深層学習の数理：
ランダム行列と統計力学的視点

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

RFM オンラインセミナー

Oct. 30

Outline

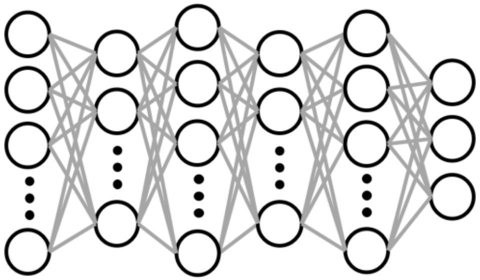
- **背景**

- **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

- **発展: Fisher情報行列とNeural Tangent Kernel**
  - Fisher情報行列
    - Loss landscapeと学習率 [KAA19a], [HK20]
    - Batch Normalizationの役割 [KAA19b]

  - Neural Tangent Kernel
    - 近似自然勾配の高速な収束 [KO20]

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 ...

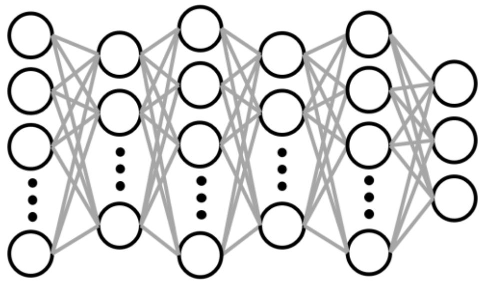
+ Adversarial attacks & defences, 解釈性,
contrastive learning ...

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

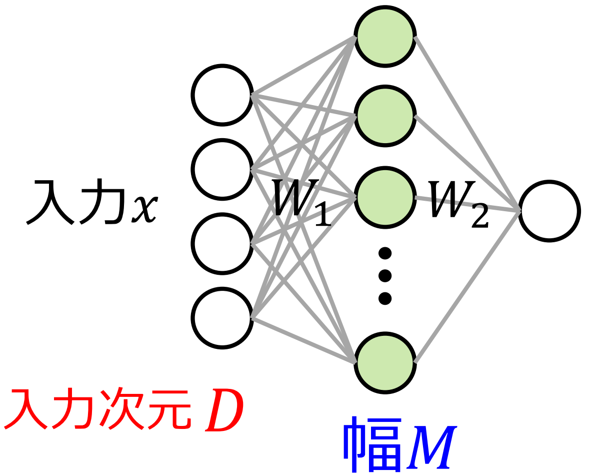
- **Barronの定理** (1993)

(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

$$f_M(x) = W_2 \mathrm{Sigmoid}(\mathrm{W}_1 \mathrm{x})$$



一方, 基底(W_1)を学習しない場合,

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

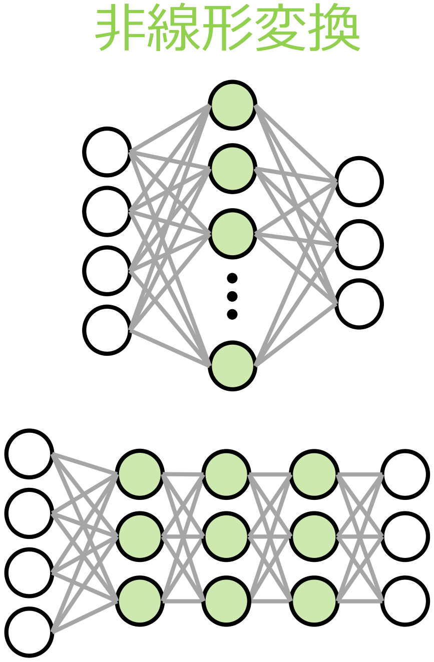
階層性が近似精度における**次元の呪いを克服**. 3層で十分？

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* [Delallea & 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]



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

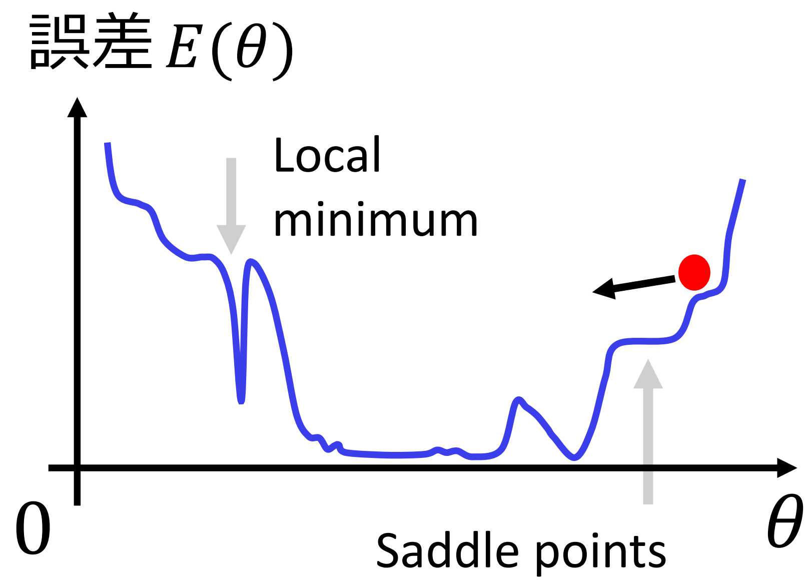
訓練性の問題

表現能力が高い $\neq$ 学習させやすい

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

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

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = g(W_L \cdots g(W_2 g(W_1 x + b_1) + b_2))$$



最急勾配法 (Backpropagation)

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

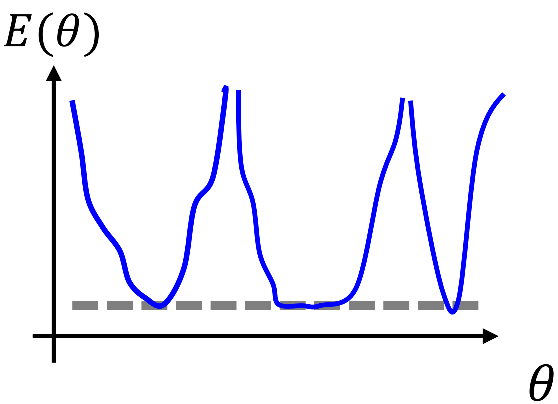
- 一般には, global minimaにたどり着く保証はない
- DNNは勾配が消失 (or 発散)しやすい

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

訓練性とloss landscape

・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) ≧ 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]

$$\theta_{t+1} = \theta_t - \frac{\partial E}{\partial \theta}(\theta, \underline{x_t})$$

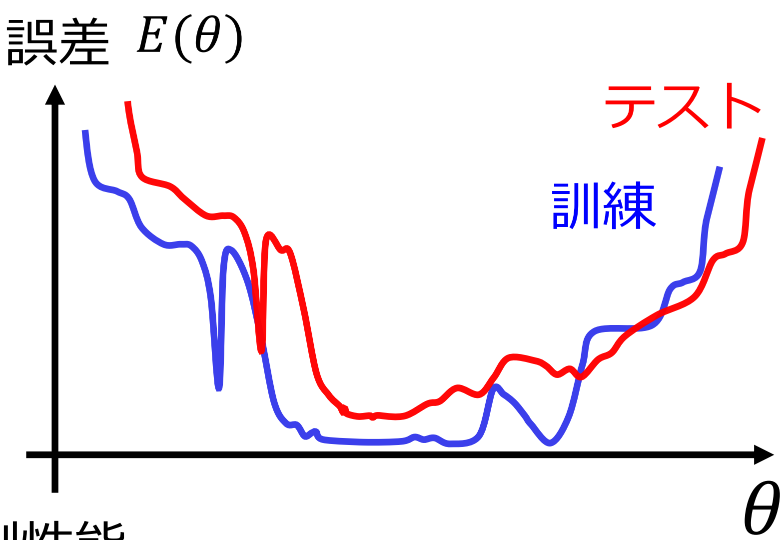
Mini-batch由来のノイズが効く

汎化の問題

表現能力が高い $\neq$ 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

Outline

- **背景**

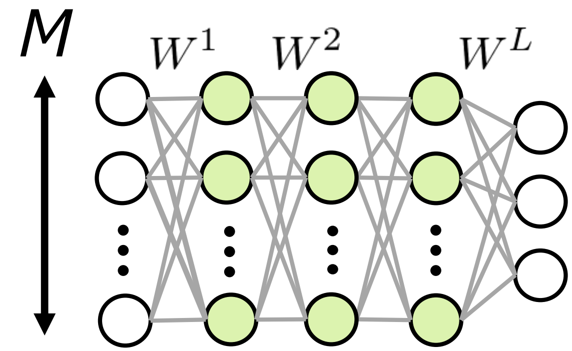
- **統計力学的アプローチの紹介**
  - 深層学習とランダムネス
  - 平均場理論
  - ランダム行列理論

- **発展: Fisher情報行列とNeural Tangent Kernel**
  - Fisher情報行列
    - Loss landscapeと学習率 [KAA19a], [HK20]
    - Batch Normalizationの役割 [KAA19b]

  - Neural Tangent Kernel
    - 近似自然勾配の高速な収束 [KO20]

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 2010]

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

ある層の素子数Mに対し, パラメータ分散は$O(1/M)$

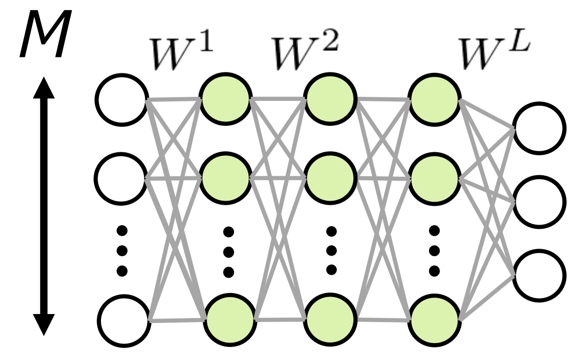
$$\sum_j^M W_{ij} x_j \sim O(1)$$

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

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

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

$M \gg 1$ →

再帰的計算

$$q^l = \sigma_w^2 \underbrace{\int Dz}_{\text{ガウス積分 (ランダム変数の和)}} \phi^2\left(\sqrt{q^{l-1}}z\right) + \sigma_b^2$$

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

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

$M \gg 1$

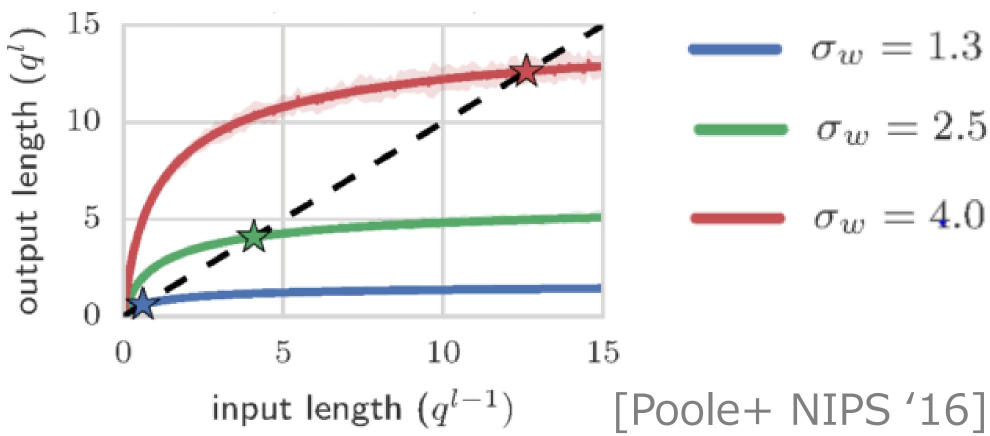
再帰的計算

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

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

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

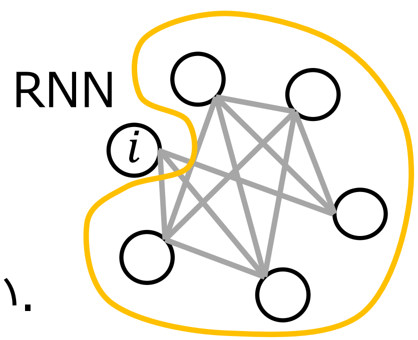
再帰的計算

$$q^l = \sigma_w^2 \underbrace{\int Dz}_{\text{ガウス積分}} \phi^2\left(\sqrt{q^{l-1}}z\right) + \sigma_b^2$$

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)]



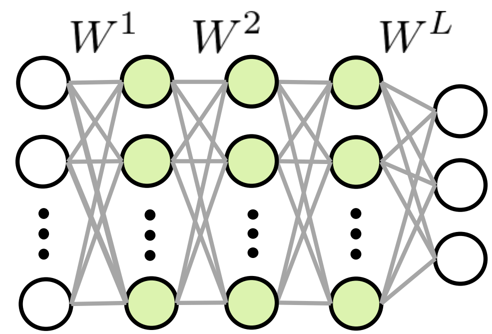
個々の素子が独立で同じ統計性
を持つと仮定

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 $h^0(a), h^0(b)$



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

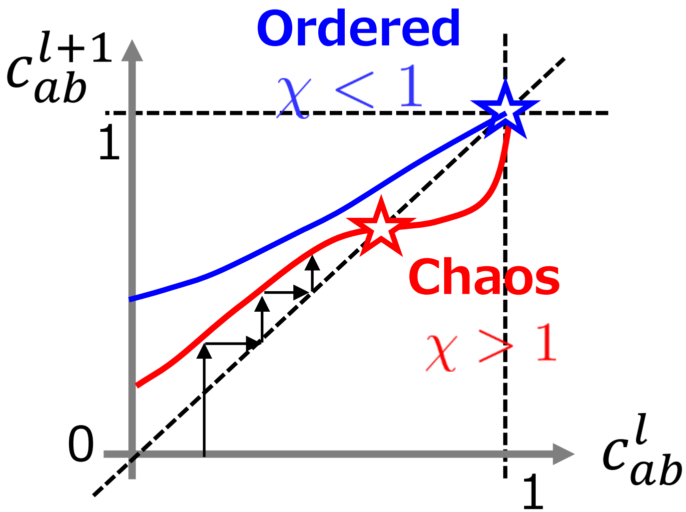
$M \gg 1$

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

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

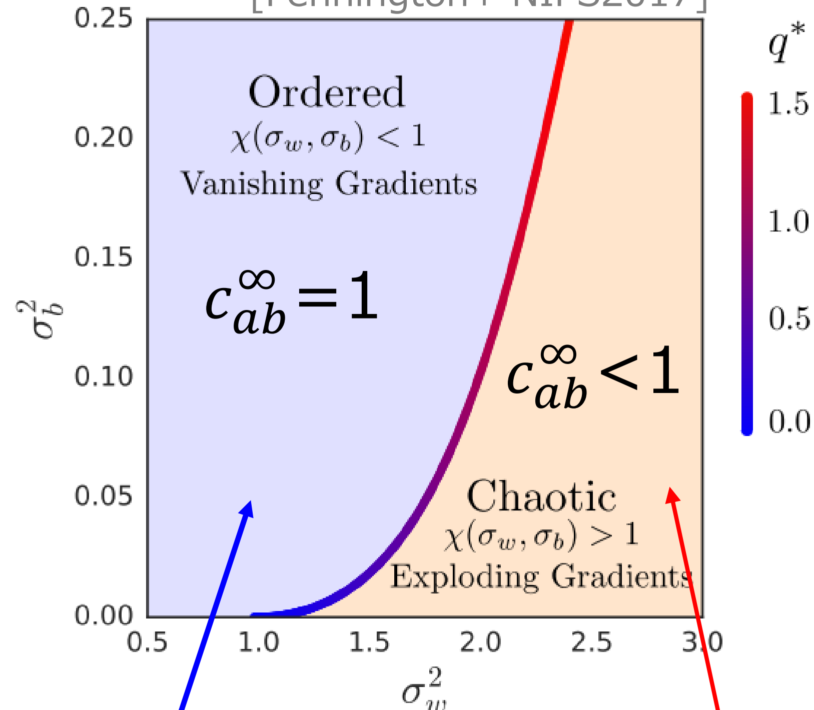
異なる入力に対して, ニューロンの活動は**独立とは限らない** (結合を共有しているため). 入力間の相関に応じて活動も**相関**.

信号相関の伝達と秩序-カオス相転移



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

[Pennington+ NIPS2017]



固定点に注目; $q^* = \lim_{l \to \infty} q^l$

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

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

- 導出は以下を仮定

Gradient Independence Assumption (GIA)

W_{ji}^{l+1} をfeedforwardと独立なガウスランダム行列に置き換えてよい

Backpropagationの平均場理論

[Schoenholz+ ICLR ‘17]

- 導出は以下を仮定

Gradient Independence Assumption (GIA)

W_{ji}^{l+1} をfeedforwardと独立なガウスランダム行列に置き換えてよい

数学的な正当化 [Yang, “Tensor Program II”, arXiv:2006.14548] e.g. $\sum_i (\delta_i^l)^n$

置き換えはGaussian conditioningに由来.

State evolutionの厳密化で類似のテクニック [Bayati&Montanari ‘12]

$A_{ij} \sim \mathcal{N}(0, \sigma^2)$ Q, Y, P, X: fixed matrix

Aが拘束条件 $Y = AQ$, $X = A^\top P$ で条件付けられているとき,

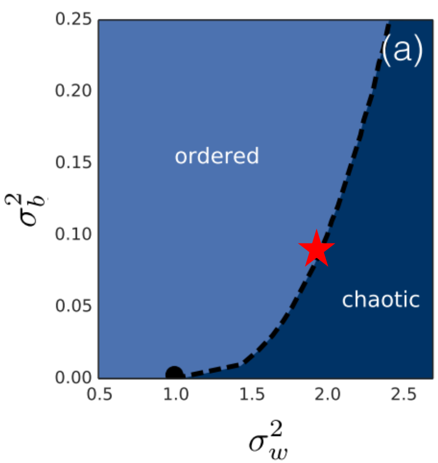
$$A \stackrel{\mathrm{d}}{=}_{Y=AQ, X=A^\top P} E + \Pi_P^\perp \tilde{A} \Pi_Q^\perp$$

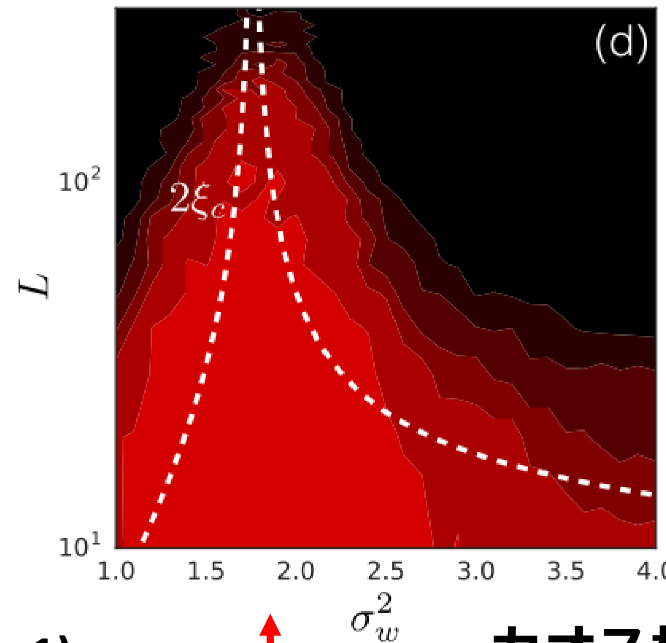
EはQ, Y, P, Xから決まる定数行列, Πは射影行列

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 (χ<1)	**カオス相 (χ>1)**
Forward:	入力信号が区別できない $c_{ab}^{\infty} = 1$	類似の信号の差も大きくなる $c_{ab}^{\infty} < 1$
Backward:	勾配ノルム消失	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: $\dfrac{\partial h_L}{\partial h_0} = \prod_{l=1}^{L} D_l W_l \quad D_l = \mathrm{diag}(\phi'(u_l))$

平均場理論: $L >> 1$で勾配の消失/発散を防ぐには χ = **1が必要**

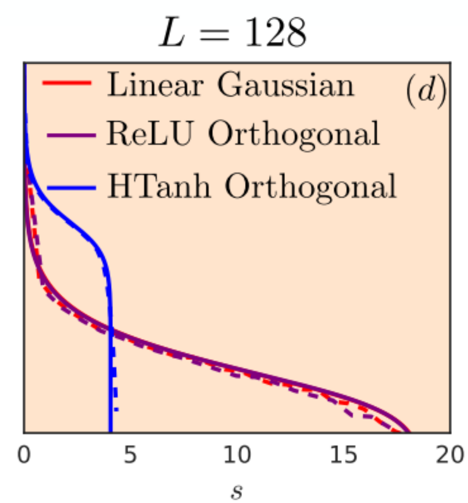
特異値の二乗平均 $\sim \chi^L$

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]

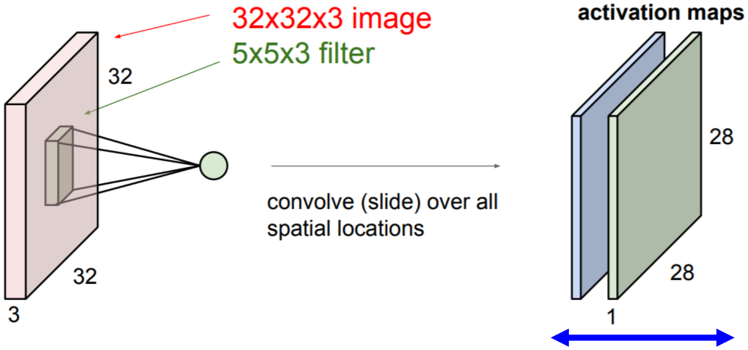
ガウス初期値ではなく, **直交初期値**が必要



実は, D_l, W_l の漸近自由独立性の成立は非自明. ガウス初期値における置き換えの正当化:
[Pastur, arXiv2001.06188] [Yang, "Tensor Program III",arXiv: 2009.10685]

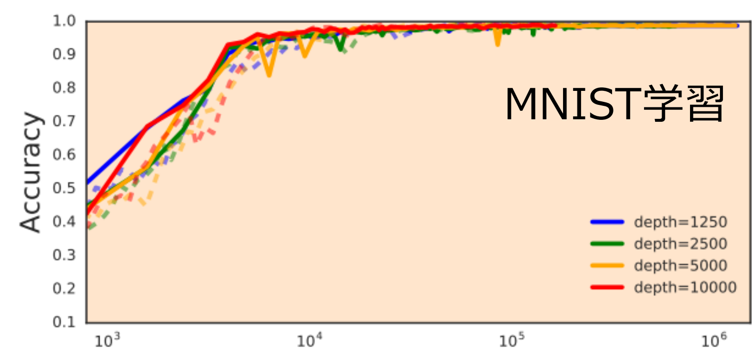
さまざまなアーキテクチャでの検証

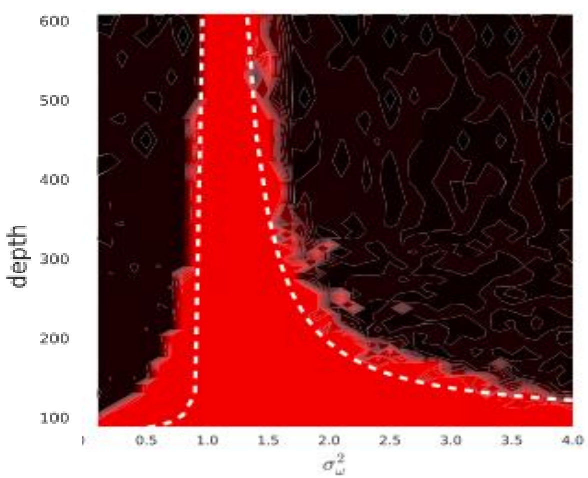
- Convolutional Neural Network [Xiao+ ICML 2018]



[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] …

[参考] DNNとランダムガウス場

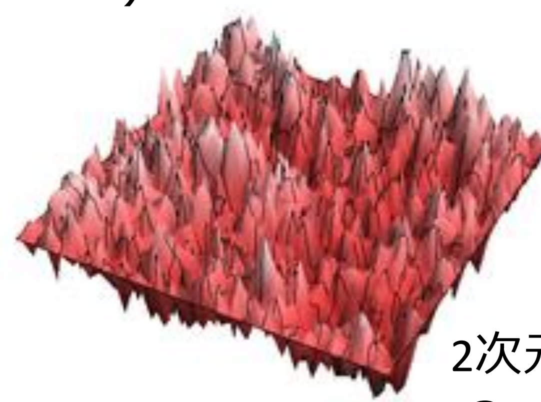
$L(\boldsymbol{w})$ ：**ランダムガウス場** (多次元ガウス過程)

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

[Fyodorov, PRL '04]など



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 ($\|\mathbf{w}_t-\mathbf{w}_0\| \sim \log t$) [Hoffer+ NIPS '17]

Outline

- **背景**

- **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

- **発展: Fisher情報行列とNeural Tangent Kernel**
  - Fisher情報行列
    - Loss landscapeと学習率
      [Karakida, Akaho & Amari, *Universal Statistics of Fisher Information in Deep Neural Networks: Mean Field Approach,* AISTATS 2019]
    - Batch Normalizationの役割
      [Karakida, Akaho & Amari, *The Normalization Method for Alleviating Pathological Sharpness in Wide Neural Networks,* NeurIPS 2019]
  - Neural Tangent Kernel
    - 近似自然勾配の高速な収束

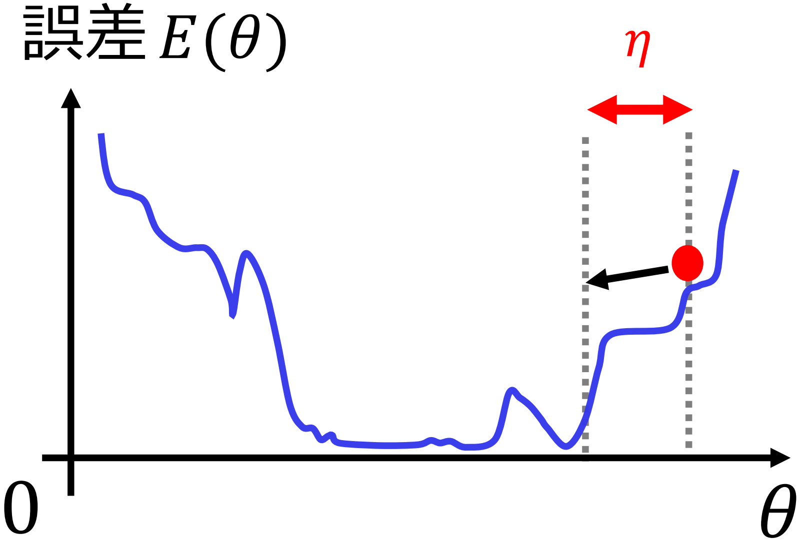
パラメータ空間の重要性

- 最急勾配における**学習率の設計**

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

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<u>訓練性の問題</u>)



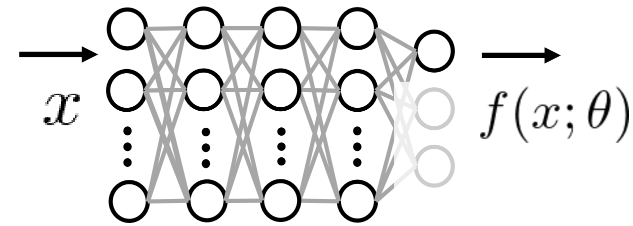
- 学習率の決め方はヒューリスティックが多い
  e.g. Training lossの減少がε以下になったら, 学習率を1/a倍にする

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

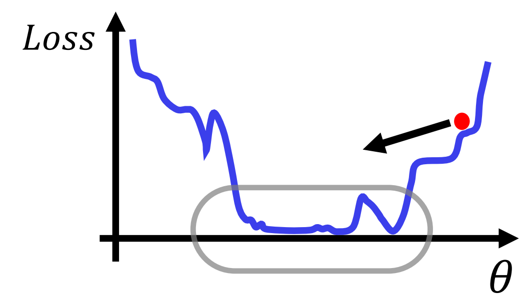


- 訓練誤差ゼロの大域解まわりのヘシアンに対応

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

Hessian:

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



- 情報幾何の基本量. パラメータ空間は曲がっている

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

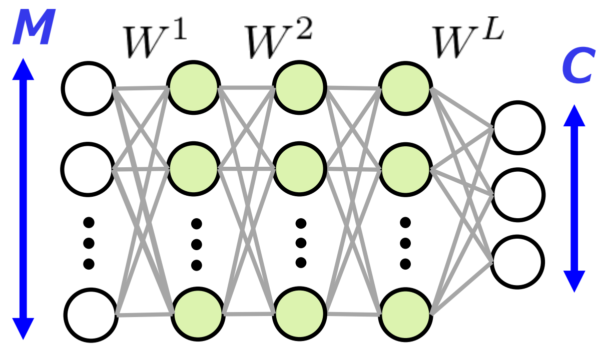
設定

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

- 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*** $(= \boldsymbol{O}(\mathbf{1}))$



- 入力もrandom Gaussian

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

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)

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

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

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

$$\boxed{\hat{q}^l := \textstyle\sum_i (h_i^l(t))^2 / M} \qquad \boxed{\hat{q}_{st}^l := \textstyle\sum_i h_i^l(s) h_i^l(t) / M}$$

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

$$\boxed{\tilde{q}^l := \textstyle\sum_i (\delta_i^l(t))^2} \qquad \boxed{\tilde{q}_{st}^l := \textstyle\sum_i \delta_i^l(s) \delta_i^l(t)}$$

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

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

- 非常に大きい孤立した最大固有値 $O(M)$

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

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

- 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

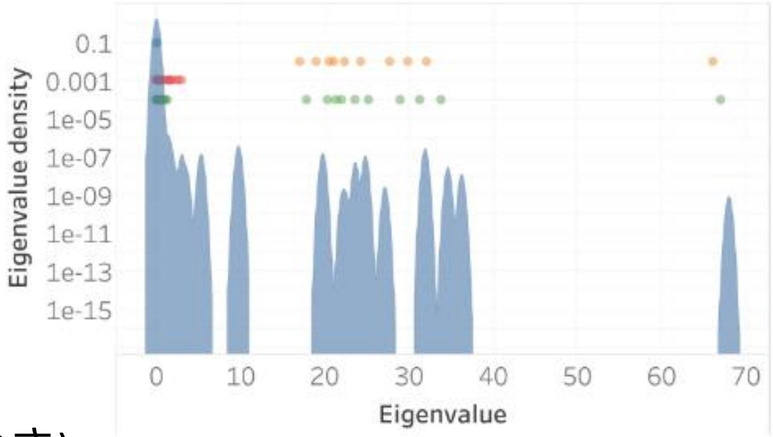
M (width) が十分に大きいとき, 漸近的に

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

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]
  [Ghorbani+, ICML '19]
  [Papyan, ICML '19]

$\underbrace{\hspace{3em}}_{C}$



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

Fisher情報行列の固有値

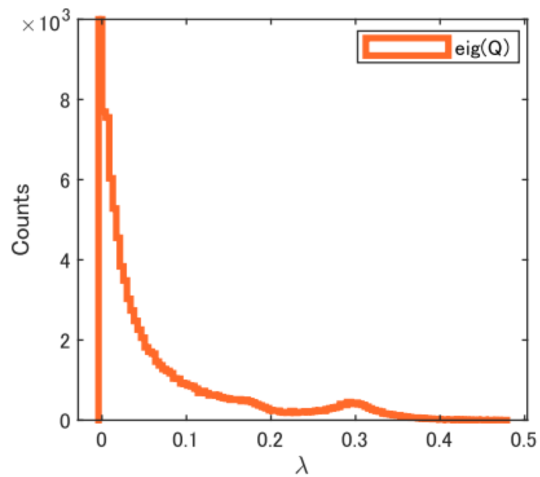
- Cross-entropy lossの場合 [RK, Akaho & Amari, arXiv:1910.05992]

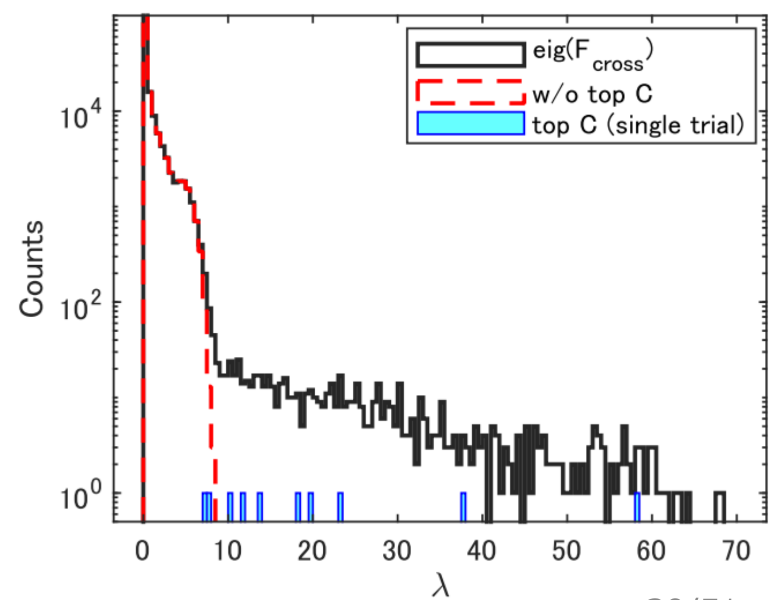
  上からC個の固有値は$O(M)$

FIM: $\underset{P}{\overset{CN}{\nabla_\theta f}} \; Q \; \nabla_\theta f^\top$

Sigmoid関数gはブロック対角行列Qに現れる

$$Q_n = \mathrm{diag}(g(n)) - g(n)g(n)^\top \quad (n = 1, \dots, N)$$





最大固有値と学習率

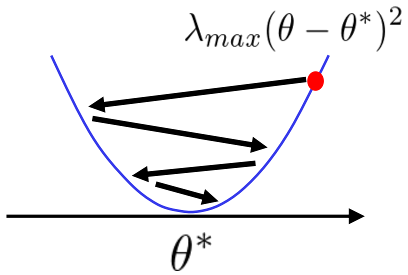
- 最急勾配 (バッチ学習) $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \underline{\eta} \dfrac{\partial E(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$

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**

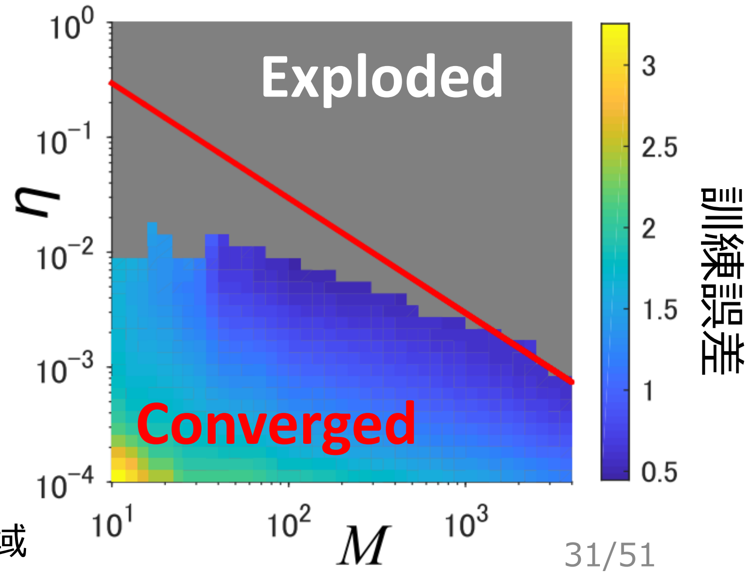
$$\boxed{\eta < 2/\lambda_{max}}$$

[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

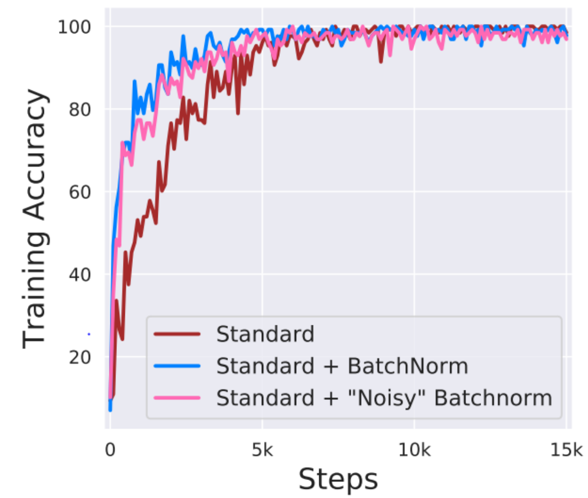
Batch Normalization (BN)

[Ioffe&Szegedy ICML '15]

各ニューロンをサンプルに対して正規化

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

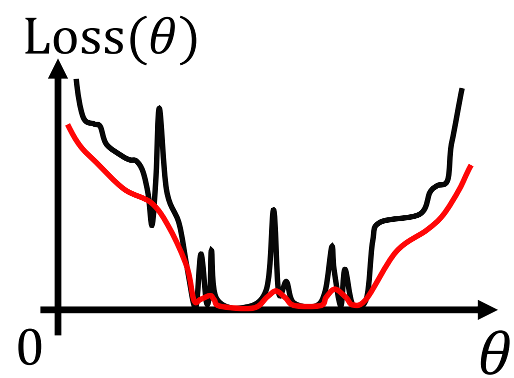
$$\mu_i^l := \mathrm{E}[u_i^l(t)] \qquad \sigma_i^l := \sqrt{\mathrm{E}[u_i^l(t)]^2 - (\mu_i^l)^2}$$



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 assumption

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

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

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

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

Batch norm in the last layer

簡単のため, **最終層のmean subtraction**のみを考える

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

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

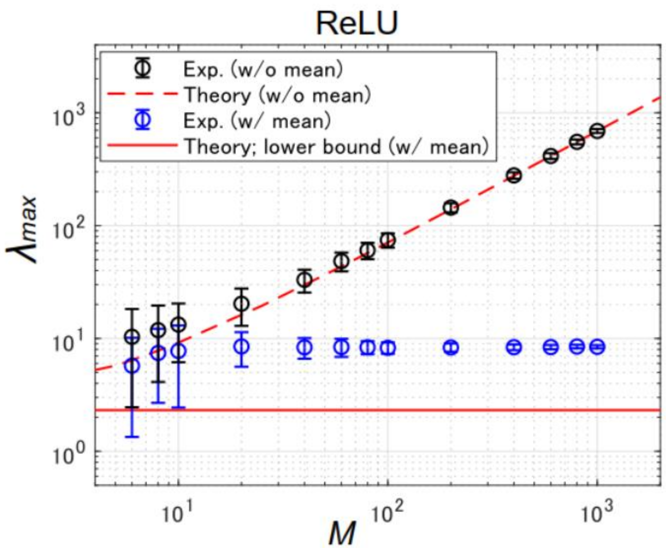
  理論の上限は保守的. 経験的には $\Theta(1)$

- 中間層のみにBN

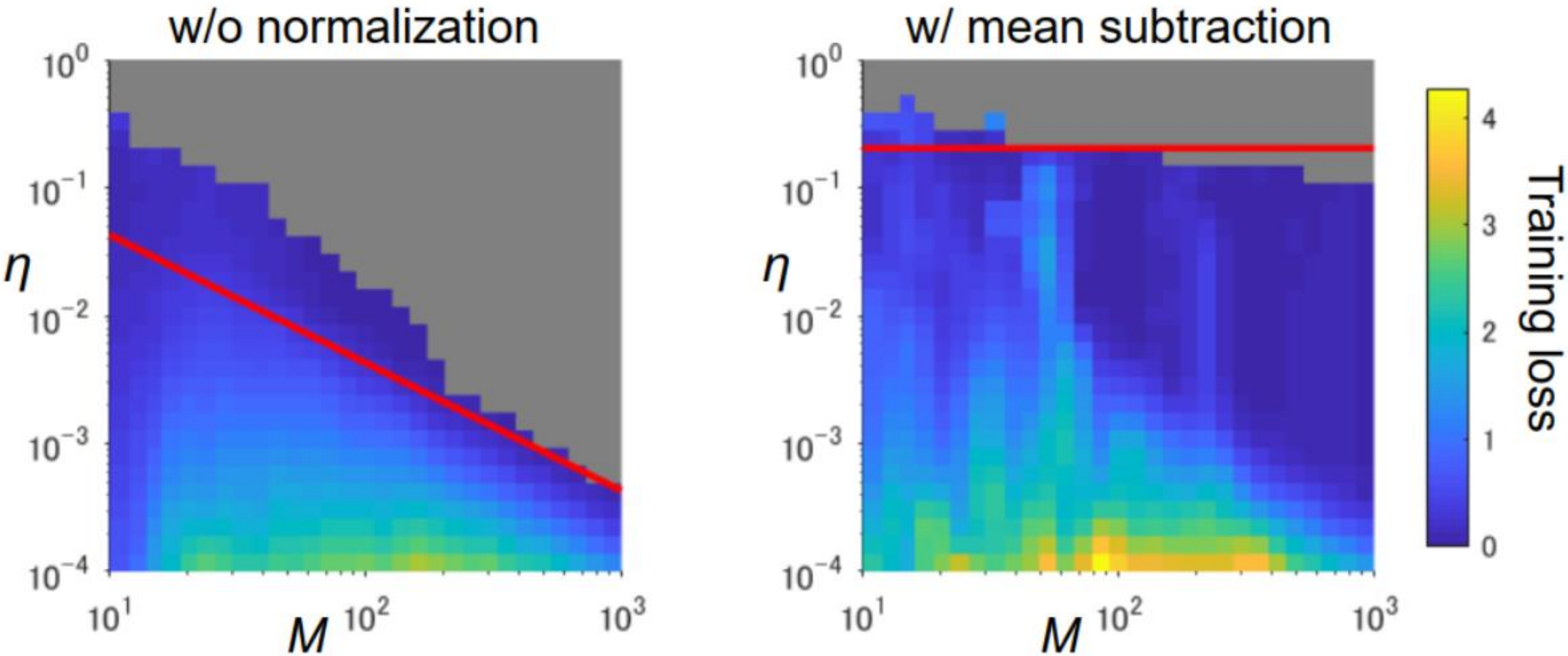
  Deep ReLU net では, $\lambda_{max} = \Theta(M)$

- 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

深層FIM(NTK)のスペクトル解析

[Fan & Wang, arXiv2005.11879 (NeurISP 2020)]

FIM = $\nabla_\theta f^\top \; \nabla_\theta f$ *NTK =*

非ゼロ固有値は同じ

$$\mathrm{NTK} = \sum_{l=1}^{L} B_l \odot A_{l-1}$$

$$= q_+ I + \sum_{l=1}^{L} \tilde{q}^l_{st} A_{l-1}$$

$$B_l = \delta_l \delta_l^\top = \begin{bmatrix} \tilde{q}^l & & \tilde{q}^l_{st} \\ & \ddots & \\ \tilde{q}^l_{st} & & \end{bmatrix}$$

中間層の共分散: $A_l = \dfrac{1}{M} H_l H_l^\top \quad H_l = \phi(W_l H_{l-1})$

$H_0 = X$

- **A_lのスペクトルはMarchenko–Pastur(MP)分布のfree convolutionに従う**

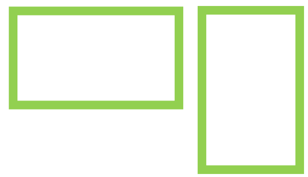
$$\mu_\ell = \rho^{\mathrm{MP}}_{\gamma_\ell} \boxtimes \left((1 - b_\sigma^2) + b_\sigma^2 \cdot \mu_{\ell-1} \right) \qquad b_\sigma = \int Dz \phi'(z)$$

関連研究 -A_1 のスペクトル[Pennigton& Worah,'17] [Laurart+ '17], Random feature回帰の bias-variance分解 [Hastie+, "Surprises … " (2019)])

\- $b_\sigma = 0$でMP則 [Benigni & Péché, arXiv1904.03090]

深層FIM(NTK)のスペクトル解析

[Fan & Wang, arXiv2005.11879 (NeurISP 2020)]

FIM = $\nabla_\theta f^\top \; \nabla_\theta f$ *NTK =*    非ゼロ固有値は同じ

- バイアス項なし, $\int Dz\phi(z) = 0$ (centered net)とする

$$\mathrm{NTK} = \sum_{l=1}^{L} B_l \odot A_{l-1} \qquad B_l = \delta_l \delta_l^\top = \begin{pmatrix} \tilde{q}^l & & \tilde{q}^l_{st} \\ & \ddots & \\ \tilde{q}^l_{st} & & \end{pmatrix}$$

$$= q_+ I + \sum_{l=1}^{L} \tilde{q}^l_{st} A_{l-1}$$

中間層の共分散: $A_l = \dfrac{1}{M} H_l H_l^\top \quad H_l = \phi(W_l H_{l-1})$

$$H_0 = X$$

- **NTKのStieltjes変換は** $m_{NTK}(z) = t_L\Big((-z + r_+, q_0, \ldots, q_{L-1}, 1), (1, 0, \ldots, 0)\Big)$

$$s_\ell(\mathbf{z}) = (1/z_\ell) + \gamma_\ell t_{\ell-1}\big(\mathbf{z}_{\text{prev}}(s_\ell(\mathbf{z}), \mathbf{z}),\ (1 - b_\sigma^2, 0, \ldots, 0, b_\sigma^2)\big),$$

$$t_\ell(\mathbf{z}, \mathbf{w}) = (w_\ell/z_\ell) + t_{\ell-1}\big(\mathbf{z}_{\text{prev}}(s_\ell(\mathbf{z}), \mathbf{z}),\ \mathbf{w}_{\text{prev}}\big)$$

$$\mathbf{z}_{\text{prev}}(s_\ell(\mathbf{z}), \mathbf{z}) \equiv \left(z_{-1} + \frac{1 - b_\sigma^2}{s_\ell(\mathbf{z})}, z_0, \ldots, z_{\ell-2}, z_{\ell-1} + \frac{b_\sigma^2}{s_\ell(\mathbf{z})}\right)$$

$$\mathbf{w}_{\text{prev}} \equiv (w_{-1}, \ldots, w_{\ell-1}) - (w_\ell/z_\ell) \cdot (z_{-1}, \ldots, z_{\ell-1})$$

Dynamical isometry成立下でのFIM

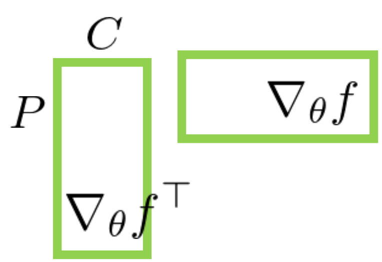
[Hayase & Karakida arXiv:2006.07814]

Input-output Jacobian:

$\frac{\partial h_L}{\partial h_0} = \prod_{l=1}^{L} D_l W_l$ の固有値が層数に非依存でも, (conditional)FIMは依存

Conditional FIM:

(given a single input *x*)



Dynamical isometryが成立するとき, 固有値は $\lambda_{\neq 0} \sim L$ に集中する

$\Theta_L(x,\theta) = \frac{1}{M}\frac{\partial f_\theta(x)}{\partial \theta}\frac{\partial f_\theta(x)}{\partial \theta}^\top$ に対して, $\Theta_{\ell+1} = \hat{q}_\ell I + W_{\ell+1} D_\ell \Theta_\ell D_\ell W_{\ell+1}^\top$

D_l, W_l の漸近自由独立性を仮定すれば, $\mu_{\ell+1} = (q_\ell + \sigma_{\ell+1}^2 \cdot)_*(\nu_\ell \boxtimes \mu_\ell)$

Free convolution

Dynamical isometry成立下でのFIM

[Hayase & Karakida arXiv:2006.07814]

Input-output Jacobian:

$$\frac{\partial h_L}{\partial h_0} = \prod_{l=1}^{L} D_l W_l$$ の固有値が層数に非依存でも, FIMは依存

Conditional FIM:

(given a single input *x*)

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

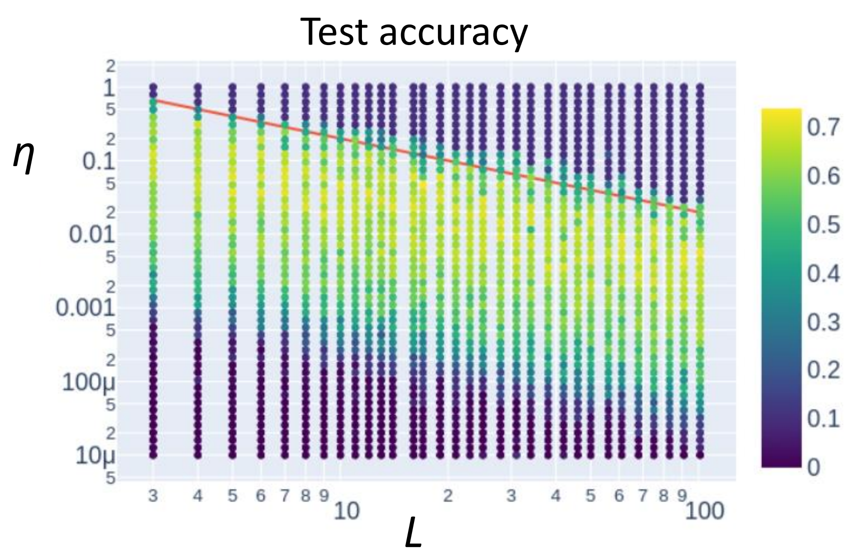
自由確率論でスペクトル解析.
固有値は集中して, $\lambda_{\neq 0} \sim L$

- 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



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
   - Loss landscapeと学習率
   - Batch Normalizationの役割

 - Neural Tangent Kernel
   - 近似自然勾配の高速な収束

[Karakida & Osawa, *Approximate Fisher Information for Fast Convergence of Natural Gradient Descent in Wide Neural Networks,* NeurIPS 2020]

Neural Tangent Kernel (NTK) 理論

[Jacot+ NeurIPS ’18]

設定

- Fully-connected DNN

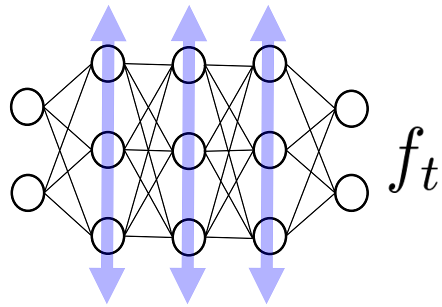
$$u_l = \frac{\sigma_w}{\sqrt{M_{l-1}}} W_l h_{l-1} + \sigma_b b_l, \quad h_l = \phi(u_l)$$

- Locally Lipschitz and non-polynomial ϕ, ϕ'

- Training samples (x_n, y_n) $(n = 1, ..., N)$, $y_n \in \mathbb{R}^C$

Normalized inputs $\|x_n\|_2 = 1$

Width $\to \infty$



- Gaussian random initialization

$$W_{l,ij}, b_{l,i} \sim \mathcal{N}(0, 1)$$

Neural Tangent Kernel (NTK) 理論

[Jacot+ NeurIPS '18]

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

$$\frac{df_t}{dt} = \eta \underbrace{\nabla_\theta f_t \nabla_\theta f_t^\top}_{=:\,\Theta_t} (y - f_t)$$

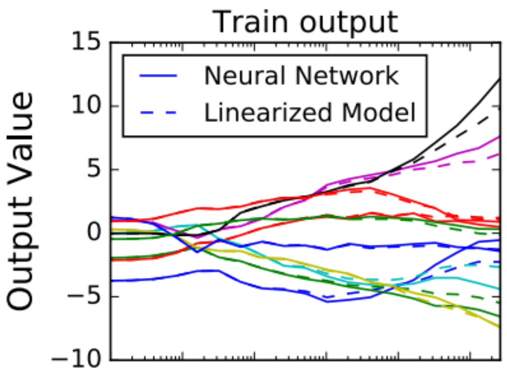
$=: \Theta_t$ NTK ($CN \times CN$ 行列)

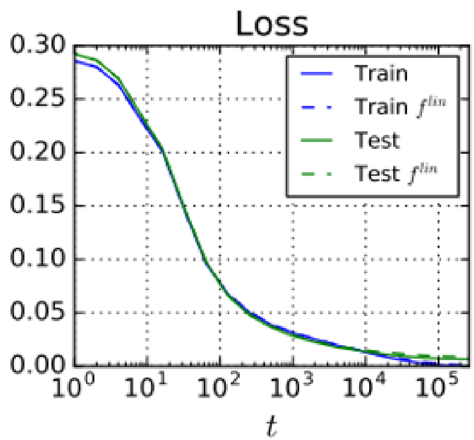
(Informal)

隠れ層幅が無限大のとき，f_tのダイナミクスは

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

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** $f_t^{\mathrm{lin}}(x) = f_0(x) + \underbrace{\nabla_\theta f_0(x)^\top}_{\sim M^{-1/2}} \underline{(\theta_t - \theta_0)}$

  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

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

- **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]

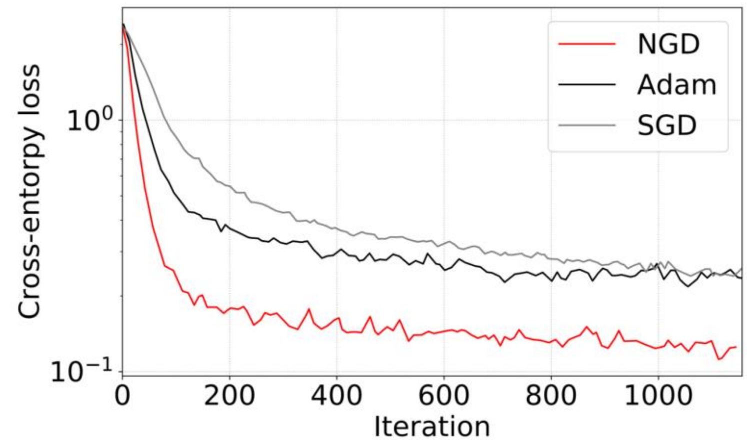
そのほかの収束証明の概要: [Zou & Gu, “An improved analysis of …”, NeurIPS ‘19]

NTK regimeにおける自然勾配法

- 自然勾配法 (Natural Gradient Descent) とは

$$\theta_{t+1} = \theta_t - \eta G_t^{-1} \nabla_\theta \mathcal{L}(\theta_t)$$

[Amari, ‘98]



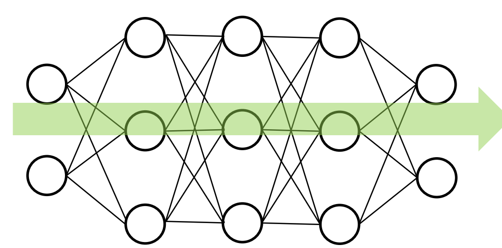
G_t : Fisher information matrix (FIM)

(=パラメータ空間のリーマン計量)

For MSE loss,

$$\mathrm{E}[\nabla_\theta f_t^\top \nabla_\theta f_t]$$

Parameter dim.
×Parameter dim.



f_t

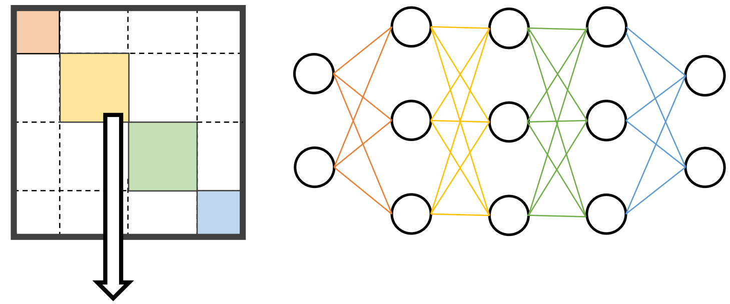
逆行列計算はコストが高く, そのままでは実用が難しい

Approximate Fisher Information Matrix (FIM)

FIMを近似して自然勾配法で利用. いくつかの方針.

- **Layer-wise FIM**
  Block diagonal, Block tri-diagonal, …

G_t :

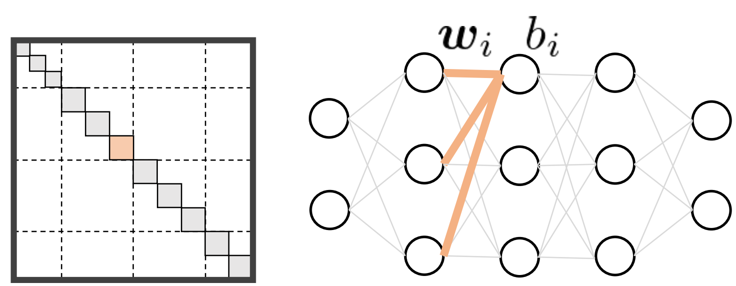


- **K-FAC** [Martens & Grosse, ’15] …

Kronecker-factored

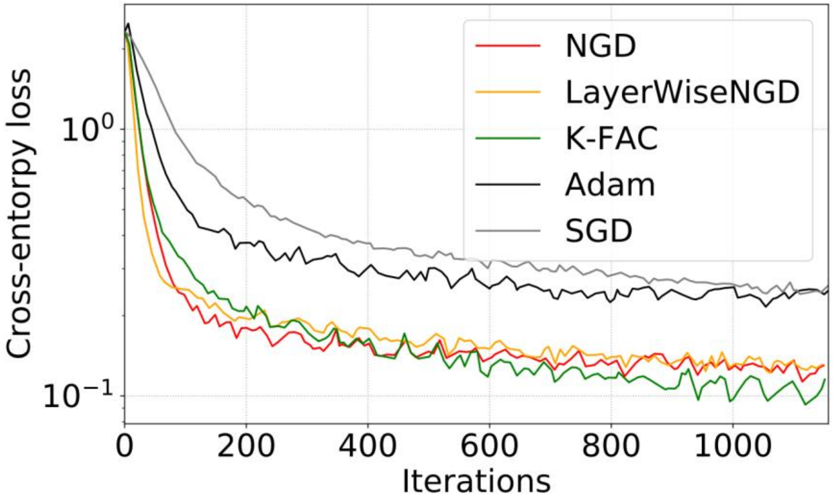
$$\mathrm{E}[(\delta_l^\top \delta_l) \otimes (h_{l-1}^\top h_{l-1})] \qquad \mathrm{E}[\delta_l^\top \delta_l] \otimes \mathrm{E}[h_{l-1}^\top h_{l-1}]$$

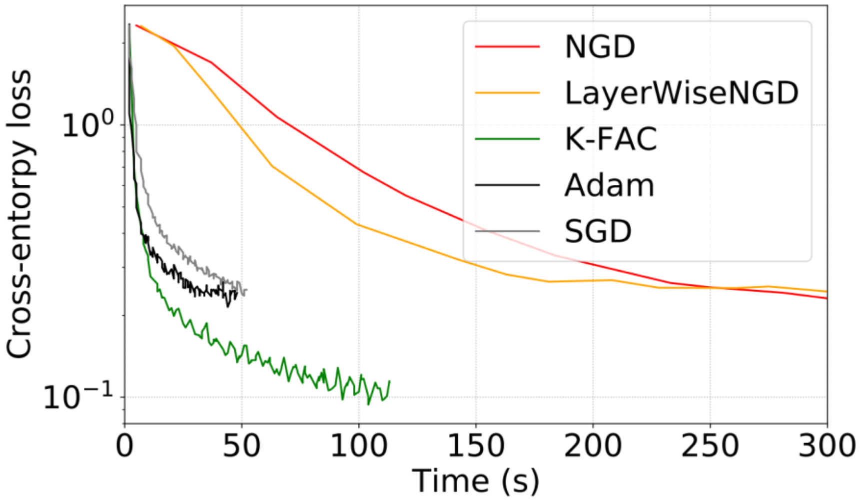
- **Unit-wise FIM**
  [Le Roux+, ’08] [Ollivier, ’15] [Amari+, ’19]



Approximate Fisher Information Matrix (FIM)

実験例





NTK regimeにおける自然勾配法

計量G_tが以下の条件を満たすとする

- **Isotropic Condition on** $\bar{\Theta} = J_0 G_0^{-1} J_0^\top / N \qquad (J_t = \nabla_\theta f_t)$

$$\bar{\Theta} = \alpha I \quad (\alpha > 0)$$

関数空間の勾配 $\frac{df}{dt} = \bar{\Theta}(y - f_t)$

- **Local Lipschitzness**

$$||G_t^{-1} J_t - G_0^{-1} J_0||_2 \lesssim ||\theta_t - \theta_0||_2 / \sqrt{M}$$

Implies $G_t^{-1} J_t \sim G_0^{-1} J_0$

このとき,

NTK dynamics of NGD (informal)

$$f_t(x') = \bar{\Theta}(x', x)\bar{\Theta}^{-1}(I - (I - \eta\bar{\Theta})^t)(y - f_0) + f_0(x')$$

$$\bar{\Theta}(x', x) := J_0(x') G_0^{-1} J_0(x)^\top / N$$

特に, training sampleにおいて, $f_t = y + (1 - \eta\alpha)^t (f_0 - y)$

NTK regimeにおける自然勾配法

(条件1) Isotropic Condition on $\bar{\Theta} = J_0 G_0^{-1} J_0^\top / N$
(条件2) Local Lipschitzness $||G_t^{-1} J_t - G_0^{-1} J_0||_2 \lesssim ||\theta_t - \theta_0||_2 / \sqrt{M}$

NTK dynamics of NGD (informal)

$$f_t(x') = \bar{\Theta}(x', x)\bar{\Theta}^{-1}(I - (I - \eta\bar{\Theta})^t)(y - f_0) + f_0(x')$$

$$\bar{\Theta}(x', x) := J_0(x')G_0^{-1}J_0(x)^\top / N$$

特に, training sampleにおいて, $f_t = y + (1 - \eta\alpha)^t(f_0 - y)$

- $0 < \eta\alpha < 2$ において大域収束. $\eta = 1/\alpha$なら1 step 収束.
- **FIM, Layer-wise FIM, K-FAC (C=1), unit-wise FIM (C=1)は条件1, 2を満たす** (証明は個別)

FIM:	α = 1	Layer-wise block diagonal NGD:	α = L
K-FAC:	α = NL	Unit-wise NGD:	α = M(L-1)/2 (for ReLU)

学習率ηを適切にスケールすれば, 近似FIMを使ったNGDは, 近似無しのNGDと全く同じ訓練ダイナミクス

NGD for over-parameterized models

- **Pseudo-inverseを使った勾配計算**

通常の自然勾配法の場合:

$$G_t = \frac{1}{N} J^\top J + \rho I$$

Parameter dim.

CN (output dim. × sample size) $J := \nabla_\theta f_t$

$$\Delta\theta = G_t^{-1} \nabla_\theta L(\theta_t)$$
$$= J^\top (J J^\top)^{-1} (f_t - y)$$

- $\nabla_\theta \mathcal{L} = J^\top (y - f_t)/N$
- **Push-through identity**
  $(J^\top J + \rho I)^{-1} J^\top = J^\top (J J^\top + \rho I)^{-1}$
- **Zero damping limit** $\rho \to 0$

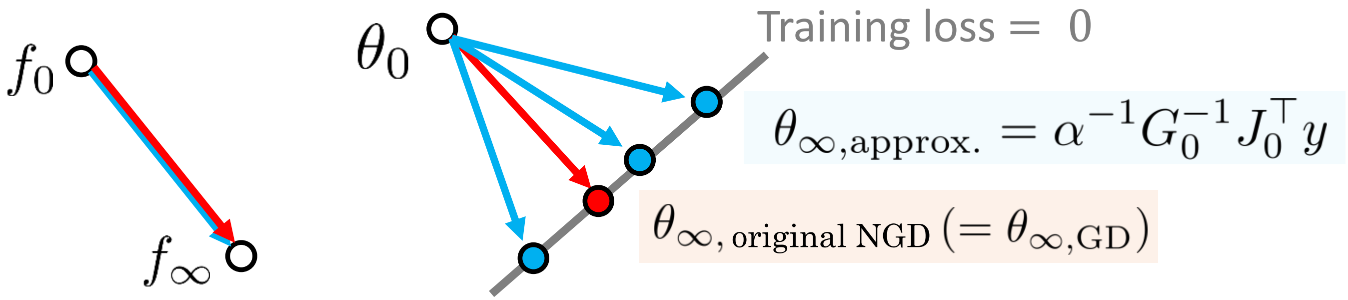
$$\Delta f = J \Delta\theta = f_t - y$$

近似自然勾配法の場合も本質的に同じ操作:

J : ∇f_t for each layer or unit, $\delta_{l,t}$ & $h_{l,t}$ for K-FAC

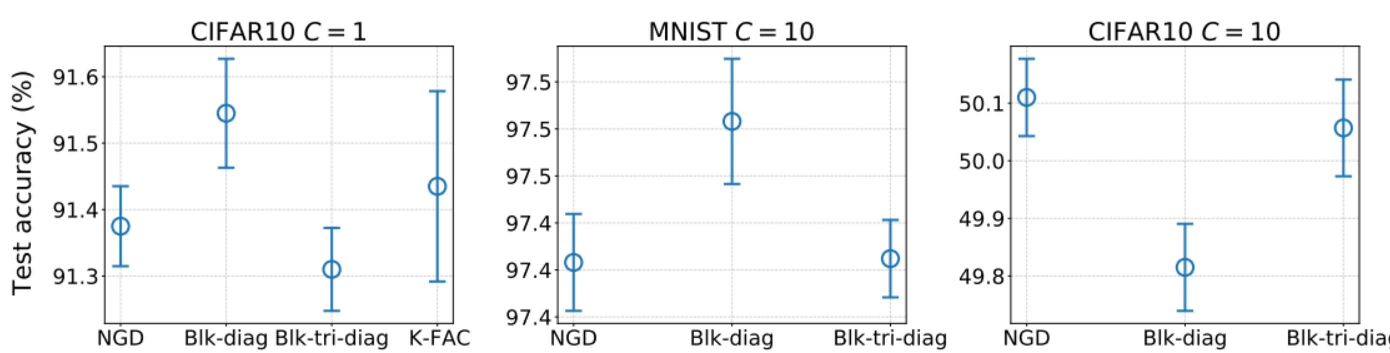
近似自然勾配の違い

- 関数としての訓練ダイナミクスは同じだが, パラメータ空間では一般には異なる



- 汎化性能の数値実験

Kernel regression: $\alpha^{-1}\bar{\Theta}(x', x)y$



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

・勾配の発散/消失を防ぐ初期値 (自由確率論が活躍)
・FIM
・NTK regime

今後の課題

- ランダム深層NTK
  - スペクトル形状の理解 (最大・最小固有値; NTK regimeの成立は $\lambda_{min} > 0$)
    層数の効果？ Batch normの効果？

  深層Random feature回帰の汎化誤差解析

$$\min_{\theta} ||y - \theta^{\top} h_L(x)||^2 + \lambda ||\theta||^2 \qquad \min_{\theta} ||y - \theta^{\top} \nabla_{\theta} f(x)||^2 + \lambda ||\theta||^2$$

- NTK regime での各種アルゴリズムの解析
- NTK regimeの外側　ランダム？特徴抽出？