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理論と汎化誤差解析

鈴木大慈

東京大学/ 理研AIP

(深層学習理論チーム)

THE UNIVERSITY OF TOKYO AIP

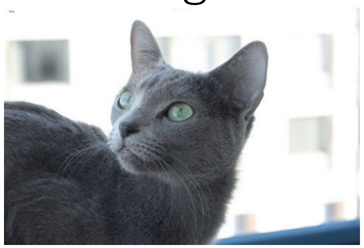
2020年9月4日@九州大学統計科学セミナー

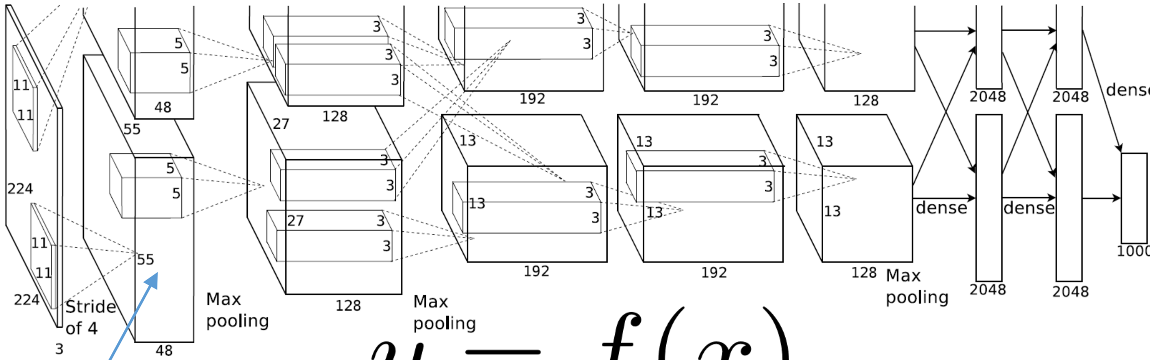
導入：問題意識

Deep Learning Model

Alex-net [Krizhevsky, Sutskever + Hinton, 2012]

Image





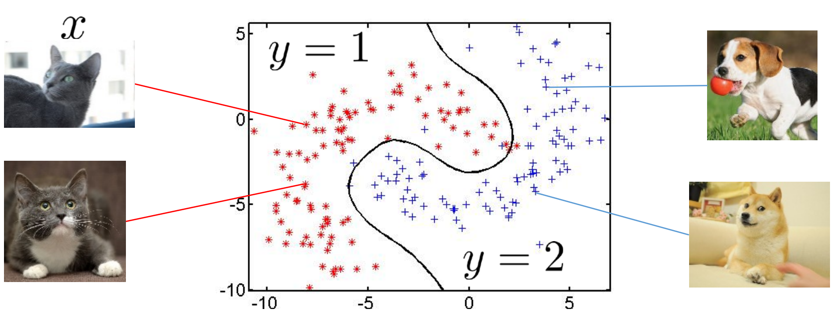
-Cat (y=1)
-Dog(y=2)

$$x$$

$$y = f(x)$$

$$y$$

$$f(x) = (W^{(L)}\eta(\cdot) + b^{(L)}) \circ (W^{(L-1)}\eta(\cdot) + b^{(L-1)}) \circ \cdots \circ (W^{(1)}x + b^{(1)})$$



- Stacking layers yields a complicated function.
- Universal approximator.

Details of each layer

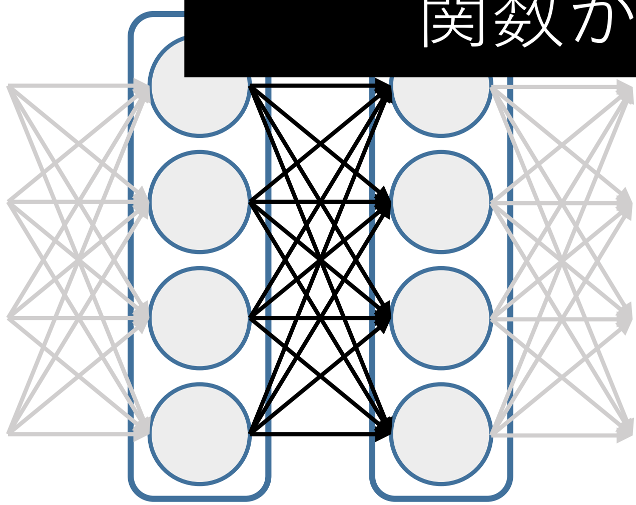
Affine transform + activation function

$$\phi_{\ell+1}(x) = \eta(W^{(\ell)}\phi_\ell(x) + b^{(\ell)})$$

$$W^{(\ell)} \in \mathbb{R}^{m_{\ell+1} \times m_\ell}$$

$b^{(\ell)} \in$

万能近似能力
(十分広い横幅のNNであらゆる
関数が近似できる)

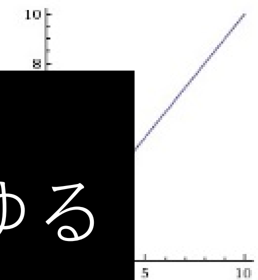


$\phi_\ell(x)$ $\phi_{\ell+1}(x)$

Activation function

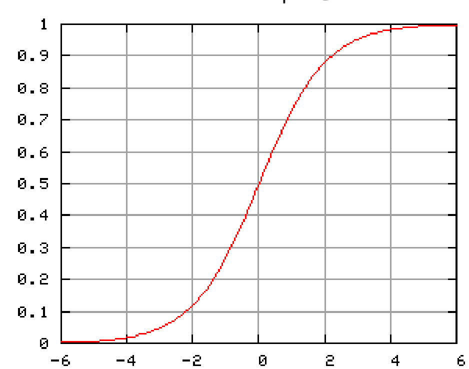
- ReLU (Rectified Linear Unit)

$$\eta(u) = \max\{u, 0\}$$



- Sigmoid function

$$\eta(u) = \frac{1}{1 + e^{-u}}$$



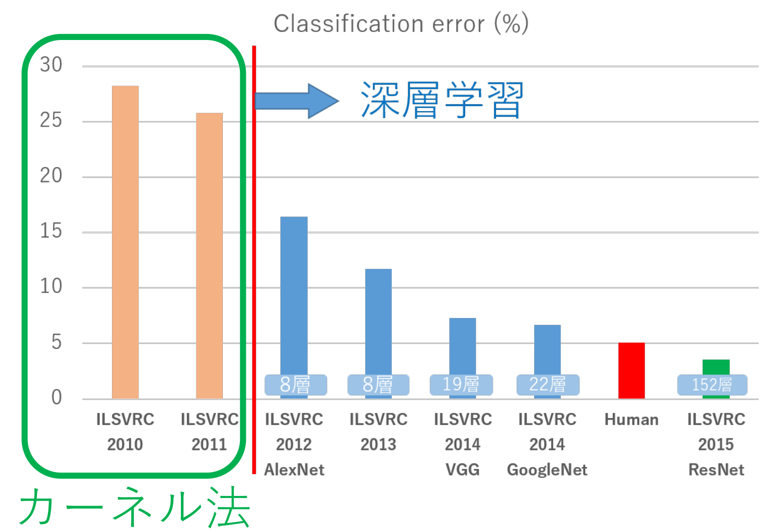
カーネル法

- 万能近似能力のある手法.
- 2000年～2010年ほどで流行.
- ILSVRCでも使われていた.

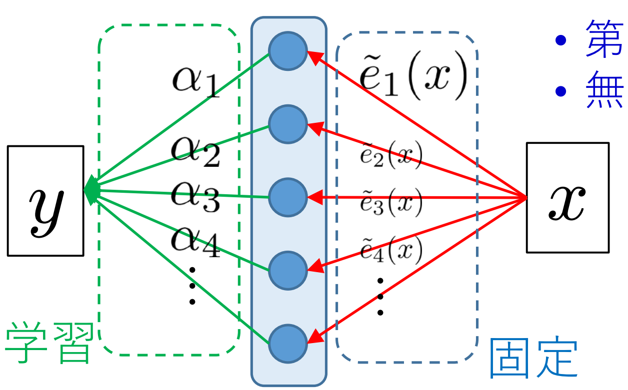
第1層目を固定した横幅無限の2層ニューラルネットワーク

(線形推定量と呼ばれるクラス)

似た手法
- スプライン法
- 局所多項式回帰
- シリーズ推定量

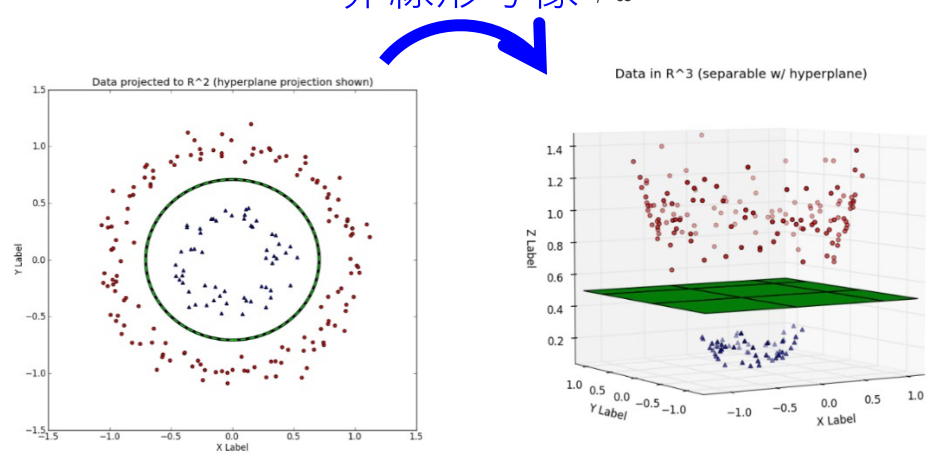


再生核ヒルベルト空間の理論



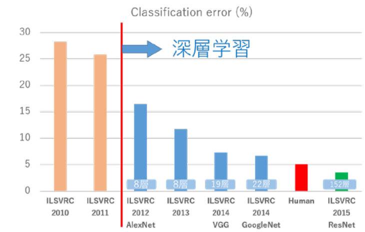
- 第一層目固定
- 無限個の素子

非線形写像 ϕ_x



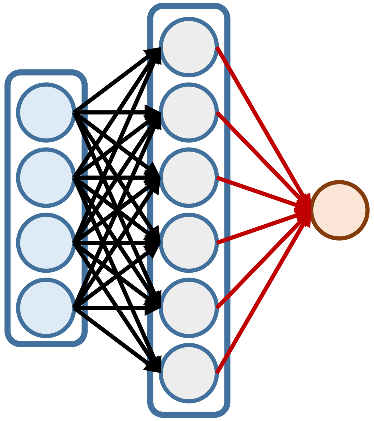
問題意識

- **[理論]** 万能近似能力という意味では浅層で十分.
- **[実際]** 実際は多層を使うことが多い.

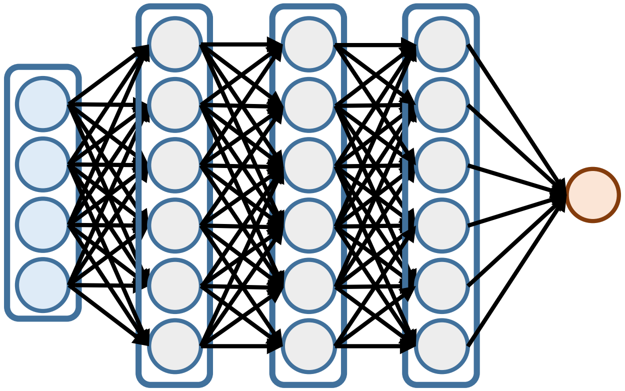


→　この差はどう埋める？

カーネル法
浅層



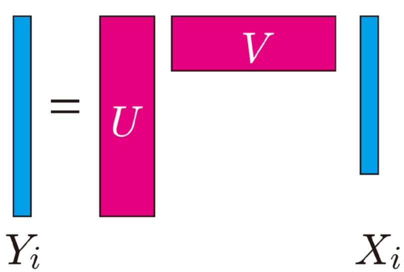
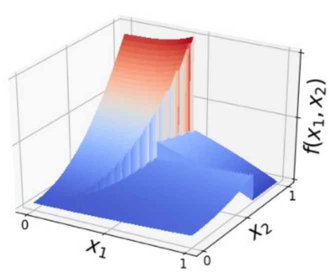
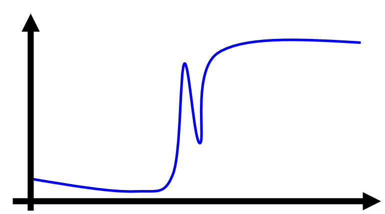
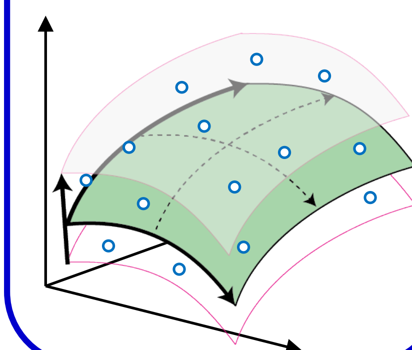
多層ニューラルネット
深層学習



→ 推定能力を比べる.

なぜ深層学習が良いのか？

- 真の関数f°の形状によって深層が有利になる

縮小ランク回帰	区分滑らかな関数	Besov空間	低次元データ
	[Imaizumi&Fukumizu, 2019]	[Suzuki, 2019]	[Schmidt-Hieber, 2019] [Nakada&Imaizumi, 2019][Chen et al., 2019][Suzuki&Nitanda, 2019]
特徴空間の次元が低い状況は深層学習が得意	不連続な関数の推定は深層学習が得意	滑らかさが非一様な関数の推定は深層学習が得意	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
			

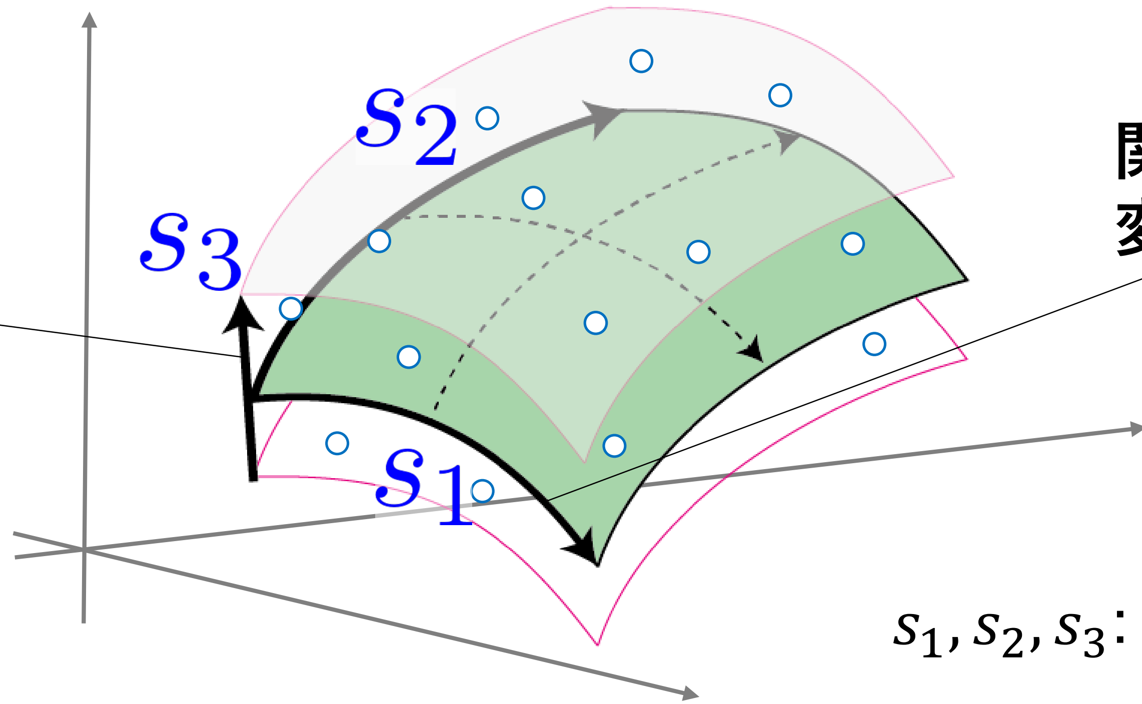
推定精度

	縮小ランク回帰	区分滑らかな関数	Besov空間	低次元データ
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	$\frac{MN}{n}$	$\frac{1}{\sqrt{n}}$	$n^{-\frac{2s-2d(1/p-1/2)_+}{2s+d-2d(1/p-1/2)_+}}$	$n^{-\frac{2(s-D/p+d/2)}{2(s-D/p+d/2)+d}} \vee n^{-\frac{2s}{2s+D}}$

例 (1): 低次元データ構造

[Suzuki&Nitanda, 2019]

関数値
ほぼ一定



関数値が
変化する方向

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

$$n^{-\frac{\tilde{s}}{\tilde{s}+1}}$$

$$\tilde{s} = \left(s_1^{-1} + s_2^{-1} + s_3^{-1}\right)^{-1}$$

浅い学習

$$n^{-\frac{s_1}{s_1+3}}$$

(次元の呪い)

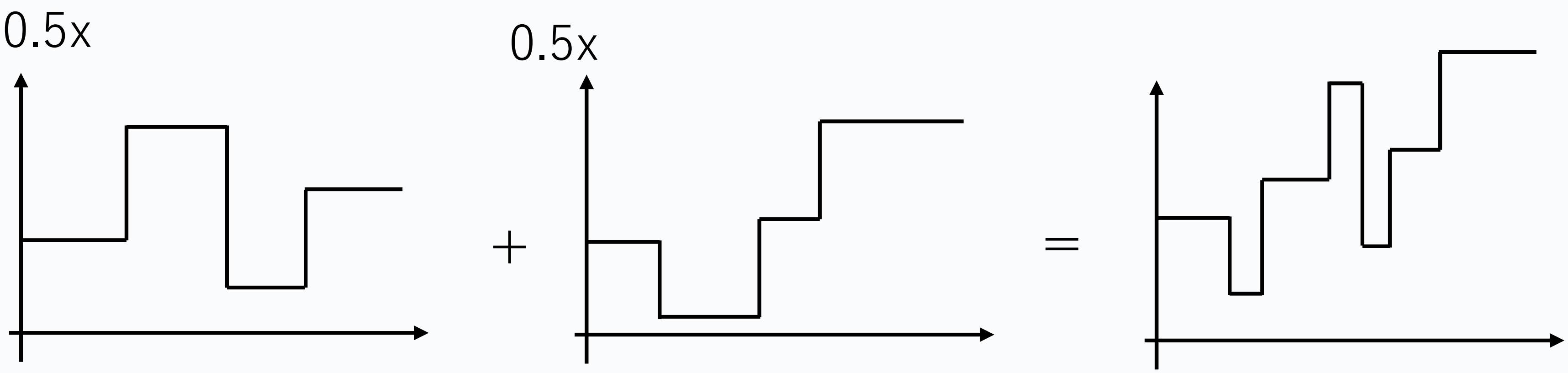
数学的に一般化

「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

→ 様々な性質を“凸性”で統一的に説明

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

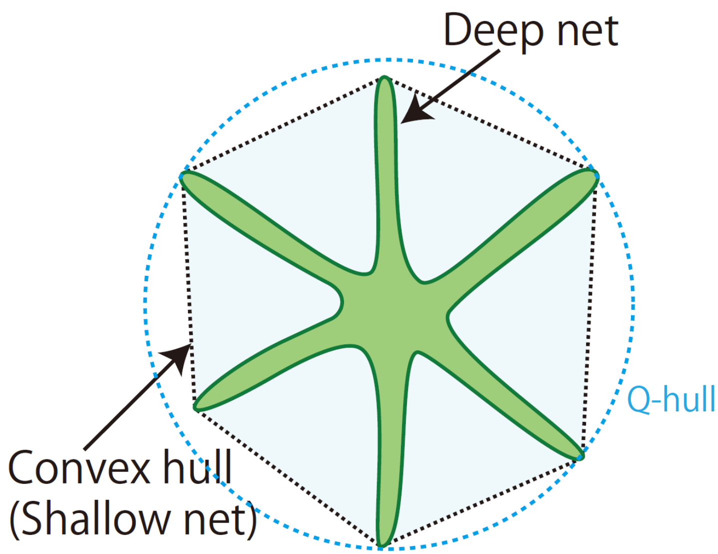
→ さらには「**スパース推定**」という観点からも説明できる．

[Satoshi Hayakawa and Taiji Suzuki: On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. arXiv:1905.09195.]

線形推定量の最悪誤差

線形推定量：$\hat{f}(x) = \sum_{i=1}^n y_i \varphi_i(x_1, \dots, x_n; x) + b$ と書ける**任意の推定量**

例: カーネルリッジ回帰 $\hat{f}(x) = K_{x,X}(K_{X,X} + \lambda \mathrm{I})^{-1} Y$ （“浅い”学習法とみなす）



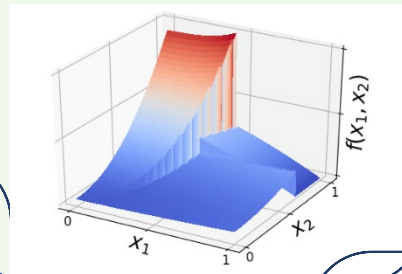
$$\inf_{\hat{f}:\text{Linear}} \sup_{f^\circ \in \mathcal{F}} \mathrm{E}[\|\hat{f} - f^\circ\|_{L_2(P)}^2] = \inf_{\hat{f}:\text{Linear}} \sup_{f^\circ \in \mathrm{conv}(\mathcal{F})} \mathrm{E}[\|\hat{f} - f^\circ\|_{L_2(P)}^2]$$

さらに条件を仮定すれば「Q-hull」まで拡張できる．

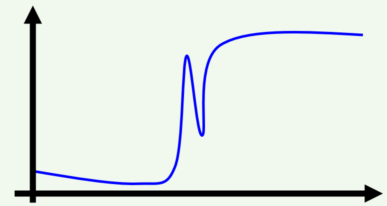
[Hayakawa&Suzuki: 2019][Donoho & Johnstone, 1994]

数学的一般化

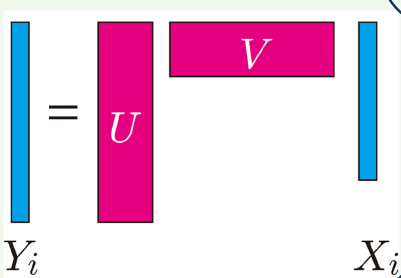
区分滑らかな関数



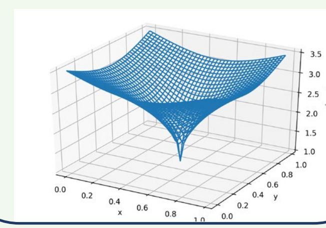
Besov空間



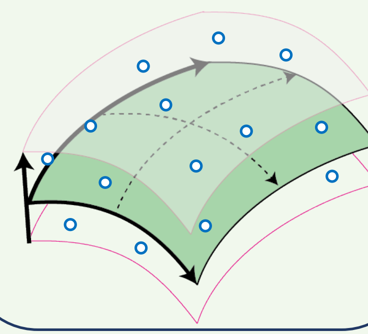
縮小ランク回帰



変動指数
Besov空間



低次元データ



非凸性
スパース性

- これら統計理論は「最適化」を考慮していない.
- 「非凸性」から逃れようとすると深層学習の本当の良さが分からなくなる.

	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	△	△
(既存の)有限次元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	**有限/無限統一的な枠組み**	**汎化ギャップ/余剰誤差ともに保証**	**○**

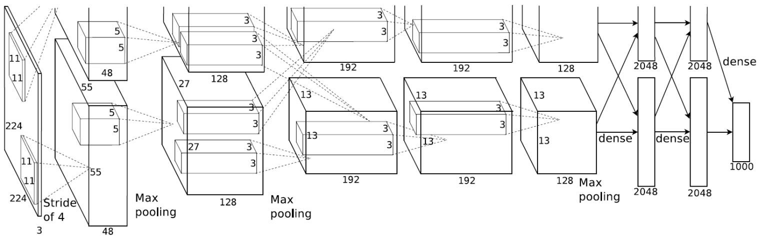
- 深層NNモデルの非凸性を失わず最適化したい．
- モデルサイズをサンプルサイズに依存させたくない．

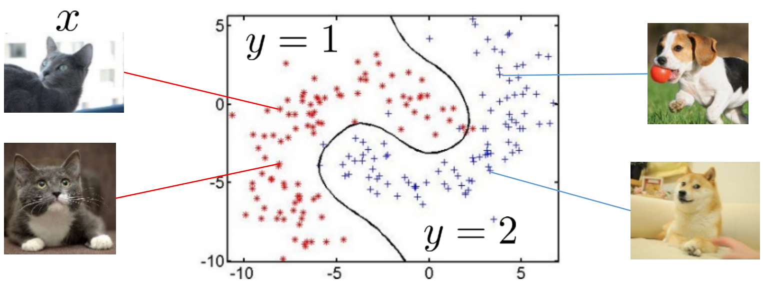
本研究

- ➢ **無限次元Langevin動力学**
- ➢ **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. arXiv:2007.05824]

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

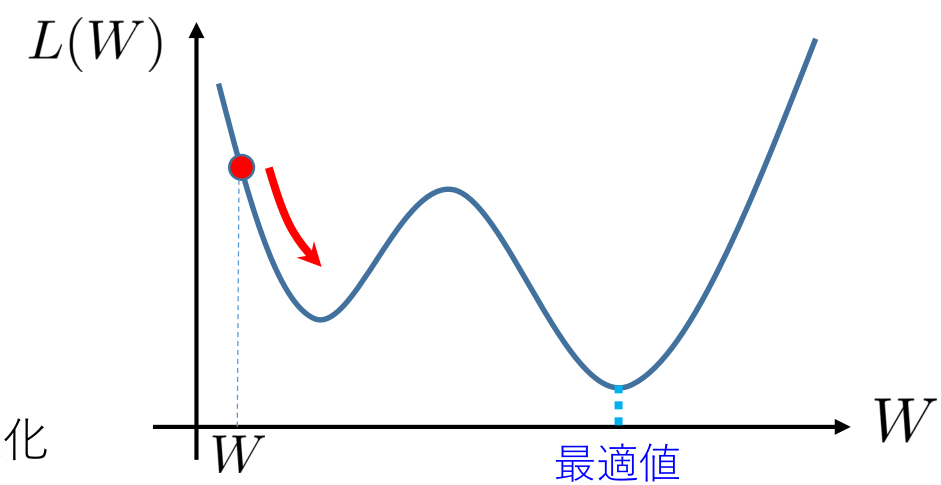
損失関数：データへの当てはまり度合い

損失関数最小化

$$\min_W L(W)$$

(Wは数十億次元)

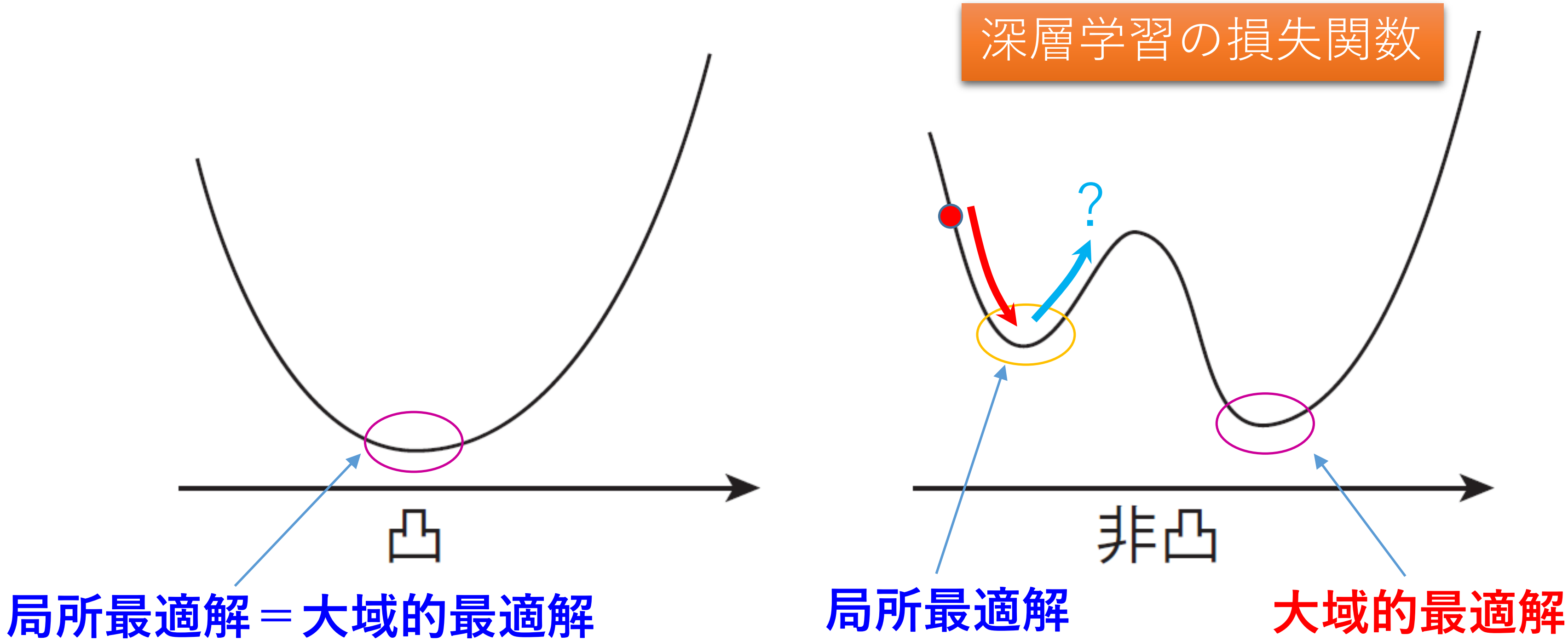
通常，**(確率的)勾配降下法**で最適化



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.

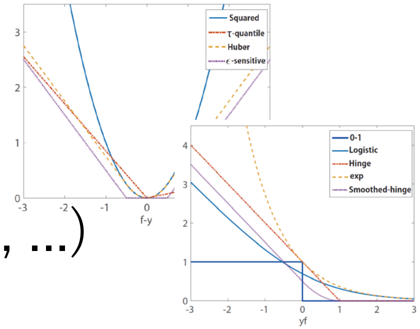
勾配法

$$f_W(x) = \sum_{j=1}^{M} \underbrace{a_j}_{\text{固定}} \eta(\underbrace{w_j^\top}_{\text{最適化}} x)$$

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

ℓ_i: 損失関数

- 回帰: 二乗損失 ($\ell_i(f_W(x_i)) = (y_i - f_W(x_i))^2$)
- 判別: 凸代理損失 (e.g., ロジスティック, 平滑化ヒンジ, ...)

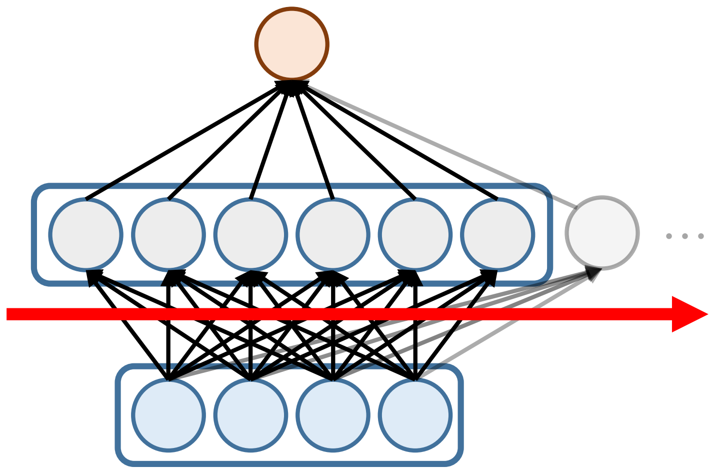


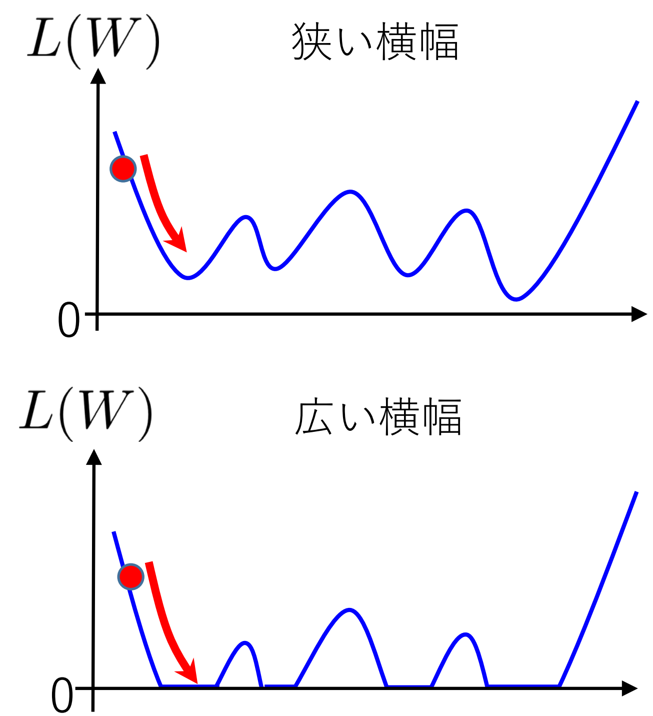
勾配降下法: $W^{(t+1)} = W^{(t)} - \alpha \nabla_W \hat{L}(W)$

確率的勾配降下法: $W^{(t+1)} = W^{(t)} - \alpha \frac{1}{B}\sum_{i\in I_t} \nabla_W \ell_i(f_W(x_i))$

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
  - Mean-field analysis (平均場解析)

二つのスケーリング

固定　最適化

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

- Neural Tangent Kernelのregime (lazy learning)
  - ➢ $a_j = O(1/\sqrt{M})$ [Jacot+ 2018][Du+ 2019][Arora+ 2019]
- 平均場解析のregime
  - ➢ $a_j = O(1/M)$ [Nitanda & Suzuki (2017), Chizat & Bach (2018), Mei, Montanari, & Nguyen (2018)]

※NTKの$1/\sqrt{M}$自体はそこまで本質ではない，$1/M$より大きいことが重要.

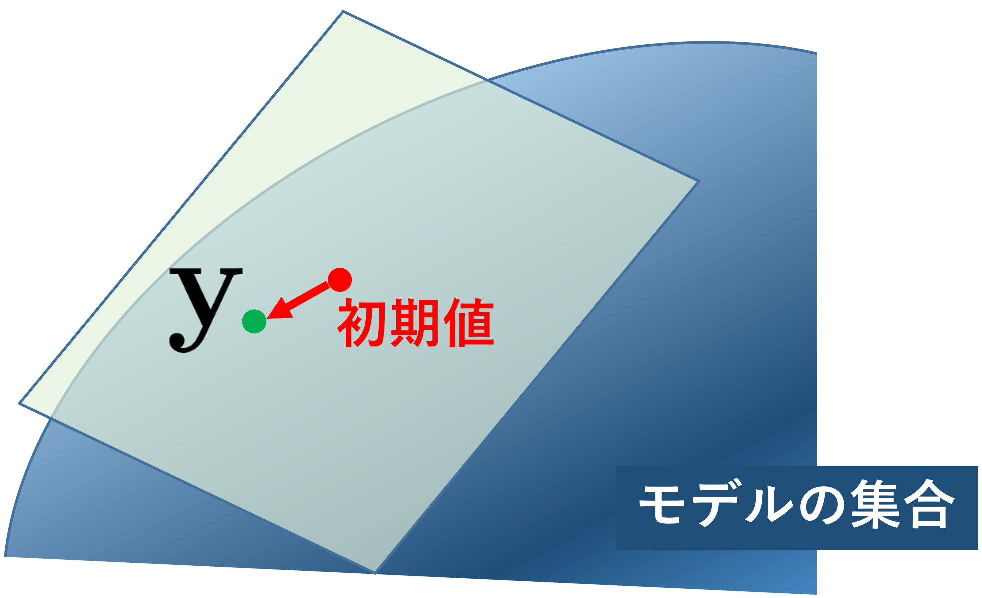
初期化のスケーリングによって，初期値と比べて学習によって動く大きさの割合が変わる.
→ 学習のダイナミクス，汎化性能に影響

(解析の難しさも違う)

NTK

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

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．



Optimization in NTK regime

以下のように初期化する:

- $a_j \sim (\pm 1)\frac{1}{\sqrt{M}}$ (+, − is generated evenly)
- $w_j \sim N(0, I)$

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

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]

- 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.**
- データに完全にフィットさせてしまうので過学習の可能性あり.

Beyond kernel

問題点：NTKは解析がしやすいが，結局カーネル法の範疇なので深層学習の“良さ”が現れない．

➢ NTKをはみ出す理論の試みがいくつかなされている．
（今後発展が予想される）

- Allen-Zhu&Li (2019,2020)

  Allen-Zhu&Li: What Can ResNet Learn Efficiently, Going Beyond Kernels? NIPS2019.
  Allen-Zhu&Li: Backward Feature Correction: How Deep Learning Performs Deep Learning. arXiv:2001.04413.
  (ResNet型ネットワークでカーネルを優越する状況)

- Li, Ma&Zhang (2019)

  Li, Ma&Zhang: Learning Over-Parametrized Two-Layer ReLU Neural Networks beyond NTK. arXiv:2007.04596.
  (テンソル分解の理論で深層学習がカーネルを優越することを示した)

- Bai&Lee (2020)

  Bai&Lee: Beyond Linearization: On Quadratic and Higher-Order Approximation of Wide Neural Networks. ICLR2020.
  (二次のテイラー展開まで使う)

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してみなす.

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

➡ (a,w)に関する確率密度ρによる平均とみなせる:

$$\boxed{f\text{の最適化} \Leftrightarrow \rho\text{の最適化}}$$

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

連続方程式

Wasserstein勾配流

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

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

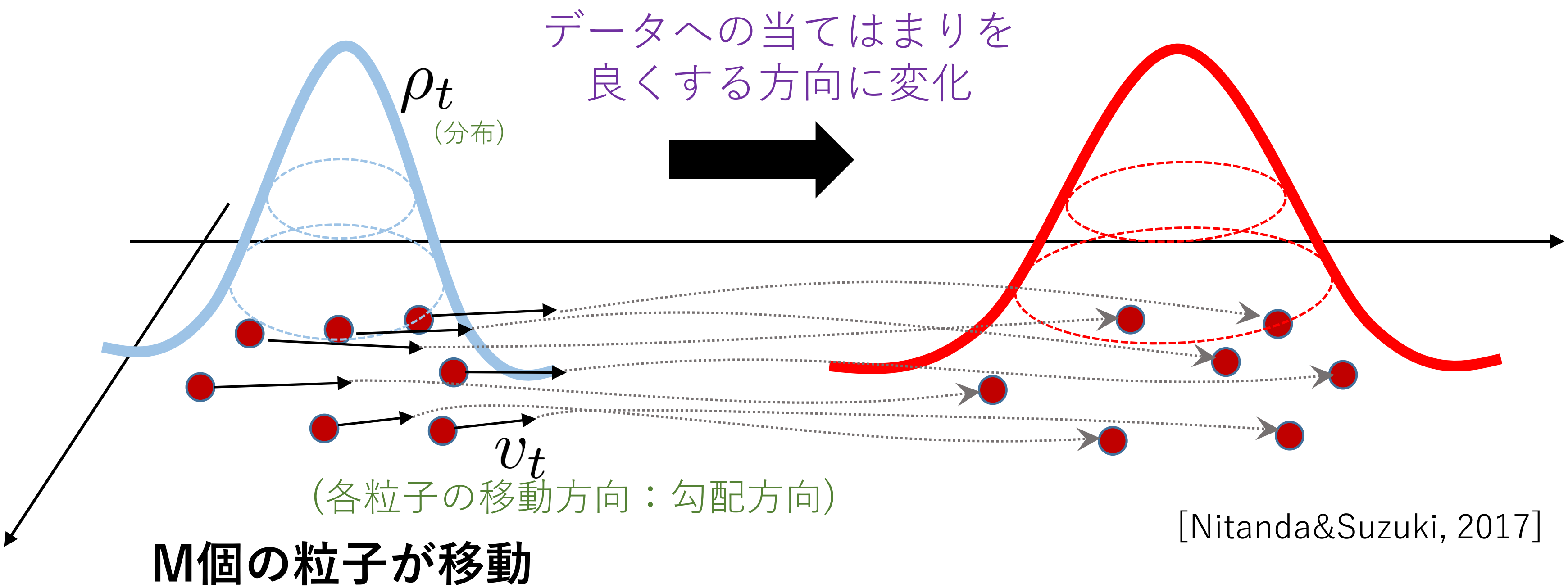
(各粒子は勾配降下方向へ移動)

粒子勾配降下法

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

1つの粒子

- 各ニューロンのパラメータを一つの粒子とみなす.
- 各粒子が誤差を減らす方向に動くことで分布が最適化される.



$M \to \infty$の極限で，最適解への収束が成り立つ場合がある.

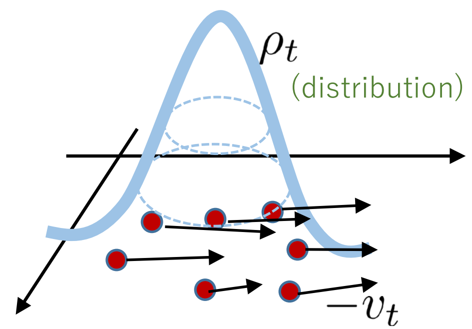
[Nitanda&Suzuki, 2017][Chizat&Bach, 2018][Chizat, 2019]

やはり横幅$M \to \infty$である必要がある

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**):

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

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

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)

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

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

$(\hat{\rho}_t = \frac{1}{M}\sum_{m=1}^{M} \delta_{X_t^m})$

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]

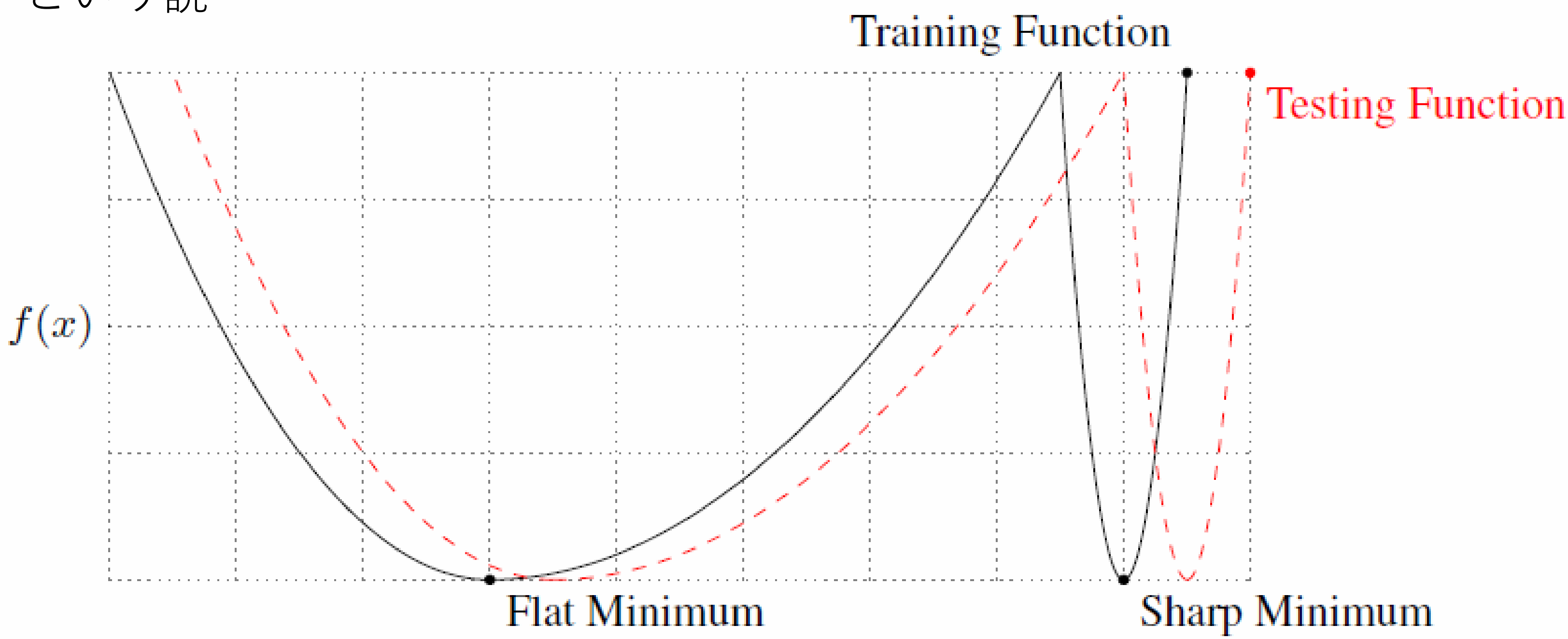
Cons:

- **“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)

Langevin動力学

Sharp minima vs flat minima

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



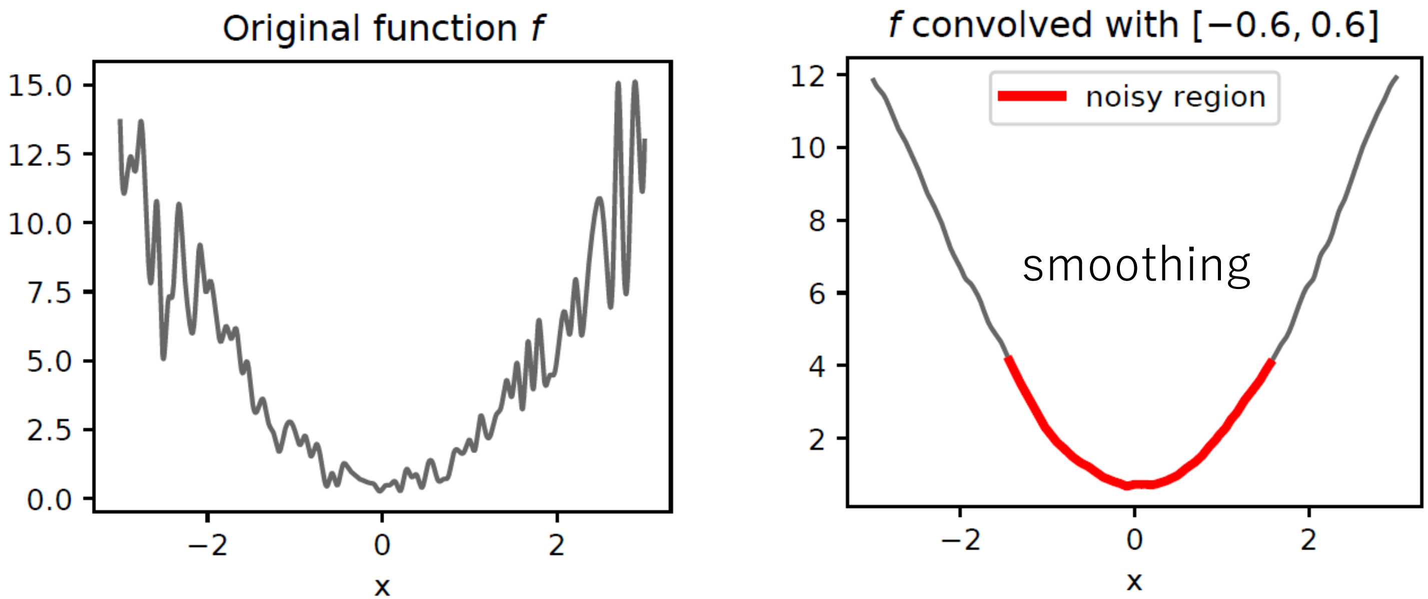
Keskar, Mudigere, Nocedal, Smelyanskiy, Tang (2017):
On large-batch training for deep learning: generalization gap and sharp minima.

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

→ ランダムウォークはフラットな領域にとどまりやすい

- 「フラット」という概念は座標系の取り方によるから意味がないという批判. (Dinh et al., 2017)
- PAC-Bayesによる解析 (Dziugaite, Roy, 2017)

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 ⇒ 目的関数の平滑化

$$x_t = x_{t-1} - \eta(\nabla L(x_{t-1}) + \xi_t) \qquad (y_t = x_t + \eta\xi_t)$$

$$\Rightarrow y_t = y_{t-1} - \eta\xi_{t-1} - \eta\nabla L(y_{t-1} - \eta\xi_{t-1})$$

$$\Rightarrow \mathbb{E}_{\xi_{t-1}}[y_t] = y_{t-1} - \eta\nabla\mathbb{E}_{\xi_{t-1}}[L(y_{t-1} - \eta\xi_{t-1})]$$

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.**

GLD/SGLD

- Stochastic Gradient Langevin Dynamics (SGLD)

$$\min_{x\in\mathbb{R}^d} L(x) = \min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ell_i(x)$$

(非凸)

β: 逆温度

$$\mathrm{d}X_t = -\nabla L(X_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}B_t$$

(勾配Langevin動力学)

定常分布： $\pi \propto \exp(-\beta L(X))$

離散化

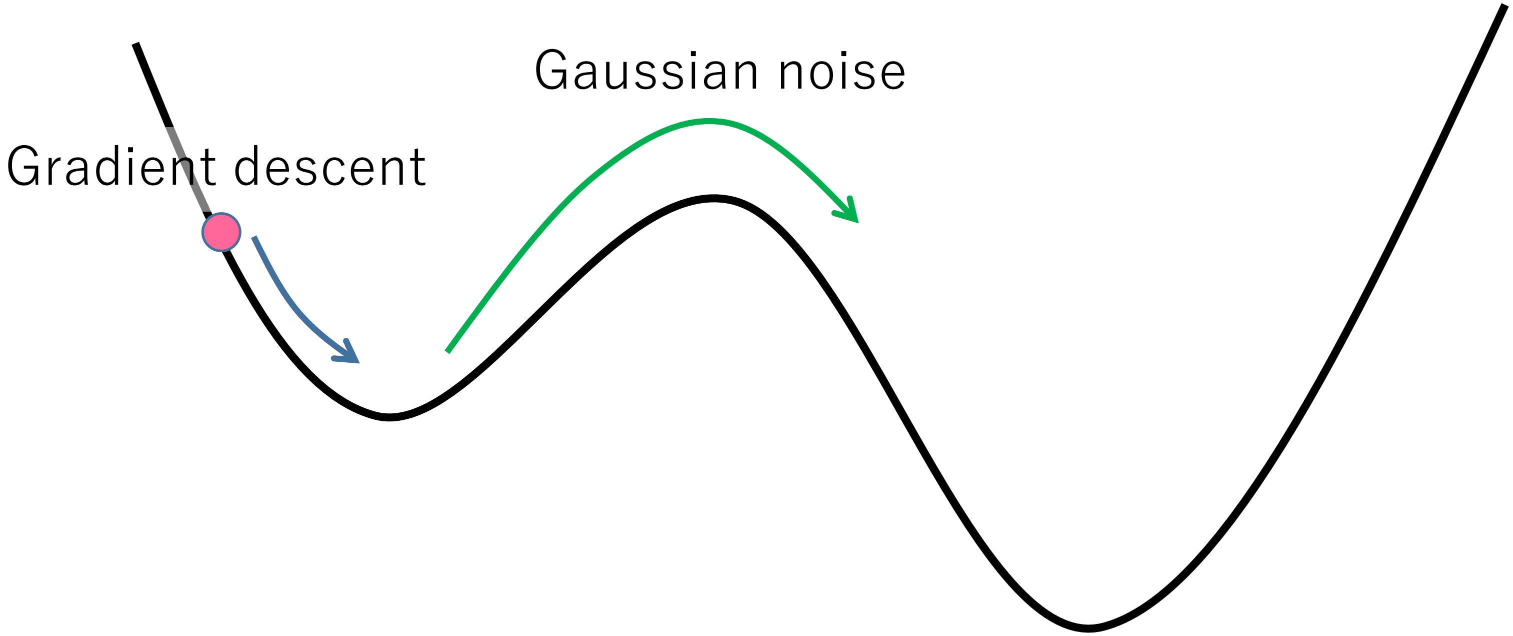
[Gelfand and Mitter (1991); Borkar and Mitter (1999); Welling and Teh (2011)]

GLD: $X_{t+1} = X_t - \eta\nabla L(X_t) + \sqrt{2\eta\beta^{-1}}\xi_t$

(Euler-Maruyama近似)

$\xi_t \sim \mathrm{N}(0, I)$

SGLD: $X_{t+1} = X_t - \eta\underbrace{\frac{1}{|I_B|}\sum_{i\in I_B}\nabla\ell_i(X_t)}_{\text{確率的勾配}} + \sqrt{2\eta\beta^{-1}}\xi_t$


Gaussian noise
Gradient descent

収束定理 (有限次元)

- f_i : 有界，Lipschitz連続，滑らかな勾配

$$\|\ell_i\|_\infty \leq A,\ \|\nabla \ell_i\|_\infty \leq B,\ \|\nabla \ell_i(x) - \nabla \ell_i(y)\| \leq M\|x-y\|$$

- **散逸条件:**

$$\langle \nabla L, w \rangle \geq m\|w\|^2 - b \quad (\forall w \in \mathbb{R}^d)$$

(+ その他細かい条件)

Thm [Raginsky, Rakhlin and Telgarsky, COLT2017]

$$\mathrm{E}[L(X_k)] - L(X^*) \leq \tilde{\mathrm{O}}\Bigg((\beta+d)k\eta^{5/4} + \frac{\beta+d}{\sqrt{\lambda^*}}\exp\left(-\tilde{\Omega}\left(\frac{\lambda^* k\eta}{\beta(d+\beta)}\right)\right) + \frac{d\log(\beta+1)}{\beta}\Bigg)$$

次元dに指数的に依存!

- λ_* はスペクトルギャップと言われる量．
  → 次元dや逆温度パラメータβに対して指数関数的に依存．
- 逆温度パラメータが十分大きくて，更新を十分な回数回せば最適解付近に近づける．
- Xu et al. (NeurIPS2018) は収束レートを改善しているが，証明にいくつかの間違いあり．

散逸条件

Infinite dim non-convex optimization by Langevin dynamics

Joint work with Boris Muzellec (CREST, ENSAE,), Kanji Sato (U-Tokyo), Mathurin Massias (INRIA).

[Boris Muzellec, Kanji Sato, Mathurin Massias, Taiji Suzuki: Dimension-free convergence rates for gradient Langevin dynamics in RKHS. arXiv:2003.00306.]

無限次元への拡張

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

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

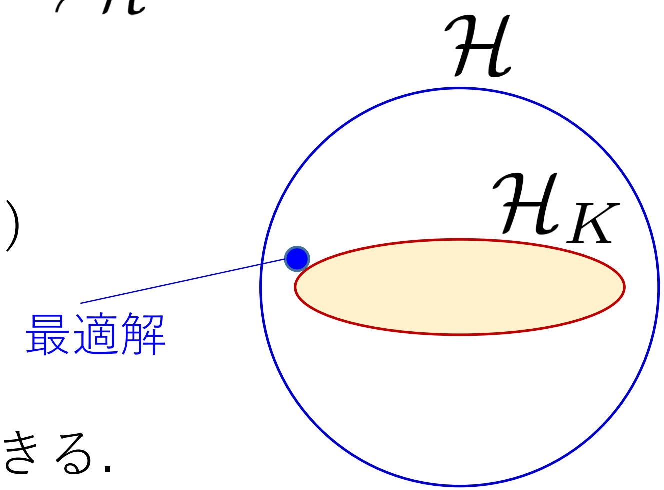
$\mathcal{H}$: Hilbert space

正則化

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

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

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



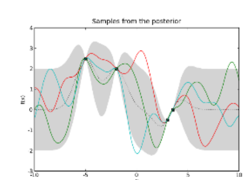
Ex.

- $\mathcal{H}$: $L^2(\rho)$
- $\mathcal{H}_K$: 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.

E.g., Bayesian optimization on infinite dimensional space

[Zimmermann and Toussaint. Bayesian functional optimization. AAAI, 2018]
[Vellanki, Rana, Gupta, de Celis Leal, Sutti, Height, and Venkatesh: Bayesian functional optimisation with shape prior. AAAI, 2019]



動機: NNの学習

- 2層ニューラルネットワーク

Idea: 分布の学習 → **<u>輸送写像の学習</u>**

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

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

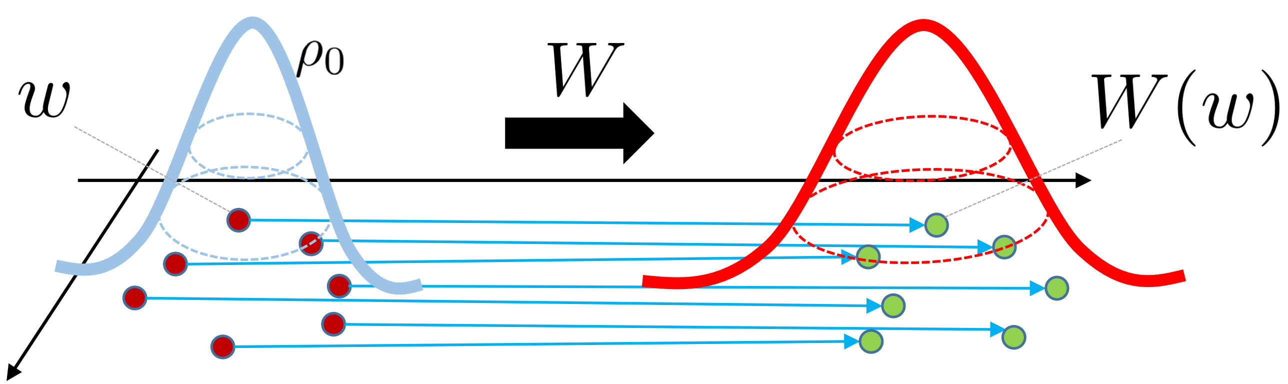
$$= \int_{\mathbb{R}^d} a(w)\sigma(w^\top x)\mathrm{d}W\sharp\rho_0(w)$$

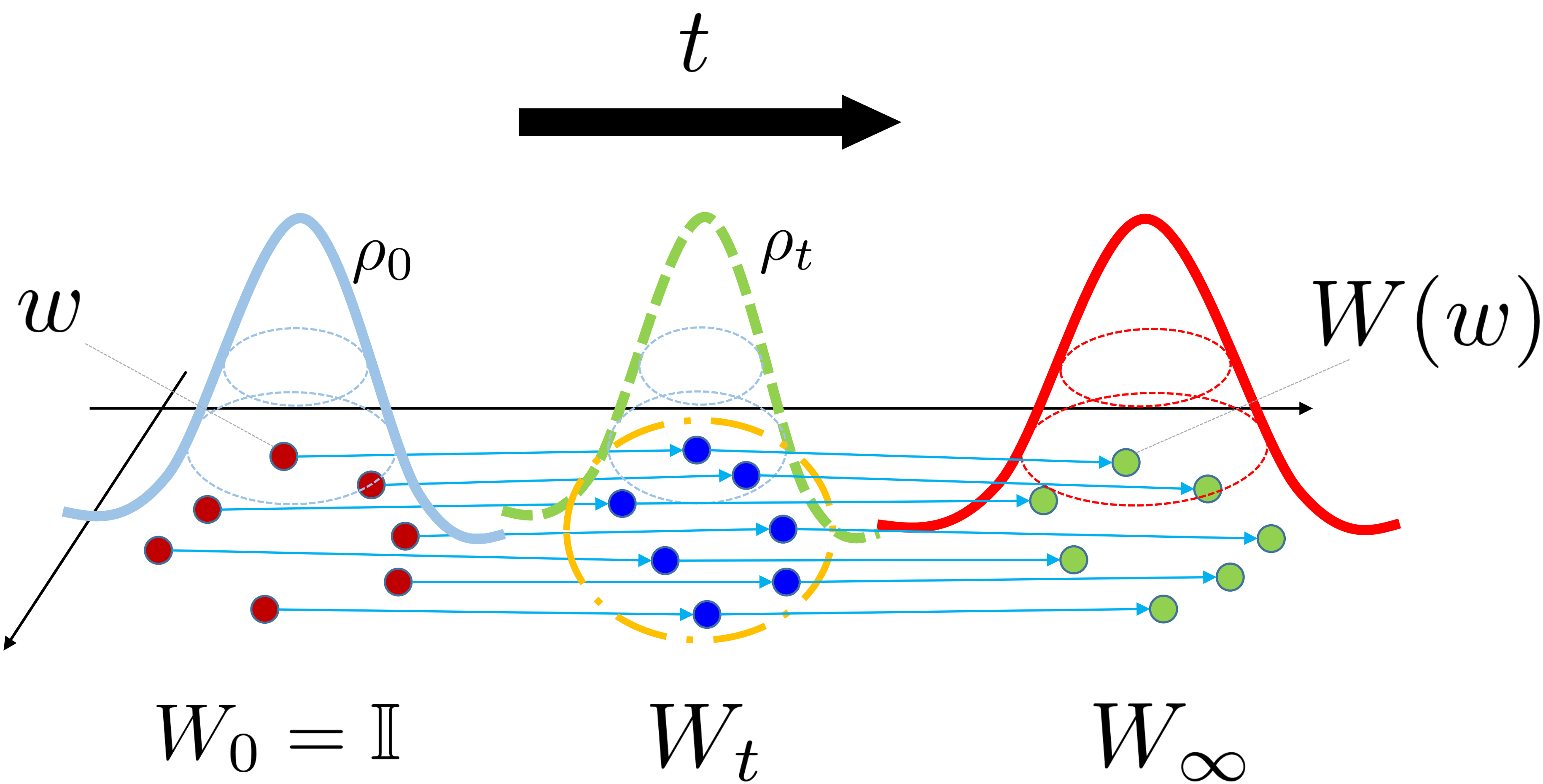
"Lift"

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

以前の表記

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



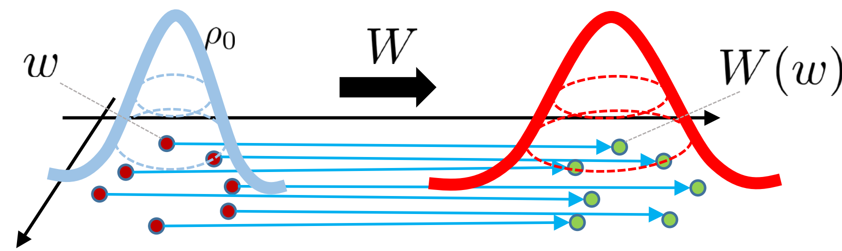


- ρ_0が有限サポートの離散測度なら<u>**有限横幅のニューラルネットワーク**</u>が扱える．しかも，横幅はサンプルサイズによらない．
  - ➢ **NTKや平均場解析と大きく異なる．**
  - ➢ 有限横幅/無限横幅を**統一的に扱える．**

利点

$$\min_{W \in \mathcal{H}} \; L(f_W)$$

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



- 初期分布ρ_0が離散有限なら**有限横幅のニューラルネットワーク**になっている．
  - ➢ **平均場やNTKと大きく異なる．**
- 初期分布ρ_0が連続なら無限横幅も扱える．

→ **有限横幅/無限横幅を統一的に扱える．**

最適化の方策:

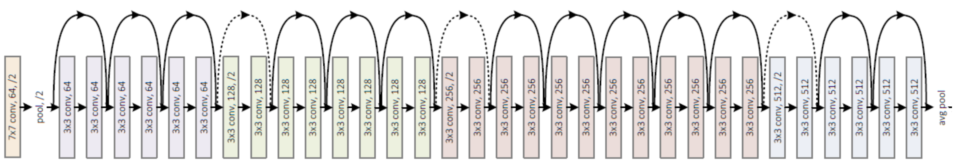
$$\mathrm{d}W_t = -(\underbrace{\overset{\text{平滑化}}{AW_t}} + \nabla_W L(f_{W_t}))\mathrm{d}t + \underbrace{\sqrt{\beta^{-1}}\mathrm{d}\xi_t}$$

1. 雑音は“関数”Wに足される．
2. 正則化により滑らかになる．

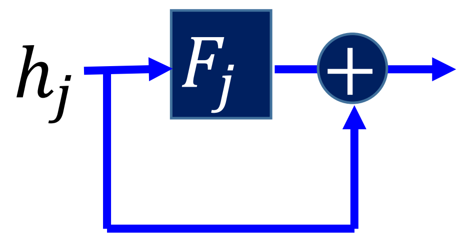
- 平滑化と合わせて粒子間の相関を表現できる．

ResNet

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



Residual block



$$h_{j+1} = h_j + F_j(h_j)$$

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

$$f_W = u^\top \Big(\mathbb{I} + \underbrace{\int_{\mathbb{R}^d} a(w,T)\sigma(W(w,T)^\top \cdot)\mathrm{d}\rho_0(w)}_{\text{Residual block}}\Big) \circ \cdots \circ \Big(\mathbb{I} + \int_{\mathbb{R}^d} a(w,1)\sigma(W(w,1)^\top x)\mathrm{d}\rho_0(w)\Big)$$

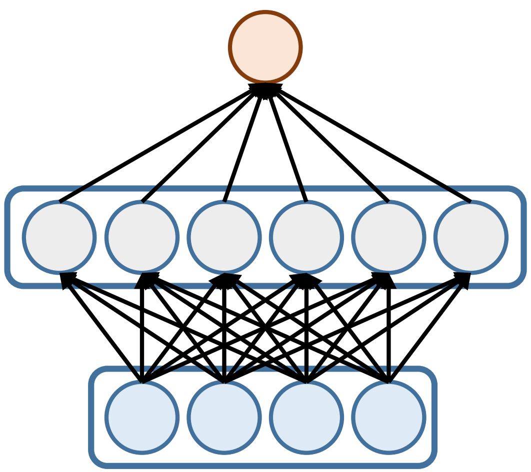
Training deep net can also be formulated as transportation map optimization.

2層NNの学習: 直接表現

$$L(W) = \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \ell_i(f_W(x_i)) + \frac{\lambda_0}{2} \|W\|_{\mathrm{F}}^2$$

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

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



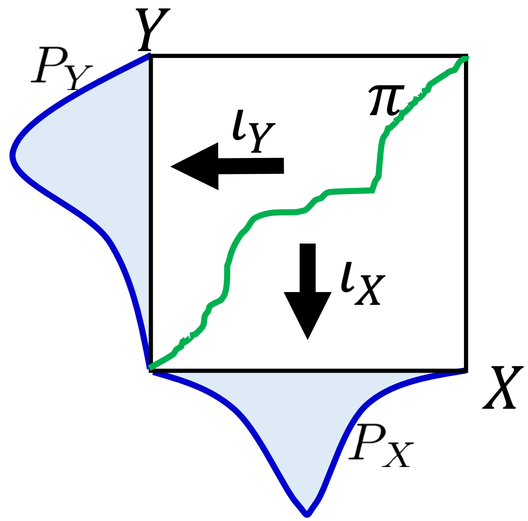
NTKと違い，a_jはデータサイズにも横幅にも依存させずスケールを固定できる．
(TNKは$a_j = 1/\sqrt{M}$とする)

Transportation map learning

- Wasserstein metric

$$W_2(X,Y) := \sqrt{\inf_{\pi:\text{coupling}} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|X-Y\|^2 \mathrm{d}\pi(X,Y)}$$

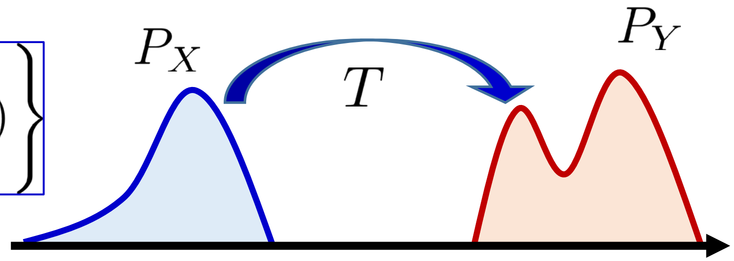
Coupling of X and Y:
distribution on $\mathbb{R}^d \times \mathbb{R}^d$ such that the maginal distribution satisfies $\iota_X \sharp \pi = P_X$, $\iota_Y \sharp \pi = P_Y$



Monge version:

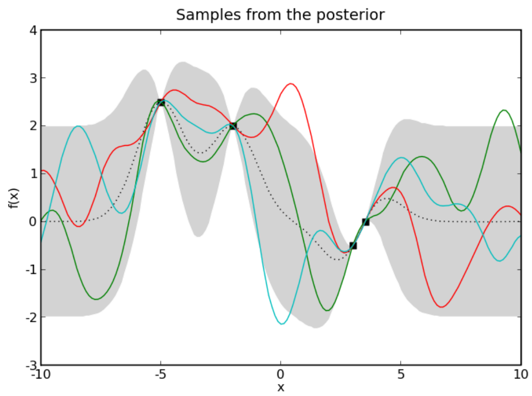
$$W_2(X,Y) = \sqrt{\inf_{T:T\sharp P_X = P_Y} \int_{\mathbb{R}^d} \|X - T(X)\|^2 \mathrm{d}P_X(X)}$$

$$\boxed{\hat{T} = \mathop{\arg\min}_{T\in\mathcal{H}_K} \left\{ \frac{1}{n}\sum_{i=1}^n (x_i - T(x_i))^2 + \lambda D(T\sharp\hat{P}_X || \hat{P}_Y) \right\}}$$



Other examples

- Bayesian optimization on infinite dimensional space



[Zimmermann and Toussaint. Bayesian functional optimization. AAAI, 2018]
[Vellanki, Rana, Gupta, de Celis Leal, Sutti, Height, and Venkatesh: Bayesian functional optimisation with shape prior. AAAI, 2019]

- Tensor decomposition on RKHS

[M. Signoretto, L. De Lathauwer, and J. A. Suykens. Learning tensors in reproducing kernel Hilbert spaces with multilinear spectral penalties. arXiv preprint arXiv:1310.4977, 2013]
[T. Suzuki, H. Kanagawa, H. Kobayashi, N. Shimizu, and Y. Tagami. Minimax optimal alternating minimization for kernel nonparametric tensor learning. In Advances in Neural Information Processing Systems 29, pages 3783–3791. 2016]

- Robust classification

[H. Masnadi-Shirazi and N. Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In Advances in Neural Information Processing Systems 21, pages 1049–1056. 2009.]

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.

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

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

ノルム: For $x = \sum_{j=1}^{\infty} x_j f_j \in \mathcal{H}$, we let $\|x\|_{\mathcal{H}_K}^2 = \sum_{j=1}^{\infty} \mu_j^{-1} x_j^2$ where $\mu_j \sim j^{-2}$.

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

時間離散化:

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

(準陰的Eulerスキーム) $A = \mathrm{diag}(\mu_1^{-1}, \mu_2^{-1}, \ldots)$

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

Galerkin近似 & 確率的勾配

- スペクトラルGalerkin近似:

無限次元ダイナミクスは実際は計算できない.
→ N-次元空間で近似.

$$H_N := \mathrm{span}\{f_0, \dots, f_N\} \quad P_N\text{: Projection to } \mathcal{H}_N$$

$$\boxed{X_{n+1}^N = S_\eta \left(X_n^N - \eta P_N \nabla L(X_n^N) + \sqrt{2\frac{\eta}{\beta}} P_N \xi_n \right)}$$

- 確率的勾配の利用:

$$\boxed{X_{n+1}^N = S_\eta \left(X_n^N - \eta P_N \frac{1}{|I_n|} \sum_{i \in I_n} \nabla \ell_i(X_n^N) + \sqrt{2\frac{\eta}{\beta}} P^N \xi_n \right)}$$

定常分布

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

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

$$\mu_* = N(0, C)$$ (Hilbert空間上のガウス過程)

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

$$\pi_\infty(x) \quad \propto \quad \exp\left(-\beta L(x) - \frac{1}{2}x^\top C^{-1}x\right)$$ 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.
→ 過学習を防ぎ汎化する

[Suzuki, arXiv:2007.05824]

無限次元の設定

ヒルベルト空間

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

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

RKHS構造

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

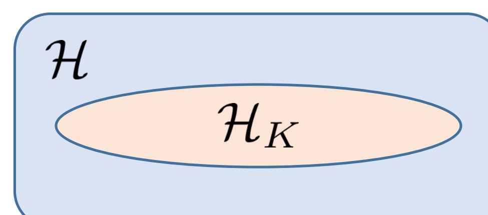
$$\langle x, y \rangle_{\mathcal{H}_K} = \sum_{k=0}^{\infty} \alpha_k \beta_k / \mu_k \quad \text{for } x = \sum_k \alpha_k f_k,\ y = \sum_k \beta_k f_k.$$

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．）

$$\min_{x \in \mathcal{H}} L(x) = \min_{x \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell_i(x) + (\tfrac{\lambda_0}{2} \|x\|^2)$$



Related work

- **Finite dimensional Langevin dynamics:**

➢ Convergence in low (convex case): Dalalyan and Tsybakov, 2012; Dalalyan, 2016; Durmus and Moulines, 2015, ..

➢ Non-convex Optimization: Raginsky et al., 2017; Xu et al., 2018; Erdogdu, Mackey and Shamir, 2018, ...

- **Infinite dimensional Langevin dynamics:**

➢ Continuous time:
  - Existence & Uniqueness of invariant measure: Da Prato and Zabczyk,1992; Maslowski, 1989; Sowers, 1992.
  - Geometric ergodicity: Jacquot and Royer, 1995; Shardlow, 1999; Hairer, 2002, Its explicit rate: Goldys and Maslowski, 2006.

➢ Discrete time:
  - Weak approximation rate of discretized scheme: Hausenblas, 2003; Debussche, 2011; Bréhier, 2014; Bréhier and Kopec 2016.

Other topics (MCMC in Hilbert space):
  - preconditioned Crank–Nicolson (pCN): Hairer et al., 2014; Eberle, 2014; Vollmer, 2015; Rudolf and Sprungk, 2018.
  - Metropolis-Adjusted Langevin Algorithm (MALA): Durmus and Moulines, 2015; Beskos et al., 2017.

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

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

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

Assumption (2)

- Smoothness:

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

- Strong smoothness condition:

（これが無い場合はレートが遅くなる）

For $\alpha \in (1/4, 1)$,

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

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

(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

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

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

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：

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

(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).

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

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.

誤差の解析 (2)

π_∞: 定常分布

$$x^* := \arg\min_{x\in\mathcal{H}} L(x) \qquad \tilde{x} := \arg\min_{x\in\mathcal{H}} \left\{ L(x) + \frac{\lambda}{2}\|x\|_{\mathcal{H}_K}^2 \right\}$$

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

上記の条件のもと，次が成り立つ：

$$L(X_n) - L(x^*) \lesssim \exp\left(-\Lambda_\eta^* n\eta\right) + \frac{c_\beta}{\Lambda_0^*}\eta^{1/2-\kappa}$$

(geometric ergodicity + time discretization)

$$+ \frac{1}{\beta}\left(\sqrt{\frac{1}{\lambda}} + 1\right) + \lambda\left(\frac{\|\tilde{x}\|_{\mathcal{H}_K}}{\sqrt{\beta}} + \|\tilde{x}\|_{\mathcal{H}_K}^2\right)$$

$$+ L(\tilde{x}) - L(x^*)$$

(bias of invariant measure)

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).

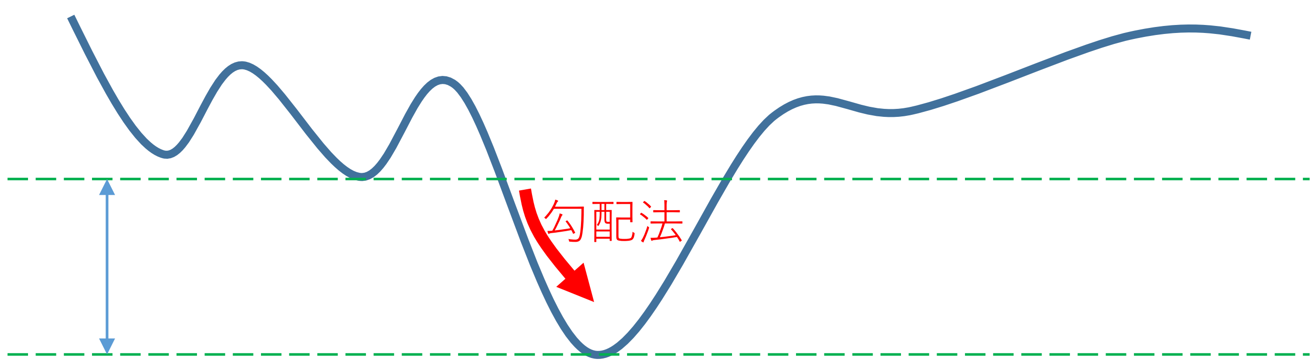
Λ_η^*: スペクトルギャップ，βに対して指数的依存がある.

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.

- 深層学習の最適化への応用と汎化誤差解析：Suzuki, arXiv:2007.05824.

ノイズのコントロール

- 大域的最適解を得るためには$\beta \to \infty$が必要.
- スペクトルギャップはβに指数的に依存.
- 大域的最適解まわりで局所的に凸になっていて, 離れた場所より目的関数値が真に小さければ途中で勾配法に切り替えても良い.
- 例えば2層NNでは訓練誤差の形状が局所的に強凸になることがある [Li and Yuan, 2017][Chizat, 2019]
(各ニューロンが適度にばらけている場合はそうなる)



誤差の分解

弱収束を示す:

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

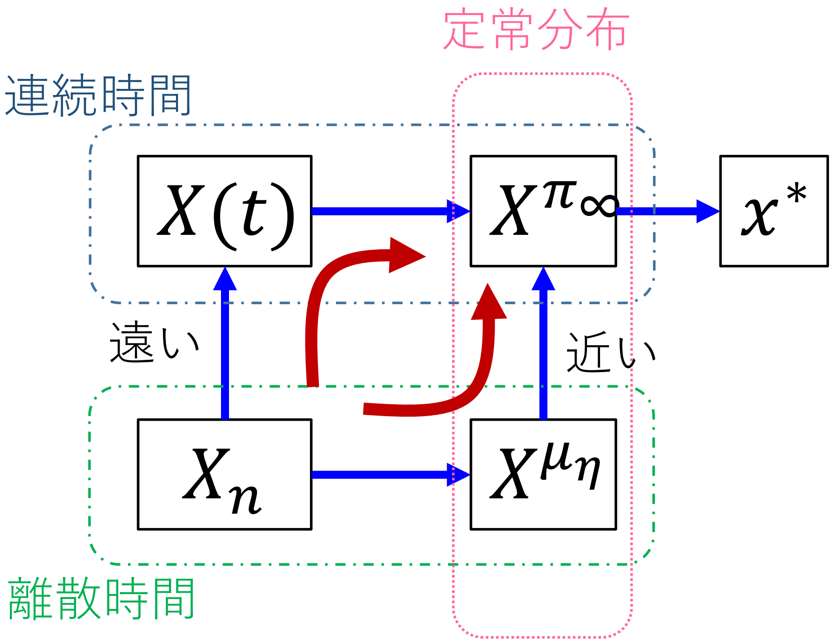
ただし，ϕは滑らかな関数.

- 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

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

補題 (離散時間ダイナミクスのGeometric ergodicity)

ある定常分布μ_ηがただ一つ存在して (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 $\quad \delta = \exp(-O(\beta))$

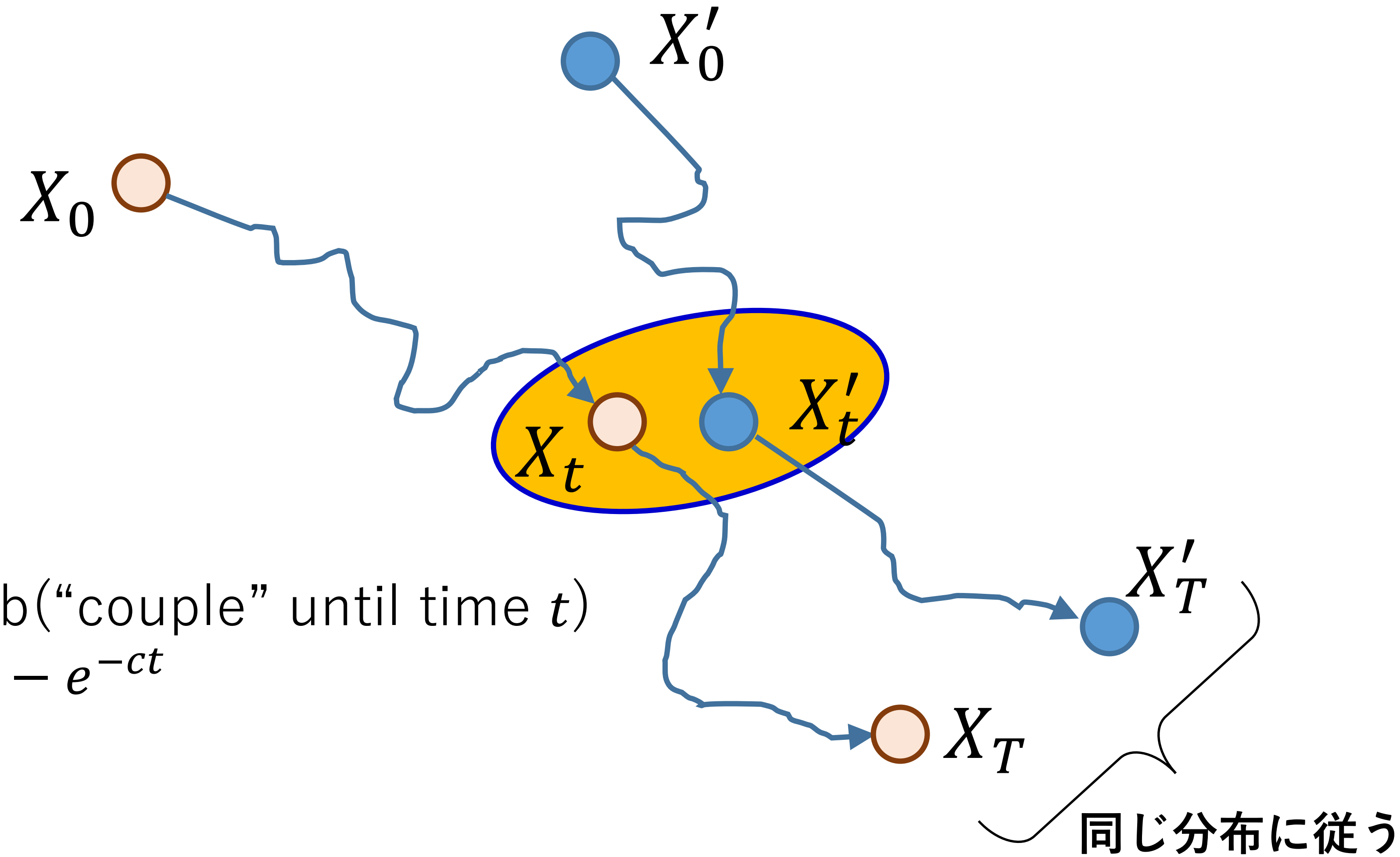
X^{μ_η}: r.v. obeying μ_η $\qquad X_0 = x_0$ (constant)

- 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)
$\geq 1 - e^{-ct}$

Second term bound

参考

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

X^{μ_η}: the invariant measure of discrete time dynamics
X^π: the invariant measure of continuous time dynamics
(existence and uniqueness are well known)

Lemma (Discrepancy between invariant measures)

For any $0<\kappa<1/2$, there exists a constant C such that

$$|\mathbb{E}[\phi(X^{\mu_\eta})-\phi(X^\pi)]| \le C\|\phi\|_{0,2}\frac{c_\beta}{\Lambda_0^*}\eta^{1/2-\kappa}.$$

- $\|\phi\|_{0,2} = \max\{\|\phi\|_\infty, \|D\phi\|_\infty, \|D^2\phi\|_\infty\}$
- $c_\beta = \sqrt{\beta}$ for bounded gradient condition, and $\beta = 1$ otherwise

- Malliavin calculus
- As the step-size goes to 0, the discrete time dynamics approaches the continuous one.
- β affects the bound through the spectral gap Λ_0^*.

既存研究との違い

参考

- 離散時間ダイナミクスの幾何的エルゴード性を示しているため，より速い弱収束.

Brehier 2014:

※ Brehierはβを考えていない.

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

Ours:

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

これは以下の追加条件による:

強平滑条件

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

(強い条件だが，機械学習応用では自然ではある)

第二項のバウンド

参考

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

X^{μ_η}: 離散時間ダイナミクスの定常分布
X^{π}: 連続時間ダイナミクスの定常分布 (存在と一意性は保証されている)

補題 (連続・離散時間ダイナミクスの定常分布の違い)

任意の$0<\kappa<1/2$に対し，ある定数Cが存在して，

$$|\mathbb{E}[\phi(X^{\mu_\eta})-\phi(X^{\pi})]| \le C\|\phi\|_{0,2}\frac{c_\beta}{\Lambda_0^*}\eta^{1/2-\kappa}.$$

- $\|\phi\|_{0,2}=\max\{\|\phi\|_\infty, \|D\phi\|_\infty, \|D^2\phi\|_\infty\}$
- $c_\beta=\sqrt{\beta}$ for bounded gradient condition, and $\beta=1$ otherwise

- **Malliavin解析**
- ステップサイズηを0に近づけると，離散時間ダイナミクスが連続時間ダイナミクスに近づく.
- βはΛ_0^*に影響している.

第三項のバウンド

参考

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

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

$$\frac{\mathrm{d}\pi}{\mathrm{d}\mu_*}(x) \quad \propto \quad \exp\left(-\beta L(x)\right)$$

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

Lemma (最適解とベイズ推定量の差分)

$$\int L\mathrm{d}\pi - L(\tilde{x}) \lesssim \frac{1}{\beta}\left(\sqrt{\frac{2M}{\lambda}}+1\right) + \lambda\left(\frac{\|\tilde{x}\|_{\mathcal{H}_K}}{\sqrt{\beta}} + \|\tilde{x}\|_{\mathcal{H}_K}^2\right).$$

- ノンパラメトリックベイズの技法と似た技法を使いながら示す.
- “Gaussian correlation inequality” を用いて$\tilde{x}$のまわりのπの大きさを評価する.

Galerkin approximation & SGLD

参考

- Assumption: Smoothness of L.

$$x^* := \arg\min_{x \in \mathcal{H}} L(x) \qquad \tilde{x} := \arg\min_{x \in \mathcal{H}} \left\{ L(x) + \frac{\lambda}{2} \|x\|_{\mathcal{H}_K}^2 \right\}$$

Thm

Under some smoothness assumption on L, we have

$$L(X_n) - L(x^*) \lesssim \exp(-\Lambda^* n\eta) + \frac{c_\beta}{\Lambda^*}\eta^{1/2-\kappa}$$

(geometric ergodicity + time discretization)

$$+ \frac{1}{\beta}\left(\sqrt{\frac{1}{\lambda}} + 1\right) + \lambda\left(\frac{\|\tilde{x}\|_{\mathcal{H}_K}}{\sqrt{\beta}} + \|\tilde{x}\|_{\mathcal{H}_K}^2\right)$$

$$+ L(\tilde{x}) - L(x^*)$$

(bias of invariant measure)

$$+ \frac{\mu_{N+1}^{1/2-\kappa}}{\Lambda_0^*} + \sqrt{\frac{n\beta\eta(n_{\mathrm{tr}} - n_{\mathrm{b}})}{n_{\mathrm{b}}(n_{\mathrm{tr}} - 1)}}$$

Galerkin approx.

Mini-batch size

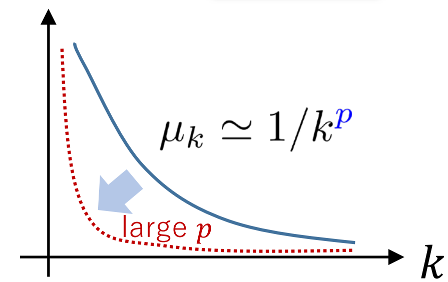
with high probability, where $\kappa > 0$ is an arbitrary small constant.

(c.f., Brehier 2014; Goldys&Maslowski, 2006)

有限次元バージョンとの関係

参考

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



- $\mu_k \simeq 1/k^2$ （我々の状況）

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

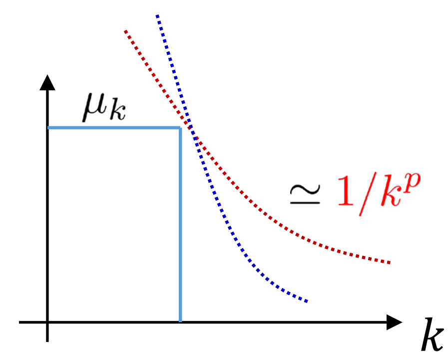
- $\mu_k \simeq 1/k^p$ （予想） see [Andersson,Kruse&Larsson, 2016] for finite time horizon.

pが大きくなるほど関数クラスは“単純”になる.

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

有限次元の解析は$p \to \infty$に対応 (定数を無視すれば):

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

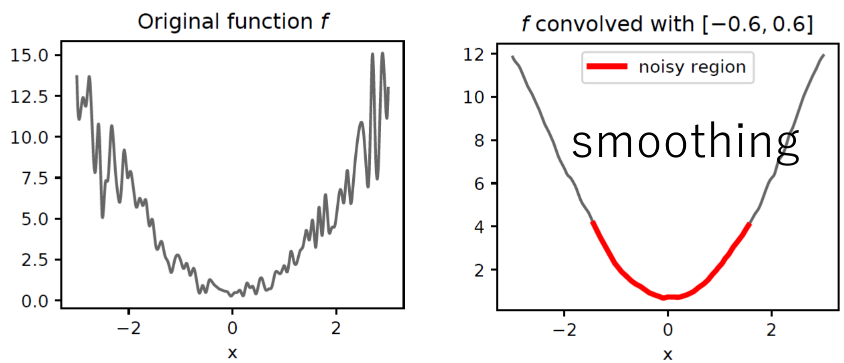


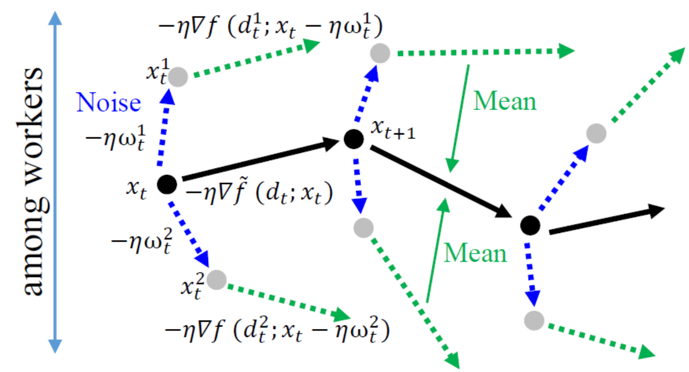
[Xu et al. (2018)]

Application: gradient noise convolution

参考

- Non-convex objective
  - ➢Smoothing the objective by noise convlution
  - ➢Noise is determined by SGD.
  - ➢**Gives good generalization error**





(Figure is from Kleinberg, Li, and Yuan, ICML2018)

Our proposal

Performance comparison

Datasets	Batch size	Baseline	RNC	GNC	GNCtoRNC
ImageNet	32,768	75.89±0.09	75.86±0.13	76.03±0.18	**76.05**±0.07
	131,072	65.57±0.45	65.14±0.86	**68.39**±0.40	68.13±0.48
CIFAR-10	4,096	93.48±0.26	93.70±0.26	93.82±0.64	**94.00**±0.60
	8,192	54.51±7.04	63.51±20.45	**91.00**±1.24	90.88±1.53
CIFAR-100	4,096	72.87±0.37	73.08±0.27	72.89±0.38	**73.79**±0.40
	8,192	69.21±2.11	70.17±1.36	71.35±0.27	**71.93**±0.21

[Haruki, Suzuki, Hamakawa, Toda, Sakai, Ozawa, Kimura: Gradient Noise Convolution (GNC): Smoothing Loss Function for Distributed Large-Batch SGD. arXiv:1906.10822]

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. arXiv:2007.05824]

問題設定

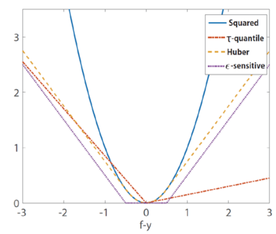
$$L(W) := \mathbb{E}[\ell(f_W, Z)]$$

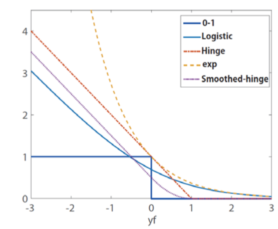
$$\widehat{L}(W) := \frac{1}{n}\sum_{i=1}^{n} \ell(f_W, z_i)$$

Wについて非凸

E.g., $\ell(f, z) = \ell(f(x), y)$

- 二乗損失
- ロジスティック損失





汎化誤差:
(汎化ギャップ)

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

残余誤差:

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

Eg. 二層ニューラルネットワークモデル:

$$f_W(x) := \int_{\mathbb{R}^d} a(w)\sigma(R[[W(w)/R]]^\top x)\mathrm{d}\rho_0(w)$$

$$[[W]] = \mathrm{sigmoid}(W)$$

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:

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

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

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

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

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

- PAC-Bayesによる汎化誤差バウンド

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

$$\Xi_k \simeq \exp(-\Lambda_\eta^*(\eta k - 1)) + \frac{\sqrt{\beta}}{\Lambda_0^*}\eta^{1/2-\kappa}.$$

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

PAC-Bayesian stability bound [Rivasplata, Kuzborskij, Szepesvári, and Shawe-Taylor, 2019]

Excess risk bound

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

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

追加の仮定:

- $\exists \gamma > 1/4$ ：モデルの複雑さを制御
  $\tilde{\ell}(W, z) = \ell(f_{T_K^{\gamma/2} W}, z)$ ➡ $L(W) = \mathbb{E}[\tilde{\ell}(W, Z)]$
- $\exists W^* \in \mathcal{H}$ s.t. $\inf_f L(f) = L(f_{W^*})$ $(= L(f^*))$
- **Bernstein条件** [Erven et al., 2015]:
  $\mathbb{E}[(\ell(f, Z) - \ell(f^*, Z))^2] \leq B(L(f) - L(f^*))^s$
  - 二乗損失 ($s = 1$)
  - ロジスティック損失 with 有界なf, f^* ($s = 1$)
- $\mathbb{E}\left[\exp\left(-\frac{\beta}{n}(\ell(f, Z) + \ell(f^*, Z))\right)\right] \leq 1$
  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

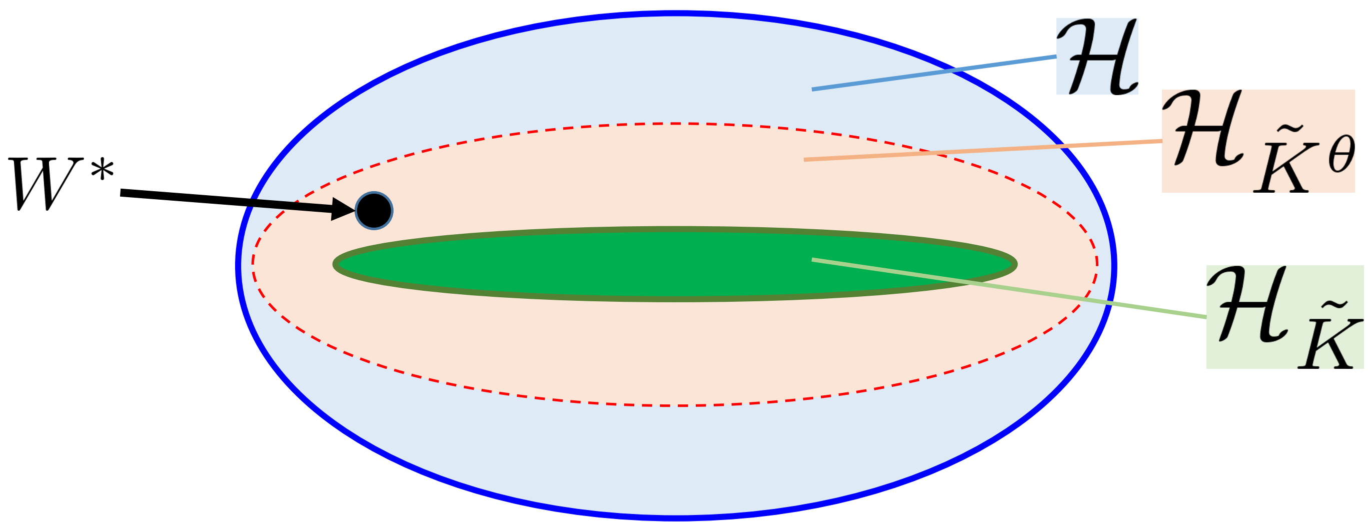
Let $\mathcal{H}_{\tilde{K}} = T_K^{(\gamma+1)/2}\mathcal{H}$ and $\mathcal{H}_{\tilde{K}^\theta} = T_K^{\theta(\gamma+1)/2}\mathcal{H}$.

Thm (Fast rate: excess risk bound)

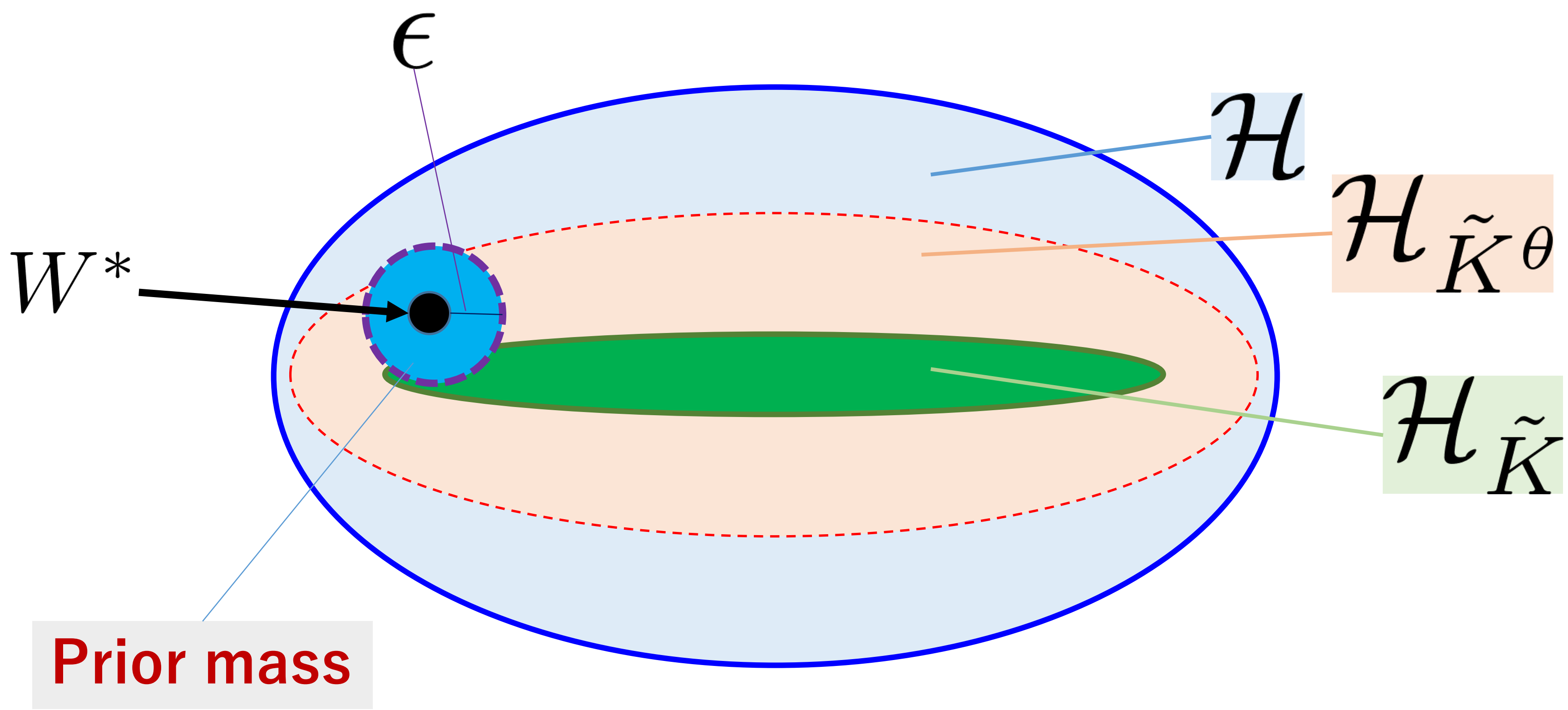
Suppose that $W^* \in \mathcal{H}_{\tilde{K}^\theta}$ for $0 < \theta < 1 - \frac{1}{2(\gamma+1)}$.

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

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



Proof idea



Concentration function:

$$\phi_{\beta,\lambda}(\epsilon) := \inf_{W \in \mathcal{H}_K : \|W - W^*\|_{\mathcal{H}} \leq \epsilon} \lambda\beta \|W\|^2_{\mathcal{H}_{\tilde{K}}} - \log \tilde{\mu}_\beta(\{W \in \mathcal{H} : \|W\|_{\mathcal{H}} \leq \epsilon\})$$

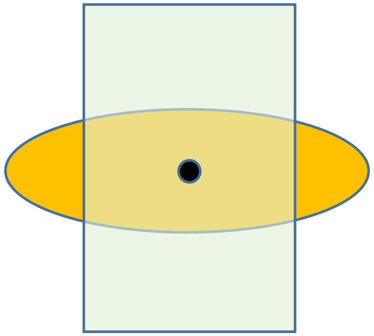
$$\tilde{\mu}_\beta\text{: dist. of } T_K^{\gamma/2} h \text{ where } h \sim \mu_\beta$$

Gaussian correlation inequality

μ : Gaussian measure in $\mathbb{R}^d$ with mean 0.

$E, F \subset \mathbb{R}^d$: convex, symmetric about origin.

$$\mu(E \cap F) \geq \mu(E) \cdot \mu(F)$$



This had been a conjecture for a long time (at least from 1972).
Royen resolved this problem in 2014 using relatively elementary tools.

- T. Royen. A simple proof of the gaussian correlation conjecture extended to multivariate gamma distributions. arXiv preprint arXiv:1408.1028, 2014.
- R. Latała and D. Matlak. Royen's proof of the Gaussian correlation inequality. pages 265–275. Springer International Publishing, 2017.

Fast rate: 回帰

$\ell(f, z) = (f(x) - y)^2$: 二乗損損失

- $\mathcal{H}$: $L_2(\rho_0)$
- $\mathcal{H}_{\tilde{K}}$: $W^{\alpha+d/2}(\mathbb{R}^d)$ (Sobolev space)
- $\theta = \frac{2\beta}{2\alpha+d}$ for $\beta < \alpha$

$\lambda^{-1} = \beta = n$ とすることで

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

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

判別問題における速い収束

Assumption

- **強低ノイズ条件:**

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

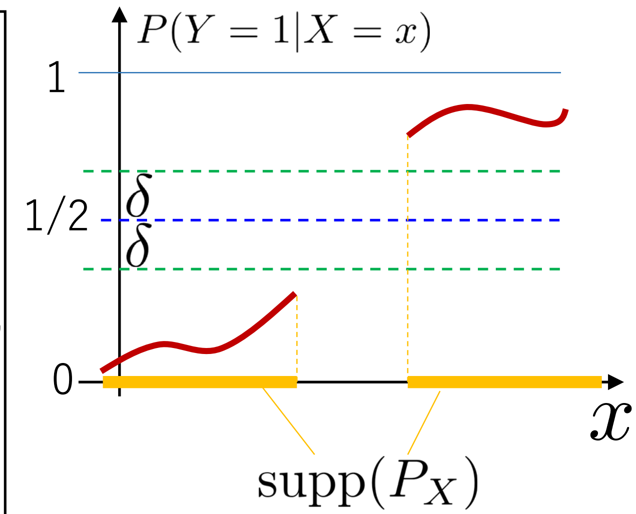
- $\mathrm{supp}(P_X) \subset [0,1]^d$ and P_X has density p such that

$$p(x) \geq c_0 \quad (\forall x \in \mathrm{supp}(P_X)).$$

- 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,

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

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

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)

Summary

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.

- 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

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**