勾配ランジュバン動力学
平均場ランジュバン動力学

微分方程式

- まずは「微分方程式」から始める.

$$\boxed{\frac{\mathrm{d}x_t}{\mathrm{d}t} = f(x_t)}$$

意味： $x_{t+\Delta t} = x_t + \Delta t \cdot f(x_t) + o(\Delta t)$

$\mathrm{d}x_t = f(x_t)\mathrm{d}t$ とも書く.

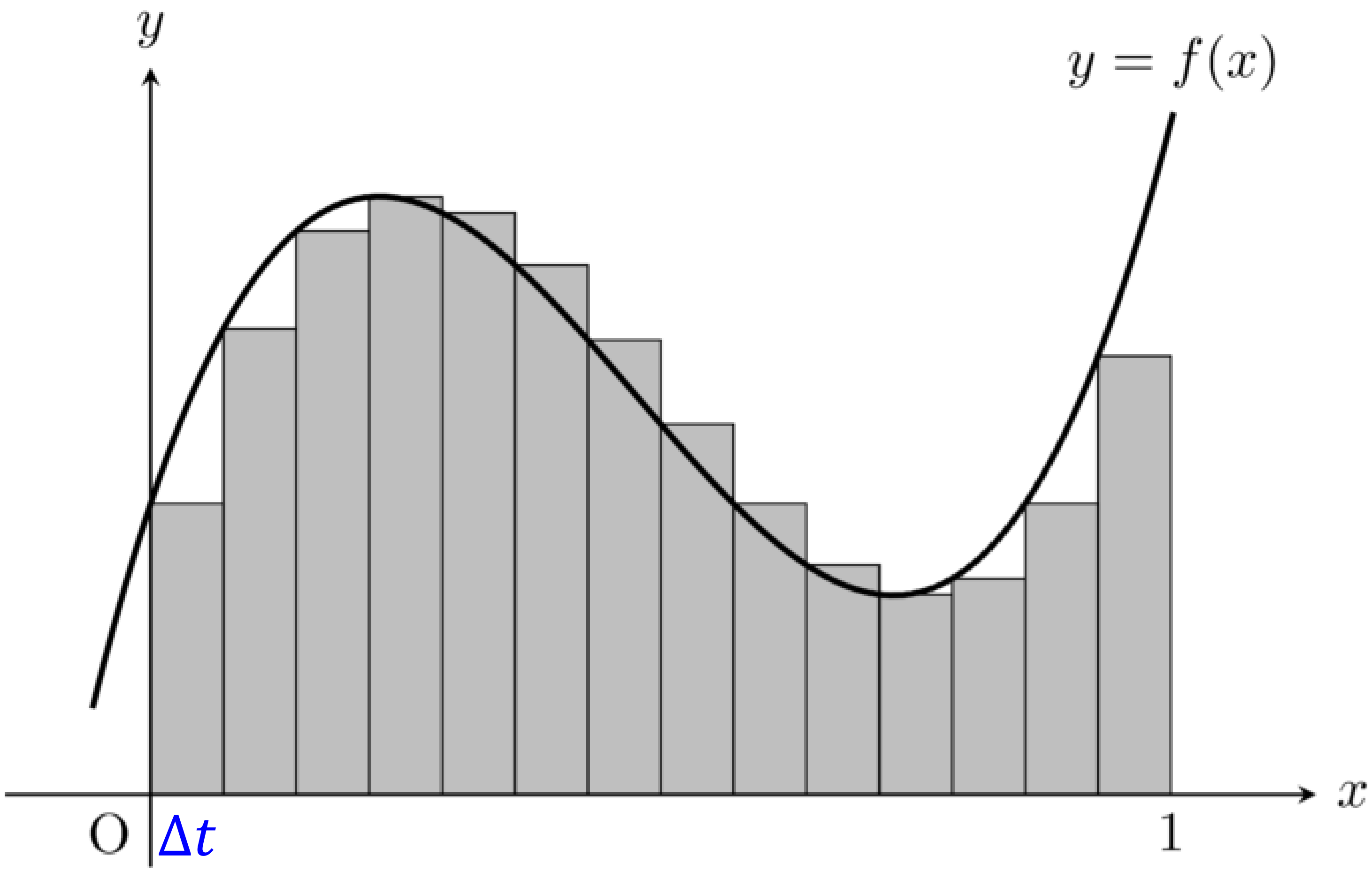
- N回和を取ると ($\Delta t = t/M$)：

$$x_t = x_0 + \frac{t}{M} f(x_0) + \frac{t}{M} f(x_{\Delta t}) + \frac{t}{M} f(x_{2\Delta t}) + \cdots + \frac{t}{M} f(x_{(M-1)\Delta t}) + o(1)$$

- $M \to \infty$とすると：

$$\boxed{x_t = x_0 + \int_0^t f(x_s)\mathrm{d}s}$$

積分表示



[積分の歴史 ～ルベーグ積分までの道のり～; https://wakara.co.jp/mathlog/20200904_2]

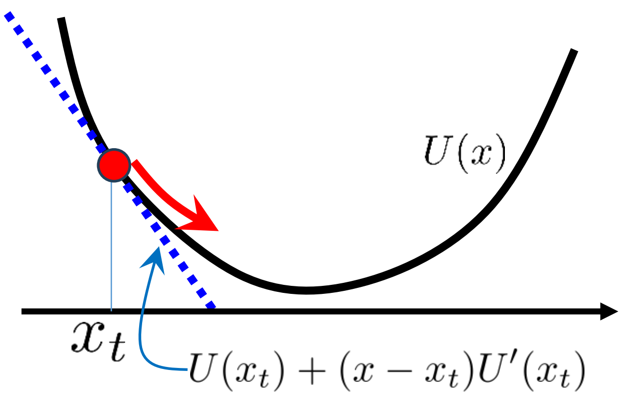
例：勾配流（勾配降下法）

- 関数$U(x)$を最小化したい．

$$f(x) = -\frac{\mathrm{d}U(x)}{\mathrm{d}x} = -U'(x)$$

とする：再急降下方向． ⇒ 「勾配流」

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = -U'(x_t)$$



$$\frac{\mathrm{d}U(x_t)}{\mathrm{d}t} = U'(x_t)\frac{\mathrm{d}x_t}{\mathrm{d}t} = -(U'(x_t))^2$$

勾配が0にならない限り$U(x_t)$は減少し続ける．

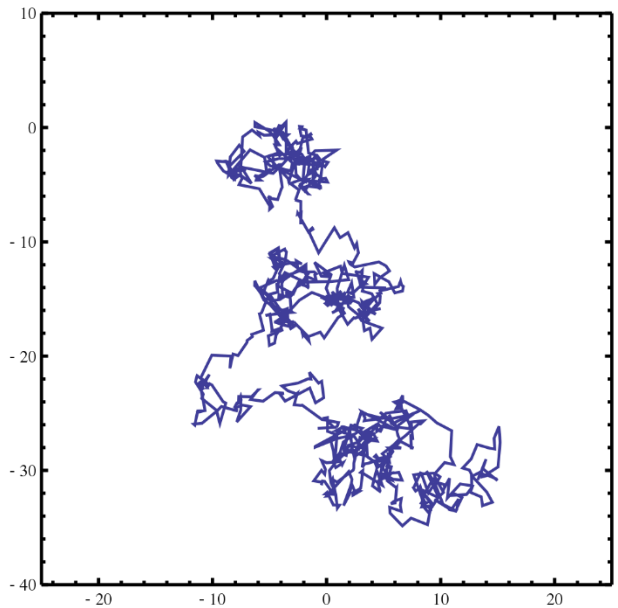
Polyak– Lojasiewicz 条件: $0 \leq U(x) \leq \frac{1}{\alpha}(U'(x))^2$
(PL-条件) $(\alpha > 0)$

例：二次関数
$U(x) = x^2$

$$\frac{\mathrm{d}U(x_t)}{\mathrm{d}t} = -(U'(x_t))^2 \leq -\alpha U(x)$$ ⇒ $$U(x_t) \leq \exp(-\alpha t)U(x_0)$$

線形収束

- ブラウン運動のような確率的にふるまうダイナミクスを記述したい



- 現実世界のダイナミクスにはノイズが含まれているはず．
 （微妙な空気のゆらぎ，電圧の揺れ，観測できない不確実性，…）
 例：ロボットの動作，金融時系列，天気・雲の動き

「確率」微分方程式

- 各更新ステップで”ノイズ”を加える：

$$X_{t+\Delta t} = X_t + \Delta t \cdot f(X_t) + \underline{\sqrt{\Delta t} \cdot \sigma_t \xi_t} \qquad (\Delta t \to 0)$$

(スモールオーダーの項は無視)

σ_t: tにのみ依存する量 (ノイズの大きさを調整)

$$\xi_t \sim N(0,1)$$ **：標準正規分布**

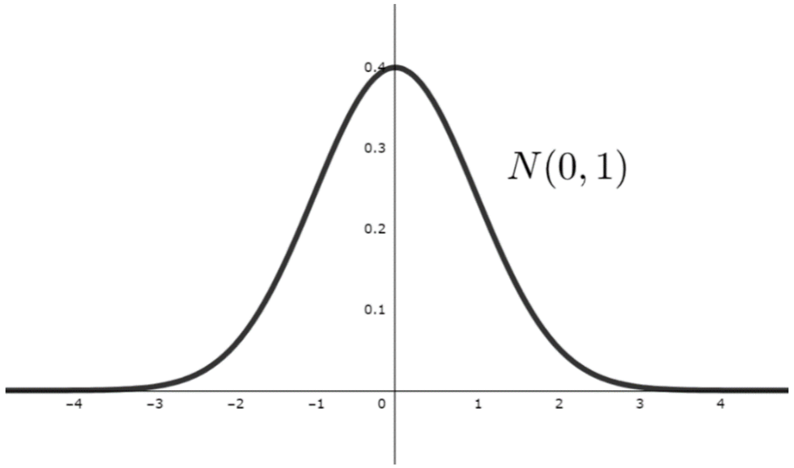
※ $\Delta t \to 0$として何らかの意味で“収束”するかは自明ではない．正確には「伊藤積分」として厳密に定義できる．

<u>$N(0,\sigma^2)$の確率密度関数 (平均0,分散σ^2)：</u>

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

d次元の場合：

$$p(x) = \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right)$$



[数学の景色: https://mathlandscape.com/normal-distrib/]

確率密度関数

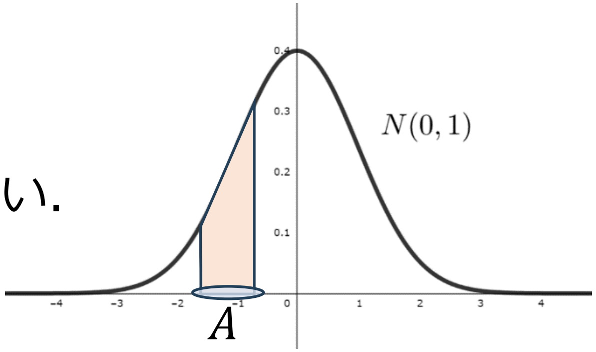
参考

$$P(X \in A) = \int_A p(x)\mathrm{d}x$$

確率変数Xが集合Aに入る確率 = 密度関数のA上での積分

$$\mathbb{E}[X] = \int xp(x)\mathrm{d}x,\ \mathbb{V}[X] = \int (x - \mathbb{E}[X])^2 p(x)\mathrm{d}x,$$

注意：
- 密度関数自体は$X = x$となる確率ではない.
- 積分してはじめて確率になる.
- $\mathbb{R}$全体での積分は１になる.



Fact: 独立な確率変数X, Yの**和 ($Z = X + Y$) の分布**の確率密度は

$$p_Z(z) = \int_{-\infty}^{\infty} p_Y(z - x)p_X(x)\mathrm{d}x$$

で与えられる. （積や和ではないことに注意！）

確認:
$$\int f(z)p_Z(z)\mathrm{d}z = \int f(z)\left(\int_{-\infty}^{\infty} p_Y(z-x)p_X(x)\mathrm{d}x\right)\mathrm{d}z = \int\int f(z)p_Y(z-x)p_X(x)\mathrm{d}x\mathrm{d}z$$
$$= \int\int f(x+y)p_Y(y)p_X(x)\mathrm{d}x\mathrm{d}y$$

$(x, y) \leftarrow (x, z - x)$ の変数変換

正規分布の性質

参考

- 正規分布の分散：

$$X \sim N(0,1) \Rightarrow \sqrt{t}X \sim N(0,t)$$

(標準偏差$\sqrt{t}$ ⇔ 分散t)

確認: $$\int f(\sqrt{t}x)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\mathrm{d}x = \int f(z)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2t}\right)\frac{1}{\sqrt{t}}\mathrm{d}z = \int f(z)\underbrace{\frac{1}{\sqrt{2\pi t}}\exp\left(-\frac{z^2}{2t}\right)}_{N(0,t)\text{の密度関数}}\mathrm{d}z$$

$z = \sqrt{t}x$と変数変換して積分

- 独立な正規分布に従う確率変数の和：

$$X \sim N(0,t),\ Y \sim N(0,s) \ \Rightarrow\ Z = X + Y \sim N(0, t+s)$$

正規分布の再生性

確認: $$\int \frac{1}{\sqrt{2\pi s}}\exp\left(-\frac{(z-x)^2}{2s}\right)\frac{1}{\sqrt{2\pi t}}\exp\left(-\frac{x^2}{2t}\right)\mathrm{d}x$$

$$= \int \frac{1}{\sqrt{(2\pi)^2 st}}\exp\left(-\frac{(x-\frac{zt}{s+t})^2}{2ts/(s+t)} - \frac{z^2}{2(s+t)}\right)\mathrm{d}x$$

$$= \int \frac{1}{\sqrt{2\pi(s+t)}}\exp\left(-\frac{z^2}{2(s+t)}\right)$$

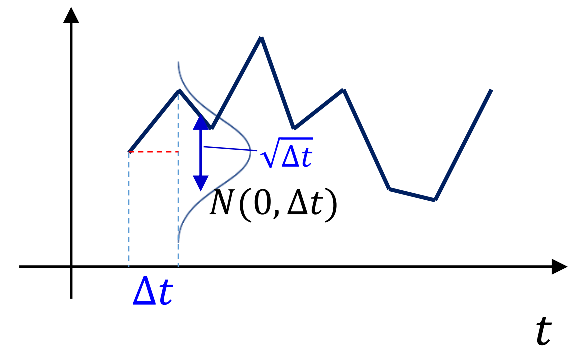
$\leftarrow N(0, s+t)$の密度

ブラウン運動

$$\mathrm{d}X_t = \mathrm{d}B_t$$ ($f(x) = 0, \sigma_t = 1$に対応．$X_0 = 0$とする)

$$X_t \approx \sum_{k=0}^{M-1} \sqrt{\Delta t}\,\xi_k$$ $(\Delta t = t/M)$

$$\xi_k \sim N(0,1)$$



(正規分布の性質より)

⇨

- $X_t \sim N(0, \underbrace{\Delta t + \Delta t + \cdots + \Delta t}_{M個}) = N(0, t)$
- $X_t - X_s \sim N(0, t-s)$

これの$\Delta t \to 0$としたものが「ブラウン運動」
ブラウン運動をB_tと書く．

※$\sqrt{\Delta t}$はちょうど良いスケールである．これより小さいオーダーなら0に収束し，大きいオーダーなら無限大に発散．

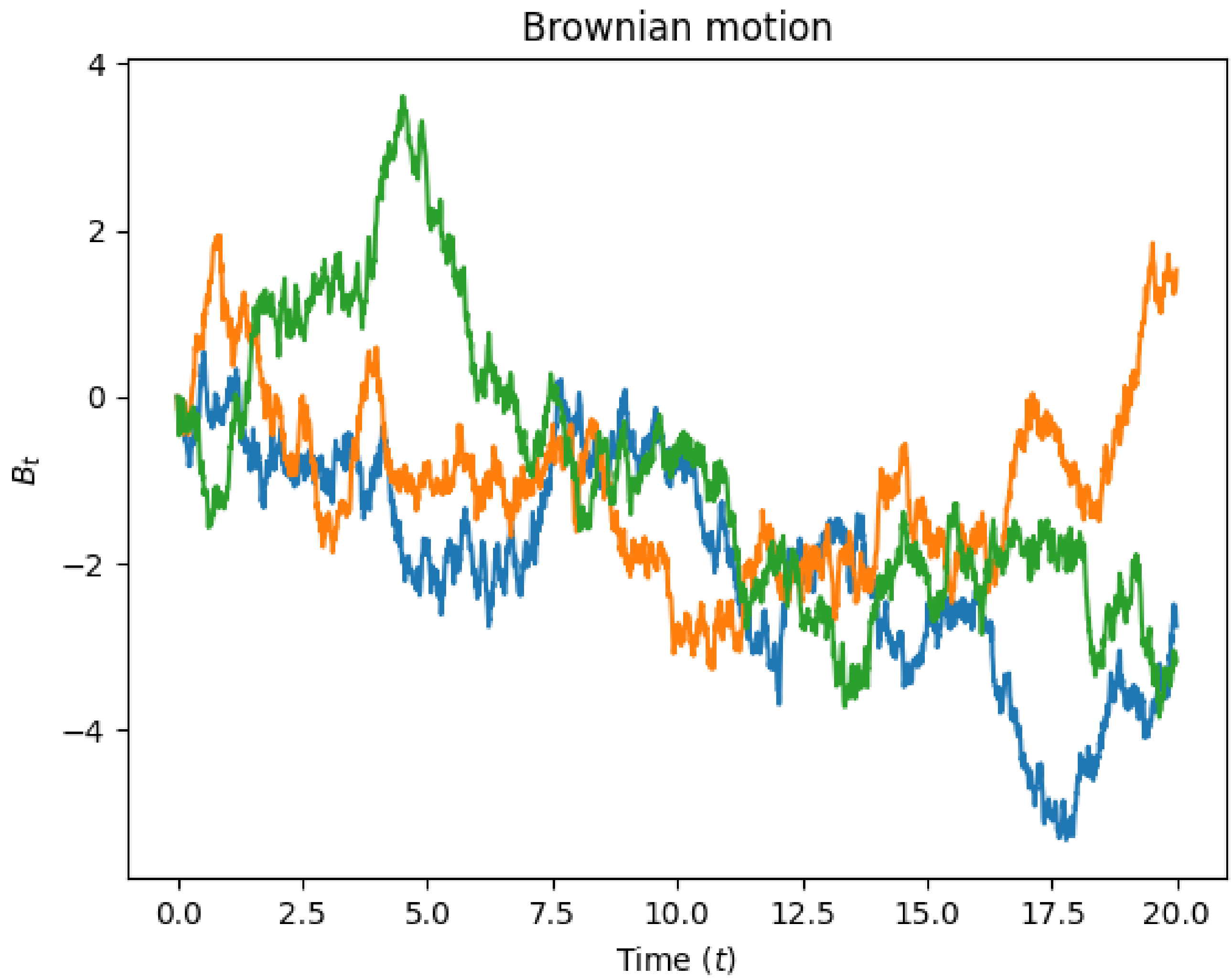
ブラウン運動の性質 (定義)

1. $B_0 = 0$
2. 任意の$0 = t_0 < t_1 < \cdots < t_n$に対して,
$$B_{t_k} - B_{t_{k-1}} \ (k = 1, \ldots, n)$$
は互いに独立
3. 任意の$t > s \geq 0$に対して,
$$B_t - B_s \sim N(0, t - s)$$

さらに, 「4. 標本路$t \mapsto B_t$は確率1で連続」を加えたものがブラウン運動の定義.

特に
$$B_{t+\Delta t} - B_t \sim N(0, \Delta t)$$

つまり
$$B_{t+\Delta t} - B_t = \sqrt{\Delta t}\,\xi_t \quad (\xi_t \sim N(0,1))$$


Brownian motion
4
2
0
−2
−4
B_t
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Time (t)

確率微分方程式

$$X_{t+\Delta t} = X_t + \Delta t \cdot f(x_t) + \sqrt{\Delta t} \cdot \sigma_t \xi_t \quad (\Delta t \to 0)$$

(スモールオーダーの項は無視)

$(\Delta t = t/M)$

$$\Rightarrow \quad X_t = X_0 + \sum_{k=0}^{M-1} \left(f(X_{k\Delta t}) \Delta t + \sigma_{k\Delta t} \sqrt{\Delta t} \xi_{k\Delta t} \right)$$

$$\Rightarrow \quad X_t = X_0 + \sum_{k=0}^{M-1} \left(f(X_{k\Delta t}) \Delta t + \sigma_{k\Delta t} \left(B_{(k+1)\Delta t} - B_{k\Delta t} \right) \right)$$

$\frac{\mathrm{d}B_t}{\mathrm{d}t} \Delta t$ と見做せるが，実際は微分できないので積分で定義.

$(\Delta t \to 0; M \to \infty)$

$$\Rightarrow \quad \boxed{X_t = X_0 + \int_0^t f(X_s) \mathrm{d}s + \int_0^t \sigma_s \mathrm{d}B_s}$$

と書く

$$\boxed{\mathrm{d}X_t = f(X_t) \mathrm{d}t + \sigma_t \mathrm{d}B_t}$$

とも書く

(伊藤積分)

生成作用素

$$\mathrm{d}X_t = v(X_t)\mathrm{d}t + \sigma_t \mathrm{d}B_t$$

p_t: X_tの確率密度関数

期待値: $\mathbb{E}[f(X_t)] = \int f(x)p_t(x)\mathrm{d}x$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[f(X_t)] = \lim_{\Delta t \to 0} \frac{\mathbb{E}[f(X_{t+\Delta})] - \mathbb{E}[f(X_t)]}{\Delta t}$$

ただし, $\xi \sim N(0,1)$でX_tとは独立.

$$= \lim_{\Delta t \to 0} \frac{\mathbb{E}[f(X_t + v_t(X_t)\Delta t + \sigma_t\sqrt{\Delta t}\xi)] - \mathbb{E}[f(X_t)]}{\Delta t}$$

テイラー展開

$$= \lim_{\Delta t \to 0} \frac{\mathbb{E}[\cancel{f(X_t)} + f'(X_t)(v_t(X_t)\Delta t + \sigma_t\sqrt{\Delta t}\xi) + \frac{1}{2}f''(X_t)(v_t(X_t)\Delta t + \sigma_t\sqrt{\Delta t}\xi)^2] - \cancel{\mathbb{E}[f(X_t)]} + o(\Delta t)}{\Delta t}$$

期待値=0

- $\lim_{\Delta t \to 0} \frac{\mathbb{E}[f'(X_t)(v_t(X_t)\cancel{\Delta t} + \cancel{\sigma_t\sqrt{\Delta t}\xi})]}{\cancel{\Delta t}} = \mathbb{E}[f'(X_t)v_t(X_t)]$
- $\lim_{\Delta t \to 0} \frac{\mathbb{E}[\frac{1}{2}f''(X_t)(v_t(X_t)\Delta t + \sigma_t\sqrt{\Delta t}\xi)^2]}{\Delta t}$

1/2 期待値=1

$$= \lim_{\Delta t \to 0} \frac{1}{2\cancel{\Delta t}}\mathbb{E}\left\{[(\Delta t)^{\cancel{2}}v_t + 2v_t\sigma_t(\sqrt{\Delta t})^{\cancel{3/2}}\xi) + \sigma_t^2(\cancel{\Delta t})\xi^2]f''(X_t)\right\} = \frac{\sigma_t^2}{2}\mathbb{E}[f''(X_t)]$$

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[f(X_t)] = \mathbb{E}\left[v_t(X_t)f'(X_t) + \frac{\sigma_t^2}{2}f''(X_t)\right]}$$

Fokker-Planck方程式

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[f(X_t)] = \frac{\mathrm{d}}{\mathrm{d}t}\int p_t(x)f(x)\mathrm{d}x = \int \underline{\frac{\partial p_t(x)}{\partial t}} f(x)\mathrm{d}x$$

でもある.

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[f(X_t)] = \mathbb{E}\left[v_t(X_t)f'(X_t) + \frac{\sigma_t^2}{2}f''(X_t)\right]$$

$$= \int \left(v_t(x)f'(x) + \frac{\sigma_t^2}{2}f''(x)\right)p_t(x)\mathrm{d}x$$

$$\overset{\text{部分積分}}{=} \int \underline{\left[\partial_x(-v_t(x)p_t(x)) + \frac{\sigma_t^2}{2}\partial_x^2 p_t(x)\right]} f(x)\mathrm{d}x$$

Fokker-Planck方程式 [p_tの時間発展を記述した偏微分方程式]：

$$\boxed{\frac{\partial p_t(x)}{\partial t} = -\partial_x(v_t(x)p_t(x)) + \frac{\sigma_t^2}{2}\partial_x^2 p_t(x)}$$

多変量の場合: $\displaystyle \frac{\partial p_t}{\partial t} = -\nabla_x(v_t p_t) + \frac{\sigma_t^2}{2}\mathrm{Tr}[\nabla_x \nabla_x^\top p_t]$

$$\mathcal{L}_t f(x) = v_t(x) f'(x) + \frac{\sigma_t^2}{2} f''(x)$$ ：**生成作用素**

- $$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[f(X_t)] = \mathbb{E}\left[v_t(X_t) f'(X_t) + \frac{\sigma_t^2}{2} f''(X_t)\right] = \mathbb{E}\left[\mathcal{L}_t f(X_t)\right]$$

- $$\frac{\partial p_t(x)}{\partial t} = -\partial_x (v_t(x) p_t(x)) + \frac{\sigma_t^2}{2} \partial_x^2 p_t(x) = \mathcal{L}_t^* p_t(x)$$ ：随伴作用素

部分積分

$$\frac{\mathrm{d}}{\mathrm{d}t}\int f(x) p_t(x) \mathrm{d}x = \int \mathcal{L}_t f(x) p_t(x) \mathrm{d}x = \int f(x) \mathcal{L}_t^* p_t(x) \mathrm{d}x$$

Ornstein–Uhlenbeck 過程 (OU-過程)

$$\mathrm{d}X_t = -X_t\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad X_0 = x_0.$$

$(v_t(x) = -x, \sigma_t = \sqrt{2})$

FP-方程式： $$\frac{\partial p_t(x)}{\partial t} = \partial_x(xp_t(x)) + \partial_x^2 p_t(x)$$

解

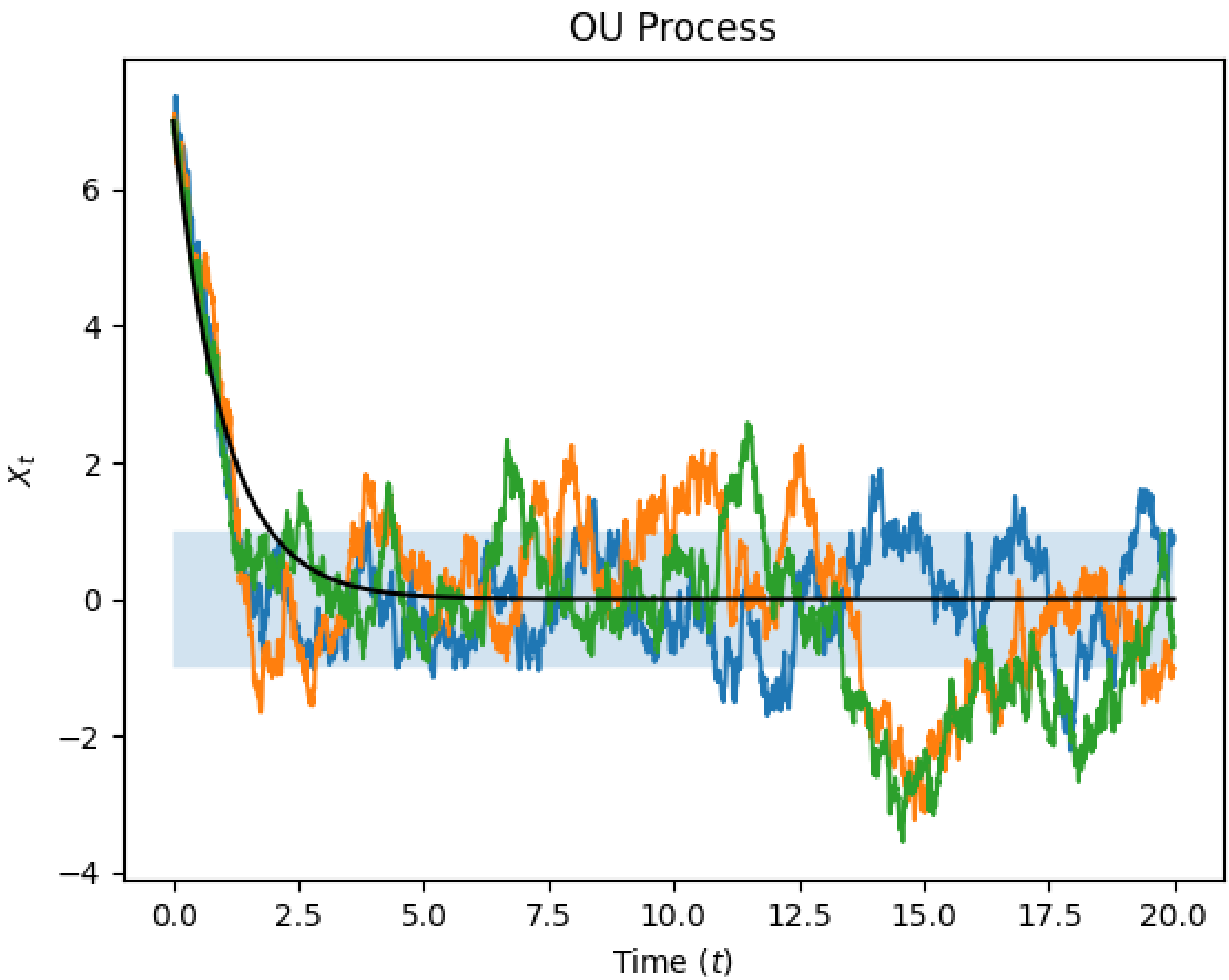
$$p_t(x|X_0 = x_0) = \frac{1}{\sqrt{2\pi(1-e^{-2t})}}\exp\left(-\frac{1}{2}\frac{(x - x_0e^{-t})^2}{1-e^{-2t}}\right)$$

$$= N(x_0e^{-t}, 1-e^{-2t})$$ (平均: x_0e^{-t}, 分散: $1-e^{-2t}$)

($X_0 = x_0$定数ではなく)
$X_0 \sim p_0$の場合: $p_t(x) = \int p_t(x|X_0 = x_0)p_0(x_0)\mathrm{d}x_0$

- 初期値x_0を指数関数的オーダーで忘れていく.
- **指数関数的速さで標準正規分布$N(0,1)$に近づいていく**.

数値実験



一般化：勾配ランジュバン動力学

$$\mathrm{d}X_t = -X_t\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

一般化

$$\Rightarrow \quad \mathrm{d}X_t = -\partial_x U(X_t)\mathrm{d}t + \sqrt{2\lambda}\mathrm{d}B_t$$

勾配ランジュバン動力学

($U(x) = x^2/2\,, \lambda = 1$ならOU-過程)

※ $U(x)$を最小化する勾配流にノイズを加えたもの.

FP-方程式： $$\frac{\partial p_t(x)}{\partial t} = \partial_x(\partial_x U(x)p_t(x)) + \lambda\partial_x^2 p_t(x)$$

定常分布：

$$\boxed{\frac{\partial p_t(x)}{\partial t} = 0 \quad \Rightarrow \quad p^*(x) \propto \exp\left(-\frac{U(x)}{\lambda}\right)}$$

これ以上分布は変化しない

($U(x) = x^2/2,\ \lambda = 1$なら確かに$p^*(x) \propto \exp(-x^2/2) = N(0,1)$の密度関数)

定常分布の導出（確認）

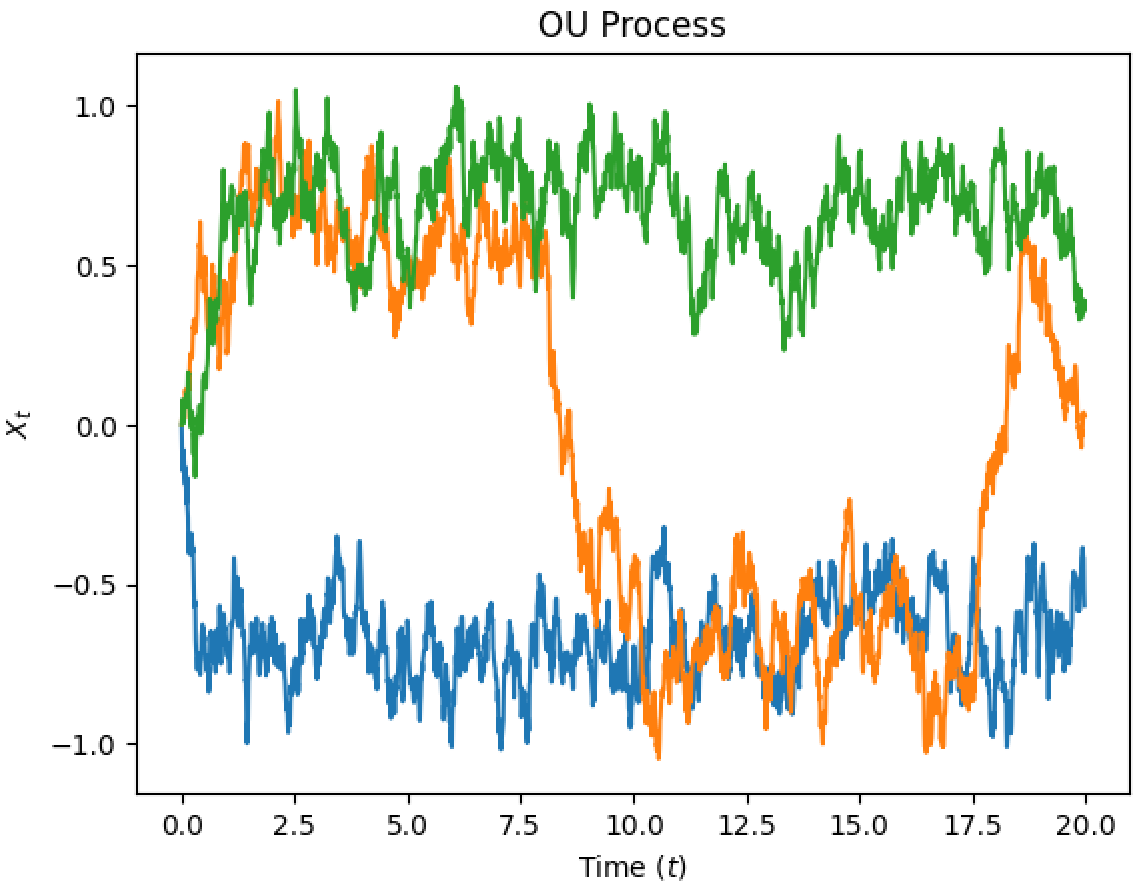
$$
\begin{aligned}
\frac{\partial p_t(x)}{\partial t} &= \partial_x(\partial_x U(x) p_t(x)) + \lambda \partial_x^2 p_t(x) \\
&= \partial_x \left(\partial_x U(x) p_t(x) + \lambda \partial_x p_t(x) \right) \\
&= \partial_x \left[\partial_x \left(\underbrace{U(x) + \lambda \log(p_t(x))}_{=\text{定数}} \right) p_t(x) \right] = 0
\end{aligned}
$$

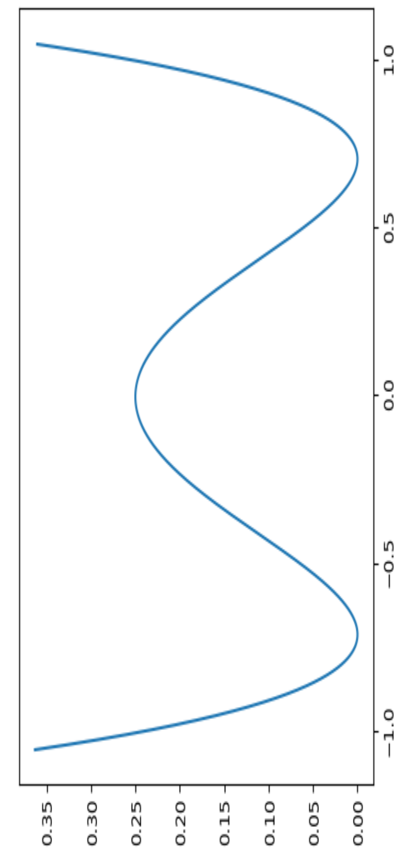
$$\Rightarrow \quad p^*(x) \propto \exp\left(-\frac{U(x)}{\lambda} \right)$$

p^*からサンプリングするためにGLDを走らせれば良い

⇒ マルコフ連鎖モンテカルロ法

$$U(x) = x^4 - x^2, \qquad \lambda = 0.08$$





応用

- ベイズ事後分布からのサンプリング (松田先生の講義を参照)

$$p(\theta|Z_n) = \frac{\prod_{i=1}^n p(z_i|\theta)\pi(\theta)}{C}$$

$$\propto \exp\left(-\sum_{i=1}^n \ell_i(\theta) - R(\theta)\right)$$

一般に分布の形は複雑になる．直接サンプリングは難しい．

$\ell_i(\theta) = -\log(p(z_i|\theta))$: 負の対数尤度（データへの当てはまり）

$R(\theta) = -\log(\pi(\theta))$: 事前分布の負の対数（事前的不確実性）

$$U(\theta) = \sum_{i=1}^n \ell_i(\theta) + R(\theta)$$

- 拡散モデル (次回講義)

収束解析

勾配ランジュバン動力学はどれくらいのスピードで定常分布に近づく？
- 分布間の「近さ」とは？

- **KL-divergence** (☆重要)

$$\mathrm{KL}(p||q) = \int \log\left(\frac{p(x)}{q(x)}\right) p(x)\mathrm{d}x$$

- **Fisher-divergence**

$$I(p||q) = \int \|\nabla \log(p(x)) - \nabla \log(q(x))\|^2 p(x)\mathrm{d}x$$

※どちらも非負で，$p = q$の時のみ0になる.

- 統計学・情報理論・機械学習で頻出.
  - 最尤推定量はデータ生成分布からのKL-divを近似的に最小化
- KL-divergenceは相対エントロピーとも言われる.

$$\frac{\partial p_t(x)}{\partial t} = \partial_x \left[\partial_x \left(\underline{U(x)} + \lambda \log(p_t(x))\right) \cdot p_t(x)\right]$$

$$= \lambda \partial_x \left[\partial_x \left(\underline{- \log(p^*(x))} + \log(p_t(x))\right) \cdot p_t(x)\right]$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(p_t || p^*) = \frac{\mathrm{d}}{\mathrm{d}t} \int \log\left(\frac{p_t(x)}{p^*(x)}\right) p_t(x) \mathrm{d}x$$

$$= \partial_t \int p_t \mathrm{d}x = \partial_t 1 = 0$$

$$= \int \log\left(\frac{p_t(x)}{p^*(x)}\right) \partial_t p_t(x) \mathrm{d}x + \int \frac{\partial_t p_t(x)}{\cancel{p_t(x)}} \cancel{p_t(x)} \mathrm{d}x$$

$$= \int \log\left(\frac{p_t(x)}{p^*(x)}\right) \cdot \lambda \partial_x \left[\partial_x \log\left(\frac{p_t(x)}{p^*(x)}\right) \cdot p_t(x)\right] \mathrm{d}x$$

部分積分

$$= -\lambda \int \left|\partial_x \log\left(\frac{p_t(x)}{p^*(x)}\right)\right|^2 p_t(x) \mathrm{d}x$$

$$= -\lambda I(p_t || p^*)$$

⇒ Fisher-divが0にならない限り，KL-divは減り続ける．

ガウシアン対数Sobolev不等式

- OU-過程で収束を示してみよう.

$$p^*(x) \propto \exp(-x^2/2)$$ (標準正規分布)

$$(U(x) = x^2/2,\ \lambda = 1)$$

勾配流のところで出た**PL-条件**に対応

定理 (ガウシアン対数ソボレフ不等式)

$p^*(x) \propto \exp(-x^2/2)$とする (標準正規分布).
任意の確率密度関数pに対して，次の不等式が成り立つ:

$$\mathrm{KL}(p||p^*) \leq \frac{1}{2} I(p||p^*)$$

よって，
$$\begin{aligned}\partial_t \mathrm{KL}(p_t||p^*) &= -I(p_t||p^*) \\ &\leq -2\mathrm{KL}(p_t||p^*).\end{aligned}$$

⇨ $$\mathrm{KL}(p_t||p^*) \leq \exp(-2t)\mathrm{KL}(p_0||p^*)$$

線形収束！

ガウシアン対数ソボレフ不等式の証明

参考

証明の方法は何通りもある．ここでは，半群を用いた方法で示す．

$\mathcal{L}f(x) := -xf'(x) + f''(x)$：OU-過程の生成作用素

$P_t f(x) := \mathbb{E}[f(X_t)|X_0 = x]$

性質 $\begin{cases} \bullet \ \partial_t P_t f = P_t \mathcal{L} f & \text{(生成作用素の性質)} \\ \bullet \ P_t P_s f = P_{t+s} f & \text{(半群性)} \end{cases}$

$p_t(\cdot\,|X_0 = x)$の形より，

$$(P_t f)' = e^{-t} P_t(f')$$

特に，両辺絶対値を取って，

$$|(P_t f)'| = e^{-t}|P_t(f')| \leq e^{-t} P_t(|f'|) \quad \cdots（1）$$

今，$\psi(r) = r\log(r)$に対して，

$$\Lambda(s) := P_s(\psi(P_{t-s} f)) \quad (s \in [0, t])$$

とする．

(証明続き)

参考

すると，生成作用素の性質より，$g = P_{t-s}f$ に対して，

$$\begin{aligned}\Lambda'(s) &= P_s(\mathcal{L}\psi(g) - \psi'(g)\mathcal{L}g) \\ &= P_s(\psi''(g)(g')^2) \\ &= P_s\left(\frac{|g'|^2}{g}\right) \qquad \cdots (2)\end{aligned}$$

今，前ページの式(1)より，

$$|g'|^2 = |(P_{t-s}f)'|^2 \leq e^{-2(t-s)}(P_{t-s}(|f'|))^2$$

であるが，コーシーシュワルツの不等式からさらに右辺は

$$(P_{t-s}(|f'|))^2 \leq P_{t-s}fP_{t-s}\left(\frac{|f'|^2}{f}\right) = gP_{t-s}\left(\frac{|f'|^2}{f}\right)$$

と抑えられるので，式(2)の右辺は次のように抑えられる:

$$\Lambda'(s) \leq e^{-2(t-s)}P_s\left(P_{t-s}\left(\frac{|f'|^2}{f}\right)\right) \underset{\text{(半群性)}}{=} e^{-2(t-s)}P_t\left(\frac{|f'|^2}{f}\right)$$

(証明続き)

参考

よって，両辺をsに関して$[0,t]$の間で積分すると，

$$\Lambda(t)-\Lambda(0)=P_t(f\log(f))-P_tf\log(P_tf)\leq\frac{1-e^{-2t}}{2}P_t\left(\frac{|f'|^2}{f}\right)$$

を得る．

$p_t(\cdot\,|X_0=x)$の形から，(適当な可積分性のもと)

$$\lim_{t\to\infty}P_tf(x)\to\mathbb{E}_{X\sim p^*}[f(X)]\quad(\forall x\in\mathbb{R})$$

なので，両辺$t\to\infty$とすると，

$$\mathbb{E}_{p^*}[f\log(f)]-\mathbb{E}_{p^*}[f]\log(\mathbb{E}_{p^*}[f])\leq\frac{1}{2}\mathbb{E}_{p^*}\left[\frac{|f'|^2}{f}\right]$$

を得る．

最後に，$f=\dfrac{p}{p^*}$ を代入すれば，$\mathrm{KL}(p||p^*)\leq\frac{1}{2}I(p||p^*)$を得る．

$\boxed{\text{証明終}}$

参考文献

確率積分および確率微分方程式については，数多くの良書があるが，たとえば，以下の書籍は参考になる:

- エクセンダール: 確率微分方程式 (日本語訳版)．丸善出版, 2012.
- 舟木：確率微分方程式．岩波書店，2005.
- Kuo: *Introduction to Stochastic Integration*. Springer, 2005.

収束解析に関する理論は，以下の本が有用である:

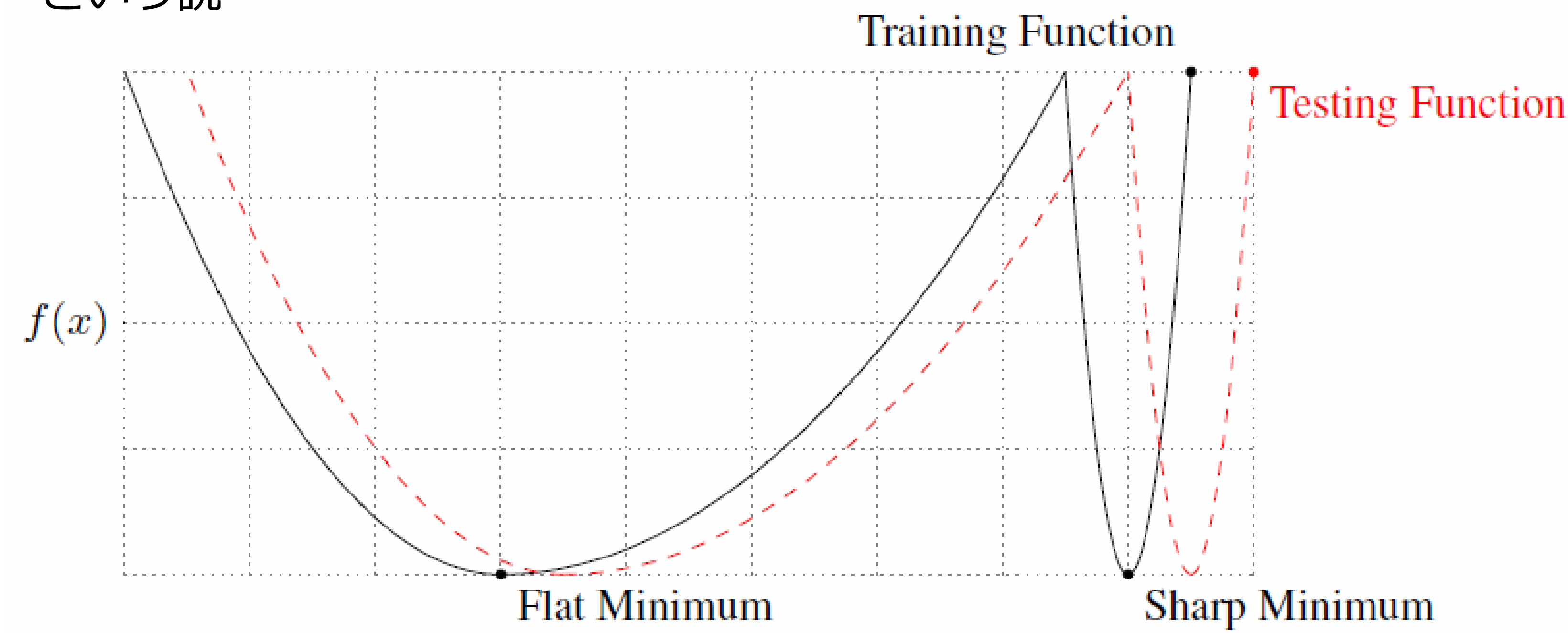
- Bakry, Gentil and Ledoux: *Analysis and Geometry of Markov Diffusion Operators*. Springer, 2014.

特に，ガウシアン対数ソボレフ不等式はこの本を参考にした．証明の元ネタは以下の論文による:

- Bakry and Emery: Diffusions hypercontractives. *Seminaire de Probabilities*, XIX, 1983/1984. Lecture Notes in Mathematics, vol. 1123, pp. 177—206, Springer.

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



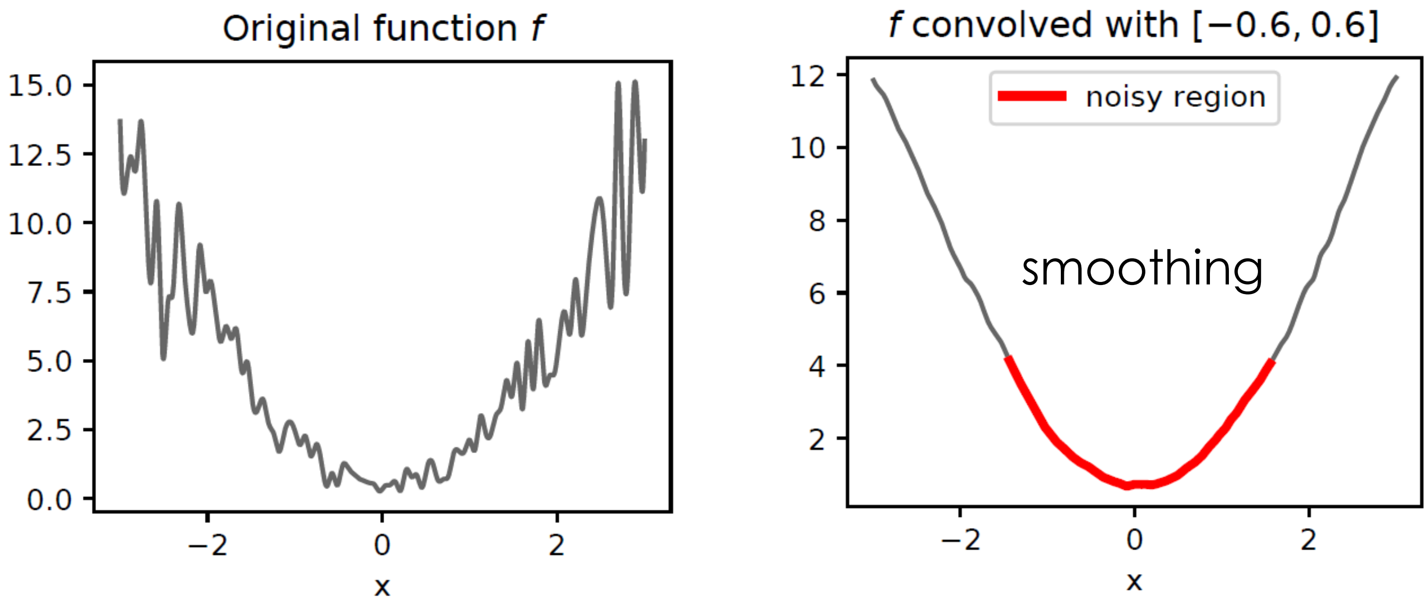
Keskar, Mudigere, Nocedal, Smelyanskiy, Tang (2017):
On large-batch training for deep learning: generalization gap and sharp minima.

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

→ 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.

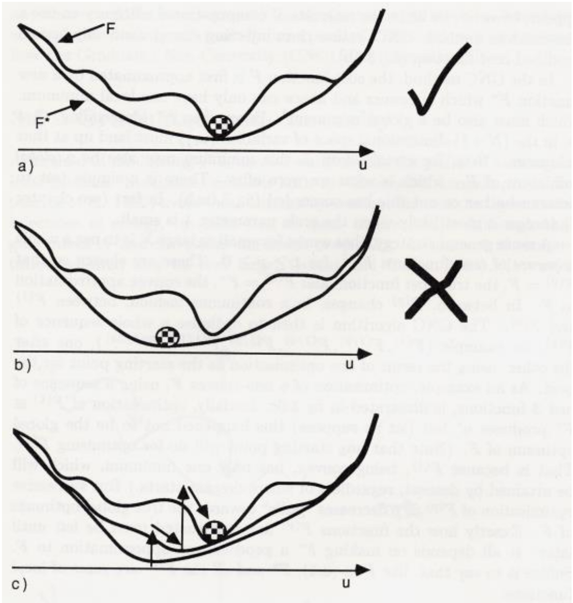
- Graduated optimization

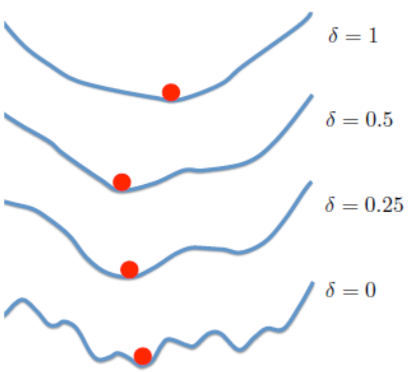
Hazan, Levy, and Shalev-Shwartz: On graduated optimization for stochastic non-convex problems. *International conference on machine learning*, pp. 1833-1841, 2016.

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

- 勾配ランジュバン動力学 (多次元版)

  Gradient Langevin Dynamics (GLD)

サンプリング：

$$\mu^*(x) \propto \exp\left(-\frac{L(x)}{\lambda}\right)$$

なる分布からサンプリングしたい.

非凸最適化： μ^*からのサンプリングは$\min_x L(x)$を近似的に解くことも出来る.

λ: 温度パラメータ

$$\mathrm{d}X_t = -\nabla L(X_t)\mathrm{d}t + \sqrt{2\lambda}\mathrm{d}B_t$$ (勾配ランジュバン動力学)

定常分布： $\pi \propto \exp(-\lambda^{-1}L(X))$

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

例

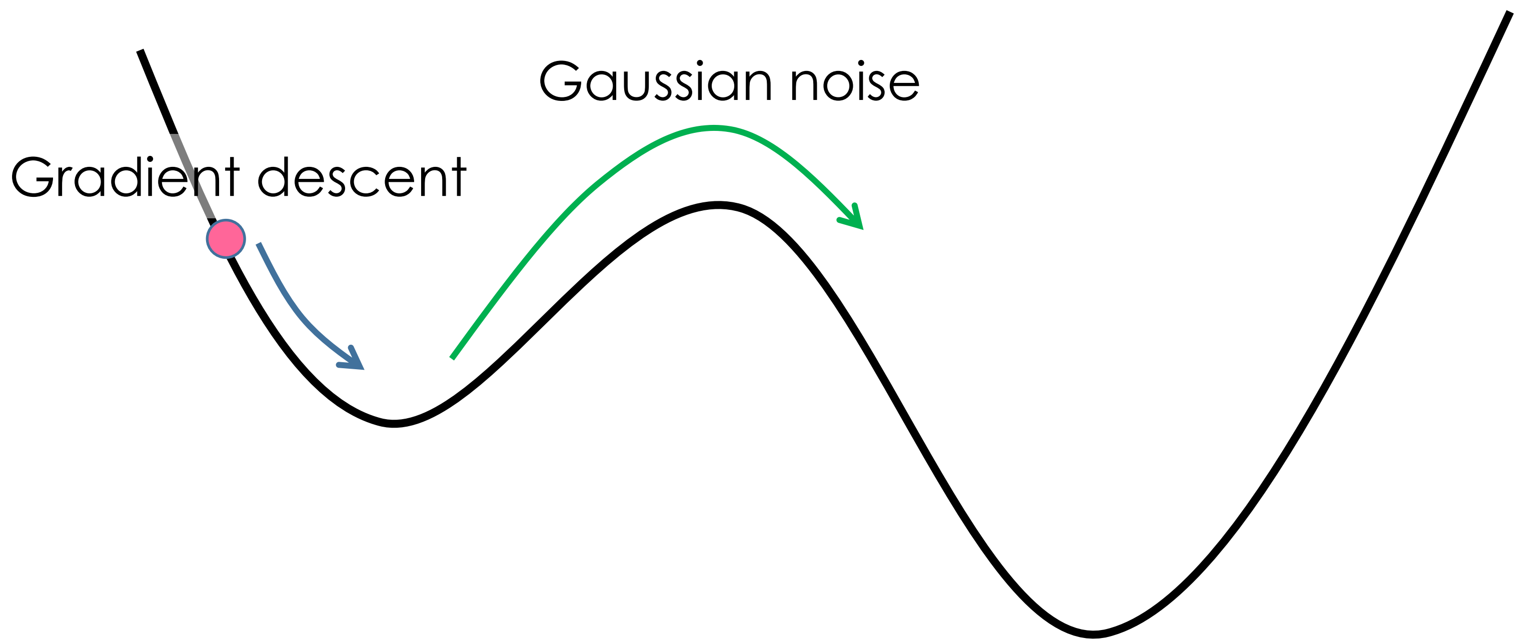
最尤推定：

$$L(x) = \sum_{i=1}^{n} -\log(p_x(z_i))$$

正則化学習：

$$L(x) = \sum_{i=1}^{n} \ell_i(x) + \lambda_1 \|x\|^2$$

$$\begin{cases} \ell_i(x) = (y_i - f_x(z_i))^2 & \text{回帰 } (y_i \in \mathbb{R}, z_i \in \mathbb{R}^d) \\ \ell_i(x) = \log(1 + \exp(-y_i f_x(z_i))) & \text{判別 } (y_i \in \{\pm 1\}, z_i \in \mathbb{R}^d) \end{cases}$$


Gaussian noise
Gradient descent

GLDのFokker-Planck方程式

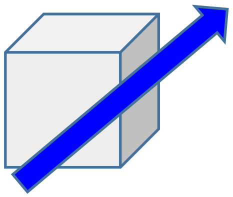
$$\mathrm{d}X_t = -\nabla L(X_t)\mathrm{d}t + \sqrt{2\lambda}\mathrm{d}B_t$$

μ_t ：X_tの分布の確率密度関数

Fokker-Planck方程式

$$\nabla \cdot [\mu_t \nabla L] = \sum_{j=1}^{d} \partial_i [\mu_t \partial_i L]$$

$$\partial_t \mu_t = \lambda \Delta_x \mu_t + \nabla \cdot [\mu_t \nabla L]$$

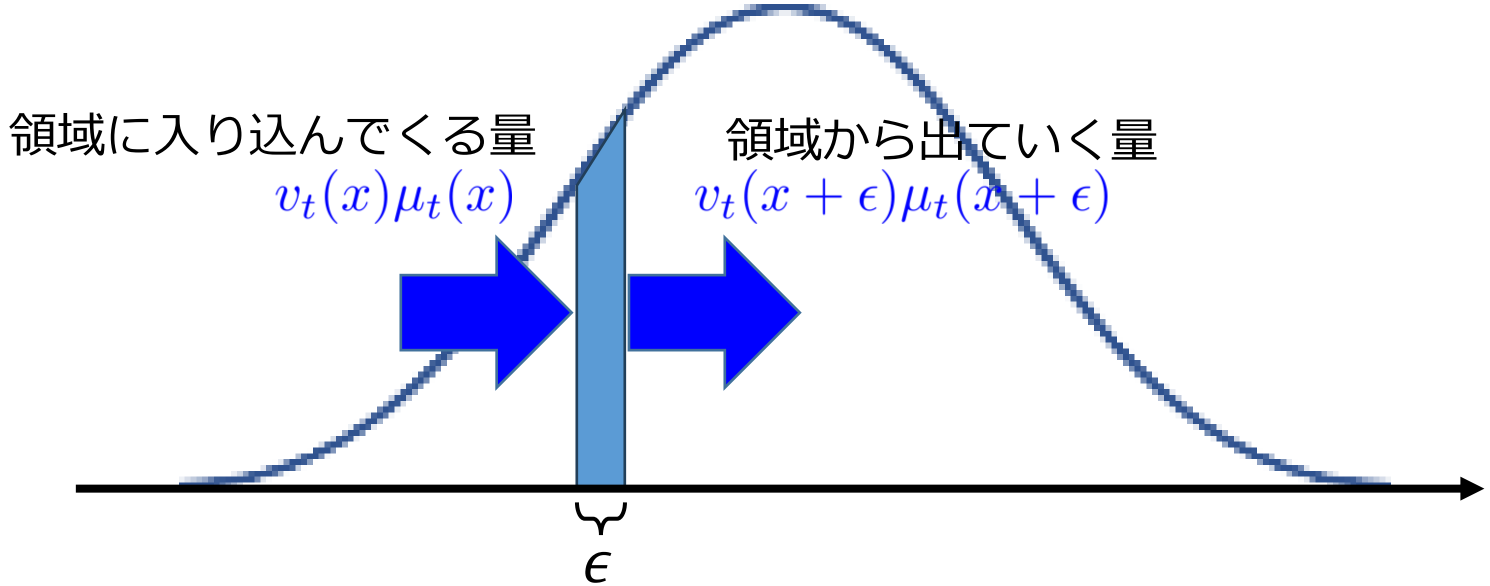


次のように解釈できる:

$$\partial_t \mu_t = \nabla \cdot [\underbrace{(\lambda \nabla \log(\mu_t) + \nabla L)}_{-v_t\text{とおく}} \mu_t] = -\nabla \cdot [v_t \mu_t]$$

[連続の方程式]

連続の方程式の意味



差分（「入り込む量」–「出ていく量」） = $-\nabla \cdot [\mu_t(x) v_t(x)]\epsilon$

連続の方程式

「連続の方程式」 $\dfrac{\partial \mu_t}{\partial t} = -\nabla \cdot (v_t \mu_t)$

この方程式の意味

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

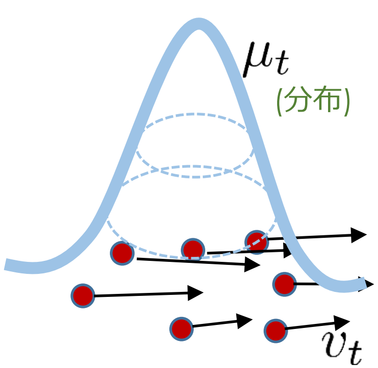
$$\left(= -\int f(x)\nabla\cdot(v_t\rho_t)\mathrm{d}x\right)$$

($\forall f$: コンパクトサポート，C^∞-級)

- ベクトル場v_tで生成される写像をT_tとする: $\frac{\mathrm{d}T_t}{\mathrm{d}t}(x) = v_t(T_t(x))$.
- μ_tは写像$T_t: R^d \to R^d$によるμ_0の押し出し：$\mu_t = T_{t\#}\mu_0$.
  つまり，$x \sim \mu_0$**に対する**$T_t(x)$**の分布が**μ_t.

(一般のt)
$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int f(x)\mu_t(x)\mathrm{d}x &= \frac{\mathrm{d}}{\mathrm{d}t}\int f(T_t(x))\mu_0(x)\mathrm{d}x \\ &= \int \nabla f(T_t(x))^\top \frac{\mathrm{d}T_t(x)}{\mathrm{d}t}\mu_0(x)\mathrm{d}x \\ &= \int \nabla f(T_t(x))^\top v_t(T_t(x))\mu_0(x)\mathrm{d}x \\ &= \int \nabla f(x)^\top v_t(x)\mu_t(x)\mathrm{d}x.\end{aligned}$$

[連続の方程式]



Wasserstein距離について

参考

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

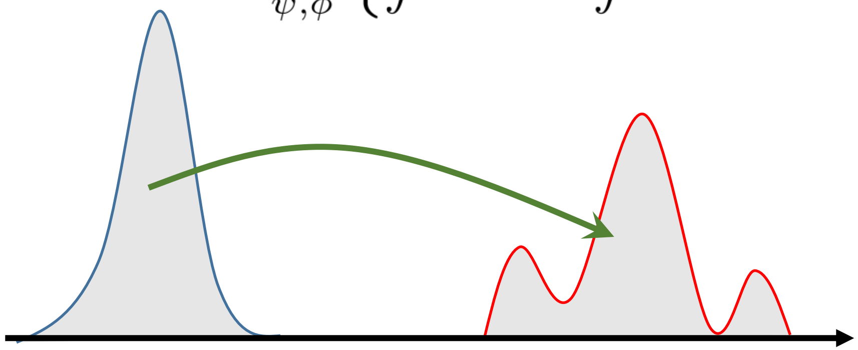
$$(\mathcal{X} = \mathbb{R}^d:\ c(x,y) = \|x-y\|)$$

- 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)

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



「輸送距離」とも言われる

接ベクトル

参考

定理

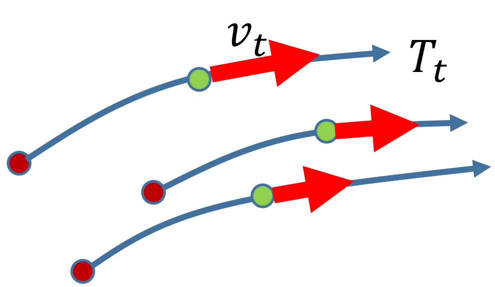
- $\rho_t = T_{t\#}\rho_0$
- $\frac{\mathrm{d}T_t}{\mathrm{d}t}(w) = v_t\big(T_t(w)\big)$
- あるϕ_tを用いて$v_t = \nabla\phi_t$と書けるとする.

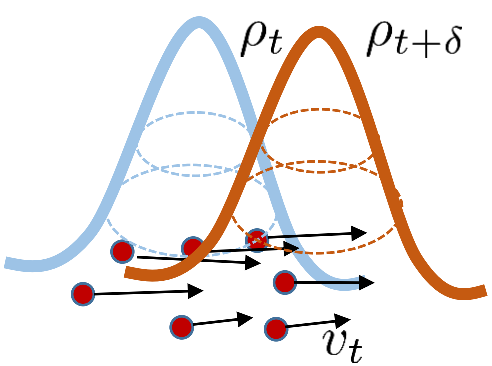
この時，以下が成り立つ:

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

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

参考

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.

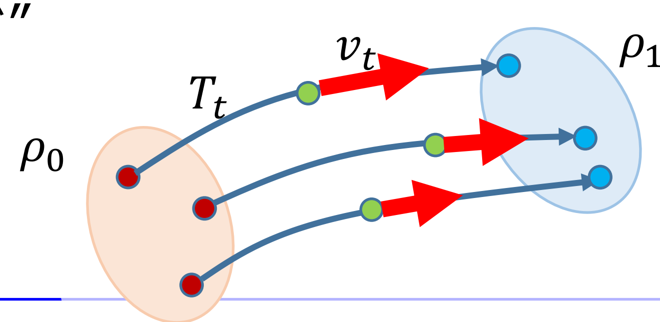
Benamou-Brenier formula (連続の方程式とW_2距離の関係):

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.

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



定常分布

$$\partial_t \mu_t = \nabla \cdot [(\lambda \nabla \log(\mu_t) + \nabla L)\, \mu_t] \;= -\nabla \cdot [v_t \mu_t]$$

定常分布: $\partial_t \mu_t = 0 \Rightarrow v_t = 0$ (分布がこれ以上動かない)

$$\boxed{\lambda \nabla \log(\mu^*) + \nabla L = 0 \quad \Rightarrow \quad \mu^*(x) \propto \exp(-L(x))}$$

実は，これは以下の目的関数を最小化するWasserstein勾配流である:

$$\mathrm{Ent}(\mu) := \int \log(\mu)\mu \mathrm{d}x$$

$$\mu^* = \mathop{\arg\min}_{\mu \in \mathcal{P}} \underbrace{\int L(x)\mathrm{d}\mu(x)}_{L\text{を最小化}} + \underbrace{\lambda \mathrm{Ent}(\mu)}_{\text{ガウスノイズによって分布を拡散させる力}} =: \mathcal{L}(\mu)$$

➡ 確かにこの最適解は定常分布と等しい:

$$\mu^*(x) \propto \exp(-\lambda^{-1} L(x))$$

Wasserstein勾配流

$$\lambda^{-1}\mathcal{L}(\mu) = \int \lambda^{-1} L(x)\mathrm{d}\mu(x) + \mathrm{Ent}(\mu)$$

$\boxed{\mu^*(x) \propto \exp(-\lambda^{-1}L(x))}$

$$= \int -\log(\mu^*)\mathrm{d}\mu + \int \log(\mu)\mathrm{d}\mu + (\mathrm{const.})$$

以下，無視

$$= \int \log\left(\frac{\mu}{\mu^*}\right)\mathrm{d}\mu = \mathrm{KL}(\mu||\mu^*)$$

連続の方程式 $\mu_t = -\nabla \cdot [v_t \mu_t]$ に従っているなら

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\mu_t||\mu^*) = \frac{\mathrm{d}}{\mathrm{d}t}\int \log\left(\frac{\mu_t(x)}{\mu^*(x)}\right)\mu_t(x)\mathrm{d}x$$

$$= \int \log\left(\frac{\mu_t(x)}{\mu^*(x)}\right)\partial_t\mu_t(x)\mathrm{d}x + \int \frac{\partial_t\mu_t(x)}{\cancel{\mu_t(x)}}\cancel{\mu_t(x)}\mathrm{d}x$$

$$= \int \log\left(\frac{\mu_t(x)}{\mu^*(x)}\right)\nabla\cdot(-v_t\mu_t(x))\mathrm{d}x$$

$$= -\int \langle v_t, \nabla\log(\mu^*) - \nabla\log(\mu_t)\rangle \mu_t(x)\mathrm{d}x$$

Wasserstein勾配流

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\mu_t||\mu^*) = -\int \langle v_t, \nabla\log(\mu^*) - \nabla\log(\mu_t)\rangle \mu_t(x)\mathrm{d}x$$

特に

$$v_t = -\left(\lambda\nabla\log(\mu_t) + \nabla L\right) = \lambda\left(\nabla\log(\mu^*) - \nabla\log(\mu_t)\right) \qquad \text{(GLD)}$$

は<u>最急降下方向</u>で，以下が成り立つ．

$$\partial_t\mu_t = \nabla\cdot\left[\underbrace{\left(\lambda\nabla\log(\mu_t) + \nabla L\right)}_{=:\,-v_t}\mu_t\right]$$

$$\boxed{\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\mu_t||\mu^*) &= -\lambda\int \|\nabla\log(\mu^*) - \nabla\log(\mu_t)\|^2\mu_t\mathrm{d}x \\ &= -\lambda I(\mu_t||\mu^*)\end{aligned}}$$

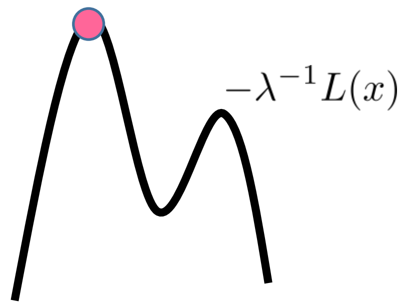
定常分布μ^*からのKL-divを最小化するWasserstein勾配流

Fisher divergence:

$$I(\mu||\nu) := \int \|\nabla\log(\nu) - \nabla\log(\mu)\|^2\mu(x)\mathrm{d}x$$

対数ソボレフ不等式と幾何的エルゴード性

定常分布 $\mu^*(x) \propto \exp(-\lambda^{-1}L(x))$



定義 (対数ソボレフ不等式 (μ^*の性質))

あるα > 0が存在して,
任意の(μ^*に対して絶対連続な)確率分布νに対し,

$$\mathrm{KL}(\nu||\mu^*) \leq \frac{1}{2\alpha} I(\nu||\mu^*)$$

例：
- 二次関数+有界関数
- Weak Morse型関数

KL-div　Fisher-div

$$\mathrm{KL}(\nu||\mu) = \int \log\left(\frac{\nu}{\mu}\right)\mu \mathrm{d}x, \quad I(\nu||\mu) = \int \left\|\nabla \log \frac{\nu}{\mu}\right\|^2 \nu \mathrm{d}x$$

➡ **幾何的エルゴード性**　μ_t: X_tの周辺分布

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\mu_t||\mu^*) = -\lambda I(\mu_t||\mu^*) \leq -2\alpha \mathrm{KL}(\mu_t||\mu^*)$$ (対数ソボレフより)

$$\mathrm{KL}(\mu_t||\mu^*) \leq \exp(-2\alpha t)\mathrm{KL}(\mu_0||\mu^*)$$

定常分布へKL-divergenceの意味で線形収束

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

対数ソボレフ不等式の十分条件

強凸な場合 (Bakry-Emery規準): $\mu^*(x) \propto \exp(-\lambda^{-1}L(x))$

$$\nabla\nabla^\top L(x) \succcurlyeq \mu I \quad \Rightarrow \quad \alpha \geq \mu/\lambda$$

[Bakry and Émery, 1985]

例：OU-過程．$L(x) = \frac{x^2}{2}$なので，$\mu = 1$で成り立つ．

(正定値対称行列)　「$\nabla\nabla^\top L(x) \succcurlyeq \mu I$」 $\Leftrightarrow$ 「$u^\top \nabla\nabla^\top L(x) u \succcurlyeq \mu \|u\|^2$」

Bounded perturbation lemma (Hollley-Stroock):

$\mu^*(x) = \mu(x) \exp(h(x))$ と書けて，μがα'-LSI条件を満たすとする．

$|h(x)| \leq B \ (\forall x)$ ➡ μのα-LSI定数は以下を満たす:

$$\alpha \geq \alpha' \exp(-4B)$$

[R. Holley and D. Stroock. Logarithmic sobolev inequalities and stochastic Ising models. Journal of statistical physics, 46(5-6):1159–1194, 1987.]

例：$L(x) = \ell(x) + \lambda_1 x^2$で$|\ell(x)| \leq B$なら，$\mu^*$は$\alpha = \frac{2\lambda_1}{\lambda}\exp(-4B/\lambda)$でLSI条件を満たす．

$\mu^*(x) \propto \exp(-\lambda^{-1}L(x))$

離散時間ダイナミクスの収束レート

過程: L はM-平滑: $\exists M, \ \|\nabla L(x) - \nabla L(y)\| \leq M\|x-y\|$

定理

[Vempala and Wibisono, 2019]

ν_k: Marginal distribution of X_k (discrete time dynamics)　($\lambda = 1$としている)

$$D(\nu_k||\pi_\infty) \leq \exp(-\alpha k\eta) D(\nu_0||\pi_\infty) + 8\frac{dM^2\eta}{\alpha}$$

定理 (informal)

散逸性と平滑性の条件のもと (and other technical condition),

$$\mathrm{E}[L(X_k)] - L(X^*) \lesssim \exp(-c_\lambda \alpha k\eta) + c_{c_\alpha,\lambda,d}\eta + \lambda d \log(\lambda^{-1}+1)$$

幾何的エルゴード性　時間離散化の誤差　$E_{\pi_\infty}[L(X)] - L(X^*)$

定常分布が最適解まわりにどれだけ集中しているか

where $c, c_{C_{LS},\beta,d} > 0$ are constants.

[Raginsky, Rakhlin and Telgarsky, 2017; Xu, Chen, Zou, and Gu, 2018; Erdogdu, Mackey and Shamir, 2018]

- 温度パラメータλが十分小さければ，目的関数が非凸でも最適解の近くに到達できる.
- ただし，一般には対数ソボレフ不等式はλ^{-1}に指数的に依存することに注意.

（そうでない場合もある：強凸目的関数，Weak Morse関数）

時間離散化 + 確率的分散縮小勾配法

研究紹介

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

定常分布: $\frac{\mathrm{d}\nu}{\mathrm{d}x}(x) \propto \exp(-\gamma f(x))$

連続時間SDE: $\mathrm{d}X_t = -\nabla f(X_t)\mathrm{d}t + \sqrt{2/\gamma}\mathrm{d}B_t$

離散時間近似 (Euler-Maruyama近似):

$$X_{k+1} = X_k - \eta \underline{\nabla f(X_k)} + \sqrt{2\eta/\gamma}\xi_k$$

計算にO(n)かかる (大規模データで困る)
→ 確率的勾配を用いる $\tilde{\nabla}_k = \frac{1}{B}\sum_{i\in I_k} \nabla f_i(X_k)$

- 全勾配は計算に時間がかかる→確率的勾配を用いる.
- 確率的勾配は分散が大きい→分散縮小法(SVRG,SARAH)と組み合わせる.

分散縮小型確率的勾配:

SVRG: $\tilde{\nabla}_k = \frac{1}{B}\sum_{i\in I_k}(\nabla f_i(X_k) - \nabla f_i(\tilde{X}) + \nabla f(\tilde{X}))$

SARAH: $\tilde{\nabla}_k = \frac{1}{B}\sum_{i\in I_k}(\nabla f_i(X_k) - \nabla f_i(X_{k-1}) + \tilde{\nabla}_{k-1})$

※ $\tilde{X}, \tilde{\nabla}_k$はm回に一回更新する. ($m = \sqrt{n}$でOK)

GLDはノイズを加えつつ最適化するので，分散縮小とやや相性が悪い.

分散縮小勾配法の収束レート

研究紹介

結果： 対数ソボレフ不等式 + 滑らかさの仮定の下，
KL-divergenceの意味での収束が分散縮小型確率的勾配を用いることで
高速化できることを証明.

意義：
- KL-divergenceは“強いノルム”.
- 目的関数の性質を対数ソボレフ不等式の定数に集約できる.

$D(\mu_t||\nu) \leq \epsilon$までの計算量

- **Vempala&Wibisono (2019): 非確率的勾配**

勾配計算量 $\tilde{O}\left(\frac{n}{\epsilon} d\gamma^2 L^2 \alpha^{-2}\right)$

- **Our result: 確率的勾配+分散縮小法**

勾配計算量 $\tilde{O}\left(\left(n + \frac{\sqrt{n}d}{\epsilon}\right)\gamma^2 L^2 \alpha^{-2}\right)$ **：$\sqrt{n}$倍高速**

“Weak Morse”条件における対数ソボレフ定数も導出

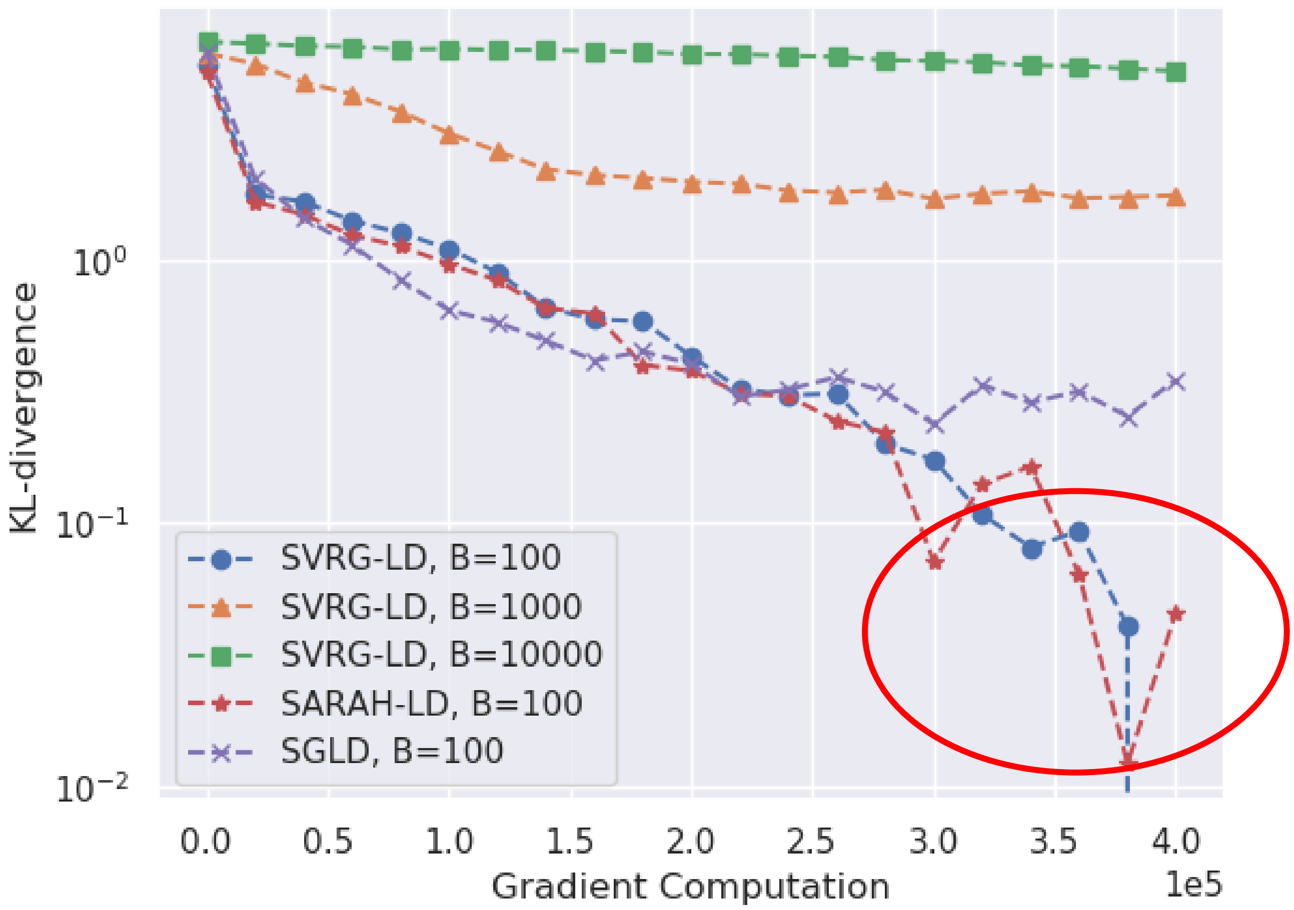
$\alpha = O(\exp(-\gamma))$

実は一般的に対数ソボレフ定数は逆温度パラメータγへ指数的に依存.
→ Weak Morse条件では多項式オーダーに緩和される.

- $0 < \exists \lambda^+ \leq$(任意の停留点のHessianの固有値の絶対値)
- 大域的最適解以外の停留点は全て鞍点かつ最小固有値が$-\lambda^+$以下

$$\alpha \gtrsim \frac{1}{\gamma}\left(1 + \frac{1}{\lambda^\dagger}\right)$$

研究紹介



KL-divergence
10^0
10^{-1}
10^{-2}
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
1e5
Gradient Computation
SVRG-LD, B=100
SVRG-LD, B=1000
SVRG-LD, B=10000
SARAH-LD, B=100
SGLD, B=100

提案手法

GLDによるNNの最適化

二層NNの最適化

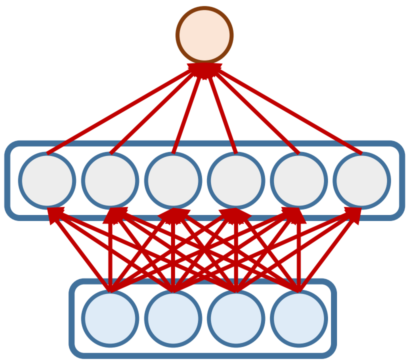
観測モデル:

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

where $x_i \sim \mathrm{Unif}(\mathbb{S}^{d-1}),\ \xi_i \sim (\text{mean } 0,\ \text{variance } \sigma^2,\ \text{bounded})$

教師生徒設定 with ReLU活性化関数:

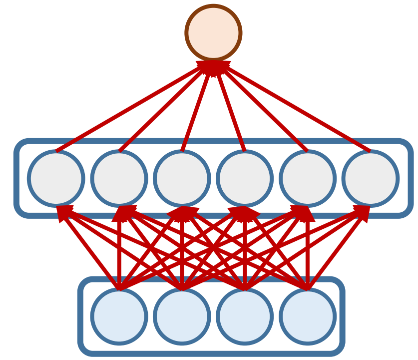
Teacher



$$f^\circ(x) = \sum_{j=1}^m a_j^\circ \sigma(\langle w_j^\circ, x\rangle)$$

σ : ReLU

Student (overparameterization)



$$f_\Theta(x) = \sum_{j=1}^m a_j \sigma(\langle w_j, x\rangle)$$

- **教師と生徒で同じサイズとする.** (おそらく緩和可能)
- 生徒は教師を多項式時間で推定できるか？

勾配ランジュバン動力学による最適化

L2-正則化あり経験損失関数:

$$\widehat{\mathcal{R}}_\lambda(\Theta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{M} a_j \sigma(\langle w_j, x\rangle)\right)^2 + \lambda \underbrace{\sum_{j=1}^{M}(a_j^2 + \|w_j\|^2)}_{\text{L2-正則化}}$$

$$\sum_{j=1}^{M} |a_j| \|w_j\| \le \frac{1}{2}\sum_{j=1}^{M}(a_j^2 + \|w_j\|^2)$$

ReLUではL2-正則化はスパース正則化 (L1-正則化) でもある.

二段階最適化

(1) 探索フェーズ: GLD

$$\Theta^{(k+1)} = \Theta^{(k)} - \eta^{(1)} \nabla \widehat{\mathcal{R}}_\lambda(\Theta^{(k)}) + \sqrt{\frac{2\eta^{(1)}}{\beta}}\underbrace{\zeta^{(k)}}_{\text{ガウスノイズ}}$$

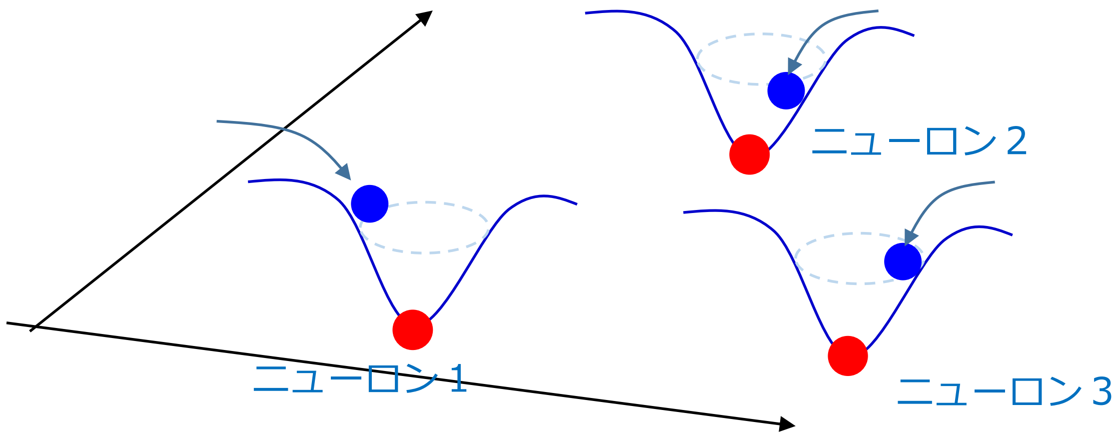
(2) 収束フェーズ: ノイズ無し・正則化なしの勾配法

$$\Theta^{(k+1)} = \Theta^{(k)} - \eta^{(1)} \nabla \widehat{\mathcal{R}}_0(\Theta^{(k)})$$

収束の様子

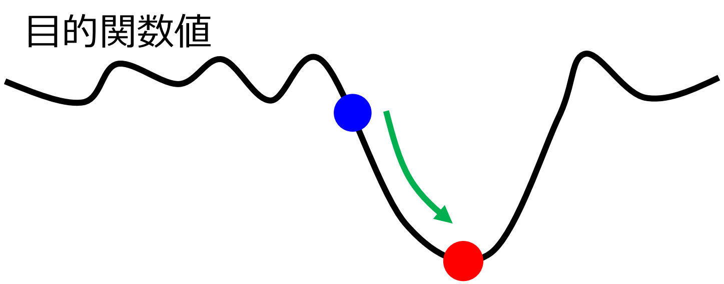
探索フェーズ:

最適解近傍への収束



収束フェーズ:

最適解まわりでは大体強凸
→ 普通の勾配法で線形収束



二段階の収束フェーズ：実験的にも観測されている (Hidden progress)

収束解析

[Akiyama and Suzuki: Excess Risk of Two-Layer ReLU Neural Networks in Teacher-Student Settings and its Superiority to Kernel Methods. arXiv:2205.14818]

- $\widehat{\mathcal{R}}_\lambda(\Theta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{M} a_j \sigma(\langle w_j, x\rangle)\right)^2 + \lambda \sum_{j=1}^{M}(a_j^2 + \|w_j\|^2)$
- $\mathcal{R}(\Theta) = \mathbb{E}[(f^\circ(X) - f_\Theta(X))^2]$
- $W^\circ = (w_1^\circ, \dots, w_m^\circ)$のm番目の特異値は$\sigma_{\min}$で下から抑えられる．

定理 (informal)

$(m, d, \sigma_{\min}$固定の元で)適当な定数時間K_1が存在して，

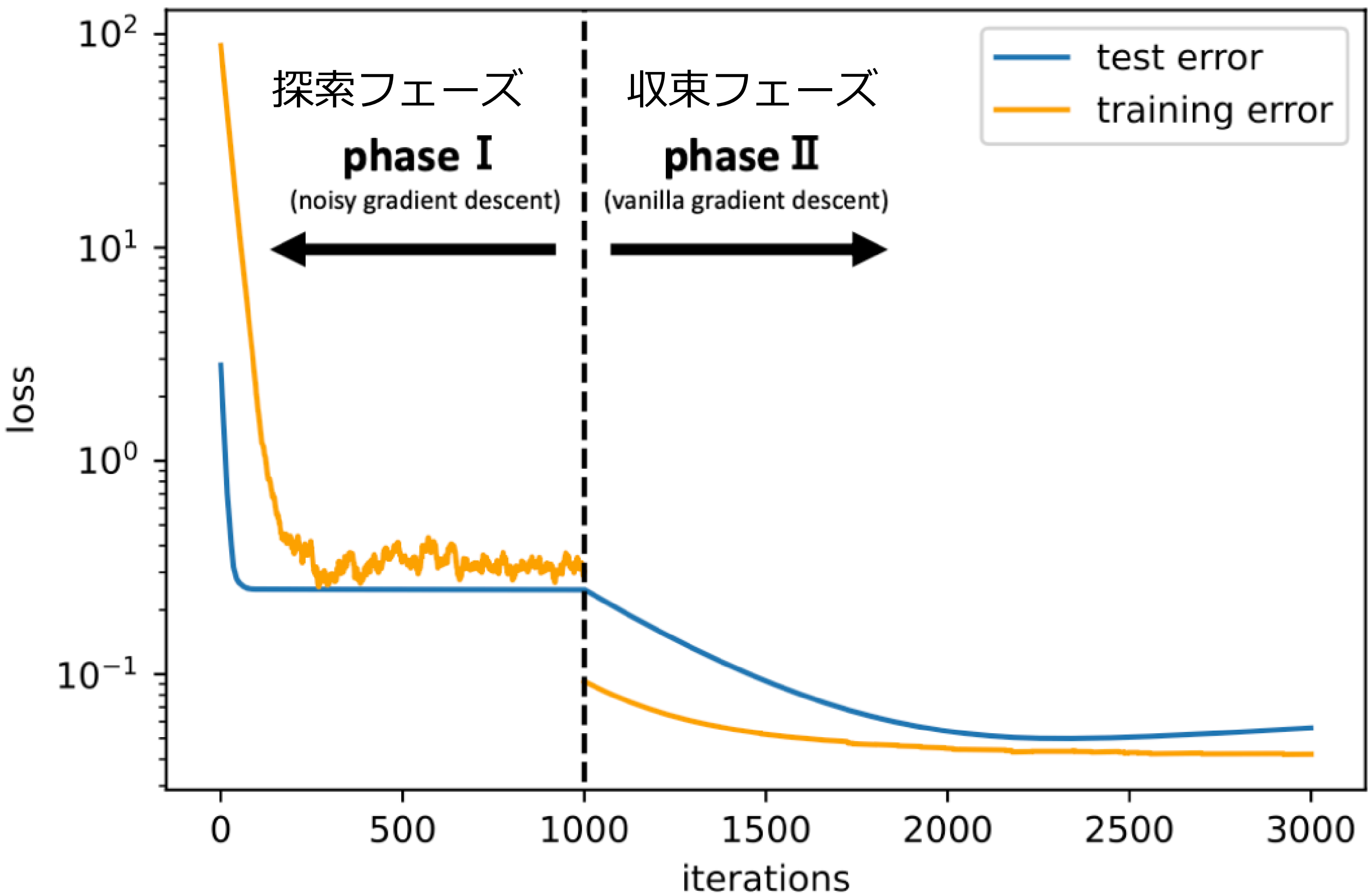
Phase 1 (大域的探索フェーズ): しばらく探索すると期待損失をある閾値以下まで減らせる (真に近くなる)

$$R\left(\Theta^{(K_1)}\right) \le \epsilon_0$$

Phase 2 (収束フェーズ): 大域的最適解へ線形収束

$$\hat{R}_0\left(\Theta^{(k)}\right) - \hat{R}_0(\Theta^*) \le c_1 \exp(-c_2(k - K_1)).$$

数値実験：二段階の学習ダイナミクス



予測誤差の比較

- 前ページのアルゴリズムで最適化して得られた解Θ^k

$$\|f_{\Theta^k} - f^\circ\|_{L^2(P_X)}^2 \lesssim \frac{\sigma_{\min}^{-4} m^5 \log(n)}{n}$$

- 線形推定量の予測誤差の下限:

$$R_{\mathrm{lin}} \gtrsim n^{-\frac{d+2}{2d+2}}$$

For $d = 2$: $n^{-2/3}$
For large d: $n^{-1/2}$

線形推定量は次元の呪いを受ける
→ 特徴量も学習することで改善

Related work

Non-overparameterized setting:

- [Li & Yuan, 2017] showed global convergence under $M = m$ and a special network structure (ResNet like structure).
- [Zhong et al., 2017] showed local convergence under $M = m$, i.e., they showed convergence when the initial solution is close to the true parameter.

Overparameterized setting:

- [Li, Ma & Zhang, 2020] showed global convergence of GD for an overparameterized setting $M > m$.

$$f^\circ(x) = \sum_{j=1}^{m} |\langle w_j, x\rangle|$$

True network

$$\mathcal{L}(\hat{f}) - \mathcal{L}(f^\circ) = O(d^{-(1+Q)})$$

where Q is a small constant.

  - ➢ Tensor decomposition technique is used.
  - ➢ The true network has a special structure.
  - ➢ The convergence is not exactly shown (it converges as $d \to \infty$).

- [Chizat, 2019]: Convergence to sparse solution with sparse reg.
  - ➢ BLASSO [De Castro & Gamboa, 2012]

2-homogeneous activation + NDSC condition

ReLU **Guarantee** ← [Akiyama&Suzuki, 2021]

平均場ランジュバン動力学

2層NNのGLDによる最適化

$$f_\Theta(x) = \frac{1}{M}\sum_{j=1}^{M} h_{\theta_j}(x)$$

例: $h_\theta(x) = r\sigma(w^\top x)$ for $\theta = (r, w)$

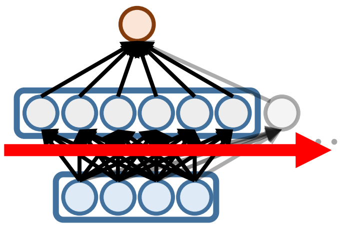
$$\mathrm{d}\theta_{j,t} = -\nabla_{\theta_j} L(f_{\Theta_t})\mathrm{d}t + \sqrt{2/\beta}\mathrm{d}B_t$$

ニューロンが沢山あると普通のGLDの理論が適用できない．
しかし，平均場ランジュバン動力学の理論により理論保証ができる．

(逆にニューロン数無限大の極限を考えると理論保証可能になる)

多粒子化（平均場）：

$$f_\Theta(x) = \frac{1}{M}\sum_{j=1}^{M} h_{\theta_j}(x) \xrightarrow{M\to\infty} f_\mu(x) = \int h_\theta(x)\mathrm{d}\mu(\theta)$$



定理 (Hu, Ren, Šiška, and Szpruch, 2021; Mei, Montanari, and Nguye, 2018)

$M \to \infty, t \to \infty$の極限で粒子$\theta_j$の分布$\mu_t$は以下の分布に収束:

$$\mu_\infty = \mathop{\arg\min}_{\mu\in\mathcal{P}} L(f_\mu) + \frac{1}{\beta}\mathrm{Ent}(\mu)$$

エントロピー
$(\mathrm{Ent}(\mu) = \int \log(\mu)\mathrm{d}\mu)$

重要：分布μに対しては凸関数！(if 損失が凸)

Objective of mean field NN

Mean field NN Model:

$$f_\mu(z) = \int h_x(z)\mathrm{d}\mu(x) \quad \text{where} \quad h_x(z) = r\sigma(w^\top z) \text{ for } x = (r, w)$$

Loss function (empirical risk + regularization):

$$F(\mu) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(f_\mu) + \lambda_1 \mathbb{E}_\mu[\|x\|^2]$$

$$\mathcal{L}(\mu) := F(\mu) + \lambda_2 \mathrm{Ent}(\mu)$$

convex + strongly convex = strongly convex

Nonlinear extension!

Vanilla GLD

$$\mathcal{L}(\mu) = \int L(x)\mathrm{d}\mu(x) + \lambda_2 \mathrm{Ent}(\mu)$$

Linear

General form of mean field LD

Mean field Langevin dynamics:

$$\mathcal{L}(\mu) = F(\mu) + \lambda_2 \mathrm{Ent}(\mu)$$

convex

- Wasserstein gradient flow to minimize $\mathcal{L}$:

$$\partial_t \mu_t = \nabla \cdot \left[\left(\nabla \frac{\delta F(\mu_t)}{\delta \mu} + \lambda_2 \nabla \log(\mu_t) \right) \mu_t \right]$$

- SDE the Fokker-Planck equation of which corresponds to this Wasserstein GF

$$\mathrm{d}X_t = -\nabla \frac{\delta F(\mu_t)}{\delta \mu}(X_t)\mathrm{d}t + \sqrt{2\lambda_2}\mathrm{d}B_t$$

$$\mu_t = \mathrm{Law}(X_t)$$

Q: Discretization error?

Definition (first variation)

The first variation $\frac{\delta F}{\delta \mu}: \mathcal{P} \times \mathbb{R}^d \to \mathbb{R}$ is defined as a continuous functional such as

$$\lim_{\epsilon \to 0} \frac{F(\epsilon\nu + (1-\epsilon)\mu) - F(\mu)}{\epsilon} = \int \frac{\delta F(\mu)}{\delta \mu}(x)\mathrm{d}(\nu - \mu)(x)$$

$$\mathcal{L}(\mu) = F(\mu) + \lambda_2 \mathrm{Ent}(\mu)$$

$$\mathrm{d}X_t = -\nabla \frac{\delta F(\mu_t)}{\delta \mu}(X_t)\mathrm{d}t + \sqrt{2\lambda_2}\mathrm{d}B_t$$

Vanilla GLD:

$$F(\mu) = \int L(x)\mathrm{d}\mu$$

(linear)

$$\Rightarrow \frac{\delta F(\mu)}{\delta \mu}(\cdot) = L(\cdot)$$

$$\Rightarrow \mathrm{d}X_t = -\nabla L(X_t)\mathrm{d}t + \sqrt{2\lambda_2}\mathrm{d}B_t$$

非線形Fokker-Planck方程式

- MF-LD obeys the following nonlinear Fokker-Planck equation:

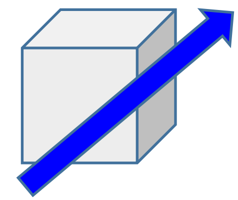
$$\partial_t \mu_t = \lambda_2 \Delta_x \mu_t + \nabla \cdot \left[\mu_t \nabla \frac{\delta F(\mu_t)}{\delta \mu}\right]$$

$$\left(= \nabla \cdot \left[\underbrace{\left(\lambda_2 \nabla \log(\mu_t) + \nabla \frac{\delta F(\mu_t)}{\delta \mu}\right)}_{=:\,-v_t} \mu_t\right]\right)$$

$$= -\nabla \cdot [v_t \mu_t]$$

[Continuity equation]

Mass: $\mu_t(x)$



Vector field: $b(x, \mu_t$

Then,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}(\mu_t) = \int \left\langle v_t, \nabla \frac{\delta \mathcal{L}(\mu_t)}{\delta \mu} \right\rangle \mathrm{d}\mu_t \quad (\because \text{continuity equation})$$

$$= \int \langle v_t, \nabla \frac{\delta F(\mu_t)}{\delta \mu} + \lambda_2 \nabla \log(\mu_t) \rangle \mathrm{d}\mu_t$$

$$= -\int \|v_t\|^2 \mathrm{d}\mu_t = -\lambda_2^2 I(\mu_t \| p_{\mu_t})$$

$$\mathcal{L}(\mu) := F(\mu) + \lambda_2 \mathrm{Ent}(\mu)$$

GLD：

$$F(\mu) = \int L(x)\mathrm{d}\mu$$

$$\Rightarrow \frac{\delta F(\mu)}{\delta \mu}(\cdot) = L(\cdot)$$

(Definition of p_{μ_t})

$$p_\mu(x) \propto \exp\left(-\frac{1}{\lambda_2}\frac{\delta F(\mu)}{\delta \mu}(x)\right)$$

This is the Wasserstein gradient flow to minimize $\mathcal{L}$.

※ Since $\frac{\delta F(\mu_t)}{\delta \mu}$ nonlinearly depends on μ_t, we say "nonlinear Fokker-Planck".

$$\text{GLD: } F(\mu) = \int L(x)\mathrm{d}\mu \Rightarrow \frac{\delta F(\mu)}{\delta \mu}(\cdot) = L(\cdot)$$

MF-LD to optimize mean field NN

$$f_\mu(z) = \int h_x(z)\mathrm{d}\mu(x)$$

$$h_x(z) = r\sigma(w^\top z) \text{ for } x = (r, w)$$

Loss function:

$$F(\mu) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(f_\mu) + \lambda_1 \mathbb{E}_\mu[r(X)]$$

$$\mathrm{d}X_t = \underbrace{-\nabla_{X_t}\left(\frac{1}{n}\sum_{i=1}^{n} \ell_i'(f_{\mu_t})h_{X_t}(z_i) + \lambda_1 r(X_t)\right)}_{\nabla \frac{\delta F(\mu_t)}{\delta\mu}(X_t)}\mathrm{d}t + \underbrace{\sqrt{2\lambda_2}\mathrm{d}B_t}_{\text{(escape from local min.)}}$$

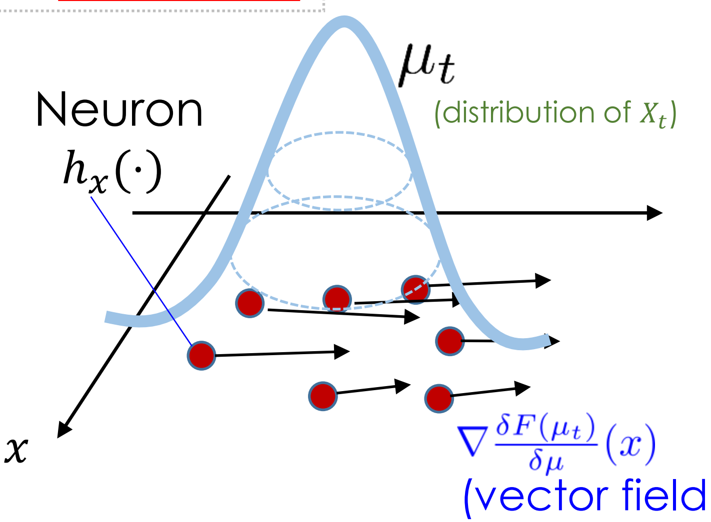
[Noise perturbation]

$$\mu_t = \mathrm{Law}(X_t)$$

$$\mathcal{L}(\mu) = \underline{F(\mu)} + \underline{\lambda_2 \mathrm{Ent}(\mu)}$$

- **Finite particle approximation:**

$$\mathrm{d}\hat{X}_t^i = -\nabla\frac{\delta F\left(\frac{1}{N}\sum_{j=1}^{N}\delta_{\hat{X}_t^j}\right)}{\delta\mu}(\hat{X}_t^i)\mathrm{d}t + \sqrt{2\lambda_2}\mathrm{d}B_t^i$$



Equivalent to GLD for optimizing a <u>finite width</u> neural network

Other applications

Mean field Langevin dynamics can be applied to several problems where a distribution is optimized.

- **Nonparametric density estimation via MMD minimization**

$$F(\mu) = \mathrm{MMD}^2(g * \mu, \hat{\mu}_n) + \lambda_1 \mathbb{E}_\mu[\|x\|^2]$$

k: positive definite kernel

$$\mathrm{MMD}^2(\nu_1, \nu_2) := \|k_{\nu_1} - k_{\nu_2}\|_{\mathcal{H}_k}^2$$

where $k_\mu = \int k(x,\cdot)\mu(\mathrm{d}x)$ (kernel embedding).

- $g(x) = \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right)$
- $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$: Empirical distribution (training data)

(see also Chizat (2022,TMLR))

- **Variational inference to approximate Bayesian posterior**

$$F(\mu) = \mathrm{KSD}(\mu) + \lambda_1 \mathbb{E}_\mu[\|x\|^2]$$

(KSD: Kernel Stein Discrepancy from a posterior distribution)

Convergence of MF-LD

Proximal Gibbs measure:

$$p_\mu(x) \propto \exp\left(-\frac{1}{\lambda_2}\frac{\delta F(\mu)}{\delta \mu}(x)\right) \qquad p_\mu = \mathop{\arg\min}_{\nu \in \mathcal{P}} (\nu - \mu)\frac{\delta F(\mu)}{\delta \mu} + \lambda_2 \mathrm{Ent}(\nu)$$

Them (Entropy sandwich) [Nitanda, Wu, Suzuki (AISTATS2022)][Chizat (2022)]

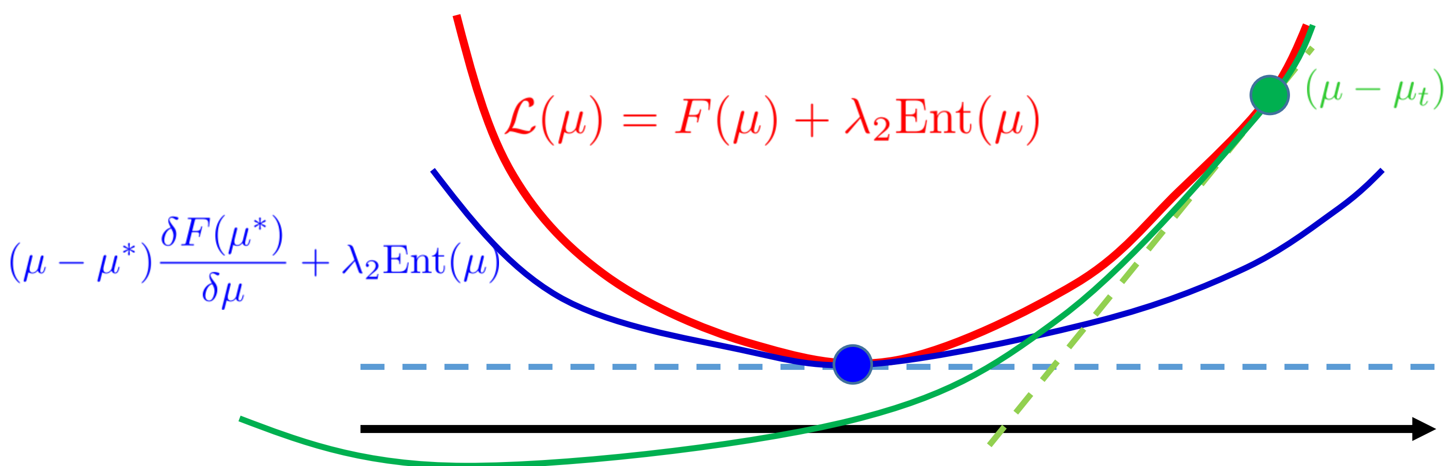
Suppose that p_μ satisfies log-Sobolev inequality with constant α, then

$$\mathcal{L}(\mu_t) - \mathcal{L}(\mu^*) \leq \exp(-2(\lambda_2/\alpha)t)(\mathcal{L}(\mu_0) - \mathcal{L}(\mu^*)),$$

$$\lambda_2 \mathrm{KL}(\mu||\mu^*) \leq \mathcal{L}(\mu) - \mathcal{L}(\mu^*) \leq \lambda_2 \mathrm{KL}(\mu||p_\mu).$$

$$\mu^* = \mathop{\arg\min}_{\mu \in \mathcal{P}} \mathcal{L}(\mu)$$

Log-Sobolev: $D(\mu||\nu) \leq \frac{1}{2\alpha} I(\mu||\nu)$ for all ν. $D(\mu||\nu) = \int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\mu,\quad I(\mu||\nu) = \int \left\|\nabla \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right\|^2 \mathrm{d}\mu$



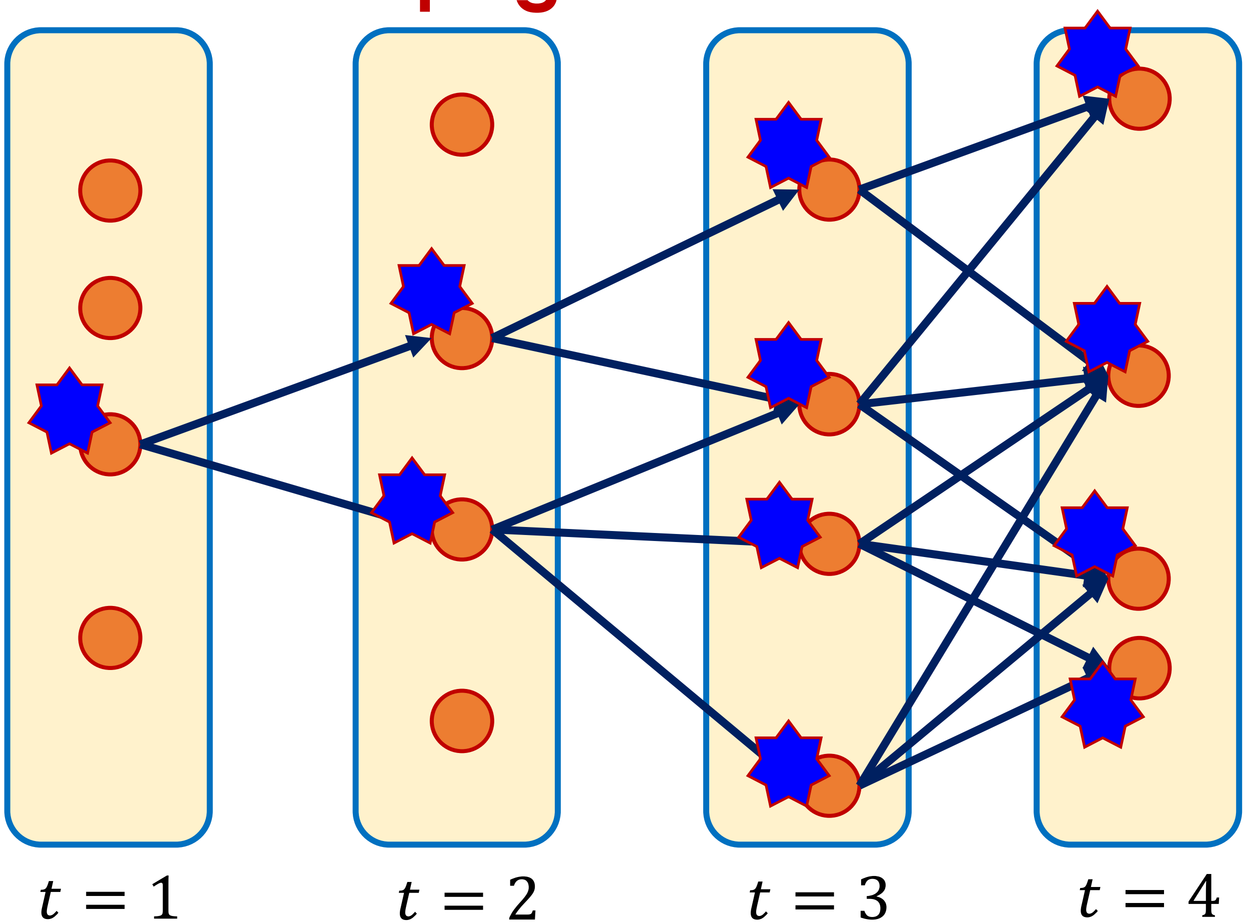
LSI is satisfied by Holley--Stroock if the loss and its gradient are bounded.

有限粒子近似

難しさ: McKean-Vlasov過程

- 粒子間相互作用のある確率微分方程式はMcKean-Vlasov過程として知られている.　(McKean, Kac,…, 60年代)
- 離散時間・有限粒子での収束を示す際にはPropagation of chaosの評価が難しい. (粒子を増やすことでそれぞれがあたかも独立に振る舞う現象)

Propagation of chaos



一つの粒子の微小な変化が他の粒子に伝播して増幅される可能性がある.

研究の流れ

平均場NNの線形収束
連続時間・無限粒子
[Nitanda, Wu, Suzuki (AISTATS2022)]
[Chizat (2022)]

時間・空間離散化：「二重ループの手法」

- PDA法 [Nitanda, Wu, Suzuki: NeurIPS2021]
- P-SDCA法 [Oko, Suzuki, Wu, Nitanda: ICLR2022]
- 無限次元拡張 [Nishikawa, Suzuki, Nitanda: NeurIPS2022]

難しい：**Propagation of chaos**
(McKean, Kac,…, 60年代より)

空間離散化・連続時間:
Uniform-in-time propagation of chaos

- Super対数Sobolev不等式
  [Suzuki, Nitanda, Wu (ICLR2023)]
- Leave-one-out型評価
  [Chen, Ren, Wang (arXiv2022)]

時間・空間離散化・<u>確率的勾配</u>：
「<u>一重ループの手法</u>」

Suzuki, Wu, Nitanda
(arXiv:2306.07221)

離散時間・有限横幅の手法

粒子双対平均化法
(Particle Dual Averaging; PDA)

[Nitanda, Wu, Suzuki: NeurIPS2021]

$$\min_{q:\text{prob.density}} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell\left(\mathbb{E}_q[h_\theta(x_i)], y_i\right) + \lambda_1 \mathbb{E}_q[\|\theta\|^2]} + \lambda_2 \mathbb{E}_q[\log(q)]$$

近似

qに関する線形汎関数で近似 (勾配を用いる)

$\mathbb{E}_{\theta\sim q}[\bar{g}^{(t)}(\theta)]$ (線形近似; $\bar{\boldsymbol{g}}^{(t)}$は基本的に勾配)

$\bar{g}^{(t)}$の決定に<u>双対平均化法</u>のルールを用いる

$$\min_{q:\text{prob.density}} \mathbb{E}_{\theta\sim q}[\bar{g}^{(t)}(\theta)] + \lambda_2 \mathbb{E}_q[\log(q)]$$

解: $q^{(t+1)}(\theta) \propto \exp(-\bar{g}^{(t)}(\theta)/\lambda_2)$ 具体形が得られる.

→ この分布からは以下の勾配ランジュバン動力学を用いてサンプリング可能:

$$\mathrm{d}\theta_t = -\nabla(\bar{g}^{(t)}(\theta)/\lambda_2)\mathrm{d}t + \sqrt{2}\mathrm{d}\xi_t.$$

時間離散化

$$\theta_k = \theta_{k-1} - \eta\nabla\bar{g}^{(t)}(\theta)/\lambda_2 + \sqrt{2\eta}\xi_{k-1}$$

計算量解析：

1. **外側ループ:** $\mathcal{L}(\hat{q}^{(t)}) - \mathcal{L}(q^*) \leq O(1/t)$
2. **内側ループ:** $T_t = \tilde{O}\left(t^2 \exp(8/\lambda_2)/(\lambda_1\lambda_2)\right)$ (GLDによる)

⇒合計: $O(\epsilon^{-3})$の勾配アップデートで十分.

➢ **<u>初の多項式オーダー最適化手法</u>**

粒子確率的双対座標上昇法
(Particle Stochastic Dual Coordinate Ascent; P-SDCA)

[Oko, Suzuki, Wu, Nitanda: ICLR2022]

主問題

$$\min_p P(p) = \frac{1}{n}\sum_{i=1}^{n} \ell_i\left(\int p(\theta)h_i(\theta)\right) + \lambda_1 \int \|\theta\|^2 p(\theta)\mathrm{d}\theta + \lambda_2 \int p(\theta)\log(p(\theta))\mathrm{d}\theta$$

|| by Fenchelの双対定理

双対問題 $\ell_i^*(g) := \sup_{u\in\mathbb{R}}\{ug - \ell_i(u)\}$

$$-\min_{g\in\mathbb{R}^n} D(g) = \frac{1}{n}\sum_{i=1}^{n} \ell_i^*(g_i) + \lambda_2 \log\left(\int q[g](\theta)\mathrm{d}\theta\right)$$

ただし $q[g](\theta) := \exp\left\{-\frac{1}{\lambda_2}\left(\frac{1}{n}\sum_{i=1}^{n} h_i(\theta)g_i + \lambda_1\|\theta\|^2\right)\right\}$

- 双対変数の座標をランダムに選択し，その座標に関して最適化.

→確率的双対座標上昇法

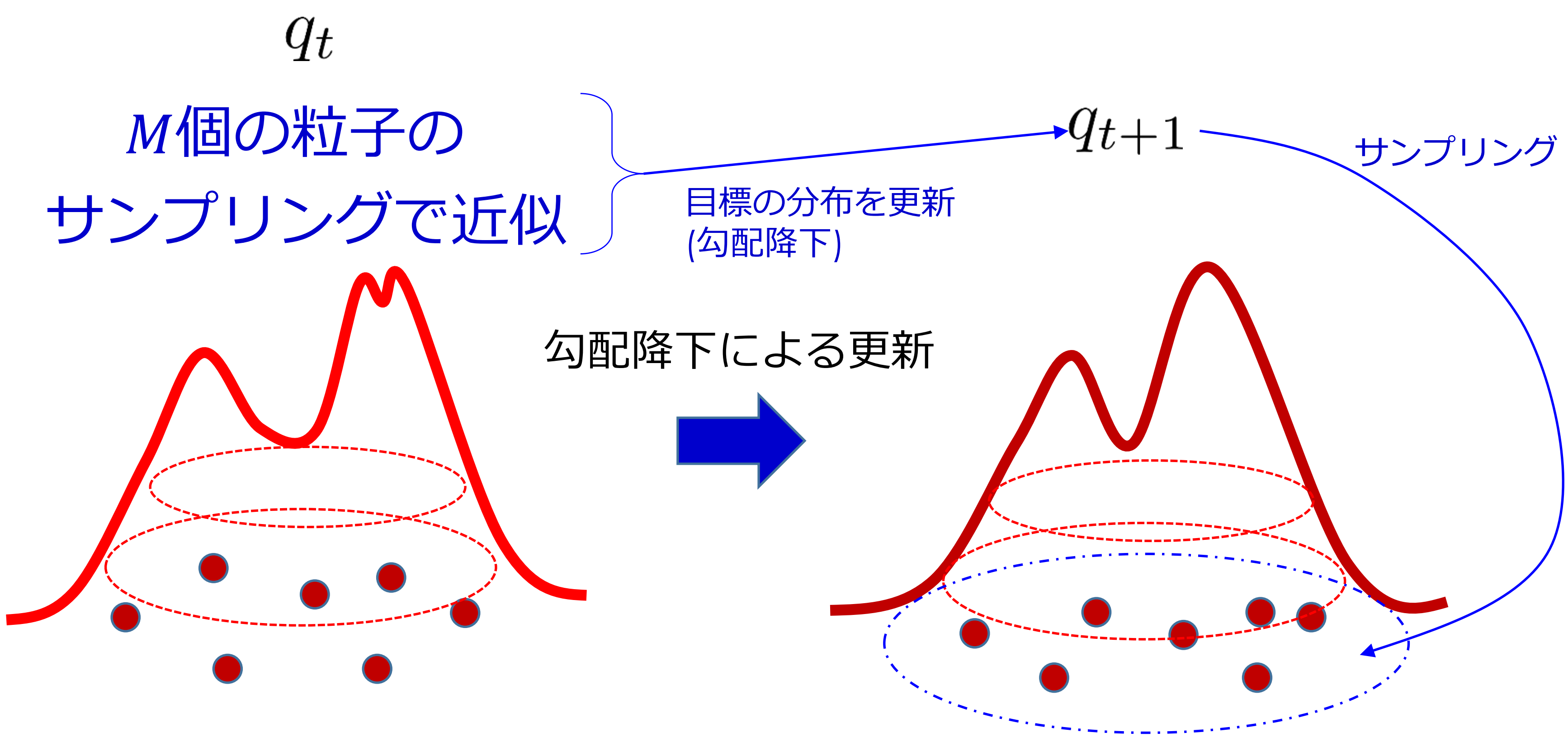
計算量解析：

双対ギャップϵ_Pを達成するのに必要な外側ループ数：

$$t_{\mathrm{end}} = 2\left(n + \frac{1}{\lambda_2\gamma}\right)\log\left(\frac{nC}{\epsilon_P}\right)$$

➢ <u>指数オーダーでの収束を達成</u>

➢ <u>サンプルサイズnへの依存を緩和</u>


q_t
M個の粒子の
サンプリングで近似
q_{t+1}
目標の分布を更新
(勾配降下)
サンプリング
勾配降下による更新

粒子双対平均化法 (Particle Dual Averaging; PDA)

[Nitanda, Wu, Suzuki: NeurIPS2021]

$$\min_{q:\mathrm{prob.density}} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\ell\Big(\mathbb{E}_q[h_\theta(x_i)], y_i\Big) + \lambda_1 \mathbb{E}_q[\|\theta\|^2]}_{} + \lambda_2 \mathbb{E}_q[\log(q)]$$

Approximation

線形近似 $\mathbb{E}_{\theta\sim q}[\bar{g}^{(t)}(\theta)]$ ($\bar{\boldsymbol{g}}^{(t)}$は基本的に勾配)

qに関する線形汎関数で近似 (勾配を用いる)

双対平均化法 (Nesterov, 2005; 2009; Xiao, 2009)

$$g^{(t)}(\theta) \leftarrow \ell'(\underline{\mathbb{E}_{\theta'\sim q^{(t)}}[h_{\theta'}(x_{i_t})]}, y_{i_t})h_\theta(x_{i_t}) + \lambda_1\|\theta\|_2^2$$

$$\bar{g}^{(t)} \leftarrow \tfrac{2}{(t+2)(t+1)}\textstyle\sum_{s=1}^{t} s g^{(s)}$$

サンプリングで近似

$$\mathbb{E}_{q^{(t)}}[h_\theta(x)] \simeq \frac{1}{M}\sum_{r=1}^{M} h_{\theta^{(r)}}(x)$$

$$\min_{q:\mathrm{prob.density}} \mathbb{E}_{\theta\sim q}[\bar{g}^{(t)}(\theta)] + \lambda_2\mathbb{E}_q[\log(q)]$$

解析解: $q^{(t+1)}(\theta) \propto \exp(-\bar{g}^{(t)}(\theta)/\lambda_2)$ 具体形が得られる

→この分布からは以下の勾配ランジュバン動力学を用いてサンプリング可能:

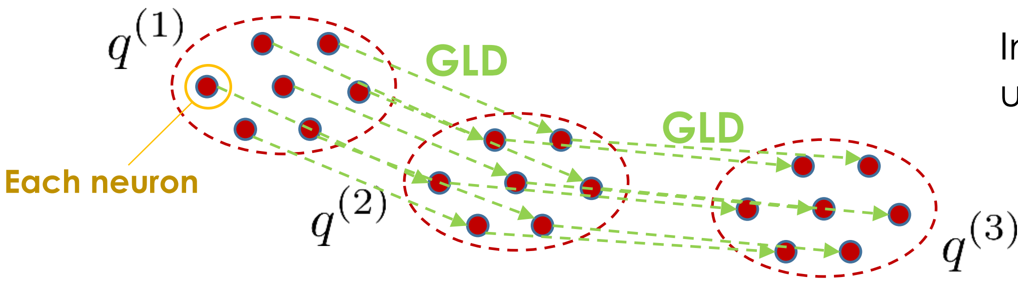
$$\mathrm{d}\theta_t = -\nabla(\bar{g}^{(t)}(\theta)/\lambda_2)\mathrm{d}t + \sqrt{2}\mathrm{d}\xi_t.$$

離散化

$$\theta_k = \theta_{k-1} - \eta\nabla\bar{g}^{(t)}(\theta)/\lambda_2 + \sqrt{2\eta}\xi_{k-1}$$

勾配ランジュバン動力学

Algorithm description



In each iteration, the potential for updating each particle is given by $\bar{g}^{(t)}$.

Algorithm 1 Particle Dual Averaging (PDA)

Randomly draw i.i.d. initial parameters $\tilde{\theta}_r^{(1)} \sim q^{(1)}(\theta)\mathrm{d}\theta$ $(r \in \{1, 2, \ldots, M\})$

$\tilde{\Theta}^{(1)} \leftarrow \{\tilde{\theta}_r^{(1)}\}_{r=1}^M$

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

Randomly draw a data index i_t from $\{1, 2, \ldots, n\}$

$g^{(t)} \leftarrow \partial_z \ell(h_{\tilde{\Theta}^{(t)}}(x_{i_t}), y_{i_t}) h(\cdot, x_{i_t}) + \lambda_1 \| \cdot \|_2^2$

$\bar{g}^{(t)} \leftarrow \frac{2}{(t+2)(t+1)} \sum_{s=1}^t s g^{(s)}$

勾配計算

$$h_{\tilde{\Theta}}(x) := \frac{1}{M} \sum_{r=1}^{M} h_{\theta_r}(x)$$

Obtain $q^{(t+1)}$ by running the Langevin algorithm to approximate the following density function:

$$q_*^{(t+1)} \propto \exp\left(-\bar{g}^{(t)}/\lambda_2\right).$$

分布の更新

$\tilde{\Theta}^{(t+1)} \leftarrow \{\tilde{\theta}_r^{(t+1)}\}_{r=1}^M$ where $\tilde{\theta}_r^{(t+1)} \sim q_*^{(t+1)}$.

サンプリング

end for

Randomly pick up t from $\{2, 3, \ldots, T+1\}$ following the probability $P[t] = \frac{2t}{T(T+3)}$ and return $h_{\tilde{\Theta}^{(t)}}$

収束解析

定理 (informal)

1. 外側ループ: (T:外側ループの回数，M:粒子数)

$$\mathcal{L}(\hat{q}) - \mathcal{L}(q^*) \leq O(1/T) + O(1/\sqrt{M})$$

2. 内側ループ:

t-回目の外部ループにおいて，以下の回数だけGLDの内部ループを更新:

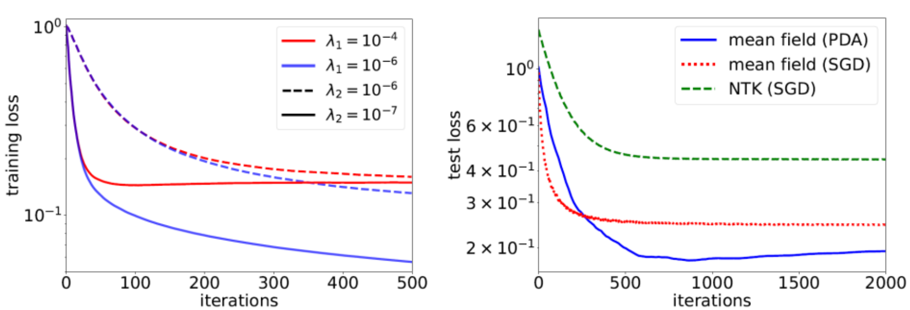
$$T_t = \tilde{O}\left(\eta_t^{-1}\right) = \tilde{O}\left(t^2 \exp(8/\lambda_2)/(\lambda_1\lambda_2)\right) \qquad \eta_t = O\left(\frac{\lambda_1\lambda_2}{t^2\exp(8/\lambda_2)}\right)$$

全体の複雑さ:

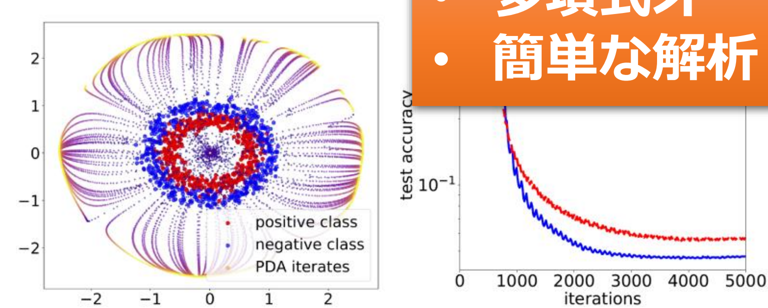
- $O(\epsilon^{-3})$ 回のGLD更新でϵ-最適解が得られる.
- ネットワークの横幅 (粒子数) は$M = \epsilon^{-2}\mathrm{poly}(n, d)$ で十分.

※厳密に離散時間・有限粒子数で多項式時間を達成した初めての方法.
※時間離散化，有限粒子近似の誤差も含んだ解析 (厳密な解析).

- **多項式オーダー**
- **簡単な解析**

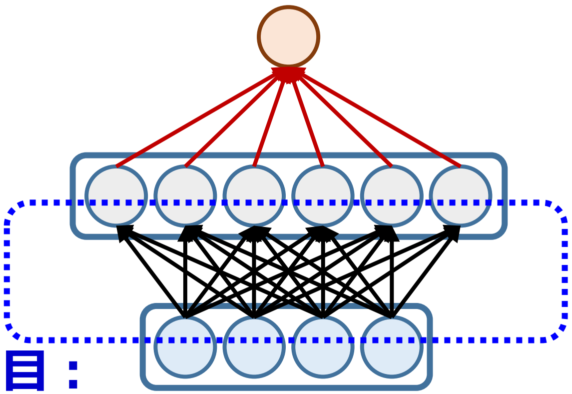


(a) training error (PDA). (b) test error comparison.



(a) trajectory of PDA. (b) test accuracy.

Kernel alignment

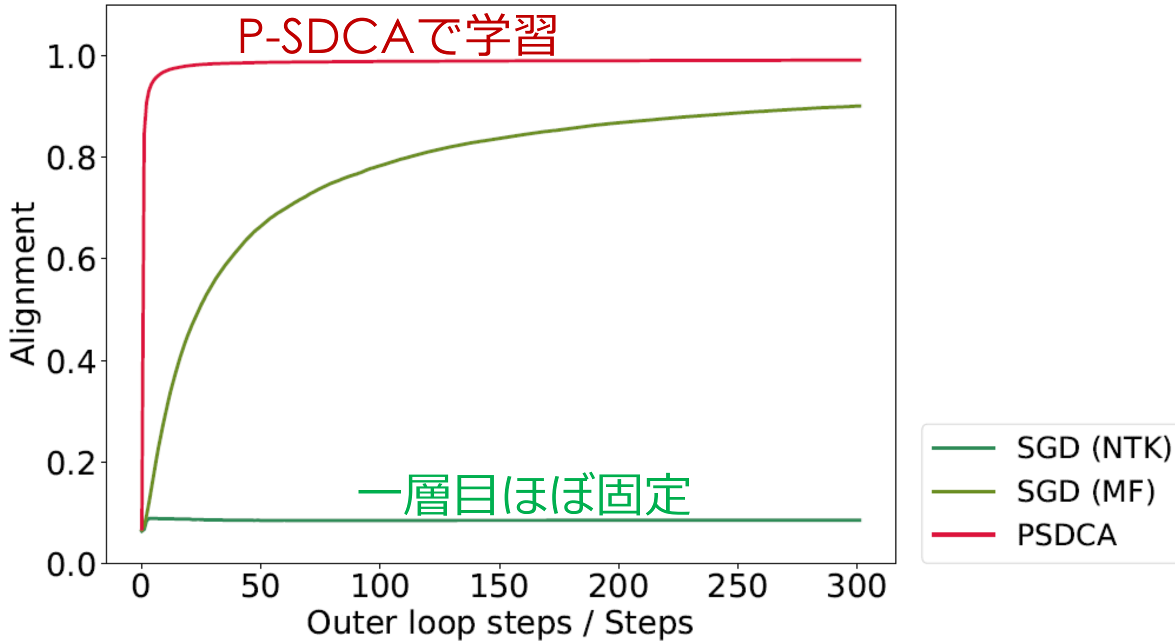


一層目：特徴抽出

カーネルalignment:

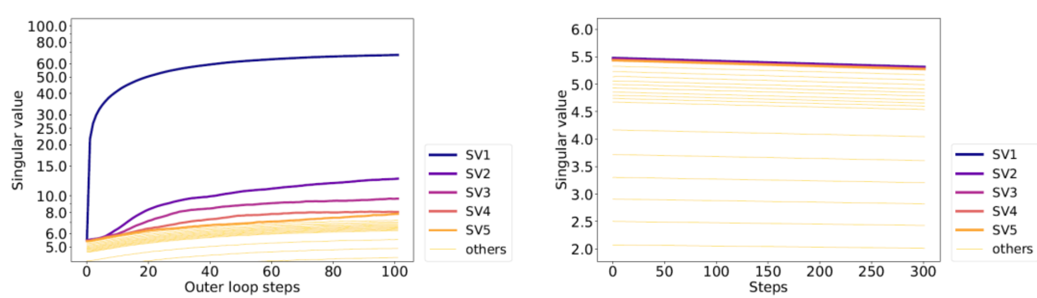
$$A(k_W) := \frac{\langle K_W, yy^\top \rangle_{\mathrm{F}}}{\sqrt{\langle K_W, K_W \rangle_{\mathrm{F}} \langle yy^\top, yy^\top \rangle_{\mathrm{F}}}}$$

一層目で抽出された特徴量が教師信号(y)とどれだけ相関しているか？
→ 高いほど特徴量が真の関数の成分を多く含んでいる．



一層目も学習することで真の関数により適合した特徴量が学習できている．

固有値の分布：



(a) P-SDCA (mean field regime)　(b) SGD (NTK regime)

有限和での収束改善

[Oko, Suzuki, Nitanda, and Wu: Particle Stochastic Dual Coordinate Ascent: Exponential convergent algorithm for mean field neural network optimization. The Tenth International Conference on Learning Representations (ICLR2022)]

Modification to SDCA

[Oko, Suzuki, Nitanda, and Wu: Particle Stochastic Dual Coordinate Ascent: Exponential convergent algorithm for mean field neural network optimization. The Tenth International Conference on Learning Representations (ICLR2022)]

$$\min_{q:\text{prob.density}} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell\left(\mathbb{E}_q[h_\theta(x_i)], y_i\right)}_{\text{訓練損失：\underline{\textbf{有限和}}}} + \lambda_1 \mathbb{E}_q[\|\theta\|^2] + \lambda_2 \mathbb{E}_q[\log(q)]$$

- 問題意識:
  - 外側ループの計算量を**有限和**の構造を活かして改善したい.
  - SDCA (Stochastic Dual Coordinate Ascent) を用いることで線形収束を達成：

$$\left(n + \frac{L}{\mu}\right)\log(1/\epsilon).$$

※ PDA: $1/\epsilon$

粒子確率的双対座標上昇法 (Particle Stochastic Dual Coordinate Ascent; P-SDCA)

主問題

$$\min_p P(p) = \frac{1}{n}\sum_{i=1}^n \ell_i\left(\int p(\theta)h_i(\theta)\mathrm{d}\theta\right) + \lambda_1 \int \|\theta\|^2 p(\theta)\mathrm{d}\theta + \lambda_2 \int p(\theta)\log(p(\theta))\mathrm{d}\theta$$

||

$$\min_{x\in\mathcal{X}} f(Ax) + g(x) = -\min_{g\in\mathcal{Y}^*} f^*(g) + g^*(-A^*g)$$ (Fenchelの双対定理)

$A:\mathcal{X}\to\mathcal{Y}$ (bounded linear)

双対問題

$$-\min_{g\in\mathbb{R}^n} D(g) = \frac{1}{n}\sum_{i=1}^n \ell_i^*(g_i) + \lambda_2 \log\left(\int q[g](\theta)\mathrm{d}\theta\right)$$

$$\ell_i^*(g) := \sup_{u\in\mathbb{R}}\{ug - \ell_i(u)\}$$

ただし

$$\begin{cases} q[g](\theta) := \exp\left\{-\frac{1}{\lambda_2}\left(\frac{1}{n}\sum_{i=1}^n h_i(\theta)g_i + \lambda_1\|\theta\|^2\right)\right\} \\ p[g](\theta) := \frac{q[g](\theta)}{\int q[g](\theta')\mathrm{d}\theta'} \end{cases}$$

双対変数の座標をランダムに選択し
その座標に関して最適化.
→双対座標上昇法

線形収束することを証明
$\exp(-(n+\kappa)^{-1}t)$

PDAが$1/t$なので大幅改善

One coordinate update

$$\min_{g_i \in \mathbb{R}} D(g) = \frac{1}{n}\sum_{i=1}^n \ell_i^*(g_i) + \lambda_2 \log\left(\int q[g](\theta)\mathrm{d}\theta\right)$$

We update just one coordinate g_i per iteration.

(ideal update)

proximal gradient descent (2nd term is linearized)

$$\Rightarrow \begin{cases} \bar{g}_i^{(t+1)} := \mathop{\arg\min}_{g_i \in \mathbb{R}} \left\{ \ell_i^*(g_i) - \underline{\int p^{(t)}(\theta) h_i(\theta)\mathrm{d}\theta (g_i - \bar{g}_i^{(t)})} + \frac{1}{2n\lambda_2}(g_i - \bar{g}_i^{(t)})^2 \right\} \\ \bar{g}_j^{(t+1)} = \bar{g}_j^{(t)} \quad (j \neq i) \\ p^{(t+1)}(\theta) := p[\bar{g}^{(t+1)}](\theta) \end{cases}$$

(requires integration)

$$p^{(t)}(\theta) \propto \exp\left\{ -\frac{1}{\lambda_2}\left(\frac{1}{n}\sum_{i=1}^n h_i(\theta) g_i^{(t)} + \lambda_1 \|\theta\|^2\right)\right\}$$

→ We can sample particles via GLD.

$$\theta_m \sim p^{(t)} \quad (m = 1, \dots, M)$$

(particle approximation)

$$\Rightarrow \int p^{(t)}(\theta) h_i(\theta)\mathrm{d}\theta \approx \sum_{m=1}^M r_m^{(t)} h_i(\theta_m)$$

$$r_m^{(0)} = 1/M, \quad \delta\bar{g}_i^{(t+1)} := \bar{g}_i^{(t+1)} - \bar{g}_i^{(t)}$$

$$\begin{cases} \tilde{r}_m^{(t+1)} = r_m^{(t)} \exp\left(-\frac{1}{n} h_i(\theta_m)\delta\bar{g}_i^{(t+1)}\right) \\ r_m^{(t+1)} = \frac{\tilde{r}_m^{(t+1)}}{\sum_{m=1}^M \tilde{r}_m^{(t+1)}} \quad (m \in [M]) \end{cases}$$

We "refresh" particles each $\tilde{n}$ iteration.

Algorithm description

参考

Algorithm 2 Dual Coordinate Descent with the particle method

Require: training data $\{(x_i, y_i)\}_{i=1}^n$ and numbers of inner-loop iterations $\tilde{n}$ and outer-loop iterations T_{end},

Choose $g_i^{(0)}$ s.t. $|\ell_i^{*\prime}(g_i^{(0)})| \leq 1$ $(i = 1, \ldots, n)$ and $\ell_i^*(g_i^{(0)}) \leq \ell_i^*(0)$

$g^{(0)} \leftarrow \mathbf{0}$,

for $T = 0, 1, \ldots, T_{\text{end}} - 1$ **do**

  Randomly (approximately) draw i.i.d. parameters θ_m $(m = 1, \ldots, M^{(\tilde{n}T)})$ from $p^{(\tilde{n}T)}(\theta)\mathrm{d}\theta$ that satisfies $\mathrm{TV}(p^{(\tilde{n}T)} \| p[g^{(\tilde{n}T)}]) \leq \epsilon_C^{(\tilde{n}T)}$.

  $r_m^{(\tilde{n}T)} \leftarrow \frac{1}{M^{(\tilde{n}T)}}$ $(m = 1, \ldots, M^{(\tilde{n}T)})$

  for $t = \tilde{n}T, \tilde{n}T + 1, \ldots, \tilde{n}T + \tilde{n} - 1$ **do**

    Randomly choose i_t from $\{1, 2, \ldots, n\}$

$$g_{i_t}^{(t+1)} \leftarrow \operatorname*{argmax}_{g_{i_t} \in \mathbb{R}} \left\{ -\ell_{i_t}^*(g_{i_t}) + \frac{\sum_{m=1}^{M^{(\tilde{n}T)}} r_m^{(t)} h_{i_t}(\theta_m)}{\sum_{m=1}^{M^{(\tilde{n}T)}} r_m^{(t)}} (g_{i_t} - g_{i_t}^{(t)}) - \frac{1}{2n\lambda_2}(g_{i_t} - g_{i_t}^{(t)})^2 \right\}.$$

$$r_m^{(t+1)} \leftarrow r_m^{(t)} \exp\left(-\frac{1}{n\lambda_2} h_{i_t}(\theta_m)(g_{i_t}^{(t+1)} - g_{i_t}^{(t)}) \right) \quad (m = 1, \ldots, M^{(\tilde{n}T)}).$$

  end for

end for

return Option (A): $g_{\text{out}}^{(A)} = g^{(\tilde{n}T_{\text{end}})}$; Option (B): $g_{\text{out}}^{(B)} = g^{(t'_{\text{end}})}$ for t'_{end} that is randomly chosen from $\{\tilde{n}T_{\text{end}} - n + 1, \ldots, \tilde{n}T_{\text{end}}\}$.

At every $\tilde{n}$ iteration, we refresh particles.

Particle weight

Dual coordinate

Discretization error

(A1) ℓ_i is γ-smooth.

(A2) $|h_i(\theta)| \leq 1$ for all θ.

(A3) Other technical conditions.

$$g_i^{(t+1)} := \arg\min_{g_i \in \mathbb{R}} \left\{ \ell_i^*(g_i) - \sum_{m=1}^{m} r_m^{(t)} h_i(\theta_m)(g_i - g_i^{(t)}) + \frac{1}{2n\lambda_2}(g_i - g_i^{(t)})^2 \right\}$$

Lemma (informal)

It holds that

(Ideal update)

$$|g_i^{(t)} - \bar{g}_i^{(t)}| \lesssim \sqrt{\frac{1}{M}\log(n/\delta)}$$

uniformly over $i \in [n], t \in [n]$ with probability $1 - \delta$.

If $t > n$, the error can exponentially diverge.
⇒ We re-sample $(\theta_m)_{m=1}^{M}$ by GLD at each $t = \tilde{n}$ updates.

収束レート

(A1) ℓ_i は$1/\gamma$-平滑.

(A2) $|h_i(\theta)| \leq 1$ が全てのθで成立.

(A3) その他テクニカルな条件.

$$P(p) = \frac{1}{n}\sum_{i=1}^{n} \ell_i\left(\int p(\theta)h_i(\theta)\right) + \lambda_1 \int \|\theta\|^2 p(\theta)\mathrm{d}\theta + \lambda_2 \int p(\theta)\log(p(\theta))\mathrm{d}\theta$$

$$D(g) = \frac{1}{n}\sum_{i=1}^{n} \ell_i^*(g_i) + \lambda_2 \log\left(\int q[g](\theta)\mathrm{d}\theta\right)$$

定理 (収束レート)

$\frac{\tilde{n}}{n\lambda_2}=o(1)$を仮定し，粒子数が以下を満たすとする:

$$M^* \gtrsim \frac{1}{\epsilon_P \lambda_2}.$$

より正確には

$$M^* \gtrsim \frac{1}{\epsilon_P \lambda_2} \exp\left\{C\left[\frac{\tilde{n}}{\lambda_2 n} + \frac{(\exp(\tilde{n}/\lambda_2 n)+1)}{n\gamma\lambda_2/\tilde{n}+1/\tilde{n}}\right]\right\}$$

すると

条件数

$$t_{\mathrm{end}} = 2\left(n + \frac{1}{\lambda_2\gamma}\right)\log\left(\frac{nC}{\epsilon_P}\right)$$

回の更新で双対ギャップϵ_Pを達成する:

$$\text{(Duality gap)} \quad \mathbb{E}[P(p^{(t_{\mathrm{end}})}) - D(g^{(t_{\mathrm{end}})})] \leq \epsilon_P$$

Total complexity:

$$M^*\left(1 + \frac{K^*}{\tilde{n}}\right)\left(n + \frac{1}{\lambda_2\gamma}\right)\log(n/$$

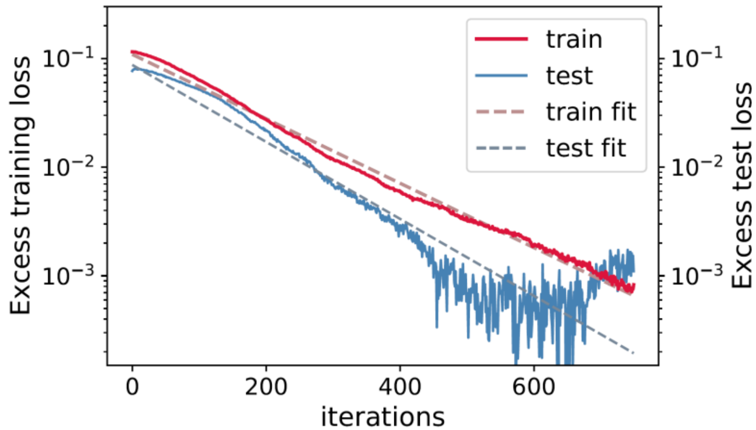
もし決定的な手法を用いたら，勾配の評価数は以下のようになる (n倍多い):

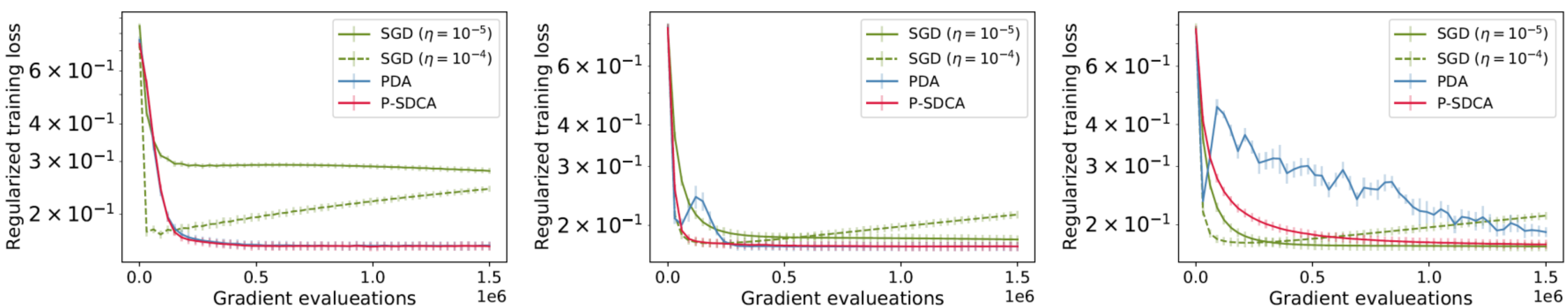
$$t_{\mathrm{end}} = O\left(\frac{n}{\lambda_2\gamma}\log(1/\epsilon_P)\right)$$

to generate M^* particles.

Experiments

$$y = \sigma(w_*^\top x + b^*) + \epsilon$$





(a) $\lambda_2 = 0.01$ (b) $\lambda_2 = 0.001$ (c) $\lambda_2 = 0.0001$

$\lambda_1 = 10^{-2}$: fixed

無限次元パラメータへの拡張

[Nishikawa, Suzuki, Nitanda, Wu: Two-layer neural network on infinite-dimensional data: global optimization guarantee in the mean-field regime. NeurIPS2022]

無限次元へ拡張

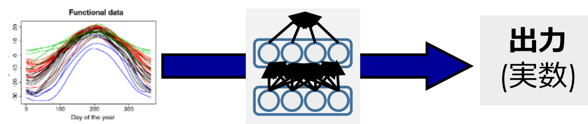
各ニューロン　$h_\theta(x) = r\sigma(w^\top x)$

↓ 無限次元ヒルベルト空間の元を入力とする

$h_\theta(x) = r\sigma(\langle w, x\rangle_{\mathcal{H}})$

推定すべきパラメータの次元も無限になる

応用例：



- 関数データ解析
- グラフを入力とするデータ解析 (グラフを表す特徴量空間を用いる)

$$\min_{\mu:\ \text{prob. measure on } \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell\Big(\mathbb{E}_\mu[h_\theta(x_i)], y_i\Big) + \lambda \mathrm{KL}(\mu \| \nu_0)$$

ν_0: $\mathcal{H}$上のガウス測度．対応する再生核ヒルベルト空間を$\mathcal{H}_K$とする．
（平均0, 分散共分散オペレーターΣ, i.e., $\mathcal{H}_K := \Sigma^{1/2}\mathcal{H}$）

<u>我々の提案手法</u>
粒子双対座標平均化法 (PDA)，粒子双対座標降下法 (PSDCA)
＋
無限次元勾配ランジュバン動力学

[Nishikawa, Suzuki, Nitanda, Wu: Two-layer neural network on infinite-dimensional data: global optimization guarantee in the mean-field regime. NeurIPS2022]

収束解析

- 無限次元版PDA

Theorem

f^*: 最適解に対応するニューラルネットワーク

η: 内部ループのステップサイズ　K: 内部ループの更新回数

$$\|f^* - \hat{f}^{(T)}\|_\infty \lesssim \exp(-\Lambda\eta K) + \eta^{1/2-\kappa} + \frac{\mathrm{KL}(\pi^*\|\nu)}{T} + \frac{1}{\sqrt{M}}$$

T: 外部ループの更新回数　M: 粒子数 (ニューラルネットワークの横幅)

Λ: 内部反復のスペクトルギャップ

- 無限次元空間では測度の絶対連続性が簡単に崩れるのでうまく対処する必要があった．
- 弱収束のみを用いて収束を保証

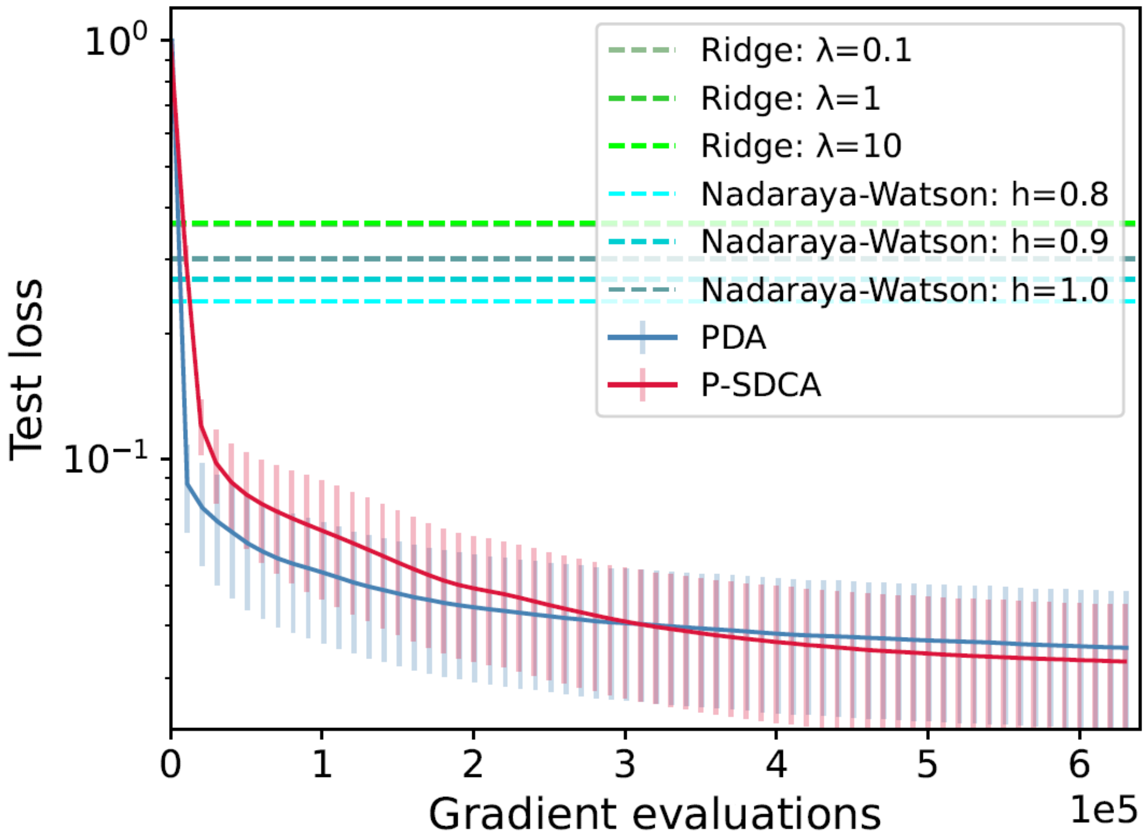
- 無限次元版P-SDCA　(n: sample size)

Theorem

$$(\text{Duality gap}) \lesssim \exp(-\Lambda\eta K) + \eta^{1/2-\kappa} + \exp\left[-\left(n + \frac{1}{\lambda}\right)^{-1} T\right] + \frac{1}{M}$$

数値実験

- 真の関数：１ニューロンの関数
- 比較対象：ヒルベルト空間内でのリッジ回帰 (線形関数)，Nadaraya-Watson推定量 (非線形関数)



研究の流れ

平均場NNの線形収束
連続時間・無限粒子
[Nitanda, Wu, Suzuki (AISTATS2022)]
[Chizat (2022)]

時間・空間離散化：「二重ループの手法」

- PDA法 [Nitanda, Wu, Suzuki: NeurIPS2021]
- P-SDCA法 [Oko, Suzuki, Wu, Nitanda: ICLR2022]
- 無限次元拡張 [Nishikawa, Suzuki, Nitanda: NeurIPS2022]

難しい：**Propagation of chaos**
(McKean, Kac,..., 60年代より)

空間離散化・連続時間:
Uniform-in-time propagation of chaos

- Super対数Sobolev不等式
  [Suzuki, Nitanda, Wu (ICLR2023)]
- Leave-one-out型評価
  [Chen, Ren, Wang (arXiv2022)]

時間・空間離散化・<u>確率的勾配</u>：
「<u>一重ループの手法</u>」

Suzuki, Wu, Nitanda
(arXiv:2306.07221)

[Suzuki, Wu, Nitanda: Convergence of mean-field Langevin dynamics: Time and space discretization, stochastic gradient, and variance reduction. arXiv:2306.07221, 2023]

一重ループの方法

$$\mathrm{d}X_t = -\nabla \frac{\delta F(\mu_t)}{\delta \mu}(X_t)\mathrm{d}t + \sqrt{2\lambda_2}\mathrm{d}B_t$$

$$\Downarrow$$

(時間離散化)

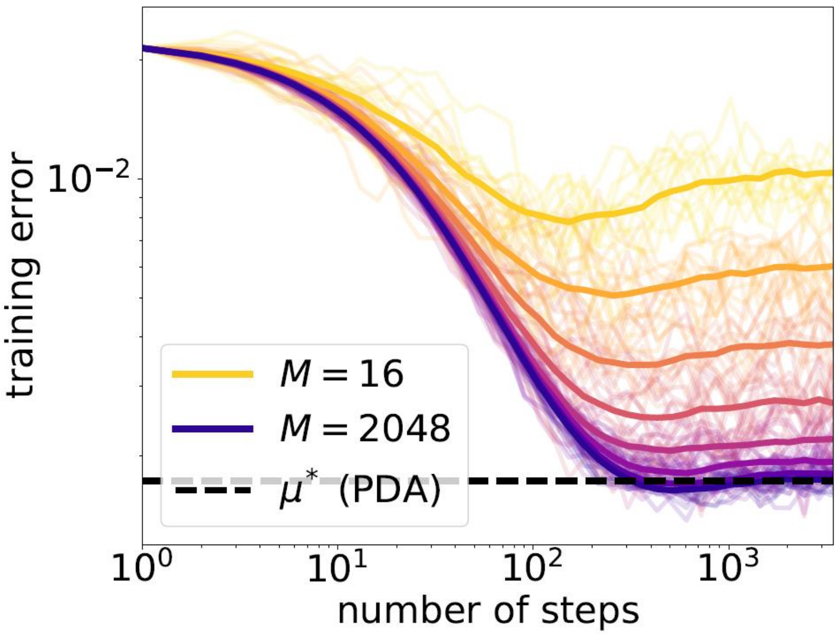
$$X_{k+1}^{(i)} = X_k^{(i)} - \eta v_k^i + \sqrt{2\eta\lambda_2}\xi_k^{(i)}$$

ただし $\mathbb{E}[v_k^i] = \nabla \frac{\delta F(\hat{\mu}_k)}{\delta \mu}(X_k^i)$ かつ $\hat{\mu}_k = \frac{1}{N}\sum_{i=1}^N \delta_{X_k^{(i)}}$

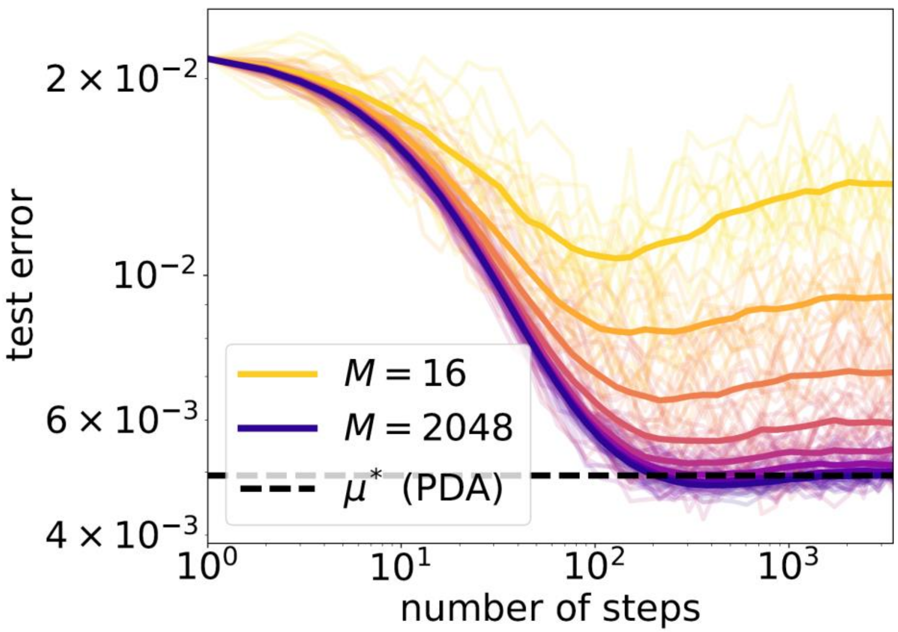
(確率的勾配)　(空間離散化)

- 時間離散化: $X_t \to X_k^{(i)}$
- 空間離散化: N粒子で近似 $(\hat{\mu}_k)$ [もっとも難しい]
- 確率的勾配: 勾配計算を軽量化 $\left(\nabla \frac{\delta F(\mu)}{\delta \mu}(x) = \mathbb{E}[v_k(x;\mu)]\right)$

Numerical experiment



Training error with $r(x) = \|x\|^4$



Test error with $r(x) = \|x\|^2$

Decomposition of objective

Continuous

$$\mathrm{d}X_t = -\nabla \frac{\delta F(\mu_t)}{\delta \mu}(X_t)\mathrm{d}t + \sqrt{2\lambda_2}\mathrm{d}B_t$$

$$\mu_t$$

Discrete

$$\mathrm{d}\hat{X}_t^i = -\nabla \frac{\delta F\left(\frac{1}{N}\sum_{j=1}^N \delta_{\hat{X}_t^j}\right)}{\delta \mu}(\hat{X}_t^i)\mathrm{d}t + \sqrt{2\lambda_2}\mathrm{d}B_t^i$$

$$\mu_t^N = \frac{1}{N}\sum_{j=1}^N \delta_{\hat{X}_t^j}$$

$\mathcal{U}(\mu)$: smooth objective (e.g., $\mathcal{U}(\mu) = \mathbb{E}[(f_\mu - f_{\mu^*})^2]$)

Decomposition of objective:

$$\mathcal{U}(\mu_t^N) - \mathcal{U}(\mu^*) = \underbrace{\mathcal{U}(\mu_t^N) - \mathcal{U}(\mu_t)}_{\text{(1) Discretization error (propagation of chaos)}} + \underbrace{\mathcal{U}(\mu_t) - \mathcal{U}(\mu^*)}_{\text{(2) Geometric ergodicity}}$$

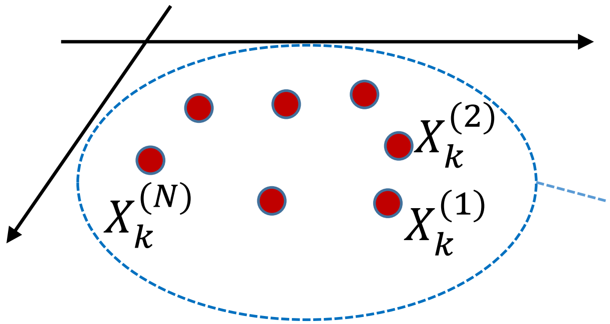
- Naïve evaluation gives **<u>exponential growth as *t*</u>**:

$$\exp(t)\, N^{-1}$$

[Mei et al. (2018, Theorem 3)]

- Otherwise, weak interaction/strong regularization has been required.

Uniform log-Sobolev inequality



$\mathscr{X}_k = \left(X_k^{(i)}\right)_{i=1}^{N} \sim \mu_k^{(N)}$: Joint distribution of N particles.

Potential of the joint distribution $\mu_k^{(N)}$ on $\mathbb{R}^{d\times N}$:

$$\mathscr{L}^N(\mu_k^{(N)}) = N\mathbb{E}_{\mathscr{X}\sim\mu_k^{(N)}}[F(\hat{\mu}_{\mathscr{X}})] + \lambda_2 \mathrm{Ent}(\mu_k^{(N)}).$$

where $\hat{\mu}_{\mathscr{X}} = \frac{1}{N}\sum_{i=1}^{N}\delta_{X^{(i)}}$ $\quad (\mathscr{X} = (X^{(i)})_{i=1}^{N})$

➢ The finite particle dynamics is the Wasserstein gradient flow that minimizes $\mathscr{L}^N$

(Approximate) Uniform log-Sobolev inequality [Chen et al. 2022]

For any N,

$$\frac{1}{N}\mathscr{L}^N(\mu_k^{(N)}) - \mathcal{L}(\mu^*) \le \frac{\lambda_2}{2\alpha}\left(\frac{1}{N} I(\mu_k^{(N)}||p^{(N)})\right) + \frac{C_{\lambda_2}}{N\lambda_2\alpha}$$

(Fisher divergence)

where $p^{(N)}(\mathscr{X}) \propto \exp(-\frac{N}{\lambda_2}F(\hat{\mu}_{\mathscr{X}}))$

Recall $\mathcal{L}(\mu) = F(\mu) + \lambda_2 \mathrm{Ent}(\mu)$

[Chen, Ren, Wang. Uniform-in-time propagation of chaos for mean field langevin dynamics. arXiv:2212.03050, 2022.]

収束解析

$$p_\mu(x) \propto \exp\left(-\frac{1}{\lambda_2}\frac{\delta F(\mu)}{\delta \mu}(x)\right)$$: proximal Gibbs measure

定理 (1ステップ更新の減少; Suzuki, Wu, Nitanda (2023))

p_μは対数Sobolev不等式を定数αで満たすとする.
損失関数の凸性と平滑性の仮定のもと,

$$\mathscr{L}^{(N)}(\hat{\mu}_{k+1}) - \mathcal{L}(\mu^*)$$
$$\leq \exp(-\lambda_2 \eta_k \alpha)\left(\mathscr{L}^{(N)}(\hat{\mu}_k) - \mathcal{L}(\mu^*)\right)$$
$$+ C\left(\underbrace{\eta_k^3 + \lambda_2 \eta_k^2}_{\text{時間離散化}} + \underbrace{\frac{\eta_k}{N}}_{\text{空間離散化}} + \underbrace{\eta_k^{\frac{3}{2}}\lambda_2^{\frac{1}{2}}\sigma_k\tilde{\sigma}_k}_{\text{確率的勾配}}\right)$$

$$\begin{cases} \sigma_k^2 = \max_i \mathbb{E}\left[\|v_k^i - \mathbb{E}[v_k^i]\|^2\right] \\ \tilde{\sigma}_k^2 = \max_i \mathbb{E}\left[\left\|\nabla v_k^{i\top}(\mathscr{X}) - \nabla\nabla^\top \frac{\delta F(\mu_{\mathscr{X}})}{\delta \mu}(X^i)\right\|_{\mathrm{op}}^2\right] \end{cases}$$

既存研究では粒子数は時間に対して指数関数的に依存
[Mei et al., 2018; Javanmard et al., 2019; De Bortoli et al., 2020]

Assumption:
1. $F: \mathcal{P} \to \mathbb{R}$ is convex and has a form of $F(\mu) = L(\mu) + \lambda_1 \mathbb{E}_\mu[\|x\|^2]$.
2. (smoothness) $\left\|\nabla \frac{\delta L(\mu)}{\delta \mu}(x) - \nabla \frac{\delta L(\nu)}{\delta \mu}(y)\right\| \leq C(W_2(\mu,\nu) + \|x-y\|)$ and
   (boundedness) $\left\|\nabla \frac{\delta L(\mu)}{\delta \mu}(x)\right\| \leq R$.

Log Sobolev for Lipschitz cont. obj

Proximal Gibbs measure:

$$p_\mu(x) \propto \exp\left(-\frac{1}{\lambda_2}\frac{\delta F(\mu)}{\delta \mu}(x)\right) \qquad p_\mu = \arg\min_{\nu \in \mathcal{P}} (\nu - \mu)\frac{\delta F(\mu)}{\delta \mu} + \lambda_2 \mathrm{Ent}(\nu)$$

Assumption: $F(\mu) = L(\mu) + \lambda_1 \mathbb{E}_\mu[\|x\|^2]$

μ satisfies the LSI if there exits $\alpha > 0$ such that for any ϕ s.t. $\mu(\phi^2) = 1$, it holds that

$$\mu(\phi^2 \log(\phi^2)) \leq \frac{2}{\alpha} \int \|\nabla \phi\|^2 \mathrm{d}\mu$$

1. Holley—Strook argument: [Bakry & Emery, 1985; Holley & Stroock, 1987]

$$\left\|\frac{\delta L(\mu)}{\delta \mu}\right\|_\infty \leq R \qquad \Rightarrow \qquad \alpha \geq \frac{\lambda_1}{\lambda_2}\exp\left(-\frac{4R}{\lambda_2}\right)$$

(New)

2. Lipschitz perturbation argument + Miclo's trick:

[Cattiaux & Guillin, 2022; Bardet et al., 2018]

$\sup_x \left\|\nabla \frac{\delta L(\mu)}{\delta \mu}(x)\right\| \leq R$ (Lipschitz continuous)

$$\Rightarrow \qquad \alpha \geq \frac{\lambda_1}{2\lambda_2}\exp\left(-\frac{4R^2}{\lambda_1 \lambda_2}\sqrt{2d/\pi}\right) \vee$$

$$\left\{\frac{4\lambda_2}{\lambda_1} + e^{\frac{R^2}{2\lambda_1\lambda_2}}\left(\frac{R}{\lambda_1} + \sqrt{\frac{2\lambda_2}{\lambda_1}}\right)^2\left[2 + d + \frac{d}{2}\log\left(\frac{\lambda_2}{\lambda_1}\right) + 4\frac{R^2}{\lambda_1\lambda_2}\right]\right\}^{-1}$$

確率的勾配の計算量

SGD-MFLD:

$F(\mu) = \frac{1}{n}\sum_{j=1}^{n} f_j(\mu)$ (有限和),

$v_k^i = \frac{1}{B}\sum_{j\in I_k} \frac{\delta f_j(\hat{\mu}_k)}{\delta \mu}(X_k^i)$ (確率的勾配)

(Mini-batch size = B)

$$\boxed{\mathscr{L}^{(N)}(\hat{\mu}_k) - \mathcal{L}(\mu^*) \lesssim \exp(-\lambda_2 \eta k \alpha) + \frac{1}{\alpha\lambda_2}\left(\underbrace{\eta^2 + \lambda_2\eta}_{\text{時間離散化}} + \underbrace{\frac{1}{N}}_{\text{空間離散化}} + \underbrace{\frac{(n-B)\sqrt{\eta\lambda_2}}{B(n-1)}}_{\text{確率的勾配}}\right)}$$

<u>更新回数のバウンド:</u>

By setting $\eta = O\left(\epsilon\alpha \wedge (\lambda_2\epsilon\alpha)^2 \frac{B^2(n-1)^2}{(n-B)^2\lambda_2} \wedge \sqrt{\lambda_2\epsilon\alpha}\right)$,
the iteration complexity becomes

$$k = O\left(\frac{1}{\epsilon\alpha} + \left(\frac{1}{\lambda_2\epsilon\alpha}\right)^2 \frac{\lambda_2(n-B)^2}{B^2(n-1)^2} + \sqrt{\frac{1}{\lambda_2\alpha\epsilon}}\right)\frac{1}{\lambda_2\alpha}\log(\epsilon^{-1})$$

to achieve $\epsilon + O(1/(\lambda_2\alpha N))$ accuracy.

- $B = n \wedge \sqrt{1/(\lambda_2\alpha\epsilon)}$ is the optimal mini-batch size. $\to k = O(\log(\epsilon^{-1})/\epsilon)$.

分散縮小勾配法

SVRG-MFLD:

$F(\mu) = \frac{1}{n}\sum_{j=1}^{n} f_j(\mu)$ (有限和),

$$v_k^i = \frac{1}{B}\sum_{j\in I_k} \nabla \frac{\delta f_j(\hat{\mu}_k)}{\delta\mu}(X_k^{(i)}) - \frac{1}{B}\sum_{j\in I_k} \nabla \frac{\delta f_j(\dot{\mu})}{\delta\mu}(\dot{X}^{(i)}) + \nabla \frac{\delta F(\dot{\mu})}{\delta\mu}(\dot{X}^{(i)})$$

(分散縮小勾配)
($\dot{X}$はm回に一回更新)

$$\begin{aligned}
&\mathscr{L}^{(N)}(\hat{\mu}_k) - \mathcal{L}(\mu^*) \\
&\lesssim \exp(-\lambda_2 \eta k \alpha) \\
&\quad + \frac{1}{\lambda_2\alpha}\left(\eta^2 + \lambda_2\eta + \frac{1}{N} + \frac{n-B}{B(n-1)}\lambda_2^{1/2}\eta\sqrt{m(\eta+\lambda_2)}\right)
\end{aligned}$$

時間離散化　空間離散化　確率的勾配の誤差

線形GLDの既存解析 [Kinoshita, Suzuki: NeurIPS2022] の非線形への拡張/改善

更新回数： $\eta = \epsilon\alpha \wedge \sqrt{\lambda_2\alpha\epsilon},$

$$k = \frac{1}{\lambda_2\alpha\eta}\log(1/\epsilon) = O\left(\frac{1}{\epsilon\alpha} + \sqrt{\frac{1}{\lambda_2\alpha\epsilon}}\right)\frac{1}{\lambda_2\alpha}\log(\epsilon^{-1})$$

ただし$B = \sqrt{m} = n^{1/3}$.

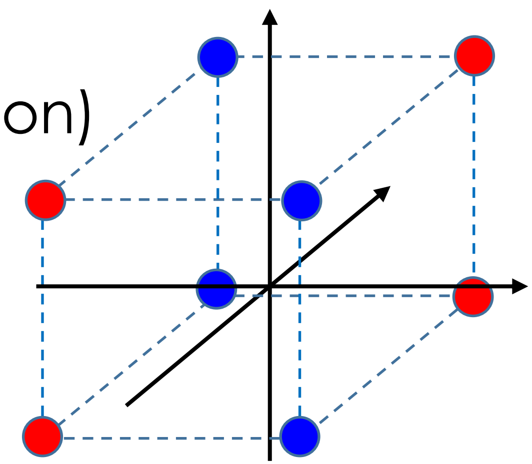
総勾配計算回数： $Bk + \frac{nk}{m} \lesssim n^{1/3}\left(\frac{1}{\alpha\epsilon} + \sqrt{\frac{1}{\lambda_2\alpha\epsilon}}\right)\frac{1}{\lambda_2\alpha}\log(\epsilon^{-1}).$

$\sqrt{n}$ in Kinoshita&Suzuki (2022)

統計的性質

- ℓ_i: ロジスティック損失
- $h_z(x) = \bar{R} \cdot [\tanh(\langle x_1, z\rangle + x_2) + x_3]/2$

- k-スパースパリティ問題
  - ➢ $X \sim \mathrm{Unif}(\{-1,1\}^d)$ (up to freedom of rotation)
  - ➢ $Y = X_{i_1}X_{i_2} \dots X_{i_k}$ for $i_j \in [d]$ with $i_j \neq i_l$.



Q: この問題設定でカーネル法を上回る？

A: Yes.

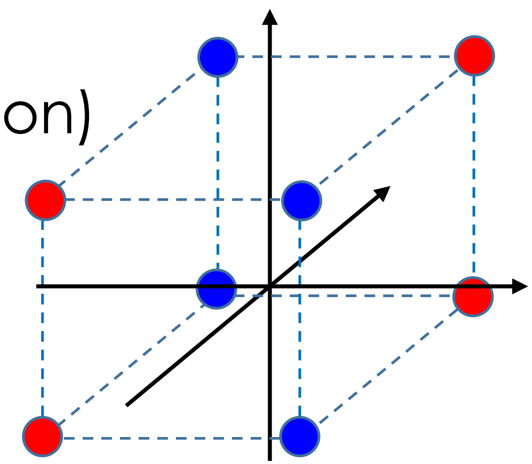
[Suzuki, Wu, Oko, Nitanda: Feature learning via mean-field Langevin dynamics: classifying sparse parities and beyond. 2023]

Authors	regime/method	width	class error	number of iterations
Ji and Telgarsky (2019)	NTK/SGD	d^8	d^2/n	d^2/ϵ
Telgarsky (2023)	NTK/SGD	d^2	d^2/n	d^2/ϵ
Barak et al. (2022)*	Two phase SGD	$O(1)$	$d^{(k+1)/2}/\sqrt{n}$	d/ϵ^2
Telgarsky (2023)	mean-field/GF	d^d	d/n	∞
Wei et al. (2019)	mean-field/WF	∞	d/n	∞
Ours*	mean-field/MFLD	$e^{O(d)}$	$\exp(-O(\sqrt{n}/d))$	$e^{O(d)}$
Ours*	mean-field/MFLD	$e^{O(d)}$	d/n	$e^{O(d)}$

統計的性質

- ℓ_i: ロジスティック損失
- $h_z(x) = \bar{R} \cdot [\tanh(\langle x_1, z\rangle + x_2) + x_3]/2$

- k-スパースパリティ問題
  - ➢ $X \sim \mathrm{Unif}(\{-1,1\}^d)$ (up to freedom of rotation)
  - ➢ $Y = X_{i_1} X_{i_2} \ldots X_{i_k}$ for $i_j \in [d]$ with $i_j \neq i_l$.



Q: この問題設定でカーネル法を上回る？

A: Yes.

[Suzuki, Wu, Oko, Nitanda: Feature learning via mean-field Langevin dynamics: classifying sparse parities and beyond. 2023]

特徴学習によって次元への依存性が改善されている.

Authors	regime/method	width	class er	
Ji and Telgarsky (2019)	NTK/SGD	d^8	$d^2/$	
Telgarsky (2023)	NTK/SGD	d^2	d^2/n	d^2/ϵ
Barak et al. (2022)*	Two phase SGD	$O(1)$	$d^{(k+1)/2}/\sqrt{n}$	d/ϵ^2
Telgarsky (2023)	mean-field/GF	d^d	d/n	∞
Wei et al. (2019)	mean-field/WF	∞	d/n	∞
Ours*	mean-field/MFLD	$e^{O(d)}$	$\exp(-O(\sqrt{n}/d))$	$e^{O(d)}$
Ours*	mean-field/MFLD	$e^{O(d)}$	d/n	$e^{O(d)}$

補足資料
「拡散モデル」

DALL·E/DALL·E 2

文章による説明から画像を生成するモデル

「An astronaut riding a horse in a photorealistic style」

DALL·E: [Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, Ilya Sutskever: Zero-Shot Text-to-Image Generation. ICML2021.]
DALL·E2:[Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen: Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125]

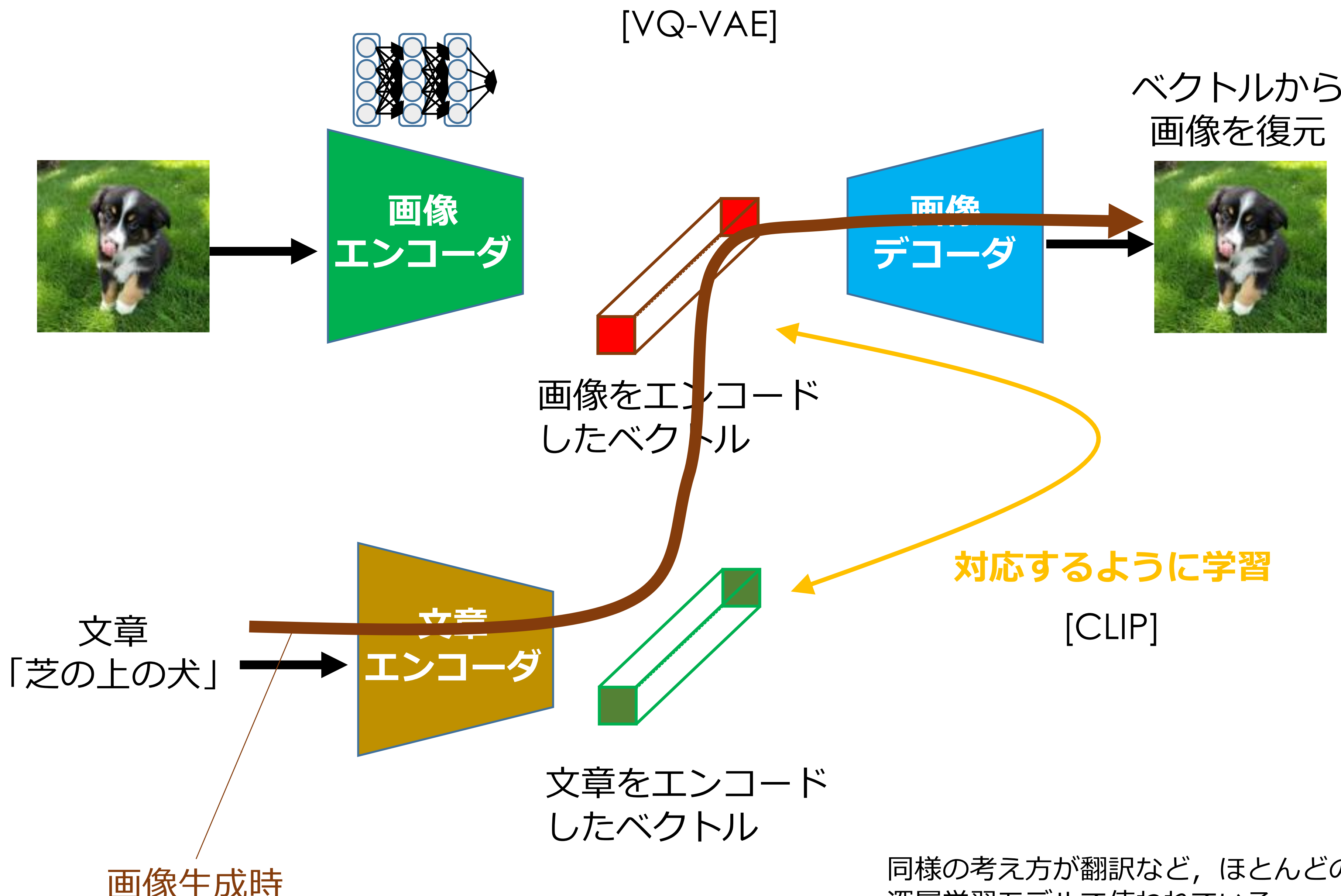
「Teddy bears shopping for groceries in the style of ukiyo-e」

他の作例

Jason Allen "Théâtre D'opéra Spatial“ generated by **Midjourney**. Colorado State Fair’s fine art competition, 1st prize in digital art category

Generated by NovelAI

文章での条件付け



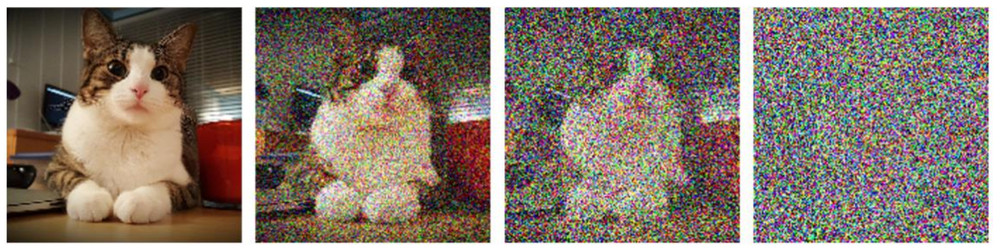
同様の考え方が翻訳など，ほとんどの
深層学習モデルで使われている．

デコーダー：拡散モデル

[Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Song et al., 2020; Ho et al., 2020; Vahdat et al., 2021]

順過程：所望の分布を正規分布に変換していく (OU-過程).

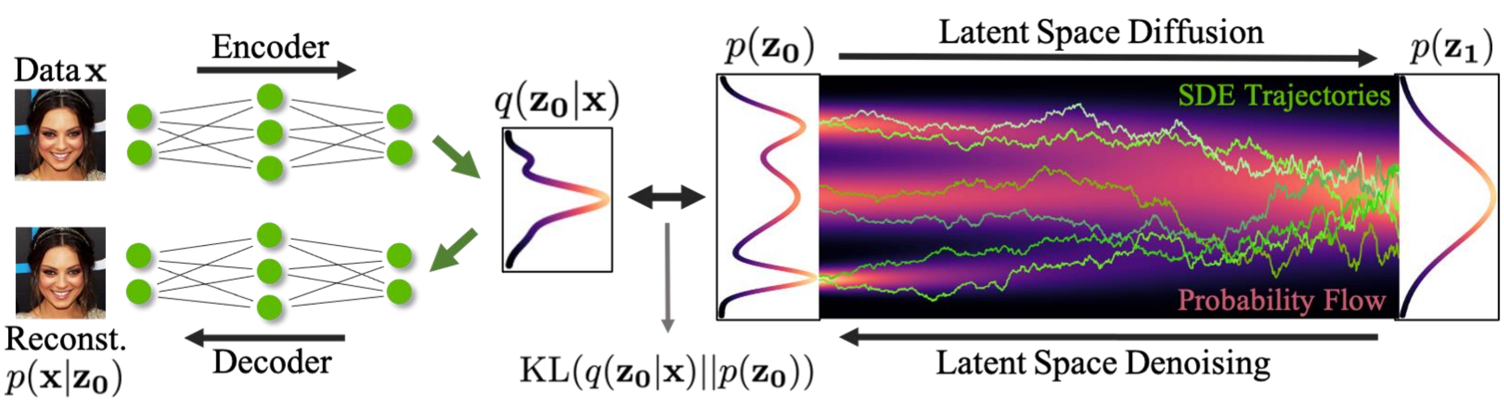
$$\mathrm{d}X_t = -X_t \mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$



$$\mathrm{d}Y_t = (Y_t + 2\nabla \log(p_{\overline{T}-t}(Y_t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

$(Y_t \sim X_{\overline{T}-t})$

逆過程：正規分布 (ノイズの分布) から逆にたどって所望の分布に逆変換していく.



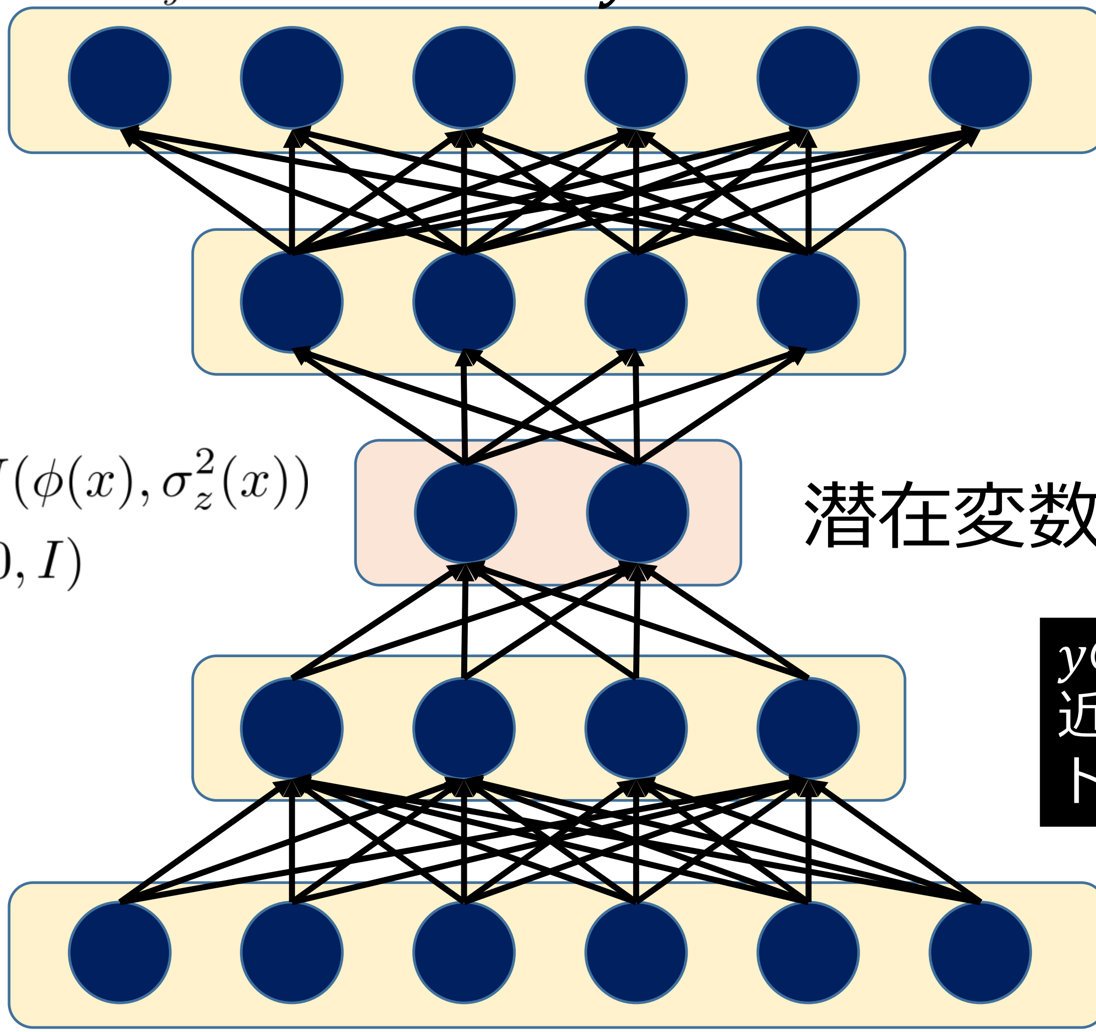
[Vahdat, Kreis, Kautz: Score-based Generative Modeling in Latent Space. arXiv:2106.05931]

変分オートエンコーダとの関係

[Kingma, Welling: Auto-Encoding Variational Bayes. 2014.]

$p(y|z) = N(\psi(z), \sigma_y^2(z))$

出力：y



$p(z|x) = N(\phi(x), \sigma_z^2(x))$

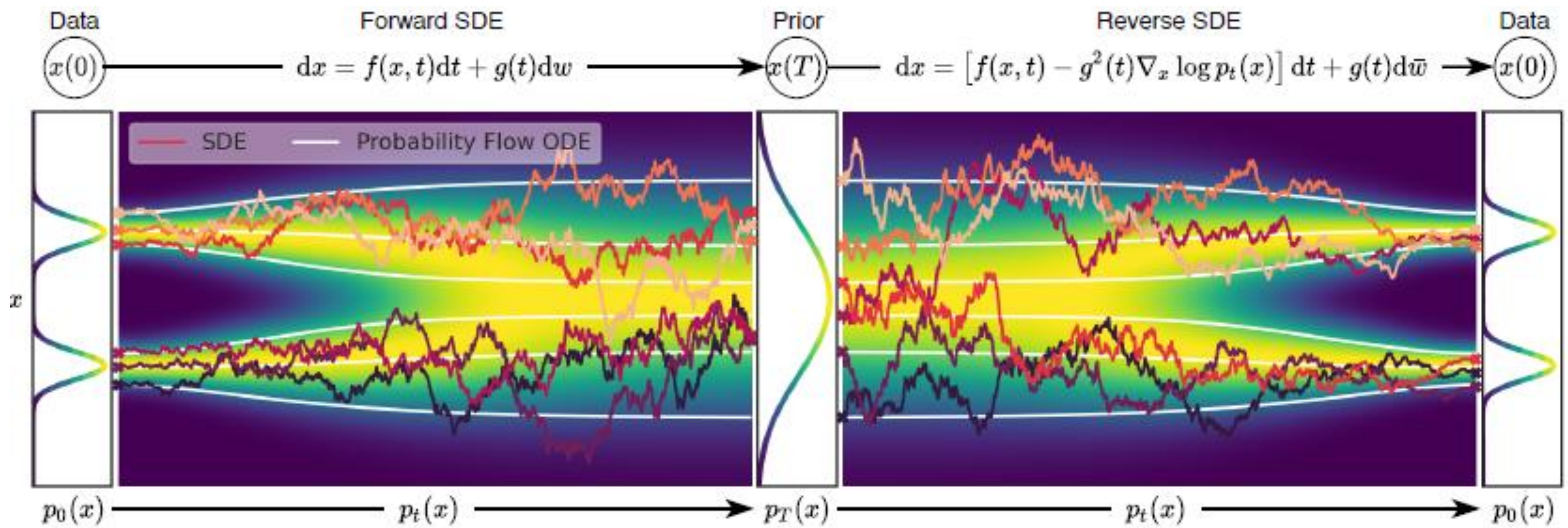
$p(z) = N(0, I)$

潜在変数：z

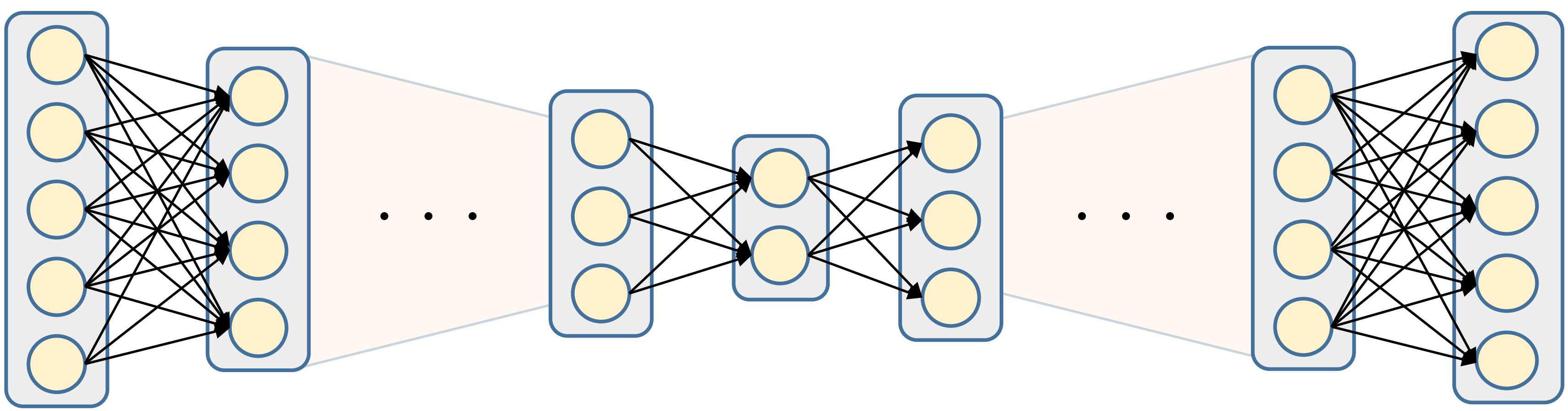
yの分布がxの分布に近くなるようにネットワークを学習

入力：x

[Song et al.: SCORE-BASED GENERATIVE MODELING THROUGH STOCHASTIC DIFFERENTIAL EQUATIONS. ICLR2021.

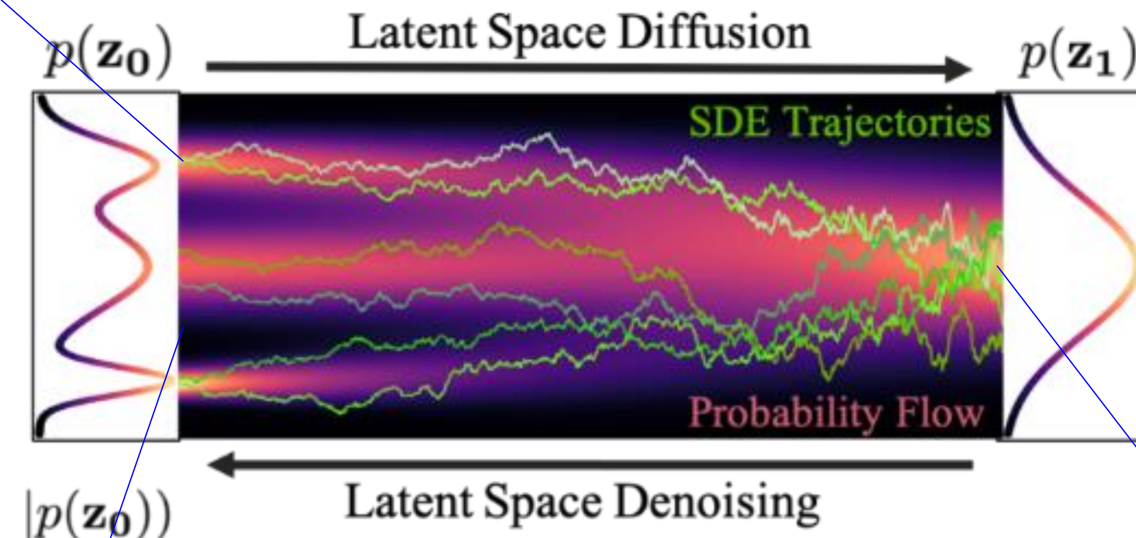


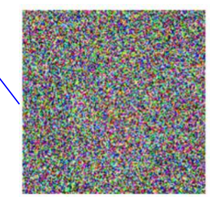
元の分布　　潜在変数(ノイズ)の分布 (正規分布)　　復元された分布



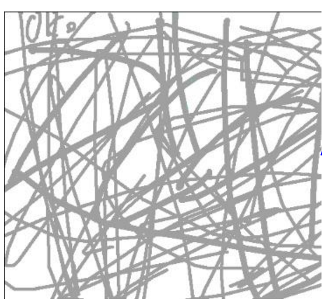
VAE

出やすい（自然な画像）





正規分布



出にくい（不自然な画像）

トイデータ例

恐竜型の分布を再現

[https://github.com/Kei18/tiny-tiny-diffusion]

注意：こうやって出てきた恐竜が生成された画像なのではなくて，
各座標が一つの画像に対応．

順過程

順過程:

$$\mathrm{d}X_t = -X_t \mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

OU-過程

形がわかっている！
x_0が与えられればx_tのサンプリングも可能

p_tをX_tの確率密度関数とする.

$$p_t(x) = \int p_0(y) \frac{1}{\sqrt{\sigma_t^{2d}(2\pi)^d}} \exp\left(-\frac{\|x-\mu_t y\|^2}{2\sigma_t^2}\right) \mathrm{d}y$$

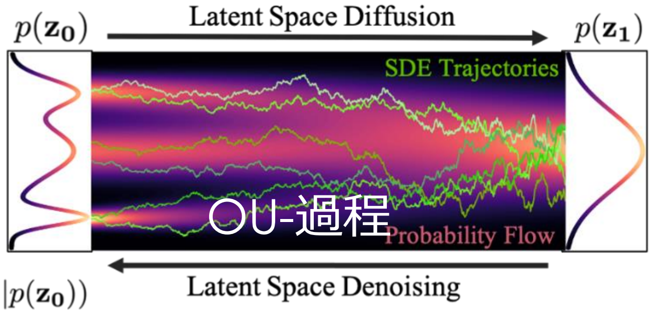
ただし，$\mu_t = \exp(-t)$, $\sigma_t^2 = 1 - \exp(-2t)$.

$$p_t = \int N(\mu_t x_0, \sigma_t^2 I) p_0(x_0) \mathrm{d}x_0$$

前回講義資料より，順過程は指数関数的に標準正規分布に近づく.

$$\mathrm{KL}(p_t \| N(0,I)) \le \exp(-2t)\mathrm{KL}(p_0 \| N(0,I))$$

元の分布



標準正規分布

[Vahdat, Kreis, Kautz: Score-based Generative Modeling in Latent Space. arXiv:2106.05931]

逆過程

逆過程:

$$Y_0 \sim p_{\bar{T}}$$

$$\mathrm{d}Y_t = (Y_t + 2\nabla \log(p_{\bar{T}-t}(Y_t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \qquad (t \in [0, \bar{T}])$$

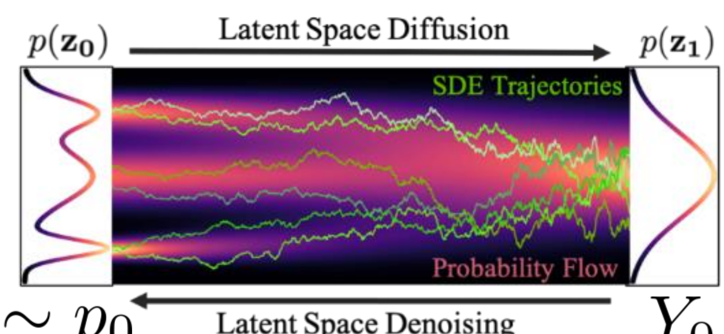
事実：Y_tの分布=$X_{\bar{T}-t}$の分布

[Haussmann & Pardoux, 1986]

すなわち，$Y_t \sim p_{\bar{T}-t}$

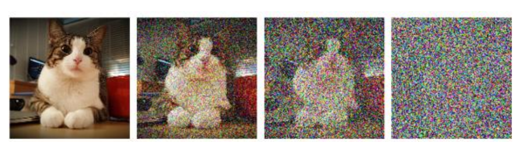
順過程を逆にたどることによって，((ほぼ)正規分布に従うノイズを徐々に修正して元の画像の分布を再現できる．

こっちで終わる



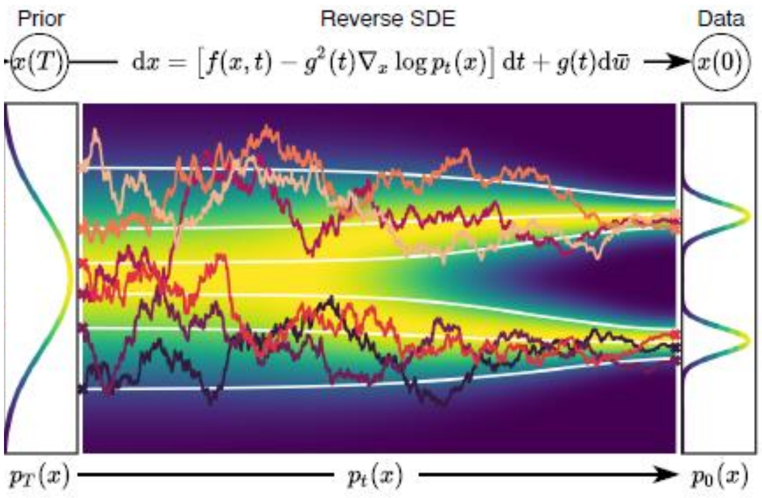
こっちから始める

$Y_{\bar{T}} \sim p_0$ $\qquad$ $Y_0 \sim p_{\bar{T}}$

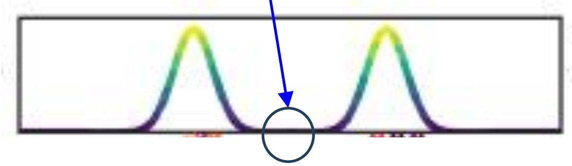


拡散モデルの利点

- 多峰な分布からのサンプリングがしやすい.
  - ➢「簡単な分布」→「難しい分布」へと変化していくことで偏りなくサンプリングできる.

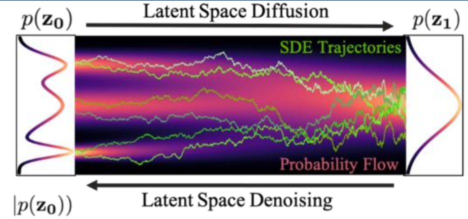


直接ターゲットの分布のスコアを推定してからGLDなどでサンプリングしようとすると谷を乗り越えられない.



- 元の分布のスコアは複雑でも，拡散させたX_tの分布は滑らか→推定しやすい→汎化しやすい.
- ノイズから元分布への写像を直接End-to-endで学習するのではなく中間的な分布p_tの情報を用いるので学習が安定する.

スコアの推定



逆過程:

(未知)

$$Y_0 \sim p_{\overline{T}}$$

(未知)

$$\mathrm{d}Y_t = (Y_t + 2\underline{\nabla \log(p_{\overline{T}-t}(Y_t))})\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \qquad (t \in [0, \overline{T}])$$

$$\Rightarrow Y_t \sim p_{\overline{T}-t}$$

[Haussmann & Pardoux, 1986]

<u>近似モデル (生成モデル):</u>

($p_{\overline{T}}$は$N(0, I)$に十分近い)

$$\hat{Y}_0 \sim N(0, I)$$

$$\mathrm{d}\hat{Y}_t = (\hat{Y}_t + 2\hat{s}(\hat{Y}_t, \overline{T} - t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \qquad (t \in [0, \overline{T}])$$

$\hat{Y}_{\overline{T}}$を生成画像として用いる.

定理 (Girsanov's theorem)

If $\hat{Y}_0 \sim p_{\overline{T}}$, then

$$\mathrm{KL}(p_0 || p_{\hat{Y}_{\overline{T}}}) \leq \frac{1}{4}\int_0^{\overline{T}} \mathbb{E}_{Y_t}[\|\nabla \log(p_{\overline{T}-t}(Y_t)) - \hat{s}(Y_t, \overline{T} - t)\|^2]\mathrm{d}t$$

$\Rightarrow$ <u>スコア関数$\nabla\log(p_t)$</u>をできるだけ正確に推定できれば良い.

スコアマッチング

$$\int_0^{\overline{T}} \mathbb{E}_{Y_t}[\|\underline{\nabla \log(p_{\overline{T}-t}(Y_t))} - \hat{s}(Y_t, \overline{T}-t)\|^2]\mathrm{d}t$$

未知，計算できない．計算できるものに置き換えたい．

$$= \int_0^{\overline{T}} \mathbb{E}_{X_t}[\|\nabla \log(p_t(X_t)) - \hat{s}(X_t, t)\|^2]\mathrm{d}t$$

($X_{\bar{T}-t}$とY_tは同じ分布)

$$= \int_0^{\overline{T}} \mathbb{E}_{X_t}[\|\nabla \log(p_t(X_t))\|^2 - 2\langle \nabla \log(p_t(X_t)), \hat{s}(X_t, t)\rangle + \|\hat{s}(X_t, t)\|^2]\mathrm{d}t$$

$$= \int_0^{\overline{T}} \mathbb{E}_{X_t}\left[\underbrace{-2\left\langle \frac{\nabla p_t(X_t)}{p_t(X_t)}, \hat{s}(X_t, t)\right\rangle + \|\hat{s}(X_t, t)\|^2}\right]\mathrm{d}t + (\mathrm{const})$$

$$\int -2\left\langle \frac{\nabla p_t(x_t)}{\cancel{p_t(x_t)}}, \hat{s}(x_t, t)\right\rangle \cancel{p_x(x_t)}\mathrm{d}x_t + \mathbb{E}\left[\|\hat{s}(X_t, t)\|^2\right]$$

$$= \int -2\left\langle \nabla p_t(x_t), \hat{s}(x_t, t)\right\rangle \mathrm{d}x_t + \mathbb{E}\left[\|\hat{s}(X_t, t)\|^2\right]$$

$$= \int -2\left\langle \nabla \int p_t(x_t|x_0)p_0(x_0)\mathrm{d}x_0, \hat{s}(x_t, t)\right\rangle \mathrm{d}x_t + \mathbb{E}\left[\|\hat{s}(X_t, t)\|^2\right]$$

$$= \int\int -2\left\langle \nabla p_t(x_t|x_0), \hat{s}(x_t, t)\right\rangle p_0(x_0)\mathrm{d}x_0\mathrm{d}x_t + \mathbb{E}\left[\|\hat{s}(X_t, t)\|^2\right]$$

$$= \int\int -2\left\langle \frac{\nabla p_t(x_t|x_0)}{\underline{p_t(x_t|x_0)}}, \hat{s}(x_t, t)\right\rangle \underline{p_t(x_t|x_0)}p_0(x_0)\mathrm{d}x_0\mathrm{d}x_t + \mathbb{E}\left[\|\hat{s}(X_t, t)\|^2\right]$$

$$= \mathbb{E}_{X_0, X_t}\left[-2\left\langle \nabla \log(p_t(X_t|X_0)), \hat{s}(X_t, t)\right\rangle + \|\hat{s}(X_t, t)\|^2\right]$$

スコアマッチング

$$\int_0^{\overline{T}} \mathbb{E}_{Y_t}[\|\nabla \log(p_{\overline{T}-t}(Y_t)) - \hat{s}(Y_t, \overline{T}-t)\|^2]\mathrm{d}t$$

$$= \int_0^{\overline{T}} \mathbb{E}_{X_t}[\|\nabla \log(p_t(X_t)) - \hat{s}(X_t, t)\|^2]\mathrm{d}t$$

$$= \int_0^{\overline{T}} \mathbb{E}_{X_t}[\|\nabla \log(p_t(X_t))\|^2 - 2\langle \nabla \log(p_t(X_t)), \hat{s}(X_t, t)\rangle + \|\hat{s}(X_t, t)\|^2]\mathrm{d}t$$

$$= \int_0^{\overline{T}} \mathbb{E}_{X_t}\left[\underbrace{-2\left\langle \frac{\nabla p_t(X_t)}{p_t(X_t)}, \hat{s}(X_t, t)\right\rangle + \|\hat{s}(X_t, t)\|^2}_{\mathbb{E}_{X_0, X_t}\left[-2\left\langle \nabla \log(p_t(X_t|X_0)), \hat{s}(X_t, t)\right\rangle + \|\hat{s}(X_t, t)\|^2\right] \quad \text{(前ページの導出より)}}\right]\mathrm{d}t + (\mathrm{const})$$

$$= \int_0^{\overline{T}} \mathbb{E}_{X_t, X_0}[\|\nabla \log(p_t(X_t|X_0)) - \hat{s}(Y_t, t)\|^2]\mathrm{d}t + (\mathrm{const})$$

スコアマッチング

$$\min_{\hat{s}} \int_0^{\overline{T}} \mathbb{E}_{X_t, X_0}[\|\nabla \log(p_t(X_t|X_0)) - \hat{s}(Y_t, t)\|^2]\mathrm{d}t$$

を解けばよい.

しかし，X_0の分布を知らないのでX_0による期待値は取れない.
→ サンプル平均で代用する (有限データからの学習).

観測値 (nデータ点, $D_n = \{x_i\}_{i=1}^n$):

$$x_i \sim p_0 \quad (i = 1, \dots, n)$$

経験スコアマッチング損失:

$$\min_{s \in \mathrm{DNN}} \frac{1}{n} \sum_{i=1}^n \int_{t=\underline{T}}^{\overline{T}} \mathbb{E}_{X_t|X_0=x_i}\left[\|s(X_t, t) - \underline{\nabla \log p_t(X_t|x_i)}\|^2\right] \mathrm{d}t$$

条件付分布はOU過程からサンプリングできる！

$$N(x_i e^{-t}, 1 - e^{-2t})$$

(正規分布)

陽に求まる！(正規分布の密度より)

$$-\frac{(X_t - e^{-t} x_i)}{1 - e^{-2t}}$$

誤差解析の理論研究

- 拡散モデルの逆向きSDEとしての定式化: Song et al. (2021)

[近似誤差解析]

- KL-divergence bound via Girsanov's theorem: Chen et al. (2022)

- Error bound with LSI: Lee et al. (2022a)
  - With smoothness: Chen et al. (2022) and Lee et al. (2022b)

- Error propagation with manifold assumption: Pidstrigach (2022)

[Generalization analysis]

- Wasserstein dist bound ($n^{-1/d}$) with manifold assumption: De Bortoli (2022)

これまでのまとめ

- 順過程：

  標準正規分布へ向かう勾配ランジュバン動力学

$$\mu_\infty \propto \exp\left(-\frac{\|x\|^2}{2}\right) = \exp(-L(x))$$

$$\Rightarrow L(x) = \frac{\|x\|^2}{2}$$

$$\Rightarrow \ \mathrm{d}X_t = -X_t \mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

（OU-過程）

Fokker-Planck方程式：$\partial_t p_t = \Delta p_t + \nabla \cdot (x p_t)$

$\boxed{\partial_t p_t = \lambda \Delta_x p_t + \nabla \cdot [p_t \nabla L]}$

- 逆過程 (reverse SDE)：

  $\tilde{t}$を逆向き時間として (t = ∞ → $t = 0$へ向かう)

$$\mathrm{d}X_{\tilde{t}} = (X_{\tilde{t}} + 2\nabla \log(p_{\tilde{t}}(x)))\mathrm{d}\tilde{t} + \sqrt{2}\mathrm{d}B_{\tilde{t}}$$

t小　順　t大

$\tilde{t}$大　逆　$\tilde{t}$小

$\mathrm{d}t = -\mathrm{d}\tilde{t}$

$b(x) = -x = -\nabla L(x)$ として次のページで導出.

前向き・後ろ向きFokker-Planck方程式

$$\mathrm{d}X_t = b_t(X_t)\mathrm{d}t + \sigma_t \mathrm{d}B_t$$

$p(t, y|s, x)$: 時刻sでの値を$X_s = x$で条件付けた時刻tにおけるX_tの確率密度.

- 前向き方程式

$$\partial_t p(t, y|s, x) = -\nabla_y(b_t(y)p(t, y|s, x)) + \frac{\sigma_t^2}{2}\Delta_y^2 p(t, y|s, x)$$

(part1より)

- 後向き方程式

$$\partial_s p(t, y|s, x) = -b_s(x)^\top \nabla_x p(t, y|s, x) - \frac{\sigma_s^2}{2}\Delta_x^2 p(t, y|s, x)$$

(後ろ向き方程式の確認)

$$\begin{aligned}
\partial_s p(t, y|s, x) &= \lim_{\delta_s \to 0} \frac{p(t, y|s+\delta_s, x) - \int p(t, y|s+\delta_s, x')p(s+\delta_s, x'|s, x)\mathrm{d}x'}{\delta_s} \\
&= \lim_{\delta_s \to 0} \frac{p(t, y|s+\delta_s, x) - \mathbb{E}_{\xi \sim N(0,I)}[p(t, y|s+\delta_s, x + b_s(x)\delta_s + \sigma_t\sqrt{\delta_s}\xi)]}{\delta_s} \\
&= -\left(b_s^\top \nabla_x + \frac{\sigma_s^2}{2}\Delta_x\right) p(t, y|s, x)
\end{aligned}$$

(∵生成作用素 (part1より))

逆向きSDEの導出

$$\partial_s p(s,x|t,y) = \partial_s\left[p(t,y|s,x)\frac{p(s,x)}{p(t,y)}\right]$$

($s < t$を想定)

$$
\begin{aligned}
&= \partial_s p(t,y|s,x)\frac{p(s,x)}{p(t,y)} + p(t,y|s,x)\frac{\partial_s p(s,x)}{p(t,y)} \\
&= [\underline{-b^\top \nabla_x p(t,y|s,x)} - \underline{\Delta_x p(t,y|s,x)}]\frac{p(s,x)}{p(t,y)} \\
&\quad + p(t,y|s,x)\frac{\underline{-\nabla_x \cdot (bp(s,x))} + \cancel{\Delta_x p(s,x)}}{p(t,y)}
\end{aligned}
$$

$$b^\top \nabla_x p(t,y|s,x)\frac{p(s,x)}{p(t,y)} = \underline{b^\top \nabla_x p(s,x|t,y)} - \cancel{p(t,y|s,x)\frac{b^\top \nabla_x p(s,x)}{p(t,y)}}$$

$$
\begin{aligned}
\Delta_x p(t,y|s,x)\frac{p(s,x)}{p(t,y)} &= \Delta_x p(s,x|t,y) - 2\nabla_x^\top p(t,y|s,x)\frac{\nabla_x p(s,x)}{p(t,y)} - p(t,y|s,x)\frac{\Delta_x p(s,x)}{p(t,y)} \\
&= \Delta_x p(s,x|t,y) - 2\nabla_x^\top p(s,x|t,y)\nabla_x \log(p(s,x)) \\
&\quad + 2p(s,x|t,y)\|\nabla_x p(s,x)\|^2 - p(t,y|s,x)\frac{\Delta_x p(s,x)}{p(t,y)} \\
&= \Delta_x p(s,x|t,y) \underline{- 2\nabla_x^\top p(s,x|t,y)\nabla_x \log(p(s,x))} \\
&\quad \underline{- 2p(s,x|t,y)\Delta_x \log(p(s,x))} + \cancel{p(t,y|s,x)\frac{\Delta_x p(s,x)}{p(t,y)}}
\end{aligned}
$$

$$p(t,y|s,x)\frac{-\nabla_x \cdot (bp(s,x))}{p(t,y)} = -(\underline{\nabla_x \cdot b} + \cancel{b^\top \nabla_x \log(p(s,x))})p(s,x|t,y)$$

まとめると，

$$\partial_s p(s,x|t,y) = -\nabla_x \cdot [(b - 2\nabla_x \log(p(s,x)))p(s,x|t,y)] - \Delta_x p(s,x|t,y)$$

時間を反転させて，$\mathrm{d}\tilde{s} \leftarrow -\mathrm{d}s$とすると，

$$\partial_{\tilde{s}} p(\tilde{s},x|t,y) = \nabla_x \cdot [(b - 2\nabla_x \log(p(\tilde{s},x)))p(\tilde{s},x|t,y)] + \Delta_x p(\tilde{s},x|t,y)$$

これはドリフト項が

$$-\Big(b - 2\nabla_x \log\big(p(s,x)\big)\Big) = x + 2\nabla_x \log(\mu_s(x))$$

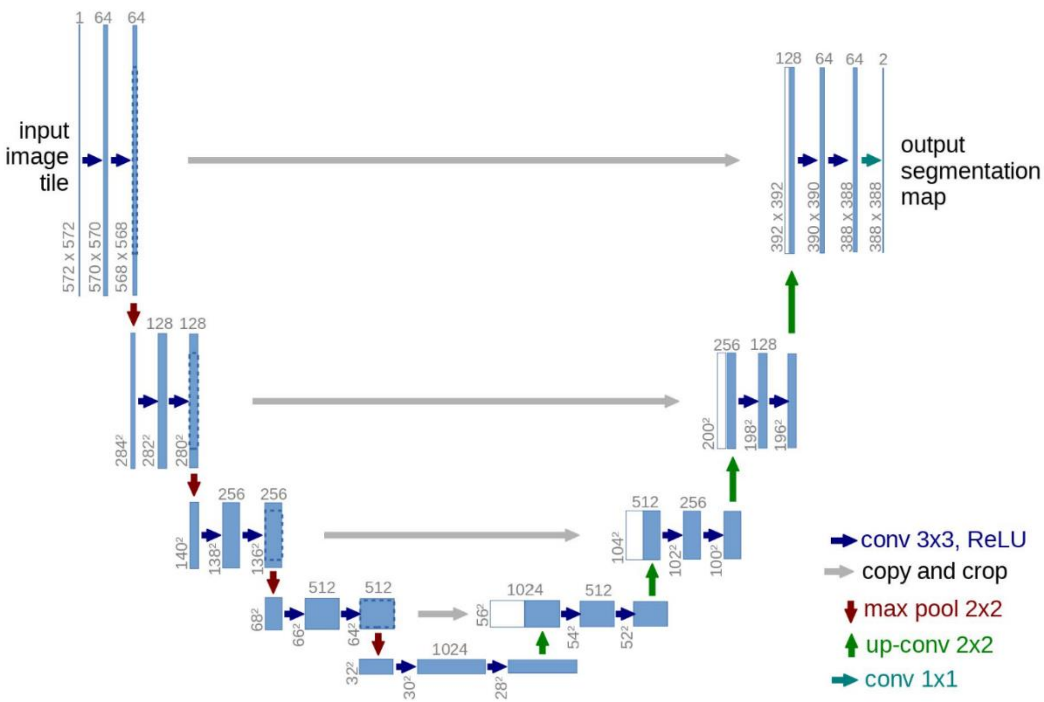
かつ$\sigma_t^2 = 2$の拡散過程の前向きFK-方程式に他ならない．

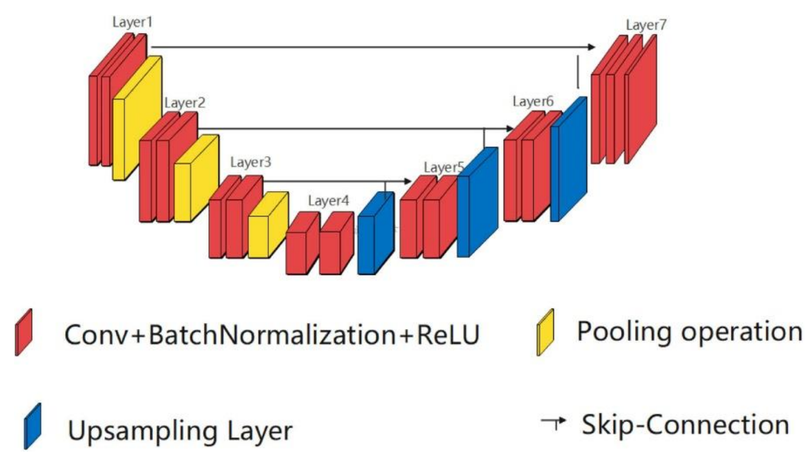
$\tilde{s} \to 0$とすることで，時刻0における分布を得ることができる．
つまり，ドリフト項をデータから推定し，逆過程を走らせることでデータの分布からのサンプリングができるようになる．

画像の場合

- U-Net
  [Olaf, Fischer, Brox: U-Net: Convolutional Networks for Biomedical Image Segmentation. MICCAI 2015]

  - ➢ 画像のsegmentationなどで標準的なネットワーク
  - ➢ 画像生成用の拡散モデルではスコア関数$(t, x) \mapsto \hat{s}(x, t)$のモデルとして最も多く利用されている.



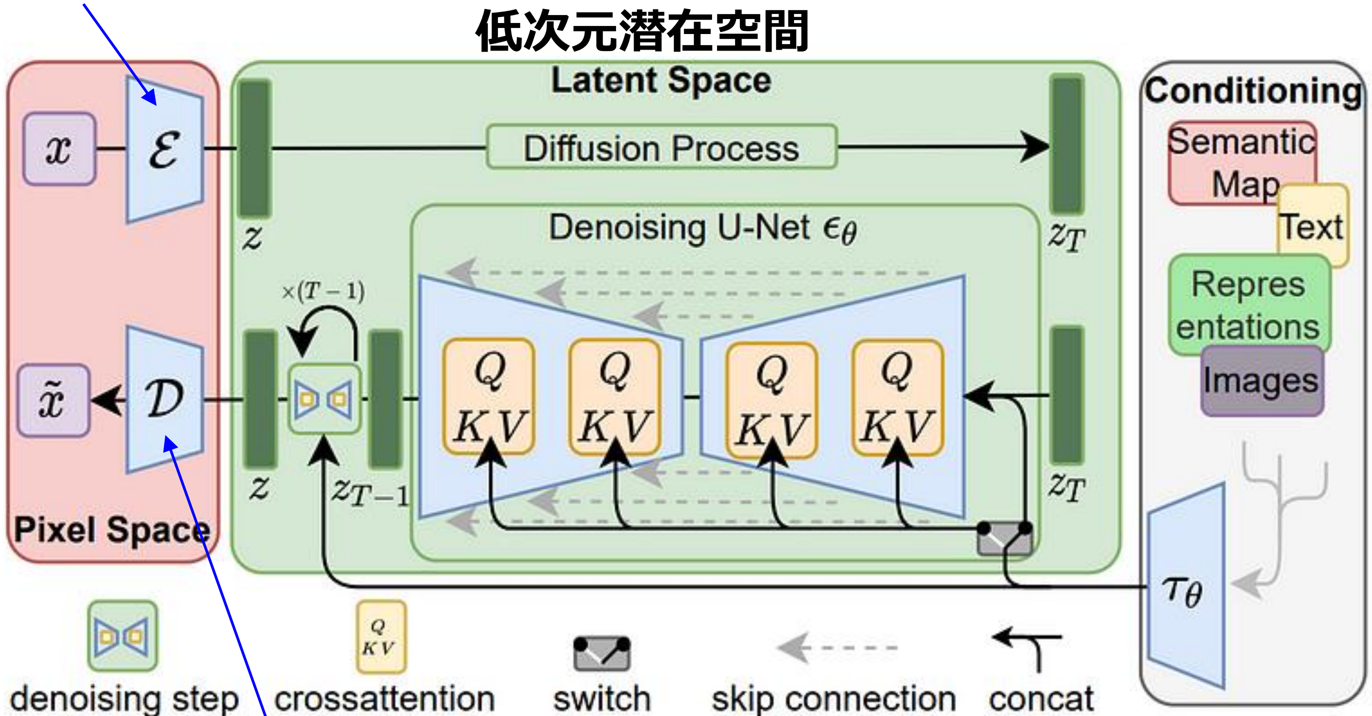


Latent diffusion model

[Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." CVPR2022.]

- 低次元潜在変数の空間で拡散モデルを走らせる.
  - ➢**計算量を削減できる**.
  - ➢汎化誤差の意味でも意義があると考えられる.
  - ➢Stable diffusionで用いられている.

潜在空間にエンコード



潜在空間からもとの空間(画像)にデコード

ODEに変換

- Probability flow ODE (PF-ODE)

$$\mathrm{d}Y_t = (Y_t + 2\nabla \log(p_{\bar{T}-t}(Y_t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

逆向きSDEのFP-方程式

$$\begin{aligned}\partial_t p_t(y) &= -\nabla_y \cdot ((y + 2\nabla \log(p_{\bar{T}-t}(y)))p_{\bar{T}-t}(y)) + \Delta_y^2 p_{\bar{T}-t}(y) \\ &= \nabla_y \cdot [(-y - 2\nabla \log(p_{\bar{T}-t}(y)) + \nabla \log(p_{\bar{T}-t}(y)))p_{\bar{T}-t}(y)] \\ &= \nabla_y \cdot [(\underbrace{-y - \nabla \log(p_{\bar{T}-t}(y))}_{= -v_t(y)})p_{\bar{T}-t}(y)]\end{aligned}$$

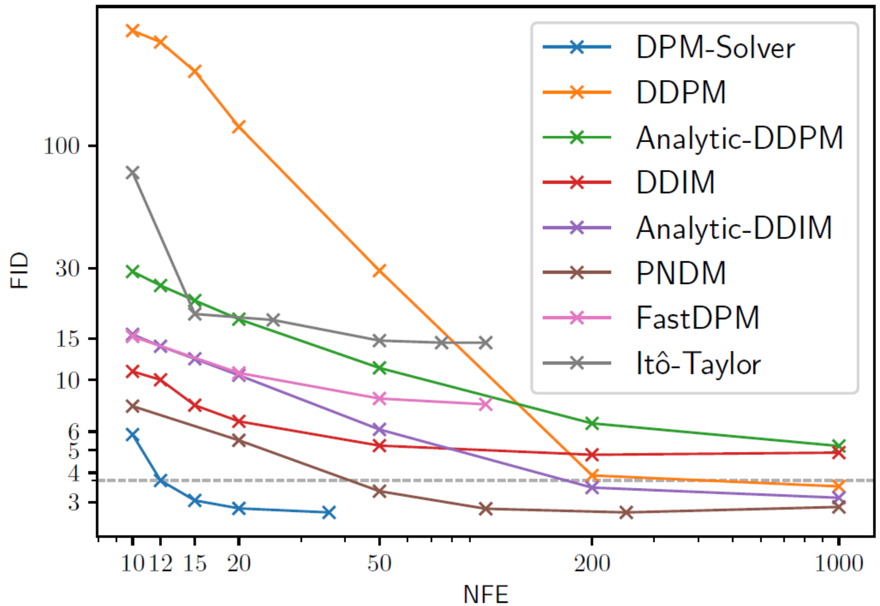
この偏微分方程式は以下のODEに対応する連続の方程式である：

$$\frac{\mathrm{d}\tilde{Y}_t}{\mathrm{d}t} = v_t(\tilde{Y}_t)$$

逆向きSDEを走らせる代わりに，$\tilde{Y}_0 \sim N(0, I)$ として
てこのODEを走らせても良い．

PF-ODEを使った手法

様々な解法が提案されています．
- ナイーブに実装すると時間離散化誤差が強く影響 [2].
- 拡散モデル用に実装を工夫する必要がある [3,4,5].
  - ➢ 線形多段法 [4]，Heun法 [2]，変形exp-Runge-Kutta法 [3]，高次漸近展開 [5]
- スコアの推定誤差には鋭敏かもしれない．



← 計算を工夫したODE型の方法はステップ数を減らしても誤差が発散しにくい．

1. Song, Meng, Ermon: Denoising Diffusion Implicit Models. ICLR2021.
2. Karas et al.: Elucidating the Design Space of Diffusion-Based Generative Models. NeurIPS2022
3. Lu et al.: DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. NeurIPS2022.
4. Liu et al.: Pseudo Numerical Methods for Diffusion Models on Manifolds. ICLR2022.
5. Dockhorn, Vahdat, Kreis: GENIE: Higher-Order Denoising Diffusion Solvers. NeurIPS2022.

- 理論：ODEベースの手法の方が「速い」（離散化誤差が小さい）

➢ Chen et al.: The probability flow ODE is provably fast. 2023.
➢ Li et al.: Towards Faster Non-Asymptotic Convergence for Diffusion-Based Generative Models. 2023.

SDE手法：$O(1/T)$

ODE手法：$O(1/T^2)$

（Tは離散化後のステップ数）

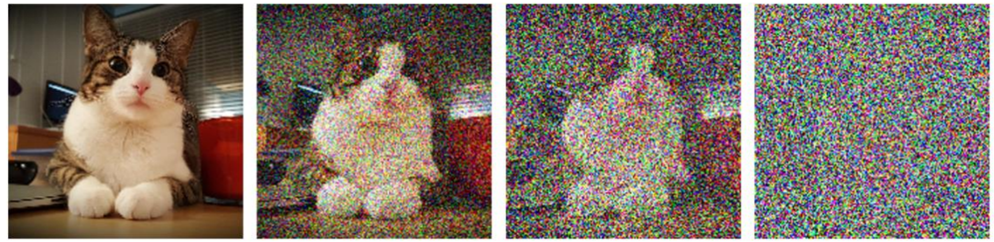
拡散モデルの統計理論

[Kazusato Oko, Shunta Akiyama, Taiji Suzuki: Diffusion Models are Minimax Optimal Distribution Estimators. ICML2023]

Stable diffusion, 2022.

$$\mathrm{d}X_t = -X_t\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

Forward process



Backward process

$$\mathrm{d}Y_t = (Y_t + 2\nabla\log(p_{\overline{T}-t}(Y_t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

$(Y_t \sim X_{\overline{T}-t})$

経験スコアマッチング推定量:

$$\hat{s} = \mathop{\arg\min}_{s\in\mathrm{DNN}} \frac{1}{n}\sum_{i=1}^{n}\int_{t=\underline{T}}^{\overline{T}} \mathbb{E}_{X_t|X_0=x_{0,i}}\left[\|s(X_t,t) - \nabla\log p_t(X_t|x_{0,i})\|^2\right]\mathrm{d}t$$

定理

Let $\hat{Y}$ be the r.v. generated by the backward process w.r.t. $\hat{s}$, then

$$\mathbb{E}_{D_n}\left[\mathrm{TV}(\hat{Y}, X_0)\right] \lesssim n^{-\frac{s}{2s+d}}\log^9(n),$$ (s: 密度関数の滑らかさ)

$$\mathbb{E}_{D_n}\left[\mathrm{W}_1(\hat{Y}, X_0)\right] \lesssim n^{-\frac{s+1-\delta}{2s+d'}}$$ (for any $\delta > 0$).

どちらも（ほぼ）ミニマックス最適 [Yang & Barron, 1999; Niles-Weed & Berthet, 2022].

(Estimator for W_1 distance requires some modification)