# CONVERGENCE OF LINEARIZED $\ell_q$ PENALTY METHODS FOR OPTIMIZATION WITH NONLINEAR EQUALITIES

Lahcen El BOURKHISSI[1], Ion NECOARA[2]

*In this paper, we consider nonconvex optimization problems with nonlinear equality constraints. We assume that the objective function and the func-tional constraints are locally smooth. To solve this problem, we introduce a linearized $\ell_q$ penalty based method, where $q \in (1, 2]$ is the parameter defining the norm used in the construction of the penalty function. Our method involves linearizing the objective function and functional constraints in a Gauss-Newton fashion at the current iteration in the penalty formulation and introduces a quadratic regularization. This approach yields an easily solvable subproblem, whose solution becomes the next iterate. By us-ing a novel dynamic rule for the choice of the regularization parameter, we establish that the iterates of our method converge to an $\epsilon$-first-order solution in $\mathcal{O}(1/\epsilon^{2+(q-1)/q})$ outer iterations. Finally, we put theory into practice and evaluate the performance of the proposed algorithm by making numerical comparisons with existing methods from literature.*

**Keywords:** Nonconvex optimization, $\ell_q$ penalty, convergence analysis.

**MSC2020:** 68Q25 . 90C06 . 90C30.

## 1. Introduction

Penalty methods have played a central role in theoretical and numerical optimization, with their historical origins dating back to at least [4]. Extensive research has explored the applications of penalty methods to a wide range of problems, as shown by works such as [1, 2, 8–11, 13, 14], among others. For example, [9] studies a polyvalent class of penalty functions, taking the form of $\|\cdot\|_q^q$, where $q > 0$, for general constrained problems. This study establishes bounds that measure the closeness of the penalty solution to the solution of original problem as a function of the penalty parameter $\rho$. In particular, under strict Mangasarian–Fromovitz constraint qualification and second-order sufficiency, a bound of the form $\mathcal{O}(\frac{1}{\rho^{q-1}})$ is derived and it becomes zero for $q \in (0, 1]$ provided that $\rho$ is sufficiently large. Paper [2] introduces an algorithm based on a Lipschitz penalty function, with dynamic quadratic regularization. This method reduces the size of a first-order criticality measure to a specified accuracy threshold $\epsilon$, in a maximum of $\mathcal{O}(1/\epsilon^2)$ functions evaluations, provided we are close to feasibility. In an alternative context, [10] focuses on the use of a quadratic penalty method to handle nonconvex composite problems with linear constraints proving convergence to an $\epsilon$-critical point in $\mathcal{O}(1/\epsilon^3)$ accelerated composite gradient steps. Furthermore, the work in [11] introduces an inexact proximal-point penalty method for solving general problems with nonconvex objective and constraints, proving convergence to an

[1] Department of Automatic Control and Systems Engineering, National University of Science and Technology Politehnica Bucharest, e-mail: lel@stud.acs.upb.ro

[2] Department of Automatic Control and Systems Engineering, National University of Science and Technology Politehnica Bucharest, 060042 Bucharest and Gheorghe Mihoc-Caius Iacob Institute of Mathematical Statistics and Applied Mathematics of the Romanian Academy, 050711 Bucharest, Romania. e-mail: ion.necoara@upb.ro

$\epsilon$-critical point within $\mathcal{O}(1/\epsilon^3)$ of functions evaluations. The result can be refined, reaching a complexity of $\mathcal{O}(1/\epsilon^{2.5})$ for nonconvex objective and convex constraints. Finally, in our previous work [6], we developed a quadratic penalty method for solving smooth nonconvex optimization problems with nonlinear constraints, where we linearize both the objective and functional constraints within the quadratic penalty function in a Gauss-Newton fashion. We established a complexity bound of $\mathcal{O}(1/\epsilon^{2.5})$ of functions evaluations. In this context, our current method is inspired by [9] and generalizes our previous work [6], as it considers an $\ell_q$ penalty approach with $q \in (1, 2]$, bridging the gap between two extremes: the exact penalty method based on $\ell_1$ norm, where the penalty parameter $\rho$ is finite but the subproblem lacks differentiability, and the quadratic penalty method, where the subproblem is smooth but the penalty parameter $\rho$ must be of the order inverse of the desired accuracy.

*Contributions:* Our approach, referred to as the linearized $\ell_q$ penalty method (qLP), effectively addresses some of the limitations of the previous studies. Notably, in [2], the subproblem is non-differentiable, due to the use of a Lipschitz penalty function, while in [10], the framework is limited to handling only linear constraints. Hence, our main contributions are as follows:

(i) At each iteration, we linearize, in a Gauss-Newton fashion, both the cost function and the nonlinear functional constraints within the $\ell_q$ penalty function, where $q \in (1, 2]$ is the parameter defining the norm used in the construction of the penalty function, and add a dynamic regularization term. This results in a new algorithm, called *the linearized $\ell_q$ penalty method* (qLP). Notably, our method considerably simplifies the computational cost of the new iterate, since each iteration reduces to minimizing a strongly convex differentiable function with Holder continuous gradient, thus making the subproblem easily solvable with e.g., an accelerated first-order scheme.

(ii) We provide rigorous proofs of global asymptotic convergence, guaranteeing that the iterates eventually converge to a critical point of the $\ell_q$ penalty function, which implies, for an appropriate choice of $\rho$, a $(0, \epsilon)$-first-order solution of the original problem. Furthermore, our method guarantees convergence to an $\epsilon$-first-order solution of the original problem in $\mathcal{O}(1/\epsilon^{2+(q-1)/q})$ outer iterations, thus improving the existing bounds.

## 2. Problem formulation and preliminaries

In this paper, we consider the following nonconvex optimization problem:

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad F(x) = 0, \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $F(x) \triangleq (f_1(x), ..., f_m(x))^T$, with $f_i : \mathbb{R}^n \to \mathbb{R}$ for all $i = 1 : m$. We assume that the functions $f, f_i \in \mathcal{C}^1$ for all $i = 1 : m$, where $f$ can be nonconvex and $F$ nonlinear. Moreover, we assume that the problem is well-posed i.e., the feasible set is nonempty and the optimal value is finite. Before introducing the main assumptions for our analysis, we would like to clarify some notations. We use $\|\cdot\|_q^q$, where $q \in (1, 2]$, to denote the $q$-norm of a vector in $\mathbb{R}^n$. For simplicity, $\|\cdot\|$ denotes the Euclidean norm of a vector or the spectral norm of a matrix. For a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, we denote by $\nabla f(x) \in \mathbb{R}^n$ its gradient at a point $x$. Moreover, we say that $x^*$ is a *critical* point of $f$ if $\nabla f(x^*) = 0$. For a differentiable vector function $F : \mathbb{R}^n \to \mathbb{R}^m$, we denote its Jacobian at a given point $x$ by $J_F(x) \in \mathbb{R}^{m \times n}$. Furthermore, for a vector $y = (y_1, \ldots, y_m)^T \in \mathbb{R}^m$ and a positive value $a$, we denote $|y|^a = (|y_1|^a, \ldots, |y_m|^a)^T$ and $\text{sign}(y) \circ |y|^a = (\text{sign}(y_1)|y_1|^a, \cdots, \text{sign}(y_m)|y_m|^a)^T \in \mathbb{R}^m$. We further introduce the notations:

$$l_f(x; \bar{x}) \triangleq f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle, \;\; l_F(x; \bar{x}) \triangleq F(\bar{x}) + J_F(\bar{x})(x - \bar{x}) \quad \forall x, \bar{x}.$$

Let us now present the main assumptions considered for problem (1):

**Assumption 1.** *Assume that $f(x)$ has compact level sets, i.e., for any $\alpha \in \mathbb{R}$, the following set is either empty or compact:*

$$\mathcal{S}_\alpha^0 \triangleq \{x : f(x) \leq \alpha\}.$$

**Assumption 2.** *Given a compact set $\mathcal{S} \subseteq \mathbb{R}^n$, there exist positive constants $M_f, M_F, \sigma, L_f, L_F$ such that $f$ and $F$ satisfy the following conditions:*

*(i)* $\|\nabla f(x)\| \leq M_f, \quad \|\nabla f(x) - \nabla f(y)\| \leq L_f\|x-y\| \quad \forall x, y \in \mathcal{S}.$
*(ii)* $\|J_F(x)\| \leq M_F, \quad \|J_F(x) - J_F(y)\| \leq L_F\|x-y\| \quad \forall x, y \in \mathcal{S}.$
*(iii) Problem* (1) *satisfies Linear Independance Constraint Qualification (LICQ) condition for all* $x \in \mathcal{S}$.

Note that these assumptions are standard in the nonconvex optimization literature, in particular in penalty type methods, see e.g., [2, 3, 6, 15]. In fact, these assumptions are not restrictive because they need to hold only locally. Indeed, large classes of problems satisfy these assumptions as discussed below.

**Remark 2.1.** *Assumption 1 holds e.g., when $f(x)$ is coercive (in particular, $f(x)$ is strongly convex), or $f(x)$ is bounded from bellow. Assumption 2 allows general classes of problems. In particular, conditions (i) hold if $f(x)$ is differentiable and $\nabla f(x)$ is locally Lipschitz continuous on a neighborhood of $\mathcal{S}$. Conditions (ii) hold when $F(x)$ is differentiable on a neighborhood of $\mathcal{S}$ and $J_F(x)$ is locally Lipschitz continuous on $\mathcal{S}$. Finally, the LICQ assumption guarantees the existence of dual multipliers and is commonly used in nonconvex optimization, see e.g., [13, 15]. For equality constraints, LICQ holds on a compact set $\mathcal{S}$ if the smallest singular value of the Jacobian matrix of the functional constraints remains strictly positive on $\mathcal{S}$.*

The following lemma is an immediate consequence of Assumption 1.

**Lemma 2.1.** *If Assumption 1 holds, then for any $\rho \geq 0$, we have:*

$$\underline{P} \triangleq \inf_{x\in\mathbb{R}^n}\{f(x) + \frac{\rho}{q}\|F(x)\|_q^q\} > -\infty \quad \text{and} \quad \bar{f} \triangleq \inf_{x\in\mathbb{R}^n}\{f(x)\} > -\infty. \tag{2}$$

We are interested in (approximate) first-order (also called KKT) solutions of optimization problem (1). Hence, let us introduce the following definitions:

**Definition 2.1.** *[First-order solution and $\epsilon$-first-order solution of* (1)*] The vector $x^*$ is said to be a first-order solution of* (1) *if $\exists \lambda^* \in \mathbb{R}^m$ such that:*

$$\nabla f(x^*) + J_F(x^*)^T\lambda^* = 0 \quad \text{and} \quad F(x^*) = 0.$$

*Moreover, $\hat{x}$ is an $(\epsilon_1, \epsilon_2)$-first-order solution of* (1) *if $\exists \hat{\lambda} \in \mathbb{R}^m$ and $\kappa_1, \kappa_2 > 0$:*

$$\|\nabla f(\hat{x}) + J_F(\hat{x})^T\hat{\lambda}\| \leq \kappa_1\epsilon_1 \quad \text{and} \quad \|F(\hat{x})\| \leq \kappa_2\epsilon_2.$$

*If $\epsilon_1 = \epsilon_2$, we refer to $\hat{x}$ as an $\epsilon$-first-order solution in the previous definition.*

## 3. A linearized $\ell_q$ penalty method

In this section, we propose a new algorithm for solving nonconvex problem (1) using the $\ell_q$ penalty framework. Let us first introduce few notations. The penalty function associated with the problem (1) is

$$\mathcal{P}_\rho^q(x) = f(x) + \frac{\rho}{q}\|F(x)\|_q^q, \tag{3}$$

where $q \in (1, 2]$. This penalty function, $\mathcal{P}_\rho^q$, is differentiable and its gradient is:

$$\nabla\mathcal{P}_\rho^q(x) = \nabla f(x) + J_F(x)^T\left(\rho\,\mathrm{sign}(F(x)) \circ |F(x)|^{q-1}\right).$$

The next lemma states that function $\mathrm{sign}(\cdot)|\cdot|^\nu$, where $\nu \in (0, 1]$, is Holder continuous:

**Lemma 3.1.** *[Holder] Let $\nu \in (0,1]$. Then, we have:*

$$|\operatorname{sign}(x)|x|^{\nu} - \operatorname{sign}(y)|y|^{\nu}| \leq 3|x-y|^{\nu} \quad \forall x, y \in \mathbb{R}.$$

From the previous lemma, using properties of norms, one can conclude that the function $v \mapsto \frac{1}{q}\|v\|_q^q$, with $q \in (1,2]$ and $v \in \mathbb{R}^m$, has the gradient Holder continuous w.r.t. Euclidean norm $\|\cdot\|$, i.e.,:

$$\|\operatorname{sign}(v) \circ |v|^{q-1} - \operatorname{sign}(w) \circ |w|^{q-1}\| \leq 3 \times m^{\frac{2-q}{2}} \|v-w\|^{q-1} \quad \forall v, w \in \mathbb{R}^m. \tag{4}$$

This implies the following inequality [5]:

$$\tfrac{1}{q}\|v\|_q^q \leq \tfrac{1}{q}\|w\|_q^q + \langle \operatorname{sign}(w) \circ |w|^{q-1}, v-w\rangle + \tfrac{3 \times m^{\frac{2-q}{2}}}{q}\|v-w\|^q, \quad \forall v, w \in \mathbb{R}^m. \tag{5}$$

Further, let us denote the following function derived from linearization in a Gauss-Newton fashion of the objective function and the functional constraints, at a given point $\bar{x}$, in the penalty function:

$$\bar{\mathcal{P}}_\rho^q(x;\bar{x}) = f(\bar{x}) + \langle \nabla f(\bar{x}), x-\bar{x}\rangle + \frac{\rho}{q}\|F(\bar{x}) + J_F(\bar{x})(x-\bar{x})\|_q^q.$$

Note that the function $\bar{\mathcal{P}}_\rho^q(\cdot;\bar{x})$ is always convex since $q > 1$. Let us also introduce the following criticality measure for the penalty function $\mathcal{P}_\rho^q$, for conducting our analysis, inspired by [2]. For $0 < r \leq 1$, we define:

$$\Psi_r(x) \triangleq \bar{\mathcal{P}}_\rho^q(x;x) - \min_{\|y-x\| \leq r} \bar{\mathcal{P}}_\rho^q(y;x) = \mathcal{P}_\rho^q(x) - \min_{\|y-x\| \leq r} \bar{\mathcal{P}}_\rho^q(y;x). \tag{6}$$

In particular, following [16], $\Psi_r(x)$ is continuous for all $x$, and $x^*$ is a critical point of penalty function $\mathcal{P}_\rho^q$ if $\Psi_r(x^*) = 0$. The next lemma states the above claim, see also Lemma 2.1 in [16].

**Lemma 3.2.** *Let $q \in (1,2]$, $0 < r \leq 1$ and $\Psi_r(\cdot)$ be as in (6). Then, $\Psi_r(x) \geq 0$ and $\Psi_r(x) = 0$ if and only if $x$ is a critical point of the penalty function $\mathcal{P}_\rho^q$. Moreover, $\Psi_r(\cdot)$ is continuous.*

Let us also introduce the following pseudo-criticality measure:

$$\bar{\Psi}(x,\beta) \triangleq \bar{\mathcal{P}}_\rho^q(x;x) - \min_{y \in \mathbb{R}^n} \left\{ \bar{\mathcal{P}}_\rho^q(y;x) + \frac{\beta}{2}\|y-x\|^2 \right\}. \tag{7}$$

We establish later a relation between these two criticality measures.

To solve the optimization problem (1) we propose the following *Linearized $\ell_q$ penalty* (qLP) algorithm, where we linearize the objective function and the functional constraints, in a Gauss-Newton fashion, within the penalty function at the current iterate and add an *adaptive* quadratic regularization. *To the best of our knowledge qLP algorithm is new and its*

**Algorithm 3.1** Linearized $\ell_q$ penalty (qLP) method

1: **Initialization:** $x_0, \rho > 0$, and $\underline{\beta} \geq 1$.
2: $k \leftarrow 0$
3: **while** stopping criterion is not satisfied **do**
4: generate a proximal parameter $\beta_{k+1} \geq \underline{\beta}$ such that
5: $x_{k+1} \leftarrow \arg\min_{x \in \mathbb{R}^n} \bar{\mathcal{P}}_\rho^q(x;x_k) + \frac{\beta_{k+1}}{2}\|x-x_k\|^2$ satisfies the descent:

$$\mathcal{P}_\rho^q(x_{k+1}) \leq \bar{\mathcal{P}}_\rho^q(x_{k+1};x_k) + \frac{\beta_{k+1}}{2}\|x_{k+1}-x_k\|^2. \tag{8}$$

6: $k \leftarrow k+1$
7: **end while**

*convergence behavior has not been analyzed before in the literature.* Note that the objective function in the subproblem of Step 5 of Algorithm 3.1 is always strongly convex since the convex function $\bar{\mathcal{P}}_\rho^q(\cdot; x_k)$ is regularized with a quadratic term. Moreover, it has a locally Holder continuous gradient with Holder constant proportial to $\rho$ and exponent $q-1$ (see Lemma 3.1). Therefore, finding a solution of the subproblem in Step 5 is easy as there are efficient methods that can deal with this type of problems (see e.g., [5]). In the sequel, we denote:

$$\Delta x_k = x_k - x_{k-1} \quad \forall k \geq 1.$$

Let us show that we can always choose an adaptive regularization parameter $\beta_{k+1}$ guaranteeing the descent property (8). Indeed, since $f$ and $F$ are smooth functions, if one chooses adaptively (i.e., depending on the current iterate $x_k$):

$$\beta_{k+1} \geq L_f + \left(3 \times m^{\frac{(2-q)(q-1)}{2q}} \times 2^{2-q}\right)^{\frac{1}{q}} q^{\frac{q-1}{q}} \rho^{\frac{1}{q}} L_F \left(\mathcal{P}_\rho^q(x_k) - \bar{f}\right)^{\frac{q-1}{q}}, \tag{9}$$

then the descent property (8) follows, as established in the following lemma.

**Lemma 3.3.** *[Existence of $\beta_{k+1}$] If the sequence $\{x_k\}_{k\geq 0}$ generated by Algorithm 3.1 is in some compact set $\mathcal{S}$ on which Assumptions 1 and 2 hold and we choose $\beta_{k+1}$ as in* (9)*, then the descent property* (8) *holds. Consequently, the following decrease condition is also satisfied:*

$$\mathcal{P}_\rho^q(x_{k+1}) \leq \mathcal{P}_\rho^q(x_k) - \frac{\beta_{k+1}}{2}\|x_{k+1} - x_k\|^2. \tag{10}$$

*Proof.* See Appendix. □

From Lemma 3.3 it follows that when using a backtracking scheme, with a geometrically increasing parameter $\mu > 1$, $\beta_{k+1}$ can be always upper bounded as:

$$\bar{\beta} \triangleq \sup_{k\geq 1} \beta_k \leq \mu \left(L_f + \left(3 \times m^{\left(\frac{(2-q)(q-1)}{2q}\right)} \times 2^{2-q}\right)^{\frac{1}{q}} q^{\frac{q-1}{q}} \rho^{\frac{1}{q}} L_F (\bar{P} - \bar{f})^{\frac{q-1}{q}}\right). \tag{11}$$

Next lemma, whose proof is straightforward, guarantees the following for $x_{k+1}$.

**Lemma 3.4.** *Let Assumption 2 hold on a compact set $\mathcal{S}$ and assume that the sequence $\{x_k\}_{k\geq 0}$ generated by Algorithm 3.1 is in $\mathcal{S}$. Then, we have:*

$$r_k \triangleq \frac{\mathcal{P}_\rho^q(x_k) - \mathcal{P}_\rho^q(x_{k+1})}{\bar{\Psi}(x_k, \beta_{k+1})} \geq 1. \tag{12}$$

Due to space limitation, the proofs of some lemmas are omitted, but the details can be found in [7].

## 4. Convergence analysis

In this section, we derive the efficiency of qLP algorithm (ALgorithm 3.1) for finding an $\epsilon$-first-order solution of the problem (1). In the sequel, we are using the $\ell_q$ penalty function, $\mathcal{P}_\rho^q$, as a Lyapunov function and denote by:

$$P_k = \mathcal{P}_\rho^q(x_k) \quad \forall k \geq 0. \tag{13}$$

It is clear, from Lemma 3.3, that $\{P_k\}_{k\geq 0}$ is decreasing and later we prove that it is bounded from bellow. The following lemma shows that if the sequence $\{x_k\}_{k\geq 0}$ generated by Algorithm 3.1 is bounded, then the Lyapunov sequence $\{P_k\}_{k\geq 0}$ is also bounded. Its proof is based on the decrease (10).

**Lemma 4.1.** *Consider Algorithm 3.1 and let* $\{P_k\}_{k\geq 0}$ *as defined in* (13). *If the sequence* $\{x_k\}_{k\geq 0}$ *generated by Algorithm 3.1 is in some compact set* $\mathcal{S}$ *on which Assumptions 1 and 2 hold. Then, there exists* $\bar{P} < \infty$ *such that:*

$$\underline{P} \leq P_k \leq \bar{P} \quad \forall k \geq 0, \tag{14}$$

*where* $\underline{P}$ *is defined in* (2).

Let us now investigate the computational complexity of Algorithm 3.1 for generating an $\epsilon$-first-order solution. First, we relate the model decrease $\bar{\Psi}(x_k, \beta_{k+1})$ to the optimality measure $\Psi_r(x_k)$ in (6).

**Lemma 4.2.** *Let* $0 < r \leq 1$ *and let* $\Psi_r(\cdot)$ *be defined by* (6). *If the sequence,* $\{x_k\}_{k\geq 0}$ *generated by Algorithm 3.1 is in some compact set* $\mathcal{S}$ *on which Assumption 2 holds, then:*

$$\bar{\Psi}(x_k, \beta_{k+1}) \geq \frac{1}{2} \min\left(1, \frac{\Psi_r(x_k)}{\beta_{k+1} r^2}\right) \Psi_r(x_k). \tag{15}$$

*Proof.* See appendix. □

Lemma 4.2 indicates that $r_k$ in (12) is well-defined whenever the current iteration is not a first-order critical point, i.e., $\Psi_r(x_k) \neq 0$. Let $0 < \epsilon \leq 1$. The following theorem demonstrates that after $K \geq \mathcal{O}(\frac{\rho^{\frac{1}{q}}}{\epsilon^2})$ iterations of Algorithm 3.1, the criticality measure $\Psi_\epsilon(x_k) \leq \epsilon^2$.

**Lemma 4.3.** *Consider Algorithm 3.1 and let* $\{P_k\}_{k\geq 0}$ *be defined as in* (13). *If the sequence* $\{x_k\}_{k\geq 0}$, *generated by Algorithm 3.1 is in some compact set* $\mathcal{S}$ *on which Assumptions 1 and 2 hold. Then, for any* $\epsilon \in (0, 1]$, *after*

$$K = \left\lceil 2\bar{\beta}\left(\bar{P} - \underline{P}\right)\epsilon^{-2}\right\rceil = \mathcal{O}\left(\frac{\rho^{\frac{1}{q}}}{\epsilon^2}\right)$$

*iterations of Algorithm 3.1, we obtain* $\Psi_\epsilon(x_k) \leq \epsilon^2$.

*Proof.* See Appendix. □

In the next theorem, we prove that when the optimality measure $\Psi_\epsilon(x_k)$ is sufficiently small, then we have an approximate critical point of the penalty function $\mathcal{P}_\rho^q(\cdot)$ and and an $\epsilon$-first-order solution for the problem (1).

**Theorem 4.1.** *If the sequence* $\{x_k\}_{k\geq 0}$ *generated by Algorithm 3.1 is in some compact set* $\mathcal{S}$ *on which Assumptions 1 and 2 hold and let* $x_k$ *be an iterate satisfying* $\Psi_\epsilon(x_k) \leq \epsilon^2$ *for a given* $0 < \epsilon \leq 1$. *Then, there exists* $\lambda_k$ *such that*

$$\|\nabla f(x_k) + J_F(x_k)^T \lambda_k\| \leq \epsilon. \tag{16}$$

*Moreover, if* $\rho = \mathcal{O}\left(\frac{1}{\epsilon^{q-1}}\right)$, *then* $x_k$ *is an* $\epsilon$*-first-order solution for* (1), *within* $k = \mathcal{O}\left(\frac{1}{\epsilon^{2+\frac{q-1}{q}}}\right)$ *iterations.*

*Proof.* Let us denote:

$$s_k^* = \arg\min_{\|s\|\leq\epsilon} \bar{\mathcal{P}}_\rho^q(x_k + s; x_k) = \arg\min_{\|s\|\leq\epsilon} f(x_k) + \langle \nabla f(x_k), s\rangle + \frac{\rho}{q}\|F(x_k) + J_F(x_k)s\|_q^q.$$

Assume that we are in the case $\|s_k^*\| < \epsilon$. Then the above problem is essentially unconstrained and convex, and first-order conditions provide that $\nabla\bar{\mathcal{P}}_\rho^q(x_k + s_k^*; x_k) = 0$, and so there exists $\lambda_k = \rho\,\mathrm{sign}(F(x_k) + J_F(x_k)s_k^*) \circ |F(x_k) + J_F(x_k)s_k^*|^{q-1}$ such that $\nabla f(x_k) + J_F(x_k)^T\lambda_k = 0$, which implies that (16) holds. It remains to consider $\|s_k^*\| = \epsilon$.

Then first-order conditions for $s_k^*$ imply that there exist $\lambda_k = \rho\,\mathrm{sign}(F(x_k) + J_F(x_k)s_k^*) \circ |F(x_k) + J_F(x_k)s_k^*|^{q-1}$, $u_k^* \geq 0$, and $z_k^* \in \partial(\|s_k^*\|)$ such that

$$\nabla f(x_k) + J_F(x_k)^T\lambda_k + u_k^* z_k^* = 0. \tag{17}$$

It follows from the definition of $\Psi_r(x_k)$ that

$$\begin{aligned}\Psi_\epsilon(x_k) &= \bar{\mathcal{P}}_\rho^q(x_k; x_k) - \bar{\mathcal{P}}_\rho^q(x_k + s_k^*; x_k)\\ &= -\langle \nabla f(x_k), s_k^*\rangle + \frac{\rho}{q}\left(\|F(x_k)\|_q^q - \|F(x_k) + J_F(x_k)s_k^*\|_q^q\right),\end{aligned}$$

and replacing $\nabla f(x_k)$ from (17) into the above, we deduce:

$$\begin{aligned}\Psi_\epsilon(x_k) &= \frac{\rho}{q}\left(\|F(x_k)\|_q^q - \|F(x_k) + J_F(x_k)s_k^*\|_q^q\right) + \langle s_k^*, J_F(x_k)^T\lambda_k\rangle + u_k^*\langle s_k^*, z_k^*\rangle\\ &= \frac{\rho}{q}\left(\|F(x_k)\|_q^q - \|F(x_k) + J_F(x_k)s_k^*\|_q^q\right) + \langle s_k^*, J_F(x_k)^T\lambda_k\rangle + u_k^* \times \epsilon,\end{aligned} \tag{18}$$

where we also used that $\langle s_k^*, z_k^*\rangle = \|s_k^*\| = \epsilon$. Let $\varphi(s) = \frac{\rho}{q}\|F(x_k) + J_F(x_k)s\|_q^q$, which is convex; then $\varphi(0) - \varphi(s_k^*) \geq (-s_k^*)^T J_F(x_k)^T\lambda_k$, where $\lambda_k = \rho\,\mathrm{sign}(F(x_k) + J_F(x_k)s_k^*) \circ |F(x_k) + J_F(x_k)s_k^*|^{q-1}$. We then deduce that:

$$\frac{\rho}{q}\left(\|F(x_k)\|_q^q - \|F(x_k) + J_F(x_k)s_k^*\|_q^q\right) + (s_k^*)^T J_F(x_k)^T\lambda_k \geq 0,$$

and thus from (18) and the fact that $\Psi_\epsilon(x_k) \leq \epsilon^2$, we have:

$$\epsilon^2 \geq \Psi_\epsilon(x_k) \geq u_k^* \times \epsilon.$$

From (17) and $\|z_k^*\| = 1$, we deduce

$$u_k^* = u_k^*\|z_k^*\| = \|\nabla f(x_k) + J_F(x_k)^T\lambda_k\| \leq \epsilon.$$

Hence, (16) holds with $\lambda_k = \rho\,\mathrm{sign}(F(x_k) + J_F(x_k)s_k^*)|F(x_k) + J_F(x_k)s_k^*|^{q-1}$. Now, let us consider a KKT point $x^* \in \mathcal{S}$. LICQ ensures the existence of a corresponding $y^*$ such that:

$$\nabla f(x^*) + J_F(x^*)^T y^* = 0 \quad \text{and} \quad F(x^*) = 0.$$

Let us analyze how much $\lambda_k$ deviates from a Lagrange multiplier $y^*$. We have:

$$y^* = -J_F(x^*)^+\nabla f(x^*).$$

Moreover, considering:

$$\|\nabla f(x_k) + J_F(x_k)^T\lambda_k\| \leq \epsilon,$$

it then follows that there exists a vector $d \in \mathbb{R}^n$ with $\|d\| \leq 1$ such that:

$$\nabla f(x_k) + J_F(x_k)^T\lambda_k = \epsilon d.$$

This implies:

$$\lambda_k = -J_F(x_k)^+\nabla f(x_k) + \epsilon J_F(x_k)^+ d.$$

Hence:

$$\begin{aligned}\|\lambda_k - y^*\| &= \| - J_F(x_k)^+\nabla f(x_k) + J_F(x^*)^+\nabla f(x^*) + \epsilon J_F(x_k)^+ d\|\\ &\leq \|J_F(x^*)^+\nabla f(x^*) - J_F(x_k)^+\nabla f(x_k)\| + \epsilon\|J_F(x_k)^+ d\|.\end{aligned}$$

Given the continuity of $J_F^+$ and $\nabla f$, along with the fact that $x_k$ and $x^*$ belong to the compact set $\mathcal{S}$ and that $\|d\| \leq 1$, we conclude that there exists a constant $M_1 \geq 0$ such that: $\|\lambda_k - y^*\| \leq M_1$. Then: $\|\lambda_k\| \leq M_1 + \|y^*\|$. Consequently:

$$\|F(x_k) + J_F(x_k)s_k^*\| = \frac{\||\lambda_k|^{\frac{1}{q-1}}\|}{\rho^{\frac{1}{q-1}}} \leq \left(\frac{\|\lambda_k\|_q}{\rho}\right)^{\frac{1}{q-1}} \leq \mathcal{O}(\epsilon). \tag{19}$$

It then follows that:

$$\begin{aligned}\|F(x_k)\| &\leq \|F(x_k) + J_F(x_k)s_k^*\| + \|J_F(x_k)s_k^*\| \\ &\leq \|F(x_k) + J_F(x_k)s_k^*\| + M_F\|s_k^*\| \leq \mathcal{O}(\epsilon) + M_F\|s_k^*\|.\end{aligned}$$

Moreover, since $\|s_k^*\| \leq \epsilon$, we get:

$$\|\nabla f(x_k) + J_F(x_k)^T\lambda_k\| \leq \epsilon \text{ and } \|F(x_k)\| \leq \mathcal{O}(\epsilon),$$

after $K = \mathcal{O}\left(\frac{\rho^{\frac{1}{q}}}{\epsilon^2}\right) = \mathcal{O}\left(\frac{1}{\epsilon^{2+\frac{q-1}{q}}}\right)$ iterations. □

**Remark 4.1.** *Note that from the convergence rate to an $\epsilon$-first-order solution of problem (1) which is of order $\mathcal{O}\left(\frac{1}{\epsilon^{2+\frac{q-1}{q}}}\right)$ outer iterations (i.e., number of functions, gradients and Jacobians evaluations), as $q$ approaches values close to 1, the convergence rate of Algorithm 3.1 becomes $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$, which coincides with the standard convergence rate for exact penalty methods [2, 12]. On the other hand, when $q = 2$, the convergence rate becomes $\mathcal{O}\left(\frac{1}{\epsilon^{2.5}}\right)$, matching the rate of quadratic penalty methods established e.g., in [6]. Furthermore, for the total complexity (inner and outer iterations), note that the subproblem in Step 5 of our algorithm is unconstrained with a strongly convex objective function whose gradient is Hölder continuous with exponent $q-1$. According to [5], solving this subproblem to an accuracy $\epsilon_{sub} = \epsilon$ using an accelerated gradient method requires $\mathcal{O}\left(\frac{\rho^{\frac{2}{3q-2}}}{\underline{\beta}^{\frac{q}{q-1}}\epsilon^{\frac{2-q}{3q-2}}}\right)$ gradient steps (i.e., products of the form Jacobian at $x_k$ times vectors), up to a logarithmic factor. By selecting $\rho = \mathcal{O}\left(\frac{1}{\epsilon^{q-1}}\right)$ and $\underline{\beta} = \mathcal{O}\left(\rho^{\frac{1}{q}}\right) = \mathcal{O}\left(\frac{1}{\epsilon^{\frac{q-1}{q}}}\right)$, the worst-case complexity of solving the subproblem simplifies to $\mathcal{O}\left(\frac{1}{\epsilon^{\frac{1}{3q-2}}}\right)$. Thus, the total complexity for obtaining an $\epsilon$-first-order solution of problem (1) is $\mathcal{O}\left(\frac{1}{\epsilon^{2+\frac{q-1}{q}+\frac{1}{3q-2}}}\right)$. In particular, as $q \to 1$, this complexity approaches $\mathcal{O}\left(\frac{1}{\epsilon^3}\right)$, while for $q = 2$, it reduces to $\mathcal{O}\left(\frac{1}{\epsilon^{2.75}}\right)$. The optimal total complexity is $\mathcal{O}\left(\frac{1}{\epsilon^{2.74}}\right)$ and it is achieved for $q = 1 + \frac{1}{\sqrt{3}} \approx 1.33$. Paper [2], which employs a Lipschitz penalty function approach, has a total complexity of order $\mathcal{O}\left(\frac{1}{\epsilon^3}\right)$, when employing the scheme in [5], which is usually higher than our total complexity for any $q \in (1, 2]$.*

**Acknowledgments**

The research leading to these results has received funding from: the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 953348.

**Appendix**

**Proof of Lemma 3.3**. Note that the subproblem's objective function $x \mapsto \bar{\mathcal{P}}_\rho^q(\cdot; x_k) + \frac{\beta_{k+1}}{2}\|x - x_k\|^2$ is strongly convex with strong convexity constant $\beta_{k+1}$. Combining this with the optimality of $x_{k+1}$ and the fact that $\bar{\mathcal{P}}_\rho^q(x_k; x_k) = \mathcal{P}_\rho^q(x_k)$, we get:

$$\bar{\mathcal{P}}_\rho^q(x_{k+1}; x_k) \leq \mathcal{P}_\rho^q(x_k) - \beta_{k+1}\|x_{k+1} - x_k\|^2. \tag{20}$$

Further, since $f$ has a Lipschitz gradient, we have the following.

$$f(x_{k+1}) - l_f(x_{k+1}; x_k) \leq \frac{L_f}{2}\|x_{k+1} - x_k\|^2. \tag{21}$$

Moreover, using the fact that the gradient of $\frac{\rho}{q}\|\cdot\|_q^q$ is Holder continuous, we obtain:

$$
\begin{aligned}
&\frac{\rho}{q}\|F(x_{k+1})\|_q^q - \frac{\rho}{q}\|l_F(x_{k+1};x_k)\|_q^q \\
&\overset{(5)}{\leq} \langle \rho \operatorname{sign}\left(l_F(x_{k+1};x_k)\right) \circ |l_F(x_{k+1};x_k)|^{q-1}, F(x_{k+1}) - l_F(x_{k+1};x_k)\rangle \\
&\quad + \frac{3\times m^{\frac{2-q}{2}}\rho}{q}\|F(x_{k+1}) - l_F(x_{k+1};x_k)\|^q \\
&\leq \langle \rho |l_F(x_{k+1};x_k)|^{q-1}, |F(x_{k+1}) - l_F(x_{k+1};x_k)|\rangle + \frac{3\times m^{\frac{2-q}{2}}\rho}{q}\|F(x_{k+1}) - l_F(x_{k+1};x_k)\|^q \\
&= \rho\left(\sum_{i=1}^m |l_{f_i}(x_{k+1};x_k)|^{q-1} \times |f_i(x_{k+1}) - l_{f_i}(x_{k+1};x_k)|\right)^{q\times\frac{1}{q}} \\
&\quad + \frac{3\times m^{\frac{2-q}{2}}\rho}{q}\|F(x_{k+1}) - l_F(x_{k+1};x_k)\|^q.
\end{aligned}
$$

Further, using Holder inequality, we get:

$$
\begin{aligned}
&\frac{\rho}{q}\|F(x_{k+1})\|_q^q - \frac{\rho}{q}\|l_F(x_{k+1};x_k)\|_q^q \\
&\leq \rho\left(\sum_{i=1}^m |l_{f_i}(x_{k+1};x_k)|^q\right)^{\frac{q-1}{q}} \times \left(\sum_{i=1}^m |f_i(x_{k+1}) - l_{f_i}(x_{k+1};x_k)|^q\right)^{\frac{1}{q}} \\
&\quad + \frac{3\times m^{\frac{2-q}{2}}\rho}{q}\|F(x_{k+1}) - l_F(x_{k+1};x_k)\|^q \\
&= \rho\|l_F(x_{k+1};x_k)\|_q^{q-1}\|F(x_{k+1}) - l_F(x_{k+1};x_k)\|_q + \frac{3\times m^{\frac{2-q}{2}}\rho}{q}\|F(x_{k+1}) - l_F(x_{k+1};x_k)\|^q \\
&\leq \rho\|l_F(x_{k+1};x_k)\|_q^{q-1} \times m^{\frac{2-q}{2q}} \times \|F(x_{k+1}) - l_F(x_{k+1};x_k)\| + \frac{3\times m^{\frac{2-q}{2}}\rho}{q}\|F(x_{k+1}) - l_F(x_{k+1};x_k)\|^q \\
&\overset{\text{Ass. 2}}{\leq} \rho\|l_F(x_{k+1};x_k)\|_q^{q-1}\frac{m^{\frac{2-q}{2q}}\times L_F}{2}\|\Delta x_{k+1}\|^2 + \frac{3\times m^{\frac{2-q}{2}}\rho}{q}\left(\frac{L_F}{2}\|\Delta x_{k+1}\|^2\right)^q, \qquad (22)
\end{aligned}
$$

where the second inequality follows from the fact that $\|v\|_2 \leq \|v\|_q \leq m^{\frac{2-q}{2q}}\|v\|_2$, $\forall q \in (1,2]$ and $v \in \mathbb{R}^m$. Furthermore, using Young's inequality, which states that for any $r, s \in (1,\infty)$ satisfying the conjugate relation $\frac{1}{r}+\frac{1}{s}=1$, the following inequality holds for all nonnegative scalars $a, b$: $ab \leq \frac{a^r}{r} + \frac{b^s}{s}$. Using this for $r = q$ and $s = \frac{q}{q-1}$, we bound $\rho\|l_F(x_{k+1};x_k)\|_q^{q-1}$

as follows:

$$\rho\|l_F(x_{k+1};x_k)\|_q^{q-1} - \tfrac{q-1}{q}\tfrac{\beta_{k+1}-L_f}{L_F} \le \tfrac{\rho}{q^{2-q}}\left(\tfrac{L_F}{\beta_{k+1}-L_f}\right)^{q-1}\left(\tfrac{\rho}{q}\|l_F(x_{k+1};x_k)\|_q^q\right)^{q-1}$$
$$= \frac{\rho}{q^{2-q}}\left(\frac{L_F}{\beta_{k+1}-L_f}\right)^{q-1}\left(\bar{\mathcal{P}}_\rho^q(x_{k+1};x_k) - f(x_{k+1}) + f(x_{k+1}) - l_f(x_{k+1};x_k)\right)^{q-1}$$
$$\overset{(20),(21)}{\le} \frac{\rho}{q^{2-q}}\left(\frac{L_F}{\beta_{k+1}-L_f}\right)^{q-1}\left(\mathcal{P}_\rho^q(x_k) - f(x_{k+1}) - \frac{2\beta_{k+1}-L_f}{2}\|\Delta x_{k+1}\|^2\right)^{q-1}$$
$$\le \frac{\rho}{q^{2-q}}\left(\frac{L_F}{\beta_{k+1}-L_f}\right)^{q-1}\left(\mathcal{P}_\rho^q(x_k) - f(x_{k+1}) - (\beta_{k+1}-L_f)\|\Delta x_{k+1}\|^2\right)^{q-1}$$
$$\overset{(f(x_{k+1})\ge\bar{f})}{\le} \frac{\rho}{q^{2-q}}\left(\frac{L_F}{\beta_{k+1}-L_f}\right)^{q-1}\left(\left(\mathcal{P}_\rho^q(x_k) - \bar{f}\right) - (\beta_{k+1}-L_f)\|\Delta x_{k+1}\|^2\right)^{q-1}$$
$$\le \frac{3\times m^{\frac{(2-q)(q-1)}{2q}}\rho}{q^{2-q}}\left(\frac{L_F}{\beta_{k+1}-L_f}\right)^{q-1}\left(\left(\mathcal{P}_\rho^q(x_k) - \bar{f}\right) - (\beta_{k+1}-L_f)\|\Delta x_{k+1}\|^2\right)^{q-1}.$$
$$\le \frac{3\times m^{\frac{(2-q)(q-1)}{2q}}\rho}{q^{2-q}}\left(\frac{L_F}{\beta_{k+1}-L_f}\right)^{q-1}\left(2^{2-q}\left(\mathcal{P}_\rho^q(x_k) - \bar{f}\right)^{q-1} - (\beta_{k+1}-L_f)^{q-1}\|\Delta x_{k+1}\|^{2(q-1)}\right)$$
$$\le \frac{3\times m^{\frac{(2-q)(q-1)}{2q}}\times 2^{2-q}\rho}{q^{2-q}}\left(\frac{L_F}{\beta_{k+1}-L_f}\right)^{q-1}\left(\mathcal{P}_\rho^q(x_k) - \bar{f}\right)^{q-1} - \frac{3\times m^{\frac{(2-q)(q-1)}{2q}}\rho L_F^{q-1}}{q^{2-q}}\|\Delta x_{k+1}\|^{2(q-1)}$$
$$\overset{(9)}{\le} \frac{\beta_{k+1}-L_f}{qL_F} - \frac{3\times m^{\frac{(2-q)(q-1)}{2q}}\rho L_F^{q-1}}{q^{2-q}}\|\Delta x_{k+1}\|^{2(q-1)}, \tag{23}$$

where in the sixth inequality we used the fact that for any $a \ge b \ge 0$ and any $\nu \in (0,1]$, we have: $(a-b)^\nu \le 2^{1-\nu}a^\nu - b^\nu$, and setting $\nu = q-1 \in (0,1]$. Further, using (23) in (22), we get:

$$\frac{\rho}{q}\|F(x_{k+1})\|_q^q - \frac{\rho}{q}\|l_F(x_{k+1};x_k)\|_q^q$$
$$\le \tfrac{\beta_{k+1}-L_f}{2}\|\Delta x_{k+1}\|^2 - 3\times m^{\frac{2-q}{2}}\rho L_F^q\left(\tfrac{1}{2q^{2-q}} - \tfrac{1}{q2^q}\right)\|\Delta x_{k+1}\|^{2q} \le \tfrac{\beta_{k+1}-L_f}{2}\|\Delta x_{k+1}\|^2. \tag{24}$$

Moreover, we have:

$$\mathcal{P}_\rho^q(x_{k+1}) - \bar{\mathcal{P}}_\rho^q(x_{k+1};x_k) = f(x_{k+1}) - l_f(x_{k+1};x_k) + \tfrac{\rho}{q}\|F(x_{k+1})\|_q^q - \tfrac{\rho}{q}\|l_F(x_{k+1};x_k)\|_q^q.$$

Using (21) and (24) in the previous relation, it follows (inequality (8)):

$$\mathcal{P}_\rho^q(x_{k+1}) \le \bar{\mathcal{P}}_\rho^q(x_{k+1};x_k) + \frac{\beta_{k+1}}{2}\|\Delta x_{k+1}\|^2.$$

Finally, using (20), we get the decrease in (10). This proves our statement. □

**Proof of Lemma 4.2** Let us first assume that $\beta_{k+1}r^2 \le \Psi_r(x_k)$. Then:

$$\min_{s\in\mathbb{R}^n}\left\{\bar{\mathcal{P}}_\rho^q(x_k+s;x_k) + \frac{\beta_{k+1}}{2}\|s\|^2\right\} \le \min_{\|s\|\le r}\left\{\bar{\mathcal{P}}_\rho^q(x_k+s;x_k) + \frac{\beta_{k+1}}{2}\|s\|^2\right\}$$
$$\le \min_{\|s\|\le r}\left\{\bar{\mathcal{P}}_\rho^q(x_k+s;x_k)\right\} + \frac{\beta_{k+1}r^2}{2} \le \min_{\|s\|\le r}\left\{\bar{\mathcal{P}}_\rho^q(x_k+s;x_k)\right\} + \frac{\Psi_r(x_k)}{2},$$

and so, from (7) and (6), it follows that:

$$\bar{\Psi}(x_k,\beta_{k+1}) \ge \bar{\mathcal{P}}_\rho^q(x_k;x_k) - \min_{\|s\|\le r}\bar{\mathcal{P}}_\rho^q(x_k+s;x_k) - \tfrac{\Psi_r(x_k)}{2} = \Psi_r(x_k) - \tfrac{\Psi_r(x_k)}{2} = \tfrac{\Psi_r(x_k)}{2},$$

which proves (15) in the case when $\beta_{k+1}r^2 \le \Psi_r(x_k)$.
Now let $\beta_{k+1}r^2 > \Psi_r(x_k)$ and $s_k^* \triangleq \arg\min_{\|s\|\le r}\bar{\mathcal{P}}_\rho^q(x_k+s;x_k)$. Then, by defining $s_k \triangleq$

$x_{k+1} - x_k$, we get:

$$\begin{aligned}\bar{\mathcal{P}}_\rho^q(x_k + s_k; x_k) + \frac{\beta_{k+1}}{2}\|s_k\|^2 &\leq \bar{\mathcal{P}}_\rho^q\left(x_k + \frac{\Psi_r(x_k)}{\beta_{k+1}r^2}s_k^*; x_k\right) + \frac{\beta_{k+1}}{2}\left\|\frac{\Psi_r(x_k)}{\beta_{k+1}r^2}s_k^*\right\|^2 \\ &\leq \bar{\mathcal{P}}_\rho^q\left(x_k + \frac{\Psi_r(x_k)}{\beta_{k+1}r^2}s_k^*; x_k\right) + \frac{\Psi_r(x_k)^2}{2\beta_{k+1}r^2},\end{aligned}$$

where, to obtain the second inequality, we used $\|s_k^*\| \leq r$. This and (7) give

$$\bar{\Psi}(x_k, \beta_{k+1}) \geq \bar{\mathcal{P}}_\rho^q(x_k; x_k) - \bar{\mathcal{P}}_\rho^q\left(x_k + \frac{\Psi_r(x_k)}{\beta_{k+1}r^2}s_k^*; x_k\right) - \frac{\Psi_r(x_k)^2}{2\beta_{k+1}r^2}. \tag{25}$$

Using $0 < \frac{\Psi_r(x_k)}{\beta_{k+1}r^2} < 1$ and the fact that $\bar{\mathcal{P}}_\rho^q$ is convex, we obtain:

$$\bar{\mathcal{P}}_\rho^q\left(x_k + \frac{\Psi_r(x_k)}{\beta_{k+1}r^2}s_k^*; x_k\right) \leq \left(1 - \frac{\Psi_r(x_k)}{\beta_{k+1}r^2}\right)\bar{\mathcal{P}}_\rho^q(x_k; x_k) + \frac{\Psi_r(x_k)}{\beta_{k+1}r^2}\bar{\mathcal{P}}_\rho^q(x_k + s_k^*; x_k),$$

which substituted into (25) gives

$$\begin{aligned}\bar{\Psi}(x_k, \beta_{k+1}) &\geq \frac{\Psi_r(x_k)}{\beta_{k+1}r^2}\left(\bar{\mathcal{P}}_\rho^q(x_k; x_k) - \bar{\mathcal{P}}_\rho^q(x_k + s_k^*; x_k)\right) - \frac{\Psi_r(x_k)^2}{2\beta_{k+1}r^2} \\ &= \frac{\Psi_r(x_k)^2}{\beta_{k+1}r^2} - \frac{\Psi_r(x_k)^2}{2\beta_{k+1}r^2} = \frac{\Psi_r(x_k)^2}{2\beta_{k+1}r^2},\end{aligned}$$

where we also used (6) and the choice of $s_k^*$. This concludes our proof. □

**Proof of Lemma 4.3** It suffices to prove that given any $\epsilon \in (0, 1]$, the total number of iterations of Algorithm 3.1 with $\Psi_\epsilon(x_k) > \epsilon^2$ is at most

$$K \leq \left\lceil 2\bar{\beta}\left(\bar{P} - \underline{P}\right)\epsilon^{-2}\right\rceil \leq \mathcal{O}\left(\frac{\rho^{\frac{1}{q}}}{\epsilon^2}\right).$$

Using Lemma 4.2 with the fact that $\beta_k \leq \bar{\beta}$ for any $k \geq 1$, it follows that:

$$\bar{\Psi}(x_k, \beta_{k+1}) \geq \frac{1}{2}\min\left(1, \frac{\Psi_\epsilon(x_k)}{\bar{\beta}\epsilon^2}\right)\Psi_\epsilon(x_k), \quad \text{for } k \geq 0.$$

Thus, while Algorithm 3.1 does not terminate, $\Psi_\epsilon(x_k) > \epsilon^2$ and $\epsilon \leq 1$ provide

$$\bar{\Psi}(x_k, \beta_{k+1}) \geq \frac{1}{2}\min\left(1, \frac{\epsilon^2}{\bar{\beta}\epsilon^2}\right)\epsilon^2 = \frac{\epsilon^2}{2\bar{\beta}},$$

where the equality follows from the fact that $\bar{\beta} \geq 1$. Combining the above inequality with (12), we get

$$\mathcal{P}_\rho^q(x_k) - \mathcal{P}_\rho^q(x_{k+1}) \geq \bar{\Psi}(x_k, \beta_{k+1}) \geq \frac{\epsilon^2}{2\bar{\beta}}.$$

Let $K > 0$. Summing up the above inequality over $k$, we get

$$\bar{P} - \underline{P} \geq \sum_{k=0}^{K}[\mathcal{P}_\rho^q(x_k) - \mathcal{P}_\rho^q(x_{k+1})] \geq K\frac{\epsilon^2}{2\bar{\beta}},$$

and so $K \leq 2\frac{\bar{\beta}(\bar{P}-\underline{P})}{\epsilon^2} \leq \mathcal{O}\left(\frac{\rho^{\frac{1}{q}}}{\epsilon^2}\right)$, which proves our claim. □

## REFERENCES


[1] D.P. Bertsekas *On penalty and multiplier methods for constrained minimization*, SIAM Journal on Control and Optimization, 14: 216–235, 1976.

[2] C. Cartis, N. Gould and P. Toint, *On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming*, SIAM Journal on Optimization, 21: 1721-1739, 2011.

[3] E. Cohen, N. Hallak and M. Teboulle, *A dynamic alternating direction of multipliers for nonconvex minimization with nonlinear functional equality constraints*, Journal of Optimization Theory and Applications, 193: 324–353, 2022.

[4] R. Courant, *Variational methods for the solution of problems of equilibrium and vibration*, Bulletin of the American Mathematical Society, 49, 1–23, 1943.

[5] O. Devolder, F. Glineur and Y. Nesterov, *First-order methods with inexact oracle: the strongly convex case*, CORE Discussion Paper, 2013.

[6] L. E. Bourkhissi and I. Necoara, *Complexity of linearized quadratic penalty for optimization with nonlinear equality constraints*, Journal of Global Optimization, 91: 483–510, 2025.

[7] L. E. Bourkhissi and I. Necoara, *Convergence analysis of linearized $\ell_q$ penalty methods for nonconvex optimization with nonlinear equality constraints*, arXiv preprint arXiv:2503.08522, 2025.

[8] R. Fletcher, *Practical Methods of Optimization*, 2nd edition, Wiley, 1987.

[9] A. F. Izmailov and M. V. Solodov, *Convergence rate estimates for penalty methods revisited*, Computational Optimization and Applications, 85: 973–992, 2023.

[10] W. Kong, J. Melo and R. Monteiro, *Complexity of a quadratic penalty accelerated inexact proximal point method for solving linearly constrained nonconvex composite programs*, SIAM Journal on Optimization, 29(4): 2566-2593, 2019.

[11] Q. Lin, R. Ma and Y. Xu, *Complexity of an inexact proximal-point penalty method for constrained smooth non-convex optimization*, Computational Optimization and Applications, 82: 175–224, 2022.

[12] Y. Nabou and I. Necoara, *Regularized higher-order Taylor approximation methods for nonlinear least-squares*, arXiv preprint arXiv:2503.02370, 2025.

[13] J. Nocedal and S.J. Wright, *Numerical Optimization*, Springer, 2006.

[14] B.T. Polyak and N. V.Tretyakov, *The method of penalty estimates for conditional extremum problems*, USSR Computational Mathematics and Mathematical Physics, 13: 42–58, 1973.

[15] Y. Xie, S.J. Wright, *Complexity of proximal augmented Lagrangian for nonconvex optimization with nonlinear equality constraints*, Journal of Scientific Computing, 86, 2021.

[16] Y. Yuan, *Conditions for convergence of trust region algorithms for non-smooth optimization*, Mathematical Programming, 31: 220–228, 1985.