Convergence of Optimizers Without Bounded Gradient Assumption

Roger Wang   **Eric Xia**
University of Washington
{rogerwyf, ericxia}@uw.edu

1   Introduction

The primary goal of most deep learning models is to minimize the underlying loss function. Optimizers are crucial for the task of updating model weights such that the model will actually converge to a minimum in a computationally efficient manner instead of overshooting or moving away from the minimum. While existing optimizers are intuitively straightforward in convex learning, in non-convex settings they (notably for Adam-type adaptive gradient methods) often require the assumption on the boundedness of gradients for achieving convergence.

While convenient, the imposition of assumptions on the boundedness of gradients can be difficult to verify in practical settings, hence in recent years there has been a trend in research efforts towards analyzing existing optimization methods and proposing novel methods without such assumptions.

In this report, we will review four recent research papers on the convergence of stochastic optimization algorithms including SGD (Section 3), SGD with Momentum (Section 4), Root Mean Squared Propagation (RMSProp) (Section 5) and Adaptive Moment Estimation (Adam) (Section 6), analyzing their approaches, methodologies, and findings as well as discussing their broader implications for the convergence problem and the general field of optimizers.

2   General Assumptions

In a stochastic setting, the optimization problem for a neural network training process can be written as a finite-sum problem:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{n} \sum_{j=0}^{n-1} f(x, s_j)$$

where $f(x, s_j)$ represents the loss function contributed by the randomly shuffled sample batch s_j.

Let $g_t := \nabla f(x_t)$ and $\tilde{g}_t := \nabla f(x_t, s_t)$ be the full gradient and the stochastic gradient with respect to the sampled batch s_t of the objective function f at time or iteration t, respectively. Below we list the common assumptions in standard stochastic optimization that are either explicitly stated or implied in the papers covered in this report:

1. ***Lower boundedness***: *f is lower-bounded by some function f^*.*
2. ***Smoothness***: *f is L-smooth or ∇f is L-Lipschitz continuous for some constant L.*
3. ***Unbiased gradients***: *$\forall t, \mathbb{E}_{s_t}[\tilde{g}_t] = g_t$.*
4. ***Independence***: *the random samples s_j's are independent.*
5. ***Bounded variance***: *$Var_{s_t}(\tilde{g}_t) = \mathbb{E}_{s_t}[\tilde{g}_t - g_t] \leq \sigma^2$ for some $\sigma > 0$.*

Throughout this report, we will refer to the above as *General Assumptions* and use the above notations. Furthermore, each paper in this report may have additional assumptions which will be

included in their corresponding section. It is worth mentioning that none of these assumptions is related to the boundedness to gradient g_t or $\tilde{g}_t$, which is often the key assumption utilized by previous theoretical work.

3 Stochastic Gradient Descent for Structured Nonconvex Functions

3.1 Previous Research & Motivation

Prior to [1], standard convergence theory for SGD in smooth non-convex settings gave a slow sublinear convergence to a stationary point. There has recently been a great interest in exploiting additional structure of classes of nonconvex function, such as error bound properties [2], quasi (strong) convexity [3–5], and quadratic growth condition [6].

In [1], Gower et al. provide a new general analysis of SGD focusing on the two weakest of these properties, quasar (strongly) convex (QC) functions and functions satisfying the Polyak-Łojasiewicz (PL) conditions [7], using the expected residual (ER) condition [8].

3.2 Additional Assumptions

On top of *General Assumptions*, the paper assumes that the *gradient noise* σ^2 is finite and that the Expected Residual condition holds, where $g \in \text{ER}(\rho)$ if

$$\mathbb{E}\left[||g(x) - g(x^*) - (\nabla f(x) - \nabla f(x^*))||^2\right] \leq 2\rho(f(x) - f(x^*)), \ \ \forall x \in \mathbb{R}^d. \tag{1}$$

3.3 Main Results & Methodology

The main results of the paper are convergence bounds using the ER condition for SGD on QC functions and minibatch SGD on PL functions.

Theorem 3.1 (QC Functions with Constant and Decreasing Step-sizes) *Assume General Assumptions, $f(x)$ is ζ-QC with respect to x^*, and $g \in ER(\rho)$. Let $0 < \gamma_k < \frac{\zeta}{2\rho+L}$ for all $k \in \mathbb{N}$ and let $r_0 := ||x^0 - x^*||^2$. Then iterates of SGD satisfy*

$$\min_{t=0,\ldots,k-1} \mathbb{E}\left[f(x^t) - f(x^*)\right] \leq \frac{1}{\sum_{i=0}^{k-1} \gamma_i(\zeta - \gamma_i(2\rho + L))} \left[\frac{r^0}{2} + \sigma^2 \sum_{t=0}^{k-1} \gamma_t^2\right].$$

Furthermore, for $\gamma < \frac{\zeta}{2\rho+L}$, we have that

1. If $\forall k \in \mathbb{N}$, $\gamma_k = \gamma \equiv \frac{1}{2}\frac{\zeta}{(2\rho+L)}$ then $\forall k \in \mathbb{N}$,

$$\min_{t=0,\ldots,k-1} \mathbb{E}[f(x^t) - f(x^*)] \leq 2r_0 \frac{2\rho + L}{\zeta^2 k} + \frac{\sigma^2}{2\rho + L}.$$

2. Suppose SGD is run for T iterations. If $\gamma_k = \frac{\gamma}{\sqrt{T}}$ for all k from 0 to $T - 1$, then

$$\min_{t=0,\ldots,T-1} \mathbb{E}[f(x^t) - f(x^*)] \leq \frac{r_0 + 2\gamma^2\sigma^2}{\gamma\sqrt{T}}.$$

3. If $\forall k \in \mathbb{N}$, $\gamma_k = \frac{\gamma}{\sqrt{k+1}}$ then $\forall k \in \mathbb{N}$,

$$\min_{t=0,\ldots,k-1} \mathbb{E}[f(x^t) - f(x^*)] \leq \frac{1}{4\gamma} \frac{r_0 + 2\gamma^2\sigma^2(\log(k) + 1)}{\zeta(\sqrt{k} - 1) - \gamma(\rho + L/2)(\log(k) + 1)},$$

which converges at a rate $\mathcal{O}\left(\frac{\log(k)}{\sqrt{k}}\right)$.

Theorem 3.2 (PL Functions with Constant Step-sizes) *Assume General Assumptions, $f \in PL(\mu)$, and $g \in ER(\rho)$. Let $\gamma_k = \gamma \leq \frac{1}{1+2\rho/\mu} \frac{1}{L}$, for all k, then SGD converges as follows:*

$$\mathbb{E}[f(x^k) - f^*] \leq (1 - \gamma\mu)^k [f(x^0) - f^*] + \frac{L\gamma\sigma^2}{\mu}. \tag{2}$$

Thus, given $\epsilon > 0$ and using step size $\gamma = \frac{1}{L} \min\{\frac{\mu\epsilon}{2\sigma^2}, \frac{1}{1+2\rho/\mu}\}$ we have that

$$k \geq \frac{L}{\mu} \max\{\frac{2\sigma^2}{\mu\epsilon}, 1 + \frac{2\rho}{\mu}\} \log\left(\frac{2(f(x^0) - f^*)}{\epsilon}\right) \implies \mathbb{E}[f(x^k) - f^*] \leq \epsilon. \tag{3}$$

When the function interpolates the data, SGD converges to the solution at a linear rate.

We will now elaborate on the methodology used to achieve the results. Taking $z = x - \frac{1}{L}(\nabla f(x) - \nabla f(x^*))$ for any $x, z \in \mathbb{R}^d$ that defines L-smoothness of f, the definition for L-smooth may be rearranged st.

$$||\nabla f(x)||^2 \leq 2L(f(x) - f(x^*)). \tag{4}$$

Using this and the definitions of Strong Growth Condition + L-smooth, Weak Growth Condition, Expected Smoothness, and ER in that successive order, it follows from simple rearrangements and taking expectations that this order also gives the descending strength of the concerned assumptions. Using (1) and (4) together gives

$$\mathbb{E}_D\left[||g(x)||^2\right] \leq 2(2\rho + L)(f(x) - f(x^*)) + 2\sigma^2, \tag{5}$$

where $g \in ER(\rho)$ and $x \in \mathbb{R}^d$. With an unbiased estimate of the gradient $g(x)$ we can then use SGD to solve the unconstrained finite-sum optimization problem by sampling $g(x^k)$ i.i.d. and iterating

$$x^{k+1} = x^k - \gamma^k g(x^k).$$

We have

$$||x^{k+1} - x^*||^2 = ||x^k - x^*||^2 - 2\gamma_k \langle g(x^k), x^k - x^* \rangle + \gamma_k^2 ||g(x^k)||^2.$$

By rearranging and taking expectation, then summing over $k = 0, \ldots, t - 1$ and using telescopic cancellation, we have that iterates of SGD satisfy the equation in Theorem 3.1 and the remaining conclusions in Theorem 3.1 follow by substitutions and integral bounds. For Theorem 3.2, smoothness of f is combined with SGD update rule, where expectation conditioned on x^k is taken and the resultant geometric series summed to get (2). Dividing the RHS of (2) into two parts and bounding each separately by $\frac{\epsilon}{2}$, we restrict the step size as in Theorem 3.2 and insert γ into a rearrangement of these bounds to get (3).

3.4 Discussion

The ER condition relies solely on L-smoothness of f and interpolation of data points and is shown above to be strictly weaker than previous assumptions, including the bounded gradient. Not only does ER hold for a larger class of assumptions than standard convergence theory in nonconvex setting which relies on bounded gradient assumption [9–11] or a growth condition [12–14], resulting convergence rates either match or exceed state-of-the-art for both QC and PL functions. Recent evidence suggests a QC structure for loss functions in NNs [15] and this work recovers a $O(1\sqrt{k})$ convergence rate for QC functions with only the ER assumption.

For PL functions, SGD was previously shown to converge at a rate of $O(1/\sqrt{k})$ assuming bounded gradients [16]. Assuming in addition the interpolation and SGC, a linear rate was achieved in Vaswani et al. 2019, but with a suboptimal dependence on the condition number of the function. This work achieves linear convergence to a neighborhood for PL functions.

Recently, Khaled and Richtarik (2020) [17] present convergence analysis of SGD in nonconvex setting, relying on the ABC assumption (A.1), of which all previously discussed assumptions SGC, WGC, ES, and ER are special cases by taking different values for A, B, C. Whereas Theorem 3.2 above is less general and therefore stronger than Theorem 3 in Khaled and Richtarik's work, unlike [17], Theorem 3.2 does not depend on total number of steps and this allows practitioners to simply observe SGD progress and stop when a desired tolerance is achieved.

4 Stochastic Gradient Descent with Momentum

4.1 Previous Research & Motivation

Prior to [18] by Liu et al., there had been some interests in investigating the convergence of SGDM. [19] provides a global convergence of SGDM but it assumes uniformly boundedness of gradients of the objection funcion. [20] presents a convergence bound of SGDM for general nonconvex functions but does not explain the competitiveness of SGDM compared to SGD. Moreover, the convergence rate of Multistage SGDM had not been established except for the classic SGD case.

In this work, Liu et al. provide a novel convergence analysis for SGDM and Multistage[1] SGDM without bounded gradient assumptions. This work also demonstrates that SGDM has the same convergence bound as SGD for both strongly convex and nonconvex functions without uniformly bounded gradient assumption, and is the first convergence guarantee for SGDM in a multistage setting.

4.2 Main Results & Methodology

The main results of this paper are the convergence bounds of SGDM and Multistage SGDM. In SGDM, let α and β be learning rate and momentum weight, we have the following result:

Theorem 4.1 (Non-convex SGDM) *Assume* $f : \mathbb{R}^d \to \mathbb{R}$ *satisfies General Assumptions, let* $\alpha \leq \min\{\frac{1-\beta}{L(4-\beta+\beta^2)}, \frac{1-\beta}{2\sqrt{2}L\sqrt{\beta+\beta^2}}\}$, *then*

$$\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}[\|g_t\|^2] \leq \frac{2(f(x_1)-f^*)}{k\alpha} + (\frac{\beta+3\beta^2}{2(1+\beta)}+1)L\alpha\sigma^2 = \mathcal{O}(\frac{f(x_1)-f^*}{k\alpha}+L\alpha\sigma^2)$$

Theorem 4.2 (Strongly Convex SGDM) *Assume* $f : \mathbb{R}^d \to \mathbb{R}$ *satisfies General Assumptions and is* μ*-strongly convex, let* $\alpha \leq \min\{\frac{1-\beta}{5L}, \frac{1-\beta}{L(3-\beta+2\beta^2+\frac{48\sqrt{\beta}}{25}\frac{2L+18\mu}{L})}\}$, *then for all* $t \geq t_0 := \lfloor\frac{\log 0.5}{\log \beta}\rfloor$,

$$\mathbb{E}[f(x_t)-f^*] = \mathcal{O}\big(\max\{1-\alpha\mu,\beta\} + \frac{L}{\mu}\alpha\sigma^2\big)$$

In a multistage setting, let α_i, β_i and T_i are learning rate (step size), momentum weight and stage length of ith stage, respectively, we have the following result:

Theorem 4.3 (Non-convex Multistage SGDM) *Assume* $f : \mathbb{R}^d \to \mathbb{R}$ *satisfies General Assumptions, restrict the parameters in each stage of Multistage SGDM so that*

$$\begin{aligned}\frac{\alpha_i\beta_i}{1-\beta_i} &\equiv A_1 \quad i=1,...,n \\ \alpha_i T_i &\equiv A_2 \quad i=1,...,n \\ 0 \leq \beta_1 &\leq \beta_2 \leq ... \leq \beta_n \leq 1\end{aligned} \tag{1}$$

and A_1, A_2 *are properly chosen constants. Let* $A_1 = \frac{1}{24\sqrt{2}L}$ *and* A_2 *be large enough so that* $\beta_i^{2T_i} \leq \frac{1}{2}$ *for* $i=1,...,n$. *In addition, let*

$$\frac{1-\beta_1}{\beta_1} \leq 12\frac{1-\beta_n}{\sqrt{\beta_n+\beta_n^2}}$$

then we have

$$\frac{1}{n}\sum_{l=1}^{n}\frac{1}{T_l}\sum_{i=T_1+...+T_{l-1}+1}^{T_1+..+T_l}\mathbb{E}[\|g_t\|^2] \leq \frac{2(f(x_1)-f^*)}{nA_2} + \frac{1}{n}\sum_{i=1}^{n}\Big(24\beta_l^2\frac{\beta_1}{\sqrt{\beta_n+\beta_n^2}}L+3L\Big)\alpha_l\sigma^2$$

$$= \mathcal{O}\Big(\frac{2(f(x_1)-f^*)}{nA_2} + \frac{1}{n}\sum_{i=1}^{n}L\alpha_l\sigma^2\Big)$$

---

[1]Multistage refers to applying a constant stepsize which is then dropped by a constant factor to encourage fine-tuning of training, and the momentum weight is either kept unchanged or gradually increased.

Next, we will summarize the key approaches used to derive the above results. Recall that the core of SGDM algorithm is the following updating rule:

$$v_t = \beta v_{t-1} + (1 - \beta)\tilde{g}_t \quad x_{t+1} = x_t - \alpha v_t$$

Therefore, assume $v_0 = 0$, then v_t can be expressed as

$$v_t = (1 - \beta)\sum_{i=1}^{t} \beta^{k-i}\tilde{g}_i \tag{2}$$

One key observation on the role of β in equation (2) from the paper is that v_t enjoys a reduced variance of $(1 - \beta)\sigma^2$ while having a controllable deviation from the full gradient g_t in expectation since v_t is a moving average of the past stochastic gradients with lower weights on the older ones, thus it makes sense to look at the deterministic version of v_t (replacing $\tilde{g}_i$ with g_i) and its deviation from the ideal descent direction g_t, which could be unbounded without further assumptions.

While previous work assumed the boundedness of g_t to circumvent above difficulty, this work constructed a novel Lyapunov function to handle this deviation:

$$L_t = (f(z_t) - f^*) + \sum_{i=1}^{k-1} c_i \|x^{k+1-i} - x^{k-i}\|^2$$

where $z_t = \begin{cases} x_t & t = 1 \\ \frac{1}{1-\beta}x_t - \frac{\beta}{1-\beta}x_{t-1} & t \geq 2 \end{cases}$

The authors then argued that by carefully defining $\{c_i\}_i^\infty$ such that it is a positive sequence in a diminishing fashion, L_t is indeed a Lyapunov function, thus one can show that $\mathbb{E}[L_{t+1} - L_t] \leq -R_1 E[\|g_t\|^2] + R_2$ for some positive constants $R_1 \geq \frac{\alpha}{2}$ and $R_2 = \mathcal{O}(L\alpha\sigma^2)$. By telescoping this inequality, the convergence of SGDM in Theorem 4.1 is then obtained, and similar techniques were utilized to derive the results in Theorem 4.2 under a strongly convex setting and in Theorem 4.3 for Multistage SGDM.

4.3 Discussion

Theorem 4.1 and Theorem 4.2 show that under both nonconvex and strongly convex settings, with a proper learning rate α, SGDM can achieve the same convergence bound as the classical convergence bound of SGD (as shown in previous work, e.g., Theorem 4.5 and 4.8 in [13]). This result only depends on General Assumptions, and the radius of the stationary distribution is smaller than the previous $\mathcal{O}(\frac{\alpha\sigma^2}{1-\beta})$ result from [19] that relies on the additional assumption of uniformly bounded gradients. It is also worth mentioning that the use of Lyapunov function is a novel approach for convergence analysis of optimization algorithms and provides some new insights throughout this paper.

For Multistage SGDM, Theorem 4.3 is the first theoretical result that guarantees its convergence. Moreover, it was demonstrated from the convergence bound that large learning rates are allowed in the first a few stages to accelerate the initial convergence, and smaller learning rates can refine the radius of the stationary distribution in the later stages, which is an advantage of stagewise training compared to plain SGDM.

However, the convergence analysis in this paper does have some weaknesses and limitations: First of all, although it is theoretically shown that SGDM is "at least as fast as" SGD, this paper did not explore the advantages of SGDM compared to SGD in detail. In addition, Theorem 4.2 assumes a lower-bound of timestamp/iteration for the result to be valid, and this lower-bound could be a problem for some choices of β (e.g, when $\beta = 0.995$, $t_0 = 138$). Finally, equation (1) from Theorem 4.3 puts a strong restriction on the choice of learning rates and momentum weights at all stages, which makes this stage-wise training setup impractical.

5 RMSProp

5.1 Previous Research & Motivation

Prior to [21] by Shi et al., there had been one line of research on the convergence of variants of Adam (which includes RMSProp) with additional assumptions. [22] provides a clean convergence result but assumed a large ϵ compared to weighted moving average of the squared gradient, which is in contrary to the spirit of RMSProp. [23] analyzes deterministic and stochastic RMSProp, but their results were based on an rather unrealistic assumption that all stochastic gradients have the same sign. Furthermore, all the above mentioned works assume the gradients to be bounded.

During the review on one famous counter-example to the convergence of Adam from [24], Shi et al. ran simulations and found that there is always a threshold of the moving average parameter above which RMSProp converges (See Figure 1 in Appendix). This observation motivated them to further investigate the relationship between this parameter and the performance of the algorithm. They discovered that the convergence of RMSProp algorithm is contingent to the choice of the moving average parameter. Moreover, they proved that RMSProp converges to stationary points for certain types of problems and to bounded region for the others, which was the first result of convergence of this algorithm with no assumption about the boundedness of the gradient norm.

5.2 Additional Assumptions

In addition to *General Assumptions*, this paper assumes that for stochastic RMSProp, the objective satisfies

$$\sum_{j=0}^{n-1} \|\nabla f_j(x)\|_2^2 \leq D_1 \|\nabla f(x)\|_2^2 + D_0 \tag{1}$$

for some non-negative constant D_0 and D_1. This can be viewed as an augment to the bounded variance assumption.

5.3 Main Results & Methodology

The main results of this paper are the convergence of RMSProp under both deterministic and stochastic setting. Let α_t be the learning rate at time/iteration t and β be the moving average parameter of the squared gradients norm, we have the following results:

Theorem 5.1 (Deterministic RMSProp) *Assume $f : \mathbb{R}^d \to \mathbb{R}$ satisfies General Assumptions, then for deterministic RMSProp (i.e, full-batch with $n = 1$, $\epsilon = 0$) with a diminishing learning rate $\alpha_t = \frac{\alpha_1}{\sqrt{t}}$ and any $\beta \in (0, 1)$, we have*

$$\min_{t \in (1,T]} \|g_t\|_1 \leq \mathcal{O}\Big(\frac{\log T}{\sqrt{T}}\Big)$$

where $T > 0$ is the total number of iterations.

Theorem 5.2 (Stochastic RMSProp - Bounded Region) *Assume $f : \mathbb{R}^d \to \mathbb{R}$ satisfies General Assumptions and (1). In addition, assume β satisfies*

$$\sqrt{\frac{10dn}{\beta^n}} dn D_1 \Big((1 - \beta) \frac{\frac{4n^2}{\beta^n} - 1}{2} + (\frac{1}{\sqrt{\beta^n}} - 1) \Big) \leq \frac{\sqrt{2} - 1}{2\sqrt{2}}$$

Then, for stochastic RMSProp with a diminishing learning rate $\alpha_t = \frac{\alpha_1}{\sqrt{t}}$, we have

$$\min_{t \in (1,T]} \min\{\|\tilde{g}_t\|_1, \|\tilde{g}_t\|_2^2 \sqrt{\frac{D_1 d}{D_0}}\} \leq \mathcal{O}\Big(\frac{\log T}{\sqrt{T}}\Big) + \mathcal{O}\Big(C\sqrt{D_0}\Big), \quad \forall\, T \geq 4$$

where C is a β-dependent constant that satisfies $\lim_{\beta \to 1} C = 0$

Corollary 5.1 (Stochastic RMSProp - Stationary Point) *Let all assumptions in Theorem 5.2 hold. In addition, assume $D_0 = 0$ in (1), then for stochastic RMSProp we have,*

$$\min_{t \in (1,T]} \|\tilde{g}_t\|_1 \leq \mathcal{O}\Big(\frac{\log T}{\sqrt{T}}\Big), \quad \forall\, T \geq 4$$

Next, we explain the methodology used to derive the above results. Recall that the core of the RMSProp algorithm is the following updating rule:

$$v_t = \beta v_{t-1} + (1-\beta)(g_t \circ g_t) \quad x_t = x_{t-1} - \alpha(v_t + \epsilon I)^{-1/2} g_t$$

For the full-batch version of RMSProp, Theorem 5.1 is an extension of results from the original Adam paper [8] for which the authors utilized the Lipschitz continuity of the gradients to remove the bounded assumption.

The derivation to the results for stochastic RMSProp is more complicated: Based on the observation from the simulation, the authors divided stochastic optimization problems into two classes: *realizable* problems when $D_0 = 0$ from equation (1) and *non-realizable* problems when $D_0 \neq 0$. They further conjectured that these two classes of problems correspond to different convergence behavior of RMSProp contingent to the choice of beta, which is summarized in Table 1:

Class	β close to 1	β close to 0
Non-realizable	Convergence to bounded region (Theorem 5.2)	Divergence
Realizable	Convergence to stationary points (Corollary 5.1)	Divergence

Table 1: Conjecture on convergence outcomes of RMSProp on classes of optimization problems.

The rationale to prove this conjecture is to track the magnitude of squared gradient terms and calculate the diminishing speed of $\|g\|_1$. Since f is L-Lipschitz, by descent lemma, one can show the following inequality

$$f(x_{t+1}) - f(x_t) \leq \langle g_t, x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \tag{2}$$

By examining the distribution of gradient norms among different sampled batches, one can find upper and lower bounds of the inner product between the gradient and difference in iterations of x_t and those of v_t, which are conditioned on the choice of β. Moreover, these bounds directly contribute to bounding the first and second term in inequality (2). Therefore, summing up the inequalities from $t_{init} = 4$ to T gives the results in Theorem 5.2, and Corollary 5.1 is a special case for when $D_0 = 0$.

5.4 Discussion

Theorem 5.2, and Corollary 5.1 provide the first convergence guarantee of RMSProp on non-convex stochastic optimization problems without bounded gradient assumptions. Unlike the previous related works, this paper not only suggests the classification of optimization problems into two kinds (realizable, non-realizable) to tackle them separately , but also proves the existence of a critical threshold of β such that the algorithm will generate reasonably good results (convergence to stationary points/bounded region).

More importantly, although in this report we have only focused on RMSProp, the extension of both theoretical and empirical results from this paper shows that the two parameters of Adam algorithm play different roles in contributing to the convergence of the algorithm and behave differently depending on the optimization problem. Specifically, the "problem-dependent" nature of β clarifies the misconception that "Adam does not converge" from [24], and the authors are able to convey this idea throughout this paper in a convincing manner.

6 Adam

6.1 Previous Research & Motivation

6.2 Additional Assumptions

6.3 Main Results & Methodology

6.4 Discussion

7 Addition Discussion & Comments

8 Future Research

References

[1] Robert M. Gower, Othmane Sebbouh, and Nicolas Loizou. Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation, 2020. URL https://arxiv.org/abs/2006.10311.

[2] Marian J. Fabian, René Henrion, Alexander Y. Kruger, and Jirí V. Outrata. Error bounds: Necessary and sufficient conditions. *Set-Valued and Variational Analysis*, 18:121–149, 2010.

[3] Oliver Hinder, Aaron Sidford, and Nimit S. Sohoni. Near-optimal methods for minimizing star-convex functions and beyond, 2019. URL https://arxiv.org/abs/1906.11985.

[4] Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018. URL http://jmlr.org/papers/v19/16-465.html.

[5] Sergey Guminov, Alexander Gasnikov, and Ilya Kuruzov. Accelerated methods for -weakly-quasi-convex problems, 2017. URL https://arxiv.org/abs/1710.00797.

[6] Mihai Anitescu. Degenerate nonlinear programming with a quadratic growth condition. *SIAM Journal on Optimization*, 10(4):1116–1135, 2000. doi: 10.1137/S1052623499359178. URL https://doi.org/10.1137/S1052623499359178.

[7] B.T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(63)90382-3. URL https://www.sciencedirect.com/science/article/pii/0041555363903823.

[8] R. Michael Gower, Peter Richtárik, and Francis R. Bach. Stochastic quasi-gradient methods: variance reduction via jacobian sketching. *Mathematical Programming*, 188:135 – 192, 2021.

[9] Feng Niu, Benjamin Recht, Christopher Re, and Stephen J. Wright. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent, 2011. URL https://arxiv.org/abs/1106.5730.

[10] Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15(71): 2489–2512, 2014. URL http://jmlr.org/papers/v15/hazan14a.html.

[11] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *ArXiv*, abs/1109.5647, 2012.

[12] Dimitri P. Bertsekas and John N. Tsitsiklis. Neuro-dynamic programming: an overview. *Proceedings of 1995 34th IEEE Conference on Decision and Control*, 1:560–564 vol.1, 1995.

[13] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning, 2016. URL https://arxiv.org/abs/1606.04838.

[14] Mark W. Schmidt, Nicolas Le Roux, and Francis R. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162:83–112, 2017.

[15] Yi Zhou, Junjie Yang, Huishuai Zhang, Yingbin Liang, and Vahid Tarokh. Sgd converges to global minimum in deep learning via star-convex path. *ArXiv*, abs/1901.00451, 2019.

[16] H. Karimi, Julie Nutini, and Mark W. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-Łojasiewicz condition. In *ECML/PKDD*, 2016.

[17] Ahmed Khaled and Peter Richtárik. URL https://arxiv.org/abs/2002.03329.

[18] Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18261–18271. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/d3f5d4de09ea19461dab00590df91e4f-Paper.pdf.

[19] Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization, 2019. URL https://arxiv.org/abs/1905.03817.

[20] Yan Yan, Tianbao Yang, Zhe Li, Qihang Lin, and Yi Yang. A unified analysis of stochastic momentum methods for deep learning. 2018. doi: 10.48550/ARXIV.1808.10396. URL https://arxiv.org/abs/1808.10396.

[21] Naichen Shi, Dawei Li, Mingyi Hong, and Ruoyu Sun. RMSprop converges with proper hyper-parameter. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=3UDSdyIcBDA.

[22] Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad, 2020. URL https://arxiv.org/abs/2003.02395.

[23] Soham De, Anirbit Mukherjee, and Enayat Ullah. Convergence guarantees for rmsprop and adam in non-convex optimization and an empirical comparison to nesterov acceleration, 2018. URL https://arxiv.org/abs/1807.06766.

[24] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond, 2019. URL https://arxiv.org/abs/1904.09237.

beginappendices

A   Appendix

We say that ABC holds with $A, B, C > 0$ if

$$\mathbb{E}\left[||g(x)||^2\right] \le 2A(f(x) - f(x^*)) + B||\nabla f(x)||^2 + C \qquad \text{(A.1)}$$
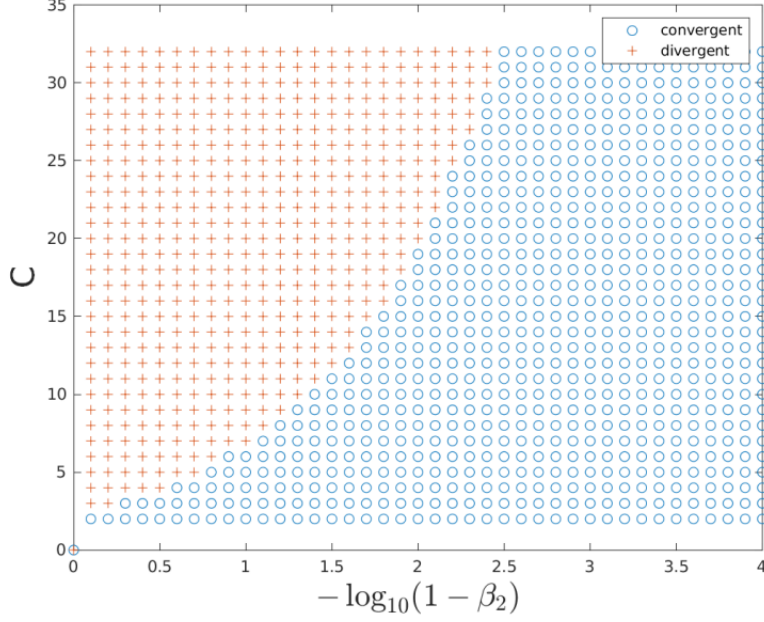
Figure 1: Phase diagram of the outcome of RMSprop on the counter example with the definition $f_t(x) = \begin{cases} Cx & \text{for } t \bmod C = 1 \\ -x & \text{Otherwise} \end{cases}$. Different marks represent different outcome: we label a data point as convergence if the distance between x and -1 is smaller than 0.01 on average after 750000 iterations and as divergence otherwise. For each choice of β, there exists a counter example, but for each counter example in which Adam diverges, there exists a larger β that can make Adam converge. Step size is set as $\alpha_t = \frac{1}{\sqrt{t}}$.