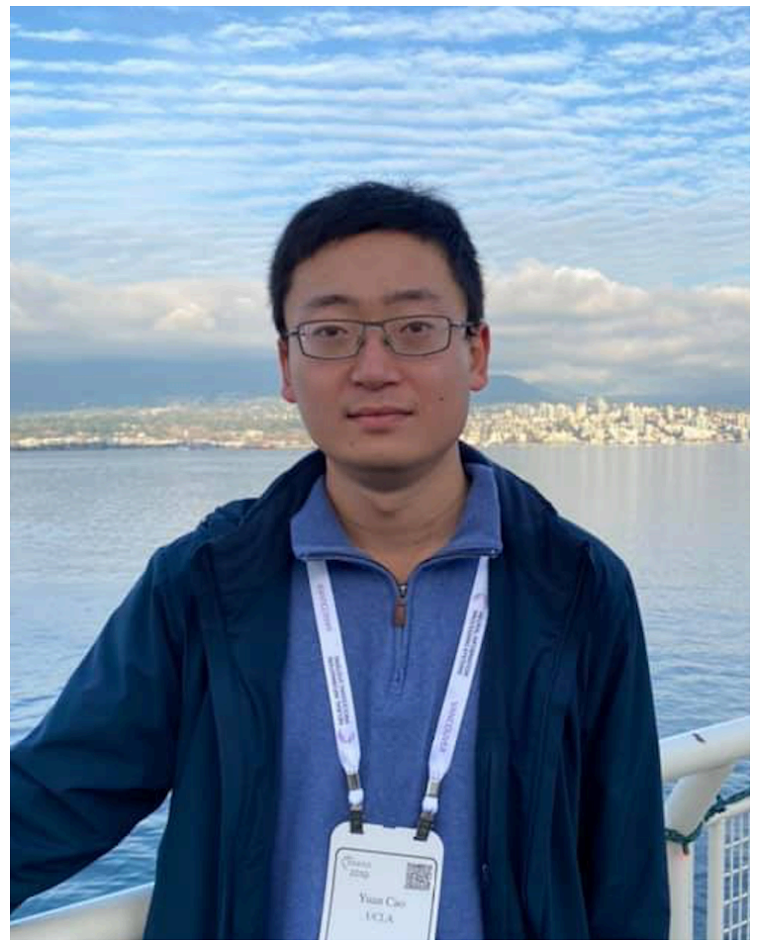
Multiple Descent in the Multiple Random Feature Model

Xuran Meng

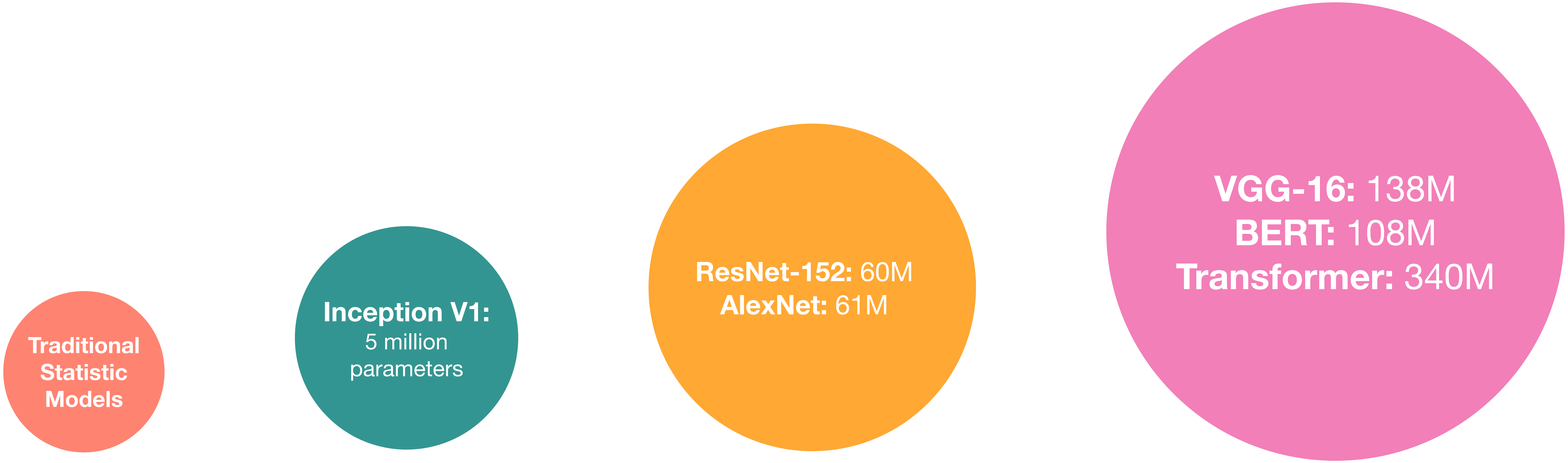
Department of Statistics and Actuarial Science

University of Hong Kong

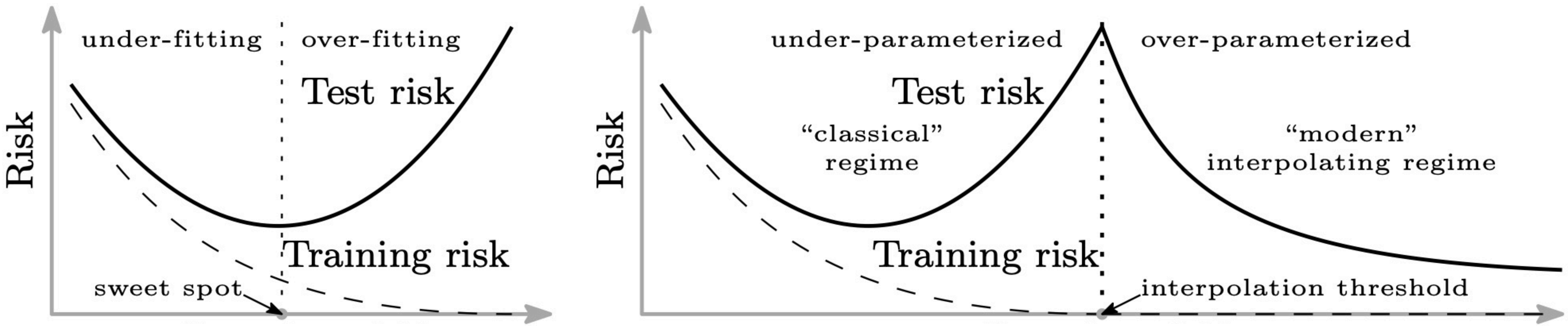


Joint work with **Jianfeng Yao** and **Yuan Cao**

Modern Neural Networks are Over-parameterized



Interesting Double Descent Phenomenon

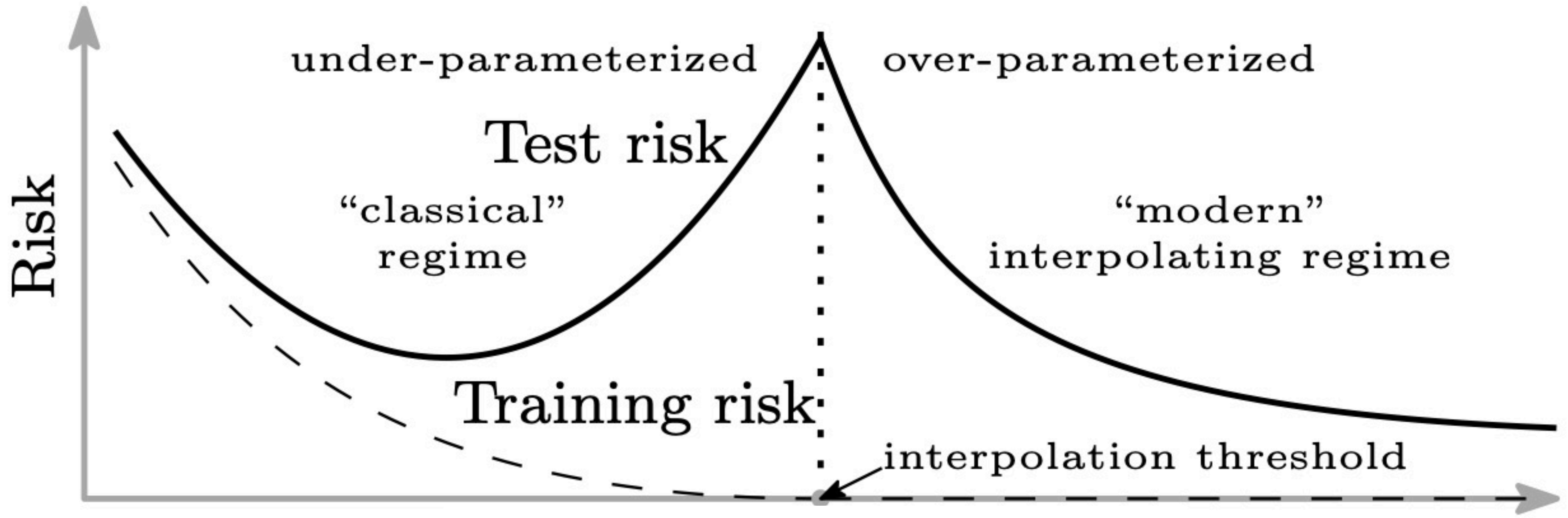


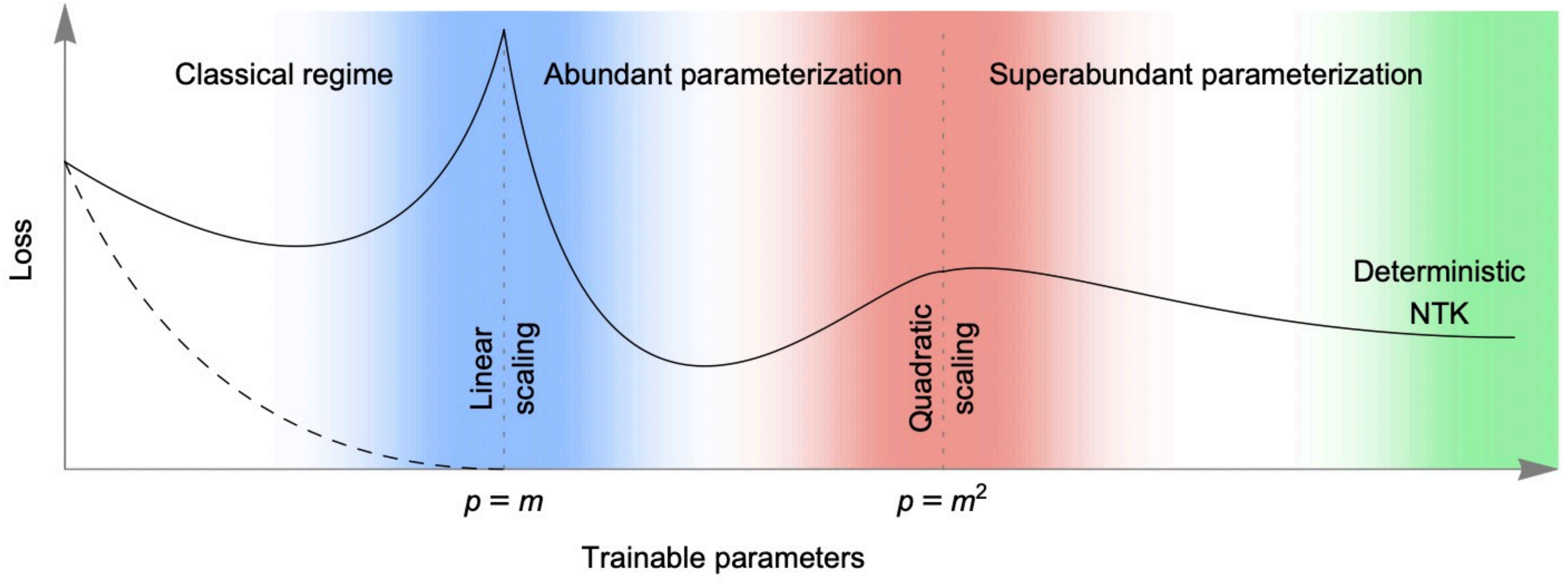
Model Complexity ∝ Number of Trainable Parameters

[1] Belkin, M., Hsu, D., Ma, S. and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *PNAS*. 2019.

[2] Belkin, M., Hsu, D. and Xu, J. Two models of double descent for weak features. SIMODS. 2020.

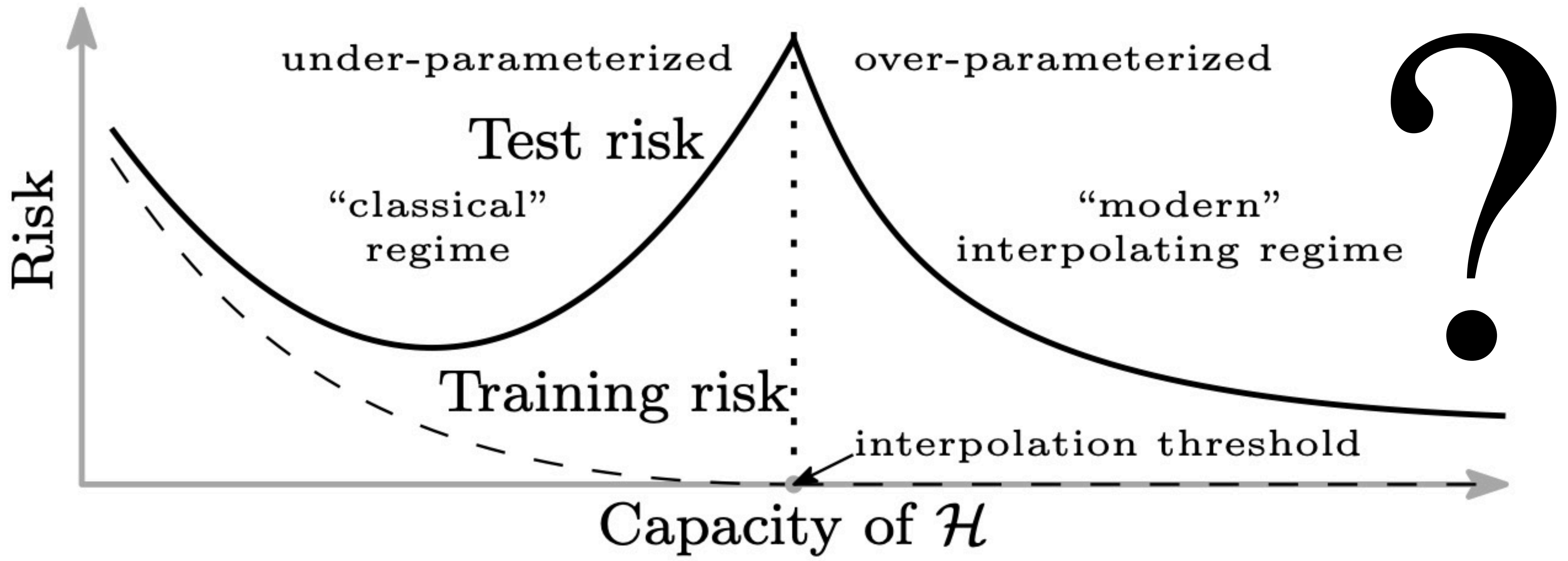
Interesting Double/Triple Descent Phenomenon





Aldam & Pennington. "The Neural Tangent Kernel in High Dimensions: Triple Descent and a Multi-Scale Theory of Generalization." *ICML, 2020.*

Always double descent?



We consider the following models

Multi-Component Prediction Models:

$$f(\mathbf{x}) = f_1(\mathbf{x}) + f_2(\mathbf{x}) + \ldots + f_K(\mathbf{x}),$$

where each $f_i(\mathbf{x})$ is an individual prediction models.

We consider the following models

Multi-Component Prediction Models:

$$f(\mathbf{x}) = f_1(\mathbf{x}) + f_2(\mathbf{x}) + \ldots + f_K(\mathbf{x}),$$

where each $f_i(\mathbf{x})$ is an individual prediction models.

- Ensemble methods

We consider the following models

Multi-Component Prediction Models:

$$f(\mathbf{x}) = f_1(\mathbf{x}) + f_2(\mathbf{x}) + \ldots + f_K(\mathbf{x}),$$

where each $f_i(\mathbf{x})$ is an individual prediction models.

- Ensemble methods
- Certain neural networks such as ResNet

We consider the following models

Multi-Component Prediction Models:

$$f(\mathbf{x}) = f_1(\mathbf{x}) + f_2(\mathbf{x}) + \ldots + f_K(\mathbf{x}),$$

where each $f_i(\mathbf{x})$ is an individual prediction models.

- Ensemble methods
- Certain neural networks such as ResNet

What can we say about the risk curves of multi-component prediction models?

Motivation

We aim to demonstrate that:

For any $K \in \mathbb{N}_+$, there exists a K-component prediction Model whose risk curve exhibits $(K+1)$-fold descent.

Motivation

We aim to demonstrate that:

For any $K \in \mathbb{N}_+$, there exists a K-component prediction

Model whose risk curve exhibits $(K+1)$-fold descent.

In the following, I will

first give some simple discussions and provide an intuitive explanation,

Motivation

We aim to demonstrate that:

For any $K \in \mathbb{N}_+$, there exists a K-component prediction Model whose risk curve exhibits $(K+1)$-fold descent.

In the following, I will

first give some simple discussions and provide an intuitive explanation,

then give some technical details for $K = 2$: how triple descent can be theoretically proved.

Motivation

We aim to demonstrate that:

For any $K \in \mathbb{N}_+$, there exists a K-component prediction

Model whose risk curve exhibits $(K+1)$-fold descent.

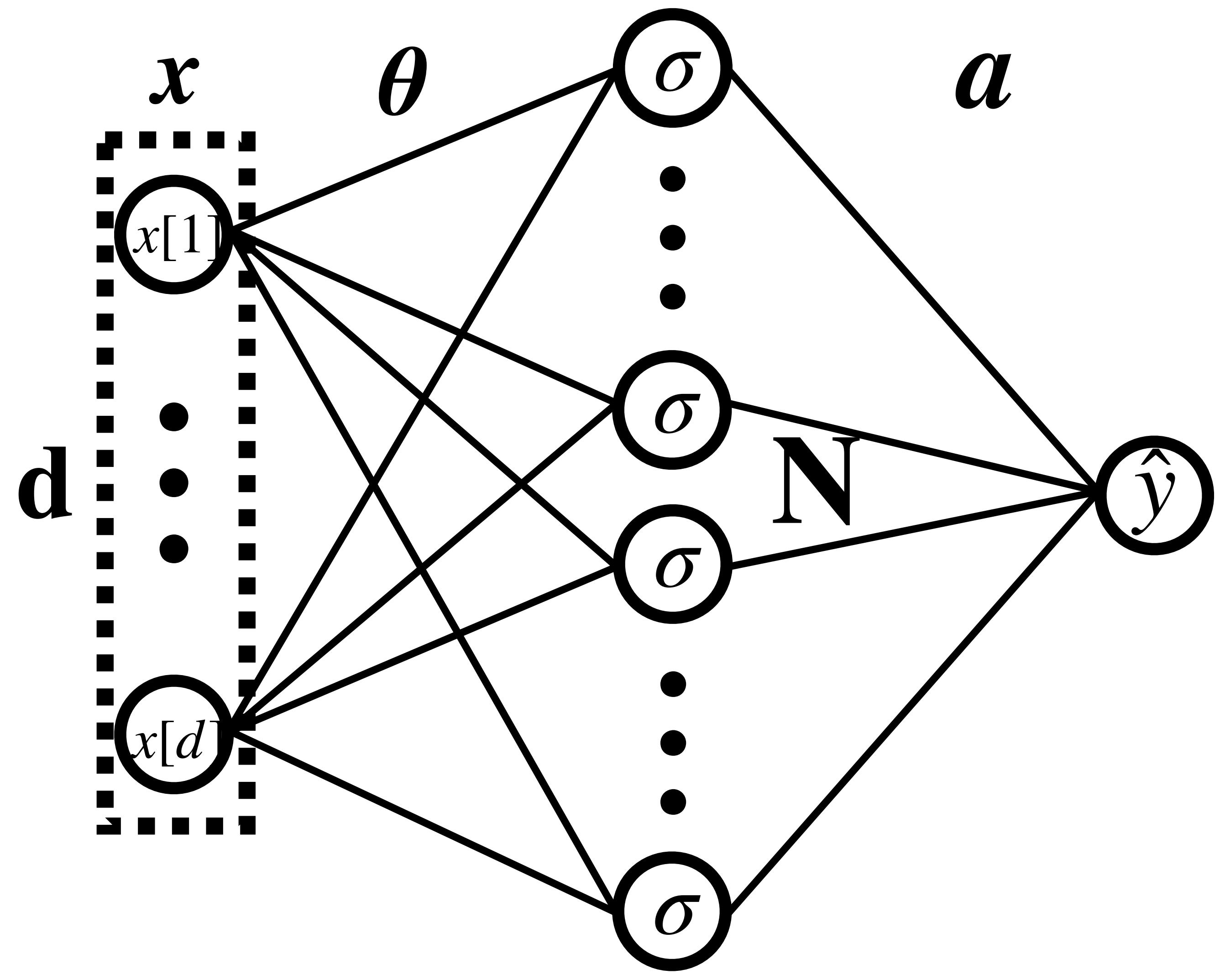
Classic random feature model:

(Mei & Montanari, 2022)

$$\mathcal{F}_{\mathrm{RF}}(\Theta) = \left\{ f(x; a, \Theta) \equiv \sum_{i=1}^{N} a_i \sigma\left(\langle \theta_i, x \rangle / \sqrt{d}\right) : a_i \in \mathbb{R} \quad \forall i \in [N] \right\}$$

Θ: fixed at randomly generated values

a: trainable parameters



Motivation

We aim to demonstrate that:

For any $K \in \mathbb{N}_+$, there exists a K-component prediction Model whose risk curve exhibits $(K+1)$-fold descent.

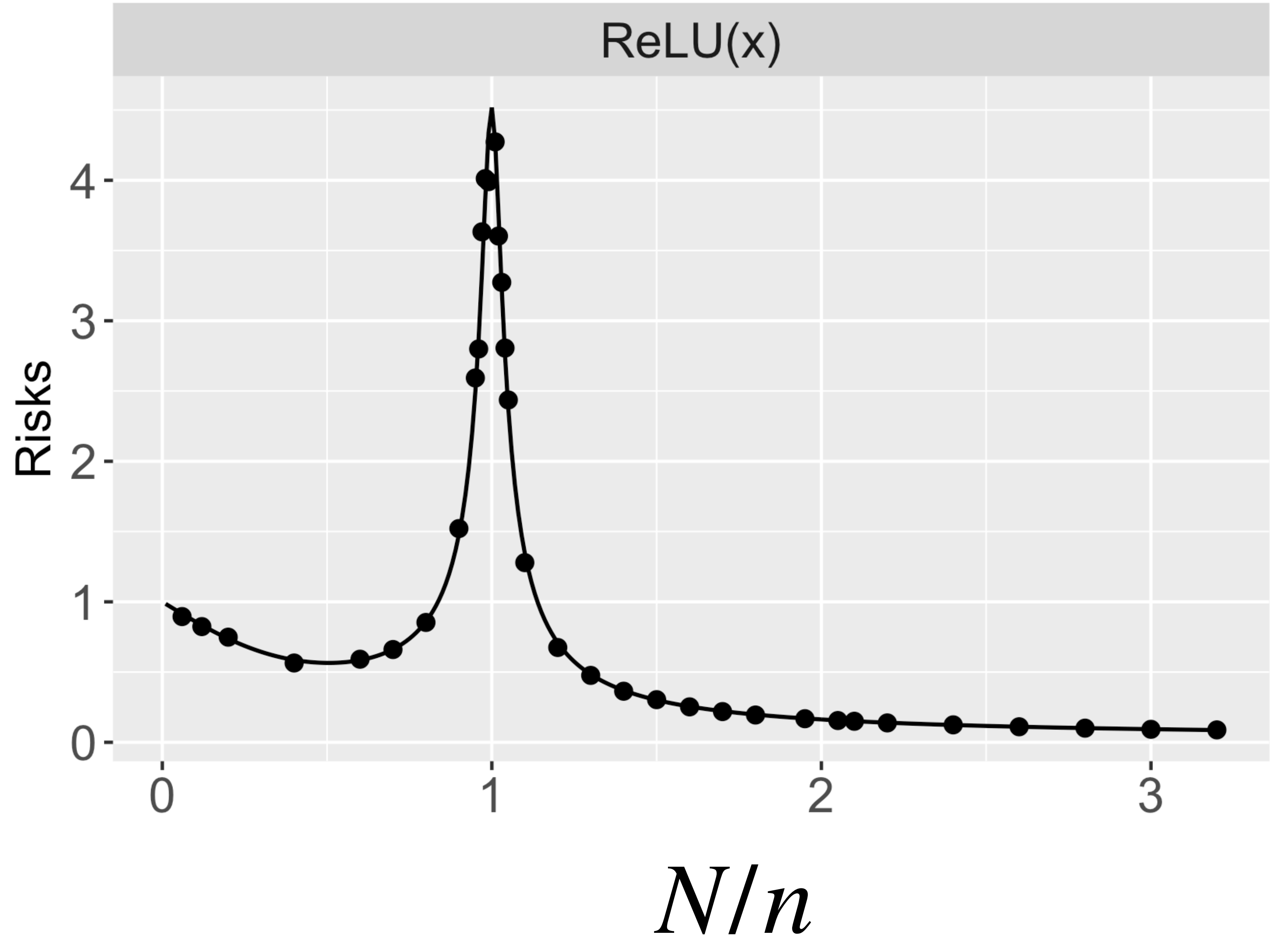
Classic random feature model:

(Mei & Montanari, 2022)

$$\mathscr{F}_{\mathrm{RF}}(\Theta) = \left\{ f(x; a, \Theta) \equiv \sum_{i=1}^{N} a_i \sigma\left(\langle \theta_i, x \rangle / \sqrt{d} \right) : a_i \in \mathbb{R} \quad \forall i \in [N] \right\}$$

Θ: fixed at randomly generated values

a: trainable parameters



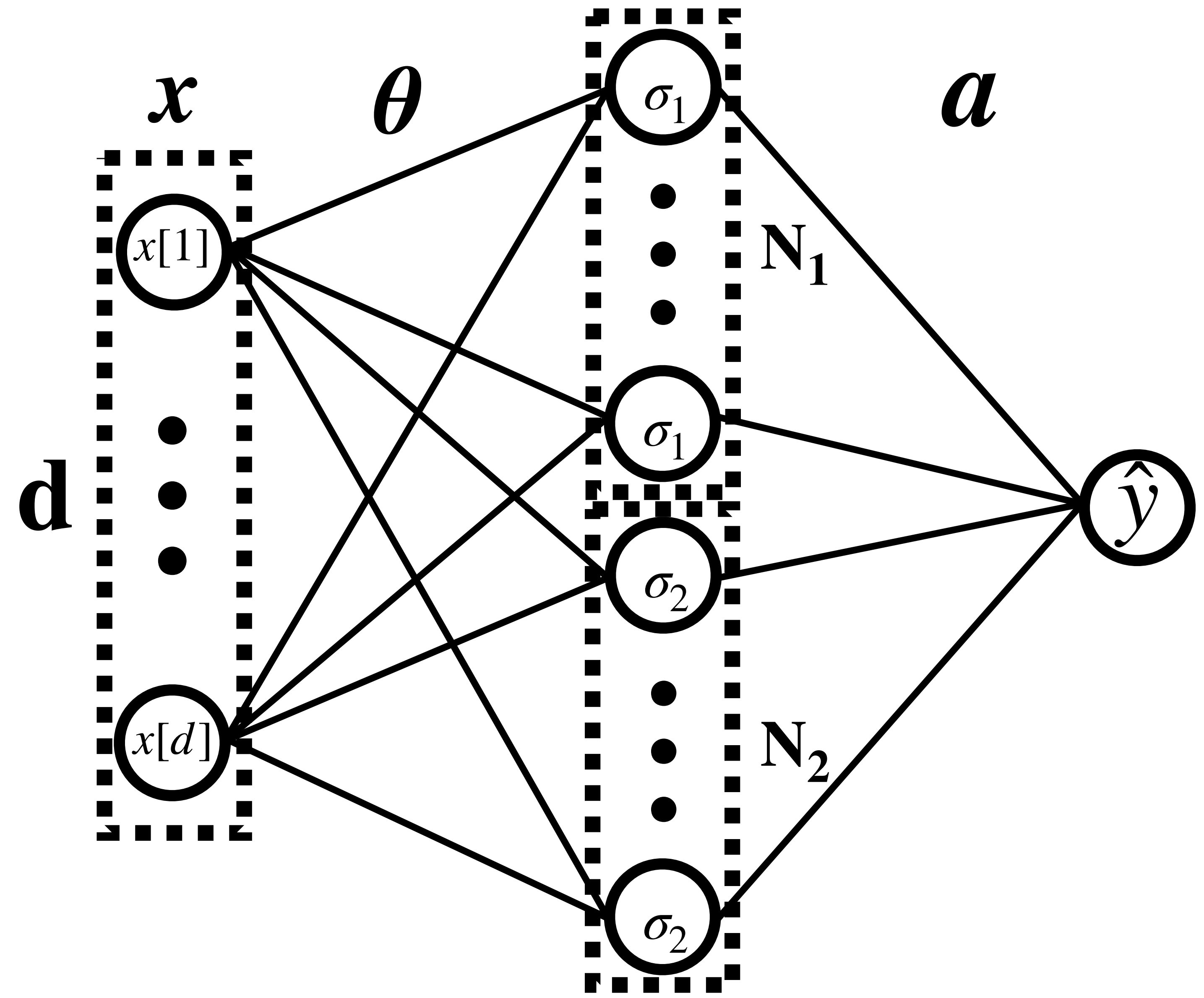
Motivation

We aim to demonstrate that:

For any $K \in \mathbb{N}_+$, there exists a K-component prediction Model whose risk curve exhibits $(K+1)$-fold descent.

Multiple random feature model:

$$\mathscr{F}_{\mathrm{MRF}}(\Theta) = \left\{ f(x; a, \Theta) \equiv \sum_{i=1}^{N_1} a_i \sigma_1 \left(\langle \theta_i, x \rangle / \sqrt{d} \right) + \sum_{i=N_1+1}^{N_1+N_2} a_i \sigma_2 \left(\langle \theta_i, x \rangle / \sqrt{d} \right) : a_i \in \mathbb{R} \quad \forall i \in [N] \right\}$$



Θ: fixed at randomly generated values

a: trainable parameters

Motivation

We aim to demonstrate that:

For any $K \in \mathbb{N}_+$, there exists a K-component prediction

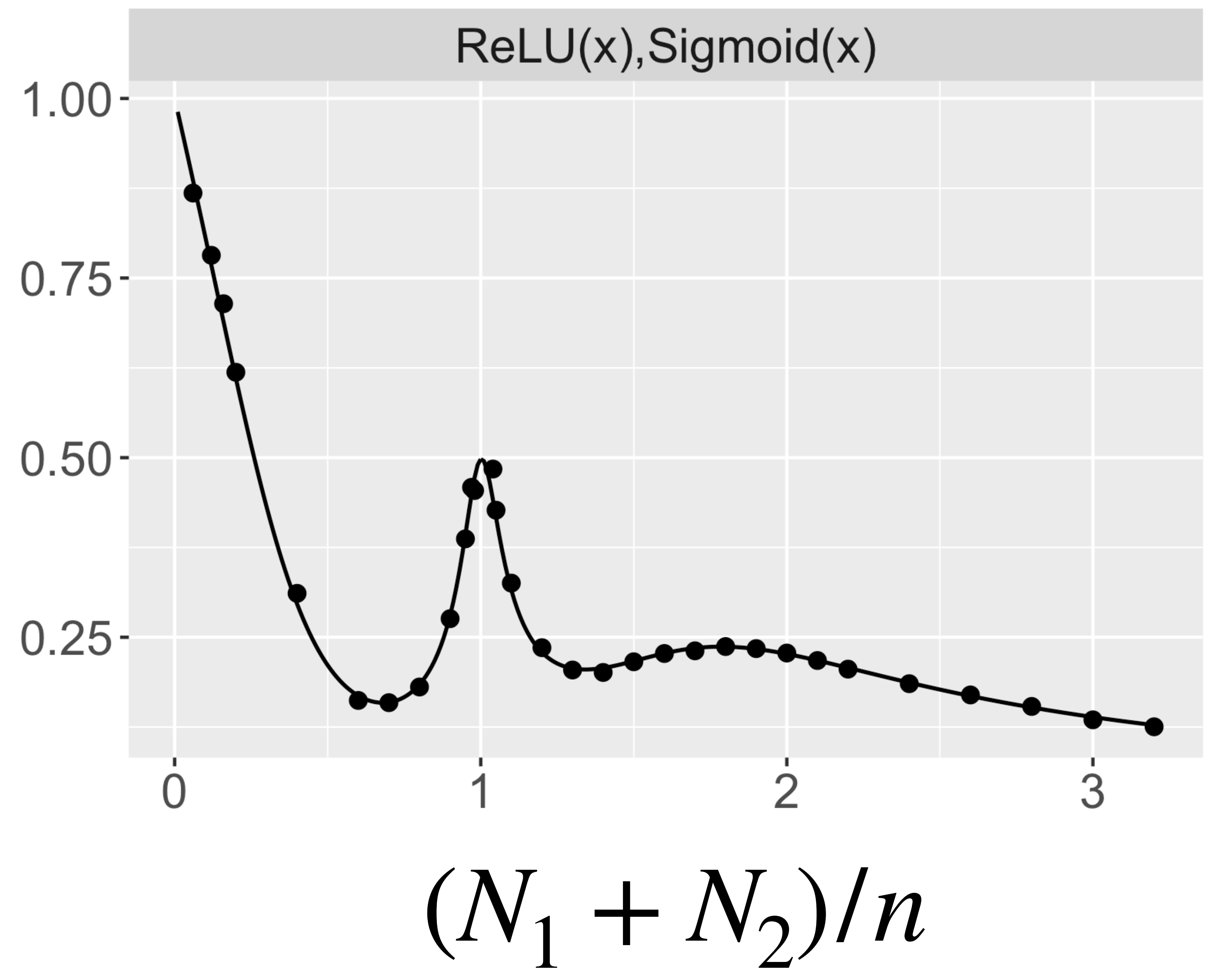
Model whose risk curve exhibits $(K+1)$-fold descent.

Multiple random feature model:

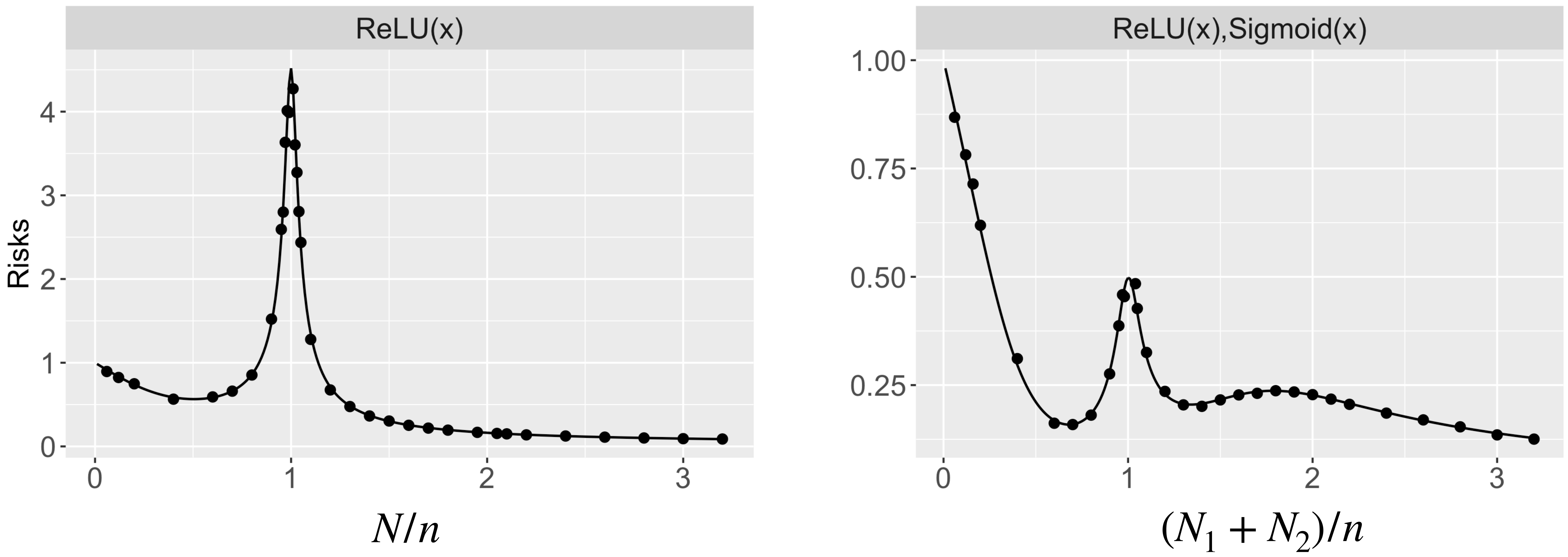
$$\mathscr{F}_{\mathrm{MRF}}(\Theta) = \left\{ f(x; a, \Theta) \equiv \sum_{i=1}^{N_1} a_i \sigma_1 \left(\langle \theta_i, x \rangle / \sqrt{d} \right) + \sum_{i=N_1+1}^{N_1+N_2} a_i \sigma_2 \left(\langle \theta_i, x \rangle / \sqrt{d} \right) : a_i \in \mathbb{R} \quad \forall i \in [N] \right\}$$

Θ: fixed at randomly generated values

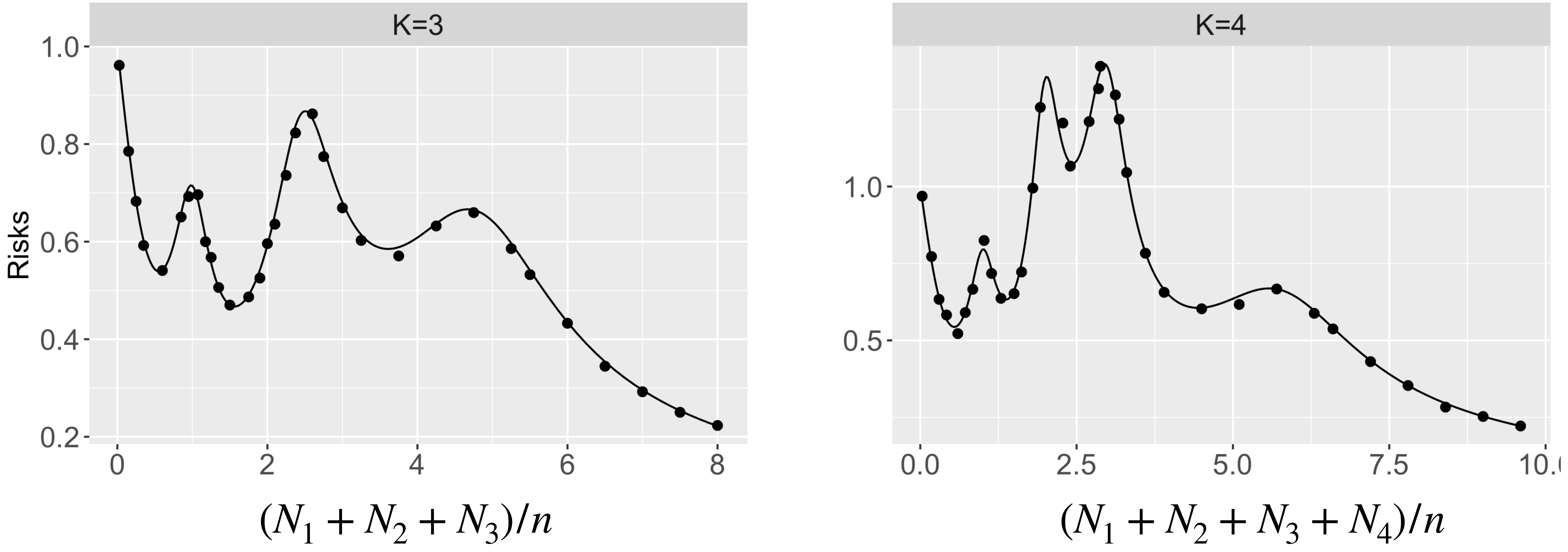
a: trainable parameters



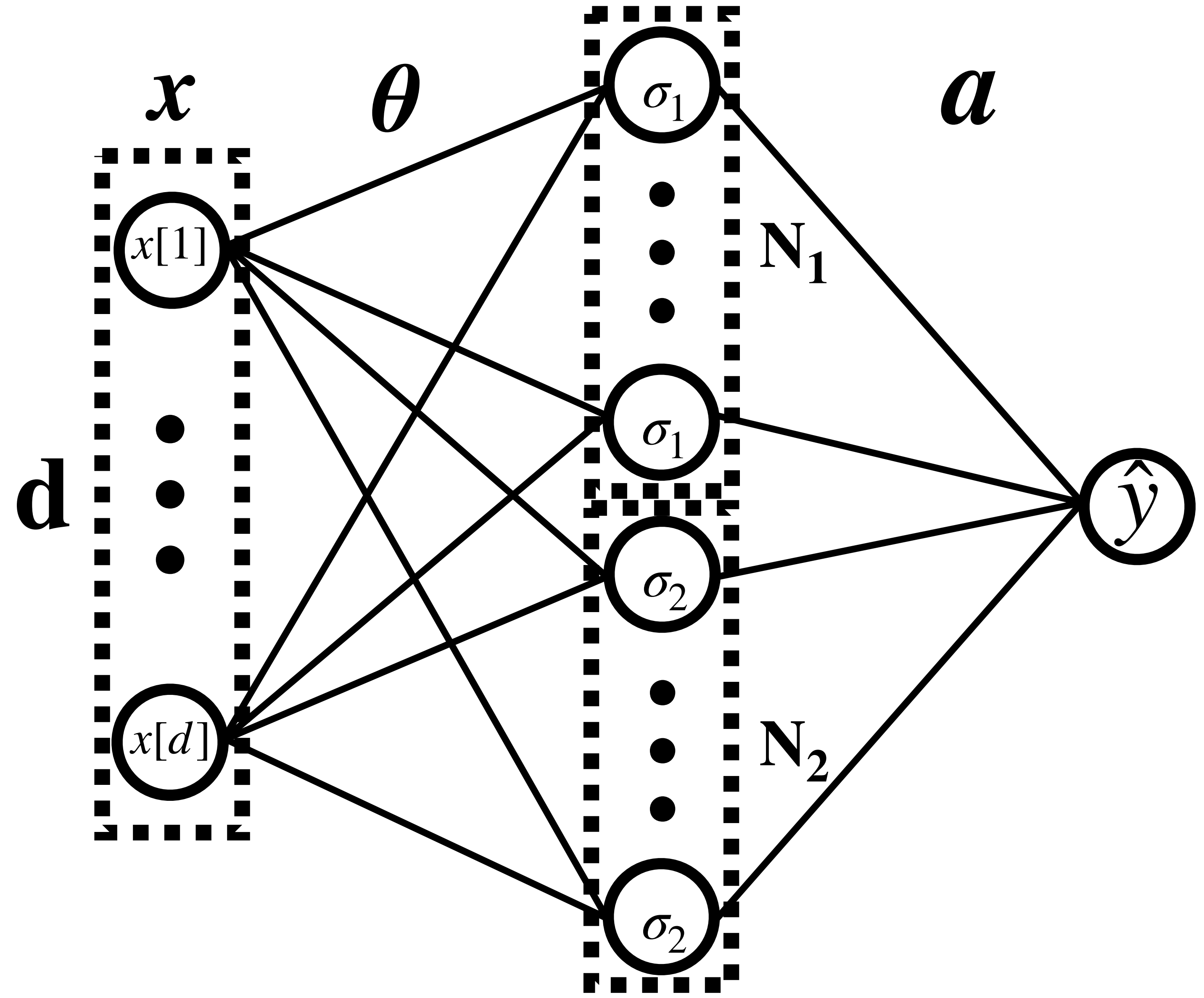
From Double Descent to Multiple Descent



From Double Descent to Multiple Descent

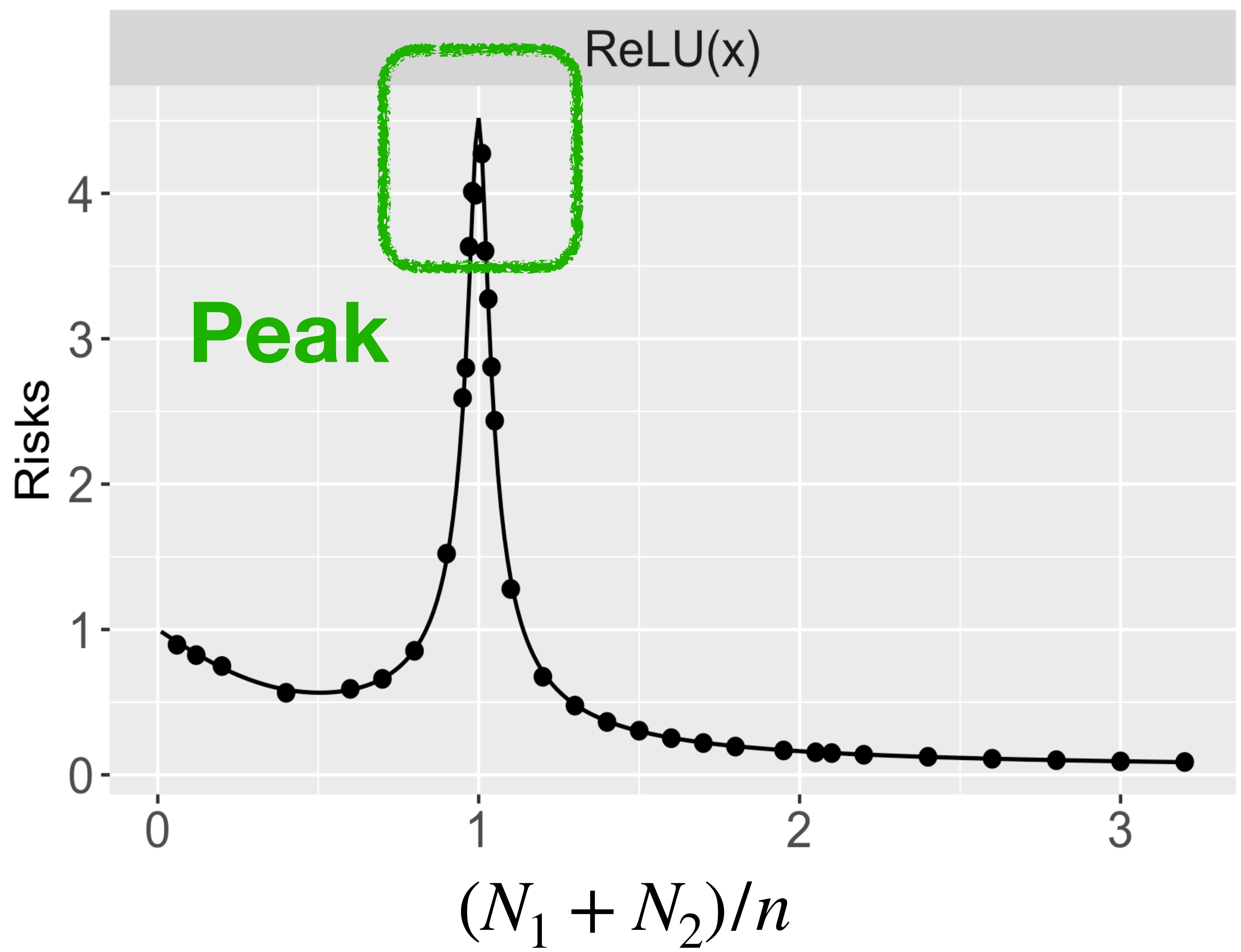


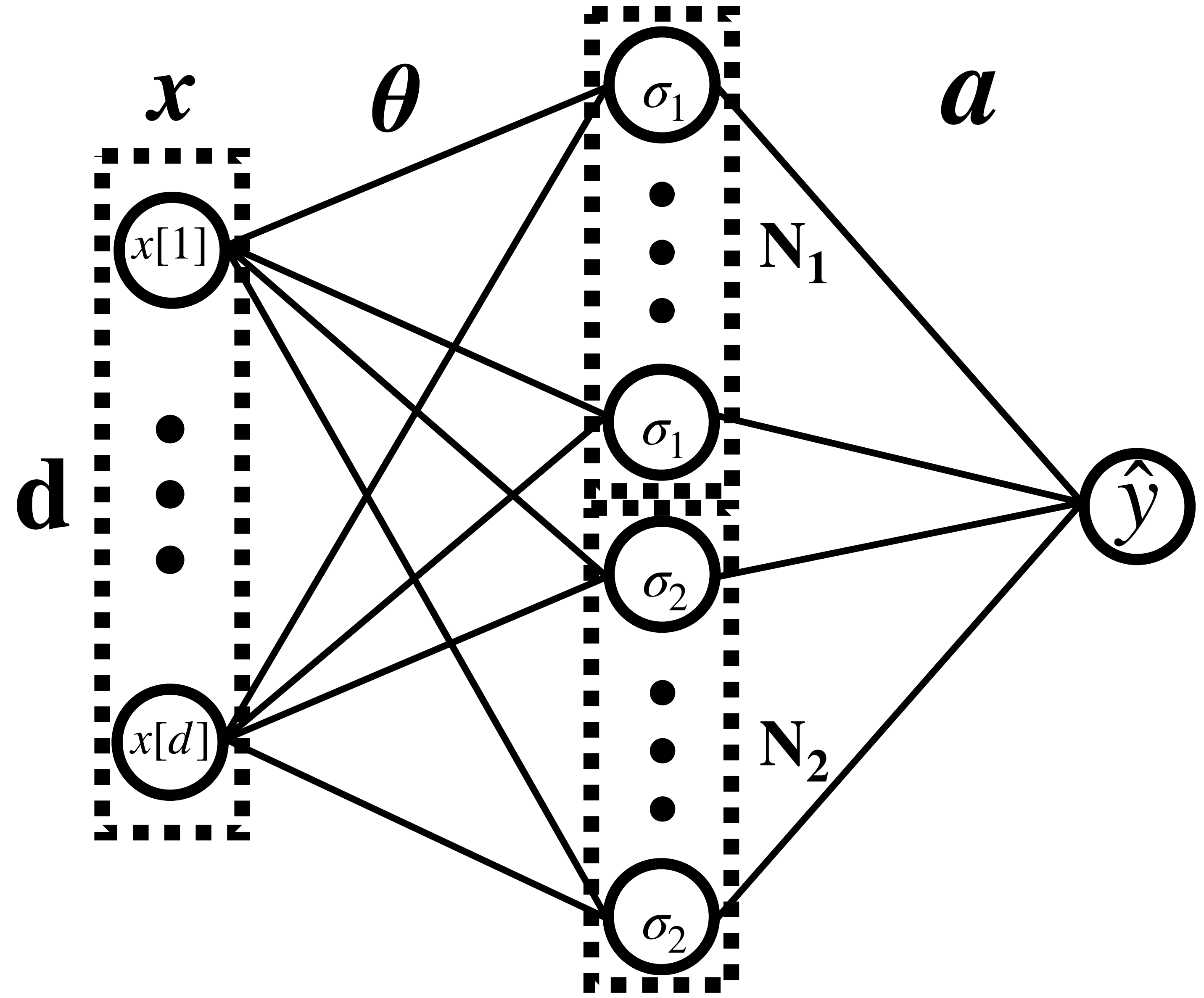
Intuition of Multiple Descent in Multi-component Models



Scale difference may be the key (consider the case $N_1 = N_2$):

Intuition of Multiple Descent in Multi-component Models

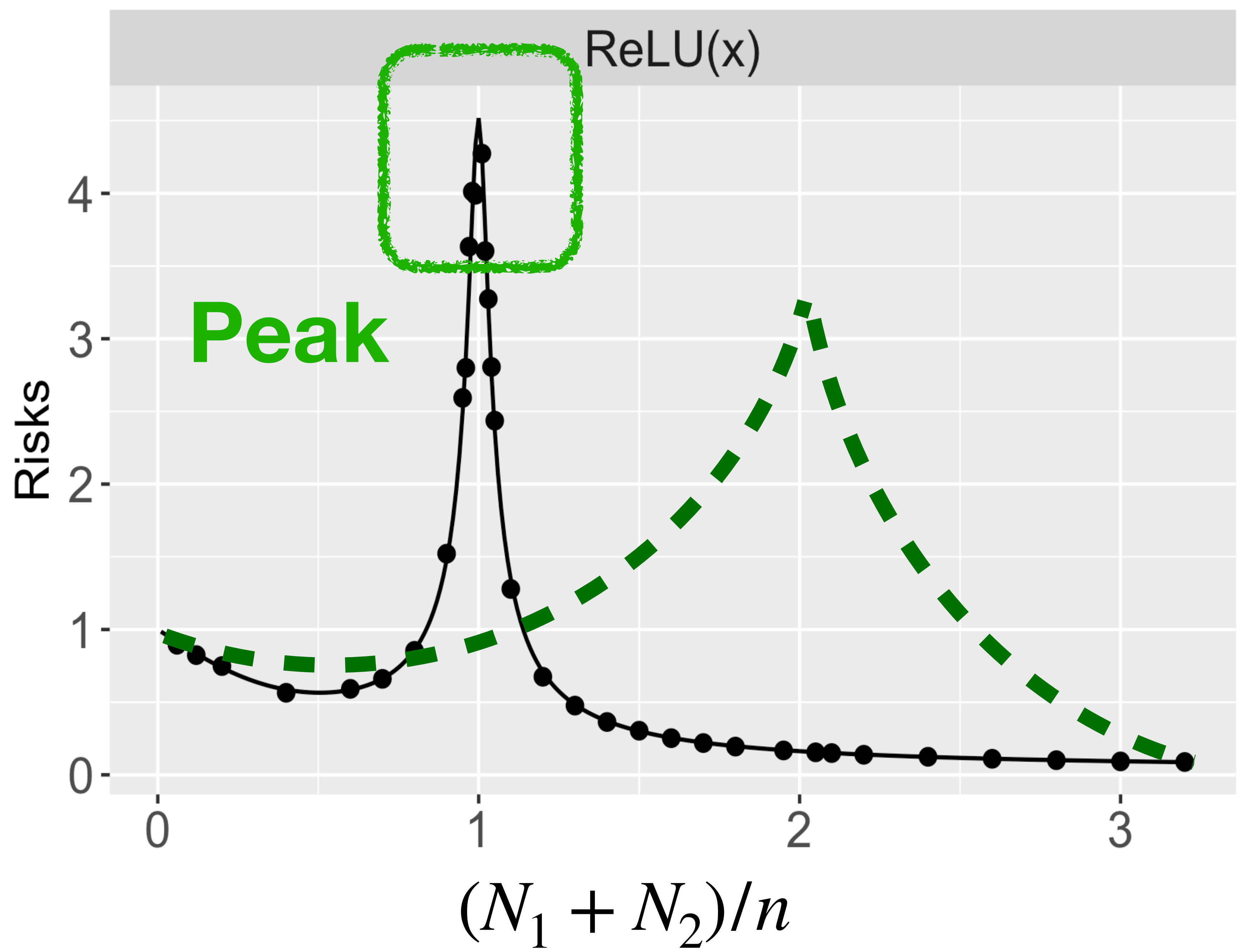


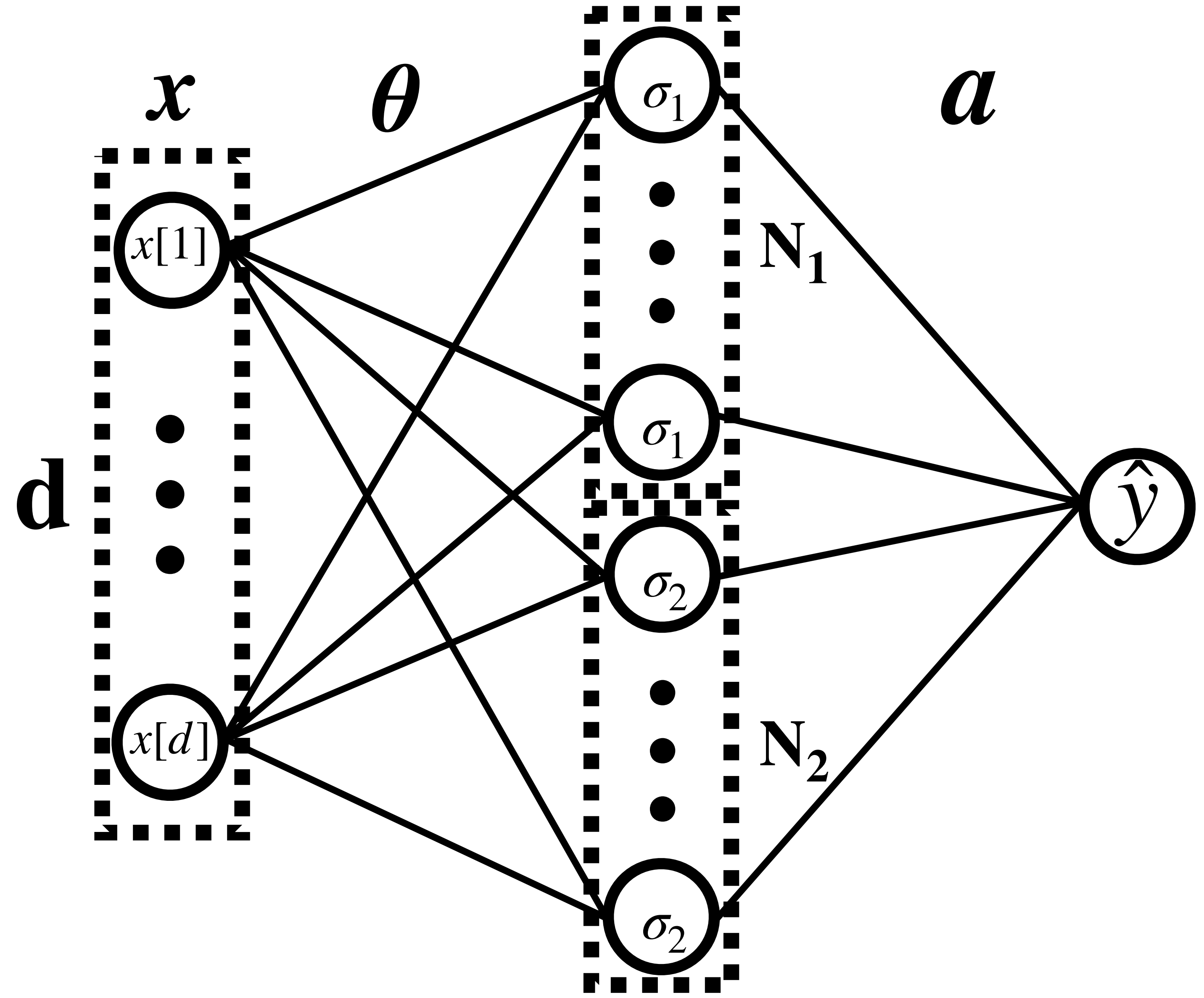


Scale difference may be the key (consider the case $N_1 = N_2$):

- If $\sigma_1(\cdot) = \sigma_2(\cdot)$, double descent exists according to [Mei & Montanari, 2022], and the peak is located at $(N_1 + N_2)/n = 1$.

Intuition of Multiple Descent in Multi-component Models

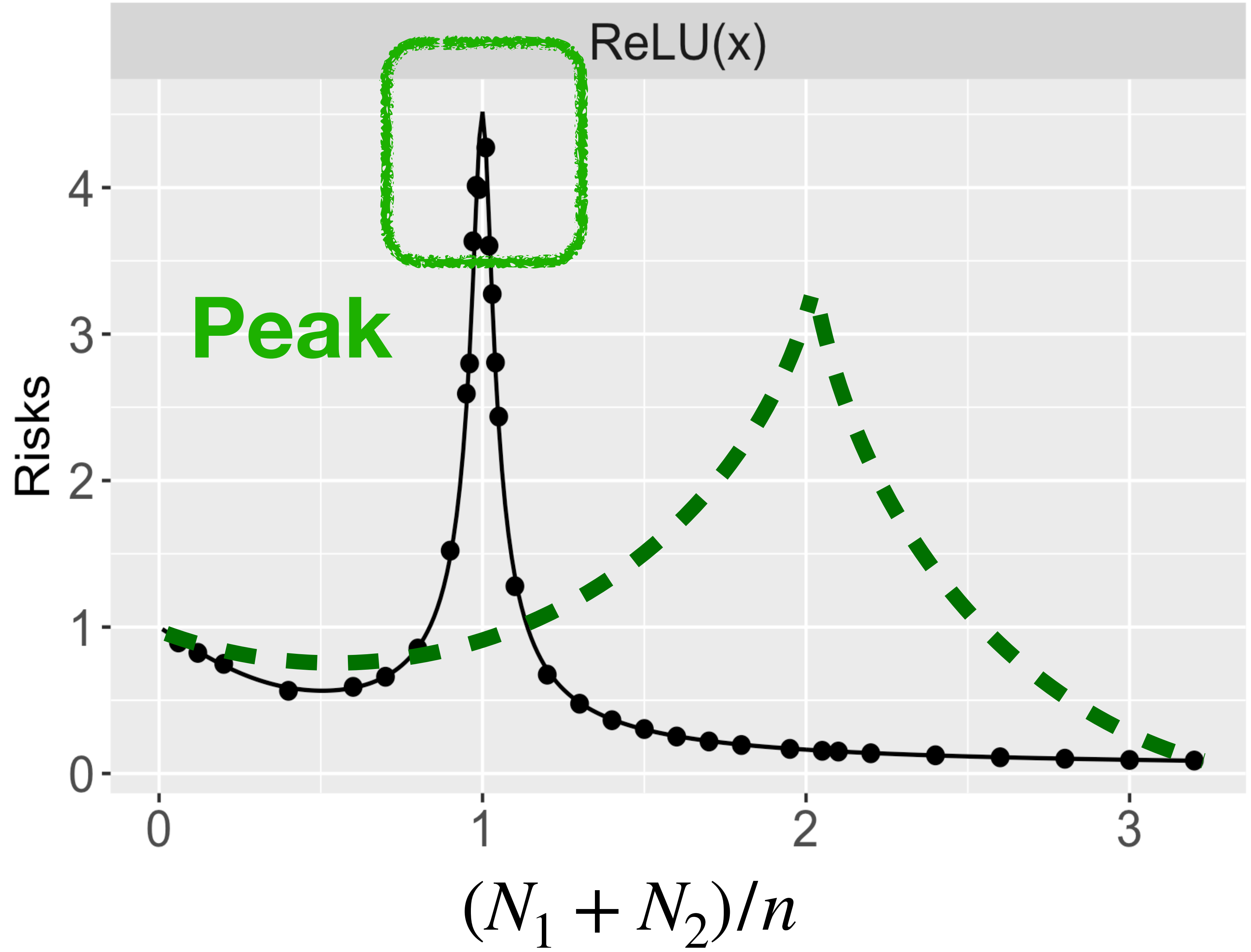


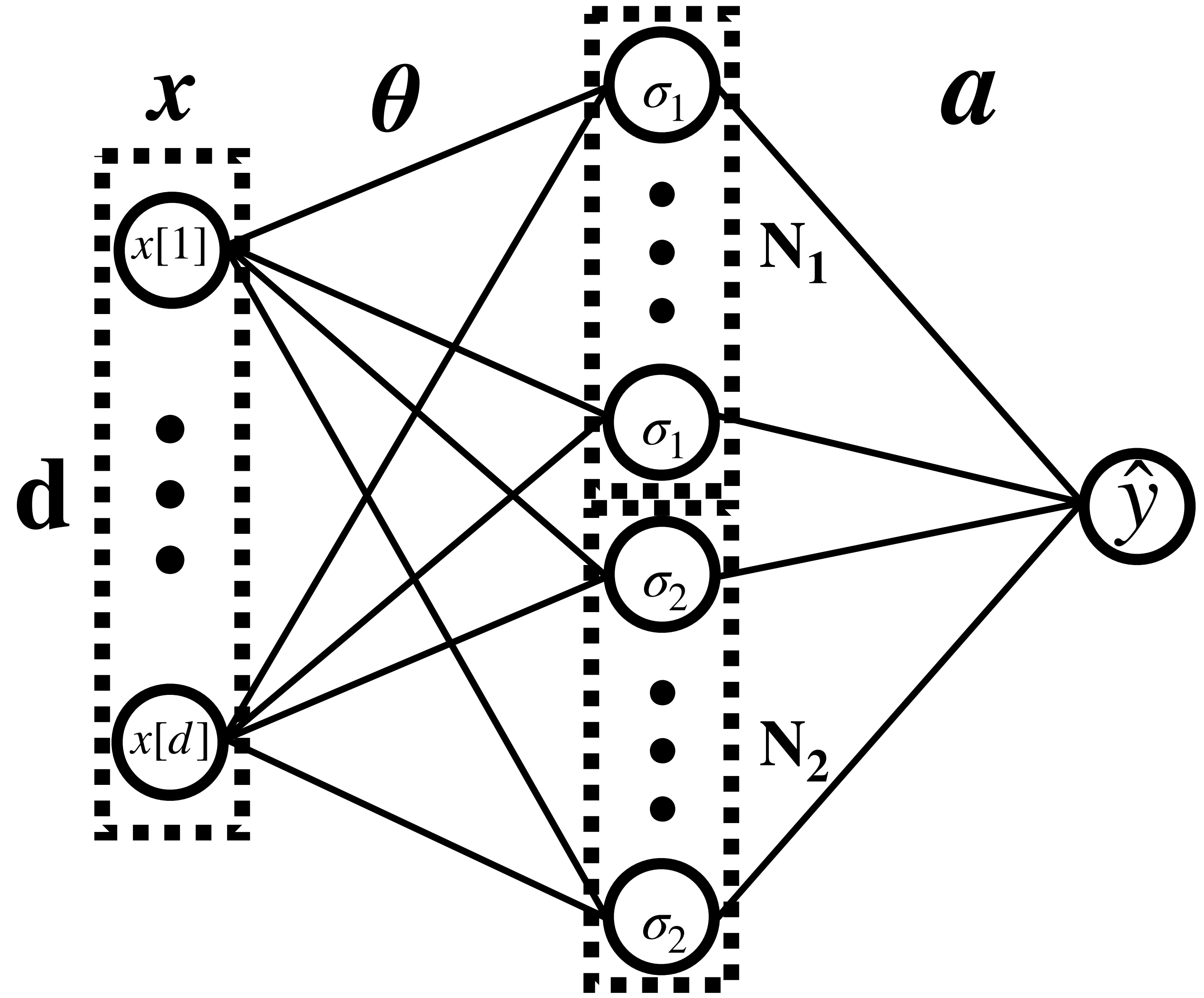


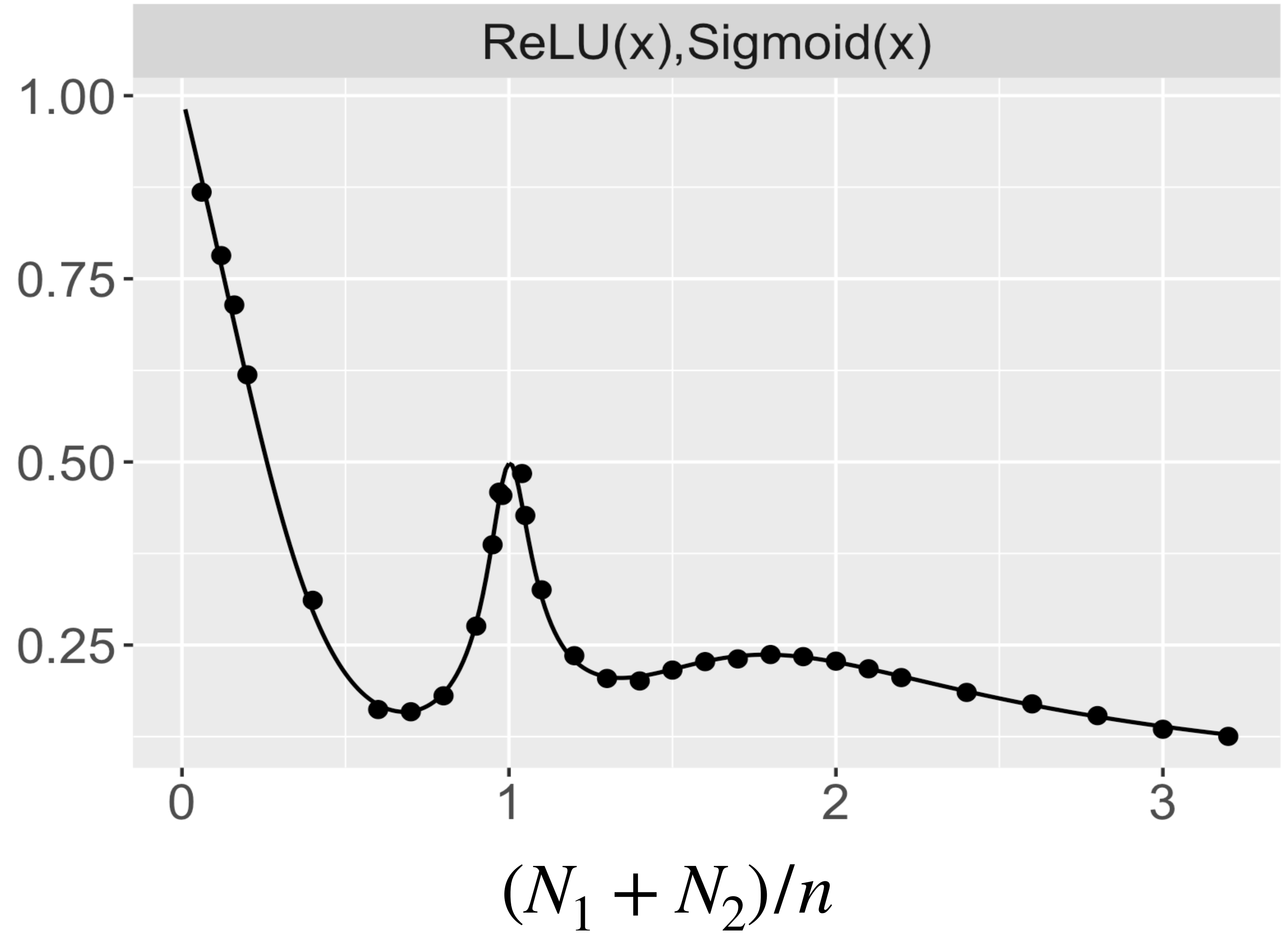
Scale difference may be the key (consider the case $N_1 = N_2$):

- If $\sigma_1(\cdot) = \sigma_2(\cdot)$, double descent exists according to [Mei & Montanari, 2022], and the peak is located at $(N_1 + N_2)/n = 1$.
- If $\sigma_2(\cdot)$ is very small compared with $\sigma_1(\cdot)$, we may also expect double descent according to [Mei & Montanari, 2022], and the peak is at $N_1/n = 1. \rightarrow (N_1 + N_2)/n = 2$

Intuition of Multiple Descent in Multi-component Models







Scale difference may be the key (consider the case $N_1 = N_2$):

An example for $\sigma_1(\cdot) = \mathrm{ReLU}(\cdot)$ and $\sigma_2(\cdot) = \mathrm{Sigmoid}(\cdot)$.

Theoretical Demonstration of Triple Descent in DRFMs

Data distribution:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta}_d + \varepsilon_i,\ i = 1,\ldots,n, \qquad \begin{cases} \mathbf{x}_i \sim \mathrm{Unif}(\sqrt{d} \cdot \mathbb{S}^{d-1}) \\ \varepsilon_i \sim \mathcal{N}(0,\tau^2) \end{cases}$$

Double random feature model

$$\mathscr{F}_{\mathrm{DRF}}(\Theta) = \left\{ f(x; a, \Theta) \equiv \sum_{i=1}^{N_1} a_i \sigma_1 \left(\langle \theta_i, x \rangle / \sqrt{d} \right) + \sum_{i=N_1+1}^{N_1+N_2} a_i \sigma_2 \left(\langle \theta_i, x \rangle / \sqrt{d} \right) : a_i \in \mathbb{R} \quad \forall i \in [N] \right\}$$

Θ: fixed at randomly generated values

a: trainable parameters

Ridge Regression & Limit of Excess Risk

Consider learning the coefficient vector $\mathbf{a}$ via the following loss function:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - f(\mathbf{x}_i; \mathbf{a}, \boldsymbol{\Theta}) \Big)^2 + \frac{d}{n} \lambda \|\mathbf{a}\|_2^2 \right\},$$

where $\lambda > 0$ is the regularization parameter. Moreover, define the excess risk

$$R_d(\mathbf{X}, \boldsymbol{\Theta}, \lambda, \boldsymbol{\beta}_d, \boldsymbol{\varepsilon}) = \mathbb{E}_{\mathbf{x} \sim \mathrm{Unif}(\sqrt{d} \cdot \mathbb{S}^{d-1})} \Big(\mathbf{x}^\top \boldsymbol{\beta}_d - f(\mathbf{x}; \hat{\mathbf{a}}, \boldsymbol{\Theta}) \Big)^2 .$$

Our goal: calculate

$$\lim_{\substack{N_1/d = \psi_1, N_2/d = \psi_2, n/d = \psi_3 \\ N_1, N_2, d, n \to +\infty}} R_d(\mathbf{X}, \boldsymbol{\Theta}, \lambda, \boldsymbol{\beta}_d, \boldsymbol{\varepsilon})$$

and investigate how this limit changes with the ratios ψ_1, ψ_2, ψ_3 when λ is small.
We collect ψ_1, ψ_2, ψ_3 into the vector $\boldsymbol{\psi}$.

Main Assumption

Assumption 1: Let $\sigma_j : \mathbb{R} \to \mathbb{R}$ $(j = 1,2)$ be weakly differentiable, with a weak derivative σ_j'. Assume $|\sigma_j(u)| \vee |\sigma_j'(u)| \leq C_0 e^{C_1|u|}$ for some constants $C_0, C_1 < +\infty$.

- Define spherical moments of σ_j.
  - For $G \sim \mathrm{N}(0,1)$, we define

$$\mu_{j,0} = \mathbb{E}\{\sigma_j(G)\}, \quad \mu_{j,1} = \mathbb{E}\{G\sigma_j(G)\}, \quad \mu_{j,*}^2 = \mathbb{E}\{\sigma_j^2(G)\} - \mu_{j,1}^2 - \mu_{j,0}^2.$$

The sphere moments are collected into the vector $\boldsymbol{\mu}$.

Main Theory for Asymptotic Excess Risk

Theorem. Under Assumption 1, it holds that

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon}} \left| R_d(\mathbf{X}, \boldsymbol{\Theta}, \lambda, \boldsymbol{\beta}_d, \boldsymbol{\varepsilon}) - \mathscr{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, \|\boldsymbol{\beta}_d\|_2, \tau) \right| = o_d(1),$$

where

$$\mathscr{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, \|\boldsymbol{\beta}_d\|_2, \tau) = \|\boldsymbol{\beta}_d\|_2^2 \left(\frac{1}{M_D^2} + \mathbf{L}_{3,4} + \mathbf{L}_{1,4} \right) + \tau^2 (\mathbf{L}_{2,3} + \mathbf{L}_{1,2}) \,.$$

$M_D \in \mathbb{R}$ and $\mathbf{L} \in \mathbb{R}^{4\times 4}$ are given as follows:

Main Theory for Asymptotic Excess Risk

Theorem. Under Assumption 1, it holds that

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon}}\left| R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\boldsymbol{\beta}_d,\boldsymbol{\varepsilon}) - \mathscr{R}(\lambda,\boldsymbol{\psi},\boldsymbol{\mu},\|\boldsymbol{\beta}_d\|_2,\tau)\right| = o_d(1),$$

where

$$\mathscr{R}(\lambda,\boldsymbol{\psi},\boldsymbol{\mu},\|\boldsymbol{\beta}_d\|_2,\tau) = \|\boldsymbol{\beta}_d\|_2^2\left(\frac{1}{M_D^2} + \mathbf{L}_{3,4} + \mathbf{L}_{1,4}\right) + \tau^2(\mathbf{L}_{2,3} + \mathbf{L}_{1,2})\,.$$

$M_D \in \mathbb{R}$ and $\mathbf{L} \in \mathbb{R}^{4\times 4}$ are given as follows:

(1) implicit functions $\nu_1(\xi), \nu_2(\xi), \nu_3(\xi) : \mathbb{C}_+ \to \mathbb{C}_+$ are defined as follows:

$$\nu_1 \cdot \left(-\xi - \mu_{1,*}^2\nu_3 - \frac{\mu_{1,1}^2\nu_3}{1-\mu_{1,1}^2\nu_1\nu_3 - \mu_{2,1}^2\nu_2\nu_3}\right) = \psi_1,$$

$$\nu_2 \cdot \left(-\xi - \mu_{2,*}^2\nu_3 - \frac{\mu_{2,1}^2\nu_3}{1-\mu_{1,1}^2\nu_1\nu_3 - \mu_{2,1}^2\nu_2\nu_3}\right) = \psi_2,$$

$$\nu_3 \cdot \left(-\xi - \mu_{1,*}^2\nu_1 - \mu_{2,*}^2\nu_2 - \frac{\mu_{1,1}^2\nu_1 + \mu_{2,1}^2\nu_2}{1-\mu_{1,1}^2\nu_1\nu_3 - \mu_{2,1}^2\nu_2\nu_3}\right) = \psi_3\,.$$

It can be proved that analytic $\nu_j(\xi)$'s exist and are unique.

Main Theory for Asymptotic Excess Risk

Theorem. Under Assumption 1, it holds that

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon}} \left| R_d(\mathbf{X}, \boldsymbol{\Theta}, \lambda, \boldsymbol{\beta}_d, \boldsymbol{\varepsilon}) - \mathscr{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, \|\boldsymbol{\beta}_d\|_2, \tau) \right| = o_d(1),$$

where

$$\mathscr{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, \|\boldsymbol{\beta}_d\|_2, \tau) = \|\boldsymbol{\beta}_d\|_2^2 \left(\frac{1}{M_D^2} + \mathbf{L}_{3,4} + \mathbf{L}_{1,4} \right) + \tau^2 (\mathbf{L}_{2,3} + \mathbf{L}_{1,2}) \, .$$

$M_D \in \mathbb{R}$ and $\mathbf{L} \in \mathbb{R}^{4\times 4}$ are given as follows:

(2) define $\nu_j^* = \nu_j(\sqrt{\lambda} i), j = 1,2,3$. Let $M_N = \nu_1^* \mu_{1,1}^2 + \nu_2^* \mu_{2,1}^2$, $M_D = \nu_3^* M_N - 1$.

$$\mathbf{H} = \begin{bmatrix} -\frac{\nu_3^{*2}\mu_{1,1}^4}{M_D^2} + \frac{\psi_1}{\nu_1^{*2}} & -\frac{\nu_3^{*2}\mu_{1,1}^2\mu_{2,1}^2}{M_D^2} & -\frac{\mu_{1,1}^2}{M_D^2} - \mu_{1,*}^2 \\ * & -\frac{\nu_3^{*2}\mu_{2,1}^4}{M_D^2} + \frac{\psi_2}{\nu_2^{*2}} & -\frac{\mu_{2,1}^2}{M_D^2} - \mu_{2,*}^2 \\ * & * & -\frac{M_N^2}{M_D^2} + \frac{\psi_3}{\nu_3^{*2}} \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} \mu_{1,*}^2 & 0 & \frac{\mu_{1,1}^2}{M_D^2} & \frac{\nu_3^{*2}\mu_{1,1}^2}{M_D^2} \\ \mu_{2,*}^2 & 0 & \frac{\mu_{2,1}^2}{M_D^2} & \frac{\nu_3^{*2}\mu_{2,1}^2}{M_D^2} \\ 0 & 1 & \frac{M_N^2}{M_D^2} & \frac{1}{M_D^2} \end{bmatrix},$$

($\mathbf{H}$ is symmetric here). Define $\mathbf{L} = \mathbf{V}^\top \mathbf{H}^{-1} \mathbf{V}$.

Theoretical Demonstration of Triple Descent

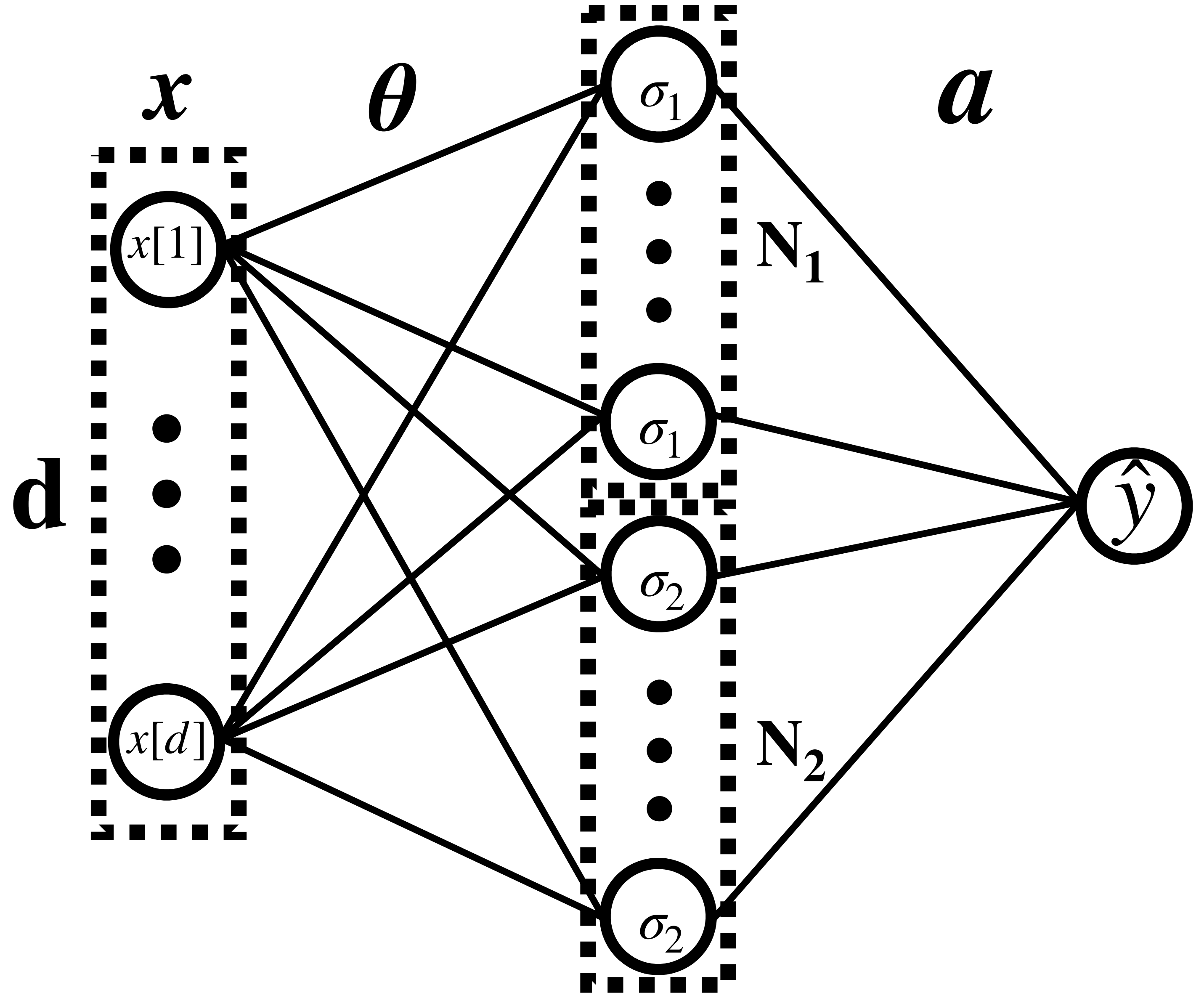
Proposition. For $\mathscr{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, \|\boldsymbol{\beta}_d\|_2, \tau)$, it holds that

1. *When* $(\psi_1 + \psi_2)/\psi_3 = c_1 < 1$, $\lim_{\lambda \to 0} \mathcal{R} < +\infty$;
2. *When* $(\psi_1 + \psi_2)/\psi_3 = 1$, $\lim_{\lambda \to 0} \mathcal{R} = +\infty$;
3. *When* $1 < (\psi_1 + \psi_2)/\psi_3 = c_2 < 1 + \psi_2/\psi_1$, $\varliminf_{\mu_{2,1}, \mu_{2,*} \to 0} \lim_{\lambda \to 0} \mathcal{R} < +\infty$;
4. *When* $(\psi_1 + \psi_2)/\psi_3 = 1 + \psi_2/\psi_1$, $\lim_{\mu_{2,1}, \mu_{2,*} \to 0} \lim_{\lambda \to 0} \mathcal{R} = +\infty$.
5. *For any* $0 < r < \infty$, $\lim_{\substack{\psi_1, \psi_2 \to \infty \\ \psi_1/\psi_2 = r}} \mathcal{R} < +\infty$

Theoretical Demonstration of Triple Descent

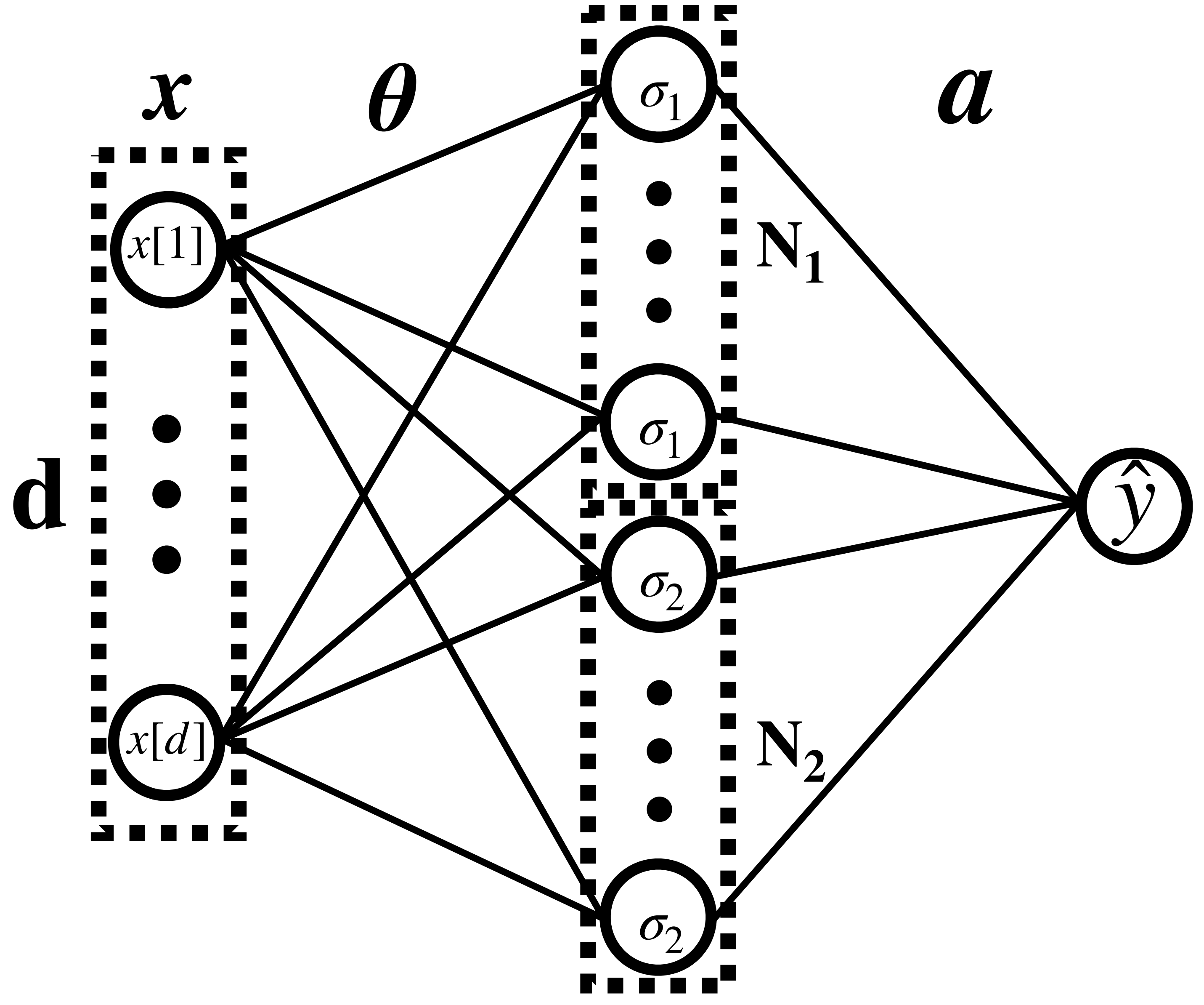
Proposition. For $\mathscr{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, \|\boldsymbol{\beta}_d\|_2, \tau)$, it holds that

1. *When* $(\psi_1 + \psi_2)/\psi_3 = c_1 < 1$, $\lim_{\lambda \to 0} \mathcal{R} < +\infty$;
2. *When* $(\psi_1 + \psi_2)/\psi_3 = 1$, $\lim_{\lambda \to 0} \mathcal{R} = +\infty$;
3. *When* $1 < (\psi_1 + \psi_2)/\psi_3 = c_2 < 1 + \psi_2/\psi_1$, $\underline{\lim}_{\mu_{2,1}, \mu_{2,*} \to 0} \lim_{\lambda \to 0} \mathcal{R} < +\infty$;
4. *When* $(\psi_1 + \psi_2)/\psi_3 = 1 + \psi_2/\psi_1$, $\lim_{\mu_{2,1}, \mu_{2,*} \to 0} \lim_{\lambda \to 0} \mathcal{R} = +\infty$.
5. *For any* $0 < r < \infty$, $\lim_{\substack{\psi_1, \psi_2 \to \infty \\ \psi_1/\psi_2 = r}} \mathcal{R} < +\infty$



$N_1/d \to \psi_1, N_2/d \to \psi_2, n/d \to \psi_3$

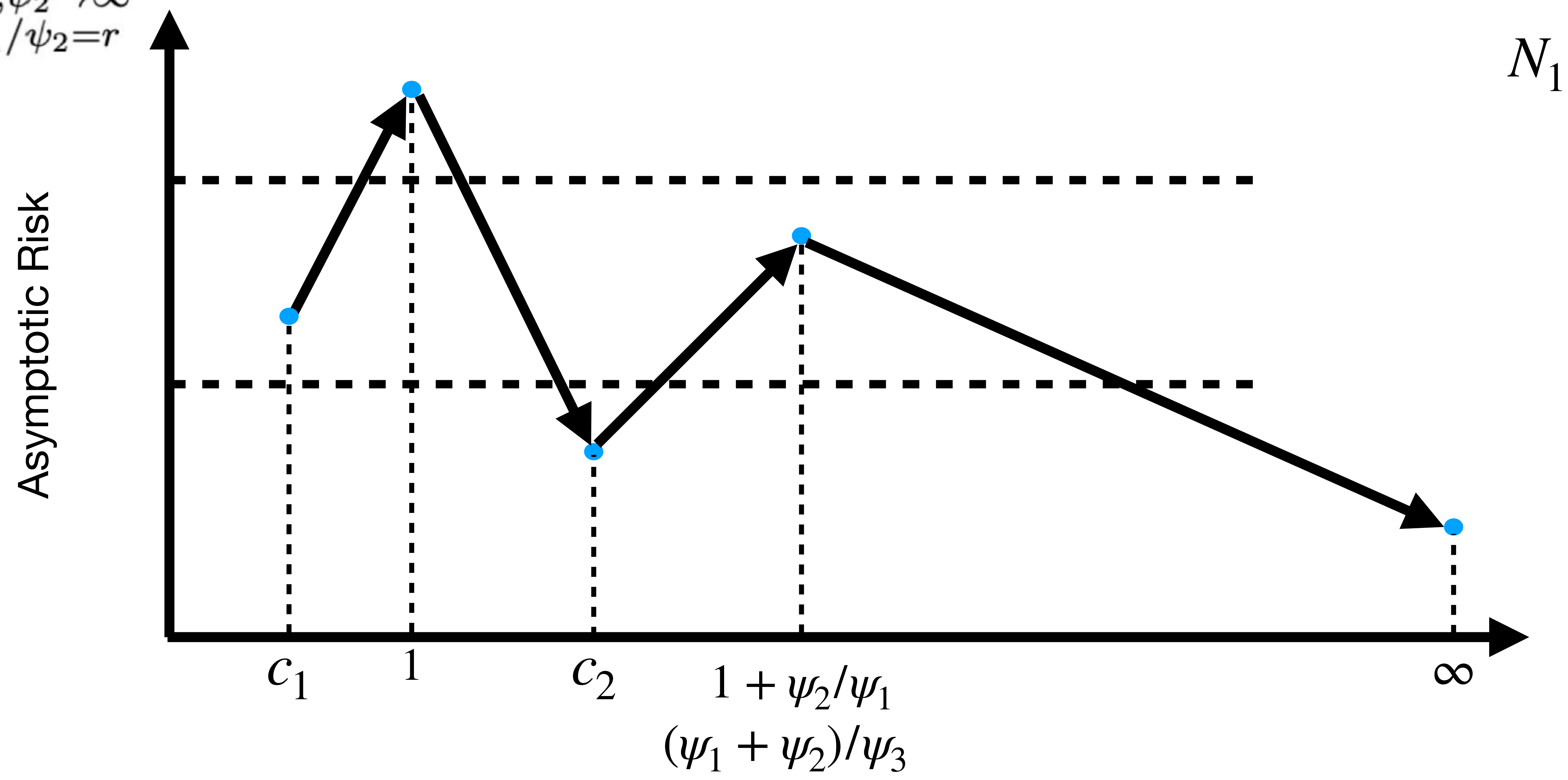
Theoretical Demonstration of Triple Descent



$N_1/d \to \psi_1, N_2/d \to \psi_2, n/d \to \psi_3$

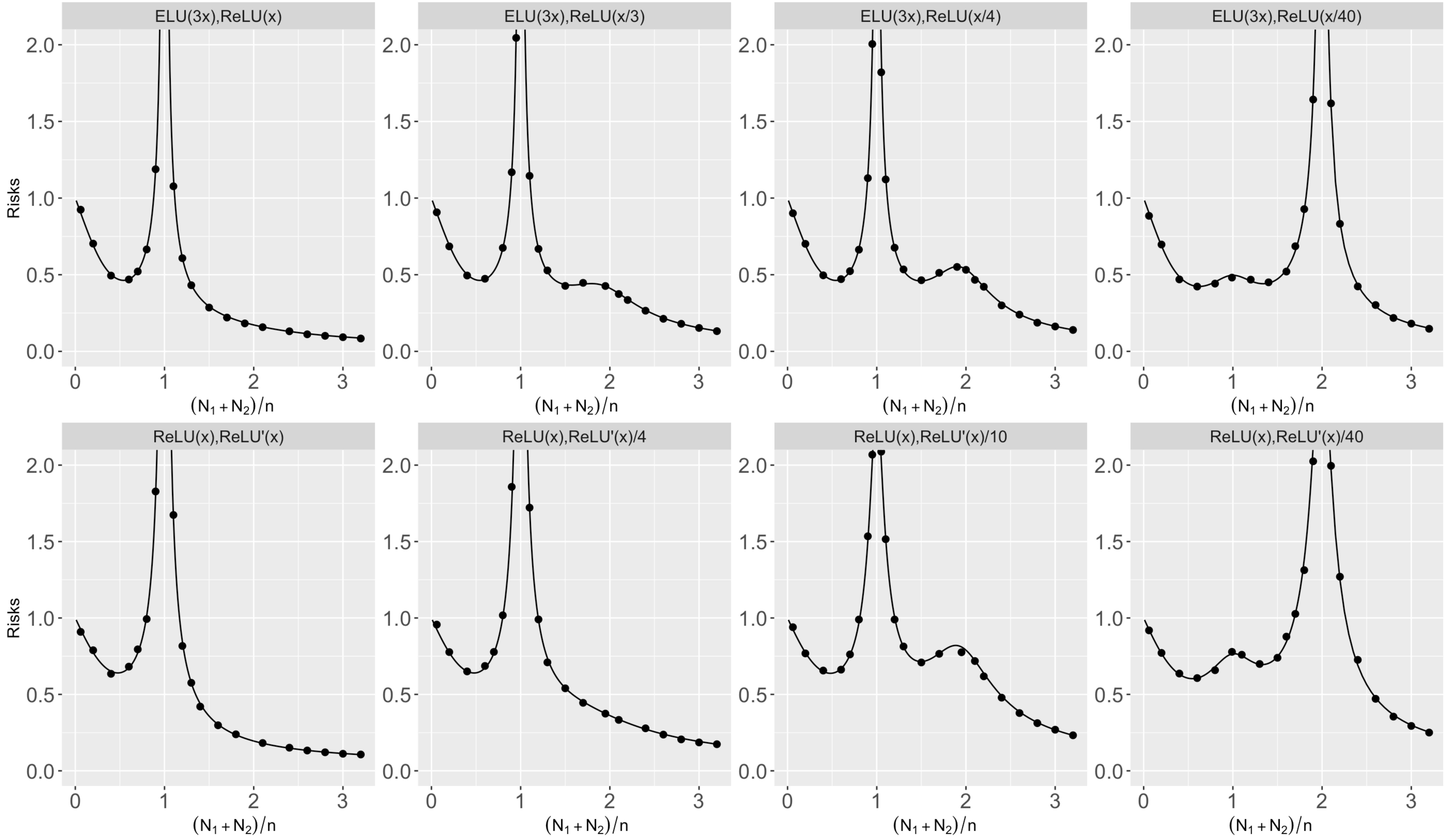
Proposition. For $\mathscr{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, \|\boldsymbol{\beta}_d\|_2, \tau)$, it holds that

1. *When* $(\psi_1+\psi_2)/\psi_3 = c_1 < 1$, $\lim_{\lambda \to 0} \mathcal{R} < +\infty$;
2. *When* $(\psi_1+\psi_2)/\psi_3 = 1$, $\lim_{\lambda \to 0} \mathcal{R} = +\infty$;
3. *When* $1 < (\psi_1+\psi_2)/\psi_3 = c_2 < 1 + \psi_2/\psi_1$, $\underline{\lim}_{\mu_{2,1},\mu_{2,*} \to 0} \lim_{\lambda \to 0} \mathcal{R} < +\infty$;
4. *When* $(\psi_1+\psi_2)/\psi_3 = 1 + \psi_2/\psi_1$, $\lim_{\mu_{2,1},\mu_{2,*} \to 0} \lim_{\lambda \to 0} \mathcal{R} = +\infty$.
5. *For any* $0 < r < \infty$, $\lim_{\substack{\psi_1,\psi_2 \to \infty \\ \psi_1/\psi_2 = r}} \mathcal{R} < +\infty$



Simulations

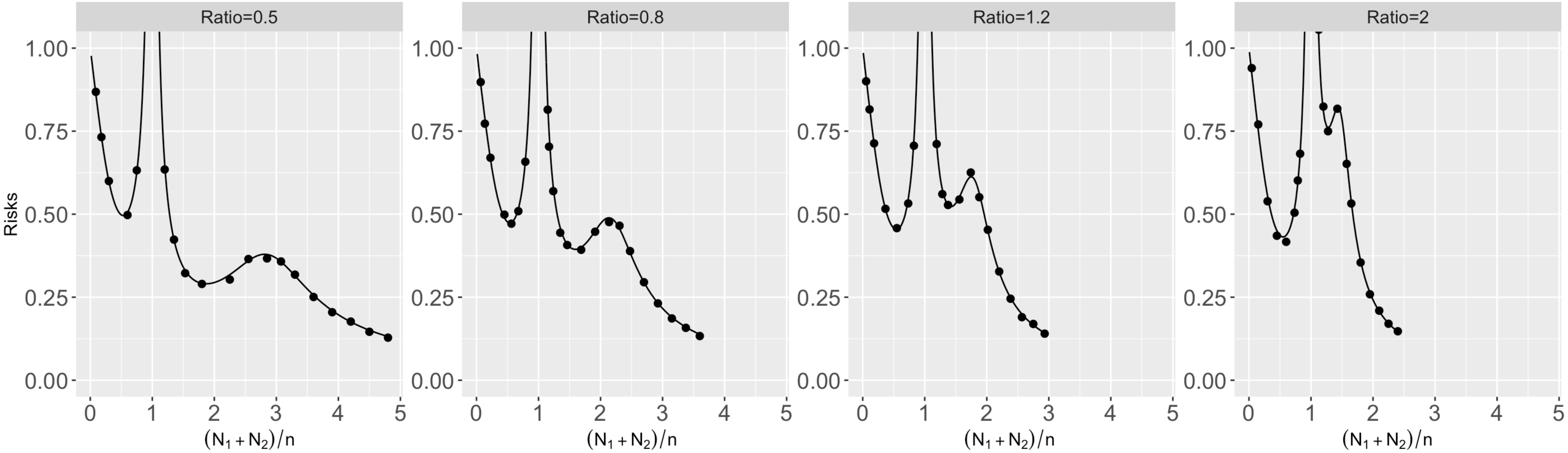
The scale difference of activation functions:



Simulations

Impact of the ratio N_1/N_2

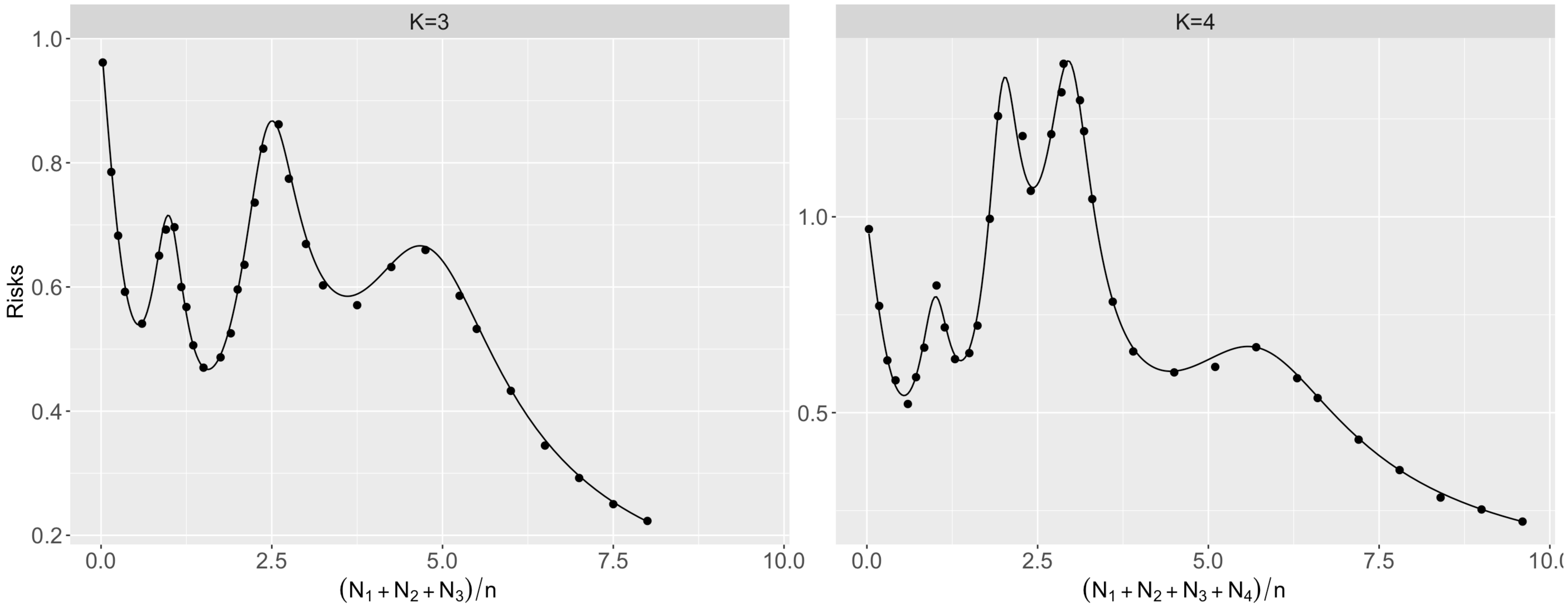
$N_1/N_2 = 0.5$ $\quad$ $N_1/N_2 = 0.8$ $\quad$ $N_1/N_2 = 1.2$ $\quad$ $N_1/N_2 = 2$



Peaks Location: $1 + N_2/N_1$ $\longrightarrow$ $(N_1 + N_2)/n = 3, \quad 9/4, \quad 11/6, \quad 3/2.$

Simulations

Multiple descent when $K > 2$.



quadruple descent **quintuple descent**

Conclusions

- We demonstrate that risk curves with a specific number of descent generally exist in learning multi-component prediction models.

Conclusions

- We demonstrate that risk curves with a specific number of descent generally exist in learning multi-component prediction models.
- We give an intuitive explanation of multiple descent and highlight that appropriate scale differences between the components may be the key.

Conclusions

- We demonstrate that risk curves with a specific number of descent generally exist in learning multi-component prediction models.
- We give an intuitive explanation of multiple descent and highlight that appropriate scale differences between the components may be the key.
- Our explanation of multiple descent can successfully predict the shapes and peak locations in simulations.

Conclusions

- We demonstrate that risk curves with a specific number of descent generally exist in learning multi-component prediction models.

- We give an intuitive explanation of multiple descent and highlight that appropriate scale differences between the components may be the key.

- Our explanation of multiple descent can successfully predict the shapes and peak locations in simulations.

- We give rigorous theoretical demonstration of multiple descent under the setting of learning “multiple random feature models”.

Thank you!