日本オペレーションズ・リサーチ学会
2021年 春季研究発表会

深層学習における適応的共役勾配法

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

経験損失最小化は、深層学習において学習の根幹を担う重要な最適化問題である．その解法として確率的勾配降下法 [1] を筆頭とする確率的勾配を用いた確率的最適化アルゴリズムが知られている．特に，適応的なアルゴリズムと呼ばれる RMSProp [2] や AdaGrad [3]，Adam [4] が広く用いられている．Adam は RMSProp や AdaGrad を組み合わせたアルゴリズムであり，局所解への高速な収束性を持つことからその改善手法も多く提案されており，AMSGrad [5] は著名な深層学習ライブラリ[1]でも実装されている．

一方，非線形共役勾配法 [6] は，無制約非線形最適化問題のための手法として広く知られている．非線形共役勾配法は，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．つまり，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．よって，本発表では既存の確率的最適化アルゴリズムの一部に確率的勾配から生成した共役勾配方向を適用したアルゴリズムを提案する．

この発表では，確率的に近似した共役勾配方向を既存の Adam や AMSGrad に取り入れた新しいアルゴリズムを示し，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. 確率的最適化問題

$\mathbb{R}^N$ を N 次元ユークリッド空間とし，$\mathbb{R}^N$ の内積 $\langle \cdot, \cdot \rangle$ から誘導されるノルムを $\|\cdot\|$ とする．$\mathcal{T} := \{1, 2, \ldots, T\}$ とする．

仮定 2.1

(A1) $\mathcal{F} \subset \mathbb{R}^N$ は空ではない凸閉集合であり，$\mathcal{F}$ への射影は容易に計算できる．

(A2) $f\colon \mathbb{R}^N \to \mathbb{R}$ は，任意の $\boldsymbol{x} \in \mathbb{R}^N$ に対して，$f(\boldsymbol{x}) := \mathbb{E}[F(\boldsymbol{x}, \boldsymbol{\xi})]$ と定義され，*well defined* である．ここで，$F\colon \mathbb{R}^N \times \Xi \to \mathbb{R}$ であり，$\boldsymbol{\xi}$ は $\Xi \subset \mathbb{R}^N$ を台とする確率分布 P に従う乱数ベクトルである．

(A3) 独立同一分布に従う乱数ベクトル $\boldsymbol{\xi}$ の実現値からなる点列 $\boldsymbol{\xi}_0, \boldsymbol{\xi}_1, \ldots$ が存在する．

(A4) 任意の $(\boldsymbol{x}, \boldsymbol{\xi}) \in \mathbb{R}^N \times \Xi$ に対して，確率的勾配 $\mathsf{G}(\boldsymbol{x}, \boldsymbol{\xi})$ は，$\mathbf{g}(\boldsymbol{x}) := \mathbb{E}[\mathsf{G}(\boldsymbol{x}, \boldsymbol{\xi})]$，かつ *well defined* であり，$\mathbf{g}(\boldsymbol{x})$ は f の $\boldsymbol{x}$ における勾配 $\nabla f(\boldsymbol{x})$ と一致する．

(A5) $M > 0$ が存在し，任意の $(\boldsymbol{x}, \boldsymbol{\xi}) \in \mathbb{R}^N \times \Xi$ に対して，$\mathbb{E}[\|\mathsf{G}(\boldsymbol{x}, \boldsymbol{\xi})\|] \leq M$ を満たす．

問題 2.1 (確率的最適化問題)

$$\text{Minimize } f(\boldsymbol{x}) := \mathbb{E}\left[F(\boldsymbol{x}, \boldsymbol{\xi})\right] = \frac{1}{T} \sum_{t \in \mathcal{T}} f_t(\boldsymbol{x})$$

$$\text{subject to } \boldsymbol{x} \in \mathcal{F}.$$

3. 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．

$$\boldsymbol{d}_n := \begin{cases} -\nabla f(\boldsymbol{x}_n), & n = 0, \\ -\nabla f(\boldsymbol{x}_n) + \gamma_n \boldsymbol{d}_{n-1}, & n \geq 1. \end{cases}$$

パラメータ γ_n の選択法としては，以下の公式がよく知られている [6]．ただし，$\boldsymbol{y}_{n-1} := \nabla f(\boldsymbol{x}_n) - \nabla f(\boldsymbol{x}_{n-1})$，$\lambda > 1/4$ とする．

[1] https://pytorch.org/docs/stable/optim.html

$$\gamma_n^{\mathrm{FR}} := \frac{\|\boldsymbol{g}_n\|^2}{\|\boldsymbol{g}_{n-1})\|^2} \quad \text{(Fletcher–Reeves)},$$

$$\gamma_n^{\mathrm{HS}} := \frac{\langle \boldsymbol{g}_n, \boldsymbol{y}_{n-1}\rangle}{\langle \boldsymbol{d}_{n-1}, \boldsymbol{y}_{n-1}\rangle} \quad \text{(Hestenes–Stiefel)},$$

$$\gamma_n^{\mathrm{PRP}} := \frac{\langle \boldsymbol{g}_n, \boldsymbol{y}_{n-1}\rangle}{\|\boldsymbol{g}_{n-1})\|^2} \quad \text{(Polak–Ribière)},$$

$$\gamma_n^{\mathrm{DY}} := \frac{\|\boldsymbol{g}_n\|^2}{\langle \boldsymbol{d}_{n-1}, \boldsymbol{y}_{n-1}\rangle} \quad \text{(Dai–Yuan)},$$

$$\gamma_n^{\mathrm{HZ}} := \frac{\langle \boldsymbol{g}_n, \boldsymbol{y}_{n-1}\rangle}{\langle \boldsymbol{d}_{n-1}, \boldsymbol{y}_{n-1}\rangle} - \lambda \frac{\|\boldsymbol{y}_{n-1}\|^2 \langle \boldsymbol{g}_n, \boldsymbol{d}_{n-1}\rangle}{\langle \boldsymbol{d}_{n-1}, \boldsymbol{y}_{n-1}\rangle^2} \quad \text{(Hager–Zhang)},$$

4. Conjugate-gradient-Based Adam

提案アルゴリズムを Algorithm 1 に示す．ここ

Algorithm 1 Conjugate-gradient-Based Adam

Require: $(\alpha_n)_{n\in\mathbb{N}} \subset (0,1)$, $(\beta_{1n})_{n\in\mathbb{N}} \subset [0,1)$, $\hat{\beta}_1 \in [0,1)$

$n \leftarrow 0, \boldsymbol{x}_0 \in \mathbb{R}^N$,$\boldsymbol{m}_{-1} := \boldsymbol{0}, \boldsymbol{g}_0 := \boldsymbol{0}$
loop
  for $t = 1, 2, \ldots, T$ **do**
    $\mathsf{G}_t := \mathsf{G}(\boldsymbol{x}_n, \boldsymbol{\xi}_t)$
    if $t = 1$ **then**
      $\mathsf{CG}_t := -\mathsf{G}_t$
    else
      $\mathsf{CG}_t \in \mathbb{R}^N$
    end if
    $\boldsymbol{m}_t := \beta_{1t}\boldsymbol{m}_{t-1} + (1-\beta_{1t})\mathsf{CG}_t$
    $\hat{\boldsymbol{m}}_n := \frac{\boldsymbol{m}_n}{1-\hat{\beta}^{n+1}}$
    $\mathsf{H}_t \in \mathrm{diag}((h_{t,i} : h_{t,i} > 0)_{i=1,2,\ldots,N})$
    Find $\mathbf{d}_t \in \mathbb{R}^N$ that solves $\mathsf{H}_t\mathbf{d} = -\hat{\boldsymbol{m}}_t$
    $\boldsymbol{x}_t := \Pi_{\mathcal{F},\mathsf{H}_t}(\boldsymbol{x}_{t-1} + \alpha_t \mathbf{d}_t)$
    $t \leftarrow t+1$,$n \leftarrow n+1$
  end for
  $\boldsymbol{g}_n := \frac{1}{T}\sum_{t\in\mathcal{T}} \nabla f_t(\boldsymbol{x}_n)$
  $\bar{\gamma}_n := \gamma(\boldsymbol{g}_n, \boldsymbol{g}_{n-1}, \bar{\boldsymbol{d}}_{n-1})$
  $\bar{\boldsymbol{d}}_n := -\boldsymbol{g}_n + \bar{\gamma}_n \bar{\boldsymbol{d}}_{n-1}$
end loop

で、確率的共役勾配方向における近似の仕方は、勾配と共役勾配方向とで，それぞれ確率的なものと決定的なものを使い分けることでいくつかの組み合わせを考えることができる．

また，$\mathsf{H}_t \in \mathbb{R}^{N\times N}$ は学習率を適応的に調整するために計算される行列であり，Adam と AMSGrad とで計算方法が異なる [7]．この際，確率的勾配 $\mathsf{G}(\boldsymbol{x}_t, \boldsymbol{\xi}_t)$ を用いている部分で確率的勾配方向 CG_t $(t \in \mathcal{T})$ を用いることもできる．

さらに，$\gamma\colon \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^N$ は（決定的）共役勾配方向の更新パラメータを求める関数である．$\{\gamma_n^{\mathrm{FR}}, \gamma_n^{\mathrm{HS}}, \gamma_n^{\mathrm{PRP}}, \gamma_n^{\mathrm{DY}}\gamma_n^{\mathrm{HZ}}\}$ から選び，$\boldsymbol{d}_{n-1} := \bar{\boldsymbol{d}}_{n-1}$ とする．

収束解析や数値実験の詳細に関しては，発表内で説明する．

参考文献

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

[2] Tieleman, T. and Hinton, G.:Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude, COURSERA: Neural networks for machine learning 4 (2) (2012) 26–31

[3] 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

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

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

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

[7] Hideaki, I.: Appropriate Learning Rates of Adaptive Learning Rate Optimization Algorithms for Training Deep Neural Networks, arXiv preprint arXiv:2002.09647 (2020)