



# Self-adaptive algorithms for quasiconvex programming and applications to machine learning

Tran Ngoc Thang<sup>1</sup> · Trinh Ngoc Hai<sup>1</sup>

Received: 26 August 2023 / Revised: 29 January 2024 / Accepted: 29 April 2024

© The Author(s) under exclusive licence to Sociedade Brasileira de Matemática Aplicada e Computacional 2024

## Abstract

For solving a broad class of nonconvex programming problems on an unbounded constraint set, we provide a self-adaptive step-size strategy that does not include line-search techniques and establishes the convergence of a generic approach under mild assumptions. Specifically, the objective function may not satisfy the convexity condition. Unlike descent line-search algorithms, it does not need a known Lipschitz constant to figure out how big the first step should be. The crucial feature of this process is the steady reduction of the step size until a certain condition is fulfilled. In particular, it can provide a new gradient projection approach to optimization problems with an unbounded constrained set. To demonstrate the effectiveness of the proposed technique for large-scale problems, we apply it to some experiments on machine learning, such as supervised feature selection, multi-variable logistic regressions and neural networks for classification.

**Keywords** Nonconvex programming · Gradient descent algorithms · Quasiconvex functions · Pseudoconvex functions · Self-adaptive step-sizes

## 1 Introduction

Gradient descent methods are a common tool for a wide range of programming problems, from convex to nonconvex, and have numerous practical applications (see Boyd and Vandenberghe 2009; Cevher et al. 2014; Lan 2020 and references therein). At each iteration, gradient descent algorithms provide an iterative series of solutions based on gradient directions and step sizes. For a long time, researchers have focused on finding the direction to improve the convergence rate of techniques, while the step-size was determined using one of the few well-known approaches (see Boyd and Vandenberghe 2009; Nesterov 2013).

---

Dedicated to Professor Pham Ky Anh on the occasion of his 75th birthday with admiration and respect.

---

✉ Trinh Ngoc Hai  
hai.trinhngoc@hust.edu.vn

Tran Ngoc Thang  
thang.tranngoc@hust.edu.vn

<sup>1</sup> School of Applied Mathematics and Informatics, Hanoi University of Science and Technology, Hai Ba Trung, Hanoi, Vietnam

Recently, new major areas of machine learning applications with high dimensionality and nonconvex objective functions have required the development of novel step-size choosing procedures to reduce the method's overall computing cost (see Cevher et al. 2014; Lan 2020). The exact or approximate one-dimensional minimization line-search incurs significant computational costs per iteration, particularly when calculating the function value is nearly identical to calculating its derivative and requires the solution of complex auxiliary problems (see Boyd and Vandenberghe 2009). To avoid the line-search, the step-size value may be calculated using prior information such as Lipschitz constants for the gradient. However, this requires using just part of their inexact estimations, which leads to slowdown convergence. This is also true for the well-known divergent series rule (see Kiwiel 2001; Nesterov 2013).

Although Kiwiel developed the gradient approach for quasiconvex programming in 2001 (Kiwiel 2001), it results in slow convergence due to the use of diminishing step-size. Following that, there are some improvements to the gradient method, such as the work of (Yu et al. 2019; Hu et al. 2020), which use a constant step-size but the objective function must fulfil the Hölder condition. The other method is the neurodynamic approach, which uses recurrent neural network models to solve pseudoconvex programming problems with unbounded constraint sets (see Bian et al. 2018; Liu et al. 2022). Its step-size selection is fixed and not adaptive.

The adaptive step-size procedure was proposed in Konnov (2018) and Ferreira and Sosa (2022). The method given in Konnov (2018), whereby a step-size algorithm for the conditional gradient method is proposed, is effective for solving pseudoconvex programming problems with the boundedness of the feasible set. It has been expanded in Ferreira and Sosa (2022) to the particular case of an unbounded feasible set, which cannot be applied to the unconstrained case.

In this research, we propose a novel and line-search-free adaptive step-size algorithm for a broad class of programming problems where the objective function is nonconvex and smooth, the constraint set is unbounded, closed and convex. A crucial component of this procedure is gradually decreasing the step size until a predetermined condition is fulfilled. Although the Lipschitz continuity of the gradient of the objective function is required for convergence, the approach does not make use of predetermined constants. The proposed change has been shown to be effective in preliminary computational tests. We perform various machine learning experiments, including multi-variable logistic regressions and neural networks for classification, to show that the proposed method performs well on large-scale tasks.

Extending the adaptive step-size technique to the case of unbounded constraint sets is one of the main contributions of our paper. This task encounters some difficulties. Firstly, the convergence of the algorithm has to be ensured without any additional auxiliary conditions. The convergent assertions have been proven by utilizing the properties of the objective function associated with inequality transformations. Secondly, projection operators ensure that the point  $x^{k+1}$  generated from  $x^k$  lies within the admissible set, but we need to prove the algorithm's convergence in the presence of this projection. Lastly, the proposed adaptive algorithm must encompass the case of fixed step-size. This case demonstrates that the proposed algorithm is a natural extension of the typical Gradient Descent (GD) algorithm and advantageous for real-world applications, particularly for non-convex objective functions. Computational examples for large-scale problems with non-convex objective functions substantiate this claim.

The rest of this paper is structured as follows: Sect. 2 provides preliminary information and details the problem formulation. Section 3 summarizes the primary results, including the proposed algorithms. Section 4 depicts numerical experiments and analyzes their computa-

tional outcomes. Section 5 presents applications to certain machine learning problems. The final section draws some conclusions.

## 2 Preliminaries

In the whole paper, we assume that  $C$  is a nonempty, closed and convex set in  $\mathbb{R}^m$ ,  $f : \mathbb{R}^m \rightarrow \mathbb{R}$  is a differentiable function on an open set containing  $C$ , the mapping  $\nabla f$  is  $L$ -Lipschitz continuous, i.e. there exists a constant  $L > 0$  such that  $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$  for all  $x, y \in C$ . We consider the optimization problem:

$$\min_{x \in C} f(x). \quad (\text{OP}(f, C))$$

Assume that the solution set of  $(\text{OP}(f, C))$  is not empty. First, we recall some definitions and basic results that will be used in the next section of the article. The interested reader is referred to Bauschke and Combettes (2011) and Rockafellar (1970) for bibliographical references.

For  $x \in \mathbb{R}^m$ , denote by  $P_C(x)$  the projection of  $x$  onto  $C$ , i.e.,

$$P_C(x) := \operatorname{argmin}\{\|z - x\| : z \in C\}.$$

**Proposition 1** (Bauschke and Combettes 2011) *It holds that*

- (i)  $\|P_C(x) - P_C(y)\| \leq \|x - y\|$  for all  $x, y \in \mathbb{R}^m$ ,
- (ii)  $\langle y - P_C(x), x - P_C(x) \rangle \leq 0$  for all  $x \in \mathbb{R}^m$ ,  $y \in C$ .

**Definition 1** (Mangasarian 1965) The function  $f : \mathbb{R}^m \rightarrow \mathbb{R}$  is said to be

- convex on  $C$  if for all  $x, y \in C$ ,  $\lambda \in [0, 1]$ , it holds that

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).$$

- pseudoconvex on  $C$  if for all  $x, y \in C$ , it holds that

$$\langle \nabla f(x), y - x \rangle \geq 0 \Rightarrow f(y) \geq f(x).$$

- quasiconvex on  $C$  if for all  $x, y \in C$ ,  $\lambda \in [0; 1]$ , it holds that

$$f(\lambda x + (1 - \lambda)y) \leq \max\{f(x); f(y)\}.$$

**Proposition 2** (Dennis and Schnabel 1983) *The differentiable function  $f$  is quasiconvex on  $C$  if and only if*

$$f(y) \leq f(x) \Rightarrow \langle \nabla f(x), y - x \rangle \leq 0.$$

It is worth mentioning that “ $f$  is convex”  $\Rightarrow$  “ $f$  is pseudoconvex”  $\Rightarrow$  “ $f$  is quasiconvex” (see Mangasarian 1965).

**Proposition 3** (Dennis and Schnabel 1983) *Suppose that  $\nabla f$  is  $L$ -Lipschitz continuous on  $C$ . For all  $x, y \in C$ , it holds that*

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2} \|y - x\|^2.$$

**Lemma 1** (Xu 2002) *Let  $\{a_k\}; \{b_k\} \subset (0; \infty)$  be sequences such that*

$$a_{k+1} \leq a_k + b_k \quad \forall k \geq 0; \quad \sum_{k=0}^{\infty} b_k < \infty.$$

*Then, there exists the limit  $\lim_{k \rightarrow \infty} a_k = c \in \mathbb{R}$ .*

### 3 Main results

**Algorithm 1** (*Gradient Descent Adaptive Algorithm-GDA*)

**Step 0.** Choose  $x^0 \in C$ ,  $\lambda_0 \subset (0, +\infty)$ ,  $\sigma, \kappa \in (0, 1)$ . Set  $k = 0$ .

**Step 1.** Given  $x^k$  and  $\lambda_k$ , compute  $x^{k+1}$  and  $\lambda_{k+1}$  as

$$x^{k+1} = P_C(x^k - \lambda_k \nabla f(x^k)),$$

If  $f(x^{k+1}) \leq f(x^k) - \sigma \langle \nabla f(x^k), x^k - x^{k+1} \rangle$  then set  $\lambda_{k+1} := \lambda_k$  else set  $\lambda_{k+1} := \kappa \lambda_k$ .

**Step 2.** Update  $k := k + 1$ . If  $x^{k+1} = x^k$  then STOP else go to Step 1.

**Remark 1** If Algorithm GDA stops at step  $k$ , then  $x^k$  is a stationary point of the problem  $\text{OP}(f, C)$ . Indeed, since  $x^{k+1} = P_C(x^k - \lambda_k \nabla f(x^k))$ , applying Proposition 1-(ii), we have

$$\langle z - x^{k+1}, x^k - \lambda_k \nabla f(x^k) - x^{k+1} \rangle \leq 0 \quad \forall z \in C. \quad (1)$$

If  $x^{k+1} = x^k$ , we get

$$\langle \nabla f(x^k), z - x^k \rangle \geq 0 \quad \forall z \in C, \quad (2)$$

which means  $x^k$  is a stationary point of the problem. If, in addition,  $f$  is pseudoconvex, from (2), it implies that  $f(z) \geq f(x^k)$  for all  $z \in C$ , or  $x^k$  is a solution of  $\text{OP}(f, C)$ .

Now, suppose that the algorithms generate an infinite sequence. We will prove that this sequence converges to a solution to the problem  $\text{OP}(f, C)$ .

**Theorem 1** Assume that the sequence  $\{x^k\}$  is generated by Algorithm GDA. Then, the sequence  $\{f(x^k)\}$  is convergent and each limit point (if any) of the sequence  $\{x^k\}$  is a stationary point of the problem. Moreover,

- if  $f$  is quasiconvex on  $C$ , then the sequence  $\{x^k\}$  converges to a stationary point of the problem.
- if  $f$  is pseudoconvex on  $C$ , then the sequence  $\{x^k\}$  converges to a solution of the problem.

**Proof** Applying Proposition 2, we get

$$f(x^{k+1}) \leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2. \quad (3)$$

In (1), taking  $z = x^k \in C$ , we arrive at

$$\langle \nabla f(x^k), x^{k+1} - x^k \rangle \leq -\frac{1}{\lambda_k} \|x^{k+1} - x^k\|^2. \quad (4)$$

Combining (3) and (4), we obtain

$$f(x^{k+1}) \leq f(x^k) - \sigma \langle \nabla f(x^k), x^k - x^{k+1} \rangle - \left( \frac{1-\sigma}{\lambda_k} - \frac{L}{2} \right) \|x^{k+1} - x^k\|^2. \quad (5)$$

We now claim that  $\{\lambda_k\}$  is bounded away from zero, or in other words, the step size changes finite times. Indeed, suppose, by contrary, that  $\lambda_k \rightarrow 0$ . From (5), there exists  $k_0 \in \mathbb{N}$  satisfying

$$f(x^{k+1}) \leq f(x^k) - \sigma \langle \nabla f(x^k), x^k - x^{k+1} \rangle \quad \forall k \geq k_0.$$

According to the construction of  $\lambda_k$ , the last inequality implies that  $\lambda_k = \lambda_{k_0}$  for all  $k \geq k_0$ . This is a contradiction. And so, there exists  $k_1 \in \mathbb{N}$  such that for all  $k \geq k_1$ , we have  $\lambda_k = \lambda_{k_1}$  and

$$f(x^{k+1}) \leq f(x^k) - \sigma \langle \nabla f(x^k), x^k - x^{k+1} \rangle. \quad (6)$$

Noting that  $\langle \nabla f(x^k), x^k - x^{k+1} \rangle \geq 0$ , we infer that the sequence  $\{f(x^k)\}$  is convergent and

$$\sum_{k=0}^{\infty} \langle \nabla f(x^k), x^k - x^{k+1} \rangle < \infty; \quad \sum_{k=0}^{\infty} \|x^{k+1} - x^k\|^2 < \infty. \quad (7)$$

From (1), for all  $z \in C$ , we have

$$\begin{aligned} \|x^{k+1} - z\|^2 &= \|x^k - z\|^2 - \|x^{k+1} - x^k\|^2 + 2 \langle x^{k+1} - x^k, x^{k+1} - z \rangle \\ &\leq \|x^k - z\|^2 - \|x^{k+1} - x^k\|^2 + 2\lambda_k \langle \nabla f(x^k), z - x^{k+1} \rangle. \end{aligned} \quad (8)$$

Let  $\bar{x}$  be a limit point of  $\{x^k\}$ . There exists a subsequence  $\{x^{k_i}\} \subset \{x^k\}$  such that  $\lim_{i \rightarrow \infty} x^{k_i} = \bar{x}$ . In (8), let  $k = k_i$  and take the limit as  $i \rightarrow \infty$ . Noting that  $\|x^k - x^{k+1}\| \rightarrow 0$ ,  $\nabla f$  is continuous, we get

$$\langle \nabla f(\bar{x}), z - \bar{x} \rangle \geq 0 \quad \forall z \in C,$$

which means  $\bar{x}$  is a stationary point of the problem. Now, suppose that  $f$  is quasiconvex on  $C$ . Denote

$$U := \left\{ x \in C : f(x) \leq f(x^k) \quad \forall k \geq 0 \right\}.$$

Note that  $U$  contains the solution set of  $\text{OP}(f, C)$ , and hence, is not empty. Take  $\hat{x} \in U$ . Since  $f(x^k) \geq f(\hat{x})$  for all  $k \geq 0$ , it implies that

$$\langle \nabla f(x^k), \hat{x} - x^k \rangle \leq 0 \quad \forall k \geq 0. \quad (9)$$

Combing (8) and (9), we get

$$\|x^{k+1} - \hat{x}\|^2 \leq \|x^k - \hat{x}\|^2 - \|x^{k+1} - x^k\|^2 + 2\lambda_k \langle \nabla f(x^k), x^k - x^{k+1} \rangle. \quad (10)$$

Applying Lemma 1 with  $a_k = \|x^{k+1} - \hat{x}\|^2$ ,  $b_k = 2\lambda_k \langle \nabla f(x^k), x^k - x^{k+1} \rangle$ , we deduce that the sequence  $\{\|x^k - \hat{x}\|\}$  is convergent for all  $\hat{x} \in U$ . Since the sequence  $\{x^k\}$  is bounded, there exist a subsequence  $\{x^{k_j}\} \subset \{x^k\}$  such that  $\lim_{j \rightarrow \infty} x^{k_j} = \bar{x} \in C$ . From (6), we know that the sequence  $\{f(x^k)\}$  is nonincreasing and convergent. It implies that  $\lim_{k \rightarrow \infty} f(x^k) = f(\bar{x})$  and  $f(\bar{x}) \leq f(x^k)$  for all  $k \geq 0$ . This means  $\bar{x} \in U$  and the sequence  $\{\|x^k - \bar{x}\|\}$  is convergent. Thus,

$$\lim_{k \rightarrow \infty} \|x^k - \bar{x}\| = \lim_{i \rightarrow \infty} \|x^{k_i} - \bar{x}\| = 0.$$

Note that each limit point of  $\{x^k\}$  is a stationary point of the problem. Then, the whole sequence  $\{x^k\}$  converges to  $\bar{x}$  - a stationary point of the problem. Moreover, when  $f$  is pseudoconvex, this stationary point becomes a solution of  $\text{OP}(f, C)$ .  $\square$

**Remark 2** In Algorithm GDA, we can choose  $\lambda_0 = \lambda$ , with the constant number  $\lambda \leq 2(1 - \sigma)/L$ . Then, we get  $(1 - \sigma)/\lambda_0 - L/2 \geq 0$ . Combined with (5), it implies that the condition  $f(x^{k+1}) \leq f(x^k) - \sigma \langle \nabla f(x^k), x^k - x^{k+1} \rangle$  is satisfied and the step-size  $\lambda_k = \lambda$  for all step  $k$ . Therefore, Algorithm GDA is still applicable for the constant step-size  $\lambda \leq 2(1 - \sigma)/L$ . For

any  $\lambda \in (0, 2/L)$ , there exists  $\sigma \in (0, 1)$  such that  $\lambda \leq 2(1 - \sigma)/L$ . As a result, if the value of Lipschitz constant  $L$  is prior known, we can choose the constant step-size  $\lambda \in (0, 2/L)$  as in the gradient descent (GD) algorithm for solving convex programming problems. This GD algorithm has been proposed in previous works. Because it is a special case of Algorithm GDA, its convergence is guaranteed as the assertions in Theorem 1.

**Algorithm 2** (*Gradient Descent Algorithm-GD*)

**Step 0.** Choose  $x^0 \in C$ ,  $\lambda \subset (0, 2/L)$ . Set  $k = 0$ .

**Step 1.** Given  $x^k$ , compute  $x^{k+1}$  as

$$x^{k+1} = P_C(x^k - \lambda \nabla f(x^k))$$

**Step 2.** Update  $k := k + 1$ . If  $x^{k+1} = x^k$  then STOP else go to Step 1.

Note that all assertions of Theorem 1 are still true for the sequence  $\{x^k\}$  generated by Algorithm GD. Now, we estimate the convergence rate of Algorithm GDA in solving unconstrained optimization problems.

**Corollary 1** Assume that  $f$  is convex,  $C = \mathbb{R}^m$  and  $\{x^k\}$  is the sequence generated by Algorithm GDA. Then,

$$f(x^k) - f(x^*) = O\left(\frac{1}{k}\right),$$

where  $x^*$  is a solution of the problem.

**Proof** Let  $x^*$  be a solution of the problem. Denote  $\Delta_k := f(x^k) - f(x^*)$ . From (6), noting that  $x^k - x^{k+1} = \lambda_k \nabla f(x^k)$ , we have

$$\Delta_{k+1} \leq \Delta_k - \sigma \lambda_{k_1} \|\nabla f(x^k)\|^2 \quad \forall k \geq k_1. \quad (11)$$

On the other hand, since the sequence  $\{x^k\}$  is bounded and  $f$  is convex, it holds that

$$\begin{aligned} 0 &\leq \Delta_k \leq \langle \nabla f(x^k), x^k - x^* \rangle \\ &\leq M \|\nabla f(x^k)\|, \end{aligned} \quad (12)$$

where  $M := \sup \{\|x^k - x^*\| : k \geq k_1\} < \infty$ . From (11) and (12), we arrive at

$$\Delta_{k+1} \leq \Delta_k - Q \Delta_k^2 \quad \forall k \geq k_1, \quad (13)$$

where  $Q := \frac{\sigma \lambda_{k_1}}{M^2}$ . Noting that  $\Delta_{k+1} \leq \Delta_k$ , from (13), we obtain

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + Q \geq \dots \geq \frac{1}{\Delta_{k_1}} + (k - k_1)Q,$$

which implies

$$f(x^k) - f(x^*) = O\left(\frac{1}{k}\right).$$

□

To conclude this section, we present a stochastic variation of Algorithm GDA for application in large scale deep learning. Consider the following problem:

$$\min_x \mathbb{E}[f_\xi(x)],$$

where  $\xi$  is the stochastic parameter and the function  $f_\xi$  is  $L$ -smooth. We are generating a stochastic gradient  $\nabla f_{\xi^k}(x^k)$  by sampling  $\xi$  at each iteration  $k$ .

The stochastic variation of the gradient descent method, especially in the context of large-scale deep learning, plays a crucial role in optimizing complex models efficiently. When we consider the optimization problem of the form:

$$\min_x \mathbb{E} [f_\xi(x)],$$

where  $x$  represents the parameters of the model (like weights in a neural network),  $\xi$  is a stochastic parameter, and  $f_\xi(x)$  is an  $L$ -smooth function, we are dealing with a scenario where the objective function is defined as the expected value of some random function  $f_\xi(x)$ . This framework is typical in machine learning, where  $f_\xi(x)$  often represents the loss function computed on a subset (batch) of the training data, and  $\xi$  represents the randomness in the choice of this subset.

The following Algorithm SGDA contains a detailed description. We leave rigorous theoretical results of Algorithm SDGA for future work.

**Algorithm 3** (*Stochastic Gradient Descent Adaptive Algorithm-SGDA*)

**Step 0.** Choose  $x^0 \in C$ ,  $\lambda_0 \subset (0, +\infty)$ ,  $\sigma, \kappa \in (0, 1)$ . Set  $k = 0$ .

**Step 1.** Sample  $\xi^k$  and compute  $x^{k+1}$  and  $\lambda_{k+1}$  as

$$x^{k+1} = P_C(x^k - \lambda_k \nabla f_{\xi^k}(x^k)),$$

If  $f_{\xi^k}(x^{k+1}) \leq f_{\xi^k}(x^k) - \sigma \langle \nabla f_{\xi^k}(x^k), x^k - x^{k+1} \rangle$  then set  $\lambda_{k+1} := \lambda_k$  else set  $\lambda_{k+1} := \kappa \lambda_k$ .

**Step 2.** Update  $k := k + 1$ . If  $x^{k+1} = x^k$  then STOP else go to Step 1.

## 4 Numerical experiments

In this section, we use two existing cases and two large-scale examples with variable sizes to validate the performance of the proposed method. Applied experiments in machine learning will be implemented in the next section. The source code is available at: <https://github.com/TranNgocThangHUST/AdaptiveGDforQCP>.

All tests are carried out in Python on a MacBook Pro M1 (3.2GHz 8-core processor, 8.00 GB of RAM). The stopping criterion in the following cases is “the number of iterations  $\leq \#Iter$ ” where  $\#Iter$  is the maximum number of iterations. Denote  $x^*$  as the limit point of iterated series  $\{x^k\}$  and Time the CPU time of the GDA algorithm using the stop criterion.

Nonconvex objective and constraint functions are illustrated in Examples 1 and 2 taken from Liu et al. (2022). In addition, Example 2 is more complex than Example 1 regarding the form of functions. Implementing Example 3 with a convex objective function and a variable dimension  $n$  allows us to evaluate the proposed method compared to Algorithm GD. Example 3 is implemented with a pseudoconvex objective function and several values for  $n$  so that we may estimate our approach compared to Algorithm RNN.

To compare to neurodynamic method in Liu et al. (2022), we consider Problem **OP(f, C)** with the constraint set is determined specifically by  $C = \{x \in \mathbb{R}^n \mid g(x) \leq 0, Ax = b\}$ , where  $g(x) := (g_1(x), g_2(x), \dots, g_m(x))^\top$  and  $g_i : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $i = 1, \dots, m$  are differential quasiconvex, the matrix  $A \in \mathbb{R}^{p \times n}$  and  $b = (b_1, b_2, \dots, b_p)^\top \in \mathbb{R}^p$ . Recall that, in Liu et al. (2022), the authors introduced the function

$$\Psi(s) = \begin{cases} 1, & s > 0; \\ [0, 1], & s = 0; \\ 0, & s < 0. \end{cases},$$

and

$$P(x) = \sum_{i=1}^m \max \{0, g_i(x)\}.$$

The neurodynamic algorithm established in Liu et al. (2022), which uses recurrent neural network (RNN) models for solving Problem  $\text{OP}(f, C)$  in the form of differential inclusion as follows:

$$\frac{d}{dt}x(t) \in -c(x(t))\nabla f(x(t)) - \partial P(x(t)) - \partial \|Ax(t) - b\|_1 \quad (\text{RNN})$$

where the adjusted term  $c(x(t))$  is

$$c(x(t)) = \left\{ \prod_{i=1}^{m+p} c_i(t) \mid c_i(t) \in 1 - \Psi(J_i(x(t))), i = 1, 2, \dots, m+p \right\}$$

with

$$J(x) = (g_1(x), \dots, g_m(x), |A_1x - b_1|, \dots, |A_p x - b_p|)^\top,$$

the subgradient term of  $P(x)$  is

$$\partial P(x) = \begin{cases} 0, & x \in \text{int}(X) \\ \sum_{i \in I_0(x)} [0, 1]\nabla g_i(x), & x \in \text{bd}(X) \\ \sum_{i \in I_+(x)} \nabla g_i(x) + \sum_{i \in I_0(x)} [0, 1]\nabla g_i(x), & x \notin X, \end{cases}$$

with

$$X = \{x : g_i(x) \leq 0, i = 1, 2, \dots, m\},$$

$$I_+(x) = \{i \in \{1, 2, \dots, m\} : g_i(x) > 0\},$$

$$I_0(x) = \{i \in \{1, 2, \dots, m\} : g_i(x) = 0\},$$

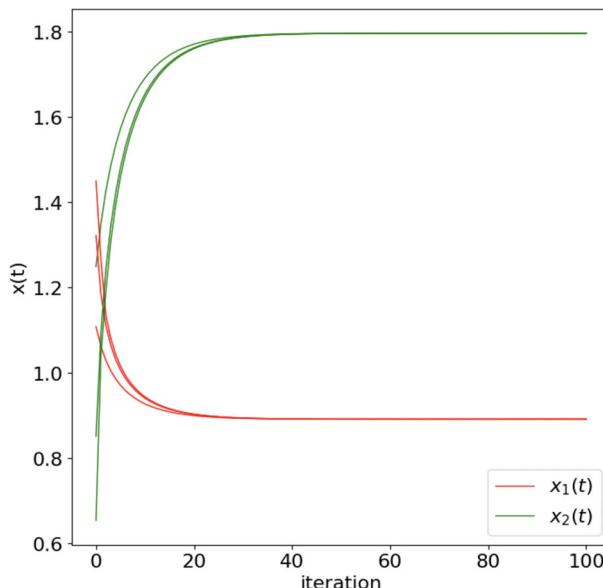
and the subgradient term of  $\|Ax - b\|_1$  is

$$\partial \|Ax - b\|_1 = \sum_{i=1}^p (2\Psi(A_i x - b_i) - 1) A_i^\top,$$

with  $A_i \in \mathbb{R}^{1 \times n}$  ( $i = 1, 2, \dots, p$ ) be the row vectors of the matrix  $A$ .

**Example 1** First, let's look at a simple nonconvex problem  $\text{OP}(f, C)$ :

$$\begin{aligned} \text{minimize } f(x) &= \frac{x_1^2 + x_2^2 + 3}{1 + 2x_1 + 8x_2} \\ \text{subject to } x &\in C, \end{aligned}$$



**Fig. 1** Computational results for Example 1

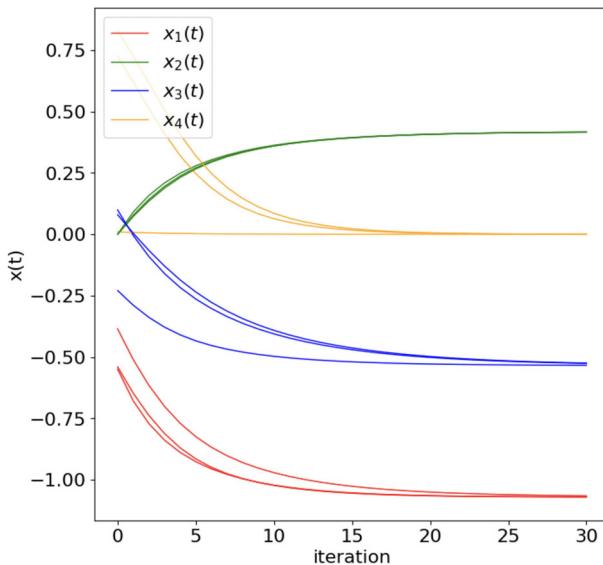
where  $C = \{x = (x_1, x_2)^\top \in \mathbb{R}^2 | g_1(x) = -x_1^2 - 2x_1x_2 \leq -4; x_1, x_2 \geq 0\}$ . It is quite evident that for this problem, the objective function  $f$  is pseudoconvex on the convex feasible set (Example 5.2 in Liu et al. (2022)).

Figure 1 illustrates the temporary solutions of the proposed method for various initial solutions. It demonstrates that the outcomes converge to the optimal solution  $x^* = (0.8922, 1.7957)$  of the given problem. The objective function value generated by Algorithm GDA is 0.4094, which is better than the optimum value 0.4101 of the neural network in Liu et al. (2022).

**Example 2** Consider the following nonsmooth pseudoconvex optimization problem with non-convex inequality constraints (Example 5.1 in Liu et al. (2022)):

$$\begin{aligned} & \text{minimize } f(x) = \frac{e^{|x_2-3|} - 30}{x_1^2 + x_3^2 + 2x_4^2 + 4} \\ & \text{subject to } g_1(x) = (x_1 + x_3)^3 + 2x_4^2 \leq 10, \\ & \quad g_2(x) = (x_2 - 1)^2 \leq 1, \\ & \quad 2x_1 + 4x_2 + x_3 = -1, \end{aligned}$$

where  $x = (x_1, x_2, x_3, x_4)^\top \in \mathbb{R}^4$ . The objective function  $f(x)$  is nonsmooth pseudoconvex on the feasible region  $C$ , and the inequality constraint  $g_1$  is continuous and quasiconvex on  $C$ , but not pseudoconvex (Example 5.1 in Liu et al. (2022)). It is easily verified that  $x_2 \neq 3$  for any  $x \in C$ . Therefore, the gradient vector of  $|x_2 - 3|$  is  $(x_2 - 3) / |x_2 - 3|$  for any  $x \in C$ . From that, we can establish the gradient vector of  $f(x)$  at a point  $x \in C$  used in the algorithm. Figure 2 shows that Algorithm GDA converges to an optimal solution  $x^* = (-1.0649, 0.4160, -0.5343, 0.0002)^\top$  with the optimal value  $-3.0908$ , which is better than the optimal value  $-3.0849$  of the neural network model in Liu et al. (2022).



**Fig. 2** Computational results for Example 2

**Example 3** Let  $e := (1, \dots, n) \in \mathbb{R}^n$  be a vector,  $\alpha > 0$  and  $\beta > 0$  be constants satisfying the parameter condition  $2\alpha > 3\beta^{3/2}\sqrt{n}$ . Consider Problem  $\text{OP}(f, C)$  (Example 4.5 in Ferreira and Sosa (2022)) with the associated function

$$f(x) := a^T x + \alpha x^T x + \frac{\beta}{\sqrt{1 + \beta x^T x}} e^T x,$$

with  $a \in \mathbb{R}_{++}^n$  is convex and the nonconvex constraint is given by

$$C := \{x \in \mathbb{R}_{++}^n : 1 \leq x_1 \dots x_n\}.$$

This example is implemented to compare Algorithm GDA with the original gradient descent algorithm (GD). We choose a random number  $\beta = 0.741271$ ,  $\alpha = 3\beta^{3/2}\sqrt{n+1}$  fulfilled the parameter condition and Lipschitz coefficient  $L = (4\beta^{3/2}\sqrt{n+1} + 3\alpha)$  suggested in Ferreira and Sosa (2022). The step size of Algorithm GD is  $\lambda = 1/L$ , and the initial step size of Algorithm GDA is  $\lambda_0 = 5/L$ . Table 1 shows the optimal value, number of loops, computational time of two algorithms through the different dimensions. From this result, Algorithm GDA is more efficient than GD at both the computational time and the optimal output value, especially for the large-scale dimensions.

**Example 4** To compare Algorithm GDA to Algorithm RNN in Liu et al. (2022), consider Problem  $\text{OP}(f, C)$  with the objective function

$$f(x) = -\exp\left(-\sum_{i=1}^n \frac{x_i^2}{\varrho_i^2}\right)$$

which is pseudoconvex on the convex constraint set

$$C := \{Ax = b, g(x) \leq 0\},$$

**Table 1** Computational results for Example 3

$n$	Algorithm GDA (proposed) $f(x^*)$	#Iter	Time	Algorithm GD $f(x^*)$	#Iter	Time
10	<b>79.3264</b>	9	<b>0.9576</b>	79.3264	15	1.5463
20	<b>220.5622</b>	10	<b>6.0961</b>	220.5622	67	34.0349
50	<b>857.1166</b>	12	<b>2.8783</b>	857.1166	16	4.6824
100	<b>2392.5706</b>	12	<b>17.2367</b>	2392.5706	17	30.8886
200	<b>7065.9134</b>	65	<b>525.1199</b>	7179.3542	200	1610.6560
500	<b>26,877.7067</b>	75	<b>2273.0011</b>	27,145.6292	500	14,113.5003

**Table 2** Computational results for Example 4

$n$	Algorithm GDA (proposed) $-\ln(-f(x^*))$	#Iter	Time	Algorithm RNN $-\ln(-f(x^*))$	#Iter	Time
10	<b>5.1200</b>	10	<b>0.3</b>	5.1506	1000	256
20	<b>2.5600</b>	10	<b>0.8</b>	2.5673	1000	503
50	<b>1.0240</b>	10	<b>2</b>	1.0299	1000	832
100	<b>0.5125</b>	10	<b>7</b>	13.7067	1000	1420
300	<b>15.7154</b>	10	<b>84</b>	39.3080	1000	3292
400	<b>20.9834</b>	10	<b>163</b>	57.6837	1000	4426
600	<b>29.0228</b>	10	<b>371</b>	83.6265	1000	6788

where  $x \in \mathbb{R}^n$ , the parameter vector  $\varrho = (\varrho_1, \varrho_2, \dots, \varrho_n)^\top$  with  $\varrho_i > 0$ , the matrix  $A = (a_1, a_2, \dots, a_n) \in \mathbb{R}^{1 \times n}$  with  $a_i = 1$  for  $1 \leq i \leq n/2$  and  $a_i = 3$  for  $n/2 < i \leq n$ , and the number  $b = 16$ . The inequality constraints are

$$g_i(x) = x_{10 \cdot (i-1)+1}^2 + x_{10 \cdot (i-1)+2}^2 + \cdots + x_{10 \cdot (i-1)+10}^2 - 20,$$

for  $i = 1, 2, \dots, n/10$ . Table 2 presents the computational results of Algorithms GDA and RNN. Note that the function  $-\ln(-z)$  is monotonically increasing by  $z \in \mathbb{R}, z < 0$ . Therefore, to compare the approximated optimal value through iterations, we compute  $-\ln(-f(x^*))$  instead of  $f(x^*)$ , with approximated optimal solution  $x^*$ . For each  $n$ , we solve Problem OP( $f, C$ ) to find the value  $-\ln(-f(x^*))$ , the number of iterations (#Iter) to terminate the algorithms, and the computation time (Time) by seconds. The computational results reveal that the proposed algorithm outperforms Algorithm RNN in the test scenarios for both optimum value and computational time, especially when the dimensions are large scale.

## 5 Applications to machine learning

The proposed method, like the GD algorithm, has many applications in machine learning. We analyze three common machine learning applications, namely supervised feature selection, regression, and classification, to demonstrate the accuracy and computational efficiency compared to other alternatives.

Firstly, the feature selection problem can be modeled as a minimization problem of a pseudoconvex fractional function on a convex set, which is a subclass of Problem  $\text{OP}(f, C)$ . This problem is used to compare the proposed approach to neurodynamic approaches. Secondly, since the multi-variable logistic regression problem is a convex programming problem, the GDA algorithm and available variants of the GD algorithm can be used to solve it. Lastly, a neural network model for an image classification problem is the same as a programming problem with neither a convex nor a quasi-convex objective function. For training this model, we use the stochastic variation of the GDA method (Algorithm SGDA) as a heuristic technique. Although the algorithm's convergence cannot be guaranteed like in the cases of pseudoconvex and quasiconvex objective functions, the theoretical study showed that if the sequence of points has a limit point, it converges into a stationary point of the problem (see Theorem 1). Computational experiments indicate that the proposed method outperforms existing neurodynamic and gradient descent methods.

## 5.1 Supervised feature selection

Feature selection problem is carried out on the dataset with  $p$ -feature set  $\mathcal{F} = \{F_1, \dots, F_p\}$  and  $n$ -sample set  $\{(x_i, y_i) \mid i = 1, \dots, n\}$ , where  $x_i = (x_{i1}, \dots, x_{ip})^T$  is a  $p$ -dimensional feature vector of the  $i$ th sample and  $y_i \in \{1, \dots, m\}$  represents the corresponding labels indicating classes or target values. In Wang et al. (2021), an optimal subset of  $k$  features  $\{F_1, \dots, F_k\} \subseteq \mathcal{F}$  is chosen with the least redundancy and the highest relevancy to the target class  $y$ . The feature redundancy is characterized by a positive semi-definite matrix  $Q$ . Then the first aim is minimizing the convex quadratic function  $w^T Q w$ . The feature relevance is measured by  $\rho^T w$ , where  $\rho = (\rho_1, \dots, \rho_p)^T$  is a relevancy parameter vector. Therefore, the second aim is maximizing the linear function  $\rho^T w$ . Combining two goals deduces the equivalent problem as follows:

$$\begin{aligned} & \text{minimize} \quad \frac{w^T Q w}{\rho^T w} \\ & \text{subject to} \quad e^T w = 1 \\ & \quad w \geq 0, \end{aligned} \tag{14}$$

where  $w = (w_1, \dots, w_p)^T$  is the feature score vector to be determined. Since the objective function of problem (14) is the fraction of a convex function over a positive linear function, it is pseudoconvex on the constraint set. Therefore, we can solve (14) by Algorithm GDA.

In the experiment, we implement the algorithms with the Parkinsons dataset, which has 23 features and 197 samples, downloaded at <https://archive.ics.uci.edu/ml/datasets/parkinsons>. The similarity coefficient matrix  $Q$  is determined by  $Q = \delta I_p + S$  (see Wang et al. 2021), where  $p \times p$ -matrix  $S = (s_{ij})$  with

$$s_{ij} = \max \left\{ 0, \frac{I(F_i; F_j; y)}{H(F_i) + H(F_j)} \right\},$$

the information entropy of a random variable vector  $\hat{X}$  is

$$H(\hat{X}) = - \sum_{\hat{x} \in \hat{X}} p(\hat{x}) \log p(\hat{x}),$$

the multi-information of three random vectors  $\hat{X}, \hat{Y}, \hat{Z}$  is  $I(\hat{X}; \hat{Y}; \hat{Z}) = I(\hat{X}; \hat{Y}) - I(\hat{X}, \hat{Y} | \hat{Z})$  with the mutual information of two random vectors  $\hat{X}, \hat{Y}$  defined by

$$I(\hat{X}; \hat{Y}) = \sum_{\hat{x} \in \hat{X}} \sum_{\hat{y} \in \hat{Y}} p(\hat{x}, \hat{y}) \log \frac{p(\hat{x}, \hat{y})}{p(\hat{x})p(\hat{y})}$$

and the conditional mutual information between  $\hat{X}, \hat{Y}$  and  $\hat{Z}$  defined by

$$I(\hat{X}; \hat{Y} | \hat{Z}) = \sum_{\hat{x} \in \hat{X}} \sum_{\hat{y} \in \hat{Y}} \sum_{\hat{z} \in \hat{Z}} p(\hat{x}, \hat{y}, \hat{z}) \log \frac{p(\hat{x}, \hat{y} | \hat{z})}{p(\hat{x} | \hat{z})p(\hat{y} | \hat{z})}.$$

The feature relevancy vector  $\rho = (\rho_1, \dots, \rho_p)^T$  is determined by Fisher score

$$\rho(F_i) = \frac{\sum_{j=1}^K n_j (\mu_{ij} - \mu_i)^2}{\sum_{j=1}^K n_j \sigma_{ij}^2},$$

where  $n_j$  denotes the number of samples in class  $j$ ,  $\mu_{ij}$  denotes the mean value of feature  $F_i$  for samples in class  $j$ ,  $\mu_i$  is the mean value of feature  $F_i$ , and  $\sigma_{ij}^2$  denotes the variance value of feature  $F_i$  for samples in class  $j$ .

The approximated optimal value of problem (14) is  $f(w^*) = 0.153711$  with the computing time  $T = 6.096260$ s for the proposed algorithm, while  $f(w^*) = 0.154013$ ,  $T = 11.030719$  for Algorithm RNN. In comparison to the Algorithm RNN, our algorithm outperforms it in terms of both accuracy and computational time.

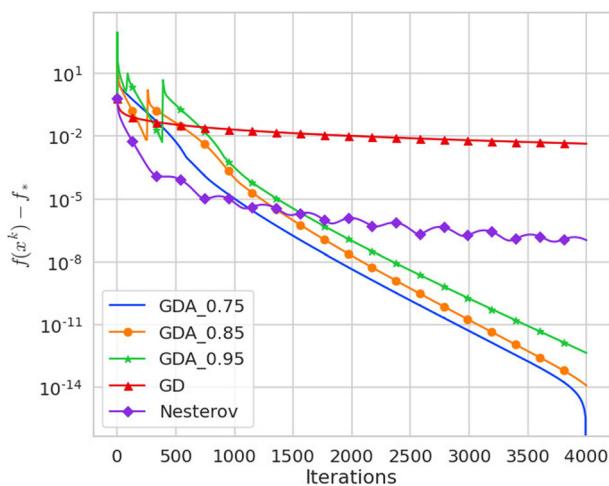
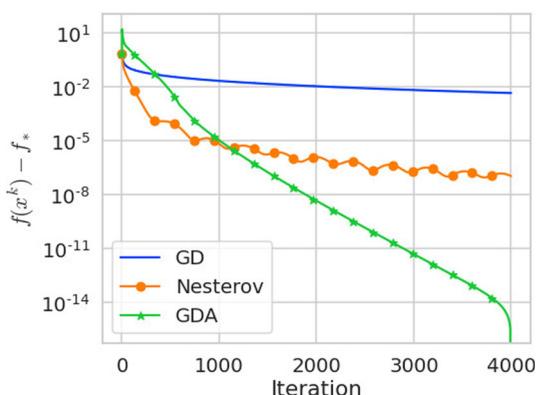
## 5.2 Multi-variable logistic regression

The experiments are performed with the dataset including  $\mathbf{N}$  observations  $(\mathbf{a}_i, b_i) \in \mathbb{R}^d \times \mathbb{R}$ ,  $i = 1, \dots, n$ . The cross-entropy loss function for multi-variable logistic regression is given by  $J(x) = -\sum_{i=1}^{\mathbf{N}} (b_i \log(\sigma(-x^T \mathbf{a}_i)) + (1 - b_i) \log(1 - \sigma(-x^T \mathbf{a}_i)))$ , where  $\sigma$  is the sigmoid function. Associated with  $\ell_2$ -regularization, we get the regularized loss function  $\bar{J}(x) = J(x) + \frac{1}{2\mathbf{N}} \|x\|^2$ . The Lipschitz coefficient  $L$  is estimated by  $\frac{1}{2\mathbf{N}}(\|A\|^2/2 + 1)$ , where  $A = (a_1^\top, \dots, a_n^\top)^\top$ . We compare the algorithms for training the logistic regression problem by using datasets Mushrooms and W8a (see Malitsky and Mishchenko 2020). The GDA method is compared to the GD algorithm with a step size of  $1/L$  and Nesterov's accelerated method. The computational results are shown in Figs. 3 and 4 respectively. The figures suggest that Algorithm GDA outperforms Algorithm GD and Nesterov's accelerated method in terms of objective function values during iterations. In particular, Fig. 4 shows the change of the objective function values according to different  $\kappa$  coefficients. In this figure, the notation "GDA\_0.75" respects to the case  $\kappa = 0.75$ . Figure 5 presents the reduction of step-sizes from an initial step-size with respect to the results in Fig. 4.

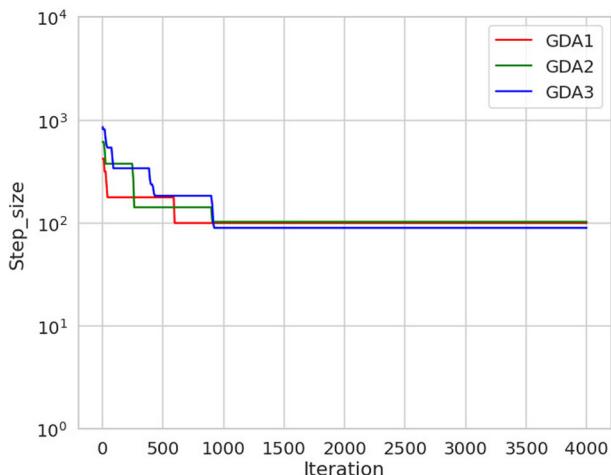
## 5.3 Neural networks for classification

In order to provide an example of how the proposed algorithm can be implemented into a neural network training model, we will use the standard ResNet-18 architectures that have been implemented in PyTorch and will train them to classify images taken from the Cifar10 dataset downloaded at <https://www.cs.toronto.edu/~kriz/cifar.html>, while taking into account

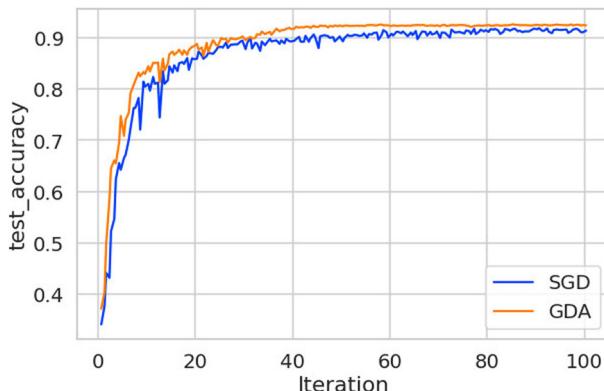
**Fig. 3** The computational results for logistic regression with dataset Mushrooms



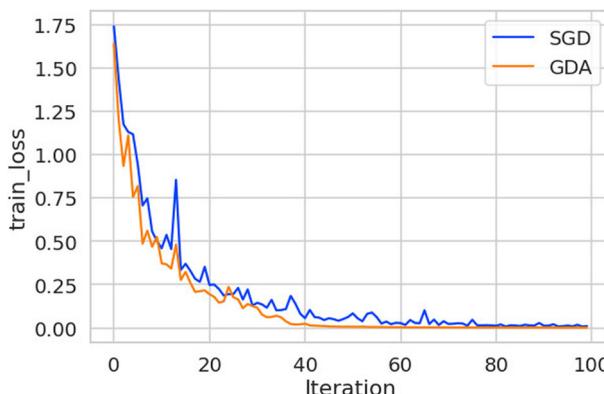
**Fig. 4** The computational results for logistic regression with dataset W8a



**Fig. 5** The step-sizes changing for each iteration in logistic regression with dataset W8a



**Fig. 6** The test accuracy through iterations for ResNet-18 model



**Fig. 7** The training loss through iterations for ResNet-18 model

the cross-entropy loss. In the studies with ResNet-18, we made use of Adam's default settings for its parameters.

For training this neural network model, we use the stochastic variation of the GDA method (Algorithm SGDA) to compare to Stochastic Gradient Descent (SGD) algorithms. The computational outcomes shown in Figs. 6 and 7 reveal that Algorithm SGDA outperforms Algorithm SGD in terms of testing accuracy and train loss over iterations.

## 6 Conclusion

We proposed a novel easy adaptive step-size process in a wide family of solution methods for optimization problems with non-convex objective functions. This approach does not need any line-searching or prior knowledge but rather takes into consideration the iteration sequence's behavior. As a result, as compared to descending line-search approaches, it significantly reduces the implementation cost of each iteration. We demonstrated technique convergence under simple assumptions. We demonstrated that this new process produces a generic foundation for optimization methods. The preliminary results of computer experiments demonstrated the new procedure's efficacy. This adaptive step-size descent algorithm

can be extended to a class of programming problems with nonsmooth objective functions or further extended to multi-objective optimization problems based on prior research (see Thang et al. 2020 and references therein).

**Acknowledgements** The authors thank the anonymous peer reviewers and the editor for their constructive comments which helped to improve the paper. This research is funded by Hanoi University of Science and Technology (HUST) under project number T2022-PC-061.

**Availability of data and materials** The data that support the findings of this study are available from the corresponding author, upon reasonable request.

## References

- Bauschke HH, Combettes PL (2011) Convex analysis and monotone operator theory in hilbert spaces. Springer Convex analysis and monotone operator theory in hilbert spaces
- Bian W, Ma L, Qin S, Xue X (2018) Neural network for nonsmooth pseudoconvex optimization with general convex constraints. *Neural Netw* 101:1–14
- Boyd SP, Vandenberghe L (2009) Convex optimization. Cambridge University Press, Cambridge
- Cevher V, Becker S, Schmidt M (2014) Convex optimization for big data. *Signal Process Mag* 31:32–4
- Dennis JE, Schnabel RB (1983) Numerical methods for unconstrained optimization and nonlinear equations. Prentice-Hall, New Jersey
- Ferreira OP, Sosa WS (2022) On the Frank-Wolfe algorithm for non-compact constrained optimization problems. *Optimization* 71(1):197–211
- Hu Y, Li J, Yu CK (2020) Convergence rates of subgradient methods for quasiconvex optimization problems. *Comput Optim Appl* 77:183–212
- Kiwiel KC (2001) Convergence and efficiency of subgradient methods for quasiconvex minimization. *Math Program Ser A* 90:1–25
- Konnov IV (2018) Simplified versions of the conditional gradient method. *Optimization* 67(12):2275–2290
- Lan GH (2020) First-order and stochastic optimization methods for machine learning. Springer series in the data sciences. Springer Nature
- Liu N, Wang J, Qin S (2022) A one-layer recurrent neural network for nonsmooth pseudoconvex optimization with quasiconvex inequality and affine equality constraints. *Neural Netw* 147:1–9
- Malitsky Y, Mishchenko K (2020) Adaptive gradient descent without descent. *Proc Mach Learn Res* 119:6702–6712
- Mangasarian O (1965) Pseudo-convex functions. *Siam Control* 8:281–290
- Nesterov Y (2013) Introductory lectures on convex optimization: a basic course, vol 87. Springer Science & Business Media
- Rockafellar RT (1970) Convex analysis. Princeton University Press, Princeton
- Thang TN, Solanki VK, Dao TA, Anh NTN, Hai PV (2020) A monotonic optimization approach for solving strictly quasiconvex multiobjective programming problems. *J Intell Fuzzy Syst* 38:6053–6063
- Wang Y, Li X, Wang J (2021) A neurodynamic optimization approach to supervised feature selection via fractional programming. *Neural Netw* 136:194–206
- Xu HK (2002) Iterative algorithms for nonlinear operators. *J Lond Math Soc* 66:240–256
- Yu CK, Hu Y, Yang X, Choy SK (2019) Abstract convergence theorem for quasi-convex optimization problems with applications. *Optimization* 68(7):1289–1304

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.