修士論文

ニューラルネットワークの連続化と確率的残差学習に対する伊藤拡散過程を用いた損失関数の滑らかさの解析

2018 年 2 月

目次

記号・用語

- $\mathbb{R}$: 実数全体の集合
- $\mathbb{C}$: 複素数全体の集合
- $\mathcal{X}$: 特徴量空間
- $\mathcal{Y}$: ラベル空間.$\mathbb{R}^m$ や $\mathbb{C}^m$
- x_i : $x \in \mathbb{R}^m$ の第 i 成分
- $D_\cdot$:Malliavin 微分
- $\mathbb{D}^{k,p}$:Malliavin 微分の意味での Sobolev 空間
- $\delta(\cdot)$:Skolohod 積分
- $\mathbb{D}^\infty$:Wiener 汎関数空間
- L : 損失関数
- k : 正定値カーネル関数 $\mathcal{X}^2 \to \mathcal{Y}$
- H_k : 再生核 k を持つ再生核ヒルベルト空間
- $p_t^X(x)$:確率過程 $\{X_t\}_{t=0}^T$ の時刻 t における密度関数 (特に誤解の恐れのない状況下では X は略記する)
- γ_T^X : 確率変数 X_T の Malliavin 共分散行列
- η : 活性化関数 $\mathbb{R}^d \to \mathbb{R}^d$
- σ : 活性化関数の射影 $\eta_i(x) = \sigma(x_i)$
- F:終端値関数。CNN の場合は全結合層と softmax 関数と損失関数の合成
- C^1 級関数 $f : \mathbb{R}^d \to \mathbb{R}^n$ の $x \in \mathbb{R}^d$ による微分を次のように定義する

$$f : \mathbb{R}^d \to \mathbb{R}^n, x \in \mathbb{R}^d \tag{1}$$

$$\frac{\partial f(x)}{\partial x} := \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \frac{\partial f_1(x)}{\partial x_2} & \cdots & \frac{\partial f_1(x)}{\partial x_d} \\ \frac{\partial f_2(x)}{\partial x_1} & \frac{\partial f_2(x)}{\partial x_2} & \cdots & \frac{\partial f_2(x)}{\partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n(x)}{\partial x_1} & \frac{\partial f_n(x)}{\partial x_2} & \cdots & \frac{\partial f_n(x)}{\partial x_d} \end{pmatrix} \tag{2}$$

- 関数 $f : \mathbb{R}^d \to \mathbb{R}^m$ の勾配 $\nabla_x f \in \mathbb{R}^d \times \mathbb{R}^m$ を次のように定義する

$$\nabla_x f := (\frac{\partial f}{\partial x})^T \tag{3}$$

- $Exp(\lambda)$: 指数 λ の指数分布
- $N(\mu, \Sigma)$: 平均ベクトル μ, 共分散行列 Σ に従う正規分布

1 序論

近年世間を騒がせている深層学習・通称 DeepLearning は, 高い性能を誇るとされているも、その数学的な解析はまだまだ発展途上の段階である。

深層ニューラルネットの数学的研究は様々な種類があるが、本修士論文では”連続化”について取り扱う。

ニューラルネットの数学的解析における”連続化”にはいくつかの解釈がある。今回取り扱うのは”wide continue”と”depth continue”、そして”learning continue”である。

2 章では、機械学習の問題設定について解説する。数学的かつ一般的に機械学習の問題を再定義し、既存の機械学習手法をこの定義に当てはめた場合の具体例を述べていく。

3 章では”Wide continue”について解説する。”ニューラルネットの積分表現理論”と呼ばれる理論で、ニューラルネットの中間層のノード数を無限大にした場合の関数解析的な解析を行う。[5] における再生核ヒルベルト空間上のリッジレット解析を、[8] の再生核ヒルベルト空間の理論を用いて再構成した。

4 章では”Depth continue”について議論する。Deep Neural Network において, 中間層の数を無限大にした場合について考える。ここでは主に、近年の Deep Learning における主流な手法となっている”ResNet”を中心に議論を行う。”中間層の数を増やす”と簡単に言うが、実際みだりに中間層の数を増やすと、性能が下落する場合が多い。しかし [2] で提唱された ResNet では、skipconnect という手法を導入することで、単純な性能向上に加えて”中間層を増やせば増やすほど性能が上がる”という状況を作り上げた。

[3] の論文は、ResNet の計算を”常微分方程式の離散化”と捉え、既存の ResNet 亜種を常微分方程式の離散化手法で分類し、そのうえで別の常微分方程式の離散化手法を用いて ResNet の改良”LM-ResNet”を提案している

[4] の論文では、さらにその考えを突き詰め、常微分方程式 $dx(t)/dt = f(t, x(t))$ の右辺 F を直接学習させる ODEnet という手法を考え、使用メモリを劇的に減少させたにも関わらず精度は微減で済むことを示している。

また [3] では ShakeShakenmodel や StochasticDepth といった確率的な ResNet は、確率微分方程式の離散化(簡易スキーム)であると提唱されている。我々はこの元々の確率微分方程式 $dX_t = f(t, X_t)dt+g(t, X_t)dB_t$ を考え、これを SDEnet と呼び、種々の解析および数値実験を行った。

以下にその解析による主結果、この修士論文の主定理その 1 を記す。

主定理 1. *SDEnet* のステップ数を N, 終端値時刻 T と置くと、適当な条件の下で初期値 x とする *SDEnet* 真の勾配との誤差 $|E[\frac{\partial}{\partial\theta}F(X_T^\theta) - \frac{\partial}{\partial\theta}F(X_N^\theta)]|$ は、ある定数 $p > 1, q > 0$ と、それぞれ f, g, F と T, x から定まる正の定数 $M(f, g, F), K(T, x)$ が存在し

$$|\frac{\partial}{\partial\theta}E[F(X_T^\theta) - F(X_T^{N,\theta})]| \leq K(T, x)M(F, g, f)||1/\det(\gamma_T)||_{L^p}^q \frac{T}{N} \tag{4}$$

が成り立つ。ただし γ_T は X_T^θ のマリアヴァン共分散行列である

この定理は [6] の定理 3.1.2 の拡張である。

[6] の定理は、活性化関数が ReLu や Swish のような非有界な場合そのまま使用することができない。しかしこれらの活性化関数は近年のニューラルネットの主流である、そのうえリプシッツ連続でありながら入力が原点から離れても勾配が消失しない点が評価されており、これは非有界性と切り離せない。また [6] では F に

微分の有界性を課しているが、ニューラルネットにおいてはここの増大度は 2 次のオーダーになるので、そのままでは使用できない。この主定理 1 では、F は多項式増大まで許容できることを示した。

今回は活性化関数に swish を想定し、[6] の定理 3.1.2 を拡張したものを主定理 1 とする。ReLu とは違い、「連続微分可能」「任意の階数の導関数も含めてほとんどいたるところ非零」という点が非常に役立つ。

5 章では"Learning continue"について議論する。一般に、パラメータ最適化では $\theta_{n+1} = \theta_n - \alpha_n \nabla_\theta L(\theta_n)$ というステップを繰り返してパラメータを最適な値に近づけていくが、これを連続的に考えて、微分方程式の離散化とみなして解析を行う。特に GLD,SGLD に対するエルゴード性を考察する。

出た結果はこの修士論文の主定理 2 とする。

主定理 2. 次のような *GLD* の連続化を考える。

$$d\theta_t = -\nabla_\theta L(\theta_T)dt + \sqrt{2\beta^{-1}} I_d dB_t \tag{5}$$

ここで連続関数 L が、有界開集合内で連続関数、その外で最小の固有値 $\lambda > 0$ を持つ正定値行列 V と $lim_{|\theta|\to\infty} |\nabla I(\theta)|/|\theta| < \lambda$ となる I を用いて $\theta^T V\theta + I(\theta)$ と書けるなら、$d\pi(\theta) := Z^{-1}e^{-\beta L(\theta)}d\theta$ という確率測度に対して

$$||\mu_t^{\theta_0} - \pi||_{var} \to 0 (t \to \infty) \tag{6}$$

この条件は例えば、パラメータに対して線形オーダーになる出力 y に対して、損失関数が適当に log 正規化された分類問題で Ridge 正則化項を付け加えればすれば成り立つ。

ここで、新たに条件を付けくわえて [24] を用いることで次の定理が言える。

定理 1.1. L が上述の条件に加えて適当な条件を満たすとする。このとき、連続化 *GLD* の離散化はステップ幅を適当にとることによって

$$\theta_n \to \theta^* (n \to \infty) \tag{7}$$

ただし θ^* は $L(\theta^*) = 0$ となる大域的最適解で、収束は確率収束である。

6 章では [3] で元の SDE とされた確率微分方程式に対して別の離散化手法を考え, 対応する ResNet 的ニューラルネットを構成した。このネットワークには Deep Swamp network という名前を付けた。その数値実験を行う。

そのメインとなるネットワーク構築は、[7] の「深澤スキーム（論文中では Moving sphere scheme）」による SDE 離散化を理論的基礎としており、Deep swamp network と呼ぶ。

7 章では今後の課題について考察を行う。[24] における大域最適への収束条件は非常に厳しく、さらに深層学習のような莫大な数のパラメータに対する最適化では計算量があまりにも増大する。そのため、今後は通常の SGD と大差ない計算量で大域最適に収束するアルゴリズムを作り上げるため、理論を拡張していく必要がある。

確率微分方程式や malliavin 解析、機械学習の基礎的な解説はほぼ行わない。また既存理論をそのまま載せる場合は証明を書かない。証明が書かれているのは、すべて「新たに作り出した定理」「既存定理の拡張」「既存定理であるものの出展に書かれている証明が不十分」のいずれかである。

2　機械学習の問題設定

特徴量空間 $\mathcal{X}(:=\mathbb{R}^d)$ から、ラベル空間 $\mathcal{Y}(:=\mathbb{R}^m\ or\ \mathbb{C}^m)$ への写像のうち、設定された条件を満たすものが成す可分ヒルベルト空間 $\mathcal{H}$ を仮説空間と呼ぶ。

初期値関数 $f_0 \in \mathcal{H}$ を更新していくことで、”良い”関数 f_k を作っていく。このあらかじめ定められた大きな自然数 k_{max} と同じ回数、更新を繰り返すことを「学習」と呼ぶ。

更新の方法は決定論的学習と確率論的学習では、勾配の計算方法が確率的か否かのみの違いである。

2.1　関数の更新方法

完備な確率空間 $(\Omega, \mathcal{F}, P)$ をおく。また、データ $(x, y) \in \mathcal{X} \times \mathcal{Y}$ の組を $\mathcal{D}(:=\{x_i, y_i\}_{i=1}^n)$ とおく。

2.1.1　決定論的な学習

写像 $L_{\mathcal{D}} : \mathcal{H} \times \Omega \to \mathbb{R}$ を考える。ただし任意の $\omega_1, \omega_2 \in \Omega$ に対して $L_{\mathcal{D}}(f, \omega_1) = L_{\mathcal{D}}(f, \omega_2)$ とする。

今後決定論的な学習においては ω を略記する。誤解の恐れがない場合は $L_{\mathcal{D}}$ を単に L と書く

$$f_{k+1} := f_k - \alpha_k \nabla_f L(f_k) \tag{8}$$

ただし、∇_f は $L(f)$ の $\mathcal{H}$ におけるフレシェ微分を表す。

α_n は学習率と呼ばれ、0.1 や 0.001 といった小さな数が代入される。

2.1.2　確率論的な学習

写像 $L : \mathcal{H} \times \Omega \to \mathbb{R}$ を考える。ω が変わることによって値が変わってもよい

$$f_{k+1} := f_k - \alpha_k \nabla_f L(f_k, \omega) \tag{9}$$

フレシェ微分、学習率については決定論的な場合と同様。

2.1.3　損失関数のパラメトライズ

$\mathcal{H}$ が $\mathcal{R}^N$ と同型である場合、$L(f)$ を $f \in \mathcal{H}$ に対応するパラメータ $\theta \in \mathcal{R}^N$ を用いて $L(\theta)$ と表記する。

2.2　問題設定と具体例

これらの $(\mathcal{X}, \mathcal{Y}, \mathcal{H}, L, (\Omega, \mathcal{F}, P), \{\alpha_k\}_{k=0}^{\infty})$ の組を「機械学習問題」と表記する。

2.2.1　具体例 1:線形回帰（フルバッチ）

上記の定義をもとに、決定論的なフルバッチの線形回帰は次のように書ける。

$$\mathcal{X} := \mathbb{R}^d \tag{10}$$
$$\mathcal{Y} := \mathbb{R} \tag{11}$$
$$\mathcal{H} := \{f : \mathcal{X} \to \mathcal{Y}, f(x) = \sum_{j=0}^{J} a_j \phi_j(x), a_j \in \mathbb{R}, \phi_j := x^j\} \tag{12}$$
$$L(f) := \frac{1}{n} \sum_{i=1}^{n} |y_i - f(x_i)|^2 \tag{13}$$

ここで、J, N, α は使用者自らが決定する変数で、「ハイパーパラメータ」と呼ばれる。
$\mathcal{H}$ は $\mathbb{R}^{J+1}$ と同型

2.2.2 具体例 2:線形回帰（確率的勾配法）

$$\mathcal{X} := \mathbb{R}^d \tag{14}$$
$$\mathcal{Y} := \mathbb{R} \tag{15}$$
$$\mathcal{H} := \{f : \mathcal{X} \to \mathcal{Y}, f(x) = \sum_{j=0}^{J} a_j \phi_j(x), a_j \in \mathbb{R}, \phi_j := x^j\} \tag{16}$$
$$L(f, \omega) := \frac{1}{|I(\omega)|} \sum_{i \in I(\omega)} |y_i - f(x_i)|^2 \tag{17}$$

$I(\omega) \subset \mathcal{D}$ はランダムに抽出してきたデータの一部で、確率空間 $(\Omega_1, \mathcal{F}_1, P_1)$ 上で定義されるものとする。

2.2.3 具体例 3:カーネル回帰（フルバッチ）

$\mathcal{X} := \mathbb{R}^d, \mathcal{Y} := \mathbb{R}$ と定める。あらかじめ正定値性を満たすカーネル関数 $k : \mathcal{X}^2 \to \mathcal{Y}$ を定めておく。

$$\mathcal{H} := \mathcal{H}_k \tag{18}$$
$$L(f) := \frac{1}{n} \sum_{i \in I(\omega)} |y_i - f(x_i)|^2 \tag{19}$$

2.2.4 具体例 4:浅いニューラルネット（回帰）

$$\mathcal{X} := \mathbb{R}^d \tag{20}$$
$$\mathcal{Y} := \mathbb{R} \tag{21}$$
$$\mathcal{H} := \{f : \mathcal{X} \to \mathcal{Y}, f(x) = W_2 \eta(W_1 x - b_1) - b_2 \phi_j(x), W_1 \in \mathbb{R}^{L \times d}, W_2 \in \mathbb{R}^{1 \times L}, b_1 \in \mathbb{R}^L, b_2 \in \mathbb{R}\} \tag{22}$$
$$L(f) := \frac{1}{n} \sum_{i=1}^{n} |y_i - f(x_i)|^2 \tag{23}$$

ただし $\eta : \mathbb{R}^L \to \mathbb{R}^L$ は活性化関数と呼ばれ、あらかじめ定めておいた非線形でリプシッツ連続な関数 $\sigma : \mathbb{R} \to \mathbb{R}$ を用いて

$$\eta_i(z) := \sigma(z_i) \tag{24}$$

として定義される。

2.2.5　具体例 5:浅いニューラルネット（分類）

分類したいクラスの集合である有限集合 $C = \{c^{(1)}, c^{(2)}, ..., c^{(m)}\}$ に対して、データ $\tilde{\mathcal{D}} = \{x_i, c_i\}_{i=1}^n$ が存在している状況で、新たな x に対してどのクラスに属するかを予測する。

$\mathcal{X} := \mathbb{R}^d, \mathcal{Y} := \mathbb{R}^m$ とおき、成形されたデータ $\mathcal{D} := \{x_i, y_i\}_{i=1}^n$ を次のように定義する。

$$y_{ij} := \begin{cases} 1(c_i = c^{(j)}) \\ 0(else) \end{cases} \tag{25}$$

このうえで、$\mathcal{H}, L$ は次のように定義される。

$$\begin{aligned} \mathcal{H} :=& \{g : \mathcal{X} \to \mathcal{Y}, f(x) = softmax(g(x)), g(x) = W_2\eta(W_1 x - b_1) - b_2\phi_j(x), \\ & W_1 \in \mathbb{R}^{L\times d}, W_2 \in \mathbb{R}^{1\times L}, b_1 \in \mathbb{R}^L, b_2 \in \mathbb{R}\} \\ L(f) :=& -\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^m y_{ij}\log f(x_i)_j \end{aligned}$$

この損失関数は交差エントロピーと呼ばれ、C 上の確率測度間の乖離度合いを表す距離のようなもの（距離の公理は満たさない）

新たな入力データ x に対して、$c_{\hat{i}}, \hat{i} = argmax_i f(x)$ を予測されるクラスとする。

2.2.6　具体例 6:深いニューラルネット（回帰）

ここでは中間層が N 層の場合を扱う。

基本的には具体例 4 と同じで、$\mathcal{H}$ のみが異なる。

$$\mathcal{H} := \{f : \mathcal{X} \to \mathcal{Y}, f(x) = V_{end}g_K \circ g_{K-1} \circ ..... \circ g_1(x) - b_{end}, g_i(x) = \eta(W_i x - b_i)\} \tag{26}$$

また、自然数列 $L := \{L_1, ..., L_K\}$ を用いて、$L_0 = d$ とおくと

$$W_i \in \mathbb{R}^{L_i \times L_{i-1}} \tag{27}$$

$$b_i \in \mathbb{R}^{L_i} \tag{28}$$

$$W_{end} \in \mathbb{R}^{1\times L_K} \tag{29}$$

$$b_{end} \in \mathbb{R} \tag{30}$$

$K = 1$ のとき、上記の浅いニューラルネットと等しくなる。

$K > 1$ のとき、このようなニューラルネットによる学習を「深層学習」という。

2.2.7　具体例 7:ResNet

近年主流になりつつある、深層学習の亜種である。ここでは、本修士論文における解析で使用する定義を書く。

具体例 7,8 においては、$T \in (0, \infty)$ に対して、有限個の数列 $t_0 = 0 < t_1 < ..... < t_{K-1} < t_K = T$ を用いて

$$V_t = V_{t_i}(t \in [t_i, t_{i+1})) \tag{31}$$

とする。W, b^1, b^2 についても同様。具体例 9,10 においてはその限りではない。

この状況下で

$$\frac{dX_t^i}{dt} = g(t, X_t^i) \tag{32}$$

$$X_0^i = x_i \tag{33}$$

と置き、終端値 X_T をさらに処理し分類する関数 h(CNN においてはプーリング層と全結合層の合成となる) を用いて損失 $L(f), f := h \bigotimes g$ は、フルバッチの回帰問題の場合

$$L(f) := \frac{1}{n} \sum_{i=1}^{n} |y_i - h(X_T^i)|^2 \tag{34}$$

と定義される。

損失関数の値は各データに対する損失の平均と考えることができ $L^{(i)}(f) := \tilde{L}(x_i, f) := |y_i - h(X_T^i)|^2$ という関数を用いて

$$L(f) := \frac{1}{n} \sum_{i=1}^{n} L^{(i)}(f) \tag{35}$$

と書き直せる。ここで 4 章で重要になる終端値関数を定義する。

定義 2.1. 終端値関数

終端値関数 $F : \mathbb{R}^d \to \mathbb{R}$ を次のように定義する。

$$F(X_T^i) := |y_i - h(X_T^i)|^2 \tag{36}$$

これは要するに「時系列 flow の終端から出力までの関数と、(世間一般で言うところの) 損失関数の合成関数」と考えればよい。

実際のところ、明らかに次の等式が成り立つ。

$$L(f) = \frac{1}{n} \sum_{i=1}^{n} F(X_T^i) \tag{37}$$

ミニバッチ法、分類問題、そして具体例 8,9,10 の場合も同様に定義する。

損失関数をこう置き換えると、ResNet や SDEnet といった時系列 flow モデルに対して、非常に数学的解析がしやすくなる。そのため 4 章では L ではなく F を用いて様々な解析を行う。

2.2.8 具体例 8:StochasticDepth

確率空間 $(\Omega_2, \mathcal{F}_2, P_2)$ 上で定義された、ベルヌーイ分布に従う独立な確率変数列 $b_1, b_2, ..., b_N$ を考える。ただし各 b_i の確率分布はあらかじめ定められた写像 p を用いて $p(i) \in (0, 1)$ に従う。

$$x_{i+1} = x_i + b_i f(i, x_i) \tag{38}$$

という形で定義する。

ここで、もしミニバッチなら $(\Omega, \mathcal{F}, P) := (\Omega_1, \mathcal{F}_1, P_1) \bigotimes (\Omega_2, \mathcal{F}_2, P_2)$ とおく。

2.2.9 具体例 9:ODENet

$$\mathcal{H} = \{f : \mathcal{H}_1 \times \mathcal{H}_2 \to \mathbb{Y}, f(x) = h_{ab}(x_T)\} \tag{39}$$
$$\mathcal{H}_1 = \{h : \mathbb{T} \times \mathcal{X} \to \mathcal{X}, h(t,x) = V_t\eta(W_t x - b_t^1) - b_t^2 \tag{40}$$
$$\mathcal{H}_2 = \{h_{ab} : \mathcal{X} \to \mathcal{Y}, h_{ab}(x) = V_{ab}\eta(W_{ab}x - b_{ab}^1) - b_{ab}^2\} \tag{41}$$

ただし、$x_0 := x, x_t := \int_0^t h(s, x_s)ds$ とおく。

2.2.10 具体例 10:SDENet

本修士論文の主題である。上記の ODEnet を確率化する。

$$\mathcal{H} = \{f : \mathcal{H}_1 \times \mathcal{H}_2 \to \mathbb{Y}, f(x) = h_{ab}(X_T)\} \tag{42}$$
$$\mathcal{H}_1 = \{h : \mathbb{T} \times \mathcal{X} \to \mathcal{X}, h_1(t,x) = V_t\eta(W_t x - b_t^1) - b_t^2, h_2 = V_t^2\eta(W_t^2 x - b_t^{12}) - b_t^{22}\} \tag{43}$$
$$\mathcal{H}_2 = \{h_{ab} : \mathcal{X} \to \mathcal{Y}, h_{ab}(x) = V_{ab}\eta(W_{ab}x - b_{ab}^1) - b_{ab}^2\} \tag{44}$$

ただし、$X_0 = x$ であり、X は $dX_t = h_1(t, X_t)dt + h_2(t, X_t)dB_t$ という確率微分方程式に従うものとする。

確率空間は $(\Omega, \mathcal{F}, P) := (\Omega_1, \mathcal{F}_1, P_1) \bigotimes (B, \mathcal{B}(B), \mu)$ と置く。ただし $(B, \mathcal{B}(B), \mu)$ は $B := C([0,T])$ とした場合の Wiener 空間である。h_1, h_2 が x に対して大域的リプシッツ連続で t に対して $1/2$ 次ヘルダー連続と置くことで、$X_T \in \mathbb{D}^\infty$ となる。つまり X_T が Wiener 汎関数と言え、上記の定義での損失関数が定義でき、またマリアヴァン微分や部分積分の議論に持ち込める。

2.3 活性化関数の具体例

上述の $\sigma(x)$ について、一応大域的リプシッツ連続で非線形であればなんでもいいことになっているが、当然よく使われるものは存在する。

2.3.1 ReLu

$$\sigma(x) := max(0, x) \tag{45}$$

と定義される。「if 文一つで書ける」「勾配が消失しない」といった利点がある。

区分的に滑らかな関数を近似するにあたってこの形が都合がいいとする研究もある

本研究では、連続的に微分可能ではないこと、零点がルベーグ測度無限大に存在すること、区分的に微分しても 2 階微分が零関数になることなどから 6 章 1 項を除いて採用しない

2.3.2 swish

$$\sigma(x) := x \cdot sigmoid(x) \tag{46}$$
$$sigmoid(x) := \frac{1}{1 + e^{-x}} \tag{47}$$

近年 Relu にとって代わって使われ始めている活性化関数。原点から離れるほど Relu に近づく。

C^∞ であること、任意の階数の導関数も含めて零点がルベーグ測度 0 であることなどから、準楕円性などを考察する本研究においては非常に都合がよく、この修士論文では原則こちらを用いることにする。

2.4 正則化

過学習を防ぐため、損失関数に付け加える項。$L \leftarrow L + L_r$ と置き、これを新たな損失関数として用いる。

2.4.1 L1 正則化

$\mathcal{H} = \mathbb{R}^N$ と置けるときに採用される。マンハッタン距離を用いる正則化で、Rasso 正則化とも呼ぶ。

パラメータが疎になりやすい（最適化したときに 0 になるパラメータが多い傾向にある）という利点がある。

2.4.2 L2 正則化

$\mathcal{H}$ のノルムの二乗で定義

ユークリッドノルムで書けるとき、これを Ridge 正則化と言う。

ユークリッドノルムとして書けないとき、チコノフ正則化とも呼ばれる。

3　ニューラルネットの積分表現理論

特徴量空間 $\mathcal{X}(:=\mathbb{R}^d)$ から、$\mathbb{C}$ への写像を構成する浅いニューラルネット考える。

$$f(x) = W_2\eta(W_1x - b_1) \tag{48}$$

ただし、W_2 は $1 \times J$ 複素行列、W_1 は $J \times d$ 実行列、b_1 は J 次元実ベクトルであるとする。

また $\eta : \mathbb{R}^J \to \mathbb{C}^J$ であり、ある非線形で大域的リプシッツ連続な写像 $\sigma : \mathbb{R} \to \mathbb{C}$ を用いて,$\eta_i(y) = \sigma(y_i)$ と書けるとする。

この計算をベクトル $a \in \mathbb{R}^d, b \in \mathbb{R}, \sigma, c \in \mathbb{C}$ を用いて書き直す

$$f(x) = \sum_{i=1}^{J} c_i\sigma(a_i \cdot x - b_i) \tag{49}$$

ここで、$J \to \infty$ とした形

$$f(x) = \int_{\mathbb{R}^{d+1}} \gamma(a, b)\sigma(a_i \cdot x - b_i)dadb \tag{50}$$

これを積分的ニューラルネットと呼ぶ。積分的ニューラルネットに対しては、正則化付き損失関数に対する大域的最適解が解析的に求められる場合がある。そのため、十分広いニューラルネットに対して、最適なパラメータの近似値が一回の数値計算で求められることになる。

この章では、特徴量空間上の測度 μ（データの分布）と、それに対して後の許容条件を満たすように自由に設定できる測度パラメータ空間上の測度 λ を扱えるようにするため、[5] で提唱された再生核ヒルベルト空間上のリッジレット解析を [6] による再生核ヒルベルト空間の理論により我流の再構成を行う。

3.1　再生核ヒルベルト空間上の積分表現理論

$\mathcal{X}$ 上の複素数値関数全体の集合を $\mathcal{F}(\mathcal{X})$ とおく。

パラメータ a, b の成す空間 $\mathbb{R}^{d+1}$ から $\mathbb{C}$ への写像のうち、$\mathbb{R}^{d+1}$ 上の測度 $\lambda(dadb)$ による L^2 空間を $\mathcal{G}(:= L^2(\mathbb{R}^{d+1} \to \mathbb{C}, \lambda(dadb)))$ とおく。

写像 $h : \mathcal{X} \to \mathcal{G}$ を固定し、次のような積分作用素 $S : \mathcal{G} \to \mathcal{F}(\mathcal{X})$ を、$F \in \mathcal{G}$ に対して、$f = SF$ となる $f \in \mathcal{F}(\mathcal{X})$ を、次の等式が成り立つ関数とすることによって定義する。

$$f(x) = \langle F, h(x)\rangle_{\mathcal{G}} \tag{51}$$

ここで、S の像空間 $\mathcal{E}(S) := S(\mathcal{G})$ に対して、次のようにノルムを入れる。

$$||f||_{\mathcal{E}(S)} := inf\{||F||_{\mathcal{G}} : SF = f\} \tag{52}$$

定理 3.1. 再生核ヒルベルト空間 *([8])*

$k : \mathcal{X}^2 \to \mathbb{C}$ を次のように定義する。

$$k(x, y) := \langle h(y), h(x)\rangle_{\mathcal{G}} \tag{53}$$

この時、$\mathcal{E}(S)$ は再生核 k を持つ再生核ヒルベルト空間
$\{h(x), x \in \mathcal{X}\}$ が $\mathcal{G}$ 上完全であることと、S が等距離写像であることは同値

今後、この $\mathcal{E}(S)$ を、再生核ヒルベルト空間であることを強調するために H_k と表記する。

定理 3.2. 等距離元の存在 *([8])*
任意の $f \in H_k$ に対し

$$||f||_{H_k} = ||F^*||_{\mathcal{G}} \tag{54}$$

を満たす $F^* \in \mathcal{G}$ が一意に存在する

この F^* を f のリッジレット変換と呼び、$\mathcal{R}f$ と表記する。
今後、定数 K を $\mathcal{X}$ 上の測度 μ を用いて,$K := \int_{\mathcal{X}} k(x,x)\mu(dx)$ と表記する。
$0 < K < \infty$ の時、この (μ, λ, h) の組は「許容条件を満たす」と呼ぶ。

定理 3.3. 積分作用素の連続性
(μ, λ, h) が許容条件を満たすとき、$H_k \subset L^2(\mathcal{X}, \mu)$ であり、$S : \mathcal{G} \to L^2(\mathcal{X}, \mu)$ は連続作用素
proof.
H_k 上の関数列 $\{u_j\}_{j=1}^{\infty}$ を、$\mathcal{G}$ の正規直交基底 $\{v_j\}_{j=1}^{\infty}$ を用いて

$$u_j(x) := \langle v_j, h(x) \rangle_{\mathcal{G}} \tag{55}$$

と定義する。両辺の $L^2(\mathcal{X}, \mu)$ 上のノルムを取る。

$$\begin{aligned}
||u_j||^2_{L^2(\mathcal{X},\mu)} &= \int_{\mathcal{X}} u_j(x)\overline{u_j}(x)d\mu(x) && (56)\\
&= \int_{\mathcal{X}} (\int_{\mathbb{R}^{d+1}} v_j(z)\overline{h(x)(z)}d\lambda(z) \overline{\int_{\mathbb{R}^{d+1}} v_j(z)\overline{h(x)(z)}d\lambda(z)})d\mu(x) && (57)\\
&\leq \int_{\mathcal{X}} (\int_{\mathbb{R}^{d+1}} v_j(z)\overline{v_j(z)}d\lambda(z) \int_{\mathbb{R}^{d+1}} h(x)(z)\overline{h(x)(z)}d\lambda(z))d\mu(x) (shwartz \text{ の不等式}) && (58)\\
&= K||v_j||^2_{\mathcal{G}} && (59)
\end{aligned}$$

$u_j = Sv_j$ であることを踏まえると、任意の $F \in \mathcal{G}$ は $\{v_j\}_{j=1}^{\infty}$ の線形和で書けるため、同じ議論により

$$||SF||_{L^2(\mathcal{X},\mu)} \leq K||F||_{\mathcal{G}} \tag{60}$$

となる。$f \in H_k$ にはすべて $f = SF$ となる F が存在するため、定理の主張が言える。

許容条件に加え、$\{h(x) : x \in \mathcal{X}\}$ が $\mathcal{G}$ 上完全であるとき、「強い意味で許容条件を満たす」と呼ぶとする。

定理 3.4. H_k の正規直交基底
強い許容条件が満たされるとき、$\{u_j\}_{j=1}^{\infty}$ は H_k の正規直交基底
proof.
上記の定理より、$||f||_{H_k} = ||F^*||_{\mathcal{G}}$ となる F^* が存在し、また内積の線形性から、複素数列 $\{c_j\}_{j=1}^{\infty}, \sum |c_j|^2 < \infty$ を用いて

$$F^* = \sum_j c_j v_j \tag{61}$$

$$f = \sum_j c_j u_j \tag{62}$$

と書ける。強い許容条件が満たされる場合、S は等距離写像。そのため $f = u_j$ としたとき $F^* = v_j$ である。
一般の f, F^* に対して、パーセバルの等式より $||F^*||_{\mathcal{G}} = \sum |c_j|^2$ で、これは $||f||_{H_k}$ と一致する。
分極公式により H_k の内積は $||\cdot||_{H_k}$ から陽に書け、$\langle u_j, u_i \rangle = \delta_{ij}$ が言える。

次はさらに強い条件を課す。この定理の条件を緩めていくことこそが、我々の再定義した積分表現理論における今後の課題となる。

定理 3.5. リッジレット変換の積分表示定理 *([8])*
(μ, λ, h) は強い意味で許容条件を満たし、さらに任意の $f \in H_k$ に対して $||f||_{H_k} = ||f||_{L^2(\mathcal{X},\mu)}$ が成り立つとする。
また、単調増大な $\mathcal{X}$ の部分集合列 $\{E_N\}_{N=1}^{\infty}$ を、$\cup_{N=1}^{\infty} E_N = \mathcal{X}$ となるようにとり

$$(\mathcal{R}_N f)(z) := \int_{E_N} f(x)\overline{h(x)(z)}\mu(dx) \tag{63}$$

と置く、任意の N に対して $F_N \in \mathcal{G}$ なら

$$||\mathcal{R}_N f - \mathcal{R} f||_{\mathcal{G}} \to 0 (N \to \infty) \tag{64}$$

この定理によって F^* はまさに既存のリッジレット変換そのもので、既存研究は $\mu(dx) = dx$ と置いた場合であることがわかる。
今回の再定義の利点は、μ をデータの分布とすることで、よりデータに沿った形でリッジレット変換を定義できることにある。
実問題では、有限個のデータ $(x_i, y_i)_{i=1}^n$ から、なるべく”良い”写像 $y = f(x)$ を構成したい。既存のリッジレット解析では、このリッジレット変換を近似するにあたって

$$(\mathcal{R} f)(a, b) := \int_{\mathbb{R}^d} f(x)\sigma(a \cdot x - b)dx \tag{65}$$

$$\approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)\sigma(a \cdot x_i - b) \tag{66}$$

$$\approx \frac{1}{n}\sum_{i=1}^{n} y_i\sigma(a \cdot x_i - b) \tag{67}$$

とするしかない。しかしこの近似は dx がルベーグ測度である場合は不自然な近似となっている。実際のデータは一様分布には程遠く、一様に近い分布になるようデータの一部を抽出するのは、d が大きい状況では非常に難しい。
ここで dx を特徴量データの分布 $\mu(x)$ に差し替えれば、上記の近似計算はモンテカルロ法として非常に自然なものとなる。

4　超深層学習の連続化

”depth continue” の章である。

深層学習の連続化を考える。

4.1　ResNet と微分方程式

深層学習においては、単純に層を深くしすぎると性能が下落することが知られていた。

そこで 2015 年 Microsoft の研究者から発表された ResNet[2] が、この問題を大きく改善し、現在の深層学習における主流となった。

既存の深層学習では、あるブロックに学習させたい関数 $h(x)$ をそのまま学習させていたが、ResNet の中核となる残差学習と呼ばれる手法では残差関数 $f(x) := h(x) - x$ を学習する。

その後 $x + f(x)$ を次のブロックの入力とすることで、実質的に $h(x)$ を学習させたことと同じになる。([2] ではさらにこの後 ReLu を通したものを次のステップの入力としていたが、後に [9] などの論文でこれを直接次のステップの入力としたほうが良いことが検証されている。そのため本修士論文では通してこちらの流儀を用いる)

n ブロック目の入力を x_n と置き、残差関数 f_n とおけば

$$x_{n+1} = x_n + f_n(x_n) \tag{68}$$

これは常微分方程式のオイラー近似に他ならない。

[4] ではこの考えのもとで、

- ResNet:オイラー法
- PolyNet:後退オイラー法
- FractalNet:2 次ルンゲクッタ法
- RevNet:連立 ODE のオイラー法

という分類をしたうえで、線形多段法を用いて新たな残差学習ネットを構成し、性能の向上に成功した。

NIPS2018 の優秀論文に選ばれた [9] は、さらにこの考えを発展させ

$$\dot{x_t} = f(t, x_t) \tag{69}$$

という常微分方程式の $f(t, x_t)$ を直接学習させるという手法を提案した。

また [4] では、ShakeShake モデルや StochasticDepth といった確率的残差学習モデルは、確率微分方程式の簡易スキームであることを主張している。

たとえば、Stochastic Depth は

$$X_{n+1} = X_n + b_n f_n(X_n) \tag{70}$$

という計算が行われる。ここで b_n は、あらかじめ定められたパラメータ $p_n \in (0,1)$ に従うベルヌーイ分布である。

この確率過程は

$$dX_t = p(t)f(t,X_t)dt + \sqrt{p(t)(1-p(t))}f(t,X_t)dB_t \tag{71}$$

という SDE の簡易スキームであるといえる。ただしブラウン運動は一次元であるとする。

より一般の確率微分方程式

$$dX_t = f(t,X_t)dt + g(t,X_t)dB_t \tag{72}$$

我々はこの確率的残差学習の連続化を SDEnet と名付け、理論的解析および SDE の数値解析手法を用いた新型残差学習ネットワークの構築を行った。

4.2 損失関数の微分可能性

この項のみ、活性化関数を ReLu $\sigma(x) := max(0,x)$ と置く。

この項では、まず活性化関数に関係なくファインマンカッツの公式を用いて SDEnet の偏微分方程式での定式化を考える。最後には toy model として一次元の SDEnet を考え、確率化によってパラメータの微分可能性が向上していることを証明する。

次のような $[0,T]$ 上で定義された連立 SDE を考える。

$$X_t^{s,x,\theta} = x + \int_s^t f(r, X_r^{s,x,\theta}, \theta)dr + \int_s^t g(t, X_r^{s,x,\theta}, \theta)dB_r \tag{73}$$

$$Y_t^{s,x} = F(X_T) + \int_t^T Z_s dB_s \tag{74}$$

活性化関数は ReLu なので、リプシッツ条件と増大条件は問題なく成り立ち、解の存在と一意性が言える。

ここで、Z_s は X, Y から定まる、マリアヴァン微分で計算できる確率過程だが、この正体自体はあまり問題ではない。重要なのは次の偏微分方程式である。

[10] の定理を用いて

$$\frac{\partial u}{\partial t}(t,x,\theta) + \mathcal{L}u(t,x,\theta) = 0 F(x) = u(T,x,\theta) \tag{75}$$

と置くと、$u(t,x) = E[F(X_T^{t,x,\theta})]$、すなわち「時刻 t で $X_t = x$ だったという条件付きの $F(X_T)$ の期待値」である。

ただし、楕円型作用素 $\mathcal{L}$ は次のように定義される。この作用素は、次の勾配法連続化議論においても重要である。

$$\mathcal{L}u(t,x,\theta) := \nabla_x u(t,x,\theta)f(t,x,\theta) + \sum_{i,j}^{d} [g^* g]_{i,j}(t,x,\theta)\frac{\partial^2 g}{\partial x_i \partial x_j}(t,x,\theta) \tag{76}$$

$u(0,x,\theta)$ の θ での微分を考えることはまさしく全体の損失のパラメータ勾配を考えることに他ならない

ここで、簡易モデルとして $g = I_d$(単位行列) とすると、この PDE の解は [11][12] より次の定理が成り立つ。

定理 4.1. $f(t,x) := w_2\eta(w_1x - b_1) + b_2$ と置く。ただし η は *Relu* であるとする。

SDE の解とブラウン運動が次の等式を満たす。

$$P(\int_0^T f^2(X_t)dt < \infty) = 1 \tag{77}$$

$$P(\int_0^T f^2(B_t)dt < \infty) = 1 \tag{78}$$

このとき、次の等式が言える。

$$u(t,x,\theta) = E^x[F(W_{T-t})exp[\int_0^{T-t} f(t+r,B_r,\theta)dB_r - \frac{1}{2}\int_0^{T-t} ||f(t+r,B_r,\theta)||^2dr]] \tag{79}$$

proof.

[11] より前半部分が言えればギルサノフ変換ができ、ギルサノフの定理を使えれば *[12]* により解を陽に書ける。

x 以外は定数なので、上記の定理を使う際は $f(x) = eta(x)$ と置いて一般性を失わない。

まずは簡単な $P(\int_0^T f(W_t)dt < \infty) = 1$ を示す。

ブラウン運動は連続な修正が存在するので、$M(\omega) := max_{0\leq t\leq T}W_t(\omega)$ と置くと、$M(\omega)$ は確率 1 で有限。

よって

$$\int_0^T f^2(W_t)dt \leq TM^2(\omega) < \infty \tag{80}$$

か確率 1 で成り立つ。

$P(\int_0^T f^2(X_t)dt < \infty) = 1$ について証明する。

$f^2(X_t) \leq X_t^2$ なので、X_t の連続性から上と同じ議論により明らか

ここからは toy model として一次元の場合、$f(t,X_t,\theta) = a\sigma(bx - c) - d, \theta = [a,b,c,d], ab \neq 0$ を考える。σ は Relu なので $\sigma(x) = max(0,x)$ すると次の定理が成り立つ

定理 4.2. 損失関数の微分可能性

PDE の解 $u(0,x,\theta)$ はパラメータ a,b,c,d に対して C^∞

proof.

a,d は明らか。b,c についてのみ証明する。

熱核の理論より、*SDE* の解 X の推移確率密度 $p(t_1,x,t_2,y)$ を用いて

$$u(t,x) = \int_{\mathbb{R}} F(y)p(t,x,T,y)dy \tag{81}$$

と書ける。

よって、$\int_0^{T-t} f(B_r,\theta)dB_r$ の密度関数に対する θ の滑らかさが言えればよい。

$\partial_x A(x,\theta) = f(x,\theta)$ となるような関数 A を考え、$A(B_t,\theta)$ に対して伊藤の公式を用いると

$$\int_0^{T-t} b(r,B_r,\theta) \odot dB_t = B(T-t,B_{T-t},\theta) - \frac{1}{2}ab\int_0^{T-t} 1_{bB_s-c>0}ds \tag{82}$$

が言える。

厳密には伊藤の公式は使えないが、$ReLu$ に近似する滑らかな関数列の極限を考えることで実現する。

ここから、b, c の微分可能性を言いたいので、$[0, T-t]$ 間での、$B_{T-t} = y$ となるブラウン橋に対して、$B_s > c/b$ の滞在時間の密度が b, c に対して滑らかであればよい。b は 0 でないので、定義域上 c/b は C^∞ なので、$B_s > l$ と置き、滞在時間の密度が l に対して滑らかであればよい。

[13] より、0 出発で時刻 r で y にたどり着くブラウン橋の l 以上の滞在時間が s 以下である確率 $P_l^s(\tau|y)$ は

$$P_l^r(\tau|y) = \begin{cases} 1-(r-\tau)e^{-\frac{c}{\tau}+\frac{y^2}{2}}(e^c(2c+1)erfc(\sqrt{c}) - 2\sqrt{\frac{c}{\pi}}) & y \leqq l \\ \int_0^\tau \frac{(\tau-u)e^{\frac{y^2}{2}-\frac{l^2}{2(r-u)}-\frac{(y-l)^2}{2u}}}{\sqrt{2\pi}(u(r-u))^{\frac{3}{2}}} \times (\frac{l(y-l)^2}{u} - \frac{(y-l)^2 l^2}{r-u} + y - 2u)du & 0 \leqq l \leqq y \\ 1 - P_{-y}^r(r-\tau) & l \leqq 0 \end{cases} \tag{83}$$

であるため、これは l に対して C^∞ である。あとはこれを τ で微分しても l に対する微分可能性は変わらないので、定理が示された。

活性化関数が $Relu$ の時、拡散項がない（確率的でない ResNet）に対応する移流方程式

$$\partial_t u(t,x,\theta) - \nabla_x u(t,x,\theta) f(t,x,\theta) = 0 \tag{84}$$

$$u(T,x) = F(x) \tag{85}$$

の解 u は、明らかに b, c に対して微分不可能である。

上記の定理は、ResNet の確率化による平滑化で、パラメータの微分可能性が工場することを示している。

4.3 Malliavin 解析を用いた勾配誤差の漸近評価

$0 = t_0 < t_1 < ..... < t_N < T$ という時間列と、次のようなオイラー丸山近似を考える。

$$X_{t_{n+1}} = X_{t_n} + f(t, X_{t_n})(t_{n+1} - t_n) + g() \tag{86}$$

定理 4.3. 近似誤差収束定理 *[22]*

f, g がリプシッツ条件と増大条件を満たすとする。このとき

任意の $p > 1$ に対して

$$sup_N E[|X_T^N|^p] < \infty \tag{87}$$

$$E[sup_{t\in[0,T]}|X_t - X_t^N|^p] \to 0 (N \to \infty) \tag{88}$$

確率的残差学習を確率微分方程式の離散化とし、このサンプリングで計算された勾配を真の SDE モデルにおける勾配の推定値とするには、ある $C(T, x, F)$ が存在し

$$|\frac{\partial}{\partial\theta}E[F(X_T)] - \frac{\partial}{\partial\theta}E[F(X_N)]| \leq C(T,x,f)\frac{T}{N} \tag{89}$$

が f を構成するパラメータ θ に対して成り立つ必要がある。これを仮定すれば、$N \to \infty$ で両辺が 0 に収束する。

次の定理が本修士論文の主定理である。

定理 4.4. 主定理 1 ：勾配誤差の漸近評価 $0 = t_0 < t_1 < ..... < t_N < T$ とおく。

確率微分方程式 X_T^θ とその離散近似確率過程 $X_T^{\theta,N}$ を考える。

$$dX_t^\theta = X_0 + \int_0^t f(s, X_s^\theta, \theta)ds + \int_0^t g(s, X_s^\theta, \theta)dB_s \tag{90}$$

$$dX_t^{\theta,N} := X_0 + \int_0^t f(\psi(s), X_{\psi(s)}^{\theta,N})ds + \int_0^t g(\psi(s), X_{\psi(s)}^{\theta,N})ds \tag{91}$$

ただし $\phi(t) := max[t_n : t > t_n]$ とする。

f, g は x に対して C^∞ かつ大域的リプシッツ連続で *1* 次の増大度を持つ。f, g は α に対して微分可能で、t に対して $1/2$ 次ヘルダー連続であるとする。

また、F は多項式増大な可微分関数であるとする。すなわち、ある n と定数 $\tilde{M}$ が存在し $F(x) \leq M(|x|^n+1)$ であるとする。

$d \times d$ 行列列 $\{L_n\}_{n=1}^{d-1}$ を次のように定義する。

$$L_1 := -(\frac{\partial f}{\partial x})^2 + \frac{\partial \frac{\partial f}{\partial x} f}{\partial x} \tag{92}$$

$$L_{n+1} : = \frac{\partial f}{\partial x} L_n - \frac{\partial L_n f}{\partial x} \tag{93}$$

$$\tag{94}$$

$[f(0,x), L_1 f(0,x), ..., L_{d-1} f(0,x)]$ が、$x \in \mathbb{R}^{d+1}$ 上ほとんどいたるところで一次独立性を持つとする。

この仮定の下で定数 $C(T, x, F)$ は存在し、ある定数 $p > 1, q > 0, K(T, x) > 0, M(f, g, F) > 0$ が存在し

$$|\frac{\partial}{\partial \theta} E[F(X_T)] - \frac{\partial}{\partial \theta} E[F(X_N)]| \leq K(T, x) M(f, g, F) ||1/\det(\gamma_T)||_p^q \frac{T}{N} \tag{95}$$

ただし、$\gamma_T := \gamma_{X_T}$ であり、確率変数 X_T のマリアヴァン共分散行列である。

proof.

他のパラメータは固定し、無作為に中質した一つのパラメータ α に対して言えれば十分である。

[6] では、$\mathcal{A}$ は α のとる定義域（一般には $\mathbb{R}$）としたうえで、ある実数 $\eta > 0$ が存在し、$v := \partial_\alpha f or\ g$ として

$$sup_{t,x,\alpha,\alpha` \in [0,T] \times \mathbb{R}^d \times \mathcal{A} \times \mathcal{A}} \frac{|v(t,x,\alpha) - v(t,x,\alpha`)|}{|\alpha - \alpha|^\eta} \tag{96}$$

が成り立つという条件の下でこの定理が成り立つことが証明されている。

今回、考えたい *ResNet* の f, g は、活性化関数を *swish* で考えているため、この式を満たさない。

さらに、$|\frac{\partial F}{\partial x}(x)|$ が x に対して有界であることを課しており、F が x に対して *2* 次のオーダーになる機械学習の実問題にはそぐわない。

そのため、この定理を拡張するために順を追って補題を積み重ねていく。

今後、この定理の証明では α を略記し、$X_t^{\alpha,N}, X_t^\alpha$ は X_t^N, X_t と書く。

補題 4.1. 勾配過程の表記 *[14]*

$f, g, \partial_\alpha f, \partial_\alpha g$ が共に任意の t, α に対して、x に大域的リプシッツ連続であるとし、f, g は 1 次の増大条件を満たすとする。

また、確率過程 $Y_t := \nabla_x X_t, Z_t = Y_t^{-1}, \dot{X}_t = \partial_\alpha X_t$ と道ごとの微分やその逆行列の存在を仮定し定義すると

$$Y_t = I_d + \int_0^t \partial_x f_s Y_s ds + \int_0^t \sum_{j=1}^n (\partial_x g_s)_j Y_s^j dB_s^j \tag{97}$$

$$Z_t = I_d - \int_0^t Z_s(\partial_x f_s - \sum_{j=1}^q (\partial_x g_s)_j^2) ds - \sum_{j=1}^n \int_0^t Z_s (\partial_x g_s)_j dB_t^j \tag{98}$$

$$\dot{X}_t = \int_0^t \partial_\alpha f_s + \partial_x f_s \dot{X}_s ds + \sum_{j=1}^n \int_0^t \partial_\alpha g + \partial_x g \dot{X}_s dB_s^j \tag{99}$$

と書ける。ただし関数 h に対し、$h_s := h(s, X_s, \alpha)$ とする。

このとき

$$\dot{X}_t = Y_t \int_0^t Z_s[(\partial_\alpha f - \sum_{j=1}^n \partial_x g_{j,s} \partial_\alpha g_{j,s}) ds + \sum_{j=1}^n \partial_\alpha g_{j,s} dB_s] \tag{100}$$

次の補題は Malliavin 解析の議論につなげていくにあたって非常に重要となる。

補題 4.2. *Wiener* 汎関数 *[15]*

上述のリプシッツ条件と増大条件に加え、f, g は x に対して C^∞ 級であるとする。また、任意の i, j に対して $f_i(t,0), g_{ij}(t,0)$ は t に対して有界であるとする。このとき $X^i \in \mathbb{D}^\infty$

$\mathbb{D}^\infty$ は wiener 汎関数空間と呼ばれ、malliavin 微分の意味でのソボレフ空間 $\mathbb{D}^{k,p}$ を用いて

$$\mathbb{D}^\infty := \cap_{k,p \geq 1} \mathbb{D}^{k,p} \tag{101}$$

と定義される。

補題 4.3. *Malliavin* 微分 *[15]*

$$\mathcal{D}_s X_t = Y_t Z_s g(s, X_s) 1_{s \leq t} \tag{102}$$

補題 4.4. *Clark* の表現定理 *[15]* $X_T \in \mathbb{D}^\infty$ のとき *(*実際には $X_T^i \in \mathbb{D}^{1,1}$ の場合まで拡張が可能*)*

$$X_T = E[X_T] + \int_0^T E[\mathcal{D}_t X_T | \mathcal{F}_t] dB_t \tag{103}$$

定義 4.1. *Malliavin* 共分散行列 *[15]*

$A = (A_1, A_2, ..., A_d)^T, A \in \mathbb{D}^\infty$ とする。

このとき、*Malliavin* 共分散行列 γ_F を次のように定義する。

$$\gamma_A := \int_0^T D_t A [D_t A]^* dt \tag{104}$$

この行列が確率 1 で正則で

$$\gamma_A \in \cap_{p\geq 1} L^p(\Omega) \tag{105}$$

であるとき、F の *Malliavin* 共分散行列は非退化であるという

f, g が次のヘルマンダー条件を満たす点 x_0 に対し、$X_0 = x_0$ となる SDE の解の malliavin 共分散行列は非退化であり、またその確率密度関数 C^∞ となることが知られている ([15])。

定義 4.2. ヘルマンダー条件 *(ブラウン運動一次元)*

次のようなベクトル場を考える。

$$D = \sum_{i=1}^{d} f_i(0, x_0)\frac{\partial}{\partial x_i} - \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} g_j(0, x_0)\frac{\partial g_i}{\partial x_j}(0, x_0)\frac{\partial}{\partial x_i} \tag{106}$$

$$C = \sum_{i=1}^{d} g_i(0, x_0)\frac{\partial}{\partial x_i} \tag{107}$$

ベクトルの無限組 $C, [C, D], [[C, D], C], [[C, D], D], [[[C, D], D], C], ...$ が、$\mathbb{R}^d$ を張るとき、*SDE* の解 X は初期点 x_0 に対してヘルマンダー条件を満たすと言う

ただし $[X, Y]$ は *Lie* 括弧積であるとする。

イメージとしては、要するに退化した SDE に対しても、ブラウン運動が全方向に”効いている”というものである。

定理 4.5. 密度関数の滑らかさの一様評価

確率変数 X_0 は確率密度関数を持ち、また *SDE* はほとんどいたるところヘルマンダー条件を満たすとする。

このとき $X_t(t > 0)$ の密度関数は C^∞ 級

proof.

X_T の場合のみを示せば十分である。

$X_0 = x_0$ の *SDE* に対して X_T の確率密度関数を $p_T(x|x_0)$ と表記する。上記の定理により、x_0 がヘルマンダー条件を満たす点なら、これは C^∞ 級

$$p_T(x) = \int_{\mathbb{R}^d} p_T(x|x_0)p_0(x_0)dx_0 \tag{108}$$

これの x に対する微分可能性を見ればよい。

さて、今回考えたい SDE を再度表記しよう

$$dX_t = p(t)f(t, X_t)dt + \sqrt{p(t)(1 - p(t))}f(t, X_t)dB_t \tag{109}$$

f が g の定数倍であること、ブラウン運動が 1 次元であることが今回のポイントである。

$p_t = 0, 1$ の時、元の離散版である stochastic depth に当てはめると「かならずその層をスキップする or かならずその層の関数を用いる」となり、確率要素が消える。実際、ヘルマンダー条件も自明に満たされない。

通常、ブラウン運動の次元が SDE と等しく ($n = d$)、被確率積分行列が単位行列のようなものだった場合、ヘルマンダー条件は自明に満たされる。今回のようなブラウン運動の次元が小さい場合は、ヘルマンダー条件は Lie 括弧積の計算をせざるを得ない。

定理 4.6. ヘルマンダー条件

$d \times d$ 行列列 $\{L_n\}_{n=1}^{d-1}$ を次のように定義する。

$$L_1 := -(\frac{\partial f}{\partial x})^2 + \frac{\partial \frac{\partial f}{\partial x} f}{\partial x} \tag{110}$$

$$L_{n+1} : = \frac{\partial f}{\partial x} L_n - \frac{\partial L_n f}{\partial x} \tag{111}$$

$$\tag{112}$$

$[f(0,x), L_1 f(0,x), ..., f_{d-1}(0,x)]$ が、$x \in \mathbb{R}^{d+1}$ 上ほとんどいたるところで一次独立性を持つとする。

このとき、$\mathbb{R}^d$ 上ほとんどいたるところヘルマンダー条件を満たす。

proof.

X^T の場合のみを示せば十分である。

あまりに乱雑なので機械学習特有のベクトルのベクトル微分表記を用いて、ベクトル場と Lie 括弧積を次のように略記する。

$$C = g \tag{113}$$

$$D = f - \frac{1}{2} \frac{\partial g}{\partial x} g \tag{114}$$

$$[X, Y] = \frac{\partial B}{\partial x} A - \frac{\partial A}{\partial x} B \tag{115}$$

ただし、$C = g$ とは $C = \sum_{i=1}^{d} g_i(0, x_0) \frac{\partial}{\partial x_i}$ の略記である。ほかのベクトルについても同様。

まず最初の Lie 括弧積を考える

$$[D, C] = \frac{\partial g}{\partial x}(f - \frac{1}{2} \frac{\partial g}{\partial x} g) - (\frac{\partial f}{\partial x} - \frac{1}{2} \frac{\partial \frac{\partial g}{\partial x} g}{\partial x}) g \tag{116}$$

$$D^0 : = D \tag{117}$$

$$D^{n+1} : = [D^n, C] \tag{118}$$

$$L_1 := -(\frac{\partial g}{\partial x})^2 + \frac{\partial \frac{\partial g}{\partial x} g}{\partial x} \tag{119}$$

$$L_{n+1} : = \frac{\partial g}{\partial x} L_n - \frac{\partial L_n g}{\partial x} \tag{120}$$

$$\tag{121}$$

として定義すると任意の自然数 n に対して

$$D^n : = L_n g \tag{122}$$

つまり $(g, L_1 g, ..., L_{d-1} g)$ が一次独立であればよい。

この定理は SDE がヘルマンダー条件を満たすための十分条件を述べているに過ぎない。(D にひたすら C のみを掛け合わせているため。回数も d 回以上であってもよい。) しかし $f = Mg(M$ は定数) という条件下では、必要十分条件に近いと思われる。

定理 4.7. 部分積分の公式 *[15]*

β を多重指数とする。また $A \in \mathbb{D}^{k_1,\infty}$ のマリアヴァン共分散行列は非退化、$B \in \mathbb{D}^{k_2,\infty}$ とする。

十分滑らかな実数値関数 F, 確率変数 $A \in \mathbb{D}^{\infty}$ に対し、β, A, B から定まるある確率変数 $H_\beta(A,B) \in L^p(\Omega)$ が存在し

$$E[\partial_\beta F(A)B] = E[F(A)H_\beta(A,B)] \tag{123}$$

定理 4.8. エラー確率変数の評価 *[6]*

測度 $\tilde{\mu}$ を次のように定義する。

$$\int_{\mathbb{R}^d} h(x)\tilde{\mu}(x) = E[h(X_T^{0,N}) + h(X_T^{1,N})] + \int_0^1 E[h(X_T^{\lambda,N})]d\lambda \tag{124}$$

ただし $\lambda \in [0,1]$ で $X_t^{\lambda,N} := \lambda X_t + (1-\lambda)X_t^N$ とする。

また、$\epsilon \in (-1,1)$ を置く *(この区間は 0 を含む小さい有界区間であれば何でも良いが、今回は便宜的に $|\epsilon| < 1$ とした)*。これを用いて新たな確率変数を $X_T^{N,\epsilon} := X_T^N + \epsilon\tilde{B}_T, X_T^{N,\epsilon} := X_T + \epsilon\tilde{B}_T$ と定義する。

$F_m \to F(in\ L^2(\tilde{\mu}))$ となるようにコンパクトな台を持つ関数列 F_m をとる。

$$\mathcal{E}_1(\epsilon, m) := E[F_m(X_T^\epsilon)H_T - F_m(X_T^{\epsilon,N})H_T] \tag{125}$$

$$\mathcal{E}_2(\epsilon, m) := E[F_m(X_T^{\epsilon,N})H_T - F_m(X_T^{\epsilon,N})H_T^N] \tag{126}$$

とすると勾配誤差 $F(X_T)H_T - F(X_T^N)H_T^N = lim_{m\to\infty,\epsilon\to 0}(\mathcal{E}_1(\epsilon,m) + \mathcal{E}_2(\epsilon,m))$ である。

$$|\mathcal{E}_1(m)| \le K_1(T,x)(||F_m(X_T^\epsilon)||_{L^2} + ||F_m(X_T^{\epsilon,N})||_{L^2} + \int_0^1 ||F_m(X_T^{\lambda,N,\epsilon})||_{L^2}d\lambda)||1/det(\gamma_T)||_{L^p}^q \frac{T}{N} \tag{127}$$

$$|\mathcal{E}_2(m)| \le K_2(T,x)(||F_m(X_T^\epsilon)||_{L^2} + ||F_m(X_T^{\epsilon,N})||_{L^2})||1/det(\gamma_T)||_{L^p}^q \frac{T}{N} \tag{128}$$

この定理を用いて主定理を証明する。

f, g は x に対して大域的リプシッツ連続かつ t に対して $1/2$ 次ヘルダー連続、さらに F はたかだか多項式増大である。そのため f, g, F から定まる定数 $M(f,g,F) := sup_{\lambda,N,\epsilon}E[|F(X_T^{\epsilon,N,\lambda})|] < \infty$ を定義することができる。

定理 3.7 の不等式右辺に対して、$m \to \infty, \epsilon \to 0$ とすることで、ルベーグの収束定理を用いて

$$
\begin{aligned}
&|F(X_T)H_T - F(X_T^N)H_T^N| \\
&= lim_{m\to\infty}|\mathcal{E}_1(m) + \mathcal{E}_2(m)| \\
&\leq lim_{m\to\infty}(|\mathcal{E}_1(m)| + |\mathcal{E}_2(m)|)(\text{三角不等式}) \\
&= lim_{m\to\infty}K_1(T,x)(||F_m(X_T^\epsilon)||_{L^2} + ||F_m(X_T^{\epsilon,N})||_{L^2} + \int_0^1 ||F_m(X_T^{\lambda,N,\epsilon})||_{L^2}d\lambda) \\
&||1/\det(\gamma_T)||_{L^p}^q \frac{T}{N} + K_1(T,x)(||F_m(X_T^\epsilon)||_{L^2} + ||F_m(X_T^{\epsilon,N})||_{L^2})||1/\det(\gamma_T)||_{L^p}^q \frac{T}{N} \\
&\leq (K_1(T,x)\cdot 3M(f,g,F) + K_2(T,x)\cdot 2M(f,g,F))||1/\det(\gamma_T)||_{L^p}^q \frac{T}{N}(\text{主定理の仮定及びルベーグの収束定理})
\end{aligned}
$$

となる。

あとは $\partial_\alpha E[F(X_T)] = E[\partial_x F(X_T)\dot{X}_T], \partial_\alpha E[F(X_T^N)] = E[\partial_x F(X_T^N)\dot{X}_T^N]$ であることに注意しつつ、定理 3.6 の部分積分公式を用い、$K(T,x) := 3K_1(T,x) + 2K_2(T,x)$ とすることで、主定理を得る。

注意 4.1. 主定理の汎用性

主定理は最もメジャーな確率的 *ResNet* である *Stochastic Depth* の連続化に対して勾配誤差の漸近性を見たが、ヘルマンダー条件以外は一般の伊藤型確率微分方程式に連続化できる確率的 *ResNet* に対しても言えることである。

他の連続化可能な確率的 *ResNet*（*shakeshake model* や *Stochastic Depth* の連続化でブラウン運動を多次元にしたものなど）に対しても、ヘルマンダー条件さえ証明すればあとは同じである。

すなわち、f, g が x に対して大域的リプシッツ連続で t に対して $1/2$ 次ヘルダー連続、かつ F がたかだか多項式増大であれば本論文主定理 1 と同じ不等式が言える。

4.4 輸送理論とポテンシャルの存在条件

今回我々は確率的残差学習について、微分方程式論で解析的な考察を行ったが、それとは別に輸送理論を用いた幾何学的な考察も存在する。

この項では、その輸送理論とその応用について軽く触れ、最後にこの研究で証明できた些末な定理を述べる。

最適輸送問題の歴史は古く、始まりは Monge の定義した最適輸送問題である

定義 4.3. 最適輸送問題（古典）

$\mathbb{R}^d$ 上の二つの確率測度 μ, τ を考える。

ここで、写像 $T : \mathbb{R}^d \to \mathbb{R}^d$ を、$\mu T^{-1} = \tau$ となるものであるとする。この T を μ から τ への輸送する写像といい、コスト関数 $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ を定義したうえで

$$
\int_{\mathbb{R}^d} c(x, T(x))\mu(dx) \tag{129}
$$

が最小になるような T の存在、そして具体的な構成を考えたい。

この問いを古典的最適輸送問題といい、解 T を最適輸送写像、もしくは単に輸送写像と呼ぶ。

古典的な最適輸送問題は輸送解析が何を問題としているかのイメージが掴みやすいが、不良設定である。そこで若干拡張した現代的な最適輸送問題がある。

定義 4.4. 最適輸送問題（現代）

$\mathbb{R}^d$ 上の二つの確率測度 μ, τ を考える。

$\mathbb{R}^d \times \mathbb{R}^d$ 上の確率測度 π が任意のボレル加速集合 A に対して、次の条件を満たすとき、π を μ, τ のカップリングであるという

$$\pi[A \times \mathbb{R}^d] = \mu(A) \tag{130}$$
$$\pi[\mathbb{R}^d \times A] = \tau(A) \tag{131}$$

このようなカップリング測度全体の集合を $\Pi(\mu, \tau)$ と書く。

そして

$$argmin_{\pi \in \Pi(\mu,\tau)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y)\pi(dxdy) \tag{132}$$

を考えることを、現代的な最適輸送問題、もしくはただ単に最適輸送問題という

最適輸送問題は Monge による提起が 1708 年と古いにも関わらず、この解の存在等に一定の成果が出たのは 1987 年と非常に新しい。

定理 4.9. 最適輸送問題の解 *[25]*

$c(x, y) = |x - y|^2$ とする。要するに距離の 2 乗分コストがかかると仮定する。（この仮定は応用上最も汎用性が高いと思われる）

μ, τ がともに 2 時モーメントが存在し、μ がルベーグ測度に絶対連続なら、最適輸送問題には解が存在し

$$\pi(\mu, \tau) = (id_{\mathbb{R}^d}, \nabla\phi)_{\#}\mu \tag{133}$$

ただし ϕ は恒等的に ∞ でない $\mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ となる凸関数である。

この $T = \nabla\phi$ を最適輸送問題の解、もしくは単に最適輸送写像と呼ぶ。

さらに、輸送写像の連続的な変形と、それに伴う分布の連続的な変形を考える。

定義 4.5. 輸送勾配流 *[1]*

$V(t, x) : [0, T] \times \mathbb{R}^d \to \mathbb{R}$ を用いて、$V_t(x) := V(t, x)$ として

$$\dot{x}_t = \nabla V_t(x_t) \tag{134}$$

という常微分方程式を考える。このとき、データ分布 μ_0 に対応する密度を p_0 と置くと、時刻 t での密度 p_t は

$$p_t(x_t)|\nabla_{x_0} x_t| = p_0(x_0) \tag{135}$$

という等式を満たす。また p_t は次の偏微分方程式を満たす

$$\partial_t p_t(x) = -\nabla \cdot [p_t(x)\nabla V_t(x)] \tag{136}$$

$\nabla\cdot$ はダイバージェンスである。

$\{\mu_t\}_{t=0}^T$ は一定の条件を満たす確率測度で構成された、wasserstein 空間という無限次元のリーマン多様体上の測地線と見做せる。これを用いた幾何学的な解析が、DAE 等に対して行われている。

ここからはこの輸送解析に対する、些細な定理を証明できたため、紹介させていただく。

決定論的な ResNet は、

$$\dot{x}_t = f(t, x_t) \tag{137}$$

$$f(t, x) = V(t)\eta(W(t)x + b_1(t)) + b_2(t) \tag{138}$$

という常微分方程式の離散化だと言えるが、これをさらに輸送勾配流と見なす場合は、どのような f ならポテンシャルを持つと言え、$\dot{x}_t = \nabla_x F(t, x_t)$ と書ける（最適輸送勾配流になりうる）かが重要となる。それについては次の定理を我々は証明した。

定理 4.10. ポテンシャルの存在条件

任意の $t \in [0, T]$ に対して $V^T(t) = W(t)$ の時、$\nabla_x F(t, x) = f(t, x)$ となるスカラーポテンシャル関数 $F : [0, T] \times \mathbb{R}^d \to \mathbb{R}$ が存在する。

また、活性化関数 $\sigma : \mathbb{R} \to \mathbb{R}$ を積分して積分定数を 0 にした関数 ψ を考える。当然 $\psi^{'} = \sigma$ である。$\Psi : \mathbb{R}^d \to \mathbb{R}$ を $\Psi(z) := \sum_{i=1}^d \psi(z_i)$ とおけば、ポテンシャルは

$$F(t, x_t) = \Psi(W_1(t)x + b_1(t)) + \langle b_2(t), x\rangle + C \tag{139}$$

ただし C は任意の定数

proof.

ポテンシャルの存在だけ証明すれば十分である。

f の定義域 $\mathbb{R}^d$ は明らかに単連結なので、ポアンカレの補題より、

$$\text{任意の } i, j \in \{1, 2, ..., d\} \text{ に対して}$$
$$\frac{\partial f_i}{\partial x_j} = \frac{\partial f_j}{\partial x_i} \tag{140}$$

がポテンシャルの存在と同値である。

$y := \eta(Wx)$ として

$$f_j(x) = V_j y \tag{141}$$

$$= \sum_{l=1}^{L} V_{j,l} y_l \tag{142}$$

$$y_l = \sigma(W_l x) \tag{143}$$

$$= \sigma(\sum_{k=0}^{d} W_{l,k} x_k) \tag{144}$$

$$\frac{\partial f_j(x)}{\partial x_i} = \sum_{l=1}^{L} V_{j,l} \frac{\partial y_l}{\partial x_i} \tag{145}$$

$$= \sum_{l=1}^{L} V_{j,l} W_{l,i} \sigma^{'}(\sum_{k=0}^{d} W_{l,k} x_k) \tag{146}$$

同様に右辺は

$$\frac{\partial f_i(x)}{\partial x_j} = \sum_{l=1}^{L} V_{i,l} W_{l,j} \sigma^{'}(\sum_{k=0}^{d} W_{l,k} x_k) \tag{147}$$

よって $V = W^T$ であればポテンシャルが存在することがわかる

この定理は、輸送勾配流の定義と併せることで「輸送勾配流構成としての側面が強いタスク（ex.GAN,AE）に対しては、ResNet+ 転置学習で学習させるのが効率がよい」ということを示唆している。

事実、GAN や AE といったタスクにおいては、$V = W^T$ とする手法がそれなりに使われている。

5　最適化アルゴリズムの連続化とエルゴード性

勾配法での最適化

$$\theta_{k+1} = \theta_k - \alpha_k \nabla_\theta L(\theta_k) \tag{148}$$

の連続化を考える。
また、通常の勾配法以外にも
SGD
ミニパッチの抽出を司るマルコフ過程 U_n を用いて

$$\theta_{k+1} = \theta_k - \alpha_k \nabla_\theta L(\theta_k, U_k) \tag{149}$$

GLD
ガウシアンノイズ $W_k \sim N(0, I_N)$ を用いて

$$\theta_{k+1} = \theta_k - \alpha_k \nabla_\theta L(\theta_k) + \sqrt{2\alpha_k} W_k \tag{150}$$

SGLD

$$\theta_{k+1} = \theta_k - \alpha_k \nabla_\theta L(\theta_k, U_k) + \sqrt{2\alpha_k} W_k \tag{151}$$

に対しても、連続化について考察していく。
通常、$\alpha > 0$ は k によって値を変化させ、$\sum_{n=1}^{\infty} \alpha_k = \infty, \sum_{k=1}^{\infty} \alpha_k^2 < \infty$ となるようにとる。この場合に局所解に収束する先行研究は数多あるため、今回はその限りではない。
また、4 章における θ は時系列 flow を構成するパラメータのみを指したが、5 章における θ は時系列 flow の構成に携わらないパラメータも含めることに注意。（CNN における全結合層のパラメータなど）

5.1　勾配法の連続化とリャプノフ安定性

後述の GLD,SGLD の研究においては、パラメータの分布の変化について考察する。
初期パラメータは $\pi_0(\theta)$ という確率密度関数を持つ確率変数からのサンプリングによって初期値決定する。実装においては、これは多くの場合平均 0 で小さい分散の共分散行列が単位行列な正規分布に従ってサンプリングされる。

$$\theta_{k+1} = \theta_k - \alpha \nabla_\theta L(\theta_k) \tag{152}$$

この勾配法に対しても、輸送解析を用いた解析が存在する。[26]
しかし本項目では輸送解析には触れない。代わりにリャプノフ安定性について簡単に解析する。
まず、次のような常微分方程式を考える

$$d\theta_t = -\nabla_\theta L(\theta_t)dt \tag{153}$$

L はあらかじめ固定された関数で、L を求めたいわけではないので偏微分方程式でないことに注意

勾配法をこの微分方程式の離散スキームとしてとらえると、学習率 α はステップ幅である。この「学習率はステップ幅」という考え方は、勾配法及びその亜種の数学的解析においては非常に重要になる。

初期値 θ_0 を変化させたものを $\tilde{\theta}_0$ と書き、初期値が $\tilde{\theta}_0$ の時の時刻 t でのパラメータを $\tilde{\theta}_t$ と表記する。

正定値行列 V とベクトル b を用いて $L(\theta) = \theta^T V\theta + \langle b, \theta\rangle + C$ と書けるなら

$$||\theta_t - \tilde{\theta}_t|| \approx e^{-\lambda t}||\theta_0 - \tilde{\theta}_0|| \tag{154}$$

ここで λ は最大リャプノフ指数と呼び、V の固有値の中で最も小さいものである。

V は正定値行列なので $\lambda > 0$ であり、初期パラメータの違いによる影響が時間が経つと共に消えていくことがわかる。

5.2 SGD の SDE 化は不可能

既存のいくつかの機械学習論文では、SGD を SDE の離散化として、伊藤の公式等を用いて解析を行っている。これに関しては、我々は否定的である。

SGD の計算を次のように表記する。

$$\theta_{k+1} = \theta_k - \alpha\nabla_\theta L(\theta_k, U_k) \tag{155}$$

通常の勾配法と違って現れる U_n は、ミニバッチの選出を表す確率変数で、勾配の確率部分を司る。この手の議論ではマルコフ性で十分だが、今回は各点独立性を仮定する。

$\Psi_1(\theta_k) := E[L(\theta_k, U_k)], \Psi_2(\theta_k) := Var[L(\theta_k, U_k)]$ と置くと、十分フルバッチサイズが大きければ、中心極限定理から近似的に

$$\theta_{k+1} = \theta_k - \Psi_1(\theta_k)\alpha + \sqrt{\Psi_2(\theta_k)}\alpha\epsilon_k \tag{156}$$

と書ける。ここで ϵ_n は各点独立な標準正規分布に従う乱数である。

先述した通り α はステップ幅なので、$\alpha \to 0$ で SDE に収束することを示したいが、ここでオーダーの問題が出てくる。

伊藤の公式で使われる $dt = (dB_t)^2$ や伊藤積分の等長性 $E[(\int_0^T f(t)dB_t)^2] = E[\int_0^T f^2(t)dt]$ からもわかる通り、ガウシアンノイズが加わる項は、時刻でとる部分の 2 乗のオーダーで収束する。

そのため、上記の近似式は、ルベーグ積分項に近似させたい項は $\Delta t = \alpha$ のオーダー、確率積分項に近似させたい項は $(\Delta t)^2 = \alpha^2$ で収束することになり、$\alpha \to 0$ としたとき、確率積分項が先に消滅し退化した SDE(=ODE) となってしまう。

この後、我々はパラメータ数が十分大きいものとして無限次元の SDE に帰着させようとしたり、クラークの表現定理でマリアヴァン微分を用いて強引に確率積分項を表記したりしようとしたが、いずれもオーダーの差という根本的な違いを埋めるには至らず、SGD の SDE 化は不可能という結論に至った。

5.3 GLD の連続化

勾配法連続化における、「ヘッセ行列が正」はパラメータ初期値を含む凸領域で凸関数になっていることと同値である。そのためもたらされる結果は良いが、非常に重い条件である。

そのため、全体の凸性より遥かに緩い条件下で、高い確率で良い値にたどり着くことを示す。

ここからは GLD 及び SGLD の SDE 化とエルゴード性、そして収束について考える。

$$\theta_{k+1} = \theta_k - \alpha \nabla_\theta L(\theta_k) - \sqrt{2\alpha}\epsilon_k \tag{157}$$

今回は、ルベーグ積分項が Δt で、確率積分項が $\sqrt{\Delta t}$ なので、収束オーダーがどちらも Δt となるため、$\alpha \to 0$ で問題なく SDE に収束させることができる。そのため SGD とは異なり SDE の離散化と見做せる。

5.3.1 エルゴード性

定理 5.1. 不変測度の存在と一意性 *[27]*

確率微分方程式

$$d\theta_t = \nabla_\theta L(\theta_t) + \sqrt{2\beta^{-1}} dB_t \tag{158}$$

に対して不変測度 $d\tilde{\pi}(\theta) = exp[-\beta L(\theta)]$ が、定数倍を除いて一意に存在する。

今回は $Z := \int_{\mathbb{R}^N} \tilde{\pi}(\theta) < \infty$ と仮定し、$d\pi(\theta) := Z^{-1} d\tilde{\pi}(\theta)$ と置くことで、上記の定理より π は一意の不変確率測度になる。

定義 5.1. *KL* 情報量

二つの確率測度 π, μ の KL 情報量 $KL(\mu|\pi)$ を次のように定義する。

$\mu << \pi$ のとき

$$KL(\mu|\pi) := \int_{\mathbb{R}^N} log(\frac{d\mu}{d\pi}(\theta)) d\pi(\theta) \tag{159}$$

絶対連続でない場合は $KL(\mu|\pi) = \infty$ とする

これは KL 距離と呼ばれることもあるが、明らかにわかる通り、距離の公理を満たさない。そのためここでは KL 情報量と呼ぶことにする。

補題 5.1. *KL* 情報量と全変動の関係 *[20]*

$KL(\mu|\pi) < \infty$ とする。このとき

$$||\mu - \pi||_{var}^2 \leq 2KL(\mu|\pi) \tag{160}$$

ただし $||\mu - \pi||_{var}^2$ は測度の全変動である

SDE の初期値変数 X_0 の測度を $d\mu_0(\theta) (= p_0(\theta) d\theta$ とおく。伊藤拡散過程の有名な事実として、時刻 t での測度 $\mu_t := p_t(\theta) d\theta$ は

$$\frac{\partial p_t}{\partial t}(\theta) = \mathcal{L}p_t(\theta) \tag{161}$$

となる。ただし、$\mathcal{L}$ は伊藤過程の生成作用素で

$$\mathcal{L} := \sum_{i=1}^{N} \{\nabla_\theta L(\theta)\}_i \frac{\partial}{\partial \theta_i} + \frac{1}{2} \sum_i \sum_j \sqrt{2\beta} \frac{\partial^2}{\partial \theta_i \partial \theta_j} \tag{162}$$

と定義される。

定理 5.2. エルゴード性

次の伊藤型 *SDE*

$$d\theta_t = \nabla_\theta L(\theta_t)dt + \sqrt{2\beta^{-1}} I_d dB_t \tag{163}$$

に対して、ある定数 $c > 0$ が存在し、$\mu << \pi$ となる任意の μ に対して、$g(\theta) := \sqrt{\frac{d\mu}{d\pi}}$ が

$$KL(\mu|\pi) \leq c \int_{\mathbb{R}^N} g(\theta)\mathcal{L}g(\theta)d\pi(\theta) \tag{164}$$

を満たすとする。(これを係数 c を持つ対数ソボレフ不等式を満たすと呼ぶ)

また、初期値の測度に対して KL 情報量は有限の値になるとする $(KL(\mu_0|\pi) < \infty)$

このとき

$$||\pi - \mu_t||_{var} \to 0 (t \to \infty) \tag{165}$$

proof.

[16] より、π が対数ソボレフ不等式を満たすなら

$$KL(\mu_t|\pi) \leq e^{-\frac{2t}{c}} KL(\mu_0|\pi) \tag{166}$$

が成り立つ。右辺は初期測度が不変測度に対して KL 情報量が有界という過程を置いているため $t \to \infty$ で $\to 0$ が言える。あとは定理 4.2 から、全変動の収束が言える。

この定理により、連続化した GLD は、十分なステップ数を経ると、初期分布に関係なく確率測度 π に近似する分布を持つ。この密度関数 p_∞ は

$$p_\infty(\theta) = Z^{-1} e^{-\beta L(\theta)} \tag{167}$$

となり、「L が最小値を取る θ、もしくは最小値に近い出力を行う入力 θ の周辺に存在する確率が高い」といえる。

事実、$L(\theta)$ の値の増大に対して、確率密度は指数的に減衰する。

次は主定理 2 の証明に重要となる定理である

定理 5.3. *SDE* のエルゴード性 *[21]*

次のような d 次元 *SDE* を考える。

$$dX_t = b(X_t)dt + I_d dB_t \tag{168}$$

ここで、ある定数 $M > 0$ が存在し、$b : \mathbb{R}^d \to \mathbb{R}^d$ が $|x| \geq M$ を満たす任意の x に対して、定数 $r > 0$ が存在し

$$\langle b(x), \frac{x}{|x|} \rangle \leq -\frac{r}{|x|} \tag{169}$$

が成り立つとする。$r > \frac{d}{2} + 1$ が成り立つなら、不変測度 π と定数 $Q, k > 0$, 自然数 m が存在し、

$$||\mu^x - \pi||_{var} \leq Q(1 + |x|^m)e^{-k+1} \tag{170}$$

この定理を用いて、次の主定理 2 が言える。

定理 5.4. 主定理 2 ：連続化 *GLD* のエルゴード性

連続な損失関数 $L(\theta)$ が、ある有界開集合 C と、正定値行列 V、C の外で定義される可微分関数 $I(\theta)$ が $\frac{|\nabla I(\theta)|}{|\theta|} \to 0(|\theta| \to \infty)$ となるものを用いて

$$L(\theta) = \begin{cases} \text{連続関数} \quad \theta \in C \\ \theta^T V \theta + I(\theta)\theta \in (\mathbb{R}^N - C) \end{cases} \tag{171}$$

と書けるとする。このとき、θ_t の時刻 t での測度を μ_t^x とすると

$$||\mu_t^x - \pi||_{var} \to 0(t \to \infty) \tag{172}$$

特に、一様な測度 μ_t を置くと、任意の自然数 m に対して $\int_{\mathbb{R}^N} |x| d\mu_0(x) < \infty$ なら

$$||\mu_t - \pi||_{var} \to 0(t \to \infty) \tag{173}$$

proof.

L に課した条件により、明らかにリプシッツ条件と増大条件を満たし、$\pi(\theta) := \frac{1}{Z} e^{-\beta L(\theta)}$ は有限な全体の測度 $Z < \infty$ で、π/Z を新たな π と置くとこれは一意の不変確率測度となる。

新たなパラメータ $d\theta_t = \frac{1}{2}\beta L(\theta)dt + I_d dB_t$ を考える。

定理 *3.4* を使うために *(32)* が成り立つ証明を行っていく。

仮定より、任意の $\epsilon > 0$ に対してある M_0 が存在し、$|\theta| \geq M_0$ となる任意の θ に対して $|\nabla I(\theta)| \leq \epsilon|\theta|$ が成り立つ。

正定値性より

$$\langle \nabla L(\theta), \theta/|\theta| \rangle = -\langle V\theta, \theta/|\theta| \rangle + \langle \nabla I(\theta), \theta/|\theta| \rangle \tag{174}$$
$$\leq -\lambda|\theta| + |\nabla I(\theta)|(shwartz) \tag{175}$$
$$\leq -\lambda|\theta| + \epsilon|\theta| \tag{176}$$
$$\leq -(\lambda - \epsilon)M_0 \tag{177}$$

$\langle V\theta, \theta \rangle = \theta^T V\theta$ で、正定値行列 V の最小の固有値 $\lambda > 0$ を用いて、$\lambda\theta^T\theta \leq \theta^T V\theta$ であることに注意
十分小さな ϵ に対して $-\lambda + \epsilon < 0$ となるので $r = d/2 + 1$ とおくと、定数 $M_1 > 0$ が存在し

$$-(\lambda - \epsilon)M_0 \leq -r/M_1 \tag{178}$$

となる。あとは $M = 2\beta max(M_0, M_1)$ とおけば、定理 *4.3* を適用でき、命題が証明される。

5.3.2 大域最適への収束

GLD は凸性より遥かに弱い仮定の下で無限ステップを経ることにより、大域的最適解に収束することが示されている。

ここではその先行研究について簡単に解説する。

確率微分方程式

$$d\theta_t = \nabla L(\theta_t)dt + \epsilon dB_t \tag{179}$$

には、$Z^\epsilon := \int_{\mathbb{R}^N} e^{-\frac{2L(\theta)}{\epsilon^2}} d\theta < \infty$ でさえあれば、一意の不変確率測度 $\pi^\epsilon(d\theta) := Z^\epsilon e^{-\frac{2L(\theta)}{\epsilon^2}}$ が存在することはすでに述べた。

ここで、$min_\theta L(\theta) = 0$ とし（下に有界でさえあればこれは平行移動で実現できる）、また $L(\theta) = 0$ を満たす θ は $\mathbb{R}^N$ 上有限個であるとする。

このとき、$\epsilon \to 0$ で、大域最適解の点のみに集中する特異測度 π に対して

$$\pi^\epsilon \to \pi \tag{180}$$

が確率収束の意味で成り立つ。

よって、GLD も加えるステップごとに正規ノイズを徐々に小さくしていくことで、無限ステップを経れば大域最適に収束することが期待できる。

2 つの条件を定義する。

定義 5.2. 条件 *1*

- L は C^2 級
- $|\nabla L(\theta)| \to \infty (|\theta| \to \infty)$
- $lim_{|\theta|\to\infty} inf |\nabla L(\theta)|^2 - \Delta L(\theta) > -\infty$

定義 5.3. 条件 *2*

- $lim_{|\theta|\to\infty} inf\langle \frac{\nabla L(\theta)}{|\nabla L(\theta)|}, \frac{\theta}{|\theta|} \rangle > \sqrt{\frac{4N-4}{4N-3}}$
- 初期値条件

  次のような離散ステップを考える。

$$\theta_{k+1} = \theta_k - a_k \nabla L(\theta_k) + b_k W_k \tag{181}$$

  ただし $a_k > 0$ で、各 W_k は互いに独立な標準正規分布に従う乱数であるとする。

  初期パラメータ $\theta_0 \in \mathbb{R}^N$ を固定する。任意の $\epsilon > 0$ に対して、あるコンパクト集合 $K \subset \mathbb{R}^N$ が存在し、任意の n に対して

$$P(\theta_n \in K) > 1 - \epsilon \tag{182}$$

  が成り立つ。

これらの条件下で、次の定理が言える。

定理 5.5. 連続時間での大域最適解への収束 *[23]*

$$d\theta_t = -\nabla L(\theta_t)dt + c(t)dB_t \tag{183}$$

ここで $c(t)$ は、$C > C_0$ となる定数 C を用いて、十分大きな t に対して $c^2(t) \approx \frac{C}{\log t}$ となる関数とする。*($C_0 > 0$ は [23] で定義された局所最適点から定まる定数)*

初期値 θ_0 が C_0 から定義されるコンパクト集合内にあり、条件 *1* が成り立つなら

$$\mu_t^{\theta_0} \to \pi(t \to \infty) \tag{184}$$

が確率収束の意味で成り立つ。

連続時間で収束するから離散時間では自明、とはいかない。$T \to \infty$ となるため、離散化の平均二乗誤差は爆発しかねない上、∇L が非大域的リプシッツな関数である場合は有限時間の場合すら誤差が爆発する可能性がある。[26]

定理 5.6. 離散時間での大域最適解への収束 *[24]*

定数 $A, B > 0$ を $C = B/A$ となるようにとる。

$$\theta_{k+1} = \theta_k - a_k \nabla L(\theta) + b_k W_k \tag{185}$$

という離散ステップを考える。ただし、$\{a_k, b_k\}_{k=0}^{\infty}$ は十分大きな k に対して $a_k \approx \frac{A}{k}, b_n^2 \approx \frac{B}{k \log\log k}$ となる数列であるとする。

ここで、条件 *1,2* が満たされるなら

$$\mu_k^{\theta_0} \to \pi(k \to \infty) \tag{186}$$

が確率収束の意味で成り立つ。

つまりこれは GLD のアルゴリズムで SDE の離散化幅を時刻によって適当に調整すれば、パラメータ θ_n は $n \to \infty$ で大域的最適解 $\theta^*(:= argmin L(\theta))$ に確率収束することを表している。

5.3.3 Ridge 正則化の効用についての考察

Ridge 正則化項が存在する場合、すなわち $\beta > 0$ を用いて

$$L(\theta) := \tilde{L}(\theta) + \beta|\theta|^2 \tag{187}$$

と置く場合、$\tilde{L}, \nabla\tilde{L}$ に適当な有界性や増大度の制限を課すだけで、SDE の解の存在と一意性、そしてエルゴード性、大域最適解への収束すべてが成り立ってしまう。

これは、エルゴード性及び大域最適解への収束が成り立つには「外側の性質のみが重要」だからだと考えられる。

十分大きなコンパクト集合の中では可微分性さえあれば良く、あとはそこから θ が出てしまった場合はコンパクト集合内に「押し戻す」力が $|\theta|$ に対して線形以上のオーダーで働けば、あとは適当にランダム要素が入ることにより、エルゴード性や大域最適解への収束が言えてしまう。

それが、$\tilde{L}$ に対する適当な増大度制限の下では、Ridge 正則化項の勾配がちょうど線形になっていることによって、十分大きな $|\theta|$ に対して ∇L が線形オーダーになることに由来する。

5.4 SGLD のエルゴード性

上記の GLD の理論に対する問題点の一つに「フルバッチの勾配を毎回計算する必要がある」というものが挙げられる。

そのため、ミニバッチに対する勾配計算にノイズを加えた SGLD に対してエルゴード性を考える必要がある。

これの連続化した SDE は regime-swiching 型と呼ばれ、有限集合 S と、S に値を取る確率過程 U_t を用いて

$$d\theta_t = \nabla_\theta L(\theta_t, U_t)dt + \sqrt{2\beta^{-1}}dB_t \tag{188}$$

と書ける。ただし U_t はマルコフ過程で、U_t, B_t は互いに独立であるとする。

フルバッチサイズを B、ミニバッチサイズを b とすると、$|S| =_B C_b$ である。

ここは既存理論のみを紹介し、今後の展望とする。

定理 5.7. *regime-switching 型 SDE のエルゴード性 [17]*

任意の固定された $k \in S$ に対して、不変測度 π_k が存在しエルゴード性が成り立つとする。

さらに、任意の $k \in S$ に対して

$$\delta_k := \int_0^\infty e^{-\tilde{I}(s,k)}\left(\int_s^\infty \frac{2\lambda e^{\tilde{I}(u,k)}}{\alpha_*(u,k)}du\right)ds < \infty \tag{189}$$

が成り立ち、加えて *[17]* の *(Hq)* が成り立つとする。（各種細かい定義は *[17]* を参照）

このときある定数 M が存在し

$$sup_{x\in\mathbb{R}^N}||\mu_t^x - \pi||_{var} \leq Me^{\beta t} \tag{190}$$

6　SDEnet ～様々な SDE 離散化手法に対応するネットワーク構築～

6.1　SDEnet とは

数値実験の章である。

前前章で、ResNet とは

$$dx_t = f(t, x_t)dt \tag{191}$$

という常微分方程式の離散化であることを見た。具体的には $0 = t_0 < t_1 < ..... < t_n = T$ に対して $f(t_i, x)$ という n 個の関数を学習していると見做せる。

[4] では、この考えをさらに発展させ、直接 $f(t, x)$ という関数を学習しようという発想に出た。実数濃度の個数の関数を学習することはできないため、hypernet を用いて学習を行う。具体的には、t を入力とする別のニューラルネットで、ResNet のパラメータもしくはフィルター係数（後述）を出力する関数 $h(t)$ を構成し、学習を行う。

[19] によるとハイパーネットにはさほど複雑性が必要ないため、中間層が少ない 3 層パーセプトロンにより実装する。

また、我々は ODEnet を発展させ

$$dX_t = f(t, X_t)dt + g(t, X_t)dB_t \tag{192}$$

という形の連続型 ResNet を SDEnet と名付け、新たな残差学習を構成した。

数値実験においては、共通の設定として CNN による CIFAR10 の実験を行う。

まず最初に畳み込み層を入れ、チャンネル数を 64 にする

そこから 3×3 のフィルター 64 枚の CNN を用いる。ブロック数は 52 ブロックで、各ブロックは入力に畳み込みを行い、活性化関数 (swish) をかけ、そのあと再度畳み込みを行う。

また、hypernet を用いてフィルター係数を導入する。各フィルターを $f_i(x)$ と置き、$f_i(t, x) = h_i(t)f_i(x)$ とすることで、連続の値をとる t に対して $f(t, x)$ を定義することができる。

損失関数は通常の交差エントロピーロスと、Ridge 正則化を行う。

6.2　離散化された SDEnet の構成

前前章では理論的解析を行った考えた stochastic depth の連続化を、今度は数値的に考える。

離散化にはいくつかの種類が考えられる。

6.2.1　簡易スキーム

既存の Stochastic Depth の場合である。全体の時刻 T と分割数 N を用いて

$$X_{n+1} = X_n + p_n f(Tn/N, X_n)\frac{T}{N} + \sqrt{p_n(1-p_n)}f(Tn/N, X_n)\Delta\tilde{B}_n \tag{193}$$

ただし $\Delta\tilde{B}_n$ とは、互いに独立な確率 1/2 で $\frac{T}{N}, \frac{T}{N}$ の値をとる離散確率変数である。

6.2.2 オイラー丸山 network

$$X_{n+1} = X_n + p_n f(Tn/N, X_n)\frac{T}{N} + \sqrt{p_n(1-p_n)} f(Tn/N, X_n)\Delta B_n \tag{194}$$

ただし ΔB_n は平均 0, 分散 $\frac{T}{N}$ の正規分布に従う乱数である

6.2.3 ミルシュタイン network

CNN のフィルターを素早く行列に変形するアルゴリズムが完成しなかったため、今回この数値実験は行わない

$$X_{n+1} = (\text{オイラー丸山スキーム項}) + p(t)(1-p(t))f(X_n)\frac{\partial f}{\partial x}(X(n))f(X_n)((\Delta W(n))^2 - \Delta t) \tag{195}$$

6.2.4 Deep swamp network

今回最も複雑な離散化である。この手法のためにハイパーネットを導入した

[7] によって証明された弱近似スキームで、誤差が 1/3 ほどになる。

大きな特徴として「分割が等間隔ではない」というものがある。

一般の SDE

$$dX_t = f(t, X_t)dt + g(t, X_t)dB_t \tag{196}$$

に対して、弱近似を考える。

$$X_{n+1} = X_n + f(t_n, X_n)\Delta t_n^N + g(t_n, X_n)\Delta^N B_n \tag{197}$$

ただし $\Delta t_n^N, \Delta^N B_n$ はブラウン運動の次元を q とするとそれぞれ

$$Z \sim Exp(1) \tag{198}$$
$$\epsilon \sim N(0, I_q) \tag{199}$$
$$a := (1 + \frac{2}{q})^{1+q/2} \tag{200}$$
$$\Delta t_n^N = \frac{1}{N} a e^{-Z} \tag{201}$$
$$\Delta^N B_n = \sqrt{\frac{1}{N} a q Z e^{-Z}} \frac{\epsilon}{|\epsilon|} \tag{202}$$

で、学習 1 回ごとに乱数を生成して構築する。

6.3 数値計算結果

Deep swamp network はパラメータを大幅に削減しつつ、Stochastic Depth とほぼ変わらない結果を出しているものの、両手法揃って早い段階で学習が収束してしまっている。

ちなみに、今回は 52 層なのでパラメータ数に倍程度の差しかつかなかったが、Deep Swamp network は SDEnet の非確率版である ODEnet の論文 [4] で指摘されている通り、層数をいくら増やしてもパラメータ数が増えない。そのためより大規模なネットワークを用いる場合は、パラメータ数の差はさらに大きくなる。

	パラメータ数	計算時間 (秒)
Stochastic Depth	8.0M	18178
Deep Swamp network	4.3M	18568

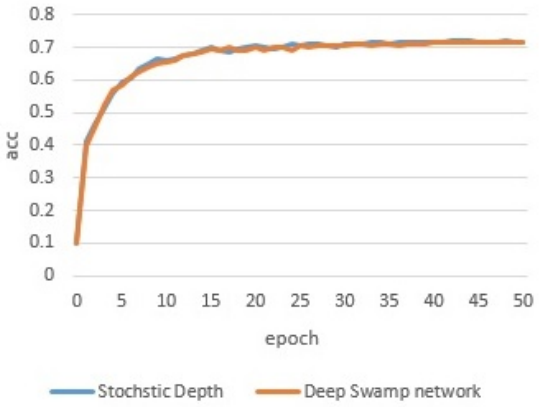

図 1　数値結果

6.4　アーキテクチャの改善案

数値結果があまりうまくいかなかったのは、「最初の層で畳み込みを行う」ということに原因があるのではないかと考えられる。

通常の画像は赤青黄の 3 チャンネルであり、これではあまりにも狭く、より wide なニューラルネットにするためこれを 64 チャンネルにしたいという状況である。ここで我々は一度畳み込みの層を挟むことで対処したが、これは学習対象のパラメータが関わり全体に悪影響をもたらすのではないかと考えられる。

そのため、ここは固定パラメータでのチャンネル数増加を行いたい。

具体的に考えられる手法は次の 2 つである。

- zero padding（残りの 61 チャンネルはすべてゼロを敷き詰める）
- 元画像のコピー

元画像のコピーとは、チャンネル数を 64 に近い 3 の倍数・63 や 66 などに設定し、元の 3 チャンネルを 21 個ないし 22 個複製することでチャンネル数を増やすという方法である。

畳み込みのパラメータ初期値はランダムに定められるため、「1 枚の画像から様々な形で特徴量を抽出する」という行為を最初から並列で行うことができ、結果の大幅な改善が期待される。

7　今後の課題

全体的に言えるのは、仮定の妥当性だと言えよう。

3-4 で証明したポテンシャルの存在条件は、活性化関数の内外でかける行列が転置であるように、その他の理論に対しても仮定を掘り下げ、軽い、もしくはコードで容易に反映できる仮定に落とし込みたい。

ここからはその点を除いた今後の課題について記す。

7.1　積分表現理論

[5] では、我々の再定義とは少々異なるが、カーネル法でのリッジレット解析を展開している。

ここでは μ はルベーグ測度に固定し

$$L(\gamma) := ||f - S\gamma||_{H_k} + \beta||\gamma||_{\mathcal{G}} \tag{203}$$

というチコノフ正則化項付きの損失関数に対して

$$argmin_{\gamma\in\mathcal{G}} L(\gamma) = \frac{1}{1+\beta}\mathcal{R}f \tag{204}$$

が成り立つことを示している。我々が再定義したカーネル上のリッジレット解析でも、同じことが成り立つのかどうか検証する必要がある。

7.2　残差学習の連続化

元々、我々は SDEnet の連続化を通して、凸性の向上が証明されることを期待していた。

伊藤の公式を用いることで、パラメータを動かしたことによる変分 $L(\alpha+\epsilon) - L(\alpha)$ は次のように書ける。

$$\begin{aligned} L(\alpha+\epsilon) - L(\alpha) &= E[F(X_T^{\alpha+\epsilon}) - F(X_T^{\alpha})] && (205)\\ &= \int_0^T E[\nabla_x u(t, X_t^{\alpha+\epsilon}(f(t, X_t^{\alpha+\epsilon}, \alpha+\epsilon) - f(t, X_t^{\alpha+\epsilon}, \alpha)) && (206)\\ &\quad + \frac{1}{2}\sum_{i,j}^{d}([gg^*]_{ij}(t, X_t^{\alpha+\epsilon}) - [gg^*]_{ij}(t, X_t^{\alpha+\epsilon}, \alpha))\frac{\partial^2}{\partial x_i \partial x_j}u(t, X_t^{\alpha+\epsilon})] && (207) \end{aligned}$$

この式に対して quesi-convex などの概念も用いて様々な解析を行ったが、特に顕著な結果は出なかった。やはり期待値で凸性を見るのは筋が悪く、下記のような学習過程に対する解析を行った場合に効力を発揮するのだと考えられる。

7.3　学習の連続化

7.3.1　エルゴード性について

機械学習の一般的状況の中でエルゴード性が言えるとしたのは大きい。

SGLD は GLD よりも大幅に計算量を抑えることができるため、SGLD に対してエルゴード性が言える状況があると分かったのは朗報である。

しかしまだ大きな問題が残されており、それは「ブラウン運動の生成」である。

ブラウン運動の次元はパラメータ数と同じであるため、数百個程度のパラメータを持つシステムに対する最適化なら問題ないが、深層学習のような億単位のパラメータを持つシステムに対するパラメータ最適化においては、ステップごとにパラメータ数と同じ数のガウスノイズを生成しなければならないという理由から、メモリと計算時間の観点により現実的ではない。

パラメータ数よりも遥かに少ない次元のブラウン運動を用いて

$$d\theta_t = \nabla_\theta L(\theta_t, U_t)dt + \Sigma dB_t \tag{208}$$

ただし $\Sigma \in \mathbb{R}^{d\times m}$ で $d >> m$

と書きたい。これであれば、ミニバッチ選出と合わせることで元の GLD アルゴリズムと比較してデータ数、パラメータ数による計算量の爆発を抑えることができ、エルゴード性も相まって非常に実用性の高い最適化アルゴリズムとなる。

しかし今度はこのような退化した SDE に対してもエルゴード性が成り立つのか考える必要がある。退化した SDE に対するエルゴード性の研究は [18] を含め少数しかなく、今後機械学習及び最適化アルゴリズムにおける確率解析学の応用で、重要な純粋数学の研究テーマである。

7.3.2 大域的最適化

エルゴード性の先には非常に重要な定理がある。

それは GLD のエルゴード性の後に解説した「あるアルゴリズムで無限のステップを踏めば、大域的最適解に収束する」である。

[23] では、連続時間での議論があり、[24] ではその離散化でも一定の条件下では大域的最適解に確率収束することが示されている。

[24] の条件 (本修士論文における条件 2) は、非退化な回転による変数変換に対する不変性がないなど、拡張する余地がある。

究極の目標は、「退化したブラウン運動による regime-swiching 型 SDE の離散化は、さほど特異ではない最適化タスクにおいて、特定のアルゴリズムによって大域的最適解に収束する」である。

これがあれば、機械学習のみならず、あらゆる最適化に応用できる非常に汎用性が高く有用なアルゴリズムとなる。

謝辞

方に感謝を申し上げます。

研究に行き詰ったときは音楽から力を頂きました。特に hosnm 氏の描くオーケストラを思わせる壮大なピアノ曲の数々、そしていかさんの歌う”ボクらの最終定理”は、研究が上手くいかない私を再び奮起させてくれました。末筆で恐縮ですが、この場でささやかな御礼の言葉を述べさせていただきます。

参考文献

[1] Sho Sonoda and Noboru Murata. Double continuum limit of deep neural networks. ICML Workshop Principled Approaches to Deep Learning, 2017

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.

[3] Yiping Lu, Aoxiao Zhong, Quanzheng Li, and Bin Dong. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. arXiv preprint, arXiv:1710.10121,2017.

[4] T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, "Neural ordinary differential equations," arXiv preprint arXiv:1806.07366, 2018.

[5] Sho Sonoda, Isao Ishikawa, Masahiro Ikeda, Kei Hagihara, Yoshihiro Sawano, Takuo Matsubara, Noboru Murata,Integral representation of shallow neural network that attains the global minimum.arXiv:1805.07517v2,2018

[6] Gobet, E., Munos, R.: Sensitivity analysis using ItoMalliavin calculus and martingales. Application to stochastic control problem. SIAM J. Control Optim. 43, 16761713,2005

[7] Masaaki Fukasawa,Jan Obloj,Efficient discretisation of stochastic dierential equations.arXiv preprint,arXiv:1506.05680,2015

[8] S. Saitoh. Integral transforms, reproducing kernels and their applications. Addison Wesley Longman, 1997

[9] Han, D., Kim, J., Kim, J.: Deep pyramidal residual networks. In: Proc. of Computer Vision and Pattern Recognition CVPR,2017

[10] Etienne Pardoux1,Shanjian Tang,Forwardbackward stochastic differential equations and quasilinear parabolic PDEs. Probab. Theory Related Fields 114 123150,1999

[11] R. S. Liptser and A. N. Shiryaev, Statistics of Random Processes I: General Theory, 2nd ed. Berlin, Germany: Springer-Verlag, 2001.

[12] Ioannis Karatzas,Steven Shreve,Brownian Motion and Stochastic Calculus. 2nd ed. Springer,Berlin Heidelberg New York.1998

[13] Roman N. Makarov and Karl Wouterloot,Exact simulation of occupation times. In: Monte Carlo and quasi-Monte Carlo Methods 2010, vol. 23, Springer Proc. Math. Stat., Springer, Heidelberg, pp. 573587.2012

[14] Philip E Protter,Stochastic Integration and Differential Equations. Springer, New York.1990

[15] Nualart, D.: Malliavin Calculus and Related Topics. (Probability and its Applications) Berlin Heidelberg New York: Springer,2000

[16] Dominique Bakry,Ivan Gentil,Analysis and Geometry of Markov Diffusion Operators. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences] Springer, Cham.2014

[17] Jinghai Shaoa,Fubao Xi,Strong ergodicity of the regime-switching diffusion processes, Stoch.Proc. Appl., 123 pp. 39033918.2019

[18] J.C. Mattinglya, A.M. Stuartb, D.J.Higham,Ergodicity for SDEs and Approximations:Locally Lipschitz Vector Fields and Degenerate Noise, Tech. Rep. 7, University of Strathclyde, Department of

Mathematics, Glasgow, UK, 2001.

[19] David Ha, Andrew Dai, Quoc V. Le,HyperNetworks. In International Conference on Learning Representations, 2017

[20] 舟木直久, 確率微分方程式. 岩波オンデマンドブックス (2004)

[21] A. Yu. Veretennikov , On polynomial mixingbounds for stochastic di;erential equations. Stochastic Process. Appl. 70, 115127.1997

[22] Peter E. Kloeden,Eckhard Platen,Numerical Solution of Stochastic Differential Equations,Springer,2011

[23] Tzuu-Shuh Chiangf,Chii-Ruey Hwangf,Shuenn-Jyi Sheu,Diffusion for global optimization in Rn,SIAM Journal on Control and Optimization 24 1031-1043. (1986)

[24] Saul B. Gelfand,Sanjoy K. Mitter, Recursive Stochastic Algorithms for Global Optimization in Rd,SIAM J. Control Optimization 29, 9991018.1991

[25] Y. Brenier, Polar factorization and monotone rearrangement of vector-valued functions, Comm. Pure Appl. Math. 44 375417,1991

[26] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for overparameterized models using optimal transport. arXiv preprint arXiv:1805.09545, 2018.

[27] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. “Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis”. In: arXiv preprint,arXiv:1702.03849 (2017).

[28] M. Hutzenthaler, A. Jentzen, P.E. Kloeden, Strong and weak divergence in finite time of Euler’s method for stochastic differential equations with nonglobally Lipschitz continuous coefficients, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science 467,2011.