日本ソフトウェア科学会第 41 回大会 (2024 年度) 講演論文集

機械学習におけるアーキテクチャ構成とその学習手法の圏論的構造化

中村 卓武，浅田 和之，菊池 健太郎，中野 圭介

本論文では勾配に基づく学習の計算の持つ構造に着目し，その学習アルゴリズムを構成する操作を圏論を用いて定義する．そしてアルゴリズムの性質や訓練手法の数学的な分析と一般化を試みる．このような学習アルゴリズムの構成は既存研究にて既に提案されていて，振舞いに基づく抽象化を重視する Fong らの構成と，一般性とモジュール性を重視する Cruttwell らの構成があり，本研究ではこれらを融合することによってより洗練された構成を提案する．そしてこの新しい構成法では機械学習におけるアーキテクチャの結合と，アーキテクチャを用いた学習アルゴリズムの構成を同時に（順不同で）行うことができ，これらの構成の操作を同じ圏で表現できる．さらに string diagram と呼ばれる形式的なグラフィカル言語を用いることで，これらの構成を一つの図の上で表現することが可能になる．これを応用すると，教師強制と呼ばれる近似を用いた訓練法も図で表現できる．そして学習アルゴリズムに限らず，あらゆる同形の計算に対して適用できるように，教師強制を一般化できる．

1 はじめに

機械学習では日々新たな学習手法が提案されているが，それらの有用性が理論的に示されることは少ない．これにより，既存の学習手法を改良し，より優れた手法を発見する試みは経験的にならざるを得ない．そこで本研究では，機械学習における複雑な計算の構造を数学的に分析することで，既存手法の有用性を明らかにすることを目的としている．

本論文ではその前段階として，計算の構造を分析するための学習アルゴリズムの数学的な表現を整備する．具体的には既存研究にて提案されている，圏論を用いた勾配に基づく学習アルゴリズムの構成法をより洗練させる形で行う．そしてその簡単な応用例として，特殊な訓練手法の持つ計算の構造を明示し，一般化する．

計算構造の分析には既存研究と同様に圏論と呼ばれる数学の理論を用いる．圏論は数学は勿論，物理学，計算機科学，経済学，確率論などにも応用されてる抽象理論である．特にソフトウェア科学に関連する強みとして，圏論は計算をモジュール化しモジュール同士の結合を代数構造に基づいて扱うことに長けている．

特に 2 節で説明する string diagram [11] は圏論で用いられる形式的なグラフィカル言語であり，これを用いるとモジュールの構成やそのモジュールの結合を図で表現できるようになる．特に機械学習ではアーキテクチャによる予測の計算だけでなく，その訓練の計算も同時に図示できるようになる．

しかし一般には，予測の計算に対しその訓練の計算は逆方向に行われる．この特殊な計算構造を扱うために，3 節で説明するレンズ [6] [10] と呼ばれる双方向の計算の一般化を圏論を介して導入する．

レンズを使用するという試みは 4 節で説明する既存研究である Fong ら [4] と Cruttwell ら [3] によって既に行われている．これらの論文ではどちらもレンズを用いた圏論的な操作によって，勾配に基づく学習による学習アルゴリズムを構成している．2 つの構成法の違いとして，Fong らは圏論を用いた振舞いの定

Categorical structuring of architecture constructions and learning methods in machine learning.

Takumu Nakamura, 東北大学情報科学研究科, Graduate School of Information Sciences, Tohoku University.
Kazuyuki Asada, Kentaro Kikuchi, Keisuke Nakano, 東北大学電気通信研究所, Research Institute of Electrical Communication, Tohoku University.

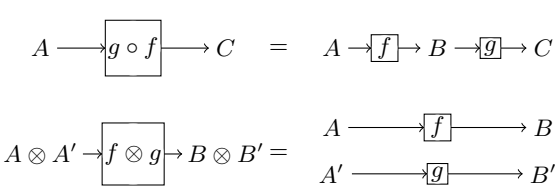
式化を重視していて，学習アルゴリズムに対する抽象的な枠組みをである Learner を提案している．一方 Cruttwell らは勾配に基づく学習における計算への忠実さを重視しつつ，その構成のモジュール化と微分計算の一般化を行っている．

本論文の 5 節で提案する新規構成法は，Fong らの提案した Learner に基づく学習アルゴリズムの構成に対し，その構成をモジュール化し，微分の計算を一般化する．この構成によって，アーキテクチャを組み合わせる操作と，アーキテクチャから学習アルゴリズムを構成する操作が同時に（順不同で）かつ一つの string diagram 上で行えるようになる．

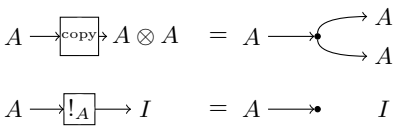
この新規構成法を応用して，6 節では学習アルゴリズムに対する操作や性質を圏論的構造に基づいて分析する．そして教師強制と呼ばれる，値の近似を用いる特殊な訓練手法を分析し，数学的な構造の変化の明示とその手法の一般化を行う．

2 String diagram

本論文では圏論における様々な射を扱うが，それらを string diagram [11] と呼ばれる図を用いて表現する．この図では，対象（集合）をワイヤで表現し，射（関数）をワイヤの上の箱や点として描画する．例として，射 $f : A \to B$，$g : B \to C$ を合成した射 $g \circ f : A \to C$ と，射 $f : A \to B$，$g : A' \to B'$ の monoidal product（関数の並列の合成） $f \otimes g : A \otimes C \to B \otimes D$ を次の図のように表現する．

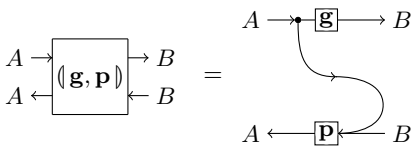


説明の都合により free cornering [8] [9] と呼ばれる圏の構成法を使用する．これによって，順方向（左から右）だけでなく，逆方向（右から左）の射を string diagram で図示することができるようになる．さらに，この string diagram では順方向から逆方向  ，逆方向から順方向  へ，折り曲がるようにワイヤを伸ばすことができる．ただし，この折り曲がるワイヤは必ず上から下への方向でなければならない．また自明なコモノイド構造として，入力を複製する関数である $\mathrm{copy}_A : A \to A \times A$ と入力を破棄する関数である $!_A : A \to 1$ を次のように表記する．

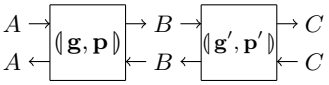


3 レンズ

レンズ [6] [10] は双方向変換と呼ばれる分野で研究されている概念である．レンズ $(\!|\mathbf{g}, \mathbf{p}|\!) : A \rightleftarrows B$ は get 関数 $\mathbf{g} : A \to B$ と put 関数 $\mathbf{p} : A \times B \to A$ の組であり，free cornering を用いて以下の図のように描画する．

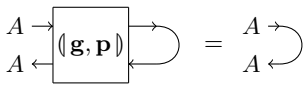


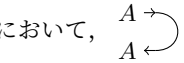
2 つの関数をこのように配置するメリットとして，レンズに対する基本的な操作であるレンズの合成を，次の図のように箱の結合で表現できるという点にある．



一般的にはレンズ $(\!|\mathbf{g}, \mathbf{p}|\!) : A \rightleftarrows B$ は，右上から出力された B の値を変更して右下へ入力するとき，右上に入力された A の値も合わせて変更し，右下から出力する，というような操作の一般化である．すると，このようなレンズは B の値が変更されないとき，A の値も変更されないという性質を満たすべきである．この性質は GetPut 則 [6] [10] と呼ばれ，free cornering を用いると，以下の string diagram 上の等

式で定義できる [1].



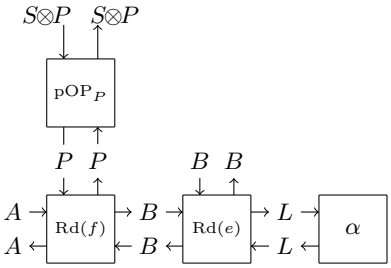
ただしこの図式において，  は上の順方向からの入力をそのまま下の逆方向に出力する操作であるから，下の逆方向の出力が上の順方向の入力と一致するという状況を表現することができる．

4 既存の学習アルゴリズムの構成

Fong ら [4] と Cruttwell ら [3] が提案した学習アルゴリズムの構成法を紹介する．ただし，説明の都合上，Fong らより後に投稿された Cruttwell らの論文の結果を先に説明する．

4.1 Cruttwell らの学習アルゴリズム

Cruttwell らは勾配に基づく学習における計算を圏論を用いて忠実に形式化した [3]．これは主に Cartesian reversed differential category [2] と呼ばれる，リバース・モードの自動微分を圏論的に一般化して得られた圏に基づいている．この形式化では学習アルゴリズムは以下の string diagram で表されるように，4 つのレンズを結合したレンズとして定義される．(ただし表記の都合上，一部のワイヤは上下の方向に描画している.)



例 4.1. 一般的な勾配に基づく学習の場合，各レンズは与えられた実数ベクトルに対して次のような順序で計算を行う．まず順方向（図式全体がなすレンズの get 関数）では，

(1) pOP_P は特に計算を行わず，入力のパラメータ p を出力する．

(2) $\mathrm{Rd}(f)$ ではパラメータ p と入力 a を受け取り，アーキテクチャと呼ばれる関数 f を用いて予測 $\hat{b} = f(p, a)$ を計算する．

(3) $\mathrm{Rd}(e)$ では予測 $\hat{b}$ と予測に対する正解 b を受け取って，その間の誤差 $L = e(\hat{b}, b)$ を誤差関数 e によって計算する．

(4) α では何も出力しない．

次に逆方向（図式全体がなすレンズの put 関数）の計算で，

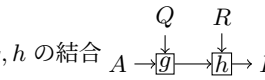
(5) α は与えられた誤差を捨て，代わりに学習率 ϵ と呼ばれるある定数を出力する．

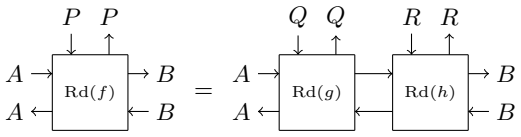
(6) $\mathrm{Rd}(e)$ では予測 $\hat{b}$ と正解 b，ϵ を受け取って，予測 $\hat{b}$ による誤差の勾配 $\epsilon\nabla_{\hat{b}}L$（ただし ϵ 倍）を出力する．この勾配は微分を用いて計算され，主に誤差を減らすための予測 $\hat{b}$ を移動させる量として使用される．

(7) $\mathrm{Rd}(f)$ では入力 a とパラメータ p を用いて，予測 $\hat{b}$ による誤差の勾配 $\epsilon\nabla_{\hat{b}}L$ に対し微分の連鎖律を適用し，パラメータ p による誤差の勾配 $\epsilon\nabla_p L$ へ変換する．

(8) 最後に pOP_P では勾配 $\epsilon\nabla_p L$ を用いてパラメータを修正し，出力する．勾配降下法と呼ばれる最適化手法を使用する場合，修正されたパラメータは $p - \epsilon\nabla_p L$ となる．またパラメータ以外にも何かしらの状態 S を入出力でき，訓練の度に更新しつつパラメータの修正に使用できる．■

Cruttwell らの構成の特徴の一つとして，パラメータの最適化に使用する学習率 ϵ の適用を，最適化を行うレンズ pOP_P で行わず，勾配の計算の段階で行っている．

もしアーキテクチャ f が，別のアーキテクチャ（レイヤー）g, h の結合  であるならば，$\mathrm{Rd}(f)$ も同様にレンズの結合で表現できる．



これにより，アーキテクチャ（レイヤー）の構成に合わせて自動微分が実行でき，誤差の勾配に対し，微分の連鎖律を段階的に適用できる．そして各パラメータによる誤差の勾配を求めるときに，段階的に計算された誤差の勾配を流用することで効率的に計算できる．そしてこの計算法は誤差逆伝播法と呼ばれる．

このように，Cruttwell らによる学習アルゴリズムの構成法ではアーキテクチャ（とその導関数）を string diagram で構成できるだけでなく，その後の学習アルゴリズムの構成も string diagram 上で行うことができる．

4.2 Fong らの学習アルゴリズム

Fong らの結果 [4] では，学習アルゴリズムの構成法や微分計算の一般化は重視されていない．しかし圏論における関手を用いて，誤差逆伝播法の拡張となる新しい訓練手法と，その手法を適用するための学習アルゴリズムの抽象的な枠組みである Learner を提案した．この新しい手法を誤差逆伝播法と区別するために訓練データ逆伝播法と呼ぶことにする．

これらの手法の違いとしては，誤差逆伝播法ではアーキテクチャ間で誤差の情報（誤差の勾配）を逆伝播するが，訓練データ逆伝播法では学習アルゴリズム間で誤差の情報（より望ましい入出力）を逆伝播する．

勾配に基づく学習を行う Learner は次のように定義される．

定義 4.2 (Fong らによる学習アルゴリズム [4])**.** 学習率を $\epsilon \in \mathbb{R}$ とし，ある関数 $d : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ が存在して，その導関数 $\frac{\partial d(x,y)}{\partial x}(z, -) : \mathbb{R} \to \mathbb{R}$ が逆関数を持つと仮定する．

任意の微分可能な実数値関数（アーキテクチャ）$f : P \times A \to B$ に対して，学習アルゴリズム $\mathrm{Lf}_{d,\epsilon}(f) : P \times A \rightleftarrows B$ を次のようにレンズとして定義する．

get 関数 $\mathbf{g}(p, a) = f(p, a)$

put 関数 $\mathbf{p}(p, a, b) = (p - \epsilon \nabla_p L, e_a(\nabla_a L))$

ただし，$(P, A, B) = (\mathbb{R}^l, \mathbb{R}^n, \mathbb{R}^m)$ とし，

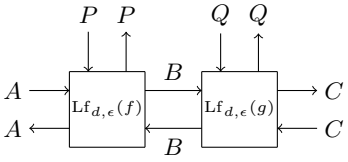
- $L = e(\hat{b}, b) = \sum_i d(\hat{b}_i, b_i)$ $(\hat{b}, b \in \mathbb{R}^m)$
- $e_a(x)_i = (\frac{\partial d(x,y)}{\partial x}(a_i, -))^{-1}(x_i)(a, x \in \mathbb{R}^n)$

とする．■

関数 e は誤差関数として扱われ，関数 d を $d(x, y) = (x - y)^2$ と定義すると逆関数に関する条件を満たし，e は誤差関数としてよく用いられる平均二乗誤差 $e(\hat{b}, b) = 1/2 \cdot \sum_i (\hat{b}_i - b_i)^2$[†1] となる．

Fong らの学習アルゴリズムは順方向の計算（get 関数）で予測を出力し，逆方向の計算（put 関数）では勾配降下法によってパラメータを修正し，同時に d の導関数から定義される逆関数を用いて入力も修正している．この入力の修正については，平均二乗誤差を用いるならば，この計算は勾配降下法による修正となる．

この学習アルゴリズムにおける訓練データ逆伝播法は，以下の図のようにアルゴリズム同士の結合によって計算される．このとき，順方向の計算ではあたえられた入力 a に $f(p, -)$ が適用されて $\hat{b}$ が得られ，さらに $g(q, -)$ が適用され予測 $\hat{c}$ が出力される．一方逆方向の計算は，直感的には，予測 $\hat{c}$ を正解 c に近づけるようなパラメータと入力（予測）b' を計算する．次に予測 $\hat{b}$ を修正された予測 b' に近づけるようにパラメータと入力 a を修正する．



より直感的な解釈として，$\mathrm{Lf}_{d,\epsilon}(g)$ は $\hat{b}$ を b に近づけることで，$\hat{c}$ が c に近づくと主張しているため，$\mathrm{Lf}_{d,\epsilon}(f)$ が $\hat{b}$ と b の間の誤差を減らすことで，間接的に $\hat{c}$ と c の間の誤差が減る，と考えることができる．

この訓練データ逆伝播法による勾配の計算が実は誤差逆伝播法の勾配の計算と一致することもあり，アーキテクチャ f から学習アルゴリズム $\mathrm{Lf}_{d,\epsilon}(f)$ を構成する操作は圏論における関手になる．それにより，学習アルゴリズム同士を結合することによって，アルゴリズムに含まれるアーキテクチャ同士の結合を表現す

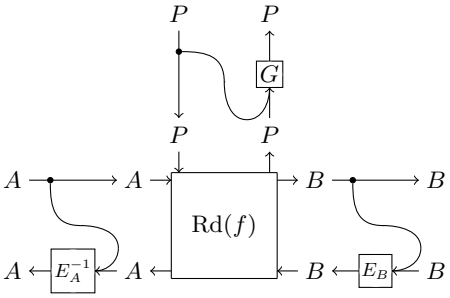
†1 計算の簡略化や圏論的な都合により，$1/m$ ではなく $1/2$ を用いる．このような定数倍は学習率で調整できるため実用上は問題にならない．

ることができる．このように Cruttwell らの構成法とは異なり，学習アルゴリズム化とアーキテクチャの構成を同時に行うことができる．

ただし Cruttwell ら [3] によれば，Fong らの論文 [4] の関手は well-defined にならないという問題があった．直感的には，等価な二つのパラメータを二つのアーキテクチャにそれぞれ部分適用した時のアーキテクチャ間の同一性を同値関係とし，アーキテクチャ全体に対し同値類を考えている．しかし，アーキテクチャから学習アルゴリズムを構成するときに，その同値関係が保たれないという問題がある．そこで本研究ではパラメータの等価性をより自明なものに限定し，この問題を回避している．例えば後の定義 5.2 で使用している smallest wide subcategory $\mathbf{Lens}(\mathbf{C})_{\mathrm{s}}$ は，主にパラメータを等価とみなすための相互変換を制限する目的で導入している．

5 学習アルゴリズムの新規構成法

定義 4.2 の Fong らの学習アルゴリズム $\mathrm{Lf}_{d,\epsilon}(f)$ をモジュール化・一般化するにあたって，その学習アルゴリズムが以下の string diagram にて構成できることは本研究において重要な観察である．



ただし，(1) $\mathrm{Rd}(f)$ は Cruttwell らの構成 (4. 1) と同様に予測の計算と微分の連鎖率を適用する操作であり，(2) $E_B : B \times B \to B$ は予測 $\hat{b}$ と対応する正解 b から誤差の勾配 $\nabla_{\hat{b}} L$ を求める関数で，(3) $G : P \times P \to P$ は現在のパラメータ p とそのパラメータによる勾配 $\nabla_p L$ を元に勾配降下法を適用する関数であり，(4) ${E_A}^{-1} : A \times A \to A$ は入力 a に対する関数 e_a と誤差の勾配 $\nabla_a L$ を用いて入力を $e_a(\nabla_a L)$ へ修正する関数である．

注意すべき点として，Cruttwell らの構成 (4. 1) では学習率を誤差の勾配を求めるときに適用していたが，この構成では勾配降下法を行うときに適用する．

そしてこの観察から $\mathrm{Rd}(f)$ の左右のレンズ $(\!|\mathrm{id}_B, E_B|\!)$ と $(\!|\mathrm{id}_A, E_A^{-1}|\!)$ に関して次の仮説を立てることができ，実際に成り立つ．

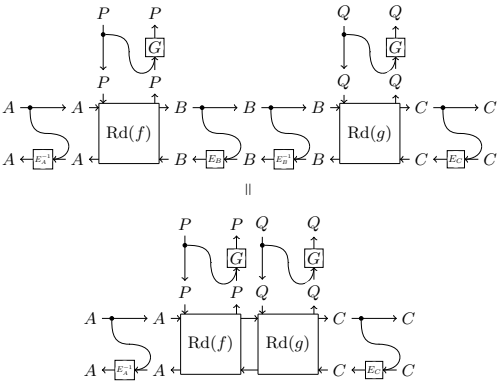
定理 5.1. レンズ $(\!|\mathrm{id}_B, E_B|\!)$ と $(\!|\mathrm{id}_B, E_B^{-1}|\!)$ は互いに逆レンズである．■

Proof.

$E_B^{-1}(b, E_B(b, b')) = b',\ E_B(b, E_B^{-1}(b, b')) = b'.$

を示せばよいが，これらの等式は E_B^{-1} の定義に用いた逆関数によって成り立つ． □

具体例として，誤差関数に平均二乗誤差を用いると $E_B^{-1}(b, b') = E_B(b, b') = b - b'$ となり，この等式を満たす．この命題によって Fong らによる学習アルゴリズムの訓練データ逆伝播法が誤差逆伝播法の計算を行うことを簡単に確認できる．なぜなら次の図のように，$\mathrm{Lf}_{d,\epsilon}(f)$ と $\mathrm{Lf}_{d,\epsilon}(g)$ による訓練データ逆伝播法を考えると，$\mathrm{Lf}_{d,\epsilon}(f)$ の $(\!|\mathrm{id}_B, E_B|\!)$ と $\mathrm{Lf}_{d,\epsilon}(g)$ の $(\!|\mathrm{id}_B, E_B^{-1}|\!)$ が打ち消し合い，$\mathrm{Rd}(f)$ と $\mathrm{Rd}(g)$ が隣接する．

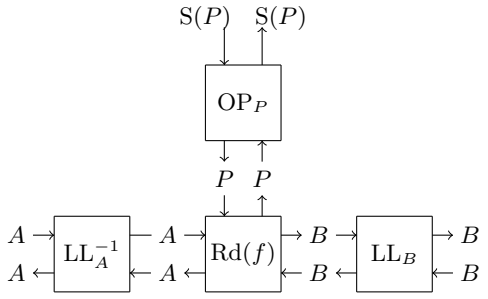


これにより，Cruttwell らの構成と同様に誤差逆伝播法の計算が行われる．

上記の観察と定理 5.1 から，Fong らの学習アルゴリズムの構成法を対称強モノイダル関手として新しい構成へ一般化する．ここで，リバース・モードの自動微分が行える射からなる圏 $\mathbf{C}$ に対し，$\mathbf{C}$ における

パラメータを持つ射（アーキテクチャとみなす）からなる圏を $\mathbf{Para}(\mathbf{C})$ とし，$\mathbf{C}$ の射から構成されるパラメータ付きレンズ（学習アルゴリズムとみなす）からなる圏を $\mathbf{Para}(\mathbf{Lens}(\mathbf{C}))$ と表す．

定義 5.2. 関手 $\mathrm{Gb}_{\mathrm{LL,S,OP}} : \mathbf{Para}(\mathbf{C}) \to \mathbf{Para}(\mathbf{Lens}(\mathbf{C}))$ を次のように定義する．圏 $\mathbf{Para}(\mathbf{C})$ の任意の射（アーキテクチャ）f に対し，パラメータ付きレンズ（学習アルゴリズム）$\mathrm{Gb}_{\mathrm{LL,S,OP}}(f)$ を次の図のように定義する．



ただし，$\mathrm{LL} = \{\mathrm{LL}_A\}_{A\in\mathbf{C}}$ は逆レンズを持つようなレンズの属とし，OP，S についてはそれぞれ $\mathrm{S} : \mathbf{Lens}(\mathbf{C})_{\mathrm{s}} \to \mathbf{Lens}(\mathbf{C})_{\mathrm{s}}$，$\mathrm{OP} : \iota \circ \mathrm{S} \Rightarrow \iota$ となるような関手と，レンズから構成される自然変換とする．また $\mathbf{Lens}(\mathbf{C})_{\mathrm{s}}$ はモノイダル圏 $\mathbf{Lens}(\mathbf{C})$ に対する smallest wide subcategory とし，$\iota : \mathbf{Lens}(\mathbf{C})_{\mathrm{s}} \to \mathbf{Lens}(\mathbf{C})$ はその包含関手とする．■

例 5.3. 定義 5.2 は定義 4.2 の Fong らによる学習アルゴリズム構成の一般化であり，それぞれ $\mathrm{LL}_A = (\!|\mathrm{id}_B, E_B|\!)$，$\mathrm{S}(P) = P$，$\mathrm{OP}_P = (\!|\mathrm{id}_P, G|\!)$ とすると Fong らによる学習アルゴリズムが得られる．また $\mathrm{S}(P) = P \times P$ とし，左の P を状態，右の P をパラメータとみなすことで，Cruttwell ら [3] が使用していたような，状態を用いる最適化手法を使用することができるようになる．■

学習アルゴリズムの構成法が関手であることで，アーキテクチャの構成と学習アルゴリズムの構成が同時に行えるというメリットがあった．この新規構成法によって，勾配に基づく学習のアルゴリズムの構成法が関手になるための条件をより明確にすることができた．

また Fong らの構成法は well-defined にならないという問題があったが，この問題を回避するためにパラメータ間の等価性をより自明なものに制限した．関手 Gb に対しても同様に制限された同値関係を保つように，smallest wide subcategory $\mathbf{Lens}(\mathbf{C})_{\mathrm{s}}$ を使用して OP_P における最適化の計算を制限している．

6 新規構成法の応用

定義 4.2 の Fong らの学習アルゴリズムを一般化・モジュール化し，これを string diagram によって構成できるようになった．この新規構成法を応用して，学習アルゴリズムの性質や，近似を用いた特殊な訓練法を string diagram を用いて議論する．

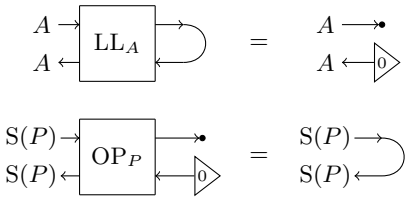
6.1 学習アルゴリズムにおける GetPut 則の圏論的性質

Fong らの学習アルゴリズムや，新規構成法によるアルゴリズムの振る舞いは 3 節にて説明した一般的なレンズとしての振る舞いに類似している．実際にこれらのアルゴリズムは予測結果が正しいものに修正されたとき，入力とパラメータをより正しいものに修正している．そこでこれらのアルゴリズムにおける GetPut 則を考えると，誤差が 0（予測と正解が一致）ならば入力，パラメータを修正しないという性質となる．これはパラメータ更新の収束性の観点から学習アルゴリズムにおいて望ましい性質である．

実際に定義 4.2 の Fong らの学習アルゴリズムでは誤差関数に平均二乗誤差を使用したとき GetPut 則を満たす [5] ことが知られているが，詳しい条件は述べられていない．そこで新規構成法（定義 5.2）を使用し，$\mathrm{Gb}_{\mathrm{LL,S,OP}}$ による一般化された学習アルゴリズムが GetPut 則を満たすための必要十分条件を導出する．

定理 6.1. 関手 $\mathrm{Gb}_{\mathrm{LL,S,OP}} : \mathbf{Para}(\mathbf{C}) \to \mathbf{Para}(\mathbf{Lens}(\mathbf{C}))$ に対し，以下の命題は同値である．

- 任意のアーキテクチャ f に対し，学習アルゴリズム $\mathrm{Gb}_{\mathrm{LL,S,OP}}(f)$ が GetPut 則を満たす．
- 任意の対象 $A, P \in \mathbf{C}$ に対し，次の図式で表される 2 つの等式が成り立つ．



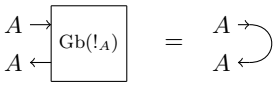
■

Proof. (⇑) はレンズにおける同型性などのレンズで構成される等式によって証明できる．
(⇓) はアーキテクチャ $P \times A \longrightarrow \bullet$ $\overset{0}{\triangleleft} \longrightarrow B$ を Gb に適用することでそれぞれ証明できる． □

ここで，$A \overset{0}{\triangleleft}$ は定数 0（零ベクトル）を出力する操作とみなすことができ，レンズ $\begin{matrix} A \longrightarrow \bullet \\ A \overset{0}{\triangleleft} \end{matrix}$ は $\mathrm{Rd}(!_A)$ と等しい．

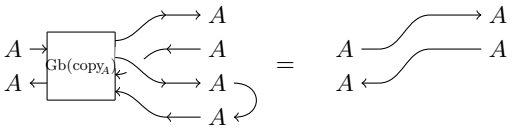
例 6.2. LL_A が平均二乗誤差から定義されるレンズ $(\!|\mathrm{id}_A, E_A|\!)$ である場合，この条件は $E_A(a,a) = 0$ となり，この等式は予測と正解が一致している場合，誤差の勾配が 0 になることを表している．一方 OP_P が勾配降下法から定義されるレンズ $(\!|\mathrm{id}_P, G|\!)$ である場合，この条件は $G(p,0) = p$ となり，この等式はパラメータによる誤差の勾配が 0 である場合，パラメータが変化させないことを表している． ■

定理 5.1 によって，Gb による学習アルゴリズムが GetPut 則を満たすとき，次の等式が成り立つ．



すなわち，値を捨てる操作 $!_A$ から構成した学習アルゴリズム $\mathrm{Gb}(!_A)$ は入力を一切修正しない学習アルゴリズムとなる．これにより，アーキテクチャが出力した値を捨てる操作は，学習アルゴリズムの出力した予測を正解として入力する操作に対応することがわかる．

また Gb は強モノイダル関手であるから，アーキテクチャの持つコモノイド構造を保ち，$\begin{matrix} A \\ A \end{matrix}$ も $!_A$ と同様にコモノイド構造におけるコユニットとなる．すなわち以下の等式が成り立つ．

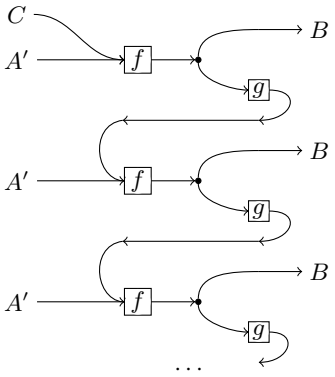


（説明の都合上，図の左辺の右側の上から 2 番目，3 番目のワイヤを交差させているが，この操作は free cornering において非可逆であり，交差させなくとも等式は成り立つ．）このように本構成法を用いることで，アーキテクチャ（予測の計算）に対する操作を，学習アルゴリズム（予測と訓練の計算）に対する操作に対応付けて考えることができる．

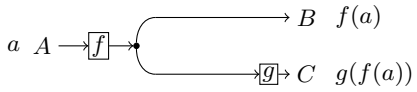
6.2 教師強制の一般化

本構成法ではアーキテクチャに対する操作を学習アルゴリズムに対する操作に対応付けることができるが，訓練の計算のみを操作することもできる．これを用いて，教師強制（teacher forcing） [7] と呼ばれる，アーキテクチャの構造を応用した訓練手法を string diagram で表現する．ただし表記の簡略化のため，以降は string diagram におけるパラメータのワイヤを省略する．

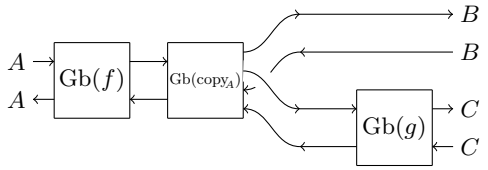
教師強制は次の図で表されるような，あるアーキテクチャ $f : P \times A \to B, g : Q \times B \to C$ から構成される回帰型ニューラルネットワーク（RNN） [7] に対して定義される．この RNN の重要な性質として，右下の B の出力は必ず右上の B よりも後に出力される．



また教師強制を考える上で，次の図のように RNN アーキテクチャの一部のみを考える．(ただし $A = C \times A'$)

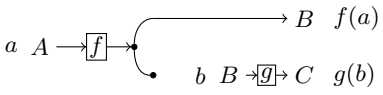


パラメータのワイヤと同様に，アーキテクチャへのパラメータの適用も同様に省略する．例としてこのアーキテクチャが入力 a を元に出力する予測を $f(a)$, $g(f(a))$ と表記する．このとき，このアーキテクチャに関手 Gb を適用して得られる学習アルゴリズムは，次のような string diagram で表すことができる．



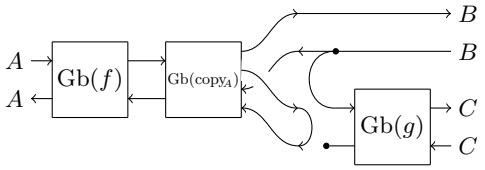
またこの図におけるワイヤの交差は，図の右の C の順方向のワイヤが B の逆方向のワイヤに合流しないことを表し，RNN アーキテクチャの性質がこの操作を可能にしている．

次に，予測 $f(a)$, $g(f(a))$ に対する正解をそれぞれ b, c とすると，教師強制を適用したアーキテクチャでは次の図のように予測の計算を行う．



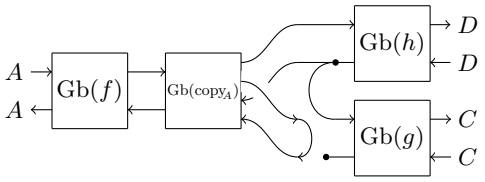
このように，教師強制では予測 $f(a)$ と正解 b が等しい，つまり f における訓練が完了しているという近似を行うことで，アーキテクチャ g では入力 $f(a)$ の変わりにより正確な正解 b を用いて予測 $g(b)$ を出力する．

関手 Gb による学習アルゴリズムが GetPut 則を満たすと仮定すると，  は $\mathrm{Gb}(!_B)$ と等しくなり，コモノイド構造におけるコユニットになる．よって教師強制を適用した後の訓練の計算は string diagram によって次のように表すことができる．



重要な点として，string diagram で表現した教師強制は機械学習の計算から一般化されている．つまり $\mathrm{Gb}(\mathrm{copy}_A)$ の順方向の計算が値を複製しているという仮定は必要であるが，レンズの形式で表せるあらゆる計算に対して教師強制のような近似手法を適用できるということである．例えば Smithe [12] によれば，ベイズの定理に基づく事前分布の更新はレンズで表すことができ，このレンズの構成法は関手的である．これにより，ベイズ更新に教師強制を適用できる可能性がある．

他にも上記の図式の右上に更に学習アルゴリズムを結合すると，次のような図式が得られる．



教師強制であれば，入力 a に対して予測 $f(a)$ が計算され，これに対応する正解 $b \in B$ をアーキテクチャ g の予測に利用していた．しかし，この図では $f(a) \in B$ に対する正解は与えられず，代わりに $h(f(a)) \in D$ に対する正解として $d \in D$ が与えられるため，教師強制によって得られるアルゴリズムではない．しかしこのアルゴリズムでは予測 $f(a)$ に対する正解 b の代わりに，$\mathrm{Gb}(h)$ によって予測 $f(a)$ から修正された予測 b' を用いている．このように訓練手法の圏論的な一般化によって，教師強制のような限られたアーキテクチャで使用されている近似手法を様々な場面に導入できる可能性がある．

7 おわりに

Cruttwell らの結果 [3] を基に，Fong らの学習アルゴリズム [4] をモジュール化・一般化した．これにより学習アルゴリズムの計算の持つ数学的な構造と，その構成要素に求められる条件が明確になった．これによりアーキテクチャの結合と学習アルゴリズムの構成を同時にかつ一つの string diagram 上で行えるようになった．そして新規構成法による計算構造を応用して，学習アルゴリズムに対する操作や性質を数学的に分析することに成功した．これにより近似を用いる特殊な訓練手法を一般化し，その適用範囲を拡張した．

しかし学習手法の有用性，例えば教師強制を適用した学習アルゴリズムの性質などはまだ明らかにできておらず，今後の課題となっている．また新規構成法に関する課題として，Cruttwell らの結果と同様に微分計算の一般化を行っているが，例として紹介した実数関数を用いた勾配に基づく学習以外の応用はできていない．また誤差関数をレンズに一般化したが，これも明確な応用は見つかっていない．Fong らの構成法で定義できなかった誤差関数にソフトマックス交差エントロピー誤差があるが，定義できない原因としては $d : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ のような各要素ごとの誤差の和で表現できないことにある．この制約もレンズに一般化したことによって解消されたが，残念ながら，この誤差関数を用いたレンズは逆レンズを持たず，関手を構成できない．Fong らの提案した学習アルゴリズムの抽象的な枠組みは勾配に基づく学習とそこまで相性が良いわけではないかもしれない．しかし勾配を用いないような学習アルゴリズムに対しても Learner の枠組みが適用でき，同様に分析できる可能性がある．これによりさらなる発展の余地が考えられる．

参 考 文 献

[1] Boisseau, G., Nester, C., and Román, M.: Cornering Optics, *Electronic Proceedings in Theoretical Computer Science*, Vol. 380(2023), pp. 97–110.

[2] Cockett, R., Cruttwell, G., Gallagher, J., Lemay, J.-S. P., MacAdam, B., Plotkin, G., and Pronk, D.: Reverse Derivative Categories, *28th EACSL Annual Conference on Computer Science Logic (CSL 2020)*, Fernández, M. and Muscholl, A.(eds.), Leibniz International Proceedings in Informatics (LIPIcs), Vol. 152, Dagstuhl, Germany, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020, pp. 18:1–18:16.

[3] Cruttwell, G. S. H., Gavranović, B., Ghani, N., Wilson, P., and Zanasi, F.: Categorical Foundations of Gradient-Based Learning, *Programming Languages and Systems*, Sergey, I.(ed.), Cham, Springer International Publishing, 2022, pp. 1–28.

[4] Fong, B., Spivak, D., and Tuyeras, R.: Backprop as Functor: A compositional perspective on supervised learning, *2019 34th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, Los Alamitos, CA, USA, IEEE Computer Society, jun 2019, pp. 1–13.

[5] Fong, B. and Johnson, M.: Lenses and Learners, *Proceedings of the Eighth International Workshop on Bidirectional Transformations*, Cheney, J. and Ko, H.-S.(eds.), 2019.

[6] Foster, J. N., Greenwald, M. B., Moore, J. T., Pierce, B. C., and Schmitt, A.: Combinators for bidirectional tree transformations: A linguistic approach to the view-update problem, *ACM Trans. Program. Lang. Syst.*, Vol. 29, No. 3(2007), pp. 17–es.

[7] Goodfellow, I., Bengio, Y., and Courville, A.: *Deep Learning*, MIT Press, 2016. `http://www.deeplearningbook.org`.

[8] Nester, C.: The Structure of Concurrent Process Histories, *Coordination Models and Languages*, Damiani, F. and Dardha, O.(eds.), Cham, Springer International Publishing, 2021, pp. 209–224.

[9] Nester, C.: Concurrent Process Histories and Resource Transducers, *Logical Methods in Computer Science*, Vol. Volume 19, Issue 1(2023).

[10] Riley, M.: Categories of Optics, 2018.

[11] Selinger, P.: *A Survey of Graphical Languages for Monoidal Categories*, Springer Berlin Heidelberg, 2011, pp. 289–355.

[12] Smithe, T. S. C.: Bayesian Updates Compose Optically, 2020.