

diffGrad: An Optimization Method for Convolutional Neural Networks

Shiv Ram Dubey, *Member, IEEE*, Soumendu Chakraborty, Swalpa Kumar Roy, *Student Member, IEEE*, Snehasis Mukherjee, *Member, IEEE*, Satish Kumar Singh, *Senior Member, IEEE*, and Bidyut Baran Chaudhuri, *Life Fellow, IEEE*

Abstract—Stochastic Gradient Decent (SGD) is one of the core techniques behind the success of deep neural networks. The gradient provides information on the direction in which a function has the steepest rate of change. The main problem with basic SGD is to change by equal sized steps for all parameters, irrespective of gradient behavior. Hence, an efficient way of deep network optimization is to make adaptive step sizes for each parameter. Recently, several attempts have been made to improve gradient descent methods such as AdaGrad, AdaDelta, RMSProp and Adam. These methods rely on the square roots of exponential moving averages of squared past gradients. Thus, these methods do not take advantage of local change in gradients. In this paper, a novel optimizer is proposed based on the difference between the present and the immediate past gradient (i.e., diffGrad). In the proposed diffGrad optimization technique, the step size is adjusted for each parameter in such a way that it should have a larger step size for faster gradient changing parameters and a lower step size for lower gradient changing parameters. The convergence analysis is done using the regret bound approach of online learning framework. Rigorous analysis is made in this paper over three synthetic complex non-convex functions. The image categorization experiments are also conducted over the CIFAR10 and CIFAR100 datasets to observe the performance of diffGrad with respect to the state-of-the-art optimizers such as SGDM, AdaGrad, AdaDelta, RMSProp, AMSGrad, and Adam. The residual unit (ResNet) based Convolutional Neural Networks (CNN) architecture is used in the experiments. The experiments show that diffGrad outperforms other optimizers. Also, we show that diffGrad performs uniformly well for training CNN using different activation functions. The source code is made publicly available at <https://github.com/shivram1987/diffGrad>.

Index Terms—Neural Networks; Optimization; Gradient Descent; Difference of Gradient; Adam, Residual Network, Image Classification.

I. INTRODUCTION

During the last few years, deep learning based techniques have gained more and more popularity in solving problems in different domains, especially where a data driven approach

S.R. Dubey and S. Mukherjee are with the Computer Vision Group, Indian Institute of Information Technology, Sri City, Chittoor, Andhra Pradesh-517646, India (e-mail: shivram1987@gmail.com, srdubey@iiits.in, snehasis.mukherjee@iiits.in).

S. Chakraborty is with the Indian Institute of Information Technology, Lucknow, Uttar Pradesh, India (email: soum.uit@gmail.com).

S.K. Roy and B.B. Chaudhuri are with the Computer Vision and Pattern Recognition Unit at Indian Statistical Institute, Kolkata-700108, India (email: swalparoy@gmail.com, bidyutbaranchaudhuri@gmail.com). Prof. Chaudhuri is also with the Techno India University, Sector V, Salt Lake City, Kolkata-700091, India.

S.K. Singh is with the Computer Vision and Biometrics Laboratory at Indian Institute of Information Technology, Allahabad-211015, India (email: sk.singh@iiita.ac.in).

is required [1]. Due to the availability of GPU-based high-end computational facilities and the huge amount of data, deep learning based approaches generally outperform the traditional hand-designed approaches to solve research problems in Computer Vision [2], [3], [4], [5], Image Processing [6], [7], Signal Processing [8], [9], Robotics [10], Natural Language Processing [11], [12], and many other diverse areas of Artificial Intelligence. Other applications where deep learning can be used include object tracking [13], [14], [15], face anti-spoofing and micro-expression recognition [16], [17], hyper-spectral image classification [18], etc.

The deep neural network has different variants to deal with the different problems, such as Convolutional Neural Networks (CNN) and Generative Adversarial Networks (GAN) for images, Recurrent Neural Network (RNN) and Long Short Term Memory Network (LSTM) for temporal sequences of data, 3D-CNN for videos, etc. Research on CNN has observed a rapid growth in recent years, especially on various image based problems. Different CNN architectures have been proposed for image related problems such as AlexNet [2], VggNet [4], GoogLeNet [19], and ResNet [20] for image classification, R-CNN [21], Fast R-CNN [22], Faster R-CNN [23], and YOLO [24] for object detection, Mask R-CNN [25] and PANet [26] for instance segmentation, RCCNet [27] for colon cancer nuclei classification, etc.

In most neural networks, the basic approach that is usually followed for finding an optimal solution is Stochastic Gradient Descent (SGD) optimization [28]. Here, initially a measure of loss over the current parameter values is computed using a loss function defined for the specific problem being solved. Then, the gradient for each parameter (i.e. in each dimension) in the network is computed and the parameter values are updated in the opposite direction of the gradient by a factor proportional to the gradient. For SGD optimization, the above two steps are repeated until convergence or until a certain number of epochs or iterations are completed. There are the following four major drawbacks in the basic SGD approach: 1) If the loss changes quickly for a set of parameters and slowly for another set of parameters, then it leads to very slow learning along shallow dimensions and a jittering effect along steep dimensions [29] as depicted in Fig. 1a. 2) If the loss function has a local minimum or a saddle point [30], then SGD gets stuck due to zero gradients in that region. These situations are illustrated in Fig. 1b and Fig. 1c. 3) The gradients are usually computed over minibatches; so the gradients can be noisy. 4) SGD takes the same step for each parameter irrespective of

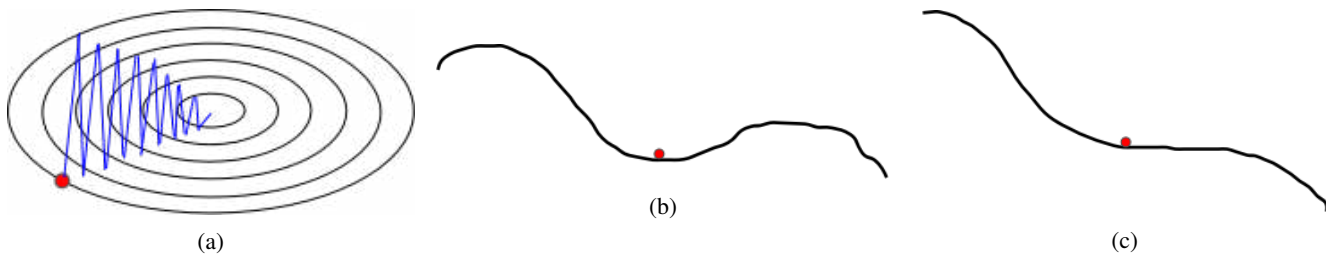


Fig. 1: (a) An optimization landscape for two parameters represented in two directions, *i.e.* horizontal and vertical directions. The loss changes slow in horizontal direction and fast in vertical direction. (b) A local minima scenario in 1-D optimization. (c) A saddle point scenario in 1-D optimization.

the iteration-wise gradient behavior for that parameter, which leads to poor optimization. Note that each parameter of the network is considered as a dimension in the optimization landscape. Therefore, both dimension and parameter are used interchangeably in this paper.

Improving SGD based optimization method for neural networks has recently become an active area of research interest. In order to address the above problems, several SGD variants have been proposed in the literature. For example, SGD with momentum (SGDM) is an extension of SGD by incorporating the past gradients in each dimension [31], [32]. SGDM maintains a momentum in each dimension as a function of previous momentum and current gradient in that dimension. The goal of SGDM is to develop high “velocity” in any dimension that has a consistent gradient. In SGDM, the jittering problem is reduced with high velocity in consistent gradient dimensions and the saddle point problem is reduced by using the past gradients, which provide some momentum, even when the current gradient is close to zero. SGDM optimization is further improved by Sutskever et al. by Nesterovs Accelerated Gradient (NAG), which guarantees a better convergence rate than SGD in certain situations [32]. In NAG, the gradients are computed based on an approximation of the next position of the parameters, which are used to update the current moments and parameters. Thus, NAG updates both the moments and gradients, based on the future dimensions.

Another widely used variant of the gradient descent method is AdaGrad [33], proposed to deal with sparse data. AdaGrad performs larger updates for infrequent parameters which leads to smaller magnitude of gradients and smaller updates for frequent parameters which leads to larger magnitude of gradients. Basically, AdaGrad divides the learning rate with the square root of the sum of the squares of the past gradients for all parameters. Apart from dealing with sparse data, AdaGrad is applied to other kinds of problems as well, such as training large-scale neural nets to recognize cats in Youtube videos [34] and training GloVe word embeddings, as infrequent words require larger updates than frequent ones [35]. But, AdaGrad accumulates the square of gradients which, in turn, may lower the learning rate drastically after some time and kill the learning process. Zeiler extended AdaGrad to AdaDelta by removing the problem of a dying learning rate, which was caused by the monotonically increasing sum of square of gradients [36]. AdaDelta also accumulates the square of past gradients, but considers only a few immediate past gradients instead of all past gradients. RMSProp is another attempt to correct the

diminishing learning rate of AdaGrad similar to AdaDelta [37], *i.e.* by accumulating the gradient as an exponentially decaying average of squared gradients. The major difference between AdaDelta and RMSProp is that AdaDelta does not use any learning rate [36], whereas RMSProp uses a learning rate [37].

One of the recent and popular variants of gradient descent is Adaptive Moment Estimation (Adam). Adam computes adaptive learning rates for each parameter [38] by utilizing both first and second moments. Adam accumulates the exponentially decaying average of past gradients similar to SGDM as first moment. It also accumulates the exponentially decaying average of square of past gradients similar to AdaDelta and RMSProp as second moment. The moment can be imagined as a ball rolling down a slope, where Adam behaves like a heavy ball with friction, which thus prefers flat minima in the error surface [39]. It is observed that Adam performs reasonably well in practice as compared to the other adaptive learning-methods. However, Adam does not utilize the change in immediate past gradient information, which is incorporated in the proposed diffGrad method. Very recently, Reddi et al. proposed AMSGrad as an improvement over Adam [40]. AMSGrad considers the maximum of past second moment (*i.e.*, “long-term memory of past gradients) in the parameter update procedure. By doing so, AMSGrad imposes more friction in order to avoid the overshooting of the minimum. AMSGrad does not change the learning rate based on the recent gradient behavior and does not deal with the saddle point problem, either. Whereas, the proposed diffGrad method controls the learning rate based on the changes in the gradient. Some of the recent works in stochastic gradient methods include the Predictive Local Smoothness based SGD (PLS-SGD) [41], the sign of each minibatch SGD (signSGD) [42], and Nostalgic Adam (NosAdam) [43], etc.

This paper proposes a difference of gradient based optimizer, which improves the well known Adam [38] with the difference of gradients (diffGrad) over the iterations. The main contributions of this paper are summarized as follows:

- This paper proposes a new diffGrad gradient descent optimization method for Convolutional Neural Networks (CNN) by considering the local gradient change information between the current and immediate past iteration.
- We show how the “short-term gradient behavior” can be utilized to control the learning rate in the optimization landscape in terms of the optimization stage, *i.e.*, near or far from an optimum solution. If change in gradient is large, it means that the optimization is not stable due to

local optima, salient region or other factors, and diffGrad allows a high learning rate. If change in gradient is small, it means that the optimization is likely to be close to the optimum solution, and diffGrad lowers the learning rate automatically.

- The proposed method also utilizes the accumulation of past gradients over iterations to deal with saddle points.
- We conduct a convergence analysis of diffGrad in terms of the regret bound using the online learning framework. We also derive a proof for diffGrad convergence.
- An empirical analysis is done by modeling the optimization problem as a regression problem to show the advantages of the proposed diffGrad optimization method over three synthetic complex non-convex functions.
- We conduct an experimental study on the proposed method and observe its improved performance for an image categorization task using the ResNet based the CNN architecture. We also experiment with different variants of diffGrad and different activation functions.

The rest of the paper is structured in the following manner: Section II presents the preliminaries in SGD optimization; Section III proposes the diffGrad optimization method; Section IV conducts the convergence analysis; Section V is devoted to the empirical analysis; Section VI presents the experimental setup; Section VII presents the experimental results, comparison and analysis; and Section VIII provides the concluding remarks.

II. PRELIMINARIES

In SGD, all parameters are updated with the same learning rate α_t in the t^{th} iteration as

$$\theta_{t+1,i} = \theta_{t,i} - \alpha_t \times g_{t,i}, \quad (1)$$

where $\theta_{t,i}$ and $\theta_{t+1,i}$ are the previous and updated values for the i^{th} parameter with $i = 1, 2, \dots, d$, where d is the number of parameters, and $g_{t,i}$ is the gradient with respect to the parameter $\theta_{t,i}$ for a loss function \mathcal{L} , defined as

$$g_{t,i} = \frac{\partial(\mathcal{L}_{t,\theta})}{\partial(\theta_{t,i})}, \quad (2)$$

where $\mathcal{L}_{t,\theta}$ is a loss function with respect to the parameters of the network (θ) in t^{th} iteration. In this paper, the cross entropy loss used for image categorization experiments is defined as

$$\mathcal{L}_{t,\theta} = \frac{1}{N_b} \sum_{j=1}^{N_b} \mathcal{L}_{t,\theta,j} + \sigma R_{t,\theta}, \quad (3)$$

where N_b is the number of training images in the batch, $\mathcal{L}_{t,\theta,j}$ is the cross entropy data loss for j^{th} training image in t^{th} iteration, $R_{t,\theta}$ is the regularization loss in t^{th} iteration, and σ is a regularization loss hyper-parameter. The cross entropy data loss $\mathcal{L}_{t,\theta,j}$ for j^{th} training image is computed as

$$\mathcal{L}_{t,\theta,j} = -\log \left(\frac{e^{S_{o_j}}}{\sum_{k=1}^{N_c} e^{S_k}} \right), \quad (4)$$

where N_c is the total number of classes in the dataset, o_j is the actual class (i.e., ground truth class) for j^{th} training image and

S_k is the computed class score for k^{th} class for j^{th} training image. The regularization loss $R_{t,\theta}$ is computed as

$$R_{t,\theta} = \sum_{i=1}^d (\theta_{t,i})^2. \quad (5)$$

In SGDM [32], the gradient in each dimension is incorporated to gain momentum for the parameters having consistent gradient, as follows:

$$m_{t,i} = \beta m_{t-1,i} + g_{t,i}, \quad (6)$$

$$\theta_{t+1,i} = \theta_{t,i} - \alpha m_{t,i}, \quad (7)$$

where $m_{t,i}$ is the moment gained at t^{th} iteration for i^{th} parameter $\theta_{t,i}$ with $m_{t,i} = 0$ for $t = 0$ and β is a hyper-parameter to control the moment.

In AdaGrad [33], the basic SGD approach is modified by normalizing the learning rate α_t as

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\alpha_t \times g_{t,i}}{\sqrt{G_{t,i} + \epsilon}}, \quad (8)$$

where ϵ is a small value to avoid division by zero and $G_{t,i}$ is the sum of the squares of the gradients of t steps for the i^{th} parameter and given as

$$G_{t,i} = \sum_{i=1}^t (g_{t,i})^2, \quad (9)$$

where $g_{t,i}$ is given by Eq. (2). Over the iterations, the value of $G_{t,i}$ may become very large due to the positive accumulation of the square of the gradients and may decrease the effective learning rate α drastically, which in turn can kill the learning process. This problem has been addressed in AdaDelta [36] and RMSProp [37] by leaking the accumulated square of gradients $G_{t,i}$ with a decay rate β . The $G_{t,i}$ in RMSProp is modified as

$$G_{t,i} = \beta G_{t-1,i} + (1 - \beta)(g_{t,i})^2 \quad (10)$$

where $G_{t-1,i} = 0$ for $t = 1$.

Adam [38] is another widely used gradient descent optimization technique that computes the learning rates at each step based on two vectors known as the 1st and 2nd order moments (i.e., mean and variance, respectively), which are recursively defined using the gradient and the square of the gradient, respectively. Basically, Adam is an improvement over RMSProp by incorporating first moment with RMSProp. Here, the 1st and 2nd order moments are defined as

$$m_{t,i} = \beta_1 m_{t-1,i} + (1 - \beta_1) g_{t,i} \quad (11)$$

$$v_{t,i} = \beta_2 v_{t-1,i} + (1 - \beta_2) g_{t,i}^2 \quad (12)$$

where β_1 and β_2 are the decay rates for first and second moments, respectively, $m_{t-1,i}$ and $v_{t-1,i}$ are the mean and variance of the gradient of the previous steps, respectively. Both $m_{t-1,i}$ and $v_{t-1,i}$ are initialized with 0 at the 1st iteration, $t = 1$. It is observed that, initially, the value of first moment is small and the value of second moment is very small, thus leading to a very large step size. In order to solve

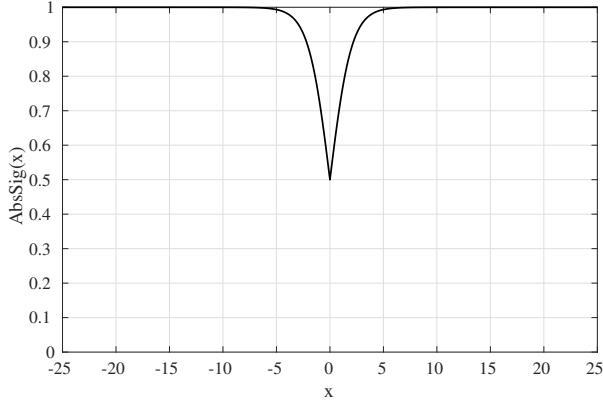


Fig. 2: The behavior of diffGrad Friction Coefficient (DFC) in terms of the $AbsSig(x)$ values. Here a DFC value of 1 means no friction and 0 means infinite friction. Note that the friction is roughly negligible if the difference in gradient is high. Here, negligible represents $DFC \approx 1$ which means there is no friction and in this case diffGrad is the same as Adam. Moreover, the minimum DFC is 0.5 when there is no gradient change. Otherwise, the DFC, close to zero, will slow down the learning process heavily.

this issue, Adam has incorporated a bias correction of the 1^{st} and 2^{nd} order moments as

$$\hat{m}_{t,i} = \frac{m_{t,i}}{(1 - \beta_1^t)} \quad \text{and} \quad \hat{v}_{t,i} = \frac{v_{t,i}}{(1 - \beta_2^t)}, \quad (13)$$

where β_1^t is β_1 power t , β_2^t is β_2 power t , and $\hat{m}_{t,i}$ and $\hat{v}_{t,i}$ are the biased corrected first and second moments, respectively. Thus, the parameter update in Adam is incorporated as

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\alpha_t \times \hat{m}_{t,i}}{\sqrt{\hat{v}_{t,i} + \epsilon}}. \quad (14)$$

A problem arises in Adam when the value of the second moment decreases significantly. Due to low values of the second moment, the friction in the optimization landscape decreases, which leads to the situation where the update process overshoots an optimum solution due to a high learning rate and diverges. This problem is addressed in AMSGrad [40] by considering the maximum of second moments in current and past iterations. In AMSGrad, the 1^{st} and 2^{nd} order moments are computed and bias-corrected similar to Adam. However, AMSGrad normalizes the learning rate α_t by the maximum of all $\hat{v}_{t,i}^{max}$ values, instead of only $\hat{v}_{t,i}$. AMSGrad memorizes the highest of 2^{nd} order moment to give more priority to those steps which update the parameter in a more accurate direction. Thus, in AMSGrad, $\hat{v}_{t,i}^{max}$ is defined as

$$\hat{v}_{t,i}^{max} = \max(\hat{v}_{t-1,i}^{max}, \hat{v}_{t,i}), \quad (15)$$

where \max is the maximum operator and $\hat{v}_{t-1,i}^{max} = 0$ for $t = 1$. Thus, the parameter update in AMSGrad is carried out using the following update rule:

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\alpha_t \times \hat{m}_{t,i}}{\sqrt{\hat{v}_{t,i}^{max} + \epsilon}}. \quad (16)$$

In practice, Adam is popular in various problems related to deep learning. Adam with 1^{st} order moment hyper-parameter $\beta_1 = 0.9$, 2^{nd} order moment hyper-parameter $\beta_2 = 0.999$, and learning rate $\alpha \in [10^{-2}, 10^{-4}]$ is a good starting choice for many models [38].

III. PROPOSED DIFFGRAD OPTIMIZATION

From the discussions in the previous section we can conclude that recent optimization techniques such as Adam and AMSGrad suffer from the problem of automatic adjustment of the learning rate. The main problem is with controlling friction for the first moment in order to avoid overshooting near to an optimum solution. In this section, we propose a new gradient descent optimization technique referred to as diffGrad to address these issues of existing gradient descent optimization techniques. The proposed diffGrad optimization technique is based on the change in short-term gradients and controls the learning rate based on the need of dynamic adjustment of learning rate. This means that diffGrad follows the norm that the parameter update should be smaller in low gradient changing regions and vice-versa. diffGrad computes the 1^{st} and 2^{nd} order moments (i.e., $m_{t,i}$ and $v_{t,i}$, respectively) as well as the 1^{st} and 2^{nd} order bias-corrected moments (i.e., $\hat{m}_{t,i}$ and $\hat{v}_{t,i}$, respectively) for the i^{th} parameter at the t^{th} iteration similar to Adam [38] using Eq. (11-13).

A diffGrad friction coefficient (DFC) is introduced in the proposed work to control the learning rate using information of short-term gradient behavior. The DFC is represented by ξ and defined as

$$\xi_{t,i} = AbsSig(\Delta g_{t,i}) \quad (17)$$

where $AbsSig$ is a non-linear sigmoid function that squashes every value between 0.5 and 1, and is defined as

$$AbsSig(x) = \frac{1}{1 + e^{-|x|}} \quad (18)$$

while $\Delta g_{t,i}$ is the change in gradient between immediate past and current iterations, given as

$$\Delta g_{t,i} = g_{t-1,i} - g_{t,i} \quad (19)$$

where $g_{t,i}$ is the computed gradient for the i^{th} parameter at the t^{th} iteration and defined in Eq. (2).

The behavior of DFC (i.e., ξ) is characterized in Fig. 2 in terms of $AbsSig(x)$ which represents the friction with respect to the change in gradient. It can be observed from Fig. 2 that large changes in the gradient incur less friction, whereas small changes in the gradient incur more friction with at most 0.5 when there is no change in the gradient. Note that $|\Delta g_{t,i}| \in \mathbb{R}_0^+$, $\forall i \in [1, d]$, and $i \in \mathbb{I}^+$ leads to $\xi_{t,i} \in [0.5, 1]$ at any iteration t . The DFC imposes more friction when gradient changes slowly and vice-versa.

In the proposed diffGrad optimization method, the steps up to the computation of bias-corrected 1^{st} order moment $\hat{m}_{t,i}$ and bias-corrected 2^{nd} order moment $\hat{v}_{t,i}$ are the same as those of Adam optimization [38]. The diffGrad optimization method updates the i^{th} parameter at the t^{th} iteration using the following update rule:

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\alpha_t \times \xi_{t,i} \times \hat{m}_{t,i}}{\sqrt{\hat{v}_{t,i} + \epsilon}}, \quad (20)$$

where $\theta_{t,i}$ is the weight for i^{th} parameter at t^{th} iteration, α_t is the learning rate for t^{th} iteration, ϵ is a very small value of approximately 10^{-7} added to avoid division by zero, $\xi_{t,i}$ is the diffGrad friction coefficient, and $\hat{m}_{t,i}$ and $\hat{v}_{t,i}$ are the

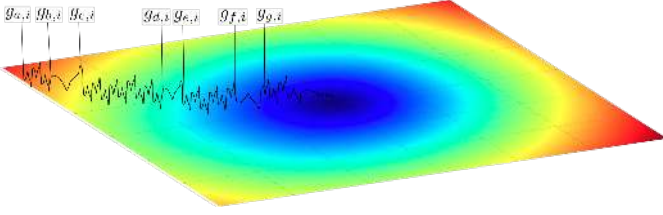


Fig. 3: The illustration of optimization landscape in order to understand the importance of short-term gradient change. The dark blue area represents the optimization goal to be reached.

bias-corrected 1^{st} order moment and the bias-corrected 2^{nd} order moment, respectively, as defined in Equation (13).

The proposed optimizer introduces the diffGrad friction coefficient (DFC) to control the oscillations. The difference in gradient reduces the learning rate by controlling the moving average near an optimum solution. The necessity of the friction coefficient is illustrated using an optimization landscape in Fig. 3. The dark blue area shows the optimization goal. As we move from dark red to dark blue, the depth of the optima tends to be reduced. Lighter shades between two colors show steep descent towards the optimization goal. Here, $g_{a,i} \rightarrow g_{g,i}$ shows the gradients at steps a to g . The color between $g_{a,i}$ and $g_{b,i}$ changes very rapidly. As a result, the learning rate increases due to the increase in the diffGrad friction coefficient. Hence, the frequency of vertical fluctuations also increases. There is a very steep descent between the pairs $g_{c,i} \rightarrow g_{d,i}$ and $g_{e,i} \rightarrow g_{f,i}$. Hence, we again see the frequent fluctuations in the vertical direction. The descents between $g_{b,i} \rightarrow g_{c,i}$ and $g_{f,i} \rightarrow g_{g,i}$ are very slow, which is evident from slow or no change in color. Here, the DFC reduces the frequency of fluctuations by decreasing the rate of learning of the network. Hence, the introduction of DFC reduces redundant learning and increases the rate of convergence. It also helps in finding an optimum solution by reducing the vertical fluctuation near local optima.

The problem of an ever decreasing learning rate in Adam [33] has been solved by Adam [38] by introducing two moments on gradient. The adaptive nature of these two moments are controlled during learning with respect to the slope of the descent of the probable solution towards a local optimum. However, a sudden decrease in second moment adversely affects the Adam optimizer and the actual solution overshoots the local optima. This problem has been resolved in AMSGrad [40] by introducing a new parameter, $\hat{v}_{t,i}^{max} = \max(\hat{v}_{t-1,i}^{max}, \hat{v}_{t,i})$, which prevents the learning rate from overshooting. Adam and AMSgrad ignore the effect of the 1^{st} moment in controlling the learning rate over the optimization landscape. The proposed diffGrad optimizer resolves this drawback by introducing a new parameter DFC (ξ), which not only allows the high learning rate for a high gradient changing surface, but also reduces the learning rate for a low gradient changing surface and prevents the probable solution from overshooting. The improved performance of diffGrad shows that the optimizer effectively introduces the required friction (\uparrow or \downarrow). Please note that the minimum DFC is considered as 0.5, as shown in Fig. 2, such that the optimization should not become stuck in local optima and

saddle regions. Due to the high moment gained, sufficient magnitude of the step size will be allowed by the DFC in the proposed diffGrad approach, so that it emerges from flat local optima and flat saddle regions.

IV. CONVERGENCE ANALYSIS

The convergence property of Adam [38] is shown using the online learning framework proposed in [44]. We also use this technique to analyze the convergence of the proposed diffGrad optimizer. Consider $f_1(\theta), f_2(\theta), \dots, f_T(\theta)$ as the unknown sequence of convex cost functions. Our aim is to predict parameter θ_t at each iteration t and evaluate over $f_t(\theta)$. For this type of problem, where the nature of the sequence is not known a priori, the algorithm can be evaluated based on the regret bound. The regret bound is computed by summing the difference between all the previous online guesses $f_t(\theta_t)$ and the best fixed point parameter $f_t(\theta^*)$ from a feasible set χ for all the previous iterations. Mathematically, the regret bound is given as

$$R(T) = \sum_{t=1}^T [f_t(\theta_t) - f_t(\theta^*)] \quad (21)$$

where $\theta^* = \arg \min_{\theta \in \chi} \sum_{t=1}^T f_t(\theta)$. We observe that diffGrad has an $O(\sqrt{T})$ regret bound. The proof is given in the appendix. Our regret bound is comparable to general convex online learning methods. We have used the following definitions: $g_{t,i}$ refers to the gradient in the t^{th} iteration for the i^{th} element, $g_{1:t,i} = [g_{1,i}, g_{2,i}, \dots, g_{t,i}] \in \mathbb{R}^t$ is the vector of gradients in the i^{th} dimension over all iterations up to t , and $\gamma \triangleq \frac{\beta_1^2}{\sqrt{\beta_2}}$.

Theorem 1. Consider the bounded gradients for function f_t (i.e., $\|g_{t,\theta}\|_2 \leq G$ and $\|g_{t,\theta}\|_\infty \leq G_\infty$) for all $\theta \in \mathbb{R}^d$. Also assume that diffGrad produces the bounded distance between any θ_t (i.e., $\|\theta_n - \theta_m\|_2 \leq D$ and $\|\theta_n - \theta_m\|_\infty \leq D_\infty$ for any $m, n \in \{1, \dots, T\}$). Let $\gamma \triangleq \frac{\beta_1^2}{\sqrt{\beta_2}}$, $\beta_1, \beta_2 \in [0, 1)$ satisfy $\frac{\beta_1^2}{\sqrt{\beta_2}} < 1$, $\alpha_t = \frac{\alpha}{\sqrt{t}}$, and $\beta_{1,t} = \beta_1 \lambda^{t-1}$, $\lambda \in (0, 1)$ where λ is typically very close to 1, e.g., $1 - 10^{-8}$. For all $T \geq 1$, the proposed diffGrad optimizer shows the following guarantee:

$$\begin{aligned} R(T) \leq & \frac{D^2}{2\alpha(1-\beta_1)} \sum_{i=1}^d (1 + e^{-|g_{1,i}|}) \sqrt{T \hat{v}_{T,i}} \\ & + \frac{\alpha(1+\beta_1)G_\infty}{(1-\beta_1)\sqrt{1-\beta_2}(1-\gamma)^2} \sum_{i=1}^d \|g_{1:T,i}\|_2 \\ & + \sum_{i=1}^d \frac{D_\infty^2 G_\infty \sqrt{1-\beta_2}}{2\alpha(1-\beta_1)(1-\lambda)^2} \end{aligned} \quad (22)$$

Note that the additive term over the dimension (d) can be much smaller than its upper bound $\sum_{i=1}^d \|g_{1:T,i}\|_2 \ll dG_\infty \sqrt{T}$ and $\sum_{i=1}^d (1 + e^{-|g_{1,i}|}) \sqrt{T \hat{v}_{T,i}} \ll d(1 + E_\infty)G_\infty \sqrt{T}$, where E_∞ is the upper bound over the exponential function and $E_\infty \gg \sum_{i=1}^d e^{-|g_{1,i}|}$. In general, $O(\log d \sqrt{T})$ is achieved by adaptive methods such as diffGrad and Adam which is improved over the $O(\sqrt{dT})$ of non-adaptive methods. The proposed diffGrad method also uses the decay of $\beta_{1,t}$ for the theoretical analysis, similar to Adam.

Finally, by using the above theorem and $\sum_{i=1}^d \|g_{1:T,i}\|_2 \ll dG_\infty\sqrt{T}$ and $\sum_{i=1}^d (1 + e^{-|g_{1,i}|})\sqrt{T}\hat{v}_{T,i} \ll d(1 + E_\infty)G_\infty\sqrt{T}$, the convergence of average regret of diffGrad can be shown as described in the following corollary.

Corollary 1. *Consider the bounded gradients for function f_t (i.e., $\|g_{t,\theta}\|_2 \leq G$ and $\|g_{t,\theta}\|_\infty \leq G_\infty$) for all $\theta \in \mathbb{R}^d$. Also, assume that diffGrad produces the bounded distance between any θ_t (i.e., $\|\theta_n - \theta_m\|_2 \leq D$ and $\|\theta_n - \theta_m\|_\infty \leq D_\infty$ for any $m, n \in \{1, \dots, T\}$). For all $T \geq 1$, the proposed diffGrad optimizer shows the following guarantee:*

$$\frac{R(T)}{T} = O\left(\frac{1}{\sqrt{T}}\right). \quad (23)$$

Thus, $\lim_{T \rightarrow \infty} \frac{R(T)}{T} = 0$.

V. EMPIRICAL ANALYSIS

In order to justify the purpose of introducing the difference of gradient based friction in diffGrad optimization, we have conducted an empirical analysis in this section. We have modeled the optimization problem as a regression problem over three one-dimensional non-convex functions. We have performed the optimization over these functions by using the widely used Adam [38] and the proposed diffGrad optimization methods.

The following non-convex functions, represented by $F1$, $F2$, and $F3$, respectively, are used for this empirical analysis:

$$F1(x) = \begin{cases} (x + 0.3)^2, & \text{for } x \leq 0 \\ (x - 0.2)^2 + 0.05, & \text{for } x > 0 \end{cases} \quad (24)$$

$$F2(x) = \begin{cases} -40x - 35.15, & \text{for } x \leq -0.9 \\ x^3 + x \sin(8x) + 0.85, & \text{for } x > -0.9 \end{cases} \quad (25)$$

$$F3(x) = \begin{cases} x^2, & \text{for } x \leq -0.5 \\ 0.75 + x, & \text{for } -0.5 < x \leq -0.4 \\ -7x/8, & \text{for } -0.4 < x \leq 0 \\ 7x/8, & \text{for } 0 < x \leq 0.4 \\ 0.75 - x, & \text{for } 0.4 < x \leq 0.5 \\ x^2, & \text{for } 0.5 < x \end{cases} \quad (26)$$

where x is the input for this function with $-\infty < x < +\infty$. Functions $F1$, $F2$, and $F3$ are shown in Fig. 4a, Fig. 4d, and Fig. 4g, respectively, for $-1 < x < +1$.

It can be observed that function $F1$ has one global minimum and one local minimum, whereas functions $F2$ and $F3$ have one global minimum and two local minima. In this experiment, for both Adam [38] as well as the proposed diffGrad optimization methods, the following are the hyper-parameter settings: the decay rate for 1st moment (β_1) is 0.95; the decay rate for 2nd moment (β_2) is 0.999; the learning rate (η) is 0.1 for each iteration; both the 1st order moment (m) and 2nd order moment (v) are initialized to 0; and the parameter θ is initialized to -1 in order to show the advantages of the proposed method. In the proposed diffGrad optimization method, the previous gradient value at the 1st iteration (i.e.,

g_0) is set to zero. We run Adam [38] and the proposed diffGrad optimization for 300 iterations for all functions. The regression loss as well as parameter value θ are recorded for both optimization methods at each iteration and analyzed.

Fig. 4b and Fig. 4c depict the regression loss and the parameter value (θ), respectively, after each iteration for function $F1$. It is discovered from these plots that Adam overshoots the global minimum due to the high moment gained so far. It can be observed in Fig. 4c that Adam overshoots the global minimum at $\theta = -0.3$ and becomes stuck in the local minimum at $\theta = 0.2$. This problem is addressed by DFC of the proposed diffGrad optimization method, which controls the momentum while reaching towards the global minimum and the overshoot does not occur. Moreover, diffGrad is able to reach a zero loss, as opposed to Adam, which saturates with reasonable loss. The same behavior is also observed for function $F2$, as shown in Fig. 4e and Fig. 4f.

Both Adam and diffGrad are able to achieve the global minimum for function $F3$, as shown in Fig. 4h and Fig. 4i. Note that one local minimum is present before the global minimum in function $F3$ (see Fig. 4g). Both Adam and diffGrad accumulate enough momentum to cross the local minimum. It can be seen in Fig. 4i that Adam oscillates with higher frequency and amplitude around the global minimum, as compared to diffGrad. Thus, better stability is obtained by diffGrad.

VI. EXPERIMENTAL SETUP FOR CLASSIFICATION

This section presents the setup and settings used in the experiments in terms of the deep architecture used, the hyper-parameter setting, and the applied dataset.

A. Deep Architecture Used

The experiments are conducted for an image categorization problem. Convolutional Neural Networks (CNN) are generally used for processing the images. The popular CNN architectures for image categorization problems are AlexNet [2], VGGNet [4], GoogleNet [19], and ResNet [20]. The ResNet based CNN architecture introduced by He et al. [20] is one of the most accurate models which won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [45] in 2015. The ResNet architecture is made with the residual unit. A residual unit considers the learning module as the residual of output w.r.t. the input by implementing a direct connection between input and output of that unit. The residual unit facilitates the training of deeper ResNet architecture which was not feasible in the earlier CNN architectures. The PyTorch implementation of ResNet, publicly available through GitHub¹, is used for our experiments. The depth of ResNet is considered as 50 in this paper. Following is the ResNet50 architecture details for the CIFAR datasets: Conv: $\{3 \times 3, 64\}$, BatchNorm, Bottleneck(3 times): $\{\text{Conv: } [1 \times 1, 64], \text{BatchNorm, Conv: } [3 \times 3, 64], \text{BatchNorm: } [1 \times 1, 256], \text{BatchNorm}\}$, Bottleneck(4 times): $\{\text{Conv: } [1 \times 1, 128], \text{BatchNorm, Conv: } [3 \times 3, 128], \text{BatchNorm, Conv: } [1 \times 1, 512], \text{BatchNorm}\}$,

¹<https://github.com/kuangliu/pytorch-cifar>

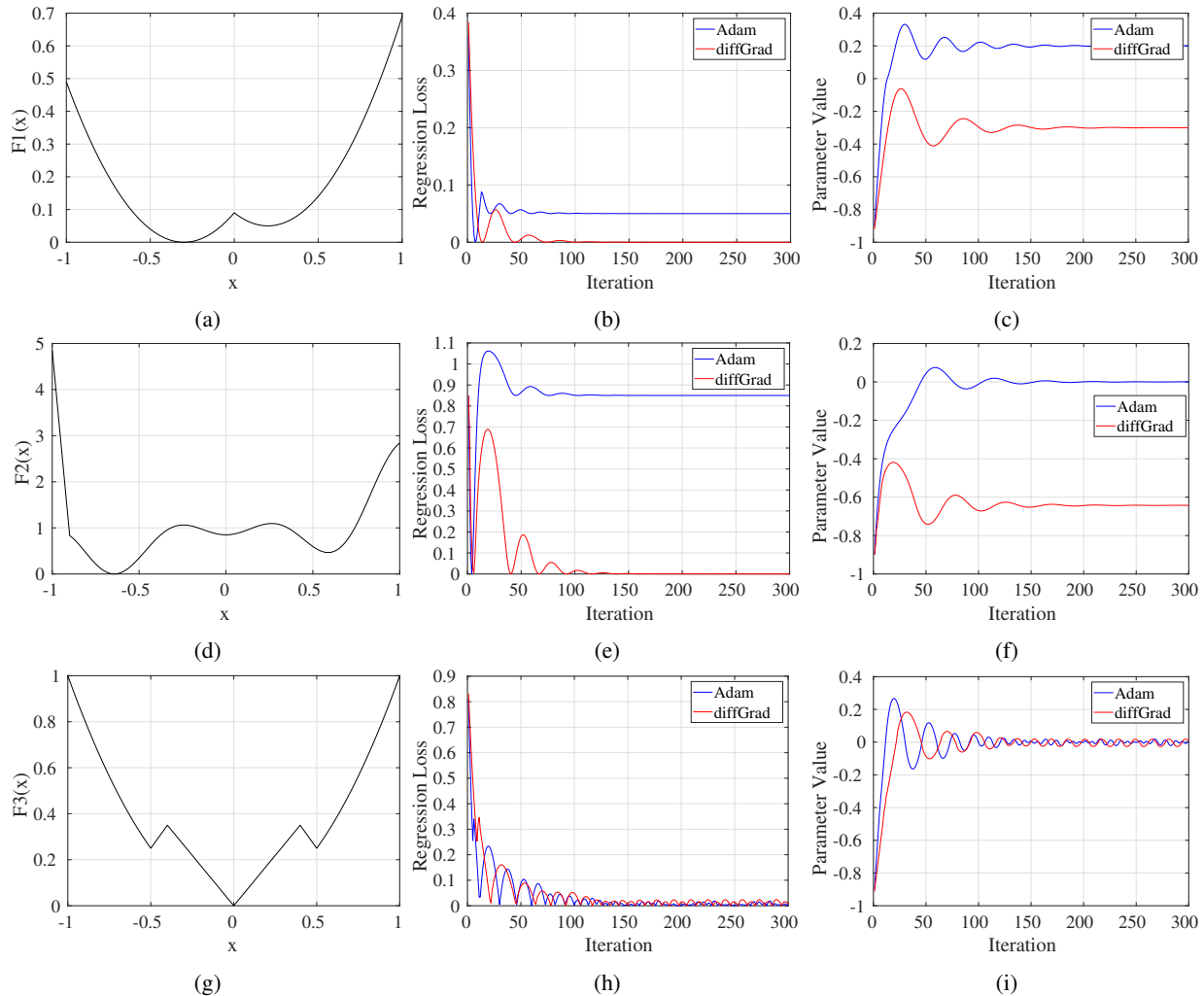


Fig. 4: The empirical results comparison among the Adam [38] and the proposed diffGrad optimization techniques over three synthetic non-convex functions. (a,d,g) The non-convex synthetic functions $F1$, $F2$, $F3$, respectively, (b,e,h) The regression loss over functions $F1$, $F2$, $F3$, respectively at each iteration, (c,f,i) The parameter value θ after each iteration during optimization of functions $F1$, $F2$, $F3$, respectively.

Bottleneck(6 times): {Conv: $[1 \times 1, 256]$, BatchNorm, Conv: $[3 \times 3, 256]$, BatchNorm, Conv: $[1 \times 1, 1024]$, BatchNorm}, Bottleneck(3 times): {Conv: $[1 \times 1, 512]$, BatchNorm, Conv: $[3 \times 3, 512]$, BatchNorm, Conv: $[1 \times 1, 2048]$, BatchNorm}, AveragePooling: $\{4 \times 4\}$. For more information regarding the used ResNet50 architecture, the above mentioned GitHub page may be consulted.

B. Hyper-Parameter Setting

The hyper-parameter settings in the experiments are illustrated in this paragraph. Stochastic Gradient Descent with momentum (SGDM) is used as the optimization technique. The batch sizes of 32, 64 and 128 are opted. The number of epochs is 100 with a learning rate of 10^{-3} for the first 80 epochs and 10^{-4} for the last 20 epochs. For all the optimizers, the default settings of PyTorch are used, excluding the moment coefficient for SGDM, which is set to 0.9.

TABLE I: The comparison results in terms of the ‘Validation Classification Accuracy’ over the CIFAR10 and CIFAR100 databases among SGDM, AdaGrad, AdaDelta, RMSProp, AMSGrad, Adam and proposed diffGrad optimization methods. The comparison is made for the batch sizes (N_b) of 32, 64 and 128. The best results among different optimization techniques are highlighted in bold.

Optimizer	CIFAR10 Database			CIFAR100 Database		
	$N_b=32$	$N_b=64$	$N_b=128$	$N_b=32$	$N_b=64$	$N_b=128$
SGDM	92.95	92.07	90.15	73.3	70.37	67.25
AdaGrad	92.04	92.01	91.3	71.01	70.8	68.38
AdaDelta	93.66	93.48	93.54	73.46	74.09	74.12
RMSProp	92.82	92.32	92.26	65.31	66.8	62.65
AMSGrad	93.73	93.4	93.51	73.06	72.45	72.86
Adam	93.78	93.81	93.72	71.82	73.31	73.72
diffGrad	94.24	94.24	94.27	75.63	76.18	75.57

C. Dataset Used

In order to conduct the image categorization experiments, the CIFAR10 and CIFAR100 datasets² [46] are used in this paper. Both the CIFAR10 and the CIFAR100 datasets consist of the same 60000 images, including 50000 images for training and 10000 images for validation. In the CIFAR10 dataset,

²<https://www.cs.toronto.edu/~kriz/cifar.html>

TABLE II: The comparison results in terms of the ‘Validation Classification Accuracy’ over the CIFAR10 and CIFAR100 databases among different variants of diffGrad. The comparison is made for the batch sizes (N_b) of 32, 64 and 128. The best results among different optimization techniques are highlighted in bold.

Optimizer	CIFAR10 Database			CIFAR100 Database		
	$N_b=32$	$N_b=64$	$N_b=128$	$N_b=32$	$N_b=64$	$N_b=128$
diffGrad	94.24	94.24	94.27	75.63	76.18	75.57
diffGrad1	94.34	94.31	94.06	75.23	75.24	75.25
diffGrad2	94.08	93.92	94.09	75.37	75.5	75.83
diffGrad3	94.01	94.2	93.58	75.6	76.15	75.39
diffGrad4	94.06	94.37	93.75	74.9	75.35	75.45
diffGrad5	94.42	94.02	94.16	75.55	75.91	75.16

all images are divided into 10 categories including ‘airplane’, ‘automobile’, ‘bird’, ‘cat’, ‘deer’, ‘dog’, ‘frog’, ‘horse’, ‘ship’, and ‘truck’. Whereas, in the CIFAR100 dataset, the same set of images is partitioned into 100 categories. The dimension of all images is $32 \times 32 \times 3$. Note that the images are pre-processed to make the RGB values zero-centered with unit standard deviation across each color channel. The data augmentation is done only over the training images by the process of flipping and cropping. The images are randomly flipped horizontally (i.e., with respect to the vertical axis) with a probability of 0.5. In case of cropping, at first the images are scaled to $40 \times 40 \times 3$ size by zero padding. Then, the images of size $32 \times 32 \times 3$ are cropped randomly from the upscaled images.

VII. CLASSIFICATION EXPERIMENTS AND ANALYSIS

The image categorization experiments over the CIFAR10 and CIFAR100 datasets are conducted to test the performance improvement of the proposed diffGrad gradient descent optimization method. The ResNet50 is used to demonstrate the suitability of diffGrad for the CNN model. In this section, at first the results of the proposed diffGrad method are presented, then the results are compared with other state-of-the-art optimization methods, and finally the stability of the proposed diffGrad optimization is tested over different activation functions. The results are computed in terms of the average top-1 validation classification accuracy.

A. Validation Results Comparison

The validation classification accuracy due to the diffGrad method is compared with state-of-the-art optimization techniques such as SGDM [32], AdaGrad [33], AdaDelta [36], RMSProp [37], AMSGrad [40], and ADAM [38]. Table I depicts the validation classification accuracy for different optimization techniques. The results are presented over the CIFAR10 and CIFAR100 datasets. The best results among the different optimization techniques are highlighted in bold. It is observed from Table I that the proposed diffGrad optimization technique outperforms all other optimization techniques over both the CIFAR10 and the CIFAR100 datasets for all tested batch sizes of 32, 64, and 128, respectively. The proposed diffGrad method utilizes all the positive characteristics of Adam. Moreover, the effect of the proposed difference of gradient based friction technique prevents the network from noisy oscillation near the minimum solution. It leads to more accurate results as compared to other optimization techniques.

TABLE III: The performance of the proposed diffGrad optimization technique with ResNet50 model for different activation functions, namely ReLU, Leaky ReLU (LReLU), ELU, and SELU. The Validation Classification Accuracy over CIFAR10 is reported for the batch sizes (N_b) of 32, 64 and 128. The best results among different activation’s are highlighted in bold.

Optimizer	CIFAR10 Database		
	$N_b=32$	$N_b=64$	$N_b=128$
ResNet50(ReLU)	94.24	94.24	94.27
ResNet50(LReLU)	94.3	94.2	94.3
ResNet50(ELU)	94.48	94.2	94.24
ResNet50(SELU)	94.03	93.62	93.96

B. Experiments with diffGrad Variants

In this section, we modify the diffGrad Friction Coefficient (DFC) of diffGrad and analyze the performance over the CIFAR10 dataset. The DFC is basically given as $DFC = 1/(1 + e^{-(|g_{t-1}-g_t|)})$ with the range $DFC \in [0.5, 1]$. We modify the DFC by removing the absolute value, which is given as $DFC1 = 1/(1 + e^{-(g_{t-1}-g_t)})$. Note that $DFC1 \in [0, 1]$. We also generate another version called DFC2, given by $DFC2 = 9/(1 + e^{-(0.5|g_{t-1}-g_t|)} - 4)$ with the range $DFC2 \in [0.5, 5]$. We also use the mean (μ) and standard deviation (ν) of absolute gradients of the batch with DFC and consider the following three more scenarios: (a) $DFC3 = 1/(1 + e^{-(\nu|g_{t-1}-g_t|-\mu)})$ with the range $DFC3 \in [0.5, 1]$, (b) $DFC4 = 1/(1 + e^{-(\nu^2|g_{t-1}-g_t|-\mu)})$ with the range $DFC4 \in [0.5, 1]$, and (c) $DFC5 = 1/(1 + e^{-(\sqrt{\nu}|g_{t-1}-g_t|-\mu)})$ with the range $DFC5 \in [0.5, 1]$. We define numbered diffGrad accordingly, i.e., diffGrad1 using DFC1 and so on. The validation classification accuracy for diffGrad using DFC, diffGrad1 using DFC1, diffGrad2 using DFC2, diffGrad3 using DFC3, diffGrad4 using DFC4, and diffGrad5 using DFC5 are presented in Table II. The results are computed over the CIFAR10 and CIFAR100 datasets with batch sizes 32, 64, and 128. It can be noticed from this result that the original diffGrad performs better over CIFAR10 dataset for high batch size and over CIFAR100 dataset for small batch sizes. The mean and standard deviation based diffGrad variants such as DFC4 and DFC5 perform better over CIFAR10 dataset for 64 and 32 batch sizes, respectively. It is due to the fact that the mean and standard deviation of the absolute gradient tends to be more accurate for higher batch size. From Table I and Table II, it is clear that original diffGrad as well as its mean and standard deviation based variants show the promising performance.

C. Performance Analysis with Activation Functions

In other experiments, the Rectified Linear Unit (ReLU) [2] is used as the default activation function in the framework of ResNet [20]. In this experiment, we have considered four activation functions, namely, ReLU [2], Leaky ReLU (LReLU) [47], Exponential Linear Unit (ELU) [48], and Scaled ELU (SELU) [49]. The performance of the proposed diffGrad optimization method is computed over the CIFAR10 dataset for different activation functions. We have used the same experimental setup as for the ResNet50 model by replacing all the activation functions with ReLU, LReLU, ELU, and SELU, one by one. The rest of the experimental setup is same as the setup of previous experiments. The leaky factor in LReLU

is considered as 0.1. The validation classification accuracy is compared in Table III. It is clear from this result that the ELU, ReLU and LReLU outperform others for the batch sizes of 32, 64 and 128, respectively.

VIII. CONCLUSION

In this paper, a new stochastic gradient descent optimization method diffGrad is proposed. diffGrad incorporates the difference of gradients of current and immediate past iteration (i.e., short term gradient change information) with Adam optimization techniques to control the learning rate based on the optimization stage. The proposed diffGrad allows a high learning rate if the gradient change is more (i.e., the optimization is far from the optimum solution), and a low learning rate if the gradient changes minimally (i.e., the optimization is near to the optimum solution). The local optima and saddle point scenarios are handled by the moment gained due to past gradients. The regret bound analysis provides a guarantee of convergence. An empirical analysis over three synthetic, non-convex functions reveals that the proposed diffGrad optimization method controls the update step in order to avoid the overshooting of global minimum and oscillation around global minimum. The proposed diffGrad optimization method is also tested with the ResNet50 model for an image categorization task over the CIFAR10 and CIFAR100 datasets. The results are compared with the state-of-the-art SGD optimization techniques such as SGDM, AdaGrad, AdaDelta, RMSProp, AMSGrad, and Adam. It is observed that the diffGrad outperforms all other optimizers. Moreover, the mean and standard deviation of the absolute gradient of a batch can be used with diffGrad for better result over CIFAR10 dataset.

ACKNOWLEDGMENTS

The authors would like to thank NVIDIA Corporation for the support of GeForce Titan X Pascal GPU donated to Computer Vision Group, IIIT Sri City. The authors would also like to thank the anonymous Associate Editor and Reviewers for their valuable comments to improve the quality of the paper. The authors would also like to thank Dr. Shrijita for editing the paper.

REFERENCES

- [1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.
- [3] X. Lan, S. Zhang, P. C. Yuen, and R. Chellappa, "Learning common and feature-specific patterns: a novel multiple-sparse-representation-based tracker," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 2022–2037, 2018.
- [4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [5] R. Shao, X. Lan, and P. C. Yuen, "Joint discriminative learning of deep dynamic textures for 3d mask face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 4, pp. 923–938, 2019.
- [6] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

- [7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.
- [8] D. Yu and L. Deng, "Deep learning and its applications to signal and information processing [exploratory dsp]," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 145–154, 2011.
- [9] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2013.
- [10] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.
- [11] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th International Conference on Machine Learning*. ACM, 2008, pp. 160–167.
- [12] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2017.
- [13] X. Lan, A. J. Ma, P. C. Yuen, and R. Chellappa, "Joint sparse representation and robust feature-level fusion for multi-cue visual tracking," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5826–5841, 2015.
- [14] S. Zhang, X. Lan, H. Yao, H. Zhou, D. Tao, and X. Li, "A biologically inspired appearance model for robust visual tracking," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2357–2370, 2016.
- [15] X. Lan, M. Ye, R. Shao, B. Zhong, P. C. Yuen, and H. Zhou, "Learning modality-consistency feature templates: A robust rgb-infrared tracking system," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9887–9897, 2019.
- [16] S. P. T. Reddy, S. T. Karri, S. R. Dubey, and S. Mukherjee, "Spontaneous facial micro-expression recognition using 3d spatiotemporal convolutional neural networks," in *Proceedings of the IEEE International Joint Conference on Neural Networks*, 2019 (Accepted).
- [17] C. Nagpal and S. R. Dubey, "A performance evaluation of convolutional neural networks for face anti spoofing," in *Proceedings of the IEEE International Joint Conference on Neural Networks*, 2019 (Accepted).
- [18] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "Hybrids: Exploring 3d-2d cnn feature hierarchy for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, 2019 (In Press).
- [19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [21] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.
- [22] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.
- [23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proceedings of the Advances in neural information processing systems*, 2015, pp. 91–99.
- [24] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.
- [25] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.
- [26] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8759–8768.
- [27] S. S. Basha, S. Ghosh, K. K. Babu, S. R. Dubey, V. Pulabaigari, and S. Mukherjee, "Recnet: An efficient convolutional neural network for histological routine colon cancer nuclei classification," in *Proceedings of the 15th International Conference on Control, Automation, Robotics and Vision*, 2018, pp. 1222–1227.
- [28] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of the COMPSTAT*, 2010, pp. 177–186.

- [29] R. Sutton, "Two problems with back propagation and other steepest descent learning procedures for networks," in *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, 1986, pp. 823–832.
- [30] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization," in *Proceedings of the Advances in Neural Information Processing Systems*, 2014, pp. 2933–2941.
- [31] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural Networks*, vol. 12, no. 1, pp. 145–151, 1999.
- [32] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proceedings of the International Conference on Machine Learning*, 2013, pp. 1139–1147.
- [33] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.
- [34] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le *et al.*, "Large scale distributed deep networks," in *Proceedings of the Advances in Neural Information Processing Systems*, 2012, pp. 1223–1231.
- [35] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1532–1543.
- [36] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.
- [37] G. Hinton, N. Srivastava, and K. Swersky, "Neural networks for machine learning," Lecture 6a overview of mini-batch gradient descent course, 2012.
- [38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [39] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a nash equilibrium," *arXiv preprint arXiv:1706.08500*, 2017.
- [40] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," in *Proceedings of the International Conference on Learning Representations*, 2018.
- [41] J. Li, H. Liu, B. Zhong, Y. Wu, and Y. Fu, "Predictive local smoothness for stochastic gradient methods," *arXiv preprint arXiv:1805.09386*, 2018.
- [42] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signsgd: compressed optimisation for non-convex problems," *arXiv preprint arXiv:1802.04434*, 2018.
- [43] H. Huang, C. Wang, and B. Dong, "Nostalgic adam: Weighing more of the past gradients when designing the adaptive learning rate," *arXiv preprint arXiv:1805.07557*, 2018.
- [44] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *Proceedings of the International Conference on Machine Learning*, 2003, pp. 928–936.
- [45] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [46] A. Krizhevsky, "Learning multiple layers of features from tiny images," *Master's thesis, University of Tront*, 2009.
- [47] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proceedings of the International Conference on Machine Learning*, 2013, pp. 1–6.
- [48] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.
- [49] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in *Proceedings of the Advances in neural information processing systems*, 2017, pp. 971–980.

APPENDIX

A. Convergence Proof

Theorem 2. *Let the bounded gradients for function f_t (i.e., $\|g_{t,\theta}\|_2 \leq G$ and $\|g_{t,\theta}\|_\infty \leq G_\infty$) for all $\theta \in R^d$. Also assume that diffGrad produces the bounded distance between any θ_t (i.e., $\|\theta_n - \theta_m\|_2 \leq D$ and $\|\theta_n - \theta_m\|_\infty \leq D_\infty$ for any $m, n \in \{1, \dots, T\}$). Let $\gamma \triangleq \frac{\beta_1^2}{\sqrt{\beta_2}}$, $\beta_1, \beta_2 \in [0, 1)$ satisfy $\frac{\beta_1^2}{\sqrt{\beta_2}} < 1$, $\alpha_t = \frac{\alpha}{\sqrt{t}}$, and $\beta_{1,t} = \beta_1 \lambda^{t-1}$, $\lambda \in (0, 1)$ with λ is typically*

close to 1, e.g $1 - 10^{-8}$. For all $T \geq 1$, the proposed diffGrad optimizer shows the following guarantee:

$$R(T) \leq \frac{D^2}{2\alpha(1-\beta_1)} \sum_{i=1}^d (1 + e^{-|g_{1,i}|}) \sqrt{T \hat{v}_{T,i}} + \frac{\alpha(1+\beta_1)G_\infty}{(1-\beta_1)\sqrt{1-\beta_2}(1-\gamma)^2} \sum_{i=1}^d \|g_{1:T,i}\|_2 + \sum_{i=1}^d \frac{D_\infty^2 G_\infty \sqrt{1-\beta_2}}{2\alpha(1-\beta_1)(1-\lambda)^2} \quad (27)$$

Proof. Using Lemma 10.2 of Adam [38], we can write as

$$f_t(\theta_t) - f_t(\theta^*) \leq g_t^T(\theta_t - \theta^*) = \sum_{i=1}^d g_{t,i}(\theta_{t,i} - \theta_{i,i}^*)$$

We can write following from the diffGrad update rule described in Eq. (20), ignoring ϵ_0 ,

$$\theta_{t+1} = \theta_t - \frac{\alpha_t \xi_t \hat{m}_t}{\sqrt{\hat{v}_t}} = \theta_t - \frac{\alpha_t \xi_t}{(1-\beta_1^t)} \left(\frac{\beta_{1,t}}{\sqrt{\hat{v}_t}} m_{t-1} + \frac{(1-\beta_{1,t})}{\sqrt{\hat{v}_t}} g_t \right) \quad (28)$$

where $\beta_{1,t}$ is the 1st order moment at t^{th} iteration and β_1^t is the t^{th} power of initial 1st order moment.

For i^{th} dimension of parameter vector $\theta_t \in R^d$, we can write

$$(\theta_{t+1,i} - \theta_{i,i}^*)^2 = (\theta_{t,i} - \theta_{i,i}^*)^2 - \frac{2\alpha_t \xi_{t,i}}{1-\beta_1^t} \left(\frac{\beta_{1,t}}{\sqrt{\hat{v}_{t,i}}} m_{t-1,i} + \frac{(1-\beta_{1,t})}{\sqrt{\hat{v}_{t,i}}} g_{t,i} \right) (\theta_{t,i} - \theta_{i,i}^*) + \alpha_t^2 \xi_{t,i}^2 \left(\frac{\hat{m}_{t,i}}{\hat{v}_{t,i}} \right)^2 \quad (29)$$

The above equation can be reordered as

$$g_{t,i}(\theta_{t,i} - \theta_{i,i}^*) = \frac{(1-\beta_1^t)\sqrt{\hat{v}_{t,i}}}{2\alpha_t \xi_{t,i}(1-\beta_{1,t})} \left((\theta_{t,i} - \theta_{i,i}^*)^2 - (\theta_{t+1,i} - \theta_{i,i}^*)^2 \right) + \frac{\beta_{1,t}}{1-\beta_{1,t}} (\theta_{i,i}^* - \theta_{t,i}) m_{t-1,i} + \frac{\alpha_t(1-\beta_1^t)\xi_{t,i}}{2(1-\beta_{1,t})} \frac{(\hat{m}_{t,i})^2}{\sqrt{\hat{v}_{t,i}}}. \quad (30)$$

Further, it can be written as

$$g_{t,i}(\theta_{t,i} - \theta_{i,i}^*) = \frac{(1-\beta_1^t)\sqrt{\hat{v}_{t,i}}}{2\alpha_t \xi_{t,i}(1-\beta_{1,t})} \left((\theta_{t,i} - \theta_{i,i}^*)^2 - (\theta_{t+1,i} - \theta_{i,i}^*)^2 \right) + \sqrt{\frac{\beta_{1,t}}{\alpha_{t-1}(1-\beta_{1,t})}} (\theta_{i,i}^* - \theta_{t,i})^2 \sqrt{\hat{v}_{t-1,i}} \sqrt{\frac{\beta_{1,t}\alpha_{t-1}(m_{t-1,i})^2}{(1-\beta_{1,t})\sqrt{\hat{v}_{t-1,i}}}} + \frac{\alpha_t(1-\beta_1^t)\xi_{t,i}}{2(1-\beta_{1,t})} \frac{(\hat{m}_{t,i})^2}{\sqrt{\hat{v}_{t,i}}} \quad (31)$$

Based on Young's inequality, $ab \leq a^2/2 + b^2/2$ and fact that $\beta_{1,t} \leq \beta_1$, the above equation can be reordered as

$$\begin{aligned}
g_{t,i}(\theta_{t,i} - \theta_{*,i}^*) &\leq \frac{1}{2\alpha_t \xi_{t,i}(1-\beta_1)} \left((\theta_{t,i} - \theta_{*,i}^*)^2 \right. \\
&\quad \left. - (\theta_{t+1,i} - \theta_{*,i}^*)^2 \right) \sqrt{\hat{v}_{t,i}} \\
&\quad + \frac{\beta_{1,t}}{2\alpha_{t-1}(1-\beta_{1,t})} (\theta_{*,i}^* - \theta_{t,i})^2 \sqrt{\hat{v}_{t-1,i}} \\
&\quad + \frac{\beta_1 \alpha_{t-1} (m_{t-1,i})^2}{2(1-\beta_1) \sqrt{\hat{v}_{t-1,i}}} + \frac{\alpha_t \xi_{t,i}}{2(1-\beta_1)} \frac{(\hat{m}_{t,i})^2}{\sqrt{\hat{v}_{t,i}}}
\end{aligned} \tag{32}$$

From Eq. (17) and Fig. 2, it is clear that $0.5 \leq \xi_{t,i} \leq 1$. So, $\xi_{t,i}$ can be removed from last term of above equation and it still satisfy the inequality. Then,

$$\begin{aligned}
g_{t,i}(\theta_{t,i} - \theta_{*,i}^*) &\leq \frac{1}{2\alpha_t \xi_{t,i}(1-\beta_1)} \left((\theta_{t,i} - \theta_{*,i}^*)^2 \right. \\
&\quad \left. - (\theta_{t+1,i} - \theta_{*,i}^*)^2 \right) \sqrt{\hat{v}_{t,i}} \\
&\quad + \frac{\beta_{1,t}}{2\alpha_{t-1}(1-\beta_{1,t})} (\theta_{*,i}^* - \theta_{t,i})^2 \sqrt{\hat{v}_{t-1,i}} \\
&\quad + \frac{\beta_1 \alpha_{t-1} (m_{t-1,i})^2}{2(1-\beta_1) \sqrt{\hat{v}_{t-1,i}}} + \frac{\alpha_t}{2(1-\beta_1)} \frac{(\hat{m}_{t,i})^2}{\sqrt{\hat{v}_{t,i}}}
\end{aligned} \tag{33}$$

We use the Lemma 10.4 of Aadm [38] and derive the regret bound by aggregating it across all the dimensions for $i \in \{1, \dots, d\}$ and all the sequence of convex functions for $t \in \{1, \dots, T\}$ in the upper bound of $f_t(\theta_t) - f_t(\theta^*)$ as

$$\begin{aligned}
R(T) &\leq \sum_{i=1}^d \frac{1}{2\alpha_1 \xi_{1,i}(1-\beta_1)} (\theta_{1,i} - \theta_{*,i}^*)^2 \sqrt{\hat{v}_{1,i}} \\
&\quad + \sum_{i=1}^d \sum_{t=2}^T \frac{1}{2(1-\beta_1)} (\theta_{t,i} - \theta_{*,i}^*)^2 \left(\frac{\sqrt{\hat{v}_{t,i}}}{\alpha_t \xi_{t,i}} - \frac{\sqrt{\hat{v}_{t-1,i}}}{\alpha_{t-1} \xi_{t-1,i}} \right) \\
&\quad + \frac{\beta_1 \alpha G_\infty}{(1-\beta_1) \sqrt{1-\beta_2}(1-\gamma)^2} \sum_{i=1}^d \|g_{1:T,i}\|_2 \\
&\quad + \frac{\alpha G_\infty}{(1-\beta_1) \sqrt{1-\beta_2}(1-\gamma)^2} \sum_{i=1}^d \|g_{1:T,i}\|_2 \\
&\quad + \sum_{i=1}^d \sum_{t=1}^T \frac{\beta_{1,t}}{2\alpha_t(1-\beta_{1,t})} (\theta_{*,i}^* - \theta_{t,i})^2 \sqrt{\hat{v}_{t,i}}
\end{aligned} \tag{34}$$

By utilizing the assumptions that $\alpha = \alpha_t \sqrt{t}$, $\|\theta_t - \theta^*\|_2 \leq D$ and $\|\theta_m - \theta_n\|_\infty \leq D_\infty$, we can write as

$$\begin{aligned}
R(T) &\leq \frac{D^2}{2\alpha(1-\beta_1)} \sum_{i=1}^d \frac{\sqrt{T \hat{v}_{T,i}}}{\xi_{1,i}} \\
&\quad + \frac{\alpha(1+\beta_1)G_\infty}{(1-\beta_1) \sqrt{1-\beta_2}(1-\gamma)^2} \sum_{i=1}^d \|g_{1:T,i}\|_2 \\
&\quad + \frac{D_\infty^2}{2\alpha} \sum_{i=1}^d \sum_{t=1}^t \frac{\beta_{1,t}}{(1-\beta_{1,t})} \sqrt{t \hat{v}_{t,i}} \\
&\leq \frac{D^2}{2\alpha(1-\beta_1)} \sum_{i=1}^d \frac{\sqrt{T \hat{v}_{T,i}}}{\xi_{1,i}} \\
&\quad + \frac{\alpha(1+\beta_1)G_\infty}{(1-\beta_1) \sqrt{1-\beta_2}(1-\gamma)^2} \sum_{i=1}^d \|g_{1:T,i}\|_2 \\
&\quad + \frac{D_\infty^2 G_\infty \sqrt{1-\beta_2}}{2\alpha} \sum_{i=1}^d \sum_{t=1}^t \frac{\beta_{1,t}}{(1-\beta_{1,t})} \sqrt{t}
\end{aligned} \tag{35}$$

It is shown in Adam [38] that $\sum_{t=1}^t \frac{\beta_{1,t}}{(1-\beta_{1,t})} \sqrt{t} \leq \frac{1}{(1-\beta_1)(1-\gamma)^2}$. Thus, the regret bound can be written as

$$\begin{aligned}
R(T) &\leq \frac{D^2}{2\alpha(1-\beta_1)} \sum_{i=1}^d \frac{\sqrt{T \hat{v}_{T,i}}}{\xi_{1,i}} \\
&\quad + \frac{\alpha(1+\beta_1)G_\infty}{(1-\beta_1) \sqrt{1-\beta_2}(1-\gamma)^2} \sum_{i=1}^d \|g_{1:T,i}\|_2 \\
&\quad + \sum_{i=1}^d \frac{D_\infty^2 G_\infty \sqrt{1-\beta_2}}{2\alpha(1-\beta_1)(1-\lambda)^2}
\end{aligned} \tag{36}$$

We know $\xi_{1,i} = 1/(1 + e^{-|g_{0,i} - g_{1,i}|}) = 1/(1 + e^{-|g_{1,i}|})$ as $g_{0,i} = 0$. Therefore, the regret bound for diffGrad is as follows:

$$\begin{aligned}
R(T) &\leq \frac{D^2}{2\alpha(1-\beta_1)} \sum_{i=1}^d (1 + e^{-|g_{1,i}|}) \sqrt{T \hat{v}_{T,i}} \\
&\quad + \frac{\alpha(1+\beta_1)G_\infty}{(1-\beta_1) \sqrt{1-\beta_2}(1-\gamma)^2} \sum_{i=1}^d \|g_{1:T,i}\|_2 \\
&\quad + \sum_{i=1}^d \frac{D_\infty^2 G_\infty \sqrt{1-\beta_2}}{2\alpha(1-\beta_1)(1-\lambda)^2}
\end{aligned} \tag{37}$$

□



Shiv Ram Dubey has been with the Indian Institute of Information Technology (IIIT), Sri City since June 2016, where he is currently the Assistant Professor of Computer Science and Engineering. He received the Ph.D. degree in Computer Vision and Image Processing from Indian Institute of Information Technology, Allahabad (IIIT Allahabad) in 2016. Before that, from August 2012-Feb 2013, he was a Project Officer in the Computer Science and Engineering Department at Indian Institute of Technology, Madras (IIT Madras). He was a recipient of several awards including the Indo-Taiwan Joint Research Grant from DST/GITA, Govt. of India, Best PhD Award in PhD Symposium, IEEE-CICT2017 at IIITM Gwalior, Early Career Research Award from SERB, Govt. of India and NVIDIA GPU Grant Award Twice from NVIDIA. He received Outstanding Certificate of Reviewing Award from Information Fusion, Elsevier in 2018. He also received the Best Paper Award in IEEE UPCON 2015, a prestigious conference of IEEE UP Section. His research interest includes Computer Vision, Deep Learning, Image Processing, Biometrics, Medical Imaging, Convolutional Neural Networks, Image Feature Description, Content Based Image Retrieval, Image-to-Image Transformation, Face Detection and Recognition, Facial Expression Recognition, Texture and Hyperspectral Image Analysis.

of the IEEE Uttar Pradesh Section from 2014. Currently he is serving as secretary Signal Processing Society Chapter, and Uttar Pradesh Section as well. He is serving as an Editorial Board Member and reviewer for many international journals. His current research interests are in the areas of digital image processing, pattern recognition, multimedia data indexing and retrieval, watermarking and biometrics.



Satish Kumar Singh is an Associate Professor with the Indian Institute of Information Technology Allahabad, Prayagraj, India. He received the Ph.D., M.Tech., and B.Tech. degrees in 2010, 2005, and 2003, respectively. He has over 13 years of experience in academic and research institutions. He has authored over 45 publications in reputed international journal and conference proceedings. He is a member of various professional societies, like the Institution of Electronics and Telecommunication Engineers. He is an Executive Committee Member

of the IEEE Uttar Pradesh Section from 2014. Currently he is serving as secretary Signal Processing Society Chapter, and Uttar Pradesh Section as well. He is serving as an Editorial Board Member and reviewer for many international journals. His current research interests are in the areas of digital image processing, pattern recognition, multimedia data indexing and retrieval, watermarking and biometrics.



Soumendu Chakraborty received his Bachelor of Engineering (B.E.) in Information Technology from University Institute of Technology, University of Burdwan, India in 2005. He did his M.Tech. in Computer Science and Engineering from GLA University, Mathura, India in 2013. He received his Ph.D. from Indian Institute of Information Technology, Allahabad, U.P., India in the year 2018. He has 12 years of teaching and research experience. Presently, he is working as an Assistant Professor in Indian Institute of Information Technology, Lucknow, India.

His research interests include Computer Vision, Machine Learning, image processing, biometric systems, image steganography and pattern recognition.



Swalpa Kumar Roy received both his Bachelor and Master degree in Computer Science & Engineering from West Bengal University of Technology, Kolkata, India, in 2012, and Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India, in 2015, respectively. He is pursuing the Ph.D. degree from under Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India, where he also worked as a Project Linked Person from July 2015 to March 2016. He is currently working as an Assistant Professor in

the Department of Computer Science & Engineering, Jalpaiguri Government Engineering College, Jalpaiguri, West Bengal, India. His research interests include Computer Vision, Deep Learning, Remote Sensing, Texture Feature Description, and Fractal Image Coding.



Snehasis Mukherjee has obtained his PhD in Computer Science from the Indian Statistical Institute in 2012. Before doctoral study, he has completed his Bachelors degree in Mathematics from the University of Calcutta and Masters degree in Computer Applications from the Vidyasagar University. He did his Post Doctoral Research works at the National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, USA. Currently he is working as an Assistant Professor in the Indian Institute of Information Technology Chittoor, SriCity (IIIT

Chittoor, SriCity). He has written several peer-reviewed research papers (in reputed journals and conferences). His research area includes Computer Vision, Machine Learning, Image and Video Processing.



Bidyut Baran Chaudhuri received the Ph.D. degree from IIT Kanpur, in 1980. He was a Leverhulme Postdoctoral Fellow with Queens University, U.K., from 1981 to 1982. He joined the Indian Statistical Institute, in 1978, where he worked as an INAE Distinguished Professor and a J C Bose Fellow at Computer Vision and Pattern Recognition Unit of Indian Statistical Institute. He is now affiliated to Techno India University, Kolkata as Pro-Vice Chancellor (Academic). His research interests include Pattern Recognition, Image Processing, Computer

Vision, Natural Language Processing (NLP), Signal processing, Digital Document Processing, Deep learning etc. He pioneered the first workable OCR system for printed Indian scripts Bangla, Assamese and Devnagari. He also developed computerized *Bharati Braille system* with speech synthesizer and has done statistical analysis of Indian language. He has published about 425 research papers in international journals and conference proceedings. Also, he has authored/edited seven books in these fields. Prof Chaudhuri received Leverhulme fellowship award, Sir J. C. Bose Memorial Award, M. N. Saha Memorial Award, Homi Bhabha Fellowship, Dr. Vikram Sarabhai Research Award, C. Achuta Menon Award, Homi Bhabha Award: Applied Sciences, Ram Lal Wadhwa Gold Medal, Jawaharlal Nehru Fellowship, J C Bose fellowship, Om Prakash Bhasin Award etc. Prof Chaudhuri is the associate editor of three international journals and a fellow of INSA, NASI, INAE, IAPR, The World Academy of Sciences (TWAS) and life fellow of IEEE (2015). He acted as General Chair and Technical Co-chair at various International Conferences.