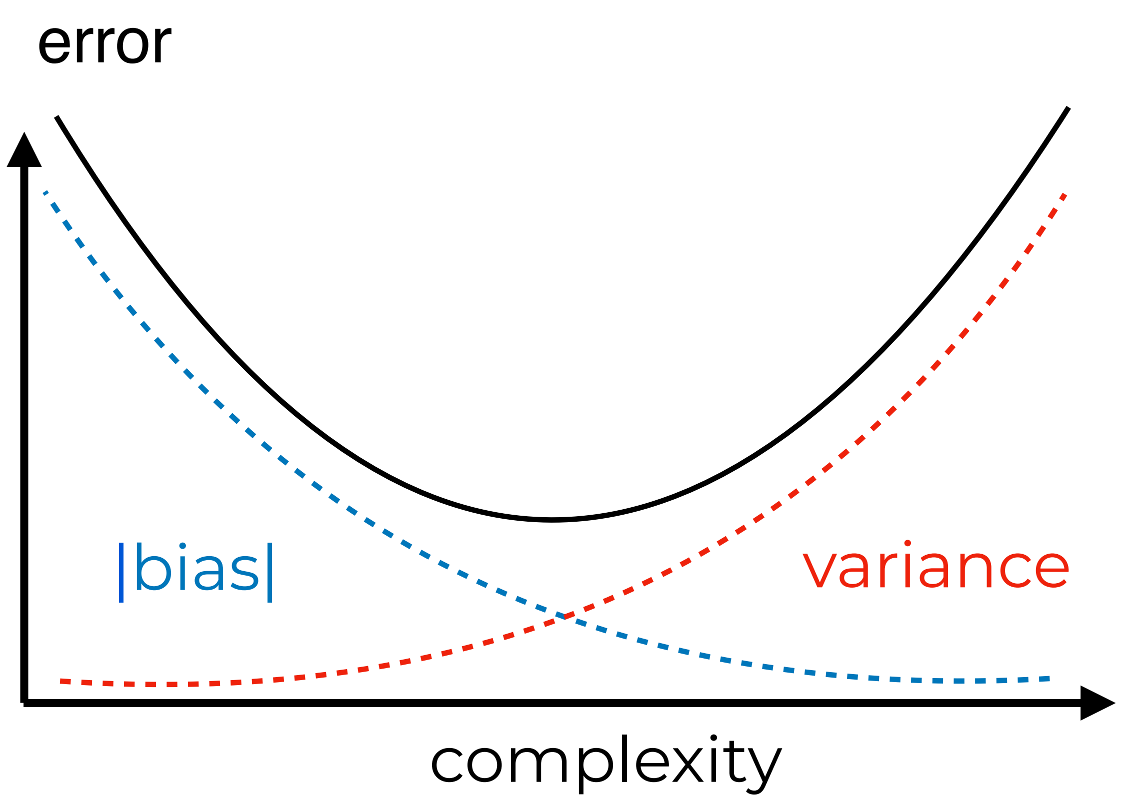
L. Zdeborová
(IPhT)

F. Krzakala
(ENS)

ICML 2020

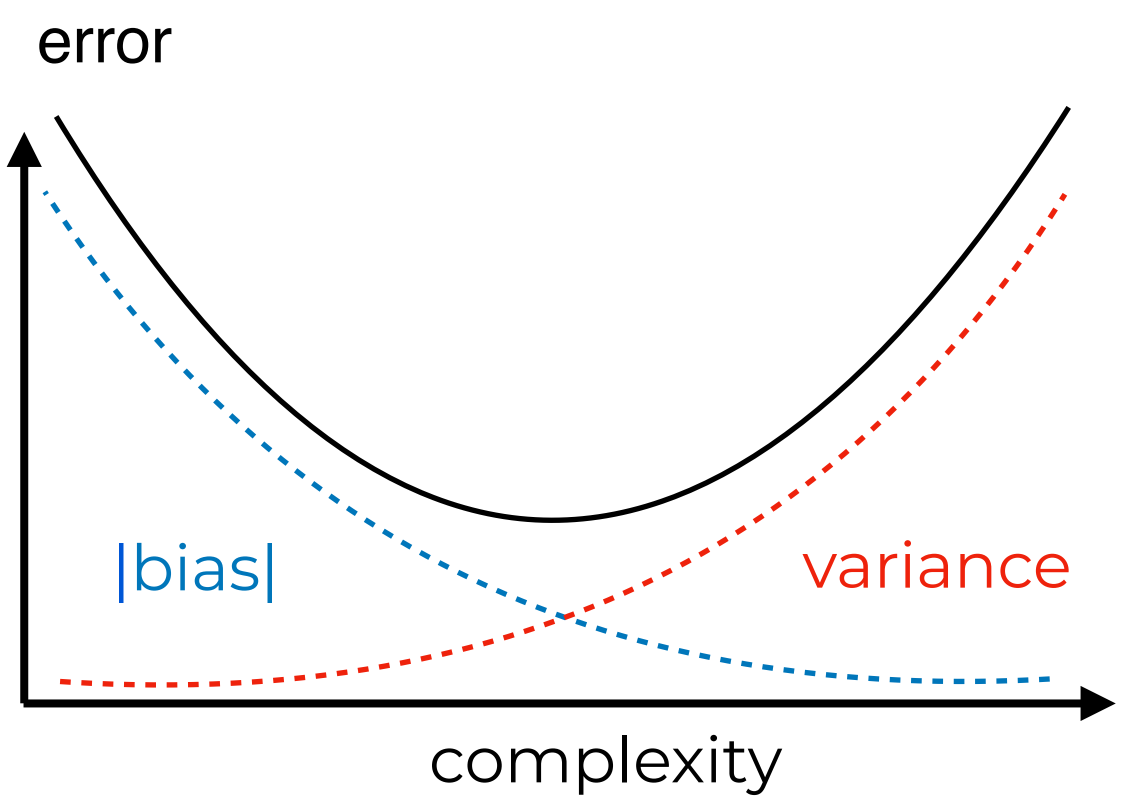
TRAINING A NEURAL NET

EXPECTATIONS

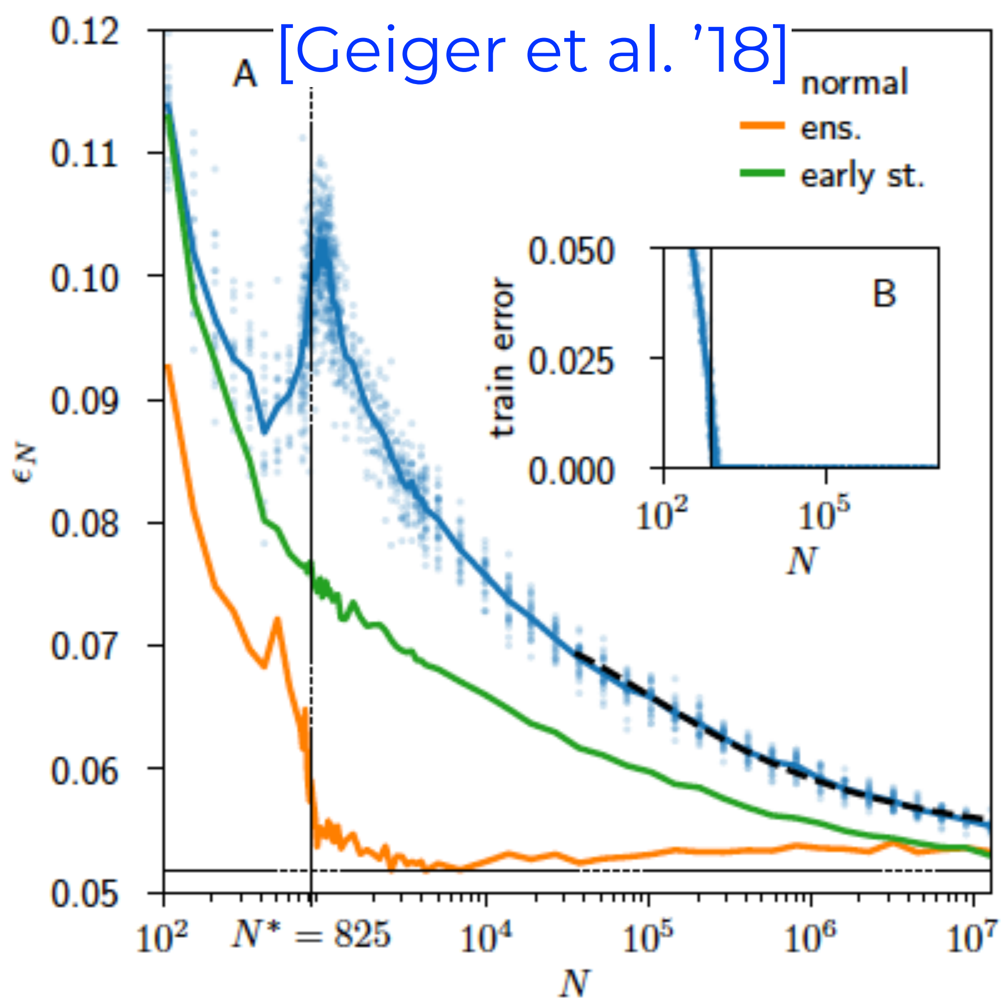


TRAINING A NEURAL NET

EXPECTATIONS



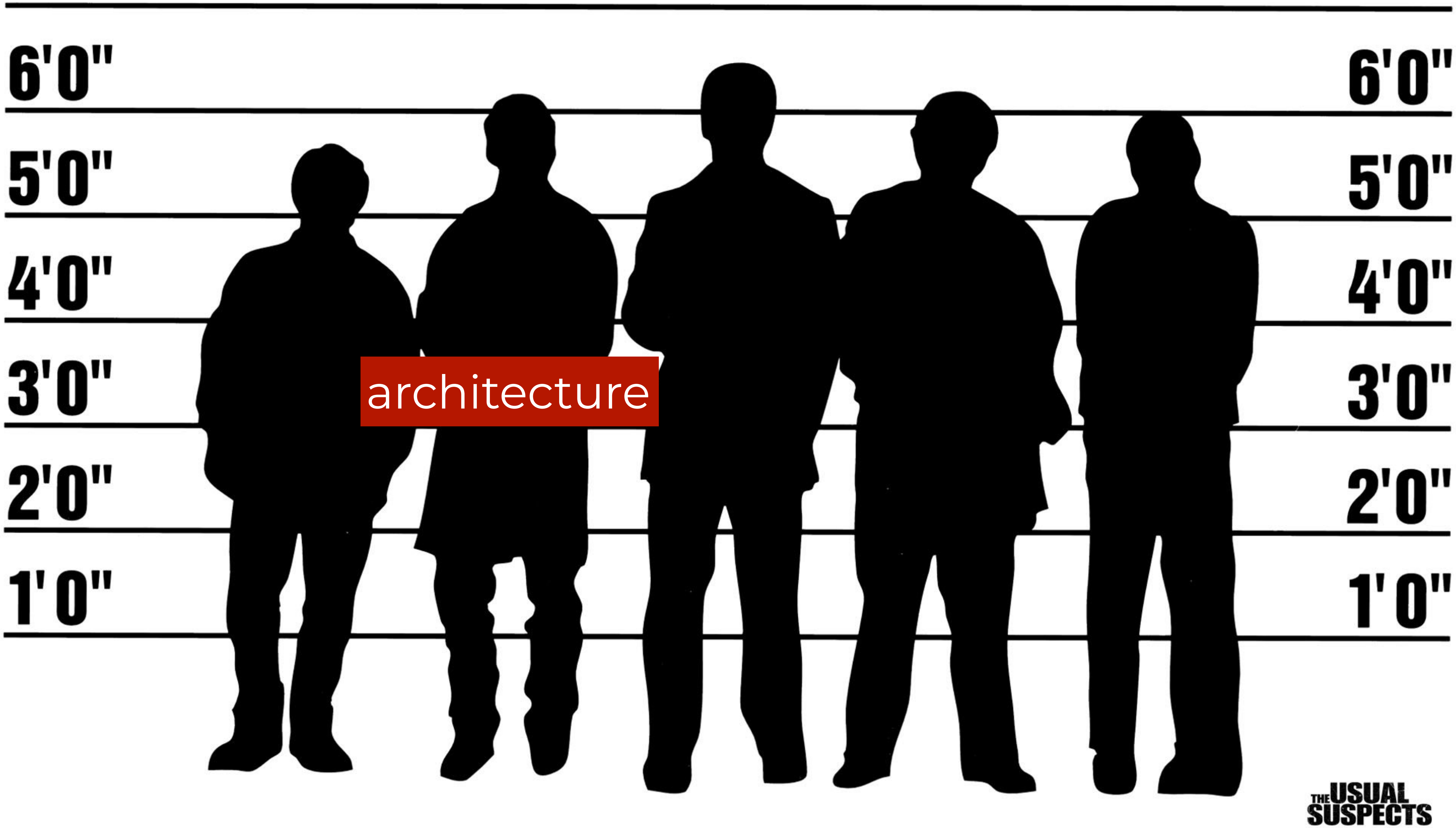
REALITY



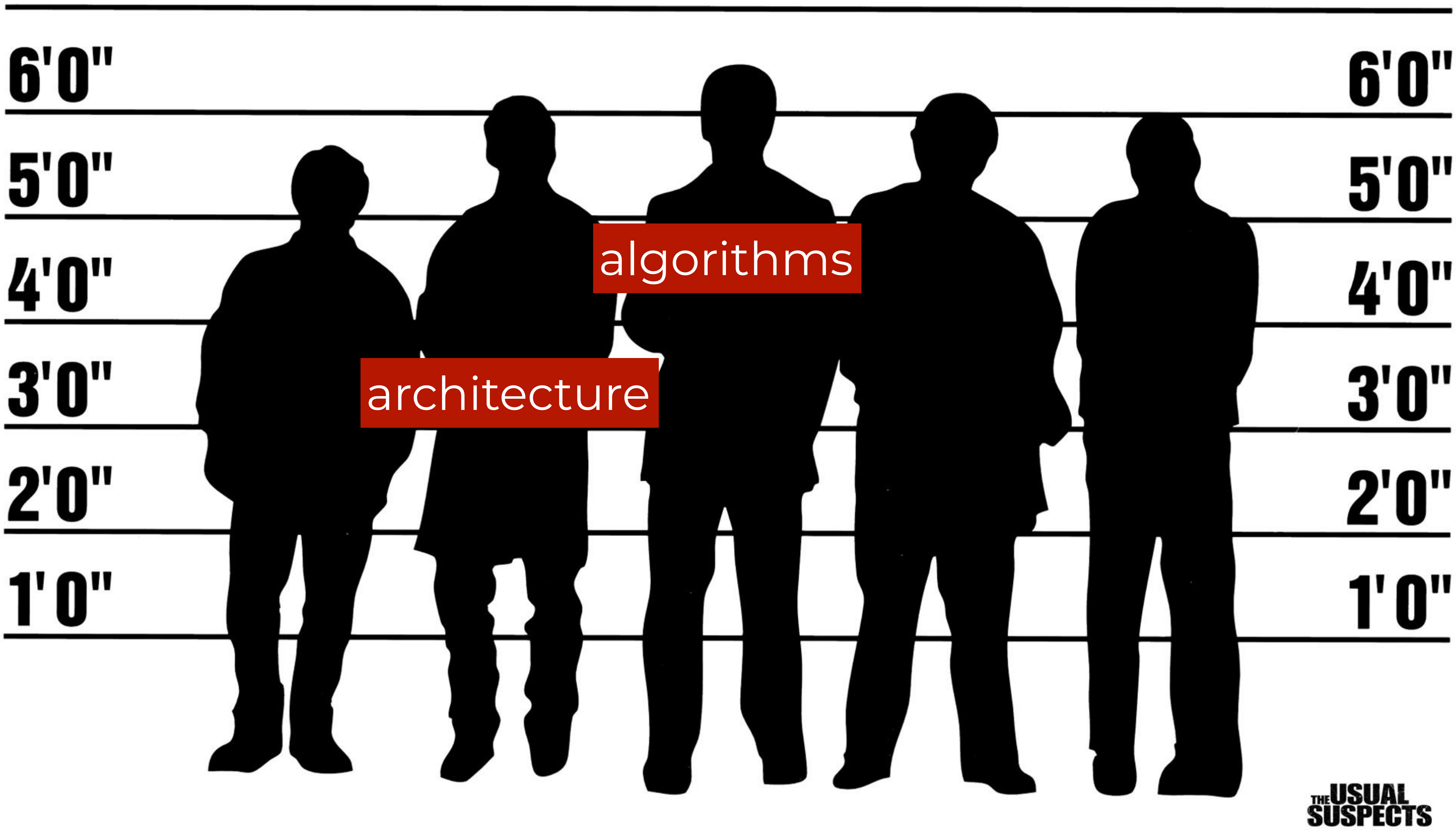
See also [Geman et al. '92; Opper '95; Neyshabur, Tomyoka, Srebro, 2015; Advani-Saxe 2017; Belkin, Hsu, Ma, Soumik, Mandal 2019; Nakkiran et al. 2019]

The usual suspects

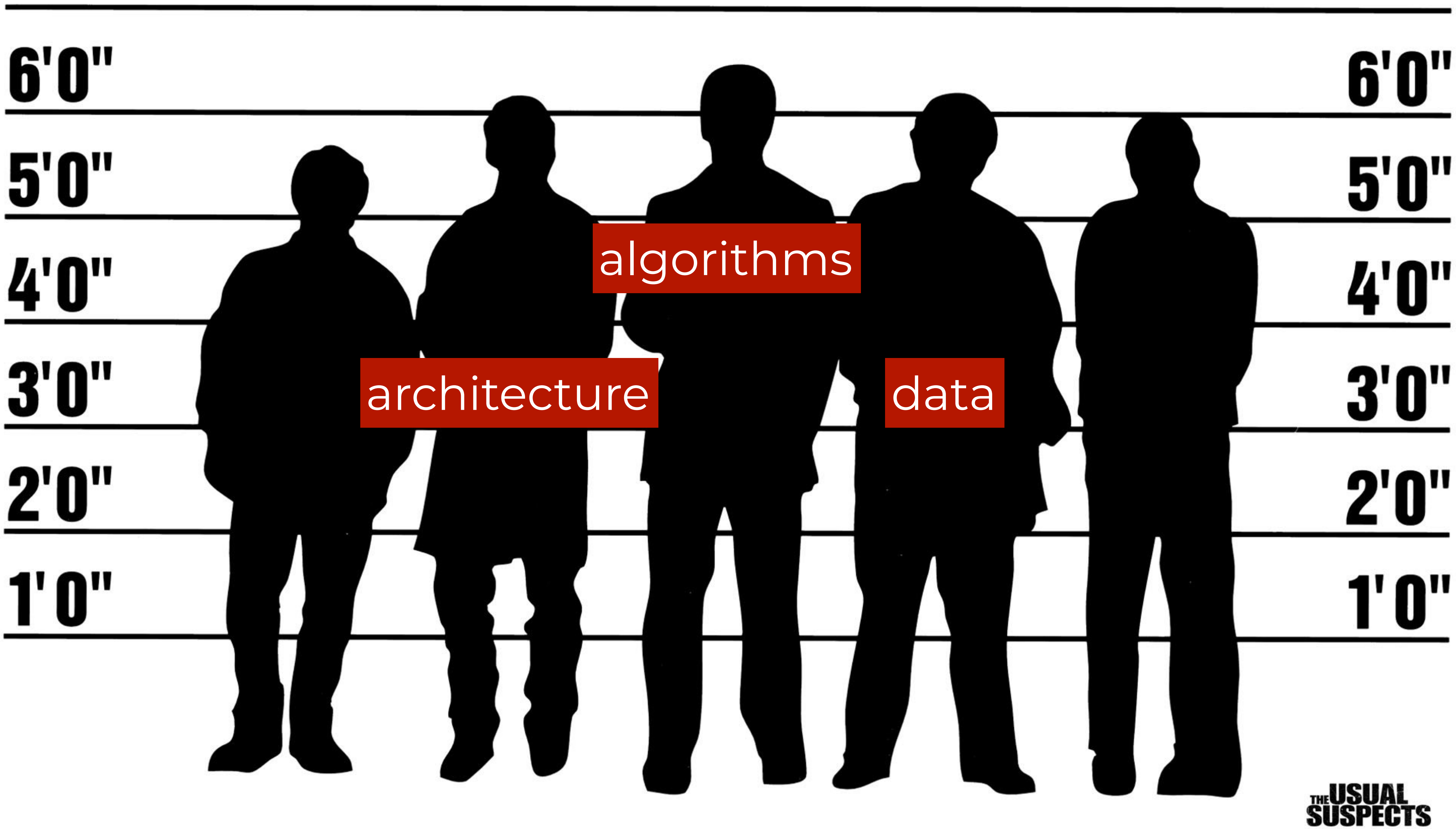
The usual suspects



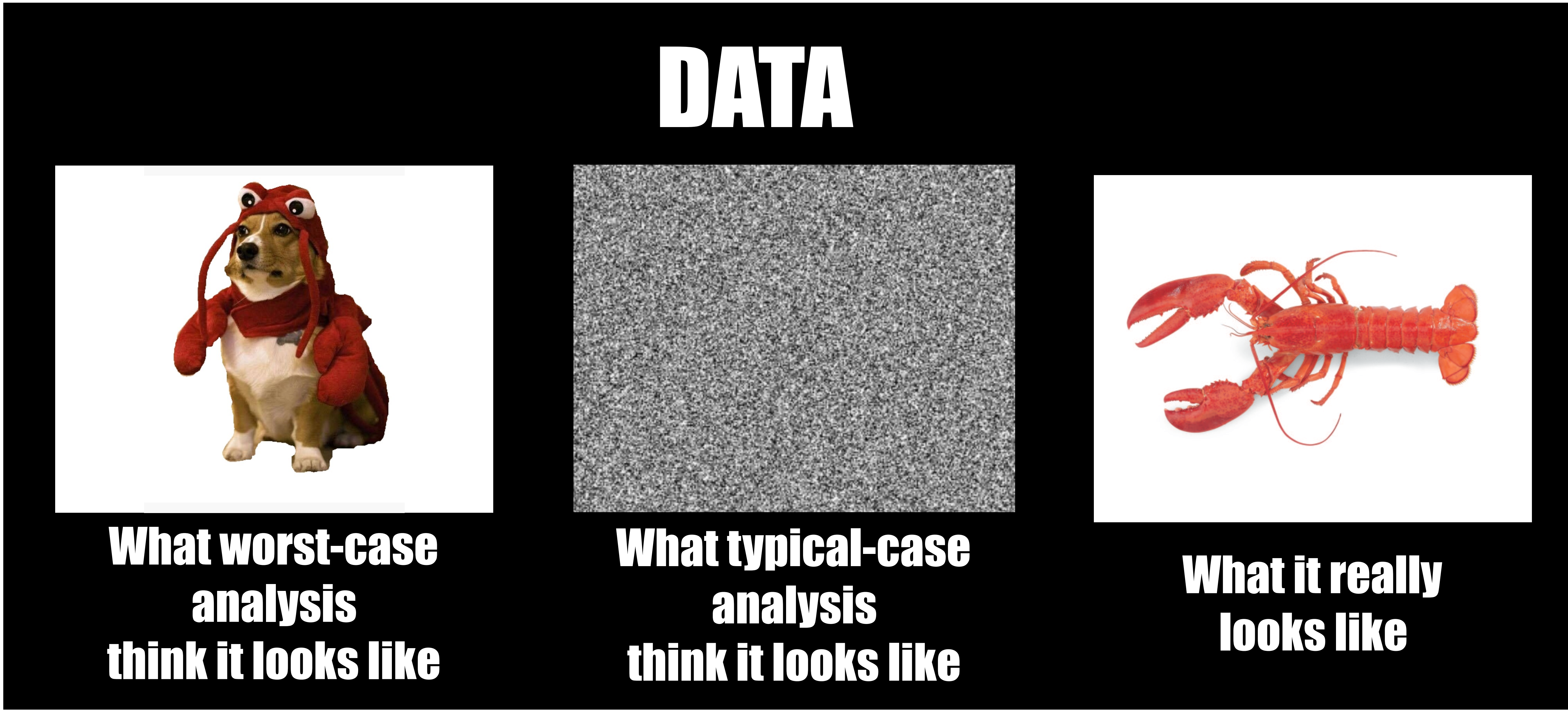
The usual suspects



The usual suspects

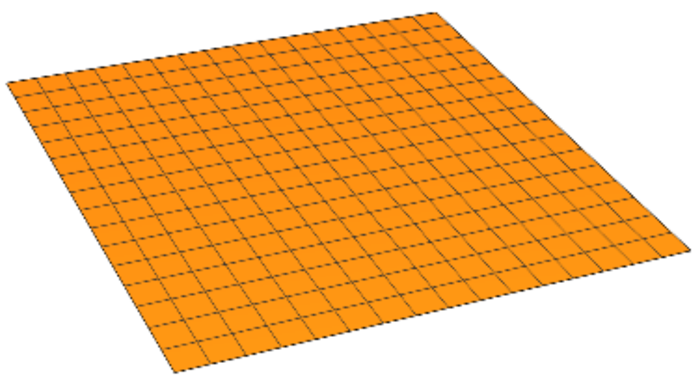


The two theory cultures



Spoiler

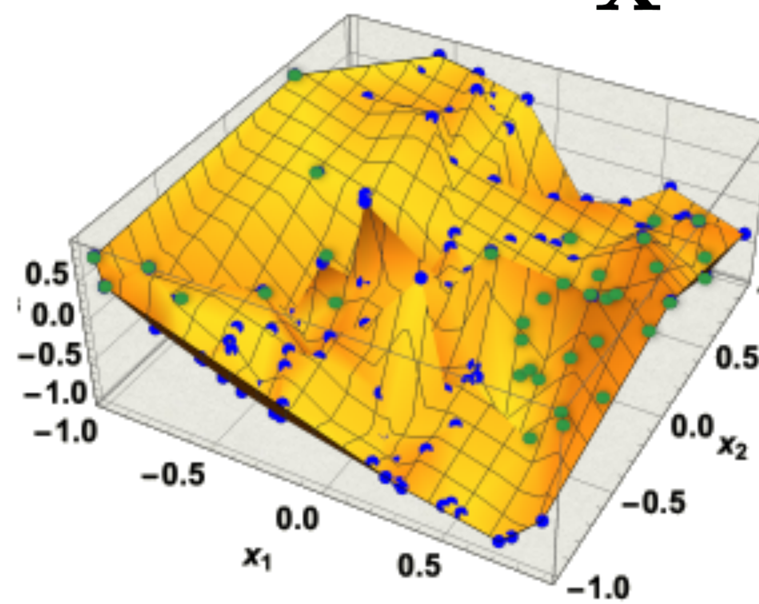
$\mathbf{c}^{\mu} \in \mathbb{R}^{d}$



Feature space

$$\sigma\left(\mathrm{F}^{\top}\cdot\right)$$

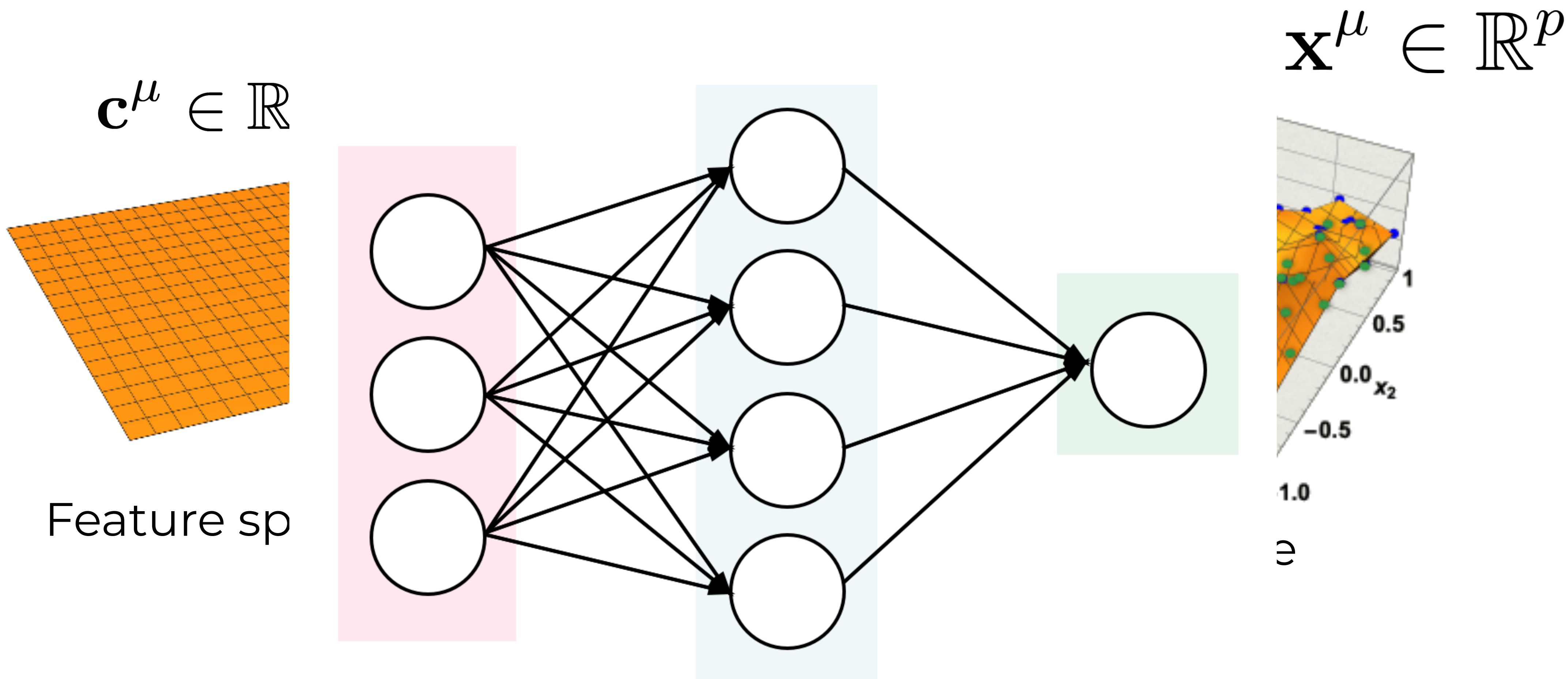
$\mathbf{x}^{\mu} \in \mathbb{R}^{p}$



Input space

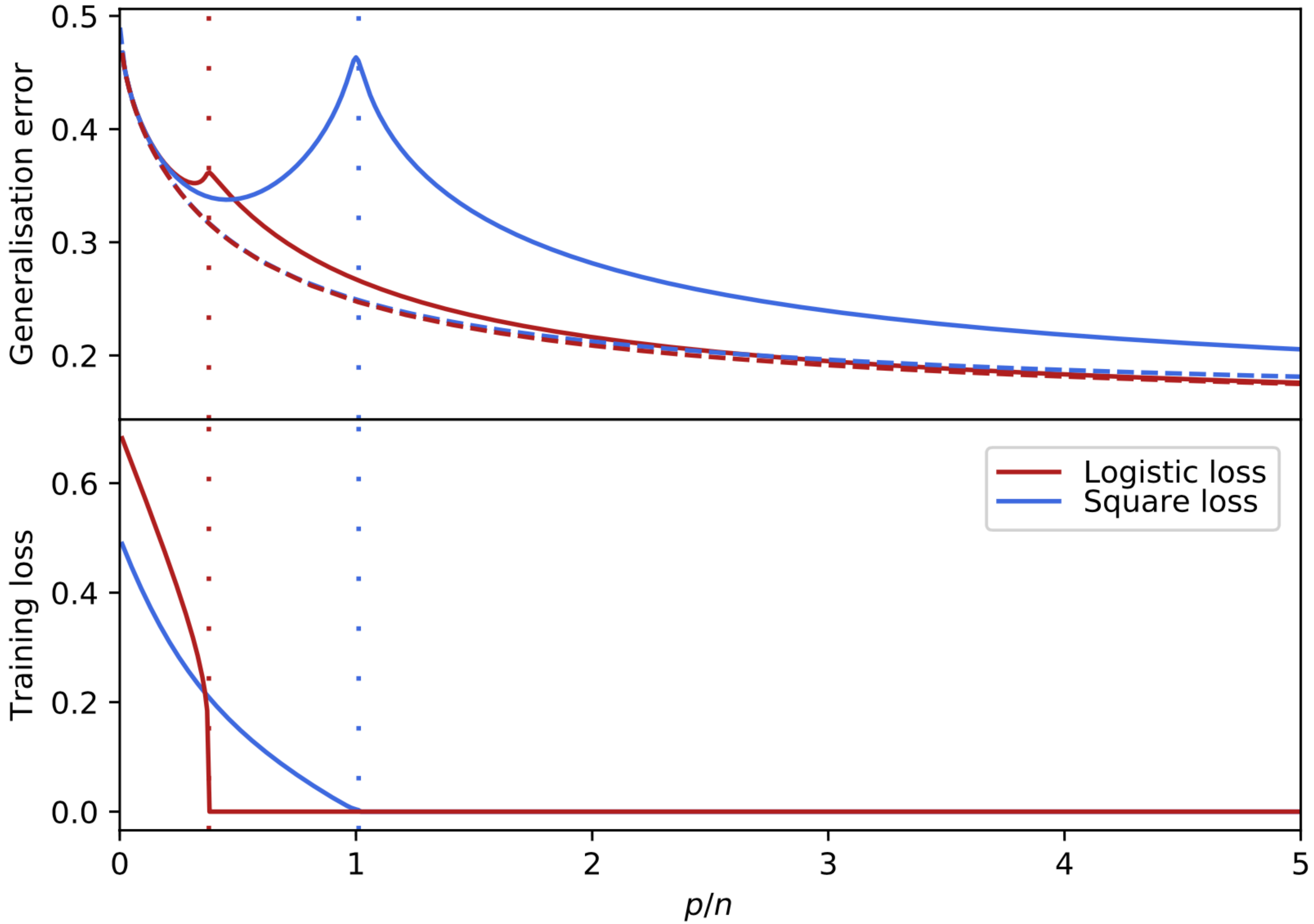
Spoiler

SPOILER ALERT



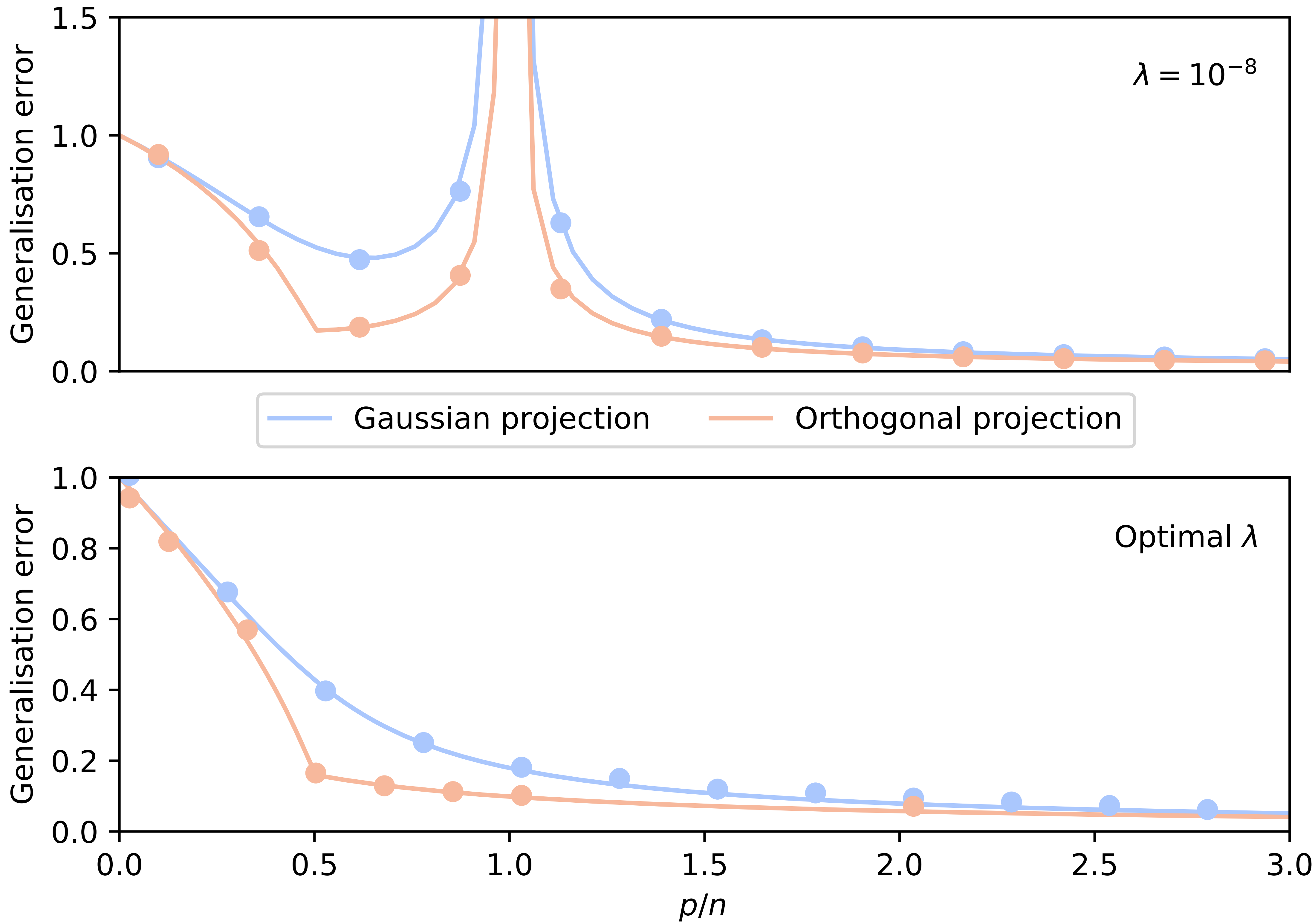
Spoiler


SPOILER
ALERT


Generalisation error
Training loss
Logistic loss
Square loss
p/n
0.5
0.4
0.3
0.2
0.6
0.4
0.2
0.0
0
1
2
3
4
5

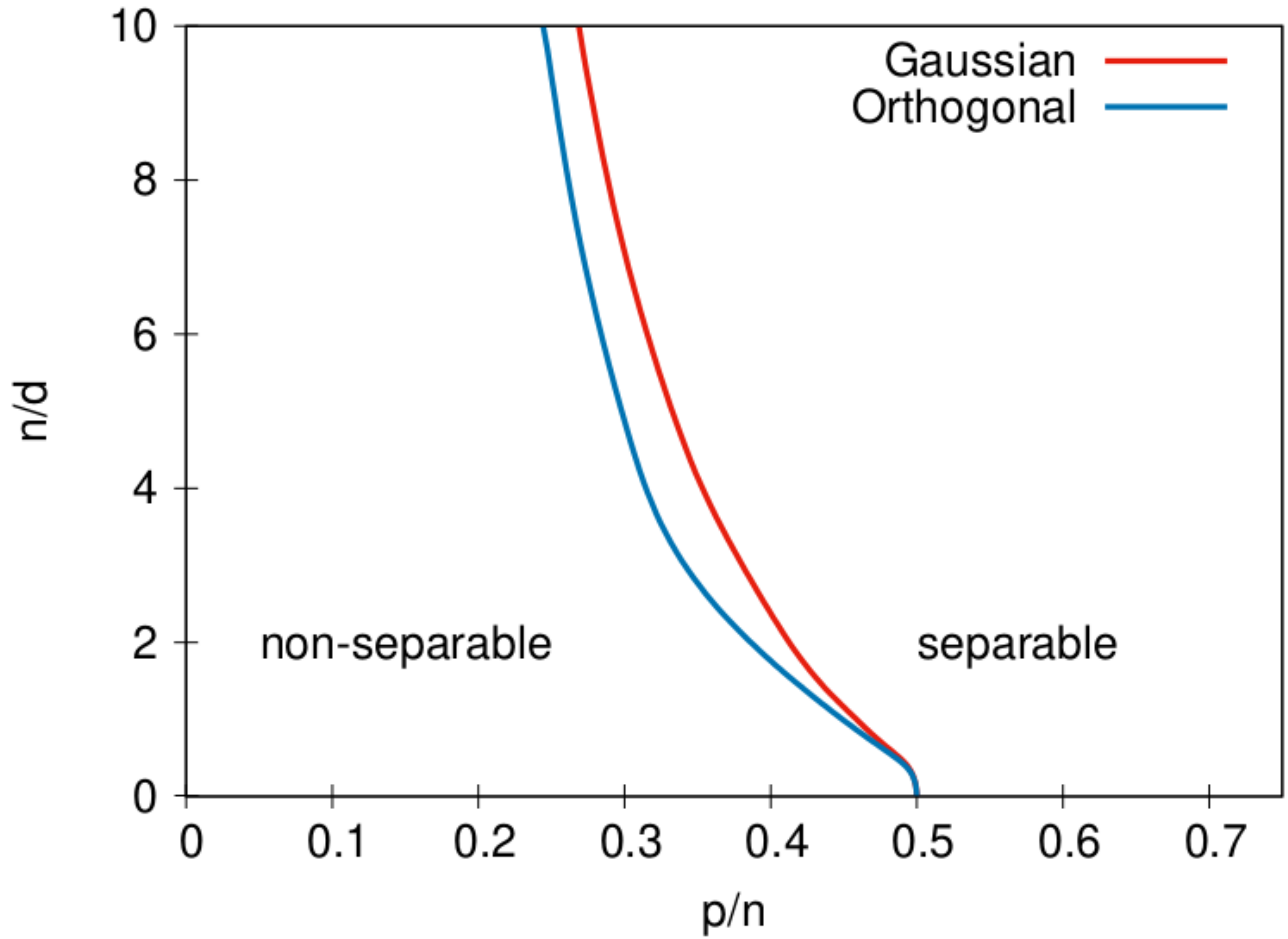
Spoiler

SPOILER ALERT



Spoiler

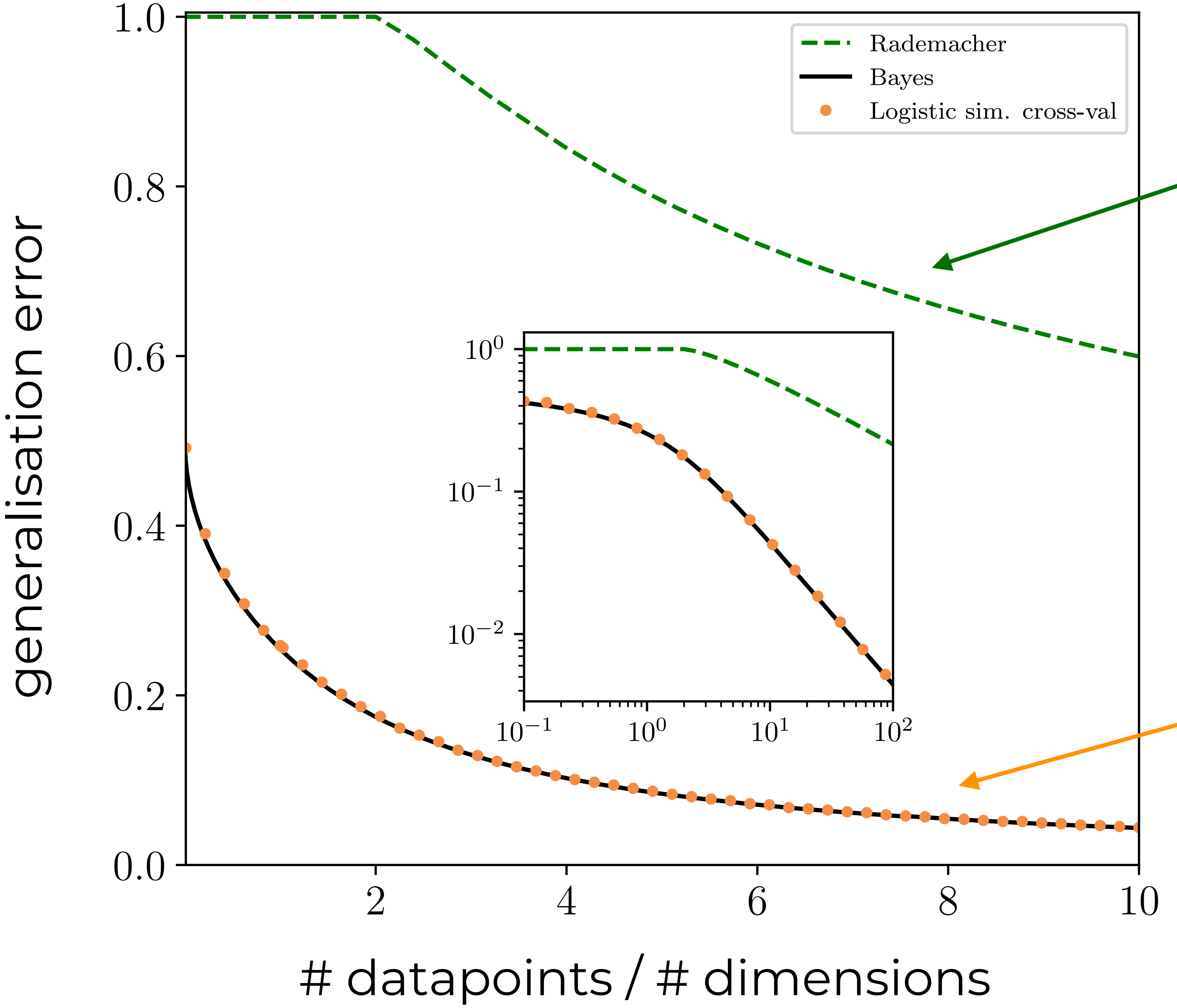

SPOILER
ALERT


Gaussian
Orthogonal
non-separable
separable
n/d
p/n
0
2
4
6
8
10
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

Worst-case vs. typical-case:
A concrete example

Concrete example

Dataset $D = \{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^n$ with $\boldsymbol{x}^\mu \sim \mathcal{N}(0, \mathrm{I_d})$ and labels $y^\mu = \mathrm{sign}\left(\boldsymbol{x}^\mu \cdot \boldsymbol{\theta}^0\right)$



Radamacher bound for function class

$$f_\theta(\boldsymbol{x}) = \mathrm{sign}\left(\boldsymbol{x} \cdot \boldsymbol{\theta}\right)$$

Out-of-the-box Logistic regression (Sklearn)

[Abbara, Aubin, Krzakala, Zdeborová '19]

Can we do better?

Hidden Manifold Model

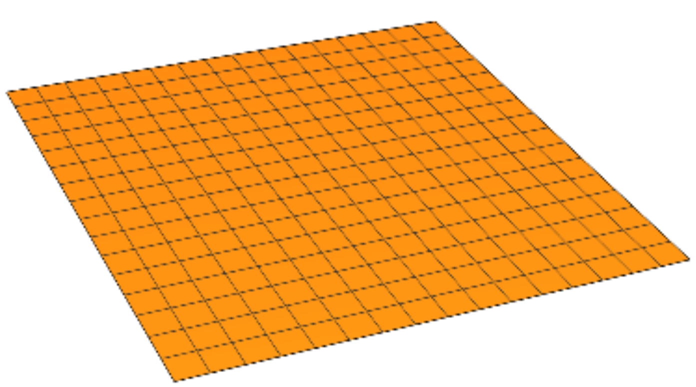
[Goldt, Mézard, Krzakala, Zdeborová '19]

Idea: dataset where both data points and labels only depend on a subset of latent variables.

$$D = \{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^{n}$$

$$y^\mu = f^0\left(\frac{\boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0}{\sqrt{d}}\right) \qquad \boldsymbol{x}^\mu = \sigma\left(\frac{\mathrm{F}^\top \boldsymbol{c}^\mu}{\sqrt{d}}\right)$$

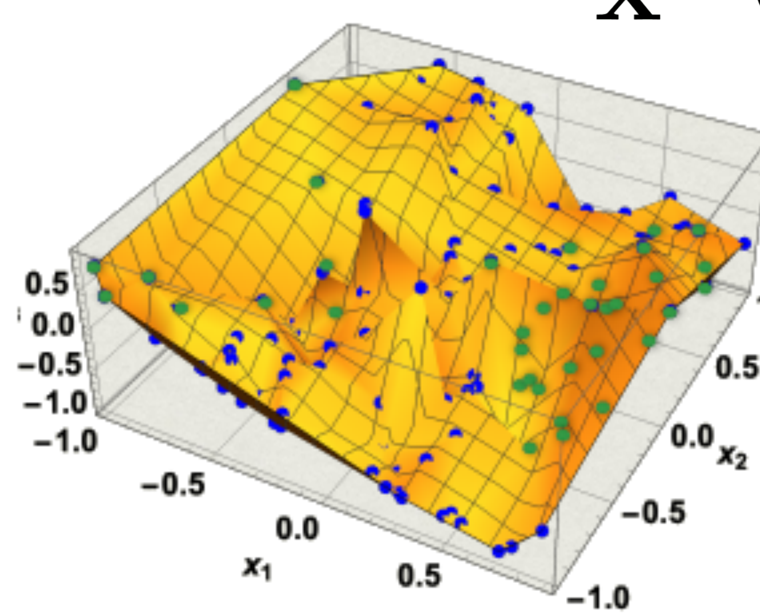
$\mathbf{c}^\mu \in \mathbb{R}^d$



Feature space

$\sigma\left(\mathrm{F}^\top \cdot\right)$

$\mathbf{x}^\mu \in \mathbb{R}^p$



Input space

Aim: study classification and regression tasks on this dataset

The task

Learn the labels using a linear model with empirical risk minimisation

$$\hat{y}^{\mu} = \hat{f}\left(\mathbf{x}^{\mu} \cdot \hat{\mathbf{w}}\right)$$

where:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \left[\frac{1}{n} \sum_{\mu=1}^{n} \ell\left(y^{\mu}, \mathbf{x}^{\mu} \cdot \mathbf{w}\right) + \frac{\lambda}{2} \|\mathbf{w}\|_2^2 \right]$$

loss function ridge penalty

examples:

- Ridge regression: $f^0(x) = \hat{f}(x) = x$ $\qquad \ell(x, y) = \frac{1}{2}(y - x)^2$
- Logistic regression: $f^0(x) = \hat{f}(x) = \operatorname{sign}(x)$ $\qquad \ell(x, y) = \log\left(1 + e^{-xy}\right)$

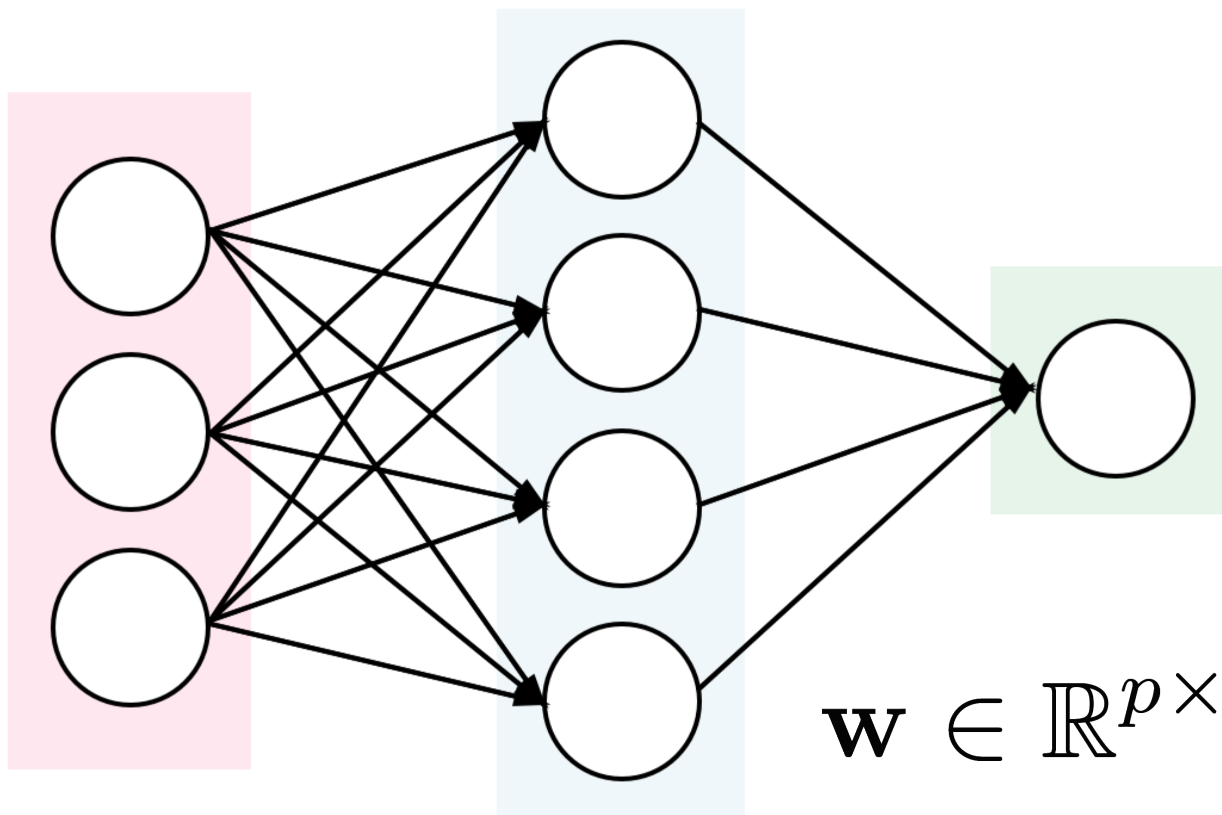
Two alternative points of view

[Williams ‘98,‘07; Retch, Raimi ‘07; Montanari, Mei 19’]

Dataset $D = \{\boldsymbol{c}^{\mu}, y^{\mu}\}_{\mu=1}^{n}$

$$\mathbf{X} = \mathbf{\Phi}_{\mathbf{F}}(\mathbf{C}) = \sigma\left(\frac{\mathbf{CF}}{\sqrt{d}}\right) \in \mathbb{R}^{n\times p}$$

$\mathbf{C} \in \mathbb{R}^{n\times d}$



$\mathbf{Y} = \hat{f}\left(\mathbf{w}\cdot\mathbf{X}\right) \in \mathbb{R}^{n\times 1}$

$\mathbf{w} \in \mathbb{R}^{p\times 1}$

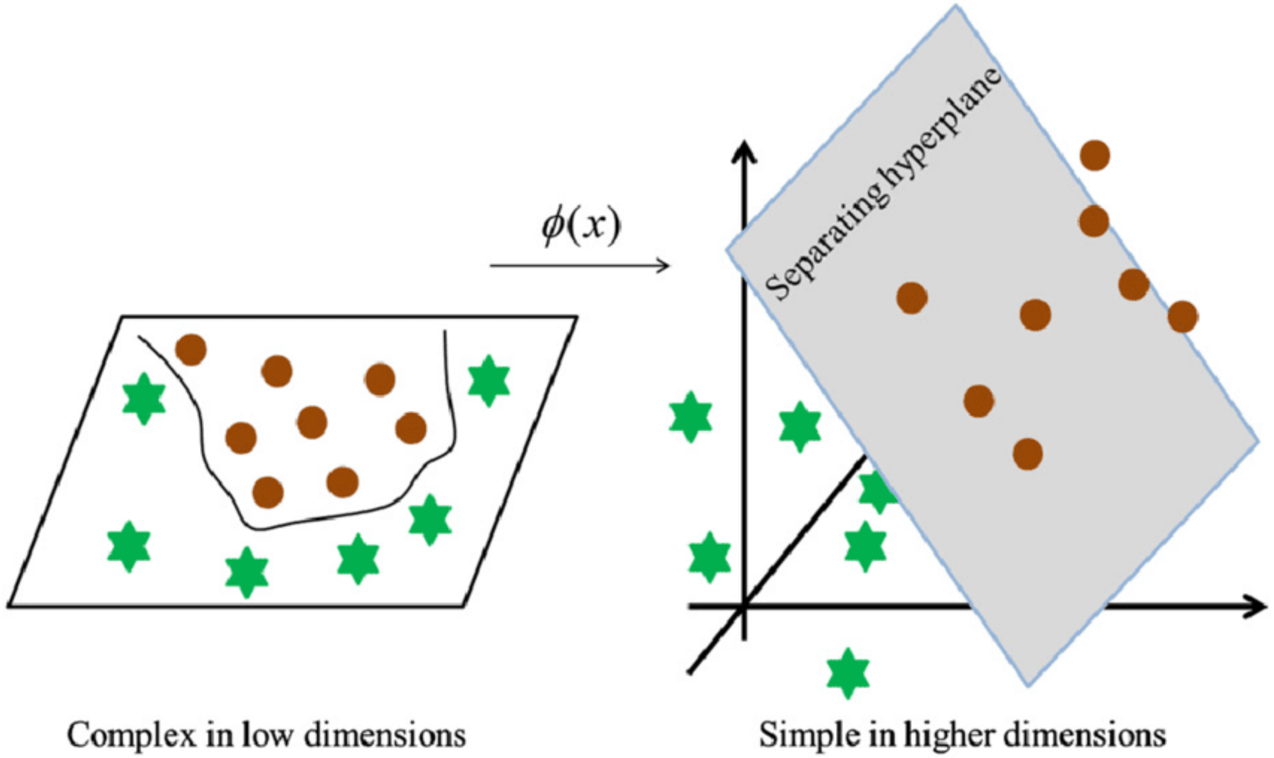
$\mathbf{F} \in \mathbb{R}^{d\times p}$

Two alternative points of view

[Williams '98,'07; Retch, Raimi '07; Montanari, Mei 19']

Dataset $D = \{\boldsymbol{c}^{\mu}, y^{\mu}\}_{\mu=1}^{n}$

Feature map $\Phi_{\mathrm{F}}(\boldsymbol{c}) = \sigma(\mathrm{F}^{\top}\boldsymbol{c})$



$$\Phi_{\mathrm{F}}(\boldsymbol{c})\Phi_{\mathrm{F}}(\boldsymbol{c}') \underset{p\to\infty}{\longrightarrow} K(\boldsymbol{c}, \boldsymbol{c}')$$

Mercer's theorem

Main result:
Asymptotic generalisation error
for arbitrary
loss ℓ and projection F

Definitions:

Consider the unique fixed point of the following system of equations

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

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

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

where $V=\kappa_1^2V_s+\kappa_\star^2V_w$, $V^0=\rho-\frac{M^2}{Q}$, $Q=\kappa_1^2q_s+\kappa_\star^2q_w$, $M=\kappa_1m_s$, $\omega_0=M/\sqrt{Q}\xi$, $\omega_1=\sqrt{Q}\xi$ and g_μ is the Stieltjes transform of FF^T

$\kappa_0=\mathbb{E}\left[\sigma(z)\right], \kappa_1\equiv\mathbb{E}\left[z\sigma(z)\right], \kappa_\star\equiv\mathbb{E}\left[\sigma(z)^2\right]-\kappa_0^2-\kappa_1^2$ and $\vec{z}^\mu\sim\mathcal{N}(\vec{0},I_p)$

In the high-dimensional limit:

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

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

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

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

Agrees with [Mei-Montanari '19] who solved a particular case using random matrix theory: linear function f^0, $\ell(x,y)=||x-y||_2^2$ & Gaussian random weights **F**

Technical note: replicated Gaussian Equivalence

An important step in the derivation of this result is the observation that the generalisation and training properties of the dataset $\{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^{n}$ are *statistically equivalent* to the following dataset $\{\tilde{\boldsymbol{x}}^\mu, y^\mu\}_{\mu=1}^{\mu}$ with the same labels but:

$$\tilde{\boldsymbol{x}}^\mu = \kappa_1 \frac{1}{\sqrt{d}} \mathrm{F}^\top \boldsymbol{c}^\mu + \kappa_\star \boldsymbol{z}^\mu \qquad \boldsymbol{z}^\mu \sim \mathcal{N}(0, \mathrm{I_p})$$

where the coefficients $\kappa_1, \kappa_\star$ are chosen to match

$$\kappa_1 = \mathbb{E}_\xi \left[\xi \; \sigma(\xi)\right] \qquad \kappa_\star^2 = \mathbb{E}_\xi \left[\sigma(\xi)^2\right] - \kappa_1^2 \qquad \xi \sim \mathcal{N}(0,1)$$

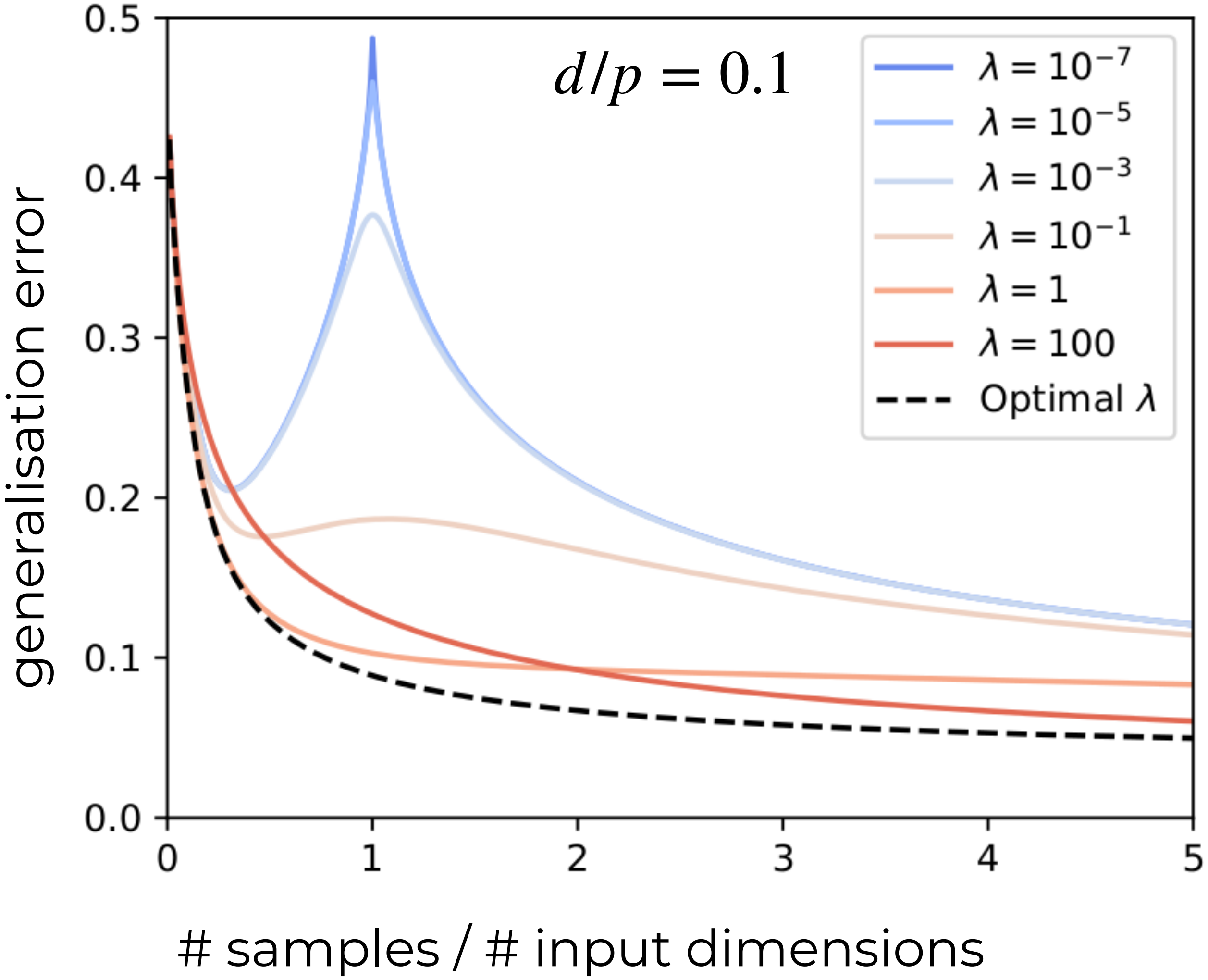
Generalisation of an observation in

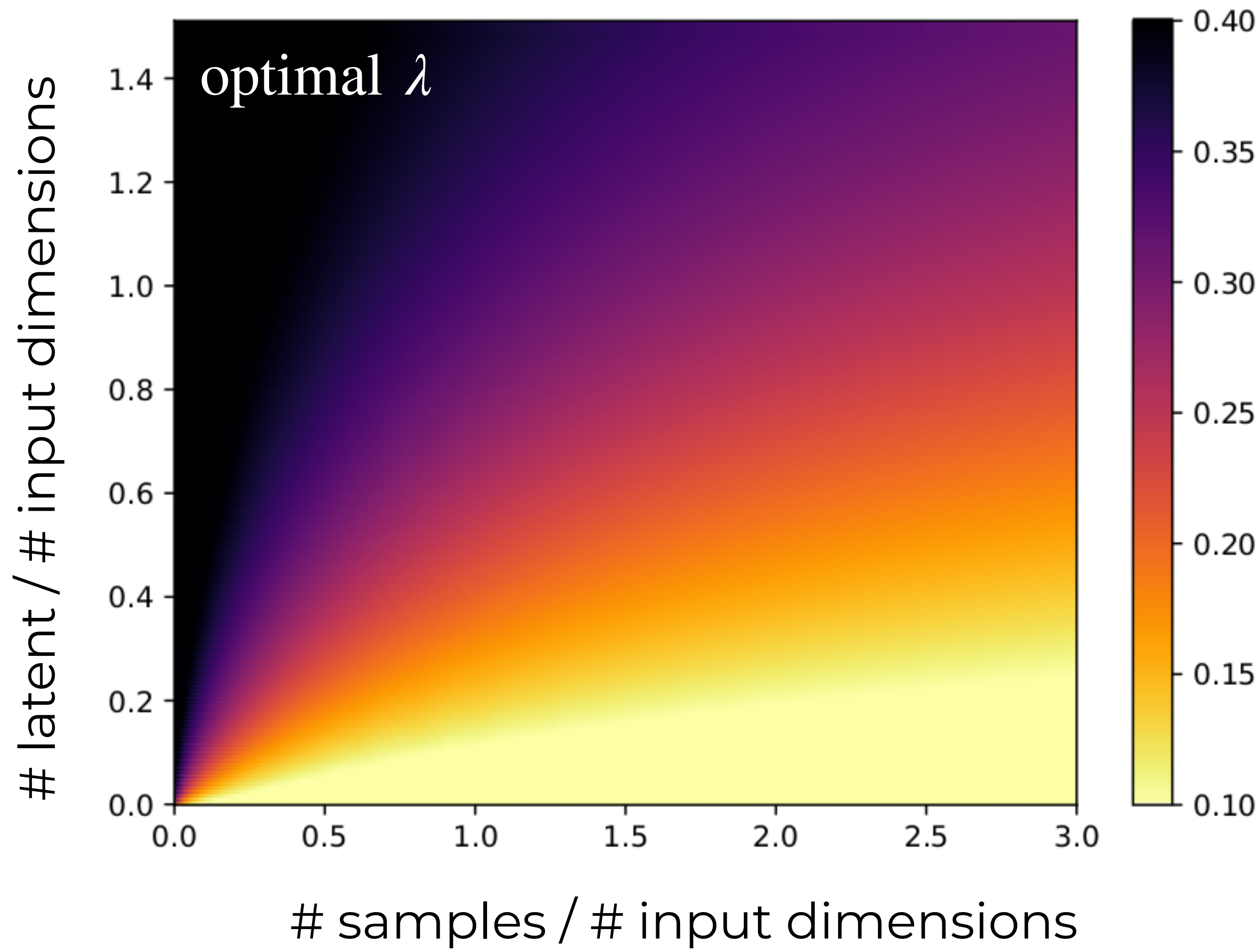
[Mei, Montanari 19'; Goldt, Mézard, Krzakala, Zdeborová '19]

Drawing the consequences of our formula

Learning in the HMM

$$f^0 = \hat{f} = \text{sign} \qquad \sigma = \text{erf} \qquad l(x, y) = \frac{1}{2}(x - y)^2 \qquad \text{Gaussian } \mathrm{F}$$

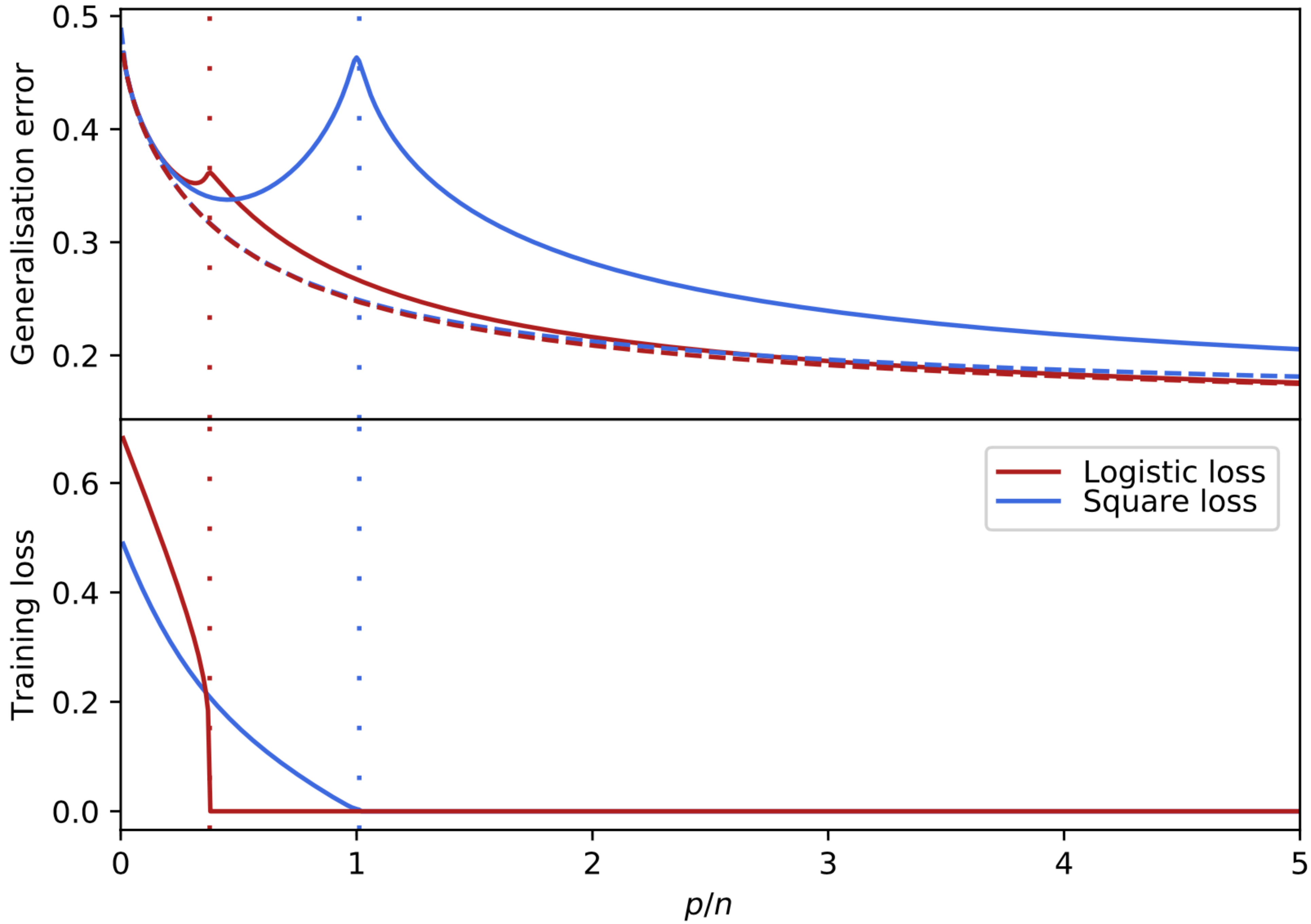




Good generalisation performance for small latent space,
even for small sample complexities

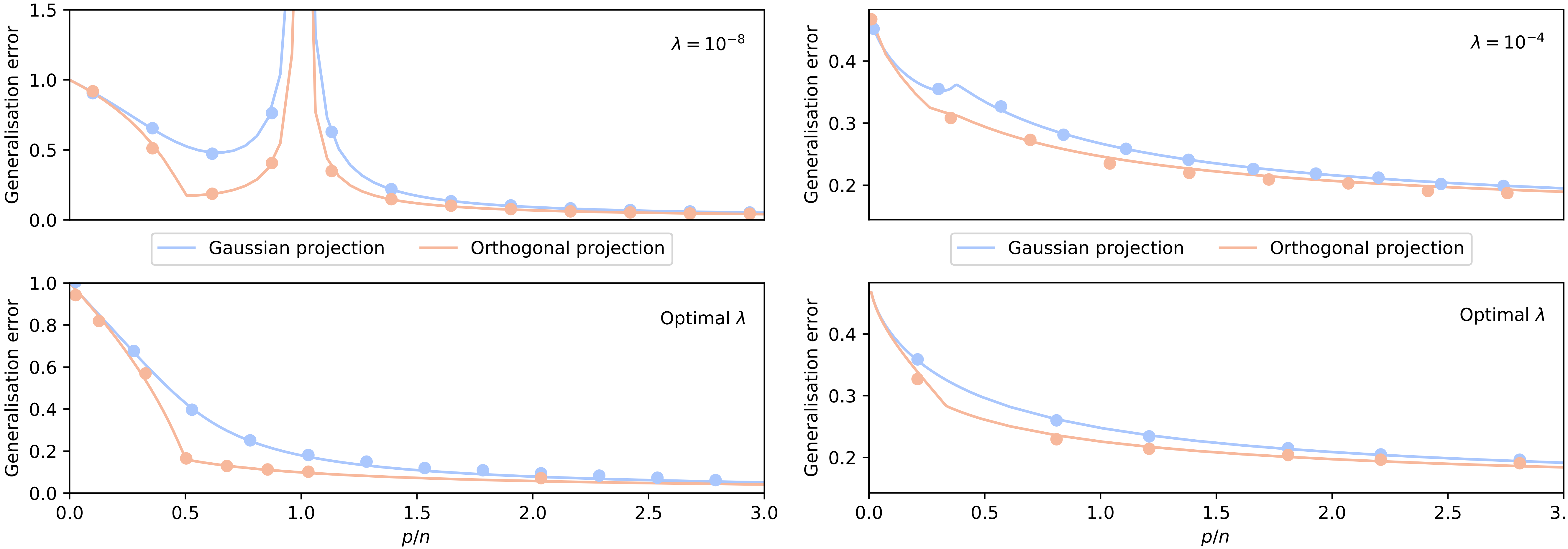
Classification tasks

$f^0 = \hat{f} = \text{sign}$ $\sigma = \text{sign}$ Gaussian F



Random vs. orthogonal projections

Ridge regression | Logistic regression



First layer: random i.i.d. Gaussian Matrix $F^i_\rho \sim \mathcal{N}(0, 1/d)$

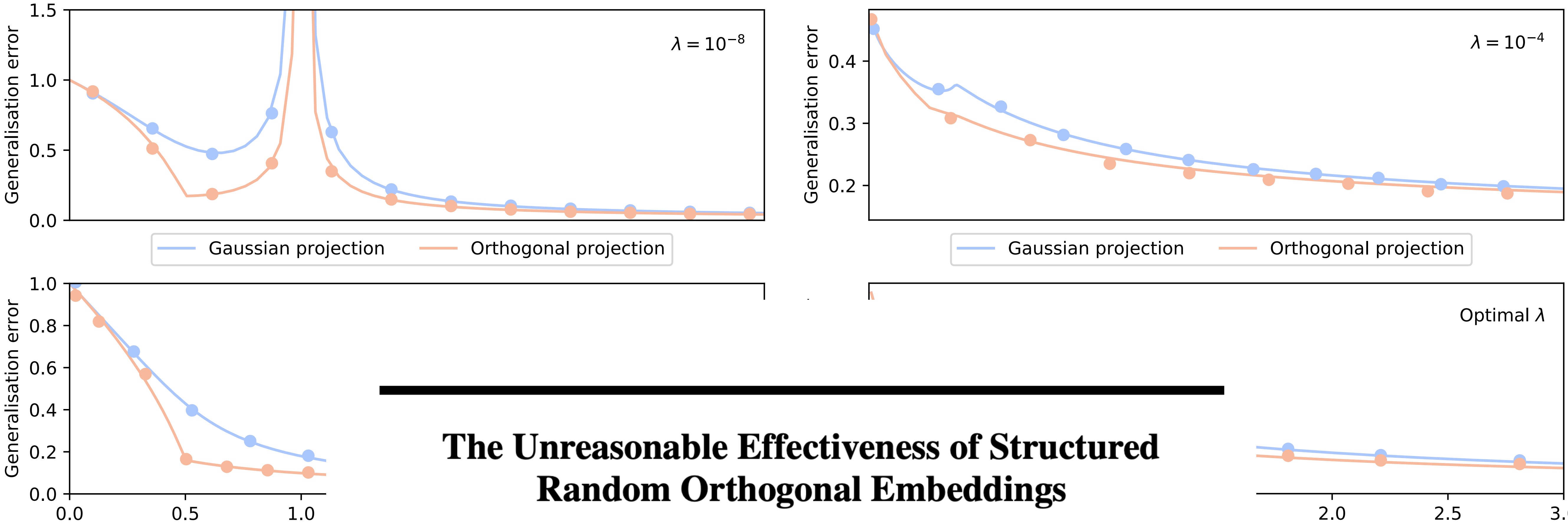
First layer: subsampled Fourier matrix $\mathrm{F} = \mathrm{U}^\top \mathrm{D} \mathrm{V}$

$\mathrm{U}, \mathrm{V} \sim \mathrm{Haar}$

Random vs. orthogonal projections

Ridge regression

Logistic regression



The Unreasonable Effectiveness of Structured Random Orthogonal Embeddings

Krzysztof Choromanski *
Google Brain Robotics
kchoro@google.com

Mark Rowland *
University of Cambridge
mr504@cam.ac.uk

Adrian Weller
University of Cambridge and Alan Turing Institute
aw665@cam.ac.uk

First

First

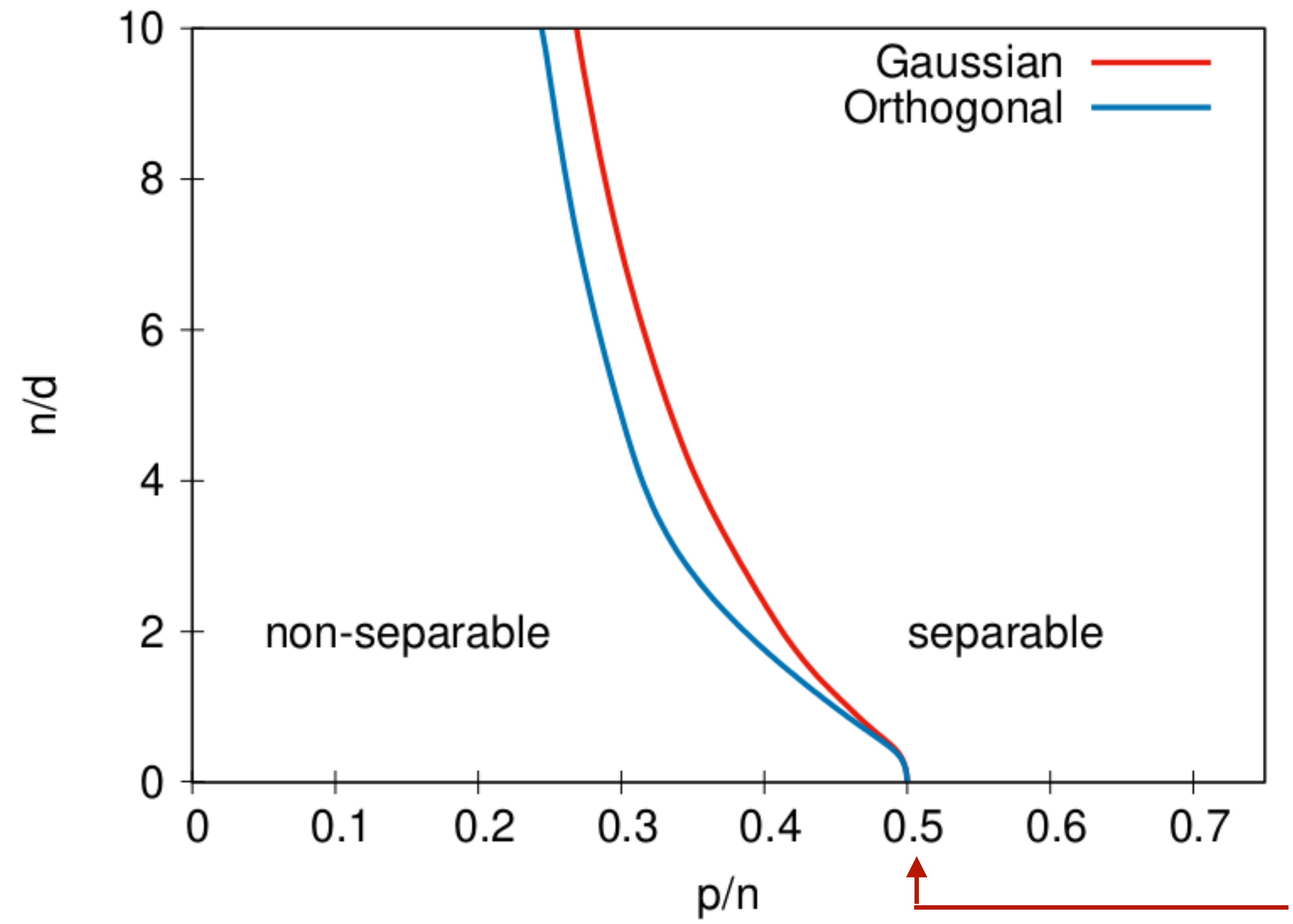
$1/d)$

[NIPS, '17]

r

Separability transition in logistic regression

$$f^0 = \hat{f} = \text{sign} \qquad \sigma = \text{erf} \qquad l(x, y) = \log\left(1 + e^{-xy}\right)$$

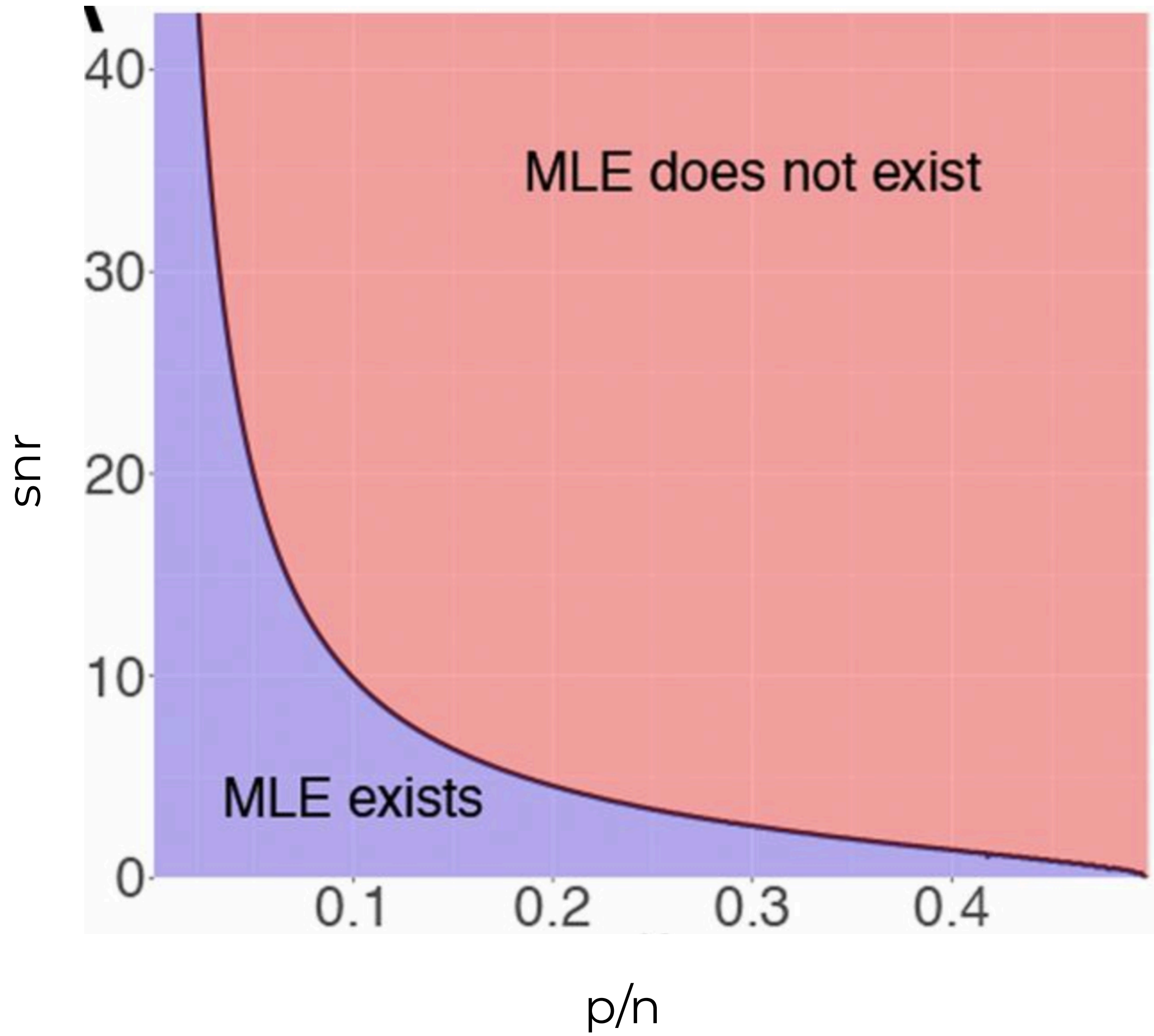


Cover theory '65

Separability transition in logistic regression

A modern maximum-likelihood theory for high-dimensional logistic regression

Pragya Sur and Emmanuel J. Candès

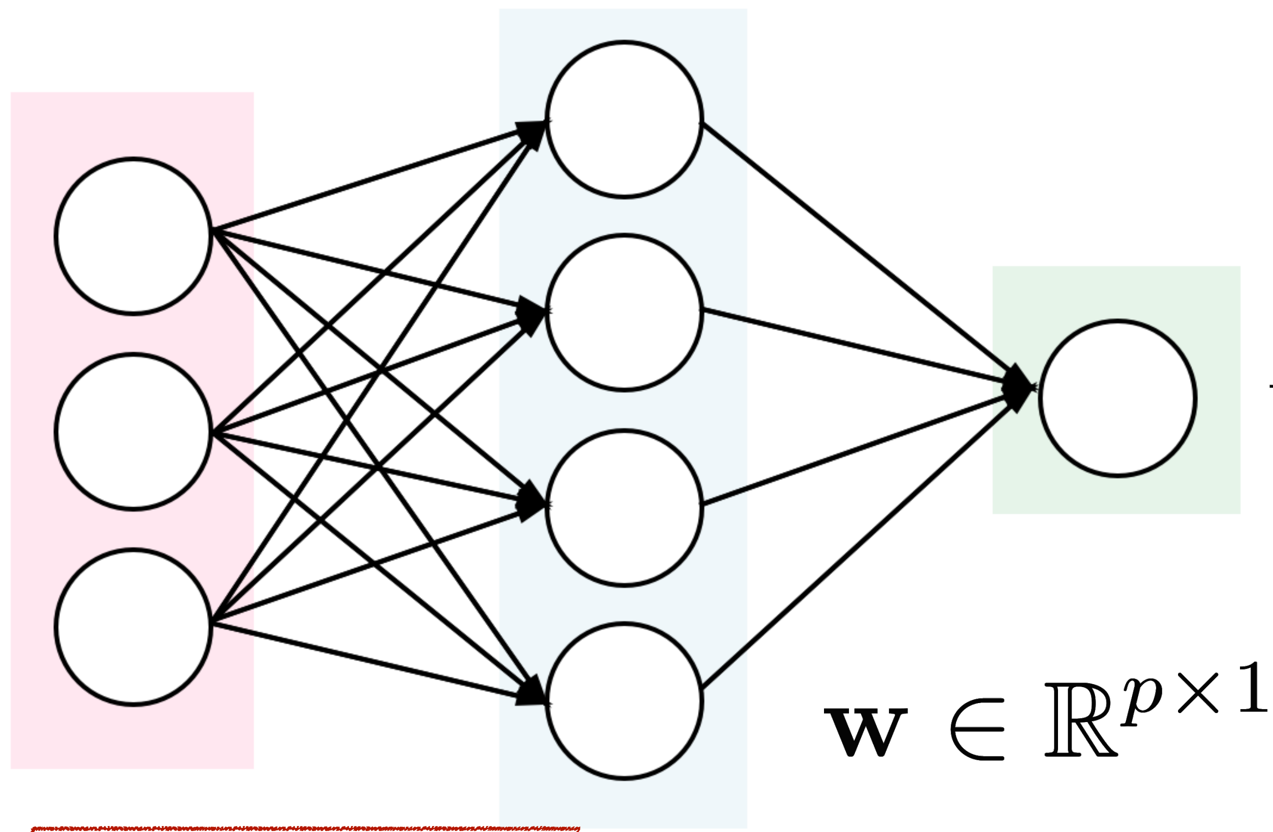


[Sur & Candes, ’18]

Next steps

$$\mathbf{X} = \mathbf{\Phi_F}\left(\mathbf{C}\right) = \sigma\left(\frac{\mathbf{CF}}{\sqrt{d}}\right) \in \mathbb{R}^{n\times p}$$

$$\mathbf{C} \in \mathbb{R}^{n\times d}$$



$$\mathbf{Y} = \hat{f}\left(\mathbf{w}\cdot\mathbf{X}\right) \in \mathbb{R}^{n\times 1}$$

$$\mathbf{w} \in \mathbb{R}^{p\times 1}$$

$$\mathbf{F} \in \mathbb{R}^{d\times p}$$

Learning F?

Thank you for your attention!

Check our paper @

arXiv: 2002.09339 [mat.ST]

contact: brloureiro@gmail.com

References in this talk

F. Gerace, B. Loureiro, F. Krzakala, M. Mézard, L. Zdeborobá, "*Generalisation error in learning with random features and the hidden manifold model*", arXiv: 2002.09339

M. Geiger, S. Spigler, S. d'Ascoli, L. Sagun, M. Baity-Jesi, G. Biroli and M. Wyart, "*Jamming transition as a paradigm to understand the loss landscape of deep neural networks*", Physical Review E, 100(1):012115

S. Goldt, M. Mézard, F. Krzakala, L. Zdeborobá, "*Modelling the influence of data structure on learning in neural networks: the hidden manifold model*", arXiv: 1909.11500

A. Abbara, B. Aubin, F. Krzakala, L. Zdeborobá, "*Rademacher complexity and spin glasses: A link between the replica and statistical theories of learning*", arXiv: 1912.02729

C. Williams , "*Computing with infinite networks*", NIPS 98'

A. Rahimi, B. Recht, "*Random Features for Large-Scale Kernel Machines*", NIPS 07'

S. Mei, A. Montanari, "*The generalization error of random features regression: Precise asymptotics and double descent curve*", arXiv: 1908.05355

K. Choromanski, M. Rowland, A. Weller, "*The Unreasonable Effectiveness of Structured Random Orthogonal Embeddings*", NIPS 07'

P. Sur and E.J. Candès, "*A modern maximum-likelihood theory for high-dimensional logistic regression*", PNAS 19'