On the convergence of the exponential multiplier method for convex programming

Paul Tseng

Department of Mathematics, University of Washington, Seattle, WA, USA

Dimitri P. Bertsekas

Department of Electrical Engineering and Computer Science, M.I.T., Cambridge, MA, USA

In this paper, we analyze the exponential method of multipliers for convex constrained minimization problems, which operates like the usual Augmented Lagrangian method, except that it uses an exponential penalty function in place of the usual quadratic. We also analyze a dual counterpart, the entropy minimization algorithm, which operates like the proximal minimization algorithm, except that it uses a logarithmic/entropy "proximal" term in place of a quadratic. We strengthen substantially the available convergence results for these methods, and we derive the convergence rate of these methods when applied to linear programs.

Key words: Convex programming, linear programming, multiplier method, exponential penalty, Augmented Lagrangian.

1. Introduction

Let $f : \mathbb{R}^n \mapsto (-\infty, \infty]$ and $g_j : \mathbb{R}^n \mapsto (-\infty, \infty], j = 1, \ldots, m$, be closed, proper, convex functions in $\mathbb{R}^n$, the n-dimensional Euclidean space. Consider the following convex program associated with f and the g_j's:

(P)   minimize $f(x)$

          subject to $g_j(x) \leq 0, \quad j = 1, \ldots, m.$ (1.1)

We make the following standing assumption about (P):

Assumption A. (a) The optimal solution set for (P) is nonempty and bounded.

(b) The effective domain of f, that is, the set $\{x \mid f(x) < \infty\}$ is contained in the effective domain $\{x \mid g_j(x) < \infty\}$ of each g_j. Furthermore, the relative interior of the effective domain of f is contained in the relative interior of the effective domain of each g_j.

(c) There exists a vector $\bar{x}$ in the relative interior of the effective domain of f, which satisfies $g_j(\bar{x}) < 0$ for all non-affine g_j.

The boundedness assumption in part (a) of Assumption A will be needed to ensure that our method is well-defined. Part (b) of Assumption A is satisfied in particular if all the constraint functions are real-valued. Parts (b) and (c) of Assumption A are constraint qualification conditions, which are needed to guarantee the existence of a Kuhn–Tucker vector for the problem (see [25, p. 277]).

We now describe the exponential multiplier method proposed by Kort and Bertsekas [15] for solving problem (P) (see also [5, Section 5.1.2]). Let $\psi : \mathbb{R} \to \mathbb{R}$ be the exponential penalty function given by

$$\psi(t) = e^t - 1. \tag{1.2}$$

We associate a multiplier $\mu_j > 0$ with the jth constraint. The method performs a sequence of unconstrained minimizations, and iterates on the multipliers at the end of each minimization. At the kth iteration ($k \geq 0$) we are given positive μ_j^k, $j = 1, \ldots, m$ (with the initial $\mu_j^0, j = 1, \ldots, m$, chosen arbitrarily); we compute x^k as

$$x^k \in \arg\min_{x \in \mathbb{R}^n} \left\{ f(x) + \sum_{j=1}^{m} \frac{\mu_j^k}{c_j^k} \psi(c_j^k g_j(x)) \right\}, \tag{1.3}$$

where each c_j^k is a positive penalty parameter, and then we update the multipliers according to

$$\mu_j^{k+1} = \mu_j^k \nabla \psi(c_j^k g_j(x^k)) = \mu_j^k e^{c_j^k g_j(x^k)}, \quad j = 1, \ldots, m. \tag{1.4}$$

Notice that for a fixed $\mu_j^k > 0$, as $c_j^k \to \infty$, the "penalty" term $(\mu_j^k / c_j^k) \psi(c_j^k g_j(x))$ tends to ∞ for all infeasible x ($g_j(x) > 0$) and to zero for all feasible x ($g_j(x) \leq 0$). On the other hand, for a fixed c_j^k, as $\mu_j^k \to 0$ (which is expected to occur if the jth constraint is inactive at the optimum), the penalty term goes to zero for all x, feasible or infeasible. This is contrary to what happens in usual exterior penalty methods [11, 17], and for this reason, much of the standard analysis for exterior penalty and multiplier methods cannot be applied to the exponential method of multipliers.

It can be shown that the minimum in (1.3) is attained for all k (see [5, p. 337]). For a brief justification, note that if this minimum were not attained, then f and the functions g_j would share a direction of recession, in which case the optimal solution set of (P) is unbounded (see [25, Section 8]), thus contradicting Assumption A.

We will consider two rules for choosing the penalty parameters c_j^k. In the first rule, which is common in muliplier methods, the c_j^k's are independent of j and are bounded from below, that is,

$$c_j^k = \omega^k \quad \forall k, \tag{1.5a}$$

where $\{\omega^k\}$ is some sequence of positive scalars satisfying

$$w^k \geq \bar{\omega} \quad \forall k, \tag{1.5b}$$

with $\bar{\omega}$ a fixed positive scalar. Note that with this rule, we can still provide for different penalization of different constraints, by multiplying the constraints with different scaling constants at the start of the computation.

In the second rule, the penalty parameters depend on the current values of the multipliers, becoming larger as these multipliers become smaller; for inactive constraints for which the associated multipliers tend to zero, the corresponding penalty parameters tend to infinity. In particular, each c_j^k is set inversely proportional to μ_j^k, that is,

$$c_j^k = c/\mu_j^k \quad \forall j, \tag{1.6}$$

where c is a fixed positive constant. The second rule is interesting because for linear programs, it leads to a superlinear rate of convergence, even though the penalty parameters corresponding to active constraints with positive multipliers remain bounded.

The principal motivation for the exponential method of multipliers is that in contrast with the usual quadratic Augmented Lagrangian function for inequality constraints [26], the minimized function in (1.3) is twice differentiable if the functions f and g_j are. As a result, Newton-type methods can be used for the corresponding unconstrained minimization more effectively, and with guaranteed superlinear convergence. This is not just a theoretical advantage; in the experience of the second author, serious difficulties arise with Newton's method when the usual quadratic Augmented Lagrangian method is used to solve linear programs [4]. By contrast, the exponential multiplier method has been used to solve fast and with consistency very large linear programs arising in production scheduling of power systems [1, 16]; simplex methods as well as the more recent interior point methods are unsuitable for the solution of these problems.

Some aspects of the convergence analysis of the exponential multiplier method have proved surprisingly difficult, even though the method has been known to be reliable in practice [1]. For nonconvex problems under second order sufficiency conditions, convergence can be analyzed using fairly standard techniques; see [22]. However, for convex problems, the sharpest result available so far, due to Kort and Bertsekas, and given in [5, p. 336], assumes (in addition to Assumption A) a mild but fairly complicated and hard to verify assumption, and asserts that when the penalty parameters c_j^k are selected according to the first rule (1.5), all cluster points of $\{\mu^k\}$ are optimal solutions of an associated dual problem. One of the contributions of the present paper, is to show using an unusual proof technique, that the entire sequence $\{\mu^k\}$ converges to an optimal solution of the dual problem, without assuming the complex assumption of [5]. The corresponding sequence $\{x^k\}$ is shown to approach optimality in an ergodic sense. As an indication of the difficulty of the analysis, we note that we have been unable to show a corresponding result when c_j^k is selected according to the second rule (1.6), even though the method in practice seems equally reliable with the rule (1.5) or (1.6).

A second contribution of the present paper is the analysis of the convergence rate of the exponential method of multipliers as applied to linear programs. The usual quadratic Augmented Lagrangian method converges in a finite number of iterations for linear programs, as shown independently by Poljak and Tretjakov [23], and Bertsekas [3] (see also [5, Section 5.4]). This is not true for the exponential method of multipliers, but we show that the rate of convergence is linear for the penalty parameter selection rule (1.5a), and quadratic for the rule (1.6).

It has been shown by Rockafellar [27] that when the quadratic Augmented Lagrangian method is dualized using the Fenchel duality theorem, one obtains the proximal minimization algorithm of Martinet [21], which is a special case of the proximal point algorithm of Rockafellar [28]. By similarly dualizing the exponential method of multipliers one obtains a method, called *entropy minimization algorithm*, which involves a logarithmic/entropy "proximal" term; see Section 2. The entropy minimization algorithm is mathematically equivalent to the exponential method of multipliers, so it is covered by our convergence results. This equivalence is also used in a substantial way in our analysis similar to several past works, which have studied nonquadratic versions of Augmented Lagrangian, proximal minimization, and proximal point algorithms [5, 12, 15, 18, 19].

Several recent works have also drawn attention to nonquadratic proximal point algorithms and the entropy minimization algorithm in particular. In particular, Censor and Zenios [7], have proposed a broad class of algorithms generalizing the proximal minimization algorithm by using Bregman functions. Eckstein [10] has generalized in an analogous manner the proximal point algorithm; see also [13]. None of these works provides a convergence or rate of convergence result for the exponential method of multipliers or its equivalent entropy minimization algorithm, although some of the analysis of [7] and [10] was helpful to us (see Section 3).[1]

Regarding notation, all our vectors are column vectors, and superscript "T" denotes transposition. For a function $h : \mathbb{R}^n \mapsto \mathbb{R}$, we denote by $\nabla h(x)$ and $\partial h(x)$ the gradient and the subdifferential of h at the vector x, respectively. For any set S and any positive integer m, we denote by S^m the m-fold Cartesian product of S with itself.

2. The entropy minimization algorithm

In this section we focus on the dual interpretation of the exponential mutiplier method (1.3)–(1.4), as worked out in [5, pp. 315–327]. Let $d : [0, \infty)^m \to [-\infty, \infty)$ be

---

[1] While this paper was under review, convergence results for the dual sequence $\{\mu^k\}$, which are similar to ours have been obtained by Censor and Zenios in a revision of their paper [7], and by Chen and Teboulle [8] by using different methods of analysis. These works have not considered rate of convergence issues or the convergence of the primal sequence $\{x^k\}$.

the dual functional associated with (P) given by

$$d(\mu) = \min_{x \in \mathbb{R}^n} \left\{ f(x) + \sum_{j=1}^{m} \mu_j g_j(x) \right\}. \tag{2.1}$$

The function d is closed, proper, and concave under Assumption A, and is the cost function of the dual problem of (P), given by

(D)   maximize $d(\mu)$

              subject to $\mu \geq 0$.

The weak duality theorem, asserts that the value $d(\mu)$ of any dual feasible vector μ is less than or equal to the cost $f(x)$ of any primal feasible vector x. Assumption A implies that there is no duality gap, that is, the optimal value of (D) is equal to f^*, the optimal cost of (P); furthermore, there exists a dual optimal solution (see [25, Theorem 28.2]).

The exponential method of multipliers (1.3) and (1.4) may be viewed alternatively as the following algorithm for solving the dual problem (D):

$$\mu^{k+1} = \arg \max_{\mu > 0} \left\{ d(\mu) - \sum_{j=1}^{m} \frac{\mu_j^k}{c_j^k} \psi^* \left(\frac{\mu_j}{\mu_j^k} \right) \right\}, \tag{2.2}$$

where ψ^* denotes the conjugate function of ψ, which is the entropy function

$$\psi^*(s) = s \ln(s) - s + 1. \tag{2.3}$$

It can be shown that the maximum is uniquely attained in (2.2) by using the strict convexity and differentiability of ψ^*, and the fact $\lim_{s \to 0} \nabla \psi^*(s) = \infty$.

One way to show the equivalence of the two methods is to use the Fenchel duality theorem. For a direct derivation, notice that, by definition, x^k satisfies the Kuhn–Tucker optimality conditions for the minimization in (1.3), so

$$0 \in \partial f(x^k) + \sum_{j=1}^{m} \mu_j^k \nabla \psi(c_j^k g_j(x^k)) \partial g_j(x^k).$$

(This equation can be justified by using Assumption A; see the subgradient calculus developed in [25, Section 23].) Then, from the multiplier update formula (1.4), we obtain

$$0 \in \partial f(x^k) + \sum_{j=1}^{m} \mu_j^{k+1} \partial g_j(x^k),$$

implying that x^k attains the minimum in the dual function definition (2.1), with μ set to μ^{k+1}. Hence,

$$d(\mu^{k+1}) = f(x^k) + \sum_{j=1}^{m} \mu_j^{k+1} g_j(x^k). \tag{2.4}$$

Furthermore, using the calculation

$$d(\mu) = \min_{x \in \mathbb{R}^n} \left\{ f(x) + \sum_{j=1}^{m} \mu_j g_j(x) \right\}$$

$$\leq f(x^k) + \sum_{j=1}^{m} \mu_j g_j(x^k)$$

$$= f(x^k) + \sum_{j=1}^{m} \mu_j^{k+1} g_j(x^k) + \sum_{j=1}^{m} (\mu_j - \mu_j^{k+1}) g_j(x^k)$$

$$= d(\mu^{k+1}) + \sum_{j=1}^{m} (\mu_j - \mu_j^{k+1}) g_j(x^k) \quad \forall \mu \in \mathbb{R}^m,$$

we have

$$\begin{bmatrix} g_1(x^k) \\ \vdots \\ g_m(x^k) \end{bmatrix} \in \partial d(\mu^{k+1}). \tag{2.5}$$

Also, from (1.4) and (2.3), it is seen that

$$g_j(x^k) = \frac{1}{c_j^k} \ln(\mu_j^{k+1}/\mu_j^k) = \frac{1}{c_j^k} \nabla \psi^*(\mu_j^{k+1}/\mu_j^k) \quad \forall j.$$

Combining this with (2.5) yields

$$0 \in \partial d(\mu^{k+1}) - \begin{bmatrix} \nabla \psi^*(\mu_1^{k+1}/\mu_1^k)/c_1^k \\ \vdots \\ \nabla \psi^*(\mu_m^{k+1}/\mu_m^k)/c_m^k \end{bmatrix},$$

which is precisely the Kuhn–Tucker optimality condition for μ^{k+1} to attain the maximum in (2.2).

We now derive some properties of the entropy function that will prove useful in our analysis. Let $q: [0, \infty) \times (0, \infty) \to \mathbb{R}$ be the function given by

$$q(u, v) = u \ln(u/v) - u + v. \tag{2.6}$$

Then by using the formula $\psi^*(s) = s \ln(s) - s + 1$ (cf. (2.3)), we have

$$q(u, v) = \psi^*(u/v)v$$

and the entropy minimization algorithm (2.2) can be rewritten as

$$\mu^{k+1} = \arg \max_{\mu > 0} \left\{ d(\mu) - \sum_{j=1}^{m} \frac{1}{c_j^k} q(\mu_j, \mu_j^k) \right\}. \tag{2.7}$$

The following lemma gives some useful properties of the function q.

Lemma 2.1.

   (a) $q(u, v) = \psi^*(u) - \psi^*(v) - \nabla \psi^*(v)(u - v) \quad \forall u \geq 0, \forall v > 0. \tag{2.8}$

   (b) q is nonnegative and $q(u, v) = 0$ if and only if $u = v$.

(c) *For any $\bar{u} \geq 0$ and any sequence of positive scalars $\{u^k\}, q(\bar{u}, u^k) \to 0$ if and only if $u^k \to \bar{u}$.*

(d) *For any $\bar{u} \geq 0$, the function $v \to q(\bar{u}, v)$ has bounded level sets.*

Proof. (a) Using the definitions of ψ^* and q (cf. (2.3) and (2.6)), we have

$$\psi^*(u) - \psi^*(v) - \nabla\psi^*(v)(u - v)$$
$$= u \ln(u) - u + 1 - (v \ln(v) - v + 1) - \ln(v)(u - v)$$
$$= u \ln(u) - u + v - u \ln(v)$$
$$= u \ln(u/v) - u + v$$
$$= ((u/v) \ln(u/v) - (u/v) + 1)v = q(u, v).$$

(b) Use part (a) and the strict convexity of ψ^* (cf. (2.3)).

(c) From (2.6), we have

$$q(\bar{u}, v) = \bar{u} \ln(\bar{u}) - \bar{u} \ln(v) - \bar{u} + v \quad \forall v > 0. \tag{2.9}$$

There are two cases to consider. If $\bar{u} = 0$, then (2.9) gives $q(\bar{u}, v) = v$, so the claim follows immediately. If $\bar{u} > 0$, then (2.9) shows that the function $v \to q(\bar{u}, v)$ is continuous at $\bar{u}$. The result follows from this continuity property, by using also part (b) to assert that $q(\bar{u}, v) = 0$ if and only if $v = \bar{u}$.

(d) If $v \to \infty$, then v, the last term in the right-hand side of (2.9), dominates (since the other terms in the right-hand side of (2.9) either remain constant or grow logarithmically in v), so v is bounded from above whenever the left-hand side of (2.9) is bounded from above. $\square$

Let $D : [0, \infty)^m \times (0, \infty)^m \to [0, \infty)$ be the function given by

$$D(\lambda, \mu) = \sum_{j=1}^{m} q(\lambda_j, \mu_j). \tag{2.10}$$

The following lemma is due to Bregman [6], and asserts that D has properties of a distance function, much like those enjoyed by the Euclidean distance function. The proof is included for completeness.

Lemma 2.2. (a) *D is nonnegative.*

(b) *For any fixed $\bar{\mu} \in [0, \infty)^m$ and any sequence $\{\lambda^k\} \subset (0, \infty)^m$, $D(\bar{\mu}, \lambda^k) \to 0$ if and only if $\lambda^k \to \bar{\mu}$.*

(c) *For any fixed $\bar{\mu} \in [0, \infty)^m$, the function $\mu \to D(\bar{\mu}, \mu)$ has bounded level sets.*

(d) *Let M be any closed convex subset of $[0, \infty)^m$ having a nonempty intersection with $(0, \infty)^m$. Then, for any $\bar{\mu} \in M$ and any $\mu \in (0, \infty)^m$, we have*

$$D(\bar{\mu}, \mu') \leq D(\bar{\mu}, \mu),$$

where $\mu' = \arg\min_{\lambda \in M} D(\lambda, \mu)$.

Proof. Parts (a) to (c) follow readily from the definition (2.10) of D and parts (b) to (d) of Lemma 2.1, respectively.

To prove part (d), note that from the form (2.10) of D and Lemma 2.1(a), we have

$$D(\lambda, \mu) = h(\lambda) - h(\mu) - \nabla h(\mu)^{\mathrm{T}}(\lambda - \mu), \tag{2.11}$$

where h is the function

$$h(\mu) = \sum_{j=1}^{m} \psi^*(\mu_j).$$

Since $\nabla \psi^*$ exists on $(0, \infty)$, we see from (2.11) and the definition of h that $\nabla_\lambda D(\cdot, \mu)$ (the partial derivative of D with respect to its first m arguments), exists on $(0, \infty)^m$. Since M intersects $(0, \infty)^m$ and μ' minimizes $D(\lambda, \mu)$ over all $\lambda \in M$, we see from the properties of D that $\mu' > 0$, so $\nabla_\lambda D(\mu', \mu)$ exists and

$$\nabla_\lambda D(\mu', \mu)^{\mathrm{T}}(\lambda - \mu') \geqslant 0 \quad \forall \lambda \in M.$$

Substituting $\bar{\mu}$ for λ in the above relation, we obtain

$$\nabla_\lambda D(\mu', \mu)^{\mathrm{T}}(\bar{\mu} - \mu') \geqslant 0,$$

so (2.11) yields

$$(\nabla h(\mu') - \nabla h(\mu))^{\mathrm{T}}(\bar{\mu} - \mu') \geqslant 0$$

or, equivalently,

$$-\nabla h(\mu')^{\mathrm{T}}(\bar{\mu} - \mu') \leqslant -\nabla h(\mu)^{\mathrm{T}}(\bar{\mu} - \mu) + \nabla h(\mu)^{\mathrm{T}}(\mu' - \mu).$$

Adding $h(\bar{\mu}) - h(\mu')$ to the left-hand side, and adding $h(\bar{\mu}) - h(\mu) + h(\mu) - h(\mu')$ to the right-hand side of the above relation, and then collecting terms using (2.11), we obtain

$$D(\bar{\mu}, \mu') \leqslant D(\bar{\mu}, \mu) - D(\mu', \mu).$$

Since the last term above is nonnegative by part (a), we obtain $D(\bar{\mu}, \mu') \leqslant D(\bar{\mu}, \mu)$. □

3. Convergence analysis

Let $\{x^k\}$ and $\{\mu^k\}$ be the primal and dual sequences generated by the exponential method of multipliers (1.3) and (1.4) with the penalty parameters being equal for all constraints and bounded away from zero (cf. (1.5a)–(1.5b)). We prove in this section that $\{\mu^k\}$ converges to an optimal solution of (D) (see Proposition 3.1).

3.1. Convergence of multipliers

Since $\{\mu^k\}$ is equivalently given by (2.7), we have from (1.5a) that

$$\mu^{k+1} = \arg\max_{\mu > 0} \left\{ d(\mu) - \frac{1}{\omega^k} \sum_{j=1}^{m} q(\mu_j, \mu_j^k) \right\}$$

$$= \arg\max_{\mu > 0} \left\{ d(\mu) - \frac{1}{\omega^k} D(\mu, \mu^k) \right\}, \tag{3.1}$$

where D is the function defined by (2.10) and q is defined by (2.6). From (3.1), the nonnegativity of q, and the fact $q(\mu_j^k, \mu_j^k) = 0$ (cf. Lemma 2.1(b)), it follows that

$$d(\mu^{k+1}) \ge d(\mu^{k+1}) - \frac{1}{\omega^k} \sum_{j=1}^m q(\mu_j^{k+1}, \mu_j^k) \ge d(\mu^k), \tag{3.2}$$

so $\{d(\mu^k)\}$ is a monotonically nondecreasing sequence. Since $\{d(\mu^k)\}$ is also bounded from above by the optimal primal value f^* (by the weak duality theorem), we see that the limit

$$d^\infty = \lim_{k \to \infty} d(\mu^k)$$

exists and is at most f^*. Let

$$M^\infty = \{\mu \in [0, \infty)^m \mid d(\mu) \ge d^\infty\}. \tag{3.3}$$

Note that M^∞ is nonempty since $d^\infty \le f^*$ and as mentioned earlier, there exists an optimal dual solution under Assumption A.

Part (a) of the following lemma is due to Censor and Zenios [7], and says that the "distance", measured in terms of the function D, from μ^k to any element of M^∞ is monotonically nonincreasing. From this, we can deduce convergence of the multipliers to a unique limit as pointed out by Eckstein [10]. The proof of these results is patterned after the one given in [2, p. 241], on the convergence of the proximal minimization algorithm, but with the quadratic "proximal" term therein replaced by D.

Lemma 3.1. (a) *For any $\bar{\mu} \in M^\infty$, the sequence $\{D(\bar{\mu}, \mu^k)\}$ is monotonically nonincreasing.*

(b) *$\{\mu^k\}$ converges to a limit.*

Proof. (a) From (3.1) we have

$$d(\mu^{k+1}) - \frac{1}{\omega^k} D(\mu^{k+1}, \mu^k) \ge d(\mu) - \frac{1}{\omega^k} D(\mu, \mu^k) \quad \forall \mu \ge 0.$$

Hence, for all $\mu \ge 0$ with $d(\mu) \ge d(\mu^{k+1})$, there holds $D(\mu^{k+1}, \mu^k) \le D(\mu, \mu^k)$, so that

$$\mu^{k+1} = \arg \min_{\mu \in M^k} D(\mu, \mu^k),$$

where M^k is the set

$$M^k = \{\mu \ge 0 \mid d(\mu) \ge d(\mu^{k+1})\}.$$

Since $\mu^{k+1} \in M^k$, M^k makes a nonempty intersection with $(0, \infty)^m$. Any $\bar{\mu}$ in M^∞ is clearly in M^k, so from Lemma 2.2(d), we obtain that

$$D(\bar{\mu}, \mu^{k+1}) \le D(\bar{\mu}, \mu^k).$$

(b) For any optimal solution $\bar{\mu}$ of (D), $\{D(\bar{\mu}, \mu^k)\}$ is bounded by part (a), so Lemma 2.2(c) implies that $\{\mu^k\}$ is bounded. Let μ^∞ be the limit of a subsequence $\{\mu^k\}_K$, so by the upper semicontinuity of d, $\mu^\infty \in M^\infty$. Using μ^∞ in place of $\bar{\mu}$ in part (a), we obtain that $\{D(\mu^\infty, \mu^k)\}$ is monotonically decreasing. By Lemma 2.2(b), $\{D(\mu^\infty, \mu^k)\}_K$ tends to zero, so the entire sequence $\{D(\mu^\infty, \mu^k)\}$ must converge to zero. Hence, by Lemma 2.2(b), $\mu^k \to \mu^\infty$. $\quad\square$

3.2. Convergence to optimality

We now show that the limit to which $\{\mu^k\}$ converges (cf. Lemma 3.1) is indeed an optimal dual solution. We remark that the standard convergence proof for multiplier methods does not apply here, owing to the ill behaviour of $\nabla\psi^*(s) = \ln(s)$ at the boundary point of its domain. Instead, we employ a novel proof technique, showing that the primal sequence $\{x^k\}$ approaches the primal feasible set in an *ergodic* sense (see Lemma 3.3). This enables us to establish convergence of the multiplier sequence $\{\mu^k\}$ to an optimal dual solution in Proposition 3.1.

We begin our proof with the following lemma stating certain properties of the sequences $\{x^k\}$ and $\{\mu^k\}$. This lemma is based on the proof of a proposition from [5] (see p. 336 therein), and in fact holds for more general multiplier iterations. The proof of the lemma makes use of the boundedness of $\{\mu^k\}$ shown in Lemma 3.1(b), as well as certain properties of the exponential function ψ and the multiplier iteration (1.3)-(1.4).

Lemma 3.2. (a) $d(\mu^k) \leq d(\mu^{k+1}) \leq f^*$ for all k.
  (b) *For all j, $\mu_j^k \psi(\omega^k g_j(x^k))/\omega^k - \mu_j^k g_j(x^k) \to 0$.*
  (c) *For all j, $\mu_j^k g_j(x^k) \to 0$.*
  (d) $d(\mu^k) - f(x^k) \to 0$.

Proof. (a) This was shown earlier; cf. (3.2).
  (b) Let L denote the Lagrangian

$$L(x, \mu) = f(x) + \sum_{j=1}^m \mu_j g_j(x),$$

and let L_ω denote the Augmented Lagrangian

$$L_\omega(x, \mu) = f(x) + \frac{1}{\omega} \sum_{j=1}^m \mu_j \psi(\omega g_j(x)).$$

Then, we have

$$d(\mu^k) \leq L(x^k, \mu^k) \leq L_{\omega^k}(x^k, \mu^k) \leq d(\mu^{k+1}), \tag{3.4}$$

where the first inequality follows from the observation $d(\mu^k) = \min_x L(x, \mu^k)$ (cf. (2.1)), the second inequality follows from the observation $t \leq \psi(t)$ for all t (cf. (1.2)), the third inequality follows from $d(\mu^{k+1}) = f(x^k) + \sum_{j=1}^m \mu_j^k \nabla\psi(\omega^k g_j(x^k)) g_j(x^k)$ (cf. (1.4), (1.5a), and (2.4)) and the observation

$\psi(t) \leq \nabla \psi(t)t$ for all t (cf. (1.2)). Hence $L_{\omega^k}(x^k, \mu^k) - L(x^k, \mu^k) \to 0$, which together with the fact $\psi(\omega^k g_j(x^k))/\omega^k \geq g_j(x^k)$ for all j (cf. $t \leq \psi(t)$ for all t) implies our claim.

(c) By (3.4), we have $d(\mu^{k+1}) - L(x^k, \mu^k) \to 0$, so the formulas for $d(\mu^{k+1})$ and $L(x^k, \mu^k)$ given in the proof of part (b) yield $\sum_{j=1}^{m} \mu_j^k g_j(x^k)(\nabla \psi(\omega^k g_j(x^k)) - 1) \to 0$. We have $t \leq \nabla \psi(t)t$ for all t, so every term inside the preceding summation is nonnegative, and we obtain that, for each j,

$$\mu_j^k g_j(x^k)(\nabla \psi(\omega^k g_j(x^k)) - 1) \to 0.$$

Fix any j. It is readily seen from the properties of ψ (cf. (1.2)) that $\nabla \psi(t) - 1 = 0$ if and only if $t = 0$. Therefore, if there existed a subsequence $K \subset \{0, 1, \dots\}$ and an $\varepsilon > 0$ such that

$$|\mu_j^k g_j(x^k)| \geq \varepsilon \quad \forall k \in K, \tag{3.5}$$

then $\{\nabla \psi(\omega^k g_j(x^k))\}_K \to 1$ and $\{\omega^k g_j(x^k)\}_K \to 0$. Since ω^k is bounded away from zero (cf. (1.5b)), it follows that $\{g_j(x^k)\}_K \to 0$, and from (3.5), $\{|\mu_j^k|\}_K \to \infty$. This contradicts the fact that μ_j^k converges (cf. Lemma 3.1(b)). Hence $\mu_j^k g_j(x^k) \to 0$.

(d) By (3.4), $d(\mu^k) - L(x^k, \mu^k) \to 0$. Since, by part (c), $L(x^k, \mu^k) - f(x^k) \to 0$, it follows that $d(\mu^k) - f(x^k) \to 0$. $\quad \square$

For each k, we define y^k to be the following weighted average (i.e. linear convex combination) of $x^k, \dots, x^0$:

$$y^k = \frac{\omega^k x^k + \cdots + \omega^0 x^0}{\omega^k + \cdots + \omega^0}. \tag{3.6}$$

If $\{y^k\}$ converges, then $\{x^k\}$ is said to converge *ergodically*. We have the following key lemma, which says that $\{y^k\}$ approaches the primal feasible set.

Lemma 3.3 (ergodic feasibility). *For $j = 1, \dots, m$, we have*

$$\limsup_{k \to \infty} g_j(y_k) \leq 0.$$

Proof. Since each g_j is convex, it suffices to show (cf. (3.6)) that

$$\limsup_{k \to \infty} \frac{\omega^k g_j(x^k) + \cdots + \omega^0 g_j(x^0)}{\omega^k + \cdots + \omega^0} \leq 0, \quad j = 1, \dots, m.$$

We will argue by contradiction. If the above relation does not hold, then there would exist some $j \in \{1, \dots, m\}$, a scalar $\theta > 0$, and a subsequence K of $\{0, 1, \dots\}$ such that

$$\gamma^k \geq \theta \quad \forall k \in K,$$

where we denote

$$\gamma^k = \frac{\omega^k g_j(x^k) + \cdots + \omega^0 g_j(x^0)}{\omega^k + \cdots + \omega^0}.$$

Then

$$\mu_j^{k+1} = \mu_j^k \, e^{\omega^k g_j(x^k)} = \mu_j^0 \, e^{\omega^k g_j(x^k) + \cdots + \omega^0 g_j(x^0)} = \mu_j^0 \, e^{(\omega^k + \cdots + \omega^0)\gamma^k}$$

$$\geq \mu_j^0 \, e^{(k+1)\bar\omega\theta} \quad \forall k \in K,$$

(cf. (1.2), (1.4), (1.5)). Hence, $\{\mu_j^{k+1}\}_K \to \infty$, a contradiction of Lemma 3.1(b). $\square$

By combining Lemma 3.1(b), Lemma 3.2(a), Lemma 3.2(d), and Lemma 3.3, we can establish the main result of this section.

Proposition 3.1. *Let $\{\mu^k\}$ be a sequence generated by the exponential multiplier method (1.3) and (1.4) with the penalty parameters chosen according to the rule (1.5a)–(1.5b). Then $\{\mu^k\}$ converges to an optimal dual solution. Furthermore, the sequence $\{y^k\}$ of (3.6) is bounded and each of its cluster points is an optimal primal solution.*

Proof. By Lemma 3.1(b), $\{\mu^k\}$ converges to some limit, say μ^∞. Since f is convex, we have from (3.6) that

$$f(y^k) \leq \frac{\omega^k f(x^k) + \cdots + \omega^0 f(x^0)}{\omega^k + \cdots + \omega^0} \quad \forall k,$$

so it follows from parts (a) and (d) of Lemma 3.2 that

$$\limsup_{k\to\infty} f(y^k) \leq \lim_{k\to\infty} d(\mu^k) \leq f^*. \tag{3.7}$$

Since the optimal solution set of (P) is bounded (cf. Assumption A(a)), f and the g_j's do not share a direction of recession, so the above relation together with Lemma 3.3 implies that $\{y^k\}$ lies in a compact set. Let y^∞ be any cluster point of $\{y^k\}$. Since f and the g_j's are closed, the above relation together with Lemma 3.3 yield

$$f(y^\infty) \leq f^*, \qquad g_j(y^\infty) \leq 0, \quad j = 1, \dots, m.$$

Since f^* is the optimal cost of (P), it follows that y^∞ is an optimal solution of (P) and

$$f(y^\infty) = f^*. \tag{3.8}$$

The lower semicontinuity of f implies that $f(y^\infty) \leq \limsup_{k\to\infty} f(y^k)$, and (3.7) and (3.8) yield $\lim_{k\to\infty} d(\mu^k) = f^*$. Hence using the upper semicontinuity of d, we obtain that $d(\mu^\infty) \geq f^*$. By weak duality, it follows that $d(\mu^\infty) = f^*$. $\square$

4. Rate of convergence analysis for linear programs

In this section we consider a special case where (P) is a linear program. In particular, we assume that f is the polyhedral function

$$f(x) = \begin{cases} b^\mathsf{T} x & \text{if } x \in X, \\ \infty & \text{otherwise,} \end{cases} \tag{4.1}$$

were b is a vector in $\mathbb{R}^n$, X is a polyhedral set in $\mathbb{R}^n$, and g_j, $j=1,\ldots,m$, are affine functions, i.e.,

$$\begin{bmatrix} g_1(x) \\ \vdots \\ g_m(x) \end{bmatrix} = a - A^\mathrm{T}x, \tag{4.2}$$

with A an $m \times n$ matrix and a a vector in $\mathbb{R}^m$. We will exploit the affine structure of the problem in analyzing the corresponding *rate* of convergence of the exponential multiplier method.

The dual functional d given by (2.1), is a concave polyhedral function. In the special case where $X = \mathbb{R}^n$, d has the form

$$d(\mu) = \begin{cases} a^\mathrm{T}\mu & \text{if } A\mu = b \text{ and } \mu \geqslant 0, \\ -\infty & \text{otherwise,} \end{cases} \tag{4.3}$$

and the corresponding dual problem is the classical linear program in standard form (see [9, 17]). We will need the following two technical lemmas for our analysis. The first gives a power series expansion of the "proximal" entropy term q:

Lemma 4.1. *For any $u \in [0, \infty)$ and any $v \in (0, \infty)$, there holds*

$$q(u, v) = \frac{1}{2}\frac{(u-v)^2}{v} - \frac{1}{6}\frac{(u-v)^3}{(v)^2} + \frac{1}{12}\frac{(u-v)^4}{(v)^3} + \cdots$$
$$+ \frac{(-1)^{k+1}}{k(k+1)}\frac{(u-v)^{k+1}}{(v)^k} + \cdots.$$

Proof. From (2.6), we have $q(u, v) = u(\ln(u) - \ln(v)) - u + v$. Replacing $u(\ln(u)-1)$ by its Taylor series expansion around v (which can be seen to be absolutely convergent over the interval $[0, 2v]$), and collecting terms, the result follows. $\square$

The second lemma will be used to relate the l_1 distance of the dual iterates from the optimal dual solution set with the corresponding dual function values. This lemma depends on the affine structure of the problem. The proof is given in the Appendix.

Lemma 4.2. *Let $\{\lambda^k\}$ be a sequence of dual feasible vectors (i.e., $d(\lambda^k) > -\infty$) that converges to some point λ^∞, and is such that $d(\lambda^k) \leqslant d(\lambda^\infty)$ for all k. For each k, let $\bar{\lambda}^k$ be an element of the set*

$$M^\infty = \{\lambda \in [0, \infty)^m \mid d(\lambda) = d(\lambda^\infty)\}$$

satisfying

$$\bar{\lambda}^k \in \arg\min\{\|\lambda - \lambda^k\|_1 \mid \lambda \in M^\infty, |\lambda_j - \lambda_j^k| \leqslant |\lambda_j^\infty - \lambda_j^k| \text{ for all } j\}, \tag{4.4}$$

where $\|\cdot\|_1$ is the usual l_1 norm in $\mathbb{R}^n$. (Notice that there exists at least one minimizing vector $\bar{\lambda}^k$ in (4.4).) Then, there exists a scalar $C_1 > 0$ and an integer $\bar{k}$ such that

$$C_1\|\lambda^k - \bar{\lambda}^k\|_1 \leqslant d(\bar{\lambda}^k) - d(\lambda^k) \quad \forall k \geqslant \bar{k}.$$

4.1. Linear and superlinear convergence

The following proposition derives the convergence rate of the method for the case where the penalty parameter is the same for all constraints:

Proposition 4.1. *Let $\{\mu^k\}$ be a sequence generated by the exponential multiplier method (1.3) and (1.4) with the penalty parameters chosen according to (1.5a) and (1.5b). Then, $\{\mu^k\}$ converges to an optimal dual solution at least linearly. If in addition the penalty parameters tend to ∞, then $\{\mu^k\}$ converges superlinearly.*

Proof. Let $\bar{M}$ denote the set of optimal dual solutions. Then, by Proposition 3.1, there exists a $\mu^\infty \in \bar{M}$ such that $\mu^k \to \mu^\infty$. For each k, let $\bar{\mu}^k$ be an element of $\bar{M}$ satisfying

$$\bar{\mu}^k \in \arg\min \{\|\mu - \mu^k\|_1 \mid \mu \in \bar{M}, |\mu_j - \mu_j^k| \le |\mu_j^\infty - \mu_j^k| \text{ for all } j\}. \tag{4.5}$$

Then by Lemma 4.2, there exists a scalar C_1 and an integer $\bar{k}$ such that

$$C_1 \|\mu^k - \bar{\mu}^k\|_1 \le d(\bar{\mu}^k) - d(\mu^k) \quad \forall k \ge \bar{k}. \tag{4.6}$$

Since $\mu^k \to \mu^\infty$ and $\mu^k > 0$ for all k, there exists an integer $k_1 \ge \bar{k}$ such that

$$\frac{|\mu_j^\infty - \mu_j^k|}{\mu_j^k} \le 1 \quad \forall j, \ \forall k \ge k_1,$$

which together with the relation $|\bar{\mu}_j^k - \mu_j^k| \le |\mu_j^\infty - \mu_j^k|$ for all j (cf. (4.5)), implies

$$\frac{|\bar{\mu}_j^k - \mu_j^k|}{\mu_j^k} \le 1 \quad \forall j, \ \forall k \ge k_1.$$

Fix any $k \ge k_1$ and any $\alpha \in [0,1]$. Then the above relation, together with Lemma 4.1, yields

$$q(\mu_j^k + \alpha(\bar{\mu}_j^k - \mu_j^k), \mu_j^k)$$
$$= \alpha^2 (\bar{\mu}_j^k - \mu_j^k) \left[\frac{1}{2} \frac{(\bar{\mu}_j^k - \mu_j^k)}{\mu_j^k} - \frac{1}{6}\alpha \frac{(\bar{\mu}_j^k - \mu_j^k)^2}{(\mu_j^k)^2} + \frac{1}{12}\alpha^2 \frac{(\bar{\mu}_j^k - \mu_j^k)^3}{(\mu_j^k)^3} + \cdots \right]$$
$$\le \alpha^2 |\bar{\mu}_j^k - \mu_j^k| (\tfrac{1}{2} + \tfrac{1}{6} + \tfrac{1}{12} + \cdots)$$
$$\le C_2 \alpha^2 |\bar{\mu}_j^k - \mu_j^k|,$$

where we let $C_2 = \tfrac{1}{2} + \tfrac{1}{6} + \tfrac{1}{12} + \cdots$. We then have from the nonnegativity of D (cf. Lemma 2.2(a)) and (3.1)–(3.2) that

$$d(\mu^{k+1}) \ge d(\mu^{k+1}) - \frac{1}{\omega^k} \sum_{j=1}^m q(\mu_j^{k+1}, \mu_j^k)$$

$$\ge d(\mu^k + \alpha(\bar{\mu}^k - \mu^k)) - \frac{1}{\omega^k} \sum_{j=1}^m q(\mu_j^k + \alpha(\bar{\mu}_j^k - \mu_j^k), \mu_j^k)$$

$$\ge d(\mu^k + \alpha(\bar{\mu}^k - \mu^k)) - C_2 \frac{\alpha^2}{\omega^k} \sum_{j=1}^m |\bar{\mu}_j^k - \mu_j^k|$$

$$= d(\mu^k + \alpha(\bar{\mu}^k - \mu^k)) - C_2 \frac{\alpha^2}{\omega_k} \|\bar{\mu}^k - \mu^k\|_1,$$

so (4.6) yields, using also the concavity of d,

$$d(\mu^{k+1}) \geq d(\mu^k + \alpha(\bar{\mu}^k - \mu^k)) + \frac{C_2}{C_1} \frac{\alpha^2}{\omega^k} (d(\mu^k) - d(\bar{\mu}^k))$$

$$= (1-\alpha) d(\mu^k) + \alpha d(\bar{\mu}^k) + \frac{C_2}{C_1} \frac{\alpha^2}{\omega^k} (d(\mu^k) - d(\bar{\mu}^k))$$

$$= d(\mu^k) - \left(\alpha - \frac{C_2}{C_1} \frac{\alpha^2}{\omega^k}\right)(d(\mu^k) - d(\bar{\mu}^k)).$$

Since $\bar{\mu}^k \in \bar{M}$, we have $d(\bar{\mu}^k) = f^*$, which in turn implies that

$$d(\mu^{k+1}) - f^* \geq \left(1 - \alpha + \frac{C_2}{C_1} \frac{\alpha^2}{\omega^k}\right)(d(\mu^k) - f^*).$$

We can choose α anywhere in $[0, 1]$, so we let

$$\alpha = \min\left\{1, \frac{C_1}{2C_2} \omega^k\right\},$$

yielding

$$d(\mu^{k+1}) - f^* \geq \left(1 - \min\left\{1, \frac{C_1}{2C_2} \omega^k\right\}\left(1 - \min\left\{\frac{C_2}{C_1\omega^k}, \frac{1}{2}\right\}\right)\right)(d(\mu^k) - f^*).$$
(4.7)

Since the above relation holds for all $k \geq k_1$, and by (1.5b), $\{\omega^k\}$ is bounded away from zero, it follows that $\{d(\mu^k)\}$ converges at least linearly to f^*. (4.6) then implies that $\{\mu^k\}$ approaches $\bar{M}$ at the rate of a geometric progression. If $\omega^k \to \infty$, then $\min\{1, (C_1/2C_2)\omega^k\} \to 1$ and $\min\{C_2/(C_1\omega^k), \frac{1}{2}\} \to 0$, so (4.7) yields

$$\frac{d(\mu^{k+1}) - f^*}{d(\mu^k) - f^*} \to 0,$$

implying that $\{d(\mu^k)\}$ converges to f^* superlinearly. It follows from (4.6) that $\{\mu^k\}$ also approaches $\bar{M}$ superlinearly. $\square$

4.2. Quadratic convergence

In this subsection we consider the exponential multiplier method with the penalty parameters chosen dynamically according to the rule (1.6). Although we do not have a convergence proof for this version of the method, in practice convergence seems to occur always. An important advantage of this version of the method is that locally it attains a *quadratic* rate of convergence without requiring the penalty parameters to tend to ∞. We state this result below. Its proof, based on Lemmas 4.1 and 4.2, is very similar to that of Proposition 4.1.

Proposition 4.2. *Let $\{\mu^k\}$ be a sequence generated by the exponential multiplier method* (1.3)–(1.4) *with the penalty parameters chosen according to* (1.6). *Suppose that* $\{\mu^k\}$ *converges to some point* μ^∞. *Then* $\{\mu^k\}$ *converges at least quadratically.*

Proof. Let $d^\infty = d(\mu^\infty)$ and let

$$M^\infty = \{\mu \in [0, \infty)^m \mid d(\mu) = d^\infty\}.$$

By using (2.7), we have analogously to (3.2) that $d(\mu^{k+1}) \geq d(\mu^k)$ for all k, so by the upper semicontinuity of d, we have $d^\infty \geq d(\mu^k)$ for all k. For each k, let $\bar{\mu}^k$ be an element of M^∞ satisfying

$$\bar{\mu}^k \in \arg\min \{\|\mu - \mu^k\|_1 \mid \mu \in M^\infty, |\mu_j - \mu_j^k| \leq |\mu_j^\infty - \mu_j^k| \text{ for all } j\}. \tag{4.8}$$

Then by Lemma 4.2, there exists a scalar C_1 and an integer $\bar{k}$ such that

$$C_1 \|\mu^k - \bar{\mu}^k\|_1 \leq d(\bar{\mu}^k) - d(\mu^k) \quad \forall k \geq \bar{k}. \tag{4.9}$$

Since $\mu^k \to \mu^\infty$ and $\mu^k > 0$ for all k, there exists an integer $k_1 \geq \bar{k}$ such that

$$|\mu_j^\infty - \mu_j^k| \leq \mu_j^k \quad \forall j, \ \forall k \geq k_1,$$

which together with $|\bar{\mu}_j^k - \mu_j^k| \leq |\mu_j^\infty - \mu_j^k|$ for all j (cf. (4.8)), implies that

$$|\bar{\mu}_j^k - \mu_j^k| \leq \mu_j^k \quad \forall j, \ \forall k \geq k_1. \tag{4.10}$$

Fix any $k \geq k_1$. We have from Lemma 4.1 and (4.10) that

$$q(\bar{\mu}_j^k, \mu_j^k) = \frac{(\bar{\mu}_j^k - \mu_j^k)^2}{\mu_j^k} \left[\frac{1}{2} - \frac{1}{6} \frac{(\bar{\mu}_j^k - \mu_j^k)}{\mu_j^k} + \frac{1}{12} \frac{(\bar{\mu}_j^k - \mu_j^k)^2}{(\mu_j^k)^2} + \cdots \right]$$

$$\leq \frac{(\bar{\mu}_j^k - \mu_j^k)^2}{\mu_j^k} \left(\frac{1}{2} + \frac{1}{6} + \frac{1}{12} + \cdots \right)$$

$$= C_2 \frac{(\bar{\mu}_j^k - \mu_j^k)^2}{\mu_j^k},$$

where $C_2 = \frac{1}{2} + \frac{1}{6} + \frac{1}{12} + \cdots$. This, together with the nonnegativity of q (cf. Lemma 2.1(b)) and (2.7) and (1.6), yield

$$d(\mu^{k+1}) \geq d(\mu^{k+1}) - \sum_{j=1}^m \frac{1}{c_j^k} q(\mu_j^{k+1}, \mu_j^k)$$

$$\geq d(\bar{\mu}^k) - \sum_{j=1}^m \frac{1}{c_j^k} q(\bar{\mu}_j^k, \mu_j^k)$$

$$\geq d(\bar{\mu}^k) - \sum_{j=1}^m \frac{C_2}{c_j^k \mu_j^k} (\bar{\mu}_j^k - \mu_j^k)^2$$

$$\geq d(\bar{\mu}^k) - \frac{C_2}{c} \sum_{j=1}^m (\bar{\mu}_j^k - \mu_j^k)^2$$

$$\geq d(\bar{\mu}^k) - \frac{C_2}{c} \|\bar{\mu}^k - \mu^k\|_1^2.$$

Using (4.9) and the fact $d(\bar{\mu}^k) = d^\infty$ for all k (cf. (4.8) and the definition of M^∞), we obtain

$$d(\mu^{k+1}) - d^\infty \geqslant -\frac{C_2}{c(C_1)^2}(d(\mu^k) - d^\infty)^2. \tag{4.11}$$

Since the choice of k above was arbitrary, (4.11) holds for all $k \geqslant k_1$ and hence $\{d(\mu^k)\}$ converges to d^∞ at least quadratically. Then, by (4.9), $\{\mu^k\}$ converges to M^∞ at least quadratically. $\square$

Appendix. Proof of Lemma 4.2

Let us express the polyhedral set X in (4.1) as

$$X = \{x \mid Bx \geqslant c\},$$

for some $p \times n$ matrix B and some vector $c \in \mathbb{R}^p$. The proof hinges on a result of Hoffman [14] on a certain upper Lipschitzian property of the solution set of a linear system with respect to perturbations in the right-hand side.

We argue by contradiction. Suppose that the claim does not hold. Then, there would exist a subsequence $K \subset \{1, 2, \ldots\}$ such that

$$\left\{\frac{d(\lambda^k) - d(\bar{\lambda}^k)}{\|\lambda^k - \bar{\lambda}^k\|_1}\right\}_K \to 0. \tag{A.1}$$

Fix any $k \in K$. Since $d(\lambda^k) > -\infty$, the minimum in the dual functional definition

$$d(\lambda^k) = \min_{x \in X}\{b^T x + (\lambda^k)^T(a - A^T x)\} \tag{A.2}$$

must be attained at some $y^k \in X$. By using (A.1), we obtain from the Kuhn–Tucker conditions for the above minimization that λ^k and y^k, together with some multiplier vector π^k associated with the constraints $Bx \geqslant c$, satisfy

$$d(\lambda^k) = b^T y^k + (\lambda^k)^T(a - A^T y^k), \tag{A.3}$$

and

$$A\lambda^k + B^T \pi^k = b, \tag{A.4}$$

$$By^k \geqslant c, \quad \pi^k \geqslant 0, \tag{A.5}$$

$$B_i y^k = c_i \quad \forall i \in I^k, \qquad \pi_i^k = 0 \quad \forall i \notin I^k, \tag{A.6}$$

for some subset $I^k \subset \{1, \ldots, p\}$, where B_i is the ith row of B, and c_i is the ith component of c. In addition, we have

$$d(\lambda^k) = (b - A\lambda^k)^T y^k + (\lambda^k)^T a = (B^T \pi^k)^T y^k + a^T \lambda^k = (\pi^k)^T c + a^T \lambda^k, \tag{A.7}$$

where the first equality is due to (A.3), the second equality follows from (A.4), and the last equality follows from (A.6).

Fix any I for which the index set $K_I = \{k \in K \mid I^k = I\}$ is infinite. For each $k \in K_I$, consider the following linear system in (y, π):

$$B^\mathrm{T}\pi = b - A\lambda^k, \qquad By \geq c, \qquad \pi \geq 0, \qquad \pi^\mathrm{T}c = d(\lambda^k) - a^\mathrm{T}\lambda^k,$$

$$B_i y = c_i \quad \forall i \in I, \qquad \pi_i = 0 \quad \forall i \notin I.$$

This system is consistent since it has a solution (y^k, π^k) (cf. (A.4)–(A.7) and $I^k = I$). By a result due to Hoffman [14] (see also [20], and [24]), it has a solution $(\hat{y}^k, \hat{\pi}^k)$ whose norm is bounded by some constant (depending on B and c only) times the norm of the right-hand side. This right-hand side is clearly bounded (recall that $\{\lambda^k\}$ converges), so the sequence $\{(\hat{y}^k, \hat{\pi}^k)\}_{K_I}$ is also bounded. Since $\lambda^k \to \lambda^\infty$, every cluster point $(\hat{y}^\infty, \hat{\pi}^\infty)$ of this sequence satisfies

$$B^\mathrm{T}\hat{\pi}^\infty = b - A\lambda^\infty, \qquad B\hat{y}^\infty \geq c, \qquad \hat{\pi}^\infty \geq 0, \qquad (\hat{\pi}^\infty)^\mathrm{T}c = d^\infty - a^\mathrm{T}\lambda^\infty,$$

$$B_i\hat{y}^\infty = c_i \quad \forall i \in I, \qquad \hat{\pi}_i^\infty = 0 \quad \forall i \notin I.$$

Hence, for each $k \in K_I$, the following linear system in (λ, y, π):

$$\lambda_j \geq 0 \text{ and } |\lambda_j - \lambda_j^k| \leq |\lambda_j^\infty - \lambda_j^k| \quad \forall j,$$

$$A\lambda + B^\mathrm{T}\pi = b, \qquad By \geq c, \qquad \pi \geq 0, \qquad \pi^\mathrm{T}c + a^\mathrm{T}\lambda = d^\infty,$$

$$B_i y = c_i \quad \forall i \in I, \qquad \pi_i = 0 \quad \forall i \notin I,$$

is consistent. Moreover, it is not difficult to see that the first m components of any solution of this system form an element of M^∞. By comparing the above linear system with (A.4)–(A.7) (and using the fact $I = I^k$), we see that (λ^k, y^k, π^k) is almost a solution of the above system, except for a difference in the right side of the last equality ($d(\lambda^k)$ instead of d^∞). By using the same result of Hoffman invoked earlier, we conclude that there exists a solution of the above system, denoted $(\tilde{\lambda}^k, \tilde{y}^k, \tilde{\pi}^k)$, whose distance to (λ^k, y^k, π^k) is bounded by the difference in the right-hand side, that is,

$$\|(\tilde{\lambda}^k, \tilde{y}^k, \tilde{\pi}^k) - (\lambda^k, y^k, \pi^k)\|_1 \leq \tau |d^\infty - d(\lambda^k)|$$

for some scalar τ depending only on A, B, a, and c. The last equation, in view of (A.1), implies that for any $k \in K_I$ sufficiently large, there holds

$$\|\tilde{\lambda}^k - \lambda^k\|_1 < \|\bar{\lambda}^k - \lambda^k\|_1,$$

a contradiction of our choice of $\bar{\lambda}^k$ (cf. (4.4)). $\square$

References

[1] D.P. Bertsekas, G.S. Lauer, N.R. Sandell, Jr. and T.A. Posbergh, "Optimal short term scheduling of large-scale power systems," *IEEE Transactions on Automatic Control* AC-28 (1982) 1–11.

[2] D.P. Bertsekas and J.N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods* (Prentice-Hall, Englewood Cliffs, NJ, 1989).

[3] D.P. Bertsekas, "Necessary and sufficient conditions for a penalty method to be exact," *Mathematical Programming* 9 (1975) 87–99.

[4] D.P. Bertsekas, "Newton's method for linear optimal control problems," in: *Proceedings of the Symposium on Large Scale Systems* (Udine, 1976) pp. 353–359.

[5] D.P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods* (Academic Press, New York, 1982).

[6] L.M. Bregman, "The relaxation method of finding the common point convex sets and its application to the solution of problems in convex programming," *USSR Computational Mathematics and Mathematical Physics* 7 (1967) 200–217.

[7] Y. Censor and S.A. Zenios, "The proximal minimization algorithm with D-functions," (1989), to appear in: *Journal of Optimization Theory and Applications*.

[8] G. Chen and M. Teboulle, "Convergence analysis of a proximal-like minimization algorithm using Bregman functions," (1990), to appear in: *SIAM Journal on Optimization*.

[9] G.B. Dantzig, *Linear Programming and Extensions* (Princeton University Press, Princeton, NJ, 1963).

[10] J. Eckstein, "Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming," (1990), to appear in: *Mathematics of Operations Research*.

[11] A.V. Fiacco and G.P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques* (Wiley, New York, 1968).

[12] E.G. Golshtein and N.V. Tretjiakov, "Modified Lagrangians in convex programming and their generalizations," *Mathematical Programming Studies* 10 (1979) 86–97.

[13] C.D. Ha, "A generalization of the proximal point algorithm," *SIAM Journal on Control and Optimization* 28 (1990) 503–512.

[14] A.J. Hoffman, "On approximate solutions of systems of linear inequalities," *Journal of Research of the National Bureau of Standards* 49 (1952) 263–265.

[15] B.W. Kort and D.P. Bertsekas, "A new penalty function method for constrained minimization," in: *Proceedings of the 1972 IEEE conference on decision and control* (New Orleans, 1972) pp. 162–166.

[16] G.S. Lauer, D.P. Bertsekas, N.R. Sandell, Jr. and T.A. Posbergh, "Optimal solution of large-scale unit commitment problems," *IEEE Transactions on Power Systems and Apparatus* 101 (1981) 79–86.

[17] D.G. Luenberger, *Linear and Nonlinear Programming* (Addison-Wesley, Reading, MA, 1984).

[18] F.J. Luque, "Nonlinear proximal point algorithms," Ph.D. Thesis, Department of Civil Engineering and Operations Research Center, Massachusetts Institute of Technology (Cambridge, MA, 1984).

[19] F.J. Luque, "The nonlinear proximal point algorithm," Report LIDS-P-1598, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology (Cambridge, MA, 1986).

[20] O.L. Mangasarian and T.-H. Shiau, "Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems," *SIAM Journal of Control and Optimization* 25 (1987) 583–595.

[21] B. Martinet, "Regularisation d'inéquations variationelles par approximations successives," *Revue Française d'Automatique et Informatique Recherche Opérationalle* 4 (1970) 154–159.

[22] V.H. Nguyen and J.J. Strodiot, "On the convergence rate of a penalty function method of exponential type," *Journal of Optimization Theory and Applications* 27 (1979) 495–508.

[23] B.T. Poljak and N.V. Tretjakov, "An iterative method for linear programming and its economic interpretation," *Matecon* 10 (1974) 34–46.

[24] S.M. Robinson, "Bounds for errors in the solution set of a perturbed linear program," *Linear Algebra and its Applications* 6 (1973) 69–81.

[25] R.T. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, NJ, 1970).

[26] R.T. Rockafellar, "New applications of duality in convex programming," in: *Proceedings of the Conference on Probability* (Brasov, 1971) pp. 73–81.

[27] R.T. Rockafellar, "A dual approach to solving nonlinear programming problems by unconstrained minimization," *Mathematical Programming* 5 (1973) 354–373.

[28] R.T. Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM Journal on Control and Optimization* 14 (1976) 877–898.