Github 版

一般的な機械学習入門

最終更新：2025/02/10

@rinnarua

目次

notation

- $\mathcal{X}$：特徴量空間。$\mathbb{R}^d$ の単連結な開部分集合
- $\mathcal{Y}$：ラベル空間。$\mathbb{R}^n$ もしくは $\mathbb{C}^n$ か、それらの上での確率測度の集合
- $(\Omega, \mathcal{F}, P)$：確率空間
- H：使う手法により異なる条件を満たす関数 $f : \mathcal{X} \to \mathcal{Y}$ の集合が為す可分ヒルベルト空間
- L：頻度論的手法における損失関数 $H \times \Omega \to \mathbb{R}$
- $\nabla_f L(f, \omega), f \in H$：頻度論的手法における勾配の定義。$\omega$ を固定した上でのフレシェ微分に f を代入したもの。
- $\mathcal{B}(A), A$ は距離空間：ボレル集合族
- $\mathcal{P}((\Omega, \mathcal{F})$：可測空間 $(\Omega, \mathcal{F})$ 上の確率測度全体の集合。$\mathcal{F}$ を略記し、Ω が距離空間である場合、$\mathcal{P}(\Omega)$ は $\mathcal{P}((\Omega, \mathcal{B}(\Omega))$ であるものとする。
- $\mathcal{R}$：リッジレット作用素
- $\mathcal{R}^*$: 双対リッジレット作用素
- D_{ucp}：ucp 位相の入った cádlág な適合過程の集合。（すなわち発展的可測）
- L_{ucp}：ucp 位相の入った cáglád な適合過程の集合。（発展的可測）
- $N(dtdz)$：ポアソンランダム測度

1　はじめに

1.1　前提知識

本資料において、入門編の前提知識は「学部レベルの解析学」(主に関数解析とルベーグ積分, 確率論)、発展編はその都度必要な知識を補完されたし。

確率解析は基本的に [1] の流儀を使用する。[*1]

1.2　無料版と有料版

後に本書は booth 等で販売する予定です。有料版は論文公開直後の最先端研究の解説や、確率論・統計学の基礎からマリアヴァン解析まで数学の簡易的な解説も付属します。投げ銭感覚でもいいので買ってくださると非常にありがたいです。価格は 1 万円未満にはします。

また、有料版発売後、github で公開しているこの無料版は、よほどのミスや重大な数学的誤り以外修正しないので、ご了承ください。有料版は発見次第なるべく早く修正します。[*2]

[*1] 数か月前まで無駄に難解なだけの読みにくい本だと思っていましたが、変な確率過程の研究をしだすとこんなにありがたい既存理論は他にないと分かります。この本がなかったら、私がここ 4 か月で出したの成果と同じ研究成果を出すのに 3 年くらいかかってそう。「ucp 位相の入った適合 cádlág 過程空間」の確率解析理論がそこそこ完成されてる事実があまりに便利すぎる。

[*2] booth は一度購入すれば制限なく新バージョン落とせます。

2　入門：機械学習の一般的問題設定

この章では、機械学習の問題設定について、関数解析的に定義する。

定義 2.1　transition kernel

二つの可測空間 $(X_1, \mathcal{F}_1), (X_2, \mathcal{F}_2)$ に対して、実数値写像 $P : X_1 \times \mathcal{F}_2 \to \mathbb{R}$ は次の条件を満たすとする。

- $A \in \mathcal{F}_2$ を固定すると、$P(A|\cdot)$ は $X_1 \to \mathbb{R}$ の写像として可測
- $x \in X_1$ を固定すると、$P(\cdot|x)$ は $(X_2, \mathcal{F}_2)$ 上の確率測度

このような P のことを transition kernel という。特に X_2 がユークリッド空間で、ルベーグ測度に対して絶対連続のとき、その確率密度関数を $p(\cdot|x)$ と書くとする。

機械学習であまりにも頻繁に出てくる概念である。しっかり使いこなせるようになっていただきたい。後の章では、もう一つ可測空間を用いて、その空間上の点を固定すると transition kernel となる、super transition kernel なる概念も登場する。ベイズ論では非常に重要だ。

定義 2.2　データの定義

多数の特徴量とラベルの組、$\mathcal{D}_n := \{(X_i, Y_i)\}_i \subset \mathcal{X} \times \mathcal{Y}$ を時刻 n のデータと呼ぶ。

また、データ観測前の σ 加法族を $\mathcal{F}_0$、データ $\mathcal{D}_n$ 観測後の σ 加法族を $\mathcal{F}_n$ と呼ぶ。また、D_n と書いた場合、σ 加法族と ω の組を略記したものとする。この定義の詳しい意義については、6 章のベイズの話で解説する。[*3]

定義 2.3　推定量

データ観測後の σ 加法族 $\mathcal{F}_t$ に対して可測な確率変数のことを推定量という。

データは $\mathcal{D}_1$ 一つだけのこともあれば、$\mathcal{D}_1, \mathcal{D}_2, ...$, と時系列に渡って増え続けることもある。

推定量とは要するに「その時刻の情報で計算可能な数値」である。

2.1　頻度論的機械学習問題の定義

2.1.1　仮説空間

特徴量空間からラベル空間への写像のうち、手法によってことなる条件を満たすものが為す可分ヒルベルト空間 H を仮説空間と呼ぶ。この H の中から、なるべく良い f を選ぶのが、機械学習問題の目的である。

2.1.2　損失関数

損失関数 L と仮説空間 H の組は次を満たす必要がある。

定義 2.4　許容可能な組み合わせ

確率的汎関数 $L : H \times \Omega \to \mathbb{R}$ が確率 1 で H 上フレシェ微分可能[*4]

$L(f, \omega)$ は $\mathcal{F}_1$ 可測な確率変数である。つまりデータがあってはじめて計算できる。

次の定理は関数解析的定義と、実際のアルゴリズムを結ぶ超重要定理である。

定理 2.1　フレシェ微分はユークリッド空間における勾配の拡張

H が有限次元ユークリッド空間 $\mathbb{R}^N$ と同型であるとき、フレシェ微分はユークリッド空間の勾配 $\nabla_\theta, \theta \in \mathbb{R}^N$ と一

[*3] これによりベイズ論で出てくる確率測度の super transition kernel としての定義域を σ 加法族の集合 $\times\Omega$ とすることができ、「データが増えるごとに定義域の次元があがっていく」という問題を解決することができそうである。可測性を定義するための σ 加法族の集合に対する位相の入れ方はまだ要検討なので、続報を待たれたし。基礎論勢に助けを求めながら考察中である。

[*4] 実アルゴリズムやその連続化や無限次元化等はすべて問題ないので、この条件を意識せずとも別に構わないのですが、非線形変換かつ線形変換を自明に含まない無限次元の仮説空間という非常に病的な機械学習問題を設定する場合に必要になってきます。重箱の隅を突くなんてレベルじゃない話ですが、問題設定は可能な限り一般的に書きたいのでこんな条件が出てくる。

致する。

この関数解析学における基本的な定理により、本書の定義が実アルゴリズムの一般化であると言うことができる。

2.1.3 更新則

初期値関数 $f_0 \in H$ を乱数等によって決定したうえで、この関数は次のように更新されていく。[*5]

$$f_{n+1} := f_n - \alpha_n \nabla_f L(f_n, \omega) \tag{2.1}$$

$\{\alpha_n\}$ の列を学習率といい、$1e-4$ など小さめの数に設定される。後述の焼きなまし法やロビンスモンロー条件等により、数値が下がっていく場合も多い。

この更新を繰り返す行為を「学習」と呼び、$n \to \infty$ で $argmin_{f \in H} E[L(f, \omega)]$ に収束することが示されているのが望ましい。[*6]

2.2 機械学習の定義

機械学習とはなんなのか

機械学習とは、AI とは、この特徴量とラベルの関係を記述する関数 $f : \mathcal{X} \to \mathcal{Y}$ が我々の観測不能な領域に存在すると考え、真の f をデータや様々な手法によって手元のコンピューター上に「近似」する手法である。

強化学習などの例外を踏まえるといささか乱暴な物言いであるが、初学者や機械学習を数理的に捉えなおしたい方はまずはこの認識を持ってほしい。

2.3 ベイズ論的問題設定

本書においてはあまりベイズ論には触れるつもりがないが、せっかくの問題設定なのでベイズ論にも応用しておく。

定義 2.5 頻度論とベイズ論

$\mathcal{Y}$ が $\mathbb{C}^n \mathrm{or} \mathbb{R}^n$ であるとき、この機械学習問題を頻度論的機械学習問題という。

$\mathcal{Y}$ が $\mathbb{C}^n \mathrm{or} \mathbb{R}^n$ を台に持つ確率測度の集合であるとき、これをベイズ論的機械楽手問題という。

$\mathcal{Y}$ が有限集合上の測度であるとき、頻度論かベイズ論かは事前分布の仮定を置くかによる。

$\mathcal{X}$ から確率測度の台 $\mathbb{C}^n \mathrm{or} \mathbb{R}^n$ への写像の集合で、適当なヒルベルト空間に制限したものを $\tilde{H}$ とおく

頻度論ではただ一つ真の f が存在すると考えたが、必ずしもそうではないと考えるのがベイズ論である。

$\tilde{H}$ 上に値を取る確率変数 $B(\omega)(f)$ を考える。当然この分布はデータとは独立ではない。

B の $\mathcal{F}_0$ 条件付き分布を事前分布[*7]、$\mathcal{F}_1$ 条件付き分布を事後分布という。[*8]

定理 2.2 ベイズの定理

B の $\mathcal{F}_t$ 条件付き分布を μ_t と置くと

$$d\mu_1(g) = \frac{\mathcal{L}(\mathcal{D}_1|g) d\mu(g)}{\int_H \mathcal{L}(\mathcal{D}_1|g') d\mu_0(g')} \tag{2.2}$$

これを用いて、ベイズ推論による f は次のように構成される。

定義 2.6 予測分布

[*5] f_0 の決定方法は手法等により様々なテクニックが存在します。

[*6] 本当は解析的に $argmin_{f \in H} E[L(f, \omega)]$ を計算できるのが望ましいですが、そのような状況はまずありません。

[*7] データを見る前に知っている情報等から構成される。

[*8] データを見たあとの分布です

ベイズ推定による f を構成する。$\mathcal{Y}$ の台上の測度 $f(x)$ を、x に対するラベルの予測分布という。

$$df(x)(y) := \int_H 1_{g(x)=y} d\mu_1(g) \tag{2.3}$$

要するに、ラベル空間の代わりにラベル空間上の測度をラベルと見做すのがベイズ論である。この考え方は、本書終盤の部分観測マルコフ決定過程と共通する。

これは謂わば最尤推定のベイズ化と言えるが、後の様々な頻度論的手法もベイズ化が可能である。具体的にどのような構成になるかは、あまりに記述が煩雑になり読者を置いてけぼりにしすぎるため、興味のある方のみ取り組んでほしい。

相当関数解析力が鍛えられるし、後の強化学習の章ではとんでもなく複雑な関数解析を乱用するため、良い練習にもなるだろう。

2.4 損失関数の構成例

ここでは、L の具体的な構成について、例を交えつつ解説する。

損失関数は基本的に経験損失項と正則化項に分けられる。

2.4.1 経験損失

回帰問題においては、平均二乗誤差という概念を用いる。

二乗誤差とはその名の通り誤差の二乗で $|f(X)-Y|^2$ とすることで計算できる。この期待値を求めたい。X は i.i.d サンプリングであるため、X の出やすさの分布を $dp_X(x)$ という測度で表現するとすると、

$$E[|f(X)-Y|^2] = \int_{\mathcal{X}} |f(x)-y|^2 dp_X(x) \tag{2.4}$$

$$= \frac{1}{|\mathcal{D}|} \sum_{X_i, Y_i \in \mathcal{D}} |f(X_i)-Y_i|^2 \quad \text{（モンテカルロ近似）} \tag{2.5}$$

期待値を計算する際はモンテカルロ法であるため、データの数が多いほど近似制度が良くなる。

次に分類問題について。こちらは基本的に $\mathcal{Y}$ が有限集合でありを台に持つ確率測度の隔たり[*9]を測る。

台となる有限集合を $\tilde{\mathcal{Y}}$ と置き、データ Y はラベル i が正解であるとき、$Y = (0, ..., 1, ...0)$(i 番目の成分のみ 1 で残りが 0) と表記することにする。有限集合上の確率測度のベクトル表記である。

近似する関数 f は x を受け取り $\tilde{\mathcal{Y}}$ 上の測度を返す。これを用いて、「正解ラベルに対して何パーセントの確率を返したか」で損失を測る。正解ラベルが 5 のとき、ラベル 5 である確率を６０％と判定した場合と、40 ％と判定した場合では、当然後者のほうが損失が多くなる。

この計算には対数 $-\log p$ を用いる。$p=1$ つまり 100 ％そのラベルが正解であると判定しているなら、損失は 0 となる。

$$E[\sum_{j \in \tilde{\mathcal{Y}}} -Y_{\cdot_j} \log f(X)(j)] = \int_{\mathcal{X}} \sum_{j \in \tilde{\mathcal{Y}}} -Y_{\cdot_j} \log f(X)(j) dp_X(x) \tag{2.6}$$

$$= -\sum_{i \in \mathcal{D}} \sum_{j \in \tilde{\mathcal{Y}}} -Y_{i_j} \log f(X)(j) \tag{2.7}$$

この測度同士の隔たりの計算方法をクロスエントロピー誤差と言う。

問 2.1　これらの損失関数のベイズ版を考察せよ。

[*9] 回帰問題はただのユークリッド距離でしたが、こちらは距離の公理を満たさないので距離ではありません。そのため「距離」ではなく「へだたり」と書きます。

2.4.2 正則化

突然だが、ここで数列クイズである。

$$0, 3, 8, 15, 24, 35, \tag{2.8}$$

と来た時、次の数字はなんだろうか。

賢明な読者なら、$a_n = n^2 - 1$ とすぐに気づき、48 と答えるだろう。しかし、本当にそれだけだろうか。

例えば $n^1938576 - 34567n^35244 + ...$ といったトンデモ式で表現される可能性は考えられないだろうか。

しかし、誰もが $n^2 - 1$ が正しそうだと考える。これは「オッカムの剃刀」（カミナリではない）と呼ばれる考え方で、「どうせ事象が表現可能なら、単純な構造の方が正しそう」という考え方だ。

これを実現するために、f の複雑さそのものを「損失」と考え、経験損失と足し合わせる。これが正則化である。

代表的なものは Ridge 正則化と Losso 正則化である。

- Ridge 正則化:$||f||_H^2$
- Losso 正則化 H がユークリッド空間であるとき、マンハッタン距離 $|\theta_1| + |\theta_2| + |\theta_3| + ...$ と計算する

Losso はパラメータが疎になりやすい（0 になるパラメータが多い）というメリットがあり、Ridge 正則化は非常に数学的に扱いやすいメリットがある。本書では特に表記なき場合正則化と言えば Ridge 正則化である。

他にも、原点周りのみ線形で原点から離れたら 2 乗するいいとこどりの手法も存在する。

正則化項は上記項を 0.0001 倍など小さな数を掛けて（正則化率という）経験損失に足し合わせる。

H がユークリッド空間と同型であるとき、これは「大きなパラメータをなるべく使わない」ということである。特に Ridge 回帰では、「コンパクト集合外では原点に押し戻す作用がパラメータの大きさに線形に働く」[*10]という性質が、GLD 等のエルゴード性の考察において非常に重要になって来る。

2.5 機械学習の具体的な問題設定

損失関数の構成を踏まえて、有名な機械学習問題等を本書定義に当てはめていく。

2.5.1 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。

$$\mathcal{X} := \mathbb{R}^d \tag{2.9}$$

$$\mathcal{Y} := \mathbb{R} \tag{2.10}$$

$$\mathcal{H} := \{f : \mathcal{X} \to \mathcal{Y}, f(x) = \sum_{j=0}^{J} a_j \phi_j(x), a_j \in \mathbb{R}, \phi_j := x^j\} \tag{2.11}$$

$$L(f) := \frac{1}{n} \sum_{i=1}^{n} |y_i - f(x_i)|^2 \tag{2.12}$$

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

[*10] 2 次なので勾配を計算すると線形オーダーになります。

2.5.2 具体例 2: 線形回帰（確率的勾配法）

$$\mathcal{X} := \mathbb{R}^d \tag{2.13}$$
$$\mathcal{Y} := \mathbb{R} \tag{2.14}$$
$$\mathcal{H} := \{f : \mathcal{X} \to \mathcal{Y}, f(x) = \sum_{j=0}^{J} a_j \phi_j(x), a_j \in \mathbb{R}, \phi_j := x^j\} \tag{2.15}$$
$$L(f, \omega) := \frac{1}{|I(\omega)|} \sum_{i \in I(\omega)} |y_i - f(x_i)|^2 \tag{2.16}$$

$I(\omega) \subset \mathcal{D}$ 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。

2.5.3 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。

$$\mathcal{H} := \mathcal{H}_k \tag{2.17}$$
$$L(f) := \frac{1}{n} \sum_{i \in I(\omega)} |y_i - f(x_i)|^2 \tag{2.18}$$

2.5.4 具体例 4: 浅いニューラルネット（回帰）

$$\mathcal{X} := \mathbb{R}^d \tag{2.19}$$
$$\mathcal{Y} := \mathbb{R} \tag{2.20}$$
$$\mathcal{H} := \{f : \mathcal{X} \to \mathcal{Y}, f(x) = W_2 \eta(W_1 x - b_1) - b_2 \phi_j(x), W_1 \in \mathbb{R}^{L \times d}, W_2 \in \mathbb{R}^{1 \times L}, b_1 \in \mathbb{R}^L, b_2 \in \mathbb{R}\} \tag{2.21}$$
$$L(f) := \frac{1}{n} \sum_{i=1}^{n} |y_i - f(x_i)|^2 \tag{2.22}$$

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 Lipsitz 連続な関数 $\sigma : \mathbb{R} \to \mathbb{R}$ を用いて

$$\eta_i(z) := \sigma(z_i) \tag{2.23}$$

として定義される。

2.5.5 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。

$$y_{ij} := \begin{cases} 1(c_i = c^{(j)}) \\ 0(else) \end{cases} \tag{2.24}$$

このうえで、$\mathcal{H}, L$ は次のように定義される。

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

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。

2.5.6 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。

$$\mathcal{H} := \{f : \mathcal{X} \to \mathcal{Y}, f(x) = V_{end} g_K \circ g_{K-1} \circ ..... \circ g_1(x) - b_{end}, g_i(x) = \eta(W_i x - b_i)\} \tag{2.25}$$

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

$$W_i \in \mathbb{R}^{L_i \times L_{i-1}} \tag{2.26}$$

$$b_i \in \mathbb{R}^{L_i} \tag{2.27}$$

$$W_{end} \in \mathbb{R}^{1 \times L_K} \tag{2.28}$$

$$b_{end} \in \mathbb{R} \tag{2.29}$$

$K = 1$ 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。

2.5.7 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。ここでは本書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

$$V_t = V_{t_i} (t \in [t_i, t_{i+1})) \tag{2.30}$$

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

$$\frac{dX_t^i}{dt} = g(t, X_t^i) \tag{2.31}$$

$$X_0^i = x_i \tag{2.32}$$

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

$$L(f) := \frac{1}{n} \sum_{i=1}^{n} |y_i - h(X_T^i)|^2 \tag{2.33}$$

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

$$L(f) := \frac{1}{n} \sum_{i=1}^{n} L^{(i)}(f) \tag{2.34}$$

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 2.7 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。

$$F(X_T^i) := |y_i - h(X_T^i)|^2 \tag{2.35}$$

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。

$$L(f) = \frac{1}{n}\sum_{i=1}^{n} F(X_T^i) \tag{2.36}$$

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。

2.5.8 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。

$$x_{i+1} = x_i + b_i f(i, x_i) \tag{2.37}$$

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。

2.5.9 具体例 9:ODENet

$$\mathcal{H} = \{f : \mathcal{H}_1 \times \mathcal{H}_2 \to \mathbb{Y}, f(x) = h_{ab}(x_T)\} \tag{2.38}$$
$$\mathcal{H}_1 = \{h : \mathbb{T} \times \mathcal{X} \to \mathcal{X}, h(t,x) = V_t\eta(W_t x - b_t^1) - b_t^2 \tag{2.39}$$
$$\mathcal{H}_2 = \{h_{ab} : \mathcal{X} \to \mathcal{Y}, h_{ab}(x) = V_{ab}\eta(W_{ab}x - b_{ab}^1) - b_{ab}^2\} \tag{2.40}$$

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

2.5.10 具体例 10:SDENet

上記の ODEnet を確率化する。

$$\mathcal{H} = \{f : \mathcal{H}_1 \times \mathcal{H}_2 \to \mathbb{Y}, f(x) = h_{ab}(X_T)\} \tag{2.41}$$
$$\mathcal{H}_1 = \{h : \mathbb{T} \times \mathcal{X} \to \mathcal{X}, h_1(t,x) = V_t\eta(W_t x - b_t^1) - b_t^2, h_2 = V_t^2\eta(W_t^2 x - b_t^{12}) - b_t^{22}\} \tag{2.42}$$
$$\mathcal{H}_2 = \{h_{ab} : \mathcal{X} \to \mathcal{Y}, h_{ab}(x) = V_{ab}\eta(W_{ab}x - b_{ab}^1) - b_{ab}^2\} \tag{2.43}$$

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 Lipschitz 連続で t に対して 1/2 次 Hölder 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 Malliavin 微分や部分積分の議論に持ち込める。

2.6 活性化関数の具体例

上述の $\sigma(x)$ について、一応大域的 Lipsitz 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。

2.6.1 ReLu

$$\sigma(x) := max(0, x) \tag{2.44}$$

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

本書では連続的に微分可能ではないこと、零点が Lebesgue 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。

2.6.2 swish

$$\sigma(x) := x \cdot sigmoid(x) \tag{2.45}$$

$$sigmoid(x) := \frac{1}{1 + e^{-x}} \tag{2.46}$$

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 Lebesgue 測度 0 であることなどから、準楕円性などを考察する状況おいては非常に都合がよく、原則こちらを用いることにする。

3　入門：ニューラルネットの積分表現理論

3.1　ニューラルネットの連続化

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。

$$f(x) = W_2\eta(W_1x - b_1) \tag{3.1}$$

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

$$f(x) = \sum_{i=1}^{J} c_i\sigma(a_i \cdot x - b_i) \tag{3.2}$$

ここで、$J \to \infty$ とした形

$$f(x) = \int_{\mathbb{R}^{d+1}} \gamma(a, b)\sigma(a_i \cdot x - b_i)dadb \tag{3.3}$$

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、[2] で提唱された再生核 Hilbert 空間上の Ridgelet 解析を [3] による再生核 Hilbert 空間の理論により我流の再構成を行う。

3.2　再生核 Hilbert 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。

$$f(x) = \langle F, h(x)\rangle_{\mathcal{G}} \tag{3.4}$$

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。

$$||f||_{\mathcal{E}(S)} := inf\{||F||_{\mathcal{G}} : SF = f\} \tag{3.5}$$

定理 3.1　再生核 Hilbert 空間 ([3])

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

$$k(x, y) := \langle h(y), h(x)\rangle_{\mathcal{G}} \tag{3.6}$$

この時、$\mathcal{E}(S)$ は再生核 k を持つ再生核 Hilbert 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 Hilbert 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 ([3])

任意の $f \in H_k$ に対し

$$||f||_{H_k} = ||F^*||_{\mathcal{G}} \tag{3.7}$$

を満たす $F^* \in \mathcal{G}$ が一意に存在する

この F^* を f の Ridgelet 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

$$u_j(x) := \langle v_j, h(x) \rangle_{\mathcal{G}} \tag{3.8}$$

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

$$\begin{aligned}
||u_j||^2_{L^2(\mathcal{X},\mu)} &= \int_{\mathcal{X}} u_j(x)\overline{u_j}(x)d\mu(x) && (3.9)\\
&= \int_{\mathcal{X}} (\int_{\mathbb{R}^{d+1}} v_j(z)\overline{h(x)(z)}d\lambda(z) \overline{\int_{\mathbb{R}^{d+1}} v_j(z)\overline{h(x)(z)}d\lambda(z)})d\mu(x) && (3.10)\\
&\leq \int_{\mathcal{X}} (\int_{\mathbb{R}^{d+1}} v_j(z)\overline{v_j(z)}d\lambda(z) \int_{\mathbb{R}^{d+1}} h(x)(z)\overline{h(x)(z)}d\lambda(z))d\mu(x) (Schwarz\ の不等式) && (3.11)\\
&= K||v_j||^2_{\mathcal{G}} && (3.12)
\end{aligned}$$

$u_j = Sv_j$ 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

$$||SF||_{L^2(\mathcal{X},\mu)} \leq K||F||_{\mathcal{G}} \tag{3.13}$$

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

$$F^* = \sum_j c_j v_j \tag{3.14}$$

$$f = \sum_j c_j u_j \tag{3.15}$$

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、Parseval 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　Ridgelet 変換の積分表示定理 ([3])

(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

$$(\mathcal{R}_N f)(z) := \int_{E_N} f(x)\overline{h(x)(z)}\mu(dx) \tag{3.16}$$

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

$$||\mathcal{R}_N f - \mathcal{R} f||_{\mathcal{G}} \to 0 (N \to \infty) \tag{3.17}$$

この定理によって F^* はまさに既存の Ridgelet 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 Ridgelet 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 Ridgelet 解析では、この Ridgelet 変換を近似するにあたって

$$(\mathcal{R} f)(a, b) : = \int_{\mathbb{R}^d} f(x)\sigma(a \cdot x - b)dx \tag{3.18}$$

$$\approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)\sigma(a \cdot x_i - b) \tag{3.19}$$

$$\approx \frac{1}{n}\sum_{i=1}^{n} y_i\sigma(a \cdot x_i - b) \tag{3.20}$$

とするしかない。しかしこの近似は積分の測度が Lebesgue 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。

3.3　(先端研究) 定義域の制限

Relu は実装が楽で計算も早いが、簡単に爆発するため扱いにくい。[*11]

[4] では

[*11] 厳密には爆発を抑えてくれない。

4 残差学習の微分方程式解釈

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[4] 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。([4] 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 [5] 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。そのため本書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

$$x_{n+1} = x_n + f_n(x_n) \tag{4.1}$$

これは常微分方程式の Euler 近似に他ならない。

[6] ではこの考えのもとで、

- ResNet:Euler 法
- PolyNet: 後退 Euler 法
- FractalNet:2 次 Runge–Kutta 法
- RevNet: 連立 ODE の Euler 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 [5] は、さらにこの考えを発展させ

$$\dot{x_t} = f(t, x_t) \tag{4.2}$$

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 [6] 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

$$X_{n+1} = X_n + b_n f_n(X_n) \tag{4.3}$$

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 Bernoulli 分布である。

この確率過程は

$$dX_t = p(t)f(t, X_t)dt + \sqrt{p(t)(1 - p(t))}f(t, X_t)dB_t \tag{4.4}$$

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

より一般の確率微分方程式

$$dX_t = f(t, X_t)dt + g(t, X_t)dB_t \tag{4.5}$$

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 Feynman–Kac 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。

$$X_t^{s,x,\theta} = x + \int_s^t f(r, X_r^{s,x,\theta}, \theta)dr + \int_s^t g(t, X_r^{s,x,\theta}, \theta)dB_r \tag{4.6}$$

$$Y_t^{s,x} = F(X_T) + \int_t^T Z_s dB_s \tag{4.7}$$

活性化関数は ReLu なので、Lipschitz 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。

[12] の定理を用いて

$$\frac{\partial u}{\partial t}(t, x, \theta) + \mathcal{L}u(t, x, \theta) = 0 F(x) = u(T, x, \theta) \tag{4.8}$$

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。

$$\mathcal{L}u(t, x, \theta) := \nabla_x u(t, x, \theta) f(t, x, \theta) + \sum_{i,j}^{d} [g^* g]_{i,j}(t, x, \theta) \frac{\partial^2 g}{\partial x_i \partial x_j}(t, x, \theta) \tag{4.9}$$

$u(0, x, \theta)$ 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 [8][9] より次の定理が成り立つ。

定理 4.1 $f(t, x) := w_2 \eta(w_1 x - b_1) + b_2$ と置く。ただし η は Relu であるとする。

SDE の解と Brown 運動が次の等式を満たす。

$$P(\int_0^T f^2(X_t)dt < \infty) = 1 \tag{4.10}$$

$$P(\int_0^T f^2(B_t)dt < \infty) = 1 \tag{4.11}$$

このとき、次の等式が言える。

$$u(t, x, \theta) = E^x[F(W_{T-t}) exp[\int_0^{T-t} f(t + r, B_r, \theta)dB_r - \frac{1}{2}\int_0^{T-t} ||f(t + r, B_r, \theta)||^2 dr]] \tag{4.12}$$

proof.

[8] より前半部分が言えれば Girsanov 変換ができ、Girsanov の定理を使えれば [9] 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。

$M(\omega) := max_{0 \le t \le T} W_t(\omega)$ と置くと、$M(\omega)$ は確率 1 で有限。

よって

$$\int_0^T f^2(W_t)dt \leq TM^2(\omega) < \infty \tag{4.13}$$

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

$$u(t, x) = \int_{\mathbb{R}} F(y)p(t, x, T, y)dy \tag{4.14}$$

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

$$\int_0^{T-t} b(r, B_r, \theta) \odot dB_t = B(T - t, B_{T-t}, \theta) - \frac{1}{2}ab\int_0^{T-t} 1_{bB_s - c > 0}ds \tag{4.15}$$

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 Brown 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。
[9] より、0 出発で時刻 r で y にたどり着く Brown 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

$$P_l^r(\tau|y) = \begin{cases} 1 - (r - \tau)e^{-\frac{c}{\tau} + \frac{y^2}{2}}(e^c(2c + 1)erfc(\sqrt{c}) - 2\sqrt{\frac{c}{\pi}}) & y \leqq l \\ \int_0^\tau \frac{(\tau - u)e^{\frac{y^2}{2} - \frac{l^2}{2(r-u)} - \frac{(y-l)^2}{2u}}}{\sqrt{2\pi}(u(r-u))^{\frac{3}{2}}} \times (\frac{l(y-l)^2}{u} - \frac{(y-l)^2 l^2}{r-u} + y - 2u)du & 0 \leqq l \leqq y \\ 1 - P_{-y}^r(r - \tau) & l \leqq 0 \end{cases} \tag{4.16}$$

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

$$\partial_t u(t, x, \theta) - \nabla_x u(t, x, \theta)f(t, x, \theta) = 0 \tag{4.17}$$
$$u(T, x) = F(x) \tag{4.18}$$

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向上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。

$$X_{t_{n+1}} = X_{t_n} + f(t, X_{t_n})(t_{n+1} - t_n) + g() \tag{4.19}$$

定理 4.3 近似誤差収束定理 [13]

f, g が Lipschitz 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

$$sup_N E[|X_T^N|^p] < \infty \tag{4.20}$$

$$E[sup_{t\in[0,T]}|X_t - X_t^N|^p] \to 0 (N \to \infty) \tag{4.21}$$

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

$$|\frac{\partial}{\partial\theta}E[F(X_T)] - \frac{\partial}{\partial\theta}E[F(X_N)]| \leq C(T, x, f)\frac{T}{N} \tag{4.22}$$

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。

定理 4.4 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。

$$dX_t^\theta = X_0 + \int_0^t f(s, X_s^\theta, \theta)ds + \int_0^t g(s, X_s^\theta, \theta)dB_s \tag{4.23}$$

$$dX_t^{\theta,N} := X_0 + \int_0^t f(\psi(s), X_{\psi(s)}^{\theta,N})ds + \int_0^t g(\psi(s), X_{\psi(s)}^{\theta,N})ds \tag{4.24}$$

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 Lipschitz 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。

$$L_1 := -(\frac{\partial f}{\partial x})^2 + \frac{\partial\frac{\partial f}{\partial x}f}{\partial x} \tag{4.25}$$

$$L_{n+1} : = \frac{\partial f}{\partial x}L_n - \frac{\partial L_n f}{\partial x} \tag{4.26}$$

$$\tag{4.27}$$

$[f(0, x), L_1 f(0, x), ..., L_{d-1}f(0, x)]$ 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

$$|\frac{\partial}{\partial\theta}E[F(X_T)] - \frac{\partial}{\partial\theta}E[F(X_N)]| \leq K(T, x)M(f, g, F)||1/\det(\gamma_T)||_p^q\frac{T}{N} \tag{4.28}$$

ただし、$\gamma_T := \gamma_{X_T}$ であり、確率変数 X_T の Malliavin 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

$$sup_{t,x,\alpha,\alpha`\in[0,T]\times\mathbb{R}^d\times\mathcal{A}\times\mathcal{A}}\frac{|v(t,x,\alpha)-v(t,x,\alpha`)|}{|\alpha-\alpha|^\eta} \tag{4.29}$$

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。

補題 4.1　勾配過程の表記 [11]

$f, g, \partial_\alpha f, \partial_\alpha g$ が共に任意の t, α に対して、x に大域的 Lipsitz 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

$$Y_t = I_d + \int_0^t \partial_x f_s Y_s ds + \int_0^t \sum_{j=1}^n (\partial_x g_s)_j Y_s^j dB_s^j \tag{4.30}$$

$$Z_t = I_d - \int_0^t Z_s(\partial_x f_s - \sum_{j=1}^q (\partial_x g_s)_j^2)ds - \sum_{j=1}^n \int_0^t Z_s(\partial_x g_s)_j dB_t^j \tag{4.31}$$

$$\dot{X}_t = \int_0^t \partial_\alpha f_s + \partial_x f_s \dot{X}_s ds + \sum_{j=1}^n \int_0^t \partial_\alpha g + \partial_x g \dot{X}_s dB_s^j \tag{4.32}$$

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

$$\dot{X}_t = Y_t \int_0^t Z_s[(\partial_\alpha f - \sum_{j=1}^n \partial_x g_{j,s}\partial_\alpha g_{j,s})ds + \sum_{j=1}^n \partial_\alpha g_{j,s} dB_s] \tag{4.33}$$

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 [10]

上述の Lipschitz 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 Sobolev 空間 $\mathbb{D}^{k,p}$ を用いて

$$\mathbb{D}^\infty := \cap_{k,p\geq 1}\mathbb{D}^{k,p} \tag{4.34}$$

と定義される。

補題 4.3　Malliavin 微分 [10]

$$\mathcal{D}_s X_t = Y_t Z_s g(s, X_s) 1_{s\leq t} \tag{4.35}$$

補題 4.4　Clark の表現定理 [10] $X_T \in \mathbb{D}^\infty$ 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)

$$X_T = E[X_T] + \int_0^T E[\mathcal{D}_t X_T | \mathcal{F}_t] dB_t \tag{4.36}$$

定義 4.1　Malliavin 共分散行列 [10]

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

$$\gamma_A := \int_0^T D_t A[D_t A]^* dt \tag{4.37}$$

この行列が確率 1 で正則で

$$\gamma_A \in \cap_{p\geq 1} L^p(\Omega) \tag{4.38}$$

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 Hörmander 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　Hörmander 条件 (Brown 運動一次元)

次のようなベクトル場を考える。

$$D = \sum_{i=1}^{d} f_i(0, x_0)\frac{\partial}{\partial x_i} - \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} g_j(0, x_0)\frac{\partial g_i}{\partial x_j}(0, x_0)\frac{\partial}{\partial x_i} \tag{4.39}$$

$$C = \sum_{i=1}^{d} g_i(0, x_0)\frac{\partial}{\partial x_i} \tag{4.40}$$

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 Hörmander 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、Brown 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 Hörmander 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 Hörmander 条件を満たす点なら、これは C^∞ 級

$$p_T(x) = \int_{\mathbb{R}^d} p_T(x|x_0)p_0(x_0)dx_0 \tag{4.41}$$

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

$$dX_t = p(t)f(t, X_t)dt + \sqrt{p(t)(1 - p(t))}f(t, X_t)dB_t \tag{4.42}$$

f が g の定数倍であること、Brown 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、Hörmander 条件も自明に満たされない。

通常、Brown 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、Hörmander 条件は自明に満たされる。今回のような Brown 運動の次元が小さい場合は、Hörmander 条件は Lie 括弧積の計算をして確認せざるをを得ない。

定理 4.6　Hörmander 条件

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

$$L_1 := -(\frac{\partial f}{\partial x})^2 + \frac{\partial \frac{\partial f}{\partial x} f}{\partial x} \tag{4.43}$$

$$L_{n+1} := \frac{\partial f}{\partial x} L_n - \frac{\partial L_n f}{\partial x} \tag{4.44}$$

$$\tag{4.45}$$

$[f(0,x), L_1 f(0,x), ..., f_{d-1}(0,x)]$ 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 Hörmander 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。

$$C = g \tag{4.46}$$

$$D = f - \frac{1}{2}\frac{\partial g}{\partial x} g \tag{4.47}$$

$$[X, Y] = \frac{\partial B}{\partial x} A - \frac{\partial A}{\partial x} B \tag{4.48}$$

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

$$[D, C] = \frac{\partial g}{\partial x}(f - \frac{1}{2}\frac{\partial g}{\partial x} g) - (\frac{\partial f}{\partial x} - \frac{1}{2}\frac{\partial \frac{\partial g}{\partial x} g}{\partial x}) g \tag{4.49}$$

$$D^0 := D \tag{4.50}$$

$$D^{n+1} := [D^n, C] \tag{4.51}$$

$$L_1 := -(\frac{\partial g}{\partial x})^2 + \frac{\partial \frac{\partial g}{\partial x} g}{\partial x} \tag{4.52}$$

$$L_{n+1} := \frac{\partial g}{\partial x} L_n - \frac{\partial L_n g}{\partial x} \tag{4.53}$$

$$\tag{4.54}$$

として定義すると任意の自然数 n に対して

$$D^n := L_n g \tag{4.55}$$

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 Hörmander 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 [10]

β を多重指数とする。また $A \in \mathbb{D}^{k_1,\infty}$ の Malliavin 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

$$E[\partial_\beta F(A) B] = E[F(A) H_\beta(A, B)] \tag{4.56}$$

定理 4.8　エラー確率変数の評価 [16]

測度 $\tilde{\mu}$ を次のように定義する。

$$\int_{\mathbb{R}^d} h(x)\tilde{\mu}(x) = E[h(X_T^{0,N}) + h(X_T^{1,N})] + \int_0^1 E[h(X_T^{\lambda,N})]d\lambda \tag{4.57}$$

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。

$$\mathcal{E}_1(\epsilon, m) : = E[F_m(X_T^\epsilon)H_T - F_m(X_T^{\epsilon,N})H_T] \tag{4.58}$$

$$\mathcal{E}_2(\epsilon, m) : = E[F_m(X_T^{\epsilon,N})H_T - F_m(X_T^{\epsilon,N})H_T^N] \tag{4.59}$$

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。

$$|\mathcal{E}_1(m)| \le K_1(T,x)(||F_m(X_T^\epsilon)||_{L^2} + ||F_m(X_T^{\epsilon,N})||_{L^2} + \int_0^1 ||F_m(X_T^{\lambda,N,\epsilon})||_{L^2}d\lambda)||1/det(\gamma_T)||_{L^p}^q \frac{T}{N} \tag{4.60}$$

$$|\mathcal{E}_2(m)| \le K_2(T,x)(||F_m(X_T^\epsilon)||_{L^2} + ||F_m(X_T^{\epsilon,N})||_{L^2})||1/det(\gamma_T)||_{L^p}^q \frac{T}{N} \tag{4.61}$$

この定理を用いて上記定理を証明する。

f, g は x に対して大域的 Lipsitz 連続かつ t に対して 1/2 次 Hölder 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、Lebesgue の収束定理を用いて

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

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、定理を得る。

問 4.1　4 章　定理の汎用性

この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、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、Hörmander 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 Lipsitz 連続で t に対して 1/2 次 Hölder 連続、かつ F がたかだか多項式増大であれば本章主定理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

$$\int_{\mathbb{R}^d} c(x, T(x))\mu(dx) \tag{4.62}$$

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 が任意のぼれる加速集合 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

$$\pi[A \times \mathbb{R}^d] = \mu(A) \tag{4.63}$$
$$\pi[\mathbb{R}^d \times A] = \tau(A) \tag{4.64}$$

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

$$argmin_{\pi \in \Pi(\mu,\tau)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y)\pi(dxdy) \tag{4.65}$$

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 [14]

$c(x, y) = |x - y|^2$ 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 Lebesgue 測度に絶対連続なら、最適輸送問題には解が存在し

$$\pi(\mu, \tau) = (id_{\mathbb{R}^d}, \nabla\phi)_{\#}\mu \tag{4.66}$$

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 [15]

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

$$\dot{x}_t = \nabla V_t(x_t) \tag{4.67}$$

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

$$p_t(x_t)|\nabla_{x_0} x_t| = p_0(x_0) \tag{4.68}$$

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

$$\partial_t p_t(x) = -\nabla \cdot [p_t(x)\nabla V_t(x)] \tag{4.69}$$

$\nabla\cdot$ 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 Riemannian 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、

$$\dot{x}_t = f(t, x_t) \tag{4.70}$$
$$f(t, x) = V(t)\eta(W(t)x + b_1(t)) + b_2(t) \tag{4.71}$$

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

$$F(t, x_t) = \Psi(W_1(t)x + b_1(t)) + \langle b_2(t), x\rangle + C \tag{4.72}$$

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、

$$\text{任意の } i, j \in \{1, 2, ..., d\} \text{ に対して}$$
$$\frac{\partial f_i}{\partial x_j} = \frac{\partial f_j}{\partial x_i} \tag{4.73}$$

がポテンシャルの存在と同値である。

$y := \eta(Wx)$ として

$$f_j(x) = V_j y \tag{4.74}$$
$$= \sum_{l=1}^{L} V_{j,l} y_l \tag{4.75}$$
$$y_l = \sigma(W_l x) \tag{4.76}$$
$$= \sigma(\sum_{k=0}^{d} W_{l,k} x_k) \tag{4.77}$$
$$\frac{\partial f_j(x)}{\partial x_i} = \sum_{l=1}^{L} V_{j,l} \frac{\partial y_l}{\partial x_i} \tag{4.78}$$
$$= \sum_{l=1}^{L} V_{j,l} W_{l,i} \sigma^{'}(\sum_{k=0}^{d} W_{l,k} x_k) \tag{4.79}$$

同様に右辺は

$$\frac{\partial f_i(x)}{\partial x_j} = \sum_{l=1}^{L} V_{i,l} W_{l,j} \sigma^{'}(\sum_{k=0}^{d} W_{l,k} x_k) \tag{4.80}$$

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とエルゴード性

この章では最適化のアルゴリズムについて解説する。

最適化は 2 章で書いた通り、なんらかの可分ヒルベルト空間 H 上で、ある関数 $f_n \in H$ に対して、

$$f_{n+1} := f_n - \alpha_n \nabla_f L(f_n) \tag{5.1}$$

といった形で更新される。ただし ∇ は損失関数に対するフレシェ微分を表す。

理論を考えるときだけならこの無限次元で考えたほうがいい場合もあるが、実際は H を非常に高次元のユークリッド空間と同相とする。

α_n は学習率と呼ばれ、小さい数字から初めて徐々にさらに下げていく。

学習は連続化すると微分方程式となり、

$$\dot{f}_t = \nabla_f L(f_t) \tag{5.2}$$

の離散化が上記の更新式となる。ここで学習率 α_n は離散化幅にあたる。

α_n は一般的に次のロビンスモンロー条件を満たすべきとされる。

$$\begin{aligned} \sum_{n=1}^{\infty} \alpha_n &= \infty \\ \sum_{n=1}^{\infty} \alpha_n^2 &< \infty \end{aligned} \tag{5.3}$$

離散化幅の和が有限になると、到達時刻 $T := \sum_{i=1}^{n}$ が $n \to \infty$ で ∞ に発散しない。つまり途中で止まってしまい、無限ステップを踏んでも時刻が一定の値を超えなくなるので、$t \to \infty$ で成り立つ理論が一切使えなくなる。

一方で二乗の和が無限になると、離散化誤差が L^2 であっても、無限に足し合わせたとき誤差が爆発してしまう。[*12]

この観点に立てば [A bayes] は、「学習率を下げずにノイズだけを下げる方法をとることによって、同じステップ数で到達できる時刻を増やすべき」としていることになる。また、ノイズが大きい確率過程に対して離散化幅がとりにくいのも道理である。ただしこの論文に記された SGD と SDE の関係は、離散化・連続化として正当化できないことには注意されたし。[*13]

5.1　様々な勾配法アルゴリズムとその連続化

5.1.1　勾配法

$$\theta_{k+1} = \theta_k - \alpha_k \nabla_\theta L(\theta_k) \tag{5.4}$$

最も基本的な手法である。大規模パラメータの機械学習はほぼすべてこの手法の改良を用いる。

5.1.2　SGD

ミニパッチの抽出を司る Markov 過程 U_n を用いて

$$\theta_{k+1} = \theta_k - \alpha_k \nabla_\theta L(\theta_k, U_k) \tag{5.5}$$

[*12] これはロビンスモンロー条件が必要とされる理由の一説である。著者はこの派閥なのでこれを記したが、絶対の真理ではない。

[*13] 詳細は後述。私の知る限り SGD を SDE の離散化と見做すすべての論文が同じ重大な問題を抱えている。離散化幅を SDE の項に入れるなど問題外である

5.1.3 GLD

Gaussian noise$W_k \sim N(0, I_N)$ を用いて

$$\theta_{k+1} = \theta_k - \alpha_k \nabla_\theta L(\theta_k) + \sqrt{2\alpha_k} W_k \tag{5.6}$$

5.1.4 SGLD

$$\theta_{k+1} = \theta_k - \alpha_k \nabla_\theta L(\theta_k, U_k) + \sqrt{2\alpha_k} W_k \tag{5.7}$$

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)

5.2 勾配法の連続化と Lyapunov 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。

$$\theta_{k+1} = \theta_k - \alpha \nabla_\theta L(\theta_k) \tag{5.8}$$

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 Lyapunov 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

$$d\theta_t = -\nabla_\theta L(\theta_t) dt \tag{5.9}$$

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

$$||\theta_t - \tilde{\theta}_t|| \approx e^{-\lambda t} ||\theta_0 - \tilde{\theta}_0|| \tag{5.10}$$

ここで λ は最大 Lyapunov 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。

5.3 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。

$$\theta_{k+1} = \theta_k - \alpha \nabla_\theta L(\theta_k, U_k) \tag{5.11}$$

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 Markov 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

$$\theta_{k+1} = \theta_k - \Psi_1(\theta_k)\alpha + \sqrt{\Psi_2(\theta_k)}\alpha\epsilon_k \tag{5.12}$$

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、Gaussian noise 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、Lebesgue 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。つまり、離散化極限として正当化できず、ただの形式的な近似である。

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、Clark の表現定理で Malliavin 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。

5.4 GLD の連続化

勾配法連続化における、「Hesse 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。[*14]

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。

$$\theta_{k+1} = \theta_k - \alpha \nabla_\theta L(\theta_k) - \sqrt{2\alpha}\epsilon_k \tag{5.13}$$

今回は、Lebesgue 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。

5.5 発展：SDE の連続化と見做せるよう SGD の改造について

5.3 で書いた通り、SGD は通常 SDE の離散化であるとは見做せない。現状の確率分布を帳尻合わせた形式的議論ではなく本当に離散化としてみなすためには、このオーダーの問題を解決する必要がある。

そのための具体的な方法の一例と、その数値実験について書く。

（2025 年中には論文を出したいところ。論文公開と共に書きます。）

[*14] 意気揚々とすごい結果出してる最適化アルゴリズムの論文を読んだら f is convex. と書かれていてキレそうになったのは一度や二度ではない。

5.6 発展：SDE のエルゴード性と最適化アルゴリズムが非凸損失関数の大域的最適解に確率収束する条件

5.7 発展：（先端研究）現実的な計算時間で 1epoch 走らせられ、かつ非凸損失関数であっても大域的最適解に確率収束するアルゴリズム発見に向けた今後の課題。

次のような確率微分方程式の離散化を考える。

$$\theta_n = \alpha_n \nabla_\theta L(\theta_n)\Delta t + \beta_n \Delta B_n) \tag{5.14}$$

さらに、$\alpha_n \sim o(A/n), \beta_n \sim o(B/n\sqrt{\log\log n})$ が成り立つとする。
次にこれを仮定する。

$$\liminf \tag{5.15}$$

このとき、$\theta_n \to \theta^*$ が確率収束の意味で成り立つ。

5.7.1 確率収束の意味で成り立つ、とは

厳密には分布収束であるが、収束先の分布が一点集中であれば分布収束と確率収束は同値である。

6　ベイズ最適化によるハイパーパラメータ調整

最初に、ベイズ論で頻発する、super transition kernel を定義する。
これは $P(A|b;c)$ と書ける実数値関数で、

- c を固定したら $P(\cdot|\cdot;c)$ は transition kernel

となるものである。A,b を固定して c に対して可測関数になることは求めない。後述の σ 加法族を入力する事情から、c に対して可測関数にすることは非常に煩雑さを伴うからである。
あまり好ましくないがないが一般的なベイズの記法 $p(y|x)$ を用いる。
厳密に言えば x によって定まる y が属する空間上の確率測度（その確率密度関数）を指す。transition kernel については解説を略するが、本来は y が属する空間上の確率測度間に位相を定義し、$x \to p(\cdot|x)$ という x を受け取り測度を返す関数が可測関数にならねばならない。
さらにこれを σ 加法族と標本に対して定義する。すなわち、$P(y|x,\mathcal{F}_n,\omega)$ といった具合で、σ 加法族とそのデータに関する標本 ω を固定すると $P(\cdot|\cdot,\mathcal{F}_n,\omega)$ となり、これは transition kernel である。
ここで、n 回試行データの集合 D_n を考える。この D_n を可測にする最小の σ 加法族を $\mathcal{F}_n$ と書けば、データに対して transition kernel が定まる形になる。
ここで、いっそ $\mathcal{F}_n,\omega$ をデータという意味で D_n と書いてしまおう。こうしてベイズ論でよく出てくる $p(y|x,D_n)$ が公理的確率論を用いて super transition kernel としてちゃんと正当化できた。
今後は、大文字の場合 (super) trantision kernel、小文字の場合はそれに対する密度関数であるとする

6.1　ベイズの定理

逆に transition kernel の $P_2(x|y,D_n)$ 及び $\mathcal{F}_n$ で可測なランダム測度 $P_3(x|D_n)$ も同様に定義できる。
これを使ってベイズの定理を書くことができる

定理 6.1（ベイズの定理） 任意の可測関数 $f:\mathcal{Y}\to\mathbb{R}$ に対して、適当な可積分性を満たす P に対しては次の式が成り立つ

$$\int_{\mathcal{Y}} f(y)dP_1(y|x,D_n)=\int_{\mathcal{Y}} f(y)\frac{dP_2(x|y,D_n)}{dP_3(x|D_n)}dP_1(y|x,D_0) \tag{6.1}$$

両辺は $\mathcal{F}_n$ 可測な確率変数である。当然だが、右辺が意味を持つために、任意の y に対して $P_3(\cdot|D_n) >> P_2(\cdot|y,D_n)$ が確率 1 で成り立つ必要がある。[*15]
特にここまで出てきた確率測度がすべてユークリッド空間上で定義され、かつルベーグ測度に対して絶対連続なら、確率密度関数に対して次の等式が任意の x,y に対して確率 1 で成り立つ。

$$p_1(y|x,D_n)=\frac{p_2(x|y,D_n)p_1(y|x,D_0)}{p_3(x|D_n)} \tag{6.2}$$

[*15] 普通にベイズやっていればまず無意識に仮定できているので、よほど奇抜なことをやりたいのでなければ特に意識する必要はない

7　強化学習と確率制御

著者の以前の飯の種であり、ライフワークともなった研究テーマである。

7.1　マルコフ決定過程

多数の定義を書く。

- 状態空間：$\mathcal{S}$　位相空間か有限集合
- 行動空間：$\mathcal{A}$　位相空間か有限集合
- 遷移・報酬測度：$P : (\mathbb{R} \times \mathcal{S}) \times (\mathcal{S} \times \mathcal{A}) \to [0, 1]$（transition kernel）
- 方策測度：$\pi : \mathcal{A} \times \mathcal{S} \to [0, 1]$（transition kernel）
- 減価率：$\gamma \in (0, 1)$

$\mathcal{S}, \mathcal{A}$ はともに有限集合でも無限集合でもよい。どちらかが有限集合でもう片方が無限集合でもよい。

π, P の定義域となる可測空間を考えるとき、σ 加法族は、$\mathcal{S}, \mathcal{A}$ が無限集合のときはボレル集合族、有限集合のときはべき集合であるものとする。

次に、π に対する状態行動価値関数 $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ を次のように定義する。

$$Q^\pi(s, a) := E[\sum_{t=0}^{\infty} \gamma^t r_t | r_t, s_{t+1} \sim P(\cdot, \cdot|s_t, a_t), a_t \sim \pi(\cdot|s_t), s_0 = s, a_0 = a] \tag{7.1}$$

要するに方策 π を使い続けると仮定したときの現在の状態 s に対し、行動 a をとることが持つ価値である。

次に行動価値関数を定義する。

$$V^\pi(s) := max_{a \in \mathcal{A}} Q^\pi(s, a) \tag{7.2}$$

すべての方策の集合を Π と置く。最適状態価値関数 V^* を次のように定義する

$$V^*(s) := \max_{\pi \in \Pi} V^\pi(s) \tag{7.3}$$

最適状態価値関数に対応する π^* を最適方策と呼ぶことにする。この正体について深堀しよう。

7.1.1　強化学習は教師無し学習→だからってヒューリスティック要素がないわけではない!

（報酬設計がいかに経験と勘によるものかを語る）

7.2　発展：部分観測マルコフ決定過程

「状態 s_t の遷移はマルコフ過程だが、それを観測することができず、観測可能な確率過程 o_t はマルコフ過程とは限らない」という非常に厄介な状況に立ち向かうためのもの。

観測可能な確率過程の値をとる空間 $\mathcal{O}$[*16]を用いて、新たな transition kernel たる $\mu : \mathcal{S} \times \mathcal{O} \to [0, 1]$ を考える。

これは見ての通り、「観測が o のとき、実際の状態は s である確率」を表す transition kernel となる。

この測度によるルベーグ積分をマルコフ決定過程に組み込む。また、この項では P の定義を変更し super transition kernel とする。増大情報系 $\mathcal{G}_t$ を、$\{o_s\}_{0 \le s \le t}$ から生成される σ 加法族として、$P(r_t, s_{t+1}, o_{t+1}|s_t, a_t; \mathcal{G}_t)$ となる super transition kernel として定義するのである。

7.3　発展：分布型強化学習

通常の強化学習では期待値しか参照しないため、分散などもちゃんと見たい場合に使える理論。

[*16] $\mathcal{S}, \mathcal{A}$ と同じく、有限集合か位相空間。σ 加法族は有限集合の場合はべき集合、位相空間の場合はボレル集合族と定義する。

7.3.1 ドロップアウトを用いたベイズ論的分布強化学習

データ検討中に論文を先に他所に出されてブちぎれた。

7.4 発展：統一理論・超一般化マルコフ決定過程

本当はこちらの定義を先に書いて、上記 3 定義はすべてこの特別な場合と言おうとしたが、強化学習初学者相手だといくら数学徒向けとはいえ鬼畜すぎるので最後に表記する。

7.5 超発展：連続時間化

ここが本書で最も難しい。

この理論は主に停止時刻でステップを踏む確率制御をやっていくのだが、連続時間強化学習であると言い張るには「等間隔でステップを踏んだら普通の強化学習になる」ことが不可欠である。

実装上離散ステップを踏むしかないため、この連続時間強化学習では、$\tau_n \to \infty \ a.e.$ が成り立つ停止時刻の列を考え、この停止時刻でステップを踏んでいく。

ここで $\tau_n = n$ とした場合に、すべての計算が通常の離散時間強化学習と同じになるように連続時間モデルを定義したい。これでこそ「自然な連続化」「通常の強化学習の拡張」と言える。ただ単に SDE に無理やり強化学習っぽいものを当てはめても、強化学習の連続化とは言えないのである。

8 大規模言語モデル（LLM）

最近流行りの手法である。著者もこちらにキャリアの軸足を移していきたいので、サーベイ的になって恐縮だが考察を併せ記述させていただく。

8.1 transformer の構成

LLM の基礎となるアーキテクトである。ここをしっかり理解するのが、LLM 理解の早道だ。

transformer の特徴、本質と言えるのは「アテンション行列 S」である。

入力を n 個の「トークン」に分割し、$n \times n$ の行列によって、「$S_{ij} := i$ 個目のトークン要素が j 個目のトークン要素にどれだけ注意を払うか」を表す。そうなるように学習する。

そしてこのアテンション行列に対して加工を行い、その加工方法も学習するのが、transformer の特徴と言える。

このアテンション行列は、いわば相関行列の拡張である。相関行列が持つべき正定値性や、対称性を持たないため、より柔軟な対応が可能となる。厳密に相関行列ではないとは言え近い概念ではあるため、後述のマルチヘッドアテンションを除けば、「対称行列に近い」状態になることが多く、また「(i,j) と (h,k) が近いとき S_{ij} と S_{hk} は近い値になる」ことが期待される。対角成分が大きくなりやすいのも相関行列と同様である。

8.1.1 マルチヘッドアテンション

入力行列 $X \in \mathbb{R}^{d \times n}$ を、Query:Q,Value:V,Key:K に変換する。

行列 $W_Q \in \mathbb{R}^{d_1 \times d}$ と $W_K \in \mathbb{R}^{d_1 \times d}$ と $W_V \in \mathbb{R}^{d_2 \times d}$ を用いて

$$Q := W_Q X \tag{8.1}$$
$$K := W_K X \tag{8.2}$$
$$V := W_V X \tag{8.3}$$

と定義する。

次にこれらを用いて、アテンション行列 S を次のように計算する。

$$S := softmax(\frac{QK^T}{\sqrt{d_1}}) \tag{8.4}$$

これに VS を計算することで、このアテンションブロックの出力の一つになる。

この W_K, W_Q, W_V の組を「アテンションヘッド」と言う。実際には d_1, d_2 を大きくするより、ここを小さくして複数個のアテンションヘッドを用意するほうが結果が良くなりやすい。このアテンションヘッドを複数用意する計算ブロックを「マルチヘッドアテンション」という。

先ほども書いた通り、アテンション行列とは相関行列の拡張であると言える。しかし厳密に相関行列ではないため、対象である必要がなく、masked Multihead attention という後述のデコーダーにおける手法は、アテンション行列 S の成分において、$S_{ij} = 0$ が $i < j$ のとき成り立たせることになる。[*17]これは文字列、動画、音声などのタスクが持つ時系列データにおいて「未来の情報は見ない」という制約となる。

8.1.2 word Embedding と Position Embedding

本来この入力 X は文章データなどの数値ではないものである。

Position Embedding の冗長性

[*17] 実装上は 0 ではなく $1e-9$ のような非常に小さい数にすることが多い。

8.1.3 エンコーダー

エンコーダーにおいて $Q.K$ と V が等価でないことは式から明らかだが、Q, K を分ける意味はなんだろうか。これについて一つの答えとして「可変長入力を受け付けるため」

8.1.4 デコーダー

デコーダーにおいては Q, K は出自が違うので分ける意味を飲み込みやすい。

8.2 LLM モデルについて

今流行りの chatgpt を含むモデルについて解説していく。

8.3 mamba モデルについて

個人的には理論整備がある程度されているこちらのほうが好ましい

8.4 発展：アテンションの連続化と拡張カーネル関数（擬内積）

アテンション行列は相関行列の拡張と言える。相関行列の連続化はカーネル関数である。

となると、アテンション行列の連続化は「カーネル関数の拡張」と考えるのが自然であろう。正定値カーネル関数全体の集合よりも大きな「アテンション関数」とでも呼ぶべき関数のクラスを考えたい。

そもそもアテンション行列は正定値行列であるとは限らず、対称性もないので、アテンション関数のクラスは対称ではなく、正定置でもないものを許容する必要がある。しかし無駄に広い許容を行ってしまえば、関数解析による考察は困難となる。

また、アテンション行列はその定義から、X に対するたかだか 2 次増大になる。つまりアテンション関数 $M(x, y)$ も、たかだか $O(|xy|)$ 以内の増大度で合ってほしい。すなわち、ある定数 C が存在し、任意の x, y に対して

$$|M(x, y)| \leq C(1 + |xy|) \tag{8.5}$$

が成り立ってほしい。これを考えるのに都合のいい関数のクラスはなんだろうか。

また、微分可能性も欲しい。近くの点のアテンションはそう大きく変化しないと考えられ、不連続な点はなるべく許容したくない。しかしマスクドマルチヘッドアテンションへの対応を行うため、不連続な線が $x = y$ で存在することは許容したい。そのため M は「ほとんどいたるところ全微分可能」とする。また、x 軸、y 軸に対して対称なコンパクトサポートを持っていると制限してもよいだろう。そうなると上記の増大度条件はむしろ原点周りで重要になり

$$|M(x, y)| \leq \frac{C}{|xy|} \tag{8.6}$$

が新たな最低限の増大度制約になる。

これらを踏まえ、ほどよく広い関数空間として、ソボレフ空間 $W^{1,2}$ を考えよう。ここではこう書いたら「コンパクトな台を持つ」という意味を内包するものとする。既存の正定値カーネル関数は皆[*18]ほとんどいたるところ微分可能なので、k, M ともにこの関数空間内で定義された関数となる。

この空間の中で、カーネル関数の集合 $\mathcal{K}$ は閉集合ではなく、その閉包 $\overline{\mathcal{K}}$ は半正定値関数の集合となる。これは部分集合として稠密ではないので、ゲルファンドトリプルのようなアプローチは不可能となる。

では、$\overline{\mathcal{K}}$ を含む $W^{1,2}$ の部分空間はなんだろうか。これを M の属するクラスと関連付けられないだろうか。

例えば $M(x, y) := k_1(x, x) - k_2(y, y)$ とすれば、これは正定置性も対称性も失う可能性がある。この二つの特徴は

[*18] というと怒られるが、例外を見たことがないので許してほしい。

壊したいのだが、あまり雑に壊すとなにもわからなくなるので、このような程よい破壊を考えたい。

$$l(x,y) := \sum_{i=1}^{n} a_{i1}k_i(x,x) + a_{i2}k_i(x,y) + a_{i3}k_i(y,y) \tag{8.7}$$

このうえで、関数空間 $L := \{l : \{k_i\}_{i\in\{1,2,.....,n\}} \subset \overline{\mathcal{K}},\ \{a_{ij}; a_{ij} \in \mathbb{R}, i \in \{1,2,.....,n\}, j \in \{1,2,3\}\},\ n \in \mathbb{N}\}$ という、上記 l で張られる集合を考えたい。[*19]

定理 8.1　アテンション関数の近似

L は $W^{1,2}$ で稠密

proof.

よく知られた事実として、コンパクトな台を持つ無限回連続微分可能な関数の集合 C_c^∞ は $W^{1,2}$ で稠密である。そのためこの任意の元を L の元で近似できればよい。

閉作用素 T を次のように定義する

8.4.1　発展：リッジレット解析とアテンション関数

3 章で定義したリッジレット作用素の再登場である。これが連続モデルにおける Q, K, V の役割を理解するのに役立つ。

ハイパーネット

8.5　発展：高次テンソル化モデル

このモデルにおいて入力するテンソルは 2 次元テンソル（行列）である必要はない、通常の transformer をここでは「2 次 transformer」と呼ぶことにする。

アインシュタインの縮約

実は CNN、RNN は 3 次 transformer モデルの特別な場合と言える。

8.6　強化学習との関係

[*19] メモ：これ $\overline{\mathcal{K}}$ でも閉包とらなくとも、この L の閉包は同じでは？

9　（有料版限定）：本書で用いられている数学の概説

非常に簡素に、かつ高速で走り抜ける。故にここで勉強するようなことは避け、イメージを掴むもの、公式や定義を忘れたときに振り返る程度のものとしてほしい。

9.1　位相・距離・極限

おすすめの書籍：

9.2　足し算、引き算、掛け算、割り算

おすすめの書籍：

9.3　大きさ、長さ、面積、体積

おすすめの書籍：

9.4　関数解析

おすすめの書籍：

9.5　発展：確率過程と確率微分方程式

おすすめの書籍：

9.6　数理統計学

おすすめの書籍：

9.7　発展: マリアヴァン解析

おすすめの書籍：

2 → 4 9. → 5 4 → 6　アーキテクチャ、7 ODEnet 11,12 → 8，9 15 → 10 14 → 1 1 10 → 12 22 → 1 3 25 → 1 4 1 → 1 5 6 → 16

参考文献

[1] P. Protter, Stochastic Integration and Differential Equations, Applications of Mathematics, Second edition, Vol. 21 (Springer-Verlag, Berlin, 2005).

[2] 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

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

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

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

[6] 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.

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

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

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

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

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

[12] Etienne Pardoux1,Shanjian Tang,Forward–backward stochastic differential equations and quasilinear parabolic PDEs. Probab. Theory Related Fields 114 123–150,1999

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

[14] Y. Brenier, Polar factorization and monotone rearrangement of vector-valued functions, Comm. Pure Appl. Math. 44 375–417,1991

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

[16] Gobet, E., Munos, R.: Sensitivity analysis using Ito–Malliavin calculus and martingales. Application to stochastic control problem. SIAM J. Control Optim. 43, 1676–1713,2005