RUTCOR
RESEARCH
REPORT

Double-Regularization Proximal Methods, with Complementarity Applications

Paulo J. S. Silva[a] Jonathan Eckstein[b]

RRR 29-2003, August 2003

RUTCOR
Rutgers Center for
Operations Research
Rutgers University
640 Bartholomew Road
Piscataway, New Jersey
08854-8003
Telephone: 732-445-3804
Telefax: 732-445-5472
Email: rrr@rutcor.rutgers.edu
http://rutcor.rutgers.edu/~rrr

[a]Department of Computer Science, Instituto de Matemática e Estatística – University of São Paulo, Brazil (rsilva@ime.usp.br).

[b]Business School and RUTCOR, 640 Bartholomew Road, Rutgers University, Piscataway, NJ 08854 USA (jeckstei@rutcor.rutgers.edu).

Rutcor Research Report
RRR 29-2003, August 2003

Double-Regularization Proximal Methods, with Complementarity Applications

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.* We apply this class of regularizations to complementarity problems using a dual formulation, leading to the broadened class of generalized 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 augmented Lagrangian based on the "neural network" smoothing function. Finally, we computationally compare this augmented Lagrangian to the already-known logarithmic-quadratic variant on the MCPLIB problem library, and show that the neural approach offers some advantages.

Acknowledgements: Bert Schaaf, via the RGMIA internet group, suggested the outline of the proof in Appendix A, more elegant and far shorter than our original proof.

1 Introduction

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

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

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.

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*

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

corresponding to some 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.

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, 12, 17, 19]: 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 in the sense of point-to-set maps, and thus guaranteed to exist. We consider the problems (4) and (6) to be duals of one another, since they have identical "optimality" conditions (5).

Applying this duality transformation to (3) with $U = F$ and $V = N_{\mathbb{R}^n_+}$ yields

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

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 $i = 1, \ldots, n$.

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

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

where ∇_1 denotes the gradient with respect to the the first argument, and $\alpha_k > 0$ is some scalar bounded away from 0. $\widetilde{D}$ 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* [21], takes $\widetilde{D}(x, y) = (1/2)\|x - y\|^2$. In general, one may have to satisfy (8) only approximately, but for simplicity we omit this complication here.

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

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

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

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

Again applying the duality transformation, but with

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

produces an equivalent subproblem

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

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

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

is single-valued, so we obtain the equivalent recursions

$$0 = F(x^{k+1}) - P'_k(-\alpha_k x^{k+1}) \tag{11}$$
$$y^{k+1} = P'_k(-\alpha_k x^{k+1}), \tag{12}$$

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'_k plays the same role as the gradient of the penalty term in augmented Lagrangian methods for optimization problems; see for example [22, 23, 11]. 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 generalsetting.

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

$$\widetilde{D}(x,y) = \sum_{i=1}^{n} \widetilde{d}(x_i, y_i) \tag{13}$$
$$\widetilde{d}(x_i, y_i) = +\infty \qquad \text{if } x_i \notin [a_i, b_i] \tag{14}$$
$$\lim_{x \downarrow a_i} \nabla_1 \widetilde{d}(x_i, y_i) = -\infty \qquad \text{if } a_i > -\infty \tag{15}$$
$$\lim_{x \uparrow b_i} \nabla_1 \widetilde{d}(x_i, y_i) = +\infty \qquad \text{if } b_i < +\infty. \tag{16}$$

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

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

which should be more convenient computationally. For example, the primal complementarity recursion (9) now reduces to

$$0 = F(x^{k+1}) + \frac{1}{\alpha_k} \nabla_1 \widetilde{D}(x^{k+1}, x^k). \tag{17}$$

This recursion is an equation, rather than an inclusion, and inherits whatever smoothness is present in F and $\nabla_1 \widetilde{D}(\cdot, x^k)$. This situation is generally preferable to the non-coercive case, where the resulting subproblem may be no easier than the original problem (3). However, the definition domain of equation (17) is constrained 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)-(12). Then, the definition (10) reduces to the much simpler

$$P'_k = \left(\nabla_1 \widetilde{D}(\cdot, y^k) \right)^{-1}. \tag{18}$$

By judicious choice of $\widetilde{D}$, one can make the single-valued function P'_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 $\widetilde{D}$ lead to nonsmooth augmented Lagrangians. Certain non-coercive choices of $\widetilde{D}$ can also lead to some limited smoothness in the augmented Lagrangian [15, 12], but result in a penalty that is very "flat" in the vicinity of the origin, which may not be ideal.

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 [5, 11, 23], 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 $\widetilde{D}(\cdot, x^k)$ was *not* coercive. The case of a general monotone T and coercive $\widetilde{D}$ remained open.

Subsequent research [6] proved convergence of certain coercive proximal algorithms when T is *paramonotone* [13], 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 $\widetilde{D}$, the logarithmic-quadratic kernel. This $\widetilde{D}$ 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 to this field of research: first, capitalizing on our related work in [18], we prove convergence for a general maximal monotone operator T of proximal methods employing a broader class of coercive distance regularization measures $\widetilde{D}$ than in [3]. These distance measures do not have take the φ-divergence form: for example, they may instead be certain kinds of rescaled Bregman distances; see Section 5.2. Second, we note a relationship between the logarithmic-quadratic penalty arising in [3] and some prior work on pure penalty (or smoothing) methods for complementarity problems by Chen and Mangasarian [7]. We study another penalty from [7], the *neural network smooth plus function*, and show that it too corresponds to a proximal algorithm and augmented Lagrangian method. The neural penalty yielded superior computational results to the logarithmic-quadratic penalty in the pure penalty environment of [7], 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 this hypothesis on a difficult, realistic set of test problems, the MCPLIB [9]. We also believe our experiments constitute the first rigorous computational testing of any coercive proximal algorithms on non-optimization complementarity problems. The experiments show that our new neural multiplier method does indeed tend to be faster than the logarithmic-quadratic multiplier method, but is less robust. A hybrid approach that executes a single initial step of the logarithmic-quadratic method, and then reverts to the neural method, cures the robustness problem and seems almost uniformly faster than the logarithmic-quadratic algorithm. Note that the MCPLIB

problems are not monotone, as required by our convergence analysis, but following the example of [12] we still use them as a computational testing library.

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

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

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 $\widetilde{D}$ must meet a set of conditions (Assumption 2.1 below) slightly reformulated from our earlier work in [18]. 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 logarithmic-quadratic measure of [2].

Section 3 then presents the heart of our convergence analysis. In particular, we show that the sequence of iterates $\{x^k\}$ produced by our proximal algorithm is Féjer monotone to the solution set of (1). If $\mu > 1$, we shortly thereafter obtain convergence of the method.

Section 4 then treats convergence in the more difficult case $\mu = 1$. Taking $\mu = 1$ tends to lead to simpler penalties than $\mu > 1$, and it is also useful in analyzing the neural network penalty function. When $\mu = 1$, we restrict ourselves to $B = \mathbb{R}^n_+$, and strengthen somewhat the assumptions on D from Section 2.

In Section 5, we give examples of distance measures meeting our assumptions. These include the Φ_2 class proposed in [3], but also other possibilities. Next, Section 6 considers how our class of distance measures manifests itself in the dual setting (11)-(12), that is, what kind of penalty terms P' correspond to $\widetilde{D}$ regularizations meeting our assumptions. We show that any penalty term having certain regularity properties and fitting inside a certain envelope corresponds to one of our allowed distance measures $\widetilde{D}$. The *upper* bound of this envelope is the logarithmic-quadratic penalty of [3]. The remainder of Section 6 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 7 presents the computational testing, and discusses possible further tuning of the algorithm.

2 Coercive separable distances and double regularizations

We begin by stating a key assumption adapted from [18]:

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

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

2.1.2. $\widetilde{d}_i$ is differentiable with respect to the first variable 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

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

2.1.3. For all $y_i \in (a_i, b_i)$, $\widetilde{d}_i(\cdot, y_i)$ is essentially smooth [20, 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 $\left|\widetilde{d}_i'(x_i, y_i)\right| \leq L\,|x_i - y_i|$.

This assumption is a simple transformation of [18, Assumption 2.1], where each $\widetilde{d}_i(\cdot, y_i)$ is divided by $\widetilde{d}_i''(y_i, y_i)$. We note that all the convergence results from [18] remain true under Assumption 2.1.

The stipulation that $\operatorname{int}\operatorname{dom} \widetilde{d}_i(\cdot, y_i) = (a_i, b_i)$ 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, the term $\widetilde{d}_i(\cdot, y_i)$ acts as a "barrier" keeping the iterates within the interval (a_i, b_i).

We will use functions $\widetilde{d}_i$ 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 $i = 1, \ldots, n$, let $d_i : \mathbb{R} \times (a_i, b_i) \to (-\infty, \infty]$ be continuously differentiable with respect to the first variable. Let*

$$\widetilde{d}_i(x_i, y_i) \stackrel{\text{def}}{=} d_i(x_i, y_i) + \frac{\mu}{2}(x_i - y_i)^2, \tag{19}$$

and

$$\widetilde{D}(x, y) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \widetilde{d}_i(x_i, y_i) = \sum_{i=1}^{n} d_i(x_i, y_i) + \frac{\mu}{2}(x_i - y_i)^2. \tag{20}$$

If each $\widetilde{d}_i$ conforms to Assumptions 2.1.1-2.1.3 we shall call $\widetilde{D}$ the double regularization based on $D(x, y) \stackrel{\text{def}}{=} \sum_{i=1}^{n} d_i(x_i, y_i)$. *Moreover, each $\widetilde{d}_i$ 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 [18], and show that it implies Assumption 2.1.4:

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

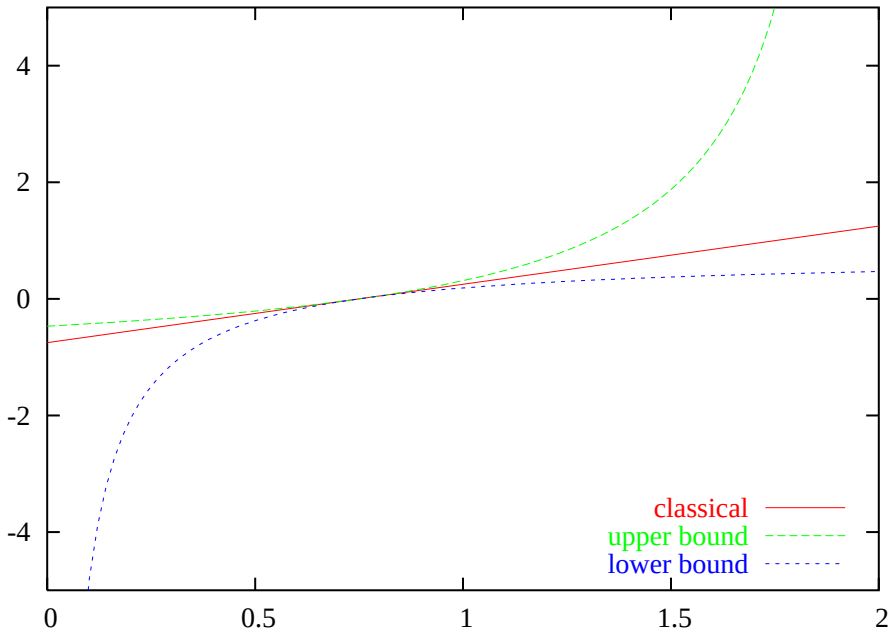

Figure 1: Bounds for the derivative of d_i, for $(a_i, b_i) = (0, 2)$ and $y_i = 0.75$. Notice that the derivative of the classical regularization $(1/2)(x_i - y_i)^2$ lies in between the limits.

1. *If a_i and b_i are both finite,*

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

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

   (a) *If only a_i is finite:*

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

   (b) *If only b_i is finite:*

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

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

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

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

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

If $-\infty < a_i \leq y_i \leq x_i < a_i + \epsilon$ 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, $|\widetilde{d}_i'(x_i, y_i)| \leq (2 + \mu)\,|x_i - y_i|$, and the double regularization component $\widetilde{d}_i$ 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. □

We now introduce our proximal method:

Proximal Method using Double Regularization (PMDR):

Let $\widetilde{D}$ be a double regularization based via (20) on a coercive term D conforming to Assumption 2.3, with $\mu \geq 1$.

1. **Initialization:** *Let $k = 0$. Choose a scalar $c > 0$, and an initial iterate $x^0 \in \operatorname{int} B$.*

2. **Iteration:**

   (a) *Choose $\alpha_k \in [c, +\infty)$.*

   (b) *Find x^{k+1} such that*[1]

$$0 \in \alpha_k T(x^{k+1}) + \nabla_1 \widetilde{D}(x^{k+1}, x^k). \tag{21}$$

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

□

[1] We assume here that we can solve each of these proximal steps exactly. Inexact versions of the calculations are possible, but for clarity in the proofs, we have omitted such variations.

3 Principle convergence analysis

To prove convergence, we will establish that the PMDR sequence is Fejér monotone to the solution set of (1). To this end, we require four technical lemmas.

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

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

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

Proof. Multiplying the inequality $\alpha \leq \beta$ by the nonnegative value $\delta - \gamma$,

$$\begin{aligned} & \alpha(\delta - \gamma) \leq \beta(\delta - \gamma) \\ \Rightarrow \quad & -\alpha\gamma - \beta\delta \leq -\beta\gamma - \alpha\delta \\ \Rightarrow \quad & \alpha\beta - \alpha\gamma - \beta\delta + \gamma\delta \leq \alpha\beta - \beta\gamma - \alpha\delta + \gamma\delta \\ \Rightarrow \quad & (\delta - \alpha)(\gamma - \beta) \leq (\delta - \beta)(\gamma - \alpha). \end{aligned}$$

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

Lemma 3.2 *Under Assumption 2.3,*

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

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 $x_i, y_i \in (a_i, b_i)$,*

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

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:

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

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

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

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

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

4. $y_i < x_i < z_i$:

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

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

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 $\widetilde{D}$ be a double regularization, A and C be subsets of* $\operatorname{int} B$*, 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 $y \in C$,

$$\langle z - x, \nabla_1 \widetilde{D}(x, y)\rangle \leq \sum_{i=1}^{n} \left(\tfrac{\mu+\zeta_i(z_i,A,C)}{2}\big((z_i - y_i)^2 - (z_i - x_i)^2\big) - \tfrac{\mu-\zeta_i(z_i,A,C)}{2}(x_i - y_i)^2 \right).$$

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

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

Using the identities

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

it follows that

$$(z_i - x_i)\nabla_1 \widetilde{D}(x, y)_i \leq \tfrac{\mu+\zeta_i(z_i,A,C)}{2}\big((z_i - y_i)^2 - (z_i - x_i)^2\big) - \tfrac{\mu-\zeta_i(z_i,A,C)}{2}(x_i - y_i)^2.$$

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

We can now establish Fejér monotonicity:

Lemma 3.5 *Let $\{x^k\}$ be a sequence computed by the PMDR. Then, $\{x^k\}$ is Fejér monotone with respect 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 (21),

$$-\left(\frac{1}{\alpha_k}\right)\nabla_1 \widetilde{D}(x^{k+1}, x^k) \in (T + N_B)(x^{k+1}).$$

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

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

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,

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

Recalling that $\mu \geq 1$, Fejér monotonicity is now proven.

Finally, if $\mu > 1$, we have

$$\left\|x^{k+1} - x^k\right\|^2 \leq \frac{\mu+1}{\mu-1}\left(\left\|z - x^k\right\|^2 - \left\|z - x^{k+1}\right\|^2\right). \tag{22}$$

Fejér monotonicity guarantees that $\|z - x^k\|$ converges for all $z \in (T + N_B)^{-1}(0)$. Thus, if the solution set is non-empty, (22) implies $x^{k+1} - x^k \to 0$. □

With the above results in hand, it is possible to apply the analysis of [18] to prove the convergence of the PMDR algorithm:

Theorem 3.6 *In the PMDR algorithm, suppose $\widetilde{D}$ 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 to the variational inequality* (1)*, if its solution set is non-empty.*

Proof. We will apply [18, 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 [18, Assumption 2.1] after applying a scaling factor. The regularity condition of Assumption 1.1 is exactly [18, Assumption 2.2]. And finally, the exact proximal step (21) and Lemma 3.5 imply [18, Assumption 2.3]. Thus, the assumptions of [18, Theorem 2.7] hold, and it guarantees convergence of $\{x^k\}$. □

4 Analysis of the case $\mu = 1$

Theorem 3.6 omits the case $\mu = 1$. However, in Sections 6 and 7, we will encounter precisely this case. By strengthening Assumption 2.3, we now obtain 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 [18, Assumption 2.3] does not hold, and the proof of Theorem 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 Theorem 3.6 will once again apply.

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

Assumption 4.1 *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 $\widetilde{D}$. We assume that $d_i'(\cdot, \cdot)$ is continuous and:*

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

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

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

4.1.2. Given any $\bar{y}_i > 0$, exist $\zeta_i(\bar{y}_i) \in (0,1)$, 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 $y_i \in C_i(\bar{y}_i) \cap \mathbb{R}_{++}$,

$$\zeta_i(\bar{y}_i)\frac{(x_i - y_i)y_i}{x_i} \leq d_i'(x_i, y_i). \tag{23}$$

Assumption 4.1.1 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 (23) automatically holds when $x_i \geq y_i$ and both sides are nonnegative, but imposes a stronger bound when $x_i < y_i$ and both sides are negative.

Lemma 4.2 *Let $d_i : \mathbb{R} \times \mathbb{R}_{++} \to (-\infty, \infty]$ be function conforming to Assumption 4.1.1. Suppose $z_i \geq 0$, 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 4.1 implies Assumption 2.3 for the case $a = 0$, $b = +\infty$. Therefore, Lemma 3.3 gives

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

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 imitate 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 following 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 4.1.1.

2. $x_i > \max(y_i, z_i)$:

   This case is proved exactly like the respective case in the proof of Lemma 3.3, but using the strict inequality from Assumption 4.1.1 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 4.1.1 in the last step.

4. $y_i < x_i < z_i$: As in the first case, we use the reasoning of the respective 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 4.3 *Let $\{x^k\}$ be a PMDR sequence where $\mu = 1$ and the double regularization is based on coercive terms d_i conforming Assumption 4.1. Moreover, assume that the solution set of the variational inequality (1) is non-empty. Then, $x^{k+1} - x^k \to 0$.*

Proof. Let z be a solution of the variational inequality (1). The Fejér monotonicity of the PMDR sequences, asserted by Lemma 3.5, implies that $\{x^k\}$ is bounded.

Thus, 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 $\widetilde{x}$ be the respective limit points of $\{x^k\}$ and $\{x^{k+1}\}$. Since $x^{k+1} - x^k \to \widetilde{x} - \bar{x}$, we need only demonstrate that $\widetilde{x} = \bar{x}$. Define two index sets

$$\begin{aligned}I(\bar{x}, \widetilde{x}) &\stackrel{\text{def}}{=} \{i \mid \bar{x}_i = \widetilde{x}_i\} \\ J(\bar{x}, \widetilde{x}) &\stackrel{\text{def}}{=} \{i \mid \bar{x}_i \neq \widetilde{x}_i\}.\end{aligned}$$

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

$$(z_i - x_i)d_i'(x_i, y_i) \leq \eta_i(z_i - y_i)(x_i - y_i). \tag{24}$$

To establish the claim, consider three possibilities:

1. $\bar{x}_i, \widetilde{x}_i > 0$. Lemma 4.2 gives

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

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

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

   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 $C_i \ni \widetilde{x}_i$ where the inequality (24) holds with $\eta_i = 1 - \epsilon$.

2. $0 = \bar{x}_i < \widetilde{x}_i$. This situation can be analyzed by subcases considering the relative position of z_i:

   - $0 = z_i = \bar{x}_i < \widetilde{x}_i$. In this case, (24) is a direct consequence of Assumption 4.1.2.
   - $0 = \bar{x}_i < z_i < \widetilde{x}_i$. Here, one may simply use the signs of the terms appearing in (24). For any x_i sufficiently close to $\bar{x}_i = 0$ and y_i sufficiently close to $\widetilde{x}_i$, one has

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

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

     $$\begin{aligned} & d_i'(x_i, y_i) \leq x_i - y_i && \text{[by Assumption 4.1.1]} \\ \Rightarrow \quad & (z_i - x_i)d_i'(x_i, y_i) \leq (z_i - x_i)(x_i - y_i) && \text{[since } z_i - x_i > 0\text{]} \\ & \leq (1/2)(z_i - y_i)(x_i - y_i), \end{aligned}$$

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

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

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

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

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

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

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

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

     By continuity, there exist neighborhoods $A_i \ni \bar{x}_i$ and $C_i \ni \widetilde{x}_i$ 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 $0 < d_i'(x_i, y_i) \leq x_i - y_i$,

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

Therefore, we conclude that (24) holds. Let A_i, C_i be the corresponding neighborhoods for all $i \in J(\bar{x}, \widetilde{x})$. For $i \in I(\bar{x}, \widetilde{x})$, define $A_i = C_i = \mathbb{R}_{++}$. Then define Cartesian product neighborhoods

$$A \stackrel{\text{def}}{=} \prod_{i=1}^{n} A_i \qquad\qquad C \stackrel{\text{def}}{=} \prod_{i=1}^{n} C_i$$

of $\bar{x}$ and $\widetilde{x}$, respectively, along with

$$\bar{\eta} \stackrel{\text{def}}{=} \max_{i \in J(\bar{x}, \widetilde{x})} \{\eta_i\} \in (0, 1).$$

Finally, let $\zeta \in \mathbb{R}^n$ be given by

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

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

$$\langle z - x^{k+1}, \nabla_1 \widetilde{D}(x^{k+1}, x^k) \rangle \leq \sum_{i=1}^{n} \tfrac{1+\zeta_i}{2}\big((z_i - x_i^k)^2 - (z_i - x_i^{k+1})^2\big) - \tfrac{1-\zeta_i}{2}(x_i^{k+1} - x_i^k)^2.$$

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

$$0 \leq \sum_{i=1}^{n} \frac{1+\zeta_i}{2}\big((z_i - x_i^k)^2 - (z_i - x_i^{k+1})^2\big) - \frac{1-\zeta_i}{2}(x_i^{k+1} - x_i^k)^2.$$

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

$$0 \leq \sum_{i \in J(\bar{x}, \widetilde{x})} \left(\frac{1+\bar{\eta}}{2}\big((z_i - \bar{x}_i)^2 - (z_i - \widetilde{x})^2\big) - \frac{1-\bar{\eta}}{2}(\widetilde{x}_i - \bar{x}_i)^2 \right).$$

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

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

or equivalently,

$$\frac{1-\bar{\eta}}{2}\|\widetilde{x} - \bar{x}\|^2 \leq \frac{1+\bar{\eta}}{2}\big(\|z - \bar{x}\|^2 - \|z - \widetilde{x}\|^2\big).$$

Fejér monotonicity implies that $\lim_{k\to\infty} \|z - x^k\|$ exists. Since both $\widetilde{x}$ and $\bar{x}$ are limit points of $\{x^k\}$, we conclude that $\|z - \widetilde{x}\| = \|z - \bar{x}\|$. Therefore, one has $\|\widetilde{x} - \bar{x}\| \leq 0$, that is, $\widetilde{x} = \bar{x}$. □

Lemma 4.3 implies that, under Assumption 4.1, the hypotheses of [18, Theorem 2.7] continue to hold when $\mu = 1$. Then, by essentially identical reasoning to Theorem 3.6, we may assert:

Theorem 4.4 *The sequence computed by the PMDR using a double regularization with $\mu = 1$, and based on coercive terms conforming to Assumption 4.1, converges to a solution to the variational inequality* (1) *if its solution set is non-empty.*

5 Examples of double regularizations

In this section, we will present example 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 5.1 *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

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

and

$$\frac{(1 - \zeta)(b - x - b + y)(b - y)}{b - x} \leq (1 - \zeta) d'_-(b - x, b - y) \leq (1 - \zeta)(b - x - b + y).$$

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

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

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

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

Using the definition of d, it follows that for $x, y \in (a, b)$,

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

□

5.1 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

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

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

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

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 logarithmic-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.

5.2 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 $h''_i(y_i)$ [18, Section 2.2.1]. One may solve monotone

variational inequality problems using such Bregman distances without resorting to additional problem assumptions like paramonotonicity [6, 13]. We note that [18, Section 4] presents similar results, but under a stronger rescaling, in which, for all $x, y \in B$,

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

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. For such coordinates, a large rescaling factor $\alpha(y)$ will "flatten" the Bregman distance excessively, and may result in numerical difficulties. Thus, the technique of [18, Section 4] is unlikely to be workable in practice. These difficulties are avoided avoided by the double regularization technique employed here.

We first introduce a lemma making it easier to verify whether Assumption 2.3 holds:

Lemma 5.2 *Let $h : \mathbb{R} \mapsto (-\infty, \infty]$, $\operatorname{int}\operatorname{dom} h = \mathbb{R}_{++}$. If $h''(x)$ is nonincreasing and $x^2 h''(x)$ is nondecreasing over $x \in \mathbb{R}_{++}$, then the 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 $0 < x < y$.

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

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

$$\begin{aligned} h'(x) - h'(y) &= \int_y^x h''(z)dz \\ &\geq \int_y^x \frac{y^2 h''(y)}{z^2} dz \qquad \text{[since } x^2 h''(x) \text{ nondecreasing]} \\ &= h''(y)(x-y)\frac{y}{x}. \end{aligned}$$

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

$$\begin{aligned} h'(y) - h'(x) &= \int_x^y h''(z)dz \\ &\geq \int_x^y h''(y)dz \qquad \text{[since } h''(x) \text{ nonincreasing]} \\ &= h''(y)(y-x). \end{aligned}$$

The case $0 < y < x$ is analogous. □

Two examples for Bregman functions that meet the hypotheses of Lemma 5.2 are:

1. $h(x) = \text{dilog}(e^x) + x\ln(e^x - 1)$, where $\text{dilog}(\cdot)$ is the dilogarithm function [16]:

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

In this case,

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

which is clearly nonincreasing.

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

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

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 $x^2 h''(x)$ is nonnegative, and therefore $x^2 h''(x)$ is nondecreasing.

2. $h(x) = x^\alpha - x^\beta$, $\alpha \geq 1,\ \beta \in (0,1)$. In this case,

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

Clearly, $x^2 h''(x)$ 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 $\beta \in (0,1)$.

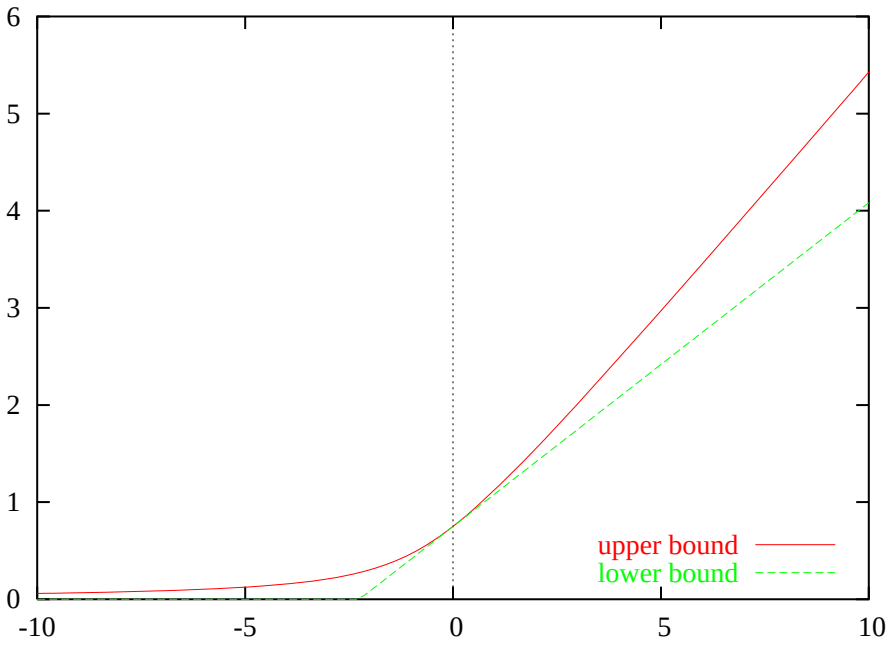

Figure 2: Limits given in Proposition 6.1 for the derivatives of a penalty based on a double regularization.

6 Penalties: conjugate distance measures

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'_k defined in (10), simplifying to (18) in the coercive case. Since P'_k is the inverse of the mapping $\nabla_1 \widetilde{D}(\cdot, y^k)$, its integral P_k is, up to a constant, equal to the *convex conjugate* [20, Chapter 12] $(\widetilde{D}(\cdot, y))^*$ of the function $\widetilde{D}(\cdot, y^k)$.

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 6.1 *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 toa the first argument. If P'_i is continuous and, for each $y_i > 0$, $P'_i(\cdot, y_i)$ is strictly increasing, strictly positive, and one has for all $u \in \mathbb{R}$ that*

$$\frac{u}{\mu + 1} + y_i \leq P'_i(u, y_i) \leq \frac{u + (\mu - 1)y_i + \sqrt{\big(u + (\mu - 1)y_i\big)^2 + 4\mu y_i^2}}{2\mu}, \tag{25}$$

then there is a double regularization component $\widetilde{d}_i$ conforming to Assumption 2.3 such that

$$P_i(\cdot, y_i) = \big(\widetilde{d}_i(\cdot, y_i)\big)^*,$$

where the symbol * denotes the convex conjugate operator [20, Chapter 12].

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

1. $\widetilde{d}_i(\cdot, y_i)$ is closed, strictly convex, and essentially smooth, since it is the conjugate of a differentiable, strictly convex function on $\mathbb{R}$ [20, Theorem 26.3].

2. $\operatorname{int}\operatorname{dom}\widetilde{d}_i(\cdot, y_i) = \operatorname{dom}\widetilde{d}_i'(\cdot, y_i) = \operatorname{rge} P_i'(\cdot, y_i) = \mathbb{R}_{++}$. Here, the first equality follows from [20, Theorem 26.1], the second from [20, Corollary 23.5.1], and the third from the bounds on $P_i'(\cdot, y_i)$.

3. $\widetilde{d}_i(\cdot, y_i)$ 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, $\widetilde{d}_i'(y_i, y_i) = 0$ [20, Corollary 23.5.1].

Now define $d_i(x_i, y_i) \stackrel{\text{def}}{=} \widetilde{d}_i(x_i, y_i) - (\mu/2)\,\|x_i - y_i\|^2$. In view of the three facts above, we need only prove that d_i meets the bounds imposed by Assumption 2.3, and that $\widetilde{d}_i'$ 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
$$\begin{aligned} & \frac{u}{\mu+1} + y_i \le P_i'(u, y_i) \\ \Leftrightarrow \qquad & x_i \le P_i'\big((\mu+1)(x_i - y_i), y_i\big). \end{aligned}$$
As $P_i'(\cdot, y_i)$ is strictly increasing, so is its inverse, $\widetilde{d}_i'(\cdot, y_i)$ [20, Corollary 23.5.1]. Applying this function to both sides of the above inequality, and using the definition of d_i,
$$\begin{aligned} & \widetilde{d}_i'(x_i, y_i) \le (\mu+1)(x_i - y_i) \\ \Leftrightarrow \qquad & d_i'(x_i, y_i) + \mu(x_i - y_i) \le (\mu+1)(x_i - y_i) \\ \Leftrightarrow \qquad & d_i'(x_i, y_i) \le (x_i - y_i). \end{aligned}$$

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

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

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

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

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

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

$$\begin{aligned} & \frac{y_i(x_i - y_i)}{y_i} + \mu(x_i - y_i) \leq \widetilde{d}_i'(x_i, y_i) \\ \Leftrightarrow \quad & \frac{y_i(x_i - y_i)}{x_i} \leq d_i'(x_i, y_i). \end{aligned}$$

Thus, the bounds on d_i are satisfied. Finally, consider the continuity of $\widetilde{d}_i'$. Let $x_i^k \to \bar{x}_i > 0$ and $y_i^k \to \bar{y}_i > 0$ be convergent sequences in $\mathbb{R}_{++} \times \mathbb{R}_{++}$. Let $u^k = \widetilde{d}_i'(x_i^k, y_i^k)$. 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,

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

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 $u^k \to_{\mathcal{K}} \bar{u}$, so the continuity of P_i' ensures that $\bar{x}_i = P_i'(\bar{u}, \bar{y}_i)$, and thus $\bar{u} = \widetilde{d}_i'(\bar{x}_i, \bar{y}_i)$. Thus, u^k is bounded and all its limit points are equal to $\widetilde{d}_i'(\bar{x}_i, \bar{y}_i)$, so it converges to $\widetilde{d}_i'(\bar{x}_i, \bar{y}_i)$. □

6.1 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 *logarithmic-quadratic multiplier method* in [3].

Examining the penalty derivative upper bound (25), we remark on a connection to prior work by Chen and Mangasarian [7]. The bound is exactly the *Chen-Harker-Kanzow-Smale plus function*, defined by

$$P'(w, \beta) \stackrel{\text{def}}{=} \frac{w + \sqrt{w^2 + 4\beta^2}}{2}, \tag{26}$$

computed at $w = u + (\mu - 1)y_i$ and $\beta = \sqrt{\mu} y_i$. In [7], Chean and Mangasarian used this function in a smoothing method — essentially a pure penalty algorithm with no explicit

Lagrange multipliers — for complementarity problems. One may consider the logarithmic-quadratic multiplier method of [3] to be a related algorithm introducing explicit duality and Lagrange multipliers.

The Chen-Harker-Kanzow-Smale plus function was not the only smoothing function studied in [7]. Thus, it is be natural to consider whether other penalties from [7] could be used 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 [7].

6.2 The neural network smooth plus function

In this section, we show that the neural network smooth plus function gives rise to a penalty conforming to the bounds of Proposition 6.1, and thus corresponds 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, as guaranteed by Proposition 6.1, but also to Assumption 4.1. Then, Theorem 4.4 will assure convergence.

Let us recall the formula for the neural network smooth plus function from [7]:

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

We now consider whether a penalty of this form can made to conform to the hypotheses of Proposition 6.1. The analysis for the case $\mu > 1$ appears difficult, so we concentrate on the $\mu = 1$ case. If one follows the transformation used to obtain the logarithmic-quadratic penalty from (26) with $\mu = 1$, one sets $w = u$ and $\beta = y_i$, producing

$$P_i'(u, y_i) = y_i \ln\bigl(e^{u/y_i} + 1\bigr). \tag{27}$$

However, this function cannot possibly conform to the bound (25), which requires $P_i'(0, y_i) = y_i$, whereas (27) 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

$$P_i'(u, y_i) = y_i \log_2\bigl(2^{u/y_i} + 1\bigr). \tag{28}$$

We proceed by letting $\widetilde{d}_i$ be the convex conjugate of $P_i(\cdot, y_i)$ as defined in (28), and then define d_i implicitly via (19) with $\mu = 1$, that is,

$$\widetilde{d}_i'(\cdot, y_i) = (P_i'(\cdot, y_i))^{-1} \tag{29}$$

$$d_i'(x_i, y_i) = \widetilde{d}_i'(x_i, y_i) - (x_i - y_i) \tag{30}$$

Since $\mu = 1$, we seek to show that the d_i implicitly defined by integrating (30) meets Assumption 4.1. Then, Theorem 4.4 will guarantee that the proximal method based on the double regularization components $\widetilde{d}_i$ is convergent. This convergence will imply convergence of the corresponding multiplier method using the penalty (28).

Inserting the definition (28) into (29) and solving for $\widetilde{d}_i'(x_i, y_i)$, we obtain the explicit expression

$$\widetilde{d}_i'(x_i, y_i) = y_i \log_2\left(2^{x_i/y_i} - 1\right).$$

Note that this regularization is a rescaled φ-divergence, as its derivative 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.

Since $\widetilde{d}_i'(\cdot,\cdot)$ is continuous on $\mathbb{R}^n_{++} \times \mathbb{R}^n_{++}$, we need only prove the validity of the the bounds in Assumption 4.1.1. It is easily confirmed that the bounds hold when $x_i = y_i$, so we wish to prove, for all $x_i, y_i > 0$, $x_i \neq y_i$, that

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

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

$$\forall t > 0,\ t \neq 1: \quad t - 1/t < \log_2(2^t - 1) \leq 2t - 2. \tag{31}$$

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 the inequality $2^t + 2^{1/t} < 2^{t+1/t}$ for $t > 0$, $t \neq 1$, which is proved in Appendix A

Finally, we need to show that Assumption 4.1.2 holds. 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,

$$\zeta\frac{(x_i - y_i)y_i}{x_i} \leq d_i'(x_i, y_i) = y_i \log_2\left(2^{x_i/y_i} - 1\right) - (x_i - y_i). \tag{32}$$

Since $\zeta < 1$ and $x_i < y_i$ for small x_i, inequality (32) is implied by

$$\begin{aligned}
& \zeta\frac{(x_i - y_i)y_i}{x_i} \leq y_i \log_2\left(2^{x_i/y_i} - 1\right) - \zeta(x_i - y_i) && (33) \\
\Leftrightarrow \quad & \zeta\frac{x_i^2 - y_i^2}{x_i} \leq y_i \log_2\left(2^{x_i/y_i} - 1\right) \\
\Leftrightarrow \quad & \zeta \geq \frac{x_i y_i \log_2\left(2^{x_i/y_i} - 1\right)}{x_i^2 - y_i^2} \\
\Leftrightarrow \quad & \zeta \geq \frac{x_i}{y_i}\log_2\left(2^{x_i/y_i} - 1\right)\frac{y_i^2}{x_i^2 - y_i^2}. && (34)
\end{aligned}$$

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

$$\lim_{t\downarrow 0} \frac{t\log_2(2^t-1)}{t^2-1} = 0. \tag{35}$$

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 (35) 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

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

Since inequality (34) is equivalent to (33), which implies (32), Assumption 4.1.2 holds.

7 Computational tests

We conclude with some preliminary computational experiments with multiplier methods using double regularizations. Our main objective is to study the behavior of the neural penalty as compared to the logarithmic-quadratic penalty. Chen and Mangasarian [7] observed better results with the neural penalty, so the question naturally arises whether the same holds true when explicit multipliers are present.

We coded MATLAB implementations of the multiplier methods using each of the two penalties, and applied them to all the nonlinear complementarity problems in the MATLAB version of MCPLIB [9]. We treat these problems as being in the form (2) from Section 1. Considering differing starting points, the test set has 83 problems, most of which are

not monotone. Even though our convergence analysis does not hold in the non-monotone case, reasonable performance on the MCPLIB may be considered important to establishing whether a method is useful in practice, as in [12].

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

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

Our augmented Lagrangian methods take the form (11)-(12). With the S scaling, these recursions become

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

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

$$\begin{aligned} 0 &= F(x^{k+1}) - P_k'(-\alpha_k S^{-1} x^{k+1}) \\ y^{k+1} &= P_k'(-\alpha_k S^{-1} x^{k+1}), \end{aligned}$$

Expanding P_k', we obtain the recursions

$$F(x^{k+1}) - \begin{bmatrix} P_i'(-(\alpha_k/S_{11})x_1^{k+1}, y_1^k) \\ \vdots \\ P_i'(-(\alpha_k/S_{nn})x_n^{k+1}, y_n^k) \end{bmatrix} = 0,$$

$$y^{k+1} = \begin{bmatrix} P_i'(-(\alpha_k/S_{11})x_1^{k+1}, y_1^k) \\ \vdots \\ P_i'(-(\alpha_k/S_{nn})x_n^{k+1}, y_n^k) \end{bmatrix}.$$

The penalty terms for the logarithmic-quadratic method are

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

whereas for the neural method, the penalty terms are

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

We employed a Newton algorithm with Armijo line search to solve system of nonlinear equations (7). In particular, we used the `nsola` code from [14]. In order to deal with nearly singular Jacobians, we incorporated the perturbation proposed in [8, Section 6.5]. All nonlinear equations were solved essentially exactly, that is, until the residual less than or equal to 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 computational results.

- We set the initial stepsize α_0 to 10. If, after successful solution of the nonlinear equation, the feasibility of the primal solution or its complementarity with the multiplier did not improve by a factor of 0.5, we divided 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 [12].

- As suggested in [7, 12], 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 [10]. Complete test results appear in Tables 1 and 2.

7.1 The impact of μ

To make a fair comparison between the logarithmic-quadratic penalty and the neural penalty, we must first study how adjusting μ affects the performance of the logarithmic-quadratic penalty. For the neural penalty, μ is fixed at 1.

We tested the logarithmic-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.[2] Clearly, performance tends to improve as μ decreases. Therefore, we should use the smallest possible μ when comparing the penalties. In our subsequent testing, we used $\mu = 1.05$, since we have only proved convergence of the logarithmic-quadratic method when $\mu > 1$. The performance of this case is very close to the limiting case $\mu = 1$.

7.2 Neural penalty performance

Next, testing of the neural method of multipliers showed it to be faster than the logarithmic-quadratic method on most problems; Figure 4 gives the performance profiles, with the "pure neural" line representing the neural method. On the other hand, the neural method appears less reliable, solving a significantly smaller fraction of the problems. After a careful review of the results, it became clear that the reliability problem lay in the solution of the first system of nonlinear equations. This system is naturally the most difficult one, since the

[2] Following [10], 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 required by any method on instance p. 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.

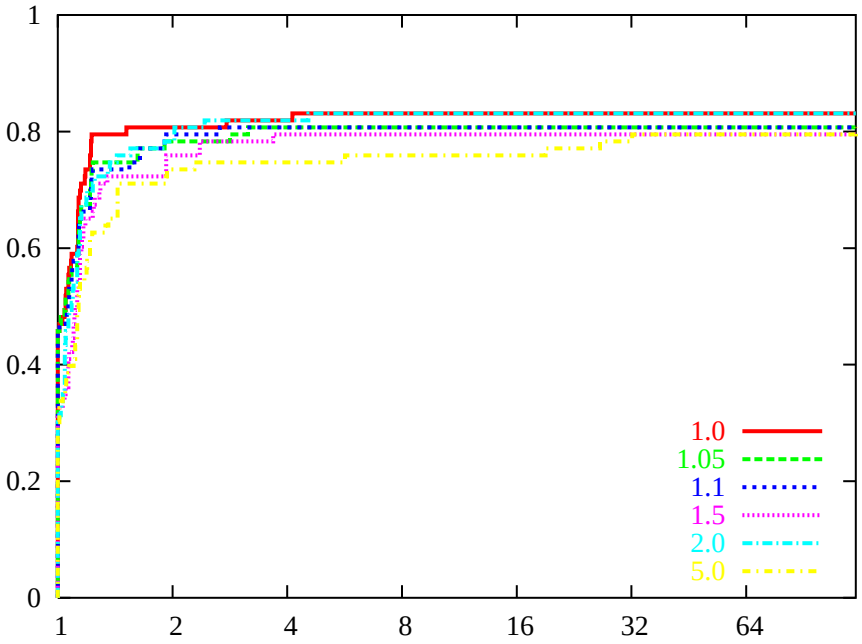

Figure 3: The impact of μ in the performance of the logarithmic-quadratic penalty

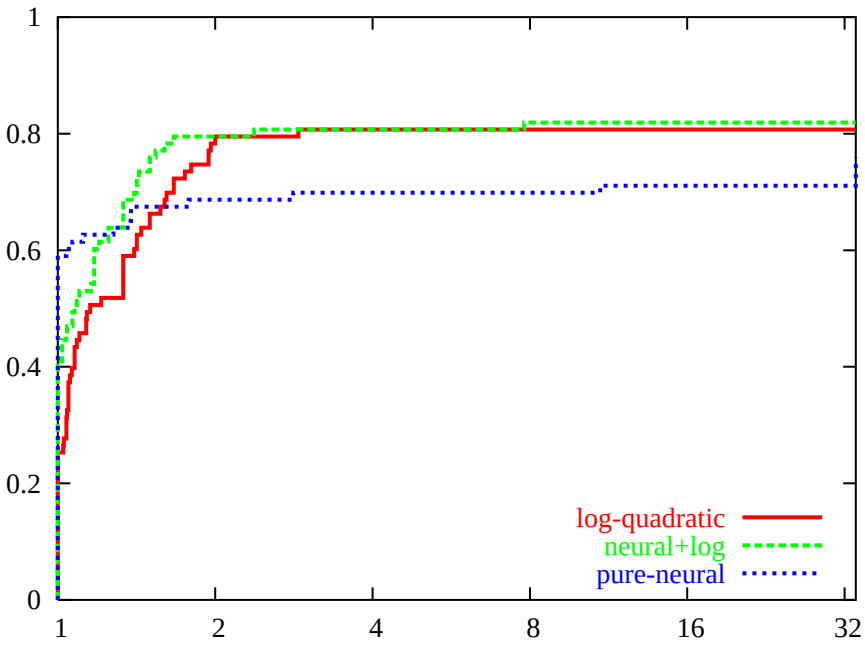

Figure 4: Neural versus logarithmic-quadratic method of multipliers

initial primal iterate x^0 is unlikely to provide a good starting point. The neural multiplier appears to suffer more than the logarithmic-quadratic method from poor starting points. On the other hand, if the first step is successful, the neural penalty turns out to require fewer Newton steps in subsequent iterations.

To overcome this first-step problem, we decided to initialize the neural multiplier method with a single logarithmic-quadratic step: we first perform one logarithmic-quadratic step and multiplier update, and subsequently use only the neural penalty. This algorithm appears Figure 4 as the "neural+log" line. This hybrid method essentially dominates the logarithmic-quadratic method, although it is significantly slower than the "pure" neural method on the easiest problems. It remains to be investigated whether other initialization procedures might have a similar effect on the reliability of the neural method, but with a smaller speed penalty. Overall, the results agree with the Chen-Mangasarian results from the pure penalty setting: both neural methods are faster for most problem instances.

Unfortunately, none of the methods tested are extremely reliable on the MCPLIB, solving at best only slightly better than 80% of the test cases. Reviewing the individual run logs, run failures usually result from non-convergence of the nonlinear equation solver. We believe that the reliability of the multiplier methods can still be improved by replacing the simple Newton nonlinear equation solver in our code with a more robust version. Another issue that deserves careful investigation is how to find a good initial multiplier. Improvements in this area could lead to faster and more reliable methods. Methods of this class have not been tested practically before, and we believe our preliminary results here show them to have some promise.

References

[1] H. Attouch and M. Théra. A general duality principle for the sum of two operators. *Journal of Convex Analysis* 3:1–24, 1996.

[2] A. Auslender, M. Teboulle, and S. Ben-Tiba. A logarithmic-quadratic proximal method for variational inequalities. *Computational Optimization and Applications* 12:31–40, 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*, 24:645–668, 1999.

[4] A. Auslender and M. Teboulle. Lagrangian duality and related multiplier methods for variational inequality problems. *SIAM Journal on Optimization*, 10:1097–1115, 2000.

[5] Y. Censor and S.A. Zenios. Proximal minimization algorithm with D-functions. *Journal of Optimization Theory and Applications* 73:451–464, 1992.

Problem	Log-quad $\mu = 5.0$	Log-quad $\mu = 1.5$	Log-quad $\mu = 1.05$	Log-quad $\mu = 1.0$	Pure Neural	Neural with One L-Q Step
bertsekas1	FAIL	99	68	42	48	68
bertsekas2	25	22	25	26	22	22
bertsekas3	73	33	101	132	268	95
bertsekas4	FAIL	99	68	42	48	68
bertsekas5	25	21	24	24	24	24
bertsekas6	27	23	25	25	24	25
colvdual1	FAIL	FAIL	FAIL	FAIL	30	FAIL
colvdual2	32	30	27	26	FAIL	26
colvdual3	17	16	17	17	15	16
colvdual4	2606	FAIL	FAIL	21	FAIL	968
colvnlp1	24	23	22	22	21	21
colvnlp2	20	21	19	20	23	18
colvnlp3	17	16	17	17	15	16
colvnlp4	20	21	22	22	29	21
colvnlp5	20	21	22	22	29	21
colvnlp6	17	14	15	15	13	14
cycle1	4	4	4	4	3	4
explcp1	17	22	21	20	13	20
hanskoop10	33	23	23	26	773	27
hanskoop2	33	23	23	26	773	27
hanskoop4	33	23	23	26	773	27
hanskoop6	33	23	23	26	FAIL	27
hanskoop8	33	23	23	26	773	27
josephy1	16	17	15	15	16	15
josephy2	20	27	FAIL	FAIL	FAIL	FAIL
josephy3	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL
josephy4	FAIL	18	16	16	16	16
josephy5	15	16	14	14	14	14
josephy6	23	FAIL	FAIL	FAIL	30	FAIL
josephy7	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL
josephy8	13	15	13	13	13	13
kojshin1	450	450	450	400	300	300
kojshin2	450	FAIL	FAIL	FAIL	FAIL	FAIL
kojshin3	450	FAIL	400	450	FAIL	300
kojshin4	450	18	14	14	14	14
kojshin5	450	21	17	17	17	17
kojshin6	495	FAIL	FAIL	FAIL	FAIL	FAIL
kojshin7	450	400	400	400	250	300
kojshin8	450	450	400	400	300	300
mathinum1	11	10	10	10	6	9
mathinum2	9	9	9	9	5	8
mathinum3	FAIL	13	13	13	9	12

Table 1: Number of Newton steps, part 1

Problem	Log-quad $\mu = 5.0$	Log-quad $\mu = 1.5$	Log-quad $\mu = 1.05$	Log-quad $\mu = 1.0$	Pure Neural	Neural with One L-Q Step
mathinum4	10	10	10	10	6	9
mathinum5	15	16	16	16	12	15
mathinum6	11	11	11	11	7	10
mathisum1	17	15	14	14	13	13
mathisum2	16	13	13	13	13	13
mathisum3	450	400	400	400	300	300
mathisum4	19	16	14	14	13	13
mathisum5	1	1	1	1	1	1
mathisum6	362	22	19	19	19	19
mathisum7	350	350	350	350	200	250
nash1	9	9	9	9	6	8
nash2	8	8	8	8	6	8
nash3	7	7	7	7	5	7
nash4	6	6	6	6	3	5
pgvon1054	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL
pgvon1055	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL
pgvon1056	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL
pgvon1064	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL
pgvon1065	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL
pgvon1066	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL
powell1	102	102	102	102	52	53
powell2	54	102	54	53	52	54
powell3	53	102	101	53	52	53
powell4	102	102	150	53	52	53
powell5	2557	1658	451	1250	FAIL	1071
powell6	102	101	101	101	52	52
scarfanum1	21	22	22	22	20	22
scarfanum2	24	25	25	25	23	25
scarfanum3	22	FAIL	27	27	27	26
scarfanum4	20	22	23	23	19	22
scarfbnum1	FAIL	42	51	51	FAIL	398
scarfbnum2	FAIL	FAIL	FAIL	151	FAIL	FAIL
scarfbsum1	43	40	36	37	FAIL	36
scarfbsum2	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL
sppe1	14	16	17	17	19	17
sppe2	15	16	16	16	16	16
sppe3	10	11	10	10	10	10
tobin1	23	24	22	22	229	21
tobin2	34	39	38	33	51	37
tobin3	37	46	42	56	73	41
tobin4	17	16	14	14	13	13

Table 2: Number of Newton steps, part 2

[6] 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* 81:373–400, 1998.

[7] C. Chen and O.L. Mangasarian. A class of smoothing functions for nonlinear and mixed complementarity problems. *Computational Optimization and Applications*, 5:97–138, 1996.

[8] J.E. Dennis and R.B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Classics in Applied Mathematics, 16. SIAM, Philadelphia, 1996.

[9] S.P. Dirkse and M.C. Ferris. Modeling and solution environments for MPEC: GAMS & MATLAB. *Reformulation: nonsmooth, piecewise smooth, semismooth and smoothing methods (Lausanne, 1997)*, 127–147, Applied Optimization, 22. Kluwer Academic Publisher, Dordrecht, 1999.

[10] E.D. Dolan and J.J. Moré. Benchmarking optimization software with performance profiles. *Mathematical Programming* 91(2): 201–213, 2002.

[11] J. Eckstein. Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming. *Mathematics of Operations Research* 18:202-226, 1993.

[12] J. Eckstein and M.C. Ferris. Smooth methods of multipliers for complementarity problems. *Mathematical Programming*, 86:65–90, 1999.

[13] A.N. Iusem. On some properties of paramonotone operators. *Journal of Convex Analysis* 5:269–278, 1998.

[14] C.T. Kelley. *Iterative methods for linear and nonlinear equations* SIAM, Philadelphia, 1995.

[15] K.C. Kiwiel. On the twice differentiable cubic augmented Lagrangian. *Journal of Optimization Theory and Applications* 88:233–236, 1996.

[16] L. Lewin, ed. *Structural Properties of Polylogarithms.* Mathematical Surveys and Monographs vol. 37, American Mathematical Society, Providence, 1991.

[17] U. Mosco. Dual variational inequalities. *Journal of Mathematical Analysis and Applications* 40:202–206, 1972.

[18] 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*, 12:238–261, 2001.

[19] S.M. Robinson. Composition duality and maximal monotonicity. *Mathematical Programming* 85:1–13, 1999.

[20] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

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

[22] R.T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research* 1:97–116, 1976.

[23] M. Teboulle. Entropic proximal mappings with applications to nonlinear programming. *Mathematics of Operations Research* 17:670–690, 1992.

A An inequality relating to the neural penalty

Lemma A.1 *For $x \in (0,1)$, $\ln(x)\ln(1-x) \leq \ln(2)^2$, with equality only for $x = 1/2$.*

Proof. Define $g(x) \stackrel{\text{def}}{=} \ln(x)\ln(1-x)$. Using L'Hôpital's rule twice,

$$\lim_{x\to 1} g(x) = \lim_{x\to 1} \frac{\ln(1-x)}{\ln(x)^{-1}} = \lim_{x\to 1} \frac{x\ln(x)^2}{1-x} = \lim_{x\to 1} \frac{\ln(x)^2 + 2\ln(x)}{-1} = 0.$$

By the symmetry $g(1-x) = g(x)$, we also have $\lim_{x\to 0} g(x) = 0$. In summary, g is strictly positive and continuous on $(0,1)$, and has $\lim_{x\to 0} g(x) = \lim_{x\to 1} g(x) = 0$, so we conclude that g must attain a maximum on $(0,1)$. A necessary condition for the maximum is

$$\begin{aligned} & g'(x) = \frac{\ln(1-x)}{x} - \frac{\ln(x)}{1-x} = 0 \\ \Leftrightarrow \quad & (1-x)\ln(1-x) = x\ln(x) \\ \Leftrightarrow \quad & (1-x)^{1-x} = x^x. \end{aligned}$$

This condition is satisfied by $x = 1/2$, and there can be no other solution because x^x is a strictly increasing function, while $(1-x)^{1-x}$ is strictly decreasing. Therefore, g's unique maximum value over $(0,1)$ is $g(1/2) = \ln(2)^2$. □

Proposition A.2 *For all $t > 0$,*

$$2^t + 2^{1/t} \leq 2^{t+1/t} \qquad\qquad 2^{-t} + 2^{-1/t} \leq 1,$$

with equality only for $t = 1$.

Proof. Multiplying or dividing through by $2^{t+1/t} > 0$ shows that the two claimed inequalities are equivalent, so it suffices to prove the second one. For any given $t > 0$, let $x = 2^{-1/t} \in (0,1)$. Using Lemma A.1 and dividing by $\ln(x) < 0$ yields $\ln(1-x) \geq \ln(2)^2/\ln(x)$, with equality only if $x = 1/2$. Substituting $x = 2^{-1/t}$ then yields

$$\ln\bigl(1 - 2^{-1/t}\bigr) \geq \frac{\ln(2)^2}{-(1/t)\ln(2)} = -t\ln(2),$$

with equality only if $t = 1$. Exponentiating both sides yields $1 - 2^{-1/t} \geq 2^{-t}$, with equality only if $t = 1$, which is equivalent to the claimed result. □