Published in
J. Optim. Theory Appl. 153 (2012), no.2, pp. 388-407.

DUAL CONVERGENCE FOR PENALTY ALGORITHMS IN CONVEX PROGRAMMING

FELIPE ALVAREZ, MIGUEL CARRASCO, THIERRY CHAMPION

Abstract. Algorithms for convex programming, based on penalty methods, can be designed to have good primal convergence properties even without uniqueness of optimal solutions. Taking primal convergence for granted, in this paper we investigate the asymptotic behavior of an appropriate dual sequence obtained directly from primal iterates. First, under mild hypotheses, which include the standard Slater condition but neither strict complementarity nor second-order conditions, we show that these dual sequence is bounded and also, each cluster point belongs to the set of Karush-Kuhn-Tucker multipliers. Then we identify a general condition on the behavior of the generated primal objective values that ensures the full convergence of the dual sequence to a specific multiplier. This dual limit depends only on the particular penalty scheme used by the algorithm. Finally, we apply this approach to prove the first general dual convergence result of this kind for penalty-proximal algorithms in a nonlinear setting.

Keywords. Convex programming, penalty function, dual problem.
AMS Classification. 90C25, 49M37.

Communicated by Jean-Pierre Crouzeix

The first author was supported by the Institute on Complex Engineering Systems (ICM: P-05-004-F, CONICYT: FBO16) of Universidad de Chile. The second author was supported by the FONDECYT grant 11090328.

Felipe Alvarez
Centro de Modelamiento Matemático (CNRS UMI 2807), Departamento de Ingeniería Matemática, FCFM, Universidad de Chile, Av. Blanco Encalada 2120, Santiago, Chile.

Miguel Carrasco
Facultad de Ingeniería y Ciencias Aplicadas, Universidad de Los Andes, Av. San Carlos de Apoquindo 2200, Las Condes, Santiago, Chile. (corresponding author)

Thierry Champion
Laboratoire Imath, U.F.R. des Sciences et Techniques, Université du Sud Toulon-Var, Avenue de l'Université, BP 20132, 83957 La Garde cedex, France.

1. Introduction

Exact penalty/barrier methods for convex programming can be designed to have good primal convergence properties even without uniqueness of optimal solutions [1, 2, 3, 4, 5]. Moreover, from the unconstrained first-order optimality condition for the auxiliary penalty problem, one may obtain a multiplier estimate that satisfies a sort of perturbed Karush-Kuhn-Tucker (KKT) system, depending on a scalar penalty parameter which is intended to be small. Indeed, under fairly general conditions, such a dual optimal path converges as the penalty parameter goes to zero to a special multiplier; see [6].

In practical computations, given a value for the penalty parameter, an iterative subroutine is applied to find an approximation of the corresponding exact primal penalty optimal solution, and then the parameter is updated. Such procedures usually require the current approximate solution to satisfy some inexact versions of the first-order optimality conditions [7], which is a key condition to ensure primal convergence.

Taking primal convergence for granted, in this paper we investigate the asymptotic behavior of an approximate dual sequence obtained directly from primal iterates generated by a general penalty algorithm, extending the convergence results that are known for the exact case. More precisely, after reviewing briefly the theory of general penalty methods, we introduce in §2 an approximate multiplier sequence and show that all its cluster points satisfy the KKT system. Then in §3, we provide a general criterion that ensures full convergence to a special multiplier in the dual optimal set; such dual limit depends only on the particular penalty scheme used by the algorithm. Finally, in §4 we apply our general approach to obtain the first result of this type in a nonlinear framework for the so called penalty-proximal algorithms, which extends significantly previous work on this subject where primal convergence has been already established; see [8, 9, 10, 3, 11]. Indeed, for purely primal penalty-proximal algorithms, to the best of our knowledge this is the first dual convergence result beyond the very restrictive case of Linear Programming.

2. Penalty Methods and General Duality Results

Let us consider the mathematical programming problem

$$(P) \qquad \min\{f_0(x) \mid f_i(x) \le 0,\, i = 1, \dots, m\},$$

where $f_i : \mathbb{R}^d \to \mathbb{R}$ is convex and continuously differentiable for every $i \in \{0, \dots, m\}$. We assume the standard Slater constraint qualification condition:

$$\exists x_0 \in \mathbb{R}^d,\ \forall i \in \{1, \dots, m\},\ f_i(x_0) < 0. \tag{S}$$

We suppose that the primal optimal solution set $S(P)$ be nonempty. In order to solve (P) approximately, we take a penalty scheme of the type

$$(P_r) \qquad \min_{x \in \mathbb{R}^d} f(x, r), \quad \text{for} \quad f(x, r) := f_0(x) + r \sum_{i=1}^{m} \theta(f_i(x)/r),$$

where $r > 0$ is a real parameter and $\theta : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ is a closed and convex function satisfying:

$$\begin{array}{ll} \text{(a)} & \operatorname{dom}\theta =]-\infty, \kappa[\text{ for some } \kappa \in [0, \infty]. \\ \text{(b)} & \theta :]-\infty, \kappa[\to \mathbb{R} \text{ is strictly convex and smooth.} \\ \text{(c)} & \theta'(u) > 0 \text{ for all } u \in]-\infty, \kappa[,\ \theta'(u) \to 0 \text{ as } u \to -\infty \text{ and } \theta'(u) \to +\infty \text{ as } u \nearrow \kappa. \end{array} \tag{H$_1$}$$

Typical choices are the exponential penalty $\theta(u) = \exp(u)$ with $\kappa = +\infty$, and the log-barrier $\theta(u) = -\ln(-u)$ or the inverse barrier $\theta(u) = -1/u$, both with $\kappa = 0$. Under mild conditions on the data functions f_i, for each $r > 0$ there exists a unique solution $x(r)$ to (P_r), and all cluster points of the optimal path $\{x(r) : r \searrow 0\}$ belong to $S(P)$; see, for instance, [6, 4]. Moreover, the primal optimal path $\{x(r) : r \searrow 0\}$ converges to some $x^\theta \in S(P)$; see [1, 2, 3, 5].

Let us introduce the Lagrangian function given by

$$L(x, \lambda) := f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \tag{1}$$

for any $(x, \lambda) \in \mathbb{R}^d \times \mathbb{R}^m$. In our setting, a point $x^* \in \mathbb{R}^d$ is optimal for (P), if and only if there is a *multiplier* $\lambda^* \in \mathbb{R}^m$ such that the pair (x^*, λ^*) satisfies the *Karush-Kuhn-Tucker* (KKT) conditions

$$\text{(KKT)} \qquad \left\{ \begin{array}{r} \nabla_x L(x^*, \lambda^*) = \nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) = 0, \\ \lambda_i^* \geq 0, \ f_i(x^*) \leq 0, \\ \lambda_i^* f_i(x^*) = 0, \quad \text{for all } i \in \{1, \ldots, m\}. \end{array} \right.$$

Recall that in our setting and under Slater's condition (S), a pair $(x^*, \lambda^*) \in \mathbb{R}^d \times \mathbb{R}^m$ satisfies the KKT conditions, if and only if x^* is an optimal solution to (P) and λ^* solves the dual problem given by

$$(D) \qquad \max_{\lambda \geq 0} p(\lambda), \qquad \text{for} \quad p(\lambda) := \inf_{x \in \mathbb{R}^d} L(x, \lambda).$$

Given $r > 0$, the unconstrained first-order optimality condition for (P_r), which is given by

$$\nabla_x f(x(r), r) = \nabla f_0(x(r)) + \sum_{i=1}^{m} \theta' \left(f_i(x(r))/r \right) \nabla f_i(x(r)) = 0, \tag{2}$$

may be written equivalently as

$$\begin{aligned} \nabla_x L(x(r), \lambda(r)) &= 0, && (3) \\ \lambda_i(r) &= \theta' \left(f_i(x(r))/r \right), \quad i = 1, ..., m. && (4) \end{aligned}$$

As $\theta' > 0$, we get that the *approximate multiplier* $\lambda(r)$ given by (4) is positive. Thus, (3)-(4) is a sort of approximate KKT system. Indeed, the dual optimal path $\lambda(r)$ converges as $r \searrow 0$ to a special multiplier λ^{θ^*}, which is called the θ^**-center* of the dual optimal set $S(D)$; see [6] and §3 for some properties on λ^{θ^*}.

We first discuss the convergence of the optimal paths $(x(r))_r$ and $(\lambda(r))_r$. Notice that the penalty objective function $f(\cdot, r)$ is convex, and at least differentiable on the set of all the Slater points for (P). For a given $x \in \mathbb{R}^d$, it is easy to verify that letting $r \searrow 0$ we get $f(x, r) \to f_0(x)$ when $f_i(x) < 0$ for all $i \in \{1, \ldots, m\}$, while $f(x, r) \to +\infty$ whenever $f_{i_0}(x) > 0$ for some $i_0 \in \{1, \ldots, m\}$; see [4]. Therefore, for a sufficiently small $r > 0$, it is natural to consider any solution $x(r)$ of (P_r) as an approximate solution of the original problem (P). The existence of $x(r)$ follows under mild condition, for example if $S(P)$ is bounded (together with (S) when $\kappa = 0$); see [6]. Moreover, under some *quasi-analytic* conditions on the data functions, $x(r)$ is unique and converges to some specific primal solution $x^\theta \in S(P)$ as $r \searrow 0$, a limit which depends on the special choice of θ in the general case when $S(P)$ is not a singleton [1, 2, 3, 5]. Without such quasi-analytic conditions, even for infinitely differentiable convex data, the optimal path may exhibit an oscillatory ill behavior [2, 12].

On the other hand, under our conditions the dual optimal set $S(D)$ turns out to be nonempty and bounded; for instance, see [13]. For a given $r > 0$, the unconstrained first-order optimality condition $\nabla_x f(x(r), r) = 0$ for $x(r)$ to be optimal for (P_r), which we have rewritten as the system (3)-(4), generates naturally an approximated dual sequence

$$\lambda_i(r) = \theta' \left(f_i(x(r))/r \right), \ i = 1, \ldots, m.$$

Note that by (H$_1$)(c), we have $\lambda_i(r) > 0$ for $i = 1, ..., m$, that is, $\lambda(r)$ is strictly feasible for (D). Furthermore, it happens that $\lambda(r)$ is the unique solution to the following problem (see [6]):

$$(D_r) \qquad \max_{\lambda \geq 0} \left\{ p(\lambda) - r \sum_{i=1}^{m} \theta^*(\lambda_i) \right\}.$$

Here, $\theta^*(\lambda) = \sup_u\{\lambda u - \theta(u)\}$ is the *Fenchel conjugate* of θ [14], which plays the role of a barrier-type penalty function for the positivity constraint $\lambda \geq 0$ in the dual problem. Indeed, for the exponential penalty we have $\theta^*(\lambda) = \lambda \log \lambda - \lambda$ if $\lambda \geq 0$ and ∞ otherwise, for the log-barrier we have $\theta^*(\lambda) = -1 - \log \lambda$ and for the inverse-barrier $\theta^*(\lambda) = 1/\sqrt{\lambda} - \sqrt{\lambda}$ if $\lambda > 0$ and ∞ otherwise.

Due to the special form of the penalty objective function in (D_r), the natural candidate to be the limit of the dual optimal path $\lambda(r)$ is the θ^*-center [15], whose definition is the following:

Definition 2.1. *The* θ^*-center *of $S(D)$ is the unique $\lambda^{\theta^*} \in S(D)$ such that*

$$\sum_{i \in I} \theta^*(\lambda_i^{\theta^*}) = \min_{\lambda^* \in S(D)} \sum_{i \in I} \theta^*(\lambda_i^*),$$

where $I := \{i \mid \exists \lambda_i^* \in S(D),\, \lambda_i^* > 0\}$. If $I = \emptyset$ then we set $\lambda^{\theta^*} = 0$.

Remark 2.1. By virtue of complementary in the KKT system, for every $i \in I$ and any $x^* \in S(P)$ we have that $f_i(x^*) = 0$. Whenever $f_i(x^*) < 0$ for some $x^* \in S(P)$, we get $i \notin I$. The case $I = \emptyset$ implies that $S(D) = \{0\}$, and the KKT system reduces to the feasibility of x^* together with the optimality condition $\nabla f_0(x^*) = 0$, and by convexity, it turns out that x^* is a global minimum of f_0. Of course, the interesting case is when $S(D)$ is not a singleton so that in particular $I \neq \emptyset$. In the general case, the existence and uniqueness of the θ^*-center follow from (S), which ensures $S(D)$ to be nonempty and bounded, and the hypotheses made on θ.

Under our conditions, the exact dual optimal path $\lambda(r)$ indeed converges to λ^{θ^*} as $r \searrow 0$; see, for instance, [6, 16].

In practical computations, an iterative subroutine is applied to find an approximation

$$x^k \approx x(r_k),$$

and then the penalty parameter is updated to some $r_{k+1} < r_k$. Such procedures usually require the current approximate solution x^k to satisfy some inexact versions of the first-order optimality conditions (see [7]). Concerning the penalty parameter sequence $(r_k)_k$, we assume that

$$r_k > 0 \text{ and } r_k \to 0 \text{ as } k \to \infty. \tag{H$_2$}$$

From now on, we assume that such a primal sequence $(x^k)_k$ exists and satisfies the following:

$$\nabla_x f(x^k, r_k) \to 0 \text{ and } x^k \to x^\infty \text{ as } k \to +\infty, \tag{H$_3$}$$

for some optimal solution $x^\infty \in S(P)$, which may depend on the starting point x^0 and on some parameters of the underlying algorithm used to find x^k. Thus, we take for granted that primal convergence holds. Of course, this is not true in general but there are several results in this direction; we will discuss this in more details for one specific framework later on (cf. §4). In such a case, the behavior of the dual sequence

$$\lambda_i^k = \theta'\left(f_i(x^k)/r_k\right), \quad i = 1, ..., m, \tag{5}$$

is important as it is expected the asymptotic convergence to a multiplier as $r_k \searrow 0$. On the one hand, dual solutions are interesting on their own: they are useful for sensitivity analysis and, moreover, in some applications they have an actual modeling interpretation. On the other hand, if the dual sequence λ^k has its cluster points in the multipliers set, this provide a theoretical basis for any global KKT-based stopping rule for the algorithm.

Lemma 2.1. *Under* (S) *together with the assumptions* (H_1)-(H_3)*, the* approximate multiplier *sequence* $(\lambda^k)_k \subset \mathbb{R}^m_{++}$ *(here* $\mathbb{R}^m_{++} := \{x \in \mathbb{R}^m \mid x_i > 0,\ i = 1, \ldots, m\}$*) defined by* (5) *is bounded and, furthermore,* $\operatorname{dist}(\lambda^k, S(D)) \to 0$ *as* $k \to +\infty$*.*

Proof. By virtue of (H_1)(c), we have that for every i such that $f_i(x^\infty) < 0$, the corresponding λ^k_i tends to 0 as $k \to \infty$. On the other hand, notice that

$$\nabla_x L(x^k, \lambda^k) \;=\; \nabla f_0(x^k) + \sum_{i=1}^m \lambda^k_i \nabla f_i(x^k) \;=\; \nabla_x f(x^k, r_k).$$

It follows from (H_3) that $\nabla_x L(x^k, \lambda^k) \to 0$ as $k \to \infty$. As a consequence, if λ^∞ is a cluster point of $(\lambda^k)_k$ then $(x^\infty, \lambda^\infty)$ satisfies the KKT conditions, so that λ^∞ is indeed a multiplier, or equivalently $\lambda^\infty \in S(D)$. To conclude, it suffices to prove the boundedness of $(\lambda^k)_k$. To this end, recall that $\nabla f_0(x^k) + \sum_{i=1}^m \lambda^k_i \nabla f_i(x^k) \to 0$ as $k \to \infty$. If the sequence $(\lambda^k)_k$ was unbounded, then dividing by $\sum_{i=1}^m \lambda^k_i$, letting $k \to +\infty$ and taking a subsequence if necessary, we would deduce that there exists some coefficients $\alpha_1, ..., \alpha_m \geq 0$ with $\sum_{i=1}^m \alpha_i = 1$ such that $\sum_{i=1}^m \alpha_i \nabla f_i(x^\infty) = 0$. In addition, as a consequence of the fact that $\lambda^k_i \to 0$ for all i such that $f_i(x^\infty) < 0$, we would have $\alpha_i = 0$ for any of such i's. Of course, this would imply that $S(D) = S(D) + \mathbb{R}_+ \cdot \alpha$, for some $\alpha \neq 0 \in \mathbb{R}^m_+$. But this would be a contradiction because Slater's condition (S) ensures that $S(D)$ is compact. This completes the proof. □

Under the hypotheses of Lemma 2.1, the dual sequence $(\lambda^k)_k$ is bounded and every cluster point belongs to the multipliers set. Now a natural but more delicate question is whether the dual sequence given by (5) fully converges to a specific multiplier as $k \to \infty$. This is the goal of the next section. This type of convergence result is interesting from a theoretical viewpoint but particularly scarce for purely primal algorithms.

3. Full Dual Convergence to the θ^*-Center

The main result of this section, Th. 3.1 below, provides a general condition ensuring the convergence for the dual sequence $(\lambda^k)_k$ given by (5). In the following, $v(P)$ stands for the optimal value of problem (P).

Theorem 3.1. *Suppose that Slater's condition* (S) *holds, together with* (H_1)-(H_3)*. Assume, in addition, the following condition:*

$$\text{The ratio } \frac{f_0(x^k) - v(P)}{r_k} \text{ is bounded from above.} \tag{C}$$

Then $(\lambda^k)_k$ *given by* (5) *converges to the* θ^**-center* λ^{θ^*} *of* $S(D)$*.*

Remark 3.1. Note that a sufficient condition for (C) is the following: $f_0(x^k) = f_0(x^\infty) + \mathcal{O}(r_k)$, where x^∞ is the primal limit of x^k as stated in (H_3) so that $f_0(x^\infty) = v(P)$. Thus, (C) is a sort of zero-order asymptotic development of the primal objective function values with respect to the penalty parameter.

Remark 3.2. Of course, condition (C) may be difficult to verify as it relies on the sequence $(x^k)_k$ and on the primal optimal value $v(P)$ which is unknown in general. It is interesting to exhibit conditions on the parameters of a given algorithm generating $(x^k)_k$ which ensure that (C) holds true *a priori*, specially without the explicit knowledge of $v(P)$. In Section 4 we show that it is possible to do so for a class of implementable versions of the penalty-proximal point method.

For the proof of Theorem 3.1 we will need some auxiliary results. First note that the interesting case is when $S(D)$ is not reduced to a singleton, and in particular we shall assume that $I \neq \emptyset$. Notice that in this case, due to (H_1), the θ^*-center λ^{θ^*} of $S(D)$ is such that $\lambda_i^{\theta^*} > 0$ for any $i \in I$. To prove the convergence of $(\lambda^k)_k$, the idea is to show that any of its cluster points satisfies the characterization of λ^{θ^*} given in the following auxiliary results:

Lemma 3.1. *Assume that $\lambda^* \in S(D)$ be such that*

$$\forall i \in I,\ \lambda_i^* \ = \ \theta'(\langle \nabla f_i(x^*), v\rangle)$$

for some $x^* \in S(P)$ and some $v \in \mathbb{R}^d$. Then $\lambda^* = \lambda^{\theta^*}$, i.e. *$\lambda^*$ is the θ^*-center of $S(D)$.*

Proof. Since λ^* and λ^{θ^*} both belong to $S(D)$, and $\lambda_i^* = \lambda_i^{\theta^*} = 0$ for any $i \in \{1, \dots, m\} \setminus I$, we get

$$\sum_{i\in I} \lambda_i^* \nabla f_i(x^*) \ = \ -\nabla f_0(x^*) \ = \ \sum_{i \in I} \lambda_i^{\theta^*} \nabla f_i(x^*).$$

Next, we compute

$$\begin{aligned}\sum_{i\in I} \theta^*(\lambda_i^{\theta^*}) - \theta^*(\lambda_i^*) \quad &\geq \quad \sum_{i\in I} \theta^{*\prime}(\lambda_i^*)(\lambda_i^{\theta^*} - \lambda_i^*) \ = \ \sum_{i\in I} \langle \nabla f_i(x^*), v\rangle (\lambda_i^{\theta^*} - \lambda_i^*) \\ &= \quad \langle \sum_{i\in I} \lambda_i^{\theta^*} \nabla f_i(x^*) - \sum_{i\in I} \lambda_i^* \nabla f_i(x^*)\,, v\rangle \ = \ 0,\end{aligned}$$

where the first inequality is by convexity of θ^*, and then we have used $\theta^{*\prime} = \theta'^{-1}$ (see [14]). This implies that λ^* minimizes $\lambda \mapsto \sum_{i\in I} \theta^*(\lambda_i)$ over $S(D)$. By uniqueness, we deduce that $\lambda^* = \lambda^{\theta^*}$. □

Lemma 3.2. *Let $x^* \in S(P)$ and consider $\Psi \colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ defined by*

$$\Psi(v) \ := \ \langle \nabla f_0(x^*), v\rangle + \sum_{i\in I} \theta(\langle \nabla f_i(x^*), v\rangle). \tag{6}$$

If $F := \mathrm{Span}\{\nabla f_i(x^*) : i \in I\}$, then the restriction of Ψ to F is coercive and strictly convex, hence has a unique minimizer v^* on F. Moreover, we have

$$\forall i \in I,\ \lambda_i^{\theta^*} \ = \ \theta'(\langle \nabla f_i(x^*), v^*\rangle). \tag{7}$$

Proof. Given $v \in F$, the recession function Ψ_∞ of Ψ at v is given by (see [17]):

$$\begin{aligned}\Psi_\infty(v) \quad &= \quad \langle \nabla f_0(x^*), v\rangle + \sum_{i\in I} \theta_\infty(\langle \nabla f_i(x^*), v\rangle) \\ &= \quad \sum_{i\in I} \Big[\theta_\infty(\langle \nabla f_i(x^*), v\rangle) - \lambda_i^{\theta^*} \langle \nabla f_i(x^*), v\rangle\Big]\end{aligned}$$

where we have used $\lambda^{\theta^*} \in S(D)$. The hypotheses made on θ yield that $\theta_\infty(u) = 0$ whenever $u \leq 0$ and $\theta_\infty(u) = +\infty$, whenever $u > 0$. Since $\lambda_i^{\theta^*} > 0$ for any $i \in I$, then, it follows that $\Psi_\infty(v) \in]0, +\infty]$ for any $v \in F \setminus \{0\}$, which proves the coercivity of Ψ. The strict convexity of Ψ on F is inherited from θ. Thus the existence and uniqueness of the minimizer v^* follows directly.

The optimality condition for the unique minimizer v^* of Ψ over F reads

$$-\nabla f_0(x^*) \ = \ \sum_{i\in I} \theta'(\langle \nabla f_i(x^*), v^*\rangle) \nabla f_i(x^*).$$

As a consequence, the vector $\lambda^* \in \mathbb{R}^m$ with coordinates: $\lambda_i^* := \theta'(\langle \nabla f_i(x^*), v^*\rangle)$ for $i \in I$ and $\lambda_i^* = 0$ otherwise, belongs to $S(D)$. Therefore, (7) follows from Lemma 3.1. □

The following Lemma is the cornerstone in the proof of Theorem 3.1.

Lemma 3.3. *Under* (S) *together with* (H$_1$)-(H$_3$), *we define*

$$w^k := \frac{x^k - x^\infty}{r_k}, \quad F_k := \mathrm{Span}\{\nabla f_i(x^k) \mid i \in \{0\} \cup I\}, \quad F_\infty := \mathrm{Span}\{\nabla f_i(x^\infty) \mid i \in I\}$$

and set $w_k^k := \mathrm{Proj}(w^k, F_k)$ *and* $w_\infty^k := \mathrm{Proj}(w^k, F_\infty)$, *where* $\mathrm{Proj}(w, F)$ *stands for the orthogonal projection of* w *onto* F. *Then:*
(i) *For some* $a > 0$, *the sequences* $(w_k^k)_k$ *and* $(w_\infty^k)_k$ *satisfy*

$$\forall k, \quad \|w_k^k\| \leq a\,(\|w_\infty^k\| + 1) \quad \text{and} \quad \|w_\infty^k\| \leq a\,(\|w_k^k\| + 1). \tag{8}$$

(ii) *Thus, up to subsequences,* $w_k^k \to v$ *if and only if* $w_\infty^k \to v$, *for some* $v \in F_\infty$.

Proof. (i) We only prove that

$$\forall k, \quad \|w_k^k\| \leq a\,(\|w_\infty^k\| + 1), \tag{9}$$

for some positive a, the other estimate in (8) is analogous. We make a proof by contradiction, and consequently, up to a change of index, we have that

$$\|w_k^k\| \to +\infty \quad \text{and} \quad \frac{\|w_\infty^k\|}{\|w_k^k\|} \to 0 \quad \text{as } k \to +\infty.$$

It follows from the convexity of the functions f_i that

$$\forall i,\ \forall k,\ \forall y \in \mathbb{R}^d, \quad \langle \nabla f_i(x^\infty), y\rangle \ \leq\ \frac{f_i(x^\infty + r_k\, y) - f_i(x^\infty)}{r_k} \ \leq\ \langle \nabla f_i(x^\infty + r_k\, y), y\rangle. \tag{10}$$

Now (10) applied with $y = w^k$ yields that

$$\forall i \in I \cup \{0\},\ \forall k, \quad \langle \nabla f_i(x^\infty), w_\infty^k\rangle \ \leq\ \langle \nabla f_i(x^\infty + r_k\, w^k), w^k\rangle \ =\ \langle \nabla f_i(x^k), w_k^k\rangle. \tag{11}$$

Thus, dividing by $\|w_k^k\|$ we get

$$\forall i \in I \cup \{0\},\ \forall k, \qquad \langle \nabla f_i(x^\infty), \frac{w_\infty^k}{\|w_k^k\|}\rangle \ \leq\ \langle \nabla f_i(x^k), \frac{w_k^k}{\|w_k^k\|}\rangle.$$

If we consider a cluster point v of $(\frac{w_k^k}{\|w_k^k\|})_k$, then v has norm 1 and belongs to F_∞, moreover

$$\forall i \in I \cup \{0\}, \quad 0 \ \leq\ \langle \nabla f_i(x^\infty), v\rangle. \tag{12}$$

Since it holds

$$\langle \nabla f_0(x^\infty), v\rangle \ =\ -\sum_{i \in I} \lambda_i^{\theta^*} \langle \nabla f_i(x^\infty), v\rangle$$

with $\lambda_i^{\theta^*} > 0$ for any $i \in I$, hence we obtain

$$\forall i \in I \cup \{0\}, \quad \langle \nabla f_i(x^\infty), v\rangle \ =\ 0$$

and then $v = 0$ since $v \in F_\infty$, but this contradicts $\|v\| = 1$. The proof of (9) is complete.

(ii) Assume that $(w_k^k)_k$ converges to some limit $w \in F_\infty$. In particular, by (8), the sequence $(w_\infty^k)_k$ is bounded and with no loss of generality we may assume that it converges to some limit $\hat{w} \in F_\infty$. Consequently, passing to the limit in (11) as $k \to +\infty$, we get

$$\forall i \in I \cup \{0\}, \quad \langle \nabla f_i(x^\infty), \hat{w}\rangle \ \leq\ \langle \nabla f_i(x^\infty), w\rangle.$$

Thus we get (12) for $v = w - \hat{w} \in F_\infty$. Reasoning exactly as before, we conclude that $v = 0$, which yields the desired conclusion. □

Now, we are in a position to prove the main result of this section.

Proof of Theorem 3.1. Let $(w^k)_k$ be as in Lemma 3.3. By taking $y = w^k$ in (10), we get

$$\forall i \in \{0\} \cup I,\ \forall k, \quad \langle \nabla f_i(x^\infty), w^k \rangle \ \leq\ \frac{f_i(x^k) - f_i(x^\infty)}{r_k} \ \leq\ \langle \nabla f_i(x^k), w^k \rangle. \tag{13}$$

Recall that $f_i(x^\infty) = 0$ for any $i \in I$ and $f_0(x^\infty) = v(P)$. Therefore, in particular we have that

$$\forall i \in I,\ \forall k, \quad \langle \nabla f_i(x^\infty), w^k \rangle \ \leq\ \frac{f_i(x^k)}{r_k} = (\theta')^{-1}(\lambda_i^k) = (\theta^*)'(\lambda_i^k).$$

By Lemma 2.1, we know that $(\lambda_i^k)_k$ is bounded in $\mathbb{R}_+^m$, hence $(\langle \nabla f_i(x^\infty), w^k \rangle)_k$ is bounded from above for every $i \in I$.

We claim that for any $i \in I$ the sequence $(\langle \nabla f_i(x^\infty), w^k \rangle)_k$ is also bounded from below when (C) holds. In fact, we have that

$$\forall k, \quad \sum_{i \in I} \lambda_i^{\theta^*} \langle \nabla f_i(x^\infty), w^k \rangle = -\langle \nabla f_0(x^\infty), w^k \rangle \geq \frac{f_0(x^\infty) - f_0(x^k)}{r_k} \geq b,$$

for some suitable constant b thanks to (C), and as $\lambda_i^{\theta^*} > 0$ for all $i \in I$, our claim follows easily.

Thus, $(w_\infty^k)_k$ is bounded. By Lemma 3.3(i), we have that $(w_k^k)_k$ is bounded as well. By Lemma 3.3(ii), up to a subsequence, $(w_\infty^k)_k$ and $(w_k^k)_k$ both converge to some v in F_∞. Passing to the limit in (13), we deduce that the corresponding cluster point λ^∞ of $(\lambda^k)_k$ satisfies the characterization of the θ^*-center given in Lemma 3.1. We thus conclude that $\lambda^\infty = \lambda^{\theta^*}$. Since this holds true for any cluster point, we conclude that the whole sequence $(\lambda^k)_k$ converges to λ^{θ^*}. □

4. Penalty-Proximal Point Algorithms

In this section, we focus on a special class of iterative methods for solving the main problem (P), in which the *Proximal Point Algorithm* (PPA) is coupled with the penalty function approach as given by (P_r). In the following, we shall apply the techniques developed in §3 to obtain the full convergence of the associated multiplier sequence towards the θ^*-center of the dual problem (D).

4.1. **A Unified Framework.**

In the sequel, we consider the sequences $(x^k)_k$ generated by the following inexact implicit iterative scheme: given the current iterate x^{k-1}, a step-size $h_k > 0$ and a tolerance parameter $\varepsilon_k \geq 0$, we find x^k and g^k such that

$$x^k = x^{k-1} - h_k g^k + \eta^k \quad \textit{with} \quad g^k \in \partial_{\varepsilon_k} f(x^k + e^k, r_k), \tag{PPPA}$$

for some errors e^k, $\eta^k \in \mathbb{R}^d$ which are intended to be small. Here $\partial_\varepsilon f(x, r)$ stands for the ε-subdifferential of $f(\cdot, r)$ at x. Note that when $e^k = \eta^k = 0$ in $\mathbb{R}^n$ and $\varepsilon_k = 0$ in $\mathbb{R}$, the Penalty Proximal Point Algorithm (PPPA) amounts to the exact optimality condition for the proximal regularized scheme:

$$x^{k+1} \ = \ \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ f(x, r_k) + \frac{1}{2h_k} \|x - x^k\|^2 \right\}. \tag{14}$$

The square norm introduced here can be viewed as a regularization term of the penalty function. In this way, the unconstrained objective function is globally strongly convex.The idea of using such an iterative process is to avoid possible numerical instabilities in the inner iterations that typically appear when r_k becomes small.

When $\varepsilon_k, r_k \searrow 0$, and under appropriate conditions on the parameters, on the functions f_i as well as on the penalty function θ, the scheme (PPPA) generates sequences $(x^k)_k$ which converge to some $x^\infty \in S(P)$ as $k \to +\infty$. In fact, the implicit iterative scheme given by (PPPA) can be viewed as a generic algorithm, which unifies several methods that have been considered in the literature. First, in the specific case where we impose $\eta^k = e^k = 0$, (PPPA) is exactly the method whose convergence is studied in [10] (see also [18, 19, 20]). More generally, for $\eta^k > 0$ and $e^k = 0$, (PPPA) was considered in [9, 3]. In order to allow $e^k \neq 0$, some variants of the proximal point algorithm have been proposed in the literature. For instance, following [21, 22], a two-step algorithm can be considered, where an inexact proximal iteration is first performed to find an auxiliary point, and then the current iterate is obtained via a projection step. Such a *hybrid* projection-proximal algorithm has the advantage of ensuring primal convergence under a fixed relative error condition at each step; we refer to [8] for results in this direction in the case of penalty methods. Similar considerations also hold for some hybrid *extragradient*-proximal method as proposed in [23].

The framework given by (PPPA) also applies to penalty methods with two parameters as developed in [11]. Indeed these methods are based on the same scheme as (PPPA), where the penalty function θ appearing in $f(\cdot, r_k)$ is replaced by $\beta_k\theta$, the varying parameter β_k being increased on each iteration for which the iterate x^k is not feasible. It is proved in [11] that in the convex case, and assuming that the Slater condition (S) holds, the sequence $(\beta_k)_k$ generated by the algorithm is stationary. In this respect, the asymptotic analysis presented in the following does apply to this type of method (with θ replaced with $(\lim \beta_k)\theta$).

From now on, we assume that (S) and (H$_1$) hold, and by analogy with (H$_3$) we shall assume that the sequences $(x^k)_k$ and $(g^k)_k$ generated by (PPPA) satisfy

$$x^k \to x^\infty \text{ and } g^k \to 0 \text{ as } k \to +\infty \tag{A$_1$}$$

for some $x^\infty \in S(P)$. In view of already known primal convergence results (see, for instance, [8, 9, 10, 3, 11]), natural but not sufficient assumptions for (A$_1$) to hold are

$$r_k \searrow 0 \text{ and } \varepsilon_k \searrow 0 \text{ as well as both residues } \eta^k \text{ and } e^k \text{ tend to 0 in } \mathbb{R}^d \text{ as } k \to \infty, \tag{A$_2$}$$

where we recall that $r_k \searrow 0$ means $(r_k)_k$ goes to 0 with positive values.

We shall also assume that

$$\exists h > 0, \forall k \geq 0, \quad h_k \geq h. \tag{A$_3$}$$

Remark 4.1. As part of the hypotheses stated in the introduction, we do assume throughout the paper that all the functions f_i are differentiable (in fact, in view of (A$_1$), local differentiability in the neighborhood of x^∞ would be sufficient). With this respect, the assumption $g^k \to 0$ made in (A$_1$) is natural; for example, when the parametrization $(r_k)_k$ is slow, this amounts to follow the central trajectory $(x(r))_r$. Moreover, this hypothesis is compatible with (A$_2$,A$_3$) in the sense that since $g^k \to 0$, we may assume that $(h_k)_k$ is bounded from below by a positive constant.

4.2. **Associated Approximate Multipliers.** In a similar way to formula (5), we can associate with the primal sequence $(x^k)_k$ obtained by iterating (PPPA) the sequence $(\lambda^k)_k$ given by

$$\lambda_i^k := \theta'\left(f_i(x^k + e^k)/r_k\right), \quad i = 1, ..., m. \tag{15}$$

Notice that for any k the vector λ^k is well defined because $g^k \in \partial_{\varepsilon_k} f(x^k + e^k, r_k)$, so that $x^k + e^k$ is in the domain of $f(\cdot, r_k)$. In our asymptotic analysis of the sequence $(\lambda^k)_k$, we shall make the following further assumption

$$\frac{\varepsilon_k}{r_k} \to 0 \quad \text{as} \quad k \to +\infty \tag{A$_4$}$$

It follows from that last assumption that the study of $(\lambda^k)_k$ reduces to the study of another sequence $(\mu^k)_k$ which in turn is associated to an iterative scheme to which the techniques developed in §2 and §3 apply, as the following shows.

Lemma 4.1. *Assume that* (A$_1$)-(A$_4$) *hold, then there exists a sequence* $(y^k)_k$ *such that*

$$y^k \;=\; y^{k-1} - h_k \nabla_x f(y^k, r_k) + \xi^k \tag{16}$$

together with

$$\nabla_x f(y^k, r_k) \to 0, \quad \xi^k \to 0, \quad y^k \to x^\infty, \quad \left\| \frac{y^k - x^\infty}{r_k} - \frac{x^k + e^k - x^\infty}{r_k} \right\| \to 0. \tag{17}$$

Moreover, if we set

$$\mu_i^k := \theta'\left(f_i(y^k)/r_k\right), \quad i = 1, ..., m$$

then $(\mu^k)_k$ *and* $(\lambda^k)_k$ *have the same asymptotic behaviour.*

Proof. When $\varepsilon_k > 0$, by the Brøndsted-Rockafellar Theorem [24], there exists y^k such that

$$\|x^k + e^k - y^k\| \le \sqrt{\varepsilon_k\, r_k} \quad \text{ and } \quad \|g^k - \nabla_x f(y^k, r_k)\| \le \sqrt{\frac{\varepsilon_k}{r_k}}. \tag{18}$$

When $\varepsilon_k = 0$, we set $y^k := x^k + e^k$. Then (16) and (17) follow from (PPPA) as well as (A$_1$)-(A$_4$). Finally, by convexity of the functions f_i, it holds

$$\langle \nabla f_i(y^k)\,,\, \frac{x^k + e^k - y^k}{r_k} \rangle \;\le\; \frac{f_i(x^k + e^k) - f_i(y^k)}{r_k} \;\le\; \langle \nabla f_i(x^k + e^k)\,,\, \frac{x^k + e^k - y^k}{r_k} \rangle.$$

for every $i \in \{1, \dots, m\}$, from which we infer that $(\mu^k)_k$ and $(\lambda^k)_k$ have the same asymptotic behaviour. □

With this lemma in hand, we can state the following result which illustrates why $(\lambda^k)_k$ may be considered as a *multiplier sequence* associated with the primal sequence $(x^k)_k$.

Proposition 4.1. *Assume that* (S) *and* (H$_1$) *hold as well as* (A$_1$)-(A$_4$)*. Then the dual sequence* $(\lambda^k)_k$ *given by* (15) *is bounded and* $\operatorname{dist}(\lambda^k, S(D)) \to 0$ *as* $k \to +\infty$.

Proof. Since the sequences $(\lambda^k)_k$ and $(\mu^k)_k$ (as obtained in Lemma 4.1) have the same asymptotic behavior, it suffices to prove the claims for the sequence $(\mu^k)_k$. In fact, since (S), (H$_1$), (A$_2$) and (17) hold, we may simply apply Lemma 2.1 to $(\mu^k)_k$, which concludes the proof. □

4.3. Convergence to the θ^*-Center.

We are now in position to state and prove the convergence result for the sequence of multipliers associated to (PPPA) defined in (15). The following convergence result generalizes that of [9, 3], which was devoted to the very restrictive case of Linear Programming. Here we only require that the functions f_i be convex and (locally) differentiable around the limit point x^∞.

Theorem 4.1. *Let* (S), (H$_1$) *as well as* (A$_1$)-(A$_4$) *hold. We also assume that* $(\frac{r_k}{r_{k-1}})_{k\ge1}$ *be bounded. Then, the dual sequence* $(\lambda^k)_k$ *defined by* (15) *converges to the* θ^**-center* λ^{θ^*} *as* $k \to +\infty$.

Proof. As in the proof of Proposition 4.1, we prove the claim for the sequence $(\mu^k)_k$ obtained via Lemma 4.1. As in Lemma 3.3, we define

$$w^k := \frac{y^k - x^\infty}{r_k}, \quad F_k := \operatorname{Span}\{\nabla f_i(y^k) \mid i \in \{0\} \cup I\}, \quad F_\infty := \operatorname{Span}\{\nabla f_i(x^\infty) \mid i \in I\}$$

and set $w_k^k := \operatorname{Proj}(w^k, F_k)$ and $w_\infty^k := \operatorname{Proj}(w^k, F_\infty)$. Thanks to (S), (H$_1$) and (A$_1$)-(A$_4$), Theorem 4.1 follows from Theorem 3.1 if we manage to show that the following analogue of condition (C) holds

$$\text{the ratio } \frac{f_0(y^k) - f_0(x^\infty)}{r_k} \text{ is bounded from above.}$$

To this end, by convexity of f_0, it is sufficient to prove that the sequence $(w^k_k)_k$ is bounded. Notice that by Lemma 3.3 this is equivalent to prove that the sequence $(w^k_\infty)_k$ is bounded, which is the goal of the rest of this proof. We shall then introduce the family of functions $\Psi^k \colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ defined as

$$\Psi^k(v) = \frac{f_0(x^\infty + r_k\, v) - f_0(x^\infty)}{r_k} + \sum_{i \in I} \theta\left(\frac{f_i(x^\infty + r_k\, v) - f_i(x^\infty)}{r_k}\right).$$

Note that it follows from (A$_2$) that for all $v \in F_\infty$

$$\Psi^k(v) \to \Psi(v) \;:=\; \langle \nabla f_0(x^\infty), v\rangle + \sum_{i \in I} \theta(\langle \nabla f_i(x^\infty), v\rangle) \tag{19}$$

where, by analogy with Lemma 3.2, we denote by v^∞ the unique minimizer of Ψ over F_∞.

Step 1. We claim that for all k it holds

$$\frac{1}{2}\|w^k_\infty - v^\infty\|^2 - \frac{1}{2}\|w^{k-1}_\infty - v^\infty\|^2 \;\le\; \frac{h_k}{r_k}\left[\Psi(v^\infty) - \Psi(w^k_\infty) + \delta_k\,(1 + \|w^k_\infty - v^\infty\|)\right] \tag{20}$$

with $\delta_k \to 0$ as $k \to +\infty$. We first compute

$$\begin{aligned}
\frac{1}{2}\|w^k_\infty - v^\infty\|^2 - \frac{1}{2}\|w^{k-1}_\infty - v^\infty\|^2 \;&=\; \langle w^k_\infty - w^{k-1}_\infty, \frac{w^k_\infty + w^{k-1}_\infty}{2} - v^\infty\rangle \\
&=\; \langle w^k_\infty - w^{k-1}_\infty, w^k_\infty - v^\infty\rangle - \frac{1}{2}\|w^k_\infty - w^{k-1}_\infty\|^2 \\
&\le\; \langle w^k_\infty - w^{k-1}_\infty, w^k_\infty - v^\infty\rangle \;=\; \langle w^k - w^{k-1}, w^k_\infty - v^\infty\rangle
\end{aligned}$$

where we used $w^k_\infty - v^\infty \in F_\infty$. We now infer from (16) and (17) that

$$\begin{aligned}
w^k - w^{k-1} \;&=\; \frac{1}{r_k}(y^k - y^{k-1}) + (\frac{1}{r_k} - \frac{1}{r_{k-1}})(y^{k-1} - x^\infty) \\
&=\; -\frac{h_k}{r_k}(\nabla_x f(y^k, r_k) + \xi^k) + \frac{h_k}{r_k} d^k_1 \;=\; -\frac{h_k}{r_k}\nabla\Psi^k(w^k) + \frac{h_k}{r_k} d^k_2
\end{aligned}$$

where $d^k_1 := \frac{1}{h_k}(1 - \frac{r_k}{r_{k-1}})(y^{k-1} - x^\infty) \to 0$ and $d^k_2 \to 0$ as $k \to +\infty$. We also infer from the fact that $\nabla\Psi^k(w^k) \in F_k$, from the convexity of Ψ, from Lemma 3.3(i) and the fact that $\nabla\Psi^k(w^k) = \nabla_x f(y^k, r_k) \to 0$ that the following holds:

$$\begin{aligned}
\langle -\nabla\Psi^k(w^k), w^k_\infty - v_\infty\rangle \;&=\; \langle \nabla\Psi^k(w^k), v^\infty - w^k_k\rangle + \langle \nabla\Psi^k(w^k), w^k_k - w^k_\infty\rangle \\
&\le\; \langle \nabla\Psi^k(w^k), v^\infty - w^k\rangle + \|\nabla\Psi^k(w^k)\|\,\|w^k_k - w^k_\infty\| \\
&\le\; \Psi^k(v^\infty) - \Psi^k(w^k) + \delta^1_k\,(1 + \|w^k_\infty - v^\infty\|)
\end{aligned}$$

with $\delta^1_k \to 0$ as $k \to +\infty$. By the convexity of the functions f_i it comes that

$$\forall k, \qquad -\Psi^k(w^k) \;\le\; -\Psi(w^k) \;=\; -\Psi(w^k_\infty).$$

Moreover, since by (19) one has $\Psi^k(v^\infty) \to \Psi(v^\infty)$ as $k \to +\infty$, the preceding computations yield

$$\begin{aligned}
\frac{1}{2}\|w^k_\infty - v^\infty\|^2 - \frac{1}{2}\|w^{k-1}_\infty - v^\infty\|^2 \;&\le\; \frac{h_k}{r_k}\left[\Psi^k(v^\infty) - \Psi(w^k_\infty) + (\delta^1_k + \|d^k_2\|)\,(1 + \|w^k_\infty - v^\infty\|)\right] \\
&\le\; \frac{h_k}{r_k}\left[\Psi(v^\infty) - \Psi(w^k_\infty) + (\delta^2_k + \delta^1_k + \|d^k_2\|)\,(1 + \|w^k_\infty - v^\infty\|)\right]
\end{aligned}$$

with $\delta^2_k \to 0$ as $k \to +\infty$, so that (20) follows.

Step 2. It follows from Lemma 3.2 that Ψ is coercive on F_∞, so that

$$\gamma \;:=\; \frac{1}{2}\inf\left\{\frac{\Psi(v)-\Psi(v^\infty)}{\|v-v^\infty\|}\,|\, v\in F_\infty,\ \|v-v^\infty\|\geq 1\right\} \;>\; 0.$$

As a consequence, we infer from (20) that for k large enough it holds

$$\|w_\infty^k - v^\infty\| \geq 1 \quad\text{implies}\quad \frac{1}{2}\|w_\infty^k - v^\infty\|^2 - \frac{1}{2}\|w_\infty^{k-1}-v^\infty\|^2 \;\leq\; -\frac{h_k}{r_k}\gamma\,.$$

It then follows from (A_2) and (A_3) that $\|w_\infty^k - v^\infty\|\leq 1$ for k large enough, so that the sequence $(w_\infty^k)_k$ is bounded, and the proof is complete. □

Remark 4.2. Note that the hypothesis that $(\frac{r_k}{r_{k-1}})_{k\geq 1}$ is a bounded sequence in the Theorem above is weaker than requiring $(r_k)_k$ to be nonincreasing. More precisely, in the setting of Linear Programming, a similar dual convergence result is proved in [9] for the particular case where $\theta\equiv\exp(\cdot)$ is the exponential penalty, and replacing (A_3) by the more general $\sum h_k = +\infty$, together with the following additional hypotheses: $(r_k)_k$ is nonincreasing, $(\frac{r_{k-1}-r_k}{r_{k-1}\,h_k})$ is bounded, $\frac{\varepsilon_k}{h_k}\to 0$, and $e^k\equiv 0$ in (A_2). In [3], and still in the setting of Linear Programming, two dual convergence results are proved for a wide class of penalty barrier functions. On the one hand, assuming that the penalty θ is bounded from below, the convergence result is obtained under the same hypotheses on the parameters as in [9]. On the other hand, allowing θ to be unbounded from below, dual convergence is proved with the additional assumptions that $\varepsilon_k\equiv 0$ and $\eta^k\equiv e^k\equiv 0$ in (A_2). Notice that in Theorem 4.1, we assume that h_k is bounded away from zero, but we have proved the convergence including the parameters $\eta^k, e^k\in\mathbb{R}^d$, $\varepsilon_k\geq 0$, and for general penalty functions θ.

Remark 4.3. It follows from Lemma 4.1 and the proof of Theorem 3.1 (or by refining the end of the argument of the proof of Theorem 4.1 above) that the following asymptotic expansion holds

$$\mathrm{Proj}(x^k+e^k, F_\infty) \;=\; \mathrm{Proj}(x^\infty, F_\infty) \;+\; r_k v^\infty \;+\; o(r_k)$$

with $F_\infty := \mathrm{Span}\{\nabla f_i(x^\infty) : i\in I\}$. This is indeed a direct consequence of the convergence of $(w_\infty^k)_k$ to v^∞. One of the main difficulties in the above proof, as well as that of Theorem 3.1, is that the sequence $(\frac{x^k+e^k-x^\infty}{r_k})_k$, or equivalently the sequence $(w^k)_k$, is necessarily unbounded in $\mathbb{R}^d$ when strict complementarity does not hold. This is illustrated by the toy example of subsection 4.4 below, and follows from the fact that, for any index i for which $f_i(x^\infty)=\lambda_i^{\theta^*}=0$, one has

$$\langle\nabla f_i(x^\infty), \frac{x^k+e^k-x^\infty}{r_k}\rangle \;\leq\; \frac{f_i(x^k+e^k)-f_i(x^\infty)}{r_k} \;=\; \theta'^{-1}(\lambda_i^k) \;\to\; -\infty.$$

4.4. **An Illustrative Toy Example.** Consider the following simple example in dimension two:

$$(P_{ex}) \qquad \min\left\{x_1 \mid x_1^2+(x_2-1)^2\leq 1,\quad -2\,x_1\leq 2,\quad x_2-x_1\leq 2\right\},$$

that is

$$f_0(x)=x_1;\ f_1(x)=x_1^2+(x_2-1)^2-1;\ f_2(x)=-2\,x_1-2,\ \text{ and }\ f_3(x)=x_2-x_1-2.$$

It is easy to see that $S(P_{ex})=\{x^\infty\}$ with $x^\infty := (-1,1)$. Notice that the three constraints are active at the optimal solution. Therefore the KKT system reads

$$\begin{cases} (1,0)+\lambda_1\,(-2,0)+\lambda_2\,(-2,0)+\lambda_3\,(-1,1)=0, \\ \hfill \lambda_1,\lambda_2,\lambda_3\geq 0. \end{cases}$$

The dual optimal set is given by the segment $S(D_{ex}) = \left[(\frac{1}{2},0,0),(0,\frac{1}{2},0)\right]$, so that $I = \{1,2\}$ and $F_\infty = \mathrm{Span}\{\nabla f_1(x^\infty), \nabla f_2(x^\infty)\} = \mathbb{R} \times \{0\}$.

Take $\theta = \exp(\cdot)$ so that $\theta^*(\lambda) = \lambda \log \lambda - \lambda$ if $\lambda \geq 0$ and ∞ otherwise. Let $(x^k)_k$ be generated by (PPPA). Under the hypotheses of Theorem 4.1, we infer that the associated dual path $(\lambda^k)_k$ converges to the θ^*-center, which is easy to compute in this case; indeed, we get $\lambda^{\theta^*} = \left(\frac{1}{4},\frac{1}{4},0\right)$. On the other hand, the sequence $(\mathrm{Proj}(w^k, F_\infty))_k$ converges to $v^\infty := (\ln 2, 0)$, the unique minimizer in $F_\infty = \mathbb{R} \times \{0\}$ of

$$\Psi(v) \ = v_1 + 2\exp(-2v_1).$$

This implies that

$$w_1^k \ \to \ln 2 \quad \text{ and } \quad w_2^k \ \to -\infty$$

where the behaviour of $(w_2^k)_k$ follows from the fact that the Lagrange multiplier for f_3 is always 0 and then $\lambda_3^k \to 0$ as $k \to +\infty$; see Remark 4.3.

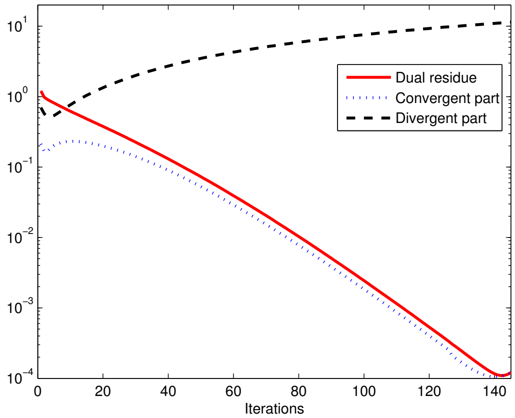

FIGURE 1. Toy example

In Figure 1, we report some numerical results obtained on this toy example, giving a particular emphasis on the convergence of the dual sequence $(\lambda^k)_k$ and on the sequences of coordinates $(w_1^k)_k$ and $(w_2^k)_k$. The continuous line (—) denotes the evolution of the dual relative residues given by $\|\lambda^k - \lambda^\theta\|/\|\lambda^\theta\|$, the dotted line $(\cdots)$ denotes the convergent direction $\left|w_1^k - \ln 2\right| / \ln 2$, finally, the dashed line (– –) denotes the divergent part $\left|w_2^k\right|$. These results were obtained using the proximal method with starting point $x^0 = (0,0)$ and prox parameter $h_k \equiv 1$, that is solving successively

$$x^{k+1} \ = \ \operatorname*{argmin}_{x \in \mathbb{R}^2} \left\{ f_0(x) + r_k \sum_{i=1}^{3} \exp\left(\frac{f_i(x)}{r_k}\right) + \frac{1}{2}\|x - x^k\|^2 \right\}. \qquad (21)$$

These problems were solved using the MATLAB[1] routine *fminunc* for unconstrained optimization, taking the sequence $r_k = 1/(1.1)^k$ with global stopping rule $r_k < 10^{-6}$. We note that routine *fminunc* uses a trust region method for each minimization problem in (21). We take x^k as a starting point of the algorithm to compute x^{k+1}, this routine stops when the gradient is small enough or when the steps between successive inner iterations are small. In our case, we take the tolerance 10^{-10} as a stopping value of inner iterations. It would be interesting to take

[1] MATLAB is a registered trademark of The MathWorks, Inc.

into account an inner stopping rule depending on r_k, in order to explore the influence of this parameter to compute approximately a minimizer.

5. Concluding Remarks

We believe that an interesting open problem be how to extend the dual convergence analysis developed here to cover the nonsmooth case, where standard gradients $\nabla f_i(x) \in \mathbb{R}^n$ are replaced with subgradients in the corresponding approximate subdifferentials $\partial_\varepsilon f_i(x) \subset \mathbb{R}^n$ for some $\varepsilon > 0$. Concerning the general result given by Theorem 3.1, the main technical difficulty appears to be a non-differentiable version of Lemma 3.3 related to the asymptotic behavior of the vector subspaces spanned by subgradients associated with the active constraints at x^∞. Similarly, our proof of Theorem 4.1 for PPPA mainly relies on the study of the sequence $(w^k_\infty)_k$ and the convergence of the functions Ψ^k to Ψ on F_∞, which also requires the differentiability of the functions f_i at x^∞. As a consequence, relaxing the differentiability hypothesis may involve quite a different approach.

Concerning the underlying algorithm used to find an approximate solution x^k of the primal problem, notice that one may consider the following global stopping rule: compute the explicit dual sequence λ^k given by (5) and then test whether the pair (x^k, λ^k) satisfies or not a relaxed version of the KKT system. For instance, we may iterate the algorithm until

$$\text{either } \|\nabla f_0(x^k)\| \leq \delta \quad \text{or} \quad \max\left\{ \frac{\|\nabla_x L(x^k, \lambda^k)\|}{\|\nabla f_0(x^k)\|}, \max_i\{f_i(x^k)\}, \max_i\{\left|\lambda_i^k f_i(x^k)\right|\}\right\} \leq \delta, \quad (22)$$

for a fixed tolerance $\delta > 0$. In the criterion above, the division by $\|\nabla f_0(x^k)\|$ is intended to normalize the equation $\nabla_x L(x^k, \lambda^k) = 0 \Leftrightarrow \nabla f_0(x^k) = -\sum_{i=1}^m \lambda_i^k \nabla f_i(x^k)$. Under the hypotheses of Lemma 2.1, such a process stops after a finite number of iterations, providing thus a theoretical basis for this stopping rule. Moreover, the full convergence of $(\lambda^k)_k$ should prevent oscillating behaviors in the computations of $(x^k)_k$ and $(\lambda^k)_k$ itself. This is particular relevant when the penalty parameter r_k is updated using a feedback rule based on the current information.

On the other hand, purely primal methods consider a suitable reformulation of the first-order condition (2) to find an approximate primal solution x^k, from which the direct dual sequence $\lambda_i^k = \theta'(f_i(x^k)/r_k)$ is obtained, while penalty primal-dual algorithms try to solve the system (3)-(4) simultaneously in the primal-dual pair. Under appropriate second-order conditions, a natural idea is to use Newton's iterations to solve these nonlinear systems approximately. But these type of systems become increasingly ill-conditioned as $r_k \searrow 0$, impairing the performance of Newton's method. To overcome such a difficulty, some *active-set identification* techniques have been proposed to reformulate the equation (2), or the system (3)-(4), to get better conditioned equivalent equations before applying Newton's method; see, for instance, [25, 26, 27]. The basic idea is to use the current dual sequence to predict the actual set I of active constraints. Again, under suitable conditions, full dual convergence ensure the identification technique to be asymptotically stable. Moreover, one could consider a global phase based on penalty methods and then use the current dual information given by (5) to switch to some separate local equality-constrained phase based on active-set identification [28, 29, 30]. But a complete numerical study is necessary to evaluate the practical performance of these kind of hybrid strategies, specially when they are compared with other techniques that one may apply for the global phase such as the augmented Lagrangian method.

Finally, it would be interesting to perform some numerical experiences changing the parameters of (14), testing the behavior of the algorithm when we take, for instance, a constant small value of r_k, or a penalty method without proximal regularization. In the case of the toy example in Section 4.4, we can confirm that taking a constant small value for r_k may produce numerical instabilities, but this example is not enough to compare the performance of PPPA with pure penalization methods or other algorithms. Therefore, it would be necessary to make this comparisons for several examples of medium and large scale problems. All these numerical studies are beyond the scope of the paper but we will certainly address them in future works.

References

1. Alvarez, F.: Absolute minimizer in convex programming by exponential penalty. J. Convex Anal. **7**(1), 197–202 (2000)
2. Champion, T.: Tubularity and asymptotic convergence of penalty trajectories in convex programming. SIAM J. Optim. **13**(1), 212–227 (2002)
3. Cominetti, R., Courdurier, M.: Coupling general penalty schemes for convex programming with the steepest descent and the proximal point algorithm. SIAM J. Optim. **13**(3), 745–765 (electronic) (2003)
4. Fiacco, A.V., McCormick, G.P.: Nonlinear programming: Sequential unconstrained minimization techniques, volume 4 of Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition (1990)
5. Gonzaga, C.G.: Path-following methods for linear programming. SIAM Review, **34**(2), 167–224 (1992)
6. Auslender, A., Cominetti, R., Haddou, M.: Asymptotic analysis for penalty methods in convex and linear programming. Math. Oper. Res., **22**(1), 43–62 (1997)
7. Nocedal, J., Wright, S.J.: Numerical optimization. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition (2006)
8. Alvarez, F., Carrasco, M., Pichard, K.: Convergence of a hybrid projection-proximal point algorithm coupled with approximation methods in convex optimization. Math. Oper. Res. **30**(4), 966–984 (2005)
9. Alvarez, F., Cominetti, R.: Primal and dual convergence of a proximal point exponential penalty method for linear programming. Math. Prog. **93**(1, Ser. A), 87–96 (2002)
10. Cominetti, R.: Coupling the proximal point algorithm with approximation methods. J. Optim. Theory Appl., **95**(3), 581–600 (1997)
11. Gonzaga, C.G., Castillo, R.A.: A nonlinear programming algorithm based on non-coercive penalty functions. Math. Program., **96**(1, Ser. A), 87–101 (2003)
12. Gilbert, J.C., Gonzaga, C.G., Karas, E.: Examples of ill-behaved central paths in convex optimization. Math. Program., **103**(1, Ser. A) 63–94 (2005)
13. Boyd, S., Vandenberghe, L.: Convex optimization. Cambridge University Press, Cambridge (2004)
14. Rockafellar, R.T.: Convex analysis. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ (1997) Reprint of the 1970 original, Princeton Paperbacks
15. Attouch, H.: Viscosity solutions of minimization problems. SIAM J. Optim. **6**(3), 769–806 (1996)
16. Iusem A.N., Svaiter, B.F., Da Cruz Neto, J.X.: Central paths, generalized proximal point methods, and Cauchy trajectories in Riemannian manifolds. SIAM J. Control Optim. **37**(2), 566–588 (1999)
17. Auslender, A., M. Teboulle, M.: Asymptotic Cones and Functions in Optimization and Variational Inequalities. Springer Monographs in Mathematics, New York (2003)
18. Auslender, A., Crouzeix, J.P., Fedit, P.: Penalty-proximal methods in convex programming. J. Optim. Theory Appl. **55**, 1–21 (1987)
19. Kaplan, A.: On a convex programming method with internal regularization. Soviet Mathematics Doklady **19**, 795–799 (1978)
20. Kaplan, A., Tichatschke, R.: Proximal point methods in view of interior-point strategies. J. Optim. Theory Appl. **98**(2), 399–429 (1998)
21. Konnov, I.V.: Combined relaxation methods for the search for equilibrium points and solutions of related problems. Izv. Vyssh. Uchebn. Zaved. Mat. **37**(2), 46–53 (1993)
22. Solodov, M.V., Svaiter, B.F.: A hybrid projection-proximal point algorithm. J. of Convex Analysis **6**(1), 59–70 (1999)
23. Solodov, M.V., Svaiter, B.F.: A hybrid approximate extragradient-proximal point algorithm using the enlargement of a maximal monotone operator. Set-Valued Anal. **7**(4), 323–345 (1999)
24. Brøndsted, A., Rockafellar, R. T.: On the subdifferentiability of convex functions. *Proc. Amer. Math. Soc.* **16**, 605–611 (1965)
25. Cominetti, R., Dussault, J.-P.: Stable exponential-penalty algorithm with superlinear convergence. J. Optim. Theory Appl. **83**(2), 285–309 (1994)
26. Cominetti, R., Pérez-Cerda, J. M.: Quadratic rate of convergence for a primal-dual exponential penalty algorithm. Optimization **39**(1), 13–32, (1997)

27. Dussault, J.-P.: Numerical stability and efficiency of penalty algorithms. SIAM Journal on Numerical Analysis **32**(1), 296–317 (1995)
28. Facchinei, F., Fischer, A., Kanzow, C.: On the accurate identification of active constraints. SIAM J. Optim. **9**(1), 14–32 (electronic) (1999)
29. Oberlin, C., Wright, S.J.: Active set identification in nonlinear programming. SIAM J. Optim. **17**(2), 577–605 (electronic) (2006)
30. Wright, S.J.: An algorithm for degenerate nonlinear programming with rapid local convergence. SIAM J. Optim. **15**(3), 673–696 (electronic) (2005)