ENSEEIHT, INPT, rue Charles Camichel, B.P. 7122 31071, Toulouse Cedex 7, France (serge.gratton@enseeiht.fr).

L. N. Vicente
CMUC, Department of Mathematics, University of Coimbra, 3001-454 Coimbra, Portugal (lnv@mat.uc.pt).

Pré-Publicações do Departamento de Matemática
Universidade de Coimbra
Preprint Number 12–03

GLOBALLY CONVERGENT EVOLUTION STRATEGIES AND CMA-ES

Y. DIOUANE, S. GRATTON AND L. N. VICENTE

Abstract: In this paper we show how to modify a large class of evolution strategies (ES) to rigorously achieve a form of global convergence, meaning convergence to stationary points independently of the starting point. The type of ES under consideration recombine the parents by means of a weighted sum, around which the offsprings are computed by random generation. One relevant instance of such ES is CMA-ES.

The modifications consist essentially of the reduction of the size of the steps whenever a sufficient decrease condition on the function values is not verified. When such a condition is satisfied, the step size can be reset to the step size maintained by the ES themselves, as long as this latter one is sufficiently large. We suggest a number of ways of imposing sufficient decrease for which global convergence holds under reasonable assumptions, and extend our theory to the constrained case.

Given a limited budget of function evaluations, our numerical experiments have shown that the modified CMA-ES is capable of further progress in function values. Moreover, we have observed that such an improvement in efficiency comes without deteriorating the behavior of the underlying method in the presence of nonconvexity.

Keywords: Evolution strategy, global convergence, sufficient decrease, covariance matrix adaptation (CMA).
AMS Subject Classification (2010): 90C30, 90C56, 90C59.

1. Introduction

Evolution strategies (ES) form a class of evolutionary algorithms for the unconstrained optimization of a continuous function without using derivatives, originally developed in [19]. ES have been widely investigated and tested (see, e.g., [2]). However, as far as we know, there are no asymptotic results regarding the convergence of the iterates generated by ES to stationary points, at least without assuming the density of the sampling procedure in a given region or set. In this paper, we focus on a large class of ES where new parents are selected as the best previous offsprings, and new offsprings

Date: January 27, 2012.
Support for this research was provided by FCT under the grant PTDC/MAT/098214/2008.

are generated around a weighted mean of the previous parents. The paper focuses first on unconstrained optimization problems of the form $\min_{x\in\mathbb{R}^n} f(x)$, addressing the constrained case separately.

Derivative-free optimization [4], on the other hand, is a field of nonlinear optimization where methods that do not use derivatives have been developed and analyzed. There are essentially two classes of algorithms, model-based methods and direct search. However, both are rigorous in the sense that one can prove some form of convergence to stationarity, in a way that is independent of the initial choice for the iterates (a feature called global convergence in the field of nonlinear optimization). In addition, model-based and direct-search methods achieve global convergence based on the principle of rejecting steps that are too large and do not provide a certain decrease in the objective function value, retracting the search to a smaller region where the quality of the model or of the sampling eventually allows some progress.

The technique that we use to globalize ES resembles what is done in direct search. In particular, given the type of random sampling used in ES, our work is inspired by direct-search methods for nonsmooth functions, where one must use a set of directions asymptotically dense in the unit sphere [1, 20]. Since the random sampling of ES will not likely provide us any integer lattice underlying structure for the iterates (like in MADS [1]), we will use a sufficient decrease condition (as opposed to just a simple decrease) to accept new iterates and ensure global convergence. By a sufficient decrease we mean a decrease of the type $f(x_{k+1}) \leq f(x_k) - o(\sigma_k)$, where σ_k stands for the step size parameter and $o(\cdot)$ obeys some properties, in particular $o(\sigma)/\sigma \to 0$ when $\sigma \to 0$.

One way of imposing sufficient decrease in ES is to apply it directly to the sequence of weighted means. However, ES are population-based algorithms where a sample set of offsprings is generated at every iteration. Other forms of imposing this type of decrease also found globally convergent involve the maximum value of the best offsprings. In fact, requiring a sufficient decrease on the sequence of maximum best offspring values renders also a globally convergent algorithm. Alternatively, demanding this maximum value to sufficiently decrease the weighted mean one, not only leads also to global convergence but seems to produce the most efficient version among the three.

The paper is organized as follows. We first describe in Section 2 the class of evolution strategies (ES) to be considered. Then, in Section 3, we show how to modify such algorithms to enable them for global convergence. Section 4

References

[1] C. Audet and J. E. Dennis Jr. Mesh adaptive direct search algorithms for constrained optimization. *SIAM J. Optim.*, 17:188–217, 2006.

[2] H.-G. Beyer and H.-P. Schwefel. Evolution strategies: A comprehensive introduction. *Natural Computing*, 1:3–52, 2002.

[3] F. H. Clarke. *Optimization and Nonsmooth Analysis*. John Wiley & Sons, New York, 1983. Reissued by SIAM, Philadelphia, 1990.

[4] A. R. Conn, K. Scheinberg, and L. N. Vicente. *Introduction to Derivative-Free Optimization*. MPS-SIAM Series on Optimization. SIAM, Philadelphia, 2009.

[5] E. D. Dolan and J. J. Moré. Benchmarking optimization software with performance profiles. *Math. Program.*, 91:201–213, 2002.

[6] E. D. Dolan, J. J. Moré, and T. S. Munson. Optimality measures for performance profiles. *SIAM J. Optim.*, 16:891–909, 2006.

[7] N. I. M. Gould, D. Orban, and P. L. Toint. CUTEr, a Constrained and Unconstrained Testing Environment, revisited. *ACM Trans. Math. Software*, 29:373–394, 2003.

[8] N. Hansen. The CMA Evolution Strategy: A Tutorial. June 28, 2011.

[9] N. Hansen, S. Fincky, R. Rosz, and A. Auger. Real-parameter black-box optimization benchmarking 2010: Noisy functions definitions. Technical report, March 22, 2010.

[10] N. Hansen, S. Fincky, R. Rosz, and A. Auger. Real-parameter black-box optimization benchmarking 2010: Noiseless functions definitions. Technical report, September 28, 2010.

[11] N. Hansen and A. Ostermeier. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In *Proceedings of the 1996 IEEE International Conference on Evolutionary Computation*, pages 312–317, 1996.

[12] N. Hansen, A. Ostermeier, and A. Gawelczyk. On the adaptation of arbitrary normal mutation distributions in evolution strategies: The generating set adaptation. In L. Eshelman, editor, *Proceedings of the Sixth International Conference on Genetic Algorithms, Pittsburgh*, pages 57–64, 1995.

[13] J. Jahn. *Introduction to the Theory of Nonlinear Optimization*. Springer-Verlag, Berlin, 1996.

[14] T. G. Kolda, R. M. Lewis, and V. Torczon. Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Rev.*, 45:385–482, 2003.

[15] R. M. Lewis and Virginia Torczon. Pattern search methods for linearly constrained minimization. *SIAM J. Optim.*, 10:917–941, 2000.

[16] S. Lucidi, M. Sciandrone, and P. Tseng. Objective-derivative-free methods for constrained optimization. *Math. Program.*, 92:37–59, 2002.

[17] J. J. Moré and S. M. Wild. Benchmarking derivative-free optimization algorithms. *SIAM J. Optim.*, 20:172–191, 2009.

[18] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer-Verlag, Berlin, second edition, 2006.

[19] I. Rechenberg. *Evolutionsstrategie: Optimierung Technischer Systeme nach Prinzipien der Biologischen Evolution*. Frommann-Holzboog, 1973.

[20] L. N. Vicente and A. L. Custódio. Analysis of direct searches for discontinuous functions. *Math. Program.*, 2011, to appear.

Y. Diouane
CERFACS, 42 Avenue Gaspard Coriolis, 31057 Toulouse Cedex 01, France (youssef.diouane@gmail.com).

S. Gratton

same value as CMA-ES. However, Figures 9(b) and 10(b)), which plot the average number of objective function evaluations taken, show that the effort of the max/mean version was all together considerably lower.

7. Conclusions and future work

We have seen that it is possible to modify ES so that they converge to stationary points without any assumption on the starting mean. The modified versions of ES promote smaller steps when the larger steps are uphill and thus lead to an improvement in the efficiency of the algorithms in the search of a stationary point. The so-called max/mean version, where the step is reduced whenever the maximum objective value of the trial offsprings does not sufficiently reduce the objective value at the current weighted mean, has emerged as the best modified version in our numerical experiments. Such a behavior seems related to the fact that it is the max/mean version the one where unsuccessful iterations can more easily occur (when compared to the mean/mean and max/max versions). Apparently, this promotion of smaller, better steps has not jeopardize the search for the global minimizer in nonconvex problems, although one probably needs further experiments to be totally sure about such a statement.

Our approach applies to all ES of the type considered in this paper (see Section 2) although we only used CMA-ES in our numerical tests. A number of issues regarding the interplay of our ES modifications (essentially the step size update based on different sufficient decrease conditions) and the CMA scheme to update the covariance matrix and corresponding step size must be better understood and investigated. In addition, we have not explored to our benefit any hidden ability of the CMA scheme to approximate or predict first or second order information (which might be used in the sufficient decrease conditions or to guide the offspring generation).

The treatment of constraints was certainly preliminary but revealed that one can extend our convergence theory to both the cases where the derivatives of the functions defining the constraints are known or unknown. We leave also to the future a full treatment of the constrained case, in particular how can one randomly generate the trial offsprings with a sample geometry that conforms to the constraints, in particular when the constraints are linear or consist of bounds on the variables.

is devoted to the analysis of global convergence of the modified ES versions. The constrained case is covered in Section 5. Our numerical experiments comparing the different modified versions of CMA-ES [11, 12] are described in Section 6. Finally, in Section 7, we draw some conclusions and describe future work.

2. A class of evolution strategies

Our working class of evolution strategies (ES) is referred to as being of the type $(\mu/\mu_W, \lambda)$–ES, in other words, it iterates using m_μ parents and m_λ offsprings (with $m_\lambda \geq m_\mu$), recombining all the m_μ parents by means of a weighted sum. The m_λ offsprings are computed by random generation around the weighted mean of the m_μ chosen parents. The parents, in turn, are selected as the best m_μ offsprings in the value of the objective function f. The weights used to compute the means belong to the simplex set $S = \{(\omega^1, \ldots, \omega^{m_\mu}) \in \mathbb{R}^{m_\mu} : \sum_{i=1}^{m_\mu} w^i = 1, w^i \geq 0, i = 1, \ldots, m_\mu\}$. The algorithmic description of such class of ES is given below.

Algorithm 2.1. A Class of Evolution Strategies

Initialization: Choose positive integers m_λ and m_μ such that $m_\lambda \geq m_\mu$. Choose an initial weighted mean x_0, an initial step length $\sigma_0^{\text{ES}} > 0$, an initial distribution $\mathcal{C}_0$, and initial weights $(\omega_0^1, \ldots, \omega_0^{m_\mu}) \in S$. Set $k = 0$.

Until some stopping criterion is satisfied:

1. **Offspring Generation:** Compute new sample points $Y_{k+1} = \{y_{k+1}^1, \ldots, y_{k+1}^{m_\lambda}\}$ such that

$$y_{k+1}^i \;=\; x_k + \sigma_k^{\text{ES}} d_k^i,$$

   where d_k^i is drawn from the distribution $\mathcal{C}_k$, $i = 1, \ldots, m_\lambda$.

2. **Parent Selection:** Evaluate $f(y_{k+1}^i)$, $i = 1, \ldots, m_\lambda$, and reorder the offspring points in $Y_{k+1} = \{\tilde{y}_{k+1}^1, \ldots, \tilde{y}_{k+1}^{m_\lambda}\}$ by increasing order: $f(\tilde{y}_{k+1}^1) \leq \cdots \leq f(\tilde{y}_{k+1}^{m_\lambda})$.

   Select the new parents as the best m_μ offspring sample points $\{\tilde{y}_{k+1}^1, \ldots, \tilde{y}_{k+1}^{m_\mu}\}$, and compute their weighted mean

$$x_{k+1} \;=\; \sum_{i=1}^{m_\mu} \omega_k^i \tilde{y}_{k+1}^i.$$

3. Updates: Update the step length σ_{k+1}^{ES}, the distribution $\mathcal{C}_{k+1}$, and the weights $(\omega_{k+1}^1, \ldots, \omega_{k+1}^{m_\mu}) \in S$. Increment k and return to Step 1.

3. A class of ES provably global convergent

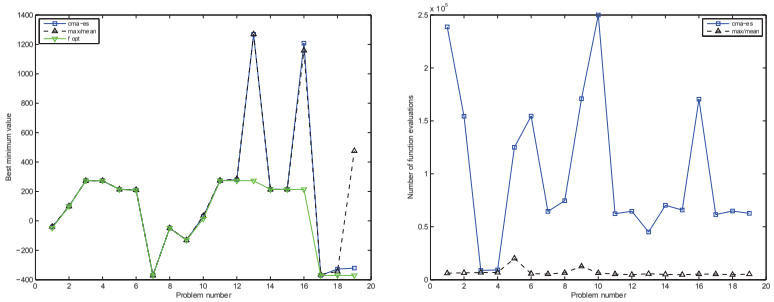
The main question we address in this paper is how to change Algorithm 2.1, in a minimal way, to make it enjoying some form of convergence properties. We will target at global convergence in the sense of nonlinear optimization, in other words we would like to prove some limit form of stationarity for any output sequence of iterates generated by the algorithm, and we would like to do this independently of the starting point.

The modifications to the algorithm will be essentially two, and they have been widely used in the field of nonlinear optimization, with and without derivatives. First we need to control the size of the steps taken, and thus we will update separately a step size parameter σ_k, letting it take the value of σ_k^{ES} whenever possible. Controlling the step size is essential as we know that most steps used in nonlinear optimization are too large away from stationarity. Secondly we need to impose some form of sufficient decrease on the objective function values to be able to declare an iteration successful and thus avoiding a step size reduction. These two techniques, step size update and imposition of sufficient decrease on the objective function values, are thus closely related since an iteration is declared unsuccessful and the step size reduced when the sufficient decrease condition is not satisfied. Moreover, this condition involves a function $\rho(\sigma_k)$ of the step size σ_k, where $\rho(\cdot)$ is a forcing function, i.e., a positive, nondecreasing function defined in $\mathbb{R}^+$ such that $\rho(t)/t \to 0$ when $t \downarrow 0$ (one can think for instance of $\rho(t) = t^2$).

Since Algorithm 2.1 evaluates the objective function at the offspring sample points but then computes new points around a weighted sum of the parents selected, it is not clear how one does impose sufficient decrease. In fact, there are several ways of proceeding. A first possibility (denoted by mean/mean) is to require the weighted means to sufficiently decrease the objective function, see the inequality (2) below, which obviously requires an extra function evaluation per iteration.

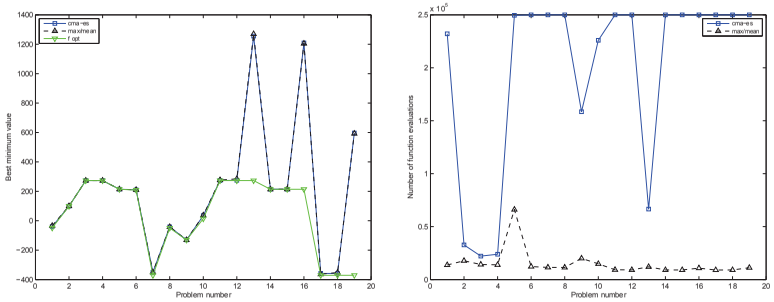
A second possibility to impose sufficient decrease (referred to as max/max), based entirely on the objective function values already computed for the parent samples, is to require the maximum of these values to be sufficiently decreased, see the inequality (3). Then, it would immediately occur to combine these first two possibilities, asking the new maximum value to reduce

large and the tolerances extremely small since we are interested in observing the asymptotic ability to determine a global minimum.



(a) Best function values (in average). (b) Number of function evaluations taken (in average).

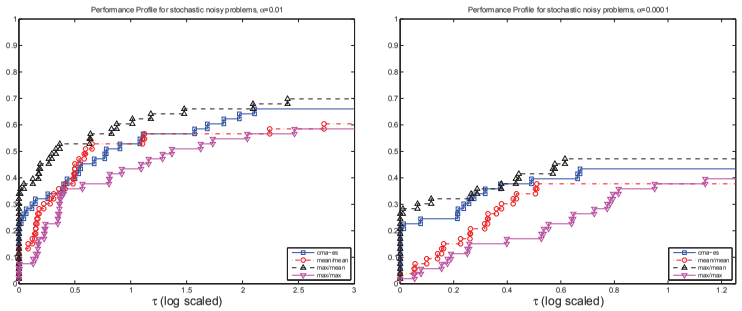
FIGURE 9. Results for the max/mean version and CMA-ES on a set of multi-modal functions of dimension 10.



(a) Best function values (in average). (b) Number of function evaluations taken (in average).

FIGURE 10. Results for the max/mean version and CMA-ES on a set of multi-modal functions of dimension 20.

Figures 9(a) and 10(a) show the averaged best objective value obtained by both the max/mean version and by the unmodified CMA-ES, as well as the global optimal value, for all the 19 problems. Except for the last problem (in dimension 10), the max/mean version seemed to have reached roughly the



(a) Accuracy level of 10^{-2}. (b) Accuracy level of 10^{-4}.

FIGURE 8. Performance profiles computed for the set of stochastic noisy problems with a logarithmic scale, considering the two levels of accuracy, 10^{-2} and 10^{-4}.

6.6. Some global optimization experiments. In this section we are interested in assessing the impact of our modifications on the ability of CMA-ES to identify the global minimum on nonconvex problems with a high number of different local minimizers.

We recall that the max/mean version exhibited the best performance among the three modified versions of CMA-ES on the test set mentioned in Section 6.3. Therefore in this section we will report a comparison of CMA-ES only against this version.

The test set is now composed of the 19 highly multi-modal problems used in [9, 10], being the last 9 noisy. We selected dimensions $n = 10$ and $n = 20$. For each dimension and using a large maximal computational budget, we ran our max/mean CMA-ES version and unmodified CMA-ES using 20 different starting points randomly chosen. We then computed the mean of all the 20 'optimal' values found for each each algorithm as well as the respective number of function evaluations taken.

Each run was ended when the stopping criterion of CMA-ES is achieved (which includes finding a function value below a certain fitness value, for our problems chosen as $f_* + 10^{-10}$, where f_* is the optimal value of the corresponding problem), when the number of function evaluations reaches 250000, and when σ_k becomes smaller than 10^{-10}. The budget is therefore extremely

sufficiently the value of the previous mean or, vice-versa, requiring the value of the new mean to reduce sufficiently the previous maximum. The lack of theoretical support of the latter possibility made us consider only the first one, called max/mean, see the inequality (4).

Version mean/mean is clear in the sense that imposes the sufficient decrease condition directly on the function values computed at the sequence of minimizer candidates, the weighted sums. It is also around these weighted sums that new points are randomly generated. Versions max/max and mean/max, however, operate based or partially based on the function values at the parents samples (on the maximum of those). Thus, in these two versions, one needs a mechanism to balance the function values at the parents samples and the function value at the weighted sum. Such a balance is unnecessary when the objective function is convex. In general we need a condition of the form (1) below.

The modified form of the ES of Algorithm 3.1 is described below. Note that one also imposes bounds on the all directions d_k^i used by the algorithm. This modification is, however, very mild since the upper bound $d_{\min}$ can be chosen very close to zero and the upper bound set to a very large number. Moreover, one can think of working always with normalized directions which entirely removes any concern.

Algorithm 3.1. A class of ES provably global convergent (versions mean/mean, max/max, and max/mean)

Initialization: Choose positive integers m_λ and m_μ such that $m_\lambda \geq m_\mu$. Select an initial weighted mean x_0, evaluate $f(x_0)$ in versions mean/mean and max/mean, and set $x_0^{m_\mu} = x_0$ for max/max. Choose initial step lengths $\sigma_0, \sigma_0^{\text{ES}} > 0$, an initial distribution $\mathcal{C}_0$, and initial weights $(\omega_0^1, \ldots, \omega_0^{m_\mu}) \in S$. Choose constants $\beta_1, \beta_2, d_{\min}, d_{\max}$ such that $0 < \beta_1 \leq \beta_2 < 1$ and $0 < d_{\min} < d_{\max}$. Select a forcing function $\rho(\cdot)$ and $\theta \in (0, 1)$. Set $k = 0$.

Until some stopping criterion is satisfied:

1. Offspring Generation: Compute new sample points $Y_{k+1} = \{y_{k+1}^1, \ldots, y_{k+1}^{m_\lambda}\}$ such that

$$y_{k+1}^i = x_k + \sigma_k d_k^i,$$

where d_k^i is drawn from the distribution $\mathcal{C}_k$ and obeys $d_{\min} \leq \|d_k^i\| \leq d_{\max}$, $i = 1, \ldots, m_\lambda$.

2. Parent Selection: Evaluate $f(y^i_{k+1})$, $i = 1, \ldots, m_\lambda$, and reorder the offspring points in $Y_{k+1} = \{\tilde{y}^1_{k+1}, \ldots, \tilde{y}^{m_\lambda}_{k+1}\}$ by increasing order: $f(\tilde{y}^1_{k+1}) \leq \cdots \leq f(\tilde{y}^{m_\lambda}_{k+1})$.

Select the new parents as the best m_μ offspring sample points $\{\tilde{y}^1_{k+1}, \ldots, \tilde{y}^{m_\mu}_{k+1}\}$, and compute their weighted mean

$$x^{trial}_{k+1} = \sum_{i=1}^{m_\mu} \omega^i_k \tilde{y}^i_{k+1}.$$

Evaluate $f(x^{trial}_{k+1})$. In versions max/max and max/mean, re-update the weights, if necessary, such that $(\omega^1_k, \ldots, \omega^{m_\mu}_k) \in S$ and

$$\sum_{i=1}^{m_\mu} \omega^i_k [f(x^{trial}_{k+1}) - f(\tilde{y}^i_{k+1})] \leq \theta\rho(\sigma_k), \quad \theta \in (0,1). \tag{1}$$

3. Imposing Sufficient Decrease:

If (version mean/mean)

$$f(x^{trial}_{k+1}) \leq f(x_k) - \rho(\sigma_k), \tag{2}$$

or (version max/max)

$$f(\tilde{y}^{m_\mu}_{k+1}) \leq f(x^{m_\mu}_k) - \rho(\sigma_k), \tag{3}$$

or (version max/mean)

$$f(\tilde{y}^{m_\mu}_{k+1}) \leq f(x_k) - \rho(\sigma_k), \tag{4}$$

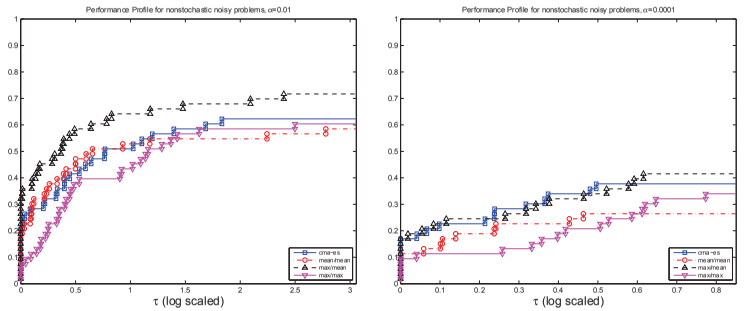
then consider the iteration successful, set $x_{k+1} = x^{trial}_{k+1}$, and $\sigma_{k+1} \geq \sigma_k$ (for example $\sigma_{k+1} = \max\{\sigma_k, \sigma^{\text{ES}}_k\}$).

Set $x^{m_\mu}_{k+1} = \tilde{y}^{m_\mu}_{k+1}$ in version max/max.

Otherwise, consider the iteration unsuccessful, set $x_{k+1} = x_k$ (and $x^{m_\mu}_{k+1} = x^{m_\mu}_k$ for max/max) and $\sigma_{k+1} = \beta_k\sigma_k$, with $\beta_k \in (\beta_1, \beta_2)$.

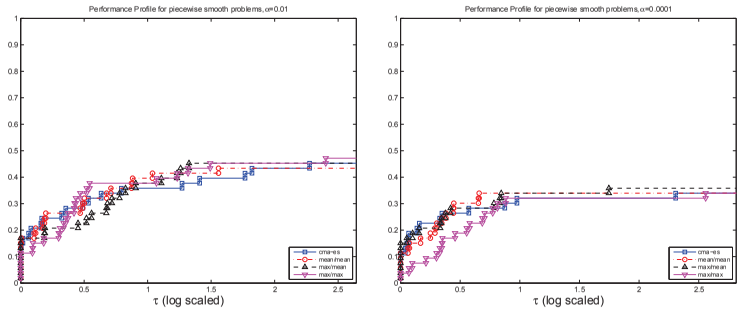
4. ES Updates: Update the ES step length σ^{ES}_{k+1}, the distribution $\mathcal{C}_k$, and the weights $(\omega^1_{k+1}, \ldots, \omega^{m_\mu}_{k+1}) \in S$. Increment k and return to Step 1.

One can see that the imposition of (1) may cost additional function evaluations per iteration. However, this condition can always be guaranteed since the ultimate choice $\omega^1_k = 1$, and $\omega^i_k = 0$, $i = 2, \ldots, m_\mu$, trivially satisfies it.



(a) Accuracy level of 10^{-2}. (b) Accuracy level of 10^{-4}.

FIGURE 6. Performance profiles computed for the set of non-stochastic noisy problems with a logarithmic scale, considering the two levels of accuracy, 10^{-2} and 10^{-4}.



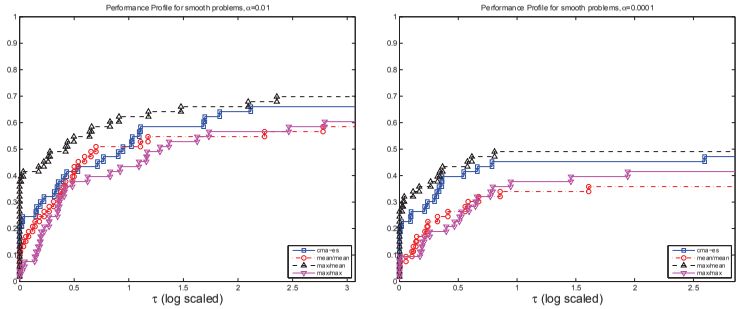
(a) Accuracy level of 10^{-2}. (b) Accuracy level of 10^{-4}.

FIGURE 7. Performance profiles computed for the set of piecewise smooth problems with a logarithmic scale, considering the two levels of accuracy, 10^{-2} and 10^{-4}.

In all types of problems and levels of accuracy considered and by looking at the profiles for large values of τ, one observes that the version max/mean is the most robust one. The other two modified versions of CMA-ES (max/max and mean/mean) seem to be less robust than the pure version in terms of efficiency.

case, it makes more sense not to select the accuracy level too small, and our tests were performed with $\alpha = 10^{-2}, 10^{-4}$.

We now examine the performance of CMA-ES and of our three modified versions on the four types of problems $\mathcal{P}$ mentioned in Section 6.3. Figures 5–8 report performance profiles obtained for the four types of problems, considering the two different levels of accuracy, $\alpha = 10^{-2}$ and $\alpha = 10^{-4}$ (Figure 5: smooth problems; Figure 6: nonstochastic noisy problems; Figure 7: piecewise smooth problems; Figure 8: stochastic noisy problems) and a maximum of 1500 function evaluations.



(a) Accuracy level of 10^{-2}. (b) Accuracy level of 10^{-4}.

FIGURE 5. Performance profiles computed for the set of smooth problems with a logarithmic scale, considering the two levels of accuracy, 10^{-2} and 10^{-4}.

On the smooth and stochastic noisy problems, the performance profiles for the lower accuracy level ($\alpha = 10^{-2}$) show that max/mean version is the fastest solvers in approximately 40% of the problems. For $\alpha = 10^{-4}$, differences in efficiency are not so visible when compared to the pure version.

The performance profiles for the nonstochastic noisy problems show that the max/mean version is the fastest solver for $\alpha = 10^{-2}$, with differences getting tighter for $\alpha = 10^{-4}$.

For the piecewise smooth problems, the performance profiles show that the four solvers exhibit a more similar behavior in terms of efficiency (perhaps with the exception of the max/max version which performs clearly worse for $\alpha = 10^{-4}$).

4. Convergence

Under appropriate assumptions we will now prove global convergence of the modified versions of the considered class of ES (again, by global convergence, we mean some form of limit first-order stationary for arbitrary starting points).

As we have seen before, an iteration is considered successful only if it produces a point that has sufficiently decreased some value of f. Insisting on a sufficient decrease will guarantee that a subsequence of step sizes will converge to zero. In fact, since $\rho(\sigma_k)$ is a monotonically increasing function of the step size σ_k, we will see that such a step size cannot be bounded away from zero since otherwise some value of f would tend to $-\infty$. Imposing sufficient decrease will make it harder to have a successful step and therefore will generate more unsuccessful poll steps. We start thus by showing that there is a subsequence of iterations for which the step size parameter σ_k tends to zero.

Lemma 4.1. *Consider a sequence of iterations generated by Algorithm 3.1 without any stopping criterion. Let f be bounded below. Then* $\liminf_{k\to+\infty} \sigma_k = 0$.

Proof: Suppose that there exists a $\sigma > 0$ such that $\sigma_k > \sigma$ for all k. If there is an infinite number of successful iterations, this leads to a contradiction to the fact that f is bounded below.

In fact, since ρ is a nondecreasing, positive function, $\rho(\sigma_k) \geq \rho(\sigma) > 0$. Let us consider the three versions separately. In the version mean/mean, we obtain $f(x_{k+1}) \leq f(x_k) - \rho(\sigma)$ for all k, which obviously contradicts the boundedness below of f. In the version max/max, we obtain $f(x_{k+1}^{m_\mu}) \leq f(x_k^{m_\mu}) - \rho(\sigma)$ for all k, which also trivially contradicts the boundedness below of f. For the max/mean version, one has

$$f(\tilde{y}_{k+1}^i) \;\leq\; f(x_{k+1}^{m_\mu}) \;\leq\; f(x_k) - \rho(\sigma_k), \quad i = 1, \ldots, m_\mu.$$

Thus, multiplying these inequalities by the weights ω_k^i, $i = 1, \ldots, m_\mu$, and adding them up, lead us to

$$\sum_{i=1}^{m_\mu} \omega_k^i f(\tilde{y}_{k+1}^i) \;\leq\; f(x_k) - \rho(\sigma_k),$$

and from condition (1) imposed on the weights in Step 2 of Algorithm 3.1, we obtain

$$f(x_{k+1}) \leq f(x_k) + (\theta - 1)\rho(\sigma_k), \quad \theta \in (0,1),$$

and the contradiction is also easily reached.

The proof is thus completed if there is an infinite number of successful iterations. However, if no more successful iterations occur after a certain order, then this also leads to a contradiction. The conclusion is that one must have a subsequence of iterations driving σ_k to zero. ■

From the fact that σ_k is only reduced in unsuccessful iterations and by a factor not approaching zero, one can then conclude the following.

Lemma 4.2. *Consider a sequence of iterations generated by Algorithm 3.1 without any stopping criterion. Let f be bounded below.*

There exists a subsequence K of unsuccessful iterates for which $\lim_{k\in K} \sigma_k = 0$.

If the sequence $\{x_k\}$ is bounded, then there exists an x_* and a subsequence K of unsuccessful iterates for which $\lim_{k\in K} \sigma_k = 0$ *and* $\lim_{k\in K} x_k = x_*$.

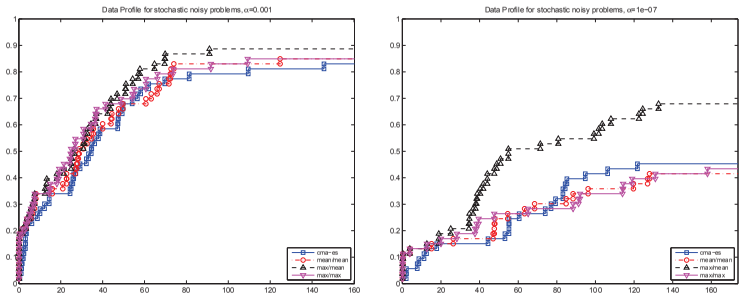
Proof: From Lemma 4.1, there must exist an infinite subsequence K of unsuccessful iterates for which σ_{k+1} goes to zero. In a such case we have $\sigma_k = (1/\beta_k)\sigma_{k+1}$, $\beta_k \in (\beta_1, \beta_2)$, and $\beta_1 > 0$, and thus $\sigma_k \to 0$, for $k \in K$, too.

The second part of the lemma is also easily proved by extracting a convergent subsequence of the subsequence K of the first part for which x_k converges to x_*. ■

The above lemma ensures under mild conditions the existence of convergent subsequences of unsuccessful iterations for which the step size tends to zero. Such type of subsequences have been called refining [1]. The global convergence results are then extracted from refining subsequences. One will assume that the function f is Lipschitz continuous near the limit point x_* of a refining subsequence, so that the Clarke generalized derivative [3]

$$f^\circ(x_*; d) \;=\; \limsup_{x \to x_*, t \downarrow 0} \frac{f(x+td) - f(x)}{t}$$

exists for all $d \in \mathbb{R}^n$. The point x_* is then Clarke stationary if $f^\circ(x_*; d) \geq 0$, $\forall d \in \mathbb{R}^n$. Our first global convergence result concerns only the mean/mean version.



(a) Accuracy level of 10^{-3}. (b) Accuracy level of 10^{-7}.

FIGURE 4. Data profiles computed for the set of stochastic noisy problems, considering the two levels of accuracy, 10^{-3} and 10^{-7}.

The performance profile of a solver $s \in \mathcal{S}$ is then defined as the fraction of problems where the performance ratio is at most τ, that is,

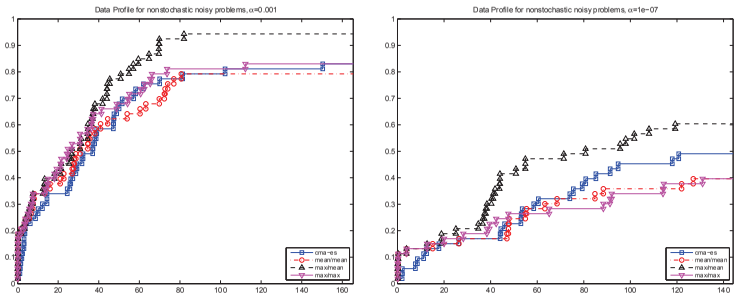
$$\rho_s(\tau) \;=\; \frac{1}{|\mathcal{P}|} \text{size}\{p \in \mathcal{P} : r_{p,s} \leq \tau\},$$

where $|\mathcal{P}|$ denotes the cardinality of $\mathcal{P}$. Performance profiles seek to capture how well the solver $s \in \mathcal{S}$ performs relatively to the others in $\mathcal{S}$ for all the problems in $\mathcal{P}$. Note, in particular, that $\rho_s(1)$ is the fraction of problems for which solver $s \in \mathcal{S}$ performs the best (efficiency), and that for τ sufficiently large, $\rho_s(\tau)$ is the fraction of problems solved by $s \in \mathcal{S}$ (robustness). In general, $\rho_s(\tau)$ is the fraction of problems with a performance ratio $r_{p,s}$ bounded by τ, and thus solvers with higher values for $\rho_s(\tau)$ are preferable.

It was suggested in [6] to use the same (scale invariant) convergence test for all solvers compared using performance profiles. The convergence test used in our experiments was

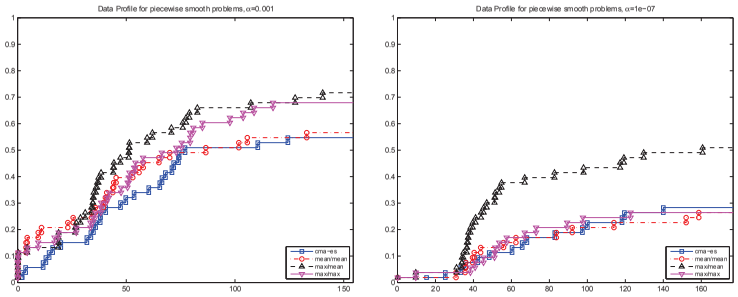
$$f(x) - f_* \;\leq\; \alpha(|f_*| + 1),$$

where α is an accuracy level and f_* is an approximation for the optimal value of the problem being tested. The convention $r_{p,s} = +\infty$ is used when the solver s fails to satisfy the convergence test on problem p. We computed f_* as the best objective function value found by the four CMA-ES solvers using an extremely large computational budget (a number of function evaluations equal to 500000). Thus, in this case, and as opposed to the data profiles



(a) Accuracy level of 10^{-3}. (b) Accuracy level of 10^{-7}.

FIGURE 2. Data profiles computed for the set of nonstochastic noisy problems, considering the two levels of accuracy, 10^{-3} and 10^{-7}.



(a) Accuracy level of 10^{-3}. (b) Accuracy level of 10^{-7}.

FIGURE 3. Data profiles computed for the set of piecewise smooth problems, considering the two levels of accuracy, 10^{-3} and 10^{-7}.

by

$$r_{p,s} \;=\; \frac{t_{p,s}}{\min\{t_{p,s} : s \in \mathcal{S}\}}.$$

Theorem 4.1. *Consider the version mean/mean and let* $a_k = \sum_{i=1}^{m_\mu} \omega_k^i d_k^i$. *Let* x_* *be the limit point of a subsequence of unsuccessful iterates* $\{x_k\}_K$ *for which* $\lim_{k\in K} \sigma_k = 0$. *Assume that* f *is Lipschitz continuous near* x_* *with constant* $\nu > 0$.

If d *is a limit point of* $\{a_k/\|a_k\|\}_K$, *then* $f^\circ(x_*; d) \geq 0$.

If the set of limit points $\{a_k/\|a_k\|\}_K$ *is dense in the unit sphere, then* x_* *is a Clarke stationary point.*

Proof: Let d be a limit point of $\{a_k/\|a_k\|\}_K$. Then it must exist a subsequence of K' of K such that $a_k/\|a_k\| \to d$ on K'. On the other hand, we have for all k that

$$x_{k+1} \;=\; \sum_{i=1}^{m_\mu} \omega_k^i \tilde{y}_{k+1}^i \;=\; x_k + \sigma_k \sum_{i=1}^{m_\mu} \omega_k^i d_k^i \;=\; x_k + \sigma_k a_k,$$

and, for $k \in K$,

$$f(x_k + \sigma_k a_k) \;>\; f(x_k) - \rho(\sigma_k).$$

Also, since the directions d_k^i and the weights are bounded above for all k and i, a_k is bounded above for all k, and so $\sigma_k\|a_k\|$ tends to zero when σ_k does.

Thus, from the definition of the Clarke generalized derivative,

$$\begin{aligned} f^\circ(x_*; d) \;&=\; \limsup_{x\to x_*, t\downarrow 0} \frac{f(x+td) - f(x)}{t} \\ &\geq\; \limsup_{k\in K'} \frac{f(x_k + \sigma_k\|a_k\|(a_k/\|a_k\|)) - f(x_k)}{\sigma_k\|a_k\|} - r_k, \end{aligned}$$

where, from the Lipschitz continuity of f near x_*,

$$r_k \;=\; \frac{f(x_k + \sigma_k a_k) - f(x_k + \sigma_k\|a_k\|d)}{\sigma_k\|a_k\|} \;\leq\; \nu \left\| \frac{a_k}{\|a_k\|} - d \right\|$$

tends to zero on K'. Finally,

$$\begin{aligned} f^\circ(x_*; d) \;&\geq\; \limsup_{k\in K'} \frac{f(x_k + \sigma_k a_k) - f(x_k) + \rho(\sigma_k)}{\sigma_k\|a_k\|} - \frac{\rho(\sigma_k)}{\sigma_k\|a_k\|} - r_k \\ &=\; \limsup_{k\in K'} \frac{f(x_k + \sigma_k a_k) - f(x_k) + \rho(\sigma_k)}{\sigma_k\|a_k\|} \\ &\geq\; 0. \end{aligned}$$

Since the Clarke generalized derivative $f^\circ(x_*;\cdot)$ is continuous in its second argument [3], it is then evident that if the set of limit points $\{a_k/\|a_k\|\}_K$ is dense in the unit sphere, $f^\circ(x_*;d) \geq 0$ for all $d \in \mathbb{R}^n$. ■

When f is strict differentiable at x_* (in the sense of Clarke [3], meaning that there exists $\nabla f(x_*)$ such that $f^\circ(x_*;d) = \langle \nabla f(x_*), d\rangle$ for all d) we immediately conclude that $\nabla f(x_*) = 0$.

A question that arises from Theorem 4.1 concerns the density of the a_k's in the unit sphere. First, we should point out that what we assume regards any refining subsequence K and not the whole sequence of iterates, but such a strengthening of the assumptions on the density of the directions seems necessary for these type of directional methods (see [1, 20]). Regarding the issue of the sum being dense in the unit sphere, notice that if, for instance, all the d_k^i are independently normally distributed, then a linear combination of them will also be normally distributed.

Now we prove global convergence for the two other versions (max/max and max/mean).

Theorem 4.2. *Consider the versions max/max and max/mean. Let x_* be the limit point of a subsequence of unsuccessful iterates $\{x_k\}_K$ for which $\lim_{k\in K} \sigma_k = 0$. Assume that f is Lipschitz continuous near x_* with constant $\nu > 0$.*

If d is a limit point of $\{d_k^{i_k}/\|d_k^{i_k}\|\}_K$, where $i_k \in \operatorname{argmax}_{1\leq i\leq m_\mu} f(\tilde{y}_{k+1}^i)$, then $f^\circ(x_*;d) \geq 0$.

If, for each $i \in \{1,\ldots,m_\mu\}$, the set of limit points $\{d_k^i/\|d_k^i\|\}_K$ is dense in the unit sphere, then x_* is a Clarke stationary point.

Proof: The proof follows the same lines of the proof of the mean/mean version. In the max/max case, one departs from the inequality that is true when $k \in K$,

$$f(x_{k+1}^{m_\mu}) \;>\; f(x_k^{m_\mu}) - \rho(\sigma_k),$$

which implies for a certain i_k

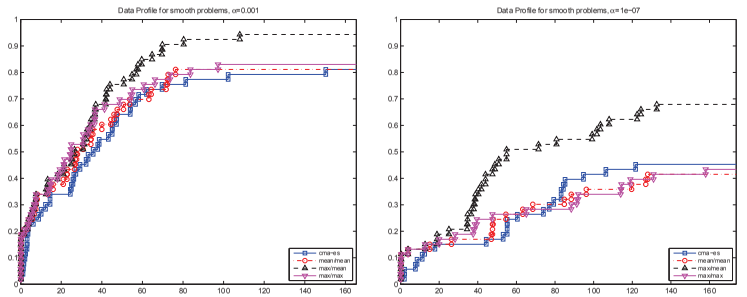
$$f(\tilde{y}_{k+1}^{i_k}) \;=\; f(x_{k+1}^{m_\mu}) \;>\; f(x_k^{m_\mu}) - \rho(\sigma_k) \;\geq\; f(\tilde{y}_k^i) - \rho(\sigma_k), \quad i = 1,\ldots,m_\mu.$$

Multiplying these inequalities by the weights ω_{k-1}^i, $i = 1,\ldots,m_\mu$, and adding them up implies

$$f(\tilde{y}_{k+1}^{i_k}) \;>\; \sum_{i=1}^{m_\mu} \omega_{k-1}^i f(\tilde{y}_k^i) - \rho(\sigma_k),$$

problems where the evaluation of the objective function is expensive. As for the levels of accuracy, we chose two values, $\alpha = 10^{-3}$ and $\alpha = 10^{-7}$. Since the best objective value f_L is chosen as the best value found by all solvers under the maximal computational budget, it makes some sense to consider a high accuracy level (like 10^{-7} or less).

We compared the results obtained with the four versions of CMA-ES (pure, mean/mean, max/max, and max/mean). Figures 1–4 report the data profiles obtained for the four types of problems, considering the two different levels of accuracy, $\alpha = 10^{-3}$ and $\alpha = 10^{-7}$ (Figure 1: smooth problems; Figure 2: nonstochastic noisy problems; Figure 3: piecewise smooth problems; Figure 4: stochastic noisy problems).



(a) Accuracy level of 10^{-3}. (b) Accuracy level of 10^{-7}.

FIGURE 1. Data profiles computed for the set of smooth problems, considering the two levels of accuracy, 10^{-3} and 10^{-7}.

Among our three modified versions of CMA-ES (mean/mean, max/max, and max/mean), the max/mean one performed clearly better than the pure one. The remaining two versions (mean/mean and max/max) did not overcome the pure one but performed competitively.

6.5. Results using performance profiles. Performance profiles [5] are defined in terms of a performance measure $t_{p,s} > 0$ obtained for each problem $p \in \mathcal{P}$ and solver $s \in \mathcal{S}$. For example, this measure could be based on the amount of computing time or the number of function evaluations required to satisfy a convergence test. Larger values of $t_{p,s}$ indicate worse performance. For any pair (p,s) of problem p and solver s, the performance ratio is defined

standard starting point tends to be too close to a solution for many of the problems.

The test set $\mathcal{P}$ is then formed by 53 different problems. No problem is overrepresented in $\mathcal{P}$ in the sense that no function k_p appears more than six times. Moreover, no pair (k_p, n_p) appears more than twice. In all cases,

$$2 \leq n_p \leq 12, \;\; 2 \leq m_p \leq 65, \;\; p = 1, \ldots, 53,$$

with $n_p \leq m_p$. For other details see [17].

The test problems have been considered in four different types, each having 53 instances: smooth (least squares problems obtained from applying the ℓ_2 norm to the vector functions); nonstochastic noisy (obtained by adding oscillatory noise to the smooth ones); piecewise smooth (as in the smooth case but using the ℓ_1 norm instead); stochastic noisy (obtained by adding random noise to the smooth ones).

6.4. Results using data profiles. To compare our modified CMA-ES versions to the pure one, we chose to work first with data profiles [17] for derivative-free optimization. Data profiles show how well a solver performs, given some computational budget, when asked to reach a specific reduction in the objective function value, measured by

$$f(x_0) - f(x) \;\geq\; (1-\alpha)[f(x_0) - f_L],$$

where $\alpha \in (0,1)$ is the level of accuracy, x_0 is the initial iterate, and f_L is the best objective value found by all solvers tested for a specific problem within a given maximal computational budget. In derivative-free optimization, such budgets are typically measured in terms of the number of objective function evaluations.

Data profiles plot the percentage of problems solved by the solvers under consideration for different values of the computational budget. These budgets are expressed in number of points $(n+1)$ required to form a simplex set, allowing the combination of problems of different dimensions in the same profile. Note that a different function of n could be chosen, but $n+1$ is natural in derivative-free optimization (since it is the minimum number of points required to form a positive basis, a simplex gradient, or a model with first-order accuracy).

We used in our experiments a maximal computational budget consisting of 500 function evaluations, as the dimension of the problems is relatively small and we are primarily interested in the behavior of the algorithms for

By using $\tilde{y}_{k+1}^{i_k} = x_k + \sigma_k d_k^{i_k}$ and condition (1) imposed on the weights in Step 2 of Algorithm 3.1, one obtains

$$f(x_k + \sigma_k d_k^{i_k}) > f(x_k) - \theta\rho(\sigma_k) - \rho(\sigma_k). \tag{5}$$

Note that in the max/mean version we arrive directly at $f(x_k + \sigma_k d_k^{i_k}) > f(x_k) - \rho(\sigma_k)$.

From this point, and for both cases (max/max and max/mean), the proof is nearly identical to the proof of Theorem 4.1. ∎

Again, when f is strict differentiable at x_*, we conclude that $\nabla f(x_*) = 0$. In Theorem 4.2 one also has the same issue regarding the density of the directions on the unit sphere being assumed for all refining subsequences K rather then for the whole sequence of iterates.

Remark 4.1. *It is important to point out that the condition imposed on the weights on Step 2 of Algorithm 3.1 is not necessary to derive Theorem 4.2 for convex functions f. In fact, under the convexity of f, condition (1) imposed on the weights would no longer be needed to prove something like (5) for the max/max version, which would result directly from $\sum_{i=1}^{m_\mu} \omega_{k-1}^i f(\tilde{y}_k^i) \geq f(\sum_{i=1}^{m_\mu} \omega_{k-1}^i \tilde{y}_k^i) = f(x_k)$ without the term $-\theta\rho(\sigma_k)$.*

The same happens also in Lemma 4.1 for the max/mean version.

In summary, versions max/max and max/mean of Algorithm 3.1 without the imposition of condition (1) are globally convergent for convex objective functions.

5. Extension to constraints

Let us consider now a constrained optimization problem of the form

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in \Omega \subset \mathbb{R}^n. \end{array}$$

The extreme barrier function associated with this problem is defined by

$$f_\Omega(x) \;=\; \begin{cases} f(x) & \text{if } x \in \Omega, \\ +\infty & \text{otherwise.} \end{cases}$$

The modifications to Algorithm 3.1 to handle a constrained set Ω are the following:

(1) One will start feasible: $x_0 \in \Omega$.
(2) One will use the extreme barrier function f_Ω instead of f in the sufficient decrease conditions (2), (3), and (4).

(3) Condition (1) is only applied when $f(\tilde{y}_{k+1}^{m_\mu}) < +\infty$.
(4) If more information is known about Ω, the generation of the poll directions can be confined to a subset of the unit sphere (see Subsection 5.3).

In fact, nothing else is needed, except to note that in Step 2 one may be considering function values that are equal to $+\infty$. In the ordering of the offspring samples this does not pose any problem, and we consider ties of $+\infty$ being broken arbitrarily. Also, the fact that condition (1) is not applied when there are $+\infty$ values is not a concern at all, since, due to (3) and (4), those iterations would be unsuccessful anyway.

By more information in the above Point 4 we mean constraints defined algebraically for which the derivatives of the functions involved are known, such as bounds on the variables or linear constraints.

Before stating the global convergence results of Algorithm 3.1 under the above modifications, we need a number of concepts specifically related to the constrained setting.

5.1. Cones and derivatives in the constrained case. A vector is said tangent to Ω at x if it satisfies the following definition.

Definition 5.1. *A vector $d \in \mathbb{R}^n$ is said to be a Clarke tangent vector to the set $\Omega \subseteq \mathbb{R}^n$ at the point x in the closure of Ω if for every sequence $\{y_k\}$ of elements of Ω that converges to x and for every sequence of positive real numbers $\{t_k\}$ converging to zero, there exists a sequence of vectors $\{w_k\}$ converging to d such that $y_k + t_k w_k \in \Omega$.*

The Clarke tangent cone to Ω at x, denoted by $T_\Omega^{Cl}(x)$, is then defined as the set of all Clarke tangent vectors to Ω at x. The Clarke tangent cone generalizes the tangent cone in Nonlinear Programming [18], but one can think about the latter one for gaining the necessary geometric motivation.

Given $x_* \in \Omega$ and $d \in T_\Omega^{Cl}(x)$, one is not sure that $x + td \in \Omega$ for $x \in \Omega$ arbitrarily close to x_*. Thus, for this purpose, one needs to consider directions in the interior of the Clarke tangent cone. The hypertangent cone appears then as the interior of the Clarke tangent cone (when such interior is nonempty).

Definition 5.2. *A vector $d \in \mathbb{R}^n$ is said to be a hypertangent vector to the set $\Omega \subseteq \mathbb{R}^n$ at the point x in Ω if there exists a scalar $\epsilon > 0$ such that*

$$y + tw \in \Omega, \quad \forall y \in \Omega \cap B(x;\epsilon), \quad w \in B(d;\epsilon), \quad \text{and} \quad 0 < t < \epsilon.$$

in [11]. CMA-ES's step length is defined as follows:

$$\sigma_{k+1}^{\text{CMA-ES}} = \sigma_k^{\text{CMA-ES}} \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|p_{k+1}^\sigma\|}{E\|\mathcal{N}(0,I)\|} - 1\right)\right),$$

where $E\|\mathcal{N}(0,I)\| = \sqrt{2}\Gamma(\frac{n+1}{2})/\Gamma(\frac{n}{2})$ is the expectation of the ℓ_2 norm of an $N(0,I)$ distributed random vector, c_σ, d_σ are positive constants, and $p_{k+1}^\sigma \in \mathbb{R}^n$ is the current state of the so-called conjugate evolution path (see [11]).

6.2. Algorithmic choices for the modified CMA-ES versions. A number of choices regarding parameters and updates of Algorithm 3.1 were made before the tests were launched.

Regarding initializations, the values of m_λ and m_μ and the initial weights followed the choices in CMA-ES (see [8]). The initial step length parameters were set to $\sigma_0 = \sigma_0^{\text{CMA-ES}} = 1$. The forcing function selected was $\rho(\sigma) = 10^{-4}\sigma^2$.

To reduce the step length in unsuccessful iterations we used $\sigma_{k+1} = 0.5\sigma_k$ which corresponds to setting $\beta_1 = \beta_2 = 0.5$. In successful iterations, we used $\sigma_{k+1} = \max\{\sigma_k, \sigma_k^{\text{CMA-ES}}\}$, in attempt to reset the step length to the ES one whenever possible.

The directions d_k^i, $i = 1, \ldots, m_\lambda$, were drawn from the multivariate normal distribution $\mathcal{C}_k$ updated by CMA-ES, scaled if necessary to obey the safeguards $d_{\min} \le \|d_k^i\| \le d_{\max}$, with $d_{\min} = 10^{-10}$, $d_{\max} = 10^{10}$.

Re-updating the weights in Step 2 of Algorithm 3.1 was not activated. In the one hand, we wanted the least amount of changes in CMA-ES. On the other hand, the weights update of Step 2 did not seem to have a real impact on the results, perhaps due to the convexity or convexity near the solutions present in many of the problems.

6.3. Test set. Our test set $\mathcal{P}$ is the one suggested in [17] and comprises 22 nonlinear vector functions from the CUTEr collection. The problems in $\mathcal{P}$ are then defined by a vector (k_p, n_p, m_p, s_p) of integers. The integer k_p is a reference number for the underlying CUTEr [7] vector function, n_p is the number of variables, m_p is the number of components $F_1, \ldots, F_{m_p}$ of the corresponding vector function F.

The integer $s_p \in \{0, 1\}$ defines the starting point via $x_0 = 10^{s_p} x_s$, where x_s is the standard starting point for the corresponding function. The use of $s_p = 1$ is helpful for testing solvers from a remote starting point since the

(see [14, 15]), where one can compute the positive generators of the appropriated tangent cones and then use them for *polling* (i.e., for evaluating the objective function at points of the form $x_k + \sigma_k d$, where d is a positive generator). The single extension to the nonlinear case that we are aware of required projections onto the feasible set at all iterations (see [16]), which may be computationally troublesome. There are, surely, a number of *hybrid* approaches using penalty or augmented Lagrangean functions or filter techniques, but without attempting to compute positive generators of the appropriated tangent cones related to the nonlinear part of the constraints. What makes the approach used in our paper successful is the combination of (i) a sufficient decrease condition for accepting new iterates (which took care of the need to drive the step size parameter to zero, a difficulty when using integer/rational lattices in the nonlinear case since the positive generators of the tangent cones in consideration would lack of rationality) with (ii) the dense generation of the directions in such tangent cones (which prevents stagnation at boundary points).

6. Numerical results

We made a number of numerical experiences to try to measure the effect of our modifications into ES. We are mainly interested in observing the changes that occur in ES in terms of an efficient and robust search of stationarity. We chose CMA-ES [11, 12] as our evolutionary strategy, on top of which we tested our globally convergent modifications.

6.1. CMA-ES. In CMA-ES (Covariance Matrix Adaptation Evolution Strategy) the distributions $\mathcal{C}_k$ are multivariate normal distributions. The weights are kept constant and besides belonging to the simplex S they also satisfy $\omega^1 \geq \cdots \geq \omega^{m_\mu} > 0$. Briefly speaking and using the notation of our paper, CMA-ES updates the covariance matrix of $\mathcal{C}_k$ as follows:

$$\begin{aligned} C_{k+1}^{\text{CMA-ES}} &= (1 - c_1 - c_\mu)\, C_k^{\text{CMA-ES}} + c_1 (p_{k+1}^c)(p_{k+1}^c)^\top \\ &\quad + c_\mu \sum_{i=1}^{m_\mu} \omega_i \, (d_k^i)(d_k^i)^\top, \end{aligned}$$

where c_1, c_μ are positive constants depending on n, and $p_{k+1}^c \in \mathbb{R}^n$ is the current state of the so-called evolution path, updated iteratively as described

The hypertangent cone to Ω at x, denoted by $T_\Omega^H(x)$, is then the set of all hypertangent vectors to Ω at x. The closure of the hypertangent cone is the Clarke tangent one (when the former is nonempty).

If we assume that f is Lipschitz continuous near x_*, we can define the Clarke-Jahn generalized derivative along directions d in the hypertangent cone to Ω at x_*,

$$f^\circ(x_*; d) = \limsup_{\substack{x \to x_*, x \in \Omega \\ t \downarrow 0, x + td \in \Omega}} \frac{f(x + td) - f(x)}{t}.$$

These derivatives are essentially the Clarke generalized directional derivatives [3], generalized by Jahn [13] to the constrained setting. Given a direction v in the tangent cone, one can consider the Clarke-Jahn generalized derivative to Ω at x_* as the limit $f^\circ(x_*; v) = \lim_{d \in T_\Omega^H(x_*), d \to v} f^\circ(x_*; d)$ (see [1]).

The point x_* is considered Clarke stationary if $f^\circ(x_*; d) \geq 0$, $\forall d \in T_\Omega^{Cl}(x_*)$.

5.2. Asymptotic results when derivatives are unknown. In this section we treat constraints as a pure black box in the sense that no information is assumed known about the constrained set Ω, rather than a yes/no answer to the question whether a given point is feasible. The changes to Algorithm 3.1 are those reported in Points 1–3 at the beginning of this section.

Now, for the analysis we start by noting that nothing changes in Lemmas 4.1 and 4.2. However, similarly to [1, 20], some background material is necessary to extend Theorems 4.1 and 4.2 to the constrained case.

The extension of Theorem 4.1 requires only the provision that d must lie in the hypertangent cone to Ω at x_* to first derive $f^\circ(x_*; d) \geq 0$ for appropriate directions called refining (see [1]). In fact, as we said before, the lim sup definition of $f^\circ(x_*; d)$ makes only sense in the constrained case when d is hypertangent to Ω at x_*. In the case of Theorem 5.1 below, such refining directions are associated with a convergent refining subsequence K, as the limit points of $\{a_k/\|a_k\|\}$ for all $k \in K$ sufficiently large such that $x_k + \sigma_k a_k \in \Omega$.

Theorem 5.1. *Consider the version mean/mean applied to the constrained setting and let $a_k = \sum_{i=1}^{m_\mu} \omega_k^i d_k^i$. Let x_* be the limit point of an unsuccessful subsequence of iterates $\{x_k\}_K$ for which $\lim_{k \in K} \sigma_k = 0$. Assume that f is Lipschitz continuous near x_* with constant $\nu > 0$.*

If $d \in T_\Omega^H(x_*)$ is a refining direction associated with $\{a_k/\|a_k\|\}_K$, then $f^\circ(x_*; d) \geq 0$.

If the set of refining directions associated with $\{a_k/\|a_k\|\}_K$ is dense in the unit sphere, then x_* is a Clarke stationary point.

Proof: The proof of the first assertion is just a repetition of the proof of Theorem 4.1. To prove the second part, we first conclude from the density of the refining directions on the unit sphere and the continuity of $f^\circ(x_*; \cdot)$ in $T_\Omega^H(x_*)$, that $f^\circ(x_*; d) \geq 0$ for all $d \in T_\Omega^H(x_*)$. Finally, we conclude that $f^\circ(x_*; v) = \lim_{d \in T_\Omega^H(x_*), d \to v} f^\circ(x_*; d) \geq 0$ for all $v \in T_\Omega(x_*)$. ∎

The extension of Theorem 4.2 requires similar precautions. In the case of Theorem 5.2 below, the refining directions are associated with a convergent refining subsequence K, as the limit points of $\{d_k^i/\|d_k^i\|\}$ for all $k \in K$ sufficiently large such that $x_k + \sigma_k d_k^i \in \Omega$.

Theorem 5.2. *Consider the versions max/max and max/mean applied to the constrained setting. Let x_* be the limit point of an unsuccessful subsequence of iterates $\{x_k\}_K$ for which $\lim_{k \in K} \sigma_k = 0$. Assume that f is Lipschitz continuous near x_* with constant $\nu > 0$.*

If $d \in T_\Omega^H(x_*)$ is a refining direction associated with $\{d_k^{i_k}/\|d_k^{i_k}\|\}_K$, where $i_k \in \operatorname{argmax}_{1 \leq i \leq m_\mu} f(\tilde{y}_{k+1}^i)$, then $f^\circ(x_*; d) \geq 0$.

If, for each $i \in \{1, \ldots, m_\mu\}$, the set of refining directions associated with $\{d_k^i/\|d_k^i\|\}_K$ is dense in the unit sphere, then x_* is a Clarke stationary point.

5.3. Asymptotic results when derivatives are known. Although the approach analyzed in Subsection 5.2 (resulting only from the modifications in Points 1–3 at the beginning of this section) can in principle be applied to any type of constraints, it is obviously more appropriate to the case where one cannot compute the derivatives of the functions algebraically defining the constraints.

Now we consider the case where we can compute tangent cones at points on the boundary of the feasible set Ω (in what can be considered as the additional information alluded to in the Point 4 of the beginning of this section). This is the case whenever Ω is defined by $\{x \in \mathbb{R}^n : \ c_i(x) \leq 0, \ i \in \mathcal{I}\}$ and the derivatives of the functions c_i are known. Two particular cases that appear frequently in practice are bound and linear constraints.

For theoretical purposes, let ϵ be a positive scalar and k_0 a positive integer. Let us also denote by T_{Ω,ϵ,k_0} the union of all Clarke tangent cones $T_\Omega(y)$ for all points y at the boundary of Ω such that $\|y - x_k\| \leq \epsilon$ for all $k \geq k_0$.

One is now ready to consider the extension of Theorems 4.1 and 4.2 to the constrained case under the presence of constrained derivative information. Such extensions can be considered as corollaries of Theorems 5.1 and 5.2. Nothing else is needed to add regarding the proofs since the Clarke tangent cone $T_\Omega(x_*)$ is contained in T_{Ω,ϵ,k_0} for any limit point x_* of a subsequence of iterates (and in particular for those consisting of unsuccessful iterations for which the step size tends to zero). The results are stated assuming that the limit point x_* is in the boundary of Ω, otherwise Theorems 5.1 and 5.2 apply as they stand.

Theorem 5.3. *Consider the version mean/mean applied to the constrained setting and let $a_k = \sum_{i=1}^{m_\mu} \omega_k^i d_k^i$. Let $x_* \in \operatorname{fr}(\Omega)$ be the limit point of an unsuccessful subsequence of iterates $\{x_k\}_K$ for which $\lim_{k \in K} \sigma_k = 0$. Assume that f is Lipschitz continuous near x_* with constant $\nu > 0$.*

If $d \in T_\Omega^H(x_*)$ is a refining direction associated with $\{a_k/\|a_k\|\}_K$, then $f^\circ(x_*; d) \geq 0$.

If the set of refining directions associated with $\{a_k/\|a_k\|\}_K$ is dense in the intersection of T_{Ω,ϵ,k_0} with the unit sphere, then x_* is a Clarke stationary point.

Theorem 5.4. *Consider the versions max/max and max/mean applied to the constrained setting. Let $x_* \in \operatorname{fr}(\Omega)$ be the limit point of an unsuccessful subsequence of iterates $\{x_k\}_K$ for which $\lim_{k \in K} \sigma_k = 0$. Assume that f is Lipschitz continuous near x_* with constant $\nu > 0$.*

If $d \in T_\Omega^H(x_*)$ is a refining direction associated with $\{d_k^{i_k}/\|d_k^{i_k}\|\}_K$, where $i_k \in \operatorname{argmax}_{1 \leq i \leq m_\mu} f(\tilde{y}_{k+1}^i)$, then $f^\circ(x_*; d) \geq 0$.

If, for each $i \in \{1, \ldots, m_\mu\}$, the set of refining directions associated with $\{d_k^i/\|d_k^i\|\}_K$ is on the intersection of T_{Ω,ϵ,k_0} with the unit sphere, then x_* is a Clarke stationary point.

It is very interesting to point out that there is a novel point in the assembly of our approach for handling the constrained case with known constrained derivative information. In fact, if one looks at the existing literature of *pure* direct-search methods (of directional type) for constraints, one sees approaches only developed for the bound or linear constrained cases