ニューラルネットワークの数理

Amuta (@Amuta151)

March 31, 2024

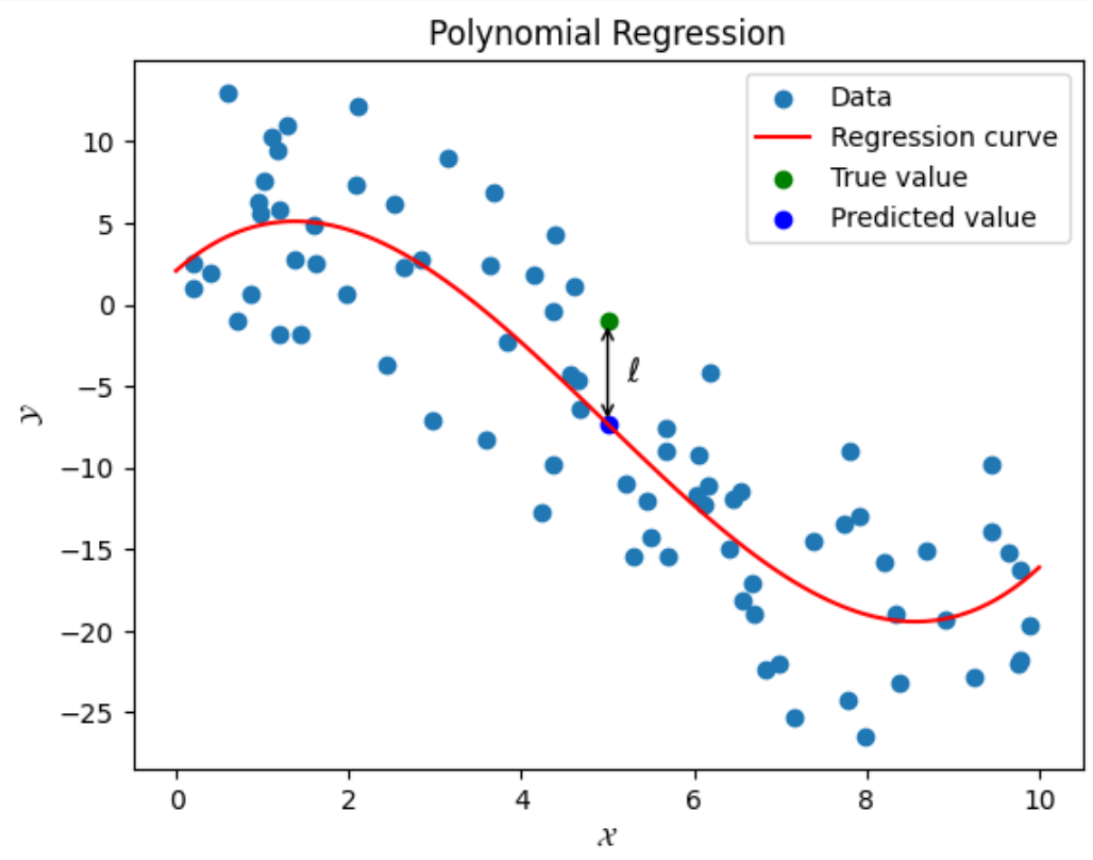
目次

Section 1：イントロダクション

目次

統計的学習（機械学習の統計解析）

$\mathcal{X}, \mathcal{Y}$ - ポーランド空間（可分かつ完備距離化可能な位相空間），
P - $\mathcal{X} \times \mathcal{Y}$ 上の確率測度,
$\mathcal{H}$ - モデル（$h : \mathcal{X} \to \mathcal{Y}$ なる可測測関数の集合），



統計的学習（機械学習の統計解析）

$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ - Lipschitz 連続関数（二つの $\mathcal{Y}$ の元の差を測る．損失（loss）関数）．$(X_1, Y_1), \cdots, (X_n, Y_n) \sim P\,(\text{i.i.d})$ を真の分布 P に従う確率変数であるとする．

経験誤差

$h \in \mathcal{H}$ に対して，

$$\hat{L}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, h(X_i)),$$

を経験誤差という．

統計的学習（機械学習の統計解析）

$\mathcal{X}, \mathcal{Y}$ - ポーランド空間,
P - $\mathcal{X} \times \mathcal{Y}$ 上の確率測度,
$\mathcal{H}$ - モデル（$h : \mathcal{X} \to \mathcal{Y}$ なる可測測関数の集合）,
$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ - Lipschitz 連続関数（二つの $\mathcal{Y}$ の元の差を測る．損失（loss）関数）．

汎化誤差

$h \in \mathcal{H}$ に対して，

$$L(h) = \mathbb{E}[\ell(Y, h(X))] = \int \ell(y, h(x)) dP(x, y),$$

を h の汎化誤差という．

統計的学習（機械学習の統計解析）

Example (二乗誤差を考えるとき)
$\mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \mathbb{R}$ とする．

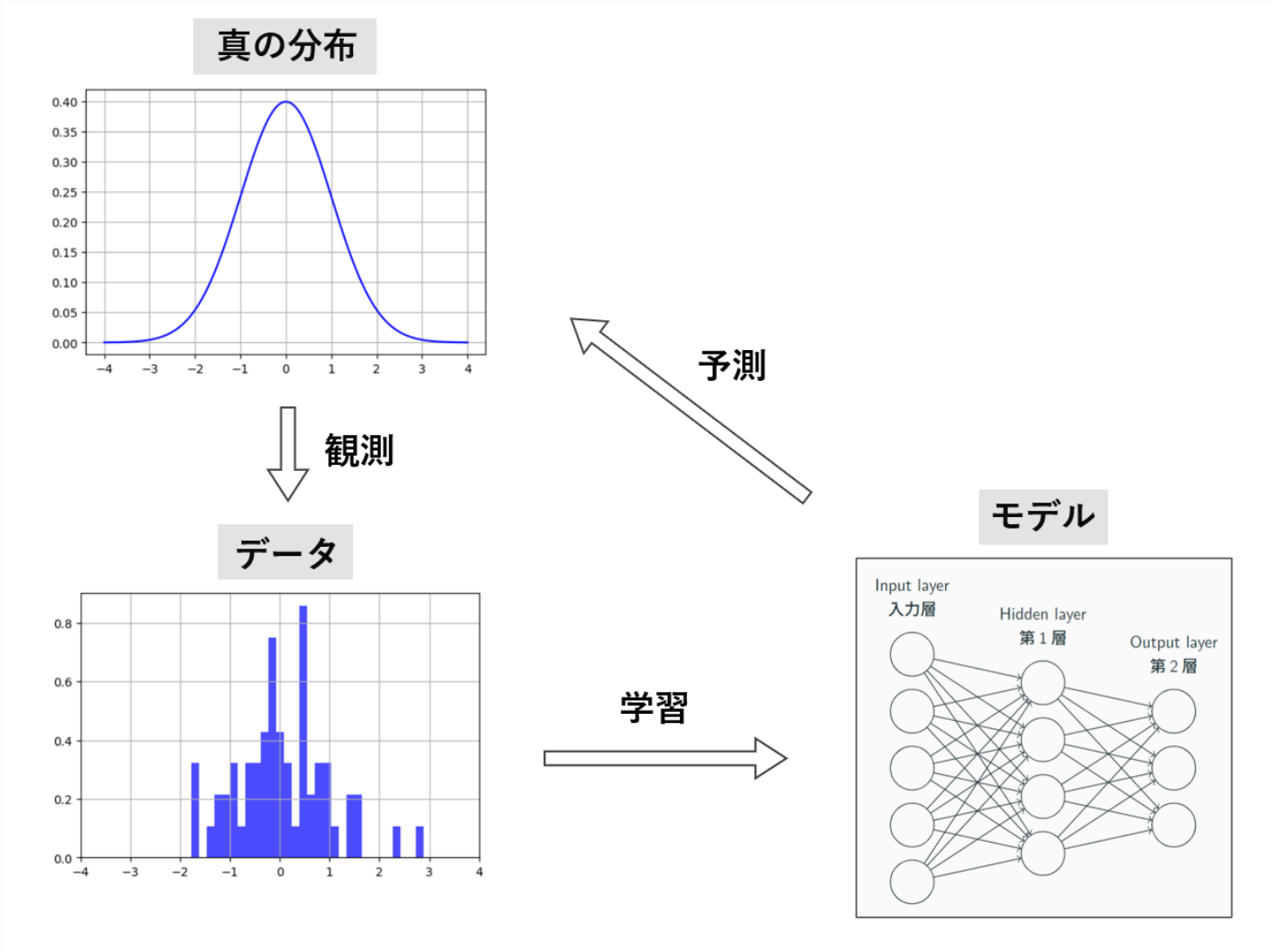
$\ell(y_1, y_2) = (y_1 - y_2)^2$ のとき，

$$\begin{aligned} L(h) &= \mathbb{E}[(Y - h(X))^2], \\ &= \int (y - h(x))^2 dP(x, y), \\ &= \|Y - h(X)\|_{L^2}^2, \end{aligned}$$

が成り立つ．いわゆる平均二乗誤差 (MSE)．

統計的学習（機械学習の統計解析）

おおざっぱなイメージを描くと……



統計的学習（機械学習の統計解析）

汎化誤差 $L(h)$ ： 未観測データも考慮した当てはまりの良さ．
経験誤差 $\hat{L}(h)$ ： すでに観測されたデータへの当てはまりの良さ．

- 汎化誤差を直接小さくすることは真の分布を知らないと不可能なので，経験誤差を用いて間接的に汎化誤差の最小化を目指す．
- 経験誤差を用いた何らかの学習アルゴリズム（NN なら確率的勾配降下法）で汎化誤差がどの程度小さくなるかを評価したい → 汎化誤差解析．

統計的学習（機械学習の統計解析）

複雑性誤差

$$L(h) = \hat{L}(h) + (L(h) - \hat{L}(h)),$$

と分解したとき $V(h) = L(h) - \hat{L}(h)$ を複雑性誤差という．

一様収束誤差

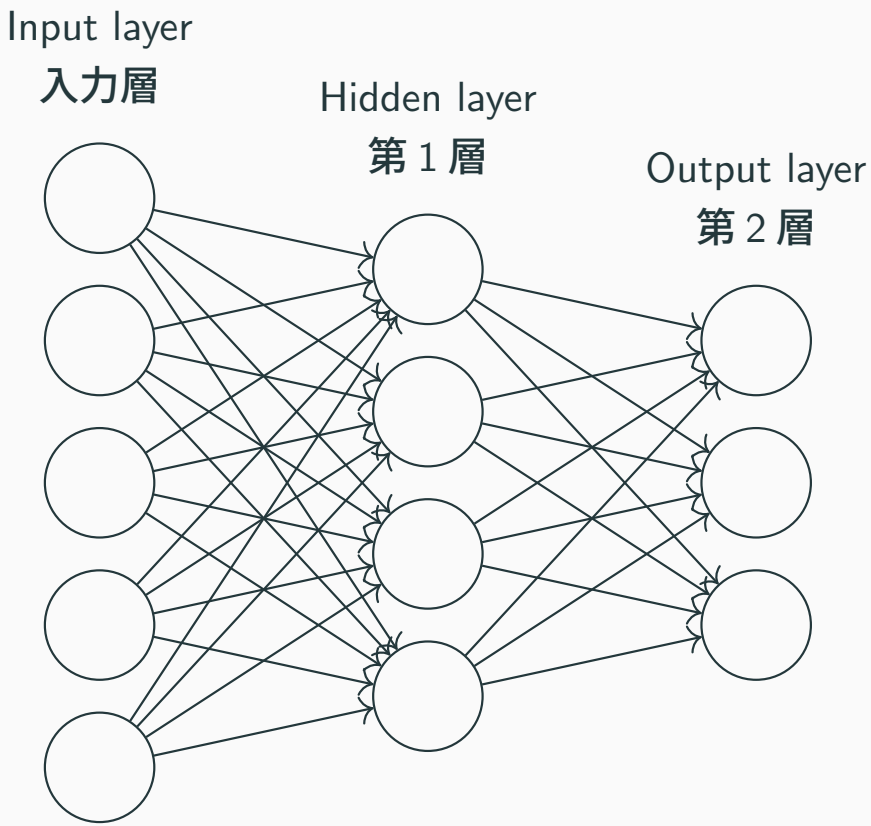
$\sup_{h \in \mathcal{H}} V(h)$ を一様収束誤差という．

- $\hat{L}(h)$ は特に NN では十分小さくできるため問題にならない．
- 一様収束誤差は $\hat{L}(h)$ を無視すればどのような学習結果でも汎化誤差を上から評価できる．

目次

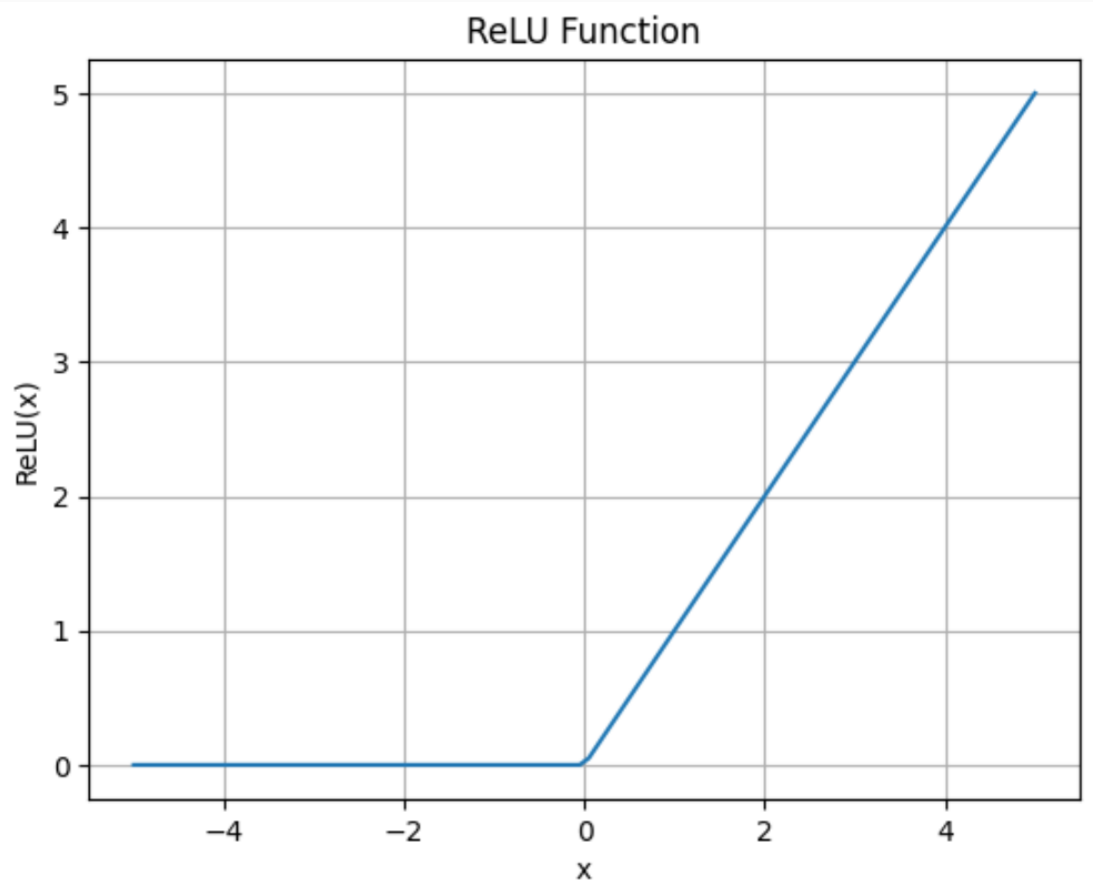
ニューラルネットワーク

二層ニューラルネットワークのイメージ図．入力次元を D，出力次元 d として関数 $f : \mathbb{R}^D \to \mathbb{R}^d$ を表す．１つのノードが１つの次元を表している．



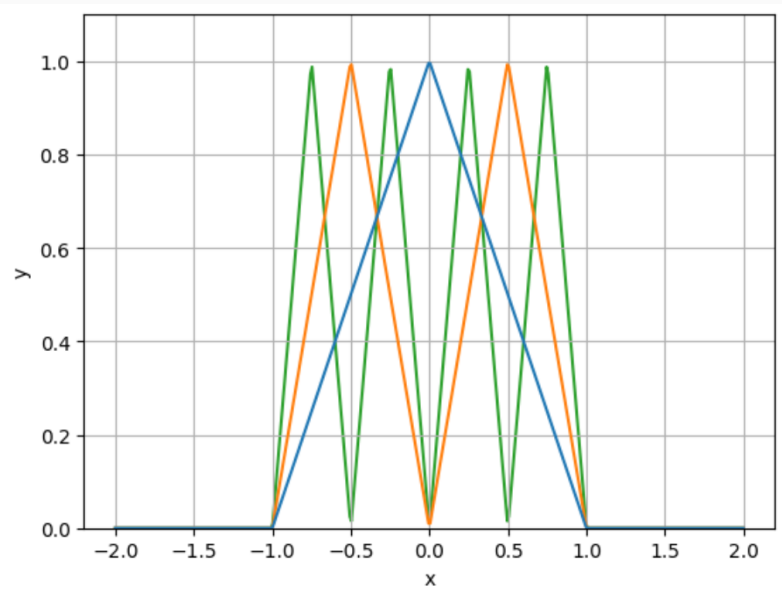
ニューラルネットワーク

代表的な活性化関数である ReLU．Yarotsky 2017 では $ReLU$ が二次関数をよく近似する理由について考察し，ReLU 関数の有効性に関して重要な示唆を与えた．



ニューラルネットワーク

ReLU の組み合わせで二次関数を近似する様子．参考：Yarotsky 2017．



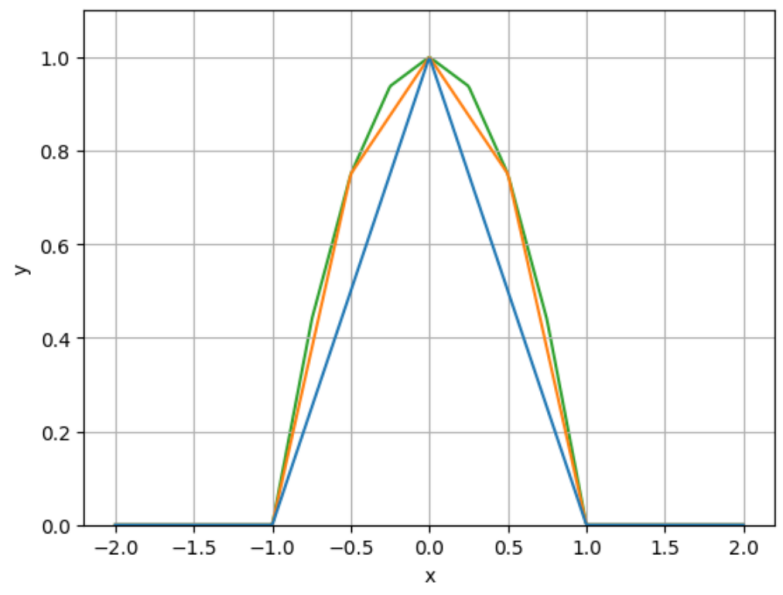
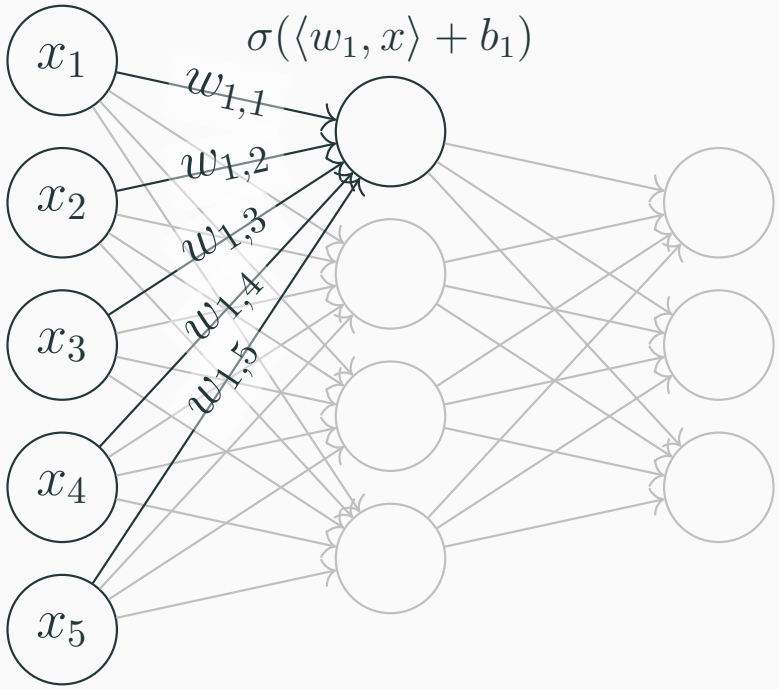


Figure 1: ReLU による二次関数の近似

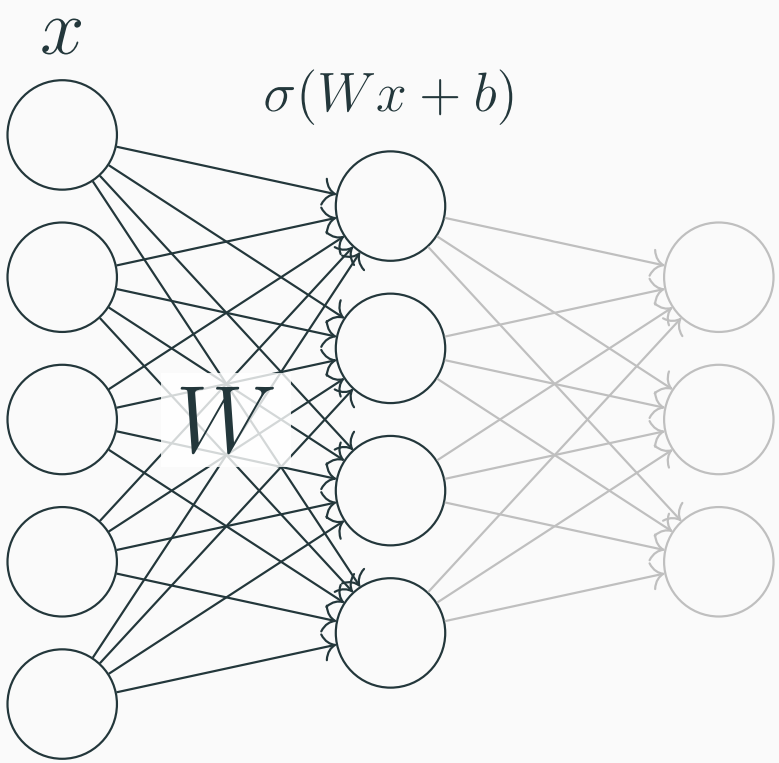
ニューラルネットワーク

入力 $x \in \mathbb{R}^D$ に対しパラメータ $(w_{i,j})$ で重みづけをし，活性化関数 $\sigma : \mathbb{R} \to \mathbb{R}$ を用いる.



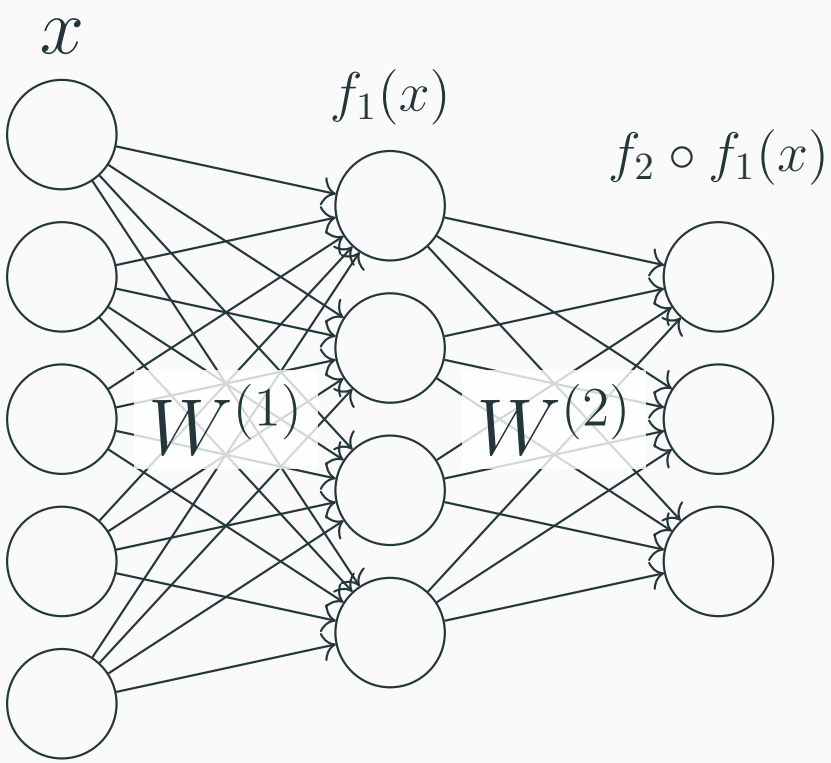
ニューラルネットワーク

パラメータからなる行列 W を用いると簡潔に表せる．ただし，$\mathbb{R}^n$ の元に活性化関数 σ を用いるときは要素ごとに作用させることを表すものとする：$\sigma(x) = (\sigma(x_1), \cdots, \sigma(x_n))^\top$.



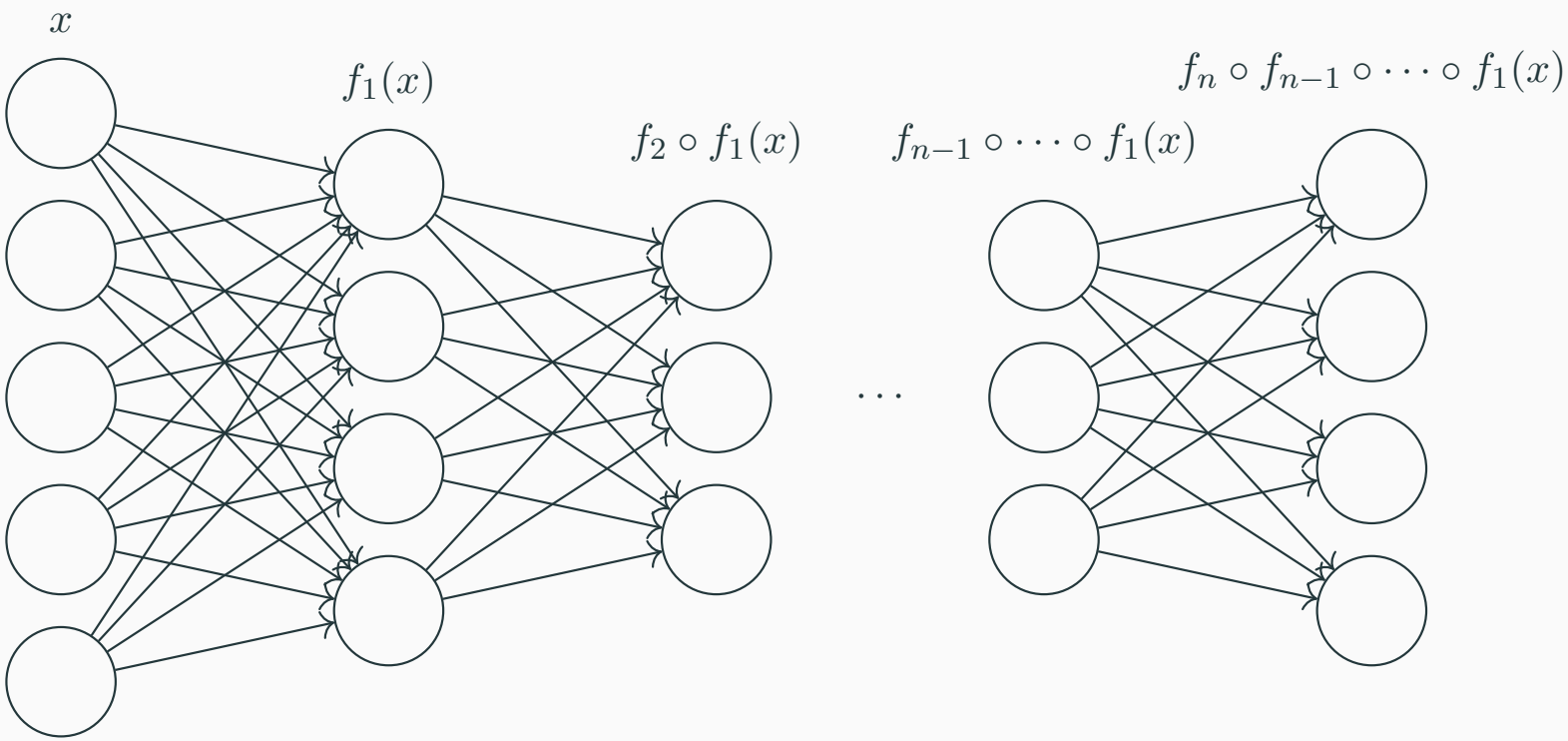
ニューラルネットワーク

第 i 層の値を f_i とする．$i = 1$ のときには $z = x$，$i > 1$ のときには $z = f_{i-1}$ として，$f_i(z) = \sigma^{(i)}(W^{(i)}z + b^{(i)})$ とする．ここで $\sigma^{(i)}$ は各層ごとの活性化関数を表し，特に出力層の活性化関数は恒等関数とする．



ニューラルネットワーク

深層ニューラルネットワークについては第 3 層以上の層を考えればよい．



ニューラルネットワークの定義

非負整数 n に対し，任意の $i \in \{1, \cdots, n-1\}$ について活性化関数という非線形関数 $\sigma^{(i)} : \mathbb{R} \to \mathbb{R}$ が定まっているとする．$i = n$ のとき便宜上 $\sigma^{(n)}$ を恒等関数であるとする．

また，d_i を第 i 層のノード数とする．ただし，第 0 層は入力層とする．

ニューラルネットワーク

n を非負整数とする．$1 \leq i \leq n$ に対し $z \in \mathbb{R}^{d_{i-1}}$ とするとき，$f_i(z) = \sigma^{(i)}(W^{(i)}z + b^{(i)})$ と定める．このとき，n 層からなるニューラルネットワークを，

$$f_n \circ f_{n-1} \circ \cdots \circ f_1(x),$$

で定義する．

目次

確率的勾配降下法

勾配降下法のイメージ：

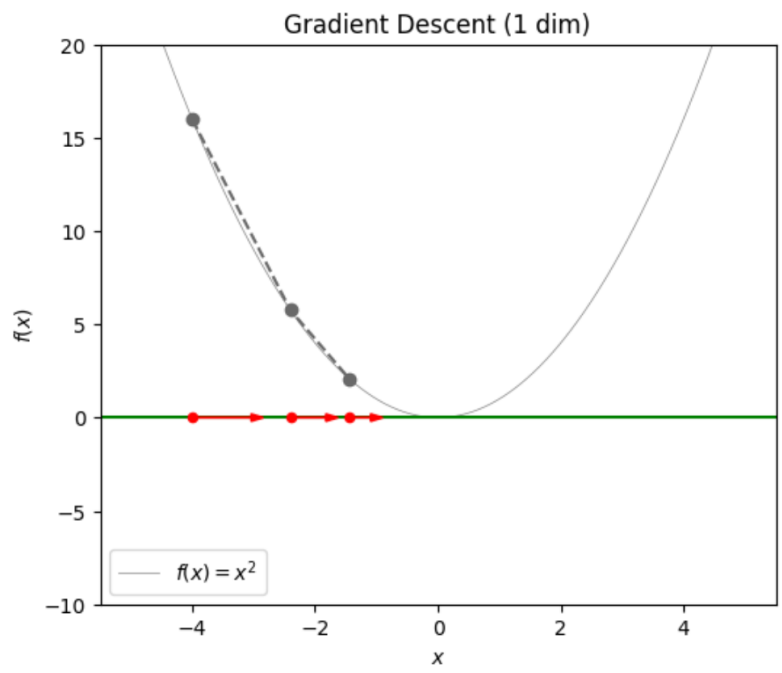


Figure 2: １次元の勾配流

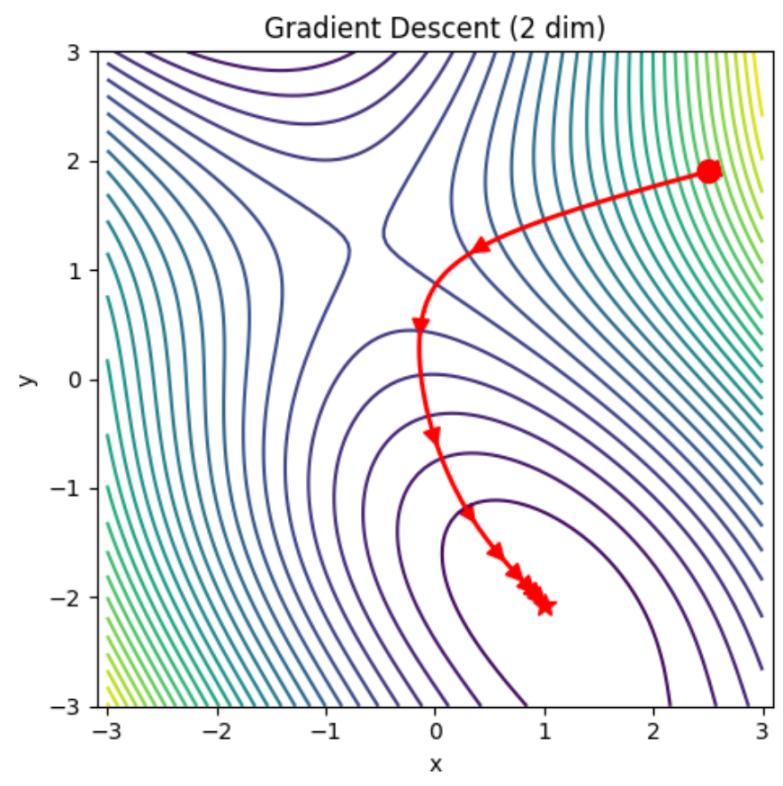


Figure 3: ２次元の勾配流

確率的勾配降下法

ニューラルネットワーク $f = f_n \circ f_{n-1} \circ \cdots \circ f_1$ に出現するパラメータ $(W^{(i)}, b^{(i)})$，$i = 1, 2, \cdots, n$ をまとめてパラメータ θ で表す．f は入力 x とパラメータ θ からなる関数 $f(x, \theta)$ である．

パラメータが θ_t のときの経験損失を $\hat{L}(f; \theta_t)$ と書くことにする．

確率的勾配降下法（Stochastic Gradient Descent; SGD）

パラメータ θ を次のように更新する：

$$\theta_{t+1} = \theta_t - \eta_t \nabla_{\theta_t} \hat{L}(f; \theta_t).$$

補足）NN を使ったことがある人向け：ミニバッチのサイズは解析手法，論文による．

NN の理論解析の現状

深層学習は様々な応用があり，非常に優れた汎化能力を持つことが知られている．

- 従来の機械学習ではモデルが複雑になりすぎることを防ぎ，与えられたデータに過剰に適応する過学習を防ぐことが重要であった．
- 一方，深層学習では極めて多くのパラメータを持つ（＝非常に複雑なモデルを使用している）にも関わらず過学習を起こしにくい性質を持つことが知られている．
- 現状，深層学習がなぜ過学習を起こしにくいのかを説明する完全な理論は整備されておらず，様々な手法が提案されている．
- 今日紹介するのはごく一部．

Section 2：NNの最適化の理論解析

目次

平均場近似による解析

Mei, Montanari, and Nguyen 2018 の紹介.

平均場近似による解析

出力層を一層目の平均とする二層ニューラルネットワークは次のように具体的に書くことができる．

$$\hat{y}(x;\theta) = \frac{1}{N}\sum_{i=1}^{N} \sigma_*(x;\theta_i),$$

$$\sigma(x;\theta_i) = a_i\sigma(\langle W_i, x\rangle + b_i),$$
$$\theta_i = (a_i, b_i, W_i) \in \mathbb{R}^D \, (i = 1, \cdots, N).$$

ただし N は中間層のノード数．

平均場近似による解析

パラメータ θ のときの二乗損失関数に関する汎化誤差は

$$L_N(\theta) = \mathbb{E}_{(X,Y)}[(Y - \hat{y}(X;\theta))^2],$$

となる．このとき，

$$L_N(\theta) = \mathbb{E}[Y^2] + \frac{2}{N}\sum_{i=1}^{N} V(\theta_i) + \frac{1}{N^2}\sum_{i,j=1}^{N} U(\theta_i, \theta_j),$$

$$V(\theta) = -\mathbb{E}[Y\sigma(X;\theta)],$$
$$U(\theta_1, \theta_2) = -\mathbb{E}[\sigma(X;\theta_1)\sigma(X;\theta_2)],$$

となる．

平均場近似による解析

汎化損失からの観察

中間層の数 N を増やしていくと，汎化損失は下側の積分の近似になってそうという予想が立つ．

$$L_N(\theta) = \mathrm{const} + \frac{2}{N}\sum_{i=1}^{N} V(\theta_i) + \frac{1}{N^2}\sum_{i,j=1}^{N} U(\theta_i, \theta_j),$$

$$L(\rho) = \mathrm{const} + 2\int V(\theta)\rho(d\theta) + \int U(\theta_1, \theta_2)\rho(d\theta_1)\rho(d\theta_2).$$

パラメータ θ がある確率測度 ρ に従う確率変数であるときのモンテカルロ積分．

平均場近似による解析

経験分布からの観察

ここで，$L_N(\theta)$ は $\theta_1, \cdots, \theta_N$ によっている．この θ_i の経験分布を，

$$\hat{\rho}(\theta) = N^{-1} \sum_{i=1}^{N} \delta_{\theta_i},$$

とする．

中間ノードの数を増やしたとき $\hat{\rho}$ が何らかの $\mathbb{R}^D$ 上の確率分布 $\rho \in \mathcal{P}(\mathbb{R}^D)$ に収束していないか？

平均場近似による解析

確率的勾配降下法からの観察

各パラメータ

$$\begin{aligned}\theta_i^{k+1} &= \theta_i^k - s_k \nabla_{\theta_i} L_N(\theta) \\ &= \theta_i^k + 2s_k(y_k - \hat{y}(x_k;\theta^k))\nabla_{\theta_i}\sigma_*(x_k;\theta_i^k),\end{aligned}$$

によって更新される (s_k は学習率).

この学習はパラメータ空間の中でのある種の"流れ"を記述しているのではないか？

平均場近似による解析

勾配降下法のイメージ：

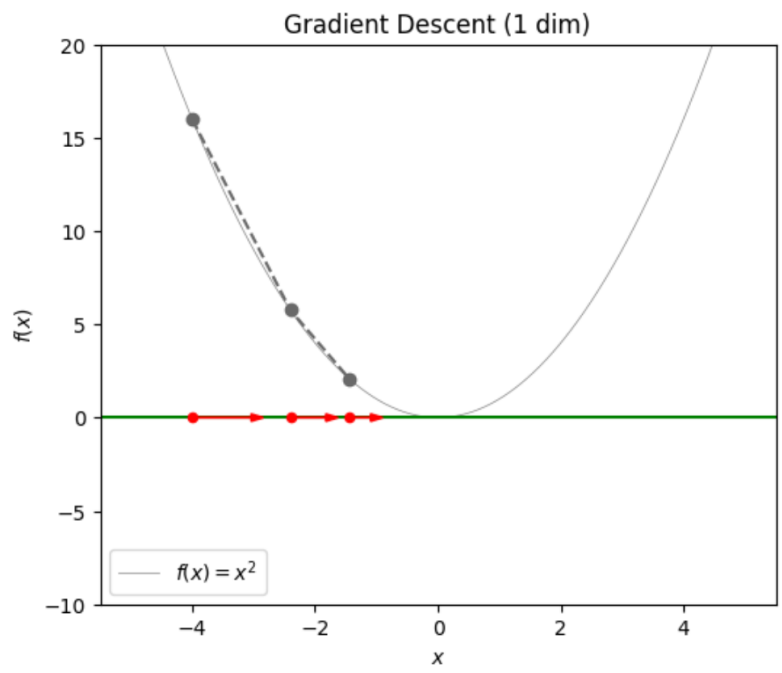


Figure 4: １次元の勾配流

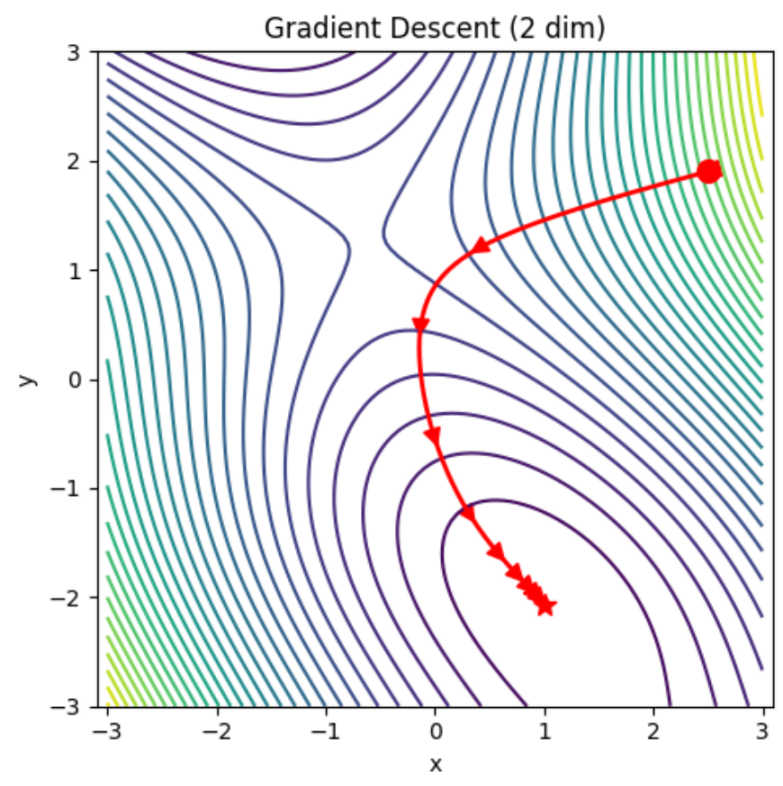


Figure 5: ２次元の勾配流

平均場近似による解析

連続の方程式

$\mathscr{P}(\mathbb{R}^D)$ を $\mathbb{R}^D$ 上の測度の集合，$\rho_t : (0,T) \to \mathscr{P}(\mathbb{R}^D)$ とする．$v_t : \mathbb{R}^D \to \mathbb{R}^D$ を $\int_0^T \int |v_t(x)| d\rho_t(x) dt < \infty$ なる Borel 可測なベクトル場とすると，

$$\frac{\partial}{\partial t}\rho_t = -\nabla \cdot (v_t \rho_t),$$

を連続の方程式[a] という．

[a] 超関数の意味で成り立つとする．

参考資料：Ambrosio, Gigli, and Savaré 2005, §8.1.

平均場近似による解析

$\varepsilon > 0$,
$s_k = \varepsilon\xi(k\varepsilon)$,
$\Psi(\theta; \rho) = V(\theta) + \int U(\theta, \theta')\rho(d\theta')$,
$\hat{\rho}_k^N = N^{-1}\sum_{i=1}^N \delta_{\theta_i^k}$,
とする．このとき，適当な仮定のもとで偏微分方程式，

$$\partial_t \rho_t = 2\xi(t)\nabla_\theta \cdot (\rho_t \nabla_\theta \Psi(\theta; \rho_t)),$$

の解 ρ_t について解の存在と一意性が言える．また N が十分大きく ε が十分小さいとき，$k = \lfloor t/\varepsilon \rfloor$ なる k に対して ρ_t は $\hat{\rho}_k^N$ を（ある意味で）よく近似する．

平均場近似による解析

偏微分方程式 $\partial_t \rho_t = 2\xi(t)\nabla_\theta \cdot (\rho_t \nabla_\theta \Psi(\theta; \rho_t))$, は損失関数 $L(\rho)$ に対する $(\mathscr{P}(\mathbb{R}^D), W_2)$ 上の勾配流を表している．ここで，W_2 は Wasserstein 距離．

Wasserstein 距離

$\mathcal{X}$ をポーランド空間とし，$\mathcal{X}$ は d で距離化されるとする．このとき $p \in [1, \infty)$ に対して，

$$W_p(\mu, \nu) = \inf \left\{ [\mathbb{E} d(\mu, \nu)^p]^{\frac{1}{p}} : \mathrm{law}(X) = \mu, \mathrm{law}(Y) = \nu \right\},$$

を p 次 Wasserstein 距離という．ここで $\mathrm{law}(X)$ は X の確率分布を表す．

参考資料：Villani et al. 2009, §6.

平均場近似による解析

Wasserstein 空間

任意の $x_0 \in \mathcal{X}$ をとって固定する．このとき，

$$\mathscr{P}_p(\mathcal{X}) := \left\{ \mu \in \mathscr{P}(\mathcal{X}) : \int_{\mathcal{X}} d(x_0, x)^p \mu(dx) < \infty \right\},$$

と定め $\mathscr{P}_p(\mathcal{X})$ を p 次 Wasserstein 空間という．

参考資料：Villani et al. 2009, §6.

平均場近似による解析

ニューラルネットワークの学習は近似的に $\mathscr{P}_2(\mathbb{R}^D)$ 上の勾配流だと思えそう.

目次

Langevin 動力学による解析

Raginsky, Rakhlin, and Telgarsky 2017 の紹介.

Langevin 動力学による解析

パラメータ $w \in \mathbb{R}^d$ に関する $L[\ell(Y, h(X))]$ はある関数 f を用いて,

$$F(w) := \mathbb{E}_P[f(w, Z)] = \int_Z f(w, z) P(dz),$$

と表せる．ここで $Z = (Z_1, \cdots, Z_n) \sim P\,(\text{i.i.d.})$.

Langevin 動力学による解析

$\eta > 0$: **学習率.**

$\beta > 0$: **定数.**

g_k : $\nabla F_Z(W_k)$ **の不変推定量.**

ξ_k : **標準正規分布に従うノイズ.**

$$W_{k+1} = W_k - \eta g_k + \sqrt{2\eta\beta^{-1}}\xi_k.$$

Langevin 動力学による解析

$$W_{k+1} = W_k - \eta g_k + \sqrt{2\eta\beta^{-1}}\xi_k.$$

$$\uparrow$$

$$dW(t) = -\nabla F_Z(W(t))dt + \sqrt{2\beta^{-1}}dB(t).$$

参考文献

参考文献 i

- Ambrosio, Luigi, Nicola Gigli, and Giuseppe Savaré (2005). *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press.
- Mei, Song, Andrea Montanari, and Phan-Minh Nguyen (2018). "A mean field view of the landscape of two-layer neural networks". In: *Proceedings of the National Academy of Sciences* 115.33, E7665–E7671.
- Oksendal, Bernt (2013). *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media.

参考文献 ii

- Raginsky, Maxim, Alexander Rakhlin, and Matus Telgarsky (2017). "Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis". In: *Conference on Learning Theory*. PMLR, pp. 1674–1703.
- Villani, Cédric et al. (2009). *Optimal transport: old and new*. Vol. 338. Springer.
- Yarotsky, Dmitry (2017). "Error bounds for approximations with deep ReLU networks". In: *Neural Networks* 94, pp. 103–114.
- 今泉允聡 (2021). "深層学習の原理解析: 汎化誤差の側面から". In: 日本統計学会誌 50.2, pp. 257–283.
- 鈴木大慈 (2018). "統計的学習理論とその深層学習への応用". In: 応用数理 28.4, pp. 28–33.