Ghost penalties in nonconvex constrained optimization: Diminishing stepsizes and iteration complexity

Francisco Facchinei

Dept. of Computer, Control, and Management Engineering
University of Rome La Sapienza
Roma, Italy

Games, Dynamics, and Optimization 2019
April 9-11, 2019
Babeș-Bolyai University, Cluj-Napoca

Co-authors

This presentation is based on work done with

Lorenzo Lampariello	University of Roma 3, Italy
Vyacheslav Kungurtsev	Czech Technical University of Prague, Czech Republic
Gesualdo Scutari	Purdue University, Indiana, USA

The problem

$$\begin{aligned} \min_x \quad & f(x) \\ & g_1(x) \leq 0 \\ & \vdots \\ & g_m(x) \leq 0 \\ & x \in K \end{aligned}$$

f, g_1, $\ldots$ g_m are C^1

$K \subseteq \Re^n$ is closed + convex

$$g(x) = \begin{pmatrix} g_1(x) \\ \vdots \\ g_m(x) \end{pmatrix}$$

and therefore

$$\begin{aligned} \min_x \quad & f(x) \\ & g(x) \leq 0 \\ & x \in K \end{aligned}$$

What are we looking for?

We do not assume constraint qualifications or any other conditions and show that the algorithm behaves sensibly even in "extreme situations"

Our target are generalized stationary points, i.e a point x that satisfies one of the three following conditions

- It is a stationary point of the feasibility problem $\min \|g_+(x)\|$
- It is a Fritz-John point
- It is a KKT point

A point satisfying any of the above three conditions is termed a (generalized) stationary point

Only after this study has been accomplished will we add conditions to exclude bad cases

---

$a_+ \triangleq \max\{0, a\}$

Two main results

- The first Diminishing Stepsize Method (DSM) for nonconvex constrained optimization problems
- The first iteration complexity result for an SQP-type method, one of the very few iteration complexity results in nonconvex constrained optimization

For both results the main tool is the (sometimes) invisible use of some penalty function

Diminishing Stepsize Methods

DSMs generate a sequence $\{x^\nu\}$ by setting

$$x^{\nu+1} = x^\nu + \gamma^\nu d^\nu,$$

where d^ν is a suitable direction and γ^ν is a positive stepsize chosen, possibly independently of the problem at hand, so that

$$\gamma^\nu \downarrow 0, \qquad \sum_{\nu=0}^{\infty} \gamma^\nu = \infty.$$

DSM (for smooth problems): pros and cons

Pros

- Very simple
- Useful when the computation of the objective function is costly
- Important when there is "noise"
- Effective in distributed and incremental methods

Cons

- Slow in practice

Recent surge of interest because of Big Data

- DSMs first introduce in the '60 in connection to nondifferentiable convex optimization
- Convergence based on decrease of iterates' distance to the solution set
- If applied to differentiable problems, alternative arguments available, based on objective function decrease, and also nonconvex problems can be tackled

For differentiable problems the situation is (roughly) the following:

	Convex	Nonconvex
Unconstrained	Yes	Yes
Constrained	Yes	No

We aim at filling the right bottom corner

The basic idea is to use an SQP-like direction: d^ν is the solution of

$$\begin{aligned} \min_d \quad & f(x^\nu) + \nabla f(x^\nu)^T d + \tfrac{1}{2}\|d\|^2 \\ & g(x^\nu) + \nabla g(x^\nu)^T d \leq 0 \\ & x^\nu + d \in K \end{aligned}$$

Possible problems:

- This subproblem could have an empty feasible set
- The subproblem could be a bad approximation of the original problem

To deal with the feasibility issue, we enlarge the feasible set

$$\begin{aligned} \min_d \quad & f(x^\nu) + \nabla f(x^\nu)^T d + \tfrac{1}{2}\|d\|^2 \\ & g(x^\nu) + \nabla g(x^\nu)^T d \leq \kappa(x) \\ & x^\nu + d \in K \end{aligned}$$

where $\kappa(x) \geq 0$ and

- to compute $\kappa(x)$ we must solve a (surely solvable) problem

$$\kappa(x) \triangleq \frac{1}{2} \max_i \{g_i(x)_+\}$$

$$+\frac{1}{2} \min_d \left\{ \max_i \{\tilde{g}(d;x)_+\} \mid \|d\|_\infty \leq \rho,\ x + d \in K \right\}$$

  where $\rho > 0$ is any positive constant

- if x is feasible we can take $\kappa(x) = 0$

For technical reasons we also add a simple bound on the magnitude of d

$$\begin{aligned} \min_d \quad & f(x^\nu) + \nabla f(x^\nu)^T d + \tfrac{1}{2}\|d\|^2 \\ & g(x^\nu) + \nabla g(x^\nu)^T d \leq \kappa(x) \\ & \|d\|_\infty \leq \beta \\ & x^\nu + d \in K \end{aligned}$$

where $\beta > \rho$ is a given constant.

To have a better approximation of the original problem and therefore a better direction d we can use approximations of f and g different from the linear/quadratic ones showed before

$$\begin{aligned} \min_d \quad & \tilde{f}(d; x^\nu) \\ & \tilde{g}(d; x^\nu) \leq \kappa(x) \\ & \|d\|_\infty \leq \beta \\ & x^\nu + d \in K \end{aligned}$$

- $\tilde{f}(\bullet; x)$ is strongly convex
- $\nabla_d \tilde{f}(0; x) = \nabla f(x)$
- $\tilde{g}(\bullet; x)$ is convex
- $\nabla_d \tilde{g}(0; x) = \nabla g(x)$
- $\tilde{g}(0; x) = g(x)$
- $+$ other technical assumptions

Example 1

Suppose

$$f(x) = f_1(x) + f_2(x)$$

with $f_1(x)$ convex.

We might take

$$\tilde{f}(x; x^k) = f_1(x) + \nabla f_2(x^k)^T (x - x^k) + \|x - x^k\|^2$$

which gives a better approximation than linearizing the whole objective function

The underlying assumptions is that solving a strongly convex optimization problem is "easy"

Example 2

Suppose that we can find $\tilde{g}_i(x; x^k)$ such that

$$\tilde{g}_i(x; x^k) \geq g(x), \qquad \forall x$$

If the starting point of the algorithm is feasible, our procedure will generate only feasible iterates ($\gamma^k \leq 1$)

- If g_i has a Lipschitz gradient, we can set

$$\tilde{g}_i(x; x^k) = g_i(x^k) + \nabla g_i(x^k)^T(x - x^k) + \frac{L_i}{2}\|x - x^k\|^2 \geq g_i(x)$$

- If g_i is a DC function $g_i = g_i^+ - g_i^-$ with both g_i^+ and g_i^- convex, we can set

$$\tilde{g}_i(x; x^k) = g_i^+(x) \nabla g_i^k(x^k)^T(x - x^k) + \|x - x^k\|^2 \geq g_i(x)$$

- Many interesting examples in applcations where structure can be suitably exploited

Convergence properties

To present the convergence properties we introduce two conditions

(a) [Positive Linear Independence of most active constraints]
We say that the extended Mangasarian-Fromovitz Constraint Qualification (eMFCQ) holds at x if

$$0 \in \nabla g(x)\xi + N_K(x),\ \xi_i \geq 0,\ \xi_i[g_i(x) - \max_i g_i(x)_+] = 0 \implies \xi = 0$$

(b)] [d-Hölder continuity] For every compact set Ω, positive constants α and θ exist such that

$$\|d(y) - d(z)\| \leq \theta \|y - z\|^\alpha$$

- eMFCQ is a generalization of the classical Mangasarian-Fromovitz condition to infeasible point. A point where eMFCQ holds can not be a FJ point or a stationary point of the violation of the constraints
- d-Hölder continuity is rather technical. It is possible to give simple conditions that guarantee its satisfaction, but since this assumption has a very ancillary role we skip a detailed discussion

Convergence properties

In the setting described so far, the algorithm is well-defined and either

(a) the sequence $\{x^\nu\}$ is unbounded or

(b) at least one limit point $\bar{x}$ of $\{x^\nu\}$ is an extended stationary point; if eMFCQ holds at $\bar{x}$, then $\bar{x}$ is a KKT point

In addition

(c) if $\{x^\nu\}$ is bounded, eMFCQ holds at every limit point of $\{x^\nu\}$, and d-Hölder continuity holds, then every limit point of $\{x^\nu\}$ is a KKT point

Of course, if K is bounded the sequence $\{x^\nu\}$ is bounded and only case (b) can happen

Example 3

$$\min \quad x^2$$

$$e^x \leq 0$$

This (convex) problem is infeasible and therefore no algorithm will ever be able to find a KKT or FJ point. The only reasonable thing the algorithm can do is to try to minimize the constraint violation e^x. But the constraint violation has neither minimum nor stationary points and any minimization method will produce an unbounded sequence.

In other words this problem has no extended stationary points and generting an unbounded sequence is the natural outcome

Where are the penalty functions?
They are used in the proofs only and therefore are not seen in the algorithm description!

We use

$$P(x;\varepsilon) = f(x) + \frac{1}{\varepsilon}\|g(x)_+\|_\infty$$

Essentially the proof of convergence boils down to showing that either there is a sufficiently small ε for which the direction d provides "uniform sufficient decrease" at each step or this is not the case. In this latter case it follows that at least one limit point is a Fritz-John point or a stationary point of the feasibility problem

- $\min_x \quad f(x) + c(x)$

  $g(x) \leq 0$

  $x \in K$

where c is a convex (non necessarily differentiable) continuous function

• Boundedness of the sequence $\{x^\nu\}$ can be obtained without assuming K boudned but using "coerciviness condition"

Iteration Complexity

Complexity results are classical for combinatorial optimization problems and for convex optimization ones

They are much rarer for nonconvex optimization problems and actually there is only one real result, in 2016, for nonconvex constrained problems.

Iteration complexity refers to the number of iterations necessary to find a $\delta-$approximate stationary point

- Iteration complexity for gradient method for smooth unconstrained minimization is "classical" (see Nesterov 2013 book)
- Nesterov and Polyak 2006 paper on cubic regularization for unconstrained optimization sparkled much interest
- In the past 10 year Cartis, Gould, and Toint produce a number of papers on iteration complexity under different settings
- Essentially, there is only one paper, by Birgin, Gardenghi, Martìnez, Santos, and Toint 2016 where iteration complexity for an algorithm for nonconvex constrained optimization is studied. There, a complexity of $O(\delta^{-3})$ is established

The setting

$$\begin{aligned} \min_x \quad & f(x) \\ & g(x) \leq 0 \\ & x \in K \end{aligned}$$

Assumption We assume that

- K is compact
- f and g all functions involved have Lipschitz gradients on K

We denote the corresponding Lipschitz constants by L_f, L_{g_i} ....

A simple case

If the eMFCQ holds at all points of K then

If the stepsize is sufficiently small then a δ-stationary point is reached in $O(\delta^{-2})$ iterations

Under different assumptions (starting feasible point is know, for example) we can give further results of this type. These results require the knowledge of the Lipschitz constants of the problem

The general case

We give a complexity bound for a "piecewise constant" choice of the stepsizes. By this we mean that we use the usual iteration

$$x^{\nu+1} = x^{\nu} + \gamma^{\nu} d^{\nu}$$

and keep γ^{ν} constant until a certain very simple test is satisfied, in which case we reduce γ^{ν} at a prescribed level and keep it constant until the test is possibly satisfied again and γ^{ν} reduced once again

This procedure finds a $\delta-$approximate generalized stationary point in $O(\delta^{-4})$ iterations

Data: $\delta > 0$, x^0, $T^{-1} \in \left(0, \frac{2\max_i\{L_{\nabla g_i}\}}{\max\{L_{\nabla f}, \eta c\}}\right]$, $\gamma^{-1} = \frac{T^{-1}\eta c}{2\max_i\{L_{\nabla g_i}\}}$, $\nu \longleftarrow 0$;

repeat

(S.1) compute $\kappa(x^\nu)$, $d(x^\nu)$ and $\theta(x^\nu) \triangleq \max_i\{g_i(x)_+\} - \kappa(x)$;

(S.2) **if** $\|d(x^\nu)\| \leq \delta$ **then**

  stop and **return** $x_\delta = x^\nu$;

end

(S.3) **if** $\nabla f(x^\nu)^T d(x^\nu) + \eta c\|d(x^\nu)\|^2 > 0$ **and** $T^{\nu-1} > \frac{\theta(x^\nu)}{\nabla f(x^\nu)^T d(x^\nu) + \eta c\|d(x^\nu)\|^2}$ **then**

(S.4) **if** $\theta(x^\nu) \leq L_{\tilde{g}}\delta$ **then**

  stop and **return** $x_\delta = x^\nu$;

else

(S.5) set $\gamma^\nu = \frac{T^\nu \eta c}{2\max_i\{L_{\nabla g_i}\}}$, $T^\nu = \frac{1}{2}\frac{\theta(x^\nu)}{\nabla f(x^\nu)^T d(x^\nu) + \eta c\|d(x^\nu)\|^2}$;

end

else

(S.6) set $T^\nu = T^{\nu-1}$ and $\gamma^\nu = \gamma^{\nu-1}$;

end

(S.7) set $x^{\nu+1} = x^\nu + \gamma^\nu d(x^\nu)$, $\nu \longleftarrow \nu + 1$;

end

Algorithm 1: Piecewise constant stepsizes

The algorithm stop at a $\delta-$approximate stationary point in at most

$$\left\lceil \frac{16B}{(\eta c)^2} \max_i \{L_{\nabla g_i}\} \left[\frac{f^M - f^m}{\delta^3} + \frac{2B\, g_+^M}{\delta^4} \right] \right\rceil$$

iterations

η: a user chosen number in (0,1)

c: a user chosen strong monotonicity constant for $\tilde{f}$

f^M: the maximum value of f over K

f^m: the minimum value of f over K

g_+^M: the maximum value of $\|g_+\|_\infty$ over K

B: the maximum value of $\|\nabla f(x)\|\beta + \eta c \beta^2$ over K

- First it is shown that the stepsize is reduced a finite number of times

$$\left\lceil \log_2 \frac{T^{-1}2B}{\delta} \right\rceil$$

- Exploiting the fact that at each iteration $\|d(x^k)\| > \delta$ the decrease of the ghost penalty function with penalty parameter T^k in the iterations where the stepsize is constant can be estimated
- Putting together the facts above, taking into account that the penalty function increases when the penalty parameter decreases and making appropriate boundings we get the desired result

What is the meaning of the two stopping criteria $\|d(x^k)\| \leq \delta$ and $\theta(x^k) \leq \delta L_{\tilde{g}}$?

To make a long story short: $\delta \to 0$ then the point where the algorithm stops tends to a generalized stationary point

- Also the algorithm above assumes that we know several Lipschitz constants and function values. What if we do not know them? An adaptive version can be devised where by making a sort of "line-search" at each iteration to determine an appropriate value of γ^k the iteration complexity of $O(\delta^{-4})$ can still be mantained

- Boundedness of K can be substituted by the coerciveness conditions discussed previously

- [Work in progress] The results can be extended to the case in which the objective function is $f(x) + c(x)$

- [Work in progress] Using the ghost penalty approach in the convergence analysis, we are able to prove convergence for the first distributed method for problems with nonconvex constraints along with a corresponding complexity analysis

Thank you!