Convergence of Gradient Methods with Deterministic and Bounded Noise

Hansi Abeynanda[1*], G. H. Jayantha Lanel[2]
[*1]Sri Lanka Institute of Information Technology
[2]University of Sri Jayewardenepura

ARTICLE INFO

Article History:
Received Date: 01 July 2022
Accepted Date: 15 September 2022

Keywords:
The gradient method; deterministic and bounded noise; distributed optimization; dual decomposition

Citation:
Hansi Abeynanda, G. H. Jayantha Lanel. (2022). *Convergence of Gradient Methods with Deterministic and Bounded Noise*. Proceedings of SLIIT International Conference on Advancements in Sciences and Humanities, (11) October, Colombo, 189 - 195.

ABSTRACT

In this paper, we analyse the effects of noise on the gradient methods for solving a convex unconstraint optimization problem. Assuming that the objective function is with Lipschitz continuous gradients, we analyse the convergence properties of the gradient method when the noise is deterministic and bounded. Our theoretical results show that the gradient algorithm converges to the related optimality within some tolerance, where the tolerance depends on the underlying noise, step size, and the gradient Lipschitz continuity constant of the underlying objective function. Moreover, we consider an application of distributed optimization, where the objective function is a sum of two strongly convex functions. Then the related convergences are discussed based on dual decomposition together with gradient methods, where the associated noise is considered as a consequence of quantization errors. Finally, the theoretical results are verified using numerical experiments.

* Hansi, kavindika.a@sliit.lk

INTRODUCTION

Mathematical optimization is an important process that determines the best possible solution corresponding to the best performance of a quantitatively well-defined system. The related theory is frequently used in many application domains such as machine learning, signal processing, data analysis and modelling, statistics, etc. In general, these large-scaled networked systems consist of a large number of subsystems that operate together to achieve a common goal. In particular, the distributed optimization methods, which enable solving a global problem collectively with many subsystems play an important role in the operation of these large-scale distributed systems. With the existing large set of data volumes in modern systems, the currently used most well-known distributed methods are the first order methods such as gradient/subgradient methods (Nedić & Bertsekas , 2010; Nedić & Ozdaglar, 2009; Nedić, Olshevsky, Ozdaglar, & Tsitsiklis, 2008). However, in the real world applications, the large-scaled distributed systems have to face many barriers when their subsystems make local decisions and exchange information to accomplish their tasks. The main bottleneck arises with communications among subsystems in terms of bandwidth limitations. Moreover, challenges such as computational errors and measurement errors (Abeynanda & Lanel, 2021) are also affecting the operation of distributed systems. Thus, in the real world applications, it prevents the application of distributed methods in the pure form. Consequently, the analysis of distributed methods over inexact settings has received much attention in many application fields.

The subgradient/gradient methods with noise were studied many years back in (Polyak, 1987; Bertsekas & Tsitsiklis, 1999; Solodov & Zavriev, 1998). Authors (Bertsekas & Tsitsiklis, 1999) have discussed the convergence properties of inexact gradient methods using the diminishing step size rule and bounded errors. The bound on the errors depends on the current step size selection, and therefore the diminishing step size conditions lead to diminishing errors which are not realistic in many real life situations. The study by (Solodov & Zavriev, 1998) has focused their work only using a compact constraint set and has shown their results only using the diminishing step size rule. Recent works (Pu, Zeilinger, & Jones, 2017; Michelusi, Scutari, & Lee, 2018; Magn´usson, Shokri-Ghadikolaei, & Li, 2020) have studied the effect of imperfect communication over distributed methods using diminishing errors. However, diminishing errors impose restrictions in many real life applications (measurement errors are persistent). Thus, the study of convergence properties of gradient methods with persistent and bounded errors is an important area of the study which requires further development. Motivated by this, the main focus of this study is to analyse the gradient methods with persistent and bounded errors. More importantly, we explicitly provide a novel convergence proof for the gradient method with bounded noise. More specifically, the main objectives of this study are as follows.

1. Analyse the convergence properties of the gradient method with deterministic and bounded errors using a convex unconstraint optimization problem (see Proposition 1).
2. Solve a problem of minimizing a sum of two strongly convex objective functions with a common decision variable using dual decomposition.
3. Verify the theoretical results using numerical experiments.

MATERIALS AND METHODS

This section is organized as follows. First, we present the notation and important definitions used in this study. Next, we introduce the main problem considered in this paper together with related assumptions. Finally, the gradient algorithm with inexact gradients is discussed.

Notation and Definitions

We use the following notation and definitions that are frequently referenced in our study.

We use $\mathbb{R}$, $\mathbb{R}^n$, $\mathbb{Z}_0^+$, and $\mathbb{R}^{n\times m}$ to denote the set of real numbers, set of real n-vectors, set of nonnegative integers, and set of real $n \times m$ matrices, respectively. For $x \in \mathbb{R}^n$, x^T and $\|x\|$ denote the transpose of x and l_2-norm, respectively. The asymptotic notation “Big oh” is denoted by $O(\cdot)$.

Definition 1 (Gradient Lipschitz Continuity):
A differentiable function $f: \mathbb{R}^n \longrightarrow \mathbb{R}$ is said to have a Lipschitz continuous gradient on $C \subseteq dom\, f$, if there exists $L > 0$ such that
$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$, for all $x, y \in C$, (1)

where L is called the gradient Lipschitz continuous constant of f.

Definition 2 (Strongly Convex Function): A differentiable function $f: \mathbb{R}^n \longrightarrow \mathbb{R}$ is strongly convex on $C \subseteq dom\, f$, if there exists $\mu > 0$ such that

$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \frac{1}{2}\mu t(1-t)\|x - y\|^2$, for all $x, y \in C$, when $0 < \mu < 1$, (2)

where μ is called the strong convexity constant of f.

Problem Formulation and related assumptions
In this paper, we consider the following unconstrained optimization problem:
$\underset{x \in \mathbb{R}^n}{\text{minimize}}\, g(x)$, (3)

where the function $g: \mathbb{R}^n \to \mathbb{R}$ is convex, differentiable, and has a finite minimizer. We let X^* as the set of optimal solutions to the problem (3) and g^* as the optimal value. The following assumption is made on the objective function g.

Assumption 1: The gradient ∇g is Lipschits continuous with constant L (See Definition 1).

Assumption 2: The gradients of g are bounded. I.e, there exists $M > 0$, s.t. $\|\nabla g(x^k)\| \leq M$ for all k.

The Gradient Method with Noise

We solve the problem (3) using the following gradient method with noise:

$x^{k+1} = x^k - \alpha_k \widehat{\nabla} g(x^k)$, (4)

where $\widehat{\nabla} g(x^k) = \nabla g(x^k) + r^k$, k denotes the iteration index, and α_k represents the step size at kth iteration. Here r^k denotes the noise imposed on the system due to the inexactness of the gradient information $\nabla g(x^k)$ at the kth iteration. The inexactness of the gradient can be caused due to numerous errors such as computational errors, quantization errors, etc. We made the following assumption on r^k.
Assumption 3: The error term r^k is deterministic and bounded, i.e., there exists $\epsilon > 0$, s.t. $\|r^k\| < \epsilon$ for all k.

Here, it is worth noting that the error term r^k is deterministic means that r^k is not random, i.e., r^k is not achieved through a stochastic process, e.g., computational errors due to roundoff in arithmetic operations on a computer are deterministic (Polyak, 1987, Section 4.1).

RESULTS AND DISCUSSION

In this section, we discuss the convergence properties of the method (4). First, we will present the following Lemma, which will then use to prove the convergence results.
Lemma 1: Let Assumptions 1, 2, and 3 hold. Then, in method (4) we have
$g(x^{k+1}) \leq g(x^k) - \alpha_k \left(1 - \frac{\alpha_k L}{2}\right) \|\nabla g(x^k)\|^2 + |\alpha_k - \alpha_k^2 L| M\epsilon + \frac{\alpha_k^2 L}{2}\epsilon^2$

Proof: Let $x, y \in \mathbb{R}^n$**.** Since g is differentiable at x, the first order linear approximation at x is given by

$$g(x+y) = g(x) + \nabla g(x)^T y + \int_0^1 [\nabla g(x+\tau y) - \nabla g(x)]^T y \, d\tau, \quad (5)$$

where the integral part represents the remainder term (Polyak, 1987). We let $x = x^k$ and $y = -\alpha_k s^k$, where $s^k = \nabla g(x^k) + r^k$. Then (5) yields

$$g(x^k - \alpha_k s^k) = g(x^k) - \alpha_k \nabla g(x^k)^T s^k - \alpha_k \int_0^1 [\nabla g(x^k - \alpha_k \tau s^k) - \nabla g(x^k)]^T s^k d\tau$$

Thus,

$$g(x^{k+1}) \le g(x^k) - \alpha_k \nabla g(x^k)^T s^k + \alpha_k \int_0^1 \|\nabla g(x^k - \alpha_k \tau s^k) - \nabla g(x^k)\| \|s^k\| d\tau \quad (6)$$

$$\le g(x^k) - \alpha_k \nabla g(x^k)^T s^k + \alpha_k L \|s^k\| \int_0^1 \|x^k - \alpha_k \tau s^k - x^k\| d\tau \quad (7)$$

$$= g(x^k) - \alpha_k \nabla g(x^k)^T s^k + {\alpha_k}^2 L \|s^k\|^2 \int_0^1 \tau \, d\tau \quad (8)$$

$$= g(x^k) - \alpha_k \nabla g(x^k)^T s^k + \frac{{\alpha_k}^2 L}{2} \|s^k\|^2 \quad (9)$$

$$= g(x^k) - \alpha_k \|\nabla g(x^k)\|^2 - \alpha_k \nabla g(x^k)^T r^k + \frac{{\alpha_k}^2 L}{2} \left(\|\nabla g(x^k)\|^2 + 2 \nabla g(x^k)^T r^k + \|r^k\|^2 \right) \quad (10)$$

$$= g(x^k) - \alpha_k \left(1 - \frac{\alpha_k L}{2}\right) \|\nabla g(x^k)\|^2 - (\alpha_k - {\alpha_k}^2 L) \nabla g(x^k)^T r^k + \frac{{\alpha_k}^2 L}{2} \|r^k\|^2 \quad (11)$$

$$\le g(x^k) - \alpha_k \left(1 - \frac{\alpha_k L}{2}\right) \|\nabla g(x^k)\|^2 + |\alpha_k - {\alpha_k}^2 L| \|\nabla g(x^k)\| \|r^k\| + \frac{{\alpha_k}^2 L}{2} \|r^k\|^2 \quad (12)$$

$$\le g(x^k) - \alpha_k \left(1 - \frac{\alpha_k L}{2}\right) \|\nabla g(x^k)\|^2 + |\alpha_k - {\alpha_k}^2 L| M\epsilon + \frac{{\alpha_k}^2 L}{2} \epsilon^2, \quad (13)$$

where (6) follows using (4) and from Cauchy-Schwarz inequality, (7) follows using Assumption 1, (8), (9), (10), and (11) follow using simplifications, (12) follows again from Cauchy-Schwarz inequality and from that $-x \le |x|$ for any $x \in \mathbb{R}$ and, finally, (13) follows using Assumptions 2 and 3.

Next, the convergence of the gradient method (4) is established concerning the norm of the gradients. In particular, we show that $\min_{i \in \{0, \cdots k\}} \|\nabla g(x^i)\|$ converges to 0 with some tolerance. The convergence proof is provided using the constant step size rule. However, it is worth noting that other variations of different step size rules such as square summable but not summable and nonsummable diminishing (Boyd S. , 2014, pp. 4-5) or improved step size schedules (Khirirat, Wang, Magnússon, & Johansson, 2021) can also be considered for related results which require further investigations.

Proposition 1: Suppose Assumptions 1, 2, and 3 hold. Let $\alpha_k = \alpha$ for all $k \in \mathbb{Z}_0^+$ with $0 < \alpha < 2/L$. Then

$$\min_{i \in \{0, \cdots k\}} \|\nabla g(x^i)\| \le \sqrt{\frac{M\epsilon |1 - \alpha L| + \frac{\alpha L}{2} \epsilon^2}{\left(1 - \frac{\alpha L}{2}\right)}}.$$

Proof: Let $\alpha_k = \alpha$**.** Then with the recursive application of Equation (13) we get

$$g(x^{k+1}) \le g(x^0) - \sum_{i=0}^k \alpha \left(1 - \frac{\alpha L}{2}\right) \|\nabla g(x^i)\|^2 + M\epsilon \sum_{i=0}^k |\alpha - \alpha^2 L| + \sum_{i=0}^k \frac{\alpha^2 L}{2} \epsilon^2. \quad (14)$$

Next, by rearranging the terms in (14) we get

$$\sum_{i=0}^k \alpha \left(1 - \frac{\alpha L}{2}\right) \|\nabla g(x^i)\|^2 \le g(x^0) - g(x^{k+1}) + M\epsilon \sum_{i=0}^k |\alpha - \alpha^2 L| + \sum_{i=0}^k \frac{\alpha^2 L}{2} \epsilon^2$$

$$\leq g(x^0) - g(x^*) + M\epsilon \sum_{i=0}^{k} |\alpha - \alpha^2 L| + \sum_{i=0}^{k} \frac{\alpha^2 L}{2} \epsilon^2 , \qquad (15)$$

Where (15) follows using that $g(x^*) \leq g(x), \forall x$ (recall that g is convex).

Then, since $\min_{i \in \{0,\cdots,k\}} \|\nabla g(x^i)\|^2 \leq \|\nabla g(x^i)\|^2$, from (15) we get

$$\min_{i \in \{0,\cdots,k\}} \|\nabla g(x^i)\|^2 \sum_{i=0}^{k} \alpha \left(1 - \frac{\alpha L}{2}\right) \leq g(x^0) - g(x^*) + M\epsilon \sum_{i=0}^{k} |\alpha - \alpha^2 L| + \sum_{i=0}^{k} \frac{\alpha^2 L}{2} \epsilon^2 . \qquad (16)$$

Finally we get

$$\min_{i \in \{0,\cdots,k\}} \|\nabla g(x^i)\|^2 \leq \frac{g(x^0) - g(x^*) + M\epsilon \sum_{i=0}^{k} |\alpha - \alpha^2 L| + \sum_{i=0}^{k} \frac{\alpha^2 L}{2} \epsilon^2}{\sum_{i=0}^{k} \alpha \left(1 - \frac{\alpha L}{2}\right)}$$

$$\leq \frac{g(x^0) - g(x^*)}{\alpha \left(1 - \frac{\alpha L}{2}\right)(k+1)} + \frac{M\epsilon |\alpha - \alpha^2 L| + \frac{\alpha^2 L}{2} \epsilon^2}{\alpha \left(1 - \frac{\alpha L}{2}\right)}. \qquad (17)$$

Then, since $\min_{i \in \{0,\cdots,k\}} \|\nabla g(x^i)\|^2 = \left(\min_{i \in \{0,\cdots,k\}} \|\nabla g(x^i)\|\right)^2$, (17) implies

$$\min_{i \in \{0,\cdots,k\}} \|\nabla g(x^i)\| \leq \sqrt{\frac{g(x^0) - g(x^*)}{\alpha \left(1 - \frac{\alpha L}{2}\right)(k+1)} + \frac{M\epsilon |\alpha - \alpha^2 L| + \frac{\alpha^2 L}{2} \epsilon^2}{\alpha \left(1 - \frac{\alpha L}{2}\right)}}$$

$$\leq \sqrt{\frac{g(x^0) - g(x^*)}{\alpha \left(1 - \frac{\alpha L}{2}\right)(k+1)}} + \sqrt{\frac{M\epsilon |1 - \alpha L| + \frac{\alpha L}{2} \epsilon^2}{\left(1 - \frac{\alpha L}{2}\right)}}. \qquad (18)$$

Thus, the Proposition 1 follows from (18), because $1/(k+1) \to 0$ as $k \to \infty$.

An Application with Numerical Results

We consider a system of two users who collectively solve an optimization problem

$$\underset{x \in \mathrm{X}}{\text{minimize}}\, f(x) = f_1(x) + f_2(x), \qquad (19)$$

where $x \in \mathbb{R}^n$ is known as the public variable, f_i: $\mathbb{R}^n \to \mathbb{R}, i = 1,2$ are strongly convex functions, and the constraint set $\mathrm{X} \subseteq \mathbb{R}^n$ is convex. Then to solve (19) in a distributed manner using dual decomposition, we reformulate the Problem (19) as

$$\underset{x_1, x_2 \in \mathrm{X}}{\text{minimize}} \quad f(x_1, x_2) = f_1(x_1) + f_2(x_2)$$

$$\text{Subject to} \quad x_1 = x_2, \qquad (20)$$

where $x_1 \in \mathrm{X}$ and $x_2 \in \mathrm{X}$ belong to user 1 and user 2, respectively. Then the dual function associated with problem (20) is given by

$$d(\lambda) = \inf_{x_1, x_2 \in \mathrm{X}} [f_1(x_1) + f_2(x_2) + \lambda^T (x_1 - x_2)]$$
$$= \inf_{x_1 \in \mathrm{X}} [f_1(x_1) + \lambda^T x_1] + \inf_{x_2 \in \mathrm{X}} [f_2(x_2) - \lambda^T x_2]. \qquad (21)$$

The dual problem is then given by $\text{maximize}_{\lambda \in \mathbb{R}^n}\, d(\lambda)$. The dual function is always concave (Boyd & Vandenberghe, 2004). Thus, $-d(\lambda)$ is always convex. We let $l(\lambda) = -d(\lambda)$. Then the equivalent minimization problem is given by

$$\underset{\lambda \in \mathbb{R}^n}{\text{minimize}} \quad l(\lambda) \qquad (22)$$

Then, the standard gradient method to solve (22) is given by

$$\lambda^{k+1} = \lambda^k - \alpha_k \nabla l(\lambda^k), \qquad (23)$$

where $\nabla l(\lambda^k) = x_2^k - x_1^k$ [see Equation (21)]. Here we note that some communication is required between user 1 and user 2 to construct $\nabla l(\lambda^k)$. Then, the dual variable λ can be updated by both users in parallel. However, the communication among users may not be perfect in practice due to unavoidable reasons such as quantization errors (Magn´usson, Shokri-Ghadikolaei, & Li, 2020), computational errors, etc. Thus, instead of the exact gradient $\nabla l(\lambda^k)$, we consider an inexact gradient $\widehat{\nabla} l(\lambda^k) = \hat{x}_2^k - \hat{x}_1^k$, where $\hat{x}_1^k = x_1^k + r_1^k$, and $\hat{x}_2^k = x_2^k + r_2^k$, are the quantized versions of x_1^k and x_2^k, respectively. Here, r_1^k and r_2^k represent errors due to quantization. Then, we can solve (22) using the inexact gradient method (4). Clearly, $\|r_1^k\|$ and $\|r_2^k\|$ are bounded because they entirely depend on the size of the quantization grid. Thus, the total error $\|r^k\| = \|\widehat{\nabla} l(\lambda^k) - \nabla l(\lambda^k)\|$ is also bounded. Hence, Assumption 3 holds. Moreover, $\|\nabla l(\lambda^k)\|$ is also bounded because $x_1^k, x_2^k \in X$ for all k**.** It follows that Assumption 2 holds. Further, the dual function $l(\lambda)$ is with Lipschitz continuous gradients since the objective function $f(x_1, x_2)$ is strongly convex (Magn´usson, Shokri-Ghadikolaei, & Li, 2020), which is then leading to Assumption 1. Thus, we can observe that the convergence properties of the method (4) with the negative dual function $l(\lambda)$ are guaranteed using Proposition 1.

We illustrate the convergence results in Proposition 1 numerically. We let $f_1$and f_2 in (19) are quadratic functions such that, $f_i = x_i^T A_i x_i + a_i^T x_i, i = 1, 2$, where each $A_i \in \mathbb{R}^{n \times n}$ is positive definite and $x_i, a_i \in \mathbb{R}^n$. Moreover, we let $X = [-2, 2]^n$, the n-fold cartesian product of $[-2, 2]$, (i.e., X is a hypercube in $\mathbb{R}^n$ with each side 4 units in length). We consider that the users transmit their quantized data at a rate of b bits per dimension in each iteration. To implement the quantization, we divide each side of X into 2^b parts**.** Then, X consists of 2^{nb} similar small hypercubes, each with a width of $4/2^b$**.** Then at each iteration - k, we choose the quantized data $\hat{x}_i^k$ of $x_i^k, i = 1,2$ as the mid-point of the associated small hypercube that x_i^k lies. Figure 1 depicts the convergence results using $n = 2$ for different bits (in this case, X is a square in $\mathbb{R}^2$). Results show that $\min_{i \in \{0, \cdots, k\}} \|\nabla l(\lambda^i)\|$ converges around zero within some tolerance, and the tolerance decreases when the number of communicated bits increases. Figure 2 depicts the convergence results using $b = 10$ for different dimensions of x_is. Results show that the related tolerance get increases when the dimension n of the decision variable x increases.

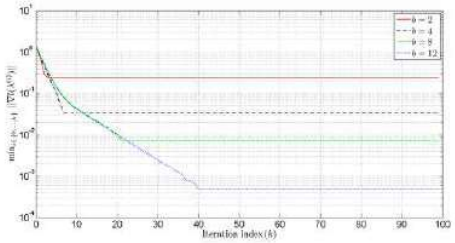

Figure 1: Convergence of dual gradients with different bits.

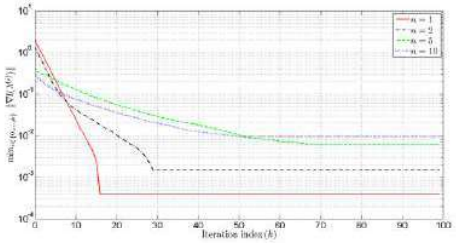

Figure 2: Convergence of dual gradients with different dimensions of the decision variable x.

CONCLUSION

In this paper, we have considered a problem of minimizing a convex function with Lipschitz continuous gradients. The gradient method with inexact gradients is considered to solve the underlying optimization problem. The error involved due to the inexactness of the gradient is considered deterministic and bounded. The convergence properties of the gradient method with inexact gradients are analysed with the constant step size rule. Our theoretical results show that the inexact gradient method can converge to the optimality within some tolerance, where the tolerance depends on the underlying inexactness, step size, and the gradient Lipschitz continuity constant of the underlying objective function. Minimizing a sum of two strongly convex functions with a common decision variable is considered with dual decomposition as an application of distributed optimization. Finally, the theoretical results were verified using numerical experiments.

REFERENCES

Abeynanda , H., & Lanel, G. (2021). A study on distributed optimization over largescale networked systems. Journal of Mathematics, 2021.

Bertsekas, D., & Tsitsiklis, J. (1999). Gradient Convergence in Gradient methods with Errors. SIAM Journal on Optimization, 10(3), 627–642.

Boyd, S. (2014). Subgradient Methods. Retrieved from http://stanford.edu/class/ee364b/lectures/subgrad_method_notes.pdf

Boyd, S., & Vandenberghe, L. (2004). Convex Optimization. USA: Cambridge University Press.

Chen, J., & Luss, R. (2018). Stochastic Gradient Descent with Biased but Consistent Gradient Estimators. Retrieved from http://arxiv.org/abs/1807.11880

Khirirat, S., Wang, X., Magnússon, S., & Johansson, M. (2021). Improved Step-Size Schedules for Noisy Gradient Methods. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3655-3659). IEEE.

Magn´usson, S., Shokri-Ghadikolaei, H., & Li, N. (2020). On Maintaining Linear Convergence of Distributed Learning and Optimization Under Limited Communication. IEEE Transactions on Signal Processing, 68, 6101–6116.

Michelusi, N., Scutari, G., & Lee, C.-S. (2018). inite rate quantized distributed optimization with geometric convergence. 2018 52nd Asilomar Conference on Signals, Systems, and Computers, (pp. 1876–1880).

Nedić, A., & Bertsekas , D. (2010). The effect of deterministic noise in subgradient methods. Mathematical Programming, 125(1), 75–99.

Nedić, A., & Ozdaglar, A. (2009). Distributed Subgradient Methods for Multi-Agent Optimization. IEEE Transactions on Automatic Control, 54(1), 48-61.

Nedić, A., Olshevsky, A., Ozdaglar, A., & Tsitsiklis, J. N. (2008). Distributed subgradient methods and quantization effects. 2008 47th IEEE Conference on Decision and Control, (pp. 4177-4184).

Polyak, B. T. (1987). Introduction to Optimization. NY: Inc., Publications Division.

Pu, Y., Zeilinger, M., & Jones, C. (2017). Quantization design for distributed optimization. EEE Transactions on Automatic Control, 62(5), 2107–2120.

Solodov , M., & Zavriev, S. (1998). Error stability properties of generalized gradient-type algorithms. Journal of Optimization Theory, 98(3), 663–680.