無限次元勾配ランジュバン動力学による深層学習の最適化と汎化誤差解析

1. [Suzuki: Generalization bound of globally optimal non-convex neural network training: Transportation map estimation by infinite dimensional Langevin dynamics. NeurIPS2020]
2. [Suzuki & Akiyama: Benefit of deep learning with non-convex noisy gradient descent: Provable excess risk bound and superiority to kernel methods. network training: Transportation map estimation by infinite dimensional Langevin dynamics. ICLR2021]

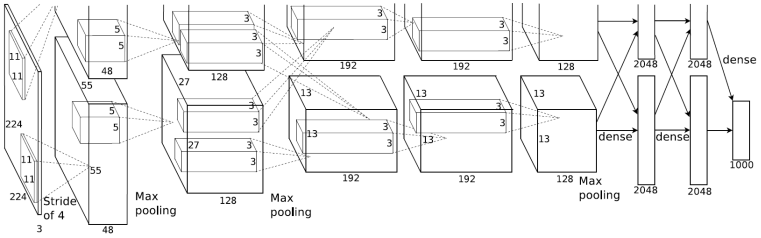
鈴木大慈

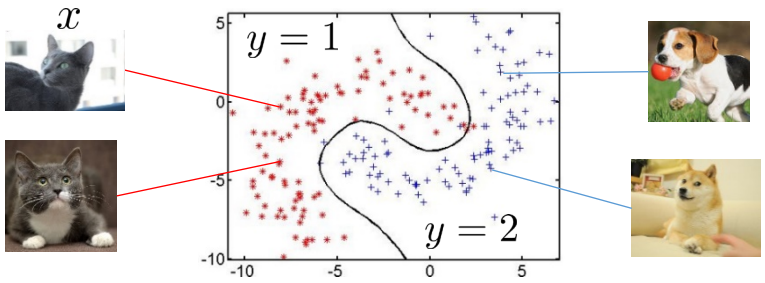
東京大学/ 理研AIP (深層学習理論チーム)

2021年1月14日

@ディープラーニングと物理学2020オンライン

深層学習の“学習”





深層ニューラルネットワークをデータにフィットさせるとは？

$$L(W) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(W)$$

W: パラメータ

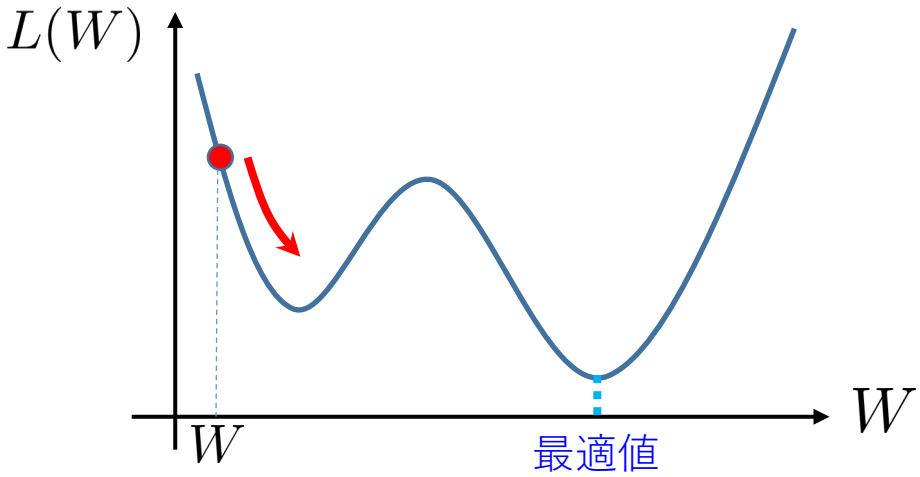
i番目のデータで正解していれば小さく，間違っていれば大きく

損失関数：データへの当てはまり度合い

損失関数最小化

$$\min_W L(W)$$

(Wは数十億次元)



通常，**(確率的)勾配降下法**で最適化

- 最適化は可能？
- その時の汎化誤差はどれくらい？

訓練誤差と汎化誤差

パラメータ θ：ネットワークの構造を表す変数
損失関数 $\ell(Y, f(X, \theta))$：パラメータ θ がデータをどれだけ説明しているか

予測誤差：損失の期待値

$$\mathrm{E}[\ell(Y, f(X, \theta))]$$

$$L(\theta)$$

本当に最小化したいもの.

訓練誤差：有限個のデータで代用

$$\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i, \theta))$$

$$\widehat{L}(\theta)$$

代わりに最小化するもの.

(訓練データはテストデータと
同じ分布に従っていると仮定)

この二つには大きなギャップがある.
[過学習]

※クラスタリング等，教師なし学習も尤度を使ってこのように書ける.

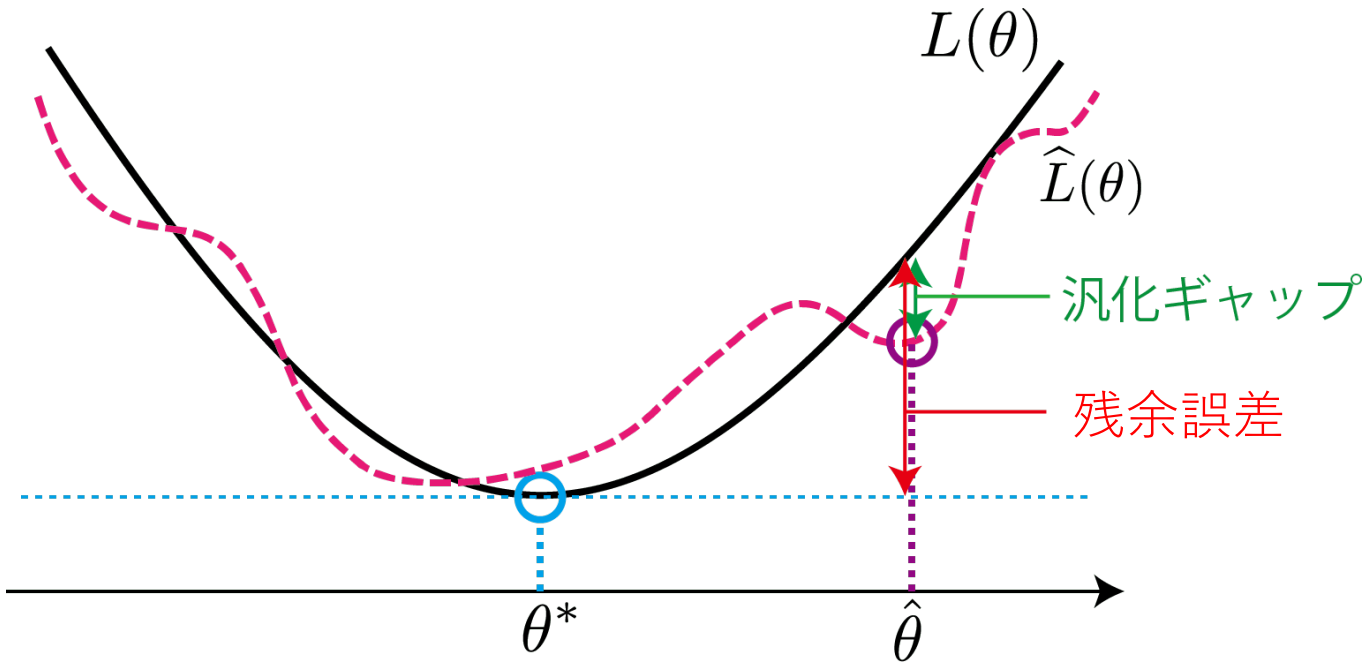
学習理論の設定

- 汎化ギャップ(汎化誤差)と余剰誤差
  Generalization gap　　Excess risk

汎化誤差: $L(\hat{\theta}) - \widehat{L}(\hat{\theta})$

残余誤差: $L(\hat{\theta}) - \inf_{\theta} L(\theta)$

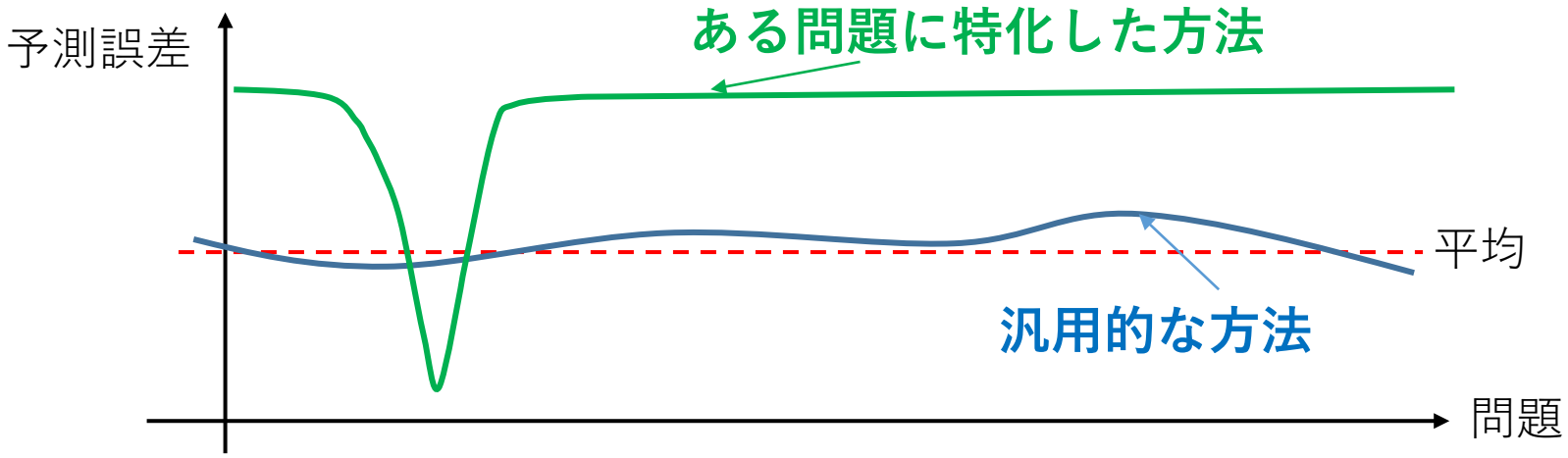
もしくは $L(\hat{\theta}) - \inf_{f} \mathbb{E}[\ell(Y, f(X))]$



学習機の複雑さと学習能力

- No free lunch theorem

「あらゆる問題で性能の良い汎用的学習機は実現不可能であり，ある問題に特殊化された手法に勝てない」



機械学習への教訓

「必要以上に複雑なモデルを当てはめると失敗する」

学習手法は「どこかを“贔屓”する必要がある」
→ モデリングの重要性 (オッカムの剃刀)

William of Ockham：1285-1347．スコラ学の神学者，哲学者．

No free lunch theorem: [D.H.Wolpert and W.G. Macready: 1995,1997][Y.C. Ho and D.L. Pepyne: 2002]

リスクの概念

- どこを贔屓するか？
  - ➢真の分布を限定せずに予測誤差の優劣を論じることは不可能
  - ➢仮に真があるモデルに入っていた場合にある推定量はどのような贔屓をする方法か？

$\mathcal{P}$ ：真の分布のモデル　　$P^* \in \mathcal{P}$ ：真の分布

$D^n = (x_i, y_i)_{i=1}^n$ ：訓練データ　　$\hat{\theta}, \tilde{\theta}$: 推定量

■ ミニマックス最適性

$$\sup_{P^* \in \mathcal{P}} \mathbb{E}_{D^n \sim P^*}[L(\hat{\theta})] = \inf_{\tilde{\theta}:\text{Estimator}} \sup_{P^* \in \mathcal{P}} \mathbb{E}_{D^n \sim P^*}[L(\tilde{\theta})]$$

■ 許容性　つぎのような推定量$\tilde{\theta}$が存在しない:

$$\mathbb{E}_{D^n \sim P^*}[L(\tilde{\theta})] \leq \mathbb{E}_{D^n \sim P^*}[L(\hat{\theta})] \quad (\forall P^* \in \mathcal{P})$$

$$\mathbb{E}_{D^n \sim P^*}[L(\tilde{\theta})] < \mathbb{E}_{D^n \sim P^*}[L(\hat{\theta})] \quad (\exists P^* \in \mathcal{P})$$

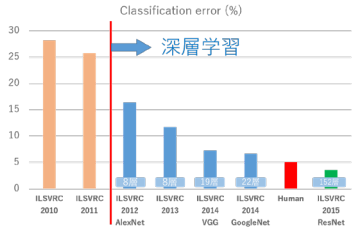
■ ベイズ最適性　　π_0: 事前分布

ベイズリスク $\int \mathbb{E}_{D^n \sim P^*}[L(\hat{\theta})] \mathrm{d}\pi_0(P^*)$ を最小にする推定法$\hat{\theta}$.

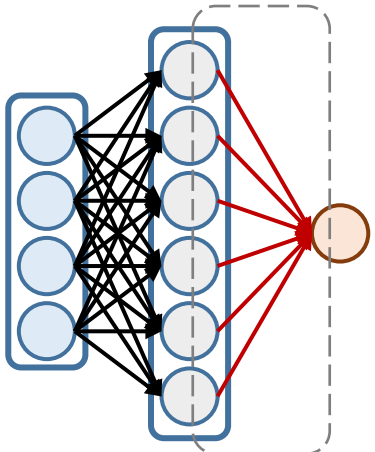
(→ ベイズ推定量)

前提となる話

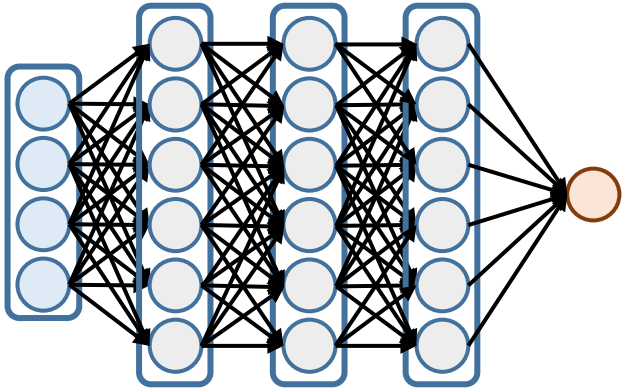
- **最適化という意味ではカーネル法は「良い」.**
- **線形モデル+凸損失関数→凸最適化**



カーネル法
浅層



多層ニューラルネット
深層学習



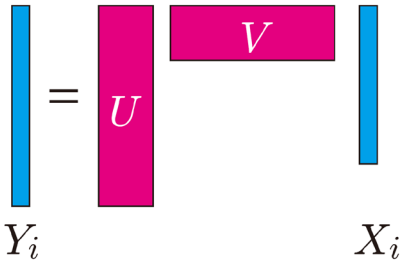
カーネル法と深層学習の違い？

なぜ深層学習が良いのか？

- 真の関数の形状によって深層が有利になる（残余誤差を比較）

縮小ランク回帰

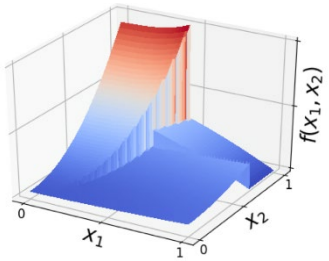
特徴空間の次元が低い状況は深層学習が得意



区分滑らかな関数

[Imaizumi&Fukumizu, 2019]

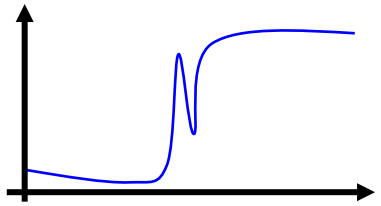
不連続な関数の推定は深層学習が得意



Besov空間

[Suzuki, 2019]

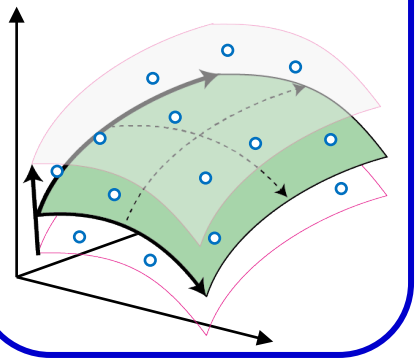
滑らかさが非一様な関数の推定は深層学習が得意



低次元データ

[Schmidt-Hieber, 2019] [Nakada&Imaizumi, 2019][Chen et al., 2019][Suzuki&Nitanda, 2019]

データが低次元部分空間上に分布していたら深層学習が有利

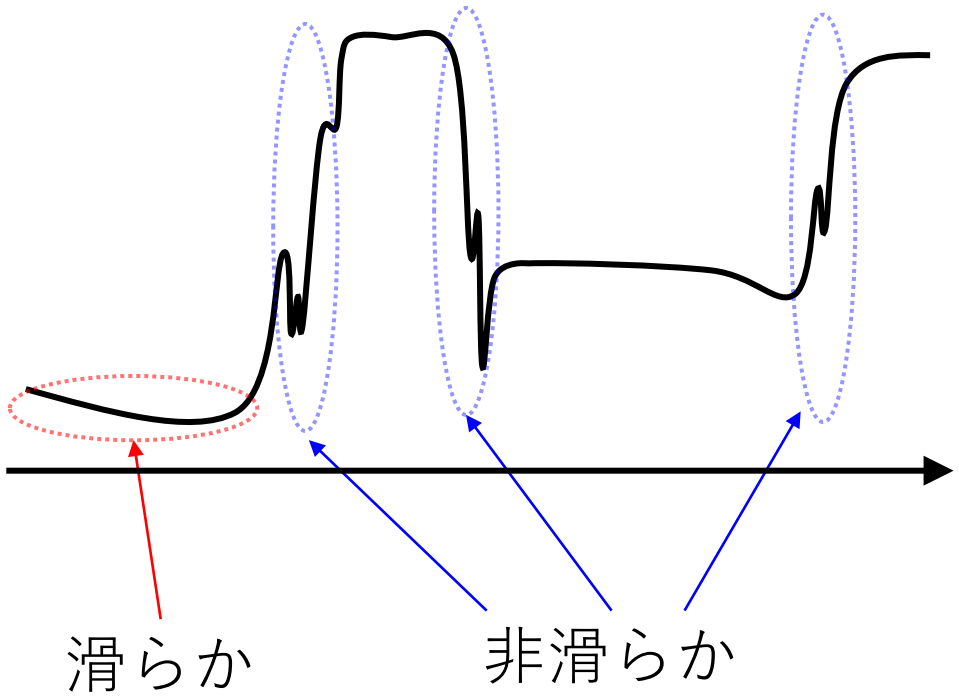


深層	$\frac{r(M+N)}{n}$	$n^{-\frac{2s}{2s+d}} \vee n^{-\frac{\alpha}{\alpha+D-1}}$	$n^{-\frac{2s}{2s+d}}$	$n^{-\frac{2s}{2s+D}}$
カーネル	$\frac{MN}{n}$	$\frac{1}{\sqrt{n}}$	$n^{-\frac{2s-2d(1/p-1/2)_+}{2s+d-2d(1/p-1/2)_+}}$	$n^{-\frac{2(s-D/p+d/2)}{2(s-D/p+d/2)+d}} \vee n^{-\frac{2s}{2s+D}}$

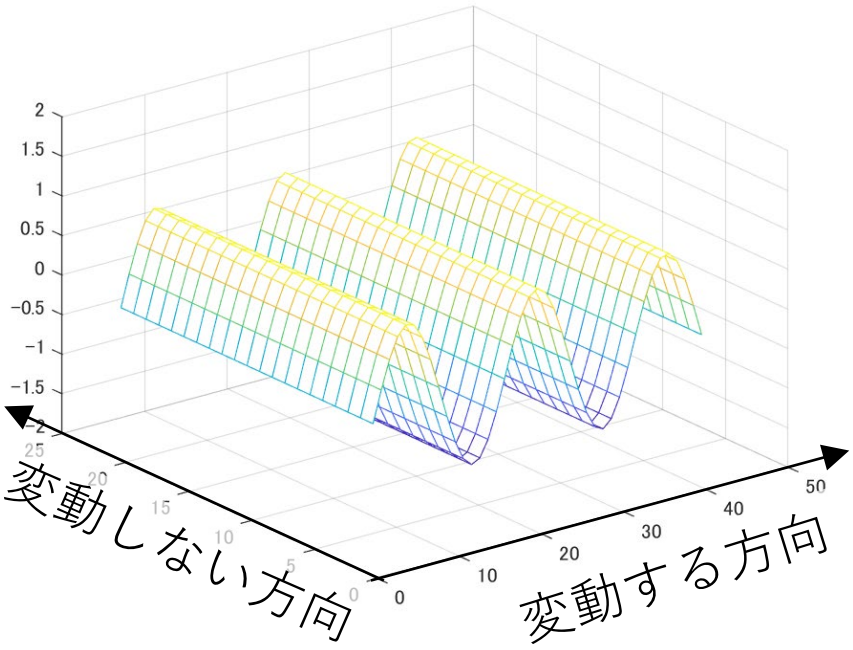
推定精度

典型的な例

滑らかな部分と
そうでない部分が混在



大きく変動する方向と
そうでない方向が混在



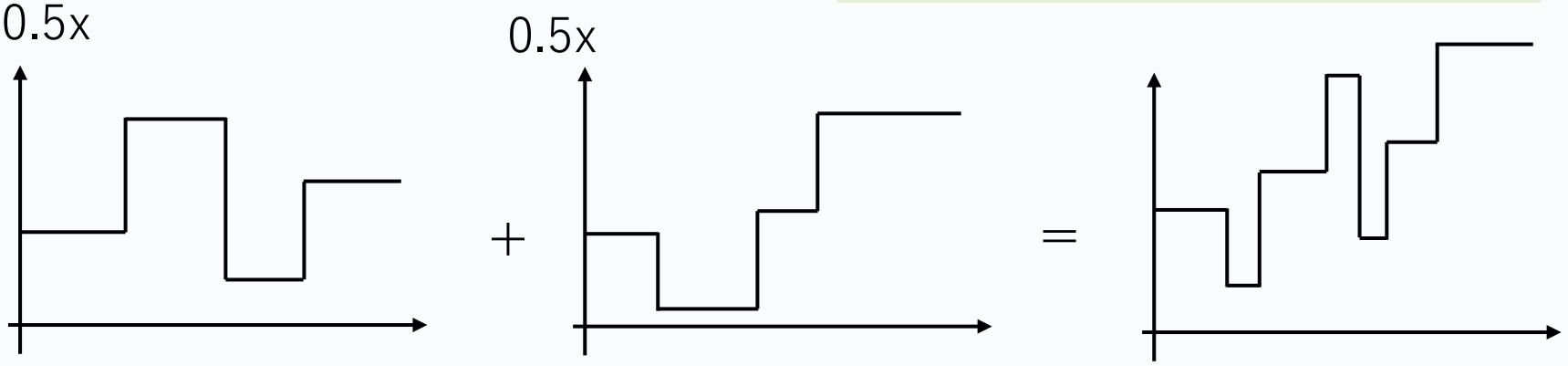
数学的に一般化

「滑らかさの非一様性」「不連続性」「データの低次元性」
凸結合を取って崩れる性質をもった関数の学習は深層学習が強い

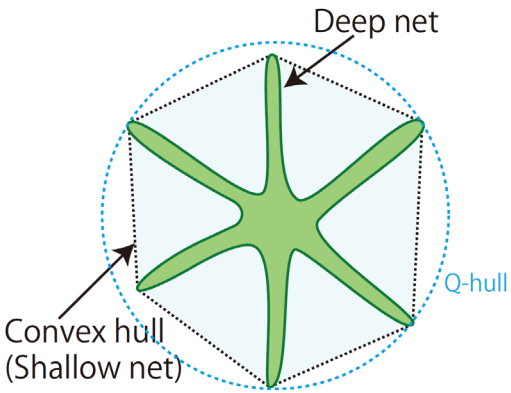
[Satoshi Hayakawa and Taiji Suzuki: 2020]

例：ジャンプが３か所の区分定数関数

深層:$1/n$, カーネル: $1/\sqrt{n}$

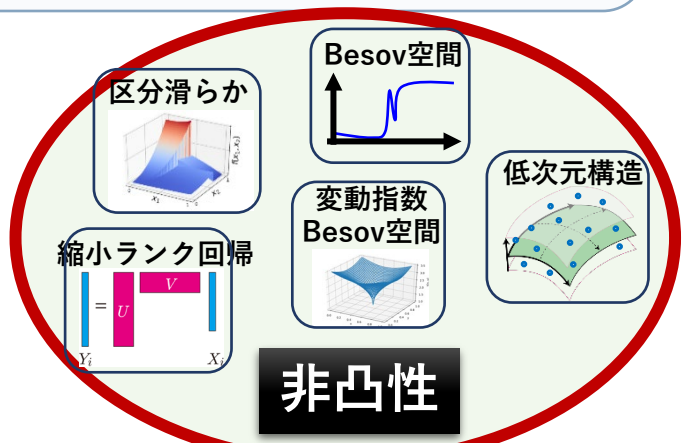


ジャンプ**3**か所　　ジャンプ**3**か所　　ジャンプ**6**か所



$$\inf_{\hat{f}:\text{Linear}} \sup_{f^o \in \mathcal{F}} \mathrm{E}[\|\hat{f} - f^o\|^2_{L_2(P)}]$$

$$= \inf_{\hat{f}:\text{Linear}} \sup_{f^o \in \mathrm{conv}(\mathcal{F})} \mathrm{E}[\|\hat{f} - f^o\|^2_{L_2(P)}]$$



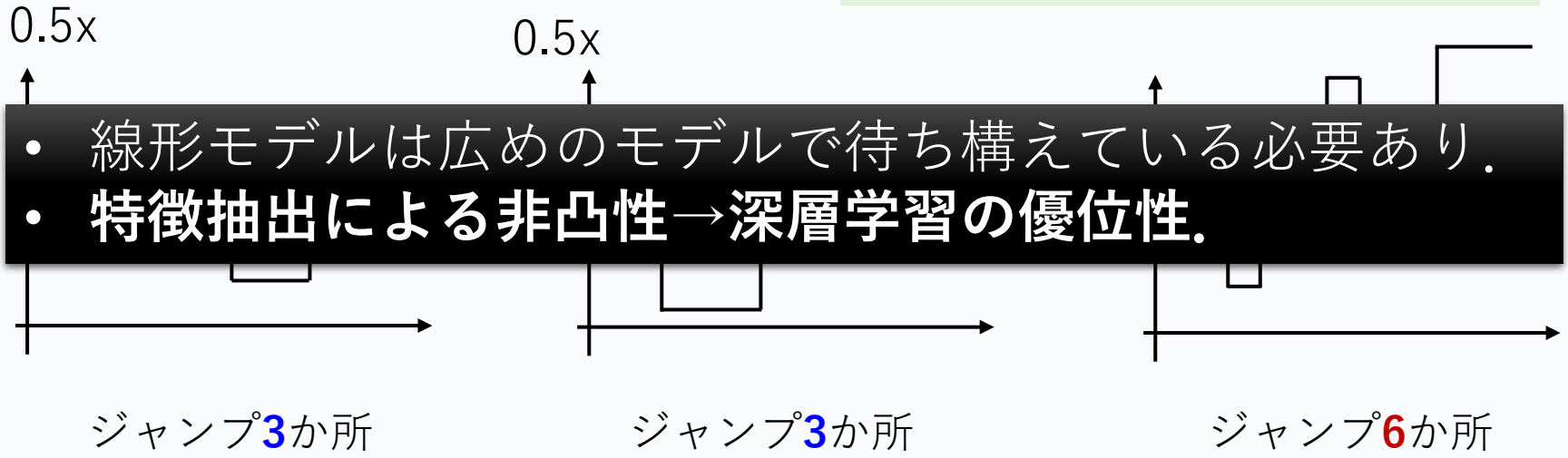
数学的に一般化

「滑らかさの非一様性」「不連続性」「データの低次元性」
凸結合を取って崩れる性質をもった関数の学習は深層学習が強い

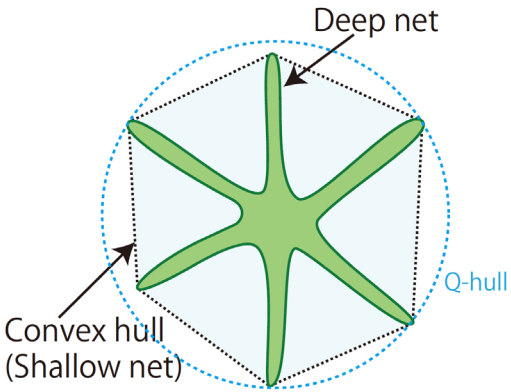
[Satoshi Hayakawa and Taiji Suzuki: 2020]

例：ジャンプが３か所の区分定数関数

深層:$1/n$, カーネル: $1/\sqrt{n}$

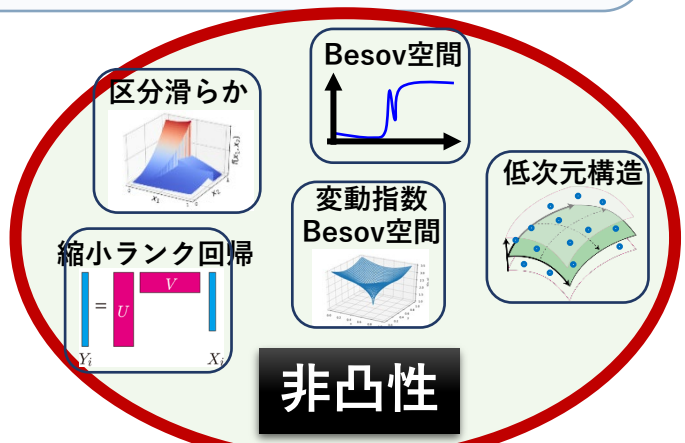


- 線形モデルは広めのモデルで待ち構えている必要あり.
- **特徴抽出による非凸性→深層学習の優位性.**



$$\inf_{\hat{f}:\text{Linear}} \sup_{f^o \in \mathcal{F}} \mathrm{E}[\|\hat{f} - f^o\|^2_{L_2(P)}]$$

$$= \inf_{\hat{f}:\text{Linear}} \sup_{f^o \in \mathrm{conv}(\mathcal{F})} \mathrm{E}[\|\hat{f} - f^o\|^2_{L_2(P)}]$$



- 深層学習の良さは「非凸性」は本質的.
- これら統計理論は「最適化」を考慮していない.

	凸性	非凸性
統計理論		
最適化理論		

本研究の目標

深層学習における非凸性を保ちつつ
「最適化」と「統計理論」
を結びつける.

既存研究との関係

大域的最適性を保証する理論的枠組み

理論的枠組み	横幅（次元）	汎化性能	多層
Neural Tangent Kernel	無限へ漸近	本質的にカーネル法 /Early stopping必要	△
平均場解析	無限へ漸近	限定された状況	△
（既存の）有限次元勾配 Langevin動力学	有限（低次元）	汎化ギャップは保証あり/大きいモデルはNG	△
本研究: 無限次元勾配 Langevin動力学	**有限/無限 統一的な枠組み**	**汎化ギャップ/残余誤差ともに保証**	**〇**

- 深層NNモデルの非凸性を失わず最適化したい.
- モデルサイズをサンプルサイズに依存させたくない.

本研究

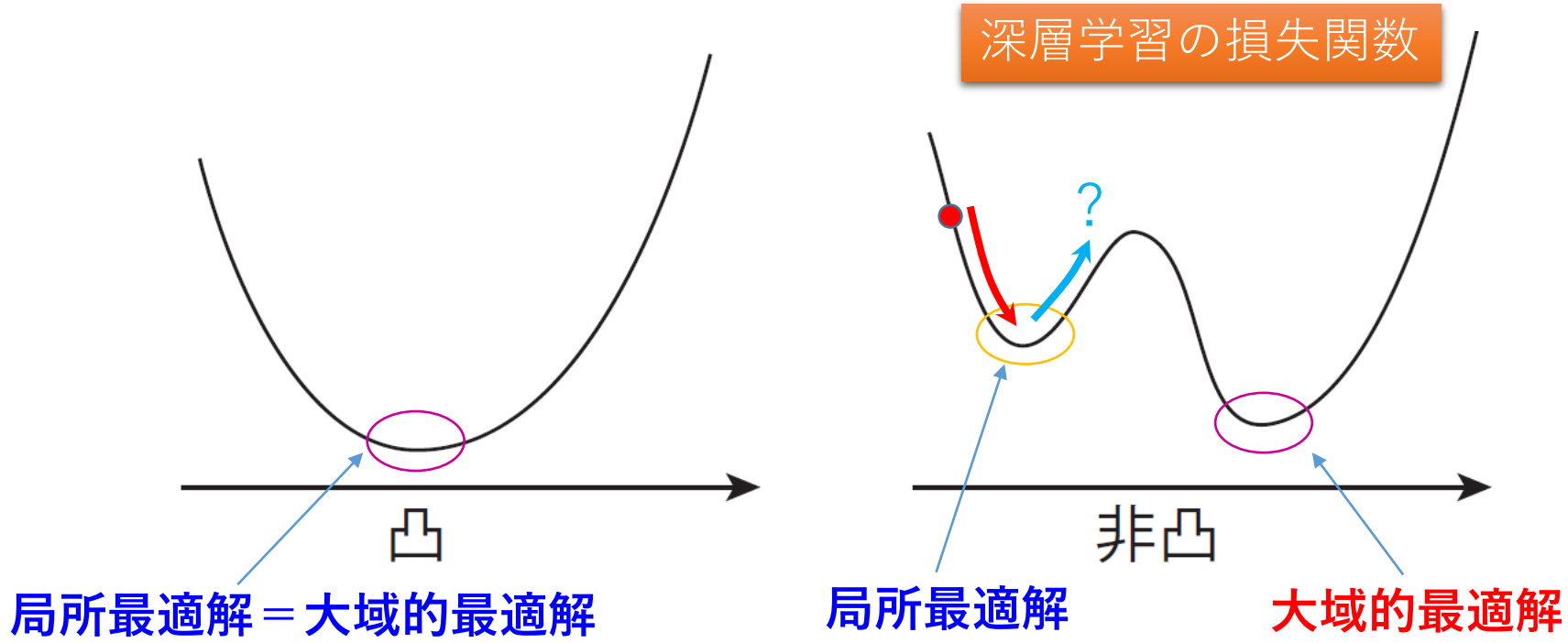
無限次元Langevin動力学
- **有限・無限横幅を統一的に扱う**
- **汎化性能保証**

[Muzellec, Sato, Massias & Suzuki: Dimension-free convergence rates for gradient Langevin dynamics in RKHS. arXiv:2003.00306]
[Suzuki: Generalization bound of globally optimal non-convex neural network training: Transportation map estimation by infinite dimensional Langevin dynamics. NeurIPS2020. arXiv:2007.05824]

問題点

目的関数が非凸関数

凸関数 $\theta f(x) + (1-\theta) f(y) \geq f(\theta x + (1-\theta) y) \quad (\forall x, y \in \mathbb{R}^p, \theta \in [0,1])$



局所最適解や鞍点にはまる可能性あり

“**狭い**”ネットワークの学習はNP-完全:

- Judd (1988), Neural Network Design and the Complexity of Learning.
- Blum&Rivest (1992), Training a 3-node neural network is NP-complete.

Loss landscape

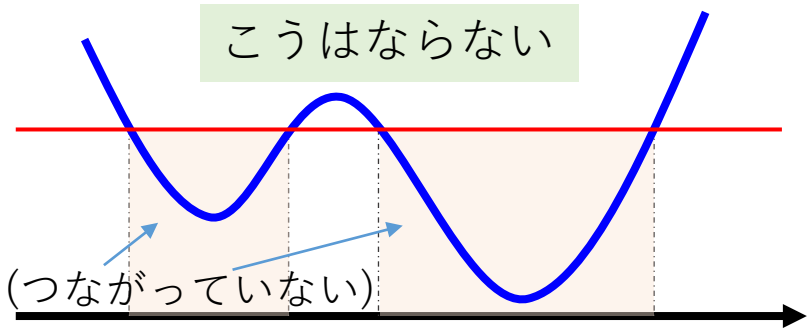
- 横幅の広いNNの訓練誤差には孤立した局所最適解がない．（局所最適解は大域的最適解とつながっている）

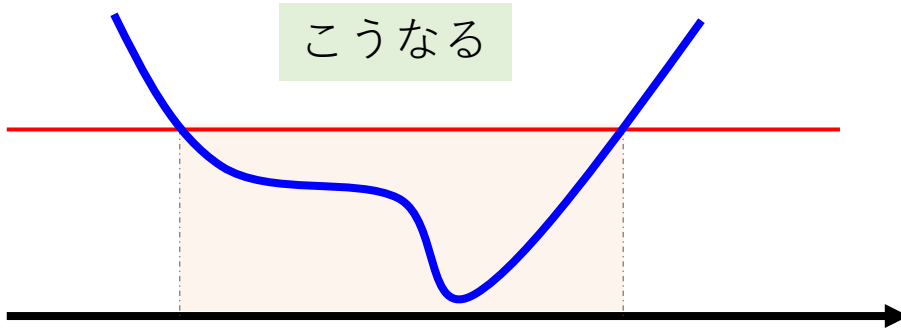
※とはいえ，勾配法で大域的最適解に到達可能かは別問題．

定理

n個の訓練データ$(x_i, y_i)_{i=1}^n$が与えられているとする．損失関数ℓは凸関数とする．
任意の連続な活性化関数について，横幅がデータサイズより広い（$M \geq n$）二層NN$f_{(a,W)}(x) = \sum_{m=1}^{M} a_m \eta(w_m^{\mathsf{T}} x)$に対する訓練誤差$\hat{L}(a, W) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f_{(a,W)}(x_i))$の任意のレベルセットの弧状連結成分は大域的最適解を含む．言い換えると，任意の局所最適解は大域的最適解である．

[Venturi, Bandeira, Bruna: Spurious Valleys in One-hidden-layer Neural Network Optimization Landscapes. JMLR, 20:1-34, 2019.]





（参考）2層NN-非線形活性化関数-

二層目の重みを固定する設定

(Tian, 2017; Brutzkus and Globerson, 2017; Li and Yuan, 2017; Soltanolkotabi, 2017; Soltanolkotabi et al., 2017; Shalev-Shwartz et al., 2017; Brutzkus et al., 2018)

固定　こちらのみ動かす

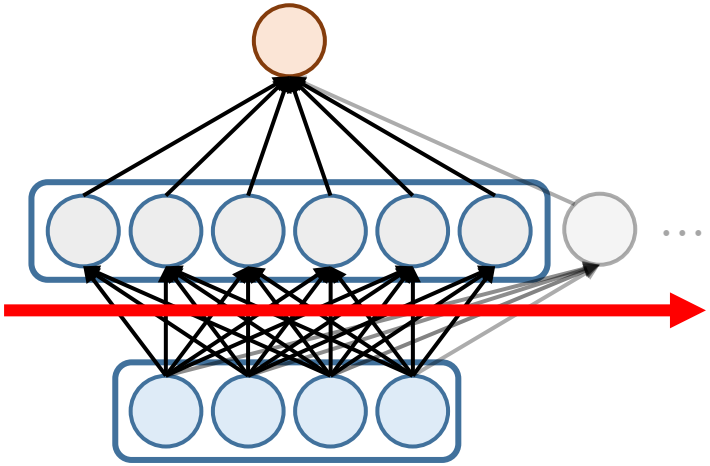
$$y = \sum_{j=1}^{k} v_j \eta(w_j^\top x + b_j)$$

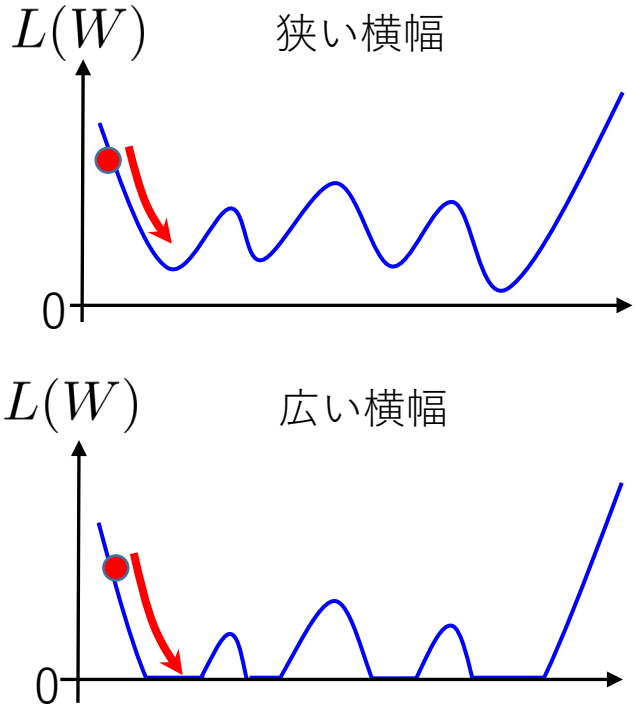
- Li and Yuan (2017): ReLU，入力はガウス分布を仮定
  - SGDは多項式時間で大域的最適解に収束
  - 学習のダイナミクスは2段階
    → 最適解の近傍へ近づく段階 + 近傍での凸最適化的段階
- Soltanolkotabi (2017): ReLU，入力はガウス分布を仮定
  - 過完備（横幅>サンプルサイズ）なら勾配法で最適解に線形収束
    (Soltanolkotabi et al. (2017)は二乗活性化関数でより強い帰結)
- Brutzkus et al. (2018): ReLU
  - 線形分離可能なデータなら過完備ネットワークで動かしたSGDは大域的最適解に有限回で収束し，過学習しない．
    (線形パーセプトロンの理論にかなり依存)

Li and Yuan (2017): Convergence Analysis of Two-layer Neural Networks with ReLU Activation.
Soltanolkotabi (2017): Learning ReLUs via Gradient Descent.
Brutzkus, Globerson, Malach and Shalev-Shwartz (2018): SGD learns over parameterized networks that provably generalized on linearly separable data.

オーバーパラメトライゼーション

横幅が広いと局所最適解が大域的最適解になる.





自由度が上がるため，初期値から最適解(完全フィット)へ到達しやすい.

- 二種類の解析手法
  - Neural Tangent Kernel [Jacot+ 2018][Du+ 2019][Arora+ 2019]
  - Mean-field analysis (平均場解析)

[Nitanda & Suzuki (2017), Chizat & Bach (2018), Mei, Montanari, & Nguyen (2018)]

Neural Tangent Kernel

$f_W(x) = \sum_{j=1}^{M} a_j \eta(w_j^\top x)$

[Jacot, Gabriel, & Hongler (2019)]

$$f_W(x) \simeq (W - W^{(0)})^\top \nabla_W f_{W^{(0)}}(x)$$

➡ $k_{\mathrm{NTK}}(x, x') = \langle \nabla_W f_{W^{(0)}}(x), \nabla_W f_{W^{(0)}}(x') \rangle$ (NTK)

初期値のスケールが大きいので，初期値周りの線形近似でデータにフィットできてしまう．

Theorem [Arora et al., 2019]

$M = \Omega(n^2 \log(n)/\lambda_{\min})$とすれば，勾配法によって大域的最適解へ線形収束し，その汎化誤差は$\sqrt{\boldsymbol{y}^\top (K_{W^{(0)}})^{-1} \boldsymbol{y}/n}$で抑えられる．

See also[Du et al., 2018; Allen-Zhu, Li & Song, 2018; Li & Liang, 2018; Zou & Gu, 2019]

- 訓練誤差0の解に線形収束する．
- 汎化誤差も一応抑えられている．

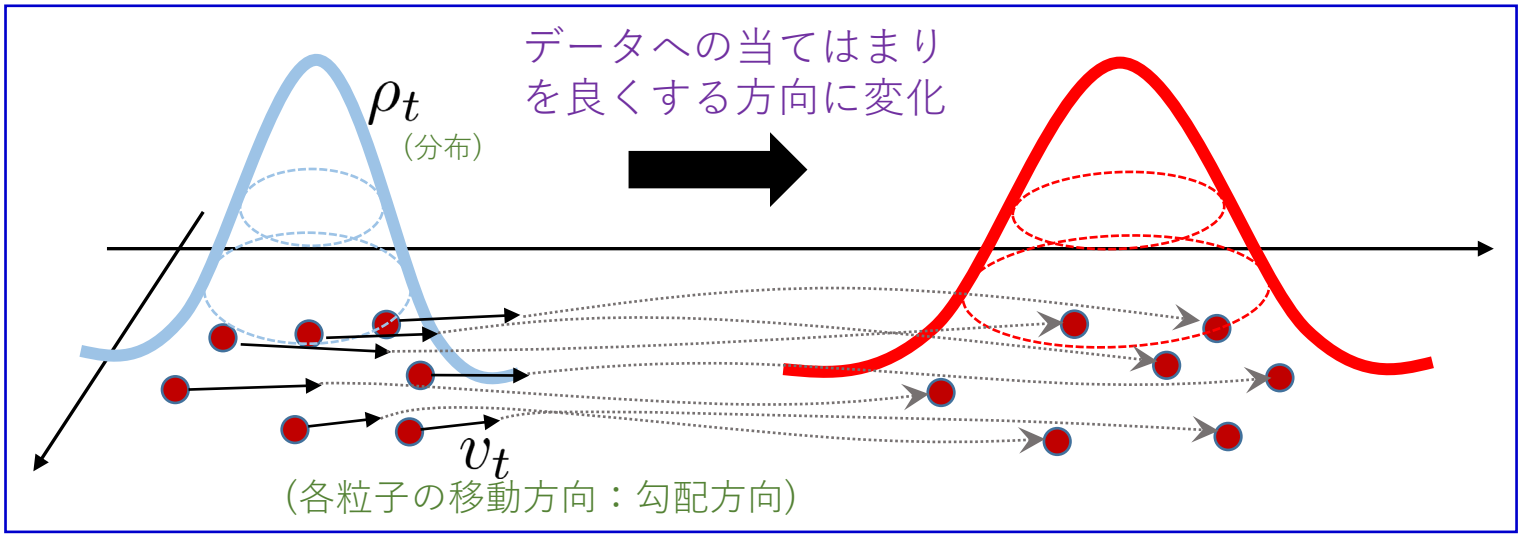
- **横幅Mはサンプルサイズnに応じて無限大へ飛ぶ必要がある．**
- **カーネル法の枠組みを抜け出せていない．**
- Early stoppingしないと過学習する．

平均場解析

- ニューラルネットワークの最適化をパラメータの分布最適化とみなす.

$$f(x) = \frac{1}{M}\sum_{j=1}^{M} a_j \eta(w_j^\top x)$$

1つの粒子



$$v_t(a,w) = -\frac{1}{n}\sum_{i=1}^{n} \nabla_{(a,w)}(a\eta(w^\top x_i))\ell'(y_i, f_{\rho_t}(x_i))$$

(各粒子は勾配降下方向へ移動)

$$f(x) = \frac{1}{M}\sum_{j=1}^{M} a_j \eta(w_j^\top x) \xrightarrow{M\to\infty} \int a\eta(w^\top x)\rho(a,w)\mathrm{d}a\mathrm{d}w$$

Wasserstein勾配流

粒子数無限大(横幅無限)でパラメータに関数確立密度ρによる期待値とみなせる.

$$f(x) = \frac{1}{M}\sum_{j=1}^{M} a_j \eta(w_j^\top x) \xrightarrow{M\to\infty} \int a\eta(w^\top x)\rho(a,w)\mathrm{d}a\mathrm{d}w$$

fの最適化 ⇔ ρの最適化

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

連続方程式

Wasserstein勾配流

[Atsushi Nitanda and Taiji Suzuki: Stochastic Particle Gradient Descent for Infinite Ensembles. arXiv:1712.05438.]

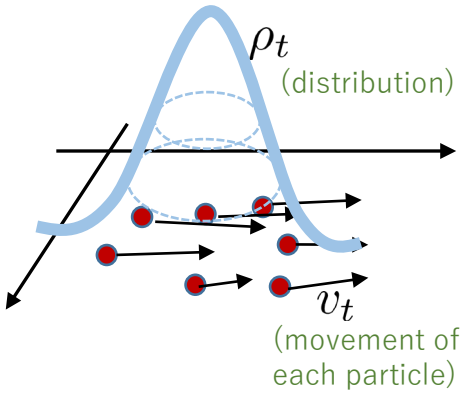
$M\to\infty$の極限で，最適解への収束が成り立つ場合がある.

- やはり横幅$M\to\infty$である必要がある
- 推定誤差理論△（状況限定）

McKean-Vlasov過程

参考

- ノイズありダイナミクス



Model: $f_{\rho_t}(x) = \int a\eta(X_t^\top x)\mathrm{d}\rho_t(X_t)$

ρ_t: law of X_t at time t

ダイナミクス (**McKean-Vlasov過程**):

$$\mathrm{d}X_t = v_t\mathrm{d}t + \sqrt{\tau}\mathrm{d}B_t$$

$$v_t(X_t, \rho_t) = \frac{1}{n}\sum_{i=1}^{n} \ell'(f_{\rho_t}(x_i), y_i)a\eta'(X_t^\top x_i)$$

X_tの値だけでなく**分布**にも依存

Fokker-Planck方程式:

$$\frac{\mathrm{d}\rho_t}{\mathrm{d}t} = -\nabla\cdot(v_t\rho_t) + \frac{\tau}{2}\Delta\rho_t$$

- **収束解析:** Mei, Montanari&Nguyen, 2018; Rotskoff &Vanden-Eijnden, 2018.
- **最適制御理論:** Weinan et al., 2019; Tzen&Raginsky, 2020; Lu et al., 2020.

離散化

参考

時空間離散化:

$$X_{t+1}^m = X_t^m - \epsilon \hat{v}_t(X_t^m, \hat{\rho}_t) + \sqrt{\epsilon\tau}\xi_t$$

(i.i.d., standard Gaussian)

$$\hat{v}_t(X_t^m, \hat{\rho}_t) = \frac{1}{n}\sum_{i=1}^{n} \ell'(f_{\hat{\rho}_t}(x_i), y_i) a \eta'(X_t^{m\top} x_i)$$

$$f_{\hat{\rho}_t}(x) = \frac{1}{M}\sum_{m=1}^{M} a\eta(X_t^{m\top} x) \qquad (\hat{\rho}_t = \tfrac{1}{M}\textstyle\sum_{m=1}^{M} \delta_{X_t^m})$$

Empirical distribution

Pros: 定常分布への収束が保証されている.

[Mei, Montanari&Nguyen, 2018][Tzen&Raginsky, 2020]

Cons:

- **“横幅”Mは$\exp(T)$である必要がある.** 時刻T, 収束を保証するには横幅は十分多くする必要あり(有限粒子数では収束保証されず).
- **ガウス雑音 ξ_t (dB_t) は各粒子ごとに独立同一に添加.**
  - 粒子間の相関・滑らかさは考慮されていない. 実際のDNNでは位置が近い粒子には値が似たノイズ.
- ρ_t は絶対連続 (有限横幅のNNは対象外)

輸送写像による学習と
無限次元ランジュバン動力学

輸送写像を用いたNNの学習

- 2層ニューラルネットワーク

Idea: 分布の学習 → **輸送写像の学習**

$$W : \mathbb{R}^d \to \mathbb{R}^d \qquad W \in L_2(\rho_0)$$

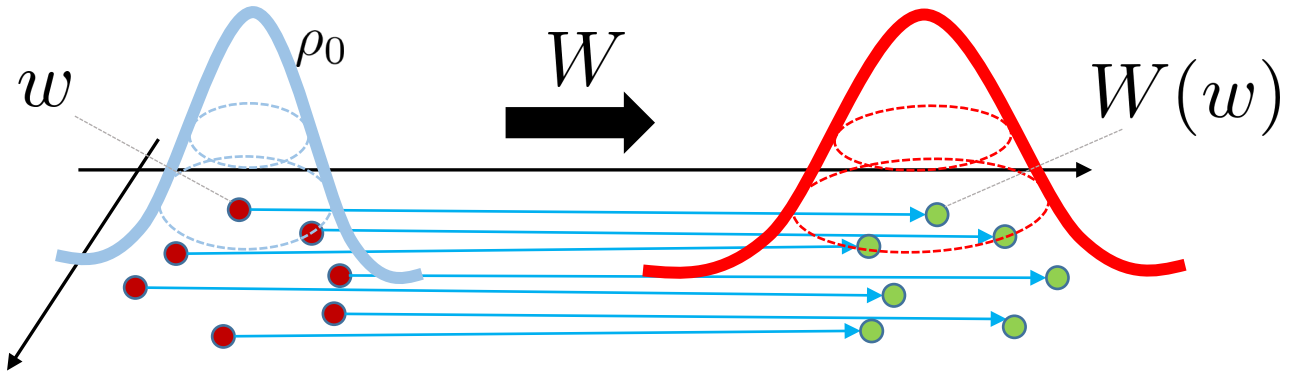
$$f_W(x) := \int_{\mathbb{R}^d} a(W(w))\sigma(W(w)^\top x)\mathrm{d}\rho_0(w) \quad \text{“Lift”}$$

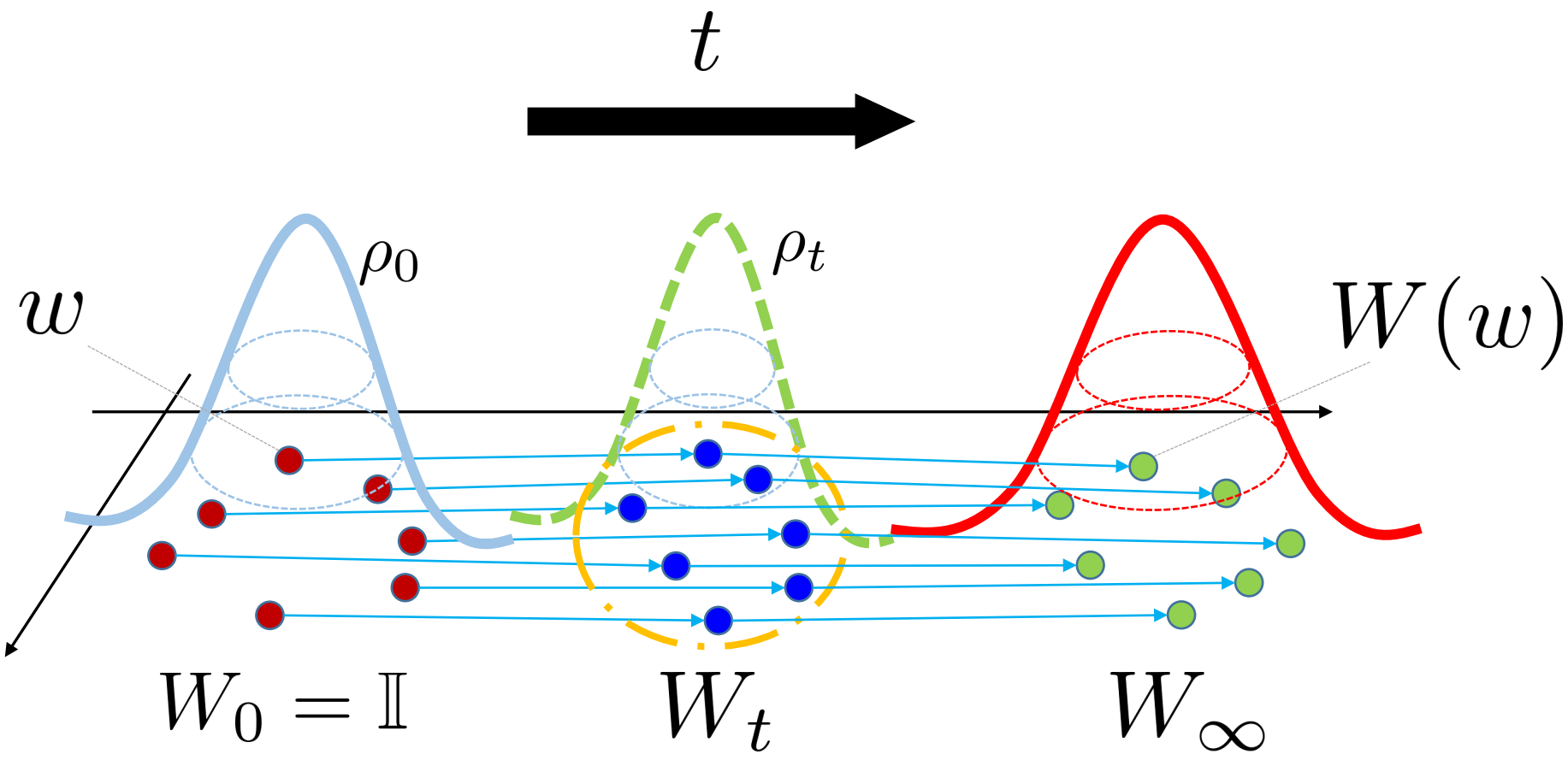
$$= \int_{\mathbb{R}^d} a(w)\sigma(w^\top x)\mathrm{d}W\sharp\rho_0(w)$$

$$f_\rho(x) = \int_{\mathbb{R}^d} a(w)\sigma(w^\top x)\mathrm{d}\rho(w)$$

以前の表記

$$\min_{\rho} \; L(f_\rho) \longrightarrow \min_{W \in \mathcal{H}} \; L(f_{W\sharp\rho_0})$$





ρ_0が有限サポートの離散測度なら**有限横幅のニューラルネットワーク**が扱える．しかも，**横幅はサンプルサイズと独立**.

- **NTKや平均場解析と大きく異なる.**
- 有限横幅/無限横幅を**統一的に扱える.**

より簡単な設定

輸送写像による表現は実は以下の簡単な2層NNの表現も含む.

- 2層NN: 直接表現

$$W : \mathbb{N} \to \mathbb{R}^d \ (m \mapsto W_m)$$

$$f_W = \sum_{m=1}^{\infty} a_m \sigma(W_m^\top x)$$

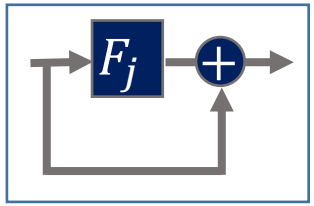
(横幅無限)

- $(a_m)_m$は$a_m \to 0 \ (m \to \infty)$なる固定された係数.
- $a_m = 0 \ (\forall m > M)$ とすれば有限横幅も含まれる.

その他定式化/応用一覧

- **ResNet**

$$f(x) = u^\top (\mathbb{I} + F_T(\cdot)) \circ (\mathbb{I} + F_{T-1}(\cdot)) \circ \cdots \circ (x + F_1(x))$$

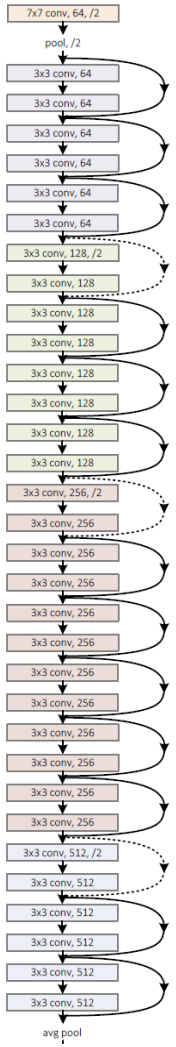


$$W : \mathbb{R}^d \times \{1, \ldots, T\} \to \mathbb{R}^d$$

$$f_W = u^\top \left(\mathbb{I} + \int_{\mathbb{R}^d} a(w,T)\sigma(W(w,T)^\top \cdot) \mathrm{d}\rho_0(w) \right) \circ \cdots \circ \left(\mathbb{I} + \int_{\mathbb{R}^d} a(w,1)\sigma(W(w,1)^\top x) \mathrm{d}\rho_0(w) \right)$$

Residual block

(輸送写像で表現可)



- **2層NNの直接表現**

$$f_W(x) = \sum_{j=1}^{\infty} a_j \sigma(w_j^\top x)$$

- $a_j \le j^{-\gamma}$ for $\gamma > 1/2$
- η is a smooth activation, e.g., sigmoid.

- **RKHS上のベイズ最適化**

[Zimmermann and Toussaint. AAAI, 2018]
[Vellanki, Rana, Gupta, de Celis Leal, Sutti, Height, and Venkatesh: AAAI2019]

- **RKHS上のテンソル分解**

[Signoretto, Lathauwer, and Suykens. arXiv:1310.4977, 2013]
[Suzuki, Kanagawa, Kobayashi, Shimizu, and Tagami. NIPS2016]

- **ロバスト判別**

[Masnadi-Shirazi and Vasconcelos. NIPS2009.]

無限次元非凸最適化

[Muzellec, Sato, Massias, Suzuki, arXiv:2003.00306][Suzuki, arXiv:2007.05824]

$$\min_{x \in \mathcal{H}} \quad L(x)$$

$\mathcal{H}$: Hilbert space

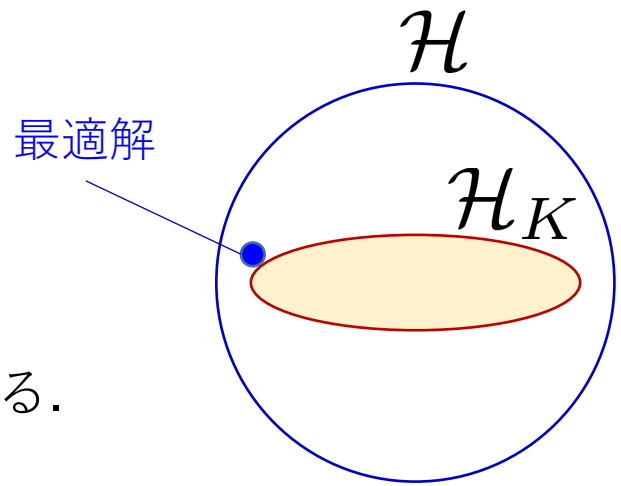
正則化

$$\min_{x \in \mathcal{H}} \quad L(x) + \lambda \|x\|^2_{\mathcal{H}_K}$$

$\mathcal{H}_K$:"smaller" Hilber space

$\mathcal{H}_K \hookrightarrow \mathcal{H}$

例：
- $\mathcal{H}$: $L^2(\rho_0)$
- $\mathcal{H}_K$: 再生核ヒルベルト空間
  (e.g., Sobolev空間)

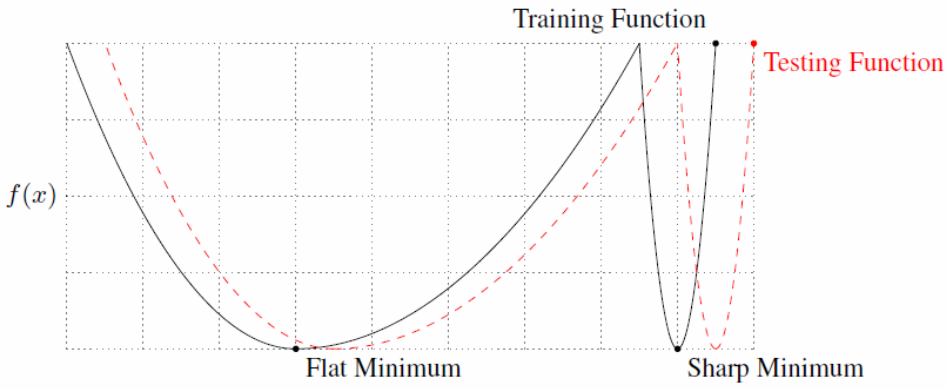


暗黙の仮定: 大域的最適解は$\mathcal{H}_K$で十分に近似できる.

→ **無限次元ランジュバン動力学**で最適化

Noisy gradient descent

SGDは「フラットな局所最適解」に落ちやすい
→良い汎化性能を示す
という説

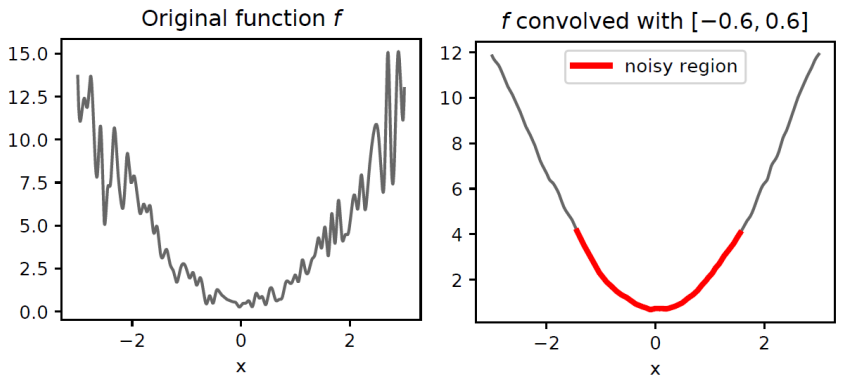


Keskar, Mudigere, Nocedal, Smelyanskiy, Tang (2017):
On large-batch training for deep learning: generalization gap and sharp minima.

$$\theta_t = \theta_{t-1} - \underbrace{\alpha_b \left(\frac{1}{b} \sum_{j=1}^{b} \nabla_\theta \ell(z_{i_j}; \theta) \right)}_{\cong \text{正規分布}}$$

→ ランダムウォークはフラットな領域にとどまりやすい

ノイズによる平滑化効果



[Kleinberg, Li, and Yuan, ICML2018]

確率的勾配を用いる
⇒ 解にノイズを乗せている
⇒ 目的関数を平滑化

$$\begin{aligned} x_t &= x_{t-1} - \eta(\nabla L(x_{t-1}) + \xi_t) \\ \Rightarrow y_t &= y_{t-1} - \eta\xi_{t-1} - \eta\nabla L(y_{t-1} - \eta\xi_{t-1}) \\ \Rightarrow \mathbb{E}_{\xi_{t-1}}[y_t] &= y_{t-1} - \eta\nabla\mathbb{E}_{\xi_{t-1}}[L(y_{t-1} - \eta\xi_{t-1})] \end{aligned}$$

ノイズを加えて平滑化した目的関数
$\bar{L}(y_t) = \mathbb{E}_{\xi_t}[L(y_t - \eta\xi_t)]$ **を最適化.**

無限次元ランジュバン動力学

$$x = \sum_{j=1}^{\infty} x_j f_j \in \mathcal{H}$$

$$\min_{x \in \mathcal{H}} L(x) \Rightarrow \min_{x \in \mathcal{H}} \left\{ L(x) + \frac{\lambda}{2} \|x\|_{\mathcal{H}_K}^2 \right\}$$

$\mathcal{H}_K$: RKHS with kernel K.

$\mathcal{H}_K \hookrightarrow \mathcal{H}$

$$\mathrm{d}X_t = -\nabla \left(L(X_t) + \frac{\lambda}{2} \|X_t\|_{\mathcal{H}_K}^2 \right) \mathrm{d}t + \sqrt{\frac{2}{\beta}} \mathrm{d}\xi_t$$

Cylindrical Brownian motion: $\xi_t = \sum_{j=1}^{\infty} \xi_{j,t} f_j$

$(f_j)_j$: $\mathcal{H}$の正規直交基底

時間離散化 (準陰的Eulerスキーム):

$$X_{n+1} = X_n - \eta \left(\nabla L(X_n) + \frac{\lambda}{2} \nabla \|X_{n+1}\|_{\mathcal{H}_K}^2 \right) + \sqrt{2\frac{\eta}{\beta}} \xi_n$$

$$\Rightarrow X_{n+1} = S_\eta \left(X_n - \eta \nabla L(X_n) + \sqrt{2\frac{\eta}{\beta}} \xi_n \right) \quad \left(S_\eta := (I + \eta\lambda A)^{-1} \right)$$

where $x^* A x = \|x\|_{\mathcal{H}_K}^2$

$\xi_n = \sum_{j=1}^{\infty} \gamma_{n,j} f_j$ where $\gamma_{n,j} \sim N(0,1)$ (i.i.d.).

定常分布

$$\mathrm{d}X_t = -\nabla\left(L(X_t) + \frac{\lambda}{2}\|X_t\|^2_{\mathcal{H}_K}\right)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}\xi_t$$

$$\mathrm{d}\pi_\infty(x) \;\propto\; \exp\big(-\beta L(x)\big)\mathrm{d}\mu_*(x)$$

尤度　　事前分布

$\mu_* = N(0, (\beta\lambda A)^{-1})$：正則化項に対応したガウス過程

where A is an operator such that $x^* A x = \|x\|^2_{\mathcal{H}_K}$.

$$\pi_\infty(x) \;\propto\; \exp\left(-\beta L(x) - \beta\frac{\lambda}{2}\|x\|^2_{\mathcal{H}_K}\right)$$ と解釈しても良い.

（無限次元）勾配ランジュバン動力学の定常分布は
ガウス過程事前分布を用いたベイズ事後分布に対応する.
→ 汎化誤差の解析が可能

無限次元の設定

ヒルベルト空間

$$\mathcal{H} = \left\{ \sum_{k=0}^{\infty} \alpha_k f_k \mid \sum_{k=0}^{\infty} \alpha_k^2 < \infty \right\}$$

$$\langle x, y \rangle = \sum_{k=0}^{\infty} \alpha_k \beta_k \quad \text{for } x = \sum_k \alpha_k f_k,\ y = \sum_k \beta_k f_k.$$

RKHS構造

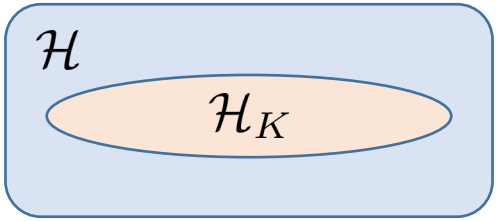
$$\mathcal{H}_K = \left\{ \sum_{k=0}^{\infty} \alpha_k f_k \mid \sum_{k=0}^{\infty} \alpha_k^2 / \mu_k < \infty \right\}$$

$$\langle x, y \rangle_{\mathcal{H}_K} = \sum_{k=0}^{\infty} \alpha_k \beta_k / \mu_k \quad \text{for } x = \sum_k \alpha_k f_k,\ y = \sum_k \beta_k f_k.$$

仮定（固有値の減少）

$$\mu_k \simeq k^{-2}$$

（あまり本質的ではない．$\mu_k \sim k^{-p}$ $(p > 1)$ としても良い．）



Assumption (1)

- It either holds:
  - (Strict Dissipativity) $\lambda > M\mu_0$, or (強):強凸
  - (Bounded gradients) $\|\nabla L(\cdot)\| \leq B$, for $B > 0$. (弱)

散逸条件:

For $C = -\frac{\lambda}{2}\nabla\|\cdot\|^2_{\mathcal{H}_K}$

$$\langle Cx - \nabla L(x), x\rangle \leq -m\|x\|^2 + c.$$

$$L(x) + \frac{\lambda}{2}\|x\|^2_{\mathcal{H}_K}$$

Assumption (2)

- Smoothness:

$$\|\nabla L(x) - \nabla L(y)\| \le M\|x - y\|$$

- Strong smoothness condition:

For $\alpha \in (1/4, 1)$, （これが無い場合はレートが遅くなる）

$$\|\nabla L(x) - \nabla L(y)\|_{-\alpha} \le M\|x - y\|$$

where $\|x\|_\varepsilon = \left(\sum_{k\ge 0}(\mu_k)^{2\varepsilon}|\langle x, f_k\rangle|^2\right)^{1/2}$.

(This is not standard, but, is satisfied in the previous examples)

- Third order smoothness:

Let $L_N = L(P_N x)$. There exists $\alpha' \in [0, 1)$ such that

$$\|D^3 L_N(x)\cdot(h,k)\|_{\alpha'} \le C_{\alpha'}\|h\|_0\|k\|_0,$$
$$\|D^3 L_N(x)\cdot(h,k)\|_0 \le C_{\alpha'}\|h\|_{-\alpha'}\|k\|_0.$$

誤差の解析

π_∞:定常分布

Thm (informal) [Muzellec, Sato, Massias, Suzuki, arXiv:2003.00306 (2020)]

上記の条件のもと，次が成り立つ：

Ξ_k とする

$$L(X_k) - \int L(x)\mathrm{d}\pi_\infty(x) \lesssim \exp\left(-\Lambda_\eta^* k\eta\right) + \frac{c_\beta}{\Lambda_0^*}\eta^{1/2-\kappa}$$

(geometric ergodicity + time discretization)

ただし $\kappa > 0$ は任意の正の実数, $c_\beta = \sqrt{\beta}$ (有界な勾配), $c_\beta = 1$ (強散逸条件), Λ_η^*はスペクトルギャップ.

Remark: $$\int L(x)\mathrm{d}\pi_\infty(x) \simeq L(\tilde{x}) \quad \text{for} \quad \tilde{x} := \operatorname*{arg\,min}_{x\in\mathcal{H}} \left\{ L(x) + \frac{\lambda}{2}\|x\|_{\mathcal{H}_K}^2 \right\}$$

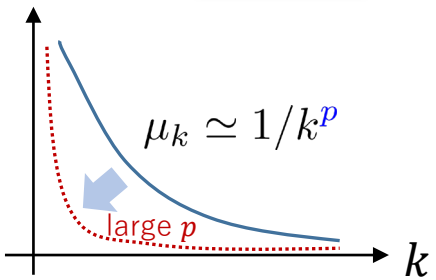
証明は以下の論文のテクニックを援用: Brehier 2014; Brehier&Kopec 2016; Mattingly et al., 2002; Goldys&Maslowski, 2006.

- Coupling argument, マリアバン解析

有限次元バージョンとの関係

参考

$$\mathcal{H}_K = \left\{ \sum_{k=0}^{\infty} \alpha_k f_k \mid \sum_{k=0}^{\infty} \alpha_k^2 / \mu_k < \infty \right\}$$



- $\mu_k \simeq 1/k^2$ （我々の状況）

$$|\mathbb{E}[\phi(X_n) - \phi(X^\pi)]| \leq C \left[\exp\left(-\Lambda_\eta^* n\eta\right) + \frac{c_\beta}{\Lambda_0^*} \eta^{1/2-\kappa} \right] \quad \text{(optimal)}$$

- $\mu_k \simeq 1/k^p$ （予想）see [Andersson,Kruse&Larsson, 2016] for finite time horizon.

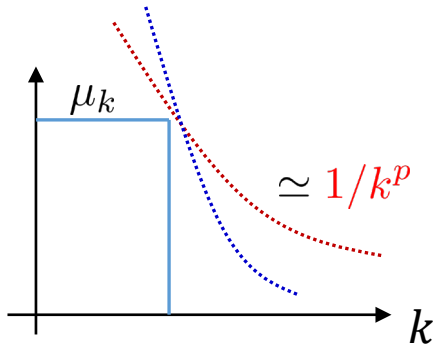
pが大きくなるほど関数クラスは“単純”になる.

$$|\mathbb{E}[\phi(X_n) - \phi(X^\pi)]| \leq C \left[\exp\left(-\Lambda_\eta^* n\eta\right) + \frac{c_\beta}{\Lambda_0^*} \eta^{\frac{p-1}{p}-\kappa} \right]$$

有限次元の解析は$p \to \infty$に対応（定数を無視すれば）:

$$|\mathbb{E}[\phi(X_n) - \phi(X^\pi)]| \leq C \left[\exp\left(-\Lambda_\eta^* n\eta\right) + \frac{c_\beta}{\Lambda_0^*} \eta \right]$$

[Xu et al. (2018)]



誤差の分解

参考

弱収束を示す:

$$|\mathbb{E}[\phi(X_n)] - \phi(x^*)| \leq ?$$

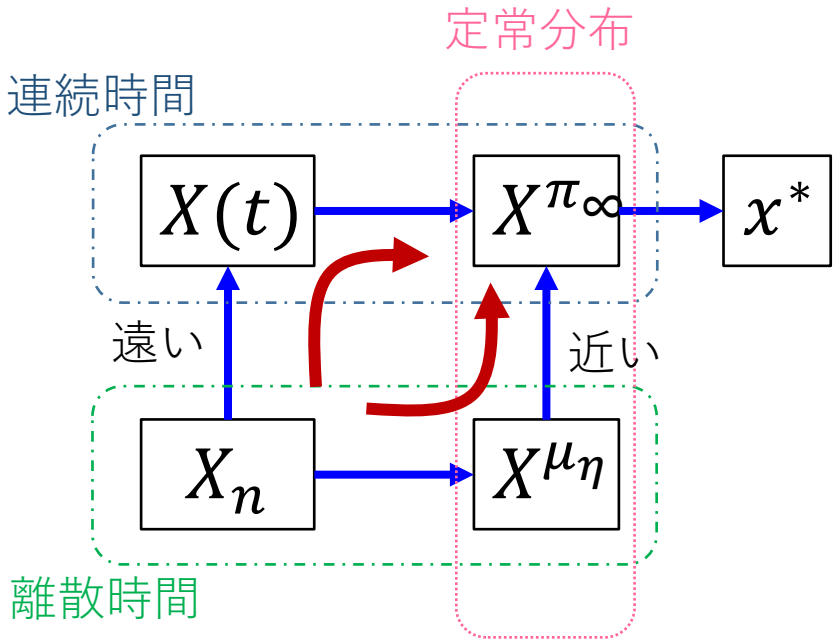
ただし，ϕは滑らかな関数.

- Raginsky et al. (2017),
  Bréhier (2014), Bréhier and Kopec (2016):

$$\mathbb{E}[\phi(X_n) - \phi(X(n\eta))] + \mathbb{E}[\phi(X(n\eta)) - \phi(X^{\pi_\infty})] + \mathbb{E}[\phi(X^{\pi_\infty}) - \phi(x^*)]$$

- Xu et al. (2018):

$$\mathbb{E}[\phi(X_n) - \phi(X^{\mu_\eta})] + \mathbb{E}[\phi(X^{\mu_\eta}) - \phi(X^{\pi_\infty})] + \mathbb{E}[\phi(X^{\pi_\infty}) - \phi(x^*)]$$



レートが速い，一方でより強い条件が必要
(Strong smoothness)

第一項のバウンド

参考

$$\mathbb{E}[\phi(X_n) - \phi(x^*)] = \mathbb{E}[\phi(X_n) - \phi(X^{\mu_\eta})] + \mathbb{E}[\phi(X^{\mu_\eta}) - \phi(X^{\pi_\infty})] + \mathbb{E}[\phi(X^{\pi_\infty}) - \phi(x^*)]$$

補題 (離散時間ダイナミクスのGeometric ergodicity)

ある定常分布μ_ηがだた一つ存在して (極限分布),
geometric ergodicity (定常分布への線形収束) が成り立つ:

$$\mathbb{E}[\phi(X_n) - \phi(X^{\mu_\eta})] \leq C(1 + \|x_0\|) \exp\left(-\Lambda_\eta^* n\eta\right)$$

ただし，“スペクトルギャップ” Λ_η^* は以下のように与えられる,

(i) (Strict dissipative)

$$\Lambda_\eta^* = \frac{\frac{\lambda}{\mu_0} - M}{1 + \eta\frac{\lambda}{\mu_0}}$$

(ii) (Bounded gradient)

$$\Lambda_\eta^* = C \min\left(\frac{\lambda}{2\mu_0}, \frac{1}{2}\right)\delta$$

for $\quad \delta = \exp(-O(\beta))$

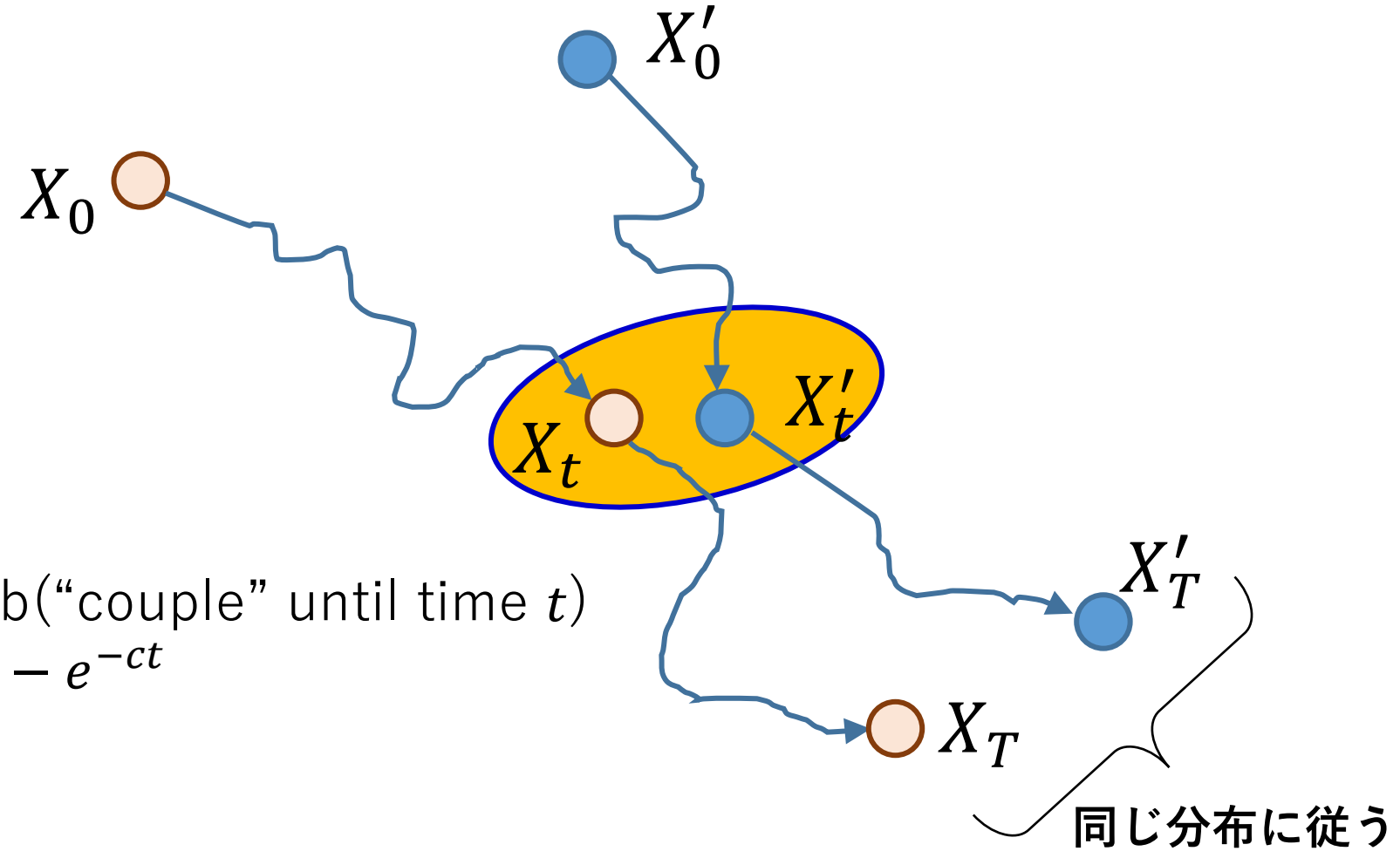
X^{μ_η}: r.v. obeying μ_η $\qquad$ $X_0 = x_0$ (constant)

- 有限次元の場合と違い，強平滑条件がないとおそらく成り立たない.
- **Coupling argument**: Lyapunov条件, majorization条件より
(Mattingly et al. (2002)とGoldys&Maslowski (2006)のテクニックを合わせる)

Geometric ergodicity

参考

- Coupling argument



Prob(“couple” until time t)
$\geq 1 - e^{-ct}$

汎化誤差解析

[Suzuki: Generalization bound of globally optimal non-convex neural network training: Transportation map estimation by infinite dimensional Langevin dynamics. NeurIPS2020. (arXiv:2007.05824)]

問題設定

$$f_W(x) := \int_{\mathbb{R}^d} W_2(w)\sigma(W_1(w)^\top x)\mathrm{d}\rho_0(w)$$

$$L(W) := \mathbb{E}[\ell(f_W, Z)] \qquad \widehat{L}(W) := \frac{1}{n}\sum_{i=1}^{n} \ell(f_W, z_i)$$

汎化誤差:

$$L(\widehat{W}) - \widehat{L}(\widehat{W})$$

残余誤差 (Excess risk):

$$L(\widehat{W}) - \inf_{f:\text{measurable}} \mathbb{E}[\ell(f, Z)]$$

学習の方法 (無限次元GLD):

$$\mathrm{d}W_t = -\nabla\left(\widehat{L}(W_t) + \frac{\lambda}{2}\|W_t\|_{\mathcal{H}_K}^2\right)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}\xi_t$$

時間の離散化

$$W_{k+1} = S_\eta\left(W_k - \eta\nabla\widehat{L}(W_k) + \sqrt{2\frac{\eta}{\beta}}\xi_k\right)$$

$$\left(S_\eta := (I + \eta\tfrac{\lambda}{2}\nabla\|\cdot\|_{\mathcal{H}_K})^{-1}\right)$$

(再掲) 連続時間ダイナミクスの定常Gibbs分布:

$$\frac{\mathrm{d}\pi_\infty}{\mathrm{d}\mu_\beta}(x) \propto \exp\left(-\beta\widehat{L}(x)\right)$$

(擬-)Bayes事後分布

$$\mu_\beta = N(0, C) \text{ where } C = (\beta\lambda)^{-1}\mathrm{diag}(\mu_0, \mu_1, \dots).$$

汎化誤差バウンド

最適化誤差： $\widehat{L}(W_k) - \int \widehat{L}(w)\mathrm{d}\pi_\infty(w) \lesssim \underbrace{\exp\left(-\Lambda_\eta^* k\eta\right) + \frac{c_\beta}{\Lambda_0^*}\eta^{1/2-\kappa}}_{\Xi_k}$

Thm (汎化誤差バウンド)

For any $\kappa > 0$,

$$\mathbb{E}_{W_k}[L(W_k)] \leq \mathbb{E}_{W_k}[\widehat{L}(W_k)] + \frac{R^2}{\sqrt{n}}\left[2\left(1+\frac{2\beta}{\sqrt{n}}\right) + \log\left(\frac{1+e^{R^2/2}}{\delta}\right)\right] + \Xi_k$$

with probability $1-\delta$, where

$O(1/\sqrt{n})$

PAC-Bayesian stability bound

[Rivasplata, Kuzborskij, Szepesvári, and Shawe-Taylor, 2019]

仮定

- 損失関数はWについて“十分に滑らか”.
- 損失関数は有界:

$0 \leq \ell(f_W, z) \leq R, \quad \|\nabla_W \ell(f_W, z)\|_{\mathcal{H}} \leq R \quad (\forall W \in \mathcal{H},\ z \in \mathrm{supp}(P))$

残余誤差の評価

$\widehat{L}(W) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_W, z_i)$　　$L(f) = \mathbb{E}[\ell(f, Z)]$

残余誤差 (Excess risk): $L(\widehat{W}) - \inf_{f:\text{measurable}} L(f)$

追加の仮定: $(T_K^a x := \sum_{k=0}^{\infty} \mu_k^a x_k e_k \text{ where } x = \sum_{k=0}^{\infty} x_k e_k \text{ and } \|x\|_{\mathcal{H}_K}^2 = \sum_{k=0}^{\infty} \mu_k x_k^2)$

- $\exists W^* \in \mathcal{H}$ s.t. $\inf_f L(f) = L(f_{W^*})\ (= L(f^*))$
- $\exists \gamma > 1/4$ **：モデルの複雑さを制御**

  $\tilde{\ell}(W, z) = \ell\left(f_{T_K^{\gamma/2} W}, z\right)$ ➡ $L(W) = \mathbb{E}[\tilde{\ell}(W, Z)]$
- **Bernstein条件** [Erven et al., 2015]:

  $\mathbb{E}[(\ell(f, Z) - \ell(f^*, Z))^2] \le B(L(f) - L(f^*))^s$
  - 二乗損失: $s = 1$
  - ロジスティック損失 with 有界なf, f^*: $s = 1$
- $\mathbb{E}\left[\exp\left(-\frac{\beta}{n}(\ell(f, Z) + \ell(f^*, Z))\right)\right] \le 1$
  - 損失関数は対数尤度である必要はない.
  - 真の分布が軽い裾を持っていることを仮定.

Fast rate: 一般形

Let $\mathcal{H}_{\tilde{K}} = T_K^{(\gamma+1)/2}\mathcal{H}$ and $\mathcal{H}_{\tilde{K}^\theta} = T_K^{\theta(\gamma+1)/2}\mathcal{H}$.

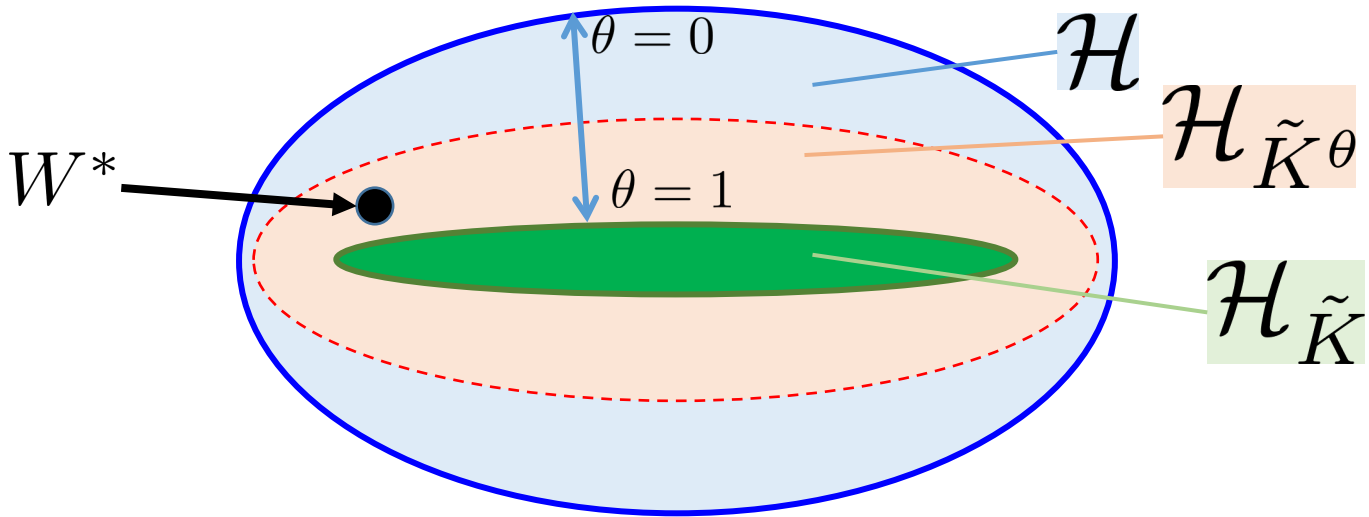
Thm (Fast rate: excess risk bound)

Suppose that $W^* \in \mathcal{H}_{\tilde{K}^\theta}$ for $0 < \theta < 1 - \frac{1}{2(\gamma+1)}$.

Then, for $\tilde{\alpha} = \frac{1}{2(\gamma+1)}$, it holds that

$$\mathbb{E}_{D_n}\left[\mathbb{E}_{W_k}[L(W_k)] - L(W^*)\right] \lesssim \max\left\{(\lambda\beta)^{\frac{2\tilde{\alpha}/\theta}{2-s(1-\tilde{\alpha}/\theta)}} n^{-\frac{1}{2-s(1-\tilde{\alpha}/\theta)}}, \lambda^{-\tilde{\alpha}}\beta^{-1}, \lambda^{\theta}\right\} + \Xi_k$$

Can be faster than $O(1/\sqrt{n})$



Fast rate: 回帰

$\ell(f,z) = (f(x) - y)^2$: 二乗損失

$$f_W(x) := \int_{\mathbb{R}^d} W_2(w)\sigma(W_1(w)^\top x)\mathrm{d}\rho_0(w)$$

- $\mathcal{H}$: $L_2(\rho_0)$
- $\mathcal{H}_{\tilde{K}}$: $W^{a+d/2}(\mathbb{R}^d)$ (Sobolev space)　（$\gamma = a/d + 1/2$に相当）
- $\theta = \frac{2b}{2a+d}$ for $b < a$

$\lambda^{-1} = \beta = n$ とすることで

$$\mathbb{E}_{D_n}\left[\mathbb{E}_{W_k}[L(W_k)] - L(W^*)\right] \lesssim n^{-\frac{2\min\{a,b\}}{2a+d}} + \Xi_k$$

Sobolev空間のミニマックス最適レートに一致する
($a = b$の時).

判別問題における速い収束

Assumption

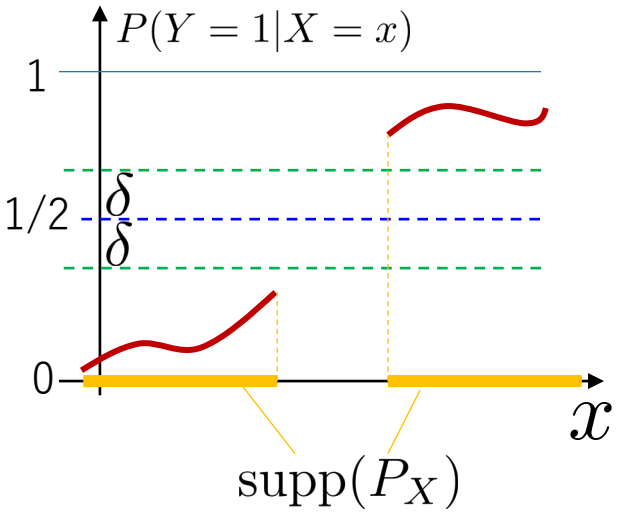
- **強低ノイズ条件:**

$$|P(Y=1|X)-1/2| \geq \delta \quad \text{(a.s.)}$$

- $\mathrm{supp}(P_X) \subset [0,1]^d$ and P_X has density p such that $p(x) \geq c_0 \quad (\forall x \in \mathrm{supp}(P_X))$.
- 活性化関数はなめらか:

$$\sigma \in \mathcal{C}^m(\mathbb{R}) \qquad \text{for } 2m > d$$

- 真の関数はモデルに入っているとする: $f^* = f_{W^*}$.



$$f_W(x) = \sum_{j=1}^{\infty} a_j \eta(w_j^\top x)$$

十分大きなnと$\beta \leq n$に対し,

$$\mathbb{E}[P_{\pi_k}(\{W_k \in \mathcal{H} \mid P_X[\mathrm{sign}(f_{W_k}(X)) = \mathrm{sign}(f^*(X))] \neq 0\})]$$

classification error

$$\lesssim \exp(-c\beta\delta^{2m/(2m-d)}) + \frac{\Xi_k}{\delta^{2m/(2m-d)}}$$

ベイズ最適な判別機が高い確率で求まる. (βは定数のままでも良い)

カーネル法(線形推定量)との比較

[Suzuki&Akiyama: Benefit of deep learning with non-convex noisy gradient descent: Provable excess risk bound and superiority to kernel methods. network training: Transportation map estimation by infinite dimensional Langevin dynamics. ICLR2021]

モデル (Teacher-student model)

教師モデル：

$$f_W(x) = \sum_{m=1}^{\infty} a_m w_{2,m} \sigma(b_m^{-1} w_{1,m}{}^{\top} x)$$

$W = (w_{1,m}, w_{2,m})_{m=1}^{\infty}$ ：学習可能パラメータ

$(a_m, b_m)_{m=1}^{\infty}$ ：固定パラメータ

$$\mathcal{H}_\gamma := \left\{ W = (w_{1,m}, w_{2,m})_{m=1}^{\infty} \mid \|W\|_{\mathcal{H}_\gamma}^2 = \textstyle\sum_{m=1}^{\infty} (w_{1,m}^2 + \|w_{2,m}\|^2)/\mu_m^\gamma < \infty \right\}$$

$$\mathcal{F}_\gamma := \{ f_W \mid W \in \mathcal{H}_\gamma,\ \|W\|_{\mathcal{H}_\gamma} \leq 1 \}$$

観測モデル： $f^\circ \in \mathcal{F}_\gamma$ (真の関数) を用いて

$$y_i = f^\circ(x_i) + \varepsilon_i \qquad (i = 1, \ldots, n)$$

$D_n = (x_i, y_i)_{i=1}^n$ (観測データ) から真の関数を推定.

条件

条件

- $\mu_m \propto m^{-2}$
- $a_m \propto \mu_m^{\alpha_1}$ for $\alpha_1 > 1/2$
- $b_m \propto \mu_m^{\alpha_2}$ for $\alpha_2 > \gamma/2$
- 活性化関数σは三回微分可能でその微分係数は有界

$$f_W(x) = \sum_{m=1}^{\infty} a_m w_{2,m} \sigma(b_m^{-1} w_{1,m})$$

深層学習： 二乗損失を採用

$$W_{k+1} = W_k - \eta \left(\nabla_W \widehat{L}(f_{W_k}) + \frac{\lambda}{2} \nabla_W \|W_{k+1}\|_{\mathcal{H}_1}^2 \right) + \sqrt{2\frac{\eta}{\beta}} \xi_k$$

(一見難解に見えるが単に各成分ごと勾配法で更新してノイズを加えているだけ)

Thm (深層学習の推定精度上界)

$$\mathbb{E}_{D^n} \left[\mathbb{E}_{W_k} [\| f_{W_k} - f^\circ \|_{L_2(P_X)}^2 | D_n] \right] \lesssim n^{-\frac{\gamma}{\alpha_1 - 3\alpha_2 + 1}} + \Xi_k$$

線形推定量

$$X_n = (x_1, \ldots, x_n)$$

$$\hat{f}(x) = \sum_{i=1}^{n} \varphi_i(x; X_n) y_i$$

線形

- Kernel ridge estimator
- Sieve estimator
- Nadaraya-Watson estimator
- k-NN estimator

Kernel ridge regression:

$$\hat{f}(x) = K_{x,X}(K_{X,X} + \lambda \mathrm{I})^{-1} Y$$

$$R_{\mathrm{lin}}(\mathcal{F}_\gamma) := \inf_{\hat{f}:\text{linear}} \sup_{f^\circ \in \mathcal{F}_\gamma} \mathbb{E}_{D_n}[\|\hat{f} - f^\circ\|^2_{L_2(P_X)}]$$

最悪誤差を最小化する推定量の推定誤差

：線形推定量のミニマックス誤差

Thm (線形推定量の推定精度下界)

$\tilde{\beta} = \frac{\alpha_1 + \alpha_2}{\alpha_2 - \gamma/2}$ として，任意の$\kappa' > 0$に対して，以下が成り立つ:

$$R_{\mathrm{lin}}(\mathcal{F}_\gamma) \gtrsim n^{-\frac{2\tilde{\beta}+d}{2\tilde{\beta}+2d} - \kappa'}$$

いかなるカーネルを用いてもこのレートは越えられない．

レートの比較

- 深層学習の残余誤差

$$n^{-\frac{\gamma}{\alpha_1-3\alpha_2+1}}$$

- 線形推定量の残余誤差下限

$$R_{\mathrm{lin}}(\mathcal{F}_\gamma) \gtrsim n^{-\frac{2\tilde{\beta}+d}{2\tilde{\beta}+2d}} \qquad \left(\tilde{\beta} = \frac{\alpha_1+\alpha_2}{\alpha_2-\gamma/2}\right)$$

線形推定量は次元dに依存→高次元で性能悪化（次元の呪い）

例: $\alpha_1 = \gamma + 3\alpha_2, \alpha_2 = 4\gamma$とする.

深層

$$n^{-\left(1+\frac{1}{\gamma}\right)^{-1}}$$

$1/n$ (large γ)

線形（カーネル法）

$$n^{-\left(1+\frac{d}{d+11.3}\right)^{-1}}$$

$1/\sqrt{n}$ (large d)

現実的な最適化の保証ありで性能差を証明

まとめ

深層学習の統計理論→非凸性が重要→非凸性を残した最適化理論.

パラメータのダイナミクス→輸送写像のダイナミクス

- 無限次元Langevin動力学
  - 弱収束の収束速度
  - 正則化を入れることで無限次元での収束を保証
- 無限次元Langevin動力学の汎化誤差理論
  - 擬-ベイズ事後分布
  - 汎化ギャップ
  - Fast rateの導出→**非凸最適化とノンパラ統計の接点**
- Teacher-studentの設定における深層学習の優位性
  - 最適化アルゴリズムも含めた理論

何がまだ足りないか?

- 深層学習の適応能力 (minimax最適性).
  - Hölder class [Schmidt-Hieber, 2017]
  - Besov space [Suzuki, 2019][Hayakawa&Suzuki, 2019]
  - Piece-wise smooth [Imaizumi&Fukumizu, 2018]
  - Anisotropic Besov [Suzuki&Nitanda, 2019]

→ **最適化理論と統計理論の融合**

- 目的関数の構造を最大限考慮に入れた最適化誤差解析