

ADAMT: A STOCHASTIC OPTIMIZATION WITH TREND CORRECTION SCHEME

Anonymous authors

Paper under double-blind review

ABSTRACT

Adam-typed optimizers, as a class of adaptive moment estimation methods with the exponential moving average scheme, have been successfully used in many applications of deep learning. Such methods are appealing for capability on large-scale sparse datasets. On top of that, they are computationally efficient and insensitive to the hyper-parameter settings. In this paper, we present a new framework for adapting Adam-typed methods, namely *AdamT*. Instead of applying a simple exponential weighted average, AdamT also includes the trend information when updating the parameters with the adaptive step size and gradients. The newly added term is expected to efficiently capture the non-horizontal moving patterns on the cost surface, and thus converge more rapidly. We show empirically the importance of the trend component, where AdamT outperforms the conventional Adam method constantly in both convex and non-convex settings.

1 INTRODUCTION

Employing first order optimization methods, such as stochastic gradient descent (SGD), is a key of solving large-scale problems. The classic gradient descent algorithm is widely used to update the model parameters, denoted by x ,

$$x_{t+1} = x_t - \eta \nabla f(x_t), \quad (1)$$

where the gradient is denoted by $\nabla f(x_t)$ and the step size by η . While the method has shown its efficiency for many contemporary tasks, the adaptive variants of SGD outperform the vanilla SGD methods on their rapid training time. Specifically, the step size η is substituted by an adaptive step size $\eta/\sqrt{v_t}$, and v_t is generated from the squared gradient $[\nabla f(x_t)]^2$.

Several variants of the popular adaptive optimizers can be summarized into such common format. These optimizers share gradients calculation and parameters updating functions, but specify different moving average schemes for calculating the parameter-wise adaptive learning rate v_t . For example, AdaGrad (Duchi et al., 2011) takes the arithmetic average of historical squared gradients $[\nabla f(x_t)]^2$. Compared with the conventional momentum method, it adapts the learning rate to each parameter to suit the sparse data structure, and thus gains a rapid convergence speed (Ruder, 2016). Later, Tieleman & Hinton (2012) proposed RMSProp to reduce the aggressiveness of the decay rate in AdaGrad. The method modifies v_t to the exponentially decayed squared gradients. Similar implementations could also be found in ADADELTA (Zeiler, 2012). Instead of the squared gradients, the method applies squared parameter updates to define the adaptive learning rate. As a result, each update guarantees the same hypothetical units as the parameter. Later, Adam (Kingma & Ba, 2015) modifies RMSProp with the idea from momentum methods (Qian, 1999). Except for the second moment moving average, the new rule also replaces the gradient $\nabla f(x_t)$ at the end of the Equation (1) to the first-moment estimation. The method has practically shown its superiority regarding the converge speed and memory requirement. While the aforementioned methods are the most famous frameworks, there are also many variants for each of them. The examples include NAdam (Dozat, 2016), AMSGrad (Reddi et al., 2018) and Adafom (Chen et al., 2019).

So far, the adaptive methods with exponential moving average gradients have gained great attention with huge success in many deep learning tasks. However, it remains unsolved whether the simple exponential smoothing results or the level information is sufficient in capturing the landscape of the cost surface. When clear upward or downward pattern could be recognized within the moving routine, it is suggested to add a trend term on top of the single level information.

In this paper, we modify the Adam rule with trend-corrected exponential smoothing schemes, namely *AdamT*, to obtain the local minima with a faster speed. To the best of our knowledge, our research is the first to apply the trend-corrected features on gradients scaling and parameters updating. It shall be emphasized that our framework is universally implementable for all adaptive update methods that apply the exponential average term, including but not restricted to ADADELTA, RM-SProp, AdaMAX and other well-recognized methods. For the sake of conciseness, in this specific paper, we focus on Adam regarding rule modification and performance comparison.

Our contributions in this paper could be summarized in three-fold:

1. We propose the notion of trend corrected exponential smoothing to modify the conventional application of exponential moving average in optimizers with adaptive gradients. Our *AdamT* method collaborates the trend information into the update rule of Adam.
2. We show the conditions for the method to converge in convex settings. The regret bound is in consistent to Adam at $\mathcal{O}(\sqrt{T})$.
3. We demonstrate AdamT’s convergence in both convex and non-convex settings. The performance is compared with Adam, where AdamT shows clear superiority on both the training set and the test set, especially for non-convex problems.

For the remainder of the paper, we present the fundamental idea of Adam and Holt’s linear methods in Section 2. In Section 3 and 4, we detail the update rules and experimental analysis, respectively. In addition, Section 5 reviews recent developments of Adam-typed optimizers. While many of them focus more on non-convex optimizations, there is a potential to incorporate our methods with such frameworks and this extension is expected for future settings.

2 PRELIMINARY

2.1 ADAPTIVE METHODS WITH EXPONENTIAL MOVING AVERAGES

For adaptive gradient methods, we apply the update rule at step $t + 1$

$$x_{t+1} = x_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t.$$

where m_t is the gradient updates, and conventionally it is defined as the last gradient value $\nabla f(x_t)$. To prevent zero division, a smoothing term ϵ is included on the denominator.

In this paper, we focus our analysis and modifications on Adam. The method was initially proposed by Kingma & Ba (2015), and quickly becomes one of the most popular optimizers in the last few years. The adaptive step size is accelerated from the previous square gradients

$$v_t = \alpha v_{t-1} + (1 - \alpha) [\nabla f(x_t)]^2. \quad (2)$$

In terms of the gradient m_t , Adam takes the exponentially weighted average of all previous gradients instead of solely relying on the last gradient ∇f_t

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) \nabla f(x_t). \quad (3)$$

While the two moment estimates from Equations (2) & (3) could potentially counteract towards zero, the series \hat{m}_t and \hat{v}_t are considered for bias-correction. Formally, the rules are defined as:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}; \quad \hat{v}_t = \frac{v_t}{1 - \beta_2^t}.$$

2.2 TREND CORRECTED EXPONENTIAL SMOOTHING

The idea of extracting a smoothed new point from all the previous information is called the exponential weighted moving average. Holt (2004) extended the method by including the trend behaviours within the series, namely trend corrected exponential smoothing or Holt’s linear method. Consider a time series $\{y_t\}$ for $t = 1, 2, \dots$. Our target is to find the smoothing results at step t . We denote the

smoothing result as $\widehat{y_{t+1}|t}}$. Holt’s linear method formulates the conditional forecasting by summing two smoothing equations up

$$\begin{aligned}\widehat{y_{t+1}|t} &= \ell_t + b_t \\ \ell_t &= \beta(\ell_{t-1} + b_{t-1}) + (1 - \beta)y_t \\ b_t &= \gamma b_{t-1} + (1 - \gamma)(\ell_t - \ell_{t-1}).\end{aligned}$$

For a new estimation, we first update the level term ℓ_t with the weighted average of the last observation y_t and its estimation $\widehat{y_{t+1}|t}}$. The trend term b_t is updated simultaneously as the weighted average of the estimated trend $\ell_t - \ell_{t-1}$ and its previous estimation b_{t-1} . The smoothing parameters for the level and the trend are denoted as α and β . Both values could be selected between 0 and 1.

Gardner Jr & McKenzie (1985) also suggested adding a damping factor ϕ , so that

$$\begin{aligned}\widehat{y_{t+1}|t} &= \ell_t + \phi b_t \\ \ell_t &= \beta(\ell_{t-1} + \phi b_{t-1}) + (1 - \beta)y_t \\ b_t &= \gamma \phi b_{t-1} + (1 - \gamma)(\ell_t - \ell_{t-1}).\end{aligned}$$

The damped method is identical to Holt’s Linear method with $\phi = 1$, and is the same as simple exponential moving average method with $\phi = 0$. When ϕ is positive, the parameter could be used to control the significance of the trend component.

The damped trend methods are considerably popular for forecasting tasks (Hyndman & Athanassopoulos, 2018). Such methods inherent both level and trend information from historical series, while stay flexible enough to adjust the influence of the trend term via ϕ . On top of that, involving the damped factor could to some extent reduce the volatility of the smoothed line.

3 METHODOLOGY

We introduce our proposed algorithm *AdamT*, which is based on Adam (Kingma & Ba, 2015) with added Holt’s linear trend information for both of the first moment estimate and the second raw moment estimate. Specifically, we use trend-corrected exponential weighted moving averages in the final parameter update step instead of the level-only estimates used in Adam.

3.1 ALGORITHM

Consider the gradient of a stochastic objective function $f(x)$ evaluated at T iterations as a time series $\{\nabla f(x_t)\}$ for $t = 1, 2, \dots, T$. According to the Holt’s linear trend method illustrated in Section 2.2, we write two series $\{\ell_t^m\}$ and $\{b_t^m\}$ as the exponential weighted moving averages which estimate the level and trend information of the first moment $\nabla f(x_t)$:

$$\begin{aligned}\ell_t^m &= \beta_1 m_{t-1} + (1 - \beta_1) \nabla f(x_t) \\ b_t^m &= \gamma_1 \phi_1 b_{t-1}^m + (1 - \gamma_1)(\ell_t^m - \ell_{t-1}^m) \\ m_t &= \ell_t^m + \phi_1 b_t^m\end{aligned}$$

where β_1, γ_1 and ϕ_1 have the same functionality as explained in Section 2.2 and these are regarded as hyperparameters in our algorithm. Equation (3.1) combines the level and the trend information of first moment, which will be used for calculating the final update rule and the trend-corrected level estimates. The procedures for the second raw moment $\nabla f(x_t) \circ \nabla f(x_t)$ is analogous:

$$\begin{aligned}\ell_t^v &= \beta_2 v_{t-1} + (1 - \beta_2) \nabla f(x_t) \circ \nabla f(x_t) \\ b_t^v &= \gamma_2 \phi_2 b_{t-1}^v + (1 - \gamma_2)(\ell_t^v - \ell_{t-1}^v) \\ v_t &= \ell_t^v + \phi_2 b_t^v\end{aligned}$$

where the operation “ \circ ” denotes an element-wise multiplication. The hyperparameters β_2, γ_2 and ϕ_2 here share the same corresponding meanings as before. The moving averages $\{\ell_t^v\}$ and $\{b_t^v\}$ estimate the level and trend of the second raw moment respectively. The term v_t combines these two information, which will be used in the calculations of final update rule and trend-corrected level estimates of the second raw moment.

In our algorithm, we set the initial values of the series $\{\ell_t^m\}$, $\{\ell_t^v\}$, $\{b_t^m\}$ and $\{b_t^v\}$ to be zero vectors, so that $\ell_0^m = \ell_0^v = b_0^m = b_0^v = 0$. The series $\{m_t\}$ and $\{v_t\}$, as a result, are also initialized as zero vectors. As observed in Kingma & Ba (2015), the exponential weighted moving averages could bias towards zero, especially during the early training stage. We perform the bias correction for the two level estimates $\{\ell_t^m\}$ and $\{\ell_t^v\}$ by following Kingma & Ba (2015). For the two trend estimates $\{b_t^m\}$ and $\{b_t^v\}$, we correct the bias in a different way by taking into account the effect of damping parameters (ϕ_1, ϕ_2) . Thus, the bias-corrected version of the series $\{m_t\}$ and $\{v_t\}$ can be written as:

$$\begin{aligned}\hat{m}_t &= \frac{\ell_t^m}{1 - \beta_1^t} + \frac{(1 - \gamma_1 \phi_1) b_t^m}{(1 - \gamma_1)(1 - (\gamma_1 \phi_1)^t)} \\ \hat{v}_t &= \frac{\ell_t^v}{1 - \beta_2^t} + \frac{(1 - \gamma_2 \phi_2) b_t^v}{(1 - \gamma_2)(1 - (\gamma_2 \phi_2)^t)}\end{aligned}$$

The justification for the two bias-corrected trend estimates is provided in Appendix A. Similar to Adam, we consider the adaptive update rule with the bias-corrected first moment estimate and the second raw moment estimate:

$$x_{t+1} = x_t - \frac{\eta}{\sqrt{|\hat{v}_t|} + \epsilon} \hat{m}_t \quad (4)$$

where ϵ is a positive tiny number added in the denominator to avoid zero-division case. Please note that the series $\{\hat{m}_t\}$ and $\{\hat{v}_t\}$ in AdamT are different from that of Adam. The two series are trend-corrected (also bias-corrected) estimates of both moments.

The direction of the effective step $\Delta_t = \eta \cdot \hat{m}_t / \sqrt{|\hat{v}_t|}$ (with $\epsilon = 0$) in the parameter space depends on the joint effect of the first moment level and trend estimates. In the update rule (4), we only care about the magnitude of \hat{v}_t by taking the absolute value and thus the ratio $\hat{m}_t / \sqrt{|\hat{v}_t|}$ can be seen as a signal-to-noise ratio. Note that the effective step Δ_t in our algorithm is also invariant to the scale of the gradients. Specifically, re-scaling the gradients $\nabla f(x_t)$ with a factor c will scale ℓ_t^m and b_t^m by a factor c , and will scale ℓ_t^v and b_t^v by a factor c^2 . This results in scaling \hat{m}_t and \hat{v}_t by a factor c and c^2 respectively, and finally cancel out in the parameter update rule $(c \cdot \hat{m}_t) / (\sqrt{c^2 \cdot \hat{v}_t}) = \hat{m}_t / \sqrt{|\hat{v}_t|}$. Note that our proposed method AdamT has two extra computational steps, that is Equations (3.1) & (3.1). However, the computational complexity of these two steps is almost linear in time. Therefore, we can conclude that AdamT yields a superior performance compared with Adam (the results will be shown in the experiment section) with a minimal additional computational cost.

In our algorithm, we set the hyperparameters $\beta_1, \gamma_1, \beta_2, \gamma_2$ according to the suggestion in Kingma & Ba (2015). The smoothing parameters for the first moment estimates are set to 0.9, that is $\beta_1 = \gamma_1 = 0.9$, while the smoothing parameters for the second raw moment estimates are set to 0.999, that is $\beta_2 = \gamma_2 = 0.999$. We empirically find that the good default values of the two damping parameters can be set to $\phi_1 = \phi_2 = 0.5$. The pseudo-code of our AdamT is provided in Algorithm 1.

3.2 REGRET BOUND

We present the key results of convergence analysis for AdamT in this section and the details is presented in Appendix B.

Theorem 3.1. *Assume that the objective function f_t has bounded gradients, that is $\|\nabla f_t(x)\|_2 \leq G$, $\|\nabla f_t(x)\|_\infty \leq G_\infty$ for all $x \in \mathbb{R}^d$. Assume that the distance between any x_t produced by AdamT is bounded, that is $\|x_n - x_m\|_2 \leq D$, $\|x_m - x_n\|_\infty \leq D_\infty$ for any $m, n \in \{1, 2, \dots, T\}$, and $\beta_1, \gamma_1, \beta_2, \gamma_2 \in [0, 1)$, $\phi_1, \phi_2 \in [0, 1]$. AdamT achieves the following guarantee for all $T \geq 1$*

$$\begin{aligned}R(T) &\leq \frac{D_\infty^2}{2\eta(1 - \beta_1)} \sum_{i=1}^d \sqrt{T \hat{v}_{T,i}} + \frac{1}{2} [\beta_1(1 - \beta_1)(1 + \phi_1(1 - \gamma_1))^2] G_\infty^2 \\ &+ \frac{D_\infty^2 2(1 - \gamma_1 \phi_1)^3}{(1 - \beta_1)((1 - \gamma_1)(1 - (\gamma_1 \phi_1)^T))^2} + \frac{\eta}{2(1 - \beta_1)} \frac{(2 - \beta_1)^2}{\sqrt{2 - \beta_2}} \sum_{i=1}^d \|g_{1:T,i}\|_2 + dD_2^2 \frac{(1 - \gamma_1)\beta_1 + 1 - \gamma_1\phi_1}{2(1 - \gamma_1)(1 - \beta_1)}.\end{aligned}$$

This result implies that AdamT has $\mathcal{O}(\sqrt{T})$ regret bound given that $\sum_{i=1}^d \|g_{1:T,i}\|_2 \leq dG_\infty \sqrt{T}$ and $\sum_{i=1}^d \sqrt{T \hat{v}_{T,i}} \leq dCG_\infty \sqrt{T}$ for some positive constant C . Hence, we can prove that the average regret of AdamT converges,

Algorithm 1 *AdamT*. The **Adam** optimizer modified with Holt’s Linear Trend method. Empirically suggested default values for the hyperparameters are $\beta_1 = \gamma_1 = 0.9, \beta_2 = \gamma_2 = 0.999, \phi_1 = \phi_2 = 0.5, \eta = 0.0001$. All of the operations on vectors are element-wise.

```

1: Require:  $\eta$ : Learning rate
2: Require:  $\beta_1, \gamma_1, \beta_2, \gamma_2 \in [0, 1)$ : Smoothing hyperparameters
3: Require:  $\phi_1, \phi_2 \in [0, 1]$ : Damping hyperparameters
4: Require:  $f(x)$ : Noisy objective function with parameters  $x$ 
5: Initialize:  $x_1$ : Initial parameter values
6:            $\ell_0^m \leftarrow 0$ : Initial first moment level estimate
7:            $\ell_0^v \leftarrow 0$ : Initial second raw moment level estimate
8:            $b_0^m \leftarrow 0$ : Initial first moment trend estimate
9:            $b_0^v \leftarrow 0$ : Initial second raw moment trend estimate
10: for  $t = 1, 2, \dots, T$  do
11:    $\ell_t^m \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) \nabla f(x_t)$ 
12:    $b_t^m \leftarrow \gamma_1 \phi_1 b_{t-1}^m + (1 - \gamma_1) (\ell_t^m - \ell_{t-1}^m)$ 
13:    $\ell_t^v \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) \nabla f(x_t) \circ \nabla f(x_t)$ 
14:    $b_t^v \leftarrow \gamma_2 \phi_2 b_{t-1}^v + (1 - \gamma_2) (\ell_t^v - \ell_{t-1}^v)$ 
15:    $\hat{m}_t \leftarrow \ell_t^m / (1 - \beta_1^t) + [(1 - \gamma_1 \phi_1) b_t^m] / [(1 - \gamma_1)(1 - (\gamma_1 \phi_1)^t)]$ 
16:    $\hat{v}_t \leftarrow \ell_t^v / (1 - \beta_2^t) + [(1 - \gamma_2 \phi_2) b_t^v] / [(1 - \gamma_2)(1 - (\gamma_2 \phi_2)^t)]$ 
17:    $x_{t+1} \leftarrow x_t - \eta \hat{m}_t / (\sqrt{|\hat{v}_t|} + \epsilon)$   $\triangleright \epsilon = 1e - 8$ 
18: end for
19: return  $x_t$ 

```

Corollary 3.1.1. *By following the assumptions in Theorem 3.1, for all $T \geq 1$ AdamT achieves*

$$\frac{R(T)}{T} = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$$

This result follows immediately from Theorem 3.1 and thus $\lim_{T \rightarrow \infty} \frac{R(T)}{T} = 0$.

4 EXPERIMENTS

We evaluate the proposed algorithm *AdamT* on both convex and non-convex real-world optimization problems with several popular types of machine learning models. The models we considered in the experiments include logistic regression which has a well-known convex loss surface, and different neural network models, including feedforward neural networks, convolutional neural networks and variational autoencoder. Neural Networks with non-linear activation function typically have an inherent non-convex loss surface which is more challenging for an optimization method.

We compare our method with Adam (Kingma & Ba, 2015) and demonstrate the effectiveness of the trend information of the gradients infused in AdamT. The experiment results show that our method has a converge faster to reach a better minimum point than Adam. The observation evidences that the added trend information effectively helps AdamT to better capture the landscape of loss surface.

In the following experiments, we prepare the same set of initial values for the models, so that the initial model losses (the loss value at epoch = 0) are identical for all the optimization methods. In terms of the hyperparameters, all the smoothing parameters (β_1, β_2 in Adam and $\beta_1, \beta_2, \gamma_1, \gamma_2$ in AdamT) are set at their corresponding default values which are provided in Algorithm 1. The damping factors (ϕ_1, ϕ_2) and the learning rate η are tuned through a dense grid search to produce the best results for both of the optimizers. All the experiments and optimizers are written in PyTorch.

4.1 LOGISTIC REGRESSION FOR FASHION-MNIST

We first evaluate AdamT on the logistic regression for multi-class classification problem with Fashion-MNIST dataset (Xiao et al., 2017) which is a MNIST-like dataset of fashion products. The dataset has 60,000 training samples and 10,000 testing samples. Each of them has 28×28 pixels. Each of the samples is classified into one of the 10 fashion products. The cross-entropy loss

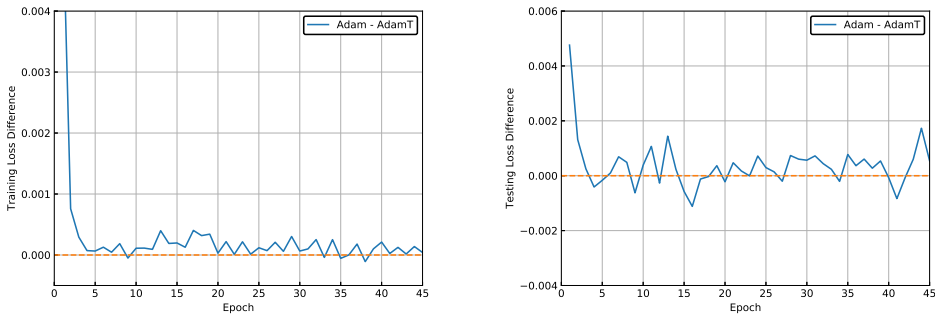


Figure 1: Training loss difference (left) and test loss difference (right) of the logistic regression on Fashion-MNIST dataset for classification task.

function has a well-behaved convex surface. The learning rate η for both Adam and AdamT is set to be constant during the training procedure. We use minibatch training with size set to 128.

The results are reported in Figure 1. Since the superiority of our method over Adam is relatively small in this experiment, the plot of loss value against epoch cannot visualize the difference. Instead, we plot the loss difference of the two optimizers, which is $(Loss_{Adam} - Loss_{AdamT})$ against training epoch. The difference above zero reflect the advantage of AdamT. Figure 1 indicates that AdamT converges faster at the early training stage and constantly outperforms Adam during the rest of the training phase, though the advantage is relatively small in this experiment. The AdamT gives a similar (slightly better) performance on the test dataset as Adam. The loss surface of logistic regression is convex and well-behaved so that the trend information of AdamT cannot further provide much useful information for optimization, which results in a small advantage in this experiment.

4.2 FEEDFORWARD NEURAL NETWORKS FOR SVHN

To investigate the performance on non-convex objective functions, we conduct the experiment with feedforward neural networks on *The Street View House Numbers* (SVHN) dataset (Netzer et al., 2011) for a digit classification problem. We pre-process this RGB image dataset into grayscale for dimension reduction by taking the average across the channels for each pixel in the image. The samples are 32×32 grayscale images. The neural network used in this experiment has two fully-connected hidden layers, each of which has 1,400 hidden units and ReLU activation function is used for the two hidden layers. We use softmax cross-entropy loss function for training the model.

To evaluate the performance of the optimizers in noisy settings, we apply a stochastic regularization method in the model for a separate experiment. Specifically, we include two dropout layers (Srivastava et al., 2014), where one is applied between the two hidden layers and the other one is used before the output layer. The dropout probability is set to 0.5 for both of the two dropout layers. In the experiments, we use a constant learning rate η and minibatch training with size set to 128.

We examine the training and test loss of the models with and without dropout layers for the two optimizers. According to Figure 2, we find that AdamT outperforms Adam significantly. In terms of the training process, AdamT yields a faster convergence and reaches a better position than Adam for the models, both with and without dropout layers. The superior performance of AdamT is also shown in the test phase, which demonstrates that AdamT also has a better generalization ability than Adam. For the model without dropout, our method performs on a par with Adam on the test dataset. Comparing to logistic regression, the loss surface in this experiment becomes complex and non-convex. The trend estimates of the gradients from AdamT can provide more meaningful information of the landscape of the loss surface, and it encourages a better performance on AdamT.

4.3 CONVOLUTIONAL NEURAL NETWORKS FOR CIFAR-10

Convolutional neural network (CNN) is the main workhorse for Computer Vision tasks. We train a CNN model on standard CIFAR-10 dataset for a multi-class classification task. The dataset contains

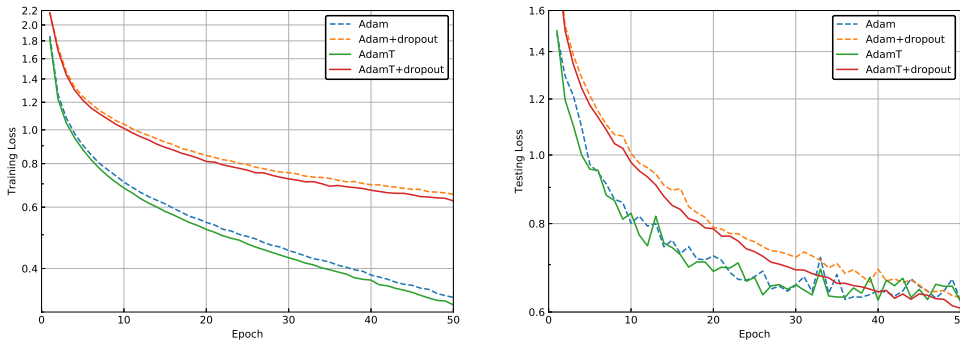


Figure 2: Training loss (left) and test loss (right) of the feedforward neural network on SVHN dataset for a classification problem. The model architecture is fc1400-fc1400.

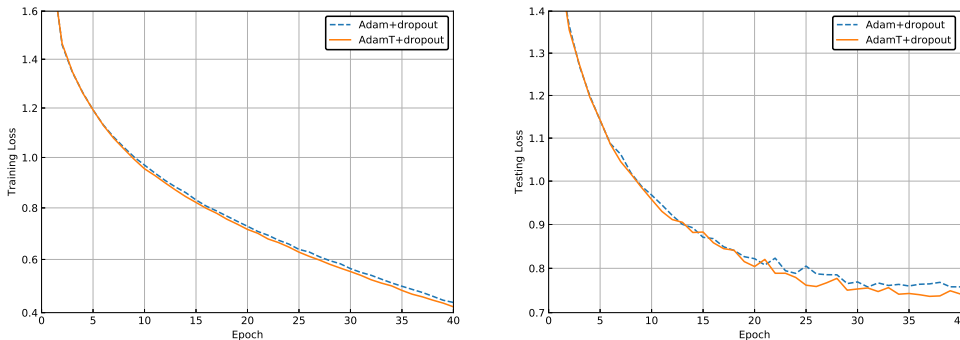


Figure 3: Training loss (left) and test loss (right) of the convolutional neural network on CIFAR-10 dataset for a classification task. The model architecture is c64-c64-fc600.

50,000 training samples and 10,000 test samples, and each sample is an RGB 32×32 image. We pre-process the dataset by normalizing the pixel values to the range $[-1, 1]$ for a more robust training. The CNN model employed in this experiment is similar to the model used in Reddi et al. (2018), which has the following architecture. There are 2 stages of alternating convolution and max pooling layers. Each convolution layer has 64 channels and kernel size 5×5 with stride 1. Each max pooling layer is applied with a kernel size of 2×2 . After that, there is a fully-connected layer with 600 hidden units and a dropout probability 0.5 followed by the output layer with 10 units. We use ReLU for the activation function and softmax cross-entropy for the loss function. The model is trained with a tuned constant learning rate and minibatch size 128 same as the previous experiments.

The experiment results are reported in Figure 3. We can observe that the proposed AdamT clearly excels Adam on the training loss, and this superiority translates into a more significant advantage of AdamT on the test loss, which again demonstrates a better generalization ability.

4.4 DEEP GENERATIVE MODELS FOR MNIST

Variational Autoencoder (VAE) (Kingma & Welling, 2014; Rezende et al., 2014) is one of the most popular deep generative models for density estimation and image generation. In this experiment, we train a VAE model on the standard MNIST dataset which contains 60,000 training samples and 10,000 test samples. Each sample is one 28×28 black-and-white image of the handwritten digit. The VAE model used in this experiment exactly matches the architecture presented in Kingma & Welling (2014): Gaussian encoder and Bernoulli decoder, both of which are implemented by feedforward neural networks with single hidden layer and there are 500 hidden units in each hidden layer. We employ the hyperbolic tangent activation function for the model and set the dimensionality of the latent space as 20. We use the constant learning rate and set the minibatch size to 150.

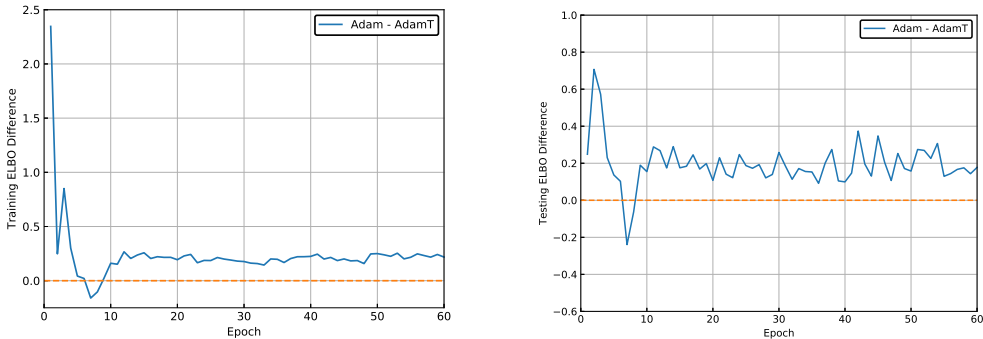


Figure 4: Training ELBO difference (left) and test ELBO difference (right) of the variational autoencoder on MNIST dataset.

We examine the Evidence Lower Bound (ELBO) of the training and testing phases for the two optimizers’ performance assessment. See Figure 4 for the experiment results. Due to the issue of different scales, we plot the difference between the ELBOs produced by the two optimizers. Similar to the first experiment, we plot the difference value ($ELBO_{Adam} - ELBO_{AdamT}$) against the epoch for training and testing. We observe that our AdamT has a much faster convergence at the early stage of training than Adam and constantly excels Adam during the rest of the training phase. The superior performance of AdamT in this experiment also translates into a clear advantage in the testing phase.

5 RELATED WORKS

We consider the class of adaptive moment estimation methods with exponential moving average scheme as Adam-type learning algorithms. The fundamental idea was proposed in Kingma & Ba (2015) and quickly extended to many variants. Some examples include AdaMax (Kingma & Ba, 2015), Nadam (Dozat, 2016) and AdamW (Loshchilov & Hutter, 2019).

Despite the efficiency in practice, the conventional Adam-type methods fail to guarantee global convergences. Reddi et al. (2018) discussed the problematic short-term memory of the gradients. For the convex settings, they proposed AMSGrad that promises a global optimization with a comparable performance. Except for some other recent studies for convex optimization (Xu et al., 2017; Chen et al., 2018; Levy et al., 2018), several works developed optimization methods for non-convex problems. Padam (Chen & Gu, 2018; Zhou et al., 2018) introduces a partial adaptive parameter p to interpolate between SGD with momentum and Adam so that adjacent learning rates could decrease smoothly. AdaUSM (Zou & Shen, 2018) appends the idea of unified momentum for non-decreasing sequence of weights. AdaFom (Chen et al., 2019) obtains first-order stationary by taking simple average on the second moment estimation. More conditions for pursuing global convergence were summarized in Zou et al. (2019), basing on the currently successful variants.

6 DISCUSSION AND CONCLUSION

In this work, we have modified the scheme to calculate the adaptive step size from exponential moving average to trend-corrected exponential smoothing. Empirical results demonstrate that our method, AdamT, works well in practice and constantly beats the baseline method Adam.

We leave some potentials for future developments. First, although we focused primarily on ADAM for theoretical and experimental analysis, we believe that similar ideas could also extend to other adaptive gradient methods, such as RMSProp (Tieleman & Hinton, 2012) and AMSGrad (Reddi et al., 2018). Also, this work, the same as the original ADAM method, relies on the theoretical assumption of convex problems settings. We have demonstrated its computational ability on the non-convex settings, and it is possible to extend the theoretical framework to non-convex scenarios. Some potential candidates in the latest research are listed in Section 5.

REFERENCES

- Jinghui Chen and Quanquan Gu. Closing the generalization gap of adaptive gradient methods in training deep neural networks. *preprint arXiv:1806.06763*, 2018.
- Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of ADAM-type algorithms for non-convex optimization. In *Proceedings of International Conference on Learning Representation (ICLR)*, 2019.
- Zaiyi Chen, Yi Xu, Enhong Chen, and Tianbao Yang. Sadagrad: Strongly adaptive stochastic gradient methods. In *International Conference on Machine Learning*, pp. 912–920, 2018.
- Timothy Dozat. Incorporating Nesterov momentum into ADAM. In *Proceedings of 4th International Conference on Learning Representations, Workshop Track*, 2016.
- John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.
- Everette S Gardner Jr and ED McKenzie. Forecasting trends in time series. *Management Science*, 31(10):1237–1246, 1985.
- Charles C Holt. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1):5–10, 2004.
- Rob J Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 2018.
- Diederik P Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. In *Proceedings of International Conference on Learning Representation (ICLR)*, 2015.
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of International Conference on Learning Representations*, 2014.
- Yehuda Kfir Levy, Alp Yurtsever, and Volkan Cevher. Online adaptive methods, universality and acceleration. In *Advances in Neural Information Processing Systems*, pp. 6500–6509, 2018.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of 4th International Conference on Learning Representations*, 2019.
- Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.
- Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151, 1999.
- Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of ADAM and beyond. In *Proceedings of International Conference on Learning Representation (ICLR)*, 2018.
- Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.
- Sebastian Ruder. An overview of gradient descent optimization algorithms. *preprint arXiv:1609.04747*, 2016.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2): 26–31, 2012.
- Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *preprint arXiv:1708.07747*, 2017.

Yi Xu, Qihang Lin, and Tianbao Yang. Stochastic convex optimization: Faster local growth implies faster global convergence. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3821–3830. JMLR. org, 2017.

Matthew D Zeiler. Adadelta: an adaptive learning rate method. *preprint arXiv:1212.5701*, 2012.

Dongruo Zhou, Yiqi Tang, Ziyang Yang, Yuan Cao, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *preprint arXiv:1808.05671*, 2018.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pp. 928–936, 2003.

Fangyu Zou and Li Shen. On the convergence of weighted AdaGrad with momentum for training deep neural networks. *preprint arXiv:1808.03408*, 2018.

Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of ADAM and RMSProp. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11127–11135, 2019.

A BIAS CORRECTION OF THE TREND ESTIMATES

To conduct the bias correction for the two trend estimates $\{b_t^m\}$ and $\{b_t^v\}$, we have to take into account the effect of the corresponding damping factors. Here, we give the justification for the trend estimate $\{b_t^m\}$ and the procedures for $\{b_t^v\}$ is analogous. Note that we can write the trend estimate b_t^m into the following compact summation form:

$$\begin{aligned} b_t^m &= \gamma_1 \phi_1 b_{t-1}^m + (1 - \gamma_1)(\ell_t^m - \ell_{t-1}^m) \\ &= (1 - \gamma_1) \sum_{i=1}^t (\gamma_1 \phi_1)^{t-i} (\ell_i^m - \ell_{i-1}^m) \end{aligned}$$

To find how the expectation of the trend estimates b_t^m relates to the expectation of the difference between the level estimates at successive timesteps $(\ell_t^m - \ell_{t-1}^m)$, we take the expectation for both sides of the above equation:

$$\begin{aligned} \mathbb{E}[b_t^m] &= \mathbb{E}[(1 - \gamma_1) \sum_{i=1}^t (\gamma_1 \phi_1)^{t-i} (\ell_i^m - \ell_{i-1}^m)] \\ &= (1 - \gamma_1) \sum_{i=1}^t (\gamma_1 \phi_1)^{t-i} \mathbb{E}[(\ell_i^m - \ell_{i-1}^m)] \\ &= \mathbb{E}[(\ell_t^m - \ell_{t-1}^m)] (1 - \gamma_1) \sum_{i=1}^t (\gamma_1 \phi_1)^{t-i} + \zeta \end{aligned}$$

where ζ can be considered as a small constant, since the factor $(\gamma_1 \phi_1)^{t-i}$ will be tiny if the associated expectation $\mathbb{E}[(\ell_i^m - \ell_{i-1}^m)]$ is too far away in the past in the case that $\mathbb{E}[(\ell_i^m - \ell_{i-1}^m)]$ is non-stationary. If $\mathbb{E}[(\ell_i^m - \ell_{i-1}^m)]$ is stationary, the constant ζ will be zero. To further simplify the above equation, we apply the formula for the sum of geometric sequence:

$$\mathbb{E}[b_t^m] = \mathbb{E}[(\ell_t^m - \ell_{t-1}^m)] (1 - \gamma_1) \left(\frac{1 - (\gamma_1 \phi_1)^t}{1 - \gamma_1 \phi_1} \right) + \zeta$$

This suggests that we can use the term $(1 - \gamma_1)[1 - (\gamma_1 \phi_1)^t]/[1 - \gamma_1 \phi_1]$ to correct the bias and close the discrepancy between the above two expectations at the presence of the damping factor ϕ_1 .

B PROOF OF REGRET BOUND

B.1 BACKGROUNDS

We investigate the convergence of AdamT with regret minimization by following Zinkevich (2003). Formally, the regret is defined as the difference between total cost and that of best point in hindsight

$$R(T) = \sum_{t=1}^T [f_t(x_t) - f_t(x^*)],$$

where $f(x_t)$ is the convex function with arguments x_t at iteration t and x^* denotes the optimal set of the parameters. For the sake of clarity, we denote $f(x_t)$ as f_t .

For a vanishing average regret, Jensen's inequality implies that the average decision $f(\bar{x}_T)$ converges to $f(x^*)$. In other words,

$$f(\bar{x}_T) - f(x^*) \leq \frac{1}{T} \sum_{t=1}^T [f(x_t) - f(x^*)] = \frac{\text{regret}}{T}.$$

Below we list essential theorems and lemmas.

Definition B.1 (Convexity). *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a bounded convex and compact set in Euclidean space. A function $f : \mathcal{K} \mapsto \mathbb{R}$ is convex if for any $x, y \in \mathcal{K}$*

$$\forall \alpha \in [0, 1], \quad f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y).$$

We denote by D an upper bound on the diameter of \mathcal{K} , then

$$x, y \in \mathcal{K}, \quad \|x - y\| \leq D$$

Definition B.2 (Subgradient). *The set of all subgradients of a function f at x , denoted $\partial f(x)$, is the set of all vectors u such that*

$$f(y) \geq f(x) + u^\top (y - x)$$

Suppose f is differentiable, for any existing gradient $\nabla f(x)$, we have $\nabla f(x) \in \partial f(x)$ and $\forall y \in \mathcal{K}$

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x).$$

We denote $G \geq 0$ and upper bound on the norm of the subgradients of f over \mathcal{K} , i.e., $\|\nabla f(x)\| \leq G$ for all $x \in \mathcal{K}$. In other words, the function f is Lipschitz continuous with parameter G , that is, for all $x, y \in \mathcal{K}$

$$|f(x) - f(y)| \leq G\|x - y\|$$

Below are the three core lemmas that applied in proving the regret bound.

Lemma B.3 (Kingma & Ba (2015)). *We define $\nabla f_{1:t,i} \in \mathbb{R}^t$ as the vector at the i^{th} dimension of the gradients over all iterations till t , $\nabla f_{1:t,i} = [\nabla f_{1,i}, \nabla f_{2,i}, \dots, \nabla f_{t,i}]$. With $\|\nabla f_t(x_t)\|_2 \leq G$, $\|\nabla f_t(x_t)\|_\infty \leq G_\infty$ we have*

$$\sum_{t=1}^T \sqrt{\frac{\nabla f_{t,i}^2}{t}} \leq 2G_\infty \|\nabla f_{1:T,i}\|_2$$

Lemma B.4. *For $\beta_1, \gamma_1, \beta_2, \gamma_2 \in [0, 1)$ and $\phi_1, \phi_2 \in [0, 1]$ the series $\{\hat{m}_{t,i}\}$ and $\{\hat{v}_{t,i}\}$ has the following summation form*

$$\begin{aligned} \hat{m}_{t,i} &\leq \left[\frac{(1 - \beta_1)}{1 - \beta_1^t} + \frac{(1 - \gamma_1 \phi_1)(1 - \beta_1)}{(1 - (\gamma_1 \phi_1)^t)} \right] \sum_{i=1}^t \nabla f_i. \\ \hat{v}_{t,i} &\leq \left[\frac{(1 - \beta_2)}{1 - \beta_2^t} + \frac{(1 - \gamma_2 \phi_2)(1 - \beta_2)}{(1 - (\gamma_2 \phi_2)^t)} \right] \sum_{i=1}^t \nabla f_i^2. \end{aligned}$$

Proof. We start from $\hat{m}_{t,i}$. It can be shown that

$$\begin{aligned} m_{t-1} &= \ell_{t-1}^m + \phi_1 b_{t-1}^m \\ \ell_t^m &= (1 - \beta_1) \nabla f_t + \beta_1 \ell_{t-1}^m + \beta_1 \phi_1 b_{t-1}^m \\ b_t^m &= (1 - \beta_1)(1 - \gamma_1) \nabla f_t - (1 - \beta_1)(1 - \gamma_1) \ell_{t-1}^m - ((1 - \beta_1)(1 - \gamma_1) - \phi_1) b_{t-1}^m \end{aligned}$$

Substitute $\ell_{t,i}^m$ and $b_{t,i}^m$, we have

$$\begin{aligned} \hat{m}_{t,i} &= \frac{1}{1 - \beta_1^t} \ell_{t,i}^m + \frac{(1 - \gamma_1 \phi_1)}{(1 - \gamma_1)(1 - (\gamma_1 \phi_1)^t)} b_{t,i}^m \\ &= \left[\frac{(1 - \beta_1)}{1 - \beta_1^t} + \frac{(1 - \gamma_1 \phi_1)(1 - \beta_1)}{(1 - (\gamma_1 \phi_1)^t)} \right] \nabla f_{t,i} \\ &\quad + \left[\frac{(1 - \beta_1)}{1 - \beta_1^t} (\beta_1(1 + \phi_1)) + \frac{(1 - \gamma_1 \phi_1)(1 - \beta_1)}{(1 - (\gamma_1 \phi_1)^t)} (\phi_1 - (1 - \beta_1)(2 - \gamma_1)) \right] \nabla f_{t-1,i} \\ &\quad + \left[\frac{(1 - \beta_1)}{1 - \beta_1^t} (\beta_1((\beta_1 - \phi_1(1 - \beta_1)(1 - \gamma_1)) + \phi_1)(\beta_1 - \phi_1 + (1 - \beta_1)(1 - \gamma_1))) \right. \\ &\quad \left. + \frac{(1 - \gamma_1 \phi_1)(1 - \beta_1)}{(1 - (\gamma_1 \phi_1)^t)} (\beta_1(1 - \beta_1)(\phi_1(\gamma_1 - 1) - 1) - (\phi_1 - (1 - \beta_1)(2 - \gamma_1))((1 - \beta_1)(1 - \gamma_1) - \phi_1)) \right] \nabla f_{t-2,i} \\ &\quad + \dots \\ &\leq \left[\frac{(1 - \beta_1)}{1 - \beta_1^t} + \frac{(1 - \gamma_1 \phi_1)(1 - \beta_1)}{(1 - (\gamma_1 \phi_1)^t)} \right] \sum_{k=1}^t \nabla f_{k,i} \end{aligned}$$

Similarly for \hat{v}_t , we have

$$\hat{v}_{t,i} \leq \left[\frac{(1-\beta_2)}{1-\beta_2^t} + \frac{(1-\gamma_2\phi_2)(1-\beta_2)}{(1-(\gamma_2\phi_2)^t)} \right] \sum_{k=1}^t \nabla f_{k,i}^2.$$

□

B.2 PROOF OF REGRET BOUND

Theorem B.5. *Assume that the objective function f_t has bounded gradients, that is $\|\nabla f_t(x)\|_2 \leq G$, $\|\nabla f_t(x)\|_\infty \leq G_\infty$ for all $x \in \mathbb{R}^d$. We further assume that the distance between any x_t produced by AdamT is bounded, that is $\|x_n - x_m\|_2 \leq D$, $\|x_m - x_n\|_\infty \leq D_\infty$ for any $m, n \in \{1, 2, \dots, T\}$, and $\beta_1, \gamma_1, \beta_2, \gamma_2 \in [0, 1)$, $\phi_1, \phi_2 \in [0, 1]$. The proposed AdamT achieves the following guarantee for all $T \geq 1$*

$$\begin{aligned} R(T) &\leq \frac{D_\infty^2}{2\eta(1-\beta_1)} \sum_{i=1}^d \sqrt{T\hat{v}_{T,i}} + \frac{1}{2} [\beta_1(1-\beta_1)(1+\phi_1(1-\gamma_1))^2] G_\infty^2 + \frac{2(1-\gamma_1\phi_1)^3}{(1-\beta_1)((1-\gamma_1)(1-(\gamma_1\phi_1)^T))^2} D_\infty^2 \\ &\quad + \frac{\eta}{2(1-\beta_1)} \frac{(2-\beta_1)^2}{\sqrt{2-\beta_2}} \sum_{i=1}^d \|g_{1:T,i}\|_2 + dD_2^2 \frac{(1-\gamma_1)\beta_1 + 1 - \gamma_1\phi_1}{2(1-\gamma_1)(1-\beta_1)} \end{aligned}$$

Proof. According to Theorem B.2, for any convex function f_t and $x^* \in \mathcal{K}$, we have

$$f_t(x_t) - f_t(x^*) \leq \nabla f_t(x_t)^\top (x_t - x^*)$$

which implies the regret upper bound

$$R(T) = \sum_{t=1}^T f_t(x_t) - \sum_{t=1}^T f_t(x^*) \leq \sum_{t=1}^T \nabla f_t(x_t)^\top (x_t - x^*).$$

From the update rules in Algorithm,

$$\begin{aligned} x_{t+1} &= x_t - \frac{\eta_t}{\sqrt{\hat{v}_t}} \hat{m}_t \\ &= x_t - \frac{\eta_t}{\sqrt{\hat{v}_t}} \left(\frac{\ell_t^m}{1-\beta_1^t} + \frac{(1-\gamma_1\phi_1)b_t^m}{(1-\gamma_1)(1-(\gamma_1\phi_1)^t)} \right) \\ &= x_t - \frac{\eta_t}{\sqrt{\hat{v}_t}} \left(\frac{\beta_1(\ell_{t-1}^m + \phi_1 b_{t-1}^m)}{1-\beta_1^t} + \frac{(1-\beta_1)\nabla f_t(x_t)}{1-\beta_1^t} + \frac{(1-\gamma_1\phi_1)b_t^m}{(1-\gamma_1)(1-(\gamma_1\phi_1)^t)} \right). \end{aligned}$$

Consider the i^{th} dimension of the parameter vector at time t

$$\begin{aligned} (x_{t+1,i} - x_i^*)^2 &= (x_{t,i} - \frac{\eta_t}{\sqrt{\hat{v}_{t,i}}} \hat{m}_{t,i} - x_i^*)^2 \\ &= (x_{t,i} - x_i^*)^2 - 2\eta_t \frac{\hat{m}_{t,i}}{\sqrt{\hat{v}_{t,i}}} (x_{t,i} - x_i^*) + \eta_t^2 \left(\frac{\hat{m}_{t,i}}{\sqrt{\hat{v}_{t,i}}} \right)^2 \\ &= (x_{t,i} - x_i^*)^2 - \frac{2\eta_t}{\sqrt{\hat{v}_t}} \left(\frac{\beta_1(\ell_{t-1}^m + \phi_1 b_{t-1}^m)}{1-\beta_1^t} + \frac{(1-\beta_1)\nabla f_t(x_t)}{1-\beta_1^t} + \frac{(1-\gamma_1\phi_1)b_t^m}{(1-\gamma_1)(1-(\gamma_1\phi_1)^t)} \right) (x_{t,i} - x_i^*) \\ &\quad + \eta_t^2 \left(\frac{\hat{m}_{t,i}}{\sqrt{\hat{v}_{t,i}}} \right)^2 \end{aligned}$$

Rearrange the equation above will get

$$\begin{aligned}
\nabla f_{t,i}(x_{t,i} - x_i^*) &= \frac{\sqrt{\hat{v}_t}(1 - \beta_1^t)}{2\eta_t(1 - \beta_1)} ((x_{t,i} - x_i^*)^2 - (x_{t+1,i} - x_i^*)^2) \\
&\quad - \left(\frac{\beta_1(\ell_{t-1}^m + \phi_1 b_{t-1}^m)}{1 - \beta_1} + \frac{(1 - \beta_1^t)(1 - \gamma_1 \phi_1) b_t^m}{(1 - \beta_1)(1 - \gamma_1)(1 - (\gamma_1 \phi_1)^t)} \right) (x_{t,i} - x_i^*) + \frac{\eta_t(1 - \beta_1^t) \hat{m}_{t,i}^2}{2(1 - \beta_1) \sqrt{\hat{v}_{t,i}}} \\
&= \frac{\sqrt{\hat{v}_t}(1 - \beta_1^t)}{2\eta_t(1 - \beta_1)} ((x_{t,i} - x_i^*)^2 - (x_{t+1,i} - x_i^*)^2) + \frac{\beta_1(\ell_{t-1}^m + \phi_1 b_{t-1}^m)}{1 - \beta_1} (x_i^* - x_{t,i}) \\
&\quad + \frac{(1 - \beta_1^t)(1 - \gamma_1 \phi_1) b_t^m}{(1 - \beta_1)(1 - \gamma_1)(1 - (\gamma_1 \phi_1)^t)} (x_i^* - x_{t,i}) + \frac{\eta_t(1 - \beta_1^t) \hat{m}_{t,i}^2}{2(1 - \beta_1) \sqrt{\hat{v}_{t,i}}}.
\end{aligned}$$

For any $\beta_1 \leq 1$ and $(1 - \beta_1^t) \leq 1$, we can get

$$\begin{aligned}
\nabla f_{t,i}(x_{t,i} - x_i^*) &\leq \frac{\sqrt{\hat{v}_t}}{2\eta_t(1 - \beta_1)} ((x_{t,i} - x_i^*)^2 - (x_{t+1,i} - x_i^*)^2) + \frac{\beta_1(\ell_{t-1}^m + \phi_1 b_{t-1}^m)}{1 - \beta_1} (x_i^* - x_{t,i}) \\
&\quad + \frac{(1 - \gamma_1 \phi_1) b_t^m}{(1 - \beta_1)(1 - \gamma_1)(1 - (\gamma_1 \phi_1)^t)} (x_i^* - x_{t,i}) + \frac{\eta_t \hat{m}_{t,i}^2}{2(1 - \beta_1) \sqrt{\hat{v}_{t,i}}}.
\end{aligned}$$

Applying Young's inequality on the second and the third term yields

$$\begin{aligned}
&\frac{\beta_1(\ell_{t-1}^m + \phi_1 b_{t-1}^m)}{1 - \beta_1} (x_i^* - x_{t,i}) \\
&= \left(\sqrt{\frac{\beta_1}{(1 - \beta_1)}} (x_i^* - x_{t,i}) \right) \left(\sqrt{\frac{\beta_1}{(1 - \beta_1)}} (\ell_{t-1}^m + \phi_1 b_{t-1}^m) \right) \\
&\leq \frac{\beta_1}{2(1 - \beta_1)} (x_i^* - x_{t,i})^2 + \frac{\beta_1}{2(1 - \beta_1)} (\ell_{t-1}^m + \phi_1 b_{t-1}^m)^2, \\
&\frac{(1 - \gamma_1 \phi_1) b_t^m}{(1 - \beta_1)(1 - \gamma_1)(1 - (\gamma_1 \phi_1)^t)} (x_i^* - x_{t,i}) \\
&= \left(\sqrt{\frac{1 - \gamma_1 \phi_1}{(1 - \beta_1)(1 - \gamma_1)}} (x_i^* - x_{t,i}) \right) \left(\sqrt{\frac{\phi_1}{(1 - \beta_1)(1 - \gamma_1^t)}} b_t^m \right) \\
&\leq \frac{1 - \gamma_1 \phi_1}{2(1 - \beta_1)(1 - \gamma_1)} (x_i^* - x_{t,i})^2 + \frac{1 - \gamma_1 \phi_1}{2(1 - \beta_1)(1 - \gamma_1)} \left(\frac{b_t^m}{(1 - \gamma_1^t)} \right)^2
\end{aligned}$$

The regret bound can be derived by taking summation across all the dimensions for $i = 1, 2, \dots, d$ and the convex functions for $t = 1, 2, \dots, T$.

$$\begin{aligned}
R(T) &\leq \sum_{i=1}^d \sum_{t=1}^T \nabla f_{t,i}(x_{t,i} - x_i^*) \\
&\leq \underbrace{\sum_{i=1}^d \frac{1}{2\eta(1 - \beta_1)} (x_{1,i} - x_i^*)^2 \sqrt{\hat{v}_{1,i}} + \sum_{i=1}^d \sum_{t=2}^T \frac{1}{2(1 - \beta_1)} (x_{t,i} - x_i^*)^2 \left(\frac{\sqrt{\hat{v}_{t,i}}}{\eta} - \frac{\sqrt{\hat{v}_{t-1,i}}}{\eta_{t-1}} \right)}_{P1} \\
&\quad + \underbrace{\sum_{i=1}^d \frac{1}{2(1 - \beta_1)} \left(\beta_1(\ell_{t-1}^m + \phi_1 b_{t-1}^m)^2 + \frac{(1 - \gamma_1 \phi_1)((1 - \gamma_1 \phi_1)(b_t^m))^2}{((1 - \gamma_1)(1 - (\gamma_1 \phi_1)^t))^2} \right)}_{P2} \\
&\quad + \underbrace{\sum_{i=1}^d \sum_{t=1}^T \frac{\eta}{2(1 - \beta_1)} \frac{\hat{m}_{t,i}^2}{\sqrt{\hat{v}_{t,i}}}}_{P3} + \underbrace{\sum_{t=1}^T \frac{1}{2(1 - \beta_1)} \left(\beta_1 + \frac{1 - \gamma_1 \phi_1}{1 - \gamma_1} \right) (x_i^* - x_{t,i})^2}_{P4}
\end{aligned}$$

According to the assumptions $\|x_t - x^*\|_2 \leq D$, $\|x_m - x_n\|_\infty \leq D_\infty$ for any $m, n \in 1, 2, \dots, T$, the above four parts in the inequality can be further upper bounded as follows respectively:

$$\begin{aligned}
P1 &= \sum_{i=1}^d \frac{1}{2\eta(1-\beta_1)} (x_{1,i} - x_i^*)^2 \sqrt{\hat{v}_{1,i}} + \sum_{i=1}^d \sum_{t=2}^T \frac{1}{2(1-\beta_1)} (x_{t,i} - x_i^*)^2 \left(\frac{\sqrt{\hat{v}_{t,i}}}{\eta_t} - \sqrt{\hat{v}_{t-1,i}} \right) \\
&\leq \frac{D_\infty^2}{2\eta(1-\beta_1)} \sum_{i=1}^d \sqrt{T\hat{v}_{T,i}} \\
P2 &= \sum_{t=1}^T \frac{1}{2(1-\beta_1)} \left(\beta_1(\ell_{t-1}^m + \phi_1 b_{t-1}^m)^2 + \frac{(1-\gamma_1\phi_1)((1-\gamma_1\phi_1)(b_t^m))^2}{((1-\gamma_1)(1-(\gamma_1\phi_1)^T))^2} \right) \\
&= \sum_{t=1}^T \frac{\beta_1}{2(1-\beta_1)} (\ell_{t-1}^m + \phi_1 b_{t-1}^m)^2 + \sum_{t=1}^T \frac{1}{2(1-\beta_1)} \frac{(1-\gamma_1\phi_1)((1-\gamma_1\phi_1)(b_t^m))^2}{((1-\gamma_1)(1-(\gamma_1\phi_1)^T))^2} \\
&\leq \frac{1}{2}(\beta_1(1-\beta_1)(1+\phi_1(1-\gamma_1))^2) \sum_{t=1}^T \nabla f_{t-1}^2 \\
&\quad + \frac{(1-\gamma_1\phi_1)^3}{2(1-\beta_1)((1-\gamma_1)(1-(\gamma_1\phi_1)^T))^2} \sum_{t=1}^T (b_t^m)^2 \\
&\leq \frac{1}{2}(\beta_1(1-\beta_1)(1+\phi_1(1-\gamma_1))^2) G_\infty^2 + \frac{2(1-\gamma_1\phi_1)^3}{(1-\beta_1)((1-\gamma_1)(1-(\gamma_1\phi_1)^T))^2} D_\infty^2 \\
P3 &= \sum_{i=1}^d \sum_{t=1}^T \frac{\eta}{2(1-\beta_1)} \frac{\hat{m}_{t,i}^2}{\sqrt{\hat{v}_{t,i}}} \leq \frac{\eta}{2(1-\beta_1)} \frac{(2-\beta_1)^2}{\sqrt{2-\beta_2}} \sum_{i=1}^d \|g_{1:T,i}\|_2 \\
P4 &= \sum_{i=1}^d \sum_{t=1}^T \frac{1}{2(1-\beta_1)} \left(\beta_1 + \frac{1-\gamma_1\phi_1}{1-\gamma_1} \right) (x_i^* - x_{t,i})^2 \\
&\leq \frac{(1-\gamma_1)\beta_1 + 1 - \gamma_1\phi_1}{2(1-\gamma_1)(1-\beta_1)} \sum_{i=1}^d D_2^2 \\
&= dD_2^2 \frac{(1-\gamma_1)\beta_1 + 1 - \gamma_1\phi_1}{2(1-\gamma_1)(1-\beta_1)}
\end{aligned}$$

Hence, the final regret upper bound can be written as:

$$\begin{aligned}
R(T) &\leq \frac{D_\infty^2}{2\eta(1-\beta_1)} \sum_{i=1}^d \sqrt{T\hat{v}_{T,i}} + \frac{1}{2}[\beta_1(1-\beta_1)(1+\phi_1(1-\gamma_1))^2] G_\infty^2 + \frac{2(1-\gamma_1\phi_1)^3}{(1-\beta_1)((1-\gamma_1)(1-(\gamma_1\phi_1)^T))^2} D_\infty^2 \\
&\quad + \frac{\eta}{2(1-\beta_1)} \frac{(2-\beta_1)^2}{\sqrt{2-\beta_2}} \sum_{i=1}^d \|g_{1:T,i}\|_2 + dD_2^2 \frac{(1-\gamma_1)\beta_1 + 1 - \gamma_1\phi_1}{2(1-\gamma_1)(1-\beta_1)}
\end{aligned}$$

□