Int. J. Appl. Math. Comput. Sci., 2001, Vol.11, No.2, 429–457

ANALYSIS OF AN ISOPE-TYPE DUAL ALGORITHM FOR OPTIMIZING CONTROL AND NONLINEAR OPTIMIZATION

Wojciech TADEJ*, Piotr TATJEWSKI*

First results concerning important theoretical properties of the dual ISOPE (Integrated System Optimization and Parameter Estimation) algorithm are presented. The algorithm applies to on-line set-point optimization in control structures with uncertainty in process models and disturbance estimates, as well as to difficult nonlinear constrained optimization problems. Properties of the conditioned (dualized) set of problem constraints are investigated, showing its structure and feasibility properties important for applications. Convergence conditions for a simplified version of the algorithm are derived, indicating a practically important threshold value of the right-hand side of the conditioning constraint. Results of simulations are given confirming the theoretical results and illustrating properties of the algorithms.

Keywords: nonlinear optimization, optimizing control, duality, condition number

1. Introduction

The aim of this paper is to analyze the convergence of a dual algorithm of the ISOPE method. The Integrated System Optimization and Parameter Estimation (ISOPE) method was originally introduced by Roberts (1979) and then extensively investigated, see, e.g. (Brdyś *et al.*, 1987) for a first basic theoretical analysis with applicability conditions, as well as (Brdyś and Tatjewski, 1994; Tatjewski, 1999) for latest developments. The method was originally designed as an on-line steady-state optimizing control algorithm for industrial processes within a multilayer structure. It also applies to nonlinear optimization problems with difficult, strongly nonlinear equality constraints (e.g. process models).

The multilayer approach is commonly used in industrial applications. It was also intensively investigated in the recent decades, see, e.g. (Findeisen *et al.*, 1980). The main idea is to decompose the original control task, which is the generation of optimized trajectories of the manipulated variables, into a sequence of different and hierarchically structured sub-tasks handled by dedicated control layers. Direct

* Institute of Control and Computation Engineering, Faculty of Electronics and Information Technology, Warsaw University of Technology, ul. Nowowiejska 15/19, 00–665 Warsaw, Poland, e-mail: {W.Tadej,P.Tatjewski}@ia.pw.edu.pl

follow-up control, optimizing control and supervisory control are the classical, well-established layers. The task of the direct control layer is to keep the process at a desired state in spite of fast disturbances acting upon it, where the process state is defined by a collection of set-points for direct controllers, usually standard industrial controllers. The task of the steady-state optimizing control is an on-line adjustment of the set-points to run the process as profitable as possible in a varying, uncertain environment. This is typically modelled as slow-varying (or abrupt but rare) changes in certain measurable and unmeasurable uncontrollable process inputs (disturbances), and uncertainty in process models.

It is assumed in the ISOPE method that the available process model is only approximate due to modelling errors and disturbances. The method copes with this uncertainty using an on-line steady-state feedback measurement information. The method is iterative with feedback from the process consisting of output measurements in subsequent steady-states. It is the only optimizing control method where iterations converge to the real process optimum in spite of uncertainty. However, to perform the next iteration, derivatives of the process outputs in the current steady-state, with respect to the set-points, are also required (set-points are decision variables at the optimization layer). This creates the main difficulty in practical implementations of the method. Originally (cf. Brdyś *et al.*, 1987; Roberts, 1979), an application of additional (possibly small) changes of the process set-points around any current value was proposed, to obtain approximations of the derivatives by a finite-difference technique. However, it is time- and cost-consuming because each additional change of the set-points means an additional dynamical transient process in the plant. Therefore, many attempts have been made to overcome this difficulty, mainly by trying to formulate stochastic or composite dynamic and steady-state versions of the algorithm (Zhang and Roberts, 1990). However, the approaches based on attempts to extract precise statics of the plant from measurements of its transient processes turned out to be difficult and of limited reliability. The reason is that more reliable algorithms for set-point optimizing control should rely on steady-state measurements; since then the measurement noise can be sufficiently well filtered. A breakthrough along these guidelines was the development of a dual-type ISOPE algorithm (Brdyś and Tatjewski, 1994). It was the first ISOPE algorithm which used the steady-state measurements only and did not require additional set-point changes for derivative approximation. Consecutive set-points for the direct controllers are generated in such a way that they constitute a sequence which not only converges to the optimal set-point, but also forms a basis for estimation of the derivatives. This is due to the fact that at every algorithm iteration the set-point is calculated in a way taking into account both the process optimality and the need for estimation of the process output derivatives in the next iteration (future identification). In this sense the algorithm is dual.

An important application area of the ISOPE method is the (off-line) optimization of problems with difficult, strongly nonlinear equality constraints on the input-output structure (e.g. complex, phenomenological models of input-output relations of different processes). The ISOPE algorithms are the same as in the previous case, only simplified (e.g. linear) constraints are used as 'models' of the original difficult constraints, which in turn serve as 'process output mappings'. Certainly, in this situation

calculation of derivatives is easy when using well-known numerical approaches and may be very accurate. However, the dual ISOPE algorithms are usually advantageous also here, as the ones resulting in a smaller number of calculations of difficult nonlinear constraints. Another area of application of the ISOPE method in optimization is the case when the performance function itself is difficult to be evaluated (in the sense of the time needed for its calculation). In this case the ISOPE algorithm with linearization of the performance function at each iteration point (algorithm ISOPEDL analyzed in Section 4) can be advantageously applied.

Theoretical results concerning the optimality and convergence of the basic version of the ISOPE method were proved formally and under reasonable assumptions in (Brdyś *et al.*, 1987). However, despite several attempts, the approach used there could not be applied to the dual method, which significantly differs from the basic one. Moreover, the theoretical analysis occurred to be extremely difficult due to a strongly nonlinear nature of an additional constraint (called the conditioning constraint) introduced to force duality. Therefore, in this paper theoretical results obtained using another, geometrical approach are presented, and they regard a two-dimensional case. To the best of our knowledge, these are first theoretical results concerning the feasibility and convergence of a dual-type ISOPE method.

First, properties of the feasibility set composed of original and conditioning constraints are investigated. The structure of the feasibility set is derived, and it is proved that adding the conditioning constraint does not cause the overall feasible set to become empty during iterations, which is an important result from the point of view of application. Second, convergence conditions for a simplified, unconstrained (except for the conditioning constraint) version of the algorithm are derived. A threshold value of the parameter on the right-hand side of the conditioning constraint is found which guarantees, under several reasonable conditions, that the gradient of the process performance function converges to zero. The results are practically important not only because convergence conditions are given, but also because the threshold value lies well outside a numerically recommended range of values. The results of numerical simulations are finally presented to confirm the obtained theoretical results and to investigate the behaviour of both the versions of the algorithm.

2. Dual-Type ISOPE Algorithm

The *steady-state optimizing control problem* (OCP) can be formulated as follows (Brdyś *et al.*, 1987; Brdyś and Tatjewski, 1994):

$$\begin{aligned} &\text{minimize} && Q(c,y) \\ &\text{subject to} && y = F_*(c), \\ & && c \in \mathcal{C}, \end{aligned} \tag{1}$$

where $c \in \mathbb{R}^n$ are set-points of direct process controllers to be optimized at the optimizing control layer, and y denote process outputs. $F_*\colon\ \mathbb{R}^n \mapsto \mathbb{R}^m$ represents a real, generally nonlinear input-output mapping (static characteristics) of the plant, and $Q(\cdot,\cdot)$ is the plant performance index describing formally the process productivity

(economic production goals) dependent on c and y. The set $\mathcal{C}$ represents inequality constraints on the set-points. Certainly, it cannot be assumed that $F_*(\cdot)$ is known exactly and consequently, only an *approximate model* of F_* is available,

$$y = F(c, \alpha),$$

with model (adjustable) parameters α. Therefore, the following *steady-state model optimization problem* (MOP) corresponds to the OCP:

$$\begin{aligned} &\text{minimize} && Q(c, y) \\ &\text{subject to} && y = F(c, \alpha), \\ & && c \in \mathcal{C}. \end{aligned} \tag{2}$$

By eliminating the output variable y, the problems (1) and (2) can be simplified to the form

$$\begin{aligned} &\text{minimize} && Q(c, F_*(c)) \\ &\text{subject to} && c \in \mathcal{C}, \end{aligned} \tag{3}$$

and

$$\begin{aligned} &\text{minimize} && Q\big(c, F(c, \alpha)\big) \\ &\text{subject to} && c \in \mathcal{C}, \end{aligned} \tag{4}$$

respectively.

Unfortunately, due to modelling inaccuracies, a solution $\widehat{c}_m$ to (4) can differ significantly from a solution $\widehat{c}_*$ to (3), leading to suboptimal control with production losses when pure model-based set-points $\widehat{c}_m$ are applied in the plant. A remedy is an iterative improvement of the set-points (starting from $\widehat{c}_m$). The ISOPE method makes it possible to iterate set-points towards $\widehat{c}_*$. The idea of the approach is to use iteratively the following modification of (3) called the *modified model optimization problem* (MMOP):

$$\begin{aligned} &\text{minimize}_c \big\{ Q\big(c, F(c, \alpha_i)\big) - \lambda(c_i, \alpha_i)^T c + \rho \|c - c_i\|^2 \big\} \\ &\text{subject to} \quad c \in \mathcal{C}, \end{aligned} \tag{5}$$

where

$$\lambda(c_i, \alpha_i)^T = Q_y'\big(c_i, F(c_i, \alpha_i)\big) \big[F_c'(c_i, \alpha_i) - F_*'(c_i) \big], \tag{6}$$

and $\rho > 0$ is a penalty coefficient of the quadratic regularizing term. The subscript 'i' is an iteration index, the point c_i constitutes a set-point which is to be improved after the current iteration of the algorithm. $Q_y'(c_i, y_i)$ denotes the partial derivative of Q with respect to y, taken at (c_i, y_i), etc.

It follows directly from the construction that *the performance function of the problem MMOP (5) has the derivative at the point c_i equal to the derivative of the performance function of the original optimizing control problem (3)*, provided the

model and process outputs are equal after an appropriate model parameter estimation (yielding the parameter values α_i) at the point c_i,

$$F(c_i, \alpha_i) = F_*(c_i). \tag{7}$$

Further, consider a situation when the modified model optimization problem MMOP is used instead of the basic model optimization problem MOP in the so-called 'iterative two-step approach', i.e. when iterations of the set-points are performed in such a way that a solution $\widehat{c}(c_i)$ to the MMOP problem becomes the next process set-point c_{i+1}, $c_{i+1} = \widehat{c}(c_i)$, etc. If the sequence $\{c_i\}$ is then convergent to a point, say $\tilde{c}$, then this point satisfies $\tilde{c} = \widehat{c}(\tilde{c})$ and thus *also satisfies necessary optimality conditions for the optimizing control problem* (3). Moreover, the reasoning is also true if instead of (MMOP) the following *simplified modified model optimization problem* (MMOPL) is used:

$$\begin{aligned}\text{minimize}_c\big\{&Q(c_i, y_i) + Q_c^{'}(c_i, y_i)(c - c_i) + Q_y^{'}(c_i, y_i)F_c^{'}(c_i, \alpha_i)(c - c_i)\\ &-\lambda(c_i, \alpha_i)^T c + \rho\|c - c_i\|^2\big\}\\ \text{subject to}\ \ &c \in \mathcal{C},\end{aligned} \tag{8}$$

where we write $y_i = F(c_i, \alpha_i)$ to shorten the notation, $y_i = F(c_i, \alpha_i) = F_*(c_i)$ due to (7). The problem (MMOPL) is a simplified version of (MMOP) using instead of $Q(c, F(c, \alpha_i))$ its *linearization* at the point c_i only.

The basic dual-type ISOPE algorithm will now be formulated:

Algorithm 1. The ISOPED (dual ISOPE) algorithm:

1. Set $i := 0$, take (or evaluate) process outputs $F_*(c_i)$ at points $c_{-n}, \ldots, c_0 \in \mathcal{C} \subset \mathbb{R}^n$ such that the matrix $A_0(c_0)$ (see (11) for the definition) is sufficiently well conditioned (see further remarks in the text).

2. Change set-points of the process direct controllers to the values c_i. Wait for the steady-state and measure the outputs $y_i = F_*(c_i)$. Estimate the model parameters α_i under the condition (7). Calculate an approximation of the derivative $F_*^{'}(c_i)$ using output measurements at points $c_{i-n}, c_{i-n+1}, \ldots, c_i$ (applying the formula (12)).

3. Solve the conditioned modified model optimization problem (CMMOP)

$$\begin{aligned}&\text{minimize}_c\big\{q_{\rho i}(c) = Q\big(c, F(c, \alpha_i)\big) - \lambda(c_i, \alpha_i)^T c + \rho\|c - c_i\|^2\big\}\\ &\text{subject to}\ \ \mathcal{C} \cap \mathcal{D},\end{aligned} \tag{9}$$

   where

$$\mathcal{D} = \left\{c \in \mathbb{R}^n :\ \frac{\sigma_{\max}\big(A_{i+1}(c)\big)}{\sigma_{\min}\big(A_{i+1}(c)\big)} \le a\right\}, \tag{10}$$

$\sigma_{\max}(A_{i+1}(c))$, $\sigma_{\min}(A_{i+1}(c))$ are respectively the maximal and minimal singular values, of the $n \times n$ matrix

$$A_{i+1}(c) = \begin{bmatrix} c - c_i & c - c_{i-1} & \cdots & c - c_{i-n+1} \end{bmatrix}, \tag{11}$$

and $a > 1$ is a parameter of the algorithm. Set $c_{i+1} := \arg\min_{\mathcal{C}\cap\mathcal{D}} q_{\rho i}(c)$.

4. If $\|c_{i+1} - c_i\| < \varepsilon$ then STOP, otherwise set $i := i + 1$ and go to Step 2.

In Step 3 of Algorithm 1 the optimization problem CMMOP is solved, which is an extension of the MMOP problem (5) of the original ISOPE method. The constraint set in CMMOP is the product $\mathcal{C} \cap \mathcal{D}$, not the set $\mathcal{C}$ only. The constraint $c_{i+1} \in \mathcal{D}$ is necessary to assure that estimation of the derivative $F_*'(c_{i+1})$ in the next iteration (based on points $c_{i-n+1}, c_{i-n+2}, \ldots, c_{i+1}$) will be well-conditioned. The derivative $F_*'(c_i) = [F_{*1}'(c_i)^T \ \cdots \ F_{*m}'(c_i)^T]^T$ at each iteration is calculated using the formula

$$A_i(c_i)^T F_{*j}'(c_i)^T \cong \begin{bmatrix} F_{*j}(c_i) - F_{*j}(c_{i-1}) \\ \vdots \\ F_{*j}(c_i) - F_{*j}(c_{i-n}) \end{bmatrix}, \quad j = 1, \cdots, m = \dim F_*, \tag{12}$$

see (Brdyś and Tatjewski, 1994) for a detailed derivation. Because the right-hand sides of the systems of linear equations (12) are vectors consisting of measurement values, the results of calculations can be strongly affected by measurement errors. Therefore, the matrix $A_i(c_i)$ must be well-conditioned, i.e. its condition number $\mathrm{cond}\,(A_i(c_i)) = \sigma_{\max}(A_i(c_i))/\sigma_{\min}(A_i(c_i))$ must not be too large. Recall that $\mathrm{cond}\,(A_i(c_i))$ measures the influence of errors in the right-hand side vector on the errors in the solution of (12), see, e.g. (Kiełbasiński, 1992). Thus, in Step 3 of the algorithm, c_{i+1} is forced to have values assuring that the matrix $A_{i+1}(c_{i+1})$ has the condition number not greater than a. The set $\mathcal{D}$ will further be called the *conditioning set*. Observe that the nature of the algorithm is dual—when solving the CMMOP both the optimality at the current iteration and estimation requirements for the next iteration are taken into account.

Two points concerning the formulated ISOPED algorithm should be explained here. First, the next point is generated using the simple formula $c_{i+1} := \widehat{c}(c_i)$, where $\widehat{c}(c_i) = \arg\min_{\mathcal{C}\cap\mathcal{D}} q_{\rho i}(c)$. In most ISOPE formulations a more general relaxation formula $c_{i+1} := c_i + k_c(\widehat{c}(c_i) - c_i)$ was used, where k_c is a gain factor affecting the convergence and the convergence rate of the algorithm. The usually case-dependable adjustment of this parameter was necessary in the original algorithm formulations where the convexifying term $\rho\|c - c_i\|^2$ was not applied. However, this term not only convexifies the problem, but it also affects the algorithm behaviour similarly as the relaxation formula influencing the distance between the current point c_i and the next one c_{i+1}. This was clearly shown in many simulation examples, where decreasing k_c or increasing ρ had a similar influence on the distances between consecutive points and the convergence rate. Therefore, using in this paper the simple case $k_c = 1$ does

not constrain significantly the generality of the analysis, on the other hand making this analysis easier.

Second, the formulated ISOPED algorithm can be successfully applied for optimization of problems with known but difficult nonlinear equality constraints of input-output type. The OCP problem (1) serve then as an original optimization problem with difficult constraints $y = F_*(c)$, and the problem (2) is its simplification easily solvable by standard solvers. Obviously, the conditioning set can be less restrictive in pure optimization applications (i.e. greater values of a possible), since the errors are now numerical errors in calculation of $F_*(c_i)$ only.

It should be mentioned that the nature of the conditioning set $\mathcal{D}$ is complex, which makes a theoretical analysis of the algorithm very difficult (the standard ISOPE method being itself difficult to analyze). Although the ISOPED algorithm was originally formulated in (Brdyś and Tatjewski, 1994), it is in the present paper that first theoretical results are published for the two-dimensional case, $\dim c = 2$.

In the next section properties of the conditioning set $\mathcal{D}$ will be investigated, and in the following section a convergence analysis of the dual algorithm with a simplified modified model optimization problem (CMMOPL) will be presented.

3. Geometric Properties of Conditioning Sets

In the two-dimensional case the conditioning set $\mathcal{D} = \mathcal{D}_a(c_{i-1}, c_i)$ is described by

$$\mathcal{D}_a(c_{i-1}, c_i) = \left\{ c \in \mathbb{R}^2 : \ \frac{\sigma_{\max}\big(A(c)\big)}{\sigma_{\min}\big(A(c)\big)} \le a \right\}$$

$$= \bigcup_{\widetilde{a} \in \langle 1, a \rangle} \left\{ c \in \mathbb{R}^2 : \ \frac{\sigma_{\max}(A(c))}{\sigma_{\min}(A(c))} = \widetilde{a} \right\} = \bigcup_{\widetilde{a} \in \langle 1, a \rangle} \widetilde{\mathcal{D}}_{\widetilde{a}}(c_{i-1}, c_i),$$

where

$$A(c) = \big[c - c_i \ \ c - c_{i-1} \big].$$

Proposition 1. *$\widetilde{\mathcal{D}}_{\widetilde{a}}(c_{i-1}, c_i)$ is composed of two circles with the same radius equal to $\widetilde{r}(|\overline{c_i c_{i-1}}|/2)$, with centres located on the bisector of the segment $\overline{c_i c_{i-1}}$ symmetrically at its both sides and at the distance $\widetilde{h}(|\overline{c_i c_{i-1}}|/2)$ from the midpoint of the segment, where*

$$\widetilde{r} = \frac{\widetilde{a}^2 - 1}{2\widetilde{a}}, \quad \widetilde{h} = \frac{\widetilde{a}^2 + 1}{2\widetilde{a}}, \quad \widetilde{l} \stackrel{\text{def}}{=} \sqrt{\widetilde{r}^2 + 2} = \sqrt{\widetilde{h}^2 + 1}. \tag{13}$$

Proof. There exists a map $\mathcal{S} : \mathbb{R}^2 \to \mathbb{R}^2$ of the form $\mathcal{S}(c) = g\,Hc + s$, where $g > 0$, H is an orthogonal matrix 2×2 and $s \in \mathbb{R}^2$ (the so-called similarity map, i.e. an affine map preserving angles), such that

$$\mathcal{S}(c_{i-1}) = [-1, 0]^T \quad \text{and} \quad \mathcal{S}(c_i) = [1, 0]^T.$$

Note that the following holds:

$$v \in \widetilde{\mathcal{D}}_{\widetilde{a}}(c_{i-1}, c_i) \Leftrightarrow v \in \left\{ c \in \mathbb{R}^2 : \ \frac{\sigma_{\max}\big(A(c)\big)}{\sigma_{\min}\big(A(c)\big)} = \widetilde{a} \right\}$$

$$\Rightarrow \mathcal{S}(v) \in \left\{ c \in \mathbb{R}^2 : \ \frac{\sigma_{\max}\big(B(c)\big)}{\sigma_{\min}\big(B(c)\big)} = \widetilde{a} \right\}$$

$$\Leftrightarrow \mathcal{S}(v) \in \widetilde{\mathcal{D}}_{\widetilde{a}}\big([-1,0]^T, [1,0]^T\big) = \widetilde{\mathcal{D}}_{\widetilde{a}}\big(\mathcal{S}(c_{i-1}), \mathcal{S}(c_i)\big)$$

and also

$$w \in \widetilde{\mathcal{D}}_{\widetilde{a}}\big(\mathcal{S}(c_{i-1}), \mathcal{S}(c_i)\big) \Rightarrow \mathcal{S}^{-1}(w) \in \widetilde{\mathcal{D}}_{\widetilde{a}}(c_{i-1}, c_i),$$

where

$$B(c) = \left[\ c - \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad c - \begin{bmatrix} -1 \\ 0 \end{bmatrix} \ \right] \tag{14}$$

because for matrices $A(v) = [v - c_i \ \ v - c_{i-1}]$ and

$$\begin{aligned} B\big(\mathcal{S}(v)\big) &= \left[\ \mathcal{S}(v) - \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \mathcal{S}(v) - \begin{bmatrix} -1 \\ 0 \end{bmatrix} \ \right] \\ &= \big[\mathcal{S}(v) - \mathcal{S}(c_i) \quad \mathcal{S}(v) - \mathcal{S}(c_{i-1})\big] \\ &= gH\big[v - c_i \ \ v - c_{i-1}\big] = gHA(v) \end{aligned}$$

the ratios of singular values are equal:

$$\frac{\sigma_{\max}\big(B(\mathcal{S}(v))\big)}{\sigma_{\min}\big(B(\mathcal{S}(v))\big)} = \frac{\sigma_{\max}\big(A(v)\big)}{\sigma_{\min}\big(A(v)\big)}.$$

Thus it is enough to find the set $\widetilde{\mathcal{D}}_{\widetilde{a}}([-1,0]^T, [1,0]^T)$. Then

$$\widetilde{\mathcal{D}}_{\widetilde{a}}(c_{i-1}, c_i) = \mathcal{S}^{-1}\left(\widetilde{\mathcal{D}}_{\widetilde{a}}\big([-1,0]^T, [1,0]^T\big)\right). \tag{15}$$

For $w \in \widetilde{\mathcal{D}}_{\widetilde{a}}([-1,0]^T, [1,0]^T)$ we have

$$w \in \widetilde{\mathcal{D}}_{\widetilde{a}}\big([-1,0]^T, [1,0]^T\big) \Leftrightarrow \frac{\sigma_{\max}\big(B(w)\big)}{\sigma_{\min}\big(B(w)\big)} = \widetilde{a}$$

$$\Leftrightarrow \exists\, k > 0 \text{ such that } \sigma_{\min}\big(B(w)\big) = k \text{ and } \sigma_{\max}\big(B(w)\big) = k\widetilde{a}$$

$$\Leftrightarrow \exists\, k > 0 : \ \ k^2 \text{ and } k^2\widetilde{a}^2 \text{ are the eigenvalues of } B(w)^T B(w).$$

This is equivalent to

$$\exists\, k>0:\ \ \det\big(B(w)^T B(w)-sI\big)=\big(s-k^2\big)\big(s-k^2\widetilde{a}^2\big)$$

$$\Leftrightarrow \exists\, k>0:\ \ \det\begin{bmatrix} b_1^T b_1 - s & b_1^T b_2 \\ b_2^T b_1 & b_2^T b_2 - s \end{bmatrix}=\big(s-k^2\big)\big(s-k^2\widetilde{a}^2\big).$$

For $w=\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]$, from (14) it follows that

$$B=[b_1\ b_2]=\begin{bmatrix} x-1 & x+1 \\ y & y \end{bmatrix}.$$

Therefore we require

$$\exists\, k>0:\ \ b_1^T b_1+b_2^T b_2=2x^2+2y^2+2=k^2\big(1+\widetilde{a}^2\big)$$

and

$$\det\begin{bmatrix} b_1^T b_1 & b_1^T b_2 \\ b_2^T b_1 & b_2^T b_2 \end{bmatrix}=\big(\det B(w)\big)^2=4y^2=k^4\widetilde{a}^2$$

$$\Leftrightarrow \exists\, k:\ \ \left(y=\frac{k^2\widetilde{a}}{2}\ \text{ or }\ y=-\frac{k^2\widetilde{a}}{2}\right)\ \text{ and }\ 2x^2+2y^2+2=k^2\big(1+\widetilde{a}^2\big)$$

$$\Leftrightarrow 2x^2+2y^2+2=\frac{2y}{\widetilde{a}}\big(1+\widetilde{a}^2\big)\ \text{ or }\ 2x^2+2y^2+2=\frac{-2y}{\widetilde{a}}\big(1+\widetilde{a}^2\big)$$

$$\Leftrightarrow x^2+\big(y-\widetilde{h}\big)^2=\widetilde{r}^2\ \text{ or }\ x^2+\big(y+\widetilde{h}\big)^2=\widetilde{r}^2, \tag{16}$$

where

$$\widetilde{h}=\frac{\widetilde{a}^2+1}{2\widetilde{a}},\quad \widetilde{r}=\frac{\widetilde{a}^2-1}{2\widetilde{a}}.$$

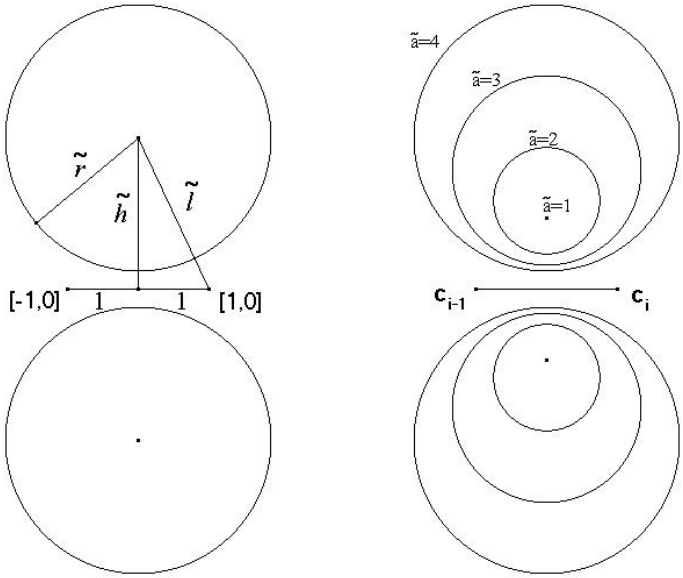
From the above it follows that the set $\widetilde{\mathcal{D}}_{\widetilde{a}}([-1,0]^T,[1,0]^T)$ is composed of two circles (see Fig. 1(a)) satisfying equations (16). Because $\mathcal{S}^{-1}$ is also a similarity map and due to (15), the set $\widetilde{\mathcal{D}}_{\widetilde{a}}(c_{i-1},c_i)$ has the properties stated in the proposition. ■

Let us rewrite the left equation of (16) as follows:

$$x^2+\left(y-\frac{\widetilde{a}^2+1}{2\widetilde{a}}\right)^2=\left(\frac{\widetilde{a}^2-1}{2\widetilde{a}}\right)^2$$

$$\Leftrightarrow \big(x^2+(y-1)^2\big)+2\left(1-\frac{\widetilde{a}^2+1}{2\widetilde{a}}\right)(y)=0. \tag{17}$$

We obtain a linear combination of equations $x^2+(y-1)^2=0$ and $y=0$. This defines a pencil of circles, which is well-known in analytical geometry (see, e.g. Stark,



(a) (b)

Fig. 1. Conditioning set.

1974). For $\widetilde{a} \in \mathbb{R}\backslash\{0\}$ (17) defines a circle of the pencil, see Fig. 1(b). Note that a negative value of $\widetilde{a}$ in (17) produces the same circle as the right equation in (16) for the corresponding positive value.

Since circles of the pencil (17) fill the plane, we have the following result:

Corollary 1. *The set*

$$\mathcal{D}_a(c_{i-1}, c_i) = \bigcup_{\widetilde{a}\in\langle 1,a\rangle} \widetilde{\mathcal{D}}_{\widetilde{a}}(c_{i-1}, c_i)$$

is composed of two discs $\mathcal{D}_1$, $\mathcal{D}_2$ *with the same radius equal to* $r(|\overline{c_i c_{i-1}}|/2)$ *and with centres located on the bisector of the segment* $\overline{c_i c_{i-1}}$ *symmetrically at its both sides and at the distance* $h(|\overline{c_i c_{i-1}}|/2)$ *from the midpoint of the segment, where*

$$r = \frac{a^2-1}{2a}, \quad h = \frac{a^2+1}{2a}, \quad l \stackrel{\text{def}}{=} \sqrt{r^2+2} = \sqrt{h^2+1}. \tag{18}$$

Here and subsequently $\mathcal{D}_1$ and $\mathcal{D}_2$ denote the component discs of the conditioning set $\mathcal{D}$.

Now we can consider properties crucial for the applicability and analysis of the considered algorithm.

Proposition 2. (Non-emptiness of the feasibility set $\mathcal{C}\cap(\mathcal{D}_1\cup\mathcal{D}_2)$.) *Assume that the algorithm feasibility set* $\mathcal{C}_\mathcal{D} = \mathcal{C}\cap(\mathcal{D}_1\cup\mathcal{D}_2)$ *is non-empty for* $i = 0$, *where the set*

$\mathcal{D}_1 \cup \mathcal{D}_2$ is defined by the points c_{-1} and c_0. Then the feasibility set is non-empty for all $i > 0$, i.e. at all next algorithm iterations where $\mathcal{D}_1 \cup \mathcal{D}_2$ is defined by points c_{i-1}, c_i, $i = 1, 2, \ldots$.

Proof. The proof is by induction and is based on Fig. 2.

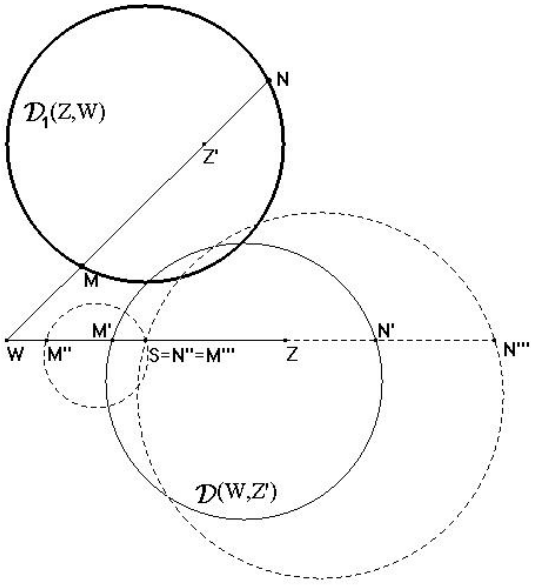


Fig. 2. Non-emptiness of $\mathcal{C} \cap (\mathcal{D}_1 \cup \mathcal{D}_2)$.

Assume that the proposition is true for $c_{i-1} = Z$, $c_i = W$. Then there exists a point $C = c_{i+1}$ somewhere within one of the discs $\mathcal{D}_1$, $\mathcal{D}_2$ defined by Z, W. Let $c_{i+1} \in \mathcal{D}_1(Z, W)$ and $\mathcal{D}_1(Z, W)$ be the upper, thick-lined disc in Fig. 2. Let M, N be the intersection points of $\mathcal{D}_1(Z, W)$ with the ray $\mathcal{R}(WC)$, the points closer to and farther from W, appropriately. Let $Z' \in \mathcal{R}(WC)$ and $|\overline{WZ'}| = |\overline{WZ}|$.

If C were equal to Z', then the thin-lined disc $\mathcal{D}(W, Z')$—the image of $\mathcal{D}_1(Z, W)$ with respect to the bisector of the angle $\sphericalangle ZWZ'$—would be one of $\mathcal{D}_i$ discs for W, C, say $\mathcal{D}_1(W, C)$. The points M, N would then be mapped to M', N'. If C were shifted to M, then $\mathcal{D}_1(W, C)$ would be dilated, accordingly, with respect to W and with factor $|\overline{WM}|/|\overline{WZ'}|$ (M', N' would be mapped to M'', N''). A similar transformation would occur when dilating C to N (factor $|\overline{WN}|/|\overline{WZ'}|$, M', N' mapped to M''', N''').

Since $\mathcal{R}(WC) \cap \mathcal{D}_1(Z, W)$ is non-empty (it contains $c_{i+1} = C$), then by reflection $\mathcal{R}(WZ) \cap \mathcal{D}_1(W, C = Z')$ is non-empty and by dilation so is $\mathcal{R}(WZ) \cap \mathcal{D}_1(W, C)$ for any $C \in \overline{MN}$. Let us calculate the distances between W and the intersection points M'', N'', M''', N'''.

Assume that $|\overline{WZ}| = 2$, $x = |\overline{WM}| = |\overline{WM'}|$, $y = |\overline{WN}| = |\overline{WN'}|$. Then it is easy to show that $x \cdot y = 2$ because, due to (18), the number 2 is the power of W

with respect to $\mathcal{D}_1(Z,W)$ circle. Then the distances are as follows:

$$|\overline{WM''}| = |\overline{WM'}|\frac{|\overline{WM}|}{|\overline{WZ'}|} = \frac{x^2}{2}, \qquad |\overline{WN''}| = |\overline{WN'}|\frac{|\overline{WM}|}{|\overline{WZ'}|} = \frac{xy}{2} = 1,$$

$$|\overline{WM'''}| = |\overline{WM'}|\frac{|\overline{WN}|}{|\overline{WZ'}|} = \frac{xy}{2} = 1, \quad |\overline{WN'''}| = |\overline{WN'}|\frac{|\overline{WN}|}{|\overline{WZ'}|} = \frac{y^2}{2}.$$

Thus $S = M''' = N''$ is the midpoint of the segment $\overline{WZ}$ and S belongs to $\mathcal{D}_1(W,C)$ for $C \in \overline{MN}$, i.e. for any $c_{i+1} = C \in \mathcal{D}_1(Z,W)$. In other words, $S = (c_i + c_{i-1})/2 \in \mathcal{D}_1(c_i, c_{i+1})$. The set $\mathcal{C}$ is convex, $W, Z \in \mathcal{C}$, so $S \in \mathcal{C}$ and

$$\frac{c_i + c_{i-1}}{2} \in \mathcal{C} \cap (\mathcal{D}_1 \cup \mathcal{D}_2), \quad \mathcal{D}_1, \mathcal{D}_2 \text{ for } c_i, c_{i+1}.$$

Therefore, the proposition is true for c_i, c_{i+1}. Now, since by assumption c_{-1}, c_0 are chosen so that it is possible to determine c_1, then by induction c_{i+1} exists for any i. ■

Corollary 2. *Let c_i and c_{i+1} be two consecutive solutions from the algorithm. Then c_{i-1} belongs to the dilated image, with respect to c_i and by factor 2, of one of the conditioning discs $\mathcal{D}_1, \mathcal{D}_2$ for c_i, c_{i+1}.*

Proof. We have shown in the proof of Proposition 2 that $S = (c_i + c_{i-1})/2 \in \mathcal{C} \cap (\mathcal{D}_1(c_i, c_{i+1}) \cup \mathcal{D}_2(c_i, c_{i+1}))$, i.e. S belongs to one of the conditioning discs, say $\mathcal{D}_1(c_i, c_{i+1})$. Since c_{i-1} is the image of S in dilation with respect to c_i and by factor 2, c_{i-1} must belong to the corresponding dilated image of $\mathcal{D}_1(c_i, c_{i+1})$. ■

Corollary 3. *Let c_i and c_{i+1} be given. Let $S = (c_i + c_{i-1})/2$ and c_{i-1} belong to $\mathcal{D}_1(c_i, c_{i+1})$ and to the corresponding dilated image $\mathcal{D}'_1(c_i, c_{i+1})$, respectively. Let E be that common point of $\mathcal{D}_1(c_i, c_{i+1})$ and the line going through c_i and tangent to $\mathcal{D}_1(c_i, c_{i+1})$ circle for which $|\sphericalangle E c_i c_{i+1}| < |\sphericalangle F c_i c_{i+1}|$, where F is the second point having this property. Then $E \in \mathcal{C} \cap \mathcal{D}_1(c_i, c_{i+1})$.*

Proof. The proof is based on Fig. 3:

Let Z, W, C, S, M denote c_{i-1}, c_i, c_{i+1}, $(c_{i-1} + c_i)/2$ and $(c_i + c_{i+1})/2$, respectively. Let D be the common point of the $\mathcal{D}_1(W,C)$ circle and the line going through M tangent to $\mathcal{D}_1(W,C)$, as shown in Fig. 3. Let P be the intersection point of the segments $\overline{MD}$ and $\overline{WE}$. Let $M' = C, E', P', D', \mathcal{D}'_1(W,C)$ denote the images of the objects $M, E, P, D, \mathcal{D}_1(W,C)$ in dilation with respect to W and by factor 2. By assumption, $\mathcal{D}_1(W,C)$ is the conditioning disc for W, C to the image of which $Z = c_{i-1}$ belongs: $Z \in \mathcal{D}'_1(W,C)$, see Corollary 2.

Since $Z \in \mathcal{D}'_1(W,C)$, Z, W, C all belong to $\mathcal{C}$, $\mathcal{C}$ is a convex set, and the following inclusions hold (where $\sphericalangle$ denotes a wedge-shaped set):

$$\sphericalangle WCP' \subseteq \sphericalangle WCZ, \quad \sphericalangle CWP' \subseteq \sphericalangle CWZ,$$

$$\triangle CWP' = \sphericalangle CWP' \cap \sphericalangle WCP', \quad \triangle CWZ = \sphericalangle CWZ \cap \sphericalangle WCZ,$$

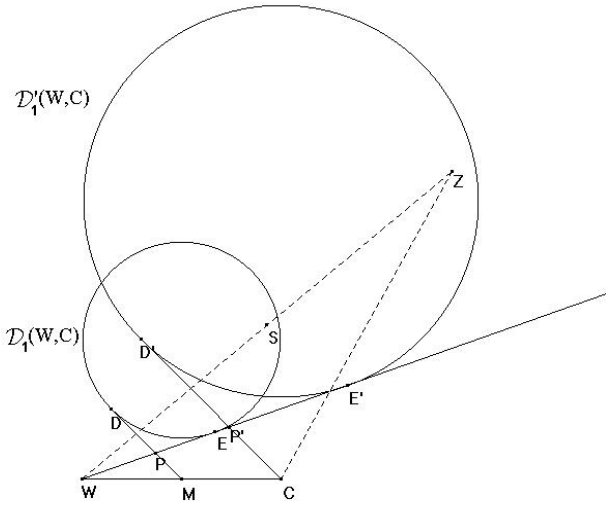


Fig. 3. Feasibility of E : $E \in \mathcal{C} \cap \mathcal{D}_1(W, C)$.

we conclude that

$$\triangle CWP' \subseteq \triangle CWZ \subseteq \mathcal{C}.$$

Assume that $|\overline{WC}| = 2$. Then from (18) we have $|\overline{WE}| = \sqrt{2}$, $|\overline{MD}| = 1$. Let $x = |\overline{WP}|$. Then $|\overline{PE}| = |\overline{PD}| = \sqrt{2} - x$, so $|\overline{PM}| = |\overline{MD}| - |\overline{PD}| = 1 - \sqrt{2} + x$. For the triangle $\triangle WMP$ we have: $|\overline{WP}| + |\overline{PM}| > |\overline{WM}| \iff 2x + 1 - \sqrt{2} > 1 \iff x > \sqrt{2}/2$. Thus $|\overline{WP'}| = 2|\overline{WP}| > \sqrt{2}$. The last inequality implies that $E \in \overline{WP'} \subset \triangle CWP' \subseteq \mathcal{C}$. This completes the proof. ■

From Proposition 2 and Corollary 3 we have that $S = (c_{i-1} + c_i)/2$ and $E(c_i, c_{i+1})$ both belong to the product $\mathcal{C} \cap \mathcal{D}_1(c_i, c_{i+1})$, where $\mathcal{D}_1$ is one of the discs of the conditioning set defined by c_i and c_{i+1}. This means that they are good candidates for a starting point of an optimization procedure necessary to perform Step 3 of the algorithm. On the other hand, nothing is known about the product $\mathcal{C} \cap \mathcal{D}_2(c_i, c_{i+1})$ and it can be empty.

4. Convergence Analysis

Writing

$$q_m(c, \alpha) = Q(c, F(c, \alpha)), \tag{19}$$

$$q(c) = Q(c, F_*(c)), \tag{20}$$

and using (6), (7) the problem MMOPL (8) can be easily reformulated to the following equivalent but simpler form (MMOPL1):

$$\begin{aligned} &\text{minimize } \left\{q(c_i) + \nabla q^T(c_i)(c - c_i) + \rho\|c - c_i\|^2\right\} \\ &\text{subject to } c \in \mathcal{C}, \end{aligned} \tag{21}$$

where

$$\nabla q(c_i)^T = \nabla_c Q(c_i, y_i)^T + \nabla_y Q(c_i, y_i)^T F_*^{'}(c_i),$$
$$F_*^{'}(c_i) = \left[\nabla F_{*1}(c_i) \cdots \nabla F_{*m}(c_i)\right]^T,$$

m denoting the number of outputs, $y \in \mathbb{R}^m$, $\nabla_x f$ denoting the gradient of a function f with respect to the variables x.

The simplified ISOPED algorithm for $\dim c = 2$, with the conditioned problem CMMOPL instead of the CMMOP, takes now the following form:

Algorithm 2. The ISOPEDL algorithm (ISOPED with Linearization of the performance function used):

1. Choose $c_{-2}, c_{-1}, c_0 \in \mathcal{C} \subset \mathbb{R}^2$ such that the condition number of the matrix $A(c_0) = [\ c_0 - c_{-1} \quad c_0 - c_{-2}\ ]$ is not greater than a, set $i := 0$,
2. Measure $F_*(c_i)$, evaluate $q(c_i) = Q(c_i, F_*(c_i))$ and estimate $\nabla q(c_i)$.
3. Solve the following CMMOPL problem:

$$\text{minimize } \left\{f(v) = q(c_i) + \nabla q(c_i)^T (v - c_i) + \rho \|v - c_i\|^2\right\}$$
$$\text{subject to } v \in \mathcal{C} \cap (\mathcal{D}_1 \cup \mathcal{D}_2), \tag{22}$$

   where

$$\mathcal{D}_1 \cup \mathcal{D}_2 = \left\{v \in \mathbb{R}^2 : \frac{\sigma_{\max}(A(v))}{\sigma_{\min}(A(v))} \le a\right\}, \tag{23}$$

   and $\sigma_{\max}(A(v))$, $\sigma_{\min}(A(v))$ are the maximal and minimal singular value, respectively, of the matrix $A(v) = [\ v - c_i \quad v - c_{i-1}\ ]$. Then set $c_{i+1} := \arg\min_{\mathcal{C} \cap (\mathcal{D}_1 \cup \mathcal{D}_2)} f(v)$.
4. If $\|c_{i+1} - c_i\| < \varepsilon$ then STOP, otherwise set $i := i + 1$ and go to Step 2.

Our intention is to estimate the gradient $\nabla q(c_{i+1})$ in the next iteration solving the set of linear equations with respect to $\nabla q(v)$, for $v = c_{i+1}$:

$$A(v)^T \nabla q(v) = \begin{bmatrix} q(v) - q(c_i) \\ q(v) - q(c_{i-1}) \end{bmatrix} \Leftrightarrow A(v)^T \nabla q(v) = b(v). \tag{24}$$

The convergence analysis presented in this section is limited to the unconstrained case, $\mathcal{C} = \mathbb{R}^2$. We will also assume that the gradient estimation is precise, i.e. the gradient $\nabla q(c_i)$ at each c_i will be assumed as known (the same was also assumed in the convergence analysis of the basic ISOPE algorithm in (Brdyś *et al.*, 1987). The novelty are first results of a convergence analysis of the dual ISOPE algorithm,

i.e. with the conditioning set $\mathcal{D}_1 \cup \mathcal{D}_2$. The results show the impact of this set on the convergence and in particular, the influence of the parameter a (see (23)) on the algorithm behaviour – a result important from the point of view of application.

Before stating the theorem a parameter $\beta_{\max}$ will be defined, using the drawing shown in Figure 4.

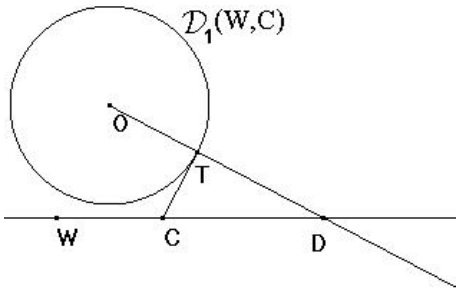


Fig. 4. Definition of $\beta_{\max}$.

$\mathcal{D}_1(W,C)$ is one of the discs of the conditioning set for $W = c_{i-1}$, $C = c_i$ (two consecutive solutions from the algorithm), and let $\mathcal{L}(C,T)$ be the line going through C and tangent to $\mathcal{D}_1(W,C)$, where T is the common point of $\mathcal{D}_1(W,C)$ and $\mathcal{L}(C,T)$. O is the centre of $\mathcal{D}_1(W,C)$, and D is the intersection point of the line $\mathcal{L}(WC)$ and the ray $\mathcal{R}(OT)$. Since there are two possible choices for $\mathcal{L}(C,T)$, we choose T to be such that $D \notin \overline{WC}$. Now $\beta_{\max} = |\overline{OD}|/|\overline{OT}|$, $1 < \beta_{\max} < \infty$ (if it exists). Naturally, $\beta_{\max}$ depends on a.

Theorem 1. *Let* $q : \mathbb{R}^2 \to \mathbb{R}$ *be a differentiable function satisfying the following conditions:*

(A1) *there exists* $\overline{q} \in \mathbb{R}$ *such that* $q(c) \geq \overline{q}$ *for all* $c \in \mathbb{R}^2$,

(A2) $\|\nabla q(x) - \nabla q(y)\| \leq L\|x - y\|$ *for a certain constant* L *and for all* $x, y \in \mathbb{R}^2$ *(Lipschitz continuity of* $\nabla q(c)$*),*

and let the parameters of the algorithm a, ρ *and a constant* B *satisfy the inequalities:*

(A3) $\beta_{\max}(a) < 5$, *i.e.* $a > \widehat{a} = \widehat{h} + \widehat{r} = \sqrt{\widehat{r}^2 + 1} + \widehat{r}$ *where* $\widehat{r} = \sqrt{1 + \sqrt{14}/4}$ $(\widehat{a} \simeq 3.11)$,

(A4) $\rho > L/(5 - \beta_{\max})$ *(it implies* $\rho(\beta_{\max} - 1)/2 < (4\rho - L)/2$ *and* $4\rho > L$ *because* $\beta_{\max} > 1$*),*

(A5) $\rho(\beta_{\max} - 1)/2 < B < (4\rho - L)/2$.

Then, defining the energy function $\mathcal{E}(c_i, c_{i-1}) = q(c_i) + B\|c_i - c_{i-1}\|^2$, *we have at consecutive iterations of the algorithm:*

1. $\mathcal{E}(c_{i+1}, c_i) < \mathcal{E}(c_i, c_{i-1})$
2. $\lim_{i\to\infty} \|c_{i+1} - c_i\| = 0$ *and* $\sum_{i=0}^{\infty} \|c_{i+1} - c_i\|^2 < \infty$.

Proof. See the Appendix. ■

Note that we are interested in convergence results for a broad range of values for the parameter a, including all values practically important from the point of view of application. Certainly, the larger a, the better the chance to prove the convergence, because a larger a means a smaller influence of the conditioning constraint on the modified model optimization problems—and the algorithm without this constraint has been proved to be convergent. Luckily, to prove the main result concerning the algorithm optimality, i.e. the convergence of the gradient to zero, only a slightly larger value of the threshold for a than that in Assumption (A3) must be assumed.

Theorem 2. *If Assumptions (A1)–(A5) of Theorem 1 are satisfied, and additionally*

$$a > \widetilde{a} = \widetilde{h} + \widetilde{r} = \sqrt{\widetilde{r}^2 + 1} + \widetilde{r} \quad \textit{where} \quad \widetilde{r} = \frac{1+\sqrt{5}}{2}, \quad \textit{i.e.} \quad \widetilde{a} \simeq 3.52,$$

then

$$\lim_{i \to \infty} \|\nabla q(c_i)\| = 0.$$

Proof. The proof is omitted since it is lengthy and uses a reasoning similar to that in the proof of Theorem 1. ■

Theorem 2 implies that if $\lim_{i\to\infty} c_i = c$, then, due to the assumed Lipschitz continuity of the gradient, c is a stationary point, i.e. $\nabla q(c) = \mathbf{0}$. The result concerning the threshold value of $a = \widetilde{a} \simeq 3.52$ is very important from the practical point of view, because $\widetilde{a}$ is quite small, allowing a very good conditioning of the matrix A (recall the smallest possible value yielding an ideal conditioning is $a = 1$). Certainly, for $a \leq \widetilde{a}$ (and also $a \leq \widehat{a}$) the norm $\|c_i - c_{i-1}\|$ may also happen to converge to zero. Moreover, for a small enough i.e. for $a < (1+\sqrt{5})/2$ (which is equivalent to $r < 1/2$) the sequence $\{c_i\}$ is convergent because $\max(\|c_{i+1} - c_i\|/\|c_i - c_{i-1}\|) = (l+r)/2 < 1$. However, we cannot then assure the optimality properties of the algorithm, because the smaller a the larger the influence of the conditioning constraint on the next point, and this influence finally destroys the optimality of the algorithm—for the smallest possible value of $a = 1$ (and the best possible conditioning of the matrix A) the feasible set is reduced to two points only, because the conditioning discs reduce then to single points (centres).

Certainly, the theorem assumptions are sufficient conditions—it has not been proved that for certain values of a smaller than $\widetilde{a}$ the convergence and optimality cannot happen. Moreover, it has happened in certain example simulations with a slightly smaller than $\widetilde{a}$.

5. Simulation Results

A nonlinear process described by the following input-output mapping, which is assumed to be unknown, is considered (cf. Brdyś and Tatjewski, 1994):

$$y = F_*(c) = F_*\left([c_{(1)}, c_{(2)}]^T\right) = 2c_{(1)}^{0.5} + c_{(2)}^{0.4} + 0.2c_{(1)}c_{(2)}.$$

The performance function to be minimized is described by

$$Q(c, y) = -y + (c_{(1)} - 0.5)^2 + (c_{(2)} - 0.5)^2.$$

The following process model is assumed to be available:

$$y = F(c, \alpha) = F\left([c_{(1)}, c_{(2)}]^T, \alpha\right) = 0.6c_{(1)} + 0.4c_{(2)} + \alpha.$$

The real optimal point is $\widehat{c} = [\widehat{c}_{(1)}, \widehat{c}_{(2)}]^T = [1.067, 0.830]^T$ with $Q(\widehat{c}, F_*(\widehat{c})) = -2.7408$. It is located within the interior of the feasibility set

$$\mathcal{C} = \left\{c = [c_{(1)}, c_{(2)}]^T \in \mathbb{R}^2 : \ 0 \le c_{(1)} \le 2, \quad 0 \le c_{(2)} \le 2\right\}.$$

To implement the ISOPED or ISOPEDL algorithms efficiently, a choice of initial points $c_{-2}, c_{-1}, c_0 \in \mathcal{C} \subset \mathbb{R}^2$ (see Algorithm 2) satisfying the requirement $\mathrm{cond}\,(A(c_0)) \le a$ must be appropriately designed. Only one current set-point corresponding to actual uncertainty conditions is usually available when starting the algorithm. It is then the task of the algorithm itself to gather the data necessary to start regular iterations. Gathering these data is called the *initial phase* of the algorithm. The initial phase should be thoroughly designed because it must apply set-point changes, and each set-point change means a transient process in the plant and is connected with plant productivity. Therefore, an optimized initial phase was proposed in (Tatjewski, 1998), for a general case with $c \in \mathbb{R}^n$. It is given below for the case $c \in \mathbb{R}^2$ to be consistent with the theoretical convergence analysis of the paper.

The optimized initial phase of the ISOPED algorithm (*Step 1*):

1.1. Choose appropriately positive parameters γ, a. Set $c_{-2} := c_0$, the actual point.

1.2. Solve the following augmented model optimization problem (MOPA):

$$\begin{aligned} &\mathrm{minimize}_c\left\{Q(c, F(c, \alpha_{-2})) + \rho_0 \|c - c_{-2}\|^2\right\} \\ &\text{subject to } \ c \in \mathcal{C}, \quad \|c - c_{-2}\| \ge \gamma, \end{aligned} \tag{25}$$

denoting the solution point by c_{-1}. Apply the set-point c_{-1} to the controlled plant and measure the corresponding outputs. Add the measurement to the data record and adapt the steady-state model (i.e. the parameters α).

1.3. Solve the following conditioned augmented model optimization problem (CMOPA):

$$\begin{aligned} &\mathrm{minimize}_c\left\{Q(c, F(c, \alpha_{-1})) + \rho_0 \|c - c_{-1}\|^2\right\} \\ &\text{subject to } \quad c \in \mathcal{C}, \quad \|c - c_{-1}\| \ge \gamma, \\ &\qquad\qquad\ \ \mathrm{cond}\,(A(c)) \le a, \end{aligned} \tag{26}$$

denoting by c_0 the solution point. Go to Step 2 of the ISOPED algorithm.

The initial phase for the ISOPEDL version is analogous and only linearization of the function $Q(c, F(c, \alpha_j))$ at every point c_j $(j = -2, -1)$ must be used. The ISOPED and ISOPEDL algorithms with optimized initial phase were used in all simulation experiments. Note that generally, a different, larger value of the penalty coefficient ρ should be used during the optimized initial phase, especially when the model uncertainty is significant. The reason is that the MOPA and CMOPA problems are model-based only (without the modifier λ carrying a feedback information from the process), therefore larger set-point changes should be avoided. The penalty coefficient for the initial phase is denoted by ρ_0, and $\rho_0 = 2\rho$ was taken in the simulation experiments.

First, the influence of the parameter a on the convergence properties of the ISOPEDL algorithm was tested. Sample results for $a = 4$, $a = 3$ and $a = 2$ (all with $\rho = 1$) are shown in Figs. 5 and 6, in the form of the process performance function values ('qre') and set-point trajectories, respectively. For $a = 4$ $(> \widetilde{a} \simeq 3.52)$ the convergence is to the optimal point and it is very good; convergence to the optimum also occurs for $a = 3$, but for $a = 2$ the algorithm loses the optimality property, converging more slowly and to a point far away from the optimal one. The results confirm the statements of the Theorem 2 and show the guaranteed convergence for $a > \widetilde{a} \simeq 3.52$. Observe an increasing zig-zag nature of the set-point trajectory with the decrease in a.

It is an interesting and important question how close the behaviour and convergence properties of the ISOPED and ISOPEDL algorithms are. Intuitively, they should differ when the process model enters the performance function and is significantly non-linear, especially in regions of larger changes in the set-points—but should be analogous for small set-point steps. To test this hypothesis, both the algorithms were simulated for different values of ρ and a; sample results for $\rho = 1$ are given in Figs. 7 and 8, and for $\rho = 0.5$ in Figs. 9 and 10. In both the cases $a = 10$ was used, a value usually sufficient under a reasonable error level in the feedback information. These results show that the ISOPEDL algorithm works properly for values of ρ slightly larger than the ISOPED algorithm ($\rho = 0.5$ is clearly too small for ISOPEDL), and that the convergence properties of both the algorithms seem to be analogous. The behaviour of the ISOPED algorithm for different values of $a = 4$, $a = 3$ and $a = 2$ (all with $\rho = .2$) shown in Figs. 11 and 12 further confirms this statement, cf. Figs. 5 and 6.

6. Conclusions

Basic theoretical properties of the dual-type ISOPE (called ISOPED) algorithm have been considered. The problem is very difficult due to the complicated nature of the algorithm, and therefore, it has not been possible until now to obtain results for the two-dimensional case, $\dim c = 2$.

First, properties of the feasibility set composed of original and conditioning constraints were investigated. The structure of the conditioning set was derived (Corollary 1). Then it was proved that adding the conditioning constraints—which vary

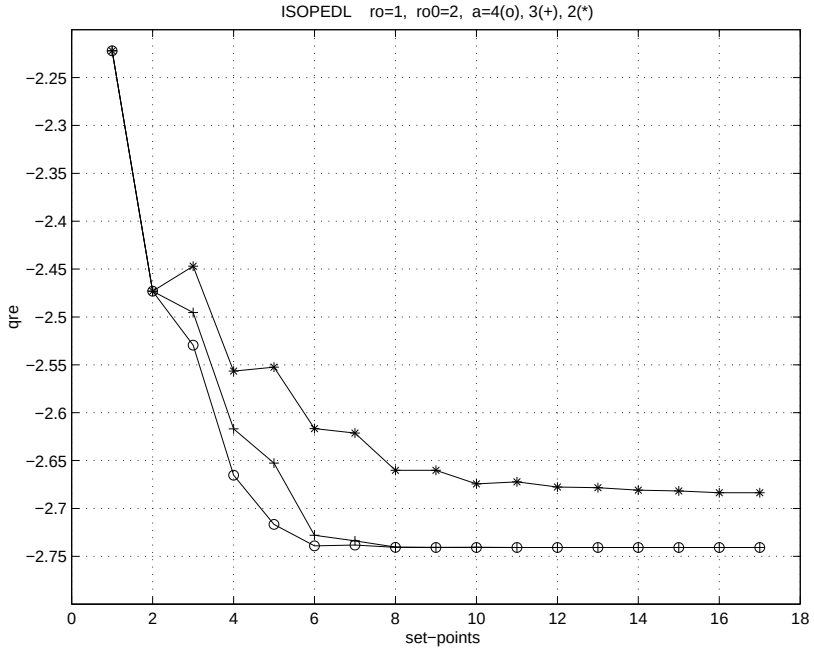


Fig. 5. Performance function trajectories for different values of a.

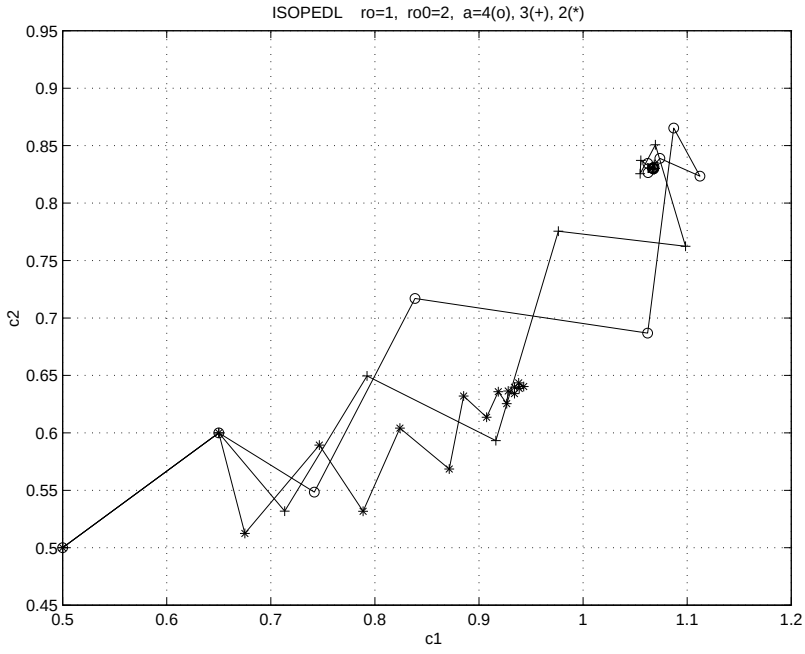


Fig. 6. Set-point trajectories for different values of a.

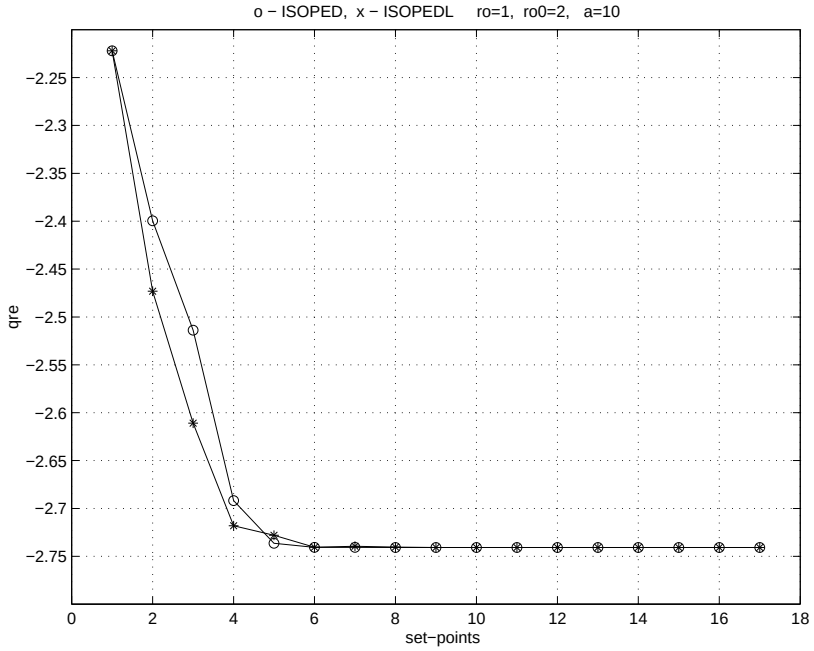


Fig. 7. Performance function trajectories of ISOPED and ISOPEDL algorithms for $\rho = 1$.

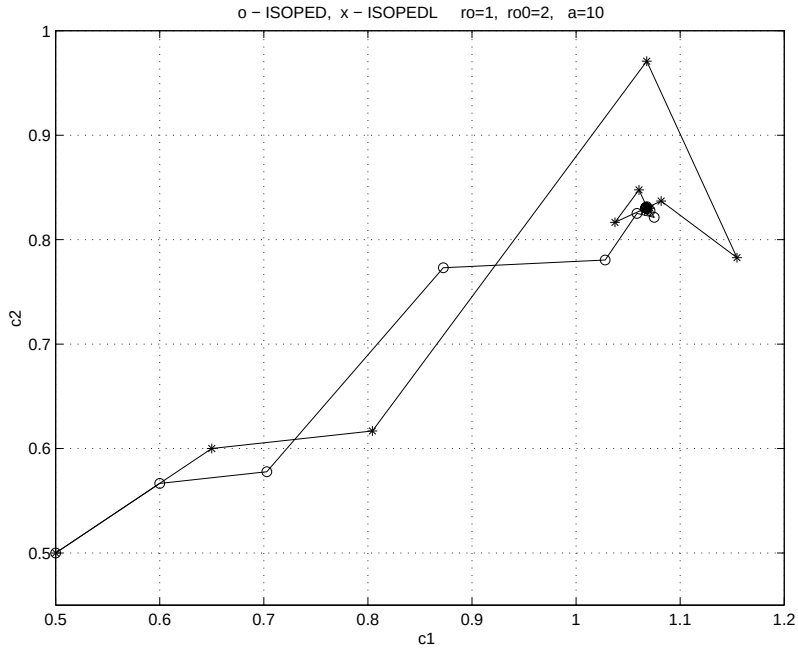


Fig. 8. Set-point trajectories of ISOPED and ISOPEDL algorithms for $\rho = 1$.

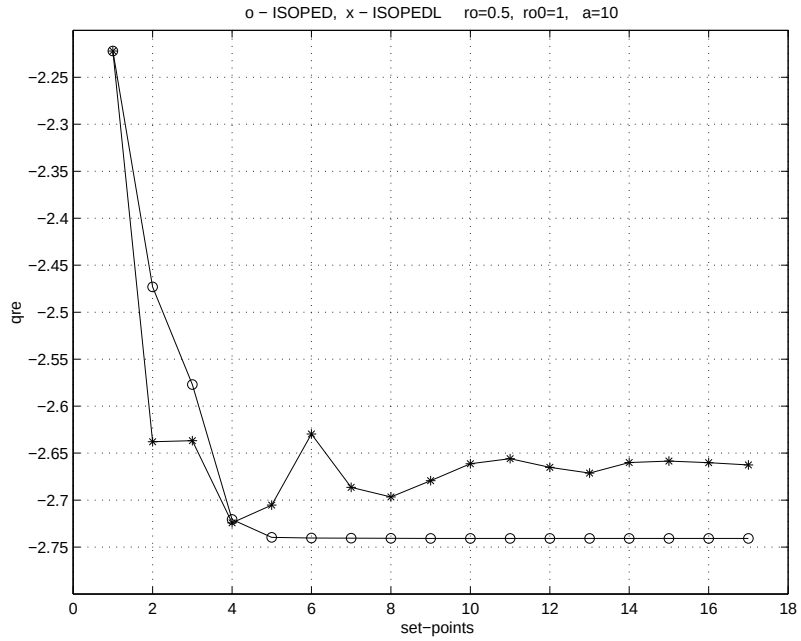


Fig. 9. Performance function trajectories of ISOPED and ISOPEDL algorithms for $\rho = 0.5$.

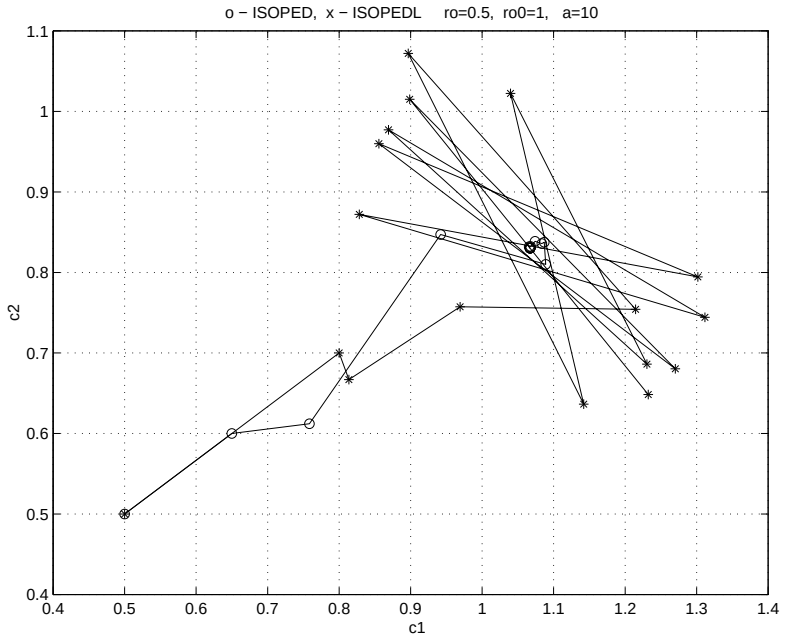


Fig. 10. Set-point trajectories of ISOPED and ISOPEDL algorithms for $\rho = 0.5$.

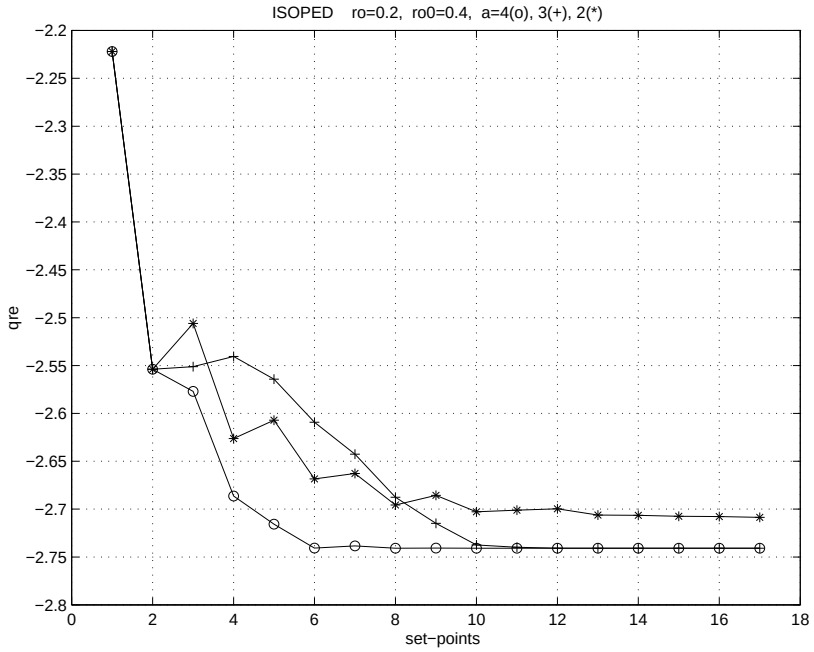


Fig. 11. Performance function trajectories for different values of a (ISOPED).

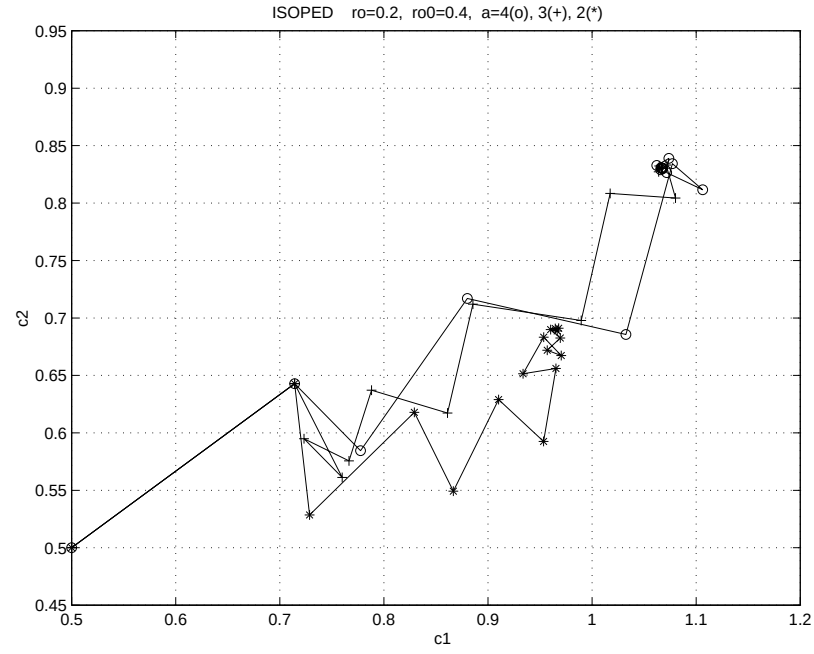


Fig. 12. Set-point trajectories for different values of a (ISOPED).

with iterations—cannot make the overall feasible set empty (Proposition 2), which constitutes a very important result from the point of view of application. Second, we managed to obtain results concerning convergence properties of the ISOPEDL algorithm, a version of the ISOPED algorithm with the performance function of the modified model optimization problems linearized at each iteration point. For the first time, a threshold value of the parameter of the right-hand side of the conditioning constraint was found ($a \simeq 3.52$) such that it guarantees, under other reasonable conditions, that the gradient of the process performance function converges to zero (Theorems 1 and 2, the unconstrained case). The result is of practical importance because the *threshold value lies well outside a numerically recommended range of values for* a (about 10, down to about 5 for high feedback error levels). The results of numerical simulations with ISOPEDL and ISOPED algorithms were also presented, fully confirming the obtained theoretical results. Moreover, these results showed similar behaviour and practically identical convergence properties of both the algorithms in the considered example.

Certainly, there is still much more to be done. We did not manage to perform a convergence analysis in the constrained case, i.e. when the optimal point lies on the boundary of the original feasibility set $\mathcal{C}$, although simulation results with simple lower-upper bound constraints indicate that these properties are similar. Second, cases with $\dim c \geq 3$ should be analyzed. The approach should then be perhaps somehow different, relying less on geometrical analysis. Finally, the convergence of the ISOPED algorithm itself should be theoretically investigated. Or perhaps it could be shown that convergence properties of ISOPEDL and ISOPED algorithms are closely connected, as simulation results indicate.

Acknowledgments

The authors are grateful to the anonymous referee for the comments which led to a significant improvement of the presentation. The work was partially supported by the Dean of the FEIT under grant No. 503/G/0180/200.

References

Bertsekas D.P. (1995): *Nonlinear Programming*. — Belmont: Athena Scientific.

Brdyś M., Ellis J.E. and Roberts P.D. (1987): *Augmented integrated system optimization and parameter estimation technique: Derivation, optimality and convergence*. — IEE Proc.-D, Vol.134, No.3, pp.201–209.

Brdyś M. and Tatjewski P. (1994): *An algorithm for steady-state optimizing dual control of uncertain plants*. — Proc. 1st IFAC Workshop *New Trends in Design of Control Systems*, Smolenice, Slovakia, pp.249–254.

Findeisen W., Bailey F.N., Brdyś M., Malinowski K., Tatjewski P. and Woźniak A. (1980): *Control and Coordination in Hierarchical Systems*. — Chichester: Wiley.

Kiełbasiński A. and Schwetlick H. (1992): *Numerical Linear Algebra*. — Warsaw: WNT (in Polish).

Roberts P.D. (1979): *An algorithm for steady-state system optimization and parameter estimation.* — Int. J. Syst. Sci., Vol.10, No.7, pp.719–734.

Stark M. (1974): *Analytical Geometry with an Introduction to Multidimensional Geometry.* — Warsaw: Polish Scientific Publishers (in Polish).

Tatjewski P. (1998): *Two-phase dual-type optimising control algorithm for uncertain plants.* — Proc. 5th Int. Symposium *Methods and Models in Automation and Robotics MMAR'98*, Międzyzdroje, Poland, pp.171–176.

Tatjewski P. (1999): *Two-phase dual-type optimising control algorithm for uncertain plants with active output constraints.* — Proc. *European Control Conference ECC'99*, Karlsruhe, Germany, paper FO 347 (published on CD-ROM).

Zhang H. and Roberts P.D. (1990): *On-line steady-state optimization of nonlinear constrained processes with slow dynamics.* — Trans. Inst. MC, Vol.12, No.5, pp.251–261.

Appendix

Proof of Theorem 1. Note that solving the CMMOPL optimization problem in Step 3 of the algorithm can be considered as a projection of the result $p = c - \nabla q(c)/2\rho$ (where $c = c + i$) of the unconstrained optimization onto the conditioning set $\mathcal{D}$ (i.e. onto the nearest from the two discs), because the level sets of $f(v)$ are circles around p.

Let w, c and v denote c_{i-1}, c_i and c_{i+1}, respectively. We will estimate the expression

$$\mathcal{E}(v,c) - \mathcal{E}(c,w) \tag{27}$$

for two cases concerning the location of v.

Case 1. *v belongs to the $\mathcal{D}_1(w,c)$ circle (the boundary of the $\mathcal{D}_1(w,c)$ disc, see Fig. 13).*

Denote by v and v' the results of projection of the unconstrained optimization result p at Step 3 of the algorithm: p ($\overrightarrow{cp} = p - c = -\nabla q(c)/2\rho$) is projected onto the discs $\mathcal{D}_1(w,c)$ and $\mathcal{D}_2(w,c)$. We assume that $v \in \mathcal{D}_1$ is the solution, i.e. $\|v-p\| < \|v'-p\|$. This means that p and v belong to the same half-plane with respect to the line $\mathcal{L}(wc)$. If p lies on this line, either of the points v and v' can be chosen as the result of Step 3 of the algorithm. From now on we will regard p as the result of dilating v with respect to o_1 (the centre point of $\mathcal{D}_1$): $p = o_1 + \beta(v - o_1)$, $\beta \geq 1$, since o_1, v, p are colinear.

We introduce new orthogonal axes in such a way that o_1 is the origin, the x-axis is parallel to the vector $\overrightarrow{wc}$ (directed to the right), and the y-axis is parallel to the vector $\overrightarrow{mo_1}$ (directed upward). The scale preserves all the distances, i.e. the transition between the original and new coordinates is isometric. In the new axes $\overline{oxy}$: $\overline{o_1} = (0,0)$; $\overline{m} = (\|w-c\|/2)(0,-h)$; $\overline{c} = (\|w-c\|/2)(1,-h)$; $\overline{v} = (\|w-c\|/2)(r\cos(\phi), r\sin(\phi))$; $\overline{p} = (\|w-c\|/2)(\beta r\cos(\phi), \beta r\sin(\phi))$ where $\beta \geq 1$, and

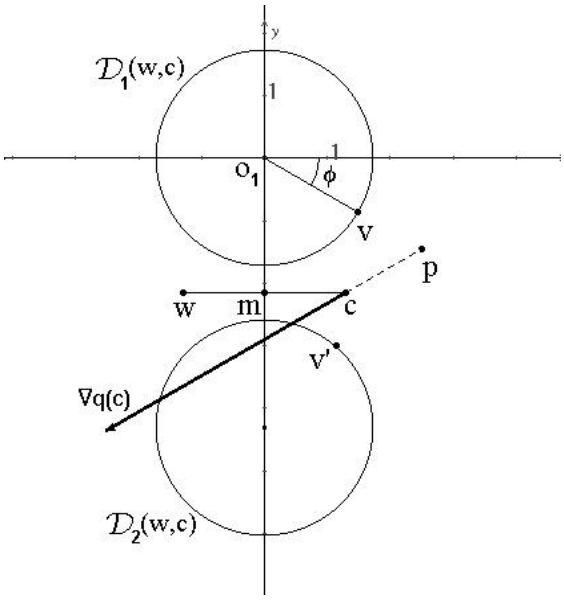


Fig. 13. Projection of p onto $\mathcal{D}_1$ and $\mathcal{D}_2$.

finally $\overline{\nabla q(c)} = 2\rho(\overline{c} - \overline{p}) = (\|w-c\|/2)2\rho(1-\beta r\cos(\phi), -h-\beta r\sin(\phi))$ where r and h satisfy (18). Further, based on (18), we have

$$\begin{aligned}\|v-c\|^2 = \|\overline{v}-\overline{c}\|^2 &= \left\|\frac{\|w-c\|}{2}\big(r\cos(\phi)-1, r\sin(\phi)+h\big)\right\|^2 \\ &= \frac{\|w-c\|^2}{4}\big(2(r^2 - r\cos(\phi) + rh\sin(\phi)) + 2\big) \\ &= \frac{\|w-c\|^2}{4}\big(2\mathcal{F}(\phi)+2\big), \end{aligned} \tag{28}$$

$$\begin{aligned}\nabla q(c)^T(v-c) &= \overline{\nabla q(c)}^T(\overline{v}-\overline{c}) \\ &= 2\rho\frac{\|w-c\|^2}{4}\big[(1-\beta r\cos(\phi))(r\cos(\phi)-1) \\ &\quad + (-h-\beta r\sin(\phi))(r\sin(\phi)+h)\big] \\ &= \frac{\|w-c\|^2}{4}\big[-4\rho - 2\rho(1+\beta)(r^2 - r\cos(\phi) + rh\sin(\phi))\big] \\ &= \frac{\|w-c\|^2}{4}\big[-4\rho - 2\rho(1+\beta)\mathcal{F}(\phi)\big]. \end{aligned} \tag{29}$$

Now, based on the estimation of the descent of a function with Lipschitz continuous gradient (see, e.g. Bertsekas, 1995), we estimate the expression (27) for an

arbitrary positive B. For notational convenience, we write $\mathcal{E}(v)$ and $\mathcal{E}(c)$ instead of $\mathcal{E}(v,c)$ and $\mathcal{E}(c,w)$, and generally, $\mathcal{E}(c_i)$ instead of $\mathcal{E}(c_i,c_{i-1})$.

We thus get

$$\begin{aligned}\mathcal{E}(v)-\mathcal{E}(c) &= q(v)-q(c)+B(\|v-c\|^2-\|c-w\|^2)\\ &\le \nabla q(c)^T(v-c)+\frac{L}{2}\|v-c\|^2+B(\|v-c\|^2-\|c-w\|^2)\\ &\le \frac{\|w-c\|^2}{4}(-4\rho-2\rho(1+\beta)\mathcal{F}(\phi))-B\|w-c\|^2\\ &\quad+\left(B+\frac{L}{2}\right)\frac{\|w-c\|^2}{4}(2\mathcal{F}(\phi)+2)\\ &= \frac{\|w-c\|^2}{4}\big(\mathcal{F}(\phi)(L+2B-2\rho(1+\beta))+L-2B-4\rho\big). \quad (30)\end{aligned}$$

Let us investigate $\mathcal{F}(\phi)=r^2-r\cos(\phi)+rh\sin(\phi)$. From (28) we have $\mathcal{F}(\phi)=2(\|v-c\|^2/\|w-c\|^2)-1$ and it is a periodic function of ϕ as v rotates around the centre of $\mathcal{D}_1$. Note that (see Fig. 14) due to the geometric properties of $\mathcal{D}_1$ with respect to w and c:

$$\begin{aligned}&\mathcal{F}(\phi) \text{ is maximal } \iff \phi=\phi_{\max} \iff v=v_{\max},\\ &\mathcal{F}(\phi) \text{ is minimal } \iff \phi=\phi_{\min} \iff v=v_{\min},\\ &\mathcal{F}(\phi)=0 \Leftrightarrow \phi=\phi_1 \text{ or } \phi=\phi_2 \iff v=t_1 \text{ or } v=t_2. \quad (31)\end{aligned}$$

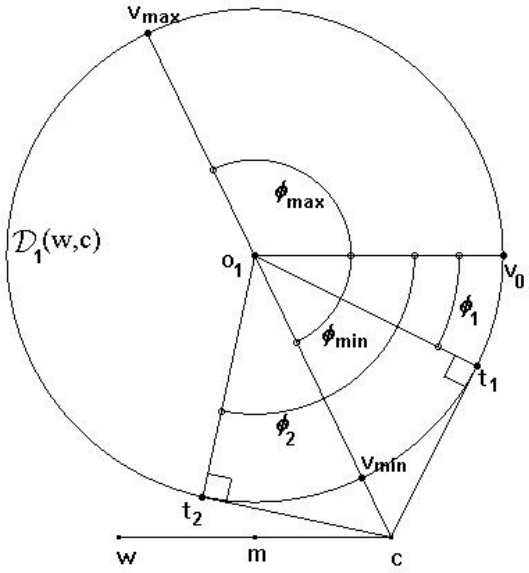


Fig. 14. $\mathcal{F}(\phi)$ depends on the position of v.

Therefore $\mathcal{F}(\phi) < 0$ or $\mathcal{F}(\phi) > 0$ when v belongs to the shorter arc t_1t_2 or the longer arc t_1t_2, respectively. Also note that if v is, in $\overline{oxy}$, not lower on the $\mathcal{D}_1$ circle than $v_0(\phi = 0)$, the factor β designating the position of p can be made arbitrarily large. Otherwise β is bounded by β' (Fig. 15) so that p stays above the line $\mathcal{L}(wc)$, in $\overline{oxy}$. Here β' is given by

$$\beta' = -\frac{h}{r\sin(\phi)}$$

for v positioned at an angle ϕ on the $\mathcal{D}_1$ circle: $\overline{v} = (\|w-c\|/2)(r\cos(\phi), r\sin(\phi))$, $\phi \in (-\pi, 0)$.

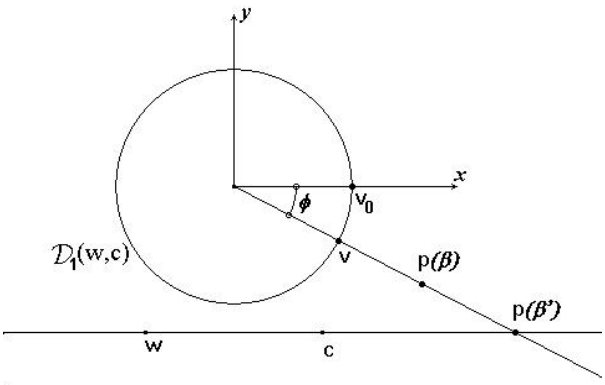


Fig. 15. Definition of β'.

Now, the second component of (30) satisfies the inequality $(\|w-c\|^2/4)(L-2B-4\rho) < 0$, because from the assumptions of Theorem 1 we have $B > 0$ and $4\rho - L > 0$.

We have to analyze two situations for the first component:

1. $\mathcal{F}(\phi) \geq 0$ (see (31)). In order to have for the first component of (30) the relation $(\|w-c\|^2/4)(L+2B-2\rho(1+\beta))\mathcal{F}(\phi) < 0$, there must be (since $\beta \geq 1$)

$$B < \frac{4\rho - L}{2} \tag{32}$$

   and it is so as assumed in (A5). Thus both the components of (30) are negative.

2. $\mathcal{F}(\phi) < 0$, i.e. v belongs to the shorter arc t_1t_2 (Fig. 14).

   We require that the parameter a of the algorithm influencing the size and position of $\mathcal{D}_1$ over the segment $\overline{wc}$ be such that t_1 is under v_0, in $\overline{oxy}$. This is equivalent to the fact that both t_1 and t_2 are under the x-axis of $\overline{oxy}$. If this requirement were not fulfilled, β (designating p) could be arbitrarily large for some v. As a result, the first component of (30) could be an arbitrarily

large positive number for $\mathcal{F}(\phi) < 0$. The requirement on a is expressed by (see Fig. 14, m is the midpoint of the segment $\overline{wc}$):

$$|\overline{o_1 t_1}| > |\overline{mc}| \Longleftrightarrow r = r(a) = \frac{a^2 - 1}{2a} > 1 \Longleftrightarrow a > 1 + \sqrt{2}.$$

Then for each v belonging to the shorter arc $t_1 t_2$ there exists β' being the upper bound for β. It is obvious that β' reaches its maximum for $v = t_1$, so if $\mathcal{F}(\phi) < 0$, then $\beta(v) \le \beta'(v) \le \beta'(t_1) = \beta_{\max}$ as defined before Theorem 1. Since we assumed in (A3) that $a > \widehat{a} \Leftrightarrow \beta_{\max} < 5$ ($\beta_{\max}$ exists) and $\beta_{\max}(a)$ is a decreasing function of a, the requirement on a stated above is met.

To estimate the expression (30), we also have to consider the inequality (see Figs. 14 and 1, as well as (18)):

$$0 < -\mathcal{F}(\phi) = -2\frac{\|v - c\|^2}{\|w - c\|^2} + 1 < -2\frac{\|v_{\min} - c\|^2}{\|w - c\|^2} + 1 = 1 - 2\frac{(l - r)^2}{2^2} < 1.$$

Now the expression (30) is estimated, taking (32) into account, for v such that $\mathcal{F}(\phi) < 0$:

$$\begin{aligned}
&\frac{\|w - c\|^2}{4}\big(\mathcal{F}(\phi)(L + 2B - 2\rho(1 + \beta)) + L - 2B - 4\rho\big) \\
&= \frac{\|w - c\|^2}{4}\big(- \mathcal{F}(\phi)(2\rho(1 + \beta) - 2B - L) + L - 2B - 4\rho\big) \\
&\le \frac{\|w - c\|^2}{4}\big(1(2\rho(1 + \beta_{\max}) - 2B - L) + L - 2B - 4\rho) \\
&= \frac{\|w - c\|^2}{4}\big(2(\beta_{\max} - 1)\rho - 4B\big). \qquad (33)
\end{aligned}$$

In order to have (33) negative, we must get

$$B > \frac{\rho(\beta_{\max} - 1)}{2} \qquad (34)$$

and this is true due to assumption (A5). Note that the set of B satisfying (A5) is non-empty due to (A3) and (A4).

We have shown so far that if Assumptions (A2)–(A5) are satisfied, then in Case 1:

$$\text{if } \mathcal{F}(\phi) \ge 0 \text{ then } (30) \le \frac{\|w - c\|^2}{4}(L - 2B - 4\rho) < 0,$$

$$\text{if } \mathcal{F}(\phi) < 0 \text{ then } (30 \le \frac{\|w - c\|^2}{4}\big(2(\beta_{\max} - 1)\rho - 4B\big) < 0.$$

Thus

$$\mathcal{E}(v) - \mathcal{E}(c) \le \frac{\|w-c\|^2}{4} \max(L - 2B - 4\rho, 2(\beta_{\max} - 1)\rho - 4B)$$
$$= \overline{\mathcal{W}}\|w-c\|^2 < 0.$$

Case 2. *v belongs to the interior of $\mathcal{D}_1(w,c)$.*

Then, in Step 3 of the algorithm we have

$$v = p = c - \frac{\nabla q(c)}{2\rho}$$

and we estimate

$$\mathcal{E}(v) - \mathcal{E}(c) = q(v) - q(c) + B\big(\|v-c\|^2 - \|c-w\|^2\big)$$
$$\le \nabla q(c)^T(v-c) + \frac{L}{2}\|v-c\|^2 + B\big(\|v-c\|^2 - \|c-w\|^2\big)$$
$$= -\frac{\|\nabla q(c)\|^2}{2\rho} + \left(\frac{L}{2} + B\right)\frac{\|\nabla q(c)\|^2}{4\rho^2} - B\|c-w\|^2$$
$$= \|\nabla q(c)\|^2 \frac{L + 2B - 4\rho}{8\rho^2} - B\|c-w\|^2 < 0$$

because of the condition (A5) forcing $L + 2B - 4\rho < 0$.

Thus in both Cases 1 and 2 there is a negative change in the energy function, and

$$\mathcal{E}(v) - \mathcal{E}(c) \le \|w-c\|^2 \max\big(\overline{\mathcal{W}}, -B\big) = \overline{\mathcal{V}}\|w-c\|^2 < 0,$$

that is

$$\mathcal{E}(c_{i+1}) - \mathcal{E}(c_i) \le -|\overline{\mathcal{V}}|\,\|c_i - c_{i-1}\|^2 < 0.$$

Therefore, $\mathcal{E}(c_i)$ is a strictly decreasing sequence, for $c_i \neq c_{i+1}$, $\mathcal{E}(c_i) = q(c_i) + B\|c_i - c_{i-1}\|^2$ is bounded by $\overline{q}$ (Assumption (A1)), thus $\mathcal{E}(c_i)$ converges to $\overline{\mathcal{E}}$ and we have

$$\lim_{i\to\infty} \|c_i - c_{i-1}\|^2 \le \lim_{i\to\infty} \frac{\mathcal{E}(c_{i+1}) - \mathcal{E}(c_i)}{-|\overline{\mathcal{V}}|} = 0 \Longrightarrow \lim_{i\to\infty} \|c_i - c_{i-1}\| = 0,$$

$$\forall\, N \ \sum_{i=0}^{N} \|c_i - c_{i-1}\|^2 \le \sum_{i=0}^{N} \frac{\mathcal{E}(c_{i+1}) - \mathcal{E}(c_i)}{-|\overline{\mathcal{V}}|}$$
$$= \frac{1}{-|\overline{\mathcal{V}}|}(\mathcal{E}(c_{N+1}) - \mathcal{E}(c_0)) < \frac{\overline{\mathcal{E}} - \mathcal{E}(c_0)}{-|\overline{\mathcal{V}}|},$$

so $\sum_{i=0}^{\infty} \|c_i - c_{i-1}\|^2 < \infty$. The proof is thus completed. ■

Received: 7 July 2000
Revised: 4 December 2000