A nonlinear programming algorithm based on non-coercive penalty functions

Clovis C. Gonzaga* Rómulo A. Castillo†

September, 2000, revised June 2002.

Abstract

We consider first the differentiable nonlinear programming problem and study the asymptotic behavior of methods based on a family of penalty functions that approximate asymptotically the usual exact penalty function. We associate two parameters to these functions: one is used to control the slope and the other controls the deviation from the exact penalty.

We propose a method that does not change the slope for feasible iterates and show that for problems satisfying the Mangasarian-Fromovitz constraint qualification all iterates will remain feasible after a finite number of iterations. The same results are obtained for non-smooth convex problems under a Slater qualification condition.

1 Introduction

Consider the nonlinear programming problem

$$(P) \qquad \begin{array}{rl} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq \quad 0 \\ & x \in \mathbb{R}^n, \end{array}$$

where $f_i : \mathbb{R}^n \to R$ for $i = 0, 1, \ldots, m$ are continuously differentiable functions. We shall study algorithms for solving the problem (P) based on penalty functions. Each iteration solves an unconstrained problem obtained by adding a penalty function to the criterion. A well known penalty function for this problem is the exact penalty function l_1, see for instance Zangwill [15], given by

$$y \in R \mapsto t^+(y) = \max\{0, y\}. \tag{1}$$

*Department of Mathematics, Federal University of Santa Catarina, Florianópolis, Brazil, e-mail clovis@mtm.ufsc.br. Research partially supported by CAPES, Brazil

†Department of Mathematics, Universidad Centroccidental Lisandro Alvarado, Barquisimeto, Venezuela. e-mail: romulo@uicm.ucla.edu.ve

The penalized function in this case is

$$x \in I\!R^n \mapsto f_\beta(x) = f_o(x) + \beta \sum_{i=1}^{m} (f_i(x))^+ ,$$

where $\beta > 0$ is a penalty parameter. Zangwill found an interesting result for the convex case: If this function has a minimizer for some $\beta_0 > 0$, then the same will be true for any $\beta > \beta_0$. It is known that under mild hypotheses, if β is sufficiently large, then the minimizers of $f_\beta(\cdot)$ will be feasible, and consequently optimal. The obvious difficulty with the exact penalty is that it is non-differentiable, which prevents the use of efficient minimization algorithms.

Differentiable approximations to the exact penalty function were obtained in different contexts, see Zang [14], Ben Tal-Teboulle [3], and Pinar-Zenios [11]. Penalty methods based on functions of this class were studied in Auslender, Cominetti and Haddou [2], who describe algorithms for the convex case. These functions are also described by Chen and Mangasarian [4, 5], who construct them by integrating probability distributions. Xavier [13] studies a specific function in this class, the hyperbolic penalty. He was the first to use a two-parameter approach, and his work is the basis for ours. In [1], Auslender studies convex and non convex programming problems and extends results given in [2] and [3] under relaxed hypotheses. He also describes results for positive semidefinite programming problems. The goal of this work is to use in a penalty approach a family of smooth approximations to the exact l_1 penalty function, getting a result similar to the one given by Zangwill.

Our method uses two parameters, r and β, and minimizes at each iteration a penalized function with the shape

$$x \in I\!R^n \mapsto f_{\beta,r}(x) = f_o(x) + \beta r \sum_{i=1}^{m} \theta\left(\frac{f_i(x)}{r}\right), \tag{2}$$

where $\theta(\cdot)$ is a smooth approximation of the exact penalty (see below for formal definitions and hypotheses). The parameters play different roles: r decreases to improve the precision of the approximation, and β is the increasing penalty weight.

The main result of this paper is that, under a Mangasarian-Fromowitz constraint qualification, for β sufficiently large and r sufficiently small, all iterates will be feasible, thus extending to the inexact case the feasibility property of the exact penalty.

Our algorithm only increases β at infeasible iterates, and hence it will stop increasing in finite time, after what feasible points will converge to optimal solutions as the precision increases (r decreases).

When Newton like methods are used for solving the penalized problems, ill-conditioning will arise when r tends to zero. In the case of the interior logarithmic barrier function, Nash and Sofer [7, 8] present a method for stabilizing the computations. We adapt their treatment to general penalties, with a very mild hypothesis on the penalty function derivatives.

Comparison with other works. Auslender, Cominetti and Haddou [2] use penalized functions which can be rewritten as (2), but they require that in the non coercive case $\beta \to \infty$. They study only the convex case. Our work is strongly based on theirs, but we treat nonlinear programming problems and prove that under a Mangasarian-Fromowitz condition, β can stop growing in finite time.

When using the traditional quadratic external penalty $\theta(y) = (y^+)^2$, the penalized function becomes

$$f_{\beta,r}(x) = f_o(x) + \frac{\beta}{r}\sum_{i=1}^{m}(f_i(x)^+)^2.$$

It is known that in non-trivial cases all iterates will be infeasible, and approach feasibility as β/r grows. Truemper [12] takes again the convex case and shows that there exists a threshold value β_0 (dependent on the optimal KKT multipliers) such that if $\beta \geq \beta_0$, then the minimizer of the penalized function will be r-active.

Pinar and Zenios [11] extend Truemper's result by considering a linear-quadratic barrier, and obtain the same behavior. They use the parameters β and r explicitly, and derive an algorithm for convex programming which changes only one of β or r at each iteration. Again, r-active solutions will be achieved for β larger than Truemper's threshold.

Note that the quadratic external penalty $\theta(\cdot)$ has second derivatives given by $\theta''(y) = u(0)$, the step function at the origin. Pinar and Zenios use $\theta''(y) = u(0) - u(1)$, the unit pulse applied at the origin. A special case of the penalties studied by us has $\theta''(y) = u(-0.5) - u(0.5)$, the unit pulse *centered* at the origin.

Xavier [13] describes a hyperbolic penalty function with two parameters, which can be put in our framework (his work gave the inspiration for ours). He proves convergence results to optimal solutions for nonlinear programming problems under general conditions.

We state the problem in Sect. 2 and develop the algorithm in Sect. 3, followed by a simple example and the discussion of ill-conditioning. In Sect. 4 we adapt the results to the non-smooth convex problem.

2 Problem and Penalty Approach

Consider the differentiable nonlinear programming problem (P). For $\epsilon \geq 0$ define the relaxed feasible set

$$\Omega_\epsilon = \{x \in I\!R^n \mid f_i(x) \leq \epsilon \ , \ i = 1, \ldots, m\}.$$

Then $\Omega \equiv \Omega_0$ is the feasible set for (P). We shall use the following hypotheses:

$(H1)$ For some $\epsilon > 0$, Ω_ϵ is bounded.
$(H2)$ All optimal solutions of problem (P) satisfy the Mangasarian-Fromovitz qualification condition, namely, for any optimal solution x^*, there exist $h \in I\!R^n$

and $\rho > 0$ such that $\nabla f_i(x^*)^T h < -\rho$ for all $i \in I(x^*) \equiv \{i = 1, \ldots, m \mid f_i(x^*) = 0\}$.

A point $x \in I\!R^n$ will be called *interior* if $f_i(x) < 0$ for $i = 1, \ldots, m$. Note that the qualification condition implies that interior points exist.

We consider algorithms for solving (P) based on approximations of the exact penalty function (1). Each iteration solves an unconstrained problem obtained by adding a parameterized penalty function to the criterion. We now describe the class of penalty functions used in our treatment.

Smoothed penalty functions. Differentiable penalty functions are obtained by smoothing the function t^+. These penalties are based on functions $\theta : I\!R \to I\!R_+$, with the following properties:

(i) $\theta(\cdot)$ is convex, differentiable and has $\theta'(0) > 0$.
(ii) $\lim_{t\to-\infty} \theta(t) = 0$.
(iii) (Recession function) $\theta'_\infty(1) := \lim_{t\to+\infty} \theta(t)/t = 1$.
Notice that (ii) implies $\theta'_\infty(-1) := \lim_{t\to+\infty} \theta(-t)/t = 0$.

We assumed for simplicity that $\theta(\cdot)$ is differentiable, although this is not needed. The assumption that $\theta'(0) > 0$ can be easily replaced by the hypothesis that 0 is not a subgradient of θ at 0.

Consider a function $\theta(\cdot)$ defined as above, and a parameter $r \in (0, 1]$. We consider the function

$$t \in I\!R \mapsto \theta_r(t) = r\theta(t/r).$$

For any $r \in (0, 1]$, $\theta_r(\cdot)$ satisfies the properties (i)–(iii). Besides this, we have:

(iv) for $t \in I\!R$, $r \in (0, 1]$, $\theta_r^{'}(t) = \theta^{'}(t/r) \geq \theta^{'}(t)$.
(v) $\theta_r(\cdot)$ converges pointwise to t^+ when $r \to 0^+$.
(vi) For $r \in (0, 1]$, $\theta_r(t) \geq \theta(t) - \theta(0)$.

The property (iv) is straightforward, and it implies (vi). Let us prove (v): For t fixed,

$$\lim_{r\to0} \theta_r(t) = \lim_{r\to0} r\theta(t/r).$$

If $t < 0$, then clearly $\lim_{r\to0} \theta(t/r) = 0 = t^+$. Otherwise, change variables by $z \equiv t/r$:

$$\lim_{r\to0} r\theta(t/r) = \lim_{z\to\infty} t\theta(z)/z = t = t^+,$$

using property (iii). This completes the proof.

Methods based on functions of this class were studied in Auslender, Cominetti and Haddou [2], for the convex case. These functions are also described by Chen and Mangasarian [4, 5], who construct them as double integrals of probability density functions.

Examples. There are many examples of penalty functions in the literature, which are surveyed and classified in [2]. Let us list some examples:
(i) $\theta(t) = \log(1 + e^t)$.
(ii) $\theta(t) = (t + \sqrt{t^2 + 4})/2$.

(iii) The double integral of a unit pulse at the origin: let $u(\tau)$ be the unit step function applied at τ. Then we set $\theta''(t) = u(t-0.5) - u(t+0.5)$. This function is interesting because it only differs from an exact penalty in the interval $(-0.5, 0.5)$.
(iv) The double integral of a triangular pulse at the origin has the same properties as (iii), and is twice differentiable: set $\theta''(t) = (t-1)^+ - 2t^+ + (t+1)^+$.

Remarks. We are assuming that the penalties are smoothings of the exact penalty t^+. Assumption (iii) can obviously be written as $\theta'_\infty(1) = \nu > 0$. With $\nu = +\infty$, we would have coercive penalties, and it would be straightforward to adapt all the results in this paper to this case by replacing t^+ by $\nu\, t^+$, with the convention that $0 \times \infty = 0$. The case of coercive penalties is actually simpler: they are smoothings of the infinite barrier, with which any problem is solved in a single iteration.

Assumption (ii) can be replaced by "$\theta(\cdot)$ is bounded from below", since adding a constant to the penalties does not affect the algorithm.

The penalized function.

The penalized function is now constructed using $\theta_r(\cdot)$:

$$f_{\beta,r}(x) = f_0(x) + \beta \sum_{i=1}^{m} \theta_r(f_i(x)),$$

where $\beta \geq 1$ is the penalty weight. We shall denote

$$p_r(x) = \sum_{i=1}^{m} \theta_r(f_i(x)) = r \sum_{i=1}^{m} \theta\left(\frac{f_i(x)}{r}\right).$$

We have two parameters, r and β. The weight β changes the inclination of the penalty $\beta\theta_r(\cdot)$. The parameter r controls the precision of the smoothing.

The algorithm will play with β and r as follows: r decreases at all iterations (by, say, $r := r/2$). At infeasible points β increases (by, say, $\beta := 2\beta$); at feasible points β is not changed (See Fig. 1). The product βr never increases, which allows us to use the following property:

(vii) If βr is bounded, then for any interior point x

$$\lim_{r\to 0^+} \beta p_r(x) = 0. \tag{3}$$

Proof.

$$\lim_{r\to 0^+} \beta p_r(x) = \lim_{r\to 0^+} \beta r \sum_{i=1}^{m} \theta\left(\frac{f_i(x)}{r}\right).$$

For an interior point x, $f_i(x) < 0$ for $i = 1, \ldots, m$. Hence $\lim_{r\to 0^+} \theta\left(\frac{f_i(x)}{r}\right) = 0$, by (ii). The result follows from the boundedness of βr. □

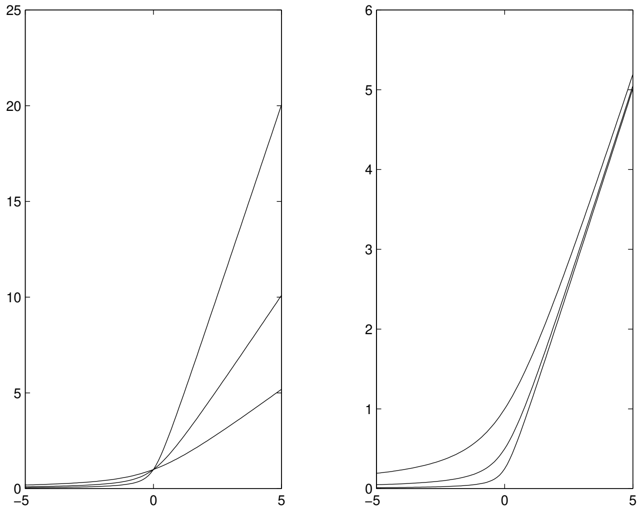

Figure 1: Behavior of $\beta\theta_r(\cdot)$. Left: β increases, r decreases; right: β fixed, r decreases.

3 The algorithm

At this point we assume that for any $\beta \geq 1$, $r \in (0, 1]$, $f_{\beta,r}(\cdot)$ has minimizers. This assumption will be discussed later.

Algorithm 3.1 *Data:* $\beta_0 = 1$, $r_0 = 1$.

For $k = 0, 1, \ldots,$

Minimization: compute

$$x^k \in \underset{x \in I\!R^n}{\operatorname{argmin}} \{f_0(x) + \beta_k \sum_{i=1}^{m} \theta_{r_k}(f_i(x))\} \tag{4}$$

$r_{k+1} := r_k/2$.

If x^k is feasible then $\beta_{k+1} := \beta_k$
Else $\beta_{k+1} := 2\beta_k$.

The following facts are trivially true for sequences generated by this algorithm:

- $r_k \to 0$ and hence by (v), $\theta_{r_k} \to t^+$ pointwise.
- $\beta_k r_k \leq 1$ and hence by (vii), $\beta_k p_{r_k}(\tilde{x}) \to 0$ for any interior point $\tilde{x}$.
- If (x^k) has an infeasible infinite subsequence then $\beta_k \to +\infty$.

This algorithm does not fall in the class studied by Auslender, Cominetti and Haddou [2]. They use the same penalized function, but require that $\beta_k r_k \to 0$ and $\beta_k \to +\infty$. Here we do not require this last limit: in fact, we will show that under our constraint qualification (β_k) is always bounded, i.e., β_k is equal to a constant $\bar{\beta}$ for large k, which means that $\beta_k p_{r_k}(\cdot)$ will converge pointwise to $\bar{\beta} \sum_{i=1}^m (f_i(\cdot))^+$, the exact penalty with a fixed parameter $\bar{\beta}$.

The minimizations.
Several conditions can be used to ensure that the minimization (4) is well defined in each step of the algorithm. We shall briefly discuss this for the convex case in Sect. 4. The discussion of an initial value for β is common to all penalty methods, and particularly to exact penalties. The hypothesis used by Zangwill [15] is precisely the one used above, that for large β a minimizer always exists. We shall use a weaker hypothesis. He comments that the initial β can be found by trial steps, and an algorithm capable of detecting that a problem has no minimizer. He also points out that if f_0 is coercive, then minimizers always exist because the penalties are non-negative.

Here we shall present what seems to be the weakest possible condition. We state the following assumption:

Hypothesis. Assume that for some $\beta \geq 1$, $r \leq 1$, $\operatorname{argmin}_{x \in I\!\!R^n} f_{\beta,r}(x) \neq \emptyset$.
We also assume that the minimization algorithm can decide that a problem has no minimizer. Note that if the penalized problem has no minimizer, then there are points arbitrarily far from the feasible set where $f_{\beta,r}(\cdot)$ is smaller than at any given feasible point.

Without loss of generality, we can simplify the treatment by assuming that the hypothesis above is satisfied for $\beta = 1$ and $r = 1$, by a suitable scaling of the problem ($f_0 \leftarrow f_0/\beta$, $p \leftarrow p_r$). We intend to prove that under this assumption, the minimization will be well defined for β sufficiently large and r sufficiently small. Instead of proving this directly, we shall change the minimization step in the algorithm to allow unsuccessful steps.

Let $\tilde{x}$ be an interior point, i. e., a point such that $f(\tilde{x}) < 0$. Such a point exists as a consequence of the Mangasarian-Fromovitz condition. The new minimization step will be:

Minimization.
Either compute $x^k \in \operatorname{argmin}_{x \in I\!\!R^n} \{f_0(x) + \beta_k \sum_{i=1}^m \theta_{r_k}(f_i(x))\}$
or decide that there exists $x^k \notin \Omega_\epsilon$ such that $f_{\beta_k,r_k}(x^k) \leq f_{\beta_k,r_k}(\tilde{x})$.

This step is well defined: if $\Gamma = \{x \in I\!\!R^n \mid f_{\beta_k,r_k}(x) \leq f_{\beta_k,r_k}(\tilde{x})\} \subset \Omega_\epsilon$, then Γ is bounded and a minimizer can be found. Otherwise, a descent algorithm is able to find a point in $\Gamma \backslash \Omega_\epsilon$. Note that the points x^k for unsuccessful iterations are not needed: they were defined only to fix the notation used in the proofs below. Actually, all that is needed is a way of establishing (guessing) that $f_{\beta_k,r_k}(\cdot)$ may be low at ϵ-infeasible points. This is easier than establishing that there is no minimizer: for instance, if an interior point is known, the internal search can be interrupted as soon as the condition in the modified step is satisfied for some given ϵ. If a lower bound L to the optimal value of the problem is known, than the minimization can also be stopped whenever $f_{\beta_k,r_k}(x^k) \leq L$ and x^k is ϵ-infeasible — it is easy to see that this condition implies the one in the modified minimization step.

From now on, assume that $(x^k), (r_k), (\beta_k)$ are sequences generated by the algorithm, keeping in mind that by hypothesis the first iteration is successful.

Lemma 3.2 *The sequence (x^k) is bounded, and all its accumulation points are feasible solutions.*

Proof. In the first iteration, $\beta_0 = r_0 = 1$ and by definition, for $k \in I\!N$

$$f_0(x^k) + p_1(x^k) \geq f_0(x^0) + p_1(x^0).$$

Using (vi), $\theta(t) \leq \theta_{r_k}(t) + \theta(0)$, for any $t \in I\!R$. Hence for $k \in I\!N$ and $i = 1, \dots, m$, $\theta_{r_k}(f_i(x^k)) \geq \theta(f_i(x^k)) - \theta(0)$, and it follows that

$$p_{r_k}(x^k) \geq p_1(x^k) - m\theta(0).$$

Adding these two inequalities,

$$f_0(x^k) + p_{r_k}(x^k) \geq f_0(x^0) + p_1(x^0) - m\theta(0). \tag{5}$$

Now consider the interior point $\tilde{x}$. By definition of x^k for $k \in I\!N$,

$$f_0(x^k) + \beta_k p_{r_k}(x^k) \leq f_0(\tilde{x}) + \beta_k p_{r_k}(\tilde{x}). \tag{6}$$

Subtracting (5) from (6) and collecting the constant terms,

$$(\beta_k - 1)p_{r_k}(x^k) \leq \beta_k p_{r_k}(\tilde{x}) + K,$$

Where K is a constant.
If (x^k) is feasible for large values of k, then we have nothing to prove. Assume then that this is not true. It follows that $\beta_k \to +\infty$. By (vii), $\beta_k p_{r_k}(\tilde{x}) \to 0$. Hence $p_{r_k}(x^k) \to 0$, and consequently $\theta_{r_k}(f_i(x^k)) \to 0$ for $i = 1, \dots, m$. This implies that $(f_i(x^k))^+ \to 0$ because for any $t > 0$, $\theta_{r_k}(t) \to t^+ = t > 0$.
— From the hypothesis (H1)[1], we see that (x^k) is bounded, otherwise for some infinite subsequence and some $i = 1, \dots, m$, $f_i(x^k) > \epsilon$.
— Let $\bar{x}$ be an accumulation point of (x^k). For $i = 1, \dots, m$, from the continuity of the functions and the fact that $(f_i(x^k))^+ \to 0$, we conclude that $(f_i(\bar{x}))^+ = 0$, completing the proof. □

Lemma 3.3 *Any accumulation point of (x^k) is an optimal solution of (P).*

Proof. Let $\bar{x}$ be an accumulation point of (x^k), and let $\mathcal{K} \subset I\!N$ be such that $x^k \xrightarrow{\mathcal{K}} \bar{x}$. By Lemma 3.2, $\bar{x}$ is feasible.
Consider now an arbitrary interior point $\tilde{x}$. By definition of x^k, for $k \in \mathcal{K}$,

$$f_0(x^k) \leq f_0(\tilde{x}) + \beta_k p_{r_k}(\tilde{x}) - \beta_k p_{r_k}(x^k).$$

[1]This is the only point in the proof where this hypothesis is used. This fact will be useful in the next section.

Taking limits for $k \in \mathcal{K}$, we have: $\beta_k p_{r_k}(\tilde{x}) \xrightarrow{\mathcal{K}} 0$, because of (vii), and always $\beta_k p_{r_k}(x^k) \geq 0$. It follows that

$$f_0(\bar{x}) \leq f_0(\tilde{x})$$

for any interior point $\tilde{x}$. By the qualification condition, all optimal solutions are in the closure of the set of interior points. This implies that $f_0(\bar{x}) \leq f_0(x^*)$ for any optimal solution x^*, completing the proof. □

Given a point $x \in I\!R^n$, let us define the sets

$$\begin{array}{rcl} I(x) & = & \{i = 1, \ldots, m \mid f_i(x) = 0\}, \\ I^+(x) & = & \{i = 1, \ldots, m \mid f_i(x) \geq 0\}. \end{array}$$

The following lemma is well known. We shall prove it for completeness.

Lemma 3.4 *If $x^k \to \bar{x}$, then for sufficiently large k, $I^+(x^k) \subset I(x^*)$.*

Proof. Assume that $x^k \to x^*$, and let $i \in \{1, \ldots, m\}$ be arbitrary. We must prove that if $i \notin I(x^*)$, then for large k, $i \notin I^+(x^k)$.
If $i \notin I(x^*)$, then $f_i(x^*) < 0$. Hence for large k, due to the continuity of $f_i(\cdot)$, $f_i(x^k) < 0$, or equivalently, $i \notin I^+(x^k)$, completing the proof. □

The following theorem is our main result. Using the Mangasarian-Fromovitz qualification condition, it states that only a finite number of iterates can be infeasible, so the slope of the penalty function remains bounded.

Theorem 3.5 *Let (x^k) be a sequence generated by the algorithm. Then there exists $\bar{k} \in I\!N$ such that for $k \geq \bar{k}$, x^k is feasible.*

Proof. Assume by contradiction that there exists an infinite set $\mathcal{K}_1$ such that for $k \in \mathcal{K}_1$, x^k is infeasible. Then by construction, $\beta_k \to \infty$. By Lemma 3.2, (x^k) is bounded, and hence $(x^k)_{k \in \mathcal{K}_1}$ has an accumulation point x^*, which is optimal by Lemma 3.3.
Let $\mathcal{K}_2 \subset \mathcal{K}_1$ be such that $x^k \xrightarrow{\mathcal{K}_2} x^*$.

Associating with each $k \in I\!N$ the set $I^+(x^k)$ defined above, the following three facts are true for large $k \in \mathcal{K}_2$:
(a) $I^+(x^k) \subset I(x^*)$, by Lemma 3.4.
If $I(x^*) = \emptyset$, then x^k is feasible, contradicting our assumption. Assume then that $I(x^*) \neq \emptyset$.
(b) For $i \in I(x^*)$, $f_i'(x^k, h) = \nabla f_i(x^k)^T h \leq -\rho/2$, where $h \in I\!R^n$ and $\rho > 0$ are given by the qualification condition (H2). This follows from the continuity of the gradients[2] and the fact that $x^k \xrightarrow{\mathcal{K}_2} x^*$.
(c) For $i \notin I(x^*)$, $f_i(x^k) \leq -\mu < 0$, where $\mu > 0$ is such that $-\mu > \max\{f_i(x^*) \mid i \notin I(x^*)\}$.

[2]This is the only point of the proof where differentiability is used. This fact will be useful in the next section.

Define the set $\mathcal{K} \subset \mathcal{K}_2$ as the set of iteration indices where these three facts are true.

Consider the directional derivatives along h. By definition of x^k, $f'_{\beta_k, r_k}(x^k, h) = 0$, and hence, using the fact that $\theta'_r(t) = \theta'(t/r)$,

$$f'_0(x^k, h) + \beta_k \sum_{i=1}^{m} \theta'(\frac{f_i(x^k)}{r_k}) f'_i(x^k, h) = 0.$$

Dividing by β_k and splitting the indices in the summation,

$$\frac{f'_0(x^k, h)}{\beta_k} + \sum_{i \notin I(x^*)} \theta'(\frac{f_i(x^k)}{r_k}) f'_i(x^k, h) = - \sum_{i \in I(x^*)} \theta'(\frac{f_i(x^k)}{r_k}) f'_i(x^k, h) \qquad (7)$$

Taking limits for $k \xrightarrow{\mathcal{K}} +\infty$, we have:
Left-hand side:

$\dfrac{f'_0(x^k, h)}{\beta_k} \to 0$, because $\beta_k \to +\infty$.
$\dfrac{f_i(x^k)}{r_k} \leq -\dfrac{\mu}{r_k}$ for $i \notin I(x^*)$, due to (c). Hence,
$\dfrac{f_i(x^k)}{r_k} \xrightarrow{\mathcal{K}} -\infty$, and $\theta'\left(\dfrac{f_i(x^k)}{r_k}\right) \xrightarrow{\mathcal{K}} 0$ by (ii).
We conclude that the left-hand side converges to 0 in $\mathcal{K}$.

Right-hand side:
For $i \in I(x^*)$, $f'_i(x^k, h) \leq -\rho/2$, by (b). Hence

$$\begin{aligned} - \sum_{i \in I(x^*)} \theta'\left(\frac{f_i(x^k)}{r_k}\right) f'_i(x^k, h) \quad &\geq \quad \frac{\rho}{2} \sum_{i \in I(x^*)} \theta'\left(\frac{f_i(x^k)}{r_k}\right) \\ &\geq \quad \frac{\rho}{2} \sum_{i \in I^+(x^k)} \theta'\left(\frac{f_i(x^k)}{r_k}\right) \\ &\geq \quad \frac{\rho}{2} \theta'(0)\ |I^+(x^k)|, \end{aligned}$$

because $I^+(x^k) \subset I(x^*)$ by (a), and for $i \in I^+(x^k)$, $f_i(x^k) \geq 0$ and $\theta(\cdot)$ is increasing. Since x^k is infeasible for $k \in \mathcal{K}$, $|I^+(x^k)| \geq 1$, and hence the right-hand side has a positive lower bound, contradicting (7) and completing the proof. $\square$

An example. Now we show a simple one-dimensional example:

$$\begin{array}{rrl} \text{minimize} & -2x & \\ \text{subject to} & (x + 0.25)^2 - 1/8 \leq & 0 \\ & x \leq & 0. \end{array}$$

We use the penalty given by $\theta(t) = \log(1 + \exp(t))$, and obtain the results shown in Figs. 2 and 3. This is a convex problem, and the interesting aspect of this example is that the iterates oscillate between feasible and infeasible: x^0 is feasible, x^1 and x^2 are infeasible, and from $k = 3$ on all iterates are feasible. If we had used the exact penalty, x^0 and x^1 would be infeasible, and x^2 would be optimal.

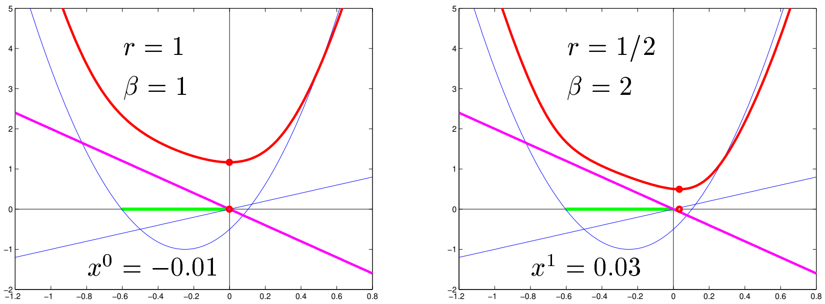

Figure 2: First two iterations. The thin lines represent constraints, and the thicker ones represent the objective and penalized functions.

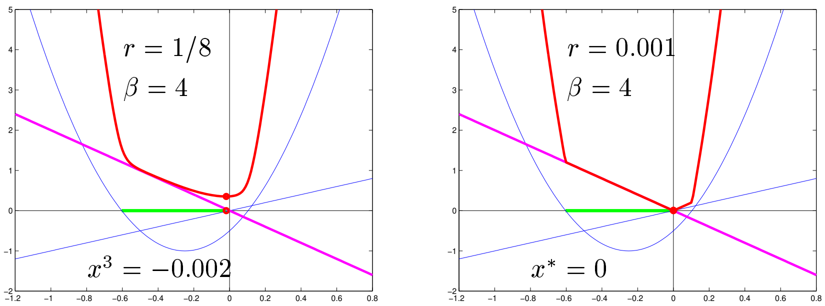

Figure 3: Fourth and tenth iterations.

Computational aspects.

Although the main scope of this paper is to extend to inexact penalties the main property of the exact penalty, we must comment a little on the computational aspects of the method. We assumed that global minimizers are found at all iterations: this is common to penalty methods in general, and this discussion does not belong in this paper. But ill-conditioning will naturally arise when we use Newton like methods with small values for the precision parameter r.

This fact has been known for a long time in connection with barrier methods. In that case, Nash and Sofer [7, 8] developed a stabilized method which can be easily extended to our penalties. We shall follow directly their approach. Consider the penalized function as above,

$$f_{\beta,r}(x) = f_0(x) + \beta r \sum_{i=1}^{m} \theta(\frac{f_i(x)}{r}).$$

Omitting the arguments for simplicity and computing derivatives,

$$\begin{aligned}
\nabla f_{\beta,r} &= \nabla f_0 + \beta \sum_{i=1}^{m} \theta'(\frac{f_i}{r}) \nabla f_i, \\
\nabla^2 f_{\beta,r} &= \nabla^2 f_0 + \beta \sum_{i=1}^{m} \theta'(\frac{f_i}{r}) \nabla^2 f^i + \frac{\beta}{r} \sum_{i=1}^{m} \theta''(\frac{f_i}{r}) \nabla f_i \nabla f_i^T.
\end{aligned}$$

We assume that all the Hessian matrices for the original problem are well-conditioned. We need some hypothesis on the behavior of $\theta''(\cdot)$. We shall make the following one, which is very reasonable and is satisfied by all existent penalty functions:

Hypothesis. There exists $T < 0$ such that $\theta''(\cdot)$ increases monotonically in $(-\infty, T)$.

Remember that $\theta''(\cdot)$ is non-negative, and that its integral satisfies

$$\int_{-\infty}^{T} \theta''(t) dt \leq \int_{-\infty}^{0} \theta''(t) dt = \theta'(0) < 1.$$

The following fact is now a simple exercise in Calculus:

Fact 3.6 *Under the hypothesis above,* $\lim_{t \to -\infty} t\, \theta''(t) = 0$.

Proof. By contradiction, assume that for some $\nu > 0$ there exists a sequence (t_i) such that $t_i \to -\infty$ and $-t_i \theta''(t_i) \geq \nu$. Choose t_i so that for all $i \in I\!N$, $t_{i+1} \leq 2t_i$. We have $t_i - t_{i+1} \geq -t_{i+1}/2$ and

$$\int_{t_{i+1}}^{t_i} \theta''(t) dt \geq \theta''(t_{i+1})(t_i - t_{i+1}) \geq -\frac{\nu(t_i - t_{i+1})}{t_{i+1}} \geq \nu/2.$$

It follows that $\int_{-\infty}^{t_1} \theta''(t) dt = +\infty$, contradicting the bound on the integral and completing the proof. □

Let us define for $i = 1, \ldots, m$ for given β and r,

$$\mu_i = \beta \theta'(\frac{f_i}{r}), \quad w_i = \beta \theta''(\frac{f_i}{r}),$$

and make the following assumptions, following our reference: the sequence (x^k) converges to an optimal solution x^*, with an active set $\mathcal{I} = \{i = 1, \ldots, m | f_i(x^*) = 0\}$ such that the active constraint gradients are linearly independent. We also assume that second order sufficient optimality conditions are satisfied at x^*.

Under these assumptions, the sequence (μ^k) associated with x^k, β^k, r_k by the expressions above converges to the unique optimal multiplier vector μ^* at x^*. Besides this, there exists a constant $\delta > 0$ such that $f_i(x^*) < -\delta$ for $i \notin \mathcal{I}$. Each iteration of the algorithm uses a *working set* of constraints, which will coincide with $\mathcal{I}$ for large k. The Hessian matrix will be decomposed in two parts, using the working set at each iteration. Here we assume that the working set is $\mathcal{I}$, and let us write the Hessian at a given iteration as

$$\nabla^2 f_{\beta,r} = \nabla^2 f_0 + \sum_{i=1}^{m} \mu_i \nabla^2 f^i + \sum_{i \notin \mathcal{I}} \frac{w_i}{r} \nabla f_i \nabla f_i^T + \frac{1}{r} \sum_{i \in \mathcal{I}} w_i \nabla f_i \nabla f_i^T. \quad (8)$$

The last term in this expression clearly leads to ill-conditioning. The first two terms are well conditioned. Let us examine the third term, and prove that it is also well-conditioned. For this, it is sufficient to prove the following fact:

Fact 3.7 *For all $i \notin \mathcal{I}$, $\lim_{k\to\infty} \frac{w_i^k}{r_k} = 0$.*

Proof. Consider $i \notin \mathcal{I}$. For large k (say, $k > k_1$), we know that $f_i(x^k) < -\delta/2$, and $\beta_k = \beta_{\bar{k}}$, where $\bar{k}$ is the iteration given by Theorem 3.5. By definition,

$$\frac{w_i^k}{r_k} = \frac{\beta_{\bar{k}}}{r_k} \theta'' \left(\frac{f_i(x^k)}{r_k} \right).$$

As $k \to \infty$, $r_k \to 0$ and $\frac{f_i(x^k)}{r_k} \leq \frac{-\delta}{2r_k} \to -\infty$. For k sufficiently large (say, $k > k_2 > k_1$), $-\delta/2r_k < T$ and then $\theta''(\cdot)$ is increasing in $(-\infty, -\delta/2r_k)$. Hence

$$\theta'' \left(\frac{f_i(x^k)}{r_k} \right) \leq \theta'' \left(\frac{-\delta}{2r_k} \right).$$

It follows that

$$\frac{w_i^k}{r_k} \leq \frac{\beta_{\bar{k}}}{2\delta} \frac{\delta}{2r_k} \theta'' \left(\frac{-\delta}{2r_k} \right).$$

The conclusion now follows from the fact that $\lim_{t\to-\infty} \theta''(t) = 0$, completing the proof. $\square$

Now, (8) can be written as

$$\nabla^2 f_{\beta,r} = H + \frac{1}{r} B,$$

where where H and B are well conditioned as in Nash and Sofer [7]. Now their procedure can be followed step by step: let N be the constraint Jacobian. The Newton directions are written as

$$p \approx p_1 + r p_2,$$

where p_1 and p_2 belong respectively to the null space of $N_{\mathcal{I}}$ and to the range space of $N_{\mathcal{I}}^T$. These directions are obtained by stable computations and result in an $O(r)$ approximation of the Newton direction.

4 The convex case

Now we study the problem (P) for the case in which $f_i(\cdot)$, $i = 0, 1, \ldots, m$, are non-smooth closed proper convex functions satisfying the following hypotheses:
(A1) The optimal set S of the problem (P) is nonempty and compact.
(A2) The Slater condition in satisfied, that is, there exists $\bar{x} \in R^n$ such that $f_i(\bar{x}) < 0$ for $i = 1, \ldots, m$.

The Slater condition is equivalent to the generalized Mangasarian-Fromovitz condition for the convex problem (see Pappalardo [10]), namely, given an optimal solution x^*, there exist $h \in R^n$ and $\rho > 0$ such that $f_i'(x^*, h) < -\rho$ for $i \in I(x^*) = \{i = 1, \ldots, m \mid f_i(x^*) = 0\}$.

We shall now study the Algorithm 3.1 applied to the convex problem with these hypotheses. Here we do not assume that the feasible set Ω is compact, but the assumption that the optimal set is compact has similar effects: all the action happens in a compact subset of Ω, as we show in the following lemma:

Lemma 4.1 *Let x^* be an optimal solution for (P), and set $v = f_0(x^*)$. Define $\alpha = m\theta(0)$. We have:*
(i) The set $\bar{\Omega} = \{x \in \Omega \mid f_0(x) \leq v + \alpha\}$ is compact.
(ii) Let $\beta \geq 1$ and $r \leq 1$ be such that $\beta r \leq 1$. Then for any $x \in \Omega \backslash \bar{\Omega}$, $f_{\beta,r}(x) > f_{\beta,r}(x^*)$.

Proof. (i) is a well known consequence of the compactness of the optimal set. Here is how it is proved: if $\bar{\Omega}$ is unbounded, then it has a recession direction h. Along h all functions f_i, $i = 0, 1, \ldots, m$ must be non-increasing. This contradicts the boundedness of the optimal set.
(ii) Since $x^* \in \Omega$, $f_i(x^*) \leq 0$ for $i = 1, \ldots, m$, and then $\theta_r(f_i(x^*)) \leq r\theta(0)$. Adding these inequalities we get $p_r(x^*) \leq mr\theta(0)$ and $\beta p_r(x^*) \leq \alpha$, because $\beta r \leq 1$. We conclude that $f_{\beta,r}(x^*) \leq v + \alpha$.

Now consider $x \in \Omega \backslash \bar{\Omega}$. Since $p_r(x) \geq 0$, we have

$$f_{\beta,r}(x) \geq f_0(x) > v + \alpha \geq f_{\beta,r}(x^*),$$

completing the proof. $\square$

So, if any iteration of the algorithm produces a feasible point x^k, then $x^k \in \bar{\Omega}$.

Now Lemmas 3.2 and 3.3 can be reproduced for the convex case with $\epsilon = 0$, with identical proofs, by using the set $\bar{\Omega}$ in the role occupied by Ω_ϵ in the proof of 3.2.

Instead of developing these proofs, we shall show how the same results can be derived directly from the work of Auslender, Cominetti and Haddou [2]. Theorem 2.2 of this reference, when adapted to our notation, says the following:

Theorem 4.2 *Consider the sequences (x^k), (r_k) and (β_k) generated by an application of the algorithm to a convex programming problem satisfying hypotheses (A1) and (A2), and assume that $\beta_k \to \infty$. Then for k sufficiently large, the sets $S_k = \operatorname{argmin}_{x \in I\!R^n} f_{\beta_k, r_k}(x)$ are compact. Furthermore the selections $x^k \in S_k$ stay bounded as $k \to \infty$ and their accumulation points belong to the optimal set of (P).*

The complete result for our algorithm (in which (β_k) is not necessarily unbounded) is the following:

Lemma 4.3 *Consider the sequences* (x^k), (r_k) *and* (β_k) *generated by an application of the algorithm to a convex programming problem satisfying hypotheses (A1) and (A2). Then* (x^k) *is bounded and its accumulation points belong to the optimal set of (P).*

Proof. Consider an application of the algorithm. If $\beta_k \to \infty$, then the result follows from Theorem 4.2. Otherwise, there exists $\bar{k}$ such that $x^k \in \Omega$ for $k \geq \bar{k}$, by construction. Lemma 4.1 ensures that these points belong to the compact set $\bar{\Omega}$, proving that the sequence is bounded. Lemma 3.3 (identically proved for the convex case) asserts that in this case also the accumulation points are optimal solutions, completing the proof. $\square$

We can finally establish our main result also for the convex case:

Theorem 4.4 *Let* (x^k) *be a sequence generated by the algorithm. Then there exists* $\bar{k} \in I\!N$ *such that for* $k \geq \bar{k}$, x^k *is feasible.*

Proof. The proof is similar to that of Theorem 3.5. The only point in that proof where differentiability was used was in item (b) which we now prove for the convex case:
(b) For k sufficiently large and $i \in I(x^*)$, $f_i'(x^k, h) \leq -\rho/2$, where h and ρ are given by the Mangasarian-Fromovitz condition. This follows from the following facts: by the qualification condition, $f_i'(x^*, h) \leq -\rho$ for $i \in I(x^*)$; by Corollary 6.2.5 in [9], the directional derivatives $x \in I\!R^n \mapsto f_i'(x, h)$ are upper semi-continuous. $\square$

Convergence of the sequence. Cominetti [6] studies the sequence (x^k) in the convex case and proves that under certain additional hypotheses, the sequence converges to a particular solution of the problem, known as the *θ-center*. Using the same hypotheses in our treatment, the same result can be proved for the two-parameter approach. To prove this fact, one should see that Cominetti only uses the condition $\beta \to \infty$ for proving the convergence to optimizers, which we proved in Lemma 4.3. Given this fact, his treatment applies directly to the two-parameter case with bounded β, with the same results.

For the non-convex case, we do not know any results about the convergence of the sequence to particular solutions.

References

[1] Auslender, A. (1997): Penalty and barrier methods: unified framework. Technical report, Laboratoire D'econometrie de L'ecole Polytechnique, Paris

[2] Auslender, A., Cominetti, R., Haddou, M. (1997): Asymptotic analysis for penalty and barrier methods in convex and linear programming. Mathematics of Operations Research **22**, 43–62

[3] Ben-Tal, A., Teboulle, M. (1989): A smoothing technique for non-differentiable optimization problems. in: Lecture Notes in Mathematics, Springer Verlag, Berlin, West–Germany, vol. 1405, 1–11

[4] Chen, C., Mangasarian, O. L. (1995): Smoothing methods for convex inequalities and linear complementary problems. Mathematical Programming **71**, 1–112

[5] Chen, C., Mangasarian, O. L. (1996): A class of smoothing functions for non-linear and mixed complementary problems. Computational Optimization and Applications **5**, 97–138

[6] Cominetti, R. (1999): Nonlinear averages and convergence of penalty trajectories in convex optimization. in: Lecture Notes in Economics and Math. Systems , Springer Verlag, Berlin, West–Germany, vol. 477, 65–78.

[7] Nash, S., Sofer, A. (1993): A barrier method for large-scale constrained optimization. ORSA Journal on Computing **5**, 40–53.

[8] Nash, S., Sofer, A. (1996): Linear and nonlinear programming. McGraw-Hill, New York.

[9] Hiriart-Urruty, J.-B., Lemarechal, C. (1993): Convex Analysis and Minimization Algorithms I. Springer-Verlag, New York.

[10] Pappalardo, M. (1967): Error bounds for generalized Lagrange multiplier in locally Lipschitz programming. Journal of Optimization Theory Applications **17**, 34–47.

[11] Pinar, M., Zenios, S. (1994): On smoothing exact penalty functions for convex constrained optimization. SIAM Journal on Optimization **4**, 486–511.

[12] Truemper, K. (1975): Note on finite convergence of exterior penalty functions. Management Science **21**, n.5, 600–606.

[13] Xavier, A. E. (1992): Hyperbolic Penalization. Ph.D. thesis, COPPE - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil (in Portuguese).

[14] Zang, I. (1980): A smoothing-out technique for min-max optimization. Mathematical Programming **19**, 61–77.

[15] Zangwill, W. I. (1967): Non-linear programming via penalty functions. Management Science **13**, 344–358.