On Convergence-Diagnostic based Step Sizes for Stochastic Gradient Descent

Scott Pesme, Aymeric Dieuleveut, Nicolas Flammarion

On Convergence-Diagnostic based Step Sizes for Stochastic Gradient Descent

Scott Pesme [1] **Aymeric Dieuleveut** [2] **Nicolas Flammarion** [1]

Abstract

Constant step-size Stochastic Gradient Descent exhibits two phases: a transient phase during which iterates make fast progress towards the optimum, followed by a stationary phase during which iterates oscillate around the optimal point. In this paper, we show that efficiently detecting this transition and appropriately decreasing the step size can lead to fast convergence rates. We analyse the classical statistical test proposed by Pflug (1983), based on the inner product between consecutive stochastic gradients. Even in the simple case where the objective function is quadratic we show that this test cannot lead to an adequate convergence diagnostic. We then propose a novel and simple statistical procedure that accurately detects stationarity and we provide experimental results showing state-of-the-art performance on synthetic and real-world datasets.

1. Introduction

The field of machine learning has had tremendous success in recent years, in problems such as object classification (He et al., 2016) and speech recognition (Graves et al., 2013). These achievements have been enabled by the development of complex optimization-based architectures such as deep-learning, which are efficiently trainable by Stochastic Gradient Descent algorithms (Bottou, 1998).

Challenges have arisen on both the theoretical front – to understand why those algorithms achieve such performance, and on the practical front – as choosing the architecture of the network and the parameters of the algorithm has become an art itself. Especially, there is no practical heuristic to set the step-size sequence. As a consequence, new optimization strategies have appeared to alleviate the tuning burden, as Adam (Kingma & Ba, 2014), together with new learning rate scheduling, such as cyclical learning rates (Smith, 2017) and warm restarts (Loshchilov & Hutter, 2016). However those strategies typically do not come with theoretical guarantees and may be outperformed by SGD (Wilson et al., 2017).

Even in the classical case of convex optimization, in which convergence rates have been widely studied over the last 30 years (Polyak & Juditsky, 1992; Zhang, 2004; Nemirovski et al., 2009; Bach & Moulines, 2011; Rakhlin et al., 2012) and where theory suggests to use the *averaged iterate* and provides optimal choices of learning rates, practitioners still face major challenges: indeed (a) averaging leads to a slower decay during early iterations, (b) learning rates may not adapt to the difficulty of the problem (the optimal decay depends on the class of problems), or may not be robust to constant misspecification. Consequently, the state of the art approach in practice remains to use the *final iterate* with decreasing step size $a/(b+t^{\alpha})$ with constants a, b, α obtained by a tiresome hand-tuning. Overall, there is a desperate need for adaptive algorithms.

In this paper, we study *adaptive step-size scheduling* based on *convergence diagnostic*. The behaviour of SGD with constant step size is dictated by (a) a *bias term*, that accounts for the impact of the initial distance $\|\theta_0 - \theta_*\|$ to the minimizer θ_* of the function, and (b) a *variance term* arising from the noise in the gradients. Larger steps allow to forget the initial condition faster, but increase the impact of the noise. Our approach is then to use the largest possible learning rate as long as the iterates make progress and to *automatically* detect when they stop making any progress. When we have reached such a *saturation*, we reduce the learning rate. This can be viewed as "restarting" the algorithm, even though only the learning rate changes. We refer to this approach as *Convergence-Diagnostic algorithm*. Its benefits are thus twofold: (i) with a large initial learning rate the bias term initially decays at an *exponential rate* (Kushner & Huang, 1981; Pflug, 1986), (ii) decreasing the learning rate when the effect of the noise becomes dominant defines an efficient and practical adaptive strategy.

Reducing the learning rate when the objective function stops decaying is widely used in deep learning (Krizhevsky et al., 2012) but the epochs where the step size is reduced are mostly hand-picked. Our goal is to select them automatically by detecting saturation. Convergence diagnostics date back to Pflug (1983), who proposed to use the inner product

[1]Theory of Machine Learning lab, EPFL [2]École Polytechnique. Correspondence to: Scott Pesme <scott.pesme@epfl.ch>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

between consecutive gradients to detect convergence. Such a strategy has regained interest in recent years: Chee & Toulis (2018) provided a similar analysis for quadratic functions, and Yaida (2018) considers SGD with momentum and proposes an analogous restart criterion using the expectation of an observable quantity under the limit distribution, achieving the same performance as hand-tuned methods on two simple deep learning models. However, none of these papers provide a convergence rate and we show that Pflug's approach provably fails in simple settings. Lang et al. (2019) introduced Statistical Adaptive Stochastic Approximation which aims to improve upon Pflug's approach by formalizing the testing procedure. However, their strategy leads to a very small number of reductions of the learning rate.

An earlier attempt to adapt the learning rate depending on the directions in which iterates are moving was made by Kesten (1958). Kesten's rule decreases the step size when the iterates stop moving consistently in the same direction. Originally introduced in one dimension, it was generalized to the multi-dimensional case and analyzed by Delyon & Juditsky (1993).

Finally, some orthogonal approaches have also been used to automatically change the learning rate: it is for example possible to consider the step size as a parameter of the risk of the algorithm, and to update the step size using another meta-optimization algorithm (Sutton, 1981; Jacobs, 1988; Benveniste et al., 1990; Sutton, 1992; Schraudolph, 1999; Kushner & Yang, 1995; Almeida et al., 1999).

Another line of work consists in changing the learning rate for each coordinate depending on how much iterates are moving (Duchi et al., 2011; Zeiler, 2012). Finally, Schaul et al. (2013) propose to use coordinate-wise adaptive learning rates, that maximize the decrease of the expected loss on separable quadratic functions.

We make the following contributions:

- We provide convergence results for the Convergence-Diagnostic algorithm when used with the oracle diagnostic for smooth and strongly-convex functions.
- We show that the intuition for Pflug's statistic is valid for all smooth and strongly-convex functions by computing the expectation of the inner product between two consecutive gradients both for an arbitrary starting point, and under the stationary distribution.
- We show that despite the previous observation the empirical criterion is provably inefficient, even for a simple quadratic objective.
- We introduce a new *distance-based diagnostic* based on a simple heuristic inspired from the quadratic setting with additive noise.
- We illustrate experimentally the failure of Pflug's statistic, and show that the distance-based diagnostic competes with state-of-the-art methods on a variety of loss functions, both on synthetic and real-world datasets.

The paper is organized as follows: in Section 2, we introduce the framework and present the assumptions. Section 3 we describe and analyse the oracle convergence-diagnostic algorithm. In Section 4, we show that the classical criterion proposed by Pflug cannot efficiently detect stationarity. We then introduce a new distance-based criterion Section 5 and provide numerical experiments in Section 6.

2. Preliminaries

Formally, we consider the minimization of a risk function f defined on $\mathbb{R}^d$ given access to a sequence of unbiased estimators of f's gradients (Robbins & Monro, 1951). Starting from an arbitrary point θ_0, at each iteration $i+1$ we get an unbiased random estimate $f'_{i+1}(\theta_i)$ of the gradient $f'(\theta_i)$ and update the current estimator by moving in the opposite direction of the stochastic gradient:

$$\theta_{i+1} = \theta_i - \gamma_{i+1} f'_{i+1}(\theta_i), \tag{1}$$

where $\gamma_{i+1} > 0$ is the *step size*, also referred to as *learning rate*. We make the following assumptions on the stochastic gradients and the function f.

Assumption 1 (Unbiased gradient estimates)**.** *There exists a filtration $(\mathcal{F}_i)_{i\geq 0}$ such that θ_0 is $\mathcal{F}_0$-measurable, f'_i is $\mathcal{F}_i$-measurable for all $i \in \mathbb{N}$, and for each $\theta \in \mathbb{R}^d$: $\mathbb{E}\left[f'_{i+1}(\theta) \mid \mathcal{F}_i\right] = f'(\theta)$. In addition $(f_i)_{i\geq 0}$ are identically distributed random fields.*

Assumption 2 (L-smoothness)**.** *For all $i \geq 1$, the function f_i is almost surely L-smooth and convex:*

$$\forall \theta, \eta \in \mathbb{R}^d, \left\|f'_i(\theta) - f'_i(\eta)\right\| \leq L \left\|\theta - \eta\right\|.$$

Assumption 3 (Strong convexity)**.** *There exists a finite constant $\mu > 0$ such that for all $\theta, \eta \in \mathbb{R}^d$:*

$$f(\theta) \geq f(\eta) + \langle f'(\eta),\, \theta - \eta\rangle + \frac{\mu}{2}\left\|\theta - \eta\right\|^2.$$

For $i > 0$ and $\theta \in \mathcal{W}$, we denote by $\varepsilon_i(\theta) = f'_i(\theta) - f'(\theta)$ the noise, for which we consider the following assumption:

Assumption 4 (Bounded variance)**.** *There exists a constant $\sigma \geq 0$ such that for any $i > 0$, $\mathbb{E}\left[\left\|\varepsilon_i(\theta^*)\right\|^2\right] \leq \sigma^2$.*

Under Assumptions 1 and 4 we define the noise covariance as the function $\mathcal{C} : \mathbb{R}^d \mapsto \mathbb{R}^{d\times d}$ defined for all $\theta \in \mathbb{R}^d$ by $\mathcal{C}(\theta) = \mathbb{E}\left[\varepsilon(\theta)\varepsilon(\theta)^T\right]$.

In the following section we formally describe the restart strategy and give a convergence rate in the omniscient setting where all the parameters are known.

3. Bias-variance decomposition and stationarity diagnostic

When the step size γ is constant, the sequence of iterates $(\theta_n)_{n\geq 0}$ produced by the SGD recursion in eq. (1) is a homogeneous Markov chain. Under appropriate conditions (Dieuleveut et al., 2017), this Markov chain has a unique stationary distribution, denoted by π_γ, towards which it converges exponentially fast. This is the *transient phase*. The rate of convergence is proportional to γ and therefore a larger step size leads to a faster convergence.

When the Markov chain has reached its stationary distribution, i.e. in the *stationary phase*, the iterates make negligible progress towards the optimum θ^* but stay in a bounded region of size $O(\sqrt{\gamma})$ around it. More precisely, Dieuleveut et al. (2017) make explicit the expansion $\mathbb{E}_{\pi_\gamma}\left[\|\theta - \theta^*\|^2\right] = b\gamma + O(\gamma^2)$ where the constant b depends on the function f and on the covariance of the noise $\mathcal{C}(\theta^*)$ at the optimum . Hence the smaller the step size and the closer the iterates $(\theta_n)_{n\geq 0}$ get to the optimum θ^*.

Therefore a clear trade-off appears between: (a) using a large step size with a fast transient phase but a poor approximation of θ^* and (b) using a small step size with iterates getting close to the optimum but taking longer to get there. This *bias-variance* trade-off is directly transcribed in the following classical proposition (Needell et al., 2014).

Proposition 5. *Consider the recursion in eq. (1) under Assumptions 1 to 4. Then for any step-size $\gamma \in (0, 1/2L)$ and $n \geq 0$ we have:*

$$\mathbb{E}\left[\|\theta_n - \theta^*\|^2\right] \leq (1-\gamma\mu)^n \mathbb{E}\left[\|\theta_0 - \theta^*\|^2\right] + \frac{2\gamma\sigma^2}{\mu}.$$

The performance of the algorithm is then determined by the sum of a *bias term* – characterizing how fast the initial condition θ_0 is forgotten and which is increasing with $\|\theta_0 - \theta^*\|$; and a *variance term* – characterizing the effect of the noise in the gradient estimates and that increases with the variance of the noise σ^2. Here the bias converges exponentially fast whereas the variance is $O(\gamma)$. Note that the bias decrease is of the form $(1-\gamma\mu)^n\delta_0$, which means that the typical number of iterations to reach stationarity is $\Theta(\gamma^{-1})$.

As noted by Bottou et al. (2018), this decomposition naturally leads to the question: which convergence rate can we hope getting if we keep a large step size as long as progress is being made but decrease it as soon as the iterates saturate? More explicitly, starting from θ_0, one could run SGD with a constant step size γ_0 for Δn_1 steps until progress stalls. Then for $n \geq \Delta n_1$, a smaller step size $\gamma_1 = r\gamma_0$ (where $r \in (0,1)$) is used in order to decrease the variance and therefore get closer to θ^* and so on. This simple strategy is implemented in Algorithm 1. However the crucial difficulty here lies in detecting the saturation. Indeed

Algorithm 1 Convergence-Diagnostic algorithm

Input: Starting point θ_0, Step size $\gamma_0 > 0$, Step-size decrease $r \in (0,1)$
Output: θ_N
$\gamma \leftarrow \gamma_0$
for $n = 1$ to N **do**
  $\theta_n \leftarrow \theta_{n-1} - \gamma f_n'(\theta_{n-1})$
  if { Saturation Diagnostic } is True **then**
    $\gamma \leftarrow r \times \gamma$
  end if
end for
Return: θ_N

Algorithm 2 Oracle diagnostic

Input: γ, δ_0, μ, L, σ^2, n
Output: Diagnostic boolean
Bias $\leftarrow (1-\gamma\mu)^n\delta_0$
Variance $\leftarrow \frac{2\gamma\sigma^2}{\mu}$
Return: { Bias $<$ Variance }

when running SGD we do not have access to $\|\theta_n - \theta^*\|$ and we cannot evaluate the successive function values $f(\theta_n)$ because of their prohibitively expensive cost to estimate. Hence, we focus on finding a statistical diagnostic which is computationally cheap and that gives an accurate restart time corresponding to saturation.

Oracle diagnostic. Following this idea, assume first we have access to all the parameters of the problem: $\|\theta_0 - \theta^*\|$, μ, L, σ^2. Then reaching saturation translates into the bias term and the variance term from Proposition 5 being of the same magnitude, i.e.

$$(1-\gamma_0\mu)^{\Delta n_1}\|\theta_0 - \theta^*\|^2 = \frac{2\gamma_0\sigma^2}{\mu}.$$

This oracle diagnostic is formalized in Algorithm 2. The following proposition guarantees its performance.

Proposition 6. *Under Assumptions 1 to 4, consider Algorithm 1 instantiated with Algorithm 2 and parameter $r \in (0,1)$. Let $\gamma_0 \in (0, 1/2L)$, $\delta_0 = \|\theta_0 - \theta^*\|^2$ and $\Delta n_1 = \frac{1}{\gamma_0\mu}\log(\frac{\mu\delta_0}{2\gamma_0\sigma^2})$. Then, we have for all $n \leq \Delta n_1$:*

$$\mathbb{E}\left[\|\theta_n - \theta^*\|^2\right] \leq (1-\gamma_0\mu)^n\delta_0 + \frac{2\gamma_0\sigma^2}{\mu},$$

and for all $n > \Delta n_1$:

$$\mathbb{E}\left[\|\theta_n - \theta^*\|^2\right] \leq \frac{8\sigma^2}{\mu^2(n-\Delta n_1)(1-r)}\ln\left(\frac{2}{r}\right).$$

The proof of this Proposition is given in Appendix B.1. We make the following observations:

- The rate $O(1/\mu^2 n)$ is optimal for last-iterate convergence for strongly-convex problem (Nguyen et al., 2019) and is also obtained by SGD with decreasing step size $\gamma_n = C/\mu n$ where $C > 2$ (Bach & Moulines, 2011). More generally, the rate $O(1/n)$ is known to be information-theoretically optimal for strongly-convex stochastic approximation (Nemirovsky & Yudin, 1983).
- To reach an ε-optimal point, $O\big(\frac{\sigma^2}{\mu^2\varepsilon} + \frac{L}{\mu}\log(\frac{\mu L \delta_0}{\sigma^2})\big)$ calls to the gradient oracle are needed. Therefore the bias is forgotten exponentially fast. This stands in sharp contrast to averaged SGD for which there is no exponential forgetting of initial conditions (Bach & Moulines, 2011).
- We present in Appendix B.2 additional results for weakly and uniformly convex functions. In this case too, the oracle diagnostic-based algorithm recovers the optimal rates of convergence. However these results hold only for the restart iterations n_k, and the behaviour in between each can be theoretically arbitrarily bad.
- Our algorithm shares key similarities with the algorithm of Hazan & Kale (2014) which halves the learning rate every 2^k iterations but with the different aim of obtaining the sharp $O(1/n)$ rate in the non-smooth setting.

This strategy is called oracle since all the parameters must be known and, in that sense, Algorithm 2 is clearly non practical. However Proposition 6 shows that Algorithm 1 implemented with a practical and suitable diagnostic is *a priori* a good idea since it leads to the optimal rate $O(1/\mu^2 n)$ without having to know the strong convexity parameter μ and the rate α of decrease of the step-size sequence $\gamma_n = O(n^{-\alpha})$. The aim of the following sections is to propose a computationally cheap and efficient statistic that detects the transition between transience and stationarity.

4. Pflug's Statistical Test for stationarity

In this section we analyse a statistical diagnostic first developed by Pflug (1983) which relies on the sign of the inner product of two consecutive stochastic gradients $\langle f'_{k+1}(\theta_k),\ f'_{k+2}(\theta_{k+1})\rangle$. Though this procedure was developed several decades ago, no theoretical analysis had been proposed yet despite the fact that several papers have recently showed renewed interest in it (Chee & Toulis, 2018; Lang et al., 2019; Sordello & Su, 2019). Here we show that whilst it is true this statistic becomes in expectation negative at stationarity, it is provably inefficient to properly detect the restart time – for the particular example of quadratic functions.

4.1. Control of the expectation of Pflug's statistic

The general motivation behind Pflug's statistic is that during the transient phase the inner product is in expectation positive and during the stationary phase, it is in expectation negative. Indeed, in the transient phase, where $\|\theta - \theta^*\| >> \sqrt{\gamma}\sigma$, the effect of the noise is negligible and the behavior of the iterates is very similar to the one of noiseless gradient descent (i.e, $\varepsilon(\theta) = 0$ for all $\theta \in \mathbb{R}^d$) which satisfies:

$$\langle f'(\theta), f'(\theta - \gamma f'(\theta))\rangle = \|f'(\theta)\|^2 + O(\gamma) > 0.$$

On the other hand, in the stationary phase, we may intuitively assume starting from $\theta_0 = \theta^*$ to obtain

$$\begin{aligned}\mathbb{E}\left[\langle f'_1(\theta_0),\, f'_2(\theta_1)\rangle\right] &= -\mathbb{E}\left[\langle \varepsilon_1,\ f'(\theta^* + \gamma\varepsilon_1)\rangle\right] \\ &= -\gamma \operatorname{Tr}\ f''(\theta^*)\mathbb{E}\left[\varepsilon_1\varepsilon_1^\top\right] + O(\gamma) < 0.\end{aligned}$$

The single values $\langle f'_{k+1}(\theta_k),\ f'_{k+2}(\theta_{k+1})\rangle$ are too noisy, which leads (Pflug, 1983) in considering the running average:

$$S_n = \frac{1}{n}\sum_{k=0}^{n-1}\langle f'_{k+1}(\theta_k),\ f'_{k+2}(\theta_{k+1})\rangle.$$

This average can easily be computed online with negligible extra computational and memory costs. Pflug (1983) then advocates to decrease the step size when the statistic becomes negative, as explained in Algorithm 1. A burn-in delay n_b can also be waited to avoid the first noisy values.

Algorithm 3 Pflug's diagnostic

Input: $(f'_k(\theta_{k-1}))_{0\le k\le n}$, $n_b > 0$
Output: Diagnostic boolean
$S \leftarrow 0$
for $k = 2$ to n **do**
  $S \leftarrow S + \langle f'_k(\theta_{k-1}),\ f'_{k-1}(\theta_{k-2})\rangle$
end for
Return : $\{S < 0\}$ AND $\{n > n_b\}$

For quadratic functions, Pflug (1988a) first shows that, when $\theta \sim \pi_\gamma$ at stationarity, the inner product of two successive stochastic gradients is negative in expectation. To extend this result to the wider class of smooth strongly convex functions, we make the following technical assumptions.

Assumption 7 (Five-times differentiability of f)**.** *The function f is five times continuously differentiable with second to fifth uniformly bounded derivatives.*

Assumption 8 (Differentiability of the noise)**.** *The noise covariance function $\mathcal{C}$ is three times continuously differentiable with locally-Lipschitz derivatives. Moreover $\mathbb{E}(\|\varepsilon_1(\theta^*)\|^6)$ is finite.*

These assumptions are satisfied in natural settings. The following proposition addresses the sign of the expectation of Pflug's statistic.

Proposition 9. *Under Assumptions 1 to 4, 7 and 8, for $\gamma \in (0, 1/2L)$, let π_γ be the unique stationary distribution. Let $\theta_1 = \theta_0 - \gamma f_1'(\theta_0)$. For any starting point θ_0, we have*

$$\mathbb{E}\left[\langle f_1'(\theta_0),\, f_2'(\theta_1)\rangle\right] \geq (1-\gamma L)\left\|f'(\theta_0)\right\|^2 - \gamma L \operatorname{Tr} \mathcal{C}(\theta_0) + O(\gamma^2).$$

And for $\theta_0 \sim \pi_\gamma$, we have:

$$\mathbb{E}_{\pi_\gamma}\left[\langle f_1'(\theta_0),\, f_2'(\theta_1)\rangle\right] = -\frac{1}{2}\gamma \operatorname{Tr}\, f''(\theta^*)\mathcal{C}(\theta^*) + O(\gamma^{3/2}).$$

Sketch of Proof. The complete proof is given in Appendix C.1. The first part relies on a simple Taylor expansion of f' around θ_0. For the second part, we decompose:

$$\mathbb{E}[\langle f_1'(\theta_0),\, f_2'(\theta_1)\rangle \mid \theta_0] = \underbrace{\mathbb{E}\left[\langle f'(\theta_0), f'(\theta_1)\rangle \mid \theta_0\right]}_{S_{\text{grad}}} + \underbrace{\mathbb{E}\left[\langle \varepsilon_1, f'(\theta_1) \mid \theta_0\rangle\right]}_{S_{\text{noise}}}.$$

Then, applying successive Taylor expansions of f' around the optimum θ^* yields for both terms:

$$S_{\text{grad}} = \operatorname{Tr}\, f''(\theta^*)^2(\theta_0 - \theta^*)^{\otimes 2} + O(\gamma^{3/2}),$$
$$S_{\text{noise}} = -\gamma \operatorname{Tr}\, f''(\theta^*)\mathcal{C}(\theta_0) + O(\gamma^{3/2}).$$

Using results from Dieuleveut et al. (2017) on $\mathbb{E}_{\pi_\gamma}\left[(\theta_0 - \theta^*)^{\otimes 2}\right]$ and $\mathbb{E}_{\pi_\gamma}\left[\mathcal{C}(\theta_0)\right]$ then leads to

$$\mathbb{E}_{\pi_\gamma}\left[S_{\text{grad}}\right] = \frac{1}{2}\gamma \operatorname{Tr}\, f''(\theta^*)\mathcal{C}(\theta^*) + O(\gamma^{3/2}),$$
$$\mathbb{E}_{\pi_\gamma}\left[S_{\text{noise}}\right] = -\gamma \operatorname{Tr}\, f''(\theta^*)\mathcal{C}(\theta^*) + O(\gamma^{3/2}). \qquad \square$$

We note that, counter intuitively, the inner product is not negative because the iterates bounce around θ^* (we still have $S_{\text{grad}} = \mathbb{E}\left[\langle f'(\theta_1),\ f'(\theta_0)\rangle\right] > 0$), but because the noise part $S_{\text{noise}} = \mathbb{E}\left[\langle \varepsilon_1,\ f'(\theta_1)\rangle\right]$ is negative and dominates the gradient part S_{grad}.

In the case where f is quadratic we immediately recover the result of Pflug (1988b). We note that Chee & Toulis (2018) show a similar result but under far more restrictive assumptions on the noise distribution and the step size.

Proposition 9 establishes that the sign of the expectation of the inner product between two consecutive gradients characterizes the transient and stationary regimes: for an iterate θ_0 far away from the optimum, i.e. such that $\|\theta_0 - \theta^*\|$ is large, the expected value of the statistic is positive whereas it becomes negative when the iterates reach stationarity. This makes clear the motivation of considering the sign of the inner products as a convergence diagnostic. Unfortunately this result does not guarantee the good performance of this statistic. Even though the inner product is negative, its value is only $O(\gamma)$. It is then difficult to distinguish $\langle f'_{k+1},\ f'_{k+2}\rangle$ from zero for small step size γ. In fact, we now show that even for simple quadratic functions, the statistical test is unable to offer an adequate convergence diagnostic.

4.2. Failure of Pflug's method for Quadratic Functions

In this section we show that Pflug's diagnostic fails to accurately detect convergence, even in the simple framework of quadratic objective functions with additive noise. While we have demonstrated in Proposition 9 that the sign of its expectation characterizes the transient and stationary regime, we show that the running average S_n does not concentrate enough around its mean to result in a valid test. Intuitively, from a restart when we leave stationarity: (1) the expectation is positive but smaller than γ , and (2) the standard deviation of S_n is not decaying with γ, but only with the number of steps over which we average, as $1/\sqrt{n}$. As a consequence, in order to ensure that the sign of S_n is the same as the sign of its expectation, we would need to average over more than $1/\gamma^2$ steps, which is orders of magnitude bigger than the optimal restart time of $\Theta(1/\gamma)$ (See Section 3). We make this statement quantitative under simple assumptions on the noise.

Assumption 10 (Quadratic semi-stochastic setting)**.** *There exists a symmetric positive semi-definite matrix H such that $f(\theta) = \frac{1}{2}\theta^T H \theta$. The noise $\varepsilon_i(\theta) = \xi_i$ is independent of θ and:*

$$(\xi_i)_{i\geq 0} \text{ are i.i.d. },\ \mathbb{E}\left[\xi_i\right] = 0,\ \mathbb{E}\left[\xi_i^T \xi_i\right] = C.$$

In addition we make a simple assumption on the noise:

Assumption 11 (Noise symmetry and continuity)**.** *The function $\mathbb{P}\left(\xi_1^T\xi_2 \geq x\right)$ is continuous in $x = 0$ and*

$$\mathbb{P}\left(\xi_1^T\xi_2 \geq x\right) = \mathbb{P}\left(\xi_1^T\xi_2 \leq -x\right) \quad \text{for all } x \geq 0.$$

This assumption is made for ease of presentation and can be relaxed. We make use of the following notations. We assume SGD is run with a constant step size γ_{old} until the stationary distribution $\pi_{\gamma_{old}}$ is reached. The step size is then decreased and SGD is run with a smaller step $\gamma = r \times \gamma_{old}$. Hence the iterates cease to be at stationarity under $\pi_{\gamma_{old}}$ and start a transient phase towards π_γ. We denote by $\mathbb{E}_{\theta_0\sim\gamma_{old}}$ (resp. $\mathbb{P}_{\theta_0\sim\gamma_{old}}$) the expectation (resp. probability) of a random variable (resp. event) when the initial θ_0 is sampled from the old distribution $\pi_{\gamma_{old}}$ and a new step size $\gamma = r \times \gamma_{old}$ is used. Note that $\mathbb{E}_{\theta_0\sim\gamma_{old}}$ and $\mathbb{E}_{\pi_\gamma}$ have different meanings, the latter being the expectation under π_γ.

We first split S_n in a γ-dependent and a γ-independent part.

Lemma 12. *Under Assumption 10, let $\theta_0 \sim \pi_{\gamma_{old}}$ and assume we run SGD with a smaller step size $\gamma = r \times \gamma_{old}$, $r \in (0, 1)$. Then, the statistic S_n can be decomposed as: $S_n = -R_{n,\gamma} + \chi_n$. The part χ_n is independent of γ and*

$$\mathbb{E}_{\theta_0\sim\pi_{\gamma_{old}}}\left[R_{n,\gamma}^2\right] \leq M\left(\frac{\gamma}{n} + \gamma^2\right);$$
$$\mathbb{E}\left[\chi_n\right] = 0\,,\ \operatorname{Var}(\chi_n) = \frac{1}{n}\operatorname{Tr}\left(C^2\right) \text{ and}$$
$$\operatorname{Var}(\chi_n^2) = \frac{\mathbb{E}\left[(\xi_1^T\xi_2)^4\right] - \operatorname{Tr}^2\, C^2}{n^3},$$

where M is independent of γ and n.

Thus the variance of χ_n does not depend on γ while, from a restart, the second moment $\mathbb{E}_{\theta_0\sim\pi_{\gamma_{old}}}\left[R^2_{n,\gamma}\right]$ is $O(\frac{\gamma}{n}+\gamma^2)$. Therefore the signal to noise ratio is high. This property is the main idea behind the proof of the following proposition.

Proposition 13. *Under Assumptions 10 and 11, let $\theta_0 \sim \pi_{\gamma_{old}}$ and run SGD with $\gamma = r \times \gamma_{old}$, $r \in (0,1)$. Then for all $0 \leq \alpha < 2$, and $n_\gamma = O(\gamma^{-\alpha})$ we have:*

$$\lim_{\gamma\to 0} \mathbb{P}_{\theta_0\sim\pi_{\gamma_{old}}}\left(S_{n_\gamma} \leq 0\right) = \frac{1}{2}.$$

Sketch of Proof. The complete proofs of Lemma 12 and Proposition 13 are given in Appendix C.2. The main idea is that the signal to noise ratio is too high. The signal during the transient phase is positive and $O(\gamma)$. However the variance of S_n is $O(1/n)$. Hence $\Omega(1/\gamma^2)$ iterations are typically needed in order to have a clean signal. Before this threshold, S_n resembles a random walk and its sign gives no information on whether saturation is reached or not, this leads to early on restarts. □

We make the following observations.

- Note that the typical time to reach saturation with a constant step size γ is of order $1/\gamma$ (see Section 3). We should expect Pflug's statistic to satisfy $\lim_{\gamma\to 0}\mathbb{P}_{\theta_0\sim\pi_{\gamma_{old}}}\left(S_{n_b} \leq 0\right) = 0$ for all constant burn-in time n_b smaller than the typical saturation time $O(1/\gamma)$ – since the statistic should not detect saturation before it is actually reached. Proposition 13 shows that this is not the case and that the step size is therefore decreased too early. This phenomenon is clearly seen in Fig. 1 in Section 6.
- We note that Pflug (1988a) describes an opposite result. We believe this is due to a miscalculation of $\mathrm{Var}(\chi_n)$ in his proof (see detail in Appendix C.3).
- Lang et al. (2019) similarly point out the existence of a large variance in the diagnostic proposed by Yaida (2018). They make the strategy more robust by implementing a formal statistical test, to only reduce the learning rate when the limit distribution has been reached with *high confidence*. Unfortunately, Proposition 13 entails that more than $O(1/\gamma^2)$ iterations are needed to accurately detect convergence for Pflug's statistic, and we thus believe that Lang's approach would be too conservative and would not reduce the learning rate often enough.

Hence Pflug's diagnostic is inadequate and leads to poor experimental results (see Section 6). We propose then a novel simple distance-based diagnostic which enjoys state-of-the art rates for a variety of classes of convex functions.

5. A new distance-based statistic

We propose here a very simple statistic based on the distance between the current iterate θ_n and the iterate from which the step size has been last decreased. Indeed, we would ideally like to decrease the step size when $\|\eta_n\| = \|\theta_n - \theta^*\|$ starts to saturate. Since the optimum θ^* is not known, we cannot track the evolution of this criterion. However it has a similar behaviour as $\|\Omega_n\| = \|\theta_n - \theta_0\|$, which we can compute. This is seen through the simple equation

$$\|\Omega_n\|^2 = \|\eta_n\|^2 + \|\eta_0\|^2 - 2\langle\eta_n,\ \eta_0\rangle.$$

The value $\|\eta_n\|^2$ is then expected to saturate roughly at the same time as $\|\Omega_n\|^2$. In addition, $\|\theta_n - \theta_0\|^2$ describes a large range of values which can be easily tracked, starting at 0 and roughly finishing around $\|\theta^* - \theta_0\|^2 + O(\gamma)$ (see Corollary 15). It is worth noting this would not be the case if a different referent point, $\tilde{\theta} \neq \theta_0$, was considered.

To find a heuristic to detect the convergence of $\|\theta_n - \theta_0\|^2$, we consider the particular setting of a quadratic objective with additive noise stated in Assumption 10. In this framework we can compute the evolution of $\mathbb{E}\left[\|\Omega_n\|^2\right]$ in closed-form .

Proposition 14. *Let $\theta_0 \in \mathbb{R}^d$ and $\gamma \in (0, 1/L)$. Let $\Omega_n = \theta_n - \theta_0$. Under Assumption 10 we have that:*

$$\mathbb{E}\left[\|\Omega_n\|^2\right] = \eta_0^T[I-(I-\gamma H)^n]^2\eta_0 + \gamma\ \mathrm{Tr}\left[I-(I-\gamma H)^{2n}\right](2I-\gamma H)^{-1}H^{-1}C.$$

The proof of this result is given in Appendix D. We can analyse this proposition in two different settings: for small values of n at the beginning of the process and when the iterates θ_n have reached stationarity.

Corollary 15. *Let $\theta_0 \in \mathbb{R}^d$ and $\gamma \in [0, 1/L]$. Let $\Omega_n = \theta_n - \theta_0$. Under Assumption 10 we have that for all $n \geq 0$:*

$$\mathbb{E}_{\pi_\gamma}\left[\|\Omega_n\|^2\right] = \|\eta_0\|^2 + \gamma\ \mathrm{Tr}\ H^{-1}C(2I-\gamma H)^{-1},$$
$$\mathbb{E}\left[\|\Omega_n\|^2\right] = \gamma^2\eta_0^T H^2\eta_0 \times n^2 + \gamma^2\ \mathrm{Tr}\ C \times n + o((n\gamma)^2).$$

From Corollary 15 we have shown the following asymptotic behaviours:

- *Transient phase.* For $n \ll 1/(\gamma L)$, in a log-log plot $\mathbb{E}\left[\|\Omega_n\|^2\right]$ has a slope bigger than 1.
- *Stationary phase.* For $n \gg 1/(\gamma\mu)$, $\mathbb{E}\left[\|\Omega_n\|^2\right]$ is constant and therefore has a slope of 0 in a log-log plot.

This dichotomy naturally leads to a distance-based convergence diagnostic where the step size is decreased by a factor

$1/r$ when the slope becomes smaller than a certain threshold smaller than 2. The slope is computed between iterations of the form q^k and q^{k+1} for $q > 1$ and $k \geq k_0$. The method is formally described in Algorithm 4. We impose a burn-in time q^{k_0} in order to avoid unwanted and possibly harmful restarts during the very first iterations of the SGD recursion, it is typically worth ~ 8 ($q = 1.5$ and $k_0 = 5$) in all our experiments, see Section 6 and Appendix A.2. Furthermore note that from Proposition 5, saturation is reached at iteration $\Theta(\gamma^{-1})$. Therefore when the step-size is decreased as $\gamma \leftarrow r \times \gamma$ then the duration of the transience phase is increased by a factor $1/r$. This shows that it is sufficient to run the diagnostic every q^k where q is smaller than $1/r$.

Algorithm 4 Distance-based diagnostic

Input: $\theta_0, \theta_n, \theta_{n/q}, n, q > 1, k_0 \in \mathbb{N}^*$, thresh $\in (0, 2]$
Output: Diagnostic boolean
if $n = q^{k+1}$ for a $k \geq k_0$ in $\mathbb{N}^*$ **then**
  $S \leftarrow \frac{\log \|\theta_n - \theta_0\|^2 - \log \left\|\theta_{n/q} - \theta_0\right\|^2}{\log n - \log n/q}$
  Return: $\{S < \text{thresh}\}$
else
  Return: False
end if

6. Experiments

In this section, we illustrate our theoretical results with synthetic and real examples. We provide additional experiments in Appendix A.2.

Least-squares regression. We consider the objective $f(\theta) = \frac{1}{2}\mathbb{E}\left[(y_i - \langle x_i,\ \theta\rangle)^2\right]$. The inputs x_i are i.i.d. from $\mathcal{N}(0, H)$ where H has random eigenvectors and eigenvalues $(1/k)_{1\leq k\leq d}$. We note $R^2 = \operatorname{Tr}\ H$. The outputs y_i are generated following $y_i = \langle x_i,\ \theta^*\rangle + \varepsilon_i$ where $(\varepsilon_i)_{1\leq i\leq n}$ are i.i.d. from $\mathcal{N}(0, \sigma^2)$. We use averaged-SGD with constant step size $\gamma = 1/2R^2$ as a baseline since it enjoys the optimal statistical rate $O(\sigma^2 d/n)$ (Bach & Moulines, 2013).

Logistic regression setting. We consider the objective $f(\theta) = \mathbb{E}\left[\log(1 + e^{-y_i\langle x_i,\ \theta\rangle}\right]$. The inputs x_i are generated the same way as in the least-square setting. The outputs $y_i \in \{-1, 1\}$ are generated following the logistic probabilistic model. We use averaged-SGD with step-sizes $\gamma_n = 1/\sqrt{n}$ as a baseline since it enjoys the optimal rate $O(1/n)$ (Bach, 2014). We also compare to online-Newton (Bach & Moulines, 2013) which achieves better performance in practice.

ResNet18. We train an 18-layer ResNet model (He et al., 2016) on the CIFAR-10 dataset (Krizhevsky, 2009) using SGD with a momentum of 0.9, weight decay of 0.0001 and batch size of 128. To adapt the distance-based step-size

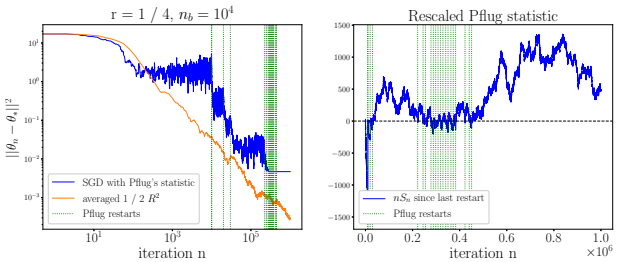

Figure 1. Least-squares on synthetic data. Left: least-squares regression. Right: Scaled Pflug's statistic nS_n. The dashed vertical lines correspond to Pflug's restarts. Note that only the left plot is in log-log scale.

statistic to this scenario, we use Pytorch's *ReduceLROnPlateau*() scheduler, created to detect saturation of arbitrary quantities. We use it to reduce the learning rate by a factor $r = 0.1$ when it detects that $\|\theta_n - \theta_{restart}\|^2$ has stopped increasing. The parameters of the scheduler are set to: patience $= 1000$, threshold $= 0.01$. Investigating if this choice of parameters is robust to different problems and architectures would be a fruitful avenue for future research. We compare our method to different step-size sequences where the step size is decreased by a factor $r = 0.1$ at various epoch milestones. Such sequences achieve state-of-the-art performances when the decay milestones are properly tuned. All initial step sizes are set to 0.1.

Inefficiency of Pflug's statistic. In order to test Pflug's diagnostic we consider the least-squares setting with $n =$ 1e6, $d = 20$, $\sigma^2 = 1$. Algorithm 3 is implemented with a conservative burn-in time of $n_b =$ 1e4 and Algorithm 1 with a discount factor $r = 1/4$. We note in Fig. 1 that the algorithm is restarted too often and abusively. This leads to small step sizes early on and to insignificant decrease of the loss afterward. The signal of Pflug's statistic is very noisy, and its sign gives no significant information on weather saturation has been reached or not. As a consequence the final step-size is very close to 0. We note that its behavior is alike the one of a random walk. On the contrary, averaged-SGD exhibits an $O(1/n)$ convergence rate. We provide further experiments on Pflug's statistic in Appendix A.1, showing its systematic failure for several values of the decay parameter r, the seed and the burn-in.

Efficiency of the distance-based diagnostic. In order to illustrate the benefit of the distance-based diagnostic, we performed extensive experiments in several settings, more precisely: (1) Least Squares regression on a synthetic dataset, (2) Logistic regression on both synthetic and real data, (3) Uniformly convex functions, (4) SVM, (5) Lasso. *In all these settings, without any tuning, we achieve the same performance as the best suited method for the problem.* These experiments are detailed in Appendix A.2. We hereafter

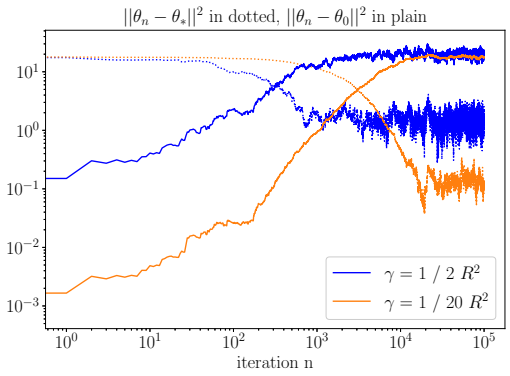

Figure 2. Logistic regression on synthetic dataset. $\|\theta_n - \theta^*\|^2$ (dotted) and $\|\theta_n - \theta_0\|^2$ (plain) for 2 different step sizes.

present results for Logistic Regression.

First, we consider the logistic regression setting with $n =$ 1e5, $d = 20$. In Fig. 2, we compare the behaviour of $\|\theta_n - \theta_0\|^2$ and $\|\theta_n - \theta^*\|^2$ for two different step sizes $1/2R^2$ and $1/20R^2$. We first note that these two quantities have the same general behavior: $\|\theta_n - \theta_0\|^2$ stops increasing when $\|\theta_n - \theta^*\|^2$ starts to saturate, and that this observation is consistent for the two step sizes. We additionally note that the average slope of $\|\theta_n - \theta_0\|^2$ is of value 2 during the transient phase and of value 0 when stationarity has been reached. This demonstrates that, even if this diagnostic is inspired by the quadratic case, the main conclusions of Corollary 15 still hold for convex non-quadratic function and the distance-based diagnostic in Algorithm 4 should be more generally valid. We also notice that the two oracle restart times are spaced by $\log(20/2) = 1$ which confirms that the transient phase lasts $\Theta(1/\gamma)$.

We further investigate the performance of the distance-based diagnostic on real-world datasets: the Covertype dataset and the MNIST dataset[1]. Each dataset is divided in two equal parts, one for training and one for testing. We then sample without replacement and perform a total of one pass over all the training samples. The loss is computed on the test set. This procedure is replicated 10 times and the results are averaged. For MNIST the task consists in classifying the parity of the labels which are $\{0, \dots, 9\}$. We compare our algorithm to: online-Newton ($\gamma = 1/10R^2$ for the Covertype dataset and $\gamma = 1/R^2$ for MNIST) and averaged-SGD with step sizes $\gamma_n = 1/2R^2\sqrt{n}$ (the value suggested by theory) and $\gamma_n = C/\sqrt{n}$ (where the parameter C is tuned to achieve the best testing error). In Fig. 3, we present the results. Top row corresponds to the Covertype dataset for two different values of the decrease coefficient $r = 1/2$ and $r = 1/4$,

[1]Covertype dataset available at archive.ics.uci.edu/ml/datasets/covertype and MNIST at yann.lecun.com/exdb/mnist.

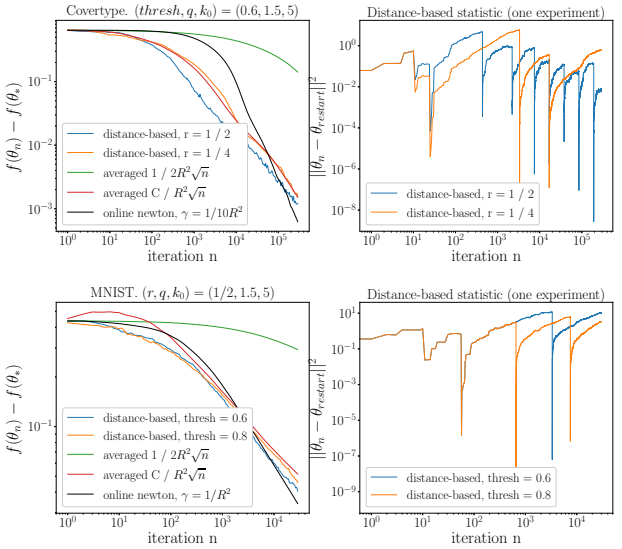

Figure 3. Top: Covertype dataset. Two different values of r are used: $1/2$, $1/4$. Bottom: MNIST dataset. Two different values of $thresh$ are used: 0.6, 0.8. Left: Logistic regression. Right: distance-based statistics $\|\theta_n - \theta_{restart}\|^2$.

the other parameters are set to $(tresh, q, k_0) = (0.6, 1.5, 5)$, left are shown the convergence rates for the different algorithms and parameters, right are plotted the evolution of the distance-based statistic $\|\theta_n - \theta_0\|^2$. Bottom row corresponds to the MNIST dataset for two different values of the threshold $thresh = 0.6$ and $thresh = 0.8$, the other parameters are set to $(r, q, k_0) = (1/2, 1.5, 5)$, left are shown the convergence rates for the different algorithms and parameters, right are plotted the evolution of the distance-based statistic $\|\theta_n - \theta_0\|^2$. The initial step size for our distance-based algorithm was set to $4/R^2$. Our adaptive algorithm obtains comparable performance as online-Newton and optimally-tuned averaged SGD, enjoying a convergence rate $O(1/n)$, and better performance than theoretically-tuned averaged-SGD. Moreover we note that the convergence of the distance-based algorithm is the fastest early stage. Thus this algorithm seems to benefit from the same exponential-forgetting of initial conditions as the oracle diagnostic (see Proposition 6). We point out that our algorithm is relatively independent of the choice of r and $thresh$. We also note (red and green curves) that the theoretically optimal step size is outperformed by the hand-tuned one with the same decay, which only confirms the need for adaptive methods. On the right is plotted the statistic during the SGD procedure. Unlike Pflug's one, the signal is very clean, which is mostly due to the large range of values that are taken.

Application to deep learning. We conclude by testing the distance-based statistic on a deep-learning problem in

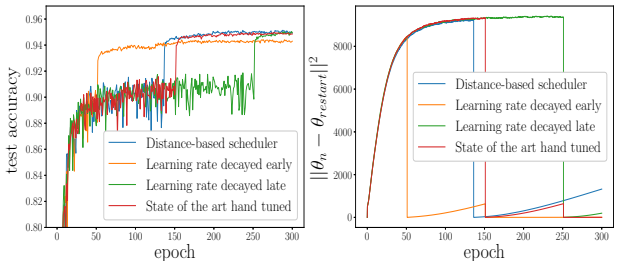

Figure 4. ResNet18 trained on Cifar10. Left: test accuracies. Right: distance-based statistic $\|\theta_n - \theta_{restart}\|^2$.

Fig. 4. In practice, the learning rate is decreased when the accuracy has stopped increasing for a certain number of epochs. In red is plotted the accuracy curve obtained when the learning rate is decreased by a factor $r = 0.1$ at epochs 150 and 250. These specific epochs have been manually tuned to obtain state of the art performance.

Looking at the red accuracy curve, it seems natural to decrease the learning rate earlier around epoch 50 when the test accuracy has stopped increasing. However doing so leads to a lower final accuracy (orange curve). On the other hand, decreasing the learning rate later, at epoch 250, leads to a good final accuracy but takes longer to reach it. If instead of paying attention to the test accuracy we focus on the metric $\|\theta_n - \theta_{restart}\|^2$ we notice that it still notably increases after epoch 50 and until epoch 150. This phenomenon manifests that this statistic contains information that cannot be simply obtained from the test accuracy curve. Hence when the ReduceLROnPlateau scheduler is implemented using the distance-based strategy, the learning rate is automatically decreased around epoch 140 and kept constant beyond (blue curve) which leads to a final state-of-the-art accuracy.

Therefore our distance-based statistic seems also to be a promising tool to adaptively set the step size for deep learning applications. We hope this will inspire further research.

Conclusion

In this paper we studied convergence-diagnostic step-sizes. We first showed that such step-sizes make sense in the smooth and strongly convex framework since they recover the optimal $O(1/n)$ rate with in addition an exponential decrease of the initial conditions. Two different convergence diagnostics are then analysed. First, we theoretically prove that Pflug's diagnostic leads to abusive restarts in the quadratic case. We then propose a novel diagnostic which relies on the distance of the final iterate to the restart point. We provide a simple restart criterion and theoretically motivate it in the quadratic case. The experimental results on synthetic and real world datasets show that our simple diagnostic leads to very satisfying convergence rates in a variety of frameworks.

An interesting future direction to our work would be to theoretically prove that our diagnostic leads to adequate restarts, as seen experimentally. It would also be interesting to explore more in depth the applications of our diagnostic in the non-convex framework.

Acknowledgements

The authors would like to thank the reviewers for useful suggestions as well as Jean-Baptiste Cordonnier for his help with the experiments.

References

Almeida, L. B., Langlois, T., Amaral, J. D., and Plakhov, A. *Parameter Adaptation in Stochastic Optimization*, pp. 111–134. Cambridge University Press, 1999.

Bach, F. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15:595–627, 2014.

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

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.

Benveniste, A., Priouret, P., and Métivier, M. *Adaptive Algorithms and Stochastic Approximations*. Springer-Verlag, 1990.

Bottou, L. Online algorithms and stochastic approximations. In Saad, D. (ed.), *Online Learning and Neural Networks*. Cambridge University Press, Cambridge, UK, 1998. URL http://leon.bottou.org/papers/bottou-98x. revised, oct 2012.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Chee, J. and Toulis, P. Convergence diagnostics for stochastic gradient descent with constant learning rate. In *International Conference on Artificial Intelligence and Statistics*, pp. 1476–1485, 2018.

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

Dieuleveut, A., Durmus, A., and Bach, F. Bridging the gap between constant step size stochastic gradient descent and markov chains. *arXiv preprint arXiv:1707.06386*, 2017.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(Jul):2121–2159, 2011.

Graves, A., Mohamed, A., and Hinton, G. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6645–6649, 2013.

Hazan, E. and Kale, S. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15: 2489–2512, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Jacobs, R. A. Increased rates of convergence through learning rate adaptation. *Neural Networks*, 1(4):295 – 307, 1988.

Kesten, H. Accelerated stochastic approximation. *Ann. Math. Statist.*, 29(1):41–59, 03 1958.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Kushner, H. J. and Huang, H. Asymptotic properties of stochastic approximations with constant coefficients. *SIAM Journal on Control and Optimization*, 19(1):87–105, 1981.

Kushner, H. J. and Yang, J. Analysis of adaptive step-size sa algorithms for parameter tracking. *IEEE Transactions on Automatic Control*, 40(8):1403–1410, 1995.

Lang, H., Xiao, L., and Zhang, P. Using statistics to automate stochastic optimization. In *Advances in Neural Information Processing Systems*, pp. 9536–9546, 2019.

Loshchilov, I. and Hutter, F. SGDR: stochastic gradient descent with restarts. *CoRR*, abs/1608.03983, 2016. URL http://arxiv.org/abs/1608.03983.

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.

Nemirovsky, A. S. and Yudin, D. B. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, 1983.

Nguyen, P., Nguyen, L., and van Dijk, M. Tight dimension independent lower bound on the expected convergence rate for diminishing step sizes in sgd. In *Advances in Neural Information Processing Systems*, pp. 3665–3674, 2019.

Pflug, G. C. On the determination of the step size in stochastic quasigradient methods. Technical report, IIASA Collaborative Paper, 1983.

Pflug, G. C. Stochastic minimization with constant stepsize: Asymptotic laws. *SIAM Journal on Control and Optimization*, 24(4):655–666, 1986.

Pflug, G. C. Adaptive stepsize control in stochastic approximation algorithms. *IFAC Proceedings Volumes*, 21(9): 787–792, 1988a.

Pflug, G. C. Stepsize rules, stopping times and their implementation in stochastic quasi-gradient algorithms. *numerical techniques for stochastic optimization*, pp. 353–372, 1988b.

Polyak, B. T. and Juditsky, A. B. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the Conference on Machine Learning (ICML)*, 2012.

Robbins, H. and Monro, S. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Schaul, T., Zhang, S., and LeCun, Y. No more pesky learning rates. In *International Conference on Machine Learning*, pp. 343–351, 2013.

Schraudolph, N. N. Local gain adaptation in stochastic gradient descent. In *In Proc. Intl. Conf. Artificial Neural Networks*, pp. 569–574, 1999.

Smith, L. N. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472. IEEE, 2017.

Sordello, M. and Su, W. Robust learning rate selection for stochastic optimization via splitting diagnostic. *arXiv preprint arXiv:1910.08597*, 2019.

Sutton, R. Adaptation of learning rate parameters. In *In: Goal Seeking Components for Adaptive Intelligence: An Initial Assessment, by A. G. Barto and R. S. Sutton.* Air Force Wright Aeronautical Laboratories Technical Report AFWAL-TR-81-1070. Wright-Patterson Air Force Base, Ohio 45433., 1981.

Sutton, R. S. Adapting bias by gradient descent: An incremental version of delta-bar-delta. In *AAAI*, 1992.

Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., and Recht, B. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, pp. 4148–4158, 2017.

Yaida, S. Fluctuation-dissipation relations for stochastic gradient descent. *arXiv e-prints*, art. arXiv:1810.00004, Sep 2018.

Zeiler, M. D. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701, 2012. URL http://arxiv.org/abs/1212.5701.

Zhang, T. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, pp. 116, 2004.