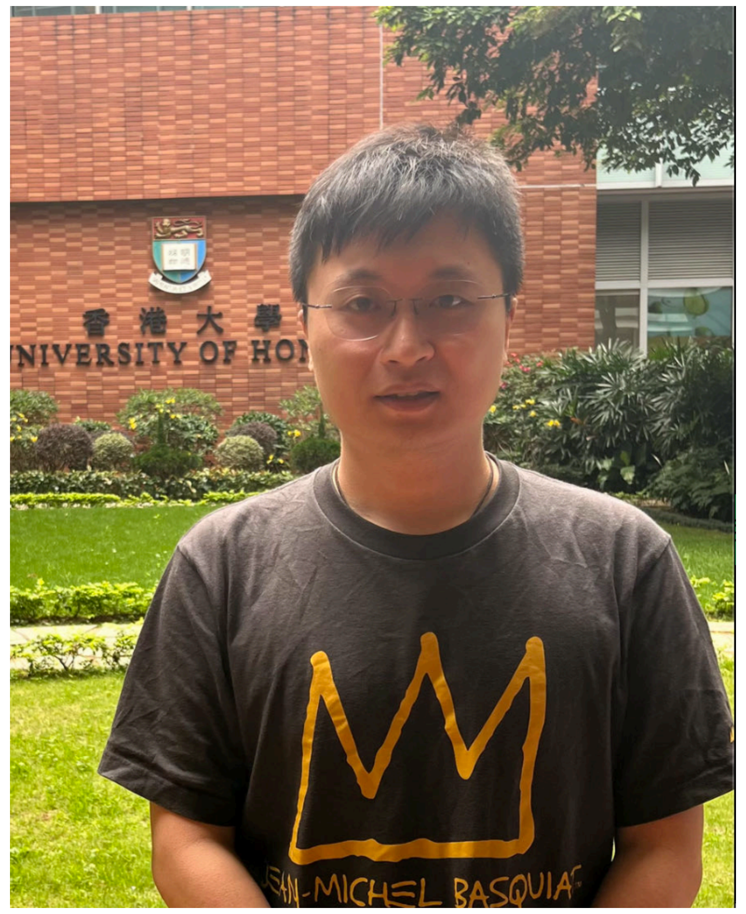
Multiple Descent in the Multiple Random Feature Model

Yuan Cao

Department of Statistics and Actuarial Science
University of Hong Kong



Joint work with **Xuran Meng** and **Jianfeng Yao**

A Simple Question in Linear Regression

Consider

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

A Simple Question in Linear Regression

Consider

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

and its linear ridgeless regression estimator (minimum ℓ_2-norm estimator) is then

$$\hat{\boldsymbol{\beta}} = \lim_{\lambda \to 0^+} \hat{\boldsymbol{\beta}}_\lambda, \quad \hat{\boldsymbol{\beta}}_\lambda = \min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\beta}^\top \mathbf{x}_i - y_i)^2 + \lambda \|\boldsymbol{\beta}\|_2^2,$$

A Simple Question in Linear Regression

Consider

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

and its linear ridgeless regression estimator (minimum ℓ_2-norm estimator) is then

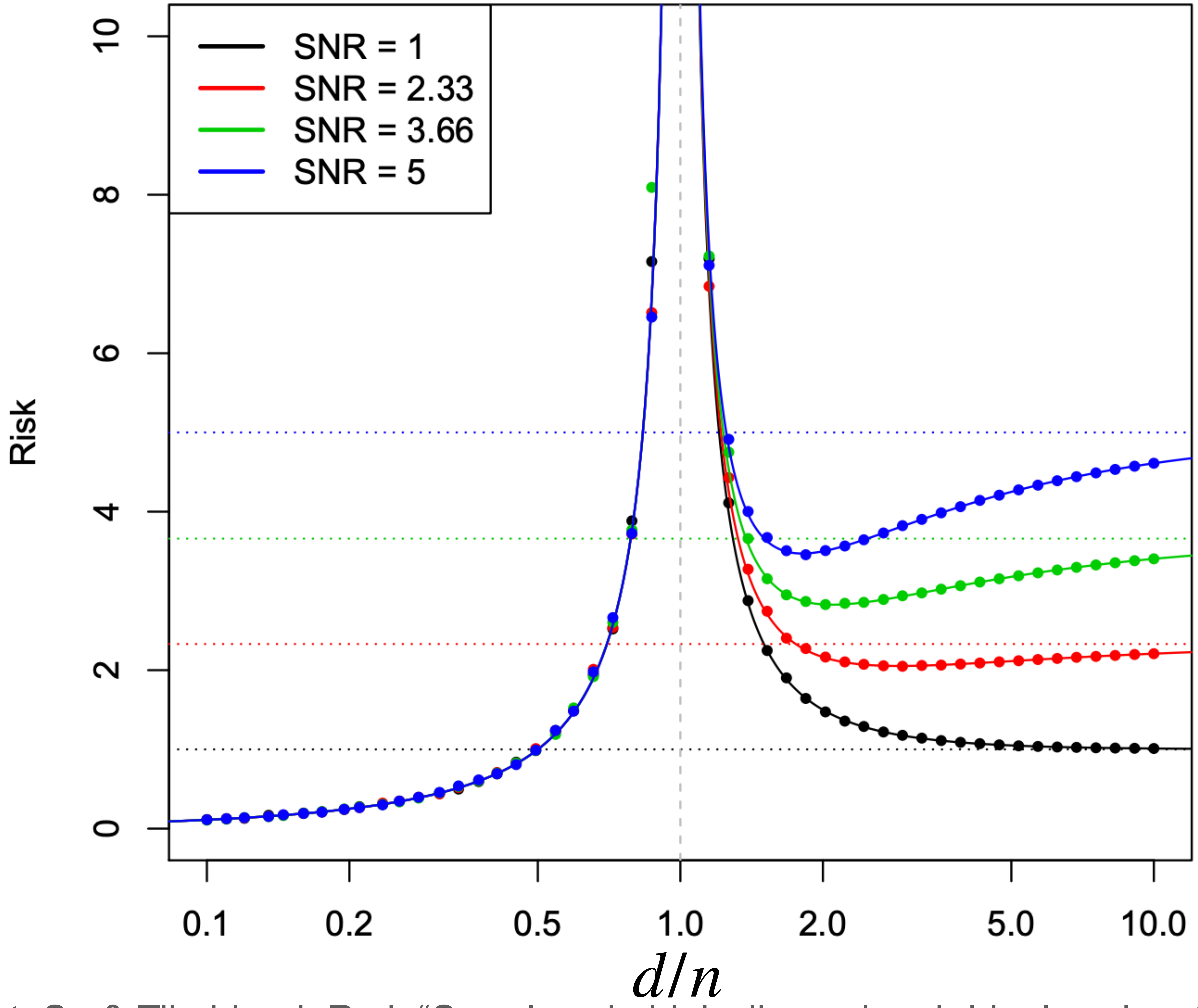
$$\hat{\boldsymbol{\beta}} = \lim_{\lambda\to 0^+} \hat{\boldsymbol{\beta}}_\lambda, \quad \hat{\boldsymbol{\beta}}_\lambda = \min_{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\beta}^\top \mathbf{x}_i - y_i)^2 + \lambda \|\boldsymbol{\beta}\|_2^2,$$

Suppose that the sample size is fixed as a large constant (e.g., $n = 200$). How will the excess risk

$$R(\hat{\boldsymbol{\beta}}) := \mathbb{E}_{\mathbf{x}_{\text{test}}}(\hat{\boldsymbol{\beta}}^\top \mathbf{x}_{\text{test}} - \boldsymbol{\beta}^\top \mathbf{x}_{\text{test}})^2$$

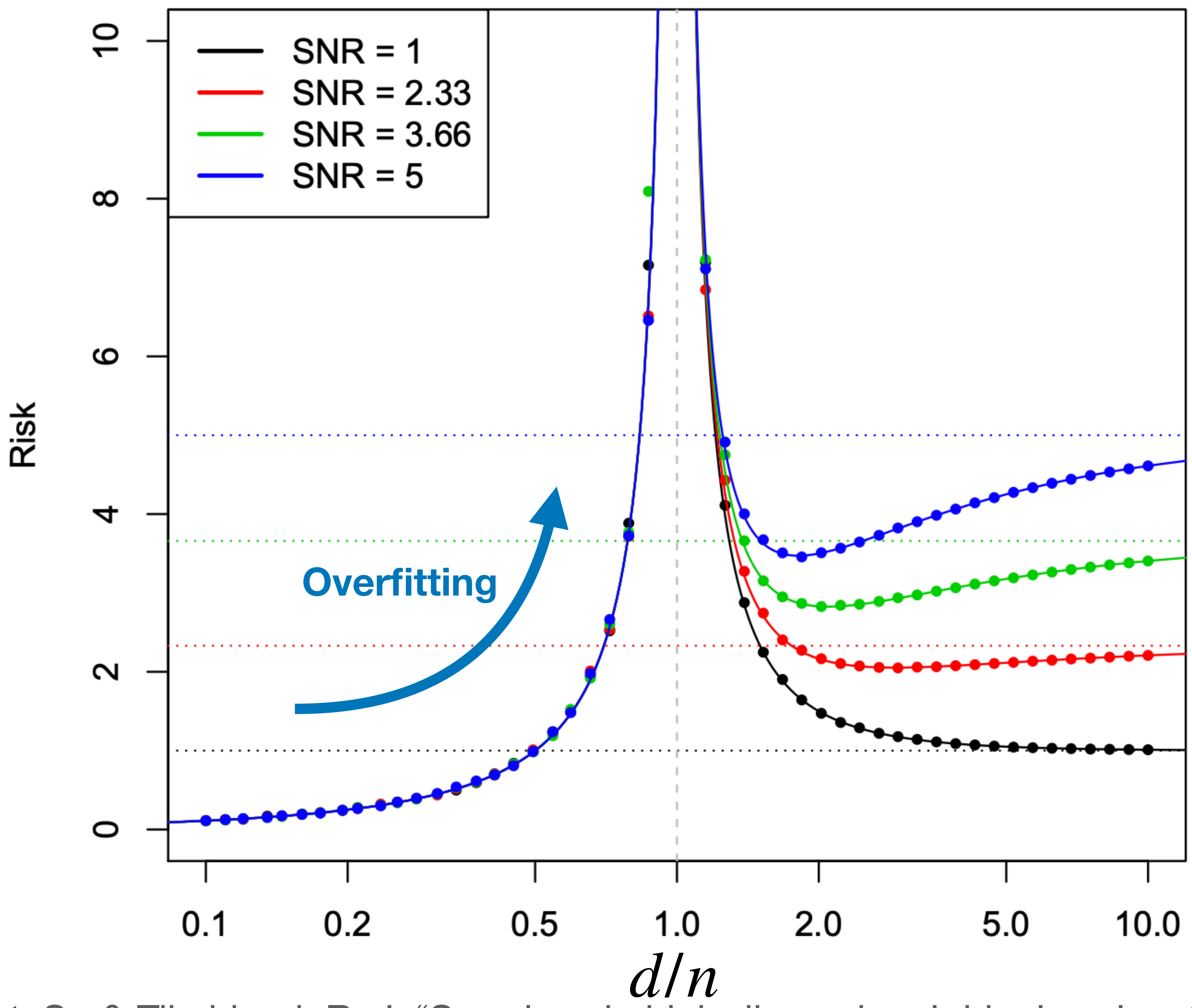
change as d grows from $d < n$ to $d = n$ then to $d > n$? ($\|\boldsymbol{\beta}\|_2$ is fixed.)

A Surprising Observation



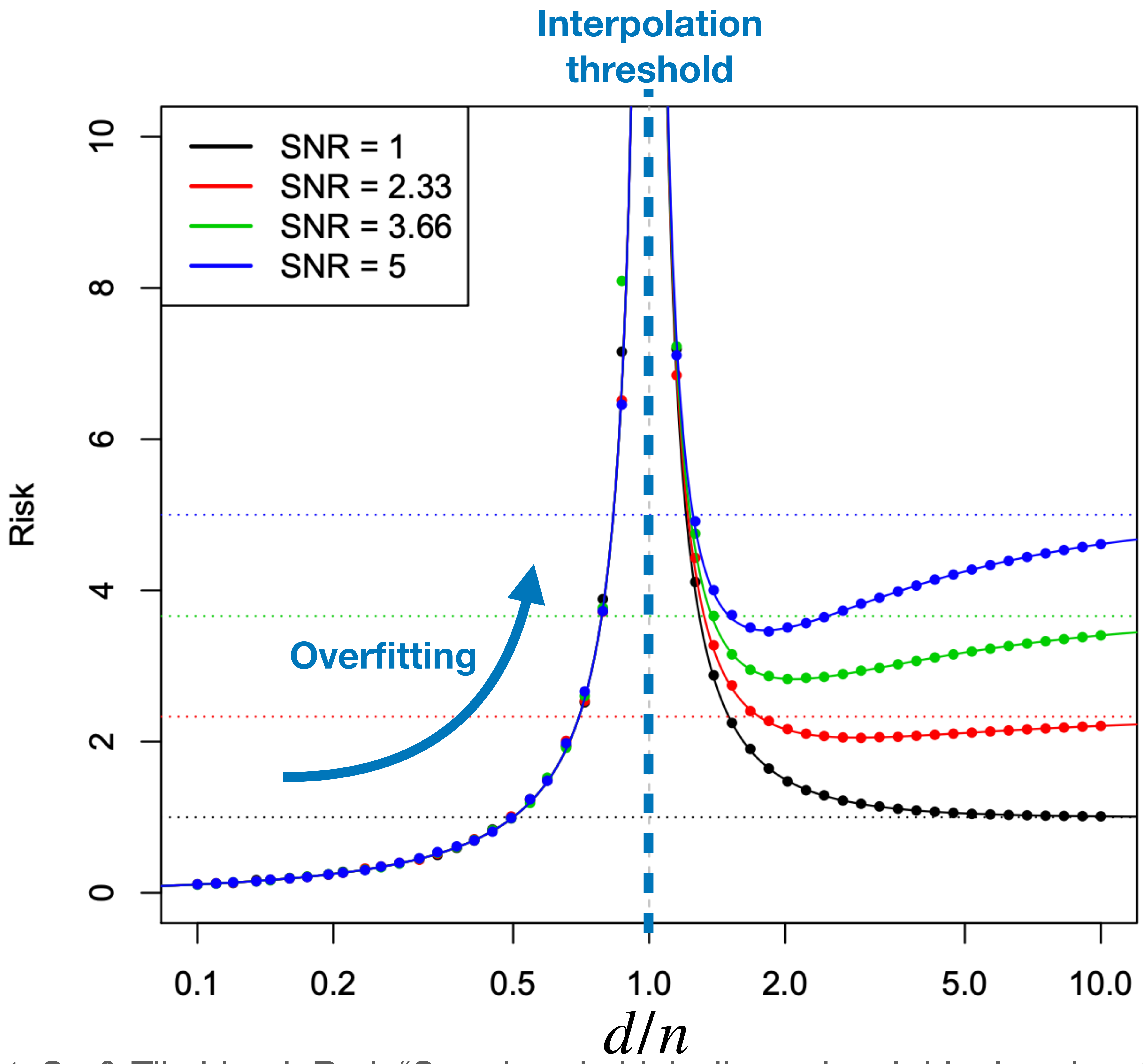
Hastie, T., Montanari, A., Rosset, S., & Tibshirani, R. J. "Surprises in high-dimensional ridgeless least squares interpolation". The Annals of Statistics, 50(2), 949-986, 2022.

A Surprising Observation



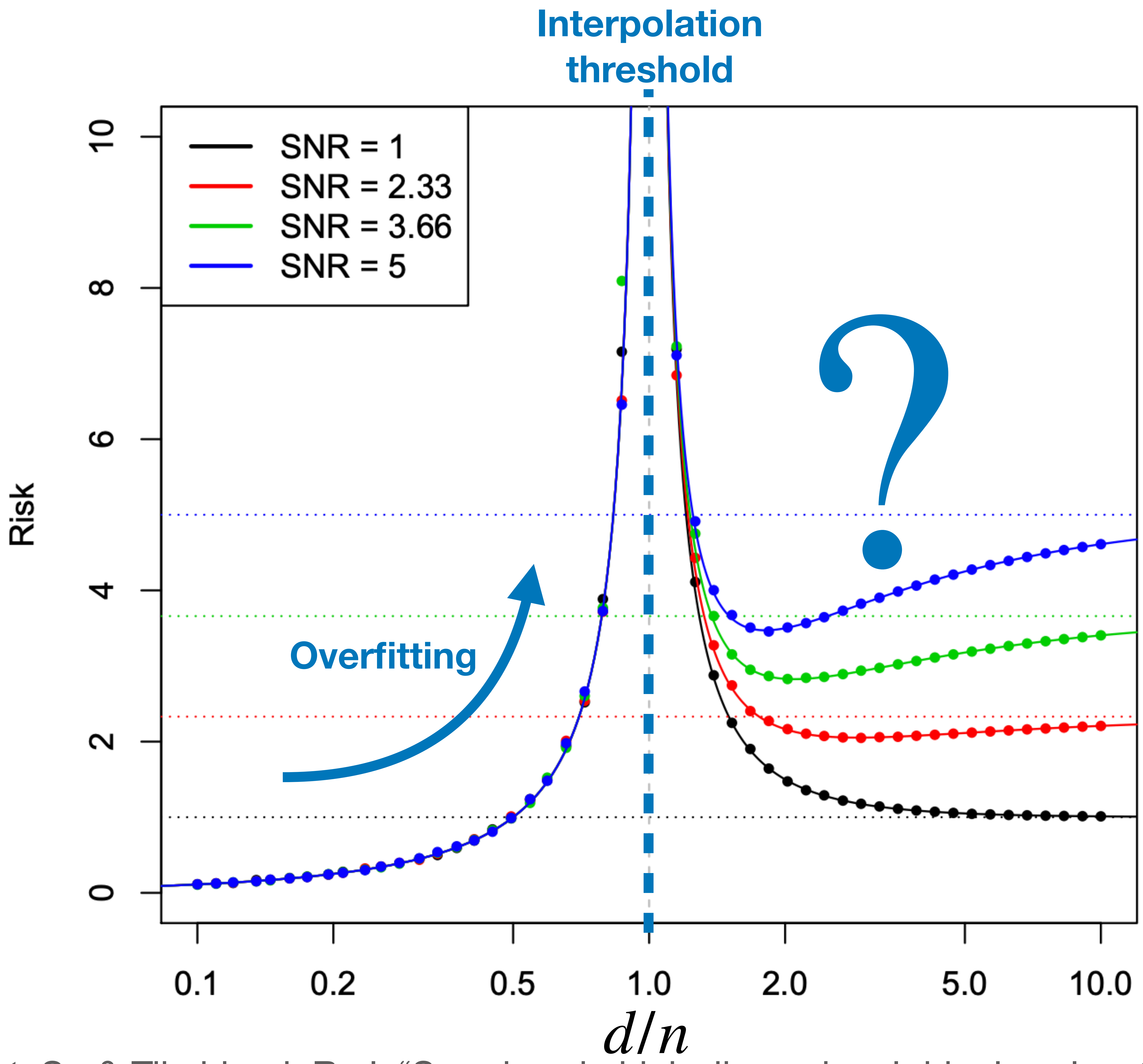
Hastie, T., Montanari, A., Rosset, S., & Tibshirani, R. J. “Surprises in high-dimensional ridgeless least squares interpolation”. The Annals of Statistics, 50(2), 949-986, 2022.

A Surprising Observation



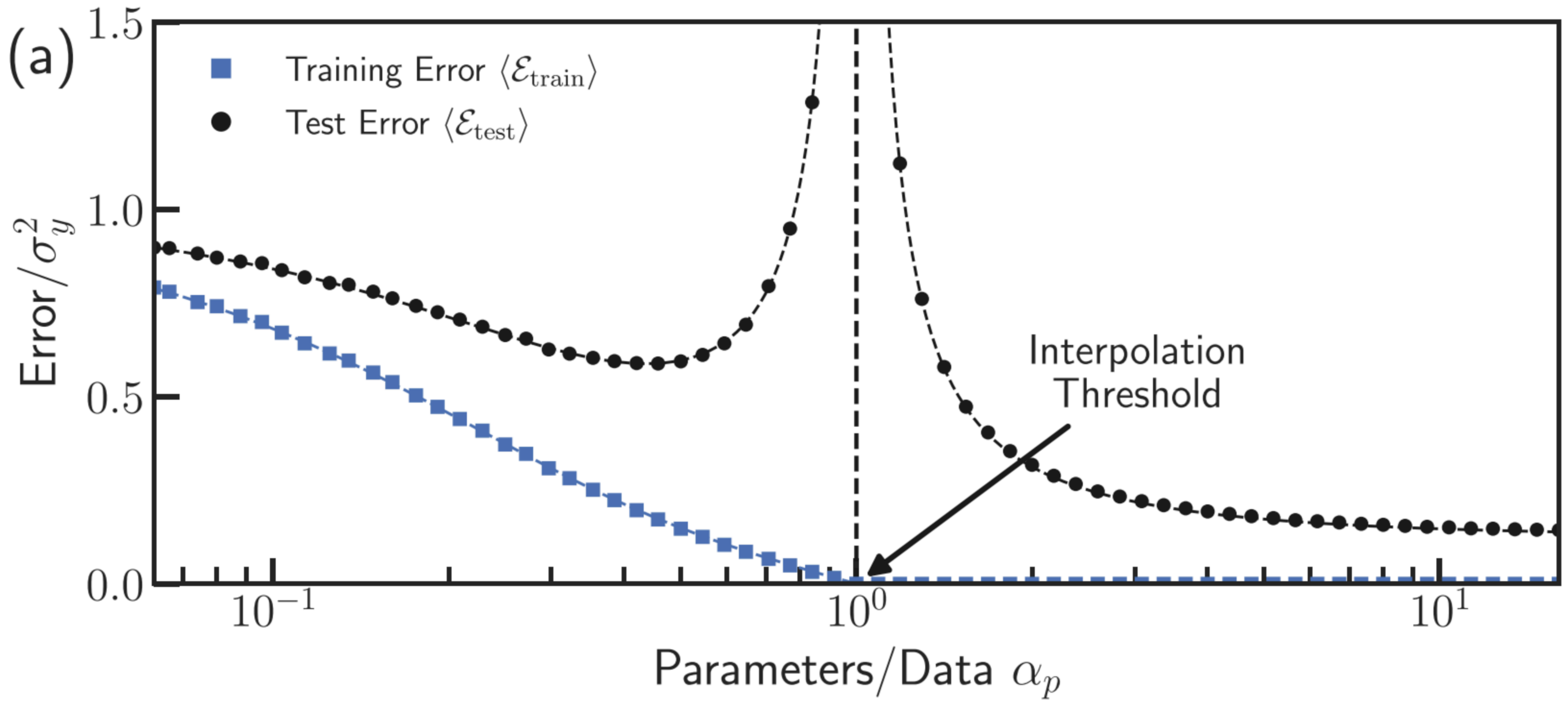
Hastie, T., Montanari, A., Rosset, S., & Tibshirani, R. J. “Surprises in high-dimensional ridgeless least squares interpolation”. The Annals of Statistics, 50(2), 949-986, 2022.

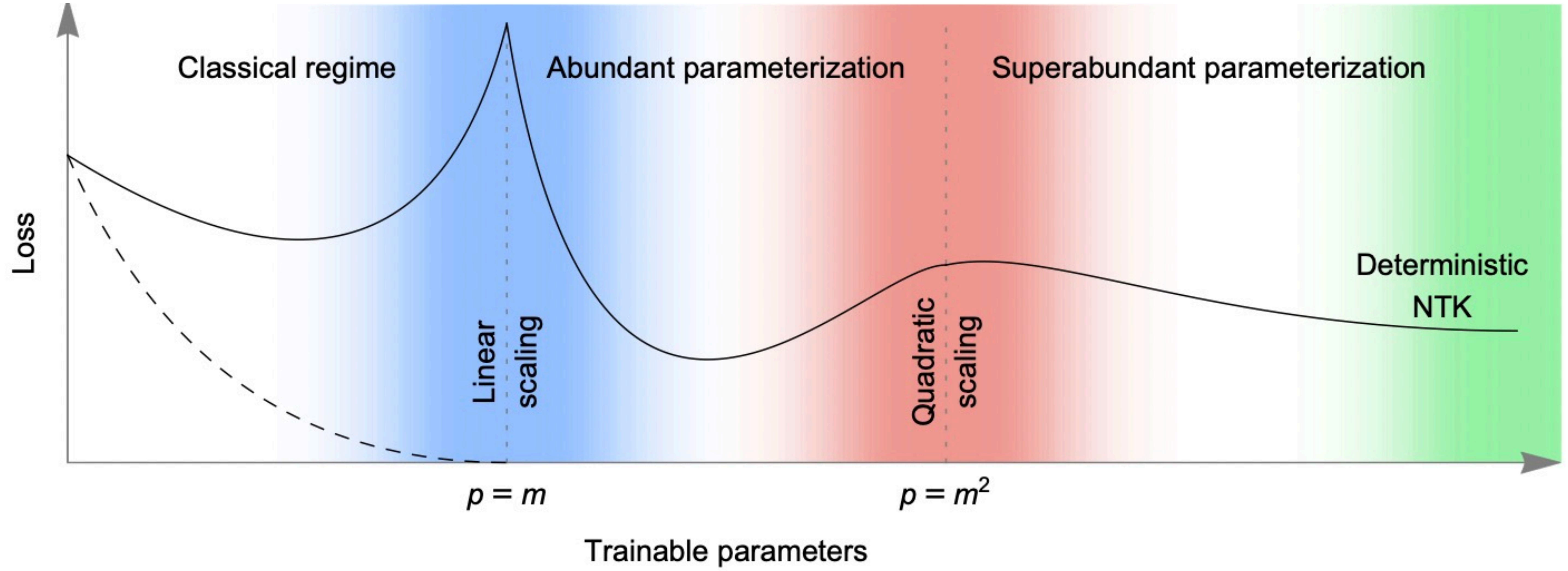
A Surprising Observation



Hastie, T., Montanari, A., Rosset, S., & Tibshirani, R. J. “Surprises in high-dimensional ridgeless least squares interpolation”. The Annals of Statistics, 50(2), 949-986, 2022.

The Double/Multiple Descent Phenomenon

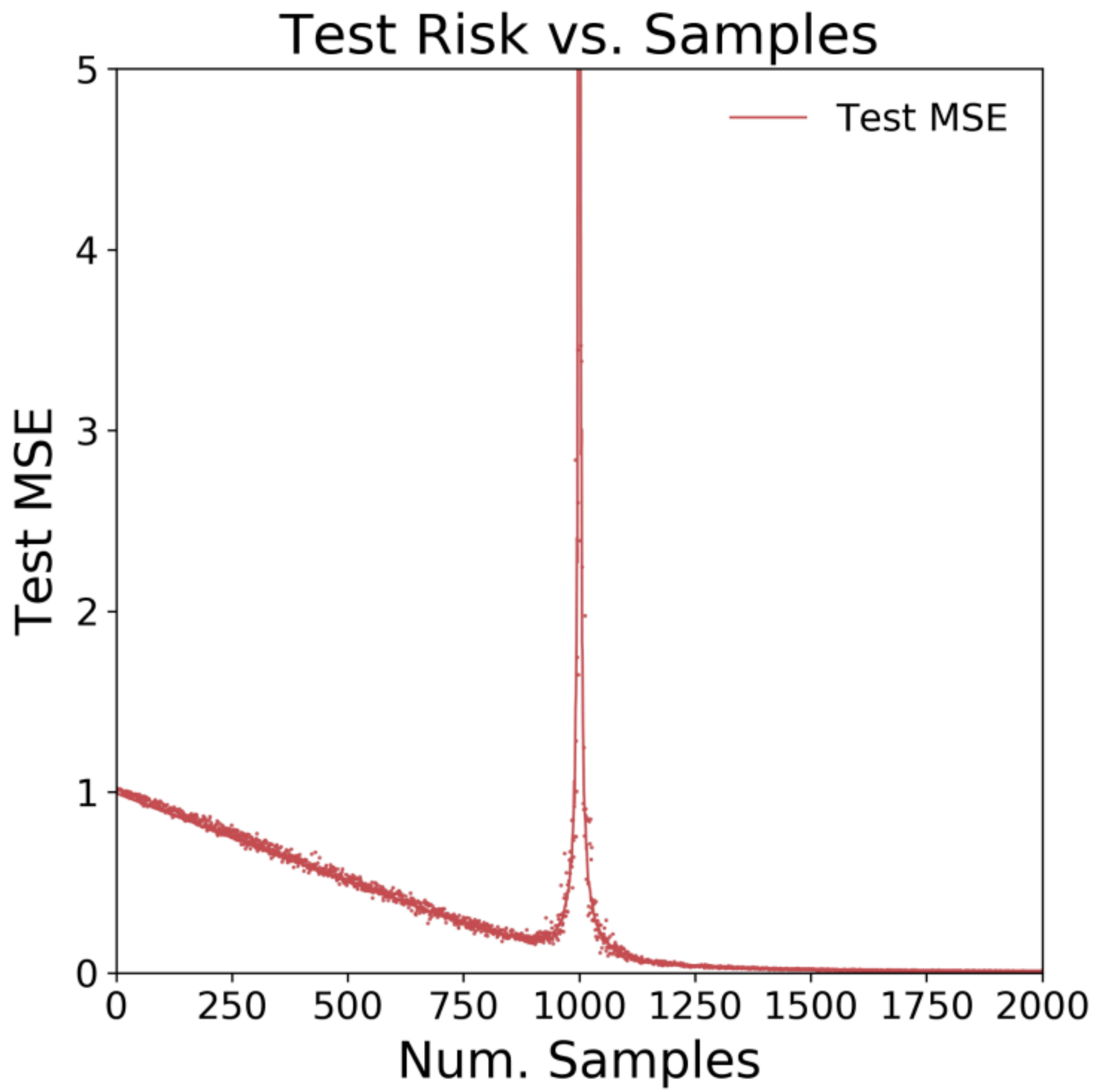




https://en.wikipedia.org/wiki/Double_descent

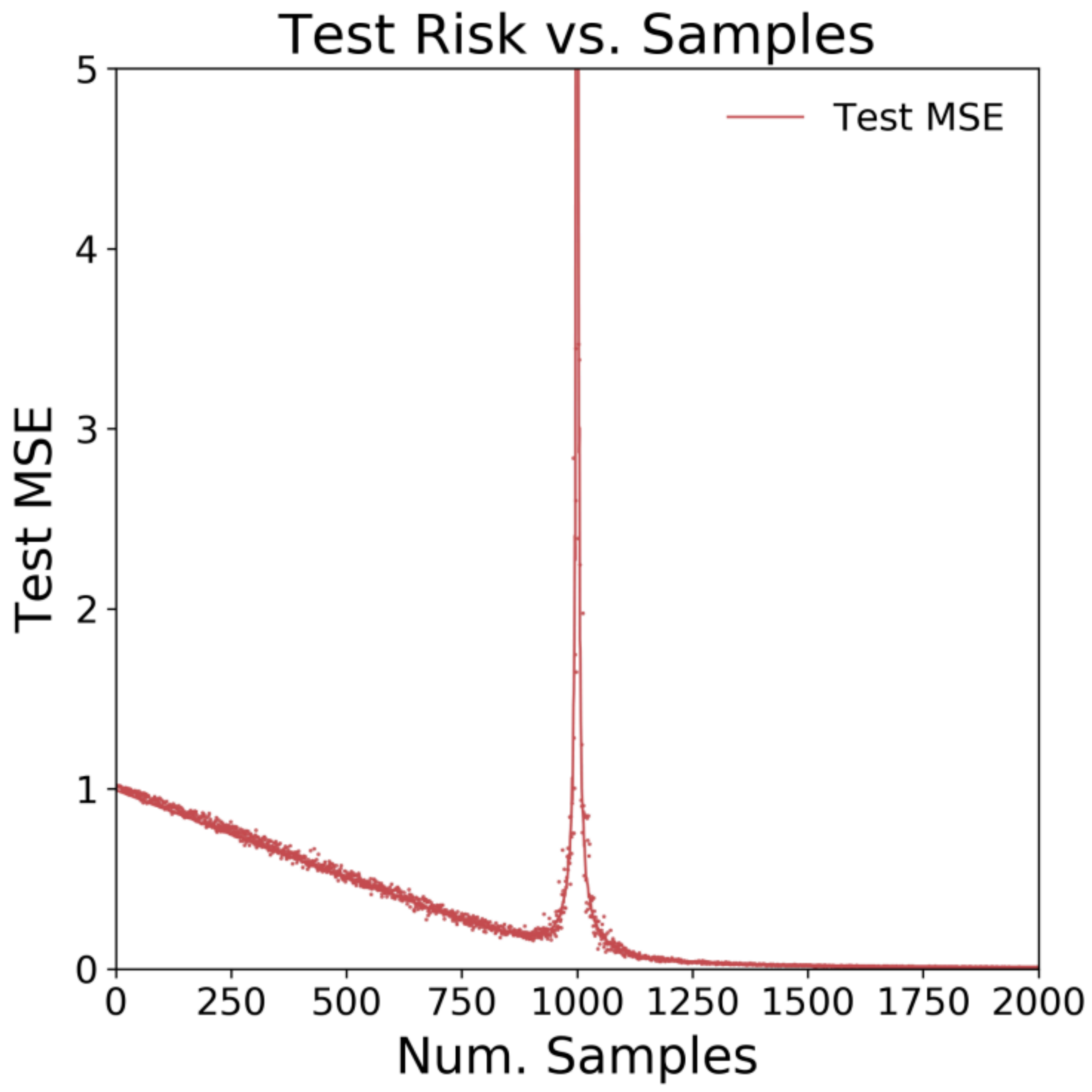
Adlam, Ben, and Jeffrey Pennington. "The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization." In *International Conference on Machine Learning*, 2020.

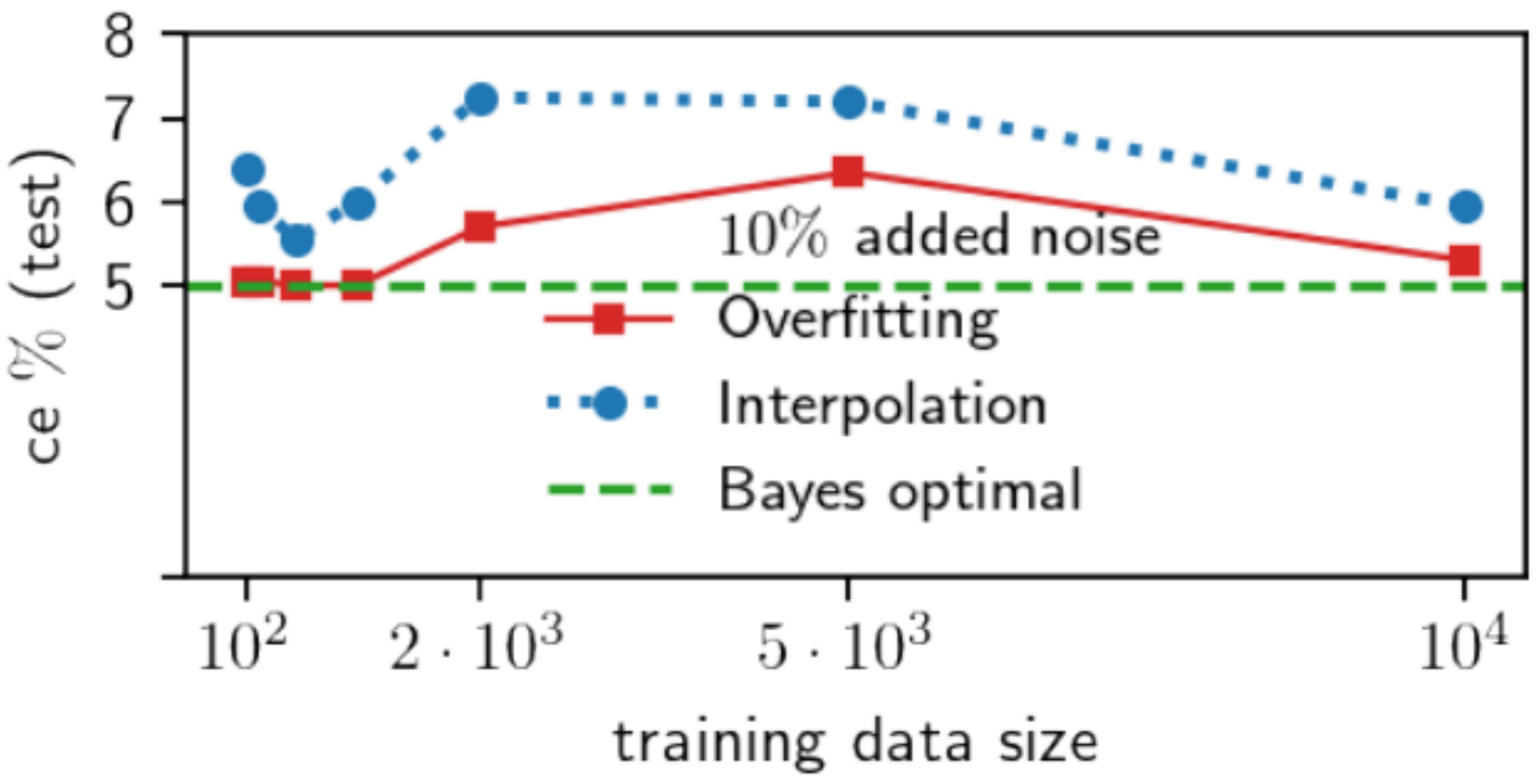
Double/Multiple Descent w.r.t. Sample Size



Nakkiran, Preetum. "More data can hurt for linear regression: Sample-wise double descent." arXiv preprint arXiv:1912.07242 (2019).

Double/Multiple Descent w.r.t. Sample Size





Nakkiran, Preetum. "More data can hurt for linear regression: Sample-wise double descent." arXiv preprint arXiv:1912.07242 (2019).

Belkin, Mikhail, Siyuan Ma, and Soumik Mandal. "To understand deep learning we need to understand kernel learning." International Conference on Machine Learning. PMLR, 2018.

What if we consider more complicated models?

Multi-component prediction models:

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

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

What if we consider more complicated models?

Multi-component prediction models:

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

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

- A class of semi-parametric models

What if we consider more complicated models?

Multi-component prediction models:

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

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

- A class of semi-parametric models
- Ensemble methods

What if we consider more complicated models?

Multi-component prediction models:

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

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

- A class of semi-parametric models
- Ensemble methods
- Certain neural network models such as ResNet

What if we consider more complicated models?

Multi-component prediction models:

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

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

- A class of semi-parametric models
- Ensemble methods
- Certain neural network models such as ResNet

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

More Specifically…

Consider again the simple learning the problem

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

More Specifically…

Consider again the simple learning the problem

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

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.

More Specifically…

Consider again the simple learning the problem

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

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,

More Specifically…

Consider again the simple learning the problem

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

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.

Multiple Descent in Multiple Random Feature Models

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

Constructed prediction model: *"multiple random feature model"*

Multiple Descent in Multiple Random Feature Models

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

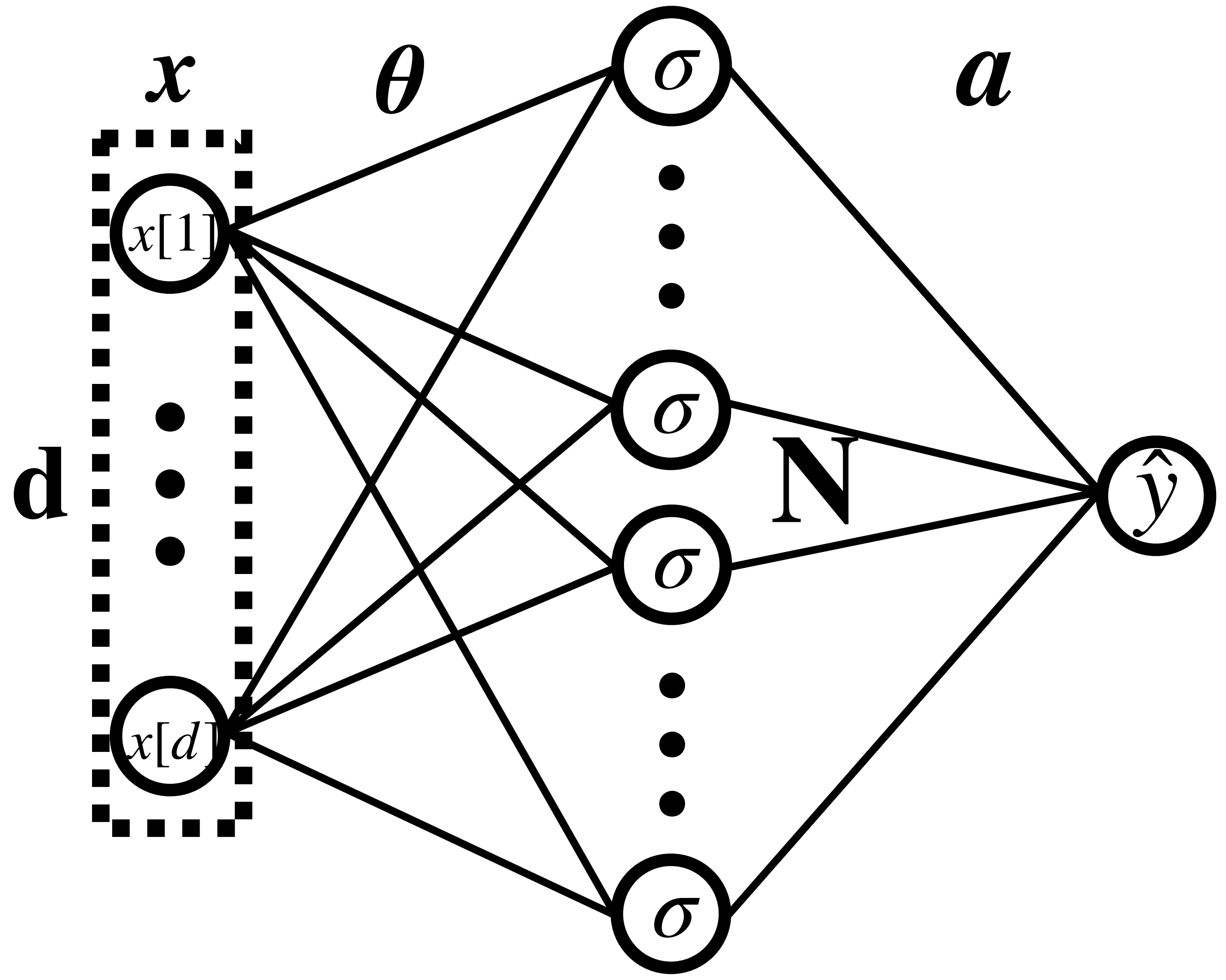
Constructed prediction model: *"multiple random feature model"*

Classic random feature model:

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

Θ: fixed at randomly generated values

a: trainable parameters



Multiple Descent in Multiple Random Feature Models

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

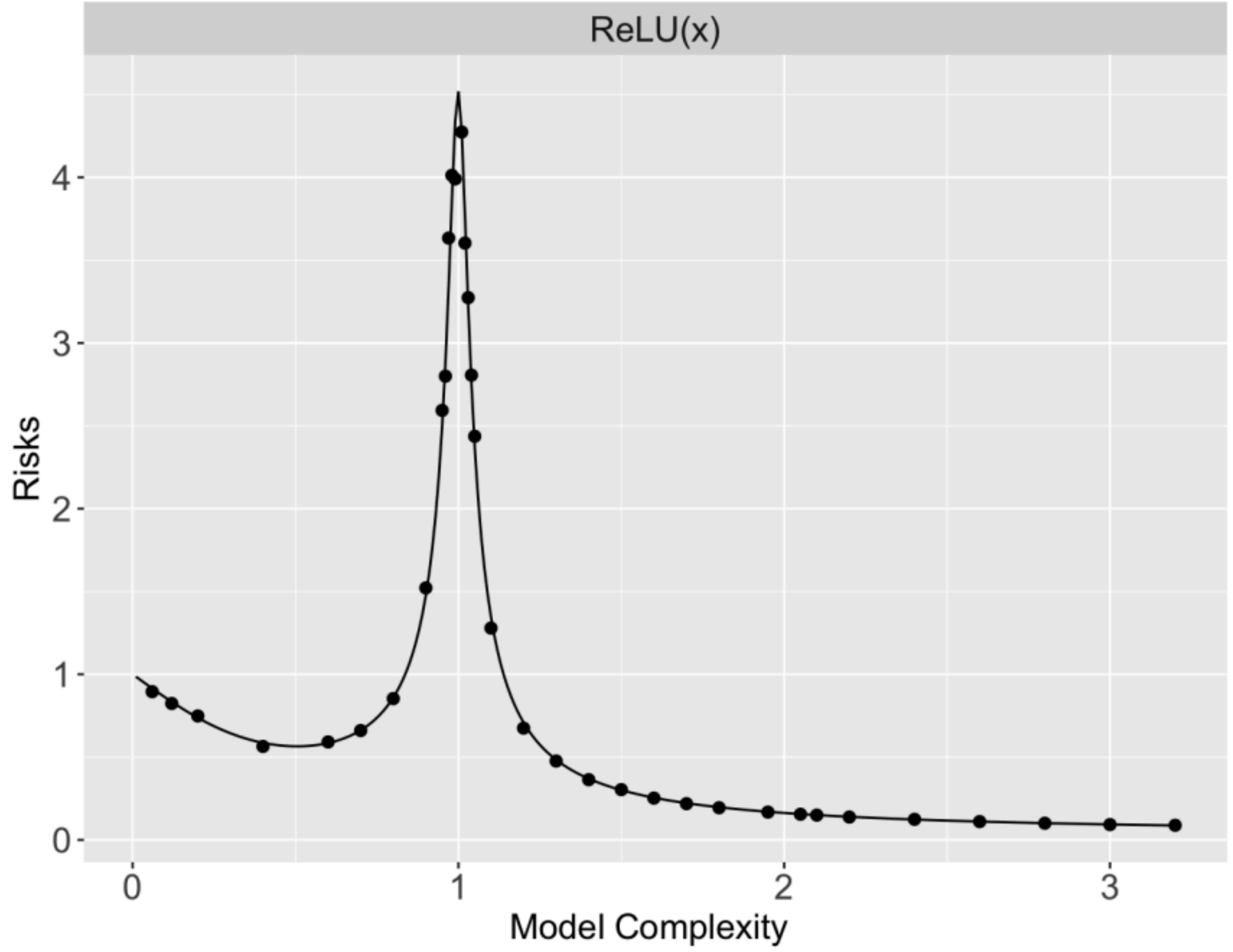
Constructed prediction model: *"multiple random feature model"*

Classic random feature model:

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

$\boldsymbol{\Theta}$: fixed at randomly generated values

a: trainable parameters



[Mei & Montanari, 2022] has demonstrated a double descent risk curve for classic random feature models.

Mei, Song, and Andrea Montanari. "The generalization error of random features regression: Precise asymptotics and the double descent curve." Communications on Pure and Applied Mathematics 75, no. 4 (2022): 667-766.

Multiple Descent in Multiple Random Feature Models

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

Constructed prediction model: *"multiple random feature model"*

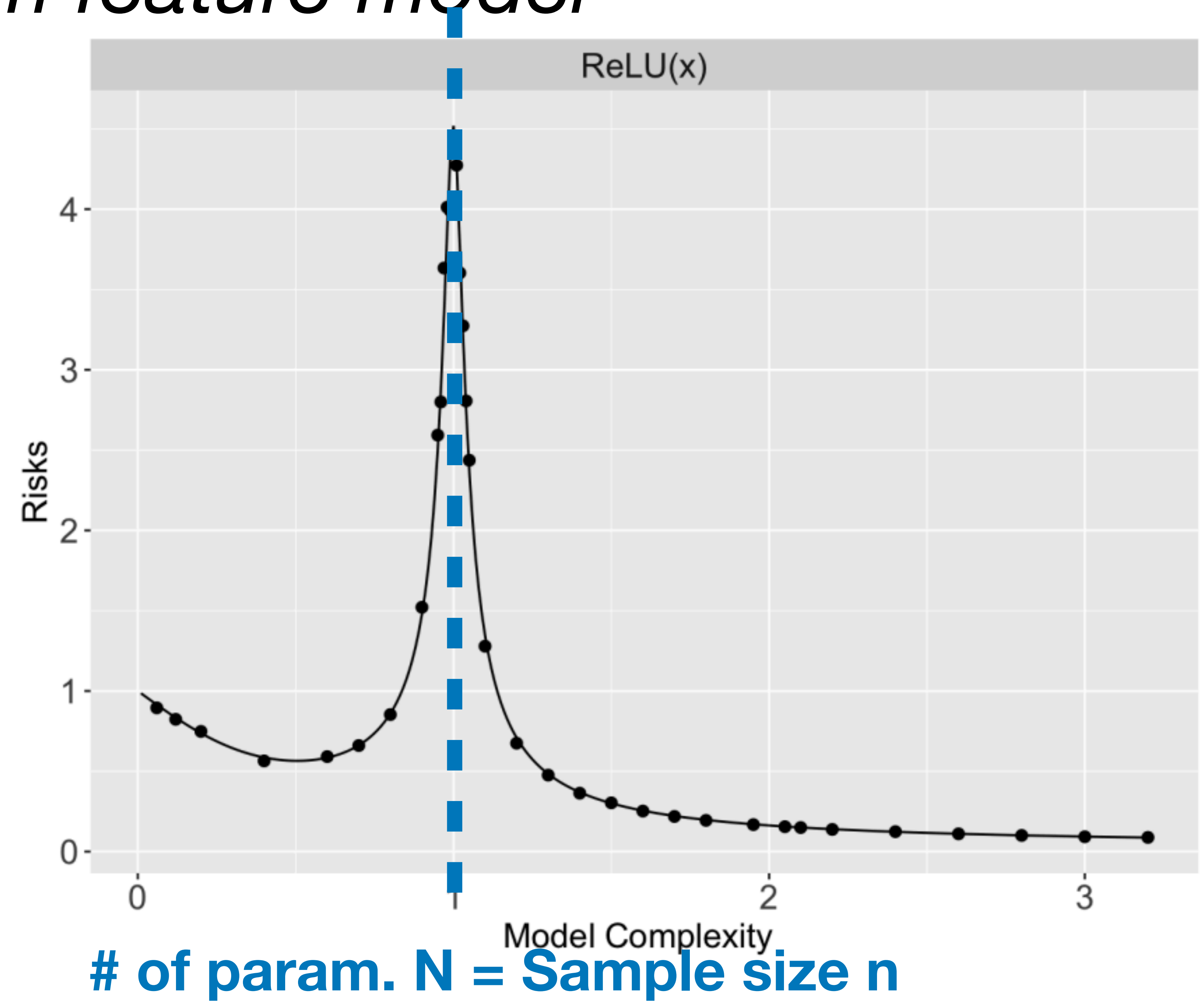
Classic random feature model:

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

$\boldsymbol{\Theta}$: fixed at randomly generated values

a: trainable parameters

[Mei & Montanari, 2022] has demonstrated a double descent risk curve for classic random feature models.



Mei, Song, and Andrea Montanari. "The generalization error of random features regression: Precise asymptotics and the double descent curve." Communications on Pure and Applied Mathematics 75, no. 4 (2022): 667-766.

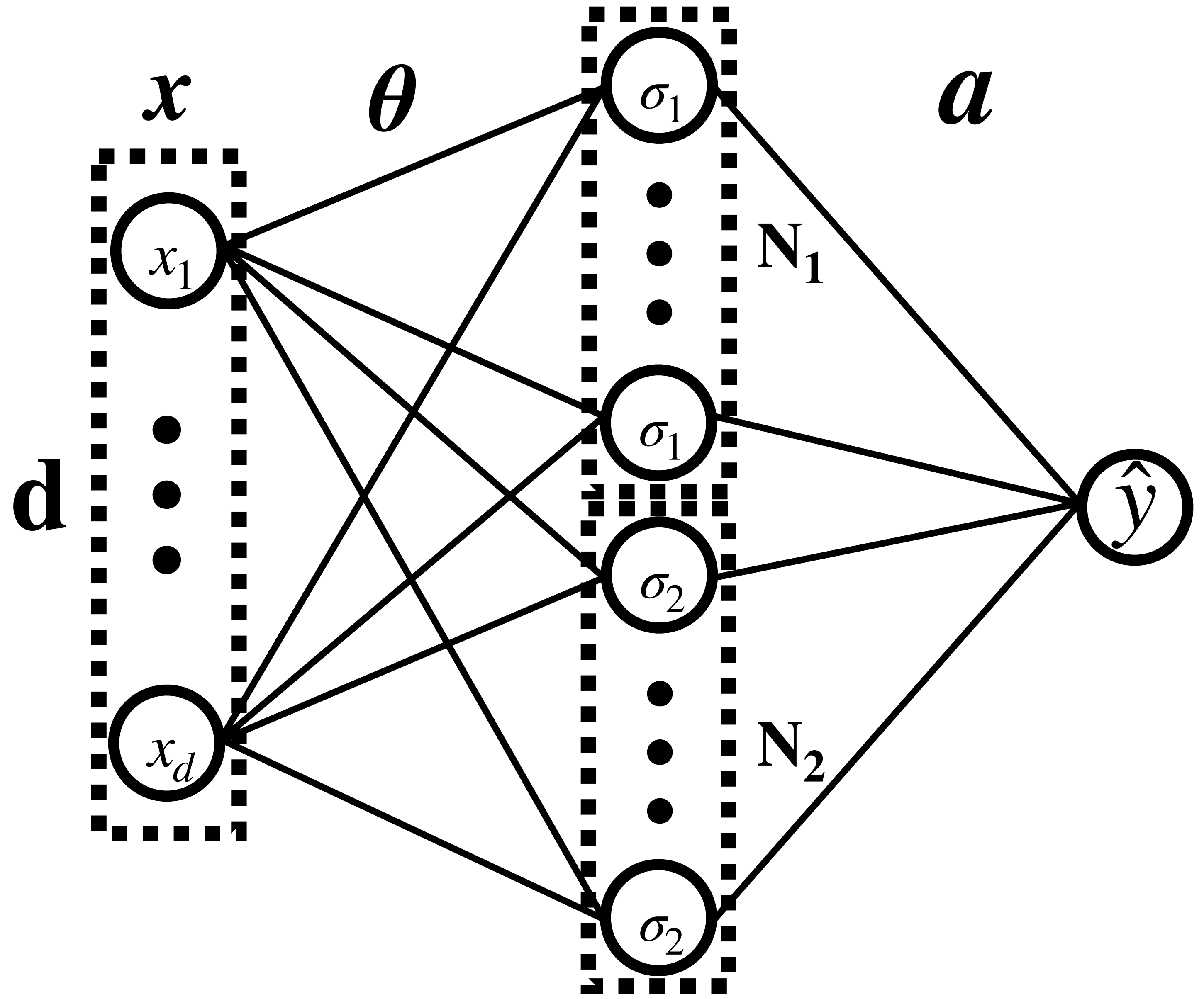
Multiple Descent in Multiple Random Feature Models

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

Constructed prediction model: *"multiple random feature model"*

Double random feature model:

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



Θ: fixed at randomly generated values

a: trainable parameters

Multiple Descent in Multiple Random Feature Models

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

Constructed prediction model: *"multiple random feature model"*

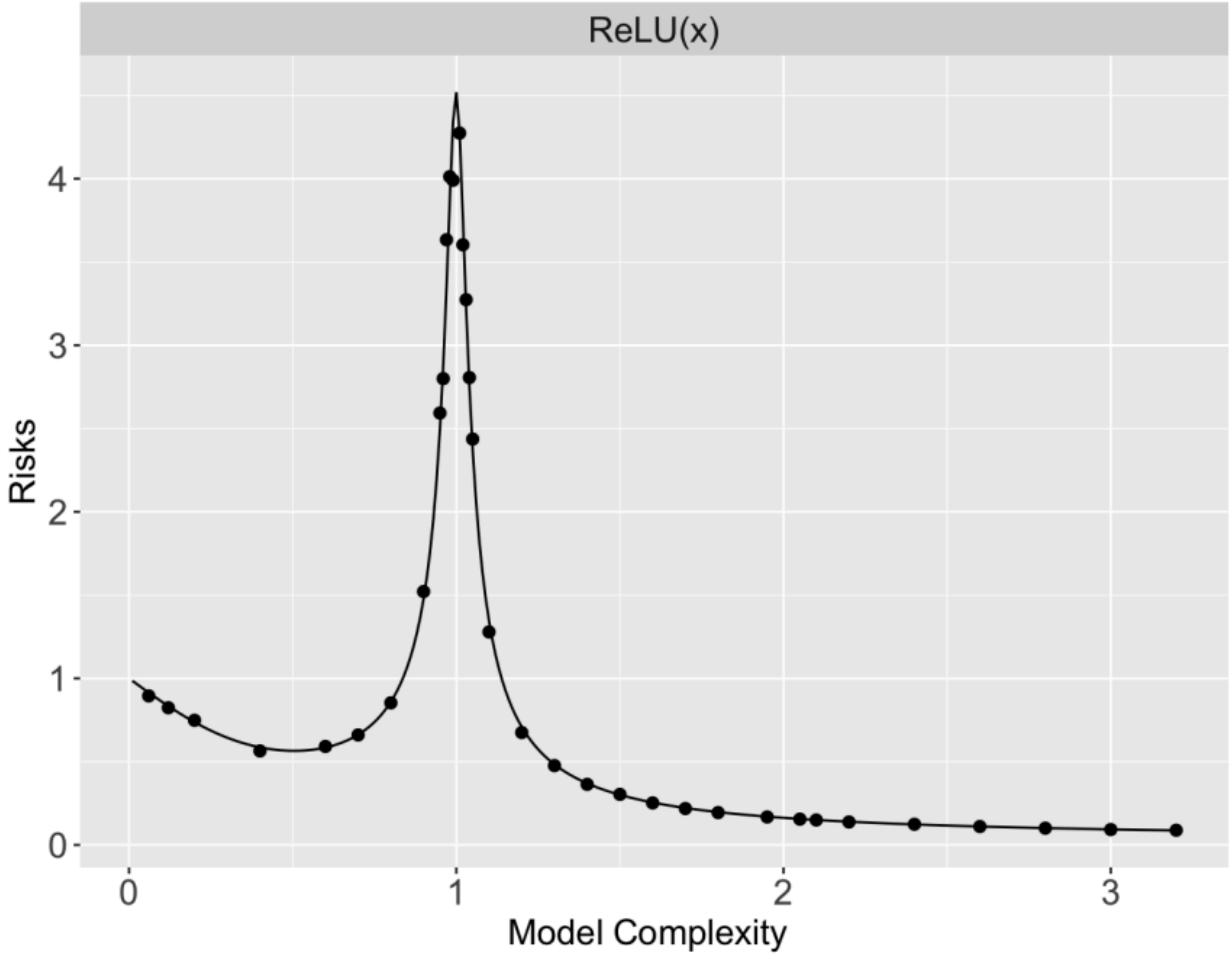
Multiple random feature model:

$$\mathscr{F}_{\mathrm{MRF}}(\Theta) = \left\{ f(\mathbf{x}; \mathbf{a}, \boldsymbol{\Theta}) \equiv \sum_{j=1}^{K} \sum_{i \in \mathcal{N}_j} a_i \sigma_j\left(\langle \boldsymbol{\theta}_i, \mathbf{x} \rangle / \sqrt{d}\right) : a_i \in \mathbb{R}, i \in [N] \right\}$$

$\boldsymbol{\Theta}$: fixed at randomly generated values

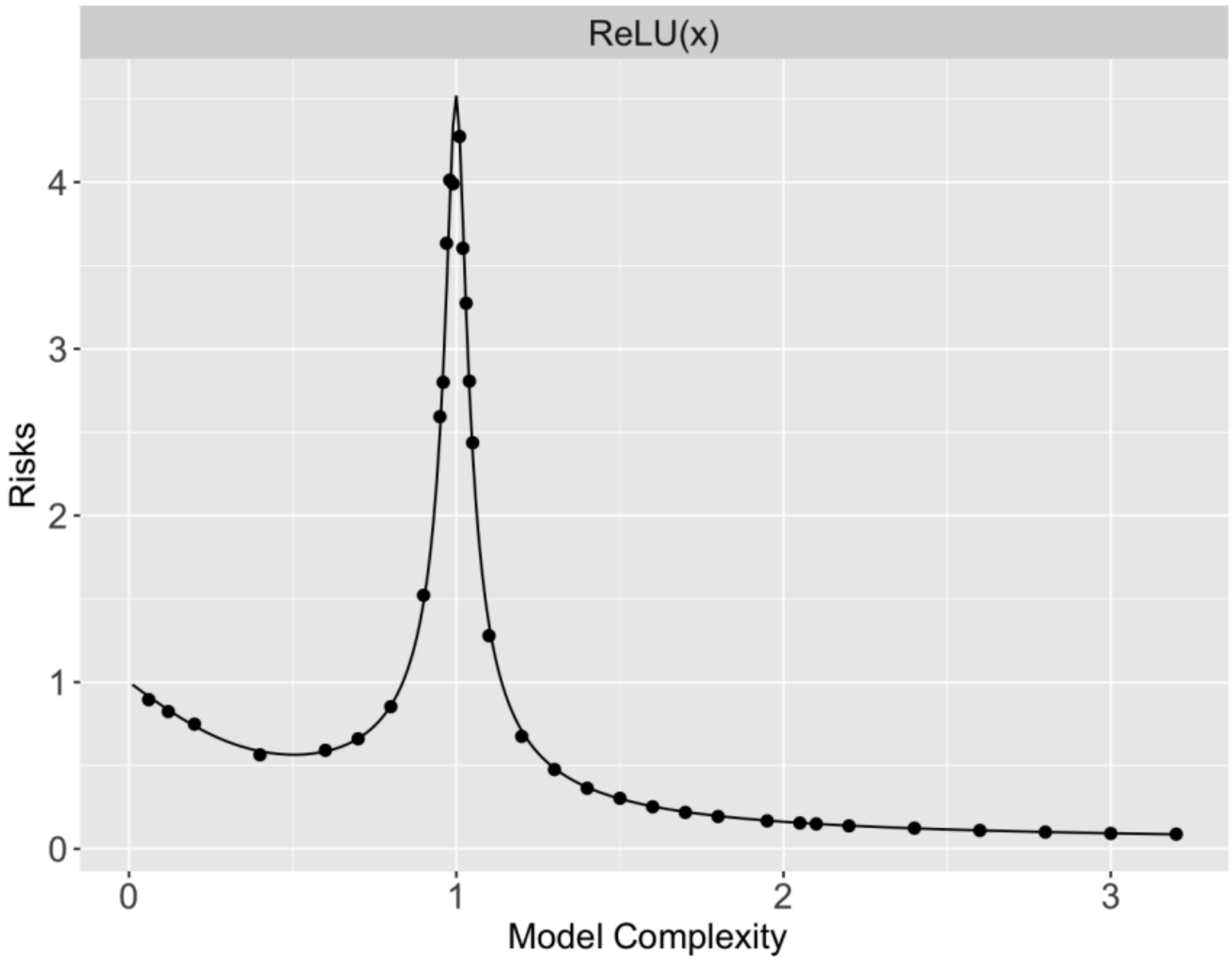
a: trainable parameters

From Double Descent to Multiple Descent

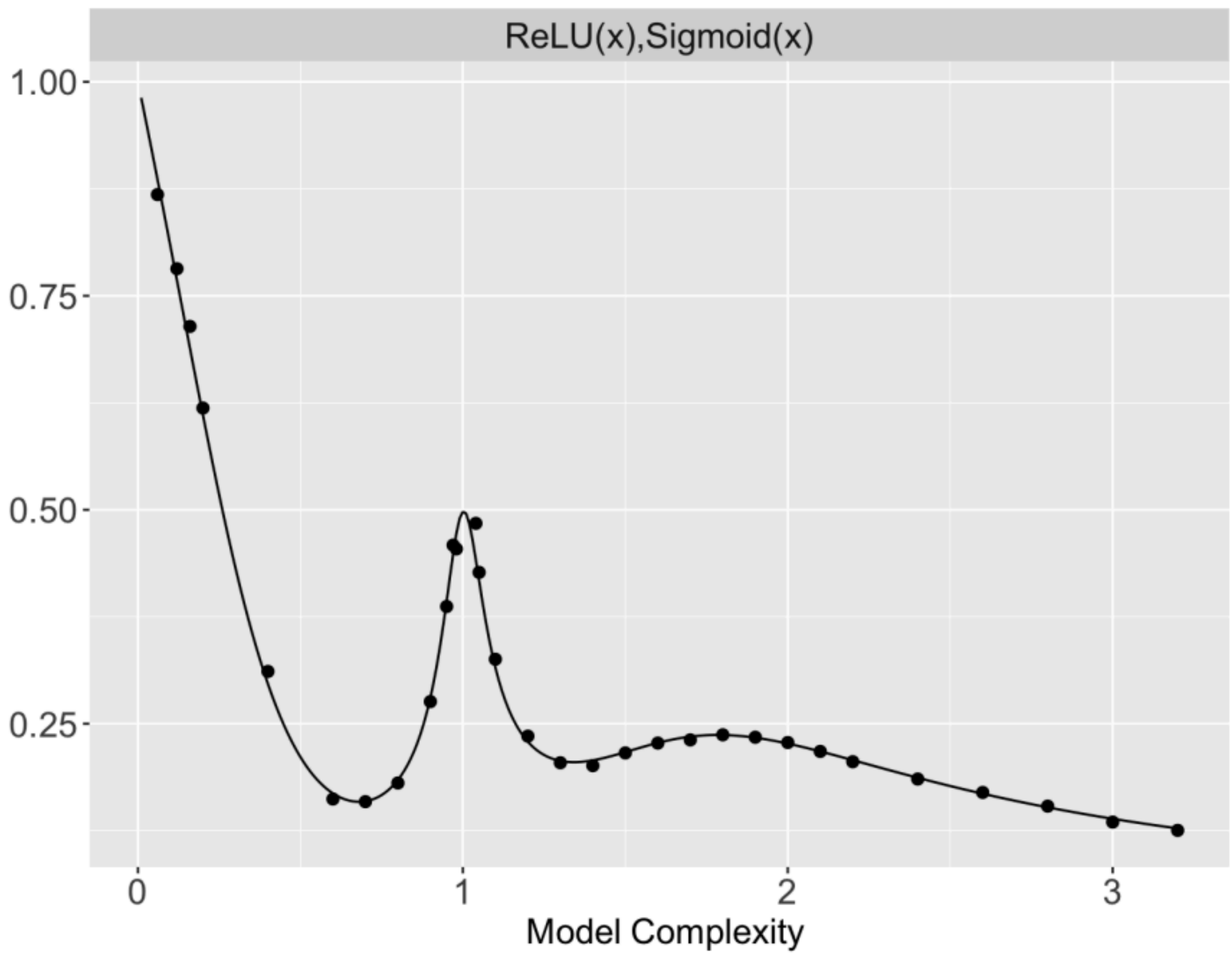


double descent

From Double Descent to Multiple Descent

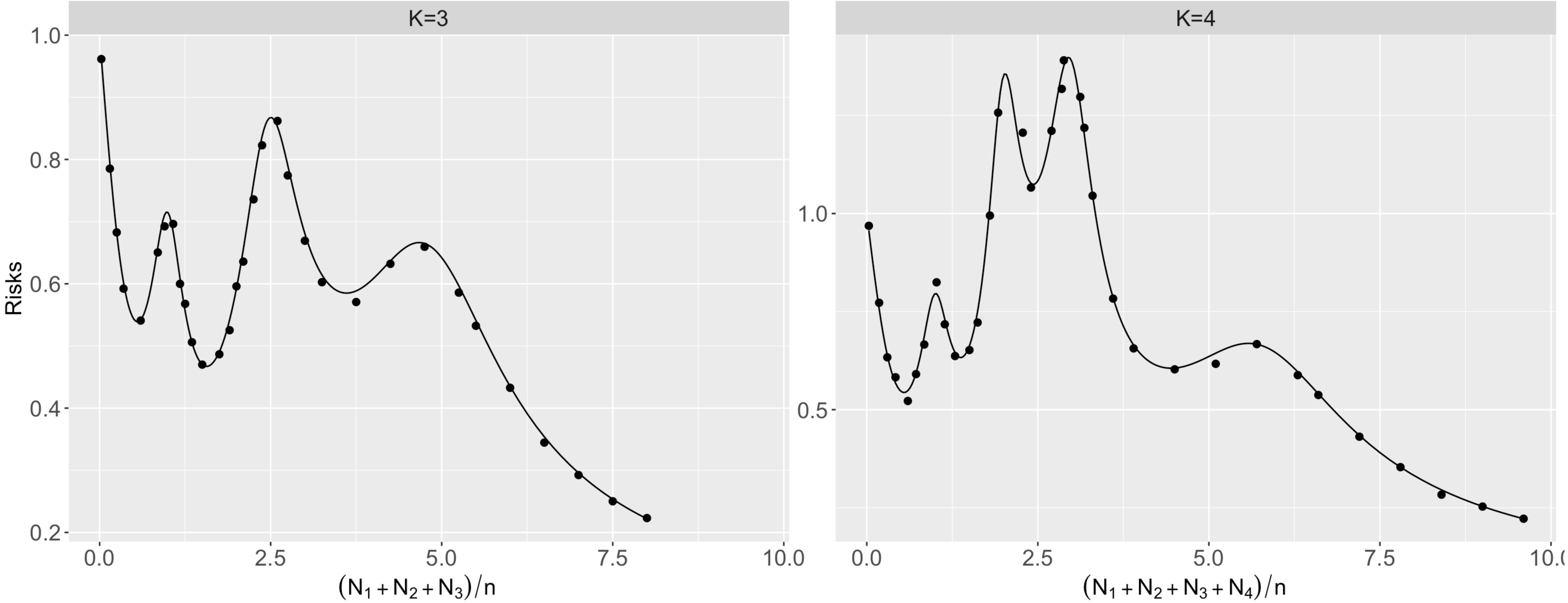


double descent



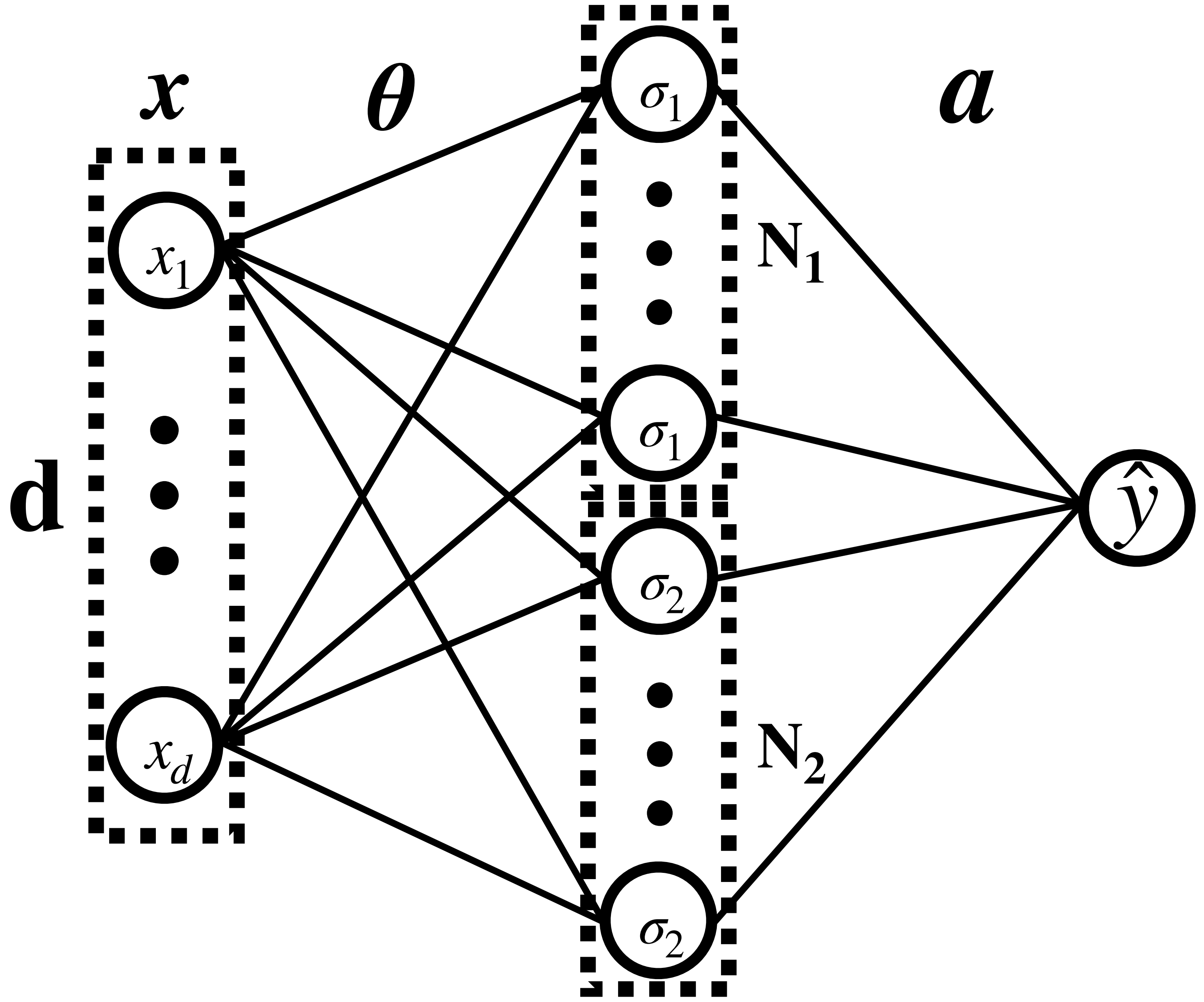
triple descent

From Double Descent to Multiple Descent



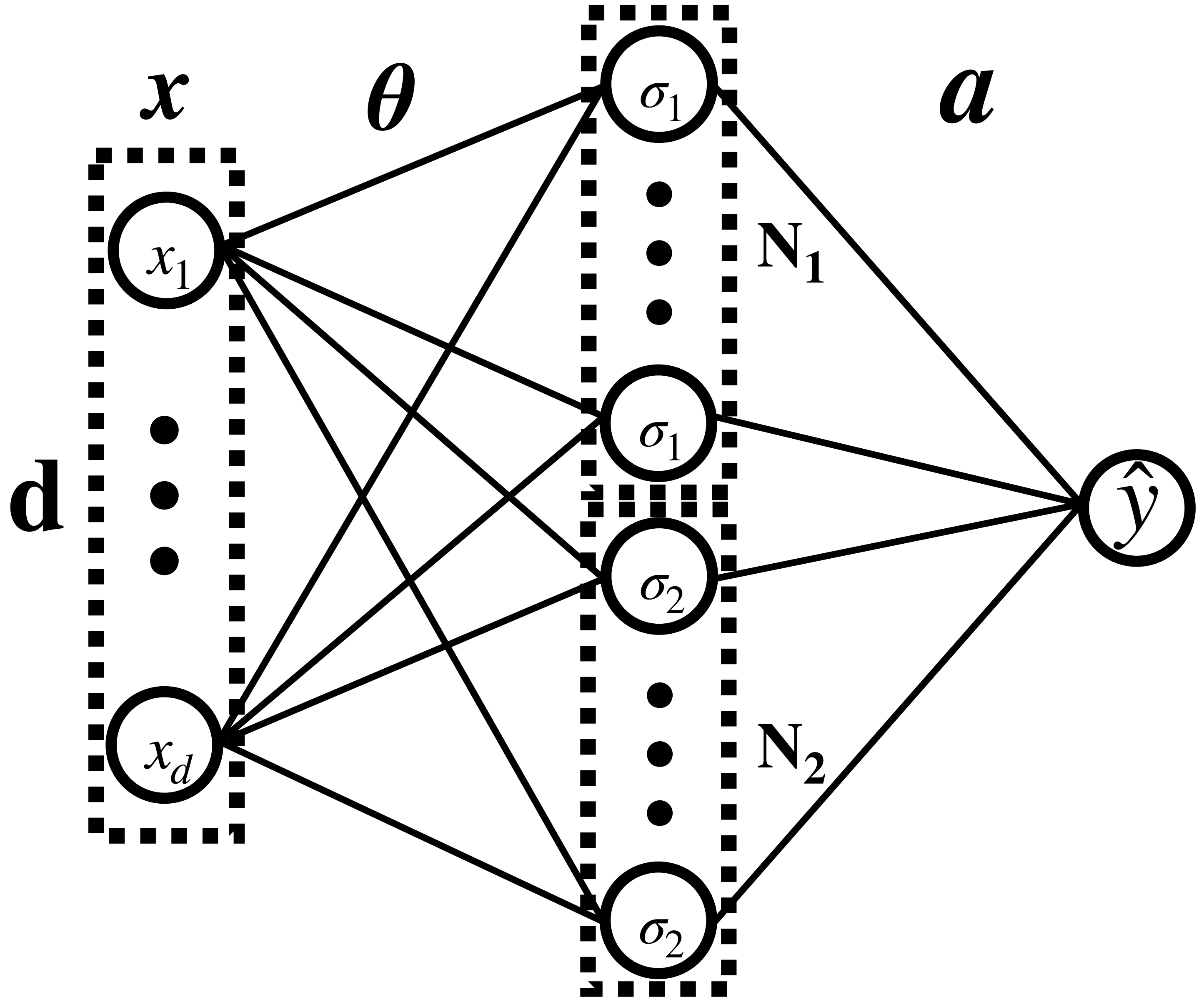
quadruple descent **quintuple descent**

Intuition of Multiple Descent in Multi-Component Models



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

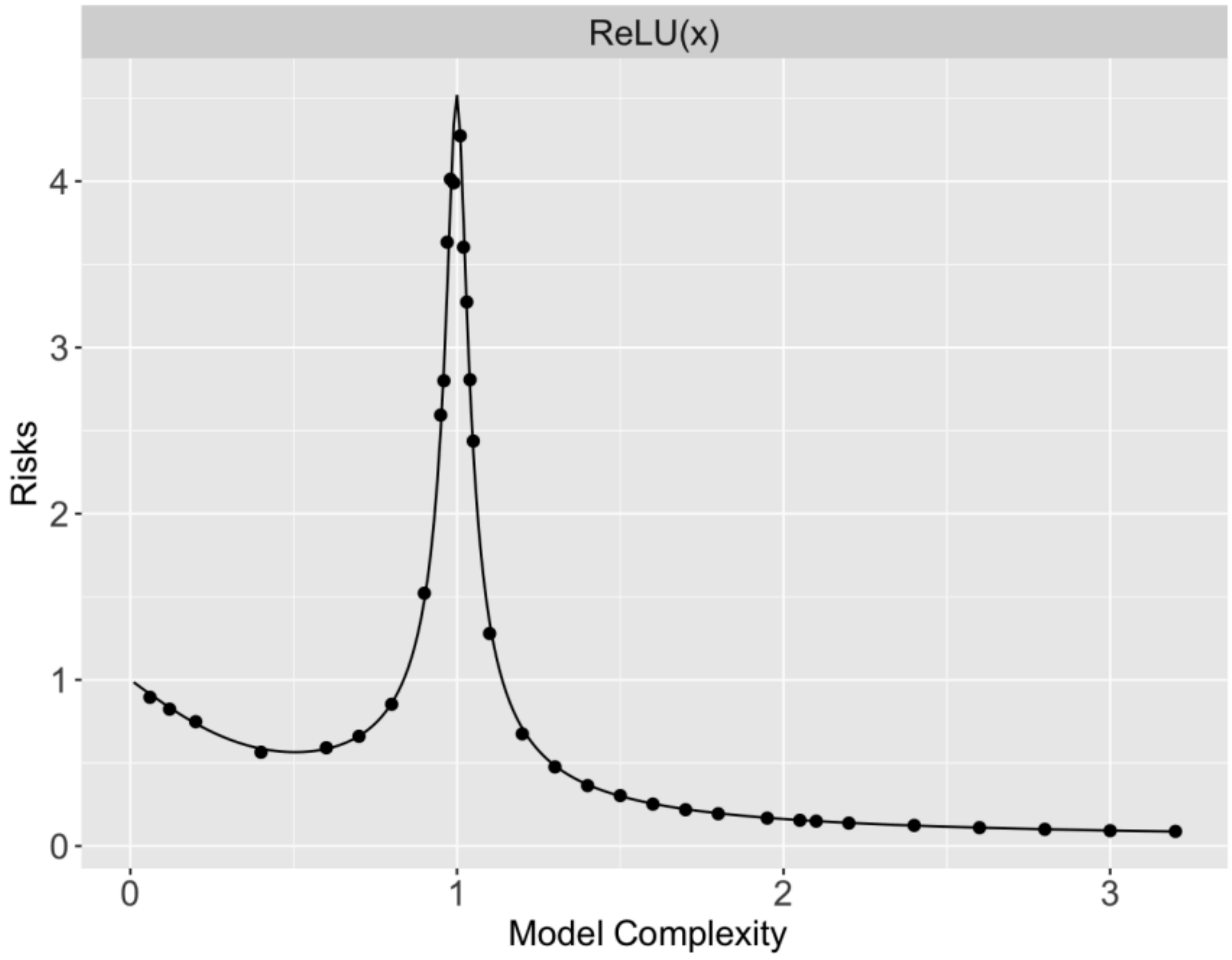
Intuition of Multiple Descent in Multi-Component Models

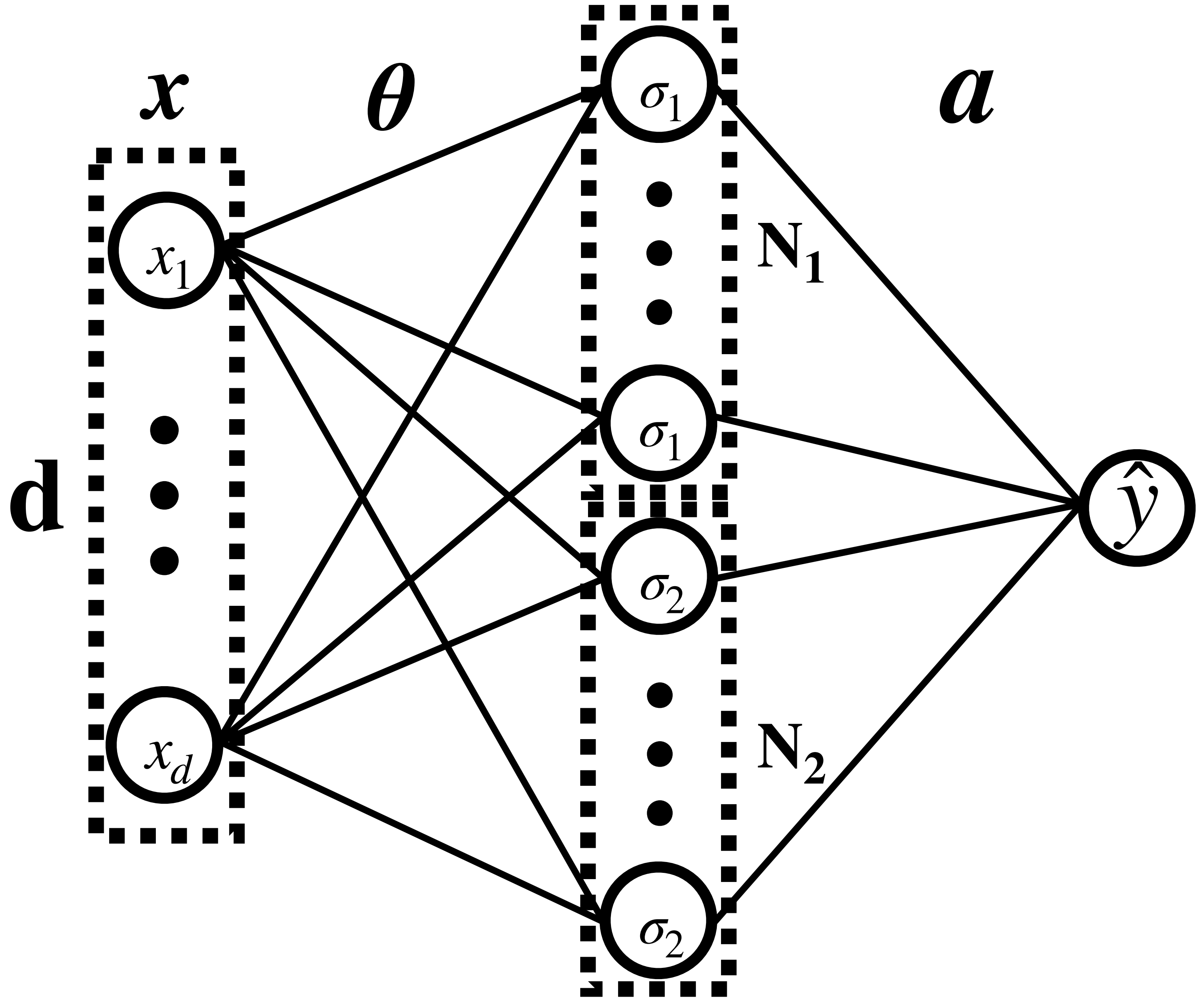


Scale difference may be the key (consider the case $N_1 = N_2$):

- If $\sigma_1(), \sigma_2()$ are the same, we may expect double descent according to existing studies [Mei & Montanari, 2022], and the peak is at $(N_1 + N_2)/n = 1$.

Intuition of Multiple Descent in Multi-Component Models

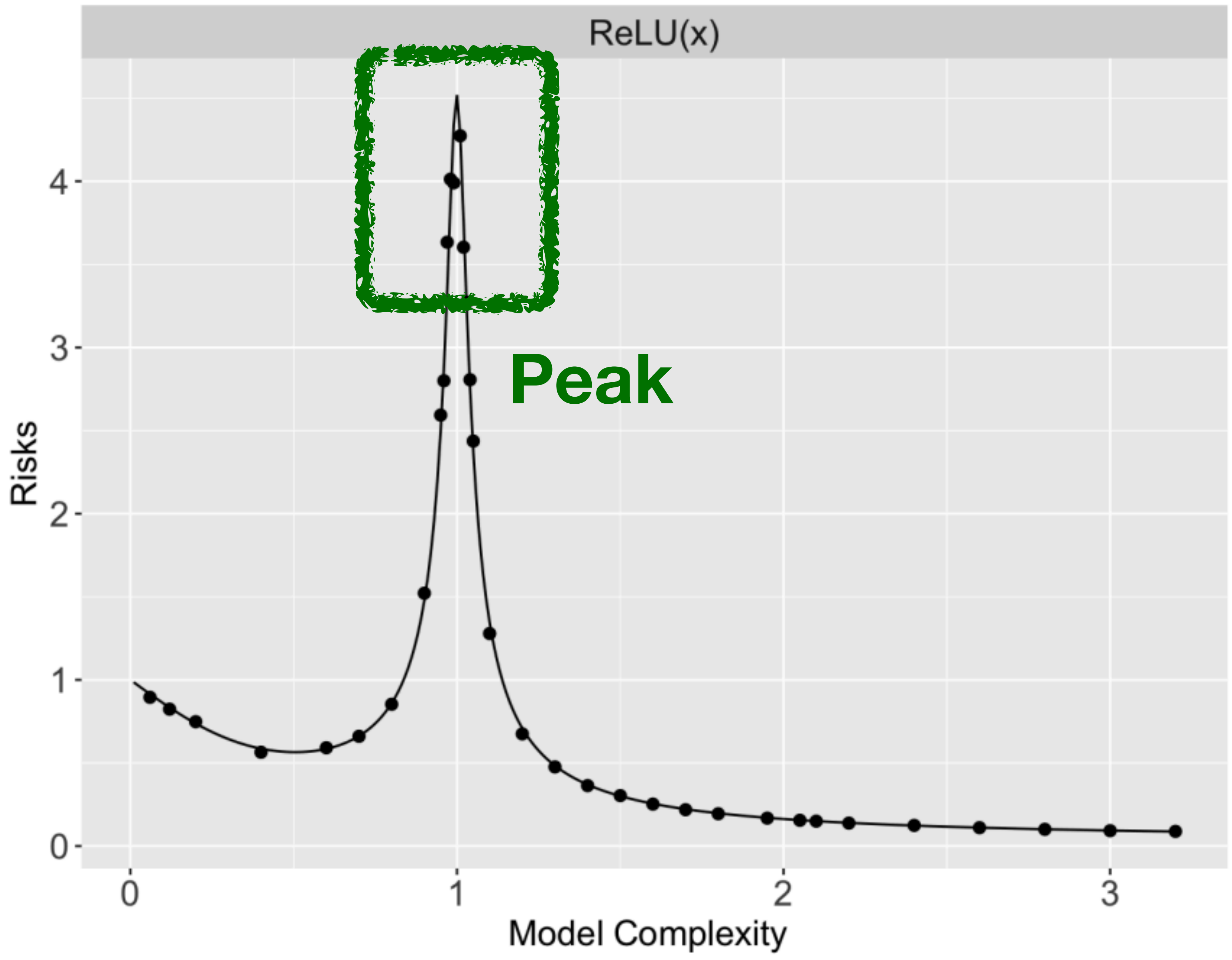


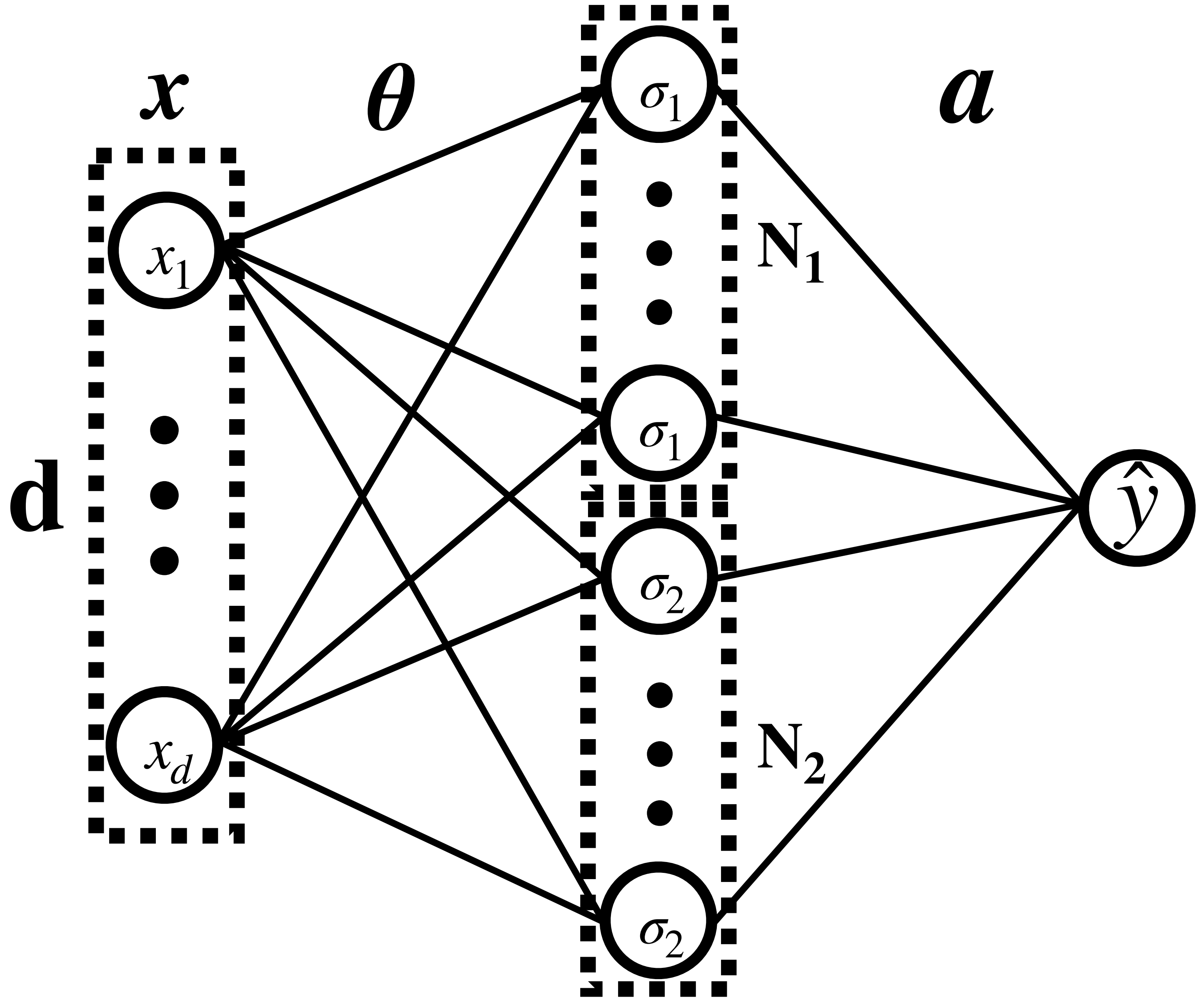


Scale difference may be the key (consider the case $N_1 = N_2$):

- If $\sigma_1(), \sigma_2()$ are the same, we may expect double descent according to existing studies [Mei & Montanari, 2022], and the peak is at $(N_1 + N_2)/n = 1$.

Intuition of Multiple Descent in Multi-Component Models

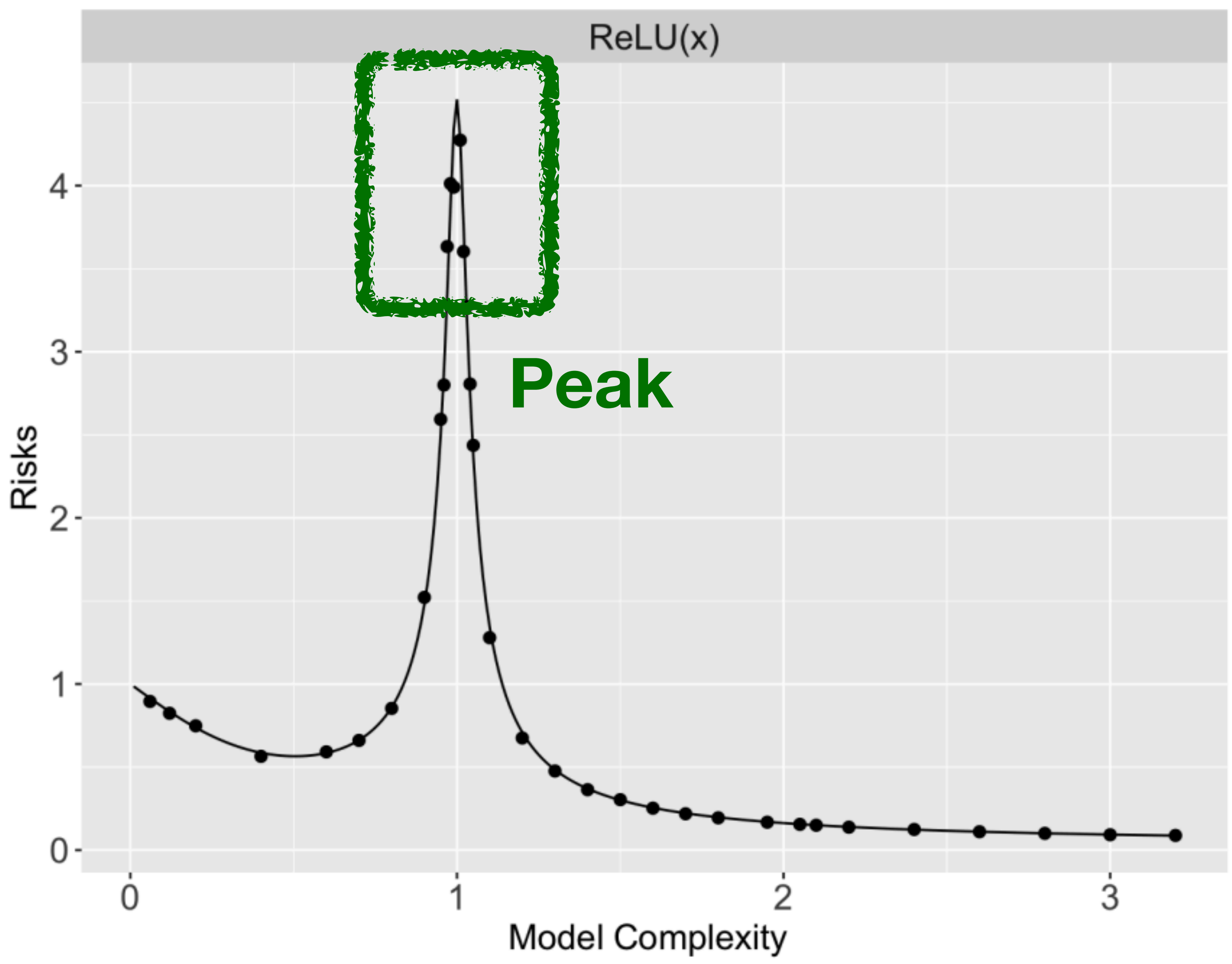


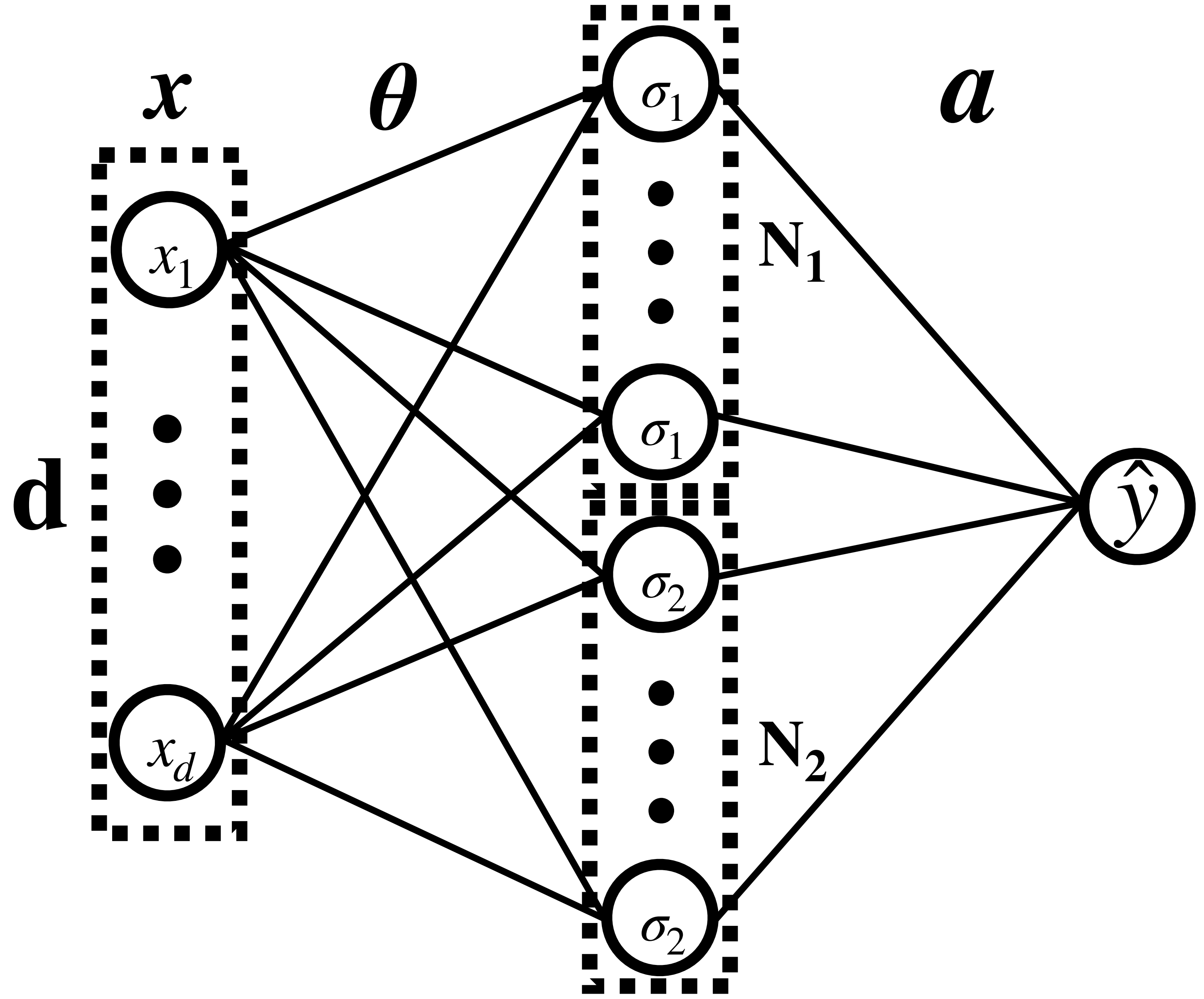


Scale difference may be the key (consider the case $N_1 = N_2$):

- If $\sigma_1(), \sigma_2()$ are the same, we may expect double descent according to existing studies [Mei & Montanari, 2022], and the peak is at $(N_1 + N_2)/n = 1$.

Intuition of Multiple Descent in Multi-Component Models

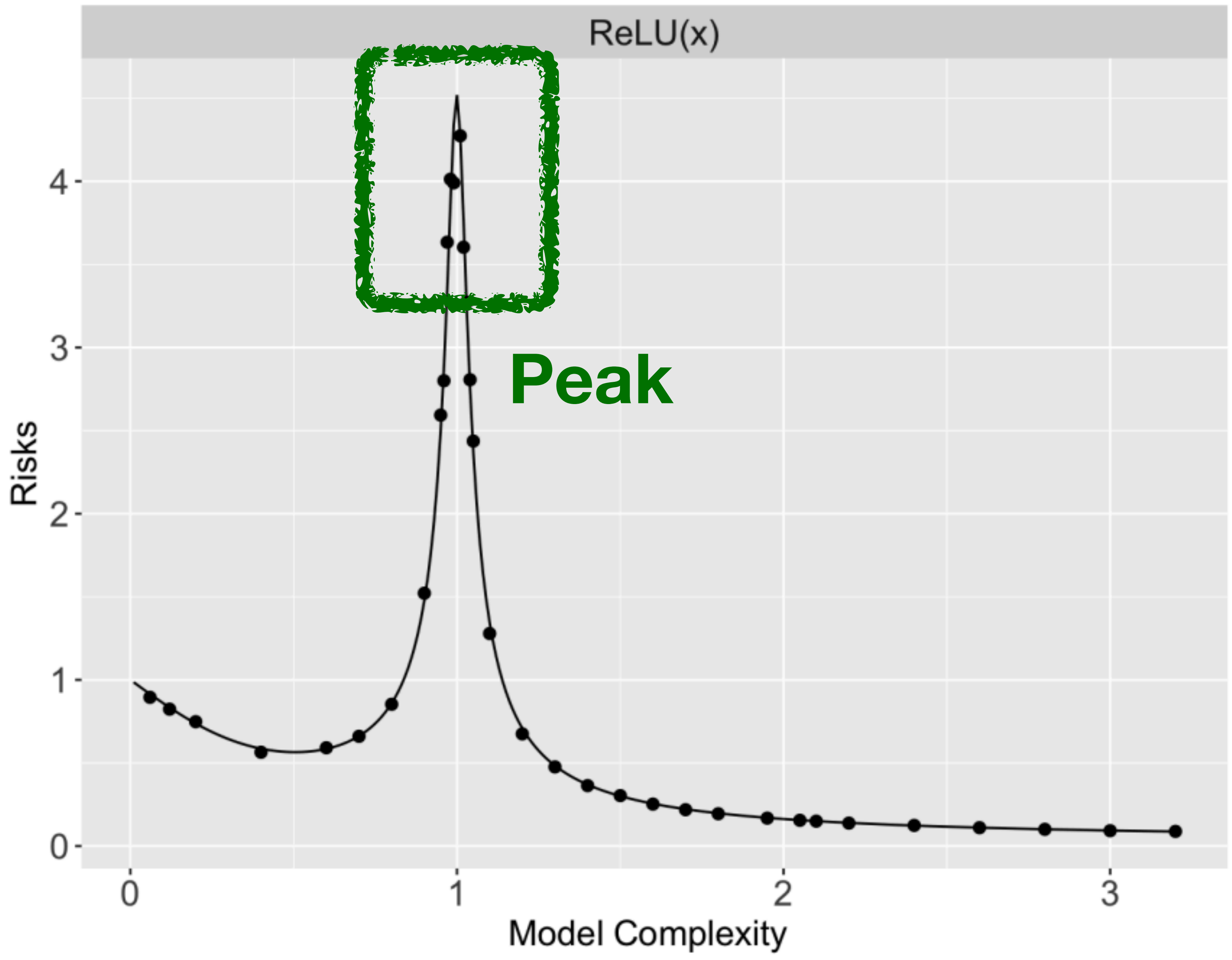


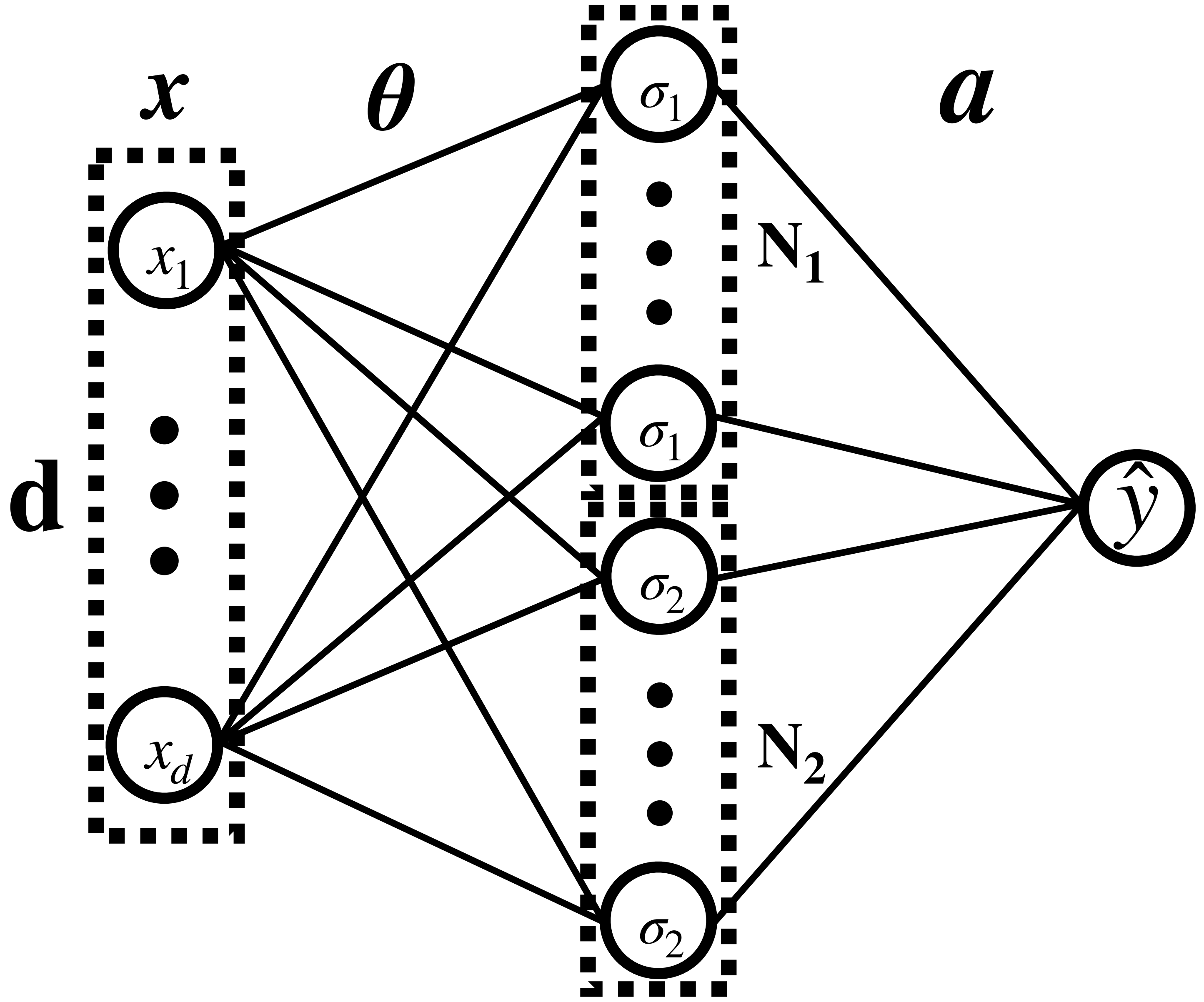


Scale difference may be the key (consider the case $N_1 = N_2$):

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

Intuition of Multiple Descent in Multi-Component Models

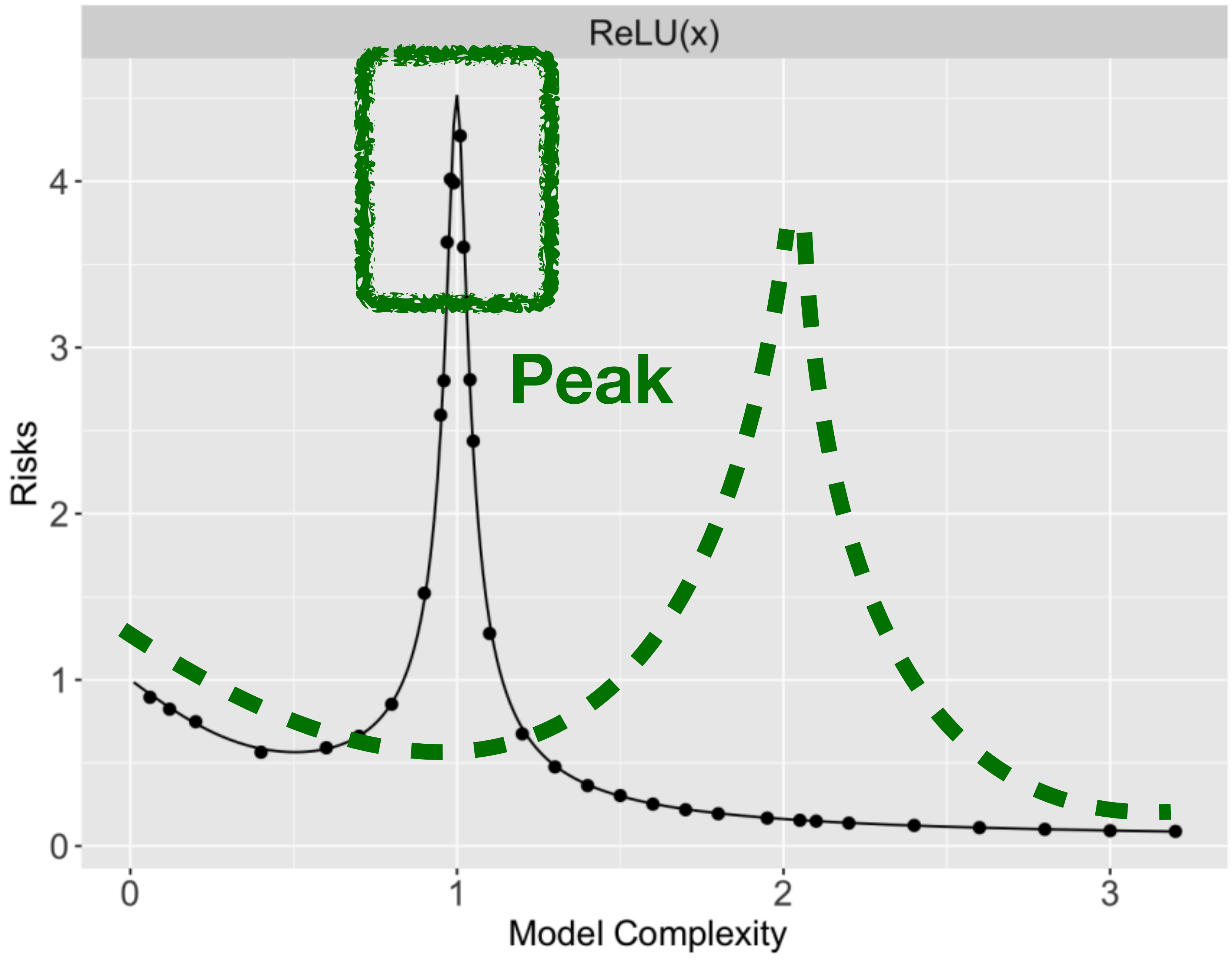


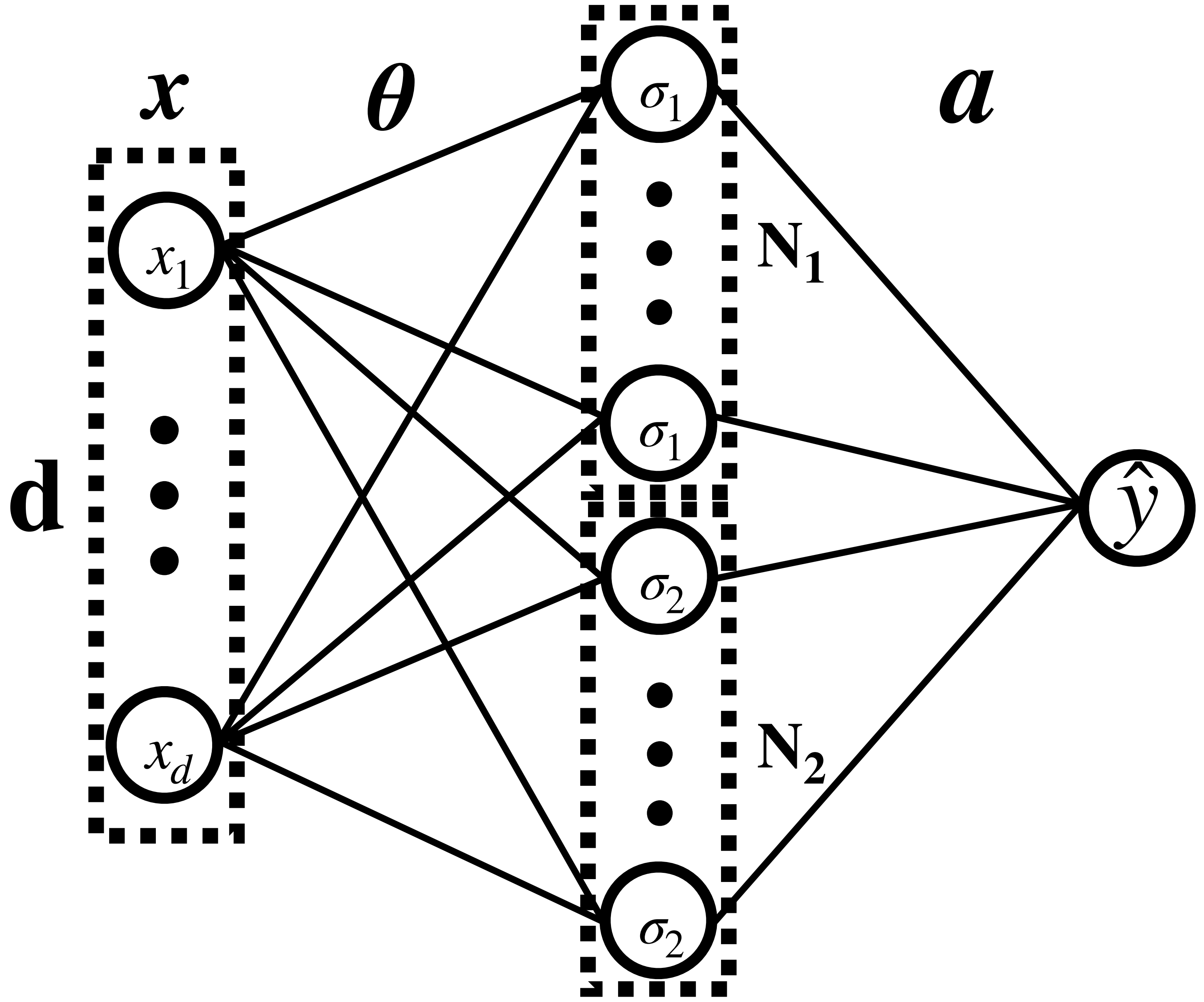


Scale difference may be the key (consider the case $N_1 = N_2$):

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

Intuition of Multiple Descent in Multi-Component Models

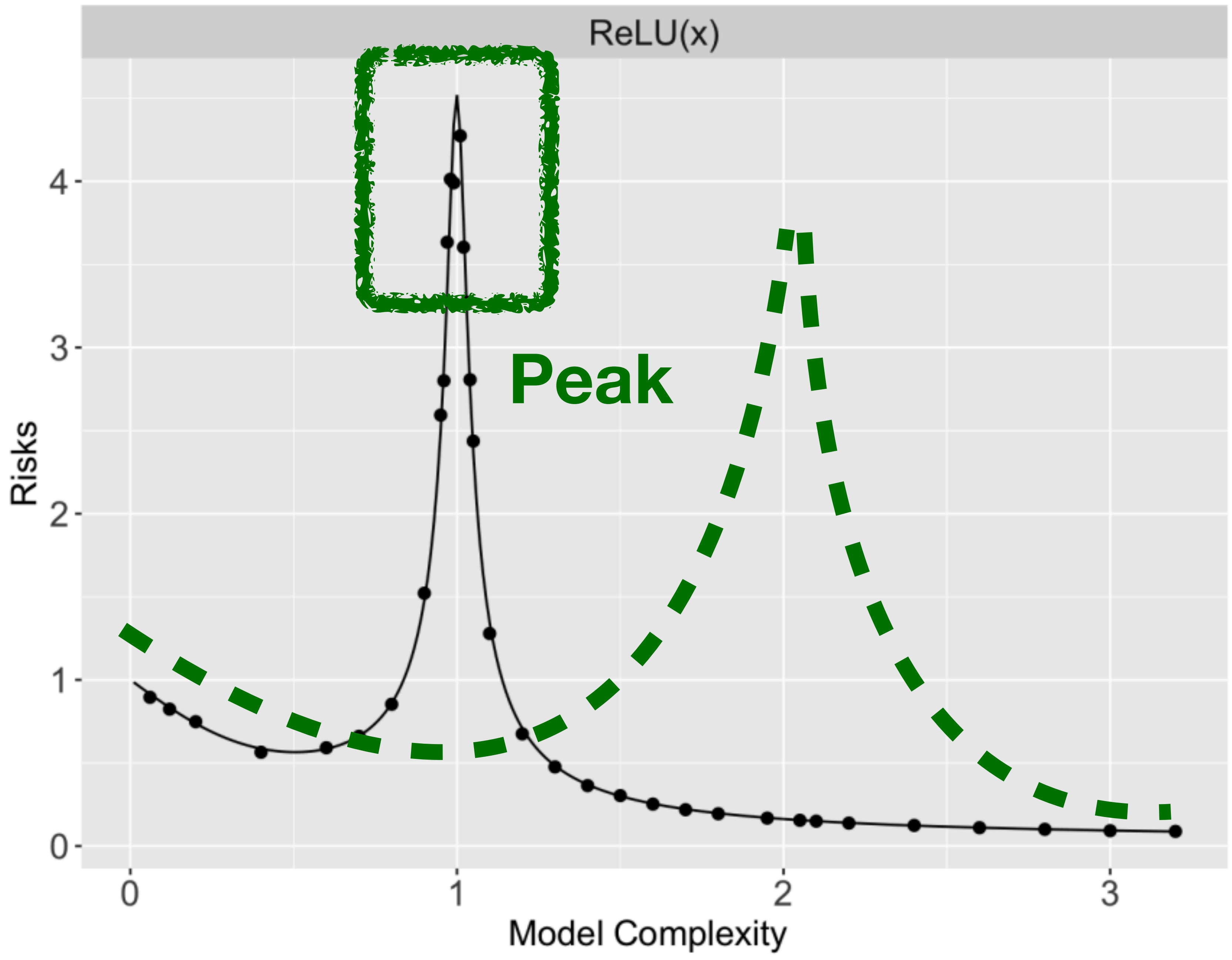


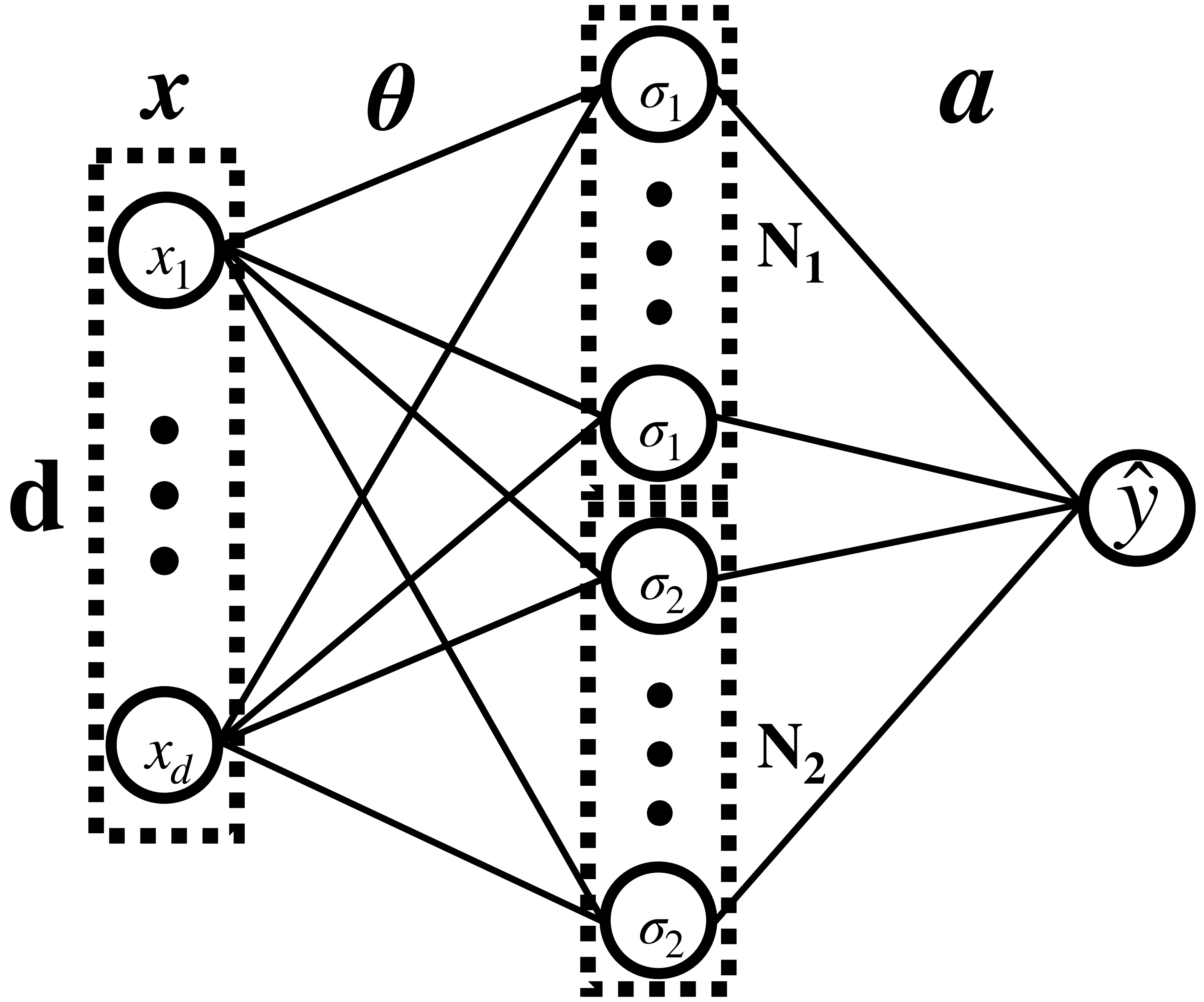


Scale difference may be the key (consider the case $N_1 = N_2$):

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

Intuition of Multiple Descent in Multi-Component Models



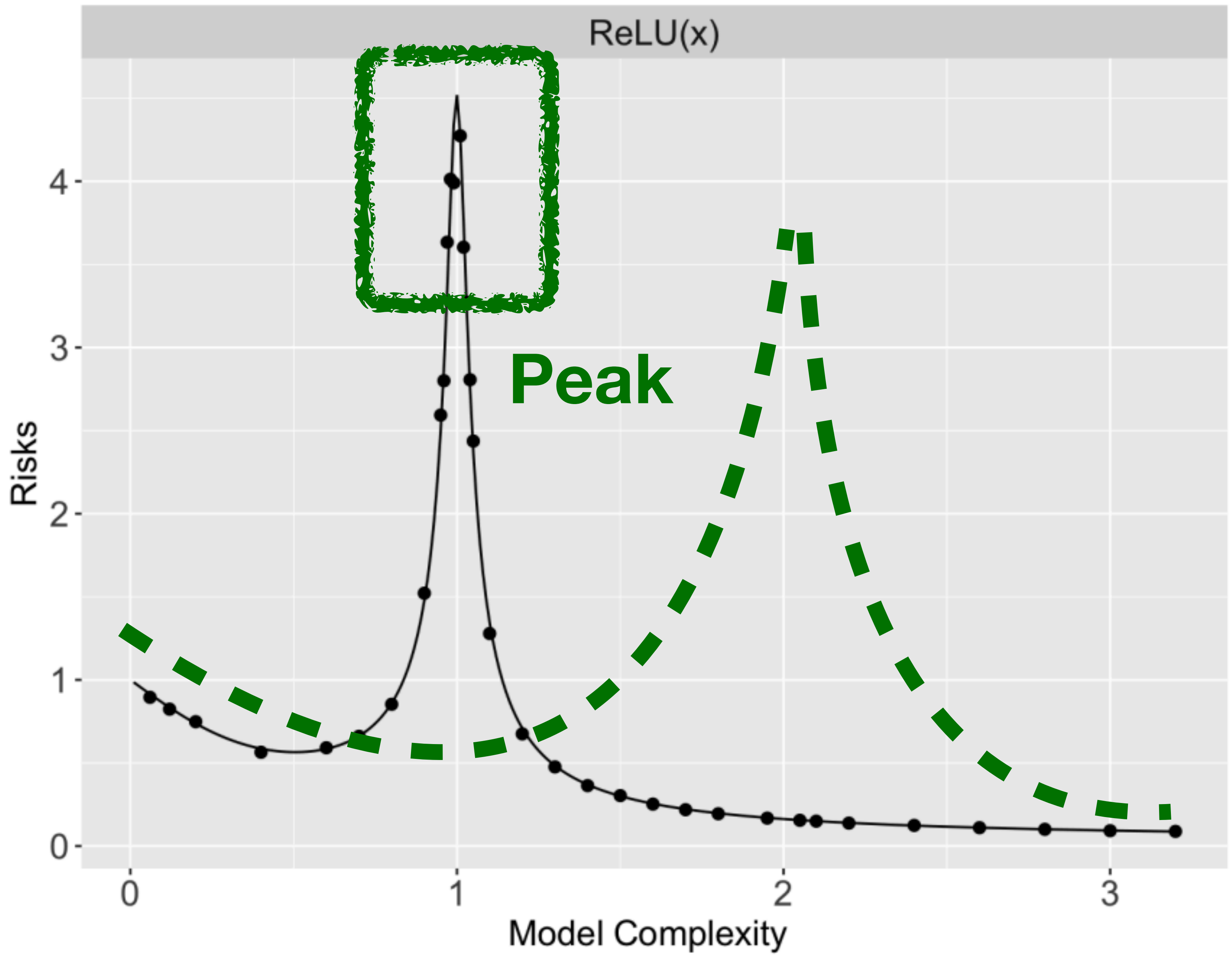


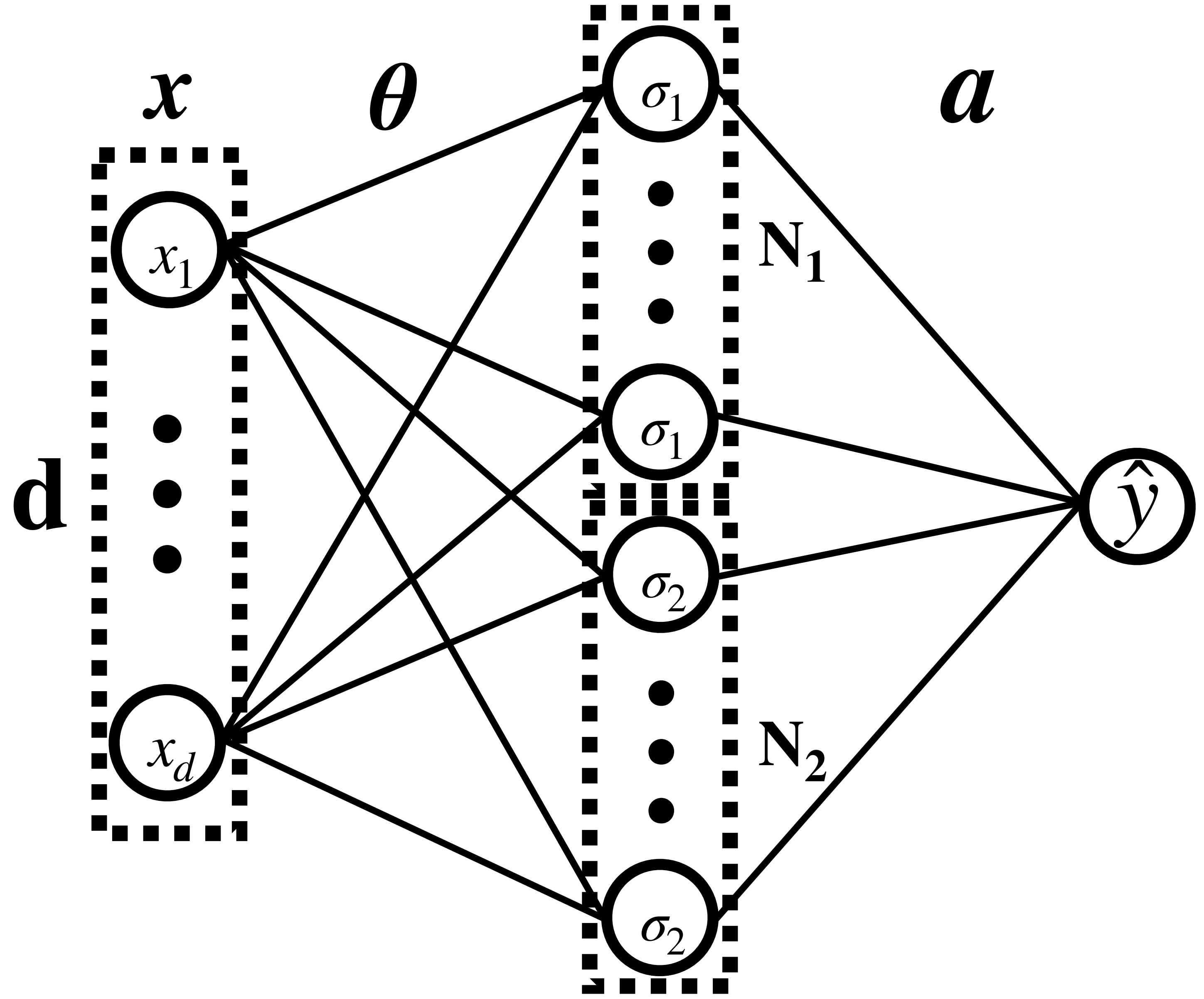
Scale difference may be the key (consider the case $N_1 = N_2$):

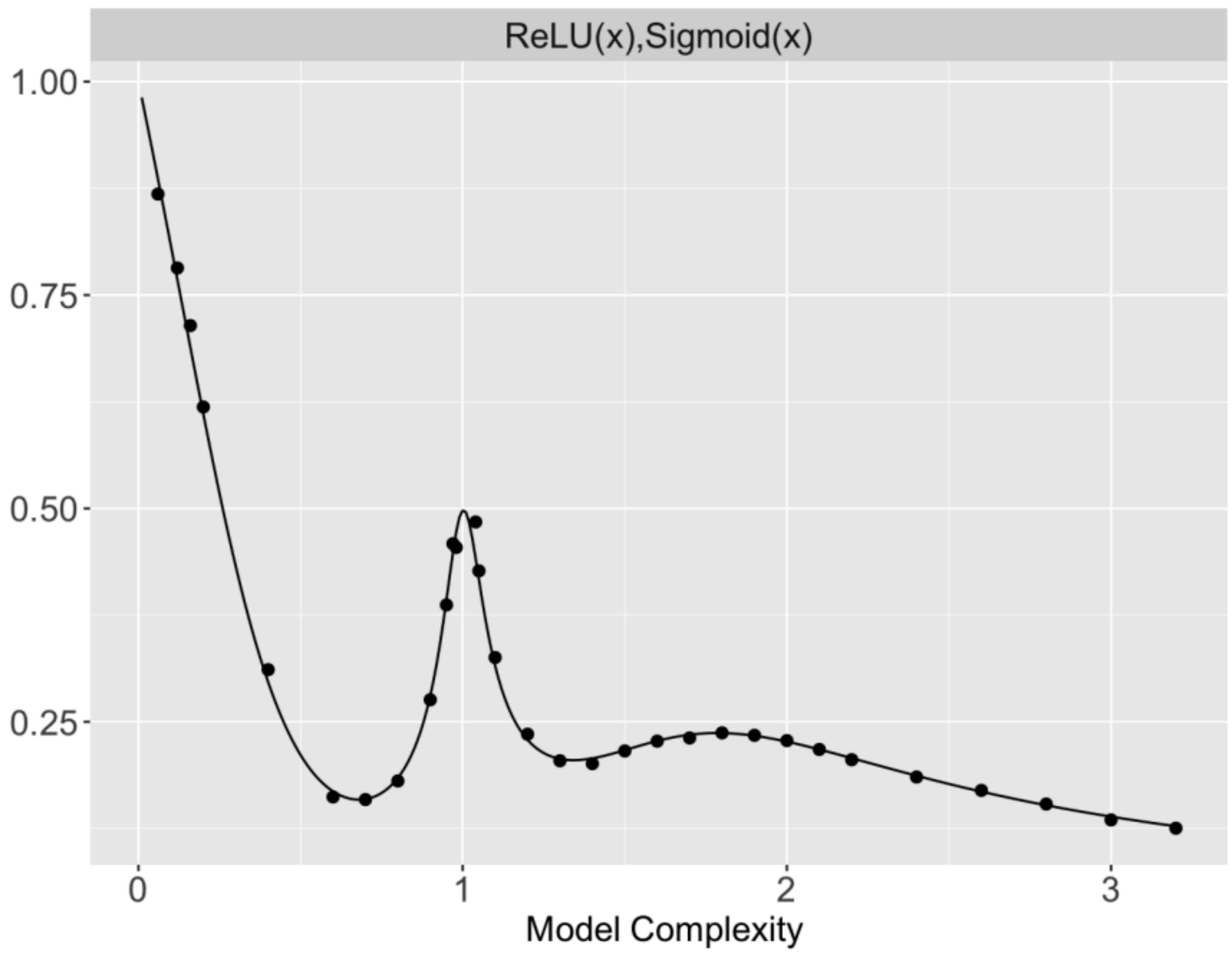
- If $\sigma_1(), \sigma_2()$ are the same, we may expect double descent according to existing studies [Mei & Montanari, 2022], and the peak is at $(N_1 + N_2)/n = 1$.
- If $\sigma_2()$ is very small compared with $\sigma_1()$, we may also expect double descent according to existing studies [Mei & Montanari,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

The above are two extreme cases, each showing double descent with different peak locations. Therefore for more appropriate scalings of $\sigma_1(), \sigma_2()$, we can expect triple descent with two peaks.

Intuition of Multiple Descent in Multi-Component Models







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

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

The above are two extreme cases, each showing double descent with different peak locations. Therefore for more appropriate scalings of $\sigma_1(), \sigma_2()$, we can expect triple descent with two peaks.

Theoretical Demonstration of Triple Descent in DRFMs

Data distribution

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

Double random feature model

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

$\mathbf{\Theta}$: fixed at randomly generated values

a: trainable parameters

Ridge(less) Regression & Limit of Excess Risk

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

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

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

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

Ridge(less) 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

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

Our goal: calculate

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

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 $\boldsymbol{\psi} = [\psi_1, \psi_2, \psi_3]$.

Main Assumptions

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

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

where

$$\mathcal{R}(\lambda,\boldsymbol{\psi},\boldsymbol{\mu},F_1,\tau)=\|\boldsymbol{\beta}\|_2^2\cdot\left(\frac{1}{M_D^2}+\mathbf{L}_{3,4}+\mathbf{L}_{1,4}\right)+\tau^2\left(\mathbf{L}_{2,3}+\mathbf{L}_{1,2}\right).$$

$M_D \in \mathbb{R}$ and $\mathbf{L} \in \mathbb{R}^{4\times4}$ are given as follows:

Main Theory for Asymptotic Excess Risk

Theorem. Under Assumption 1, it holds that

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

where

$$\mathcal{R}(\lambda,\boldsymbol{\psi},\boldsymbol{\mu},F_1,\tau)=\|\boldsymbol{\beta}\|_2^2\cdot\left(\frac{1}{M_D^2}+\mathbf{L}_{3,4}+\mathbf{L}_{1,4}\right)+\tau^2\left(\mathbf{L}_{2,3}+\mathbf{L}_{1,2}\right).$$

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

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

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

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

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

Main Theory for Asymptotic Excess Risk

Theorem. Under Assumption 1, it holds that

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

where

$$\mathcal{R}(\lambda,\boldsymbol{\psi},\boldsymbol{\mu},F_1,\tau) = \|\boldsymbol{\beta}\|_2^2 \cdot \left(\frac{1}{M_D^2} + \mathbf{L}_{3,4} + \mathbf{L}_{1,4}\right) + \tau^2\left(\mathbf{L}_{2,3} + \mathbf{L}_{1,2}\right).$$

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

(1) implicit functions $\nu_1, \nu_2, \nu_3 : \mathbb{C}_+ \to \mathbb{C}_+$ 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 ν_j's exist and are unique.

Main Theory for Asymptotic Excess Risk

Theorem. Under Assumptions 1 and 2, it holds that

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

where

$$\mathcal{R}(\lambda,\boldsymbol{\psi},\boldsymbol{\mu},F_1,\tau) = \|\boldsymbol{\beta}\|_2^2 \cdot \left(\frac{1}{M_D^2} + \mathbf{L}_{3,4} + \mathbf{L}_{1,4}\right) + \tau^2\left(\mathbf{L}_{2,3} + \mathbf{L}_{1,2}\right).$$

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

(2) Denote $\nu_j^* = \nu_j(\sqrt{\lambda} i), j = 1,2,3$. 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

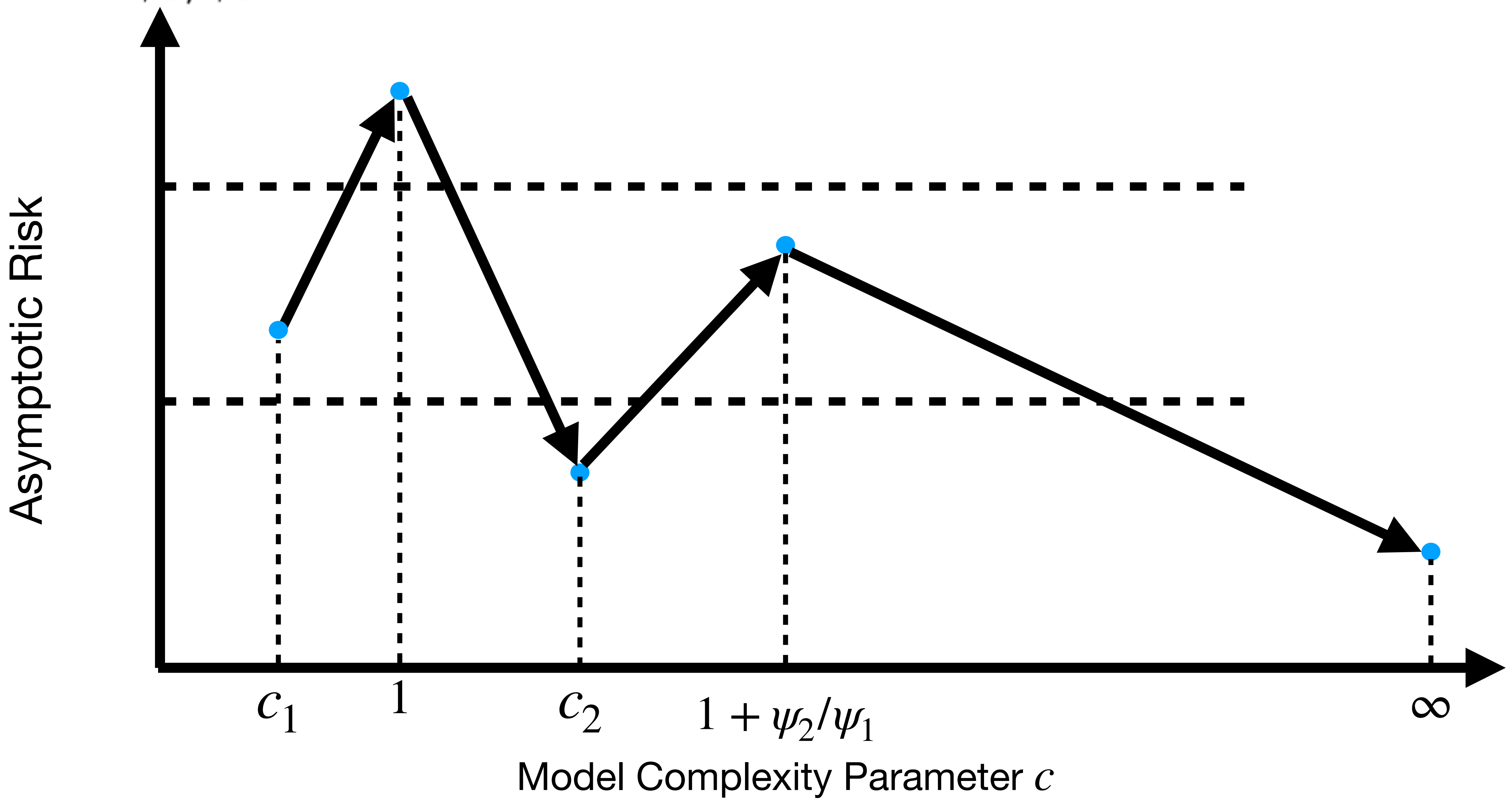
Proposition. Under Assumptions 1 and 2, it holds that

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* $0 < r < \infty$, $\lim_{\substack{\psi_1, \psi_2 \to \infty \\ \psi_1/\psi_2 = r}} \mathcal{R} < +\infty$

Theoretical Demonstration of Triple Descent

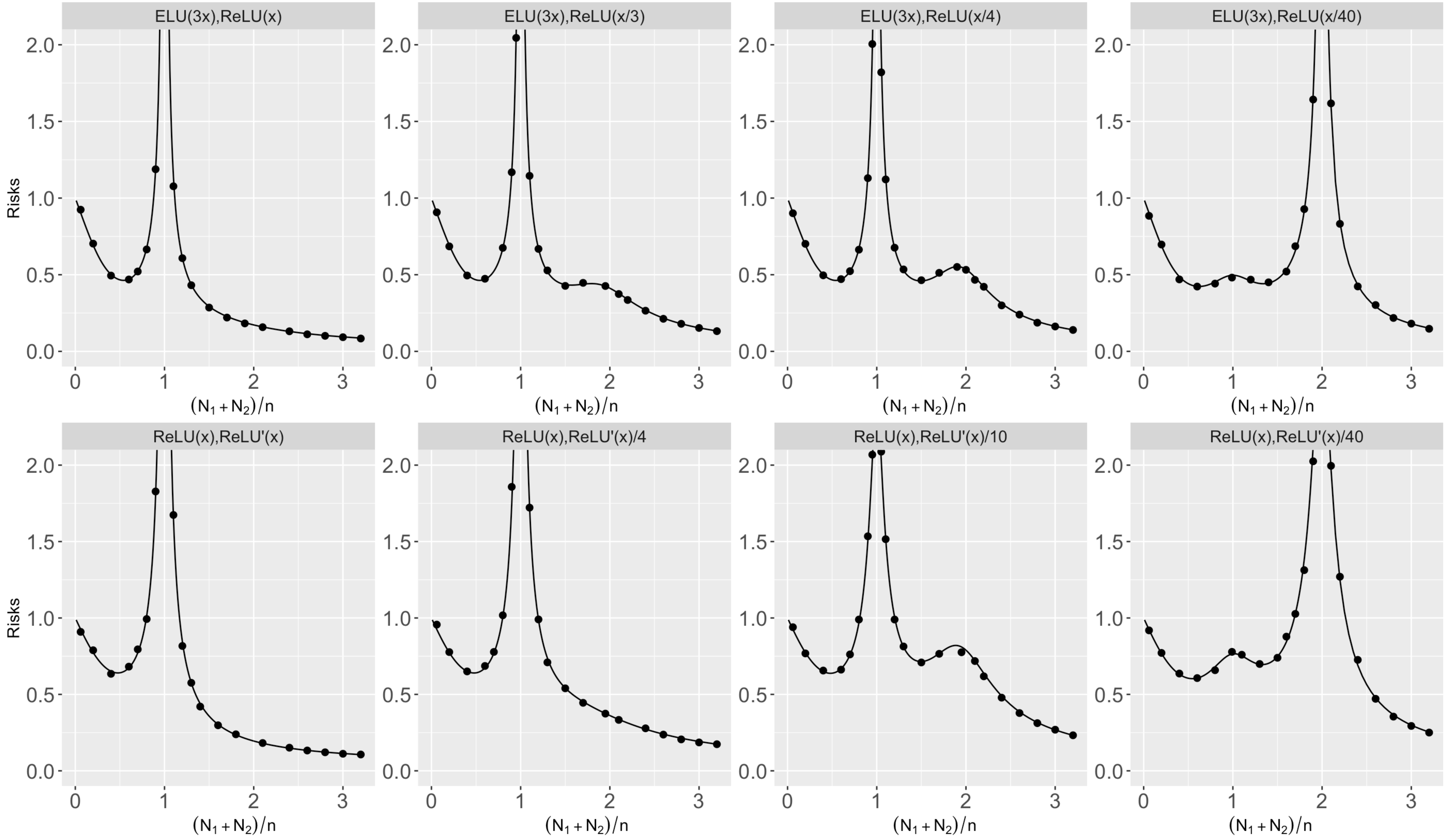
Proposition. Under Assumptions 1 and 2, it holds that

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* $0 < r < \infty$, $\lim_{\substack{\psi_1,\psi_2 \to \infty \\ \psi_1/\psi_2 = r}} \mathcal{R} < +\infty$



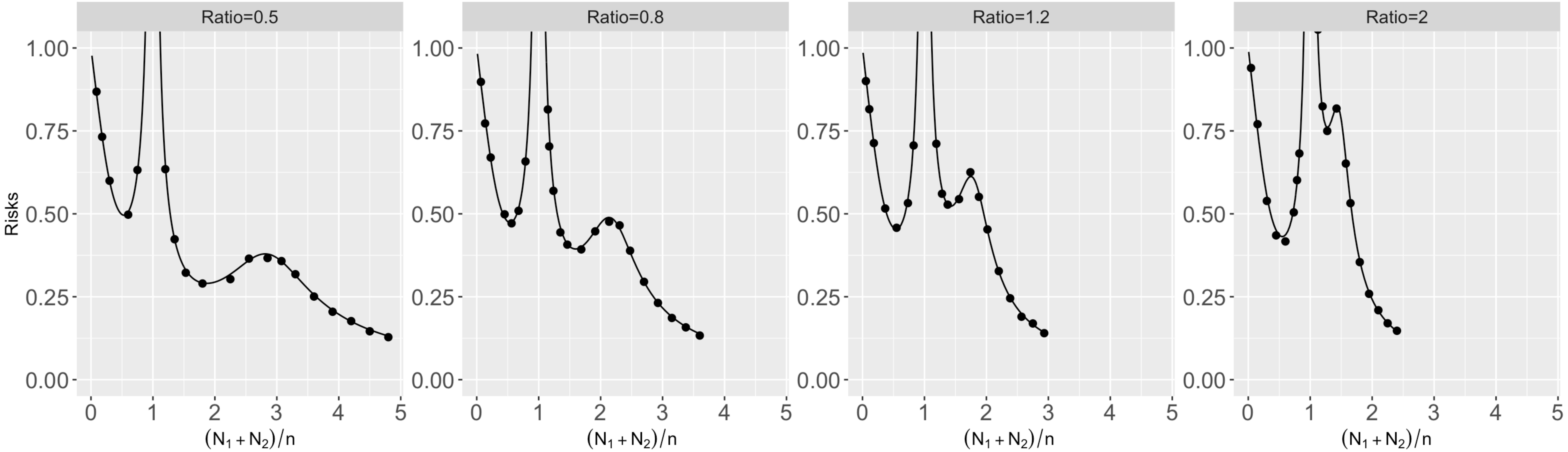
Simulations

The scale difference of activation functions:



Simulations

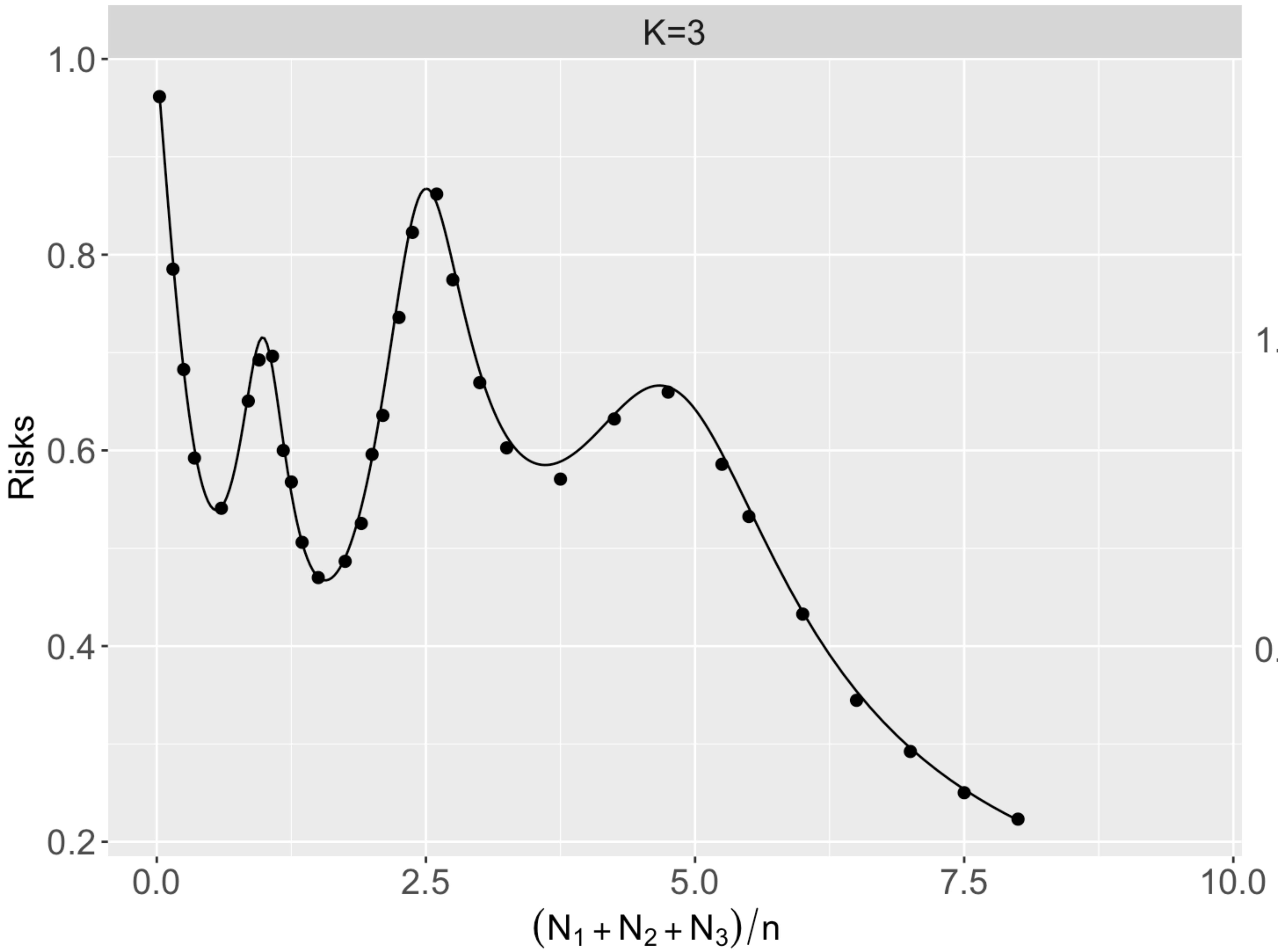
Impact of the ratio N_1/N_2:



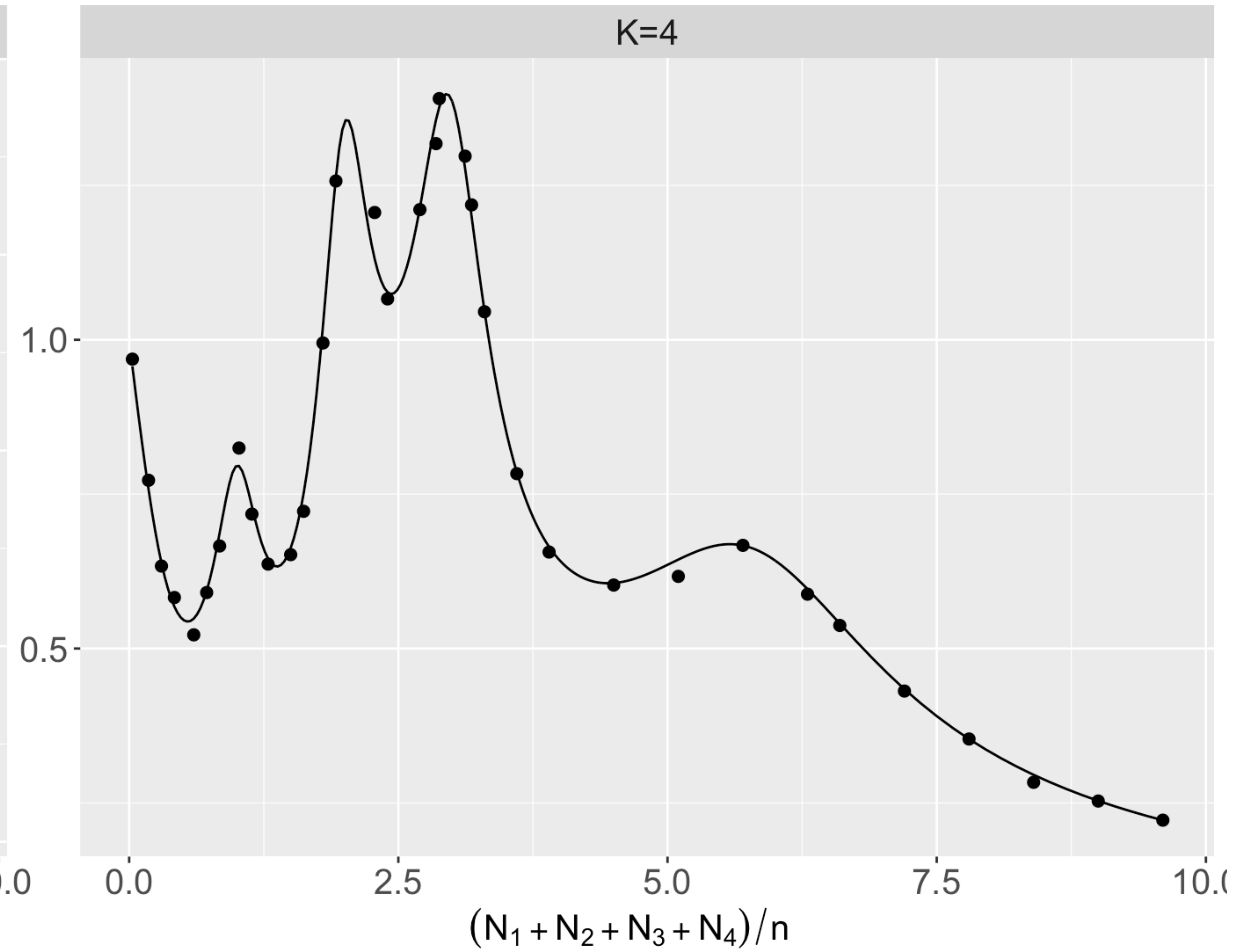
Peaks Location: $N_1/n = 1 \longrightarrow (N_1 + N_2)/n = 3, \quad 9/4, \quad 11/6, \quad 3/2.$

Simulations

Multiple descent with 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”

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!