論文内容の要旨

論文題目

Geometric Theory of Weight Dynamics for Training Hierarchical Neural Networks

(階層型神経回路モデルにおける学習力学の幾何学的理論）

氏名　　唐木田　亮

ニューラルネットワークは生物の神経回路を模した機械学習のモデルであり，モデルのパラメータを入力データに応じて決める学習を行う．特に，構成素子を層状に並べた階層構造を持つモデルを使った学習は深層学習と呼ばれる．近年，深層学習ではこれまでにない学習手法の提案と開発が進み，画像認識や音声認識といった分野では既存の機械学習を大きく上回る性能が実現されている[1]．しかしながら，深層学習では性能重視の手法開発が進んでおり，理論的には妥当性や最適性が保証されていないヒューリスティクスも多く存在するのが現状である．深層学習で使われているヒューリスティクスの性能や挙動のメカニズムを数理的に解き明かすことは，深層学習の技術をより発展させるために欠かせない知見である．

本論文では，経験的に性能が良いことが報告されている学習アルゴリズムに対して理論解析を試み，性能保証を与える．具体的には，Restricted Boltzmann Machine モデルの最尤学習とその近似である Contrastive Divergence 学習[2,3]，および勾配法の一種として知られる Weight Normalization[4]に注目し解析を行った．さらに各々の学習アルゴリズムの解析で明らかになったパラメータ空間の幾何学構造の知見に基づき，効率的なアルゴリズムの開発を行い，その有効性を理論計算および数値実験で検証した．

まず，第 2 章では, Restricted Boltzmann Machine (RBM) 学習において，最尤学習のヒューリスティクスとして知られる Contrastive Divergence (CD) 学習[2]が厳密手法と解が一致す

る保証を与える．RBM を含む生成モデルの典型的な学習方法は尤度を評価関数とした最尤学習であるが，通常の RBM では尤度が解析的に計算できない．そのため最尤学習では MCMC に代表されるサンプリング手法が必要となるが，膨大な計算時間がかかるため実用は困難である．そこで，最尤学習の粗い近似として，サンプリングを有限回で止める CD 学習が提案された．CD 学習は短時間で最尤学習に近い解へ収束すると経験的には指摘されているが，一般的な理論保証は知られていない．CD 解と最尤解の差あるいは共通性を解き明かすことは，CD 学習を実用する上で重要な知見である．

そこで本章では，CD 学習と最尤学習に共通な解とその特徴を平衡点解析[5]によって明らかにした．より具体的には，入力素子が連続値をとる 2 種類の RBM モデルにおいて，CD 学習のアルゴリズムを解析的に書き換えた学習方程式の安定平衡点を求め，最尤学習方程式の安定平衡点と一致することを明らかにした．まず，隠れ素子が連続値の Gaussian-Gaussian RBM モデルでは両学習方程式の安定平衡点はすべて一致し，解は主成分分析に基づいた入力データの次元削減に対応することを示す．次に，隠れ素子が離散値の Gaussian-Bernoulli RBM モデルでは，一定の条件下で両学習方程式の安定平衡点のひとつが求まり，解が入力の独立成分分析に対応することを示す．我々の解析結果は，最尤学習の近似である CD 学習で最尤学習と同じ特徴成分を抽出できることを示している．以上の結果は英文論文誌[7]にて発表された．

次に第 3 章では，Gaussian-Bernoulli (G-B) RBM の学習アルゴリズムの開発を行う．第 2 章の解析では G-B RBM において最尤学習と CD 学習の解の一部が明らかとなったが，この解析では G-B RBM の結合行列を直交に制約すると尤度とその勾配が解析的に計算できることを重要な性質として用いた．この性質を用いて，我々は結合行列に直交制約を課すことで，解析的に尤度と勾配が計算可能でサンプリング不要な学習アルゴリズムを提案した．具体的には，直交制約下でのパラメータ更新を実現するために測地線勾配を使ったアルゴリズムを提案し，自然画像を使った数値実験で制約無しの既存の学習アルゴリズムと同等な性能を高速に達成できることを示す．本手法は直交空間上の最尤推定値へ最急勾配に従って収束することが保証されている点，学習中の尤度が容易に計算できるため収束や最大化を確認しやすい点で，CD 学習のようなヒューリスティクスよりも効率のよい最尤学習であるといえる．また，直交性の仮定が隠れ変数の独立性と等価であることを示し，制約に直観的な解釈を与える．以上の結果は国際会議[8]にて発表された．

深層学習の発展に欠かせない要素として勾配アルゴリズムの開発が挙げられる． AdaGrad や Adam に代表される確率的最急降下法は，学習の収束を高速化する方法として開発が進んだ．特に，近年になって開発された手法が Weight Normalization (WN) 法である[4]．この手法は，深層モデルの結合ベクトルを動径と方向の座標成分に分けて最急降下を行う勾配法で，学習の高速化が畳み込みネットワーク，変分 Auto-Encoder，リカレントネットワークといった

幅広い深層学習のモデルで経験的に報告されている．しかしながら，高速化のメカニズムは理論的に未解明であった．

そこで第 4 章では, WN 法の理論解析を行い，高速化のメカニズムの解明を試みた．具体的には，パラメータ変換のヤコビ行列[7]に注目することで, WN 法の勾配が通常座標の空間で持つ実効的な学習係数を導出する．この実効的な学習係数は学習ステップに対して単調減少し，学習係数の補正を行うことを明らかにする．この補正のため，初期値として大きな学習係数を与えることが可能となり，学習の高速化が実現できる．また，多層 ReLU モデルでは, WN 法の勾配がパラメータのスケール変換に対する不変性を持つことを示す．通常の多層 ReLU モデルでは勾配のスケールに対する不変性が成り立たないため，パラメータのスケールが大きくなると学習が停滞する問題があった. WN 法はこの学習の停滞も防ぎながら，学習の高速化に寄与していると考えられる．

第 5 章では, WN 法をさらに高速化できるアルゴリズムを自然勾配法によって実現する．自然勾配法は，パラメータ空間のリーマン幾何構造を表す計量行列，すなわちフィッシャー情報行列を使って最急勾配を補正する方法である[10]．本章では, WN 法の動径パラメータに自然勾配法を適用した．通常の自然勾配法は全パラメータに関する計量の逆行列計算を必要とするため計算量が大きくなり，大規模なニューラルネットワークでの実用が難しい問題がある．この問題に対して，動径パラメータのみに関する計量を考えた本手法は自然勾配にかかる計算量を減らせる点で利点がある．また，第 4 章で述べた WN 法の実効的な学習係数による補正は，方向成分のパラメータに強く働くことが簡単な計算と数値実験から示唆される．よって，動径成分にも補正を加える本手法の方が学習を高速化する可能性がある． 我々は動径パラメータの計量行列が厳密な自然勾配のフィッシャー情報行列の幾何構造の一部を反映していることを明らかにする．具体的には多層パーセプトロンにおいて動径計量とフィッシャー情報行列の類似を示す．最後に，深層学習の性能評価で用いられる画像データセット MNIST と CIFAR10 を多層パーセプトロンに学習させた場合に，本提案手法が WN 法の収束を高速化することを実証する．

本研究で RBM 学習の解析に用いた平衡点解析，および WN 法の解析に用いた力学系のパラメータ変換は，本論文が解析した学習アルゴリズムに限らず広く適用できる枠組みである．また，アルゴリズムの拡張で提案した測地線勾配や自然勾配のような幾何学的な最適化の手法も，様々な既存の学習アルゴリズムの拡張に利用できるだろう．今後，階層型ニューラルネットワークの学習アルゴリズムを系統的に解析し，背後に潜む幾何学的な構造を活かした効率のよいアルゴリズムを考える上で，本論文が基礎となることを期待する．

参考文献

[1] LeCun, Bengio, and Hinton, Nature, 521(7553):436–444. (2015)

[2] Hinton, Neural Computation, 14(8):1771–1800. (2002)

[3] Hinton and Salakhutdinov, Science, 313(5786):504–507. (2006)

[4] Salimans and Kingma, NIPS2016. (2016)

[5] Baldi and Hornik, Neural networks, 2(1):53–58. (1989)

[6] Dauphin, Pascanu, Gulcehre, Cho, Ganguli, and Bengio, NIPS2014. (2014)

[7] Karakida, Okada, and Amari, Neural Networks, 79:78-87. (2016)

[8] Karakida, Okada, and Amari, ESANN2016. (2016)

[9] Cousseau, Ozeki, and Amari, IEEE Transactions on Neural Networks, 19(8):1313–1328. (2008)

[10] Amari, Neural computation, 10(2):251–276. (1998)