深層学習の汎化の謎をめぐって

森 貴司（理研CEMS）

T. Mori and M. Ueda, "*Is deeper better? It depends on locality of relevant features*", arXiv:2005.12488
T. Mori and M. Ueda, "*Improved generalization by noise enhancement*", arXiv:2009.13094

自己紹介

専門：統計物理学　　厳密な結果の証明と多体系の数値計算

- 長距離相互作用系の平衡統計力学
- 孤立量子系の熱平衡化（統計力学の基礎づけ）
- 周期駆動量子多体系の非平衡ダイナミクス
- 開放量子系の非平衡ダイナミクス

機械学習の原理に興味

モデルの自由度が大きいほど汎化が良くなるように見える
そこに新しい「物理」はないか？
「理解する」とはどういうことか？

くりこみ群の拡張？
大自由度系のダイナミクス

深層学習の大きな謎

訓練データ　$\mathscr{D} = \{(x^{(\mu)}, y^{(\mu)}) : \mu = 1,2,\ldots N\}$　　$x^{(\mu)} \in \mathbb{R}^D, \quad y^{(\mu)} \in \{\pm 1\}$

ネットワーク　$f(x, w)$　　パラメータ $w \in \mathbb{R}^{N_{\text{para}}}$

過剰パラメータ領域 (overparameterized regime)

パラメータの数 $N_{\text{para}} \sim 10^8 \gg$ 訓練データの数 $N \sim 10^5$

- すべての訓練データを完全にfitting可能
- にもかかわらず深刻な過学習は起こらない
- ネットワークを広くすることでパラメータ数を増やすほど汎化性能が向上

C. Zhang et al.,"*Understanding Deep Learning Requires Rethinking of Generalization*", ICLR 2017
S. Spigler et al.,"*A jamming transition from under- To over-parametrization affects generalization in deep learning*", J. Phys. A (2019)

訓練誤差と汎化誤差

訓練データ　$\mathscr{D} = \{(x^{(\mu)}, y^{(\mu)}) : \mu = 1,2,\ldots N\}$　　$x^{(\mu)} \in \mathbb{R}^D, \quad y^{(\mu)} \in \{\pm 1\}$

ネットワーク　$f(x, w)$　　パラメータ $w \in \mathbb{R}^{N_{\text{para}}}$

予測クラス $\hat{y}$　　$f(x^{(\mu)}, w) \geq 0 \to \hat{y}^{(\mu)} = 1$　　$f(x^{(\mu)}, w) < 0 \to \hat{y}^{(\mu)} = -1$

訓練誤差 $E_{\text{train}} = \dfrac{1}{N}\sum_{\mu=1}^{N}\left(1 - \delta_{\hat{y}^{(\mu)}, y^{(\mu)}}\right)$

訓練データ内で誤った分類をしたサンプルの割合

汎化誤差 $E_{\text{test}} = \mathbb{E}_{x,y}\left[1 - \delta_{\hat{y},y}\right]$

未知のデータの誤り確率

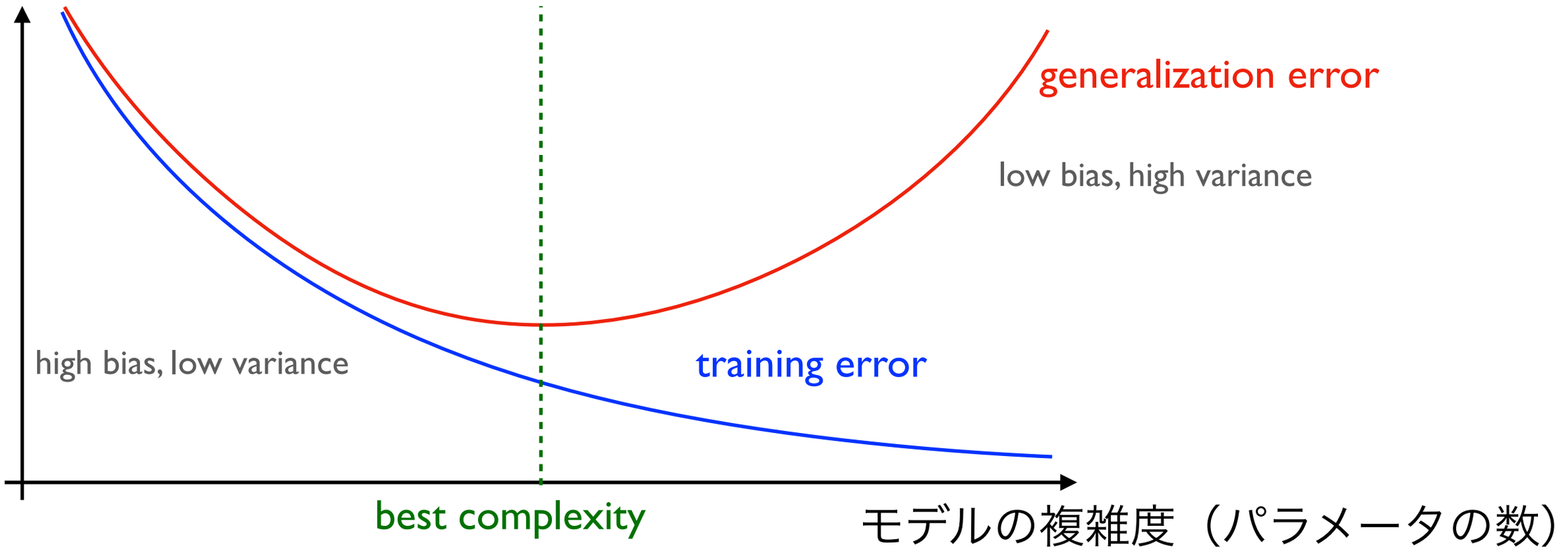
戦略：訓練誤差を下げることで低い汎化誤差を実現

損失関数　$L(w) = \dfrac{1}{N}\sum_{\mu=1}^{N}\ell(f(x^{(\mu)}, w); y^{(\mu)})$

$$w_{t+1} = w_t - \eta \nabla_w L(w_t)$$

bias-variance tradeoff

訓練誤差を下げても汎化誤差が下がる保証はない



モデル（ネットワーク）が単純すぎると訓練誤差を小さくできない（bias大）

モデルが複雑すぎると訓練データの無用な特徴（ノイズなど）をも忠実に再現しようとする

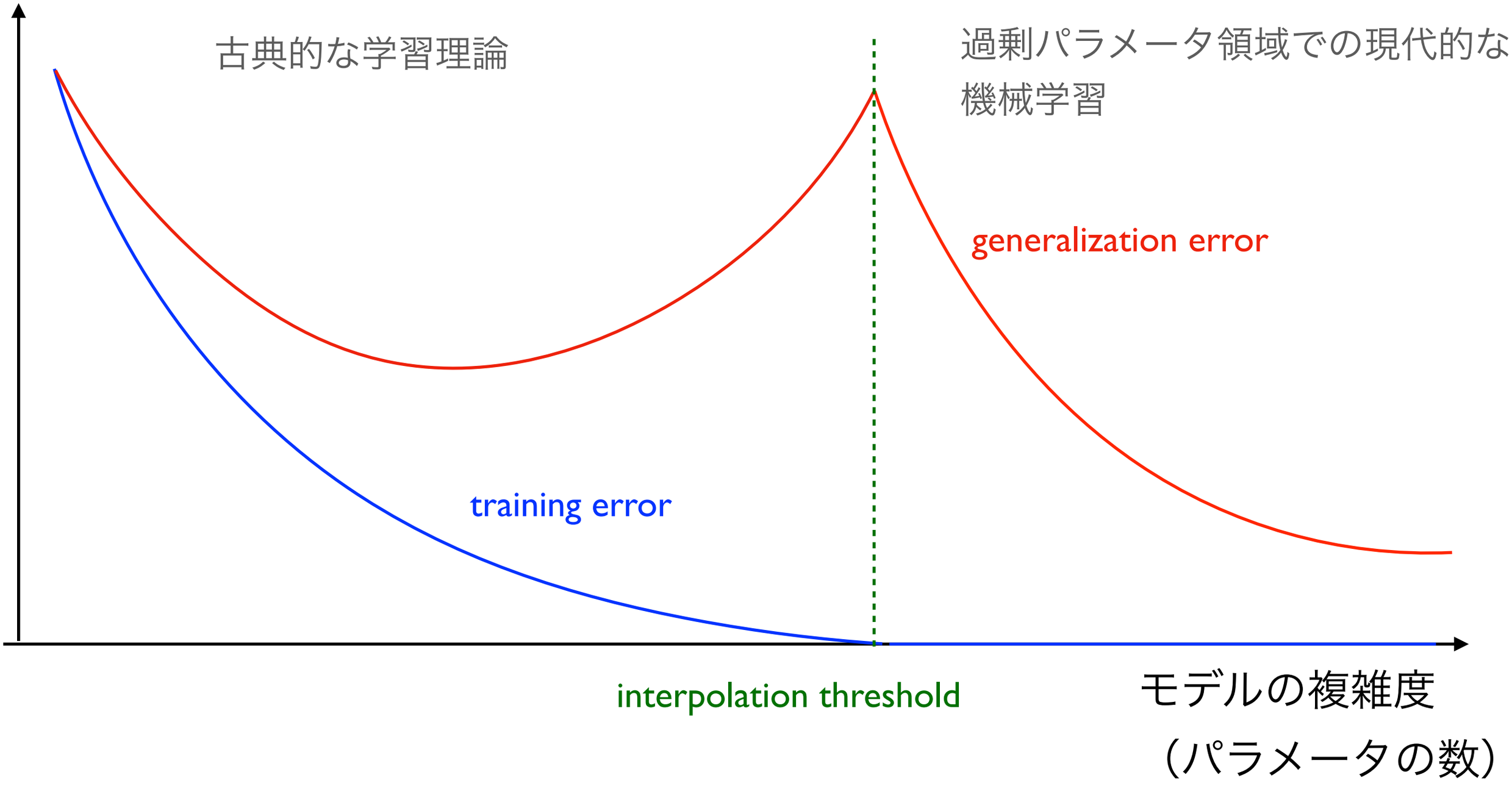
→学習した関数は特定の訓練データ，初期状態に強く依存（variance大）

→訓練誤差と汎化誤差のギャップが大きくなる（過学習）

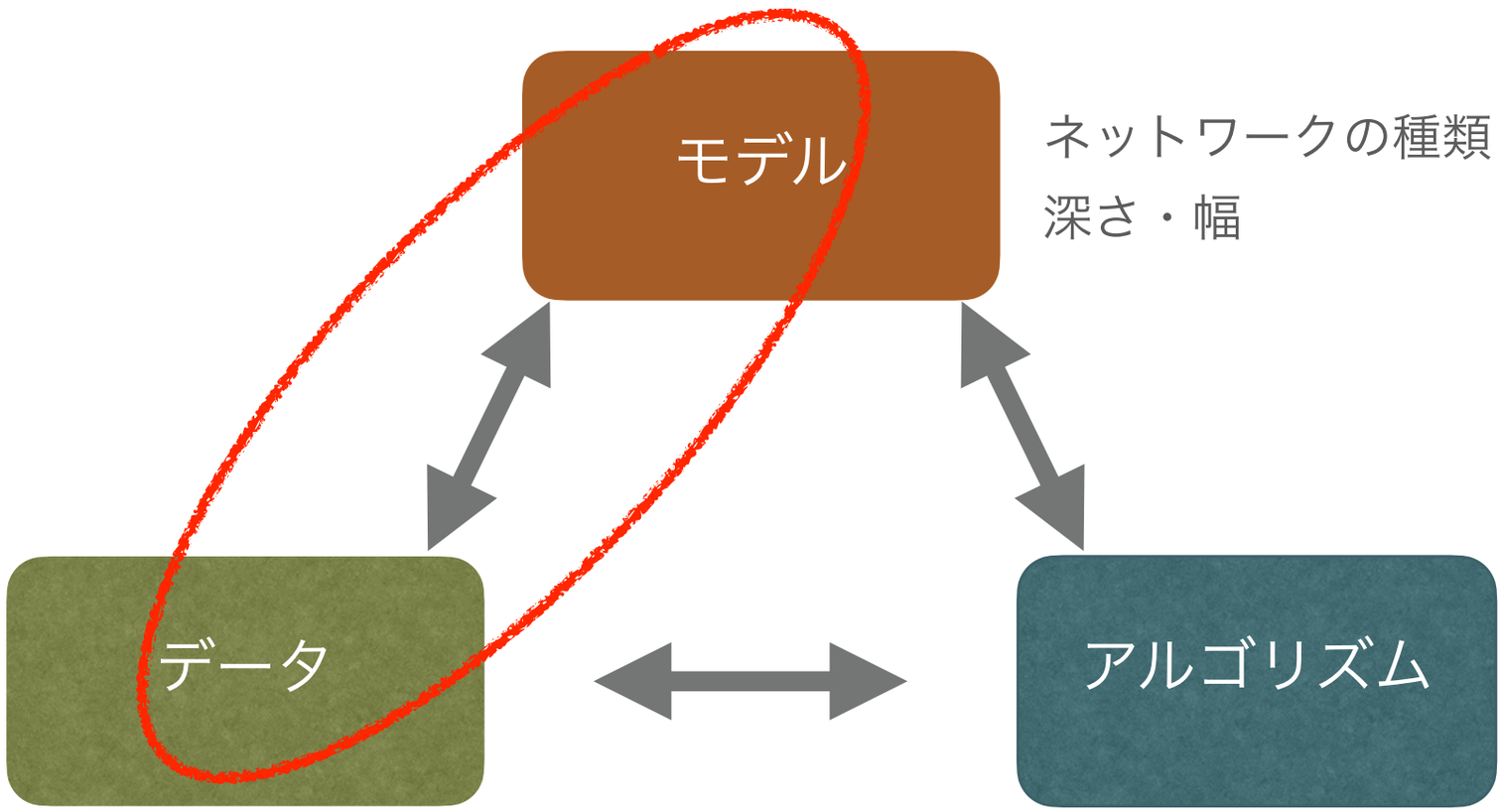
過剰パラメータ領域とdouble descent curve

S. Spigler, M. Geiger, S. d'Ascoli, L. Sagun, G. Biroli, and M. Wyart, J. Phys. A (2019)
M. Belkin, D. Hsu, S. Ma, and S. Mandal, Proc. Natl. Acad. Sci. USA (2019)

パラメータの数をさらに増やしていくと汎化誤差が**再び減少**



過剰パラメータ領域での汎化の理解に向けて



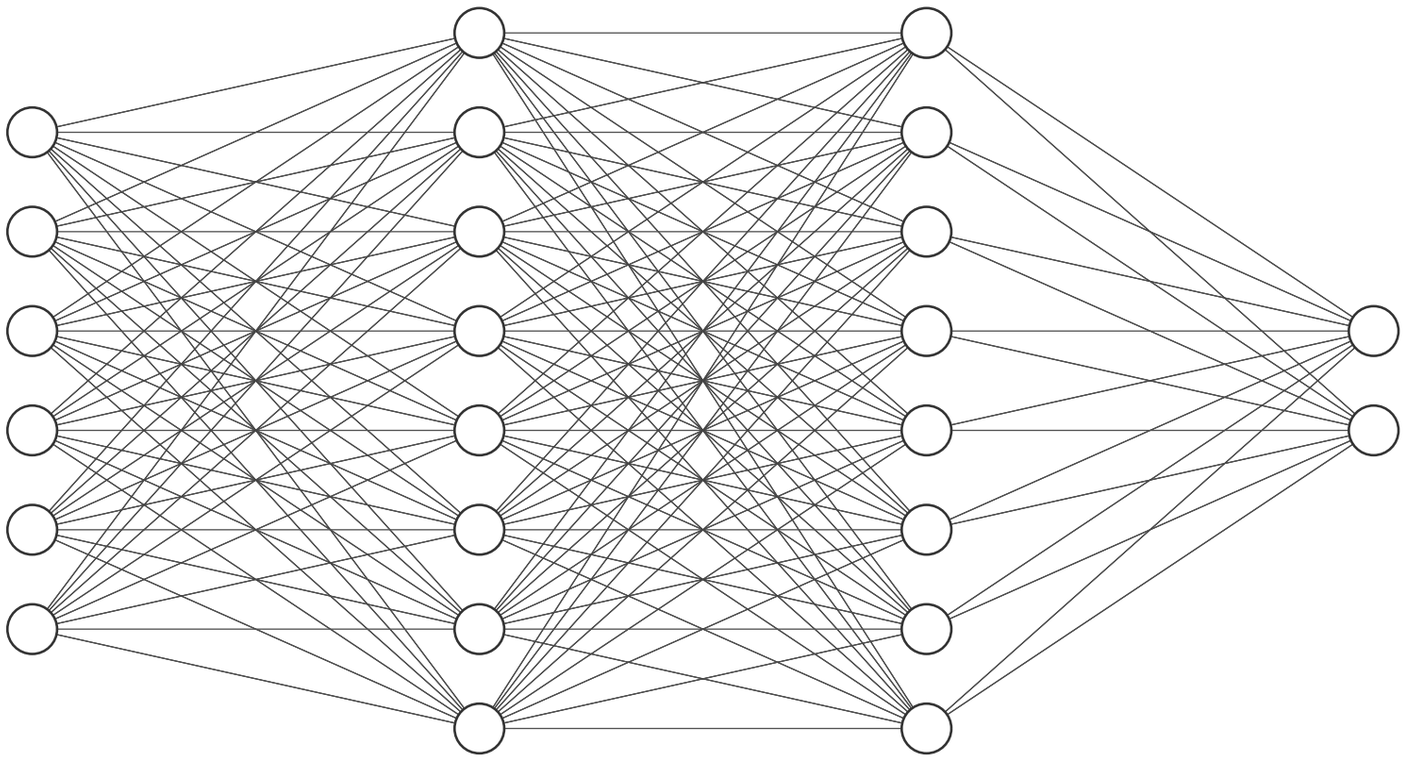
有効次元
対称性・局所性

ダイナミクス
(SGD: learning rate, minibatch size)

モデル：全結合ニューラルネットワーク

深さ L 幅 H の全結合ニューラルネット

入力層　　隠れ層 $\times L$　　出力層



softmax
+ cross-entropy loss

各隠れ層のユニット数 H

過剰パラメータ領域：隠れ層の幅を増やす

(i) ダイナミクス：十分広い隠れ層を持つニューラルネットでは局所解に陥ることなく損失関数の最小点に到達

Z. Allen-Zhu, Y. Li, and Z. Song, arXiv:1811.03962
S. S. Du, J. D. Lee, H. Li, L. Wang, and X. Zhai, arXiv:1811.03804
K. Kawaguchi and J. Huang, arXiv:1908.02419

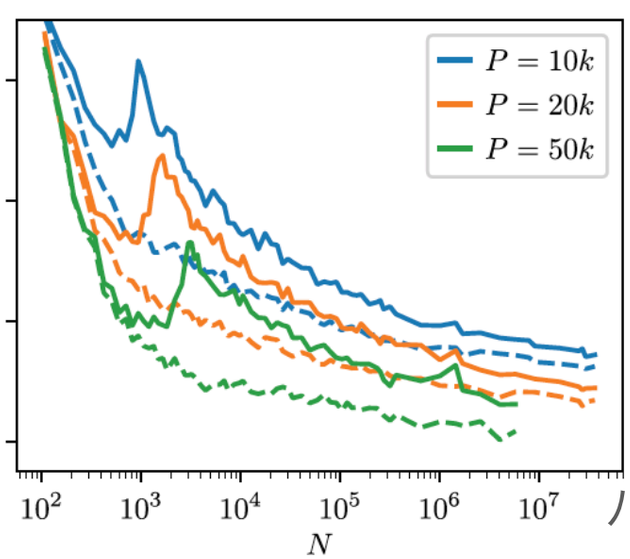
幅無限大の極限でカーネル法と等価（Neural Tangent Kernel）
線形の最適化問題となり(i)をよく説明

A. Jacot, F. Gabriel, C. Hongler, NeurIPS 2018

(ii) 汎化：隠れ層の幅が広いほど汎化性能が向上 double descent curve

S. Spigler, M. Geiger, S. d'Ascoli, L. Sagun, G. Biroli, and M. Wyart, J. Phys. A (2019)

汎化誤差
(test error)

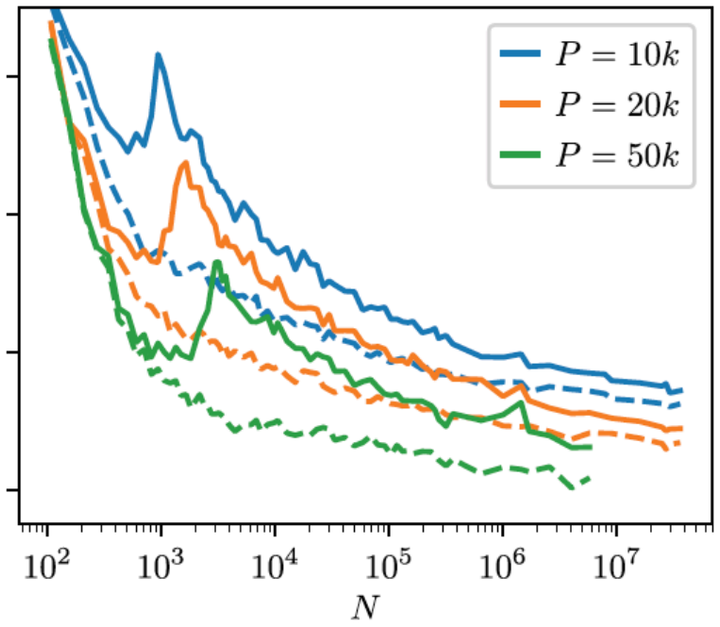


P:訓練データのサンプル数

パラメータの数

ジャミング転移の物理との対応

S. Spigler, M. Geiger, S. d'Ascoli, L. Sagun, G. Biroli, and M. Wyart, J. Phys. A (2019)

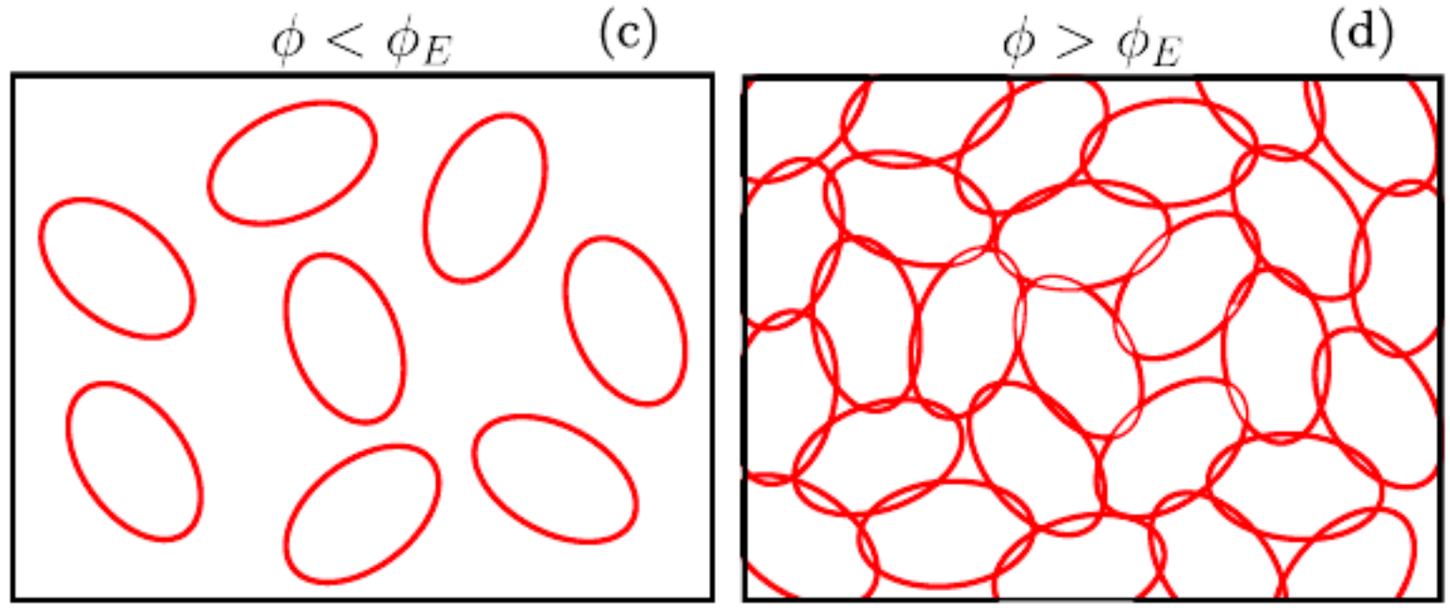


汎化誤差に見られるピークの場所で**相転移**

粒子系の**ジャミング転移**として理解可能

M. Geiger, S. Spigler, S. d'Ascoli, L. Sagun, M. Baity-Jesi, G. Biroli, and M. Wyart, PRE (2019)

over-parameterized phase　　under-parameterized phase



簡単にglobal minimaに到達　ダイナミクスが凍結

球状粒子の場合

r_{ij} 粒子iと粒子jの距離

R 粒子の直径

$r_{ij} > R$ 相互作用なし

$r_{ij} < R$ 斥力相互作用

$$V_{ij} = \max\{0, R - r_{ij}\}^2$$

ヒンジ損失

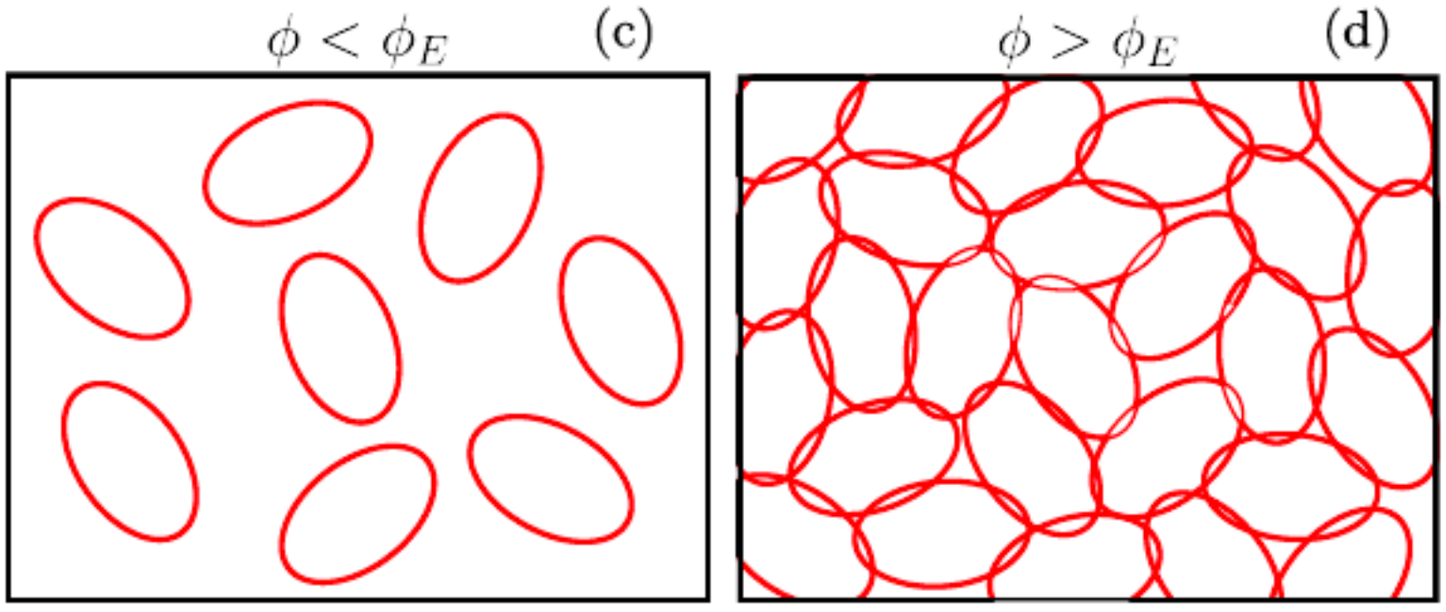
$$\ell(f(x^{(\mu)}, w); y^{(\mu)}) = \max\left\{0, 1 - y^{(\mu)} f(x^{(\mu)}, w)\right\}^2$$

粒子数 = データのサンプル数

ダイナミクスの変化

M. Geiger, S. Spigler, S. d'Ascoli, L. Sagun, M. Baity-Jesi, G. Biroli, and M. Wyart, PRE (2019)

over-parameterized phase　　under-parameterized phase



液体的　　ガラス的

（簡単にglobal minimaに到達）（ダイナミクスが凍結）

過剰パラメータ領域では実際に損失関数をほぼ0にできる

過剰パラメータ領域での深層学習のダイナミクスはガラス的**ではない**

M. Baity-Jesi, L. Sagun, M. Geiger, S. Spigler, G. Ben Arous, C. Cammarota, Y. LeCun, M. Wyart, and G. Biroli (2017)

幅無限大の極限：Neural Tangent Kernel (NTK)

最近のDeep Learningの理論的進展の一つ　　cf. 第4回 唐木田さんの講演

A. Jacot, F. Gabriel, C. Hongler, "*Neural tangent kernel: Convergence and generalization in neural networks*", NeurIPS 2018

隠れ層の幅 H と学習率 η を $H\eta = \text{const.}$ で $H \to \infty$ の極限　　$w_{t+1} = w_t - \eta \nabla_w L(w_t)$

損失関数：平均2乗誤差 $\ell_\mu = [y^{(\mu)} - f(x^{(\mu)}, w)]^2$

ネットワークはランダムに初期化 $w \sim \mathcal{N}(0, H^{-1})$

重みwは訓練の間初期値w_0からほとんど変化しない

⇒ 線形最適化問題に帰着　$f(x, w) \approx f(x, w_0) + [\nabla_w f(x, w_0)] \cdot (w - w_0)$

物理とのアナロジー：ネットワークを熱浴と見立てる　熱浴を調和振動子の集まりと考える=線形近似
個々の調和振動子はブラウン粒子の影響をほとんど受けない（パラメータに変化なし）

無限に多くのランダムな特徴量を使った線形回帰　$\nabla_w f(x, w_0) \in \mathbb{R}^{N_{\text{para}}}$

カーネル関数 $K_{\text{NTK}}(x, x') = \left[\nabla_w f(x, w_0)\right]^{\mathrm{T}} \left[\nabla_w f(x', w_0)\right]$ を用いたカーネル法と等価

（Neural Tangent Kernel, NTK）

ダイナミクスは線形：Lossのglobal minimumにたどり着ける

ネットワークの深さの役割

なぜ深いネットワークが良い汎化性能を示すかは**理論的に不明**

深いネットワークの優越性：表現能力

- 隠れ層を一つ以上持つニューラルネットワークは隠れ層の幅を十分広くすれば任意の連続関数を実現可能（**万能近似定理**）
- 深いネットワークほどより少ないパラメータで同じ表現能力

B. Poole et al.,“*Exponential expressivity in deep neural networks through transient chaos*”, NIPS 2016

表現能力の高さは本当に重要か？

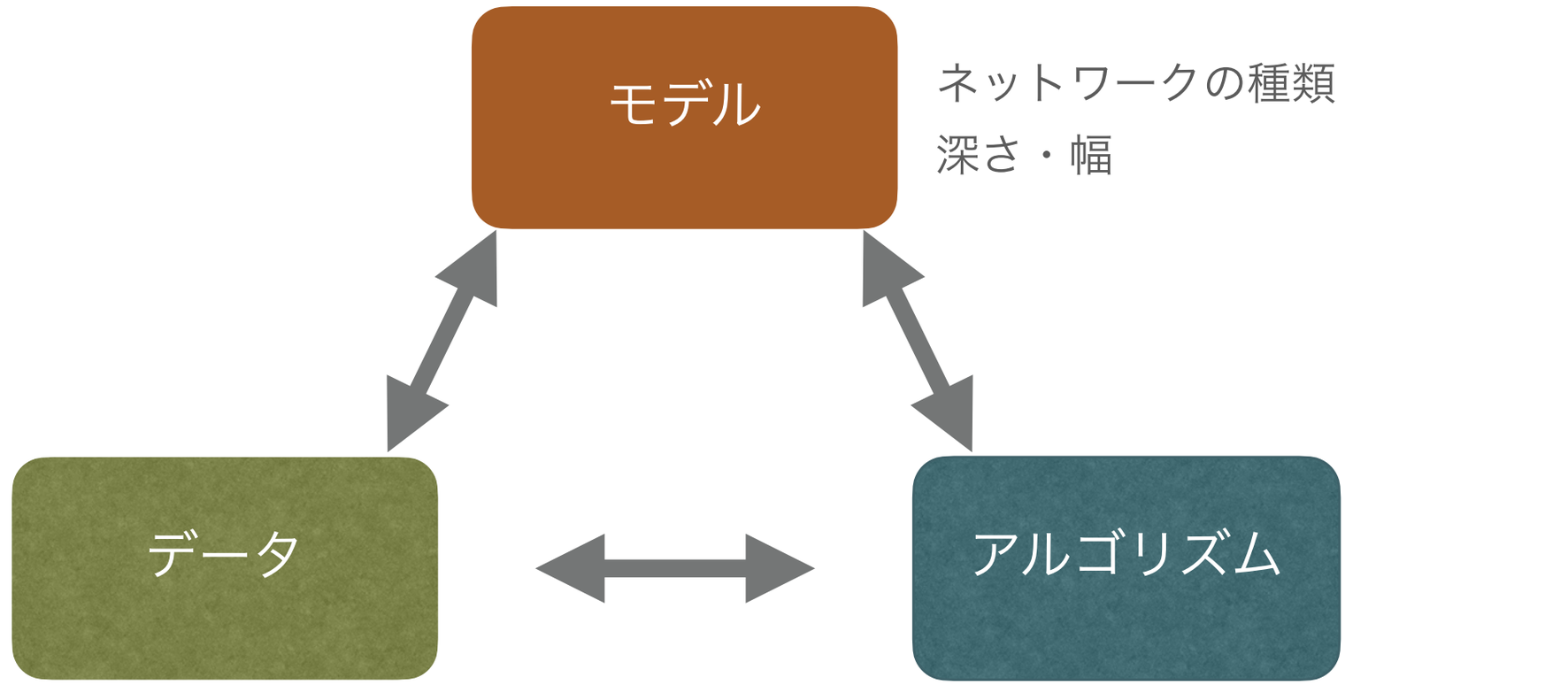
- 現実の分類問題を解くのにそれほど複雑な関数が必要か？

R. Eldan and O. Shamir,“*The Power of Depth for Feedforward Neural Networks*”, PMLR 2016

- 深いネットワークが複雑な関数を実現可能だとしてもそれを学習可能とは限らない

E. Malach and S. Shalev-Shwartz,“*Is Deeper Better only when Shallow is Good?*”, NeurIPS 2019

過剰パラメータ領域での汎化の理解に向けて



データに隠れた低次元多様体

多様体仮説：入力データ$x \in \mathbb{R}^D$は次元$d_{\text{eff}} \ll D$の多様体の近くに集中している

cf. 第7回 本武陽一さんの講演

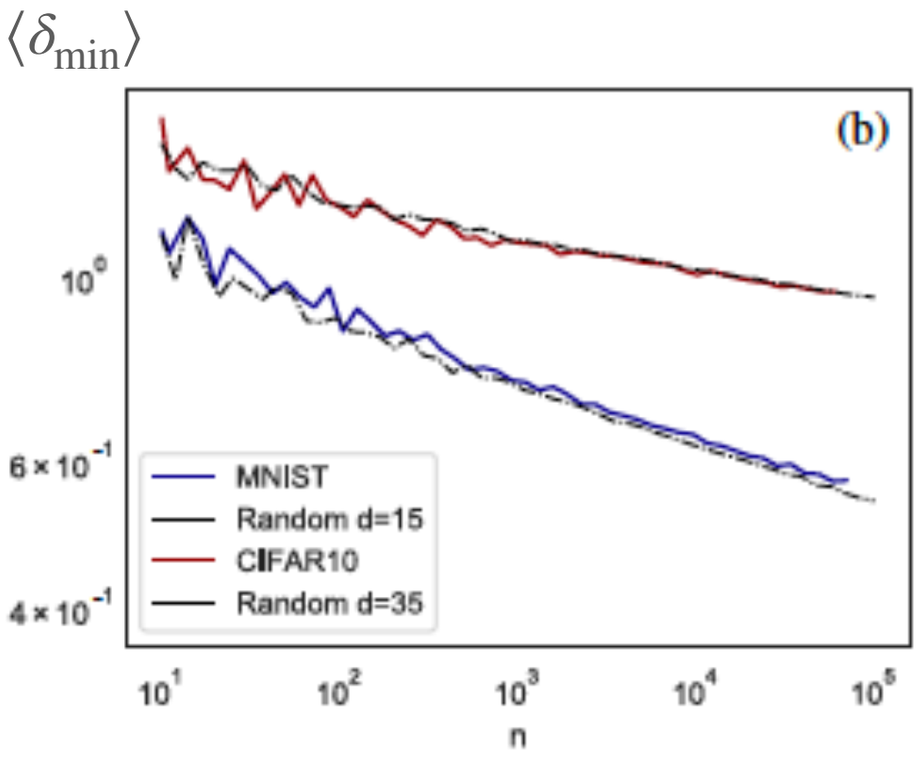
データ多様体の次元d_{eff}の見積もり

n個のデータ点をランダムに選ぶ

あるサンプルから最も近いサンプルの距離は典型的に

$$\delta_{\min} \sim n^{-1/d_{\text{eff}}}$$

距離r以内に$r^d n$に比例した数のサンプル



S. Spigler, M. Geiger, and M. Wyart, arXiv:1905.10843

MNIST $D = 784, \quad d_{\text{eff}} \simeq 15$

Cifar-10 $D = 3072, \quad d_{\text{eff}} \simeq 35$

入力データ$x \in \mathbb{R}^D$の局所性や対称性がd_{eff}を小さくする

統計力学的アプローチ

teacher-student scenario　入力$x \in \mathbb{R}^D$　ラベルyは実数

教師ネットワーク：$y = f(w^*, x) + \varepsilon$　w^*「真の」パラメータ

生徒：教師ネットワークに近い出力を返すように学習 $\hat{y} = f(w, x)$

入力$x \in \mathbb{R}^D$は各要素が独立なランダム変数

教師ネットワークのパラメータw^*もランダムに固定

データに構造がない！

hidden-manifold model　S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, arXiv:1909.11500

データが低次元多様体に分布することを取り入れた簡単な設定

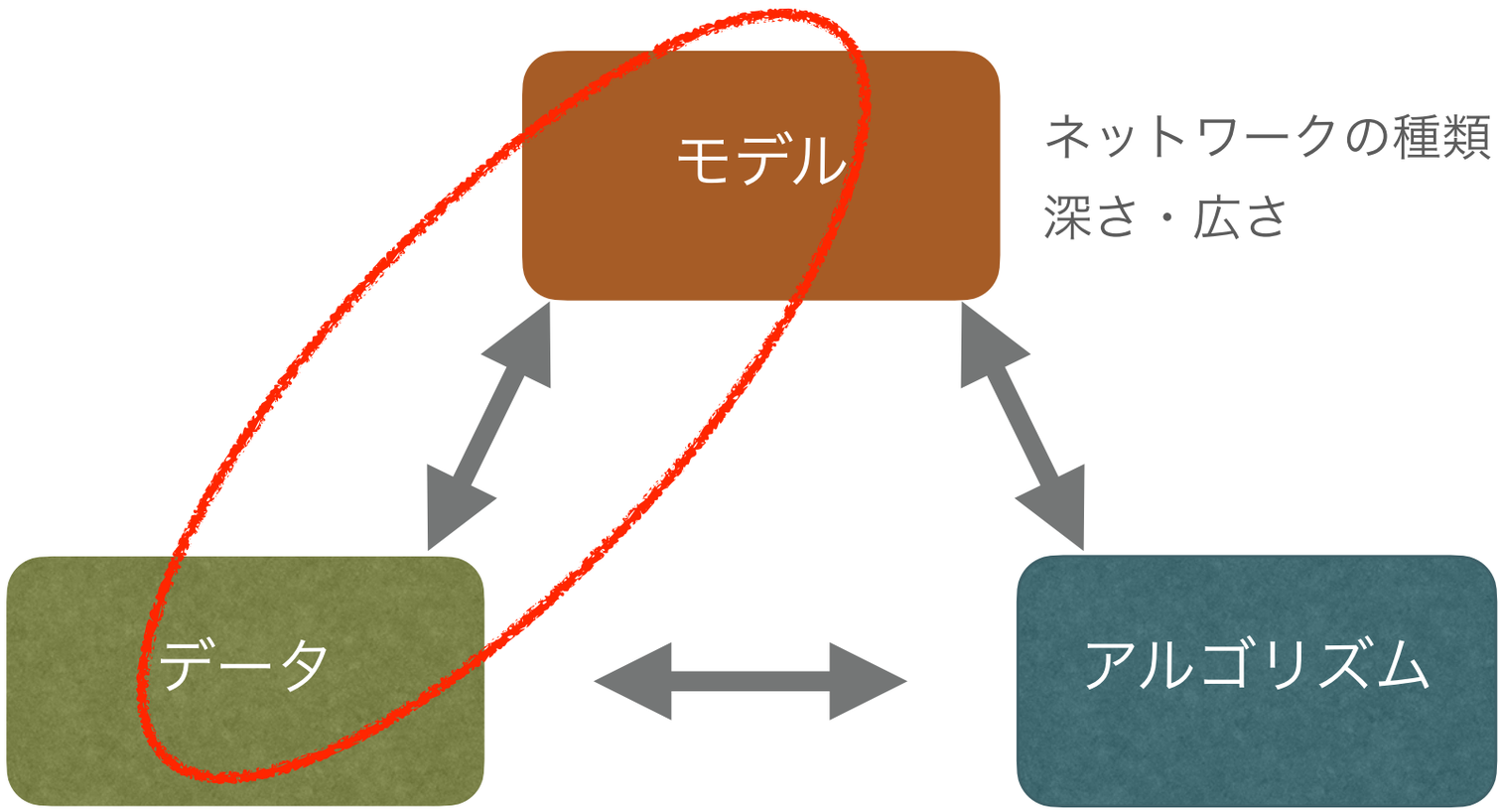
$$x^{(\mu)} = \sigma\left(\frac{1}{\sqrt{d_{\text{eff}}}}\hat{F}c^{(\mu)}\right)$$

$\hat{F} \in \mathbb{R}^{D \times d_{\text{eff}}}$ ランダムに固定

$c^{(\mu)} \in \mathbb{R}^{d_{\text{eff}}}$ i.i.d. ランダムベクトル

$\sigma(\cdot)$ 非線形関数

過剰パラメータ領域での汎化の理解に向けて



有効次元
対称性・局所性

ダイナミクス
(SGD: learning rate, minibatch size)

研究の目的

データ構造のどういう要素が**深さ**に利点をもたらすか

現実のデータは複雑すぎる：簡単な構造を持つ**人工データ**を使う

分類問題では入力の構造だけでなく**入力と分類ラベルの関係**が重要

分類ラベルは入力データのどの特徴に着目するかを決める

熱統計力学：

$N = 10^{23}$個の粒子の座標と運動量の組がデータとして与えられるが

マクロ物理量だけに着目することで普遍的な体系を得る

特に平衡状態に限定すると，系のエネルギー，粒子数，体積だけですべてが決まる

分類ラベルの局所性に注目

分類ラベルの局所性

訓練データ　$\mathscr{D} = \{(x^{(\mu)}, y^{(\mu)}) : \mu = 1,2,\ldots N\}$

$x^{(\mu)} = (x_1^{(\mu)}, x_2^{(\mu)}, \ldots, x_D^{(\mu)})^{\mathrm{T}}$: i.i.d. ガウシアンランダムベクトル

入力データには何の構造もない

$y^{(\mu)} = \mathrm{sgn}(M)$: ここで特徴量 M は $x^{(\mu)}$ と簡単な関数関係

$1 \le i_1 < i_2 < \ldots < i_k \le D$を一つ固定

k-localラベル

$$M = x_{i_1}^{(\mu)} x_{i_2}^{(\mu)} \ldots x_{i_k}^{(\mu)}$$

ミクロな相関関数

k-globalラベル

$$M = \sum_{j=1}^{D} x_{j+i_1}^{(\mu)} x_{j+i_2}^{(\mu)} \ldots x_{j+i_k}^{(\mu)} \qquad x_{i+D}^{(\mu)} = x_i^{(\mu)}$$

マクロな熱力学量

学習の詳細

損失関数 $L(w) = \frac{1}{N}\sum_{\mu=1}^{N} \ell_\mu(w)$　　$\ell_\mu(w) = \ell(f(x^{(\mu)}, w); y^{(\mu)})$ cross-entropy loss

確率的勾配法(SGD)　　$w_{t+1} = -\eta \nabla_w L_{\mathcal{B}_t}(w)$　　$L_{\mathcal{B}}(w) = \frac{1}{B}\sum_{\mu\in\mathcal{B}} \ell_\mu$

- 訓練データ全体をサイズ B のミニバッチ $\mathcal{B}$ にランダムに分割
- 各ステップでは一つのミニバッチの損失関数の勾配を計算
- 訓練データの誤答率が完全に0%になるまで学習させる

学習率 η を最適化：10-fold cross validation （交差検証）

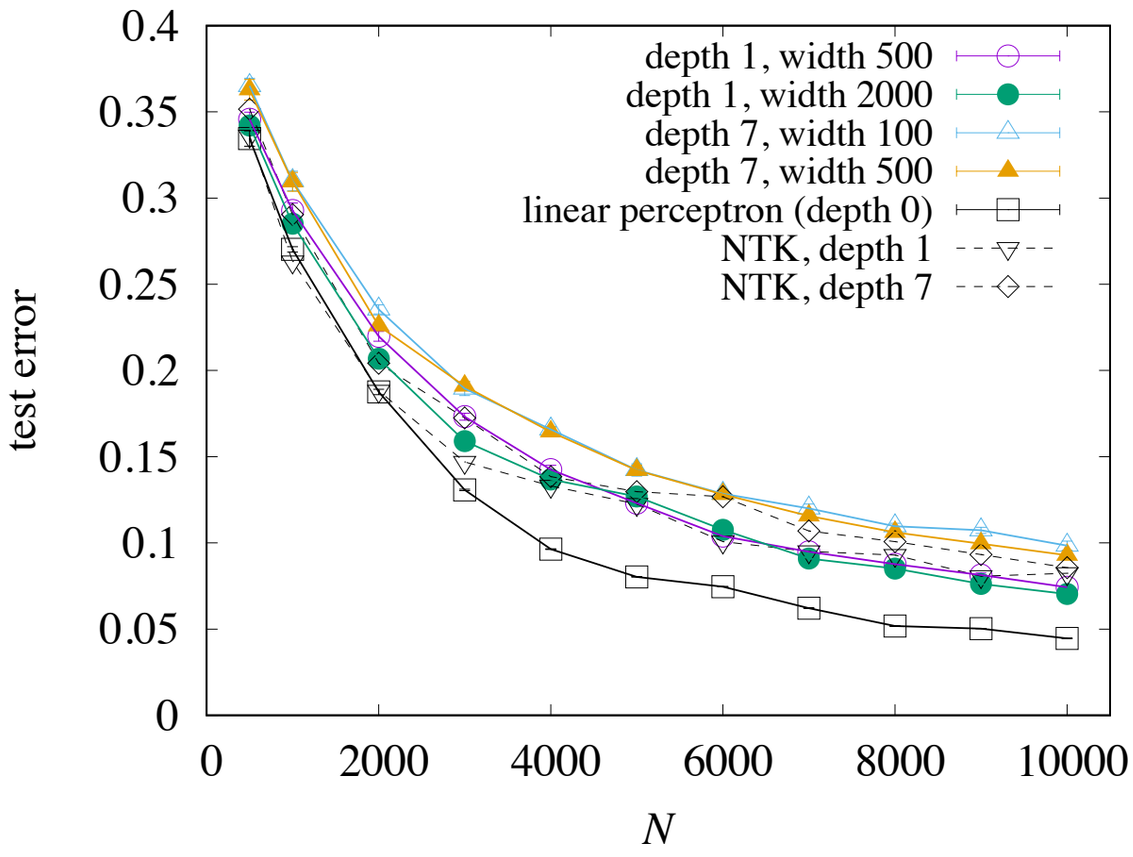
- データを10分割：一つを検証用として残りを訓練データとして使用
- 学習後の検証誤差（検証データの誤答率）を計算
- これを検証用データを選ぶ10通りの仕方について繰り返す
- 平均検証誤差が小さくなる学習率をベイズ最適化の手法で取得

すべての訓練データを使って最適な学習率のSGDで訓練

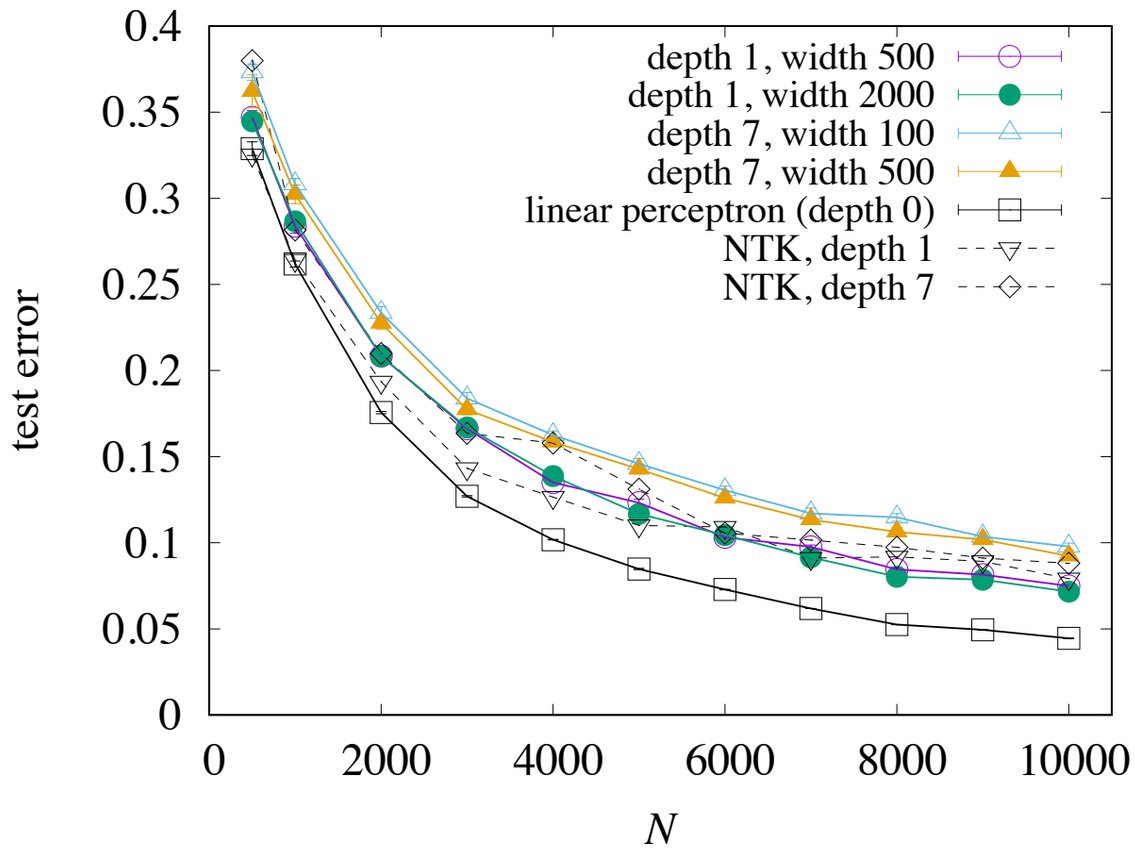
10回繰り返してテスト誤差（テストデータの誤答率）の平均と標準偏差を計算）

$k = 1$: localかglobalかで違いはない

1-local ($D = 1000$) $M = x_i^{(\mu)}$



1-global ($D = 1000$) $M = \sum_{i=1}^{N} x_i^{(\mu)}$

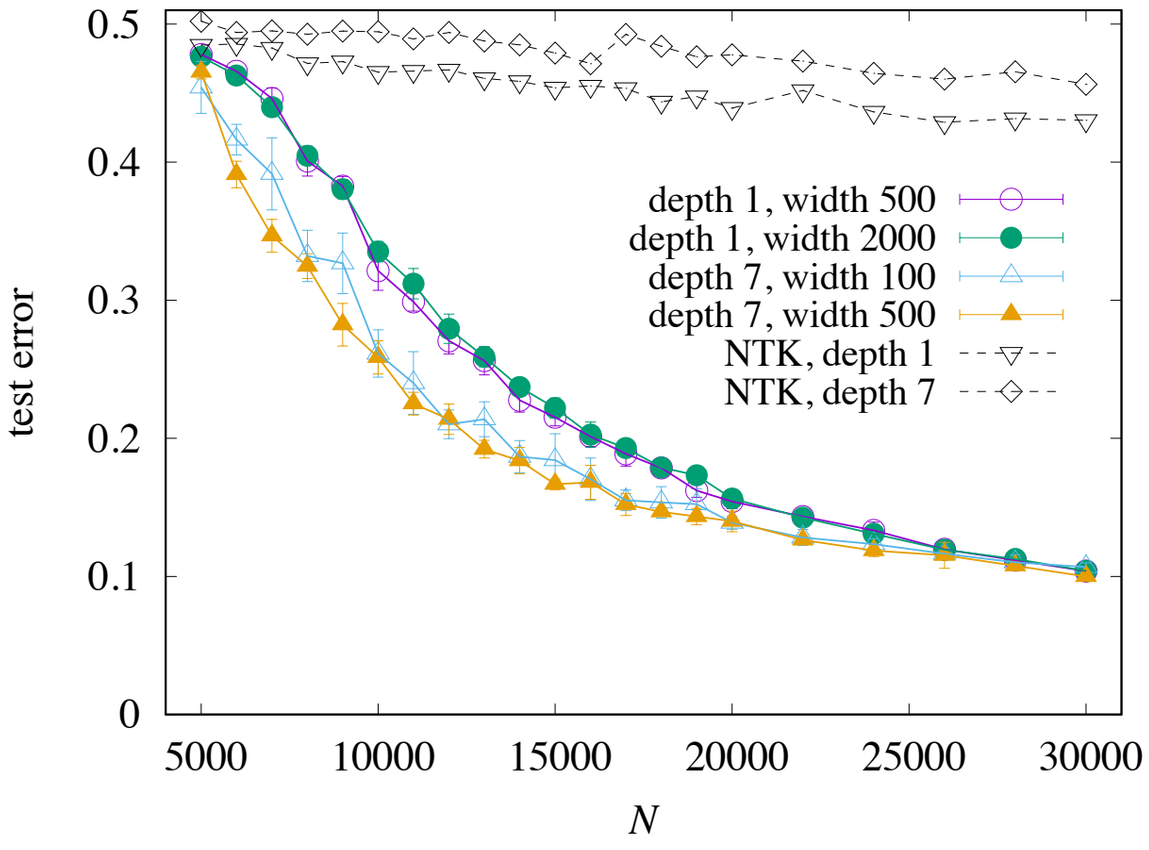


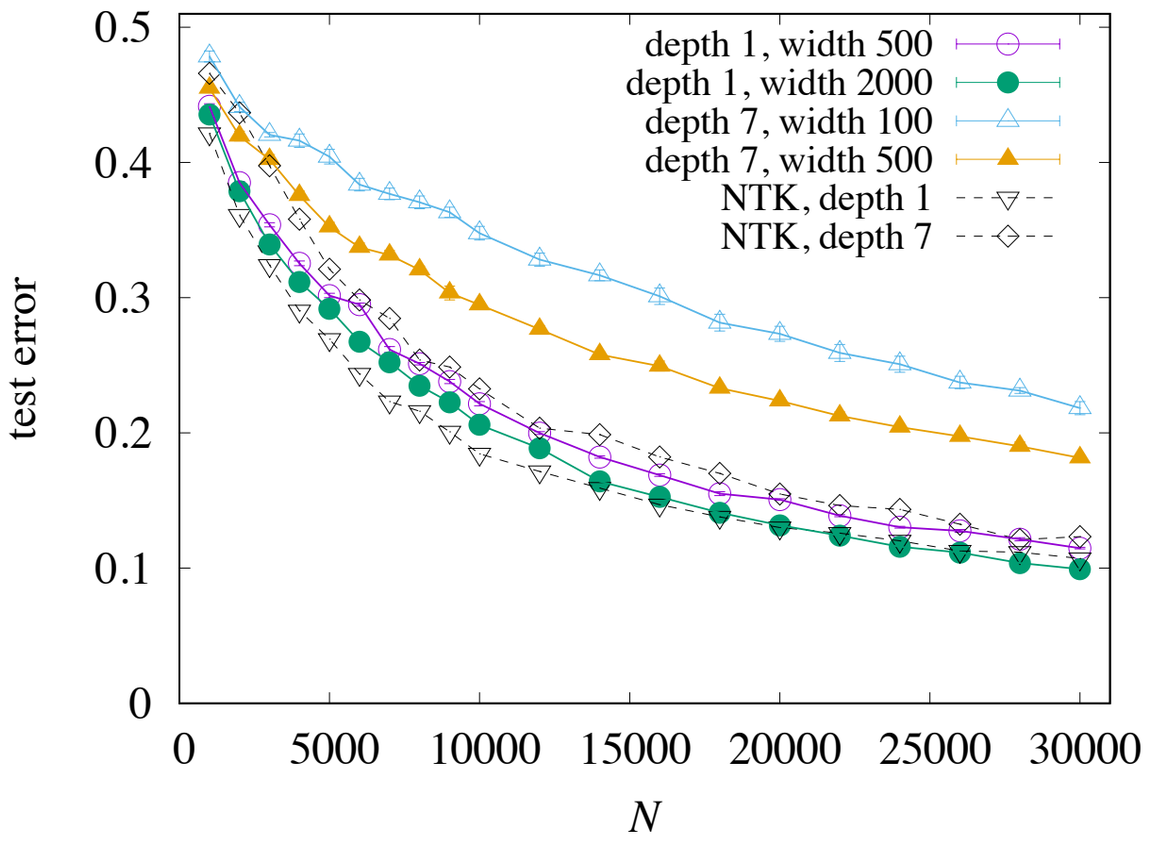
- 定量的にもほぼ同じ結果
  （1-localと1-globalはフーリエ変換の関係にあることから自然）
- 浅いネットワークの方が深いネットワークより良い

$k \geq 2$: localかglobalかで逆の深さ依存性

2-local ($D = 500$) $M = x^{(\mu)}_{i_1} x^{(\mu)}_{i_2}$

2-global ($D = 100$) $M = \sum_{j=1}^{N} x^{(\mu)}_{j+i_1} x^{(\mu)}_{j+i_2}$



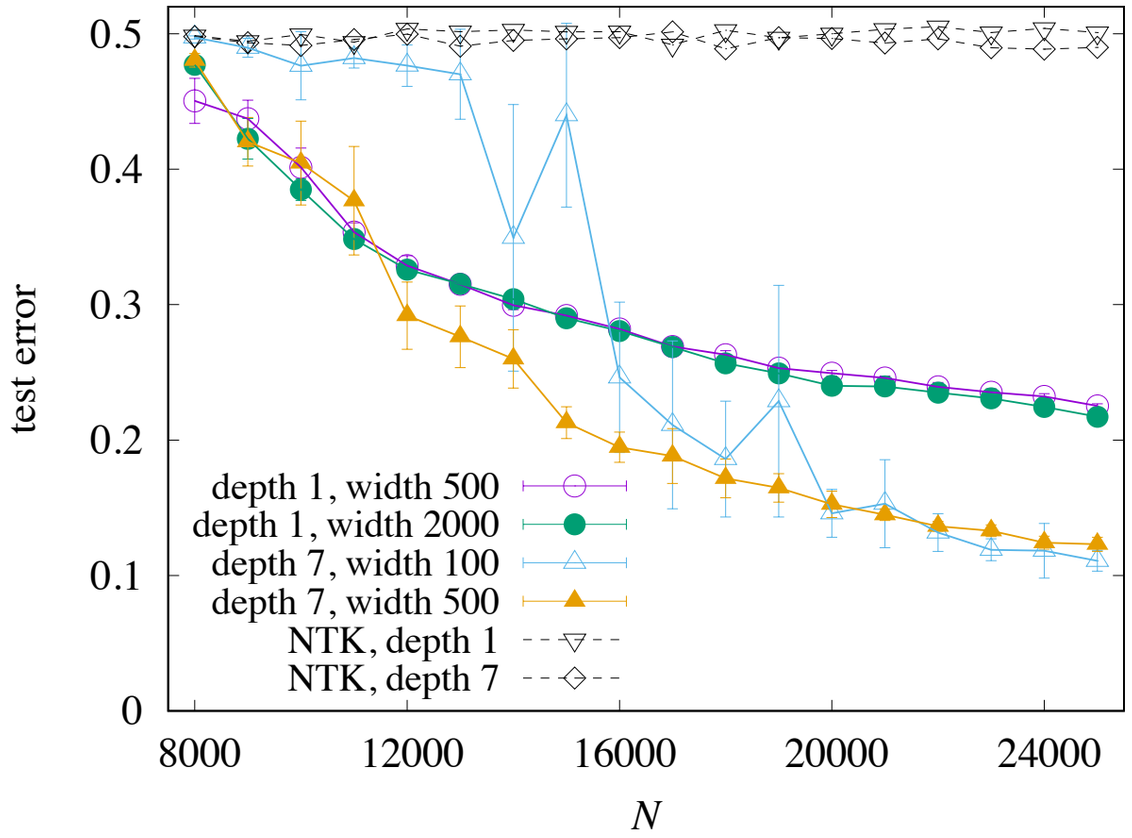


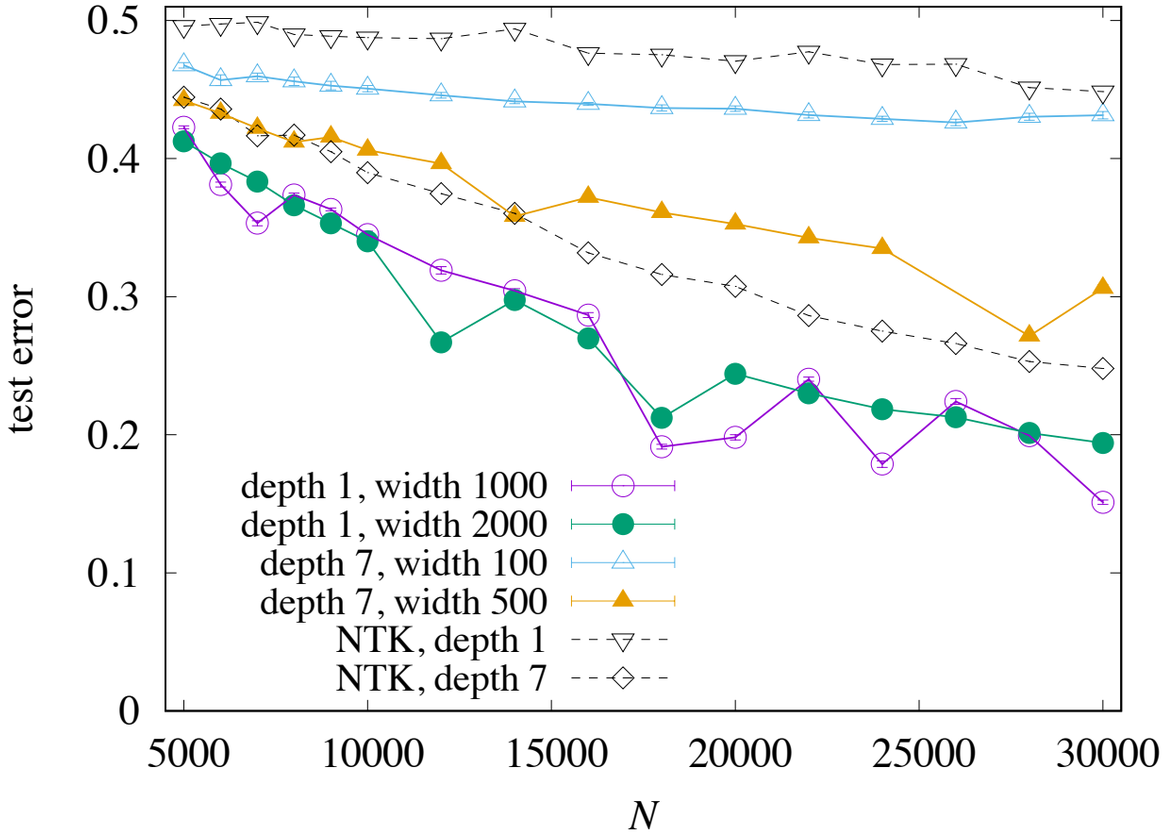
- 2-localラベル：**深い**ネットワークの方良い汎化性能
- 2-globalラベル：**浅い**ネットワークの方が良い汎化性能

$k \geq 2$: localかglobalかで逆の深さ依存性

3-local ($D = 100$) $M = x_{i_1}^{(\mu)} x_{i_2}^{(\mu)} x_{i_3}^{(\mu)}$

3-global ($D = 40$) $M = \sum_{j=1}^{N} x_{j+i_1}^{(\mu)} x_{j+i_2}^{(\mu)} x_{j+i_3}^{(\mu)}$



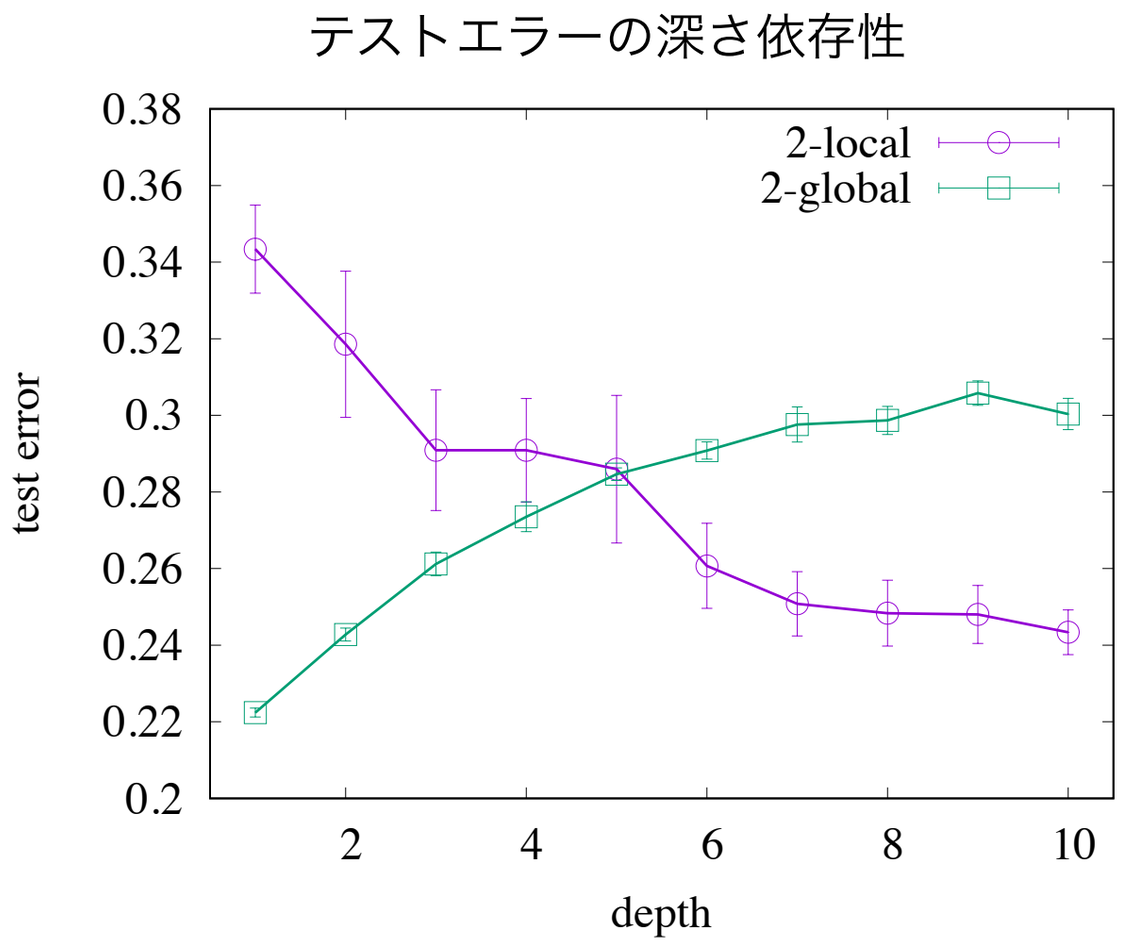


- 3-localラベル：**深い**ネットワークの方が良い汎化性能
- 3-globalラベル：**浅い**ネットワークの方が良い汎化性能

結論：特徴量の局所性が重要

k-localラベルでは深い方がよいがk-globalラベルでは浅い方がよい

常に深い方が良いというわけではない

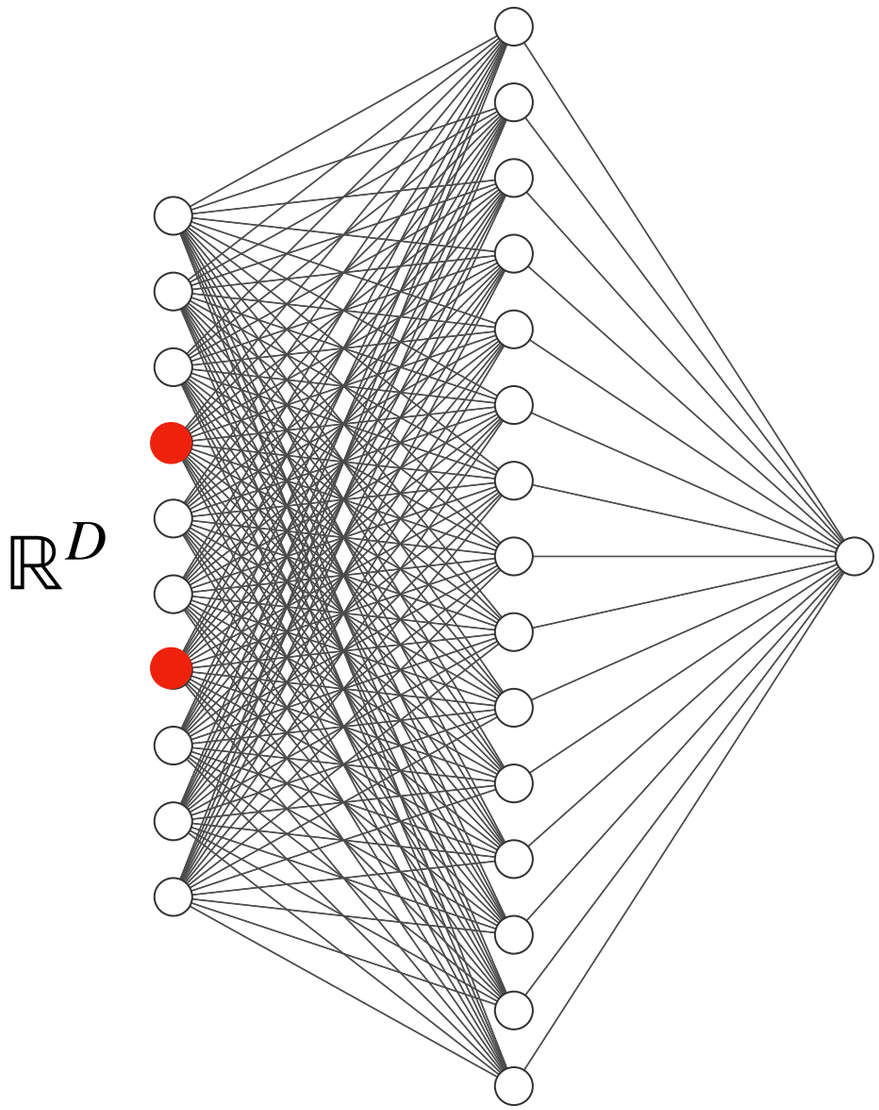


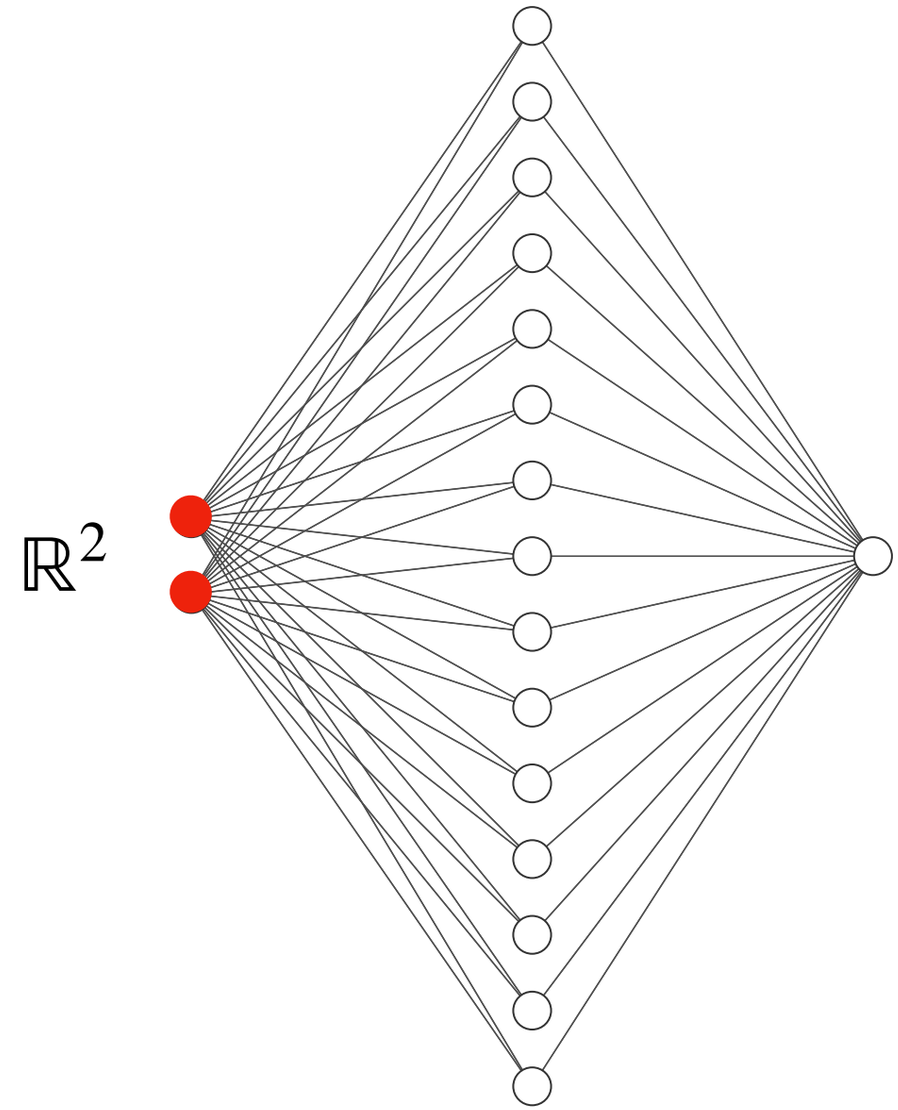
深さの利点は表現能力の高さによるものではない

表現能力の高さはネットワークの深さと密接に関係

しかしそのことはk-localラベルでの深さの利点と関係がない

2-localラベル：広さ5〜10の隠れ層が一つあれば十分正確に表現可能





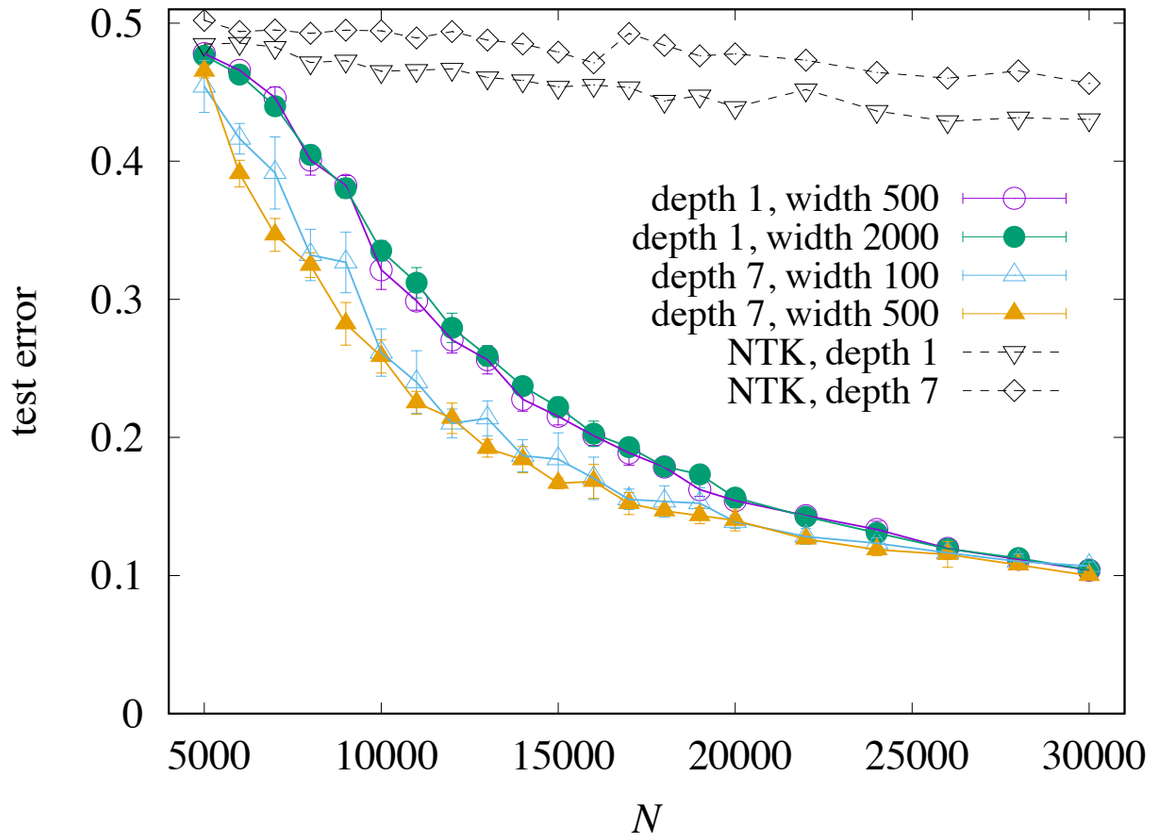
NTKは幅無限大の極限の理論？

NTKではほとんど汎化しない
ネットワークを広くしていってもNTKの結果に漸近しない

NTKの理論は隠れ層の幅Hを大きくするのと同時に学習率ηを小さくすることで得られる

$$H \to \infty \text{ with } H\eta \text{ fixed}$$

2-localラベルの結果



この極限ではネットワークの重みwは初期値からほぼ変化しない
データの特徴を抽出しているのではなく無限個のランダムな特徴量を使って分類問題を解いている

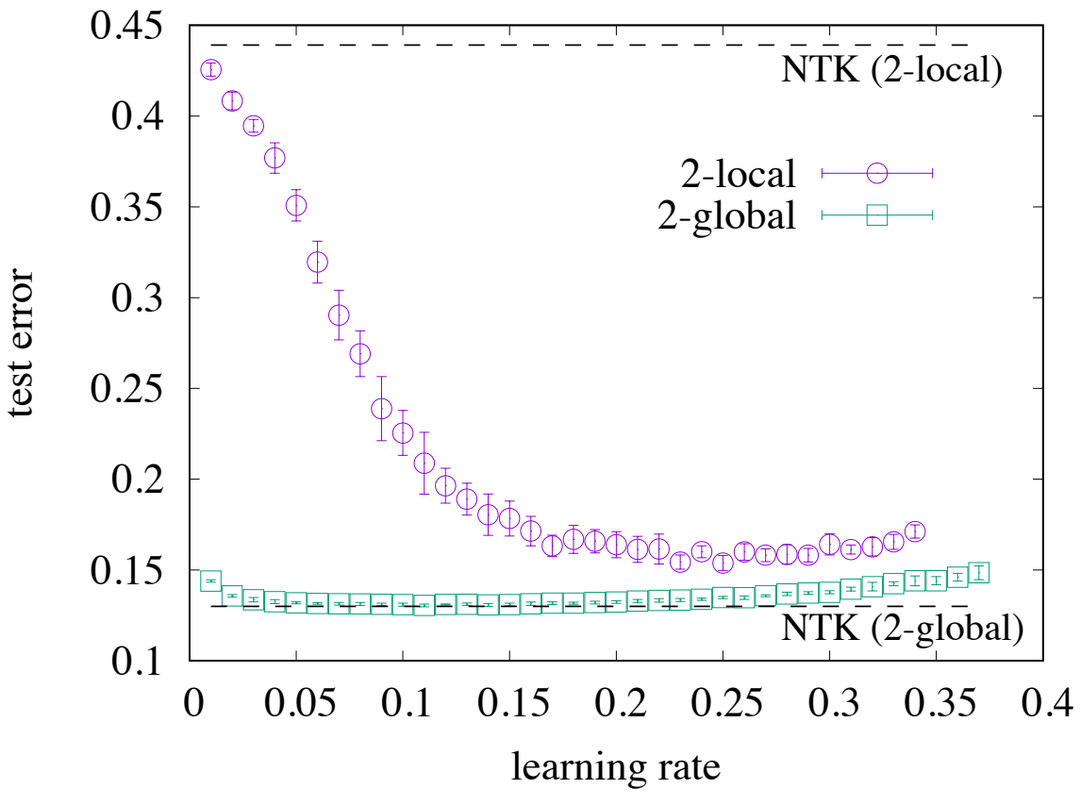
lazy learning ⇔ *feature learning*

交差検証によって最適化された学習率はfeature learning regime

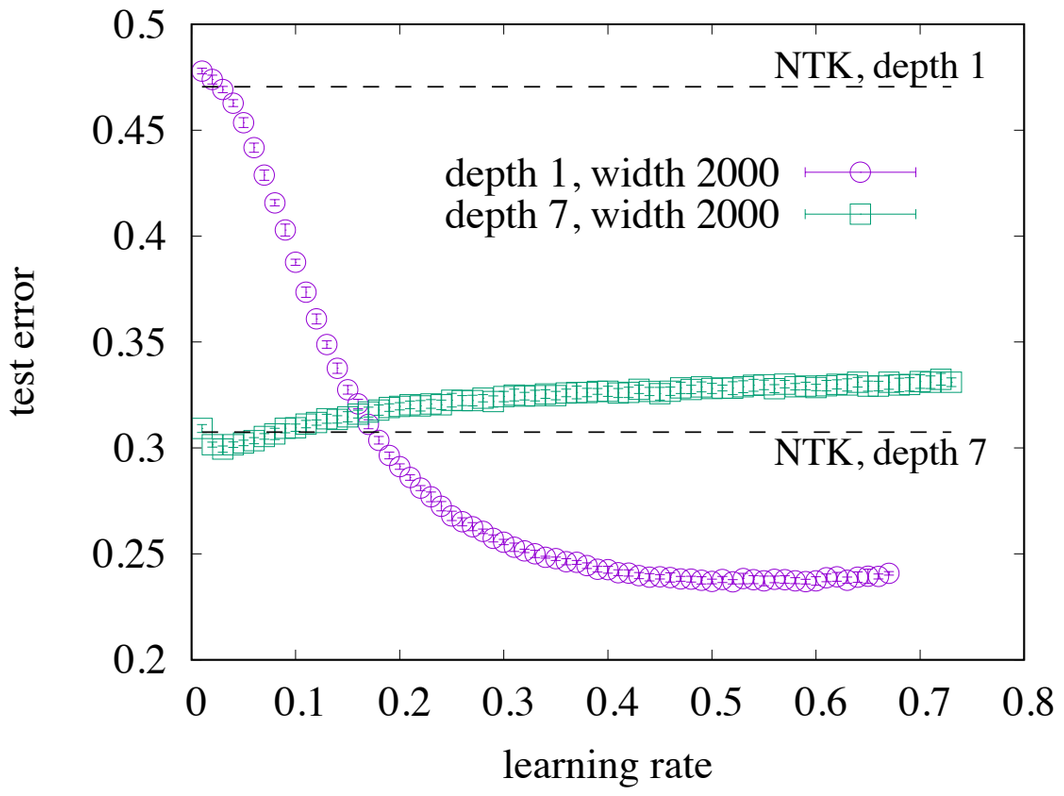
lazy learningではなくfeature learningが重要

テストエラーの学習率依存性

2-local and 2-global



3-global label



- 学習率を0に近づけるとNTKの結果に近づく
- 最適な学習率でのテストエラーは隠れ層の幅を広くしても必ずしもNTKの結果に漸近しない
- 3-globalラベルの場合：学習率が0に近い領域では深いネットワークの方が良い結果だが最適な学習率では浅いネットワークの方が良い結果

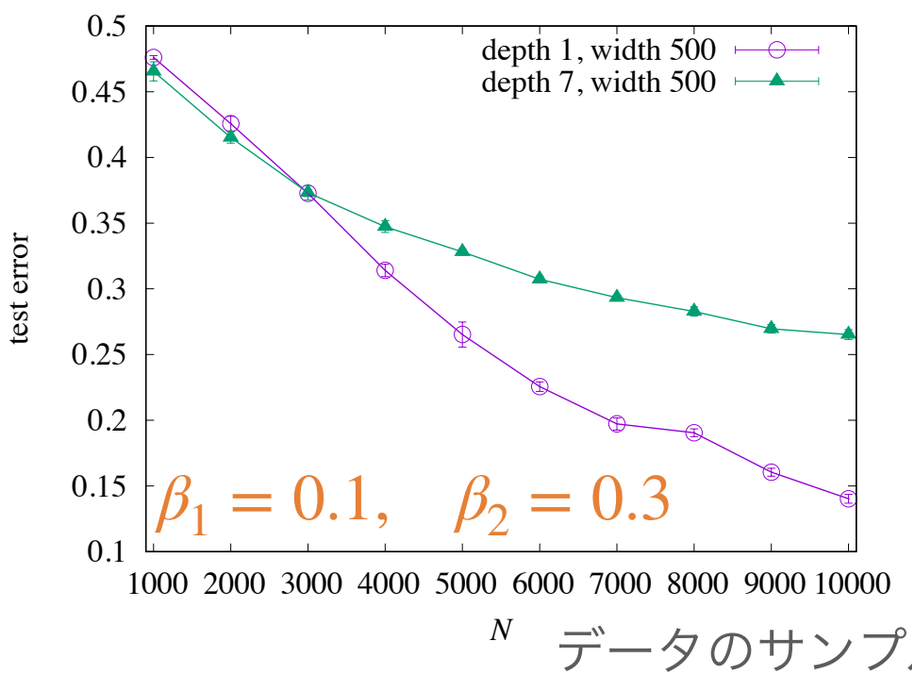
温度の推定

ガウシアンモデル　$P_\beta(x) = e^{\beta \sum_{i=1}^{N} x_i x_{i+1} - \frac{1}{2} x_i^2}$　　$\beta = \frac{1}{k_B T}$ 逆温度　　$|\beta| < \frac{1}{2}$

$x^{(\mu)} = (x_1^{(\mu)}, x_2^{(\mu)}, \ldots, x_D^{(\mu)})^{\mathrm{T}}$: i.i.d. ガウシアンランダムベクトル　温度無限大に相当

与えられたサンプルxが二つの逆温度β_1, β_2のどちらから生成されたかを予測する機械

温度の情報は熱力学量に入ってくる：エネルギー$H = -\sum_{i=1}^{N} x_i x_{i+1}$



エネルギーは2-globalラベル：浅い方が有利？

数値結果：浅いネットワークの方が良い結果

まとめ T. Mori and M. Ueda, "*Is deeper better? It depends on locality of relevant features*", arXiv:2005.12488

- ネットワークの深さがどういうときに有利に働くかを明らかにするために，簡単な構造を持つデータを用いた分類問題でネットワークの汎化性能を調べた
- 分類ラベルの局所性に注目し，k-localラベルとk-globalラベルを導入した
- k-localラベルでは深いネットワークの方が良い汎化性能
- k-globalラベルでは浅いネットワークの方が良い汎化性能
- ここで見られた汎化性能の深さ依存性は表現能力の高低によるものではない
- 深さ依存性の理解にはfeature learning regimeを考えるべき