SCW-SGD: STOCHASTICALLY CONFIDENCE-WEIGHTED SGD

Takumi Kobayashi

National Institute of Advanced Industrial Science and Technology
1-1-1 Umezono, Tsukuba, Ibaraki, Japan

ABSTRACT

As the field of deep learning has been rapidly growing, the optimization methods (optimizers) gain keen attention for efficiently training neural networks. While SGD exhibits practically favorable performance on various tasks, adaptive methods, such as ADAM, are also formulated to equip the gradient-based updating with adaptive scaling in a sophisticated way. In this paper, we propose a novel optimizer to integrate those two approaches of the adaptive method and SGD through assigning stochastic confidence weights to the gradient-based updating. We define statistical uncertainty of the gradients which is implicitly embedded in the adaptive scaling of ADAM, and then based on the uncertainty, naturally incorporate stochasticity into the optimizer as a bridge between SGD and ADAM. Thereby, the proposed optimizer, SCWSGD, endows the parameter updating with two types of stochasticity regarding multiplicative scaling for the gradient and mini-batch sampling to compute the gradient, for improving generalization performance. In the experiments on image classification using various CNNs, the proposed optimizer produces favorable performance in comparison to the other optimizers.

***Index Terms*—** Neural Network, Optimization, SGD, Stochastic weighting

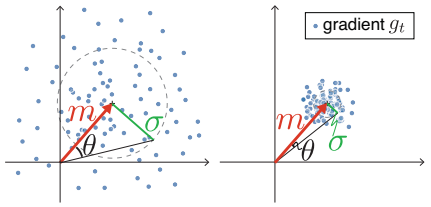
1. INTRODUCTION

Deep neural networks have made great progress in this decade with promising performance on various fields including image recognition and signal processing [1, 2]. While various architectural improvements are applied to effectively train the networks [3, 4], the optimization methods (*optimizers*) have also attracted keen attention for addressing the issues to train the large network models on large-scale training data. The optimizer of stochastic gradient descent (SGD) [5, 6] is arguably one of the most successful approaches to optimize the networks. It is simply formulated by leveraging mini-batches to update network parameters, while exhibiting intriguing property regarding optimization [7, 8]. The simplicity of SGD induces the more sophisticated optimizers which adaptively scale the gradients, such as RMSPROP [9], ADADELTA [10] and ADAM [11]. Those adaptive methods are advantageous in terms of rapid training time and automatically adjusting the learning rate. Some recent works [12, 13], however, pointed out that the generalization and out-of-sample performance of the adaptive methods are not fully understood and actually inferior to that of the simple SGD on some practical tasks. To mitigate the problem, in [13], the optimization process of ADAM is theoretically analyzed to propose a novel optimizer of AMSGRAD by manipulating ADAM process in a simple yet effective way. While the AMSGRAD is endowed with favorable theoretical property regarding convergence, it can be empirically found that there is still a performance gap between SGD and AMSGRAD.

In the another line of research for improving the adaptive methods, there are some works to integrate both approaches of SGD and the adaptive method, especially ADAM. The adaptive method consists of two factors of the gradient-based updating and its scaling. ADABOUND [14] imposes the lower and upper bounds on the scaling factor to make the optimizer behave similarly to SGD at the later training epochs and successfully boosts the performance of ADAM. In M-SVAG [15], the adaptive scaling factor is applied to the SGD-based update through carefully analyzing the variance of the gradients.

In this paper, we propose a novel optimizer by integrating SGD and ADAM through further embedding stochastic characteristics into the optimizer. From the statistical viewpoint, we define the uncertainty of gradients which is implicitly embedded in the scaling factor of ADAM and fundamentally controls the adaptivity. The uncertainty is naturally compatible with stochasticity and thus they are combined to construct the stochastic confidence weight for effectively scaling SGD-based update (gradient momentum), connecting SGD and ADAM in a stochastic manner, to establish our optimizer of stochastically confidence-weighted SGD (SCWSGD). The SCWSGD is endowed with *two* types of stochasticity regarding the mini-batch sampling of SGD and the confidence weighting based on the uncertainty.

The proposed method is related to M-SVAG [15] which applies the adaptive scaling derived from ADAM formulation to the gradient momentum in a *deterministic* way. Our method, however, exhibits clear difference from M-SVAG in that the gradient momentum is scaled in a *stochastic* manner according to its statistical uncertainty through the confidence weighting for rendering stochastic perturbation toward robust



(a) Larger σ (b) smaller σ

Fig. 1. Uncertainty $\theta = \arctan\frac{\sigma}{|m|}$. It reflects the reliability of the estimate m.

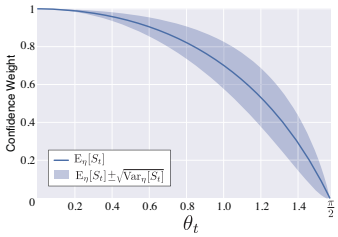


Fig. 2. Mean and deviation of the stochastic confidence weights S_t with $U = 1$ in (9).

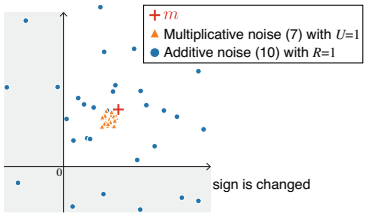


Fig. 3. Perturbation by two types of noise injected into 2-D gradients (m_t) in case of $|\hat{m}_t| = \hat{\sigma}_t$.

parameter updating [7]. While it is also possible to realize the perturbation in an *additive* form such as by injecting Gaussian noise [16] into gradients, the proposed method embeds *multiplicative* perturbation such that the gradient (momentum) sign is not changed; the sign direction of the gradient conveys critical information to update parameters [11, 15].

2. MINI-BATCH BASED OPTIMIZATION

The neural network, e.g., CNN, equipped with parameters $\Psi \in \mathbb{R}^D$ is generally trained in an iterative manner;

$$\psi_{t+1} = \psi_t - \alpha\Delta_t, \tag{1}$$

which updates one parameter $\psi \in \Psi$ at the t-th iteration with a learning rate α. Since it is computationally inefficient to compute the full gradient over the whole training samples, SGD [5] leverages mini-batch to produce Δ_t efficiently;

$$\Delta_t = \frac{1}{|\mathcal{B}_t|}\sum_{i\in\mathcal{B}_t}\frac{\partial}{\partial\psi}l(x_i;\Psi_t) \triangleq g_t, \tag{2}$$

where $\mathcal{B}$ is the mini-batch (index) set, x_i is the i-th training sample and the loss function is denoted by l. In the SGD (2), the gradient g_t is *stochastically* given through sampling the mini-batch $\mathcal{B}_t$ [7] from the whole training set. To stabilize the update Δ_t, the momentum [6] can also be widely applied by

$$\Delta_t = m_t, \tag{3}$$

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)g_t = (1-\beta_1^t)\sum_{\tau=0}^{t-1}\hat{\beta}_\tau g_{t-\tau}, \tag{4}$$

where the momentum is computed by means of an exponential moving average (EMA)[1] for g_t, and $m_0 = 0$, $\hat{\beta}_\tau = \frac{1-\beta_1}{1-\beta_1^t}\beta_1^\tau$, and $\sum_{\tau=0}^{t-1}\hat{\beta}_\tau = 1$. ADAM [11], the adaptive optimizer, further exploits the second order statistics of g_t in the form of

$$\Delta_t = \frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon},\ v_t = \beta_2 v_{t-1} + (1-\beta_2)g_t^2, \tag{5}$$

where ϵ is a small constant, say $\epsilon = 10^{-8}$, and the normalized statistics are given by $\hat{m}_t = \frac{m_t}{1-\beta_1^t}$ and $\hat{v}_t = \frac{v_t}{1-\beta_2^t}$.

[1] In (4), our momentum multiplies the gradient g_t by $1-\beta_1$, which affects the learning rate α; in our setting with $\beta_1 = 0.9$, $\alpha = 1$ corresponds to the learning rate of 0.1 in the standard SGD.

3. SCW-SGD

We first reformulate the ADAM updating (5) as in [15] to define the statistic uncertainty of g_t, and then propose stochastic confidence weight to scale the gradient momentum (4).

3.1. Uncertainty of Gradients

The ADAM updating (5) is rewritten by

$$\frac{\hat{m}_t(\beta_1)}{\sqrt{\hat{v}_t(\beta_2)}} = \frac{\hat{m}_t(\beta_1)}{\sqrt{\hat{m}_t^2(\beta_2)+\hat{\sigma}_t^2(\beta_2)}} = \frac{1}{\sqrt{1+\tan^2\theta_t(\beta_2)}}\frac{\hat{m}_t(\beta_1)}{|\hat{m}_t(\beta_2)|},$$
$$\hat{\sigma}_t^2(\beta_2) = \hat{v}_t(\beta_2) - \hat{m}_t^2(\beta_2),\ \tan\theta_t(\beta_2) = \frac{\hat{\sigma}_t(\beta_2)}{|\hat{m}_t(\beta_2)|}, \tag{6}$$

where we omit ϵ and explicitly show the dependency on β_1 and β_2, and $\hat{\sigma}_t$ indicates the standard deviation of gradient g_t. In case of $\beta_1 = \beta_2$, the ADAM optimizer is reduced to the sign of gradient moment [15], $\frac{\hat{m}_t(\beta_1)}{|\hat{m}_t(\beta_2)|} = \mathrm{sign}(\hat{m}_t) \in \{-1, +1\}$ which is scaled by $\frac{1}{\sqrt{1+\tan^2\theta_t}} = \cos\theta_t$.

We define the *uncertainty* of gradients by θ_t in (6) which statistically measures the confidence of the gradient mean $\hat{m}_t$ estimated via EMA (4) in a robust manner to the scale of gradients g_t; it is connected to coefficient of variance [17] and signal-to-noise-ratio. As shown in Fig. 1, the smaller θ indicates that gradients g_t are concentrated around the mean $\hat{m}_t$, therefore suggesting that the update $\hat{m}_t$ surely contributes to optimizing the parameter ϕ by effectively decreasing the loss. In contrast, the mean $\hat{m}_t$ of the larger θ is not so reliable as to be used for updating the parameter ϕ. Therefore, the measure $\theta_{t(\beta_2)}$ indicates the uncertainty of $\hat{m}_{t(\beta_2)}$, the mean of gradients $\{g_\tau\}_{\tau=1}^t$ via EMA with the rate β_2; in this study, we consider $\theta_{t(\beta_1)}$ to measure the uncertainty of the momentum $\hat{m}_{t(\beta_1)}$ used for the parameter updating in (3) and thus drop the subscript (β_1) for brevity in what follows.

3.2. Stochastic Confidence Weighting

The uncertainty θ_t contributes to ADAM [11] in the form of $\cos\theta_t$ (6) for scaling the update $\frac{\hat{m}_t(\beta_1)}{|\hat{m}_t(\beta_2)|}$ which is close to $\mathrm{sign}[m_{t(\beta_1)}]$, while in M-SVAG [15] it applies to the weighting via $\cos^2\theta_t$ through the sophisticated variance estimation

of $\hat{\sigma}_t$. In contrast to those *deterministic* weighting, we propose *stochastic* weighting based on the uncertainty θ_t which adaptively scales the momentum m_t, leading to the method of stochastically confidence-weighted SGD (SCWSGD).

The perturbation to the gradient is effective for robust training [7, 16], and the stochastic perturbation is related to the uncertainty θ; it would be natural to introduce the larger perturbation to the less-confident m_t of the larger uncertainty θ_t. We thus formulate SCWSGD by

$$\Delta_t = \frac{m_t}{\sqrt{1+\eta\tan^2\theta_t}+\epsilon'} = \frac{|\hat{m}_t| m_t}{\sqrt{(1-\eta)\hat{m}_t^2+\eta\hat{v}_t}+\epsilon}, \quad (7)$$

$$\text{where} \quad \eta \sim \mathcal{U}[0,U], \quad (8)$$

and ϵ is a small constant as in (5); say $\epsilon = 10^{-8}$. The random noise η is uniformly drawn from $[0,U]$ in which U indicates the upper bound of the uniform distribution. Note that the statistics m_t and v_t are computed by EMA with the identical rate β_1 in (4). In SCWSGD (7), the stochastic confidence weight S_t scales m_t, being statistically characterized by[2]

$$S_t = \frac{1}{\sqrt{1+\eta\tan^2\theta_t}}, \quad \mathrm{E}_\eta[S_t] = \frac{2}{\sqrt{1+U\tan^2\theta_t}+1}, \quad (9)$$

$$\mathrm{Var}_\eta[S_t] = \frac{\log(1+U\tan^2\theta_t)}{U\tan^2\theta_t} - \frac{4}{(\sqrt{1+U\tan^2\theta_t}+1)^2},$$

which are depicted in Fig. 2. The (moderately) larger uncertainty θ_t increases the variance of the perturbation, while the confident update of low uncertainty ($\theta_t \approx 0$) works in a manner similar to the deterministic momentum-SGD (3). For the extremely large uncertainty $\theta_t \approx \frac{\pi}{2}$, the update is simply suppressed by the small weight $S_t \approx 0$ with low perturbation.

In contrast to the additive Gaussian noise [16], the proposed SCWSGD introduces the uniform noise η into the weighting to produce *multiplicative* perturbation which is advantageous in the following two points: 1) The sign of the updating direction m_t is unchanged in (7), $\mathrm{sign}(\Delta_t) = \mathrm{sign}(m_t)$, and 2) the extreme cases of $\eta = 0$ and $\eta = U$ lead to the momentum-SGD (3) and the one close to ADAM (5), respectively. As shown in [11, 15], the sign of gradient momentum m_t is an important clue for updating the parameter ϕ. As shown in Fig. 3, the additive noise might change the sign by the large perturbation, while the multiplicative perturbation surely maintain it. And, the proposed method stochastically combines the two approaches of SGD and ADAM-based adaptive method through the stochastic confidence weight (9). Note that SCWSGD endows the parameter updating with two types of stochasticity; one is derived from mini-batch sampling and the other is parametrically given by η in (9).

4. EXPERIMENTAL RESULTS

We apply the proposed optimizer SCWSGD to train various CNNs on image classification tasks [18]; similarly to the other optimizers, the proposed method is applicable to optimize various parameterized models including neural networks.

[2]For simplicity, we omit ϵ.

4.1. Ablation Study

We evaluate SCWSGD from various aspect through training the CNN of ResNet-34 [2] on Cifar-100 dataset [18]. According to the standard practice, the CNN is trained by the optimizer with a batch size of 128, weight decay of 0.0005, and the initial learning rate of $\alpha = 1$ which is then divided by 10 at the 150th epoch over 200 training epochs. We set the hyper-parameters of SCWSGD to $\beta_1 = 0.9$ and $U = 0.5$ which are analyzed in the following experiments. We evaluate the classification performance on the test split provided in Cifar-100 dataset and repeat the evaluation three times with different initial random seeds for CNN parameters to report the average and the standard deviation of error rates (%).

Stochastic vs deterministic. In SCW-SGD, the stochasticity is embedded by means of the uniform random noise η in (7) drawn from the uniform distribution $\mathcal{U}[0,U]$, and the stochasticity can be controlled by the upper bound $U \in \{0.25, 0.5, 1\}$. Table 1a shows a comparison with the deterministic optimizer that fixes $\eta \in \{0,1\}$ during the training. The stochastic method outperforms the deterministic ones both of $\eta = 0$ and $\eta = 1$, validating the effectiveness of the stochasticity by which the the proposed method combines those two deterministic optimizers of $\eta = 0$ and $\eta = U$; the method of $U = 0.5$ produces the better performance.

Degree of stochasticity. As shown in (7), the uniform random noise η is sampled at each parameter $\psi \in \Psi$ in an *i.i.d.* manner. In the alternative approach, the noise η is *shared* across parameters at each layer to exhibit coherent (or correlated) stochasticity; namely, we sample single random number η for each convolution filter, bias and so on. Those two types of stochasticity are empirically compared in Table 1b. The *i.i.d.* sampling of η is superior to the coherent one due to the high stochasticity producing uncorrelated randomness for respective parameter updating toward robust training.

Uncertainty. The uncertainty measure θ_t constructs the confidence weighting in (7, 9) which is applied to scale the momentum m_t, and those θ_t and m_t are built upon the same statistic process via EMA (4); that is, $\beta_1 = \beta_2 = 0.9$ for the momentum $m_{t(\beta_1)}$ and the uncertainty $\theta_{t(\beta_2)}$. On the other hand, it could be possible to apply the different hyper-parameters, $\beta_1 \neq \beta_2$, as in ADAM [11]. In Table 1c, we test those hyper-parameter settings on the two approaches of stochastic ($\eta \sim \mathcal{U}[0,0.5]$) and deterministic ($\eta = 1$) ones. The performance comparison demonstrates the importance of computing the uncertainty θ_t and the momentum m_t in the identical statistical process (EMA) for improving performance; the ones produced by inconsistent process of $\beta_1 \neq \beta_2$ significantly degrade performance. The importance of the statistical consistency also indicates that the performance improvement by SCWSGD comes from not only the stochasticity

Table 1. Ablation study on Cifar-100 [18] by ResNet-34 [2]. The performance is measured by classification error rate (%).

(a) Stochastic vs deterministic

Method	Error (%)
Deterministic $\eta = 1$	22.35±0.11
Deterministic $\eta = 0$ (SGD)	23.00±0.42
Stochastic $\eta \sim \mathcal{U}[0, 1]$	21.83±0.24
Stochastic $\eta \sim \mathcal{U}[0, 0.5]$	**21.40**±0.26
Stochastic $\eta \sim \mathcal{U}[0, 0.25]$	21.84±0.05

(b) Degree of Stochasticity

Sampling η	Error (%)
i.i.d.	**21.40**±0.26
coherently	21.75±0.05

(c) Uncertainty $\theta_{t(\beta_2)}$

β_1	β_2	η	Error (%)
0.9	0.9	$\mathcal{U}[0, 0.5]$	**21.40**±0.26
0.9	0.999	$\mathcal{U}[0, 0.5]$	25.63±0.40
0.9	0.999	1	28.41±0.76

(d) Perturbation type

Method	Error (%)
Multiplicative (7)	**21.40**±0.26
Additive (10) $R = 1$	25.13±0.22
Additive (10) $R = 0.1$	22.73±0.16

Table 2. Performance comparison on various CNNs. The performance is measured by classification error rate (%).

	Cifar-10			Cifar-100		
Optimizer	ResNet34 [2]	DenseNet121 [19]	WRN28-10 [20]	ResNet34 [2]	DenseNet121 [19]	WRN28-10 [20]
SGD [6]	5.45±0.14	5.51±0.07	4.52±0.08	23.00±0.42	21.25±0.13	19.51±0.11
ADAM [11]	6.73±0.22	6.74±0.15	7.96±0.09	27.03±0.35	25.70±0.24	27.55±0.57
AMSGRAD [13]	6.28±0.04	6.41±0.07	7.70±0.22	26.70±0.48	25.72±0.16	27.97±0.32
ADABOUND [14]	5.27±0.05	5.13±0.15	4.33±0.12	23.64±0.30	22.96±0.37	20.86±0.24
M-SVAG [15]	4.90±0.10	**4.52**±0.19	4.15±0.06	22.57±0.24	20.49±0.16	20.45±0.21
SCWSGD	**4.69**±0.14	4.55±0.13	**3.97**±0.10	**21.40**±0.26	**19.99**±0.14	**19.41**±0.15

via η but also the uncertainty measure θ_t of m_t; the uncertainty should properly reflect the statistical characteristics of the updating direction m_t.

Perturbation The proposed stochastic form can be compared with the Gaussian-based stochastic form [16] of

$$\Delta_t = m_t + \eta', \ \eta' \sim \mathcal{N}(0, R\hat{\sigma}_t), \tag{10}$$

where $\hat{\sigma}_t$ is the standard deviation computed in (6) and R controls the degree of perturbation. This stochastic optimizer (10) *adds* Gaussian noise η' with the scale $R\hat{\sigma}_t$ in a manner similar to [16], while the proposed SCWSGD considers the multiplicative perturbation, as shown in Fig. 3. Table 1d shows the performance comparison, demonstrating the effectiveness of the multiplicative perturbation in SCWSGD. As described in Sec. 3.2 and Fig. 3, the *additive* noise could change the updating direction, $\mathrm{sign}(\Delta_t) \neq \mathrm{sign}(m_t)$, by some large deviation due to $R\hat{\sigma}_t$, and it is hard to tune the noise scale (R) for gradients of various scales. On the other hand, the proposed multiplicative form (7) effectively scales m_t according to the statistical confidence (uncertainty) of the gradients without affecting their signs.

4.2. Comparison to Other Optimizers

Finally, we compare the proposed SCWSGD with the other optimizers; momentum-SGD [6], ADAM [11], AMSGRAD [13], ADABOUND [14] and M-SVAG [15]. We follow the standard practice to determine the hyper-parameters in those optimizers; in SGD and M-SVAG, momentum of $\beta_1 = 0.9$, and in ADAM and AMSGRAD, $\beta_1 = 0.9, \beta_2 = 0.999$, and in ADABOUND, $\beta_1 = 0.9, \beta_2 = 0.999, \gamma = 0.001$; for detail of the hyper-parameters, refer to the respective papers. The initial learning rate is 1 for {SGD, M-SVAG}[3] and 0.001 for {ADAM, AMSGRAD, ADABOUND}, and then the learning rate is divided by 10 at the 150th epoch as in our setting for fair comparison. We apply these optimizers to train the deep CNNs of ResNet34 [2], DenseNet121 [19] and wide-ResNet (WRN28-10) [20] on Cifar-10/100 datasets [18]. The performance results are shown in Table 2. The proposed SCWSGD produces favorably competitive performance to the others, outperforming them on most cases. From the viewpoint of computation cost, SCWSGD is also comparable to the adaptive methods such as ADAM with a negligible extra cost for sampling η since the method is based on just two moving statistics of m_t and v_t.

5. CONCLUSION

In this paper, we have proposed a new optimizer of stochastically confidence-weighted SGD (SCWSGD) for training neural networks. The proposed method focuses on the statistical uncertainty of the gradient mean (momentum) estimated via exponential moving average over mini-batches. Based on the uncertainty, we formulate a stochastic weighting scheme for the gradients to provide the gradient-based parameter updating with multiplicative perturbation for improving generalization performance. In the experiments on image classification tasks using various CNNs, the proposed method exhibits favorable performance in comparison with the other optimizers including SGD and ADAM.

[3] As described in Sec. 2, the momentum is defined as (4) multiplying g_t by $1 - \beta_1$, and thus it should be noted that the learning rate $\alpha = 1$ with $\beta_1 = 0.9$ corresponds to the learning rate of 0.1 in the standard SGD setting.

6. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012, pp. 1097–1105.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[3] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout : A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[4] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *Journal of Machine Learning Research*, vol. 37, pp. 448–456, 2015.

[5] Herbert Robbins and Sutton Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.

[6] Boris T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.

[7] Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma, "The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects," in *ICML*, 2019.

[8] Chiyuan Zhang, Qianli Liao, Alexander Rakhlin, Brando Miranda, Noah Golowich, and Tomaso Poggio, "Theory of deep learning III: Generalization properties of sgd," CBMM Memo 67, 2017.

[9] Tijmen Tieleman and Geoffrey E. Hinton, "Rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.

[10] Matthew D. Zeiler, "Adadelta: An adaptive learning rate method," *CoRR, abs/1212.5701*, 2012.

[11] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[12] Ashia C. Wilson, Rebecca Roelofs, Mitchell Stern, Nathan Srebro, and Benjamin Recht, "The marginal value of adaptive gradient methods in machine learning," in *NeurIPS*, 2017.

[13] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar, "On the convergence of adam and beyond," in *ICLR*, 2018.

[14] Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun, "Adaptive gradient methods with dynamic bound of learning rate," in *ICLR*, 2019.

[15] Lukas Balles and Philipp Hennig, "Dissecting adam: The sign, magnitude and variance of stochastic gradients," in *ICML*, 2018.

[16] Arvind Neelakantan, Luke Vilnis, Quoc V. Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens, "Adding gradient noise improves learning for very deep networks," *arXiv:1511.06807*, 2015.

[17] Charles E. Brown, *Applied Multivariate Statistics in Geohydrology and Related Sciences*, Springer, 1998.

[18] Alex Krizhevsky and Geoffrey E. Hinton, "Learning multiple layers of features from tiny images," Technical report, University of Toronto, 2009.

[19] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017, pp. 2261–2269.

[20] Sergey Zagoruyko and Nikos Komodakis, "Wide residual networks," in *BMVC*, 2016.