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 $\varepsilon$ centered at a unique invariant distribution in the 1-Wasserstein metric where $\kappa$ 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 $\varepsilon$-optimal solution where $\kappa$ 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 $\nabla f(x)$.
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 $x \in \mathcal{X}$.

(H1) $E\left[\nabla f(x) | x\right] = \nabla f(x),$

(H2) $E\left[\|\tilde{\nabla} f(x) - \nabla f(x)\|^2 | x\right] \leq \sigma^2,$

(see e.g. Bubeck (2014); Lan (2012)). A standard example of this in machine learning is the familiar prediction scenario when $f(x) = E_\theta \ell(x, \theta)$ where $\ell(x, \theta)$ is the (instantaneous) loss of the predictor $x$ on the example $\theta$ 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 $\theta$ from the unknown underlying distribution, and outputs $\tilde{\nabla} f(x) = \nabla_x \ell(x, \theta)$ 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 $\ell$ (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 $\alpha_k = O(1/k^\beta)$ 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., 2017) 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 to 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 $\pi_\alpha$ in the 2-Wasserstein distance requiring $O(\kappa \log(1/\varepsilon))$ iterations to $\varepsilon$ 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 \( \sigma^2 \). 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 \( O(\sqrt{n} \log(1/\varepsilon)) \) in the \( p \)-Wasserstein distance for any \( p \geq 1 \) with explicit constants. The convergence results hold regardless of the noise magnitude \( \sigma \), although \( \sigma \) 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 \( \sigma^2 \) so that ASG method converges linearly to a unique stationary distribution with the accelerated linear rate \( O(\sqrt{n} \log(1/\varepsilon)) \) in the 1-Wasserstein distance. Our results provide convergence rates as a function of \( \alpha, \beta \) and \( \sigma^2 \) that recovers the convergence rate of the AG algorithm as the noise level \( \sigma^2 \) 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 \( O\left( \frac{1}{\sqrt{\varepsilon}} \log(1/\varepsilon) \right) \) 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 \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) is called \( L \)-smooth if its gradient is Lipschitz with constant \( L \). A function \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) is \( \mu \)-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 \( S_{\mu,L} \) denote the class of functions that are convex and \(\mu\)-smooth for some \( L > 0 \). We use \( S_{\mu,L} \) to denote functions that are both \(L\)-smooth and \( \mu \)-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 \( P_p(\mathbb{R}^{2d}) \) as the space consisting of all the Borel probability measures \( \nu \) on \( \mathbb{R}^{2d} \) with the finite \( p \)-th moment (based on the Euclidean norm). For any two Borel probability measures \( \nu_1, \nu_2 \in P_p(\mathbb{R}^{2d}) \), we define the standard \( p \)-Wasserstein metric (see e.g. Villani (2009)):

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

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:

\[
||z_1 - z_2||_S := ((z_1 - z_2)^T S (z_1 - z_2))^{1/2}.
\]

Define \( P_{2,S}(\mathbb{R}^{2d}) \) as the space consisting of all the Borel probability measures \( \nu \) on \( \mathbb{R}^{2d} \) with the finite second moment (based on the \( \| \cdot \|_2 \) norm). For any two Borel probability measures \( \nu_1 \) and \( \nu_2 \) in the space \( P_{2,S}(\mathbb{R}^{2d}) \), the weighted 2-Wasserstein distance is defined as

\[
W_{2,S}(\nu_1, \nu_2) := \left( \inf_{Z_1 \sim \nu_1, Z_2 \sim \nu_2} E[||Z_1 - Z_2||^2_S] \right)^{1/2},
\]

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 \( \nu_1 \) and \( \nu_2 \). Equipped with the 2-Wasserstein distance (1), \( P_{2,S}(\mathbb{R}^{2d}) \) forms a complete metric space (see e.g. Villani (2009)).
Let \( P_{\alpha, \beta}(z, \cdot) \) be a Markov transition kernel (with parameters \( \alpha, \beta \)) associated to a time-homogeneous Markov chain \( \{\xi_k\}_{k \geq 0} \) on \( \mathbb{R}^d \). A Markov transition kernel is the analogue of the transition matrix for finite state spaces. In particular, if \( \xi_0 \) has probability law \( \nu_0 \) then we use the notation that \( \xi_k \) has probability law \( P_k^{\alpha, \beta, \nu_0} \). Given a Borel measurable function \( \varphi : \mathbb{R}^d \to [0, +\infty] \), we also define
\[
(\mathcal{P}_{\alpha, \beta} \varphi)(z) = \int_{\mathbb{R}^d} \varphi(y) P_{\alpha, \beta}(z, dy).
\]
Therefore, it holds that \( E[\varphi(\xi_{k+1})|\xi_k = z] = (\mathcal{P}_{\alpha, \beta} \varphi)(z) \).

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
\[
x_{k+1} = y_k - \alpha \nabla f(y_k), \quad y_k = (1 + \beta)x_k - \beta x_{k-1},
\]
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
\[
\xi_k := (x_k^T \quad x_{k-1}^T)^T \in \mathbb{R}^{2d},
\]
and rewrite the AG iterations in terms of \( \xi_k \). 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
\[
\xi_{k+1} = A \xi_k + B w_k,
\]
where \( A = \tilde{A} \otimes I_d \) and \( B = \tilde{B} \otimes I_d \) with
\[
\tilde{A} := \begin{pmatrix} 1 + \beta & -\beta \\ \alpha & 0 \end{pmatrix}, \quad \tilde{B} := \begin{pmatrix} -\alpha \\ 0 \end{pmatrix},
\]
and \( w_k := \nabla f ((1 + \beta)x_k - \beta x_{k-1}) \). The standard analysis of deterministic AG is based on the following Lyapunov function that combines the state vector and function values:
\[
V_P(\xi_k) := (\xi_k - \xi_*)^T P (\xi_k - \xi_*) + f(x_k) - f_*,
\]
where \( \xi_* = (x_*^T \quad x_*^T)^T \) and \( P \in \mathbb{R}^{2d \times 2d} \) is positive semidefinite 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 \( \rho \) 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 \( 2 \times 2 \) matrix \( P \) (that may depend on \( \rho \)) such that
\[
\begin{pmatrix}
\tilde{A}^T \tilde{P} \tilde{A} - \rho \tilde{P} & \tilde{A}^T \tilde{P} \tilde{B} \\
\tilde{B}^T \tilde{P} \tilde{A} & \tilde{B}^T \tilde{P} \tilde{B}
\end{pmatrix} - \tilde{X} \preceq 0,
\]
where \( \tilde{X} := \rho \tilde{X}_1 + (1 - \rho) \tilde{X}_2 \in \mathbb{R}^{3 \times 3} \) with
\[
\tilde{X}_1 := \begin{pmatrix} \frac{\beta^2 \mu}{2} & -\beta^2 \mu & -\beta \\ -\beta^2 \mu & \frac{\beta^2 \mu}{2} & \beta \\ -\beta & \beta & \frac{2}{\mu} \end{pmatrix},
\]
\[
\tilde{X}_2 := \begin{pmatrix} (1 + \beta)^2 \mu & -\beta (1 + \beta) \mu & -(1 + \beta) \\ -\beta (1 + \beta) \mu & \beta^2 \mu & \beta \\ -(1 + \beta) & \beta & (2\mu - L \alpha) \end{pmatrix},
\]
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
\[
P_{\text{AG}} := \tilde{P}_{\text{AG}} \otimes I_d, \quad \tilde{P}_{\text{AG}} := \tilde{u} \tilde{u}^T,
\]
and \( (\alpha, \beta) = (\alpha_{\text{AG}}, \beta_{\text{AG}}) \) with
\[
\alpha_{\text{AG}} := \frac{1}{L}, \quad \beta_{\text{AG}} := \frac{\sqrt{\gamma} - 1}{\sqrt{\gamma} + 1},
\]
in Theorem 1, we obtain the accelerated convergence rate of \( \rho_{\text{AG}} := 1 - \sqrt{\mu/L} = 1-1/\sqrt{\gamma} \).

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 \( \nabla f(y_k) = \tilde{\nabla} f(y_k) + \epsilon_{k+1} \), where \( \epsilon_{k+1} \) is the random gradient noise. AG algorithm with stochastic gradients has the form
\[
x_{k+1} = y_k - \alpha \tilde{\nabla} f(y_k) + \epsilon_{k+1},
\]
y_k = (1 + \beta)x_k - \beta x_{k-1},
\]
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 \( \xi_k \).

**Assumption 2** (Formal statement of (H1)–(H2)). On some probability space \( (\Omega, \mathcal{F}, \mathbb{P}) \) with a filtration \( \mathcal{F}_k \) the noise \( \epsilon_{k} \)'s are \( \mathcal{F}_k \)-measurable, stationary and
\[
E[\epsilon_k | \mathcal{F}_{k-1}] = 0 \quad \text{and} \quad E[\|\epsilon_k\|^2 | \mathcal{F}_{k-1}] \leq \sigma^2.
\]
Under Assumption 2, the iterations \( \xi_k \) forms a time-homogeneous Markov chain which we will study further in Sections 3 and 4.

### 2.3 HB method

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

\[
x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta (x_k - x_{k-1}),
\]

where \( \alpha > 0 \) is the step size and \( \beta \) 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 \( f \in S_{\mu, L} \) 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

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

Then, \( \|x_k - x_*\| \leq (\rho_{HB} + \delta_k)^k \cdot \|x_0 - x_*\| \), where \( \delta_k \) is a non-negative sequence that goes to zero and

\[
\rho_{HB} := \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} = 1 - \frac{2}{\sqrt{\kappa} + 1}.
\]

Furthermore, \( f(x_k) - f(x_*) \leq \frac{\mu}{2} (\rho_{HB} + \delta_k)^{2k} \cdot \|x_0 - x_*\|^2 \).

This result has an asymptotic nature as the sequence \( \delta_k \) 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 \( \rho_{HB} \); 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 \( \rho_{HB} \). Note that the asymptotic rate \( \rho_{HB} \) of HB in (13) on quadratic problems is strictly (smaller) faster than the rate \( \rho_{AG} \) 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 \( \alpha_{HB} \) and \( \beta_{HB} \) (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:

\[
x_{k+1} = x_k - \alpha (\nabla f(x_k) + \xi_{k+1}) + \beta (x_k - x_{k-1}),
\]

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 \( f \in S_{\mu, L} \) is a quadratic function of the form

\[
f(x) = \frac{1}{2} x^T Q x + a^T x + b,
\]

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 \( \mu I_d \preceq Q \preceq LI_d \) so that \( f \in S_{\mu, L} \). In this section, we assume the noise \( \xi_k \) 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

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

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

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

where \( P_{\alpha, \beta} := P_{\alpha, \beta} \otimes I_d \) and \( \tilde{P}_{\alpha, \beta} \) is a non-zero symmetric positive definite \( 2 \times 2 \) matrix (that may depend on the parameters \( \alpha \) and \( \beta \)) 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 \( \nu_1 \) and \( \nu_2 \) 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 \( P_{\alpha, \beta} \) associated to this chain. Recall that if \( \nu \) is the distribution of \( \xi_0 \), the distribution of \( \xi_k \) is denoted...
by $\mathcal{P}_{\alpha,\beta}^k\nu$. The following theorem shows that this Markov Chain admits a unique equilibrium distribution $\pi_{\alpha,\beta}$ and the distribution of the ASG iterations 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 $W_{2,S_{\alpha,\beta}}$). The proof is given in the supplementary file and it is based on studying the contractivity properties of the map $\nu \mapsto \mathcal{P}_{\alpha,\beta}^k\nu$ in the Wasserstein space.

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

$$W_{2,S_{\alpha,\beta}}(\nu_{k,\alpha,\beta},\pi_{\alpha,\beta}) \leq \mathcal{P}_{\alpha,\beta}^k W_{2,S_{\alpha,\beta}}(\nu_{0,\alpha,\beta},\pi_{\alpha,\beta}),$$

where $W_{2,S_{\alpha,\beta}}$ 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:

$$W_{2,S_{\alpha,\beta}}(\nu_{k,\alpha,\beta},\pi_{\alpha,\beta}) \leq \rho_{AG}^k W_{2,S_{\alpha,\beta}}(\nu_{0,\alpha,\beta},\pi_{\alpha,\beta}),$$

with $\rho_{AG} = 1 - 1/\sqrt{\kappa}$ 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

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

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 $\rho_{AG}$ (Lasserd et al., 2016); improving the iteration complexity by a factor of $4/\sqrt{5} \approx 2.3$ when $\kappa$ is large. However, these results are asymptotic. Below we provide a first non-asymptotic bound with the faster rate $\rho^*_AG$.

**Theorem 5.** Let $f \in S_{\mu,L}$ 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,

$$\|x_k - x_*\| \leq C_k(\rho^*_AG)^{2k} \|\xi_0 - \xi_*\|,$$

$$f(x_k) - f(x_*) \leq \frac{L}{2} (C_k)^2(\rho^*_AG)^{2k} \|\xi_0 - \xi_*\|^2,$$

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

$$C_k^* := \max \left\{ C^*, \sqrt{k^2(\rho^*_AG)^2 + 1} + 2(\rho^*_AG)^2 \right\},$$

with $\tilde{C}^* := \frac{\sqrt{5\kappa + 1} + 2}{\sqrt{5\kappa + 1} - 1} \sqrt{\mu(3L + \mu)}$, where $\{\lambda_i\}_{i=1}^d$ are the eigenvalues of the Hessian $Q$.

**Remark 6.** The constants $C^*_k$ grow 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 $\rho_{AG}$ 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 $f \in S_{\mu,L}$ be a quadratic function (15). Consider the AG iterations $\{x_k\}_{k \geq 0}$ defined by the recursion (10). Let $\nu_{k,\alpha,\beta}$ be the distribution of the $k$-th iterate $\xi_k$ for $k \geq 0$, where $\xi_k := (x_k, x_{-1})$ and parameters $\alpha,\beta = (\alpha_{AG},\beta_{AG})$ as in (19). Also assume that $\nu_{0,\alpha^*,\beta^*} \in \mathcal{P}_p(\mathbb{R}^d)$ and the noise $\varepsilon_k$ has finite $p$-th moment. Then, there exists a unique stationary distribution $\pi_{\alpha,\beta}$ and for any $p \geq 1$,

$$W_p(\nu_{k,\alpha,\beta},\pi_{\alpha,\beta}) \leq C_k^*(\rho^*_AG)^{p} \cdot W_p(\nu_{0,\alpha,\beta},\pi_{\alpha,\beta}),$$

where $\rho^*_AG = 1 - \frac{2}{\sqrt{3\kappa + 1}}$, $C_k^*$ is defined in (21) and $W_p$ 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,

$$\mathbb{E}[f(x_k)] - f(x_*) \leq \frac{L}{2} tr(X^*_AG) + V^*_AG(\xi_0)(C_k^*)^2(\rho^*_AG)^{2k},$$

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 $\xi_0$ 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
imply a bound on the suboptimality $f(x_k) - f(x_*)$. This refines the asymptotic results in the literature (Theorem 3).

**Theorem 9.** Let $f \in S_{\mu,L}$ 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,

$$
\|x_k - x_*\| \leq C_k \rho_{HB}^k \cdot \|\xi_0 - \xi_*\|, \quad (24)
$$

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

where $\rho_{HB}$ is defined by (13) and

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

with $\bar{C} := \max_{i: \mu_i < L} \frac{\mu_i + \lambda_i}{2\sqrt{L - \mu_i}}$, 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 $f \in S_{\mu,L}$ 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 $\xi_k$ for $k \geq 0$, where $\xi_k := (x_k^T, x_{k-1}^T)$ and parameters $(\alpha, \beta) = (\alpha_{HB}, \beta_{HB})$ where $(\alpha_{HB}, \beta_{HB})$ is defined as in (12). Also assume that $\nu_{0,\alpha_{HB},\beta_{HB}} \in \mathcal{P}(\mathbb{R}^{2d})$ and the noise $\xi_k$ has finite $p$-moment. Then, there exists a unique stationary distribution $\pi_{\alpha,\beta}$ and for $p \geq 1$,

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

where $\rho_{HB} = 1 - \frac{1}{\sqrt{k+1}}$ 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 $E[f(x_k)] - f(x_*)$ decays linearly in $k$ with the fast rate $\rho_{HB}$ to a constant determined by the variance of the equilibrium distribution.

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

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

where $\rho_{HB} = 1 - \frac{2}{\sqrt{k+1}}$ 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 $\xi_0$ 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 $\rho_{AG}$ with a $\sqrt{k}$ 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 $\mu_1, \mu_2$ on $\mathbb{R}^{2d}$, and any positive constant $\psi$, 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 $\psi$ and $V_P$ are non-negative, $d_\psi (\mu_1, \mu_2) \geq \|\mu_1 - \mu_2\|_{TV}$, where $\|\cdot\|_{TV}$ is the standard total variation norm. Moreover, when $P(2, 2) \neq 0$, we will show in the supplementary file (Lemma 27 and Proposition 26) that

$$
\mathcal{W}_1 (\mu_1, \mu_2) \leq \epsilon_0^{-1} d_\psi (\mu_1, \mu_2),
$$

for some explicit constant $\epsilon_0$ (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 $\varepsilon_k$ admits a continuous density so that conditional on $\xi_k = (x_k^T, x_{k-1}^T)$, $x_{k+1}$ also admits a continuous density, i.e. $p(x_{k+1} \in dx | \xi_k = \xi) = p(x, x) dx$, where $p(x, x) > 0$ is continuous in both $x$ and $x$.

#### 4.1. Accelerated linear convergence of ASG

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

**Theorem 13.** Given any $\eta \in (0, 1)$ and $M > 0$ so that

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

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

$$
\mathcal{W}_1 (\nu_{k,\alpha,\beta}, \pi_{\alpha,\beta}) \leq \epsilon_0^{-1} d_\psi (\nu_{0,\alpha,\beta}, \pi_{\alpha,\beta}) \leq (1 - \eta)^k \epsilon_0^{-1} d_\psi (\nu_{0,\alpha,\beta}, \pi_{\alpha,\beta}),
$$
where \( \mathcal{W}_1 \) is the standard 1-Wasserstein distance and\( \psi := \frac{n}{2K_{\alpha,\beta}}, K_{\alpha,\beta} := \left( \frac{2}{\alpha + \beta}(1, 1) \right) \alpha^2 \sigma^2 \) and \( \bar{\eta} := \min \left\{ \frac{n}{2}, \left( \frac{1 - \alpha \beta}{8\sqrt{\kappa}} - \frac{K_{\alpha,\beta}}{R} \right) \right\} \).

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

\[
\mathcal{W}_1(\nu_k, \pi_{\alpha,\beta}) \leq c_0^{-1} d_\psi(\nu_k, \pi_{\alpha,\beta}) = \left( 1 - \frac{1}{8\sqrt{\kappa}} \right)^k c_0^{-1} d_\psi(\nu_0, \pi_{\alpha,\beta}),
\]

where \( \mathcal{W}_1 \) is the standard 1-Wasserstein distance and for any initial state \( \xi_0 \),

\[
\mathbb{E}[f(x_k)] - f(x^*) \leq V_{\psi}(\xi_0) \left( 1 - \frac{1}{\sqrt{\kappa}} \right)^k + \frac{\sqrt{\kappa} \sigma^2}{L}.
\]

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 \( \sigma^2 \) 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 \( \varepsilon_k \) are i.i.d. Gaussian \( \mathcal{N}(0, \Sigma) \), where \( \Sigma \prec L^2 I_d \). Then, Proposition 14 holds with

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

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

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

We note that Proposition 14 and Corollary 15 provide explicit bounds on the admissible noise level \( \sigma^2 \) 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 \( D_{\mathcal{C}} := \sup_{x,y \in \mathcal{C}} \|x - y\|_2 \). 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 \( d_\psi \) 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 allows 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 in \( \varepsilon \) 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 AG 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 ASPG 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


Stochastic Momentum Methods


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.


