Workshop on
Nonsmooth Optimization and Applications
NOPTA 8-12 April 2024
University of Antwerp, Belgium

Penalty interior approaches for constrained nonsmooth nonconvex problems

Alberto De Marchi
University of the Bundeswehr Munich, Germany
alberto.demarchi@unibw.de

Universität der Bundeswehr München

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. We provide a theoretical characterization of the algorithm and its asymptotic properties, deriving convergence results for fully nonconvex problems. 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.

We foresee a combination with (exact) penalty methods to accommodate equality constraints and to be able to invoke generic (possibly accelerated) prox-grad subsolvers.

Joint work with Andreas Themelis (Kyushu University).

Introduction

Problem:

$$\begin{aligned} &\text{minimize} && q(x) := f(x) + g(x) && \text{over} \quad x \in \mathbb{R}^n \qquad \text{(P)} \\ &\text{subject to} && c(x) \leq 0, \end{aligned}$$

Assumptions:

(A1) $f\colon \mathbb{R}^n \to \mathbb{R}$ has a locally Lipschitz-continuous gradient;
(A2) $g\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is proper, lsc, prox-bounded;
(A3) $c\colon \mathbb{R}^n \to \mathbb{R}^m$ has locally Lipschitz-continuous Jacobian;
(A4) well-posed: $\inf\{q(x) \mid c(x) \leq 0\} \in \mathbb{R}$;
(A5) strictly feasible: $\operatorname{dom} q \cap \{x \in \mathbb{R}^n \mid c(x) < 0\} \neq \emptyset$.

Objective $q := f + g$ can be nonconvex, as well as f and g, but g has easily computable proximal mapping.

Methodology

Combines IP and proximal algorithms for coping with nontrivial constraints and large-scale nonsmooth problems.

The IP framework builds upon a *barrier* function b to replace the *inequality* constraints. We fix a nonnegative and smooth barrier function b that complies with

(B1) $\operatorname{dom} b = (-\infty, 0)$; $b(t) \to \infty$ as $t \to 0^-$;
(B2) b is twice cont. diff. with $b' > 0$ on its domain.

A prominent example for b is $b(t) := -1/t$ for $t < 0$, $b(t) := \infty$ otherwise. Sequence of "unconstrained" subproblems

$$\text{minimize} \quad q_\mu(z) := f_\mu(z) + g(z) \qquad \text{over} \quad z \in \mathbb{R}^n, \qquad (\mathrm{P}_\mu)$$

whose differentiable term $f_\mu\colon \mathbb{R}^n \to \mathbb{R}$ includes the barrier weighted by $\mu > 0$:

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

Instances of (P_μ) are *somehow* suited for prox-grad solvers, but f_μ has not full domain. Need tailored prox-grad scheme!

▹ Algorithm 1 addresses (P) by constructing and solving a sequence of barrier subproblems (P_μ).

▹ Algorithm 2 provides a barrier-friendly proximal gradient method that is well suited as inner solver for Algorithm 1.

Contributions

⇝ (P) brings together structured objectives with nontrivial constraints in the fully nonconvex setting.
⇝ Warm-startable scheme with convergence guarantees under mild conditions.
⇝ Eventually output an approximate KKT-optimal pair.
⇝ Adaptive prox-grad steps generate (strictly) feasible iterates while guaranteeing a descent-type condition.
⇝ Cope with lack of full domain of locally smooth f_μ.
⇝ No artificial bound on stepsize sequence for prox-grad.

Preliminaries

Definition 1 (ε-stationarity) *Relative to* (P_μ)*, a point $z^\star$ is ε-stationary for some $\varepsilon \geq 0$ if $\operatorname{dist}(0, \partial q_\mu(z^\star)) \leq \varepsilon$. When $\varepsilon = 0$, i.e., when $0 \in \partial q_\mu(z^\star)$, $z^\star$ is said to be* stationary.

Definition 2 (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 $c(x^\star) < 0$.*

Definition 3 ($(\epsilon_\mathrm{p}, \epsilon_\mathrm{d})$-KKT optimality) *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*

$$\begin{aligned} &\operatorname{dist}\big(-\nabla c(x^\star)^\top y^\star,\, \partial q(x^\star)\big) \leq \epsilon_\mathrm{d} \\ \text{and} \quad &\min\{-c_i(x^\star), y_i^\star\} \leq \epsilon_\mathrm{p} \qquad \forall i = 1, \ldots, m. \end{aligned}$$

Proposition 4 *Informally, an asymptotic counterpart of approximate KKT-optimality is necessary for optimality.*

Numerical methods

Algorithm 1: Interior point method for (P) using IP-FB as inner subsolver

require : x^0 strictly feasible starting point,
$\epsilon_\mathrm{p}, \epsilon_\mathrm{d} > 0$ primal-dual tolerances
provide : $x^\star$ $(\epsilon_\mathrm{p}, \epsilon_\mathrm{d})$-KKT optimal point for (P)
initialize: $\varepsilon_0, \mu_0 > 0$ tolerance and barrier parameters,
$\theta_\varepsilon, \theta_\mu \in (0, 1)$ update coefficients

begin step $k = 0, 1, \ldots$
    $x^{k+1} \leftarrow$ IP-FB$(x^k, \mu_k, \varepsilon_k)$ [ε_k-stationary for q_{μ_k}]
    Set $y_i^{k+1} \leftarrow \mu_k b'(c_i(x^{k+1}))$ for all $i = 0, \ldots, m$
    If $\varepsilon_k \leq \epsilon_\mathrm{d}$ and $\max_i \min\{-c_i(x^{k+1}), y_i^{k+1}\} \leq \epsilon_\mathrm{p}$
        return $(x^\star, y^\star) \leftarrow (x^{k+1}, y^{k+1})$
    Set $\varepsilon_{k+1} \leftarrow \max\{\epsilon_\mathrm{d}, \theta_\varepsilon \varepsilon_k\}$ and $\mu_{k+1} \leftarrow \theta_\mu \mu_k$
end

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

require : z strictly feasible starting point,
$\mu > 0$ barrier coefficient, $\varepsilon > 0$ tolerance
provide : $z^\star$ strictly feasible ε-stationary point for (P_μ)
initialize: $\gamma_0 \in (0, \gamma_g)$ initial stepsize, $r \geq 1$ regret factor,
$\alpha, \beta \in (0, 1)$ backtracking parameters

set $z^0 \leftarrow z$ and **begin** step $j = 0, 1, \ldots$
    If $j \geq 1$, $\gamma_j \leftarrow r\gamma_{j-1}$ and $z^j \leftarrow \bar{z}^{j-1}$
    while *true* **do**
        Compute $\bar{z}^j \in \operatorname{prox}_{\gamma_j g}(z^j - \gamma_j \nabla f_\mu(z^j))$
        If $q_\mu(\bar{z}^j) \leq q_\mu(z^j) - \frac{1-\alpha}{2\gamma_j}\|\bar{z}^j - z^j\|^2$ and
        $\|\nabla f_\mu(\bar{z}^j) - \nabla f_\mu(z^j)\| \leq \frac{\alpha}{\gamma_j}\|\bar{z}^j - z^j\|$
            break; else $\gamma_j \leftarrow \beta\gamma_j$
    endw
    If $\|\frac{1}{\gamma_j}(z^j - \bar{z}^j) - \nabla f_\mu(z^j) + \nabla f_\mu(\bar{z}^j)\| \leq \varepsilon$
        return $z^* \leftarrow \bar{z}^j$
end

Main Results

Algorithm 2: barrier-friendly prox-grad

Barrier subproblems are solved, up to approximate stationarity, with (tailored) proximal gradient iterations.

Theorem 5 (Asymptotic analysis) *The iterates generated by Algorithm 2 with $\varepsilon = 0$ satisfy the following:*

1. $\{q_\mu(z^j)\}$ converges to a finite value $q_\mu^\star \geq \inf q_\mu$.
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. If q_μ is level bounded, then $\{\bar{z}^j\}$ and $\{z^j\}$ are bounded and $\{\gamma_j\}$ remains bounded away from zero.
5. $\sum_{j \in \mathbb{N}} \gamma_j = \infty$ and $\liminf_{j \to \infty} \frac{1}{\gamma_j}\|\bar{z}^j - z^j\| = 0$.
6. $\liminf_{j \to \infty} \|\frac{1}{\gamma_j}(z^j - \bar{z}^j) - \nabla f_\mu(z^j) + \nabla f_\mu(\bar{z}^j)\| = 0$.
7. If the iterates remain bounded, then the set ω of accumulation points of $\{\bar{z}^j\}$ is made of stationary points for q_μ, and q_μ is constantly equal to $q_\mu^\star$ on ω.

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

Algorithm 1: interior point scheme

Convergence to feasible and KKT-optimal pairs in the fully nonconvex setting $\implies$ finite termination.

Lemma 6 (Algorithmic behavior) *Consider a sequence $\{x^k, y^k\}$ generated by Algorithm 1. Then, for every $k \geq 0$,*

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 strictly feasible and ε_k-stationary for q_{μ_k}.
3. $y^{k+1} \geq 0$ and $\operatorname{dist}\big(-\nabla c(x^{k+1})^\top y^{k+1}, \partial q(x^{k+1})\big) \leq \varepsilon_k$.

Theorem 7 (Asymptotic analysis) *Consider a sequence $\{x^k, y^k\}$ generated by Algorithm 1. Then,*

1. any limit point of $\{x^k\}$ is feasible.
2. If $\epsilon_\mathrm{p} = 0$ or $\epsilon_\mathrm{d} = 0$, then $\lim_{k \to \infty} \min\{-c(x^k), y^k\} = 0$.

If $\epsilon_\mathrm{d} = 0$, Algorithm 1 runs indefinitely with $\varepsilon_k, \mu_k \to 0$ and the following also hold for a subsequence $\{x_k\}_{k \in K}$ converging to some point $x^\star$:

3. $x^\star$ is feasible and asymptotically KKT-optimal for (P)*.*
4. If $\{y^k\}_{k \in K}$ is bounded, then $x^\star$ is KKT-optimal for (P)*.*

Download

webpage   poster   preprint

Simulations

Nonnegative PCA

Principal component analysis (PCA) aims at estimating the direction of maximal variability of a high-dimensional dataset. Here we impose nonnegativity of entries as prior knowledge, and solve PCA restricted to the positive orthant:

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

Nonnegative PCA is an NP-hard nonconvex problem that cannot be addressed by standard SVD directly.

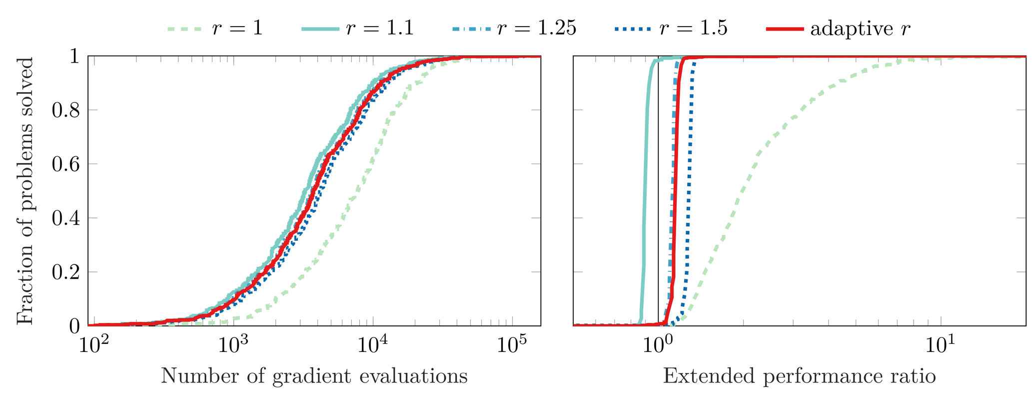

Figure 1: *Comparison for different regret factors r. Data profiles (left) and extended performance profiles (right) relative to the number of gradient evaluations.*

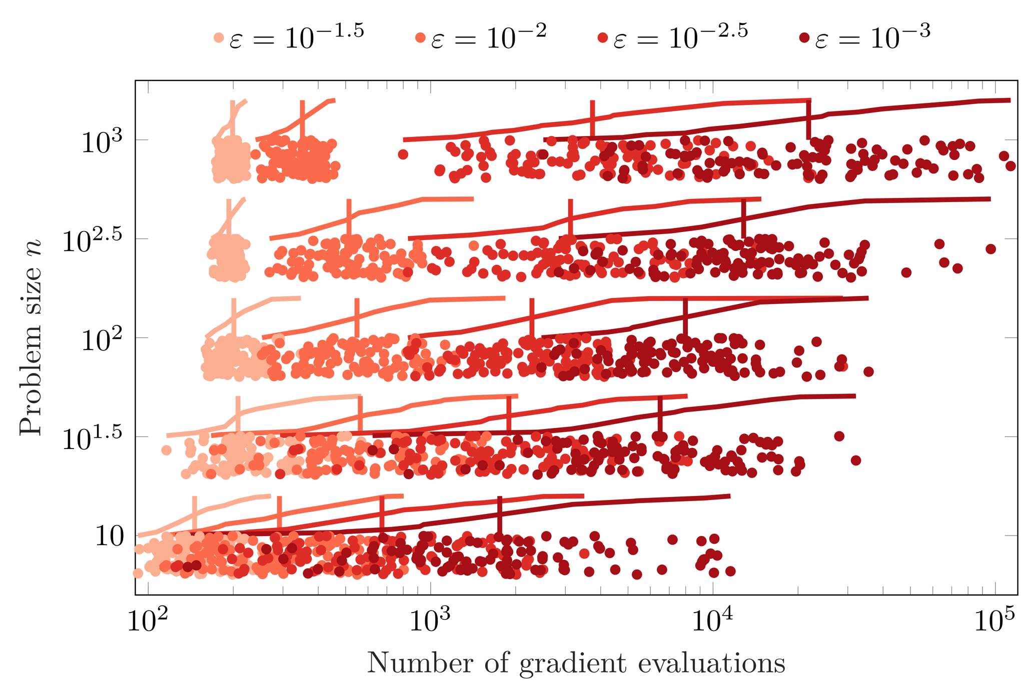

Figure 2: *Comparison for increasing accuracy requirements (decreasing tolerances $\epsilon_\mathrm{p} = \epsilon_\mathrm{d} = \varepsilon$) and problem sizes n. Combination of jitter plot (dots) and cumulative distribution function (solid line) with median (vertical line).*

Final remarks

⇝ IP + prox-grad method for nonsmooth nonconvex minimization subject to smooth inequality constraints.

Outlook

How to handle equalities? How to overcome the slow tail convergence that is typical of first-order methods?

Relax and marginalize

Given $\alpha > 0$, consider an L^1-relaxation of (P), equivalently cast by introducing a slack variable $s \in \mathbb{R}^m$ as

$$\begin{aligned} &\text{minimize} && q(x) + \alpha\langle 1, s\rangle && \text{over} \quad x \in \mathbb{R}^n,\, s \in \mathbb{R}^m_+ \\ &\text{subject to} && c(x) \leq s. \end{aligned}$$

Build the barrier problem as (P_μ) and marginalize s.
⇝ Equalities and bounds are easy to include too.
⇝ Penalty-barrier subproblems are prox-friendly structured and the smooth term has full domain.
⇝ Any prox-grad solver (capable of handling local, not global, smoothness) can do as subsolver.
⇝ Infeasible method! but possible to steer convergence with penalty and barrier parameters α, μ.

References

[1] Emilie Chouzenoux, Marie-Caroline Corbineau, and Jean-Christophe Pesquet. A proximal interior point algorithm with applications to image processing. *Journal of Mathematical Imaging and Vision*, 62(6):919–940, 2020.

[2] Alberto De Marchi and Andreas Themelis. An interior proximal gradient method for nonconvex optimization. *arXiv:2208.00799*, 2022.

[3] Geoffroy Leconte and Dominique Orban. An interior-point trust-region method for nonsmooth regularized bound-constrained optimization. *arXiv:2402.18423*, 2024.

[4] Tuomo Valkonen. Interior-proximal primal-dual methods. *Applied Analysis and Optimization*, 3(1):1–28, 2019.