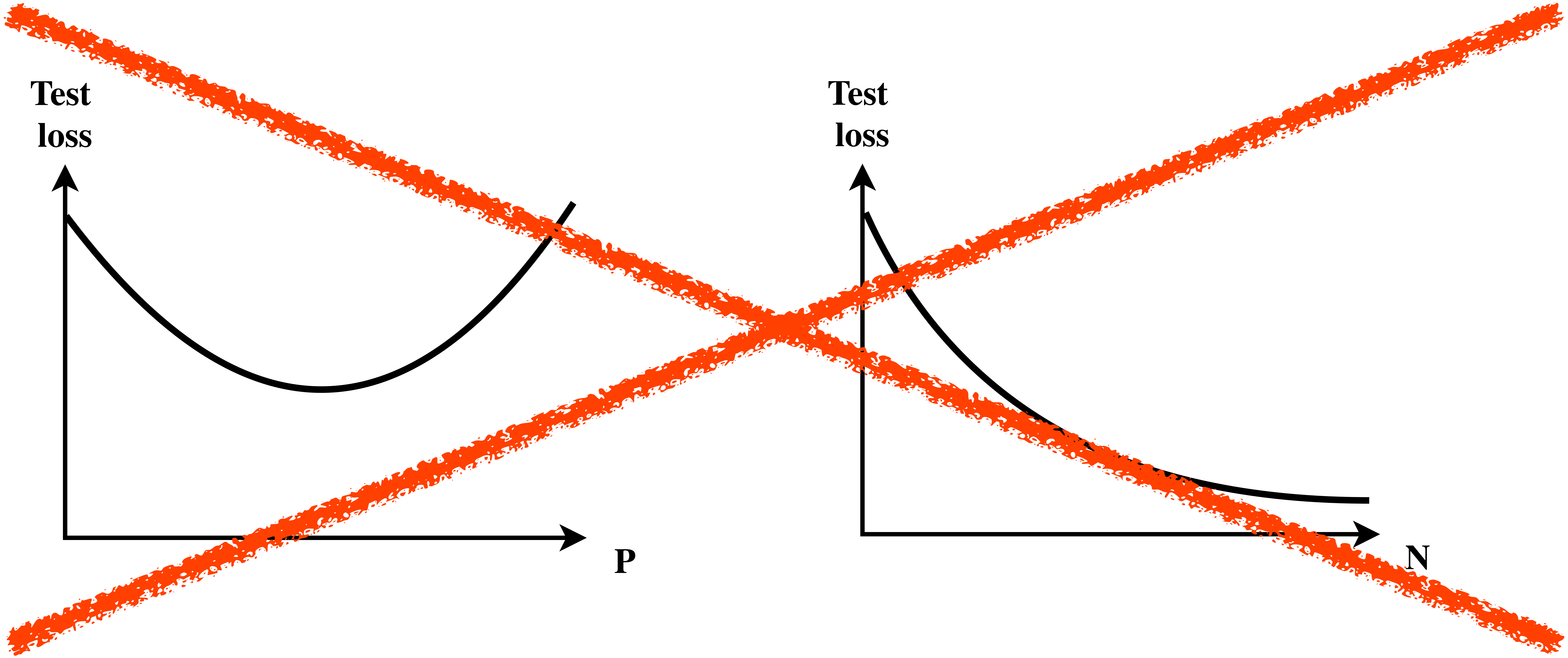
TRIPLE DESCENT : THE TWO KINDS OF OVERFITTING

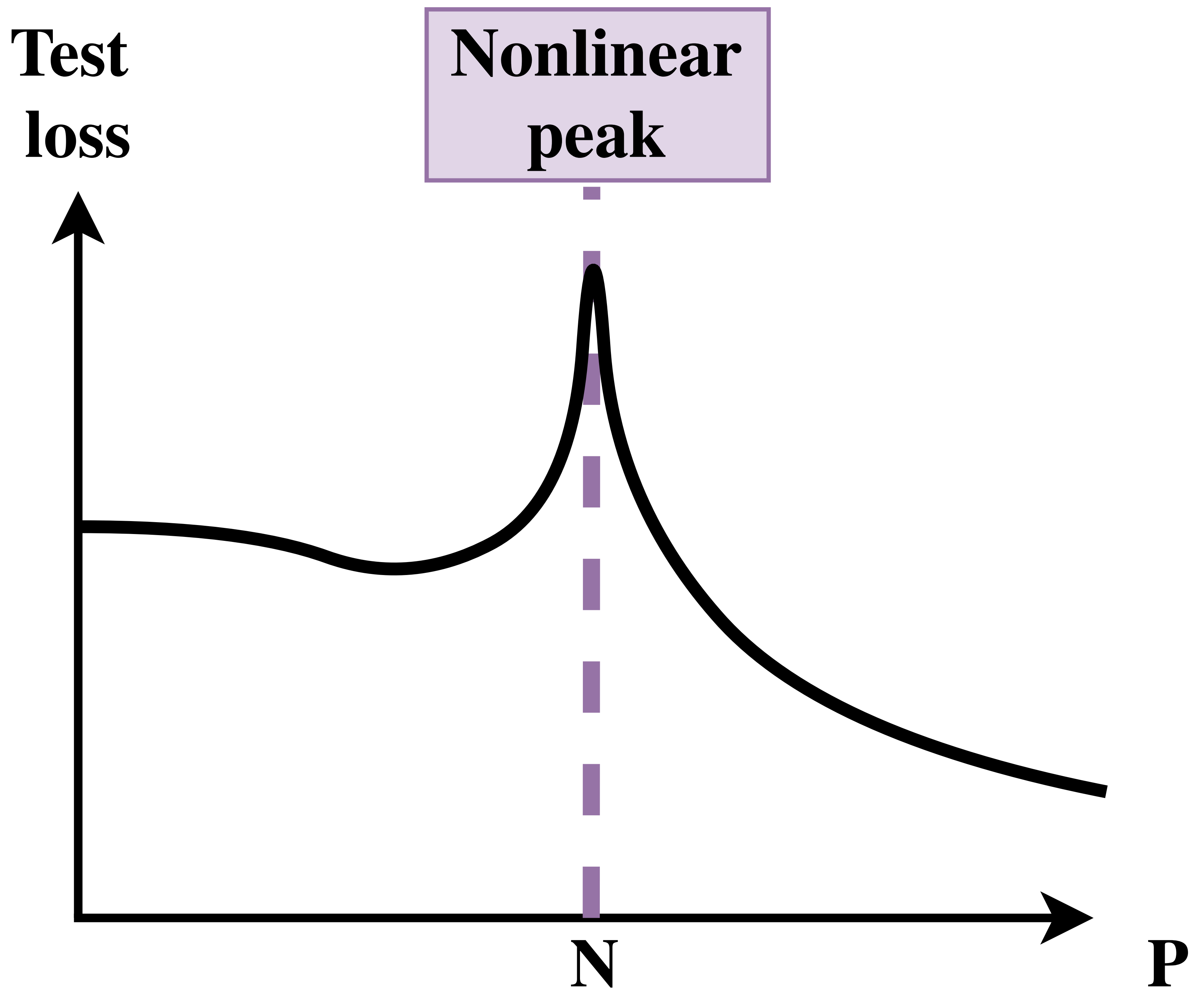
STÉPHANE D'ASCOLI, LEVENT SAGUN, GIULIO BIROLI

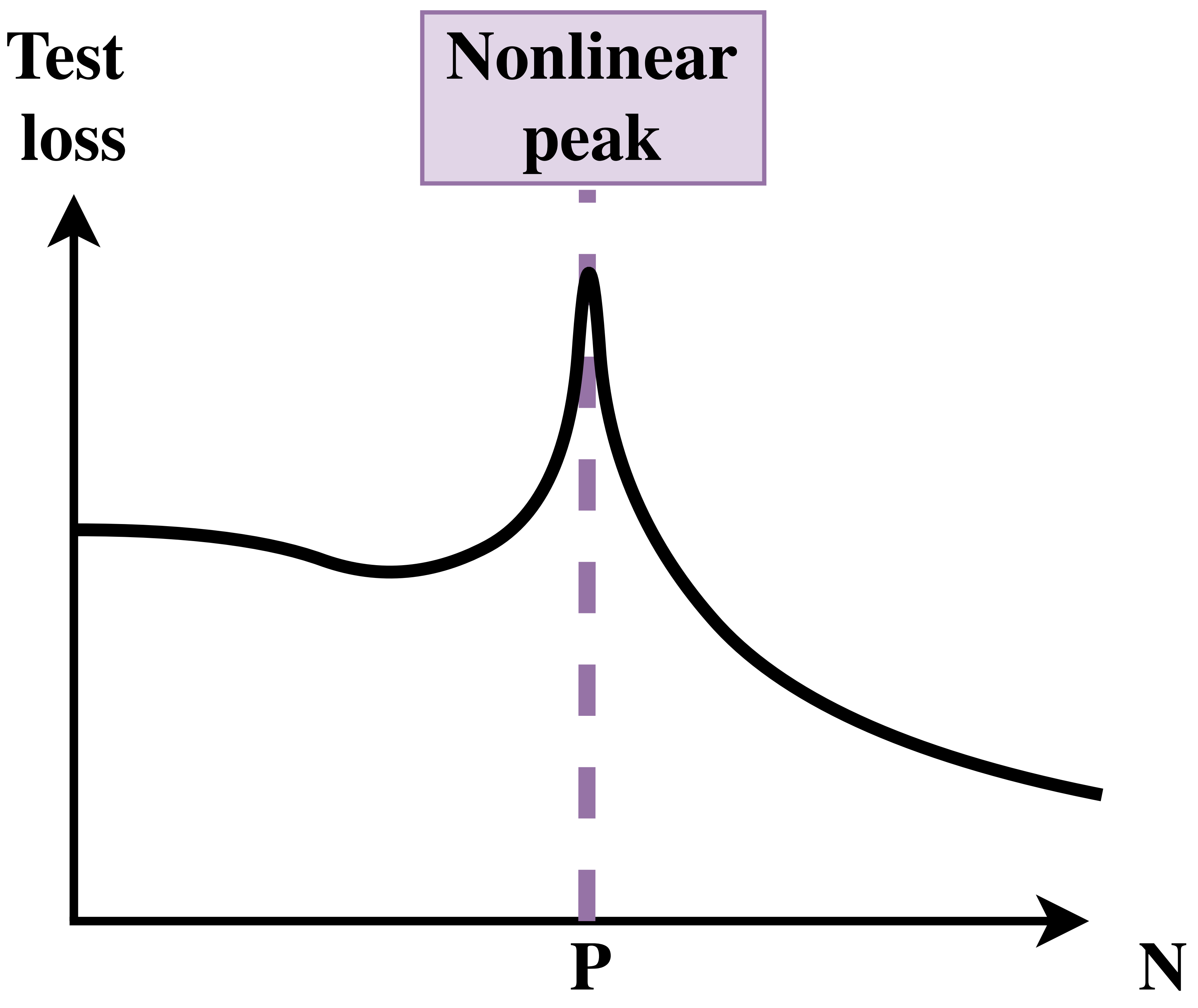
ÉCOLE NORMALE SUPÉRIEURE & FACEBOOK AI RESEARCH

PARAMETER-WISE AND SAMPLE-WISE



PARAMETER-WISE AND SAMPLE-WISE



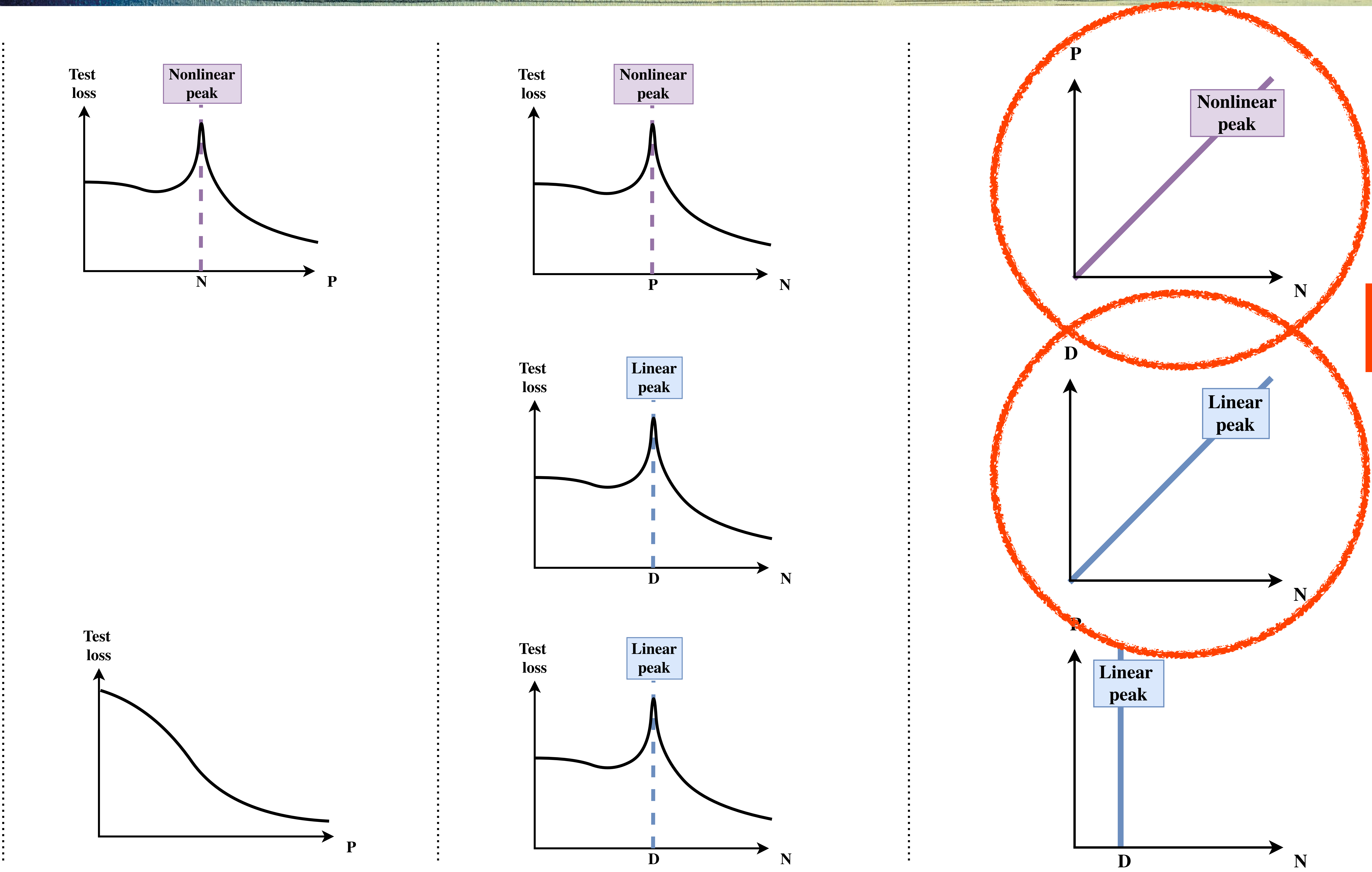


FOR LINEAR TO NONLINEAR

NONLINEAR NETWORKS

LINEAR MODELS

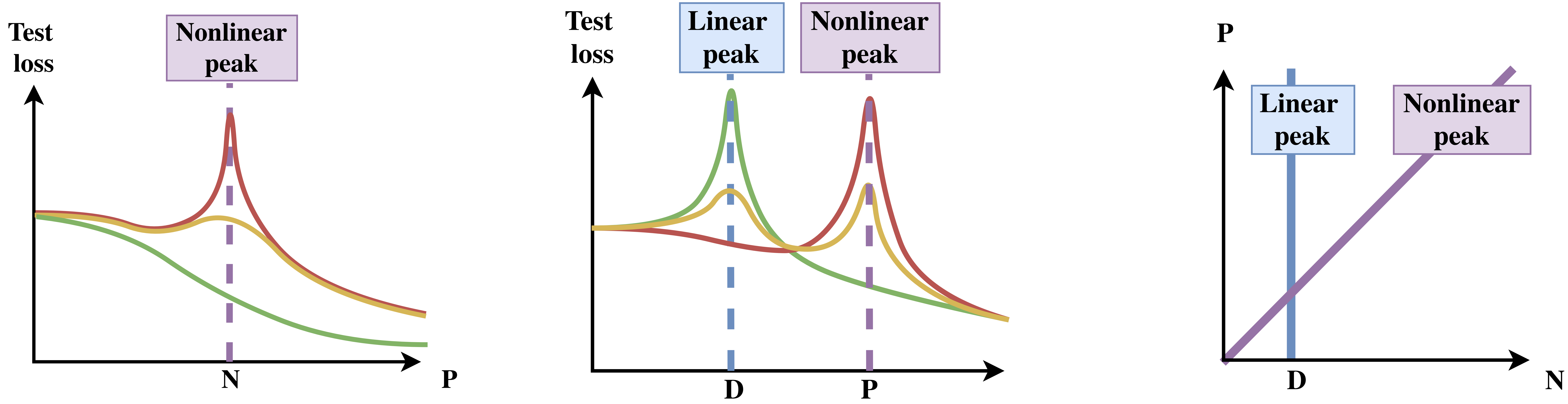
LINEAR NETWORKS

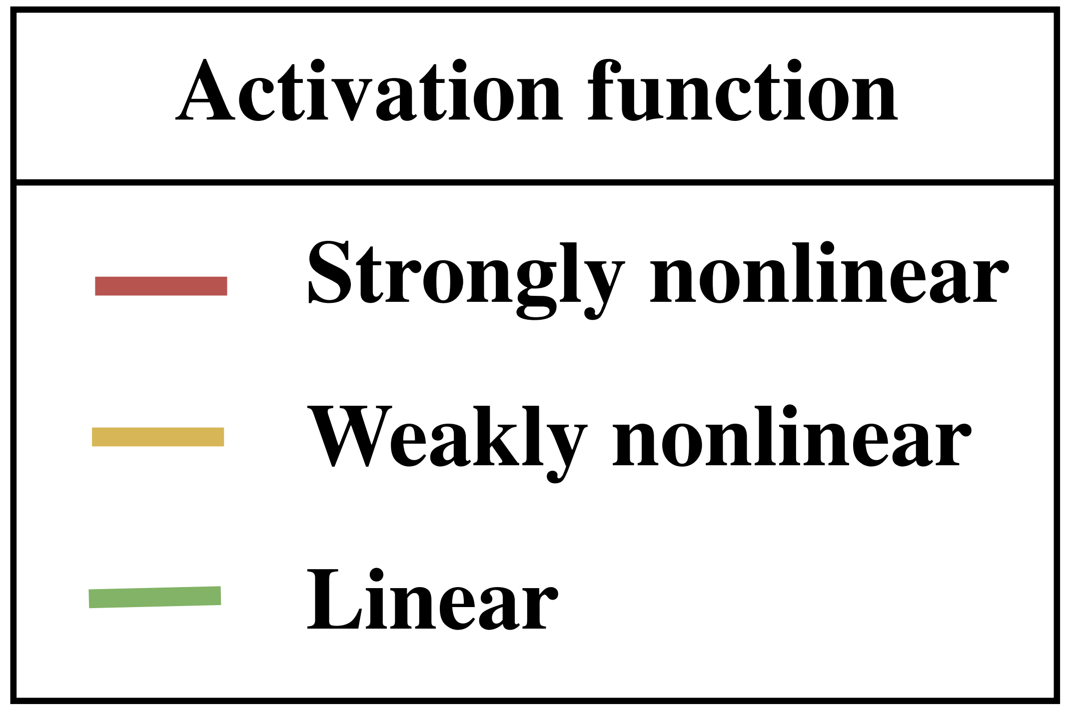


ARE THESE THE SAME ?

ANSWER : NO !

FROM LINEAR TO LINEAR





WHAT MECHANISMS UNDERLIE THESE PEAKS ?
HOW ARE THEY DIFFERENT ?

THE TWO MODELS

DATASET

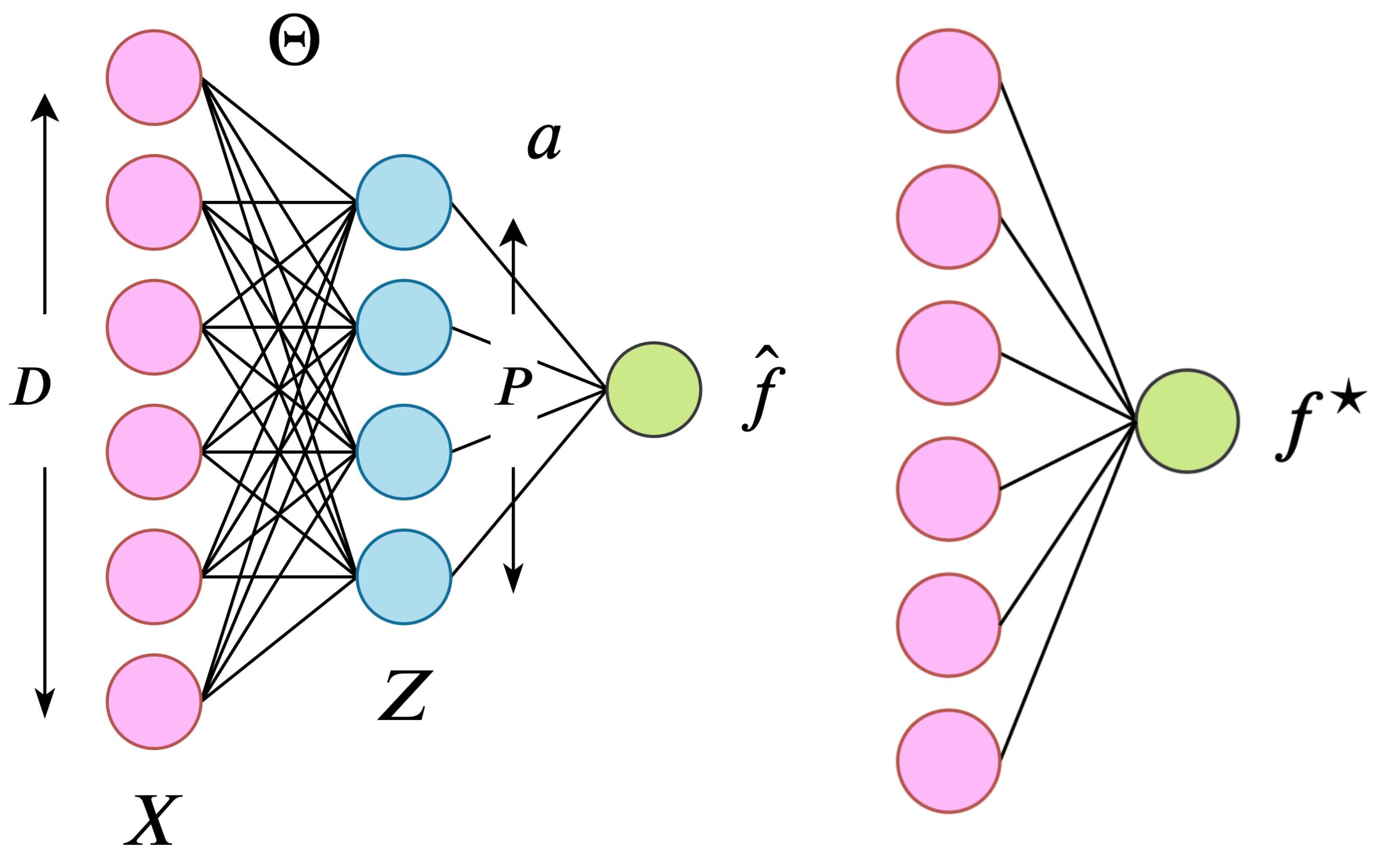
$$\boldsymbol{X} \sim \mathcal{N}(0,1) \in \mathbb{R}^{N \times D}$$

$$y = f^\star(\boldsymbol{x}) + \epsilon$$

$$\epsilon \sim \mathcal{N}(0,1/\mathrm{SNR})$$

$$\mathcal{L}_g = \mathop{\mathbb{E}}_{x}\left[\left(f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})\right)^2\right]$$

RF MODEL

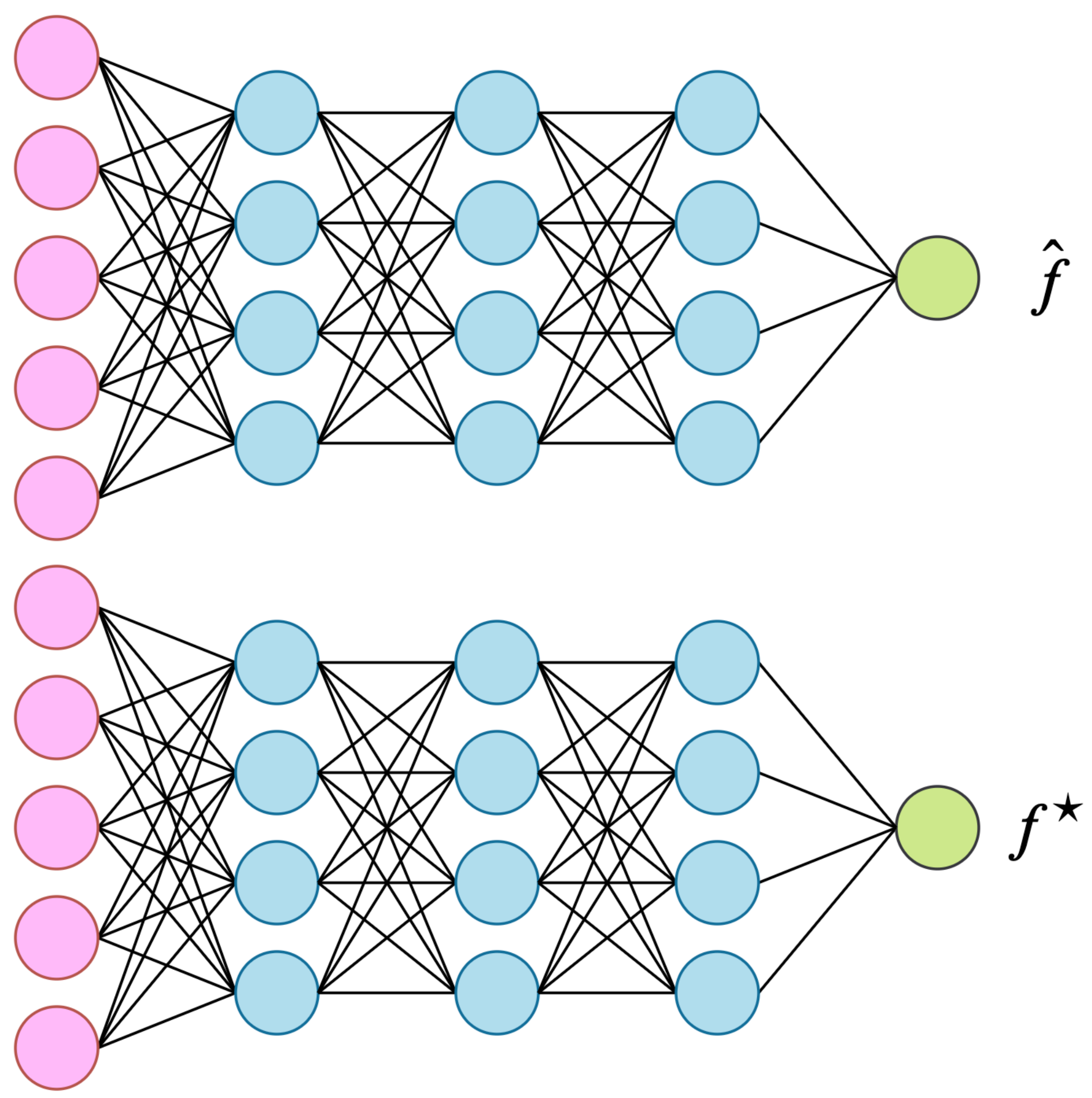


$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{P} a_i \sigma\left(\frac{\langle \boldsymbol{\Theta}_i, \boldsymbol{x} \rangle}{\sqrt{D}}\right)$$

$$\hat{\boldsymbol{a}} = \mathop{\arg\min}_{\boldsymbol{a} \in \mathbb{R}^P} \left[\frac{1}{N}\left(\boldsymbol{y} - \boldsymbol{a}\mathbf{Z}^\top\right)^2 + \frac{P\gamma}{D}\|\boldsymbol{a}\|_2^2\right]$$

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

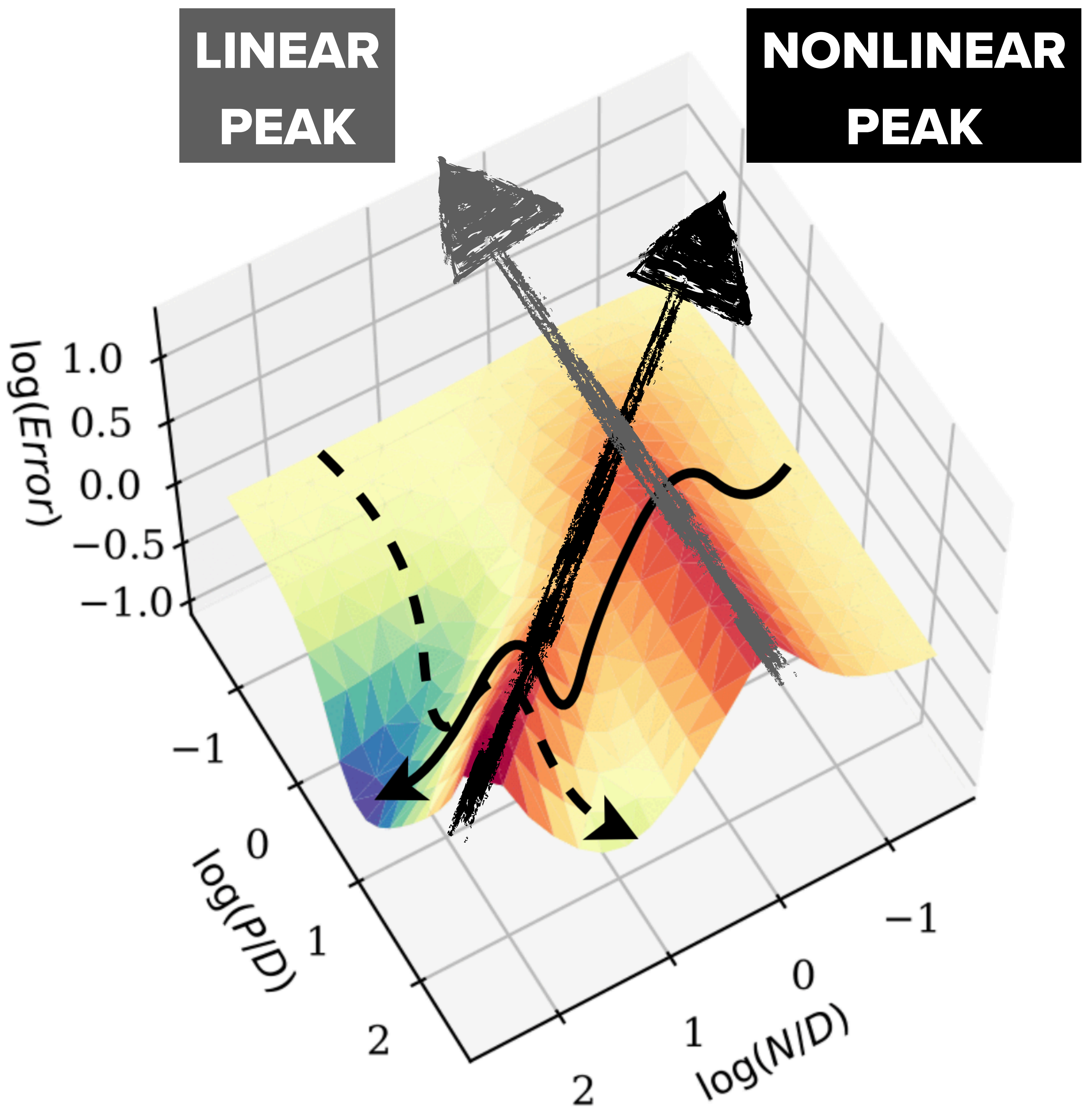
DNN MODEL



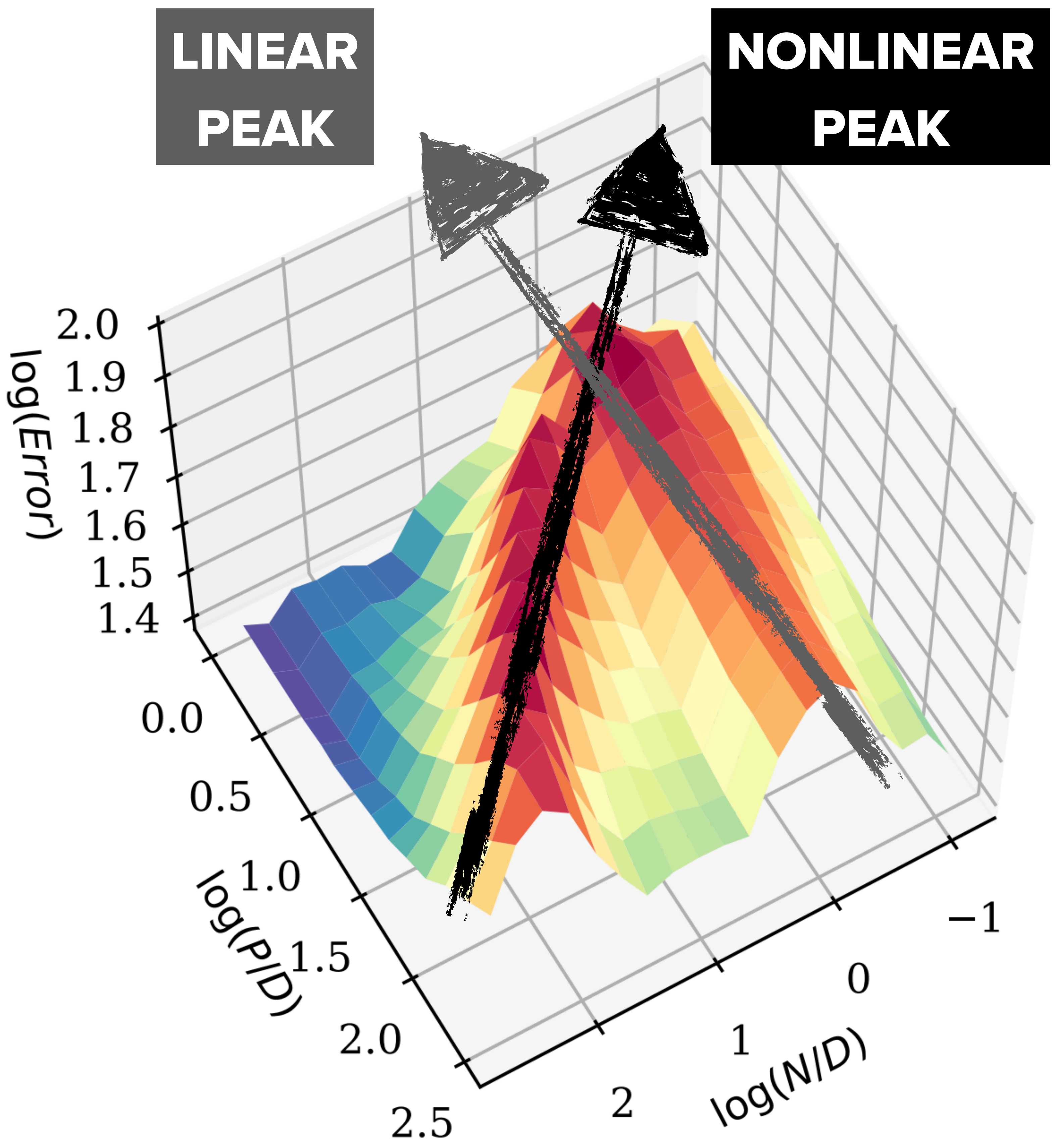
Random teacher

Student trained by GD

EVIDENCE OF TRIPLE DESCENT



RF MODEL



DNN MODEL

ANALYTICAL DESCRIPTION

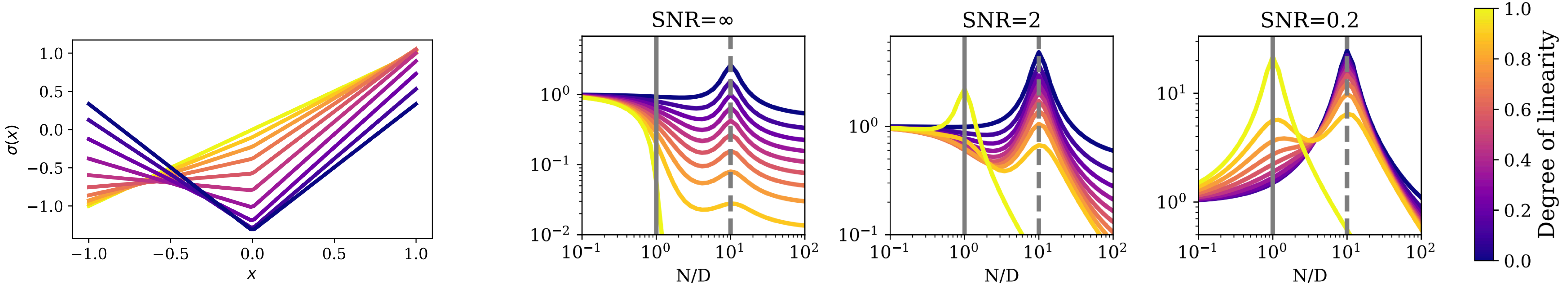
HIGH-DIMENSIONAL LIMIT

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

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

DEGREE OF LINEARITY

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



ANALYTICAL SPECTRUM

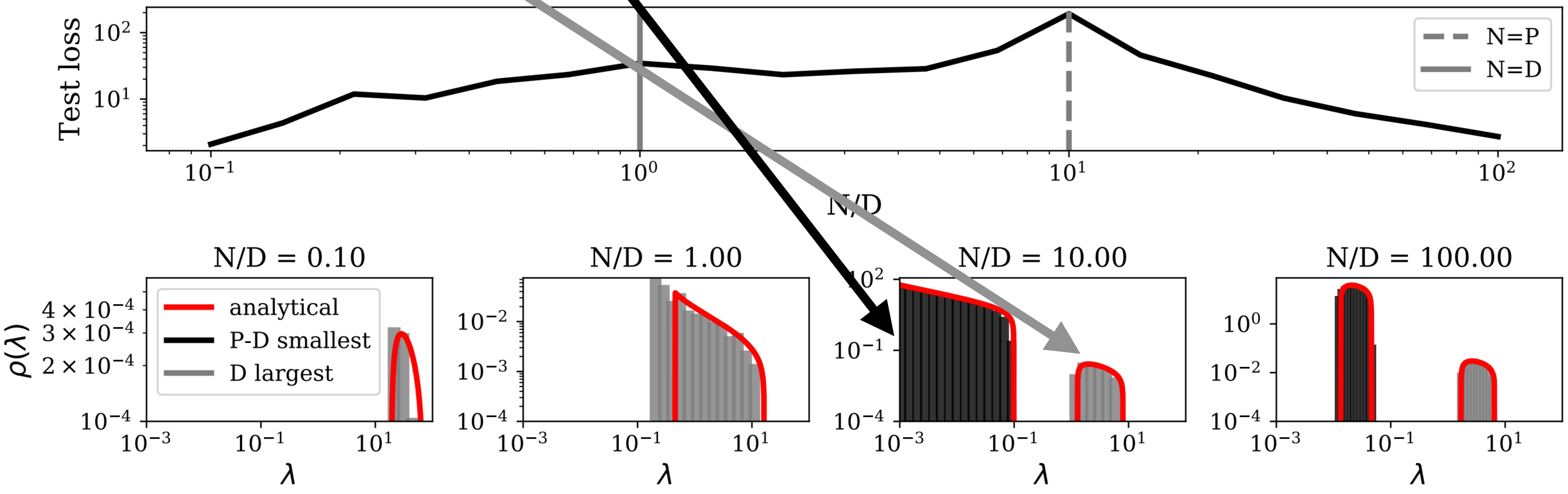
LINEAR PART **NONLINEAR PART**

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

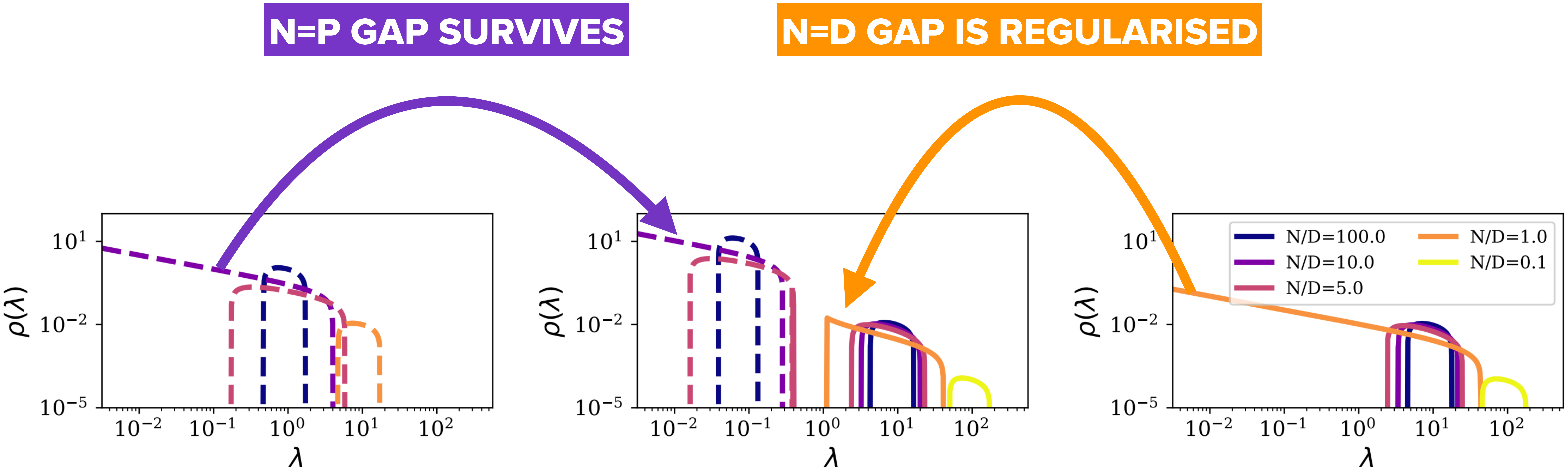
$$\rho(\lambda) = \frac{1}{\pi}\lim_{\epsilon\to0^+} \mathrm{Im}G(\lambda - i\epsilon), \qquad G(z) = \frac{\psi}{z}A\left(\frac{1}{z\psi}\right) + \frac{1-\psi}{z}$$

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

[Pennington & Worah 2017]



ANALYTICAL SPECTRUM



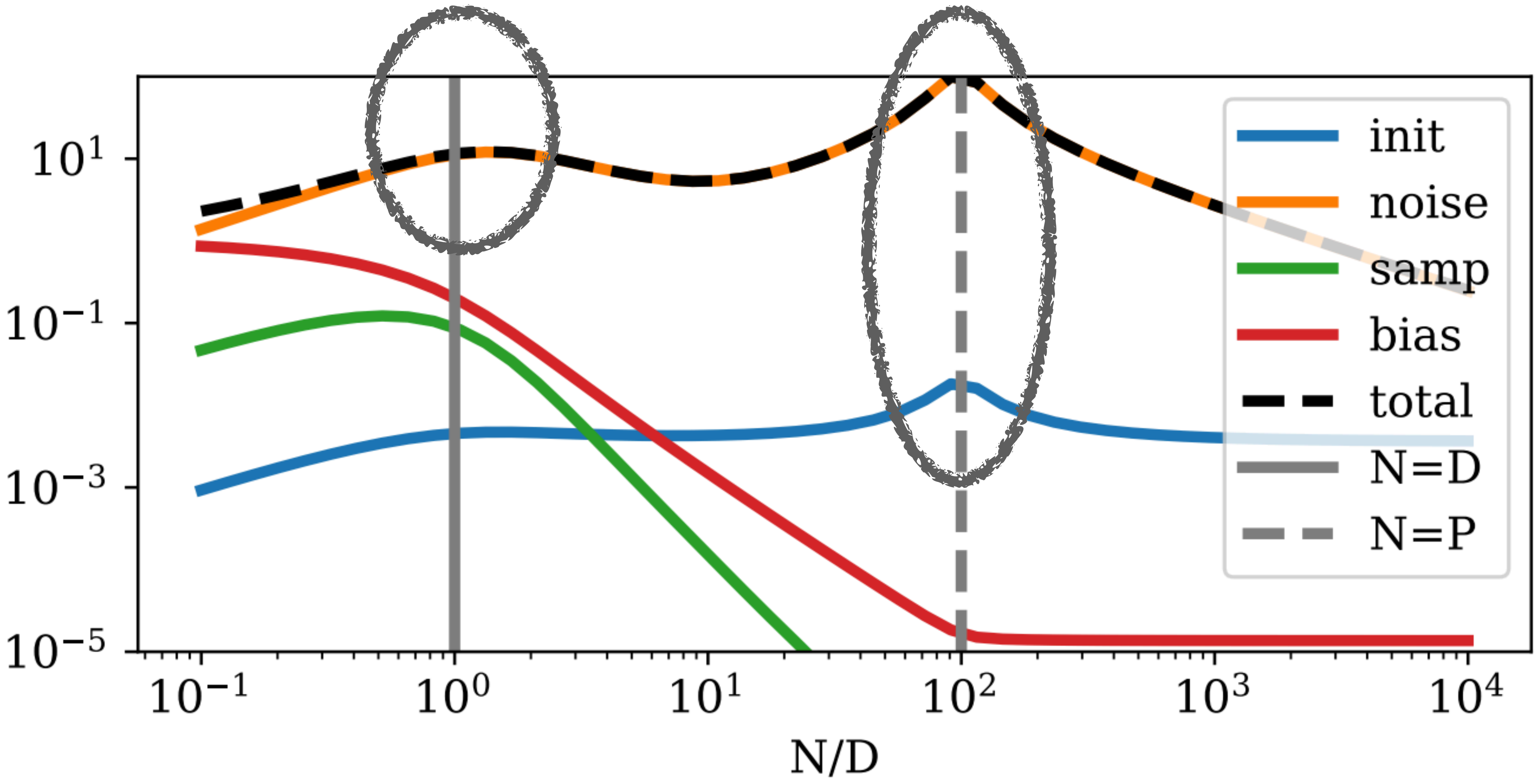
(a) Absolute value ($r = 0$)

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

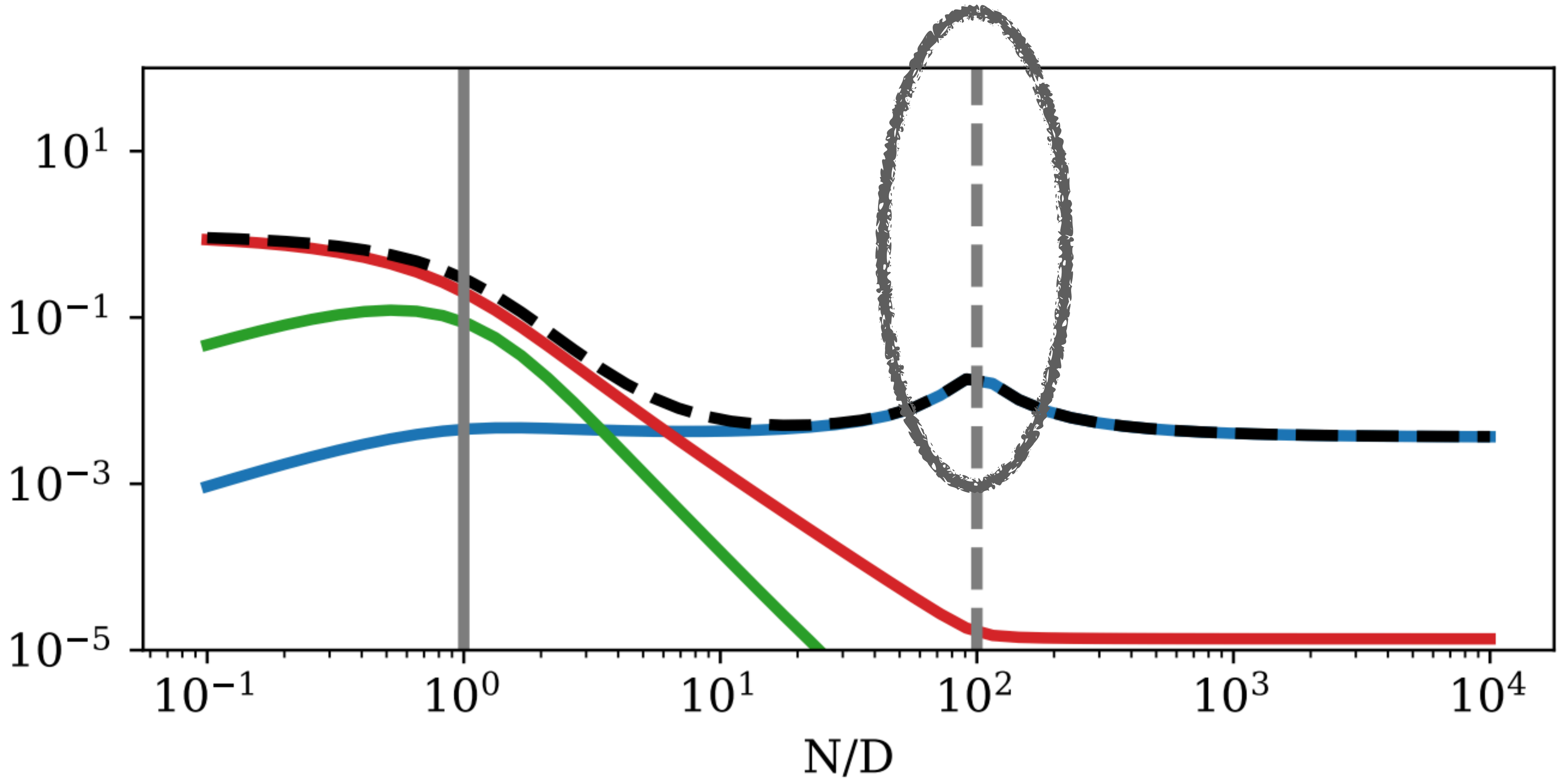
(c) Linear ($r = 1$)

BIAS AND VARIANCES

LINEAR PEAK CAUSED BY NOISE | **NONLINEAR PEAK CAUSED BY NOISE & INIT** → **NONLINEAR PEAK SURVIVES IN ABSENCE OF NOISE**

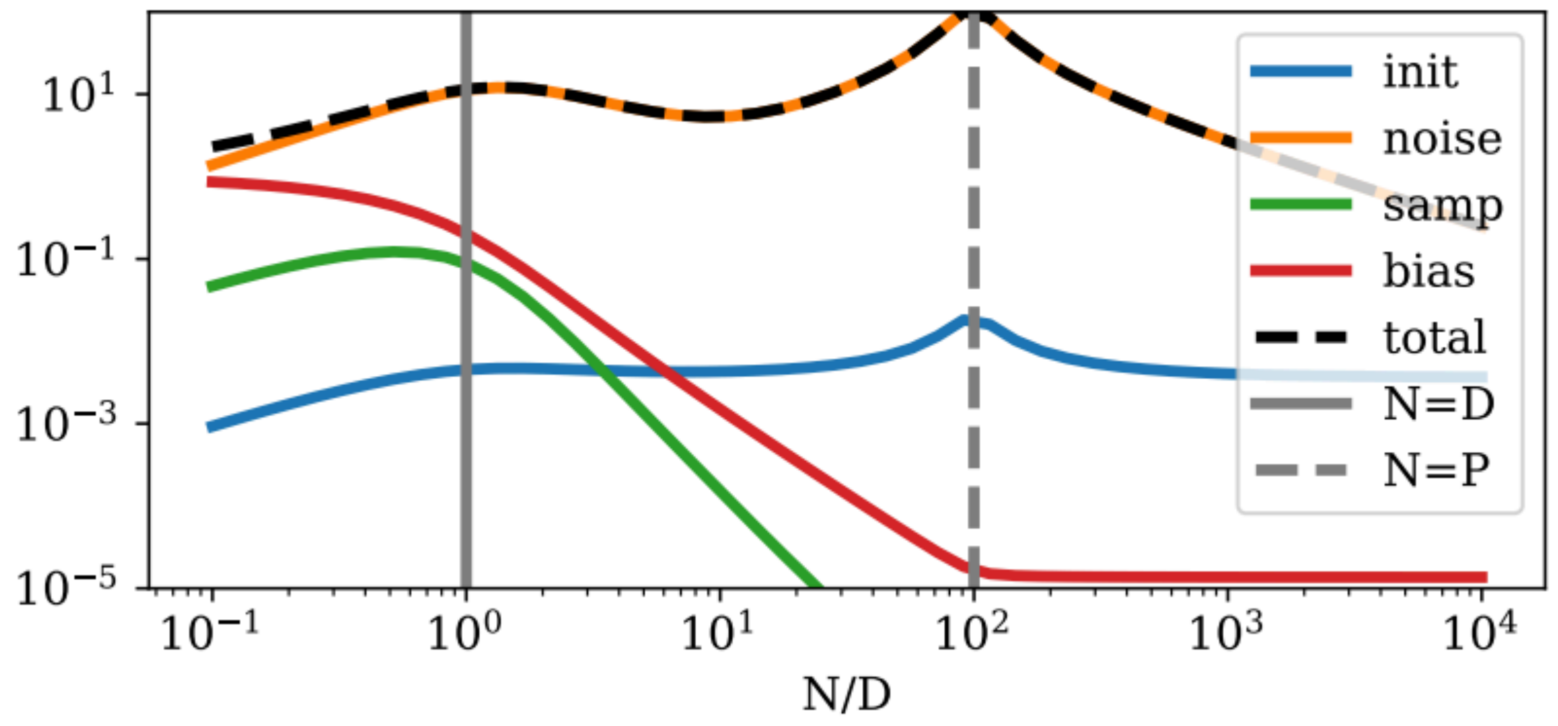


NOISY

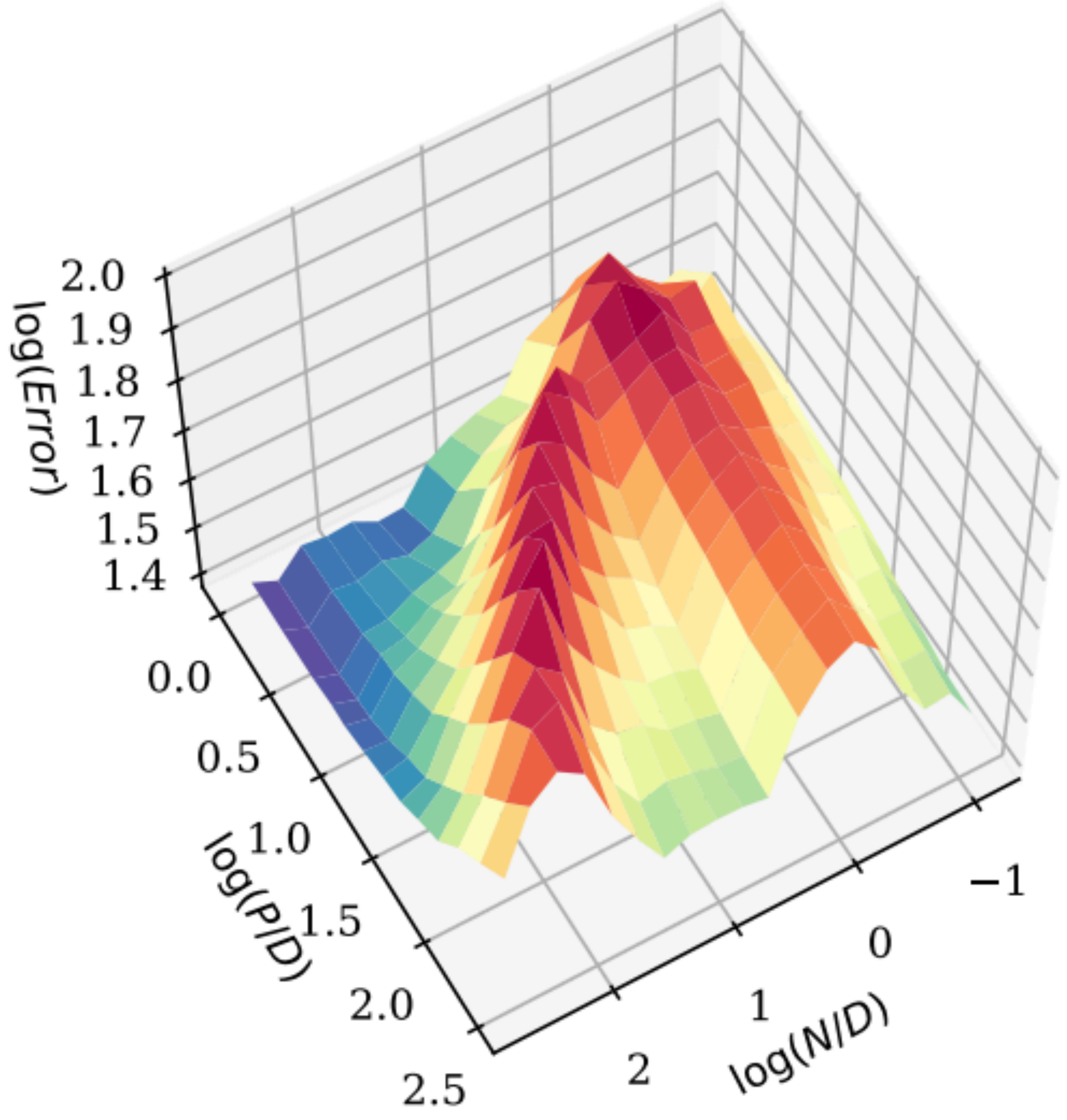


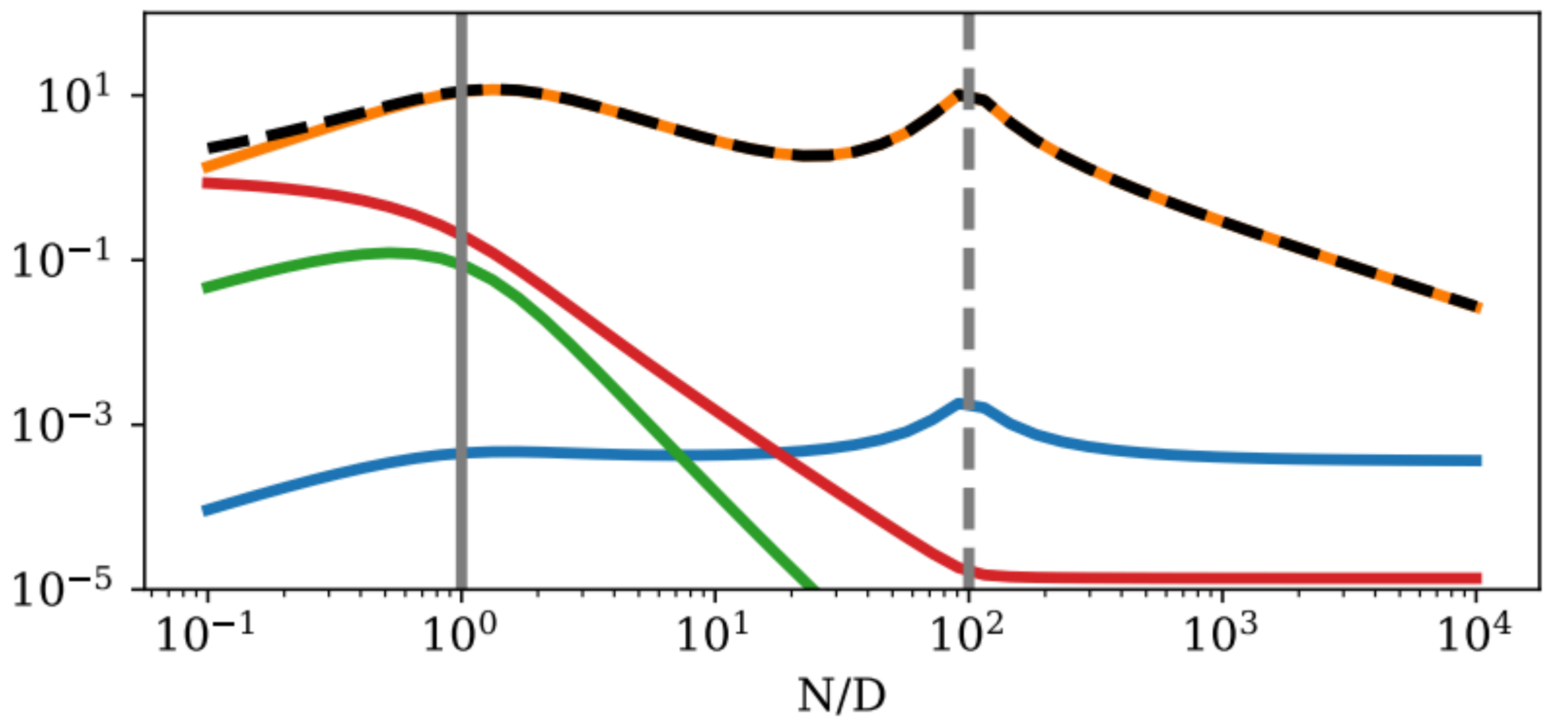
NOISELESS

EFFECT OF REGULARISATION

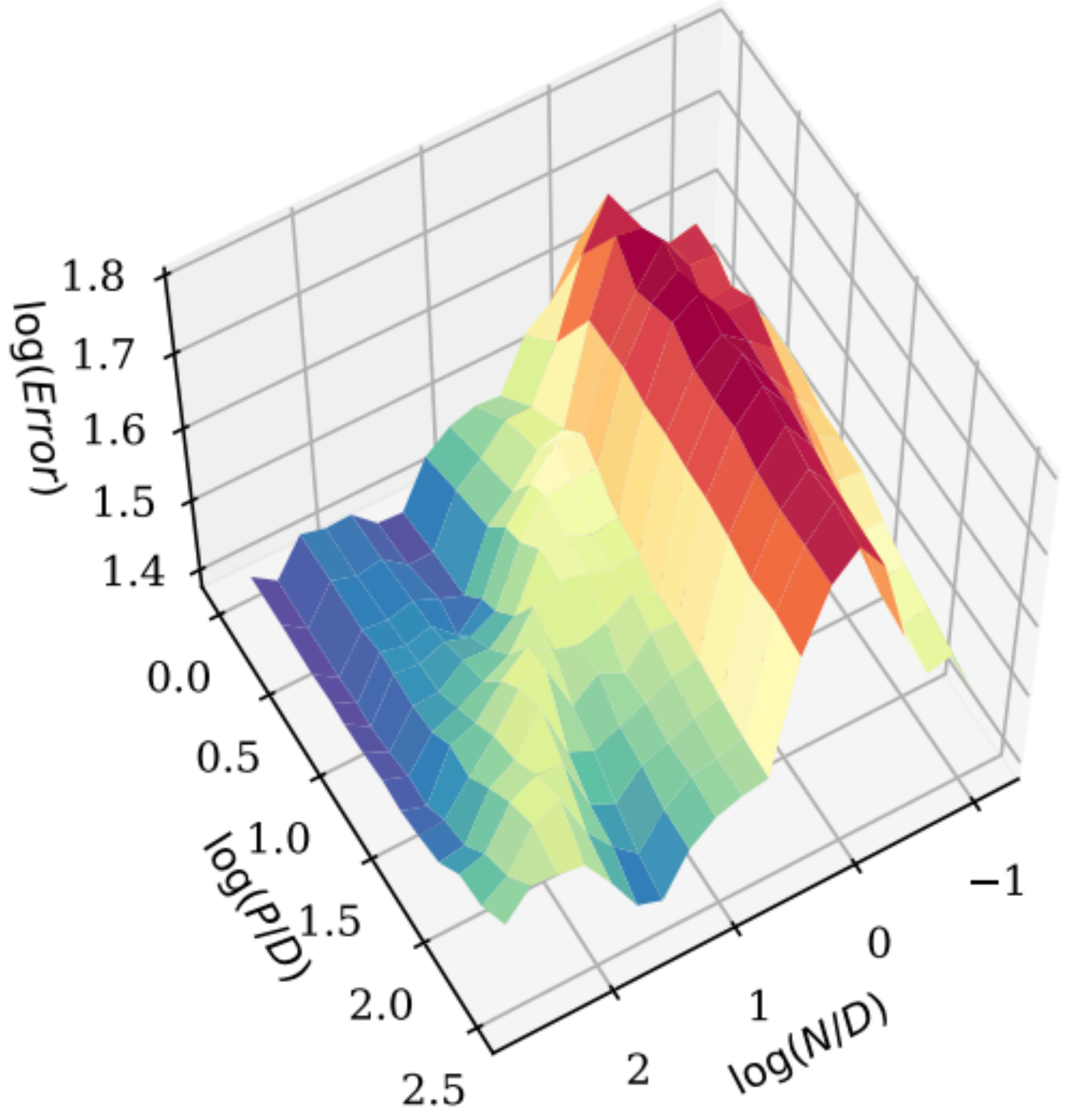


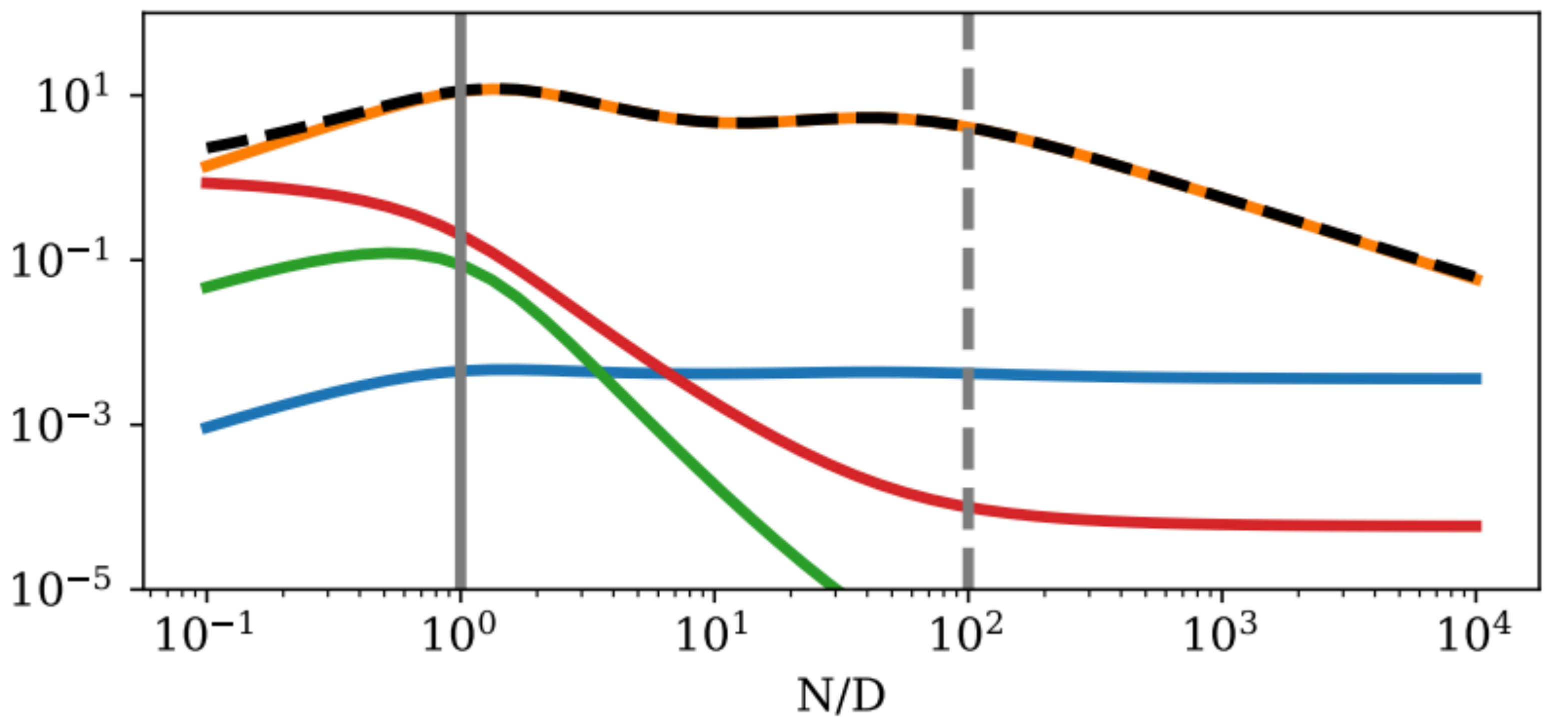
VANILLA



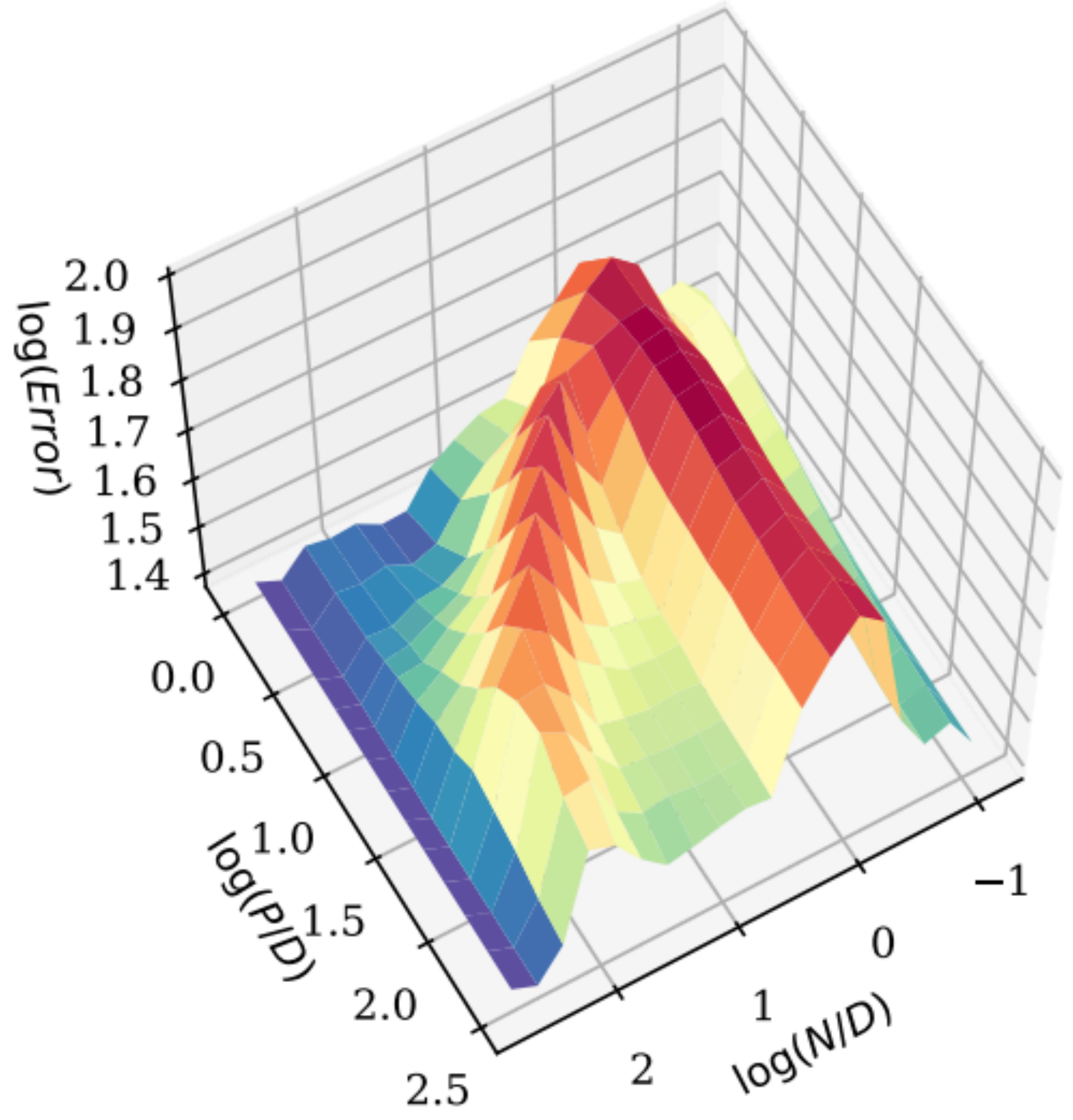


ENSEMBLING

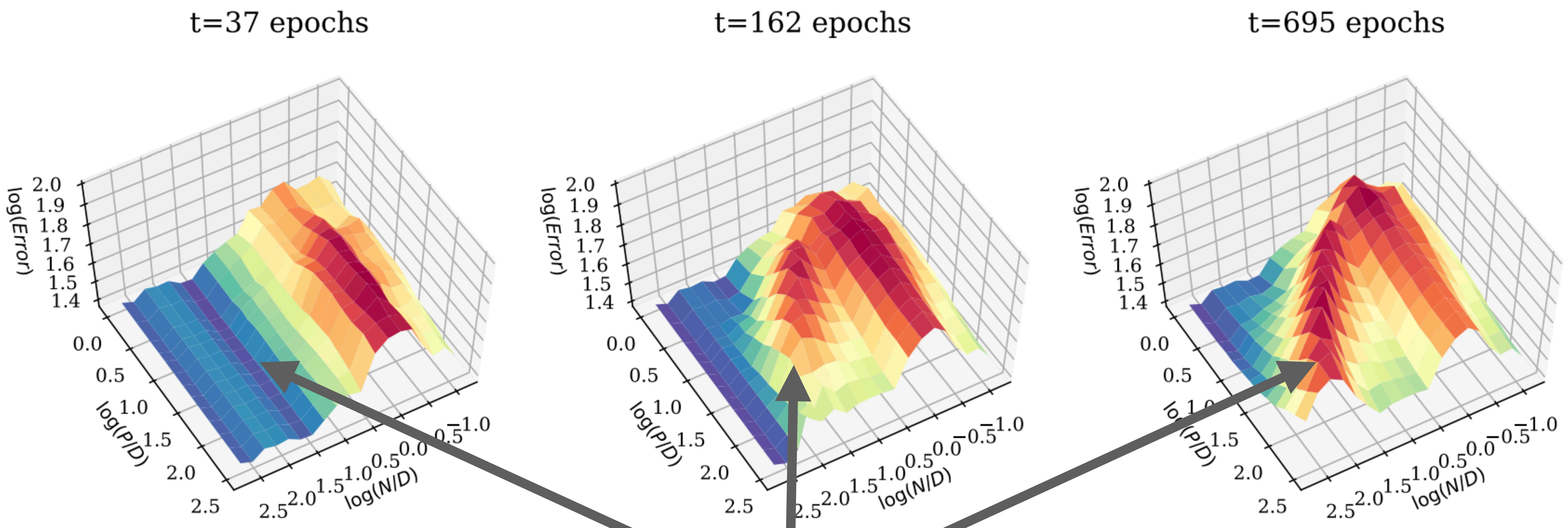




REGULARIZING

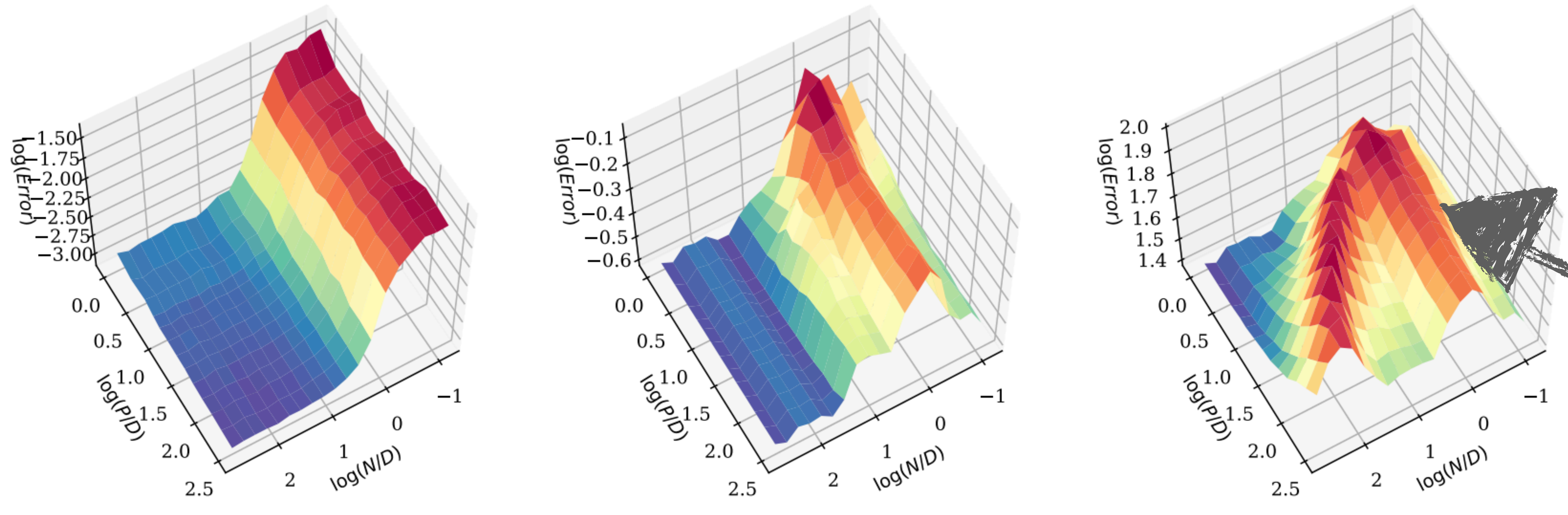


TIME DEPENDENCE



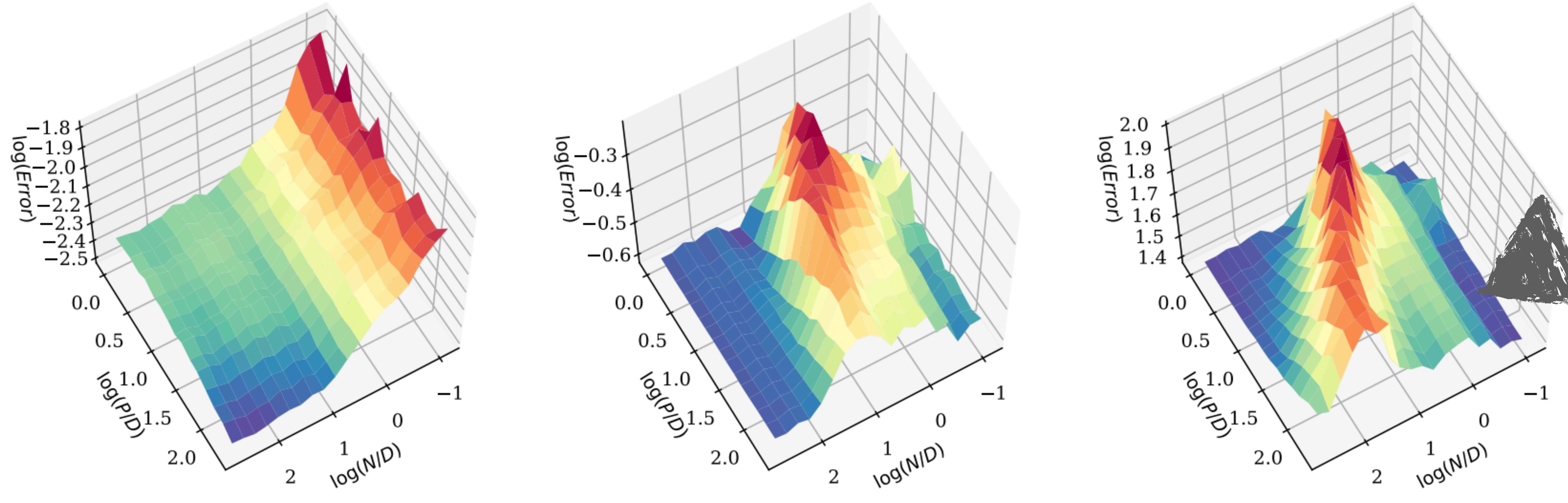
THE NONLINEAR PEAK FORMS AT LATE TIMES

EFFECT OF NOISE AND NONLINEARITY



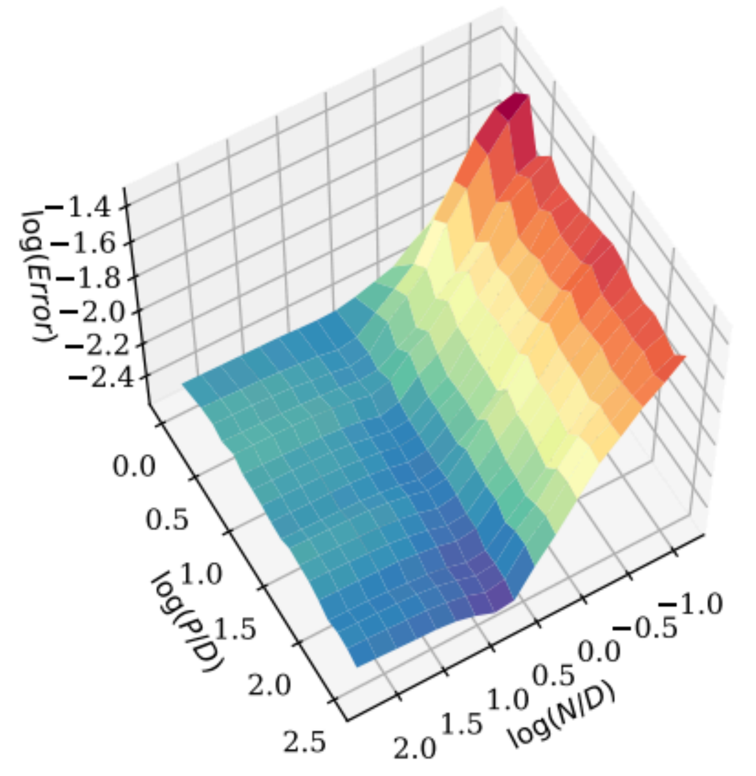
(a) Tanh, $SNR = \infty$ (b) Tanh, $SNR = 2$ (c) Tanh, $SNR = 0.2$

LINEAR PEAK IS WEAKER FOR RELU

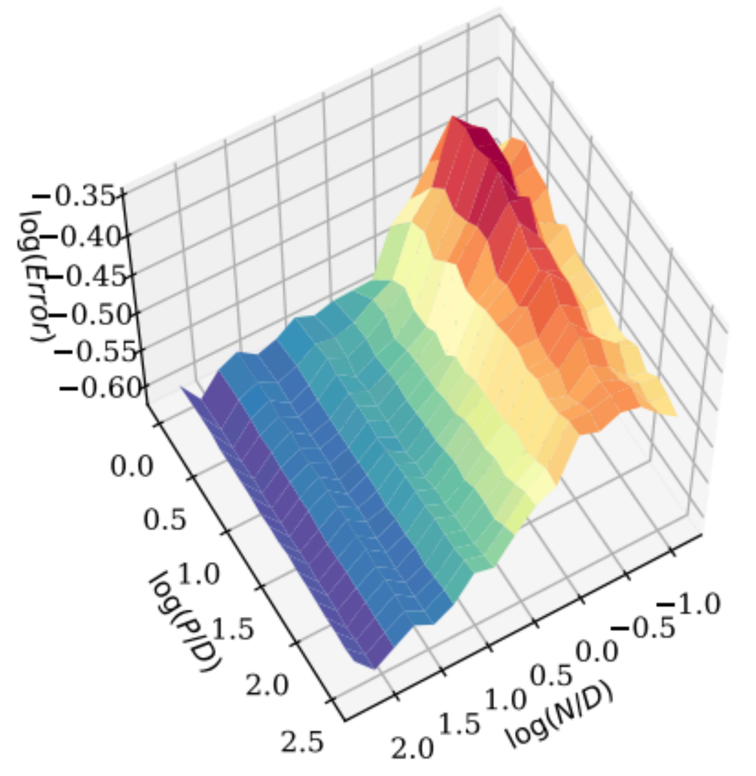


(d) ReLU, $SNR = \infty$ (e) ReLU, $SNR = 2$ (f) ReLU, $SNR = 0.2$

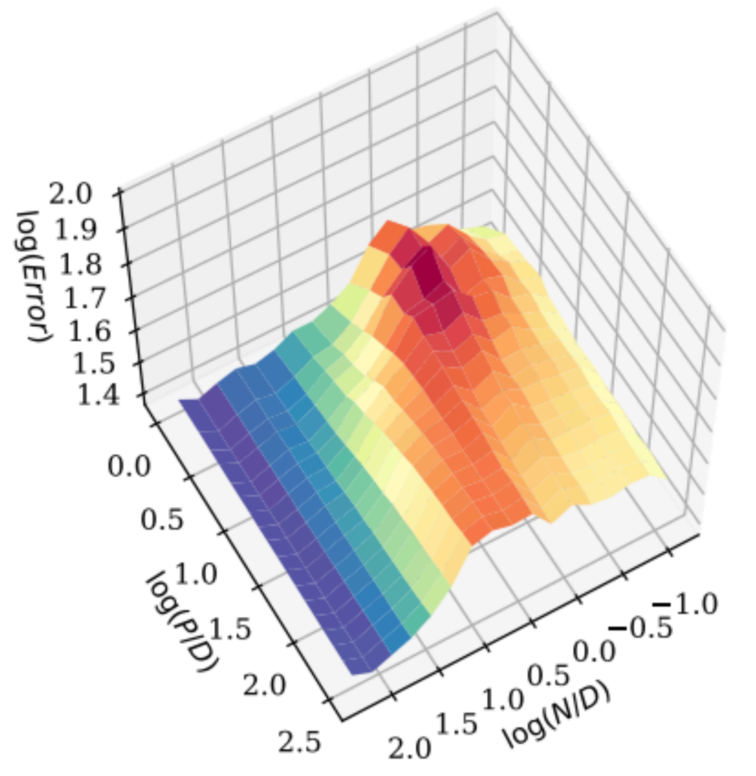
STRUCTURED DATASETS



(a) MNIST, $SNR = \infty$

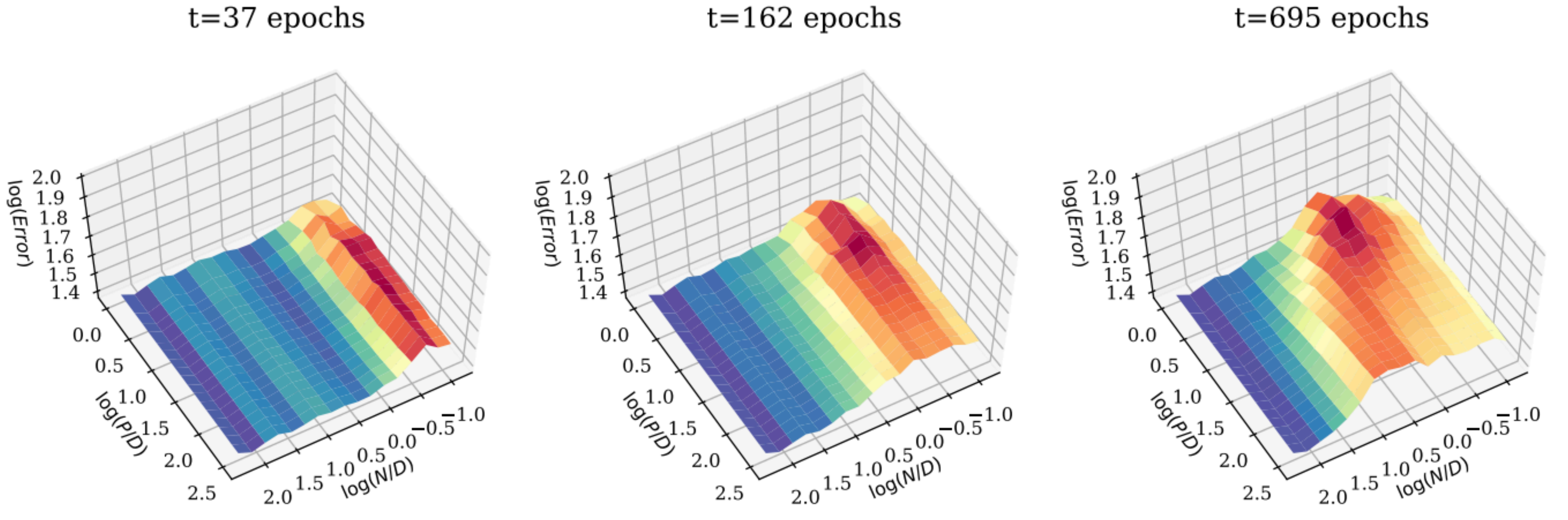


(b) MNIST, $SNR = 2$



(c) MNIST, $SNR = 0.2$

LINEAR AND NONLINEAR PEAK ARE MERGED TOGETHER



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

SHIFT FROM LINEAR TO NONLINEAR DURING TRAINING