On the Convergence of Nesterov’s Accelerated Gradient Method in Stochastic Settings

Mahmoud Assran1 2 3 Michael Rabbat2 3

Abstract

We study Nesterov’s accelerated gradient method with constant step-size and momentum parameters in the stochastic approximation setting (unbiased gradients with bounded variance) and the finite-sum setting (where randomness is due to sampling mini-batches). To build better insight into the behavior of Nesterov’s method in stochastic settings, we focus throughout on objectives that are smooth, strongly-convex, and twice continuously differentiable. In the stochastic approximation setting, Nesterov’s method converges to a neighborhood of the optimal point at the same accelerated rate as in the deterministic setting. Perhaps surprisingly, in the finite-sum setting, we prove that Nesterov’s method may diverge with the usual choice of step-size and momentum parameters. Given an initial point $x_0$, and with $x_{-1} = x_0$, the AG method repeats, for $k \geq 0$,

\[
y_{k+1} = x_k + \beta (x_k - x_{k-1})
\]

\[
x_{k+1} = y_{k+1} - \alpha g_{k+1},
\]

where $\alpha$ and $\beta$ are the step-size and momentum parameters, respectively, and in the deterministic setting, $g_{k+1} = \nabla f(y_{k+1})$. When the momentum parameter $\beta = 0$, AG simplifies to standard gradient descent (GD). When $\beta > 0$ it is possible to achieve accelerated rates of convergence for certain combinations of $\alpha$ and $\beta$ in the deterministic setting.

1. Previous Work with Deterministic Gradients

Suppose that the objective function in (1) is $L$-smooth and $\mu$-strongly-convex. Then $f$ is minimized at a unique point $x^*$, and we denote its minimum by $f^* = f(x^*)$. Let $Q := L/\mu$ denote the condition number of $f$. In the deterministic setting, where $g_k = \nabla f(y_k)$ for all $k$, GD with constant step-size $\alpha = 1/(L+\mu)$ converges at the rate (Polyak, 1987)

\[
f(x_k) - f^* \leq \frac{L}{2} \left( \frac{Q - 1}{Q + 1} \right)^{2k} \|x_0 - x^*\|^2.
\]

The AG method with constant step-size $\alpha = 1/L$ and momentum parameter $\beta = \frac{\sqrt{Q} - 1}{\sqrt{Q} + 1}$ converges at the rate (Nesterov, 2004)

\[
f(x_k) - f^* \leq L \left( \frac{\sqrt{Q} - 1}{\sqrt{Q}} \right)^{k} \|x_0 - x^*\|^2.
\]

The rate in (5) matches (up to constants) the tightest-known worst-case lower bound achievable by any first-order black-box method for $\mu$-strongly-convex and $L$-smooth objectives:

\[
f(x_k) - f^* \geq \frac{\mu}{2} \left( \frac{\sqrt{Q} - 1}{\sqrt{Q} + 1} \right)^{2k} \|x_0 - x^*\|^2.
\]

The lower bound (6) is proved in Nesterov (2004) in the infinite dimensional setting under the assumption $Q > 1$. However, the theoretical understanding of accelerated methods remains limited when used with stochastic gradients.

This paper studies the accelerated gradient (AG) method of Nesterov (1983) with constant step-size and momentum parameters. Given an initial point $x_0$, and with $x_{-1} = x_0$, the AG method repeats, for $k \geq 0$,

\[
y_{k+1} = x_k + \beta (x_k - x_{k-1})
\]

\[
x_{k+1} = y_{k+1} - \alpha g_{k+1},
\]

where $\alpha$ and $\beta$ are the step-size and momentum parameters, respectively, and in the deterministic setting, $g_{k+1} = \nabla f(y_{k+1})$. When the momentum parameter $\beta = 0$, AG simplifies to standard gradient descent (GD). When $\beta > 0$ it is possible to achieve accelerated rates of convergence for certain combinations of $\alpha$ and $\beta$ in the deterministic setting.

1. Introduction

First-order stochastic methods have become the workhorse of machine learning, where many tasks can be cast as optimization problems,

\[
\min_{x \in \mathbb{R}^d} f(x).
\]

Methods incorporating momentum and acceleration play an important role in the current practice of machine learning (Sutskever et al., 2013; Bottou et al., 2018), where they are commonly used in conjunction with stochastic gradients.

1Department of Electrical & Computer Engineering, McGill University, Montreal, QC, Canada 2Facebook AI Research, Montreal, QC, Canada 3Mila – Quebec Artificial Intelligence Institute, Montreal, QC, Canada. Correspondence to: Mahmoud Assran <mussran@fb.com>, Michael Rabbat <mikerabbat@fb.com>.

Accordingly, Nesterov’s Accelerated Gradient method is considered optimal in the sense that the convergence rate in (5) depends on $\sqrt{\kappa}$ rather than $Q$.

The proof of (5) presented in Nesterov (2004) uses the method of estimate sequences. Several works have set out to develop better intuition for how the AG method achieves acceleration though other analysis techniques.

One line of work considers the limit of infinitessimally small step-sizes, obtaining ordinary differential equations (ODEs) that model the trajectory of the AG method (Su et al., 2014; Defazio, 2019; Laborde & Oberman, 2019). Allen-Zhu & Orecchia (2014) view the AG method as an alternating iteration between mirror descent and gradient descent and show sublinear convergence of the AG method for smooth convex objectives.

Lessard et al. (2016) and Hu & Lessard (2017) frame the AG method and other popular first-order optimization methods as linear dynamical systems with feedback and characterize their convergence rate using a control-theoretic stability framework. The framework leads to closed-form rates of convergence for strongly-convex quadratic functions with deterministic gradients. For more general (non-quadratic) deterministic problems, the framework provides a means to numerically certify rates of convergence.

1.2. Previous Work with Stochastic Gradients

When Nesterov’s method is run with stochastic gradients $g_{k+1}$, typically satisfying $\mathbb{E}[g_{k+1}] = \nabla f(y_{k+1})$, we refer to it as the accelerated stochastic gradient (ASG) method. In this setting, if $\beta = 0$ then ASG is equivalent to stochastic gradient descent (SGD).

Despite the widespread interest in, and use of, the ASG method, there are no definitive theoretical convergence guarantees. Wiegerinck et al. (1994) study the ASG method in an online learning setting and show that optimization can be modelled as a Markov process but do not provide convergence rates. Yang et al. (2016) study the ASG method in the smooth strongly-convex setting, and show an $O(1/\sqrt{k})$ convergence rate when employed with a diminishing step-size and bounded gradient assumption, but the rates obtained are slower than those for SGD.

Recent work establishes convergence guarantees for the ASG method in certain restricted settings. Aybat et al. (2019) and Kulunchakov & Mairal (2019a) consider smooth strongly-convex functions in a stochastic approximation model with gradients that are unbiased and have bounded variance, and they show convergence to a neighborhood when running the method with constant step size and momentum. Can et al. (2019) further establish convergence in Wasserstein distribution under a stochastic approximation model. Laborde & Oberman (2019) study a perturbed ODE and show convergence for diminishing step-size. Vaswani et al. (2019) study the ASG method with constant step-size and diminishing momentum, and show linear convergence under a strong-growth condition, where the gradient variance vanishes at a stationary point.

Some results are available for other momentum schemes. Loizou & Richtárik (2017) study Polyak’s heavy-ball momentum method with stochastic gradients for randomized linear problems and show that it converges linearly under an exactness assumption. Gitman et al. (2019) characterize the stationary distribution of the Quasi-Hyperbolic Momentum (QHM) method (Ma & Yarats, 2019) around the minimizer for strongly-convex quadratic functions with bounded gradients and bounded gradient noise variance.

The lack of general convergence guarantees for existing momentum schemes, such as Polyak’s and Nesterov’s, have led many authors to develop alternative accelerated methods specifically for use with stochastic gradients (Lan, 2012; Ghadimi & Lan, 2012; 2013; Allen-Zhu, 2017; Kidambi et al., 2018; Cohen et al., 2018; Kulunchakov & Mairal, 2019b; Liu & Belkin, 2020).

Accelerated first-order methods are also known to be sensitive to inexact gradients when the gradient errors are deterministic (possibly adversarial) and bounded (d’Aspremont, 2008; Devolder et al., 2014).

1.3. Contributions

We provide additional insights into the behavior of Nesterov’s accelerated gradient method when run with stochastic gradients by considering two different settings. We first consider the stochastic approximation setting, where the gradients used by the method are unbiased, conditionally independent from iteration to iteration, and have bounded variance. We show that Nesterov’s method converges at an accelerated linear rate to a region of the optimal solution for smooth strongly-convex quadratic problems.

Next, we consider the finite-sum setting, where $f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$, under the assumption that each term $f_i$ is smooth and strongly-convex, and the only randomness is due to sampling one or a mini-batch of terms at each iteration. In this setting we prove that, even when all functions $f_i$ are quadratic, Nesterov’s ASG method with the usual choice of step-size and momentum cannot be guaranteed to converge without making additional assumptions on the condition number and data distribution. When coupled with convergence guarantees in the stochastic approximation setting, this impossibility result illuminates the dichotomy between our understanding of momentum-based methods in the stochastic approximation setting, and practical implementations of these methods in a finite-sum framework.

Our results also shed light as to why Nesterov’s method may
fail to converge or achieve acceleration in the finite-sum setting, providing further insight into what has previously been reported based on empirical observations. In particular, the bounded-variance assumption does not apply in the finite-sum setting with quadratic objectives.

We also suggest choices of the step-size and momentum parameters under which the ASG method is guaranteed to converge for any smooth strongly-convex finite-sum, but where accelerated rates of convergence are no longer guaranteed. Our analysis approach leads to new bounds on the convergence rate of SGD in the finite-sum setting, under the assumption that each term \( f_i \) is smooth, strongly-convex, and twice continuously differentiable.

2. Preliminaries and Analysis Framework

In this section we establish a basic framework for analyzing the AG method. Then we specialize it to the stochastic approximation and finite-sum setting settings, respectively, in Sections 3 and 4.

Throughout this paper we assume that \( f \) is twice-continuously differentiable, \( L \)-smooth, and \( \mu \)-strongly convex, with \( 0 < \mu \leq L \); see e.g., Nesterov (2004); Bubeck (2015). Examples of typical tasks satisfying these assumptions are \( \ell_2 \)-regularized logistic regression and \( \ell_2 \)-regularized least-squares regression (i.e., ridge regression). Taken together, these properties imply that the Hessian \( \nabla^2 f(x) \) exists, and for all \( x \in \mathbb{R}^d \) the eigenvalues of \( \nabla^2 f(x) \) lie in the interval \([\mu, L]\). Also, recall that \( x^* \) denotes the unique minimizer of \( f \) and \( f^* = f(x^*) \).

In contrast to all previous work we are aware of, our analysis focuses on the sequence \( (y_k)_{k\geq 0} \) generated by the method (2)–(3). Let \( r_k := y_k - x^* \) denote the suboptimality of the current iterate, and let \( v_k := x_k - x_{k-1} \) denote the velocity.

Substituting the definition of \( y_{k+1} \) from (2) into (3) and rearranging, we obtain

\[
v_{k+1} = \beta v_k - \alpha g_{k+1}.
\]  

(7)

By using the definition of \( v_k \), substituting (7) and (3) into (2), and rearranging, we also obtain that

\[
r_{k+1} = r_k + \beta^2 v_{k-1} - \alpha (1 + \beta) g_k.
\]  

(8)

Combining (7) and (8), we get the recursion

\[
\begin{bmatrix} r_{k+1} \\ v_k \end{bmatrix} = \begin{bmatrix} I & \beta^2 I \\ 0 & \beta I \end{bmatrix} \begin{bmatrix} r_k \\ v_{k-1} \end{bmatrix} - \alpha \begin{bmatrix} (1 + \beta) I \\ I \end{bmatrix} g_k.
\]  

(9)

Note that \( r_1 = x_0 - x^* \) and \( v_0 = 0 \) based on the common convention that \( x_{-1} = x_0 \).

Our analysis below will build on the recursion (9) and will also make use of the basic fact that if \( f : \mathbb{R}^d \rightarrow R \) is twice continuously differentiable then for all \( x, y \in \mathbb{R}^d \)

\[
\nabla f(y) = \nabla f(x) + \int_0^1 \nabla^2 f(x + t(y - x)) dt \ (y - x).
\]  

(10)

3. The Stochastic Approximation Setting

Now consider the stochastic approximation setting. We assume, for all \( k \), that \( g_k \) is a random vector satisfying

\[
\mathbb{E}[g_k] = \nabla f(y_k)
\]

and that there is a finite constant \( \sigma^2 \) such that

\[
\mathbb{E} \left[ \|g_k - \nabla f(y_k)\|^2 \right] \leq \sigma^2.
\]

Let \( \zeta_k = g_k - \nabla f(y_k) \) denote the gradient noise at iteration \( k \), and suppose that these gradient noise terms are mutually independent. Applying (10) with \( y = y_k \) and \( x = x^* \), we get that

\[
g_k = H_k r_k + \zeta_k, \quad \text{where} \quad H_k = \int_0^1 \nabla^2 f(x^* + tr_k) dt.
\]

Using this in (9), we find that \( r_k \) and \( v_k \) evolve according to

\[
\begin{bmatrix} r_{k+1} \\ v_k \end{bmatrix} = A_k \begin{bmatrix} r_k \\ v_{k-1} \end{bmatrix} - \alpha \begin{bmatrix} (1 + \beta) I \\ I \end{bmatrix} \zeta_k,
\]  

(11)

where

\[
A_k = \begin{bmatrix} I - \alpha (1 + \beta) H_k & \beta^2 I \\ -\alpha H_k & \beta I \end{bmatrix}.
\]  

(12)

Unrolling the recursion (11), we get that

\[
\begin{bmatrix} r_{k+1} \\ v_k \end{bmatrix} = \left( A_k \cdots A_1 \right) \begin{bmatrix} x_0 - x^* \\ 0 \end{bmatrix} - \alpha \begin{bmatrix} (1 + \beta) I \\ I \end{bmatrix} \zeta_k
\]

\[
- \alpha \sum_{j=1}^{k-1} (A_k \cdots A_{j+1}) \begin{bmatrix} (1 + \beta) I \\ I \end{bmatrix} \zeta_j,
\]  

(13)

from which it is clear that we may expect convergence properties to depend on the matrix products \( A_k \cdots A_j \).

3.1. The quadratic case

We can explicitly bound the matrix product in the specific case where \( f(x) = \frac{1}{2} x^T H x - b^T x + c \), for a symmetric matrix \( H \in \mathbb{R}^{d \times d} \), and with \( b \in \mathbb{R}^d \) and \( c \in \mathbb{R} \). In this case, (13) simplifies to

\[
\begin{bmatrix} r_{k+1} \\ v_k \end{bmatrix} = A^k \begin{bmatrix} x_0 - x^* \\ 0 \end{bmatrix} - \alpha \sum_{j=1}^{k} A^{k-j} \begin{bmatrix} (1 + \beta) I \\ I \end{bmatrix} \zeta_j,
\]  

(14)

where

\[
A = \begin{bmatrix} I - \alpha (1 + \beta) H & \beta^2 I \\ -\alpha H & \beta I \end{bmatrix}.
\]  

(15)
We obtain an error bound by ensuring that the spectral radius \( \rho(A) \) of \( A \) is less than 1. In this case we recover the well-known rate for ASG in the deterministic setting. Let \( \Delta = (1 + \beta)^2(1 - \alpha\lambda)^2 - 4\beta(1 - \alpha\lambda) \) and define

\[
\rho(\alpha, \beta) = \begin{cases} 
\frac{1}{2} (1 + \beta)(1 - \alpha\lambda) + \frac{1}{2} \sqrt{\Delta} & \text{if } \Delta \geq 0, \\
\beta (1 - \alpha\lambda) & \text{otherwise}.
\end{cases}
\]

**Theorem 1.** Let \( \rho(\alpha, \beta) = \max\{\rho_A(\alpha, \beta), \rho_L(\alpha, \beta)\} \). If \( \alpha \) and \( \beta \) are chosen so that \( \rho(\alpha, \beta) < 1 \), then for any \( \epsilon > 0 \), there exists a constant \( C_\epsilon \) such that, for all \( k \),

\[
\mathbb{E}\left[\|y_{k+1} - x^*\|^2\right] \leq C_\epsilon \left( (\alpha, \beta) + \epsilon \right)^{2k} \|x_0 - x^*\|^2 + \frac{\alpha^2(1 + \beta)^2 + \beta^2}{1 - \rho(\alpha, \beta)^2} \sigma^2.
\]

Theorem 1 holds with respect to all norms; the constant \( C_\epsilon \) depends on \( \epsilon \) and the choice of norm. Theorem 1 shows that ASG converges at a linear rate to a neighborhood of the minimizer of \( f \), that is proportional to \( \sigma^2 \). The proof is given in Appendix A of the supplementary material, and we provide numerical experiments in Section 3.2 to analyze the tightness of the convergence rate and coefficient multiplying \( \sigma^2 \) in Theorem 1. Comparing to Aybat et al. (2019), we recover the same rate, despite taking a different approach, and the coefficient multiplying \( \sigma^2 \) in Theorem 1 is smaller.

**Corollary 1.1.** Suppose that \( \alpha = \frac{1}{L} \) and \( \beta = \frac{\sqrt{\gamma - 1}}{\sqrt{\gamma + 1}} \). Then for all \( k \),

\[
\mathbb{E}[f(y_{k+1})] - f^* \leq \frac{L}{2} \left( \frac{\sqrt{\gamma} - 1}{\sqrt{\gamma} + 1} + \epsilon_k \right)^{2k} \|x_0 - x^*\|^2 + C_\epsilon \frac{5Q^2 + 2Q^{3/2} + Q}{2L(2\sqrt{\gamma} - 1)(\sqrt{\gamma} + 1)} \sigma^2,
\]

where \( \epsilon_k \sim \left( \sqrt{\gamma} - 1 \right) \).

**Theorem 2.** Let \( f \) be \( L \)-smooth, \( \mu \)-strongly-convex, and twice continuously-differentiable (not necessarily quadratic). Suppose that \( \alpha = \frac{2}{\mu + L} \) and \( \beta = 0 \). Then for all \( k \),

\[
\mathbb{E}[f(y_{k+1})] - f^* \leq \frac{L}{2} \left( \frac{Q - 1}{Q + 1} \right)^{2k} \|x_0 - x^*\|^2 + \frac{Q\sigma^2}{2L}.
\]

Corollary 1.1 confirms that, with the standard choice of parameters, ASG converges at an accelerated rate to a region of the optimizer. Comparing with Theorem 2, which is proved in Appendix B, we see that in the stochastic approximation setting, with bounded variance, ASG not only converges at a faster rate than SGD, the factor multiplying \( \sigma^2 \) also scales more favorably, \( O\left( \sqrt{\gamma}\sigma^2 \right) \) for ASG vs. \( O\left( Q\sigma^2 \right) \) for SGD.

### 3.2. Numerical Experiments

In Figure 1 we visualize runs of the ASG method on a least-squares regression problem for different problem condition numbers \( Q \). The objective \( f \) corresponds to the worst-case quadratic function used to construct the lower bound (6) (Nesterov, 2004), for dimension \( d = 100 \). Stochastic gradients are sampled by adding zero-mean Gaussian noise with variance \( \sigma^2 = 0.0025 \) to the true gradient. The left plots in each sub-figure depict theoretical predictions from Theorem 1, while the right plots in each sub-figure depict empirical results. Each pixel corresponds to an independent run of the ASG method for a specific choice of constant step-size and momentum parameters. In all figures, the area enclosed by the red contour depicts the theoretical stability region from Theorem 1 for which \( \rho(\alpha, \beta) < 1 \).

Figures 1a/1c/1e showcase the coefficient multiplying the variance term, which is taken to be \( \frac{\sigma^2(1 + \beta)^2 + 1}{1 - \rho(\alpha, \beta)^2} \) in theory. Brighter regions correspond to smaller coefficients, while darker regions correspond to larger coefficients. All sets of figures (theoretical and empirical) use the same color scale. We can see that the coefficient of the variance term in Theorem 1 provides a good characterization of the magnitude of the neighbourhood of convergence. The constant \( C_\epsilon \) is approximated as \( 1 + (1 - \rho(\alpha, \beta)^2)(|A|^2 - \rho(\alpha, \beta)^2) \), where \( |A| \) denotes the largest singular value of \( A \) in (15), and \( \rho(\alpha, \beta) \) is the largest eigenvalue of \( A \). More detail on this simple approximation is provided in Appendix A.1 of the supplementary material.

Figures 1b/1d/1f showcase the linear convergence rate in theory and in practice. Brighter regions correspond to faster rates, and darker regions correspond to slower rates. Again, all figures (theoretical and empirical) use the same color scale. We can see that the theoretical linear convergence rates in Theorem 1 provide a good characterization of the empirical convergence rates. Moreover, the theoretical conditions for convergence in Theorem 1 depicted by the red contour appear to be tight.

In short, the theory developed in this section appears to provide an accurate characterization of the ASG method in the stochastic-approximation setting. As we will see in the subsequent section, this theoretical characterization does not reflect its behavior in the finite-sum setting, which is typically closer to practical machine-learning setups, where randomness is due to mini-batching.

### 4. The Finite-Sum Setting

Now consider the finite-sum setting, with

\[
f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x),
\]
where each function \( f_i \) is \( \mu \)-strongly convex, \( L \)-smooth, and twice continuously differentiable. In this setting, stochastic gradients \( g_k \) are obtained by sampling a subset of terms. This can be seen as approximating the gradient \( \nabla f(y_k) \) with a mini-batch gradient

\[
g_k = \sum_{i=1}^{n} \nu_{k,i} \nabla f_i(y_k),
\]

where \( \nu_k \in \mathbb{R}^n \) is a sampling vector with components \( \nu_{k,i} \) satisfying \( E[\nu_{k,i}] = \frac{1}{n} \) (Gower et al., 2019). To simplify the discussion, let us assume that the mini-batch sampled at every iteration \( k \) has the same size, and all elements are given the same weight, so \( \sum_{i=1}^{n} \nu_{k,i} = 1 \), those indices \( i \) which are sampled have \( \nu_{k,i} = \frac{1}{m} \) where \( m \) is the mini-batch size (1 \( \leq m \leq n \)), and \( \nu_{k,i} = 0 \) for all other indices.

4.1. An Impossibility Result

Next we show that even when each function \( f_i \) is well-behaved, the ASG method may diverge when using the standard choice of step-size and momentum. Instability of Nesterov’s method for convex (but not strongly convex) functions with unbounded eigenvalues is shown in Liu & Belkin (2020). This section employs a different proof technique to strengthen this result to the case where each function \( f_i \) is \( \mu \)-strongly-convex and \( L \)-smooth (all eigenvalues bounded between \( \mu \) and \( L \)).

Let us assume that we do not see the same mini-batch twice consecutively; i.e.,

\[
P(||\nu_{k+1} - \nu_k|| > 0) = 1 \quad \text{for all } k.
\]
It is typical in practice to perform training in epochs over the data set, and to randomly permute the data set at the beginning of each epoch, so it is unlikely to see the same mini-batch twice in a row. Note we have not assumed that the sample vectors \( \nu_k \) are independent. We do assume that \( \mathbb{E}_k[\nu_{k, i}] = \frac{1}{n} \), where \( \mathbb{E}_k \) denotes expectation with respect to the marginal distribution of \( \nu_k \).

The interpolation condition is said to hold if the minimizer \( x^* \) of \( f \) also minimizes each \( f_i \); i.e., if \( \nabla f_i(x^*) = 0 \) for all \( i = 1, \ldots, n \). It has been observed in some settings that stronger convergence guarantees can also be obtained when interpolation or a related assumption holds; e.g., (Schmidt & Le Roux, 2013; Loizou & Richtárik, 2017; Ma et al., 2018; Vaswani et al., 2019).

**Theorem 3.** Suppose we run the ASG method (2)–(3) in a finite-sum setting where \( n \geq 3 \) and the sampling vectors \( \nu_k \) satisfy the condition (18). For any initial point \( x_0 \in \mathbb{R}^d \), there exist L-smooth, convex, \( \mu \)-strongly convex quadratic functions \( f_1, \ldots, f_n \) such that \( f \) is also \( L \)-smooth and \( \mu \)-strongly convex, and if we run the ASG method with \( \alpha = 1/L \) and \( \beta = \sqrt{\frac{L}{\mu} - 1} \), then

\[
\lim_{k \to \infty} \mathbb{E}[\|y_k - x^*\|] = \infty.
\]

This is true even if the functions \( f_1, \ldots, f_n \) are required to satisfy the interpolation condition.

**Proof.** We will prove this claim constructively. Given the initial vector \( x_0 \), choose \( x^* \in \mathbb{R}^d \) to be any vector \( x^* \neq x_0 \).

Let \( U \) be an orthogonal matrix. Let the Hessian matrices \( H_i \), \( i = 1, \ldots, n \), be chosen so that they are all diagonalized by \( U \), and let \( \Lambda_i \) denote the diagonal matrix of eigenvalues of \( H_i \); i.e., \( H_i = U \Lambda_i U^\top \). Denote by \( \Lambda_{\nu_k} \) the matrix

\[
\Lambda_{\nu_k} = \sum_{i=1}^{n} \nu_{k, i} \Lambda_i.
\]

It follows that \( \Lambda_{\nu_k} \in \mathbb{R}^{d \times d} \) is also diagonal, and all of its diagonal entries are in \([\mu, L]\).

Recall that we have assumed that the functions \( f_i \) are quadratic: \( f_i(x) = \frac{1}{2} x^\top H_i x - b_i^\top x + c_i \). Let us assume that \( b_i \in \mathbb{R}^d \) and \( c_i \in \mathbb{R} \) are chosen so that all functions \( f_i \) are minimized at the same point \( x^* \), satisfying the interpolation condition. Then from (10), we have

\[
g_k = U \Lambda_{\nu_k} U^\top r_k.
\]

Using this in (9) and unrolling, we obtain that

\[
\begin{bmatrix} r_{k+1} \\ u_k \end{bmatrix} = \begin{bmatrix} A_k & A_k-1 & \cdots & A_1 & v_1 \\ v_0 \end{bmatrix} \begin{bmatrix} r_1 \\ u_0 \end{bmatrix},
\]

where

\[
A_j = \begin{bmatrix} I - \alpha(1 + \beta)U \Lambda_{\nu_j} U^\top & \beta^2 I \\ -\alpha U \Lambda_{\nu_j} U^\top & \beta I \end{bmatrix}.
\]

For fixed \( n \) and \( m \), there are a finite number of sampling vectors \( \nu_k \) (precisely \( \binom{n}{m} \)), and therefore the matrices \( A_j \) belong to a bounded set \( \mathcal{A} \). It follows that the trajectory \( \{\|r_k\|, \nu_k\} \) is stable if the joint spectral radius of the set of matrices \( \mathcal{A} \) is less than one (Rota & Strang, 1960). Conversely, if \( \mathbb{E}[\rho(A_k \cdots A_1)^{1/k}] > 1 \) for all \( k \) sufficiently large, then \( \lim_{k \to \infty} \|y_k - x^*\| = \infty \).

Based on the construction above, the norm of the matrix product \( A_k \cdots A_1 \) in (21) can be characterized by studying products of smaller \( 2 \times 2 \) matrices of the form

\[
B(\lambda_{k, j}) = \begin{bmatrix} 1 - \alpha(1 + \beta)\lambda_{k, j} & \beta^2 \alpha \\ -\alpha \lambda_{k, j} & \beta \end{bmatrix},
\]

where \( \lambda_{k, j} \) is a diagonal entry of \( \Lambda_{\nu_k} \). To see this, observe that there is a permutation matrix \( P \in \{0, 1\}^{2d \times 2d} \) such that (see Appendix C)

\[
P \begin{bmatrix} U \top & 0 \\ 0 & U \top \end{bmatrix} A_j \begin{bmatrix} U \top & 0 \\ 0 & U \top \end{bmatrix} P^\top
\]

\[
= \begin{bmatrix} B(\lambda_{j, 1}) & \cdots & 0 \\ 0 & B(\lambda_{j, 2}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B(\lambda_{j, d}) \end{bmatrix},
\]

where \( \lambda_{j, i} \) is the \( i \)th diagonal entry of \( \Lambda_{\nu_j} \).

Furthermore, since all matrices \( H_i \) have the same eigenvectors, we have that

\[
P \begin{bmatrix} U \top & 0 \\ 0 & U \top \end{bmatrix} A_k A_{k-1} \cdots A_1 \begin{bmatrix} U \top & 0 \\ 0 & U \top \end{bmatrix} P^\top
\]

\[
= \begin{bmatrix} T_{k, 1} & 0 & \cdots & 0 \\ 0 & T_{k, 2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & T_{k, d} \end{bmatrix},
\]

where \( T_{k, j} = B(\lambda_{k, j}) \cdots B(\lambda_{1, j}) \). Hence, the spectral radius of the product \( A_k \cdots A_1 \) corresponds to the maximum spectral radius of any of the \( 2 \times 2 \) matrices \( T_{k, j} \), \( j = 1, \ldots, d \).

Let \( j \) index a subspace such that \( u_j^\top \nu_1 \neq 0 \), where \( u_j \) is the \( j \)th column of \( U \). To simplify the discussion, suppose that all mini-batches are of size \( m = 1 \), and assume \( n > 1 \).

Since we can define the Hessians of the functions \( f_i \) such that the eigenvalues pair together arbitrarily, consider matrix products of the form

\[
T_{k, j} = B(L)B(\mu)^{k_1}B(L)B(\mu)^{k_2} \cdots B(L)B(\mu)^{k_s},
\]

where \( k = k_1 + \cdots + k_s + s \). That is, all but one of the functions \( f_i \) have the eigenvalue \( \mu \) in this subspace, and the remaining one has eigenvalue \( L \) in this subspace. Hence, most of the time we sample mini-batches corresponding to \( B(\mu) \),
The divergence result in Theorem 3 stems from the fact that the algorithm acquires momentum along a low-curvature direction, and then, suddenly, a high-curvature mini-batch is sampled that overshoots along the current trajectory. Momentum prevents the iterates from immediately adapting to the overshoot, and propels the iterates away from the minimizer for several consecutive iterations.

To illustrate this effect, consider the following example finite-sum problem with $d = 3$, where each function $f_i$ is a strongly-convex quadratic with gradient

$$\nabla f_i(x) = U A_i U^T (x - x^*).$$

For simplicity, take $U = I$, and let

$$A_i = \begin{bmatrix} L & 0 & 0 \\ 0 & \mu & 0 \\ 0 & 0 & \lambda_i \end{bmatrix}.$$ 

The scalar $\lambda_i$ is equal to $\mu$ for all $i \neq n$, and is equal to $L$ for $i = n$. Therefore, each function $f_i$ is $\mu$-strongly convex, $L$-smooth, and minimized at $x^*$, and the global objective $f$ is also $\mu$-strongly convex, $L$-smooth, and minimized at $x^*$. Moreover, the functions $f_i$ are nearly all identical, except for $f_n$, which we refer to as the inconsistent mini-batch.

From the proof of Theorem 3, the growth rate of the iterates along the third coordinate direction, with the usual choice of parameters $(\alpha = 1/L, \beta = \sqrt[Q-1]{Q/(Q+1)})$, is

$$\mathbb{E}[\rho(A_k \cdots A_1)^{1/k}] \sim \left(\frac{\sqrt[Q-1]{Q}}{\sqrt[Q]{Q}}\right) (n-1)^{\frac{1}{n}}.$$

Notice that the term $(n-1)^{\frac{1}{n}}$ goes to 1 as $n$ grows to infinity. Hence, for a fixed condition number $Q$, the ASG method exhibits an increased probability of convergence as

and once in a while we sample mini-batches with $B(L)$. Moreover, since we do not sample the same mini-batch twice consecutively, we never see back-to-back $B(L)$'s. For this case, and with the standard choice of step-size and momentum parameters, we can precisely characterize the spectral radius of $T_{k,j}$.

**Lemma 1.** If $\alpha = 1/L$ and $\beta = \frac{\sqrt{Q} - 1}{\sqrt{Q} + 1}$, then

$$\rho(T_j) = \left(\frac{\sqrt{Q} - 1}{\sqrt{Q}}\right)^k \times k_1 \cdots k_s.$$ 

The proof of Lemma 1 is given in Appendix D. Since we do not sample the same mini-batch twice in a row, it follows that $k_{j} \geq 1$ for all $j = 1, \ldots, s$. Based on the assumption that $\mathbb{E}[\nu_{k,i}] = \frac{1}{n}$, we have $\mathbb{E}[s] = \frac{k}{n}$. Moreover, since $\mathbb{E}[\nu_{k,i}] = \frac{1}{n}$, for large $k$ ($\gg n$) we have $\mathbb{E}[(k_1 \cdots k_s)^{\frac{1}{k}}] \approx (n - 1)^{1/n}$. Thus, for sufficiently large $Q$ and sufficiently large $k$,

$$\mathbb{E}[\rho(A_k \cdots A_1)^{1/k}] > 1.$$ 

Therefore, $\lim_{k \to \infty} \mathbb{E}[\|y_k - x^*\| = \infty$.

Recall that we assumed the interpolation condition holds in order to get $g_k$ of the form (20). If we relax this and do not require interpolation, then $g_k$ will have an additional term involving $\nabla f_i(x^*)$, and the expression (21) will also have additional terms, akin to the $z_k$ terms in (13). The same arguments still apply, leading to the same conclusion. 

**4.2. Example**

The divergence result in Theorem 3 stems from the fact that the algorithm acquires momentum along a low-curvature coordinate, and then, suddenly, a high-curvature mini-batch is sampled that overshoots along the current trajectory. Momentum prevents the iterates from immediately adapting to the overshoot, and propels the iterates away from the minimizer for several consecutive iterations.

To illustrate this effect, consider the following example finite-sum problem with $d = 3$, where each function $f_i$ is a strongly-convex quadratic with gradient

$$\nabla f_i(x) = U A_i U^T (x - x^*).$$

For simplicity, take $U = I$, and let

$$A_i = \begin{bmatrix} L & 0 & 0 \\ 0 & \mu & 0 \\ 0 & 0 & \lambda_i \end{bmatrix}.$$ 

The scalar $\lambda_i$ is equal to $\mu$ for all $i \neq n$, and is equal to $L$ for $i = n$. Therefore, each function $f_i$ is $\mu$-strongly convex, $L$-smooth, and minimized at $x^*$, and the global objective $f$ is also $\mu$-strongly convex, $L$-smooth, and minimized at $x^*$. Moreover, the functions $f_i$ are nearly all identical, except for $f_n$, which we refer to as the inconsistent mini-batch.

From the proof of Theorem 3, the growth rate of the iterates along the third coordinate direction, with the usual choice of parameters $(\alpha = 1/L, \beta = \frac{\sqrt{Q} - 1}{\sqrt{Q} + 1})$, is

$$\mathbb{E}[\rho(A_k \cdots A_1)^{1/k}] \sim \left(\frac{\sqrt[Q-1]{Q}}{\sqrt[Q]{Q}}\right) (n-1)^{\frac{1}{n}}.$$ 

Notice that the term $(n-1)^{\frac{1}{n}}$ goes to 1 as $n$ grows to infinity. Hence, for a fixed condition number $Q$, the ASG method exhibits an increased probability of convergence as

and once in a while we sample mini-batches with $B(L)$. Moreover, since we do not sample the same mini-batch twice consecutively, we never see back-to-back $B(L)$'s. For this case, and with the standard choice of step-size and momentum parameters, we can precisely characterize the spectral radius of $T_{k,j}$.

**Lemma 1.** If $\alpha = 1/L$ and $\beta = \frac{\sqrt{Q} - 1}{\sqrt{Q} + 1}$, then

$$\rho(T_j) = \left(\frac{\sqrt{Q} - 1}{\sqrt{Q}}\right)^k \times k_1 \cdots k_s.$$ 

The proof of Lemma 1 is given in Appendix D. Since we do not sample the same mini-batch twice in a row, it follows that $k_{j} \geq 1$ for all $j = 1, \ldots, s$. Based on the assumption that $\mathbb{E}[\nu_{k,i}] = \frac{1}{n}$, we have $\mathbb{E}[s] = \frac{k}{n}$. Moreover, since $\mathbb{E}[\nu_{k,i}] = \frac{1}{n}$, for large $k$ ($\gg n$) we have $\mathbb{E}[(k_1 \cdots k_s)^{\frac{1}{k}}] \approx (n - 1)^{1/n}$. Thus, for sufficiently large $Q$ and sufficiently large $k$,

$$\mathbb{E}[\rho(A_k \cdots A_1)^{1/k}] > 1.$$ 

Therefore, $\lim_{k \to \infty} \mathbb{E}[\|y_k - x^*\| = \infty$.

Recall that we assumed the interpolation condition holds in order to get $g_k$ of the form (20). If we relax this and do not require interpolation, then $g_k$ will have an additional term involving $\nabla f_i(x^*)$, and the expression (21) will also have additional terms, akin to the $z_k$ terms in (13). The same arguments still apply, leading to the same conclusion. 

**4.2. Example**

The divergence result in Theorem 3 stems from the fact that the algorithm acquires momentum along a low-curvature coordinate, and then, suddenly, a high-curvature mini-batch is sampled that overshoots along the current trajectory. Momentum prevents the iterates from immediately adapting to the overshoot, and propels the iterates away from the minimizer for several consecutive iterations.

To illustrate this effect, consider the following example finite-sum problem with $d = 3$, where each function $f_i$ is a strongly-convex quadratic with gradient

$$\nabla f_i(x) = U A_i U^T (x - x^*).$$

For simplicity, take $U = I$, and let

$$A_i = \begin{bmatrix} L & 0 & 0 \\ 0 & \mu & 0 \\ 0 & 0 & \lambda_i \end{bmatrix}.$$ 

The scalar $\lambda_i$ is equal to $\mu$ for all $i \neq n$, and is equal to $L$ for $i = n$. Therefore, each function $f_i$ is $\mu$-strongly convex, $L$-smooth, and minimized at $x^*$, and the global objective $f$ is also $\mu$-strongly convex, $L$-smooth, and minimized at $x^*$. Moreover, the functions $f_i$ are nearly all identical, except for $f_n$, which we refer to as the inconsistent mini-batch.

From the proof of Theorem 3, the growth rate of the iterates along the third coordinate direction, with the usual choice of parameters $(\alpha = 1/L, \beta = \frac{\sqrt{Q} - 1}{\sqrt{Q} + 1})$, is

$$\mathbb{E}[\rho(A_k \cdots A_1)^{1/k}] \sim \left(\frac{\sqrt[Q-1]{Q}}{\sqrt[Q]{Q}}\right) (n-1)^{\frac{1}{n}}.$$ 

Notice that the term $(n-1)^{\frac{1}{n}}$ goes to 1 as $n$ grows to infinity. Hence, for a fixed condition number $Q$, the ASG method exhibits an increased probability of convergence as
Then for all $k$, Annotated red points indicate iterations at which the mini-batch is sampled. The shaded windows illustrate that immediately after the inconsistent mini-batch is sampled, the gradient and momentum buffer have opposite signs for several consecutive iterations.

### 4.3. Convergent Parameters

Next we turn our attention to finding alternative settings for the parameters $\alpha$ and $\beta$ in the ASG method which guarantee convergence in the finite-sum setting. Vaswani et al. (2019) obtain linear convergence under a strong growth condition using an alternative formulation of ASG which has multiple momentum parameters by keeping the step-size constant and having the momentum parameters vary. Here we focus on constant step-size and momentum and make no assumptions about growth.

Our approach is to bound the spectral norm of the products $\|A_k \cdots A_j\|$ using submultiplicativity of matrix norms. This recovers linear convergence to a neighborhood of the minimizer, but the rate is no longer accelerated.

Define the quantities

$$C_\lambda(\alpha, \beta) = (1 - \alpha(1 + \beta))\lambda^2 + \alpha^2\lambda^2 + \beta(\beta^2 + 1)$$

$$\Delta_\lambda(\alpha, \beta) = C_\lambda(\alpha, \beta)^2 - 4\beta^2(1 - \alpha\lambda)^2$$

$$R_\lambda(\alpha, \beta) = \frac{1}{\sqrt{2}} \left( C_\lambda(\alpha, \beta) + \sqrt{\Delta_\lambda(\alpha, \beta)} \right)^{1/2}$$

and let $R(\alpha, \beta) = \max_{\lambda \in [\mu, L]} R_\lambda(\alpha, \beta)$.

**Theorem 4.** Let $\alpha$ and $\beta$ be chosen so that $R(\alpha, \beta) < 1$. Then for all $k \geq 0$,

$$\mathbb{E}\|y_{k+1} - x^*\| \leq R(\alpha, \beta)^k \|y_0 - x^*\| + \frac{\alpha}{1 - R(\alpha, \beta)} \sigma,$$

where $\sigma = \frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(x^*)\|$.

Theorem 4 is proved in Appendix E for general $L$-smooth $\mu$-strongly-convex functions. Note that if an interpolation condition holds (a weaker assumption than the strong growth condition), then $\sigma = 0$.

Theorem 4 shows that the ASG method can be made to converge in the finite-sum setting for $L$-smooth $\mu$-strongly convex objective functions when run with constant step-sizes. In particular, the algorithm converges at a linear rate to a neighborhood of the minimizer that is proportional to the variance of the noise terms. Note that this theorem also allows for negative momentum parameters. Using the spectral norm to guarantee stability is restrictive, in that it is sufficient but not necessary. There may be values of $\alpha$ and $\beta$ for which $R(\alpha, \beta) \geq 1$ and the algorithm still converges. Having $R(\alpha, \beta) < 1$ ensures that $\|r_k\| + \|\epsilon_k\|$ decreases at every iteration.

**Corollary 4.1.** Suppose that $\alpha < \frac{1}{2} L$ and $\beta = 0$. Then for all $k \geq 0$

$$\mathbb{E}\|y_k - x^*\| \leq \phi(\alpha)k \|y_0 - x^*\| + \frac{\alpha}{1 - \phi(\alpha)} \sigma,$$

where $\phi(\alpha) = \max_{\lambda \in [\mu, L]} |1 - \alpha\lambda|$.

Corollary 4.1 is proved in Appendix F for smooth strongly-convex functions, and shows the convergence of SGD in the finite-sum setting without making any assumptions on the noise distribution.

**Corollary 4.2.** Suppose that $\alpha = \frac{1}{2} L$ and $\beta = 0$. Then for all $k \geq 0$

$$\mathbb{E}\|y_k - x^*\| \leq \left( \frac{Q - 1}{Q + 1} \right)^k \|y_0 - x^*\| + \frac{1}{\mu} \sigma.$$

Corollary 4.2 is proved in Appendix F for smooth strongly-convex functions, and shows that SGD converges to a neighborhood of $x^*$ at the same linear rate as GD, viz. (4), in the finite-sum setting, without making any assumptions on the noise distribution, such as the strong-growth condition; a novel result to the best of our knowledge. Moreover, when the interpolation condition holds, we have that $\sigma = 0$.

Figure 3 illustrates the tightness of the convergence rate and variance bound in Corollary 4.2 when minimizing randomly generated least-squares problems with various condition numbers. The finite-sum least-squares problem consists of 25000 data samples, with 2 features each, partitioned into 50 mini-batches, each with condition number $Q$. At each iteration, one of the 50 mini-batches is sampled to compute a stochastic gradient step. Dashed lines indicate the theoretical convergence rate and variance bound from Corollary 4.2. Solid lines indicate the empirical convergence observed in practice. The convergence rate and variance bound in Corollary 4.2 provide a tight characterization of the SGD convergence observed in practice.
We also believe that future work understanding the role that negative momentum parameters play in practice may lead to improved optimization of machine learning models. All convergence guarantees and variance bounds in this paper hold for both positive and negative momentum parameters.

Our variance bounds and theoretical rates support the observation that negative momentum parameters may slow-down convergence, but can also lead to non-trivial variance reduction. Previous work has found negative momentum to be useful in asynchronous distributed optimization (Mitliagkas et al., 2016) and for stabilizing adversarial training (Gidel et al., 2018). Although it is almost certainly not possible (in general) to obtain zero variance solutions by only using negative momentum parameters, for Deep Learning practitioners that already use the ASG method to train their models, perhaps momentum schedules incorporating negative values towards the end of training can improve performance.

Acknowledgements

We thank Leon Bottou, Aaron Defazio, Alexandre Defossez, Tom Goldstein, and Mark Tygert for feedback and conversations about earlier versions of this work.

References


On the Convergence of Nesterov’s Accelerated Gradient Method


A. Proof of Theorem 1

We begin from (14). By taking the squared norm on both sides, and recalling that the random vectors $\zeta_k$ have zero mean and are mutually independent, we have

$$E \|y_{k+1} - x^*\|^2 \leq E \left\| \frac{r_{k+1}}{v_k} \right\|^2$$

$$= E_{\zeta_k, \ldots, \zeta_1} \left[ \left\| A^k \begin{bmatrix} x_1 - x^* \\ 0 \end{bmatrix} - \alpha \sum_{j=1}^k A^{k-j} \begin{bmatrix} (1 + \beta)I \\ I \end{bmatrix} \zeta_j \right\|^2 \right]$$

$$= E_{\zeta_k} \left[ \cdots \left\| A^{k} \begin{bmatrix} x_1 - x^* \\ 0 \end{bmatrix} - \alpha \sum_{j=1}^k A^{k-j} \begin{bmatrix} (1 + \beta)I \\ I \end{bmatrix} \zeta_j \right\|^2 \right] \cdots$$

$$= \left\| A^{k} \begin{bmatrix} x_1 - x^* \\ 0 \end{bmatrix} \right\|^2 + \alpha^2 \sum_{j=1}^k E_{\zeta_j} \left\| A^{k-j} \begin{bmatrix} (1 + \beta)I \\ I \end{bmatrix} \zeta_j \right\|^2$$

$$\leq \left\| A^{k} \right\|^2 \left\| x_1 - x^* \right\|^2 + \alpha^2(1 + \beta)^2(1 + \beta \sigma^2) \sum_{j=1}^k \left\| A^{k-j} \right\|^2.$$  \hspace{1cm} (25)

Recall that the spectral radius of a square matrix $A \in \mathbb{R}^{2d \times 2d}$ is defined as $\max_{i=1, \ldots, 2d} |\lambda_i(A)|$, where $\lambda_i(A)$ is the $i$th eigenvalue of $A$. The spectral radius satisfies (Horn & Johnson, 2013)

$$\rho(A)^k \leq \left\| A^k \right\| \quad \text{for all } k,$$

and (Gelfand’s theorem)

$$\lim_{k \to \infty} \left\| A^k \right\|^{1/k} = \rho(A).$$

Hence, for any $\epsilon > 0$, there exists a $K_\epsilon$ such that $\left\| A^k \right\|^{1/k} \leq (\rho(A) + \epsilon)$ for all $k \geq K_\epsilon$. Let

$$C_\epsilon = \max_{k < K_\epsilon} \left\{ 1, \frac{\left\| A^k \right\|}{(\rho(A) + \epsilon)^k} \right\}. \hspace{1cm} (26)$$

Then $\left\| A^k \right\| \leq C_\epsilon (\rho(A) + \epsilon)^k$ for all $k$. Moreover, if $\left\| A^k \right\|^{1/k}$ converges monotonically to $\rho(A)$, then $C_\epsilon \leq \|A\| / \rho(A)$.

Now, recall that we have assumed $f(x) = \frac{1}{2} x^T H x - b^T x + c$ where $H \in \mathbb{R}^{d \times d}$ is symmetric, and we have also assumed that $f$ is $L$-smooth and $\mu$-strongly convex. Thus all eigenvalues of $H$ satisfy $\mu \leq \lambda_i(H) \leq L$.

**Lemma 2.** For $A$ as defined in (15), we have $\rho(A) = \max \{\rho_\mu(\alpha, \beta), \rho_L(\alpha, \beta)\}$ where

$$\rho_\lambda(\alpha, \beta) = \begin{cases} \frac{1}{2} (1 + \beta)(1 - \alpha \lambda) + \frac{1}{2} \sqrt{\Delta \lambda} & \text{if } \Delta \lambda \geq 0, \\ \frac{1}{\sqrt{\beta(1 - \alpha \lambda)}} & \text{otherwise}, \end{cases}$$

and $\Delta \lambda = (1 + \beta)^2(1 - \alpha \lambda)^2 - 4\beta(1 - \alpha \lambda)$. 


Proof. Since $H$ is real and symmetric, it has a real eigenvalue decomposition $H = UΛ_H U^T$, where $U \in \mathbb{R}^{d \times d}$ is an orthogonal matrix and $Λ_H$ is the diagonal matrix of eigenvalues of $H$. Observe that $A$ can be viewed as a $2 \times 2$ block matrix with $d \times d$ blocks that all commute with each other, since each block is an affine matrix function of $H$. Thus, by Polyak (1964, Lemma 5), $ξ$ is an eigenvalue of $A$ if and only if there is an eigenvalue $λ$ of $H$, such that $ξ$ is an eigenvalue of the $2 \times 2$ matrix

$$B(λ) := \begin{bmatrix} 1 - α(1 + β)λ & β^2 \\ -αλ & β \end{bmatrix}. \quad (27)$$

The characteristic polynomial of $B(λ)$ is

$$ξ^2 - (1 + β)(1 - αλ)ξ + β(1 - αλ) = 0,$$

from which it follows that eigenvalues of $B(λ)$ are given by $ρ(α, β)$; see, e.g., Lessard et al. (2016, Appendix A). Note that the characteristic polynomial of $B(λ)$ is the same as the characteristic polynomial of a different matrix appearing in Lessard et al. (2016), that arises from a different analysis of the AG method. Finally, as discussed in Lessard et al. (2016), for any fixed values of $α$ and $β$, the function $ρ(α, β)$ is quasi-convex in $λ$, and hence the maximum over all eigenvalues of $A$ is achieved at one of the extremes $λ = μ$ or $λ = L$.

To complete the proof of Theorem 1, use Lemma 2 with (25) to obtain that, for any $ε > 0$, there is a positive constant $C_ε$ such that

$$\mathbb{E}[∥y_{k+1} - x^*∥^2] \leq C_ε \left( (ρ(λ) + ε)^{2k} ∥x_0 - x^*∥^2 + α^2((1 + β)^2 + 1)σ^2(ρ(λ) + ε)^{2(k-j)} \right) \leq C_ε \left( (ρ(λ) + ε)^{2k} ∥x_0 - x^*∥^2 + \frac{α^2((1 + β)^2 + 1)}{1 - (ρ(λ) + ε)^2}σ^2 \right).$$

A.1. Estimating the constant $C_ε$

For the theoretical plots in the numerical experiments in Section 3.2 and in Appendix G below, we estimate the constant $C_ε$ by taking $K_ε \approx 2$ in (26). That is, for arbitrarily small $ε$ and all $k \geq 2$, we approximate $∥A^k∥^{1/k}$ by $(ρ(λ) + ε)$. Therefore, the summation term in (25) is approximated as

$$α^2((1 + β)^2 + 1) \left( \frac{1}{1 - ρ(α, β)^2} + (∥A∥^2 - ρ(α, β)^2) \right), \quad (28)$$

where $∥A∥$ denotes the largest singular value of $A$ in (15), and $ρ(α, β)$ is the largest eigenvalue of $A$. The first term in (28) corresponds to the geometric limit of the summation term in (25) after taking matrix norms and approximating the norms of matrix products by powers of the spectral radius for all products $k \geq 2$. The difference term in (28) is simply used to correct for the case $k = 1$. Setting $C_ε \frac{α^2((1 + β)^2 + 1)}{1 - ρ(α, β)^2}$ equal to (28) and solving for $C_ε$ gives us the approximate expression for $C_ε$ used in the theoretical plots in Section 3.2.

B. Proofs of Corollary 1.1 and Theorem 2

Taking $α = 1/L$ and $β = \frac{\sqrt{Q} - 1}{\sqrt{Q} + 1}$, we find that $ρ(α, β) = \frac{\sqrt{Q} - 1}{\sqrt{Q} + 1}$. Since $f(x) = \frac{1}{2}x^THx - b^Tx + c$ is an $L$-smooth $μ$-strongly convex quadratic, all eigenvalues of $H$ are bounded between $μ$ and $L$. Therefore, from Polyak (1964, Lemma 5), we have that $∥A^k∥_2 \leq \max_{λ∈[μ, L]} ∥B(λ)^k∥_2 \leq \max_{λ∈[μ, L]} \sqrt{d} ∥B(λ)^k∥_∞$, where $B(λ)$ is as defined in (27). The eigenvalues of $B(λ)^k$ are maximized at $λ = μ$ for $k > 1$, therefore, for large $k$, $∥B(λ)^k∥_∞$ is maximized at $λ = μ$.

Note that the Jordan form of $B(μ)$ is given by $VJV^{-1}$, where

$$V = \begin{bmatrix} \frac{\sqrt{Q} - 1}{\sqrt{Q} + 1} & Q \\ -1 & 0 \end{bmatrix} \quad \text{and} \quad J = \begin{bmatrix} \frac{\sqrt{Q} - 1}{\sqrt{Q}} & 1 \\ 0 & \frac{1}{\sqrt{Q}} \end{bmatrix}.$$
On the Convergence of Nesterov’s Accelerated Gradient Method

Using the Jordan form, we determine that \( B(\mu)^k \) is

\[
B(\mu)^k = \begin{bmatrix}
(1 + \frac{k}{\sqrt{Q}+1}) \left( \frac{\sqrt{Q}-1}{\sqrt{Q}} \right)^k & k \left( \frac{\sqrt{Q}-1}{\sqrt{Q}+1} \right)^2 \left( \frac{\sqrt{Q}-1}{\sqrt{Q}} \right)^{k-1} \\
- \frac{k}{Q} \left( \frac{\sqrt{Q}-1}{\sqrt{Q}} \right)^{k-1} & (1 - \frac{k}{\sqrt{Q}+1}) \left( \frac{\sqrt{Q}-1}{\sqrt{Q}} \right)^k
\end{bmatrix}.
\]

Therefore, we have that

\[
\| B(\mu)^k \|_\infty \leq (1 + \frac{k}{\sqrt{Q}+1}) \left( \frac{\sqrt{Q}-1}{\sqrt{Q}} \right)^k + k \max \left\{ \frac{1}{Q}, \left( \frac{\sqrt{Q}-1}{\sqrt{Q}+1} \right)^2 \right\} \left( \frac{\sqrt{Q}-1}{\sqrt{Q}} \right)^{k-1}. \tag{29}
\]

Therefore for large \( k \)

\[
\| A^k \|_2 \leq d \| B(\mu)^k \|_\infty = \left( \frac{\sqrt{Q}-1}{\sqrt{Q}} + \epsilon_k \right)^{2k},
\]

where \( \epsilon_k \sim (\sqrt{Q} - 1) \). Also observe that

\[
\frac{\alpha^2((1 + \beta)^2 + 1)}{1 - \rho(\alpha, \beta)^2} = \frac{1}{L^2} \left[ \frac{\alpha^{2\sqrt{Q}}}{1 - \left( \frac{\sqrt{Q}-1}{\sqrt{Q}} \right)^2} \right]^2 = \frac{1}{L^2} \frac{5Q^2 + 2Q^{3/2} + Q}{(\sqrt{Q} + 1)^2(2\sqrt{Q} - 1)}.
\]

Since \( f \) is \( L \)-smooth,

\[
f(y_{k+1}) - f^* \leq \frac{L}{2} \| y_{k+1} - x^* \|^2.
\]

Thus, by Theorem 1 we have

\[
\mathbb{E}[f(y_{k+1})] - f^* \leq \frac{L}{2} \left( \frac{\sqrt{Q}-1}{\sqrt{Q}} + \epsilon_k \right)^{2k} \| x_0 - x^* \|^2 + C \epsilon_k \frac{5Q^2 + 2Q^{3/2} + Q}{2L(2\sqrt{Q} - 1)(\sqrt{Q} + 1)^2} \sigma^2,
\]

which completes the proof of Corollary 1.1.

To prove Theorem 2, first observe that when \( \beta = 0 \), the recursion simplifies significantly. Specifically, then \( y_{k+1} = x_k \), \( v_k = -\alpha g_k \), and we have (using similar notation as in the proof of Theorem 1)

\[
r_{k+1} = (I - \alpha H_k)r_k - \alpha \zeta_k
\]

\[
= \prod_{j=1}^{k} (I - \alpha H_j)r_1 - \alpha \zeta_k - \alpha \sum_{j=1}^{k-1} \prod_{l=j+1}^{k} (I - \alpha H_l)\zeta_l,
\]

where

\[
H_j = \int_0^1 \nabla f^2(x^* - t r_j) dt.
\]

Of course, since \( f \) is \( L \)-smooth and \( \mu \)-strongly convex, all eigenvalues of \( H_j \) lie in the interval \([\mu, L]\) for all \( j \geq 0 \).

Now, taking the squared norm on both sides, and recalling that the random vectors \( \zeta_k \) have zero mean and are mutually
independent, we have
\[
\mathbb{E} \|y_{k+1} - x^*\|_2^2 = \mathbb{E} \|r_{k+1}\|_2^2
\]
\[
= \mathbb{E}_{\zeta_1, \ldots, \zeta_k} \left[ \prod_{j=1}^k (I - \alpha H_j) r_1 - \alpha \zeta_k - \alpha \sum_{j=1}^{k-1} \prod_{l=j+1}^k (I - \alpha H_l) \zeta_j \right]^2
\]
\[
= \mathbb{E}_{\zeta_1} \cdots \mathbb{E}_{\zeta_k} \left[ \prod_{j=1}^k (I - \alpha H_j) r_1 - \alpha \zeta_k - \alpha \sum_{j=1}^{k-1} \prod_{l=j+1}^k (I - \alpha H_l) \zeta_j \right]^2 \cdots
\]
\[
= \left( \prod_{j=1}^k \|I - \alpha H_j\|_2 \right) \|x_1 - x^*\|_2^2 + \alpha^2 \mathbb{E}_{\zeta_k} \|\zeta_k\|_2^2 + \alpha^2 \sum_{j=1}^{k-1} \mathbb{E}_{\zeta_j} \left( \prod_{l=j+1}^k \|I - \alpha H_l\|_2 \right) \zeta_j^2
\]
\[
\leq \left( \prod_{j=1}^k \|I - \alpha H_j\|_2 \right) \|x_1 - x^*\|_2^2 + \alpha^2 \sigma^2 + \alpha^2 \sum_{j=1}^{k-1} \left( \prod_{l=j+1}^k \|I - \alpha H_l\|_2 \right)^2.
\]

Now, since \( I - \alpha H_j \) is symmetric, we have \( \| (I - \alpha H_j) \|_2^2 = \rho(I - \alpha H_j)^2 \), where \( \rho(I - \alpha H_j) \) denotes the spectral radius of \( I - \alpha H_j \) (the largest magnitude of an eigenvalue of \( I - \alpha H_j \)). For \( \alpha = \frac{2}{\mu + L} \), and since the eigenvalues of \( H_j \) lie in the interval \([\mu, L]\), it is straightforward to show that \( \rho(I - \alpha H_j) = \frac{2}{Q+1} \).

Therefore we have
\[
\mathbb{E} \left[ \|y_{k+1} - x^*\|_2^2 \right] \leq \left( \frac{Q - 1}{Q + 1} \right)^{2k} \|x_0 - x^*\|_2^2 + \alpha^2 \sum_{j=1}^{k} \left( \frac{Q - 1}{Q + 1} \right)^{2(k-j)}
\]
\[
\leq \left( \frac{Q - 1}{Q + 1} \right)^{2k} \|x_0 - x^*\|_2^2 + \frac{\alpha^2 \sigma^2}{1 - \left( \frac{Q - 1}{Q + 1} \right)^2}
\]
\[
= \left( \frac{Q - 1}{Q + 1} \right)^{2k} \|x_0 - x^*\|_2^2 + \frac{Q}{2L} \sigma^2,
\]

which completes the proof of Theorem 2.

### C. Permutation Matrix Construction

For a vector \( x \in \mathbb{R}^d \), let \( \text{diag}(x) \) denote a \( d \times d \) diagonal matrix with its \( i \)th diagonal entry equal to \( x_i \). Let \( a, b, c, d \in \mathbb{R}^d \) and suppose \( M \in \mathbb{R}^{2d \times 2d} \) is the matrix
\[
M = \begin{bmatrix}
\text{diag}(a) & \text{diag}(b) \\
\text{diag}(c) & \text{diag}(d)
\end{bmatrix}.
\]

Let \( P \in \{0, 1\}^{2d \times 2d} \) be the permutation matrix with entries \( P_{i,j} \) for \( i, j = 1, \ldots, 2d \) given by
\[
P_{i,j} = \begin{cases}
1 & \text{if } i \text{ is odd and } j = (i - 1)/2 + 1 \\
1 & \text{if } i \text{ is even and } j = d + \lfloor \frac{i-1}{2} \rfloor + 1 \\
0 & \text{otherwise}.
\end{cases}
\]

Then one can verify that
\[
PMP^T = \begin{bmatrix}
T_1 & 0 & \cdots & 0 \\
0 & T_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & T_d
\end{bmatrix}
\]
where, for \( j = 1, \ldots, d \), \( T_j \) is the \( 2 \times 2 \) matrix
\[
T_j = \begin{bmatrix}
a_j & b_j \\
c_j & d_j
\end{bmatrix}.
\]
D. Proof of Lemma 1

Recall that \( \alpha = 1/L \) and \( \beta = \frac{\sqrt{Q}-1}{\sqrt{Q}+1} \). For matrices of the form

\[
T_k = B(L)B(\mu)^{k_1}B(L)B(\mu)^{k_2} \cdots B(L)B(\mu)^{k_s}B(L),
\]

where

\[
B(\lambda) = \begin{bmatrix}
1 - \alpha(1 + \beta) & \beta \\
-\alpha \lambda & \beta
\end{bmatrix},
\]

we would like to show that the spectral radius \( \rho(T_k) \) is equal to

\[
\rho(T_k) = \left( \frac{\sqrt{Q} - 1}{\sqrt{Q}} \right)^k \times k_1k_2 \cdots k_s.
\]

To see this, first note that the Jordan form of \( B(\mu) \) is given by \( VJV^{-1} \), where

\[
V = \begin{bmatrix}
\frac{\sqrt{Q}(\sqrt{Q} - 1)}{\sqrt{Q} + 1} & Q \\
0 & 0
\end{bmatrix}
\]

and

\[
J = \begin{bmatrix}
\frac{\sqrt{Q} - 1}{\sqrt{Q}} & 1 \\
0 & \frac{\sqrt{Q} - 1}{\sqrt{Q}}
\end{bmatrix}.
\]

Using the Jordan form, we determine that \( B(\mu)^{k_\ell} \) is

\[
B(\mu)^{k_\ell} = \begin{bmatrix}
\left( 1 + \frac{k_\ell}{\sqrt{Q}+1} \right)^k & \left( \frac{\sqrt{Q} - 1}{\sqrt{Q}} \right)^k \left( \frac{\sqrt{Q} - 1}{\sqrt{Q}} \right)^{k_\ell-1} \\
-\frac{k_\ell}{Q} \left( \frac{\sqrt{Q} - 1}{\sqrt{Q}} \right)^{k_\ell-1} & \left( 1 - \frac{k_\ell}{\sqrt{Q}+1} \right) \left( \frac{\sqrt{Q} - 1}{\sqrt{Q}} \right)^k
\end{bmatrix}.
\]

Through direct matrix multiplication

\[
B(L)B(\mu)^{k_\ell}B(L) = - \left( \frac{\sqrt{Q} - 1}{\sqrt{Q}} \right)^{k_\ell+1} k_\ell B(L).
\]

Therefore,

\[
T_j = (-1)^{s-1} \left( \frac{\sqrt{Q} - 1}{\sqrt{Q}} \right)^{k_1 \cdots k_s} k_1k_2 \cdots k_{s-1} B(L)B(\mu)^{k_s}.
\]

Finally, the spectral-radius of \( B(L)B(\mu)^{k_s} \) is

\[
\rho \left( B(L)B(\mu)^{k_s} \right) = \left( \frac{\sqrt{Q} - 1}{\sqrt{Q}} \right)^{k_s+1} k_s,
\]

and hence

\[
\rho(T_k) = \left( \frac{\sqrt{Q} - 1}{\sqrt{Q}} \right)^k k_1k_2 \cdots k_s.
\]

E. Proof of Theorem 4

Since the functions \( f_i \) are assumed to be twice continuously differentiable, by (10) we can express the mini-batch gradients as

\[
g_k = \tilde{H}_k r_k + z_k, \tag{30}
\]

where

\[
\tilde{H}_k = \sum_{i=1}^n v_{k,i} \int_0^1 \nabla^2 f_i(x^* + tr_k)dt.
\]
On the Convergence of Nesterov’s Accelerated Gradient Method

and

\[ z_k = \sum_{i=1}^{n} v_{k,i} \nabla f_i(x^*). \]

By convexity of norms,

\[ \|z_k\| \leq \sum_{i=1}^{n} v_{k,i} \|\nabla f_k(x^*)\|. \]

Hence, taking expectations gives

\[ \mathbb{E}_k[\|z_k\|] \leq \frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(x^*)\| = \sigma. \]

Using (30) in (9) and unrolling, we obtain

\[
\begin{bmatrix}
  r_{k+1} \\
  v_k
\end{bmatrix} = A_k \cdots A_1 \begin{bmatrix} r_1 \\ v_0 \end{bmatrix} - \alpha \begin{bmatrix} (1 + \beta)I \\ I \end{bmatrix} z_k - \alpha \sum_{j=1}^{k-1} (A_k \cdots A_{j+1}) \begin{bmatrix} (1 + \beta)I \\ I \end{bmatrix} z_j,
\]

(31)

where

\[ A_k = \begin{bmatrix} I - \alpha(1 + \beta)\tilde{H}_k & \beta^2I \\ -\alpha \tilde{H}_k & \beta I \end{bmatrix}. \]

By submultiplicativity of matrix norms, \( \|A_k \cdots A_{j+1}\| \leq \prod_{l=j+1}^{k} \|A_l\|. \) Thus we turn our attention to bounding the spectral norm of \( A_k. \)

**Lemma 3.**

\[ \|A_k\| \leq \max_{\lambda \in [\mu, L]} \|B(\lambda)\| = R(\alpha, \beta). \]

**Proof.** For all \( k \geq 0, \) every eigenvalue of \( \tilde{H}_k \) lies in the interval \([\mu, L],\) based on the assumption that each function \( f_i \) is \( L \)-smooth and \( \mu \)-strongly convex. It follows from Polyak (1964, Lemma 5) that there exists an eigenvalue \( \lambda \) of \( \tilde{H}_k \) such that \( \|A_k\| \) is equal to the spectral norm of

\[ B(\lambda) = \begin{bmatrix} 1 - \alpha(1 + \beta)\lambda & \beta^2 \\ -\alpha \lambda & \beta \end{bmatrix}. \]

We next compute \( \|B(\lambda)\|, \) which is equal to the square root of the largest eigenvalue of

\[ B(\lambda)^\top B(\lambda) = \begin{bmatrix} (1 - \alpha(1 + \beta)\lambda)^2 + \alpha^2\lambda^2 & \beta^2(1 - \alpha(1 + \beta)\lambda) - \alpha\beta\lambda \\ \beta^2(1 - \alpha(1 + \beta)\lambda) - \alpha\beta\lambda & \beta^2(\beta^2 + 1) \end{bmatrix}. \]

The characteristic polynomial of \( B(\lambda)^\top B(\lambda) \) is

\[ \xi^2 - C_\lambda(\alpha, \beta)\xi + \beta^2(1 - \alpha\lambda)^2 = 0, \]

where

\[ C_\lambda(\alpha, \beta) = (1 - \alpha(1 + \beta)\lambda)^2 + \alpha^2\lambda^2 + \beta^2(\beta^2 + 1). \]

The largest root of the characteristic polynomial is equal to

\[ R_\lambda(\alpha, \beta)^2 = \frac{1}{2} \left( C_\lambda(\alpha, \beta) + \sqrt{C_\lambda(\alpha, \beta)^2 - 4\beta^2(1 - \alpha\lambda)^2} \right) \]

which is equal to \( \|B(\lambda)\|^2. \) Therefore

\[ \|A_k\| \leq \max_{\lambda \in [\mu, L]} R_\lambda(\alpha, \beta). \]

\[ \square \]
Assume that $\alpha$ and $\beta$ have been chosen so that $R(\alpha, \beta) < 1$. Then for all $k$ and $j + 1$, $\|A_k \cdots A_{j+1}\| \leq \prod_{l=j+1}^{k} \|A_l\| \leq R(\alpha, \beta)^{k-j}$.

Taking the norm on both sides of (31) and using the triangle inequality, we have

$$\left\| \begin{bmatrix} r_{k+1} \\ v_k \end{bmatrix} \right\| \leq R(\alpha, \beta)^k \left\| \begin{bmatrix} r_1 \\ v_0 \end{bmatrix} \right\| + \alpha \sqrt{1 + \beta^2} + 1 \sum_{j=1}^{k} R(\alpha, \beta)^{k-j} \|z_k\|. \tag{32}$$

Taking the expectation gives

$$E_k \|y_{k+1} - x^*\| \leq E_k \left\| \begin{bmatrix} r_{k+1} \\ v_k \end{bmatrix} \right\| \leq R(\alpha, \beta)^k \|x_0 - x^*\| + \frac{\alpha \sqrt{1 + \beta^2} + 1}{1 - R(\alpha, \beta)^2} \sigma. \tag{33}$$

Taking the expectation gives

$$E_k \|y_{k+1} - x^*\| \leq R(\alpha, \beta)^k \|x_0 - x^*\| + \frac{\alpha \sqrt{1 + \beta^2} + 1}{1 - R(\alpha, \beta)^2} \sigma. \tag{34}$$

**F. Proof of Corollaries 4.2 and 4.1**

When $\beta = 0$, we have $y_{k+1} = x_k$ and $v_k = -\alpha g_k$ for all $k$. In this case we have

$$r_{k+1} = r_k - \alpha g_k.$$ 

Since the objectives $f_i$ are twice continuously differentiable, the mini-batch gradients can again be written as (using the same notation as in the proof of Theorem 4)

$$g_k = H_k r_k + z_k.$$ 

Thus, with $A_k = I - \alpha H_k$, we have

$$r_{k+1} = A_k r_k - \alpha z_k = A_k \cdots A_1 r - \alpha z_k - \alpha \sum_{j=1}^{k-1} (A_k \cdots A_{j+1}) z_k.$$ 

Since $H_k$ is symmetric, it follows that $A_k$ is also symmetric, and so $\|A_k\|$ is equal to the largest magnitude of any eigenvalue of $A_k$. Recall that all eigenvalues of $H_k$ lie in the interval $[\mu, L]$. Therefore, $\|A_k\| \leq \max_{\lambda \in [\mu, L]} |1 - \alpha \lambda| = \max \{|1 - \alpha \mu|, |1 - \alpha L|\}$. Choosing $\alpha < \frac{1}{2}$ and taking the norm and expectation thus yields that

$$E_k \|x_k - x^*\| = E_k \|r_{k+1}\| \leq \left|1 - \alpha \lambda\right|^k \|x_0 - x^*\| + \frac{\alpha}{1 - \alpha \lambda} \sigma, \tag{35}$$

where $\lambda := \text{argmax}_{\lambda \in [\mu, L]} |1 - \alpha \lambda|$. When $\alpha = \frac{2}{\mu + L}$, we have that $\max_{\lambda \in [\mu, L]} |1 - \alpha \lambda| = \frac{Q - 1}{Q + 1}$, and equation (35) simplifies as

$$E_k \|x_k - x^*\| = E_k \|r_{k+1}\| \leq \left(\frac{Q - 1}{Q + 1}\right)^k \|x_0 - x^*\| + \frac{1}{\mu} \sigma.$$ 

**G. Additional Experiments**

**G.1. Least Squares**

To provide additional experiments illustrating the relationship between empirical observations and the theory developed in Section 3 for the stochastic approximation setting, we conduct additional experiments on randomly-generated least-squares problems. We generate the least-squares problem using the approach described in (Lenard & Minkoff, 1984). Visualizations are shown in Figure G.1.
Figure G.1. Visualizing the accuracy with which the theory predicts the coefficient of the variance term and the convergence rate for different choices of constant step-size and momentum parameters, and various objective condition numbers $Q$. Plots labeled “Theoretical” depict theoretical results from Theorem 1. Plots labeled “Empirical” depict empirical results when using the ASG method to solve a least-squares regression problem with additive Gaussian noise; each pixel corresponds to an independent run of the ASG method for a specific choice of constant step-size and momentum parameters. In all figures, the area enclosed by the red contour depicts the theoretical stability region from Theorem 1 depicted by the red-contour are tight.

We run the ASG method on least-squares regression problems with various condition numbers $Q$. The objectives $f$ correspond to randomly generated least squares problems, consisting of 500 data samples with 10 features each. Stochastic gradients are sampled by adding zero-mean Gaussian noise, with standard-deviation $\sigma = 0.25$, to the true gradient. The left plots in each sub-figure depict theoretical predictions from Theorem 1, while the right plots in each sub-figure depict empirical results. Each pixel corresponds to an independent run of the ASG method for a specific choice of constant step-size and momentum parameters. In all figures, the area enclosed by the red contour depicts the theoretical stability region from Theorem 1 for which $\rho(\alpha, \beta) < 1$.

Figures G.1a/G.1c/G.1e show the coefficient multiplying the variance term, which is taken to be $\frac{\sigma^2((1+\beta)^2+1)}{1-\rho(\alpha, \beta)^2}$ in theory. Brighter regions correspond to smaller coefficients, while darker regions correspond to larger coefficients. All sets of figures (theoretical and empirical) use the same color scale. We can see that the coefficient of the variance term in Theorem 1 provides a good characterization of the magnitude of the neighborhood of convergence. The constant $C_\epsilon$ is approximated as $1 + (1 - \rho(\alpha, \beta)^2)(\rho(\alpha, \beta)^2 - \rho(\alpha, \beta)^2)$, where $\rho(\alpha, \beta)$ is defined as the largest singular value of $A$ in (15), and $\rho(\alpha, \beta)$ is the largest eigenvalue of $A$.

Figures G.1b/G.1d/G.1f show the linear convergence rate in theory and in practice. Brighter regions correspond to
We randomly generate multi-class classification problems consisting of 5 classes and 100 data samples with 10 features each, only 5 of which are discriminative. We create one data-cluster per class, and vary the cluster separation and regularization parameter to vary the condition number $Q$. For reporting purposes, we estimate the condition number $Q$ during training by evaluating the eigenvalues of the Hessian at each iteration. The smoothness constant $L$ is taken to be the maximum eigenvalue seen during training, and the modulus of strong-convexity $\mu$ is taken to be the minimum eigenvalue seen during training. The faster convergence rates (brighter regions) correspond to $\beta > 0$, indicating that the ASG method provides acceleration over SGD in this stochastic approximation setting. Moreover, for a given step-size, the contrast between the brighter regions ($\beta > 0$) and darker regions ($\beta = 0$) increases as the condition number grows, supporting theoretical findings that the convergence rate of the ASG method exhibits a better dependence on the condition number than SGD.

faster rates, and darker regions correspond to slower rates. Again, all figures (theoretical and empirical) use the same color scale. We can see that the theoretical linear convergence rates in Theorem 1 provide a good characterization of the empirical convergence rates. Moreover, the theoretical conditions for convergence in Theorem 1 depicted by the red-contour appear to be tight.

**G.2. Multinomial Logistic Regression**

Next we conduct experiments on $\ell_2$ regularized multinomial logistic regression problems with additive Gaussian noise, to examine whether the ASG method still achieves acceleration over SGD for these problems in the stochastic approximation setting, as is predicted by the theory in Section 3. These problems are smooth and strongly-convex, but non-quadratic. Tight estimates of the smoothness constant $L$ and the modulus of strong-convexity $\mu$ cannot be computed definitively since the eigenvalues of the Hessian vary throughout the parameter space.

We randomly generate multi-class classification problems consisting of 5 classes and 100 data samples with 10 features each, only five of which are discriminative. We create one data cluster per class, and vary the cluster separation and regularization parameter to vary the condition number $Q$. For reporting purposes, we estimate the condition number $Q$ during training by evaluating the eigenvalues of the Hessian at each iteration. The smoothness constant $L$ is taken to be the maximum eigenvalue seen during training, and the modulus of strong-convexity $\mu$ is taken to be the minimum eigenvalue seen during training. We use the `make_classification()` function in scikit-learn (Pedregosa et al., 2011) to generate random classification problem instances.

Visualizations are provided in Figure G.2. Each pixel corresponds to an independent run of the ASG method for a specific choice of constant step-size and momentum parameters. Pixel intensities denote the linear convergence rates observed in practice. Brighter regions correspond to faster rates, and darker regions correspond to slower rates.

The parameter setting $\beta$ equals 0 corresponds to SGD, and the parameter setting $\beta > 0$ corresponds to the ASG method. The faster convergence rates (brighter regions) correspond to $\beta > 0$, indicating that the ASG method provides acceleration over SGD in this stochastic approximation setting. Moreover, for a given step-size, the contrast between the brighter regions ($\beta > 0$) and darker regions ($\beta = 0$) increases as the condition number grows, supporting theoretical findings that the convergence rate of the ASG method exhibits a better dependence on the condition number than SGD.