Non-monotone Inexact Restoration Method for nonlinear programming

Juliano B. Francisco[*], Douglas S. Gonçalves[†], Fermín S. V. Bazán[‡], Lila L. T. Paredes[§]

Department of Mathematics, Federal University of Santa Catarina, Florianópolis, SC, Brazil, 88040–600

October 15, 2018

Abstract

This paper deals with a new variant of the Inexact Restoration Method of Fischer and Friedlander (COAP, 46, pp. 333-346, 2010). We propose an algorithm that replaces the monotone line-search performed in the tangent phase (with regard to the penalty function) by a non-monotone one. Convergence to feasible points satisfying the approximate gradient projection (AGP) condition is proved under mild assumptions. Numerical results on representative test problems validate and show that the proposed approach outperforms the monotone version when a suitable non-monotone parameter is chosen.

Keywords. Inexact Restoration Method · Non-monotone line-search · Nonlinear Programming · Approximate Gradient Projection

AMS Classification. 49M37 · 65K05 · 90C30

1 Introduction

This work is concerned with the nonlinear programming problem,

$$\text{Minimize } f(x), \quad \text{subject to } H(x) = 0 \text{ and } x \in \Omega, \tag{1}$$

wherein $f : \mathcal{R}^n \to \mathcal{R}$ and $H : \mathcal{R}^n \to \mathcal{R}^m$ are continuous differentiable functions and Ω is a convex and compact set. Hereafter we denote the feasible set by Γ, that is,

$$\Gamma = \{x \in \Omega \mid H(x) = 0\},$$

[*]juliano.francisco@ufsc.br
[†]douglas@mtm.ufsc.br
[‡]fermin.bazan@ufsc.br
[§]lilablossom@gmail.com

and we will assume that the nonlinear programming problem (1) satisfies the following assumption:

(H1) ∇f e ∇H (the Jacobian of H) are Lipschitz on Ω, that is, there exists $K > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \le K\|x - y\| \tag{2}$$
$$\|\nabla H(x) - \nabla H(y)\| \le K\|x - y\|, \tag{3}$$

for all $x, y \in \Omega$.

Several numerical schemes can be applied for solving (1). These include the Augmented Lagrangian Method [2], the Sequential Quadratic Programming [23], Reduced-Gradient Algorithms [16], Sequential Restoration Algorithm [20], and the Inexact Restoration Method [3, 7, 17, 18]. In this paper we propose a numerical method for finding feasible points which satisfies the Aproximate Gradient Projection (AGP) condition [19], that means stationary points of (1), since some constraint qualification is required.

Our numerical scheme follows the line of analysis of the Inexact Restoration Method (IRM) [7]. As far as we know, the Inexact Restoration approach was introduced first by Martínez and Pilotta in [18] to solve a general nonlinear programming problem. Such a numerical optimization algorithm does not require feasible iterates and it is suitable especially when strongly nonlinear equations are present. Besides, in the case of well behaved constraints, the approach is promising when the geometry of the feasible set allows a natural (and reasonably cheap) way to restore an infeasible point back to feasibility.

In our approach we improve the tangent phase of the IRM by means of a non-monotone strategy devised in [27]. The motivation for this comes from numerical evidence reported in several papers that non-monotone line search strategies (or even in the context of trust-regions) outperforms the monotone ones in a large class of optimization problems, especially with respect to number of evaluations of the objective function (and thus in CPU time) [5, 9, 10, 11, 24, 25]. Also, we emphasize that the standard version of the Inexact Restoration Method of [7] has been applied successfully in a large class of optimization problems, as seen, for example, in [8, 15, 1, 4]. The present work introduces a new algorithm for problem (1) that combines non-monotone line search strategies and Inexact Restoration schemes. As we will testify, this new version of IRM (now non-monotone) will accept more frequently full tangent steps (and so less back-tracking) and, consequently, the local convergence of the generated sequence can be speed up when a suitable minimization strategy is employed in the tangent phase. Besides, such a non-monotone line-search effectively reduces the overall number of function evaluations when compared with the classical version, as we will display in our numerical experiments. To the best of our knowledge, our approach in the context of Inexact Restoration scheme has not been investigated so far.

Let us define the penalty function $\phi : \Omega \times [0, 1) \to \mathcal{R}$ of (1) by

$$\phi(x, \theta) = \theta f(x) + (1 - \theta)h(x), \tag{4}$$

wherein $h : \mathcal{R}^n \to \mathcal{R}$ is a continuous function such that $h(x) \geq \|H(x)\|$ for all $x \in \Omega$. In summary, from an iterate $x_k \in \Omega$ and a penalty parameter $\theta_k \in (0, 1)$, the IRM accomplishes two phases in order to compute the new iterate, namely, the restoration phase and the minimization (or tangent) phase. In the first phase a point $y_k \in \Omega$ is computed such that $h(y_k) \leq rh(x_k)$ and $f(y_k) \leq f(x_k) + \beta h(x_k)$ (with $r \in [0, 1)$ and $\beta > 0$ being input parameters). Then a penalty parameter $\theta_{k+1} \in (0, 1)$ is adjusted in order to balance $\phi(y_k, \theta_{k+1})$ and $\phi(x_k, \theta_{k+1})$. Next, in the minimization phase, a tangent (or almost tangent) direction d_k is computed and followed by a backtracking scheme where a parameter $t_k \in (0, 1]$ is obtained in such a way that

$$\phi(y_k + t_k d_k, \theta_{k+1}) \leq \phi(x_k, \theta_{k+1}) + \frac{(1 - r)}{2}(h(y_k) - h(x_k)). \tag{5}$$

Lastly, the new iterated is update as $x_{k+1} = x_k + t_k d_k$.

Under the hypothesis of f and H being Lipschitz continuous on Ω (this means Assumption **(H1)** is satisfied), it is proved in [7] that $(h(x_k), h(y_k)) \to (0, 0)$. Besides, if the sufficient decrease condition $f(x_{k+1}) \leq f(y_k) - \alpha t_k \|d_k\|^2$ is asked in the minimization phase, it is assured that $\|d_k\| \to 0$. In this work we replace $\phi(x_k, \theta_{k+1})$ in the right-hand side of (5) by a non-monotone term $\mathcal{T}_k$ based on the non-monotone scheme of [27]. With this slight modification we will prove that this new version of the IRM preserve the same theoretical properties given in [7].

In the following, we will introduce some definitions that will be used throughout the paper.

For all $x \in \Omega, \lambda \in \mathcal{R}^m$, we define the Lagrangian function as

$$L(x, \lambda) = f(x) + H(x)^T \lambda. \tag{6}$$

Thus, a pair $(x^*, \lambda^*) \in \Omega \times \mathcal{R}^m$ is called a Karush-Kuhn-Tucker (or KKT) pair of (1) if $x^* \in \Gamma$ and

$$\text{Proj}_\Omega(x^* - \nabla L(x^*, \lambda^*)) - x^* = 0$$

where $\text{Proj}_\Omega(z)$ denotes the projection of z onto the closed convex set Ω. Here, $\lambda^* \in \mathcal{R}^m$ contains the Lagrange multipliers associated to the stationary point x^*. In addition, for all $y \in \mathcal{R}^n$ we define the *tangent set* and the *tangent subspace* in y respectively by:

$$T(y) = \{z \in \Omega \mid \nabla H(y)(z - y) = 0\}. \tag{7}$$

and

$$S(y) = \{z \in \Omega \mid \nabla H(y)z = 0\}, \tag{8}$$

wherein we recall that $\nabla H(y)$ denotes the Jacobian matrix of H evaluated in y.

This paper is organized as follows. Section 2 describes a non-monotone variant of the Inexact Restoration Method while its convergence properties are established in Section 3. To assess the performance of the non-monotone scheme on several test problems, the outcome of numerical experiments are presented in Section 4. The paper ends with concluding remarks in Section 5.

2 Description of the method

We establish in Algorithm 1 the proposed non-monotone Inexact Restoration Method.

Algorithm 1 - Non-monotone Inexact Restoration Method

Let $r, \eta_{min}, \eta_{max} \in [0,1)$ (with $\eta_{min} \leq \eta_{max}$), $\beta > 0, \gamma > 0, \bar{\gamma} > 0, \tau > 0$, $\theta_0 \in (0,1)$ and $\alpha \in (0,1)$.

Pick $x_0 \in \Omega$ and $\eta_0 \in [\eta_{min}, \eta_{max}]$. Set $C_0 = \phi(x_0, \theta_0)$, $Q_0 = 1$ and $k = 0$.

Step 1. Restoration Phase: Compute $y_k \in \Omega$ such that

$$h(y_k) \leq r h(x_k) \tag{9}$$
$$f(y_k) \leq f(x_k) + \beta h(x_k). \tag{10}$$

Step 2. Penalty Parameter: Compute $\theta_{k+1} = \max\{\theta_k / 2^j \mid j \in \mathbb{N}\}$ such that

$$\phi(y_k, \theta_{k+1}) \leq \phi(x_k, \theta_{k+1}) + \frac{(1-r)}{2}(h(y_k) - h(x_k)). \tag{11}$$

Step 3. Tangent Direction: Compute $d_k \in \mathcal{R}^n$ such that $y_k + d_k \in \Omega$,

$$f(y_k + t d_k) \leq f(y_k) - \gamma t \|d_k\|^2, \tag{12}$$
$$h(y_k + t d_k) \leq h(y_k) + \bar{\gamma} t^2 \|d_k\|^2, \qquad \forall t \in [0, \tau]. \tag{13}$$

Step 4. Minimization Phase: Choose $\eta_k \in [\eta_{min}, \eta_{max}]$, let $\mathcal{T}_k = \max\{C_k, \phi(x_k, \theta_{k+1})\}$ and compute $t_k = \max\{1/2^j \mid j \in \mathbb{N}\}$ such that

$$\phi(y_k + t_k d_k, \theta_{k+1}) \leq \mathcal{T}_k + \frac{(1-r)}{2}(h(y_k) - h(x_k)) \tag{14}$$

Step 5. Define $x_{k+1} = y_k + t_k d_k$, update Q_{k+1} and C_{k+1} as (15) and (16). Set $k = k + 1$ and go back to Step 1.

Let $x_{k+1} \in \Omega$ and $\theta_{k+1} \in (0,1)$ be the current iterated and penalty parameter, respectively, and choose $\eta_k \in [\eta_{\min}, \eta_{\max}] \subseteq [0,1)$. Initializing $C_0 = \Phi(x_0, \theta_0)$ and $Q_0 = 1$, define $\mathcal{T}_k = \max\{C_k, \Phi(x_k, \theta_{k+1})\}$ (Step 4) and the sequences $\{Q_k\}$ and $\{C_k\}$ (Step 5) by the recurrence relation

$$Q_{k+1} = \eta_k Q_k + 1, \tag{15}$$
$$C_{k+1} = (\eta_k Q_k \mathcal{T}_k + \phi(x_{k+1}, \theta_{k+1}))/Q_{k+1}. \tag{16}$$

Note that at Step 3 of Algorithm 1, the tangent direction d_k is assumed to satisfy conditions (12) and (13). We give in the following some remarks concerning these conditions. First, we notice that for all $x, y \in \Omega$, it follows from (2) that

$$|f(x) - f(y) - \langle \nabla f(y), x - y \rangle| \quad \leq \quad \frac{K}{2}\|x - y\|^2. \tag{17}$$

Thus, for all $y, d \in \mathcal{R}^n$ and for all $t > 0$,

$$f(y + td) \leq f(y) + t\langle \nabla f(y), d \rangle + t^2 \frac{K}{2}\|d\|^2.$$

Now, if $\mu \in (0, 1]$ is such that

$$\langle \nabla f(y), d \rangle \leq -\mu\|d\|^2, \tag{18}$$

for all $t \in [0, \min\{1, \mu/K\}]$, we have

$$f(y + td) \leq f(y) + t(-\mu + tK/2)\|d\|^2 \leq f(y) - \frac{\mu t}{2}\|d\|^2,$$

that is, the requirement (12) is fulfilled. Now, on the other hand, if $d \in S(y)$, from (3) we have that

$$\|H(y + td)\| \leq \|H(y)\| + t^2 \frac{K}{2}\|d\|^2. \tag{19}$$

for all $t \in [0, 1]$. Therefore, if $d_k \in S(y_k)$ satisfies condition (18) at Step 3 of Algorithm 1, both requirements (12) and (13) hold true with $\tau = \min\{1, \mu/K\}$, $\gamma = \mu/2$ and $\bar{\gamma} = K/2$. Consequently, although the parameter τ appears in the initialization of the Algorithm 1, it should be chosen sufficiently small or even omitted if a proper d_k direction is chosen at Step 3 (e.g., $d_k = \sigma_k \text{Proj}_{S(y)}(-\nabla f(y))$, for some bounded parameter $\sigma_k > 0$).

Section 3 will show that if Step 1 and Step 3 are accomplished then Steps 2 and 4 (and so Algorithm 1) are well-defined.

3 Convergence Analysis

This section establishes technical results regarding convergence properties of the sequences generated by Algorithm 1. Under an additional requirement at Step 4, we assure that the sequence of tangent directions d_k goes to zero as k grows. In addition, as elucidated at the end of the section, with a proper choice of these directions d_k at Step 3, we establishes global convergence to stationary points of (1).

Next result shows that our proposed scheme is well defined and that both sequences, $\{\theta_k\}_k$ and $\{t_k\}_k$, are bounded away from zero. The proof is similar to that given in [7, Lemma 3].

Theorem 1. *Algorithm 1 is well defined. In addition, there exist $k_0 \in \mathbb{N}$ and $\bar{t} > 0$ such that*

$$\theta_k = \theta_{k_0}, \quad \textit{for all } k \geq k_0,$$
$$t_k \geq \bar{t}, \qquad \textit{for all } k \in \mathbb{N}.$$

Proof: Let x_k be an iterated of Algorithm 1. From (9) and (10) we have

$$\begin{aligned}\phi(y_k,\theta) - \phi(x_k,\theta) &= \theta(f(y_k) - f(x_k)) + (1-\theta)(h(y_k) - h(x_k)) \\ &\leq \theta\beta h(x_k) - (1-\theta)(1-r)h(x_k) \\ &= h(x_k)\big(\theta(\beta+1-r) - (1-r)\big).\end{aligned}$$

Then, if $0 \leq \theta \leq \tilde{\theta} = (1-r)/(2(\beta+1-r))$,

$$\phi(y_k,\theta) - \phi(x_k,\theta) \leq -\frac{1}{2}(1-r)h(x_k) \leq \frac{1}{2}(1-r)(h(y_k) - h(x_k)).$$

Thus, for all $k \in \mathbb{N}$, θ_{k+1} can be chosen as

$$\theta_{k+1} \geq \bar{\theta} = \min\{\theta_0, \tilde{\theta}/2\}.$$

Since $\theta_{k+1} = \max\{\theta_k/2^j\}_{j\in\mathbb{N}}$ (and so non-increasing), there exists $k_0 \in \mathbb{N}$ such that

$$\theta_k = \theta_{k_0} \geq \bar{\theta}, \text{ for all } k \geq k_0. \tag{20}$$

Now we prove that Step 4 of Algorithm 1 is well defined. Since $\mathcal{T}_k \geq \phi(x_k, \theta_{k+1})$,

$$\phi(y_k + td_k, \theta_{k+1}) - \mathcal{T}_k \leq \phi(y_k + td_k, \theta_{k+1}) - \phi(x_k, \theta_{k+1})$$

Therefore, from (11), (12), (13) and (20), for all $t \in (0, \tau]$ (here τ comes from Step 3),

$$\begin{aligned}&\phi(y_k + td_k, \theta_{k+1}) - \phi(x_k, \theta_{k+1}) \\ &\quad = \big(\phi(y_k + td_k, \theta_{k+1}) - \phi(y_k, \theta_{k+1})\big) + \big(\phi(y_k, \theta_{k+1}) - \phi(x_k, \theta_{k+1})\big) \\ &\quad \leq \theta_{k+1}(f(y_k + t_k d_k) - f(y_k)) + (1-\theta_{k+1})\big(h(y_k + td_k) - h(y_k)\big) \\ &\qquad + \frac{1}{2}(1-r)(h(y_k) - h(x_k)) \\ &\quad \leq -\bar{\theta}\gamma t\|d_k\|^2 + \bar{\gamma}t^2\|d_k\|^2 + \frac{1}{2}(1-r)(h(y_k) - h(x_k)) \\ &\quad = t\|d_k\|^2(-\bar{\theta}\gamma + \bar{\gamma}t) + \frac{1}{2}(1-r)(h(y_k) - h(x_k)).\end{aligned}$$

Then, if $0 \leq t \leq \tilde{t} = \min\{\tau, \bar{\theta}\gamma/\bar{\gamma}\}$, we have $-\bar{\theta}\gamma + \bar{\gamma}t \leq 0$ and

$$\phi(y_k + td_k, \theta_{k+1}) - \phi(x_k, \theta_{k+1}) \leq \frac{1}{2}(1-r)(h(y_k) - h(x_k)).$$

Therefore, (14) is well defined and the sequence t_k satisfies

$$t_k \geq \bar{t} = \frac{\tilde{t}}{2}, \text{ for all } k \in \mathbb{N},$$

as stated. □

Next lemma gives positive bounds for sequence $\{Q_k\}_k$.

Lemma 1. *Let $\{Q_k\}$ be the sequence generated by Algorithm 1. Then*

$$1 \leq Q_{k+1} \leq \sum_{j=0}^{k} \eta_{max}^j \leq \frac{1}{1-\eta_{max}}, \quad \textit{for all} \quad k \geq 0.$$

Proof: Since $Q_0 = 1$, $Q_{k+1} = \eta_k Q_k + 1$ and $\eta_k \in [\eta_{min}, \eta_{max}]$,

$$\eta_{min} + 1 \leq Q_1 = \eta_0 Q_0 + 1 \leq \eta_{max} + 1.$$

From induction principle over k, we have that

$$Q_{k+1} = \eta_k Q_k + 1 \leq \eta_{max}(\sum_{j=0}^{k-1} \eta_{max}^j) + 1 = \sum_{j=0}^{k} \eta_{max}^j$$

Therefore, since $\eta_{min}, \eta_{max} \in [0,1)$, the proof follows straightforwardly. □

Next lemma states that sequence $\{\mathcal{T}_k\}$ is monotone nonincreasing and bounded from below.

Lemma 2. *Let $k_0 \in \mathbb{N}$ from Theorem 1. Then, for all $k \geq k_0$, $\phi(x_{k+1}, \theta_{k_0}) \leq \mathcal{T}_{k+1} \leq \mathcal{T}_k$. Further,*

$$\mathcal{T}_{k+1} - \mathcal{T}_k \leq -(1-\eta_{max})\frac{(1-r)^2}{2}h(x_k), \quad \textit{for all } k \geq k_0. \tag{21}$$

Proof: From Theorem 1, we have that $\theta_k = \theta_{k_0}$, for all $k \geq k_0$, and therefore $\mathcal{T}_{k+1} = \max\{C_{k+1}, \phi(x_{k+1}, \theta_{k_0})\}$, for all $k \geq k_0$. Now, it follows from (9), (14), (16) and Lemma 1 that, for all $k \geq k_0$,

$$\begin{aligned} C_{k+1} &= \frac{\eta_k Q_k \mathcal{T}_k + \phi(x_{k+1}, \theta_{k_0})}{Q_{k+1}} \\ &\leq \frac{\eta_k Q_k \mathcal{T}_k + \mathcal{T}_k + ((1-r)/2)(h(y_k) - h(x_k))}{Q_{k+1}} \\ &\leq \frac{(\eta_k Q_k + 1)\mathcal{T}_k - ((1-r)^2/2)h(x_k)}{Q_{k+1}} \\ &= \mathcal{T}_k - \frac{(1-r)^2}{2Q_{k+1}}h(x_k) \leq \mathcal{T}_k - (1-\eta_{max})\frac{(1-r)^2}{2}h(x_k). \end{aligned} \tag{22}$$

Also, for all $k \geq k_0$, from Step 4 of Algorithm 1,

$$\begin{aligned} \phi(x_{k+1}, \theta_{k+2}) &\leq \mathcal{T}_k - \frac{(1-r)^2}{2}h(x_k) \\ &\leq \mathcal{T}_k - (1-\eta_{max})\frac{(1-r)^2}{2}h(x_k). \end{aligned} \tag{23}$$

From (22), (23) and $\mathcal{T}_{k+1}$ definition, for all $k \geq k_0$ it follows that

$$\mathcal{T}_{k+1} \leq \mathcal{T}_k - (1-\eta_{\max})\frac{(1-r)^2}{2}h(x_k).$$

Therefore, for all $k \geq k_0$, we have that $\phi(x_{k+1}, \theta_{k_0}) = \phi(x_{k+1}, \theta_{k+2}) \leq \mathcal{T}_{k+1} \leq \mathcal{T}_k$, from where the result follows. □

Theorem 2. *$\{\mathcal{T}_k\}_{k\in\mathbb{N}}$ is convergent.*

Proof: By continuity of f and h it follows from Lemma 2 that $\{\mathcal{T}_k\}_{k\geq k_0}$ is a monotone nonincreasing sequence bounded from below (since Ω is compact). Then, it is convergent. □

In what follows, we prove that $\lim_{k\to\infty} h(x_k) = \lim_{k\to\infty} h(y_k) = 0$, in other words, all limit points of $\{x_k\}_k$ and $\{y_k\}_k$ are feasible.

Lemma 3. *Let $\{x_k\}_{k\in\mathbb{N}}$ and $\{y_k\}_{k\in\mathbb{N}}$ be the sequences generated by Algorithm 1. Then,*

$$\sum_{k=0}^{\infty} h(x_k) < +\infty$$

and

$$\lim_{k\to\infty} h(x_k) = \lim_{k\to\infty} h(y_k) = 0.$$

Proof: By Theorem 1 and (21), for all $l \geq k_0$,

$$\mathcal{T}_{l+1} - \mathcal{T}_{k_0} = \sum_{k=k_0}^{l} (\mathcal{T}_{k+1} - \mathcal{T}_k) \leq -(1-\eta_{max})\frac{(1-r)^2}{2}\sum_{k=k_0}^{l} h(x_k) < 0.$$

Thus, since $\{\mathcal{T}_k\}_k$ is convergent, from Theorem 2,

$$\sum_{k=k_0}^{\infty} h(x_k) < +\infty$$

and so $\lim_{k\to\infty} h(x_k) = 0$. Now, from (9), we have $\lim_{k\to\infty} h(y_k) = 0$. □

Lemma 4. *Consider $k_0 \in \mathbb{N}$ from Theorem 1. Then, for all $k \geq k_0 + 1$, we have $\mathcal{T}_k = C_k$.*

Proof: Note that, by (16), for all $k \in \mathbb{N}$,

$$\eta_k Q_k (C_{k+1} - \mathcal{T}_k) = \phi(x_{k+1}, \theta_{k+1}) - C_{k+1}. \tag{24}$$

Now, from (22), for all $k \geq k_0$ we have,

$$C_{k+1} - \mathcal{T}_k \leq -(1/2)(1-\eta_{max})(1-r)^2 h(x_k) \leq 0.$$

Hence, for all $k \geq k_0$,

$$\phi(x_{k+1}, \theta_{k+1}) - C_{k+1} = -\frac{\eta_k Q_k}{2}(1-\eta_{max})(1-r)^2 h(x_k) \leq 0.$$

Since $\theta_{k+2} = \theta_{k+1} = \theta_{k_0}$, it results that

$$\phi(x_{k+1}, \theta_{k+2}) \leq C_{k+1}.$$

Thus, $\mathcal{T}_{k+1} = \max\{\phi(x_{k+1}, \theta_{k+2}), C_{k+1}\} = C_{k+1}$, for all $k \geq k_0$. □

From the compactness of set Ω, it follows that the generated sequence $\{y_k\}$ admits at least one limit point. Next theorem is essential for proving that this cluster point satisfies the AGP (Aproximate Gradient Projection) condition [19].

Lemma 5. *Let $d_k \in \mathcal{R}^n$ computed at Step 3 of Algorithm 1. Then, there exists $\widetilde{t} > 0$ such that*

$$\theta_{k+1} f(y_k + td_k) \leq \mathcal{T}_k + \theta_{k+1}\alpha t\langle \nabla f(y_k), d_k\rangle \tag{25}$$

for all $t \in (0, \widetilde{t}]$.

Proof: First, note that

$$\lim_{t\to 0^+} \frac{f(y_k + td_k) - f(y_k)}{t\langle \nabla f(y_k), d_k\rangle} = 1.$$

Since $\alpha \in (0,1)$ and $\langle \nabla f(y_k), d_k\rangle < 0$, there exists $\widetilde{t} > 0$ such that $f(y_k + td_k) \leq f(y_k) + t\alpha\langle \nabla f(y_k), d_k\rangle$, for all $t \in (0, \widetilde{t}]$. So, for $t \geq \widetilde{t}$, from (9) and (11) it follows that

$$\begin{aligned}
\theta_{k+1} f(y_k + td_k) &\leq \theta_{k+1} f(y_k) + t\theta_{k+1}\alpha\langle \nabla f(y_k), d_k\rangle \\
&\leq \phi(y_k, \theta_{k+1}) + t\theta_{k+1}\alpha\langle \nabla f(y_k), d_k\rangle - (1-\theta_{k+1})h(y_k) \\
&\leq \phi(x_k, \theta_{k+1}) + \frac{(1-r)}{2}(h(y_k) - h(x_k)) + t\theta_{k+1}\alpha\langle \nabla f(y_k), d_k\rangle \\
&\leq \mathcal{T}_k + t\theta_{k+1}\alpha\langle \nabla f(y_k), d_k\rangle - \frac{(1-r)^2}{2}h(x_k) \\
&\leq \mathcal{T}_k + t\theta_{k+1}\alpha\langle \nabla f(y_k), d_k\rangle,
\end{aligned}$$

from where the result follows. □

Theorem 3. *Suppose that, at Step 4, t_k is computed such that condition (25) is satisfied at every $k \in \mathbb{N}$. Then, $\lim_{k\to\infty} \|d_k\| = 0$.*

Proof: From Step 3, note that $\langle \nabla f(y_k), d_k\rangle \leq -\gamma\|d_k\|^2$, Then, by Lemma 4 and Theorem 2, $\lim_{k\to\infty} \mathcal{T}_k = \lim_{k\to\infty} C_k$. By using Lemmas 1, 2, 4 and Equation (24), for all $k \geq k_0$ it results that

$$\frac{\eta_{max}}{(1-\eta_{max})}(\mathcal{T}_{k+1} - \mathcal{T}_k) = \frac{\eta_{max}}{(1-\eta_{max})}(C_{k+1} - C_k) \leq \phi(x_{k+1}, \theta_{k+1}) - C_{k+1} \leq 0.$$

Now, since $\{C_k\}_k$ is convergent, we have that

$$\lim_{k\to\infty} \mathcal{T}_k = \lim_{k\to\infty} C_k = \lim_{k\to\infty} \phi(x_k, \theta_k) = \lim_{k\to\infty} \phi(x_k, \theta_{k_0}).$$

Since $\lim_{k\to\infty} h(x_k) = 0$ and , for all $k \geq k_0$, $\phi(x_k, \theta_k) = \phi(x_k, \theta_{k_0}) = \theta_{k_0} f(x_k) + (1 - \theta_{k_0}) h(x_k)$, it follows that $\{\theta_k f(x_k)\}_k$ is convergent and

$$\lim_{k\to\infty} \mathcal{T}_k = \lim_{k\to\infty} \phi(x_k, \theta_k) = \lim_{k\to\infty} \theta_k f(x_k).$$

Let us to prove that $\|d_k\| \to 0$. For this, note from (25) and Theorem 1 that

$$0 = \lim_{k\to\infty} (\theta_{k+1} f(x_{k+1}) - \mathcal{T}_k) \leq -\bar{t}\theta_{k_0}\alpha\gamma \lim_{k\to\infty} \|d_k\|^2 \leq 0,$$

that is, $\lim_{k\to\infty} \|d_k\| = 0$. □

As we have mentioned above, this last theorem has a noteworthy relation with the global convergence of the Algorithm 1. To be clear, if we require at Step 3 of Algorithm 1 that $\|d_k\| \geq \bar{\mu}\|\mathrm{Proj}_{S(y_k)}(-\nabla f(y_k))\|$ for some $\bar{\mu} > 0$, by Theorem 3 it turns out that all accumulation point y_* of $\{y_k\}$ is feasible (by Lemma 3) and fulfils the AGP (Approximate Gradient Projection) optimality condition, that is,

$$\mathrm{Proj}_{S(y_*)}(-\nabla f(y_*)) = 0.$$

In addition, if such a cluster point y_* satisfies the Mangasarian-Fromovitz (MF) constraint qualification, it follows that y_* is a KKT point of (1) (see [19] for further details). Therefore, from the practical point of view, we advocate that a numerical implementation of Algorithm 1 should consider the requirement (25) at Step 4 and, at Step 3, to compute $d_k \in S(y_k)$ so that

$$\langle \nabla f(y_k), d_k \rangle \leq -\mu\|d_k\|^2 \tag{26}$$

and

$$\|d_k\| \geq \bar{\mu}\|\mathrm{Proj}_{S(y_k)}(-\nabla f(y_k))\|, \tag{27}$$

where $\mu \in (0, 1]$ and $\bar{\mu} > 0$. Thus, with these choices we assure that both requirements (12) and (13) are fulfilled and, in addition to, the convergence of a subsequence of $\{y_k\}$ to a KKT point under the MF constraint qualification requirement over the limit point.

We observe that condition (25) might be required at Step 4 of Algorithm 1 to assure that $\|d_k\| \to 0$. In case it is not verified, one guarantees only that $\lim_{k\to\infty} \|h(x_k)\| = \lim_{k\to\infty} \|h(x_k)\| = 0$, according Lemma 3.

4 Numerical Results

In order to illustrate the effectiveness of the proposed non-monotone approach, specially with regard to the significant decrease of the total number of function evaluations, we consider three sets of problems described in what follows. For each set, we tested both the (monotone) Inexact Restoration algorithm by Fischer and Friedlander [7] and our proposed (non-monotone) variant with $\eta_k = \eta$, for all k, wherein $\eta = 0.99, 0.85, 0.5, 0.15, 0$. Although our algorithm with $\eta_k = 0$ is monotone, we emphasize that it is different from the model algorithm given in [7] (due to the term $\mathcal{T}_k$). However, if the penalty parameters sequence

$\{\theta_k\}_k$ stabilizes soon in the earlier iterations, both strategies are numerically quite similar (see Figure 1, for example).

With regard to Step 2 of Algorithm 1, in order to avoid an under estimate of θ_{k_0} (see Theorem 1), for practical purposes it is convenient to restart $\theta_{k+1} \in (\theta_k, \theta_0]$ in some iterations. In our implementation we use $\theta_{k+1} = \min\{\theta_0, 4\theta_k\}$ at each ten iterations. Also, in all instances the restoration phase is performed exactly and so we set $r = 0$ in all cases.

In what follows, we describe each test problem and present the performance profile [6] in a representative set of instances by considering the number of function evaluations as the performance measure.

4.1 Hard-spheres problem

According to [18], and references therein, the hard-spheres problem consists of finding a distribution of n points in the unit sphere in $\mathcal{R}^k$ such that the minimum pairwise distance is maximized.

After a suitable reformulation this problem reads

$$\begin{aligned} \min_{z,w} \quad & z \\ \text{s.t.} \quad & w_i^T w_j - z \leq 0, \quad \forall i \neq j \\ & \|w_i\|^2 = 1, \quad \text{for } i = 1, 2, \ldots, n, \end{aligned} \tag{28}$$

where, for each i, $w_i \in \mathcal{R}^k$.

Let us explain in few lines some details on how we addressed both phases of our algorithm. In the restoration phase, if $x_k = (w_1, \ldots, w_n, z) \in \mathcal{R}^{nk+1}$ is the current iterate, the restored point is given by $y_k = (w'_1, \ldots, w'_n, z')$, where $w'_i = w_i/\|w_i\|$, for $i = 1, 2, \ldots, n$ and $z' = \max_{i \neq j}\{(w'_i)^T w'_j\}$. The search direction in the tangent phase is computed by solving a linear programming obtained from (28) by linearizing the constraints. The reader is referred to [18] for further details on this reformulation.

We create a set of instances by varying the dimension $k \in \{2, 3\}$ and the number of points $n \in \{5, 10, 15, 20, 30\}$. For each pair (n, k), we generate 10 different initial points at random, such that $\|w_i\| = 1$ and $z = 0$. In fact, we regard each of this trials as an instance, so that 100 instances are considered in the performance profile.

Figure 1 shows the performance profile related to this test set. We can observe that the variants of Algorithm 1 with $\eta_k = 0.99$ and $\eta_k = 0.85$ were slightly more efficient than the variant with $\eta_k = 0.5$, whereas the later was slightly more robust. We also remark that our approach with $\eta_k = 0$ and the algorithm proposed in [7] (henceforth called FF) behaved quite similarly with this set of problems.

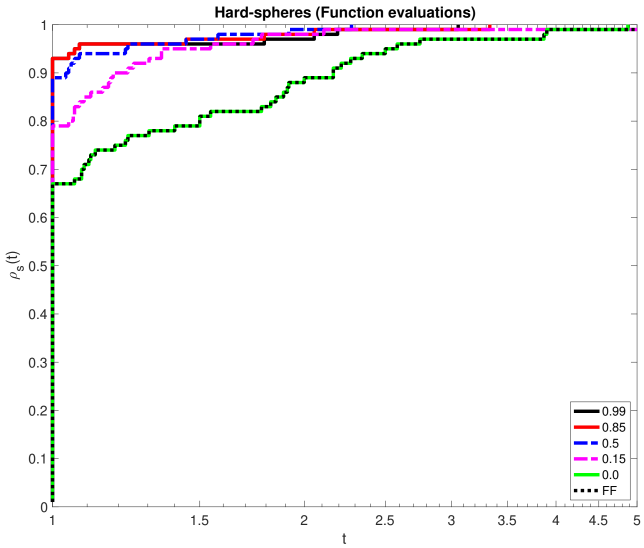

Figure 1: Performance Profile using function evaluations as performance measure. Hard-spheres test set.

4.2 Minimization on Grassmann manifolds

We now consider the problem of minimizing a differentiable functional over a Grassmann manifold, which can be stated as

$$\begin{aligned} \min_{X \in \mathcal{R}^{n\times n}} \quad & f(X) \\ \text{s.t.} \quad & X^2 = X, \;\; \mathrm{tr}(X) = p \quad \text{and} \quad X^T = X. \end{aligned} \tag{29}$$

Such an optimization problem has attracted the interest/attention of researches from different areas, e.g., system identification, geometric computer vision, computational physics (for problems from density functional theory), matrix completion and others [8, 21, 22].

In our implementation, the restored point is computed as follows: given a symmetric matrix $X_{k+1} = Y_k + t_k D_k$ in the tangent set we consider its eigendecompostion $X_{k+1} = Q_{k+1}\Lambda_{k+1}Q_{k+1}^T$. Then, in the restoration phase the feasible point is computed as $Y_{k+1} = \tilde{Q}_{k+1}\tilde{Q}_{k+1}^T$, wherein $\tilde{Q}_{k+1} \in \mathcal{R}^{n\times p}$ contains the columns of Q_{k+1} corresponding to the largest eigenvalues of X_k. It is worth remarking that, with such a choice, $Y_{k+1} = \mathrm{Proj}_\Gamma(X_{k+1})$. The search direction D_k is computed as the projection of a scaled negative gradient onto the tangent subspace at Y_k. See [8] for further details on orthogonal projection onto Grassmann manifold.

We consider $f(X) = (1/2)\mathrm{tr}(LX) + (\alpha/4)\rho(X)^T L^{-1}\rho(X)$ as is usual in density functional theory [14, 8], where L is a discrete Laplacian, $\rho(X) = \mathrm{diag}(X)$

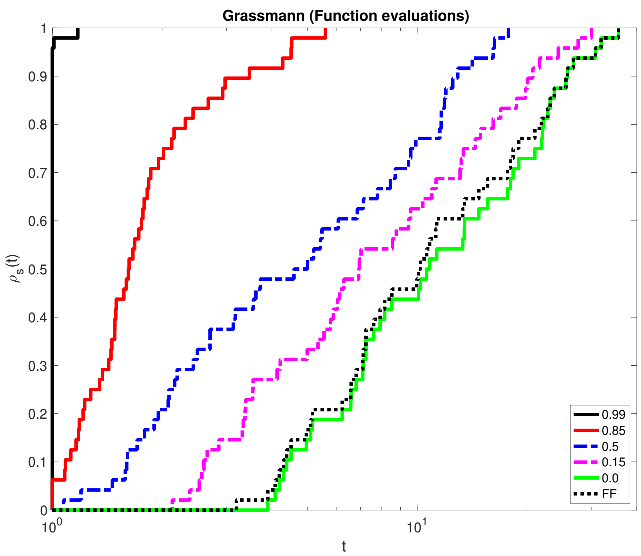

Figure 2: Performance Profile with respect to functions evaluations. Grassmann test set.

and $\alpha > 0$ controls the degree of non-linearity of f. A total of 48 instances were generated by varying $p \in \{5, 10, 15, 20\}$ and $n \in \{50, 100, 200, 300\}$, for each $\alpha = 0.01, 0.1, 1.00$.

The performance profile in Figure 2 indicates that the variant $\eta_k = 0.99$ is the most efficient and robust for this kind of problem.

4.3 Minimization with orthogonality constraints

For the last test set, we consider the problem of minimizing a continuously differentiable functional on the set of (rectangular) orthogonal matrices $X \in \mathcal{R}^{n\times p}$ ($n \geq p$) with orthogonal columns, also called Stiefel manifold. Namely,

$$\begin{aligned} \min_X \quad & f(X) \\ \text{s.t.} \quad & X^T X = I. \end{aligned}$$

In fact, we consider three classes of this problem:

- Linear eigenvalue problem: $f(X) = -\mathrm{tr}(X^T A X)$, where A is a symmetric matrix;
- Nonlinear eigenvalue problem: $f(X) = (1/2)\mathrm{tr}(X^T L X) + (\alpha/4)\rho(X)^T L^{-1}\rho(X)$, where L is a discrete Laplacian, $\alpha > 0$ and $\rho(X) = \mathrm{diag}(XX^T)$;
- Orthogonal Procrustes problem: $f(X) = \|AX - B\|_F^2$, where $A \in \mathcal{R}^{n\times n}$, $B \in \mathcal{R}^{n\times p}$, $n \geq p$.

For geometric properties and computational difficulties concerning this problem the reader is referred to [13, 26].

With regard to the experiments, we ran our algorithm on a set of 48 random problems with the following features:

- Linear eigenvalue problem: 20 instances with $n \in \{500, 1000, 2000, 3000\}$ and $p \in \{10, 50, 100, 200, 300\}$. In these problems $A = B^T B$ where $B \in \mathcal{R}^{n\times n}$ has entries drawn from a standard normal distribution.

- Nonlinear eigenvalue problem: 16 instances with $n \in \{200, 400, 800, 1000\}$, $p \in \{10, 20, 30, 40\}$ and $\alpha = 0.01$.

- Orthogonal Procrustes problem: 12 instances with $n \in \{500, 1000, 2000\}$, $p \in \{10, 20, 50, 100\}$ and $A = U\Sigma V^T$ where U and V are random orthogonal matrices and Σ is a non-negative diagonal matrix. The diagonal entries of Σ are drawn from a uniform distribution in the interval $[10, 12]$. Then, a random $\tilde{Q} \in \mathcal{R}^{n\times p}$ with orthonormal columns is generated and $B = A\tilde{Q}$.

The (exact) restoration phase is implemented by computing the orthogonal projection of X_k onto the feasible set, which requires a thin SVD factorization and can be computed as follows. Let $X_{k+1} = Y_k + t_k D_k = U_{k+1}\Sigma_{k+1}V_{k+1}^T$ be the thin SVD factorization of the point in the tangent set (D_k is the tangent direction). Then the restored point is $Y_{k+1} = \mathrm{Proj}_\Gamma(X_{k+1}) = U_{k+1}V_{k+1}^T$. Again, D_k is the scaled projected gradient onto the tangent subspace at Y_k. See [12] for further details on projections onto this feasible set .

From the performance profile shown in Figure 3, the variant of Algorithm 1 with $\eta_k = 0.99$ is clearly the more efficient and robust in terms of function evaluations. We also notice that the FF scheme is slightly less efficient than Algorithm 1 with $\eta_k = 0$.

From the numerical results reported in the above subsections we see that the proposed non-monotone restoration algorithm performs better than its monotone counterpart, mainly in terms of function evaluations. Hence, we conclude the section by emphasizing that our algorithm may be appropriate for problems in which the evaluation of the objective function is expensive.

5 Conclusion

This paper have proposed a non-monotone Inexact Restoration Method based on the work of Fischer and Friedlander [7]. In brief, we have replaced the (monotone) line-search in the tangent phase by a non-monotone one in the sense of Zhang and Hager [27]. Under the same mild assumptions used in [27], we have proved that such a non-monotone variation still guarantees the main convergence properties of the generated sequences, specifically, convergence to limit points satisfying the AGP (Approximate Gradient Projection) condition and, consequently, convergence to a stationary point of (1) under a proper constraint qualification. Numerical experiments have shown that our proposal is reliable

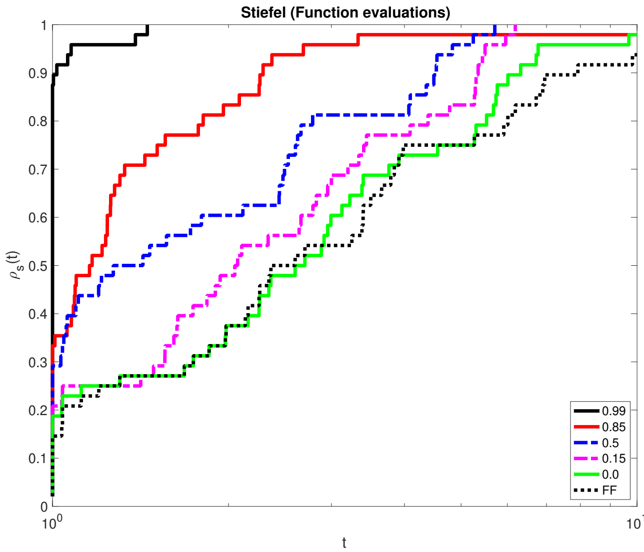

Figure 3: Performance Profile using function evaluations as performance measure. Orthogonality constraints test set.

and particularly efficient in reducing the number of function evaluations when compared to the classical version. We believe that the method becomes attractive when evaluations of the objective function are computationally expensive, as well as, because of allowing inexactness in the restoration phase, when highly nonlinear constraints are present.

Acknowledgements

J.B.F., D.S.G and F.S.V.B are grateful to CNPq by the financial support (Grant n. 421386/2016-9 and 308523/2017-2). L.L.T.P would like to thank to CAPES by the Ph.D. scholarship. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001

References

[1] AROUXÉT, B., ECHEBEST, N.E., PILOTTA, E.A.: Inexact restoration method for nonlinear optimization without derivatives. Journal of Computational and Applied Mathematics **290**(15), 26–43 (2015)

[2] BIRGIN, E., MARTÍNEZ, J.M.: Practical augmented Lagrangian methods for constrained optimization. Fundamental of algorithms. SIAM, Philadelphia (2014)

[3] BIRGIN, E.G., MARTÍNEZ, J.M.: Local convergence of an inexact-restoration method and numerical experiments. Journal of Optimization Theory and Applications **127**, 229 – 247 (2005)

[4] BUENO, L.F., FRIEDLANDER, A., MARTÍNEZ, J.M., SOBRAL, F.N.C.: Inexact restoration method for derivative-free optimization with smooth constraints. SIAM Journal on Optimization **23**(2), 1189–1213 (2013)

[5] DAI, Y.H.: On the nonmonotone line search. Journal of Optimization Theory and Application **112**(2), 315–330 (2002)

[6] DOLAN, E.D., MORÉ, J.J.: Benchmarking optimization software with performance profiles. Math. Prog. **91**, 201–213 (2002)

[7] FISCHER, A., FRIEDLANDER, A.: A new line search inexact restoration approach for nonlinear programming. Comput Optim Appl **46**, 333–346 (2010)

[8] FRANCISCO, J.B., MARTÍNEZ, J.M., MARTÍNEZ, L., PISNITCHENKO, F.I.: Inexact restoration methods for minimization problems that arise in electronic structure calculations. Computational Optimization and Applications **50**, 555–590 (2011)

[9] FRANCISCO, J.B., VILOCHE-BAZÁN, F.S.: Nonmonotone algorithm for minimization on closed sets with application to minimization on stiefel manifolds. Journal of Computational and Applied Mathematics **236**(10), 2717–2727 (2012)

[10] FRANCISCO, J.B., VILOCHE-BAZÁN, F.S., WEBER-MENDONÇA, M.: Non-monotone algorithm for minimization on arbitrary domains with applications to large-scale orthogonal procrustes problem. Applied Numerical Mathematics **112**, 51–64 (2017)

[11] FU, J., SUN, W.: Nonmonotone adaptive trust-region method for unconstrained optimization problems. Applied Mathematics and Computation **163**, 489–504 (2005)

[12] GOLUB, G.A., VAN LOAN, C.F.: Matrix Computations, 3 edn. The John Hopkins University Press Ltda, London (1996)

[13] JIANG, B., DAI, Y.H.: A framework of constraint preserving update schemes for optimization on stiefel manifold. Mathematical Programming **153**(2), 535–575 (2015)

[14] KOHN, W.: Nobel lecture: Electronic structure of matter?wave functions and density functionals. Review of Modern Physics **71**(5), 1253–1266 (1999)

[15] KREJIĆ, N., MARTÍNEZ, J.M.: Inexact restoration approach for minimization with inexact evaluation of the objective function. Mathematics of Computation **85**, 1775–1791 (2016)

[16] LASDON, L.S., FOX, R.L., RATNER, M.W.: Nonlinear optimization using the generalized reduced gradient method. R.A.I.R.O. Operations research **8**(3), 73–103 (1974)

[17] MARTÍNEZ, J.M.: Inexact restoration method with lagrangian tangent decrease and new merit function for nonlinear programming. Journal of Optimization Theory and Applications **111**, 39–58 (2001)

[18] MARTÍNEZ, J.M., PILOTTA, E.A.: Inexact restoration algorithms for constrained optimization. Journal of Optimization Theory and Applications **104**, 135–163 (2000)

[19] MARTÍNEZ, J.M., SVAITER, B.: A practical optimality condition without constraint qualifications for nonlinear programming. Journal of Optmization Theory and Application **118**(1), 117–133 (2003)

[20] MIELE, A., HUANG, H.Y., HEIDEMAN, J.C.: Sequential gradient-restoration algorithm for the minimization of constrained functions - ordinary and conjugate gradient versions. Journal of Optimization Theory and Applications **4**(4), 213–243 (1969)

[21] MITTAL, S., MEER, P.: Conjugate gradient on Grassmann manifolds for robust subspace estimation. Image and Vision Computing **30**(2), 417–427 (2012)

[22] NGO, T., SAAD, Y.: Scaled gradients on Grassmann manifolds for matrix completion. Advances in Neural Information Processing Systems **25**, 1412–1420 (2012)

[23] NOCEDAL, J., WRIGHT, S.J.: Numerical Optimization. Springer Series in Operations Research. Springer Verlag, New York (1999)

[24] TOINT, P.L.: An assessment of non-monotone linesearch techniques for unconstrained optimization. SIAM Journal on Scientific Computing **17**, 725–739 (1996)

[25] TOINT, P.L.: Non-monotone trust region algorithm for nonlinear optimization subject to convex constraints. Mathematical Programming **77**, 69–94 (1997)

[26] WEN, Z., YIN, W.: A feasible method for optimization with orthogonality constraints. Mathematical Programming **142**(1), 397–434 (2013)

[27] ZHANG, H., HAGER, W.: A nonmonotone line search technique and its application to unconstrained optimization. SIAM Journal on Optimization **14**(4), 1043–1056 (2004)