Journal of Global Optimization

On a smoothed penalty-based algorithm for global optimization

--Manuscript Draft--

Manuscript Number:	JOGO-D-16-00039R2
Full Title:	On a smoothed penalty-based algorithm for global optimization
Article Type:	S.I. : EURO-2015
Keywords:	Global optimization; Penalty function; Artificial fish swarm; Markov chains.
Corresponding Author:	Ana Maria A.C. Rocha, Ph.D. Algoritmi Research Centre, University of Minho Braga, PORTUGAL
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Algoritmi Research Centre, University of Minho
Corresponding Author's Secondary Institution:	
First Author:	Ana Maria A.C. Rocha, Ph.D.
First Author Secondary Information:	
Order of Authors:	Ana Maria A.C. Rocha, Ph.D.
	M. Fernanda P. Costa, Ph.D.
	Edite M.G.P. Fernandes, Aggr.
Order of Authors Secondary Information:	
Funding Information:	
Abstract:	This paper presents a coercive smoothed penalty framework for nonsmooth and nonconvex constrained global optimization problems. The properties of the smoothed penalty function are derived. Convergence to an ϵ-global minimizer is proved. At each iteration k, the framework requires the $\epsilon^{(k)}$-global minimizer of a subproblem, where $\epsilon^{(k)} \rightarrow \epsilon$. We show that the subproblem may be solved by well-known stochastic metaheuristics, as well as by the artificial fish swarm (AFS) algorithm. In the limit, the AFS algorithm convergence to an $\epsilon^{(k)}$-global minimum of the real-valued smoothed penalty function is guaranteed with probability one, using the limiting behavior of Markov chains. In this context, we show that the transition probability of the Markov chain produced by the AFS algorithm, when generating a population where the best fitness is in the $\epsilon^{(k)}$-neighborhood of the global minimum, is one when this property holds in the current population, and is strictly bounded from zero when the property does not hold. Preliminary numerical experiments show that the presented penalty algorithm based on the coercive smoothed penalty gives very competitive results when compared with other penalty-based methods.
Response to Reviewers:	Dear Editor All the issues have been addressed. Best regards, The authors

Journal of Global Optimization manuscript No.
(will be inserted by the editor)

On a smoothed penalty-based algorithm for global optimization

Ana Maria A.C. Rocha · **M. Fernanda P. Costa** · **Edite M.G.P. Fernandes**

Received: date / Accepted: date

Abstract This paper presents a coercive smoothed penalty framework for nonsmooth and nonconvex constrained global optimization problems. The properties of the smoothed penalty function are derived. Convergence to an ε-global minimizer is proved. At each iteration k, the framework requires the $\varepsilon^{(k)}$-global minimizer of a subproblem, where $\varepsilon^{(k)} \to \varepsilon$. We show that the subproblem may be solved by well-known stochastic metaheuristics, as well as by the artificial fish swarm (AFS) algorithm. In the limit, the AFS algorithm convergence to an $\varepsilon^{(k)}$-global minimum of the real-valued smoothed penalty function is guaranteed with probability one, using the limiting behavior of Markov chains. In this context, we show that the transition probability of the Markov chain produced by the AFS algorithm, when generating a population where the best fitness is in the $\varepsilon^{(k)}$-neighborhood of the global minimum, is one when this property holds in the current population, and is strictly bounded from zero when the property does not hold. Preliminary numerical experiments show that the presented penalty algorithm based on the coercive smoothed penalty gives very competitive results when compared with other penalty-based methods.

Keywords Global optimization · Penalty function · Artificial fish swarm · Markov chains

Ana Maria A.C. Rocha
Algoritmi Research Centre,
Department of Production and Systems,
University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal
E-mail: arocha@dps.uminho.pt

M. Fernanda P. Costa
Centre of Mathematics,
Department of Mathematics and Applications,
University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal
E-mail: mfc@math.uminho.pt

Edite M.G.P. Fernandes
Algoritmi Research Centre
E-mail: emgpf@dps.uminho.pt

1 Introduction

This paper aims to contribute to the research area concerned with smoothed penalty function methods in constrained global optimization (CGO), by using well-established metaheuristics to compute a sequence of approximations to the solution of the CGO problem. The problem to be addressed is of the form:

$$\begin{array}{rl} \min\limits_{x \in \Omega} & f(x) \\ \text{subject to} & g_i(x) \leq 0, i = 1, \ldots, p \end{array} \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, p$ are nonlinear continuous functions, possibly nondifferentiable, and $\Omega = \{x \in \mathbb{R}^n : -\infty < l \leq x \leq u < \infty\}$. When the problem has equality constraints, $h(x) = 0$, they may be reformulated into the above form by using a couple of inequality constraints $h(x) - \upsilon \leq 0$ and $-h(x) - \upsilon \leq 0$, where υ is a small positive relaxation parameter. Functions f and g may be nonconvex and many local minima may exist in the feasible region.

When solving global optimization problems, penalty-type methods combined with stochastic techniques have been appearing in the literature [1–7]. They are simple to implement and provide in general high quality solutions. Penalty functions require in general the use of a positive penalty parameter that aims to balance function and constraint violation values. Setting the initial value for the penalty parameter and tuning its values throughout the iterative process are not easy tasks. Furthermore, the performance of the algorithm is greatly affected by their values. In some cases, the optimal solution of the problem is attained only when the penalty parameter approaches infinity. This is the case with the quadratic penalty [8]. Popular penalties when combined with stochastic heuristics are the dynamic penalty, in which the parameter depends continuously on the iteration counter [1,3,5,6], and the adaptive penalty function. These make the performance of the algorithm less sensitive to penalty variations [2,9,7]. Penalties that rely on two parameters have appeared in the literature. A two-parameter hyperbolic penalty is presented and the convergence properties are discussed in [10]. A family of non-coercive smoothed penalty functions and an algorithm for nonlinear constrained optimization are presented in [11]. The convergence properties are also shown. This last work is based on [12] where the therein studied penalty functions can be rewritten like those in [11]. Exact penalty functions are very efficient since every global minimizer of the penalty function is a global solution of the related constrained problem, and conversely, for some finite values of the penalty. In [13], an improved version of the well-known global deterministic DIRECT algorithm, tailored for global optimization with simple bounds, is incorporated into an exact penalty approach. It is proved that every accumulation point of a sequence of iterates produced by the algorithm is a global minimizer of the constrained problem. Furthermore, under a weak assumption, it is proved that the penalty parameter is updated a finite number of times.

In this paper, we aim to further explore penalty-based approaches for solving CGO problems. The contribution of the present study is a coercive smoothed penalty framework for nonsmooth and nonconvex CGO problems. The penalty added to the objective function is a smoothed penalty function and depends on two parameters, one is a penalty weight for constraint violation and the other is a smoothing parameter. Further, convergence of a sequence of iterates to an ε-global minimizer is proved. At each iteration k, the penalty framework requires an $\varepsilon^{(k)}$-global minimizer of a bound constrained optimization subproblem, where $\varepsilon^{(k)} \to \varepsilon$. The subproblems are solved by well-known stochastic metaheuristics. One particular metaheuristic – the artificial fish swarm (AFS) algorithm [14] – is further analyzed and its convergence to an $\varepsilon^{(k)}$-global minimum of the real-valued smoothed penalty function is guaranteed with probability one, using the limiting behavior of Markov chains. We also show that the proposed smoothed penalty framework, when a selected set of well-known metaheuristics (including the AFS algorithm) are used to solve the subproblems, is effective in finding global optimal solutions to the CGO problem.

Metaheuristics are general-purpose, usually stochastic and population-based, and define a set of solutions using nature-inspired operations. They are important and effective strategies to solve hard optimization problems. This type of problems cannot be solved to optimality by any exact method within a reasonable time limit [15]. Metaheuristics are specially tailored for black-box optimization problems, where there is no explicit form for the involved functions. Objective and constraint functions are evaluated through physical or computational simulation, thus preventing the use of automatic differentiation techniques to obtain the derivatives. Numerical differentiation is also avoided due to high cost of function evaluations. Challenging problems of this class are optimal control problems, simulation-based optimization and parameter estimation over differential equations, where a function model has to be obtained by sampling or by a third party software library. In general, metaheuristics are simple to implement and use, do not require transformation of the original problem, perform quite well and generate good quality solutions in less time than the traditional optimization techniques. However, when derivative information is available, gradient-based algorithms are substantially superior in terms of convergence speed since fast global convergence may be guaranteed. With stochastic metaheuristics only stochastic convergence can be ensured. A state-of-art concerning metaheuristics in optimization is presented in [16]. Interesting and challenging issues related to the use of metaheuristics are addressed in [17].

The paper is organized as follows. In Section 2, the new smoothed penalty function is presented, and the penalty-based algorithm with its convergence properties are derived. In Section 3, we refer to the metaheuristics that have been selected to solve the bound constrained optimization subproblem and discuss the asymptotic convergence properties of the AFS algorithm, Section 4 presents some numerical experiments and we conclude the paper in Section 5.

2 Penalty Method

The feasible set of the problem (1) is defined as

$$\mathscr{F} = \{x \in \Omega \subset \mathbb{R}^n : g_i(x) \leq 0, i = 1,\ldots,p\}. \tag{2}$$

In this paper, we aim at converging to a global solution of the nonlinear optimization problem (1) using a penalty framework. In [11], a smoothed penalty function that depends on a smooth parameter $\mu \in (0,1]$, and is defined as $P_\mu(t) = \mu P\left(\frac{t}{\mu}\right)$, where

$$P(t) = \log(1+\exp(t)) \tag{3}$$

is used, in the sense that a differentiable function is used by smoothing the function $t^+ = \max\{0,t\}$ (see also [12,18]). Thus, in [11], the penalty term associated with the constraint $g_i(x) \leq 0$ has the form

$$P_\mu(g_i(x)) = \mu \log\left(1+\exp\left(\frac{g_i(x)}{\mu}\right)\right) \tag{4}$$

and aims to penalize constraints violation of the problem (1). Thus, at each iteration k, where k denotes the iteration counter of the outer cycle, and for fixed values of the smoothing parameter, $\mu^{(k)} \in (0,1]$, and of a penalty weight, $\tau^{(k)} \geq 1$, the subproblem takes the form:

$$\min_{x \in \Omega} \phi(x;\tau^{(k)},\mu^{(k)}) \equiv f(x) + \tau^{(k)}\mu^{(k)} \sum_{i=1}^{p} P\left(\frac{g_i(x)}{\mu^{(k)}}\right). \tag{5}$$

Hence, this penalty method consists in solving a family of bound constrained optimization subproblems of the form (5) where $\mu^{(k)} \to 0$ and the function P satisfies certain properties. It is proved in [12] that a path of optimal solutions $\{x_{\mu^{(k)}}\}_{\mu^{(k)}>0}$ defined by $x_{\mu^{(k)}} \equiv \arg\min_{x\in\Omega} \phi(x;\tau^{(k)},\mu^{(k)})$ converges towards the optimal set of the original convex constrained optimization problem (likewise (1)) when $\mu^{(k)} \to 0$. We note that the parameter μ aims to control the precision of the smoothing and τ aims to change the inclination of $\tau P_\mu(t)$. Thus, the method used for solving the subproblems should ensure that the bound constraints are always satisfied and global optimal solutions are obtained. Some interesting differences between penalty algorithms are located on the framework used to find an approximate solution to the subproblem (5). Our work is inspired by the smoothed penalty functions studied in [11]. However, we consider nonconvex optimization problems and we do not assume that the objective f and the constraints $g_i, i = 1,\ldots,p$ are differentiable. Thus, we do not assume that the Mangasarian-Fromovitz condition holds. We use the following definition:

Definition 1 ($\varepsilon^{(k)}$-global minimizer) Let $\phi^k(x) \equiv \phi(x;\tau^{(k)},\mu^{(k)})$ be a continuous objective function defined over a bounded space $\Omega \subset \mathbb{R}^n$. The point $x^{(k)} \in \Omega$ is an $\varepsilon^{(k)}$-global minimizer of the subproblem (5) if $\phi^k(x^{(k)}) \leq \min_{y\in\Omega} \phi^k(y) + \varepsilon^{(k)}$, where $\varepsilon^{(k)} > 0$ is the error bound which reflects the accuracy required for the approximation.

Our penalty-based algorithm requires that at each iteration k of the outer cycle, an $\varepsilon^{(k)}$-global solution of the subproblem (5) is obtained. When the optimization problem is nonconvex, a global optimization method is required to approximately solve the subproblem (5), so that the algorithm has some guarantee to converge to a global solution instead of being trapped in a local one. To compute an $\varepsilon^{(k)}$-global minimizer of the subproblem (5), for fixed values of $\tau^{(k)}$ and $\mu^{(k)}$, we propose the use of well-known stochastic metaheuristics.

2.1 Smoothed penalty function

We now present a new smoothed penalty term for the subproblem described in (5):

$$P_\mu(g_i(x)) \equiv \mu P\left(\frac{g_i(x)}{\mu}\right) = \mu\left(1 + \theta\left(\frac{g_i(x)}{\mu}\right)\right) \tag{6}$$

where the function $\theta(t)$ is defined, for $t \in \mathbb{R}$, by

$$\theta(t) = \begin{cases} \tanh(t), & \text{if } t \leq 0, \\ \sinh(t), & \text{otherwise,} \end{cases} \quad \text{being } P(t) = 1 + \theta(t). \tag{7}$$

Figure 1 displays plots of $P(t)$, $P'(t) = \theta'(t)$ as well as of the penalty term $P(t)$ defined in (3), and its first derivative, for comparison.

As previously defined, we have $P_\mu(t) = \mu P\left(\frac{t}{\mu}\right)$ where $P : \mathbb{R} \to \mathbb{R}^+$. For any $\mu \in (0,1]$ and $t \in \mathbb{R}$, the proposed $P_\mu(t)$ defined by (6) with $\theta(t)$ as in (7) satisfy the following properties.

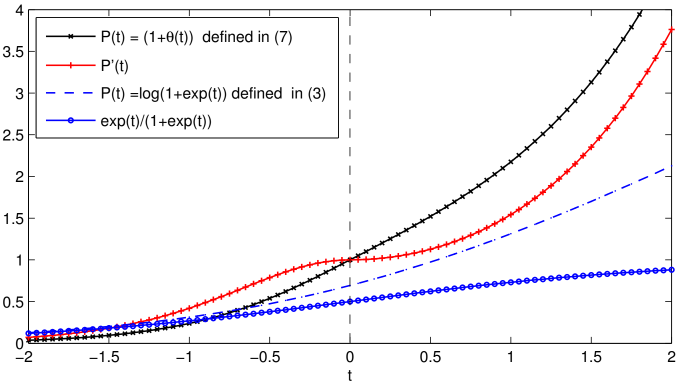


Fig. 1 Plots of $P(t)$ in (3) and (7), and first derivatives

Property 1 $P_\mu(t)$ is convex, continuous and differentiable as a function of t, and $P'_\mu(0) > 0$.

Proof Function $\theta(t)$ is convex, $P(t)$ is convex and since $\mu \in (0,1]$, $P_\mu(t)$ is also convex. Similarly $P_\mu(t)$ is differentiable since $\theta(t)$ is differentiable.

Since $P'_\mu(t) = \mu\theta'\left(\frac{t}{\mu}\right)$,

$$P'_\mu(t) = \begin{cases} 1/\cosh^2\left(\frac{t}{\mu}\right), & \text{if } t \leq 0, \\ \cosh\left(\frac{t}{\mu}\right), & \text{otherwise,} \end{cases} \tag{8}$$

and we get $P'_\mu(0) = 1$. Furthermore, $P'_\mu(0^+) = 1$. □

Property 2 $\lim_{t\to-\infty} P(t) = 0$, $\lim_{t\to-\infty} P_\mu(t) = 0$ and $\lim_{t\to 0^-} P_\mu(t) = \mu$.

Proof

$$\lim_{t\to-\infty} P(t) = \lim_{t\to-\infty} (1+\tanh(t)) = 0 \text{ and } \lim_{t\to-\infty} P_\mu(t) = \lim_{t\to-\infty} \mu\left(1+\tanh\left(\frac{t}{\mu}\right)\right) = 0.$$

Furthermore,

$$\lim_{t\to 0^-} P_\mu(t) = \lim_{t\to 0^-} \mu\left(1+\tanh\left(\frac{t}{\mu}\right)\right) = \mu.$$

□

Property 3 $\lim_{t\to\infty} \dfrac{P_\mu(t)}{t} = +\infty$ and $\lim_{t\to\infty} \dfrac{P_\mu(-t)}{t} = 0$.

Proof Recalling the definition of θ in (7),

$$\begin{aligned} \lim_{t\to\infty} \frac{P_\mu(t)}{t} &= \lim_{t\to\infty} \frac{\mu}{t}\left(1+\sinh\left(\tfrac{t}{\mu}\right)\right) = \lim_{t\to\infty} \cosh\left(\tfrac{t}{\mu}\right) \\ &= \nu > 0 \end{aligned}$$

where $\nu = +\infty$, defining a coercive penalty. To prove the second part of the property, we use (7) and:

$$\lim_{t\to\infty} \frac{P_\mu(-t)}{t} = \lim_{t\to\infty} \frac{\mu}{t}\left(1+\tanh\left(\tfrac{-t}{\mu}\right)\right) = 0.$$

□

Property 4 For $\mu \in (0,1]$, $P'_\mu(t) \le P'(t)$ when $t \le 0$ and $P'_\mu(t) \ge P'(t)$ when $t > 0$.

Proof For $\mu \in (0,1]$, we have $\frac{t}{\mu} \le t$ when $t \le 0$, and $\frac{t}{\mu} \ge t$ when $t > 0$. From (8), we have (recalling that the hyperbolic cosine is strictly decreasing for negative arguments and strictly increasing for positive arguments) for $t \le 0$:

$$P'_\mu(t) = 1/\cosh^2\left(\frac{t}{\mu}\right) \le 1/\cosh^2(t) = P'(t)$$

and for $t > 0$

$$P'_\mu(t) = \cosh\left(\frac{t}{\mu}\right) \ge \cosh(t) = P'(t).$$

□

Figure 2 displays the plots of $P_\mu(t)$, $P(t)$ and $P(t) - P(0)$ as well as their first derivatives, to check Properties 4 and 5.

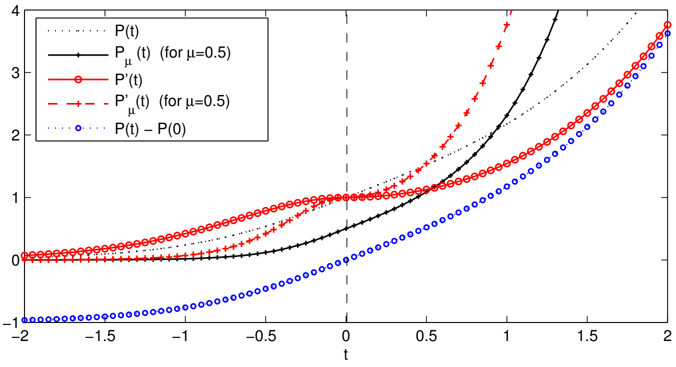


Fig. 2 Plots of $P_\mu(t)$, $P(t) - P(0)$, $P'_\mu(t)$ and $P'(t)$

Property 5 For $t \in \mathbb{R}$ and $\mu \in (0,1]$, $P_\mu(t) = \mu P\left(\frac{t}{\mu}\right) > P(t) - P(0)$.

Proof We note that $P(0) = 1$ and $P(t) - P(0) = \theta(t)$. Consider $\mu \in (0,1]$. When $t \le 0$, $\frac{t}{\mu} \le t$; on the other hand, when $t > 0$, $\frac{t}{\mu} \ge t$. First, we consider the case $t \le 0$, where

$$1 + \tanh\left(\frac{t}{\mu}\right) > 0 \ge \tanh(t)$$

(recalling that $-1 < \tanh(t) \le 0$ for $t \le 0$) and thus

$$\mu\left(1 + \tanh\left(\frac{t}{\mu}\right)\right) > \tanh(t).$$

Now, to prove that $P_\mu(t) > P(t) - P(0) = \theta(t)$ is true for $t > 0$, we use $P'_\mu(t) \ge P'(t)$ (when $t > 0$) from Property 4, and get

$$\int_0^t \cosh\left(\frac{x}{\mu}\right) dx \ge \int_0^t \cosh(x)\, dx$$

which implies

$$\mu \sinh\left(\frac{t}{\mu}\right) \geq \sinh(t)$$

and therefore

$$\mu \sinh\left(\frac{t}{\mu}\right) + \mu > \sinh(t).$$

□

Property 6 For $t \in \mathbb{R}$, $P_\mu(t)$ converges pointwise to $\nu t^+ = \nu \max\{0, t\}$ when $\mu \to 0^+$ where $\nu = +\infty$ (using the convention that $\infty \times 0 = 0$).

Proof First, we consider the case $t \leq 0$, where

$$\lim_{\mu\to 0} P_\mu(t) = \lim_{\mu\to 0} \mu \left(1 + \tanh\left(\frac{t}{\mu}\right)\right) = 0 = \nu t^+.$$

However, when $t > 0$, using the change of variable $z = \frac{t}{\mu}$, and Property 3 we get

$$\begin{aligned}
\lim_{\mu\to 0} P_\mu(t) = \lim_{\mu\to 0} \mu \left(1 + \sinh\left(\tfrac{t}{\mu}\right)\right) &= \lim_{z\to\infty} \frac{t}{z}\left(1 + \sinh(z)\right) \\
&= \lim_{z\to\infty} \frac{t}{z} + t \lim_{z\to\infty} \frac{\sinh(z)}{z} \\
&= t \lim_{z\to\infty} \cosh(z) \\
&= \underbrace{t\nu}_{\text{with } \nu=+\infty} = \nu t^+
\end{aligned}$$

□

Figure 3 shows the plots of $P_\mu(t)$ for values of $\mu = 1, 0.5, 0.25$ and 0.125, as well as the plots of penalty (3) (identified with $P_{\log}$ in the legend) for $\mu = 1$ and 0.125.

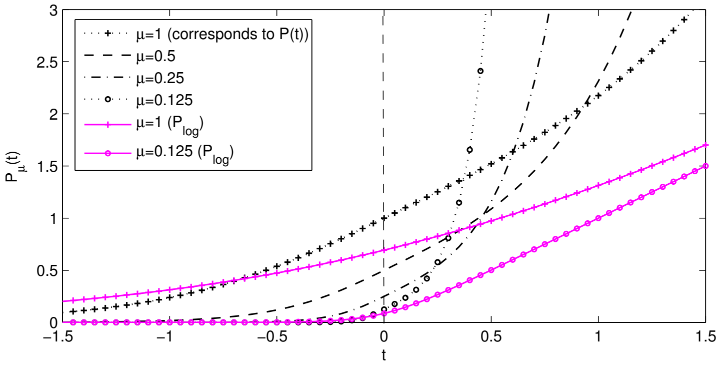


Fig. 3 Behavior of $P_\mu(t)$ for four different values of μ

We now define the penalty term in (5) as

$$p_\mu(x) = \sum_{i=1}^{p} P_\mu(g_i(x)) = \mu \sum_{i=1}^{p} P\left(\frac{g_i(x)}{\mu}\right) \tag{9}$$

and if the algorithm never increases the value of the product $\tau\mu$, we have:

Property 7 If $\tau\mu$ is bounded, then $\lim_{\mu\to 0^+} \tau p_\mu(x) = 0$, for any x in the interior of $\mathscr{F}$.

Proof For any x in the interior of $\mathscr{F}$, $g_i(x) < 0$, $i = 1,\ldots,p$ and

$$\lim_{\mu\to 0^+} \tau p_\mu(x) = \lim_{\mu\to 0^+} \tau\mu \sum_{i=1}^{p} \left(1 + \theta\left(\frac{g_i(x)}{\mu}\right)\right) = \lim_{\mu\to 0^+} \tau\mu \sum_{i=1}^{p} P\left(\frac{g_i(x)}{\mu}\right) = 0$$

by Property 2 and boundedness of $\tau\mu$. □

Property 8 If the sequence $\{\tau\mu\} \to 0$, then $\lim_{\mu\to 0^+} \tau p_\mu(x) = 0$, for any $x \in \mathscr{F}$.

Proof For any $x \in \mathscr{F}$:

$$\lim_{\mu\to 0^+} \tau p_\mu(x) = \lim_{\mu\to 0^+} \tau\mu \sum_{i=1}^{p} \left(1 + \theta\left(\frac{g_i(x)}{\mu}\right)\right) = \lim_{\mu\to 0^+} \tau\mu \sum_{i=1}^{p} P\left(\frac{g_i(x)}{\mu}\right) = 0,$$

since $\{\tau\mu\} \to 0$ and $P\left(\frac{g_i(x)}{\mu}\right)$ is bounded when $g_i(x) \le 0$. □

2.2 The penalty-based algorithm

A formal description of the penalty algorithm for solving the original problem (1) is presented in Algorithm 2.1. We remark the following:

Remark 1 When finding the global minimum of a continuous objective function $f(x)$ over a bounded space $\mathscr{F} \subset \mathbb{R}^n$, the point $\bar{x} \in \mathscr{F}$ is an ε-global minimizer if $f(\bar{x}) \le \min_{y\in\mathscr{F}} f(y) + \varepsilon$, where ε is the error bound which reflects the accuracy required for the solution.

We analyze feasibility at iteration k using the vector g^+, componentwise defined by

$$g_i^+(x^{(k)}) = \max\left\{0, g_i(x^{(k)})\right\}, i = 1,\ldots,p. \tag{10}$$

Algorithm 2.1 Penalty algorithm

```
Require: $\tau^{(1)} \ge 1, 0 < \mu^{(1)} \le 1, \gamma_\tau > 1, 0 < \gamma_\mu < (\gamma_\tau)^{-1}, \gamma_g < 1, \gamma_\varepsilon < 1, 0 < \varepsilon_{\min} \ll 1, \eta_{\min} \ll 1, \varepsilon^{(1)} > \varepsilon_{\min}$,
    $k_{\max}, LB$
Set $k = 1$, randomly generate a set of solutions in $\Omega$ and select $x^{(0)}$
while $\left(\|g^+(x^{(k-1)})\| > \eta_{\min}$ or $f(x^{(k-1)}) > LB + \varepsilon_{\min}\right)$ and $k \le k_{\max}$ do
    Find an $\varepsilon^{(k)}$-global minimizer of subproblem (5), $x^{(k)}$, so that

        $\phi(x^{(k)};\tau^{(k)},\mu^{(k)}) \le \phi(x;\tau^{(k)},\mu^{(k)}) + \varepsilon^{(k)}$ for all $x \in \Omega$    (11)

    Set $\mu^{(k+1)} = \gamma_\mu \mu^{(k)}$
    if $\|g^+(x^{(k)})\| \le \gamma_g \|g^+(x^{(k-1)})\|$ then
        Set $\tau^{(k+1)} = \tau^{(k)}$, $\varepsilon^{(k+1)} = \max\left\{\varepsilon_{\min}, \gamma_\varepsilon \varepsilon^{(k)}\right\}$
    else
        Set $\tau^{(k+1)} = \gamma_\tau \tau^{(k)}$, $\varepsilon^{(k+1)} = \varepsilon^{(k)}$
    end if
    Set $k = k + 1$
end while
```

We note that the penalty parameter μ decreases at all iterations and parameter τ is not updated if some progress in the feasibility is observed, summarized by the condition

$$\|g^+(x^{(k)})\| \leq \gamma_g \|g^+(x^{(k-1)})\|, \text{ for } 0 < \gamma_g < 1. \tag{12}$$

This condition says that the level of constraint violation at the point $x^{(k)}$ is lower than that of the previous approximation, and for this reason the penalty value τ may remain constant. Constants $\gamma_\tau > 1$ and $0 < \gamma_\mu < 1$ aim to increase and decrease the penalties $\tau^{(k)}$ and $\mu^{(k)}$ respectively, throughout the iterative process. We note that

(i) choosing $\gamma_\tau = (\gamma_\mu)^{-1}$ in the algorithm, and setting $\tau^{(1)} = 1$ and $\mu^{(1)} = 1$ then $\tau^{(k)}\mu^{(k)} \leq 1$ for all k and
$$\lim_k \tau^{(k)} p_{\mu^{(k)}}(\tilde{x}) = 0$$
for any interior point $\tilde{x}$. In [11] $\gamma_\tau = 2$ and $\gamma_\mu = 0.5$ are proposed for a non-coercive penalty;

(ii) choosing $\gamma_\mu < (\gamma_\tau)^{-1}$ in the algorithm, $\{\tau^{(k)}\mu^{(k)}\}$ is a bounded monotonically decreasing sequence that converges to zero.

Algorithm 2.1 terminates when a solution $x^{(k)}$ that is feasible and has an objective function value within $\varepsilon_{\min}$ of the known minimum is found, i.e., when $\|g^+(x^{(k)})\| \leq \eta_{\min}$ and $f(x^{(k)}) \leq LB + \varepsilon_{\min}$, where LB is a known lower bound to the optimal value of the problem, for sufficiently small positive tolerances $\eta_{\min}$ and $\varepsilon_{\min} > 0$.

The main properties of our Algorithm 2.1 are the following:

Property A1 $\mu^{(k)} \to 0$ (see Step 4 in the algorithm).

Property A2 $\tau^{(k)}\mu^{(k)} \leq 1$ and $\gamma_\mu < (\gamma_\tau)^{-1}$ (see initial values and Steps 4, 6 and 8 in the algorithm).

Property A3 if $\{x^{(k)}\}$ has an infeasible infinite subsequence then $\tau^{(k)} \to +\infty$ (a consequence of Step 8).

Properties A1 and A2 aim to ensure that $P_{\mu^{(k)}}(t) \to \nu t^+$ (see Property 6) and $\tau^{(k)} p_{\mu^{(k)}}(x) \to 0$ for any $x \in \mathscr{F}$ (see Properties 7 and 8), respectively.

2.3 Convergence of the penalty algorithm

We assume that $\{x^{(k)}\}$, $\{\tau^{(k)}\}$ and $\{\mu^{(k)}\}$ are sequences generated by the Algorithm 2.1 and that there exists a subset of indices $\mathscr{N} \subseteq \mathbb{N}$ so that $\lim_{k\in\mathscr{N}} x^{(k)} = x^*$. Here we aim to prove that every accumulation point, x^*, of the sequence $\{x^{(k)}\}$ produced by Algorithm 2.1 is an ε-global minimizer of problem (1). Since the set Ω is compact and the penalty function $\phi(x; \tau^{(k)}, \mu^{(k)})$ is continuous, the $\varepsilon^{(k)}$-global minimizer of subproblem (5), $x^{(k)}$, does exist. We also assume that it is possible to find an $\varepsilon^{(k)}$-global minimizer of subproblem (5) using well-known and recent stochastic metaheuristics, and in particular the AFS algorithm, whose stochastic convergence results are shown in the next section. Now, we follow the methodology presented in [19], where differentiability is not required. We investigate the properties of accumulation points of the sequence $\{x^{(k)}\}$.

Theorem 1 *Every accumulation point x^* of the sequence $\{x^{(k)}\}$ generated by the Algorithm 2.1 is feasible for the problem* (1).

Proof Since $x^{(k)} \in \Omega$ and Ω is closed then $x^* \in \Omega$. We now consider two cases: (a) $\{\tau^{(k)}\}$ is bounded; (b) $\tau^{(k)} \to \infty$.

In case (a), there exists an index K and a value $\bar{\tau} > 0$ such that $\tau^{(k)} = \bar{\tau}$ for all $k \geq K$. This means that for all $k \geq K$, condition (12) holds, which implies that $g_i^+(x^{(k)}) \to 0$ for all i. Thus $g_i(x^*) \leq 0$ for all i, and the accumulation point is feasible.

The proof in case (b) is by contradiction assuming that x^* is not feasible and that a global minimizer z exists in Ω (the same for all k) such that $g_i(z) \leq 0$ for all i. Thus there exists at least one i such that $g_i(x^*) > 0 \geq g_i(z)$. Without loss of generality we assume that $g_i(x^*) \leq g_i(z)$ for just one i, say $i = 1$, and $g_i(x^*) > 0 \geq g_i(z)$ for all $i \neq 1$. Then, we have

$$\sum_{i=2}^{p} P_{\mu^{(k)}}(g_i(x^*)) > \sum_{i=2}^{p} P_{\mu^{(k)}}(g_i(z))$$

and

$$0 < P_{\mu^{(k)}}(g_1(x^*)) \leq P_{\mu^{(k)}}(g_1(z))$$

or, for a large enough $k \in \mathscr{N}$, continuity implies that there exists a positive constant c such that

$$\tau^{(k)} \sum_{i=2}^{p} P_{\mu^{(k)}}(g_i(x^{(k)})) > \tau^{(k)} \sum_{i=2}^{p} P_{\mu^{(k)}}(g_i(z)) + \tau^{(k)} c$$

and therefore

$$\tau^{(k)} \sum_{i=1}^{p} P_{\mu^{(k)}}(g_i(x^{(k)})) > \tau^{(k)} \left(\sum_{i=1}^{p} P_{\mu^{(k)}}(g_i(z)) + P_{\mu^{(k)}}(g_1(x^{(k)})) - P_{\mu^{(k)}}(g_1(z)) \right) + \tau^{(k)} c.$$

Now, using Property 2, we get

$$\tau^{(k)} \sum_{i=1}^{p} P_{\mu^{(k)}}(g_i(x^{(k)})) > \tau^{(k)} \left(\sum_{i=1}^{p} P_{\mu^{(k)}}(g_i(z)) - \mu^{(k)} \right) + \tau^{(k)} c$$

or

$$\begin{aligned} f(x^{(k)}) + \tau^{(k)} \sum_{i=1}^{p} P_{\mu^{(k)}}(g_i(x^{(k)})) > f(z) + \tau^{(k)} \sum_{i=1}^{p} P_{\mu^{(k)}}(g_i(z)) + f(x^{(k)}) - f(z) \\ + \tau^{(k)}(c - \mu^{(k)}). \end{aligned}$$

Thus, for large enough k, $\tau^{(k)} \to \infty$, $\tau^{(k)}\mu^{(k)} \to 0$ and $f(x^{(k)}) - f(z) + \tau^{(k)} c > \varepsilon^{(k)} > 0$ which implies $\phi(x^{(k)}; \tau^{(k)}, \mu^{(k)}) > \phi(z; \tau^{(k)}, \mu^{(k)}) + \varepsilon^{(k)}$ which contradicts the definition of $x^{(k)}$ in (11). □

We now prove that a sequence of iterates generated by the algorithm converges to an ε-global minimizer of problem (1).

Theorem 2 *Every accumulation point x^* of a sequence $\{x^{(k)}\}$ generated by Algorithm 2.1 is an ε-global minimizer of the problem* (1).

Proof We consider the cases: (a) $\{\tau^{(k)}\}$ is bounded; (b) $\tau^{(k)} \to \infty$. First, the case (a). By the definition of $x^{(k)}$ in (11), and since $\tau^{(k)} = \bar{\tau}$ for all $k \geq K$, we have

$$f(x^{(k)}) + \bar{\tau} \sum_{i=1}^{p} P_{\mu^{(k)}}(g_i(x^{(k)})) \leq f(z) + \bar{\tau} \sum_{i=1}^{p} P_{\mu^{(k)}}(g_i(z)) + \varepsilon^{(k)} \tag{13}$$

where $z \in \Omega$ is a global minimizer of problem (1). Since $g_i(z) \leq 0$ for all i, and $\mu^{(k)}, \bar{\tau} > 0$, for all $k \geq K$ we have

$$f(x^{(k)}) + \bar{\tau} \sum_{i=1}^{p} P_{\mu^{(k)}}(g_i(x^{(k)})) \leq f(z) + \bar{\tau} p \mu^{(k)} + \varepsilon^{(k)}$$

and taking limits for $k \in \mathscr{N}$ and using $\lim_{k \in \mathscr{N}} \varepsilon^{(k)} = \varepsilon$, we obtain

$$f(x^*) \leq f(x^*) + \bar{\tau} \sum_{i=1}^{p} P_{\mu^{(k)}}(g_i(x^*)) \leq f(z) + \bar{\tau} p \mu^{(k)} + \varepsilon$$

since by Theorem 1 x^* is feasible, and using $\lim_{k \in \mathbb{N}} \mu^{(k)} = 0$, we finally get

$$f(x^*) \leq f(z) + \varepsilon$$

proving that x^* is an ε-global minimizer since z is a global minimizer of problem (1).

For case (b), we have for all $k \in \mathbb{N}$

$$\begin{aligned} f(x^{(k)}) + \tau^{(k)} \sum_{i=1}^{p} P_{\mu^{(k)}}(g_i(x^{(k)})) &\leq f(z) + \tau^{(k)} \sum_{i=1}^{p} P_{\mu^{(k)}}(g_i(z)) + \varepsilon^{(k)} \\ &= f(z) + \tau^{(k)} p_{\mu^{(k)}}(z) + \varepsilon^{(k)} \end{aligned}$$

replacing $\bar{\tau}$ by $\tau^{(k)}$ in the above relation (13). Taking limits for $k \in \mathscr{N}$ as in case (a), using Theorem 1 and $\lim_{k \in \mathbb{N}} \tau^{(k)} \mu^{(k)} = 0$ (recalling Property 8 for z feasible), we have

$$f(x^*) \leq f(z) + \varepsilon.$$

□

3 Solving the bound constrained subproblems

In this paper, we use stochastic metaheuristics to find an approximate global minimizer $x^{(k)}$ of subproblem (5) that satisfies (11). Three evolutionary biology-based algorithms are tested: the genetic algorithm (GA) [20], probably the most well-known and used evolutionary technique [15,17], a differential evolution technique, that uses specific strategies to set the control parameters self-adaptive according to the available information (jDE) [21], and a recent evolution strategy with covariance matrix adaptation (CMA-ES) [22,23], that has shown to be reliable and very competitive. Another metaheuristic that is inspired from the collective behavior of insect colonies and animal societies, known as the ant colony optimization (ACO) algorithm [24] (a modified version of the ant colony method for continuous domains), and the single solution-based metaheuristic with origins in statistical mechanics, called simulated annealing (SA) [25], are also tested. The other swarm intelligence-based algorithm that is selected for the tests is the AFS algorithm [14,26], a recent metaheuristic successfully applied during the last decade in engineering system design, control, signal processing, data mining, neural networks, scheduling and other areas of operations research [27]. With the exception of the SA, they all are population-based algorithms. For the sake of brevity, details of the GA, jDE, CMA-ES, ACO, SA and AFS algorithms are not included and the reader is referred to the literature.

In the previous section, convergence to an ε-global minimizer of the penalty-based algorithm has been guaranteed, provided that the subproblems are $\varepsilon^{(k)}$-globally solved, where

$\varepsilon^{(k)} \to \varepsilon$. Thus, the issue here is to guarantee that an $\varepsilon^{(k)}$-global minimizer $x^{(k)}$ of subproblem (5) is found so that condition (11) is satisfied for all $x \in \Omega$. We now further elaborate on the AFS algorithm [14,28,29] by focusing on the complete convergence of the algorithm using Markov chains. Stochastic convergence relies on a probability measure and is different from the convergence concept of classical analysis. Convergence of random variables can be defined in various ways, being the convergence with probability one (or almost sure convergence) the probabilistic version of pointwise convergence known from elementary real analysis. A stronger convergence concept is the complete convergence that is sometimes more convenient to establish [30]. Complete convergence implies almost sure convergence.

Convergence properties of AFS-based algorithms have been analyzed in the literature. In [28], convergence in mean square of the AFS algorithm is ensured, and in [29], probability theory is used to prove that an improved version of the AFS algorithm converges to an $\varepsilon^{(k)}$-global minimizer of a hyperbolic augmented Lagrangian function with probability one. We note that convergence in mean square does not imply almost sure convergence, and the converse does not hold true either [30].

The AFS algorithm relies on a swarm intelligence based paradigm to construct fish/point movements over the search space. Since the solution of each subproblem (5) is obtained by a stochastic method that generates a population of solutions at each iteration, each point in the population is considered a stochastic vector. At each iteration t, each point of the current population of m points, herein denoted by $X = [x_1 \dots x_m] \equiv [x_1^{(t)} \dots x_m^{(t)}]$, is used to generate the corresponding trial point by means of a uniformly distributed random variable in $(0,1]$, where $Y = [y_1 \dots y_m]$ represents the trial population. (We note that all points of the population are modified, in contrast to some evolutionary algorithms which often select only a subset of points to apply the evolutionary operators.) An elitist and deterministic selection procedure is then carried out and the current point is replaced by its trial to define the current point for the next iteration, if its fitness is not worse than the current fitness $\phi^k(y_j) \le \phi^k(x_j)$; otherwise the current point is preserved.

Notation 1 *In the context of solving subproblem* (5)*, let*

$$\phi_{best}^{k,t}(X) = \min\{\phi^k(x_j^{(t)}), j = 1, \dots, m\} \tag{14}$$

denote the best function value attained by a point in the population of m points, at iteration t, where t is the iteration counter of the inner cycle, and x_{best} be the corresponding best point of the population. $\phi^{k,*} = \min\{\phi^k(x) : x \in \Omega\}$ is the global minimum of $\phi^k(x)$, when $\tau^{(k)}$ and $\mu^{(k)}$ are fixed.

Let $\{X^{(t)} : t \in T\}$ with index set T be a stochastic process with discrete time on a probability space, characterized by the triple $(E, \mathscr{A}, \mathbf{P})$, which takes values in the set E of a measurable space $(E, \mathscr{A})$. The probability measure function $\mathbf{P}$ associates a number $\mathbf{P}(A)$ to each set $A \in \mathscr{A}$ [30]. The image space E of the stochastic process is called the state space of the process. In this paper, the index set T is identical to $\mathbb{N}_0$ and the indices $t \in T$ are iteration numbers (or time instants).

Definition 2 (Markov chain) A stochastic process $\{X^{(t)} : t \ge 0\}$ with index set T is said to be a Markov chain if for $0 < t_1 < t_2 < \cdots < t_l < t$ with some $l \in \mathbb{N}$ and $A \in \mathscr{A}$

$$\mathbf{P}[X^{(t)} \in A | X^{(t_1)}, X^{(t_2)}, \dots, X^{(t_l)}] = \mathbf{P}[X^{(t)} \in A | X^{(t_l)}]$$

with probability one.

This property is equivalent to the statement that the conditional probability of any future/trial event $X^{(t)}$, given any past event $X^{(t_1)}, X^{(t_2)}, \ldots, X^{(t_{l-1})}$, and the current $X^{(t_l)}$, is independent of the past event. If $\mathbf{P}[X^{(t+j)} \in A | X^{(s+j)}] = \mathbf{P}[X^{(t)} \in A | X^{(s)}]$ for arbitrary t, s, j with $s \leq t$, then the Markov chain is termed homogeneous.

Definition 3 (Markovian/Stochastic kernel) Let $\{X^{(t)} : t \geq 0\}$ be a homogeneous Markov chain on a probability space $(E, \mathscr{A}, \mathbf{P})$ with image space $(E, \mathscr{A})$. The map

$$\mathbf{K} : E \times \mathscr{A} \to [0,1]$$

is called a stochastic kernel (or a transition probability function) for the Markov chain if

i) $\mathbf{K}(\cdot, A)$ is measurable for any fixed set $A \in \mathscr{A}$;
ii) $\mathbf{K}(X, \cdot)$ is a probability measure on $(E, \mathscr{A})$ for any fixed state/event $X \in E$.

In particular, $\mathbf{K}(X^{(t)}, A) = \mathbf{P}[X^{(t+1)} \in A | X^{(t)} = x^{(t)}]$ [30,31].

Since a relation between the limiting behavior of a Markov chain and the stochastic convergence of random sequences can be established [30], the AFS algorithm can be described by a Markov chain. In the AFS algorithm, the position of each trial point in the population only depends on the position of the current point of the population, in a probabilistic manner. Considering the whole population, we say that the state of the next iteration only depends on the state of the previous iteration. Thus, Markov chains are appropriate to model and analyze the AFS algorithm convergence properties. At each iteration, the AFS algorithm generates a population of m points $X^{(t)}$, denoted as a state, where each solution is $x_j^{(t)} \in \Omega$ $(j = 1, \ldots, m)$ and the initial state $X^{(0)} = [x_1^{(0)} \ldots x_m^{(0)}]$ is generated according to some initial distribution. A population at iteration $t \geq 0$ is modified to generate the population at iteration $t+1$. The outcome only depends on the population of the previous iteration and the probabilistic modifications are described by the stochastic kernel $\mathbf{K}(\cdot, \cdot)$.

Remark 2 Using a population-based solution method, with m points, the state space is $E = \Omega^m$.

If the sequence $\{\phi_{\text{best}}^{k,t}\}$ converges in some stochastic manner to the global optimum $\phi^{k,*}$, then the likelihood is that the population $X^{(t)}$ generates better and better solutions of $\min\{\phi^k(x) : x \in \Omega\}$, as t increases. The following definition of convergence of a stochastic algorithm is used.

Definition 4 (Convergence of a stochastic algorithm) A stochastic algorithm converges to an $\varepsilon^{(k)}$-global minimizer of the function $\phi^k : \Omega \to \mathbb{R}$ if the random sequence $\{D^{(t)}\}$, where

$$D^{(t)} \equiv D(X^{(t)}) = \phi_{\text{best}}^{k,t}(X) - \phi^{k,*}$$

and $\phi_{\text{best}}^{k,t}(X)$ is defined in (14), converges completely to zero, i.e., if for any $\varepsilon^{(k)} > 0$

$$\lim_{t \to \infty} \sum_{i=1}^{t} \mathbf{P}[|D^{(i)}| > \varepsilon^{(k)}] < \infty. \tag{15}$$

It should be noted that $\{D^{(t)}\}$ converging completely to zero implies that $\{D^{(t)}\}$ converges almost surely to zero. Further, the two types of convergence are equivalent if $\{D^{(t)}\}$ is a sequence of independent random variables. Hence, conditions for which the AFS algorithm converges to an $\varepsilon^{(k)}$-global optimizer, in the sense of Definition 4, are herein discussed. Let

$$A_{\varepsilon^{(k)}} = \{X^{(t)} \in E : D(X^{(t)}) \leq \varepsilon^{(k)}\}$$

be the set of $\varepsilon^{(k)}$-optimal states (population of points), for some $\varepsilon^{(k)} > 0$.

The below provided results are similar to those established in [31] for a particular class of evolutionary algorithm. In order to state the convergence properties of the AFS algorithm, we need the following lemma and theorem.

Lemma 1 *(Lemma 1 in [31]) Consider $A_{\varepsilon^{(k)}}$ the set of $\varepsilon^{(k)}$-optimal states. If $\boldsymbol{K}(X,A_{\varepsilon^{(k)}}) \geq \lambda > 0$ for all $X \in A^c_{\varepsilon^{(k)}} \equiv E \backslash A_{\varepsilon^{(k)}}$ and $\boldsymbol{K}(X,A_{\varepsilon^{(k)}}) = 1$ for $X \in A_{\varepsilon^{(k)}}$, then for all $t \geq 1$:*

$$\boldsymbol{K}^{(t)}(X,A_{\varepsilon^{(k)}}) \geq 1-(1-\lambda)^t. \tag{16}$$

Proof The proof is by induction [31]. □

Theorem 3 *(similar to Theorem 1 in [31]) Assume that the real-valued objective function of subproblem* (5), *$\phi^k : \Omega \to \mathbb{R}$, is bounded below, that is, $\phi^k(x) > -\infty$ for all $x \in \Omega$. An algorithm whose stochastic kernel satisfies the preconditions of Lemma 1 will converge to an $\varepsilon^{(k)}$-global minimizer of ϕ^k, in the sense of Definition 4, regardless of the initial distribution.*

We note that the word preconditions refers to the stated necessary conditions in Lemma 1 to prove the result (16). From Theorem 3, we may conclude that instead of analyzing the stochastic properties of the iterates, it is sufficient to investigate the properties of the stochastic kernel associated with the AFS algorithm. Hence, in the next theorem we state our main result concerned with the stochastic kernel of the Markov chain produced by the AFS algorithm.

Theorem 4 *The stochastic kernel of the Markov chain produced by the AFS algorithm satisfies the preconditions of Lemma 1.*

Proof To show that the Markov chain produced by the AFS algorithm has a stochastic kernel that satisfies the preconditions of Lemma 1, each iteration of the AFS algorithm is split into two phases: the movement phase and the selection one. In the movement phase, each trial point $y_j^{(t)}$ is generated based on the current point $x_j^{(t)}$. In the selection phase, a comparison between each current and the trial point is made to select the point for transition to the next iteration. These phases may be described separately by their own stochastic kernels since selection begins after movement has being terminated. Thus, the stochastic kernel of the AFS algorithm is described as the product kernel:

$$\mathbf{K}(X^{(t)},A) = (\mathbf{K}_m\mathbf{K}_s)(X^{(t)},A) = \int_E \mathbf{K}_m(X^{(t)},dY)\mathbf{K}_s(Y^{(t)},A)$$

where $\mathbf{K}_m, \mathbf{K}_s$ stand for the movement kernel and selection kernel respectively.

To derive the selection kernel $\mathbf{K}_s$, we first assume that the population has just one point ($m = 1$), $x^{(t)} \in E$. This is the simplest case. We note that a trial point $y^{(t)} \in E$ will be the current point $x^{(t+1)}$ for the next iteration if $y^{(t)}$ is better than or equal to $x^{(t)}$, i.e., if

$$y^{(t)} \in B(x^{(t)}) = \{v^{(t)} \in E : \phi^k(v^{(t)}) \leq \phi^k(x^{(t)})\};$$

otherwise, the current point is preserved to the next iteration ($x^{(t+1)} \leftarrow x^{(t)}$). Let $B^c(x^{(t)})$ be defined as $B^c(x^{(t)}) \equiv E \backslash B(x^{(t)})$. Since selection depends on the current point $x^{(t)} \in E$, the kernel incorporates an additional parameter and is given by:

$$\begin{aligned} \mathbf{K}_s(y^{(t)},A;x^{(t)}) &= \mathbf{I}_{B(x^{(t)})}(y^{(t)})\mathbf{I}_A(y^{(t)}) + \mathbf{I}_{B^c(x^{(t)})}(y^{(t)})\mathbf{I}_A(x^{(t)}) \\ &= \mathbf{I}_{A\cap B(x^{(t)})}(y^{(t)}) + \mathbf{I}_{B^c(x^{(t)})}(y^{(t)})\mathbf{I}_A(x^{(t)}) \end{aligned} \tag{17}$$

where $\mathbf{I}_A(y^{(t)})$ denotes the indicator function for the set A, which returns 1 if $y^{(t)} \in A$, and returns 0 otherwise. The selection kernel in (17) is interpreted as follows. If the trial point $y^{(t)} \in E$ is better than or equal to the current point $x^{(t)}$, i.e. if $y^{(t)} \in B(x^{(t)})$, and is also in the set A, then it will be a current point for the next population $x^{(t+1)} \leftarrow y^{(t)}$ ($y^{(t)}$ will transition to the set A, in particular to the set $A \cap B(x^{(t)})$ with probability one). However, if $y^{(t)}$ is worse than $x^{(t)}$, i.e. if $y^{(t)} \in B^c(x^{(t)})$, then the current point will be preserved: $x^{(t+1)} \leftarrow x^{(t)}$. The stochastic kernel of the algorithm is then:

$$\begin{aligned}\mathbf{K}(x^{(t)},A) &= \int_E \mathbf{K}_m(x^{(t)},dy)\,\mathbf{I}_{A\cap B(x^{(t)})}(y^{(t)}) + \mathbf{I}_A(x^{(t)}) \int_E \mathbf{K}_m(x^{(t)},dy)\,\mathbf{I}_{B^c(x^{(t)})}(y^{(t)}) \\ &= \int_{A\cap B(x^{(t)})} \mathbf{K}_m(x^{(t)},dy) + \mathbf{I}_A(x^{(t)}) \int_{B^c(x^{(t)})} \mathbf{K}_m(x^{(t)},dy) \\ &= \mathbf{K}_m(x^{(t)},A\cap B(x^{(t)})) + \mathbf{I}_A(x^{(t)})\mathbf{K}_m(x^{(t)},B^c(x^{(t)})).\end{aligned} \tag{18}$$

We now address the case of a population with more than one point (with $E = \Omega^m$ and $m > 1$). The set of states better than or equal to the state $X^{(t)}$ is re-defined, according to (14), as

$$B(X^{(t)}) = \{Y^{(t)} \in E : \phi^k_{\text{best}}(Y^{(t)}) \leq \phi^k_{\text{best}}(X^{(t)})\}.$$

If $Y^{(t)} \in E$ is in $A \cap B(X^{(t)})$, the population transitions to A. However, if $Y^{(t)} \in B^c(X^{(t)})$, meaning that the best point of the population $Y^{(t)}$ is worse than the best point of population $X^{(t)}$, the entire population $Y^{(t)}$ may be rejected and all points in $X^{(t)}$ are selected. This is equivalent to the above structure with one point in the population (see (18)). In the other case, where the selected population $X^{(t+1)}$ will have some trial points from $Y^{(t)}$ as well as points from $X^{(t)}$, the best point in $X^{(t)}$ is in the selected population and consequently $X^{(t+1)} \in B(X^{(t)})$.

In the sequence of expression (18), the kernel has the following structure:

$$\mathbf{K}(X^{(t)},A) = \mathbf{K}_m(X^{(t)},A\cap B(X^{(t)})) + \mathbf{I}_A(X^{(t)}) \int_{B^c(X^{(t)})} \mathbf{K}_m(X^{(t)},dY)\,\mathbf{I}_A(X^{(t+1)}). \tag{19}$$

If we restrict the analysis to the set $A_{\varepsilon^{(k)}}$ of $\varepsilon^{(k)}$-optimal states, then using (19) we obtain for the general case:

i) if $A_{\varepsilon^{(k)}} \subset B(X^{(t)})$ then $X^{(t)} \notin A_{\varepsilon^{(k)}}$, $A_{\varepsilon^{(k)}} \cap B(X^{(t)}) = A_{\varepsilon^{(k)}}$ and $\mathbf{I}_{A_{\varepsilon^{(k)}}}(X^{(t)}) = 0$; thus $\mathbf{K}(X^{(t)},A_{\varepsilon^{(k)}}) = \mathbf{K}_m(X^{(t)},A_{\varepsilon^{(k)}})$;

ii) if $B(X^{(t)}) \subseteq A_{\varepsilon^{(k)}}$ then $X^{(t)} \in A_{\varepsilon^{(k)}}$, $A_{\varepsilon^{(k)}} \cap B(X^{(t)}) = B(X^{(t)})$ and $\mathbf{I}_{A_{\varepsilon^{(k)}}}(X^{(t)}) = 1$; thus
$$\begin{aligned}\mathbf{K}(X^{(t)},A_{\varepsilon^{(k)}}) &= \mathbf{K}_m(X^{(t)},B(X^{(t)})) + \int_{B^c(X^{(t)})} \mathbf{K}_m(X^{(t)},dY)\,\mathbf{I}_{A_{\varepsilon^{(k)}}}(X^{(t+1)}) \\ &= \mathbf{K}_m(X^{(t)},B(X^{(t)})) + \mathbf{K}_m(X^{(t)},B^c(X^{(t)})) = 1\end{aligned}$$
since $X^{(t+1)} \in B(X^{(t)}) \subseteq A_{\varepsilon^{(k)}}$ and $\mathbf{I}_{A_{\varepsilon^{(k)}}}(X^{(t+1)}) = 1$.

Therefore the stochastic kernel restricted to the set $A_{\varepsilon^{(k)}}$ is given by

$$\mathbf{K}(X^{(t)},A_{\varepsilon^{(k)}}) = \mathbf{K}_m(X^{(t)},A_{\varepsilon^{(k)}})\,\mathbf{I}_{A^c_{\varepsilon^{(k)}}}(X^{(t)}) + \mathbf{I}_{A_{\varepsilon^{(k)}}}(X^{(t)})$$

satisfying the preconditions of Lemma 1 if $\mathbf{K}_m(X^{(t)},A_{\varepsilon^{(k)}}) \geq \lambda > 0$, a value strictly bounded from zero, although it may be $\varepsilon^{(k)}$ dependent, for all $X^{(t)} \in A^c_{\varepsilon^{(k)}}$.

Thus, we must guarantee that the set $A_{\varepsilon^{(k)}}$ can be reached from everywhere outside $A_{\varepsilon^{(k)}}$ with some probability $\lambda > 0$. The argument is the following. Each point of the population is moved at random and independently. Let

$$M_{\varepsilon^{(k)}} = \{x^{(t)} \in \Omega : \phi^k(x^{(t)}) - \phi^{k,*} \leq \varepsilon^{(k)}\}$$

be the set of $\varepsilon^{(k)}$-optimal solutions, for some $\varepsilon^{(k)} > 0$, where $\phi^k(x^{(t)})$ is the penalty function value computed at $x^{(t)}$. If the probability that a point is moved to a point in the set $M_{\varepsilon^{(k)}} \subseteq \Omega$ is larger than $\beta \equiv \beta(\varepsilon^{(k)}) > 0$, then the probability that a point is moved to another point outside $M_{\varepsilon^{(k)}}$ is smaller or equal to $1 - \beta$. Now, the probability that a population of m points moves to another one which is outside $A_{\varepsilon^{(k)}}$, i.e., to a population whose best point satisfies $\phi^k_{\text{best}}(X^{(t)}) - \phi^{k,*} > \varepsilon^{(k)}$, is smaller or equal to $\prod_{i=1}^m (1-\beta) = (1-\beta)^m$, since each point moves at random and independently. Therefore, the probability that a population enters the set $A_{\varepsilon^{(k)}} \in \mathscr{A}$ is at least $\lambda \equiv 1 - (1-\beta)^m > 0$. Since $\beta > 0$ implies $\lambda > 0$, we must have $\mathbf{K}_m(x^{(t)}, M_{\varepsilon^{(k)}}) \geq \beta > 0$. We note here that the uniform distribution used to create the trial point or, for that matter, any other distribution with density function that is nowhere zero over the set E satisfy the condition. □

Using the results of Lemma 1 and both Theorems 3 and 4, the convergence property of the AFS algorithm follows. (This result corresponds to Theorem 2 in [31].)

Theorem 5 *[31] Assume that the objective function of subproblem* (5), $\phi^k : \Omega \to \mathbb{R}$, *is bounded below. Then, the stochastic AFS algorithm will converge to an* $\varepsilon^{(k)}$*-global minimizer of* ϕ^k*, in the sense of Definition 4, regardless of the initial distribution.*

4 Numerical Experiments

For a preliminary practical validation of the proposed algorithm based on the smoothed penalty function, two sets of benchmark constrained global optimization problems are used. The C programming language is used in this real-coded algorithm and the numerical experiments were performed on a PC with a a 2.7 GHz Core i7-4600U and 8 Gb of memory.

4.1 Using metaheuristics to solve subproblem (5)

First, we analyze the performance of Algorithm 2.1 when different metaheuristics are used to solve subproblem (5). For this experiment, the following parameter values are used. Due to their restrictive conditions, the parameters $\tau^{(1)}$ and $\mu^{(1)}$ have been naturally set both to one. The parameter γ_g is set to 0.5, the most used value in the literature. Parameters γ_τ and γ_μ must be related so that Property 8 holds and after fixing $\gamma_\tau = 2$, γ_μ is set to 0.25. Other parameters are set as follows: $\varepsilon^{(1)} = 1$, $\gamma_\varepsilon = 0.1$, $\varepsilon_{\min} = 10^{-5}$, $\eta_{\min} = 10^{-6}$ and $k_{\max} = 20$. For population-based algorithms, we set $m = 50$ and a maximum number of iterations, $t_{\max} = 100$, is allowed; in the SA we set $t_{\max} = 2500$. Each problem was solved 30 times. The metaheuristics GA, jDE, CMA-ES, ACO, SA and AFS are tested using the parameter values as suggested in the above cited papers. We use 11 problems that have only inequality constraints and are a subset of the well-known g-suite [32]. We note that g02, g08 and g12 are maximization problems that were converted into minimization ones. We report the number of the problem, ‘Prob.’, the median (as a measure of the central tendency of the distribution) of the 30 solutions, ‘f_{median}’, and the average number of function evaluations,

Table 1 Results produced by Algorithm 2.1 using GA, jDE, CMA-ES, ACO, SA and AFS algorithms to solve subproblem (5)

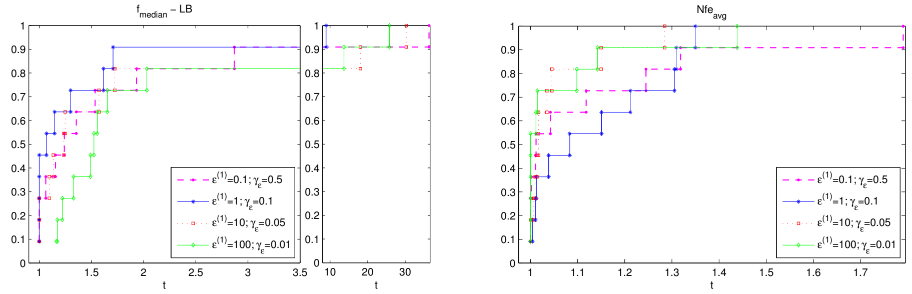
Prob.	*LB*		GA	jDE	CMA-ES	ACO	SA	AFS
g01	-15.000000	f_{median}	*-14.999992*	-14.998716	-14.999944	-14.999991	-14.852036	-14.999994
		Nfe_{avg}	75969	101001	102021	74023	90210	88770
g02	-0.803619	f_{median}	*-0.757711*	-0.800193	-0.717933	-0.754064	-0.360098	-0.540234
		Nfe_{avg}	98199	96214	102021	96061	87034	171397
g04	-30665.53867	f_{median}	-30665.52283	*-30665.53568*	-30662.61083	-30665.53798	-30664.91162	-30665.47540
		Nfe_{avg}	102933	100994	44955	97528	100061	112993
g06	-6961.813876	f_{median}	-6912.430838	-6943.599554	-6958.370917	-6961.673093	*-6961.312079*	-6961.075018
		Nfe_{avg}	103001	101001	41083	101001	100061	106871
g07	24.306209	f_{median}	25.136433	26.269085	24.325982	26.162025	24.588419	*24.337138*
		Nfe_{avg}	103001	101001	102021	101001	100061	136988
g08	-0.095825	f_{median}	-0.095824	-0.095823	-0.095825	*-0.095824*	-0.095823	-0.095822
		Nfe_{avg}	11534	11916	21085	*10736*	15367	10951
g09	680.630057	f_{median}	680.666794	680.655094	680.630513	*680.632149*	680.690420	680.721112
		Nfe_{avg}	103001	101001	102021	101001	100061	125957
g10	7049.248021	f_{median}	*7243.461699*	7514.086284	7135.130208	8550.501792	7541.187987	7826.758564
		Nfe_{avg}	103001	101001	102021	101001	100061	129351
g12	-1.000000	f_{median}	-1.000000	-1.000000	-1.000000	-1.000000	-0.999990	*-1.000000*
		Nfe_{avg}	7599	5269	19045	7013	63425	*6661*
g18	-0.866025	f_{median}	-0.865728	-0.828130	*-0.866001*	-0.833600	-0.855751	-0.866016
		Nfe_{avg}	103001	101001	102021	101001	100061	101619
g24	-5.508013	f_{median}	*-5.508009*	-5.508007	-5.508011	-5.508008	-5.508006	-5.508005
		Nfe_{avg}	66703	64454	70225	52894	83331	64589

'Nfe_{avg}', after the 30 runs. For comparative purposes, the best-known solution available in the literature [32], '*LB*', is shown. From the results summarized in Table 1, we conclude that CMA-ES comes five times in 1st place (results presented "underlined") and once in 2nd place (presented in "italic" style), both ACO and AFS come twice in 1st place and twice in 2nd place, jDE comes twice in 1st place and once in 2nd place, GA comes four times in 2nd place and SA comes once in 2nd place.

To analyze the statistical significance of the results, we perform a Friedman test. The Matlab$^{\text{TM}}$ (Matlab is a registered trademark of the MathWorks, Inc.) function `friedman` is used. This is a non-parametric statistical test for multiple comparisons to determine significant differences in mean for one independent variable with two or more levels (the results achieved by two or more methods) and a dependent variable (the matched groups taken as the problems) [33]. When the assumption of normality of the data is absent, non-parametric tests are the preferred ones. The null hypothesis in this test is that the mean ranks assigned to the results of the methods under testing are the same. We apply the statistical procedure in the joint analysis of the results of the six distributions of f_{median} values. The Friedman's chi-square value is 11.06 and the corresponding p-value = 0.0502 indicates for a significance level of 5% (or even 1%) that we do not have enough evidence to reject the null hypothesis of "no significant differences on mean ranks", i.e., the observed differences between the six distributions of f_{median} values are not statistically significant. From the results in the table, we can also conclude that the proposed penalty algorithm is effective in finding the global optimal solutions to CGO problems.

4.2 Sensitivity analysis of parameters

The Algorithm 2.1 using the AFS metaheuristic for solving the subproblem (5), undergoes a sensitivity analysis on the parameters $\varepsilon^{(1)} > 0$, $0 < \gamma_\varepsilon < 1$ and $0 < \gamma_\mu < 1$, which is related to $\gamma_\tau > 1$ by the condition $\gamma_\mu < (\gamma_\tau)^{-1}$. For this set of experiments, we set $k_{\max} = 20$, $m = \min\{5n, 50\}$, $t_{\max} = 50$ and solve each problem 30 times. The effect of the pair $(\varepsilon^{(1)}, \gamma_\varepsilon)$ on the performance of the Algorithm 2.1 is analyzed by using four different pairs of values (0.1, 0.5), (1, 0.1), (10, 0.05) and (100, 0.01), after fixing $\gamma_\tau = 2$ and $\gamma_\mu = 0.25$. We use a graphical procedure to visualize the performance differences among the results produced by these four sets of parameters, in relative terms on the previously selected 11 problems, known as performance profiles [34]. These profiles correspond to (cumulative) distribution functions for a performance metric. The plot reports (on the vertical axis) the percentage of problems solved with the competing variants of the algorithm that is within a certain threshold, t, (on the horizontal axis) of the best result. The performance profiles of the four distributions of the metric $f_{\text{median}} - LB$ (error of the median f value to the LB), for the selected problems, are shown in the Figure 4(a). Figure 4(b) corresponds to the metric Nfe_{avg}. The higher the percentage the better. A higher value for $t = 1$ means that the algorithm based on that pair of values achieves the best median values (when f_{median} is analyzed) and takes the lowest computational efforts (when Nfe_{avg} is analyzed) mostly. Therefore, the pair (1, 0.1) is more successful since it has the highest probability of being the best setting for $(\varepsilon^{(1)}, \gamma_\varepsilon)$ as far as the median of f values is concerned. From the plot, the probability that the pair (1, 0.1) is the winner on a given problem is about 0.45 since it produces the best f_{median} in five of the 11 problems. (Each one of the pairs (0.1, 0.5) and (10, 0.05) produces the best f_{median} in three out of 11 problems and the pair (100, 0.01) does not produce any best result.) On the other hand, the right end of the profile gives the percentage of the problems that are successfully solved by each variant. From the results relative to the metric Nfe_{avg}, we may conclude that the par (1, 0.1) is the best in one of the 11 problems, being the pair (100, 0.01) the best in six out of 11 problems. Thus, the pair that is able to reach better f values is on average computationally more expensive.



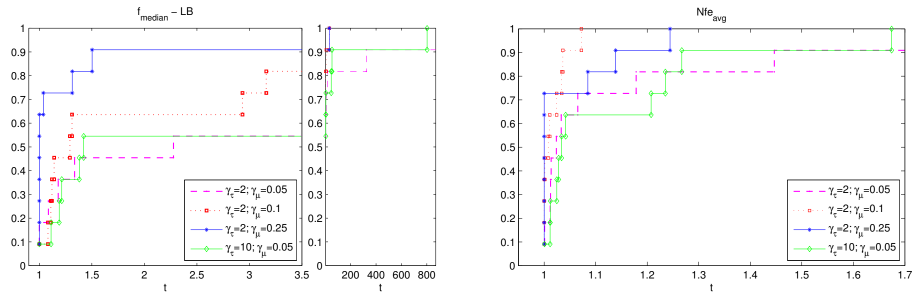
(a) $f_{\text{median}} - LB$ metric

(b) Nfe_{avg} metric

Fig. 4 Performance profiles for four different pairs of $(\varepsilon^{(1)}, \gamma_\varepsilon)$, using $\gamma_\tau = 2$ and $\gamma_\mu = 0.25$

We now aim to analyze the effect of the reduction factor γ_μ on the performance of Algorithm 2.1. Besides 0.25, we also test the values 0.1 and 0.05 (with γ_τ fixed at 2) and $\gamma_\mu = 0.05$ with $\gamma_\tau = 10$. From the profiles in the Figure 5(a) we may conclude that $\gamma_\mu = 0.25$

with $\gamma_\tau = 2$ gives the best median solutions in seven of the 11 problems. The variant with $\gamma_\tau = 2$ and $\gamma_\mu = 0.05$ is not able to find a feasible solution to one of the problems and this is noted in its profile that does not reach the top at the right end of the plot. From the Figure 5(b) it is possible to conclude that the variant with $\gamma_\mu = 0.25$ and $\gamma_\tau = 2$ also is on average computationally less expensive.



(a) $f_{\text{median}} - LB$ metric

(b) Nfe_{avg} metric

Fig. 5 Performance profiles for different γ_μ and γ_τ, using $\varepsilon^{(1)} = 1$ and $\gamma_\varepsilon = 0.1$

4.3 Comparison with penalty-based algorithms

We now compare our results with those obtained by recent penalty-based algorithms [13, 29, 35, 36]. To solve the subproblem (5), in the context of the proposed Algorithm 2.1, the AFS algorithm is used. A genetic algorithm based augmented Lagrangian (GAAL) method is proposed in [35], a hyperbolic augmented Lagrangian (HAL) algorithm based on the improved AFS metaheuristic is presented in [29], a shifted hyperbolic augmented Lagrangian (sh-HAL) combined with an enhanced two-swarm AFS algorithm is analyzed in [36], and in [13], an exact penalty (ex-PEN) method based on the DIRECT solver is proposed.

Table 2 contains our results and those obtained by GAAL, available in [35]. During these experiments we use $\gamma_\tau = 2$, $\gamma_\mu = 0.25$, $\varepsilon^{(1)} = 1$, $\gamma_\varepsilon = 0.1$, $k_{\max} = 20$, $t_{\max} = 500$ and consider similar conditions to those reported in the cited paper, as follows: $m = \max\{10n, 50\}$, each problem was solved 25 times and the algorithm terminates when the absolute difference between the function values of two consecutive iterations is less than or equal to 10^{-4}. The other parameters of Algorithm 2.1 are set as previously defined. We report 'f_{best}' (the best solution obtained after the 25 runs), 'f_{median}', 'f_{worst}' (the worst solution of the 25 runs) and the percentage of successful runs, 'SR', where a run is considered to be successful if the obtained feasible solution, f_{sol}, satisfies $|f_{\text{sol}} - LB| \leq 10^{-4}$. The solutions produced by GAAL have high quality [35]. This is an expectable behavior since the GAAL algorithm employs, at the end of each set of five iterations, the gradient-based `fmincon` local search function (from Matlab$^{\text{TM}}$ Optimization Toolbox) starting from the best found solution. We have also invoked a local search procedure, at each iteration, to be able to obtain higher quality solutions. However, a derivative-free local solver named Hooke-and-Jeeves (HJ) is used instead [37]. The search HJ is allowed to run for 50 iterations. We conclude that the results produced by our algorithm based on the smoothed penalty function are very competitive. Applying the Friedman test for the comparison of the distributions of f_{best} values relative to the Algorithm 2.1 and GAAL, the chi-square statistical value of 2.78 and a p-value of 0.0956 are obtained. Thus, there is not enough evidence to reject the null hypothesis of "no significant differences on mean ranks", at a significance level of 5%. When the Friedman test is applied to the distributions of f_{median} values, the chi-square statistical value is 7.36 and the p-value = 0.0067. Hence, the null hypothesis is rejected, at a significance level of 5%, and we conclude that the distributions of f_{median} values have statistically significant differences.

Table 2 Comparison of Algorithm 2.1 with GAAL

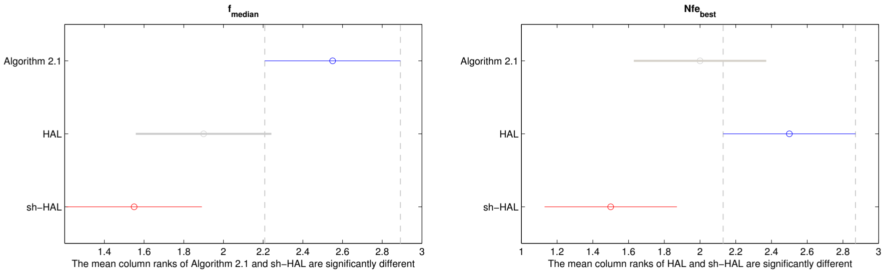
	Algorithm 2.1				GAAL in [35]			
Prob.	f_{best}	f_{median}	f_{worst}	SR	f_{best}	f_{median}	f_{worst}	SR
g01	-15.000001	-14.999999	-14.999990	100	-15	-15	-12.453125	96
g02	-0.798167	-0.712454	-0.643880	0	-0.803619	-0.803619	-0.744767	92
g04	-30665.538671	-30665.536658	-30665.424037	20	-30665.538672	-30665.538672	-30665.538672	100
g06	-6961.813263	-6961.687392	-6961.641895	0	-6961.813876	-6961.813876	-6961.813876	100
g07	24.306302	24.306700	24.310923	8	24.306209	24.306209	24.306209	100
g08	-0.095825	-0.095823	-0.095802	100	-0.095825	-0.095825	0	32
g09	680.630143	680.630792	680.636586	8	680.630057	680.630057	680.630057	100
g10	7073.392696	7452.558621	8110.527953	0	7049.24802	7049.24802	7049.24802	100
g12	-1.000000	-1.000000	-1.000000	100	-0.999375	-0.999375	-0.999375	100
g18	-0.866015	-0.866004	-0.865955	100	-0.866025	-0.866025	-0.674981	96
g24	-5.508003	-5.508001	-5.507999	100	-5.508013	-5.508013	-5.508013	100

For the comparison of the Algorithm 2.1 with the HAL algorithm in [29], the sh-HAL algorithm in [36] and the ex-PEN method [13], a different set of 20 small constrained global optimization problems is tested (the number of variables ranges from 2 to 6 and the number of constraints ranges from 1 to 12) [19]. Table 3 lists the number of the problem and the best-known solution available in the literature, '*LB*', as shown in [19]; the best solution obtained by Algorithm 2.1 (during the 30 runs), 'f_{best}'; the median, 'f_{median}'; and the number of function evaluations required by the best solution, 'Nfe_{best}'. The table also displays the results produced by HAL, by sh-HAL, the solution found by the ex-PEN, 'f_{sol}' and the number of function evaluations, 'Nfe' available in [13]. The parameters for this experiment are set as follows: $m = \max\{50, 10n\}$, $t_{\max} = 20$, $k_{\max} = 10$ and the parameter υ for the equality constraints of the problems was set to 10^{-15}. From the results we may conclude that Algorithm 2.1 based on the proposed smoothed penalty function performs reasonably well. The comparison with the results in [13] is very favorable. From the comparison with the results in [29], we conclude that Algorithm 2.1 is able to reach best and median solutions similar to its competitor using a lower computational effort. The comparison with the results in [36] is slightly favorable to sh-HAL. The statistical test, for multiple comparisons based on the Friedman test, on the three distributions of f_{median} values relative to the Algorithm 2.1, HAL and sh-HAL (penalty-based algorithms combined with variants of the AFS metaheuristic), gives a chi-square statistical value of 6.4, with p-value = 0.0408. Hence, the null hypothesis of "no significant differences on mean ranks" is rejected, at a significance level of 5%, and there is evidence that the three distributions of f_{median} values have statistically significant differences. (However, at a significance level of 1% – a smaller probability of making a wrong decision – there is no evidence to reject the null hypothesis.)

Table 3 Comparison of Algorithm 2.1 with HAL, sh-HAL and ex-PEN

		results of the Algorithm 2.1			HAL in [29]			sh-HAL in [36]			ex-PEN in [13]	
Prob.	LB	f_{best}	Nfe_{best}	f_{median}	f_{best}	Nfe_{best}	f_{median}	f_{best}	Nfe_{best}	f_{median}	f_{sol}	Nfe
1	0.029313	0.0293	4719	0.0295	0.0342	9608	0.1204	0.0294	9723	0.0450	0.0625	39575
2(a)	-400.00	-362.9380	7565	-313.5667	-380.674	15813	-369.111	-400.0000	2434	-400.0000	-134.1127	115107
2(b)	-600.00	-530.9222	7472	-306.3665	-385.051	15808	-360.786	-600.0000	4038	-400.0000	-768.4569	120057
2(c)	-750.00	-749.9787	7662	-740.9088	-743.416	15612	-693.743	-750.0000	2433	-750.0000	-82.9774	102015
2(d)	-400.00	-399.9999	7627	-399.2205	-399.910	15394	-399.492	-400.0000	2711	-400.0000	-385.1704	229773
3(a)	-0.38880	-0.3853	9382	-0.3552	-0.3880	18928	-0.3849	-0.3840	19144	-0.3691	-0.3861	48647
3(b)	-0.38881	-0.3888	2295	-0.3888	-0.3888	2589	-0.3888	-0.3888	548	-0.3888	-0.3888	3449
4	-6.6666	-6.6667	2537	-6.6667	-6.6667	2242	-6.6667	-6.6667	698	-6.6667	-6.6666	3547
5	201.16	201.1593	3099	201.1593	201.159	2926	201.159	201.1593	2717	201.1593	201.1593	14087
6	376.29	376.2925	2689	376.4515	376.292	5617	376.293	376.2919	1578	376.2919	0.4701	1523
7	-2.8284	-2.8284	2436	-2.8284	-2.8284	3434	-2.8284	-2.8284	886	-2.8284	-2.8058	13187
8	-118.70	-118.7048	2908	-118.7036	-118.705	2884	-118.705	-118.7049	995	-118.7049	-118.7044	7621
9	-13.402	-13.4019	4508	-13.4017	-13.4018	5732	-13.4017	-13.4019	1437	-13.4019	-13.4026	68177
10	0.74178	0.7418	2646	0.7419	0.7418	6342	0.7418	0.7418	1155	0.7418	0.7420	6739
11	-0.50000	-0.5000	2228	-0.5000	-0.5000	3313	-0.5000	-0.5000	1043	-0.5000	-0.5000	3579
12	-16.739	-16.7389	227	-16.7389	-16.7389	98	-16.7389	-16.7389	267	-16.7389	-16.7389	3499
13	189.35	189.3421	4432	189.3445	189.345	9230	189.347	189.3449	9703	189.3449	195.9553	8085
14	-4.5142	-4.5142	4117	-4.5142	-4.5142	6344	-4.5142	-4.5142	1170	-4.5142	-4.3460	19685
15	0.0000	0.0000	4700	0.0000	0.0000	2546	0.0000	0.0000	3187	0.0000	0.0000	1645
16	0.70492	0.7049	366	0.7049	0.7049	1850	0.7049	0.7049	371	0.7049	0.7181	22593

Pairwise comparisons may be carried out to determine which mean ranks are significantly different. The Matlab function `multcompare` is applied. The estimates of the 95% confidence intervals are shown in the Figure 6(a) for each case under testing. Two compared distributions of f_{median} are significantly different if their intervals are disjoint and are not significantly different if their intervals overlap. Hence, we conclude that the mean ranks produced by Algorithm 2.1 is significantly different from that of sh-HAL. For the other two pairs of comparison there are no significant differences on the mean ranks.



(a) f_{median} metric

(b) Nfe_{best} metric

Fig. 6 Estimates of the 95% confidence intervals

When the Friedman test is applied to the distributions of f_{best} values, the obtained chi-square statistical value is 2.4762 with p-value = 0.2899. Thus, there is no evidence that the three distributions of f_{best} have statistically significant differences. Further, the Friedman test applied to the other metric Nfe_{best}, gives a chi-square statistical value of 10 and a p-value = 0.0067, and we conclude that there is evidence that the three sets of Nfe_{best} have statistically significant differences, at a significance level of 5% (or even 1%). Using `multcompare`, the estimates of the 95% confidence intervals are shown in the Figure 6(b). It is possible to conclude that the statistically significant differences in the mean ranks are only between HAL and sh-HAL.

5 Conclusions

We have demonstrated that a penalty-based algorithm, that uses the coercive smoothed penalty (6), has guaranteed convergence to an ε-global minimum of a CGO problem. The newly developed smoothed penalty is continuous, differentiable and relies on the hyperbolic tangent and hyperbolic sine functions. Approximations to the global optimal solutions of the subproblems are obtained by some well-established and well-known stochastic meta-heuristics, as well as by the AFS algorithm. Further, using a relation between the limiting behavior of a Markov chain and the stochastic convergence of random sequences of iterates, we have also proved that the transition probability of the Markov chain produced by the AFS algorithm satisfies the property of Lemma 1 (see also [31]). This property is guaranteed by the movement kernel positiveness and the selection kernel elitism of the AFS algorithm. Then, we can conclude that in the limit, the AFS algorithm convergence to an $\varepsilon^{(k)}$-global minimum of the real-valued smoothed penalty function is guaranteed with probability one, where $\varepsilon^{(k)} \to \varepsilon$.

Some preliminary numerical experiments and a comparison with other penalty-based frameworks show that our proposed penalty algorithm is very competitive. A more exhaustive comparison with other techniques for CGO remains to be done. Large dimensional problems will be considered in a near future.

Acknowledgments

The authors would like to thank two anonymous referees for their valuable comments and suggestions to improve the paper.

This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT - Fundação para a Ciência e Tecnologia within the projects UID/CEC/00319/2013 and UID/MAT/00013/2013.

References

1. Ali, M.M., Golalikhani, M., Zhuang, J.: A computational study on different penalty approaches for solving constrained global optimization problems with the electromagnetism-like method, Optimization, **63**(3), 403–419 (2014)
2. Ali, M.M., Zhu, W.X.: A penalty function-based differential evolution algorithm for constrained global optimization. Comput. Optim. Appl. **54**(3), 707–739 (2013)
3. Barbosa, H.J.C., Lemonge, A.C.C.: An adaptive penalty method for genetic algorithms in constrained optimization problems. In: H. Iba (Ed.), Frontiers in Evolutionary Robotics pp. 9–34, I-Tech Education Publ., Austria (2008)
4. Coello, C.A.C.: Theoretical and numerical constraint-handling techniques used with evolutionary algorithms: a survey of the state of the art. Comput. Method. Appl. M. **191**(11–12), 1245–1287 (2002)
5. Liu, J.-L., Lin, J.-H.: Evolutionary computation of unconstrained and constrained problems using a novel momentum-type particle swarm optimization. Eng. Optimiz. **39**(3), 287–305 (2007)
6. Petalas, Y.G., Parsopoulos, K.E., Vrahatis, M.N.: Memetic particle swarm optimization. Ann. Oper. Res. **156**(1), 99–127 (2007)
7. Silva E.K., Barbosa H.J.C., Lemonge A.C.C.: An adaptive constraint handling technique for differential evolution with dynamic use of variants in engineering optimization. Optimization and Engineering, **12**(1–2), 31–54 (2011)
8. Bertsekas, D.P.: Nonlinear Programming, 2nd edn. Athena Scientific, Belmont (1999)
9. Mezura-Montes, E., Coello, C.A.C.: Constraint-handling in nature-inspired numerical optimization: Past, present and future. Swarm Evolut. Comput. **1**(4), 173–194 (2011)
10. Xavier, A.E.: Hyperbolic penalty: a new method for nonlinear programming with inequalities. Intl. Trans. in Op. Res. **8**(6), 659–671 (2001)
11. Gonzaga, C.C., Castillo, R.A.: A nonlinear programming algorithm based on non-coercive penalty functions. Math. Program. Ser. A, **96**(1), 87–101 (2003)
12. Auslender, A., Cominetti, R., Haddou, M.: Asymptotic analysis for penalty and barrier methods in convex and linear programming. Math. Oper. Res. **22**(1), 43–62 (1997)
13. Di Pillo, G., Lucidi, S., Rinaldi, F.: An approach to constrained global optimization based on exact penalty functions. J. Glob. Optim. **54**(2), 251–260 (2012)
14. Rocha, A.M.A.C., Fernandes, E.M.G.P., Martins, T.F.M.C.: Novel fish swarm heuristics for bound constrained global optimzation problems. In: B. Murgante et al. (Eds.) Computational Science and Its Applications – ICCSA 2011, LNCS, Vol. 6784, Part III pp. 185–199 (2011)
15. Boussaïd, I., Lepagnot, J., Siarry, P.: A survey on optimization metaheuristics. Inf. Sci. **237**, 82–117 (2013)
16. Gendreau, M., Potvin, J.-Y. (Eds.): Handbook of Metaheuristics, 2nd edn. International Series in Operations Research and Management Science, Springer, (2010)
17. Sörensen, K.: Metaheuristics – the metaphor exposed. Intl. Trans. in Op. Res. **22**, 3–18 (2015)
18. J. D. Griffin, J.D., and T. G. Kolda, T.G.: Nonlinearly constrained optimization using heuristic penalty methods and asynchronous parallel generating set search. Appl. Math. Res. Express **2010**(1), 36–62 (2010)

19. Birgin, E.G., Floudas, C.A., Martínez, J.M.: Global minimization using an Augmented Lagrangian method with variable lower-level constraints. Math. Program. Ser. A, **125**(1), 139–162 (2010)
20. Holland, J.H.: Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, MI, USA, (1975)
21. Brest, J., Greiner, S., Bošković, B., Mernik, M., Žumer, V.: Self-adapting control parameters in differential evolution: a comparative study on numerical benchmark problems. IEEE Trans. Evol. Comput. **10**, 646–657 (2006)
22. Hansen, N., Ostermeier, A.: Completely derandomized self-adaptation in evolution strategies. Evol. Comput. **9**(2), 159–195 (2001)
23. Hansen, N.: The CMA evolution strategy: a comparing review. In Lozano, J.A., Larranaga, P., Inza, I., Bengoetxea, E. (Eds.), Towards a New Evolutionary Computation. Advances on Estimation of Distribution Algorithms, Springer, pp. 75–102 (2006)
24. Socha, K., Dorigo, M.: Ant colony optimization for continuous domains. Eur. J. Oper. Res. **185**(3), 1155–1173 (2008)
25. Kirkpatrick, S., Gelatt, C., Vecchi, M.: Optimization by simulated annealing. Science **220**, 671–680 (1983)
26. Jiang, M., Wang, Y., Pfletschinger, S., Lagunas, M.A., Yuan, D.: Optimal multiuser detection with artificial fish swarm algorithm. In D.-S. Huang, L. Heutte, M. Loog (Eds.) CCIS 2, ICIC 2007, Springer-Verlag, pp. 1084–1093 (2007)
27. Neshat, M., Sepidnam, G., Sargolzaei, M., Toosi, A.N.: Artificial fish swarm algorithm: a survey of the state-of-the-art, hybridization, combinatorial and indicative applications. Artif. Intell. Rev. **42**(4), 965–997 (2014)
28. Rocha, A.M.A.C., Martins, T.F.M.C., Fernandes, E.M.G.P.: An augmented Lagrangian fish swarm based method for global optimization. J. Comput. Appl. Math. **235**(16), 4611–4620 (2011)
29. Costa, M.F.P., Rocha, A.M.A.C., Fernandes, E.M.G.P.: An artificial fish swarm algorithm based hyperbolic augmented Lagrangian method. J. Comput. Appl. Math. **259**(Part B), 868–876 (2014)
30. Karr, A. F., Probability, Springer Texts in Statistics, Springer-Verlag, New York, (1993)
31. Rudolph, G.: Convergence of evolutionary algorithms in general search spaces. Proceedings of Third IEEE Conference on Evolutionary Computation, IEEE Press, NJ, pp. 50–54 (1996)
32. Liang, J.J., Runarsson, T.P., Mezura-Montes, E., Clerc, M., Suganthan, P.N., Coello Coello, C.A., Deb, C.: In: Problem Definition and Evolution Criteria for the CEC 2006 Special Session on Constrained Real-Parameter Optimization, IEEE Congress on Evolutionary Computation, Vancouver, Canada, 17–21 July (2006)
33. Derrac, J., García, S., Molina, D., Herrera, F.: A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. Swarm Evolut. Comput. **1**(1), 3–18 (2011)
34. Dolan, E.D., Moré, J.J.: Benchmarking optimization software with performance profiles. Math. Program. Ser. A, **91**(2), 201–213 (2002)
35. Deb, K., Srivastava, S.: A genetic algorithm based augmented Lagrangian method for constrained optimization. Comput. Optim. Appl. **53**(3), 869–902 (2012)
36. Rocha, A.M.A.C., Costa, M.F.P., Fernandes, E.M.G.P.: A shifted hyperbolic augmented Lagrangian-based artificial fish two-swarm algorithm with guaranteed convergence for constrained global optimization. Eng. Optimiz. **48**(12), 2114–2140 (2016)
37. Hooke, R., Jeeves, T.A.: Direct search solution of numerical and statistical problems. J. Ass. Comput. Mach. **8**(2), 212–229 (1961)