深層学習の数理：
統計力学的アプローチ

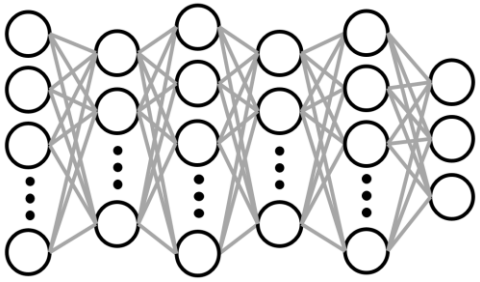
産業技術総合研究所 人工知能研究センター
機械学習研究チーム研究員

唐木田 亮

Deep learning and Physics 2020
June. 18

目次

背景: 深層学習



Deep Network Network (DNN):
Fully-connected, CNN,
ResNet, Transformer …

使いやすい訓練法: SGD, Adam, K-FAC …
Dropout, Batch Normalization …

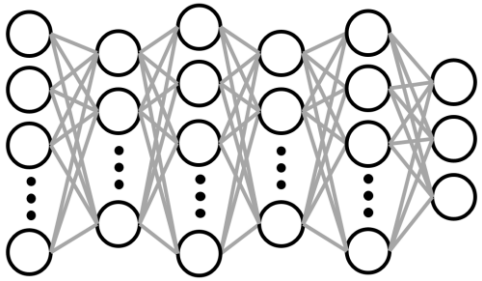
+ Adversarial attacks & defences, 解釈性,
contrastive learning …

深層生成モデル:
VAE, GAN, Flow-based …

既存の機械学習手法との**融合**:
強化学習, カーネル法 …

DNNを基盤とした機械学習の集合

背景: 深層学習



Deep Network Network (DNN):
Fully-connected, CNN,
ResNet, Transformer …

自然な疑問:
深層ニューラルネットワークはなぜ/どのような設定で
性能が高いのか

“手持ちのデータで性能が出ない”

原因は, データ？正規化？層数？アルゴリズム？ステップ数？…

数理の必要性. 近年発展が著しい. 大きく分けると,

表現能力 (Expressivity/Representation power),
訓練性 (Trainability),
汎化能力 (Generalization)

表現能力: 次元の呪いと階層性

- **万有近似性** (Universal Approximation) [1990年代～]
  3層(shallow)ニューラルネットは入力を任意の精度で近似可能

- **Barronの定理** (1993)

(ある特定のクラスの) 関数$f(x)$ に対して,

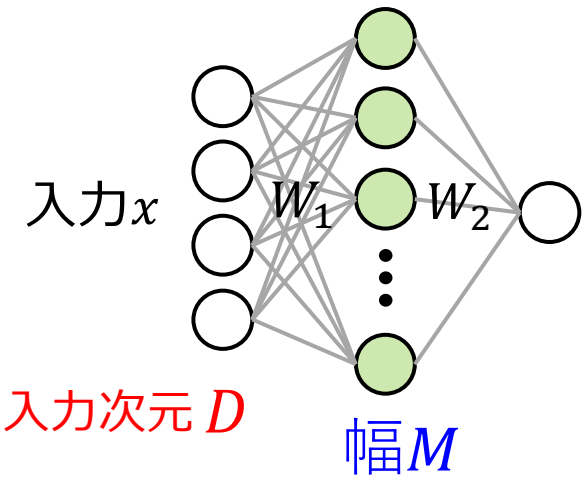
$$\mathrm{E}_x[(f(x) - f_M(x))^2] \leq C_f/M$$

を満たすW_1, W_2が存在

一方, 基底(W_1)を学習しない場合,

$$\mathrm{E}_x[(f(x) - f_M(x))^2] \geq C'_f/M^{2/D}$$

$$f_M(x) = W_2\mathrm{Sigmoid}(\mathrm{W_1x})$$



階層性が近似精度における**次元の呪いを克服**. 3層で十分？

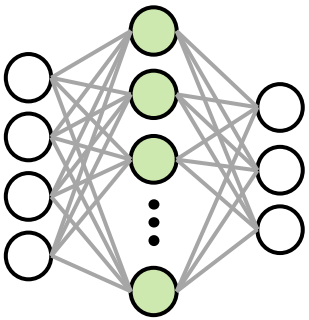
表現能力

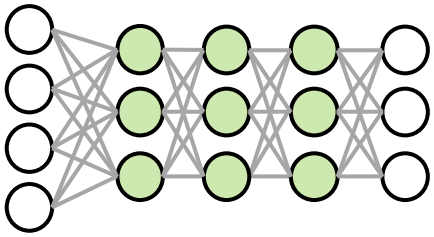
Deep vs. Shallow

非線形性や入力の種類(位相)に応じた様々な理論

- The number of *monomials* [Delalleau, Bengio, NIPS2011]
- The number of *linear regions* [Montufar+, NIPS2014]
- *Betti numbers* [Bianchini, Scarselli, IEEE (2014)]　etc…
- Barronの定理の拡張 [Lee+ COLT2017]

非線形変換





階層型ニューラルネットが表現できる関数の
複雑さは,
幅に対してベキ的に増加,
層数に対して**指数関数的に増加**

有限の計算資源(メモリ)では, 深層モデルの方が効率的

訓練性の問題

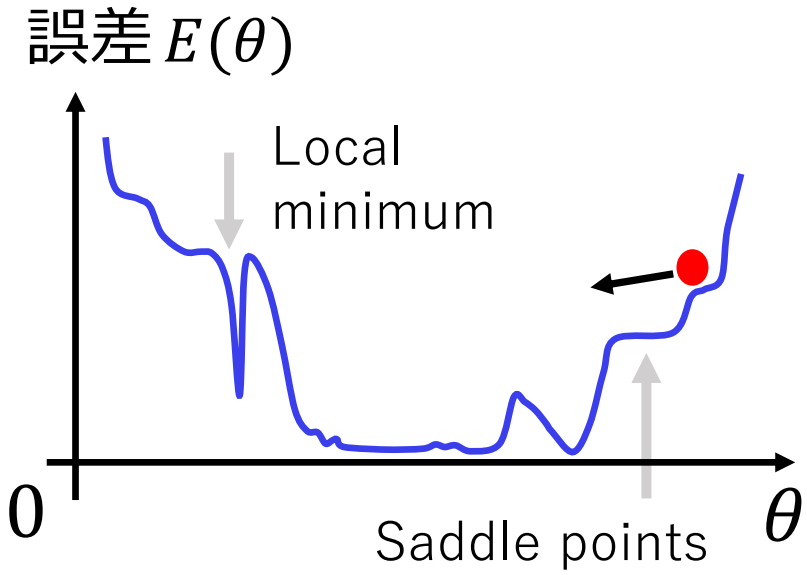
表現能力が高い ≠ 学習させやすい

訓練データ $\{\boldsymbol{x}^{(i)}, y^{(i)}\}$ $(i = 1, \ldots, \mathrm{n})$

$$E(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - f_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right)\right)^2$$

最急勾配法 (Backpropagation)

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta\frac{\partial E(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$



- 一般には, global minimaにたどり着く保証はない
- DNNは勾配が消失 (or 発散)しやすい

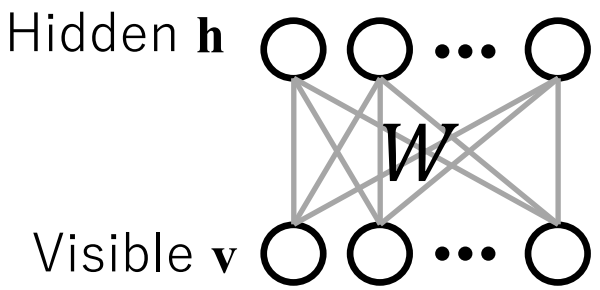
$$D_l W_l^\top D_{l+1} W_{l+1}^\top \cdots W_L^\top (y - f)$$

訓練性とloss Landscape

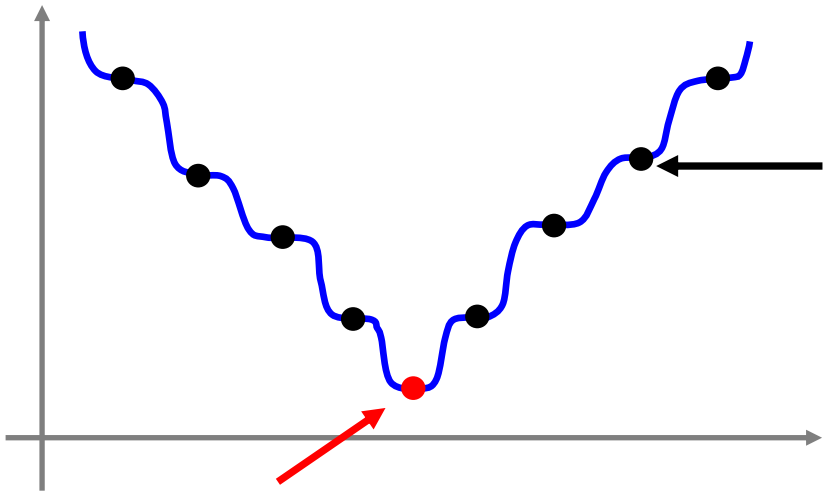
[RK, Okada, Amari, Neural Networks (2016)]

統計モデル

$$p(\mathbf{h}, \mathbf{v}) = \exp(\mathbf{h}^T W \mathbf{v} - ||\mathbf{h}||^2/2 - ||\mathbf{v}||^2/2)/Z$$



負の対数尤度



他の固定点
= saddle points

唯一のlocal minima = global minima

同様のランドスケープはテンソル分解, 線形ネットなど
でも知られる. わかりやすい描像.

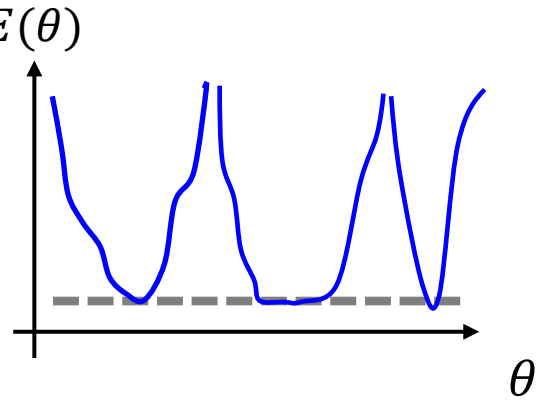
訓練性とloss landscape

・いくつかの描像

- DNNの**幅**が十分に大きければ, すべての固定点がglobal minima

Extremely wide nets
[Nguyen&Hein, ICML '17&'18] [Gori&Tesi IEEE 1992]

$M(DNNの幅) \geqq T$ (訓練サンプル数)



Sigmoid, softplus (極限としてReLU), CNN 幅広く成立

- モデルだけでなくアルゴリズムも貢献している
SGDがLocal minimaを抜け出す条件 [Kleinberg+ ICML '18]

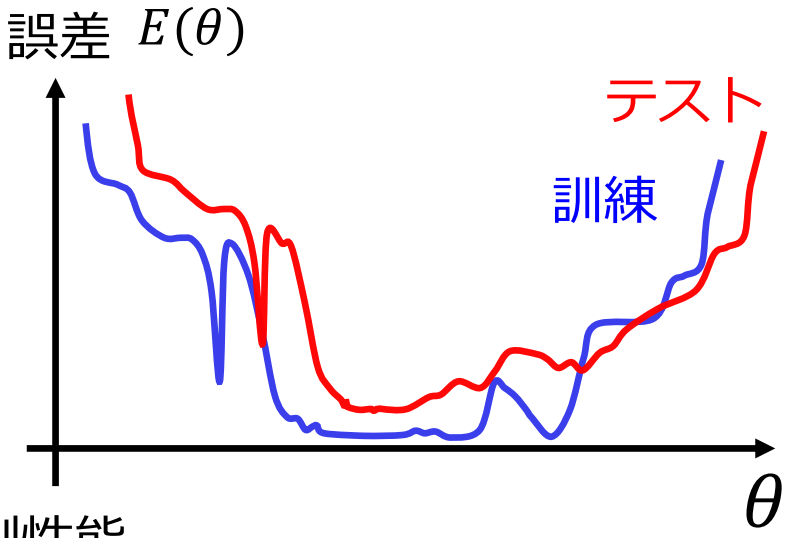
$$\theta_{t+1} = \theta_t - \frac{\partial E}{\partial \theta}(\theta, x_t)$$ Mini-batch由来のノイズが効く

汎化の問題

表現能力が高い ≠ 学習させやすい

訓練データ $\{\boldsymbol{x}^{(i)}, y^{(i)}\}\ (i = 1, \ldots, \mathrm{n})$

$$E(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - f_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right)\right)^2$$



- テストデータは
  訓練データと異なる未知データ

汎化性能: テストデータでの誤差, 予測性能.

DNNの数理の最前線. 汎化性能を測るさまざまな指標の提案.
網羅的な実験検証も行われつつある:

"Fantastic generalization measures and …" [Jiang+ ICLR 2020]

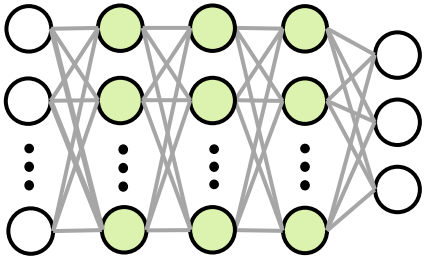
40以上の汎化指標

目次

深層学習の多様性

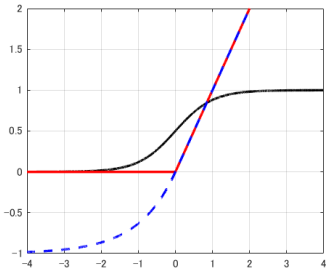
- ネットワーク構造

素子数, 層数, skip connection, 畳み込み …



- 活性化関数

Sigmoid, (Leaky-)ReLU, ELU …



- 勾配法

constant rate, step-wise, Adam, K-FAC …

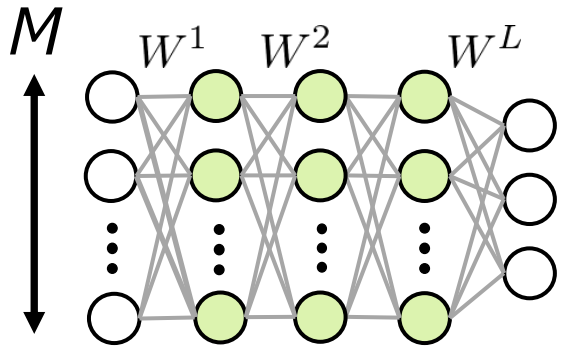
$$\theta_{t+1} = \theta_t - \eta_t \frac{\partial E}{\partial \theta}$$

個別のモデルでの個別の数理 (表現能力, 訓練性, 汎化性能, …)

多様な深層学習に**統一的・普遍的な視点**は与えられないか？

深層学習とランダムネス

- DNN最適化の**初期値**は
  ランダム行列



広く使われている初期値の例:　一様乱数 [Glorot&Bengio AISTATS 2010]

$$W \sim U\Big[-\frac{\sqrt{6}}{\sqrt{n_j+n_{j+1}}}, \frac{\sqrt{6}}{\sqrt{n_j+n_{j+1}}}\Big]$$

ある層の素子数Mに対し, パラメータ分散は$O(1/M)$

$$\sum_j^M W_{ij}x_j \sim O(1)$$

次の層の**出力**が素子数に依存しない.
ほどよいスケールに規格化される

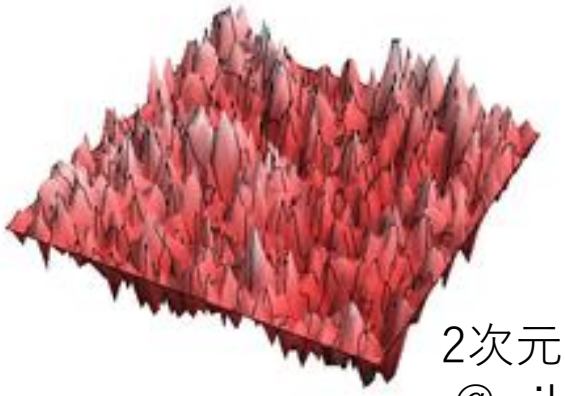
DNNとランダムネスの数理

$L(\boldsymbol{w})$ ：**ランダムガウス場** (多次元ガウス過程)

平均ゼロ, 共分散：

$$\mathbb{E}\left(L\left(\mathbf{w}_1\right)L\left(\mathbf{w}_2\right)\right) = f\left(\left\|\mathbf{w}_1 - \mathbf{w}_2\right\|^2\right)$$

f: 原点まわりで滑らかな関数



2次元グリッド
@wikipedia

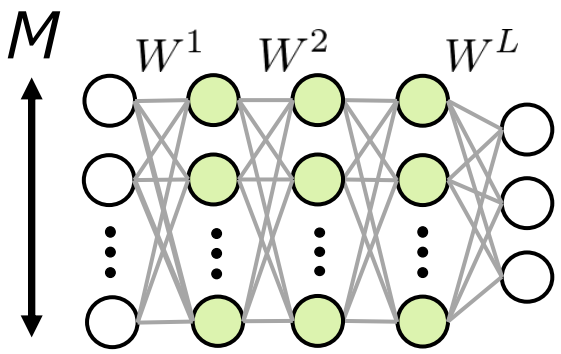
深層ネットの誤差関数ランドスケープとの定性的な類似

- 固定点は鞍点が大半で, local minimaはほとんどglobal minima.
  誤差が大きいほど負の固有値の割合が増加. [Dauphin+ NIPS ‘15]
- 学習中における結合パラメータのノルムの増加はランダムガウス場上のランダムウォークと同じ ($\left\|\mathbf{w}_t - \mathbf{w}_0\right\| \sim \log t$) [Hoffer+ NIPS ‘17]

ランダム深層ニューラルネットの平均場理論

[Amari (1970-)][Poole+ NIPS '16]

$$u_i^l = \sum_j W_{ij}^l h_j^{l-1} + b_i^l, \quad h_i^l = \phi(u_i^l)$$



- **ランダム**パラメータ　　e.g. ガウス分布

$$W_{ij}^l \sim \mathcal{N}(0, \sigma_w^2/M) \quad b_i^l \sim \mathcal{N}(0, \sigma_b^2)$$

- ニューラルネットの典型的な挙動を表す秩序変数の導出

第l層の平均活動度

$$q^l = \sum_i^M (u_i^l)^2/M \quad \xrightarrow{M \gg 1}$$

再帰的計算

$$q^l = \sigma_w^2 \int Dz \phi^2\left(\sqrt{q^{l-1}}z\right) + \sigma_b^2$$

ガウス積分 (ランダム変数の和)

Mean Field Theory　(統計神経力学ともいう)

ランダム深層ニューラルネットの平均場理論

第l層の平均活動度

$$q^l = \sum_i^M (u_i^l)^2/M$$

$M \gg 1$

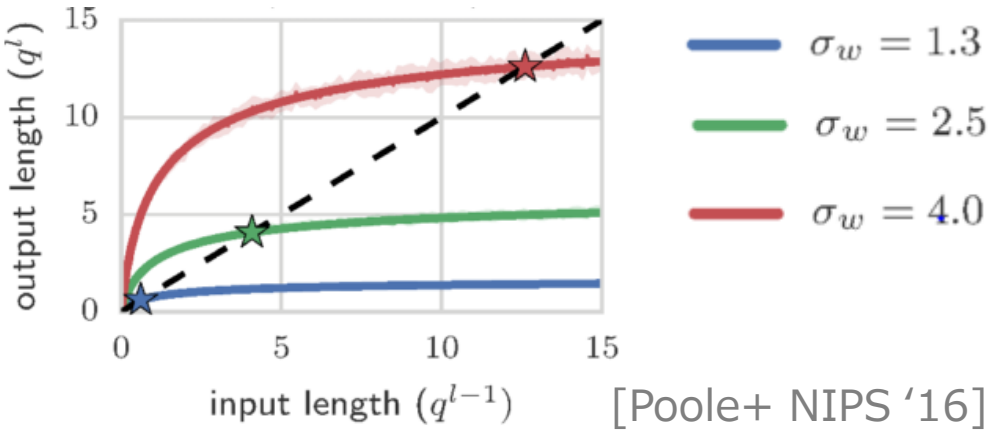
再帰的計算

$$q^l = \sigma_w^2 \int Dz\phi^2\left(\sqrt{q^{l-1}}z\right) + \sigma_b^2$$

ガウス積分

基本的には,
深層学習の ”平均場”理論 = 大数の法則と中心極限定理

$\phi(\cdot) = \tanh(\cdot)$ のとき



[Poole+ NIPS ‘16]

様々なarchitecture (shallow&deep, sigmoid, ReLU, ResNet, CNN ...) に対して共通して行える計算

ランダム深層ニューラルネットの平均場理論

第l層の平均活動度

$$q^l = \sum_i^M (u_i^l)^2/M$$

$M \gg 1$

再帰的計算

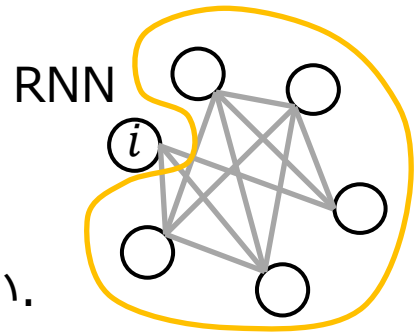
$$q^l = \sigma_w^2 \int Dz\phi^2\left(\sqrt{q^{l-1}}z\right) + \sigma_b^2$$

ガウス積分

基本的には,
深層学習の ”平均場”理論 = 大数の法則と中心極限定理

参考: Recurrent Neural Network (RNN)の平均場 ”**近似**”
[Rozonoer (1969)] [Amari (1970-)][Sompolinsky+ PRL(1988)]

秩序変数の更新則が深層ネットとRNNで類似しているので, アナロジーで“平均場”理論と呼ばれている.
RNNの(真の意味の)平均場は, セルフコンシストに解く必要がある. また, 解が厳密な計算と一致するとは限らない.
“Amari solution” [Crisanti, Sommers, Sompolinsky (2008)]



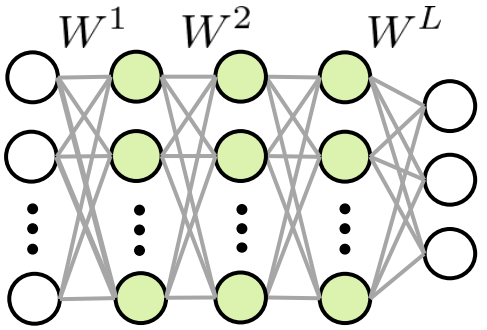
個々の素子が独立で同じ統計性を持つと仮定

ランダム深層ニューラルネットの平均場理論

[Amari (1970-)][Poole+ NIPS '16]

$$u_i^l = \sum_j W_{ij}^l h_j^{l-1} + b_i^l, \quad h_i^l = \phi(u_i^l)$$

異なる2入力 $h^0(a), h^0(b)$



$$q_{ab}^l = \sum_i^M u_i^l(a) u_i^l(b)/M$$

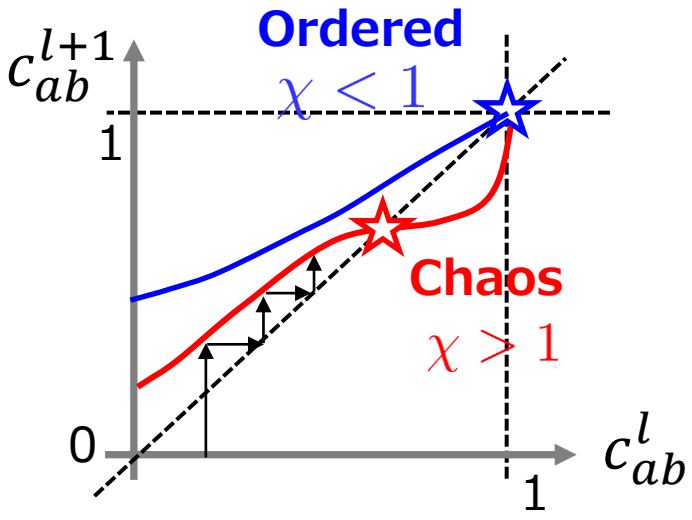
$M \gg 1$

$$q_{ab}^l = \sigma_w^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi(u_1)\phi(u_2) + \sigma_b^2 \qquad \int \mathcal{D}z = \frac{1}{\sqrt{2\pi}} \int dz e^{-\frac{1}{2}z^2}$$

$$u_1 = \sqrt{q^{l-1}} z_1 \qquad u_2 = \sqrt{q^{l-1}} \left(c_{ab}^{l-1} z_1 + \sqrt{1-(c_{ab}^{l-1})^2} z_2 \right) \qquad c_{ab}^l = q_{ab}^l / q^l$$

異なる入力に対して, ニューロンの活動は**独立とは限らない** (結合を共有しているため). 入力間の相関に応じて活動も**相関**.

信号相関の伝達と秩序-カオス相転移

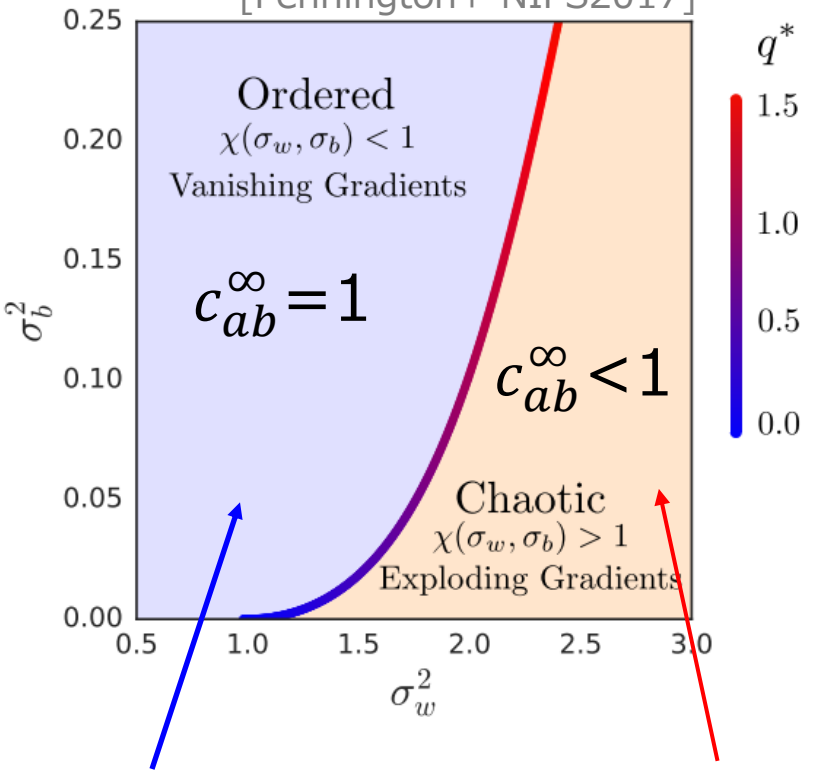


$$q_{ab}^l = \sum_i^M u_i^l(a) u_i^l(b)/M \qquad c_{ab}^l = q_{ab}^l / q^l$$

固定点に注目; $q^* = \lim_{l\to\infty} q^l$

$$\chi = \left.\frac{\partial c_{ab}^l}{\partial c_{ab}^{l-1}}\right|_{c_{ab}=1} = \sigma_w^2 \int \mathcal{D}z \left[\phi'\left(\sqrt{q^*}z\right)\right]^2$$

[Pennington+ NIPS2017]



同一の発火パターンに引き込まれ，
入力信号が区別できない

異なる信号間の差を
拡大する処理

Backpropagationの平均場理論

[Schoenholz+ ICLR '17]

勾配: $\frac{\partial f}{\partial W_{ij}^l} = \delta_i^l \phi(z_j^{l-1})$　　逆誤差伝播: $\delta_i^l = \phi'(z_i^l) \sum_j \delta_j^{l+1} W_{ji}^{l+1}$

勾配ノルム $\tilde{q}^l := \sum_i^M (\delta_i^l)^2$

$$\tilde{q}^l = \tilde{q}^{l+1} \sigma_w^2 \int \mathcal{D}z [\phi'(\sqrt{q_l} z)]^2$$

- 指数関数的な**勾配の消失・発散**が起こる

  転移点はfeedforwardと同じ　$\chi = \sigma_w^2 \int \mathcal{D}z \left[\phi'\left(\sqrt{q^*} z\right)\right]^2$

- 実は, 導出は非自明

仮定: Gradient independence

W_{ji}^{l+1} をfeedforwardと独立なアンサンブルにおきかえる

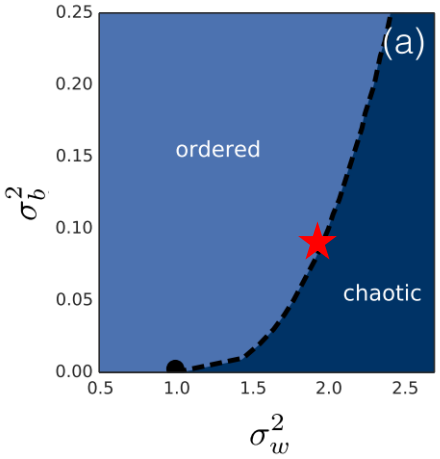
厳密導出: 拘束条件($u_l = W_l h_{l-1}$)つきでのランダムガウス平均

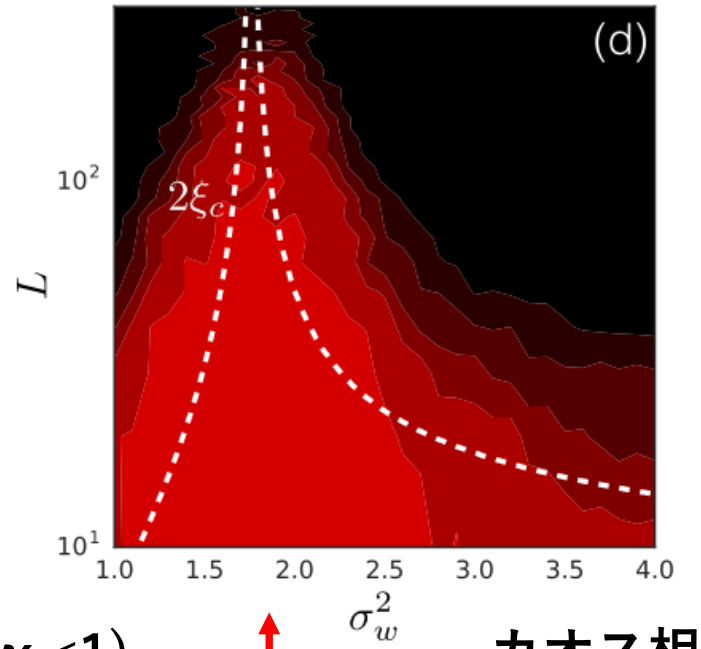
[Yang 2019] [Arora+ 2019]

Backpropagationの平均場理論

[Schoenholz+ ICLR '17]

(右図) MNIST実験
L-layer net, σ_w^2 は初期値の分散





訓練誤差
Red: high
Black: low

白点線:
理論線

秩序相 ($\chi<1$)

Forward: 入力信号が区別できない

$$c_{ab}^{\infty} = 1$$

Backward: 勾配ノルム消失

カオス相 ($\chi>1$)

類似の信号の差も大きくなる

$$c_{ab}^{\infty} < 1$$

勾配ノルム発散

転移点をパラメータ初期値にすることで高い訓練性を実現

ランダム行列理論とDynamical Isometry

$$u_l = W_l h_{l-1} + b_l, \quad h_l = \phi(u_l)$$

Input-output Jacobian: $$\frac{\partial h_L}{\partial h_0} = \prod_{l=1}^{L} D_l W_l \quad D_l = \mathrm{diag}(\phi'(u_l))$$

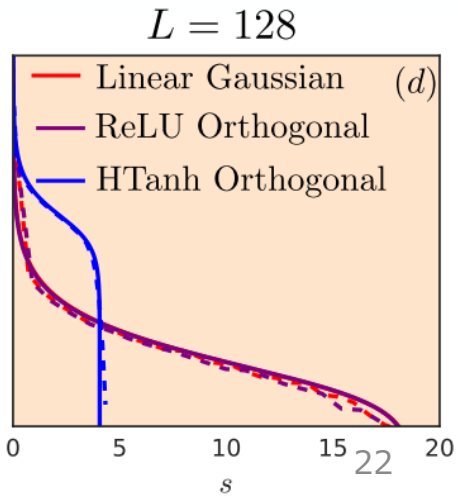
平均場理論: 勾配の消失/発散を防ぐには $\chi = 1$**が必要**

特異値の二乗平均 $\sim \chi^L$

Dynamical Isometry:
特異値の平均だけでなく, 分布の形状を層数と独立にしたい

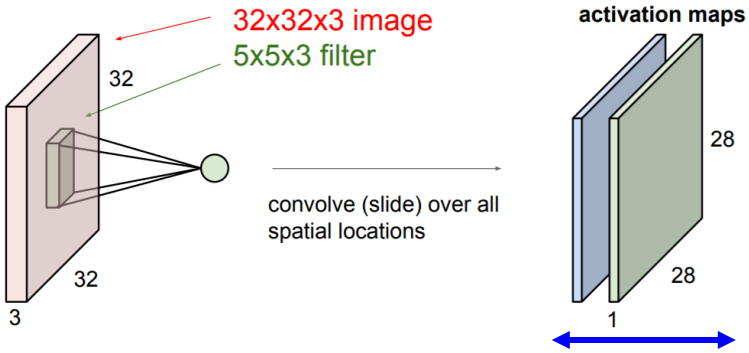
- 自由確率論によるJacobianのスペクトル解析
  [Pennington+ NIPS '17, AISTATS '18]
  [Pastur, arXiv2001.06188]

  - ガウス初期値ではなく, **直交初期値**が必要
  - ReLUには補正(e.g. shift)が必要



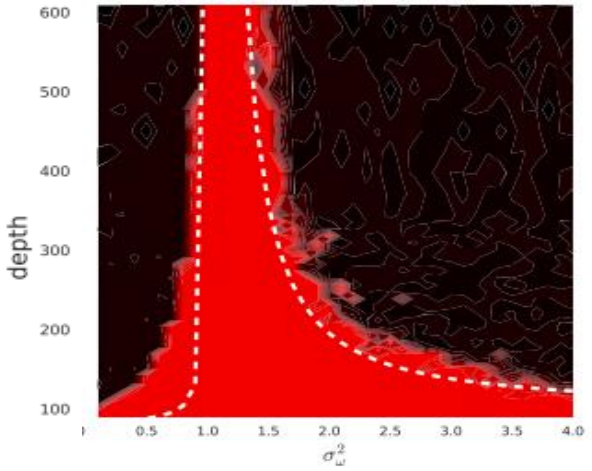
さまざまなアーキテクチャでの検証

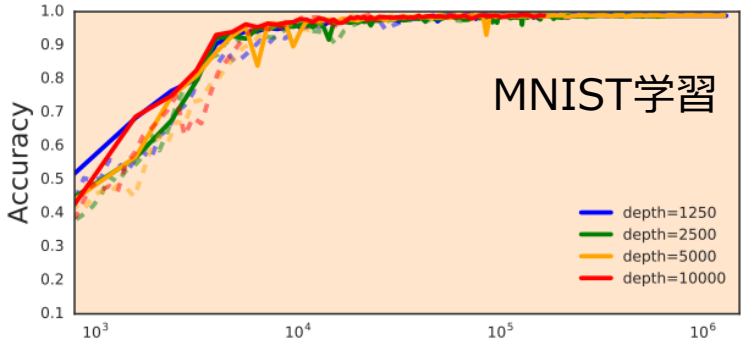
- Convolutional Neural Network [Xiao+ ICML 2018]



[http://cs231n.stanford.edu/slides/2018/cs231n_2018_lecture05.pdf]

フィルタ数 (チャネル数) →∞ の極限





Dynamical isometryを満たす初期値で

10,000層CNNの訓練に成功

そのほか ResNet [Yang+ NIPS '17], (gated) RNNs [Chen+ ICML '18] ...

おまけ - 誰が平均場理論に再注目したか？ -

Career opportunities [edit]

[https://en.wikipedia.org/wiki/Google_Brain]

Google Brain Residency Program [edit] (2016-2017期)

Google Brain Residency Program[29] is targeted at people who are eager to devote own passion to machine learning and artificial intelligence. This is an opportunity to get hands-on experience in Google team and have chance to keep in touch with professional researchers with Google Brain team. The program lasted 12 months.

Within the program were groups of new graduates from top universities with degree of BAs or Ph.Ds in computer science, physics, mathematics, and neuroscience, or others who come from years of industry experience. They were picked to work with researchers in Google Brain Team at the forefront of machine learning.

The breadth of topics in this program allowed team members to flexibly combine their professional knowledge with the application of algorithms, natural language understanding, robotics, neuroscience, genetics and more. Just several months, these new future researchers have already made a great impact in the research field.

Some of the recently published technical papers resulting from the residency program are listed below:

Unrolled Generative Adversarial Networks

Conditional Image Synthesis with Auxiliary Classifier GANs

Regularizing Neural Networks by Penalizing Their Output Distribution

Mean Field Neural Networks

Learning to Remember

Towards Generating Higher Resolution Images with Generative Adversarial Networks

Multi-Task Convolutional Music Models

Audio DeepDream: Optimizing Raw Audio With Convolutional Networks

理論を検証できる数値実験が必要．系統的かつ大規模に．

目次

- **イントロダクション**

- **統計力学的アプローチの紹介**
  - 深層学習とランダムネス
  - 平均場理論
  - ランダム行列理論

- **さらなる発展: Fisher情報行列とNeural Tangent Kernel**
  - Fisher情報行列と学習率

    [Karakida, Akaho & Amari, *Universal Statistics of Fisher Information in Deep Neural Networks: Mean Field Approach,* AISTATS 2019]
  - Neural Tangent Kernelとの対応
  - Batch Normalizationの解析

    [Karakida, Akaho & Amari, *The Normalization Method for Alleviating Pathological Sharpness in Wide Neural Networks,* NeurIPS 2019]

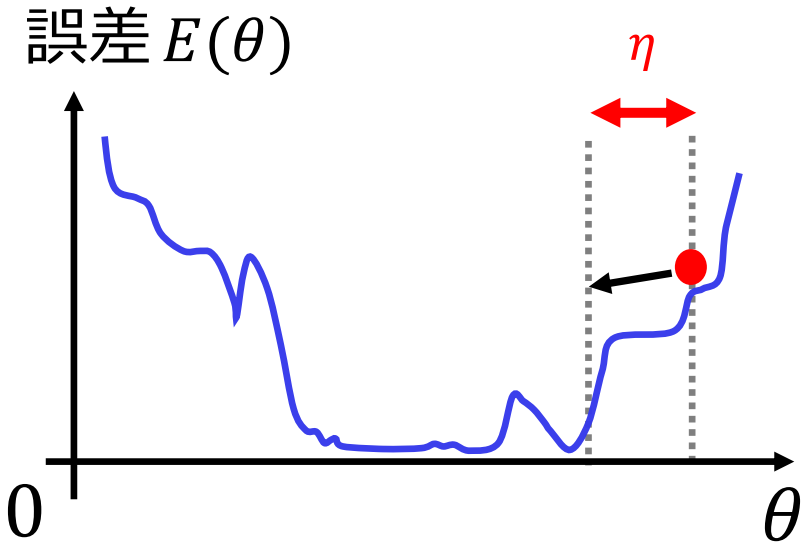
パラメータ空間の重要性

- 最急勾配における**学習率の設計**

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \frac{\partial E(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

学習率 (ステップ幅)

収束する学習率は誤差ランドスケープの形で決まる
(勾配の消失とは別の訓練性の問題)



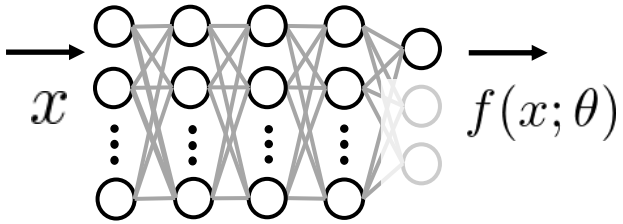
- 学習率の決め方はヒューリスティックが多い
  e.g. Training lossの減少がε以下になったら, 学習率を1/a倍にする

Fisher情報行列

二乗回帰では,

$$F = \mathrm{E}[\nabla_\theta f(x;\theta) \nabla_\theta f(x;\theta)^\top]$$

E[・]：入力サンプル平均 x

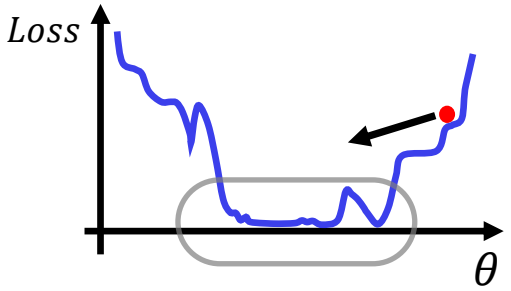


- 訓練誤差ゼロの大域解まわりのヘシアンに対応

$$Loss(\theta) = \mathrm{E}[||y - f||^2]$$

Hessian:

$$\nabla_\theta \nabla_\theta Loss(\theta) = F - \sum_{k=1}^{C} \mathrm{E}[(y_k - f_k) \nabla_\theta \nabla_\theta f_k]$$



- 情報幾何の基本量. パラメータ空間は曲がっている

$$D_{\mathrm{KL}}(p(x,y;\theta)||p(x,y;\theta+d\theta)) \sim d\theta^\top F d\theta$$

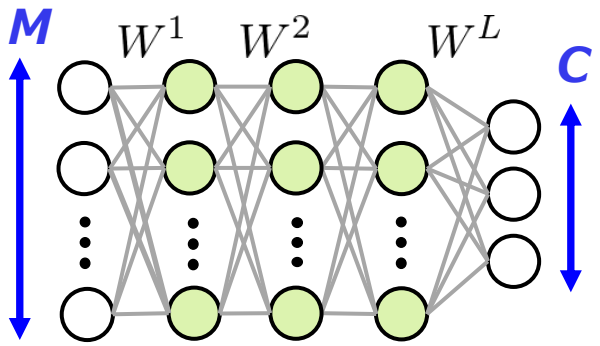
設定

$$u_i^l = \sum_j W_{ij}^l h_j^{l-1} + b_i^l, \quad h_i^l = \phi(u_i^l) \quad (l = 1, 2, ..., L)$$

- パラメータはrandom Gaussian

$$W_{ij}^l \sim \mathcal{N}(0, \sigma_w^2/M) \quad b_i^l \sim \mathcal{N}(0, \sigma_b^2)$$

入力/中間層: **素子数 M ($\gg 1$)**
出力層: **素子数 C** ($= \boldsymbol{O(1)}$)



- 入力もrandom Gaussian

$$x_i(t) \sim \mathcal{N}(0, 1) \qquad (t = 1, ..., T)$$

訓練サンプル数

- non-centered network (e.g. ReLU, Tanh with bias terms, …)

バイアスが非ゼロ($\sigma_b \neq 0$)あるいは活性化関数のガウス平均が非ゼロ ($\int Dz\phi(z) \neq 0$)

Fisher情報行列の巨視的理解

・FIMの計算方法 = chain rule (backprop.と同様)

$$F = \mathrm{E}[\nabla_\theta h^L(x)^\top \nabla_\theta h^L(x)]$$

$$\frac{\partial h_k^L}{\partial W_{ij}^l} = \delta_i^l \phi(u_j^{l-1}), \quad \delta_i^l = \phi'(u_i^l) \sum_j \delta_j^{l+1} W_{ji}^{l+1} \quad (l = 1, ..., L-1)$$

(i) Feedforwardの秩序変数 [Amari I970-, Poole+ NIPS '16]

$$\hat{q}^l := \sum_i (h_i^l(t))^2 / M$$

$$\hat{q}_{st}^l := \sum_i h_i^l(s) h_i^l(t) / M$$

(ii) Backpropagationの秩序変数 [Schoenholz+ ICLR '17]

$$\tilde{q}^l := \sum_i (\delta_i^l(t))^2$$

$$\tilde{q}_{st}^l := \sum_i \delta_i^l(s) \delta_i^l(t)$$

Fisher情報行列の固有値

M (width) が十分に大きいとき, 漸近的に

$$\lambda_{max} \sim (L-1)\left(\frac{T-1}{T}\kappa_2 + \frac{\kappa_1}{T}\right)M$$

L: 層数
T: 訓練サンプル数

$$\kappa_1 := \sum_{l=1}^{L} \tilde{q}^l \hat{q}^{l-1}/(L-1) \qquad \kappa_2 := \sum_{l=1}^{L} \tilde{q}^l_{st} \hat{q}^{l-1}_{st}/(L-1)$$

- 非常に大きい孤立した最大固有値 $\boldsymbol{O(M)}$

$$\mathrm{E}[\nabla_\theta f \nabla_\theta f^\top] = \mathrm{Cov}(\nabla_\theta f, \nabla_\theta f) + \mathrm{E}[\nabla_\theta f]\mathrm{E}[\nabla_\theta f]^\top$$

$$\rightarrow \lambda_{max} \sim ||\mathrm{E}[\nabla_\theta f]||^2$$

- クラス数Cだけ λ_{max} に縮退

Eigenvectors $\mathrm{E}[\nabla_\theta f_k] \quad (k = 1, ..., C)$

Fisher情報行列の固有値

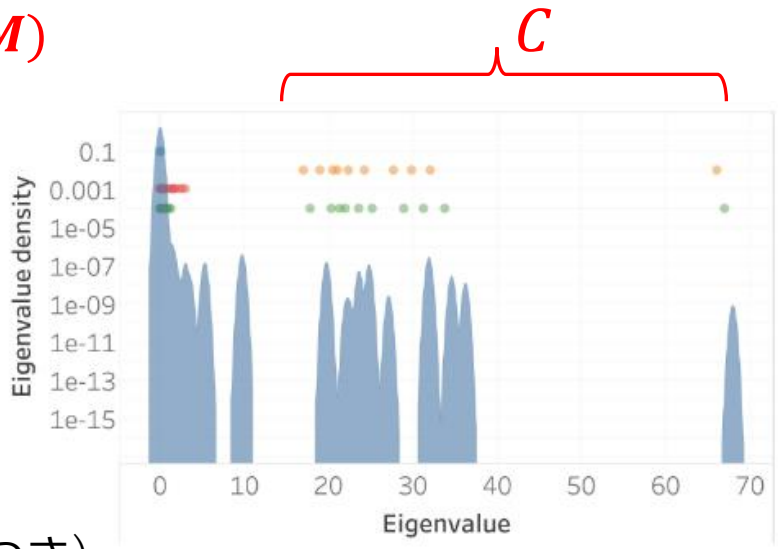
M (width) が十分に大きいとき, 漸近的に

$$\lambda_{max} \sim (L-1)\left(\frac{T-1}{T}\kappa_2 + \frac{\kappa_1}{T}\right) M$$

L: 層数
T: 訓練サンプル数

- 同様にして, 固有値の平均値は $\boldsymbol{O(1/M)}$
- クラス数だけ大きな固有値の実験的報告

  [Sagun+ 2017] [Papyan, ICML '19]



Cross-entropy [Papyan, 2019]

Remark: Lossの種類 (or データのばらつき) によって固有値はばらつきうる

最大固有値と学習率

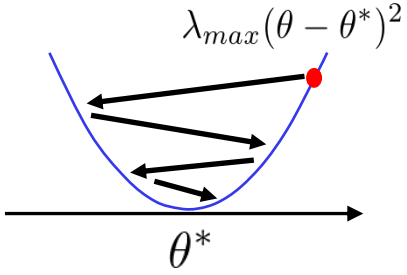
- 最急勾配 (バッチ学習)　$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \dfrac{\partial E(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$

学習率

最急勾配が大域解　$\theta^* = \{W^{l*}, b^{l*}\}$ s.t. $E(\theta^*) = 0$
の近傍で収束する**必要条件**

$$\eta < 2/\lambda_{max}$$

[LeCun, Kanter & Solla, PRL 1991]など

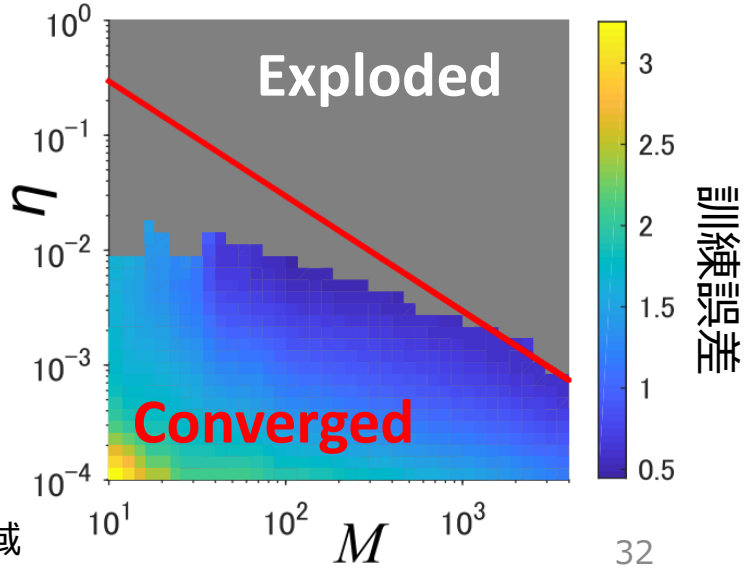


(右図) 1 epoch後の訓練誤差

L=4 ReLU on MNIST, SGD

- 発散/収束する学習率に明確な境界

理論線$(= 2/\lambda_{max})$



グレー:
誤差が発散した領域

Neural Tangent Kernel (NTK) 理論

[Jacot+ NeurIPS '18]

$$\frac{d\theta_t}{dt} = \eta \nabla_\theta f_t^\top (y - f_t) \quad \nabla_\theta f_t : CN \times P(\text{パラメータ数})\text{行列}$$

- 学習率のオーダーを$1/M$でとる (or NTK parameterizationを使う)
- 対応する関数勾配をみる

$$\frac{df_t}{dt} = \eta \nabla_\theta f_t \nabla_\theta f_t^\top (y - f_t)$$

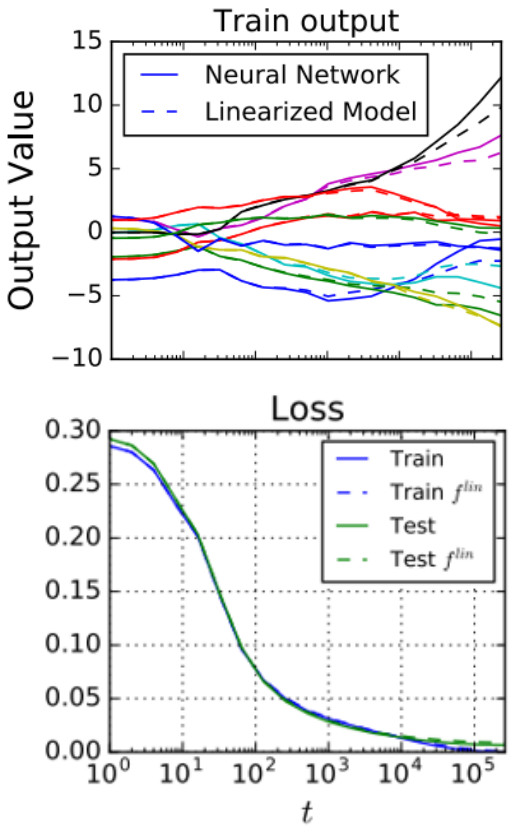
$$=: \Theta_t \quad \text{NTK } (CN \times CN \text{ 行列})$$

(Informal)

隠れ層幅が無限大のとき，f_tのダイナミクスは

$$\frac{df_t^{\text{lin}}}{dt} = \eta \Theta_0 (y - f_t^{\text{lin}})$$

のダイナミクスと一致



MNIST; ReLU (L=2), M=2048, SGD+momentum [Lee+ NeurIPS '19]

Neural Tangent Kernel (NTK) 理論

[Jacot+ NeurIPS '18] [Lee+ NeurIPS '19]

- **線形モデルの訓練と等価** $f_t^{\mathrm{lin}}(x) = f_0(x) + \underbrace{\nabla_\theta f_0(x)^\top}_{\sim M^{-1/2}} (\theta_t - \theta_0)$

微小変化するパラメータがたくさんあるので
出力は $\mathcal{O}(1)$で変化

$\eta \nabla_\theta f \sim M^{-1} \cdot M^{-1/2}$

$P \cdot M^{-2} \sim 1$

- **大域収束** $f_t = f_0 + (I - \exp(-\Theta_0 t))(y - f_0)$

- 未知データx'に対しても可解. Gaussian Processと等価.
  特に, 訓練されたモデルは**Kernel ridge-less regression**

$$\langle f_\infty(x') \rangle_{\mathrm{ini.}} = \Theta(x', x)\Theta(x, x)^{-1} y$$

※ NTKの最小固有値は正と仮定 (たとえば, 入力の正規化とnon-poly.のactivationで成立)
※ 非漸近論 (十分大きいが有限のM) による収束証明も多数

$M \gtrsim T^3$ [Huang & Yau ICML 2020]

そのほかの収束証明の概要: [Zou & Gu, "An improved analysis of …", NeurIPS '19]

NTKの構造

- カーネルの具体的な計算は秩序変数のrecurrence eq.を使っている

$$\Theta_{ab} = \sum_{l=1}^{L} \tilde{q}_l(a,b)\hat{q}_{l-1}(a,b)$$

backward path:

$$\tilde{q}_l(a,b) = \sum_{i=1}^{M} \delta_{l,i}(a)\delta_{l,i}(b)$$

Feedforward path:

$$\hat{q}_l(a,b) = \sum_{i=1}^{M} h_{l,i}(a)h_{l,i}(b)/M$$

- **FIMと固有値を共有**

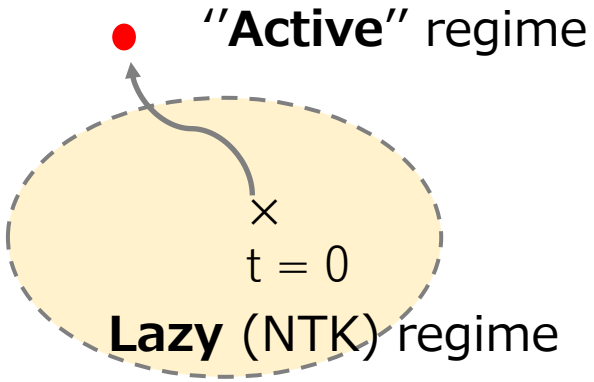
- **''Duality''** $FIM = \nabla_\theta f^\top \nabla_\theta f$ $NTK =$

パラメータ空間　　関数空間

NTK regimeの ”外側”

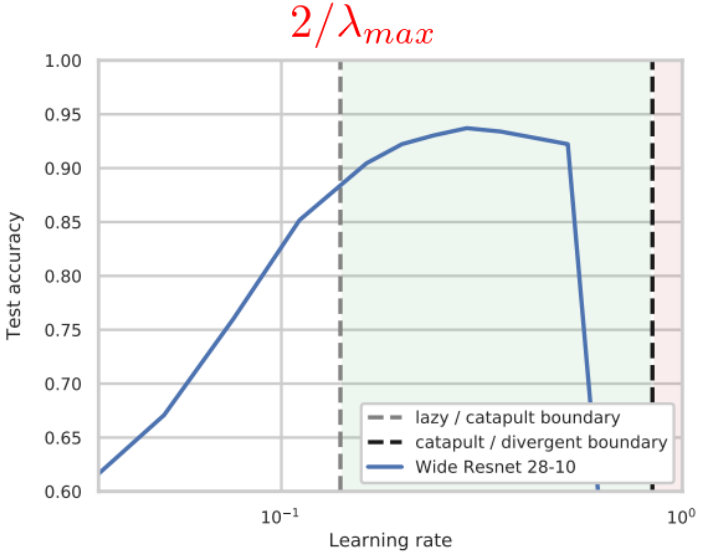
- 初期値から大きく離れる状況では NTK理論の成立が非自明
  (e.g. 有限幅の効果, ノイズのような外乱 …)

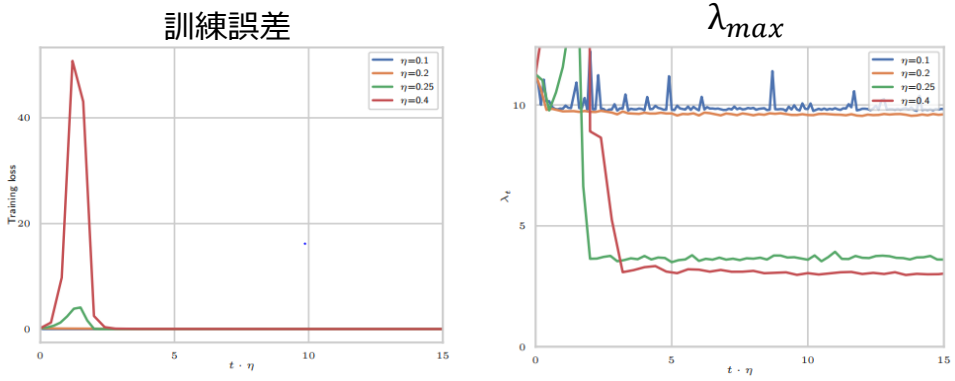
  - 外側こそが積極的に”特徴抽出”



- “Large learning rate phase”

  [Lewkowycz+ arXiv2003.02218]





実験 @Wide ResNet

Toy modelも解析. 有限幅が$\lambda_{max}(t)$を減少させて, 収束をもたらす; $t \sim \ln M$

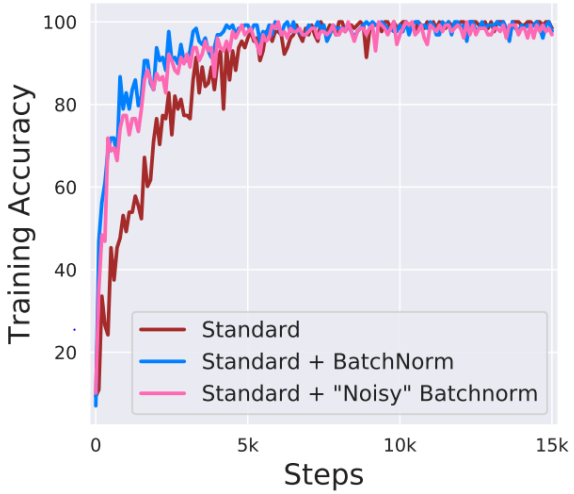
Batch Normalization (BN)

[Ioffe&Szegedy ICML '15]

各ニューロンをサンプルに対して正規化

$$u_i^l(t) \leftarrow \frac{u_i^l(t) - \mu_i^l}{\sigma_i^l}\gamma_i^l + \beta_i^l$$

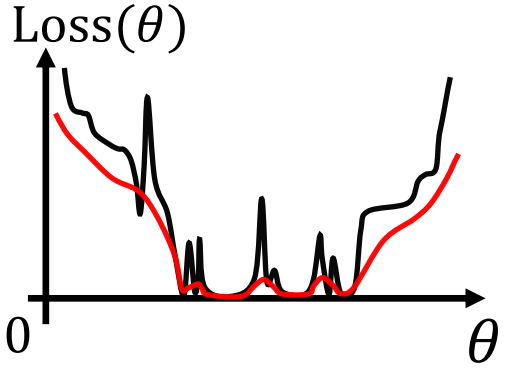
$$\mu_i^l := \mathrm{E}[u_i^l(t)] \quad \sigma_i^l := \sqrt{\mathrm{E}[u_i^l(t)]^2 - (\mu_i^l)^2}$$



経験的には - 大きい学習率での高速な収束
- 汎化しやすい

BNのメカニズム [Santurkar+ NeurIPS '18]

・通説 (Internal covariate shiftの抑制)への反証
・Loss landscapeの急激な変化を抑えている可能性



Batch norm in the last layer

簡単のため, **最終層のmean subtraction**のみを考える

$$\bar{f}_k(t) := (u_k^L(t) - \mu_k(\theta))\gamma_k + \beta_k \qquad k = 1, \ldots, C \text{ (クラス数)}$$

仮定(I) 幅Mと訓練サンプル数Tが十分大きく, $\rho := M/T$(固定)
(II) gradient independence

$$\rho\alpha(\kappa_1 - \kappa_2) + c_1 \leq \lambda_{max} \leq \sqrt{(C\alpha^2\rho(\kappa_1 - \kappa_2)^2 + c_2)M}$$

κ_1, κ_2 : 秩序変数から計算, $\alpha = L - 1$
c_1, c_2 : 非負値定数

- λ_{max} のオーダーが $\Theta(M)$ から高々 $\Theta(\sqrt{M})$ へ減少

導出の方針: $\mathrm{E}[\nabla_\theta f \nabla_\theta f^\top] = \mathrm{Cov}(\nabla_\theta f, \nabla_\theta f) + \mathrm{E}[\nabla_\theta f]\mathrm{E}[\nabla_\theta f]^\top$

$$\bar{f}_k(t) = f_k(t) - \mathrm{E}[f_k] \qquad \mathrm{E}[\nabla_\theta \bar{f}_k] = 0$$

Batch norm in the last layer

簡単のため, **最終層のmean subtraction**のみを考える

$$\rho\alpha(\kappa_1-\kappa_2)+c_1 \le \lambda_{max} \le \sqrt{(C\alpha^2\rho(\kappa_1-\kappa_2)^2+c_2)M}$$

κ_1, κ_2：秩序変数から計算, $\alpha = L-1$

c_1, c_2：非負値定数

- λ_{max} のオーダーが $\Theta(M)$ から高々 $\Theta(\sqrt{M})$ へ減少

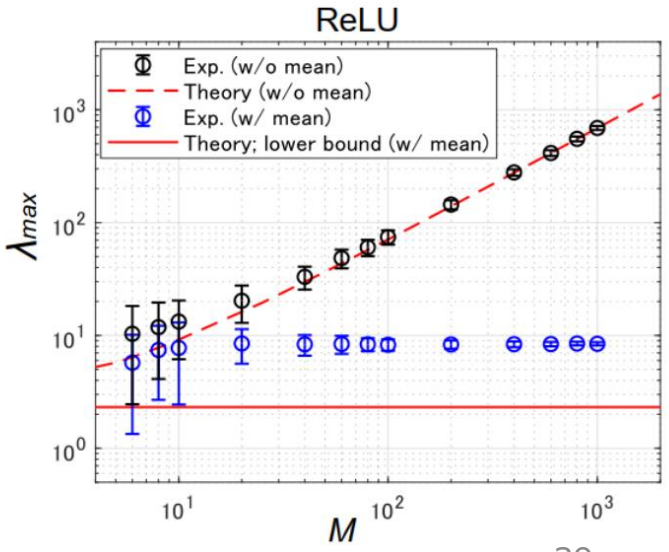
  理論の上限は保守的. 経験的には $\Theta(1)$

- 中間層のみにBN

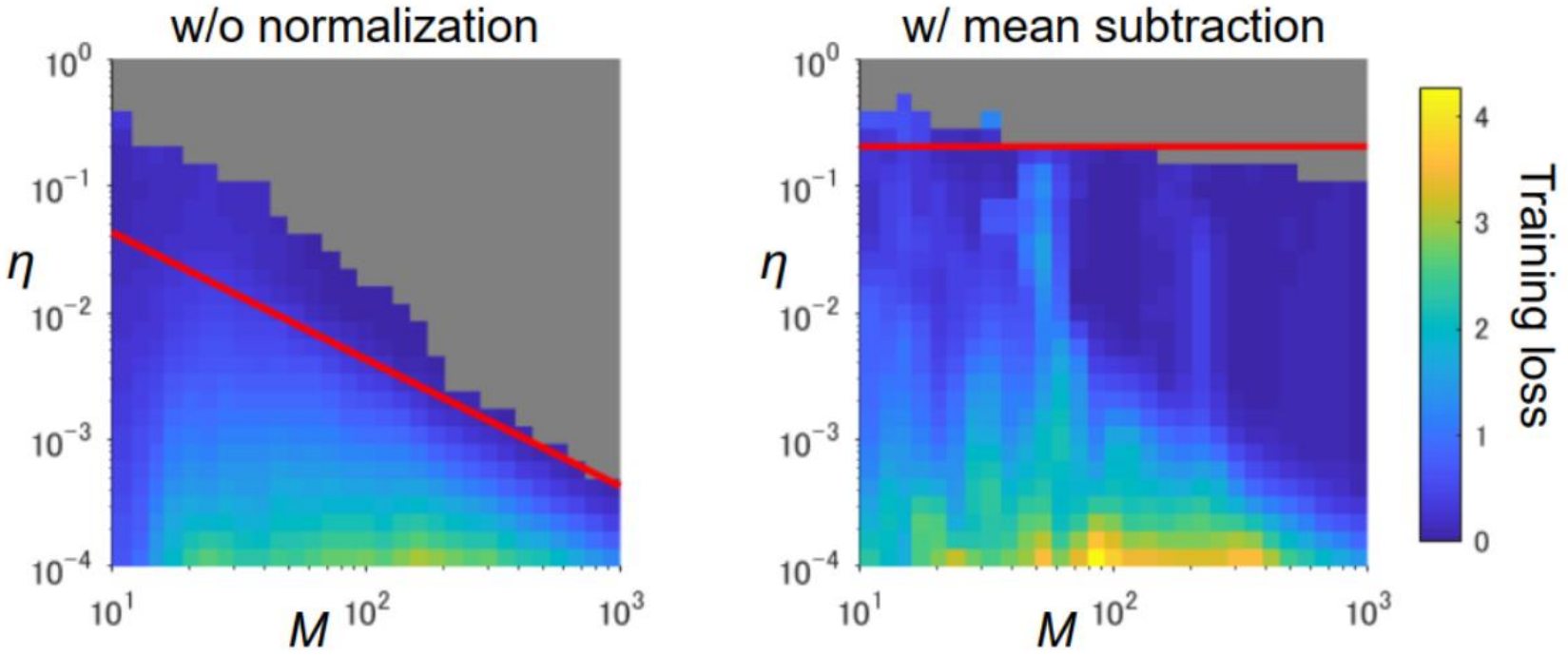
  Deep ReLU net では, $\lambda_{max} = \Theta(M)$

- Layer normalization $\lambda_{max} = \Theta(M)$

$$\mathrm{E}[\nabla_\theta f_k(t)] \neq \frac{1}{C}\sum_{k=1}^{C}[\nabla_\theta f_k(t)]$$



学習率η とλ_{max}



理論線(赤): $\eta = 2/\lambda_{max}$

$\eta = 2/(\rho\alpha(\kappa_1 - \kappa_2))$

最終層BNだけで, 幅に依存しない大きな学習率が許容された

Dynamical isometry成立下でのFIM

[Hayase & Karakida arXiv:2006.07814]

Input-output Jacobian:

$\frac{\partial h_L}{\partial h_0} = \prod_{l=1}^{L} D_l W_l$ の統計性が層数に非依存でも，FIMは依存する

Conditional FIM:
(given a single input x)

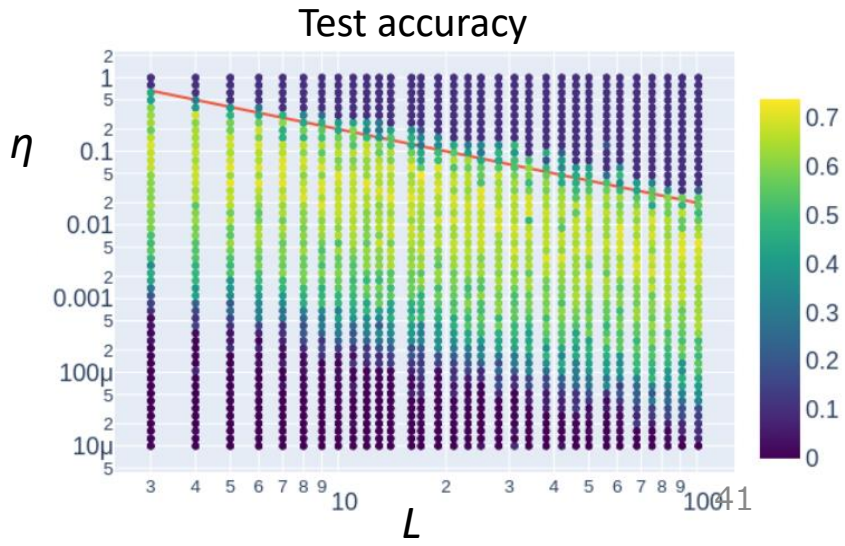
$$P \overset{C}{\boxed{\nabla_\theta f^\top}} \boxed{\nabla_\theta f}$$

自由確率論でスペクトル解析.
固有値は集中して，$\lambda_{\neq 0} \sim L$

- オンライン学習 (mini-batch size =1)

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta [M^{-1} \mathcal{L}(f_{\theta_t}(x(t) - \ y(t))]$$

[右図] 学習初期(500 step)の挙動
Hard Tanh on Fashion-MNIST



まとめ

ランダムネスに基づいた, 大自由度(幅無限大)極限
における深層ネットの数理を紹介した

- DNNの挙動を統一的にとらえられる, **普遍性**のある枠組み
- さらに, 個別の現象に**定量的な示唆**を与える

・勾配の発散/消失を防ぐ初期値
・学習率の設定
・NTK regimeの発見

今後の課題

- 各種モデル・アルゴリズムのNTK regimeでの理解
- 固有値解析 ($2/\lambda_{max}$境界，NTK regime の成立条件 $\lambda_{min} > 0$) の精緻化
- NTK regimeの外側への展開