RECONCILING DOUBLE DESCENT WITH OLDER IDEAS

STÉPHANE D'ASCOLI

ÉCOLE NORMALE SUPÉRIEURE & FACEBOOK AI RESEARCH, PARIS

JOINT WORK WITH MARIA REFINETTI, LEVENT SAGUN, GIULIO BIROLI, FLORENT KRZAKALA

PART 1
RECONCILING DOUBLE DESCENT WITH THE BIAS-VARIANCE TRADEOFF

ARXIV 2003.01054

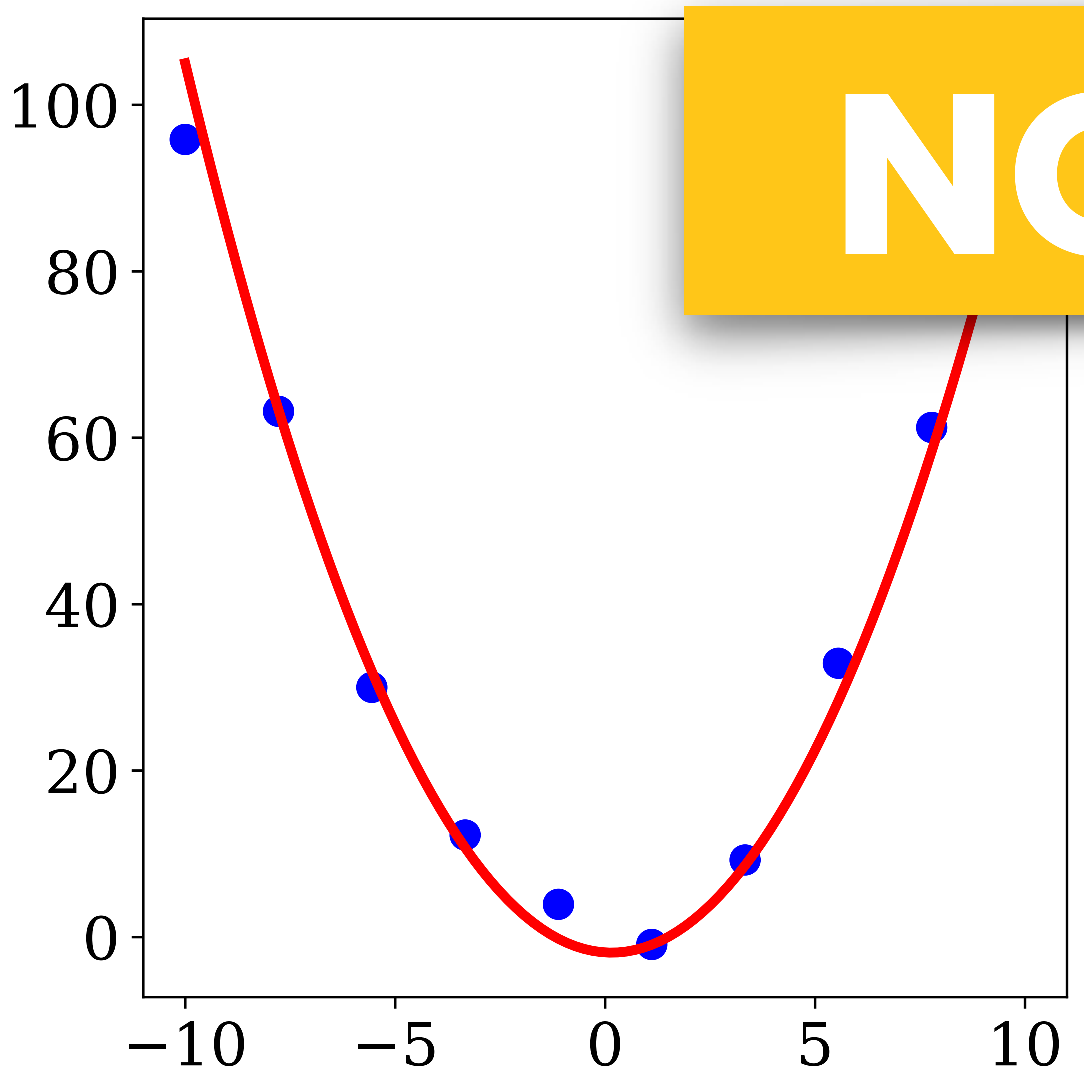
WITH MARIA REFINETTI, GIULIO BIROLI, FLORENT KRZAKALA

FROM CLASSICAL THEORY

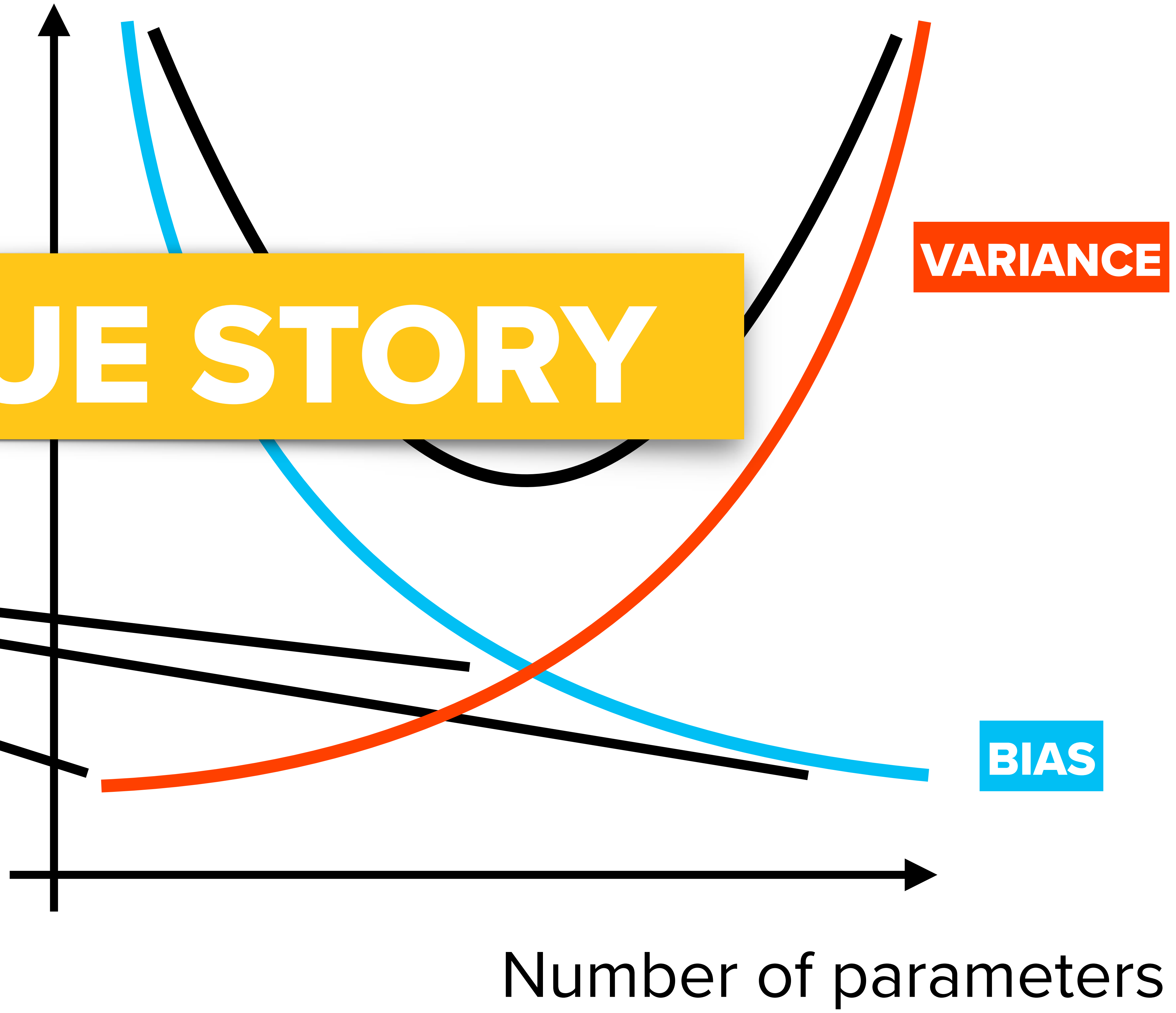
BIAS-VARIANCE TRADEOFF

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

BALANCED



NOT THE TRUE STORY



TO DOUBLE DESCENT

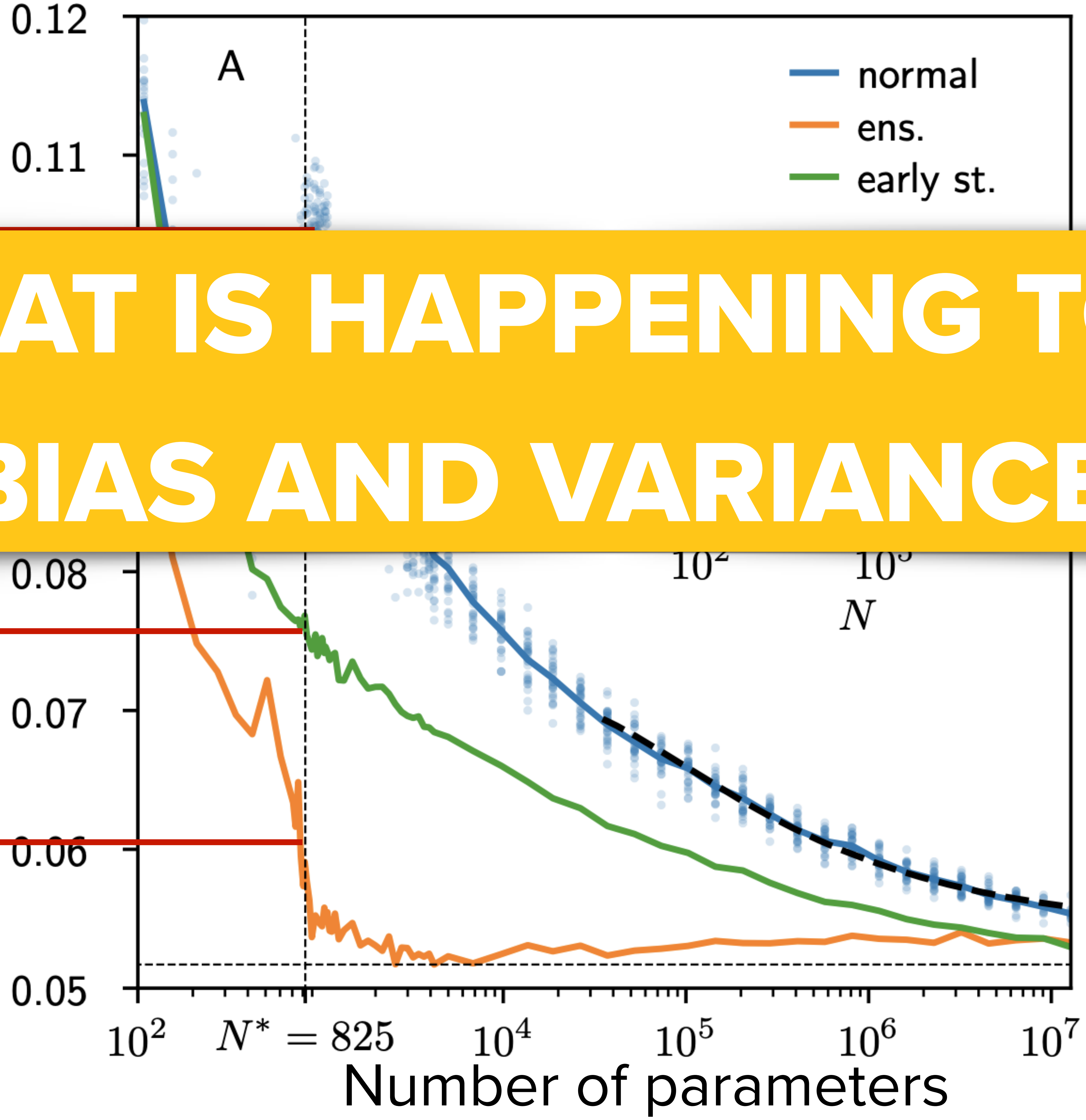
Peak in Test Error
At interpolation t

"U" shaped cur
underparametrize

Monotonous decreas
overparametrized r

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

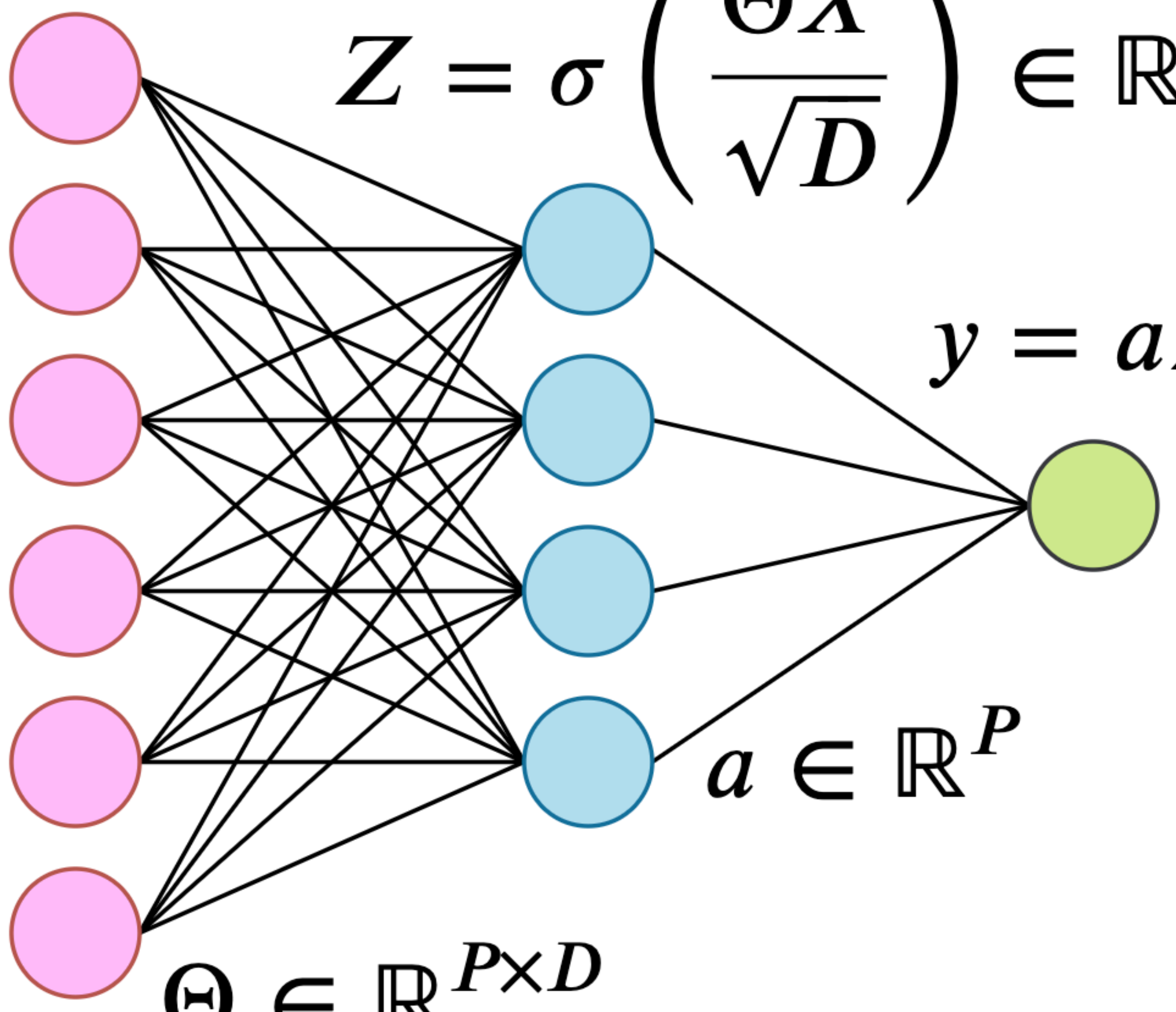
….

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?

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

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



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

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

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

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.

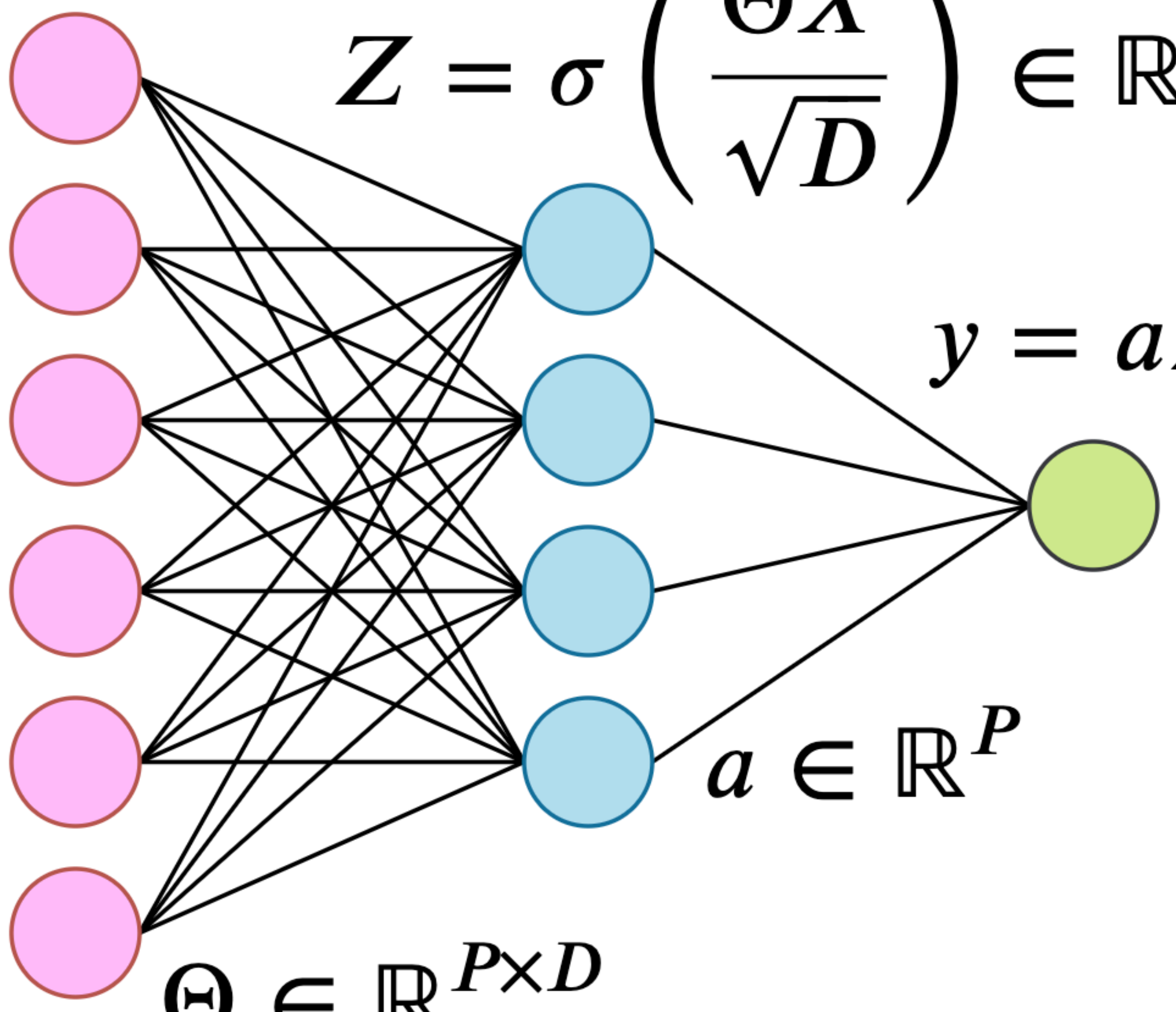
[4] Chizat, Lenaic, and Francis Bach. "**A note on lazy training in supervised differentiable programming**." arXiv preprint arXiv:1812.07956 8 (2018).

THE RANDOM FEATURES MODEL

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

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

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



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

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

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

LEARNER: ONE HIDDEN LAYER

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

GROUND TRUTH: LINEAR WITH NOISE

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

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

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

TRAIN ERROR

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

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

TEST ERROR

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

BIAS AND VARIANCE(S)

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

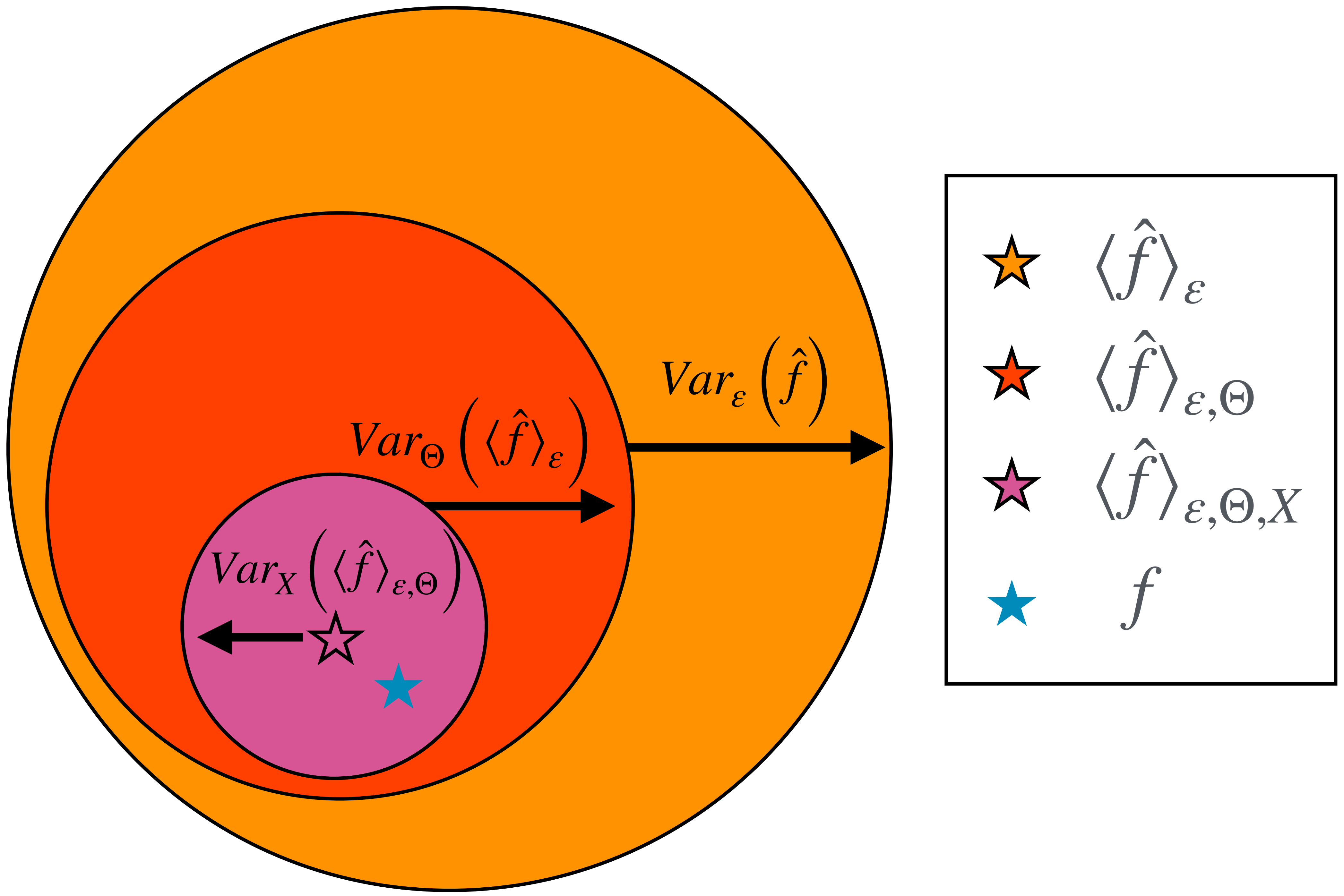
$$= F^2 + 2\mathbb{E}\left[f\hat{f}\right] + \mathbb{E}\left[\hat{f}^2\right]$$

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

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

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

Noise **Initialization** **Sampling**



BIAS AND VARIANCE(S)

HIGH DIMENSIONAL LIMIT

$$N, D, P \to \infty, \quad \frac{D}{P} = \mathcal{O}(1), \quad \frac{D}{N} = \mathcal{O}(1)$$

TEST ERROR

Noise Variance

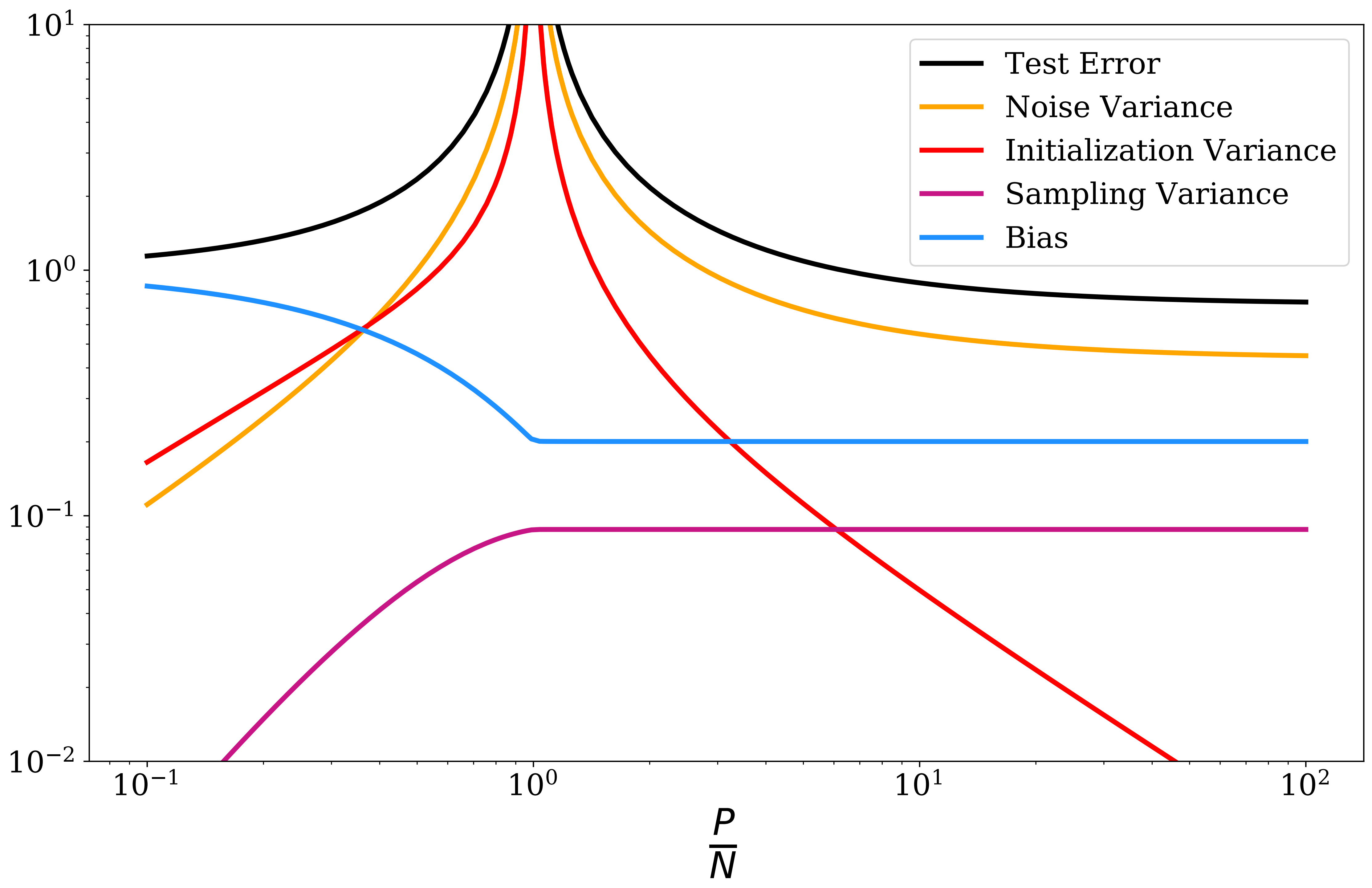
Initialisation Variance

Sampling Variance

Bias

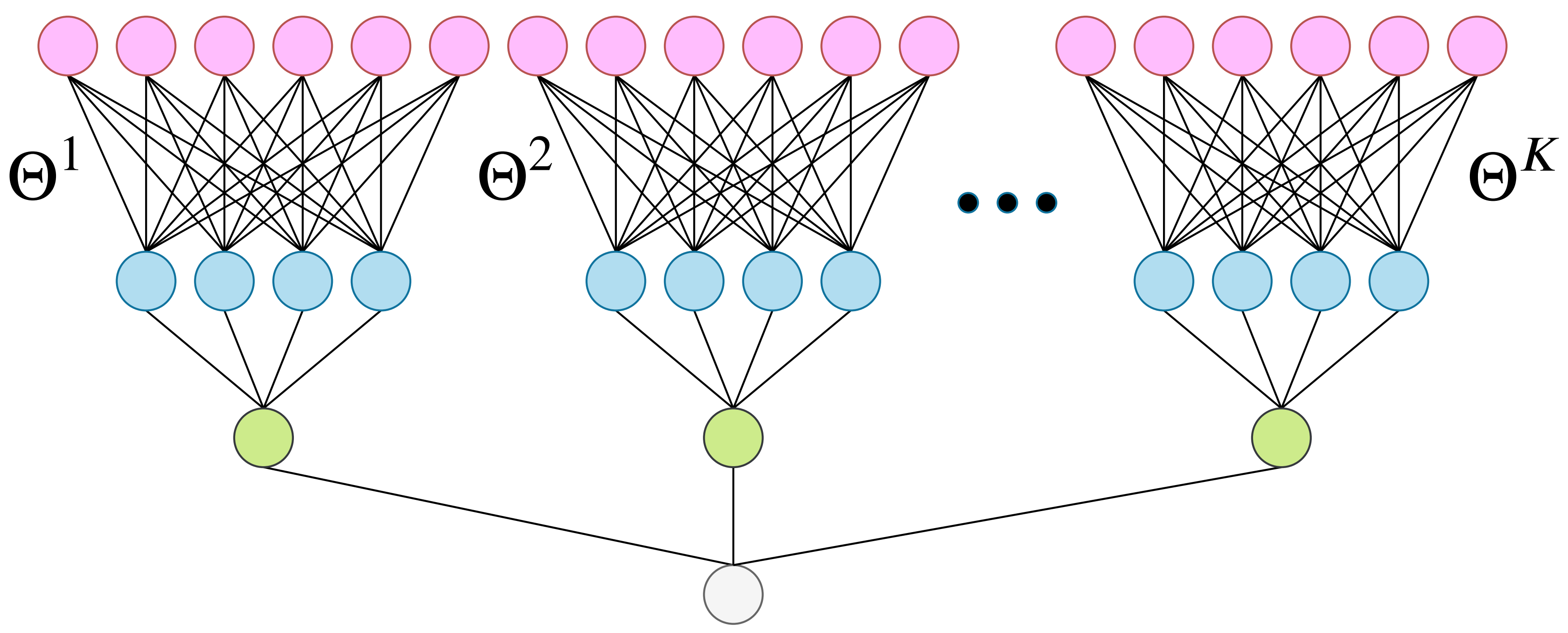
DOUBLE TROUBLE

HARMLESS



Analytical results for Random Feature Networks

ENSEMBLING

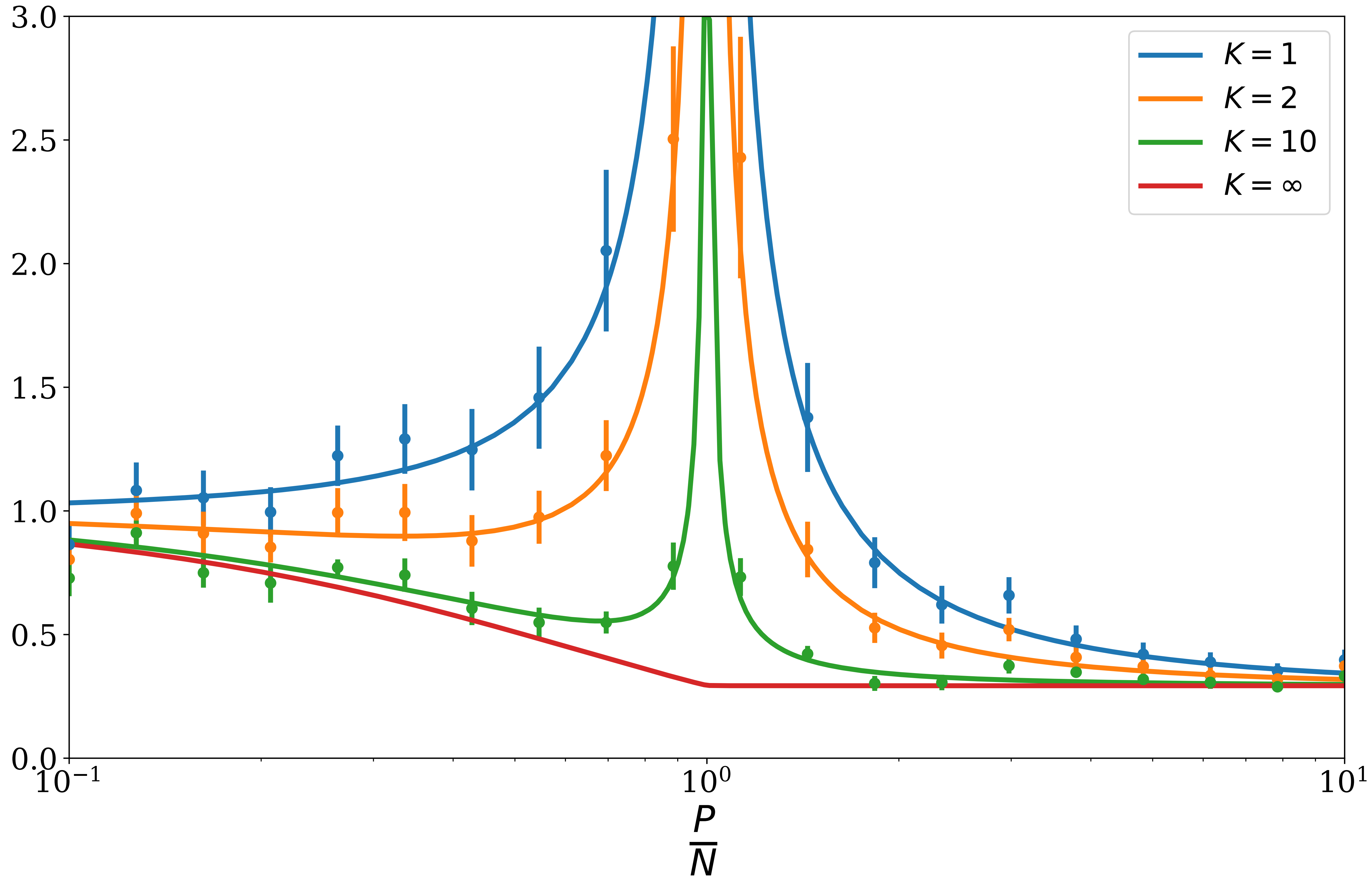


$$f(x) = \frac{1}{K}\sum_{k=1}^{K} f_{\Theta^k}(x)$$

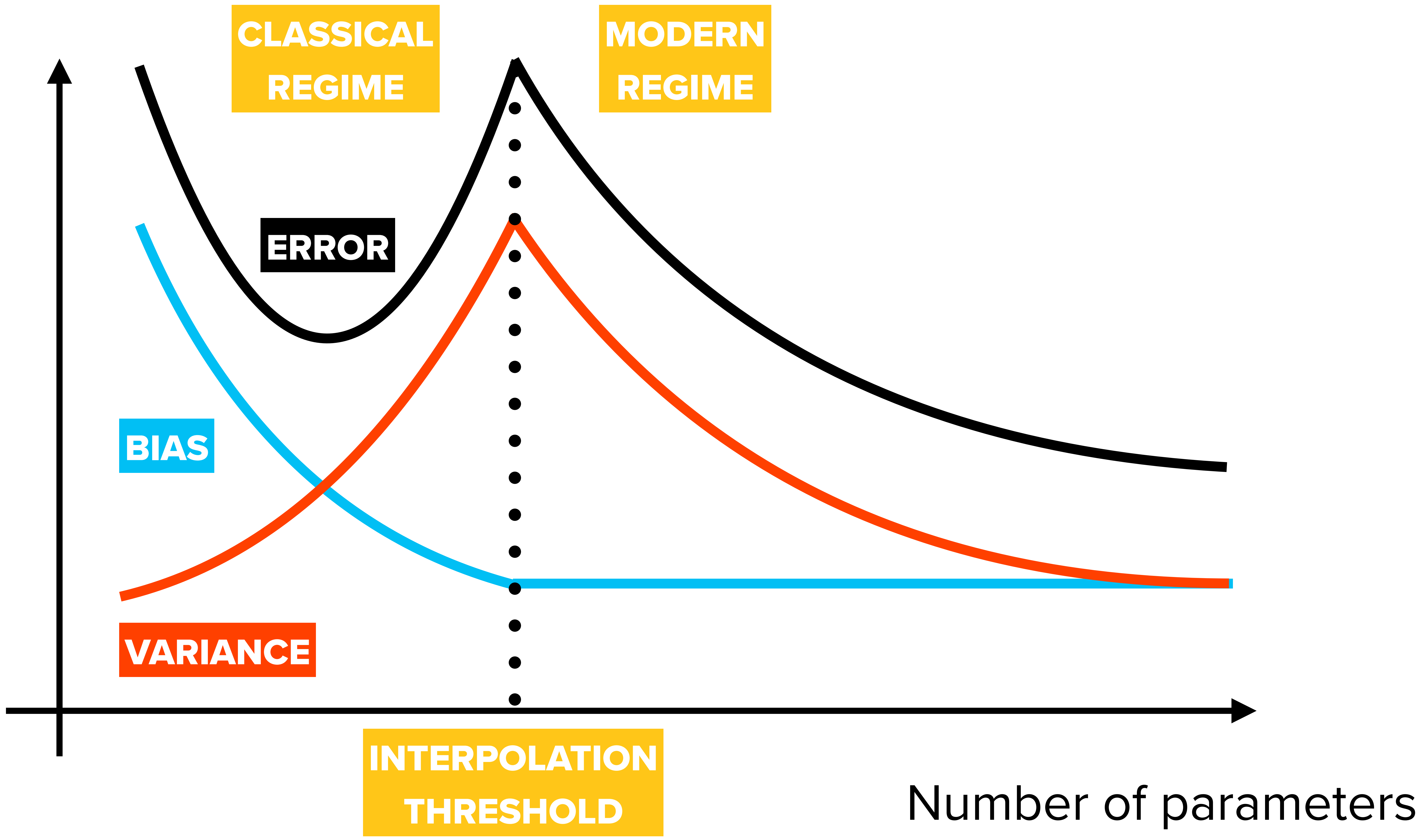
TEST ERROR

- ~~Noise Variance~~ — Attenuated
- ~~Initialisation Variance~~ — Suppressed
- Sampling Variance
- Bias

$K \to \infty$



TAKEAWAY


CLASSICAL REGIME
MODERN REGIME
ERROR
BIAS
VARIANCE
INTERPOLATION THRESHOLD
Number of parameters

PART 2
RECONCILING MODERN AND OLD SCHOOL DOUBLE DESCENT

ARXIV 2006.03509

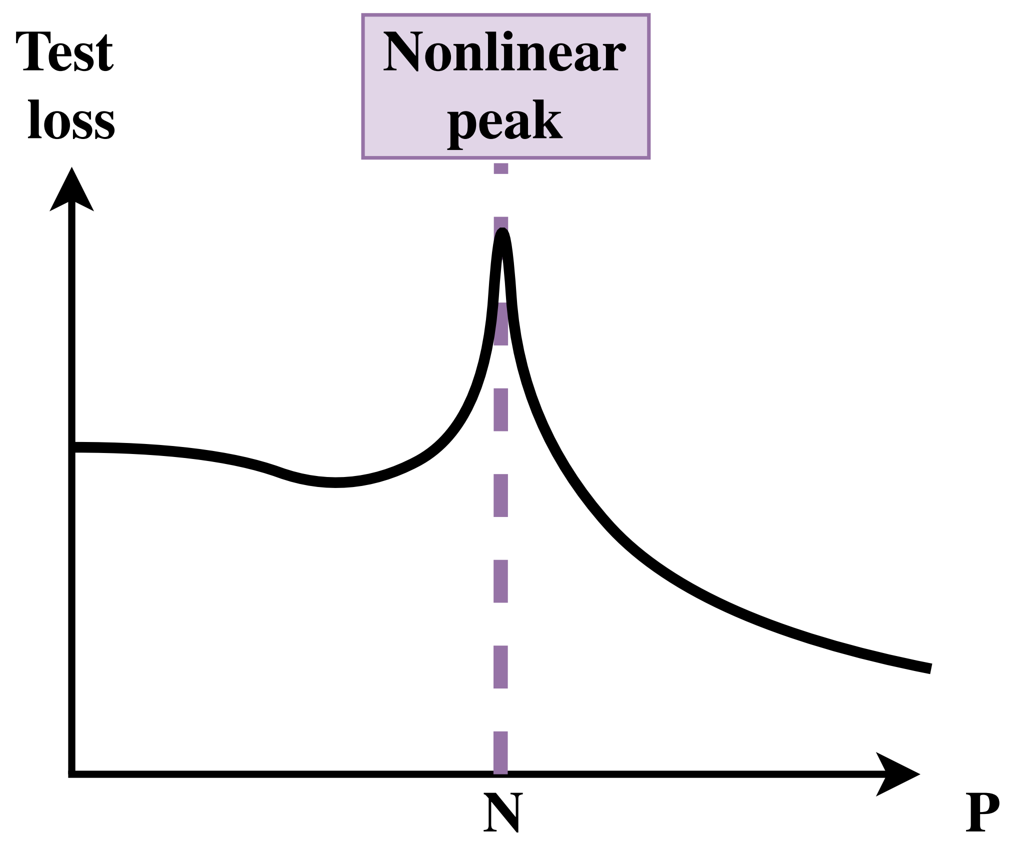
WITH LEVENT SAGUN AND GIULIO BIROLI

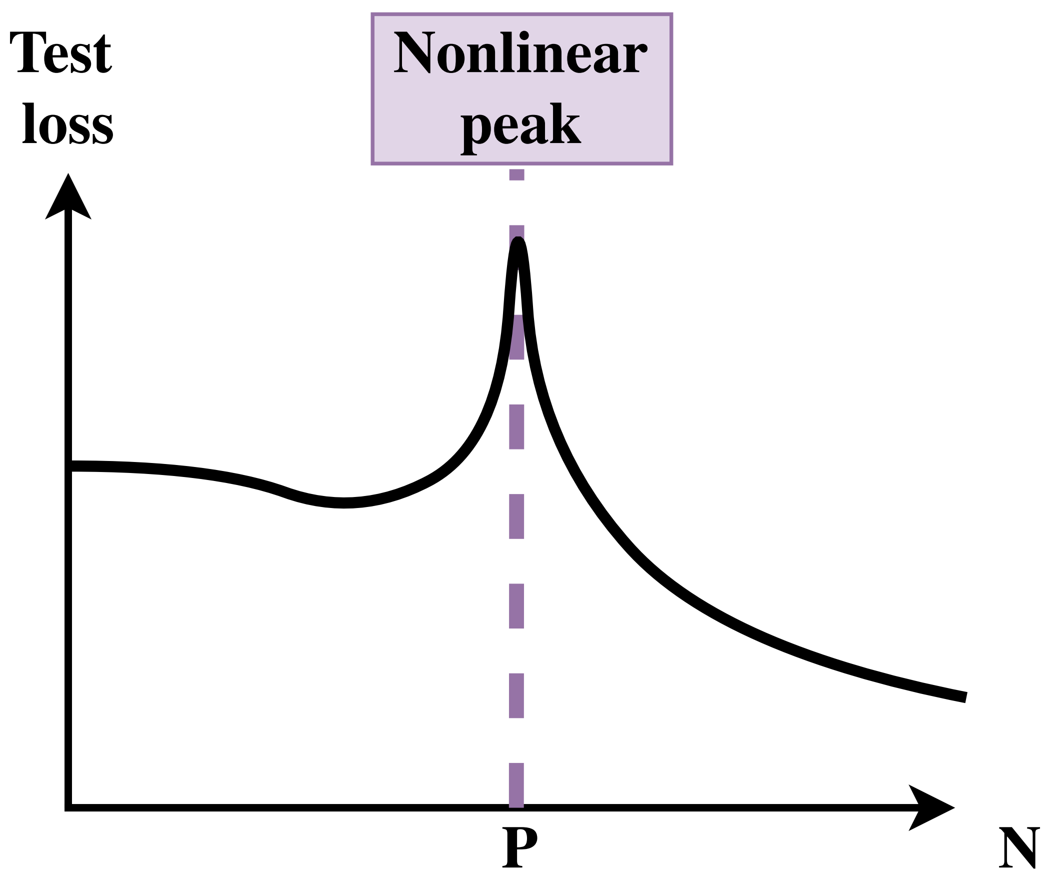
LINEAR VS NONLINEAR MODELS

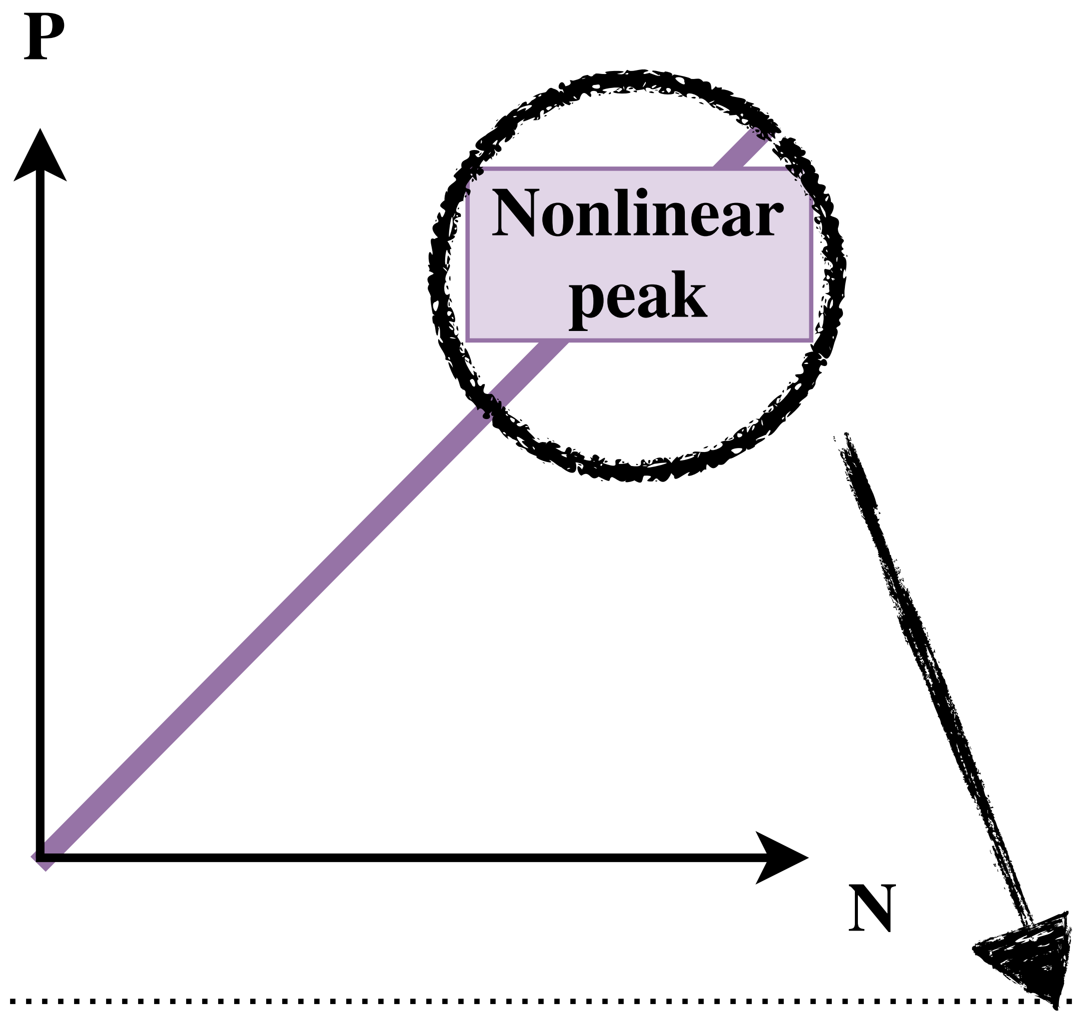
	PARAMETER-WISE	SAMPLE-WISE	PHASE-SPACE

NONLINEAR NETWORKS

“MODERN” DOUBLE DESCENT [GEIGER ’19]

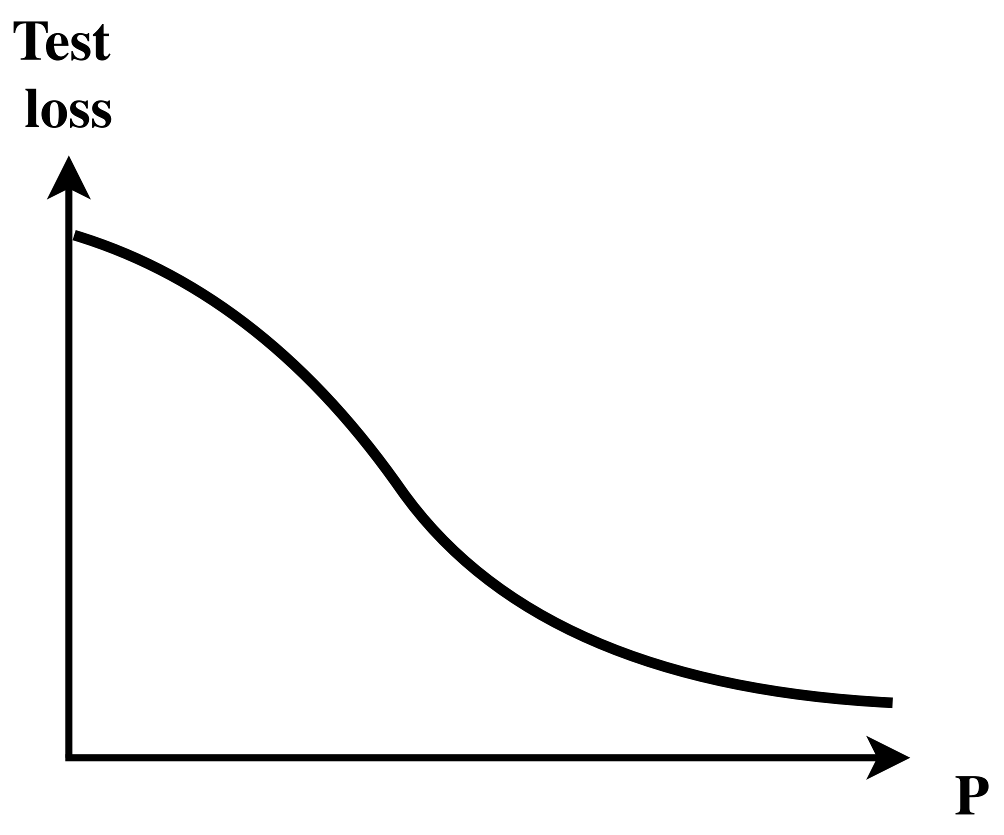


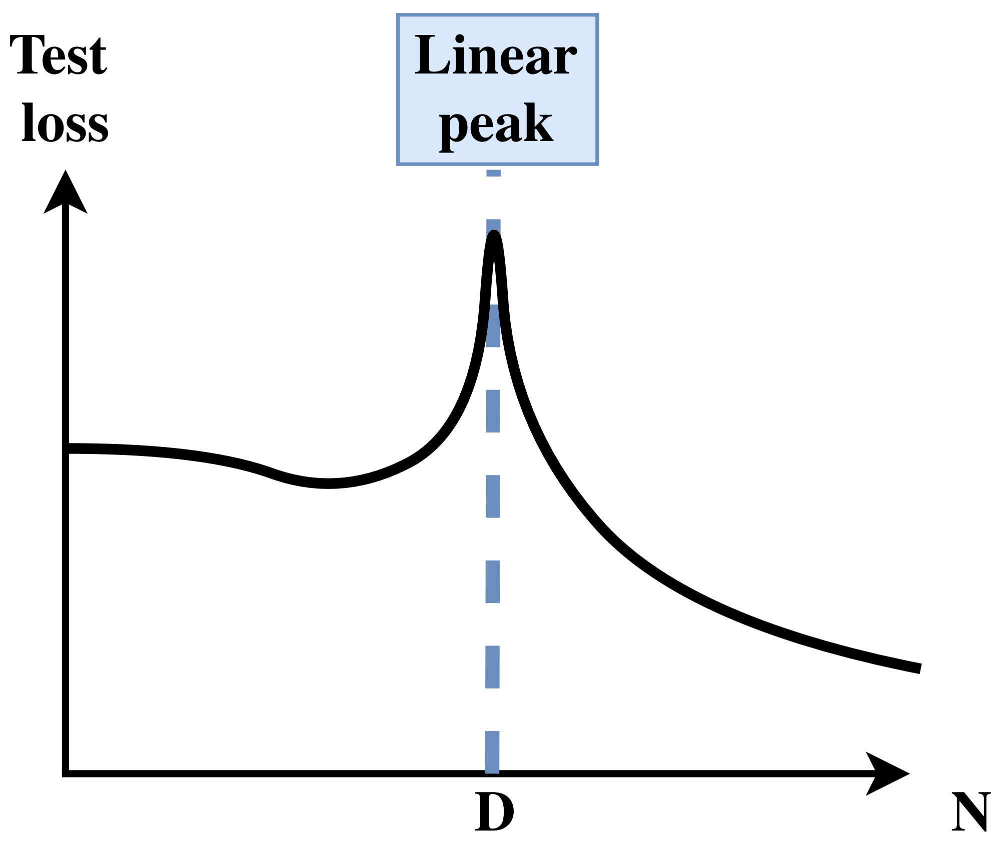


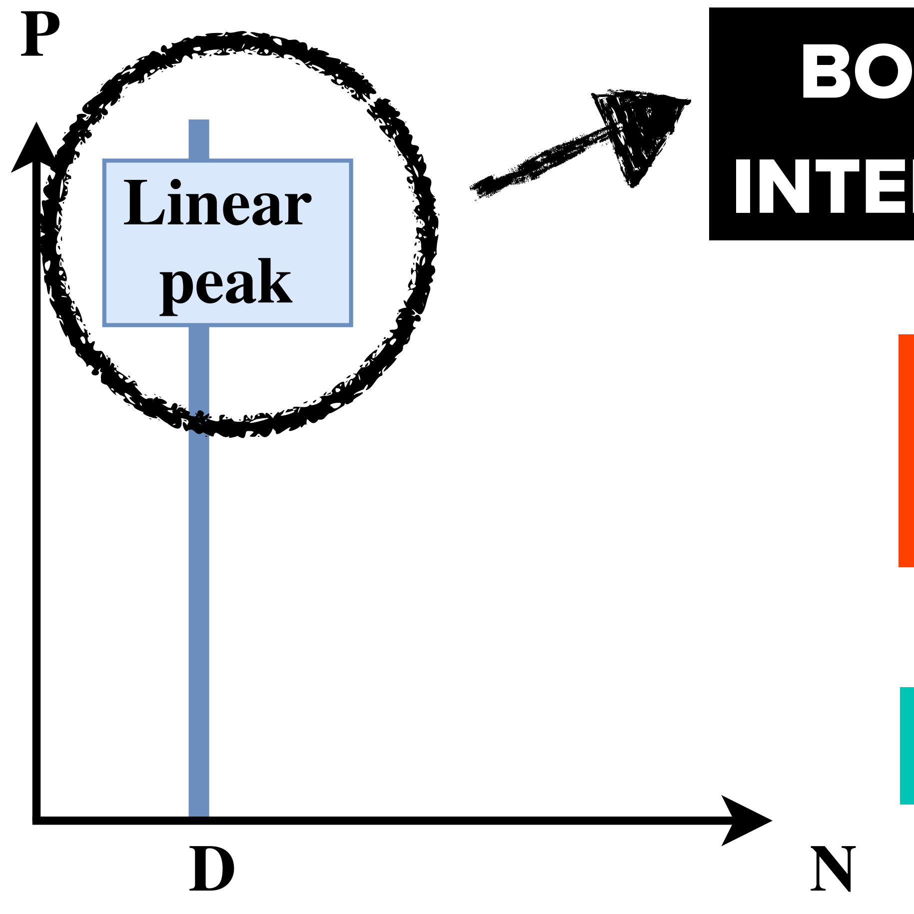


LINEAR NETWORKS

“OLD SCHOOL” DOUBLE DESCENT [OPPER ’97]







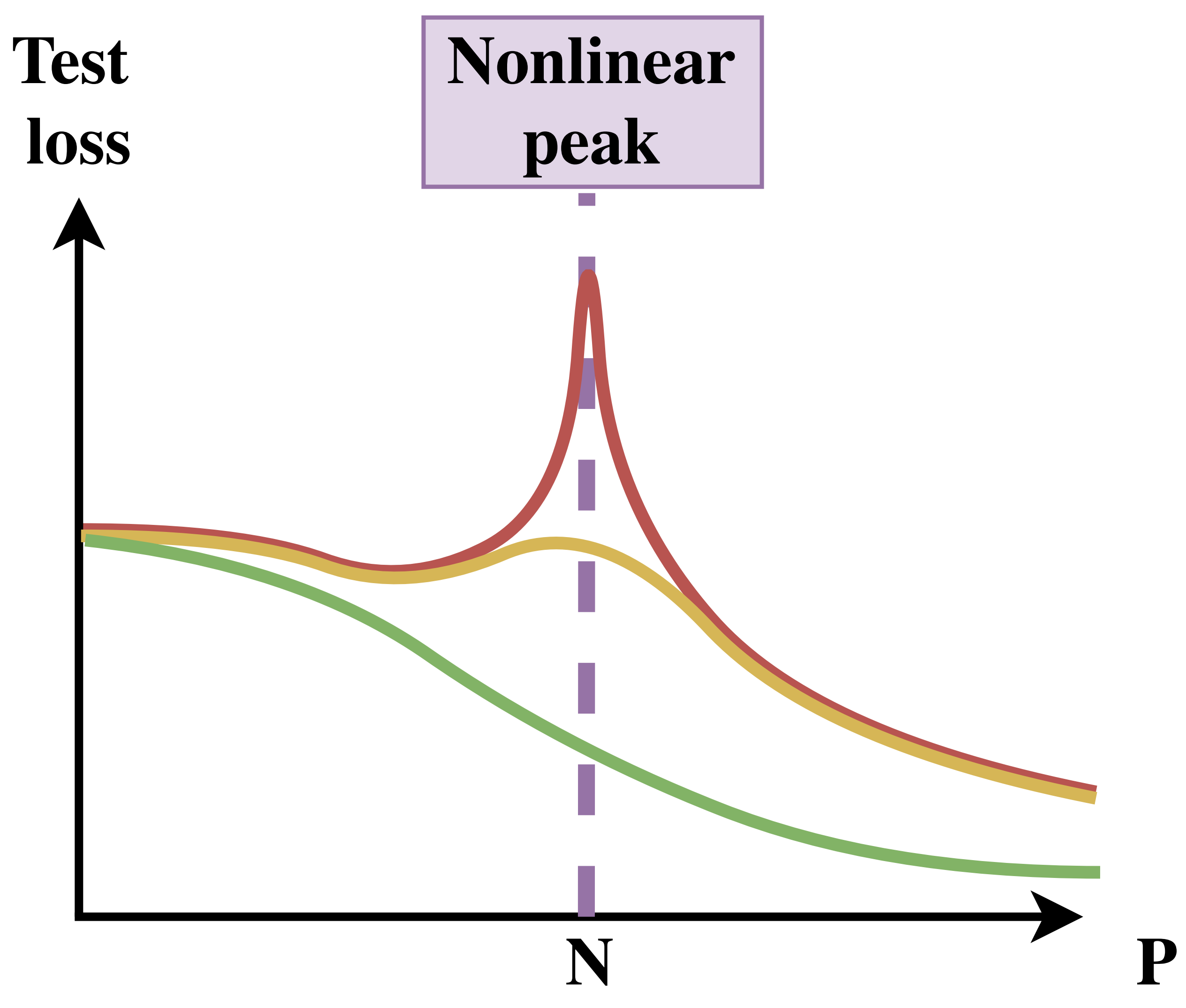
BOTH OCCUR AT INTERP THRESHOLD

ARE THEY THE SAME ?

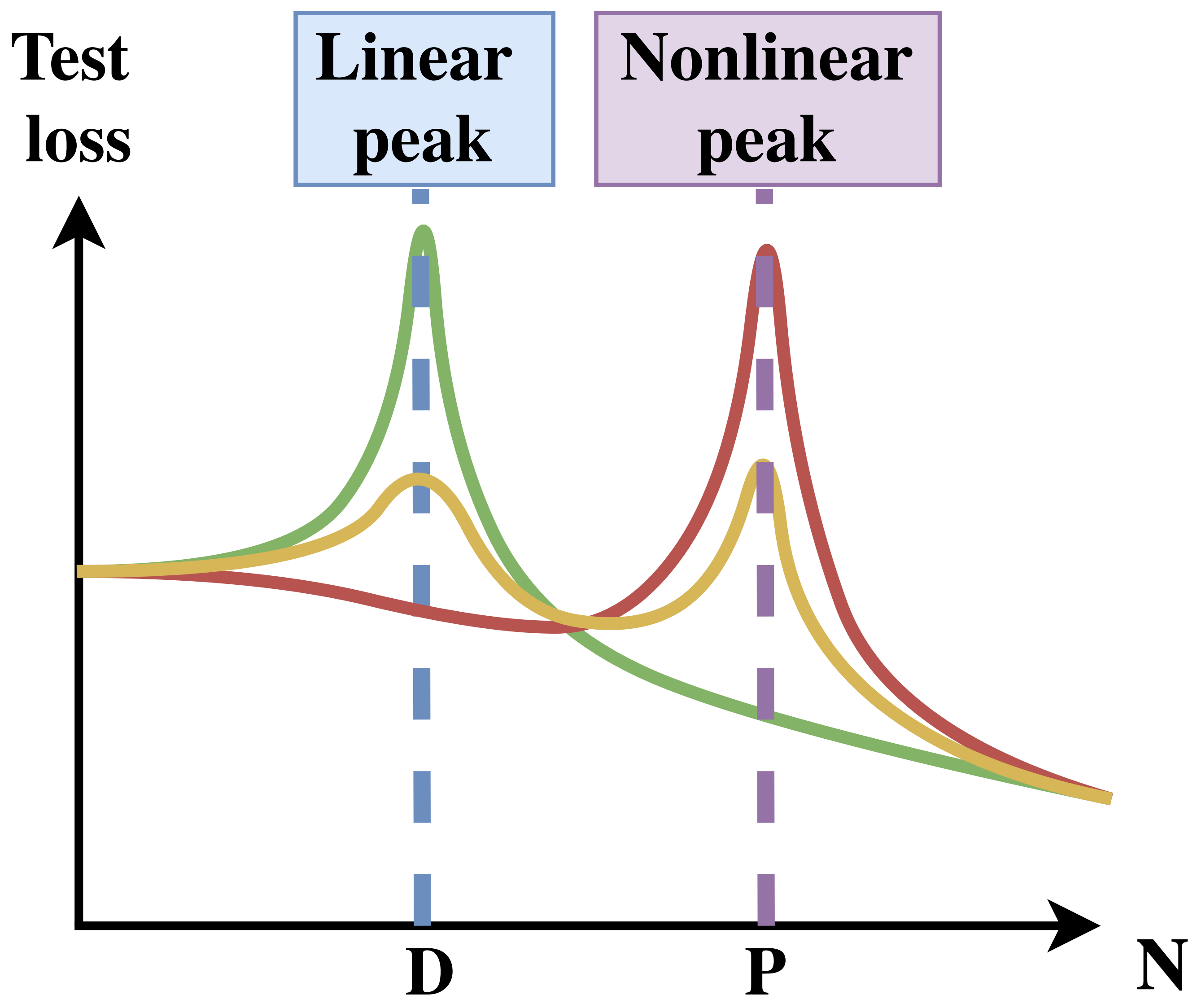
NO !

FROM LINEAR TO LINEAR

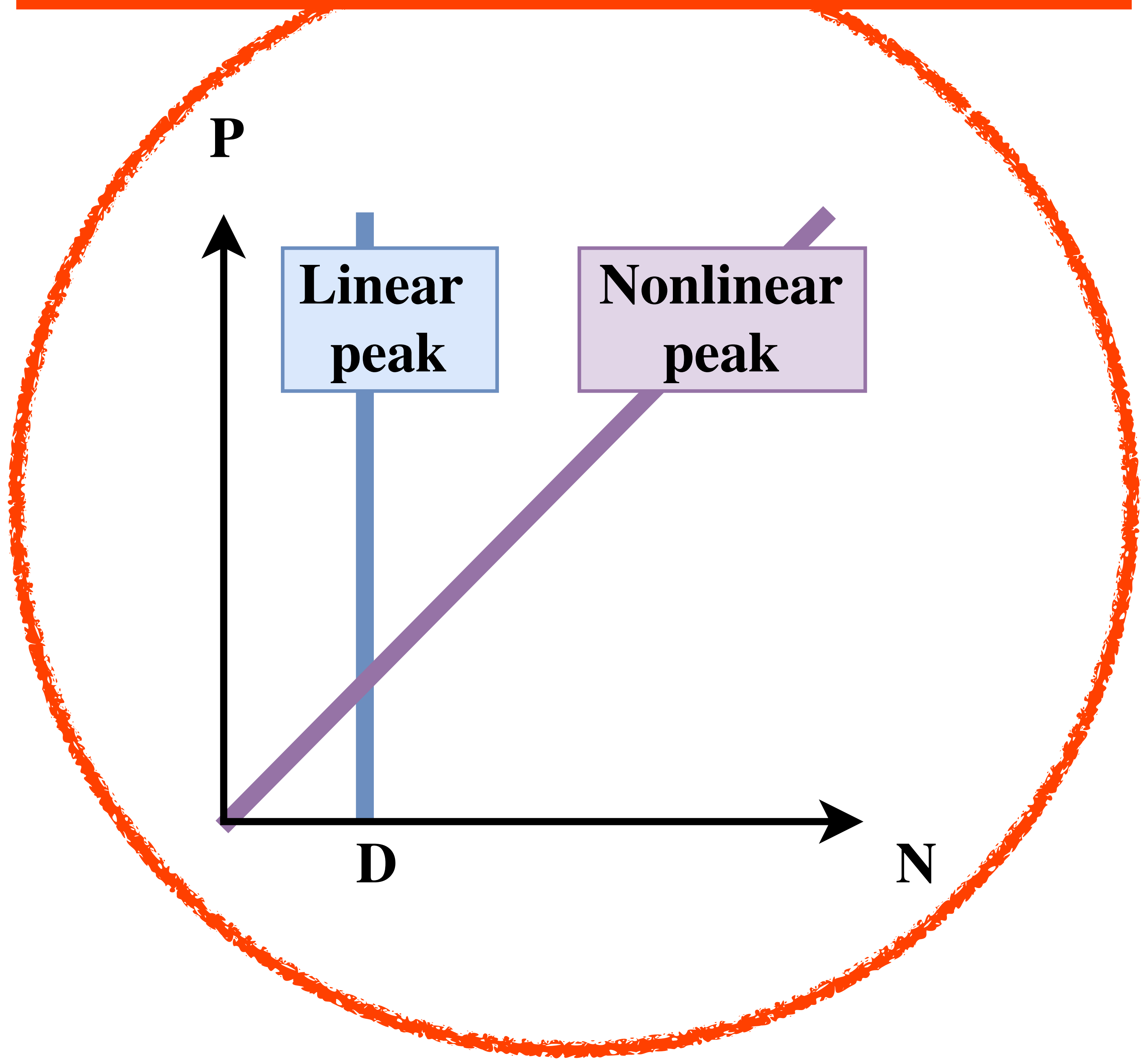
PARAMETER-WISE



SAMPLE-WISE



BOTH PEAKS COEXIST AT HIGH NOISE



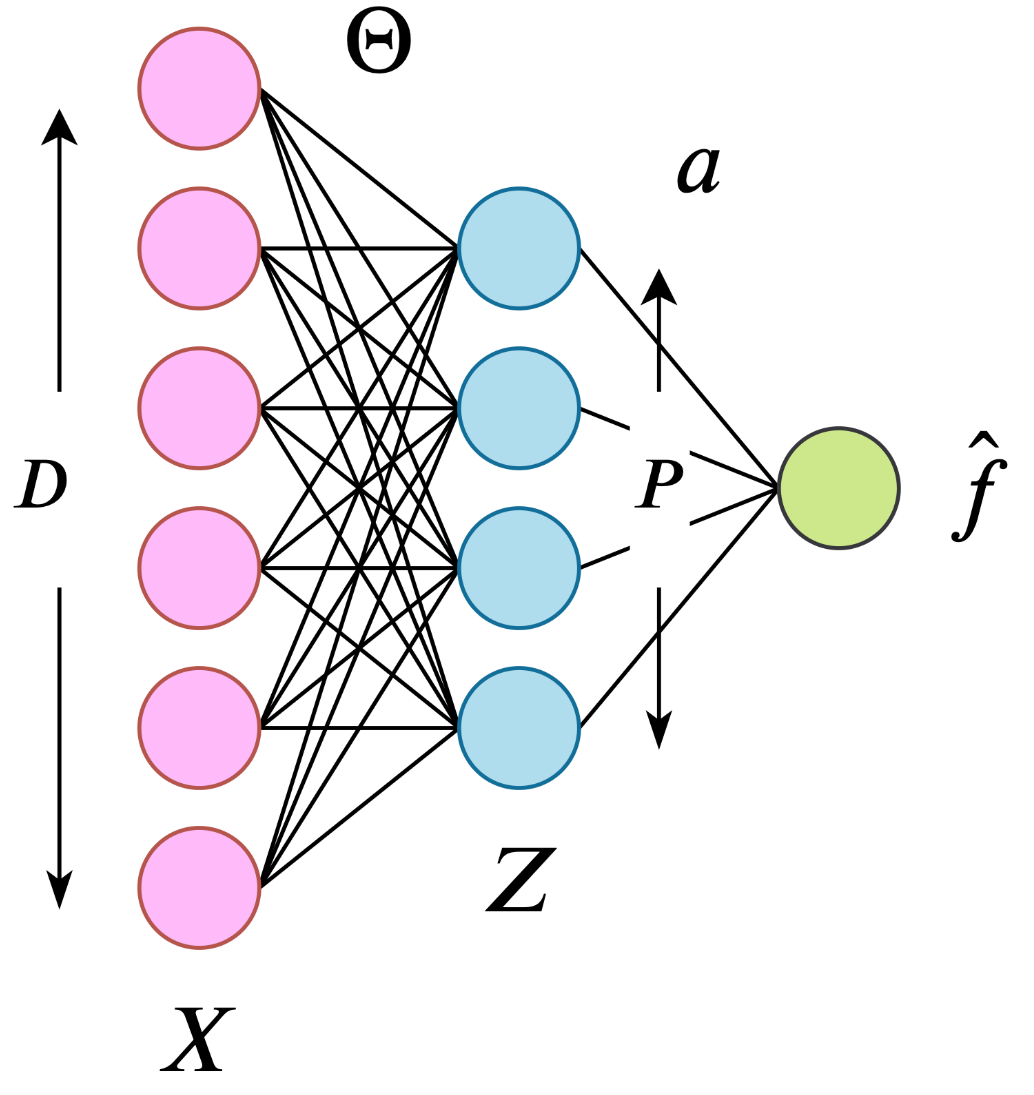
Activation function
Strongly nonlinear
Weakly nonlinear
Linear

WHAT MECHANISMS UNDERLIE THESE PEAKS ?
HOW ARE THEY DIFFERENT ?

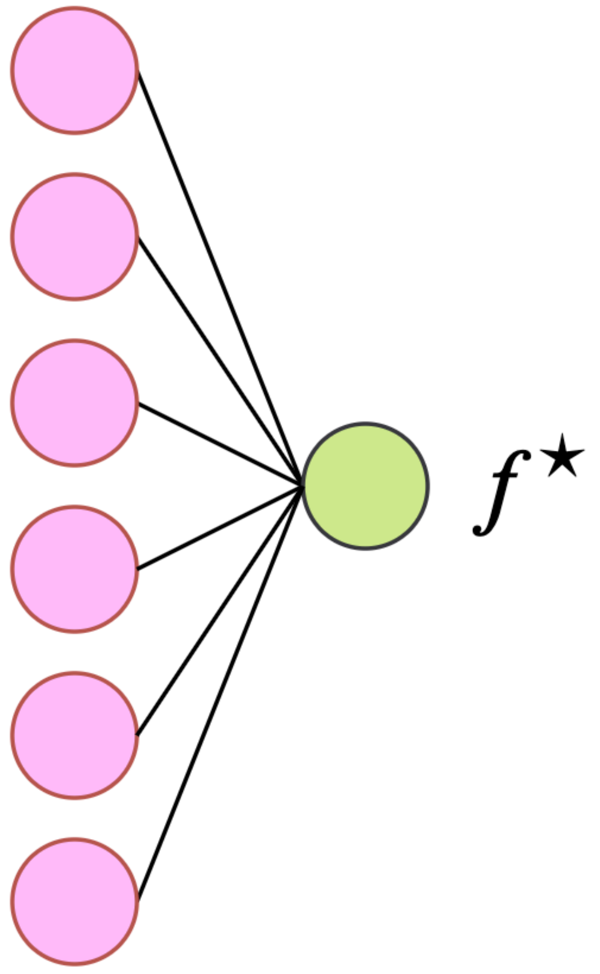
THE TWO MODELS

RF MODEL

RF STUDENT



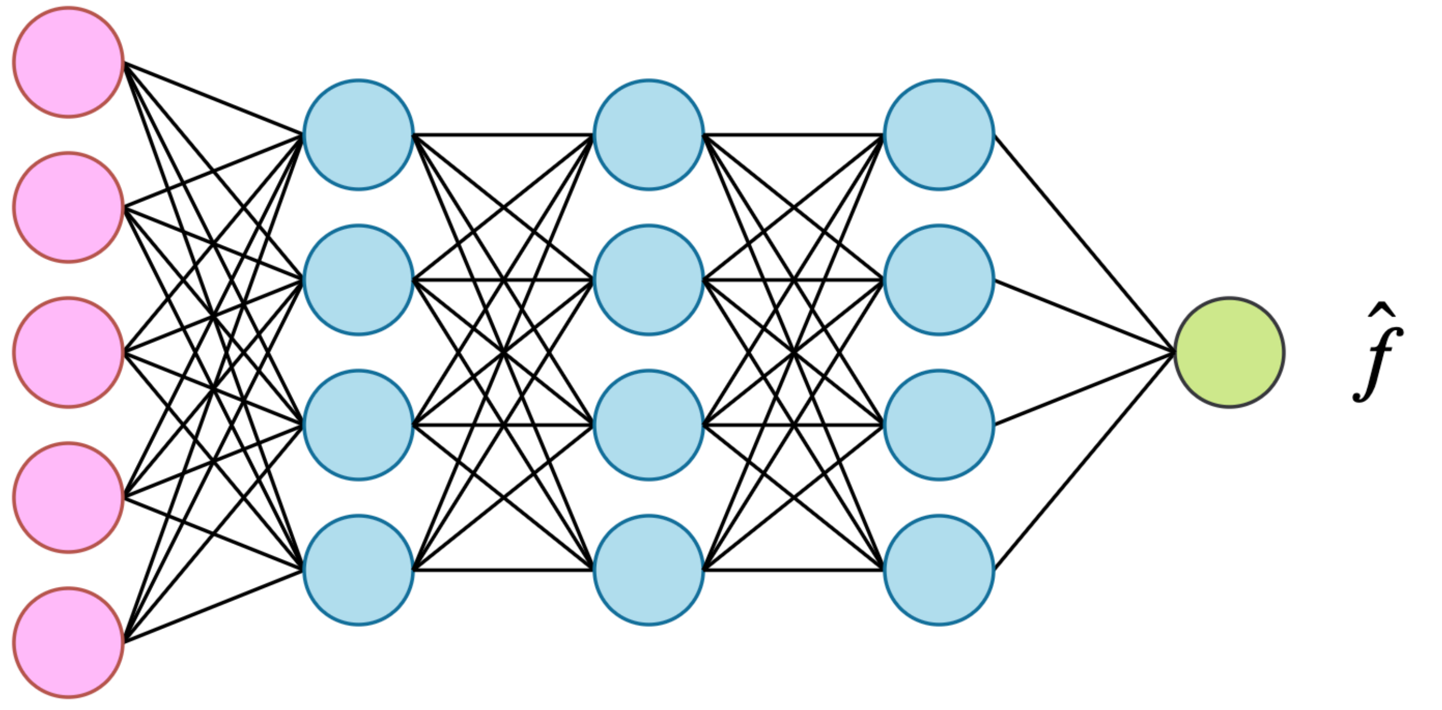
LINEAR TEACHER



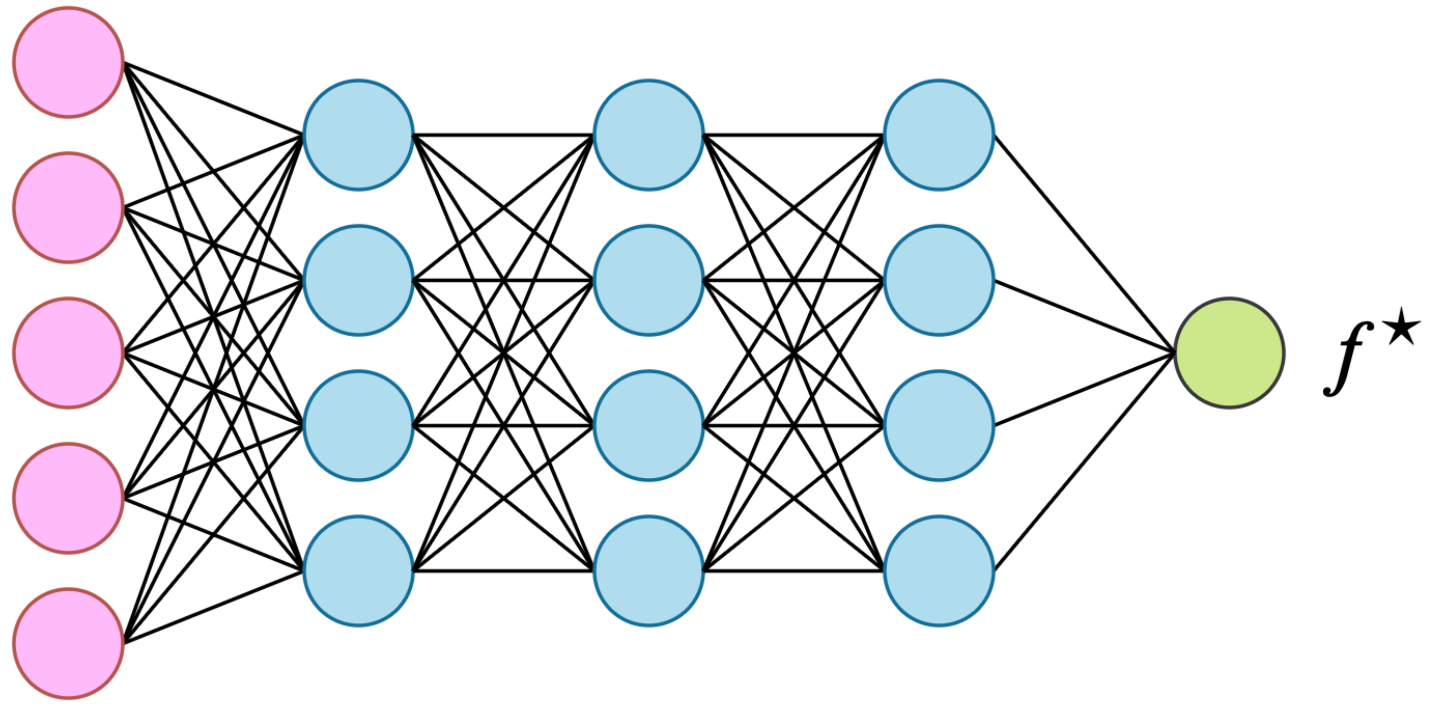
$$Z_i^\mu = \sigma\left(\frac{\langle \Theta_i, X_\mu \rangle}{\sqrt{D}}\right) \in \mathbb{R}^{N\times P}, \quad \Sigma = \frac{1}{N} Z^\top Z \in \mathbb{R}^{P\times P}$$

BAD CONDITIONING CAUSES PEAKS

DNN MODEL

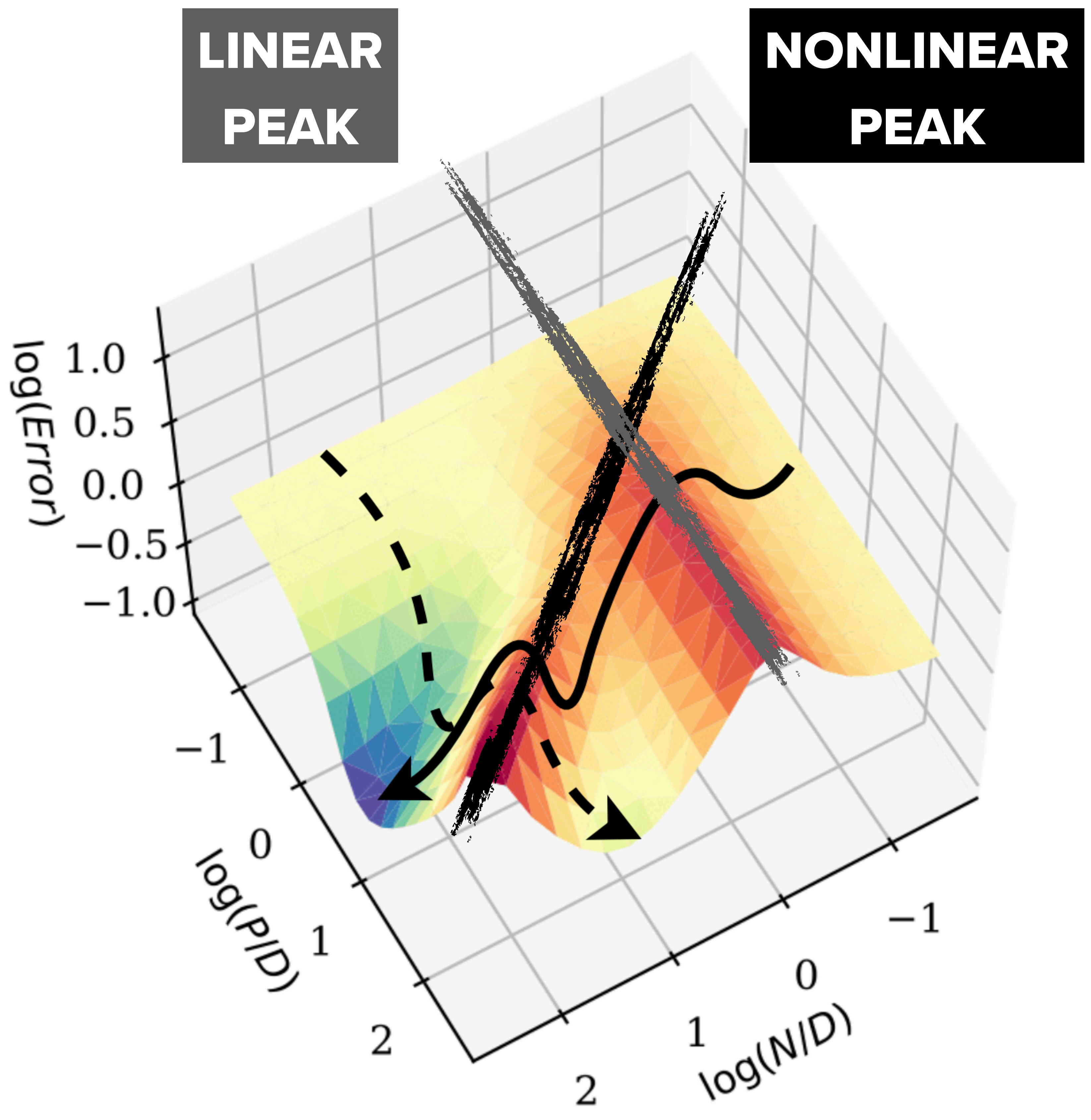


GD TRAINED STUDENT

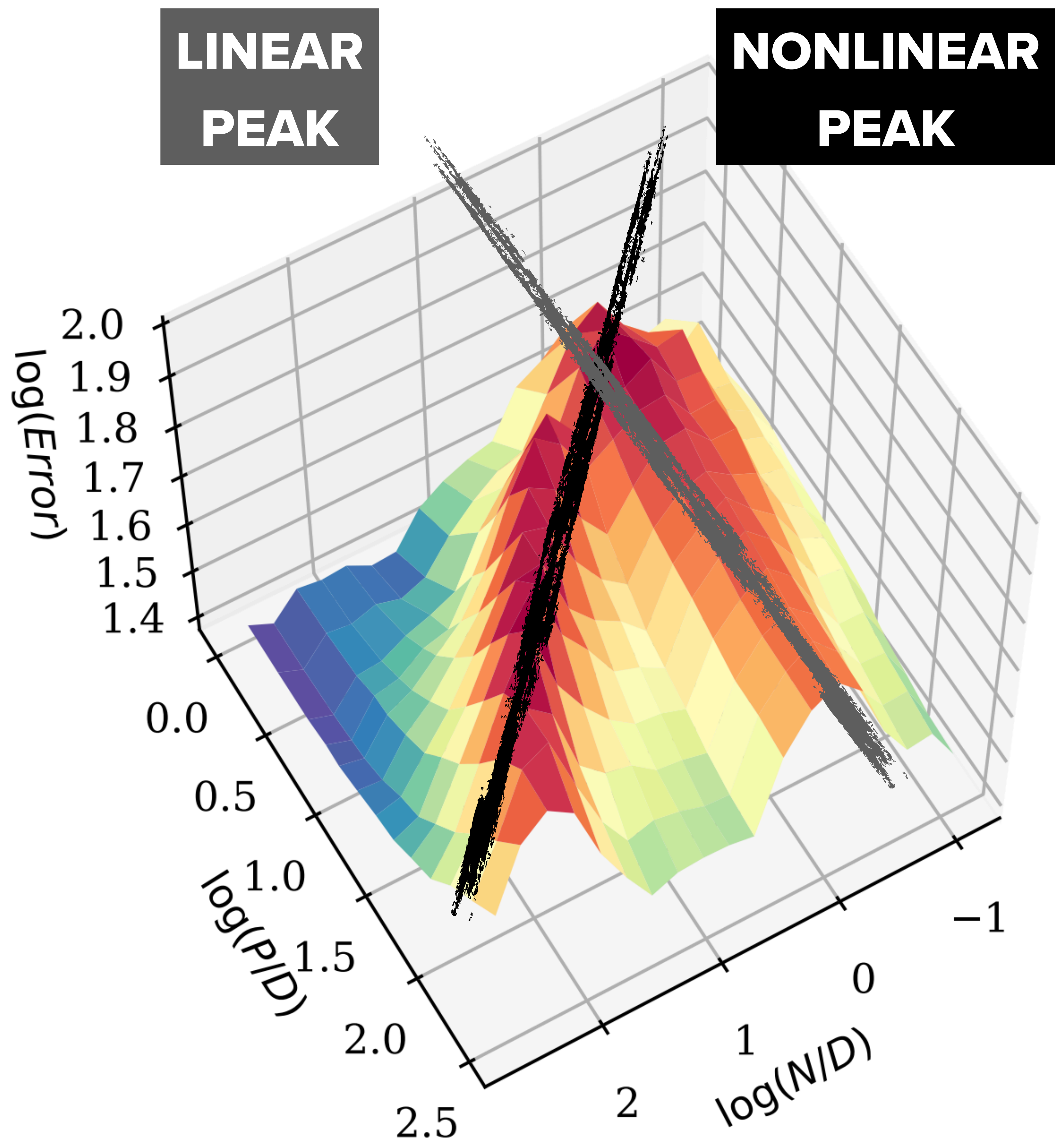


RANDOM TEACHER

EVIDENCE OF TRIPLE DESCENT



RF MODEL



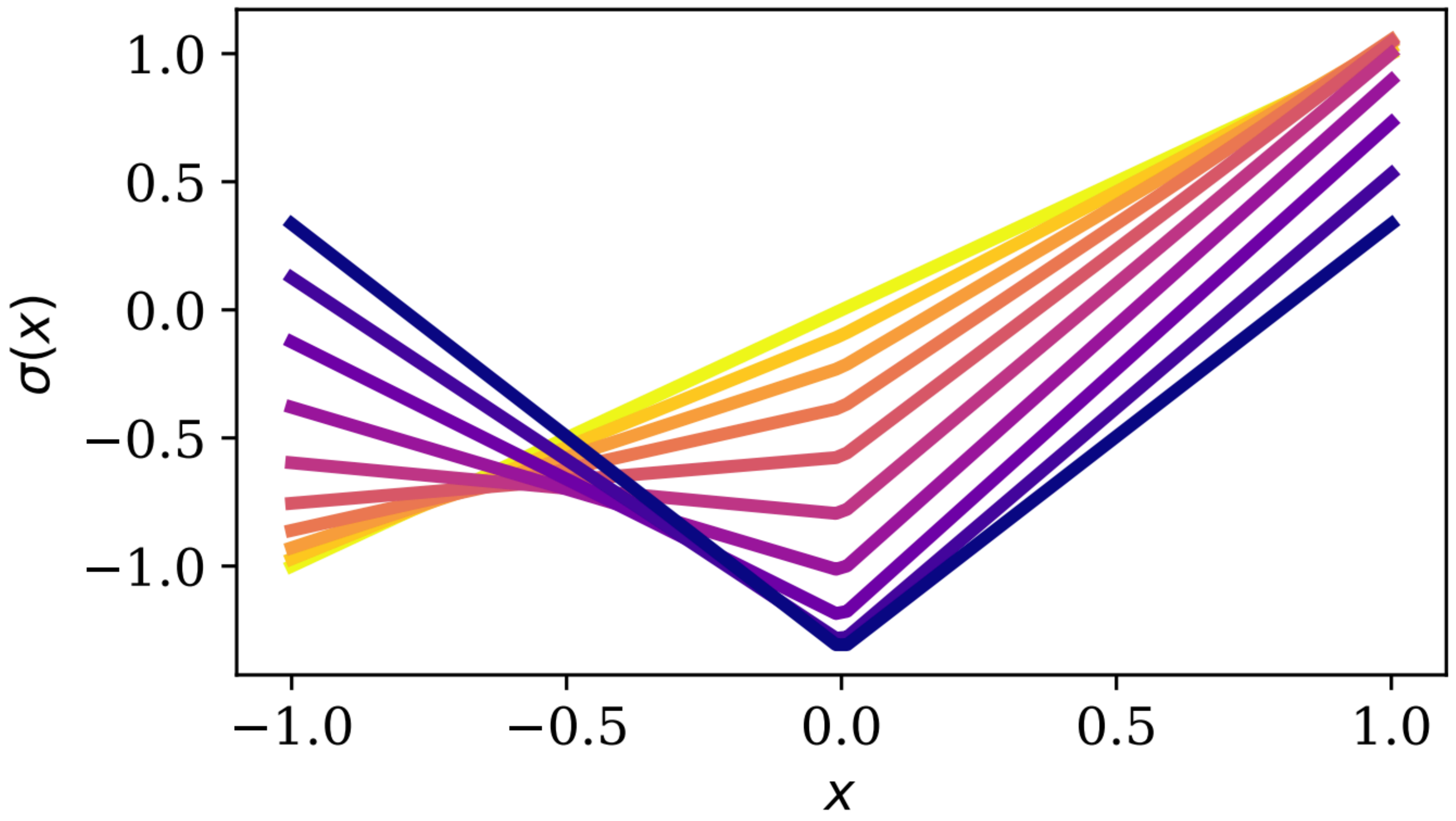
DNN MODEL

ANALYTICAL DESCRIPTION

HIGH-DIMENSIONAL LIMIT

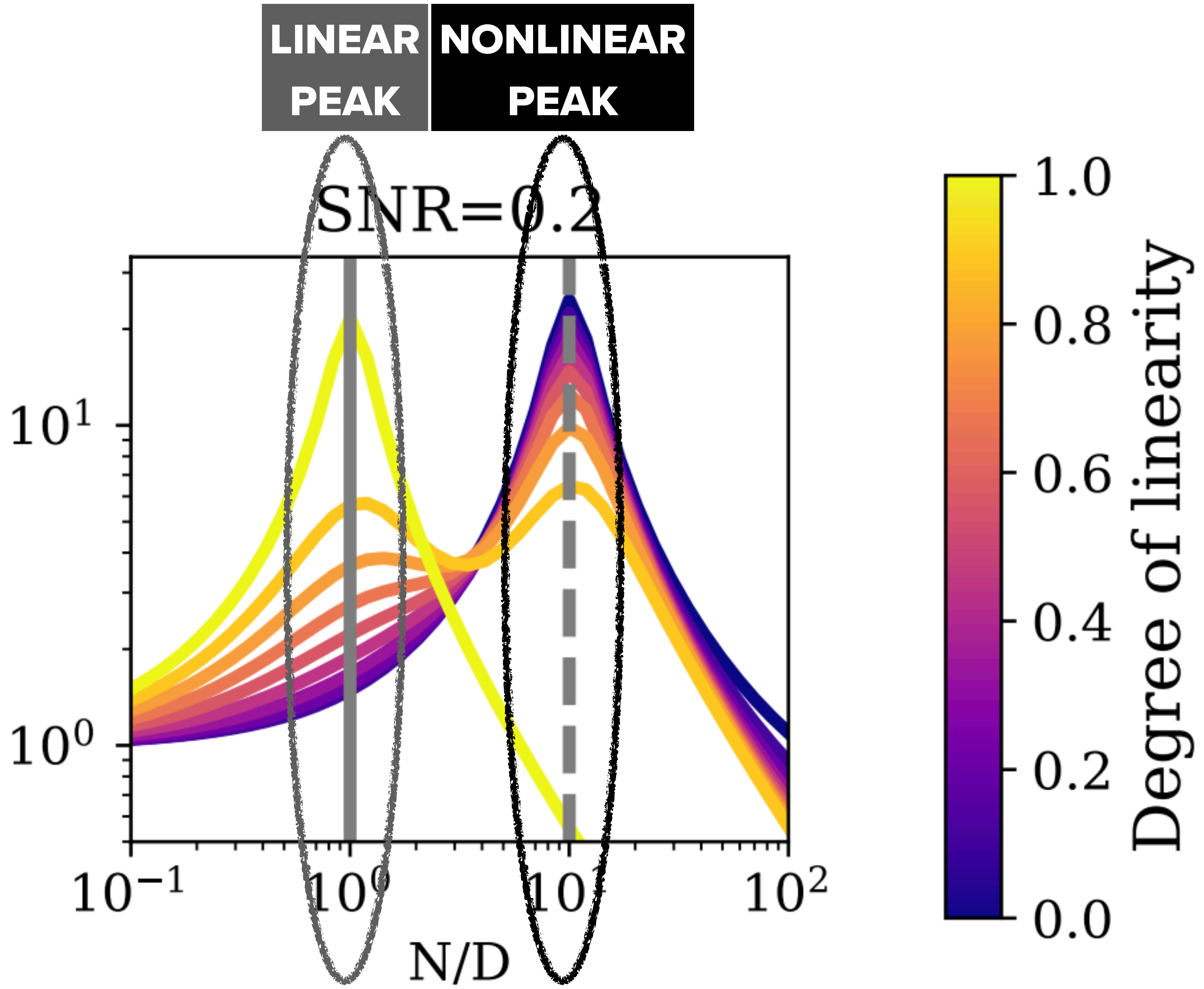
$$N, D, P \to \infty, \quad \frac{D}{P} = \mathcal{O}(1), \quad \frac{D}{N} = \mathcal{O}(1)$$

$$\eta = \int dz \frac{e^{-z^2/2}}{\sqrt{2\pi}} \sigma^2(z), \quad \zeta = \left[\int dz \frac{e^{-z^2/2}}{\sqrt{2\pi}} \sigma'(z) \right]^2$$



DEGREE OF LINEARITY

$$r = \frac{\zeta}{\eta}$$



ANALYTICAL SPECTRUM

NONLINEAR = LINEAR + NOISE

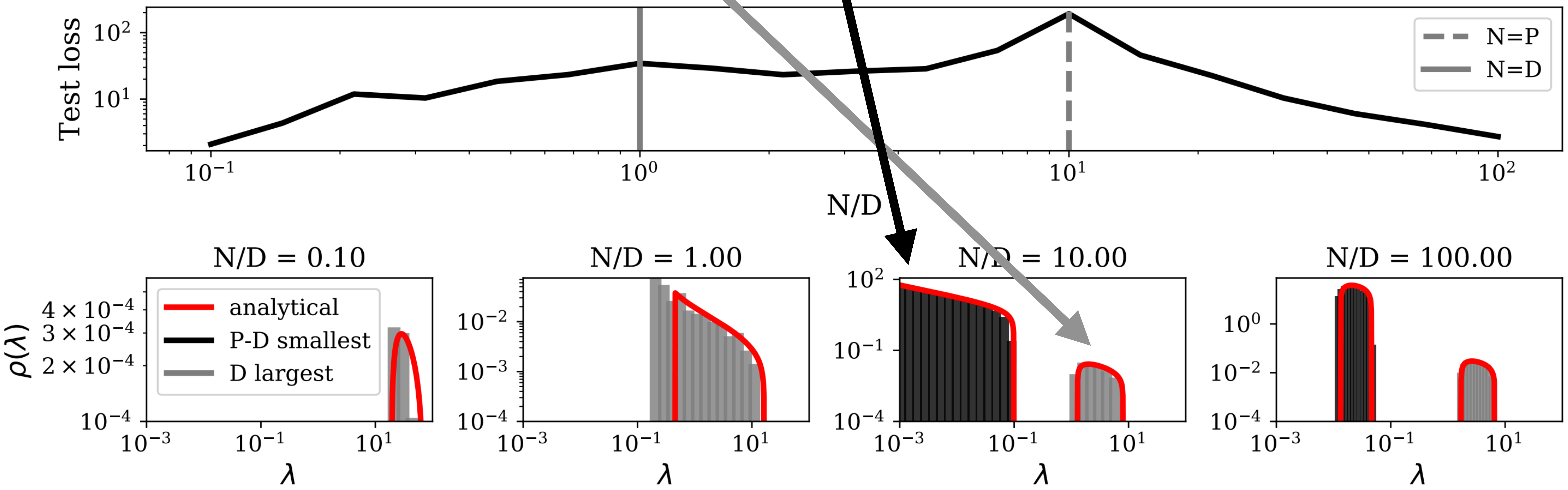
LINEAR PART **NONLINEAR PART**

$$\rho(\lambda) = \frac{1}{\pi} \lim_{\epsilon \to 0^+} \mathrm{Im} G(\lambda - i\epsilon), \qquad G(z) = \frac{\psi}{z} A\left(\frac{1}{z\psi}\right) + \frac{1-\psi}{z}$$

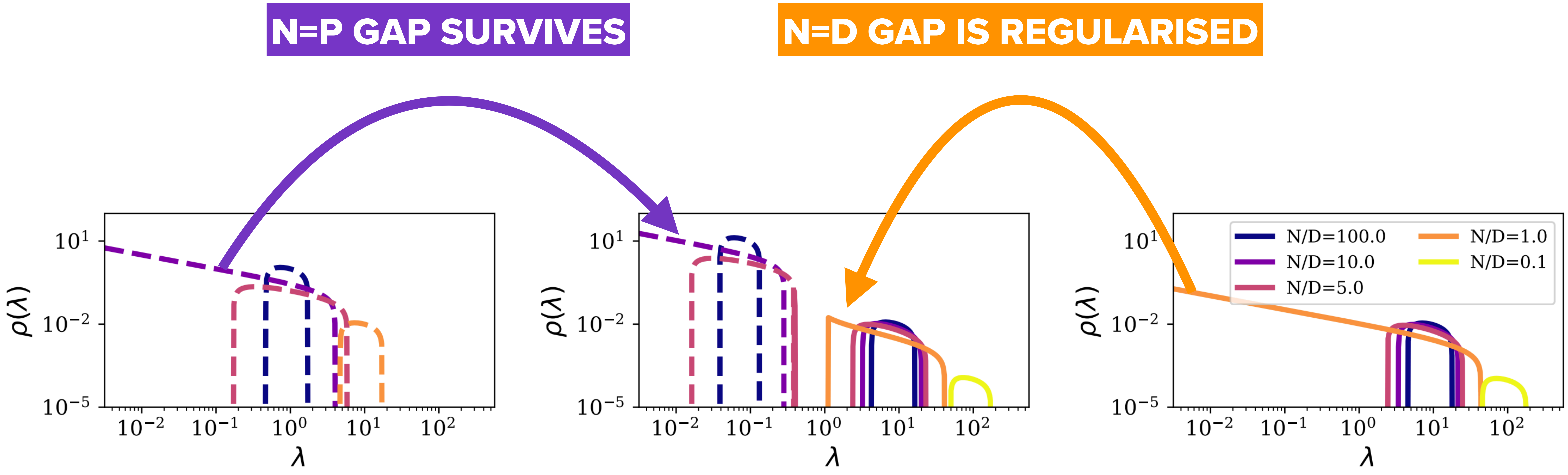
$$A(t) = 1 + (\eta - \zeta) t A_\phi(t) A_\psi(t) + \frac{A_\phi(t) A_\psi(t) t \zeta}{1 - A_\phi(t) A_\psi(t) t \zeta}$$

$$Z = \sigma\left(\frac{X\Theta^\top}{\sqrt{D}}\right) \to \sqrt{\zeta}\frac{X\Theta^\top}{\sqrt{D}} + \sqrt{\eta - \zeta}\, W, \quad W \sim \mathcal{N}(0,1)$$

[Pennington & Worah 2017]



ANALYTICAL SPECTRUM



(a) Absolute value ($r = 0$)

(b) Tanh ($r \simeq 0.92$)

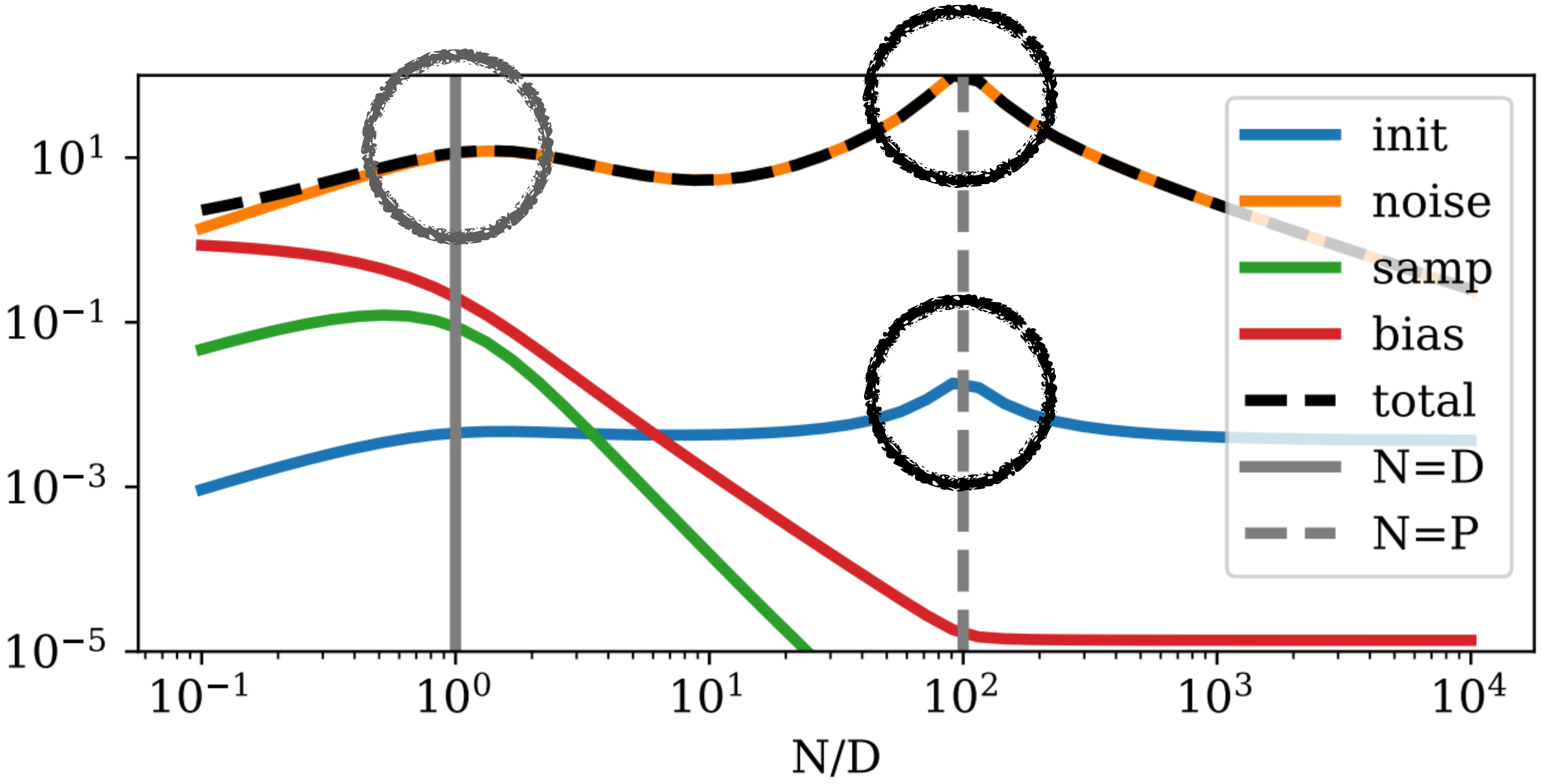
(c) Linear ($r = 1$)

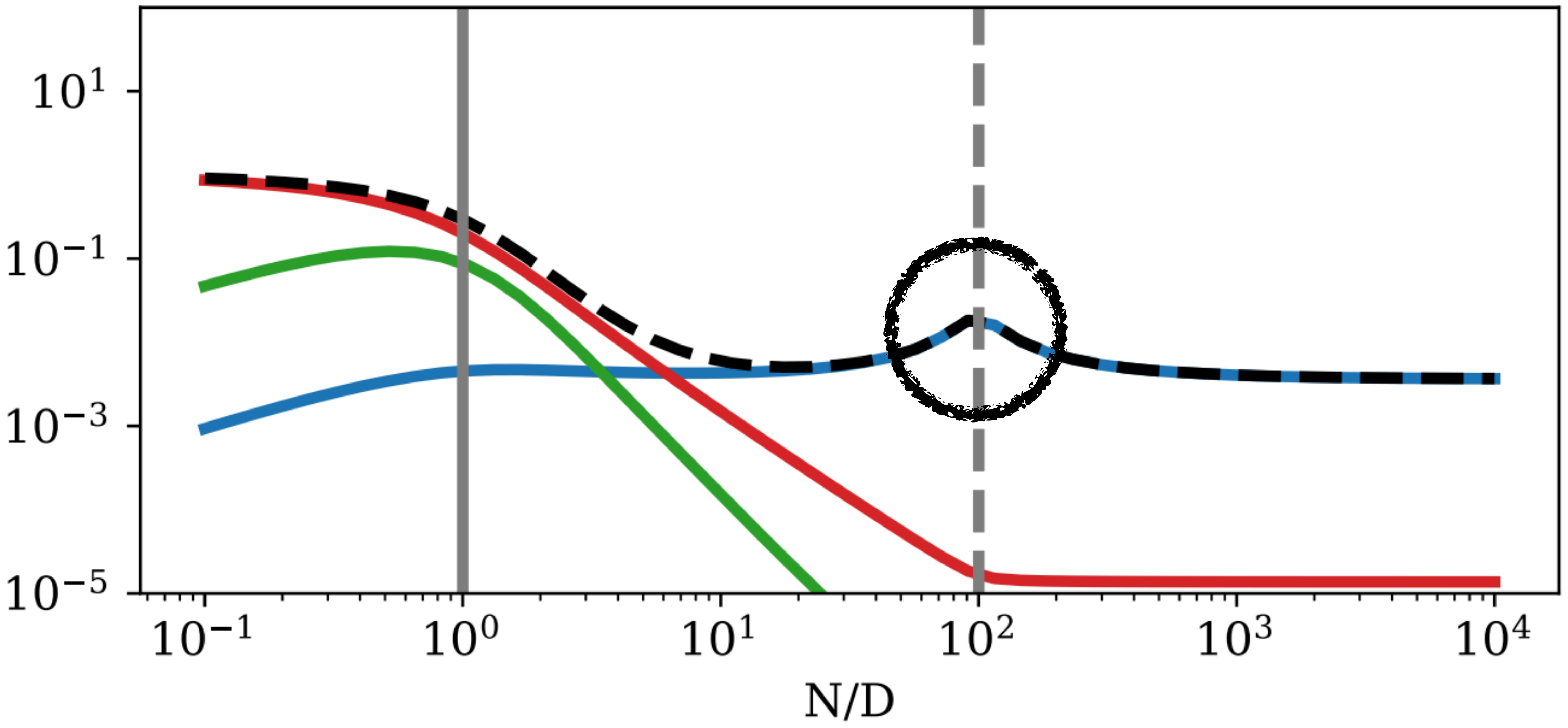
BIAS AND VARIANCES

LINEAR PEAK CAUSED BY NOISE

NONLINEAR PEAK CAUSED BY NOISE & INIT

NONLINEAR PEAK SURVIVES IN ABSENCE OF NOISE





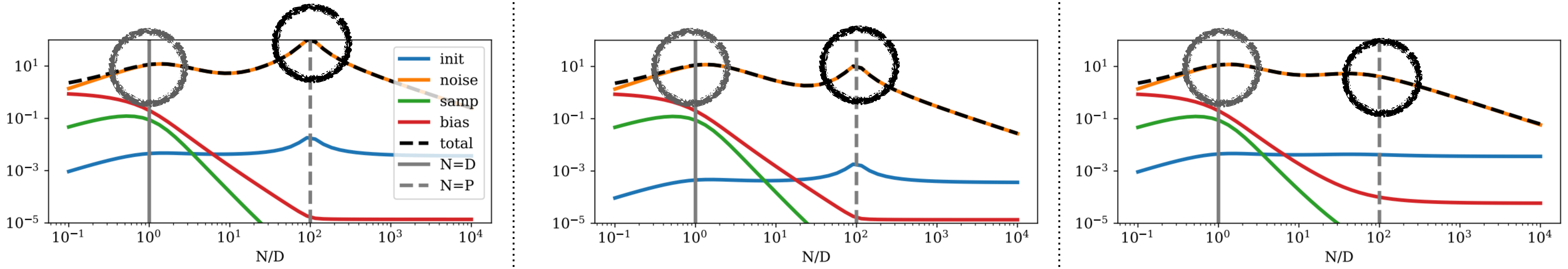
NOISY

NOISELESS

EFFECT OF REGULARISATION

VANILLA | **ENSEMBLING** | **REGULARIZING**

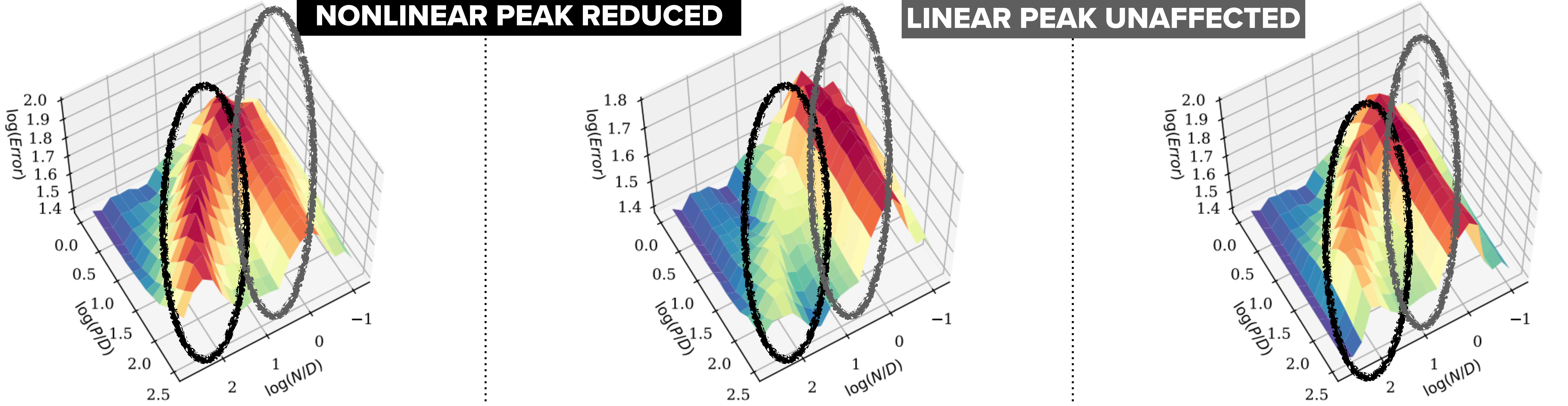
RF MODEL



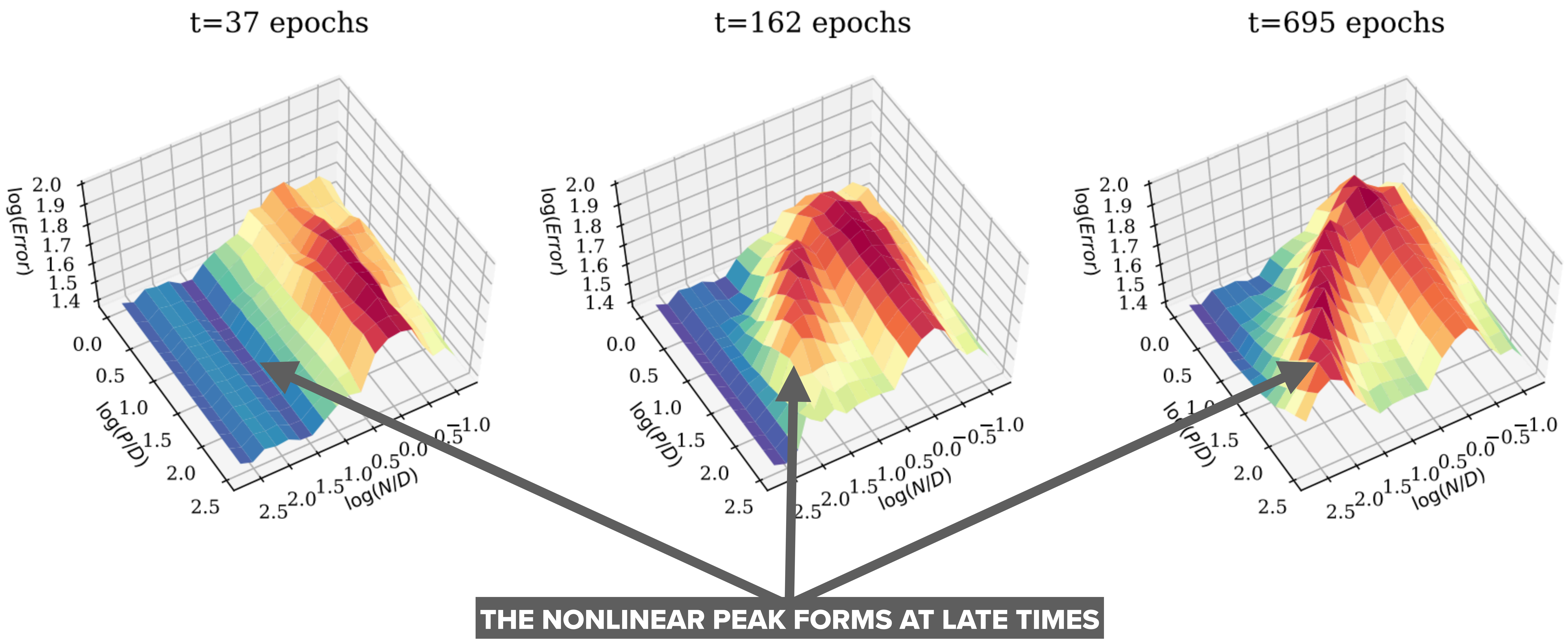
NONLINEAR PEAK REDUCED

LINEAR PEAK UNAFFECTED

DNN MODEL



TIME DEPENDENCE



THANK YOU

Stéphane d'Ascoli

Levent Sagun

Maria Refinetti

Giulio Biroli

Florent Krzakala

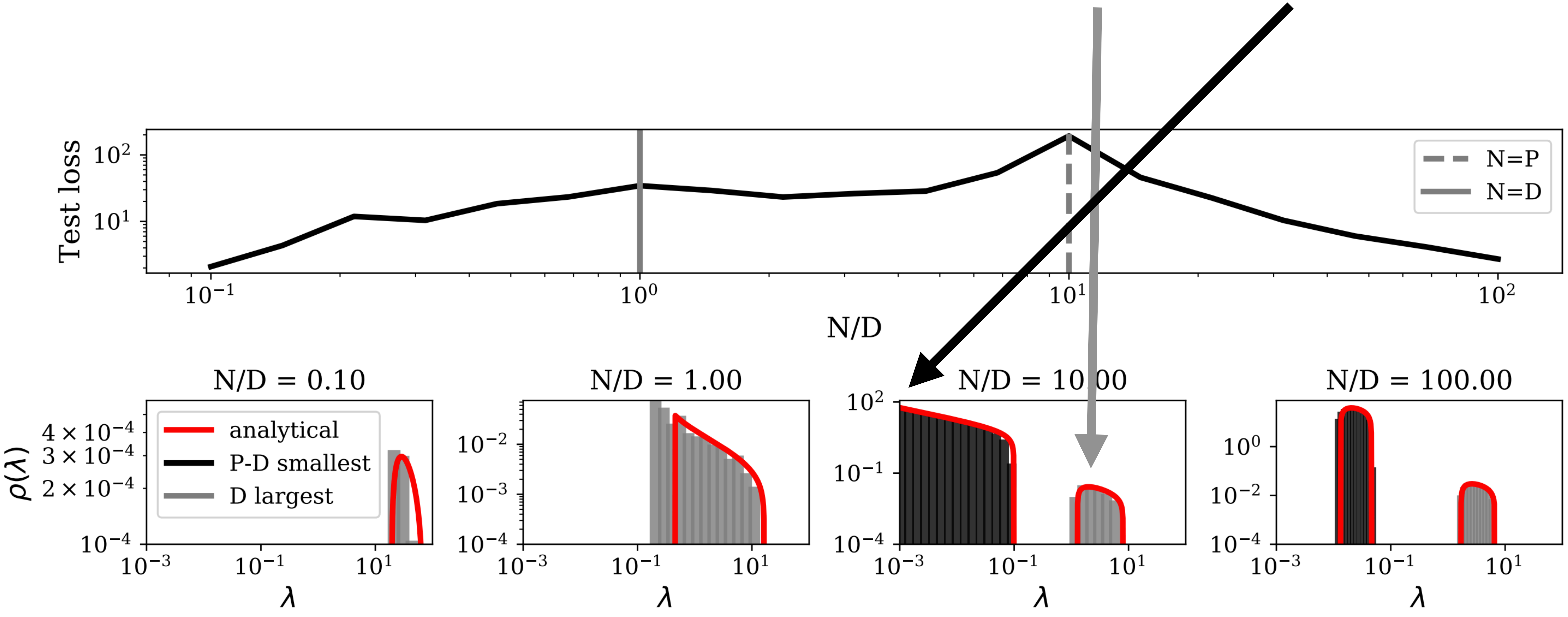
ANALYTICAL SPECTRUM

NONLINEAR = LINEAR + NOISE

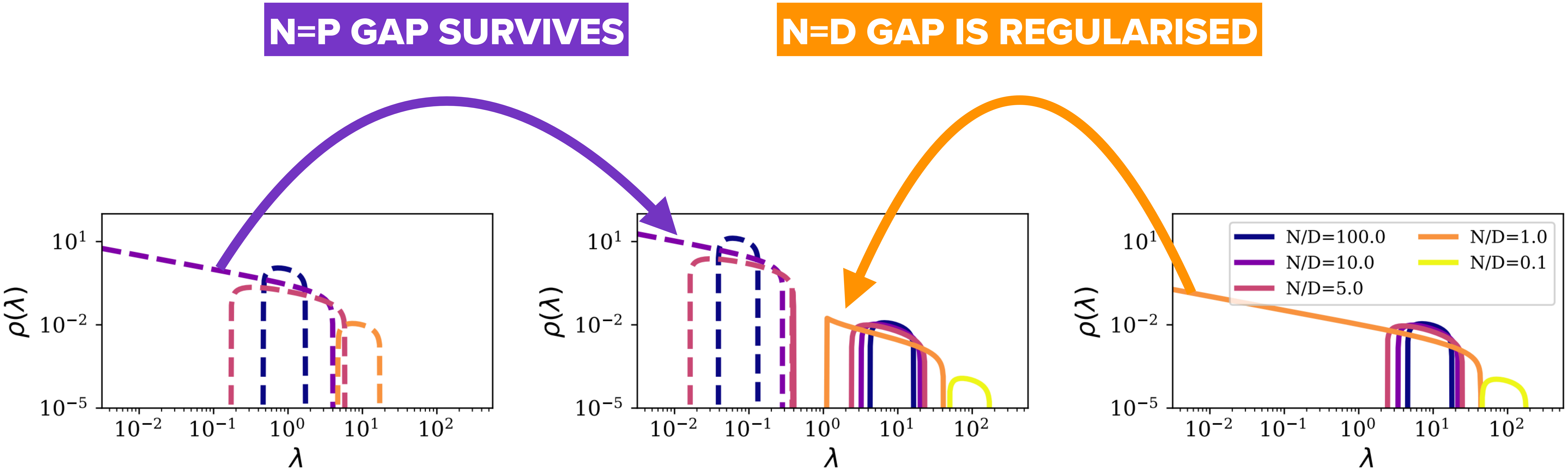
LINEAR PART

NONLINEAR PART

$$Z = \sigma\left(\frac{X\Theta^\top}{\sqrt{D}}\right) \rightarrow \sqrt{\zeta}\frac{X\Theta^\top}{\sqrt{D}} + \sqrt{\eta - \zeta}W, \quad W \sim \mathcal{N}(0,1)$$



ANALYTICAL SPECTRUM

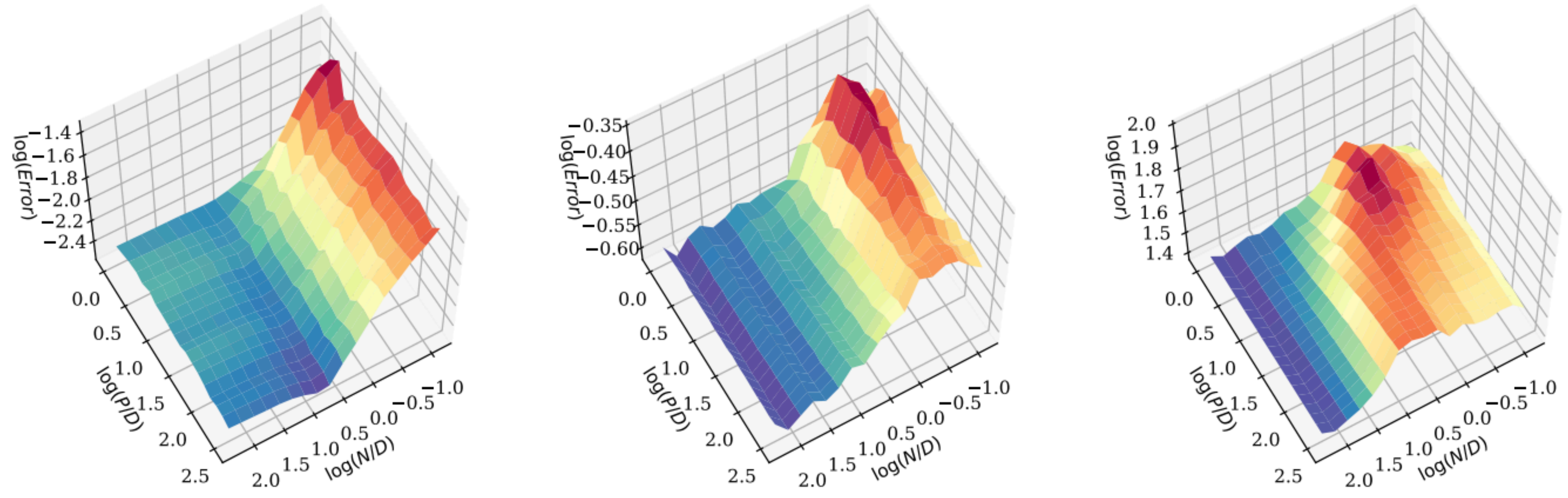


(a) Absolute value ($r = 0$)

(b) Tanh ($r \simeq 0.92$)

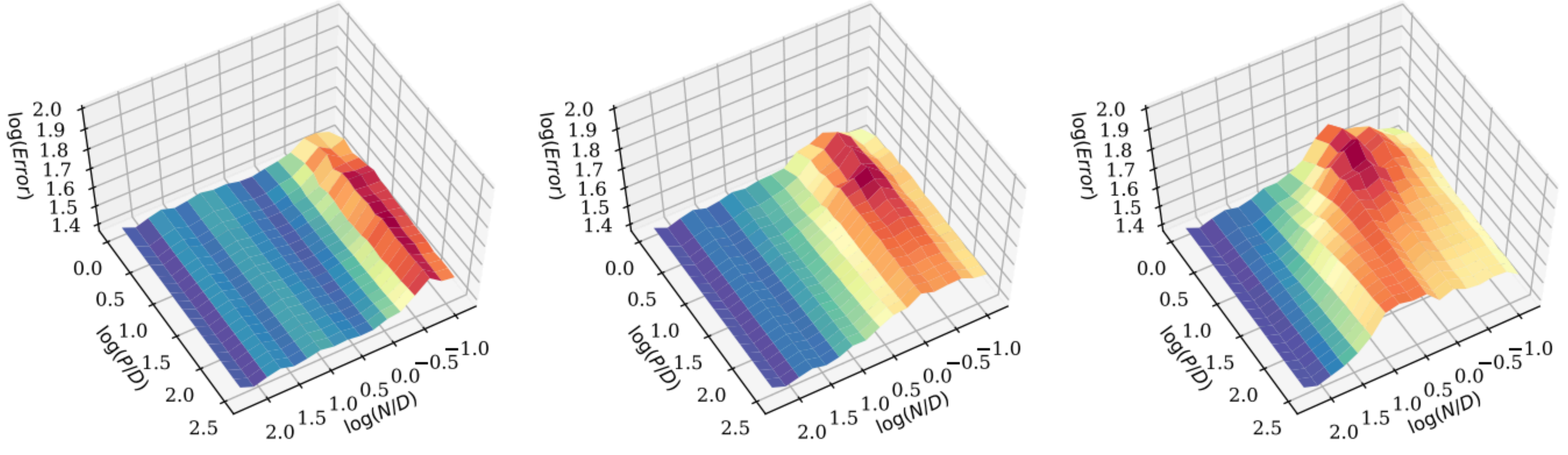
(c) Linear ($r = 1$)

STRUCTURED DATASETS



(a) MNIST, $SNR = \infty$ (b) MNIST, $SNR = 2$ (c) MNIST, $SNR = 0.2$

t=37 epochs t=162 epochs t=695 epochs

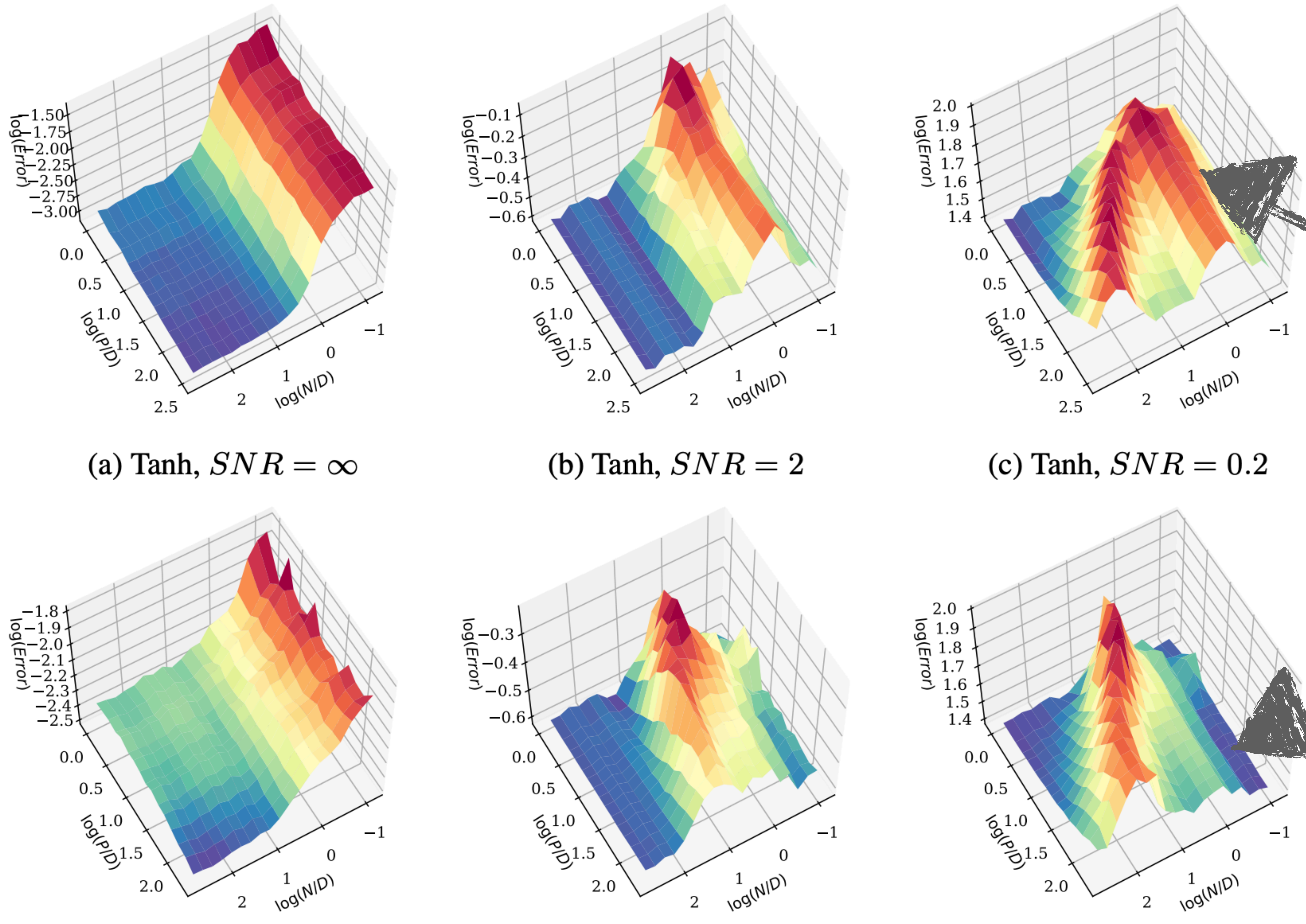


(d) Dynamics on MNIST at $SNR = 0.2$

LINEAR AND NONLINEAR PEAK ARE MERGED TOGETHER

SHIFT FROM LINEAR TO NONLINEAR DURING TRAINING

EFFECT OF NOISE AND NONLINEARITY



(a) Tanh, $SNR = \infty$ (b) Tanh, $SNR = 2$ (c) Tanh, $SNR = 0.2$

(d) ReLU, $SNR = \infty$ (e) ReLU, $SNR = 2$ (f) ReLU, $SNR = 0.2$

LINEAR PEAK IS WEAKER FOR RELU