Understanding and Detecting Convergence for Stochastic Gradient Descent with Momentum

Jerry Chee, Ping Li

Cognitive Computing Lab
Baidu Research
10900 NE 8th St. Bellevue, WA 98004, USA
jerry9567@gmail.com, ping98@gmail.com

***Abstract*—Convergence detection of iterative stochastic optimization methods is of great practical interest. This paper considers stochastic gradient descent (SGD) with a constant learning rate and momentum. We show that there exists a transient phase in which iterates move towards a region of interest, and a stationary phase in which iterates remain bounded in that region around a minimum point. We construct a statistical diagnostic test for convergence to the stationary phase using the inner product between successive gradients and demonstrate that the proposed diagnostic works well. We theoretically and empirically characterize how momentum can affect the test statistic of the diagnostic, and how the test statistic captures a relatively sparse signal within the gradients in convergence. Finally, we demonstrate an application to automatically tune the learning rate by reducing it each time stationarity is detected, and show the procedure is robust to mis-specified initial rates.**

I. Introduction

Consider the problem in stochastic optimization

$$\theta_\star = \arg\min_{\theta\in\Theta} \mathbb{E}[\ell(\theta, \xi)]. \tag{1}$$

The loss ℓ is parameterized by $\Theta \subseteq \mathbb{R}^p$, and ξ is a source of randomness like a randomly sampled point. For example, the quadratic loss is $\ell(\theta, \xi) = (1/2)(y - x^\top\theta)^2$ with $\xi = (x, y)$. When the data size N and parameter size p are large, classical optimization methods can fail to estimate $\theta_\star$. In such large-scale settings, the method of stochastic gradient descent (SGD)

$$\theta_{n+1} = \theta_n - \gamma\nabla\ell(\theta_n, \xi_{n+1}) \tag{2}$$

is a powerful alternative [4], [5], [29], [35]. θ_{n+1} is the estimate of $\theta_\star$ at the $(n+1)$-th iteration, and $\gamma > 0$ the learning rate. ξ_{n+1} represents randomly sampled data used to compute the stochastic gradient. A mini-batch can reduce the variance of the stochastic gradients and aid in performance [25].

Momentum or Heavy Ball SGD (SGDM) can typically offer significant speedups [24]:

$$\theta_{n+1} = \theta_n - \gamma\nabla\ell(\theta_n, \xi_{n+1}) + \beta(\theta_n - \theta_{n-1}) \tag{3}$$

where $\beta \in [0, 1)$ is the momentum. The momentum term $\beta(\theta_n - \theta_{n-1})$ accumulates movements in a common direction. The performance of stochastic gradient methods is greatly influenced by the learning rate γ, which can be decreasing (e.g., $\propto 1/n$) or constant. Decreasing learning rates are commonly used in the literature to attain theoretical convergence guarantees. In practice, however, constant learning rates are common due to their ease of tuning and speed of convergence.

Stochastic iterative procedures start from an initial point and then move from a transient phase to a stationary phase [20]. With a decreasing learning rate, the transient phase can be long, and impractically so if the learning rate is just slightly misspecified [22], [30]. But, the stationary phase is convergence to $\theta_\star$. With a constant learning rate the transient phase is much shorter and more robust to the learning rate. The stationary phase is not true convergence but oscillation within a bounded region containing $\theta_\star$. In this study, we develop a statistical convergence diagnostic for SGDM with constant learning rate. Constant learning rate is commonly used in practice, makes the transition from transient to stationary phase clear, and it is pointless to keep running the procedure once the stationary phase has been reached.

A. Related work

The idea that stochastic gradient methods can be separated into a transient and stationary phase (or search and convergence phase) is not new [20]. However, until recently there has been little work in developing principled statistical methods for convergence detection which can guide empirical practice. Heuristics from optimization theory are commonly used, such as stopping when $\|\theta_n - \theta_{n-1}\|$ is small according to some threshold, or when updates of the loss function have reached machine precision [6], [11]. These methods are more suited for deterministic rather than stochastic procedures as they do not account for the sampling variation in stochastic gradient estimates. A more statistically motivated approach is to concurrently monitor test error on a hold-out validation set and stop when validation error begins increasing [3], [5]. But, the validation error is also a stochastic process, and estimating whether it is increasing presents similar, if not greater, challenges to detecting convergence to the stationary phase.

In stochastic approximation, classical theory of stopping times addresses the detection of stationarity [23], [34]. One noteworthy method by [23] forms the basis for our work. It keeps a running average of the inner product of successive gradients $\nabla\ell(\theta_n, \xi_{n+1})^\top\nabla\ell(\theta_{n-1}, \xi_n)$. At a high level, in the transient phase the stochastic gradients generally point in the same direction, resulting in a positive inner product.

In the stationary phase the stochastic gradients roughly point in different directions due to their oscillation in a bounded region containing $\theta_\star$, resulting in a negative inner product. Accelerated methods in stochastic approximation share the underlying intuition that a negative inner product of successive gradients indicates convergence [10], [15], [26].

There has been a recent interest in principled convergence detection for stochastic gradient methods and automated step decay learning rates. Work by [7] developed a principled convergence diagnostic for SGD in Eq. (2) based on Pflug's procedure. We generalize Pflug's procedure to the momentum setting, which introduces challenges to the theoretical justification and practicality of the convergence diagnostic. [14] developed a procedure to automatically switch from Adam [16] to SGD. [27] proposed a modified splitting procedure [28] to detect the stationary phase for SGD and implement a robust learning rate schedule. [18] used a Markov chain t-test and a stationarity condition by [32] to automatically reduce the learning rate for SGDM. [12] analyzed the step decay learning rate schedule in least squares regression and show its optimality over polynomially decaying rates.

B. Our contributions

Section II presents a statistical convergence diagnostic for SGDM (stochastic gradient descent with momentum) and explains the significance of the challenges introduced by momentum. In Section III we provide theoretical and empirical support for the design choices of the convergence diagnostic, and demonstrate the effect of momentum on the test statistic of the diagnostic. We investigate in Section IV what drives the test statistic by analyzing the distribution of the inner product of successive gradients in the stationary phase. In Section V we provide empirical type I and type II error rates on simulated data experiments. Section VI presents an application of the convergence diagnostic to an automatically tuned learning rate schedule, with experiments on benchmark datasets.

II. Convergence diagnostic

Our convergence diagnostic aims to detect the transition from the transient to the stationary phase. We first present the theory which supports the existence of these two phases for SGDM. The expected difference in loss to the minimum has bias terms due to initial conditions, and a variance term due to noise in the stochastic gradients.

Theorem 1 ([33]). *If the expected loss $f(\theta) = \mathbb{E}[\ell(\theta, \xi)]$ is convex, under additional assumptions of the loss, there are positive constants $Q_\beta, R_\beta, S_\beta$ such that for every n, we have*

$$\mathbb{E}[f(\hat{\theta}_n) - f(\theta_\star)] \leq \frac{Q_\beta}{n+1}(f(\theta_0) - f(\theta_\star)) + \frac{R_\beta}{\gamma(n+1)}\|\theta_0 - \theta_\star\|^2 + \gamma S_\beta.$$

□

Remarks. $\hat{\theta}_n = \sum_{t=0}^{n} \theta_t/(n+1)$, $Q_\beta = \beta/(1-\beta)$, $R_\beta = (1-\beta)/2$, and $S_\beta = (G^2 + \delta^2)/2(1-\beta)$ where G is a bound on the gradients and δ^2 a bound on the variance of the stochastic gradients. For large enough n the bias contributions from the transient phase are negligible, and thus a bounded $\mathbb{E}[f(\hat{\theta}_n) - f(\theta_\star)]$ indicates a bounded $\mathbb{E}[f(\theta_n) - f(\theta_\star)]$ in the stationary phase.

Theorem 1 suggests that the constant rate SGDM moves quickly through the transient phase discounting initial conditions $f(\theta_0) - f(\theta_\star)$ and $\|\theta_0 - \theta_\star\|^2$, and then enters the stationary phase where the distance from $\theta_\star$ is bounded $\propto O(\gamma)$. We observe a widely noted trade-off for stochastic gradient methods: a larger learning rate speeds up the transient phase by discounting bias from initial conditions at a higher rate, but increases the radius of the stationary region [1], [21].

We cite [33] in Theorem 1 because their convergence rate consists of a reducible and irreducible term with respect to the number of updates, and best matches our empirical observations. Though [19] showed a linear convergence rate with constant stepsize, their restriction on the momentum (β) makes their convergence rate difficult to realize in practice. For example, using the formula in [19] with min and max eigenvalues $= 0.5$ and stepsize $= 0.1$, then one can check that we have $\beta < 0.2$, which is very restrictive.

While convergence analyses such as Theorem 1 offer valuable theoretical insight, they provide limited practical guidance. One could try to declare convergence when the bias due to initial conditions has been discounted to 1% of the variance, choosing n for $\left[\frac{Q_\beta}{n+1}(f(\theta_0) - f(\theta_\star)) + \frac{R_\beta}{\gamma(n+1)}\|\theta_0 - \theta_\star\|^2\right] = 0.01\gamma S_\beta$. But estimating $f(\theta_0) - f(\theta_\star)$, $\|\theta_0 - \theta_\star\|^2$, G^2, and δ^2 is difficult. We provide an alternative, by developing a practical statistical diagnostic test to estimate the phase transition and detect convergence of SGDM in a much simpler way.

A. Modified Pflug diagnostic

We present a convergence diagnostic for SGDM in Algorithm 1. We draw upon Pflug's procedure in stochastic approximation [23], and generalize the procedure in [7] to momentum. In the transient phase SGDM moves quickly towards $\theta_\star$ by discarding initial conditions, and so gradients likely point in the same direction. This implies on average a positive inner product. In the stationary phase SGDM oscillates in a region around $\theta_\star$, indicating the gradients point in different directions. This implies on average a negative inner product. Thus a change in sign from positive to negative inner products is a good indicator that convergence has been reached.

Momentum introduces two significant challenges to the development of a convergence diagnostic. First, the test statistic of the diagnostic needs to be constructed. Pflug's procedure takes an inner product between successive gradients, which can be rewritten as $\frac{1}{\gamma^2}(\theta_{n+1} - \theta_n)^\top(\theta_n - \theta_{n-1})$ since by Eq. (2), $\theta_n - \theta_{n+1} = \gamma\nabla\ell(\theta_n, \xi_{n+1})$. But with momentum the updates become $(\theta_n - \theta_{n+1}) = \gamma\nabla\ell(\theta_n, \xi_{n+1}) - \beta(\theta_n - \theta_{n-1})$, by Eq. (3). It is unclear what linear combination of the gradient $\nabla\ell(\theta_n, \xi_{n+1})$ and momentum term $\beta(\theta_n - \theta_{n-1})$ should be included in the inner product. Second, regardless of what linear

Algorithm 1: Convergence diagnostic for SGDM.

input: Initial point θ_0, data $\{(x_1, y_1), (x_2, y_2), \dots\}$, $\gamma > 0$, $\beta \in [0, 1)$, final momentum $\beta' \in [0, \beta)$, heuristic convergence h, threshold $T > 0$, checking period $c > 0$, `burnin` > 0.

1 $S \leftarrow 0$; $\alpha \leftarrow 0$
2 Sample $\xi_1 \leftarrow (x_1, y_1)$
3 $\theta_1 \leftarrow \theta_0 - \gamma \nabla \ell(\theta_0, \xi_1)$
4 **for** $n \in \{2, 3, \dots\}$ **do**
5 Sample $\xi_n = (x_n, y_n)$
6 $\theta_n \leftarrow \theta_{n-1} - \gamma \nabla \ell(\theta_{n-1}, \xi_n) + \beta(\theta_{n-1} - \theta_{n-2})$
7 $\alpha, \beta \leftarrow$ `momentum_switch`$(n, \alpha, \beta, \nabla \ell(\theta_0, \xi_1), \dots, \nabla \ell(\theta_{n-1}, \xi_n))$
8 **if** $\alpha > 0$ *and* $n > \alpha +$ `burnin` **then**
9 $S \leftarrow S + \nabla \ell(\theta_{n-1}, \xi_n)^\top \nabla \ell(\theta_{n-2}, \xi_{n-1})$
10 **if** $S < 0$ *and* $n \bmod c = 0$ **then**
11 **return** θ_n
12 **end**
13 **end**
14 **end**
15 **function** `momentum_switch`$(n, \alpha, \beta, \nabla \ell(\theta_1, \xi_1), \dots, \nabla \ell(\theta_{n-1}, \xi_n))$**:**
16 **if** $h(\nabla \ell(\theta_0, \xi_1), \dots \nabla \ell(\theta_{n-1}, \xi_n)) < T$ *and* $n \bmod c = 0$ *and* $\alpha = 0$ **then**
17 $\alpha \leftarrow n$
18 $\beta \leftarrow \beta'$
19 **end**
20 **return** α, β

combination is chosen, with high momentum the test statistic can have positive expectation in the stationary phase. This is a serious issue because a negative expectation allows the use of a threshold of zero. A zero threshold is attractive because it is independent of data distribution and loss function. If the inner products are expected positive in the stationary phase, the threshold to declare convergence now depends on these factors and would require an additional estimation problem to set.

The convergence diagnostic for SGDM in Algorithm 1 effectively resolves these issues from momentum. It is defined by a random variable S (line 9) which keeps the running average of the inner product $\nabla \ell(\theta_n, \xi_{n+1})^\top \nabla \ell(\theta_{n-1}, \xi_n)$ of the gradient at successive iterates. In Section III we provide theory which shows that this choice of inner product is more easily able to attain the desired negative expectation, and is more robust to the momentum β. To remedy the high momentum issue, β is automatically reduced (lines 15-20) at a point determined by an optimization heuristic to be close to the stationary phase. This does not greatly affect the convergence rate as momentum is most useful in the transient phase. We can think of the momentum reduction point as a noisy estimate of convergence, followed by a more accurate estimate with the convergence diagnostic. A gradient norm based heuristic function is used (line 16). Often convergence heuristics are calculated in an online manner, hence storage is not an issue.

III. The difficulty with momentum

We now provide our theoretical and empirical justifications for the design choices in Algorithm 1 regarding the challenges due to momentum. The overall goal is to have a test statistic with negative expectation to enable the use of a practical zero threshold. We first present two theorems to address the choice of what combination of gradient $\nabla \ell(\theta_n, \xi_{n+1})$ and momentum $\beta(\theta_n - \theta_{n-1})$ should be used to construct the test statistic of the convergence diagnostic. Then we provide a corollary, and theoretical and empirical results for the quadratic loss to study the effects of high momentum on the expectation of the chosen test statistic. We now list the assumptions for the analysis.

Assumption 1. *The expected loss $f(\theta) = \mathbb{E}[\ell(\theta, \xi)]$ is strongly convex with constant c.*

Assumption 2. *The expected loss $f(\theta) = \mathbb{E}[\ell(\theta, \xi)]$ is Lipschitz-smooth with constant L.*

Assumption 3. *Theorem 1 [33] holds s.t. $\mathbb{E}[f(\theta_n) - f(\theta_\star)] \leq \gamma M$ for some $M > 0$ and large enough n.*

Assumption 4. *$\exists\ \sigma_0^2 > 0$ s.t. $\mathbb{E}[\|\nabla \ell(\theta, \xi)\|^2] > \sigma_0^2$.*

Assumption 5. *$\exists\ K > 1$ s.t. $\mathbb{E}[(\theta_n - \theta_{n-1})^\top (\theta_{n-1} - \theta_{n-2})] \geq -K\mathbb{E}[\|\theta_n - \theta_{n-1}\|^2]$ for large enough n.*

Assumptions 1 and 2 are standard in stochastic gradient methods analysis [1], [2]. Assumption 3 requires $\|\nabla f(\theta)\| \leq G$ and $\mathbb{E}[\|\nabla \ell(\theta, \xi) - \nabla f(\theta)\|^2] \leq \delta^2$ [33]. Assumption 4 posits a minimum amount of noise in the stochastic gradients. We also assume K is not too large in Assumption 5 since $\|\theta_n - \theta_{n-1}\|^2 \approx \|\theta_{n-1} - \theta_{n-2}\|^2$ in the stationary phase.

A. Constructing a test statistic

We select as the test statistic a running mean of

$$\nabla \ell(\theta_n, \xi_{n+1})^\top \nabla \ell(\theta_{n-1}, \xi_n). \tag{4}$$

In the following Theorems 2 and 3 we derive the upper bounds on the expected values of different inner products, and use these results to choose the test statistic.

Theorem 2. *Suppose Assumptions 1, 3, and 4 hold. Define $A_\beta = 1/(1 + 2\beta K + \beta^2)$. The test statistic in Eq. (4) for the diagnostic in Algorithm 1 for SGDM in Eq. (3) is bounded*

$$\mathbb{E}[\nabla \ell(\theta_n, \xi_{n+1})^\top \nabla \ell(\theta_{n-1}, \xi_n)] \leq (1 + \beta) \left[M - \frac{c}{2} \gamma \sigma_0^2 A_\beta \right].$$

□

Theorem 3. *Suppose that Assumptions 1, 3, and 4 hold. Define $\nabla \ell_{n+1} = \nabla \ell(\theta_n, \xi_{n+1})$ and $\Delta_n = (\theta_n - \theta_{n-1})$. The expectation of the alternative test statistic is bounded*

$$\begin{aligned} &\mathbb{E}[(\nabla \ell_{n+1} + \beta \Delta_n)^\top (\nabla \ell_n + \beta \Delta_{n-1})] \\ &< \left(\frac{1}{\gamma} + \frac{\beta}{\gamma} + 2\beta + \beta^2\right) \left[\gamma M - \frac{c}{2} \gamma^2 \sigma_0^2 A_\beta \right] + \beta^3 \gamma M. \end{aligned} \tag{5}$$

□

Remarks. By Theorem 3 the alternative test statistic in Eq. (5) is less likely to achieve a negative expectation in

stationarity, and thus be able to use a zero threshold, due to the added $\beta^3\gamma M > 0$ term. A choice of another constant $t \neq \beta$ in the linear combination (5) would not change this conclusion since its sign is not controlled by t. The last term would be $t^2\beta\gamma M > 0$ and not change sign. The convergence threshold for a good test statistic should not depend too much on momentum, but the alternative test statistic has increased dependence due to the $2\beta, \beta^2$ terms. Also, note that the test statistic still has dependence on the momentum. Eq. (4) reads

$$\frac{\beta}{\gamma}[\nabla\ell_{n+1}^\top(\theta_{n-1}-\theta_{n-2})] - \frac{1}{\gamma}[\nabla\ell_{n+1}^\top(\theta_n-\theta_{n-1})]. \quad (6)$$

B. Effect of high momentum

The test statistic in Eq. (4) is chosen to ensure a negative expectation in the stationary phase. The next corollary guarantees this negativity under certain conditions on the learning rate.

Corollary 4. *Consider SGDM in Eq. (3). If the learning rate satisfies $\gamma > 2M/c\sigma_0^2 A_\beta$, then,*

$$\mathbb{E}[\nabla\ell(\theta_n,\xi_{n+1})^\top\nabla\ell(\theta_{n-1},\xi_n)] < 0$$

as $n \to \infty$. The convergence diagnostic activates a.s. □

Remarks. Again, $A_\beta = 1/(1+2\beta K+\beta^2)$ is a monotonically decreasing function where $A_{\beta|\beta=0} = 1$ and $A_{\beta|\beta=1} = 1/(2+2K)$. Higher β reduces A_β, restricting the condition $\gamma > 2M/c\sigma_0^2 A_\beta$. Corollary 4 suggests that too large momentum may make the learning rate condition too prohibitive, invalidating the negative expectation in practice.

C. Quadratic loss model

After we have constructed our test statistic, we would like to gain further insight into the convergence diagnostic and the effect of high momentum. We do this by considering quadratic loss $\ell(\theta,y,x) = \frac{1}{2}(y-x^\top\theta)^2$ with gradient $\nabla\ell(\theta,y,x) = -(y-x^\top\theta)x$. Let $y = x^\top\theta_\star + \epsilon$, where ϵ are zero mean random variables $\mathbb{E}[\epsilon|x] = 0$. $\theta_0 = \theta_\star$; the procedure has started in the stationary region. The first three iterates are:

$$\begin{aligned}
\theta_1 &= \theta_\star + \gamma(y_1 - x_1^\top\theta_\star)x_1\\
\theta_2 &= \theta_1 + \gamma(y_2 - x_2^\top\theta_1)x_2 + \beta(\theta_1-\theta_\star)\\
\theta_3 &= \theta_2 + \gamma(y_3 - x_3^\top\theta_2)x_3 + \beta(\theta_2-\theta_1)
\end{aligned}$$

Three steps are taken in order for the momentum to effect both terms of the inner product. The expected value of the test statistic at θ_3 is:

$$\begin{aligned}
&\mathbb{E}[\nabla\ell(\theta_2,y_3,x_3)^\top\nabla\ell(\theta_1,y_2,x_2)]\\
=&-\gamma\mathbb{E}[\epsilon_2^2]\mathbb{E}[(x_3^\top x_2)^2] - \gamma^3\mathbb{E}[\epsilon_1^2]\mathbb{E}[(x_2^\top x_1)^2(x_3^\top x_2)^2]\\
&+\gamma^2(1+\beta)\mathbb{E}[\epsilon_1^2]\mathbb{E}[(x_2^\top x_1)(x_3^\top x_1)(x_3^\top x_2)] \quad (7)
\end{aligned}$$

From Eq. (7) in the $\gamma^2(1+\beta)$ term we see that momentum contributes positively to the test statistic, and if it is too large the expectation is positive. $\mathbb{E}[(x_2^\top x_1)(x_3^\top x_1)(x_3^\top x_2)] = tr(\mathbb{E}[(x_1x_1^\top)(x_2x_2^\top)(x_3x_3^\top)]) > 0$ by application of trace and $x_1x_1^\top$ is positive definite. The results are generalized in the following theorem.

Theorem 5. *Suppose that the loss is quadratic, $\ell(\theta) = 1/2(y-x^\top\theta)^2$. Let x_n and x_{n+1} be two iid vectors from the distribution of x. Let $A = \mathbb{E}[(x_nx_{n+1}^\top)(x_n^\top x_{n+1})]$, $B = \mathbb{E}[(x_nx_n^\top)(x_n^\top x_{n+1})^2]$, $\sigma_{quad}^2 = \mathbb{E}[\epsilon_n^2]$, $d^2 = \mathbb{E}[(x_n^\top x_{n+1})^2]$. Then for $\gamma > 0$, we have*

$$\begin{aligned}
&\mathbb{E}[\nabla\ell(\theta_n,\xi_{n+1})^\top\nabla\ell(\theta_{n-1},\xi_n)|\theta_{n-1},\theta_{n-2}]\\
=&(\theta_{n-1}-\theta_\star)^\top(A-\gamma B)(\theta_{n-1}-\theta_\star) - \gamma\sigma_{quad}^2 d^2\\
&+(\theta_{n-1}-\theta_\star)^\top(\beta A)(\theta_{n-1}-\theta_{n-2}).
\end{aligned}$$

□

Remarks. The momentum term βA only becomes significant in the stationary phase when $\|\theta_{n-1}-\theta_\star\|^2 \approx \|\theta_{n-1}-\theta_{n-2}\|^2$. It makes an expected positive contribution as θ_{n-1} and θ_{n-2} are more likely to be on opposite sides of $\theta_\star$ in the stationary phase. Otherwise, progress towards $\theta_\star$ is still being made in the transient phase.

In the transient phase the bias dominates, resulting in an expected positive contribution from $(A-\gamma B)$ to the test statistic. In the stationary phase the variance dominates, resulting in an expected negative contribution from $-\gamma\sigma_{quad}^2 d^2$. Theorem 5 supports that in stationarity momentum β contributes positively to the test statistic, and β which is too high makes the expected value of the test statistic positive. We empirically validate this effect quadratic loss.

We sample 1000 data points $x \sim N(0, I_{20})$, set $y = x^\top\theta_\star + \epsilon$ with $\epsilon \sim N(0,1)$, and $\theta_{\star,i} = (-1)^i\, 2\exp(-0.7i)$ for $i = 1,2,\ldots,20$. SGDM is run with batch size 25 for 50 epochs. The stationary phase is marked when the MSE with respect to $\theta_\star$ has flattened out. We set $\beta = 0.2$ and $\beta = 0.9$ to contrast low and high momentum. Both settings attain equivalent MSE. After 25 independent runs of SGDM we retrieve values of the test statistic in stationarity of -6.71 and 2.77 for $\beta = 0.2$ and $\beta = 0.9$ respectively, supporting the observations from Corollary 4 and Theorem 5: the expectation of the test statistic becomes positive with too large momentum.

IV. Distribution of Inner Products

The convergence diagnostic crucially relies upon the negative expectation of its test statistic. An important question emerges: **What drives the expectation of the inner product negative? What is its relation to momentum?** At a high level, there is an oscillation in the stationary phase driven by the dominating variance of the stochastic gradients. This oscillation interacts with the curvature of the loss function around $\theta_\star$, driving the expectation of inner products negative. We propose a more refined view which also helps explain the observed sensitivity of the expectation to high momentum.

Proposition 6. *In the stationary phase there are a small number of key iterates which drive the expected inner product in Eq. (4) negative. Consider the decomposition of the stochastic gradient $\nabla\ell(\theta_n,\xi_{n+1}) = \mathbb{E}[\nabla\ell(\theta_n,\xi_{n+1})]+\sigma^2$ into its true gradient and noise. A majority of inner products $\nabla\ell(\theta_n,\xi_{n+1})^\top\nabla\ell(\theta_{n-1},\xi_n)$ are mean zero due to the dominance of the noise term σ^2.* ***The expectation is negative due to a relatively sparse number of inner products which have high magnitude and negative sign.*** □

Remarks. There are some iterates θ_n in the stationary phase which get relatively farther from $\theta_\star$. This would require a relatively large gradient $\nabla \ell(\theta_{n-1}, \xi_n)$ pointing generally away from $\theta_\star$. The following iteration would result in an also large gradient $\nabla \ell(\theta_n, \xi_{n+1})$ pointing generally back towards $\theta_\star$ due to the curvature of the loss.

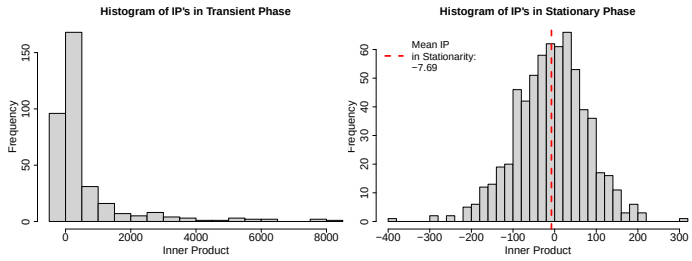

Fig. 1: Histogram of the inner product of successive gradients (Eq. (4)). Left: Transient phase. Right: Stationary phase. Training settings from the low β quadratic setting in Section III-C.

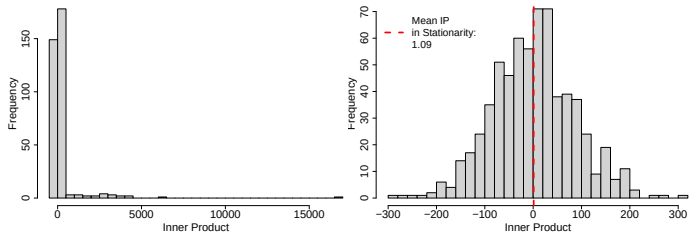

Fig. 2: Histogram of the inner product of successive gradients with high momentum $\beta = 0.8$ and no momentum reduction. Left panel: Transient phase. Right panel: Stationary phase. Quadratic loss model from Section III-C.

We first provide empirical evidence to support Proposition 6. The low β quadratic setting from Section III-C is used, and SGDM run for 20 epochs. Fig. 1 plots histograms of the inner products from Eq. (4) in the transient and stationary phase. The phase transition is chosen by monitoring MSE with respect to $\theta_\star$. These results are robust across loss functions and parameter settings. In the transient phase the inner products are in majority positive with positive skew. This makes sense as when θ_n is far from $\theta_\star$, the bias dominates and gradients are likely pointed in the same direction. In the stationary phase we observe a unimodal distribution around zero with a longer negative tail. **The distribution of inner products in the stationary phase is of utmost interest.** The expectation is negative consistently across many experiments. Yet, the magnitude of the variance exceeds the magnitude of the mean, indicating a high frequency of iterates with mean zero inner product. The larger negative tail–specifically the small number of inner products around -400–supports the existence of a small number of key iterates as stated in Proposition 6. This empirical observation is even more striking when you compare similar histograms when the momentum is high $\beta = 0.8$ and keeping all other settings the same, in Fig. 2. The previously observed asymmetry in the stationary phase is now gone. Thus with high momentum (or without momentum reduction), the convergence diagnostic cannot work as there is no clear signal from the test statistic.

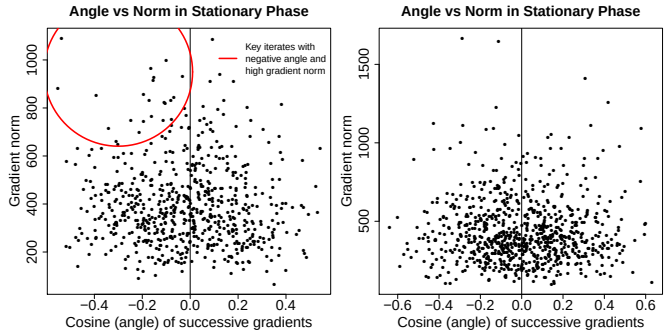

Fig. 3: Cosine similarity vs. gradient norm for SGDM in the stationary phase. The red circle indicates those key inner products with negative angle and high gradient norm. Left: Training settings from the low β quadratic setting in Section III-C. Right: high momentum $\beta = 0.8$. The inner product distribution is symmetric in this case.

In Fig. 3, we provide the empirical evidence by plotting $\|\nabla \ell(\theta_n, \xi_{n+1})\|_2^2$ and cosine$(\nabla \ell(\theta_n, \xi_{n+1}), \nabla \ell(\theta_{n-1}, \xi_n))$ of successive gradients in the stationary phase. We first look at the low β quadratic setting in Section III-C. A red circle is drawn to identify iterates with high magnitude and negative angle, the key iterates described in Proposition 6. We observe such key iterates exist and drive the expectation negative, consistently across loss functions and parameter settings. With high momentum we have empirically observed that these key inner products with high magnitude and negative angle disappear. Thus Proposition 6 helps explaining the observed sensitivity of the expected test statistic to high momentum.

A. *Variance bounds*

We now provide the theory to support Proposition 6. We show that in the stationary phase, the magnitude of the variance dominates the magnitude of the mean for the test statistic. A relatively large variance of the inner products $\nabla \ell(\theta_n, \xi_{n+1})^\top \nabla \ell(\theta_{n-1}, \xi_n)$ suggests that a majority of iterates are dominated by the variance of the stochastic gradients. A relatively small mean for the inner products suggests that a minority of iterates drive the expectation. Even though in the stationary phase SGDM is trapped in a bounded region, and the expected test statistic driven by a sparse number of key iterates, there is still significant room for random motion.

Theorem 7. *Consider the SGDM procedure in Eq. (3). Suppose that Assumptions 1, 2, 3, 4, and 5 hold. Define* $IP = \nabla \ell(\theta_n, \xi_{n+1})^\top \nabla \ell(\theta_{n-1}, \xi_n)$. *Then,*

$$\frac{Var[IP]}{\mathbb{E}[IP]^2} \geq \frac{(M - L\gamma\sigma_0^2 A_\beta)^2}{M^2(1 + 8L/c)^2} - 1.$$

□

Corollary 8. *Consider the SGDM procedure in Eq. (3) and a fixed scaling factor $\lambda > 2$. Set the learning rate $\gamma = 2tM/L\sigma_0^2 A_\beta$ with $t \geq 1 + \sqrt{\lambda}(1 + 4L/c)$. Then, $Var[IP] \geq (\lambda - 1)\ \mathbb{E}[IP]^2$.* □

Remarks. The results hold regardless of the sign of $\mathbb{E}[IP]$, and show that the variance of the statistic upper bounds the squared mean. Corollary 8 specifies that a greater learning rate increases the variance bound. A larger learning rate indeed increases the radius of the stationary region.

We have seen that Theorem 7 and Corollary 8, along with Figures 1, 2 and 3, provide theoretical and empirical support for Proposition 6. While the bound in Theorem 7 is more robust to β, it is still unable to provide practical guidance as the data dependent constants M, L, c, σ_0^2, and A_β must still be estimated. The convergence diagnostic monitors a certain signal in the gradients. In Section III we have shown that this signal can be sensitive to high momentum, and in this Section we have shown that the signal may be sparse within other gradient noise. Currently, we believe that an empirical mean is still the best way to capture this gradient signal, with the simple but effective automatic reduction in Algorithm 1 to combat the negative effects of high momentum.

V. Numerical Experiments

We now evaluate the convergence diagnostic in Algorithm 1 on synthetic data with quadratic loss and phase retrieval [9]. For phase retrieval let $\ell(\theta, y, x) = 1/4[(x^\top\theta)^2 - y]^2$ with $x \sim N(0, I_{20})$ and $y = (x^\top\theta_\star)^2$. $\theta_{\star,i} = (-1)^i \times 2\exp(-0.7i)$ for $i = 1, \ldots, 20$ and $N = 10^3$. The checking period c is every epoch. Due to non-convexity, we record the training runs where SGDM has entered a good minima.

There are two failure modes: the convergence diagnostic can activate too early, or too late. Let θ_n be the estimate when the convergence diagnostic has activated. If the diagnostic activates too early then the error is too high, i.e., $\|\theta_n - \theta_\star\|^2 > \eta$ for some threshold value. η is set as a tight upper bound on the error observed in the stationary phase across many runs. Let $K = (n-k)/n$ such that $\|\theta_k - \theta_n\|^2 = \eta$ and $k \leq n$. If the diagnostic activates too late, it can waste unnecessary computation and we expect θ_n to be far into the stationary phase, and thus $n - k$ to be a significant portion of n, i.e., $K > \kappa$ for some threshold value. κ is set as a tight lower bound on the K calculated by running SGDM into the stationary phase. Table I displays the results of 100 independent runs with the percentage of type I errors (too early), type II errors (too late), and good diagnostic activations. Low and high momentum settings are used for quadratic loss and phase retrieval. Empirically the type I errors are small, while the type II errors are a larger concern. This observation on the higher frequency of type II errors is corroborated by [7]. Encouragingly we see that the automatic momentum reduction in Algorithm 1 has enabled the convergence diagnostic to be robust to momentum. High momentum settings have little or no effect on the Type I and II error rates of the convergence diagnostic. Contrast Table I with the results in Section III, where the sign of the test statistic in the stationary phase was shown to be sensitive to momentum. Additionally, in approximately 50-70% of all Type II errors the diagnostic activated only moderately late and in 10-20% of Type II errors the diagnostic activated late.

TABLE I: Empirical evaluation of the convergence diagnostic. SGDM run for 20 epochs with batch size 20 over 100 independent runs. Quadratic low β (Q-Low) set $\beta = 0.2$, $\gamma = 10^{-2}$, $\eta = 10^{-3}$, $\kappa = 0.65$. Quadratic high β (Q-High) set $\beta = 0.8$, $\gamma = 10^{-2}$, $\eta = 2 \times 10^{-3}$, $\kappa = 0.30$. Phase retrieval low β (PR-Low) set $\beta = 0.2$, $\gamma = 10^{-2}$, $\eta = 10^{-2}$, $\kappa = 0.6$. Phase retrieval high β (PR-High) set $\beta = 0.8$, $\gamma = 10^{-2}$, $\eta = 10^{-2}$, $\kappa = 0.65$.

	Type I (too early)	Type II (too late)	Good activation
Q-Low	1%	22%	77%
Q-High	0%	17%	83%
PR-Low	1%	17%	82%
PR-High	0%	16%	84%

VI. Application: An Automatic Learning Rate Schedule

The convergence diagnostic has a natural application in automating the learning rate schedule. Learning rate tuning has a major impact on the performance of optimization methods. Tuning is typically a manual process requiring many training runs. The benefit of an automatic learning rate is to greatly reduce the amount of supervision and number of training runs.

Algorithm 2: SGDM with automatic learning rate

input: Algorithm 1 $\texttt{SGDM}(\theta, \gamma)$, initial and minimum stepsize γ_0, γ_{min}, reduction $\rho \in (0, 1)$

$\gamma \leftarrow \gamma_0$
while $\gamma > \gamma_{min}$ **do**
  $\theta \leftarrow \texttt{SGDM}(\theta, \gamma)$ and $\gamma \leftarrow \rho \times \gamma$
end
return θ

Algorithm 2 displays an automatic learning rate based on our diagnostic algorithm 1. SGDM with constant learning rate moves quickly towards $\theta_\star$ but cannot improve beyond distance $O(\gamma)$, as suggested by Theorem 1. The convergence diagnostic is used to detect stationarity, after which the learning rate is reduced $\gamma \leftarrow \rho\gamma$ and a smaller radius $O(\rho\gamma)$ of stationarity achieved, with $\rho \in (0, 1)$. Algorithm 2 takes advantage of the speedup afforded to constant rate in the transient phase, while avoiding the trade-off cost of a larger stationary region by reducing the learning rate. It is common to use a constant learning rate and manually decay several times [13], [17].

A major benefit of automatic hyper-parameter tuning is robustness to a potentially misspecified initial setting. We train logistic regression on benchmark datasets MNIST and Online News Popularity (from UCI repository), for a variety of initial

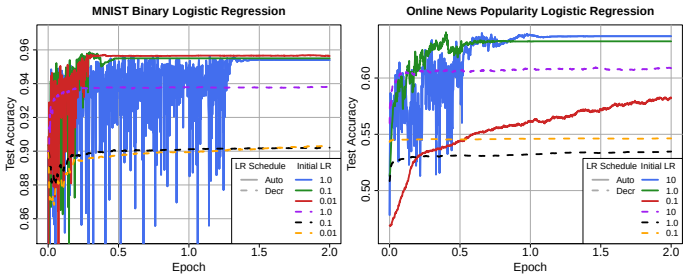

Fig. 4: Binary logistic regression with SGDM using Algorithm 2 and decreasing rate $\gamma = \gamma_0/n$, $\beta = 0.8$. Left: MNIST. Right: Online News Popularity.

learning rates γ_0. In Fig. 4 the accuracy on a held out test set is compared between the automatic rate in Algorithm 2 and a decreasing rate $\gamma = \gamma_0/n$ for $\gamma_0^{mnist} \in \{1.0, 0.1, 0.01\}$ in MNIST and $\gamma_0^{news} \in \{10, 1.0, 0.1\}$ in Online News. The findings are consistent across both datasets. The automatic learning rate is significantly more robust to initial conditions and achieves higher test accuracy than the decreasing rate.

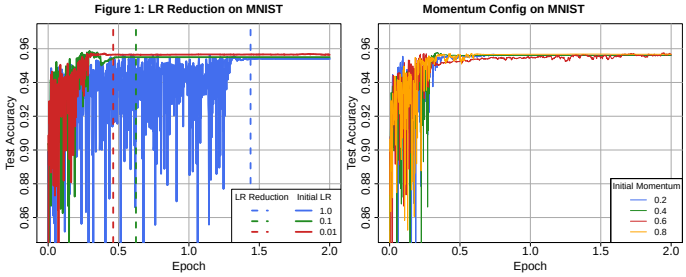

Fig. 5: Left: Vertical lines marks the diagnostic activation and learning rate reduction on MNIST. Right: SGDM using Algorithm 2 and varying momentum $\beta \in \{0.2, 0.4, 0.6, 0.8\}$.

In Fig. 5 we plot SGDM using Algorithm 2 and vertical lines to mark the diagnostic activation. Fig. 5 also shows SGDM using Algorithm 2 with varying momentum constant, illustrating that Algorithm 2 is robust to the choice of momentum. The function $h()$ in Algorithm 1 to set the change in momentum was the mean squared distance between successive iterates. However, any convergence heuristic can be used for $h()$. We show that $h()$ works consistently in Fig. 6. Regardless of the initial momentum, SGDM using Algorithm 2 on MNIST reduces the initial momentum at a consistent point, helping to ensure a consistent activation of the convergence diagnostic.

Experiments were performed with constant rate $\gamma = \gamma_0$, however the stationary region was large enough to result in significant test accuracy fluctuations for a given γ_0. The fact that SGDM using Algorithm 2 is able to achieve competitive performance validates our convergence diagnostic. As further evidence, Fig. 7 plots the test statistic of the convergence diagnostic along with the test accuracy for SGDM on MNIST. The vertical line indicates the activation of the convergence diagnostic, when the test statistic becomes negative. The diagnostic activates just as the test accuracy flattens out.

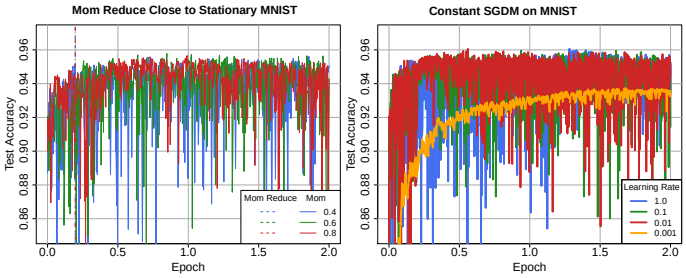

Fig. 6: Left: Vertical line marks a consistent momentum reduction using convergence heuristic of $\|\theta_n - \theta_{n-1}\|^2$. Right: Binary logistic regression with SGDM using Algorithm 2 and constant learning rate $\gamma_0 \in \{1.0, 0.1, 0.01, 0.001\}$ on MNIST.

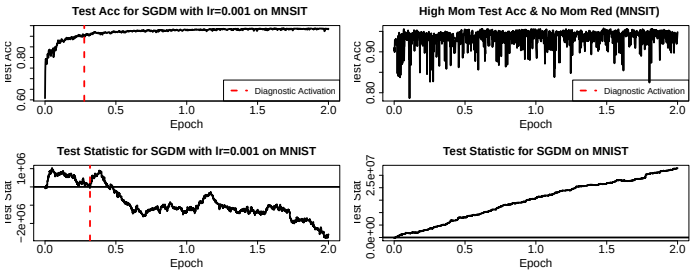

Fig. 7: Test accuracy and convergence diagnostic test statistic from Algorithm 1 on MNIST. Left: With momentum reduction, the vertical line marks the diagnostic's activation, which coincides with the test accuracy flattening out. Right: Without momentum reduction and $\beta = 0.8$, flat test accuracy and convergence of SGDM, no activation since the test is positive.

We conduct an ablation study to understand the negative effects of high momentum on our convergence diagnostic. Fig. 7 plots the test accuracy and test statistic by removing the momentum reduction component of Algorithm 1 and using SGDM with high momentum $\beta = 0.8$ on MNIST. While the test accuracy has plateaued, the convergence diagnostic does not activate because its test statistic is perpetually positive.

We plot the test statistic of Algorithm 1 in Fig. 8 with no momentum reduction and increasing momentum $\beta \in \{0.2, 0.4, 0.6, 0.8\}$. Momentum has a proportional relationship

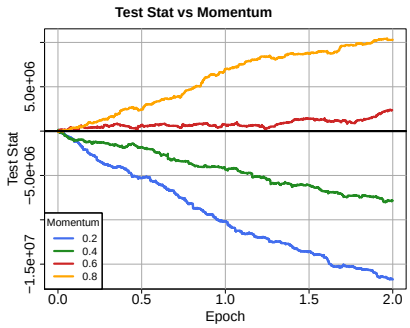

Fig. 8: Test statistic for different values of momentum $\beta \in \{0.2, 0.4, 0.6, 0.8\}$. Higher momentum increases the slope of the test statistic, indicating an even greater difficulty for the convergence diagnostic to detect the stationary phase.

with the slope of the test statistic. A positive slope indicates that the test statistic is not negative upon convergence, and thus the convergence diagnostic without the momentum reduction is ineffective for higher momentum $\beta \in \{0.6, 0.8\}$ on MNIST.

VII. Conclusion

In this paper we focus on detecting the transition phase of SGDM to the stationary phase, inspired by the stopping times techniques in stochastic approximation. Momentum introduces challenges in the construction and operation of the test statistic for the convergence diagnostic. We present the theory and experiments which support that high momentum alters the trajectory of the stochastic gradients which the diagnostic monitors. In addition we show the dynamics of SGDM in stationarity are largely random with a sparse number of key iterates behaving in an informative way, captured by the diagnostic. The proposed automatic momentum reduction technique resolves the issues with high momentum. Empirical results demonstrate that the diagnostic has few type I errors and a reasonably small number of type II errors, and thus reliably detects the stationary phase. We present an application on automatic learning rate, robust to initial conditions. Future works including extensions to adaptive stochastic gradient methods such as "Local AMSGrad" [8] and "Optimistic AMSGrad" [31] are of great interest.

References

[1] Francis R. Bach and Eric Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 451–459, Granada, Spain, 2011.

[2] Francis R. Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In *Advances in Neural Information Processing Systems (NIPS)*, pages 773–781, Lake Tahoe, NV, 2013.

[3] Avrim Blum, Adam Kalai, and John Langford. Beating the hold-out: Bounds for k-fold and progressive cross-validation. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (COLT)*, pages 203–208, Santa Cruz, CA, 1999.

[4] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT)*, pages 177–186, 2010.

[5] Léon Bottou. Stochastic gradient descent tricks. In Grégoire Montavon, Genevieve B. Orr, and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade - Second Edition*, volume 7700 of *Lecture Notes in Computer Science*, pages 421–436. Springer, 2012.

[6] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[7] Jerry Chee and Panos Toulis. Convergence diagnostics for stochastic gradient descent with constant learning rate. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1476–1485, Playa Blanca, Lanzarote, Canary Islands, Spain, 2018.

[8] Xiangyi Chen, Xiaoyun Li, and Ping Li. Toward communication efficient adaptive gradient method. In *Proceedintgs of ACM-IMS Foundations of Data Science Conference (FODS)*, pages 119–128, Virtual Event, 2020.

[9] Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: fast global convergence for nonconvex phase retrieval. *Math. Program.*, 176(1-2):5–37, 2019.

[10] Bernard Delyon and Anatoli B. Juditsky. Accelerated stochastic approximation. *SIAM Journal on Optimization*, 3(4):868–881, 1993.

[11] Yu M Ermoliev and RJ-B Wets. *Numerical techniques for stochastic optimization*. Springer-Verlag, 1988.

[12] Rong Ge, Sham M. Kakade, Rahul Kidambi, and Praneeth Netrapalli. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 14951–14962, 2019.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, 2016.

[14] Nitish Shirish Keskar and Richard Socher. Improving generalization performance by switching from Adam to SGD. *arXiv preprint arXiv:1712.07628*, 2017.

[15] Harry Kesten. Accelerated stochastic approximation. *The Annals of Mathematical Statistics*, 29(1):41–59, 1958.

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, 2015.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, 2017.

[18] Hunter Lang, Lin Xiao, and Pengchuan Zhang. Using statistics to automate stochastic optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9536–9546, Vancouver, Canada, 2019.

[19] Nicolas Loizou and Peter Richtárik. Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *Computational Optimization and Applications*, pages 1–58, 2020.

[20] Noboru Murata. A statistical study of on-line learning. *Online Learning and Neural Networks. Cambridge University Press, Cambridge, UK*, pages 63–92, 1998.

[21] Deanna Needell, Nathan Srebro, and Rachel Ward. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Math. Program.*, 155(1-2):549–573, 2016.

[22] Arkadi Nemirovski, Anatoli B. Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[23] Georg Ch Pflug. Non-asymptotic confidence bounds for stochastic approximation algorithms with constant step size. *Monatshefte für Mathematik*, 110(3):297–314, 1990.

[24] B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[25] Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos, and Alexander J. Smola. On variance reduction in stochastic gradient descent and its asynchronous variants. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2647–2655, Montreal, Canada, 2015.

[26] Nicolas Le Roux, Mark Schmidt, and Francis R. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, Lake Tahoe, NV, 2012.

[27] Matteo Sordello and Weijie Su. Data-adaptive learning rate selection for stochastic gradient descent using convergence diagnostic. Joint Statistics Meeting, 2019.

[28] Weijie J Su and Yuancheng Zhu. Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent. *arXiv preprint arXiv:1802.04876*, 2018.

[29] Panos Toulis and Edoardo M. Airoldi. Scalable estimation strategies based on stochastic approximations: classical results and new insights. *Stat. Comput.*, 25(4):781–795, 2015.

[30] Panos Toulis, Edoardo M Airoldi, et al. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, 2017.

[31] Jun-Kun Wang, Xiaoyun Li, Belhal Karimi, and Ping Li. An optimistic acceleration of amsgrad for nonconvex optimization. *arXiv preprint arXiv:1903.01435*, 2020.

[32] Sho Yaida. Fluctuation-dissipation relations for stochastic gradient descent. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, 2019.

[33] Tianbao Yang, Qihang Lin, and Zhe Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. *arXiv preprint arXiv:1604.03257*, 2016.

[34] George Yin. Stopping times for stochastic approximation. In *Modern Optimal Control: A Conference in Honor of Solomon Lefschetz and Joseph P. LaSalle*, pages 409–420, 1989.

[35] Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-first International Conference on Machine Learning (ICML)*, Banff, Canada, 2004.