COMMUN. MATH. SCI.
Vol. 19, No. 3, pp. 851–863


*FAST COMMUNICATION*

# A SHARP CONVERGENCE RATE FOR A MODEL EQUATION OF THE ASYNCHRONOUS STOCHASTIC GRADIENT DESCENT*

YUHUA ZHU† AND LEXING YING‡

**Abstract.** We give a sharp convergence rate for the asynchronous stochastic gradient descent (ASGD) algorithms when the loss function is a perturbed quadratic function based on the stochastic modified equations introduced in [An et al. Stochastic modified equations for the asynchronous stochastic gradient descent, arXiv:1805.08244]. We prove that when the number of local workers is larger than the expected staleness, then ASGD is more efficient than stochastic gradient descent. Our theoretical result also suggests that longer delays result in slower convergence rate. Besides, the learning rate cannot be smaller than a threshold inversely proportional to the expected staleness.

**Keywords.** Asynchronous Stochastic Gradient Descent; Stochastic Modified Equations; Distributed Learning.

**AMS subject classifications.** 90C15; 65K05; 68W15; 68W20.

## 1. Introduction

Thanks to the availability of large datasets and modern computing resources, optimization-based machine learning has achieved state-of-the-art results in many applications of artificial intelligence. As the datasets continue to increase, distributed optimization algorithms have received more attention for solving large scale machine learning problems. Parallel *stochastic gradient descent* (SGD) is arguably the most popular one among them.

Based on the interaction between different working nodes, there are two types of parallel SGD algorithms: synchronous v.s. asynchronous SGD (SSGD v.s. ASGD). Both methods compute the gradient of the loss function for a given mini-batch on local workers. In SSGD, the local workers pause the training process until the gradients from all local works have been added into the *shared* parameter variable. While in ASGD, the local workers interact with the shared parameter independently without any synchronization, i.e., each local worker continues to compute the next gradient right after their own gradients have been added to the shared parameter. Therefore, ASGD is presumably more efficient than SSGD since the overall training speed is not affected by the slow local workers. On the other hand, ASGD can potentially suffer from the problem of delayed gradients, i.e., the gradients that a local worker sends to the shared parameter are often computed with respect to the parameter of an older version of the model. Therefore, extra stochasticity is introduced in ASGD due to this delay. An interesting mathematical problem is how the delayed gradient affects the training process.

There have been a few papers in the literature that analyze the convergence rate of ASGD. Most of them are from an optimization perspective. For example, [7, 12, 14] proved that ASGD can achieve a nearly optimal rate of convergence when the optimiza-

*Received: January 26, 2020; Accepted (in revised form): November 28, 2020. Communicated by Shi Jin.

†Department of Mathematics, Stanford University, Stanford, CA 94305-2125, USA (yuhuazhu@stanford.edu).

‡Department of Mathematics, Stanford University, Stanford, CA 94305-2125, USA (lexing@stanford.edu).

tion problem is sparse; [13] studies the relationship between ASGD and momentum SGD.

*Related work.* This note follows a perspective based on partial differential equation (PDE) and stochastic differential equation (SDE). In [10], Li et al. first introduced the stochastic modified equations (SME) for modeling the dynamics of SGD in a continuous time approximation. Following this work, there have been quite a few papers along this direction [5, 8, 9, 11] for SGD. Specifically, [3, 4, 6] use the approximated SDE and PDE to study the convergence rate of SGD and SGD with momentum. However, literatures about ASGD are limited. Recently, in [1], An et al. applied the SME approach to the study of ASGD. Based on their paper, we study the convergence rate of ASGD.

This note studies the convergence rate of the time-dependent probability distribution function (PDF) of ASGD to its steady state distribution by using PDE techniques of the stochastic modified equation. The main focus is on the case where the loss function is a perturbed quadratic function. There are mainly two difficulties in this analysis. The first one is that asynchrony results in a degenerate diffusion operator in the corresponding PDE. No trivial analysis is able to give an exponential decay rate for the convergence to the steady state. Thanks to the association of a degenerate diffusion operator and a conservative operator, the decay rate can be recovered through "hypocoercivity" [15]. The key here is to construct a Lyapunov functional to prove the exponential decay of this functional. The second difficulty is to obtain a sharp convergence rate. Such a sharp rate is important to understand the influence of the asynchrony quantitatively. Though our analysis is based on the framework introduced by [2], the current case is more complicated because one has to bound extra terms introduced by the perturbed loss function around the quadratic function. The main contributions of this paper are the following:

- We give a sharp convergence rate for ASGD when the loss function is a perturbed quadratic function.
- For a fixed learning rate, longer delays result in slower convergence rate.
- The learning rate should not be smaller than a threshold and the threshold is inversely proportional to the staleness rate. See Remark 3.1 for details.
- When the number of local workers is larger than the expectated staleness, then ASGD is more efficient than SGD. See Remark 3.2 for details.

The rest of the paper is organized as follows. Section 2 summarizes the results of [1] and derives the PDE for the probability density function (PDF) of ASGD based on these results. Section 3 presents the main results, while the proof of the main theorem is given in Section 4.

## 2. PDE for the probability density function of ASGD

We consider the minimization problem,

$$\min_{\theta\in\mathbb{R}^d} f(\theta) := \frac{1}{n}\sum_{i=1}^{n} f_i(\theta). \tag{2.1}$$

where $\theta$ represents the model parameters, $f_i(\theta)$ denotes the loss function at the $i$-th training sample and $n$ is the size of the training sample set. In the asynchronous stochastic gradient descent (ASGD) algorithm, the parameter $\theta$ is updated with

$$\theta_{k+1} = \theta_k - \eta \nabla_\theta f_{\gamma_k}(\theta_{k-\tau_k}),$$

where $\gamma_k$ is i.i.d. uniform random variable from $\{1,2,\cdots,n\}$ and $\theta_{k-\tau_k}$ is the delayed read of the parameter $\theta$ used to update $\theta_{k+1}$ with a random staleness $\tau_k$.

In [1], An et al. derived the modified stochastic differential equation for the algorithm, under the assumption that $\tau_k$ follows the geometric distribution, i.e., $\tau_k = l$ with probability $(1-\kappa)\kappa^l$ for $\kappa \in (0,1)$. We call $\kappa$ the *staleness rate*. Note that if $\kappa$ is larger, then there is a longer delay. Besides, the expectation of the random staleness $\tau_k$ is $\frac{1}{1-k}$, so we call $\frac{1}{1-k}$ the *expected staleness*. By introducing

$$y_k = -\sqrt{\frac{\eta}{1-\kappa}}\mathbb{E}_{\tau_k}\nabla f(\theta_{\tau_k}),$$

when the learning rate $\eta$ is small, $(\theta_k, y_k)$ can be approximated by time discretizations of a continuous-time stochastic process $(\Theta_{k\delta_t}, Y_{k\delta_t})$ for $\delta_t = \sqrt{\eta(1-\kappa)}$. $(\Theta_t, Y_t)$ satisfies the stochastic differential equation (SDE)

$$\begin{aligned} d\Theta_t &= Y_t dt + \tau dB_t, \\ dY_t &= -\nabla f(\Theta_t)dt - \gamma Y_t dt, \end{aligned} \tag{2.2}$$

where $\tau = \eta^{3/4}/(1-\kappa)^{1/4}\Sigma$, $\gamma = \sqrt{((1-\kappa)/\eta)}$, and $\Sigma$ is the covariance matrix conditioned on $\tau_k$, that is, $\Sigma$ is the covariance between $\Theta_t$ and $Y_t$. We refer to [1] for more details of the above SDE. In this paper, we assume $\Sigma$ is a constant for simplicity. Note that although the theoretical upper bound of the approximation error of the above SDE increases in time, the numerical experiments in [1] show that the approximation error remains small until ASGD converges. So we think it is meaningful to analyze the convergence rate of the continuous form.

We first formally derive the partial differential equation for the probability density function $\psi(t,\theta,y)$ of $(\Theta_t, Y_t)$. For any compactly supported $C^\infty$ function $\phi(\Theta_t, Y_t)$, by Itô's formula,

$$d\phi(\Theta_t, Y_t) = \left(Y_t \cdot \nabla_\theta \phi + (-\nabla f(\Theta_t) - \gamma Y_t)\cdot \nabla_y \phi + \frac{1}{2}\tau^2 \nabla_\theta \cdot \nabla_\theta \phi\right) dt + \tau \nabla_\theta \phi dB_t.$$

Taking expectation of this equation and integrating over $[t, t+h]$ leads to

$$\begin{aligned} &\frac{1}{h}\mathbb{E}(\phi(\Theta_{t+h}, Y_{t+h}) - \phi(\Theta_t, Y_t)) \\ =&\frac{1}{h}\int_t^{t+h} \mathbb{E}\left(Y_s \cdot \nabla_\theta \phi + (-\nabla f(\Theta_s) - \gamma Y_s)\cdot \nabla_y \phi + \frac{\tau^2}{2}\nabla_\theta \cdot \nabla_\theta \phi\right) ds, \end{aligned}$$

which further gives,

$$\begin{aligned} &\frac{1}{h}\int \phi(\theta,y)\left(\psi(t+h,\theta,y) - \psi(t,\theta,y)\right) d\theta\, dy \\ =&\frac{1}{h}\int_t^{t+h}\int \left(y\cdot\nabla_\theta\phi + (-\nabla f(\theta) - \gamma y)\cdot\nabla_y\phi + \frac{\tau^2}{2}\nabla_\theta\cdot\nabla_\theta\phi\right)\psi(t,\theta,y)\, d\theta\, dy\, ds. \end{aligned}$$

Integrating by parts and letting $h \to 0$ results in

$$\int \phi(\theta,y)\partial_t\psi\, d\theta\, dy = \int \phi\left(-y\cdot\nabla_\theta\psi + \nabla f(\theta)\cdot\nabla_y\psi + \nabla_y\cdot(\gamma y\psi) + \frac{\tau^2}{2}\nabla_\theta\cdot\nabla_\theta\psi\right) d\theta\, dy,$$

which is true for any test function. Therefore, the PDF $\psi(t,\theta,y)$ satisfies

$$\partial_t\psi + y\cdot\nabla_\theta\psi - \nabla f(\theta)\cdot\nabla_y\psi = \nabla_y\cdot(\gamma y\psi) + \frac{\tau^2}{2}\nabla_\theta\cdot\nabla_\theta\psi. \tag{2.3}$$

In what follows, we consider the case where $\nabla f$ is a perturbed linear function,

$$\nabla f(\theta) = \omega_0^2\theta + \varepsilon(\theta).$$

A further change-of-variable of

$$x = y, \quad v = -\omega_0^2\theta - \gamma y,$$

turns (2.3) into

$$\partial_t g + v\cdot\nabla_x g - \omega_0^2 x\cdot\nabla_v g = \gamma\nabla_v\cdot\left(vg + \frac{1}{\beta}\nabla_v g\right) + \varepsilon\left(-\frac{1}{\omega_0^2}(v+\gamma x)\right)\cdot(\nabla_x g - \gamma\nabla_v g) \tag{2.4}$$

with $g(t,x,v) = \psi\left(t, -\frac{1}{\omega_0^2}(v+\gamma x), x\right)$, $\beta = \frac{2\gamma}{\tau^2\omega_0^4}$.

General loss functions are much harder than the quadratic loss in the SDE for ASGD. First, different from the PDE for the SGD, or SGD with momentum, which is a Fokker-Planck equation or a Vlasov-Fokker-Planck equation, the steady-state is explicitly given for general loss function. The steady-state of (2.3) cannot be explicitly calculated except for the case where $f$ is a quadratic loss. Second, the sharp decay rate for general potentials is still an open question even for Fokker-Planck equations or Vlasov-Fokker-Planck equations, so it is even harder for the sharp decay rate for equations like (2.3) . In this paper, we use the sharp decay rate to study the difference between ASGD compared with SGD. If the decay rate is not sharp, then the relationship between the parameters will be unclear.

## 3. Main results and proof sketch

When $\varepsilon = 0$, the steady state $M(x,v)$ of (2.4),

$$M(x,v) = M_x M_v := \left(\frac{1}{Z_1}e^{-\frac{\beta\omega_0^2}{2}|x|^2}\right)\left(\frac{1}{Z_2}e^{-\frac{\beta}{2}|v|^2}\right), \tag{3.1}$$

where $Z_1, Z_2$ are the normalization constants such that $\int M_v dv = \int M_x dx = 1$. However, for general $f(\theta)$, unfortunately there is no explicit form of the steady state. By denoting $F(x,v)$ as the steady state of (2.4), the weighted fluctuation function

$$h(t,x,v) = \frac{1}{M}[g(t,x,v) - F(x,v)]$$

satisfies the following equation,

$$\partial_t h + Th = Lh + Rh, \tag{3.2}$$

where

$$\begin{aligned} &T = v\cdot\nabla_x - \omega_0^2 x\cdot\nabla_v \text{ is the transport operator;}\\ &L = \frac{\gamma}{\beta}\frac{1}{M}\nabla_v\cdot(M\nabla_v) \text{ is the linearized Fokker-Planck operator;}\\ &Rh = \varepsilon\cdot(\nabla_x h - \gamma\nabla_v h) - \beta\varepsilon\cdot\left(\omega_0^2 x - \gamma v\right)h \text{ are the perturbation terms.} \end{aligned} \tag{3.3}$$

The above Equation (3.2) is typically called the *microscopic equation* in the literature. It is also convenient for the forthcoming analysis to define the inner product $\langle\cdot,\cdot\rangle$ and the norm $\|\cdot\|_*$ as

$$\langle h,g\rangle_* = \int hgM\,dxdv, \quad \|h\|_*^2 = \langle h,h\rangle_*. \tag{3.4}$$

In addition, $\|\cdot\|^2$ is the standard $L^2$ norm with respect to the Lebesgue measure.

Under the above Gaussian measure $M$, the following Poincare inequality holds

$$\|h\|_*^2 \le \frac{1}{d\beta\min\{\omega_0^2,1\}}\left(\|\nabla_x h\|_*^2 + \|\nabla_v h\|_*^2\right), \quad \text{for } \forall h \text{ s.t. } \int hM dxdv = 0 \tag{3.5}$$

The following key assumption ensures various bounds of the perturbation $\varepsilon(\theta)$.

ASSUMPTION 3.1. *There exists a small constant $\epsilon_0 > 0$, such that,*

$$\max_i \|\varepsilon_i\|_{L^\infty}, \|\varepsilon\cdot x\|_{L^\infty}, \|\varepsilon\cdot v\|_{L^\infty}, d\max_i \|\varepsilon_i'\|_{L^\infty}, \sum_i \|\varepsilon_i'\|_{L^\infty}, \|\varepsilon'\cdot x\|_{L^\infty}, \|\varepsilon'\cdot v\|_{L^\infty} \le \epsilon_0,$$

*where $\varepsilon'(\theta)$ is the derivative of $\varepsilon(\theta)$.*

The following theorem states an exponential decay bound for the fluctuation $h$.

THEOREM 3.1. *Under Assumption 3.1 with $\epsilon_0$ small enough, the fluctuation $h$ decays exponentially as follows,*

$$\|h(t)\|_*^2 \lesssim e^{-2(\mu-\epsilon)t} H(0),$$

*where $H(0) = \|\nabla_x h(0)\|_*^2 + C\|\nabla_v h(0)\|_*^2 + 2\hat{C}\langle\nabla_x h(0), \nabla_v h(0)\rangle_*$, $\epsilon = \epsilon_0 C_1$ for a constant $C_1$ depending on $\omega_0, \gamma$ and*

$$\begin{cases} \text{when } \gamma < 2\omega_0: & \mu = \gamma, \quad C = \omega_0^2, \quad \hat{C} = \gamma/2; \\ \text{when } \gamma > 2\omega_0: & \mu = \gamma - \sqrt{\gamma^2 - 4\omega_0^2}, \quad C = \gamma^2/2 - \omega_0^2, \quad \hat{C} = \gamma/2; \\ \text{when } \gamma = 2\omega_0: & \forall \delta > 0, \text{ there exists } C(\delta), \hat{C}(\delta), \text{ such that the decay rate } \mu = \gamma - \delta. \end{cases} \tag{3.6}$$

*More specifically* $C_1 = \left(11 + 11C + 15\hat{C}\right)\epsilon_0 \cdot C_2^2 \cdot \frac{\max\{1,C\}}{C - \hat{C}^2}$,
*where* $C_2 = \frac{\max\{1,\gamma,\gamma^2,\beta\gamma,\beta\omega_0^2\}}{\min\{1,\omega_0^2\}}$.

REMARK 3.1. **How the learning rate and staleness affect the convergence rate?** When the perturbation $\epsilon_0$ is small, the decay rate is dominated by $e^{-2\mu t}$. By the definition of $\delta_t$, one has $t = k\delta_t = k\sqrt{\eta(1-\kappa)}$ with $k$ the number of steps, $\eta$ the learning rate and $\kappa$ the staleness rate. Inserting the definition of $\gamma = \sqrt{((1-\kappa)/\eta)}$ into (3.6), the dominated decay rate $e^{-2\mu t}$ can also be written as,

$$\begin{cases} \text{when } \eta > \dfrac{1}{4\omega_0^2}(1-\kappa): & \mu t = (1-\kappa)k; \\ \text{when } \eta < \dfrac{1}{4\omega_0^2}(1-\kappa): & \mu t = (1-\kappa)k - \left(\sqrt{(1-\kappa)^2 - 4\omega_0^2(1-\kappa)\eta}\right)k. \end{cases} \tag{3.7}$$

From the above discussion, we make two observations:

- The learning rate should not be smaller than $\frac{1}{4\omega_0^2}(1-\kappa)$. For a fixed staleness rate $\kappa$, when the learning rate is larger than the threshold $\frac{1}{4\omega_0^2}(1-\kappa)$, the convergence rate is a constant only depending on $(1-\kappa)$. While the learning rate is smaller than this threshold, the convergence rate will become slower as the learning rate becomes smaller.
- Longer delays result in slower convergence rate. For a fixed learning rate, the optimal decay rate $e^{-2(1-\kappa)k}$ only relates to the staleness of the system. If the system has more delayed readings from the local workers, i.e., $(1-\kappa)$ is smaller, then the convergence rate is slower.

All the above discussion is based on the assumption that $\eta$ is small enough so that the SME-ASGD is a good approximation for ASGD. In other words, we assume $\omega_0$ is large here, hence the threshold $\frac{1}{4\omega_0^2}(1-\kappa)$ is still in the valid regime.

REMARK 3.2. **When is ASGD more efficient than SGD?** Assume we have $m$ local workers and the learning rate is larger than the threshold $\eta > \frac{1}{4\omega_0^2}(1-\kappa)$. When the perturbation $\epsilon_0$ is small, for single batch SGD, the decay rate is $e^{-2k}$ after $k$ steps, while for ASGD, the decay rate is $e^{-2(1-\kappa)k}$ after calculating $k$ gradients. Since now we have $m$ local workers, for the same amount of time, the decay rate for ASGD becomes $e^{-2(1-\kappa)mk}$. Therefore, as long as $(1-\kappa)m > 1$, ASGD will be more efficient than SGD. Since the expectation of the random staleness $\tau_k$ is $\frac{1}{1-\kappa}$, in other words, when the number of local workers is larger than the expected staleness, then ASGD is more efficient than SGD.

We run a simple numerical experiment in Figure 3.1 to verify the above conclusions. (a) When $\kappa = 0.98$, the threshold for the learning rate is $(1-\kappa)/4 = 0.005$. One can see that the blue and red lines spend similar time to converge, which verifies that the convergence rate is the same when the learning rate is above the threshold. When the learning rate is below the threshold, as the learning rate becomes smaller, the convergence of ASGD becomes slower. (b) When the learning rates are all above the threshold, as the staleness rate becomes larger, it takes a longer time for the ASGD to converge. (c) When the staleness rate is 0.96, it takes 2 local workers for ASGD to converge faster than SGD in time. Remark 3.2 gives a conservative estimate for the number of local workers. In practice, it actually requires fewer local workers for ASGD to be more efficient than SGD.

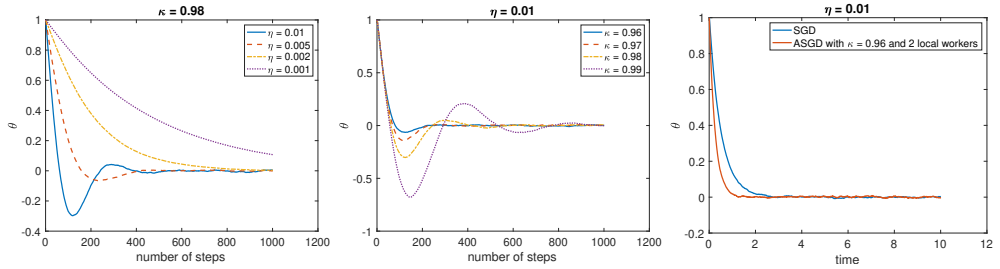
The proof of the theorem is given in Section 4. The main ingredient of the proof is the following Lyapunov functional $H(t)$,

$$H(t) = \|\nabla_x h\|_*^2 + C\|\nabla_v h\|_*^2 + 2\hat{C}\langle \nabla_x h, \nabla_v h\rangle_* \tag{3.8}$$

where $C, \hat{C}$ are constants to be determined later. Note that,

$$\frac{d}{dt}H(t) = \frac{d}{dt}\left(\|\nabla_x h\|_*^2 + C\|\nabla_v h\|_*^2\right) + 2\hat{C}\frac{d}{dt}\langle \nabla_x h, \nabla_v h\rangle_*. \tag{3.9}$$

The first two terms can be estimated with an energy estimation given in Lemma 4.1 of $\nabla_x(3.2)$, $\nabla_v(3.2)$, while the estimation of the last term is given in Lemma 4.2. Actually, $\frac{d}{dt}\|\nabla_v h\|_*^2$ will give the dissipation of $\|\nabla_v h\|_*^2$, and $\frac{d}{dt}\langle \nabla_x h, \nabla_v h\rangle_*$ will give the dissipation of $\|\nabla_x h\|_*^2$. The term $\|\nabla_x h\|_*^2$ and the constants $C, \hat{C}$ in the Lyapunov functional are to make sure the functional is always positive and after combining the results in




(a) (b) (c)

FIG. 3.1. *Apply ASGD to minimize the quadratic function* $f(\theta)=\theta^2$ *with two components* $f_1(\theta)=(\theta-1)^2-1$ *and* $f_2(\theta)=(x+1)^2-1$. *All the plots are averaged results over* 1000 *simulations with initialization* $\theta_0=1$. *(a) Compare the convergence of ASGD with different learning rate when the staleness rate is* $\kappa$. *(b) Compare the convergence of ASGD with different staleness rate when the learning rate is* $\eta=0.01$. *(c) Compare the convergence of ASGD and SGD in time.*

Lemmas 4.1 and 4.2, one could have,

$$\frac{1}{2}\partial_t H(t)+\tilde{C}H(t)\le 0. \tag{3.10}$$

Finally, the exponential decay of $\|h(t)\|_*^2$ can be derived from this inequality and the relationship between $H(t)$ and $\|h(t)\|^2$.

## 4. Proof of Theorem 3.1

The following proposition summarizes a few equalities and inequalities, which will be used frequently in the proof of the main theorem. The proofs are provided in the Appendix.

PROPOSITION 4.1. *For* $\forall g(t,x,v),h(t,x,v)$, *the following statements hold*

*(a)* $\langle Tg,h\rangle_*=-\langle g,Th\rangle_*,\quad \langle Th,h\rangle_*=0.$

*(b)* $\langle Lg,h\rangle_*=-\frac{\gamma}{\beta}\langle\nabla_v g,\nabla_v h\rangle_*.$

*(c)* $\langle Rg,g\rangle_*\le\epsilon_0 C_2\|g\|_*^2,\quad \langle Rg,h\rangle_*+\langle Rh,g\rangle_*\le\epsilon_0 C_2\left(\|g\|_*^2+\|h\|_*^2\right),$
*where* $C_2=\frac{\max\{1,\gamma,\gamma^2,\beta\gamma,\beta\omega_0^2\}}{\min\{1,\omega_0^2\}}.$

*(d)* $\langle\nabla_x(Rh),\nabla_x h\rangle_*\le\frac{11}{2}\epsilon_0 C_2^2\|\nabla_x h\|_*^2+2\epsilon_0 C_2^2\|\nabla_v h\|_*^2.$

*(e)* $\langle\nabla_v(Rh),\nabla_v h\rangle_*\le\frac{11}{2}\epsilon_0 C_2^2\|\nabla_v h\|_*^2+2\epsilon_0 C_2^2\|\nabla_x h\|_*^2.$

*(f)* $\langle\nabla_x(Rh),\nabla_v h\rangle_*+\langle\nabla_v(Rh),\nabla_x h\rangle_*\le\frac{15}{2}\epsilon_0 C_2^2\|\nabla_x h\|_*^2+\frac{15}{2}\epsilon_0 C_2^2\|\nabla_v h\|_*^2.$

The following lemma is the energy estimation of $\nabla_x(3.2)$ and $\nabla_v(3.2)$.

LEMMA 4.1. *The weighted fluctuation function* $h(t,x,v)$ *satisfies*

$$\begin{aligned}&\frac{d}{dt}\left(\|\nabla_x h\|_*^2+C\|\nabla_x h\|_*^2\right)+2\frac{\gamma}{\beta}\sum_i\int M(|\partial_{v_i}\nabla_x h|^2+C|\partial_{v_i}\nabla_v h|^2)dv\\ \le&-2(C-\omega_0^2)\langle\nabla_x h,\nabla_v h\rangle_*-2\gamma C\|\nabla_v h\|_*^2\\ &+(11+4C)\epsilon_0 C_2\|\nabla_x h\|_*^2+(4+11C)\epsilon_0 C_2\|\nabla_v h\|_*^2.\end{aligned}$$

*Proof.* After taking $\nabla_x$ and $\nabla_v$ to (3.2), multiplying them with $\nabla_x hM$ and $\nabla_v hM$ respectively, and integrating over $dxdv$, one obtains

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\nabla_x h\|_*^2 + \langle T\nabla_x h, \nabla_x h\rangle_* - \omega_0^2\langle \nabla_v h, \nabla_x h\rangle_* = \langle L\nabla_x h, \nabla_x h\rangle_* + \langle \nabla_x(Rh), \nabla_x h\rangle_*.\\
&\frac{1}{2}\frac{d}{dt}\|\nabla_v h\|_*^2 + \langle T\nabla_v h, \nabla_v h\rangle_* + \langle \nabla_x h, \nabla_v h\rangle_* = \langle L\nabla_v h, \nabla_v h\rangle_* - \gamma\langle \nabla_v h, \nabla_v h\rangle_*\\
&\qquad\qquad + \langle \nabla_v(Rh), \nabla_v h\rangle_*.
\end{aligned}$$

By invoking Proposition 4.1/(a), the second term of the LHS of each equation vanishes. Multiplying these two equations with 1 and $C$ respectively and adding them together gives rise to

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\left(\|\nabla_x h\|_*^2 + C\|\nabla_x h\|_*^2\right) - \left(\langle L\nabla_x h, \nabla_x h\rangle_* + C\langle L\nabla_v h, \nabla_v h\rangle_*\right)\\
&\le -(C-\omega_0^2)\langle \nabla_x h, \nabla_v h\rangle_* - \gamma C\|\nabla_v h\|_*^2 + \langle \nabla_x(Rh), \nabla_x h\rangle_* + C\langle \nabla_v(Rh), \nabla_v h\rangle_* \qquad (4.1)
\end{aligned}$$

After applying Proposition 4.1/(b),(d),(e), one obtains

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\left(\|\nabla_x h\|_*^2 + C\|\nabla_x h\|_*^2\right) + \frac{\gamma}{\beta}\left(\sum_i \|\partial_{v_i}\nabla_x h\|_*^2 + C\|\partial_{v_i}\nabla_v h\|_*^2\right)\\
&\le -(C-\omega_0^2)\langle \nabla_x h, \nabla_v h\rangle_* - \gamma C\|\nabla_v h\|_*^2\\
&\quad + \left(\frac{11}{2} + 2C\right)\epsilon_0 C_2\|\nabla_x h\|_*^2 + \left(2 + \frac{11}{2}C\right)\epsilon_0 C_2\|\nabla_v h\|_*^2.
\end{aligned}$$

∎

LEMMA 4.2.

$$\begin{aligned}
&\frac{d}{dt}\langle \nabla_x h, \nabla_v h\rangle_* + 2\frac{\gamma}{\beta}\sum_i \langle \partial_{v_i}\nabla_x h, \partial_{v_i}\nabla_v h\rangle_*\\
&\le -\gamma\langle \nabla_x h, \nabla_v h\rangle_* + \omega_0^2\|\nabla_v h\|_*^2 - \|\nabla_x h\|_*^2 + \frac{15}{2}\epsilon_0 C_2^2\|\nabla_x h\|_*^2 + \frac{15}{2}\epsilon_0 C_2^2\|\nabla_v h\|_*^2.
\end{aligned}$$

*Proof.* Taking $\nabla_x$ and $\nabla_v$ to (3.2), multiplying them by $\nabla_v hM$ and $\nabla_x hM$ respectively, and integrating over $dxdv$, one obtains

$$\begin{aligned}
&\left\langle \frac{d}{dt}\nabla_x h, \nabla_v h\right\rangle_* + \langle T\nabla_x h, \nabla_v h\rangle_* - \omega_0^2\langle \nabla_v h, \nabla_v h\rangle_* = \langle L\nabla_x h, \nabla_v h\rangle_* + \langle \nabla_x(Rh), \nabla_v h\rangle_*.\\
&\left\langle \frac{d}{dt}\nabla_v h, \nabla_x h\right\rangle_* + \langle T\nabla_v h, \nabla_x h\rangle_* + \langle \nabla_x h, \nabla_x h\rangle_* = \langle L\nabla_v h, \nabla_x h\rangle_* - \gamma\langle \nabla_v h, \nabla_x h\rangle_*\\
&\qquad\qquad + \langle \nabla_v(Rh), \nabla_x h\rangle_*.
\end{aligned}$$

From Proposition 4.1/(a), the sum of the second terms on the LHS of both equations vanishes. Applying Proposition 4.1/(b) to the first term on the RHS of both equations combine them into a single term. Finally, summing the above two equations and applying Proposition 4.1/(f) to the last terms leads to

$$\begin{aligned}
&\frac{d}{dt}\langle \nabla_x h, \nabla_v h\rangle_* + 2\frac{\gamma}{\beta}\sum_i \langle \partial_{v_i}\nabla_x h, \partial_{v_i}\nabla_v h\rangle_*\\
&\le -\gamma\langle \nabla_x h, \nabla_v h\rangle_* + \omega_0^2\|\nabla_v h\|_*^2 - \|\nabla_x h\|_*^2 + \frac{15}{2}\epsilon_0 C_2^2\|\nabla_x h\|_*^2 + \frac{15}{2}\epsilon_0 C_2^2\|\nabla_v h\|_*^2. \qquad (4.2)
\end{aligned}$$

□

*Proof.* **(The proof of Theorem 3.1.)** By combining the results in Lemma 4.1 and Lemma 4.2, one concludes that

$$\begin{aligned}
&\frac{d}{dt}H(t)+\int[\nabla_x h,\nabla_v h]K[\nabla_x h,\nabla_v h]^\top\frac{1}{M}dxdv\\
&\quad+\int\frac{2\gamma}{\beta}\sum_i[\partial_{v_i}\nabla_x h,\partial_{v_i}\nabla_v h]P[\partial_{v_i}\nabla_x h,\partial_{v_i}\nabla_v h]^\top\frac{1}{M}dxdv\\
&\le\left(11+4C+15\hat{C}\right)\epsilon_0 C_2^2\|\nabla_x h\|_*^2+\left(4+11C+15\hat{C}\right)\epsilon_0 C_2^2\|\nabla_v h\|_*^2,
\end{aligned}\tag{4.3}$$

where

$$K=\begin{bmatrix}2\hat{C}I_d & (C-\omega_0^2+\gamma\hat{C})I_d\\ (C-\omega_0^2+\gamma\hat{C})I_d & (2\gamma C-2\omega_0^2\hat{C})I_d\end{bmatrix},\quad P=\begin{bmatrix}I_d & \hat{C}I_d\\ \hat{C}I_d & CI_d\end{bmatrix}.\tag{4.4}$$

Note that $K$ can be decomposed as,

$$K=QP+PQ^\top,\quad \text{wtih } Q=\begin{bmatrix}0I_d & I_d\\ -\omega_0^2 I_d & \gamma I_d\end{bmatrix}.\tag{4.5}$$

By invoking Lemma 4.3 in [2], we know that there exists a positive definite matrix $P$ such that,

$$K=QP+Q^\top P\ge 2\mu P,\quad \text{with } \mu=\min\{\mathrm{Re}(\lambda):\lambda \text{ is an eigenvalue of } Q\}.\tag{4.6}$$

The value of $\mu$, $C$, and $\hat{C}$ can be separated into three cases.

- case 1: $\gamma<2\omega_0$: $\quad\mu=\gamma,\quad C=\omega_0^2,\quad \hat{C}=\gamma/2$
- case 2: $\gamma>2\omega_0$: $\quad\mu=\gamma-\sqrt{\gamma^2-4\omega_0^2},\quad C=\gamma^2/2-\omega_0^2,\quad \hat{C}=\gamma/2$
- case 3: $\gamma=2\omega_0$: $\quad$For $\forall\delta>0$, there exists $\mu=\gamma-\delta, C(\delta),\hat{C}(\delta)$, such that (4.6) holds.

Inserting $K\ge 2\mu P$ and using the fact that $H(t)=\int[\nabla_x h,\nabla_v h]P[\nabla_x h,\nabla_v h]^\top\frac{1}{M}dxdv$, we can bound the second term in (4.3) from below by $2\mu H(t)$. By the positive definiteness of $P$, the third term is always positive. Therefore,

$$\begin{aligned}
&\frac{d}{dt}H(t)+2\mu H(t)\le\left(11+4C+15\hat{C}\right)\epsilon_0 C_2^2\|\nabla_x h\|_*^2+\left(4+11C+15\hat{C}\right)\epsilon_0 C_2^2\|\nabla_v h\|_*^2,\\
&\frac{d}{dt}H(t)+2(\mu-\epsilon)H(t)\le-2\epsilon H(t)+\left(11+11C+15\hat{C}\right)\epsilon_0 C_2^2\left(\|\nabla_x h\|_*^2+\|\nabla_v h\|_*^2\right),\\
&\frac{d}{dt}H(t)+2(\mu-\epsilon)H(t)\le-\epsilon\left\|\nabla_x h+\hat{C}\nabla_v h\right\|_*^2-\epsilon\left\|\sqrt{C}\nabla_v h+\frac{\hat{C}}{\sqrt{C}}\nabla_x h\right\|_*^2\\
&\qquad\qquad-\epsilon\left(C-\hat{C}^2\right)\|\nabla_v h\|_*^2-\frac{\epsilon}{C}\left(C-\hat{C}^2\right)\|\nabla_x h\|_*^2\\
&\qquad\qquad+\left(11+11C+15\hat{C}\right)\epsilon_0 C_2^2\left(\|\nabla_x h\|_*^2+\|\nabla_v h\|_*^2\right).
\end{aligned}$$

The RHS is less than 0 for all $\nabla_x h,\nabla_v h$ if

$$\left(11+11C+15\hat{C}\right)\epsilon_0 C_2^2\le\epsilon(C-\hat{C}^2)\min\left\{1,\frac{1}{C}\right\},$$

which implies that as long as $\epsilon_0$ is sufficiently small one has

$$\frac{d}{dt}H(t)+2(\mu-\epsilon)H(t)\leq 0 \tag{4.7}$$

for $\epsilon=\left(11+11C+15\hat{C}\right)\epsilon_0 C_2^2\frac{\max\{1,C\}}{C-\hat{C}^2}$. By integrating (4.7) over time and applying Grönwall's inequality to it, we obtain

$$H(t)\leq H(0)e^{-2(\mu-\epsilon)t}.$$

By the Poincare inequality (3.5) and the positive definiteness of $P$, one can bound $H(t)$ from below by $\|h\|_*^2$ up to a constant,

$$H(t)\gtrsim(\|\nabla_x h\|_*^2+\|\nabla_v h\|_*^2)\gtrsim\|h\|_*^2.$$

Therefore, we can conclude

$$\|h(t)\|_*^2\lesssim H(0)e^{-2(\mu-\epsilon)t}.$$

□

**Appendix. The proof of Proposition 4.1.** *Proof.*

(a) The first equation can be proved via integration by parts,

$$\begin{aligned}\langle Tg,h\rangle_* &= \int\left(v\cdot\nabla_x g-\omega_0^2 x\cdot\nabla_v g\right)hM\,dxdv\\ &= \int\left(-v\cdot\nabla_x h+\omega_0^2 x\cdot\nabla_v h\right)g+gh(-v\cdot\nabla_x M+\omega_0^2 x\cdot\nabla_v M)\,dxdv\\ &= -\langle Th,g\rangle_*+\int gh\left(v\cdot(\beta\omega_0^2 x)-\omega_0^2 x\cdot(\beta v)\right)dxdv=-\langle Th,g\rangle_*.\end{aligned}$$

The second equation is directly followed by $\langle Th,h\rangle_*=-\langle Th,h\rangle_*$.

(b) This equation can be obtained also by integration by parts,

$$\begin{aligned}\langle Lg,h\rangle_*=\frac{\gamma}{\beta}\int\frac{1}{M}\nabla_v\cdot(M\nabla_v g)hM\,dxdv &= -\frac{\gamma}{\beta}\int M\nabla_v g\cdot\nabla_v h\,dxdv\\ &= -\frac{\gamma}{\beta}\langle\nabla_v g,\nabla_v h\rangle_*.\end{aligned}$$

(c) The first equation can be written as

$$\begin{aligned}\langle Rg,g\rangle_* &= \langle\varepsilon\cdot(\nabla_x g-\gamma\nabla_v g),g\rangle_*-\left\langle\beta\varepsilon\cdot\left(\omega_0^2 x-\gamma v\right)g,g\right\rangle_*\\ &= \frac{1}{2}\left\langle\varepsilon,(\nabla_x-\gamma\nabla_v)g^2\right\rangle_*-\left\langle\beta\varepsilon\cdot\left(\omega_0^2 x-\gamma v\right)g,g\right\rangle_*.\end{aligned} \tag{A.1}$$

By integrating by parts the first term and using the definition of $\varepsilon=\varepsilon\left(-\frac{1}{\omega_0^2}(v+\gamma x)\right),M=\exp(-\beta\omega_0^2|x|^2/2+\beta|v|^2/2)$, one has

$$\begin{aligned}&\frac{1}{2}\left\langle\varepsilon,(\nabla_x-\gamma\nabla_v)g^2\right\rangle_*=-\frac{1}{2}\left\langle(\nabla_x-\gamma\nabla_v)\cdot\varepsilon,g^2\right\rangle_*-\frac{1}{2}\left\langle\varepsilon\cdot(\nabla_x-\gamma\nabla_v)M,g^2\right\rangle\\ =&-\frac{1}{2}\left\langle-\frac{\gamma}{\omega_0^2}\nabla\cdot\varepsilon-\gamma(-\frac{1}{\omega_0^2})\nabla\cdot\varepsilon,g^2\right\rangle_*+\frac{1}{2}\left\langle\beta\varepsilon\cdot\left(\omega_0^2 x-\gamma v\right),g^2\right\rangle_*\end{aligned}$$

$$=\frac{1}{2}\left\langle \beta\varepsilon\cdot\left(\omega_0^2 x-\gamma v\right),g^2\right\rangle_*. \tag{A.2}$$

Inserting the above equation into (A.1) gives rise to

$$\langle Rg,g\rangle_* = -\frac{1}{2}\left\langle \beta\varepsilon\cdot\left(\omega_0^2 x-\gamma v\right),g^2\right\rangle_* \le \frac{\epsilon_0}{2}(\beta\omega_0^2+\beta\gamma)\|g\|_*^2 \le \epsilon_0 C_2\|g\|_*^2,$$

where we use the assumption $\|\varepsilon\cdot x\|_{L^\infty},\|\varepsilon\cdot v\|_{L^\infty}\le\epsilon_0$ in Assumption 3.1 at the first inequality.

Now we estimate $\langle Rg,h\rangle_*+\langle Rh,g\rangle_*$. First, similar to (A.2), by the definition of $\varepsilon,M$, one has

$$\begin{aligned}
&\langle Rg,h\rangle_* = \langle \varepsilon\cdot(\nabla_x g-\gamma\nabla_v g),h\rangle_* - \left\langle \beta\varepsilon\cdot\left(\omega_0^2 x-\gamma v\right)g,h\right\rangle_* \\
=&-\langle(\nabla_x-\gamma\nabla_v)\varepsilon g,h\rangle_* - \langle \varepsilon g,(\nabla_x-\gamma\nabla_v)h\rangle_* - \langle\varepsilon(\nabla_x-\gamma\nabla_v)Mg,h\rangle \\
&-\left\langle \beta\varepsilon\cdot\left(\omega_0^2 x-\gamma v\right)g,h\right\rangle_* \\
=&0-\langle \varepsilon g,(\nabla_x-\gamma\nabla_v)h\rangle_*+0.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\langle Rg,h\rangle_*+\langle Rh,g\rangle_* \\
=&-\langle\varepsilon(\nabla_x-\gamma\nabla_v)h,g\rangle_*+\langle\varepsilon\cdot(\nabla_x h-\gamma\nabla_v h),g\rangle_* - \left\langle\beta\varepsilon\cdot\left(\omega_0^2 x-\gamma v\right)h,g\right\rangle_* \\
=&-\left\langle\beta\varepsilon\cdot\left(\omega_0^2 x-\gamma v\right)h,g\right\rangle_* \le \frac{\epsilon_0}{2}\left(\beta\omega_0^2+\beta\gamma\right)\left(\|h\|_*^2+\|g\|_*^2\right) \\
\le&\epsilon_0 C_2\left(\|h\|_*^2+\|g\|_*^2\right).
\end{aligned}$$

(d) By the definition of $R$ in (3.3),

$$\begin{aligned}
\langle\nabla_x(Rh),\nabla_x h\rangle_* =& \langle R\nabla_x h,\nabla_x h\rangle_* + \langle\nabla_x\varepsilon(\nabla_x h-\gamma\nabla_v h),\nabla_x h\rangle_* \\
&-\beta\left\langle\nabla_x\varepsilon(\omega_0^2 x-\gamma v)h,\nabla_x h\right\rangle_* - \beta\omega_0^2\langle\varepsilon h,\nabla_x h\rangle_* \\
\le&\epsilon_0 C_2\|\nabla_x h\|_*^2+\langle\nabla_x\varepsilon(\nabla_x h-\gamma\nabla_v h),\nabla_x h\rangle_* \\
&-\beta\left\langle\nabla_x\varepsilon(\omega_0^2 x-\gamma v)h,\nabla_x h\right\rangle_* + \frac{1}{2}\beta\omega_0^2\epsilon_0(d\|h\|_*^2+\|\nabla_x h\|_*^2),
\end{aligned} \tag{A.3}$$

where we apply the inequality $(c)$ to the first term and Assumption 3.1 $\max_i\|\varepsilon_i\|_{L^\infty}\le\epsilon_0$ to the last term of the above inequality. We will then estimate the second and third terms.

$$\begin{aligned}
&\langle\nabla_x\varepsilon(\nabla_x h-\gamma\nabla_v h),\nabla_x h\rangle_* = \sum_{i,j}\left\langle\partial_{x_j}\varepsilon_i(\partial_{x_i}h-\gamma\partial_{v_i}h),\partial_{x_j}h\right\rangle_* \\
=&-\frac{\gamma}{\omega_0^2}\sum_{i,j}\left\langle\varepsilon_i'(\partial_{x_i}h-\gamma\partial_{v_i}h),\partial_{x_j}h\right\rangle_* \\
\le&\frac{\gamma}{\omega_0^2}\sum_{i,j}\|\varepsilon_i'\|_{L^\infty}\left(\frac{1}{2}\|\partial_{x_i}h\|_*^2+\frac{\gamma}{2}\|\partial_{v_i}h\|_*^2+\left\|\partial_{x_j}h\right\|_*^2\right) \\
\le&\frac{\gamma}{\omega_0^2}\left(d\max_i\|\varepsilon_i'\|_{L^\infty}\left(\frac{1}{2}\|\nabla_x h\|_*^2+\frac{\gamma}{2}\|\nabla_v h\|_*^2\right)+\left(\sum_i\|\varepsilon_i'\|_{L^\infty}\right)\|\nabla_x h\|_*^2\right) \\
\le&\frac{\gamma}{\omega_0^2}\epsilon_0\left(\frac{1}{2}\|\nabla_x h\|_*^2+\frac{\gamma}{2}\|\nabla_v h\|_*^2+\|\nabla_x h\|_*^2\right)\le\epsilon_0 C_2\left(\frac{3}{2}\|\nabla_x h\|_*^2+\frac{1}{2}\|\nabla_v h\|_*^2\right)
\end{aligned} \tag{A.4}$$

where $d\max_i \|\varepsilon_i'\|_{L^\infty}, \sum_i \|\varepsilon_i'\|_{L^\infty} \le \epsilon_0$, are used in the second inequality from last. The third term in (A.3) can be bounded by,

$$\begin{aligned}
&-\beta\left\langle \nabla_x \varepsilon(\omega_0^2 x-\gamma v)h, \nabla_x h\right\rangle_* = -\beta\sum_{i,j}\left\langle \partial_{x_j}\varepsilon_i(\omega_0^2 x_i-\gamma v_i)h, \partial_{x_j} h\right\rangle_* \\
&= -\frac{\gamma\beta}{\omega_0^2}\sum_j\left\langle \sum_i \varepsilon_i'(\omega_0^2 x_i-\gamma v_i)h, \partial_{x_j} h\right\rangle_* \\
&\le \frac{\gamma\beta}{\omega_0^2}\epsilon_0\left(\frac{d}{2}\omega_0^2\|h\|_*^2+\frac{d}{2}\gamma\|h\|_*^2+\|\nabla_x h\|_*^2\right) \\
&\le \epsilon_0 C_2\left(d\beta\|h\|_*^2+\|\nabla_x h\|_*^2\right), \qquad \text{(A.5)}
\end{aligned}$$

where $\|\varepsilon'\cdot x\|_{L^\infty}, \|\varepsilon'\cdot v\|_{L^\infty} \le \epsilon_0$ are used in last inequality. Then inserting (A.4), (A.5) into (A.3) leads to

$$\langle \nabla_x(Rh), \nabla_x h\rangle_* \le 4\epsilon_0 C_2\|\nabla_x h\|_*^2+\frac{1}{2}\epsilon_0 C_2\|\nabla_v h\|_*^2+\frac{3}{2}\epsilon_0 C_2\left(d\beta\|h\|_*^2\right). \qquad \text{(A.6)}$$

Now applying the Poincare inequality (3.5) under the Gaussian measure to (A.6) gives rise to

$$\langle \nabla_x(Rh), \nabla_x h\rangle_* \le \left(4\epsilon_0 C_2+\frac{3}{2}\epsilon_0 C_2^2\right)\|\nabla_x h\|_*^2+\left(\frac{1}{2}\epsilon_0 C_2+\frac{3}{2}\epsilon_0 C_2^2\right)\|\nabla_v h\|_*^2.$$

(e) Similar to the proof of (d), one has

$$\begin{aligned}
&\langle \nabla_v(Rh), \nabla_v h\rangle_* \\
&= \langle R\nabla_v h, \nabla_v h\rangle_* + \langle \nabla_v \varepsilon(\nabla_x h-\gamma\nabla_v h), \nabla_v h\rangle_* \\
&\quad -\beta\left\langle \nabla_v \varepsilon(\omega_0^2 x-\gamma v)h, \nabla_v h\right\rangle_* + \beta\gamma\langle \varepsilon h, \nabla_v h\rangle_* \\
&\le \epsilon_0 C_2\|\nabla_v h\|_*^2+\epsilon_0 C_2\left(\frac{1}{2}\|\nabla_x h\|_*^2+\frac{3}{2}\|\nabla_v h\|_*^2\right) \\
&\quad +\epsilon_0 C_2\left(d\beta\|h\|_*^2+\|\nabla_v h\|_*^2\right)+\epsilon_0 C_2\left(\frac{1}{2}d\beta\|h\|_*^2+\frac{1}{2}\|\nabla_v h\|_*^2\right) \\
&\le \left(4\epsilon_0 C_2+\frac{3}{2}\epsilon_0 C_2^2\right)\|\nabla_v h\|_*^2+\left(\frac{1}{2}\epsilon_0 C_2+\frac{3}{2}\epsilon_0 C_2^2\right)\|\nabla_x h\|_*^2.
\end{aligned}$$

(f) Similar to the proof of (d), one has

$$\begin{aligned}
&\langle \nabla_x(Rh), \nabla_v h\rangle_* + \langle \nabla_v(Rh), \nabla_x h\rangle_* \\
&= (\langle R\nabla_x h, \nabla_v h\rangle_* + \langle R\nabla_v h, \nabla_x h\rangle_*) + \langle \nabla_x \varepsilon(\nabla_x h-\gamma\nabla_v h), \nabla_v h\rangle_* \\
&\quad -\beta\left\langle \nabla_x \varepsilon(\omega_0^2 x-\gamma v)h, \nabla_v h\right\rangle_* - \beta\omega_0^2\langle \varepsilon h, \nabla_v h\rangle_* + \langle \nabla_v \varepsilon(\nabla_x h-\gamma\nabla_v h), \nabla_x h\rangle_* \\
&\quad -\beta\left\langle \nabla_v \varepsilon(\omega_0^2 x-\gamma v)h, \nabla_x h\right\rangle_* + \beta\gamma\langle \varepsilon h, \nabla_x h\rangle_* \\
&\le \left(\frac{9}{2}\epsilon_0 C_2+3\epsilon_0 C_2^2\right)\|\nabla_x h\|_*^2+\left(\frac{9}{2}\epsilon_0 C_2+3\epsilon_0 C_2^2\right)\|\nabla_v h\|_*^2.
\end{aligned}$$

□

## REFERENCES


[1] J. An, J. Lu, and L. Ying, *Stochastic modified equations for the asynchronous stochastic gradient descent*, Inf. Inference, 9(4):851–873, 2020. 1, 2, 2

[2] A. Arnold and J. Erb, *Sharp entropy decay for hypocoercive and non-symmetric Fokker-Planck equations with linear drift*, arXiv preprint, arXiv:1409.5425, 2014. 1, 4

[3] B. Shi, W. Su, and M. Jordan, *On learning rates and Schrödinger operators*, arXiv preprint, arXiv:2004.06977, 2020. 1

[4] C. Pratik, O. Adam, O. Stanley, S. Stefano, and C. Guillaume, *Deep relaxation: partial differential equations for optimizing deep neural networks*, Res. Math. Sci., 5(3):30, 2018. 1

[5] P. Chaudhari and S. Soatto, *Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks*, in 2018 Information Theory and Applications Workshop (ITA), IEEE, 1–10, 2018. 1

[6] X. Dai and Y. Zhu, *On large batch training and sharp minima: A Fokker-Planck perspective*, J. Stat. Theory Pract., 14(3):53, 2020. 1

[7] J. Duchi, M.I. Jordan, and B. McMahan, *Estimation, optimization, and parallelism when data is sparse*, Adv. Neural Inf. Process. Syst., 2832–2840, 2013. 1

[8] W. Hu, C.J. Li, L. Li, and J.-G. Liu, *On the diffusion approximation of nonconvex stochastic gradient descent*, Ann. Math. Sci. Appl., 4(1):3–32, 2019. 1

[9] S. Jastrzebski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey, *Three factors influencing minima in SGD*, arXiv preprint, arXiv:1711.04623, 2017. 1

[10] Q. Li, C. Tai, and W. E., *Stochastic modified equations and adaptive stochastic gradient algorithms*, Proceedings of the 34th International Conference on Machine Learning, 70:2101–2110, 2017. 1

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

[12] H. Mania, X. Pan, D. Papailiopoulos, B. Recht, K. Ramchandran, and M.I Jordan, *Perturbed iterate analysis for asynchronous stochastic optimization*, SIAM J. Optim., 27(4):2202–2229, 2017. 1

[13] I. Mitliagkas, C. Zhang, S. Hadjis, and C. Ré, *Asynchrony begets momentum, with an application to deep learning*, in 54th Annual Allerton Conference on Communication, Control and Computing (Allerton), IEEE, 997–1004, 2016. 1

[14] B. Recht, C. Re, S. Wright, and F. Niu, *Hogwild: A lock-free approach to parallelizing stochastic gradient descent*, Adv. Neural Inf. Process. Syst., 693–701, 2011. 1

[15] C. Villani, *Hypocoercivity*, Memoirs of the American Mathematical Society, 202(950), 2009. 1