A Continuous-Time Analysis of Distributed Stochastic Gradient

Nicholas M. Boffi
boffi@g.harvard.edu
John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, U.S.A.

Jean-Jacques E. Slotine
jjs@mit.edu
Nonlinear Systems Laboratory, MIT, Cambridge, MA 02139, U.S.A.

We analyze the effect of synchronization on distributed stochastic gradient algorithms. By exploiting an analogy with dynamical models of biological quorum sensing, where synchronization between agents is induced through communication with a common signal, we quantify how synchronization can significantly reduce the magnitude of the noise felt by the individual distributed agents and their spatial mean. This noise reduction is in turn associated with a reduction in the smoothing of the loss function imposed by the stochastic gradient approximation. Through simulations on model nonconvex objectives, we demonstrate that coupling can stabilize higher noise levels and improve convergence. We provide a convergence analysis for strongly convex functions by deriving a bound on the expected deviation of the spatial mean of the agents from the global minimizer for an algorithm based on quorum sensing, the same algorithm with momentum, and the elastic averaging SGD (EASGD) algorithm. We discuss extensions to new algorithms that allow each agent to broadcast its current measure of success and shape the collective computation accordingly. We supplement our theoretical analysis with numerical experiments on convolutional neural networks trained on the CIFAR-10 data set, where we note a surprising regularizing property of EASGD even when applied to the non-distributed case. This observation suggests alternative second-order in time algorithms for nondistributed optimization that are competitive with momentum methods.

1 Introduction

Stochastic gradient descent (SGD) and its variants have become the de facto algorithms for large-scale machine learning applications such as deep neural networks (Bottou, 2010; Goodfellow, Bengio, & Courville, 2016; LeCun,

Neural Computation **32, 36–96** (2020)
https://doi.org/10.1162/neco_a_01248

Bengio, & Hinton, 2015; Mallat, 2016). SGD is used to optimize finite-sum loss functions, where a stochastic approximation to the gradient is computed using only a random selection of the input data points. Well-known results on almost-sure convergence rates to global minimizers for strictly convex functions and to stationary points for non-convex functions exist under sufficient regularity conditions (Bottou, 1998; Robbins & Siegmund, 1971). Classic work on iterate averaging for SGD (Polyak & Juditsky, 1992) and other more recent extensions (Bach & Moulines, 2013; Defazio, Bach, & Lacoste-Julien, 2014; Roux, Schmidt, & Bach, 2012; Schmidt, Le Roux & Bach, 2017) can improve convergence under a set of reasonable assumptions typically satisfied in the machine learning setting. Convergence proofs rely on a suitably chosen decreasing step size; for constant step sizes and strictly convex functions, the parameters ultimately converge to a distribution peaked around the optimum.

For large-scale machine learning applications, parallelization of SGD is a critical problem of significant modern research interest (Chaudhari et al., 2017; Dean et al., 2012; Recht & Ré, 2013; Recht, Re, Wright, & Niu, 2011). Recent work in this direction includes the elastic averaging SGD (EASGD) algorithm, in which *p* distributed agents coupled through a common signal optimize the same loss function. EASGD can be derived from a single SGD step on a global variable consensus objective with a quadratic penalty, and the common signal takes the form of an average over space and time of the parameter vectors of the individual agents (Boyd, Parikh, Chu, Peleato, & Eckstein, 2010; Zhang, Choromanska & LeCun, 2015). At its core, the EASGD algorithm is a system of identical, coupled, discrete-time dynamical systems. And indeed, the EASGD algorithm has exactly the same structure as earlier mathematical models of synchronization (Chung & Slotine, 2009; Russo & Slotine, 2010) inspired by quorum sensing in bacteria (Miller & Bassler, 2001; Waters & Bassler, 2005). In these models, which have typically been analyzed in continuous-time, the dynamics of the common (quorum) signal can be arbitrary (Russo & Slotine, 2010), and in fact they may consist simply of a weighted average of individual signals. Motivated by this immediate analogy, we present here a continuous-time analysis of distributed stochastic gradient algorithms, of which EASGD is a special case. A significant focus of this work is the interaction between the degree of synchronization of the individual agents, characterized rigorously by a bound on the expected distance between all agents and governed by the coupling strength, and the amount of noise induced by their stochastic gradient approximations.

The effect of coupling between identical continuous-time dynamical systems has a rich history. In particular, synchronization phenomena in such coupled systems have been the subject of much mathematical (Wang & Slotine, 2005), biological (Russo & Slotine, 2010), neuroscientific (Tabareau, Slotine & Pham, 2010), and physical interest (Javaloyes, Perrin, & Politi, 2008). In nonlinear dynamical systems, synchronization has been shown to play a crucial role in protection of the individual systems from independent sources of noise (Tabareau et al., 2010). The interaction between synchronization and noise has also been posed as a possible source of regularization in biological learning, where quorum sensing–like mechanisms could be implemented between neurons through local field potentials (Bouvrie & Slotine, 2013). Given the significance of stochastic gradient (Y. Zhang, Saxe, Advani & Lee, 2018) and externally injected (Neelakantan et al., 2015) noise in regularization of large-scale machine learning models such as deep networks (Zhang, Bengio, Hardt, Recht & Vinyals, 2017), it is natural to expect that the interplay between synchronization of the individual agents and the noise from their stochastic gradient approximations is of central importance in distributed SGD algorithms.

Recently, there has been renewed interest in a continuous-time view of optimization algorithms (Betancourt, Jordan, & Wilson, 2018; Wibisono & Wilson, 2015; Wibisono, Wilson & Jordan, 2016; Wilson, Recht & Jordan, 2016). Nesterov's accelerated gradient method (Nesterov, 1983) was fruitfully analyzed in continuous time in Su, Boyd and Candes (2014), and a unifying extension to other algorithms can be found in Wibisono et al. (2016). Continuous-time analysis has also enabled discrete-time algorithm development through classical discretization techniques from numerical analysis (Zhang, Mokhtari, Sra & Jadbabaie, 2018). This article adds to this line of work by deriving new results with the mathematical tools afforded by the continuous-time view, such as stochastic calculus and nonlinear contraction analysis (Lohmiller & Slotine, 1998).

The article is organized as follows. In section 2, we provide some necessary mathematical preliminaries: a review of SGD in continuous time, a continuous-time limit of the EASGD algorithm, a review of stochastic nonlinear contraction theory, and a statement of some needed assumptions. In section 3, we demonstrate that the effect of synchronization of the distributed SGD agents is to reduce the magnitude of the noise felt by each agent and by their spatial mean. We derive this for an algorithm where all-to-all coupling is implemented through communication with the spatial mean of the distributed parameters, and we refer to this algorithm as quorum SGD (QSGD). The appendix presents a similar derivation with arbitrary dynamics for the quorum variable, of which EASGD is a special case. In section 4, we connect this noise reduction property with a recent analysis in Kleinberg, Li, and Yuan (2018), which shows that SGD can be interpreted as performing gradient descent on a smoothed loss in expectation. We use this derivation to garner intuition about the qualitative performance of distributed SGD algorithms as the coupling strength is varied, and we verify this intuition with simulations on model non-convex loss functions in low and high dimensions. In section 5, we provide new convergence results for QSGD, QSGD with momentum, and EASGD for a strongly convex objective. In section 6, we explore the properties of EASGD and QSGD for training deep neural networks and, in particular, test the stability and performance of variants proposed throughout the article. We also propose a new class of second-order in time algorithms motivated by the EASGD algorithm with a single agent, which consists of standard SGD coupled in feedback to the output of a nonlinear filter of the parameters. We provide some concluding remarks in section 7.

2 Mathematical Preliminaries

In this section, we provide a brief review of the necessary mathematical tools employed in this work.

2.1 Convex Optimization.

For the convergence proofs in section 5 and for synchronization of momentum methods, we require a few standard definitions from convex optimization.

Definition 1 (Strong Convexity). *A function $f \in \mathcal{C}^2(\mathbb{R}^n, \mathbb{R})$ is l-strongly convex with $l > 0$ if its Hessian is uniformly lower bounded by $l\mathbf{I}$ with respect to the positive semidefinite order, $\nabla^2 f(\boldsymbol{x}) > l\mathbf{I}$ for all $\boldsymbol{x} \in \mathbb{R}^n$.*

Definition 2 (L-Smoothness). *A function $f \in \mathcal{C}^2(\mathbb{R}^n, \mathbb{R})$ is L-smooth with $L > 0$ if its Hessian is uniformly upper bounded by $L\mathbf{I}$ with respect to the positive semidefinite order, $\nabla^2 f(\boldsymbol{x}) < L\mathbf{I}$ for all $\boldsymbol{x} \in \mathbb{R}^n$.*

2.2 Stochastic Gradient Descent in Discrete-Time.

Minibatch SGD has been essential for training large-scale machine learning models such as deep neural networks, where empirical risk minimization leads to finite-sum loss functions of the form

$$f(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} l(\mathbf{x}, \mathbf{y}^i).$$

Above, $\mathbf{y}^i$ is the ith input data example, and the vector $\mathbf{x}$ holds the model parameters. In the typical machine learning setting where N is very large, the gradient of f requires N gradient computations of l, which is prohibitively expensive.

To avoid this calculation, a stochastic gradient is computed by taking a random selection $\mathcal{B}$ of size $b < N$, typically known as a minibatch. It is simple to see that the stochastic gradient,

$$\hat{\mathbf{g}}(\mathbf{x}) = \frac{1}{b} \sum_{\mathbf{y} \in \mathcal{B}} \nabla l(\mathbf{x}, \mathbf{y}),$$

is an unbiased estimator of the true gradient. The parameters are updated according to the iteration

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \hat{\mathbf{g}}(\mathbf{x}).$$

By adding and subtracting the true gradient, the SGD iteration can be rewritten as

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t) - \frac{\eta}{\sqrt{b}} \boldsymbol{\zeta}_t, \tag{2.1}$$

where $\boldsymbol{\zeta}_t \sim N(0, \boldsymbol{\Sigma}(\mathbf{x}_t))$ is a data-dependent noise term. $\boldsymbol{\zeta}_t$ can be taken to be gaussian under a central limit theorem argument, assuming that the size of the minibatch is large enough (Jastrzębski et al., 2017; Mandt, Hoffman, & Blei, 2015). $\boldsymbol{\Sigma}(\mathbf{x})$ is then given by the variance of a single-element stochastic gradient:

$$\boldsymbol{\Sigma}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} \left[\left(\nabla l\left(\mathbf{x}, \mathbf{y}^i\right) - \nabla f\left(\mathbf{x}\right) \right) \left(\nabla l\left(\mathbf{x}, \mathbf{y}^i\right) - \nabla f\left(\mathbf{x}\right) \right)^T \right].$$

2.3 Stochastic Gradient Descent in Continuous-Time. A significant difficulty in a continuous-time analysis of SGD is formulating an accurate stochastic differential equation (SDE) model. Recent work has proved rigorously (Feng, Li, & Liu, 2018; Hu, Li, Li, & Liu, 2017; Li, Tai, & Weinan, 2018) that the sequence of values $\mathbf{x}(k\eta)$ generated by the SDE,

$$d\mathbf{x} = \left(-\nabla f(\mathbf{x}) - \frac{1}{4} \eta \nabla \left\| \nabla f(\mathbf{x}) \right\|^2 \right) dt + \sqrt{\frac{\eta}{b}} \mathbf{B}(\mathbf{x}) d\mathbf{W},$$

approximates the SGD iteration with weak error $\mathcal{O}(\eta^2)$, where $\mathbf{W}$ is a Wiener process, $\| \cdot \|$ denotes the Euclidean 2-norm,[1] and $\mathbf{B}\mathbf{B}^T = \boldsymbol{\Sigma}$. Dropping the small term proportional to η reduces the weak error to $\mathcal{O}(\eta)$ (Hu et al., 2017). This leads to the SDE:

$$d\mathbf{x} = -\nabla f(\mathbf{x}) dt + \sqrt{\frac{\eta}{b}} \mathbf{B}(\mathbf{x}) d\mathbf{W}. \tag{2.2}$$

Equation 2.2 has appeared in a number of recent works (Chaudhari, Oberman, Osher, Soatto, & Carlier, 2018; Chaudhari & Soatto, 2018; Jastrzębski et al., 2017; Mandt et al., 2015; Mandt, Hoffman, & Blei, 2016, 2017) and is generally obtained by making the replacement $\eta \rightarrow dt$ and $\sqrt{\eta} \boldsymbol{\zeta} \rightarrow \mathbf{B} d\mathbf{W}$ in equation 2.1 as a sort of reverse Euler-Maruyama discretization (Kloeden & Platen, 1992).

[1] For the remainder of this article, unless otherwise specified, we will use $\| \cdot \|$ to denote the 2-norm.

2.4 EASGD in Continuous-Time. Following Zhang et al. (2015), we provide a brief introduction to the EASGD algorithm and convert the resulting sequences to continuous-time. We imagine a distributed optimization setting with $p \in \mathbb{N}$ agents and a single master. We are interested in solving a stochastic optimization problem,

$$\min_{\mathbf{x}} F(\mathbf{x}) = \mathbb{E}_{\boldsymbol{\zeta}}\left[f(\mathbf{x}, \boldsymbol{\zeta})\right]$$

where $\mathbf{x} \in \mathbb{R}^n$ is the vector of parameters and $\boldsymbol{\zeta}$ is a random variable representing the stochasticity in the objective. This is equivalent to the distributed optimization problem (Boyd et al., 2010),

$$\min_{\mathbf{x}^1,\ldots,\mathbf{x}^p,\tilde{\mathbf{x}}} \sum_{i=1}^{p} \left(\mathbb{E}_{\boldsymbol{\zeta}^i}[f(\mathbf{x}^i, \boldsymbol{\zeta}^i)] + \frac{k}{2}\|\mathbf{x}^i - \tilde{\mathbf{x}}\|^2\right), \tag{2.3}$$

where each $\mathbf{x}^i$ is a local vector of parameters and $\tilde{\mathbf{x}}$ is the quorum variable. The quadratic penalty ensures that all local agents remain close to $\tilde{\mathbf{x}}$, and k sets the coupling strength. Smaller values of k allow for more exploration, while larger values ensure a greater degree of synchronization. Intuitively, the interaction between agents mediated by $\tilde{\mathbf{x}}$ is expected to help individual trajectories escape local minima, saddle points, and flat regions unless they all fall into the same deep or wide minimum together.

We assume the expectation in equation 2.3 is approximated by a sum over input data points and that the stochastic gradient is computed by taking a minibatch of size b. After taking an SGD step, the updates for each agent and the quorum variable become

$$\begin{aligned}
\mathbf{x}^i_{t+1} &= \mathbf{x}^i_t - \eta\nabla f(\mathbf{x}^i_t) + \eta k\left(\tilde{\mathbf{x}}_t - \mathbf{x}^i_t\right) - \frac{\eta}{\sqrt{b}}\boldsymbol{\zeta}^i_t,\\
\tilde{\mathbf{x}}_{t+1} &= \tilde{\mathbf{x}}_t + \eta pk\left(\mathbf{x}^{\bullet}_t - \tilde{\mathbf{x}}_t\right),
\end{aligned}$$

where $\mathbf{x}^{\bullet}_t = \frac{1}{p}\sum_{i=1}^{p}\mathbf{x}^i_t$ and $\mathbb{E}\left[\boldsymbol{\zeta}^i_t(\boldsymbol{\zeta}^i_t)^T\right] = \boldsymbol{\Sigma}(\mathbf{x}^i_t)$. Transferring to the continuous-time limit, these equations become,

$$d\mathbf{x}^i = \left(-\nabla f(\mathbf{x}^i) + k\left(\tilde{\mathbf{x}} - \mathbf{x}^i\right)\right)dt + \sqrt{\frac{\eta}{b}}\mathbf{B}(\mathbf{x}^i)d\mathbf{W}^i, \tag{2.4}$$

$$d\tilde{\mathbf{x}} = kp\left(\mathbf{x}^{\bullet} - \tilde{\mathbf{x}}\right)dt, \tag{2.5}$$

with $\mathbf{B}\mathbf{B}^T = \boldsymbol{\Sigma}$. Note that in equation 2.5, the dynamics of $\tilde{\mathbf{x}}$ represent a simple low-pass filter of the center of mass (spatial mean) variable $\mathbf{x}^{\bullet}$. In the limit of large p, the dynamics of this filter will be much faster than the SGD dynamics, and the continuous-time EASGD system can be approximately replaced by

$$d\mathbf{x}^i = \left(-\nabla f(\mathbf{x}^i) + k\left(\mathbf{x}^\bullet - \mathbf{x}^i\right)\right) dt + \sqrt{\frac{\eta}{b}}\mathbf{B}(\mathbf{x}^i) d\mathbf{W}^i. \tag{2.6}$$

We refer to equation 2.6 as quorum SGD (QSGD) and it will be a significant focus of this work.

2.5 Background on Nonlinear Contraction Theory. The main mathematical tool used in this work is nonlinear contraction theory, a form of incremental stability for nonlinear systems. In particular, we specialize to the case of time- and state-independent metrics (further details can be found in Lohmiller & Slotine, 1998).

Definition 3 (Contraction). *The nonlinear dynamical system*

$$\dot{x} = f(x, t), \tag{2.7}$$

with $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{f} \in \mathcal{C}^1(\mathbb{R}^n \times \mathbb{R}, \mathbb{R}^n)$, is said to be contracting with rate $\lambda > 0$ and invertible metric transformation $\boldsymbol{\Theta} \in \mathbb{R}^{n\times n}$ if the symmetric part of the generalized Jacobian,

$$\left(\boldsymbol{\Theta}\nabla f(x, t)\boldsymbol{\Theta}^{-1}\right)_s \leq -\lambda \boldsymbol{I}, \tag{2.8}$$

is uniformly negative definite for all $\boldsymbol{x} \in \mathbb{R}^n$ and all $t \in \mathbb{R}$. Above, subscript s denotes the symmetric part of a matrix, $\boldsymbol{A}_s = \frac{1}{2}\left(\boldsymbol{A} + \boldsymbol{A}^T\right)$. Equivalently, the system is said to be contracting in the corresponding metric $\boldsymbol{M} = \boldsymbol{\Theta}^T\boldsymbol{\Theta}$.

If condition 2.8 is satisfied, all trajectories exponentially converge to one another regardless of initial conditions. That is, for two solutions $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ of equation 2.7,

$$\|\mathbf{x}_1(t) - \mathbf{x}_2(t)\|_{\mathbf{M}} \leq e^{-\lambda t}\|\mathbf{x}_1(0) - \mathbf{x}_2(0)\|_{\mathbf{M}}, \tag{2.9}$$

where $\|\mathbf{x}\|_{\mathbf{M}} = \sqrt{\mathbf{x}^T\mathbf{M}\mathbf{x}}$. Intuitively, because of property 2.9, a nonlinear system is called contracting if differences in system trajectories due to initial conditions and temporary disturbances are exponentially forgotten. This behavior is proved differentially by considering the time evolution of the squared Euclidean norm of the virtual displacement $\delta\mathbf{z} = \boldsymbol{\Theta}\delta\mathbf{x}$, which formally obeys the differential equation $\delta\dot{\mathbf{z}} = \boldsymbol{\Theta}\nabla\mathbf{f}(\mathbf{x})\boldsymbol{\Theta}^{-1}\delta\mathbf{z}$ (Lohmiller & Slotine, 1998). As an immediate and powerful corollary, if the system is contracting and a single trajectory is known, then all trajectories must converge to the single known trajectory exponentially.

In this work, we interchangeably refer to $\mathbf{f}$, the system, and the generalized Jacobian as contracting depending on the context. In particular, for stochastic differential equations, we refer to $\mathbf{f}$ as contracting if the deterministic system is contracting. Two specific robustness results for contracting systems needed for the derivations in this work are summarized below.

Lemma 1. *Consider the dynamical system 2.7, and assume that it is contracting with metric transformation $\boldsymbol{\Theta}$ and contraction rate λ. Let $\chi = \|\boldsymbol{\Theta}^{-1}\|\|\boldsymbol{\Theta}\|$ denote the condition number of $\boldsymbol{\Theta}$, where $\|\boldsymbol{\Theta}\| = sup_{\|\boldsymbol{y}\|=1}\|\boldsymbol{\Theta y}\|$ denotes the induced matrix 2-norm. Consider the perturbed dynamical system:*

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, t) + \boldsymbol{\epsilon}(\boldsymbol{x}, t). \tag{2.10}$$

Then, for a solution $\boldsymbol{x}_1$ of equation 2.7 and a solution $\boldsymbol{x}_2$ of equation 2.10, with $R = \|\boldsymbol{\Theta}\left(\boldsymbol{x}_1 - \boldsymbol{x}_2\right)\|$,

$$\dot{R} + \lambda R \leq \|\boldsymbol{\Theta\epsilon}(\boldsymbol{x}, t)\|. \tag{2.11}$$

Furthermore, if $\|\boldsymbol{\epsilon}\| \leq Ae^{-at} + B$ with $A, B \in \mathbb{R}$ and $a \in \mathbb{R}^+$, then after exponential transients of rates a and λ,

$$\|\boldsymbol{x}_1(t) - \boldsymbol{x}_2(t)\| \leq \frac{\chi B}{\lambda}. \tag{2.12}$$

Proof. See point vii of "linear properties of generalized contraction analysis" in Lohmiller and Slotine (1998) for the derivation of equation 2.11. From equation 2.11, $\dot{R} + \lambda R \leq \|\boldsymbol{\Theta}\|\|\boldsymbol{\epsilon}\| \leq \|\boldsymbol{\Theta}\|\left(Ae^{-at} + B\right)$. Convolving $e^{-\lambda t}$ with the right-hand side yields the inequality

$$R(t) \leq \|\boldsymbol{\Theta}\|\left(\frac{B}{\lambda} + \frac{Ae^{-\lambda t}}{a-\lambda} - \frac{Be^{-\lambda t}}{\lambda} - \frac{Ae^{-at}}{a-\lambda}\right). \tag{2.13}$$

Noting that $\|\mathbf{x}_1(t) - \mathbf{x}_2(t)\| = \|\boldsymbol{\Theta}^{-1}\boldsymbol{\Theta}\left(\mathbf{x}_1 - \mathbf{x}_2\right)\| \leq \|\boldsymbol{\Theta}^{-1}\|\|\boldsymbol{\Theta}\left(\mathbf{x}_1 - \mathbf{x}_2\right)\| = \|\boldsymbol{\Theta}^{-1}\|R$ yields the equation 2.12. □

Theorem 1. *Consider the stochastic differential equation,*

$$d\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, t)dt + \boldsymbol{\sigma}(\boldsymbol{x}, t)d\boldsymbol{W}, \tag{2.14}$$

with $\boldsymbol{x} \in \mathbb{R}^n$ and where $\boldsymbol{W}$ denotes an n-dimensional Wiener process. Assume that there exists a positive-definite metric $\boldsymbol{M} = \boldsymbol{\Theta}^T\boldsymbol{\Theta}$ such that $\boldsymbol{x}^T\boldsymbol{M}\boldsymbol{x} \geq \beta\|\boldsymbol{x}\|^2$ with $\beta > 0$ and that $\boldsymbol{f}$ is contracting in this metric. Further assume that $Tr\left(\boldsymbol{\sigma}(\boldsymbol{x}, t)^T\boldsymbol{M}\boldsymbol{\sigma}(\boldsymbol{x}, t)\right) \leq C$ where $C \in \mathbb{R}^+$. Then, for two trajectories $\boldsymbol{a}(t)$ and $\boldsymbol{b}(t)$ driven by independent sources of noise with stochastic initial conditions given by a probability distribution $p(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2)$,

$$\mathbb{E}\left[\|\boldsymbol{a}(t) - \boldsymbol{b}(t)\|^2\right] \leq \frac{1}{\beta}\left(\mathbb{E}\left[\left(\|\boldsymbol{a}(0) - \boldsymbol{b}(0)\|_{\boldsymbol{M}}^2 - \frac{C}{\lambda}\right)^+\right]e^{-2\lambda t} + \frac{C}{\lambda}\right)$$

where $(\cdot)^+$ denotes the unit ramp (or ReLU) function. The expectation on the left-hand side is over the noise $d\mathbf{W}(s)$ for all $s < t$, and the expectation on the right-hand side is over the distribution of initial conditions.

See Pham, Tabareau, and Slotine (2009, theorem 2) for a proof. The following corollary will be useful in section 5.

Corollary 1. *Assume that the conditions of theorem 1 are satisfied. Then for a trajectory $\boldsymbol{x}_{nf}(t)$ of equation 2.7 and a trajectory $\boldsymbol{x}(t)$ of equation 2.14,*

$$\mathbb{E}\left[\|\boldsymbol{x}(t) - \boldsymbol{x}_{nf}(t)\|^2\right] \leq \frac{1}{\beta}\left(\mathbb{E}\left[\left(\|\boldsymbol{x}(0) - \boldsymbol{x}_{nf}(0)\|_{\boldsymbol{M}}^2 - \frac{C}{2\lambda}\right)^+\right]e^{-2\lambda t} + \frac{C}{2\lambda}\right).$$

Corollary 1 is obtained by following the proof of theorem 2 in Pham et al. (2009), with the restriction that one system is deterministic. To reduce the appearance of decaying exponential terms, in applications of theorem 1, corollary 1, and other related contraction-based bounds, we will simply state the final constant and the corresponding rate of exponential transients. The conditions of theorem 1 are worthy of their own definition.

Definition 4 (Stochastic Contraction). *If the conditions of theorem 1 are satisfied, system 2.14 is said to be stochastically contracting in the metric $\boldsymbol{M}$ (or with metric transformation $\boldsymbol{\Theta}$) with bound C and rate λ.*

In this work, we will also make use of an extension of contraction known as partial contraction originally introduced in Wang and Slotine (2005). The procedure is summarized in theorem 2:

Theorem 2. *Consider the nonlinear dynamical system 2.7—not assumed to be contracting—and consider a contracting auxiliary system of the form*

$$\dot{\boldsymbol{y}} = \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{x}, t) \tag{2.15}$$

with the requirement that $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{x}, t) = \boldsymbol{f}(\boldsymbol{x}, t)$.[2] *Assume a single trajectory $\boldsymbol{y}(t)$ of equation 2.15 is known. Then all trajectories of equation 2.7 converge to $\boldsymbol{y}(t)$.*

Proof. By assumption, equation 2.15 is contracting, and so all trajectories converge to $\mathbf{y}(t)$. Because $\mathbf{g}(\mathbf{x}, \mathbf{x}, t) = \mathbf{f}(\mathbf{x}, t)$, any solution $\mathbf{x}(t)$ of equation 2.7 is also a solution of equation 2.15, and hence must converge to $\mathbf{y}(t)$. □

We will commonly refer to the auxiliary $\mathbf{y}$ system in theorem 2 as a virtual system, and $\mathbf{f}$ is said to be partially contracting. Theorem 2 enables the

[2] For example, say $\mathbf{f}(\mathbf{x}, t) = -\mathbf{P}(\mathbf{x})\mathbf{x}$ with $\mathbf{P}(\mathbf{x})$ a symmetric and uniformly positive-definite matrix. Then $\mathbf{g}(\mathbf{y}, \mathbf{x}, t) = -\mathbf{P}(\mathbf{x})\mathbf{y}$ satisfies this restriction requirement. The $\mathbf{y}$ system is also contracting in $\mathbf{y}$, as the symmetric part of the Jacobian $\mathbf{J}_s = -\mathbf{P}(\mathbf{x}) < 0$ uniformly. The $\mathbf{f}(\mathbf{x}, t)$ system has Jacobian $\frac{\partial f_i}{\partial x_j} = -P_{ij}(\mathbf{x}) - \sum_k \frac{\partial P_{ik}(\mathbf{x})}{\partial x_j} x_k$, which has a symmetric part with unknown definiteness without further assumptions on $\mathbf{P}$.

application of contraction to systems that in themselves are not contracting but can be embedded in a virtual system that is.

This notion also extends to stochastic systems through the use of stochastic contraction. If a stochastically contracting system,

$$d\mathbf{y} = g(\mathbf{y}, \mathbf{x}, t)dt + \Xi(\mathbf{y}, \mathbf{x}, t)d\mathbf{W}, \tag{2.16}$$

can be found such that $\mathbf{g}(\mathbf{x}, \mathbf{x}, t) = \mathbf{f}(\mathbf{x}, t)$ and $\Xi(\mathbf{x}, \mathbf{x}) = \boldsymbol{\sigma}(\mathbf{x}, t)$, then trajectories of equation 2.16 can be compared to trajectories of equation 2.7 through the application of corollary 1, or to the trajectories of equation 2.14 through the application of theorem 1.

2.6 Assumptions. We require two main assumptions about the objective function $f(\mathbf{x})$, both of which have been employed in previous work analyzing synchronization and noise in nonlinear systems (Tabareau et al., 2010). The first is an assumption on the nonlinearity of the components of the gradient.

Assumption 1. Assume that the Hessian matrix of each component of the negative gradient has bounded maximum eigenvalue, $\nabla^2\left[(-\nabla f(\mathbf{x}))_j\right] \leq Q\mathbf{I}$ for all j.

The second assumption is a condition on the robustness of the distributed gradient flows studied in this work to small, potentially stochastic perturbations.

Assumption 2. Consider two dynamical systems,

$$\dot{\mathbf{x}} = -\nabla f(\mathbf{x}) + k(\mathbf{z} - \mathbf{x}), \tag{2.17}$$

$$d\mathbf{y} = (-\nabla f(\mathbf{y}) + k(\mathbf{z} - \mathbf{y}) + \mathbf{P}_l)\, dt + \beta_q d\mathbf{W}, \tag{2.18}$$

where $\mathbf{P}_l$ is a continuous-time stochastic process dependent on a parameter l and $\beta_q \in \mathbb{R}$ is a real coefficient dependent on a parameter q. Denote by $\mathbf{x}(t)$ the solution to equation 2.17 and by $\mathbf{y}_{l,q}(t)$ the solution to equation 2.18 with the same initial condition, $\mathbf{x}(0) = \mathbf{y}_{l,q}(0)$. We assume that $\lim_{l\to\infty} \mathbb{E}\left(\|\mathbf{P}_l\|\right) = 0$ and $\lim_{q\to\infty} \beta_q = 0$ implies that $\lim_{l\to\infty} \lim_{q\to\infty} \|\mathbf{x} - \mathbf{y}_{l,q}\| = 0$ almost surely.

Continuous dependence of trajectories on the parameters of the dynamics in the sense of assumption 2 can be characterized for deterministic systems through continuity assumptions on the dynamics (see, e.g., section 3.2 in Khalil, 2002). Here we assume a natural stochastic extension. Assumption 2 has been verified for FitzHugh-Nagumo oscillators where $\mathbf{P}_l$ is a white noise process (Tuckwell & Rodriguez, 1998) and validated in simulation for more complex nonlinear oscillators (Tabareau et al., 2010). We remark that

$\mathbb{E}[\|\mathbf{P}\|] \to 0$ implies that $\|\mathbf{P}\| \to 0$ almost surely, and hence that $\mathbf{P} \to \mathbf{0}$ almost surely.

3 Synchronization and Noise

In this section, we analyze the interaction between synchronization of the distributed QSGD agents and the noise they experience. We begin with a derivation of a quantitative measure of synchronization that applies to a class of distributed SGD algorithms involving coupling to a common external signal with no communication delays. We then present the section's primary contribution, which will serve as a basis for the theory in the remainder of the article, as well as for the intuition for various experiments.

3.1 A Measure of Synchronization.

We now present a simple theorem on synchronization in the deterministic setting, which will allow us to prove a bound on synchronization in the stochastic setting using theorem 1.

Theorem 3. *Consider the coupled gradient descent system,*

$$\dot{x}^i = -\nabla f(x^i) + k(z - x^i), \tag{3.1}$$

where z represents a common external signal. Let $\bar{\lambda}$ denote the maximum eigenvalue of $-\nabla^2 f(x)$. For $k > \bar{\lambda}$, the individual x^i trajectories synchronize exponentially with rate $k - \bar{\lambda}$ regardless of initial conditions.

Proof. Consider the auxiliary virtual system,

$$\dot{\mathbf{y}} = -\nabla f(\mathbf{y}) + k(\mathbf{z} - \mathbf{y}), \tag{3.2}$$

where $\mathbf{z}$ is an external input. Note that with $\mathbf{y} = \mathbf{x}^i$, we recover equation 3.1—that is, equation 3.2 admits the trajectories of each agent $\mathbf{x}^i$ as particular solutions. The Jacobian of equation 3.2 is given by

$$\mathbf{J} = -\nabla^2 f(\mathbf{y}) - k\mathbf{I}. \tag{3.3}$$

Equation 3.3 is symmetric and negative definite for $k > \bar{\lambda}$ for any external input $\mathbf{z}$. Because the individual $\mathbf{x}^i$ are particular solutions of this virtual system, contraction implies that for all i and j, $\|\mathbf{x}^i - \mathbf{x}^j\| \to 0$ exponentially. The contraction rate is given by $k - \bar{\lambda}$. □

This theorem motivates a definition.

Definition 5 (Global Exponential Synchronization). *We will say the agents in a distributed algorithm globally exponentially synchronize if they all converge to one another exponentially regardless of initial conditions.*

Theorem 3 gives a simple condition on the coupling gain k for synchronization of the individual agents in equation 3.1. Because $\mathbf{z}$ can represent any input, theorem 3 applies to any dynamics of the quorum variable: with $\mathbf{z} = \mathbf{x}^{\bullet}$, it applies to the QSGD algorithm, and with $\mathbf{z} = \tilde{\mathbf{x}}$, it applies to the EASGD algorithm. Under the assumption of a contracting deterministic system, we can use the stochastic contraction results in theorem 1 to bound the expected distance between individual agents in the stochastic setting.

Lemma 2. *Assume that $k > \bar{\lambda}$ and that $Tr(\boldsymbol{BB}^T) = Tr(\boldsymbol{\Sigma}) < C$ uniformly. Then, after exponential transients of rate $2(k - \bar{\lambda})$,*

$$\mathbb{E}\left[\sum_i \|\boldsymbol{x}^i - \boldsymbol{x}^{\bullet}\|^2\right] \leq \frac{(p-1)C\eta}{2b\,(k - \bar{\lambda})}, \tag{3.4}$$

where each $\mathbf{x}^i$ is a solution of equation 2.4 or 2.6.

Proof. Consider the systems for $i = 1, \ldots, p$,

$$d\mathbf{x}^i = \left(-\nabla f(\mathbf{x}^i) + k\left(\mathbf{z} - \mathbf{x}^i\right)\right) dt + \sqrt{\frac{\eta}{b}}\mathbf{B}(\mathbf{x}^i)d\mathbf{W}^i, \tag{3.5}$$

which reproduces equation 2.4 with $\mathbf{z} = \tilde{\mathbf{x}}$ and equation 2.6 with $\mathbf{z} = \mathbf{x}^{\bullet}$. Each solution $\mathbf{x}^i$ to equation 3.5 is a solution of the stochastic virtual system,

$$d\mathbf{y} = (-\nabla f(\mathbf{y}) + k\,(\mathbf{z} - \mathbf{y}))\,dt + \sqrt{\frac{\eta}{b}}\mathbf{B}(\mathbf{y})d\mathbf{W},$$

which has contracting deterministic part under the assumptions of the lemma and by theorem 3. For fixed i and j, applying the results of theorem 1 in the Euclidean metric leads to

$$\mathbb{E}\left[\|\mathbf{x}^i - \mathbf{x}^j\|^2\right] \leq \frac{C\eta}{b(k - \bar{\lambda})} \tag{3.6}$$

after exponential transients of rate $2(k - \bar{\lambda})$. Summing equation 3.6 over i and j leads to

$$\mathbb{E}\left[\sum_{i<j} \|\mathbf{x}^i - \mathbf{x}^j\|^2\right] \leq \frac{p\,(p-1)\,\eta C}{2b\,(k - \bar{\lambda})}.$$

Finally, as in Tabareau et al. (2010), we can rewrite

$$\sum_{i<j} \|\mathbf{x}^i - \mathbf{x}^j\|^2 = p \sum_i \|\mathbf{x}^i - \mathbf{x}^\bullet\|^2,$$

which proves the result. □

We will refer to equation 3.4 as a synchronization condition.

3.2 Reduction of Noise Due to Synchronization. We now provide a mathematical characterization of how synchronization reduces the amount of noise felt by the individual QSGD agents. The derivation follows the mathematical procedure first employed in Tabareau et al. (2010) in the study of neural oscillators.

Theorem 4 (The Effect of Synchronization on Stochastic Gradient Noise). *Let $x^\bullet_{k,p}(t)$ denote the center of mass trajectory of the continuous-time QSGD system 2.6 with coupling gain k and p agents. In the simultaneous limits $k \to \infty$ and $p \to \infty$, the difference between $\mathbf{x}^\bullet_{k,p}(t)$ and a trajectory of the noise-free dynamics,*

$$\dot{x}_{nf} = -\nabla f(x_{nf}), \tag{3.7}$$

tends to zero, $lim_{k\to\infty}\, lim_{p\to\infty} \|x^\bullet_{k,p} - x_{nf}\| \to 0$ almost surely, with $x_{nf}(0) = x^\bullet_{k,p}(0)$.

Proof. Summing the stochastic dynamics in equation 2.6 over p, we find

$$d\mathbf{x}^\bullet = \left[-\frac{1}{p} \sum_i \nabla f(\mathbf{x}^i) \right] dt + \sqrt{\frac{\eta}{bp^2}} \sum_i \mathbf{B}(\mathbf{x}^i) d\mathbf{W}^i. \tag{3.8}$$

To make clear the dependence of the dynamics on $\mathbf{x}^\bullet$, we define the disturbance term,

$$\boldsymbol{\epsilon} = -\frac{1}{p} \sum_i \nabla f(\mathbf{x}^i) + \nabla f(\mathbf{x}^\bullet),$$

so that we can rewrite equation 3.8 as

$$d\mathbf{x}^\bullet = \left[-\nabla f(\mathbf{x}^\bullet) + \boldsymbol{\epsilon} \right] dt + \sqrt{\frac{\eta}{bp^2}} \sum_i \mathbf{B}(\mathbf{x}^i) d\mathbf{W}^i. \tag{3.9}$$

Each term $\sqrt{\frac{\eta}{bp^2}}\mathbf{B}(\mathbf{x}^i) d\mathbf{W}^i$ is a gaussian random variable with covariance $\frac{\eta}{bp^2}\boldsymbol{\Sigma}(\mathbf{x}^i)$, and each $d\mathbf{W}^i$ is independent of all other $d\mathbf{W}^j$. Hence, the sum over the noise terms in equation 3.9 can also be written as a single gaussian random variable with covariance $\frac{\eta}{bp^2}\sum_i \mathbf{\Sigma}(\mathbf{x}^i)$,

$$\sqrt{\frac{\eta}{bp^2}}\sum_i \mathbf{B}(\mathbf{x}^i)d\mathbf{W}^i = \sqrt{\frac{\eta}{bp^2}}\mathbf{T}d\mathbf{W}, \tag{3.10}$$

where $\mathbf{T} = \mathbf{T}(\mathbf{x}^1, \ldots, \mathbf{x}^p)$ and $\mathbf{T}\mathbf{T}^T = \sum_i \mathbf{\Sigma}(\mathbf{x}^i)$. Equation 3.10 leads to an additional simplification of equation 3.9:

$$d\mathbf{x}^\bullet = \left[-\nabla f(\mathbf{x}^\bullet) + \boldsymbol{\epsilon}\right] dt + \sqrt{\frac{\eta}{bp^2}}\mathbf{T}d\mathbf{W}. \tag{3.11}$$

Equation 3.11 shows that the effect of the additive noise is eliminated as the number of agents $p \to \infty$.[3] We now let $\mathbf{F}_j$ denote the gradient of $(-\nabla f(\mathbf{x}))_j$, and we let $\mathbf{H}_j$ denote its Hessian. We apply the Taylor formula with integral remainder to $(-\nabla f(\mathbf{x}))_j$:

$$\begin{aligned}&\left(-\nabla f(\mathbf{x}^i)\right)_j + \left(\nabla f(\mathbf{x}^\bullet)\right)_j - \mathbf{F}_j^T(\mathbf{x}^\bullet)(\mathbf{x}^i - \mathbf{x}^\bullet)\\ &\quad = \int_0^1 (1-s)\left(\mathbf{x}^i - \mathbf{x}^\bullet\right)^T \mathbf{H}_j\left((1-s)\mathbf{x}^i + s\mathbf{x}^\bullet\right)\left(\mathbf{x}^i - \mathbf{x}^\bullet\right). \end{aligned}\tag{3.12}$$

Summing equation 3.12 over i and applying the assumed bound $\mathbf{H}_j \le Q\mathbf{I}$ leads to the inequality

$$\left| p\left(\nabla f(\mathbf{x}^\bullet)\right)_j - \sum_i \left(\nabla f(\mathbf{x}^i)\right)_j \right| \le \frac{Q}{2}\sum_i \|\mathbf{x}^i - \mathbf{x}^\bullet\|^2.$$

The left-hand side of the above inequality is $p|\boldsymbol{\epsilon}_j|$. Squaring both sides and summing over j provides a bound on $p^2\|\boldsymbol{\epsilon}\|^2$. Taking a square root of this bound, we find

$$\|\boldsymbol{\epsilon}\| \le \frac{\sqrt{n}Q}{2p}\sum_i \|\mathbf{x}^i - \mathbf{x}^\bullet\|^2,$$

[3] Indeed, the covariance $\frac{\eta}{bp^2}\sum_i \mathbf{\Sigma}(\mathbf{x}^i) \le \frac{\eta}{bp}\bar{\mathbf{\Sigma}}$, where $\bar{\mathbf{\Sigma}} = \max_i \mathbf{\Sigma}(\mathbf{x}^i)$ and the max and $\le$ are with respect to the positive semidefinite order. The covariance $\frac{\eta}{bp}\bar{\mathbf{\Sigma}}$ tends to zero as $p \to \infty$, so that gaussian random variables drawn from a distribution with this covariance will become increasingly concentrated around zero with increasing p. Because the true covariance $\frac{\eta}{bp^2}\mathbf{T}\mathbf{T}^T$ is less positive semidefinite, random variables drawn from the true distribution will also become concentrated around zero as $p \to \infty$.

where the factor of $\sqrt{n}$ originates from the sum over the components of $\boldsymbol{\epsilon}$. Performing an expectation over the noise $d\mathbf{W}(s)$ for all $s < t$ and using the synchronization condition in equation 3.4, we conclude that after exponential transients of rate $2(k - \bar{\lambda})$,

$$\mathbb{E}\left[\|\boldsymbol{\epsilon}\|\right] \leq \frac{(p-1)\sqrt{n}QC\eta}{4p\,(k-\bar{\lambda})\,b}. \tag{3.13}$$

The bound in equation 3.13 depends on the synchronization rate of the agents $k - \bar{\lambda}$, the dimensionality of space n, the bound on the third derivative of the objective Q, and the bound on the noise strength $\frac{\eta}{b}C$. In the limit of large p, the dependence on p becomes negligible. The expected effect of the disturbance term $\boldsymbol{\epsilon}$ tends to zero as the coupling gain k tends to infinity, corresponding to the fully synchronized limit.

By assumption 2 and theorem 1, as $k \to \infty$ and $p \to \infty$, the difference between trajectories of equation 3.11 and the unperturbed, noise-free system tends to zero almost surely, as the effects of both the stochastic disturbance $\boldsymbol{\epsilon}$ and the additive noise term are eliminated in this simultaneous limit. □

3.3 Discussion. Theorem 4 demonstrates that for distributed SGD algorithms, roughly speaking, the noise strength is set by the ratio parameter $\frac{\eta}{bp}$ at the expense of a distortion term, which tends to zero with synchronization. Whether this noise reduction is a benefit or a drawback for non-convex optimization depends on the problem at hand.

If the use of a stochastic gradient is purely as an approximation of the true gradient (e.g., due to single-node or single-GPU memory limitations), then synchronization can be seen as improving this approximation and eliminating undesirable noise while simultaneously parallelizing the optimization problem. The analysis in this section then gives rigorous bounds on the magnitude of noise reduction. The $\boldsymbol{\epsilon}$ term could be measured in practice to understand the empirical size of the distortion, and k could be increased until $\boldsymbol{\epsilon}$ tends approximately to zero and the noise is reduced to a desired level.

Many studies have reported the importance of stochastic gradient noise in deep learning, particularly in the context of generalization performance (Poggio et al., 2017; Zhu, Wu, Yu, Wu & Ma, 2018; Chaudhari & Soatto, 2018; Zhang et al., 2017). Furthermore, large batches are known to cause issues with generalization, and this has been hypothesized to be due to a reduction in the noise magnitude due to a higher b in the ratio $\frac{\eta}{b}$ (Keskar, Mudigere, Nocedal, Smelyanskiy, & Tang, 2016). In this context, reduction of noise may be undesirable, and one may be interested only in the parallelization of the problem. Our analysis then suggests choosing k high enough such that the quorum variable represents a meaningful average of the parameters, but low enough that the noise in the SGD iterations is not reduced. Indeed, in section 6, we will find the best generalization performance for low values of k that still result in convergence of the quorum variable. For deep networks, the level of synchronization for a given value of k will be both architecture and data set dependent.

The condition in theorem 3 is merely a sufficient condition for synchronization, and synchronization may occur for significantly lower values of k than predicted by contraction in the Euclidean metric. However, independent of when synchronization exactly occurs, so long as there is a fixed upper bound as in equation 3.4, the results in this section will apply with the corresponding estimate of $\mathbb{E}[\|\boldsymbol{\epsilon}\|]$.

3.4 Extension to Multiple Learning Rates. Our analysis can be extended to the case when each individual agent has a different learning rate η_i (or, equivalently, different batch size), and thus a different noise level. In effect, this is because each agent still follows the same dynamics, though with different integration errors and at a different rate. In this case, the synchronization condition, equation 3.4, is modified to

$$\mathbb{E}\left[\sum_i \|\mathbf{x}^i - \mathbf{x}^\bullet\|^2\right] \le \frac{C}{2pb(k-\bar{\lambda})} \sum_{i<j} \left(\eta_i + \eta_j\right),$$

so that

$$\mathbb{E}\left[\|\boldsymbol{\epsilon}\|\right] \le \frac{\sqrt{n}QC}{4p^2b(k-\bar{\lambda})} \sum_{i<j} \left(\eta_i + \eta_j\right). \tag{3.14}$$

The noise term $\sum_i \sqrt{\frac{\eta^i}{bp^2}}\mathbf{B}(\mathbf{x}^i)d\mathbf{W}^i$ becomes a sum of p independent gaussians, each with covariance $\frac{\eta^i}{bp^2}\boldsymbol{\Sigma}(\mathbf{x}^i)$, and can be written as a single gaussian random variable $\sqrt{\frac{1}{bp^2}}\mathbf{T}d\mathbf{W}$ with $\mathbf{T}\mathbf{T}^T = \sum_i \eta^i \boldsymbol{\Sigma}(\mathbf{x}^i)$. An analogous argument as given in section 3.2 shows that the effect of this additive noise will tend to zero as $p \to \infty$. This could allow, for example, for multiresolution optimization, where agents with larger learning rates may help avoid sharper local minima, saddle points, and flat regions of the parameter space, while agents with finer learning rates may help converge to robust local minima that generalize well. Standard learning rate schedules can also be applied agent-wise using the validation loss of individual agents rather than decreasing all learning rates using the validation loss of the quorum variable.

3.5 Extension to Momentum Methods. Our analysis can also be extended to momentum methods, modeled using the differential equation (Su et al., 2014),

$$\ddot{\mathbf{x}}^i + \mu(t)\dot{\mathbf{x}}^i + \nabla f(\mathbf{x}^i) = 0,$$

in component-wise form:

$$\begin{aligned}\dot{\mathbf{x}}_1^i &= \mathbf{x}_2^i,\\ \dot{\mathbf{x}}_2^i &= -\nabla f(\mathbf{x}_1^i) - \mu(t)\mathbf{x}_2^i.\end{aligned}$$

Coupling the agents in both position and velocity leads to the dynamics,

$$\dot{\mathbf{x}}_1^i = \mathbf{x}_2^i + k_1(\mathbf{x}_1^\bullet - \mathbf{x}_1^i), \tag{3.15}$$

$$\dot{\mathbf{x}}_2^i = -\nabla f(\mathbf{x}_1^i) - \mu(t)\mathbf{x}_2^i + k_2(\mathbf{x}_2^\bullet - \mathbf{x}_2^i), \tag{3.16}$$

where $\mathbf{x}_l^\bullet = \frac{1}{p}\sum_j \mathbf{x}_l^j$.

Lemma 3. *Consider the QSGD with momentum system given by equations 3.15 and 3.16. Assume that f is $\underline{\lambda}$-strongly convex and $\bar{\lambda}$-smooth. For $k_1 > \frac{1}{4\left(\inf_t \mu(t)+k_2\right)} max\left((1-\bar{\lambda})^2, (1-\underline{\lambda})^2\right)$, the individual $\mathbf{x}^i$ systems globally exponentially synchronize with rate ξ, where*

$$\xi \geq \frac{k_1 + \inf_t \mu(t) + k_2}{2} - \sqrt{\left(\frac{k_1 - (\inf_t \mu(t) + k_2)}{2}\right)^2 + \frac{\max\left((1-\bar{\lambda})^2, (1-\underline{\lambda})^2\right)}{4}}. \tag{3.17}$$

Proof. The virtual system,

$$\dot{\mathbf{y}}_1 = \mathbf{y}_2 + k_1(\mathbf{x}_1^\bullet - \mathbf{y}_1), \tag{3.18}$$

$$\dot{\mathbf{y}}_2 = -\nabla f(\mathbf{y}_1) - \mu(t)\mathbf{y}_2 + k_2(\mathbf{x}_2^\bullet - \mathbf{y}_2), \tag{3.19}$$

has system Jacobian,

$$\mathbf{J} = \begin{pmatrix} -k_1\mathbf{I} & \mathbf{I} \\ -\nabla^2 f(\mathbf{y}_1) & -\left(\mu(t)+k_2\right)\mathbf{I}\end{pmatrix},$$

and will be contracting for $(\inf_t \mu(t) + k_2)\,k_1 > \sup_{\mathbf{x}}\left(\sigma^2\left(\frac{1}{2}\left(-\nabla^2 f(\mathbf{x}) + \mathbf{I}\right)\right)\right)$, where $\sigma^2(\cdot)$ denotes the largest squared singular value (Wang & Slotine, 2005). Because $\mathbf{I} - \nabla^2 f$ is symmetric, the squared singular values are simply the squared eigenvalues. This leads to the condition $(\inf_t \mu(t) + k_2)\,k_1 > \frac{1}{4}\max\left((1-\bar{\lambda}^2), (1-\underline{\lambda})^2\right)$, which may be rearranged to yield the condition in the theorem.

Equations 3.18 and 3.19 also admit the $\mathbf{x}_l^i$ as particular solutions, so that the agents globally exponentially synchronize with a rate $\xi = |\lambda_{\max}(\mathbf{J})|$. The lower bound on ξ can be obtained by application of the result in Slotine (2003, example 3.8). □

Hence, a bound similar to equation 3.4 can be derived just as in lemma 2. Because the $\mathbf{x}_1^\bullet$ dynamics are linear and because the $\mathbf{x}_2^\bullet$ dynamics are nonlinear only through the gradient of the loss, assumption 1 does not need to be modified. For $\inf_t \mu(t) > 0$, k_2 can be set to zero, so that coupling is only through the position variables.

4 An Alternative View of Distributed Stochastic Gradient Descent

In this section, we connect the discussion of synchronization and noise reduction with the analysis in Kleinberg et al. (2018), which interprets SGD as performing gradient descent on a smoothed loss in expectation. Specifically, we show that the reduction of noise due to synchronization can be viewed as a reduction in the smoothing of the loss function. This provides further geometrical intuition for the effect of synchronization on distributed SGD algorithms. It furthermore sheds light as to why one may want to use low values of k to prevent noise reduction in learning problems involving generalization, where optimization of the empirical risk rather than the expected risk introduces spurious defects into the loss function that may be removed by sufficient smoothing.

Defining the auxiliary sequence $\mathbf{y}_t = \mathbf{x}_t - \eta\nabla f(\mathbf{x}_t)$ and comparing with equation 2.1 shows that $\mathbf{x}_{t+1} = \mathbf{y}_t - \frac{\eta}{\sqrt{b}}\boldsymbol{\zeta}_t$, yielding

$$\mathbf{y}_{t+1} = \mathbf{y}_t - \eta\nabla f\left(\mathbf{y}_t - \frac{\eta}{\sqrt{b}}\boldsymbol{\zeta}_t\right) - \frac{\eta}{\sqrt{b}}\boldsymbol{\zeta}_t,$$

so that

$$\mathbb{E}_{\boldsymbol{\zeta}_t}\left[\mathbf{y}_{t+1}\right] = \mathbf{y}_t - \eta\nabla\mathbb{E}_{\boldsymbol{\zeta}_t}\left[f\left(\mathbf{y}_t - \frac{\eta}{\sqrt{b}}\boldsymbol{\zeta}_t\right)\right].$$

This demonstrates that the $\mathbf{y}$ sequence performs gradient descent on the loss function convolved with the $\frac{\eta}{\sqrt{b}}$-scaled noise in expectation.[4] Using this argument, Kleinberg et al. (2018) show that SGD can converge to minimizers for a much larger class of functions than just convex functions, though the convolution operation can disturb the locations of the minima.

4.1 The Effect of Synchronization on the Convolution Scaling. The analysis in section 3 suggests that synchronization of the $\mathbf{x}^i$ variables should reduce the convolution prefactor for a $\mathbf{y}$ variable related to the center of

[4] Kleinberg et al. (2018) group the factor of $\sqrt{1/b}$ with the covariance of the noise.

mass, and we now make this intuition more precise for the QSGD algorithm. We have that

$$\Delta \mathbf{x}_t^i = -\eta \nabla f(\mathbf{x}_t^i) + \eta k(\mathbf{x}_t^\bullet - \mathbf{x}_t^i) - \frac{\eta}{\sqrt{b}} \boldsymbol{\zeta}_t^i,$$

so that

$$\Delta \mathbf{x}_t^\bullet = -\eta \nabla f(\mathbf{x}_t^\bullet) + \eta \boldsymbol{\epsilon}_t - \frac{\eta}{\sqrt{bp}} \boldsymbol{\zeta}_t,$$

with $\boldsymbol{\epsilon}_t = \nabla f(\mathbf{x}_t^\bullet) - \frac{1}{p}\sum_i \nabla f(\mathbf{x}_t^i)$ as usual. Define the auxiliary variable $\mathbf{y}_t^\bullet = \mathbf{x}_t^\bullet - \eta \nabla f(\mathbf{x}_t^\bullet)$, so that

$$\mathbf{x}_{t+1}^\bullet = \mathbf{y}_t^\bullet + \eta \boldsymbol{\epsilon}_t - \frac{\eta}{\sqrt{bp}} \boldsymbol{\zeta}_t. \tag{4.1}$$

Equation 4.1 can then be used to state

$$\begin{aligned} \mathbf{y}_{t+1}^\bullet &= \mathbf{y}_t^\bullet - \eta \nabla f(\mathbf{x}_{t+1}^\bullet) + \eta \boldsymbol{\epsilon}_t - \frac{\eta}{\sqrt{bp}} \boldsymbol{\zeta}_t, \\ &= \mathbf{y}_t^\bullet - \eta \nabla f\left(\mathbf{y}_t^\bullet - \frac{\eta}{\sqrt{bp}} \boldsymbol{\zeta}_t + \eta \boldsymbol{\epsilon}_t\right) + \eta \boldsymbol{\epsilon}_t - \frac{\eta}{\sqrt{bp}} \boldsymbol{\zeta}_t. \end{aligned}$$

Taylor-expanding the gradient term, we find

$$\begin{aligned} \nabla f\left(\mathbf{y}_t^\bullet - \frac{\eta}{\sqrt{bp}} \boldsymbol{\zeta}_t + \eta \boldsymbol{\epsilon}_t\right) &= \nabla f\left(\mathbf{y}_t^\bullet - \frac{\eta}{\sqrt{bp}} \boldsymbol{\zeta}_t\right) \\ &\quad + \eta \nabla^2 f\left(\mathbf{y}_t^\bullet - \frac{\eta}{\sqrt{bp}} \boldsymbol{\zeta}_t\right) \boldsymbol{\epsilon}_t + \mathcal{O}(\eta^2), \end{aligned}$$

which alters the discrete $y^\bullet$ update to

$$\begin{aligned} \Delta \mathbf{y}_t^\bullet &= -\eta \nabla f\left(\mathbf{y}_t^\bullet - \frac{\eta}{\sqrt{bp}} \boldsymbol{\zeta}_t\right) + \eta\left(1 - \eta \nabla^2 f\left(\mathbf{y}_t^\bullet - \frac{\eta}{\sqrt{bp}} \boldsymbol{\zeta}_t\right)\right) \boldsymbol{\epsilon}_t \\ &\quad - \frac{\eta}{\sqrt{bp}} \boldsymbol{\zeta}_t + \mathcal{O}(\eta^3). \end{aligned} \tag{4.2}$$

Equation 4.2 says that in expectation, $y^\bullet$ performs gradient descent on a convolved loss with noise scaling reduced by a factor of $\frac{1}{\sqrt{p}}$. The reduced scaling comes at the expense of the usual disturbance term $\boldsymbol{\epsilon}$, which decreases to zero with increasing synchronization in expectation over the noise ζ_s for

$s < t$. Equation 4.2 differs from the non-distributed case by an additional $\mathcal{O}(\eta^2)$ factor of the Hessian.

4.2 Discussion. To better understand the interplay of synchronization and noise in SGD, we can consider several limiting cases. Consider a choice of η corresponding to a fairly high noise level, so that the loss function is sufficiently smoothed for the iterates of SGD ($k = 0$) to avoid local minima, saddle points, and flat regions, but so that the iterates would not reliably converge to a desirable region of parameter space, such as a deep and robust minimum.

For $k \to \infty$ and p sufficiently large, the quorum variable will effectively perform gradient descent on a minimally smoothed loss and will converge to a local minimum of the true loss function close to its initialization. Due to the strong coupling, the agents will likely get pulled into this minimum, leading to convergence as if a single agent had been initialized using deterministic gradient descent at $\mathbf{x}^{\bullet}(t = 0)$, despite the high value of η.

With an intermediate value of k so that the agents remain in close proximity to each other, but not so strong that $\|\boldsymbol{\epsilon}\| \to 0$, the $\mathbf{x}$ variables will be concentrated around the minima of the smoothed loss (the coupling will pull the agents together, but because $\|\boldsymbol{\epsilon}\| \neq 0$, the smoothing will not be reduced in the sense of equation 4.2). The stationary distribution of SGD is thought to be biased toward concentration around degenerate minima of high volume (Banburski et al., 2019); the coupling force should thus amplify this effect and lead to an accumulation of agents in wider and deeper minima in which all agents can approximately fit. Eventually, if sufficiently many agents arrive in a single minimum, it will be extremely difficult for any one agent to escape, leading to a consensus solution chosen by the agents even at a high noise level.

4.3 Numerical Simulations in Non-Convex Optimization. In this section, we consider simulations on a model one-dimensional non-convex loss function, as well as one possible high-dimensional generalization. There are several goals of the discussion. The first is to show that the intuition presented in section 4.2 is correct. The second is to provide a setting where visualization of the loss function, its analytically smoothed counterpart, and the distribution of possible convergent points is straightforward. The third is to elucidate qualitative trends in distributed non-convex optimization as a function of k in low- and high-dimensional settings, and to show to what extent properties of the low-dimensional setting translate to the high-dimensional setting. We consider the loss function

$$f(x) = \frac{\left(x^4 - 4x^2 + \frac{1}{5}x + \frac{2}{5}\left(3\sin(20x) - \frac{7}{2}\sin(2\pi x) + \cos\left(\frac{10ex}{3}\right)\right)\right)}{F}, \tag{4.3}$$

where the sinusoidal oscillations in equation 4.3 introduce spurious local minima. The constant factor $F \in \mathbb{R}^+$ is used for numerical stability for a wider range of η values in order to reduce the large gradient magnitudes introduced by the high-frequency modes. We simulate the dynamics of QSGD using a forward Euler discretization,

$$x_{t+1}^i = x_t^i - \eta \nabla f(x^i) + \eta k(x_t^\bullet - x_t^i) - \eta \zeta_t^i \qquad i = 1 \ldots p, \tag{4.4}$$

with $f(x)$ given by equation 4.3. We include 1000 agents in each of 250 simulations per k value. Each simulation is allowed to run for 20,000 iterations with $\eta = .15$.[5] The corresponding distributions of final points, computed via a kernel density estimate, are plotted over a range of k values in Figure 1. In each panel of the figure, the true loss function is plotted in orange and the loss function convolved with the noise distribution is in blue. The loss functions are normalized so they can appear on the same scale as the distributions, and the y scale is thus omitted. The agents are initialized uniformly over the interval $[-3, 3]$, and each experiences an independent and identically distributed (i.i.d.) uniform noise term $\zeta_t^i \sim U(-1.5, 1.5)$ per iteration. F is fixed at 150.

In Figure 1a, there is no coupling and the distribution of final iterates for the agents is nearly uniform across the parameter space, with a slightly increased probability of convergence to the two deepest regions. The distribution of the quorum variable is sharply peaked around zero.[6] As k increases to $k = 0.4$ in Figure 1b, the agents concentrate around the wide basins of the convolved loss function and avoid the sharp local minima of the true loss function. The distribution for the quorum variable is similar, but is too wide to imply reliable convergence to a minimum with loss near the global optimum.

As k is increased further to $k = 0.8$ in Figure 1c and $k = 1.0$ in Figure 1d, performance increases significantly. The distribution of the agents is centered around the global optimum of the smoothed loss, and the distribution of the quorum variable is very sharp around the same minimum; this represents the regime in which the agents have chosen a consensus solution. As demonstrated by Figure 1a, this improved convergence is not possible with standard SGD. As k is increased again in Figures 1e and 1f, the coupling force becomes too great, and performance decreases. There is no

[5] We choose a relatively high value of η so that the convolved loss will be qualitatively different from the true loss to a degree that is visible by eye. This enables us to distinguish convergence to true minima from convergence to minima of the convolved loss. An alternative and equivalent choice would be to choose η smaller, with a correspondingly wider distribution of the noise.

[6] Note that without coupling, each agent performs basic SGD. Hence, the results in Figure 1a are equivalent to $p \times n$ single-agent SGD simulations, where n is the total number of simulations and p is the number of agents per simulation.

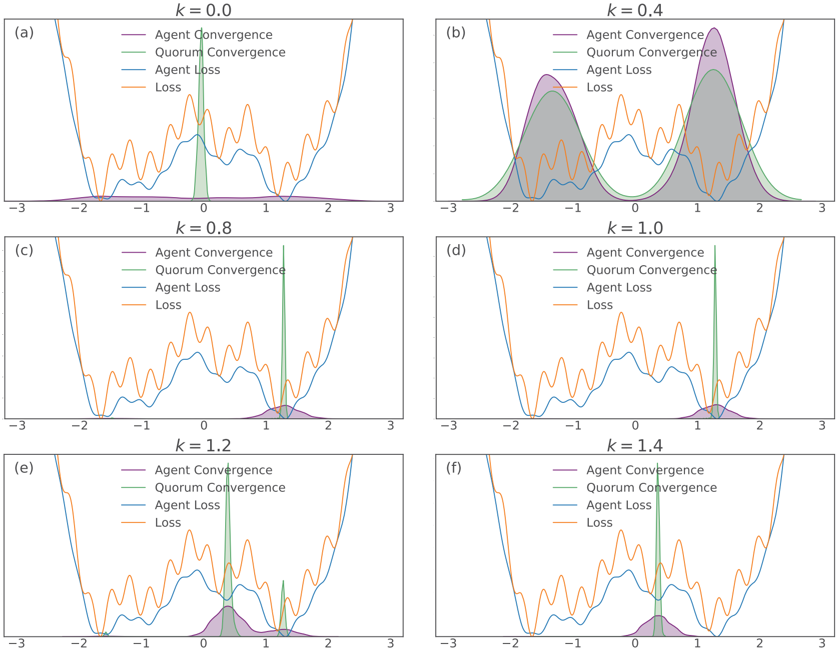

Figure 1: A demonstration of the effect of coupling in the high-noise regime. As the gain is increased, the agents transition from uniform convergence across parameter space, to sharply peaked convergence around deep minima of the smoothed loss, to convergence around minima of the smoothed loss near the initialization. The true loss is shown in orange, the smoothed loss is shown in blue, and the distributions of final iterates for the agents and the quorum variable are shown in purple and green, respectively. These simulations use a value of $\eta = 0.15$. Each plot contains the final iterates over 250 simulations with 20,000 iterations each and 1000 agents per simulation. The figure is best viewed in color.

initial exploratory phase to find the deeper regions of the landscape, and convergence is simply near the initialization of $x^{\bullet}$.

These simulation results suggest a useful combination of high noise, coupling, and traditional learning rate schedules. High noise levels can lead to rapid exploration and avoidance of problematic regions in parameter space, such as local minima, saddle points, or flat regions, while coupling can stabilize the dynamics toward a distribution around a wide and deep minimum of the convolved loss. The learning rate can then be decreased to improve convergence to minima of the true loss that lie within the spread of the distribution. In the uncoupled case, similar levels of noise would lead to a random walk.

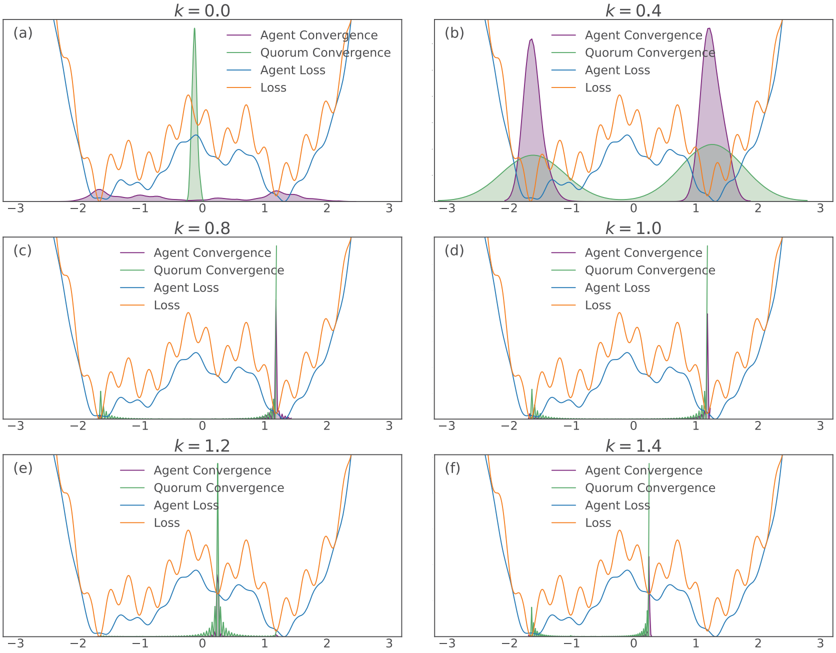

Figure 2: A demonstration of the effect of combining a learning rate schedule with coupling in the high-noise regime. The combination of coupling and learning rate scheduling significantly improves convergence for values of k that concentrate around the global optimum of the smoothed loss in the non-annealed case ($k = 0.8$ and $k = 1.0$), and the combination leads to sharp peaks around the minima of the true loss function. The true loss is shown in orange, the smoothed loss is shown in blue, and the distributions of final iterates for the agents and the quorum variable are shown in purple and green, respectively. These simulations use an initial learning rate of $\eta = 0.15$. Each plot contains the final iterates over 250 simulations with 20,000 iterations each and 1000 agents per simulation. The figure is best viewed in color.

This intuition is supported by the simulation results in Figure 2. The same simulation parameters are used, except the learning rate is now decreased by a factor of two every 4000 iterations until $\eta \leq 0.001$, where it is fixed. In the uncoupled case in Figure 2a, the schedule only slightly improves convergence around minima of the smoothed loss when compared to Figure 1a. Figure 2b again reflects a mild improvement relative to Figure 1b. For the two best values of $k = 0.8$ and $k = 1.0$ in Figures 2c and 2d, convergence of the agents and the quorum variable around the deepest minimum of the true loss that lies within the distribution of the agents in Figures 1c and 1d is excellent. In the very high k regime in Figures 2e and

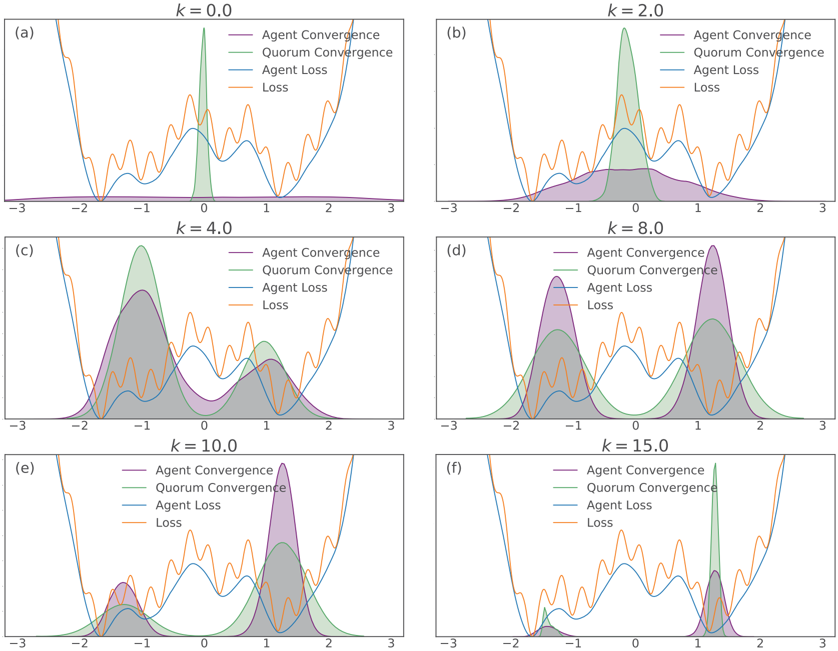

Figure 3: Simulations for the momentum method iteration given by equations 4.5 and 4.6 with $\eta = 0.1$ and $\delta = 0.9$. The true loss is shown in orange, the smoothed loss is shown in blue, and the distributions of final iterates for the agents and the quorum variable are shown in purple and green, respectively. The results are qualitatively similar to QSGD without momentum, except that higher k values are tolerated without degradation of performance. Each plot contains the final iterates over 250 simulations with 20,000 iterations each and 1000 agents per simulation. The figure is best viewed in color.

2f, the coupling force is too strong to enable exploration, and convergence is again near the initialization of $\mathbf{x}^\bullet$, but now to the minima of the true loss.

The preceding results also qualitatively apply to momentum methods. We now turn to simulate the following iteration

$$v_{t+1}^i = \delta v_t^i - \eta \nabla f(x_t^i + \delta v_t^i) - \eta \zeta_t^i, \tag{4.5}$$

$$x_{t+1}^i = x_t^i + v_{t+1}^i + \eta k \left(x_t^\bullet - x_t^i\right), \tag{4.6}$$

with the loss function again given by equation 4.3. The distributions of final iterates after 20,000 steps with $\eta = 0.1$, computed from 250 simulations per k value with 1000 agents per simulation, are shown in Figure 3.

Figure 3a is identical to Figure 1a except for the difference in learning rate: the agents converge uniformly across the parameter space. As k is increased to $k = 2$ in Figure 3b, the distribution of the agents becomes more localized around the center of parameter space but not around any minima. When k is increased to $k = 4$ in Figure 3c, $k = 8$ in Figure 3d, and $k = 10$ in Figure 3e, the distributions of the agents and the quorum variable become localized on the two deepest minima of the convolved loss but are still too wide for reliable convergence. The value $k = 15$ in Figure 3f leads to reliable convergence around the deep minimum on the right and would combine well with a learning rate schedule as in Figure 2. Overall, the trend is similar to the case without momentum, though much higher values of k are tolerated before degradation in performance. Despite high k values rapidly pulling the agent positions close to $\mathbf{x}^{\bullet}(t = 0)$, significant differences in the velocities of the agents prevent convergence to a local minimum nearby $\mathbf{x}^{\bullet}(t = 0)$ in the high k regime.

To demonstrate that these qualitative results also hold in higher dimensions, we now consider a d-dimensional objective function inspired by the one-dimensional objective function in equation 4.3. The loss function is given by

$$f(\mathbf{x}) = \left(\sum_{i=1}^{d} x_i^4 - 4x_i^2 + \frac{1}{5}x_i + \frac{2}{5}\left(\sum_{i,j=1}^{d} 3\sin(20x_i)\sin(20x_j) \right.\right.$$
$$\left.\left. + \cos\left(\frac{10e}{3}x_i\right)\cos\left(\frac{10e}{3}x_j\right) - \frac{7}{2}\sin(2\pi x_i)\sin(2\pi x_j)\right)\right)/F. \quad (4.7)$$

Equation 4.7 represents a separable sum of double well loss functions with pairwise sinusoidal coupling between all parameters. We include 1000 agents in each of 250 simulations per k value with $d = 250$. Each simulation is allowed to run for 10,000 steps with 1000 agents per simulation. The parameters are updated according to the vector forms of equations 4.5 and 4.6 with $\eta = .15$ and $\delta = .9$. No learning rate schedule is used. The agents are all randomly initialized uniformly in $[-4, 4] \times [-4, 4]$, and each experiences an i.i.d. noise term $\zeta_t^i \sim U(-.75, .75)$. F is fixed at 50.

For visualization purposes, we plot the contours of a two-dimensional cross section of the loss function by evaluating the last $d - 2$ coordinates at the value -1.2. This value was chosen to represent the bottom-left cluster apparent in Figures 5 and 6; it also lies close to the global minimum of the uncorrupted loss function $(-1.426, -1.426, \ldots, -1.426)^T \in \mathbb{R}^d$. Visualization of high-dimensional loss functions is difficult, and using such a cross section has its drawbacks; in particular, a saddle point may show up as a local minimum, correctly as a saddle point, or as a local maximum depending on the cross section taken. Nevertheless, the employed cross sections enable qualitative visualization of the clustering of the quorum variable and the individual agents and provide assurance that the general phenomena seen in one dimension in Figures 1 to 3 generalize naturally to higher dimensions.

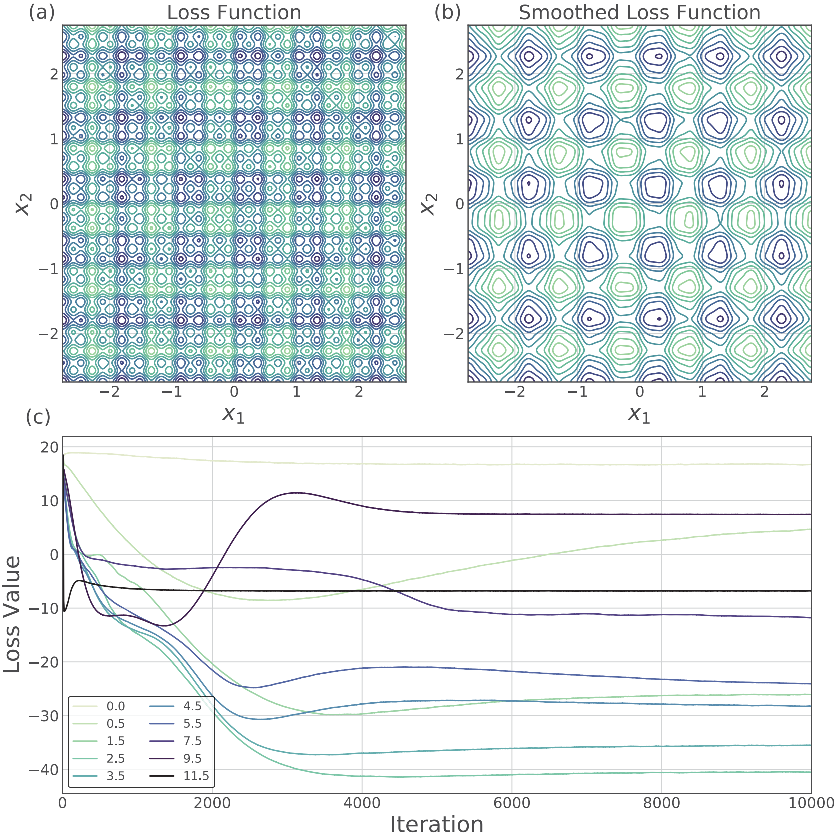

Figure 4: (a) A cross section of the loss function equation 4.7, evaluated at $x_i = -1.2$ for $i = 3, \ldots, d$. (b) The same cross section as in panel a, now of the smoothed loss function given by equation 4.7 convolved with the η-scaled uniform noise distribution. (c) The loss function value over time for the quorum variable, averaged over all simulations (see text), for a range of k values. The curves demonstrate that there is an optimum value of coupling, in this case around $k = 2.5$, for minimizing the loss function. The figure is best viewed in color.

The loss function itself is shown in Figure 4a, and the smoothed loss is shown in Figure 4b, which has significantly reduced complexity. Figure 4c displays the loss value of the quorum variable, averaged over all simulations, as a function of iteration number for a set of possible k values. The results are much the same as was described qualitatively in one dimension. Low values of k such as $k = 0$ and $k = 0.5$ do not successfully minimize the loss function as the agents are too spread out. Despite a significant ability to explore the loss landscape with such small coupling, the agents are not concentrated enough for $\mathbf{x}^\bullet$ to represent a meaningful average. As k increases, the ability to optimize the loss function at first significantly improves. While better than $k = 0$ and $k = 0.5$, $k = 1.5$ still represents the regime of too little coupling. $k = 2.5$ and $k = 3.5$ obtain much lower loss values than $k = 0$ and $k = 0.5$, with $k = 2.5$ achieving the lowest loss of the displayed k values. As k is increased further, performance starts to degrade. $k = 4.5$ performs worse than $k = 2.5$, and $k = 3.5$ obtains similar performance to $k = 1.5$. Increasing k to $k = 7.5$, $k = 9.5$, and $k = 11.5$ continues to deteriorate the ability of the algorithm to minimize the loss. The optimum k value represents, for the given noise level and loss function, the correct balance of exploration and resistance to noise.

As in the case of any algorithmic hyperparameter, it is natural to expect that there will be an optimum value of k. To see that the manifestation of this optimum is precisely a high-dimensional analog of the qualitative behavior observed in the one-dimensional simulations in Figures 1 to 3, we visualize the final points found by the quorum variable and a random selection of 25 agents per simulation in Figures 5 and 6, respectively, for a representative subset of the k values seen in Figure 4c.

Figure 5a shows that $k = 0$ results in essentially uniform convergence of the agents across the parameter space to local minima and saddle points, and hence the quorum variable simply converges near the origin in Figure 6a. The small amount of coupling $k = 0.5$ in Figure 5b leads to increased, but still insufficient, clustering of the agents. This manifests itself in Figure 6b as a shift of the ball of quorum convergence points toward the bottom left corner. $k = 1.5$ and $k = 2.5$ in Figures 5c and 5d have significantly improved convergence, with strong clustering of the agents in four balls around $(\pm 1.2, \pm 1.2)^T$. These clusters are located near the minima of the uncorrupted loss function, which occur at $(\pm 1.426, \pm 1.426, \ldots, \pm 1.426)^T$.

$k = 1.5$ and $k = 2.5$ have similar quorum convergence plots in Figures 6c and 6d, though the value of the loss in Figure 4c is noticeably different at iteration 10,000. The differences in the loss function values for the quorum variables are likely hidden by the low-dimensional visualization method. Figures 5c and 5d show that $k = 1.5$ has more "straggler" agents between the four corner clusters than $k = 2.5$, which may shift the quorum convergence points uphill. From a qualitative perspective, both are good choices for tracking minima of the uncorrupted or the non-smoothed loss functions and could be combined with a learning rate schedule to improve convergence from the cloud of "starting points" in Figures 5c and 5d.

As k is increased further to $k = 7.5$, the coupling begins to grow too strong. The distinct agent clusters attempt to merge, as seen in Figure 5e. The result of this is seen in Figure 6e, where there are scattered quorum convergence points between the clusters. Finally, for $k = 11.5$, the coupling is too great, and convergence of both the agents and the quorum variables in Figures 5f and 6f, respectively, are both near the origin.

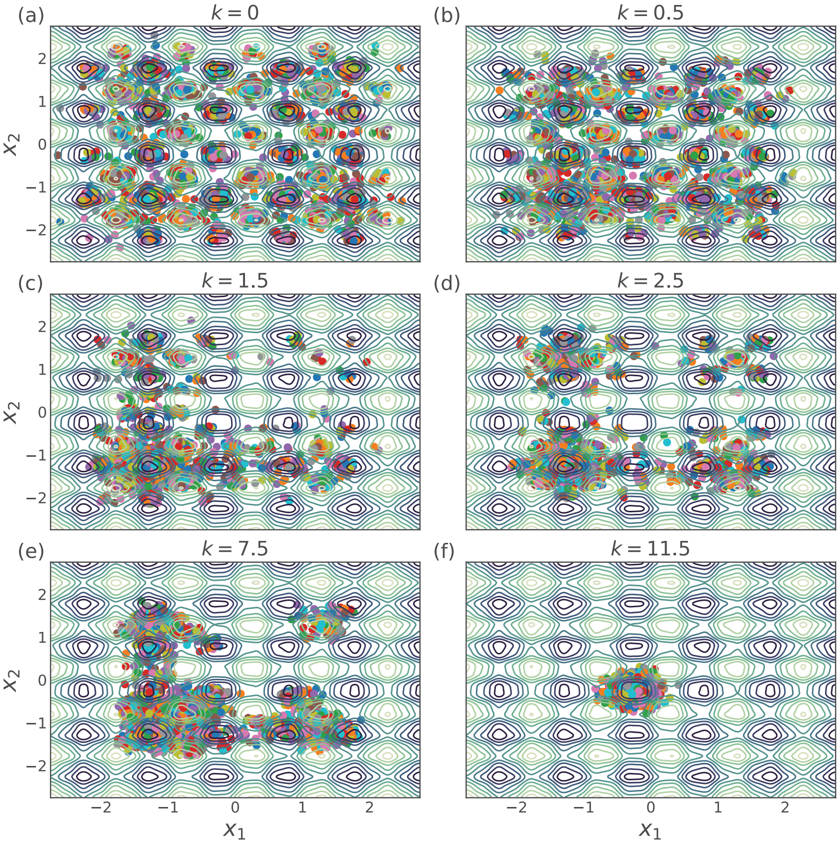

Figure 5: Contour plots displaying the location of 25 agents per simulation (multicolored dots) at the final iteration on top of the smoothed loss. See the text for the overall simulation setup. The figure is best viewed in color.

Taken together, Figures 1 to 6 provide significant qualitative insight into the convergence of distributed SGD algorithms, both with and without momentum. In one-dimensional and high-dimensional simulations, there is an optimum level of coupling that represents an ideal balance between the ability of the agents to explore the loss function and the concentration of the distribution of final iterates. Pushing k too high will lead to convergence near the initialization of $\mathbf{x}^{\bullet}$ and ultimately to reduced smoothing of the loss function, while setting k too low will lead to poor convergence of the quorum variable due to a lack of clustering of the agents. Intermediate values of k lead to concentration of the agents around deep and wide minima of the smoothed loss, which will generally lie close to the minima of the

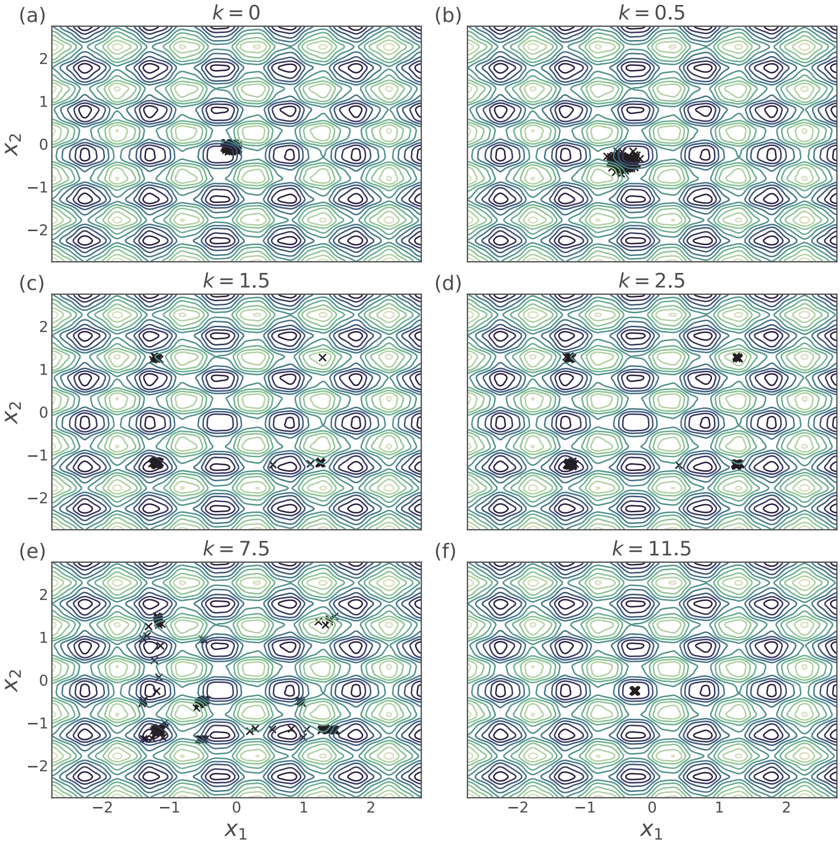

Figure 6: Contour plots displaying the location of the quorum variable (black x) in each simulation at the final iteration on top of the smoothed loss. See the text for the overall simulation setup. The figure is best viewed in color.

uncorrupted loss; convergence can be improved from here with a learning rate schedule.

The optimum value of k is set by the size of the gradients in comparison to the noise level. In the simulation setup used here, this corresponds to a trade-off between the value of F, which sets the gradient magnitudes, and the width of the noise distribution. By setting the width of the noise distribution very high, the optimum k value can be shifted to a large value, so that numerical stability issues arise before performance begins to degrade. Similarly, with small width and small F, the optimum value of k can be very small. In section 6, we will see a manifestation of a similar phenomenon in deep networks for the testing loss.

5 Convergence Analysis

We now provide contraction-based convergence proofs for QSGD and EASGD in the strongly convex setting. In the original work on EASGD, rigorous bounds were found for multivariate quadratic objectives in discrete-time, and the analysis for a general strongly convex objective was restricted to an inequality on the iteration for several relevant variances (Zhang et al., 2015). The results in this section thus extend previously available convergence results for EASGD and contain new results for QSGD. We furthermore present convergence results for QSGD with momentum.

A significant theme of this section is that the general methodology of theorem 4 can be applied to produce bounds on the expected distance of the quorum variable from the global minimizer of a strongly convex function, again split into a sum of two terms—one based on the averaged noise and one based on bounding the distortion vector $\boldsymbol{\epsilon}$. We also demonstrate in this section that an optimality result obtained for EASGD in discrete-time in Zhang et al. (2015) can be obtained through a straightforward application of stochastic calculus in continuous-time, and that the same result applies for QSGD.

5.1 QSGD Convergence Analysis. We first present a simple lemma describing the convergence of deterministic distributed gradient descent with arbitrary coupling.

Lemma 4. *Consider the all-to-all coupled system of ordinary differential equations,*

$$\dot{\boldsymbol{x}}^i = -\nabla f(\boldsymbol{x}^i) + \sum_{j \neq i} \left(\boldsymbol{u}\left(\boldsymbol{x}^j\right) - \boldsymbol{u}\left(\boldsymbol{x}^i\right) \right), \tag{5.1}$$

with $\boldsymbol{x}^i \in \mathbb{R}^n$ for $i = 1, \ldots, p$. Assume that $-\nabla f - p\boldsymbol{u}$ is contracting in some metric with rate λ_1, and that $-\nabla f$ is contracting in some (not necessarily the same) metric with rate λ_2. Then all $\boldsymbol{x}^i$ globally exponentially converge to a critical point of f.

Proof. Consider the virtual system:

$$\dot{\mathbf{y}} = -\nabla f(\mathbf{y}) - p\mathbf{u}(\mathbf{y}) + \sum_{j=1}^{p} \mathbf{u}(\mathbf{x}^j).$$

This system is contracting by assumption, and each of the individual agents is a particular solution. The agents therefore globally exponentially synchronize with rate λ_1. After this exponential transient, the dynamics of each agent is described by the reduced-order virtual system,

$$\dot{\mathbf{y}} = -\nabla f(\mathbf{y}).$$

By assumption, this system is contracting in some metric with rate λ_2 and has a particular solution at any critical point $\mathbf{x}^*$ such that $\nabla f(\mathbf{x}^*) = 0$. □

Remark 1. This simple lemma demonstrates that any form of coupling can be used so long as the quantity $-\nabla f(\mathbf{y}) - p\mathbf{u}(\mathbf{y})$ is contracting to guarantee exponential convergence to a critical point. A simple choice is $\mathbf{u}(\mathbf{x}^j) = \frac{k}{p}\mathbf{x}^j$, where k is the coupling gain, corresponding to balanced and equal-strength all-to-all coupling. Then equation 5.1 can be simplified to

$$\dot{\mathbf{x}}^i = -\nabla f(\mathbf{x}^i) + k\left(\mathbf{x}^\bullet - \mathbf{x}^i\right), \tag{5.2}$$

which is QSGD without noise. Note that all-to-all coupling can thus be implemented with only $2p$ directed connections by communicating with the center of mass variable.

Remark 2. If f is l-strongly convex, $-\nabla f$ will be contracting in the identity metric with rate l.

Remark 3. If f is locally l-strongly convex, $-\nabla f$ will be locally contracting in the identity metric with rate l. For example, for a nonconvex objective with initializations $\mathbf{x}^i(0)$ in a strongly convex region of parameter space, we can conclude exponential convergence to a local minimizer for each agent.

If f is strongly convex, the coupling between agents provides no advantage in the deterministic setting, as they would individually contract toward the minimum regardless. For stochastic dynamics, however, coupling can improve convergence. We now demonstrate the ramifications of the results in section 3 in the context of QSGD with the following theorem.

Theorem 5. *Consider the QSGD algorithm,*

$$d\boldsymbol{x}^i = \left(-\nabla f(\boldsymbol{x}^i) + k(\boldsymbol{x}^\bullet - \boldsymbol{x}^i)\right) dt + \sqrt{\frac{\eta}{b}}\boldsymbol{B}(\boldsymbol{x}^i)d\boldsymbol{W},$$

with $\boldsymbol{x}^i \in \mathbb{R}^n$ for $i = 1, \dots, p$. Assume that the conditions in assumption 1 hold, that $\boldsymbol{B}\boldsymbol{B}^T = \boldsymbol{\Sigma}$ is bounded such that $Tr(\boldsymbol{\Sigma}) \leq C$ uniformly, and that f is λ-strongly convex. Then, after exponential transients of rate λ and $\lambda + k$, the expected difference between the center of mass trajectory $\boldsymbol{x}^\bullet$ and the global minimizer $\boldsymbol{x}^*$ of f is given by

$$\mathbb{E}\left[\|\boldsymbol{x}^* - \boldsymbol{x}^\bullet\|\right] \leq \frac{Q(p-1)C\sqrt{n}\eta}{4pb\lambda(\lambda + k)} + \sqrt{\frac{\eta C}{2bp\lambda}}. \tag{5.3}$$

Proof. We first sum the dynamics of the individual agents to compute the dynamics of the center of mass variable. This leads to the SDE,

$$d\mathbf{x}^\bullet = (-\nabla f(\mathbf{x}^\bullet) + \boldsymbol{\epsilon})\, dt + \sqrt{\frac{\eta}{bp^2}}\mathbf{T}d\mathbf{W},$$

with $\boldsymbol{\epsilon} = \nabla f(\mathbf{x}^\bullet) - \frac{1}{p}\sum_i \nabla f(\mathbf{x}^i)$ and $\mathbf{T}\mathbf{T}^T = \sum_i \boldsymbol{\Sigma}(\mathbf{x}^i)$ defined exactly as in section 3. Consider the hierarchy of virtual systems:

$$\begin{aligned}
\dot{\mathbf{y}}^1 &= -\nabla f(\mathbf{y}^1),\\
\dot{\mathbf{y}}^2 &= -\nabla f(\mathbf{y}^2) + \boldsymbol{\epsilon}(\mathbf{x}^1, \ldots, \mathbf{x}^p),\\
d\mathbf{y}_t^3 &= \left(-\nabla f(\mathbf{y}^3) + \boldsymbol{\epsilon}(\mathbf{x}^1, \ldots, \mathbf{x}^p)\right) dt + \sqrt{\frac{\eta}{bp^2}}\mathbf{T}(\mathbf{x}^1, \ldots, \mathbf{x}^p) d\mathbf{W}.
\end{aligned}$$

The $\mathbf{y}^1$ system is contracting by assumption, and admits a particular solution $\mathbf{y}^1 = \mathbf{x}^*$. As in the proof of lemma 1, we can write with $R = \|\mathbf{y}^1 - \mathbf{y}^2\|$,

$$\dot{R} + \lambda R \leq \|\boldsymbol{\epsilon}\|, \tag{5.4}$$

which shows that R is bounded. Hence, by dominated convergence,

$$\dot{\overline{\mathbb{E}[R]}} + \lambda \mathbb{E}[R] \leq \mathbb{E}[\|\boldsymbol{\epsilon}\|]]. \tag{5.5}$$

As shown in section 3, $\mathbb{E}[\|\boldsymbol{\epsilon}\|] \leq \frac{Q(p-1)C\eta\sqrt{n}}{4p(\lambda+k)b}$ after exponential transients of rate $\lambda + k$.[7] Hence by lemma 1, the difference between the $\mathbf{y}^1$ and $\mathbf{y}^2$ systems can be bounded as

$$\mathbb{E}[\|\mathbf{y}^2 - \mathbf{x}^*\|] \leq \frac{Q(p-1)C\eta\sqrt{n}}{4p(\lambda+k)\lambda b}$$

after exponential transients of rate λ. The $\mathbf{y}^2$ system is contracting for any input $\boldsymbol{\epsilon}$, and the $\mathbf{y}^3$ system is identical with the addition of an additive noise term. By corollary 1, after exponential transients of rate λ,

$$\mathbb{E}[\|\mathbf{y}^3 - \mathbf{y}^2\|^2] \leq \frac{\eta C}{2bp\lambda}.$$

By Jensen's inequality and noting that $\sqrt{\cdot}$ is a concave, increasing function,

$$\mathbb{E}[\|\mathbf{y}^3 - \mathbf{y}^2\|] \leq \sqrt{\mathbb{E}[\|\mathbf{y}^3 - \mathbf{y}^2\|^2]} \leq \sqrt{\frac{\eta C}{2bp\lambda}}.$$

[7] In section 3, the denominator contained the factor $k - \bar{\lambda}$ rather than $k + \lambda$. Strong convexity of f was not assumed, so that the contraction rate of the coupled system was $k - \bar{\lambda}$. In this proof, strong convexity of f implies that the contraction rate of the coupled system is $k + \lambda$.

Finally, note that $\mathbf{x}^{\bullet}$ is a particular solution of the $\mathbf{y}^3$ virtual system. From these observations and an application of the triangle inequality, after exponential transients,

$$\mathbb{E}[\|\mathbf{x}^{\bullet} - \mathbf{x}^{*}\|] \leq \frac{Q(p-1)C\sqrt{n}\eta}{4pb\lambda(\lambda+k)} + \sqrt{\frac{\eta C}{2bp\lambda}}. \qquad \square$$

As in section 3, the bound 5.3 consists of two terms. The first term originates from a lack of complete synchronization and can be decreased by increasing k. The second term comes from the additive noise and can be decreased by increasing the number of agents. Both terms can be decreased by decreasing $\frac{\eta}{b}$, as this ratio sets the magnitude of the noise, and hence the size of both the disturbance and the noise term.

State- and time-dependent couplings of the form $k(\mathbf{x}^{\bullet}, t)$ are also immediately applicable with the proof methodology above. For example, increasing k over time can significantly decrease the influence of the first term in equation 5.3, leaving only a bound essentially equivalent to linear noise averaging. For non-convex objectives, this suggests choosing low values of $k(\mathbf{x}^{\bullet}, t)$ in the early stages of training for exploration and larger values near the end of training to reduce the variance of $\mathbf{x}^{\bullet}$ around a minimum. By the synchronization and noise argument in section 3 and the considerations in section 4, this will also have the effect of improving convergence to a minimum of the true loss function rather than the smoothed loss. If accessible, local curvature information could be used to determine when $\mathbf{x}^{\bullet}$ is near a local minimum and therefore when to increase k. Using state- and time-dependent couplings would change the duration of exponential transients, but the result in theorem 5 would still hold.

It is worth comparing equation 5.3 to a bound obtained with the same methodology for standard SGD. With the stochastic dynamics,

$$d\mathbf{x} = -\nabla f(\mathbf{x})dt + \sqrt{\frac{\eta}{b}}\mathbf{B}d\mathbf{W},$$

and the same assumptions as in theorem 5, the expected difference after exponential transients between a critical point of f and the stochastic $\mathbf{x}$ is given by corollary 1 and an application of Jensen's inequality as

$$\mathbb{E}\left[\|\mathbf{x} - \mathbf{x}^{*}\|\right] \leq \sqrt{\frac{\eta C}{2b\lambda}}.$$

In the distributed, synchronized case described by theorem 5, the deviation is reduced by a factor of $\frac{1}{\sqrt{p}}$ in exchange for an additional additive term. This additive term is related to the noise strength $\frac{C\eta}{b}$, the bound Q, and the number of parameters n, and is divided by $\lambda(\lambda + k)$—that is, it is smaller for more strongly convex functions and with more synchronized dynamics.

5.2 EASGD Convergence Analysis. We now incorporate the additional dynamics present in the EASGD algorithm. First, we prove a lemma demonstrating convergence to the global minimum of a strongly convex function in the deterministic setting.

Lemma 5. *Consider the deterministic continuous-time EASGD algorithm,*

$$\begin{aligned} \dot{\mathbf{x}}^i &= -\nabla f(\mathbf{x}^i) + k(\tilde{\mathbf{x}} - \mathbf{x}^i), \\ \dot{\tilde{\mathbf{x}}} &= kp\,(\mathbf{x}^\bullet - \tilde{\mathbf{x}})\,, \end{aligned}$$

with $\mathbf{x}^i \in \mathbb{R}^n$ for $i = 1, \ldots, p$. Assume f is λ-strongly convex. Then all agents and the quorum variable $\tilde{\mathbf{x}}$ globally exponentially converge to the unique global minimum $\mathbf{x}^*$ with rate

$$\gamma \geq \frac{\lambda + k + kp}{2} - \sqrt{\left(\frac{\lambda + k - kp}{2}\right)^2 + k^2 p}. \tag{5.6}$$

Proof. By theorem 3 and strong convexity of f, the individual $\mathbf{x}^i$ trajectories globally exponentially synchronize with rate $\lambda + k$. On the synchronized subspace, the system can be described by the reduced-order virtual system:

$$\begin{aligned} \dot{\mathbf{y}} &= -\nabla f(\mathbf{y}) + k(\tilde{\mathbf{x}} - \mathbf{y}), \\ \dot{\tilde{\mathbf{x}}} &= kp\,(\mathbf{y} - \tilde{\mathbf{x}})\,. \end{aligned}$$

The system Jacobian is then given by

$$\mathbf{J} = \begin{pmatrix} -\nabla^2 f(\mathbf{y}) - k\mathbf{I} & k\mathbf{I} \\ kp\mathbf{I} & -kp\mathbf{I} \end{pmatrix}.$$

When a metric transformation $\boldsymbol{\Theta} = \begin{pmatrix} \sqrt{p}\mathbf{I} & 0 \\ 0 & \mathbf{I} \end{pmatrix}$ is chosen, the generalized Jacobian becomes

$$\boldsymbol{\Theta}\mathbf{J}\boldsymbol{\Theta}^{-1} = \begin{pmatrix} -\nabla^2 f(\mathbf{y}) - k\mathbf{I} & \sqrt{p}k\mathbf{I} \\ \sqrt{p}k\mathbf{I} & -kp\mathbf{I} \end{pmatrix},$$

which is clearly symmetric. A sufficient condition for negative definiteness of this matrix is that $(\lambda + k)\,kp > k^2 p$ (Wang & Slotine, 2005; Horn & Johnson, 2012). Rearranging leads to the condition $\lambda > 0$, which is satisfied by the strong convexity of f. The virtual system is therefore contracting. Finally, note that $\mathbf{y} = \tilde{\mathbf{x}} = \mathbf{x}^*$, where $\mathbf{x}^*$ is the unique global minimum, is a particular solution. All trajectories thus globally exponentially converge to this minimum. The lower bound on the contraction rate in the statement of the theorem can be found by applying the result in Slotine (2003, example 3.8). □

Just as in theorem 5, we now turn to a convergence analysis for the EASGD algorithm using the results of lemma 5:

Theorem 6. *Consider the continuous-time EASGD algorithm,*

$$d\boldsymbol{x}^i = \left(-\nabla f\left(\boldsymbol{x}^i\right) + k\left(\tilde{\boldsymbol{x}} - \boldsymbol{x}^i\right)\right) dt + \sqrt{\frac{\eta}{b}}\boldsymbol{B}(\boldsymbol{x}^i)d\boldsymbol{W}^i$$

$$d\tilde{\boldsymbol{x}} = kp\left(\boldsymbol{x}^\bullet - \tilde{\boldsymbol{x}}\right) dt,$$

for $i = 1, \ldots, p$. Assume that f is λ-strongly convex and that the conditions in assumption 1 are satisfied. Let γ denote the contraction rate of the deterministic, fully synchronized EASGD system in the metric $\boldsymbol{M} = \boldsymbol{\Theta}^T\boldsymbol{\Theta}$ with $\boldsymbol{\Theta}$ the metric transformation from lemma 5, as lower bounded in equation 5.6. Further assume that $Tr(\boldsymbol{B}^T\boldsymbol{M}\boldsymbol{B}) \leq C(p)$ with C a positive constant potentially dependent on p through the dependence of $\boldsymbol{M}$ on p. Then, after exponential transients of rate γ and $\lambda + k$,

$$\mathbb{E}\left[\|\boldsymbol{z} - \boldsymbol{z}^*\|\right] \leq \frac{Q(p-1)C(p)\sqrt{n}\eta}{4b\sqrt{p}\gamma(\lambda + k)} + \sqrt{\frac{\eta C(p)}{2bp\gamma}}, \tag{5.7}$$

where $\boldsymbol{z} = (\boldsymbol{x}^\bullet, \tilde{\boldsymbol{x}})$ and $\boldsymbol{z}^* = (\boldsymbol{x}^*, \boldsymbol{x}^*)$.

Proof. Adding up the agent dynamics, the center of mass trajectory follows,

$$d\mathbf{x}^\bullet = \left(-\nabla f(\mathbf{x}^\bullet) + \boldsymbol{\epsilon} + k\left(\tilde{\mathbf{x}} - \mathbf{x}^\bullet\right)\right) dt + \sqrt{\frac{\eta}{bp^2}}\mathbf{T}d\mathbf{W},$$

with the usual definitions of $\boldsymbol{\epsilon}$ and $\mathbf{T}$. Consider the hierarchy of virtual systems:

$$\dot{\mathbf{y}}^1 = -\nabla f(\mathbf{y}^1) + k\left(\tilde{\mathbf{y}}^1 - \mathbf{y}^1\right),$$

$$\dot{\tilde{\mathbf{y}}}^1 = kp\left(\mathbf{y}^1 - \tilde{\mathbf{y}}^1\right),$$

$$\dot{\mathbf{y}}^2 = -\nabla f(\mathbf{y}^2) + k\left(\tilde{\mathbf{y}}^2 - \mathbf{y}^2\right) + \boldsymbol{\epsilon}(\mathbf{x}^1, \ldots, \mathbf{x}^p),$$

$$\dot{\tilde{\mathbf{y}}}^2 = kp\left(\mathbf{y}^2 - \tilde{\mathbf{y}}^2\right),$$

$$d\mathbf{y}^3 = \left(-\nabla f(\mathbf{y}^3) + k\left(\tilde{\mathbf{y}}^3 - \mathbf{y}^3\right)\right)dt + \boldsymbol{\epsilon}(\mathbf{x}^1, \ldots, \mathbf{x}^p) + \sqrt{\frac{\eta}{bp^2}}\mathbf{T}(\mathbf{x}^1, \ldots, \mathbf{x}^p)d\mathbf{W},$$

$$d\tilde{\mathbf{y}}^3 = kp\left(\mathbf{y}^3 - \tilde{\mathbf{y}}^3\right)dt.$$

The first system is contracting toward the unique global minimum with rate γ by the assumptions of the theorem and lemma 5. The second system is contracting for any external input $\boldsymbol{\epsilon}$, and we have already bounded $\mathbb{E}[\|\boldsymbol{\epsilon}\|]$ in section 3 (the application of the bound is independent of the dynamics of the quorum variable; see the appendix for details). Let $\mathbf{z}^i = \left(\mathbf{y}^i, \tilde{\mathbf{y}}^i\right)$ and $\mathbf{z}^* = \left(\mathbf{x}^*, \mathbf{x}^*\right)$. By an identical argument as in the proof of theorem 5 and noting that the condition number of $\boldsymbol{\Theta}$ is $\sqrt{p}$,

$$\mathbb{E}\left[\left\|\mathbf{z}^2 - \mathbf{z}^*\right\|\right] \leq \frac{\sqrt{p}}{\gamma}\mathbb{E}\left[\|\boldsymbol{\epsilon}\|\right] \leq \frac{Q(p-1)C(p)\eta\sqrt{n}}{4\sqrt{p}(\lambda+k)\gamma b}$$

after exponential transients of rate γ and $\lambda + k$. Note that $\lambda_{\min}(\mathbf{M}) = 1$. Hence we can take $\beta = 1$ in corollary 1 and

$$\mathbb{E}\left[\|\mathbf{z}^3 - \mathbf{z}^2\|\right] \leq \sqrt{\frac{\eta C(p)}{2bp\gamma}}$$

after exponential transients of rate γ. Combining these results via the triangle inequality and noting that $\mathbf{x}^\bullet$, $\tilde{\mathbf{x}}$ is a solution to the $\mathbf{y}^3$, $\tilde{\mathbf{y}}^3$ virtual system, we find that after exponential transients of rate γ,

$$\mathbb{E}\left[\|\mathbf{z} - \mathbf{z}^*\|\right] \leq \frac{Q(p-1)C(p)\sqrt{n}\eta}{4b\sqrt{p}\gamma(\lambda+k)} + \sqrt{\frac{\eta C(p)}{2bp\gamma}},$$

where $\mathbf{z} = \left(\mathbf{x}^\bullet, \tilde{\mathbf{x}}\right)$. □

Theorem 6 demonstrates an explicit bound on the expected deviation of both the center of mass variable $\mathbf{x}^\bullet$ and the quorum variable $\tilde{\mathbf{x}}$ from the global minimizer of a strongly convex function. As in the discussion after theorem 5, the results will still hold with state- and time-dependent couplings of the form $k = k(\tilde{\mathbf{x}}, t)$, and the same ideas suggested for QSGD based on increasing k over time can be used to eliminate the effect of the first term in the bound.

Theorem 6 is strictly weaker than theorem 5. The metric transformation used adds a factor of $\sqrt{p}$ to the first quantity in the bound, and the assumption $Tr(\mathbf{B}^T\mathbf{M}\mathbf{B}) \leq C(p)$ now depends on p through the factor of p in the top-left block of $\mathbf{M}$. Indeed, writing the matrix $\mathbf{B}$ in $n \times n$ block form, $Tr(\mathbf{B}^T\mathbf{M}\mathbf{B}) = C + (p-1)Tr(\mathbf{B}_{11}^T\mathbf{B}_{11} + \mathbf{B}_{12}^T\mathbf{B}_{12})$ where $C = Tr(\mathbf{B}^T\mathbf{B})$ as in theorem 5. Thus, the dependence of $C(p)$ on p is in general linear.

Because of this linear dependence on p, the first term in the bound scales like $p^{3/2}$, while the second is asymptotically independent of p. This is not the case in theorem 5, where the first term is asymptotically independent of p and the second term scales like $\frac{1}{p}$. The unfavorable scaling of the bound in theorem 6 with p implies that higher values of p do not improve convergence for EASGD as they do for QSGD. These issues can be avoided by reformulating lemma 5 in the Euclidean metric, but this leads to the fairly strong restriction $k < \frac{4\lambda p}{(p-1)^2}$.

These observations highlight potential convergence issues for EASGD with large p which are not present with QSGD. In line with these theoretical conclusions, we will empirically find stricter stability conditions on k for EASGD when compared to QSGD for training deep networks in section 6. Nevertheless, in the context of non-convex optimization, higher values of p can still lead to improved performance by affording increased parallelization of the problem and exploration of the landscape.

Less significantly, unlike in theorem 5, the bound in theorem 6 is applied to the combined vector $\mathbf{z}$ rather than the quorum variable $\tilde{\mathbf{x}}$ itself, and the contraction rate γ is used rather than λ in the virtual system bounds.[8] Both of these facts weaken the result when compared to theorem 5. γ will in general be less than λ, as exemplified by the lower bound, equation 5.6.

5.3 QSGD with Momentum Convergence Analysis. We now present a proof of convergence for the QSGD algorithm with momentum. We first prove a lemma demonstrating convergence to the global minimum of a strongly convex, $\bar{\lambda}$-smooth function. We consider the case of coupling only in the position variables; coupling additionally through the momentum variables is similar. We also restrict to the case of constant momentum coefficient for simplicity.

Lemma 6. *Consider the deterministic continuous-time QSGD with momentum algorithm*

$$\dot{x}_1^i = x_2^i + k\left(x^\bullet - x_1^i\right),$$
$$\dot{x}_2^i = -\nabla f(x_1^i) - \mu x_2,$$

with $\boldsymbol{x}_j^i \in \mathbb{R}^n$ *for* $i = 1, \ldots, p$. *Assume that* f *is* $\underline{\lambda}$*-strongly convex and* $\bar{\lambda}$*-smooth. For* $\mu > 2\sqrt{\underline{\lambda} + \bar{\lambda} - 2\sqrt{\underline{\lambda}\bar{\lambda}}}$ *and* $k > \frac{1}{4\mu} \max\left((1-\bar{\lambda})^2, (1-\underline{\lambda})^2\right)$*, all agents globally exponentially converge to the unique minimum with zero velocity,* $(\boldsymbol{x}_1^i, \boldsymbol{x}_2^i) \rightarrow$

[8] The factor of $\lambda + k$ in the first term remains, as this factor originates in the derivation of the bound on $\mathbb{E}\left[\|\boldsymbol{\epsilon}\|\right]$, where the synchronization rate is $\lambda + k$.

$(x^*, 0)$ *for all i. The exponential convergence rate* κ *can be lower-bounded as*

$$\kappa \geq \frac{\delta\mu + (1-\delta)\mu}{2} - \sqrt{\left(\frac{\delta\mu - (1-\delta)\mu}{2}\right)^2 + \frac{1}{4}\left(\frac{(\bar{\lambda} - \underline{\lambda})^2}{2\left(\underline{\lambda} + \bar{\lambda} + 2(\delta-1)\delta\mu^2\right)}\right)} \tag{5.8}$$

with $\delta = \delta(\mu) \in (0, 1)$.

Proof. By lemma 3 and according to the assumption on k, the agents will globally exponentially synchronize with rate ξ, where ξ may be lower bounded as in equation 3.17. On the synchronized subspace, the overall system can be described by the virtual system

$$\begin{aligned} \dot{\mathbf{x}}_1 &= \mathbf{x}_2, \\ \dot{\mathbf{x}}_2 &= -\nabla f(\mathbf{x}_1) - \mu\mathbf{x}_2, \end{aligned}$$

where the superscript has been omitted and the coupling term vanishes. Note that this system admits the particular solution $(\mathbf{x}_1, \mathbf{x}_2) = (\mathbf{x}^*, 0)$. This system has Jacobian

$$\mathbf{J} = \begin{pmatrix} 0 & \mathbf{I} \\ -\nabla^2 f & -\mu\mathbf{I} \end{pmatrix},$$

which is clearly not contracting. Define the metric transformation $\mathbf{\Theta} = \begin{pmatrix} a\mathbf{I} & 0 \\ \delta\mu\mathbf{I} & \mathbf{I} \end{pmatrix}$ with $0 < \delta < 1$ and $a \in \mathbb{R}$. The resulting symmetric part of the generalized Jacobian is given by

$$\left(\mathbf{\Theta}\mathbf{J}\mathbf{\Theta}^{-1}\right)_s = \begin{pmatrix} -\delta\mu\mathbf{I} & \frac{1}{2}\left(a\mathbf{I} - \frac{1}{a}\nabla^2 f - \frac{(\delta-1)\delta}{a}\mu^2\mathbf{I}\right) \\ \frac{1}{2}\left(a\mathbf{I} - \frac{1}{a}\nabla^2 f - \frac{(\delta-1)\delta}{a}\mu^2\mathbf{I}\right) & (\delta-1)\mu \end{pmatrix}.$$

For contraction, we require that

$$\delta(1-\delta)\mu^2 > \frac{1}{4}\max\left(\left(a - \frac{\delta(\delta-1)}{a}\mu^2 - \frac{\underline{\lambda}}{a}\right)^2, \left(a - \frac{\delta(\delta-1)}{a}\mu^2 - \frac{\bar{\lambda}}{a}\right)^2\right).$$

Choosing

$$a = \sqrt{\frac{1}{2}\left(\underline{\lambda} + \bar{\lambda} + 2(\delta - 1)\delta\mu^2\right)} \tag{5.9}$$

ensures that the two arguments of the max are equal. For a to be real, we require that $\mu < \sqrt{\frac{\underline{\lambda}+\bar{\lambda}}{2(1-\delta)\delta}}$. The condition for contraction then reads that

$$4\delta(1-\delta)\mu^2 > \left(\frac{(\bar{\lambda} - \underline{\lambda})^2}{2\left(\underline{\lambda} + \bar{\lambda} + 2(\delta - 1)\delta\mu^2\right)}\right),$$

leading to the condition on μ,

$$\frac{1}{2}\sqrt{\frac{\underline{\lambda} + \bar{\lambda} - 2\sqrt{\underline{\lambda}\bar{\lambda}}}{\delta(1-\delta)}} < \mu < \min\left(\frac{1}{2}\sqrt{\frac{\underline{\lambda} + \bar{\lambda} + 2\sqrt{\underline{\lambda}\bar{\lambda}}}{\delta(1-\delta)}}, \sqrt{\frac{\underline{\lambda} + \bar{\lambda}}{2(1-\delta)\delta}}\right).$$

The lower bound is always real and positive by the arithmetic-geometric mean inequality. There is always a gap between the lower and upper bound, regardless of which argument of the min is chosen in the upper bound. The lower bound is minimized for $\delta = \frac{1}{2}$, leading to the condition that $\mu > 2\sqrt{\underline{\lambda} + \bar{\lambda} - 2\sqrt{\bar{\lambda}\underline{\lambda}}}$. With μ satisfying this minimal lower bound, the valid range of μ can be shifted arbitrarily large by choosing

$$\delta(\mu) = \left(\frac{\sqrt{\mu^2 + 4\left(2\sqrt{\bar{\lambda}\underline{\lambda}} - (\bar{\lambda} + \underline{\lambda})\right)} + \mu}{2\mu} - \alpha\right) \in (0, 1)$$

with $\alpha > 0$ an arbitrarily small positive constant, thus eliminating the upper bound. The lower bound on the contraction rate κ of the system can be obtained by application of the result in Slotine (2003, example 3.8). □

Note that in general, so long as μ is chosen to satisfy the lower bound of the preceding lemma, the QSGD with momentum system will be contracting in some metric. The given metric will depend on the value of $\delta(\mu)$—for example, chosen as suggested in the proof.

With Lemma 6 in hand, we can now state a convergence result for QSGD with momentum.

Theorem 7. *Consider the continuous-time QSGD with momentum algorithm,*

$$d\boldsymbol{x}_1^i = \big(\boldsymbol{x}_2^i + k(\boldsymbol{x}^\bullet - \boldsymbol{x}_1^i)\big)\,dt,$$

$$d\boldsymbol{x}_2^i = \big(-\nabla f(\boldsymbol{x}_1^i) - \mu\boldsymbol{x}_2\big)\,dt + \sqrt{\frac{\eta}{b}}\boldsymbol{B}(\boldsymbol{x}_1^i)d\boldsymbol{W}^i,$$

for $i = 1, \ldots, p$. Assume that the conditions of lemma 6 and assumption 2 are satisfied. Let κ denote the contraction rate of the deterministic, fully synchronized QSGD with momentum system as lower bounded in equation 5.8 and let ξ denote the synchronization rate of the QSGD with momentum system as lower bounded in equation 3.17. Further assume that $Tr(\boldsymbol{B}^T\boldsymbol{M}\boldsymbol{B}) \le C$ with $C > 0$ where $\boldsymbol{M} = \boldsymbol{\Theta}^T\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}$ is the metric transformation from lemma 6. Let $\psi = \frac{1}{2}(1 + a^2 + \delta^2\mu^2 - \sqrt{(1 + a^2 + \delta^2\mu^2)^2 - 4a^2})$ denote the minimum eigenvalue of $\boldsymbol{M}$ with a given by equation 5.9 and let $\Psi = \frac{1}{2}(1 + a^2 + \delta^2\mu^2 + \sqrt{(1 + a^2 + \delta^2\mu^2)^2 - 4a^2})$ denote the maximum eigenvalue. Then, after exponential transients of rate κ and ξ, with $\boldsymbol{z} = (\boldsymbol{x}_1^\bullet, \boldsymbol{x}_2^\bullet)$ and $\boldsymbol{z}^* = (x^*, 0)$

$$\mathbb{E}\left[\|\boldsymbol{z} - \boldsymbol{z}^*\|\right] \le \frac{Q\sqrt{\Psi}(p-1)C\sqrt{n}\eta}{\sqrt{\psi}4bp\kappa\xi} + \sqrt{\frac{\eta C}{2bp\psi\kappa}}. \tag{5.10}$$

Proof. Summing the agent dynamics, the center of mass trajectory follows

$$d\mathbf{x}_1^\bullet = \mathbf{x}_2^\bullet dt$$

$$d\mathbf{x}_2^\bullet = (-\nabla f(\mathbf{x}_1^\bullet) + \boldsymbol{\epsilon} - \mu\mathbf{x}_2^\bullet)\,dt + \sqrt{\frac{\eta}{bp^2}}\mathbf{T}(\mathbf{x}^1, \ldots, \mathbf{x}^p)d\mathbf{W}$$

with the usual definition of $\boldsymbol{\epsilon}$ and $\mathbf{T}$. Consider an analogous hierarchy of virtual systems as in theorems 5 and 6:

$$\dot{\mathbf{y}}_1^1 = \mathbf{y}_2^1,$$

$$\dot{\mathbf{y}}_2^1 = -\nabla f(\mathbf{y}_1^1) - \mu\mathbf{y}_2^1,$$

$$\dot{\mathbf{y}}_1^2 = \mathbf{y}_2^2,$$

$$\dot{\mathbf{y}}_2^2 = -\nabla f(\mathbf{y}_1^2) - \mu\mathbf{y}_2^2 + \boldsymbol{\epsilon}(\mathbf{x}_1^1, \ldots, \mathbf{x}_1^p),$$

$$d\mathbf{y}_1^3 = \mathbf{y}_2^3 dt,$$

$$d\mathbf{y}_2^3 = \big(-\nabla f(\mathbf{y}_1^3) + \boldsymbol{\epsilon}(\mathbf{x}_1^1, \ldots, \mathbf{x}_1^p) - \mu\mathbf{y}_2^3\big)\,dt + \sqrt{\frac{\eta}{bp^2}}\mathbf{T}(\mathbf{x}^1, \ldots, \mathbf{x}^p)d\mathbf{W}.$$

The first system is contracting toward the global minimum with zero velocity and will arrive after exponential transients of rate κ by the assumptions of the theorem and by lemma 6. The second system is contracting for any external input $\boldsymbol{\epsilon}$, and as argued in section 3, the bound on $\mathbb{E}\left[\|\boldsymbol{\epsilon}\|\right]$ can be applied as is to the momentum system with a suitable replacement of contraction rates. As in theorem 5, and noting that the condition number of $\boldsymbol{\Theta}$ is $\sqrt{\frac{\Psi}{\psi}}$,

$$\mathbb{E}\left[\|\mathbf{z}^2 - \mathbf{z}^*\|\right] \leq \frac{(p-1)C\eta\sqrt{n}Q\sqrt{\Psi}}{4p\xi\kappa b\sqrt{\psi}},$$

after exponential transients of rate κ and ξ. Similarly, an application of corollary 2 gives

$$\mathbb{E}\left[\|\mathbf{z}^3 - \mathbf{z}^2\|\right] \leq \sqrt{\frac{\eta C}{2bp\psi\kappa}}$$

after exponential transients of rate κ, where we have noted that $\mathbf{x}^T\mathbf{M}\mathbf{x} \geq \psi\|\mathbf{x}\|^2$. An application of the triangle inequality leads to the result. □

Equation 5.10 is similar to the results for EASGD and QSGD. The bound is closer in spirit to the bound for QSGD without momentum, in that the two terms do not have poor dependencies on p as they do for EASGD. However, the statement of the theorem is complicated by the expressions for the contraction rates κ and ξ, the expressions for the minimum and maximum eigenvalues of the metric ψ and Ψ, and the expression for a in the metric transformation. Together, these four quantities create a more complex dependence of the bound on hyperparameters such as μ and k. Nevertheless, the spirit is still the same as theorem 5, in that the first term originates from the $\boldsymbol{\epsilon}$ disturbance and can be eliminated with synchronization, while the second term originates from the additive noise and can be eliminated by including additional agents.

5.4 Extensions to Other Distributed Structures. Similar results can be derived for many other possible distributed structures in an identical manner. We present one general formalism here, involving local state- and time-dependent couplings.

Lemma 7. *The state-dependent all-to-all coupled system,*

$$\dot{x}^i = -\nabla f(x^i) + \sum_j k_j(x^j, t)(x^j - x^i), \quad i = 1, \ldots, p, \tag{5.11}$$

will globally exponentially synchronize with rate

$$\underset{x^1,\ldots,x^p,t}{inf}\left\{\sum_j k_j(x^j,t)\right\} - \underset{y}{sup}\left\{\lambda_{max}\left(-\nabla^2 f(y)\right)\right\} \tag{5.12}$$

whenever this value is positive.

Proof. The weighted sum $\sum_j k_j(\mathbf{x}^j,t)\mathbf{x}^j$ now plays the role of the quorum variable, so that one has

$$\dot{\mathbf{x}}^i = -\nabla f(\mathbf{x}^i) - \sum_j k_j(\mathbf{x}^j,t)\,\mathbf{x}^i + \sum_j k_j(\mathbf{x}^j,t)\mathbf{x}^j \quad i = 1,\ldots,p. \tag{5.13}$$

The virtual system,

$$\dot{\mathbf{y}} = -\nabla f(\mathbf{y}) - \sum_j k_j(\mathbf{x}^j,t)\,\mathbf{y} + \sum_j k_j(\mathbf{x}^j,t)\mathbf{x}^j,$$

shows that the individual $\mathbf{x}^i$ trajectories globally exponentially synchronize if the conditions of the theorem are met. □

We note that the condition 5.12 is independent of the number of agents. With noise, the center of mass of equation 5.11 satisfies

$$d\mathbf{x}^\bullet = (-\nabla f(\mathbf{x}^\bullet) + \boldsymbol{\epsilon})dt + \sqrt{\frac{\eta}{pb}}\mathbf{B}d\mathbf{W},$$

where now $\boldsymbol{\epsilon} = \nabla f(\mathbf{x}^\bullet) - \frac{1}{p}\sum_i \nabla f(\mathbf{x}^i) + \sum_j k_j(\mathbf{x}^j,t)\mathbf{x}^j - \mathbf{x}^\bullet \sum_j k_j(\mathbf{x}^j,t)$. As usual, $\boldsymbol{\epsilon} \to 0$ in the fully synchronized state.

Individually state-dependent couplings of the form 5.11 or its quorum-mediated equivalent, equation 5.13, allow for individual gain schedules that depend on local cost values or other local performance measures. This can allow each agent to broadcast its current measure of success and shape the quorum variable accordingly. For example, the classification accuracy on a validation set for each $\mathbf{x}^i$ could be use to select the current best parameter vectors and increase the corresponding k_i values to pull other agents toward them.

5.5 Specialization to a Multivariate Quadratic Objective. In the original discrete-time analysis of EASGD in Zhang et al. (2015), it was proven that iterate averaging (Polyak & Juditsky, 1992) of $\tilde{\mathbf{x}}$ leads to an optimal variance around the minimum of a quadratic objective. We now derive an identical result in continuous-time for the QSGD algorithm, demonstrating that this optimality is independent of the additional dynamics in the EASGD algorithm.

For a multivariate quadratic $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$ with $\mathbf{A}$ symmetric and positive definite, the stochastic dynamics of each agent can be written as

$$d\mathbf{x}^i = \left(-\mathbf{A}\mathbf{x}^i + k\left(\mathbf{x}^\bullet - \mathbf{x}^i\right)\right)dt + \mathbf{B}d\mathbf{W}^i.$$

To make the optimal result clearer, we group the factor of $\sqrt{\frac{\eta}{b}}$ into the definition of $\mathbf{B}$, unlike in previous sections. We furthermore relax the state dependence of $\mathbf{B}$ in this section and assume it to be a constant matrix; this matches the case handled in Zhang et al. (2015).

The assumption of state-independence can be justified in several ways. Theoretical analyses have demonstrated that the specific form of positive semidefinite $\mathbf{B}$ does not affect the $\mathcal{O}(\eta)$ weak accuracy of the approximating stochastic differential equation 2.2 for SGD (Feng et al., 2018; Hu et al., 2017; Li et al., 2018), though it does affect the constant.[9] For relevance to general non-convex optimization, we can assume that all agents have arrived sufficiently close to a minimum of the loss function that it can be approximately represented as a quadratic and that the noise covariance is approximately constant (Mandt et al., 2016, 2017). For deep networks, the noise covariance has been empirically shown to align with the Hessian of the loss (Sagun, Evci, Guney, Dauphin, & Bottou, 2017; Zhu et al., 2018), with theoretical justification for when this is valid provided in appendix A of Jastrzębski et al. (2017). For all agents in an approximately quadratic basin of a local minimum of a deep network, $\mathbf{B}$ can then be taken to be constant such that $\mathbf{B}\mathbf{B}^T = \mathbf{A}$, where $\mathbf{A}$ is the approximately state-independent Hessian.

With this assumption, $\mathbf{x}^\bullet$ satisfies

$$d\mathbf{x}^\bullet_t = -\mathbf{A}\mathbf{x}^\bullet_t dt + \frac{1}{\sqrt{p}}\mathbf{B}d\mathbf{W}.$$

This is a multivariate Ornstein-Uhlenbeck process with solution

$$\mathbf{x}^\bullet(t) = e^{-\mathbf{A}t}\mathbf{x}^\bullet(0) + \frac{1}{\sqrt{p}}\int_0^t e^{-\mathbf{A}(t-s)}\mathbf{B}d\mathbf{W}_s. \tag{5.14}$$

By assumption, $-\mathbf{A}$ is negative definite, so that the stationary expectation $\lim_{t\to\infty}\mathbb{E}[\mathbf{x}^\bullet(t)] = 0$. The stationary variance $\mathbf{V}$ is given by

$$\mathbf{A}\mathbf{V} + \mathbf{V}\mathbf{A}^T = \frac{1}{p}\boldsymbol{\Sigma}$$

[9] The state-dependent version used earlier in this work has been empirically shown to have a lower constant (Li et al., 2018), and is closer to the $\mathcal{O}(\eta^2)$ approximating SDE, which is why it has been used up to this point.

(see, e.g., Gardiner (2009, p. 107). We now define

$$\mathbf{z}(t) = \frac{1}{t}\int_0^t \mathbf{x}^\bullet(t')dt',$$

and can immediately state the following lemma:

Lemma 8. *The averaged variable $\boldsymbol{z}(t)$ converges weakly to a normal distribution with mean zero and standard deviation $\frac{1}{p}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-T}$:*

$$\lim_{t\to\infty} \sqrt{t}\left(\boldsymbol{z}(t) - \boldsymbol{x}^*\right) \to \mathcal{N}\left(0, \frac{1}{p}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-T}\right).$$

In particular, for the single-variable case with $\boldsymbol{A} = h$ and $\boldsymbol{\Sigma} = \sigma^2$,

$$\lim_{t\to\infty} \sqrt{t}\left(z(t) - x^*\right) \to \mathcal{N}\left(0, \frac{\sigma^2}{ph^2}\right).$$

Proof. From equation 5.14,

$$\sqrt{t}\mathbf{z}(t) = \frac{1}{\sqrt{t}}\left(\mathbf{A}^{-1}\left(1 - e^{-\mathbf{A}t}\right)\mathbf{x}^\bullet(0)\right) + \frac{1}{\sqrt{tp}}\int_0^t dt' \int_0^{t'} e^{-\mathbf{A}(t'-s)}\mathbf{B}d\mathbf{W}_s,$$

the mean of which is asymptotically zero. In computing the variance, only the stochastic integral remains. Interchanging the order of integration,

$$\begin{aligned}\int_0^t \int_0^{t'} e^{-\mathbf{A}(t'-s)}\mathbf{B}d\mathbf{W}_s dt' &= \int_0^t \int_s^t e^{-\mathbf{A}(t'-s)}dt'\mathbf{B}d\mathbf{W}_s,\\ &= \mathbf{A}^{-1}\int_0^t \left(1 - e^{-\mathbf{A}(t-s)}\right)\mathbf{B}d\mathbf{W}_s.\end{aligned}$$

After an application of Ito's isometry, the variance is given by

$$\begin{aligned}&\mathbb{V}\left[\sqrt{t}\mathbf{z}(t)\right] =\\ &\frac{\mathbf{A}^{-1}}{tp}\left(\int_0^t \left(\boldsymbol{\Sigma} - e^{-\mathbf{A}(t-s)}\boldsymbol{\Sigma} - \boldsymbol{\Sigma}e^{-\mathbf{A}^T(t-s)} + e^{-\mathbf{A}(t-s)}\boldsymbol{\Sigma}e^{-\mathbf{A}^T(t-s)}\right)ds\right)\mathbf{A}^{-T}.\end{aligned}$$

In the limit, the only nonvanishing quantity after the computation of the integral is the linear term $\boldsymbol{\Sigma}t$. Then,

$$\lim_{t\to\infty} \mathbb{V}\left[\sqrt{t}\mathbf{z}(t)\right] = \frac{1}{p}\mathbf{A}^{-1}\boldsymbol{\Sigma}\mathbf{A}^{-T}. \tag{5.15}$$

□

As in the discrete-time EASGD analysis, equation 5.15 is optimal in the sense of achieving the Fisher information lower bound and is independent of the coupling strength k (Polyak & Juditsky, 1992; Zhang et al., 2015). The lack of dependence on the coupling k is less surprising in this case, as it is not present in the $\mathbf{x}^{\bullet}$ dynamics. The optimality of this result, together with the comparison of theorems 5 and 6, suggests that the extra $\tilde{\mathbf{x}}$ dynamics may not provide any benefit over coupling simply through the spatial average variable $\mathbf{x}^{\bullet}$ from the perspective of convex optimization. However, in section 6, we will show through numerical experiments on deep networks that EASGD tends to find networks that generalize better than QSGD. The benefits of EASGD must then go beyond basic optimization, and the extra dynamics may have a regularizing effect.

We can also make a slightly stronger statement about equation 5.15, as in Mandt et al. (2017).[10] If we precondition the stochastic gradients for each agent by the same constant invertible matrix $\mathbf{Q}$, then the stationary variance remains optimal. To see this, note that we can account for this preconditioning simply by modifying the derivation so that $\mathbf{A} \rightarrow \mathbf{QA}$ and $\mathbf{B} \rightarrow \mathbf{QB}$. Then,

$$\begin{aligned}\lim_{t\to\infty} \mathbb{V}\left[\sqrt{t}\mathbf{z}(t)\right] &= \frac{1}{p}(\mathbf{QA})^{-1}(\mathbf{QB})(\mathbf{QB})^{T}(\mathbf{QA})^{-T},\\ &= \frac{1}{p}\mathbf{A}^{-1}\mathbf{Q}^{-1}\mathbf{QBB}^{T}\mathbf{Q}^{T}\mathbf{Q}^{-T}\mathbf{A}^{-T},\\ &= \frac{1}{p}\mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-T}.\end{aligned}$$

If different agents are preconditioned by different matrices $\mathbf{Q}^{i}$, this result will not hold. Using adaptive algorithms based on past iterations for each agent such as AdaGrad (Duchi, Hazan, & Singer, 2011) thus may eliminate the optimality, as each agent would compute a different preconditioner.

6 Deep Network Simulations

We now turn to evaluate EASGD, QSGD, and one possible state-dependent variant of QSGD, equation 5.13, as learning algorithms for training deep neural networks on the CIFAR-10 data set. A significant goal of the section is to understand the role of synchronization and noise in training deep neural networks. We also seek to test the extensions proposed throughout this article, such as multiple learning rates, synchronization bounds

[10] A similar continuous-time analysis for the averaging scheme considered here was performed in Mandt et al. (2017) for the non-distributed case; the derivation here is simpler and provides asymptotic results.

allowing for independent initial conditions of the agents, and state-dependent coupling.

We obtain two primary results. The first is that less synchronization, when it still leads to reliable convergence of the quorum variable, results in the best generalization capabilities of the learned network. This is similar to the results of the model experiments performed in section 4.3, though those experiments revealed this to be true for general optimization rather than generalization. The observation of better generalization with reduced synchronization is in line with the comments of section 3.3 regarding noise and generalization in deep networks.

Our second primary result is the observation of an interesting regularizing property of EASGD, even in the single-agent case. Unlike QSGD with a single agent, EASGD does not reduce to standard SGD. We find that EASGD without momentum outperforms SGD with momentum and EASGD with momentum in the nondistributed setting.

6.1 Experimental Setup. We use a three-layer convolutional neural network based on the experiments in Zhang et al. (2015); each layer consists of a two-dimensional convolution, an ReLU nonlinearity, 2×2 max-pooling with a stride of two, and BatchNorm (Ioffe & Szegedy, 2015) with batch statistics in both training and evaluation. The first convolutional layer has kernel size nine, the second has kernel size five, and the third has kernel size three. All convolutions use a stride of one and zero padding. Following the three convolutional layers, there is a single fully connected layer to which we apply dropout with a probability of 0.5. The input data are normalized to have mean zero and standard deviation one in each channel in both the training and test sets. Because we are interested in qualitative trends rather than state-of-the-art performance, we do not employ any data augmentation strategies. We use an 80/20 training/validation set split, and we use the cross-entropy loss. The stochastic gradient is computed using minibatches of size 128. The learning rate is set to $\eta = 0.05$ initially unless otherwise specified. This value was chosen as the highest initial value of η that remained stable throughout training for most values of p, and the qualitative trends presented here were robust to the choice of learning rate (further simulations demonstrating this robustness are available in the supplemental information). We decrease the learning rate three times when the validation loss stalls:[11] first by a factor of five, then a factor of two the second and third times. This is done on an agent basis: the agents are allowed

[11] More precisely, we keep track of the validation loss for each agent at a reference point, beginning with the validation loss at the first epoch. If the validation loss at the next epoch changes by greater than 1% of the reference point, the reference loss is set to the newly computed validation loss. If the validation loss changes by less than 1%, the reference point is unchanged. When the reference point has been unchanged for five epochs, we decrease the learning rate.

to maintain different learning rates. Because we are focused on the behavior of the algorithms rather than efficiency from the standpoint of a parallel implementation, the agents communicate with the quorum variable after each update.

In all methods, we use a Nesterov-based momentum scheme unless otherwise specified (Nesterov, 1983, 2004) with a momentum parameter $\delta = 0.9$ unless otherwise specified and coupling only in the position variables. For EASGD, this takes the form (Zhang et al., 2015),

$$\begin{aligned}
\mathbf{v}_{t+1}^{i} &= \delta \mathbf{v}_{t}^{i} - \eta_i \mathbf{g}(\mathbf{x}_t^i + \delta \mathbf{v}_t^i), \\
\mathbf{x}_{t+1}^{i} &= \mathbf{x}_t^i + \mathbf{v}_{t+1}^i + \eta_i k(\tilde{\mathbf{x}}_t - \mathbf{x}_t^i), \\
\tilde{\mathbf{x}}_{t+1} &= \tilde{\mathbf{x}}_t + k \sum_i \eta_i (\mathbf{x}_t^i - \tilde{\mathbf{x}}_t),
\end{aligned}$$

where $\mathbf{g}$ is the stochastic gradient. The equivalent form for QSGD can be obtained by the replacement $\tilde{\mathbf{x}}_t \to \mathbf{x}_t^{\bullet}$ and by dropping the dynamics for $\tilde{\mathbf{x}}$. The update for SD-QSGD is similar:

$$\begin{aligned}
\mathbf{v}_{t+1}^{i} &= \delta \mathbf{v}_{t}^{i} - \eta_i \mathbf{g}(\mathbf{x}_t^i + \delta \mathbf{v}_t^i), \\
\mathbf{x}_{t+1}^{i} &= \mathbf{x}_t^i + \mathbf{v}_{t+1}^i + \eta_i \left(\sum_j k_j(\mathbf{x}^j, t)\mathbf{x}^j - \mathbf{x}^i \sum_j k_j(\mathbf{x}^j, t) \right).
\end{aligned}$$

In SD-QSGD, we use state-dependent gains $k_j = k_j(\mathbf{x}^j, t)$ inspired by a spiking winner-take-all formalism (Denève & Machens, 2016; Wang & Slotine, 2006). At the start of each epoch, we find the agent with the current minimum validation loss value. Denoting the index of this agent by j^*, we define

$$k_{j*} = \frac{k}{p} + (Mp - 1)\frac{k}{p} e^{-t/\tau}, \tag{6.1}$$

$$k_j = \frac{k}{p} \left(\frac{e^{t/\tau} - 1}{e^{t_f/\tau} - 1} \right) \quad \text{for } k_j \neq k_{j^*}, \tag{6.2}$$

for $t < t_f$, with k, τ, t_f and $M \geq 1$ fixed constants, and where t is reset to zero at the start of each epoch. Equations 6.1 and 6.2 shape the quorum variable to be entirely composed of the single best agent instantaneously at the start of an epoch. The constant M is a magnification factor and sets the size of the force all other agents feel in the direction of the best agent. The gains relax exponentially back to the QSGD formalism, which is obtained when $k_j = k/p$ for all j. The constant τ sets the speed of relaxation, and t_f defines the duration of the spike. At $t = t_f$, all k_j will have relaxed back to the original value $\frac{k}{p}$ for all $j \neq j^*$, and with proper choice of τ, k_{j^*} will be very close. We introduce a small discontinuity measured by the magnitude of $(Mp-1)\frac{k}{p}e^{-t_f/\tau}$ and simply set $k_{j^*}=\frac{k}{p}$ at $t=t_f$. We use a value of $M=10$, choose $t_f=N_b/4$ where N_b is the number of batches in an epoch, and choose $\tau=t_f/16$, corresponding to a rather rapid spike.[12]

In each of the following simulations, the fully connected weights and biases are initialized randomly and uniformly $W_{ij},\ b_i\sim\mathcal{U}(-\frac{1}{\sqrt{m}},\frac{1}{\sqrt{m}})$ where m is the number of inputs. The convolutional weights use Kaiming initialization (He, Zhang, Ren, & Sun, 2015). In each comparison, the methods are initialized from the same points in parameter space, but the agents are not required to be initialized at the same location. In QSGD and SD-QSGD, the quorum variable is exponentially weighted $\bar{\mathbf{x}}_{t+1}=\gamma\mathbf{x}^{\bullet}_t+(1-\gamma)\bar{\mathbf{x}}_t$ with $\gamma=.1$, and we test the convergence of $\bar{\mathbf{x}}$. Note that because this variable is not coupled to the dynamics of the individual agents, this is still distinct from EASGD. Because we use momentum in nearly all experiments, we will refer simply to QSGD and EASGD. The non-momentum variant of EASGD, when used, will be referred to as EASGD-WM (EASGD without momentum).

6.2 Experimental Results. We first analyze the effect of k on classification performance. We find that the best performance is obtained for the lowest possible fixed values of k that still lead to convergence of the quorum variable. This is demonstrated in Figure 7 for the EASGD algorithm with $\eta=0.05$ initially and $p=8$, where we observe the general trend that test accuracy improves as the coupling gain is decreased. Note that $k=0.01$ and $k=0.02$, as well as $k=0$ (not shown), have too little synchronization for the quorum variable to reflect a meaningful average, and hence do not lead to good performance. Similar results hold for QSGD (not shown). We found not only the best performance for low, fixed k but also the best scaling with the number of agents.[13]

There are several plausible explanations for the observation of improved generalization with reduced coupling. Lower values of k allow for greater exploration of the optimization landscape, which intuitively should lead to better performance. As the measure of synchronization in Figure 7d tends to zero, the $\boldsymbol{\epsilon}$ term in the $\mathbf{x}^{\bullet}$ dynamics will also tend to zero, and synchronization will begin to reduce the amount of noise felt by the individual agents. In neural networks, it is expected that this noise reduction will favor convergence to minima that do not generalize as well as those obtained with higher amounts of noise, as seen in Figure 7c.

[12] Another option would be to set $k_j=(k-k_{j^*})/(p-1)$ when this is positive and zero otherwise. This ensures, outside of the initial spiking period, that the total sum of the k_j is constant. We found similar empirical results with both choices.

[13] The improvement in test accuracy and in the minimization of the test loss with increasing number of agents is demonstrated in later plots. We found that this trend was maximized with lower values of k.

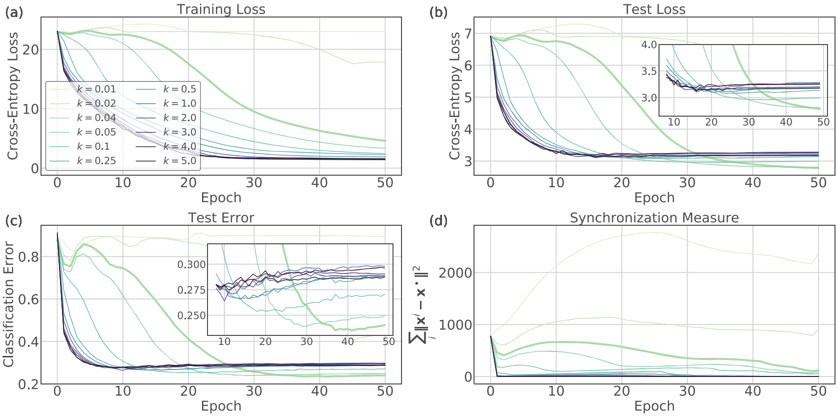

Figure 7: The effect of varying k on the learning procedure for the EASGD algorithm with $\eta = 0.05$ initially and $p = 8$. In general, lower test errors and lower test loss values are seen for lower values of k so long as convergence is still obtained. $k = 0.01$ and $k = 0.02$ have too little synchronization for the quorum variable to represent a meaningful average. Insets display a more finely grained view near the end of learning. The best-performing curve is shown in bold. The figure is best viewed in color.

Results for a comparison of QSGD and SD-QSGD are shown in Figure 8 for $p = 1, 4, 8, 16, 32$, and 64 with $k = 0.04$. QSGD is shown in solid lines, while SD-QSGD is shown in dashed; color indicates the number of agents (see the key in Figure 8a). Note that $p = 1$ simply corresponds to SGD for both SD-QSGD and QSGD, as the coupling term vanishes for a single agent. In both cases, we see significant improvement in accuracy as the number of agents increases, most likely due to an improved ability of the agents to explore the landscape, along with a decrease in synchronization. The test loss and test error curves display interesting differences between the two algorithms; for $p = 8$ and $p = 16$, the state-dependent formalism obtains mildly improved generalization relative to QSGD, as expected by the bias toward minima with lower validation loss. QSGD performs better for $p = 32$ and $p = 64$; SD-QSGD does not converge for $p = 64$.

We display a comparison of QSGD and EASGD in Figure 9, again for $k = 0.04$. QSGD tends to decrease the training loss further and more rapidly than EASGD; this is in line with earlier comments that, from an optimization perspective, the extra dynamics of the quorum variable offer no clear theoretical benefit. However, consistently across all experiments except for $p = 16$ where it does not converge, EASGD generalizes better: the test loss is driven lower, and the test accuracy is higher than QSGD. A particularly interesting result is the single-agent case, where EASGD actually performs

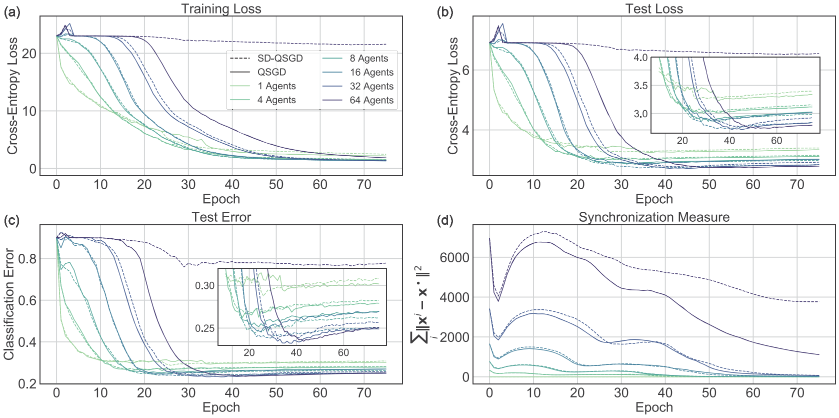

Figure 8: A comparison of SD-QSGD using a spiking winner-take-all formalism (see the text) to QSGD with a value of $k = 0.04$. The state-dependent formalism obtains improved accuracy for the intermediate values of $p = 8$ and $p = 16$. QSGD and SD-QSGD perform similarly for $p = 4$, and QSGD performs better for $p = 32$. SD-QSGD does not converge for $p = 64$ while QSGD does. Insets display a more finely grained view near the end of learning. The figure is best viewed in color.

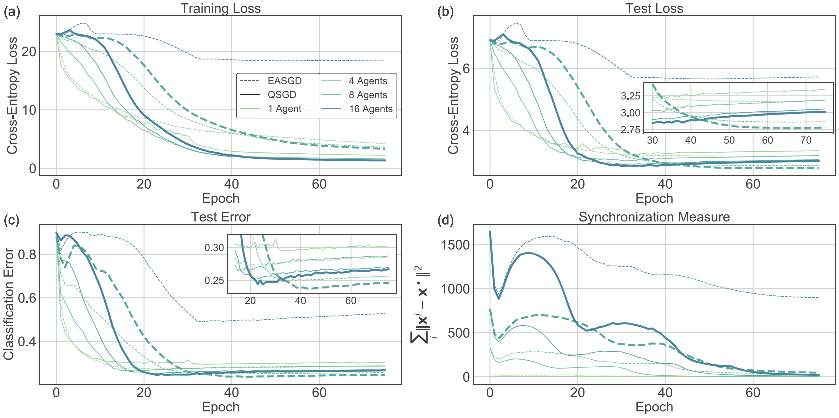

Figure 9: A comparison of the performance of QSGD to EASGD with $k = 0.04$ (see the text). QSGD optimizes the training loss further and faster than EASGD, leading to overfitting. The two algorithms respond differently to fixed k and have different levels of synchronization. For $p = 16$, EASGD fails to converge, though QSGD continues to converge. Nevertheless, for fewer agents, EASGD obtains improved performance. Insets display a more finely grained view near the end of learning. The best-performing curves for each algorithm are shown in bold. The figure is best viewed in color.

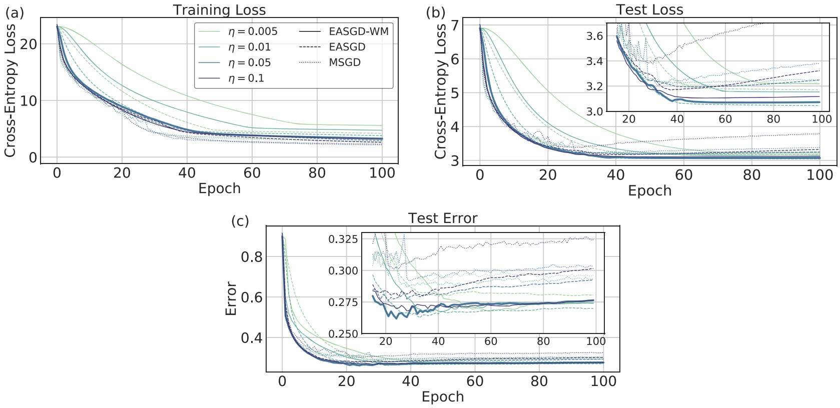

Figure 10: A comparison of EASGD, EASGD without momentum (EASGD-WM) and SGD with momentum (MSGD) over a range of learning rates for momentum parameter $\delta = 0.9$ and coupling gain $k = 0.054$. Surprisingly, EASGD and EASGD-WM perform better than MSGD in general, and in many cases, EASGD-WM performs better than EASGD. This motivates considering alternative dynamics for the quorum variable even for non-distributed optimization. Insets display a more finely grained view near the end of learning. The best-performing curve is shown in bold. The figure is best viewed in color.

better than SGD with momentum.[14] These observations suggest that the extra dynamics of the quorum variable may impose a form of implicit regularization that, to our knowledge, has not been observed before.

Motivated by this observation, we now compare the $p = 1$ EASGD algorithm with momentum, without momentum, and basic SGD with momentum in Figure 10 across a range of initial learning rates. Each algorithm is initialized from the same location, and each curve represents an average over three runs to eliminate stochastic variability. The momentum algorithms use $\delta = 0.9$, and the two EASGD variants use $k = 0.054$. In general, EASGD with and without momentum (dashed and solid lines, respectively) both achieve higher test accuracy than SGD with momentum (dotted lines). Surprisingly, EASGD without momentum often performs better than EASGD with momentum.

To show that this trend is not an artifact of incorrectly choosing the momentum parameter, we have compiled additional data in Table 1 over a range of momentum parameters and learning rates. Each data point reported is again the result of an average over three independent runs, and

[14] Note that unlike QSGD with a single agent, EASGD with a single agent is a different algorithm from basic SGD. It can be seen as SGD coupled in feedback to a low-pass filter of its output.

Table 1: Comparison of Minimum Test Loss Achieved and Minimum Error Achieved for EASGD-WM, EASGD, and MSGD on the CIFAR-10 Data Set (Each with $p = 1$, Providing Details on the Effect of Hyperparameter Choices Not Seen in Figure 10).

		Minimum Test Loss						Minimum Error					
		$\delta = .1$	$\delta = .25$	$\delta = .5$	$\delta = .75$	$\delta = .9$	$\delta = .99$	$\delta = .1$	$\delta = .25$	$\delta = .5$	$\delta = .75$	$\delta = .9$	$\delta = .99$
$\eta = .005$	EASGD-WM	**4.25**	**4.26**	**4.28**	4.29	3.22	3.24	**.267**	**.266**	**.264**	**.269**	**.268**	**.270**
	EASGD	4.72	4.83	4.56	**4.28**	**3.17**	**3.11**	.304	.313	.301	.282	.280	.277
	MSGD	4.75	4.87	4.64	4.33	3.21	3.29	.310	.323	.306	.286	.292	.295
$\eta = .01$	EASGD-WM	**4.09**	**5.15**	**4.03**	**4.12**	3.15	**3.10**	**.262**	**.259**	**.253**	**.261**	.267	**.257**
	EASGD	4.57	5.75	4.27	4.14	**3.04**	3.22	.297	.300	.280	.275	**.263**	.283
	MSGD	4.59	5.81	4.48	4.46	3.22	3.33	.300	.307	.294	.294	.287	.301
$\eta = .05$	EASGD-WM	**3.95**	**3.96**	**3.86**	**3.97**	**3.07**	**3.00**	**.252**	**.258**	**.250**	**.255**	**.262**	**.253**
	EASGD	4.41	4.27	4.05	4.06	3.19	4.04	.286	.283	.265	.267	.276	.417
	MSGD	4.46	4.57	4.48	4.43	3.21	6.91	.294	.307	.295	.290	.292	0.9
$\eta = .1$	EASGD-WM	**4.08**	**4.01**	**4.04**	**4.05**	**3.11**	**3.15**	**.267**	**.264**	**.268**	**.265**	**.268**	**.269**
	EASGD	4.24	4.23	4.14	4.13	3.17	6.91	.282	.283	.277	.272	.280	0.9
	MSGD	4.62	4.55	4.22	4.47	3.38	6.91	.288	.307	.287	.288	.301	0.9

Notes: Each experiment was run three times, and the minimum was taken over the average trajectory. In each run, the algorithms were initialized from the same starting location. Surprisingly, EASGD-WM consistently achieves the lowest test error (all but one setting) and the lowest test loss (all but four settings) in comparison to EASGD and MSGD. For high learning rate and high δ, MSGD and EASGD eventually run into convergence issues, while EASGD-WM does not (error of .9 and test loss of 6.91 indicate convergence issues). Bold indicates the top performance of the three algorithms for choice of η and δ.

each algorithm is initialized from the same location in each run. For simplicity, we simply report the testing loss and testing error rather than the results on the training data. For all but one choice of η and δ, EASGD-WM outperforms both EASGD and MSGD in classification accuracy, demonstrating that the trend is robust to choice of learning rate and momentum value.

Much like SGD with momentum, single-agent EASGD-WM is a second-order system in time. It also maintains a similar computational complexity and requires storing only one extra set of parameters for the quorum variable.

Indeed, this motivates a new class of second-order in time algorithms for non-distributed optimization given by the feedback interconnection

$$\dot{\mathbf{x}} = -\nabla f(\mathbf{x}) + k(\tilde{\mathbf{x}} - \mathbf{x}), \tag{6.3}$$

$$\dot{\tilde{\mathbf{x}}} = \mathbf{g}(\tilde{\mathbf{x}}, \mathbf{x}), \tag{6.4}$$

where $\mathbf{g}$ represents arbitrary dynamics for the quorum variable (Russo & Slotine, 2010), and in general might be chosen as a nonlinear filter. The simple linear filter $\mathbf{g}(\tilde{\mathbf{x}}, \mathbf{x}) = k(\mathbf{x} - \tilde{\mathbf{x}})$ recovers EASGD. Figure 9 shows that while EASGD obtains better performance than QSGD, QSGD maintains better stability properties. Designing nonlinear filters $\mathbf{g}$ that can combine the regularization of EASGD with the stability of QSGD is an interesting direction of future research.

Returning to the distributed case, Figure 9d shows that EASGD and QSGD respond differently to the choice of k.[15] EASGD is less synchronized than QSGD in all cases. Hence, in the context of Figure 7, a possible explanation for the improved performance of EASGD when compared to QSGD is simply the observation that it tends to remain less synchronized.

To answer this question, we use a scaling factor $k_{EASGD} = r \times k_{QSGD}$ to roughly match the levels of synchronization between EASGD and QSGD. Results for $r = 1.35$ are shown in Figure 11, and the synchronization curves are either approximately equal or EASGD remains more synchronized across all values of p. Additional values of $p = 32$ and $p = 64$ are shown, and EASGD now converges for all attempted values of $p < 64$. QSGD continues to perform worse than EASGD on the test data due to an increased tendency to overfit. As the number of agents is increased, QSGD improves up to $p = 32$; $p = 64$ obtains roughly the same test performance. EASGD improves up to around $p = 16$ and does not converge for $p = 64$ (see Figure 11a; the curves in Figures 11b and 11d are covered by the insets, but EASGD obtains roughly 55% testing accuracy). In general, EASGD with p agents obtains roughly the same performance as QSGD with $2p$ agents.

[15] Figure 9d shows the distance from $\tilde{\mathbf{x}}$ for EASGD. The distance from $\mathbf{x}^{\bullet}$ for EASGD is nearly identical.

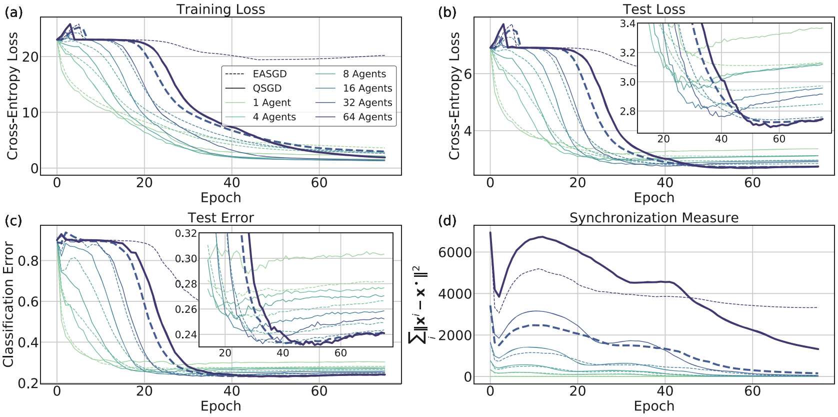

Figure 11: A comparison of QSGD and EASGD with $k_{QSGD} = 0.04$ and $k_{EASGD} = r \times k_{QSGD}$ with $r = 1.35$. In this case, EASGD converges and performs better for all values of p up to $p = 64$, where it again fails to converge. Nevertheless, performance for EASGD with $p = 16$ and $p = 32$ approximately matches that of QSGD with $p = 64$. Insets display a more finely grained view near the end of learning. The best-performing curves for each algorithm are shown in bold. The figure is best viewed in color.

Interestingly, Figure 11d shows that the high p stability issues for EASGD are not simply due to a lack of synchronization, as EASGD actually remains more synchronized than QSGD for $p = 64$ for much of the training time. We offer a simple possible explanation for these stability issues in the supplemental information by analyzing discrete-time optimization of a one-dimensional quadratic objective. Another explanation is afforded by theorems 5 and 6, which reveal poor scaling with p of both terms in the bound for EASGD when compared to QSGD. Together, these observations highlight stability issues in both continuous and discrete-time.

As discussed in the text and the description of the experimental setup, our theory allows the agents to be initialized in different locations and to use distinct learning rates through individual learning rate schedules. In the original work on EASGD, it was postulated that starting the agents at different locations would break symmetry and lead to instability (Zhang et al., 2015). Similarly, a single learning rate was used for all agents. The above simulations demonstrate that starting from distinct locations and decreasing the learning rate on an individual basis is nonproblematic. We show in Figure 12 that starting from a single location leads to decreased performance. Surprisingly, Figure 12 also highlights that initializing the agents from multiple locations is critical for optimal improvement as the number of agents is increased.

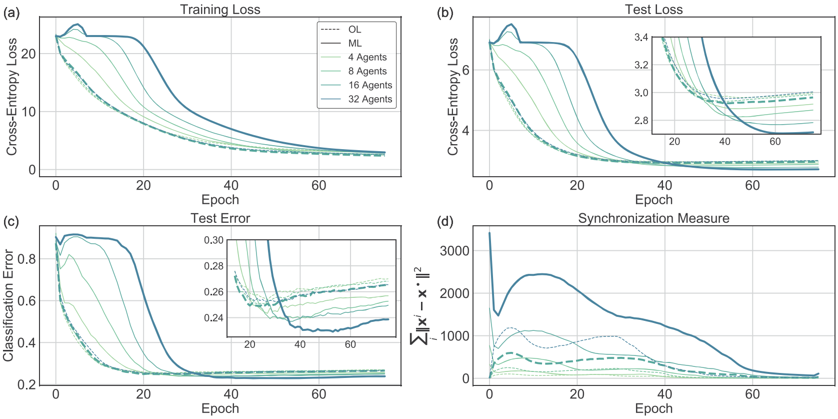

Figure 12: A comparison between starting the agents from multiple locations (ML) and one location (OL) for EASGD with a value of $k = 0.054$. Starting from multiple locations exhibits better test accuracy, lower test loss, and greater improvement as the number of agents is increased. Insets display a more finely grained view near the end of learning. The best-performing curves for each setting are shown in bold. The figure is best viewed in color.

7 Conclusion

In this article, we presented a continuous-time analysis of distributed stochastic gradient algorithms within the framework of stochastic nonlinear contraction theory. Through analogy with quorum-sensing mechanisms, we analyzed the effect of synchronization of the individual SGD agents on the noise generated by their stochastic gradient approximations. We demonstrated that synchronization can effectively reduce the noise felt by each of the individual agents and by their spatial mean. We further demonstrated that synchronization can be seen to reduce the amount of smoothing imposed by SGD on the loss function. Through simulations on model non-convex optimization problems, we provided insight into how the distributed and coupled setting affects convergence to minima of the smoothed loss and the true loss. We introduced a new distributed algorithm, QSGD, and proved convergence results for a strongly convex objective for QSGD, QSGD with momentum, and EASGD. We further introduced a state-dependent variant of QSGD and constructed one specific example of the algorithm to show how the formalism can be used to bias exploration. We presented experiments on deep neural networks and compared the properties of QSGD, SD-QSGD, and EASGD for generalization performance. We noted an interesting regularizing property of EASGD even in the single-agent case and compared it to basic SGD with momentum, showing that it can lead to improved generalization. Research into similar higher-order in time optimization algorithms formed as coupled dynamical systems is an interesting direction of future work.

Appendix: Interaction between Synchronization and Noise: Extra Quorum Dynamics

We provide a mathematical characterization of how synchronization reduces the noise felt by the agents with arbitrary quorum dynamics. This is a generalization of what was shown in section 3.2 and does not depend on the dynamics of the quorum variable. In addition to the assumptions stated in section 2.6, we require that the gradient workers are stochastically contracting with rate $\lambda = k - \bar{\lambda}$ and bound $\frac{\eta}{b}C$, so that the synchronization condition 3.4 derived in section 3.1 can be applied. For completeness, we consider

$$d\mathbf{x}^i = \left(-\nabla f(\mathbf{x}^i) + k\left(\tilde{\mathbf{x}} - \mathbf{x}^i\right)\right) dt + \sqrt{\frac{\eta}{b}}\mathbf{B}(\mathbf{x}^i)d\mathbf{W}^i, \tag{A.1}$$

$$d\tilde{\mathbf{x}} = \mathbf{g}(\mathbf{x}^\bullet, \tilde{\mathbf{x}})dt. \tag{A.2}$$

As in the main text, we define $x^\bullet = \frac{1}{p}\sum_i x_i$. Adding up the stochastic dynamics in equation A.2, we find

$$d\mathbf{x}^\bullet = \left[-\frac{1}{p}\sum_i \nabla f(\mathbf{x}^i) + k(\tilde{\mathbf{x}} - \mathbf{x}^\bullet)\right] dt + \sqrt{\frac{\eta}{bp^2}}\sum_i \mathbf{B}(\mathbf{x}^i)d\mathbf{W}^i.$$

We then define

$$\boldsymbol{\epsilon} = -\frac{1}{p}\sum_i \nabla f(\mathbf{x}^i) + \nabla f(\mathbf{x}^\bullet),$$

so that we can rewrite

$$d\mathbf{x}^\bullet = \left[-\nabla f(\mathbf{x}^\bullet) + \boldsymbol{\epsilon} + k(\tilde{\mathbf{x}} - \mathbf{x}^\bullet)\right] dt + \sqrt{\frac{\eta}{bp^2}}\sum_i \mathbf{B}(\mathbf{x}^i)d\mathbf{W}^i.$$

Applying the Taylor formula with integral remainder to the components of the gradient $(-\nabla f(x))_j$, we have, with $\mathbf{F}_j$ denoting the gradient of $(-\nabla f(x))_j$, and $\mathbf{H}_j$ denoting its Hessian:

$$\begin{aligned}&\left(-\nabla f(\mathbf{x}^i)\right)_j + \left(\nabla f(\mathbf{x}^\bullet)\right)_j - \mathbf{F}_j^T(\mathbf{x}^\bullet)(\mathbf{x}^i - \mathbf{x}^\bullet)\\ &\quad = \int_0^1 (1-s)\left(\mathbf{x}^i - \mathbf{x}^\bullet\right)^T \mathbf{H}_j\left((1-s)\mathbf{x}^i + s\mathbf{x}^\bullet\right)\left(\mathbf{x}^i - \mathbf{x}^\bullet\right).\end{aligned}$$

Summing over i and applying the assumed bound $\mathbf{H}_j \leq Q\mathbf{I}$ leads to the inequality

$$\left|\sum_i \left[\left(-\nabla f(\mathbf{x}^i)\right)_j + \left(\nabla f(\mathbf{x}^\bullet)\right)_j\right]\right| \leq \frac{Q}{2}\sum_i \|\mathbf{x}^i - \mathbf{x}^\bullet\|^2.$$

The left-hand side of the above inequality is $p|\boldsymbol{\epsilon}_j|$. Squaring both sides and summing over j provides a bound on $p^2\|\boldsymbol{\epsilon}\|^2$. Squaring both sides, performing this sum, noting that j runs from 1 to n, taking a square root, taking an expectation over the noise, and using the synchronization condition in equation 3.4,

$$\mathbb{E}\left[\|\boldsymbol{\epsilon}\|\right] \leq \frac{(p-1)\eta QC\sqrt{n}}{4pb\,(k-\bar{\lambda})}.$$

As a sum of p independent gaussian random variables with mean zero and standard deviations $\frac{\eta}{bp^2}\boldsymbol{\Sigma}(\mathbf{x}^i)$, the quantity

$$\sqrt{\frac{\eta}{bp^2}}\sum_i \mathbf{B}(\mathbf{x}^i)d\mathbf{W}^i = \sqrt{\frac{\eta}{bp^2}}\mathbf{T}d\mathbf{W}$$

can be rewritten as a single gaussian random variable with $\mathbf{T}\mathbf{T}^T = \sum_i \boldsymbol{\Sigma}(\mathbf{x}^i)$ as in the main text. Thus, for a given noise covariance $\boldsymbol{\Sigma}$ and corresponding bound C, the difference between the dynamics followed by $\mathbf{x}^\bullet$ and the noise-free dynamics

$$\begin{aligned}\dot{\mathbf{x}}^i_{nf} &= -\nabla f(\mathbf{x}^i_{nf}) + k\left(\tilde{\mathbf{x}} - \mathbf{x}^i_{nf}\right),\\ \dot{\tilde{\mathbf{x}}} &= \mathbf{g}(\tilde{\mathbf{x}}, \mathbf{x}^\bullet_{nf}),\end{aligned}$$

tends to zero almost surely as $k \to \infty$ and $p \to \infty$. The limit $k \to \infty$ is needed to increase the degree of synchronization to eliminate the effect of $\boldsymbol{\epsilon}$ on $\mathbf{x}^\bullet$, while the limit $p \to \infty$ is needed to eliminate the effect of the additive noise.

Acknowledgments

N.B. was supported by a Department of Energy Computational Science Graduate Fellowship. We graciously thank the reviewers for helpful feedback and for suggestions to improve the work.

References

Bach, F., & Moulines, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate $o(1/n)$. In C. J. C. Burgess, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems, 26* (pp. 773–781). Red Hook, NY: Curran.

Banburski, A., Liao, Q., Miranda, B., Rosasco, L., Liang, B., Hidary, J., & Poggio, T. A. (2019). Theory III: Dynamics and generalization in deep networks. *CoRR*, abs/1903.04991.

Betancourt, M., Jordan, M. I., & Wilson, A. C. (2018). *On symplectic optimization*. doi:10.1109/LPT.2005.844008

Bottou, L. (1998). *On-line learning in neural networks*. New York: Cambridge University Press.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of the 19th International Conference on Computational Statistics* (pp. 177–187). Berlin: Springer.

Bouvrie, J., & Slotine, J.-J. (2013). *Synchronization and noise: A mechanism for regularization in neural systems*. arXiv:1312.1632.

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, *3*(1), 1–122.

Chaudhari, P., Baldassi, C., Zecchina, R., Soatto, S., Talwalkar, A., & Oberman, A. (2017). *Parle: Parallelizing stochastic gradient descent*. arXiv:1707.00424.

Chaudhari, P., Oberman, A., Osher, S., Soatto, S., & Carlier, G. (2018). Deep relaxation: Partial differential equations for optimizing deep neural networks. *Research in the Mathematical Sciences*, *5*.

Chaudhari, P., & Soatto, S. (2018). Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *Proceedings of the 2018 Information Theory and Applications Workshop* (pp. 1–10). doi:10.1109/ITA.2018.8503224

Chung, S., & Slotine, J. E. (2009). Cooperative robot control and concurrent synchronization of Lagrangian systems. *IEEE Transactions on Robotics*, *25*(3), 686–700. doi:10.1109/TRO.2009.2014125

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., . . . Ng, A. Y. (2012). Large scale distributed deep networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems, 25* (pp. 1223–1231). Red Hook, NY: Curran.

Defazio, A., Bach, F., & Lacoste-Julien, S. (2014). Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems, 27* (pp. 1646–1654). Red Hook, NY: Curran.

Denève, S., & Machens, C. K. (2016). Efficient codes and balanced networks. *Nature Neuroscience*, *19*, 375–382. https://doi.org/10.1038/nn.4243

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, *12*, 2121–2159.

Feng, Y., Li, L., & Liu, J. G. (2018). Semigroups of stochastic gradient descent and online principal component analysis: Properties and diffusion approximations. *Communications in Mathematical Sciences*, *16*(3), 777–789. doi:10.4310/CMS.2018.v16.n3.a7

Gardiner, C. (2009). *Stochastic methods* (4th ed.). Berlin: Springer-Verlag.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*. abs/1502.01852.

Horn, R. A., & Johnson, C. R. (2012). *Matrix analysis* (2nd ed.). New York: Cambridge University Press.

Hu, W., Li, C. J., Li, L., & Liu, J.-G. (2017). *On the diffusion approximation of nonconvex stochastic gradient descent*. arXiv:1705.07562v2.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*. abs/1502.03167.

Jastrzębski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., & Storkey, A. (2017). *Three factors influencing minima in SGD*. arXiv:1711.04623v3

Javaloyes, J., Perrin, M., & Politi, A. (2008). Collective atomic recoil laser as a synchronization transition. *Phys. Rev. E*, *78*, 011108. doi:10.1103/PhysRevE.78.011108

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2016). *On large-batch training for deep learning: Generalization gap and sharp minima*. arXiv:1609.04836.

Khalil, H. K. (2002). *Nonlinear systems* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Kleinberg, R., Li, Y., & Yuan, Y. (2018). *An alternative view: When does SGD escape local minima?* arXiv:1802.06175.

Kloeden, P., & Platen, E. (1992). *Numerical solution of stochastic differential equations*. Berlin: Springer-Verlag.

LeCun, Y., Bengio, Y., & Hinton, G. (2015, 27). Deep learning. *Nature*, *521*, 436–444.

Li, Q., Tai, C., & Weinan, E. (2018). *Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations*. arXiv:1811.01558.

Lohmiller, W., & Slotine, J.-J. E. (1998). On contraction analysis for non-linear systems. *Automatica*, *34*(6), 683–696. doi:10.1016/S0005-1098(98)00019-3

Mallat, S. (2016). Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150203. doi:10.1098/rsta.2015.0203

Mandt, S., Hoffman, M. D., & Blei, D. M. (2015). Continuous-time limit of stochastic gradient descent revisited. *NIPS Workshop on Optimization for Machine Learning*.

Mandt, S., Hoffman, M. D., & Blei, D. M. (2016). A variational analysis of stochastic gradient algorithms. In *Proceedings of the 33rd International Conference on Machine Learning*, vol. 48 (pp. 354–363).

Mandt, S., Hoffman, M. D., & Blei, D. M. (2017). Stochastic gradient descent as approximate Bayesian inference. *J. Mach. Learn. Res.*, *18*(1), 4873–4907.

Miller, M. B., & Bassler, B. L. (2001). Quorum sensing in bacteria. *Annual Review of Microbiology*, *55*(1), 165–199. doi:10.1146/annurev.micro.55.1.165

Neelakantan, A., Vilnis, L., Le, Q. V., Sutskever, I., Kaiser, L., Kurach, K., & Martens, J. (2015). *Adding gradient noise improves learning for very deep networks*. arXiv:1511.06807.

Nesterov, Y. (1983). A method for solving a convex programming problem with *convergence rate* $O(1/k^2)$. *Soviet Mathematics Doklady*, *26*, 367–372.

Nesterov, Y. (2004). *Introductory lectures on convex optimization*. Berlin: Springer.

Pham, Q., Tabareau, N., & Slotine, J. (2009). A contraction theory approach to stochastic incremental stability. *IEEE Transactions on Automatic Control*, *54*(4), 816–820. doi:10.1109/TAC.2008.2009619

Poggio, T., Kawaguchi, K., Liao, Q., Miranda, B., Rosasco, L., Boix, X., . . . Mhaskar, H. (2017). *Theory of deep learning III: Explaining the non-overfitting puzzle*. arXiv:1801.00173.

Polyak, B., & Juditsky, A. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, *30*(4), 838–855. doi:10.1137/0330046

Recht, B., & Ré, C. (2013, 01). Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, *5*(2), 201–226. doi:10.1007/s12532-013-0053-8

Recht, B., Re, C., Wright, S., & Niu, F. (2011). Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, *24* (pp. 693–701). Red Hook, NY: Curran.

Robbins, H., & Siegmund, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In J. S. Rustagi (Ed.), *Optimizing methods in statistics* (pp. 233–257). New York: Academic Press. https://doi.org/10.1016/B978-0-12-604550-5.50015-8

Roux, N. L., Schmidt, M., & Bach, F. R. (2012). A stochastic gradient method with an exponential convergence rate for finite training sets. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, *25* (pp. 2663–2671). Red Hook, NY: Curran.

Russo, G., & Slotine, J. J. E. (2010). Global convergence of quorum-sensing networks. *Phys. Rev. E*, *82*, 041919. doi:10.1103/PhysRevE.82.041919

Sagun, L., Evci, U., Guney, V. U., Dauphin, Y., & Bottou, L. (2017). *Empirical analysis of the Hessian of over-parameterized neural networks*. arXiv:1706.04454. doi:10.1016/J.ACTPSY.2012.07.016

Schmidt, M., Le Roux, N., & Bach, F. (2017). Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, *162*(1), 83–112. doi:10.1007/s10107-016-1030-6

Slotine, J.-J. E. (2003). Modular stability tools for distributed computation and control. *International Journal of Adaptive Control and Signal Processing*, *17*(6), 397–416. doi:10.1002/acs.754

Su, W., Boyd, S., & Candes, E. (2014). A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems, 27* (pp. 2510–2518). Red Hook, NY: Curran.

Tabareau, N., Slotine, J.-J., & Pham, Q.-C. (2010). How synchronization protects from noise. *PLOS Computational Biology*, *6*(1), 1–9. doi:10.1371/journal.pcbi.1000637

Tuckwell, H. C., & Rodriguez, R. (1998, Mar 01). Analytical and simulation results for stochastic Fitzhugh-Nagumo neurons and neural networks. *Journal of Computational Neuroscience*, *5*(1), 91–113. https://doi.org/10.1023/A:1008811814446

Wang, W., & Slotine, J. J. E. (2005). On partial contraction analysis for coupled nonlinear oscillators. *Biological Cybernetics*, *92*(1), 38–53. doi:10.1007/s00422-004-0527-x

Wang, W., & Slotine, J.-J. E. (2006). Fast computation with neural oscillators. *Neurocomputing*, *69*(16), 2320–2326. https://doi.org/10.1016/j.neucom.2005.04.012.

Waters, C. M., & Bassler, B. L. (2005). Quorum sensing: Cell-to-cell communication in bacteria. *Annual Review of Cell and Developmental Biology*, *21*(1), 319–346. doi:10.1146/annurev.cellbio.21.012704.131001

Wibisono, A., & Wilson, A. C. (2015). *On accelerated methods in optimization*. arXiv:1509.03616.

Wibisono, A., Wilson, A. C., & Jordan, M. I. (2016). A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, *113*(47), E7351–E7358. doi:10.1073/pnas.1614734113

Wilson, A. C., Recht, B., & Jordan, M. I. (2016). *A Lyapunov analysis of momentum methods in optimization*. arXiv:1611.02635.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). *Understanding deep learning requires rethinking generalization*. arXiv:1611.03530.

Zhang, J., Mokhtari, A., Sra, S., & Jadbabaie, A. (2018). Direct Runge-Kutta discretization achieves acceleration. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems, 31* (pp. 3904–3913). Red Hook, NY: Curran.

Zhang, S., Choromanska, A. E., & LeCun, Y. (2015). Deep learning with elastic averaging SGD. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems*, *28* (pp. 685–693).

Zhang, Y., Saxe, A. M., Advani, M. S., & Lee, A. A. (2018). Energy-entropy competition and the effectiveness of stochastic gradient descent in machine learning. *Molecular Physics*, *116*(21–22), 3214–3223. doi:10.1080/00268976.2018.1483535

Zhu, Z., Wu, J., Yu, B., Wu, L., & Ma, J. (2018). *The anisotropic noise in stochastic gradient descent: Its behavior of escaping from minima and regularization effects*. arXiv:1803.00195.

Received April 9, 2019; accepted August 26, 2019.