Accelerated Linear Convergence of Stochastic Momentum Methods in Wasserstein Distances

Bugra Can [1] **Mert Gurbuzbalaban** [1] **Lingjiong Zhu** [2]

Abstract

Momentum methods such as Polyak's heavy ball (HB) method, Nesterov's accelerated gradient (AG) as well as accelerated projected gradient (APG) method have been commonly used in machine learning practice, but their performance is quite sensitive to noise in the gradients. We study these methods under a first-order stochastic oracle model where noisy estimates of the gradients are available. For strongly convex problems, we show that the distribution of the iterates of AG converges with the accelerated $O(\sqrt{\kappa}\log(1/\varepsilon))$ linear rate to a ball of radius ε centered at a unique invariant distribution in the 1-Wasserstein metric where κ is the condition number as long as the noise variance is smaller than an explicit upper bound we can provide. Our analysis also certifies linear convergence rates as a function of the stepsize, momentum parameter and the noise variance; recovering the accelerated rates in the noiseless case and quantifying the level of noise that can be tolerated to achieve a given performance. To the best of our knowledge, these are the first linear convergence results for stochastic momentum methods under the stochastic oracle model. We also develop finer results for the special case of quadratic objectives, extend our results to the APG method and weakly convex functions showing accelerated rates when the noise magnitude is sufficiently small.

1. Introduction

Many key problems in machine learning can be formulated as convex optimization problems. Prominent examples in supervised learning include linear and non-linear regression problems, support vector machines, logistic regression or more generally risk minimization problems (Vapnik, 2013). Accelerated first-order optimization methods based on momentum averaging and their stochastic and proximal variants have been of significant interest in the machine learning community due to their scalability to large-scale problems and good performance in practice both in convex and non-convex settings, including deep learning (see e.g. Sutskever et al. (2013); Nitanda (2014); Hu et al. (2009); Xiao (2010)).

Accelerated optimization methods for unconstrained problems based on momentum averaging techniques go back to Polyak who proposed the *heavy ball* (HB) method (Polyak, 1964) and are closely related to Tschebyshev acceleration, conjugate gradient and under-relaxation methods from numerical linear algebra (Varga, 2009; Karimi & Vavasis, 2017). Another popular momentum-based method is the Nesterov's *accelerated gradient* (AG) method (Nesterov, 2004). For deterministic strongly convex problems, with access to the gradients of the objective, there is a well-established convergence theory for momentum methods. In particular, for minimizing strongly convex smooth objectives with Lipschitz gradients AG method requires $O(\sqrt{\kappa}\log(1/\varepsilon))$ iterations to find an ε-optimal solution where κ is the condition number, this improves significantly over the $O(\kappa\log(1/\varepsilon))$ complexity of the gradient descent (GD) method. HB method also achieves a similar accelerated rate asymptotically in a local neighborhood around the global minimum. Also, for the special case of quadratic objectives, HB method can achieve the accelerated linear rate globally. In the absence of strong convexity, for convex functions, AG has an iteration complexity of $O(1/\sqrt{\varepsilon})$ in function values which accelerates the standard $O(1/\varepsilon)$ convergence rate of GD. In particular, it can be argued that AG method achieves an optimal convergence rate among all the methods that has access to only first-order information (Nesterov, 2004). For constrained problems, a variant of AG, the *accelerated projected gradient* (APG) method (O'Donoghue & Candes, 2015) can also achieve similar accelerated rates (Nesterov, 2004; Fazlyab et al., 2017).

On the other hand, in many applications, the true gradient of the objective function $\nabla f(x)$ is not available but we have access to a noisy but unbiased estimated gradient $\hat{\nabla} f(x)$

[1] Department of Management Science and Information Systems, Rutgers Business School, Piscataway, NJ-08854, United States of America [2] Department of Mathematics, Florida State University, 1017 Academic Way, Tallahassee, FL-32306, United States of America. The authors are in alphabetical order. Correspondence to: Mert Gürbüzbalaban <mg1366@rutgers.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

of the true gradient instead. The common choice of the noise that arises frequently in (stochastic oracle) models is the centered, statistically independent noise with a finite variance where for every $x \in \mathcal{X}$,

$$\textbf{(H1)} \qquad \mathbb{E}\left[\hat{\nabla} f(x) \middle| x\right] = \nabla f(x),$$

$$\textbf{(H2)} \qquad \mathbb{E}\left[\|\hat{\nabla} f(x) - \nabla f(x)\|^2 \middle| x\right] \leq \sigma^2,$$

(see e.g. Bubeck (2014); Lan (2012)). A standard example of this in machine learning is the familiar prediction scenario when $f(x) = \mathbb{E}_\theta \ell(x, \theta)$ where $\ell(x, \theta)$ is the (instantaneous) loss of the predictor x on the example θ with an unknown underlying distribution where the goal is to find a predictor with the best expected loss. In this case, given x, the stochastic oracle draws a random sample θ from the unknown underlying distribution, and outputs $\hat{\nabla} f(x) = \nabla_x \ell(x, \theta)$ which is an unbiased estimator of the gradient. In fact, linear regression, support vector machine and logistic regression problems correspond to particular choices of this loss function ℓ (see e.g. Vapnik (2013)). A second example is where an independent identically distributed (i.i.d.) Gaussian noise with a controlled magnitude is added to the gradients of the objective intentionally, for instance in *private risk minimization* to guarantee privacy of the users' data (Bassily et al., 2014), to escape a local minimum (Ge et al., 2015) or to steer the iterates towards a global minimum for non-convex problems (Gao et al., 2018a;b; Raginsky et al., 2017). Such additive gradient noise arises also naturally when gradients are estimated from noisy data (Cohen et al., 2018; Birand et al., 2013) or the true gradient is estimated from a subset of its components as in (mini-batch) stochastic gradient descent (SGD) methods and their variants.

It is well recognized that momentum-based accelerated methods are quite sensitive to gradient noise (Hardt, 2014; Devolder et al., 2014; Flammarion & Bach, 2015; Devolder et al., 2013), and need higher accuracy of the gradients to perform well (d'Aspremont, 2008; Devolder et al., 2014) compared to standard methods like GD. In fact, with the standard choice of their stepsize and momentum parameter, numerical experiments show that they lose their superiority over a simple method like GD in the noisy setting (Hardt, 2014), yet alone they can diverge (Flammarion & Bach, 2015). On the other hand, numerical studies have also shown that carefully tuned constant stepsize and momentum parameters can lead to good practical performance for both HB and AG under noisy gradients in deep learning (Sutskever et al., 2013). Overall, there has been a growing interest for obtaining convergence guarantees for *stochastic momentum* methods, i.e. momentum methods subject to noise in the gradients.

Several works provided sublinear convergence rates for stochastic momentum methods. Lan (2012); Ghadimi & Lan (2012) developed the AC-SA method which is an adaptation of the AG method to the stochastic composite convex and strongly convex optimization problems and obtained an optimal $O(1/\sqrt{k})$ for the convex case. In a follow-up paper, Ghadimi & Lan (2013) obtained an optimal $O(1/k)$ convergence bound for the *constrained* strongly convex optimization employing a domain shrinking procedure. However, these results do not apply to stochastic HB (SHB). Yang et al. (2016) provided a uniform analysis of SHB and accelerated stochastic gradient (ASG) showing $O(1/\sqrt{k})$ convergence rate for weakly convex stochastic optimization. Gadat et al. (2018) obtained a number of sublinear convergence guarantees for SHB, showing that with decaying stepsize $\alpha_k = O(1/k^\theta)$ for some $\theta \in (0, 1]$, SHB method converges with rate $O(1/k^\theta)$. Several other works focused on proper averaging for reducing the variance of the gradient error in the iterates for strongly convex linear regression problems (Jain et al., 2017; Flammarion & Bach, 2015; Dieuleveut et al., 2017b) and obtained a $O(1/k)$ convergence rate that achieves the minimax estimation rate. Recently, Loizou & Richtárik (2017) studied the SHB algorithm for optimizing the least squares problems arising in the solution of consistent linear systems where the gradient noise comes from sampling the rows of the associated linear system and therefore the gradient errors have a multiplicative form vanishing at the optimum (see Loizou & Richtárik (2017, Sec 2.5)), in which case SGD enjoys linear rates to the optimum with constant stepsize. The authors show that using a constant stepsize the expected SHB iterates converge linearly to a global minimizer with the accelerated rate and provide a first linear (but not an accelerated linear) rate for the expected suboptimality in function values, however the rate provided is not better than the linear rate of SGD and does not reflect the acceleration behavior compared to SGD. We note however that the results of this paper do not apply to our setting as our noise assumptions **(H1)–(H2)** are more general. In our setting, due to the persistence of the noise, it is not possible for the iterates of stochastic momentum methods converge to a global minimum, but rather converge to a stationary distribution around the global minimum. To our knowledge, a linear convergence result for momentum-based methods has never been established under this setting. For SGD, Dieuleveut et al. (2017a) showed that when f is strongly convex, the distribution of the SGD iterates with constant stepsize converges linearly to a unique stationary distribution π_α in the 2-Wasserstein distance requiring $O(\kappa \log(1/\varepsilon))$ iterations to be ε close to the stationary distribution when $\alpha = 1/L$ which is similar to the iteration complexity of (deterministic) gradient descent. A natural question is whether stochastic momentum methods admit a stationary distribution, if so whether the convergence to this distribution can happen faster compared to SGD. As the momentum methods are quite sensitive to gradient noise (Hardt, 2014; Cohen et al., 2018) in terms of performance; a precise characterization of how much noise can be tolerated to achieve accelerated convergence rates under stochastic momentum methods remains understudied.

Contributions: We obtain a number of accelerated convergence guarantees for the SHB, ASG and accelerated stochastic projected gradient (ASPG) methods on both (weakly) convex and strongly convex smooth problems. We note that existing convergence bounds obtained for finite-sum problems that approximate stochastic optimization problems (Nitanda, 2014) do not apply to our setting as our noise is more general, allowing us to deal directly with the stochastic optimization problem itself.

First, for illustrative reasons, we focus on the special case when f is a strongly convex quadratic on $\mathcal{X} = \mathbb{R}^d$ and the gradient noise is additive, statistically independent and i.i.d. with a finite variance σ^2. We obtain accelerated linear convergence results for the ASG method in the weighted 2-Wasserstein distances. Building on the framework of Hu & Lessard (2017) which simplifies the analysis of momentum-based deterministic methods, our analysis shows that all the existing convergence rates and constants can be translated from the deterministic setting to the stochastic setting. Building on novel non-asymptotic convergence guarantees in function values we develop for both the deterministic HB and AG methods, we show that the Markov chain corresponding to the stochastic HB and AG iterates is geometrically ergodic and the distribution of the iterates converges to a unique equilibrium distribution (whose first two moments we can estimate) with the accelerated linear rate $O(\sqrt{\kappa}\log(1/\varepsilon))$ in the p-Wasserstein distance for any $p \geq 1$ with explicit constants. The convergence results hold regardless of the noise magnitude σ, although σ scales the standard deviation of the equilibrium distribution linearly. We also provide improved non-asymptotic estimates for the suboptimality of the HB and AG methods both for deterministic and stochastic settings.

Second, we consider (non-quadratic) stochastic strongly convex optimization problems on $\mathbb{R}^d$ under the stochastic oracle model **(H1)–(H2)**. We derive explicit bounds on the noise variance σ^2 so that ASG method converges linearly to a unique stationary distribution with the accelerated linear rate $O(\sqrt{\kappa}\log(1/\varepsilon))$ in the 1-Wasserstein distance. Our results provide convergence rates as a function of α, β and σ^2 that recovers the convergence rate of the AG algorithm as the noise level σ^2 goes to zero. Therefore, for different parameter choices, we can provide bounds on how much noise can be tolerated to maintain linear convergence.

Third, we focus on the accelerated stochastic projected gradient (ASPG) algorithm for constrained stochastic strongly convex optimization on a bounded domain. We obtain fast accelerated convergence rate to a stationary distribution in the p-Wasserstein distance for any $p \geq 1$. Finally, we extend our results to the weakly convex setting where we show an accelerated $O(\frac{1}{\sqrt{\varepsilon}}\log(1/\varepsilon))$ convergence rate as long as the noise level is smaller than explicit bounds we provide. To our knowledge, accelerated rates in the presence of non-zero noise was not reported in the literature before.

2. Preliminaries

2.1. Notation

We use the notation I_d and 0_d to denote the $d \times d$ identity and zero matrices. The entry at row i and column j of a matrix A is denoted by $A(i,j)$. Kronecker product of two matrices A and B are denoted by $A \otimes B$. A continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is called L-smooth if its gradient is Lipschitz with constant L. A function $f : \mathbb{R}^d \to \mathbb{R}$ is *μ-strongly convex* if the function $x \mapsto f(x) - \frac{\mu}{2}\|x\|^2$ is convex for some $\mu > 0$, where $\|\cdot\|$ denotes the Euclidean norm. Following the literature, let $\mathcal{S}_{0,L}$ denote the class of functions that are convex and L-smooth for some $L > 0$. We use $\mathcal{S}_{\mu,L}$ to denote functions that are both L-smooth and μ-strongly convex for $0 < \mu < L$ (we exclude the trivial case $\mu = L$ in which case the Hessian of f is proportional to the identity matrix where both deterministic gradient descent, HB and AG can converge in one iteration with proper choice of parameters). The ratio $\kappa := L/\mu$ is known as the *condition number*. We denote the global minimum of f on $\mathbb{R}^d$ by f_* and the minimizer of f on $\mathbb{R}^d$ by x_*, which is unique by strong convexity. For any $p \geq 1$, define $\mathcal{P}_p(\mathbb{R}^{2d})$ as the space consisting of all the Borel probability measures ν on $\mathbb{R}^{2d}$ with the finite p-th moment (based on the Euclidean norm). For any two Borel probability measures $\nu_1, \nu_2 \in \mathcal{P}_p(\mathbb{R}^{2d})$, we define the standard p-Wasserstein metric (see e.g. Villani (2009)):

$$\mathcal{W}_p(\nu_1,\nu_2) := \left(\inf_{Z_1\sim\nu_1, Z_2\sim\nu_2} \mathbb{E}[\|Z_1 - Z_2\|^p]\right)^{1/p}.$$

Let $S \in \mathbb{R}^{2d \times 2d}$ be a symmetric positive definite matrix. For any two vectors $z_1, z_2 \in \mathbb{R}^{2d}$, consider the following weighted L_2 norm:

$$\|z_1 - z_2\|_S := \left((z_1 - z_2)^T S (z_1 - z_2)\right)^{1/2}.$$

Define $\mathcal{P}_{2,S}(\mathbb{R}^{2d})$ as the space consisting of all the Borel probability measures ν on $\mathbb{R}^{2d}$ with the finite second moment (based on the $\|\cdot\|_S$ norm). For any two Borel probability measures ν_1 and ν_2 in the space $\mathcal{P}_{2,S}(\mathbb{R}^{2d})$, the weighted 2-Wasserstein distance is defined as

$$\mathcal{W}_{2,S}(\nu_1,\nu_2) := \left(\inf_{Z_1\sim\nu_1, Z_2\sim\nu_2} \mathbb{E}\left[\|Z_1 - Z_2\|_S^2\right]\right)^{1/2}, \tag{1}$$

where the infimum is taken over all random couples (Z_1, Z_2) taking values in $\mathbb{R}^{2d} \times \mathbb{R}^{2d}$ with marginals ν_1 and ν_2. Equipped with the 2-Wasserstein distance (1), $\mathcal{P}_{2,S}(\mathbb{R}^{2d})$ forms a complete metric space (see e.g. Villani (2009)).

Let $\mathcal{P}_{\alpha,\beta}(z,\cdot)$ be a Markov transition kernel (with parameters α, β) associated to a time-homogeneous Markov chain $\{\xi_k\}_{k\geq 0}$ on $\mathbb{R}^{2d}$. A Markov transition kernel is the analogue of the transition matrix for finite state spaces. In particular, if ξ_0 has probability law ν_0 then we use the notation that ξ_k has probability law $\mathcal{P}^k_{\alpha,\beta}\nu_0$. Given a Borel measurable function $\varphi : \mathbb{R}^{2d} \to [0, +\infty]$, we also define

$$(\mathcal{P}_{\alpha,\beta}\varphi)(z) = \int_{\mathbb{R}^{2d}} \varphi(y)\mathcal{P}_{\alpha,\beta}(z, dy).$$

Therefore, it holds that $\mathbb{E}[\varphi(\xi_{k+1})|\xi_k = z] = (\mathcal{P}_{\alpha,\beta}\varphi)(z)$. We refer the readers to Çınlar (2011) for more on the basic theory of Markov chains.

2.2. AG method

For $f \in \mathcal{S}_{\mu,L}$, the deterministic AG method consists of the iterations

$$x_{k+1} = y_k - \alpha\nabla f(y_k),\ y_k = (1+\beta)x_k - \beta x_{k-1}, \quad (2)$$

starting from the initial points $x_0, x_{-1} \in \mathbb{R}^d$, where $\alpha > 0$ is the stepsize and $\beta > 0$ is the momentum parameter (Nesterov, 2004). Since the AG iterate x_{k+1} depends on both x_k and x_{k-1}, it is standard to define the state vector

$$\xi_k := \begin{pmatrix} x_k^T & x_{k-1}^T \end{pmatrix}^T \in \mathbb{R}^{2d}, \quad (3)$$

and rewrite the AG iterations in terms of ξ_k. To simplify the presentation and the analysis, we build on the representation of optimization algorithms as a dynamical system from Hu & Lessard (2017) and rewrite the AG iterations as

$$\xi_{k+1} = A\xi_k + Bw_k,$$

where $A = \tilde{A} \otimes I_d$ and $B = \tilde{B} \otimes I_d$ with

$$\tilde{A} := \begin{pmatrix} (1+\beta) & -\beta \\ 1 & 0 \end{pmatrix},\ \tilde{B} := \begin{pmatrix} -\alpha \\ 0 \end{pmatrix}, \quad (4)$$

and $w_k := \nabla f\left((1+\beta)x_k - \beta x_{k-1}\right)$. The standard analysis of deterministic AG is based on the following Lyapunov function that combines the state vector and function values:

$$V_P(\xi_k) := (\xi_k - \xi_*)^T P(\xi_k - \xi_*) + f(x_k) - f_*, \quad (5)$$

where $\xi_* = (x_*^T\ x_*^T)^T$ and $P \in \mathbb{R}^{2d\times 2d}$ is positive semi-definite matrix to be appropriately chosen. In particular, a linear convergence $f(\xi_{k+1}) - f(\xi_*) \leq V_P(\xi_{k+1}) \leq \rho V_P(\xi_k)$ with rate ρ can be guaranteed if P satisfies a certain matrix inequality precised as follows.

Theorem 1. *(Hu & Lessard, 2017) Let $\rho \in [0,1)$ be given. If there exists a symmetric positive semi-definite 2×2 matrix $\tilde{P}$ (that may depend on ρ) such that*

$$\begin{pmatrix} \tilde{A}^T\tilde{P}\tilde{A} - \rho\tilde{P} & \tilde{A}^T\tilde{P}\tilde{B} \\ \tilde{B}^T\tilde{P}\tilde{A} & \tilde{B}^T\tilde{P}\tilde{B} \end{pmatrix} - \tilde{X} \preceq 0, \quad (6)$$

where $\tilde{X} := \rho\tilde{X}_1 + (1-\rho)\tilde{X}_2 \in \mathbb{R}^{3\times 3}$ with

$$\tilde{X}_1 := \begin{pmatrix} \frac{\beta^2\mu}{2} & \frac{-\beta^2\mu}{2} & \frac{-\beta}{2} \\ \frac{-\beta^2\mu}{2} & \frac{\beta^2\mu}{2} & \frac{\beta}{2} \\ \frac{-\beta}{2} & \frac{\beta}{2} & \frac{\alpha(2-L\alpha)}{2} \end{pmatrix},$$

$$\tilde{X}_2 := \begin{pmatrix} \frac{(1+\beta)^2\mu}{2} & \frac{-\beta(1+\beta)\mu}{2} & \frac{-(1+\beta)}{2} \\ -\frac{\beta(1+\beta)\mu}{2} & \frac{\beta^2\mu}{2} & \frac{\beta}{2} \\ \frac{-(1+\beta)}{2} & \frac{\beta}{2} & \frac{\alpha(2-L\alpha)}{2} \end{pmatrix},$$

and $\tilde{A}, \tilde{B}$ are given by (4), then the deterministic AG iterates defined by (2) for minimizing $f \in \mathcal{S}_{\mu,L}$ satisfies $f(x_k) - f(x_*) \leq V_P(\xi_k) \leq \rho^k V_P(\xi_0)$ where V_P is defined by (5) and $P = \tilde{P} \otimes I_d$.

In particular, Theorem 1 can recover existing convergence rate results for deterministic AG. For example, for the particular choice of

$$\begin{aligned} P_{AG} &:= \tilde{P}_{AG} \otimes I_d,\ \ \tilde{P}_{AG} := \tilde{u}\tilde{u}^T, \quad (7) \\ \tilde{u} &:= \begin{pmatrix} \sqrt{L/2} & \sqrt{\mu/2} - \sqrt{L/2} \end{pmatrix}^T, \end{aligned}$$

and $(\alpha, \beta) = (\alpha_{AG}, \beta_{AG})$ with

$$\alpha_{AG} := \frac{1}{L}, \qquad \beta_{AG} := \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}, \quad (8)$$

in Theorem 1, we obtain the accelerated convergence rate of

$$\rho_{AG} := 1 - \sqrt{\mu/L} = 1 - 1/\sqrt{\kappa}. \quad (9)$$

However, as outlined in the introduction, in a variety of applications in machine learning and stochastic optimization, we do not have access to the true gradient $\nabla f(y_k)$ as in the deterministic AG iterations but we have access to a (noisy) stochastic version $\tilde{\nabla} f(y_k) = \nabla f(y_k) + \varepsilon_{k+1}$, where ε_{k+1} is the random gradient noise. AG algorithm with stochastic gradients has the form

$$\begin{aligned} x_{k+1} &= y_k - \alpha[\nabla f(y_k) + \varepsilon_{k+1}], \quad (10) \\ y_k &= (1+\beta)x_k - \beta x_{k-1}, \end{aligned}$$

which is called the *accelerated stochastic gradient (ASG)* method (see e.g. Jain et al. (2017)). We note that due to the existence of noise, the standard Lyapunov analysis from the literature (see e.g. Wilson et al. (2016); Su et al. (2014)) does not apply directly. We make the assumption that the random gradient errors are centered, statistically independent from the past iterates and have a finite second moment following the literature (Cohen et al., 2018; Hardt, 2014; Neelakantan et al., 2015; Aybat et al., 2018; Flammarion & Bach, 2015). The following assumption is a more formal statement of **(H1)–(H2)** adapting to the iterations ξ_k.

Assumption 2 (Formal statement of **(H1)–(H2)**). *On some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with a filtration $\mathcal{F}_k$ the noise ε_k's are $\mathcal{F}_k$-measurable, stationary and*

$$\mathbb{E}[\varepsilon_k|\mathcal{F}_{k-1}] = 0 \quad \textit{and} \quad \mathbb{E}[\|\varepsilon_k\|^2|\mathcal{F}_{k-1}] \leq \sigma^2.$$

Under Assumption 2, the iterations ξ_k forms a time-homogeneous Markov chain which we will study further in Sections 3 and 4.

2.3. HB method

For $f \in \mathcal{S}_{\mu,L}$, the HB method was proposed by Polyak (1964). It consists of the iterations

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}), \tag{11}$$

where $\alpha > 0$ is the step size and β is the momentum parameter. The following asymptotic convergence rate result for HB is well known.

Theorem 3 (Polyak (1987), see also Recht (2012))**.** *Let the objective function $f \in \mathcal{S}_{\mu,L}$ be a strongly convex quadratic function. Consider the deterministic HB iterations $\{x_k\}_{k\geq 0}$ defined by the recursion (11) from an initial point $x_0 \in \mathbb{R}^d$ with parameters $(\alpha,\beta) = (\alpha_{HB}, \beta_{HB})$ where*

$$\alpha_{HB} := \frac{4}{(\sqrt{\mu}+\sqrt{L})^2}, \quad \beta_{HB} := \left(\frac{\sqrt{L/\mu}-1}{\sqrt{L/\mu}+1}\right)^2. \tag{12}$$

Then, $\|x_k - x_*\| \leq (\rho_{HB} + \delta_k)^k \cdot \|\xi_0 - \xi_*\|$, where δ_k is a non-negative sequence that goes to zero and

$$\rho_{HB} := \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} = 1 - \frac{2}{\sqrt{\kappa}+1}. \tag{13}$$

Furthermore, $f(x_k) - f(x_*) \leq \frac{L}{2}(\rho_{HB}+\delta_k)^{2k} \cdot \|\xi_0 - \xi_*\|^2$.

This result has an asymptotic nature as the sequence δ_k is not explicit. There exist non-asymptotic linear convergence results for HB, but to our knowledge, known linear rate guarantees are slower than the accelerated rate ρ_{HB}; with a rate similar to the rate of gradient descent (Ghadimi et al., 2014). In Section 3.2, we will derive a new non-asymptotic version of this theorem that can guarantee suboptimality for finite k with explicit constants and the accelerated rate ρ_{HB}. Note that the asymptotic rate ρ_{HB} of HB in (13) on quadratic problems is strictly (smaller) faster than the rate ρ_{AG} of AG from (9) in general (except in the particular special case of $\kappa = 1$, we have $\rho_{AG} = \rho_{HB} = 0$). However, for strongly convex functions, HB iterates given by (11) is not globally convergent with parameters α_{HB} and β_{HB} (Lessard et al., 2016), but if the iterates are started in a small enough neighborhood around the global minimum of a strongly convex function, this rate can be achieved asymptotically (Polyak, 1987). Since known guarantees for deterministic AG is stronger than deterministic HB on non-quadratic strongly convex functions, we will focus on the AG method for non-quadratic objectives in our paper.

We will analyze the HB method under noisy gradients:

$$x_{k+1} = x_k - \alpha\left(\nabla f(x_k) + \varepsilon_{k+1}\right) + \beta(x_k - x_{k-1}), \tag{14}$$

where the noise satisfies Assumption 2. This method is called the *stochastic HB* method (Gadat et al., 2018; Loizou & Richtárik, 2018; Flåm, 2004).

In the next section, we show that stochastic momentum methods admit an invariant distribution towards which they converge linearly in a sense we make precise. For illustrative purposes, we first analyze the special case when the objective is a quadratic function, and then move on to the more general case when f is smooth and strongly convex. Also, for quadratic functions we can obtain stronger guarantees exploiting the linearity properties of the gradients.

3. Special case: strongly convex quadratics

First, we assume that the objective $f \in \mathcal{S}_{\mu,L}$ and is a quadratic function of the form

$$f(x) = \frac{1}{2}x^T Q x + a^T x + b, \tag{15}$$

where $x \in \mathbb{R}^d$, $Q \in \mathbb{R}^{d\times d}$ is symmetric positive definite, $a \in \mathbb{R}^d$ is a column vector and $b \in \mathbb{R}$ is a scalar. We also assume $\mu I_d \preceq Q \preceq L I_d$ so that $f \in \mathcal{S}_{\mu,L}$. In this section, we assume the noise ε_k are i.i.d. which is a special case of Assumption 2. We next show that both accelerated stochastic gradient and stochastic HB admit a unique invariant distribution towards which the iterates converge linearly in the 2-Wasserstein metric.

3.1. Accelerated linear convergence of AG and ASG

Given vectors, $z_1, z_2 \in \mathbb{R}^{2d}$, we consider

$$\|z_1 - z_2\|_{S_{\alpha,\beta}} := \left((z_1 - z_2)^T S_{\alpha,\beta}(z_1 - z_2)\right)^{1/2}. \tag{16}$$

where $S_{\alpha,\beta} \in \mathbb{R}^{2d\times 2d}$ is defined as the symmetric matrix

$$S_{\alpha,\beta} := P_{\alpha,\beta} + \begin{pmatrix} \frac{1}{2}Q & 0_d \\ 0_d & 0_d \end{pmatrix}, \tag{17}$$

where $P_{\alpha,\beta} := \tilde{P}_{\alpha,\beta} \otimes I_d$ and $\tilde{P}_{\alpha,\beta}$ is a non-zero symmetric positive definite 2×2 matrix (that may depend on the parameters α and β) with the entry $\tilde{P}_{\alpha,\beta}(2,2) \neq 0$. It can be shown that $S_{\alpha,\beta}$ is positive definite on $\mathbb{R}^{2d}$ (see Lemma 18 in the supplementary file), even though $\tilde{P}_{\alpha,\beta}$ can be rank deficient. In this case, due to the positive definiteness of $S_{\alpha,\beta}$, (16) defines a weighted L_2 norm on $\mathbb{R}^{2d}$. Therefore, if we set $S_{\alpha,\beta}$ in (1), we can consider the 2-Wasserstein distance between two Borel probability measures ν_1 and ν_2 defined on $\mathbb{R}^{2d}$ with finite second moments (based on the $\|\cdot\|_{S_{\alpha,\beta}}$ norm.

The ASG iterates $\{\xi_k\}_{k\geq 0}$ defined by (3) and (10) forms a time-homogeneous Markov chain on $\mathbb{R}^{2d}$. Consider the Markov kernel $\mathcal{P}_{\alpha,\beta}$ associated to this chain. Recall that if ν is the distribution of ξ_0, the distribution of ξ_k is denoted by $\mathcal{P}^k_{\alpha,\beta}\nu$. The following theorem shows that this Markov Chain admits a unique equilibrium distribution $\pi_{\alpha,\beta}$ and the distribution of the ASG iterates converges to this distribution exponentially fast with (linear) rate $\rho_{\alpha,\beta}$. This rate achieved by ASG is the same as the rate of the deterministic AG method, except that it is achieved in a different notion (with respect to convergence in $\mathcal{W}_{2,S_{\alpha,\beta}}$). The proof is given in the supplementary file and it is based on studying the contractivity properties of the map $\nu \mapsto \mathcal{P}^k_{\alpha,\beta}\nu$ in the Wasserstein space.[1]

Theorem 4. *Let $f \in \mathcal{S}_{\mu,L}$ be a quadratic function (15). Consider the Markov chain $\{\xi_k\}_{k\geq 0}$ defined by the ASG recursion (10) with parameters α and β and let $\nu_{k,\alpha,\beta}$ denote the distribution of ξ_k with $\nu_{0,\alpha,\beta} \in \mathcal{P}_{2,S_{\alpha,\beta}}(\mathbb{R}^{2d})$. Let any convergence rate $\rho_{\alpha,\beta} \in [0,1)$ be given. If there exists a matrix $\tilde{P}_{\alpha,\beta}$ with $\tilde{P}_{\alpha,\beta}(2,2) \neq 0$ satisfying inequality (6) with $P = P_{\alpha,\beta}$ and $\rho = \rho_{\alpha,\beta}$, then there exists a unique stationary distribution $\pi_{\alpha,\beta}$.*

$$\mathcal{W}_{2,S_{\alpha,\beta}}(\nu_{k,\alpha,\beta}, \pi_{\alpha,\beta}) \leq \rho^k_{\alpha,\beta}\mathcal{W}_{2,S_{\alpha,\beta}}(\nu_{0,\alpha,\beta}, \pi_{\alpha,\beta}),$$

where $\mathcal{W}_{2,S_{\alpha,\beta}}$ is the 2-Wasserstein distance (1) equipped with the $\|\cdot\|_{S_{\alpha,\beta}}$ norm. In particular, with $(\alpha,\beta) = (\alpha_{AG}, \beta_{AG})$ and $P = P_{AG}$ with P_{AG} defined in (7), we obtain the optimal accelerated linear rate of convergence:

$$\mathcal{W}^2_{2,S_{\alpha,\beta}}(\nu_{k,\alpha,\beta}, \pi_{\alpha,\beta}) \leq \rho^k_{AG}\mathcal{W}^2_{2,S_{\alpha,\beta}}(\nu_{0,\alpha,\beta}, \pi_{\alpha,\beta}), \tag{18}$$

with $\rho_{AG} = 1 - \frac{1}{\sqrt{\kappa}}$ as in (9).

For the AG method, the choice of $(\alpha,\beta) = (\alpha_{AG}, \beta_{AG})$ is popular in practice, however a faster rate can be achieved asymptotically if

$$\alpha^*_{AG} := \frac{4}{3L+\mu}, \qquad \beta^*_{AG} := \frac{\sqrt{3\kappa+1}-2}{\sqrt{3\kappa+1}+2}, \tag{19}$$

so that the asymptotic linear convergence rate in distance to the optimality becomes $\rho^*_{AG} := 1 - \frac{2}{\sqrt{3\kappa+1}}$, which translates into the rate $(\rho^*_{AG})^2$ in function values that is (smaller) faster than ρ_{AG} (Lessard et al., 2016); improving the iteration complexity by a factor of $4/\sqrt{3} \approx 2.3$ when κ is large. However, these results are asymptotic. Below we provide a first non-asymptotic bound with the faster rate ρ^*_{AG}.

Theorem 5. *Let $f \in \mathcal{S}_{\mu,L}$ be a quadratic function (15). Consider the deterministic AG iterations $\{x_k\}_{k\geq 0}$ defined by the recursion (3) with initialization $x_0, x_{-1} \in \mathbb{R}^d$ and parameters $(\alpha,\beta) = (\alpha^*_{AG}, \beta^*_{AG})$ as in (19). Then,*

$$\|x_k - x_*\| \leq C^*_k(\rho^*_{AG})^k \cdot \|\xi_0 - \xi_*\|, \tag{20}$$

$$f(x_k) - f(x_*) \leq \frac{L}{2}(C^*_k)^2(\rho^*_{AG})^{2k} \cdot \|\xi_0 - \xi_*\|^2,$$

where $\rho^*_{AG} = 1 - \frac{2}{\sqrt{3\kappa+1}}$ and

$$C^*_k := \max\left\{\bar{C}^*, \sqrt{k^2((\rho^*_{AG})^2+1)^2 + 2(\rho^*_{AG})^2}\right\}, \tag{21}$$

with $\bar{C}^* := \frac{\sqrt{3\kappa+1}+2}{2}((\rho^*_{AG})^2+1)\tilde{C}^*$ and

$$\tilde{C}^* := \max_{i:\mu<\lambda_i<L, \lambda_i \neq \frac{3L+\mu}{4}} \frac{\sqrt{\mu(3L+\mu)}}{\sqrt{(\lambda_i-\mu)|3L+\mu-4\lambda_i|}},$$

where $\{\lambda_i\}_{i=1}^d$ are the eigenvalues of the Hessian Q.

Remark 6. *The constants C^*_k grows linearly with k in Theorem 5 and this dependency is tight in the sense that there are examples achieving it (see the proof in the supplementary file). Our bounds improves the existing results that provide a slower rate ρ_{AG} with bounded constants in front of the linear rate (Nesterov, 2004; Bubeck, 2014), if k is large enough (larger than a constant that can be made explicit).*

Building on this non-asymptotic convergence result for the deterministic AG method, we obtain similar non-asymptotic convergence guarantees for the ASG method in p-Wasserstein distances towards convergence to a stationary distribution.

Theorem 7. *Let $f \in \mathcal{S}_{\mu,L}$ be a quadratic function (15). Consider the ASG iterations $\{x_k\}_{k\geq 0}$ defined by the recursion (10). Let $\nu_{k,\alpha,\beta}$ be the distribution of the k-th iterate ξ_k for $k \geq 0$, where $\xi_k^T := (x_k^T, x_{k-1}^T)$ and parameters $(\alpha,\beta) = (\alpha^*_{AG}, \beta^*_{AG})$ as in (19). Also assume that $\nu_{0,\alpha^*_{AG},\beta^*_{AG}} \in \mathcal{P}_p(\mathbb{R}^{2d})$ and the noise ε_k has finite p-th moment. Then, there exists a unique stationary distribution $\pi_{\alpha,\beta}$ and for any $p \geq 1$,*

$$\mathcal{W}_p(\nu_{k,\alpha,\beta}, \pi_{\alpha,\beta}) \leq C^*_k(\rho^*_{AG})^k \cdot \mathcal{W}_p(\nu_{0,\alpha,\beta}, \pi_{\alpha,\beta}), \tag{22}$$

where $\rho^*_{AG} = 1 - \frac{2}{\sqrt{3\kappa+1}}$, C^*_k is defined in (21) and $\mathcal{W}_p$ is the standard the p-Wasserstein distance.

We can also control the expected suboptimality $\mathbb{E}[f(x_k)] - f(x_*)$ after k iterations.

Theorem 8. *With the same assumptions as in Theorem 7,*

$$\mathbb{E}[f(x_k)] - f(x_*) \leq \frac{L}{2}Tr(X^*_{AG}) + V^*_{AG}(\xi_0)(C^*_k)^2(\rho^*_{AG})^{2k}, \tag{23}$$

where $\rho^*_{AG} = 1 - \frac{2}{\sqrt{3\kappa+1}}$, C^*_k is defined in (21), X^*_{AG} is the covariance matrix of $\xi_\infty - \xi_*$ and $V^*_{AG}(\xi_0)$ is a constant depending on any initial state ξ_0 and both X and $V^*_{AG}(\xi_0)$ will be spelled out in explicit form in the supplementary file.

3.2. Accelerated linear convergence of HB and SHB

We first give a non-asymptotic convergence result for the deterministic HB method with explicit constants, which also

[1] We also provide numerical experiments in the supplementary file to illustrate the results of Theorem 4.

implies a bound on the suboptimality $f(x_k) - f(x_*)$. This refines the asymptotic results in the literature (Theorem 3).

Theorem 9. *Let $f \in \mathcal{S}_{\mu,L}$ be a quadratic function (15). Consider the deterministic HB iterations $\{x_k\}_{k\geq 0}$ defined by the recursion (11) with initialization $x_0, x_{-1} \in \mathbb{R}^d$ and parameters $(\alpha, \beta) = (\alpha_{HB}, \beta_{HB})$ as in (12). Then,*

$$\begin{aligned} \|x_k - x_*\| \leq & C_k \rho_{HB}^k \cdot \|\xi_0 - \xi_*\|, \qquad (24) \\ f(x_k) - f(x_*) \leq & \frac{L}{2} C_k^2 \rho_{HB}^{2k} \cdot \|\xi_0 - \xi_*\|^2, \end{aligned}$$

where ρ_{HB} is defined by (13) and

$$C_k := \max\left\{\bar{C}, \sqrt{4k^2\left(\frac{L+\mu}{L-\mu}\right)^2 + 2}\right\}, \qquad (25)$$

with $\bar{C} := \max_{i:\mu<\lambda_i<L} \frac{\mu+L}{2\sqrt{(\lambda_i-\mu)(L-\lambda_i)}}$, where $\{\lambda_i\}_{i=1}^d$ are the eigenvalues of the Hessian matrix of f.

Remark 10. *It is clear from the definition of C_k in Theorem 9 that the leading coefficient C_k grows at most linearly in the number of iterates k and this dependency cannot be removed in the sense that there are some examples achieving our upper bounds in terms of k dependency (see the supplementary file).*

Building on this non-asymptotic convergence result for the deterministic HB method, we obtain similar non-asymptotic convergence guarantees for the SHB method in Wasserstein distances towards convergence to a stationary distribution.

Theorem 11. *Let $f \in \mathcal{S}_{\mu,L}$ be a quadratic function (15). Consider the HB iterations $\{x_k\}_{k\geq 0}$ defined by the recursion (14). Let $\nu_{k,\alpha,\beta}$ be the distribution of the k-th iterate ξ_k for $k \geq 0$, where $\xi_k^T := (x_k^T, x_{k-1}^T)$ and parameters $(\alpha, \beta) = (\alpha_{HB}, \beta_{HB})$ where $(\alpha_{HB}, \beta_{HB})$ is defined as in (12). Also assume that $\nu_{0,\alpha_{HB},\beta_{HB}} \in \mathcal{P}_p(\mathbb{R}^{2d})$ and the noise ε_k has finite p-th moment. Then, there exists a unique stationary distribution $\pi_{\alpha,\beta}$ and for any $p \geq 1$,*

$$\mathcal{W}_p\left(\nu_{k,\alpha,\beta}, \pi_{\alpha,\beta}\right) \leq C_k \rho_{HB}^k \cdot \mathcal{W}_p\left(\nu_{0,\alpha,\beta}, \pi_{\alpha,\beta}\right), \qquad (26)$$

where $\rho_{HB} = 1 - \frac{2}{\sqrt{k}+1}$ as defined in (13), C_k is defined in (25) and $\mathcal{W}_p$ is the standard the p-Wasserstein distance.

Similarly, for SHB we can show that the suboptimality $\mathbb{E}[f(x_k)] - f(x_*)$ decays linearly in k with the fast rate ρ_{HB} to a constant determined by the variance of the equilibrium distribution.

Theorem 12. *With the same assumptions as in Theorem 11,*

$$\mathbb{E}[f(x_k)] - f(x_*) \leq \frac{L}{2} Tr(X_{HB}) + V_{HB}(\xi_0) \cdot C_k^2 \cdot \rho_{HB}^{2k}, \qquad (27)$$

where $\rho_{HB} = 1 - \frac{2}{\sqrt{\kappa}+1}$ as in (13), C_k is defined in (25), X_{HB} is the covariance matrix of $\xi_\infty - \xi_*$, $V_{HB}(\xi_0)$ is a constant depending on any initial state ξ_0 and both X and $V_{HB}(\xi_0)$ will be spelled out in explicit form in the supplementary file.

4. Strongly convex smooth optimization

In this section, we study the more general case when the objective function f is strongly convex, but not necessarily a quadratic. The proof technique we use for Wasserstein distances can be adapted to obtain a linear rate for a strongly convex objective but this approach does not yield the accelerated rates ρ_{AG} with a $\sqrt{\kappa}$ dependency to the condition number even if the noise magnitude is small. However, we can show accelerated rates in the following alternative metric which implies convergence in the 1-Wasserstein metric. For any two probability measures μ_1, μ_2 on $\mathbb{R}^{2d}$, and any positive constant ψ, we define the weighted total variation distance (introduced by Hairer & Mattingly (2011)) as

$$d_\psi(\mu_1, \mu_2) := \int_{\mathbb{R}^{2d}} (1 + \psi V_P(\xi))|\mu_1 - \mu_2|(d\xi).$$

where V_P is the Lyapunov function defined in (5). Moreover, since ψ and V_P are non-negative, $d_\psi(\mu_1, \mu_2) \geq 2\|\mu_1 - \mu_2\|_{TV}$, where $\|\cdot\|_{TV}$ is the standard total variation norm. Moreover, when $\tilde{P}(2,2) \neq 0$, we will show in the supplementary file (Lemma 27 and Proposition 26) that

$$\mathcal{W}_1(\mu_1, \mu_2) \leq c_0^{-1} d_\psi(\mu_1, \mu_2),$$

for some explicit constant c_0 (to be given in the supplementary file), where $\mathcal{W}_1$ is the standard 1-Wasserstein distance.

We will consider the accelerated stochastic gradient (ASG) method for unconstrained optimization problems. We will also assume in this section that the random gradient error ε_k admits a continuous density so that conditional on $\xi_k = (x_k^T, x_{k-1}^T)^T$, x_{k+1} also admits a continuous density, i.e. $\mathbb{P}(x_{k+1} \in dx|\xi_k = \xi) = p(\xi, x)dx$, where $p(\xi, x) > 0$ is continuous in both ξ and x.

4.1. Accelerated linear convergence of ASG

For the ASG method with any given α, β so that $\rho_{\alpha,\beta}$, $P_{\alpha,\beta}$ satisfy the LMI inequality (6). Let $\nu_{k,\alpha,\beta}$ be the distribution of the k-th iterate ξ_k for $k \geq 0$, where $\xi_k^T := (x_k^T, x_{k-1}^T)$ and the iterates x_k are given in (10) so that $\mathbb{E}[V_{P_{\alpha,\beta}}(\xi_0)]$ is finite. The next result gives a bound of k-th iterate to stationary distribution in the weighted total variation distance d_ψ. We also control the expected suboptimality $\mathbb{E}[f(x_k)] - f(x_*)$ after k iterations.

Theorem 13. *Given any $\eta \in (0,1)$ and $M > 0$ so that $\int_{\|x-x_*\|\leq M} p(\xi_*, x)dx \geq \sqrt{\eta}$, and any $R > 0$ so that*

$$\inf_{\xi\in\mathbb{R}^{2d}, x\in\mathbb{R}^d: V_{P_{\alpha,\beta}}(\xi)\leq R, \|x-x_*\|\leq M} \frac{p(\xi, x)}{p(\xi_*, x)} \geq \sqrt{\eta}.$$

Then there is a unique stationary distribution $\pi_{\alpha,\beta}$ so that

$$\begin{aligned} \mathcal{W}_1(\nu_{k,\alpha,\beta}, \pi_{\alpha,\beta}) &\leq c_0^{-1} d_\psi(\nu_{k,\alpha,\beta}, \pi_{\alpha,\beta}) \\ &\leq (1 - \bar{\eta})^k c_0^{-1} d_\psi(\nu_{0,\alpha,\beta}, \pi_{\alpha,\beta}), \end{aligned}$$

where $\mathcal{W}_1$ is the standard 1*-Wasserstein distance and* $\psi := \frac{\eta}{2K_{\alpha,\beta}}$, $K_{\alpha,\beta} := \left(\frac{L}{2} + \tilde{P}_{\alpha,\beta}(1,1)\right)\alpha^2\sigma^2$ *and* $\bar{\eta} := \min\left\{\frac{\eta}{2}, \left(\frac{1}{2} - \frac{\rho_{\alpha,\beta}}{2} - \frac{K_{\alpha,\beta}}{R}\right)\frac{R\eta}{4K_{\alpha,\beta}+R\eta}\right\}$.

Next, we obtain the optimal convergence rate and provide a bound on the expected suboptimality by choosing $(\alpha, \beta) = (\alpha_{AG}, \beta_{AG})$.

Proposition 14. *Given* $(\alpha, \beta) = (\alpha_{AG}, \beta_{AG})$*. Define* M *and* R *as in Theorem 13 with* $\eta = 1/\kappa^{1/2}$*. Also assume that the noise has small variance, i.e.* $\sigma^2 \leq RL/(4\sqrt{\kappa})$*. Then, with* $\psi := \frac{L}{2\sqrt{\kappa}\sigma^2}$*, we have*

$$\begin{aligned}\mathcal{W}_1(\nu_{k,\alpha,\beta}, \pi_{\alpha,\beta}) &\leq c_0^{-1} d_\psi(\nu_{k,\alpha,\beta}, \pi_{\alpha,\beta}) \qquad (28)\\ &\leq \left(1 - \frac{1}{8\sqrt{\kappa}}\right)^k c_0^{-1} d_\psi(\nu_{0,\alpha,\beta}, \pi_{\alpha,\beta}),\end{aligned}$$

where $\mathcal{W}_1$ *is the standard* 1*-Wasserstein distance and for any initial state* ξ_0*,*

$$\mathbb{E}[f(x_k)] - f(x_*) \leq V_{P_{AG}}(\xi_0)\left(1 - \frac{1}{\sqrt{\kappa}}\right)^k + \frac{\sqrt{\kappa}\sigma^2}{L}. \qquad (29)$$

The bound (29) is similar in spirit to Corollary 4.7. in Aybat et al. (2018) but with a different assumption on noise. We can see that the expected value of the objective with respect to the k-th iterate is close to the true minimum of the objective if k is large, and the variance of the noise σ^2 is small. In the special case when the noise are i.i.d. Gaussian, one can compute the constants in closed-form.

Corollary 15. *If the noise* ε_k *are i.i.d. Gaussian* $\mathcal{N}(0, \Sigma)$*, where* $\Sigma \prec L^2 I_d$*. Then, Proposition 14 holds with*

$$M := \left(-2\log\left(\left(1 - \frac{1}{\kappa^{1/4}}\right)\sqrt{\det(I_d - L^{-2}\Sigma)}\right)\right)^{1/2},$$

$$R := \left(-M + \sqrt{M^2 + \frac{\log(L/\mu)}{2L^2\|\Sigma^{-1}\|}}\right)^2 \frac{(L-\mu)^2}{8(3\sqrt{L} - \sqrt{\mu})^3}.$$

If we take $\mu = \Theta(1)$*, then* $L = \Theta(\kappa)$ *and it follows that we have* $M = O(\kappa^{-1/8})$ *and* $R = O\left(\kappa^{-13/4}\log^2(\kappa)\right)$*.*

We note that Proposition 14 and Corollary 15 provide explicit bounds on the admissable noise level σ^2 to ensure accelerated convergence with respect to Wasserstein distances and expected suboptimality after k iterations.

5. ASPG and the weakly convex setting

Constrained optimization and ASPG. Our analysis for AG can be adapted to study the *accelerated stochastic projected gradient* (ASPG) method for constrained optimization problems $\min_{x \in \mathcal{C}} f(x)$, where $\mathcal{C} \subset \mathbb{R}^d$ is a compact set with diameter $\mathcal{D}_{\mathcal{C}} := \sup_{x,y \in \mathcal{C}} \|x - y\|_2$. Theorem 13, Proposition 14 and Corollary 15 extends to ASPG in a natural fashion with modified constants that reflect the diameter of the constraint set (see the supplementary file). Furthermore, due to the finiteness of the diameter, it can be shown that the metric d_ψ implies the standard p-Wasserstein metric for any $p \geq 1$. We also provide bounds in expected suboptimality for ASPG.

Weakly convex functions. If the objective is (weakly) convex but not strongly convex and the constraint set is bounded, our analysis for the strongly convex case can be adapted with minor modifications. Following standard regularization techniques (see e.g. Lessard et al. (2016); Bubeck (2014)), that allow to approximate a weakly convex function with a strongly convex function, we provide explicit bounds on the noise level to obtain the accelerated $O(\varepsilon^{-1/2})$ rate up to a log factor on ε in expected suboptimality in function values (see the supplementary file).

6. Conclusion

We have studied accelerated convergence guarantees for a number of stochastic momentum methods (SHB, ASG, ASPG) for strongly and (weakly) convex smooth problems. First, we studied the special case when the objective is quadratic and the gradient noise is additive and i.i.d. with a finite second moment. Non-asymptotic guarantees for accelerated linear convergence are obtained for the deterministic and stochastic AG and HB methods for any p-Wasserstein distance ($p \geq 1$), and also for the ASG method in the weighted 2-Wasserstein distance, which builds on the dissipativity theory from the deterministic setting. Our analysis for HB and AG also leads to improved non-asymptotic convergence bounds in suboptimality after k iterations for both deterministic and stochastic settings which is of independent interest. Second, we studied the (non-quadratic) strongly convex optimization under the stochastic oracle model **(H1)–(H2)**. Accelerated linear convergence rate is obtained for the ASG method in the 1-Wasserstein distance. Third, we studied the ASPG method for constrained stochastic strongly convex optimization on a bounded domain. Accelerated linear convergence rate is obtained in any p-Wasserstein distance ($p \geq 1$), and extension to the (weakly) convex setting will be discussed in the supplementary file. Our results provide performance bounds for stochastic momentum methods in expected suboptimality and in Wasserstein distances. Finally, the proofs of all the results in our paper will be given in the supplementary file.

Acknowledgements

Mert Gürbüzbalaban and Bugra Can acknowledge support from the grants NSF DMS-1723085 and NSF CCF-1814888. Lingjiong Zhu is grateful to the support from the grant NSF DMS-1613164.

References

Agarwal, A., Wainwright, M. J., Bartlett, P. L., and Ravikumar, P. K. Information-theoretic lower bounds on the oracle complexity of convex optimization. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 22*, pp. 1–9. Curran Associates, Inc., 2009.

Aybat, N. S., Fallah, A., Gürbüzbalaban, M., and Ozdaglar, A. Robust accelerated gradient methods for smooth strongly convex functions. *arXiv preprint arXiv:1805.10579*, 2018.

Aybat, N. S., Fallah, A., Gurbuzbalaban, M., and Ozdaglar, A. A universally optimal multistage accelerated stochastic gradient method. *arXiv preprint arXiv:1901.08022*, 2019.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pp. 464–473. IEEE, 2014.

Beck, A. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017. doi: 10.1137/1.9781611974997.

Birand, B., Wang, H., Bergman, K., and Zussman, G. Measurements-based power control-a cross-layered framework. In *National Fiber Optic Engineers Conference*, pp. JTh2A–66. Optical Society of America, 2013.

Bubeck, S. Theory of Convex Optimization for Machine Learning. *arXiv preprint arXiv:1405.4980*, May 2014.

Chatterjee, S., Duchi, J. C., Lafferty, J., and Zhu, Y. Local minimax complexity of stochastic convex optimization. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 3423–3431. Curran Associates, Inc., 2016.

Çınlar, E. *Probability and Stochastics*, volume 261. Springer Science & Business Media, New York, 2011.

Cohen, M. B., Diakonikolas, J., and Orecchia, L. On Acceleration with Noise-Corrupted Gradients. *arXiv e-prints*, art. arXiv:1805.12591, May 2018.

Combettes, P. L. and Wajs, V. R. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

d'Aspremont, A. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19(3):1171–1183, 2008. doi: 10.1137/060676386.

Devolder, O., Glineur, F., and Nesterov, Y. Intermediate gradient methods for smooth convex problems with inexact oracle. Technical report, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2013.

Devolder, O., Glineur, F., and Nesterov, Y. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1-2):37–75, 2014.

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

Dieuleveut, A., Flammarion, N., and Bach, F. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1): 3520–3570, 2017b.

Fazlyab, M., Ribeiro, A., Morari, M., and Preciado, V. M. A dynamical systems perspective to convergence rate analysis of proximal algorithms. In *Communication, Control, and Computing (Allerton), 2017 55th Annual Allerton Conference on*, pp. 354–360. IEEE, 2017.

Flåm, S. D. Optimization under uncertainty using momentum. In *Dynamic Stochastic Optimization*, pp. 249–256. Springer, 2004.

Flammarion, N. and Bach, F. From averaging to acceleration, there is only a step-size. In *Conference on Learning Theory*, pp. 658–695, 2015.

Gadat, S., Panloup, F., and Saadane, S. Stochastic heavy ball. *Electronic Journal of Statistics*, 12(1):461–529, 2018.

Gao, X., Gürbüzbalaban, M., and Zhu, L. Global Convergence of Stochastic Gradient Hamiltonian Monte Carlo for Non-Convex Stochastic Optimization: Non-Asymptotic Performance Bounds and Momentum-Based Acceleration. *arXiv preprint arXiv:1809.04618*, September 2018a.

Gao, X., Gurbuzbalaban, M., and Zhu, L. Breaking Reversibility Accelerates Langevin Dynamics for Global Non-Convex Optimization. *arXiv preprint arXiv:1812.07725*, December 2018b.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points–online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pp. 797–842, 2015.

Ghadimi, E., Feyzmahdavian, H. R., and Johansson, M. Global convergence of the Heavy-ball method for convex optimization. *arXiv e-prints*, art. arXiv:1412.7457, December 2014.

Ghadimi, S. and Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

Ghadimi, S. and Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: Shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013. doi: 10.1137/110848876.

Golub, G. H. and Van Loan, C. F. *Matrix computations*. Johns Hopkins University Press, Baltimore, 3rd edition, 1996.

Hairer, M. and Mattingly, J. C. Yet another look at Harris' ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI*, pp. 109–118, Basel, 2011.

Hardt, M. Robustness versus acceleration., August 2014. URL `http://blog.mrtz.org/2014/08/18/robustness-versus-acceleration.html`.

Harris, T. E. The existence of stationary measures for certain Markov processes. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954-1955, vol. II*, pp. 113–124, Berkeley and Los Angeles, 1956.

Hu, B. and Lessard, L. Dissipativity theory for Nesterov's accelerated method. *arXiv preprint arXiv:1706.04381*, 2017.

Hu, C., Pan, W., and Kwok, J. T. Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems*, pp. 781–789, 2009.

Jain, P., Kakade, S. M., Kidambi, R., Netrapalli, P., and Sidford, A. Accelerating stochastic gradient descent. *arXiv preprint arXiv:1704.08227*, 2017.

Karimi, S. and Vavasis, S. A single potential governing convergence of conjugate gradient, accelerated gradient and geometric descent. *arXiv e-prints*, art. arXiv:1712.09498, December 2017.

Lan, G. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133:365–397, 2012.

Lessard, L., Recht, B., and Packard, A. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

Loizou, N. and Richtárik, P. Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods. *arXiv preprint arXiv:1712.09677*, 2017.

Loizou, N. and Richtárik, P. Accelerated gossip via stochastic heavy ball method. *arXiv preprint arXiv:1809.08657*, 2018.

Meyn, S. P. and Tweedie, R. L. *Markov Chains and Stochastic Stability*. Communications and Control Engineering Series. Springer-Verlag, London, 1993.

Meyn, S. P. and Tweedie, R. L. Computable bounds for gemeometric convergence rates of Markov chains. *Annals of Applied Probability*, 4(4):981–1011, 1994.

Neelakantan, A., Vilnis, L., Le, Q. V., Sutskever, I., Kaiser, L., Kurach, K., and Martens, J. Adding gradient noise improves learning for very deep networks. *CoRR*, abs/1511.06807, 2015.

Nesterov, Y. *Introductory Lectures on Convex Optimization. Applied Optimization, Vol. 87*. Kluwer Academic Publishers, Boston, 2004.

Nitanda, A. Stochastic proximal gradient descent with acceleration techniques. In *Advances in Neural Information Processing Systems*, pp. 1574–1582, 2014.

O'Donoghue, B. and Candes, E. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15(3):715–732, 2015.

Parikh, N., Boyd, S., et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

Polyak, B. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1 – 17, 1964. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(64)90137-5.

Polyak, B. T. *Introduction to optimization*. Translations series in mathematics and engineering. Optimization Software, 1987.

Raginsky, M. and Rakhlin, A. Information-based complexity, feedback and dynamics in convex programming. *IEEE Transactions on Information Theory*, 57(10):7036–7056, 2011.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. *arXiv preprint arXiv:1702.03849*, 2017.

Recht, B. Lyapunov analysis and the heavy ball method. *Online lecture notes*, 2012.

Simsekli, U., Sagun, L., and Gurbuzbalaban, M. A tail-index analysis of stochastic gradient noise in deep neural networks. *arXiv preprint arXiv:1901.06053*, 2019.

Su, W., Boyd, S., and Candes, E. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pp. 2510–2518, 2014.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, pp. 1139–1147, 2013.

Vapnik, V. *The nature of statistical learning theory*. Springer science & business media, 2013.

Varga, R. S. *Matrix iterative analysis*, volume 27. Springer Science & Business Media, 2009.

Villani, C. *Optimal Transport: Old and New*. Springer, Berlin, 2009.

Williams, K. S. The nth power of a 2×2 matrix. *Mathematics Magazine*, 65(5):336–336, 1992. doi: 10.1080/0025570X.1992.11996049.

Wilson, A., Recht, B., and Jordan, M. A Lyapunov analysis of momentum methods in optimization. *arXiv preprint arXiv:1611.02635*, 2016.

Xiao, L. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596, 2010.

Yang, T., Lin, Q., and Li, Z. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. *arXiv preprint arXiv:1604.03257*, 2016.