Computational and statistical complexity analysis of learning problems based on first-order gradient information

（1 次勾配情報に基づく学習問題の統計的・計算量的解析）

数理情報学専攻 48226208　大古 一聡
指導教員　鈴木 大慈 准教授

1　はじめに

大規模な学習問題には 1 次勾配情報に基づく手法が用いられる．そうした手法は実データにおいて特徴学習に成功しているが，その理論は発展途上である．

本論文では，近年台頭したいくつかの設定を例とし，1 次勾配情報に基づく手法の効率性を統計的・計算量的観点から調べた．同時に，1 次勾配情報に基づく学習の本質的困難性を示し，最悪ケースを与える例が普遍的にデータの高次元性に起因することを見た．更に，現実のデータの持つ構造を定式化して追加で仮定し，統計的・計算量的上下界のデータの構造への依存性を示した．

Part I: 非凸有限和設定の計算量的複雑度

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

の形の問題を有限和設定という．f_i が非凸・L-平滑である問題で，多くの従来手法 (例: [6]) が確率的勾配だけでなく正確な $\nabla f(x)$ の値を何度も計算する必要があった．そこで，確率的勾配のみを用い，最適計算量を達成しつつ従来手法と同様の汎用性を持つ手法を提案した．

次に f_i の間の類似性を特徴づけるパラメータ ζ を導入し解析を行った．最悪計算量のケースは ∇f_i が別々の方向を向く例から与えられるが [7]，ζ が異なる次元として取れる数を制限する低次元性を誘導し，ζ 依存の下界を得る．提案手法はこの ζ を加味した最適計算量も達成し，この性質は分散学習への応用で有効である．

定理 1. 提案手法は ε-*1* 次最適点を $O\left(\frac{L+\zeta\sqrt{n}}{\varepsilon^2}\right)$ 回の勾配計算で求める．また，*2* 次最適点を求め，*PL* 条件下で線形収束する．一方，最悪計算量は $\Omega\left(\frac{L+\zeta\sqrt{n}}{\varepsilon^2}\right)$.

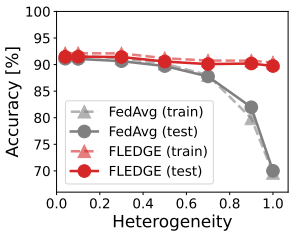

図 1. 連合学習でクライアント間の不均一性を変えて提案手法 (赤) と既存手法 (灰) の精度をプロットしたもの．これは有限和設定で ζ を変えることに対応する．提案手法は不均一性が増加しても精度を保つ．

Part II: 拡散モデルの統計的複雑度

2　スコアベース拡散モデル

スコアベース拡散モデルは，Stable Diffusion 等で有名な代表的なデータ生成手法である．

生成したいデータが d 次元であるとし，データの従う分布の密度関数を p_0 とする．初期分布を p_0 とする Ornstein–Uhlenbeck 過程 ($\theta = \sigma = 1$) を考え，その時刻 t での分布を p_t とする．十分大きな $\overline{T}$ を取り，d 次元 Brown 運動を $(B_t)_{t>0}$ とし，確率過程 $\{Y_t\}_{t\in[0,\overline{T}]}$ を

$$\mathrm{d}Y_t = (Y_t + 2\nabla \log p_{\overline{T}-t}(Y_t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,\ Y_0 \sim \mathcal{N}(0, I_d)$$

で定めると，$Y_{\overline{T}}$ は p_0 に従う．これを時間離散化したものがスコアベース拡散モデルの仕組みである [4]．なおスコアとは p_t の対数 1 次勾配 $\nabla \log p_t(x)$ のこと．

実際には真の分布 p_0 は未知であり，スコア $\nabla \log p_t(x)$ は有限の訓練データで学習したニューラルネットワーク $\hat{s}(x,t)$ によって近似される．多くの理論研究は $\nabla \log p_t(x)$ と $\hat{s}(x,t)$ の誤差に関する仮定の下，その誤差が生成データの分布と真の分布 p_0 の距離にどのように伝播するかを調べてきた [2]．しかし，そもそも誤差の仮定を満たすのに必要なサンプルサイズやネットワークサイズは不明のままだった．

3　分布推定問題としての定式化

本論文ではこれに対処し，拡散モデルが有限の訓練データからどれほど効率的に特徴を学習できるか調べるため，拡散モデルを分布推定問題として初めて定式化し統計的複雑度を解析した．具体的には，p_0 が典型的な関数空間に属するという仮定の下，p_0 からサンプルした n データから経験損失を最小化するニューラルネットワークを選び，これを用いて生成したデータの分布 q と p_0 の距離を調べた．

分布のクラス.　以下の Besov ノルムの仮定は p_0 の平滑性に関するもので，s 回まで微分係数が有界なら十分．

仮定 2. $0 < p, q \le \infty, 0 < s, 1/p - 1/2 < s$ とする．p_0 は $\Omega = [-1,1]^d$ に台を持ち，Ω 上の関数として Besov ノルム $\|p_0\|_{B^s_{p,q}([-1,1]^d)}$ 及び $p_0(x), (p_0(x))^{-1}$ が有界である．更に台の端の幅 $n^{-\frac{1}{2s+d}}$ では十分滑らかとする．

ニューラルネットワークのクラス. 非 0 パラメータの数 S を制限したスパースなネットワークを考える.

定義 3. $\Phi(L,W,S,B)$ を次のような $d+1$ 次元入力 d 次元出力の ReLU ネットワークの集合とする. $\Phi(L,W,S,B) := \{(A^{(L)}\mathrm{ReLU}(\cdot)+b^{(L)})\circ\cdots\circ(A^{(1)}x+b^{(1)})|\ A^{(i)}\in\mathbb{R}^{W_i\times W_{l+1}}, b^{(l)}\in\mathbb{R}^{W_{l+1}}, \sum_{l=1}^{l}(\|A^{(l)}\|_0+\|b^l\|_0)\le S, \max_l\|A^{(l)}\|_\infty\vee\|b^{(l)}\|_\infty\le B\}$.

スコアマッチング. 以下の経験スコアマッチング損失を最小化するニューラルネットワークを $\Phi(L,W,S,B)$ から選び, $\hat{s}(x,t)$ とする. ここで条件付き分布 $\nabla\log p_{t_j}(x_j|x_{0,i_j})$ は陽に計算可能である.

$$\frac{1}{n}\sum_{i=1}^{n}\int_{\underline{T}}^{\overline{T}}\mathop{\mathbb{E}}_{x_t\sim p_t(x_t|x_{0,i})}[\|s(x_t,t)-\nabla\log p_t(x_t|x_{0,i})\|^2]\mathrm{d}t.$$

4 主結果

定理 4. $L,W,S,B,\overline{T}$ を適切に定めた下で生成データの分布 q と真の分布 p の全変動距離は,

$$\mathbb{E}[\mathrm{TV}(p_0,q)]\lesssim n^{-s/(2s+d)}\log^8 n.$$

一方 n データに基づく任意の推定量 $\hat{\mu}$ に対して,

$$\inf_{\hat{\mu}}\sup_{p:\ \|p\|_{B^s_{p,q}}\le C}\mathbb{E}[\mathrm{TV}(\hat{\mu},p)]\gtrsim n^{-s/(2s+d)}.$$

つまり拡散モデルは (ほぼ) ミニマックス最適レートを達成する優れた分布推定能力を持つ. Wasserstein 距離 W_1 に対しても同様の結果が得られる.

上界の証明は ReLU ネットワークの近似理論 [5] を拡張した, 拡散 B-spline 基底 $E_{k,j}(x,t)$ によるスコア近似のための新たな基底展開に基づく. 例えば $p_t(x)$ は,

$$p_t(x)\approx\sum_{(k,j)}\alpha_{(k,j)}\underbrace{\int M^d_{k,j}(y)K_t(x|y)\mathrm{d}y}_{=:E_{k,j}(x,t)}.$$

と近似でき, ここで $M^d_{k,j}(y)$ は B-spline 基底, $K_t(x|y)$ は拡散項に対応する. 下界の証明は Besov 空間の L^1-ノルムに関する被覆数の評価に基づく.

5 低次元性を持つ分布での次元の呪いの回避

定理 4 では n の指数に次元 d が現れるが, 現実に倣って $d\gg 1$ の状況を考えると推定誤差の減衰は遅くなり, 次元の呪いと呼ばれる. 現実の拡散モデルが次元の呪いを回避するのは, データの低次元構造ゆえと考えられる. そこで, 真の分布が $d'(\le d)$ 次元部分空間に局在し部分空間の座標の関数として仮定 2 を満たすとする. すると, n の指数は d' のみに依存する.

定理 5. 任意の定数 $\delta>0$ に対し, n データを用いてニューラルネットワークを訓練すると以下が成り立つ:

$$\mathbb{E}[W_1(p_0,q)]\lesssim n^{-\frac{(s+1-\delta)}{d'+2s}}.$$

Part III: 構造を持つ高次元学習問題の統計的・計算量的複雑度

高次元データは本質的低次元性を持ち, その構造に適応することで効率的な学習ができる [3]. 統計的・計算量的解析を両立すべく, そうした関数の学習を考えた.

非等方的入力の k-スパースパリティ 入力の k 方向の射影成分の積の正負を学習する問題を考え, 目的関数が少数の方向にのみ依存するスパース性を定式化した. 更に, 説明変数が重要な方向を向く非等方性を持つ状況を考え, 非等方性の強さによって定まる実質的次元 d_{eff} を用いて統計的・計算量的複雑度を評価した.

定理 6. 手法毎に学習に必要なサンプル数は以下:

ニューラルネット: $\tilde{O}(d_{\mathrm{eff}})$, カーネル法: $\Omega(d_{\mathrm{eff}}^{k-o(1)})$.

スパース関数の有限和 最後に, (1) で各 f_i が 1 次元に依存する p 次多項式である問題を考えた. 勾配法による学習を記述し, いくつかの下界を示した. 1 種類の f_i を学習する問題は [1] で調べられ, 必要サンプル数のオーダーは p に依存しないことが分かっている. 一方, (1) では p を大きくすると下界がいくらでも悪化することが示され, (1) の本質的な難しさが明らかになった.

6 結論

本論文では 1 次勾配情報に基づく手法の理論的評価を与えた. また, 学習の本質的困難性が高次元の最悪ケースから導かれ, 現実のデータの持つ構造が実質的な次元を減らすことで上界・下界に影響する様子をそれぞれの例に普遍的に見ることができた.

参考文献

[1] Chen, S. and Meka, R. Learning polynomials of few relevant dimensions. COLT 2020.

[2] Chen, S., Chewi, S., Li, J., Li, Y., Salim, A. and Zhang, A. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. ICLR 2023.

[3] Damian, A., Lee, J., and Soltanolkotabi, M. Neural Networks can Learn Representations with Gradient Descent. COLT 2022.

[4] Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. ICLR 2020.

[5] Suzuki, T. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. ICLR 2018.

[6] Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. SARAH: A novel method for machine learning problems using stochastic recursive gradient. ICML 2017.

[7] Zhou, D. and Gu, Q. Lower bounds for smooth nonconvex finite-sum optimization. ICML 2019.