

Eigenvectors of Fisher Information Matrix for Simple ReLU Networks

武石, 啓成
九州大学

飯田, 昌澄
九州大学

竹内, 純一
九州大学

<https://hdl.handle.net/2324/7174507>

出版情報 : 2021-11. The Institute of Electronic, Information and Communication Engineers
バージョン :
権利関係 :

Eigenvectors of Fisher Information Matrix for Simple ReLU Networks

Abstract

In learning neural network parameters, it is important to examine the Fisher information matrix (FIM). We investigate the FIM of a one hidden layer network with the ReLU activation function and identify the main eigenvalues and eigenvectors of the FIM under certain conditions.

武石 啓成, 飯田 昌澄, 竹内 純一 (九州大学)

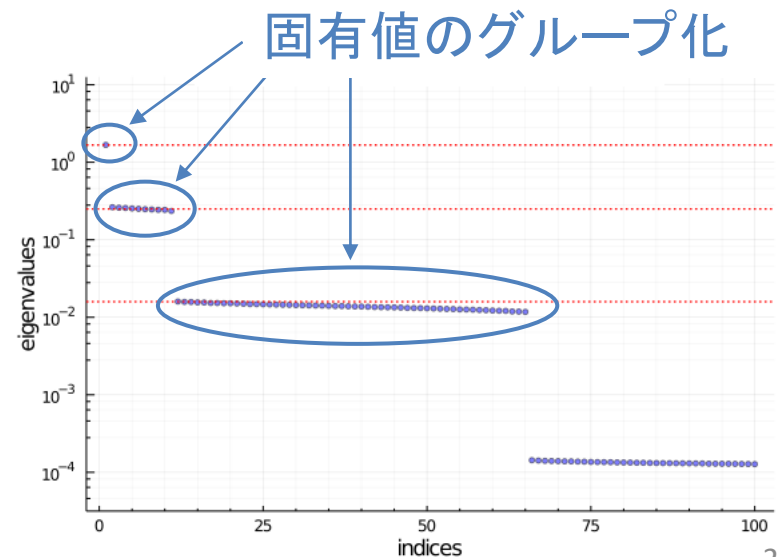
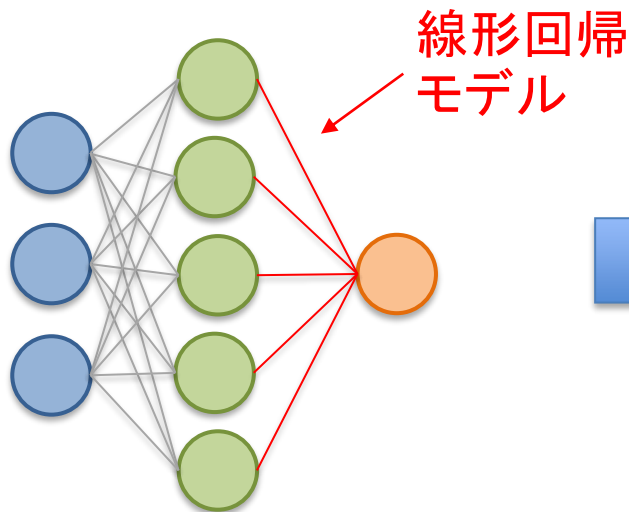
有益な助言を頂いた長岡浩司氏, 村田昇氏, 三村和史氏に謝意を表する

研究背景・概要

- 深層学習→理論的性能評価が課題（汎化誤差の評価）
- 汎化誤差の解析→Fisher情報行列が重要な役割

研究成果

簡単なReLUニューラルネットワークにおいて、Fisher情報行列の主要な固有値と固有ベクトルを特定



問題設定

ベクトルは特に断りが
ない限り, 行ベクトルで
定義する

- 隠れ層1層のニューラルネットワーク
(バイアス項無し)
- 隠れ層の j 番目の出力

$$X_j = \varphi \left(\sum_{i=1}^d x_i W_{ij} \right) \text{ for } j = 1, 2, \dots, p$$

$$\varphi(z) = \begin{cases} z & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases} \quad (\text{ReLU})$$

$$W \in \mathbb{R}^{d \times p} \quad (\text{重み行列 } p \gg d)$$

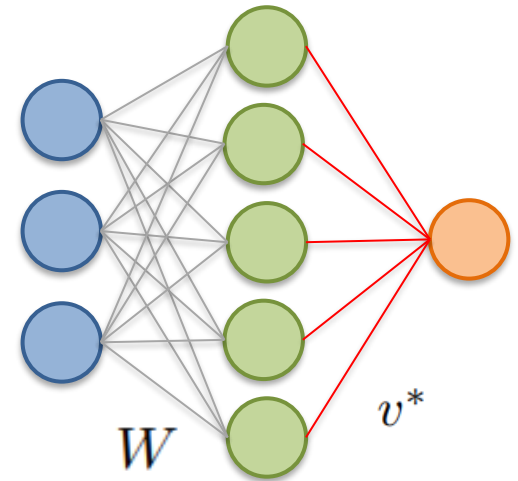
- 出力層

$$y = X v^{*T} + \epsilon \quad v^* \in \mathbb{R}^p \quad \epsilon \sim N(0, \sigma^2)$$

目標

学習データ $\{x_{(1)}, y_{(1)}\}, \dots, \{x_{(n)}, y_{(n)}\}$ から v^* を推定したい

入力層 隠れ層 出力層
 $x \in \mathbb{R}^d$ $X \in \mathbb{R}^p$ $y \in \mathbb{R}$



汎化誤差

- v^* の推定結果を v とすると, 汎化誤差は

$$\begin{aligned} E[(y - Xv^T)^2] &= E[(X(v^* - v)^T + \epsilon)^2] \\ &= (v - v^*)J(v - v^*)^T + \sigma^2 \end{aligned}$$

- 期待値は x と ϵ に対して取る ($x \sim N(0, I_d)$)
- $J = E[X^T X]$

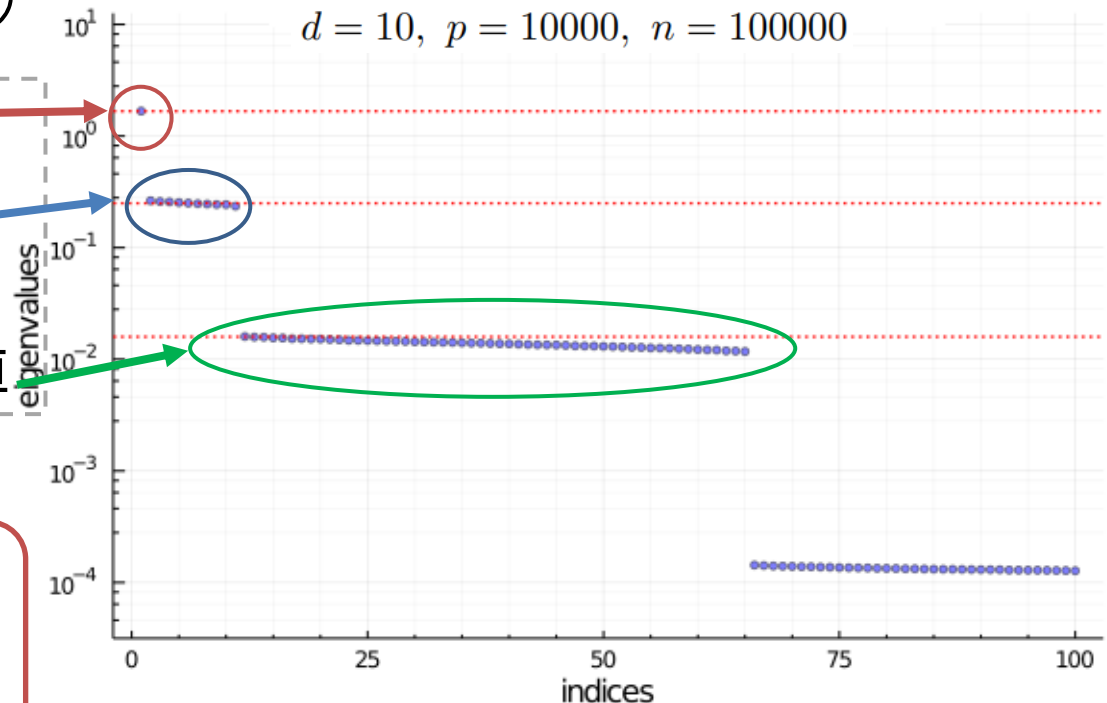
- J の主要な固有ベクトルの方向に対して, パラメータ v の学習を進める必要あり
- J はFisher情報行列 I の定数倍

$$I_{ij} = E \left[\frac{\partial}{\partial v_i} \log p(x, y|v) \frac{\partial}{\partial v_j} \log p(x, y|v) \right] = \frac{J_{ij}}{\sigma^2}$$

J の固有値分布

- $J^{(n)} = \frac{1}{n} \sum_{t=1}^n X_{(t)}^T X_{(t)}$ ($J = E[X^T X]$ の近似)
- 数値シミュレーションにより, $J^{(n)}$ の固有値を計算
($W_{ij} \sim N(0, 1/p)$ で生成)

1. 最大固有値
2. d 個の固有値
3. $d(d+1)/2 - 1$ 個の固有値



本研究の動機

各グループの固有ベクトルを
特定し, 定式化したい

各ベクトルの定義

- $W_l \in \mathbb{R}^p$: 行列 W の l 番目の行ベクトル

2.

- $W^{(i)} \in \mathbb{R}^d$: 行列 W の i 番目の列ベクトル

- $v^{(0)} \in \mathbb{R}^p$: i 番目の要素が以下で定義される行ベクトル

$$v_i^{(0)} = \frac{\|W^{(i)}\|}{\sqrt{d}} \quad (W_{ij} \sim N(0, 1/p) \text{ のとき } v^{(0)} \text{ は } \frac{1}{\sqrt{p}}(1, \dots, 1) \text{ に近い})$$

1.

- $v^{(\alpha, \beta)} \in \mathbb{R}^p$: i 番目の要素が以下で定義される行ベクトル

$$v_i^{(\alpha, \beta)} = \frac{\sqrt{d}W_{\alpha i}W_{\beta i}}{\|W^{(i)}\|} \quad (1 \leq \alpha \leq \beta \leq d)$$

- $v^{(\gamma)} \in \mathbb{R}^p$: $v^{(\gamma)} = \frac{v^{(\gamma, \gamma)} - v^{(0)}}{\sqrt{2}} \quad (1 \leq \gamma \leq d) \quad \left(\sum_{\gamma=1}^d v^{(\gamma)} = 0 \right)$

3.

$$- \sum_{\gamma=1}^d v_i^{(\gamma, \gamma)} = \sum_{\gamma=1}^d \frac{\sqrt{d}W_{\gamma i}^2}{\|W^{(i)}\|} = \frac{\sqrt{d}\|W^{(i)}\|^2}{\|W^{(i)}\|} = dv_i^{(0)} \text{ より, } v^{(0)} \text{ は } v^{(\gamma, \gamma)} \text{ の重心}$$

行列 J の各要素の計算

補題1

$\theta_{ij} = \angle(W^{(i)}, W^{(j)}) \in [0, \pi]$ とすると, $i \neq j$ のとき以下が成り立つ

$$J_{ij} = \frac{1}{2\pi} ((\pi - \theta_{ij}) \cos \theta_{ij} + \sin \theta_{ij}) \|W^{(i)}\| \|W^{(j)}\|$$

対角成分は, 上式で $\theta_{ii} = 0$ とすると以下となる

$$J_{ii} = \frac{1}{2} \|W^{(i)}\|^2$$

↑ 導出

$e = W^{(i)T} / \|W^{(i)}\|$
 $w = W^{(j)T}$ を代入

定理1 (Tian 2017)

行ベクトル $e, w, x \in \mathbb{R}^d$ (e は単位ベクトル) と,

$D(w) = \begin{cases} 1 & \text{if } x \cdot w > 0 \\ 0 & \text{if } x \cdot w \leq 0 \end{cases}$ に対して, $F(e, w) = x^T D(e) D(w) x \cdot w$ を定義する

$x \sim N(0, I_d)$ としたとき, $\theta = \angle(e, w) \in [0, \pi]$ に対して, 以下が成り立つ

$$E[F(e, w)] = \frac{1}{2\pi} [(\pi - \theta)w^T + (\|w\| \sin \theta)e^T]$$

行列 J の固有値分解

補題1

$$J_{ij} = \frac{1}{2\pi} ((\pi - \theta_{ij}) \cos \theta_{ij} + \sin \theta_{ij}) A_{ij}$$

$$A_{ij} = \|W^{(i)}\| \|W^{(j)}\|$$

$$\cos \theta_{ij} = W^{(i)} \cdot W^{(j)} / A_{ij}$$

$$\arcsin z = \sum_{n=0}^{\infty} \binom{2n}{n} \frac{z^{2n+1}}{2^{2n}(2n+1)}$$

$$J_{ij} = \frac{A_{ij}}{2\pi} + \frac{W^{(i)} \cdot W^{(j)}}{4} + \frac{(W^{(i)} \cdot W^{(j)})^2}{4\pi A_{ij}} + \frac{1}{2\pi} \sum_{n=1}^{\infty} \binom{2n}{n} \frac{(W^{(i)} \cdot W^{(j)})^{2n+2}}{2^{2n}(2n+1)(2n+2)A_{ij}^{2n+1}}$$

$$A_{ij} = \|W^{(i)}\| \|W^{(j)}\| = d \left(v^{(0)T} v^{(0)} \right)_{ij}$$

$$W^{(i)} \cdot W^{(j)} = \sum_{l=1}^d W_{l,i} W_{l,j} = \sum_{k=1}^d (W_l^T W_l)_{ij} \quad \text{etc.}$$

定理2

$$J = \underbrace{\frac{2d+1}{4\pi} v^{(0)T} v^{(0)}}_{1.} + \underbrace{\frac{1}{4} \sum_{k=1}^d W_k^T W_k}_{2.} + \underbrace{\frac{1}{2\pi d} \left(\sum_{\gamma=1}^d v^{(\gamma)T} v^{(\gamma)} + \sum_{\alpha < \beta} v^{(\alpha,\beta)T} v^{(\alpha,\beta)} \right)}_{3.} + \textcircled{R} \text{ 半正定値}$$

正規直交性の評価

補題2

$W_{ij} \sim N(0, 1/p)$ とし, $D = (d+1)(d+2)(d^2 + 3d + 4)/8$ とする.

このとき, 任意の正数 δ に対して, 確率 $1 - CD/(\delta^2 p)$ 以上で下記が成立

$$(1) \quad \begin{aligned} \left| \|v^{(0)}\|^2 - 1 \right| &\leq \delta, \\ \left| \|W_l\|^2 - 1 \right| &\leq \delta, \quad (1 \leq l \leq d) \\ \left| \|v^{(\alpha, \beta)}\|^2 - 1 \right| &\leq \delta + \xi(d_1), \quad (1 \leq \alpha < \beta \leq d) \\ \left| \|v^{(\gamma)}\|^2 - 1 \right| &\leq \delta + \xi(d_2). \quad (1 \leq \gamma \leq d) \end{aligned} \quad (\text{正規性})$$

$$(2) \quad \begin{aligned} V_1 &= \{v^{(0)}, W_l, v^{(\alpha, \beta)} \mid 1 \leq l \leq d, 1 \leq \alpha < \beta \leq d\} \\ V_2 &= \{v^{(\gamma)} \mid 1 \leq \gamma \leq d\}, \quad V = V_1 \cup V_2 \text{ とすると, 相異なるベクトル} \\ & v, v' \in V \text{ について, 以下が成り立つ} \end{aligned}$$

$$|v \cdot v'| \leq \delta \quad (v \notin V_2 \text{ or } v' \notin V_2)$$

(直交性)

$$\left| v \cdot v' - \left(-\frac{1}{d-1} \right) \right| \leq \delta + \frac{\xi(d_2)}{d-1} \quad (v, v' \in V_2)$$

($d_1 = d - 2, d_2 = d - 1$. $\xi(d)$ は任意の $\eta \in (0, 1/2)$ について, $O(1/d^{1/2-\eta})$)

Wに関する確率的評価

定理3

$W_{ij} \sim N(0, 1/p)$ とし, $D = (d+1)(d+2)(d^2 + 3d + 4)/8$ とする.
 このとき, 任意の正数 δ に対して, 確率 $1 - CD/(\delta^2 p)$ 以上で下記が成立

(1) $\text{tr}(J) \leq \frac{d}{2}(1 + \delta)$ (固有値和の上界)

(2) $\frac{v^{(0)}}{\|v^{(0)}\|} J \frac{v^{(0)T}}{\|v^{(0)}\|} \geq \frac{2d+1}{4\pi}(1 - \delta)$ (1. 方向の固有値の下界)

$\frac{W_l}{\|W_l\|} J \frac{W_l^T}{\|W_l\|} \geq \frac{1}{4}(1 - \delta)$ (2. 方向の固有値の下界)

$\frac{v^{(\alpha,\beta)}}{\|v^{(\alpha,\beta)}\|} J \frac{v^{(\alpha,\beta)T}}{\|v^{(\alpha,\beta)}\|} \geq \frac{1}{2\pi d}(1 - \delta - \xi(d_1))$ (3. 方向の固有値の下界)

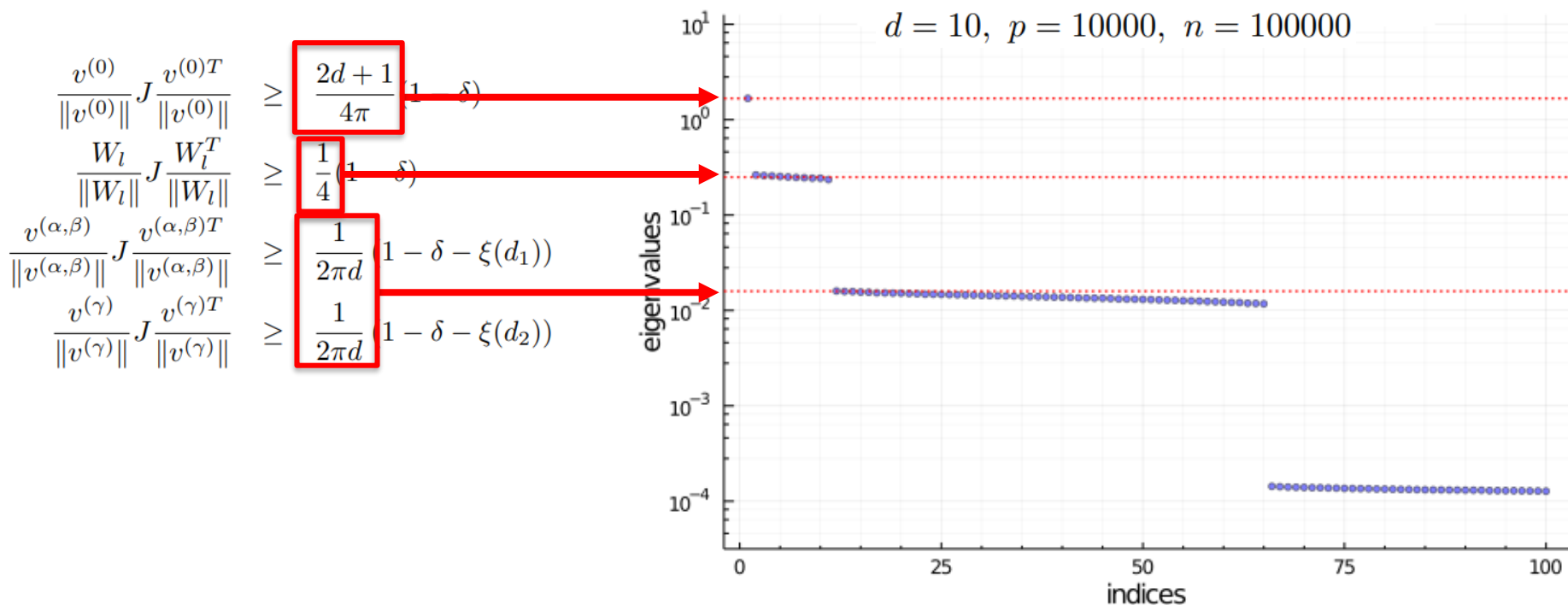
$\frac{v^{(\gamma)}}{\|v^{(\gamma)}\|} J \frac{v^{(\gamma)T}}{\|v^{(\gamma)}\|} \geq \frac{1}{2\pi d}(1 - \delta - \xi(d_2))$

($d_1 = d - 2, d_2 = d - 1$. $\xi(d)$ は任意の $\eta \in (0, 1/2)$ について, $O(1/d^{1/2-\eta})$)

$$\frac{2d+1}{4\pi} + \frac{1}{4} \cdot d + \frac{1}{2\pi d} \cdot \left(\frac{d(d-1)}{2} + d - 1 \right) \simeq 0.977 \cdot \frac{d}{2}$$

数値シミュレーション

- 定理3による固有値の下界の見積もり値を，数値シミュレーション結果と比較



まとめ

- 簡単なReLUニューラルネットワークにおいて, Fisher情報行列の主要な固有値と固有ベクトルを特定
 - 固有値の大きさの「グループ化」



今後の展望

- 多層のニューラルネットワークや、ReLU以外の活性化関数への拡張
- 汎化誤差の評価に本研究の成果を活用

参考文献 (1/2)

- [1] S. Amari, “Any Target Function Exists in a Neighborhood of Any Sufficiently Wide Random Network: A Geometrical Perspective,” *Neural Computation* 32, pp.1431-1447, 2020
- [2] R. Bapat, *Nonnegative Matrices and Applications*, Cambridge University Press, 1997
- [3] A. R. Barron and T. M. Cover, “Minimum complexity density estimation,” *IEEE Trans. Inf. Theory*, vol. 37, no. 4, pp. 1034-1054, Jul. 1991.
- [4] M. Belkin, D. Hsu, S. Ma, and S. Mandal, “Reconciling modern machine learning practice and the bias-variance trade-off,” *In Proc. of the National Academy of Sciences*, 116.32, pp. 15849-15854, 2019.
- [5] M. Belkin, D. Hsu, and J. Xu, “Two models of double descent for weak feature,” *SIAM Journal on Mathematics of Data Science*, 2(4):1167-1180, 2020.
- [6] I. S. Gradshteyn and I.M. Ryzhik, *Table of Integrals, Series, and Products*, fourth edition, Academic Press, 1980
- [7] A. Jacot, F. Gabriel, and C. Hongler, “Neural tangent kernel: Convergence and generalization in neural networks,” Presented at the 32nd Conference on Neural Information Processing Systems, arXiv:1806.07572v3, 2018.

参考文献 (2/2)

- [8] R. Karakida, S. Akaho, and S. Amari, “Universal Statistics of Fisher Information in Deep Neural Networks: Mean Field Approach,” Presented at the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS), arXiv:1806.01316, 2019.
- [9] M. Kawakita and J. Takeuchi, “Minimum Description Length Principle in Supervised Learning with Application to Lasso,” *IEEE Trans. Inf. Theory*, vol. 66, no. 7, pp. 4245-4269, Jul. 2020.
- [10] Y. LeCun, L. Bottou, G. B. Orr, and K. Muller, “Efficient backprop,” In *Neural networks: Tricks of the trade*, pp. 9-50. Springer, 1998.
- [11] J. Rissanen, “Modeling by shortest data description,” *Automatica*, vol. 14, no. 5, pp. 465-471, Sep. 1978.
- [12] Y. Tian, “An analytical formula of population gradient for two-layered ReLU network and its applications in convergence and critical point analysis,” in *Proc. International Conference on Machine Learning*, pp. 3404-3413, 2017.
- [13] M. J. Wainwright, *High-Dimensional Statistics*, Cambridge University Press, 2019