修士論文 要旨

二次最適性を保証する分散確率的最適化手法

数理情報学専攻 48196203　荒毛 大輔
指導教員　鈴木 大慈 准教授

1　はじめに

本論文では，次のような形式の制約なし非凸最適化問題を，P ワーカーを並列に用いる分散環境において解くことを考える:

$$\min_{x\in\mathbb{R}^d} f(x) = \frac{1}{P}\sum_{p=1}^{P} f^p(x),$$
$$f^p(x) = \mathrm{E}_{z\sim\mathcal{D}_p}[f_z^p(x)].$$

ただし，各ワーカー $p \in [P]$ が直接扱えるのはデータ $\mathcal{D}_p$ およびデータ点 $z \in \mathcal{D}_p$ における損失関数 $f_z^p(x)$ のみである．

非凸最適化問題において大域的最適解を得る問題は一般に NP 困難であるため，局所最適解を得ることを目標とする．そのために，近似的局所最適解の一種である近似的一次停留点と近似的二次停留点を次のように定義する．

Definition 1.　$\epsilon > 0$, 一階微分可能な関数 $f : \mathcal{X} \to \mathbb{R}$ に対し $x \in \mathcal{X}$ が f の ϵ-一次停留点であるとは,

$$\|\nabla f(x)\| \le \epsilon$$

が成り立つことをいう．

Definition 2.　$\epsilon, \delta > 0$, 二階微分可能な関数 $f : \mathcal{X} \to \mathbb{R}$ に対し $x \in \mathcal{X}$ が f の (ϵ, δ)-二次停留点であるとは,

$$\|\nabla f(x)\| \le \epsilon \text{ かつ } \lambda_{\min}(\nabla^2 f(x)) \ge -\delta$$

が成り立つことをいう．

一次停留点は極小点, 極大点, 鞍点からなるが, Matrix Sensing [2]，Matrix Completion [4] といった問題やある種の深層ニューラルネットワークの学習 [5] において，任意の極小点が大域的最適解になる．このような背景から，性質の悪い停留点である極大点や鞍点（本論文ではこれらをまとめて鞍点と呼ぶ）への収束を回避することもまた重要である．そのために二次停留点を得る手法が研究されている．二次停留点は Strict Saddle Property [3] を仮定することによりただちに極小点となる．

2　既存研究

まず非分散設定における二次最適性を保証する手法について説明する．二次停留点を得るためには，鞍点を訪れた際に効率的に脱出すること，すなわち Hessian の最小固有値に対応する固有ベクトルの方向に座標を更新することが重要である．このために Hessian を直接用いる手法 [7] や Hessian の最小固有値方向を探索するサブルーチン (NC-Search) を用いる手法 [1] が知られている．直接用いる場合は毎更新に $\nabla^2 f(x)$ と同じサイズの計算量がかかるため高次元のタスクには適用が難しく，NC-Search を用いる手法は計算量的な問題は回避しているもののなお Hessian に関する計算を行うために複雑な実装が必要であるという問題がある．これに対して SSRGD (Simple Stochastic Recursive Gradient Descent) [6] は Hessian に関する情報を用いずに二次最適性を保証するアルゴリズムである．これを Algorithm 1 に示す．SSRGD の特徴は各エポックの初期点 $\tilde{x}_s$ において一次最適性を満たす，すなわち $\|\nabla f(\tilde{x}_s)\| \le \epsilon$（無限和の場合はミニバッチを用いる）ならばノイズを加え，スーパーエポックと呼ばれる特殊なエポックに突入することである．スーパーエポックは，あるステップ数の更新を行うか関数値がある程度減少した場合に終える．ノイズを加えることで，$\tilde{x}_s$ が鞍点である場合にはスーパーエポックを経て高確率で鞍点を脱出する効果がある．

次に分散設定において一次最適性を保証する手法について説明する．分散環境では，SGD の素朴な拡張として，毎回の更新においてミニバッチをワーカー数で分割し，勾配情報を集約して全体の勾配評価とする方法がある．この場合，毎回の更新において通信が発生する．それに対し PR-SGD (Parallel Restarted SGD) は，インターバル K を定め，各ワーカーが K 回の独立な SGD による更新を行うごとに座標の平均を取って全体の座標を更新する手法である．PR-SGD は，K を適切に定めれば，素朴な分散型 SGD と同等の勾配計算量で一次最適性を保証しながら，通信コストを $1/K$ 倍にできる．

Algorithm 1 SSRGD

input 初期値 $\tilde{x}_0$，ステップサイズ η，エポック数 S，エポック長 K，ミニバッチ数 b.

for $s = 1$ to S **do**
  $v_0 = \nabla f(\tilde{x}_{s-1})$ を計算する.
  if スーパーエポック中でなく，$\|v_0\| \leq \epsilon$ **then**
    スーパーエポックを開始する.
    $u \sim \mathrm{Unif}(\mathcal{B}(\tilde{x}_{s-1}))$ をサンプルし $x_0 = u$ とする.
  end if
  if スーパーエポック中で, 停止条件を満たす **then**
    スーパーエポックを終了する.
  end if
  for $\tau = 1$ to K **do**
    $v_{\tau-1} = \frac{1}{b}\sum_{z\in I_b}(\nabla f_z(x_{\tau-1}) - \nabla f_z(x_{\tau-2})) + v_{\tau-1}$ とする.
    $x_\tau = x_{\tau-1} - \eta v_{\tau-1}$ と更新する.
  end for
  $\tau \sim [K]$ を一様にサンプルし $\tilde{x}_s = x_\tau$ とする.
end for

3 Distributed SSRGD

分散環境において二次最適性を保証する手法である D-SSRGD (Distributed SSRGD) を提案する．D-SSRGD では，エポックの最初で $\nabla f^p(x)$ の計算を並列に行い，その結果を集約して $v_0 = \nabla f(x) = \frac{1}{P}\sum_{p=1}^P \nabla f^p(x)$ とする．そして一次最適性 $\|v_0\| \leq \epsilon$ を満たすかを確認し，満たすならばノイズを加えてスーパーエポックを経ることにより，二次最適性を満たさない場合は鞍点を脱出する．また，エポック中の座標の更新は並列せず代表的なワーカーが単独で計算する．D-SSRGD の収束レートは次の系にまとめられる．ただし $L, \rho > 0$ は勾配と Hessian の Lipschitz 性，$\zeta > 0$ はワーカー間の損失関数の Hessian の近さからの仮定からくる定数で，$\Delta_0 = f(x_0) - \min_{x\in\mathbb{R}^d} f(x)$ である．

Corollary 1. $t_{\mathrm{thres}}, \eta, f_{\mathrm{thres}}, r, S > 0$ を適切に定め，特に $\zeta \leq L\sqrt{P/n}$ の時に $b = K = \tilde{\Theta}(\sqrt{n/P})$ とすれば，各ワーカーの総勾配計算量

$$\tilde{\Theta}\left(\sqrt{\frac{n}{P}}\frac{L\Delta_0}{\epsilon^2} + \sqrt{\frac{n}{P}}\frac{L\rho^2\Delta_0}{\delta^4} + \frac{n}{P}\frac{\rho^2\Delta_0}{\delta^3}\right)$$

によって D-SSRGD の出力 $\{\tilde{x}_s^{\mathrm{out}}\}_{s=1}^S$ は (ϵ, δ)-二次最適性条件を満たす解を高い確率で含む.

Corollary 2. $t_{\mathrm{thres}}, \eta, f_{\mathrm{thres}}, r, S > 0$ を適切に定め，特に $\zeta \leq L\sqrt{P\epsilon^2/\sigma^2}$ の時に $b = K = \tilde{\Theta}(\sqrt{\sigma^2/P\epsilon^2})$ とすれば，各ワーカーの総勾配計算量

$$\tilde{\Theta}\left(\sqrt{\frac{\sigma^2}{P\epsilon^2}}\frac{L\Delta_0}{\epsilon^2} + \sqrt{\frac{\sigma^2}{P\epsilon^2}}\frac{L\rho^2\Delta_0}{\delta^4} + \frac{\sigma^2}{P\epsilon^2}\frac{\rho^2\Delta_0}{\delta^3}\right)$$

によって D-SSRGD の出力 $\{\tilde{x}_s^{\mathrm{out}}\}_{s=1}^S$ は (ϵ, δ)-二次最適性条件を満たす解を高い確率で含む.

これを SSRGD の結果と比較すると，有限和の場合には n を n/P で，無限和の場合は σ^2/ϵ^2 を $\sigma^2/P\epsilon^2$ で置き換えた形となる．これは $v_0 = \frac{1}{P}\sum_{p=1}^P \nabla f^p(x)$ を計算する際に各ワーカーがサンプルするデータ数に対応しており，自然な線形高速化である.

また, 座標の更新で毎回通信をする場合と比較すると，通信回数は $1/K = \tilde{\Theta}(\sqrt{P/n})$ 倍または $\tilde{\Theta}(\sqrt{P\epsilon^2/\sigma^2})$ 倍に削減される.

参考文献

[1] Z. Allen-Zhu and Y. Li. Neon2: Finding local minima via first-order oracles. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 3716–3726. Curran Associates, Inc., 2018.

[2] S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 3873–3881. Curran Associates, Inc., 2016.

[3] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points — online stochastic gradient for tensor decomposition. In P. Grünwald, E. Hazan, and S. Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 797–842, Paris, France, 03–06 Jul 2015. PMLR.

[4] R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 2973–2981. Curran Associates, Inc., 2016.

[5] R. Ge, J. D. Lee, and T. Ma. Learning one-hidden-layer neural networks with landscape design, 2017.

[6] Z. Li. Ssrgd: Simple stochastic recursive gradient descent for escaping saddle points. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 1523–1533. Curran Associates, Inc., 2019.

[7] Y. Nesterov and B. T. Polyak. Cubic regularization of newton method and its global performance. *Math. Program.*, 108(1):177–– 205, Aug. 2006.