BLOCK-NORMALIZED GRADIENT METHOD: AN EMPIRICAL STUDY FOR TRAINING DEEP NEURAL NETWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we propose a generic and simple strategy for utilizing stochastic gradient information in optimization. The technique essentially contains two consecutive steps in each iteration: 1) computing and normalizing each block (layer) of the mini-batch stochastic gradient; 2) selecting appropriate step size to update the decision variable (parameter) towards the negative of the block-normalized gradient. We conduct extensive empirical studies on various non-convex neural network optimization problems, including multi layer perceptron, convolution neural networks and recurrent neural networks. The results indicate the block-normalized gradient can help accelerate the training of neural networks. In particular, we observe that the normalized gradient methods having constant step size with occasionally decay, such as SGD with momentum, have better performance in the deep convolution neural networks, while those with adaptive step sizes, such as Adam, perform better in recurrent neural networks. Besides, we also observe this line of methods can lead to solutions with better generalization properties, which is confirmed by the performance improvement over strong baselines.

1 INTRODUCTION

Continuous optimization is a core technique for training non-convex sophisticated machine learning models such as deep neural networks (Bengio, 2009). Compared to convex optimization where a global optimal solution is expected, non-convex optimization usually aims to find a stationary point or a local optimal solution of an objective function by iterative algorithms. Among a large volume of optimization algorithms, first-order methods, which only use (stochastic) gradient information of objective functions to update solution, are most widely used due to its relatively low requirement on memory space and computation time.

In this paper, we are particularly interested in solving the deep neural network training problems with first order method. However, compared to other non-convex problems, deep neural network training has the following challenges:

1. Gradient computation is expensive. A full gradient is usually difficult to compute, especially when applied to machine learning problems over big data, because its calculation requires going through the whole dataset.
2. Gradient may be vanishing and/or exploding. The gradient may become very small over some plateaus regions while sharply becoming large on the cliffs of objective function. As the number of layers in the neural network increases, this phenomenon of vanishing or exploding gradients becomes more severe so that the iterative solution will converge very slowly or diverge quickly.

These two challenges indicate that for each iteration, stochastic gradient might be the best practical first order information we can get. Once we obtain this information, we should develop certain mechanism to avoid its exploding or vanishing phenomenon and eventually ensure faster convergence.

In this paper we aim to solve those challenges from the two key ingredients of first-order optimization algorithm per iteration, i.e., an updating direction and a step size (a.k.a. learning rate). The

updating direction indicates where the next iterated point should move to and the step size determines how far the move should end up with.

We propose to use stochastic *normalized gradient* which is constructed by dividing the (stochastic) gradient by its norm. Compared to the regular gradient, normalized gradient only provides an updating direction but does not incorporate the local steepness of the objective through its magnitude, which help to control the change of the solution through a well-designed step length. Intuitively, as it constrains the magnitude of the gradient to be 1, it should to some extent prevent the gradient vanishing or exploding phenomenon. In fact, as showed in (Hazan et al., 2015; Levy, 2016), normalized gradient descent (NGD) methods are more numerically stable and have better theoretical convergence properties than the regular gradient descent method in non-convex optimization.

Once the updating direction is determined, step size is the next important component in the design of first-order methods. While for convex problems there are some well studied strategies to find a stepsize ensuring convergence, for non-convex optimization, the choice of step size is more difficult and critical as it may either enlarge or reduce the impact of the aforementioned vanishing or exploding gradients. When a normalized gradient is used, the step size fully controls the progress of the solution so that the convergence is possible only when the step size is carefully designed.

Among different choices of step sizes, the adaptive *feature-dependent* step size has attracted lots of attention due to its great performance when applied to training deep neural networks (Duchi et al., 2011; Kingma & Ba, 2014). Different from the standard step-size rule which multiplies a same number to each coordinate of gradient, the adaptive feature-dependent step-size rule multiplies different numbers to coordinates of gradient so that different parameters in the learning model can be updated in different paces. For example, the adaptive step size invented by (Duchi et al., 2011) is constructed by aggregating each coordinates of historical gradients. As discussed by (Duchi et al., 2011), this method can dynamically incorporate the frequency of features in the step size so that frequently occurring features will have a small step size while infrequent features have a long step size. The similar adaptive step size is proposed in (Kingma & Ba, 2014) but the historical gradients are integrated into feature-dependent step size by a different weighting scheme. While the adaptive stepsize is useful in the recurrent neural network, as we observe in the experiments, we also find that methods having constant stepsize works better in the deep convolution neural networks.

In this paper, we propose a generic framework using the mini-batch stochastic normalized gradient as the updating direction (like (Hazan et al., 2015; Levy, 2016)) and the step size is either constant or adaptive to each coordinate as in (Duchi et al., 2011; Kingma & Ba, 2014). In this way, we hope the resulting methods can benefit from both techniques and generate a even better performance than the existing methods. Our framework starts with computing regular mini-batch stochastic gradient, which is immediately normalized. The normalized version is then plugged in the current popular adaptive step size methods, such as Adam (Kingma & Ba, 2014) and AdaGrad (Duchi et al., 2011), or the constant stepsize with occasional decay, such as SGD+momentum. The numerical results shows that normalized gradient always helps to improve the performance of the original methods especially when the network structure is deep. It seems to be the first thorough empirical study on various types of neural networks with this normalized gradient idea. Besides, although we focus our empirical studies on deep learning where the objective is highly non-convex, we also provide a convergence proof under this framework when the problem is convex and the stepsize is adaptive in the appendix.

The rest of the paper is organized as follows. In Section 2, we briefly go through the previous work that are related to ours. In Section 3, we formalize the problem to solve and propose the generic algorithm framework. In Section 4, we conduct comprehensive experimental studies to compare the performance of different algorithms on various neural network structures. We conclude the paper in Section 5. Finally, in the appendix, we provide a concrete example of this type of algorithm and show its convergence property under the convex setting.

2 Related Work

A pioneering work on normalized gradient descent (NGD) method was by Nesterov (Nesterov, 1984) where it was shown that NGD can find a ϵ-optimal solution within $O(\frac{1}{\epsilon^2})$ iterations when the objective function is quasi-convex. Kiwiel (Kiwiel, 2001) and Hazan et al (Hazan et al., 2015)

extended NGD for upper semi-continuous quasi-convex objective functions and local-quasi-convex objective functions, respectively, and achieved the same iteration complexity. Moreover, Hazan et al (Hazan et al., 2015) showed that NGD's iteration complexity can be reduced to $O(\frac{1}{\epsilon})$ if the objective function is local-quasi-convex and locally-smooth. A stochastic NGD algorithm is also proposed by Hazan et al (Hazan et al., 2015) which, if a mini-batch is used to construct the stochastic normalized gradient in each iteration, finds ϵ-optimal solution with a high probability for locally-quasi-convex functions within $O(\frac{1}{\epsilon^2})$ iterations. Levy (Levy, 2016) proposed a Saddle-Normalized Gradient Descent (Saddle-NGD) method, which adds a zero-mean Gaussian random noise to the stochastic normalized gradient periodically. When applied to strict-saddle functions with some additional assumption, it is shown (Levy, 2016) that Saddle-NGD can evade the saddle points and find a local minimum point approximately with a high probability.

Analogous yet orthogonal to the gradient normalization ideas have been proposed for the deep neural network training. For example, batch normalization (Ioffe & Szegedy, 2015) is used to address the internal covariate shift phenomenon in the during deep learning training. It benefits from making normalization a part of the model architecture and performing the normalization for each training mini-batch. Weight normalization (Salimans & Kingma, 2016), on the other hand, aims at a reparameterization of the weight vectors that decouples the length of those weight vectors from their direction. Recently (Neyshabur et al., 2015) proposes to use path normalization, an approximate path-regularized steepest descent with respect to a path-wise regularizer related to max-norm regularization to achieve better convergence than vanilla SGD and AdaGrad. Perhaps the most related idea to ours is Gradient clipping. It is proposed in (Pascanu et al., 2013) to avoid the gradient explosion, by pulling the magnitude of a large gradient to a certain level. However, this method does not do anything when the magnitude of the gradient is small.

Adaptive step size has been studied for years in the optimization community. The most celebrated method is the line search scheme. However, while the exact line search is usually computational infeasible, the inexact line search also involves a lot of full gradient evaluation. Hence, they are not suitable for the deep learning setting. Recently, provably correct algorithms with adaptive step sizes in the convex scenario start to be applied to the non-convex neural network training, such as AdaGrad (Duchi et al., 2012), Adam (Kingma & Ba, 2014) and RMSProp. However they directly use the unnormalized gradient, which is different from our framework. To the best of our knowledge, our work seems to be the first study on combining normalized gradient with adaptive stepsize for deep neural network training.

3 Algorithm Framework

In this section, we present our algorithm framework to solve the following general problem:

$$\min_{x\in\mathbb{R}^d} f(x) = \mathbb{E}(F(x,\xi)), \tag{1}$$

where $x = (x^1, x^2, \ldots, x^B) \in \mathbb{R}^d$ with $x^i \in \mathbb{R}^{d_i}$ and $\sum_{i=1}^B d_i = d$, ξ is a random variable following some distribution $\mathbb{P}$, $F(\cdot,\xi)$ is a loss function for each ξ and the expectation $\mathbb{E}$ is taken over ξ. Now our goal is to minimize the objective function f over x, where x can be the parameters of a machine learning model when (1) corresponds to a training problem. Here, the parameters are partitioned into B blocks. The problem of training a neural network is an important instance of (1), where each block of parameters x^i can be viewed as the parameters associated to the ith layer in the network.

We propose the generic optimization framework in Algorithm 1. In iteration t, it firstly computes the partial (sub)gradient $F_i'(x_t,\xi_t)$ of F with respect to x^i for $i = 1, 2, \ldots,$ at $x = x_t$ with a mini-batch data ξ_t, and then normalizes it to get a partial direction $g_t^i = \frac{F_i'(x_t,\xi_t)}{\|F_i'(x_t,\xi_t)\|_2}$. We define $g_t = (g_t^1, g_t^2, \ldots, g_t^B)$. The next is to find d adaptive step sizes $\tau_t \in \mathbb{R}^d$ with each coordinate of τ_t corresponding to a coordinate of x. We also partition τ_t in the same way as x so that $\tau_t = (\tau_t^1, \tau_t^2, \ldots, \tau_t^B) \in \mathbb{R}^B$ with $\tau_t^i \in \mathbb{R}^{d_i}$. We use τ_t as step sizes to update x_t to x_{t+1} as $x_{t+1} = x_t - \tau_t \circ g_t$, where $\circ$ represents coordinate-wise (Hadamard) product. In fact, our framework can be customized to most of existing first order methods with fixed or adaptive step sizes, such as SGD, AdaGrad(Duchi et al., 2011), RMSProp and Adam(Kingma & Ba, 2014), by adopting their step size rules respectively.

Algorithm 1 Generic Block-Normalized Gradient (BNG) Descent

1: Choose $x_1 \in \mathbb{R}^d$.
2: **for** $t = 1, 2, ...,$ **do**
3: Sample a mini-batch of data ξ_t and compute the partial stochastic gradient $g_t^i = \frac{F_i'(x_t,\xi_t)}{\|F_i'(x_t,\xi_t)\|_2}$
4: Let $g_t = (g_t^1, g_t^2, \ldots, g_t^B)$ and choose step sizes $\tau_t \in \mathbb{R}^d$.
5: $x_{t+1} = x_t - \tau_t \circ g_t$
6: **end for**

4 Numerical Experiments

Basic Experiment Setup In this section, we conduct comprehensive numerical experiments on different types of neural networks. The algorithms we are testing are SGD with Momentum (SGDM), AdaGrad (Duchi et al., 2013), Adam (Kingma & Ba, 2014) and their block-normalized gradient counterparts, which are denoted with suffix "NG". Specifically, we partition the parameters into block as $x = (x_1, x_2, \ldots, x_B)$ such that x_i corresponds to the vector of parameters (including the weight matrix and the bias/intercept coefficients) used in the ith layer in the network.

Our experiments are on four diverse tasks, ranging from image classification to natural language processing. The neural network structures under investigation include multi layer perceptron, long-short term memory and convolution neural networks.

To exclude the potential effect that might be introduced by advanced techniques, in all the experiments, we only adopt the basic version of the neural networks, unless otherwise stated. The loss functions for classifications are cross entropy, while the one for language modeling is log perplexity. Since the computational time is proportional to the epochs, we only show the performance versus epochs. Those with running time are similar so we omit them for brevity. For all the algorithms, we use their default settings. More specifically, for Adam/AdamNG, the initial step size scale $\alpha = 0.001$, first order momentum $\beta_1 = 0.9$, second order momentum $\beta_2 = 0.999$, the parameter to avoid division of zero $\epsilon = 1e^{-8}$; for AdaGrad/AdaGradNG, the initial step size scale is 0.01.

4.1 MNIST Image Classification with Feedforward Neural Network

The first network structure we are going to test upon is the Multi Layer Perceptron (MLP). We will adopt the handwritten digit recognition data set MNIST[1]Lecun et al. (1998), in which, each data is an image of hand written digits from $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 0\}$. There are 60k training and 10k testing examples and the task is to tell the right number contained in the test image. Our approach is applying MLP to learn an end-to-end classifier, where the input is the raw 28×28 images and the output is the label probability. The predicted label is the one with the largest probability. In each middle layer of the MLP, the hidden unit number are 100, and the first and last layer respectively contain 784 and 10 units. The activation functions between layers are all sigmoid and the batch size is 100 for all the algorithms.

We choose different numbers of layer from $\{6, 12, 18\}$. The results are shown in Figure 1. Each column of the figures corresponds to the training and testing objective curves of the MLP with a given layer number. From left to right, the layer numbers are respectively 6, 12 and 18. We can see that, when the network is as shallow as containing 6 layers, the normalized stochastic gradient descent can outperform its unnormalized counterpart, while the Adam and AdaGrad are on par with or even slightly better than their unnormalized versions. As the networks become deeper, the acceleration brought by the gradient normalization turns more significant. For example, starting from the second column, AdamNG outperforms Adam in terms of both training and testing convergence. In fact, when the network depth is 18, the AdamNG can still converge to a small objective value while Adam gets stuck from the very beginning. We can observe the similar trend in the comparison between AdaGrad (resp. SGDM) and AdaGradNG (resp. SGDMNG). On the other hand, the algorithms with adaptive step sizes can usually generate a stable learning curve. For example, we can see from the last two column that SGDNG causes significant fluctuation in both training and testing

[1] http://yann.lecun.com/exdb/mnist/

curves. Finally, under any setting, AdamNG is always the best algorithm in terms of convergence performance.

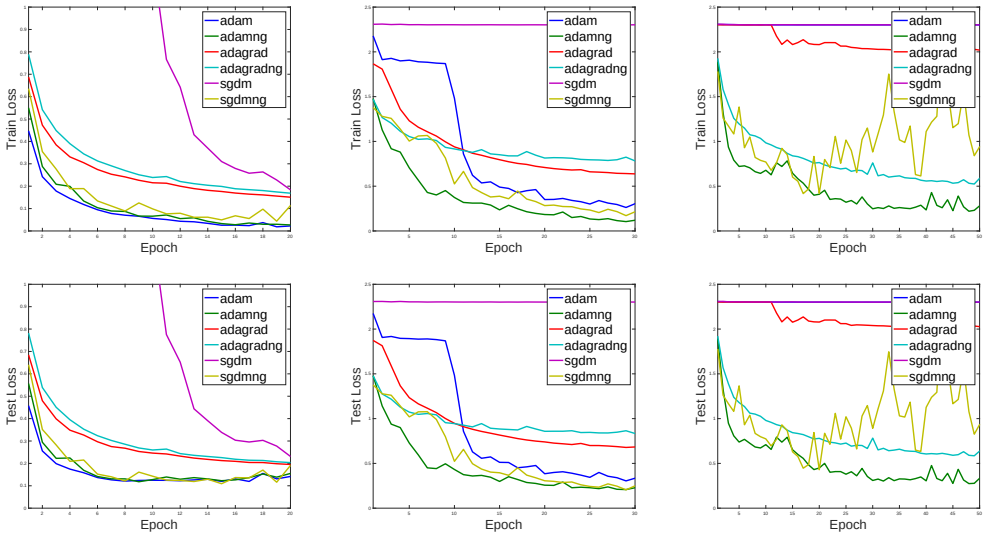

Figure 1: The training and testing objective curves on MNIST dataset with multi layer perceptron. From left to right, the layer numbers are 6, 12 and 18 respectively. The first row is the training curve and the second is testing.

4.2 Residual Network on Cifar 10 and Cifar100

In the following section, we compare the methods on CIFAR (both CIFAR-10 and CIFAR-100) datasets over the residual network He et al. (2016). CIFAR-10 consists of 50,000 training images and 10,000 test images from 10 classes, while CIFAR-100 from 100 classes. Each input image consists of 32×32 pixels. The dataset was preprocessed as described in He et al. (2016) by subtracting the means and dividing the variance for each channel. We follow the simple data augmentation that 4 pixels are padded on each side, and a 32×32 crop is randomly sampled from the padded image or its horizontal flip as described in He et al. (2016).

We compare two major methods, namely SGD and ADAM, who respectively represents the constant step size with occasional decay and adaptive step size. All the algorithms are combined with batch normalization technique. We compare different gradient processing schemes under these two methods, where "BN" is the plain batch normalization without any processing , "Grad clipping" is the standard gradient clipping method, "NG$_{UNIT}$" is the default normalized gradient method, while "NG" (we abuse the notation a bit) is a variant of the NG$_{UNIT}$ method, i.e. NG = NG$_{UNIT}\times$ Norm of variable $\times$ ratio. We empirically observe that NG has better performance than NG$_{UNIT}$.

We follow the exactly same experimental protocol as described in He et al. (2016) and adopt the publicly available Torch implementation[2] for residual network. That is (1) For training, SGD is used with momentum of 0.9, weight decay of 0.0001 and mini-batch size of 128. The initial learning rate is set as 0.1 and dropped by a factor of 0.1 at 80, 120 with a total training of 160 epochs; (2) The weight initionization method is used as He et al. (2015).

We use the residual network architecture with varied depths $L = \{20, 32, 44, 56, 110\}$ and over CIFAR-10 and CIFAR100 dataset. For the extra hyper-parameters: thresh of Clipping, and ratio of NG, we chose the best heyper-paerameter from the 56 layer residual network. Particularly, for the thresh of Clipping, the search values are $0.05, 0.1, 0.5, 1, 5$, and the best value is 0.1. For the ratio of NG, the search values are $0.01, 0.02, 0.05$, and the best value is 0.1

[2] `https://github.com/facebook/fb.resnet.torch`

The results are shown in Figure 2 and Table 1 and 2. We can see that the methods with normalized gradient converges much faster than those without, no matter the step size is constant or adaptive. We also observe that the normalized gradient method, especially the version "NG" has a better generalization properties.

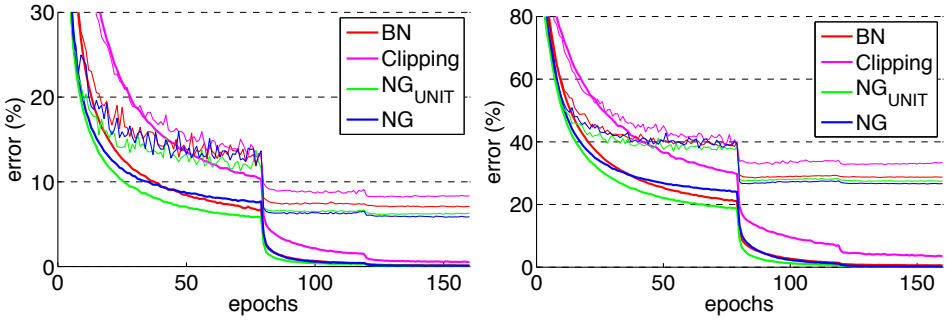

Figure 2: The training and testing objective curves on CIFAR10 and CIFAR100 datasets with Res-110. Left: CIFAR10; Right: CIFAR100. The thick curves are the training while the thin are testing.

Algorithm	Res-20	Res-32	Res-44	Res-56	Res-110
SGD					
BN	7.93	7.15	7.09	7.34	7.07
Grad Clipping	9.03	8.44	8.55	8.30	8.35
NG_{UNIT}	7.82	7.09	6.60	6.59	6.28
NG	**7.71**	**6.90**	**6.43**	**6.19**	**5.87**
ADAM					
BN	9.14	8.33	7.794	7.33	6.75
Grad Clipping	10.18	9.18	8.89	9.24	9.96
NG_{UNIT}	9.42	8.50	8.06	7.69	7.29
NG	**8.52**	**7.62**	**7.28**	**7.04**	**6.71**

Table 1: Error rates of ResNets with different depths on CIFAR 10.

Algorithm	Res-20	Res-32	Res-44	Res-56	Res-110
SGD	32.28	30.62	29.96	29.07	28.79
Grad Clipping	35.06	34.49	33.36	34.00	33.38
NG_{UNIT}	32.46	31.16	30.05	29.42	27.49
NG	**31.43**	**29.56**	**28.92**	**28.48**	**26.72**

Table 2: Error rates of ResNets with different depths on CIFAR 100.

4.3 34 LAYER PRE-RESIDUAL NETWORK FOR IMAGENET CLASSIFICATION

In this section, we further test our methods on ImageNet 2012 classification challenges, which consists of more than 1.2M images from 1,000 classes. We use the given 1.28M labeled images for training and the validation set with 50k images for testing. We employ the validation set as the test set, and evaluate the classification performance based on top-1 and top-5 error. The pre-activation version of ResNet is adopted in our experiments to perform the classification task.

We run our experiments on one GPU and use single scale and single crop test for simplifying discussion. We keep all the experiments settings the same as the publicly available Torch implementation [3], (that is, we apply stochastic gradient descent with momentum of 0.9, weight decay of 0.0001, and

[3] We use the public Torch implementation: `https://github.com/facebook/fb.resnet.torch`

set the initial learning rate to 0.1.). The exception is that we use mini-batch size of 64 and 50 training epochs considering the GPU memory limitations and training time costs. Regarding learning rate annealing, we use exponential decay to 0.001.

Due to the time-consuming nature of training the networks (which usually takes one week) in this experiment, we only compare SGD with our default NG_{UNIT} method on the testing error of the classification. From Table 3, we can see NG_{UNIT} has a non-trivial improvement on the testing error over SGD.

method	Top-1	Top-5
SGD	29.05	9.95
NG_{UNIT}	**28.43**	**9.57**

Table 3: Top-1 and Top 5 error rates of ResNets on ImageNet classification.

4.4 Language Modeling with Recurrent Neural Network

Now we move on to test the algorithms on Recurrent Neural Networks (RNN). In this section, we test the performance of the proposed algorithm on the word-level language modeling task with a popular type of RNN, i.e. single directional Long-Short Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997). The data set under use is Penn Tree Bank (PTB) (Marcus et al., 1993) data, which, after preprocessed, contains 929k training words, 73k validation and 82k test words. The vocabulary size is about 10k. The LSTM has 2 layers, each containing 200 hidden units. The word embedding has 200 dimensions which is trained from scratch. The batch size is 100. We vary the length of the backprop through time (BPTT) within the range $\{40, 400, 1000\}$. To prevent overfitting, we add a dropout regularization with rate 0.5 under all the settings.

The results are shown in Figure 3. The conclusions drawn from those figures are again similar to those in the last two experiments. However, the slightly different yet cheering observations is that the AdamNG is uniformly better than all the other competitors with any training sequence length. The superiority in terms of convergence speedup exists in both training and testing.

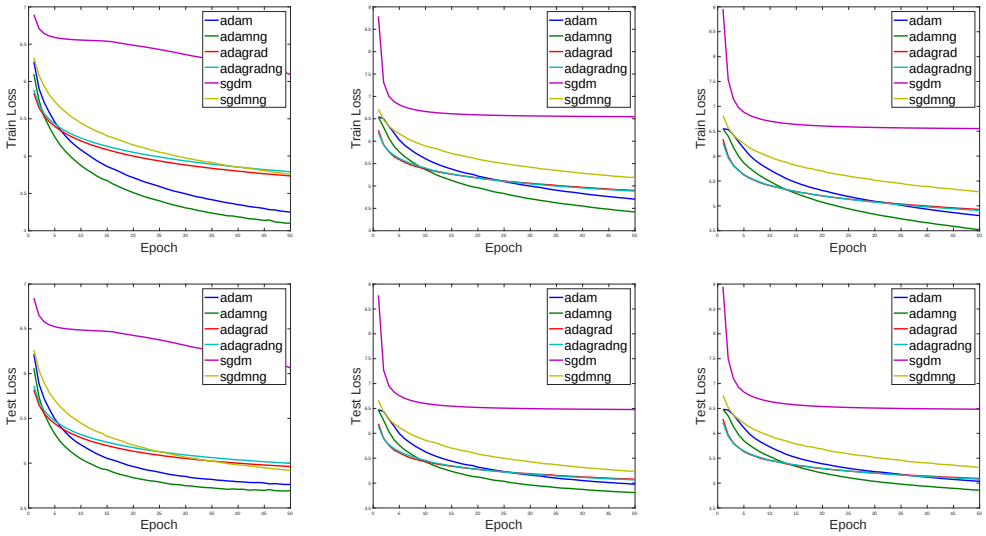

Figure 3: The training and testing objective curves on Penn Tree Bank dataset with LSTM recurrent neural networks. The first row is the training objective while the second is the testing. From left to right, the training sequence (BPTT) length are respectively 40, 400 and 1000. Dropout with 0.5 is imposed.

4.5 SENTIMENT ANALYSIS WITH CONVOLUTION NEURAL NETWORK

The task in this section is the sentiment analysis with convolution neural network. The dataset under use is Rotten Tomatoes[4] Pang & Lee (2005), a movie review dataset containing 10,662 documents, with half positive and half negative. We randomly select around 90% for training and 10% for validation. The model is a single layer convolution neural network that follows the setup of (Kim, 2014). The word embedding under use is randomly initialized and of 128-dimension.

For each algorithm, we run 150 epochs on the training data, and report the best validation accuracy in Table 4. The messages conveyed by the table is three-fold. Firstly, the algorithms using normalized gradient achieve much better validation accuracy than their unnormalized versions. Secondly, those with adaptive stepsize always obtain better accuracy than those without. This is easily seen by the comparison between Adam and SGDM. The last point is the direct conclusion from the previous two that the algorithm using normalized gradient with adaptive step sizes, namely AdamNG, outperforms all the remaining competitors.

Algorithm	AdamNG	Adam	AdaGradNG	AdaGrad	SGDMNG	SGDM
Validation Accuracy	**77.11%**	74.02%	71.95%	69.89%	71.95%	64.35%

Table 4: The Best validation accuracy achieved by the different algorithms.

5 CONCLUSION

In this paper, we propose a generic algorithm framework for first order optimization. It is particularly effective for addressing the vanishing and exploding gradient challenge in training with non-convex loss functions, such as in the context of convolutional and recurrent neural networks. Our method is based on normalizing the gradient to establish the descending direction regardless of its magnitude, and then separately estimating the ideal step size adaptively or constantly. This method is quite general and may be applied to different types of networks and various architectures. Although the primary application of the algorithm is deep neural network training, we provide a convergence for the new method under the convex setting in the appendix.

Empirically, the proposed method exhibits very promising performance in training different types of networks (convolutional, recurrent) across multiple well-known data sets (image classification, natural language processing, sentiment analysis, etc.). In general, the positive performance differential compared to the state of the art is most striking for very deep networks, as shown in our comprehensive experimental study.

REFERENCES

Yoshua Bengio. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

John Duchi, Michael I Jordan, and Brendan McMahan. Estimation, optimization, and parallelism when data is sparse. In *NIPS*, pp. 2832–2840, 2013.

John C. Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: convergence analysis and network scaling. *Automatic Control, IEEE Transactions on*, 57 (3):592–606, 2012.

Elad Hazan, Kfir Y. Levy, and Shai Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. In *NIPS*, pp. 1594–1602, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015.

[4] `http://www.cs.cornell.edu/people/pabo/movie-review-data/`

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pp. 448–456, 2015.

Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL `http://arxiv.org/abs/1412.6980`.

Krzysztof C Kiwiel. Convergence and efficiency of subgradient methods for quasiconvex minimization. *Mathematical programming*, 90(1):1–25, 2001.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Yann Lecun, Leon Bottou, Yoshua Bengio, and Patrick Haffner? Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pp. 2278–2324, 1998.

Kfir Y. Levy. The power of normalization: Faster evasion of saddle points. *CoRR*, abs/1611.04831, 2016. URL `http://arxiv.org/abs/1611.04831`.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Nesterov. Minimization methods for nonsmooth convex and quasiconvex functions. *Matekon*, 29: 519–531, 1984.

Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-sgd: Path-normalized optimization in deep neural networks. In *NIPS*, pp. 2422–2430, 2015.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pp. 115–124. Association for Computational Linguistics, 2005.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *ICML*, pp. 1310–1318, 2013.

Tim Salimans and Diederik P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *NIPS*, 2016.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. URL `http://arxiv.org/abs/1409.1556`.

APPENDIX

As a concrete example for Algorithm 1, we present a simple modification of AdaGrad using block-wise normalized stochastic gradient in Algorithm 2, where $g_{1:t}$ is a matrix created by stacking g_1, g_2,... and g_t in columns and $g_{1:t,j} \in \mathbb{R}^t$ represents the jth row of $g_{1:t}$ for $j = 1, 2, \ldots, d$. Assuming the function F is convex over x for any ξ and following the analysis in (Duchi et al., 2011), it is straightforward to show a $O(\frac{1}{\sqrt{T}})$ convergence rate of this modification. In the following, we denote $\|x\|_W := \sqrt{x^\top W x}$ as the the Mahalanobis norm associated to a $d \times d$ positive definite matrix W. The convergence property of Block-Normalized AdaGrad is presented in the following theorem.

Algorithm 2 AdaGrad with Block-Normalized Gradient (AdaGradBNG)

1: Choose $x_1 \in \mathbb{R}^d$, $\delta > 0$ and $\eta > 0$.
2: **for** $t = 1, 2, ...,$ **do**
3: Sample a mini-batch data ξ_t and compute the stochastic partial gradient $g_t^i = \frac{F_i'(x_t,\xi_t)}{\|F_i'(x_t,\xi_t)\|_2}$
4: Let $g_t = (g_t^1, g_t^2, \ldots, g_t^B)$, $g_{1:t} = [g_1, g_2, \ldots, g_t]$ and $s_t = (\|g_{1:t,1}\|_2, \|g_{1:t,2}\|_2, \ldots, \|g_{1:t,d}\|_2)$
5: Partition $s_t = (s_t^1, s_t^2, \ldots, s_t^B)$ in the same way as g_t.
6: Let $\tau_t = (\tau_t^1, \tau_t^2, \ldots, \tau_t^B)$ with[5] $\tau_t^i = \eta \|F_i'(x_t, \xi_t)\|_2 (\delta \mathbf{1}_{d_i} + s_t^i)^{-1}$.
7: $x_{t+1} = x_t - \tau_t \circ g_t$
8: **end for**

Theorem 1 *Suppose F is convex over x, $\|F'(x_t, \xi_t)\|_2 \le M$ and $\|x_t - x^*\|_\infty \le D_\infty$ for all t for some constants M and $D_\infty > 0$ in Algorithm 2. Let $H_t = \delta I_d + diag(s_t)$ and $H_t^i = \delta I_{d_i} + diag(s_t^i)$ for $t = 1, 2, \ldots$ and $\bar{x}_T := \frac{1}{T}\sum_{t=1}^T x_t$. Algorithm 2 guarantees*

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] \quad \le \quad \frac{\|x_1 - x^*\|_{H_1}^2}{2\eta T} + \frac{D_\infty^2 \sqrt{Bd}}{2\eta\sqrt{T}} + \sum_{i=1}^B \frac{\eta M^2 \sqrt{d_i}}{\sqrt{T}}.$$

Proof: It is easy to see that H_t is a positive definite and diagonal matrix. According to the updating scheme of x_{t+1} and the definitions of H_t, τ_t and g_t, we have

$$\begin{aligned}
\|x_{t+1} - x^*\|_{H_t}^2 &= (x_t - \tau_t \circ g_t - x^*)^\top H_t (x_t - \tau_t \circ g_t - x^*) \\
&= \|x_t - x^*\|_{H_t}^2 - 2(x_t - x^*)^\top H_t (\tau_t \circ g_t) + \|\tau_t \circ g_t\|_{H_t}^2 \\
&\le \|x_t - x^*\|_{H_t}^2 - 2\eta (x_t - x^*)^\top F'(x_t, \xi_t) + \|\tau_t \circ g_t\|_{H_t}^2.
\end{aligned}$$

The inequality above and the convexity of $F(x, \xi)$ in x imply

$$\begin{aligned}
F(x_t, \xi_t) - F(x^*, \xi_t) &\le \frac{\|x_t - x^*\|_{H_t}^2}{2\eta} - \frac{\|x_{t+1} - x^*\|_{H_t}^2}{2\eta} + \frac{\|\tau_t \circ g_t\|_{H_t}^2}{2\eta} \\
&= \frac{\|x_t - x^*\|_{H_t}^2}{2\eta} - \frac{\|x_{t+1} - x^*\|_{H_t}^2}{2\eta} + \sum_{i=1}^B \frac{\eta \|F_i'(x_t, \xi_t)\|_2^2 \|g_t^i\|_{(H_t^i)^{-1}}^2}{2} \qquad (2)
\end{aligned}$$

Taking expectation over ξ_t for $t = 1, 2, \ldots$ and averaging the above inequality give

$$\begin{aligned}
&\mathbb{E}[f(\bar{x}_T) - f(x^*)] \\
\le\ & \frac{1}{T}\sum_{t=1}^T \left[\frac{\mathbb{E}\|x_t - x^*\|_{H_t}^2}{2\eta} - \frac{\mathbb{E}\|x_{t+1} - x^*\|_{H_t}^2}{2\eta} \right] + \sum_{t=1}^T \sum_{i=1}^B \frac{\eta \mathbb{E}\left[\|F_i'(x_t, \xi_t)\|_2^2 \|g_t^i\|_{(H_t^i)^{-1}}^2 \right]}{2T} \\
\le\ & \frac{1}{T}\sum_{t=1}^T \left[\frac{\mathbb{E}\|x_t - x^*\|_{H_t}^2}{2\eta} - \frac{\mathbb{E}\|x_{t+1} - x^*\|_{H_t}^2}{2\eta} \right] + \sum_{t=1}^T \sum_{i=1}^B \frac{\eta M^2 \mathbb{E}\left[\|g_t^i\|_{(H_t^i)^{-1}}^2 \right]}{2T} \qquad (3)
\end{aligned}$$

where we use the fact that $\|F'(x_t, \xi_t)\|_2^2 \le M^2$ in the second inequality.

According to the equation (24) in the proof of Lemma 4 in (Duchi et al., 2011), we have

$$\sum_{t=1}^{T} \|g_t^i\|^2_{(H_t^i)^{-1}} = \sum_{t=1}^{T} \sum_{j=d_1+d_2+\cdots+d_{i-1}+1}^{d_1+d_2+\cdots+d_i} \frac{g_{t,j}^2}{\delta + \|g_{1:t,j}\|_2} \leq \sum_{j=d_1+d_2+\cdots+d_{i-1}+1}^{d_1+d_2+\cdots+d_i} 2\|g_{1:T,j}\|_2 \leq 2\sqrt{Td_i} \quad (4)$$

where the second inequality is due to Cauchy-Schwarz inequality and the fact that $\|g_t^i\|_2 = 1$. Following the analysis in the proof of Theorem 5 in (Duchi et al., 2011), we show that

$$\begin{aligned} \|x_{t+1} - x^*\|^2_{H_{t+1}} - \|x_{t+1} - x^*\|^2_{H_t} &= \langle x^* - x_{t+1}, \text{diag}(s_{t+1} - s_t)(x^* - x_{t+1})\rangle \\ &\leq D_\infty^2 \|s_{t+1} - s_t\|_1 = D_\infty^2 \langle s_{t+1} - s_t, \mathbf{1}\rangle . \end{aligned} \quad (5)$$

After applying (4) and (5) to (3) and reorganizing terms, we have

$$\begin{aligned} \mathbb{E}[f(\bar{x}_T) - f(x^*)] &\leq \frac{\|x_1 - x^*\|^2_{H_1}}{2\eta T} + \frac{D_\infty^2 \mathbb{E}\langle s_T, \mathbf{1}\rangle}{2\eta T} + \sum_{i=1}^{B} \frac{\eta M^2 \sqrt{d_i}}{\sqrt{T}} \\ &\leq \frac{\|x_1 - x^*\|^2_{H_1}}{2\eta T} + \frac{D_\infty^2 \sqrt{TBd}}{2\eta T} + \sum_{i=1}^{B} \frac{\eta M^2 \sqrt{d_i}}{\sqrt{T}} \\ &= \frac{\|x_1 - x^*\|^2_{H_1}}{2\eta T} + \frac{D_\infty^2 \sqrt{Bd}}{2\eta \sqrt{T}} + \sum_{i=1}^{B} \frac{\eta M^2 \sqrt{d_i}}{\sqrt{T}}, \end{aligned}$$

where the second inequality is because $\langle s_T, \mathbf{1}\rangle = \sum_{j=1}^{d} \|g_{1:T,j}\|_2 \leq \sqrt{TBd}$ which holds due to Cauchy-Schwarz inequality and the fact that $\|g_t^i\|_2 = 1$. ■