Understanding and Detecting Convergence for Stochastic Gradient Descent with Momentum

Anonymous Authors[1]

Abstract

Convergence detection of iterative stochastic optimization methods is of great practical interest. In this paper, we consider stochastic gradient descent (SGD) with a constant learning rate and momentum. Our study shows 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 a potentially misspecified initial rate.

1. 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 stochastic gradient descent (SGD):

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

is a powerful alternative (Bottou, 2010; 2012; Toulis & Airoldi, 2015; Zhang, 2004). 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 (Reddi J. Sashank, 2015).

Momentum or Heavy Ball SGD (SGDM) can offer significant speedups (Polyak, 1964):

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

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. The learning rate can be decreasing (e.g., $\propto 1/n$), or constant. Decreasing learning rates are commonly used in the literature to attain theoretical convergence guarantees. However in practice constant learning rates are commonly used 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 (Murata, 1998). With a decreasing learning rate, the transient phase can be long, and impractically so if the learning rate is just slightly misspecified (Nemirovski et al., 2009; Toulis et al., 2017). 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.

1.1. 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 (Murata, 1998). However, until recently there has been little work in developing principled statistical methods for convergence detection which can guide empirical practice. Heuristics from optimization the-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

ory 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 (Bottou et al., 2016; Ermoliev & Wets, 1988). 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 (Blum et al., 1999; Bottou, 2012). 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 (Pflug, 1990; Yin, 1989). One noteworthy method by Pflug (1990) 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 (Delyon & Juditsky, 1993; Kesten, 1958; Roux et al., 2012).

There has been recent interest in principled convergence detection for stochastic gradient methods, and automated step decay learning rates. Work by Chee & Toulis (2018) 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. Keskar & Socher (2017) developed a procedure to automatically switch from Adam (Kingma & Ba, 2015) to SGD. Sordello & Su (2019) propose a modified splitting procedure (Su & Zhu, 2018) to detect the stationary phase for SGD and implement a robust learning rate schedule. Lang et al. (2019) use a Markov chain t-test and a stationarity condition by Yaida (2019) to automatically reduce the learning rate for SGDM. Ge et al. (2019) analyze the step decay learning rate schedule in least squares regression and show its optimality over polynomially decaying rates.

1.2. Our contributions

Section 2 presents a statistical convergence diagnostic for SGDM (stochastic gradient descent with momentum) and explains the significance of the challenges introduced by momentum. In Section 3 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 4 what drives the test statistic by analyzing the distribution of the inner product of successive gradients in the stationary phase. In Section 5 we provide empirical type I and type II error rates on simulated data experiments. Section 6 presents an application of the convergence diagnostic to an automatically tuned learning rate schedule, with experiments on benchmark data sets.

2. Convergence diagnostic

The proposed convergence diagnostic aims to detect the transition from the transient to the stationary phase. We first present 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 2.1 ((Yang et al., 2018)). *Suppose that 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, it holds that*

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

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 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 2.1 suggests that 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 (Moulines & Bach, 2011; Needell et al., 2014).

While convergence analyses such as Theorem 2.1 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 route. In the next section we develop a practical statistical diagnostic test to estimate the phase transition and detect convergence of SGDM in a much simpler way.

2.1. Modified Pflug diagnostic

We present a convergence diagnostic for SGDM in Algorithm 1. We draw upon Pflug's procedure in stochastic approximation (Pflug, 1990), and generalize the procedure in Chee & Toulis (2018) 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 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 3) 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 3 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 5-8) at a point determined by an optimization heuristic to be close to the stationary phase. This does not greatly effect 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 was used (line 6). Often convergence heuristics are calculated in an online manner, so storage is not an issue.

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.

$S \leftarrow 0$; $\alpha \leftarrow 0$
Sample $\xi_1 \leftarrow (x_1, y_1)$
$\theta_1 \leftarrow \theta_0 - \gamma\nabla\ell(\theta_0, \xi_1)$
for $n \in \{2, 3, \dots\}$ **do**
  Sample $\xi_n = (x_n, y_n)$
  $\theta_n \leftarrow \theta_{n-1} - \gamma\nabla\ell(\theta_{n-1}, \xi_n) + \beta(\theta_{n-1} - \theta_{n-2})$
  $\alpha, \beta \leftarrow$ `momentum_switch`$(n, \alpha, \beta, \nabla\ell(\theta_0, \xi_1), \dots, \nabla\ell(\theta_{n-1}, \xi_n))$
  if $\alpha > 0$ *and* $n > \alpha +$ `burnin` **then**
    $S \leftarrow S + \nabla\ell(\theta_{n-1}, \xi_n)^\top\nabla\ell(\theta_{n-2}, \xi_{n-1})$
    if $S < 0$ *and* $n \bmod c == 0$ **then**
      return θ_n
function `momentum_switch`$(n, \alpha, \beta, \nabla\ell(\theta_1, \xi_1), \dots, \nabla\ell(\theta_{n-1}, \xi_n))$ **:**
  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**
    $\alpha \leftarrow n$
    $\beta \leftarrow \beta'$
  return α, β

3. The difficulty with momentum

We now provide theoretical and empirical justification for the design choices in Algorithm 1 regarding the challenges due to momentum. The overall goal is 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 in quadratic loss to study the effects of high momentum on the expectation of the chosen test statistic. We now list the assumptions.

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 2.1 (Yang et al., 2018) 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.*

Remarks. Assumptions 1 and 2 are standard in analysis of stochastic gradient methods (Bach & Moulines, 2013; Moulines & Bach, 2011). Assumption 3 requires $\|\nabla f(\theta)\| \leq G$ and $\mathbb{E}[\|\nabla\ell(\theta,\xi) - \nabla f(\theta)\|^2] \leq \delta^2$ (Yang et al., 2018). Assumption 4 posits a minimum amount of noise present in the stochastic gradients. In Assumption 5 it is reasonable to assume K is not too large because $\|\theta_n - \theta_{n-1}\|^2 \approx \|\theta_{n-1} - \theta_{n-2}\|^2$ in the stationary phase.

3.1. 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 3.1 and 3.2 we derive upper bounds on the expected values of different inner products, and use these results to choose the test statistic.

Theorem 3.1. *Suppose that Assumptions 1, 3, and 4 hold. Define $A_\beta = 1/(1+2\beta K+\beta^2)$. The test statistic in Eq. (4) for the convergence diagnostic in Algorithm 1 for SGDM in Eq. (3) is bounded*

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

□

Theorem 3.2. *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*

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

Remarks. By Theorem 3.2 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. This is because of the additional $\beta^3\gamma M > 0$ term. A choice of another constant t other than β in the linear combination in Eq. (5) would not change this conclusion. The sign of the bound is not controlled by t, and 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.

Note that the test statistic still has dependence on the momentum. Eq. (4) can be rewritten as

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

3.2. Effect of high momentum

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

Corollary 3.3. *Consider SGDM in Eq. (3). 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 $n \to \infty$. And thus the convergence diagnostic activates almost surely. □

Remarks. Again, $A_\beta = 1/(1 + 2\beta K + \beta^2)$. A_β 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 3.3 suggests that too large momentum may make the learning rate condition too prohibitive, invalidating the negative expectation in practice.

3.3. Quadratic loss model

Now that we have constructed our test statistic, we want to gain 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 $\mathbb{E}[\epsilon|x] = 0$. Consider when $\theta_0 = \theta_\star$; 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:

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

The full derivation can be found in the supplement. 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.

	Low β	High β
Test Statistic in Stationarity	-6.71	2.77

Table 1. Mean test statistic in stationarity across 25 independent runs with $\gamma = 10^{-2}$. Low β setting has $\beta = 0.2$. High β setting has $\beta = 0.8$.

Theorem 3.4. *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 $A = \mathbb{E}[(x_n x_{n+1}^\top)(x_n^\top x_{n+1})]$, $B = \mathbb{E}[(x_n x_n^\top)(x_n^\top x_{n+1})^2]$, $\sigma^2_{quad} = \mathbb{E}[\epsilon_n^2]$, $d^2 = \mathbb{E}[(x_n^\top x_{n+1})^2]$. Then for $\gamma > 0$, we have*

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

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 $-\gamma\sigma^2_{quad}d^2$. Theorem 3.4 supports that in stationarity momentum β contributes positively to the test statistic, and β that is too high makes the expected value of the test statistic positive.

We empirically validate the effect of high momentum on the test statistic for 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 $i = 1, \ldots, 20$. 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. Table 1 reports the mean test statistic in stationarity across 25 independent runs of SGDM. We set $\beta = 0.2$ and $\beta = 0.9$ to contrast low and high momentum. Both settings attain equivalent MSE. The experimental results in Table 1 support the observations from Corollary 3.3 and Theorem 3.4: that the expectation of the test statistic becomes positive with too large momentum.

4. A distribution analysis 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?** 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 4.1. *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.*** *In these few key inner products the true gradient dominates because of an interaction with the loss curvature.* □

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.

We first provide empirical evidence to support Proposition 4.1. The low β quadratic setting from Section 3.3 is used, and SGDM run for 20 epochs. In Figure 1 are 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 have been observed to be robust across loss functions and parameter settings. In the transient phase the inner products are majority positive and have positive skew. This makes sense as when θ_n is far away 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. **What is interesting is the distribution of inner products in the stationary phase.** The expectation is negative, and consistently so across many experiments. But the magnitude of the variance exceeds the magnitude of the mean by an order of magnitude. The high magnitude variance indicates 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 4.1.

Figure 2 provides further empirical evidence by plotting the magnitude $\|\nabla\ell(\theta_n, \xi_{n+1})\|_2^2$ and angle $Cosine(\nabla\ell(\theta_n, \xi_{n+1}), \nabla\ell(\theta_{n-1}, \xi_n))$ of successive gradients for SGDM in the stationary phase. Again, we use settings from the low β quadratic setting in Section 3.3. A red circle is drawn to identify iterates with high magnitude and negative angle, exactly those key iterates described in

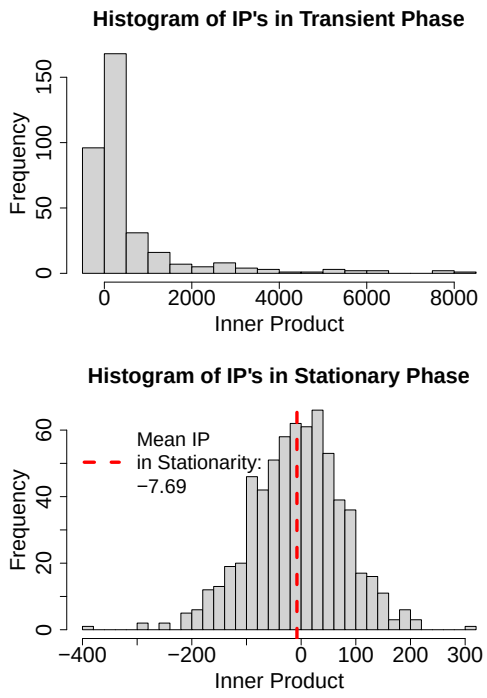

Figure 1. (Top) Histogram of the inner product of successive gradients (Eq. (4)) for SGDM in the transient phase. (Bottom) Histogram of the inner products for SGDM in the stationary phase. Training settings from the low β quadratic setting in Section 3.3.

Proposition 4.1. We see such key iterates exist and drive the expectation negative. These results have been observed to be robust 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 4.1 helps explain the observed sensitivity of the expected test statistic to high momentum.

4.1. Variance bounds

We now provide theory which supports Proposition 4.1. We show that in the stationary phase, the magnitude of the variance dominates the magnitude of the mean for the test statistic of the convergence diagnostic. 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 gradient. 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.

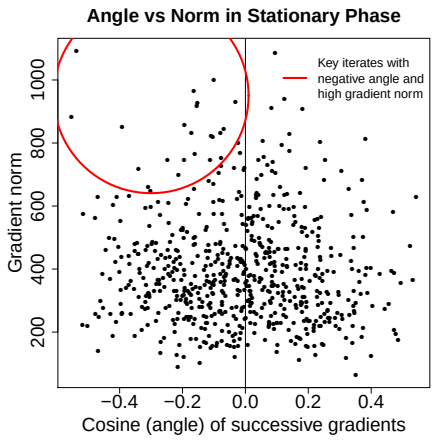

Figure 2. Cosine similarity vs gradient norm for SGDM in the stationary phase. The red circle indicates those key inner products with negative angle and high gradient norm. Training settings from the low β quadratic setting in Section 3.3.

Theorem 4.2. *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 4.3. *Consider the SGDM procedure in Eq. (3). Fix a 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 the expectation of inner products, and show that the variance of the test statistic upper bounds the squared mean. Corollary 4.3 specifies that a greater learning rate increases the variance bound. This makes sense as a larger learning rate increases the radius of the stationary region.

We have seen that Theorem 4.2 and Corollary 4.3, along with Figures 1 and 2, provide theoretical and empirical support for Proposition 4.1. While the bound in Theorem 4.2 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 3 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.

	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%

Table 2. Empirical evaluation of the convergence diagnostic in Algorithm 1 over 100 independent runs for each experimental setting. SGDM run for 20 epochs with batch size 20. 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 $\beta = 0.8$, $\gamma = 10^{-2}$, $\eta = 10^{-2}$, $\kappa = 0.65$.

5. Numerical experiments

We now evaluate the convergence diagnostic in Algorithm 1 on synthetic data experiments in settings of quadratic loss and phase retrieval (Chen et al., 2018). The quadratic loss setting is described in Section 3.3. 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 $i = 1, \ldots, 20$. The number of data points $N = 10^3$. The checking period c is every epoch. Due to non-convexity in phase retrieval we only record the training runs where SGDM has entered a good minima to eliminate the need to tune the parameters for our error rate procedure for different 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. If the diagnostic activates too late, it can waste unnecessary computation. Let $K = (n-k)/n$ such that $\|\theta_k - \theta_n\|^2 = \eta$ and $k \leq n$. If the diagnostic activates too late, then 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 2 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 Chee & Toulis (2018). 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 to no effect on the Type I and II error rates of the convergence diagnostic. Contrast Table 2 with the results in Table 1, 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. In approximately 10-20% of Type II errors the diagnostic activated late.

Algorithm 2 SGDM with automatic learning rate

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

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

6. Application: an automatic learning rate schedule

The convergence diagnostic has a natural application in automating the learning rate schedule. Hyper-parameter tuning, especially for the learning rate, has a major effect on the performance of stochastic gradient methods. Tuning is typically a very manual process requiring many training reruns. The benefit of an automatic learning rate is to greatly reduce the amount of supervision and number of training runs.

In Algorithm 2 we present an automatic learning rate based on the convergence diagnostic in Algorithm 1. SGDM with constant learning rate moves quickly towards $\theta_\star$ but cannot improve beyond distance $O(\gamma)$, as suggested by Theorem 2.1. 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. In practice it is common to use a constant learning rate and decrease several times at hand chosen points (He et al., 2016; Krizhevsky et al., 2012).

A major benefit of automatic hyper-parameter tuning is robustness to a potentially misspecified initial setting, in this case γ_0. We train logistic regression on benchmark data

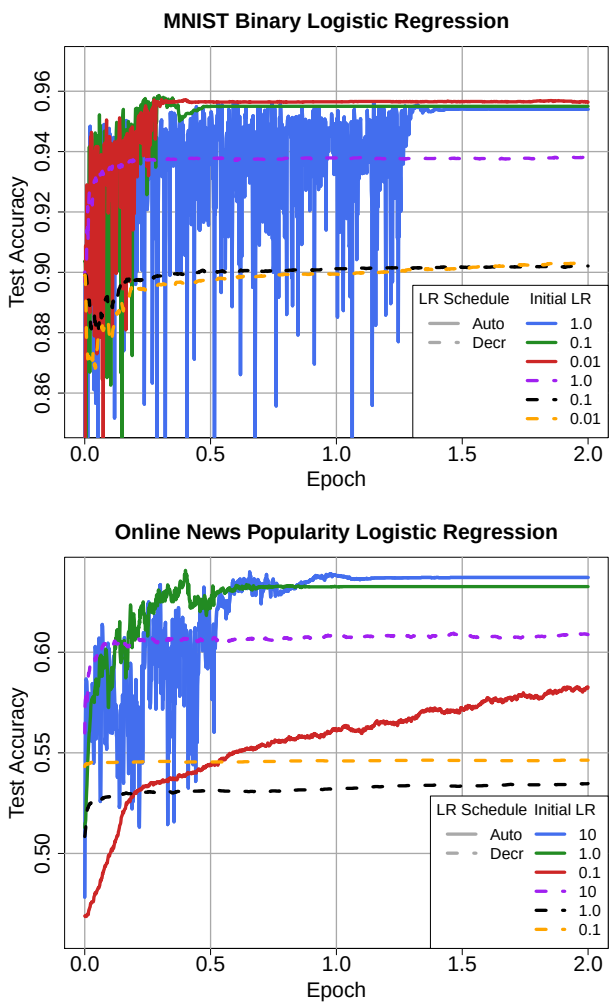

Figure 3. Binary logistic regression trained on MNIST and Online News Popularity data sets with SGDM using Algorithm 2 and decreasing rate $\gamma = \gamma_0/n$. $\beta = 0.8$.

sets MNIST[1] and Online News Popularity[2] for a variety of initial learning rates γ_0. In Figure 3 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 data sets. The automatic learning rate was significantly more robust to initial conditions than the decreasing rate. In addition, the automatic learning rate achieved significantly higher test accuracy than the decreasing rate. 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.

[1] http://yann.lecun.com/exdb/mnist/

[2] https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity

7. Conclusion

In this paper we focus on detecting the phase transition of SGDM (stochastic gradient descent with momentum) to the stationary phase. Inspiration is drawn from literature on stopping times in stochastic approximation. Momentum introduces challenges in the construction and operation of the test statistic for the convergence diagnostic. We present 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, which the diagnostic is able to capture. 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 convergence to the stationary phase. We present an application to an automatic learning rate and show that it is robust to initial conditions.

Future work should focus on development of test statistics with easy to implement thresholds which are robust to momentum, in addition to theoretical analysis of type I and II errors. Extensions to other stochastic gradient methods such as adaptive gradient methods or distributed methods such as k-step are also of interest. Finally, the application of the automatic learning rate to tune deep neural networks would be of great interest.

References

Bach, F. and Moulines, E. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In *Advances in Neural Information Processing Systems*, pp. 773–781, 2013.

Blum, A., Kalai, A., and Langford, J. Beating the hold-out: Bounds for k-fold and progressive cross-validation. In *Proceedings of the twelfth annual conference on Computational learning theory*, pp. 203–208. ACM, 1999.

Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pp. 177–186. Springer, 2010.

Bottou, L. Stochastic Gradient Descent Tricks. In *Neural Networks: Tricks of the Trade*, volume 1, pp. 421–436. 2012.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *arXiv preprint arXiv:1606.04838*, 2016.

Chee, J. and Toulis, P. Convergence diagnostics for stochastic gradient descent with constant learning rate. In *21st*

International Conference on Artificial Intelligence and Statistics, 2018.

Chen, Y., Chi, Y., Fan, J., and Ma, C. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1-2):5–37, 2018.

Delyon, B. and Juditsky, A. Accelerated stochastic approximation. *SIAM J. Optimization*, 3(4):868–881, 1993.

Ermoliev, Y. M. and Wets, R.-B. *Numerical techniques for stochastic optimization*. Springer-Verlag, 1988.

Ge, R., Kakade, S. M., Kidambi, R., and Netrapalli, P. The step decay schedule: A near optimal, geometrically decaying learning rate procedure. In *Advances in Neural Information Processing Systems*, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, 2016.

Keskar, N. S. and Socher, R. Improving generalization performance by switching from adam to sgd. *arXiv preprint arXiv:1712.07628*, 2017.

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

Kingma, D. P. and Ba, J. L. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

Lang, H., Zhang, P., and Xiao, L. Using statistics to automate stochastic optimization. In *Advances in Neural Information Processing Systems*, 2019.

Moulines, E. and Bach, F. R. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pp. 451–459, 2011.

Murata, N. A statistical study of on-line learning. *Online Learning and Neural Networks. Cambridge University Press, Cambridge, UK*, pp. 63–92, 1998.

Needell, D., Ward, R., and Srebro, N. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In *Advances in Neural Information Processing Systems*, pp. 1017–1025, 2014.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

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

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

Reddi J. Sashank, Hefny Ahmed, S. S. P. B. S. A. On variance reduction in stochastic gradient descent and its asynchronous variants. In *Advances in Neural Information Processing Systems*, 2015.

Roux, N. L., Schmidt, M., and Bach, F. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems*, pp. 2663–2671, 2012.

Sordello, M. and Su, W. Data-adaptive learning rate selection for stochastic gradient descent using convergence diagnostic. Joint Statistics Meeting, 2019. URL https://ww2.amstat.org/meetings/jsm/2019/onlineprogram/AbstractDetails.cfm?abstractid=300067.

Su, W. J. and Zhu, Y. Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent. *arXiv preprint arXiv:1802.04876*, 2018.

Toulis, P. and Airoldi, E. M. Scalable estimation strategies based on stochastic approximations: classical results and new insights. *Statistics and computing*, 25(4):781–795, 2015.

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

Yaida, S. Fluctuation-dissipation relations for stochastic gradient descent. In *International Conference on Learning Representations*, 2019.

Yang, T., Lin, Q., and Li, Z. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. In *International Joint Conferences on Artificial Intelligence*, 2018.

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

Zhang, T. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 116. ACM, 2004.