Computational Optimization and Applications, 33, 115–156, 2006

DOI: 10.1007/s10589-005-3065-0

Double-Regularization Proximal Methods, with Complementarity Applications

PAULO J. S. SILVA* pjssilva@ime.usp.br
Department of Computer Science, Instituto de Matemática e Estatística, University of São Paulo, Brazil

JONATHAN ECKSTEIN jeckstei@rutcor.rutgers.edu
Business School and RUTCOR, 640 Bartholomew Road, Rutgers University, Piscataway, NJ, 08854, USA

Received August 12, 2003; Revised February 25, 2005

Published online: 18 October 2005

Abstract. We consider the variational inequality problem formed by a general set-valued maximal monotone operator and a possibly unbounded "box" in $\mathbb{R}^n$, and study its solution by proximal methods whose distance regularizations are coercive over the box. We prove convergence for a class of *double regularizations* generalizing a previously-proposed class of Auslender et al. Using these results, we derive a broadened class of augmented Lagrangian methods. We point out some connections between these methods and earlier work on "pure penalty" smoothing methods for complementarity; this connection leads to a new form of augmented Lagrangian based on the "neural" smoothing function. Finally, we computationally compare this new kind of augmented Lagrangian to three previously-known varieties on the MCPLIB problem library, and show that the neural approach offers some advantages. In these tests, we also consider primal-dual approaches that include a primal proximal term. Such a stabilizing term tends to slow down the algorithms, but makes them more robust.

Keywords: proximal algorithms, variational inequalities, complementarity

1. Introduction

Let $B \subseteq \mathbb{R}^n$ denote the possibly unbounded n-dimensional "box",

$$B \stackrel{\text{def}}{=} ([a_1, b_1] \times \cdots \times [a_n, b_n]) \cap \mathbb{R}^n,$$

where $-\infty \leq a_i < b_i \leq +\infty$, $i = 1, \ldots, n$. This paper will consider the generalized variational inequality problem

$$0 \in T(x) + N_B(x), \tag{1}$$

where T is a (possibly set-valued) maximal monotone operator, and $N_B(x)$ denotes the cone of vectors normal to the set B at x.

*This author was partially supported by CNPq, Grant PQ 304133/2004-3 and PRONEX-Optimization.

Throughout, we will make the standard regularity assumption:

Assumption 1.1. $\operatorname{dom} T \cap \operatorname{int} B \neq \emptyset$.

As an application of this general problem setting, we will be particularly interested in the *complementarity problem*

$$F(x) \geq 0 \quad x \geq 0 \quad \langle F(x), x \rangle = 0 \tag{2}$$

corresponding to some continuous single-valued function $F : \mathbb{R}^n \to \mathbb{R}^n$. In our analysis, we assume that F is monotone, although we will drop this assumption in later computational experiments. Note that since F is continuous and monotone on $\mathbb{R}^n$, it must be maximal monotone; see for example [30, Example 12.7].

A straightforward application of (1) to (2) is to set $T(x) = \{F(x)\}$ for all $x \in \mathbb{R}^n$, and $a_i = 0, b_i = +\infty$ for $i = 1, \ldots, n$. Then $B = \mathbb{R}^n_+$, the nonnegative orthant in $\mathbb{R}^n$, and (1) reduces to

$$0 \in F(x) + N_{\mathbb{R}^n_+}(x), \tag{3}$$

which is equivalent to 2, and called its *primal formulation*.

One can obtain an alternative formulation of 3 in the form 1 by applying a simple duality transformation [1, 16, 23, 26]: given arbitrary set-valued maps U and V, x is a solution of

$$0 \in U(x) + V(x) \tag{4}$$

if and only if there exists y such that

$$y \in U(x) \qquad -y \in V(x). \tag{5}$$

On the other hand, given some y, the existence of an x such that (5) holds is equivalent to y solving

$$0 \in U^{-1}(y) - V^{-1}(-y), \tag{6}$$

where the inverses are taken as point-to-set maps. Applying this duality transformation to (3) with $U = F$ and $V = N_{\mathbb{R}^n_+}$ yields

$$0 \in F^{-1}(y) - (N_{\mathbb{R}^n_+})^{-1}(-y),$$

where inverses are again taken as point-to-set maps. It is easily confirmed that $(-I) \circ (N_{\mathbb{R}^n_+})^{-1} \circ (-I) = N_{\mathbb{R}^n_+}$, so this problem is identical to the *dual formulation*

$$0 \in F^{-1}(y) + N_{\mathbb{R}^n_+}(y), \tag{7}$$

which is also of the form (1) by letting $T(y) = \{x \mid F(x) = y\}$ and $a_i = 0, b_i = +\infty$ for $i = 1, \ldots, n$. As above, F's being continuous and monotone implies it is maximal, which in turn implies $T = F^{-1}$ is maximal monotone; see e.g. [30, Exercise 12.4].

This paper will study generalized proximal methods for (1). These methods are conceptual algorithms in which one takes some generalized distance measure $\tilde{D} : \mathbb{R}^n \times \mathbb{R}^n \to (-\infty, +\infty]$, strictly convex in its first argument, and computes a sequence of iterates $\{x^k\}$ via the recursion

$$0 \in \alpha_k T(x^{k+1}) + N_B(x^{k+1}) + \nabla_1 \tilde{D}(x^{k+1}, x^k), \tag{8}$$

where ∇_1 denotes the gradient with respect to the the first argument, and $\alpha_k > 0$ is some scalar bounded away from 0. $\tilde{D}$ should be finite on $\operatorname{int} B \times \operatorname{int} B$, but may be finite elsewhere as well. The original method of this form, the classical *proximal point algorithm* [28], takes $\tilde{D}(x, y) = (1/2)\|x - y\|^2$. In general, one may have to satisfy (8) only approximately, but for simplicity we defer this complication to Section 2 .

Applying such an algorithm to the primal formulation (3) of the complementarity problem, one obtains the recursion

$$0 \in F(x^{k+1}) + N_{\mathbb{R}^n_+}(x^{k+1}) + \frac{1}{\alpha_k}\nabla_1 \tilde{D}(x^{k+1}, x^k). \tag{9}$$

Applying the same algorithm to the dual formulation (7) leads to subproblem recursion

$$0 \in F^{-1}(y^{k+1}) + N_{\mathbb{R}^n_+}(y^{k+1}) + \frac{1}{\alpha_k}\nabla_1 \tilde{D}(y^{k+1}, y^k).$$

Again applying the duality transformation, but with

$$U = F^{-1} \qquad\qquad V = N_{\mathbb{R}^n_+} + (1/\alpha_k)\nabla_1 \tilde{D}(\cdot, y^k),$$

produces an equivalent subproblem

$$0 \in F(x^{k+1}) - \left(N_{\mathbb{R}^n_+} + \nabla_1 \tilde{D}(\cdot, y^k)\right)^{-1}(-\alpha_k x^{k+1}).$$

The strict convexity of $\tilde{D}(\cdot, y^k)$ in its first argument implies that the mapping

$$P'(\cdot, y^k) \stackrel{\text{def}}{=} \left(N_{\mathbb{R}^n_+} + \nabla_1 \tilde{D}(\cdot, y^k)\right)^{-1} \tag{10}$$

is single-valued, so we obtain the equivalent recursions

$$0 = F(x^{k+1}) - P'(-\alpha_k x^{k+1}, y^k) \tag{11}$$

$$y^{k+1} = P'(-\alpha_k x^{k+1}, y^k), \tag{12}$$

which is known as a *method of multipliers* or *generalized augmented Lagrangian* method. First, one solves the system of nonlinear equations (11) —the augmented Lagrangian—to obtain x^{k+1}, and then one updates the Lagrange multiplier estimates via (12). We use the letter P because $P'(\cdot, y^k)$ plays the same role as the gradient of the penalty term in augmented Lagrangian methods for optimization problems; see for example [14, 29, 32]. Algorithms of this class exist for problems where the constraint set takes a much more general form than a box, but we focus here on the simple complementarity case. The augmented Lagrangian methods of this paper are easily adapted to the more general setting; see for example [4].

The main subject of this paper is when $\tilde{D}$ is separable and *coercive* on B, that is,

$$\tilde{D}(x, y) = \sum_{i=1}^{n} \tilde{d}(x_i, y_i) \tag{13}$$

$$\tilde{d}(x_i, y_i) = +\infty \qquad \text{if } x_i \notin [a_i, b_i] \tag{14}$$

$$\lim_{x \downarrow a_i} \nabla_1 \tilde{d}(x_i, y_i) = -\infty \qquad \text{if } a_i > -\infty \tag{15}$$

$$\lim_{x \uparrow b_i} \nabla_1 \tilde{d}(x_i, y_i) = +\infty \qquad \text{if } b_i < +\infty. \tag{16}$$

In this case, $\nabla_1 \tilde{D}(\cdot, x^k)$ acts as a kind of "barrier" in algorithms like (8), keeping successive iterates within $\operatorname{int} B$. In particular, $N_B + \nabla_1 \tilde{D}(\cdot, x^k) = \nabla_1 \tilde{D}(\cdot, x^k)$, so (8) reduces to the simpler recursion

$$0 \in \alpha_k T(x^{k+1}) + \nabla_1 \tilde{D}(x^{k+1}, x^k),$$

which should be more convenient computationally. For example, the primal complementarity recursion (9) now reduces to $0 = F(x^{k+1}) + (1/\alpha_k)\nabla_1 \tilde{D}(x^{k+1}, x^k)$, which is an equation rather than an inclusion, and inherits whatever smoothness is present in F and $\nabla_1 \tilde{D}(\cdot, x^k)$. This situation may be preferable to the non-coercive case, where the resulting subproblem may be no easier than the original problem (3). However, much as in interior-point methods, we must still constrain the definition domain of x^{k+1} to the positive orthant, which we denote $\mathbb{R}^n_{++}$, presenting possible computational difficulties.

An even more important property of coercive separable distances emerges when they are applied to the dual formulation (7), and used in the corresponding multiplier method (11) and (12). Then, the definition (10) reduces to the much simpler

$$P'(\cdot, y^k) = (\nabla_1 \tilde{D}(\cdot, y^k))^{-1}. \tag{17}$$

By judicious choice of $\tilde{D}$, one can make the single-valued function $P'(\cdot, y^k)$ finite everywhere, with any desired degree of smoothness. The augmented Lagrangian equation system (11) can then be made to have the same definition domain and degree of smoothness as F. This property may in turn allow solution by standard Newton methods, a significant advantage. Classical choices of $\tilde{D}$ lead to nonsmooth augmented Lagrangians. Note, however, that certain non-coercive choices of $\tilde{D}$ can still lead to limited smoothness of the augmented Lagrangian [16, 21]; Section 6 considers one such choice and empirically compares it to some coercive choices.

Given the attractive properties of coerciveness, an unfortunate gap existed for some time in the theory of coercive proximal algorithms. In early analyses such as [8, 14, 32], convergence of methods like (8) was demonstrated only when *either* T was the subgradient map of some convex function $f : \mathbb{R}^n \to (-\infty, +\infty]$, *or* when the distance regularization $\tilde{D}(\cdot, x^k)$ was *not* coercive. The case of a general monotone T and coercive $\tilde{D}$ remained open.

Subsequent research [9] proved convergence of certain coercive proximal algorithms when T is *paramonotone* [19], a condition less stringent than T being a subgradient, but more restrictive than general monotonicity.

A breakthrough then came with Auslender et al.'s publication of [2], which proved convergence of a proximal method with a general monotone operator T, and a specific, coercive form of $\tilde{D}$, the log-quadratic kernel. This $\tilde{D}$ is a weighted sum of a logarithmic term, one of the standard coercive choices, and a traditional quadratic term.

Very shortly thereafter, the authors of [2] generalized its analysis in [3]. There, they proved convergence for a family of possible distance regularizations, the class Φ_2 of rescaled φ-divergences, and also included analyses of dual algorithms like (11)–(12).

This paper makes three contributions: first, capitalizing on our related work in [31], we prove convergence for a general maximal monotone operator T of proximal methods employing a broader class of coercive distance regularization measures $\tilde{D}$ than in [3]. These distance measures do not have to take the φ-divergence form: for example, they may instead be certain kinds of rescaled Bregman distances; see Section 4.2. Second, we note a relationship between the log-quadratic penalty arising in [3] and some prior work on pure penalty (or smoothing) methods for complementarity problems [10]. We study another penalty from [10], the *neural network smooth plus function*, and show that it too corresponds to a proximal algorithm and augmented Lagrangian method. We call the penalty term for this method the *neural penalty*, and note that it is essentially the integral of the augmented Lagrangian penalty term proposed in [25]. The neural penalty yielded superior computational results to the log-quadratic penalty in the pure penalty environment of [10], so it is natural to consider whether it might also be superior in an augmented Lagrangian setting.

The third contribution of this paper is to test a variety of augmented Lagrangian algorithms on a difficult, realistic test set of complementarity problems, the MCPLIB [11, 12]. Note that the MCPLIB problems are not monotone, as required by our convergence analysis, but, following the example of [16], we still use them as a computational testing library. In addition to evaluating the relative merits of the log-quadratic and neural penalties, we compare them to a variant of the classic exponential penalty and to the cubic penalty of [16], and consider in each case the effect of adding a primal proximal term to the augmented Lagrangian.

The remainder of this paper is structured as follows: Section 2 sets forth the class of distance regularizations $\tilde{D}$ that we analyze, and our *PMDR algorithm* that employs them. Generalizing [3], we study distances $\tilde{D}$ of the form

$$\tilde{D}(x, y) = D(x, y) + \frac{\mu}{2}\|x - y\|^2,$$

that is, the sum of a coercive term D and μ times the traditional squared Euclidean distance; we assume $\mu \geq 1$. We then make two sets of assumptions about this distance: first, $\tilde{D}$ must meet a set of conditions (Assumption 2.1 below) slightly reformulated from our earlier work in [31]. Next, we introduce a set of conditions on the coercive term $D(\cdot, y)$ (Assumption 2.3 below), that constrain its derivative to lie within a certain envelope. In the case $B = \mathbb{R}^n_+$, the lower bound of this envelope corresponds exactly to the log-quadratic measure of [2].

Section 3 then presents our convergence analysis. We first prove convergence in the case $\mu > 1$, as in [3]. In Section 3.1 , however, we consider the case $\mu = 1$, which requires some strengthened assumptions, but is needed to analyze the neural penalty. Section 3.2 ends Section 3 by demonstrating a special-case quadratic convergence rate result like those of [3].

Section 4 gives examples of distance measures meeting our assumptions. These include the Φ_2 class proposed in [3], but also other possibilities. Next, Section 5 considers how our class of distance measures manifests itself in the dual setting (11)–(12), and also (briefly) in primal-dual settings. We show that any penalty term having certain regularity properties and fitting inside a certain envelope corresponds to one of our allowed distance measures $\tilde{D}$. The *upper* bound of this envelope is the log-quadratic penalty of [3]. The remainder of Section 5 develops the relationship with the work of Chen and Mangasarian, as well as the properties of the neural penalty, which uses $\mu = 1$. Finally, Section 6 presents the computational testing.

2. Coercive separable distances and double regularizations

We begin by stating a key set of assumptions adapted from [31]:

Assumption 2.1. For $i = 1, \ldots, n$, the function $\tilde{d}_i : \mathbb{R} \times (a_i, b_i) \to (-\infty, \infty]$ has the following properties:

2.1.1. For all $y_i \in (a_i, b_i)$, $\tilde{d}_i(\cdot, y_i)$ is closed and strictly convex, with its minimum at y_i. Moreover, $\operatorname{int}\operatorname{dom} \tilde{d}_i(\cdot, y_i) = (a_i, b_i)$.

2.1.2. $\tilde{d}_i$ is differentiable with respect to its first argument on $(a_i, b_i) \times (a_i, b_i)$, and this partial derivative is continuous at all points of the form $(x_i, x_i) \in (a_i, b_i) \times (a_i, b_i)$. Moreover, we will use the notation

$$\tilde{d}_i'(x_i, y_i) \stackrel{\text{def}}{=} \frac{\partial \tilde{d}_i}{\partial x_i}(x_i, y_i).$$

2.1.3. For all $y_i \in (a_i, b_i)$, $\tilde{d}_i(\cdot, y_i)$ is essentially smooth [28, Chapter 26].

2.1.4. There exist $L, \epsilon > 0$ such that if either $-\infty < a_i < y_i \leq x_i < a_i + \epsilon$ or $b_i - \epsilon < x_i \leq y_i < b_i < +\infty$, then $|\tilde{d}_i'(x_i, y_i)| \leq L\, |x_i - y_i|$.

This assumption is a simple transformation of [31, Assumption 2.1], where each $\tilde{d}_i(\cdot, y_i)$ is divided by $\tilde{d}_i''(y_i, y_i)$. We note that all the convergence results from [31] remain true under Assumption 2.1.

The stipulation that int dom $\tilde{d}_i(\cdot, y_i) = (a_i, b_i)$ and Assumption 2.1.3's requirement of essential smoothness imply that (14)–(16) hold—that is, they guarantee that $d_i(\cdot, y_i)$ is coercive on $[a_i, b_i] \cap \mathbb{R}$. Within a proximal algorithm, such a $\tilde{d}_i(\cdot, y_i)$ acts as a "barrier" keeping the iterates within the interval (a_i, b_i).

We will use functions $\tilde{d}_i$ of this sort as proximal kernels; however, we will obtain such functions by adding a simple quadratic function to another coercive function d_i, as follows:

Definition 2.2. Let $\mu \geq 1$. For $i = 1, \dots, n$, let $d_i : \mathbb{R} \times (a_i, b_i) \to (-\infty, \infty]$ be continuously differentiable with respect to the first variable. Let

$$\tilde{d}_i(x_i, y_i) \stackrel{\text{def}}{=} d_i(x_i, y_i) + \frac{\mu}{2}(x_i - y_i)^2, \tag{18}$$

and

$$\tilde{d}(x, y) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \tilde{d}_i(x_i, y_i) = \sum_{i=1}^{n} d_i(x_i, y_i) + \frac{\mu}{2}(x_i - y_i)^2. \tag{19}$$

If each $\tilde{d}_i$ conforms to Assumptions 2.1.1–2.1.3, we shall call $\tilde{d}$ the *double regularization* based on $D(x, y) \stackrel{\text{def}}{=} \sum_{i=1}^{n} d_i(x_i, y_i)$. Moreover, each $\tilde{d}_i$ will be called the *double regularization component* based on d_i.

Note that we did not directly require Assumption 2.1.4. Instead, we make a further assumption with no analog in [31], and show that it implies Assumption 2.1.4:

Assumption 2.3. For $i = 1, \dots, n$, let $d_i : \mathbb{R} \times (a_i, b_i) \to (-\infty, \infty]$ and $x_i, y_i \in (a_i, b_i)$. Then,

2.3.1. If a_i and b_i are both finite as illustrated in Figure 1,

$$\frac{(x_i - y_i)(y_i - a_i)}{x_i - a_i} \leq d_i'(x_i, y_i) \leq \frac{(x_i - y_i)(b_i - y_i)}{b_i - x_i}.$$

2.3.2. Otherwise, we take the respective limits as $a_i \to -\infty$ or $b_i \to \infty$ in the above relation:

2.3.2a. If only a_i is finite:

$$\frac{(x_i - y_i)(y_i - a_i)}{x_i - a_i} \leq d_i'(x_i, y_i) \leq x_i - y_i.$$

2.3.2b. If only b_i is finite:

$$x_i - y_i \leq d_i'(x_i, y_i) \leq \frac{(x_i - y_i)(b_i - y_i)}{b_i - x_i}.$$

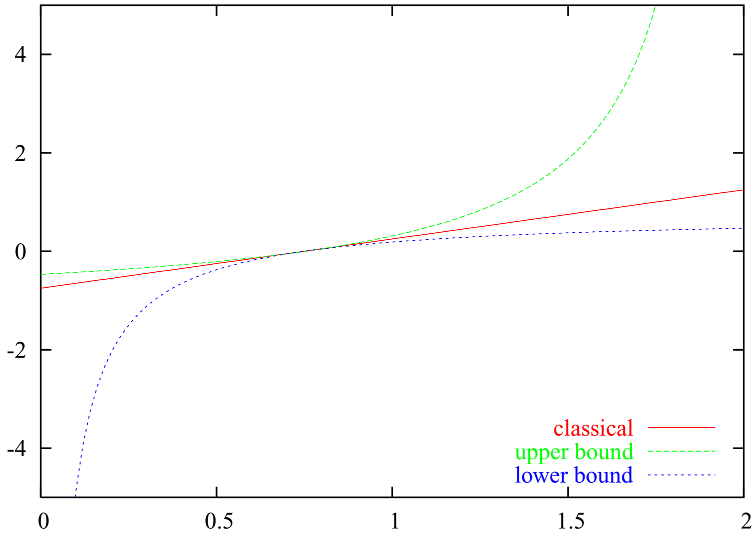

Figure 1. Bounds for the derivatives of d_i for $(a_i, b_i) = (0, 2)$ and $y_i = 0.75$. The derivative of the classical regularization $(1/2)(x_i - y_i)^2$ lies in between the limits.

2.3.2c. $(a_i, b_i) = \mathbb{R}$:

$$d_i'(x_i, y_i) = x_i - y_i.$$

Lemma 2.4. *Suppose d_i conforms to Assumption 2.3. Let*

$$\epsilon \stackrel{\text{def}}{=} \min_{i=1,\dots,n} \left\{ \frac{b_i - a_i}{2} \right\} \in (0, +\infty].$$

If $-\infty < a_i < y_i \leq x_i < a_i + \epsilon$ or $b_i - \epsilon < x_i \leq y_i < b_i < +\infty$, then $|d_i'(x_i, y_i)| \leq 2\,|x_i - y_i|$. Therefore, $|\tilde{d}_i'(x_i, y_i)| \leq (2 + \mu)\,|x_i - y_i|$, and the double regularization component $\tilde{d}_i$ based on d_i meets Assumption 2.1.4. with $L = 2 + \mu$.

Proof: Suppose $-\infty < a_i < y_i \leq x_i < a_i + \epsilon$. If $b_i = +\infty$, we have

$$|d_i'(x_i, y_i)| = d_i'(x_i, y_i) \leq x_i - y_i = |x_i - y_i| \leq 2\,|x_i - y_i|\,.$$

On the other hand, if $b_i \in \mathbb{R}$, we get

$$\begin{aligned} |d_i'(x_i, y_i)| = d_i'(x_i, y_i) &\leq \frac{(x_i - y_i)(b_i - y_i)}{b_i - x_i} = \frac{|x_i - y_i|(b_i - y_i)}{b_i - x_i} \\ &\leq \frac{|x_i - y_i|(b_i - a_i)}{b_i - a_i - \epsilon} \leq 2|x_i - y_i|. \end{aligned}$$

The analysis of the situation $b_i - \epsilon < x_i \leq y_i < b_i < +\infty$ is analogous. $\square$

We now introduce our proximal method, generalizing the discussion of Section 1 by allowing approximate computation of the iterates:

Proximal method using double regularization (PMDR): Let $\tilde{d}$ be a double regularization based via (19) on a coercive term D conforming to Assumption 2.3, with $\mu \geq 1$.

1. **Initialization:** Let $k = 0$. Choose a scalar $\underline{\alpha} > 0$, and an initial iterate $x^0 \in \operatorname{int} B$.
2. **Iteration:**
   (a) Choose $\alpha_k \in [\underline{\alpha}, +\infty)$.
   (b) Find $x^{k+1} \in \mathbb{R}^n$ and some "small" $e^{k+1} \in \mathbb{R}^n$ such that

$$e^{k+1} \in \alpha_k T(x^{k+1}) + \nabla_1 \tilde{d}(x^{k+1}, x^k). \tag{20}$$

   (c) Let $k \leftarrow k+1$, and repeat.

In order to ensure convergence, we need some conditions on the error sequence $\{e^k\}$, so that it is indeed sufficiently "small". We adopt the error criterion from [15]:

Assumption 2.5. The error sequence $\{e^k\}$ should conform to:

$$\sum_{k=1}^{\infty} \|e^k\| < \infty, \qquad \sum_{k=1}^{\infty} \langle e^k, x^k \rangle \text{ exists and is finite.}$$

This error assumption will allow us to prove that the PMDR sequence is quasi-Fejér convergent to the solution set of (1). That is, for any $z \in (T + N_B)^{-1}(0)$, there is a summable sequence $\{\epsilon_k(z)\}$ such that

$$\|z - x^{k+1}\|^2 \leq \|z - x^k\|^2 + \epsilon_k(z).$$

The usefulness of quasi-Fejér convergence is summarized in the following result; see [6, Theorem 1] and [15, Lemma 4].

Proposition 2.6. *Let $Z \subset \mathbb{R}^n$ be a nonempty set, and $\{x^k\}$ be a sequence such that*

$$\|z - x^{k+1}\|^2 \leq \|z - x^k\|^2 + \epsilon_k(z),$$

for all $z \in Z$ and all k, with $\{\epsilon_k(z)\}$ being summable for all $z \in Z$. Then $\{x^k\}$ is bounded, the sequence $\{|z - x^k\}|$ converges for all $z \in Z$, and if any cluster point of $\{x^k\}$ is in Z, the entire sequence $\{x^k\}$ converges to that point.

3. Convergence analysis

To prove the quasi-Fejér convergence of a PMDR sequence, will need four technical lemmas.

Lemma 3.1. *Let $\alpha \le \beta$ and $\gamma \le \delta$ be real numbers. Then*

$$(\delta - \alpha)(\gamma - \beta) \le (\delta - \beta)(\gamma - \alpha),$$

and this inequality is strict if $\alpha \ne \beta$ and $\gamma \ne \delta$.

Proof: Multiplying the inequality $\alpha \le \beta$ by the nonnegative value $\delta - \gamma$,

$$\begin{array}{rc} & \alpha(\delta - \gamma) \le \beta(\delta - \gamma) \\ \Rightarrow & -\alpha\gamma - \beta\delta \le -\beta\gamma - \alpha\delta \\ \Rightarrow & \alpha\beta - \alpha\gamma - \beta\delta + \gamma\delta \le \alpha\beta - \beta\gamma - \alpha\delta + \gamma\delta \\ \Rightarrow & (\delta - \alpha)(\gamma - \beta) \le (\delta - \beta)(\gamma - \alpha). \end{array}$$

The strict inequality assertion follows from the same reasoning, observing that the first two inequalities are strict when $\alpha \ne \beta$ and $\gamma \ne \delta$. □

Lemma 3.2. *Under Assumption 2.3,*

$$\operatorname{sgn}(d_i'(x_i, y_i)) = \operatorname{sgn}(x_i - y_i),$$

where

$$\operatorname{sgn}(x) \stackrel{\text{def}}{=} \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ +1 & \text{if } x > 0. \end{cases}$$

Proof: Examining Assumption 2.3, the signs of all the upper and lower bounds on $d_i'(x_i, y_i)$ are identical to the sign of $x_i - y_i$. □

Lemma 3.3. *Let $d_i : \mathbb{R} \times (a_i, b_i) \to (-\infty, \infty]$ be a function conforming to Assumption 2.3 . For all $z_i \in [a_i, b_i] \cap \mathbb{R}^n$ and $x_i, y_i \in (a_i, b_i)$,*

$$(z_i - x_i) d_i'(x_i, y_i) \le (z_i - y_i)(x_i - y_i).$$

Proof: With the help of Lemma 3.2, one can easily confirm the inequality whenever $x_i = y_i$, $x_i = z_i$ or $y_i = z_i$. So, from now on, suppose that x_i, y_i, and z_i are all distinct.

Suppose $a_i, b_i \in \mathbb{R}$. Then we divide the proof into four cases:

1. $x_i < \min(y_i, z_i)$:
   If $z_i < y_i$, it follows that $(z_i - x_i) d_i'(x_i, y_i) < 0 < (z_i - y_i)(x_i - y_i)$.
   If $y_i < z_i$, we apply Lemma 3.1 with $\alpha = x_i,\ \beta = y_i, \gamma = z_i, \delta = b_i$ and get:

$$\begin{array}{rl} & (b_i - x_i)(z_i - y_i) \le (b_i - y_i)(z_i - x_i) \\ \Rightarrow & (x_i - y_i)(z_i - y_i) \ge \dfrac{x_i - y_i}{b_i - x_i}(b_i - y_i)(z_i - x_i) \\ \Rightarrow & (x_i - y_i)(z_i - y_i) \ge d_i'(x_i, y_i)(z_i - x_i). \qquad \text{[Assumption 2.3]} \end{array}$$

2. $x_i > \max(y_i, z_i)$ (very similar to case 1):
   If $z_i > y_i$, we have $(z_i - x_i)d_i'(x_i, y_i) < 0 < (z_i - y_i)(x_i - y_i)$.
   If $z_i < y_i$, apply Lemma 3.1 with $\alpha = a_i, \beta = z_i, \gamma = y_i, \delta = x_i$, yielding

$$\begin{aligned} & (x_i - a_i)(y_i - z_i) \leq (x_i - z_i)(y_i - a_i) \\ \Rightarrow \quad & (x_i - y_i)(y_i - z_i) \leq \frac{x_i - y_i}{x_i - a_i}(x_i - z_i)(y_i - a_i) \\ \Rightarrow \quad & (x_i - y_i)(y_i - z_i) \leq d_i'(x_i, y_i)(x_i - z_i). \qquad \text{[Assumption 2.3]} \end{aligned}$$

3. $z_i < x_i < y_i$:
   Apply Lemma 3.1 with $\alpha = a_i, \beta = z_i, \gamma = x_i, \delta = y_i$, resulting in

$$\begin{aligned} & (y_i - a_i)(x_i - z_i) \leq (y_i - z_i)(x_i - a_i) \\ \Rightarrow \quad & \frac{x_i - y_i}{x_i - a_i}(y_i - a_i)(x_i - z_i) \geq (x_i - y_i)(y_i - z_i) \\ \Rightarrow \quad & d_i'(x_i, y_i)(x_i - z_i) \geq (x_i - y_i)(y_i - z_i). \qquad \text{[Assumption 2.3]} \end{aligned}$$

4. $y_i < x_i < z_i$:
   Again, we apply Lemma 3.1, but now with $\alpha = y_i, \beta = x_i, \gamma = z_i, \delta = b_i$:

$$\begin{aligned} & (b_i - y_i)(z_i - x_i) \leq (b_i - x_i)(z_i - y_i) \\ \Rightarrow \quad & \frac{x_i - y_i}{b_i - x_i}(b_i - y_i)(z_i - x_i) \leq (x_i - y_i)(z_i - y_i) \\ \Rightarrow \quad & d_i'(x_i, y_i)(z_i - x_i) \leq (x_i - y_i)(z_i - y_i). \qquad \text{[Assumption 2.3]} \end{aligned}$$

It remains only to consider what occurs if $a_i = -\infty$ or $b_i = \infty$. These unbounded cases follow, similarly to the respective inequalities in Assumption 2.3, by taking limits in the bounded cases above. □

Lemma 3.4. *Let $\tilde{d}$ be a double regularization, A and C be subsets of* int B*, and $z \in B$. If for each $i = 1, \ldots, n$, there exists some $\zeta_i(z_i, A, C) > 0$ such that for all $x \in A$ and $y \in C$,*

$$(z_i - x_i)d_i'(x_i, y_i) \leq \zeta_i(z_i, A, C)(z_i - y_i)(x_i - y_i),$$

then for all $x \in A$ and $y \in C$,

$$\begin{aligned} & \langle z - x, \nabla_1 \tilde{d}(x, y) \rangle \\ & \quad \leq \sum_{i=1}^{n} \left(\frac{\mu + \zeta_i(z_i, A, C)}{2}((z_i - y_i)^2 - (z_i - x_i)^2) - \frac{\mu - \zeta_i(z_i, A, C)}{2}(x_i - y_i)^2 \right). \end{aligned}$$

Proof: If $x \in A$ and $y \in C$,

$$\begin{aligned} (z_i - x_i)\nabla_1 \tilde{d}(x, y)_i &= (z_i - x_i)(d_i'(x_i, y_i) + \mu(x_i - y_i)) \\ &\leq \zeta_i(z_i, A, C)(z_i - y_i)(x_i - y_i) + \mu(z_i - x_i)(x_i - y_i). \end{aligned}$$

Using the identities

$$(z_i - y_i)(x_i - y_i) = \frac{(z_i - y_i)^2 - (z_i - x_i)^2 + (x_i - y_i)^2}{2}$$
$$(z_i - x_i)(x_i - y_i) = \frac{(z_i - y_i)^2 - (z_i - x_i)^2 - (x_i - y_i)^2}{2},$$

it follows that

$$(z_i - x_i)\nabla_1 \tilde{d}(x, y)_i \leq \frac{\mu + \zeta_i(z_i, A, C)}{2}((z_i - y_i)^2 - (z_i - x_i)^2) - \frac{\mu - \zeta_i(z_i, A, C)}{2}(x_i - y_i)^2.$$

The result follows by adding this inequality for $i = 1, \ldots, n$. □

We can now establish quasi-Fejér convergence:

Lemma 3.5. *Let $\{x^k\}$ be a sequence computed by the PMDR conforming to Assumption 2.5. Then $\{x^k\}$ is quasi-Fejér convergent to the solution set of 1. Moreover, if $\mu > 1$ and the solution set is non-empty, then $x^{k+1} - x^k \to 0$.*

Proof: Let $z \in (T + N_B)^{-1}(0)$. From (20),

$$\frac{-\nabla_1 \tilde{d}(x^{k+1}, x^k) + e^k}{\alpha_k} \in (T + N_B)(x^{k+1}).$$

Using the monotonicity of $T + N_B$, it follows that

$$0 \leq \langle z - x^{k+1}, \nabla_1 \tilde{d}(x^{k+1}, x^k) - e^{k+1} \rangle.$$

From Lemma 3.3, it is possible to apply Lemma 3.4 with $A = C = \operatorname{int} B$, $x = x^{k+1}$, $y = x^k$ and $\zeta_i(z_i, A, C) = 1$ for all $i = 1, \ldots, n$. Therefore,

$$0 \leq \frac{\mu + 1}{2}\left(\|z - x^k\|^2 - \|z - x^{k+1}\|^2\right) - \frac{\mu - 1}{2}\|x^{k+1} - x^k\|^2 + \langle e^{k+1}, x^{k+1} - z \rangle.$$

Rearranging, multiplying by $2/(\mu + 1) > 0$, and using the Cauchy-Schwartz inequality,

$$\begin{aligned} \|z - x^{k+1}\|^2 &\leq \|z - x^k\|^2 - \frac{\mu - 1}{\mu + 1}\|x^{k+1} - x^k\|^2 + \frac{2}{\mu + 1}\langle e^{k+1}, x^{k+1} - z \rangle \\ &\leq \|z - x^k\|^2 - \frac{\mu - 1}{\mu + 1}\|x^{k+1} - x^k\|^2 + \epsilon_k(z), \end{aligned} \quad (21)$$

where we set

$$\epsilon_k(z) = \frac{2}{\mu + 1}(\langle e^{k+1}, x^{k+1} \rangle + \|e^{k+1}\| \|z\|).$$

We have $\mu \geq 1$, and Assumption 2.5 implies $\{\epsilon_k(z)\}$ is summable for any z, so (21) establishes quasi-Fejér convergence.

Finally, to prove that $x^{k+1} - x \to 0$ whenever $\mu > 1$, we follow the proof of [15, Lemma 3]. Applying induction to (21), we obtain for all $l \geq 1$ that

$$\|z - x^l\|^2 \leq \|z - x^0\|^2 - \frac{\mu - 1}{\mu + 1} \sum_{k=0}^{l-1} \|x^{k+1} - x^k\|^2 + \sum_{k=0}^{l-1} \epsilon_k(z).$$

Since $\{\epsilon_k(z)\}$ is summable, it follows that $\bar{E}(z) \overset{\text{def}}{=} \sup_{l \geq 1}\{\sum_{k=0}^{l-1} \epsilon_k(z)\}$ must be finite. Therefore,

$$\sum_{k=0}^{\infty} \|x^{k+1} - x^k\|^2 \leq \frac{\mu + 1}{\mu - 1} \|z - x^0\|^2 + \bar{E}(z) < \infty.$$

Thus, $\|x^{k+1} - x^k\|$ is square summable, and therefore $x^{k+1} - x^k \to 0$. □

With the above results in hand, it is possible to apply the analysis of [26] to prove the convergence of the PMDR algorithm:

Proposition 3.6. *In the PMDR algorithm, suppose $\tilde{d}$ is a double regularization based on a distance D conforming to Assumption 2.3, where $\mu > 1$. Then the resulting sequence $\{x^k\}$ converges to a solution of 1, if such a solution exists.*

Proof: We will apply [31, Theorem 2.7]. First, we have already shown that any double regularization that conforms to Assumption 2.3 also conforms to Assumption 2.1, which is identical to [31, Assumption 2.1] after applying a scaling factor. The regularity condition of Assumption 1.1 is exactly [31, Assumption 2.2]. And finally, Assumption 2.5 and Lemma 3.5 imply [31, Assumption 2.3] with $\beta^k = \|e^k\|_\infty$. Thus, the assumptions of [31, Theorem 2.7] hold, and so that result asserts subsequential convergence of $\{x^k\}$. Via Proposition 2.6, quasi-Fejér convergence implies convergence of the whole sequence. □

3.1. Analysis of the case $\mu = 1$

Proposition 3.6 omits the case $\mu = 1$. However, in Sections 5 and 6, we will encounter precisely this case. By strengthening Assumption 2.3, we now develop a convergence result for the $\mu = 1$ case.

When $\mu = 1$, Lemma 3.5 does not guarantee that the difference of successive iterates goes to zero, so [31, Assumption 2.3] does not hold, and the proof of Proposition 3.6 is not valid. By strengthening Assumption 2.3, we seek to reestablish the condition $x^{k+1} - x^k \to 0$, so the logic of Proposition 3.6 will once more apply.

For simplicity, we consider only the case $a = 0$, $b = +\infty$.

Assumption 3.7. Let $B = \mathbb{R}^n{}_+$. Let $d_i : \mathbb{R} \times \mathbb{R}_{++} \to (-\infty, \infty]$, $i = 1, \ldots, n$, be the coercive terms used to construct a double regularization $\tilde{d}$. We assume that $d_i'(\cdot, \cdot)$ is continuous and:

3.7.1 For all $x_i, y_i \in \mathbb{R}_{++}$,

$$\frac{(x_i - y_i)y_i}{x_i} \le d_i'(x_i, y_i) \le x_i - y_i,$$

and the lower bound is strict if $x_i \neq y_i$.

3.7.2. Given any $\bar{y}_i > 0$, there exist a constant $\zeta_i(\bar{y}_i) \in (0, 1)$, a neighborhood $A_i(\bar{y}_i)$ of 0, and a neighborhood $C_i(\bar{y}_i)$ of $\bar{y}_i$ such that, for all $x_i \in A_i(\bar{y}_i) \cap \mathbb{R}_{++}$ and $y_i \in C_i(\bar{y}_i) \cap \mathbb{R}_{++}$,

$$\zeta_i(\bar{y}_i)\frac{(x_i - y_i)y_i}{x_i} \le d_i'(x_i, y_i). \tag{22}$$

Assumption 3.7.1 simply restates Assumption 2.3 for the $(a, b) = \mathbb{R}_{++}$ case, with the additional stipulation of strict inequality for $x_i \neq y_i$. Note that (22) automatically holds when $x_i \ge y_i$ and both sides are nonnegative, but imposes a stronger bound when $x_i < y_i$ and both sides are negative.

Lemma 3.8. *Let $d_i : \mathbb{R} \times \mathbb{R}_{++} \to (-\infty, \infty]$ be function conforming to Assumption 3.7.1. Suppose $z_i \ge 0$, and $x_i, y_i > 0$ with $x_i \neq y_i$. Then,*

$$(z_i - x_i)d_i'(x_i, y_i) < (z_i - y_i)(x_i - y_i).$$

Proof: Assumption 3.7.1 implies Assumption 2.3 for the case $a = 0, b = +\infty$. Therefore, Lemma 3.3 gives

$$(z_i - x_i)d_i'(x_i, y_i) \le (z_i - y_i)(x_i - y_i).$$

Thus, we need only show that this inequality is strict when $x_i \neq y_i$. First, if $z_i = x_i$, we have:

$$(z_i - x_i)d_i'(x_i, y_i) = 0 < (x_i - y_i)^2 = (z_i - y_i)(x_i - y_i).$$

Similarly, if $z_i = y_i$,

$$(z_i - x_i)d_i'(x_i, y_i) = (y_i - x_i)d_i'(x_i, y_i) < 0 = (z_i - y_i)(x_i - y_i).$$

Now, we can assume x_i, y_i, and z_i are distinct, and thus we can proceed as in the proof of Lemma 3.3:

1. $x_i < \min(y_i, z_i)$:
   If $z_i < y_i$, the strict inequality is already present in the proof of Lemma 3.3. If $y_i < z_i$, we repeat the reasoning of the respective case for Lemma 3.3, but take δ to

be any number strictly greater than x_i, y_i, and z_i. Then,

$$\begin{aligned}(x_i - y_i)(z_i - y_i) &\geq \frac{\delta - y_i}{\delta - x_i}(x_i - y_i)(z_i - x_i)\\ &> (x_i - y_i)(z_i - x_i)\\ &\geq d_i'(x_i, y_i)(z_i - x_i),\end{aligned}$$

the strict inequality coming from $0 < (\delta - y_i)/(\delta - x_i) < 1$ and $(x_i - y_i)(z_i - x_i) < 0$. The last inequality follows from Assumption 3.7.1.

2. $x_i > \max(y_i, z_i)$:
   This case follows the corresponding case in the proof of Lemma 3.3, but using the strict inequality from Assumption 3.7.1 in the last step.
3. $z_i < x_i < y_i$:
   Again, we follow the respective case in Lemma 3.3, but use the strict inequality from Assumption 3.7.1 in the last step.
4. $y_i < x_i < z_i$:
   As in the first case, we use the reasoning of the corresponding part of Lemma 3.3, but take any $\delta > x_i, y_i, z_i$. Then

$$\begin{aligned}(x_i - y_i)(z_i - y_i) &\geq \frac{\delta - y_i}{\delta - x_i}(x_i - y_i)(z_i - x_i)\\ &> (x_i - y_i)(z_i - x_i)\\ &\geq d_i'(x_i, y_i)(z_i - x_i).\end{aligned}$$

□

Lemma 3.9. *Let $\{x^k\}$ be a PMDR sequence where $\mu = 1$ and the double regularization is based on coercive terms d_i conforming Assumption 3.7. Then, if the solution set of (1) is non-empty, one has $x^{k+1} - x^k \to 0$.*

Proof: Let z be a solution of the variational inequality (1). The quasi-Fejér convergence of the PMDR sequence, shown in Lemma 3.5, implies that $\{x^k\}$ is bounded, so the sequence $\{x^{k+1} - x^k\}$ is also bounded. Thus, it suffices to show that 0 is its only possible limit point. Let $\mathcal{K} \subset \mathbb{N}$ be any infinite index set over which $x^{k+1} - x^k$ is convergent. By passing to subsequences, we may assume without loss of generality that $\{x^k\}$ and $\{x^{k+1}\}$ converge over $\mathcal{K}$ as well. Let $\bar{x}$ and $\tilde{x}$ be the respective limit points of $\{x^k\}$ and $\{x^{k+1}\}$. Since $x^{k+1} - x^k \to_{\mathcal{K}} \tilde{x} - \bar{x}$, we need only demonstrate that $\tilde{x} = \bar{x}$. Define two index sets

$$I(\bar{x}, \tilde{x}) \stackrel{\text{def}}{=} \{i \mid \bar{x}_i = \tilde{x}_i\} \qquad J(\bar{x}, \tilde{x}) \stackrel{\text{def}}{=} \{i \mid \bar{x}_i \neq \tilde{x}_i\}.$$

We claim that for any $i \in J(\bar{x}, \tilde{x})$, there exist a scalar $\eta_i \in (0, 1)$, a neighborhood A_i of $\bar{x}_i$, and a neighborhood C_i of $\tilde{x}_i$ such that, for all $x_i \in A_i \cap \mathbb{R}_{++}$ and $y_i \in C_i \cap \mathbb{R}_{++}$,

$$(z_i - x_i)d_i'(x_i, y_i) \leq \eta_i(z_i - y_i)(x_i - y_i). \tag{23}$$

To establish the claim, consider three possibilities:

1. $\bar{x}_i, \tilde{x}_i > 0$. Lemma 3.8 gives

$$(z_i - \bar{x}_i)d_i'(\bar{x}_i, \tilde{x}_i) < (z_i - \tilde{x}_i)(\bar{x}_i - \tilde{x}_i).$$

Thus, there exists an $\epsilon \in (0, 1)$ such that

$$(z_i - \bar{x}_i)d_i'(\bar{x}_i, \tilde{x}_i) < (1 - \epsilon)(z_i - \tilde{x}_i)(\bar{x}_i - \tilde{x}_i).$$

Since both sides of this inequality are continuous in x_i and y_i, there must be neighborhoods $A_i \ni \bar{x}_i$ and $C_i \ni \tilde{x}_i$ where the inequality (23) holds with $\eta_i = 1 - \epsilon$.
2. $0 = \bar{x}_i < \tilde{x}_i$. This situation can be analyzed by subcases considering the relative position of z_i:

   - $0 = z_i = \bar{x}_i < \tilde{x}_i$. In this case, (23) is a direct consequence of Assumption 3.7.2.
   - $0 = \bar{x}_i < z_i < \tilde{x}_i$. Here, one may simply use the signs of the terms appearing in (23). For any x_i sufficiently close to $\bar{x}_i = 0$ and y_i sufficiently close to $\tilde{x}_i$, one has

$$(z_i - x_i)d_i'(x_i, y_i) < 0 < (1/2)(z_i - y_i)(x_i - y_i).$$

   - $0 = \bar{x}_i < \tilde{x}_i \leq z_i$. For x_i close enough to $\bar{x}_i = 0$ and y_i close enough to $\tilde{x}_i$,

$$\begin{aligned} & d_i'(x_i, y_i) \leq x_i - y_i && \text{[by Assumption 3.7.1]} \\ \Rightarrow\ & (z_i - x_i)d_i'(x_i, y_i) \leq (z_i - x_i)(x_i - y_i) && [\text{since } z_i - x_i > 0] \\ & \leq (1/2)(z_i - y_i)(x_i - y_i), \end{aligned}$$

   where the last inequality follows from $x_i - y_i < 0$ and $0 < (1/2)(z_i - x_i) < z_i - x_i$.

3. $0 = \tilde{x}_i < \bar{x}_i$. Once again, we consider the relative position of z_i:

   - $0 = \tilde{x}_i = z_i < \bar{x}_i$. For x_i close enough to $\bar{x}_i$ and y_i close enough to $\tilde{x}_i = 0$, $y_i < x_i$, then

$$\begin{aligned} & \frac{y_i(x_i - y_i)}{2x_i} < \frac{y_i(x_i - y_i)}{x_i} \leq d_i'(x_i, y_i) \\ \Rightarrow\quad & (1/2)y_i(x_i - y_i) \leq x_i d_i'(x_i, y_i) \\ \Rightarrow\quad & (1/2)(z_i - y_i)(x_i - y_i) \geq (z_i - x_i)d_i'(x_i, y_i). \end{aligned}$$

   - $0 = \tilde{x}_i < z_i < \bar{x}_i$. Once again, an argument based only on signs suffices. For x_i sufficiently close to $\bar{x}_i$ and y_i sufficiently close to $\tilde{x}_i$,

$$(z_i - x_i)d_i'(x_i, y_i) < 0 < (1/2)(z_i - y_i)(x_i - y_i).$$

   - $0 = \tilde{x}_i < \bar{x}_i \leq z_i$. Let $\epsilon \in (0, 1)$ be small enough that

$$(z_i - \bar{x}_i) < (1 - \epsilon)(z_i - \tilde{x}_i).$$

By continuity, there exist neighborhoods $A_i \ni \bar{x}_i$ and $C_i \ni \tilde{x}_i$ such that for $x_i \in A_i$ and $y_i \in C_i \cap \mathbb{R}_{++}$, one has $(z_i - x_i) < (1-\epsilon)(z_i - y_i)$, and thus, since $0 < d_i'(x_i, y_i) \le x_i - y_i$,

$$(z_i - x_i) d_i'(x_i, y_i) < (1-\epsilon)(z_i - y_i) d_i'(x_i, y_i) \le (1-\epsilon)(z_i - y_i)(x_i - y_i).$$

Therefore, we conclude that (23) holds. For $i \in I(\bar{x}, \tilde{x})$, define $A_i = C_i = \mathbb{R}_{++}$. Then define Cartesian product neighborhoods

$$A \stackrel{\text{def}}{=} A_1 \times A_2 \times \cdots \times A_n \quad C \stackrel{\text{def}}{=} C_1 \times C_2 \times \cdots \times C_n$$

of $\bar{x}$ and $\tilde{x}$, respectively, along with $\bar{\eta} \stackrel{\text{def}}{=} \max_{i \in J(\bar{x}, \tilde{x})}\{\eta_i\} \in (0, 1)$. Finally, let $\zeta \in \mathbb{R}^n$ be given by

$$\zeta_i \stackrel{\text{def}}{=} \begin{cases} 1, & \text{if } i \in I(\bar{x}, \tilde{x}), \\ \bar{\eta}, & \text{if } i \in J(\bar{x}, \tilde{x}). \end{cases}$$

Then, using Lemma 3.4 with the above definitions of A and C, we observe that for $k \in \mathcal{K}$ large enough,

$$\begin{aligned} &\langle z - x^{k+1}, \nabla_1 \tilde{d}(x^{k+1}, x^k) - e^{k+1} \rangle \\ &\le \sum_{i=1}^n \left\{ \frac{1+\zeta_i}{2}\left(\left(z_i - x_i^k\right)^2 - \left(z_i - x_i^{k+1}\right)^2\right) - \frac{1-\zeta_i}{2}\left(x_i^{k+1} - x_i^k\right)^2 \right\} + \langle e^{k+1}, x^{k+1} - z \rangle. \end{aligned}$$

Recalling that z is a solution of (1), we may use reasoning similar to Lemma 3.5's to establish, for sufficiently large $k \in \mathcal{K}$, that

$$\begin{aligned} 0 \le \sum_{i=1}^n &\left\{ \frac{1+\zeta_i}{2}\left(\left(z_i - x_i^k\right)^2 - \left(z_i - x_i^{k+1}\right)^2\right) - \frac{1-\zeta_i}{2}\left(x_i^{k+1} - x_i^k\right)^2 \right\} \\ &+ \langle e^{k+1}, x^{k+1} - z \rangle. \end{aligned}$$

Taking limits over $k \in \mathcal{K}$, and recalling that $\bar{x}_i = \tilde{x}_i$ for $i \in I(\bar{x}, \tilde{x})$, one obtains:

$$0 \le \sum_{i \in J(\bar{x}, \tilde{x})} \left\{ \frac{1+\bar{\eta}}{2}((z_i - \bar{x}_i)^2 - (z_i - \tilde{x})^2) - \frac{1-\bar{\eta}}{2}(\tilde{x}_i - \bar{x}_i)^2 \right\} + \langle 0, \tilde{x} - z \rangle.$$

Using once more the definition of $I(\bar{x}, \tilde{x})$, we recover

$$0 \le \sum_{i=1}^n \left\{ \frac{1+\bar{\eta}}{2}((z_i - \bar{x}_i)^2 - (z_i - \tilde{x})^2) - \frac{1-\bar{\eta}}{2}(\tilde{x}_i - \bar{x}_i)^2 \right\},$$

or equivalently,

$$\frac{1-\bar{\eta}}{2}\|\tilde{x}-\bar{x}\|^2 \le \frac{1+\bar{\eta}}{2}(\|z-\bar{x}\|^2 - \|z-\tilde{x}\|^2).$$

Quasi-Fejér convergence implies, via Proposition 2.6, that $\lim_{k\to\infty}\|z-x^k\|$ exists. Since both $\tilde{x}$ and $\bar{x}$ are limit points of $\{x^k\}$, we conclude that $\|z-\tilde{x}\| = \|z-\bar{x}\|$. Therefore, one has $\|\tilde{x}-\bar{x}\| \le 0$, that is, $\tilde{x}=\bar{x}$. □

Lemma 3.9 implies that, under Assumption 3.7, the hypotheses of [31, Theorem 2.7] continue to hold when $\mu = 1$. Then, by essentially identical reasoning to Proposition 3.6, we may assert:

Proposition 3.10. *The sequence computed by the PMDR using a double regularization with $\mu = 1$, and based on coercive terms conforming to Assumption 3.7, converges to a solution of (1), if any exist.*

3.2. A quadratic convergence rate result

We now consider the special case of applying the PMDR when $T = \partial f$, the subgradient mapping of a closed proper convex function *f*. In this case, problem (1) is equivalent to minimizing *f* over *B*. Suppose further that the solution set $X^* = \operatorname{Arg\,min}_{x\in B}\{f(x)\}$ constitutes a set of *weak sharp minima* [7, 17], that is, there exists a scalar $\nu > 0$ such that

$$\operatorname{dist}(x, X^*) \le \nu(f(x) - f^*) \qquad \forall x \in B, \tag{24}$$

where f^* is the value of *f* at any optimal solution. Results similar to the following proposition are established in [3] for the Φ_2 class of φ-divergence regularizations, which we show in Section 4.1 below to be a special case of our general double regularization approach.

Proposition 3.11. *Suppose the PMDR algorithm is applied in the case that T is the subgradient of a closed proper convex function f whose nonempty set of minima X^* over B is weak sharp. If all iterates are computed exactly, that is, $e^k \equiv 0$, then $\{f(x^k)\}$ and $\{x^k\}$ converge globally Q-quadratically to the optimal value f^* and to a point in X^*, respectively.*

Proof: The proof resembles that of [3, Theorem 6.1]. For a given *k*, let w^k denote the unique point in X^* such that $\|w^k - x^k\| = \operatorname{dist}(x^k, X^*)$. From (20) with $e^{k+1} = 0$ and $T = \partial f$, we obtain $-(1/\alpha_k)\nabla_1 D(x^{k+1}, x^k) \in \partial f(x^{k+1})$. Applying the subgradient inequality,

$$\begin{aligned} & f^* \ge f(x^{k+1}) + \langle w^k - x^{k+1}, -(1/\alpha_k)\nabla_1 D(x^{k+1}, x^k)\rangle \\ \Leftrightarrow \quad & f(x^{k+1}) - f^* \le \frac{1}{\alpha_k}\langle w^k - x^{k+1}, \nabla_1 D(x^{k+1}, x^k)\rangle. \end{aligned}$$

Employing Lemma 3.4 with $x = x^{k+1}$, $y = x^k$, $z = w^k$, and $\zeta_i(y^k, A, C) = 1$ for all $i = 1, \ldots, n$, which is possible by Lemma 3.3, and using the lower bound $\alpha_k \geq \underline{\alpha}$,

$$\begin{aligned} f(x^{k+1}) - f^* &\leq \frac{\mu + 1}{2\alpha_k}\left(\|w^k - x^k\|^2 - \|w^k - x^{k+1}\|^2\right) \\ &\leq \frac{\mu+1}{2\underline{\alpha}}\|w^k - x^k\|^2 \\ &= \frac{\mu+1}{2\underline{\alpha}}\operatorname{dist}(x^k, X^*)^2. \end{aligned}$$

Substituting the weak sharp minimum condition (24) on the right and left sides of this inequality, respectively, we obtain

$$\begin{aligned} f(x^{k+1}) - f^* &\leq \frac{(\mu+1)\nu^2}{2\underline{\alpha}}\left(f(x^k) - f^*\right)^2 \\ \operatorname{dist}(x^{k+1}, X^*) &\leq \frac{(\mu+1)\nu}{2\underline{\alpha}}\operatorname{dist}(x^k, X^*)^2. \end{aligned}$$

Since k is arbitrary, these inequalities establish global Q-quadratic convergence. □

We remark that in the case $B = \mathbb{R}^n$, in which case our assumptions require the classical choice $D(x, y) = (1/2)\|x - y\|^2$, [17] shows that the weak sharp minimum condition implies *finite* convergence.

Sufficient conditions implying that (24) holds include f being a proper polyhedral convex function [7, Corollary 3.6]. Thus, the quadratic convergence result will hold in both primal and dual applications of the PMDR to linear programming problems. In fact, the results of [3, Section 6] only explicitly consider linear programming, but the analysis immediately generalizes to any situation satisfying (24).

4. Examples of double regularizations

This section presents examples of coercive regularizations that conform to Assumption 2.3, and may thus be used to build double regularizations for convergent proximal methods.

We will focus on the case $(a, b) = \mathbb{R}_{++}$; given a regularization for $\mathbb{R}_{++}$, it is straightforward to use argument translations and sign changes to produce regularizations for the cases (a, ∞) and $(-\infty, b)$, where $a, b \in \mathbb{R}$. For an arbitrary finite interval (a, b), the following simple construction applies:

Lemma 4.1. *Let $d_+, d_- : \mathbb{R} \times \mathbb{R}_{++} \to (-\infty, \infty]$ be functions conforming to Assumption 2.3 for the domain $\mathbb{R}_{++}$. Then, given $a, b \in \mathbb{R}$, $a < b$ and $\zeta \in (0, 1)$,*

$$d(x, y) \stackrel{\text{def}}{=} \zeta d_+(x - a, y - a) + (1 - \zeta) d_-(b - x, b - y)$$

conforms to Assumption 2.3, but for (a, b).

Proof: Let $x, y \in (a, b)$. Using Assumption 2.3 for d we have

$$\frac{\zeta(x-a-y+a)(y-a)}{x-a} \le \zeta d'_+(x-a, y-a) \le \zeta(x-a-y+a),$$

and

$$\begin{aligned}\frac{(1-\zeta)(b-x-b+y)(b-y)}{b-x} &\le (1-\zeta)d'_-(b-x, b-y)\\ &\le (1-\zeta)(b-x-b+y).\end{aligned}$$

Simplifying, multiplying the second inequality by -1, and adding, we arrive at

$$\begin{aligned}&\frac{\zeta(x-y)(y-a)}{x-a} + (1-\zeta)(x-y)\\ &\quad\le \zeta d'_+(x-a, y-a) - (1-\zeta)d'_-(b-x, b-y)\\ &\quad\le \frac{(1-\zeta)(x-y)(b-y)}{b-x} + \zeta(x-y).\end{aligned}$$

On the other hand, since $x, y \in (a, b)$, we also have

$$\frac{(x-y)(y-a)}{x-a} \le x-y \qquad x-y \le \frac{(x-y)(b-y)}{b-x}.$$

Hence,

$$\begin{aligned}\frac{(x-y)(y-a)}{x-a} &\le \zeta d'_+(x-a, y-a) - (1-\zeta)d'_-(b-x, b-y)\\ &\le \frac{(x-y)(b-y)}{b-x}.\end{aligned}$$

Finally, differentiating the definition of d with respect to x and observing that the chain rule inverts the sign of the d_- term, we obtain

$$\frac{(x-y)(y-a)}{x-a} \le d'(x, y) \le \frac{(x-y)(b-y)}{b-x},$$

for all $x, y \in (a, b)$. □

4.1. φ-divergences

As already discussed, the results of this paper may be seen as generalizing ideas in [2, 3]. There, Auslender et al. obtain double regularizations for the positive orthant by adding the squared Euclidean norm to rescaled *φ-divergences*.

We now show that Assumption 2.3 generalizes the Φ_2 class from [3]. There, the coercive part of the double regularization components have the form

$$d_i(x, y) = y^2\varphi\left(\frac{x}{y}\right),$$

for some $\varphi : \mathbb{R}_+ \to \mathbb{R}$. In this case, Assumption 2.3 becomes

$$\forall x, y \in \mathbb{R}_{++} : \quad \frac{(x-y)y}{x} \le y\varphi'\left(\frac{x}{y}\right) \le x - y.$$

The simple change of variables $t = x/y$ converts this condition into

$$\forall t > 0 : \quad (1 - 1/t) \le \varphi'(t) \le (t - 1),$$

which is precisely the condition defining the Φ_2 class. From [3], we have the following examples of functions conforming to this last inequality:

1. $\varphi(t) = t\ln(t) - t + 1$;
2. $\varphi(t) = 2(\sqrt{t} - 1)^2$;
3. $\varphi(t) = -\ln(t) + t - 1$.

In particular, the function $\varphi(t) = -\ln(t) + t - 1$ generates the log-quadratic regularization, the first double regularization studied in the literature [2]. Moreover, this regularization has

$$d'_\varphi(x, y) = \frac{(x-y)y}{x},$$

so its derivative coincides with the lower bound imposed by Assumption 2.3.

4.2. Bregman distances

Another standard construction for producing regularization distances for proximal methods is the *Bregman distance*

$$d(x, y) = h(x) - h(y) - h'(y)(x - y),$$

where h is some strictly convex function.

We now present some functions that can be used to derive Bregman distances conforming to Assumption 2.3 after rescaling by $h_i''(y_i)$ [31, Section 2.2.1]. One may solve monotone variational inequality problems using such Bregman distances without resorting to additional problem assumptions like paramonotonicity [9, 19]. We note that [31, Section 4] presents similar results, but under the stronger rescaling

$$\alpha(y) = \max_{i=1,\ldots,n} \{h_i''(y_i)\}, \qquad \tilde{d}(x, y) = \sum_{i=1}^{n} \frac{h_i(x_i) - h_i(y_i) - h_i'(y_i)(x_i - y_i)}{\alpha(y)}.$$

In this case, the rescaling factor $\alpha(y)$ may go to infinity very quickly, and uniformly for all coordinates, including coordinates that remain bounded away from their interval endpoints. We expect that this approach would not be practical in comparison to the double regularization technique suggested here. We first introduce a lemma making it easier to verify whether Assumption 2.3 holds:

Lemma 4.2. *Let* $h : \mathbb{R} \to (-\infty, \infty]$, $\operatorname{int}\operatorname{dom} h = \mathbb{R}_{++}$. *If* $h''(x)$ *is nonincreasing and* $x^2 h''(x)$ *is nondecreasing over* $x \in \mathbb{R}_{++}$, *then the rescaled Bregman distance*

$$d(x, y) \stackrel{\text{def}}{=} \frac{h(x) - h(y) - h'(y)(x - y)}{h''(y)}$$

conforms to Assumption 2.3 for $(a, b) = \mathbb{R}_{++}$.

Proof: Letting $(a, b) = \mathbb{R}_{++}$, and substituting the definition of $d(x, y)$ above, the lower bound for $d'(x, y)$ in Assumption 2.3 reduces to

$$h''(y)(x - y)\frac{y}{x} \leq (h'(x) - h'(y)).$$

To show that this inequality holds, we consider two cases:

1. If $0 < x < y$,

$$\begin{aligned} h'(y) - h'(x) &= \int_x^y h''(z)dz \\ &\leq \int_x^y \frac{y^2 h''(y)}{z^2} dz \qquad [\text{since } x^2 h''(x) \text{ nondecreasing}] \\ &= h''(y)\left(-\frac{1}{y} + \frac{1}{x}\right) y^2 \\ &= h''(y)\frac{y - x}{xy} y^2 \\ &= h''(y)(y - x)\frac{y}{x}. \end{aligned}$$

2. If $0 < y < x$, similar reasoning produces

$$h'(x) - h'(y) = \int_y^x h''(z)dz \geq \int_y^x \frac{y^2 h''(y)}{z^2} dz = h''(y)(x - y)\frac{y}{x},$$

   where the inequality again results from $x^2 h''(x)$ being nondecreasing.

The *upper* bound from Assumption 2.3 reduces to $h'(x) - h'(y) \leq h''(y)(x - y)$. Once again, we analyze two possibilities: If $0 < x_i < y_i$,

$$h'(y) - h'(x) = \int_x^y h''(z)dz \geq \int_x^y h''(y)dz = h''(y)(y - x),$$

where the inequality follows from $h''(x)$ being nonincreasing. The case $0 < y < x$ is analogous. □

Two examples for Bregman functions that meet the hypotheses of Lemma 4.2 are:

1. $h(x) = \text{dilog}(e^x) + x\ln(e^x - 1)$, where dilog$(\cdot)$ is the dilogarithm function [22]:

$$\text{dilog}(z) \stackrel{\text{def}}{=} \int_1^z \frac{\ln(t)}{1-t}dt.$$

In this case,

$$h''(x) = \frac{e^x}{e^x - 1},$$

which is clearly nonincreasing.
To show that $x^2h''(x)$ is nondecreasing, we calculate

$$\frac{d}{dx}x^2h''(x) = \frac{e^x x(2e^x - x - 2)}{(e^x - 1)^2}.$$

For $x > 0$, this function has the same sign as $2e^x - x - 2$. Now, $2e^x - x - 2$ evaluates to 0 at $x = 0$, and is strictly increasing for $x > 0$. Hence, the derivative of $x^2h''(x)$ is nonnegative, and therefore $x^2h''(x)$ is nondecreasing.
2. $h(x) = x^\alpha - x^\beta$, $\alpha \geq 1$, $\beta \in (0, 1)$. In this case,

$$h''(x) = \alpha(\alpha - 1)x^{\alpha-2} + \beta(1-\beta)x^{\beta-2}.$$
$$x^2h''(x) = \alpha(\alpha - 1)x^{\alpha} + \beta(1-\beta)x^{\beta}.$$

Clearly, $x^2h''(x)$ is nondecreasing on $\mathbb{R}_{++}$. Also $h''(x)$ is nonincreasing if and only if $\alpha \leq 2$. Hence, the Bregman distance given by this choice of h conforms to Assumption 2.3 when $\alpha \in [1, 2]$ and $\beta \in (0, 1)$.

5. Penalties and multiplier methods

Suppose one applies a generalized proximal method with distance kernel $d(x, y)$ to the dual of a complementarity problem or variational inequality. Then, as noted in Section 1, one obtains a generalized augmented Lagrangian method involving the penalty term $P'(\cdot, y^k)$ defined in (10), simplifying to (17) in the coercive case. Since $P'(\cdot, y^k)$ is the inverse of the mapping $\nabla_1\tilde{d}(\cdot, y^k)$, its componentwise integral $P(\cdot, y^k)$ is, up to a constant, equal to the *convex conjugate* [27, Chapter 12] $(\tilde{d}(\cdot, y))^*$ of the function $\tilde{d}(\cdot, y^k)$.

We now investigate the properties of such conjugates. In particular, we consider which functions $P(\cdot, y)$ can be expressed as conjugates of double regularizations conforming to Assumption 2.3.

Proposition 5.1. *Let $\mu \geq 1$. Let $P_i : \mathbb{R} \times \mathbb{R}_{++} \to \mathbb{R}$, and denote by $P_i'(\cdot, y_i)$ its derivative with respect to the first argument. If $P_i'(\cdot, y_i)$ is continuous, $P_i'(\cdot, y_i)$ is both* *strictly increasing and strictly positive for each* $y_i > 0$*, and one has for all* $u \in \mathbb{R}$ *that*

$$\frac{u}{\mu+1} + y_i \le P_i'(u, y_i) \le \frac{u + (\mu - 1)y_i + \sqrt{\left(u + (\mu - 1)y_i\right)^2 + 4\mu y_i^2}}{2\mu}, \quad (25)$$

then there is a double regularization component $\tilde{d}_i$ *conforming to Assumption 2.3 such that* $P_i(\cdot, y_i) = (\tilde{d}_i(\cdot, y_i))^*$*, where the symbol* * *denotes the convex conjugacy operator [27, Chapter 12].*

Proof: Take any $y_i > 0$. Since $P_i'(\cdot, y_i)$ is strictly increasing, $P_i(\cdot, y_i)$ is strictly convex. Let us denote the convex conjugate of this function by $\tilde{d}_i(\cdot, y_i)$. We then have:

1. $\tilde{d}_i(\cdot, y_i)$ is closed, strictly convex, and essentially smooth, since it is the conjugate of a differentiable, strictly convex function on [27, Theorem 26.3].
2. $\operatorname{int}\operatorname{dom}\tilde{d}_i(\cdot, y_i) = \operatorname{dom}\tilde{d}_i'(\cdot, y_i) = \operatorname{rge}P_i'(\cdot, y_i) = \mathbb{R}_{++}$. Here, the first equality follows from [27, Theorem 26.1], the second from [27, Corollary 23.5.1], and the third from the bounds on $P_i'(\cdot, y_i)$.
3. $\tilde{d}_i(\cdot, y_i)$ attains its minimum at y_i: since both bounds on $P_i'(\cdot, y_i)$ are equal to y_i at 0, we have $P_i'(0, y_i) = y_i$. Then, $\tilde{d}_i'(y_i, y_i) = 0$ [27, Corollary 23.5.1].

Now define $d_i(x_i, y_i) \stackrel{\text{def}}{=} \tilde{d}_i(x_i, y_i) - (\mu/2)\|x_i - y_i\|^2$. In view of the three facts above, we need only prove that d_i meets the bounds imposed by Assumption 2.3, and that $\tilde{d}_i'$ is continuous. We begin with the bounds:

1. Take any $x_i > 0$, and let

$$u = (\mu + 1)(x_i - y_i).$$

The lower bound on $P_i'(\cdot, y_i)$ implies that

$$\frac{u}{\mu+1} + y_i \le P_i'(u, y_i) \quad \Leftrightarrow \quad x_i \le P_i'((\mu + 1)(x_i - y_i), y_i).$$

As $P_i'(\cdot, y_i)$ is strictly increasing, so is its inverse, $\tilde{d}_i'(\cdot, y_i)$ [27, Corollary 23.5.1]. Applying this function to both sides of the above inequality, and using the definition of d_i,

$$\begin{aligned} & \tilde{d}_i'(x_i, y_i) \le (\mu + 1)(x_i - y_i) \\ \Leftrightarrow \quad & d_i'(x_i, y_i) + \mu(x_i - y_i) \le (\mu + 1)(x_i - y_i) \\ \Leftrightarrow \quad & d_i'(x_i, y_i) \le (x_i - y_i). \end{aligned}$$

2. Again, take any $x_i > 0$. We follow similar logic, but define u via

$$u = \frac{y_i(x_i - y_i)}{x_i} + \mu(x_i - y_i).$$

Multiplying through by $x > 0$, we obtain a quadratic equation in x. Applying the quadratic formula,

$$x_i = \frac{u + (\mu - 1)y_i \pm \sqrt{(u + (\mu - 1)y_i)^2 + 4\mu y_i^2}}{2\mu}.$$

Since $\mu \geq 1$ and $y_i > 0$, there is only one positive solution, and we obtain

$$x_i = \frac{u + (\mu - 1)y_i + \sqrt{(u + (\mu - 1)y_i)^2 + 4\mu y_i^2}}{2\mu}.$$

The hypothesized upper bound on $P_i'(\cdot, y_i)$ guarantees

$$P_i'(u, y_i) \leq \frac{u + (\mu - 1)y_i + \sqrt{(u + (\mu - 1)y_i)^2 + 4\mu y_i^2}}{2\mu}.$$

Substituting the definition of u and applying the strictly increasing function $\tilde{d}_i'(\cdot, y_i)$ to both sides yields

$$\frac{y_i(x_i - y_i)}{y_i} + \mu(x_i - y_i) \leq \tilde{d}_i'(x_i, y_i) \quad \Leftrightarrow \quad \frac{y_i(x_i - y_i)}{x_i} \leq d_i'(x_i, y_i).$$

Thus, the bounds on d_i are satisfied. Finally, consider the continuity of $\tilde{d}_i'$. Let $x_i^k \to \bar{x}_i > 0$ and $y_i^k \to \bar{y}_i > 0$ be convergent sequences in $\mathbb{R}_{++}$. Let $u^k = \tilde{d}_i'(x_i^k, y_i^k)$. Then by the inverse properties of the conjugate, $x_i^k = P_i'(u^k, y_i^k)$. By the bounds we have just established,

$$\frac{y_i^k\left(x_i^k - y_i^k\right)}{y_i^k} + \mu\left(x_i^k - y_i^k\right) \leq u^k \leq x_i^k - y_i^k,$$

so $\{u^k\}$ is bounded. Let $\bar{u}$ be one of its limit points and $\mathcal{K} \subset \mathbb{N}$ be the respective index set. We then have $x_i^k \to_{\mathcal{K}} \bar{x}_i$, $y_i^k \to_{\mathcal{K}} \bar{y}_i$ and $u^k \to_{\mathcal{K}} \bar{y}_i$, so the continuity of P_i' ensures that $\bar{x}_i = P_i'(\bar{u}, \bar{y}_i)$, and thus $\bar{y}_i = \tilde{d}_i'(\bar{x}_i, \bar{y}_i)$. Thus, u^k is bounded and all its limit points are equal to $\tilde{d}_i'(\bar{x}_i, \bar{y}_i)$, so it converges to $\tilde{d}_i'(\bar{x}_i, \bar{y}_i)$. □

Hence, if $P : \mathbb{R}^n \times \mathbb{R}^n{}_{++} \to \mathbb{R}^n$ is composed of components conforming to the assumptions of the above Proposition, the generalized augmented Lagrangian method given by (11)–(12) is equivalent to the PMDR applied to the dual problem (7), with $e^k \equiv 0$. Figure 2 illustrates Proposition 5.1's penalty bounds. We summarize the properties of the augmented Lagrangian method in the following proposition:

Proposition 5.2. *Let F be a continuous monotone function $\mathbb{R}^n \to \mathbb{R}^n$ and assume that the complementarity problem (2) has a solution. Suppose $P : \mathbb{R}^n \times \mathbb{R}^n_{++} \to \mathbb{R}$*

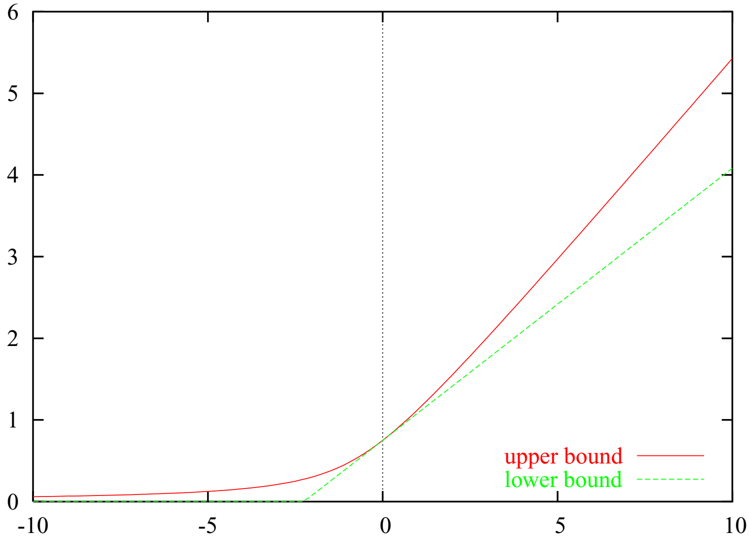

Figure 2. Limits given in Proposition 5.1 for the derivatives of a penalty based on a double regularization.

conforms to the hypotheses of Proposition 5.1 with $\mu > 1$*, and* $\{x^k\}, \{y^k\} \subset \mathbb{R}^n$ *conform to the recursions (11)–(12). Then* $\{y^k\}$ *converges to a dual solution of (2), that is, to a limit* $y^* \geq 0$ *such that* $y^* = F(x^*)$ *and* $\langle x^*, y^*\rangle = 0$ *for some* $x^* \geq 0$ *solving (2). Furthermore,* $\liminf_{k\to 0} x^k \geq 0$ *(interpreted componentwise),* $F(x^k) \to y^*$*,* $\langle x^k, y^k\rangle \to 0$*, and all limit points of* $\{x^k\}$ *are solutions of (2).*

Proof: Proposition 5.1 asserts that $P(\cdot, y^k) = (\tilde{d}(\cdot, y))^*$, and hence that (17) holds, where $\tilde{d}(\cdot, \cdot)$ is a double regularization conforming to Assumption 2.1. Thus, as demonstrated in Section 1, the recursions (11)–(12) are equivalent to applying the PMDR method to the dual formulation (7) of (2). As also seen in Section 1, existence of solution to (2) implies a solution to (7) exists. Since $\mu > 1$, Proposition 3.6 asserts the convergence of $\{y^k\}$ to a solution y^* of (7).

It remains to prove the assertions about $\{x^k\}$. Substituting (12) into (11), we obtain $y^{k+1} = F(x^{k+1})$, and thus $y^k = F(x^k)$ for all $k \geq 1$, and $F(x^k) \to y^*$.

Next, substituting (12) into the first inequality in (25), we obtain for $i = 1, \ldots, n$ and all $k > 0$ that

$$\frac{-\alpha_k x_i^k}{\mu + 1} + y_i^k \leq y_i^{k+1} \quad \Leftrightarrow \quad x_i^k \geq \frac{\mu + 1}{\alpha_k}\big(y_i^k - y_i^{k+1}\big).$$

Since $\{y_i^k\}$ converges and α_k is bounded below, we obtain that $\liminf_{k\to\infty} x_i^k \geq 0$, so $\liminf_{k\to\infty} x^k \geq 0$ componentwise.

Lemma 3.3 allows us to apply Lemma 3.4 under the substitutions $B \leftarrow \mathbb{R}^n{}_+$, $A, C \leftarrow \mathbb{R}^n{}_{++} = \operatorname{int} B$, $z \leftarrow 0 \in B$, $x \leftarrow y^{k+1}$, $y \leftarrow y^k$, and $\zeta(z_i, A, C) \leftarrow 1$ for all

$i = 1, \ldots, n$ to obtain

$$\langle -y^{k+1}, \nabla_1 \tilde{d}(y^{k+1}, y^k)\rangle \leq \frac{\mu+1}{2}(\|y^k\|^2 - \|y^{k+1}\|^2) - \frac{\mu-1}{2}\|y^{k+1} - y^k\|^2.$$

Next, note that applying the conjugacy relation to (12) yields $\nabla_1 \tilde{d}(y^{k+1}, y^k) = -\alpha_k x^{k+1}$, which we substitute into the above inequality to obtain

$$\langle y^{k+1}, x^{k+1}\rangle \leq \frac{\mu+1}{2\alpha_k}(\|y^k\|^2 - \|y^{k+1}\|^2).$$

Since $\{y^k\}$ converges and $\{\alpha_k\}$ is bounded below, we obtain $\limsup_{k\to\infty}\langle x^k, y^k\rangle \leq 0$. On the other hand, since $y^k \to y^* \geq 0$, $y^k > 0$ for all k by the positivity of the $P_i'(\cdot,\cdot)$, and $\liminf_{k\to\infty} x^k \geq 0$, it follows that $\liminf_{k\to 0}\langle x^k, y^k\rangle \geq 0$. Combining the lim inf and lim sup inequalities, we have $\langle x^k, y^k\rangle \to 0$.

Finally, let x^∞ be any limit point of $\{x^k\}$. Taking limits over an appropriate subsequence, and using the continuity of F, the just-established properties of $\{x^k\}$ imply $F(x^\infty) = y^* \geq 0$, $x^\infty \geq 0$, and $\langle x^\infty, y^*\rangle = 0$, so x^∞ is a solution to (2). □

In the interest of simplicity, Proposition 5.2 assumes exact solution of (11). However, by allowing $e^k \neq 0$ in the PMDR algorithm, the analysis can be extended to allow approximate solution of (11) using the criterion proposed in [16, Theorem 2].

We note also that our multiplier method convergence results should still hold in the more general setting where F is multivalued—that is, we simply take $B = \mathbb{R}^n_+$ in (1), and consider the dual problem $T^{-1}(y) + N_{\mathbb{R}^n_+}(y) \ni 0$ in the manner of (7). In this case, we obtain the augmented Lagrangian inclusion $0 \in T(x) - P'(-\alpha_k x^k, y^k)$ in place of (11); we omit the details of this generalization in the interest of simplicity.

5.1. *Proximal methods of multipliers*

We now briefly describe an alternative way to apply the PMDR algorithm to (2). Writing the "optimality" conditions for (3) in the form (5), we have $y = F(x)$, $-y \in N_{\mathbb{R}^n_+}(x)$. Rewriting the latter condition as $x \in N^{-1}_{\mathbb{R}^n_+}(-y)$ and then $-x \in N_{\mathbb{R}^n_+}(y)$, followed by some rearrangment, we obtain

$$0 = F(x) - y \qquad\qquad 0 \in x + N_{\mathbb{R}^n_+}(y).$$

This set of conditions is equivalent to problem (1) formulated in $\mathbb{R}^{2n}$, where $T : (x, y) \mapsto \{(F(x) - y, x)\}$ and $B = \mathbb{R}^n \times \mathbb{R}^n_+$. It is easily confirmed that if $F : \mathbb{R}^n \to \mathbb{R}^n$ is continuous and monotone, T defined in this manner is maximal monotone. We now consider applying the PMDR in $\mathbb{R}^{2n}$ to this problem. Note that for $i = 1, \ldots, n$, we have $a_i = -\infty$ and $b_i = \infty$, so Assumption 2.3.2c requires $d_i'(x_i, y_i) = x_i - y_i$. For $i = n+1, \ldots, 2n$, on the other hand, the d_i' should take the customary forms for $\mathbb{R}_+$. Shifting notation slightly so that the iterate sequence is denoted $\{(x^k, y^k)\}$ and $\tilde{d}(\cdot,\cdot)$ denotes the distance measure for just the last n components, the PMDR recursions become

$$0 = F(x^{k+1}) - y^{k+1} + \frac{\mu + 1}{\alpha_k}(x^{k+1} - x^k) \tag{26}$$
$$0 = x + \frac{1}{\alpha_k}\nabla_1 \tilde{d}(y^{k+1}, y^k). \tag{27}$$

Rewriting (27) as $-\alpha_k x^k = \nabla_1 \tilde{d}(y^{k+1}, y^k)$, applying the conjugacy transformation to obtain $y^{k+1} = P'(-\alpha_k x^{k+1}, y^k)$, and substituting into (26), we arrive at

$$0 = F(x^{k+1}) - P'(-\alpha_k x^{k+1}, y^k) + \frac{\mu + 1}{\alpha_k}(x^{k+1} - x^k) \tag{28}$$
$$y^{k+1} = P'(-\alpha_k x^{k+1}, y^k), \tag{29}$$

which is identical to (11)–(12), except for the additional term $((\mu + 1)/\alpha_k)(x^{k+1} - x^k)$. This kind of method is known as a *proximal method of multipliers*, with a history extending back to (29). We omit a formal result in interest of brevity, but it is easily shown in this case that $\{(x^k, y^k)\}$ converges to $(x^*, F(x^*))$, where x^* is some solution to (2); see for example [16, Theorem 4] or [4, Theorem 4.6] for similar results. Approximate solution of (28) is straightforward to include, as in [16, Theorem 4].

5.2. *Connections to the work of Chen and Mangasarian*

The penalty derivative upper bound in (25) corresponds to the lower bound in Assumption 2.3, and is proposed as the penalty term for a *log-quadratic multiplier* method in (3).

Examining the penalty derivative upper bound (25), we remark on a connection to [10]. The bound is exactly the *Chen-Harker-Kanzow-Smale plus function*, defined by

$$P'(w, \beta) \stackrel{\text{def}}{=} \frac{w + \sqrt{w^2 + 4\beta^2}}{2}, \tag{30}$$

computed at $w = u + (\mu - 1)y_i$ and $\beta = \sqrt{\mu}\, y_i$. The experiments in [10] use this function in a smoothing method—essentially a pure penalty algorithm with no explicit Lagrange multipliers—for complementarity problems. One may consider the log-quadratic multiplier method of [3] to be a related algorithm introducing explicit duality and Lagrange multipliers. Incidentally, this kind of penalty can be traced back even earlier, to the unpublished work of Xavier [34].

The Chen-Harker-Kanzow-Smale plus function was not the only smoothing function studied in [10]. Thus, it is natural to consider whether other penalties from [10] could be used to generate double regularizations and associated methods of multipliers. In particular, we consider the *neural network smooth plus function*, since it yielded the best numerical results in [10].

5.3. *The neural network smooth plus function*

In this section, we show that the neural network smooth plus function gives rise to a penalty corresponding to a double regularization. To do so, however, it appears necessary to set $\mu = 1$. To be assured of convergence, we must thus check whether the corresponding distance conforms not only to Assumption 2.3, but also to Assumption 3.7. Then, Proposition 3.10 will assure convergence.

Let us recall the formula for the neural network smooth plus function from [10]:

$$P'(w, \beta) \stackrel{\text{def}}{=} \beta \ln(e^{w/\beta} + 1).$$

We now consider whether a penalty of this form can made to conform to the hypotheses of Proposition 5.1. The analysis for the case $\mu > 1$ appears difficult, so we concentrate on $\mu = 1$. If one follows the transformation used to obtain the log-quadratic penalty from (30) with $\mu = 1$, one sets $w = u$ and $\beta = y_i$, producing

$$P_i'(u, y_i) = y_i \ln(e^{u/y_i} + 1). \tag{31}$$

However, this function cannot possibly conform to the bound (25), which requires $P_i'(0, y_i) = y_i$, whereas (31) implies $P_i'(0, y_i) = \ln(2) y_i$. However, a simple change of scale $w = u$, $\beta = y_i / \ln(2)$ remedies this difficulty, producing

$$P_i'(u, y_i) = y_i \log_2(2^{u/y_i} + 1). \tag{32}$$

We proceed by letting $\tilde{d}_i$ be the convex conjugate of $P_i(\cdot, y_i)$ as defined in (32), and then define d_i implicitly via (18) with $\mu = 1$, that is,

$$\tilde{d}_i'(\cdot, y_i) = (P_i'(\cdot, y_i))^{-1} \tag{33}$$

$$d_i'(x_i, y_i) = \tilde{d}_i'(x_i, y_i) - (x_i - y_i) \tag{34}$$

Since $\mu = 1$, we seek to show that the d_i implicitly defined by integrating meets Assumption 3.7. Then, Proposition 3.10 will guarantee that the proximal method based on the double regularization components $\tilde{d}_i$ is convergent. This convergence will imply convergence of the corresponding multiplier method using the penalty (32).

Lemma 5.3 *The function d_i defined by (32)–(34) conforms to Assumption 3.7.*

Proof: Inserting the definition (32) into (33) and solving for $\tilde{d}_i'(x_i, y_i)$, we obtain the explicit expression

$$\tilde{d}_i'(x_i, y_i) = y_i \log_2(2^{x_i/y_i} - 1). \tag{35}$$

As $\tilde{d}_i'(\cdot, \cdot)$ is clearly continuous, it remains to confirm Assumptions 3.7.1 and 3.7.2.

Let us first consider Assumption 3.7.1. It is easily confirmed that the bounds hold when $x_i = y_i$, so is suffices to prove, for all $x_i, y_i > 0$, $x_i \neq y_i$, that

$$
\begin{aligned}
& \frac{(x_i - y_i)y_i}{x_i} < d_i'(x_i, y_i) \leq x_i - y_i \\
\Leftrightarrow \quad & \frac{(x_i - y_i)y_i}{x_i} < y_i \log_2(2^{x_i/y_i} - 1) - (x_i - y_i) \leq x_i - y_i \\
\Leftrightarrow \quad & \frac{x_i^2 - y_i^2}{x_i} < y_i \log_2(2^{x_i/y_i} - 1) \leq 2(x_i - y_i) \\
\Leftrightarrow \quad & \frac{x_i}{y_i} - \frac{y_i}{x_i} < \log_2(2^{x_i/y_i} - 1) \leq 2\left(\frac{x_i}{y_i} - 1\right).
\end{aligned}
$$

If we define $t \stackrel{\text{def}}{=} x_i/y_i$, these bounds are equivalent to

$$\forall t > 0, t \neq 1: \quad t - 1/t < \log_2(2^t - 1) \leq 2t - 2. \tag{36}$$

The upper bound is easily proved, as $2t - 2 - \log_2(2^t - 1)$ is a strictly convex function with its minimum at 1 and minimum value 0. The lower bound is equivalent to $2^{-t} + 2^{-1/t} \leq 1$, with equality only for $t = 1$, as established in [24, Problem 23]. Thus, Assumption 3.7.1 holds.

Finally, we turn to Assumption 3.7.2. Take any $\bar{y}_i > 0$, select some $\zeta \in (0, 1)$, and define $C_i \stackrel{\text{def}}{=} (\bar{y}_i/2, 2\bar{y}_i)$. We will show for any $y_i \in C_i$ and x_i small enough,

$$\zeta \frac{(x_i - y_i)y_i}{x_i} \leq d_i'(x_i, y_i) = y_i \log_2(2^{x_i/y_i} - 1) - (x_i - y_i). \tag{37}$$

Since $\zeta < 1$ and $x_i < y_i$ for small x_i, inequality (37) is implied by

$$\zeta \frac{(x_i - y_i)y_i}{x_i} \leq y_i \log_2(2^{x_i/y_i} - 1) - \zeta(x_i - y_i) \tag{38}$$

$$
\begin{aligned}
\Leftrightarrow \quad & \zeta \frac{x_i^2 - y_i^2}{x_i} \leq y_i \log_2(2^{x_i/y_i} - 1) \\
\Leftrightarrow \quad & \zeta \geq \frac{x_i y_i \log_2(2^{x_i/y_i} - 1)}{x_i^2 - y_i^2} \\
\Leftrightarrow \quad & \zeta \geq \frac{x_i}{y_i} \log_2(2^{x_i/y_i} - 1) \frac{y_i^2}{x_i^2 - y_i^2}.
\end{aligned} \tag{39}
$$

Once again, we introduce the change of variables $t = x_i/y_i$, which reduces the above expression in x_i and y_i to $t \log_2(2^t - 1)/(t^2 - 1)$. We next claim that

$$\lim_{t \downarrow 0} \frac{t \log_2(2^t - 1)}{t^2 - 1} = 0. \tag{40}$$

As $\lim_{t\downarrow 0} 1/(t^2-1) = -1$, it suffices to show that

$$\lim_{t\downarrow 0} t\log_2(2^t-1) = 0.$$

Writing $t\log_2(2^t-1) = \log_2(2^t-1)/(1/t)$ and applying L'Hôpital's rule, one obtains

$$\lim_{t\downarrow 0} t\log_2(2^t-1) = \lim_{t\downarrow 0} \frac{-t^2 2^t}{2^t-1}.$$

Since $\lim_{t\downarrow 0} 2^t = 1$, it is sufficient to prove that

$$\lim_{t\downarrow 0} \frac{-t^2}{2^t-1} = 0,$$

which follows from a second use of L'Hôpital's rule. Thus, we have verified that the limit (40) holds. Therefore, there exists a $\bar{t} > 0$ such that for $0 < t \leq \bar{t}$

$$\frac{t\log_2(2^t-1)}{t^2-1} \leq \zeta.$$

Define $A_i \stackrel{\text{def}}{=} (0, \bar{t}\bar{y}_i/2)$. For $x_i \in A_i$ and $y_i \in C_i$, we have $x_i/y_i < \bar{t}$, and hence

$$\frac{x_i}{y_i}\log_2(2^{x_i/y_i}-1)\frac{y_i^2}{x_i^2-y_i^2} \leq \zeta.$$

Since inequality (39) is equivalent to (38), which implies (37), we conclude that Assumption 3.7.2 holds. □

Proposition 5.4. *Let F be a continuous monotone function $\mathbb{R}^n \to \mathbb{R}^n$ and assume that the complementarity problem (2) has a solution. Suppose that $\{x^k\}, \{y^k\} \subset \mathbb{R}^n$ conform to the recursions (11)–(12) with $P_i'(u, y_i) = y_i\log_2(2^{u/y_i}+1)$. Then $\{y^k\}$ converges to a dual solution of (2), $\liminf_{k\to 0} x^k \geq 0$, $F(x^k) \to y^*$, $\langle x^k, y^k\rangle \to 0$, and all limit points of $\{x^k\}$ are solutions of (2).*

Proof: Lemma 5.3 establishes that $P_i'(u, y_i) = y_i\log_2(2^{u/y_i}+1)$ corresponds to a $\mu = 1$ distance kernel $\tilde{d}_i'(x_i, y_i) = d_i'(x_i, y_i) + x_i - y_i$ meeting Assumption 3.7. so Proposition 3.10 implies convergence of $\{y^k\}$ to a solution y^* of (7).

It remains to prove the properties of the primal sequence $\{x^k\}$. We note that if the first inequality in (25) holds with $\mu = 1$, the properties of $\{x^k\}$ follow in exactly the same manner as in the proof Proposition 5.2, since those arguments did not depend on $\mu > 1$.

Therefore, it is sufficient to establish that $u/2 + y_i \leq y_i\log_2(2^{u/y_i}+1)$ for all $u \in \mathbb{R}$, $y_i > 0$. Substituting $s = u/y_i$, this condition is equivalent to $s/2 + 1 \leq \log_2(2^s+1)$ for all $s \in \mathbb{R}$. This inequality in turn is equivalent to the upper bound in (36) by observing that if $f, g : \mathbb{R} \to \mathbb{R}$ are two strictly increasing functions, $f(t) \leq g(t)$ for all t is equivalent to $f^{-1}(s) \geq g^{-1}(s)$ for all s. □

One can also show convergence under approximate computation of (11) and/or an additional proximal term as in (28).

We conclude this section by relating our neural penalty method to the prior literature. First, we note from (35) that the regularization kernel $\tilde{d}_i$ of our neural method is a rescaled φ-divergence, as its derivative $\tilde{d}_i'(x_i, y_i)$ has the form $y_i\varphi'(x_i/y_i)$ for $\varphi'(t) = \log_2(2^t - 1)$. However, we have used $\mu = 1$, and thus the convergence results of [3], which require $\mu > 1$, do not apply.

Finally, there is also a connection with the log-sigmoid method of multipliers proposed in [25]. The method of [25] uses a penalty of the general form $\beta \ln(e^{w/\beta} + 1)$, and thus a penalty derivative of the form $e^{w/\beta}/(e^{w/\beta} + 1)$; here, in rough terms, we may think of β as relating to the Lagrange multiplier, and w to the constraint violation. To make the penalty derivative approach infinity as $w \to \infty$ in [25], it is truncated and smoothly continued with a linear function; thus the penalty itself is smoothly continued with a quadratic function. The difference in our approach is that the neural penalty *derivative*, not the penalty, takes the form $\beta \ln(e^{w/\beta} + 1)$. Thus, the penalty of [25] is essentially the derivative of our neural penalty, and conversely our proposed neural penalty is essentially the integral of [25]'s. For this reason, our penalty derivative does not require truncation and replacement by a linear function for large w.

6. Computational tests

We conclude with some preliminary computational experiments with augmented Lagrangian methods for complementarity problems. Our original objective was to study the behavior of the neural penalty as compared to the log-quadratic penalty. However, we also include in the results two other penalties, the smooth cubic penalty of [16] and a variant of the classic exponential penalty. Our goal here is to compare double regularization methods to other proximal approaches, both coercive and non-coercive. We did not include the classic quadratic penalty, since it would cause the system of equations (11) to be nonsmooth.

Our tests of the cubic penalty in (11)–(12) revealed sizable discrepancies from [16], mainly because [16] instead tests a variant of the primal-dual algorithm (28)–(29). To better understand these differences, we tested not only an implementation of (11)–(12), but also of (28)–(29), using all four penalties.

We coded the algorithms in MATLAB and applied them to all the nonlinear complementarity problems in the MATLAB version of MCPLIB [11, 12] except the `pgvon` problems, which are especially badly behaved, with F ill-defined at the solution. We treated the remaining problems as being in the form (2). Considering differing starting points, the test set has 77 problems, most of which are not monotone. Even though our convergence analysis requires monotonicity, performance on the MCPLIB may still be considered a reasonable benchmark of practical performance, as in [16].

To improve the numerical behavior of these models, we introduce a positive diagonal scaling matrix S, with diagonal elements S_{ii}, along with a change of variables $w = S^{-1}x$, and cast the problem as

$$F(Sw) \geq 0 \quad w \geq 0 \quad \langle F(Sw), w\rangle = 0.$$

Under this scaling, the recursions (11)–(12) become

$$\begin{aligned} 0 &= F(Sw^{k+1}) - P'(-\alpha_k w^{k+1}, y^k) \\ y^{k+1} &= P'(-\alpha_k w^{k+1}, y^k), \end{aligned}$$

or, after changing back to the original variables $x = Sw$,

$$0 = F(x^{k+1}) - P'(-\alpha_k S^{-1} x^{k+1}, y^k) \tag{41}$$

$$y^{k+1} = P'(-\alpha_k S^{-1} x^{k+1}, y^k). \tag{42}$$

In the case of an additional proximal term, we similarly alter (28)–(29) to

$$0 = F(x^{k+1}) - P'(-\alpha^{\mathrm{d}}{}_k S^{-1} x^{k+1}, y^k) + \frac{1}{\alpha_k^{\mathrm{p}}} S(x^{k+1} - x^k) \tag{43}$$

$$y^{k+1} = P'(-\alpha_k^{\mathrm{d}} S^{-1} x^{k+1}, y^k). \tag{44}$$

Note also that we use different stepsizes α_k^{p} and α_k^{d} for the primal and dual portions of the regularization. Briefly, by altering our analysis to use a scaled distance kernel and quasi-Féjer convergence in a scaled norm, this modification is compatible with theoretical convergence if α_k^{p} and α_k^{d} remain in a fixed proportion, or $\alpha_k^{\mathrm{d}}/\alpha^{\mathrm{p}}{}_k$ only changes a finite number of times. We tested the following penalties:

1. The log-quadratic penalty:

$$P_i'(u_i, y_i) = \frac{u_i + (\mu - 1)y_i + \sqrt{(u_i + (\mu - 1)y_i)^2 + 4\mu y_i^2}}{2\mu}.$$

2. The neural penalty:

$$P_i'(u_i, y_i) = y_i \log_2(2^{u_i/y_i} + 1).$$

3. The cubic penalty of [16]:

$$P_i'(u_i, y_i) = \max\{\sqrt{y_i} + u_i, 0\}^2.$$

   This penalty is not coercive, but is specially constructed so that the system of equations (41) or (43) will nonetheless be first-order smooth if F is.
4. A modified exponential penalty:

$$P_i'(u_i, y_i) = \begin{cases} y_i e^{u_i/y_i}, & \text{if } u_i/y_i \le 1 \\ u_i e, & \text{if } u_i/y_i \ge 1. \end{cases}$$

We initially tested the classic exponential penalty $P_i'(u_i, y_i) = y_i e^{u_i/y_i}$ for all u_i (see for example [5, 33]), but this function grows so rapidly for large u_i that its value was often not representable in MATLAB. Therefore, we truncated it and continued it smoothly with a linear function. Note that, with or without this modification, theoretical convergence with this penalty requires more assumptions than F being monotone, and may not permit the additional proximal term of (43).

We employed a Newton algorithm with Armijo line search, specifically the `nsola` code of [20], to solve the systems of nonlinear equations (41) and (43). In order to deal with nearly singular Jacobians, we incorporated the modified Cholesky Factorization described in [18]. All nonlinear equations were solved essentially exactly, with residual no more than 10^{-8}. Other details of the implementation implementation are as follows:

- The initial multipliers were set to 1, since they must be strictly positive and this choice gave us good empirical results.
- For the dual method (41)–(42), we set the initial stepsize α_0 to 10. If, after successful solution of the nonlinear equations, the feasibility of the primal solution or its complementarity with the multipliers did not improve by a factor of 0.5, we multiplied the stepsize by 10. Otherwise, we multiplied it by 1.05 in order to speed up convergence. Such strategies are usual in multiplier methods, see for example [16].
- For the primal-dual method (43)–(44), we adjust the stepsizes α_k^{p}, α_k^{d} similarly to [16]. Initially, we set $\alpha_0^{\mathrm{p}} = \max\{10, |x^0|\}$ and $\alpha_0^{\mathrm{d}} = 10$. If the nonlinear equation solver fails, we divide α_k^{p} by 10 and set α_k^{d} to its initial value. Otherwise, if $\|x^{k+1} - x^k\| > 100\|y^{k+1} - y^k\|$, α_k^{d} is multiplied by 5. If $100\|x^{k+1} - x^k\| < \|y^{k+1} - y^k\|$, α_k^{d} is set to $\max\{\|y^k\|, 1.0\}$. If α_k^{p} and α_k^{d} are still unchanged, we multiply both α_k^{p} and α_k^{d} by 1.05 or 5 depending on whether the primal feasibility and complementarity slackness has improved by a factor of at least 0.5 or not. In our experiments, as in [16], the number of times α_k^{p} and α_k^{d} are updated idependently is very small, and hence our convergence theory still holds after the last independent update.
- As suggested in [10, 16], the scaling matrix S was determined by the initial solution x^0 via

$$S_{ii} \stackrel{\text{def}}{=} \frac{1}{\max\left(0.1\|\nabla F_{ii}(x^0)\|, 10\right)}.$$

Finally, we have chosen the total number of Newton steps as our benchmark, since our code is preliminary and MATLAB is an interpreted language, meaning that reporting run time may be misleading. We graphically present our test results using performance profiles [13]. Complete test results appear in Tables 1–4.

To make a fair comparison between the log-quadratic penalty and the other penalties, we must first study how adjusting μ affects the log-quadratic penalty performance. For the neural penalty, μ is fixed at 1. The other two penalties are not double regularizations and therefore do not have a μ parameter.

We tested the log-quadratic penalty with $\mu = 5$, 1.5, 1.05, and 1. Figure 3 displays the performance profile of this test, in terms of Newton iterations.[1] Clearly, performance tends to improve as μ decreases. Therefore, we should use the smallest possible μ when

Table 1. Number of Newton steps for pure dual method, part 1.

Problem	Log-quad $\mu = 5.0$	Log-quad $\mu = 1.5$	Log-quad $\mu = 1.05$	Log-quad $\mu = 1.0$	Cubic	Expon	Neural
bertsekas1	fail	40	78	41	78	82	48
bertsekas2	26	22	25	26	21	54	22
bertsekas3	26	36	41	37	408	216	221
bertsekas4		40	78	41	78	82	48
bertsekas5	27	24	24	24	21	39	24
bertsekas6	29	26	25	25	22	39	24
colvdual1	fail	fail	670	fail	fail	110	fail
colvdual2	32	fail	fail	fail	fail	64	20
colvdual3	16	16	17	17	11	22	15
colvdual4	fail	fail	fail	21	fail	125	fail
colvnlp1	24	21	22	23	24	88	22
colvnlp2	19	19	21	21	21	37	19
colvnlp3	16	16	17	17	11	22	15
colvnlp4	18	20	22	22	34	41	25
colvnlp5	18	20	22	22	33	38	25
colvnlp6	16	14	15	15	13	23	13
cycle1	4	4	4	4	3	3	3
explcp1	17	23	22	21	6	15	13
hanskoop10	24	25	23	23	fail	fail	34
hanskoop2	24	25	23	23	fail	fail	34
hanskoop4	24	25	23	23	fail	fail	34
hanskoop6	24	25	23	23	fail	fail	34
hanskoop8	24	25	23	23	fail	fail	34
josephy1	16	18	16	16	fail	65	15
josephy2	19	fail	fail	fail	fail	fail	fail
josephy3	fail	fail	fail	fail	fail	fail	fail
josephy4	fail	18	17	17	fail	60	16
josephy5	15	16	14	14	11	58	14
josephy6	19	22	18	20	19	fail	63
josephy7	fail	fail	fail	fail	fail	fail	fail
josephy8	13	15	13	13	11	58	13
kojshin1	450	450	450	450	fail	300	300
kojshin2	450	fail	fail	fail	fail	fail	fail
kojshin3	450	fail	fail	fail	fail	fail	fail
kojshin4	450	18	14	14	13	63	14
kojshin5	450	21	16	16	fail	109	17
kojshin6	554	fail	fail	244	fail	fail	fail
kojshin7	450	450	450	450	fail	fail	300
kojshin8	450	450	450	450	fail	250	300

comparing the penalties. In our subsequent testing, we used $\mu = 1.05$, since we have only proved convergence of the log-quadratic method when $\mu > 1$. The performance of this case is very close to the limiting case $\mu = 1$.

Table 2. Number of Newton steps for pure dual method, part 2.

Problem	Log-quad $\mu = 5.0$	Log-quad $\mu = 1.5$	Log-quad $\mu = 1.05$	Log-quad $\mu = 1.0$	Cubic	Expon	Neural
mathinum1	10	9	9	9	6	5	6
mathinum2	9	9	9	9	5	5	5
mathinum3	fail	13	13	13	9	9	9
mathinum4	10	10	10	10	6	6	6
mathinum5	16	16	16	17	12	12	13
mathinum6	11	11	11	11	7	7	7
mathisum1	17	15	14	14	10	57	13
mathisum2	16	13	13	13	11	26	13
mathisum3	450	450	450	450	fail	250	300
mathisum4	19	16	14	14	11	57	13
mathisum5	1	1	1	1	1	1	1
mathisum6	30	22	19	19	14	63	19
mathisum7	400	350	350	350	1082	200	250
nash1	9	9	9	9	6	6	6
nash2	8	8	8	8	6	6	6
nash3	7	7	7	7	6	6	5
nash4	6	6	6	6	3	3	3
powell1	103	103	103	103	17	51	52
powell2	51	105	51	51	16	51	100
powell3	51	151	103	51	20	21	29
powell4	143	51	151	51	19	51	50
powell5	fail	fail	758	fail	fail	fail	fail
powell6	103	103	103	103	14	50	52
scarfanum1	21	23	22	22	20	23	19
scarfanum2	23	26	26	26	23	27	24
scarfanum3	23	fail	27	fail	27	33	26
scarfanum4	21	23	23	22	14	20	18
scarfbnum1	fail	36	51	52	88	61	89
scarfbnum2	206	fail	209	141	62	fail	fail
scarfbsum1	37	34	41	41	992	fail	fail
scarfbsum2	fail	fail	fail	fail	fail	fail	fail
sppe1	14	16	17	17	18	28	17
sppe2	14	16	15	15	17	21	17
sppe3	10	11	10	10	11	18	10
tobin1	23	24	23	23	33	112	45
tobin2	28	36	34	41	95	962	50
tobin3	50	43	53	45	179	fail	55
tobin4	17	16	14	14	11	39	13

Next, Figure 4 presents a performance profile comparing all four penalties for the dual method (41)–(42). It also shows three two-method profiles, each comparing one of the other penalties to the neural penalty. In these comparisons, the neural penalty acquits itself well, comparing favorably to the other methods in most respects; however,

Table 3. Number of Newton steps for primal-dual method, part 1.

Problem	Cubic	Exponential	Log-quad ($\mu = 1.05$)	Neural
bertsekas1	54	78	79	45
bertsekas2	37	69	29	32
bertsekas3	140	114	36	142
bertsekas4	54	78	79	45
bertsekas5	22	33	27	26
bertsekas6	33	54	30	30
colvdual1	898	168	467	504
colvdual2	621	92	453	28
colvdual3	13	31	22	20
colvdual4	621	124	444	1706
colvnlp1	35	121	23	23
colvnlp2	24	78	23	26
colvnlp3	13	31	21	20
colvnlp4	27	35	25	24
colvnlp5	27	35	25	24
colvnlp6	16	39	20	18
cycle1	7	7	7	7
explcp1	15	24	12	23
hanskoop10	41	22	23	31
hanskoop2	41	22	23	31
hanskoop4	41	22	23	31
hanskoop6	77	22	23	31
hanskoop8	41	22	23	31
josephy1	227	58	16	17
josephy2	228	1593	227	224
josephy3	450	801	442	449
josephy4	227	59	17	17
josephy5	12	58	15	15
josephy6	35	299	226	46
josephy7	226	647	226	228
josephy8	12	58	14	13
kojshin1	235	477	634	228
kojshin2	249	833	344	392
kojshin3	358	762	470	547
kojshin4	13	1010	17	16
kojshin5	fail	963	20	19
kojshin6	276	620	335	483
kojshin7	220	492	634	215
kojshin8	fail	442	634	227

the log-quadratic method is slightly more reliable on this test set. The cubic method fails most often, but tends to run quickly when it does not fail. The modified exponential method is clearly the least desirable, tending to be very slow, but only slightly more reliable than the cubic.

Table 4. Number of Newton steps for primal-dual method, part 2.

Problem	Cubic	Exponential	Log-quad ($\mu = 1.05$)	Neural
mathinum1	11	13	13	13
mathinum2	10	10	11	10
mathinum3	15	15	13	15
mathinum4	12	12	12	12
mathinum5	19	20	17	18
mathinum6	13	13	13	13
mathisum1	11	57	15	14
mathisum2	12	22	13	13
mathisum3	fail	576	421	511
mathisum4	12	57	15	14
mathisum5	1	1	1	1
mathisum6	16	61	20	20
mathisum7	97	495	297	378
nash1	10	10	10	10
nash2	10	10	9	10
nash3	9	9	9	9
nash4	5	5	6	5
powell1	14	15	322	21
powell2	58	19	322	19
powell3	20	19	259	20
powell4	60	19	18	20
powell5	fail	fail	fail	fail
powell6	15	14	323	14
scarfanum1	22	27	29	25
scarfanum2	28	36	32	28
scarfanum3	33	48	31	35
scarfanum4	18	25	28	24
scarfbnum1	103	154	fail	395
scarfbnum2	89	151	fail	140
scarfbsum1	100	146	56	65
scarfbsum2	fail	fail	196	fail
sppe1	27	50	23	23
sppe2	19	35	17	18
sppe3	12	57	11	11
tobin1	40	141	24	37
tobin2	64	183	37	42
tobin3	103	179	53	56
tobin4	11	39	14	14

Figure 5 shows the effect of adding a "prox" term to the both the neural and exponential penalties, that is, executing (43)–(44) instead of (41)–(42). Although the proximal term slows down the method somewhat, it appears to have a significant stabilizing effect, resulting in much greater robustness; note in particular the much greater reliability it imparts to the exponential method.

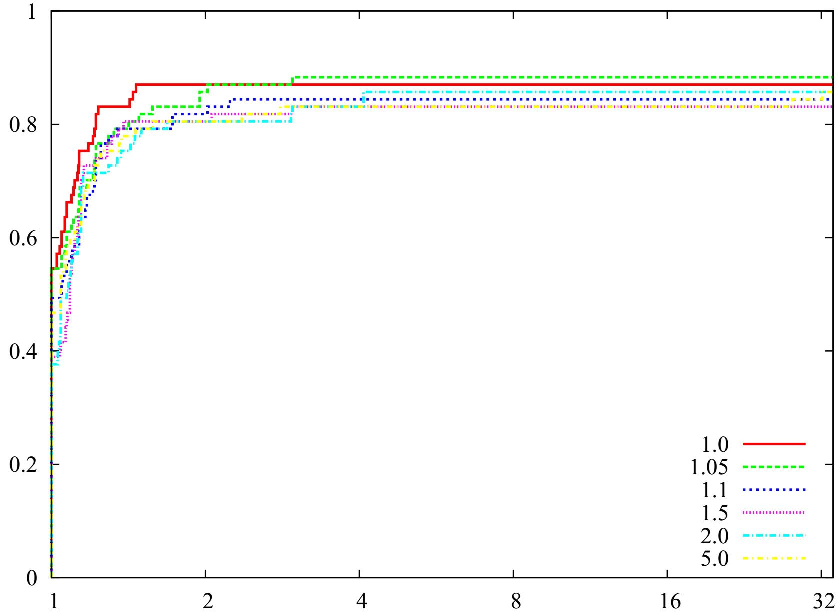

Figure 3. The impact of μ on the performance of the log-quadratic penalty.

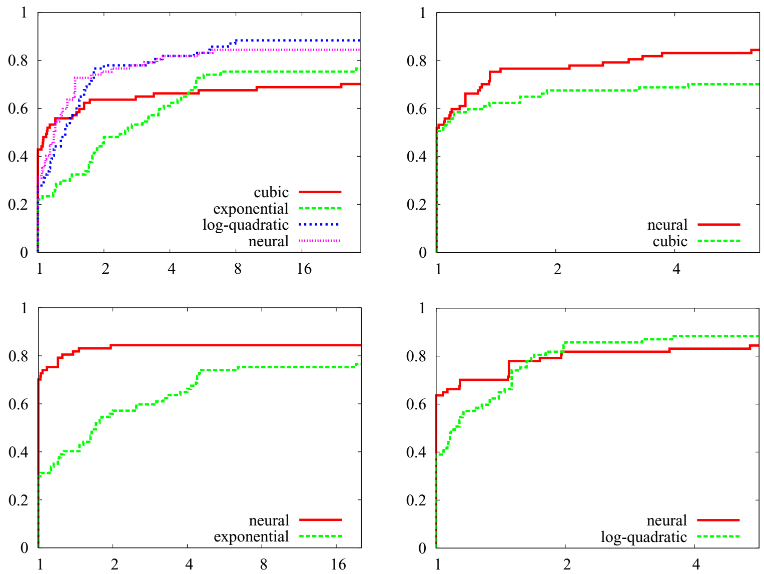

Figure 4. Pure dual augmented Lagrangian methods.

Figure 6 compares all four penalties in the primal-dual setting (43)–(44). The robustness effect of the additional prox term applies to all the methods, greatly increasing their reliability. The exponential methods remains the slowest, and while the other methods are quite comparable, the neural penalty still seems to be the preferred approach.

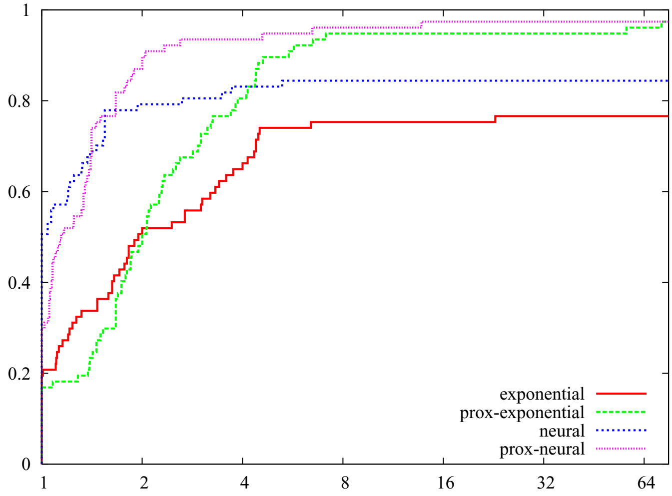

Figure 5. Comparision between primal-dual and pure dual generalized augmented Lagrangians.

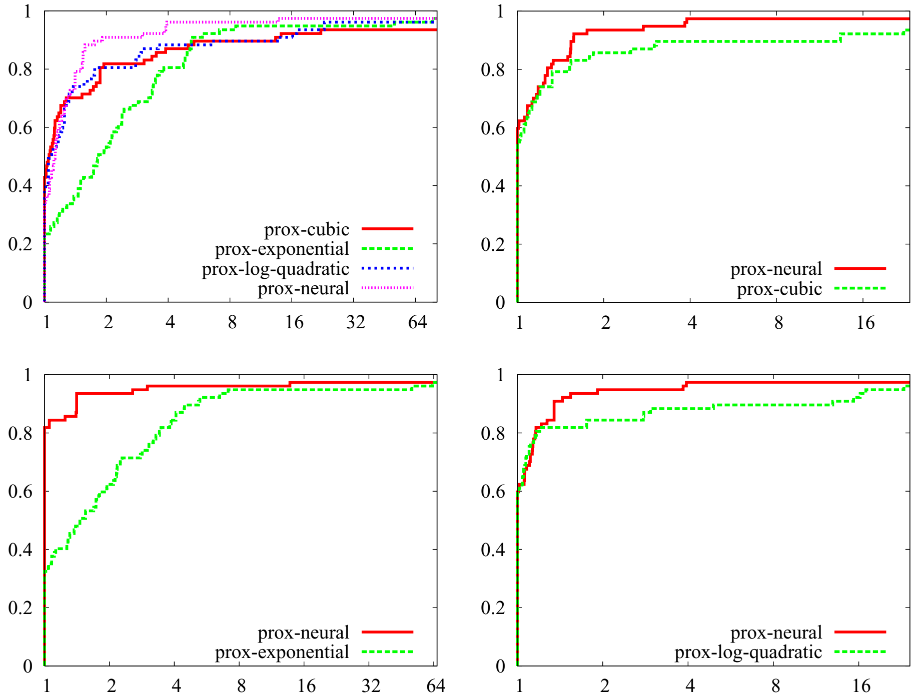

Figure 6. Primal-dual augumented Lagrangian methods.

Note

1. Following [13], let $s(p, m)$ denote the number of Newton steps required by method m on problem instance p, and let $s^*(p) = \min_m\{s(p, m)\}$ be the smallest number of steps on instance p required by any method

in the profile. Define $r(p, m) = s(p, m)/s^*(p)$. The plots display the fraction of problems p for which $r(p, m) \leq r$, r being displayed on the horizontal axis.

References

1. H. Attouch and M. Théra, "A general duality principle for the sum of two operators," Journal of Convex Analysis, vol. 3, pp. 1–24, 1996.
2. A. Auslender, M. Teboulle, and S. Ben-Tiba, "A logarithmic-quadratic proximal method for variational inequalities," Computational Optimization and Applications, vol. 12, pp. 31–40, 1999.
3. A. Auslender, M. Teboulle, and S. Ben-Tiba, "Interior proximal and multiplier methods based on second order homogeneous kernels," Mathematics of Operations Research, vol. 24, pp. 645–668, 1999.
4. A. Auslender and M. Teboulle, "Lagrangian duality and related multiplier methods for variational inequality problems," SIAM Journal on Optimization, vol. 10, pp. 1097–1115, 2000.
5. D.P. Bertsekas, Nonlinear Programming. Athena Scientific, 1995.
6. R. Burachick, M.G. Drummond, A.N. Iusem, and B.F. Svaiter, "Full convergence of the steepest descent method with inexact line searches," Optimization, vol. 10, pp. 1196–1211, 1995.
7. J.V. Burke and M.C. Ferris, "Weak sharp minima in mathematical programming," SIAM Journal on Control and Optimization, vol. 31, pp. 1340–1356, 1993.
8. Y. Censor and S.A. Zenios, "Proximal minimization algorithm with D-functions," Journal of Optimization Theory and Applications, vol. 73, pp. 451–464, 1992.
9. Y. Censor, A.N. Iusem, and S.A. Zenios, "An interior point method with Bregman functions for the variational inequality problem with paramonotone operators," Mathematical Programming, vol. 81, pp. 373–400, 1998.
10. C. Chen and O.L. Mangasarian, "A class of smoothing functions for nonlinear and mixed complementarity problems," Computational Optimization and Applications, vol. 5, pp. 97–138, 1996.
11. S.P. Dirkse and M.C. Ferris, "MCPLIB: A collection on nonlinear mixed complementarity problems," Optimization Methods and Software, vol. 2, pp. 319–345, 1995.
12. S.P. Dirkse and M.C. Ferris, "Modeling and solution environments for MPEC: GAMS & MATLAB," in Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods (Lausanne, 1997), Applied Optimization, vol. 22. Kluwer Academic Publisher, Dordrecht, 1999, pp. 127–147.
13. E.D. Dolan and J.J. Moré, Benchmarking optimization software with performance profiles, Mathematical Programming, vol. 91, no. 2, pp. 201–213, 2002.
14. J. Eckstein, "Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming," Mathematics of Operations Research, vol. 18, pp. 202–226, 1993.
15. J. Eckstein, "Approximate iterations in Bregman-function-based proximal algorithms," Mathematical programming, vol. 83, pp. 113–123, 1998.
16. J. Eckstein and M.C. Ferris, "Smooth methods of multipliers for complementarity problems," Mathematical Programming, vol. 86, pp. 65–90, 1999.
17. M.C. Ferris, "Finite termination of the proximal point algorithm," Mathematical Programming, vol. 50, pp. 359–366, 1991.
18. P.E. Gill, W. Murray, and M.H. Wright, Practical Optimization. Academic Press, 1981.
19. A.N. Iusem, "On some properties of paramonotone operators," Journal of Convex Analysis, vol. 5, pp. 269–278, 1998.
20. C.T. Kelley, Iterative Methods for Linear and Nonlinear Equations. SIAM: Philadelphia, 1995.
21. K.C. Kiwiel, "On the twice differentiable cubic augmented Lagrangian," Journal of Optimization Theory and Applications, vol. 88, pp. 233–236, 1996.
22. L. Lewin (Ed.) Structural Properties of Polylogarithms. Mathematical Surveys and Monographs, vol. 37, American Mathematical Society, Providence, 1991.
23. U. Mosco, "Dual variational inequalities," Journal of Mathematical Analysis and Applications, vol. 40, pp. 202–206, 1972.
24. D.J. Newman, A Problem Seminar. Springer-Verlag: New York, 1982.
25. R.A. Polyak, "Nonlinear rescaling vs. smoothing techinique in convex optimization," Mathematical Programming, vol. 92, pp. 197–235, 2002.

26. S.M. Robinson, “Composition duality and maximal monotonicity,” Mathematical Programming, vol. 85, pp. 1–13, 1999.
27. R.T. Rockafellar, Convex Analysis. Princeton University Press: Princeton, 1970.
28. R.T. Rockafellar, “Monotone operators and the proximal point algorithm,” SIAM Journal on Control and Optimization, vol. 14, pp. 877–898, 1976.
29. R.T. Rockafellar, “Augmented Lagrangians and applications of the proximal point algorithm in convex programming,” Mathematics of Operations Research, vol. 1, pp. 97–116, 1976.
30. R.T. Rockafellar and R.J.-B. Wets, Variational Analysis. Springer-Verlag: Berlin, 1998.
31. P.J.S. Silva, J. Eckstein, and C. Humes Jr., “Rescaling and stepsize selection in proximal methods using separable generalized distances,” SIAM Journal on Optimization, vol. 12, pp. 238–261, 2001.
32. M. Teboulle, “Entropic proximal mappings with applications to nonlinear programming,” Mathematics of Operations Research, vol. 17, pp. 670–690, 1992.
33. P. Tseng and D. Bertsekas, “On the convergence of the exponential multiplier method for convex programming,” Mathematical Programming, vol. 60, pp. 1–19, 1993.
34. A.E. Xavier, Hyperbolic Penalty: A New Method for Solving Optimization Problems,” M.Sc. Thesis, COPPE, Federal University of Rio de Janeiro, Brazil (in Portuguese), 1982.