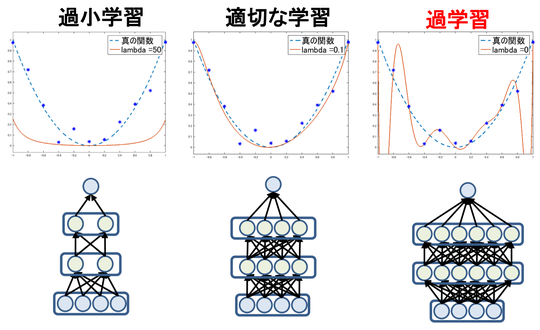
深層学習理論チーム
Deep Learning Theory Team

鈴木 大慈
Taiji Suzuki
RIKEN

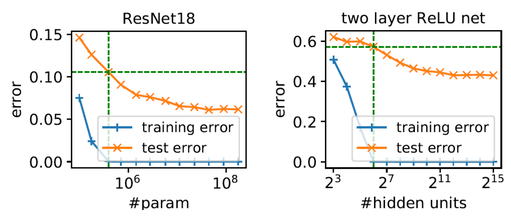
チームの目標:
- 深層学習の理論的理解の促進
- 理論に基づいた新しい手法の開発
- 関連する機械学習問題の解法と最適化手法の開発

深層学習の圧縮型汎化誤差バウンド

古典的な学習理論



実際は...

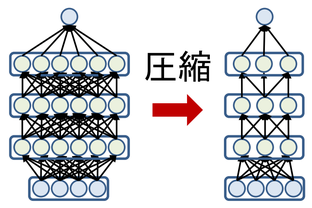


[Neyshabur et al., ICLR2019]

ネットワークのサイズを大きくしても過学習しない

データサイズ:130万
モデルパラメータサイズ:10億

圧縮型バウンド:
「圧縮」できるネットワークは過学習しない



[Taiji Suzuki: Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network, ICLR2020]

定理（一般論）

$$\Psi(\widehat{f}) \le \hat{\Psi}(\widehat{f}) + \sqrt{\frac{t}{n}(\hat{r}^2 + r_*^2)} + \dot{R}_{\hat{r}+r_*}(\widehat{\mathcal{F}} - \mathcal{F}^\sharp) + \bar{R}_n(\mathcal{F}^\sharp) + \sqrt{\frac{2t}{n}}$$

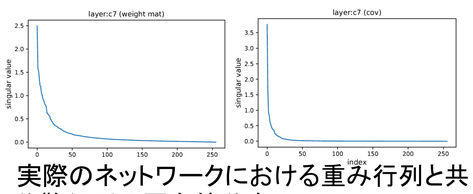
NNモデル: $f(x) = (W^{(L)}\eta(\cdot)) \circ (W^{(L-1)}\eta(\cdot)) \circ \cdots \circ (W^{(1)}x)$

中間層の重み行列・分散共分散行列がほぼ低ランクなら圧縮可能:

$$\sigma_j(W^{(\ell)}) \le C_1 j^{-\alpha} \qquad \sigma_j(\widehat{\Sigma}^{(\ell)}) \le C_2 j^{-\beta}$$

汎化誤差バウンド

$$\Psi(\widehat{f}) \le \hat{\Psi}(\widehat{f}) + O\left(\left(L\frac{\sum_{\ell=1}^{L} m_\ell}{n}\log(n)\right)^{\frac{2\alpha}{1+2\alpha}} + \sqrt{L^{1+\delta}\frac{\left(\sum_{\ell=1}^{L} m_\ell\right)^{\frac{4/\beta}{4/\beta+2(1-1/2\alpha)}}}{n}\log(n)^3}\right)$$



実際のネットワークにおける重み行列と共分散行列の固有値分布 (VGG-19, CIFAR10)

圧縮されたネットワークの大きさ (統計的自由度)

layer	In/Out	Orig	Arora et al. (2018)	Cov		
				$\nu: 10^{-1}$	10^{-2}	10^{-3}
c0	3/64	1,728	1,645	135	567	1,377
c3	128/128	147,456	644,654	891	75,240	147,456
c5	256/256	589,824	3,457,882	1,989	394,200	589,824
c8	256/512	1,179,648	36,920	540	10,395	91,719
c11	512/512	2,359,296	22,735	36	756	3,312
c14	512/512	2,359,296	26,584	108	882	4,536

我々

[Jingling Li, Yanchao Sun, Ziyin Liu, Taiji Suzuki and Furong Huang: Understanding of Generalization in Deep Learning via Tensor Methods. AISTATS2020]

→「**テンソル分解**」によるCNNの圧縮と詳細な理論評価も実現.

再生核ヒルベルト空間における確率的最適化の判別誤差に関する速い収束レート

[Nitanda&Suzuki: Stochastic Gradient Descent with Exponential Convergence Rates of Expected Classification Errors. AISTATS2019]

判別問題において確率的勾配降下法(SGD)は判別誤差を指数関数的に減少させる.

判別誤差

$$\mathcal{R}(g) \overset{\text{def}}{=} \mathbb{E}_{(X,Y)\sim\rho}[\mathbb{1}(\mathrm{sgn}(g(X)) \neq Y)]$$

- 本当に小さくした量
- 0-1損失は最適化が難しい

凸損失誤差

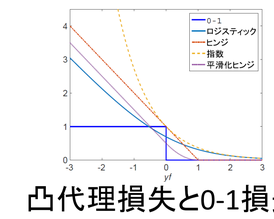
$$\mathcal{L}(g) \overset{\text{def}}{=} \mathbb{E}_{(X,Y)\sim\rho}[\ell(g(X), Y)]$$

- 実際に最小化する量
- ℓは凸関数→最適化が簡単

再生核ヒルベルト空間でのSGD (凸損失を最小化)

SGD on RKHS

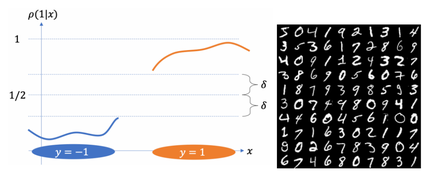
Input: 正則化係数λ, イテレーション数T, 学習率$\{\eta_t\}_{t=1}^T$, 重み$\{\theta_t\}_{t=1}^{T+1}$
Output: 識別器$\overline{g}_{T+1}$
1: initialize $g_1 \in \mathcal{H}$
2: for $t = 1, \ldots, T$:
3: sample $(x_t, y_t) \sim \rho$
4: $g_{t+1} \leftarrow g_t - \eta_t(\nabla\ell(g_t(x_t), y_t)k(x_t, \cdot) + \lambda g_t)$
5: return $\overline{g}_{T+1} = \sum_{t=1}^{T+1}\theta_t g_t$



凸代理損失と0-1損失

強低ノイズ条件

$$\exists\delta \in (0, 1/2), \quad |\rho(Y = 1|x) - 1/2| > \delta \quad (\rho_X\text{-a.s.})$$

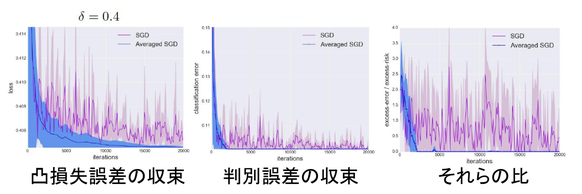


$\mathcal{L}(g_T) - \mathcal{L}(g^*) \lesssim \frac{1}{\sqrt{T}}$:凸損失の収束レート

定理 (収束レート解析)

$$\mathcal{R}(g_T) - \mathcal{R}(g^*) \lesssim \exp(-CT)$$

0-1損失の収束レート



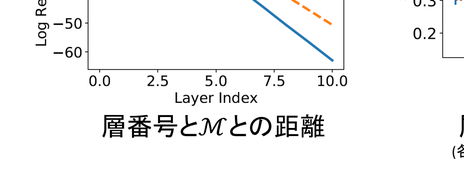
凸損失誤差の収束　判別誤差の収束　それらの比

0-1損失を考えると多項式オーダーが指数オーダーに改善

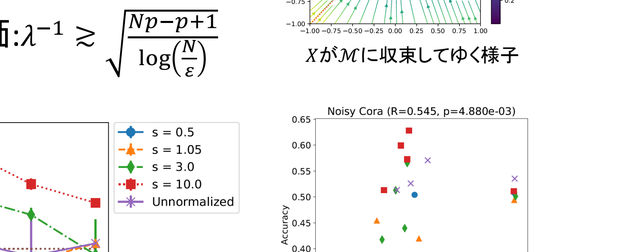
深層学習の近似理論

[Hayakawa&Suzuki: On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. Neural Networks, 2020]

深層学習が優れていることを示す様々な理論が存在:低次元特徴量, 不連続関数推定, Besov空間, ...
→ これらを「スパース性・非凸性」という統一的概念で説明することに成功.



深層学習はデータの低次元性をとらえ次元の呪いを回避できることを解明.



s_1, s_2, s_3: smoothness
(non-smooth) $s_1, s_2 \ll s_3$ (smooth)

推定精度

深層学習: $n^{-\frac{\tilde{s}}{\tilde{s}+1}}$, $\tilde{s} = (s_1^{-1} + s_2^{-1} + \cdots + s_d^{-1})^{-1}$

浅い学習: $n^{-\frac{s_1}{s_1+d}}$ (次元の呪い)

[Suzuki&Nitanda: Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space, arXiv:1910.12799.]

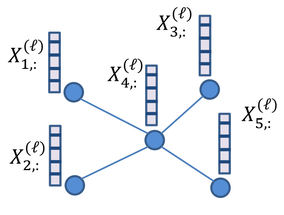
グラフ深層学習の過平滑化現象

[Oono and Suzuki: Graph Neural Networks Exponentially Lose Expressive Power for Node Classification, ICLR2020.]

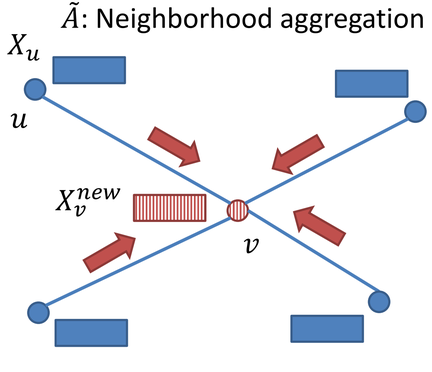
グラフ上の畳み込みNN (GCN) は層を重ねるごとにノード特有の情報を指数関数的に喪失し, ノードを見分けられなくなる.
→ ノード分類の精度が悪くなる.

GCNの更新式 (N:ノード数, C:チャネル数)

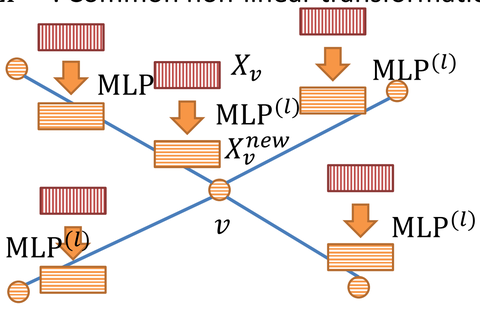
$$X^{(\ell+1)} \coloneqq \mathrm{MLP}^{(\ell)}(\tilde{A}X^{(\ell)})$$
$$X^{(\ell)} \in \mathbb{R}^{N\times C}$$



$\tilde{A}$: Neighborhood aggregation



$\mathrm{MLP}^{(l)}$: Common non-linear transformation



ノードの判別, リンク予測などに有用

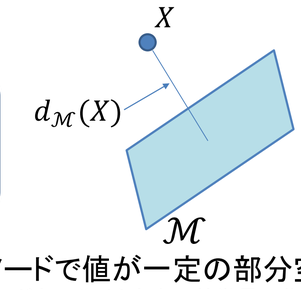
情報が指数関数的に減少

定理

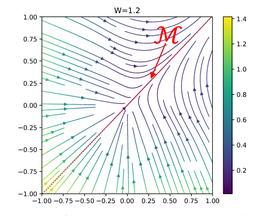
$$d_{\mathcal{M}}(X^{(\ell)}) \le (\prod_{k=1}^{\ell} s_{(k)})\lambda^\ell d_{\mathcal{M}}(X^{(0)})$$

$d_{\mathcal{M}}$: $\mathcal{M}$からの距離 (Frobenius norm)
$s_h^{(\ell)}$: $\mathrm{MLP}^{(\ell)}$の第h層の重み行列$W_h^{(l)}$の最大特異値
$s_{(\ell)} \coloneqq \prod_h s_{(\ell),h}$: それらの積 (リプシッツ連続性)
$\lambda \coloneqq 1 -$ (Aで決まる正規化ラプラシアン$\tilde{\Delta}$の第二固有値)

グラフが密に繋がっていればいるほど層を経るごとに情報が急速に失われる.

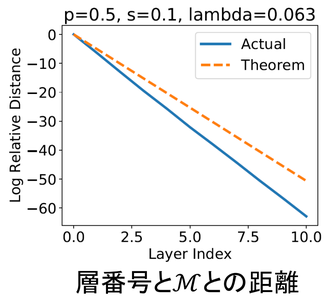


全ノードで値が一定の部分空間
(ノードを見分けられない特徴量の集合)

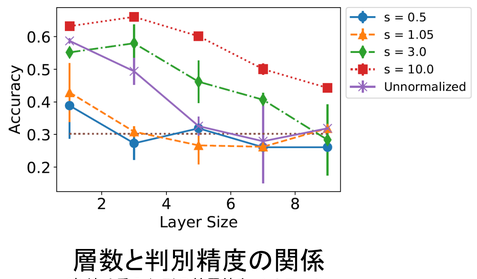


Xが$\mathcal{M}$に収束してゆく様子

ランダムグラフでλの値を理論的に評価:$\lambda^{-1} \gtrsim \sqrt{\frac{Np-p+1}{\log(\frac{N}{\varepsilon})}}$



層番号と$\mathcal{M}$との距離



層数と判別精度の関係
(各線は重み行列の特異値をコントロール)

$\mathcal{M}$と直交する成分の大きさと精度の関係