DOUBLE TROUBLE IN DOUBLE DESCENT

BIAS AND VARIANCE(S) IN THE LAZY REGIME

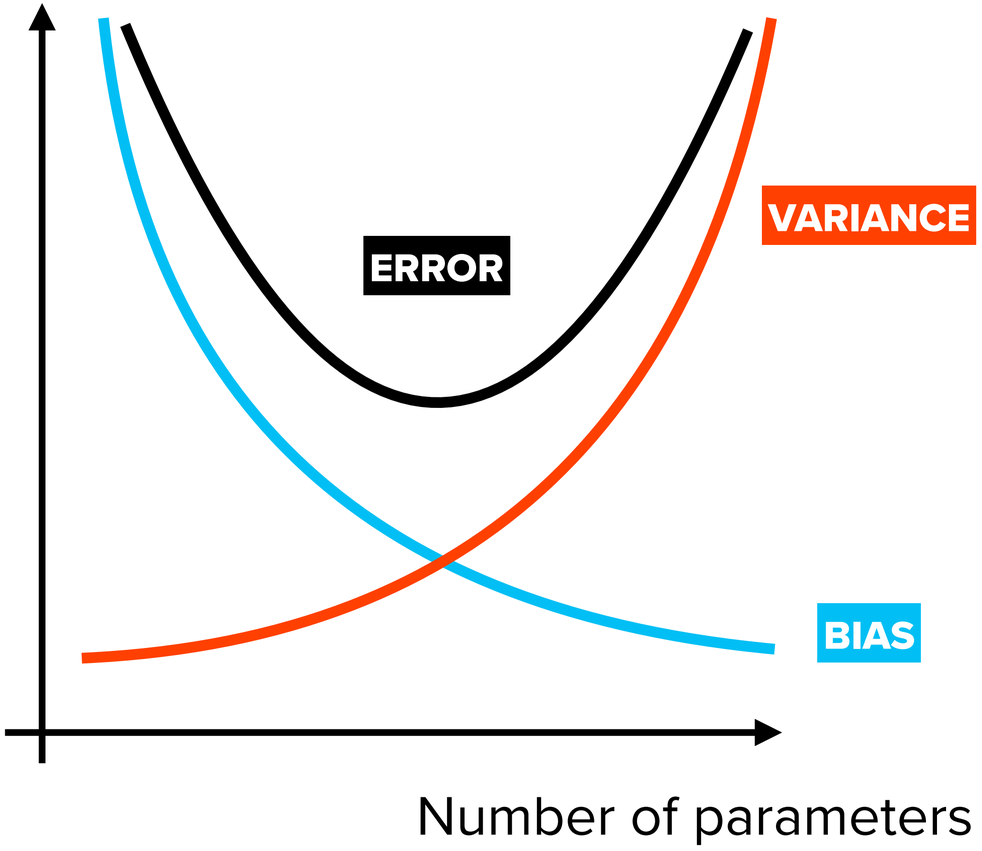
STÉPHANE D'ASCOLI, MARIA REFINETTI,
GIULIO BIROLI, FLORENT KRZAKALA

ÉCOLE NORMALE SUPÉRIEURE DE PARIS

FROM CLASSICAL THEORY

BIAS-VARIANCE TRADEOFF

"The price to pay for achieving low bias is high variance"
(Geman et al., 1992)

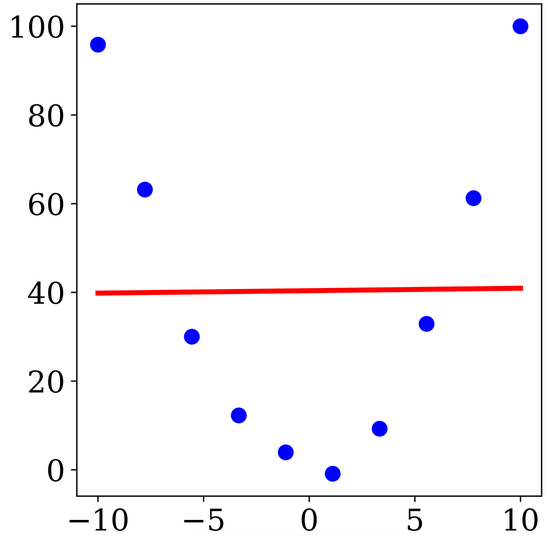


FROM CLASSICAL THEORY

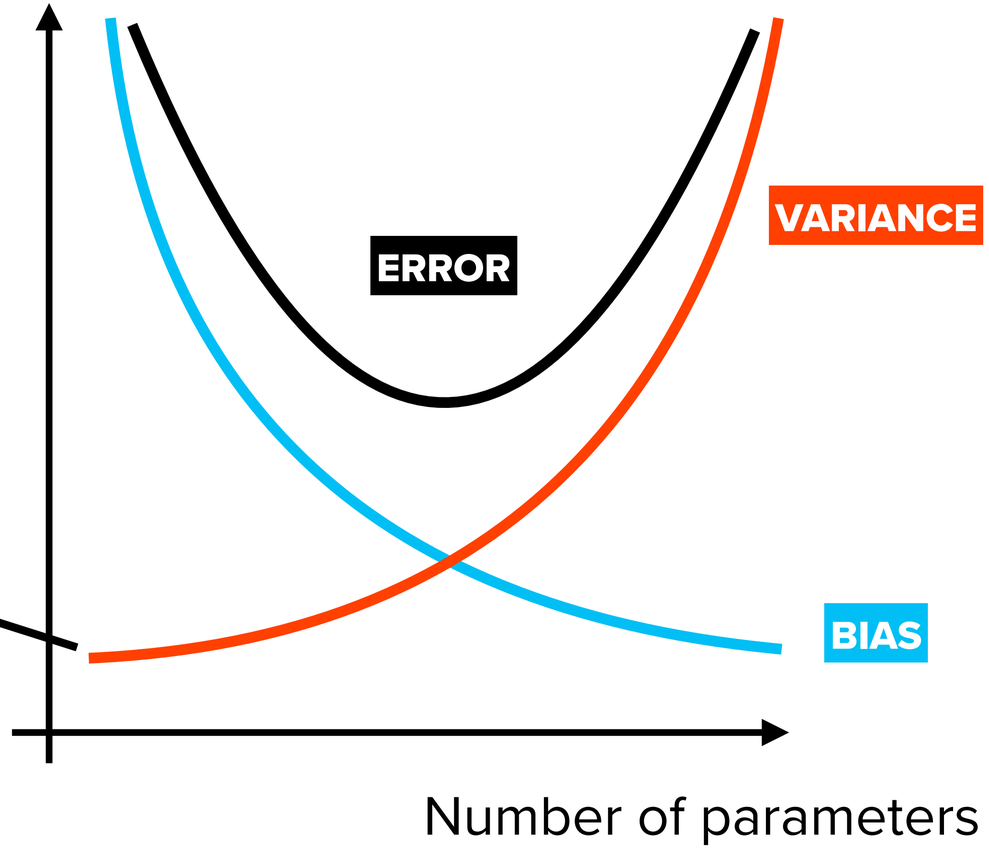
BIAS-VARIANCE TRADEOFF

"The price to pay for achieving low bias is high variance"
(Geman et al., 1992)

TOO FEW PARAMETERS:
HIGH BIAS, LOW VARIANCE



UNDERFIT

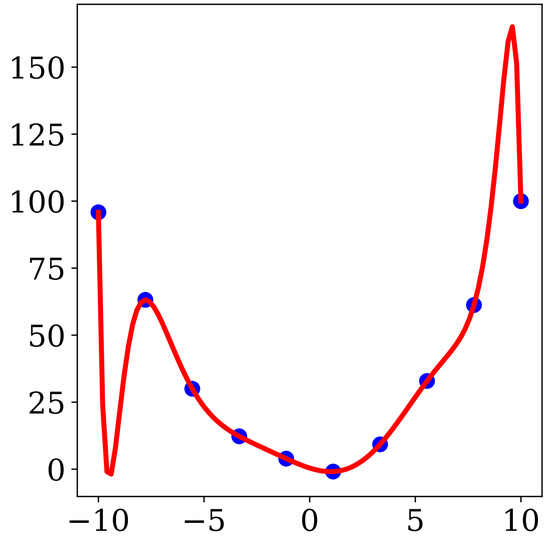


FROM CLASSICAL THEORY

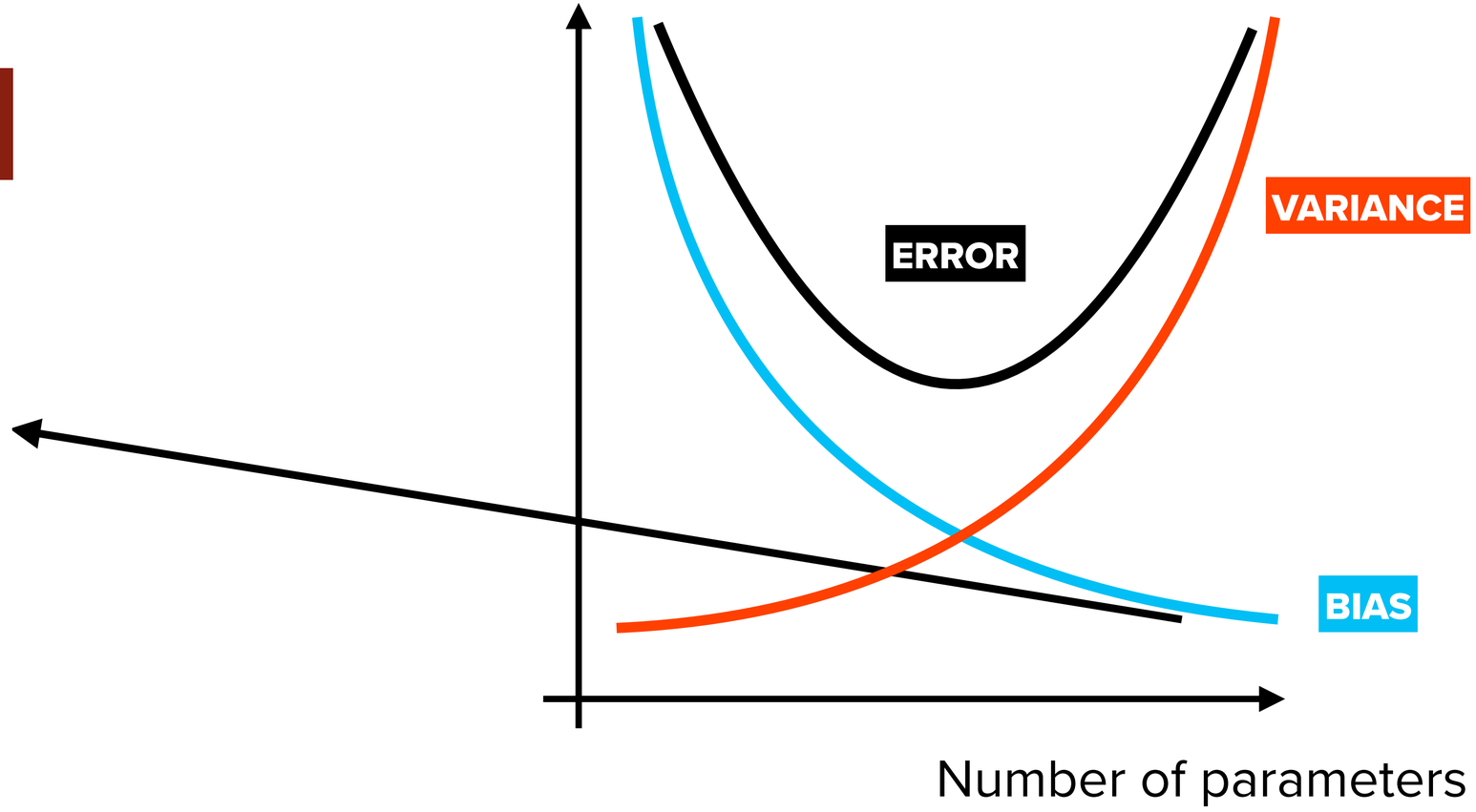
BIAS-VARIANCE TRADEOFF

"The price to pay for achieving low bias is high variance"
(Geman et al., 1992)

TOO MANY PARAMETERS: LOW BIAS, HIGH VARIANCE



OVERFIT

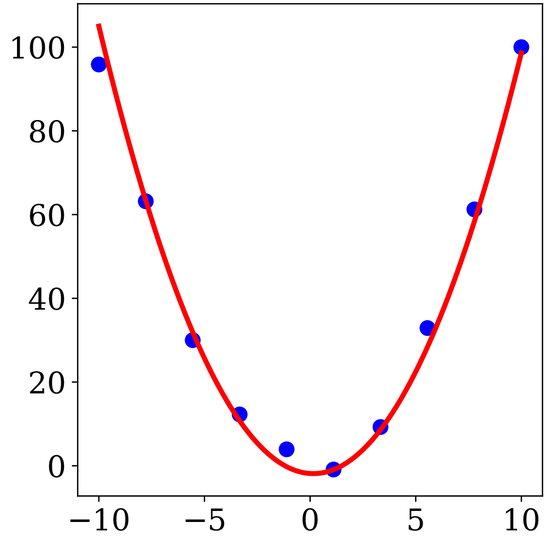


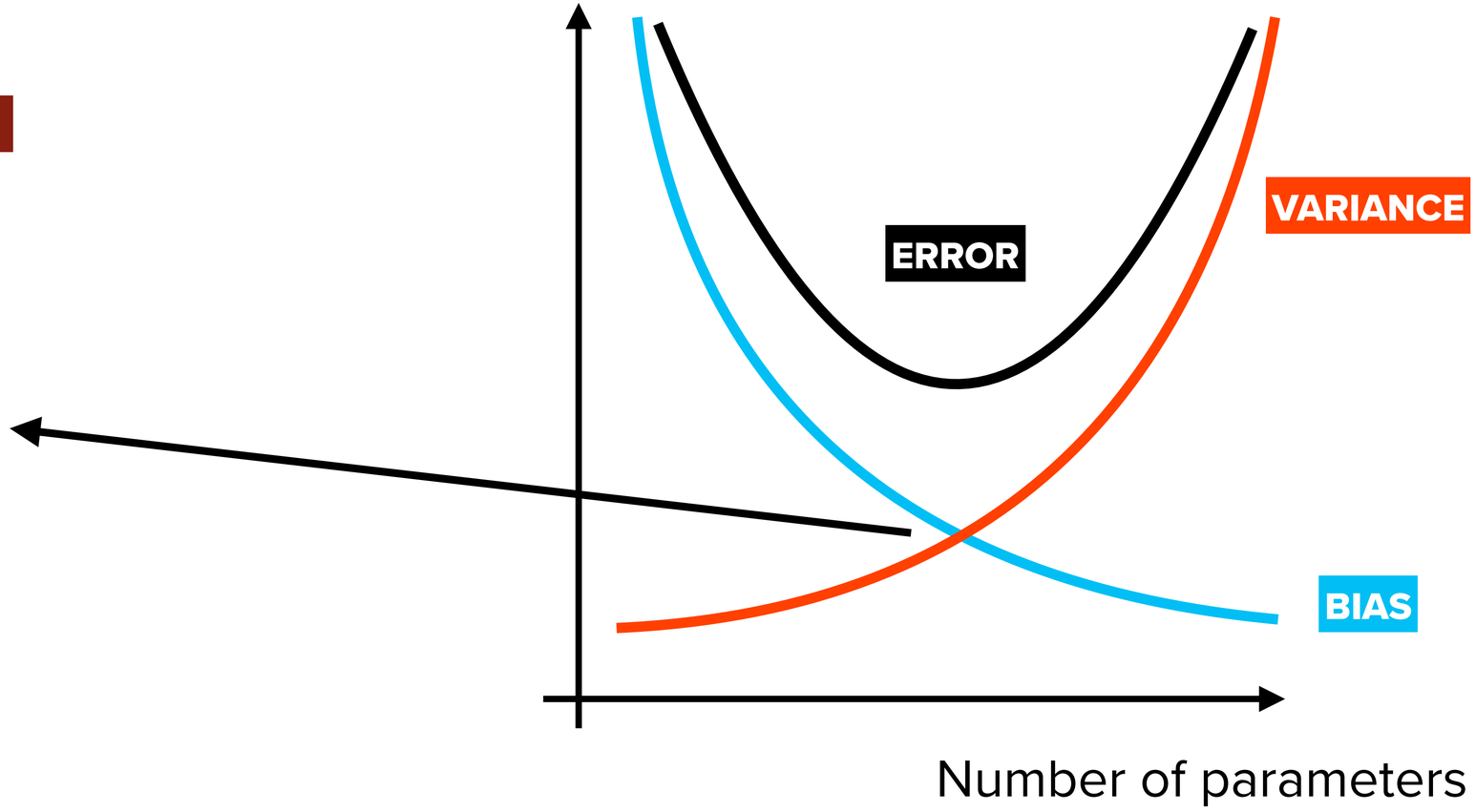
FROM CLASSICAL THEORY

BIAS-VARIANCE TRADEOFF

"The price to pay for achieving low bias is high variance"
(Geman et al., 1992)

BALANCED



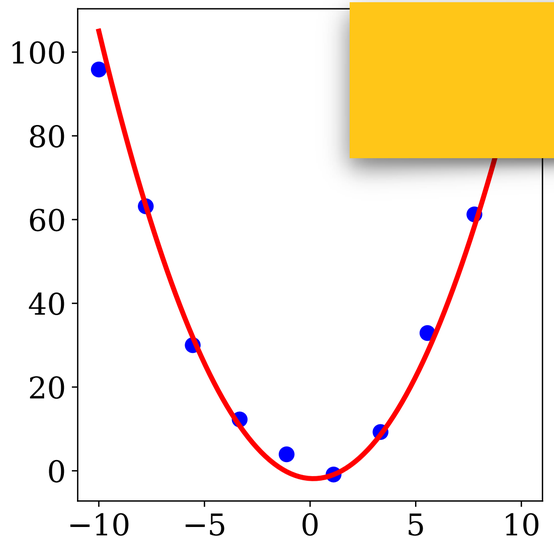


FROM CLASSICAL THEORY

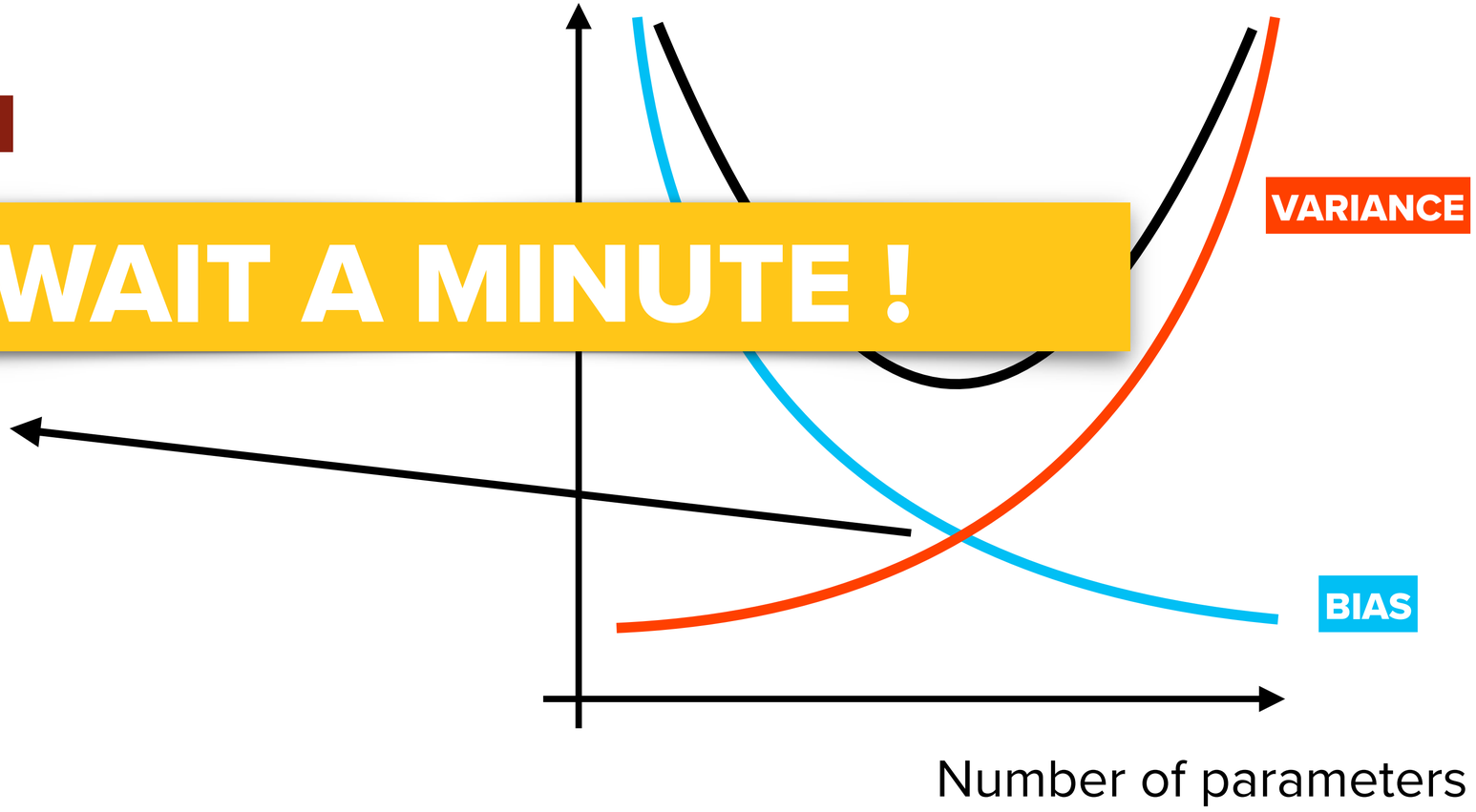
BIAS-VARIANCE TRADEOFF

"The price to pay for achieving low bias is high variance"
(Geman et al., 1992)

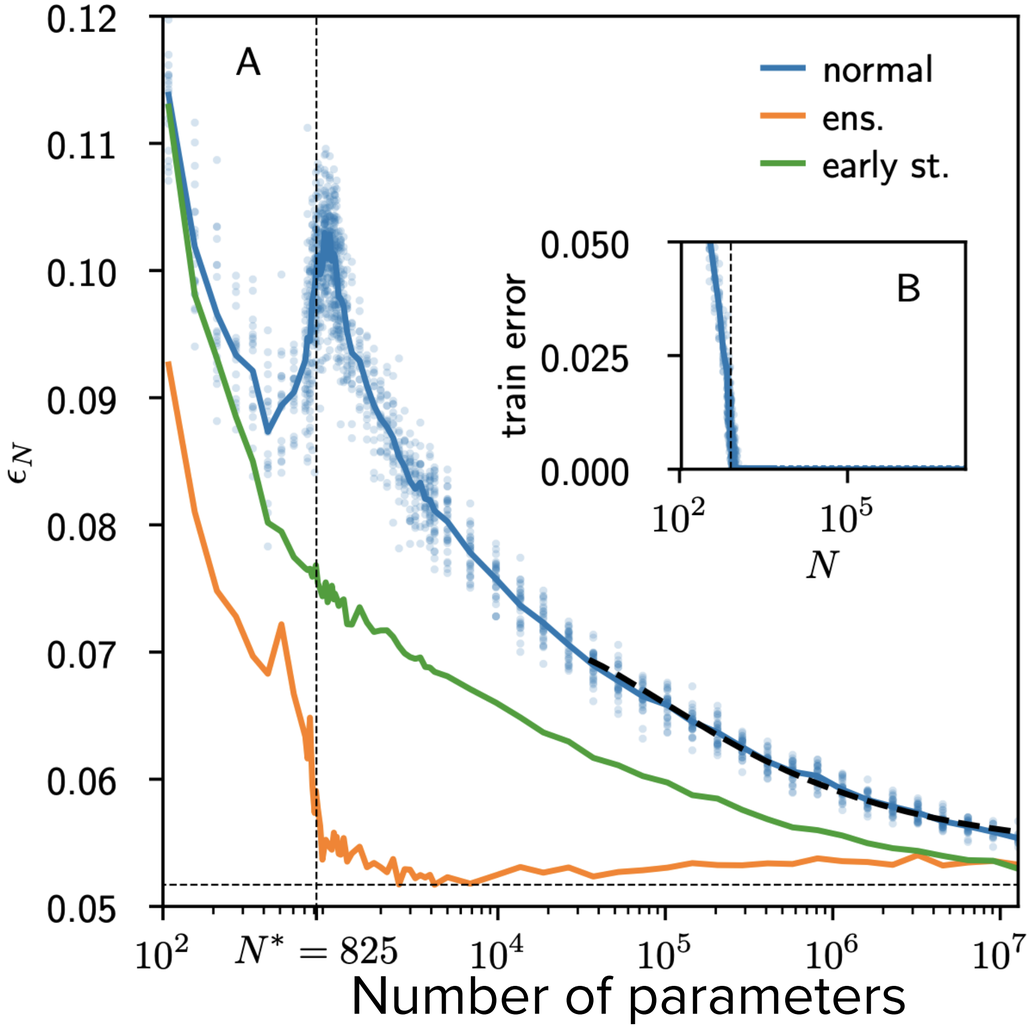
BALANCED



WAIT A MINUTE !



TO DOUBLE DESCENT



See also

Belkin et al. 2019

Nakkiran et al. 2019

Hastie et al. 2019

Mei & Montanari 2019

Kini et al. 2020

Gerace et al. 2020

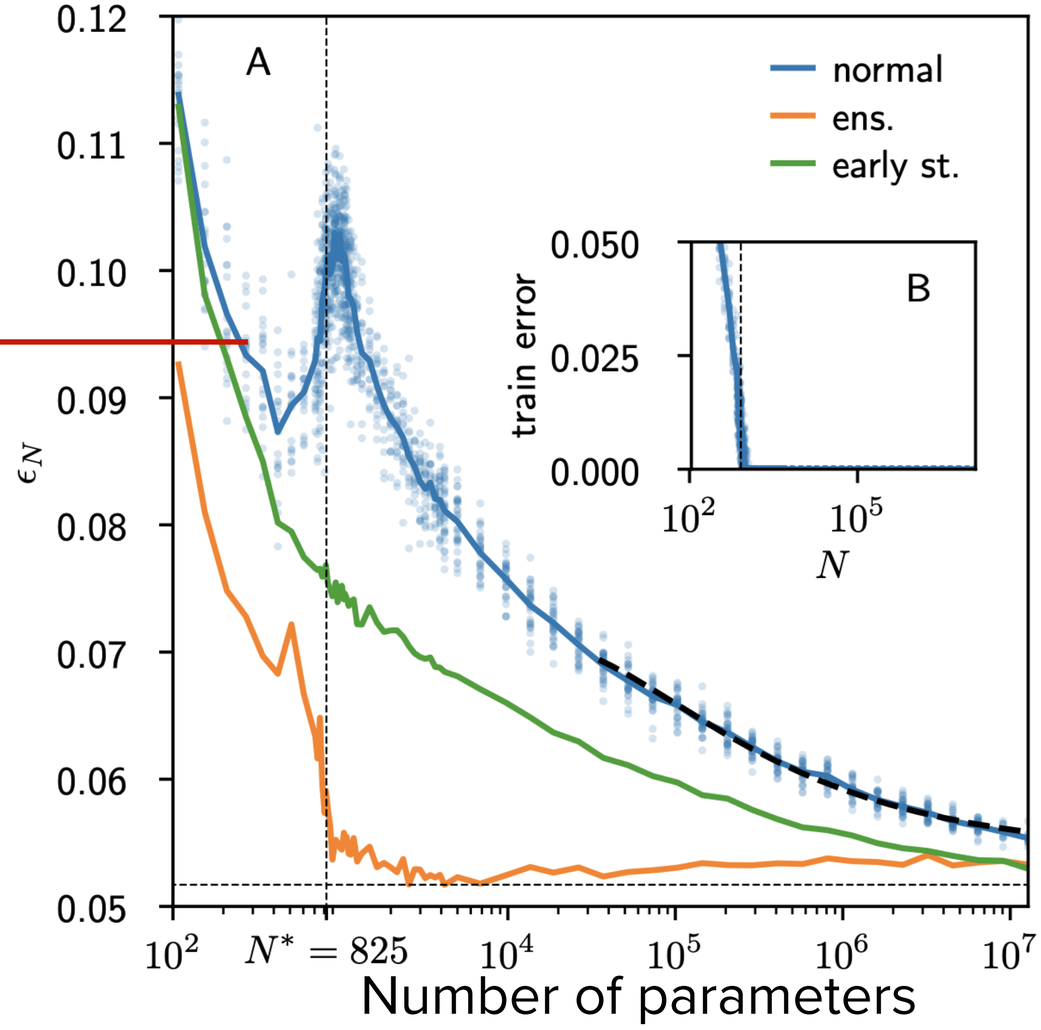
Ba et al. 2020

....

Geiger, M., Jacot, A., Spigler, S., Gabriel, F., Sagun, L., d'Ascoli, S., Biroli, G., Hongler, C., and Wyart, M. Scaling description of generalization with number of parameters in deep learning. *arXiv preprint arXiv:1901.01608*, 2019a.

TO DOUBLE DESCENT

"U" shaped curve in the underparametrized regime



See also

Belkin et al. 2019

Nakkiran et al. 2019

Hastie et al. 2019

Mei & Montanari 2019

Kini et al. 2020

Gerace et al. 2020

Ba et al. 2020

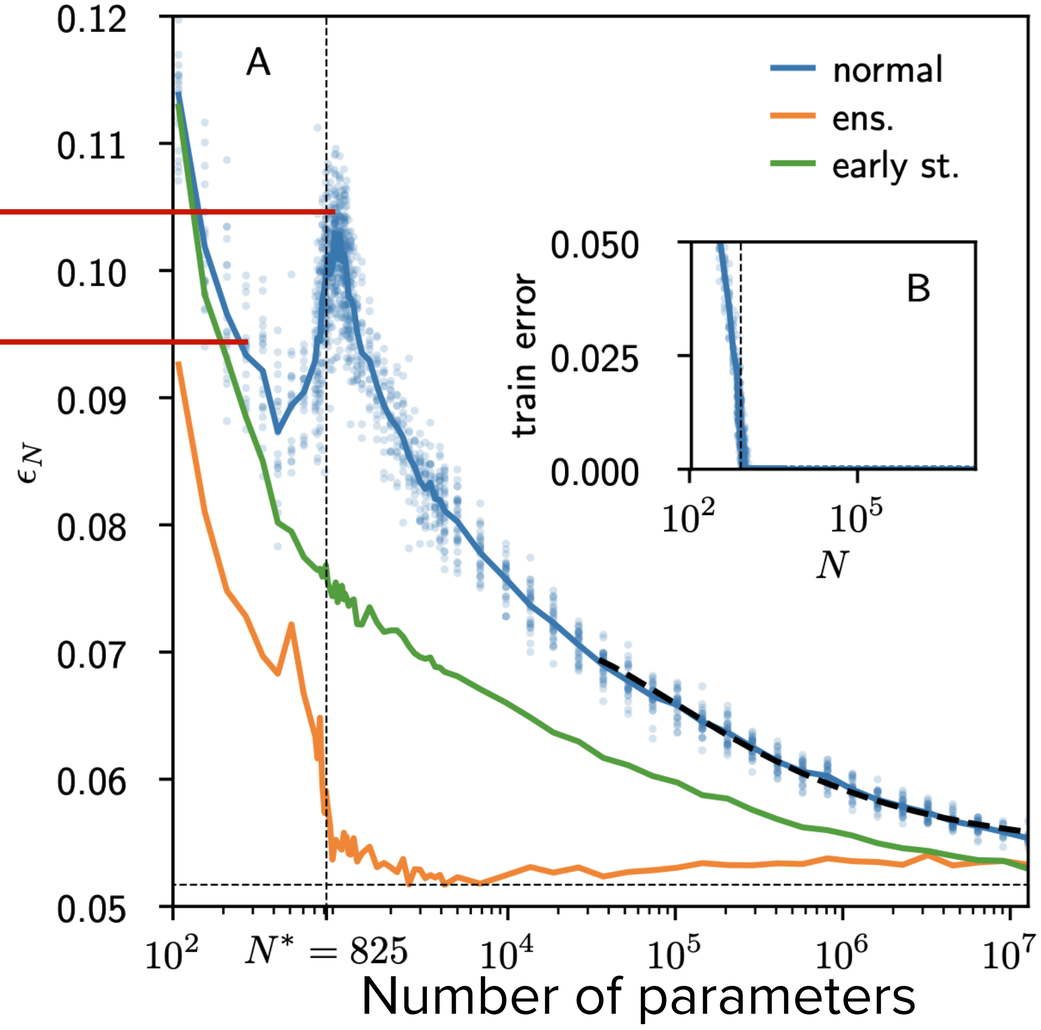
....

Geiger, M., Jacot, A., Spigler, S., Gabriel, F., Sagun, L., d'Ascoli, S., Biroli, G., Hongler, C., and Wyart, M. Scaling description of generalization with number of parameters in deep learning. *arXiv preprint arXiv:1901.01608*, 2019a.

TO DOUBLE DESCENT

Peak in Test Error
At interpolation threshold

"U" shaped curve in the
underparametrized regime



See also

Belkin et al. 2019

Nakkiran et al. 2019

Hastie et al. 2019

Mei & Montanari 2019

Kini et al. 2020

Gerace et al. 2020

Ba et al. 2020

....

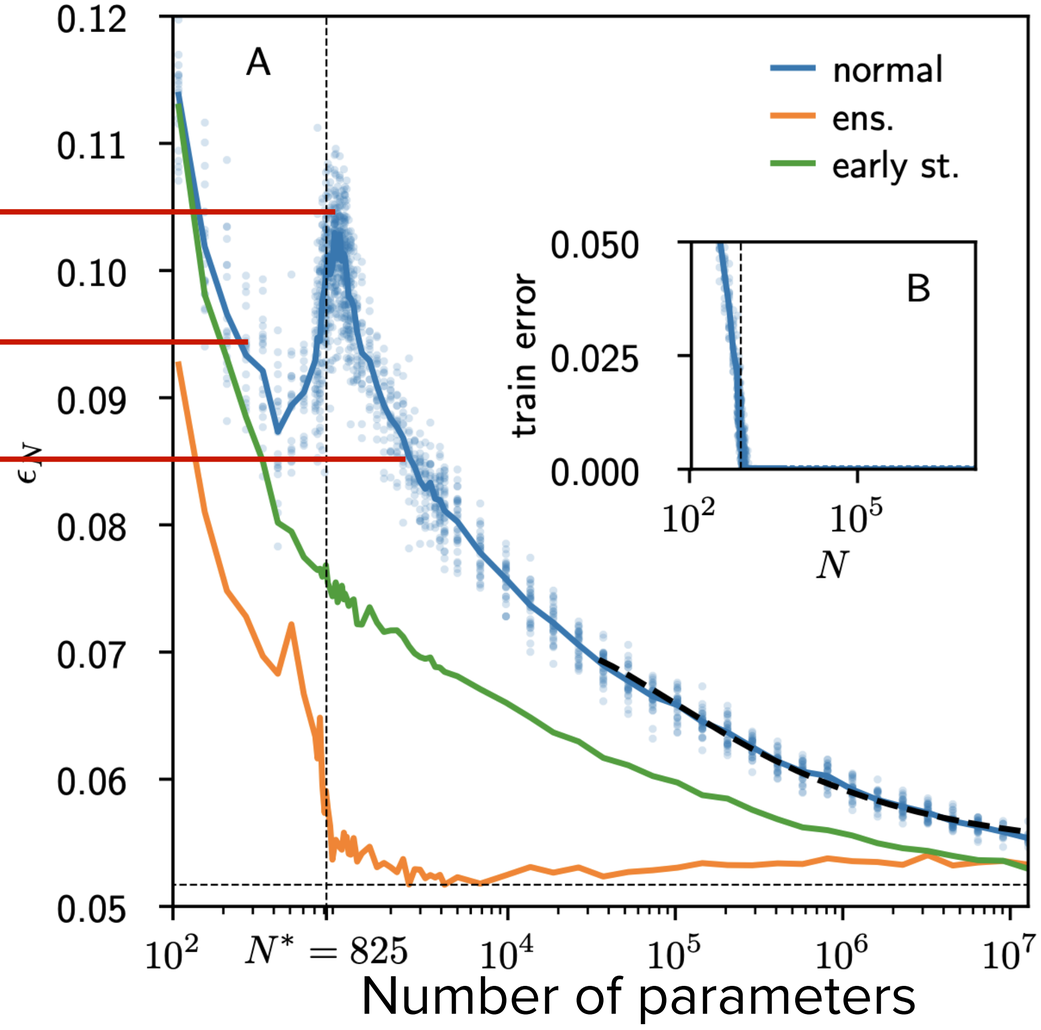
Geiger, M., Jacot, A., Spigler, S., Gabriel, F., Sagun, L., d'Ascoli, S., Biroli, G., Hongler, C., and Wyart, M. Scaling description of generalization with number of parameters in deep learning. *arXiv preprint arXiv:1901.01608*, 2019a.

TO DOUBLE DESCENT

Peak in Test Error
At interpolation threshold

"U" shaped curve in the underparametrized regime

Monotonous decrease in the overparametrized regime



See also

Belkin et al. 2019

Nakkiran et al. 2019

Hastie et al. 2019

Mei & Montanari 2019

Kini et al. 2020

Gerace et al. 2020

Ba et al. 2020

....

Geiger, M., Jacot, A., Spigler, S., Gabriel, F., Sagun, L., d'Ascoli, S., Biroli, G., Hongler, C., and Wyart, M. Scaling description of generalization with number of parameters in deep learning. *arXiv preprint arXiv:1901.01608*, 2019a.

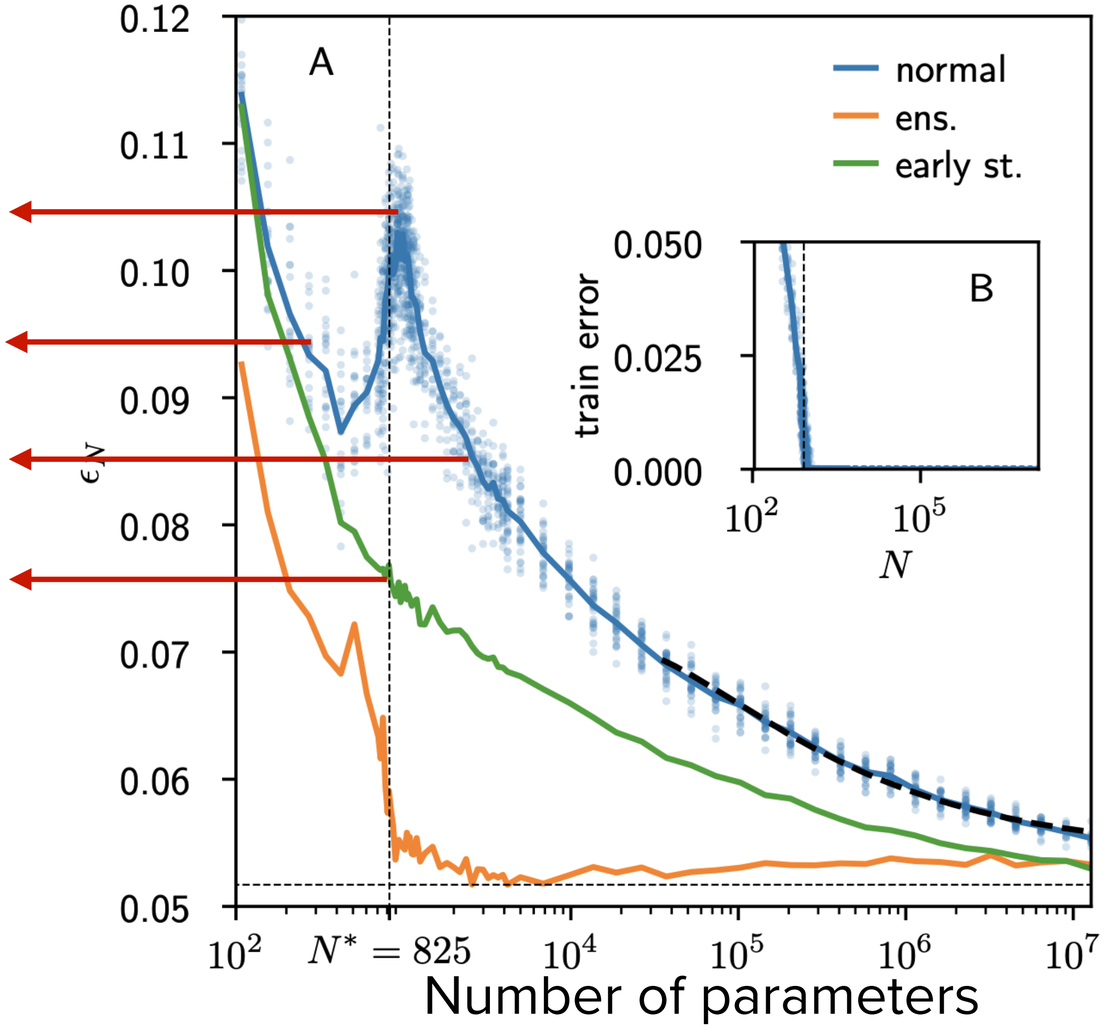
TO DOUBLE DESCENT

Peak in Test Error
At interpolation threshold

"U" shaped curve in the underparametrized regime

Monotonous decrease in the overparametrized regime

Peak suppressed by **regularisation**



See also

Belkin et al. 2019

Nakkiran et al. 2019

Hastie et al. 2019

Mei & Montanari 2019

Kini et al. 2020

Gerace et al. 2020

Ba et al. 2020

....

Geiger, M., Jacot, A., Spigler, S., Gabriel, F., Sagun, L., d'Ascoli, S., Biroli, G., Hongler, C., and Wyart, M. Scaling description of generalization with number of parameters in deep learning. *arXiv preprint arXiv:1901.01608*, 2019a.

TO DOUBLE DESCENT

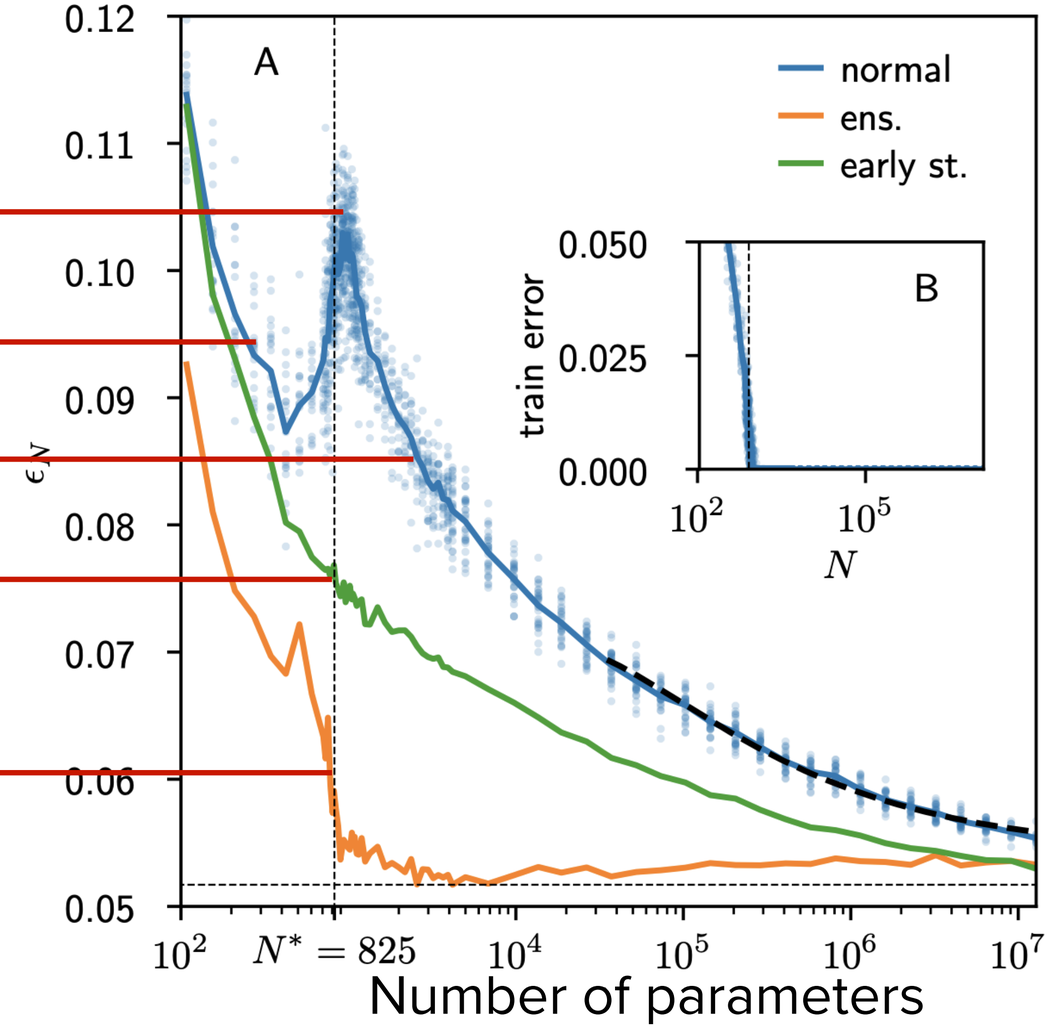
Peak in Test Error
At interpolation threshold

"U" shaped curve in the underparametrized regime

Monotonous decrease in the overparametrized regime

Peak suppressed by **regularisation**

Peak suppressed by **ensembling**



See also

Belkin et al. 2019

Nakkiran et al. 2019

Hastie et al. 2019

Mei & Montanari 2019

Kini et al. 2020

Gerace et al. 2020

Ba et al. 2020

....

Geiger, M., Jacot, A., Spigler, S., Gabriel, F., Sagun, L., d'Ascoli, S., Biroli, G., Hongler, C., and Wyart, M. Scaling description of generalization with number of parameters in deep learning. *arXiv preprint arXiv:1901.01608*, 2019a.

TO DOUBLE DESCENT

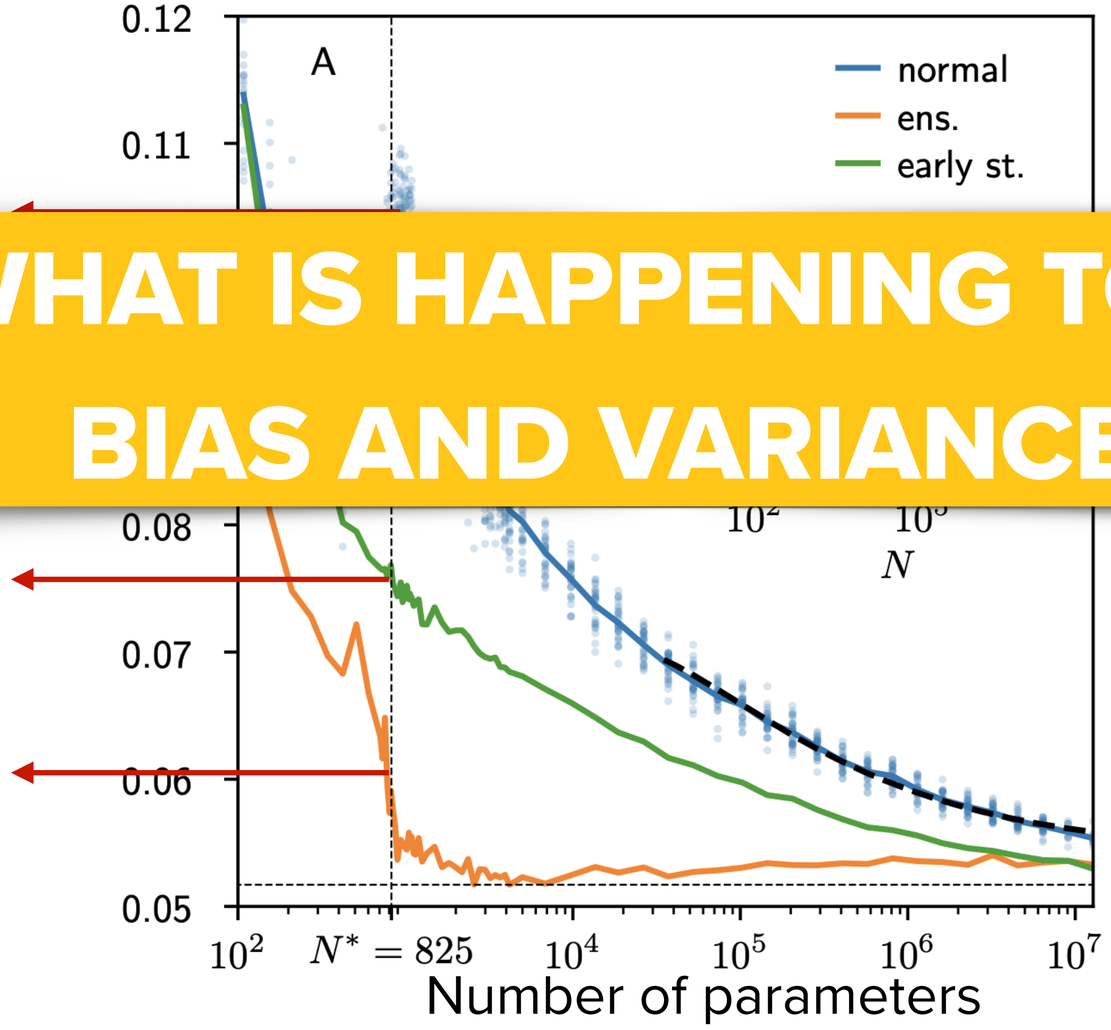
Peak in Test Error
At interpolation t

"U" shaped curv
underparametrize

Monotonous decreas
overparametrized re

Peak suppressed by
regularisation

Peak suppressed by
ensembling



WHAT IS HAPPENING TO THE BIAS AND VARIANCES ?

See also

Belkin et al. 2019

Nakkiran et al. 2019

Hastie et al. 2019

nari 2019

20

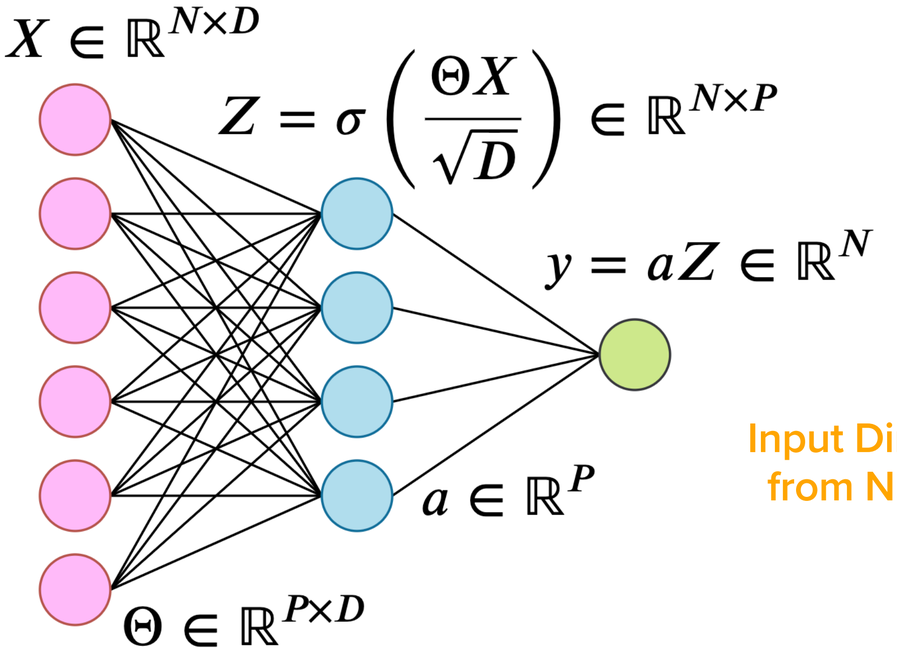
2020

Ba et al. 2020

....

Geiger, M., Jacot, A., Spigler, S., Gabriel, F., Sagun, L., d'Ascoli, S., Biroli, G., Hongler, C., and Wyart, M. Scaling description of generalization with number of parameters in deep learning. *arXiv preprint arXiv:1901.01608*, 2019a.

RANDOM FEATURES: WHY?



First introduced as **approximation** for **kernel methods [1]**

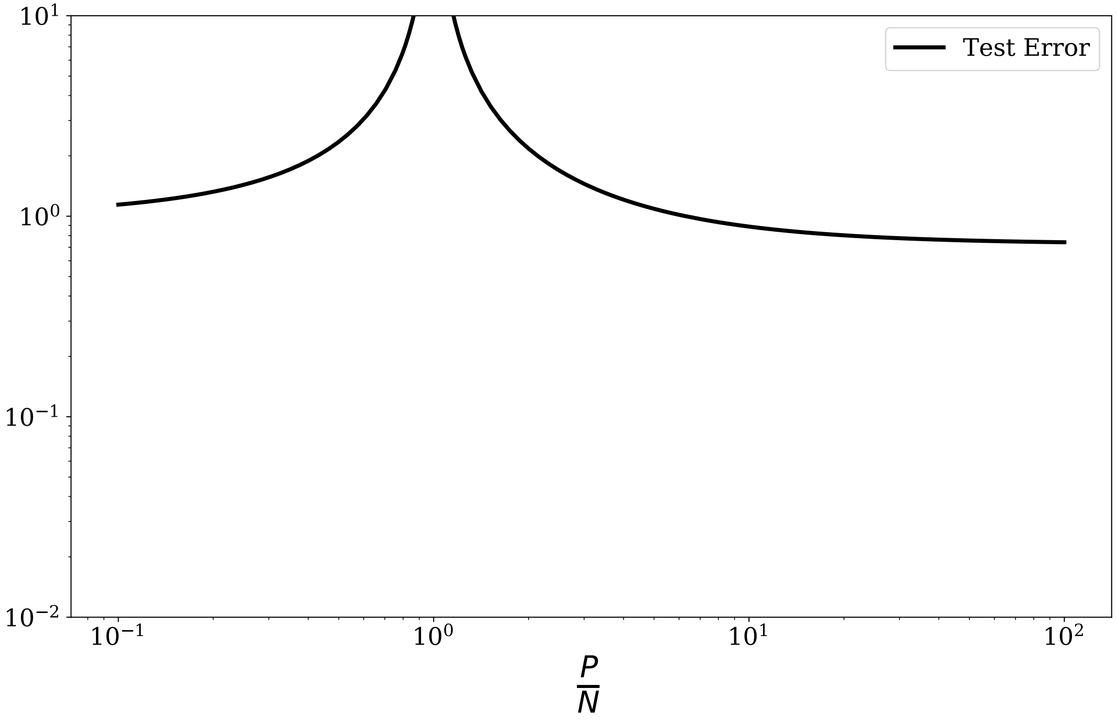
ANALYTICALLY TRACTABLE MODEL

- Input Dimension disentangled from Number of Parameters
- Double descent curve [2]
- Effect of Initialisation: Study Ensembling
- Relevant to the lazy regime of neural networks [3,4]

[1] Rahimi, Ali, and Benjamin Recht. "**Random features for large-scale kernel machines**." *Advances in neural information processing systems*. 2008.
[2] Mei, Song, and Andrea Montanari. "**The generalization error of random features regression: Precise asymptotics and double descent curve**." arXiv preprint arXiv:1908.05355 (2019).
[3] Arthur Jacot, Franck Gabriel, and Clément Hongler. **Neural tangent kernel: Convergence and generalization in neural networks**. In Advances in neural information processing systems, pages 8571–8580, 2018.
[4] Chizat, Lenaic, and Francis Bach. "**A note on lazy training in supervised differentiable programming**." arXiv preprint arXiv:1812.07956 8 (2018).

BIAS AND VARIANCE(S)

Analytical decomposition of the test error in random features regression



Analytical results for Random Feature Networks

BIAS AND VARIANCE(S)

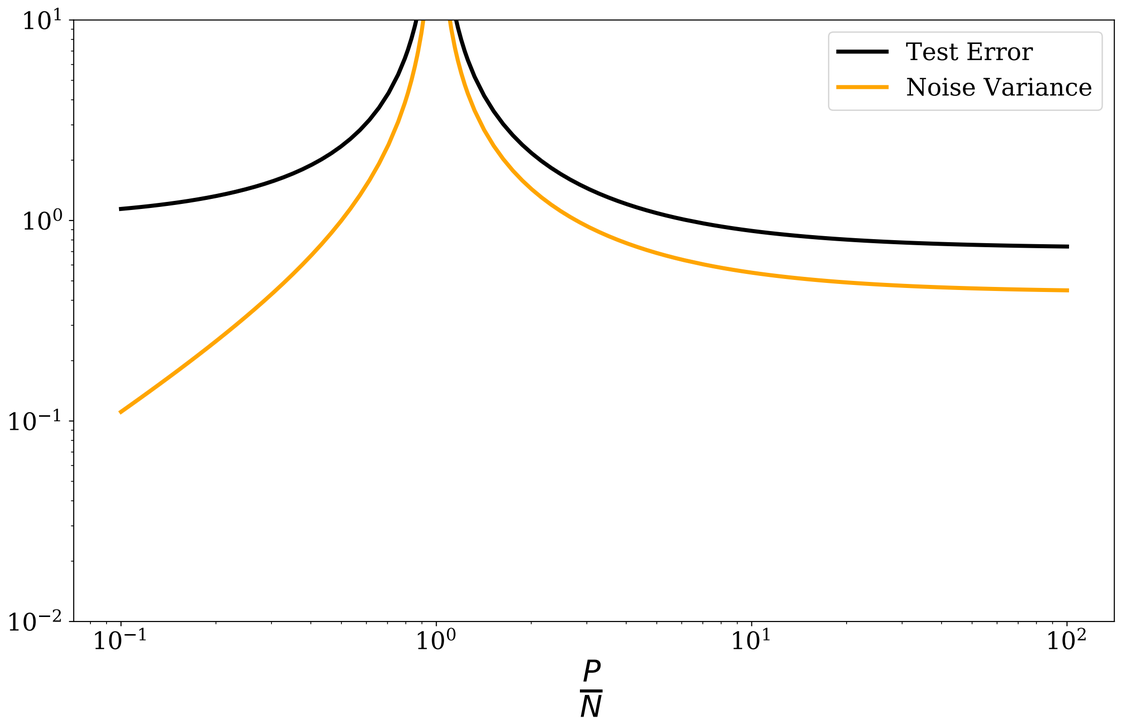
Analytical decomposition of the test error in random features regression

SOURCES OF RANDOMNESS

Label Noise

TEST ERROR

Noise Variance



Analytical results for Random Feature Networks

BIAS AND VARIANCE(S)

Analytical decomposition of the test error in random features regression

SOURCES OF RANDOMNESS

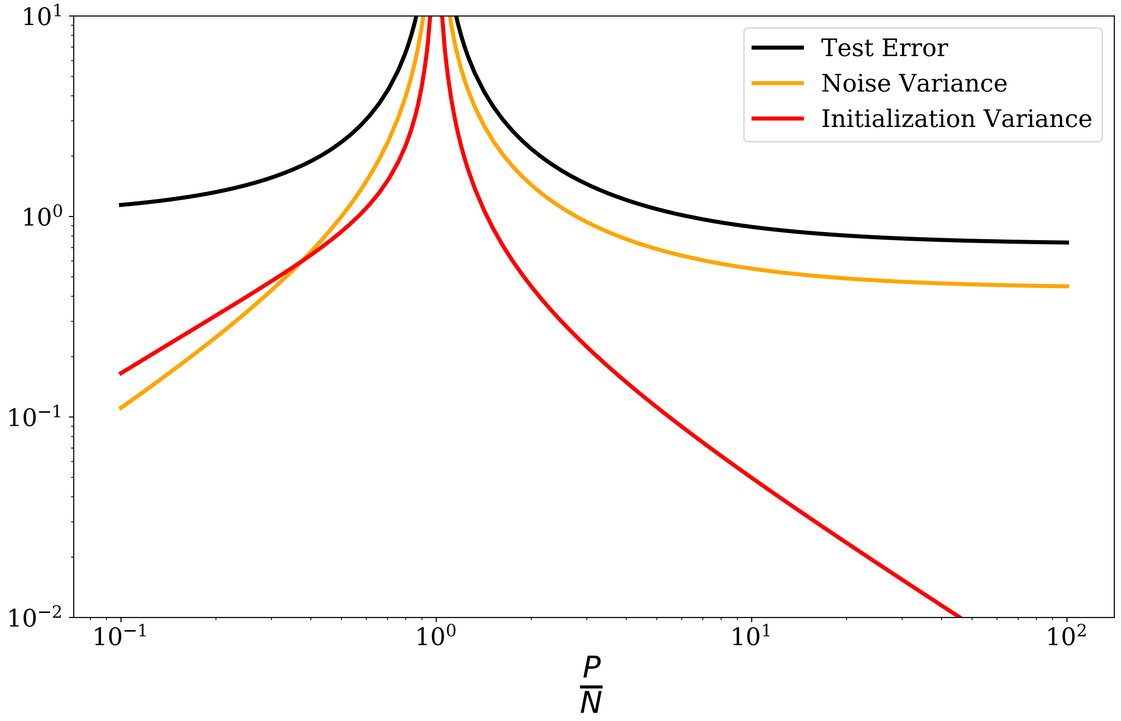
Label Noise

Features Initialisation

TEST ERROR

Noise Variance

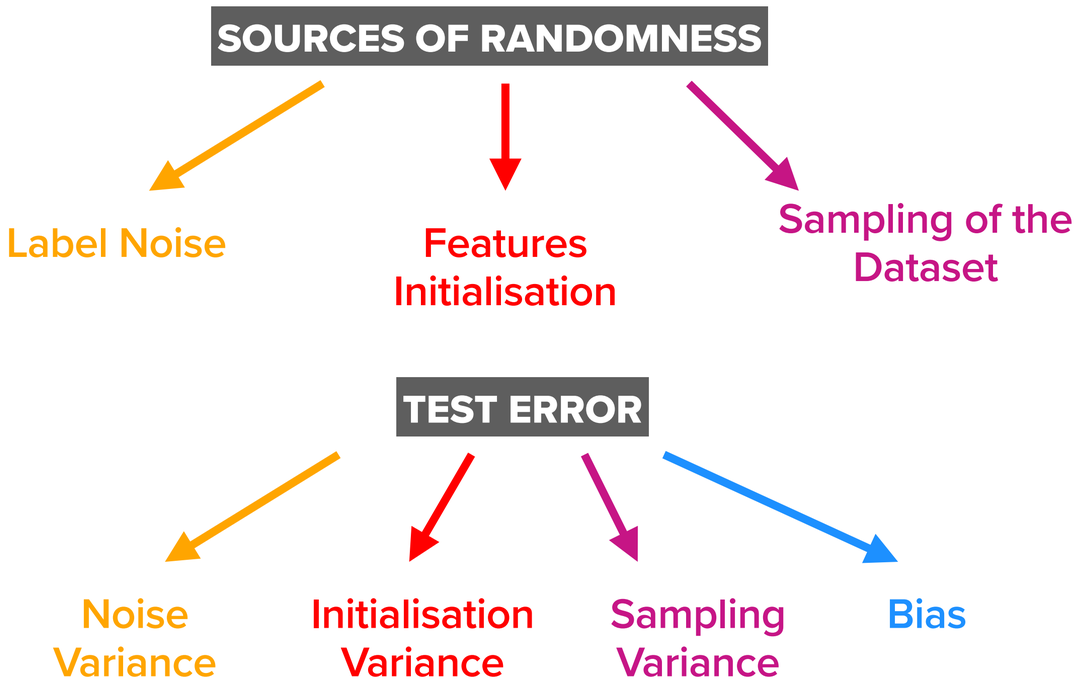
Initialisation Variance

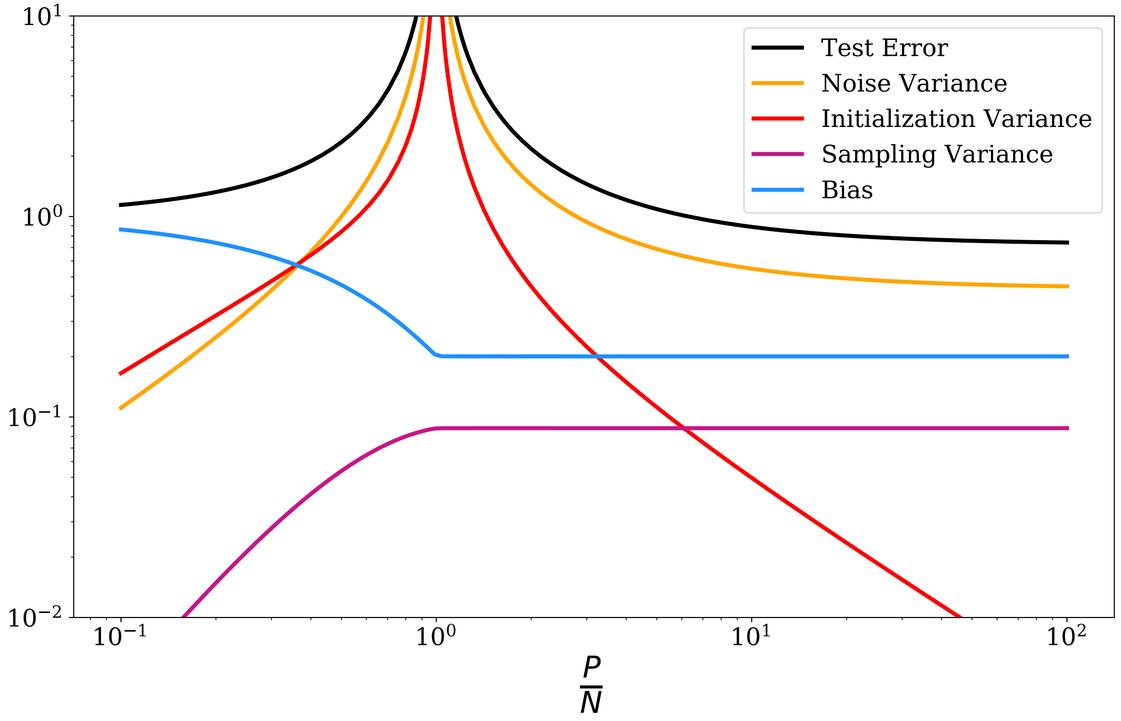


Analytical results for Random Feature Networks

BIAS AND VARIANCE(S)

Analytical decomposition of the test error in random features regression

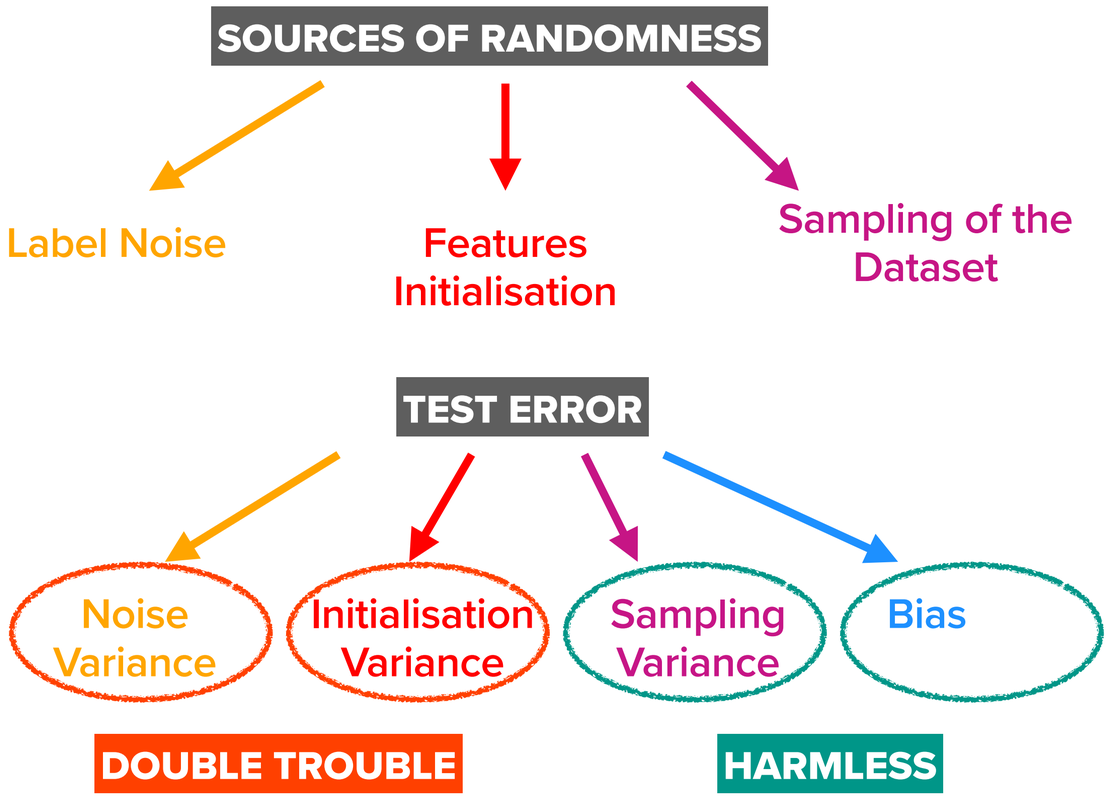


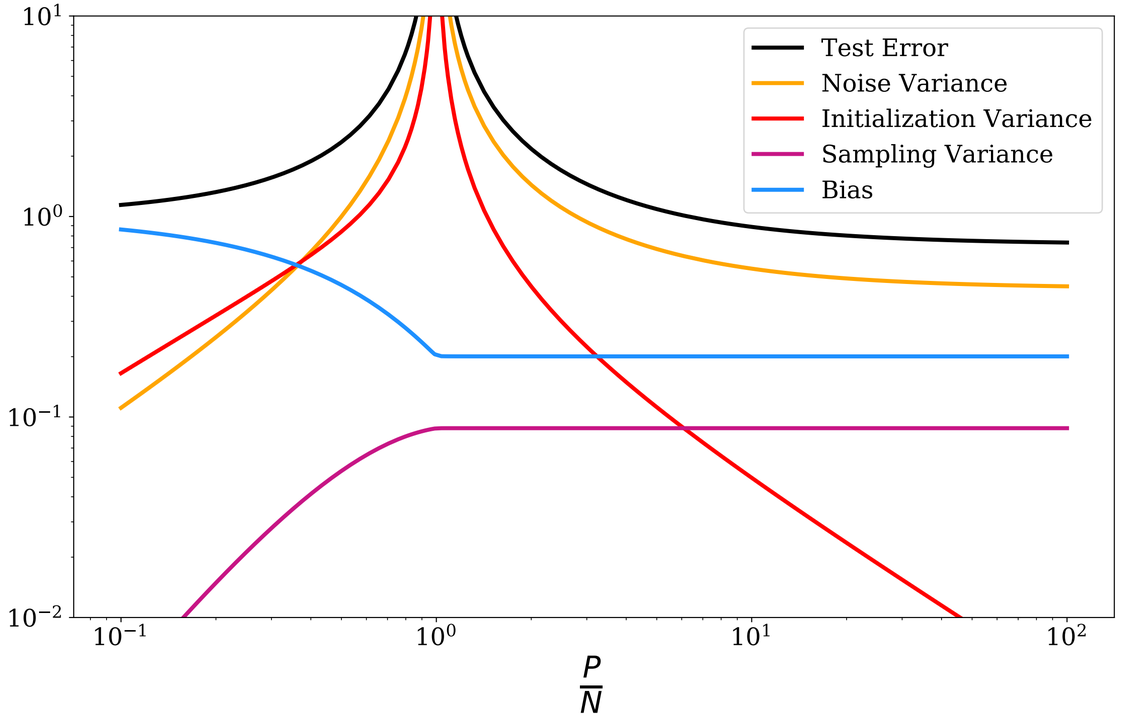


Analytical results for Random Feature Networks

BIAS AND VARIANCE(S)

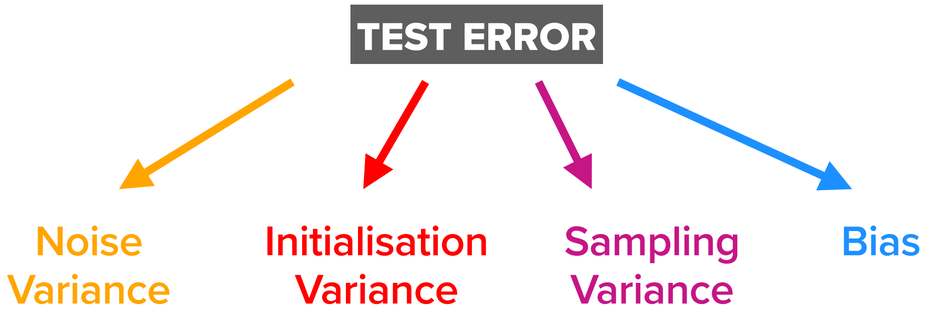
Analytical decomposition of the test error in random features regression

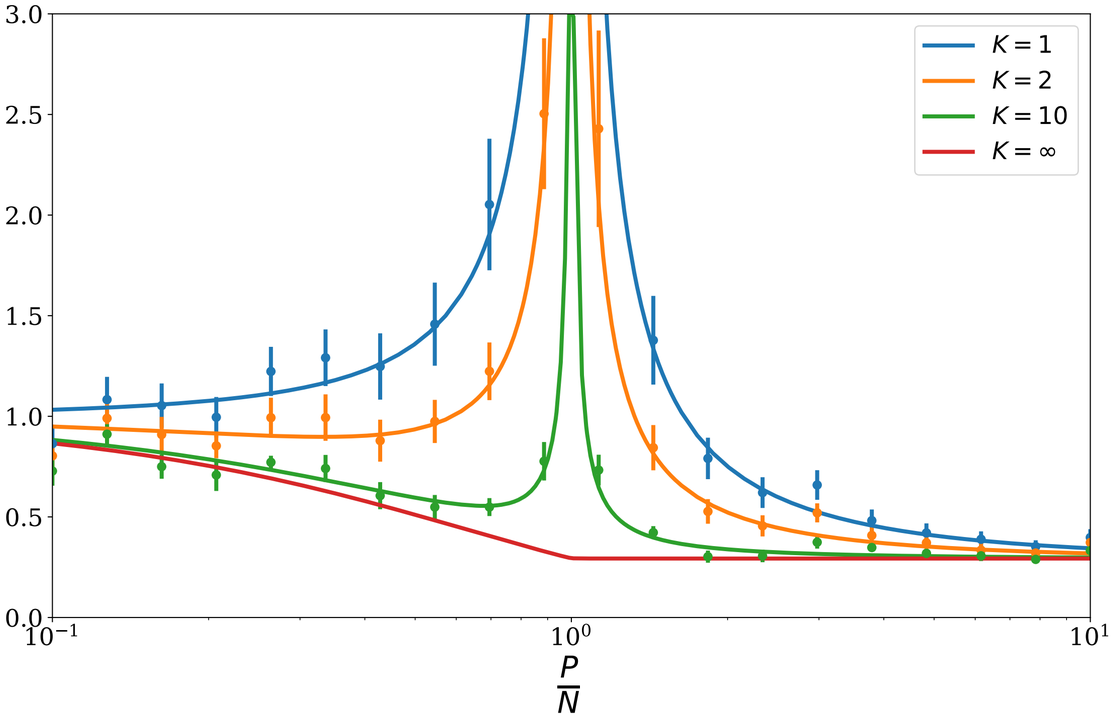




Analytical results for Random Feature Networks

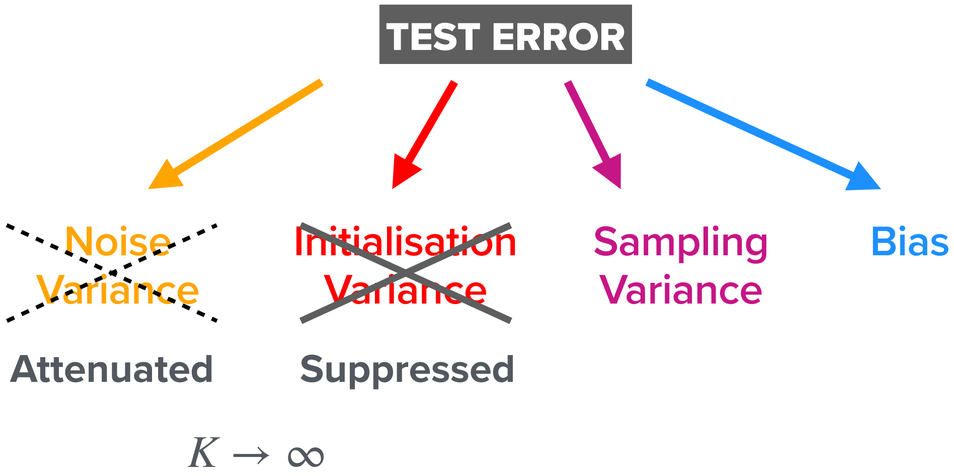
ENSEMBLING

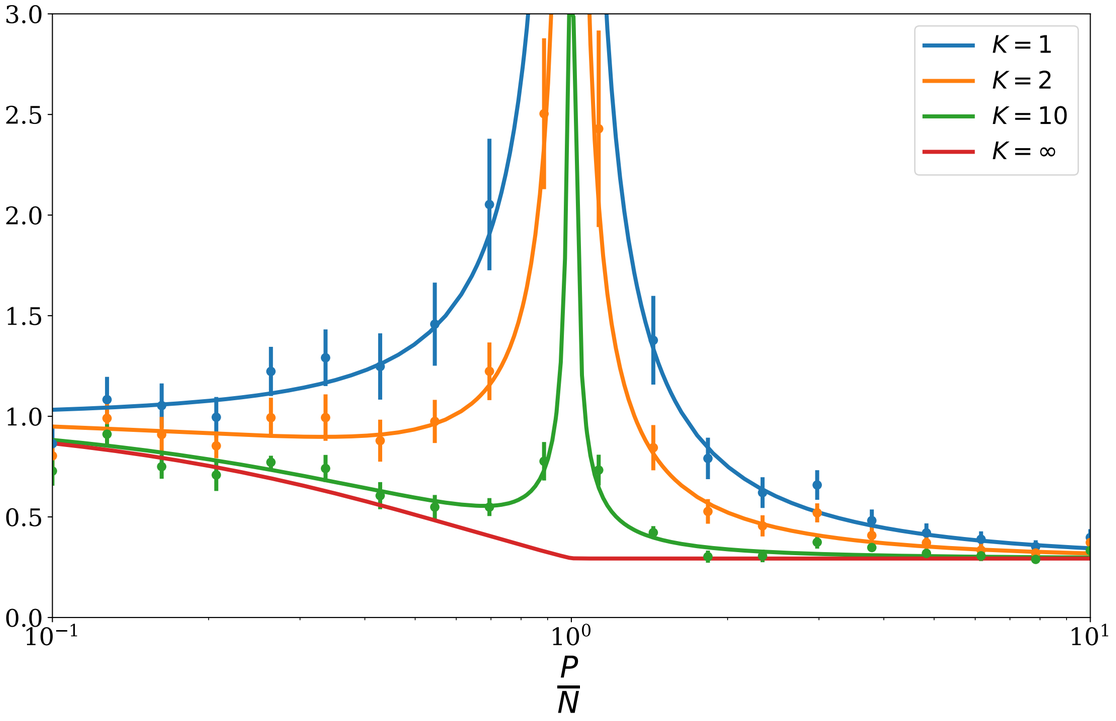




Analytical results for Random Feature Networks

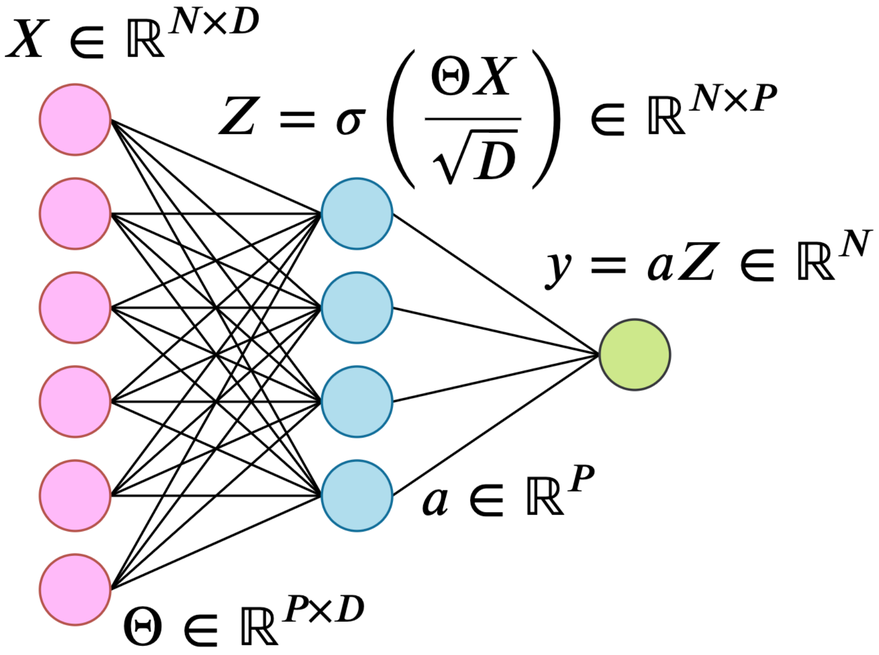
ENSEMBLING





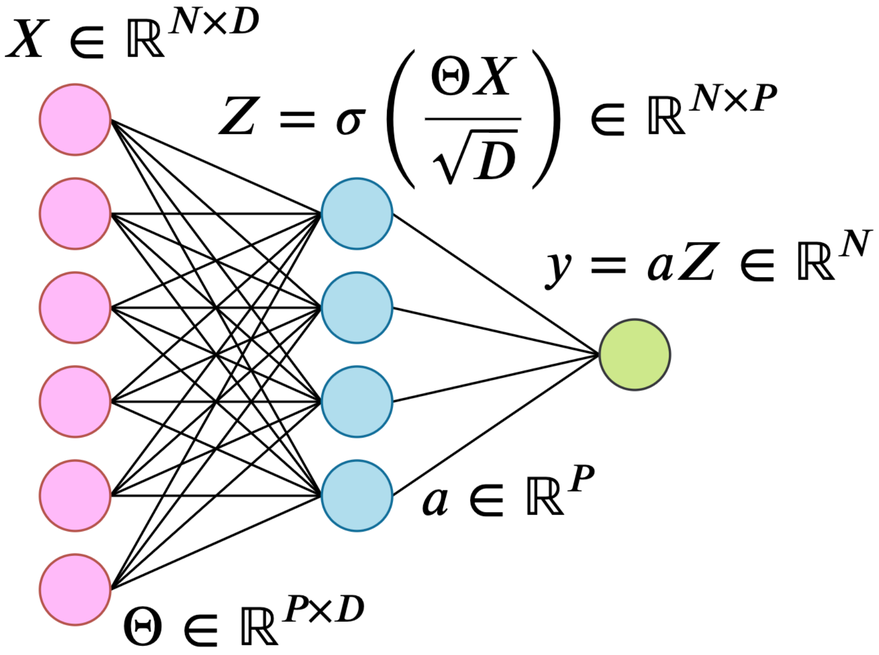
Analytical results for Random Feature Networks

THE RANDOM FEATURES MODEL



$\sigma = \text{ReLU}$ $\qquad$ $X, \boldsymbol{\Theta}, \boldsymbol{\beta} \sim \mathcal{N}(0,1)$

THE RANDOM FEATURES MODEL



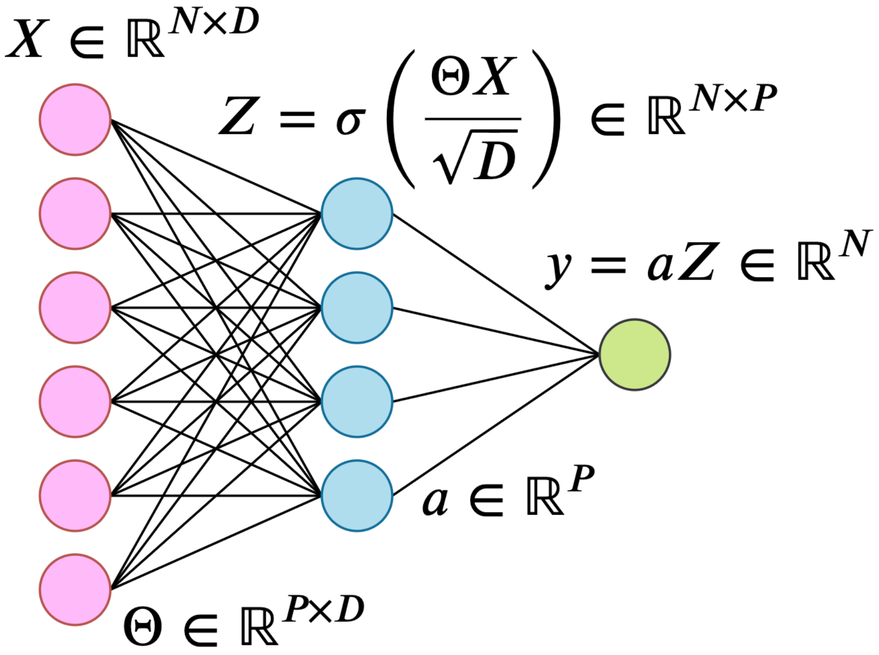
LEARNER: ONE HIDDEN LAYER

$$\hat{f}(X_\mu) = \sum_{i=1}^{P} a_i \sigma\left(\frac{\langle \Theta_i, X_\mu \rangle}{\sqrt{D}}\right)$$

$\sigma = \text{ReLU}$

$X, \Theta, \beta \sim \mathcal{N}(0,1)$

THE RANDOM FEATURES MODEL



LEARNER: ONE HIDDEN LAYER

$$\hat{f}(X_\mu) = \sum_{i=1}^{P} a_i \sigma\left(\frac{\langle \Theta_i, X_\mu \rangle}{\sqrt{D}}\right)$$

GROUND TRUTH: LINEAR WITH NOISE

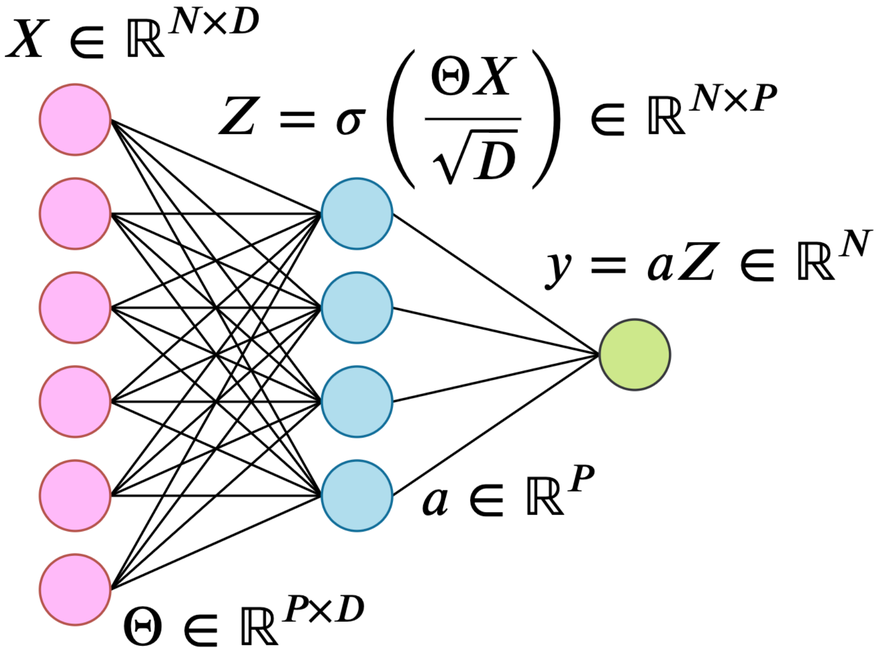
$$y_\mu = f(X_\mu) = \langle \boldsymbol{\beta}, X_\mu \rangle + \epsilon_\mu$$

$$\|\boldsymbol{\beta}\| = F, \quad \epsilon_\mu \sim \mathcal{N}(0,\tau)$$

$$\mathrm{SNR} = F/\tau$$

$\sigma = \mathrm{ReLU}$ $\qquad$ $X, \Theta, \boldsymbol{\beta} \sim \mathcal{N}(0,1)$

THE RANDOM FEATURES MODEL



LEARNER: ONE HIDDEN LAYER

$$\hat{f}(\boldsymbol{X}_\mu) = \sum_{i=1}^{P} a_i \sigma\left(\frac{\left\langle \boldsymbol{\Theta}_i, \boldsymbol{X}_\mu \right\rangle}{\sqrt{D}}\right)$$

GROUND TRUTH: LINEAR WITH NOISE

$$y_\mu = f(\boldsymbol{X}_\mu) = \left\langle \boldsymbol{\beta}, \boldsymbol{X}_\mu \right\rangle + \epsilon_\mu$$

$$\|\boldsymbol{\beta}\| = F, \quad \epsilon_\mu \sim \mathcal{N}(0,\tau)$$

$$\mathrm{SNR} = F/\tau$$

TRAIN ERROR

$$\mathscr{L}_{\mathrm{RF}}(\boldsymbol{a}) \equiv \frac{1}{N}\sum_{\mu=1}^{N}\left(y_\mu - \hat{f}(\boldsymbol{X}_\mu)\right)^2 + \frac{P\lambda}{D}\|\boldsymbol{a}\|_2^2$$

$$\hat{\boldsymbol{a}} \equiv \arg\min_{\boldsymbol{a}\in\mathbb{R}^P} \mathscr{L}_{\mathrm{RF}}(\boldsymbol{a})$$

$\sigma = \mathrm{ReLU}$ $\qquad$ $\boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{\beta} \sim \mathcal{N}(0,1)$

THE RANDOM FEATURES MODEL

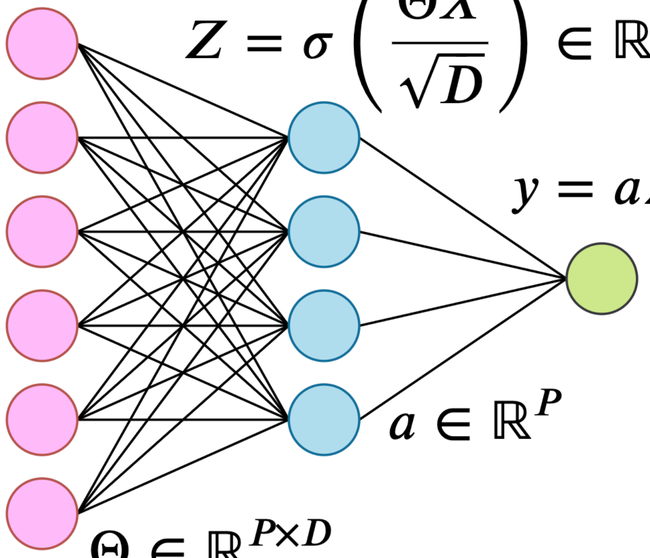
$X \in \mathbb{R}^{N \times D}$

$Z = \sigma\left(\frac{\Theta X}{\sqrt{D}}\right) \in \mathbb{R}^{N \times P}$

$y = aZ \in \mathbb{R}^{N}$

$a \in \mathbb{R}^{P}$

$\Theta \in \mathbb{R}^{P \times D}$



$\sigma = \mathrm{ReLU}$

$X, \Theta, \beta \sim \mathcal{N}(0,1)$

LEARNER: ONE HIDDEN LAYER

$$\hat{f}(X_\mu) = \sum_{i=1}^{P} a_i \sigma\left(\frac{\langle \Theta_i, X_\mu \rangle}{\sqrt{D}}\right)$$

GROUND TRUTH: LINEAR WITH NOISE

$$y_\mu = f(X_\mu) = \langle \beta, X_\mu \rangle + \epsilon_\mu$$

$$\|\beta\| = F, \quad \epsilon_\mu \sim \mathcal{N}(0,\tau)$$

$$\mathrm{SNR} = F/\tau$$

TRAIN ERROR

$$\mathcal{L}_{\mathrm{RF}}(a) \equiv \frac{1}{N} \sum_{\mu=1}^{N} \left(y_\mu - \hat{f}(X_\mu)\right)^2 + \frac{P\lambda}{D} \|a\|_2^2$$

$$\hat{a} \equiv \arg\min_{a \in \mathbb{R}^P} \mathcal{L}_{\mathrm{RF}}(a)$$

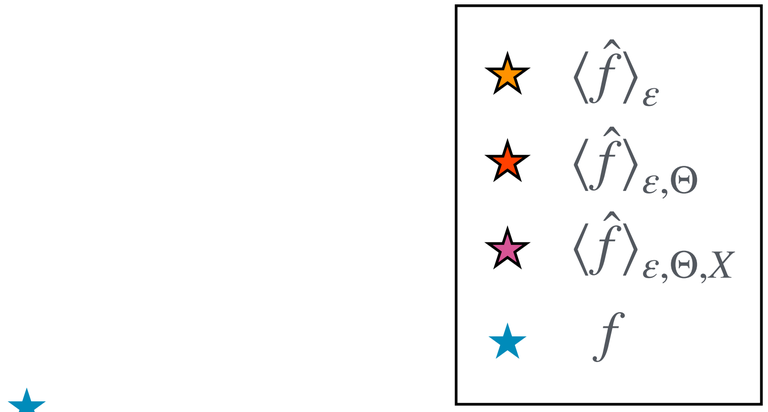
TEST ERROR

$$\mathcal{R}_{\mathrm{RF}} = \mathbb{E}_x \left[\left(f(x) - \hat{f}(x)\right)^2\right]$$

BIAS AND VARIANCE(S)

$$\mathscr{R}_{\mathrm{RF}} = \mathbb{E}\left[\left(f - \hat{f}\right)^2\right]$$

$$= F^2 + 2\mathbb{E}\left[f\hat{f}\right] + \mathbb{E}\left[\hat{f}^2\right]$$

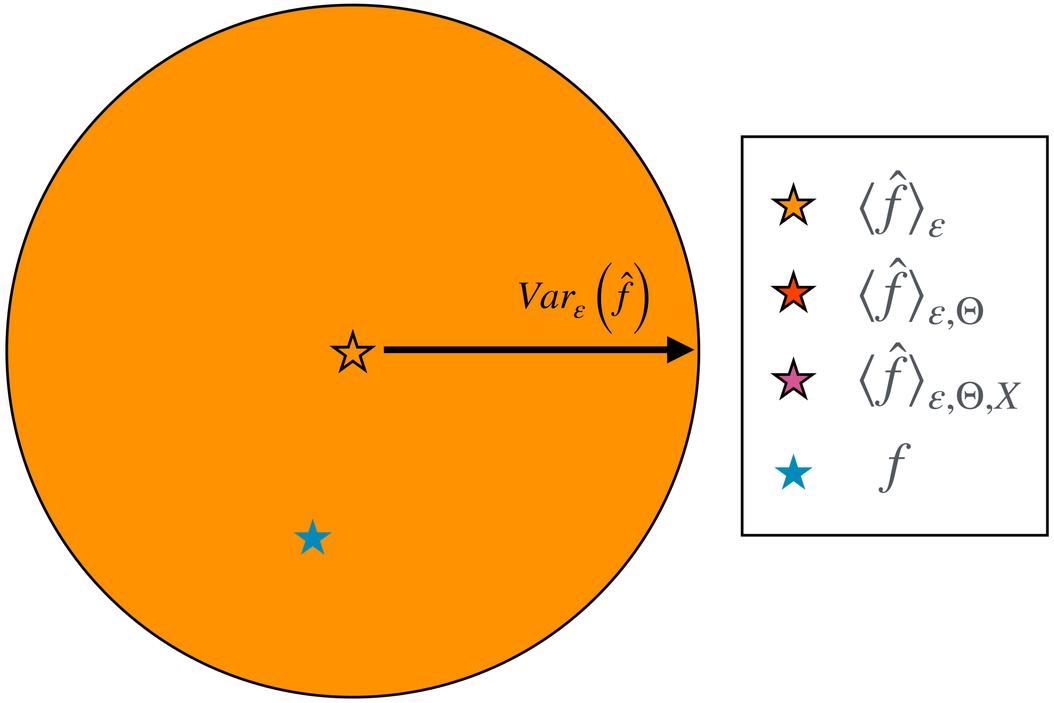


BIAS AND VARIANCE(S)

$$\mathscr{R}_{\mathrm{RF}} = \mathbb{E}\left[\left(f - \hat{f}\right)^2\right]$$

$$= F^2 + 2\mathbb{E}\left[f\hat{f}\right] + \mathbb{E}\left[\hat{f}^2\right]$$

$$= F^2 + 2\mathbb{E}\left[f\hat{f}\right] + \mathbb{E}\left[Var_\varepsilon\left(\hat{f}\right) + \left\langle \hat{f} \right\rangle_\varepsilon^2\right]$$



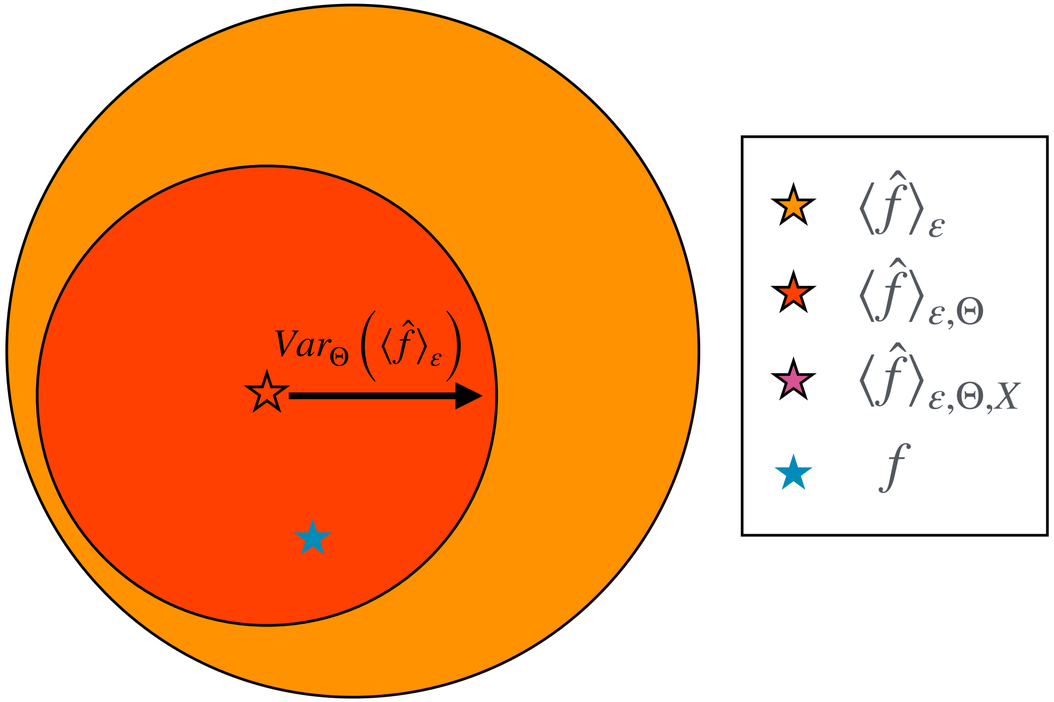
BIAS AND VARIANCE(S)

$$\mathscr{R}_{\mathrm{RF}} = \mathbb{E}\left[\left(f - \hat{f}\right)^2\right]$$

$$= F^2 + 2\mathbb{E}\left[f\hat{f}\right] + \mathbb{E}\left[\hat{f}^2\right]$$

$$= F^2 + 2\mathbb{E}\left[f\hat{f}\right] + \mathbb{E}\left[Var_{\varepsilon}\left(\hat{f}\right) + \left\langle \hat{f} \right\rangle_{\varepsilon}^2\right]$$

$$= F^2 + 2\mathbb{E}\left[f\hat{f}\right] + \mathbb{E}\left[Var_{\varepsilon}\left(\hat{f}\right) + Var_{\Theta}\left(\langle \hat{f} \rangle_{\varepsilon}\right) + \left\langle \hat{f} \right\rangle_{\varepsilon,\Theta}^2\right]$$



BIAS AND VARIANCE(S)

$$\mathscr{R}_{\mathrm{RF}} = \mathbb{E}\left[\left(f - \hat{f}\right)^2\right]$$

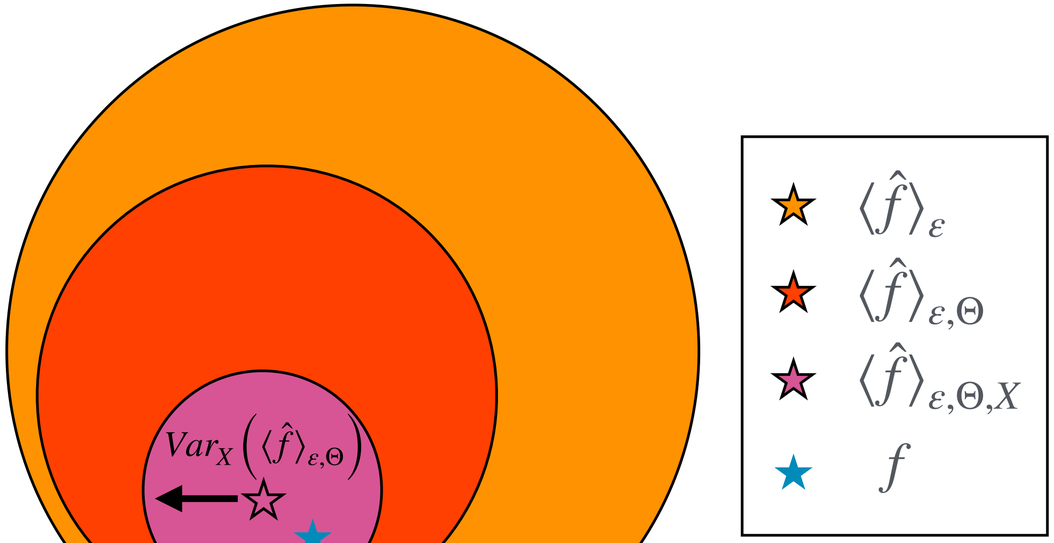
$$= F^2 + 2\mathbb{E}\left[f\hat{f}\right] + \mathbb{E}\left[\hat{f}^2\right]$$

$$= F^2 + 2\mathbb{E}\left[f\hat{f}\right] + \mathbb{E}\left[Var_{\varepsilon}\left(\hat{f}\right) + \left\langle \hat{f} \right\rangle_{\varepsilon}^2\right]$$

$$= F^2 + 2\mathbb{E}\left[f\hat{f}\right] + \mathbb{E}\left[Var_{\varepsilon}\left(\hat{f}\right) + Var_{\Theta}\left(\langle \hat{f} \rangle_{\varepsilon}\right) + \left\langle \hat{f} \right\rangle_{\varepsilon,\Theta}^2\right]$$

$$= F^2 + 2\mathbb{E}\left[f\hat{f}\right] + \mathbb{E}\left[Var_{\varepsilon}\left(\hat{f}\right) + Var_{\Theta}\left(\langle \hat{f} \rangle_{\varepsilon}\right) + Var_{X}\left(\langle \hat{f} \rangle_{\varepsilon,\Theta}\right) + \left\langle \hat{f} \right\rangle_{\varepsilon,\Theta,X}^2\right]$$

Noise **Initialization** **Sampling**



ANALYTIC THEORY

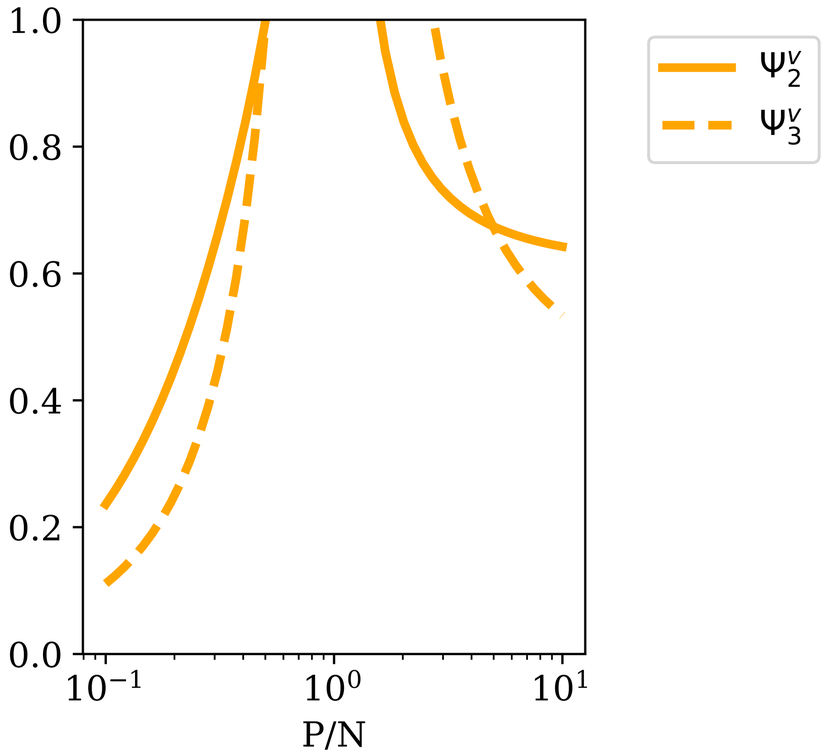
We provide an asymptotic formula in the high-dimensional limit $N, P, D \to \infty, \quad \frac{P}{D} = \mathcal{O}(1), \quad \frac{N}{D} = \mathcal{O}(1)$

- **Vanilla** terms
  mean & fluct of vanilla learner:
  both **diverge** at the interpolation threshold

$$\begin{cases} \left\langle \hat{f} \right\rangle_{\varepsilon}^{2} = F^2 \Psi_2^v \\ Var_{\varepsilon} \hat{f} = \tau^2 \Psi_3^v \end{cases}$$

- **Ensembling** terms
  mean & fluct of init-averaged learner:
  both are **finite** at the interpolation threshold

- **Divide and conquer** terms
  mean & fluct of init+samp-averaged learner:
  no more fluctuations



[1] Mei, Song, and Andrea Montanari. "**The generalization error of random features regression: Precise asymptotics and double descent curve**." arXiv preprint arXiv:1908.05355 (2019).

ANALYTIC THEORY

We provide an asymptotic formula in the high-dimensional limit $N, P, D \to \infty, \quad \frac{P}{D} = \mathcal{O}(1), \quad \frac{N}{D} = \mathcal{O}(1)$

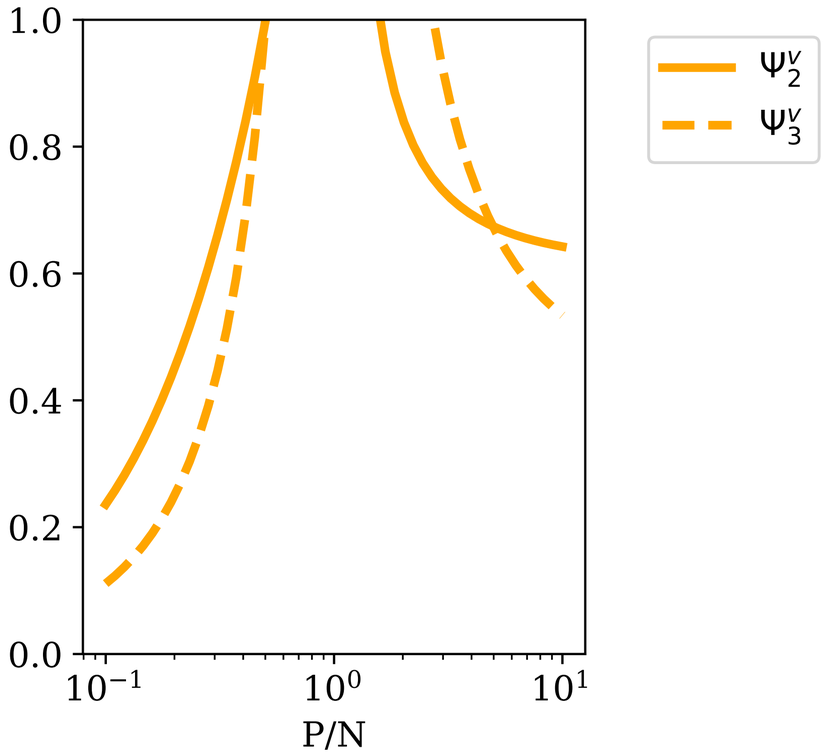
- **Vanilla** terms
  mean & fluct of vanilla learner:
  both **diverge** at the interpolation threshold

- **Ensembling** terms
  mean & fluct of init-averaged learner:
  both are **finite** at the interpolation threshold

- **Divide and conquer** terms
  mean & fluct of init+samp-averaged learner:
  no more fluctuations

CALCULATED IN [1]

$$\begin{cases} \left\langle \hat{f} \right\rangle_{\varepsilon}^{2} = F^2 \Psi_2^v \\ Var_{\varepsilon} \hat{f} = \tau^2 \Psi_3^v \end{cases}$$



[1] Mei, Song, and Andrea Montanari. "**The generalization error of random features regression: Precise asymptotics and double descent curve**." arXiv preprint arXiv:1908.05355 (2019).

ANALYTIC THEORY

We provide an asymptotic formula in the high-dimensional limit $N, P, D \to \infty, \quad \frac{P}{D} = \mathcal{O}(1), \quad \frac{N}{D} = \mathcal{O}(1)$

CALCULATED IN [1]

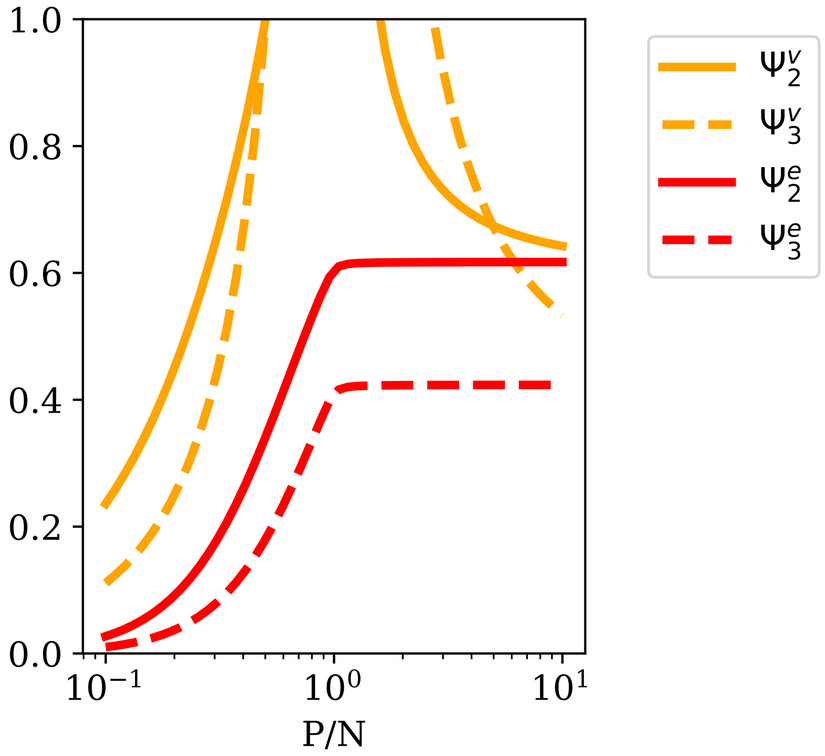
$$\begin{cases} \left\langle \hat{f} \right\rangle_{\varepsilon}^{2} = F^2 \Psi_2^v \\ Var_{\varepsilon} \hat{f} = \tau^2 \Psi_3^v \end{cases}$$

- **Vanilla** terms
  mean & fluct of vanilla learner:
  both **diverge** at the interpolation threshold

$$\begin{cases} \left\langle \hat{f} \right\rangle_{\varepsilon,\Theta}^{2} = F^2 \Psi_2^e \\ Var_{\varepsilon,\Theta} \hat{f} = \tau^2 \Psi_3^e \end{cases}$$

- **Ensembling** terms
  mean & fluct of init-averaged learner:
  both are **finite** at the interpolation threshold

- **Divide and conquer** terms
  mean & fluct of init+samp-averaged learner:
  no more fluctuations



[1] Mei, Song, and Andrea Montanari. "**The generalization error of random features regression: Precise asymptotics and double descent curve**." arXiv preprint arXiv:1908.05355 (2019).

ANALYTIC THEORY

We provide an asymptotic formula in the high-dimensional limit $N, P, D \to \infty, \quad \frac{P}{D} = \mathcal{O}(1), \quad \frac{N}{D} = \mathcal{O}(1)$

- **Vanilla** terms
  mean & fluct of vanilla learner:
  both **diverge** at the interpolation threshold

- **Ensembling** terms
  mean & fluct of init-averaged learner:
  both are **finite** at the interpolation threshold

- **Divide and conquer** terms
  mean & fluct of init+samp-averaged learner:
  no more fluctuations

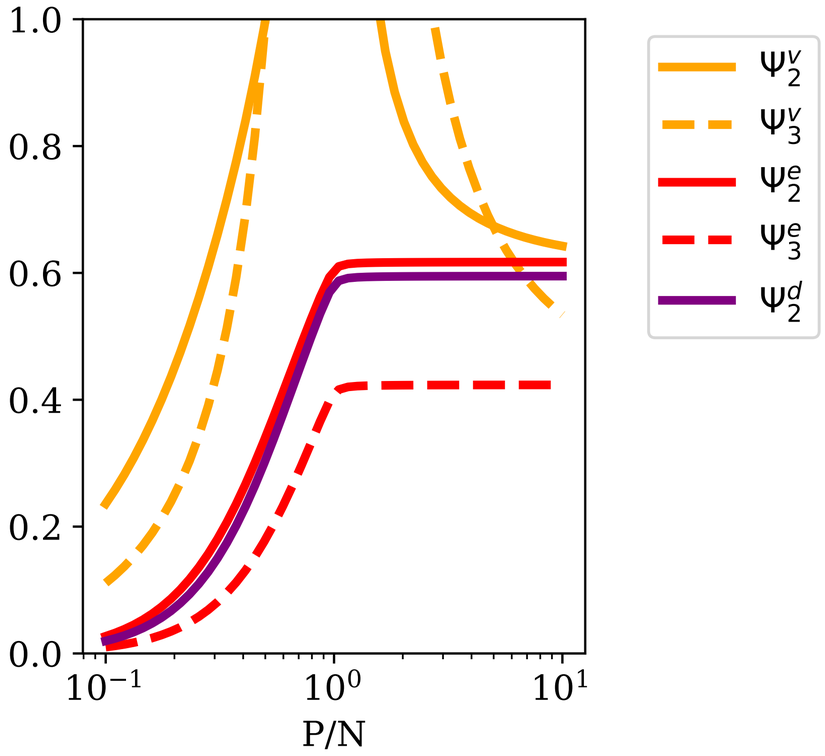
CALCULATED IN [1]

$$\begin{cases} \left\langle \hat{f} \right\rangle_{\varepsilon}^{2} = F^2 \Psi_2^v \\ Var_{\varepsilon} \hat{f} = \tau^2 \Psi_3^v \end{cases}$$

$$\begin{cases} \left\langle \hat{f} \right\rangle_{\varepsilon,\Theta}^{2} = F^2 \Psi_2^e \\ Var_{\varepsilon,\Theta} \hat{f} = \tau^2 \Psi_3^e \end{cases}$$

$$\begin{cases} \left\langle \hat{f} \right\rangle_{\varepsilon,\Theta,X}^{2} = F^2 \Psi_2^d \\ Var_{\varepsilon,\Theta,X} \hat{f} = 0 \end{cases}$$

NEW



[1] Mei, Song, and Andrea Montanari. "**The generalization error of random features regression: Precise asymptotics and double descent curve**." arXiv preprint arXiv:1908.05355 (2019).

BIAS AND VARIANCES

Decomposition of the test error $\mathop{\mathbb{E}}_{\boldsymbol{\Theta},X,\boldsymbol{\varepsilon}}\left[\mathscr{R}_{\mathrm{RF}}\right] = Var_{\mathrm{Noise}} + Var_{\mathrm{Init}} + Var_{\mathrm{Samp}} + Bias$

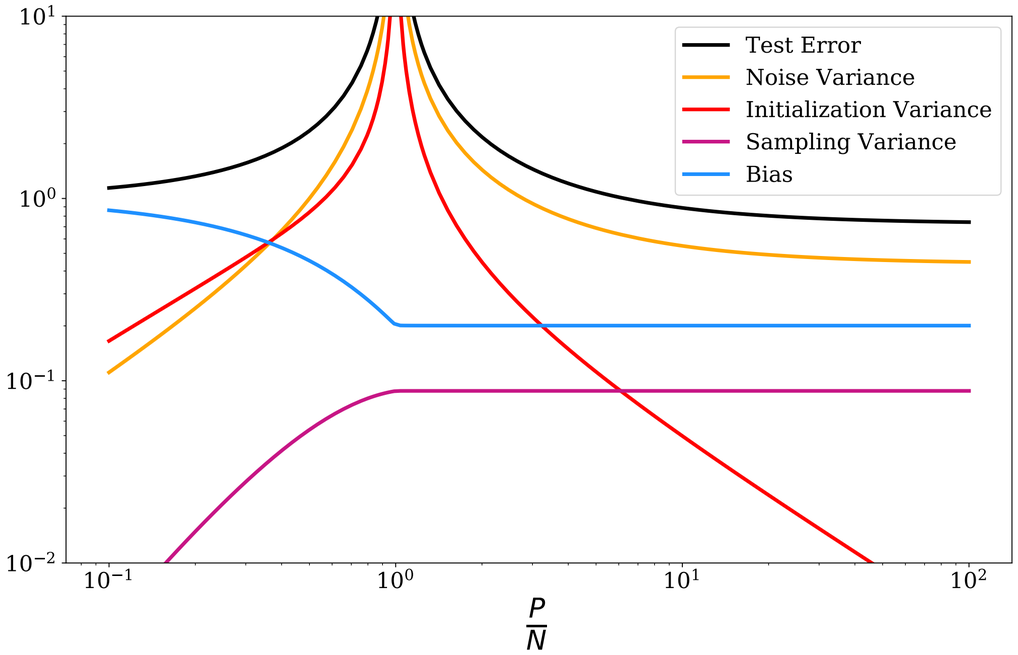
[1] Mei, Song, and Andrea Montanari. "**The generalization error of random features regression: Precise asymptotics and double descent curve**." arXiv preprint arXiv:1908.05355 (2019).

[2] Neal, Brady, et al. "**A modern take on the bias-variance tradeoff in neural networks**." arXiv preprint arXiv:1810.08591 (2018).

BIAS AND VARIANCES

Decomposition of the test error

$$\mathop{\mathbb{E}}_{\boldsymbol{\Theta},X,\boldsymbol{\varepsilon}}\left[\mathscr{R}_{\mathrm{RF}}\right] = Var_{\mathrm{Noise}} + Var_{\mathrm{Init}} + Var_{\mathrm{Samp}} + Bias$$

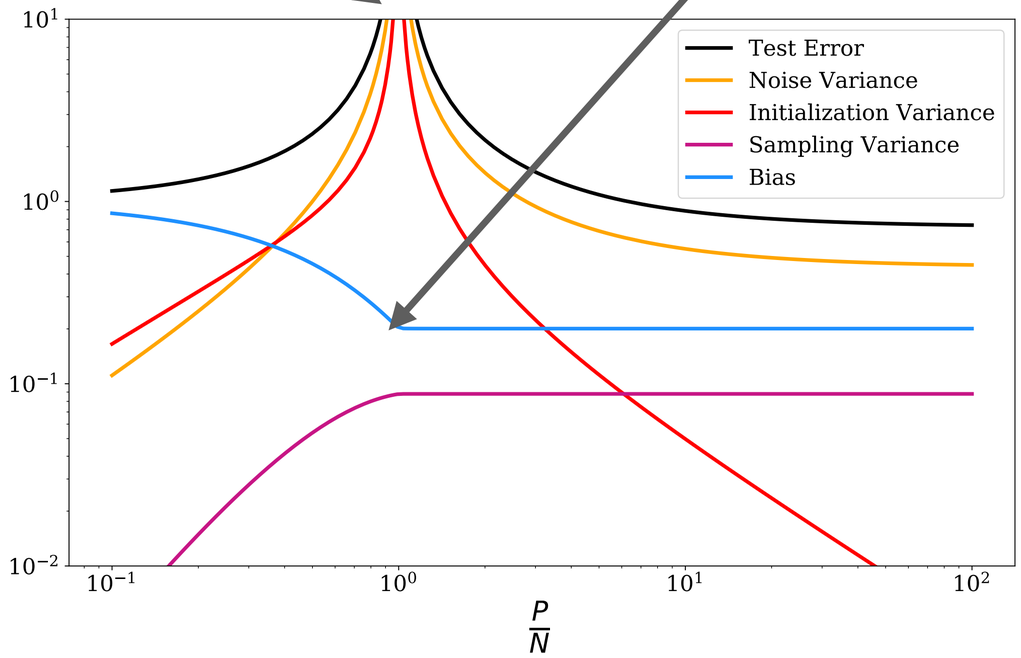


[1] Mei, Song, and Andrea Montanari. "**The generalization error of random features regression: Precise asymptotics and double descent curve**." arXiv preprint arXiv:1908.05355 (2019).
[2] Neal, Brady, et al. "**A modern take on the bias-variance tradeoff in neural networks**." arXiv preprint arXiv:1810.08591 (2018).

BIAS AND VARIANCES

Decomposition of the test error

$$\mathop{\mathbb{E}}_{\boldsymbol{\Theta},X,\boldsymbol{\varepsilon}}\left[\mathscr{R}_{\mathrm{RF}}\right] = \underbrace{Var_{\mathrm{Noise}} + Var_{\mathrm{Init}}}_{\text{DIVERGENCE}} + \underbrace{Var_{\mathrm{Samp}} + Bias}_{\text{PHASE TRANSITION}}$$



[1] Mei, Song, and Andrea Montanari. "**The generalization error of random features regression: Precise asymptotics and double descent curve**." arXiv preprint arXiv:1908.05355 (2019).

[2] Neal, Brady, et al. "**A modern take on the bias-variance tradeoff in neural networks**." arXiv preprint arXiv:1810.08591 (2018).

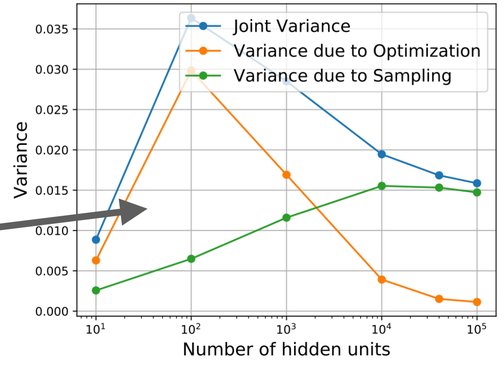
BIAS AND VARIANCES

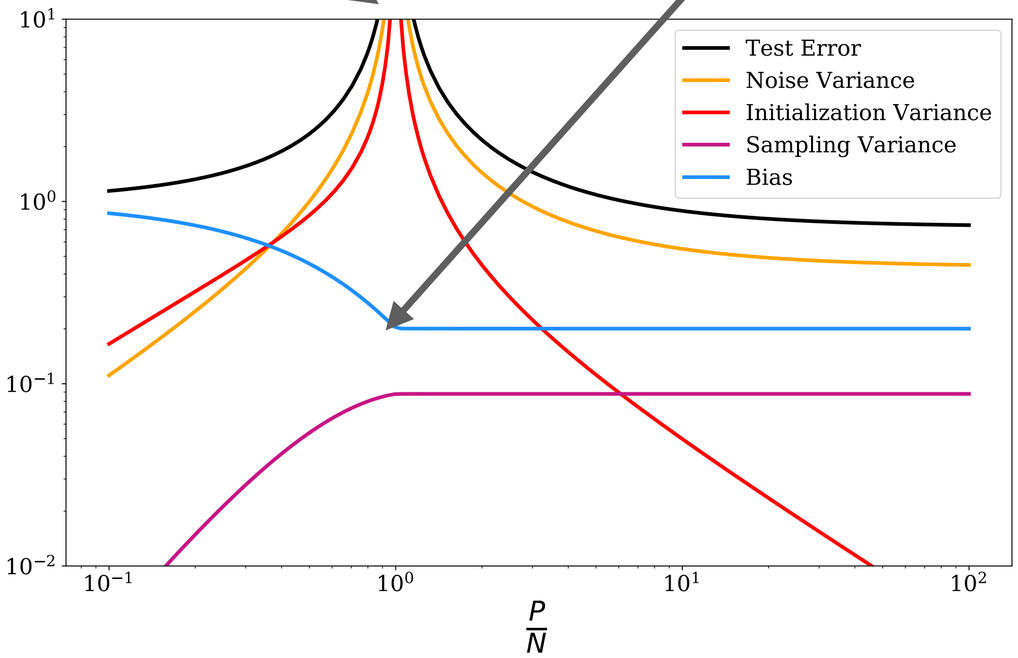
Decomposition of the test error

$$\mathop{\mathbb{E}}_{\boldsymbol{\Theta},X,\boldsymbol{\varepsilon}}\left[\mathscr{R}_{\mathrm{RF}}\right] = \underbrace{Var_{\mathrm{Noise}} + Var_{\mathrm{Init}}}_{\text{DIVERGENCE}} + \underbrace{Var_{\mathrm{Samp}} + Bias}_{\text{PHASE TRANSITION}}$$

CONSISTENT WITH EMPIRICAL STUDY IN NEURAL NETWORKS [2]

Init variance peaks,
Sampling variance plateaus



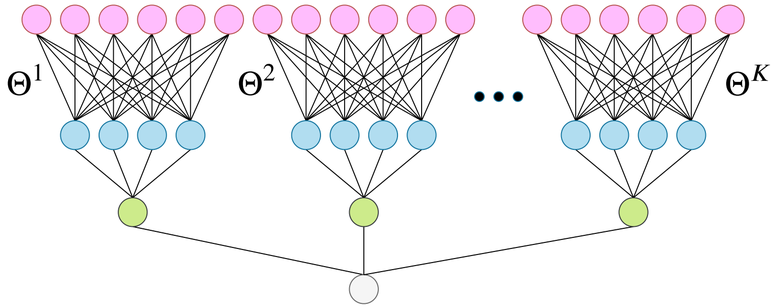


[1] Mei, Song, and Andrea Montanari. "**The generalization error of random features regression: Precise asymptotics and double descent curve**." arXiv preprint arXiv:1908.05355 (2019).
[2] Neal, Brady, et al. "**A modern take on the bias-variance tradeoff in neural networks**." arXiv preprint arXiv:1810.08591 (2018).

EFFECT OF ENSEMBLING

[1] Geiger, Mario, et al. "**Scaling description of generalization with number of parameters in deep learning**." arXiv preprint arXiv:1901.01608 (2019).

EFFECT OF ENSEMBLING



$$f(x) = \frac{1}{K} \sum_{k=1}^{K} f_{\Theta^k}(x)$$

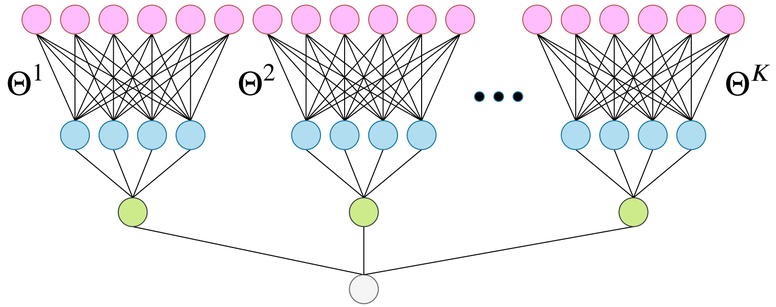
[1] Geiger, Mario, et al. "**Scaling description of generalization with number of parameters in deep learning**." arXiv preprint arXiv:1901.01608 (2019).

EFFECT OF ENSEMBLING

$$\mathop{\mathbb{E}}_{\Theta,X,\varepsilon}\left[\mathscr{R}_{\mathrm{RF}}\right] = F^2 + 2\mathbb{E}[f\hat{f}] + \frac{1}{K}\mathbb{E}\left[\hat{f}^2\right] + \left(1-\frac{1}{K}\right)\mathbb{E}\left[\left\langle\hat{f}\right\rangle_{\Theta}^2\right]$$

$$Var_{Noise} = \frac{1}{K}Var_{\varepsilon}\hat{f} + \left(1-\frac{1}{K}\right)Var_{\varepsilon}\left\langle\hat{f}\right\rangle_{\Theta}$$

$$Var_{Init} = \frac{1}{K}Var_{\Theta}\left\langle\hat{f}\right\rangle_{\varepsilon}$$



$$f(x) = \frac{1}{K}\sum_{k=1}^{K} f_{\Theta^k}(x)$$

[1] Geiger, Mario, et al. "**Scaling description of generalization with number of parameters in deep learning**." arXiv preprint arXiv:1901.01608 (2019).

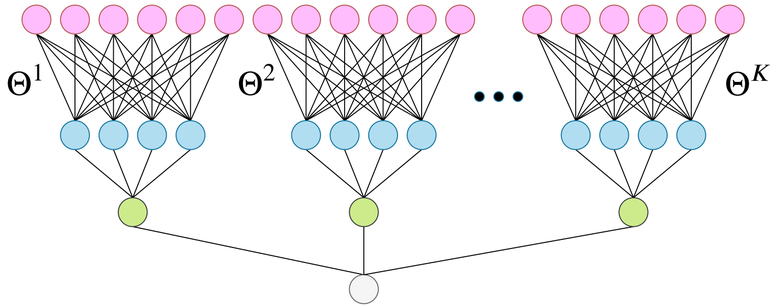
EFFECT OF ENSEMBLING

$$\mathop{\mathbb{E}}_{\Theta,X,\varepsilon}\left[\mathscr{R}_{\mathrm{RF}}\right] = F^2 + 2\mathbb{E}[f\hat{f}] + \frac{1}{K}\mathbb{E}\left[\hat{f}^2\right] + \left(1-\frac{1}{K}\right)\mathbb{E}\left[\left\langle\hat{f}\right\rangle_{\Theta}^2\right]$$

$$Var_{Noise} = \frac{1}{K}Var_{\varepsilon}\hat{f} + \left(1-\frac{1}{K}\right)Var_{\varepsilon}\left\langle\hat{f}\right\rangle_{\Theta}$$

$$Var_{Init} = \frac{1}{K}Var_{\Theta}\left\langle\hat{f}\right\rangle_{\varepsilon}$$

VANILLA TERMS SUPPRESSED AS $1/K$



$$f(x) = \frac{1}{K}\sum_{k=1}^{K} f_{\Theta^k}(x)$$

[1] Geiger, Mario, et al. "**Scaling description of generalization with number of parameters in deep learning**." arXiv preprint arXiv:1901.01608 (2019).

EFFECT OF ENSEMBLING

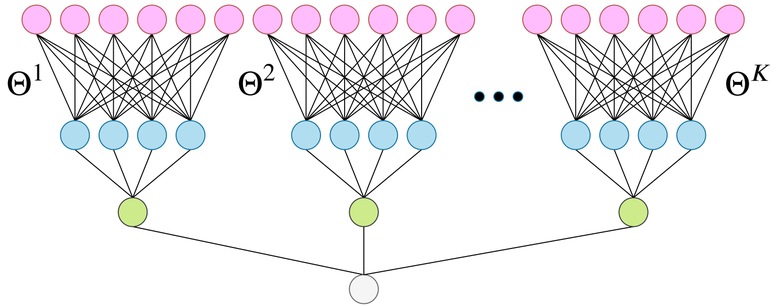
$$\mathop{\mathbb{E}}_{\Theta,X,\varepsilon}\left[\mathscr{R}_{\mathrm{RF}}\right] = F^2 + 2\mathbb{E}[f\hat{f}] + \frac{1}{K}\mathbb{E}\left[\hat{f}^2\right] + \left(1-\frac{1}{K}\right)\mathbb{E}\left[\left\langle\hat{f}\right\rangle_\Theta^2\right]$$

$$Var_{Noise} = \frac{1}{K}Var_\varepsilon \hat{f} + \left(1-\frac{1}{K}\right)Var_\varepsilon\left\langle\hat{f}\right\rangle_\Theta$$

$$Var_{Init} = \frac{1}{K}Var_\Theta\left\langle\hat{f}\right\rangle_\varepsilon$$

LINEAR INTERPOLATION TO ENSEMBLING TERMS

VANILLA TERMS SUPPRESSED AS $1/K$



$$f(x) = \frac{1}{K}\sum_{k=1}^{K} f_{\Theta^k}(x)$$

[1] Geiger, Mario, et al. "**Scaling description of generalization with number of parameters in deep learning**." arXiv preprint arXiv:1901.01608 (2019).

EFFECT OF ENSEMBLING

$$\mathop{\mathbb{E}}_{\Theta,X,\varepsilon}\left[\mathscr{R}_{\mathrm{RF}}\right] = F^2 + 2\mathbb{E}[f\hat{f}] + \frac{1}{K}\mathbb{E}\left[\hat{f}^2\right] + \left(1-\frac{1}{K}\right)\mathbb{E}\left[\left\langle\hat{f}\right\rangle_\Theta^2\right]$$

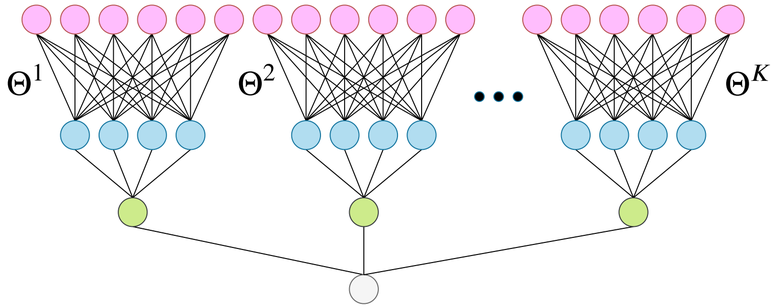
$$Var_{Noise} = \frac{1}{K}Var_\varepsilon \hat{f} + \left(1-\frac{1}{K}\right)Var_\varepsilon\left\langle\hat{f}\right\rangle_\Theta$$

$$Var_{Init} = \frac{1}{K}Var_\Theta\left\langle\hat{f}\right\rangle_\varepsilon$$

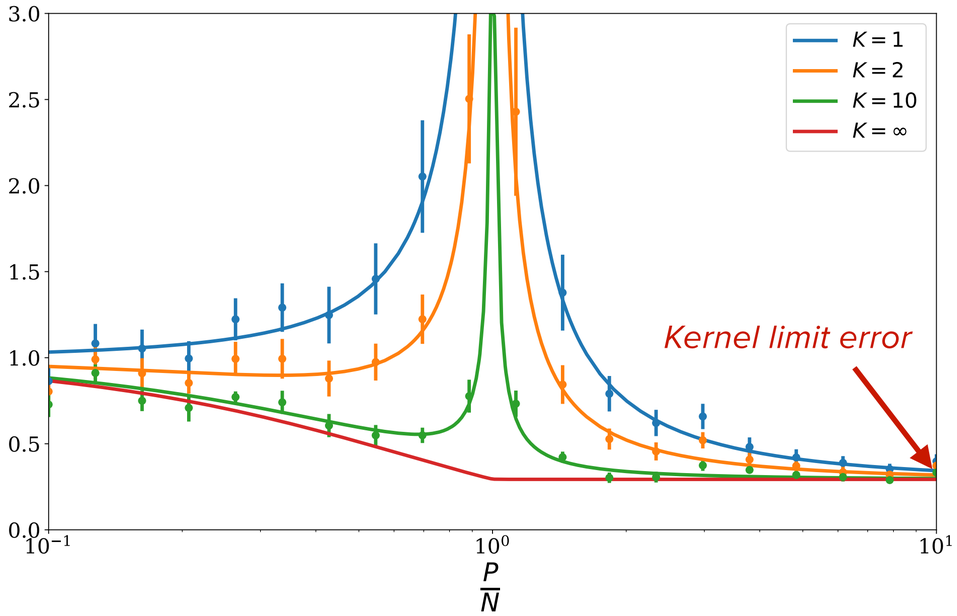
LINEAR INTERPOLATION TO ENSEMBLING TERMS

VANILLA TERMS SUPPRESSED AS $1/K$

CONSISTENT WITH EMPIRICAL STUDY IN NEURAL NETWORKS [1]



$$f(x) = \frac{1}{K}\sum_{k=1}^{K} f_{\Theta^k}(x)$$

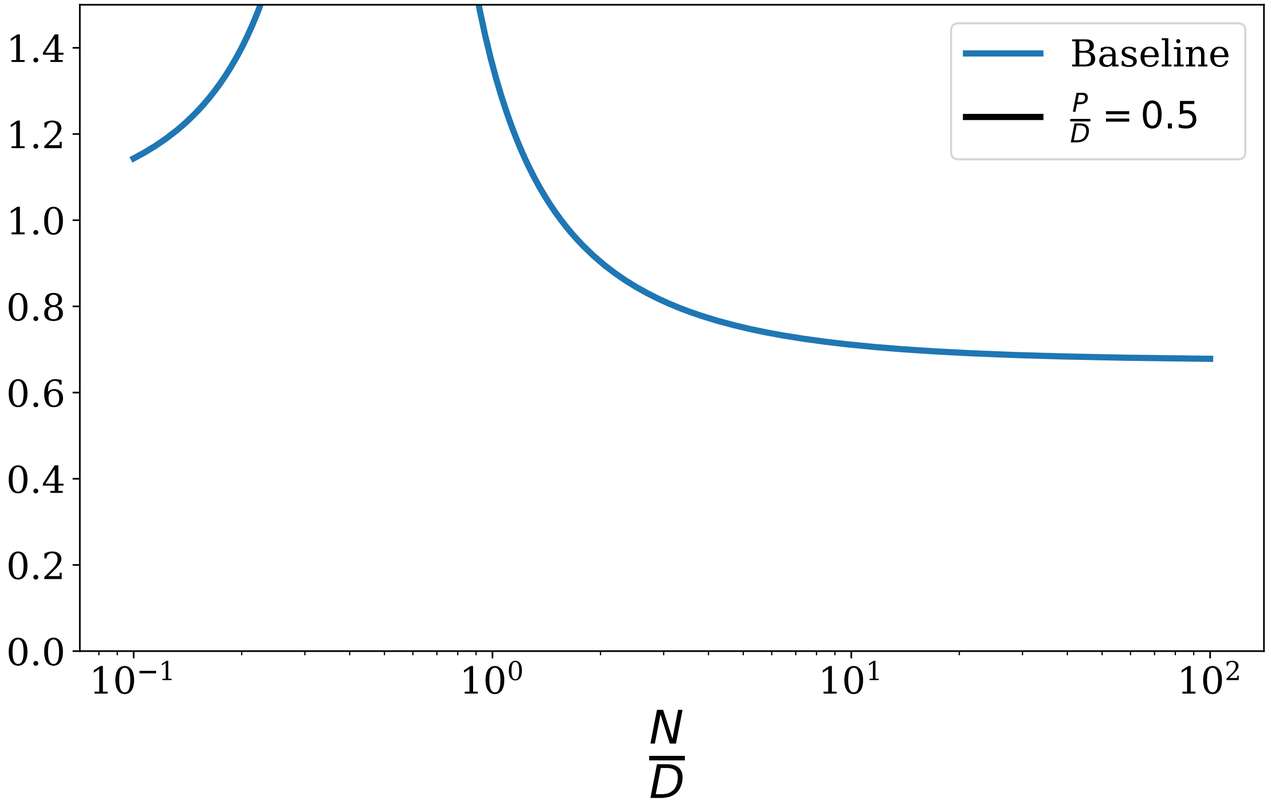


[1] Geiger, Mario, et al. "**Scaling description of generalization with number of parameters in deep learning**." arXiv preprint arXiv:1901.01608 (2019).

ENSEMBLING VS OVERPARAMETRIZING

Both **ensembling** and **overparametrizing** amounts to approaching the kernel limit.

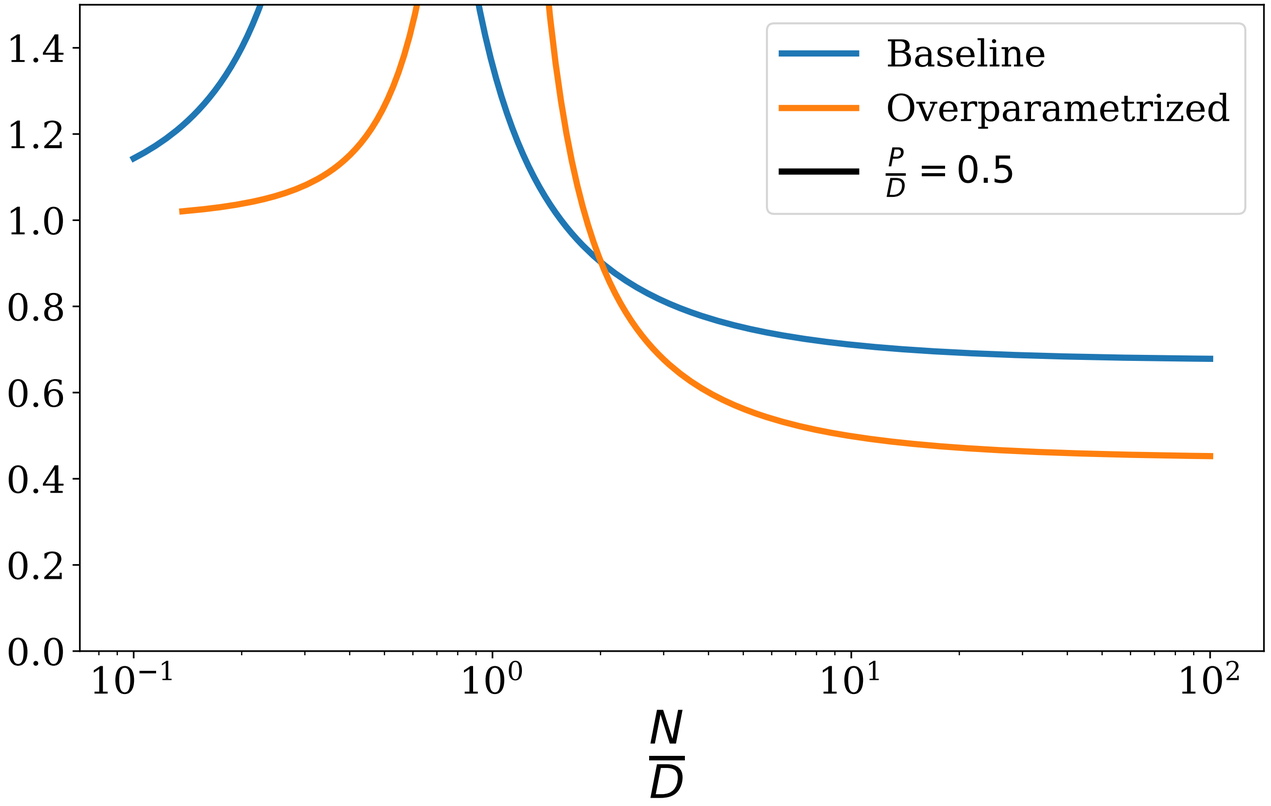
Which is better: doubling the number of features or ensembling 2 networks ?



ENSEMBLING VS OVERPARAMETRIZING

Both **ensembling** and **overparametrizing** amounts to approaching the kernel limit.

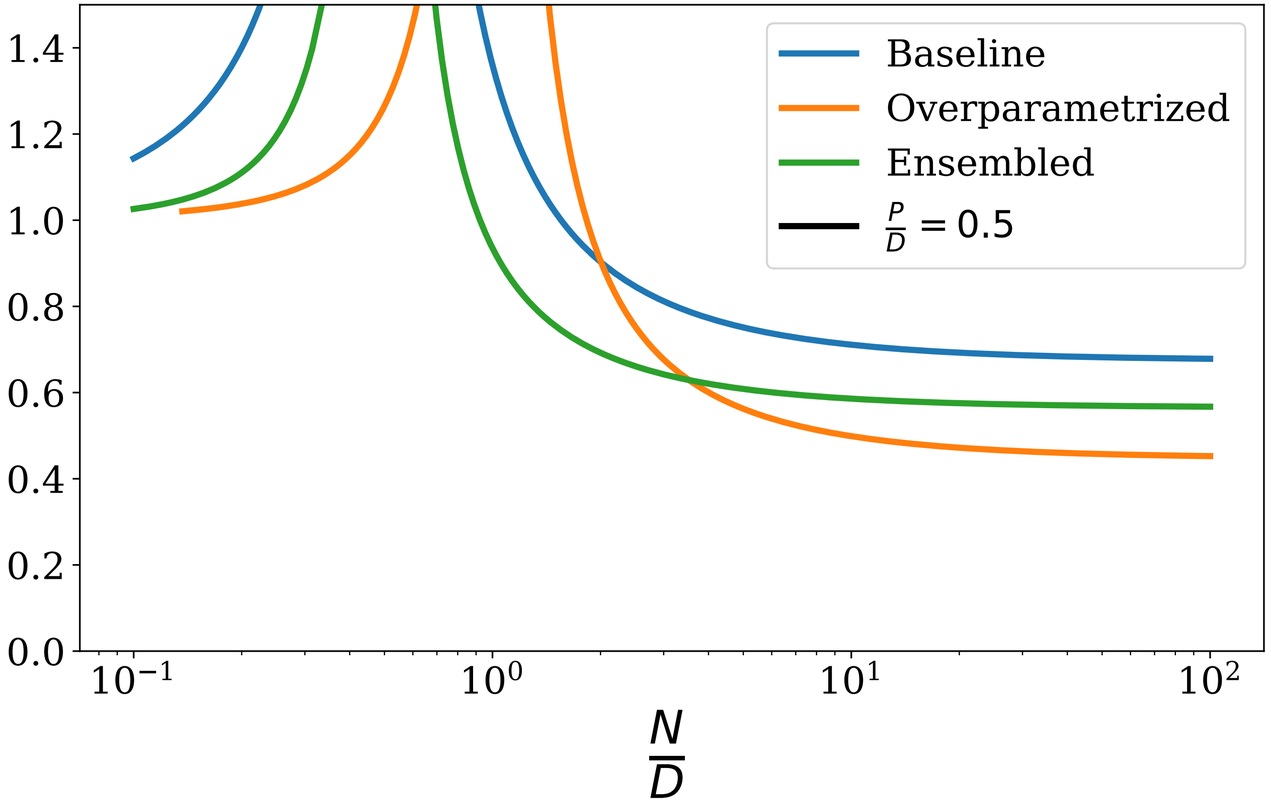
Which is better: doubling the number of features or ensembling 2 networks ?



ENSEMBLING VS OVERPARAMETRIZING

Both **ensembling** and **overparametrizing** amounts to approaching the kernel limit.

Which is better: doubling the number of features or ensembling 2 networks ?

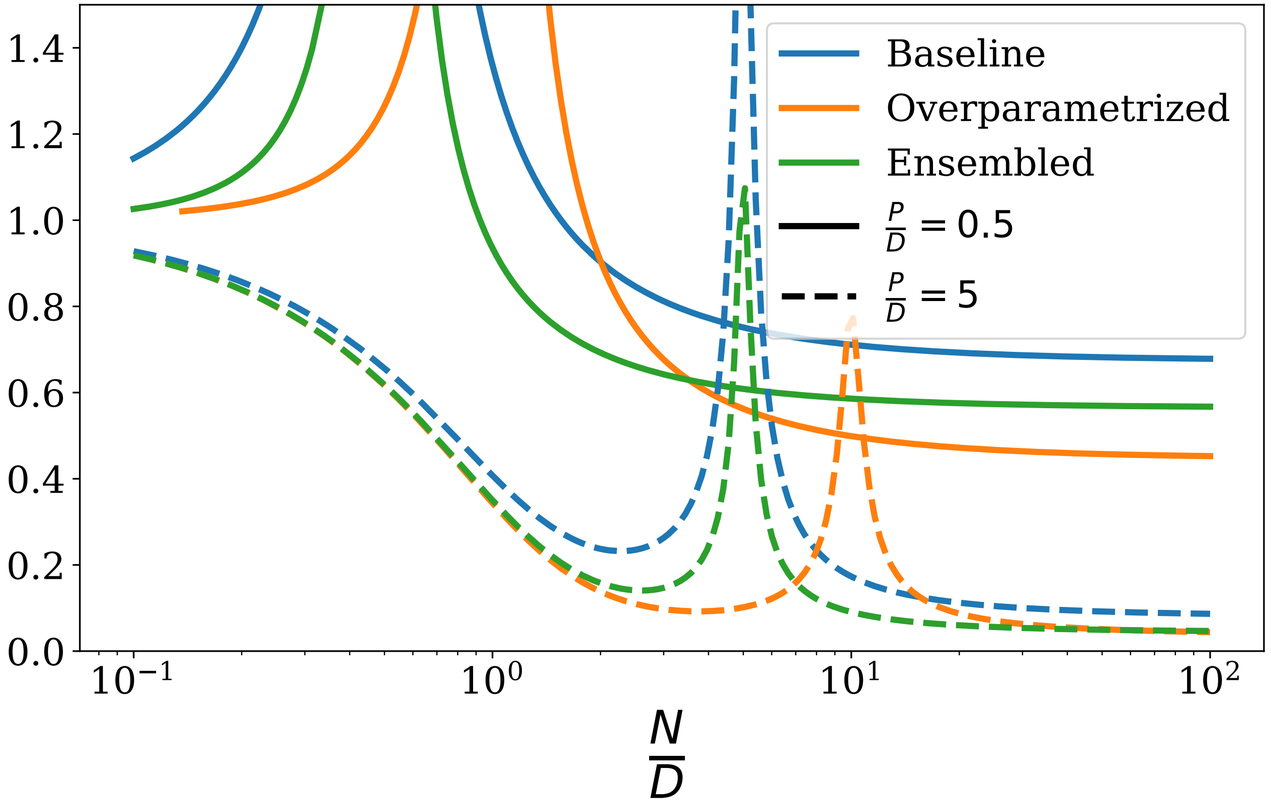


Overparametrizing is better (but slower!)

ENSEMBLING VS OVERPARAMETRIZING

Both **ensembling** and **overparametrizing** amounts to approaching the kernel limit.

Which is better: doubling the number of features or ensembling 2 networks ?



Overparametrizing is better (but slower!)

Except when already overparametrized

ENSEMBLING VS DIVIDE AND CONQUER

Ensembling:

Train K learners on same dataset

$$\mathop{\mathbb{E}}_{\boldsymbol{\Theta},\boldsymbol{X},\boldsymbol{\varepsilon}}\left[\mathscr{R}_{\mathrm{RF}}\right] = F^2\left(1-2\Psi_1^v\right) + \frac{1}{K}\left(F^2\Psi_2^v + \tau^2\Psi_3^v\right) + \left(1-\frac{1}{K}\right)\left(F^2\Psi_2^e + \tau^2\Psi_3^e\right)$$

Divide and conquer:

Train on different splits of the dataset

$$\mathop{\mathbb{E}}_{\boldsymbol{\Theta},\boldsymbol{X},\boldsymbol{\varepsilon}}\left[\mathscr{R}_{\mathrm{RF}}\right] = F^2\left(1-2\Psi_1^v\right) + \frac{1}{K}\left(F^2\Psi_2^v + \tau^2\Psi_3^v\right) + \left(1-\frac{1}{K}\right)F^2\Psi_2^d$$

ENSEMBLING VS DIVIDE AND CONQUER

Ensembling:

Train K learners on same dataset

$$\mathop{\mathbb{E}}_{\boldsymbol{\Theta},\boldsymbol{X},\boldsymbol{\varepsilon}}\left[\mathscr{R}_{\mathrm{RF}}\right] = F^2\left(1-2\Psi_1^v\right) + \frac{1}{K}\left(F^2\Psi_2^v + \tau^2\Psi_3^v\right) + \left(1-\frac{1}{K}\right)\left(F^2\Psi_2^e + \tau^2\Psi_3^e\right)$$

Divide and conquer:

Train on different splits of the dataset

$$\mathop{\mathbb{E}}_{\boldsymbol{\Theta},\boldsymbol{X},\boldsymbol{\varepsilon}}\left[\mathscr{R}_{\mathrm{RF}}\right] = F^2\left(1-2\Psi_1^v\right) + \frac{1}{K}\left(F^2\Psi_2^v + \tau^2\Psi_3^v\right) + \left(1-\frac{1}{K}\right)F^2\Psi_2^d$$

Each learner has less data BUT noise term disappears as $K \to \infty$

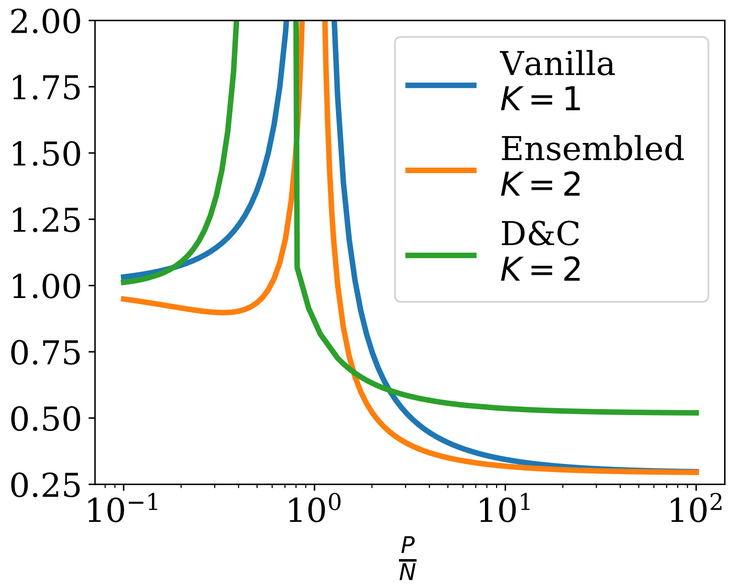
ENSEMBLING VS DIVIDE AND CONQUER

Ensembling:

Train K learners on same dataset

$$\mathop{\mathbb{E}}_{\Theta,X,\varepsilon}\left[\mathscr{R}_{\mathrm{RF}}\right] = F^2\left(1-2\Psi_1^v\right) + \frac{1}{K}\left(F^2\Psi_2^v + \tau^2\Psi_3^v\right) + \left(1-\frac{1}{K}\right)\left(F^2\Psi_2^e + \tau^2\Psi_3^e\right)$$

BETTER AT LOW NOISE



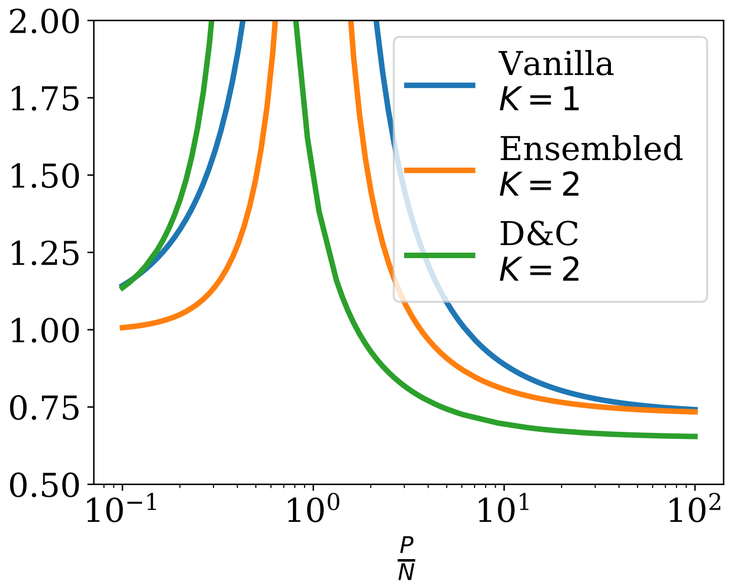
Divide and conquer:

Train on different splits of the dataset

$$\mathop{\mathbb{E}}_{\Theta,X,\varepsilon}\left[\mathscr{R}_{\mathrm{RF}}\right] = F^2\left(1-2\Psi_1^v\right) + \frac{1}{K}\left(F^2\Psi_2^v + \tau^2\Psi_3^v\right) + \left(1-\frac{1}{K}\right)F^2\Psi_2^d$$

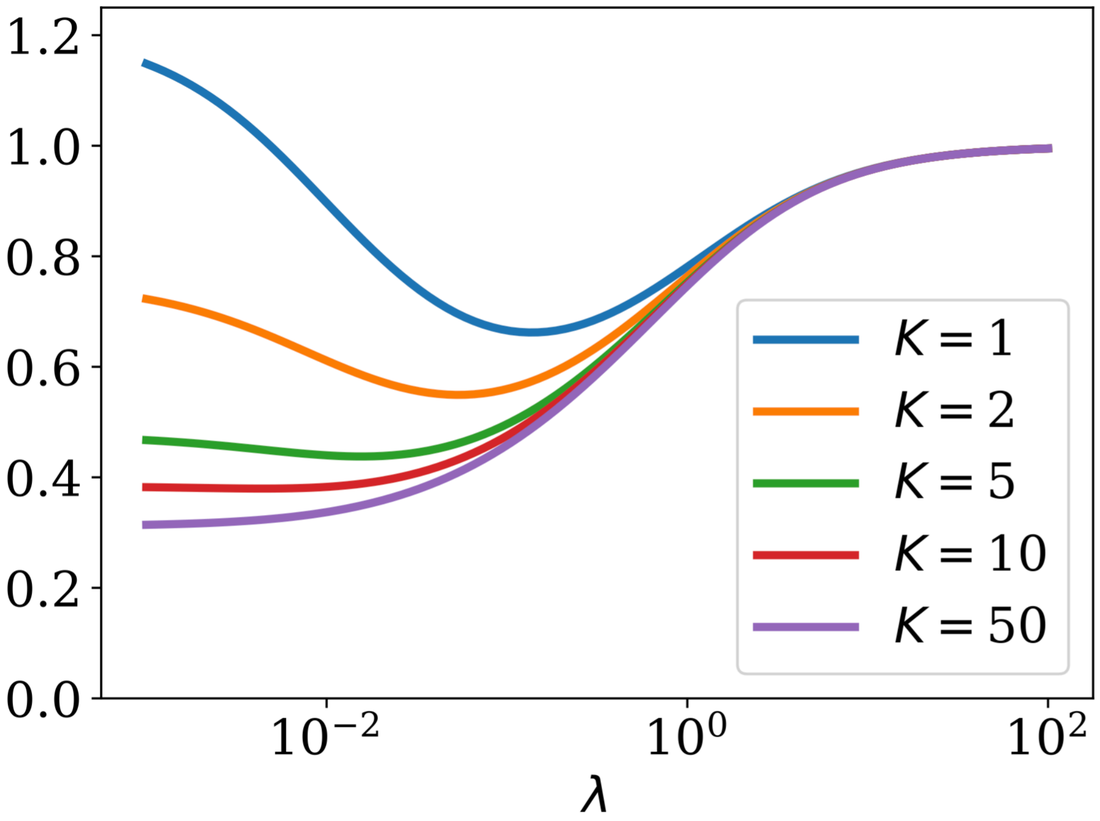
Each learner has less data BUT noise term disappears as $K \to \infty$

BETTER AT HIGH NOISE

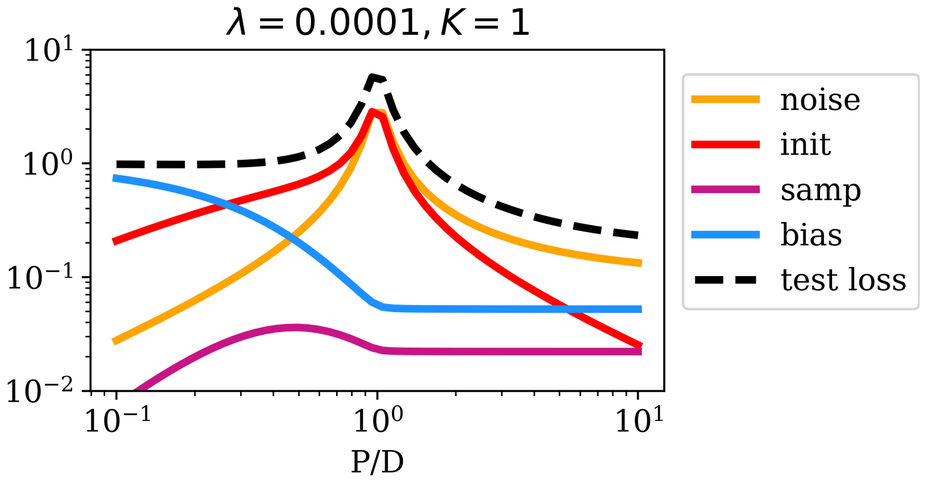


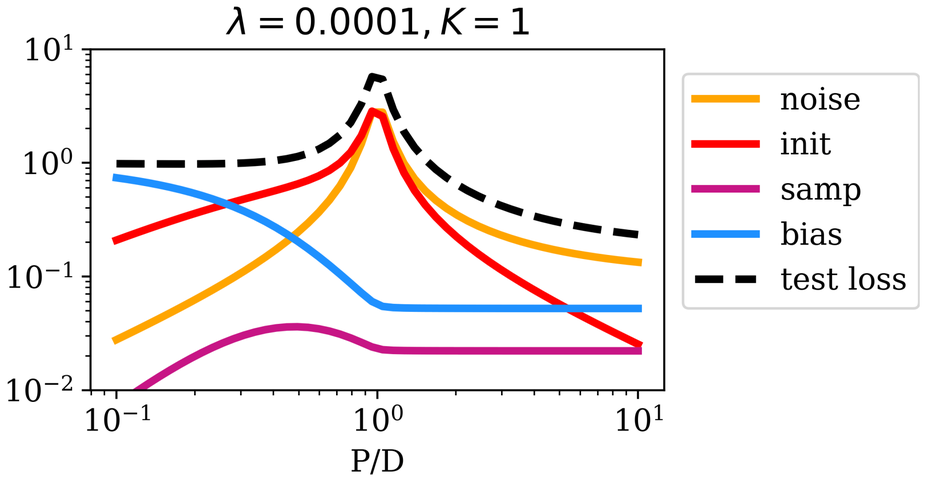
ENSEMBLING VS REGULARIZING

Ensembled learners are better off **under-regularised**



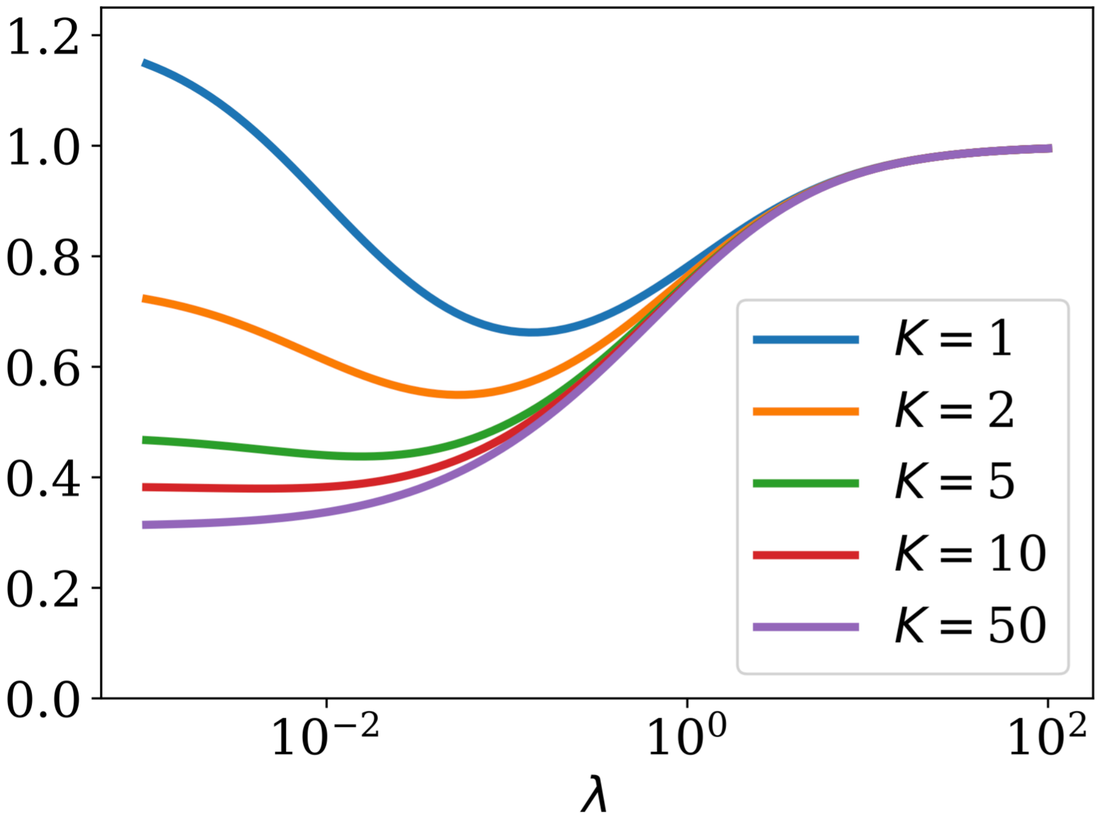
Ensembling also **reduces the variance**, but it also **increases the bias**



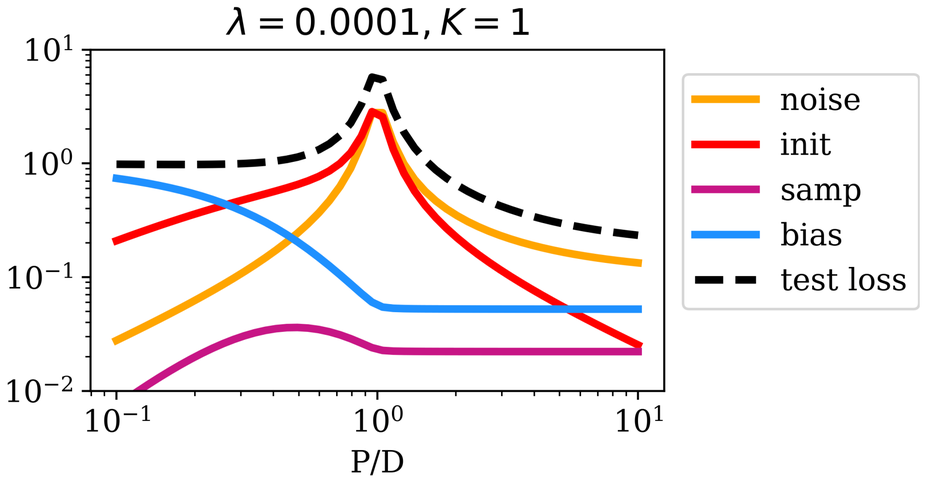


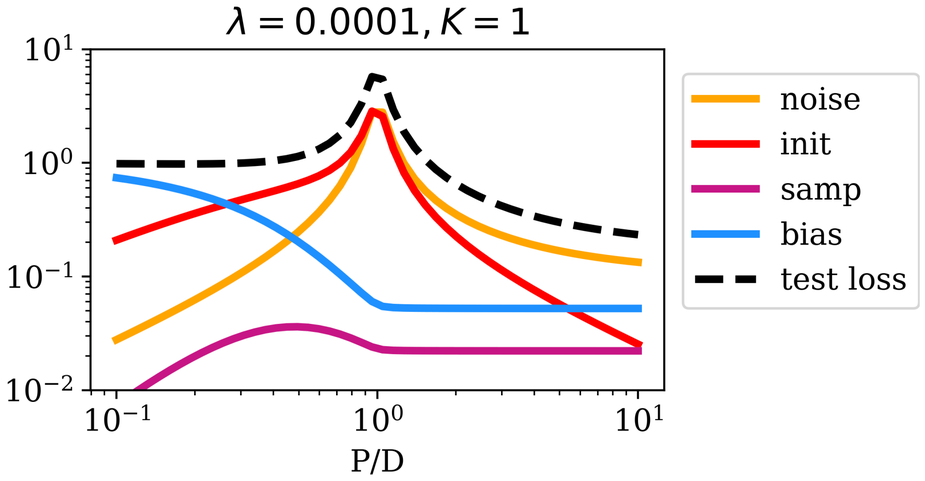
ENSEMBLING VS REGULARIZING

Ensembled learners are better off **under-regularised**



Ensembling also **reduces the variance**, but it also **increases the bias**





LAZY REGIME OF NEURAL NETWORKS

The lazy regime of neural network (weights move little) is equivalent to random feature regression:

$$f_\theta(\boldsymbol{x}) \approx \underbrace{f_{\theta_0}(\boldsymbol{x})}_{\text{Assume to be = 0}} + \underbrace{\nabla_\theta f_\theta(\boldsymbol{x})\Big|_{\theta=\theta_0}}_{\text{Random features}} \cdot \underbrace{(\theta - \theta_0)}_{\text{Vector to be learnt}}$$

We can enforce lazy behaviour by scaling the output of the network:

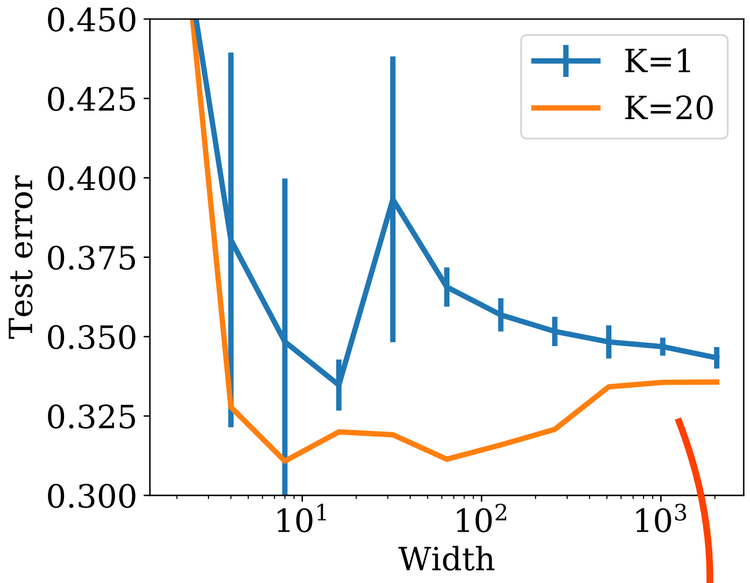
$$\underbrace{\alpha(f_\theta(\boldsymbol{x}) - f_{\theta_0}(\boldsymbol{x}))}_{\text{Must be of order one}} \approx \alpha \nabla f_{\theta_0}(\boldsymbol{x}) \cdot \underbrace{(\theta - \theta_0)}_{\text{Must scale as } 1/\alpha \ll 1}$$

$\gg 1$

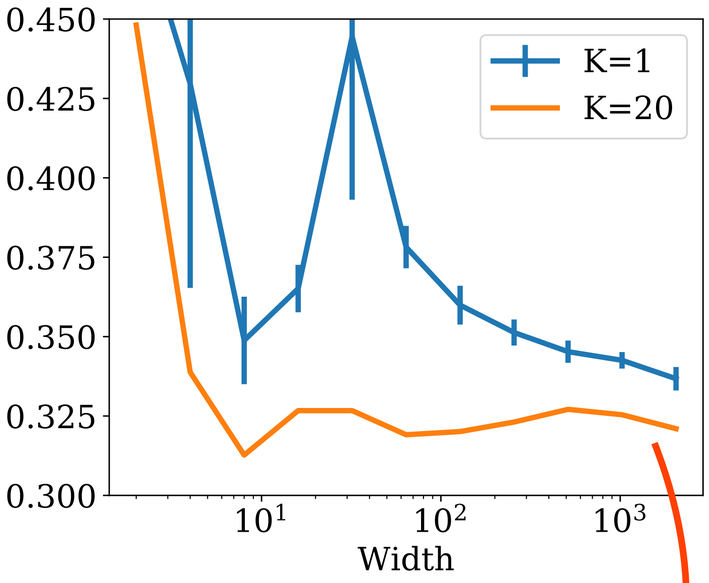
Hence we expect to observe the phase transition of the bias in neural networks at high α

LAZY REGIME OF NEURAL NETWORKS

5-layer MLP on binary 10-PCA CIFAR-10



$\alpha = 10$

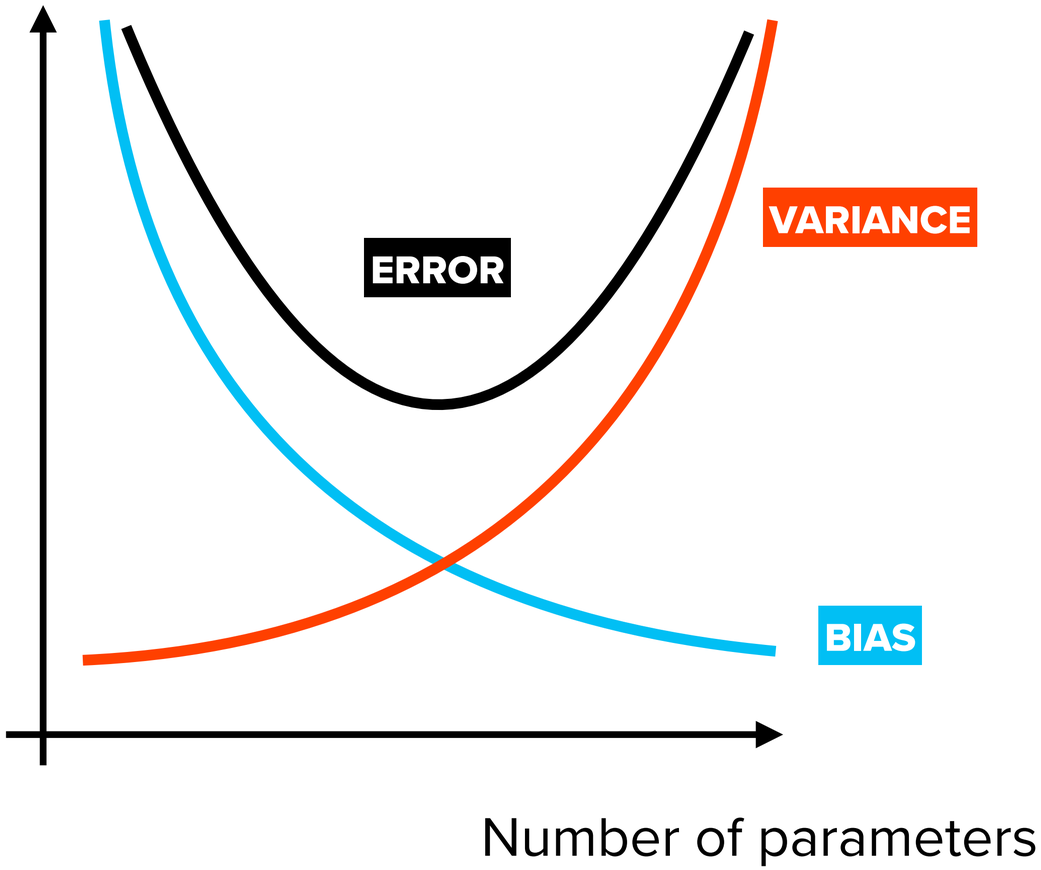


$\alpha = 100$

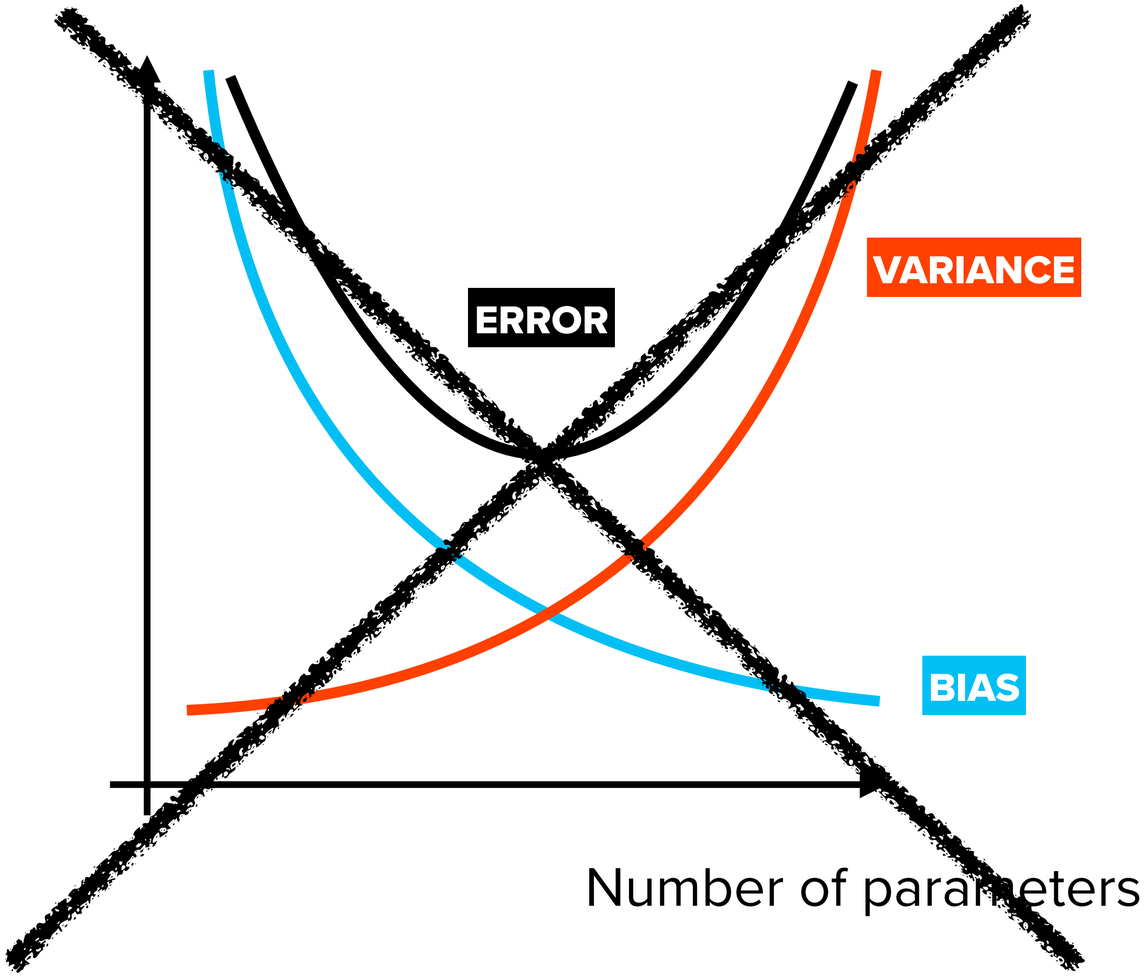
Ensembled error increases
(transition from feature learning to lazy learning)

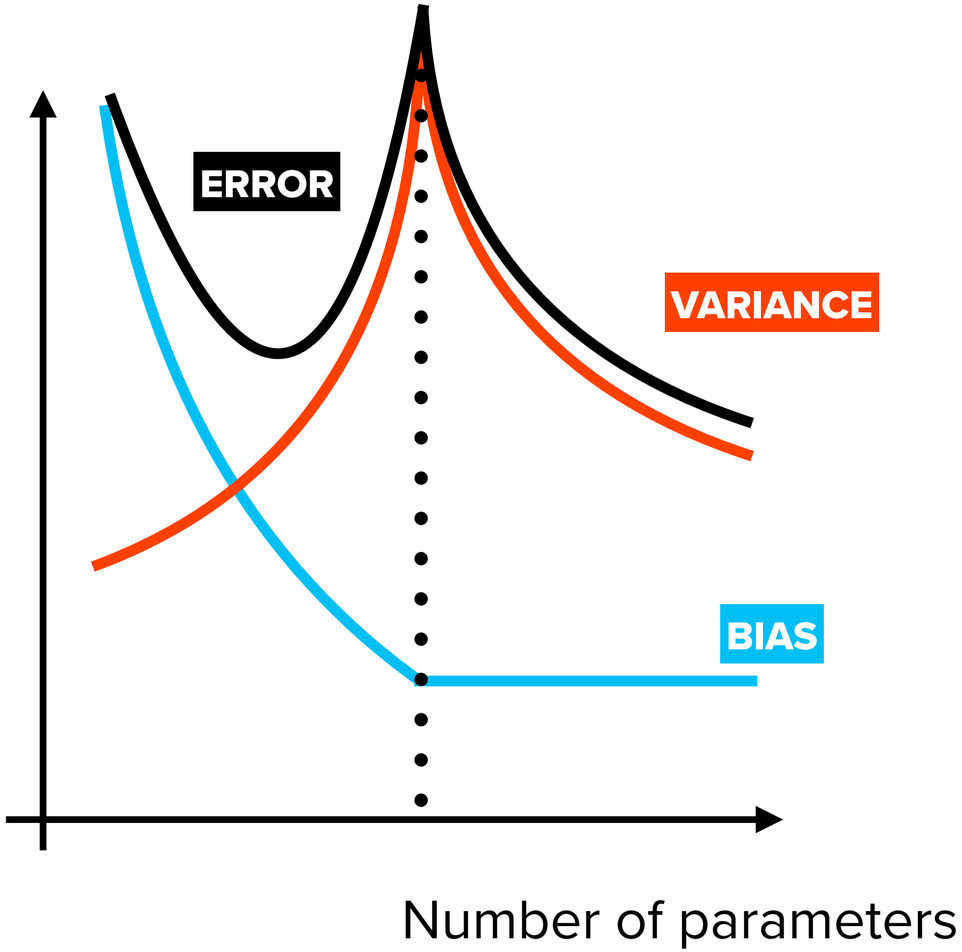
Ensembled error stays flat after interpolation threshold

TAKEAWAY



TAKEAWAY


ERROR
VARIANCE
BIAS
Number of parameters


ERROR
VARIANCE
BIAS
Number of parameters

THANK YOU FOR YOUR ATTENTION

Stéphane d’Ascoli

Maria Refinetti

Giulio Biroli

Florent Krzakala