Turk. J. Math. Comput. Sci.
15(1)(2023) 89–95
© MatDer
DOI : 10.47000/tjmcs.1037384

On the Convergence of Stochastic Aggregated Gradient Method

Figen Öztoprak

Department of Industrial Engineering, Faculty of Engineering, Gebze Technical University, 41400, Kocaeli, Turkey.

Received: 16-12-2021 • Accepted: 26-01-2023

Abstract. The minimization problem of the sum of a large set of convex functions arises in various applications. Methods such as incremental gradient, stochastic gradient, and aggregated gradient are popular choices for solving those problems as they do not require a full gradient evaluation at every iteration. In this paper, we analyze a generalization of the stochastic aggregated gradient method via an alternative technique based on the convergence of iterative linear systems. The technique provides a short proof for the $O(\kappa^{-1})$ linear convergence rate in the quadratic case. We observe that the technique is rather restrictive for the general case, and can provide weaker results.

2010 AMS Classification: 90C25

Keywords: Unconstrained optimization, stochastic gradient, incremental methods.

1. Introduction

We consider unconstrained minimization of the sum of a finite number of smooth functions $f_i : \mathbb{R}^n \to \mathbb{R}$; i.e.

$$\min_{x\in\mathbb{R}^n} F(x), \quad \text{with} \quad F(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x), \tag{1.1}$$

such that the number of component functions N is large as compared to the dimension n of the variable vector x. This problem structure arises in variety of applications, including machine learning. In the setting of parameter inference, for instance, $f_i(.)$ has the form $f_i(x) = l(x; a_i, b_i)$, where $l(.)$ is a loss function stating the misfit of a model parametrized by x for a given data point (a_i, b_i), and N is the number of data points [2].

Among popular methods for solving this problem are inexact gradient-type methods, which avoid computing the full gradient $\nabla F(x) = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(x)$ since N is large. The general form of the k^{th} step for those methods is

$$x_{k+1} = x_k - \alpha_k y_k, \qquad \alpha_k > 0.$$

Thus, the conventional gradient descent algorithm is modified by replacing $\nabla F(x_k)$ with y_k.

- The *incremental gradient* method [1] chooses one index $i_k \in \{1, ..., N\}$ and sets $y_k = \nabla f_{i_k}(x_k)$. The choice of i_k could be at random, or it can make a pass through all indices $\{1, ..., N\}$ in order. Note that the method is equivalent to the stochastic gradient algorithm if the choice of i_k is random with probability $\frac{1}{N}$ for all component functions, as the objective of (1.1) in this case can be seen as an expected value statement.

Email address: figen.topkaya@gtu.edu.tr (F. Öztoprak)

- The *aggregated gradient* method computes at each iteration the gradient of only one component function as in the incremental approach. However, it reuses the previously computed gradients in computing the search direction. That is, it sets $y_k = \frac{1}{N}\sum_{i=1}^N \nabla f_i(x_{k(i)})$. Here, $k(i)$ is the largest iteration number where the function $\nabla f_i(.)$ was evaluated. In the *incremental aggregated gradient* (IAG) variant of the method [3], the component gradient to be updated at each iteration is selected in a cyclic manner whereas in the stochastic variant that we explain in the next item it is selected at random.
- The recently proposed *stochastic aggregated gradient* (SAG) method [5] follows the ideas of the aggregated gradient method, but chooses the function index for gradient update at random; i.e.

$$y_k = \frac{1}{N}(\nabla f_j(x_k) - \nabla f_j(x_{k(j)}) + \sum_{i=1}^{N} \nabla f_i(x_{k(i)})),$$

  where $j \in \{1, 2, \cdots, N\}$ is chosen at random.

In this paper, we study a generalization of the SAG algorithm that might update multiple component functions at each iteration (Section 2). We present convergence analysis of the method in Section 3; our analysis follows a different technique as compared to [5] and [3].

Notation. We denote unspecified eigenvalues and the spectral radius of a square matrix M with λ and $\rho(M)$, respectively. $||.||$ indicates l_2-norm unless stated otherwise.

2. A Generalization of the Stochastic Aggregated Gradient Algorithm

We consider a generalization of the SAG algorithm such that each gradient component is updated with probability η at each iteration. So, we set

$$x_{k+1} = x_k - \alpha y_k, \quad k = 0, 1, \dots \tag{2.1}$$

and

$$y_k = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(x_{k(i)}), \tag{2.2}$$

where

$$k(i) = \begin{cases} k & \text{with probability } \eta, \\ [k-1](i) & \text{with probability } 1-\eta. \end{cases}$$

We do not specify the choice of the constant steplength parameter $\alpha > 0$ yet. We set α to a sufficiently small constant positive value in our algorithm. More details on the value of α will be provided in our analysis.

Define error terms

$$e_k^i = \nabla f_i(x_{k(i)}) - \nabla f_i(x_k) \qquad \text{and} \qquad e_k = \frac{1}{N}\sum_{i=1}^{N} e_k^i$$

so that we can state

$$y_k = \frac{1}{N}\sum_{i=1}^{N}(\nabla f_i(x_k) + e_k^i) = \nabla F(x_k) + e_k.$$

Note that given e_{k-1} and x_{k-1}, we have the conditional expectation

$$E[e_k^i] = (1-\eta)(e_{k-1}^i + \nabla f_i(x_{k-1}) - \nabla f_i(x_k)).$$

Assumptions.

A.1. $F(.)$ is twice continuously differentiable.

A.2. $F(.)$ is strongly convex, and x_* is its unique minimizer.

Let us define two average Hessian matrices, H_k and $\bar{H}_k$, for each k, such that the mean value theorem applies as follows.

$$\nabla F(x_k) = \nabla F(x_{k-1}) + H_k(x_k - x_{k-1}),$$
$$\nabla F(x_k) = \bar{H}_k(x_k - x_*).$$

A.3. For some $\mu > 0$ and $L > 1$,

$$\mu I \preceq H_k \preceq LI, \quad \text{and} \quad \mu I \preceq \bar{H}_k \preceq LI, \quad \text{for all } k.$$

A.4. The choice of the component functions are independent identically distributed random events.

Lemma 2.1. *Suppose that assumptions (A.1)-(A.4) hold. The conditional expectation of the error in gradient estimation $E[e_k]$ and the distance to the solution $x_k - x_*$ evolves with respect to the system*

$$\begin{pmatrix} \frac{1}{L}E[e_k] \\ x_k - x_* \end{pmatrix} = M_k \begin{pmatrix} \frac{1}{L}e_{k-1} \\ x_{k-1} - x_* \end{pmatrix}, \quad \textit{for} \quad M_k = \begin{pmatrix} (1-\eta)(I+\alpha H_k) & \frac{1}{L}(1-\eta)\alpha H_k\bar{H}_{k-1} \\ -\alpha LI & I - \alpha\bar{H}_{k-1} \end{pmatrix}. \tag{2.3}$$

Proof. We will state two relations on the change of e_k and on the change of $x_k - x_*$, respectively, and then merge the two.

$$\begin{aligned} E[e_k] &= (1-\eta)(e_{k-1} + \nabla F(x_{k-1}) - \nabla F(x_k)) \\ &= (1-\eta)(e_{k-1} - H_k(x_k - x_{k-1})) \\ &= (1-\eta)(e_{k-1} - H_k(-\alpha y_{k-1})) \\ &= (1-\eta)\left(e_{k-1} + \alpha H_k(\nabla F(x_{k-1}) + e_{k-1})\right) \\ &= (1-\eta)\left((I+\alpha H_k)e_{k-1} + \alpha H_k \nabla F(x_{k-1})\right) \\ &= (1-\eta)(I+\alpha H_k)e_{k-1} + (1-\eta)\alpha H_k\bar{H}_{k-1}(x_{k-1} - x_*). \end{aligned}$$

We also have

$$\begin{aligned} x_k - x_* &= x_{k-1} - x_* + (x_k - x_{k-1}) \\ &= x_{k-1} - x_* - \alpha y_{k-1} \\ &= x_{k-1} - x_* - \alpha(\nabla f_{k-1} + e_{k-1}) \\ &= x_{k-1} - x_* - \alpha\bar{H}_{k-1}(x_{k-1} - x_*) - \alpha e_{k-1} \\ &= (I - \alpha\bar{H}_{k-1})(x_{k-1} - x_*) - \alpha e_{k-1}. \end{aligned}$$

Merging the two equations yields the desired system. □

Lemma 2.1 states the progress of the algorithm as an *iterative linear system*, and suggests that the product of $(2nx2n)$ nonsymmetrical matrices M_k determine the convergence behavior of our algorithm. In particular, since the random processes at each iteration are independent, if

$$\prod_{k=1}^{\infty} M_k \to 0$$

holds, then we get $x_k \to x_*$.

Note that when F is a quadratic function, the matrix M_k does not depend on the iteration; i.e. since $\bar{H}_k = H_k = H$ for a constant positive definite H for all k, we have $M_k = M$ for all k where

$$M = \begin{pmatrix} (1-\eta)(I+\alpha H) & \frac{1}{L}(1-\eta)\alpha HH \\ -\alpha LI & I - \alpha H \end{pmatrix}.$$

It is also possible to state the progress of the algorithm via a constant matrix when F is non-quadratic, as we explain in the following lemma.

Lemma 2.2. *If assumptions (A.1)-(A.4) hold, then*

$$\begin{pmatrix} \frac{1}{L}\|E[e_k]\| \\ \|x_k - x_*\| \end{pmatrix} \le \bar{M} \begin{pmatrix} \frac{1}{L}\|e_{k-1}\| \\ \|x_{k-1} - x_*\| \end{pmatrix} \quad \textit{for} \quad \bar{M} = \begin{pmatrix} (1-\eta)(1+\alpha L) & (1-\eta)\alpha L \\ \alpha L & 1-\alpha\mu \end{pmatrix}.$$

Proof. Recall that we have

$$E[e_k] = (1-\eta)(I + \alpha H_k)e_{k-1} + (1-\eta)\alpha H_k \bar{H}_{k-1}(x_{k-1} - x_*),$$

and

$$x_k - x_* = (I - \alpha\bar{H}_{k-1})(x_{k-1} - x_*) - \alpha e_{k-1}.$$

That implies

$$\|E[e_k]\| \le (1-\eta)(1+\alpha L)\|e_{k-1}\| + (1-\eta)\alpha L^2\|x_{k-1} - x_*\|,$$

and

$$\|x_k - x_*\| \le (1-\alpha\mu)\|x_{k-1} - x_*\| + \alpha\|e_{k-1}\|,$$

respectively. □

In the next section, we provide convergence results for the extended SAG algorithm based on the systems given in Lemma 2.1 and Lemma 2.2.

3. Convergence Results

In this section, we will first discuss the properties of M_k and $\bar{M}$. Then, we will provide convergence results for the generalized SAG algorithm.

Theorem 3.1. *Define $\kappa = L/\mu$. If assumption (A.3) holds, then the matrix M_k defined in (2.3) satisfies the following provided that $\alpha = \frac{\theta\eta}{L}$ for $\theta \in (0, 1]$.*

(a) *All eigenvalues λ of M_k satisfy $\lambda \ge 0$.*
(b) *The largest eigenvalue λ_1 of M_k satisfies $\lambda_1 \le 1$.*
(c) $\rho(M_k) \le 1 - \eta^2\theta\kappa^{-1}$.

Proof. Let us first note that we have all eigenvalues of $H_k, \bar{H}_k$ in $[\mu, L]$, and $\kappa \ge 1$ by assumption (A.3).

(a) Since blocks (2,1) and (2,2) of M_k commute, by Theorem 1 of [6] we have

$$\begin{aligned} \det(M_k - \lambda I) &= \det\Big([(1-\eta)(I+\alpha H_k) - \lambda I]\Big[I - \alpha\bar{H}_{k-1} - \lambda I\Big] + (1-\eta)\alpha^2 H_k\bar{H}_{k-1}\Big) \\ &= \det\Big((1-\eta)(I - \alpha^2 H_k\bar{H}_{k-1} + \alpha H_k - \alpha\bar{H}_{k-1}) + \lambda^2 I - \lambda\Big[(1-\eta)(I+\alpha H_k) + I - \alpha\bar{H}_{k-1}\Big] \\ &\quad + (1-\eta)\alpha^2 H_k\bar{H}_{k-1}\Big) \\ &= \det\Big((1-\eta)(I + \alpha H_k - \alpha\bar{H}_{k-1}) - \lambda\Big[(1-\eta)(I+\alpha H_k) + I - \alpha\bar{H}_{k-1}\Big] + \lambda^2 I\Big). \end{aligned}$$

Define

$$C(\lambda) = (1-\eta)(I + \alpha H_k - \alpha\bar{H}_{k-1}) - \lambda\Big[(1-\eta)(I+\alpha H_k) + I - \alpha\bar{H}_{k-1}\Big].$$

Then, for each eigenvalue λ of M_k, there exists an eigenvalue σ of $C(\lambda)$ such that $\sigma = -\lambda^2$. To derive a contradiction, suppose $\lambda < 0$. Then, by using the fact that $\alpha L \le 1$ we have

$$\begin{aligned} \sigma &\ge (1-\eta)(1+\alpha\mu - \alpha L) - \lambda\left[(1-\eta)(1+\alpha\mu) + 1 - \alpha L\right] \\ &\ge (1-\eta)\alpha\mu - \lambda(1-\eta)(1+\alpha\mu) > 0. \end{aligned}$$

This cannot be true since $\sigma = -\lambda^2$. Thus, $\lambda \ge 0$.

(b) Suppose for contradiction that $\lambda_1 > 1$. Note that, we can state $C(\lambda)$ as

$$\begin{aligned} C(\lambda) &= (1-\eta)(I + \alpha(H_k - \bar{H}_{k-1})) - \lambda\left[(1-\eta)(I + \alpha H_k) + I - \alpha\bar{H}_{k-1} - (1-\eta)\alpha\bar{H}_{k-1} + (1-\eta)\alpha\bar{H}_{k-1}\right] \\ &= (1-\eta)(I + \alpha(H_k - \bar{H}_{k-1})) - \lambda\left[(1-\eta)(I + \alpha(H_k - \bar{H}_{k-1})) + I - \alpha\bar{H}_{k-1} + (1-\eta)\alpha\bar{H}_{k-1}\right] \\ &= (1-\eta)(1-\lambda)(I + \alpha(H_k - \bar{H}_{k-1})) - \lambda I + \lambda\eta\alpha\bar{H}_{k-1}. \end{aligned}$$

Since $-\lambda_1^2$ is an eigenvalue of $C(\lambda_1)$, it will be the smallest one; i.e. $\sigma_n = -\lambda_1^2$. As we know from part (a) that $\lambda_1 \geq 0$, and by the contradiction hypothesis that $\lambda_1 > 1$, we have

$$\begin{aligned} -\lambda_1^2 &\geq (1-\eta)(1-\lambda_1)(1 + \alpha(L-\mu)) - \lambda_1 + \lambda_1\eta\alpha\mu \\ \Rightarrow &\lambda_1(1-\lambda_1) \geq (1-\eta)(1-\lambda_1)(1 + \alpha(L-\mu)) \\ \Rightarrow &\lambda_1 \leq (1-\eta)(1 + \alpha(L-\mu)) = (1-\eta)(1 + \theta\eta(1-\kappa^{-1})) \leq 1 - \eta^2 \\ \Rightarrow &\lambda_1 \leq 1 \end{aligned}$$

for $\alpha = \dfrac{\theta\eta}{L}$, $\theta \in (0, 1]$, since $\kappa^{-1} = \dfrac{\mu}{L} \leq 1$. The contradiction is established.

(c) Let us use the representation of $C(\lambda)$ derived in part (b), and re-state the inequality on the smallest eigenvalue of $C(\lambda_1)$ based on the fact that $0 \leq \lambda_1 \leq 1$ as proven in parts (a) and (b).

$$\begin{aligned} -\lambda_1^2 &\geq (1-\eta)(1-\lambda_1)(1 + \alpha(\mu - L)) - \lambda_1 + \lambda_1\eta\alpha\mu \\ &= (1-\eta)(1-\lambda_1)(1 + \theta\eta(\kappa^{-1} - 1)) - \lambda_1(1 - \eta\alpha\mu) \\ &\geq -\lambda_1(1 - \eta\alpha\mu) \end{aligned}$$

since the fact $0 \leq \kappa^{-1} \leq 1$ implies

$$0 \leq (1-\eta)(1-\lambda_1)(1-\theta\eta) \leq (1-\eta)(1-\lambda_1)(1 + \theta\eta(\kappa^{-1} - 1)) \leq (1-\eta)(1-\lambda_1).$$

Therefore, we get $\lambda_1 \leq 1 - \eta\alpha\mu = 1 - \eta^2\theta\kappa^{-1}$.

□

The next theorem provides a similar result for $\bar{M}$. Although it describes a homogeneous transition for the non-quadratic case, we observe that a smaller steplength value can be employed and the bound on the largest eigenvalue of this matrix is larger than M_k.

Theorem 3.2. *Consider the matrix $\bar{M}$ defined in Lemma 2.2 with $0 < \mu \leq L$ and $\eta \in (0, 1]$. If $\alpha = \dfrac{\theta\eta}{\kappa L}$ for $\theta \in (0, 0.5]$, then both eigenvalues of $\bar{M}$ are nonnegative and we have $\rho(\bar{M}) \leq 1 - \eta^2\theta\kappa^{-2}$.*

Proof. The eigenvalues of matrix $\bar{M}$ are the roots of the polynomial

$$\begin{aligned} p(\lambda) &= \left[(1-\eta)(1+\alpha L) - \lambda\right](1 - \alpha\mu - \lambda) - (1-\eta)\alpha^2 L^2 \\ &= \lambda^2 - \left[(1-\eta)(1+\alpha L) + 1 - \alpha\mu\right]\lambda + (1-\eta)(1+\alpha L)(1-\alpha\mu) - (1-\eta)\alpha^2 L^2. \end{aligned}$$

So,

$$\lambda_{1,2} = \frac{1}{2}\left((1-\eta)(1+\alpha L) + 1 - \alpha\mu \mp \sqrt{\Delta}\right),$$

where

$$\begin{aligned} \Delta &= \left[(1-\eta)(1+\alpha L) + 1 - \alpha\mu\right]^2 - 4\left[(1-\eta)(1+\alpha L)(1-\alpha\mu) - (1-\eta)\alpha^2 L^2\right] \\ &= \left[(1-\eta)(1+\alpha L) - (1-\alpha\mu)\right]^2 + 4(1-\eta)\alpha^2 L^2. \end{aligned}$$

Note that

$$\begin{aligned} (1-\eta)(1+\alpha L) - (1-\alpha\mu) &= -\eta + \alpha[(1-\eta)L + \mu] \\ &\leq -\eta + 2L\alpha \leq (-1 + \kappa^{-1})\eta \leq 0. \end{aligned}$$

Also, for

$$\delta = \left[(1-\eta)(1+\alpha L) - (1-\alpha\mu) - 2(1-\eta)\alpha\mu\right]^2 - \Delta$$

we have

$$\begin{aligned}\delta &= 4(1-\eta)\alpha\mu\left[(1-\eta)\alpha\mu - [(1-\eta)(1+\alpha L)-(1-\alpha\mu)]\right] - 4(1-\eta)\alpha^2 L^2\\ &= 4(1-\eta)\alpha\mu\left[(1-\eta)\alpha\mu - [(1-\eta)(1+\alpha L)-(1-\alpha\mu)] - \alpha\frac{L^2}{\mu}\right]\\ &= 4(1-\eta)\alpha\mu\left[\eta-\eta\alpha\mu-(1-\eta)\alpha L-\alpha\frac{L^2}{\mu}\right]\\ &= 4(1-\eta)\alpha\mu\eta\left(1+\theta\left[-\eta\frac{\mu^2}{L^2}-(1-\eta)\frac{\mu}{L}-1\right]\right)\\ &\geq 4(1-\eta)\alpha\mu\eta\left(1+\theta\left[-\eta-(1-\eta)-1\right]\right) = 4(1-\eta)\alpha\mu\eta(1-2\theta)\geq 0\end{aligned}$$

as $\theta \leq 0.5$. Therefore,

$$[(1-\eta)(1+\alpha L)-(1-\alpha\mu)]^2 \leq \Delta \leq [(1-\eta)(1+\alpha L)-(1-\alpha\mu)-2(1-\eta)\alpha\mu]^2.$$

For the smaller root λ_2 this implies

$$\begin{aligned}\lambda_2 &= \frac{1}{2}\left((1-\eta)(1+\alpha L)+1-\alpha\mu-\sqrt{\Delta}\right)\\ &\geq \frac{1}{2}\left((1-\eta)(1+\alpha L)+1-\alpha\mu+(1-\eta)(1+\alpha L)-(1-\alpha\mu)-2(1-\eta)\alpha\mu\right)\\ &= (1-\eta)(1+\alpha L-\alpha\mu)\geq 0.\end{aligned}$$

As for the larger root λ_1 we have

$$\begin{aligned}\lambda_1 &= \frac{1}{2}\left((1-\eta)(1+\alpha L)+1-\alpha\mu+\sqrt{\Delta}\right)\\ &\leq \frac{1}{2}\left((1-\eta)(1+\alpha L)+1-\alpha\mu-(1-\eta)(1+\alpha L)+(1-\alpha\mu)+2(1-\eta)\alpha\mu\right)\\ &= 1-\alpha\mu+(1-\eta)\alpha\mu = 1-\eta\alpha\mu.\end{aligned}$$

Placing the value of α we get $\lambda_1 \leq 1-\theta\eta^2\kappa^{-2}$.

□

We are now ready to give the convergence rate results. First, we study the quadratic case. Then, we observe convergence for the general case.

3.1. **Quadratic Case.** Recall that in the quadratic case the result of Lemma 2.1 reduces to

$$\begin{pmatrix}\frac{1}{L}E[e_k]\\ x_k-x_*\end{pmatrix} = M^k\begin{pmatrix}\frac{1}{L}e_0\\ x_0-x_*\end{pmatrix}, \quad \text{for} \quad M=\begin{pmatrix}(1-\eta)(I+\alpha H) & \frac{1}{L}(1-\eta)\alpha HH\\ -\alpha LI & I-\alpha H\end{pmatrix}. \tag{3.1}$$

The linear convergence of the generalized SAG algorithm in the quadratic case follows by the next result.

Corollary 3.3. *Suppose the sequence $\{x_k\}$ is produced by (2.1) with y_k defined as in (2.2), and $\alpha = \frac{\theta\eta}{L}$ with $\theta \in (0,1]$. If Assumptions (A.1)-(A.4) hold and F is a quadratic function, then $\|x_k - x_*\| = O(\rho^k)$ with $\rho \leq 1-\eta^2\theta\kappa^{-1}$, and $\kappa = \frac{L}{\mu}$.*

Proof. Considering the Jordan canonical form of M^k [8], we observe that the largest term defining the right hand side of (3.1) has norm $O(\rho^k)$. Therefore,

$$\|x_k-x_*\| \leq \left\|\begin{pmatrix}\frac{1}{L}E[e_k]\\ x_k-x_*\end{pmatrix}\right\| = \left\|M^k\begin{pmatrix}\frac{1}{L}E[e_0]\\ x_0-x_*\end{pmatrix}\right\| = O(\rho^k).$$

The bound on ρ follows from Theorem (3.1). □

We note that a similar line of convergence analysis based on (nonsymmetric) iterative linear systems have long been established for iterative methods for solving linear systems of equations such as Gauss-Seidel [7].

3.2. **General Case.** We can follow the same steps as in the quadratic case to observe convergence of the algorithm in minimizing non-quadratic functions. Lemma 2.2 implies that we have

$$\begin{pmatrix} \frac{1}{L}\|E[e_k]\| \\ \|x_k - x_*\| \end{pmatrix} \leq \bar{M}^k \begin{pmatrix} \frac{1}{L}\|e_0\| \\ \|x_0 - x_*\| \end{pmatrix}, \qquad \text{for} \quad \bar{M} = \begin{pmatrix} (1-\eta)(1+\alpha L) & (1-\eta)\alpha L \\ \alpha L & 1-\alpha\mu \end{pmatrix},$$

since we assume that the random selection of a subset of component functions (with probability η) is independent at each iteration. The next result is a corollary of Theorem 3.2, and can be shown following the same steps as in the proof of Corollary 3.3.

Corollary 3.4. *Suppose the sequence $\{x_k\}$ is produced by (2.1) with y_k defined as in (2.2), and $\alpha = \dfrac{\theta\eta\mu}{L^2}$ with $\theta \in (0, 0.5]$. If Assumptions (A.1)-(A.4) hold, then $\|x_k - x_*\| = O(\rho^k)$ with $\rho \leq 1 - \eta^2\theta\kappa^{-2}$, and $\kappa = \frac{L}{\mu}$.*

Let us finally note that it is in fact possible to consider the system (2.3), and employ M_k (rather than $\bar{M}$) to show convergence in the non-quadratic case. In particular, we can show that M_k satisfies all requirements of the *slowly varying* theorem. We refer to Chapter 12 of [4] for a statement and proof of this theorem. However, the result that we obtain as a consequence of this theorem is valid under stricter conditions as compared to what we get with $\bar{M}$, and is weaker in terms of the rate of convergence.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this article.

Authors Contribution Statement

The author has read and agreed to the published version of the manuscript.

Acknowledgement: The research questions of this paper have been raised during the discussions in early stages of a joint project with Richard Byrd, Jorge Nocedal, and Stefan Solntsev. The generalization of the stochastic aggregated gradient algorithm we analyze in this paper is a *finite-sample* version of the algorithm studied in that project.

References

[1] Bertsekas, D.P., Incremental Gradient, Subgradient, and Proximal Methods for Convex Optimization: A Survey, In: Optimization for Machine Learning, MIT Press, 2012.

[2] Bottou, L., Curtis, F.E., Nocedal, J., *Optimization methods for large-scale machine learning*, SIAM Review, **60**(2018), 223–311.

[3] Gurbuzbalaban, M., Ozdaglar, A., Parrilo, P.A., *On the convergence rate of incremental aggregated gradient algorithms*, SIAM Journal on Optimization, **27**(2017), 1035–1048.

[4] Hartfiel, D.J., Nonhomogeneous Matrix Products, World Scientific, 2002.

[5] Schmidt, M., Le Roux, N., Bach, F., *Minimizing finite sums with the stochastic average gradient*, Mathematical Programming, **162**(2017), 83–112.

[6] Silvester, J.R., *Determinants of block matrices*, The Mathematical Gazette, **84**(2000), 460–467.

[7] Strang, G., Introduction to Applied Mathematics, Wellesley-Cambridge Press, 1986.

[8] Van Loan, C.F., Golub, G., Matrix Computations, The Johns Hopkins University Press, 1996.