Convergence diagnostics for stochastic gradient descent with constant learning rate

Jerry Chee
University of Chicago

Panos Toulis
University of Chicago, Booth School of Business

Abstract

Many iterative procedures in stochastic optimization exhibit a transient phase followed by a stationary phase. During the transient phase the procedure converges towards a region of interest, and during the stationary phase the procedure oscillates in that region, commonly around a single point. In this paper, we develop a statistical diagnostic test to detect such phase transition in the context of stochastic gradient descent with constant learning rate. We present theory and experiments suggesting that the region where the proposed diagnostic is activated coincides with the convergence region. For a class of loss functions, we derive a closed-form solution describing such region. Finally, we suggest an application to speed up convergence of stochastic gradient descent by halving the learning rate each time stationarity is detected. This leads to a new variant of stochastic gradient descent, which in many settings is comparable to state-of-art.

1 Introduction

We consider a classical problem in stochastic optimization stated as

$$\theta_\star = \arg\min_{\theta\in\Theta} \mathbb{E}(\ell(y, x^\top\theta)), \tag{1}$$

where ℓ is the loss function, $y \in \mathbb{R}$ denotes the response, $x \in \mathbb{R}^p$ are the features, and θ are parameters in $\Theta \subseteq \mathbb{R}^p$. For example, the quadratic loss function is defined as $\ell(y, x^\top\theta) = (1/2)(y - x^\top\theta)^2$. In estimation problems we typically have a finite data set $\{(x_i, y_i)\}$, $i = 1, 2, \ldots, N$, from which we wish to estimate $\theta_\star$ by solving the empirical version of Equation (1):

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} \ell(y_i, x_i^\top\theta).$$

When data size, N, and parameter size, p, are large classical methods for computing $\hat{\theta}$ fail. Stochastic gradient descent (SGD) is a powerful alternative [5, 6, 24, 32] because it solves the problem in an iterative fashion through the procedure:

$$\theta_n = \theta_{n-1} - \gamma\nabla\ell(y_n, x_n^\top\theta_{n-1}). \tag{2}$$

Here, θ_{n-1} is the estimate of $\theta_\star$ prior to the nth iteration, (x_n, y_n) is a random sample from the data, and $\nabla\ell$ is the gradient of the loss with respect to θ. Classical stochastic approximation theory [1, 4, 21] suggests that SGD converges to a value θ_∞ such that $\mathbb{E}(\nabla\ell(y, x^\top\theta_\infty)) = 0$, which under typical regularity conditions is equal to $\theta_\star$ when N is infinite (streaming setting), or is equal to $\hat{\theta}$ when N is finite. Going forward we assume the streaming setting for simplicity, but our results hold for finite N as well.

Typically, stochastic iterative procedures start from some starting point and then move through a transient phase and towards a stationary phase [15]. In stochastic gradient descent this behavior is largely governed by parameter $\gamma > 0$, which is known as the learning rate, and can either be decreasing over n (e.g., $\propto 1/n$), or constant. In the decreasing rate case, the transient phase is usually long, and can be impractically so if the rate is slightly misspecified [17, 25], whereas the stationary phase involves SGD converging in quadratic mean to $\theta_\star$. When γ is constant the transient phase is much shorter and less sensitive to the learning rate, whereas the stationary phase involves SGD oscillating within a region that contains $\theta_\star$. In this paper, we focus on statistical convergence diagnostics for constant rate SGD because in this setting a convergence diagnostic can be utilized to identify when there is no benefit in running the procedure longer.

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

1.1 Related work and contributions

The idea that SGD methods are composed of a transient phase and a stationary phase (also known as search phase and convergence phase, respectively), has been expressed before [15]. However, no principled statistical methods have been developed to address stationarity issues, and thereby guide empirical practice of SGD. Currently, guidance is based on heuristics originating from optimization theory that aim to evaluate the magnitude of SGD updates. For example, a popular method is to stop when $||\theta_n - \theta_{n-1}||$ is small according to some threshold, or when updates of the loss function have reached machine precision [7, 9]. These methods, however, do not take into account the sampling variation in SGD estimates, and are suited for deterministic procedures but not stochastic ones.

A more statistically motivated approach is to monitor the test error of SGD iterates on a hold-out validation test, concurrently with the main SGD iteration [3, 6]. One idea here is to stop the procedure when the validation error starts increasing. An important problem with this approach is that the validation error is also a stochastic process, and estimating when it actually starts increasing presents similar, if not greater, challenges to the original problem of detecting convergence to stationary phase. Furthermore, cross validation can be computationally costly in large data sets.

In stochastic approximation, methods to detect stationarity can be traced back to classical theory of stopping times [20, 31]. One important method, which forms the basis of this paper, is Pflug's procedure [20] that keeps a running average of the inner product of successive gradients $\nabla_{n-1}\ell^\top \nabla_n \ell$, where we defined $\nabla_j \ell = \nabla \ell(y_j, x_j^\top \theta_{j-1})$. The underlying idea is that in the transient phase the stochastic gradients point roughly to the same direction, and thus their inner product is positive. In the stationary phase, SGD with constant rate moves haphazardly around $\theta_\star$, and so the gradients point to different directions making the inner product negative.

The intuition that a negative inner product of successive gradients indicates convergence underlies accelerated methods in stochastic approximation [8, 11, 23]. The accelerated methods, however, take this intuition as a given, whereas we develop theory for it to define a formal convergence testing procedure. Recently, another related idea is that of gradient diversity [30], which is used to understand why speedup gains in batch SGD saturate with increasing batch size. An important difference is that gradient diversity calculates the inner products at a fixed parameter value θ, whereas stochastic approximation methods, including this paper, use successive parameter values.

1.2 Overview of results and contributions

Our contributions in this paper can be summarized as follows. In Section 2, we develop a formal convergence diagnostic test for SGD, which combines Pflug's stopping time procedure [20] with SGD in Equation (2) to detect when SGD exits the transient phase and enters the stationary phase. We note that by convergence of SGD with constant rate we do not mean convergence to a single point but convergence to the stationarity region. We prove a general result that the diagnostic indeed is activated almost surely. We illustrate through an example, where conditional on the diagnostic being activated, the distance $||\theta_n - \theta_\star||$ is uncorrelated with the starting distance $||\theta_0 - \theta_\star||$, implying that the diagnostic captures the transition from transient to stationary phase. In Section 3, we develop theory for quadratic loss, and derive a closed-form solution describing the region where the diagnostic is activated. In Section 4.1, we present extensions beyond the quadratic loss. In Section 4.2 we suggest an application of the diagnostic in speeding up SGD by halving the learning rate each time convergence is detected. This leads to a new SGD procedure, named $\texttt{SGD}^{1/2}$, which is comparable to state-of-art procedures, such as variance-reduced SGD [10, SVRG] and averaged SGD [5, 29], in Sections 4.3 and 4.4.

2 Convergence diagnostic

Before we develop the formal diagnostic, we present theory that supports the existence of a transient and stationary phase of SGD. The theory suggests that the mean squared error of SGD has a bias term from distance to the starting point, and a variance term from noise in stochastic gradients.

Theorem 1 ([14, 16]) *Under certain assumptions on the loss function, there are positive constants A_γ, B such that, for every n, it holds that*

$$\mathbb{E}(||\theta_n - \theta_\star||^2) \leq \mathbb{E}(||\theta_0 - \theta_\star||^2)e^{-A_\gamma n} + B\gamma.$$

Remarks. The constants A_γ, B differ depending on the analysis. For example, Bach and Moulines [14] use $A_\gamma \approx \gamma\mu/4 - \gamma^2 L^2$, where μ is the strong convexity constant of expected loss $f(\theta) = \mathbb{E}(\ell(y, x^\top\theta)|\theta)$, and L is the Lipschitz constant of $\nabla \log \ell(y, x^\top\theta)$; and $B = \sigma^2/\mu$, where σ^2 is an upper bound for the variance of $||\nabla \log \ell(y, x^\top\theta_\star)||^2$. Needell and Srebro [16] use $A_\gamma \approx 2\gamma\mu - 2\gamma^2\mu L$ and $B = \sigma^2/\mu(1 - \gamma L)$.

Despite such differences, all analyses suggest that the SGD procedure with constant rate goes through a transient phase exponentially fast during which it forgets the initial conditions $\mathbb{E}(||\theta_0 - \theta_\star||^2)$, and then

enters a stationary phase during which it oscillates around $\theta_\star$, roughly at a region of radius $R_\gamma = O(\sqrt{\gamma})$. A trade-off exists here: reducing γ will make the oscillation radius, R_γ, smaller but escaping the transient phase becomes much slower; for instance, in the extreme case where $\gamma = 0$ the procedure will never exit the transient phase.

Despite the theoretical insights it offers, Theorem 1 has limited practical utility for estimating the phase transition in SGD. One approach could be to find the value of n for which $\mathbb{E}(||\theta_0 - \theta_\star||^2)e^{-A_\gamma n} = 0.01B\gamma$, that is, the initial conditions have been discounted to 1% of the stationary variance. That, however, requires estimating μ, L, σ^2, and $\mathbb{E}(||\theta_0 - \theta_\star||^2)$, which is challenging. In the following section, we develop a concrete statistical diagnostic to estimate the phase transition and detect convergence of SGD in a much simpler way.

2.1 `Pflug` diagnostic

In this section, we develop a convergence diagnostic for SGD procedures that relies on Pflug's procedure [19] in stochastic approximation. The diagnostic is presented as Algorithm 1 and concrete instances under quadratic loss along with theoretical analysis are presented in Section 3, with extensions in Section 4.

The diagnostic is defined by a random variable S_n that keeps the running sum of the inner product of successive stochastic gradients, as shown in Line 7. The idea is that in the transient phase SGD moves towards $\theta_\star$ by discarding initial conditions, and so the stochastic gradients point to the same direction, on average. This implies that the inner product of successive stochastic gradients is likely positive in the transient phase. In the stationary phase, however, SGD is oscillating around $\theta_\star$ at a distance bounded by Theorem 1, and so the gradients point to different directions. This implies a negative inner product on average during the stationary phase. When the statistic S_n changes sign from positive to negative, this is a good signal that the procedure has exited the transient phase.

Since our convergence diagnostic is iterative we need to show that it eventually terminates with an answer. In Theorem 2 that follows we prove that $\mathbb{E}(S_n - S_{n-1}) < 0$ as $n \to \infty$, and so Algorithm 1 indeed terminates almost surely. For brevity, we state the theorem without technical details. The full assumptions and proof can be found in the supplementary material.

Theorem 2 *Under certain assumptions, the convergence diagnostic in Algorithm 1 for constant rate SGD procedure in Equation* (2) *satisfies* $\mathbb{E}(S_n - S_{n-1}) < 0$ *as* $n \to \infty$*, and so the algorithm terminates almost surely.*

Algorithm 1 `Pflug` diagnostic for convergence of stochastic gradient descent.

Input: starting point θ_0, data $\{(y_1, x_1), (y_2, x_2), \ldots\}$, $\gamma > 0$, `burnin > 0`.

Output: Iteration when SGD in Equation (2) is estimated to enter stationary phase.

1: $S_0 \leftarrow 0$
2: $\theta_1 \leftarrow \theta_0 - \gamma \nabla \ell(y_1, x_1^\top \theta_0)$
3: **for all** $n \in \{2, 3, \cdots\}$ **do**
4: Sample (x_n, y_n)
5: Define $\nabla \ell_n = \nabla \ell(y_n, x_n^\top \theta_{n-1})$.
6: $\theta_n \leftarrow \theta_{n-1} - \gamma \nabla \ell_n$.
7: $S_n \leftarrow S_{n-1} + \nabla \ell_n^\top \nabla \ell_{n-1}$.
8: **if** $n >$ `burnin` and $S_n < 0$ **then**
9: **return** n
10: **end if**
11: **end for**

Remarks. Theorem 2 shows that the inner product of successive gradients is negative in expectation as the iteration number increases. Roughly speaking, when θ_n is very close to $\theta_\star$ the dominant force is the variance in the stochastic gradient pulling the next iterates away from $\theta_\star$; when θ_n is far from $\theta_\star$ the dominant force is the bias in the stochastic gradient, which instead pulls the next iterates towards $\theta_\star$. This implies that the running sum of successive gradients will eventually become negative at a finite iteration number, and so by the law of large numbers the diagnostic returns a value almost surely.

3 Quadratic loss model

In this section, we attempt to gain analytical insight into our convergence diagnostic of Algorithm 1 by assuming simple quadratic loss function, i.e., $\ell(y, x^\top \theta) = (1/2)(y - x^\top \theta)^2$ and $\nabla \ell(y, x^\top \theta) = -(y - x^\top \theta)x$. Consider the case where $\theta_0 = \theta_\star$, i.e., the procedure starts in the stationary region. Let $y_n = x_n^\top \theta_\star + \varepsilon_n$, where ε_n are zero-mean random variables, $\mathbb{E}(\varepsilon_n | x_n) = 0$. Then,

$$\theta_1 = \theta_\star + \gamma(y_1 - x_1^\top \theta_\star)x_1 = \theta_\star + \gamma \varepsilon_1 x_1,$$

from which it follows that

$$\begin{aligned} S_2 - S_1 &= (y_2 - x_2^\top \theta_1)(y_1 - x_1^\top \theta_0) x_2^\top x_1 \\ &= (\varepsilon_2 - \gamma \varepsilon_1 x_2^\top x_1)\varepsilon_1 x_2^\top x_1. \\ \mathbb{E}(S_2 - S_1) &= -\gamma \mathbb{E}(\varepsilon_1^2)\mathbb{E}((x_2^\top x_1)^2) < 0. \end{aligned} \tag{3}$$

Thus, when the procedure starts at true parameter value, $\theta_\star$, the diagnostic is decreased in expectation,

and eventually at some iteration τ the statistic S_τ becomes negative and the diagnostic is activated at τ. We generalize this result in the following theorem.

Theorem 3 *Suppose that the loss is quadratic, $\ell(y, x^\top\theta) = (1/2)(y - x^\top\theta)^2$. Let x_1 and x_2 be two iid vectors from the distribution of x, and define: $\sigma^2 = \mathbb{E}((y - x^\top\theta_\star)^2)$; $c^2 = \mathbb{E}((x_1^\top x_2)^2)$; $C = \mathbb{E}(x_1 x_2^\top (x_1^\top x_2))$; $D = \mathbb{E}(x_1 x_1^\top (x_1^\top x_2)^2)$, and suppose that all such constants are finite. Then, for $\gamma > 0$,*

$$\begin{aligned}\Delta_n(\theta) &= \mathbb{E}(S_{n+2} - S_{n+1}|\theta_n = \theta) \\ &= (\theta - \theta_\star)^\top (C - \gamma D)(\theta - \theta_\star) - \gamma c^2\sigma^2.\end{aligned}$$

Remarks. Theorem 3 shows that the boundary surface that separates the two regions where the test statistic S_n increases or decreases in expectation looks like an ellipse, for large enough γ. Regardless of the choice of γ, when θ_n is close enough to $\theta_\star$, the diagnostic is guaranteed to decrease in expectation since the only remaining term is $-\gamma c^2\sigma^2 < 0$.

The result also shows the various competing forces between bias and variance in the stochastic gradients as they relate to how the diagnostic behaves. For instance, when θ_n is very close $\theta_\star$, larger σ^2 (noise in stochastic gradient) contributes to a faster decrease of the diagnostic in expectation, but at the cost of higher variance. The contribution of the other term, c^2, is less clear. For instance, c is large when there is strong collinearity in features x, which may contribute to decreasing S_n. But strong collinearity also implies that C is almost positive definite which contributes positive values to S_n, thus counteracting the contribution of c. Note that D is a positive definite matrix but C may not be. This implies that careful selection of γ may be necessary for the diagnostic to work well. For example, when γ is very small and C is positive definite, then S_n will converge to a negative number slowly. One way to alleviate this sensitivity to the learning rate is through implicit updates [26], which we explore in the following section.

3.1 Implicit update

As mentioned above the `Pflug` diagnostic is sensitive to the choice of learning rate γ. When γ is small and C is positive definite, S_n will be mostly increasing during the transient phase, which makes convergence slower. But choosing a large learning rate can easily lead to numerical instability. One way to alleviate such sensitivity to the learning rate is to use the SGD procedure with an implicit update as follows:

$$\theta_n = \theta_{n-1} - \gamma\nabla\ell(y_n, x_n^\top\theta_n). \tag{4}$$

Note that θ_n appears on both sides of the equation. In the quadratic loss model we can solve exactly the implicit equation as follows:

$$\theta_n = (I + \gamma x_n x_n^\top)^{-1}(\theta_{n-1} + \gamma y_n x_n). \tag{5}$$

Implementing the procedure in Eq. (5) is fast since it is equivalent to $\theta_n = (\theta_{n-1} + \gamma y_n x_n)/(1 + \gamma||x_n||^2)$. More generally, the implicit update in Equation 4 can be computed efficiently in many settings through a one-dimensional root-finding procedure [26]. Previous work on implicit SGD (ISGD) has shown that implicit procedures have similar asymptotic properties with standard SGD procedures with numerical stability as an added benefit. Since most related work on ISGD methods is with respect to decreasing learning rate procedures [2, 12, 25, 26], we provide an analysis for constant rate ISGD as in Equation (4) in the supplementary material. We note that ISGD procedures are related to proximal updates in stochastic optimization [18, 22, 28], but these methods differ from ISGD methods in that they employ a combination of classical SGD with deterministic proximal operators, whereas ISGD's proximal operator is purely stochastic.

The following theorem shows that the implicit update in the linear model mitigates the sensitivity of the `Pflug` diagnostic to the choice of the learning rate.

Theorem 4 *Let $\lambda_\gamma = \mathbb{E}(1/(1 + \gamma||x||^2)) \in (0, 1]$. Under the assumptions of Theorem 3 applied on the implicit procedure in Equation* (4)*, it holds that*

$$\begin{aligned}\Delta_n^{\text{im}}(\theta) &= \mathbb{E}(S_{n+2} - S_{n+1}|\theta_n = \theta) \\ &= a_\gamma\Delta_n(\theta) + b_\gamma\left[(\theta - \theta_\star)^\top D(\theta - \theta_\star) + \sigma^2 c^2\right],\end{aligned}$$

where $a_\gamma = \lambda_\gamma^2$, $b_\gamma = \gamma\lambda_\gamma^2(1 - \lambda_\gamma)$.

Remarks. Theorem 4 shows that the diagnostic is more stable with the ISGD procedure than with the classical SGD procedure. By stability we mean two things. First, even when classical SGD diverges the convergence diagnostic may still declare convergence. Consider, for example, the simple model $\theta_n = \theta_{n-1} + \gamma(y_n - \theta_{n-1})$, where $y \sim N(\theta_\star, 1)$. If $\gamma > 1$ the classical SGD procedure will diverge. However, the diagnostic will declare convergence almost immediately because by Theorem 3 it decreases, in expectation, for every θ. Such inconsistency due to instability of classical SGD cannot happen with implicit SGD.

Second, generally speaking, empirical performance of the diagnostic under implicit SGD matches theory better than under classical SGD. This is illustrated in the following section, where the region of diagnostic convergence is smooth and elliptical under implicit SGD, as predicted by Theorem 4; under classical SGD, the corresponding region does not follow Theorem 3 as closely due to sensitivity to learning rate specification.

3.2 Illustration

Here, we illustrate the main results of Theorem 4 through Figure 1, which can be described as follows. The shaded areas in the figure show how the `Pflug` diagnostic changes in expectation when the SGD iterate falls in the region. In other words, every point θ in the figure is shaded by the value $\Delta_n^{\text{im}}(\theta)$, as defined in Theorem 4.

Various shades of grey indicate the magnitude of change. The darkest-shaded region corresponds to the region where the diagnostic decreases in expectation, that is, $\Delta_n^{\text{im}}(\theta) \leq 0$ for all θ in that region. We call this the `Pflug` region. Note that the `Pflug` region is centered roughly around $\theta_\star$, the true parameter value. Inside the `Pflug` region the diagnostic is decreased in expectation, and outside of the region it is increased. Furthermore, the expected change in the diagnostic is uniform in distance to the center of the `Pflug` region, which is roughly $\theta_\star$: the farther we move away from the center $\theta_\star$ the larger the expected increase of the diagnostic becomes.

The blue polygon shaded with diagonal lines corresponds to empirical estimations of the convergence region of SGD, defined as the region where SGD iterates have oscillated around for 95% of the time calculated over 1000 simulations. The polygon shows that the `Pflug` region approximates very well the actual convergence region of SGD. This is remarkable because the `Pflug` region can be calculated from data using the convergence diagnostic, whereas by Theorem 1 the SGD convergence region cannot be calculated without knowledge of $\theta_\star$ and other unknown parameters.

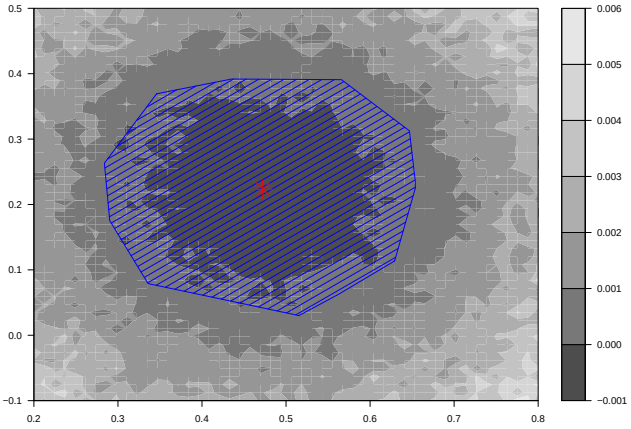

Figure 1: Shaded area in the center: region where `Pflug` diagnostic is decreased in expectation. Polygon around shaded area: convergence region of SGD where iterates oscillate around (empirically calculated). Color legend on the right: values of expected increase (or decrease) of the diagnostic.

3.3 Simulated example

Next, we test the `Pflug` diagnostic through a simulated experiment. The experimental setup is as follows. We set $p = 20$ as the parameter dimension, and also set $N = 5000$ as the data set size and fix $\theta_\star \in \mathbb{R}^p$ with $\theta_{\star,j} = 10e^{-0.75j}$; this ensures some variation and sparsity in the parameter values. We sample features as $x_i \sim \mathcal{N}_p(0, I)$, where $\mathcal{N}_p$ denotes a p-variate normal distribution, I is the identity matrix, and $i = 1, 2, \ldots N$. We sample outcomes as $y_i = x_i^\top \theta_\star + \mathcal{N}(0, \sigma^2)$, where $\sigma = 3$.

For given γ we run Algorithm 1 with `burnin` $= 0.1N$, for various values of the starting point θ_0 sampled as $\mathcal{N}_p(\theta_\star, \sigma_0^2 I)$, where $\sigma_0 = 2$. Let $E_n = ||\theta_n - \theta_\star||^2$, then for each run we store the tuple

$$(\gamma, \tau, E_0, E_{\tau/2}, E_{2\tau}),$$

where τ is the output of Algorithm 1, i.e., the iteration at which the `Pflug` diagnostic detected convergence. The idea in this experimental evaluation is that if the convergence diagnostic detects convergence accurately, iterates earlier than convergence, say, $\theta_{\tau/2}$, will depend on the initial conditions θ_0 more than iterates later than convergence, say, $\theta_{2\tau}$. Thus, for given γ and τ, we should expect a much higher correlation between $E_{\tau/2}$ and E_0 than between $E_{2\tau}$ and E_0. To test this hypothesis, for a given value of γ we draw 100 independent samples of $(E_0, E_{\tau/2}, E_{2\tau})$. With these samples we regress $E_{\tau/2}$ on E_0 and $E_{2\tau}$ on E_0 in two normal linear regression models. Table 1 summarizes the regression results from this experiment. In the second and third column of Table 1 we report the regression coefficients of E_0 in the two model fits, respectively, and also report statistical significance.

From the table, we see that the regression coefficient corresponding to $E_{\tau/2}$ is always positive and statistically significant, whereas the coefficient is mostly not significant for $E_{2\tau}$. This suggests that $E_{\tau/2}$ depends on initial conditions E_0, and thus stationarity has not yet been reached at iteration $\tau/2$. In contrast, $E_{2\tau}$ does not depend on initial conditions E_0, and thus stationarity has likely occurred after iteration τ. This is evidence indicating that the `Pflug` diagnostic performs reasonably well in estimating the switch of SGD from its transient phase to its stationary phase.

We note that in the regression evaluation we had to control for τ (by using it as a regressor) because the iteration number is correlated with mean-squared error (larger values for τ are correlated with smaller error).

Table 1: Experimental evaluation of convergence diagnostic over 100 runs per learning rate value. Significance levels: *** = < 0.1%; ** = 1%; * = 5%; . = 10%

γ	$E_{\tau/2} = \beta_{\tau/2}E_0 + \varepsilon$ $\beta_{\tau/2}$	$E_{2\tau} = \beta_{2\tau}E_0 + \varepsilon$ $\beta_{2\tau}$
0.02	0.17 **	0.01 .
0.05	0.20 ***	−0.008
0.1	0.09 **	−0.0007
0.2	0.06 **	0.005
0.5	0.09 ***	−0.008
1.0	0.06 *	0.02 *
2.0	0.06 **	0.009
5.0	0.07 **	−0.012

4 Extensions and applications

In this section we consider extensions of the `Pflug` diagnostic to a more broad family of loss functions inspired by generalized linear models (GLMs). We also consider an application of the diagnostic to speed up convergence of SGD with constant learning rate.

4.1 Generalized linear loss

Here, we consider the loss based on the GLM formulation [13, 26] where $\ell(y, x^\top\theta) = -y \cdot x^\top\theta + f(x^\top\theta)$. For example, the quadratic loss is equivalent to $f(u) = u^2/2$. The logistic loss is when y is binary and $f(u) = \log(1 + e^u)$. In general, f cannot be chosen arbitrarily—one standard choice is to define f such that $e^{-\ell(y,x^\top\theta)}$ is a proper density, i.e., it integrates to one. The following theorem generalizes the results in Section 3 on the quadratic loss.

Theorem 5 *Consider the GLM loss defined as* $\ell(y, x^\top\theta) = -y \cdot x^\top\theta + f(x^\top\theta)$. *Let* $h(u) = f'(u)$ *and suppose that* $h'(x^\top\theta) \geq k > 0$, *almost surely for all* θ. *Let* x_1, x_2 *be two iid vectors from the distribution of* x. *Define* $\sigma^2 = \mathbb{E}((y - h(x^\top\theta_\star)^2)$; $c^2 = \mathbb{E}((x_1^\top x_2)^2)$; $C(\theta, \theta_\star) = \mathbb{E}([h(x_1^\top\theta) - h(x_1^\top\theta_\star)]x_1)$; $D^2(\theta, \theta_\star) = \mathbb{E}([h(x_1^\top\theta) - h(x_1^\top\theta_\star)]^2(x_1^\top x_2)^2)$. *Then, for small enough* γ,

$$\begin{aligned}\Delta_n^{glm}(\theta) &= \mathbb{E}(S_{n+2} - S_{n+1}|\theta_n = \theta) \\ &\leq ||C(\theta, \theta_\star)||^2 - \gamma k[\sigma^2 c^2 + D^2(\theta, \theta_\star)].\end{aligned}$$

Remarks. The result in Theorem 5 has the same structure as in Theorem 3 so a direct analogy can be helpful. The terms σ^2, c^2 in the two theorems are identical, if we consider that for the quadratic loss it holds that $h(u) = u$. The term $||C(\theta, \theta_\star)||^2$ in Theorem 5 corresponds to the term $(\theta - \theta_\star)^\top C(\theta - \theta_\star)$ in Theorem 3, and $D^2(\theta, \theta_\star)$ corresponds to $(\theta - \theta_\star)^\top D(\theta - \theta_\star)$. The terms are equal when we set $h(u) = u$, in which case $k = 1$. Thus, the diagnostic with the more general GLM loss has familiar properties. For example, when $\theta \approx \theta_\star$, i.e., when SGD is near the truth, $||C(\theta, \theta_\star)||^2 \approx 0$ and $D^2(\theta, \theta_\star) \approx 0$, in which case the negative constant term dominates, and the test statistic decreases in expectation leading to activation of the diagnostic. One difference with the quadratic loss, however, is that as we move farther from $\theta_\star$ the statistic may change in a nonlinear way. Therefore the boundary separating the positive and negative regions of the diagnostic will generally not have the familiar smooth elliptical shape as in the quadratic loss (see Figure 1). This may lead to more complex behavior for the diagnostic, which is open to future work.

Regarding the assumptions of Theorem 5, we note that the constraint on derivative h' is not particularly strict because in the GLM formulation h' is guaranteed to be positive. The assumption is made to simplify the analysis, but can be improved by analyzing the quantity $h'(x^\top\theta_n)$ through existing analyses of θ_n.

4.2 `SGD`$^{1/2}$ for fast convergence

We now switch gears from analyzing the behavior of the `Pflug` diagnostic to using it in a practical application. Our suggested application is to use the diagnostic within a SGD loop where the learning rate is halved and the procedure restarted each time convergence is detected. We emphasize that our goal here is to illustrate the utility of our convergence diagnostic and not to exhaustively demonstrate the performance of the new procedure. A full analysis of the proposed procedure is open to future work.

More specifically, the SGD procedure with constant rate has linear convergence to a stationary distance from $\theta_\star$ of $R_\gamma = O(\sqrt{\gamma})$, as suggested by Theorem 1. It would therefore be beneficial to reduce the learning rate when we know that SGD iterates are oscillating around $\theta_\star$ in a ball of radius R_γ, so that the procedure moves to a ball with a smaller radius. To implement such a procedure, however, would require knowing $\theta_\star$, and also knowing all parameters required to calculate R_γ. Our solution employs the `Pflug` convergence diagnostic to detect stationarity. Algorithm 2 describes such a procedure, named `SGD`$^{1/2}$, where the learning rate is halved upon detection of convergence (Line 10).

Note that implicit updates can be used in this algorithm as well; we call this modified algorithm `ISGD`$^{1/2}$. In our experiments in the following section, we employ `ISGD`$^{1/2}$ because of the benefits in numerical stability from using implicit updates, as described earlier.

Algorithm 2 Procedure $\texttt{SGD}^{1/2}$.
Input: θ_0, data $\{(y_1, x_1), (y_2, x_2), \ldots\}$, $\gamma > 0$, `burnin, maxit > 0`.
Output: Iteration $\tau > 0$, when SGD is estimated to have converged.

$s \leftarrow 0$
$\tau \leftarrow 0$
$\theta_1 \leftarrow \theta_0 - \gamma \nabla \ell(y_1, x_1^\top \theta_0)$
for all $n \in \{2, 3, \cdots\}$ **do**
  $\theta_n \leftarrow \theta_{n-1} - \gamma \nabla \ell(y_n, x_n^\top \theta_{n-1})$
  $s \leftarrow s + (\theta_n - \theta_{n-1})^\top (\theta_{n-1} - \theta_{n-2})/\gamma^2$
  if $n > \tau +$ `burnin` and $s < 0$ **then**
    $\tau \leftarrow n$
    $s \leftarrow 0$
    $\gamma \leftarrow \gamma/2$
    if $\gamma <$ `1e-10` **and** $n >$ `maxit` **then**
      return θ_n.
    end if
  end if
end for

4.3 Simulated data experiments

To evaluate the effectiveness of $\texttt{ISGD}^{1/2}$, we compare to other classical and state-of-the-art SGD methods. We first experiment on simulated data to better understand the performance of $\texttt{ISGD}^{1/2}$ and its competition under various parameter settings. In particular, we compare the performance of procedure $\texttt{ISGD}^{1/2}$ in Algorithm 2 against SVRG and classical ISGD on simulated data. The classical ISGD uses a learning rate of $O(1/n)$, which is optimized through pre-processing. The basic experimental setup is as follows.

We consider settings of high and low signal to noise ration (SNR), and high and low dimension and test under the four combinations of these settings. For the high SNR case, we set SNR = 5, where SNR = $\text{trace}(Var(x))/pVar(y|x)$, and for the low SNR case we set SNR = 2. For the high dimension case we set $p = 150$ as the parameter dimension, and for the low dimension case we set $p = 10$. Given p, we fix $\theta_* \in \mathbb{R}^p$ such that $\theta_{\star,j} = 10e^{-0.75j}$. We set $N = 5000$ as the size of the data set. We sample features as $x_i \sim \mathcal{N}_p(0, I)$, where $i = 1, 2, \ldots N$. We sample outcomes as $y_i \sim \mathcal{N}(x_i^\top \theta_\star, \sigma^2)$ for the normal model, and $y_i \sim \text{Binom}(\exp(x_i^\top \theta_\star)/(1+\exp(x_i^\top \theta_\star))$ for the logistic model, where Binom(q) denotes the binomial random variable with mean q. The learning parameters for each SGD method were tuned to provide best performance through pre-processing.

From simulations with the normal model in the left half of Figure 2 we see that $\texttt{ISGD}^{1/2}$ attains a comparable performance to SVRG. In general, SVRG attains an overall better performance for these experiments, which we believe is related to our convergence diagnostic being aggressive in a couple of cases, which are essentially cases of Type-I error.

From simulations with the logistic model in the right half of Figure 2 we see that $\texttt{ISGD}^{1/2}$ attains an even better performance than before as there are fewer cases of Type-I error. With high SNR and low dimension parameter settings, $\texttt{ISGD}^{1/2}$ achieves consistently better performance than SVRG. We note that such comparisons do not take into account the sensitivity of SVRG to misspecifications of the learning rate (large enough learning rates can easily make the procedure diverge); or that SVRG requires periodic calculations over the entire data set, which here is easy because we are using only 5,000 data points, but may be a problem in more realistic settings. We also note that there are several improvements available for $\texttt{ISGD}^{1/2}$ by allowing a larger `burnin` period or by discounting the learning rate less aggressively. An interesting direction for future work is to understand the performance of our diagnostic test in terms of statistical validity and power, and thus address some of the aforementioned tuning issues in a principled manner.

4.4 Benchmark data sets

In addition to simulated experiments we conduct experiments on benchmark data sets MNIST (binary) and COVERTYPE (binary) to evaluate real world performance.[1] In particular, we perform binary logistic regression using $\texttt{ISGD}^{1/2}$, SVRG, classical ISGD, and averaged ISGD [27]. We plot the prediction error on a held-out test set in Figure 3 relative to the number of passes over the data.

Overall, we see that $\texttt{ISGD}^{1/2}$ convergences very quickly, after going over less than a quarter of the data, and achieves best performance in the COVERTYPE data set. We currently do not have a theoretical justification for this, but we have verified that the aforementioned result is consistently observed across multiple experiments. $\texttt{ISGD}^{1/2}$ was also very stable to specifications of the learning rate parameter, as expected from the analysis of Theorem 4. In contrast, even though SVRG performed comparably to $\texttt{ISGD}^{1/2}$, its performance was unstable, especially in the COVERTYPE data set, and required careful fine tuning of the learning rate through trial and error. Averaged SGD performed well on the MNIST data set, but flattened out very fast in the COVERTYPE data, possibly due to non-strong convexity of the objective function.

[1] Data sets can be found at `https://archive.ics.uci.edu/ml/databases/mnist/` and `https://archive.ics.uci.edu/ml/datasets/covertype`, respectively.

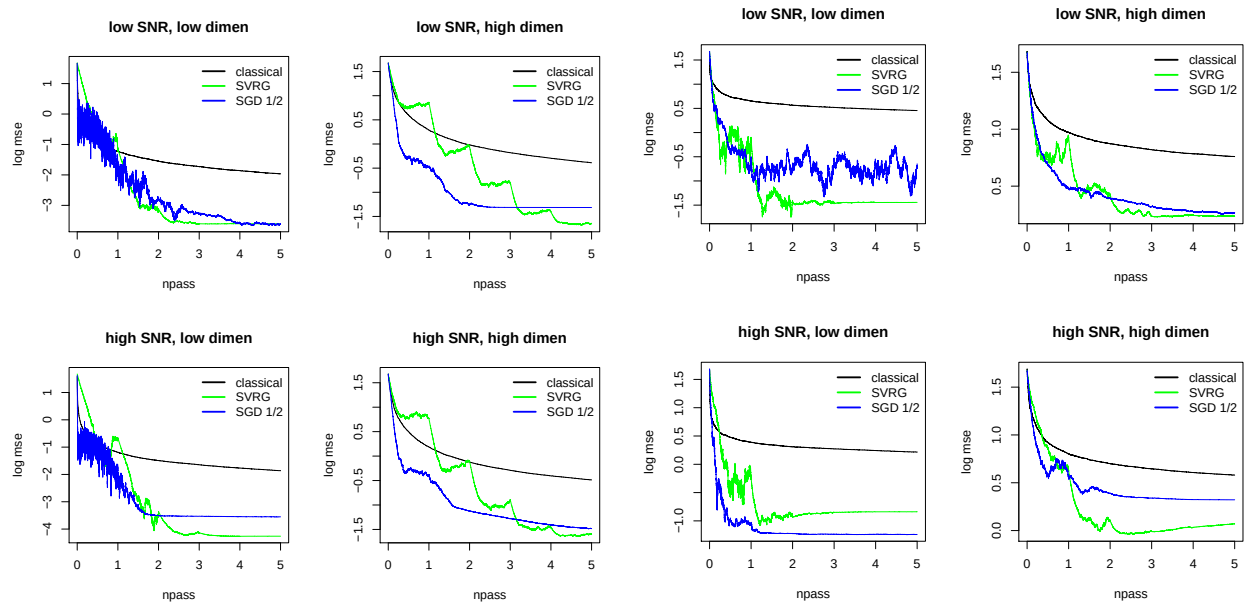

Figure 2: Simulated data experiments, comparing the performance of our procedure $\texttt{ISGD}^{1/2}$ against SVRG and classical ISGD. The left four plots are with the normal model, the right four plots with the logistic model.

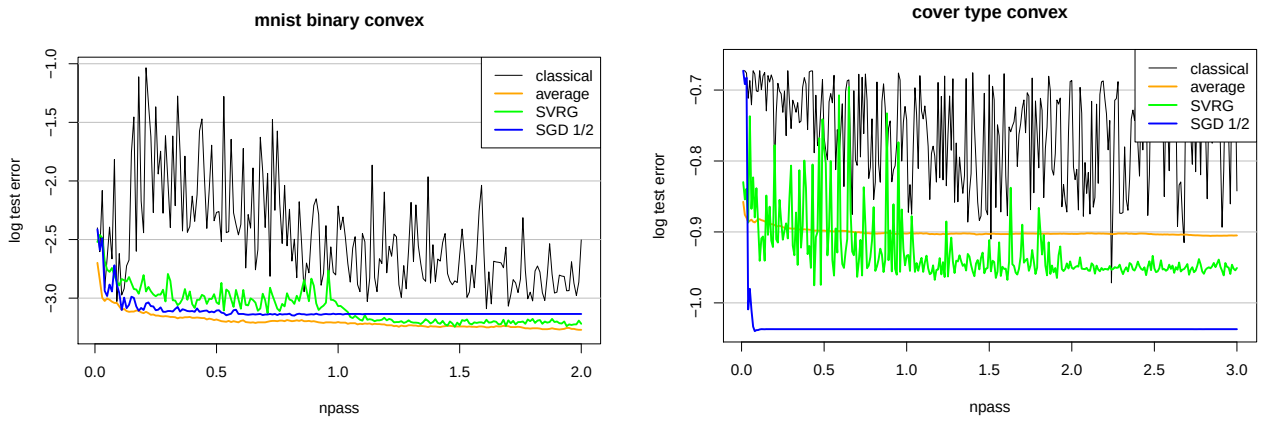

Figure 3: Benchmark data sets with binary logistic regression using $\texttt{ISGD}^{1/2}$, SVRG, classical ISGD, and averaged ISGD. Prediction error on a held out test set. MNIST (binary) on the left, COVERTYPE (binary) on the right.

5 Conclusion

In this paper we focused on detecting convergence of SGD with constant learning rate to its convergence phase. This is an important practical task because statistical properties of iterative stochastic procedures are better understood under stationarity. We borrowed from the theory of stopping times in stochastic approximation to develop a simple diagnostic that uses the inner product of successive gradients to detect convergence. Theoretical and empirical results suggest that the diagnostic reliably detects the phase transition, which can speed up classical procedures.

Future work needs to focus on analysis of errors $||\theta_n - \theta_\star||^2$ conditional on the diagnostic being activated. This could show that the error is uncorrelated with the initial starting point conditional on the test being activated, and so provide theoretical support to the empirical results in Table 1. It would also be interesting to focus more on $\texttt{ISGD}^{1/2}$ and analyze its behavior. Another idea is to run parallel $\texttt{ISGD}^{1/2}$ chains and aggregate the iterates. At stationarity we expect iterates from different chains to be uncorrelated with each other, and so averaging may help. It would also be interesting to use the diagnostic in problems with non-convex loss, such as neural networks.

References

[1] Albert Benveniste, Pierre Priouret, and Michel Métivier. *Adaptive algorithms and stochastic approximations*. Springer-Verlag New York, Inc., 1990.

[2] Dimitri P Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical programming*, 129(2):163–195, 2011.

[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*, pages 203–208. ACM, 1999.

[4] Vivek S Borkar. Stochastic approximation. *Cambridge Books*, 2008.

[5] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[6] Leon Bottou. Stochastic Gradient Descent Tricks. In *Neural Networks: Tricks of the Trade*, volume 1, pages 421–436. 2012.

[7] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *arXiv preprint arXiv:1606.04838*, 2016.

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

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

[10] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

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

[12] Brian Kulis and Peter L Bartlett. Implicit online learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 575–582, 2010.

[13] Peter McCullagh. Generalized linear models. *European Journal of Operational Research*, 16(3):285–292, 1984.

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

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

[16] Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In *Advances in Neural Information Processing Systems*, pages 1017–1025, 2014.

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

[18] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):123–231, 2013.

[19] G Ch Pflug. Gradient estimates for the performance of markov chains and discrete event processes. *Annals of Operations Research*, 39(1):173–194, 1992.

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

[21] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[22] Lorenzo Rosasco, Silvia Villa, and Bang Công Vũ. Convergence of stochastic proximal gradient algorithm. *arXiv preprint arXiv:1403.5074*, 2014.

[23] Nicolas Le Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence _rate for finite training sets. In *Advances in Neural Information Processing Systems*, pages 2663–2671, 2012.

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

[25] 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.

[26] Panos Toulis, Jason Rennie, and Edoardo Airoldi. Statistical analysis of stochastic gradient methods for generalized linear models. In *31st International Conference on Machine Learning*, 2014.

[27] Panos Toulis, Dustin Tran, and Edo Airoldi. Towards stability and optimality in stochastic gradient descent. In *Artificial Intelligence and Statistics*, pages 1290–1298, 2016.

[28] Lin Xiao and Tony Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24:2057–2075, 2014.

[29] Wei Xu. Towards optimal one pass large scale learning with averaged stochastic gradient descent. *arXiv preprint arXiv:1107.2490*, 2011.

[30] Dong Yin, Ashwin Pananjady, Max Lam, Dimitris Papailiopoulos, Kannan Ramchandran, and Peter Bartlett. Gradient diversity: a key ingredient for scalable distributed learning. In *Proceedings of 21st International Conference on Artificial Intelligence and Statistics (AISTATS'18)*, 2018.

[31] 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.

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