OJMO

Open Journal of Mathematical Optimization

Alberto De Marchi & Andreas Themelis

An interior proximal gradient method for nonconvex optimization

Volume 5 (2024), article no. 3 (22 pages)

https://doi.org/10.5802/ojmo.30

Article submitted on March 21, 2023, revised on January 28, 2024,
accepted on April 28, 2024.

CENTRE MERSENNE

Open Journal of Mathematical Optimization is a member of the
Centre Mersenne for Open Scientific Publishing
http://www.centre-mersenne.org/
e-ISSN: 2777-5860

An interior proximal gradient method for nonconvex optimization

Alberto De Marchi
University of the Bundeswehr Munich
Department of Aerospace Engineering, Institute of Applied Mathematics and Scientific Computing
Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany
alberto.demarchi@unibw.de

Andreas Themelis
Kyushu University
Faculty of Information Science and Electrical Engineering (ISEE)
744 Motooka, Nishi-ku, 819-0395 Fukuoka, Japan
andreas.themelis@ees.kyushu-u.ac.jp

Abstract

We consider structured minimization problems subject to smooth inequality constraints and present a flexible algorithm that combines interior point (IP) and proximal gradient schemes. While traditional IP methods cannot cope with nonsmooth objective functions and proximal algorithms cannot handle complicated constraints, their combined usage is shown to successfully compensate the respective shortcomings. We provide a theoretical characterization of the algorithm and its asymptotic properties, deriving convergence results for fully nonconvex problems, thus bridging the gap with previous works that successfully addressed the convex case. Our interior proximal gradient algorithm benefits from warm starting, generates strictly feasible iterates with decreasing objective value, and returns after finitely many iterations a primal-dual pair approximately satisfying suitable optimality conditions. As a byproduct of our analysis of proximal gradient iterations we demonstrate that a slight refinement of traditional backtracking techniques waives the need for upper bounding the stepsize sequence, as required in existing results for the nonconvex setting.

Digital Object Identifier 10.5802/ojmo.30

Keywords Nonsmooth nonconvex optimization, interior point methods, proximal algorithms, locally Lipschitz gradient.

Acknowledgments A. Themelis acknowledges the support of the Japan Society for the Promotion of Science (JSPS) KAKENHI grants JP21K17710 and JP24K20737.

1 Introduction

We consider structured minimization problems

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad q(x) \coloneqq f(x) + g(x) \qquad \text{subject to} \quad c(x) \leq 0, \tag{P}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $c : \mathbb{R}^n \to \mathbb{R}^m$ are continuously differentiable and $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ has easily computable proximal mapping. The structured objective $q \coloneqq f + g$ is allowed to be nonconvex, as well as each component f and g, and the constraint function c can be nonlinear. When the set induced by $c(x) \leq 0$ is "simple", one may lift the inequality constraints to the objective of (P), enforcing them via an indicator function. But in many cases, projection onto the constraint set $\{x \in \mathbb{R}^n \mid c(x) \leq 0\}$ can be expensive to compute, and even more so when coupled with the proximal mapping of g, motivating us to seek a method able to handle inequalities explicitly.

Starting from polynomial algorithms for linear programming [22, 23], interior point (IP) methods have shaken up the field of mathematical optimization and continue to spark renewed interest; see [17, 20, 46, 47] for a historical overview. It started by solving linear optimization problems with a nonlinear programming technique, based on the use of a barrier function [18] and sequential unconstrained minimization [16]. The remarkable practical success was soon corroborated by deeper understanding of the major role played by the logarithmic barrier function [19, 37], and similar methodologies were applied to solve quadratic and nonlinear optimization problems [2, 3, 12, 43, 44]. However, the focus has almost exclusively been on smooth optimization and gradient-based or Newton-type methods. Some recent exceptions are the works on derivative-free [9] and

Riemannian [24] interior point methods for constrained optimization problems, as well as a closely related proximal gradient-based method [11].

Recalling the basic idea of introducing a barrier function, the reader should observe that the IP rationale is independent of the smoothness of the functions defining the problem. Analogously to penalty and augmented Lagrangian methods [7, §4.1], this feature contributes to the *spirit of unification* that followed the interior point revolution [17]. But as far as we are aware, only a few articles consider IP approaches in the context of nonsmooth optimization problems such as (P).

The combination of IP and splitting methods has been discussed by Valkonen [42] for a class of saddle point problems, associated with structured problems in the form $\min f + g \circ A$, where both f and g are possibly nonsmooth but convex, and A is a bounded linear operator. More closely related to our approach, and associated with (P), is the proximal interior point algorithm (PIPA) presented in [11]. Other works that depart from the classical Newton-type IP approach include [28], which focuses on linear programs, and [49], which addresses convex-constrained variational inequalities involving monotone operators. These works focus, however, on the convex setting and are not directly applicable if any of the problem data functions is nonconvex. Our work aims at filling this gap in the literature by developing and analyzing an interior point method for nonsmooth nonconvex problems. By extending the combination of splitting and IP methods to the fully nonconvex setting, we aim at bringing together and binding areas of optimization that seemed unrelated there.

The constraint smoothening enabled by the adoption of suitably regular barriers in (P) results in IP-type subproblems that seemingly retain a structure that proximal gradient iterations can address, namely the sum of a differentiable and a prox-friendly function. Seemingly, for both components are, in general, extended-real valued: the barrier term smoothens the (indicator of the) feasible set from the interior, thereby shrinking the domain of the differentiable term, as opposed to penalty (or augmented Lagrangian) schemes where the constraints are relaxed and the feasible set enlarged. Although sufficiently small stepsizes can be chosen to make gradient steps remain in the differentiable region, the composition with proximal operations precludes this possibility. Unless different techniques to deal with constraints are proposed, additional structural assumptions to prevent pathological instances are necessary. In the proximal interior point algorithm (PIPA) of [11], convexity is the key.

Dropping these convexity assumptions, this work aims to be a first step toward wider applicability and more versatile modeling. In particular, we show that mere continuity of g *relative to its domain* is sufficient, with no convexity restriction on any term of (P). This is achieved by leveraging an adaptive strategy that enables the use of proximal gradient both in absence of convexity and global Lipschitz differentiability requirements [15, 21]. With a detailed analysis around boundary points, where the barriers escape to infinity, *local* properties are exploited to prove well definedness of the backtracking search. Then, we demonstrate that adaptive proximal gradient steps can generate (strictly) feasible iterates while guaranteeing a descent-type condition at the same time, eventually yielding an approximate KKT-optimal output. When specialized to the case $c = 0$ in (P), yet without g being necessarily continuous relative to its domain, it is shown that through a minor modification of the backtracking strategy no artificial bound on the stepsize sequence is necessary to recover standard convergence results for proximal gradient iterations, cf. Theorem 16. To the best of our knowledge, boundedness of the stepsize sequence is a standing assumption of any existing work dealing with the nonconvex case.

We also point out the usage of non-Euclidean geometries induced by Bregman distances as another proximal-gradient-based alternative to account for ambient constraints [8, 25, 30, 45]. Of this kind, Newton-type extensions also exist that can significantly speed up convergence and even attain superlinear rates, under assumptions at the limit point [1, 5]. All these methods are however subject to (and thus limited in applicability by) the identification of a distance-generating function enabling a so-called Lipschitz-like convexity condition, making induced proximal operations tractable, and whose domain agrees with the constraint set, which must thus be convex. Our focus is instead on addressing problem (P) in the full generality of Assumption 1, stated next.

1.1 Problem setting and proposed methodology

We consider (P) under the following standing assumptions. Technical definitions are given in Section 1.3.

Assumption 1. The following hold in problem (P):

1. $f : \mathbb{R}^n \to \mathbb{R}$ has a locally Lipschitz-continuous gradient.
2. $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is proper, lsc, γ_g-prox-bounded, and continuous relative to $\operatorname{dom} g$.
3. $c : \mathbb{R}^n \to \mathbb{R}^m$ has locally Lipschitz-continuous Jacobian.
4. $\inf \{q(x) \mid c(x) \leq 0\} \in \mathbb{R}$.

5. The problem is strictly feasible: namely, $\operatorname{dom} q \cap \{x \in \mathbb{R}^n \mid c(x) < 0\} \neq \emptyset$.

From a computational point of view, it is assumed that one strictly feasible point can be retrieved explicitly, and that g has an easily computable proximal mapping. Continuity of g relative to its domain is meant in the sense that whenever $\operatorname{dom} g \ni x^k \to x$ it holds that $g(x^k) \to g(x)$. Few exceptions apart, such as functions involving 0-norms, most nonsmooth functions widely used in practice comply with this requirement. For instance, g can be the indicator of any nonempty and closed set, and thus enforce arbitrary closed constraints.

The IP framework builds upon a barrier function $b : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ to replace the *inequality* constraints [16, 18]. We will henceforth fix a nonnegative and smooth barrier function b that complies with the following requirements, assumed throughout.

Assumption 2. The *barrier function* $b : \mathbb{R} \to [0, \infty]$ is such that

1. $\operatorname{dom} b = (-\infty, 0)$.
2. b is twice continuously differentiable with $b' > 0$ on its domain.
3. $b(t) \to \infty$ as $t \to 0^-$.

Equality constraints should be considered carefully and treated e.g. via penalty [12, §4.1.4] or augmented Lagrangian [14] approaches. In the spirit of IP methods [6, 16, 18, 44], we consider a sequence of "unconstrained" barrier problems

$$\operatorname*{minimize}_{z \in \mathbb{R}^n} \quad q_\mu(z) \coloneqq f_\mu(z) + g(z), \tag{P$_\mu$}$$

whose differentiable cost function $f_\mu : \mathbb{R}^n \to \mathbb{R}$ includes the barrier terms weighted by a barrier parameter $\mu > 0$:

$$f_\mu(z) \coloneqq f(z) + \mu \sum_{i=1}^{m} b(c_i(z)). \tag{1}$$

The presence of the possibly nonsmooth term g prevents the employment of traditional IP methods which address the barrier subproblems by means of (smooth) Newton-type techniques. Instead, whenever g has an easily computable proximal mapping, instances of (P$_\mu$) are well suited for solvers based on proximal gradient. This is the rationale originally pursued in [11] and that we here further extend beyond convexity assumptions.

The procedure detailed in Algorithm 1 advances by minimizing the cost function at each iteration and updating the barrier parameter between iterations. At Step 1.1 a point x^{k+1} is retrieved by invoking the proximal gradient method IP-FB, outlined in Algorithm 2, that provides a suitable numerical routine for addressing this task. Its definition requires some preliminary material and the introduction of some notation, and is therefore deferred to Section 3. The iterates $(y^k)_{k \in \mathbb{N}}$ defined by Step 1.2 are solely involved in the termination criterion; as we will show, they relate to the Lagrange multipliers associated with the inequality constraints; cf. Section 2.

Algorithm 1 provides a flexible template of an IP method for inequality constrained problems. It features warm-starting, inexact subsolves, and is subsolver-agnostic, meaning that one can run specialized routines for the problem at hand. In this work we focus on the proximal-gradient-based IP-FB (Algorithm 2), shown to be a suitable candidate for arbitrary formulations as (P) whenever the proximal mapping of g is easily computable.

1.2 Contribution

We present an interior point proximal method (Algorithm 1) for addressing inequality-constrained structured minimization problems. Relying on suitable barrier functions and avoiding the need for slack variables to treat inequalities, our algorithm deviates from those based on penalty-type schemes [14, 40], and always generates feasible iterates while reducing the objective value. Convergence is guaranteed from arbitrary strictly feasible starting points (cf. Theorem 18 and Corollary 19). To our knowledge, this work offers the first (feasible) IP method for addressing problem (P) in the fully nonconvex setting.

As a certified solver for the IP inner subproblems, we propose IP-FB, a proximal gradient method capable of handling barrier problems, whose well definedness is guaranteed through a suitable linesearch (cf. Lemma 13). We establish convergence guarantees in the full generality of problems (P$_\mu$) (cf. Theorem 14 and Corollary 15), coping in particular with the lack of full domain of the smooth function therein. As a byproduct of our analysis, in Theorem 16 we present the first convergence result of proximal gradient iterations with backtracking linesearch in a fully nonconvex regime that does not require any bound on the generated stepsize sequence.

Algorithm 1 Interior point method for (P)
using IP-FB (Algorithm 2, page 8) as inner subsolver

```
REQUIRE     $x^0$                                 strictly feasible starting point (i.e., $x^0 \in \operatorname{dom} g$ with $c(x^0) < 0$)
            $\epsilon_{\rm p}, \epsilon_{\rm d} > 0$             primal-dual tolerances
PROVIDE     $x^\star$                              $(\epsilon_{\rm p}, \epsilon_{\rm d})$-KKT-optimal point for (P) (cf. Definition 6)
INITIALIZE  $\varepsilon_0, \mu_0 > 0$                     initial tolerance and barrier parameters
            $\theta_\varepsilon, \theta_\mu \in (0,1)$                tolerance and barrier update coefficients

REPEAT FOR $k = 0, 1, 2 \dots$
    $x^{k+1} \leftarrow$ IP-FB$(x^k, \mu_k, \varepsilon_k)$          ▷ $\varepsilon_k$-stationary for $q_{\mu_k}$ (see Corollary 15)
    Set $y_i^{k+1} \leftarrow \mu_k b'(c_i(x^{k+1}))$ for all $i$
    IF $\varepsilon_k \leq \epsilon_{\rm d}$ AND $\max_{i=1,\dots,m} \min\{-c_i(x^{k+1}), y_i^{k+1}\} \leq \epsilon_{\rm p}$ THEN
        RETURN $(x^\star, y^\star) \leftarrow (x^{k+1}, y^{k+1})$
    END IF
    Select $0 < \varepsilon_{k+1} \leq \max\{\epsilon_{\rm d}, \theta_\varepsilon \varepsilon_k\}$ and $0 < \mu_{k+1} \leq \theta_\mu \mu_k$
```

1.3 Notation and known facts

With $\mathbb{N}$, $\mathbb{R}$, $\mathbb{R}_+ \coloneqq [0, \infty)$ and $\overline{\mathbb{R}} \coloneqq \mathbb{R} \cup \{\infty\}$ we denote the natural, real, positive real, and extended-real numbers, respectively. Given a point $p \in \mathbb{R}^n$ and a nonempty set $E \subset \mathbb{R}^n$, $\operatorname{dist}(p, E) \coloneqq \inf\{\|x - p\| \mid x \in E\}$ denotes the distance of p from E. The closed ball of radius r centered at p is denoted as $\overline{\mathrm{B}}(p, r) \coloneqq \{x \mid \|x - p\| \leq r\}$. For a sequence $(x^k)_{k\in\mathbb{N}}$ and a set of indices $K \subseteq \mathbb{N}$, $x^k \to_K x$ indicates that the subsequence $(x^k)_{k\in K}$ converges to x.

Let $F : A \to \mathbb{R}^m$ be a function defined on a set $A \subseteq \mathbb{R}^n$, and $\bar{x} \in A$. Following [38, Def. 9.1], we say that F is *locally Lipschitz* (or *strictly*) *continuous at* $\bar{x}$ if $\bar{x} \in \operatorname{int} A$ and the value

$$\operatorname{lip} F(\bar{x}) \coloneqq \limsup_{\substack{x, x' \to \bar{x} \\ x \neq x'}} \frac{\|F(x) - F(x')\|}{\|x - x'\|} \tag{2}$$

is finite; here, $\operatorname{lip} F(\bar{x})$ denotes the Lipschitz constant of F at $\bar{x}$.

The notation $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ indicates a point-to-set operator T that maps each $x \in \mathbb{R}^n$ into a set $T(x) \subseteq \mathbb{R}^n$. The *domain* of T is $\operatorname{dom} T \coloneqq \{x \in \mathbb{R}^n \mid T(x) \neq \emptyset\}$, and we say that T is *outer semicontinuous (osc)* if its *graph* $\operatorname{gph} T \coloneqq \{(x, y) \mid y \in T(x)\}$ is a closed subset of $\mathbb{R}^n \times \mathbb{R}^n$. T is said to be *locally bounded* if for any bounded set $E \subset \mathbb{R}^n$ it holds that $\bigcup_{x\in E} T(x)$ is bounded. For a set-valued mapping, we use the lim sup notation to indicate the *outer limit* [38, Def. 4.1], namely

$$\bar{y} \in \limsup_{x \to \bar{x}} T(x) \quad \overset{\text{(def)}}{\Leftrightarrow} \quad \exists\, (x^k, y^k)_{k\in\mathbb{N}} \subseteq \operatorname{gph} T : \; (x^k, y^k) \to (\bar{x}, \bar{y}).$$

In particular, T is osc if and only if $T(\bar{x}) = \limsup_{x\to\bar{x}} T(x)$ for all $\bar{x} \in \mathbb{R}^n$.

The *effective domain* of an extended-real-valued function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is $\operatorname{dom} h \coloneqq \{x \in \mathbb{R}^n \mid h(x) < \infty\}$. We say that h is *proper* if $\operatorname{dom} h \neq \emptyset$ and *lower semicontinuous* (lsc) if $h(\bar{x}) \leq \liminf_{x\to\bar{x}} h(x)$ for all $\bar{x} \in \mathbb{R}^n$. For some constant $\tau \in \mathbb{R}$, $\operatorname{lev}_{\leq\tau} h \coloneqq \{x \in \mathbb{R}^n \mid h(x) \leq \tau\}$ denotes the τ-*sublevel set* associated with h. Following [38, Def. 8.3] and [35, §1.3], we denote by $\widehat{\partial} h : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ the *regular subdifferential* of h, where

$$\bar{v} \in \widehat{\partial} h(\bar{x}) \quad \overset{\text{(def)}}{\Leftrightarrow} \quad \liminf_{\substack{x \to \bar{x} \\ x \neq \bar{x}}} \frac{h(x) - h(\bar{x}) - \langle \bar{v}, x - \bar{x} \rangle}{\|x - \bar{x}\|} \geq 0. \tag{3}$$

The (*limiting*) *subdifferential* of h is $\partial h : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, where $\bar{v} \in \partial h(\bar{x})$ if and only if $\bar{x} \in \operatorname{dom} h$ and there exist sequences $(x^k)_{k\in\mathbb{N}}$ and $(v^k)_{k\in\mathbb{N}}$ such that $(x^k, v^k, h(x^k)) \to (\bar{x}, \bar{v}, h(\bar{x}))$ and $v^k \in \widehat{\partial} h(x^k)$ for all k. By considering a constant sequence $x^k \equiv \bar{x}$, the inclusion $\widehat{\partial} h(\bar{x}) \subseteq \partial h(\bar{x})$ readily follows. The subdifferential of h at $\bar{x}$ satisfies $\partial(h + h_0)(\bar{x}) = \partial h(\bar{x}) + \nabla h_0(\bar{x})$ for any $h_0 : \mathbb{R}^n \to \overline{\mathbb{R}}$ continuously differentiable around $\bar{x}$ [38, Ex. 8.8].

The *proximal mapping* of h with stepsize $\gamma > 0$ is the set-valued operator $\operatorname{prox}_{\gamma h} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ defined as

$$\operatorname{prox}_{\gamma h}(x) \coloneqq \arg\min_{z\in\mathbb{R}^n} \left\{ h(z) + \tfrac{1}{2\gamma}\|z - x\|^2 \right\}, \tag{4}$$

and we say that h is *prox-bounded* if it is proper and $h + \frac{1}{2\gamma}\|\cdot\|^2$ is bounded below on $\mathbb{R}^n$ for some $\gamma > 0$. The supremum of all such γ is the *threshold* γ_h *of prox-boundedness* for h. In particular, if h is bounded below by an affine function, then $\gamma_h = \infty$. When h is lsc, for any $\gamma \in (0, \gamma_h)$ and $x \in \mathbb{R}^n$ it holds that [38, Thm. 1.25]

$$\emptyset \neq \limsup_{(x',\gamma') \to (x,\gamma)} \operatorname{prox}_{\gamma' h}(x') \subseteq \operatorname{prox}_{\gamma h}(x). \tag{5}$$

2 Stationarity and optimality concepts

Iterative minimization methods typically approach local solutions only asymptotically, while in finitely many iterations can only yield points that satisfy some relaxed, or approximate, optimality conditions. In the case of the minimization of a proper function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$, the inclusion $0 \in \partial h(x^\star)$ (in fact, $0 \in \widehat{\partial} h(x^\star)$) is necessary for local minimality of $x^\star$ for h [38, Thm. 10.1]. An approximate counterpart can be formulated by bounding the distance of the zero vector from the subdifferential. The following definition introduces a terminology tailored for inner problem instances (P$_\mu$).

Definition 3 (ε-stationarity for (P$_\mu$))**.** *Relative to* (P$_\mu$)*, a point* $x^\star$ *is* ε-stationary *for some* $\varepsilon \geq 0$ *if*

$$\operatorname{dist}(0, \partial q_\mu(x^\star)) \leq \varepsilon.$$

When $\varepsilon = 0$*, i.e., when* $0 \in \partial q_\mu(x^\star)$*,* $x^\star$ *is said to be* stationary.[1]

Considering the minimization problem defining the proximal mapping as in (4), the necessary stationarity condition reads

$$\frac{x - \bar{x}}{\gamma} \in \widehat{\partial} h(\bar{x}) \subseteq \partial h(\bar{x}) \qquad \forall\, \bar{x} \in \operatorname{prox}_{\gamma h}(x). \tag{6}$$

Notice that whenever $x^\star$ is an (approximate) stationary point for (P$_\mu$), it necessarily belongs to the domain of q_μ, for otherwise $\partial q_\mu(x^\star)$ would be empty. In particular, $c(x^\star) < 0$, a stronger condition than that prescribed by the constraint in the original problem (P). To emphasize the difference, we will talk in terms of feasibility and *strict* feasibility, as defined next.

Definition 4 (Strict feasibility)**.** *Relative to problem* (P)*, a point* $x^\star \in \operatorname{dom} q$ *is called* feasible *if* $c(x^\star) \leq 0$*, and* strictly feasible *if* $c(x^\star) < 0$.

The given notion of (strict) feasibility imposes the inclusion $x^\star \in \operatorname{dom} q$ so as to also account for implicit constraints encoded in the cost function. Problem (P) can equivalently be expressed as the "unconstrained" minimization of the extended-real-valued function

$$q_0 \coloneqq q + \delta_{\mathbb{R}^m_-} \circ c, \tag{7}$$

where for a set $E \subseteq \mathbb{R}^m$ we denote by $\delta_E : \mathbb{R}^m \to \overline{\mathbb{R}}$ the *indicator function* of E, defined as $\delta_E(x) = 0$ if $x \in E$ and ∞ otherwise. In these terms, feasibility of $x^\star$ can be expressed as the inclusion $x^\star \in \operatorname{dom} q_0$, whereas strict feasibility as the inclusion $x^\star \in \operatorname{dom} q_\mu$ for some (in fact, any) $\mu > 0$. The notion of feasibility is therefore independent of how the problem is formulated, whereas the set of strictly feasible points depends on the specific representation of g and c.

Similarly, in addressing problem (P) one could in principle seek for (approximate) stationary points of q_0. In practice, however, complications may arise in resolving the nonsmooth subdifferential chain rule involved in the evaluation of ∂q_0. For this reason, following the nonlinear programming approach we will consider KKT-type optimality conditions when dealing with (P). These constitute a relaxed stationarity condition, and are in fact equivalent under suitable constraint and epigraphical qualifications.

Definition 5 (KKT optimality for (P))**.** *Relative to* (P)*, a point* $x^\star \in \mathbb{R}^n$ *is* KKT optimal *if it is feasible and there exists* $y^\star \in \mathbb{R}^m_+$ *such that*

$$-\nabla c(x^\star)^\top y^\star \in \partial q(x^\star) \tag{8a}$$

and

$$y_i^\star c_i(x^\star) = 0 \quad \forall\, i = 1, \ldots, m. \tag{8b}$$

[1] The equivalence of $\operatorname{dist}(0, \partial q_\mu(x^\star)) = 0$ and $0 \in \partial q_\mu(x^\star)$ follows from closedness of $\partial q_\mu(x^\star)$, see [38, Thm. 8.6].

Mirroring the concept of ε-stationarity for "unconstrained" minimization problems such as (P_μ), the next definition gives a characterization of approximate KKT optimality for problems subject to (explicit) constraints. This notion allows us to qualify the output of Algorithm 1 in relation to (P); similarly, approximate stationarity will serve as the counterpart for the "unconstrained" inner subproblems (P_μ).

Definition 6 (($\epsilon_{\mathrm{p}}, \epsilon_{\mathrm{d}}$)-KKT optimality for (P))**.** *Relative to* (P)*, a point* $x^\star \in \mathbb{R}^n$ *is said to be* ($\epsilon_{\mathrm{p}}, \epsilon_{\mathrm{d}}$)-KKT optimal *for some* $\epsilon_{\mathrm{p}}, \epsilon_{\mathrm{d}} \geq 0$ *if it is feasible and there exists* $y^\star \in \mathbb{R}^m_+$ *such that*

$$\operatorname{dist}\bigl(-\nabla c(x^\star)^\top y^\star,\, \partial q(x^\star)\bigr) \leq \epsilon_{\mathrm{d}} \tag{9a}$$

and

$$\min\{-c_i(x^\star), y_i^\star\} \leq \epsilon_{\mathrm{p}} \quad \forall\, i = 1, \ldots, m. \tag{9b}$$

Notice that, together with feasibility of $x^\star$ and nonnegativity of $y^\star$, condition (9b) imposes a constraint of approximate complementarity. In general, it is not weaker nor stronger than the more classical condition $|y_i^\star c_i(x^\star)| \leq \epsilon_{\mathrm{p}}$, which could be considered as well.

Similarly to what remarked for approximate stationarity, ($\epsilon_{\mathrm{p}}, \epsilon_{\mathrm{d}}$)-KKT optimality naturally reduces to KKT optimality when $\epsilon_{\mathrm{p}} = \epsilon_{\mathrm{d}} = 0$. There is, however, a substantial difference in the behavior of approximate stationary and approximate KKT-optimal points when the tolerances approach zero in the limit. Suppose that $(z^k)_{k\in\mathbb{N}}$ is an ε_k-stationary point for (P_μ), with $\varepsilon_k \searrow 0$ and $z^k \to z^\star$. Under Assumption 1, we may immediately deduce that $z^\star$ is stationary.[2] On the contrary, having $x^k \to x^\star$ with x^k ($\epsilon_{\mathrm{p},k}, \epsilon_{\mathrm{d},k}$)-KKT optimal for (P) and $\epsilon_{\mathrm{p},k}, \epsilon_{\mathrm{d},k} \searrow 0$ does not guarantee KKT optimality of the limit $x^\star$. This issue raises the need of explicitly defining an asymptotic version of approximate KKT optimality, on the vein of [7, Def. 3.1] and [14, Def. 2.4].

Definition 7 (A-KKT optimality)**.** *Relative to* (P)*, a point* $x^\star \in \mathbb{R}^n$ *is said to be* asymptotically KKT (A-KKT) optimal *if it is feasible and there exist* $(y^k)_{k\in\mathbb{N}} \subset \mathbb{R}^m_+$ *and a feasible sequence* $(x^k)_{k\in\mathbb{N}} \to x^\star$ *such that*

$$\operatorname{dist}\bigl(-\nabla c(x^k)^\top y^k,\, \partial q(x^k)\bigr) \to 0 \tag{10a}$$

and

$$y_i^k c_i(x^\star) = 0 \quad \forall\, i = 1, \ldots, m. \tag{10b}$$

Having $y_i^k c_i(x^\star) = 0$ in condition (10b) causes no loss of generality over $y_i^k c_i(x^\star) \to 0$, a seemingly more natural asymptotic counterpart of (8b). This equivalence will be useful in the sequel, and is formally stated in the following lemma for future reference.

Lemma 8. *Suppose that Assumption 1 holds, and let a feasible sequence* $(x^k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ *converging to a feasible point* $x^\star$ *and a sequence* $(\widetilde{y}^k)_{k\in\mathbb{N}} \subset \mathbb{R}^m_+$ *be such that*

$$\operatorname{dist}\bigl(-\nabla c(x^k)^\top \widetilde{y}^k,\, \partial q(x^k)\bigr) \to 0 \tag{11a}$$

and

$$\widetilde{y}_i^k c_i(x^\star) \to 0 \quad \forall\, i = 1, \ldots, m. \tag{11b}$$

Then, $x^\star$ *is A-KKT optimal.*

Proof. For all $k \in \mathbb{N}$ and $i = 1, \ldots, m$, define $y_i^k = \widetilde{y}_i^k$ if $c_i(x^\star) = 0$ and $y_i^k = 0$ otherwise. Then, observing that $\|\widetilde{y}_i^k - y_i^k\| \to 0$, it is immediate to verify that $(x^k)_{k\in\mathbb{N}}$ and $(y^k)_{k\in\mathbb{N}}$ comply with Definition 7. ◀

It is also worth remarking that usual notions of A-KKT optimality do not require feasibility of the points x^k; nevertheless, in our setting where these points are retrieved through inner IP procedures, feasibility (in fact, *strict*) comes at no cost since it is always inherently satisfied.

While KKT clearly implies A-KKT, the discrepancy between the two notions is again to be found in unmet qualifications, in absence of which local minimizers may fail to be KKT optimal, even for convex problems; A-KKT optimality, on the contrary, is necessary. In referring the reader to the well documented [7, §3] for examples and a thorough discussion, we point out that the feature of A-KKT optimality allowing it to encompass any local solution lies in the possible unboundedness of the sequence $(y^k)_{k\in\mathbb{N}}$ in Definition 7, in absence of which the notion reduces to the nonasymptotic KKT counterpart.

[2] In absence of continuity of g on its domain, the claim still holds true provided that z^k converges q_μ-attentively, namely in such a way that $q_\mu(z^k) \to q_\mu(z^\star)$.

▶ Remark 9. If the sequence $(y^k)_{k\in\mathbb{N}}$ in Definition 7 admits a subsequence $(y^k)_{k\in K}$ converging to a cluster point $y^\star$, as is the case when it is bounded, then the point $x^\star$ therein is KKT optimal, not only asymptotically. This simply follows from the continuity of q on its domain, implying that $\limsup_{k\to\infty}\partial q(x^k)\subseteq\partial q(x^\star)$, and hence that

$$0=\lim_{k\to\infty}\operatorname{dist}\bigl(-\nabla c(x^k)^\top y^k,\,\partial q(x^k)\bigr)\geq\limsup_{K\ni k\to\infty}\operatorname{dist}\bigl(-\nabla c(x^k)^\top y^k,\,\partial q(x^\star)\bigr)=\operatorname{dist}\bigl(-\nabla c(x^\star)^\top y^\star,\,\partial q(x^\star)\bigr)$$

by continuity of ∇c and of the distance function.

3 A barrier-friendly proximal gradient method

In this section we elaborate upon Step 1.1 of Algorithm 1, that aims at solving the barrier problem (P_μ) via proximal gradient iterations. Specifically, we will show that at every (outer) iteration k, the call to IP-FB yields a point x^{k+1} which is ε_k-stationary for problem (P_{μ_k}) and such that $q_{\mu_k}(x^{k+1})\leq q_{\mu_k}(x^k)$, as commented at Step 1.1. IP-FB, outlined in Algorithm 2, is adapted from [15, Alg. 3] so as to cope with the lack of the full domain of the locally smooth function f_μ. In fact, improving upon [13, 15, 21] we here remove boundedness impositions on the stepsize sequence. This flexibility is captured, at the beginning of every iteration j, by initializing the stepsize as $\gamma_j=r\gamma_{j-1}$ (as opposed to $\gamma_j=\gamma_{j-1}$, or selecting γ_j from a fixed bounded interval), where the factor $r\geq 1$ quantifies the stepsize enlargement. Large values of r aim at expediting convergence in terms of number of iterations by testing large stepsizes first, at the expense of potentially more backtrackings and, consequently, gradient evaluations per iteration. Small values instead result in fewer backtrackings at the expense of more conservative stepsize choices. By compensating for the possibly overly cautious estimate obtained by previous reductions, this stepsize redemption has been denominated *"regret"* in the FOM toolbox [4], a terminology that we also adopt in this work. Although the tuning of r may be problem dependent, recent results for the convex case provide insights on parameter-free and problem-independent choices; we refer to the commentary after Theorem 16 for the details.

Relative to (P_μ), we consider the proximal gradient operator with stepsize $\gamma\in(0,\gamma_g)$, with γ_g being the prox-boundedness threshold of g as in Assumption 1.2, defined by

$$\mathrm{T}^{\mathrm{FB}}_{\mu,\gamma}(z):=\operatorname{prox}_{\gamma g}(z-\gamma\nabla f(z)) \tag{12}$$

which is compact valued, and relative to

$$\operatorname{dom}\mathrm{T}^{\mathrm{FB}}_{\mu,\gamma}=\operatorname{dom}f_\mu=\{z\in\mathbb{R}^n\mid c(z)<0\}$$

it is outer semicontinuous (osc) and locally bounded.[3] Notice that, in general, the range of $\mathrm{T}^{\mathrm{FB}}_{\mu,\gamma}$ need not be contained in its domain; as such, fixed-point iterations of $\mathrm{T}^{\mathrm{FB}}_{\mu,\gamma}$ may be ill defined.

Beyond the introduction of the regret factor r, the results and proofs stated in the following closely pattern those presented in [15], where proximal gradient with an adaptively tuned stepsize is shown to work under a mere local Lipschitz differentiability assumption of the smooth term. Although Algorithm 2 is effectively a classical adaptive proximal gradient method, the challenge here is twofold. First, the range of the proximal gradient operator may fail to be contained in its domain, which precludes the possibility of a naïve fixed-point approach. Second, the adaptive strategy considered in [15] revolves around the fact that in any bounded set a finite modulus of Lipschitz continuity of the gradient of the smooth function exists; this property dramatically fails for f_μ in the IP setting here investigated, as its gradient explodes whenever approaching the boundary of the constraint set $\{z\in\mathbb{R}^n\mid c(z)\leq 0\}$. While these issues have been examined and well resolved in [11] for the convex case, no successful attempt appears to have been accomplished in the nonconvex setting.

The key difference with traditional proximal gradient settings is that here, under Assumption 1, the function f_μ as defined in (1) has (locally) Lipschitz-continuous gradient *on its domain*, as opposed to on the entire space. This means that for every *convex and compact* set $\Omega\subset\operatorname{dom}f_\mu$ there exists $L_{f_\mu,\Omega}\geq 0$ such that

$$\begin{cases}\|\nabla f_\mu(z')-\nabla f_\mu(z)\|\leq L_{f_\mu,\Omega}\|z'-z\|\\ f_\mu(z')\leq f_\mu(z)+\langle\nabla f_\mu(z),z'-z\rangle+\frac{L_{f_\mu,\Omega}}{2}\|z'-z\|^2\end{cases}\qquad\forall\,z,z'\in\Omega, \tag{13}$$

[3] Local boundedness relative to $\operatorname{dom}f_\mu$ indicates that for every compact set $Z\subset\operatorname{dom}f_\mu$ the set $\bigcup_{z\in Z}\mathrm{T}^{\mathrm{FB}}_{\mu,\gamma}(z)$ is bounded. Moreover, for any $z\in\operatorname{dom}f_\mu$ and $\gamma\in(0,\gamma_g)$ it follows from (5) that $\emptyset\neq\limsup_{(z',\gamma')\to(z,\gamma)}\mathrm{T}^{\mathrm{FB}}_{\mu,\gamma'}(z')\subseteq\mathrm{T}^{\mathrm{FB}}_{\mu,\gamma}(z)$.

Algorithm 2 IP-FB(z, μ, ε)
Forward Backward solver for Inner Problem (P_μ)

```
REQUIRE     $z$                        strictly feasible starting point (i.e., $z \in \operatorname{dom} g$ with $c(z) < 0$)
            $\mu > 0$                  barrier coefficient
            $\varepsilon > 0$          termination tolerance
PROVIDE     $z^*$                      (strictly feasible) $\varepsilon$-stationary point for $(\mathrm{P}_\mu)$
INITIALIZE  $\gamma_0 \in (0, \gamma_g)$      initial stepsize
            $\alpha, \beta \in (0,1)$  stepsize backtracking parameters
            $r \geq 1$                 stepsize regret factor

Set $z^0 \leftarrow z$ and REPEAT FOR $j = 0, 1, \dots$
2.1: IF $j \geq 1$ THEN $\gamma_j \leftarrow r\gamma_{j-1}$ and $z^j \leftarrow \bar{z}^{j-1}$; END IF   ▷ (or $\gamma_j \leftarrow \min\{r\gamma_{j-1}, \gamma_g - \delta\}$ for some $\delta > 0$ if $\gamma_g \neq \infty$)
2.2: WHILE TRUE DO
2.3:     Compute $\bar{z}^j \in \mathrm{T}^{\mathrm{FB}}_{\mu,\gamma_j}(z^j)$
2.4:     IF { (a) $c(\bar{z}^j) < 0$
              (b) $q_\mu(\bar{z}^j) \leq q_\mu(z^j) - \frac{1-\alpha}{2\gamma_j}\|\bar{z}^j - z^j\|^2$
              (c) $\|\nabla f_\mu(\bar{z}^j) - \nabla f_\mu(z^j)\| \leq \frac{\alpha}{\gamma_j}\|\bar{z}^j - z^j\|$ }  THEN BREAK; ELSE $\gamma_j \leftarrow \beta\gamma_j$; END IF
2.5: END WHILE
2.6: IF $\|\frac{1}{\gamma_j}(z^j - \bar{z}^j) - \nabla f_\mu(z^j) + \nabla f_\mu(\bar{z}^j)\| \leq \varepsilon$ THEN RETURN $z^* \leftarrow \bar{z}^j$ END IF
```

see [38, Thm. 9.2] and [6, Prop. A.24]. In fact, as detailed in the former reference, one can take $L_{f_\mu,\Omega} = \sup_{z \in \Omega} \operatorname{lip} \nabla f_\mu(z)$ in this case. Nevertheless, an elementary compactness argument shows that a finite $L_{f_\mu,\Omega}$ exists for any compact *but not necessarily convex* $\Omega \subset \operatorname{dom} f_\mu$. This observation suggests that, inasmuch as the iterates are confined sufficiently far away from the troublesome boundary of $\{z \mid c(z) \leq 0\}$, issues originating from the lack of full domain of f_μ can be circumvented. A simple proof for the validity of (13) for any compact $\Omega \subset \operatorname{dom} f_\mu$ is detailed for completeness. Note that the interpretation of $L_{f_\mu,\Omega}$ as a Lipschitz constant is ill posed when the set Ω is not convex, and the supremum formula only furnishes a lower bound to $L_{f_\mu,\Omega}$ in this case.

Lemma 10. *Let $\mu > 0$ be fixed. For any compact set $\Omega \subset \operatorname{dom} f_\mu = \{z \mid c(z) < 0\}$ there exists a constant $L_{f_\mu,\Omega} \geq 0$ satisfying* (13)*.*

Proof. Contrary to the claim, suppose that for any $j \in \mathbb{N}$ there exist $z_j, z_j' \in \Omega$ violating either one of the two conditions in (13) with $L_{f_\mu,\Omega} = j$ therein. By compactness of Ω, there exists an infinite index set $J \subseteq \mathbb{N}$ together with $z, z' \in \Omega$ such that $z_j \to z$ and $z_j' \to z'$ as $J \ni j \to \infty$. Since $z, z' \in \Omega \subset \operatorname{dom} f_\mu$ and Ω is compact, necessarily $z = z'$ (for otherwise finiteness of either $f_\mu(z)$, $f_\mu(z')$, $\nabla f_\mu(z)$, or $\nabla f_\mu(z')$ would be violated). As a consequence, up to discarding early terms if necessary openness of $\operatorname{dom} f_\mu$ entails the existence of $\delta > 0$ such that $z_j, z_j' \in \overline{\mathrm{B}}(z, \delta) \subset \operatorname{dom} f_\mu$ holds for all $j \in J$. This is a contradiction, since for any $j \geq L_{f_\mu,\overline{\mathrm{B}}(z,\delta)}$ both conditions hold, where the existence of $L_{f_\mu,\overline{\mathrm{B}}(z,\delta)} \geq 0$ is guaranteed by compactness *and convexity* of $\overline{\mathrm{B}}(z, \delta) \subset \operatorname{dom} f_\mu$. ◀

3.1 Algorithm outline

Although retaining the core features of the adaptive proximal gradient method [15, Alg. 3], see Corollary 4.7 therein, IP-FB includes checks in order to generate iterates that are strictly feasible for $c(z) \leq 0$ and exhibit a sufficient decrease on the cost function. These conditions are enforced at Step 2.4; notice that condition 2.4(a) is implied by condition 2.4(b), and could thus be safely removed without affecting the algorithm. We however prefer to explicitly include the former as well both for clarity and algorithmic convenience: assessing condition 2.4(b) requires evaluating $c(\bar{z}^j)$ in the first place and, if condition 2.4(a) is found to fail, the whole IF statement can already be resolved to be false without further unnecessary function evaluations. Notice further that, since $\bar{z}^j = z^{j+1}$, $\nabla f_\mu(\bar{z}^j)$ evaluated within the j-th iteration can be stored and used in the next one to save computations.

Finite termination of the linesearch occurring at Step 2.4 hinges on the strict feasibility of the previous iterate, which is why the condition must be satisfied in the first place by the initial point z^0 fed in input to Algorithm 2. When called within the IP routine of Algorithm 1 at Step 1.1, this condition is always inherently satisfied, since the initial point x^k prescribed therein is the ouput of a previous call to IP-FB, and is thus strictly feasible by construction. By estimating the local Lipschitz constant of ∇f_μ and monitoring the cost function q_μ, the algorithm is shown to generate iterates $(\bar{z}^j)_{j\in\mathbb{N}}$ that remain bounded away from the barrier at $\{z \in \mathbb{R}^n \mid c(z) = 0\}$. As mentioned in the foreword to Lemma 10, this is the key feature to circumvent the lack of full domain of f_μ.

As will be shown in Corollary 15, the termination criterion at Step 2.6 is satisfied in finitely many iterations and entails ε-stationarity of the output $\bar{z}^j$ for q_μ. The condition is clearly satisfied if $\bar{z}^j = z^j$, in which case $\bar{z}^j$ is stationary, not only approximately so. For this reason, without loss of generality we may avoid trivialities by assuming throughout that $\bar{z}^j \neq z^j$ holds for every j.

3.2 Well definedness

We start by observing that each problem instance (P_μ) is well posed, and also list some important structural properties as placeholders for future reference. The proof of the assertions is a trivial consequence of Assumptions 1 and 2.

Lemma 11. *For any $\mu > 0$, the following hold:*

1. $q_\mu : \mathbb{R}^n \to \overline{\mathbb{R}}$ *is proper, lsc, with* $\operatorname{dom} q_\mu = \operatorname{dom} g \cap \{z \mid c(z) < 0\}$ *and* $\inf q_\mu \in \mathbb{R}$.
2. $f_\mu : \mathbb{R}^n \to \overline{\mathbb{R}}$ *has locally Lipschitz gradient on* $\operatorname{dom} f_\mu = \{z \mid c(z) < 0\}$.

We proceed to show that IP-FB is well defined, namely that each iteration successfully terminates without getting stuck in infinite loops at Step 2.2. Our argument is based on the fact that the proximal mapping converges to the identity as the stepsize tends to zero, a claim that is formalized in the following auxiliary result.

Lemma 12. *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be lsc, and $(z^\ell)_{\ell\in\mathbb{N}} \subset \mathbb{R}^n$ be a sequence converging to a point $z \in \operatorname{dom} h$. Let $\bar{z}^\ell \in \operatorname{prox}_{\gamma_\ell h}(z^\ell)$ with $\gamma_\ell \searrow 0$. Then, $\bar{z}^\ell \to z$.*

Proof. We start by observing that the existence of $\bar{z}^\ell$ guarantees prox-boundedness (hence properness) of h. For every ℓ, the optimality of $\bar{z}^\ell$ in the proximal minimization subproblem reads

$$h(\bar{z}^\ell) + \tfrac{1}{2\gamma_\ell}\|\bar{z}^\ell - z^\ell\|^2 \leq h(z) + \tfrac{1}{2\gamma_\ell}\|z - z^\ell\|^2.$$

By invoking the triangle and Young's inequalities, this implies that

$$\begin{aligned}\|\bar{z}^\ell - z\|^2 &\leq \|\bar{z}^\ell - z^\ell\|^2 + 2\|\bar{z}^\ell - z^\ell\|\|z - z^\ell\| + \|z - z^\ell\|^2\\ &\leq 2\|\bar{z}^\ell - z^\ell\|^2 + 2\|z - z^\ell\|^2\\ &\leq 4\gamma_\ell h(z) + 2\|z - z^\ell\|^2 - 4\gamma_\ell h(\bar{z}^\ell) + 2\|z - z^\ell\|^2\\ &\leq 4\left[\gamma_\ell h(z) - \gamma_\ell h(\bar{z}^\ell) + \|z - z^\ell\|^2\right].\end{aligned}$$

By rearranging, we obtain

$$\gamma_\ell h(\bar{z}^\ell) + \tfrac{1}{4}\|\bar{z}^\ell - z\|^2 \leq \gamma_\ell h(z) + \|z - z^\ell\|^2. \tag{14}$$

The right-hand side vanishes as $\ell \to \infty$; it suffices to show that $(\bar{z}^\ell)_{\ell\in\mathbb{N}}$ remains bounded, so that $\inf_{l\in\mathbb{N}} h(\bar{z}^l) > -\infty$ by properness and lsc of h, as this would imply that each term on the left-hand side too vanishes as $\ell \to \infty$. Contrary to the claim, up to extracting, suppose that $\|\bar{z}^\ell\| \to \infty$. Then, dividing both sides of (14) by $\|\bar{z}^\ell\|^2$ yields

$$\liminf_{\ell\to\infty} \gamma_\ell \tfrac{h(\bar{z}^\ell)}{\|\bar{z}^\ell\|^2} \leq -\tfrac{1}{4}, \quad \text{hence} \quad \liminf_{\ell\to\infty} \tfrac{h(\bar{z}^\ell)}{\|\bar{z}^\ell\|^2} = -\infty.$$

By virtue of [38, Ex. 1.24], this contradicts prox-boundedness of h. ◀

Lemma 13 (Well definedness)**.** *Consider (P_μ) and the iterates generated by Algorithm 2. The following hold:*

1. *At every iteration, the number of backtrackings at Step 2.4 is finite.*
2. *At the j-th iteration ($j \geq 1$), one has $z^j = \bar{z}^{j-1}$ and*

$$q_\mu(z^j) = q_\mu(\bar{z}^{j-1}) \leq q_\mu(z^{j-1}) - \tfrac{1-\alpha}{2\gamma_{j-1}}\|\bar{z}^{j-1} - z^{j-1}\|^2. \tag{15}$$

3. *Every iterate $\bar{z}^j$ remains within* $\operatorname{lev}_{\leq q_\mu^0} q_\mu$, *where* $q_\mu^0 := q_\mu(z^0) < \infty$.

Proof. Let us index by j, ℓ the variables defined at the ℓ-th attempt within the j-th iteration.

1. Let us show that from some strictly feasible z^{j-1}, $j \geq 1$, the iteration terminates (in finite time) yielding a strictly feasible z^j. Terminating an iteration requires to satisfy the conditions at Step 2.4. To arrive to a contradiction, suppose that this never happens, hence that $\gamma_{j,\ell} = \beta^\ell r \gamma_{j-1} \searrow 0$ as $\ell \to \infty$. By openness of $\operatorname{dom} f_\mu \ni z^{j-1}$, there exists $\delta_j > 0$ such that $\Omega_j \coloneqq \overline{\mathrm{B}}(z^{j-1}, \delta_j) \subset \operatorname{dom} f_\mu$. Since $z^{j-1} - \gamma_{j,\ell} \nabla f_\mu(z^{j-1}) \to z^{j-1} \in \operatorname{dom} g$ as $\gamma_{j,\ell} \searrow 0$, Lemma 12 applies and yields the existence of $\ell_j \geq 0$ such that $\bar{z}^{j,\ell} \in \Omega_j$ for all $\ell \geq \ell_j$. On the other hand, by convexity and compactness of $\Omega_j \subset \operatorname{dom} f_\mu$, for any given $\alpha \in (0,1)$ there also exists $\ell'_j \geq 0$ such that $\alpha/\gamma_{j,\ell} \geq L_{f_\mu, \Omega_j}$ for all $\ell \geq \ell'_j$. From Lemma 10 we then conclude that for any $\ell \geq \max\{\ell_j, \ell'_j\}$ both conditions at Step 2.4 are satisfied. In particular, for $\ell \geq \max\{\ell_j, \ell'_j\}$ we have

$$f_\mu(\bar{z}^{j,\ell}) \leq f_\mu(z^{j,\ell}) + \langle \nabla f_\mu(z^{j,\ell}), \bar{z}^{j,\ell} - z^{j,\ell} \rangle + \tfrac{\alpha}{2\gamma_{j,\ell}} \|\bar{z}^{j,\ell} - z^{j,\ell}\|^2.$$

Meanwhile, the minimizing property of $\bar{z}^{j,\ell}$ at Step 2.3 implies

$$g(\bar{z}^{j,\ell}) + \langle \nabla f_\mu(z^{j,\ell}), \bar{z}^{j,\ell} - z^{j,\ell} \rangle + \tfrac{1}{2\gamma_{j,\ell}} \|\bar{z}^{j,\ell} - z^{j,\ell}\|^2 \leq g(z^{j,\ell}).$$

Combining these inequalities, the linesearch condition 2.4(b) is eventually satisfied, whence the contradiction.

2. The assertion follows from the failure of the condition at Step 2.4 and the fact that the value of z^j is not updated after its definition at Step 2.1.

3. Follows from assertion 2, with $q_\mu(z^0) < \infty$ since z^0 is strictly feasible. ◀

3.3 Convergence analysis

The remainder of the section is devoted to showing that for every strictly feasible initial point z and $\mu, \varepsilon > 0$ IP-FB(z, μ, ε) returns an ε-stationary point $z^\star$ for q_μ satisfying $q_\mu(z^\star) \leq q_\mu(z)$. To this end, we will provide an asymptotic analysis where we show that with $\varepsilon = 0$ the algorithm runs indefinitely and produces iterates satisfying $\liminf_{j\to\infty} \|\frac{1}{\gamma_j}(z^j - \bar{z}^j) - \nabla f_\mu(z^j) + \nabla f_\mu(\bar{z}^j)\| = 0$, see Theorem 14.6. The claimed successful finite termination can then be deduced, as will ultimately be formalized in Corollary 15. The entire proof of Theorem 14 will actually be carried out without assuming continuity of g on its domain as required in Assumption 1.2. In allowing the stepsize regret parameter r to be strictly greater than 1, and without imposing any upper bound on the stepsizes γ_j (other than staying bounded away from the prox-boundedness threshold γ_g, should this be finite), this theorem constitutes a refinement of [15, Cor. 4.7] and other related works on proximal gradient algorithms such as [13, 21, 39] which rely on boundedness of $(\gamma_j)_{j\in\mathbb{N}}$.

Theorem 14 (Asymptotic analysis of IP-FB)**.** *The iterates generated by Algorithm 2 with termination tolerance $\varepsilon = 0$ satisfy the following:*

1. $(q_\mu(z^j))_{j\in\mathbb{N}}$ *converges to a finite value* $q_\mu^\star \geq \inf q_\mu$ *from above.*
2. $\sum_{j\in\mathbb{N}} \frac{1}{\gamma_j} \|\bar{z}^j - z^j\|^2 < \infty$.
3. $\sup_{j\in\mathbb{N}} \max\{c_i(\bar{z}^j), c_i(z^j)\} < 0$, *for every* $i = 1, \ldots, m$.
4. *Consider the following assertions:*
  a. q_μ *is level bounded;*
  b. $(\bar{z}^j)_{j\in\mathbb{N}}$ *is bounded;*
  c. $(z^j)_{j\in\mathbb{N}}$ *is bounded;*
  d. $(\gamma_j)_{j\in\mathbb{N}}$ *is bounded away from zero, i.e., there exists* $\gamma_{\min} > 0$ *such that* $\gamma_j \geq \gamma_{\min}$ *for every* j.
  One has $a \Rightarrow b \Leftrightarrow c \Rightarrow d$.
5. $\sum_{j\in\mathbb{N}} \gamma_j = \infty$.
6. $\liminf_{j\to\infty} \frac{1}{\gamma_j} \|\bar{z}^j - z^j\| = \liminf_{j\to\infty} \|\frac{1}{\gamma_j}(z^j - \bar{z}^j) - \nabla f_\mu(z^j) + \nabla f_\mu(\bar{z}^j)\| = 0$.
7. *If the iterates remain bounded, then the set* ω *of accumulation points of* $(\bar{z}^j)_{j\in\mathbb{N}}$ *is made of stationary points for* q_μ*, and* q_μ *is constantly equal to* $q_\mu^\star$ *as in assertion 1 on* ω.

All these claims hold without g being necessarily continuous relative to its domain.

Proof. We begin by observing that (the proofs of) all the claims of Lemmas 11, 12 and 13 that we shall refer to hereafter are indipendent of whether g is continuous on its domain or not.

1. Follows from Lemmas 13.2 and 11.1.

2. Follows from a telescoping argument on (15), having

$$(1-\alpha)\sum_{j\in\mathbb{N}}\tfrac{1}{2\gamma_j}\|\bar z^j - z^j\|^2 \leq q_\mu(z^0) - \inf q_\mu < \infty. \tag{16}$$

3. Let $i \in \{1,\dots,m\}$ be fixed. For every $j \in \mathbb{N}$ we have

$$\inf\{q(z) \mid c(z) \leq 0\} + \mu b(c_i(\bar z^j)) \leq q(\bar z^j) + \mu b(c_i(\bar z^j)) \leq q_\mu(\bar z^j) \leq q_\mu(z^0),$$

where the infimum attains a finite value by Assumption 1.4, since $b \geq 0$, the second inequality too uses nonnegativity of b, and the last one follows from Lemma 13.3. Therefore, the sequence $(b(c_i(\bar z^j)))_{j\in\mathbb{N}}$ remains bounded, which implies that $(c_i(\bar z^j))_{j\in\mathbb{N}}$ is bounded away from 0. In turn, since $z^j = \bar z^{j-1}$ by Lemma 13.2, so is $(c_i(z^j))_{j\in\mathbb{N}}$.

4. The first implication follows from Lemma 13.3, and the second one from Lemma 13.2. Suppose now that $(z^j)_{j\in\mathbb{N}}$ is bounded, and thus that so is $(\bar z^j)_{j\in\mathbb{N}}$. From assertion 3 we then infer the existence of a compact set $\Omega \subset \operatorname{dom} f_\mu$ that contains both sequences. As argued in the proof of Lemma 13.1, any value $\gamma_j \leq {}^{\alpha}/_{L_{f_\mu,\Omega}}$ will pass all conditions at Step 2.4 and will thus not be subject to any backtracking.

5. By iteratively applying the triangle inequality — recall that $z^j = \bar z^{j-1}$, cf. Lemma 13.2 — we obtain

$$\begin{aligned}\|z^j - z^0\| \leq \sum_{\ell=0}^{j-1}\|\bar z^\ell - z^\ell\| &= \sum_{\ell=0}^{j-1}\gamma_\ell^{-1/2}\|\bar z^\ell - z^\ell\|\gamma_\ell^{1/2}\\ &\leq \sqrt{\sum_{\ell=0}^{j-1}\gamma_\ell^{-1}\|\bar z^\ell - z^\ell\|^2}\sqrt{\sum_{\ell=0}^{j-1}\gamma_\ell} \overset{(16)}{\leq} \sqrt{2\frac{q_\mu(z^0)-\inf q_\mu}{1-\alpha}}\sqrt{\sum_{\ell=0}^{j-1}\gamma_\ell}.\end{aligned}$$

Contrary to the claim, if $\sum_{j\in\mathbb{N}}\gamma_j < \infty$ holds, then $(z^j)_{j\in\mathbb{N}}$ is bounded. From assertion 4 we then infer that γ_j is bounded away from zero, thus contradicting the finiteness of $\sum_{j\in\mathbb{N}}\gamma_j$.

6. That $\liminf_{j\to\infty}\frac{1}{\gamma_j}\|\bar z^j - z^j\| = 0$ follows from assertions 2 and 5. In turn, the other limit follows from the fact that $\|\nabla f_\mu(z^j) - \nabla f_\mu(\bar z^j)\| \leq \frac{\alpha}{\gamma_j}\|\bar z^j - z^j\|$, enforced by condition 2.4(c).

7. It follows from assertions 3 and 4 that the iterates z^j and $\bar z^j$ are contained in a compact set $\Omega \subset \operatorname{dom} f_\mu$, and that $\gamma_j \geq \gamma_{\min} > 0$ holds for all j. Let $z^\star \in \omega$ be fixed and let an infinite set of indices $J \subseteq \mathbb{N}$ be such that $\bar z^j \to_J z^\star$. Observe that optimality of $\bar z^j$ in the minimization problem defining $\mathrm{T}^{\text{FB}}_{\mu,\gamma_j}(z^j)$ implies

$$g(\bar z^j) + \tfrac{1}{2\gamma_j}\|\bar z^j - z^j + \gamma_j\nabla f_\mu(z^j)\|^2 \leq g(z^\star) + \tfrac{1}{2\gamma_j}\|z^\star - z^j + \gamma_j\nabla f_\mu(z^j)\|^2,$$

which after expanding the squares and using the fact that $\gamma_j \geq \gamma_{\min} > 0$ gives

$$g(\bar z^j) \leq g(z^\star) + \tfrac{1}{2\gamma_{\min}}\|\overbrace{z^\star - \bar z^j}^{\to_J 0}\|^2 + \langle\overbrace{\nabla f_\mu(z^j)}^{\text{bounded}}, \overbrace{z^\star - \bar z^j}^{\to_J 0}\rangle - \tfrac{1}{2\gamma_j}\|\bar z^j - z^j\|^2.$$

Therefore, $\limsup_{J\ni j\to\infty} g(\bar z^j) \leq g(z^\star)$. Because of lsc, necessarily $g(\bar z^j) \to_J g(z^\star)$, which together with continuity of f_μ on Ω leads to $q_\mu(\bar z^j) \to_J q_\mu(z^\star)$. From the definition of $q^\star_\mu$ in assertion 1 it then follows that $q_\mu(z^\star) = q^\star_\mu$, and the arbitrarity of $z^\star \in \omega$ yields that $q_\mu \equiv q^\star_\mu$ on ω.

To prove stationarity, we consider two cases. If, up to extracting, $\gamma_j \to_J \gamma < \gamma_g \leq \infty$, then the vanishing of $\frac{1}{\gamma_j}\|z^j - \bar z^j\|^2$ implies that

$$z^\star = \lim_{J\ni j\to\infty}\bar z^j \in \limsup_{J\ni j\to\infty}\mathrm{T}^{\text{FB}}_{\mu,\gamma_j}(z^j) \subseteq \mathrm{T}^{\text{FB}}_{\mu,\gamma}(z^\star) \overset{\text{(def)}}{=} \operatorname{prox}_{\gamma g}(z^\star - \gamma\nabla f(z^\star))$$

with the last inclusion owing to outer semicontinuity of $\mathrm{T}^{\text{FB}}_{\mu,\gamma}$ on Ω (cf. footnote 3). The inclusion $z^\star \in \operatorname{prox}_{\gamma g}(z^\star - \gamma\nabla f(z^\star))$ together with (6) yields the claimed stationarity $0 \in \hat\partial q_\mu(z^\star) \subseteq \partial q_\mu(z^\star)$. If, instead, $\gamma_j \to_J \infty$, then since $z^j, \bar z^j$ range in a bounded set, $\|\nabla f_\mu(z^j) - \nabla f_\mu(\bar z^j)\| \leq \frac{\alpha}{\gamma_j}\|\bar z^j - z^j\| \to_J 0$, where the first inequality is enforced at condition 2.4(c). It then follows that $v^j := \frac{1}{\gamma_j}(z^j - \bar z^j) - \nabla f_\mu(z^j) + \nabla f_\mu(\bar z^j) \to_J 0$. Noticing that $v^j \in \nabla f(\bar z^j) + \hat\partial g(\bar z^j) = \hat\partial q_\mu(\bar z^j)$, cf. (6), and recalling that $q_\mu(z^j) \to_J q_\mu(z^\star)$ as shown above, we conclude that $0 \in \partial q_\mu(z^\star)$. ◀

As anticipated, we can now easily infer termination of IP-FB in finitely many steps for any tolerance $\varepsilon > 0$, which confirms that the output of IP-FB complies with the requirements for the outer IP framework of Algorithm 1, as commented in Step 1.1 therein.

Corollary 15 (IP-FB as inner solver for Algorithm 1)**.** *For any strictly feasible starting point z and $\mu, \varepsilon > 0$, in finitely many steps IP-FB(z, μ, ε) returns an ε-stationary point $z^\star$ for (P_μ) satisfying $q_\mu(z^\star) \leq q_\mu(z)$.*

Proof. That the algorithm terminates in finitely many iterates, say j many, follows from Theorem 14.6. Since $\bar{z}^j \in \mathrm{T}^{\mathrm{FB}}_{\mu,\gamma_j}(z^j) = \operatorname{prox}_{\gamma_j g}(z^j - \gamma_j \nabla f(z^j))$, it follows from (6) that the output $z^\star = \bar{z}^j$ satisfies

$$\tfrac{1}{\gamma_j}(z^j - \bar{z}^j) - \nabla f_\mu(z^j) + \nabla f_\mu(\bar{z}^j) \in \widehat{\partial} g(\bar{z}^j) + \nabla f_\mu(\bar{z}^j) = \widehat{\partial} q_\mu(\bar{z}^j) \subseteq \partial q_\mu(\bar{z}^j).$$

The magnitude of such subgradient is no more than ε as enforced by the termination criterion, implying that $\bar{z}^j$ is ε-stationary for q_μ. Finally, that $q_\mu(\bar{z}^j) \leq q_\mu(z)$ follows from Lemma 13.3. ◀

Incidentally, when specialized to the case $c = 0$, Theorem 14 offers insights on plain proximal gradient (PG) iterations that, to the best of our knowledge, are novel. Specifically, it shows that enforcing a Lipschitz-like condition in addition to the standard quadratic upper bound allows one to waive any artificial cap on the stepsize sequence, which is a standing assumption in related literature. The chosen terminology *"unconstrained stepsizes"* emphasizes this distinction.

Theorem 16 (Convergence of PG with unconstrained stepsizes)**.** *Let $\varphi := f + g$ for a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ with locally Lipschitz-continuous gradient and a proper, lsc, and γ_g-prox bounded function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$. Starting from $z^0 \in \mathbb{R}^n$ and $\gamma_0 \in (0, \gamma_g)$, and given some $\alpha, \beta \in (0,1)$ and $r \geq 1$, consider the following scheme:*

```
FOR $j = 1, 2, \ldots$ DO
 1: WHILE TRUE DO
 2:     $z^j \in \operatorname{prox}_{\gamma_j g}(z^{j-1} - \gamma_j \nabla f(z^{j-1}))$
 3:     IF $\varphi(z^j) \leq \varphi(z^{j-1}) - \frac{1-\alpha}{2\gamma_j}\|z^j - z^{j-1}\|^2$ AND $\|\nabla f(z^j) - \nabla f(z^{j-1})\| \leq \frac{\alpha}{\gamma_j}\|z^j - z^{j-1}\|$ THEN
 4:         BREAK
 5:     $\gamma_j \leftarrow \beta\gamma_j$
 6: $\gamma_{j+1} \leftarrow r\gamma_j$                ▷ (or $\gamma_{j+1} \leftarrow \min\{r\gamma_j, \gamma_g - \delta\}$ for some $\delta > 0$ in case $\gamma_g \neq \infty$)
```

Then, $\sum_{j\in\mathbb{N}} \gamma_j = \infty$ and $\liminf_{j\to\infty} \big\| \frac{1}{\gamma_j}(z^{j-1} - z^j) - \big(\nabla f(z^{j-1}) - \nabla f(z^{j-1})\big) \big\| = 0$. If $(z^j)_{j\in\mathbb{N}}$ is bounded (e.g. when φ is level bounded), then its cluster set ω is made of stationary points for φ, $\varphi|_\omega \equiv \lim_{j\to\infty} \varphi(z^j)$, and $\inf_{j\in\mathbb{N}} \gamma_j > 0$.

Proof. We shall see this as a special case of IP-FB with $c = 0$ and $\mu = 0$, resulting in $\operatorname{dom} f_\mu = \operatorname{dom} f = \mathbb{R}^n$ and thus with condition 2.4(a) vacuously satisfied at any backtracking test. If $z^0 \notin \operatorname{dom} g$, then in the first iteration the first condition at Step 3 is also vacuously satisfied for (any candidate) iterate z^1. On the other hand, the second condition is satisfied for γ_j small enough, because of local Lipschitz continuity of ∇f (and the fact that all the iterates z^1 tested in the backtracking remain in a bounded set). Then, for any $j \geq 1$ (regardless of whether $z^0 \in \operatorname{dom} g$ or not) it holds that $z^j \in \operatorname{dom} g$, it being the output of a proximal mapping of g. From iteration $j = 1$ on, then, we may invoke the proof of Theorem 14. ◀

Some comments are in order. The Lipschitz-like condition $\|\nabla f(z^j) - \nabla f(z^{j-1})\| \leq \frac{\alpha}{\gamma_j}\|z^j - z^{j-1}\|$ at Step 3 in the PG scheme synopsized in Theorem 16, this being the refinement that allows for unbounded stepsizes, comes at a price, for every failed assessment incurs a wasted evaluation of $\nabla f(z^j)$.

We also remark that the first condition $\varphi(z^j) \leq \varphi(z^{j-1}) - \frac{1-\alpha}{2\gamma_j}\|z^j - z^{j-1}\|^2$ is implied by the usual local quadratic upper bound $f(z^j) \leq f(z^{j-1}) + \langle \nabla f(z^{j-1}), z^j - z^{j-1} \rangle + \frac{\alpha}{2\gamma_j}\|z^j - z^{j-1}\|^2$, cf. the proof of Lemma 13.2. The validity of Theorem 16 is thus unaffected if within the backtracking the latter inequality is adopted instead, which has the advantage of saving evaluations of g at the expense of a slight additional conservatism.

Notice that the regret factor r, that is, the ratio between the initial stepsize at any iteration and the accepted value at the previous one, is chosen constant for notational convenience and simplicity of exposition, but any sequence $(r_j)_{j\in\mathbb{N}} \subset [1, \infty)$ would be an equally valid option. In other words, the stepsize initialization at Step 6 can be replaced by any $\gamma_{j+1} \geq \gamma_j$ (as long as this choice is bounded away from γ_g, should this threshold be finite). Nevertheless, a small parameter in the range $r \in (1, 2]$ is found to work particularly well in practice, an observation that recent results in the convex setting, advocating an adaptive $r_j = \sqrt{1 + \gamma_{j-1}/\gamma_{j-2}}$, may shed some light upon;

see [32] for the pioneering analysis in the smooth case and the follow-up proximal extensions [26, 27, 33], in particular the discussion surrounding [33, Thm. 1]. This parallel is further emphasized by the Lipschitz-like condition $\gamma_j \frac{\|\nabla f(z^j)-\nabla f(z^{j-1})\|}{\|z^j-z^{j-1}\|} \leq \alpha$, though the stepsize index is shifted in the cited references which allows one to waive any backtrack altogether in the convex case. In the analysis of Theorem 14 and its special case Theorem 16, this Lipschitz-like condition is the key for lifting boundedness requirements on the stepsize sequence.

4 The outer interior point framework

In the nonsmooth setting associated to (P), a proximal gradient algorithm such as IP-FB can be adopted for computing an approximate solution of subproblems in the form of (P_μ), as shown in Section 3. The choice of the first parameter (i.e., the initial point for the inner problem) in the call to IP-FB at Step 1.1 is dictated by the following rationale. Practical performances of both inner and outer procedure may benefit from warm-starting. The similarity between inner problem instances in subsequent iterations, namely instances of (P_μ) solely differing by a slight variation of the parameter μ, suggests that the (approximate) solution x^k of the previous inner problem is an educated choice as initial iterate for the starting point of the current one. Furthermore, being the output of a call to IP-FB, x^k is guaranteed to be strictly feasible (for $k = 0$ this is true by initialization), and its employment as starting point for IP-FB is thus also theoretically supported.

We proceed with a characterization of the iterates generated by Algorithm 1, in terms of objective value, feasibility and stationarity.

Lemma 17 (Algorithmic behavior)**.** *Consider a sequence $(x^k, y^k)_{k\in\mathbb{N}}$ generated by Algorithm 1. For every $k \geq 0$, the following hold:*

1. $q(x^{k+1}) \leq q_{\mu_k}(x^{k+1}) \leq q_{\mu_k}(x^k) \leq q_{\mu_{k-1}}(x^k)$.
2. x^{k+1} *is* ε_k*-stationary for* q_{μ_k}*, and is in particular strictly feasible:* $x^{k+1} \in \operatorname{dom} q$ *and* $c(x^{k+1}) < 0$.
3. $y^{k+1} \geq 0$.
4. $\operatorname{dist}\bigl(-\nabla c(x^{k+1})^\top y^{k+1}, \partial q(x^{k+1})\bigr) \leq \varepsilon_k$.

Proof. We remind that x^{k+1} is the output of IP-FB$(x^k, \mu_k, \varepsilon_k)$, cf. Step 1.1.

1. The second inequality follows from Corollary 15, and the other two from the fact that $b \geq 0$ and $0 \leq \mu_k \leq \mu_{k-1}$.

2. Follows from Corollary 15.

3. Follows from the fact that $b' \geq 0$ and $\mu_k \geq 0$.

4. ε_k-stationarity of x^{k+1} for q_{μ_k} reads $\operatorname{dist}(0, \partial q_{\mu_k}(x^{k+1})) \leq \varepsilon_k$. The claim then follows by observing that

$$\begin{aligned}\partial q_{\mu_k}(x^{k+1}) &= \partial q(x^{k+1}) + \mu_k \sum_{i=1}^m b'(c_i(x^{k+1}))\nabla c_i(x^{k+1}) \\ &= \partial q(x^{k+1}) + \nabla c(x^{k+1})^\top y^{k+1},\end{aligned}$$

where the last identity uses the definition of y^{k+1} at Step 1.2. ◀

We next turn our attention to finite termination and output qualification for Algorithm 1. Similarly to the analysis carried out for the inner IP-FB in the previous section, we will obtain the results as a simple consequence of a more general asymptotic analysis in which the tolerances are driven to zero.

Theorem 18 (Asymptotic analysis of Algorithm 1)**.** *Consider a sequence $(x^k, y^k)_{k\in\mathbb{N}}$ of iterates generated by Algorithm 1. Then,*

1. *If the problem is coercive, in the sense that* q_0 *as in* (7) *is level bounded, then* $(x^k)_{k\in\mathbb{N}}$ *is bounded.*
2. *Any limit point of* $(x^k)_{k\in\mathbb{N}}$ *is feasible.*
3. *If either* $\epsilon_{\mathrm{p}} = 0$ *or* $\epsilon_{\mathrm{d}} = 0$*, then* $\lim_{k\to\infty} \min\bigl\{-c(x^k), y^k\bigr\} = 0$.

If $\epsilon_{\mathrm{d}} = 0$, so that the algorithm runs indefinitely with $\varepsilon_k, \mu_k \to 0$, the following also hold for a subsequence $(x_k)_{k\in K}$ converging to a point $x^\star$:

4. $x^\star$ *is a (feasible) A-KKT-optimal point for* (P).
5. *If* $(y^k)_{k\in K}$ *remains bounded, then* $x^\star$ *is a KKT-optimal point for* (P).

Proof.

1. It follows from Lemma 17.1 that $q(x^k) \leq q_{\mu_0}(x^1) < \infty$ holds for every $k \geq 1$. Since $c(x^k) < 0$ (because $x^k \in \operatorname{dom} q_{\mu_{k-1}}$), one has that $q(x^k) = q_0(x^k)$, hence that for every $k \geq 1$ x^k belongs to the sublevel set $\operatorname{lev}_{\leq q_{\mu_0}(x^1)} q_0$, which is bounded by assumption.

2. That $c(x^\star) \leq 0$ follows from Lemma 17.2 in light of continuity of c. Similarly, since $(q(x^k))_{k\in\mathbb{N}}$ is upper bounded as shown in Lemma 17.1, the inclusion $x^\star \in \operatorname{dom} q$ owes to lsc of q.

3. Among the two possibilities, the algorithm terminates in finite time only if $\epsilon_{\rm p} = 0$ and the returned pair $(x^\star, y^\star)$ satisfies $\min\{-c(x^\star), y^\star\} = 0$. Excluding this ideal situation, we may assume that it runs indefinitely and that consequently $\mu_k \to 0$. By Lemmas 17.2 and 17.3, one has $c(x^k) < 0$ and $y^k \geq 0$ for all $k \in \mathbb{N}$. If for some $\delta > 0$ and $i \in \{1, \dots, m\}$ a subsequence $(x^k)_{k\in K'}$ satisfies $-c_i(x^k) \geq \delta$ for all $k \in K'$, then $(b'(c_i(x^k)))_{k\in K'}$ is bounded and therefore $y_i^k = \mu_{k-1} b'(c_i(x^k)) \to 0$ as $K' \ni k \to \infty$. The claim then follows from the arbitrarity of the subsequence.

4. As shown in assertion 2, $x^\star$ is feasible. Also, Lemmas 17.3 and 17.4 together with the fact that $\varepsilon_k \to 0$ ensure that the sequence $(x^k, y^k)_{k\in K}$ satisfies condition (10a). Condition (10b) follows from assertion 3 together with Lemma 8.

5. Follows from the previous assertion together with Remark 9. ◀

Corollary 19 (Finite termination of Algorithm 1)**.** *For any strictly feasible starting point x^0 and primal-dual tolerance parameters $\epsilon_{\rm p}, \epsilon_{\rm d} > 0$, in finitely many steps Algorithm 1 returns an $(\epsilon_{\rm p}, \epsilon_{\rm d})$-KKT-optimal point $x^\star$ for* (P) *satisfying $q(x^\star) \leq q(x^0)$.*

Notice that the coercivity assumption of q_0 in Theorem 18.1 needed to ensure boundedness of the sequence generated by Algorithm 1 also guarantees that the cost q_{μ_k} in each subproblem is level bounded, which is a trivial consequence of the fact that $q_0 \leq q_\mu$ for any $\mu > 0$. This in particular guarantees that each subproblem (P_μ), for any $\mu > 0$, admits global minimizers. Nevertheless, the successful termination of each call to IP-FB at Step 1.1 is independent of whether or not this assumption is met, as demonstrated in Corollary 15, nor is the termination of Algorithm 1 affected (as long as strictly positive tolerances $\epsilon_{\rm p}, \epsilon_{\rm d}$ are chosen), as commented in the previous corollary.

5 Numerical examples

In this section we present some experimental results on an ill-conditioned toy problem to illustrate the numerical behavior of Algorithms 1 and 2. Then, considering a data analysis task, we investigate the influence of hyperparameters and discuss the performance on larger scale problems.

To graphically summarize our numerical results and compare different solvers, we display *epi-profiles*, *data profiles*, and *(extended) performance profiles*. For $\mathcal{P}$ the set of problems and $\mathcal{S}$ the set of solvers, let $t_{s,p}$ denote the user-defined metric for the computational effort required by solver $s \in \mathcal{S}$ to solve instance $p \in \mathcal{P}$ (lower is better). We will monitor the (total) number of gradient evaluations, so that the computational overhead triggered by backtracking is fairly accounted for.

- *Epi-profiles* display the evaluation metric for individual problems in the problem set $\mathcal{P}$, ordered in such a way that for a user-specified base solver $s \in \mathcal{S}$ the evaluation metric monotonically increases with the problem number. The lowest point in each column corresponds to the best solver on the respective instance.
- *Data profiles* display the cumulative distribution function $f_s\colon [0,\infty) \mapsto [0,1]$ of the evaluation metric, namely
$$f_s(t) := \frac{|\{p \in \mathcal{P} \mid t_{s,p} \leq t\}|}{|\mathcal{P}|}.$$
Each data profile reports the fraction of problems $f_s(t)$ solved by solver s with a budget t of evaluation metric [36], and is therefore independent of the other solvers.
- *Extended performance profiles* address the relative performance of solvers [31, §4.1]. Let $\tau_{s,p}$ denote the (extended) *performance ratio* of solver $s \in \mathcal{S}$ on a certain instance $p \in \mathcal{P}$ in comparison to the best solver, other than s itself, on that same instance. Then, an extended performance profile $\rho_s\colon [0,\infty) \mapsto [0,1]$ is the cumulative distribution function of the performance ratio of solver s, namely
$$\rho_s(\tau) := \frac{|\{p \in \mathcal{P} \mid \tau_{s,p} \leq \tau\}|}{|\mathcal{P}|} \qquad \text{where} \qquad \tau_{s,p} := \frac{t_{s,p}}{\min\{t_{i,p} \mid i \in \mathcal{S}, i \neq s\}}.$$

Thus, an extended performance profile indicates the probability (or fraction of problems) $\rho_s(\tau)$ that a given solver $s \in \mathcal{S}$ is faster or slower than any other solver by a given factor τ.

Implementation details

We describe here details pertinent to the implementation of Algorithms 1 and 2, defining particular choices left equivocal there, such as the initialization and update of algorithmic parameters. These numerical features tend to improve the practical performances, without compromising the convergence guarantees established in previous sections.

- The initial tolerance ε_0 for Algorithm 1 is chosen adaptively, based on the starting point x^0 and barrier parameter μ_0: we set $\varepsilon_0 = \max\{\epsilon_{\mathrm{d}}, \kappa_\varepsilon \eta_0\}$, where $\kappa_\varepsilon \in (0,1)$ is a user-specified parameter and η_0 is the norm evaluated for $j = 0$ at Step 2.6 of Algorithm 2 invoked at (x^0, μ_0).
- We relax the barrier parameter update rule at Step 1.5: we set $\mu_{k+1} \leftarrow \mu_k$ if (x^{k+1}, y^{k+1}) satisfies approximate complementarity, namely $\left\|\min\{-c(x^{k+1}), y^{k+1}\}\right\|_\infty \leq \epsilon_{\mathrm{p}}$, otherwise we reduce the barrier parameter as indicated.
- The initial stepsize $\gamma_0 \in (0, \gamma_g)$ in Algorithm 2 is selected adaptively, based on an estimate L_z of $\operatorname{lip} \nabla f_\mu(z)$. We set $\gamma_0 = \alpha / L_z$, where $L_z := \frac{\|\nabla f_\mu(z^+) - \nabla f_\mu(z)\|}{\|z^+ - z\|}$ is a lower bound on the smoothness constant around z. The point $z^+ := z + h$ is obtained by backtracking, starting from $h = 1$ and reducing h by a factor β until $z^+ \in \operatorname{dom} f_\mu$. This procedure is well defined since $z \in \operatorname{dom} f_\mu$ and c is continuous.[4]
- The algorithmic parameters have been set with the following (default) values: $\kappa_\varepsilon = 10^{-2}$, $\mu_0 = 1$, $\theta_\varepsilon = \theta_\mu = 1/4$ in Algorithm 1, $\alpha = 0.9$, $\beta = 1/2$, $r = 1.1$ in Algorithm 2.
- At Step 1.5 of Algorithm 1 we always select the respective upper bounds, namely we set $\varepsilon_{k+1} \leftarrow \max\{\epsilon_{\mathrm{d}}, \theta_\varepsilon \varepsilon_k\}$ and $\mu_{k+1} \leftarrow \theta_\mu \mu_k$ (or $\mu_{k+1} = \mu_k$ as described above).
- Finally, for constructing the subproblems (P_μ), we consider the barrier function b defined by $b(t) = -1/t$ for $t < 0$, and ∞ otherwise. This choice complies with our requirements for a barrier function, having $b'(t) = 1/t^2 > 0$ for $t < 0$ and $b \geq b(-\infty) = 0$.

To ensure the reproducibility of the numerical results presented in this paper, our implementation adheres to the steps detailed in Algorithms 1 and 2, incorporating the practical mechanisms just delineated, but without introducing any safeguards such as tolerances to mitigate the effects of machine precision. Furthermore, the source code of our implementation has been made available on Zenodo at doi: 10.5281/zenodo.6890044.

5.1 Nonsmooth Rosenbrock with inequalities

As an illustrative toy example, we consider a two-dimensional optimization problem involving a nonsmooth Rosenbrock-like objective function and inequality constraints. Considering the ℓ_p-quasinorm $\|\cdot\|_p$ with $p := 1/2$ and a circle with radius $r_C := 1/2$ centered at $x_C := (-1/4, 1/4)$, it reads

$$\operatorname*{minimize}_{x \in \mathbb{R}^2} \; 100\big(x_2 + 1 - (x_1+1)^2\big)^2 + \|x\|_p^p \qquad \text{subject to} \;\; \|x - x_C\|^2 \geq r_C^2. \tag{17}$$

The proximal mapping of $\|\cdot\|_p^p \colon x \mapsto \sum_{i=1}^2 |x_i|^p$ can be evaluated elementwise based on explicit formulas given in [10, 48], namely

$$\left[\operatorname{prox}_{\gamma\|\cdot\|_{1/2}^{1/2}}(x)\right]_i \ni \begin{cases} \frac{2}{3}\Big(1 + \cos\Big(\frac{2}{3}\arccos\big(-\frac{\gamma}{4}\big(\frac{3}{|x_i|}\big)^{3/2}\big)\Big)\Big) & \text{if } |x_i| > \frac{3}{2}\gamma^{2/3} \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, casting (17) into the form of (P), the problem data functions satisfy the conditions in Assumption 1. In particular, f has locally (and not globally) Lipschitz continuous gradient and g is continuous relative to its domain $\operatorname{dom} g = \mathbb{R}^2$.

We invoked the proposed algorithm on the same problem instance, with $\epsilon_{\mathrm{p}} = \epsilon_{\mathrm{d}} = 10^{-5}$, starting from 20 different (strictly feasible) points $x^0 \in \mathbb{R}^2$. These have been generated as $x^0 = (0, 1/4) + 4/5(\cos\vartheta, \sin\vartheta)$ where $\vartheta \in \mathbb{R}$ is sampled from a uniform grid over $[0, 2\pi]$.

Figures 1 and 2 summarize the outcomes of these simulations. Superimposed to the objective contour lines and the (in)feasible set, the numerical trajectories are depicted in Figure 1, concatenating over $k = 1, 2, \ldots$ the

[4] In case the prox-boundedness threshold γ_g is finite, the value should be then projected onto $[\delta, \gamma_g - \delta]$ for some $\delta > 0$. If $L_z = 0$, the choice of γ_0 can be arbitrary. These minor technicalities are not part of the implementation.

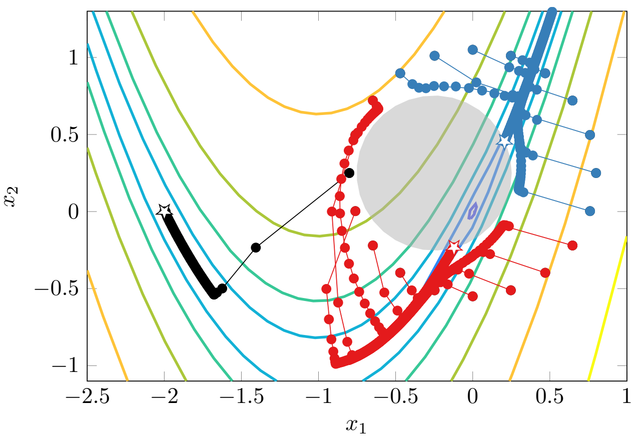

Figure 1 Rosenbrock problem (17): contour lines of the objective function, circular infeasible set (gray), trajectories of inner and outer iterations for different starting points, and limit points thereof (stars). Trajectories are colored based on the limit point: $x^{[1]}$ (red), $x^{[2]}$ (blue) or $x^{[3]}$ (black).

iterates $(x^{k,j})_{j\in\mathbb{N}}$ generated by IP-FB. Depending on the starting point, Algorithm 1 returns one among three stationary points of (P), which are indeed the global minimizer $x^{[1]} \approx (-0.12, -0.23)$ or two local minimizers $x^{[2]} \approx (0.21, 0.45)$ and $x^{[3]} \approx (-2.00, 0)$, see Figure 1. Notice that the feasible set is not simply connected (hence is nonconvex) and that the constraint is active for two minimizers. We stress that the iterates remain strictly feasible while reducing the objective value.

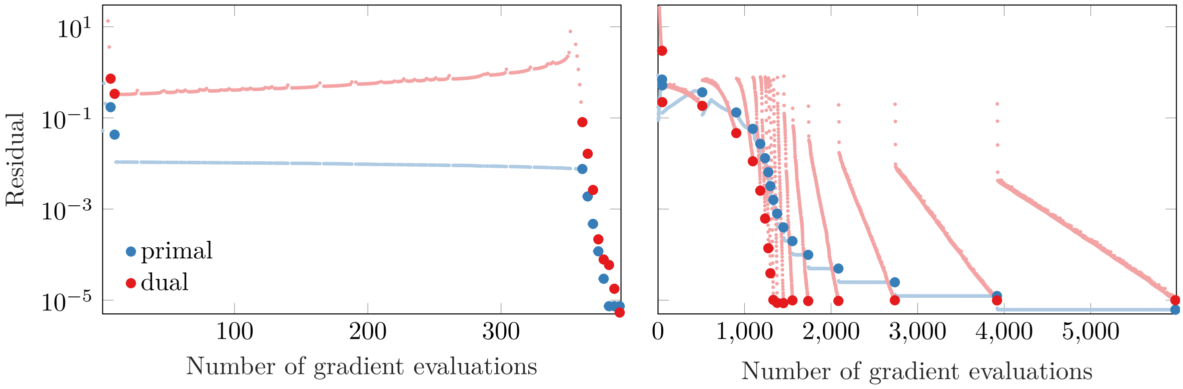

Figure 2 Rosenbrock problem (17): comparison of primal and dual residuals against the number of gradient evaluations, for the starting points $x^0 = (-0.8, 0.25)$ (left) and $x^0 = (0.8, 0.25)$ (right) whose associated limit points are $x^{[3]}$ and $x^{[2]}$, respectively. Larger dots correspond to the outer iterations.

The algorithm performance in terms of optimality and complementarity measures is illustrated in Figure 2 for two different starting points. We monitored the outer dual residual (associated to the inner residual of Step 2.6) and the outer primal residual of Step 1.3 at all iterations. In accordance with Lemma 17.4, the dual residual decreases as dictated by the sequence of inner tolerances $(\varepsilon_k)_{k\in\mathbb{N}}$. It is interesting to notice that, even though Theorem 18.3 only implies the vanishing of the primal residual, in our simulations it is also monotonically decreasing along outer iterations.

5.2 Nonnegative PCA

Principal component analysis (PCA) aims at estimating the direction of maximal variability of a high-dimensional dataset. Arguably the most successful of dimensionality reduction techniques [34], classical PCA aims to recover a signal z from finding the eigenvector that corresponds to the largest eigenvalue of a given matrix Z [29]. A recurring idea is to use additional structural information about the principal eigenvector, such as its signature or sparsity [34]. Here we impose nonnegativity of entries as prior knowledge, and solve PCA restricted to the positive orthant:

$$\operatorname*{maximize}_{x\in\mathbb{R}^n} \; x^\top Z x \qquad \text{subject to } \; \|x\| = 1, \; x \geq 0. \tag{18}$$

This task falls within the scope of (P), with $f(x) \coloneqq -x^\top Zx$, $g(x) \coloneqq \delta_{\|\cdot\|=1}(x)$, and $c(x) = -x$. Nonnegative PCA is an NP-hard nonconvex problem [34] that cannot be addressed by standard singular value decomposition.

Setup

We synthetically generate problem data following [29]. For a problem size $n \in \mathbb{N}$, let $Z = \sqrt{\sigma}zz^\top + N \in \mathbb{R}^{n\times n}$, where $N \in \mathbb{R}^{n\times n}$ is a random symmetric noise matrix and $\sigma > 0$ is the signal-to-noise ratio. The off-diagonal entries of N follow a Gaussian distribution $\mathcal{N}(0, {}^1\!/\!_n)$ and its diagonal entries follow a Gaussian distribution $\mathcal{N}(0, {}^2\!/\!_n)$. Furthermore, we let the support $S \subseteq \{1, \dots, n\}$ of the true principal direction z be uniformly random, with cardinality $|S| = \lfloor sn \rfloor$, and set $z_i = {}^1\!/\!\sqrt{|S|}$ if $i \in S$, $z_i = 0$ otherwise. We consider some dimensions n and, for each dimension, the set of problems parametrized by $\sigma \in \{0.05, 0.1, 0.25, 0.5, 1.0\}$ and $s \in \{0.1, 0.3, 0.7, 0.9\}$, which control the noise and sparsity level, respectively. A strictly feasible starting point x^0 is generated by sampling a uniform distribution over $[0, 3]^n$ and projecting onto $\operatorname{dom} g = \{x \in \mathbb{R}^n \mid \|x\| = 1\}$. There are 5 choices for σ, 4 for s, and, for each set of parameters, 5 instances are generated with different problem data Z and starting point x^0. Overall, each solver-settings pair is invoked on 100 different instances for each dimension n.

Hyperparameters tuning

Algorithms 1 and 2 are controlled by several hyperparameters, such as the initial barrier parameter μ_0, reduction factors $\theta_\mu, \theta_\varepsilon$, and the regret factor r. Investigating the influence of hyperparameters is not only interesting to effectively tune the solvers, but also to appreciate how sensitive (or robust) the performance is with respect to their values.

We now focus on the effect of $\theta_\mu, \theta_\varepsilon \in (0, 1)$, considering problem dimensions $n \in \{10, 15, 20, 25, 30\}$ and all combinations of $\theta_\mu, \theta_\varepsilon \in \{{}^1\!/\!_2, {}^1\!/\!_4, {}^1\!/\!_8\}$, for a total of 4500 calls to Algorithm 1, with tolerances $\epsilon_\text{p} = \epsilon_\text{d} = 10^{-3}$. Lower values of θ_μ (θ_ε) yield a faster decrease of the barrier parameters μ_k (inner tolerances ε_k) toward zero.

All instances are solved up to the desired primal-dual tolerances. The results are graphically summarized in Figures 3 and 4, showing that the majority of selected tunings yield comparable results. The settings $(\theta_\mu, \theta_\varepsilon) = ({}^1\!/\!_8, {}^1\!/\!_4)$, $({}^1\!/\!_4, {}^1\!/\!_2)$, and $({}^1\!/\!_8, {}^1\!/\!_2)$ are increasingly worse, whereas $(\theta_\mu, \theta_\varepsilon) = ({}^1\!/\!_4, {}^1\!/\!_4)$ seems to dominate. This value agrees with the default settings chosen for the solver, as mentioned in the beginning of this section.

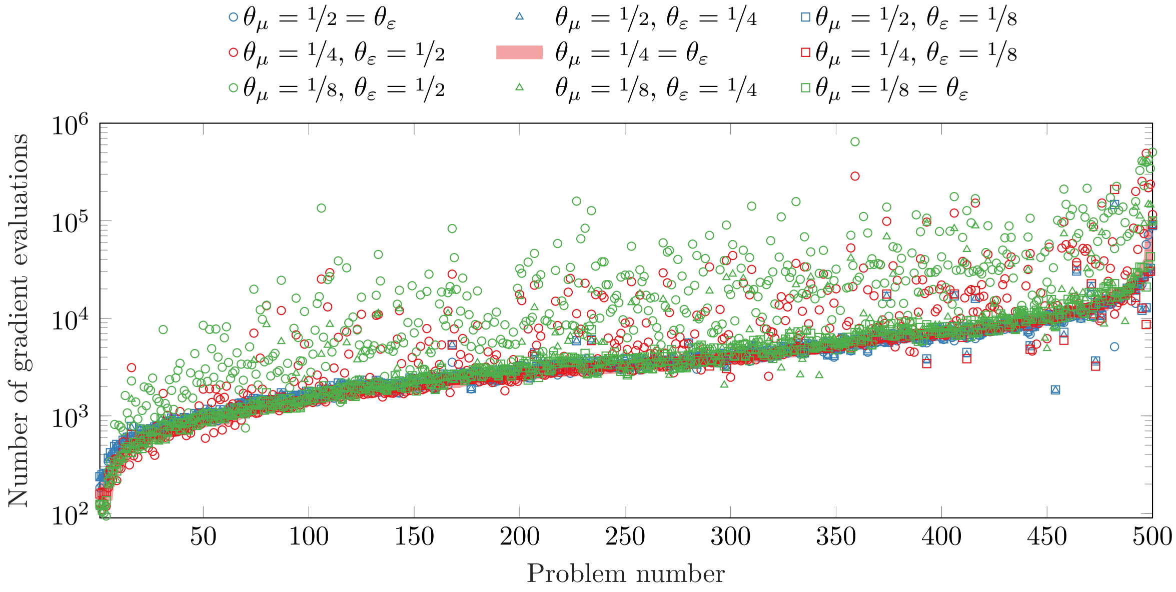

Figure 3 Nonnegative PCA problem (18): comparison for different barrier parameter and inner tolerance reduction factors $\theta_\mu, \theta_\varepsilon \in (0, 1)$. Epi-profiles ordered in relation to the default values $\theta_\mu = {}^1\!/\!_4 = \theta_\varepsilon$ (red thick line).

Let us now examine the influence of the regret factor $r \geq 1$ in Algorithm 2, considering the same problem setup and the values $r \in \{1, 1.1, 1.25, 1.5\}$, for a total of 2000 calls to Algorithm 1, including the adaptive variant described below. As commented in the beginning of Section 3, higher values of r allow the stepsize to recover faster from low values that compromise convergence speed, when the local geometry of f allows. On

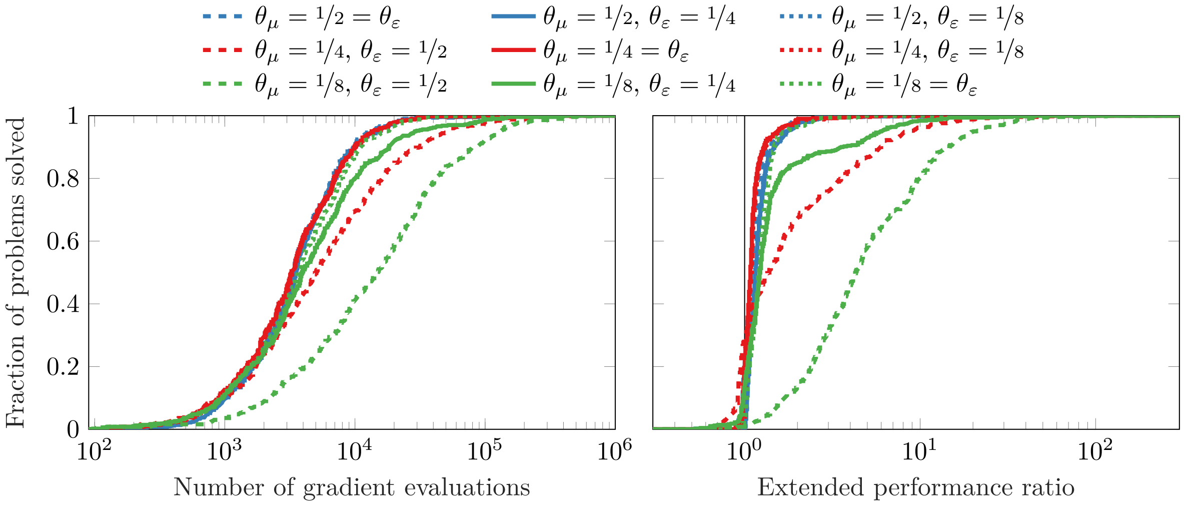

Figure 4 Nonnegative PCA problem (18): comparison for different barrier parameter and inner tolerance reduction factors $\theta_\mu, \theta_\varepsilon \in (0, 1)$. Data profiles (left) and extended performance profiles (right) relative to the number of gradient evaluations.

the other hand, lower values of r reduce the number of backtrackings at every step, and thus the number of gradient evaluations per iteration. However, other than keeping r constant, it is possible to consider any sequence $(r_j)_{j\in\mathbb{N}} \subset [1, \infty)$, as mentioned in the discussion after Theorem 16. This motivates testing also Algorithm 2 with an adaptive regret: on the line of [32], we consider the sequence generated by $r_j = \sqrt{1 + \gamma_{j-1}/\gamma_{j-2}}$ for all $j \geq 2$, with the initialization $\gamma_j = r_j\gamma_{j-1}$ at Step 2.1.

All instances are solved up to the desired primal-dual tolerances and computational results are graphically summarized in Figures 5 and 6. According to these profiles, a suitable tuning for the regret factor in Algorithm 2 appears to be around the value $r = 1.1$, in agreement with the default settings chosen for the solver. These results illustrate the significant potential benefits of a regret factor $r > 1$, as revealed by the considerable gap with the monotone stepsize initialization ($r = 1$). Furthermore, all the tested values of $r > 1$ yield consistent improvements over the choice $r = 1$, indicating that the good performance of Algorithm 2 may be robust with respect to the regret factor for values of r strictly larger than (but close to) 1. This is also true for the adaptive choice, suggesting that it could also constitute a conveniently parameter-free strategy.

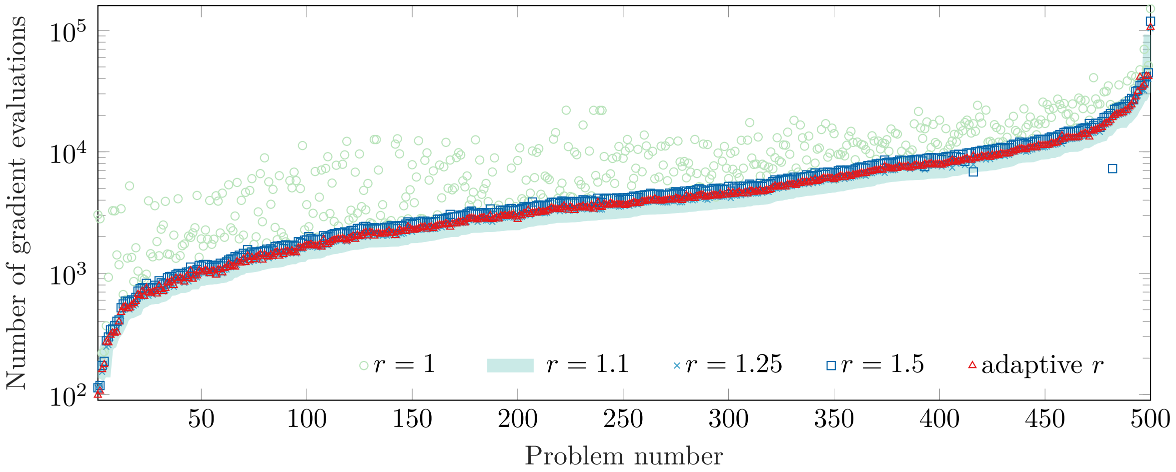

Figure 5 Nonnegative PCA problem (18): comparison for different regret factors $r \geq 1$. Epi-profiles ordered in relation to the default value $r = 1.1$ (thick line).

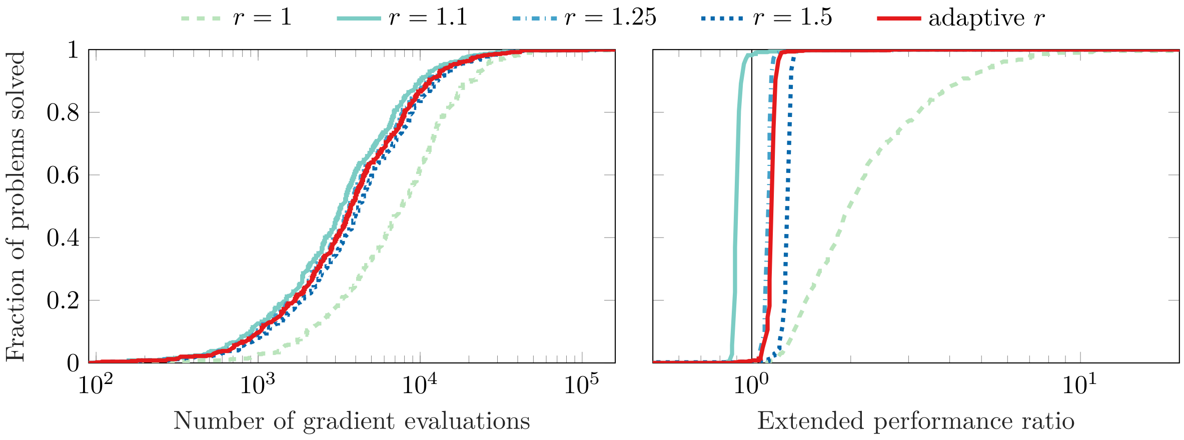

Figure 6 Nonnegative PCA problem (18): comparison for different regret factors $r \geq 1$. Data profiles (left) and extended performance profiles (right) relative to the number of gradient evaluations.

Problem size and tolerance

To investigate scalability and influence of accuracy requirements, we consider instances of (18) with dimensions $n \in \{10, \lceil 10^{1.5} \rceil, 10^2, \lceil 10^{2.5} \rceil, 10^3\}$ and tolerances $\epsilon_{\rm p} = \epsilon_{\rm d} = \varepsilon \in \{10^{-1.5}, 10^{-2}, 10^{-2.5}, 10^{-3}\}$. Each of these tolerance parameters is tested on 500 problem instances, for a total of 2000 calls to Algorithm 1.

All instances are solved up to the desired primal-dual tolerances. The results are graphically summarized in Figures 7 and 8, where it is clear that stricter tolerances demand more effort, as expected. However, it is interesting to look at how the computational cost significantly increases with the accuracy requirement, because of the slow tail convergence typical of first-order methods such as IP-FB. The influence of tolerance and problem size is depicted in Figure 9, which displays for each pair (n, ε) the number of gradient evaluations with a jitter plot and reports an estimate of the cumulative distribution function with the associated median value.[5] This chart visualizes how problem size and accuracy requirement affect the solution process, and reveals the stark effect of both n and ε.

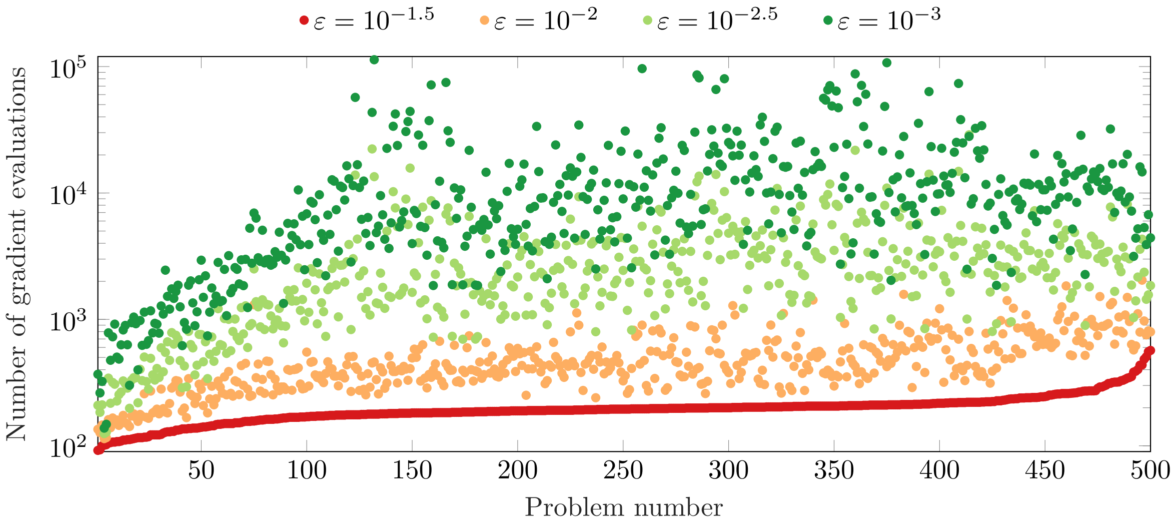

Figure 7 Nonnegative PCA problem (18): comparison for increasing accuracy requirements (decreasing tolerances $\epsilon_{\rm p} = \epsilon_{\rm d} = \varepsilon$). Epi-profiles ordered in relation to $\varepsilon = 10^{-1.5}$.

[5] Jitter plots offer a simple way of visualizing the distribution of numerical values over categories. Sample values are plotted as dots along one axis, shifted randomly along the other axis; the jittering has no meaning in itself data-wise, but allows a better view of overlapping data points. Jitter plots are complemented with the cumulative distribution function, as opposed to the probability density function, since a robust estimate of the former does not require additional assumptions. The combined plot thus conveys information on the number of data points and their density distribution in an honest and comprehensible format.

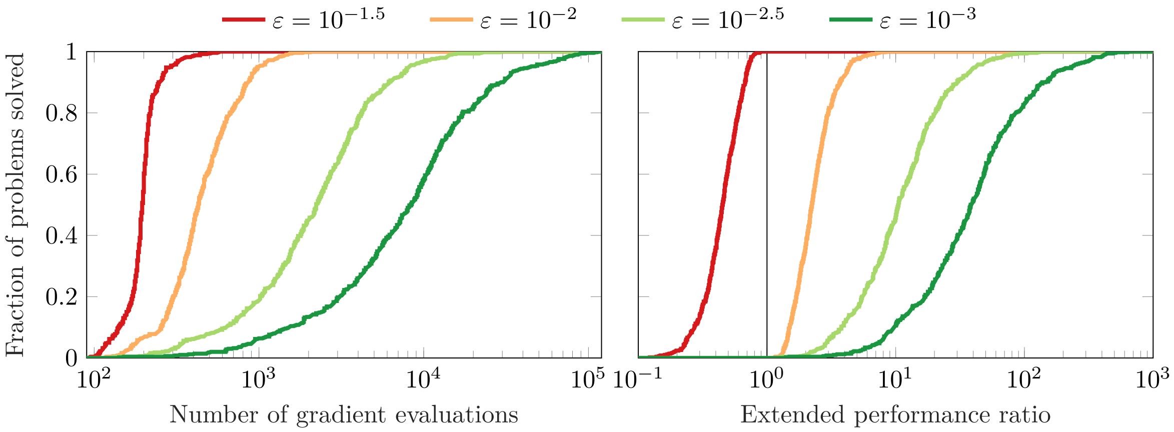

Figure 8 Nonnegative PCA problem (18): comparison for increasing accuracy requirements (decreasing tolerances $\epsilon_p = \epsilon_d = \varepsilon$). Data profiles (left) and extended performance profiles (right) relative to the number of gradient evaluations.

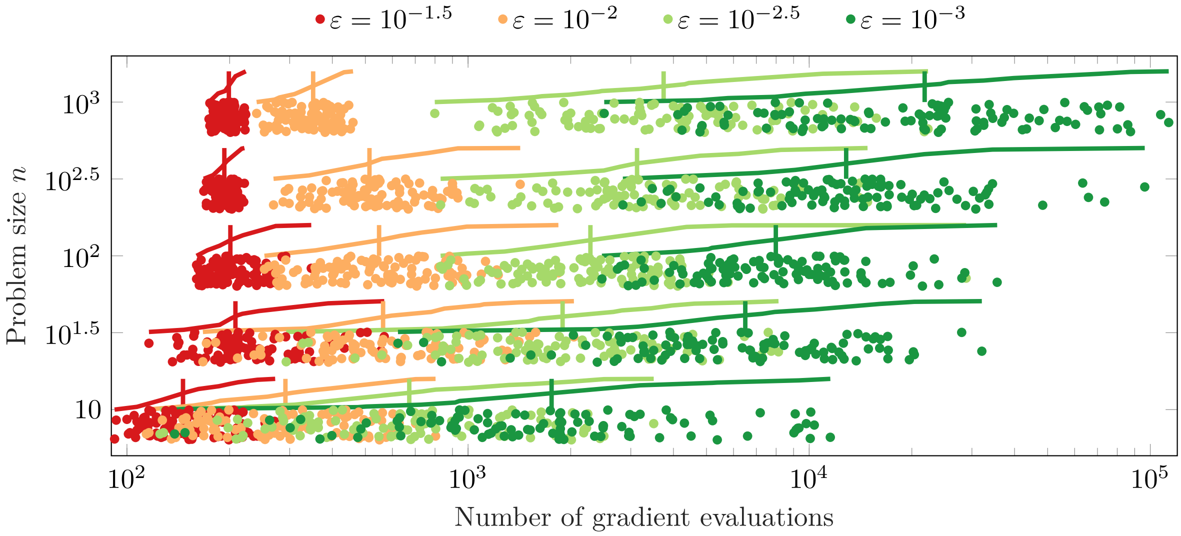

Figure 9 Nonnegative PCA problem (18): comparison for increasing accuracy requirements (decreasing tolerances $\epsilon_p = \epsilon_d = \varepsilon$) and problem sizes n. Combination of jitter plot (dots) and cumulative distribution function estimate (solid line) with median value (vertical line).

6 Conclusions

We proposed an interior point (IP) method for nonsmooth minimization subject to smooth inequality constraints, where the inner barrier subproblems are addressed by means of proximal gradient iterations. The methodology is an extension to a fully nonconvex setting of the PIPA algorithm proposed in [11], and aims at bridging the gap between IP and proximal algorithms, the former being the methods of choice for coping with complex constraints and the latter being well suited for large-scale nonsmooth problems. The result is a warm-startable iterative scheme whose output are approximate KKT-optimal pairs for the problem. Our analysis of proximal gradient iterations is novel, offering weaker conditions to ensure convergence results in the fully nonconvex setting.

Despite the benefits of adopting nonmomontone stepsize sequences demonstrated by our numerical simulations, the method suffers from the slow tail convergence that is typical of first-order methods. These observations motivate future research directions toward integrating the methodology with more adaptive and higher-order schemes. While the direct adoption of accelerated solvers along the lines of [15, 41] seems far from trivial, variable-metric or proximal-Newton approaches could be viable options for coping with the ill conditioning inherent to the barrier subproblems, as observed in [11]. Other interesting developments include gaining a deeper understanding on the choice of barrier parameters and inner tolerances to improve convergence and output quality. Finally, a non-asymptotic analysis of Algorithm 1 and 2 is left for future work, to shed light on whether there is a uniform upper bound on the number of steps, or under which conditions. In particular, as condition 2.4(a) affects the linesearch procedure, maintaining strict feasibility seems to hinder complexity estimates in the nonconvex setting of Assumption 1, suggesting that additional assumptions may be required for the purpose.

References

1 Masoud Ahookhosh, Andreas Themelis, and Panagiotis Patrinos. A Bregman Forward-Backward Linesearch Algorithm for Nonconvex Composite Optimization: Superlinear Convergence to Nonisolated Local Minima. *SIAM J. Optim.*, 31(1):653–685, 2021.

2 Anna Altman and Jacek Gondzio. Regularized symmetric indefinite systems in interior point methods for linear and quadratic optimization. *Optim. Methods Softw.*, 11(1–4):275–302, 1999.

3 Paul Armand and Riadh Omheni. A Mixed Logarithmic Barrier-Augmented Lagrangian Method for Nonlinear Optimization. *J. Optim. Theory Appl.*, 173(2):523–547, 2017.

4 Amir Beck and Nili Guttmann-Beck. FOM – a MATLAB toolbox of first-order methods for solving convex optimization problems. *Optim. Methods Softw.*, 34(1):172–193, 2019.

5 Pourya Behmandpoor, Puya Latafat, Andreas Themelis, Marc Moonen, and Panagiotis Patrinos. SPIRAL: A Superlinearly Convergent Incremental Proximal Algorithm for Nonconvex Finite Sum Minimization. *Comput. Optim. Appl.*, 88(1):71–106, 2024.

6 Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

7 Ernesto G. Birgin and José Mario Martínez. *Practical Augmented Lagrangian Methods for Constrained Optimization*. Society for Industrial and Applied Mathematics, 2014.

8 Jérôme Bolte, Shoham Sabach, Marc Teboulle, and Yakov Vaisbourd. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM J. Optim.*, 28(3):2131–2151, 2018.

9 Andrea Brilli, Giampaolo Liuzzi, and Stefano Lucidi. An interior point method for nonlinear constrained derivative-free optimization. `https://arxiv.org/abs/2108.05157v2`, 2022.

10 Feishe Chen, Lixin Shen, and Bruce W. Suter. Computing the proximity operator of the ℓ_p norm with $0 < p < 1$. *IET Signal Process.*, 10(5):557–565, 2016.

11 Emilie Chouzenoux, Marie-Caroline Corbineau, and Jean-Christophe Pesquet. A Proximal Interior Point Algorithm with Applications to Image Processing. *J. Math. Imaging Vis.*, 62(6):919–940, 2020.

12 Frank E. Curtis. A penalty-interior-point algorithm for nonlinear constrained optimization. *Math. Program. Comput.*, 4(2):181–209, 2012.

13 Alberto De Marchi. Proximal gradient methods beyond monotony. *Journal of Nonsmooth Analysis and Optimization*, 4, 2023.

14 Alberto De Marchi, Xiaoxi Jia, Christian Kanzow, and Patrick Mehlitz. Constrained composite optimization and augmented Lagrangian methods. *Math. Program.*, 201(1):863–896, 2023.

15 Alberto De Marchi and Andreas Themelis. Proximal Gradient Algorithms under Local Lipschitz Gradient Continuity: A Convergence and Robustness Analysis of PANOC. *J. Optim. Theory Appl.*, 194(3):771–794, 2022.

16 Anthony V. Fiacco and Garth P. McCormick. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. John Wiley & Sons, 1968.

17 Anders Forsgren, Philip E. Gill, and Margaret H. Wright. Interior Methods for Nonlinear Optimization. *SIAM Rev.*, 44(4):525–597, 2002.

18 Ragnar Frisch. The logarithmic potential method of convex programming. Technical report, University Institute of Economics, 1955.

19 Philip E. Gill, Walter Murray, Michael A. Saunders, John A. Tomlin, and Margaret H. Wright. On projected Newton barrier methods for linear programming and an equivalence to Karmarkar's projective method. *Math. Program.*, 36(2):183–209, 1986.

20 Jacek Gondzio. Interior point methods 25 years later. *Eur. J. Oper. Res.*, 218(3):587–601, 2012.

21 Christian Kanzow and Patrick Mehlitz. Convergence Properties of Monotone and Nonmonotone Proximal Gradient Methods Revisited. *J. Optim. Theory Appl.*, 195(2):624–646, 2022.

22 Narendra Karmarkar. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4(4):373–395, 1984.

23 Leonid G. Khachiyan. A polynomial algorithm in linear programming. *Sov. Math., Dokl.*, 20:191–194, 1979.

24 Zhijian Lai and Akiko Yoshise. Riemannian Interior Point Methods for Constrained Optimization on Manifolds. *J. Optim. Theory Appl.*, 201(1):433–469, 2024.

25 Puya Latafat, Andreas Themelis, Masoud Ahookhosh, and Panagiotis Patrinos. Bregman Finito/MISO for nonconvex regularized finite sum minimization without Lipschitz gradient continuity. *SIAM J. Optim.*, 32(3):2230–2262, 2022.

26 Puya Latafat, Andreas Themelis, and Panagiotis Patrinos. On the convergence of adaptive first order methods: proximal gradient and alternating minimization algorithms. https://arxiv.org/abs/2311.18431v1, 2023.

27 Puya Latafat, Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos. Adaptive proximal algorithms for convex optimization under local Lipschitz continuity of the gradient. https://arxiv.org/abs/2301.04431v4, 2023.

28 Tianyi Lin, Shiqian Ma, Yinyu Ye, and Shuzhong Zhang. An ADMM-based interior-point method for large-scale linear programming. *Optim. Methods Softw.*, 36(2–3):389–424, 2021.

29 Changshuo Liu and Nicolas Boumal. Simple Algorithms for Optimization on Riemannian Manifolds with Constraints. *Appl. Math. Optim.*, 82(3):949–981, 2020.

30 Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively Smooth Convex Optimization by First-Order Methods, and Applications. *SIAM J. Optim.*, 28(1):333–354, 2018.

31 Ashutosh Mahajan, Sven Leyffer, and Christian Kirches. Solving Mixed-Integer Nonlinear Programs by QP-Diving. Technical report, Mathematics and Computer Science Division, Argonne National Laboratory, 2012. Preprint ANL/MCS-P2071-0312.

32 Yura Malitsky and Konstantin Mishchenko. Adaptive Gradient Descent without Descent. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 6702–6712. PMLR, 2020.

33 Yura Malitsky and Konstantin Mishchenko. Adaptive Proximal Gradient Method for Convex Optimization. https://arxiv.org/abs/2308.02261v2, 2023.

34 Andrea Montanari and Emile Richard. Non-Negative Principal Component Analysis: Message Passing Algorithms and Sharp Asymptotics. *IEEE Trans. Inf. Theory*, 62(3):1458–1484, 2016.

35 Boris S. Mordukhovich. *Variational Analysis and Applications.* Springer, 2018.

36 Jorge J. Moré and Stefan M. Wild. Benchmarking Derivative-Free Optimization Algorithms. *SIAM J. Optim.*, 20(1):172–191, 2009.

37 Yurii Nesterov and Arkadii Nemirovkii. *Interior-Point Polynomial Algorithms in Convex Programming.* Society for Industrial and Applied Mathematics, 1994.

38 R. Tyrrell Rockafellar and Roger J. B. Wets. *Variational analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften.* Springer, 1998.

39 Saverio Salzo. The Variable Metric Forward-Backward Splitting Algorithm Under Mild Differentiability Assumptions. *SIAM J. Optim.*, 27(4):2153–2181, 2017.

40 Pantelis Sopasakis, Emil Fresk, and Panagiotis Patrinos. OpEn: Code Generation for Embedded Nonconvex Optimization. *IFAC-PapersOnLine*, 53(2):6548–6554, 2020. 21st IFAC World Congress.

41 Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos. Forward-Backward Envelope for the Sum of Two Nonconvex Functions: Further Properties and Nonmonotone Linesearch Algorithms. *SIAM J. Optim.*, 28(3):2274–2303, 2018.

42 Tuomo Valkonen. Interior-proximal primal-dual methods. *Appl. Anal. Optim.*, 3(1):1–28, 2019.

43 Robert J. Vanderbei and David F. Shanno. An Interior-Point Algorithm for Nonconvex Nonlinear Programming. *Comput. Optim. Appl.*, 13(1):231–252, 1999.

44 Andreas Wächter and Lorenz T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math. Program.*, 106(1):25–57, 2006.

45 Xianfu Wang and Ziyuan Wang. A Bregman inertial forward-reflected-backward method for nonconvex minimization. *J. Glob. Optim.*, 89(2):327–354, 2023.

46 Margaret H. Wright. The interior-point revolution in optimization: history, recent developments, and lasting consequences. *Bull. Am. Math. Soc.*, 42(1):39–56, 2005.

47 Stephen J. Wright. *Primal-Dual Interior-Point Methods.* Society for Industrial and Applied Mathematics, 1997.

48 Zongben Xu, Xiangyu Chang, Fengmin Xu, and Hai Zhang. $L_{1/2}$ Regularization: A Thresholding Representation Theory and a Fast Solver. *IEEE Trans. Neural Netw. Learn. Syst.*, 23(7):1013–1027, 2012.

49 Tong Yang, Michael I. Jordan, and Tatjana Chavdarova. Solving Constrained Variational Inequalities via a First-order Interior Point-based Method. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.