機械学習における最適化理論と学習理論的側面 (第三部：深層学習の最適化)

鈴木大慈
東京大学大学院情報理工学系研究科数理情報学専攻
理研AIP

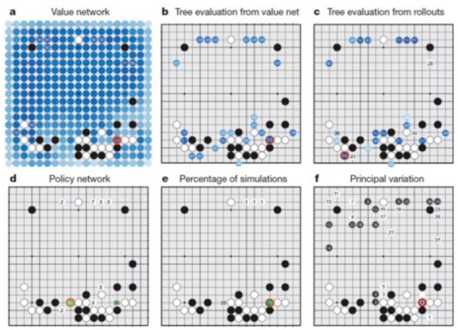
2020年8月6日

組合せ最適化セミナー2020 (COSS2020)

深層学習の広がり

様々なタスクで高い精度

AlphaGo/Zero

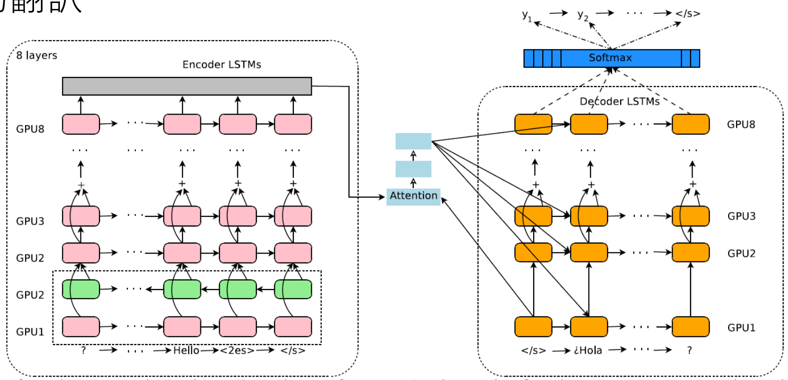


[Silver et al. (Google Deep Mind): Mastering the game of Go with deep neural networks and tree search, Nature, 529, 484—489, 2016]

画像認識

[He, Gkioxari, Dollár, Girshick: Mask R-CNN, ICCV2017]

自動翻訳



[Wu et al.: Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv:1609.08144]

画像の生成

[Glow: Generative Flow with Invertible 1x1 Convolutions. Kingma and Dhariwal, 2018]

画像の変換

[Zhu, Park, Isola, and Efros: Unpaired image-to-image translation using cycle-consistent adversarial networks. ICCV2017.]

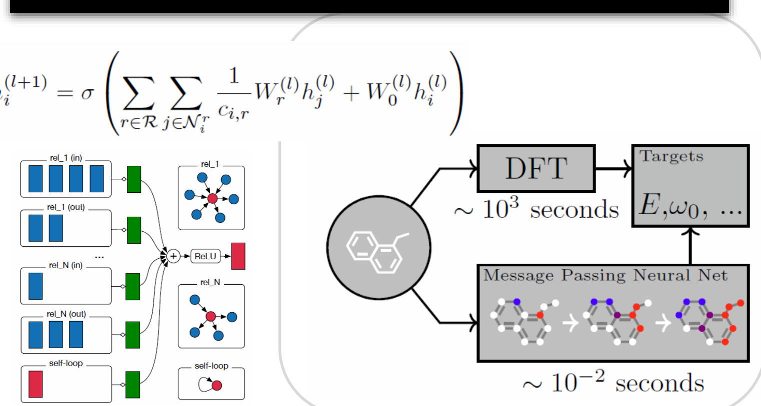
諸分野への波及

ロボット

[Google AI Blog, “Deep Learning for Robots: Learning from Large-Scale Interaction,” 2016/5/8]

量子化学計算，分子の物性予測

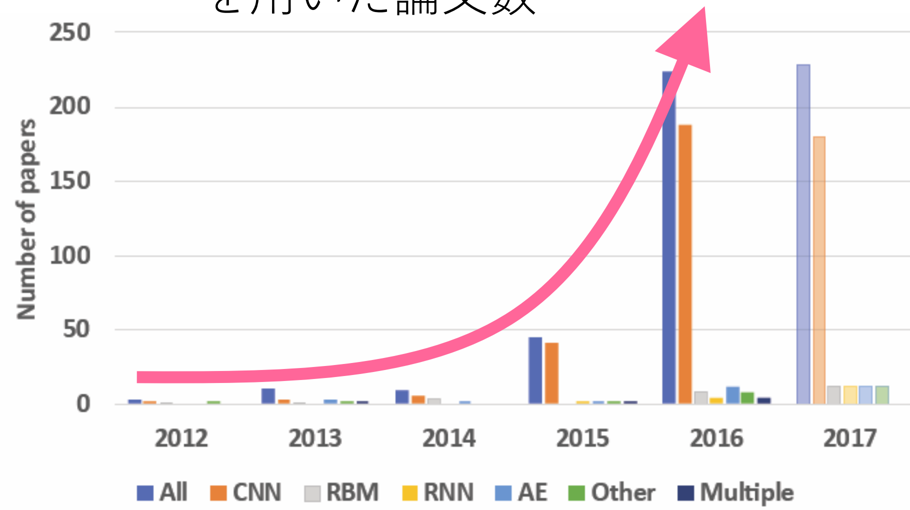
$$h_i^{(l+1)} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r}\frac{1}{c_{i,r}}W_r^{(l)}h_j^{(l)} + W_0^{(l)}h_i^{(l)}\right)$$



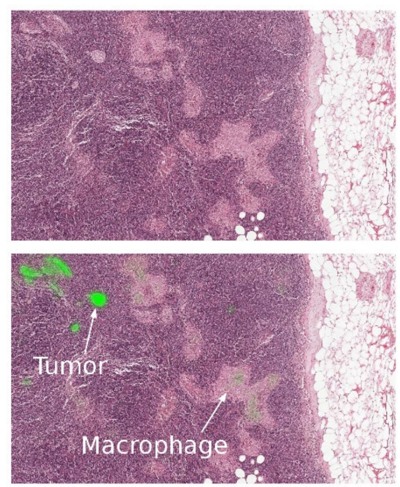
[Niepert, Ahmed&Kutzkov: Learning Convolutional Neural Networks for Graphs, 2016]
[Gilmer et al.: Neural Message Passing for Quantum Chemistry, 2017]
[Faber et al.:Machine learning prediction errors better than DFT accuracy, 2017.]

医療

医療分野における「深層学習」を用いた論文数



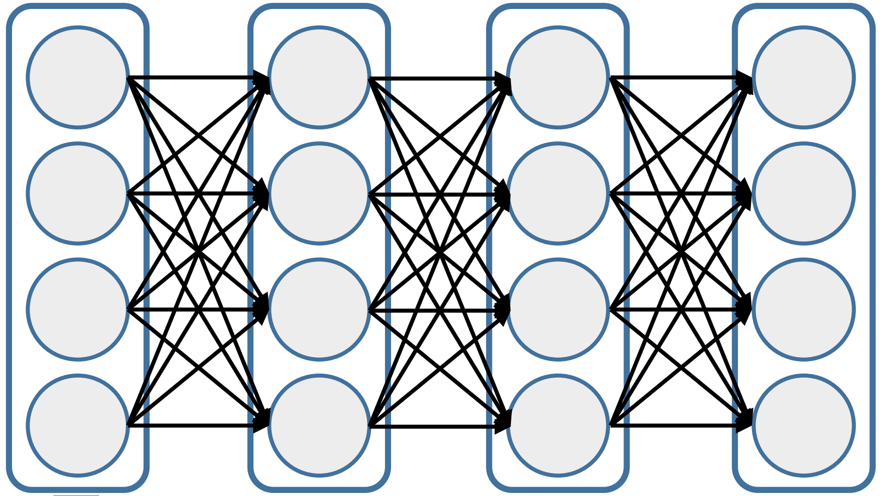
[Litjens, et al. (2017)]



- 人を超える精度（FROC73.3% -> 87.3%）
- 悪性腫瘍の場所も特定

[Detecting Cancer Metastases on Gigapixel Pathology Images: Liu et al., arXiv:1703.02442, 2017]

深層NNの構造



x W_1 h_1 W_2 h_2 W_3 h_3

基本的に「線形変換」と「非線形活性化関数」の繰り返し．

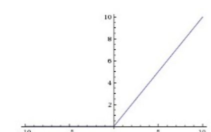
$x \rightarrow W_1 x \rightarrow h_1(W_1 x) \rightarrow W_2 h_1(W_1 x) \rightarrow h_2(W_2 h_1(W_1 x))$

入力　線形変換　非線形変換　線形変換　非線形変換

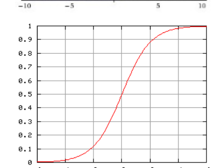
$h_1(u) = [h_{11}(u_1), h_{12}(u_2), \ldots, h_{1d}(u_d)]^T$

活性化関数は通常要素ごとにかかる．Poolingのように要素ごとでない非線形変換もある．

- ☆ReLU (Rectified Linear Unit)： $h(u) = \max\{u, 0\}$



- シグモイド関数： $h(u) = \dfrac{1}{1+e^{-u}}$

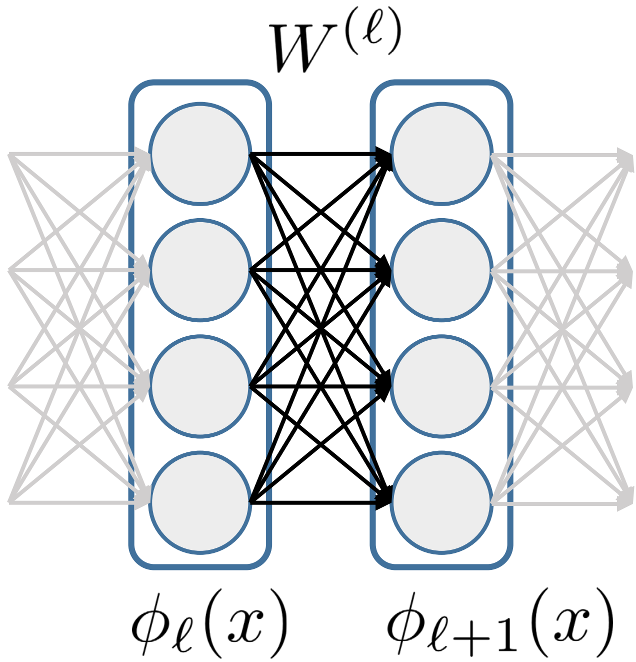


全結合モデル

- 第 ℓ 層

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

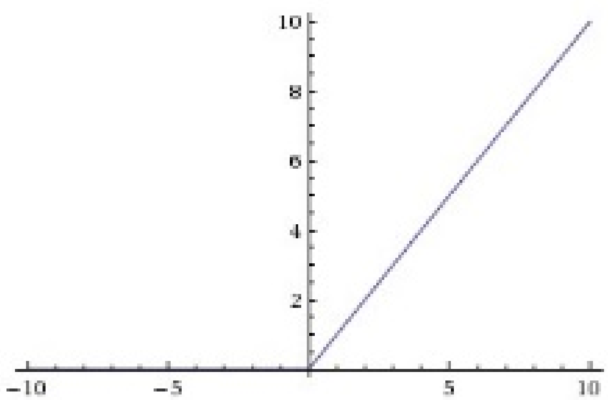
$$W^{(\ell)} \in \mathbb{R}^{m_{\ell+1}\times m_\ell} \quad b^{(\ell)} \in \mathbb{R}^{m_{\ell+1}}$$



活性化関数の例

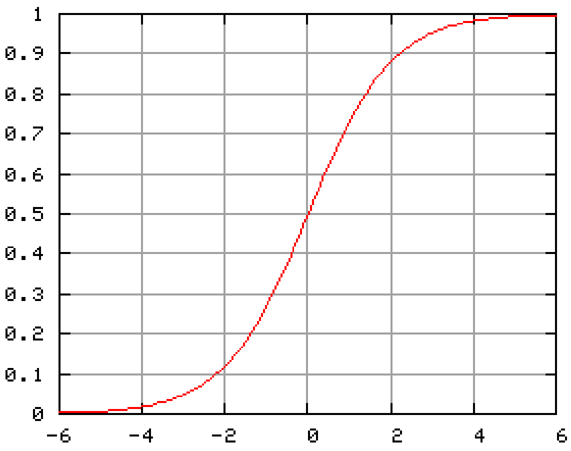
☆ReLU (Rectified Linear Unit)

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

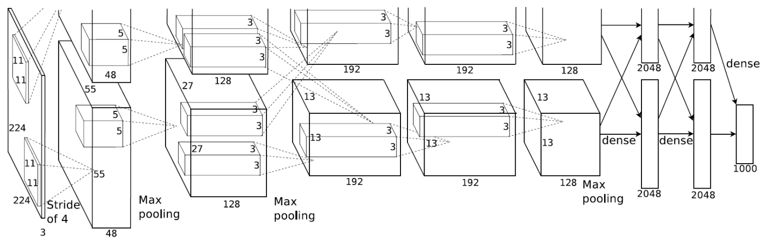


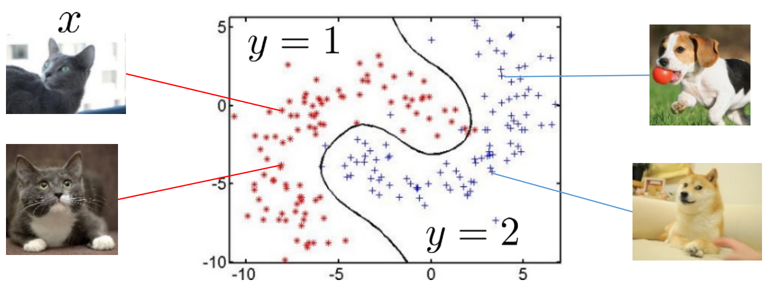
シグモイド関数

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



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

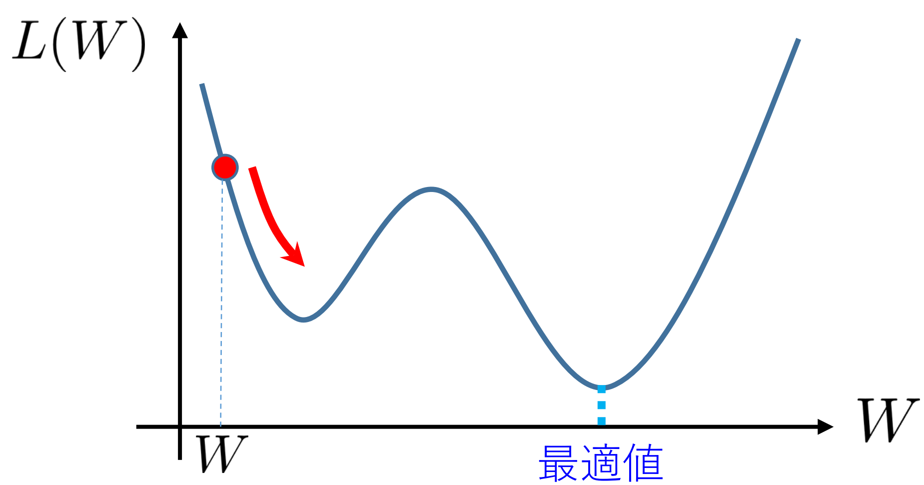
損失関数：データへの当てはまり度合い

損失関数最小化

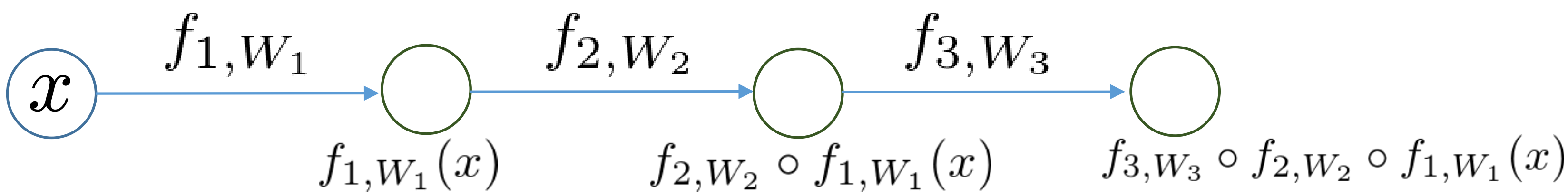
$$\min_{W} L(W)$$

(Wは数十億次元)

通常，**確率的勾配降下法**で最適化



誤差逆伝搬法



例： $f_{1,W_1}(x) = h(W_1 x)$

合成関数

$$
\begin{aligned}
f(x; W) &= f_{3,W_3}(f_{2,W_2}(f_{1,W_1}(x))) \\
&= f_{3,W_3} \circ f_{2,W_2} \circ f_{1,W_1}(x)
\end{aligned}
$$

合成関数の微分

$$
\frac{\partial f}{\partial W_1}(x) = \frac{\partial f_{3,W_3}}{\partial f_{2,W_2}} \frac{\partial f_{2,W_2}}{\partial f_{1,W_1}} \frac{\partial f_{1,W_1}}{\partial W_1}(x)
$$

深層学習で主に使われる確率的最適化法

- **SGD**
- **モーメンタム SGD** (Nesterov の加速法と類似，一番よく利用されている)

$$
\begin{aligned}
g_t &= \nabla L(w_t) \\
\Delta w_t &= \theta \Delta w_{t-1} - (1-\theta)\eta g_t \\
w_{t+1} &= w_t + \Delta w_t
\end{aligned}
$$

- **Nesterov の加速法** (凸関数における加速法と同様，パラメータの設定は異なる)

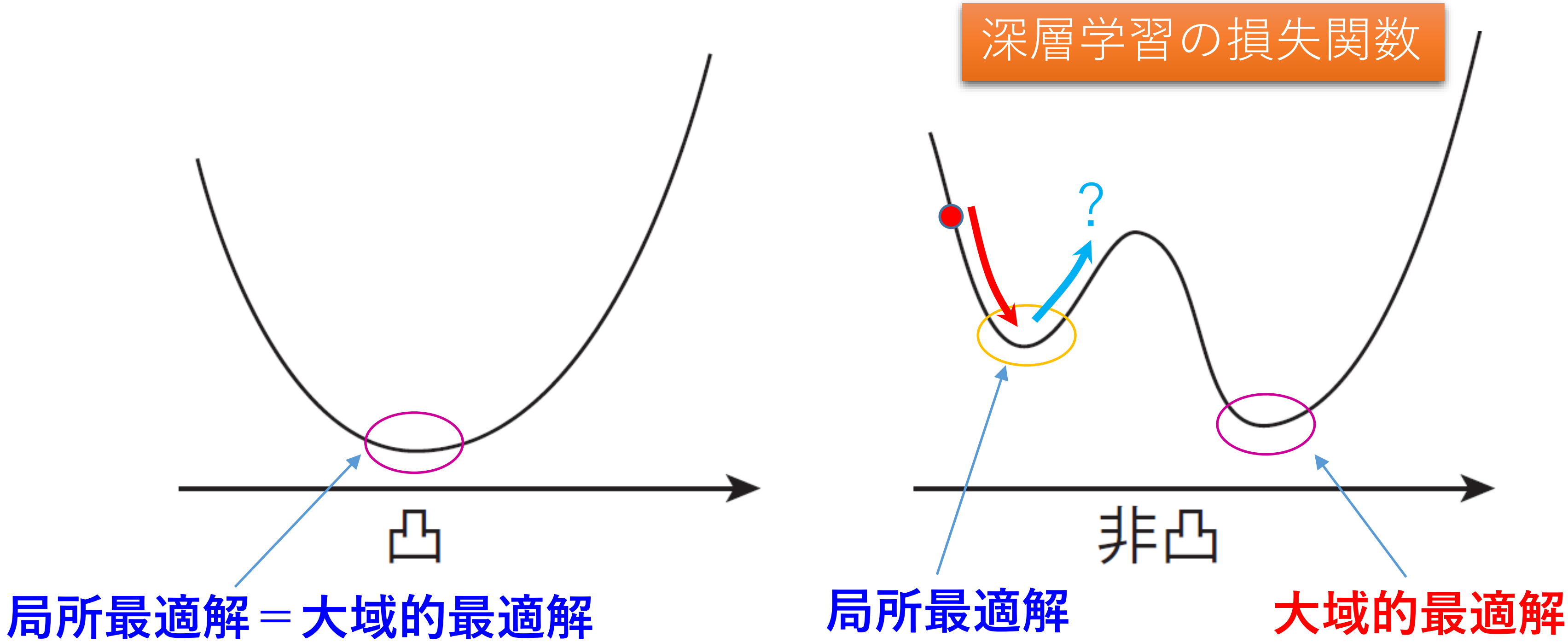
$$
g_t = \nabla L(w_t + \theta \Delta w_{t-1}), \Delta w_t = \theta \Delta w_{t-1} - (1-\theta)\eta g_t, w_{t+1} = w_t + \Delta w_t
$$

- **AdaGrad** (Duchi et al., 2011)
- **Adam** (Kingma and Ba, 2014)：AdaGrad と加速法を組み合わせたような方法．AdaGrad と違い，勾配のノルムを減衰させて次に伝える．モーメンタム SGD と並んでよく使われている．
- **RMSprop** (Hinton et al.)：AdaGrad において勾配のノルムを減衰させて和を取る方法．

問題点

目的関数が非凸関数

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



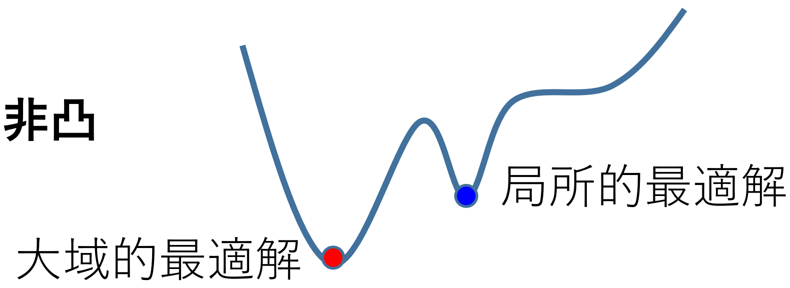
局所最適解や鞍点にはまる可能性あり

“**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.

大域的最適性

深層学習の目的関数は非凸



- 線形深層NNの局所的最適解は全て大域的最適解：
Kawaguchi, 2016; Lu&Kawaguchi, 2017.

※ただし対象は線形NNのみ.

→ 臨界点が大域的最適解であることの条件も出されている
(Yun, Sra&Jadbabaie, 2018)

- 低ランク行列補完の局所的最適解は全て大域的最適解：
Ge, Lee&Ma, 2016; Bhojanapalli, Neyshabur&Srebro, 2016.

$$\min_{U\in\mathbb{R}^{M\times k}} \sum_{(i,j)\in E} (Y_{i,j} - (UU^\top)_{i,j})^2$$

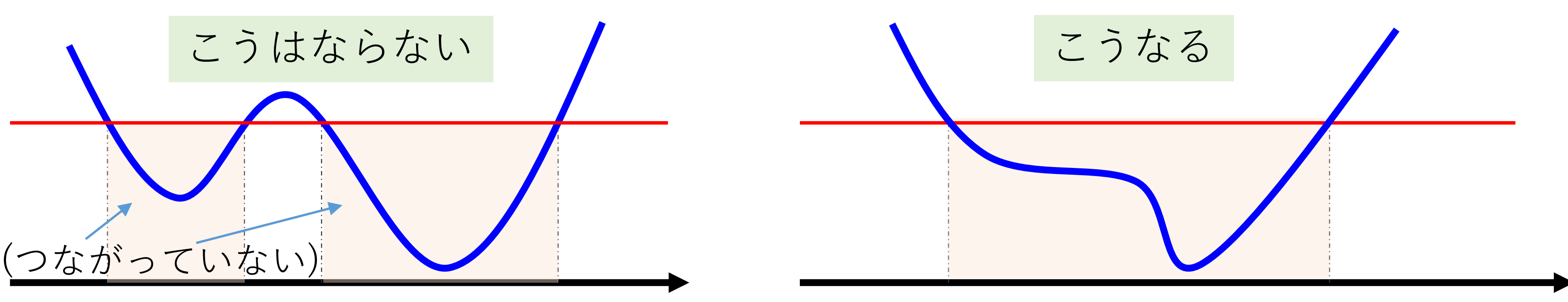
Loss landscape

- 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$f_{(a,W)}(x)=\sum_{m=1}^M a_m\,\eta(w_m^\top x)$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

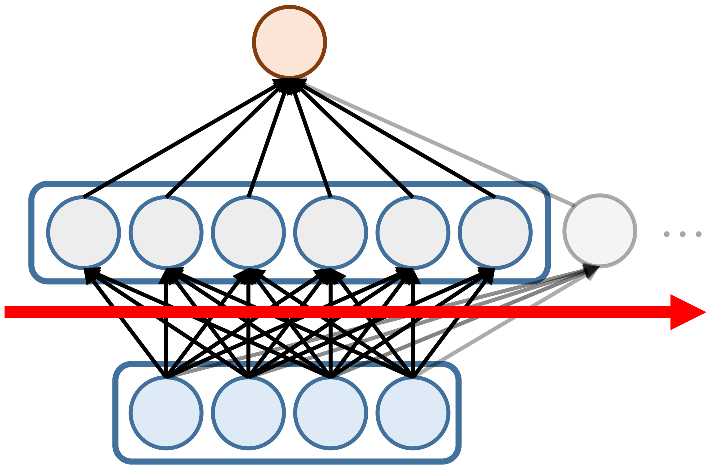
$$y = \sum_{j=1}^{k} \overbrace{v_j}^{\text{固定}} \eta(w_j^\top x + b_j)$$

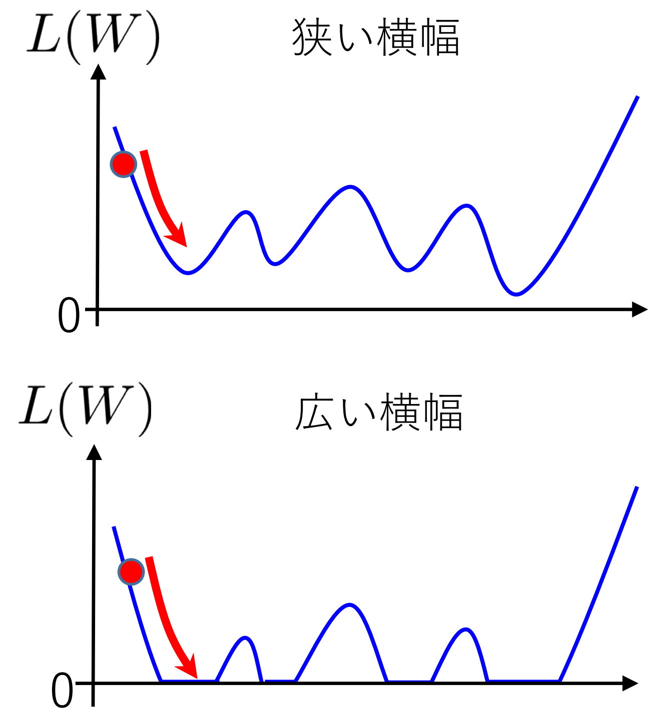
- Li and Yuan (2017): ReLU，入力はガウス分布を仮定
  - ➢ SGDは多項式時間で大域的最適解に収束
  - ➢ 学習のダイナミクスは2段階
    → 最適解の近傍へ近づく段階 + 近傍での凸最適化的段階
- Soltanolkotabi (2017): ReLU，入力はガウス分布を仮定
  - ➢ 過完備（横幅>サンプルサイズ）なら勾配法で最適解に線形収束
    (Soltanolkotabi et al. (2017)は二乗活性化関数でより強い帰結)
- Brutzkus et al. (2018): ReLU
  - ➢ 線形分離可能なデータなら過完備ネッ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．
    (線形パーセプロ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
  - ➢ Mean-field analysis (平均場解析)

二つのスケーリング

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

- 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.

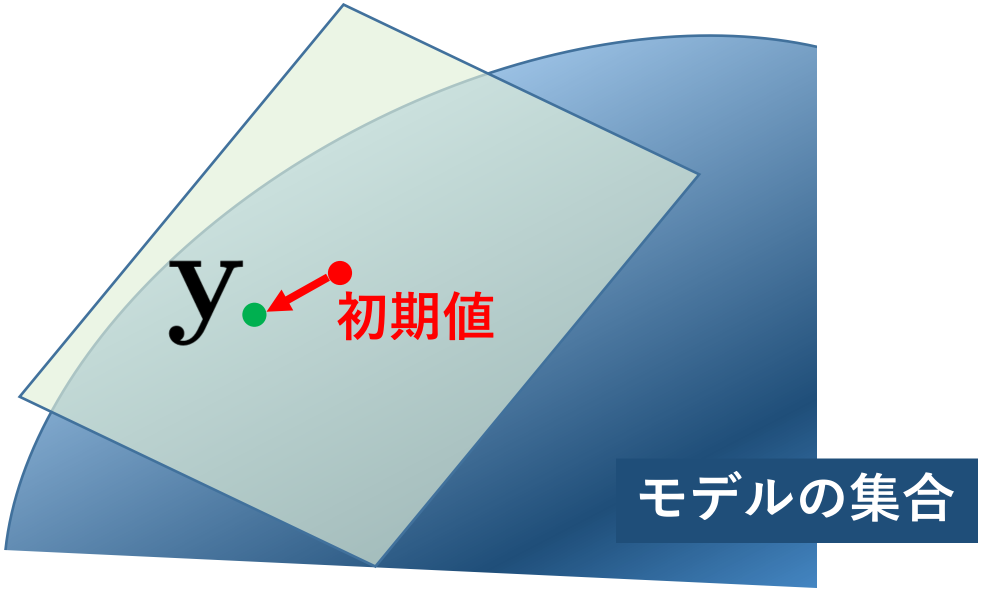
初期化のスケーリングによって，初期値と比べて学習によって動く大きさの割合が変わる.
→ 学習のダイナミクス，汎化性能に影響

(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．



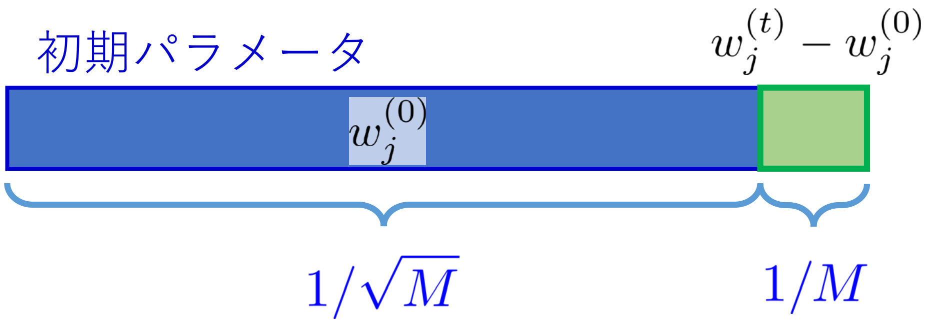
NTKと平均場の違い

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

η: ReLUとする． $a_j = O(1), w_j = O(1/\sqrt{M})$
または$w_j = O(1/M)$とスケール変換

- 各w_jが$O(1/M)$だけ動けば，全体としてO(1)の変化(データにフィットできる).
- 横幅は十分大きく取る: $M \gg n$ (overparameterization)

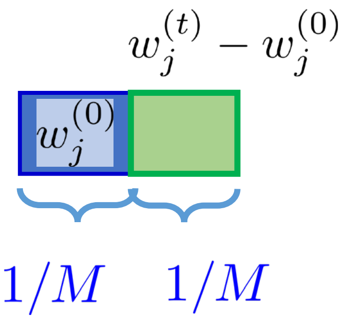
NTK：相対的変化小



$$\eta(w_j^{(t)\top} x) - \eta(w_j^{(0)\top} x)$$
$$\simeq (w_j^{(t)} - w_j^{(0)})^\top \underbrace{x\eta'(w_j^{(0)\top} x)}_{\text{NTKの特徴写像}}$$

テイラー展開により線形モデルとみなせる
→カーネル法の理論に帰着できる

平均場：相対的変化大



相対的な変化が大きいのでテイラー展開ができない.
→ 本質的に非凸最適化になる.

（原理的には展開しても良いが，
グラム行列の正定値性が保証されない）

Neural Tangent Kernel

連続時間ダイナミクスを考える.

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

- a_j は固定
- w_j を学習

[Jacot, Gabriel&Hongler, NeurIPS2018]

$$\frac{\mathrm{d}w_j}{\mathrm{d}t} = -\nabla_{w_j} \hat{L}(f_W) \qquad \text{(Gradient descent, GD)}$$

$$= -\frac{1}{n} \sum_{i=1}^{n} \ell_i'(f_W(x_i)) a_j \nabla_{w_j} \eta(w_j^\top x_i)$$

$$\boxed{\nabla_{w_j} \eta(w_j^\top x_i) = x_i \eta'(w_j^\top x_i)}$$

(勾配法による更新)

$$\frac{\mathrm{d}f_W(x)}{\mathrm{d}t} = \sum_{j=1}^{M} a_j \nabla_{w_j}^\top \eta(w_j^\top x) \frac{\mathrm{d}w_j}{\mathrm{d}t}$$

$$= -\frac{1}{n} \sum_{i=1}^{n} \Bigg(\underbrace{\sum_{j=1}^{M} a_j^2 \nabla_{w_j}^\top \eta(w_j^\top x) \nabla_{w_j} \eta(w_j^\top x_i)}_{k_W(x, x_i)} \Bigg) \underbrace{\ell_i'(f_W(x_i))}_{\text{residual}}$$

O(1/M)：
特徴写像の内積の平均

(関数勾配)

$k_W(x, x_i)$
Neural Tangent Kernel

目的関数の減少速度

$$\begin{aligned}
\frac{\mathrm{d}\hat{L}(f_W)}{\mathrm{d}t} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{d}f_W(x_i)}{\mathrm{d}t}\ell_i'(f_W(x_i)) \\
&= -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\ell_i'(f_W(x_i))\underbrace{k_W(x_i,x_j)}_{(K_W)_{i,j}}\ell_j'(f_W(x_j)) \\
&= -\frac{1}{n^2}\|\nabla_f \hat{L}(f_W)\|_{K_W}^2 \\
&\le -\lambda_{\min}\frac{1}{n^2}\|\nabla_f \hat{L}(f_W)\|^2
\end{aligned}$$

($\lambda_{\min}$：グラム行列の**最小固有値**)

Fact [Du et al., 2018; Allen-Zhu, Li & Song, 2018]

- ランダム初期化しておけば, $K_{W(0)} \succ \epsilon I$ が高確率で成立.
- 最適化の最中に最小固有値は正のまま ($\geq \epsilon/2$).

→ **線形収束** ($\exp(-\boldsymbol{\lambda}_{\min}\boldsymbol{t})$)

ランダム初期値とNTKの正定値性

$$K_{\infty,i,j} = \mathbb{E}_{w\sim N(0,I)}[x_i^\top x_j \eta'(w^\top x_i)\eta'(w^\top x_j)]$$

(横幅無限大のNTK)

補題

$$\|x_i\| = 1,\ \|x_i - x_j\| \geq \phi \quad \Rightarrow \quad K_\infty \succeq C\phi n^{-2}$$

Hoeffdingの不等式より

$$P\left(|K_{W^{(0)},i,j} - K_{\infty,i,j}| \leq \sqrt{\frac{\log(2/\delta')}{2M}}\right) \geq 1-\delta'$$

一様バウンドを取って

$$P\left(\|K_{W^{(0)}} - K_\infty\|_{\mathrm{F}}^2 \leq n^2\frac{\log(2n^2/\delta)}{2M}\right) \geq 1-\delta$$

十分横幅Mが広ければ，ランダム初期化した$K_{W^{(0)}}$の正定値性が保証される．

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]

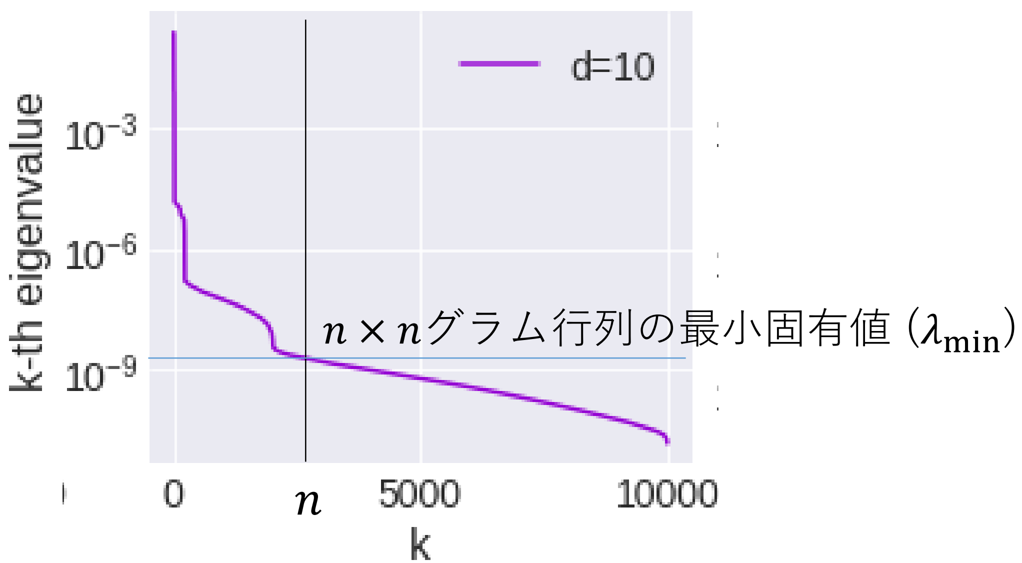
- 訓練誤差0の解に線形収束する.
- 汎化誤差も一応抑えられている.

- 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.
- Early stoppingや正則化を入れれば過学習を防げる. (次ページ)

Spectral bias

- 最適化の観点からはoverparameterizationは有用に見える.
- 汎化誤差はどうであろうか?

NTKの固有値分布



- グラム行列の最小固有値は小さい ($1/\mathrm{poly}(n)$).
- 固有値の減少レートは多項式オーダー (理論+実験).

→ Spectral bias: 汎化の意味では好ましい.

Kernelによる平滑化という視点

- Frechet 微分 in $L_2(P_n)$: $\nabla_f \hat{L}(f)$

$$\nabla_f \hat{L}(f) = (\ell_i'(f(x_i)))_{i=1}^n$$

$$\hat{L}(f+h) = \hat{L}(f) + \langle \nabla_f \hat{L}(f), h \rangle_{L_2(P_n)} + o(\|h\|_{L_2(P_n)}^2)$$

- **平滑化**積分作用素:

$$T_k f(x) := \int k(x,x') f(x') \mathrm{d}P_n(x')$$

$$\boxed{T_{k_W} \phi_j = \mu_j \phi_j}$$

- NTKにおける勾配は関数勾配を平滑化したもの:

$$\frac{\mathrm{d}f_W}{\mathrm{d}t} = -T_{k_W} \nabla_f \hat{L}(f_W)$$

勾配を平滑化！

$$\left(= -\frac{1}{n} \sum_{i=1}^n k_W(\cdot, x_i) \ell_i'(f_W(x_i))\right)$$

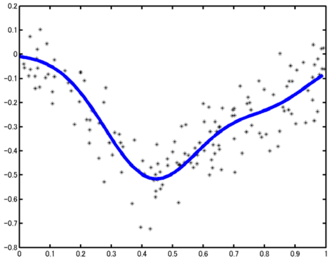
k_Wが高周波成分に小さな固有値を持てば，T_{k_W}は平滑化作用素として働く→ 帰納的バイアス (inductive bias).

NTK regimeでのSGD

モデル:

$$f_{a,W}(x) = \frac{1}{\sqrt{M}} \sum_{j=1}^{M} a_j \eta(w_j^\top x)$$

(We train both of first and second layers)



目的関数:

$$L(a,W) = \underbrace{\mathbb{E}[(Y - f_{a,W}(X))^2]}_{\text{期待損失}} + \underbrace{\frac{\lambda}{2}(\|a - a^{(0)}\|^2 + \|W - W^{(0)}\|_{\mathrm{F}}^2)}_{\text{初期値からのずれ}}$$

$Y = f^*(X) + \epsilon$ （ノイズありの観測）

Averaged Stochastic Gradient Descent

for $t = 0$ **to** $T-1$ **do**

Randomly draw a sample $(x_t, y_t) \sim \rho$

Perform SGD update for all $j \in \{1, \dots, M\}$:

$$a_j^{(t+1)} = a_j^{(t)} - \alpha_t [\nabla_a \ell(y_t, f_{a^{(t)},W^{(t)}}(x_t)) + \lambda(a^{(t)} - a^{(0)})]$$

$$W_j^{(t+1)} = W_j^{(t)} - \alpha_t [\nabla_W \ell(y_t, f_{a^{(t)},W^{(t)}}(x_t)) + \lambda(W^{(t)} - W^{(0)})]$$

end for

Return $\bar{a}^{(T)} = \frac{1}{T}\sum_{t=0}^{T-1} a^{(t)}, \ \bar{W}^{(T)} = \frac{1}{T}\sum_{t=0}^{T-1} W^{(t)}.$

NTKにおける余剰誤差の速い収束

[Nitanda&Suzuki: Fast Convergence Rates of Averaged Stochastic Gradient Descent under Neural Tangent Kernel Regime, 2020.]

仮定：真の関数がNTKの作るRKHSに入っているとする.

NTK設定で適切な正則化を入れたSGDは“速い学習レート”を達成できる.
→ NTKによるsmoothingのおかげ.

Thm (速い収束レート)

f_T: T回更新後の解

$$\mathbb{E}[\|f_T - f^*\|_{L_2}^2] \leq \epsilon_M + O(T^{-\frac{2r\overbrace{\beta}^{\text{NTKの固有値の減衰レート}}}{2r\beta+1}})$$

$M \to \infty$で0に収束する項

速い学習レート
($O(1/\sqrt{T})$より速い)

→ $T^{-\frac{2r\beta}{2r\beta+1}}$ は**ミニマックス最適レート.**
(各種パラメータの意味は次ページに詳細)

仮定

2層NNのNTK:

$$k_\infty(x,x') = \mathbb{E}_{w^{(0)}}[\eta(w^{(0)\top}x)\eta(w^{(0)\top}x')] + \mathbb{E}_{w^{(0)}}[\eta'(w^{(0)\top}x)\eta'(w^{(0)\top}x')x^\top x]$$

横幅無限における積分作用素:

$$T_{k_\infty}f(x) = \int k_\infty(x,x')f(x')\mathrm{d}P_X$$

population

スペクトル分解: $T_{k_\infty}\phi_j = \mu_j\phi_j\,,\ k_\infty(x,x') = \sum_{j=1}^\infty \mu_j\phi_j(x)\phi_j(x')$

仮定

- $f^*(x) = \mathrm{E}[Y|X=x]$ が次のように書ける:
$$T_{k_\infty}^r h = f^*$$
for $h \in L_2(P_X)$, and $r \in [1/2,1]$.

- 固有値減衰条件:
$$\mu_j = O(j^{-\beta}).$$

真の関数の平滑性

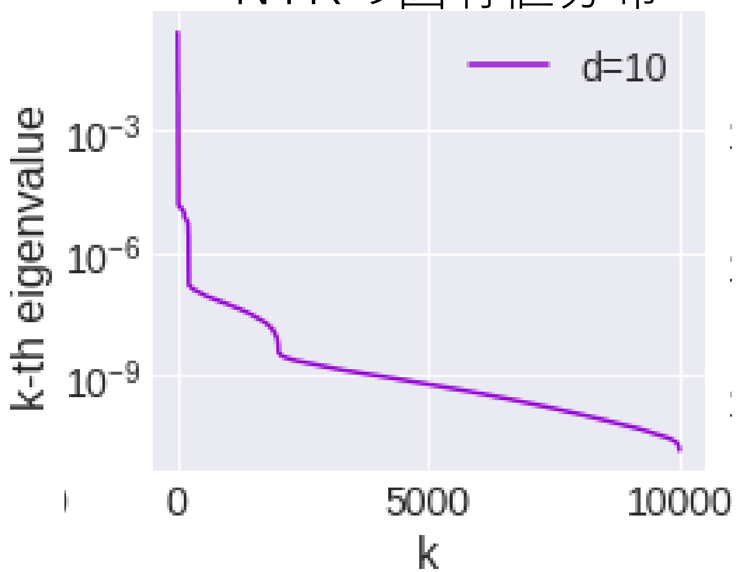
カーネル関数の“複雑さ”

カーネルリッジ回帰の解析における標準的な仮定; see, e.g., Dieuleveut et al. (2016); Caponnetto and De Vito (2007) (rの条件はやや強め).

$$k_\infty(x,x') = \sum_{m=1}^{\infty} \lambda_m \phi_m(x)\phi_m(x')$$

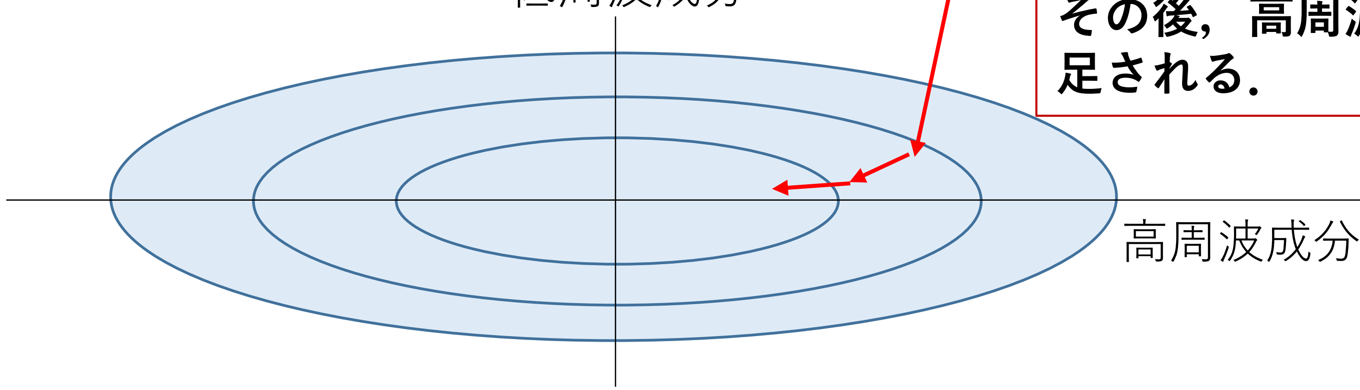
NTKの固有値固有関数分解
$(\phi_m)_{m=1}^{\infty}$: 固有関数. $L_2(P_X)$内の正規直交基底.

NTKの固有値分布



実際のNTKの固有値は多項式オーダーで減衰する.

[Bietti&Mairal (2019); Cao et al. (2019); Ronen et al. (2019)]



低周波成分が最初に補足される. その後, 高周波成分が徐々に補足される.

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:

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.]

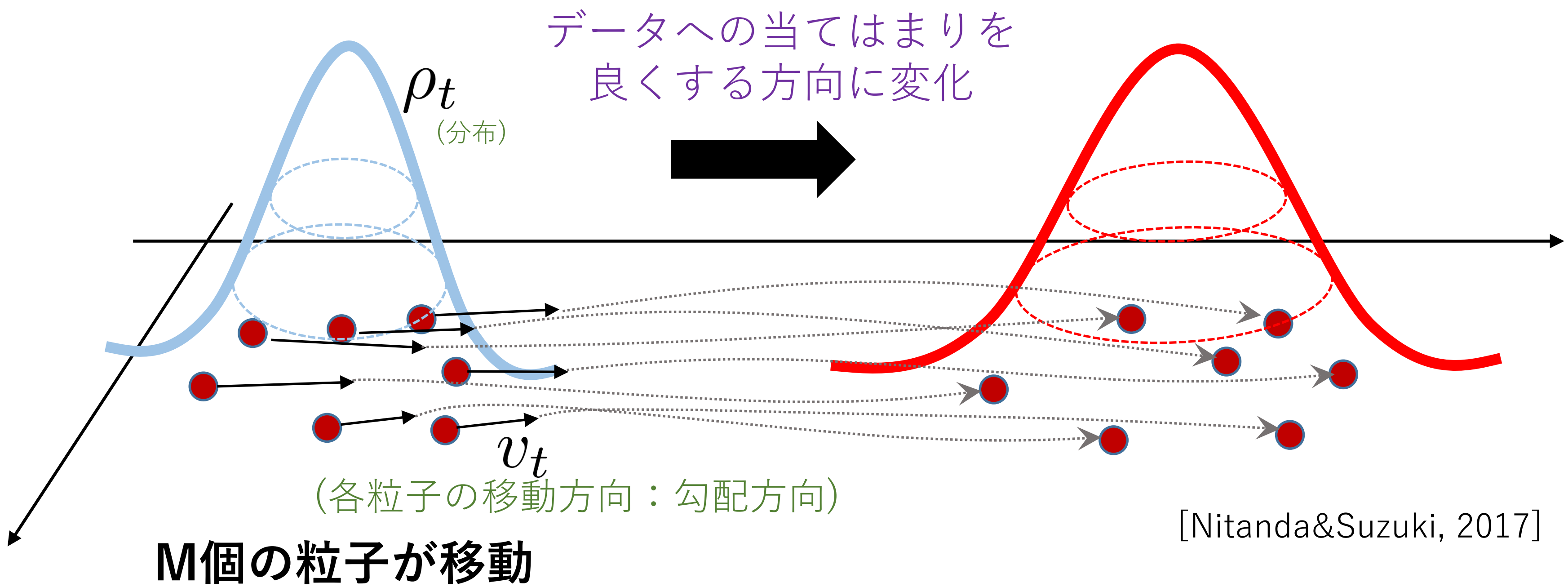
$$v_t(a,w) = -\frac{1}{n}\sum_{i=1}^{n} \nabla_{(a,w)}(a\eta(w^\top x_i))\ell'(y_i, f_{\rho_t}(x_i))$$ （各粒子は勾配降下方向へ移動）

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]

ノイズありのダイナミクス: McKean-Vlasov過程

[Mei, Montanari&Nguyen, 2018]

Wasserstein距離について

μ,ν:距離空間$(\mathcal{X},c)$上の確率測度(通常$\mathcal{X}$はPoland空間)

$$W_p(\mu,\nu)=\left(\inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathcal{X}\times\mathcal{X}}c(x,y)^p\mathrm{d}\pi(x,y)\right)^{1/p}$$

$\Pi(\mu,\nu)$: 周辺分布がμ,νである$\mathcal{X}\times\mathcal{X}$上の同時分布の集合

周辺分布を固定した同時分布の中で最小化

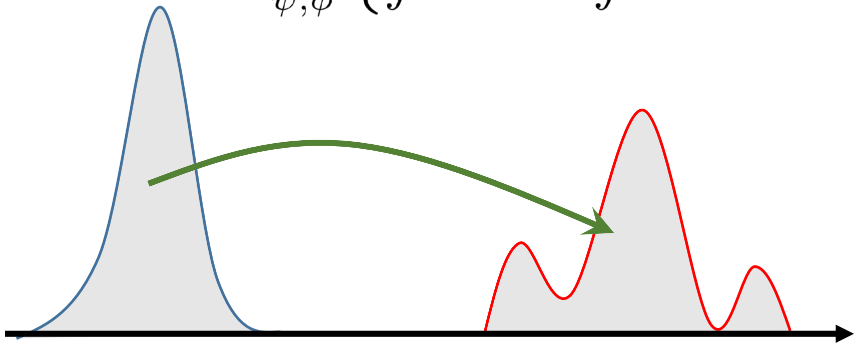
$$(\mathcal{X}=\mathbb{R}^d:\ c(x,y)=\|x-y\|)$$

- 分布のサポートがずれていてもwell-defined
- 底空間の距離が反映されている

※KL-divergenceは距離が反映されない.

(双対表現: Kantorovich双対)

$$\inf_{\pi\in\Pi(\mu,\nu)}\int c(x,y)^p\mathrm{d}\pi(x,y)=\sup_{\psi,\phi}\left\{\int\psi\mathrm{d}\mu+\int\phi\mathrm{d}\nu\ \middle|\ \psi(x)+\phi(y)\le c(x,y)^p\right\}$$



「輸送距離」とも言われる

W_2距離と粒子勾配降下法の関係

W_2距離による近接点アルゴリズムを考える：
（δは十分小さいとする）

$f_\nu(x) = \int h(w,x)\mathrm{d}\nu(w)$

$$
\begin{aligned}
&L(\nu) + \frac{W_2^2(\mu,\nu)}{2\delta} \\
&= \mathbb{E}_X\left[\ell\left(\int h(w,X)\mathrm{d}\nu(w)\right)\right] + \frac{W_2^2(\mu,\nu)}{2\delta} \\
&\simeq \Big\langle \underbrace{\ell'\left(\int h(w,X)\mathrm{d}\mu(w)\right)}_{=:\Delta_\mu(X)}, \int h(v,X)\mathrm{d}(\nu-\mu)(v) \Big\rangle_{L^2(P)} + \inf_{\pi\in\Pi(\mu,\nu)} \int \frac{\|w-v\|^2}{2\delta}\mathrm{d}\pi(w,v) \\
&\simeq \inf_{\pi\in\Pi(\mu,\nu)} \int \underbrace{\left(\Big\langle \Delta_\mu, h(v,\cdot) \Big\rangle_{L^2(P)} + \frac{\|w-v\|^2}{2\delta}\right)} \mathrm{d}\pi(w,v) + \mathrm{const.}
\end{aligned}
$$

νについて最小化→各wごとにvの条件付分布を最小化 →Dirac測度（∵局所的に強凸）

$$
\begin{aligned}
v &= \operatorname*{arg\,min}_{v'}\left\{\langle \Delta_\mu, h(v',\cdot)\rangle_{L^2(P)} + \frac{\|w-v'\|^2}{2\delta}\right\} \\
&\simeq w - \delta\nabla_w\langle h(w,\cdot), \ell'(f_\mu)\rangle
\end{aligned}
$$

：最急降下法

- 各粒子ごとにみると単純な最急降下法.
- 粒子勾配降下法はW_2距離を近接項とした近接点アルゴリズムの一次近似
  → $\delta \to 0$の極限（連続時間）：Wasserstein gradient flow

連続の方程式と勾配流

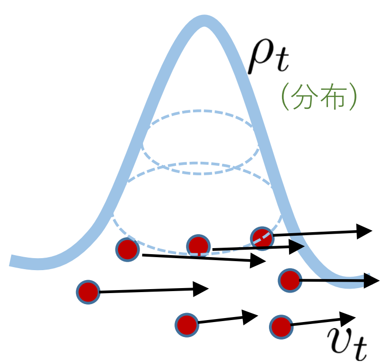
「連続の方程式」 $\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (v_t \rho_t)$ の意味

$$\frac{\mathrm{d}}{\mathrm{d}t}\int f(w)\mathrm{d}\rho_t(w) = \int (\nabla f(w))^\top v_t(w)\mathrm{d}\rho_t(w)$$

$\left(= -\int f(w)\mathrm{d}[\nabla \cdot (v_t\rho_t)]\right)$　　（$\forall f$: コンパクトサポート，C^∞-級）

- 今，ρ_tは写像$T_t: R^d \to R^d$によるρ_0の<u>押し出し</u>であるとする：$\rho_t = T_{t\#}\rho_0$. つまり，$w \sim \rho_0$に対する$T_t(w)$の分布がρ_tであるとする．
- 写像T_tを生成するベクトル場を$\frac{\mathrm{d}T_t}{\mathrm{d}t}(w) = v_t(T_t(w))$とする．

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int f(w)\mathrm{d}\rho_t(w) &= \frac{\mathrm{d}}{\mathrm{d}t}\int f(T_t(w))\mathrm{d}\rho_0(w) \\ &= \int \nabla f(T_t(w))^\top \frac{\mathrm{d}T_t(w)}{\mathrm{d}t}\mathrm{d}\rho_0(w) \\ &= \int \nabla f(T_t(w))^\top v_t(T_t(w))\mathrm{d}\rho_0(w) \\ &= \int \nabla f(w)^\top v_t(w)\mathrm{d}\rho_t(w).\ \text{(連続の方程式)}\end{aligned}$$



$w_t = T_t(w)$ に対し，$v_t(w_t) = -\nabla_w \langle h(w,\cdot), \ell'(f_{\rho_t})\rangle|_{w=w_t}$ としたのが前ページの更新式．

接ベクトル

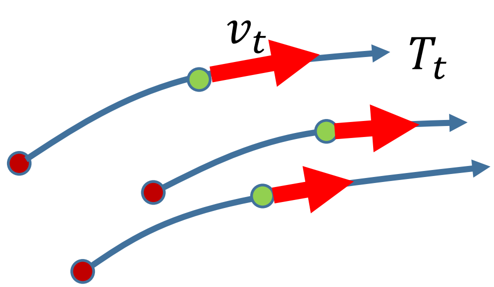
定理

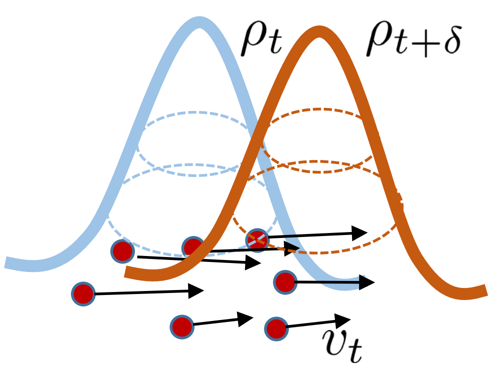
- $\rho_t = T_{t\#}\rho_0$
- $\frac{\mathrm{d}T_t}{\mathrm{d}t}(w) = v_t\big(T_t(w)\big)$
- あるϕ_tを用いて$v_t = \nabla\phi_t$と書けるとする．

この時，以下が成り立つ:

$$\lim_{\delta\to 0}\frac{W_2(\rho_{t+\delta}, (\mathrm{id}+\delta v_t)_{\#}\rho_t)}{\delta} = 0$$

詳細は以下を参照:
Ambrosio, Gigli, and Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2008.





輸送写像

Brenierの定理

ρ_0, ρ_1が確率密度関数を持つ時，以下が成り立つ:

$$W_2^2(\rho_0, \rho_1) = \inf_{T: T_\# \rho_0 = \rho_1} \mathbb{E}_{X \sim \rho_0}[\|X - T(X)\|^2]$$

- Infを達成する写像T^*が存在する．
- しかも，ある凸関数ψが存在して$T^*(x) \in \partial\psi(x)$と書ける．
- このT^*を最適輸送写像という．

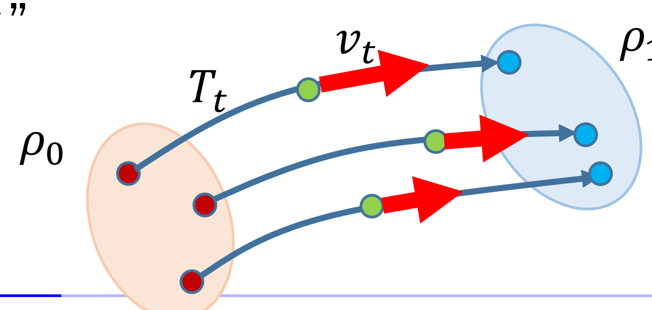
Benamou-Brenier formula (連続の方程式とW_2距離の関係):

同条件のもと

$$W_2^2(\rho_0, \rho_1) = \inf_{\{v_t\}_t} \int_0^1 \|v_t\|_{L_2(\rho_t)}^2 \mathrm{d}t$$

ただし，infはρ_0からρ_1へ連続の方程式で“繋ぐ”
全ての速度ベクトル場v_tに関して取る．

- $\rho_t = T_{t\#}\rho_0$
- $\frac{\mathrm{d}T_t}{\mathrm{d}t}(w) = v_t(T_t(w))$



- Wasserstein勾配流はW_2距離を用いた近接点アルゴリズムで特徴づけられることが分かった.
- 目的関数がW2距離に関する凸性（displacement convexity）が成り立つなら，大域的最適解への収束が示せる（エントロピーなど）：

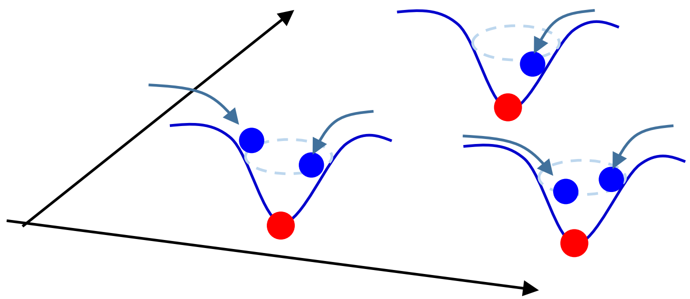
$$W_2(\rho_t, \rho^*) \leq e^{-\lambda t} W_2(\rho_0, \rho^*).$$

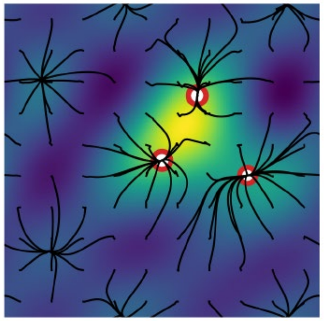
- しかし，NNの最適化では凸性は成り立たない.

そのため，大域収束を示すことが難しい.

- もっとも，局所的には凸性が成り立ちうる.

例：スパースな最適解 [Chizat, 2019]





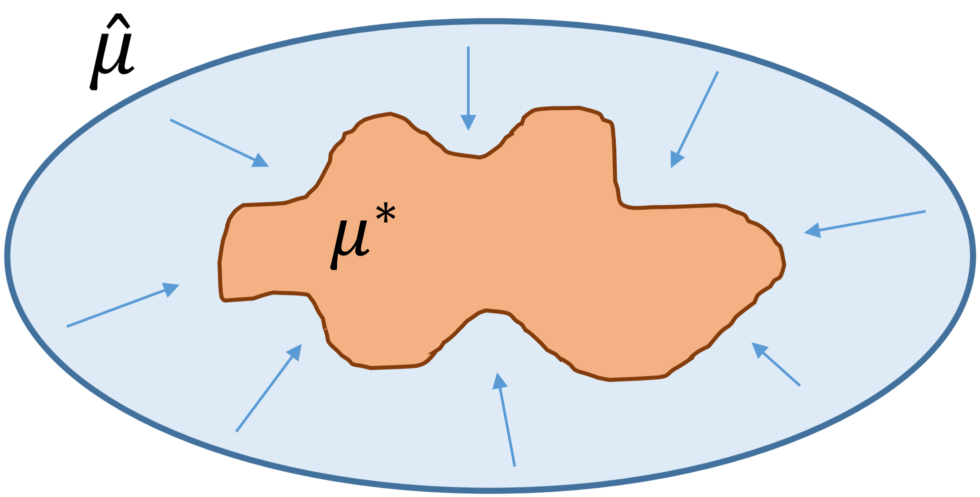
局所最適性条件

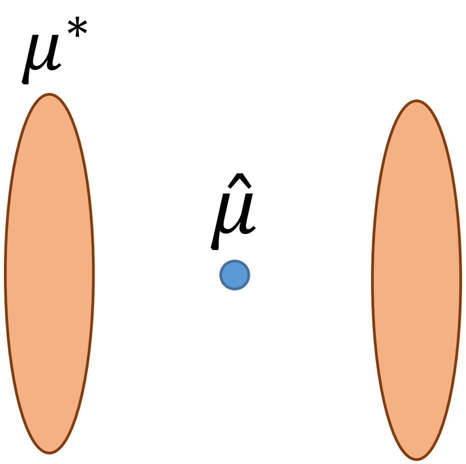
定理 (Nitanda&Suzuki, 2017)

ある解$\hat{\mu}$がコンパクトな台の確率密度関数を持つとする．
この時，あるμ^* s.t. $L(\mu^*) < L(\hat{\mu})$が存在して

- $\mathrm{supp}(\mu^*) \subseteq \mathrm{supp}(\hat{\mu})$かつ$\mu^*$は確率密度を持つ，or
- μ^*は密度を持たず$\mathrm{supp}(\mu^*)$は$\mathrm{supp}(\hat{\mu})$に内部に含まれる，

が満たされるとき，降下方向が存在して粒子降下法によって目的関数値を減らすことができる．





このような場合は降下方向は無い

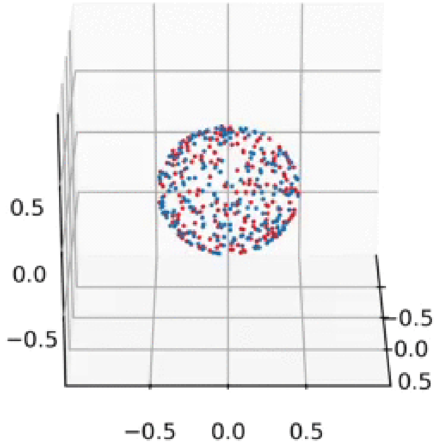
（参考）平均場解析と陰的正則化

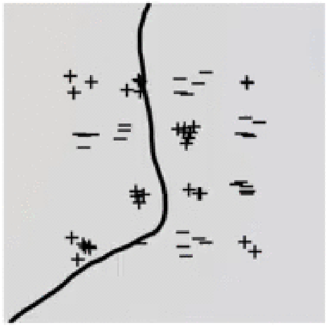
二値判別をexp-損失を用いて解く (ラベルノイズなしとする):

$$\min_{\rho} \sum_{i=1}^{n} \exp\left(-y_i f_\rho(x)\right) \qquad \text{ただし} \qquad f_\rho(x) = \int \eta(w^\top x) \mathrm{d}\rho(w)$$

符号付測度の中で最適化

平均場解析の設定で最適化する.
初期値が小さいので判別に必要なニューロンだけが「生えてくる」.





[Chizat&Bach:Implicit Bias of Gradient Descent for Wide Two-layer Neural Networks Trained with the Logistic Loss. COLT2020.]

→ **スパースな解：<u>陰的正則化</u>**

最適化の結果として「単純な」解が求まってしまう.

判別平面はL1-正則化解マージン最大化元に収束する:

$$\max_{\rho:\|\rho\|_{\mathcal{F}_1}} \min_{i\in\{1,\dots,n\}} y_i f_\rho(x_i) \qquad \|\rho\|_{\mathcal{F}_1} = |\rho|(\mathbb{R}^d)$$

（参考）勾配法と陰的正則化

- 小さな初期値から勾配法を始めるとノルム最小化点に収束しやすい→陰的正則化

多層の線形NNを判別問題でGDすると“スパースな解”が得られる．

[Gunasekar et al.: Implicit Regularization in Matrix Factorization, NIPS2017]
[Soudry et al.: The implicit bias of gradient descent on separable data. JMLR2018]
[Gunasekar et al.: Implicit Bias of Gradient Descent on Linear Convolutional Networks, NIPS2018]
[Moroshko et al.: Implicit Bias in Deep Linear Classification: Initialization Scale vs Training Accuracy, arXiv:2007.06738]

（参考）各regimeにおける陰的正則化

各regimeにおける陰的正則化の種類

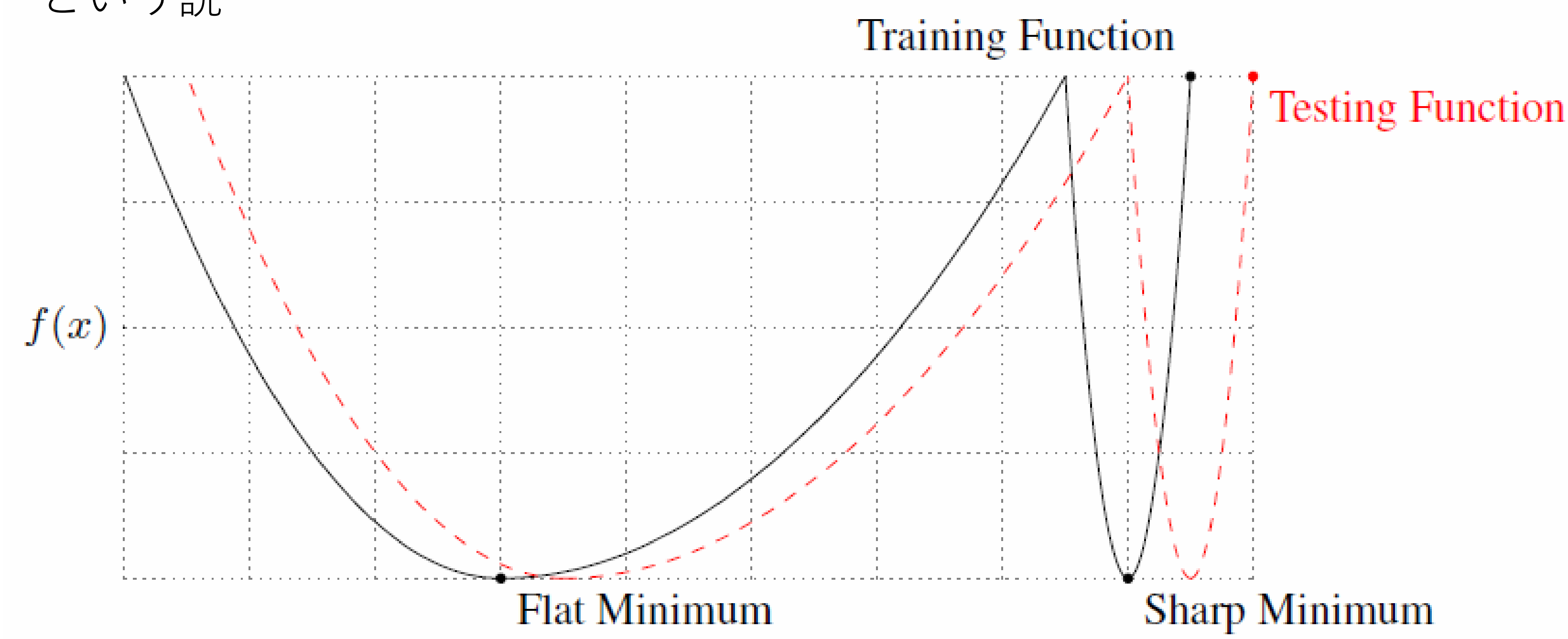
Regime	対応する正則化
NTK, カーネル法 with early stopping	L2-正則化
平均場理論	L1-正則化

- ニューラルネットワークの学習では様々な「**陽的正則化**」を用いる：バッチノーマリゼーション，Dropout，Weight decay，MixUp，…
- 一方で，深層学習の構造が自動的に生み出す「**陰的正則化**」も強く効いていると考えられる．
  → オーバーパラメタライズしても過学習しない．

ノイズあり勾配法と大域的最適性

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



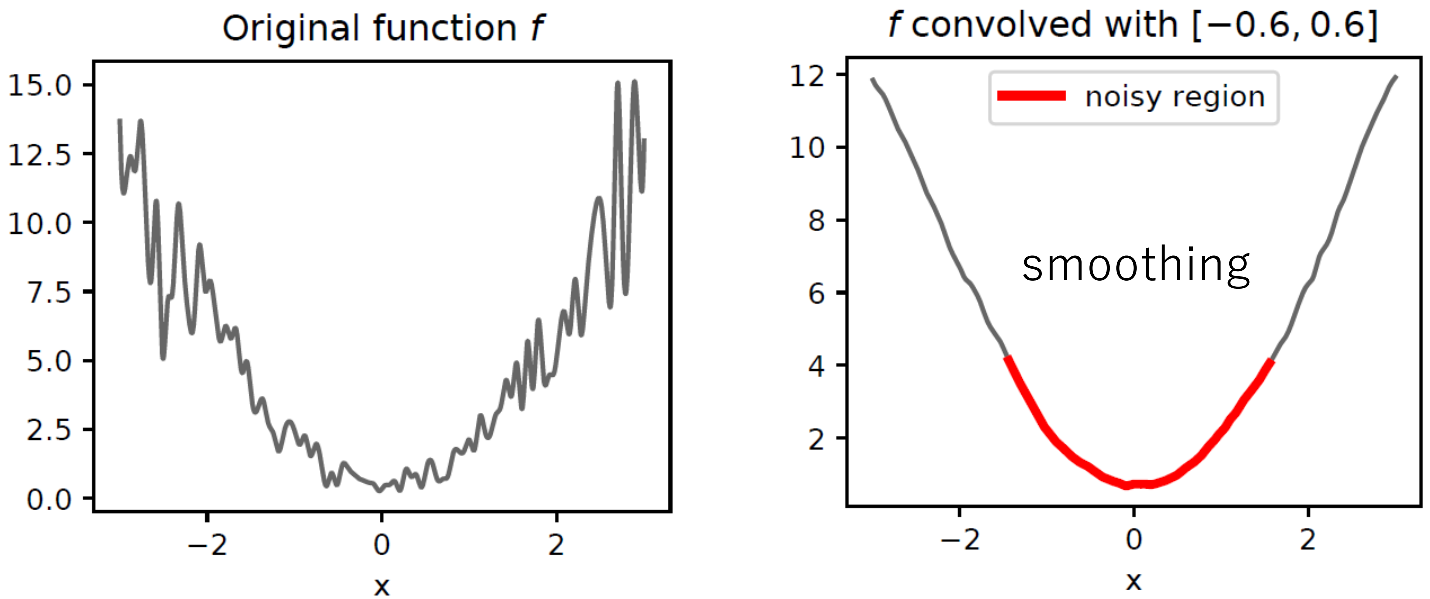
Keskar, Mudigere, Nocedal, Smelyanskiy, Tang (2017):
On large-batch training for deep learning: generalization gap and sharp minima.

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

→ 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

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

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

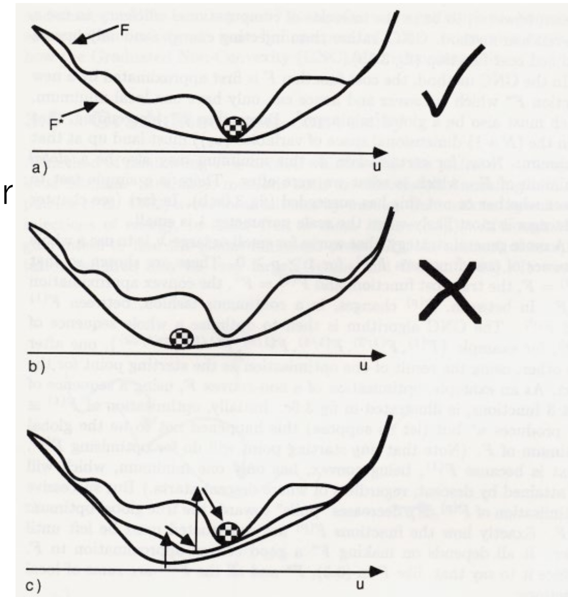
$$\Rightarrow \mathbb{E}_{\xi_{t-1}}[y_t] = y_{t-1} - \eta\nabla\mathbb{E}_{\xi_{t-1}}[L(y_{t-1} - \eta\xi_{t-1})]$$

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.**

関連研究: Graduated optimization

- Graduated non-convexity

Blake and Zisserman: *Visual reconstruction*, volume 2. MIT press Cambridge, 1987.



- Gaussian kernelとの畳み込み

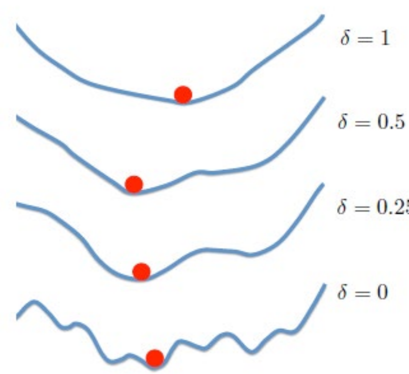
Z. Wu. The effective energy transformation scheme as a special continuation approach to global optimization with application to molecular conformation. SIAM Journal on Optimization, 6(3):748-768, 1996.

- Graduated optimization

Hazan, Levy, and Shalev-Shwartz: On graduated optimization for stochastic non-convex problems. *International conference on machine learning*, pp. 1833-1841, 2016.

σ-nice性の導入．多項式オーダーでの収束.

$$\hat{L}_\delta(x) = \mathrm{E}_{u \sim U(\mathrm{B}(\mathrm{R}^d))}[L(x + \delta u)]$$



Survey:
Mobahi and Fisher III. On the link between gaussian homotopy continuation and convex envelopes. *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pp. 43-56, 2015.

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)]

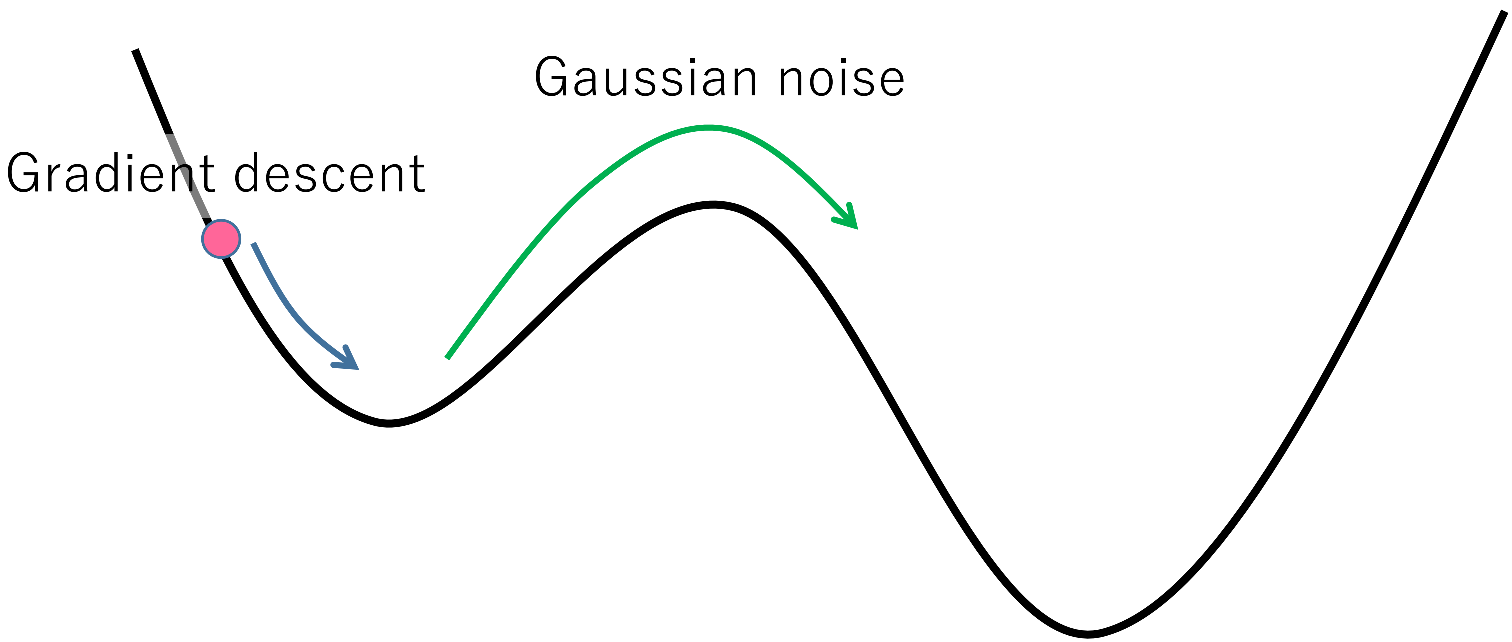
GLD: $$X_{t+1} = X_t - \eta \nabla L(X_t) + \sqrt{2\eta\beta^{-1}}\xi_t$$

(Euler-Maruyama近似)

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

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

確率的勾配


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]

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

- 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

対数Sobolev不等式

$\pi_\infty(\mathrm{d}x) \propto \exp(-\beta L(x))\mathrm{d}x$ ：連続時間ダイナミクスの定常分布

1. 散逸条件 (→Poincareの不等式)
2. 平滑性

➡ **対数Sobolev不等式**

$$\mathrm{d}\nu = f \ \mathrm{d}\pi_\infty \quad \text{(probability)}$$

$$\int f \log(f)\mathrm{d}\pi_\infty \leq 2c_{\mathrm{LS}} \int \frac{\|\nabla f\|^2}{f}\mathrm{d}\pi_\infty \quad (D(\nu||\pi_\infty) \leq 2c_{\mathrm{LS}} I(\nu||\pi_\infty))$$

➡ **Geometric ergodicity**

ρ_t: X_tの周辺分布

$$D(\rho_t||\pi_\infty) \leq \exp(-2t/c_{\mathrm{LS}}) D(\rho_0||\pi_\infty)$$

定常分布へ線形収束

[Bakry, Gentil, and Ledoux: *Analysis and Geometry of Markov Diffusion Operators*. Springer, 2014. Th. 5.2.1]

連続の方程式再び

- 勾配ランジュバン動力学に対応する連続の方程式

$$\partial_t \rho_t = \nabla \cdot (\rho_t \nabla \log(\rho_t / \pi_\infty))$$

- 勾配ランジュバン動力学は相対エントロピー（KL-ダイバージェンス）をWasserstein勾配流で最適化していることに対応:

$$D(\rho || \pi_\infty) = \int \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi_\infty}\right) \mathrm{d}\rho$$

Remark:
通常の（相対でない）エントロピーはW_2-距離に関して凸（displacement convexity）．つまり，W_2-距離に関する測地線上で凸関数になる．

無限次元への拡張

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

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

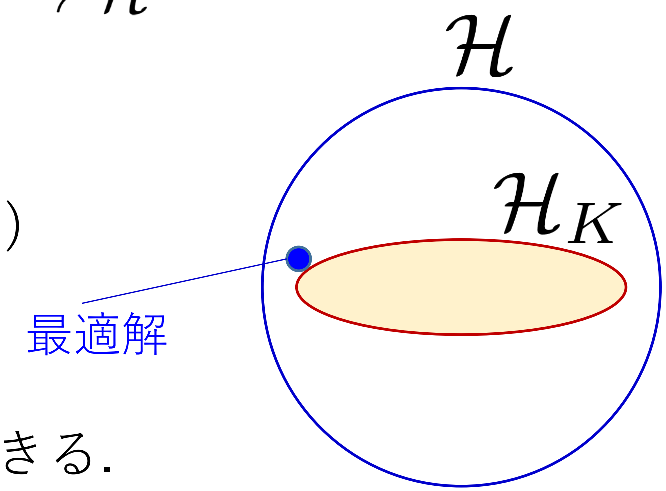
$\mathcal{H}$: Hilbert space

正則化

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

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

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



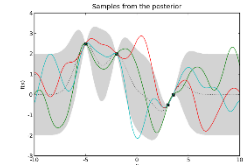
Ex.

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



例: 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>**

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

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

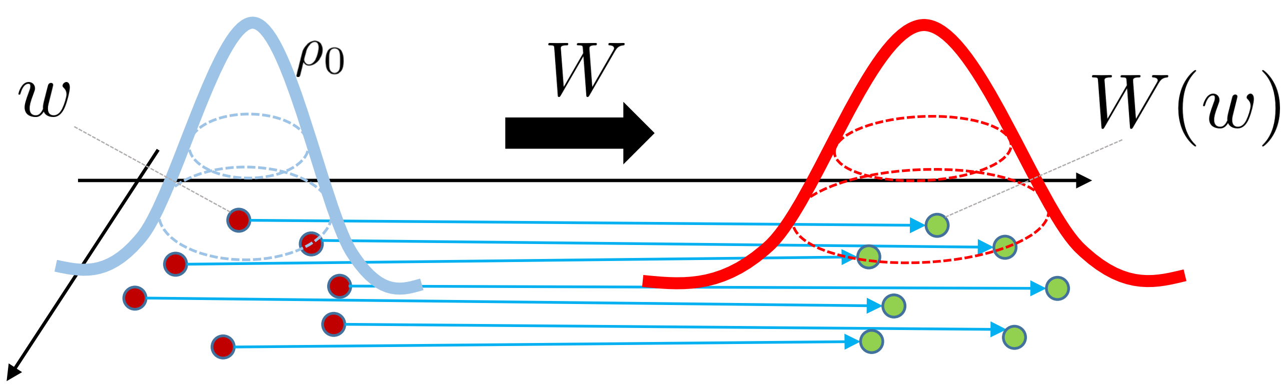
$$= \int_{\mathbb{R}^d} a(w)\sigma(w^\top x)\mathrm{d}W\sharp\rho_0(w)$$

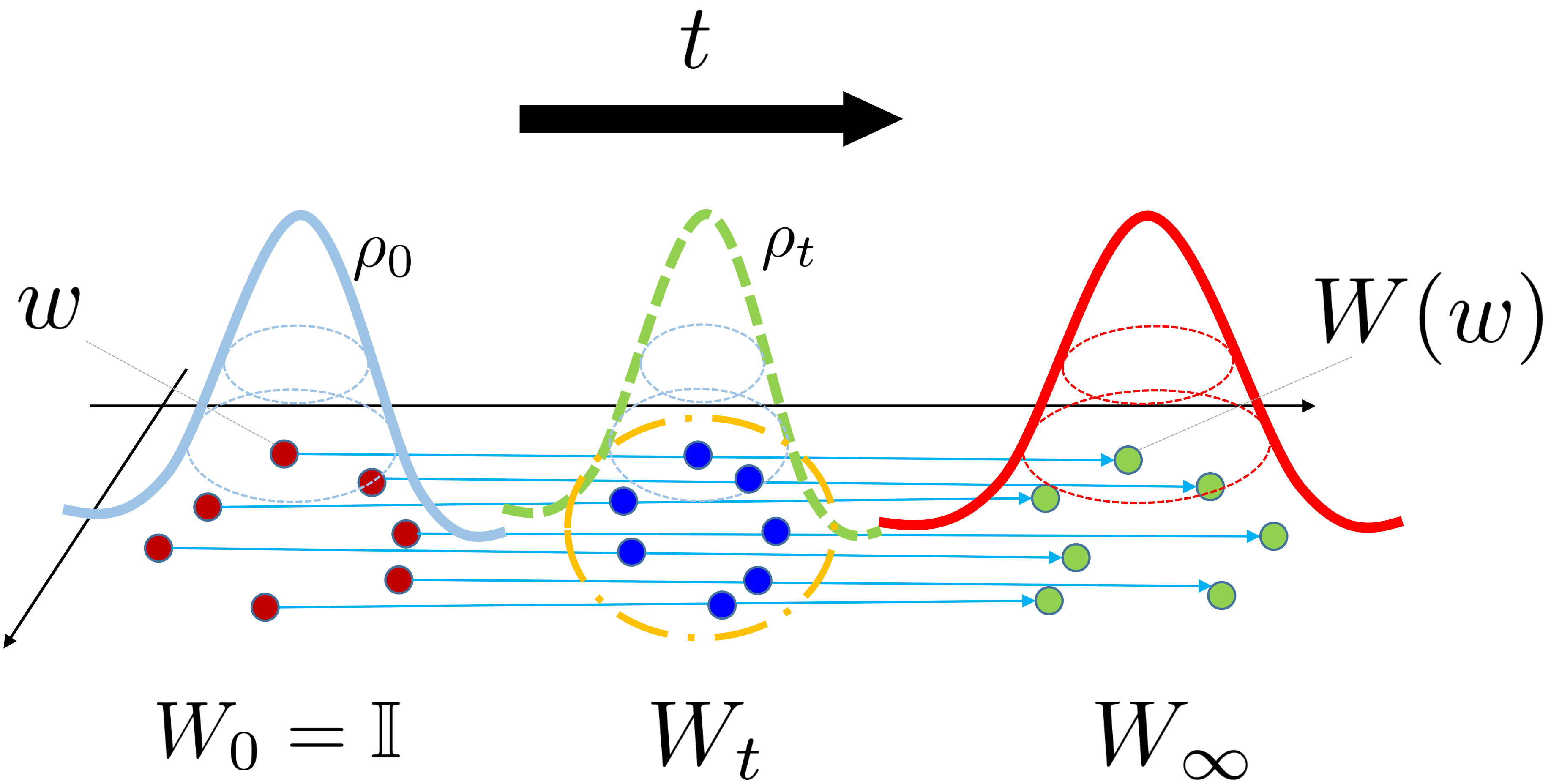
"Lift"

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

以前の表記

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





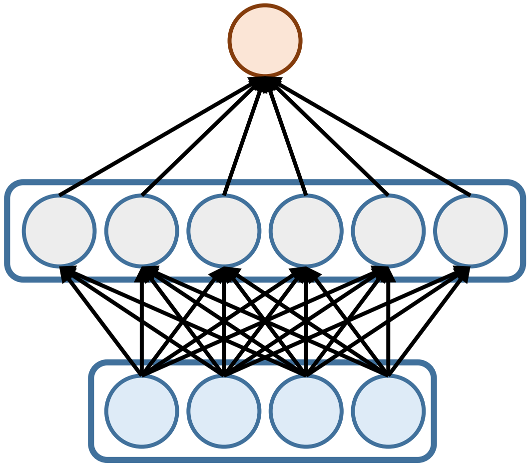
t
ρ_0
ρ_t
w
$W(w)$
$W_0 = \mathbb{I}$
W_t
W_∞

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)

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:

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

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

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

定常分布

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

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

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

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

$$\pi_\infty(x) \;\propto\; \exp\left(-\beta L(x) - \frac{1}{2}x^\top C^{-1}x\right)$$ 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
<u>ガウス過程事前分布を用いたベイズ事後分布</u>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

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

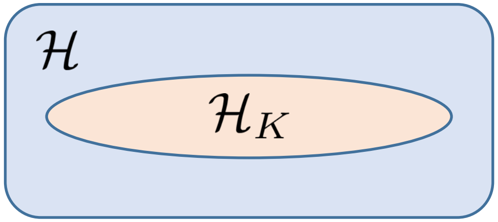
$$\langle x, y \rangle_{\mathcal{H}_K} = \sum_{k=0}^{\infty} \alpha_k \beta_k / \mu_k \quad \text{for } x = \sum_k \alpha_k f_k,\ y = \sum_k \beta_k f_k.$$

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．）

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



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

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

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

Assumption (2)

- Smoothness:

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

- 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）

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

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

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

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 [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).

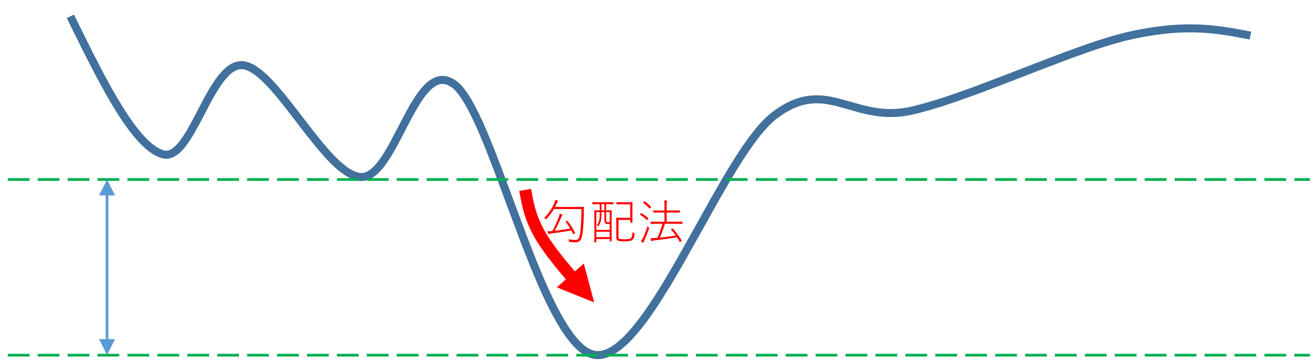
Λ_η^*: スペクトルギャップ，βに対して指数的依存がある．

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:

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

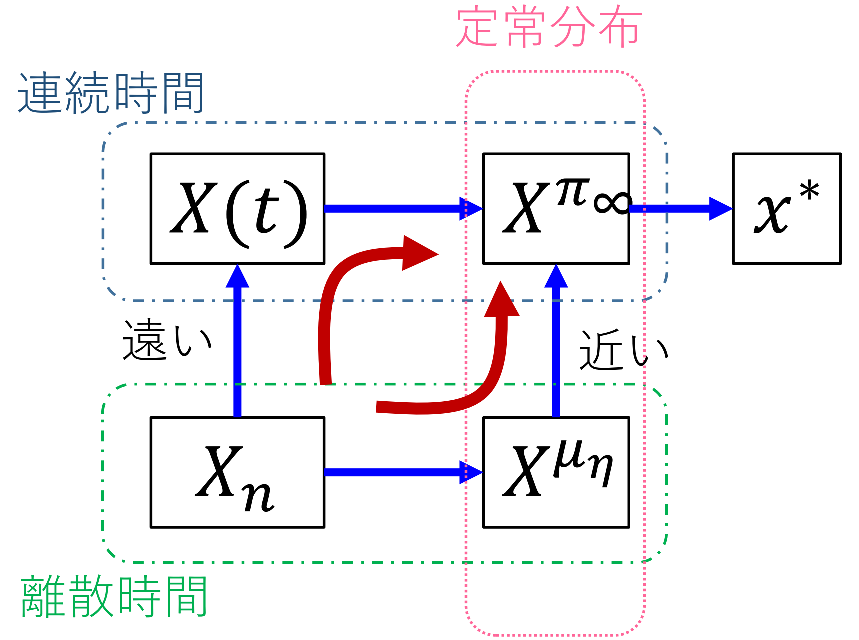
for a smooth function ϕ.

- 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

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

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)

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

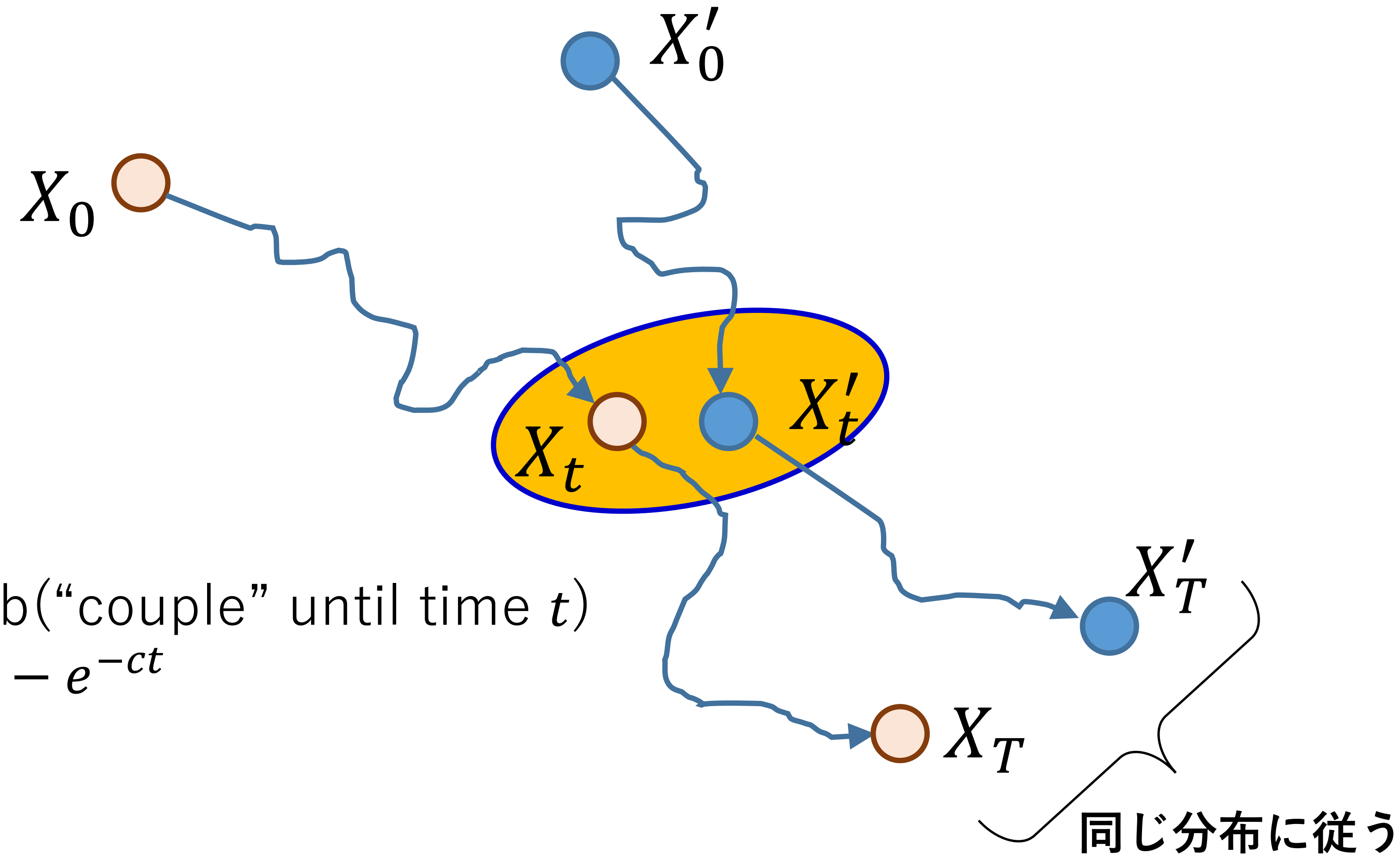
for $\quad \delta = \exp(-O(\beta))$

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

- Coupling argument



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

第二項のバウンド

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

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，

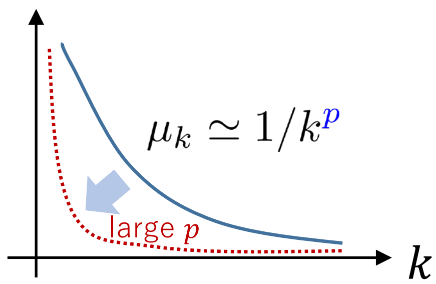
$$|\mathbb{E}[\phi(X^{\mu_\eta}) - \phi(X^{\pi})]| \le C\|\phi\|_{0,2}\frac{c_\beta}{\Lambda_0^*}\eta^{1/2-\kappa}.$$

- $\|\phi\|_{0,2} = \max\{\|\phi\|_\infty, \|D\phi\|_\infty, \|D^2\phi\|_\infty\}$
- $c_\beta = \sqrt{\beta}$ 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．

有限次元バージョンとの関係

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



- $\mu_k \simeq 1/k^2$ （我々の状況）

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

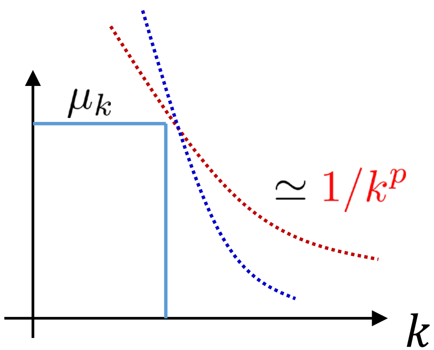
- $\mu_k \simeq 1/k^p$ （予想） see [Andersson,Kruse&Larsson, 2016] for finite time horizon.

pが大きくなるほど関数クラスは“単純”になる.

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

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)]

（参考）判別問題における速い収束

Assumption

- **強低ノイズ条件:**

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

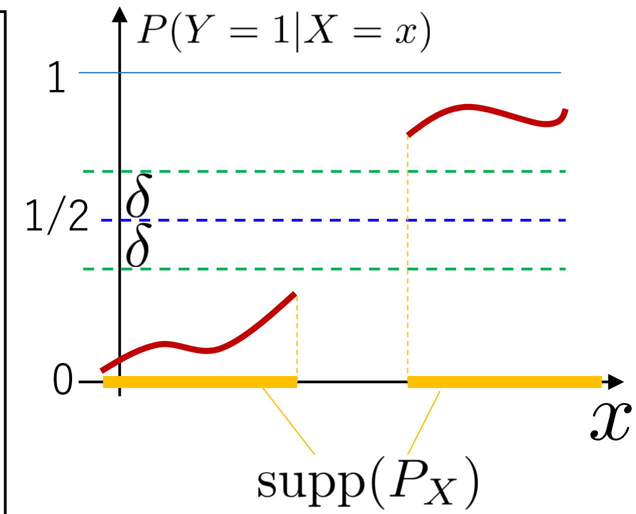
- $\mathrm{supp}(P_X) \subset [0,1]^d$ and P_X has density p such that

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

- 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,

$$\mathbb{E}[P_{\pi_k}(\{W_k \in \mathcal{H} \mid P_X[\mathrm{sign}(f_{W_k}(X)) = \mathrm{sign}(f^*(X))] \neq 0\})]$$
$$\lesssim \exp(-c\beta\delta^{2m/(2m-d)}) + \frac{\Xi_k}{\delta^{2m/(2m-d)}}$$

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)

まとめ

- Overparameterizationの状況では最適解への収束が比較的示しやすい.
  - ➢Neural Tangent Kernel
  - ➢平均場解析
- ノイズを乗せた最適化手法 (SGDは擬似的にこれを実現)
  - ➢局所解から抜けて大域解へ収束
  - ➢無限次元でも理論展開可能
- Wasserstein幾何等の道具を用いて確率測度の収束へ議論を押し付ける

→条件によっては収束が示せる.

総括

- 機械学習の最適化の特徴
  - いかに問題を簡単にして解くか.
- スパース正則化学習：問題を簡単な凸最適化に帰着して解く.
  - 一時期「問題を凸にすれば勝ち」という雰囲気があった. (凸最適化で定式化するのがかっこいい)
- 最近：深層学習の再興により非凸最適化への忌避感が薄まった.
  - きっちり非凸最適化手法を構築する方向性
  - 深層学習の非凸最適化の中にも何らかの形で「凸性」を見つける方向性（NTK, Wasserstein幾何）
- 最適化のダイナミクスと汎化誤差の関係を見出す研究も盛ん (陰的正則化).

今後の方向性：

- 汎化誤差と最適化の良い落としどころ
  - 深層学習では非凸性がカーネル法への優位性を示す鍵
  - どこに“丁度良い”場所はあるのか？

  → 有限次元/無限次元のノイズあり勾配降下法で出てくる対数Sobolev不等式等の“凸的な性質”を利用した大域的最適性と汎化誤差解析
- 深層学習のような大規模データ・高次元モデルの効率的最適化手法は今後も重要：確率的最適化/オンライン最適化，非凸最適化，分散環境最適化，二次最適化法