日本オペレーションズ・リサーチ学会
2019年 秋季研究発表会

共役勾配を適用した確率的最適化アルゴリズムと深層学習への応用

明治大学大学院　*小林悠　KOBAYASHI Yu
01016200　明治大学　飯塚秀明　IIDUKA Hideaki

1. はじめに

確率的最適化アルゴリズムは，機械学習や深層学習の分野において，誤差平方和関数の最小化などの最適化問題を解くための手法として重要な役割を担ってきた．特に，確率的勾配を用いたAdam[1]などの最適化アルゴリズムがこれらの分野で広く用いられている．一方，非線形共役勾配法 [2] は非線形無制約最適化問題を解くための手法として知られている．この発表では，機械学習の領域で注目を集めているAdamのアルゴリズムに対する収束性の保証のための新たなアプローチとして，確率的勾配から計算した共役勾配を用いたアルゴリズムを提案する．

機械学習の分野では，確率的勾配を用いた確率的最適化アルゴリズムが多く知られているが，そのほとんどは確率的勾配降下法 [3] から派生したものであると捉えることができる．例えば，Momentum[4] は確率的勾配降下法の更新式において，過去の確率的勾配から計算した指数移動平均の項を追加したものである．また，AdaGrad[5] は確率的勾配を要素ごとに二乗したものをステップ幅に反映させたもので，AdaGradを改善したAdamは機械学習や深層学習において広く用いられている．しかしながら，Adamの収束性は一般には保証されず，AMSGrad[6] などの修正されたアルゴリズムも提案されている．

非線形共役勾配法は，ニュートン法や準ニュートン法のようにヘッセ行列の計算を必要としないため，大規模な問題を解く場合でも各反復ごとの計算量も非常に少ないが，最急降下法よりも高速に最適解へ収束するという利点がある．よって，機械学習のような大規模な訓練データを必要とする問題に対しても有効な手法であることが期待できる．

この発表では，まず従来のAdamに共役勾配を取り入れた新しいアルゴリズムを示す．次にその収束性について議論する．さらに，深層学習を用いた数値実験によって提案手法の性質や性能について議論する．

2. 準備

$f:\mathbb{R}^N\to\mathbb{R}$ は N 次元ユークリッド空間 $\mathbb{R}^N$ 上で微分可能でノイズを含む確率的な目的関数であり，確率分布 P に従う乱数 ξ の実現値を $\xi_1,\xi_2,\ldots$ とする．このとき，各ステップ $t=1,2,\ldots,T$ における f の関数値を $f_1(\boldsymbol{x}),f_2(\boldsymbol{x}),\ldots,f_T(\boldsymbol{x})$ とし，$\mathbf{g}_t=[\mathrm{g}_{t,1},\mathrm{g}_{t,2},\ldots,\mathrm{g}_{t,N}]^\top:=\nabla f_{\xi_t}(\boldsymbol{x}_t)$ を確率的勾配とする．$\mathrm{g}_{t,i}$ は $\mathbf{g}_t$ の i 番目の成分であり，$\tilde{\mathbf{g}}_t:=[\mathrm{g}_{t,1}^2,\mathrm{g}_{t,2}^2,\ldots,\mathrm{g}_{t,N}^2]^\top$ とする．

2.1. 確率的最適化問題

以降，$f_1(\boldsymbol{x}),f_2(\boldsymbol{x}),\ldots,f_T(\boldsymbol{x})$ に対して，次の無制約最小化問題を考える．

$$\text{Minimize } \sum_{t=1}^{T} f_t(\boldsymbol{x}),$$
$$\text{subject to } \boldsymbol{x}\in\mathbb{R}^N.$$

2.2. 非線形共役勾配法

反復法は任意の初期点 $\boldsymbol{x}_0\in\mathbb{R}^N$ から出発し，次の反復式によって点列を更新する．

$$\boldsymbol{x}_{n+1}:=\boldsymbol{x}_n+\alpha_n\boldsymbol{d}_n.$$

非線形共役勾配法は反復法の反復式において探索方向 $\boldsymbol{d}_n$ を次の式で計算したものである．

$$\boldsymbol{d}_n:=\begin{cases}-\boldsymbol{g}_n, & n=0,\\ -\boldsymbol{g}_n+\beta_n\boldsymbol{d}_{n-1}, & n\geq 1.\end{cases}$$

ここで，$\boldsymbol{g}_n$ は目的関数の $\boldsymbol{x}_n$ における勾配である．パラメータ β_n の選択法としては，以下の公式がよく知られている [2]．ただし，$\boldsymbol{y}_{n-1}:=\boldsymbol{g}_n-\boldsymbol{g}_{n-1}$

とする．

$$\beta^{\mathrm{FR}} := \frac{\|\boldsymbol{g}_n\|^2}{\|\boldsymbol{g}_{n-1}\|^2} \quad \text{(Fletcher-Reeves)},$$
$$\beta^{\mathrm{HS}} := \frac{\langle \boldsymbol{g}_n, \boldsymbol{y}_{n-1}\rangle}{\langle \boldsymbol{d}_{n-1}, \boldsymbol{y}_{n-1}\rangle} \quad \text{(Hestenes-Stiefel)},$$
$$\beta^{\mathrm{PRP}} := \frac{\langle \boldsymbol{g}_n, \boldsymbol{y}_{n-1}\rangle}{\|\boldsymbol{g}_{n-1}\|^2} \quad \text{(Polak-Ribière)},$$
$$\beta^{\mathrm{DY}} := \frac{\|\boldsymbol{g}_n\|^2}{\langle \boldsymbol{d}_{n-1}, \boldsymbol{y}_{n-1}\rangle} \quad \text{(Dai-Yuan)}.$$

3. 提案アルゴリズム

まず，既存の手法である Adam のアルゴリズムを以下に示す．

Algorithm 1 Adam[1]

Require: $\boldsymbol{x}_0 \in \mathbb{R}^N, f_1, \ldots, f_T : \mathbb{R}^N \to \mathbb{R}, \alpha \in \mathbb{R}^+, \beta_1, \beta_2 \in [0,1), \epsilon := 10^{-8}$
$t \leftarrow 0, \boldsymbol{m}_0 := \boldsymbol{0}, \boldsymbol{v}_0 := \boldsymbol{0}$
while $\boldsymbol{x}_t$ not converged **do**
$\mathbf{g}_{t+1} := \nabla_{\boldsymbol{x}_t} f_{t+1}(\boldsymbol{x}_t)$
$\tilde{\mathbf{g}}_{t+1} := [\mathsf{g}^2_{t+1,1}, \mathsf{g}^2_{t+1,2}, \ldots, \mathsf{g}^2_{t+1,N}]^\top$
$\boldsymbol{m}_{t+1} := \beta_1 \boldsymbol{m}_t + (1-\beta_1)\mathbf{g}_{t+1}$
$\boldsymbol{v}_{t+1} := \beta_2 \boldsymbol{v}_t + (1-\beta_2)\tilde{\mathbf{g}}_{t+1}$
$\widehat{\boldsymbol{m}}_{t+1} := (1-\beta_1^t)^{-1}\boldsymbol{m}_{t+1}$
$\widehat{\boldsymbol{v}}_{t+1} := (1-\beta_2^t)^{-1}\boldsymbol{v}_{t+1}$
$\boldsymbol{d}_{t+1} := [\widehat{m}_{t+1,1}/(\widehat{v}_{t+1,1}+\epsilon), \widehat{m}_{t+1,2}/(\widehat{v}_{t+1,2}+\epsilon), \ldots, \widehat{m}_{t+1,N}/(\widehat{v}_{t+1,N}+\epsilon)]^\top$
$\boldsymbol{x}_{t+1} := \boldsymbol{x}_t - \alpha \boldsymbol{d}_{t+1}$
$t \leftarrow t+1$
end while

これに対して，提案アルゴリズムは共役勾配を適用したものである．提案アルゴリズムの詳細については発表内で議論する．

4. 数値実験

この発表では，既存手法と提案手法の比較のための数値実験として，再帰的ニューラルネットワーク (RNN) と呼ばれる深層ニューラルネットワークを用いた深層学習を行った．データセットは Penn Treebank (PTB) データセット [7] を用いて，出現単語を予測する言語モデルの学習を行い，以下の式で示されるパープレキシティと呼ばれる指標によって各エポックにおけるモデルの評価を行った．

$$L(\boldsymbol{x}) := \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} y_{tk} \log z_{tk}(\boldsymbol{x}),$$
$$P := \exp L(\boldsymbol{x}).$$

ここで，K は出力の個数，y_{tk} は正解データ，$z_{tk}(\boldsymbol{x})$ は活性化関数の出力，$L(\boldsymbol{x})$ は損失関数である．

参考文献

[1] Kingma, D. P. and Ba, J. L.: Adam, A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[2] Hager, W. W. and Zhang, H.: A survey of nonlinear conjugate gradient methods, Pacific journal of Optimization, 2 (1) (2006) 35–58

[3] Robbins, H. and Monro, S.: A stochastic approximation method, Herbert Robbins Selected Papers, Springer (1985) 102–109

[4] Qian, N.: On the momentum term in gradient descent learning algorithms, Neural Networks 12 (1) (1999) 145–151

[5] Duchi, J. and Hazan, E. and Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization, Journal of Machine Learning Research 2 (2011) 2121–2159

[6] Reddi, S. J., Kale, S., and Kumar, S.: On the convergence of adam and beyond, arXiv preprint arXiv:1904.09237 (2019)

[7] Mikolov, T.: Penn Treebank dataset, http://www.fit.vutbr.cz/ imikolov/rnnlm/