# Differential Equations for Modeling Asynchronous Algorithms

**Li He**[1,2*], **Qi Meng**[3*], **Wei Chen**[4], **Zhi-Ming Ma**[1,2] and **Tie-Yan Liu**[4]
[1] University of Chinese Academy of Sciences
[2] Academy of Mathematics and Systems Science, Chinese Academy of Sciences
[3] Peking University
[4] Microsoft Research
heli@amss.ac.cn, qimeng13@pku.edu.cn, {wche, tie-yan.liu}@microsoft.com, mazm@amt.ac.cn

## Abstract

Asynchronous stochastic gradient descent (ASGD) is a popular parallel optimization algorithm in machine learning. Most theoretical analysis on ASGD take a discrete view and prove upper bounds for their convergence rates. However, the discrete view has its intrinsic limitations: there is no characterization of the optimization path and the proof techniques are induction-based and thus usually complicated. Inspired by the recent successful adoptions of stochastic differential equations (SDE) to the theoretical analysis of SGD, in this paper, we study the continuous approximation of ASGD by using stochastic differential delay equations (SDDE). We introduce the approximation method and study the approximation error. Then we conduct theoretical analysis on the convergence rates of ASGD algorithm based on the continuous approximation. There are two methods: *moment estimation* and *energy function minimization* can be used to analyze the convergence rates. Moment estimation depends on the specific form of the loss function, while energy function minimization only leverages the convex property of the loss function, and does not depend on its specific form. In addition to the convergence analysis, the continuous view also helps us derive better convergence rates. All of this clearly shows the advantage of taking the continuous view in gradient descent algorithms.

## 1 Introduction

Asynchronous stochastic gradient descent (ASGD) is a popular parallel optimization algorithm in machine learning [Langford *et al.*, 2009; Agarwal and Duchi, 2011; Recht *et al.*, 2011; Lian *et al.*, 2015; Zhang *et al.*, 2015]. It has been broadly used in solving deep neural network and received many successes in practice recently, which significantly reduce the communication overhead by avoiding idleness. The main issue with asynchronous algorithms lies on using delayed stochastic gradient information. Suppose $\{(a_1,b_1),\cdots,(a_i,b_i),\cdots,(a_n,b_n)\}$ is the training data set, where the input vector $a_i \in \mathbb{R}^d$ and the output $b_i \in \mathbb{R}$. Supervised machine learning algorithms aim to minimize the empirical risk, i.e.,

$$\min_x F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{1}$$

where $f_i(x) = l(x; a_i, b_i)$ is the loss of model $x$ for the training instance $(a_i, b_i)$. ASGD uses multiple threads or local workers to calculate the gradient of the loss function using a mini-batch of training instances. Then the local worker push the gradient to a master and pull the latest parameter from the master. The master uses the received gradient to update the parameter $x$. Because there is no synchronization between the local workers, the gradient which the master received may be a delayed information for the parameter [Agarwal and Duchi, 2011; Recht *et al.*, 2011; Lian *et al.*, 2015]. The update rule for ASGD at iteration $k$ can be described as:

$$x_{k+1} = x_k - \eta_k g_{M_k}(b, x_{k-l_k}), \tag{2}$$

where $\eta_k$ is the learning rate, $g_{M_k}(b, x_{k-l_k})$ is the stochastic gradient calculated using the minibatch $M_k$ with minibatch size $b$, $l_k \geq 0$ denotes the stochastic delay. In general, it assumes that $l_k, \forall k$ or $\mathbb{E}l_k$ is upper bounded [Agarwal and Duchi, 2011; Lian *et al.*, 2015].

The delayed information will influence the convergence rate of the optimization algorithms. Theoretical analyses have been conducted on ASGD for various problem settings, mostly from a discrete view, in which the convergence rate is proved by induction for the sequence of iterative of the optimization algorithm. Then they compare it with the convergence rate of sequential stochastic gradient descent to get the speedup condition. For example, [Recht *et al.*, 2011] shows that for convex problem, if the training data is sparse enough, ASGD can achieve linear speedup. [Lian *et al.*, 2015] shows that for nonconvex problems, if the delay is upper bounded by the stochastic sampling variance, ASGD can achieve linear speedup.

However, the discrete view has its intrinsic limitations. (1) It cannot provide an explicit characterization of the optimization path, thus lacks insights about the optimization process. (2) The proofs of optimization algorithms are usually induction-based and somehow unavoidably complicated.

In recent years, researchers study the dynamics of optimization algorithms by taking a continuous view. They

[*] This work was done when the authors were visiting Microsoft Research Asia.

derive the corresponding differential equations of the existing discrete-time optimization methods by regarding the optimization rule as the numerical solution of a differential equation. Many works [Raginsky and Bouvrie, 2012; Wibisono *et al.*, 2016; Su *et al.*, 2014; Li *et al.*, 2016; Yang *et al.*, 2017; Mandt *et al.*, 2016; Krichene *et al.*, 2015] study sequential optimization algorithms using continuous view from various aspects. For example, [Raginsky and Bouvrie, 2012] used the continuous time analysis to study mirror descent algorithm. In [Wibisono *et al.*, 2016] and [Su *et al.*, 2014], some accelerated methods were studied by the continuous techniques such as second-order ordinary differential equations. In [Li *et al.*, 2017], the authors developed the method of stochastic modified equations (SME), in which stochastic gradient algorithms are approximated in the weak sense by continuous-time SDE. These works provide a clearer understanding of the dynamic optimization process than the previous works that only take the discrete view, thus have the potential in addressing the aforementioned limitations of the discrete view. However, most of the above works are focusing on sequential algorithms, and using continuous techniques to study asynchronous algorithms has not been studied yet.

Inspired by these works, in this paper, we study the continuous approximation of asynchronous stochastic gradient descent. First, we propose a procedure to approximate asynchronous stochastic gradient based algorithms by using stochastic differential delay equations. Then we analyze the approximation error and show that the approximation error is related to the number of iteration $K$ and mini-batch size $b$.

Second, within this approximation, we study the convergence rates by using continuous techniques and show the following results. (1) For the linear regression problem which is solved by SGD with constant learning rate, we can attain full information of the optimization path, including the first and second-order moments. (2) For SDDEs, although it is hard to obtain full information, we can still analyze the optimization path by using *moment estimation* and *energy function minimization*. Moment estimation depends on the specific form of the loss function but has nothing to do with its convexity property, whereas energy function minimization leverages the convexity property but does not depend on its specific form. (3) By using these two techniques, in addition to characterizing the optimization path, we also get the convergence rates of optimization algorithms, with a much simpler proof than that from discrete view. Specifically, we prove a tighter convergence rate for ASGD than the other existing results [Zheng *et al.*, 2017; Recht *et al.*, 2011].

All of these results clearly demonstrate the advantages of taking the continuous view in analyzing ASGD algorithm.

## 2 Backgrounds

In this section, we briefly review the basic settings of asynchronous stochastic gradient descent algorithm, and introduce the stochastic differential delay equations.

### 2.1 Basic Settings of ASGD

**Asynchronous Stochastic Gradient Descent** is an efficient parallel algorithm. Local workers or threads do not need to wait for others to do synchronization, thus the model is updated faster compared with the synchronous SGD. ASGD updates the parameter using a delayed gradient and has been proved achieving linear or sub-linear speedup under certain conditions such as the training data are sparse or the delay can be upper bounded [Agarwal and Duchi, 2011; Recht *et al.*, 2011; Lian *et al.*, 2015]. In this paper, we consider the update rule of ASGD under the consistent read setting [Lian *et al.*, 2015]. Let $M_k$ be the index of a mini-batch with size $b$ and $\eta_k$ be the learning rate for $k$-th iteration. We denote $g_{M_k}(b, x_k) = \frac{1}{b}\sum_{i\in M_k}\nabla f_i(x_k)$ as the averaged gradient calculated by a mini-batch. The update rule of ASGD is

$$x_{k+1} = x_k - \eta_k g_{M_k}(b, x_{k-l_k}), \tag{3}$$

where $l_k$ is a random delay satisfying $0 \le l_k \le l$.[1] Sequential SGD is a special case of ASGD with $l_k = 0, \forall k$.

### 2.2 Stochastic Differential Delay Equation

Stochastic differential delay equation (SDDE) for a $d$-dimensional continuous process $X = (X(t))_{t\ge 0}$ is a natural generalization of stochastic differential equation (SDE) by allowing the coefficients depending on values in the past, which has been studied by researchers [Mao, 2007; Bao *et al.*, 2016]. We consider the following form:

$$dX(t) = g_1(t, X_t)dt + g_2(t, X_t)dB(t),\ t \in [t_0, T],$$

where $X_t = \{X(t-\theta(t)) : 0 \le \theta(t) \le \tau\}$ is the segment (or the functional solution), $\theta(t)$ is a random variable describing the random delay and $\tau$ is its upper bound. The symbol $B(t)$ denotes $r$-dimensional Brownian motion [Durrett, 2010]. Denote by $C([-\tau, 0]; \mathbb{R}^d)$ the family of continuous functions $\xi$ from $[-\tau, 0]$ to $\mathbb{R}^d$ with the norm $\|\xi\| = sup_{s\in[-\tau,0]}\|\xi(s)\|$. Then $X_t$ is regarded as a $C([-\tau, 0]; \mathbb{R}^d)$-valued stochastic process. The vector $g_1 \in \mathbb{R}^d$ and the matrix $g_2 \in \mathbb{R}^d \times \mathbb{R}^r$ are appropriate functions. For the sake of simplicity, we consider the case of $d = r$ in this paper. SDE can be regarded as a special case of SDDE with $\theta(t) = 0, \forall t$.

## 3 Continuous Approximation of Asynchronous Optimization Methods

In this section, we first propose the continuous approximation of ASGD by using SDDE and describe the approximation procedure in detail. Then we analyze the approximation error for the continuous approximation of ASGD.

First, we introduce some notations. We relate the discrete iterations with continuous stochastic process by introducing the ansatz $x_k \approx X(k\delta)$ where $X(t) : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^d$ is a continuous function of $t$. Given $t \in [0, T]$, we have $T = \delta K$, where $K$ is the total number of discrete iterations and $\delta$ is the interval to do discretization of $X(t)$ which we call the time step. The discrete index of iterations and continuous index of time have the relation: $X(t+\delta) \approx x_{(t+\delta)/\delta} := x_{k+1}$, $X(t) \approx x_{t/\delta} := x_k$. We define the covariance matrix of $\nabla f_i(x_k)$ as

$$\Sigma(x_k) := \mathbb{E}(\nabla f_i(x_k) - \mathbb{E}\nabla f_i(x_k))(\nabla f_i(x_k) - \mathbb{E}\nabla f_i(x_k))^T.$$

Since $g_{M_k}(b, x_k)$ is a sum of $b$ i.i.d random vectors $\nabla f_i(x_k)$, then the covariance matrix of $g_{M_k}(b, x_k)$ is $\frac{\Sigma(x_k)}{b}$ [Balles *et al.*, 2017; Li *et al.*, 2017].

[1] ASGD under inconsistent read setting will be studied as the future work.

### 3.1 Asynchronous Gradient Methods

Most existing approximations are built for sequential algorithms [Li *et al.*, 2017; Mandt *et al.*, 2016], while asynchronous gradient methods are widely used due to its efficiency in utilizing multiple computational nodes [Recht *et al.*, 2011; Agarwal and Duchi, 2011; Dean *et al.*, 2012; Zhang *et al.*, 2015; Zheng *et al.*, 2017]. Therefore, in this section, we describe how to approximate the update rule of ASGD using SDDEs.

Consider the following asynchronous gradient methods

$$x_{k+1} = x_k - \eta u_k g_{M_k}(b, x_{k-l_k}), \tag{4}$$

where $\eta$ is a deterministic learning rate and $u_k \in [0,1]$ is an adjustment function of $k$. The symbol $l_k$ denotes the random delay and $g_{M_k}(b, x_{k-l_k})$ denotes the delayed information such as delayed gradient.

We first give the following proposition (Theorem 3.9.6, [Durrett, 2010], [Balles *et al.*, 2017]), on which our transformation and approximation error analysis are based.

**Proposition 3.1** *The $\nabla F(x_{k-l_k}) - g_{M_k}(b, x_{k-l_k})$ converges weakly to a multivariate normal distribution with mean $\vec{0}$ and covariance matrix $\Sigma(x_{k-l_k})/b$.*

Moreover, we assume positive definiteness and a square root decomposition of the covariance matrix of $\Sigma(x_{k-l_k})/b$, i.e., $\frac{\Sigma(x_{k-l_k})}{b} = \sigma(x_{k-l_k})\sigma(x_{k-l_k})^T$. Reformulating Eq.(4) as:

$$\begin{aligned} x_{k+1} =& x_k - \eta u_k \nabla F(x_{k-l_k}) \\ &+ \eta u_k(\nabla F(x_{k-l_k}) - g_{M_k}(b, x_{k-l_k})). \end{aligned}$$

Let $z_k$s be i.i.d random vectors which follow standard normal distribution $N(\vec{0}, I_{d\times d})$. Thus, we can use $\sigma(x_{k-l_k})z_k$ to approximate the noise term (by Central Limit Theorem). Therefore, the above equation can be approximated by

$$x_{k+1} = x_k - \eta u_k \nabla F(x_{k-l_k}) + \eta u_k \sigma(x_{k-l_k})z_k. \tag{5}$$

Choosing a precision $\delta$, we can view the Eq.(5) as an Euler-Maruyama approximation of the following SDDE [Mao, 2007]:

$$\begin{aligned} \mathrm{d}X(t) =& -\frac{\eta}{\delta}U(t)\nabla F(X(t-\theta(t)))\mathrm{d}t \\ &+ \frac{\eta}{\delta}\sqrt{\delta}U(t)\sigma(X(t-\theta(t)))\mathrm{d}B(t), \end{aligned} \tag{6}$$

where $\theta(t) = l_k\delta$. We assume that $0 \le l_k \le l$ and denote $\tau := l\delta$. In the following, we give several possible choices for the time step under different cases. (a) For ASGD with constant learning rate $\eta = \eta_0$ and $u_k = 1$, we choose the $\delta = \eta_0$. (b) For decreasing learning rate $\eta = \eta_0$ and $u_k = \frac{1}{k}$, we have $U(t) = \frac{\delta}{t}$. We can choose any precision. (c) For $\eta = \eta_0$ and $u_k = \frac{1}{\sqrt{k}}$, since $U(t) = \frac{\sqrt{\delta}}{\sqrt{t}}$, we choose $\delta = \eta_0$.

In summary, we used SDDE to approximate ASGD through Eq.(4), Eq.(5) and Eq.(6). Many asynchronous gradient methods can be represented by Eq.(4), such as ASGD, asynchronous SVRG, etc.

Using above similar technique, Nesterov's accelerated and momentum stochastic gradient methods can be transformed into two stage SDEs. Besides, the continuous approximation can be described by a second-order SDE. We omit the explicit transform because it is not the main case which will be studied in this paper. For more details, it can be referred in [Su *et al.*, 2014; Li *et al.*, 2017].

### 3.2 Approximation Error

In this subsection, we study the approximation error between $X(t_k)$, produced by the continuous process and $x_k$ which is produced by the discrete process. We applied Central Limit Theorem and Euler-Maruyama approximation of stochastic delay differential equation in analyzing approximation error. Now we give the following Theorem 3.2-3.4 and proof sketches, which show the approximation error between the update produced by ASGD (Eq.(4)) and its corresponding S-DDE (Eq.(6)) in different cases.

**Theorem 3.2** *Assume that $\nabla F(x)$ is $L$-Lipschitz continuous, and $b$ is the mini-batch size. Consider delay $l_k$ with upper bound $l$ and $u_k = 1$ for $k = 1, \ldots, K$. If we consider ASGD with $\delta = \eta$, $\eta = \frac{1}{K}$ and $\eta L(l+1) \le 1$, we have*

$$\mathbb{E}\|X(t_K) - x_K\| \le C_1(\frac{1}{\sqrt{K}} + \frac{l}{K^{3/2}}) + C_2\frac{e^{L(l+1)}}{\sqrt{b}(l+1)},$$

*where $C_1, C_2$ are constants.*

*Proof sketch:* (1) For SDDE with $U(t) = 1$, the Euler-Maruyama scheme with $\delta = \eta$ is

$$\tilde{x}_{k+1} = \tilde{x}_k - \eta\nabla F(\tilde{x}_{k-l_k}) + \sqrt{\eta}(B(t_{k+1}) - B(t_k))\sigma(\tilde{x}_{k-l_k}),$$

where $B(t_{k+1}) - B(t_k)$ is usually modeled by $\sqrt{\eta}z_k$. It is well-known that the Euler-Maruyama scheme is strongly convergent with order $\frac{1}{2}$. According to (Theorem 5.5.5 [Mao, 2007]), we have $\mathbb{E}(\|X(t_k) - \tilde{x}_k\|) \le \tilde{C}_1\eta^{3/2}(k+l)e^{2(\eta k)^2}$. Let $\eta = \frac{1}{K}$, we can obtain

$$\mathbb{E}(\|X(t_K) - \tilde{x}_K\|) \le C_1(\frac{1}{\sqrt{K}} + \frac{l}{K^{3/2}}).$$

(2) Next, let us consider the relationship between $x_{k+1}$ and $\tilde{x}_{k+1}$. By the central limit theorem (CLT), we can get that $\nabla F(x_{k-l_k}) - g_{M_k}(b, x_{k-l_k})$ converges to $\sigma(x_{k-l_k})z_k$ in distribution with the increasing minibatch size $b$. Assume that $\nabla F(x_k)$ is $L$-Lipschitz continuous and using Theorem 3.4.9 in [Durrett, 2010], the gap between their distribution functions goes to zero at rate $b^{-\frac{1}{2}}$. Since ASGD uses a delayed gradient, it will cause the mismatch between $x_k$ and $g_{M_k}(b, x_{k-l_k})$ when we expand the series. We have

$$\begin{aligned} &\mathbb{E}\|x_{k+1} - \tilde{x}_{k+1}\| \\ =&\mathbb{E}\|x_k - \tilde{x}_k - \eta(\nabla F(x_{k-l_k}) - \nabla F(\tilde{x}_{k-l_k})) \\ &- \eta[\sigma(\tilde{x}_{k-l_k})z_k - \nabla F(x_{k-l_k}) + g_{M_k}(b, x_{k-l_k})]\| \\ \le&\mathbb{E}\|x_k - \tilde{x}_k - \eta(\nabla F(x_{k-l_k}) - \nabla F(\tilde{x}_{k-l_k}))\| \\ &+ \mathbb{E}\|\eta[\sigma(\tilde{x}_{k-l_k})z_k - \nabla F(x_{k-l_k}) + g_{M_k}(b, x_{k-l_k})]\|. \end{aligned}$$

Denote $\mathbb{E}\|\eta[\sigma(\tilde{x}_{k-l_k})z_k - \nabla F(x_{k-l_k}) + g_{M_k}(b, x_{k-l_k})]\|$ as $\Phi(k)$, then we have

$$\begin{aligned} \Phi(k) \le&\eta(\mathbb{E}\|\sigma(\tilde{x}_{k-l_k}))z_k\|^2)^{\frac{1}{2}} \\ &+ \eta(\mathbb{E}\|g_{M_k}(b, x_{k-l_k}) - \nabla F(x_{k-l_k})\|^2)^{\frac{1}{2}} \\ =&\eta\sqrt{\frac{tr(\Sigma(\tilde{x}_{k-l_k}))}{b}} + \eta\sqrt{\frac{tr(\Sigma(x_{k-l_k}))}{b}}, \end{aligned}$$

where $tr(\Sigma)$ denotes the trace of the matrix and we derived the last step by simple matrix computations.

Denote $\frac{\eta\Phi}{\sqrt{b}} = sup_{s=0}^{k}\Phi(s)$ as an upper bound. Let $A_k = \mathbb{E}\|x_k - \tilde{x}_k\|$ and $A_0 = 0$. We consider a probability $q_j$ is assumed for each $l_k = j,\ j = 0, 1, \cdots, l$. [2] Taking expectation and using upper bound, it becomes

$$\begin{aligned} A_k &\le A_{k-1} + \eta L \sum_{j=1}^{l+1} q_j A_{k-j} + \frac{\eta\Phi}{\sqrt{b}} \\ &\le A_{k-1} + \eta L \sum_{j=1}^{l+1} A_{k-j} + \frac{\eta\Phi}{\sqrt{b}}. \end{aligned}$$

Thus we have $A_k \le \eta L \sum_{j=1}^{l+1}(A_{1-j} + \cdots + A_{k-j}) + k\frac{\eta\Phi}{\sqrt{b}} \le \eta L(l+1)(A_0 + A_1 + \cdots + A_{k-1}) + k\frac{\eta\Phi}{\sqrt{b}}$, where $A_0 = 0, A_1 = \eta L(l+1)A_0 + \frac{\eta\Phi}{\sqrt{b}}$. Using this recursion, we obtain

$$\begin{aligned} A_k &\le k\frac{\eta\Phi}{\sqrt{b}} + \eta L(l+1)\frac{\eta\Phi}{\sqrt{b}} \sum_{j=0}^{k-2}(1 + \eta L(l+1))^j (k - j - 1) \\ &\le k\frac{\eta\Phi}{\sqrt{b}} + \eta L(l+1)\frac{\eta\Phi}{\sqrt{b}} \frac{(1 + \eta L(l+1))^k}{(\ln(1 + \eta L(l+1)))^2} \\ &\le k\frac{\eta\Phi}{\sqrt{b}} + \eta L(l+1)\frac{\eta\Phi}{\sqrt{b}} \frac{e^{\eta L k(l+1)}}{(\ln(1 + \eta L(l+1)))^2} \\ &\le k\frac{\eta\Phi}{\sqrt{b}} + \frac{4\frac{\eta\Phi}{\sqrt{b}} e^{\eta L k(l+1)}}{\eta L(l+1)}, \end{aligned}$$

where we used integration approximation. In the last step, we assumed that $\eta L(l+1) \le 1$ and used $\frac{\eta L(l+1)}{2} \le \eta L(l+1) - \frac{(\eta L(l+1))^2}{2} \le \ln(1 + \eta L(l+1))$. Let $\eta = \frac{1}{K}$, we have

$$\mathbb{E}\|X(t_K) - x_K\| \le C_1(\frac{1}{\sqrt{K}} + \frac{l}{K^{3/2}}) + C_2\frac{e^{L(l+1)}}{\sqrt{b}(l+1)},$$

where $C_1, C_2$ are constants. □

If we consider the upper bound $l = 0$, ASGD reduces to SGD. We give the following corollary.

**Corollary 3.3** *Assume that $\nabla F(x)$ is L-Lipschitz continuous, and $b$ is the mini-batch size. Consider delay $l_k$ with upper bound $l = 0$ and $u_k = 1$ for $k = 1, \ldots, K$. The SGD with $\delta = \eta$, $\eta = \frac{1}{K}$ and $\eta L \le 1$, we have*

$$\mathbb{E}\|X(t_K) - x_K\| \le C_1\frac{1}{\sqrt{K}} + C_2\frac{e^L}{\sqrt{b}},$$

*where $C_1, C_2$ are constants.*

The above theorem just assumes that the delay can be upper bounded. If we further assume that the randomness of the delay can be neglected, for example, a cyclic delayed update architecture [Agarwal and Duchi, 2011], the results can be improved. In this case, we give the following theorem.

**Theorem 3.4** *Assume that $\nabla F(x)$ is L-Lipschitz continuous, and $b$ is the mini-batch size. Consider delay $l_k = l$ and $u_k = 1$ for $k = 1, \ldots, K$. If we consider ASGD with $\delta = \eta$, $\eta = \frac{1}{K}$ and $\eta Ll \le 1$, we have*

$$\mathbb{E}\|X(t_K) - x_K\| \le C_3(\frac{1}{\sqrt{K}} + \frac{l}{K^{3/2}}) + C_4\frac{e^L}{\sqrt{b}}.$$

*where $C_3, C_4$ are constants.*

*Proof sketch:* The analysis between $X(t_k)$ and $\tilde{x}_k$ follows from the proof of Theorem 3.2. We analyze the relationship between $x_k$ and $\tilde{x}_k$. Using the above inequality and $A_0 = 0$, we can get

$$A_{kl} \le \eta L \sum_{i=1}^{(k-1)l} A_i + \frac{\eta\Phi}{\sqrt{b}} \cdot kl.$$

Furthermore, we have the following recursion relation:

$$(A_{(k-1)l} + \cdots + A_{(k-2)l+1}) \le \eta Ll \sum_{i=1}^{(k-2)l} A_i + \frac{\eta\Phi}{\sqrt{b}} \cdot (k-1)l^2.$$

Using that $A_l + \cdots + A_1 \le \frac{2(1+\eta L)^l}{\ln(1+\eta L)}$, we conclude that

$$\begin{aligned} A_{kl} \le& \eta L(1 + \eta Ll)^{k-1} \cdot \frac{2(1+\eta L)^l \cdot \frac{\eta\Phi}{\sqrt{b}}}{\ln(1+\eta L)} \\ &+ \frac{\eta^2 l^2 L\Phi}{\sqrt{b}} \sum_{i=0}^{k-1}(1 + \eta Ll)^i (k - 1 - i) \\ \le& \eta L(1 + \eta Ll)^{k-1} \cdot \frac{2(1+\eta L)^l \cdot \frac{\eta\Phi}{\sqrt{b}}}{\ln(1+\eta L)} + \frac{\eta^2 l^2 L\Phi(1+\eta Ll)^k}{(\ln(1+\eta Ll))^2\sqrt{b}} \\ \le& \frac{2\eta L e^{\eta Llk} \cdot \frac{\eta\Phi}{\sqrt{b}}}{\ln(1+\eta L)} + \frac{\eta^2 l^2 L\Phi e^{\eta Llk}}{(\ln(1+\eta Ll))^2\sqrt{b}}. \end{aligned}$$

Assume that $\eta Ll \le 1$ and using $\frac{\eta Ll}{2} \le \ln(1+\eta Ll)$, we obtain

$$A_{kl} \le 4e^{\eta Llk} \cdot \frac{\eta\Phi}{\sqrt{b}} + \frac{4\Phi e^{\eta Llk}}{L\sqrt{b}} \le C_4\frac{e^{\eta Llk}}{\sqrt{b}}.$$

Let $\eta = \frac{1}{K}$, now we have the following bound

$$\mathbb{E}\|X(t_K) - x_K\| \le C_3(\frac{1}{\sqrt{K}} + \frac{l}{K^{3/2}}) + C_4\frac{e^L}{\sqrt{b}},$$

where $C_3, C_4$ are constants. □

The Lipschitzs coefficient is relatively a small constant. For example, we have $L \le 1$ and $L \le 0.25$ for linear regression and logistic regression, respectively. Theorem 3.2 shows that the approximation error of ASGD will be small if the number of iterations $K$ and the minibatch size $b$ are large. The delay $l$ has influence on the two parts. This is consistent with the intuition that delay will cause mismatch of the updates. From Theorem 3.4, when the delay equals to a constant, its approximation error is similar to that of SGD since $l$ does not have influence on the part of $\frac{e^L}{\sqrt{b}}$.

# 4 Convergence Analysis: Techniques and Examples

In this section, we show some techniques using SDDE for convergence analysis. Firstly, we introduce the measure for convergence analysis. Taking SDE as a simple example, for fixed $t$, the SDE becomes an ODE after taking expectation of $X(t)$. We can get the optimum $x^*$ by taking limits $\lim_{t\to\infty}\mathbb{E}X(t)$. Secondly, we can calculate $\mathbb{E}(X(t) - x^*)$ and $\mathbb{E}||X(t) - \mathbb{E}X(t)||^2$ for each fixed $t$. Combining them we can get the convergence of $\mathbb{E}||X(t) - x^*||^2$.

Usually, SDEs can be classified into two cases: analytic solution case and non-analytic solution case. An example of SDE with analytic solution is linear regression solved

[2] It should be noted that for $k \le l$, the random delay $l_k$ only take values from the set $\{0, 1, \cdots, k\}$.

by SGD with constant learning rate. The objective function is $F(x) = \frac{1}{2n}\|Ax - B\|^2$, with $A = (a_1^T, \ldots, a_n^T)^T$ and $B = (b_1, \ldots, b_n)^T$, $x \in \mathbb{R}^d, a_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$. Denote $\tilde{A} = \frac{1}{n}A^TA$. The solution of its SDE is an *Ornstein-Uhlenbeck* process [Uhlenbeck and Ornstein, 1930]:

$$X(t) = x^* + e^{-\tilde{A}t}(X(0) - x^*) + \int_0^t e^{\tilde{A}(s-t)}\sqrt{\eta}G\mathrm{d}B(s).$$

We can calculate the moments directly, i.e.,

$$\mathbb{E}[X(t) - x^*|X(0) = x_0] = e^{-\tilde{A}t}(x_0 - x^*),$$
$$\mathbb{E}[\|X(t) - \mathbb{E}X(t)\|^2|X(0) = x_0] \le \int_0^t \|e^{\tilde{A}(s-t)}\sqrt{\eta}G\|^2\mathrm{d}s.$$

From the above results, we can get many insights about the dynamics of SGD, such as the oscillation of the sample path which has been discussed in [Li *et al.*, 2017].

However, SDDEs are more difficult to obtain analytic solution. For non-analytic solution case, it is hard to calculate the exact moments of $X(t) - x^*$, we can use two methods to estimate them to achieve the convergence for ASGD: *moment estimation* and *energy function minimization*.

## 4.1 Moment Estimation

We show the moment estimation technique by taking linear regression as an example. It can also be used for nonconvex loss functions. Before showing the details, we need to guarantee the existence and uniqueness of the solutions for SDDEs. Uniform Lipschitz condition and linear growth condition are two conditions to guarantee the existence and uniqueness of the solutions for SDDE [Mao, 2007]. We should see that these two conditions have no direct relation with convexity. For our continuous approximation, we do not care the explicit form of $\sigma(X(t - \theta(t)))$ (except for variance reduced techniques are involved such as SVRG [Johnson and Zhang, 2013]) and just assume that it has an upper bound thus it is a constant about $X(t - \theta(t))$. In this case, the diffusion term satisfies linear growth and Lipschitz conditions. We just need to check whether $-\nabla F(X(t - \theta(t)))$ satisfies the two conditions.

We use Theorem 5.2.2 in [Mao, 2007] to guarantee the existence and uniqueness of the solution to SDDE. We consider the ASGD with constant learning rate $\eta$ for the linear regression. The corresponding SDDE is

$$\begin{aligned}&\mathrm{d}X(t) = -(\tilde{A}X(t - \theta(t)) - \tilde{B})\mathrm{d}t + \sqrt{\eta}\sigma(X(t - \theta(t)))\mathrm{d}B(t),\\ &X(s) = \xi(s), s \in [-\tau, 0],\end{aligned} \tag{7}$$

where $\theta(t) \in [0, \tau]$. Here we consider $\theta(t) = \tau$ and $\tau$ is its upper bound. Following the techniques in [Bao *et al.*, 2016; 2014], we can bound the first and second moments and give the following theorem.

**Theorem 4.1** *Define* $V = sup\{Re(\beta) : \beta \in \mathbb{C}, det(\beta I_{d\times d} + \tilde{A}e^{-\beta\tau}) = 0\}$. *For any given training error* $\epsilon$, *let* $\eta = \frac{\epsilon}{\tau^2}$. *If* $V < 0$, *then for* $\lambda \in (0, -V)$, *the convergence rate for ASGD for linear regression Eq.(7) by using moment estimation is*

$$\mathbb{E}\|X(t) - x^*\|^2 \le C_5e^{-2\lambda(t-\tau)} + C_6\frac{\epsilon}{\lambda\tau^2},$$

*where* $C_5$ *and* $C_6$ *are constants.* [3]

*Proof sketch:* Through some lengthy derivation [Bao *et al.*, 2016], we have $\|\mathbb{E}(X(t)) - x^*\| \le a_1e^{-\lambda t}$ and $\mathbb{E}\|X(t) - \mathbb{E}X(t)\|^2 \le a_2(1 - e^{-2\lambda t})$ where the coefficient $a_1$ is $c_\lambda\|\tilde{\xi}\| + c_\lambda\|\tilde{\xi}\|\|\tilde{A}\|(e^{\lambda\tau} - 1)\frac{1}{\lambda}$, $a_2$ is $\frac{c_\lambda^2\mathbb{E}\|\sqrt{\eta}\sigma\|^2}{2\lambda}$ and $x^* = \tilde{A}^{-1}\tilde{B}$. Moreover, we can obtain

$$\mathbb{E}\|X(t) - x^*\|^2 = \mathbb{E}\|X(t) - \mathbb{E}X(t)\|^2 + \|\mathbb{E}X(t) - x^*\|^2.$$

Thus by putting $a_1$ and $a_2$ in it , we have

$$\mathbb{E}\|X(t) - x^*\|^2 \le$$
$$\frac{c_\lambda^2}{\lambda^2}\left[(\|\tilde{\xi}\|^2(\|\tilde{A}\| + \lambda)^2 + \lambda\eta\mathbb{E}||\sigma||^2)e^{2\lambda\tau}\right]e^{-2\lambda t} + \frac{c_\lambda^2\mathbb{E}\|\sqrt{\eta}\sigma\|^2}{2\lambda}.$$

Putting $\eta = \frac{\epsilon}{\tau^2}$ in the above equation, we can get

$$\begin{aligned}&\mathbb{E}\|X(t) - x^*\|^2\\ \le&\frac{c_\lambda^2 \cdot e^{-2\lambda t}}{\lambda^2}\left[(\|\tilde{\xi}\|^2(\|\tilde{A}\| + \lambda)^2 + \lambda\epsilon\mathbb{E}||\sigma||^2/\tau^2)e^{2\lambda\tau}\right]\\ &+ \epsilon c_\lambda^2\mathbb{E}\|\sigma\|^2/2\tau^2\lambda \le C_5e^{-2\lambda(t-\tau)} + C_6\frac{\epsilon}{\lambda\tau^2}. \qquad \square\end{aligned}$$

**Discussion:** We compare the results with the existing convergence rates of ASGD under consistent read setting. If we don't assume the data are sparse, the number of iterations is no less than $\mathcal{O}(\frac{\tau}{\epsilon})$ to achieve $\mathbb{E}\|x_k - x^*\|^2 \le \epsilon$ [Zheng *et al.*, 2017]. Another well-known theoretical result is $\mathcal{O}(\frac{\log(1/\epsilon)}{\epsilon} \cdot \tau^2)$ if we let the sparse coefficients be 1 and set $\eta = \frac{\epsilon}{\tau^2}$ [Recht *et al.*, 2011] [4]. Theorem 4.1 shows that we need $e^{-2\lambda(t-\tau)} \le \epsilon$ in order to make $\mathbb{E}\|X(t) - x^*\|^2 \le \mathcal{O}(\epsilon)$. Thus we need $t \ge \mathcal{O}(\tau + \frac{\log(1/\epsilon)}{2\lambda})$, which is in common faster than the previous two results $\mathcal{O}(\frac{\tau}{\epsilon})$ and $\mathcal{O}(\frac{\log(1/\epsilon)}{\epsilon} \cdot \tau^2)$.

## 4.2 Energy Function Minimization

In this section, we show the *energy function minimization*. This technique has been explored in many works [Su *et al.*, 2014; Krichene *et al.*, 2015; Wibisono *et al.*, 2016]. Firstly, we need to define proper *energy function* for corresponding differential equation. Energy function is related to the measure which is used for the convergence rate of optimization, such as $\|X(t) - x^*\|^2$, $F(X(t)) - F(x^*)$ and the expected convergence rate for the optimization algorithms. Suppose that we use ASGD to minimize a strongly convex loss function. We can define the energy function for SDDE as $\mathcal{E}(t) = \frac{t-1}{2}\|X(t) - x^*\|^2 - \frac{(\delta+1)(H+2L^2D^2\tau)\ln t}{2\mu^2}$ and calculate $\mathrm{d}\mathbb{E}\mathcal{E}(\mathrm{t})$ by using Itô formula. If $\mathrm{d}\mathbb{E}\mathcal{E}(\mathrm{t}) \le 0$, we can get $\mathbb{E}\mathcal{E}(t)$ is a decreasing function about $t$. Then the convergence rate for SDDE is obtained by using $\mathbb{E}\mathcal{E}(t) \le \mathcal{E}(t_0)$. Thus, the design for energy function aims to make $\mathrm{d}\mathbb{E}\mathcal{E}(\mathrm{t}) \le 0$. Theorem 4.2 shows the convergence rate for SDDE.

**Theorem 4.2** *For any* $F \in \mathcal{F}_\infty$ *with smooth coefficient* $L$, *let* $X(t)$ *be the unique global solution to (6) with initial conditions* $X(s) = \xi(s), s \in [-\tau, 0]$ *and* $X(t) \in d(x_0, D)$, $\forall t$. *Assume that* $F(x)$ *is strongly convex about* $x$ *and* $\eta_k = \frac{1}{\mu k}$. *Let* $H$ *be a constant which satisfies* $H \ge \mathbb{E}(Tr(\sigma\sigma^T)) + 2\mu LD^2$. *For any* $t > 1$,

$$\mathbb{E}\|X(t) - x^*\|^2 \le \frac{(\delta + 1)(H + 2L^2D^2\tau)\ln t}{(t - 1)\mu^2}.$$

[3] The coefficients are related to the root of characteristic function of the SDDE.

[4] Please notice that the theoretical analysis in [Recht *et al.*, 2011] is proved under consistent read setting.

*Proof sketch:* For strongly convex case, $\eta_k$ is set as $\eta_k = \frac{1}{\mu k}$. Then we have $\eta = 1$ and $U(t) = \frac{\delta}{\mu t}$. The SDDE is

$$\mathrm{d}X(t) = -\frac{1}{\mu t}\nabla F(X(t-\theta(t)))\mathrm{d}t + \frac{\sqrt{\delta}}{\mu t}\sigma(X(t-\theta(t)))\mathrm{d}B(t).$$

We define the energy function

$$\mathcal{E}(t) = \frac{t-1}{2}\|X(t)-x^*\|^2 - \frac{(\delta+1)(H+2L^2D^2\tau)\ln t}{2\mu^2},$$

where $H \geq \mathbb{E}(Tr(\sigma\sigma^T)) + 2\mu LD^2$. Using Itô formula and taking expectation, we can get

$$\begin{aligned}
&\mathrm{d}\mathbb{E}\mathcal{E}\\
=&\frac{1}{2}\mathbb{E}\|X(t)-x^*\|^2\mathrm{d}t - \frac{1}{\mu}\mathbb{E}\langle\nabla F(X(t-\theta(t))), X(t)-x^*\rangle\mathrm{d}t\\
&+\frac{1}{t\mu}\mathbb{E}\langle\nabla F(X(t-\theta(t))), X(t)-x^*\rangle\mathrm{d}t\\
&+\frac{(t-1)\delta}{2\mu^2t^2}\mathbb{E}Tr(\sigma\sigma^T)\mathrm{d}t - \frac{(\delta+1)(H+2L^2D^2\tau)}{2t\mu^2}\mathrm{d}t\\
\leq&\left(\frac{L^2D^2\theta(t)}{\mu^2 t} + \frac{LD^2}{t\mu}\right)\mathrm{d}t + \frac{\delta}{2\mu^2 t}\mathbb{E}Tr(\sigma\sigma^T)\mathrm{d}t\\
&-\frac{(\delta+1)(H+2L^2D^2\tau)}{2t\mu^2}\mathrm{d}t \leq 0.
\end{aligned}$$

The first equality comes from Itô formula. The first inequality applied the positiveness of $\frac{1}{2\mu^2t^2}\mathbb{E}Tr(\sigma\sigma^T)$ and strongly convex assumption. Therefore, we have

$$\mathbb{E}\|X(t)-x^*\|^2 \leq \frac{(\delta+1)(H+2L^2D^2\tau)\ln t}{(t-1)\mu^2},$$

where $\tau$ is the upper bound of $\theta(t)$. □

When we consider the special case of $\tau = 0$, ASGD algorithm reduces to SGD algorithm. We can similarly define an energy function $\mathcal{E}(t) = \frac{t-1}{2}\|X(t)-x^*\|^2 - \frac{(\delta+1)H\ln t}{2\mu^2}$ and we give the following corollary.

**Corollary 4.3** *For any $F \in \mathcal{F}_\infty$ with smooth coefficient $L$, let $X(t)$ be the unique global solution to*

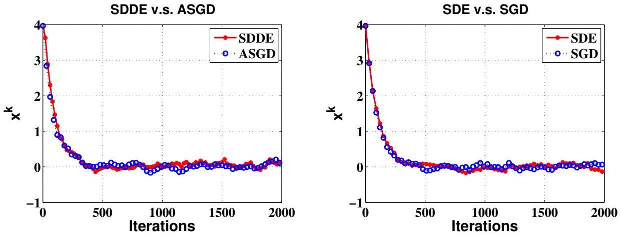
$$\mathrm{d}X(t) = -\frac{\eta}{\delta}U(t)\nabla F(X(t))\mathrm{d}t + \frac{\eta}{\delta}\sqrt{\delta}U(t)\sigma(X(t))\mathrm{d}B(t). \quad (8)$$

*with initial conditions $X(1) = x_0$ and $X(t) \in d(x_0, D)$, $\forall t$. Assume that $F(x)$ is strongly convex about $x$ and $\eta_k = \frac{1}{\mu k}$. Let $H \geq \mathbb{E}(Tr(\sigma\sigma^T)) + 2\mu LD^2$. For any $t > 1$,*

$$\mathbb{E}\|X(t)-x^*\|^2 \leq \frac{(\delta+1)H\ln t}{(t-1)\mu^2}.$$

**Discussions:** Since the approximation error is relatively small, the convergence rate of SDDE can be approximated as the result of ASGD. (1) We prove a tighter convergence rate for ASGD than the other existing results. The result in Theorem 4.2 is no slower than the convergence rate $\mathcal{O}(\tau/\epsilon)$ [Zheng *et al.*, 2017]. If $2L^2D^2\tau \leq H$, the negative influence of $\tau$ can be neglected, which means that it is comparable with the serial SGD and it can achieve linear speedup. Compared with the results in [Recht *et al.*, 2011], the speedup condition does not rely on the sparsity condition. If the stochastic variance $\mathbb{E}(Tr(\sigma\sigma^T))$ is large, the condition $2L^2D^2\tau \leq H$ is easier to be satisfied, which is consistent to the results for non-convex case studied in [Lian *et al.*, 2015].

(2) Corollary 4.3 shows that $\mathbb{E}\|x_k - x^*\|^2 \leq \frac{(\delta+1)H\ln(\delta k)}{(\delta k-1)\mu^2}$ since $t = \delta k$. It is comparable with the existing convergence rate of $\frac{4H}{k\mu^2}$ [Rakhlin *et al.*, 2011] under the same setting.



(a) (b)

Figure 1: Path comparison

## 5 Experiments

In this section, we take $F(x) = \frac{1}{2}(x+1)^2 + \frac{1}{2}(x-1)^2$ as a simple example. We compared the discrete optimization algorithms with the Euler-Maruyama schemes of stochastic differential equations.

For optimization iteration and Euler-Maruyama scheme, the learning rate is set to be $\eta = 0.005$. We run the experiments within 2000 iterations. The figures include ASGD (resp. SGD) path and a sample path for SDDE (resp. SDE) approximation. First, we analyzed SDDE for ASGD algorithm. The time interval of the SDDE is $[0, T]$ where $T = \eta K$ and $K$ is the total number of ASGD iterations. For SDDE, we set the initial function as $\xi(\theta) = 4$ for any $\theta \in [-\tau, 0]$. For ASGD, we assume a constant delay as $l = 10$ and let $X(0) = 4$. Since the delay $l = 10$, we run the SGD during the first ten steps and ASGD iterations are used from ten steps on. It is observed that the two paths are close. Second, from Fig.1(b) we can see that the SDE approximation and SGD iteration are well matched.

## 6 Conclusions

In this paper, we studied the continuous approximation of asynchronous stochastic gradient descent algorithm. We analyzed the approximation error and conducted theoretical analyses on the convergence rates of asynchronous stochastic gradient decent algorithms by continuous methods: moment estimation and energy function minimization. From the aspect of continuous view, we cannot only obtain existing results by the discrete view, but also get new results. For applications of this view, it can also be referred at [Mandt *et al.*, 2016; Li *et al.*, 2016]. To the best of our knowledge, we firstly give a continuous description to asynchronous stochastic gradient descent, and proved a tighter bound. In the future, we will apply this unified continuous view to analyze more optimization algorithms in more tasks.

## Acknowledgements

We would like to thank Liping Li and Jianhai Bao for their helpful discussions and suggestions. Zhi-Ming Ma was partially supported by National Center for Mathematics and Interdisciplinary Sciences of Chinese Academy of Sciences.

## References

[Agarwal and Duchi, 2011] Alekh Agarwal and John C Duchi. Distributed delayed stochastic optimization. In *Advances in Neural Information Processing Systems*, pages 873–881, 2011.

[Balles *et al.*, 2017] Lukas Balles, Javier Romero, and Philipp Hennig. Coupling adaptive batch sizes with learning rates. *Conference on Uncertainty in Artificial Intelligence*, 2017.

[Bao *et al.*, 2014] Jianhai Bao, George Yin, and Chenggui Yuan. Ergodicity for functional stochastic differential equations and applications. *Nonlinear Analysis: Theory, Methods & Applications*, 98:66–82, 2014.

[Bao *et al.*, 2016] Jianhai Bao, George Yin, and Chenggui Yuan. Asymptotic analysis for functional stochastic differential equations, 2016.

[Dean *et al.*, 2012] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Andrew Senior, Paul Tucker, Ke Yang, Quoc V Le, et al. Large scale distributed deep networks. In *Advances in neural information processing systems*, pages 1223–1231, 2012.

[Durrett, 2010] Rick Durrett. *Probability: theory and examples*. Cambridge university press, 2010.

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

[Krichene *et al.*, 2015] W Krichene, Am Bayen, and Pl Bartlett. Accelerated mirror descent in continuous and discrete time. pages 2845–2853, 2015.

[Langford *et al.*, 2009] John Langford, Alex J Smola, and Martin Zinkevich. Slow learners are fast. *Advances in Neural Information Processing Systems*, 22:2331–2339, 2009.

[Li *et al.*, 2016] Chris Junchi Li, Zhaoran Wang, and Han Liu. Online ica: Understanding global dynamics of nonconvex optimization via diffusion processes. In *Advances in Neural Information Processing Systems*, pages 4961–4969, 2016.

[Li *et al.*, 2017] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2101–2110, 2017.

[Lian *et al.*, 2015] Xiangru Lian, Yijun Huang, Yuncheng Li, and Ji Liu. Asynchronous parallel stochastic gradient for nonconvex optimization. In *International Conference on Neural Information Processing Systems*, pages 2737–2745, 2015.

[Mandt *et al.*, 2016] Stephan Mandt, Matthew D Hoffman, and David M Blei. A variational analysis of stochastic gradient algorithms. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 354–363, 2016.

[Mao, 2007] Xuerong Mao. *Stochastic differential equations and applications*. Elsevier, 2007.

[Raginsky and Bouvrie, 2012] Maxim Raginsky and J Bouvrie. Continuous-time stochastic mirror descent on a network: Variance reduction, consensus, convergence. 23(1):6793–6800, 2012.

[Rakhlin *et al.*, 2011] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

[Recht *et al.*, 2011] Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 693–701, 2011.

[Su *et al.*, 2014] Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

[Uhlenbeck and Ornstein, 1930] George E Uhlenbeck and Leonard S Ornstein. On the theory of the brownian motion. *Physical review*, 36(5):823, 1930.

[Wibisono *et al.*, 2016] Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, page 201614734, 2016.

[Yang *et al.*, 2017] Lin F. Yang, R. Arora, V. Braverman, and Tuo Zhao. The physical systems behind optimization algorithms. 2017.

[Zhang *et al.*, 2015] Sixin Zhang, Anna E Choromanska, and Yann LeCun. Deep learning with elastic averaging sgd. In *Advances in Neural Information Processing Systems*, pages 685–693, 2015.

[Zheng *et al.*, 2017] Shuxin Zheng, Qi Meng, Taifeng Wang, Wei Chen, Nenghai Yu, Zhi-Ming Ma, and Tie-Yan Liu. Asynchronous stochastic gradient descent with delay compensation. In *International Conference on Machine Learning*, pages 4120–4129, 2017.