Fluctuations, bias, variance & ensemble of learners:

Exact Asymptotics for Convex Losses in High-Dimension

Bruno Loureiro
(EPFL)

Cédric Gerbelot
(ENS)

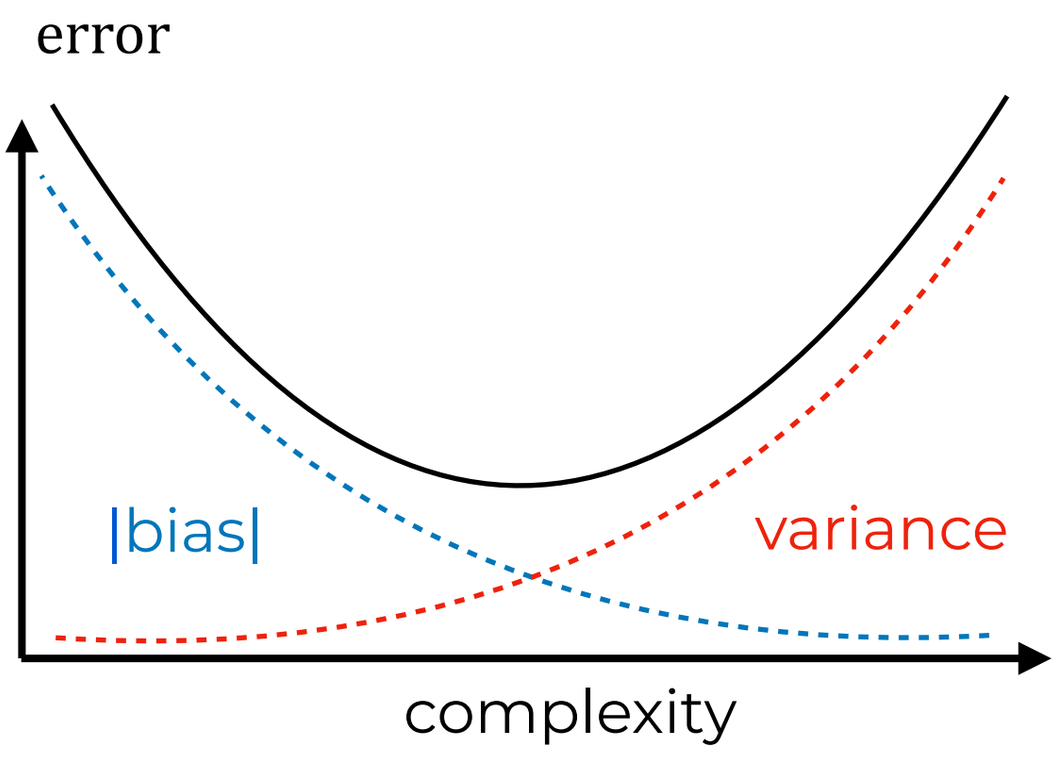
Maria Refinetti
(ENS)

Gabriele Sicuro
(KCL)

Florent Krzakala
(EPFL)

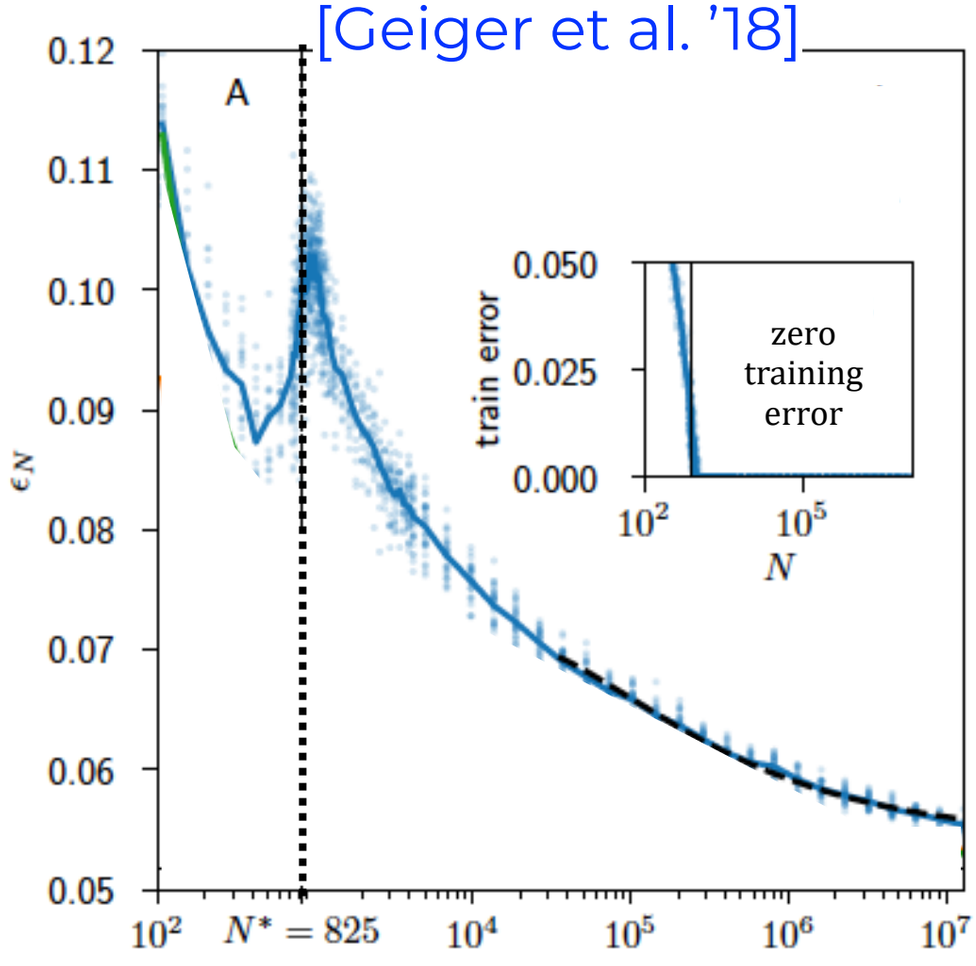
TRAINING A NN

EXPECTATIONS



REALITY

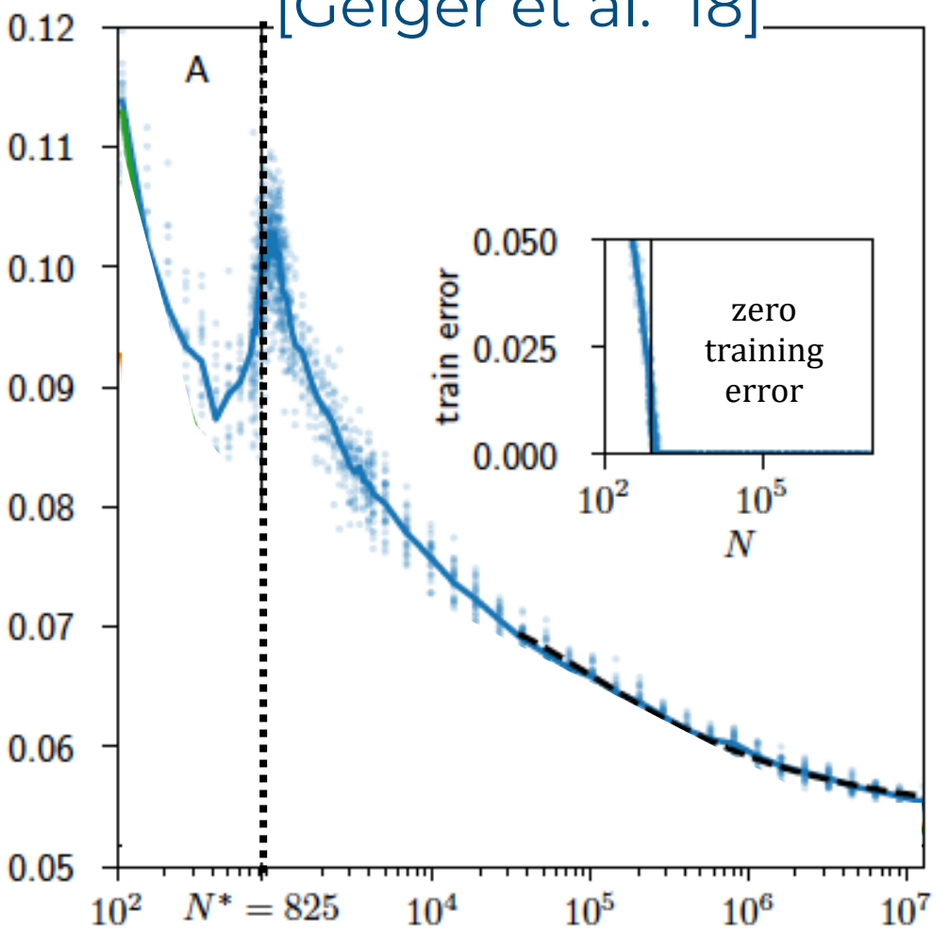
[Geiger et al. '18]



Number of parameters

The “double descent” [Belkin et al. ’18]

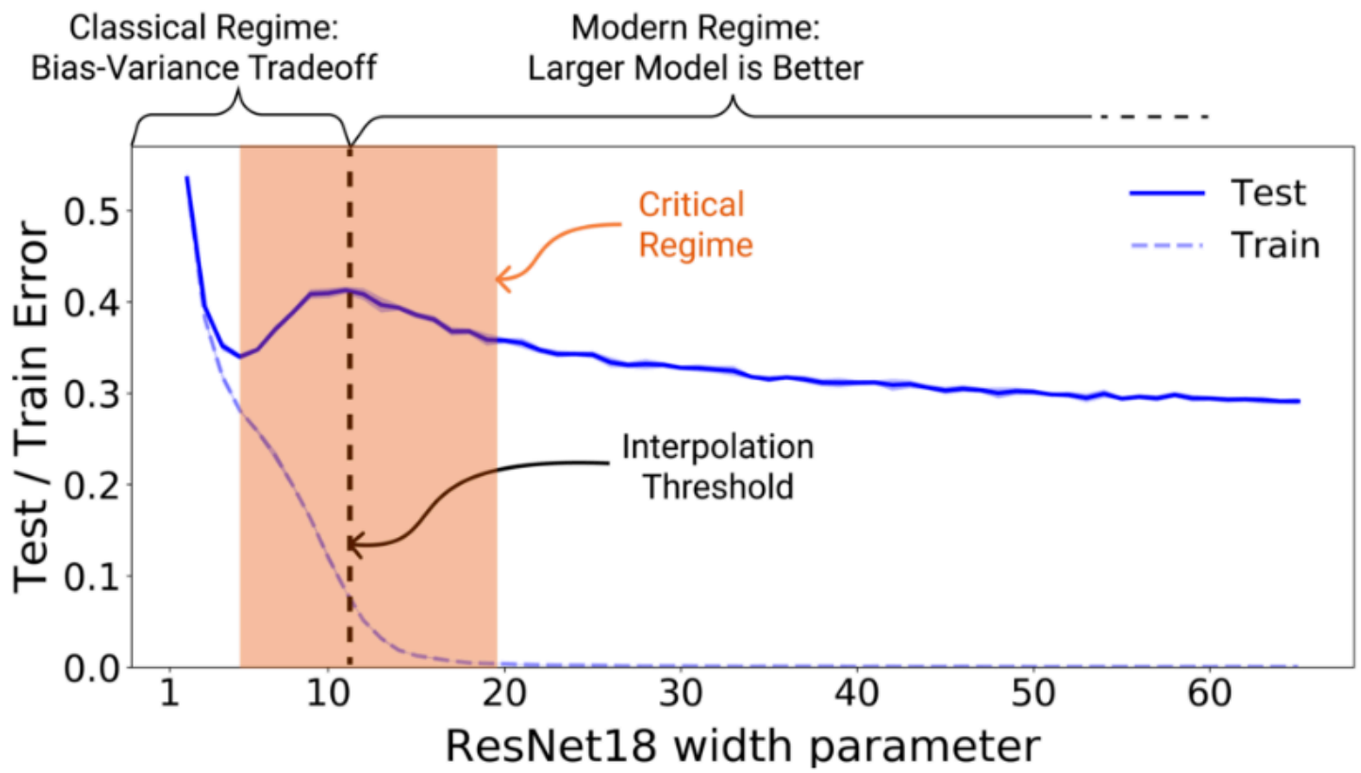
[Geiger et al. ’18]



Number of parameters

Parity-MNIST, 5 layers, fully-connected, no regularisation

[Nakkiran et al. ’19]



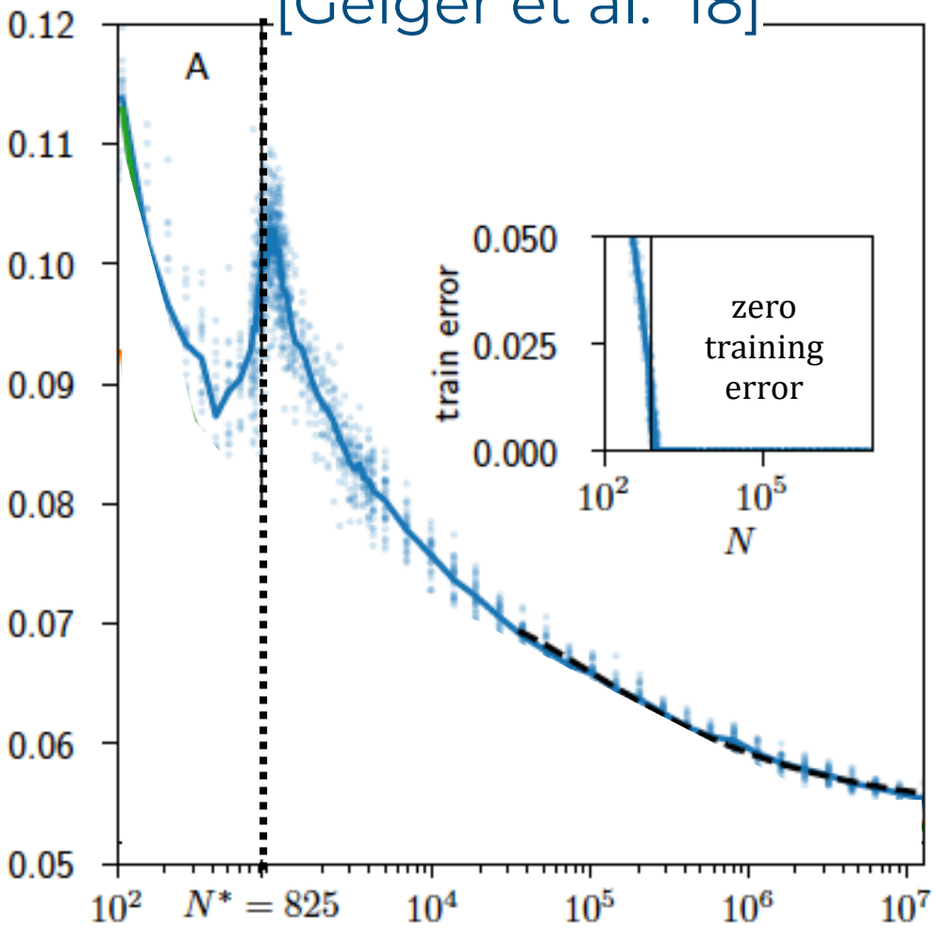
CIFAR10, no regularisation

See also [Opper 89’; Krogh, Hertz ’92; Geman et al. ’92; Opper ’95; Neyshabur, Tomyoka, Srebro ’15; Advani, Saxe ’17]

The “double descent”

[Belkin et al. ’18]

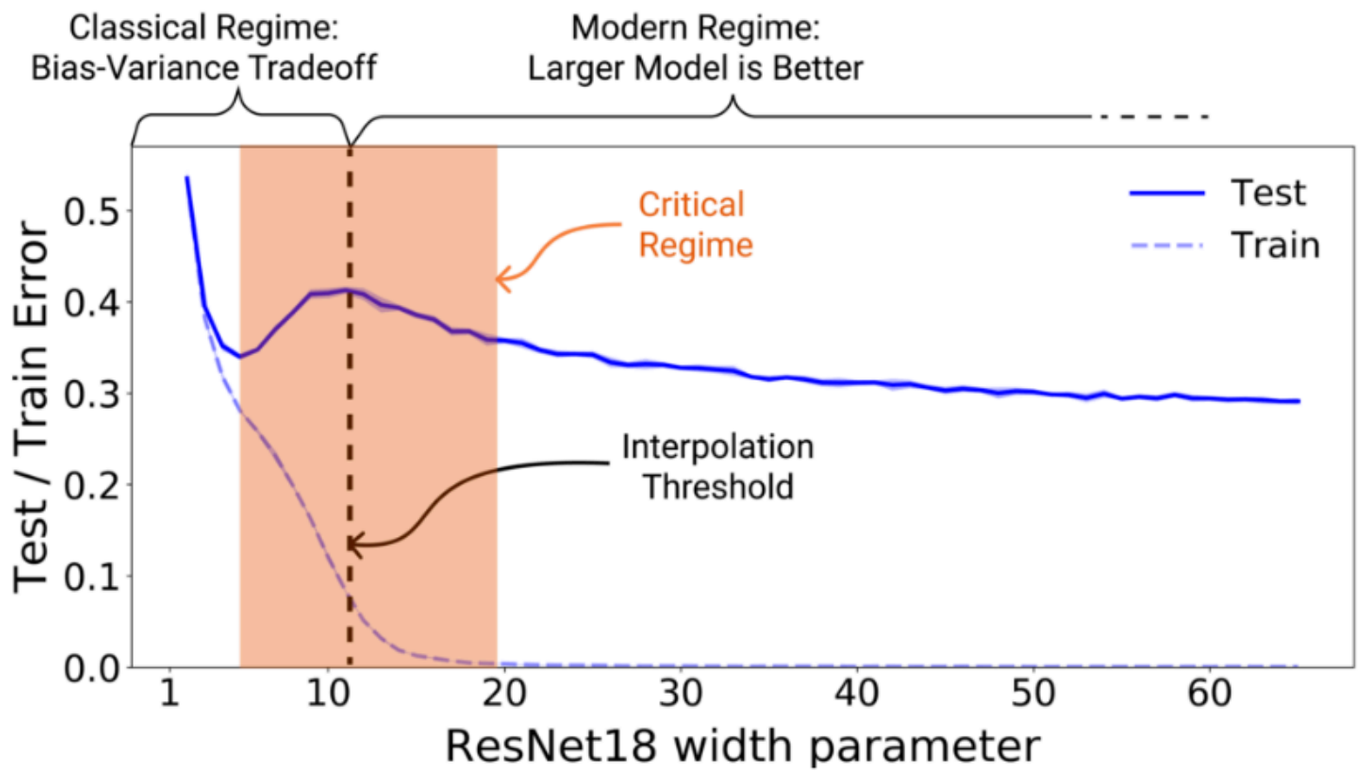
[Geiger et al. ’18]



Number of parameters

Parity-MNIST, 5 layers, fully-connected, no regularisation

[Nakkiran et al. ’19]



CIFAR10, no regularisation

Questions:

1. Why error goes down?
2. Why the peak?

See also [Opper 89’; Krogh, Hertz ’92; Geman et al. ’92; Opper ’95; Neyshabur, Tomyoka, Srebro ’15; Advani, Saxe ’17]

Questions:
1. **Why error goes down?**
2. Why the peak?

Recent development:
"Benign overfitting"
in simple convex models.

[Bartlett et al. '19; Hastie et al. 19']

Questions:
1. **Why error goes down?**
2. Why the peak?

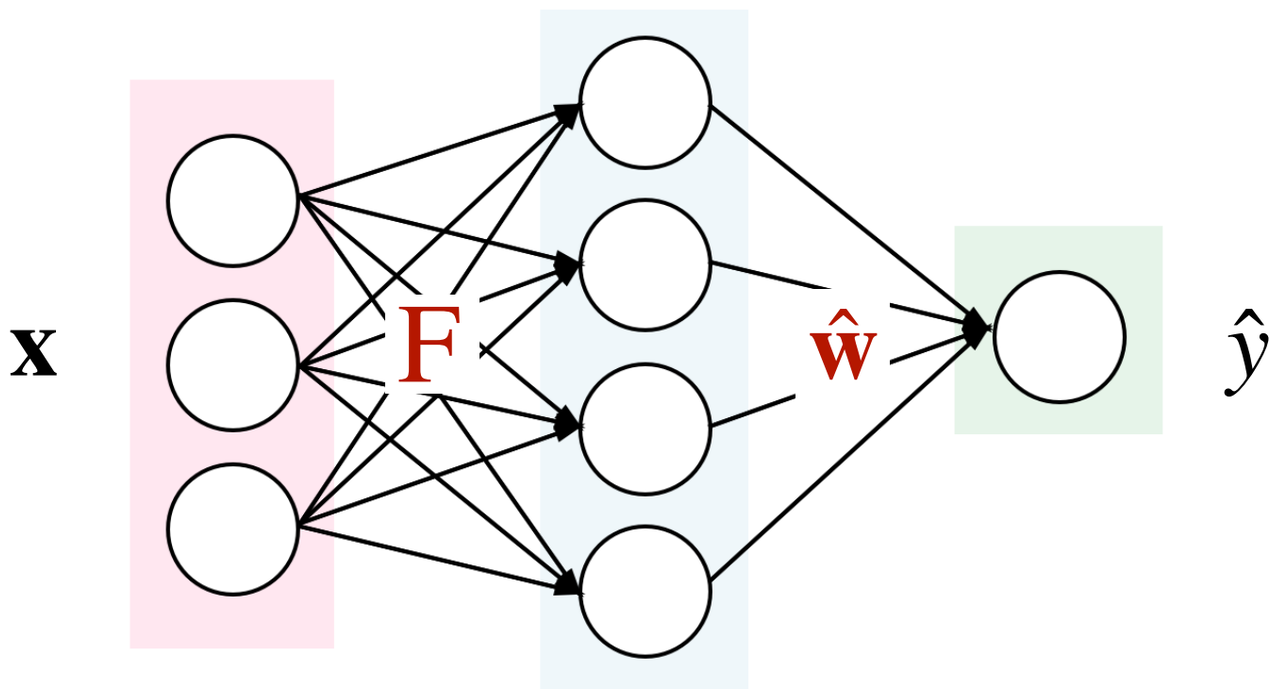
Recent development:
"Benign overfitting"
in simple convex models.

[Bartlett et al. '19; Hastie et al. 19']

The random features model

[Williams '98,'07; Retch, Raimi '07]

Fit $\hat{y}(\mathbf{x}) = \hat{f}(\hat{\mathbf{w}}^\top \varphi(\mathbf{x}))$ by ERM. Feature map $\varphi(\mathbf{x}) = \sigma(\mathbf{F}\mathbf{x})$, random $\mathbf{F} \in \mathbb{R}^{p \times d}$



GP / NTK / Lazy training:
≈ of deep networks

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

Exact asymptotics

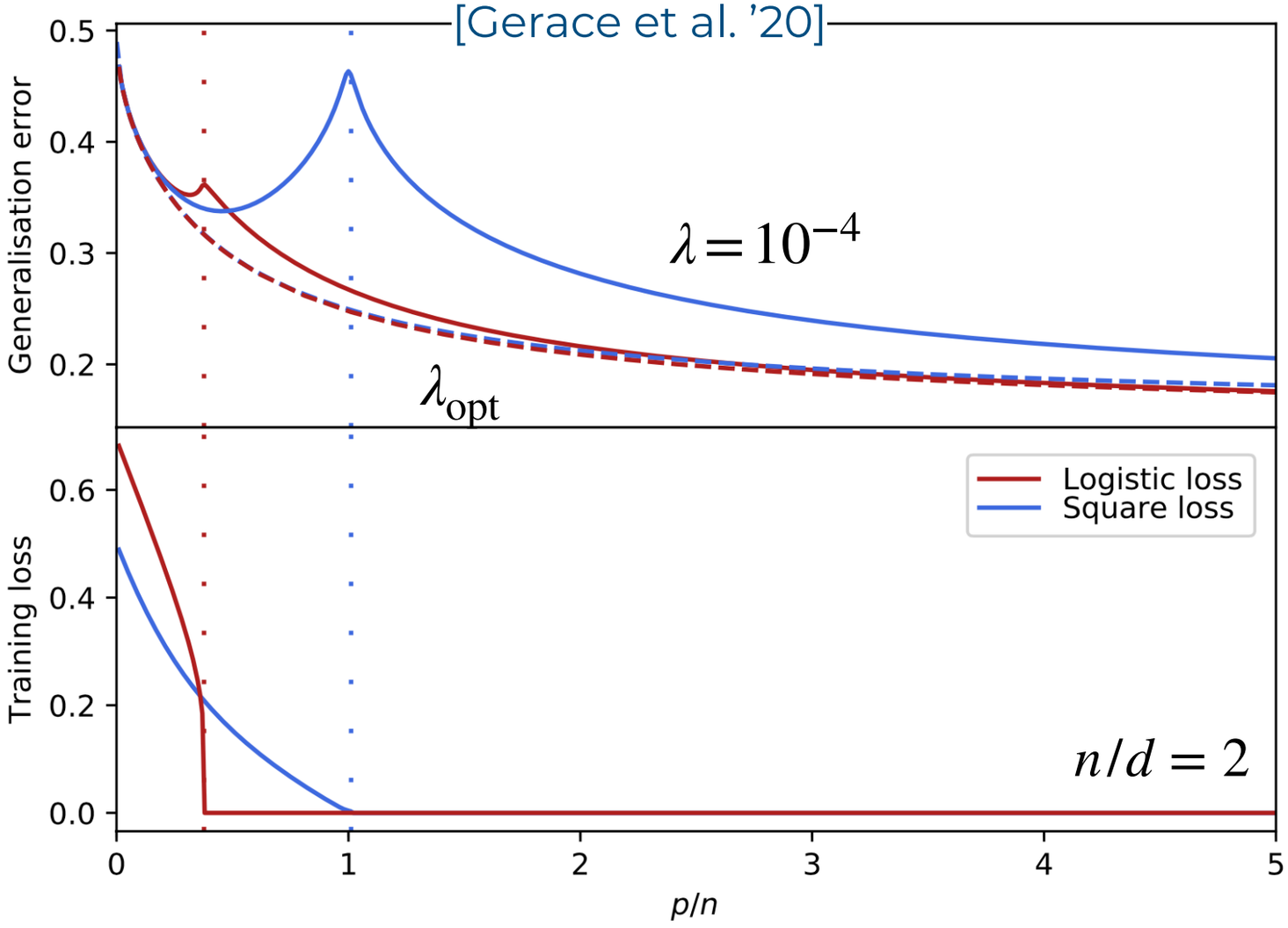
Key result:

Exact asymptotics when $n, d, p \to \infty$ with fixed n/d and p/n for convex risks, generic F and data $\mathbf{x} \sim \mathcal{N}(0, \mathsf{I}_d)$,

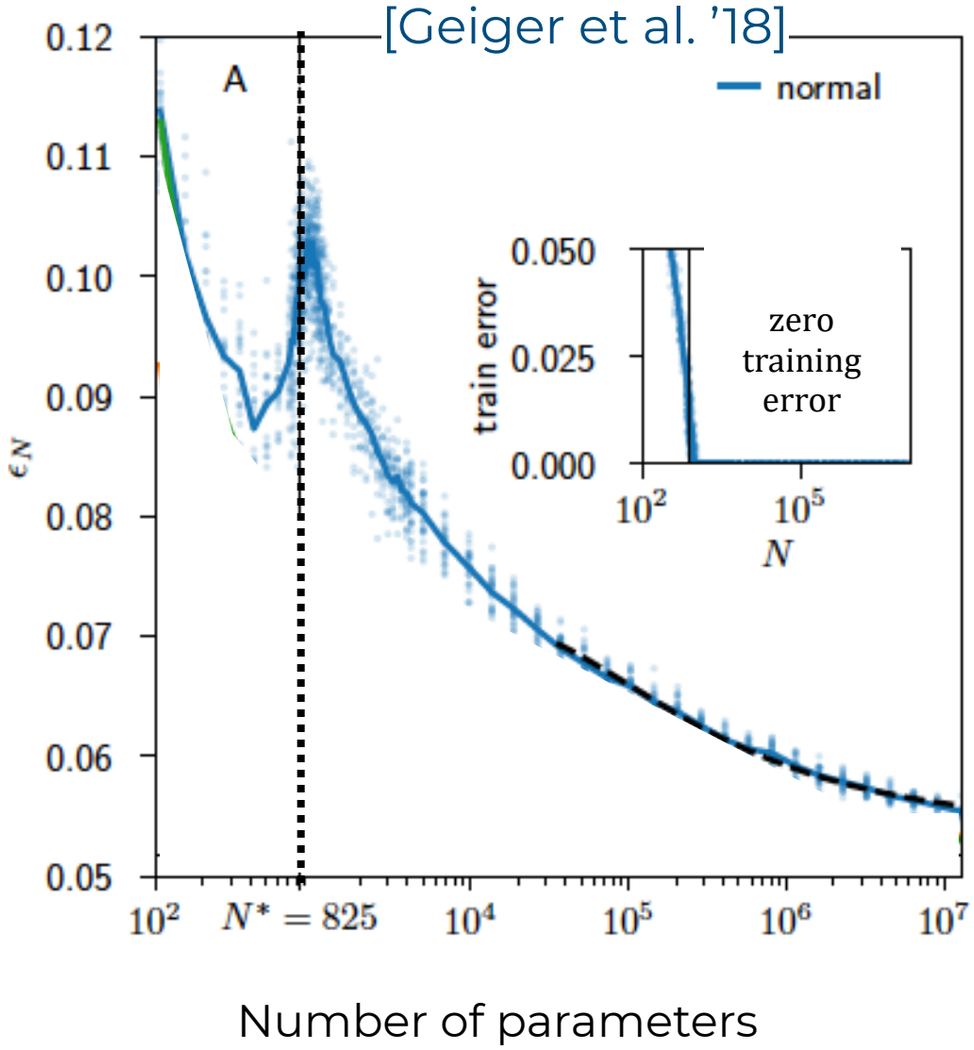
$$y^\mu = f_0(\theta_0^\top \mathbf{x}^\mu)$$

[Montanari, Mei 19'; Gerace et al '20; Goldt et al '20; Hu, Lu, '20; Dhifallah, Lu '20]

[Gerace et al. '20]



[Geiger et al. '18]

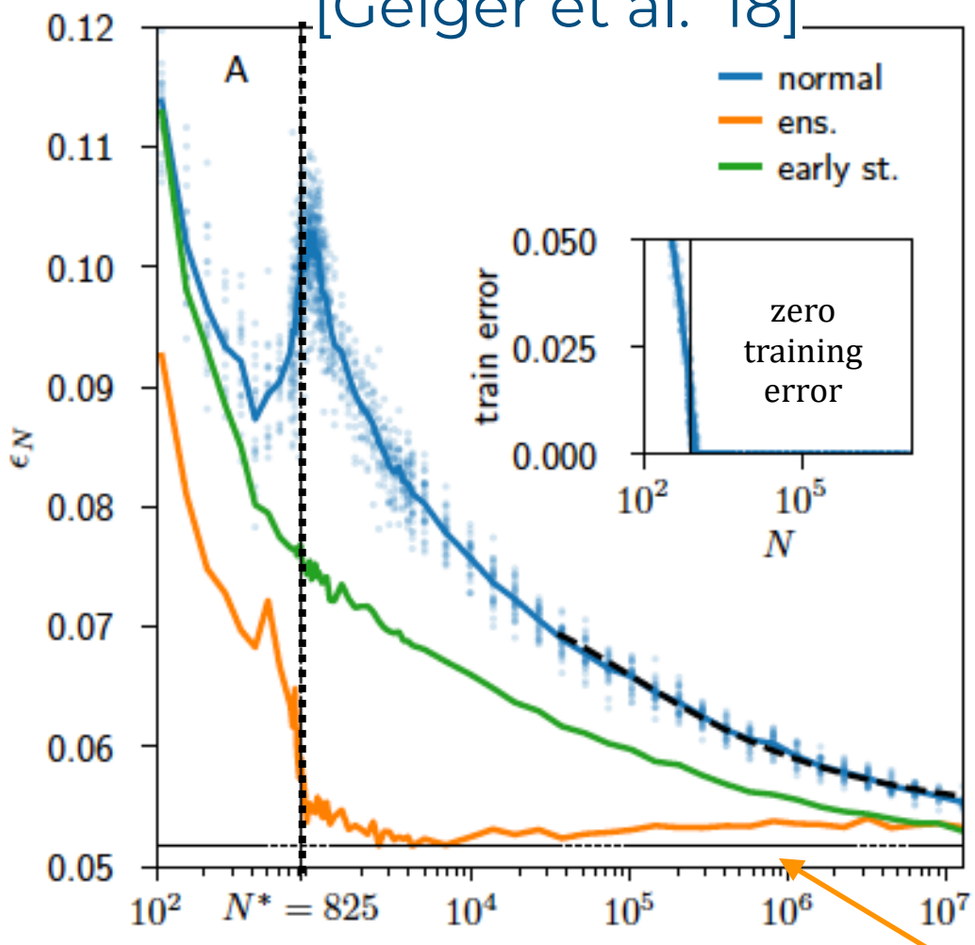


Number of parameters

$y^\mu = \text{sign}(\theta_0^\top \mathbf{x})$ $\quad \hat{y} = \text{sign}\left(\hat{\mathbf{w}}^\top \text{erf}\left(\mathbf{F}\mathbf{x}\right)\right)$ $\quad \mathbf{x}, \theta_0 \sim \mathcal{N}(0, \mathbf{I}_d)$ $\quad F$ Gaussian i.i.d

1. Why error goes down?
2. **Why the peak?**

[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 independently trained NNs

Ensemble of random features

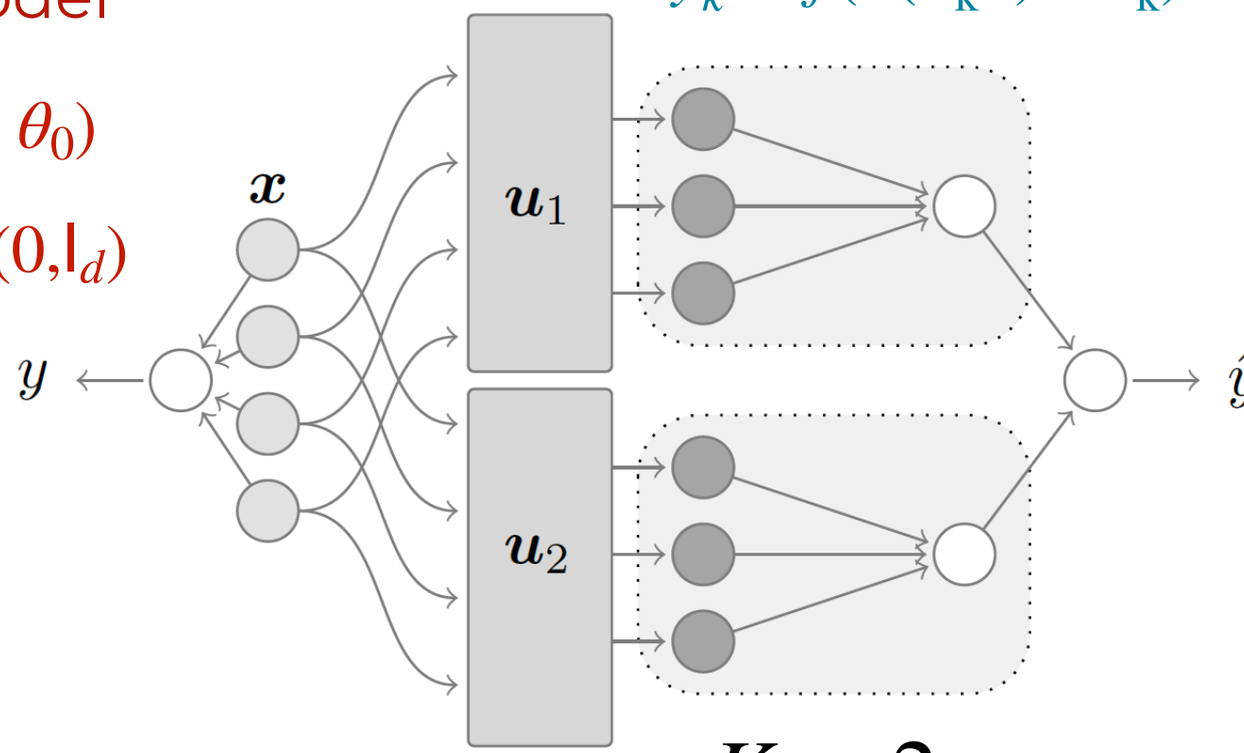
Predictors

$$\hat{y}_k = \hat{f}(\sigma(\mathrm{F}_k\mathbf{x}) \cdot \mathbf{w}_k)$$

Data model

$$y = f_0(\mathbf{x} \cdot \theta_0)$$

$$\mathbf{x}, \theta_0 \sim \mathcal{N}(0, \mathrm{I}_d)$$



$K = 2$

Final predictor

$$\hat{y} = f(\{\sigma(\mathrm{F}_k\mathbf{x}) \cdot \mathbf{w}_k\}_{k=1}^{K})$$

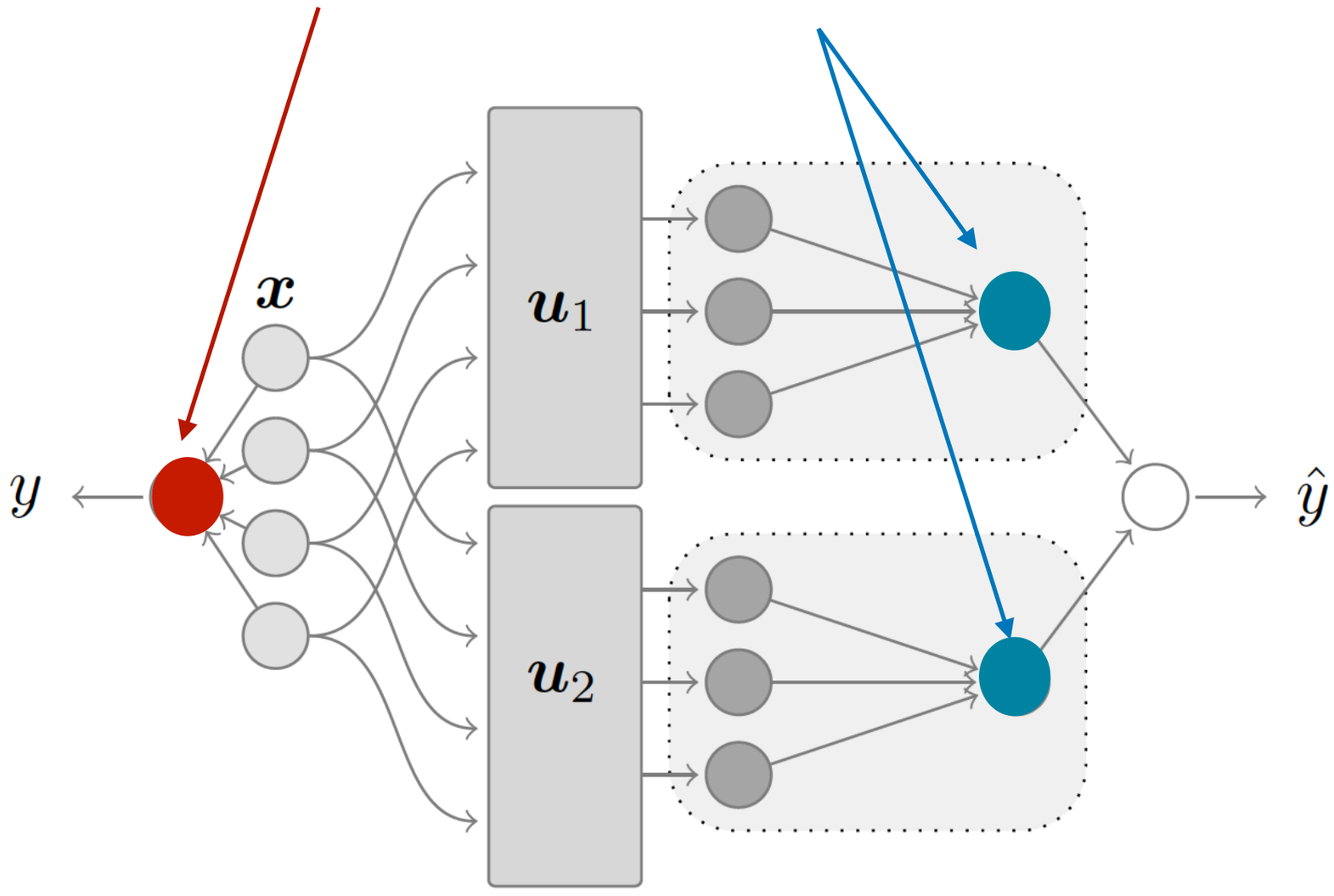
Learning: $$\hat{\mathscr{R}}_n(\mathbf{w}_k) = \frac{1}{n}\sum_{\mu=1}^{n} \ell\left(y^\mu, \mathbf{w}_k^\top \sigma(\mathrm{F}\mathbf{x}^\mu)\right) + \frac{\lambda}{2}||\mathbf{w}_k||_2^2$$

In the limit $n, p, d \to \infty$, with K, $\alpha \equiv n/d$, $\gamma \equiv p/d$ fixed

Main result

Theorem (informal):

The pre-activations $\nu = \theta_0^\top \mathbf{x}^\mu, \quad \mu_k = \hat{\mathbf{w}}_k^\top \sigma(\mathrm{F_k x})$



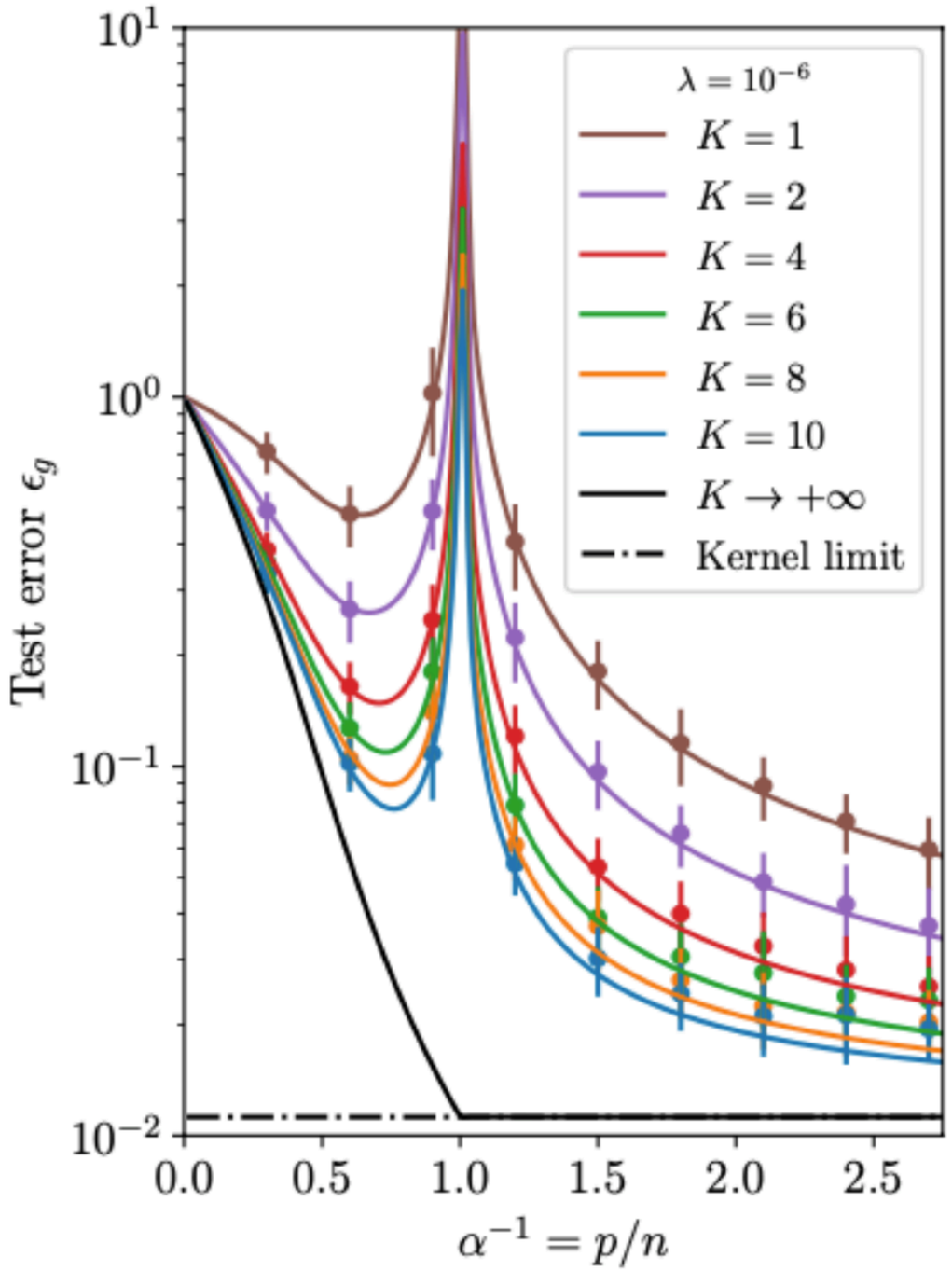
are jointly r.v. with covariances given by self-consistent equation.

(Theorem proven in more general setting)

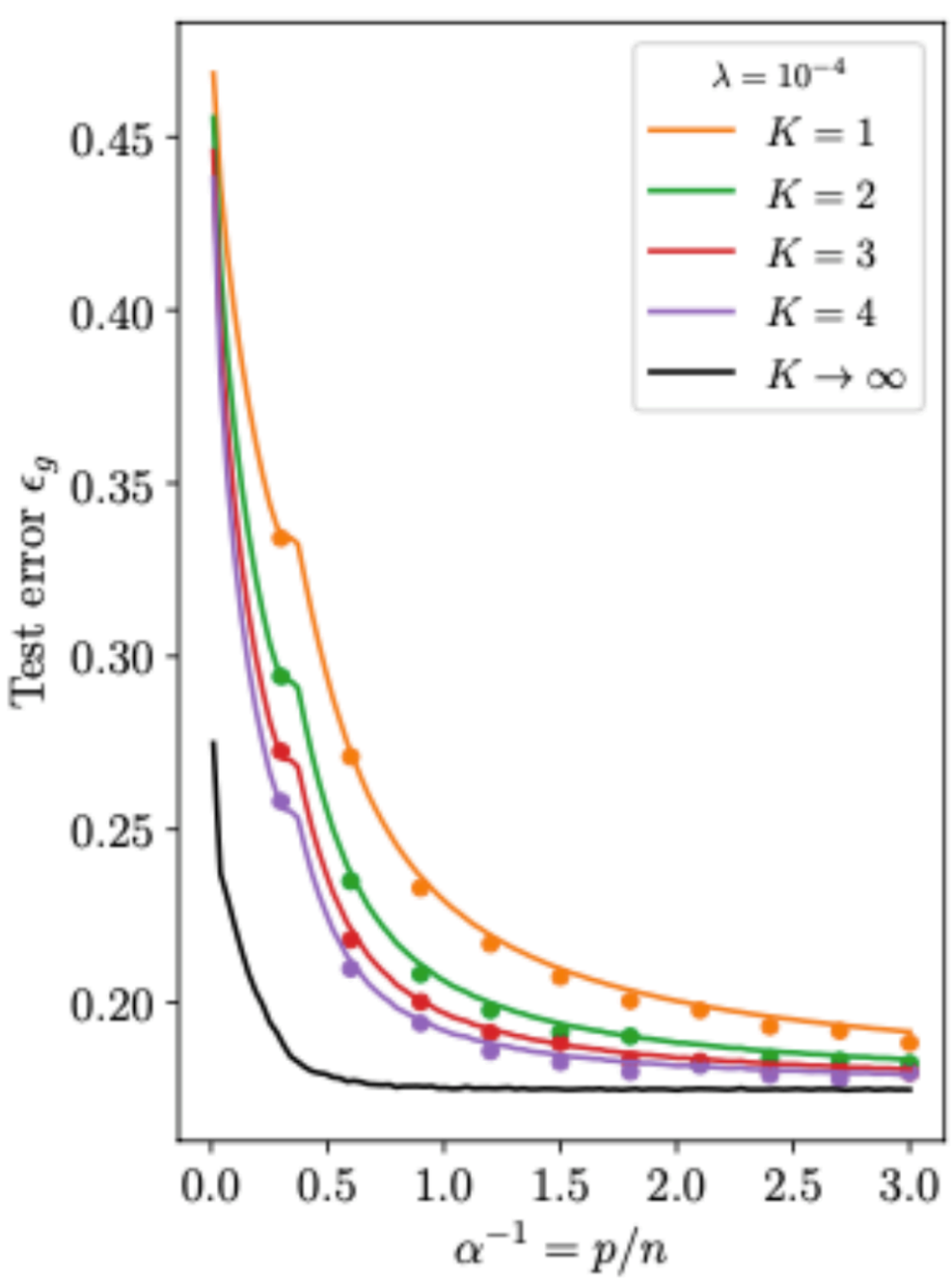
Generalise [Biroli et al '20; Adlam, Pennington '20; Lin, Dobriban '20;]

Suppressing fluctuations

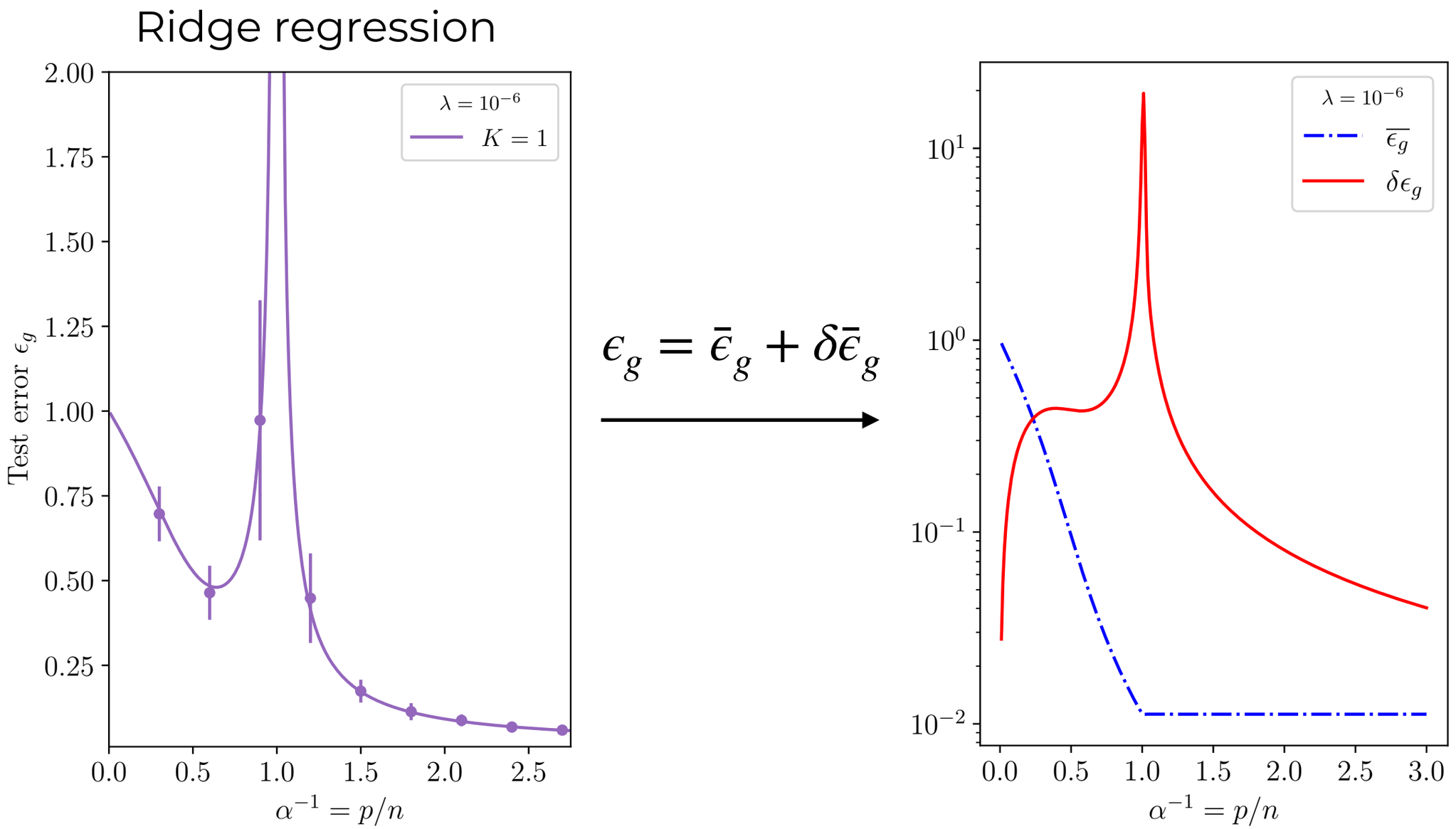
Ridge regression



Logistic regression



Bias / variance Trade-off



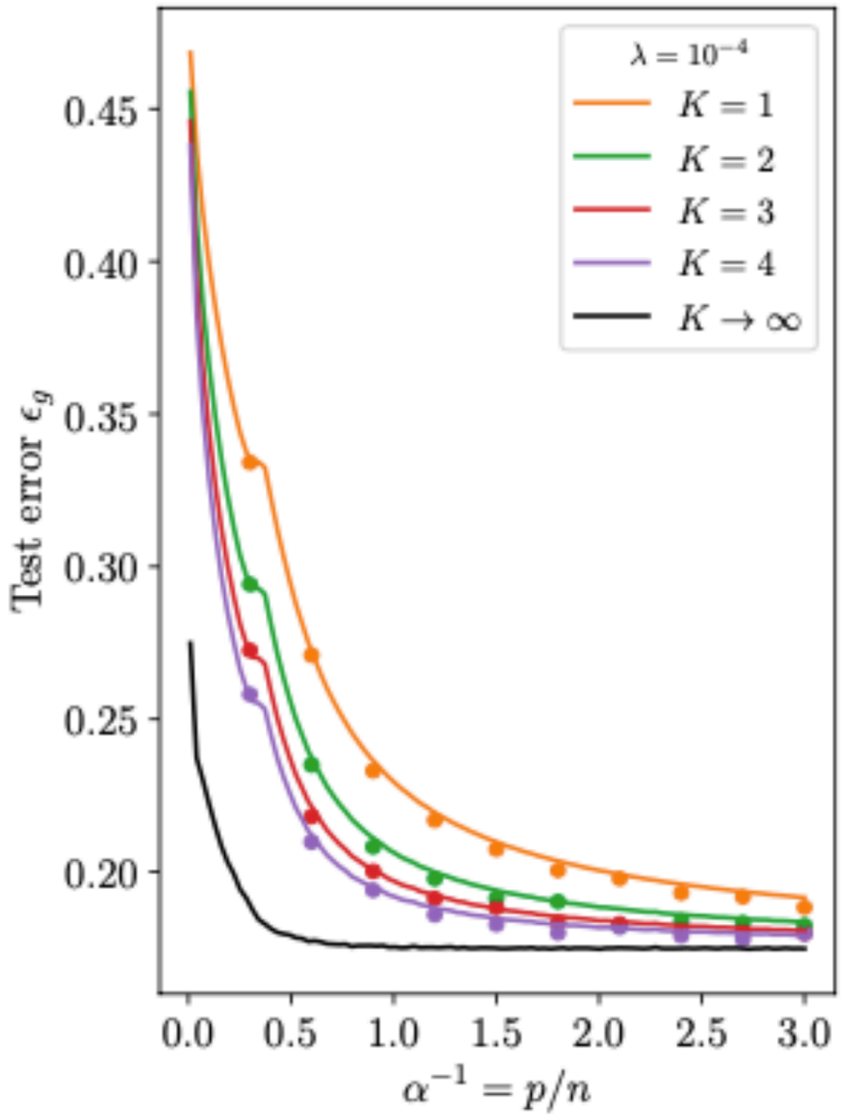
Bias (approximation error): decrease and vanishes at interpolation

Variance: overfitting of random weights $\mathbf{F}$ fluctuations

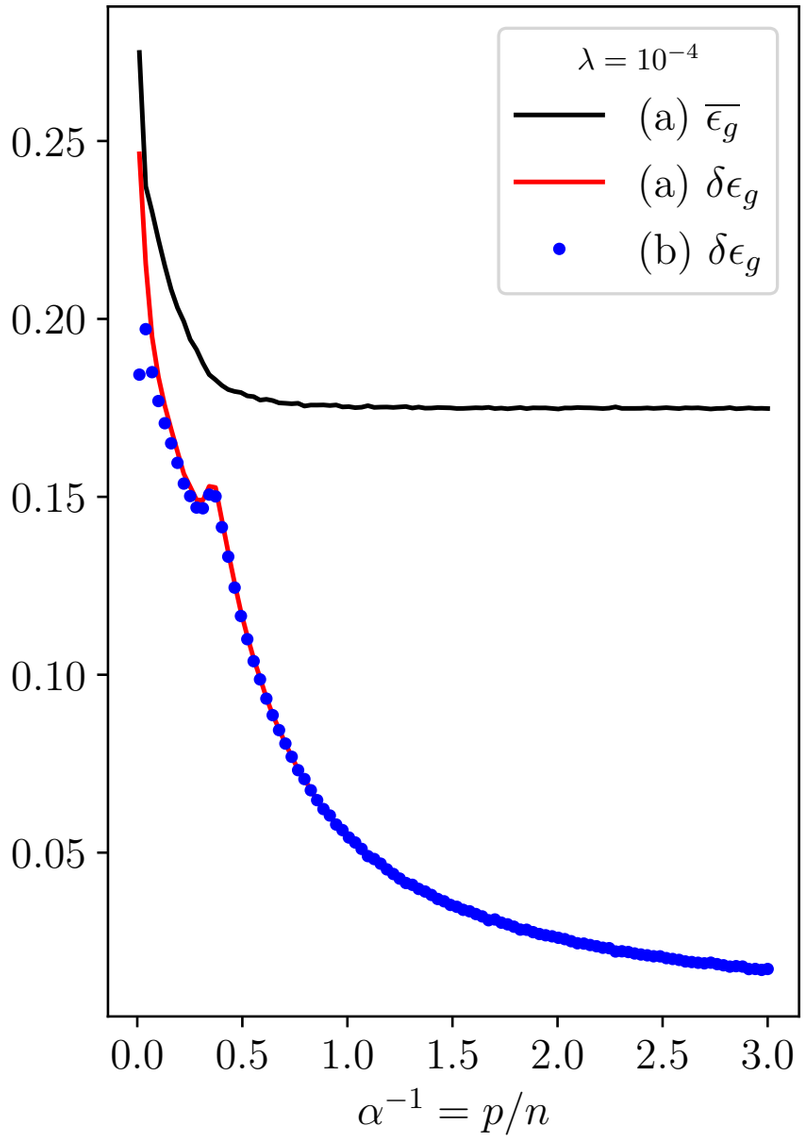
[Krogh, Vedelsby ’95; Biroli et al ’20]

Bias / variance Trade-off

Logistic regression



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



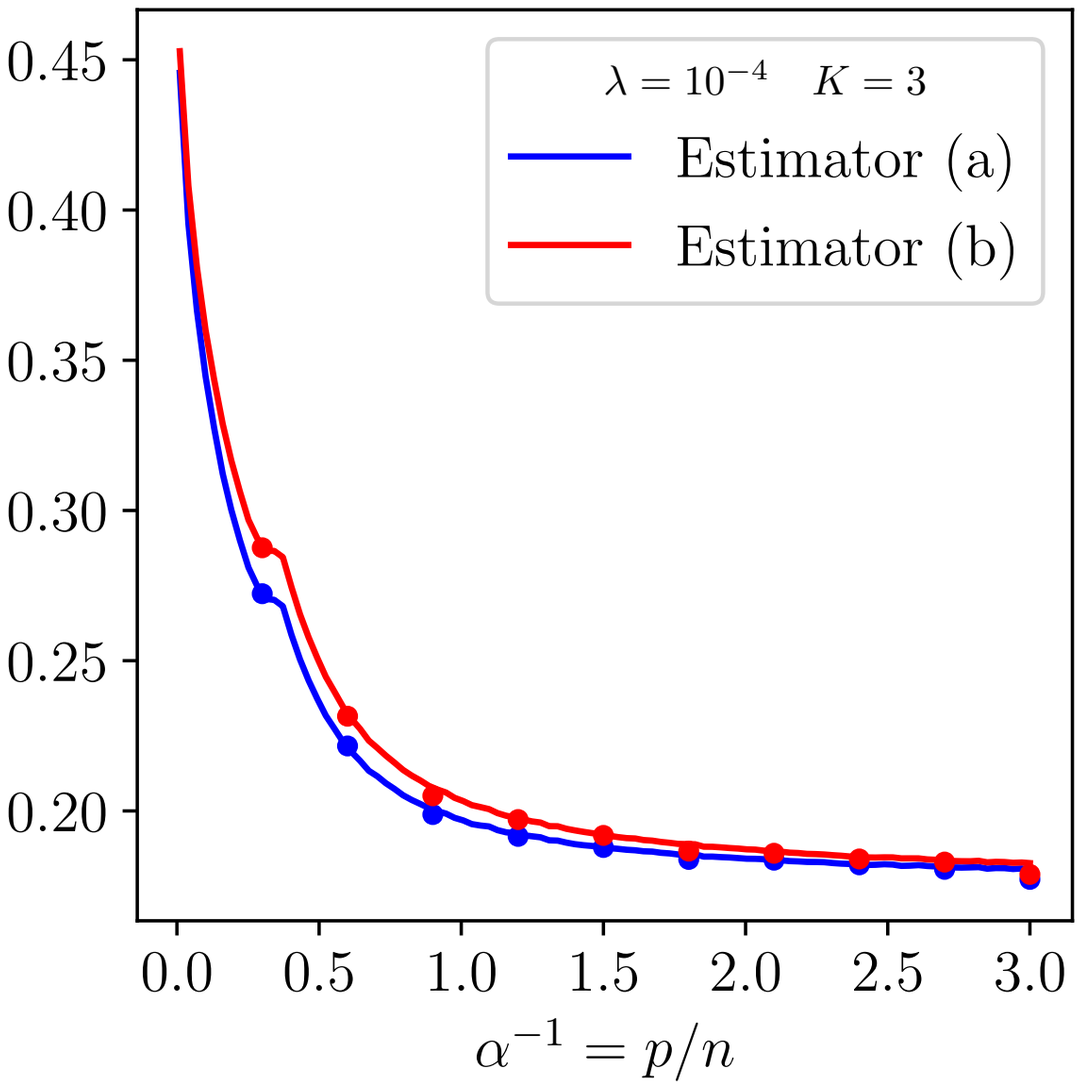
Bias (approximation error): decrease and vanishes at interpolation

Variance: overfitting of random weights $\mathbf{F}$ fluctuations

[Krogh, Vedelsby ’95; Biroli et al ’20]

Ensembling methods

Logistic regression, $\lambda \to 0^+$



a) $\mathbf{v} \in \mathbb{R}^k \mapsto \text{sign}\left(\sum_{k=1}^{K} v_k\right)$

b) $\mathbf{v} \in \mathbb{R}^k \mapsto \text{sign}\left(\sum_{k=1}^{K} \text{sign}(v_k)\right)$

"Majority vote"

Conclusion

- Exact asymptotics for an ensemble of random features.
- Valid for: strongly convex risks, generic weight distribution
- Why the peak (for RFs)? Overfitting of fluctuations of weights.

Thank you!

brloureiro@gmail.com