A THEORY OF GENERALIZED PROXIMITY FOR ADMM

François D. Côté, Ioannis N. Psaromiligkos, and Warren J. Gross

Department of Electrical and Computer Engineering, McGill University, Montreal, Canada

ABSTRACT

The alternating direction method of multipliers has become in recent years the most widely used proximal method for signal processing. In this paper, we lay the groundwork for a new notion of proximity and use it to illustrate that the method (ADMM) is actually somewhat of a maladroit rearrangement of a new, more practical procedure that generalizes the Douglas-Rachford algorithm. Compared to ADMM, the algorithm that we propose enjoys not only a more sensible form, but also a more general convergence result.

Index Terms— ADMM, Douglas-Rachford, proximal methods

1. INTRODUCTION

The alternating direction method of multipliers (ADMM) is a very popular approach for decomposing a complex optimization problem into an iterative procedure involving a pair of manageable problems. Being proximal in nature, the two problems can often themselves be decomposed into problems that can be tackled in a parallel manner, making the method especially useful in modern, distributed settings. The method has found wide application, for example, in distributed implementations of regression and classification. Examples of such treatments include [1] and [2]. The method is remarkably effective, and even rivals specialized solutions. See [3] for a survey.

There are several versions of ADMM, and it is well known that some can be derived from the Douglas-Rachford algorithm (DR) [4]. In this work, we revisit this idea. Through a new notion of proximity that we develop in this paper, we show that a very general version of ADMM arises from a new version of DR that we also propose.

Our algorithm offers both practical and theoretical advantages. Compared to ADMM, it runs more efficiently, affording relaxation without requiring additional memory, and under assumptions that can be verified without detouring through duality, converges more generally, allowing for infinite dimensions and inexact steps.

1.1. Related work and contributions

Our paper is unavoidably mathematical. At the surface, it touches on convex analysis, and at a deeper level, on monotone operator theory. For a brief review, see [5, Sec. 2]; for a complete treatment, see [6]. Most relevant are Problem 4, Proposition 14, and Theorem 20 in [5], which our study completely generalizes.

Our investigation develops the notion of generalized proximity, a concept that we have flirted with before [7] and recently used for dimensionality reduction [8]. Without refinement, the idea was also used by Combettes and Pesquet [9], who, while describing ADMM, coined the notation that we employ herein for the general operator. (The concept should not be confounded with the proximal mapping relative to a metric [6, Prop. 24.24], which has the same notation.) Although the idea is not new, no precedent has related the idea to the operation of infimal postcomposition—an instrumental operation in the development of our algorithm.

In this paper, we use infimal postcomposition to reformulate the problem that ADMM solves. Such a reformulation was inspired by Yan and Yin, who proposed it to study the many ways of applying ADMM [10]. Their analysis, however, does not explicitly yield our algorithm, nor is it framed by a theory of proximity.

Our key finding is that under the assumption behind our theory, infimal postcomposition is ensured to produce a proper convex lower semicontinuous function. Such a guarantee (which recently became a textbook result [6, Cor. 25.44 (i)]) was also given by Becker and Combettes [11], but under a different assumption, based on duality. In contrast, our assumption is totally primal.

Two popular ways exist for proving the convergence of ADMM. One, due to Gabay, involves DR and a Fenchel dual problem [4], and the other, due to Fortin and Glowinski, utilizes the Lagrangian [12]. These strategies are based on duality theories. For a treatise on the convergence of ADMM, see [13]. Our theory promotes a third way. Similar to Gabay's strategy, it allows us to obtain ADMM from DR. But our treatment is more direct, dealing only with a primal problem.

Let us point out our contributions:

- We generalize the problem that DR solves, including both its qualification condition and objective. See Problem 1.
- We describe scenarios in which the qualification condition holds. This is Proposition 2—also a generalization.
- We introduce a new theory of proximity, furnishing several new results, such as Theorem 7 (our key finding mentioned earlier), useful for deriving our algorithm. See Section 3.
- Our algorithm, which generalizes DR, is given in Theorem 8.
- We show how existing versions of ADMM can be recovered from our algorithm. See Section 5.

This paper is expository in nature. We focus on dispensing the results, omitting lengthy proofs.

We stress that our theory of proximity may not only be useful for developing our algorithm; it may also have broad implications for convex analysis. An immediate consequence of our work is that the methods that our algorithm subsumes inherit our convergence result, and therefore converge more generally than previously known.

1.2. Some notation

Before we begin our study, it is useful to introduce a bit of notation. Our notation is rather standard. We use $\mathbb{N}$ for the set of nonnegative integers, and $\mathcal{F}$, $\mathcal{G}$, and $\mathcal{H}$ for real Hilbert spaces; we use $\|\cdot\|$ for the norms induced by the inner products on these spaces, and Id for the identity operators on the spaces; by $\Gamma_0(\mathcal{H})$ we understand the class comprising every proper convex lower semicontinous function from $\mathcal{H}$ to $(-\infty, +\infty]$; we use $\ker$ for the null space of a mapping, and dom for the effective domain of a function; we use cone for the conical hull, int for the interior, and ri for the relative interior of a set; and we recognize $\mathcal{B}(\mathcal{G}, \mathcal{H})$ as the space of all bounded linear operators from $\mathcal{G}$ to $\mathcal{H}$. We will introduce other notation as needed throughout the course of our study.

Proc. 2017 IEEE Global Conf. on Signal and Information Processing (GlobalSIP '17)

2. PROBLEMS

While ADMM is widely known to the signal processing community, DR has received less attention. To understand the advantages of our method, it is important to understand how the two approaches relate to each other in the context of the different problems that they solve.

The purpose of both approaches is to minimize a sum of two functions by treating only one function at a time. Until now, ADMM has been preferred over DR for linearly constrained problems.

DR originated as a procedure for solving linear equations [14], and Varga gave an early account with relaxation [15]. The nonlinear form of the method, which solves the problem

$$\min_{x\in\mathcal{H}} \ f(x)+g(x), \tag{1}$$

where f and g belong to $\Gamma_0(\mathcal{H})$, is due to Lions and Mercier [16]. Eckstein and Bertsekas generalized the approach, adding relaxation and inexact steps [17]. Combettes improved this version, weakening the assumptions for convergence [18], and together with Pesquet, introduced it to the signal processing community [5]. Other versions (incorporating inertia [19], for example) have also surfaced.

ADMM was conceived by Glowinski and Marroco for solving nonlinear problems [20]. For solving the problem

$$\min_{(y,z)\in\mathcal{F}\times\mathcal{H}} \ f(y)+g(z) \quad \text{s.t.} \quad Ay=z, \tag{2}$$

where g is still as in (1), but $f\in\Gamma_0(\mathcal{F})$, and $A\in\mathcal{B}(\mathcal{F},\mathcal{H})$, the method is due to Gabay and Mercier [21]. Proving its convergence, Gabay showed that the method can be derived by applying DR to the dual problem [4]. Following this approach, Eckstein and Bertsekas developed the ADMM analog to their version of DR [17], assuming, in addition to finite dimensions, that A is a full column rank matrix. Fortin and Glowinski established the convergence of ADMM using the Lagrangian [12]. Boyd et al. thusly proved another version of the method [3], tailored to the problem

$$\min_{(y,z)\in\mathcal{F}\times\mathcal{G}} \ f(y)+g(z) \quad \text{s.t.} \quad Ay+Bz=c, \tag{3}$$

where f and A remain unchanged from (2), but $g\in\Gamma_0(\mathcal{G})$, and $B\in\mathcal{B}(\mathcal{G},\mathcal{H})$ and $c\in\mathcal{H}$. Boyd et al. also mentioned how to add relaxation, and the outcome was considered by Nishihara et al. [22]. Whereas Boyd et al. and Nishihara et al. restricted their attention to finite dimensions, Davis and Yin redeveloped the method from DR, establishing a relaxed version of ADMM in infinite dimensions [23]. As is the case with DR, other adaptations of ADMM—such as one combining inertia and relaxation [24]—have also been proposed.

Of all these variations, the algorithm of Combettes and Pesquet enjoys the most general convergence result. But the inability of this method to directly handle (3) has, ostensibly, made it less appealing even than the approach described by Boyd et al., which does, albeit only explicitly in finite dimensions.

Our goal is to generalize DR, making it conducive to problems that ADMM solves, while keeping the general convergence result.

2.1. A more general problem for DR

In this paper, we develop a version of DR that expressly handles the following problem, which generalizes Problem 4 in [5]:

Problem 1. *Let f be a function in $\Gamma_0(\mathcal{F})$, and g a function in $\Gamma_0(\mathcal{G})$. Let A be an operator in $\mathcal{B}(\mathcal{F},\mathcal{H})$, and B an operator in $\mathcal{B}(\mathcal{G},\mathcal{H})$. Finally, let c, d, and e be vectors in $\mathcal{H}$. Suppose that*

$$f+\|A\cdot\|^2 \quad \textit{and} \quad g+\|B\cdot\|^2 \quad \textit{are strongly convex} \tag{4}$$

and that the following so-called qualification condition is satisfied:

$$\operatorname{cone}(A\operatorname{dom} f + B\operatorname{dom} g - c) \quad \textit{is a closed subspace.} \tag{5}$$

The objective is to solve the problem

$$\min_{(y,z)\in\mathcal{F}\times\mathcal{G}} \ f(y)+g(z) \quad \text{s.t.} \quad Ay+Bz=c=d+e.$$

Remark. There is a reason for expressing the constant c as $d+e$. Unlike ADMM, our algorithm will depend on d and e separately.

Problem 4 in [5] is a special case of our problem with $A=\operatorname{Id}$, $B=-\operatorname{Id}$, and $c=0$. In this case, (4) holds automatically.

In general, (4) holds in various cases. Consider $f+\|A\cdot\|^2$ for instance. This function is strongly convex provided that f or $\|A\cdot\|^2$ manifests the property. For the latter, this property is equivalent to A being bounded below, or in finite dimensions, to A being a full column rank matrix.

For the qualification condition, we can establish the following result, which generalizes Proposition 14 in [5]:

Proposition 2. *Condition* (5) *holds in any of the following cases:*

1) $\operatorname{int}(A\operatorname{dom} f)\cap(-B\operatorname{dom} g+c)\neq\emptyset$ *or* $\big(-\operatorname{int}(B\operatorname{dom} g)\big)\cap(A\operatorname{dom} f-c)\neq\emptyset$.
2) *f is finite and A is surjective, or g is finite and B is surjective.*
3) *$\mathcal{H}$ is finite dimensional and* $\operatorname{ri}(A\operatorname{dom} f)\cap\big(-\operatorname{ri}(B\operatorname{dom} g)+c\big)\neq\emptyset$.

Proposition 14 in [5] is a special case of this result with $A=\operatorname{Id}$, $B=-\operatorname{Id}$, and $c=0$. In this case, A and B are inevitably surjective.

Since it includes problems to which ADMM applies, Problem 1 covers many practical examples. We stress that the conditions in the problem are not hard to satisfy, as the next example illustrates.

Example 3. Consider the following signal recovery problem:

$$\min_{y\in L^2(\Omega)} \ \|\Phi y-v\|^2+\alpha\|y\|_{L^1(\Omega,w)},$$

where $L^p(\Omega,w)$ is the space of Lebesgue measurable real-valued signals and $\|y\|_{L^p(\Omega,w)}=\left(\int_\Omega|y(t)|^p w(t)\,dt\right)^{\frac{1}{p}}<+\infty$ for $p\geq 1$. When w is equal to 1 everywhere, the space is the usual $L^p(\Omega)$ space of real-valued functions and the norm is the usual L^p norm, $\|\cdot\|_{L^p}$. In the example, w is a bounded continuous function that is bounded below away from zero, Ω is a bounded open subset of $\mathbb{R}^n$, the scalar α is positive, and Φ is a bounded linear operator from $L^2(\Omega)$ to a real Hilbert space containing v.

We can express the example in the form of Problem 1 by setting $\mathcal{F}$, $\mathcal{G}$, and $\mathcal{H}$ to $L^2(\Omega)$, defining $f=\|\Phi\cdot-v\|^2$ and $g=\alpha\|\cdot\|_{L^1}$ together with $(Ay)(t)=y(t)w(t)$ for $t\in\Omega$ and $B=-\operatorname{Id}$, and letting $c=0$. It follows that all the stipulations in Problem 1 hold. In particular, (4) holds because A and B are bounded below, and (5) holds because g is finite and B is surjective.

3. GENERALIZED PROXIMITY

In this section, we develop our notion of generalized proximity and connect it to infimal postcomposition. Unless mentioned otherwise, we will assume the context of the following problem:

Problem 4. *Let γ be a positive real number, and x a vector in $\mathcal{H}$, and let f be a function in $\Gamma_0(\mathcal{F})$, and A an operator in $\mathcal{B}(\mathcal{F},\mathcal{H})$. Suppose that*

$$f+\|A\cdot\|^2 \quad \textit{is strongly convex.}$$

The objective is to solve the problem

$$\min_{y\in\mathcal{F}} \ \gamma f(y) + \frac{1}{2}\|Ay - x\|^2.$$

This problem generalizes the problem that defines the proximal mapping. In fact, we can prove that Problem 4 must always have a unique solution. We denote this solution by

$$\operatorname{prox}^A_{\gamma f} x,$$

which generalizes the proximal mapping, defined according to

$$\operatorname{prox}_{\gamma f} = \operatorname{prox}^{\mathrm{Id}}_{\gamma f}.$$

The proximal mapping, which was formulated by Moreau [25], possesses several attractive properties. One that will be useful to us is the implication of a proximal point solving (1). Specifically, we have the following result [5, Prop. 18 (iii)]:

Lemma 5. *Consider* (1) *in the context of Problem 4 in [5], and let x^* be a solution to* (1)*. Then,*

$$x^* = \operatorname{prox}_{\gamma g} x \implies x^* = \operatorname{prox}_{\gamma f}(2x^* - x).$$

This relation is inherent in DR, and we can easily recognize it by the form that the iterates take. See [5, Th. 20].

The generalized operator inherits many properties of the vanilla operator (continuity, for example). But it also manifests new ones, such as allowing a local inversion of the possibly non-injective A. We formalize this property in the following lemma:

Lemma 6. *Let y be a vector in $\mathcal{F}$. Suppose that*

$$f(y) \le f(y + \Delta y) \quad \textit{for every } \Delta y \textit{ in } \ker A.$$

Then,

$$A \operatorname{prox}^A_{\gamma f} x = Ay \implies \operatorname{prox}^A_{\gamma f} x = y.$$

Proof. By the left side of the implication, the proximal point must be equal to y plus a vector Δy in $\ker A$. To justify the implication, we will prove that $\Delta y = 0$. Suppose for contradiction that $\Delta y \neq 0$. Then, by the definition of the proximal mapping,

$$\gamma f(y) + \frac{1}{2}\|Ay - x\|^2 > \gamma f(y + \Delta y) + \frac{1}{2}\|A(y + \Delta y) - x\|^2$$

(where the inequality is strict because the proximal point is unique). Since $A\Delta y = 0$ and $\gamma > 0$, we end up with $f(y) > f(y + \Delta y)$, contradicting the assumption in the statement of the lemma. □

3.1. Infimal postcomposition

An operation that is interestingly related to generalized proximity is infimal postcomposition. In this study, we will apply this operation with respect to an affine operator. Still in the context of Problem 4, let c be a vector in $\mathcal{H}$. The operation is defined by

$$(A\cdot + c) \triangleright f(x) = \inf_{\substack{y\in\mathcal{F}\\ Ay + c = x}} f(y), \quad x \in \mathcal{H},$$

where we understand that $\inf \emptyset = +\infty$. Functions that result from this operation have nice properties, such as [6, Prop. 12.36 (i)]

$$\operatorname{dom}\big((A\cdot + c) \triangleright f\big) = A \operatorname{dom} f + c. \tag{6}$$

We can also establish the following important fact:

Theorem 7. *In the context of Problem 4, it holds that*

$$(A\cdot + c) \triangleright f \in \Gamma_0(\mathcal{H}).$$

Theorem 7 is important because it confirms that functions that result from infimal postcomposition belong to the class of functions with which we are concerned. Moreover, proximal points relative to such functions are well defined. In fact, we can show that they can be expressed by

$$\operatorname{prox}_{\gamma((A\cdot+c)\triangleright f)} x = A \operatorname{prox}^A_{\gamma f}(x - c) + c, \quad x \in \mathcal{H}. \tag{7}$$

Combined with Lemma 6, this relation is useful in generalizing DR.

4. GENERALIZING DR

To obtain our algorithm, we will first re-express Problem 1. Looking at the problem, we notice that to any pair of vectors y and z satisfying the constraint, we can associate a vector x in $\mathcal{H}$ such that

$$x = Ay - d \tag{8}$$

and

$$x = -Bz + e. \tag{9}$$

Now consider finding an x^* associated to a solution to the problem. For each x in $\mathcal{H}$, it suffices to retain among the values of y satisfying (8), the vectors minimizing f, and similarly, among the values of z satisfying (9), those minimizing g. The problem, then, is to find an x for which a retained pair of y and z (if it exists) minimizes $f + g$. This nested procedure can be expressed as

$$\min_{x\in\mathcal{H}} \ (A\cdot - d) \triangleright f(x) + (-B\cdot + e) \triangleright g(x). \tag{10}$$

Recalling both (6) and Theorem 7, and turning to Problem 4 in [5], we see that Problem 1 can be solved using Theorem 20 in [5].

The point of this strategy is that the resulting algorithm can be described using generalized proximity so that it produces a solution to Problem 1. The benefit of the approach is significant: while the decision space is $\mathcal{H}$, the method yields a solution in $\mathcal{F} \times \mathcal{G}$. Such an approach is especially practical in contexts of machine learning, where $\mathcal{H}$ is finite dimensional and $\mathcal{F}$ and $\mathcal{G}$ are infinite dimensional; such is the case in nonlinear classification. See, for example, [8].

We now present our algorithm. The following result generalizes Theorem 20 in [5]:

Theorem 8. *Let γ be a positive real number, and let $(\lambda_n)_{n\in\mathbb{N}}$ be a sequence in $(0, 2)$. Let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be two sequences in $\mathcal{F}$ and $\mathcal{G}$, respectively. Suppose that*

(i) *Problem 1 has at least one solution, (y^*, z^*);*

(ii) $\sum_{n\in\mathbb{N}} \lambda_n(2 - \lambda_n) = +\infty$*;*

(iii) $\sum_{n\in\mathbb{N}} \lambda_n(\|Aa_n\| + \|Bb_n\|) < +\infty$.

Take any x_0 in $\mathcal{H}$, and for every n in $\mathbb{N}$, repeat

$$\begin{aligned}
z_n &= \operatorname{prox}^{-B}_{\gamma g}(x_n - e) + b_n,\\
y_n &= \operatorname{prox}^{A}_{\gamma f}\big(2(-Bz_n + e) - x_n + d\big) + a_n, \quad \textit{and}\\
x_{n+1} &= x_n + \lambda_n(Ay_n + Bz_n - c).
\end{aligned}$$

Then,

(a) *$(x_n)_{n\in\mathbb{N}}$ converges weakly to some point x in $\mathcal{H}$,*

(b) $\operatorname{prox}^{-B}_{\gamma g}(x - e) = z^*$ *and*
$\operatorname{prox}^{A}_{\gamma f}\big(2(-Bz^* + e) - x + d\big) = y^*$,

and if $\mathcal{H}$ is finite dimensional and $a_n \to 0$ and $b_n \to 0$, then

(c) $y_n \to y^*$ *and* $z_n \to z^*$.

Proof. Applying [5, Th. 20] to (10), we obtain Theorem 8 through (7) after redefining a_n and b_n in [5, Th. 20] as Aa_n and $-Bb_n$, respectively. What remains to show is that (b) and (c) hold.

Recalling (7) and (9), we must, according to [5, Th. 20], have

$$-B \operatorname{prox}_{\gamma g}^{-B}(x-e) + e = -Bz^* + e.$$

However, since $g(z^*) \leq g(z^* + \Delta z)$ for any Δz in $\ker B$ (since otherwise z^* would not be optimal), there is only one possibility for z^* according to Lemma 6. Through Lemma 5, we obtain a similar result for y^*. The conclusion is that (b) holds.

Finally, if $\mathcal{H}$ is finite dimensional, then weak convergence is the same as strong convergence, so $x_n \to x$. By the continuity of the proximal mapping, and since $a_n \to 0$ and $b_n \to 0$, we reach (c). □

Our algorithm converges like DR. Through a_n and b_n, it allows for inexact steps, and through λ_n, it allows for relaxation.

With $A = \mathrm{Id}$, $B = -\mathrm{Id}$, and $c = d = e = 0$, Theorem 8 reduces to Theorem 20 in [5]. In the next section, we will see how we can also recover several other algorithms.

5. RELATION TO ADMM

Our algorithm leads to the following version of ADMM:

Theorem 9. *In the context of Theorem 8, consider this procedure: Take any u_0 in $\mathcal{H}$ and any z_0 in $\mathcal{G}$, and for every n in $\mathbb{N}$, repeat*

$$y_n = \operatorname*{arg\,min}_{y \in \mathcal{F}} \Big(f(y) + \frac{1}{2\gamma} \|Ay + Bz_n - c + \gamma u_n\|^2 \Big) + a_n \,;$$

$$z_{n+1} = \operatorname*{arg\,min}_{z \in \mathcal{G}} \Big(g(z) + \frac{1}{2\gamma} \|Ay_n + Bz - c + \gamma u_n + (\lambda_n - 1)(Ay_n + Bz_n - c)\|^2 \Big) + b_{n+1};$$

$$u_{n+1} = u_n + \frac{1}{\gamma}\big(Ay_n + Bz_{n+1} - c + (\lambda_n - 1)(Ay_n + Bz_n - c)\big).$$

Then, $(y_n)_{n\in\mathbb{N}}$ and $(z_n)_{n\in\mathbb{N}}$ converge as in Theorem 8.

Proof. Our plan is to switch the order of y_n and z_n in Theorem 8 by using a new iterate u_n with carefully chosen x_0 and b_0.

We begin by defining

$$u_n = \frac{1}{\gamma}(x_n + Bz_n - e), \quad n \geq 1.$$

Then, having initialized u_0 and z_0, let $x_0 = \gamma u_0 - Bz_0 + e$ and $b_0 = z_0 - \operatorname{prox}_{\gamma g}^{-B}(x_0 - e)$. It follows that

$$\begin{aligned} y_n &= \operatorname{prox}_{\gamma f}^{A}\big(2(-Bz_n + e) - x_n + d\big) + a_n \\ &= \operatorname{prox}_{\gamma f}^{A}(-Bz_n + c - \gamma u_n) + a_n, \end{aligned}$$

and with

$$\begin{aligned} x_{n+1} &= x_n + \lambda_n(Ay_n + Bz_n - c) \\ &= \gamma u_n - Bz_n + e + \lambda_n(Ay_n + Bz_n - c), \end{aligned}$$

it follows that

$$\begin{aligned} z_{n+1} &= \operatorname{prox}_{\gamma g}^{-B}(x_{n+1} - e) + b_{n+1} \\ &= \operatorname{prox}_{\gamma g}^{-B}\big(\gamma u_n - Bz_n + \lambda_n(Ay_n + Bz_n - c)\big) + b_{n+1} \\ &= \operatorname{prox}_{\gamma g}^{-B}\big(Ay_n - c + \gamma u_n \\ &\qquad + (\lambda_n - 1)(Ay_n + Bz_n - c)\big) + b_{n+1}. \end{aligned}$$

Since

$$\begin{aligned} \gamma u_{n+1} - Bz_{n+1} + e &= x_{n+1} \\ &= \gamma u_n - Bz_n + e + \lambda_n(Ay_n + Bz_n - c), \end{aligned}$$

it follows that

$$\begin{aligned} u_{n+1} &= u_n + \frac{1}{\gamma}\big(\lambda_n(Ay_n + Bz_n - c) + B(z_{n+1} - z_n)\big) \\ &= u_n + \frac{1}{\gamma}\big(Ay_n + Bz_{n+1} - c \\ &\qquad + (\lambda_n - 1)(Ay_n + Bz_n - c)\big). \end{aligned}$$

Finally, by expressing the proximal points as optimization problems, we obtain the algorithm in the theorem. □

Theorem 9 corresponds to a version of ADMM that combines the techniques in [3, Secs. 3.4.3 and 3.4.4]. Note that the u_n update, unlike the x_n update in Theorem 8, requires memory: it depends on both z_n and z_{n+1}. This is the cost of relaxation in ADMM, and is an artifact of the ADMM formalism.

Remark. We saw in the proof of Theorem 9 that b_0, which appears nowhere in the description of the algorithm, depends on u_0 and z_0. Since a particular b_0 does not influence the requirement on $(b_n)_{n\in\mathbb{N}}$ (that is, (iii) in Theorem 8), we can disregard b_0.

We now use Theorem 9 to recover four versions of ADMM:

1) If we fix $a_n = 0$, $b_{n+1} = 0$, and $\lambda_n = 1$ for every n in $\mathbb{N}$, then we recover the version of Boyd et al. [3, eqs. (3.2)–(3.4)].
2) With $A = \mathrm{Id}$, $B = -\mathrm{Id}$, and $c = 0$, and a_n, b_{n+1}, and λ_n set as in 1), we reach the version of Parikh and Boyd [26, Sec. 4.4].
3) For the case where $B = -\mathrm{Id}$ and $c = 0$, we obtain the version of Eckstein and Bertsekas [17, Th. 8].
4) Choosing a_0 so that $y_0 = 0$, setting $z_0 = 0$, letting $\lambda_0 = 1$, and fixing $a_n = 0$ and $b_n = 0$ for $n \geq 1$, we arrive at the version of Davis and Yin [23, Algorithm 2].

We emphasize that these special cases enjoy both the context of Problem 1 and the convergence described in Theorem 8 and are therefore more straightforward and general than previously reported. In particular, validating the assumptions for convergence does not require a detour through duality, and convergence does not require finite dimensions or the injectivity of A.

6. CONCLUSION

In this paper, we have introduced a theory of proximity and used it to generalize DR, making it more attractive than ADMM, both in form and convergence. From this generalization, we recovered ADMM, imparting to it the primal context and general convergence result of our algorithm. We believe that our work shows that there is merit in breaking away from ADMM and paying closer attention to DR.

7. REFERENCES

[1] G. Mateos, J. A. Bazerque, and G. B. Giannakis, "Distributed sparse linear regression," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5262–5276, 2010.

[2] P. A. Forero, A. Cano, and G. B. Giannakis, "Consensus-based distributed support vector machines," *J. Mach. Learn. Res.*, vol. 11, pp. 1663–1707, 2010.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[4] D. Gabay, "Applications of the method of multipliers to variational inequalities," in *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*, M. Fortin and R. Glowinski, Eds., pp. 299–331. North-Holland, Amsterdam, 1983.

[5] P. L. Combettes and J.-C. Pesquet, "A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 564–574, 2007.

[6] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, Cham, Switzerland, 2nd edition, 2017.

[7] F. D. Côté, I. N. Psaromiligkos, and W. J. Gross, "In-network linear regression with arbitrarily split data matrices," in *2016 IEEE Global Conf. Signal and Information Processing*, Washington, DC, 2016, pp. 580–584.

[8] F. D. Côté, I. N. Psaromiligkos, and W. J. Gross, "A distributed constrained-form support vector machine," in *2017 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, New Orleans, 2017, pp. 4242–4246.

[9] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pp. 185–212. Springer, New York, 2011.

[10] M. Yan and W. Yin, "Self equivalence of the alternating direction method of multipliers," in *Splitting Methods in Communication, Imaging, Science, and Engineering*, R. Glowinski, S. J. Osher, and W. Yin, Eds., pp. 165–194. Springer, Cham, Switzerland, 2016.

[11] S. R. Becker and P. L. Combettes, "An algorithm for splitting parallel sums of linearly composed monotone operators, with applications to signal recovery," *J. Nonlinear Convex Anal.*, vol. 15, no. 1, pp. 137–159, 2014.

[12] M. Fortin and R. Glowinski, "On decomposition-coordination methods using an augmented Lagrangian," in *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*, M. Fortin and R. Glowinski, Eds., pp. 97–146. North-Holland, Amsterdam, 1983.

[13] J. Eckstein and W. Yao, "Understanding the convergence of the alternating direction method of multipliers: Theoretical and computational perspectives," *Pac. J. Optim.*, vol. 11, no. 4, pp. 619–644, 2015.

[14] J. Douglas and H. H. Rachford, "On the numerical solution of heat conduction problems in two and three space variables," *Trans. Am. Math. Soc.*, vol. 82, no. 2, pp. 421–439, 1956.

[15] R. S. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[16] P. L. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM J. Numer. Anal.*, vol. 16, no. 6, pp. 964–979, 1979.

[17] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Program.*, vol. 55, no. 1–3, pp. 293–318, 1992.

[18] P. L. Combettes, "Solving monotone inclusions via compositions of nonexpansive averaged operators," *Optimization*, vol. 53, no. 5–6, pp. 475–504, 2004.

[19] P. Patrinos, L. Stella, and A. Bemporad, "Douglas-Rachford splitting: Complexity estimates and accelerated variants," in *53rd IEEE Conf. Decision and Control*, Los Angeles, 2014, pp. 4234–4239.

[20] R. Glowinski and A. Marroco, "Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires," *Rev. Fr. Automat. Infor.*, vol. 9, no. R2, pp. 41–76, 1975.

[21] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximations," *Comput. Math. Appl.*, vol. 2, no. 1, pp. 17–40, 1976.

[22] R. Nishihara, L. Lessard, B. Recht, A. Packard, and M. I. Jordan, "A general analysis of the convergence of ADMM," in *Proc. 32nd Int. Conf. Machine Learning*, Lille, France, 2015, pp. 343–352.

[23] D. Davis and W. Yin, "Convergence rate analysis of several splitting schemes," in *Splitting Methods in Communication, Imaging, Science, and Engineering*, R. Glowinski, S. J. Osher, and W. Yin, Eds., pp. 115–163. Springer, Cham, Switzerland, 2016.

[24] F. Iutzeler and J. M. Hendrickx, "A generic online acceleration scheme for optimization algorithms via relaxation and inertia," arXiv:1603.05398, 2017.

[25] J.-J. Moreau, "Poximité et dualité dans un espace hilbertien," *Bull. Soc. Math. France*, vol. 93, pp. 273–299, 1965.

[26] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, 2014.