pSGLDの前処理行列最適化手法の比較調査

吉澤亜斗武　　山本和英
長岡技術科学大学
{yoshizawa, yamamoto}@.jnlp.org

1　はじめに

近年、過剰適合を防ぎ不確実性を考慮するモデルの一つとしてベイズ深層学習が注目されている。ベイズ深層学習に用いられるベイズ推論の手法は様々存在するが、自然言語処理の分野において用いられている一つの手法として確率的勾配ランジュバン動力学法がある。この手法においてパラメータの曲率を考慮して学習率を減少させる方法が提案されており、従来の研究ではRMSpropを用いて近似的に求められてきた。しかし、新たな最適化手法は年々提案されているのにも関わらず、多くの研究ではRMSpropを用いているのが現状である。そこで本研究ではRMSprop以降に提案されたAdadelta, Adam, RmspropGravesなどの最適化手法と比較し、確率的勾配ランジュバン動力学法における前処理行列最適化手法の比較調査を行った。

2　関連研究

ベイズ推論の一つの手法として確率的勾配ランジュバン動力学法(stochastic gradient Langevin dynamics ; SGLD)がある。この方法では学習率に応じた分散をもつガウシアンノイズを加えることでよりパラメータ空間を探索でき、学習率を次第に0に近づけていくことでパラメータの事後分布が推定できる。従来はステップ回数に応じたスケジューリングを行っていたが、パラメータの成分のスケールが大きく異なる場合に収束が遅くなることがあるため、リーマン計量であるフィッシャー情報量の逆行列を用いたSGRLD (stochastic gradient Riemannian Langevin dynamics)[1]が提案された。だが多くのモデル場合はフィッシャー情報量の逆行列を求めるのは一般に困難である。一方で、行列の要素のスケールが大きく異なっていても高い精度で逆行列が近似的に求まるESGD (equilibrated stochastic gradient descent)とRMSpropが同様の挙動をすることが実験的に知られている[2]。そこでフィッシャー情報量の逆行列をRMSpropにより近似して得られた前処理行列を用いたpSGLD (preconditioned SGLD)を用いたベイズ深層学習が提案された[3]。近年、このベイズ深層学習を使った手法が自然言語処理の分野にも見られるようになった[4]。

しかしRMSprop以外の最適化手法を用いたpSGLDはほとんど研究されていない。そこで本研究ではRMSprop以降に提案され、その改良にあたるAdadelta、Adam、RMSpropGravesの最適化手法を用いて前処理行列を近似してpSGLDに適用した。また、RMSpropGra vesはモーメンタム項を加えたものを使うことが多いため同様に実験を行った。

3　手法

3.1　pSGLDを用いたベイズ深層学習

pSGLDを用いたベイズ深層学習において損失関数は以下のように表される。

$$U(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta}) - \sum_{n=1}^{N} \log p(\boldsymbol{D_n}|\boldsymbol{\theta}) \qquad (1)$$

ここで$\boldsymbol{\theta}$はパラメータ、Nはデータサイズ、D_nは学習データであり、入力X_nと出力Y_nの組$D_n = (X_n, Y_n)$である。

ミニバッチを用いて確率的勾配法を行うため、バッチサイズ$M \ll N$の時に成立する以下の式を実際には用いる。

$$\tilde{U}_t(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta}) - \frac{N}{M} \sum_{m=1}^{M} \log p(\boldsymbol{D_m}|\boldsymbol{\theta}) \qquad (2)$$

pSGLDにおいてパラメータの更新式は次のように表される。

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\eta}{2} G_t^{-1} \tilde{\boldsymbol{f}}_t + G_t^{-\frac{1}{2}} \mathcal{N}(0, \eta \boldsymbol{I}) \qquad (3)$$

ここで $\tilde{\boldsymbol{f}}_t = \nabla \tilde{U}_t(\theta)$ であり、η はステップサイズ、$\boldsymbol{I}$ は単位行列、$\mathcal{N}(0, \eta \boldsymbol{I})$ はガウシアンノイズ、G_t^{-1} は前処理行列である。

G_t^{-1} を RMSprop で求める際のアルゴリズムは以下のようになる。

Algorithm 1 pSGLD with RMSprop

Input: η, λ, α

Output: $\{\boldsymbol{\theta}_t\}_{t=1:T}$

Initialize:$\boldsymbol{v}_0 \leftarrow 0, \boldsymbol{\theta}_1 \sim \mathcal{N}(0, \boldsymbol{I})$

1: **for** $t = 1$ to T **do**
2: % Estimate gradient from minibatch
3: $\tilde{\boldsymbol{f}}_t = \nabla \tilde{U}_t(\theta)$;
4: % Preconditioning with RMSprop
5: $\boldsymbol{v}_t \leftarrow \alpha \boldsymbol{v}_{t-1} + (1-\alpha) \tilde{\boldsymbol{f}}_t \odot \tilde{\boldsymbol{f}}_t$;
6: $G_t^{-1} \leftarrow \mathrm{diag}\Big(1 \oslash (\lambda 1 + \boldsymbol{v}_t^{\frac{1}{2}})\Big)$;
7: % Parameter update
8: $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\eta}{2} G_t^{-1} \tilde{\boldsymbol{f}}_t + G_t^{-\frac{1}{2}} \mathcal{N}(0, \eta \boldsymbol{I})$
9: **end for**

3.2 最適化手法

本研究では RMSprop を用いて前処理行列を求めるところ、別の最適化手法を用いて比較する。RMSprop 以降に提案され、その改良とされている Adadelta、Adam、RMSpropGraves、RMSpropGraves にモーメンタム項を加えたものを用いる。

- Adadelta
  RMSprop では次元の不整合が生じているという問題点がある。Adadelta ではパラメータの差分の二乗の指数移動平均を用いることでこの問題を解決し、同時に学習率の設定を不要とした。本実験では形式的に $\eta = 1.0$ と設定した。

- Adam
  RMSprop や Adadelta では過去の勾配の情報そのものを考慮していなかった。Adam では過去の勾配との指数移動平均をとり、指数移動平均が有するバイアスを取り除いた不変推定量を用いる。

- RMSpropGraves
  RMSprop は最新の二乗勾配の影響を大きく受ける。RMSpropGraves は最新の二乗勾配の影響をより抑え、前処理行列の値を RMSprop に比べて大きめに推定する傾向にある。

4 実験

本実験では単語レベルの言語モデルタスク、画像キャプション生成タスク、文の分類タスクに関して実験を行った。プログラムは theano により実装し、GPU は NVIDIA GeForce GTX 1080 Ti を用いた。また、本実験ではガウシアンノイズを用いたドロップコネクトや開発データを用いた early-stopping を行っている。

- 言語モデルタスク
  言語モデルタスクでは、Penn Treebank(PTB) コーパス [5] を用い、訓練/開発/テストの単語数は 929K/73K/82K である。また、語彙数は 10K である。ミニバッチサイズは 32、エポック数 55 とし、二層 LSTM を隠れユニット数 1500 として用い、ドロップ率 0.65 とした。また先行研究 [4] において SGLD でも高い精度であるため、同様の実験を行い perplexity を用いて評価した。

- 画像キャプション生成タスク
  画像キャプション生成タスクでは Flickr8k[6] を用いた。このデータセットは 8k の画像にそれぞれ 5 つの文が与えらたものであり、訓練/開発/テストは 6K/1K/1K 枚の画像を用いる。バッチサイズを 64、エポック数を 20 とし、単層 LSTM を隠れ層 512 として用いた。ドロップ率は 0.5 とし、言語モデルタスクとの比較のため SGLD でも実験を行い、perplexity を用いて評価した。

- 分類タスク
  文の分類タスクでは MR[7]、CR[8]、MPQA[9] を用い、十分割交差検証を行った。ミニバッチサイズは 50、エポック数は 20 とし、単層双方向 LSTM を隠れユニット数 400 として用いた。ドロップ率を 0.5 とし、分類誤差を用いて評価した。

最適化手法に用いられるハイパーパラメータはいくつか決めた候補の組み合わせを実験して、各タスクの評価指標が最も良くなるような組み合わせを選んだ。学習率は Adadelta、SGLD、モーメンタム項を加えた RMSpropGraves を除いて、画像キャプション生成タスク、分類タスクで 1.0×10^{-3}、言語モデルタスクでは RMSprop、Adam が 1.0×10^{-3}、RMSpropGraves は 1.0×10^{-2} とした。モーメンタム項を加える場合はその十分の一にし、SGLD の学習率は 1 として焼きなましを行う。

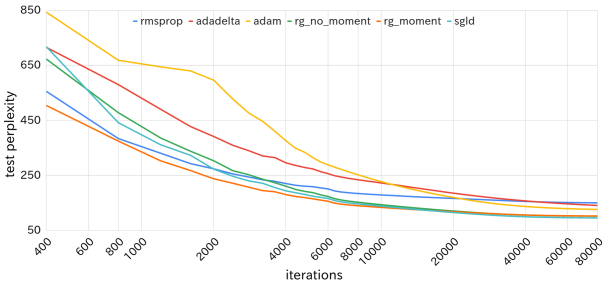

図 1: 学習過程（PTB）

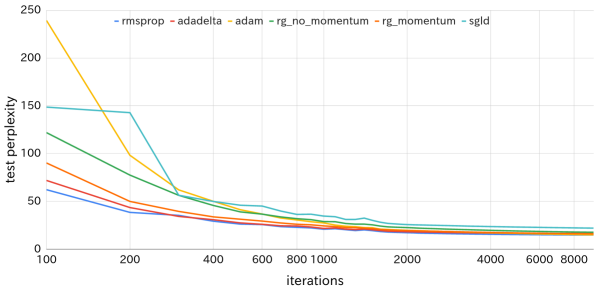

図 2: 学習過程（Flickr8k）

5　結果・考察

一定のイテレーション数ごとにテストデータを入れて評価し、学習過程を可視化した。

5.1　言語モデルタスク

図 1 より RMSprop を用いた場合では、初めは SGLD よりも学習が早く進むが最終的には最も精度が悪化する。Adadelta と Adam は学習が遅いが RMSprop よりも良い精度となった。RMSpropGraves(rg) は SGLD と同様の振る舞いをしているが、モーメンタム項を追加することで RMSprop よりも学習が早く、SGLD との最終的にほぼ同じ精度になった。

これは RMSpropGraves は最新の二乗勾配の影響を抑えるため学習初期は RMSprop の方が早く perplexity が小さくなるが、タスクが難しく鞍点や局所解などに陥った場合に最新の二乗勾配の影響が比較的大きい RMSprop は抜けにくいことを示している。しかし、RMSpropGraves は最新の二乗勾配の影響を抑えるため局所解を比較的に抜け出しやすい。またモーメンタム項を加えることでパラメータ空間の探索をより促し、学習速度の改善ができたのではないかと考えられる。

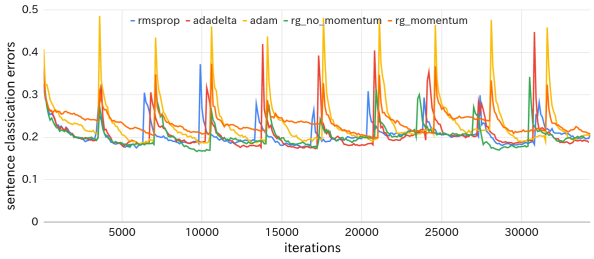

図 3: 学習過程（MR）

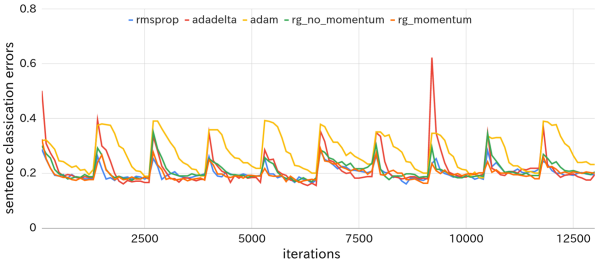

図 4: 学習過程（CR）

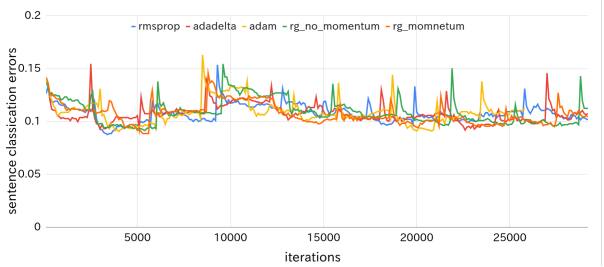

図 5: 学習過程（MPQA）

5.2　画像キャプション生成

図 2 より SGLD は最終的な精度が悪く、学習速度も Adam を除いた pSGLD の方が速い。また言語モデルタスクに比べて全体として学習初期で perplexity がある程度低いことが分かる。そのため RMSpropGraves にモーメンタム項を加えた手法は学習速度は大きくならず、単純な RMSprop や次元の不整合がない Adadelta の方が学習速度は速くなった。

5.3 分類タスク

図3、4、5において最適化手法ごとに異なるイテレーションから十分割交差検証が起こっているのは early-stopping のためである。Adam、Adadelta はデータセットによっては学習初期の誤差が大きいが、Adadelta の方が誤差の減少が速い。図 2 においては RMSprop-Graves にモーメンタム項を加えたものが誤差の減少が遅く、収束しきっていないことが分かる。

これは Adam やモーメンタム項を加えた最適化手法がパラメータ空間の探索をより行っていることを表している。表 1 より各データセットにおける平均誤差が最も少ないのは MR では RMSpropGraves、CR では Adadelta、MPQA では RMSprop となっており、分類タスクのような易しいタスクでは、RMSprop や RMSpropGraves のように過去の勾配の情報を大きく考慮しない手法が学習が速く、収束して early-stopping と比較的になりやすいと考えられる。

表 1: 分類タスクにおける平均誤差と標準偏差

pSGLD with	MR	CR	MPQA
RMSprop	19.11±1.14	18.52±1.48	10.51±0.70
Adadelta	19.29±1.11	18.45±1.58	10.55±0.70
Adam	19.91±0.98	20.61±1.67	10.67±0.87
RMSpropGraves	19.09±1.09	19.03±1.75	10.74±1.02
+momentum	20.16±1.05	18.49±1.35	10.64±0.82

6 おわりに

本研究では pSGLD の前処理行列の最適化手法について RMSprop の他に Adadelta、Adam、RMSprop-Graves を用いて 3 つのタスクにおいて比較実験を行った。その結果、言語モデルタスクでは初めは RMSprop の方が学習が速いが、最終的な精度が一番悪くなった。一方で RMSpropGraves にモーメンタム項を加えたモデルは SGLD よりも学習が速く同様の精度が得られ、RMSprop 以外の最適化手法を検討する必要性を示した。一方、比較的に易しいタスクである像キャプション生成タスクにおいては従来の RMSprop の方が学習が速かった。分類タスクにおいてもデータセットによって誤差が少なくなる最適化手法が異なっており、最適化手法のデータ依存性及びタスク依存性を示した。

謝辞

本研究は、平成 29〜31 年学術研究助成基金助成金挑戦的研究 (萌芽) 課題番号 17K18481 の助成を受けています。

参考文献

[1] Sam Patterson and Yee Whye Teh. Stochastic Gradient Riemannian Langevin Dynamics on the Probability Simplex. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pp. 3102–3110. Curran Associates, Inc., 2013.

[2] Yann Dauphin, Harm de Vries, and Yoshua Bengio. Equilibrated adaptive learning rates for non-convex optimization. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pp. 1504–1512. Curran Associates, Inc., 2015.

[3] Chunyuan Li, Changyou Chen, David Carlson, and Lawrence Carin. Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks. In *Thirtieth AAAI Conference on Artificial Intelligence*, pp. 1788–1794, February 2016.

[4] Zhe Gan, Chunyuan Li, Changyou Chen, Yunchen Pu, Qinliang Su, and Lawrence Carin. Scalable Bayesian Learning of Recurrent Neural Networks for Language Modeling: Supplementary Material. p. 3, 2017.

[5] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, Vol. 19, No. 2, pp. 313–330, 1993.

[6] M. Hodosh, P. Young, and J. Hockenmaier. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research*, Vol. 47, pp. 853–899, August 2013.

[7] Bo Pang and Lillian Lee. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. pp. 115–124, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[8] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pp. 168–177, Seattle, WA, USA, August 2004. Association for Computing Machinery.

[9] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, Vol. 39, No. 2, pp. 165–210, May 2005.