# Convergence of Optimizers Without Bounded Gradient Assumption

**Roger Wang, Eric Xia**
rogerwyf@uw.edu, 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.

For example, for an objective function $f(x)$ satisfying Polyak-Łojasiewicz conditions [1], convergence rates of $O(1/T)$ were established for Stochastic Gradient Descent (SGD) under the assumption that $\mathbb{E}[||\nabla f_i(x_k)||^2] \leq C^2$ for all $x_k$ and some $C$, where $f_i$ typically represents the fit on an individual training sample [2].

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 survey and analyze results from 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), comparing and contrasting 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 $\nabla f$ 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.*

Preprint. Under review.

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 [3], 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 [4], quasi (strong) convexity [5–7], and quadratic growth condition [8].

In [3], Gower et al. provide a new general analsysis of SGD focusing on the two weakest of these properties, quasar (strongly) convex (QC) functions and functions satisfying the Polyak-Łojasiewicz (PL) conditions [1], using the expected residual (ER) condition [9].

## 3.2 Additional Assumptions

On top of *General Assumptions*, the paper makes the mild assumption that the *gradient noise* $\sigma^2$ is finite:

$$\sigma^2 \coloneqq \sup_{x^* \in \mathcal{X}^*} \mathbb{E}\left[||g(x^*)||^2\right] < \infty. \tag{1}$$

## 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.

**Definition 3.1 (Expected Residual)** *We say $g \in ER(\rho)$ if*

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

**Theorem 3.1 (QC Functions with Constant and Decreasing Step-sizes)** *Assume General Assumptions, $f(x)$ is $\zeta$-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 $r_0 \coloneqq ||x^0 - x^*||^2$. Then iterates of SGD satisfy*

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

*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 $\forall k \in \mathbb{N}$,*

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

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

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

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

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

*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:*

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

*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*

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

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

### 3.4 Discussion

## 4 Stochastic Gradient Descent with Momentum

### 4.1 Previous Research & Motivation

Prior to this work, there had been some interests in investigating the convergence of SGDM. [10] provides a global convergence of SGDM but it assumes uniformly boundedness of gradients of the objection funcion. [11] 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 [12], 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

The main results of this paper are the convergence bounds of SGDM and Multistage SGDM. In SGDM, let $\alpha$ and $\beta$ 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 $\mu$-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 $t \geq t_0 := \lfloor\frac{\log 0.5}{\log\beta}\rfloor$,*

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

[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.

In a multistage setting, let $\alpha_i$, $\beta_i$ and $T_i$ are learning rate (step size), momentum weight and stage length of $i$th 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*

$$\begin{gathered}\frac{\alpha_i\beta_i}{1-\beta_i} \equiv A_1 \quad i = 1, ..., n\\ \alpha_i T_i \equiv A_2 \quad i = 1, ..., n\\ 0 \le \beta_1 \le \beta_2 \le ... \le \beta_n \le 1\end{gathered} \tag{2}$$

*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 $\beta_i^{2T_i} \le \frac{1}{2}$ for $i = 1, ..., n$. In addition, let*

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

*then we have*

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

### 4.3 Key Approaches & Insights

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

$$v_t = (1-\beta)\sum_{i=1}^{t}\beta^{k-i}\tilde{g}_i \tag{3}$$

One key observation on the role of $\beta$ 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:

$$L_t = (f(z_t) - f^*) + \sum_{i=1}^{k-1} c_i\|x^{k+1-i} - x^{k-i}\|^2 \tag{4}$$

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

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 $\mathbb{E}[L_{t+1} - L_t] \le -R_1E[\|g_t\|^2] + R_2$ for some positive constants $R_1 \ge \frac{\alpha}{2}$ 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.

### 4.4 Discussion

Theorem 4.1 and Theorem 4.2 show that under both nonconvex and strongly convex settings, with a proper learning rate $\alpha$, 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 [10] 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 $\beta$ (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 this work, there has been one line of research on the convergence of variants of Adam (which includes RMSProp) with additional assumptions. [14] provided a clean convergence result but assumed a large $\epsilon$ compared to weighted moving average of the squared gradient, which is in contrary to the spirit of RMSProp. [15] analyzed 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.

Shi et al. ran simulations for one counter-example to the convergence of Adam from [16] and found that there is always a threshold of the moving average parameter above which RMSProp converges, thus motivating them to investigate the relationship between this parameter and the performance of the algorithm. They discovered [17] that the convergence of RMSProp algorithm is contingent to the choice of the moving average parameter. 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 for stochastic RMSProp that

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

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

The main results of this paper are the convergence of RMSProp under both deterministic and stochastic setting. Let $\alpha_t$ be the learning rate at time/iteration $t$ and $\beta$ 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*

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

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 $\beta$ satisfies*

$$\sqrt{\frac{10dn}{\beta^n}} dnD_1\Big((1-\beta)\frac{\frac{4n^2}{\beta^n}-1}{2} + (\frac{1}{\sqrt{\beta^n}} - 1)\Big) \leq \frac{\sqrt{2}-1}{2\sqrt{2}}$$

*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 $\beta$-dependent constant that satisfies $\lim\limits_{\beta \to 1} C = 0$*

**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,*

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

### 5.4 Key Approaches & Insights

### 5.5 Discussion

# 6 Adam

## 6.1 Previous Research & Motivation

## 6.2 Additional Assumptions

## 6.3 Main Results

## 6.4 Key Approaches & Insights

## 6.5 Discussion

# 7 Addition Discussion & Comments

# 8 Future Research

## References

[1] 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.

[2] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. *CoRR*, abs/1608.04636, 2016. URL http://arxiv.org/abs/1608.04636.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[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] 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.

[15] 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.

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

[17] Tijmen Tieleman, Geoffrey Hinton, et al. 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.