Projection onto Minkowski Sums with Application to Constrained Learning

Kenneth Lange [1] **Joong-Ho Won** [2] **Jason Xu** [3]

Abstract

We introduce block descent algorithms for projecting onto Minkowski sums of sets. Projection onto such sets is a crucial step in many statistical learning problems, and may regularize complexity of solutions to an optimization problem or arise in dual formulations of penalty methods. We show that projecting onto the Minkowski sum admits simple, efficient algorithms when complications such as overlapping constraints pose challenges to existing methods. We prove that our algorithm converges linearly when sets are strongly convex or satisfy an error bound condition, and extend the theory and methods to encompass non-convex sets as well. We demonstrate empirical advantages in runtime and accuracy over competitors in applications to $\ell_{1,p}$-regularized learning, constrained lasso, and overlapping group lasso.

1. Introduction

One of the most prevalent approaches to estimation, prediction, and inference problems arising in machine learning is to frame the task as an optimization problem, possibly subject to constraints. Examples of such constraints include rank, sparsity, positivity, or sum-to-zero constraints. For instance, consider the generic penalized estimation problem of minimizing a measure of fit plus a penalty term regularizing solution complexity. Regularizing the parameter vector $\boldsymbol{x}$ by way of an ℓ_p or $\ell_{1,p}$ norm can be viewed as a set constraint in the dual space. Projections onto the constraint sets appear naturally in many algorithms for these problems.

When multiple constraint sets $\{C_i\}_{i=1}^N$ are considered simultaneously, it is often straightforward to project onto their intersection. These feasibility-seeking problems are well studied (von Neumann, 1950; Dykstra, 1983; Bauschke & Borwein, 1993). In many tasks, however, the *Minkowski sum* is a more appropriate or useful operation to consider. The Minkowski sum $A+B$ of two sets A and B in Euclidean space $\mathbb{R}^d$ is defined to be the set $\{\boldsymbol{a}+\boldsymbol{b} : \boldsymbol{a} \in A,\, \boldsymbol{b} \in B\}$. Many penalized or constrained estimation problems involve a Minkowski sum of convex sets:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) + \sigma_{C_1+\cdots+C_k}(\boldsymbol{x}) = \min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) + \sum_{i=1}^k \sigma_{C_i}(\boldsymbol{x}), \tag{1}$$

where $\sigma_C(\boldsymbol{x}) = \sup_{\boldsymbol{y}\in C}\langle \boldsymbol{x}, \boldsymbol{y}\rangle$ is the support function of set C, and f is a (closed and proper) convex function. The equality above offers a more familiar form on the right-hand side and holds since support functions are additive over Minkowski sums (Hiriart-Urruty & Lemaréchal, 2012). If $C_1 + \cdots + C_k$ is closed, then $\sigma^*_{C_1+\cdots+C_k}(\boldsymbol{y}) = \iota_{C_1+\cdots+C_k}(\boldsymbol{y})$, where $g^*(\boldsymbol{y}) = \sup_{\boldsymbol{x}}\langle \boldsymbol{x}, \boldsymbol{y}\rangle - g(\boldsymbol{x})$ denotes the Fenchel conjugate of function g, and ι_S is the $0/\infty$ indicator function of set S. This relation suggests the immediate applications of Euclidean projection of a point in $\mathbb{R}^d$ onto $C_1 + \cdots + C_k$ featured below.

Multiple/overlapping norm penalties Suppose C_i has the form $C_i = \{\boldsymbol{y} = (\boldsymbol{y}_{i1}, \boldsymbol{y}_{i2}) : \|\boldsymbol{y}_{i1}\|_q \le \lambda,\ \boldsymbol{y}_{i2} = \boldsymbol{0}\}$, where $\boldsymbol{y}_{i1}$ is a vector of a predefined, coordinate-aligned subspace of $\mathbb{R}^d$ and $\boldsymbol{y}_{i2}$ is the residual coordinates filling in $\mathbb{R}^d$. If $q \ge 1$, then problem (1) corresponds to the well-known $\ell_{1,p}$ group lasso problem, where p is the conjugate exponent satisfying $1/p + 1/q = 1$ (Yuan & Lin, 2006):

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) + \lambda \sum_{i=1}^k \|\boldsymbol{x}_{i1}\|_p. \tag{2}$$

While the standard group lasso formulation assumes no overlap between groups, the relation between (1) and (2) holds whether or not the indices among C_i overlap. When overlap is present, problem (2) is no longer separable, leading to substantially more complicated routines (Yuan et al., 2011; Yang & Zou, 2015). Fortunately, given a method for projecting onto the Minkowski sum, straightforward solution methods remain unchanged regardless of overlap.

Conic constraints Let $K^* = \{\boldsymbol{y} : \langle \boldsymbol{x}, \boldsymbol{y}\rangle \le 0, \forall \boldsymbol{x} \in K\}$ denote the polar cone of a cone K. If C_i is a closed convex cone (e.g., a subspace), then $\sigma_{C_i}(\boldsymbol{x}) = \iota_{C_i^*}(\boldsymbol{x})$, so that problem (1) also includes conically constrained problems.

[1]University of California, Los Angeles [2]Department of Statistics, Seoul National University [3]Department of Statistical Science, Duke University. Correspondence to: Joong-Ho Won <wonj@stats.snu.ac.kr>, Jason Xu <jason.q.xu@duke.edu>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

This includes the constrained lasso (James et al., 2013):

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) + \lambda\|\boldsymbol{x}\|_1 \text{ subject to } \boldsymbol{B}\boldsymbol{x} = \boldsymbol{0},\ \boldsymbol{C}\boldsymbol{x} \le \boldsymbol{0}, \quad (3)$$

which subsumes the generalized lasso (Tibshirani & Taylor, 2011) as a special case (Gaines et al., 2018). Problem (3) is written as (1) with

$$\begin{aligned} C_1 &= \{\boldsymbol{x} : \boldsymbol{B}\boldsymbol{x} = \boldsymbol{0}\}^* = \{\boldsymbol{x} : \boldsymbol{B}\boldsymbol{x} = \boldsymbol{0}\}^\perp, \\ C_2 &= \{\boldsymbol{x} : \boldsymbol{C}\boldsymbol{x} \le \boldsymbol{0}\}^*, \quad C_3 = \{\boldsymbol{x} : \|\boldsymbol{x}\|_\infty \le \lambda\}, \end{aligned} \quad (4)$$

where $V^\perp$ is the orthogonal complement of subspace V.

Additional applications such as constraint relaxation are discussed in the Supplement (Sect A).

To illustrate the utility of Minkowski projection, consider the contemporary situation in which we seek to solve an $\ell_{1,p}$ (possibly overlapping) group lasso (2) or constrained lasso (3) problem using first-order methods such as proximal gradient descent (Combettes & Wajs, 2005). A core subroutine of these methods requires computing the proximity operator of $g = \sigma_{C_1+\cdots+C_k}$, that is,

$$\operatorname{prox}_{\gamma g}(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^d} \sigma_{C_1+\cdots+C_k}(\boldsymbol{u}) + \frac{1}{2\gamma}\|\boldsymbol{u}-\boldsymbol{x}\|_2^2$$

for $\gamma > 0$. Efficient computation of $\operatorname{prox}_{\gamma g}(\boldsymbol{x})$ is necessary in rendering these methods practical. The connection to projection onto the Minkowski sum begins with Moreau's decomposition

$$\boldsymbol{x} = \operatorname{prox}_{\gamma g}(\boldsymbol{x}) + \gamma \operatorname{prox}_{\gamma^{-1} g^*}(\gamma^{-1}\boldsymbol{x}), \quad (\text{M})$$

Because $C_1 + \cdots + C_k$ is closed for both problems (see Sect 2), we see that if a fast way to compute the Euclidean projection $P_{C_1+\cdots+C_k}(\boldsymbol{x})$ of $\boldsymbol{x}$ onto $C_1 + \cdots + C_k$ is given, then problem (2) or (3) can be solved efficiently since

$$P_{C_1+\cdots+C_k}(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{u}\in C_1+\cdots+C_k} \frac{\gamma}{2}\|\boldsymbol{u}-\boldsymbol{x}\|_2^2 = \operatorname{prox}_{\gamma^{-1} g^*}(\boldsymbol{x}). \quad (\text{P})$$

From this view, overlaps among C_i are automatically handled without distinction from the non-overlapping case of (2). In contrast, it is nontrivial to adapt algorithms for fitting the standard group lasso to the overlapping case.

Motivated by these classes of problems, the theme of this paper is to study an algorithm for solving (P), i.e., projecting a point $\boldsymbol{x}$ onto a Minkowski sum of closed sets, when the projections onto each set are computable. The exposition focuses on the case of two sets; we first study projecting an external point in the case that both sets are convex. For completeness, we also treat internal points in the Supplement (Sect C). The algorithm and results immediately generalize to sums of any finite collection of sets, and thus can be applied to solve problem (1). We prove that our algorithm converges linearly if either set is strongly convex. We then proceed to analyze projections onto possibly nonconvex sets (including convex but non-strongly convex sets). By adding viscosity penalties, a simple modification of the algorithm generates iterates that converge to a critical point for fairly general classes of sets. Importantly, linear convergence is attained globally for polyhedral sets. Finally, we showcase the merits of our algorithm (highlighted as Algorithm 1) in empirical studies with application to the constrained lasso and overlapping group lasso problems. Compared to recent methods tailored to each problem as well as highly optimized generic solvers, our method is not only more transparent and simple to code, but outperforms competitors in terms of both accuracy and runtime.

Algorithm 1 Projection onto a Minkowski sum of sets

Input: Projection operator P_{C_i} onto set C_i, $i = 1, \ldots, k$; initial value $\boldsymbol{a}_0^i$, $i = 1, \ldots, k$; viscosity parameter $\rho \ge 0$.
Initialization: $n \leftarrow 0$
repeat
  for $i = 1, 2, \ldots, k$ **do**
    $\boldsymbol{a}_{n+1}^{(i)} = P_{C_i}\Big(\frac{1}{1+\rho}\big(\boldsymbol{x} - \sum_{j=1}^{i-1}\boldsymbol{a}_{n+1}^{(j)} - \sum_{j=i+1}^{k}\boldsymbol{a}_n^{(j)}\big) + \frac{\rho}{1+\rho}\boldsymbol{a}_n^{(i)}\Big)$ // See §2 for $\rho = 0$; §3 for $\rho > 0$
  end for
  $n \leftarrow n + 1$
until convergence
Output: $\sum_{i=1}^{k}\boldsymbol{a}_N^{(i)}$

2. Minkowski Projections for Convex Sets

We begin by recalling several elementary properties of the Minkowski sum $A + B \equiv \{\boldsymbol{a} + \boldsymbol{b} : \boldsymbol{a} \in A,\ \boldsymbol{b} \in B\}$. It is easy to show that $A + B$ is convex whenever A and B are both convex and is closed if at least one of the two sets is compact and the other is closed. It is also closed when A and B are both polyhedral (intersections of half spaces). Indeed, any polyhedral set is closed, and the Minkowski-Weyl representation of a polyhedral set (Lange, 2013) implies that the Minkowski sum of two polyhedral sets is polyhedral.

Before studying how to project a point $\boldsymbol{x}$ onto $A + B$ given both convexity and closure, note that the two summands $\boldsymbol{a} \in A$ and $\boldsymbol{b} \in B$ in the representation $\boldsymbol{a} + \boldsymbol{b}$ of the closest point need not be unique. For instance, consider in dimension $d = 2$ the sets $A = B = [-1, 1] \times \{0\}$ and the external point $\boldsymbol{e}_2 = (0, 1)^T$. The closest point can be represented $\boldsymbol{0} = \binom{-u}{0} + \binom{u}{0} \in A + B$ for any $u \in [-1, 1]$. On the other hand, it is easy to verify that any valid pair $\boldsymbol{a}$ and $\boldsymbol{b}$ must be boundary points from A, B.

2.1. Block Descent Algorithm and Convergence

Suppose that the projection operators $P_A(\boldsymbol{a})$ and $P_B(\boldsymbol{b})$ are both known and $A + B$ is closed with A and B convex. In this case, we employ a block descent algorithm for finding the closest point to $\boldsymbol{x}$, which consists of alternating

$$\begin{aligned} \boldsymbol{b}_{n+1} &= P_B(\boldsymbol{x} - \boldsymbol{a}_n) \\ \boldsymbol{a}_{n+1} &= P_A(\boldsymbol{x} - \boldsymbol{b}_{n+1}). \end{aligned} \tag{5}$$

Our proof that the sequence $\boldsymbol{a}_n + \boldsymbol{b}_n$ converges to the closest point makes use of the notion of a paracontractive operator. A continuous operator $P(\boldsymbol{a})$ on $\mathbb{R}^d$ is paracontractive if for every fixed point $\tilde{\boldsymbol{a}}$ of $P(\boldsymbol{a})$, the inequality

$$\|P(\boldsymbol{a}) - \tilde{\boldsymbol{a}}\|_2 \quad < \quad \|\boldsymbol{a} - \tilde{\boldsymbol{a}}\|_2$$

holds unless $\boldsymbol{a}$ itself is a fixed point. The theorem of Elsner, Koltract, and Neumann (Elsner et al., 1992) states that whenever a paracontractive map $P(\boldsymbol{a})$ possesses one or more fixed points, then the sequence of iterates $\boldsymbol{a}_{n+1} = P(\boldsymbol{a}_n)$ converges to a fixed point regardless of the initial value $\boldsymbol{a}_0$ of the sequence. In our case, the relevant map is

$$P(\boldsymbol{a}) \quad = \quad P_A[\boldsymbol{x} - P_B(\boldsymbol{x} - \boldsymbol{a})],$$

corresponding to the $\boldsymbol{a}$ iterates of block descent.

Proposition 2.1. *The map $P(\boldsymbol{a})$ is nonexpansive and paracontractive.*

Proof. Because the set projections are nonexpansive,

$$\begin{aligned} &\|P_A[\boldsymbol{x} - P_B(\boldsymbol{x} - \boldsymbol{a})] - P_A[\boldsymbol{x} - P_B(\boldsymbol{x} - \tilde{\boldsymbol{a}})]\|_2 \\ &\leq \|P_B(\boldsymbol{x} - \boldsymbol{a}) - P_B(\boldsymbol{x} - \tilde{\boldsymbol{a}})\|_2 \leq \|\boldsymbol{a} - \tilde{\boldsymbol{a}}\|_2. \end{aligned} \tag{6}$$

This proves that $P(\boldsymbol{a})$ is nonexpansive. Now suppose $\tilde{\boldsymbol{a}}$ is a fixed point and equality holds throughout these inequalities. The standard proof that a convex set projection is paracontractive (Lange, 2013, pp. 389–390) indicates that equality is achieved in the previous two inequalities only if

$$\begin{aligned} &P_A[\boldsymbol{x} - P_B(\boldsymbol{x} - \boldsymbol{a})] - [\boldsymbol{x} - P_B(\boldsymbol{x} - \boldsymbol{a})] \\ &\quad = P_A[\boldsymbol{x} - P_B(\boldsymbol{x} - \tilde{\boldsymbol{a}})] - [\boldsymbol{x} - P_B(\boldsymbol{x} - \tilde{\boldsymbol{a}})], \\ \text{and} \quad &P_B(\boldsymbol{x} - \boldsymbol{a}) - (\boldsymbol{x} - \boldsymbol{a}) = P_B(\boldsymbol{x} - \tilde{\boldsymbol{a}}) - (\boldsymbol{x} - \tilde{\boldsymbol{a}}). \end{aligned}$$

Subtracting the second of these equalities from the first gives

$$P_A[\boldsymbol{x} - P_B(\boldsymbol{x} - \boldsymbol{a})] - \boldsymbol{a} = P_A[\boldsymbol{x} - P_B(\boldsymbol{x} - \tilde{\boldsymbol{a}})] - \tilde{\boldsymbol{a}} = \boldsymbol{0}.$$

It follows that equality in inequalities (6) is achieved only if $\boldsymbol{a}$ is also a fixed point. □

This paves the way for our next convergence proof:

Proposition 2.2. *Assuming both A and B are closed and convex, and $A + B$ is closed, the block descent iterates $\boldsymbol{a}_n + \boldsymbol{b}_n$ converge to the closest point in $A + B$ to the external point $\boldsymbol{x}$.*

Proof. It suffices to show that the map $P(\boldsymbol{a}) = P_A[\boldsymbol{x} - P_B(\boldsymbol{x} - \boldsymbol{a})]$ possesses a fixed point, and any fixed point furnishes a minimum of the convex function $f(\boldsymbol{a}, \boldsymbol{b}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{a} - \boldsymbol{b}\|_2^2$ on the set $A \times B$. Given the closedness of $A + B$, there exists a closest point $\tilde{\boldsymbol{a}} + \tilde{\boldsymbol{b}}$ to $\boldsymbol{x}$. Since block descent cannot improve $f(\boldsymbol{a}, \boldsymbol{b})$ starting from $(\tilde{\boldsymbol{a}}, \tilde{\boldsymbol{b}})$, it is clear that $\tilde{\boldsymbol{a}} = P(\tilde{\boldsymbol{a}})$. Now suppose $\tilde{\boldsymbol{a}}$ is any fixed point, and define $\tilde{\boldsymbol{b}} = P_B(\boldsymbol{x} - \tilde{\boldsymbol{a}})$. To prove that $\tilde{\boldsymbol{a}} + \tilde{\boldsymbol{b}}$ minimizes the distance to $\boldsymbol{x}$, it suffices to show that for every tangent vector $\boldsymbol{v} = \boldsymbol{a} + \boldsymbol{b} - \tilde{\boldsymbol{a}} - \tilde{\boldsymbol{b}}$ at $\tilde{\boldsymbol{a}} + \tilde{\boldsymbol{b}}$, the directional derivative

$$\begin{aligned} d_{\boldsymbol{v}} \frac{1}{2}\|\boldsymbol{x} - \tilde{\boldsymbol{a}} - \tilde{\boldsymbol{b}}\|_2^2 &= -(\boldsymbol{x} - \tilde{\boldsymbol{a}} - \tilde{\boldsymbol{b}})^T \boldsymbol{v} \\ &= -(\boldsymbol{x} - \tilde{\boldsymbol{a}} - \tilde{\boldsymbol{b}})^T (\boldsymbol{a} - \tilde{\boldsymbol{a}}) - (\boldsymbol{x} - \tilde{\boldsymbol{a}} - \tilde{\boldsymbol{b}})^T (\boldsymbol{b} - \tilde{\boldsymbol{b}}) \end{aligned}$$

is nonnegative. However, the inequalities $-(\boldsymbol{x} - \tilde{\boldsymbol{a}} - \tilde{\boldsymbol{b}})^T(\boldsymbol{a} - \tilde{\boldsymbol{a}}) \geq 0$ and $-(\boldsymbol{x} - \tilde{\boldsymbol{a}} - \tilde{\boldsymbol{b}})^T(\boldsymbol{b} - \tilde{\boldsymbol{b}}) \geq 0$ hold because $\tilde{\boldsymbol{a}}$ minimizes $\boldsymbol{a} \mapsto \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{a} - \tilde{\boldsymbol{b}}\|_2^2$ and $\tilde{\boldsymbol{b}}$ minimizes $\boldsymbol{b} \mapsto \frac{1}{2}\|\boldsymbol{x} - \tilde{\boldsymbol{a}} - \boldsymbol{b}\|_2^2$. □

For smooth sets, the local rate of convergence is determined by the dominant eigenvalue of the differential

$$\begin{aligned} dP(\boldsymbol{a}) &= dP_A[\boldsymbol{x} - P_B(\boldsymbol{x} - \boldsymbol{a})] dP_B(\boldsymbol{x} - \boldsymbol{a}) \\ &= dP_A(\boldsymbol{x} - \boldsymbol{b}) dP_B(\boldsymbol{x} - \boldsymbol{a}), \end{aligned}$$

at the fixed point $\boldsymbol{a} + \boldsymbol{b}$, provided these differentials exist. Existence is guaranteed if either set is *strongly convex*:

Definition 2.1. *A set $C \subset \mathbb{R}^d$ is* α-strongly convex *with respect to norm $\|\cdot\|$ if there is a constant $\alpha > 0$ such that for any $\boldsymbol{a}$ and $\boldsymbol{b}$ in C and any $\gamma \in [0, 1]$, C contains a ball of radius $r = \gamma(1-\gamma)\frac{\alpha}{2}\|\boldsymbol{a} - \boldsymbol{b}\|^2$ centered at $\gamma\boldsymbol{a} + (1-\gamma)\boldsymbol{b}$. In other words, for any unit vector $\boldsymbol{z} \in \mathbb{R}^d$, we have*

$$\gamma\boldsymbol{a} + (1-\gamma)\boldsymbol{b} + \gamma(1-\gamma)\frac{\alpha}{2}\|\boldsymbol{a} - \boldsymbol{b}\|^2 \boldsymbol{z} \in C. \tag{7}$$

For instance, an ℓ_p norm ball of radius r for $1 < p \leq 2$ is $(p-1)/r$-strongly convex with respect to $\|\cdot\|_p$, which in turn is $(p-1)d^{1/2-1/p}/r$-strongly convex with respect to $\|\cdot\|_2$ (Garber & Hazan, 2015). We now show that when C is strongly convex, projection onto C is locally strictly contractive. A similar result is presented by Balashov & Golubev (2012).

Lemma 2.1. *If $C \in \mathbb{R}^d$ is α-strongly convex with respect to $\|\cdot\|_2$, then on the complement of C the projection operator $P_C(\boldsymbol{x})$ is a strict contraction. In particular, for $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d \setminus C$, we have*

$$\|P_C(\boldsymbol{a}) - P_C(\boldsymbol{b})\|_2^2 \quad \leq \quad \kappa\|\boldsymbol{a} - \boldsymbol{b}\|_2^2,$$

where $\kappa = \dfrac{1}{1 + \frac{\alpha}{4}\|\boldsymbol{a} - P_C(\boldsymbol{a})\|_2 + \frac{\alpha}{4}\|\boldsymbol{b} - P_C(\boldsymbol{b})\|_2} < 1.$

Proof. Consider the midpoint of the line segment between $P_C(\boldsymbol{a})$ and $P_C(\boldsymbol{b})$ corresponding to the choice $\gamma = \frac{1}{2}$ in

condition (7). Given the strong convexity of C, the obtuse angle property of convex set projection implies

$$\Big\langle \bar{\boldsymbol{a}} + \frac{\alpha}{8}\|P_C(\boldsymbol{a}) - P_C(\boldsymbol{b})\|^2 \boldsymbol{z}_a, \boldsymbol{a} - P_C(\boldsymbol{a}) \Big\rangle \le 0$$
$$\Big\langle \bar{\boldsymbol{b}} + \frac{\alpha}{8}\|P_C(\boldsymbol{a}) - P_C(\boldsymbol{b})\|^2 \boldsymbol{z}_b, \boldsymbol{b} - P_C(\boldsymbol{b}) \Big\rangle \le 0,$$

where $\bar{\boldsymbol{a}} = \frac{1}{2}P_C(\boldsymbol{a}) + \frac{1}{2}P_C(\boldsymbol{b}) - P_C(\boldsymbol{a})$ and $\bar{\boldsymbol{b}} = \frac{1}{2}P_C(\boldsymbol{a}) + \frac{1}{2}P_C(\boldsymbol{b}) - P_C(\boldsymbol{b})$. Adding these two inequalities, putting $\boldsymbol{z}_a = \frac{\boldsymbol{a} - P_C(\boldsymbol{a})}{\|\boldsymbol{a} - P_C(\boldsymbol{a})\|_2}$ and $\boldsymbol{z}_b = \frac{\boldsymbol{b} - P_C(\boldsymbol{b})}{\|\boldsymbol{b} - P_C(\boldsymbol{b})\|_2}$, and rearranging produce

$$\|P_C(\boldsymbol{a}) - P_C(\boldsymbol{b})\|_2^2 \Big[1 + \frac{\alpha}{4}\|\boldsymbol{a} - P_C(\boldsymbol{a})\|_2 + \frac{\alpha}{4}\|\boldsymbol{b} - P_C(\boldsymbol{b})\|_2\Big] \le \Big\langle P_C(\boldsymbol{a}) - P_C(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \Big\rangle.$$

Replacing the left-hand side by its Cauchy-Schwarz upper bound, then we find that

$$\|P_C(\boldsymbol{a}) - P_C(\boldsymbol{b})\|_2^2 \Big[1 + \frac{\alpha}{4}\|\boldsymbol{a} - P_C(\boldsymbol{a})\|_2 + \frac{\alpha}{4}\|\boldsymbol{b} - P_C(\boldsymbol{b})\|_2\Big] \le \|P_C(\boldsymbol{a}) - P_C(\boldsymbol{b})\|_2 \cdot \|\boldsymbol{a} - \boldsymbol{b}\|_2.$$

A simple cross division completes the proof. □

We are now ready to establish the linear convergence rate.

Theorem 2.1. *For any exterior point, the block descent algorithm (5) converges at a linear rate if either of the sets A and B is strongly convex with respect to $\|\cdot\|_2$.*

Proof. First note that $\boldsymbol{x} \notin A + B$ implies that $\boldsymbol{x} - \boldsymbol{a} \notin B$ and $\boldsymbol{x} - \boldsymbol{b} \notin A$ whenever $\boldsymbol{a} \in A$ and $\boldsymbol{b} \in B$. Given that the algorithm converges, denote its limiting values by $\lim_{n\to\infty} \boldsymbol{a}_n = \boldsymbol{a}^*$ and $\lim_{n\to\infty} \boldsymbol{b}_n = \boldsymbol{b}^*$. In view of the nonexpansiveness of the two projection operators,

$$\begin{aligned}
&\|P_A[\boldsymbol{x} - P_B(\boldsymbol{x} - \boldsymbol{a}_n)] - \boldsymbol{a}^*\|_2 \\
&= \|P_A[\boldsymbol{x} - \overbrace{P_B(\boldsymbol{x} - \boldsymbol{a}_n)}^{\boldsymbol{b}_{n+1}}] - P_A[\boldsymbol{x} - \overbrace{P_B(\boldsymbol{x} - \boldsymbol{a}^*)}^{\boldsymbol{b}^*}]\|_2 \\
&\le \kappa_1 \|\boldsymbol{x} - P_B(\boldsymbol{x} - \boldsymbol{a}_n) - \boldsymbol{x} + P_B(\boldsymbol{x} - \boldsymbol{a}^*)\|_2 \\
&\le \kappa_1 \|P_B(\boldsymbol{x} - \boldsymbol{a}^*) - P_B(\boldsymbol{x} - \boldsymbol{a}_n)\|_2 \\
&\le \kappa_1 \kappa_2 \|\boldsymbol{a}_n - \boldsymbol{a}^*\|_2
\end{aligned}$$

for κ_1 and κ_2 belonging to $[0, 1]$. In a neighborhood of the limit, Lemma 2.1 indicates that $\kappa_1 < 1$ (or $\kappa_2 < 1$) whenever A (or B) is a strongly convex set. Furthermore, the inequalities hold uniformly, and a linear rate of convergence is achieved. □

3. Modification for Possibly Nonconvex Sets

For closed, nonconvex sets A and B, establishing convergence is more delicate. (This includes convergence rate analysis for convex but not strongly convex sets, e.g., polyhedra.) To this end, we assume either A or B is compact and equivalently pose the projection problem as minimizing the objective

$$\phi(\boldsymbol{a}, \boldsymbol{b}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{a} - \boldsymbol{b}\|_2^2 + \iota_A(\boldsymbol{a}) + \iota_B(\boldsymbol{b}). \tag{8}$$

Toward defining the update of $\boldsymbol{b}$, it will be useful to define a surrogate function which *majorizes* or lies above $\phi(\boldsymbol{a}, \boldsymbol{b})$ via adding a viscosity penalty of the form $\frac{\rho}{2}\|\boldsymbol{b} - \boldsymbol{b}_n\|_2^2$ with $\rho > 0$. In view of the identity

$$\begin{aligned}
&\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{a}_n - \boldsymbol{b}\|_2^2 + \frac{\rho}{2}\|\boldsymbol{b} - \boldsymbol{b}_n\|_2^2 \\
&= \frac{1+\rho}{2}\Big\|\frac{1}{1+\rho}(\boldsymbol{x} - \boldsymbol{a}_n) + \frac{\rho}{1+\rho}\boldsymbol{b}_n - \boldsymbol{b}\Big\|_2^2 + c_n,
\end{aligned}$$

where c_n is an irrelevant constant that depends only on $\boldsymbol{b}_n$, we derive a viscosity-modified update by minimizing this surrogate. The resulting majorization-minimization (MM) update is given by any instance of the set-valued map

$$\boldsymbol{b}_{n+1} \in P_B\Big[\frac{1}{1+\rho}(\boldsymbol{x} - \boldsymbol{a}_n) + \frac{\rho}{1+\rho}\boldsymbol{b}_n\Big]. \tag{9}$$

By the MM principle (Lange, 2016), this update is guaranteed to decrease the original objective $\phi(\boldsymbol{a}, \boldsymbol{b})$; more details on MM are included in the Supplement (Sect D). Likewise, we include the analogous non-negative penalty term $\frac{\rho}{2}\|\boldsymbol{a} - \boldsymbol{a}_n\|_2^2$ to the objective in updating $\boldsymbol{a}$, yielding

$$\boldsymbol{a}_{n+1} \in P_A\Big[\frac{1}{1+\rho}(\boldsymbol{x} - \boldsymbol{b}_{n+1}) + \frac{\rho}{1+\rho}\boldsymbol{a}_n\Big]. \tag{10}$$

Because the norm $\|\boldsymbol{x} - \boldsymbol{a} - \boldsymbol{b}\|_2$ consistently decreases and at least one of the sets A, B is compact, the sequences $\boldsymbol{a}_n$ and $\boldsymbol{b}_n$ are bounded, and the sequence $\phi(\boldsymbol{a}_n, \boldsymbol{b}_n)$ possesses a finite limit $\bar{\phi}$. The next lemma formalizes the value of adding the viscosity penalties.

Lemma 3.1. *The modified block descent steps (9),(10) satisfy*

$$\begin{aligned}
&\|\boldsymbol{a}_{n+1} - \boldsymbol{a}_n\|_2^2 + \|\boldsymbol{b}_{n+1} - \boldsymbol{b}_n\|_2^2 \\
&\le \frac{2}{\rho}\Big[\phi(\boldsymbol{a}_n, \boldsymbol{b}_n) - \phi(\boldsymbol{a}_{n+1}, \boldsymbol{b}_{n+1})\Big].
\end{aligned} \tag{11}$$

Hence, $\lim_{n\to\infty} \|\boldsymbol{b}_{n+1} - \boldsymbol{b}_n\|_2 = 0$ and $\lim_{n\to\infty} \|\boldsymbol{a}_{n+1} - \boldsymbol{a}_n\|_2 = 0$. If $\phi(\boldsymbol{a}_n, \boldsymbol{b}_n) = \bar{\phi}$ for any n, then $(\boldsymbol{a}_m, \boldsymbol{b}_m) = (\boldsymbol{a}_n, \boldsymbol{b}_n)$ for all $m \ge n$.

Proof. The inequality (due to the MM principle)

$$\phi(\boldsymbol{a}_n, \boldsymbol{b}_{n+1}) + \frac{\rho}{2}\|\boldsymbol{b}_{n+1} - \boldsymbol{b}_n\|_2^2 \le \phi(\boldsymbol{a}_n, \boldsymbol{b}_n)$$

implies $\|\boldsymbol{b}_{n+1} - \boldsymbol{b}_n\|_2^2 \le \frac{2}{\rho}\Big[\phi(\boldsymbol{a}_n, \boldsymbol{b}_n) - \phi(\boldsymbol{a}_n, \boldsymbol{b}_{n+1})\Big]$. A similar inequality exists for the $\boldsymbol{a}$ update, and adding these

gives inequality (11). Since the difference in objective values on the right of inequality (11) tends to 0, the stated limits follow. The second assertion is an obvious consequence of these considerations. □

To establish convergence of this modified algorithm, we must define a few additional concepts which are explored in more detail in the references (Attouch et al., 2010; Kruger, 2003; Rockafellar & Wets, 2009).

Definition 3.1. *((Limiting) Fréchet subdifferential). A vector $\boldsymbol{g} \in \mathbb{R}^d$ is a Fréchet subgradient of a lower semicontinuous function ψ at the point $\boldsymbol{x} \in \mathbf{dom}(\psi)$ if*

$$\liminf_{\boldsymbol{y} \to \boldsymbol{x}} \frac{\psi(\boldsymbol{y}) - \psi(\boldsymbol{x}) - \langle \boldsymbol{g}, \boldsymbol{y} - \boldsymbol{x} \rangle}{\|\boldsymbol{y} - \boldsymbol{x}\|} \geq 0$$

for some norm $\|\cdot\|$. The set $\partial^F \psi(\boldsymbol{x})$ of Fréchet subgradients of ψ at $\boldsymbol{x}$ is called the Fréchet subdifferential. For any $\boldsymbol{x} \notin \mathbf{dom}(\psi)$, $\partial^F \psi(\boldsymbol{x}) = \emptyset$. The limiting Fréchet subdifferential, or simply subdifferential, is defined by $\partial \psi(\boldsymbol{x}) = \{g : \exists \boldsymbol{x}_n \to \boldsymbol{x}, \psi(\boldsymbol{x}_n) \to \psi(\boldsymbol{x}), \boldsymbol{g}_n \in \partial^F \psi(\boldsymbol{x}_n)$ and $\boldsymbol{g}_n \to \boldsymbol{g}\}$. The set $\partial \psi(\boldsymbol{x})$ is closed, convex, and possibly empty. If ψ is convex, then $\partial \psi(\boldsymbol{x})$ reduces to its convex subdifferential. If ψ is differentiable, then $\partial \psi(\boldsymbol{x})$ reduces to its ordinary differential.

The domain $\mathbf{dom}(\psi)$ of an extended real-valued function ψ and properness thereof are defined in the Supplement (Def. B.2). The following Global Convergence Theorem (Zangwill, 1969) establishes subsequence convergence.

Lemma 3.2. *Every limit point $(\tilde{\boldsymbol{a}}, \tilde{\boldsymbol{b}})$ of the modified block descent iterates (9) and (10) is a critical point of the objective $\phi(\boldsymbol{a}, \boldsymbol{b}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{a} - \boldsymbol{b}\|_2^2 + \iota_A(\boldsymbol{a}) + \iota_B(\boldsymbol{b})$. Furthermore, $\lim_{n \to \infty} \operatorname{dist}(\boldsymbol{a}_n + \boldsymbol{b}_n, W) = 0$, where W denotes the set of limit points and $\operatorname{dist}(\boldsymbol{x}, W) = \inf\{\|\boldsymbol{x} - \boldsymbol{y}\|_2 : \boldsymbol{y} \in W\}$. The set W is compact and connected.*

Proof. Fermat's principle implies

$$\mathbf{0} = -(\boldsymbol{x} - \boldsymbol{a}_{n-1} - \boldsymbol{b}_n) + \rho(\boldsymbol{b}_n - \boldsymbol{b}_{n-1}) + \boldsymbol{u}_n \quad (12)$$

$$\mathbf{0} = -(\boldsymbol{x} - \boldsymbol{a}_n - \boldsymbol{b}_n) + \rho(\boldsymbol{a}_n - \boldsymbol{a}_{n-1}) + \boldsymbol{v}_n, \quad (13)$$

where $\boldsymbol{u}_n \in \partial^F \iota_B(\boldsymbol{b}_n)$ and $\boldsymbol{v}_n \in \partial^F \iota_A(\boldsymbol{a}_n)$. If we take limits along a convergent subsequence $(\boldsymbol{a}_{n_m}, \boldsymbol{b}_{n_m}) \to (\tilde{\boldsymbol{a}}, \tilde{\boldsymbol{b}})$ and apply Lemma 3.1, then we can conclude that

$$\mathbf{0} = -(\boldsymbol{x} - \tilde{\boldsymbol{a}} - \tilde{\boldsymbol{b}}) + \tilde{\boldsymbol{u}}$$
$$\mathbf{0} = -(\boldsymbol{x} - \tilde{\boldsymbol{a}} - \tilde{\boldsymbol{b}}) + \tilde{\boldsymbol{v}},$$

where $\tilde{\boldsymbol{u}} = \lim_{m \to \infty} \boldsymbol{u}_{n_m}$ and $\tilde{\boldsymbol{v}} = \lim_{m \to \infty} \boldsymbol{v}_{n_m}$ necessarily exist. Because $\iota_B(\boldsymbol{b}_n) = \iota_A(\boldsymbol{a}_n) = 0$ for all n, and A and B are closed, the definition of the limiting subdifferential implies $\tilde{\boldsymbol{u}} \in \partial \iota_B(\tilde{\boldsymbol{b}})$ and $\tilde{\boldsymbol{v}} \in \partial \iota_A(\tilde{\boldsymbol{a}})$. In view of the the Cartesian product formula

$$\begin{aligned} \partial \phi(\boldsymbol{a}, \boldsymbol{b}) = &\{-(\boldsymbol{x} - \boldsymbol{a} - \boldsymbol{b}) + \partial \iota_A(\boldsymbol{a})\} \\ &\times \{-(\boldsymbol{x} - \boldsymbol{a} - \boldsymbol{b}) + \partial \iota_B(\boldsymbol{b})\} \end{aligned} \quad (14)$$

from Proposition 3 of (Attouch et al., 2010), it follows that $(\mathbf{0}, \mathbf{0}) \in \partial \phi(\tilde{\boldsymbol{a}}, \tilde{\boldsymbol{b}})$.

To prove the second assertion, note that W is nonempty because the sequence $\boldsymbol{a}_n + \boldsymbol{b}_n$ is bounded. If $\lim_{n \to \infty} \operatorname{dist}(\boldsymbol{a}_n + \boldsymbol{b}_n, W) = 0$ fails, then there exists an $\epsilon > 0$ such that $\operatorname{dist}(\boldsymbol{a}_n + \boldsymbol{b}_n, W) \geq \epsilon$ for infinitely many n. The subsequence defined by this condition has a convergent subsubsequence whose limit falls outside W, contradicting the definition of W. The compactness and connectedness of W follow from Proposition 7.3.4 of (Lange, 2016). □

Finally, we will need to invoke the Kurdyka-Łojasiewicz (KL) inequality (Bierstone & Milman, 1988; Bochnak et al., 2013), which applies to all subanalytic functions, to show that the whole sequence converges. For exposition we focus on the subclass of semialgebraic functions. The class of semialgebraic subsets of $\mathbb{R}^d$ is the smallest class that: *a)* contains all sets of the form $\{\boldsymbol{x} : q(\boldsymbol{x}) > 0\}$ for a polynomial $q(\boldsymbol{x})$ in p variables; *b)* is closed under the formation of finite unions, finite intersections, and set complementation. A function $a : \mathbb{R}^d \mapsto \mathbb{R}^r$ is said to be semialgebraic if its graph is a semialgebraic set of $\mathbb{R}^{p+r}$. The class of real-valued semialgebraic functions contains all polynomials $p(\boldsymbol{x})$ and all $0/\infty$ indicators of algebraic sets. It is closed under the formation of sums, products, absolute values, reciprocals when $a(\boldsymbol{x}) \neq 0$, nth roots when $a(\boldsymbol{x}) \geq 0$, and maxima $\max\{a(\boldsymbol{x}), b(\boldsymbol{x})\}$ and minima $\min\{a(\boldsymbol{x}), b(\boldsymbol{x})\}$. For our purposes, it is important to note that the Euclidean distance $\operatorname{dist}(\boldsymbol{x}, S)$ from vector $\boldsymbol{x}$ to a set S is a semialgebraic function whenever S is a semialgebraic set.

Definition 3.2. *(Kurdyka-Łojasiewicz inequality (Bolte et al., 2007)) Let $\psi(\boldsymbol{x})$ be a closed (lower semicontinuous) and subanalytic function with a closed domain. If $\boldsymbol{y}$ is a critical point of $\psi(\boldsymbol{x})$, i.e., $\mathbf{0} \in \partial \psi(\boldsymbol{y})$, then*

$$|\psi(\boldsymbol{x}) - \psi(\boldsymbol{y})|^\theta \leq c\|\boldsymbol{v}\| \quad (15)$$

for the same norm $\|\cdot\|$ as in Definition 3.1, all $\boldsymbol{x} \in B_r(\boldsymbol{y}) \cap \{\tilde{\boldsymbol{x}} : \partial \psi(\tilde{\boldsymbol{x}}) \neq \emptyset\}$ satisfying $\psi(\boldsymbol{x}) > \psi(\boldsymbol{y})$ and all $\boldsymbol{v}$ in $\partial \psi(\boldsymbol{x})$. Here the exponent $\theta \in [0, 1)$, the radius $r > 0$, and the constant $c \geq 0$ depend on $\boldsymbol{y}$.

We now apply the KL inequality to characterize the limit points of our viscosity modified block descent algorithm (Attouch et al., 2010; Bolte et al., 2007).

Theorem 3.1. *Assume the two closed sets A and B are subanalytic and at least one of them is bounded. Then the modified block descent iterates (9) and (10) converge to a critical point of the objective function $\phi(\boldsymbol{a}, \boldsymbol{b})$ regardless of their initial values.*

Proof. Lemmas 3.1 and 3.2 together imply the result if $\phi(\boldsymbol{a}_n, \boldsymbol{b}_n) = \bar{\phi}$ for any n, so assume the contrary. The optimality conditions (12) and (13) identify a vector $\boldsymbol{u}_n \in \partial^F \iota_B(\boldsymbol{b}_n) \subset \partial \iota_B(\boldsymbol{b}_n)$ and a vector $\boldsymbol{v}_n \in \partial^F \iota_A(\boldsymbol{a}_n) \subset \partial \iota_A(\boldsymbol{b}_n)$. Hence, the product formula (14) identifies the limiting subgradient

$$\begin{aligned}\boldsymbol{w}_n &= \begin{pmatrix} -(\boldsymbol{x} - \boldsymbol{a}_n - \boldsymbol{b}_n) + \boldsymbol{0} + (\boldsymbol{x} - \boldsymbol{a}_{n-1} - \boldsymbol{b}_n) \\ -(\boldsymbol{x} - \boldsymbol{a}_n - \boldsymbol{b}_n) + \boldsymbol{0} + (\boldsymbol{x} - \boldsymbol{a}_n - \boldsymbol{b}_n) \end{pmatrix} \\ &\quad - \rho \begin{pmatrix} \boldsymbol{b}_n - \boldsymbol{b}_{n-1} \\ \boldsymbol{a}_n - \boldsymbol{a}_{n-1} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{a}_n - \boldsymbol{a}_{n-1} - \rho(\boldsymbol{b}_n - \boldsymbol{b}_{n-1}) \\ -\rho(\boldsymbol{a}_n - \boldsymbol{a}_{n-1}) \end{pmatrix}.\end{aligned}$$

For this choice of $\boldsymbol{w}_n \in \partial\phi(\boldsymbol{a}_n, \boldsymbol{b}_n)$, we have

$$\begin{aligned}\|\boldsymbol{w}_n\| &\leq \left\| \begin{pmatrix} \boldsymbol{a}_n - \boldsymbol{a}_{n-1} \\ \boldsymbol{0} \end{pmatrix} \right\| + \rho \left\| \begin{pmatrix} \boldsymbol{b}_n - \boldsymbol{b}_{n-1} \\ \boldsymbol{a}_n - \boldsymbol{a}_{n-1} \end{pmatrix} \right\| \\ &\leq (1+\rho) \left\| \begin{pmatrix} \boldsymbol{a}_n - \boldsymbol{a}_{n-1} \\ \boldsymbol{b}_n - \boldsymbol{b}_{n-1} \end{pmatrix} \right\| \end{aligned} \tag{16}$$

for the norm $\|\cdot\|$ used in the KL inequality (15). We now set $\boldsymbol{y} = (\boldsymbol{a}, \boldsymbol{b})$ and consider the subanalytic function $\phi(\boldsymbol{y}) - \bar{\phi}$. According to the KL inequality, around each limit point $\boldsymbol{z} \in W$ of the algorithm there exists an open ball $B_{r(\boldsymbol{z})}(\boldsymbol{z})$ around $\boldsymbol{z}$ and an exponent $\theta(\boldsymbol{z}) \in [0, 1)$ such that

$$|\phi(\boldsymbol{y}) - \phi(\boldsymbol{z})|^{\theta(\boldsymbol{z})} = |\phi(\boldsymbol{y}) - \bar{\phi} - \phi(\boldsymbol{z}) + \bar{\phi}|^{\theta(\boldsymbol{z})} \leq c(\boldsymbol{z})\|\boldsymbol{w}\| \tag{17}$$

for all $\boldsymbol{y} \in B_{r(\boldsymbol{z})}(\boldsymbol{z})$ and all $\boldsymbol{w} \in \partial(\phi - \bar{\phi})(\boldsymbol{y}) = \partial(\phi)(\boldsymbol{y})$. We will apply this inequality to the sequence $\boldsymbol{y}_n = (\boldsymbol{a}_n, \boldsymbol{b}_n)$ and the limiting subgradients $\boldsymbol{w}_n$ identified above. In so doing, we would like to assume that the exponent $\theta(\boldsymbol{z})$ and constant $c(\boldsymbol{z})$ do not depend on $\boldsymbol{z}$. With this end in mind, cover the compact set W by a finite number of balls $B_{r(\boldsymbol{z}_i)}(\boldsymbol{z}_i)$ and take $\theta = \max_i \theta(\boldsymbol{z}_i) < 1$ and $c = \max_i c(\boldsymbol{z}_i)$. For a sufficiently large N, every $\boldsymbol{y}_n$ with $n \geq N$ falls within one of these balls and satisfies $|\phi(\boldsymbol{y}_n) - \bar{\phi}| < 1$. Without loss of generality assume $N = 0$. In combination with the concavity of the function $t^{1-\theta}$ on $[0, \infty)$, inequalities (11), (16), and (17) imply

$$\begin{aligned}&[\phi(\boldsymbol{y}_{n-1}) - \bar{\phi}]^{1-\theta} - [\phi(\boldsymbol{y}_n) - \bar{\phi}]^{1-\theta} \\ &\quad \geq \frac{1-\theta}{[\phi(\boldsymbol{y}_n) - \bar{\phi}]^{\theta}} [\phi(\boldsymbol{y}_{n-1}) - \phi(\boldsymbol{y}_n)] \\ &\quad \geq \frac{1-\theta}{c\|\boldsymbol{w}_n\|} \frac{\rho}{2} \|\boldsymbol{y}_n - \boldsymbol{y}_{n-1}\|_2^2 \geq \frac{(1-\theta)\rho}{2C(1+\rho)} \|\boldsymbol{y}_n - \boldsymbol{y}_{n-1}\|,\end{aligned}$$

where $C > 0$ is a scaled version of c due to the equivalence of norms in $\mathbb{R}^d$. Rearranging and summing over n yield

$$\sum_{n=1}^{\infty} \|\boldsymbol{y}_n - \boldsymbol{y}_{n-1}\| \quad \leq \quad \frac{2C(1+\rho)}{(1-\theta)\rho} [\phi(\boldsymbol{y}_0) - \bar{\phi}]^{1-\theta}.$$

Thus, the sequence $\boldsymbol{y}_n$ is a fast Cauchy sequence and converges to a unique limit in W. □

Having established convergence of the viscosity-modified algorithm even for non-convex sets, we note that the rate of convergence will depend on the value of the exponent θ appearing in (15). A prototypical result on the local convergence rate of first-order algorithms (including ours) via the KL inequality takes the following form (Li & Pong, 2018):

1. If $\theta = 0$, then $\{\boldsymbol{x}_k\}$ converges finitely.
2. If $\theta \in (0, \frac{1}{2}]$, then $\{\boldsymbol{x}_k\}$ converges locally linearly.
3. If $\theta \in (\frac{1}{2}, 1)$, then $\{\boldsymbol{x}_k\}$ converges locally sublinearly.

Fortunately, recent work provides a useful result for determining θ via an error bound condition:

Definition 3.3 (Luo-Tseng Error Bound)**.** *Let $\phi(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x})$, where f is proper and closed with Lipschitz continuous gradient on its open domain, and g is proper, closed and convex. Let $W \neq 0$ be the set of stationary points of ϕ. We say the Luo-Tseng Error Bound holds if for any $\eta \geq \inf \phi$, there exist $\kappa, \epsilon > 0$ such that*

$$\operatorname{dist}(\boldsymbol{x}, W) \leq \kappa \|\operatorname{prox}_g[\boldsymbol{x} - \nabla f(\boldsymbol{x})] - \boldsymbol{x}\| \tag{18}$$

whenever $\|\operatorname{prox}_g[\boldsymbol{x} - \nabla f(\boldsymbol{x})] - \boldsymbol{x}\| < \epsilon$ and $\phi(\boldsymbol{x}) < \eta$.

Theorem 3.2 ((Li & Pong, 2018))**.** *Let $W \neq \emptyset$ be the set of stationary points of ϕ. Suppose that for any $\tilde{\boldsymbol{x}} \in W$, there exists $\delta > 0$ so that $\phi(\boldsymbol{y}) = \phi(\tilde{\boldsymbol{x}})$ whenever $\boldsymbol{y} \in W$ and $\|\boldsymbol{y} - \tilde{\boldsymbol{x}}\| < \delta$. Further assume the Luo-Tseng Error Bound holds. Then ϕ satisfies KL with exponent $\theta = 1/2$.*

Theorem 3.2 provides a verifiable sufficient condition for determining whether the viscosity-modified algorithm attains a linear rate; the assumption that ϕ is locally constant within the set of stationary points is satisfied, e.g., if f is convex. Indeed, we see that the objective in (8) for projection onto a Minkowski sum has the form $\phi(\boldsymbol{a}, \boldsymbol{b}) = f(\boldsymbol{a}, \boldsymbol{b}) + g(\boldsymbol{a}, \boldsymbol{b})$ with $f(\boldsymbol{a}, \boldsymbol{b}) = h(\boldsymbol{M}(\boldsymbol{a}, \boldsymbol{b}))$, where h is a strongly convex function, $\boldsymbol{M}$ is a linear map, and g is the sum of set indicators. In fact, for such a function ϕ the Luo-Tseng inequality (18) holds *uniformly* whenever the constraint set is polyhedral (Karimi et al., 2018). In such cases, convergence of the viscosity-modified algorithm is *globally* linear. This observation has a direct consequence to the $\ell_{1,\infty}$-(overlapping) group lasso, as shown in Sect 5.1.

4. Alternative Algorithms

There are other ways problem (P) could be tackled, especially when the summand sets are convex. Splitting methods such as the ADMM (Boyd et al., 2010) or Davis-Yin three-operator splitting (Davis & Yin, 2017) can be considered. However, we do not know whether these methods can achieve a linear convergence rate under strong convexity of a summand set as Algorithm 1 does. While the Davis-Yin method enjoys a linear rate if one objective term is strongly

convex and another smooth, the smooth term in objective (8) is not strongly convex; the indicator of a strongly convex set is neither smooth nor strongly convex. Sublinear rates for non-strongly convex sets can be achieved with our viscosity algorithm as well. Further, ADMM and Davis-Yin do not produce descent algorithms, and introduce additional variables as well as intermediate steps. In recent related work, Qin & An (2018) considered projection onto a Minkowski sum of affine transforms of convex and compact sets. Their approach differs in focus from ours, and appeals to a smooth approximation to the dual, yielding a slower than $O(1/n^2)$ rate. In practical settings (Qin & An, 2018, Table 1), oddly it is slower than the base (Gilbert's) algorithm. We expect our simpler primal algorithm to be faster.

5. Applications and Empirical Results

5.1. $\ell_{1,p}$-Overlapping Group Lasso

Including an $\ell_{1,p}$ penalty with the objective is a flexible and popular approach to impose structured sparsity. Recall this takes the general form (1) with ℓ_q-norm disks

$$C_i = \{\boldsymbol{y} \in V_i : \|\boldsymbol{y}\|_q \leq \lambda\}, \quad (1/p + 1/q = 1),$$

where $V_i = \{\boldsymbol{y} = (\boldsymbol{y}_{i1}, \boldsymbol{y}_{i2}) \in \mathbb{R}^d : \boldsymbol{y}_{i2} = \boldsymbol{0}\}$ for an appropriate coordinate-aligned splitting of vector $\boldsymbol{y}$. The resulting optimization problem is typically solved via first-order methods. In particular, for proximal gradient descent (assuming f is smooth), we require projection onto $C_1 + \cdots + C_k$ as discussed in Sect. 1, for which we can apply Algorithm 1. A fast and reliable algorithm for projection onto ℓ_q-norm disks is available (Liu & Ye, 2010). If $p \in [2, \infty)$, C_i is strongly convex with respect to $\|\cdot\|_2$ due to Lemma 3 of (Garber & Hazan, 2015), strong convexity of $\|\cdot\|_q^2$ on subspace V_i (Shalev-Shwartz, 2007, p. 130), and the fact that α-strongly convex sets w.r.t. $\|\cdot\|_q$ are $\alpha d^{1/q-1/2}$-strongly convex w.r.t. $\|\cdot\|_2$. Therefore, Algorithm 1 with $\rho = 0$ (corresponding to Eq. (5)) achieves a linear rate. If $p = \infty$, then each C_i is polyhedral, and Algorithm 1 with $\rho > 0$ (Eqs. (9)–(10)) converges linearly by the discussion following Theorem 3.2. We re-emphasize that no special considerations in the presence of overlaps are required.

We compare this Minkowski projection-based method with a popular method by Yuan et al. (2011) developed for $p = 2$. As both methods employ proximal gradient descent, it suffices to compare the performance of computing the proximal map $\operatorname{argmin}_{\boldsymbol{u} \in \mathbb{R}^d} \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{x}\|_2^2 + \lambda \sum_{i=1}^g \|\boldsymbol{u}_{i1}\|_2$. Briefly, this computation in Yuan et al. (2011) relies heavily on a preprocessing step to screen out zero groups, followed by dual projected gradient descent as well as a duality gap computation to certify convergence. Our empirical study follows their experimental setup: each group is of size $2d/g$ and overlaps half of the previous group, where d is the dimension and g denotes the number of groups. For example, if $d = 100$ and $g = 10$, group indices are $\{1, \ldots, 20\}$, $\{11, \ldots, 30\}, \ldots, \{91, \ldots, 100, 1, \ldots, 10\}$. For each combination of $d = 10^3, 10^4, 10^5, 10^6$ and $g = 10, 20, 50, 100$, proximal maps were computed using both methods for 50 randomly generated inputs $\boldsymbol{x}$; $\lambda = 2.1$ was used. We compare a MATLAB implementation of our algorithm to the competing method implemented in the MATLAB package SLEP (Liu et al., 2011). The simulation was run on a Linux machine with two Intel Xeon E5-2650v4 (2.20GHz) CPUs.

Results of the simulation study are summarized in Figure 1. The Minkowski projection method is much faster than SLEP when the dimension is high and the number of groups is moderate (top left). The number of summands appearing in the Minkowski sum grows is equal to the number of groups, and we show that our method slows down if the number of groups grows large (top right). Recall SLEP screens out zero groups in a preprocessing step implemented in C, explaining its advantage in the top right of Figure 1. Except for the d=1000, g=100 case, the algorithm terminated within tens of iterations (mostly less than 10, bottom left). Recall that in this experiment only the proximal operator was computed, which means the iteration count is for the "inner" iteration by the Minkowski projection. Thus the bottom left plot shows that the linear rate is achieved in practice. We observe that the accuracy of SLEP deteriorates when either the dimension or number of groups is high (bottom right), while Minkowski projections remain stable. That a naïve implementation of our method is already more accurate and competitive in runtime with a highly optimized package speaks to its promise. Further, our algorithm is both more transparent and more general, applying to any $p \geq 1$.

5.2. Constrained Lasso

Since the sets (4) involved with the constrained lasso (3) are closed polyhedra, Algorithm 1 with $\rho > 0$ converges linearly. Again, efficient projections onto subspace C_1 or $C_1^\perp$ and cone C_2 or C_2^* are the key to success. Following Gaines et al. (2018), we consider two important combinations of C_1 and C_2. The first is the zero-sum constrained lasso, which has been used in the analysis of compositional data such as those arising in microbiome informatics (Lin et al., 2014; Altenbuchinger et al., 2017). This case corresponds to $C_1 = \{\boldsymbol{x} : \sum_{j=1}^d x_j = 0\}^\perp$ and $C_2 = \{\boldsymbol{0}\}$. Projection onto C_1 merely averages the input components. The second is the nonnegative lasso (Efron et al., 2004; El-Arini et al., 2013; Wu et al., 2014), which corresponds to $C_1 = \{\boldsymbol{0}\}$ and $C_2 = \{\boldsymbol{x} : -\boldsymbol{x} \leq \boldsymbol{0}\}^*$. Projection onto C_2 corresponds to taking the negative part of the input.

For both examples, we followed the simulation setup of Gaines et al. (2018), and compare a MATLAB implementation of our method to a quadratic programming formulation solved via Gurobi (Gurobi Optimization, LLC, 2018), the al-

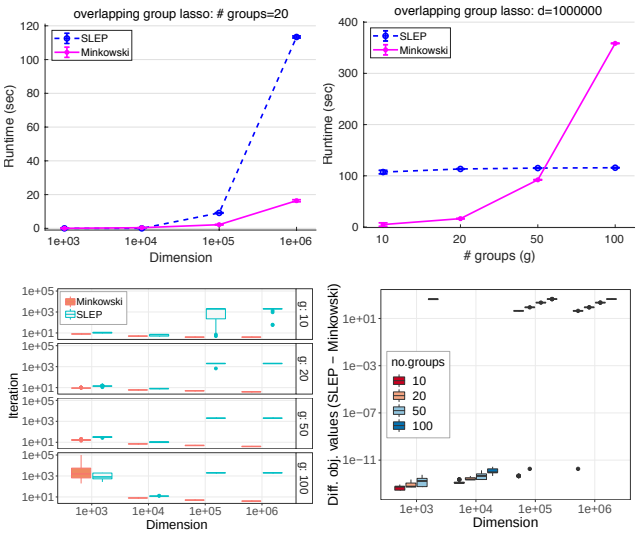

Figure 1. Comparison of the proposed Minkowski method and the dual projected gradient method by Yuan et al. (2011) for fitting the group lasso with overlap. Top left: runtime by dimension. Top right: runtime by number of groups. Bottom left: number of iterations until convergence by dimension and group. Bottom right: difference in final objective values by dimension and group. Additional timing results can be found in the Supplement.

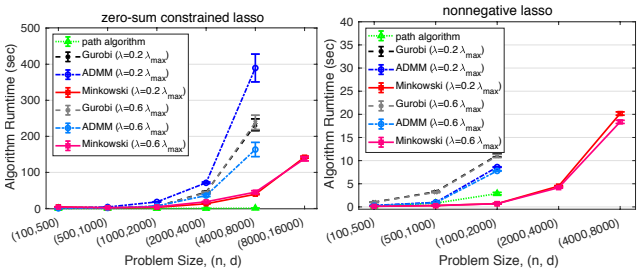

Figure 2. Comparison of the proposed Minkowski projection method and the other methods considered by Gaines et al. (2018) for fitting constrained lasso problems. Left: runtime for the zero-sum constrained lasso. Right: runtime for the nonnegative lasso.

ternating directions method of multipliers (ADMM), and the path-following algorithm developed in Gaines et al. (2018) as implemented in their MATLAB package. Note ADMM iterations require solving a lasso subproblem by calling a Fortran subroutine. We consider the regression objective $f(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2/2$ for various combinations of sample size n and dimension d. Each setting is repeated over 20 trials with randomly sampled $\boldsymbol{A}$ and noisy response $\boldsymbol{b}$. Four sparsity levels were tried: $\lambda/\lambda_{\max} = 0.2, 0.4, 0.6, 0.8$, where $\lambda_{\max}$ is the maximal sparsity level found by solving a linear program via Gurobi (Gaines et al., 2018, Sect. 3). The simulation was run on a Linux machine with two Intel Xeon E5-2680v2 (2.80GHz) CPUs with 256GB memory.

The average runtime of the methods are shown in Figure 2 for $\lambda/\lambda_{\max} = 0.2$ and 0.6. As the problem size (n, d) grows, the Minkowski method outperformed all other methods. The runtime for the path-following algorithm was normalized by the number of knots in the piecewise linear path as it computes the entire solutions for all λ. Though the path algorithm was the fastest up to $(n, d) = (4000, 8000)$ in the zero-sum lasso by this measure, together with the ADMM and Gurobi it did not terminate within four days in the run with $(n, d) = (8000, 16000)$. This is because all of these methods need to solve direct matrix inversion subproblems, which do not scale well as the problem size grows. This phenomenon becomes severe in the nonnegative lasso which imposes d inequality constraints. All but the Minkowski method hit a ceiling as early as $(n, d) = (2000, 4000)$, and the path algorithm loses its edge by $(500, 1000)$. It is worth noting that the Minkowski method was also less sensitive to the sparsity level than the ADMM, which slowed down for smaller λ. The accuracy of both first-order methods were similar ($< 0.0001\%$ difference relative to a Gurobi baseline); these comparisons and additional timing results for a range of λ values are detailed in the Supplement.

6. Discussion

We propose an efficient algorithm for projecting points onto Minkowski sums of sets, and provide a thorough convergence analysis in both convex and non-convex settings. In particular, the method achieves a linear rate of convergence whenever at least one summand is strongly convex or the Luo-Tseng error bound condition is satisfied. The algorithm can immediately be applied to several cornerstones of machine learning with competitive performances. Our method equips researchers to reconsider problems where structural complexities such as non-separability may now be handled gracefully via formulations involving Minkowski sums.

We have demonstrated that converting structurally complex penalties into set constraints via duality (M) can be effective within algorithms such as proximal gradient methods. Since we envision Algorithm 1 to serve as an inner loop in these methods, our emphasis on fast convergence is warranted.

Finally, our algorithm does not require projections onto the sets at hand to be exact. From the MM perspective (Sect 3), each projection minimizes the surrogate, but doing so is not necessary for convergence — any descent step is enough. We observed similar success using inexact projections: for instance, when the ball projection is computed via bisection when $p > 1$ for $\ell_{1,p}$ minimization. In light of the analysis by Schmidt et al. (2011), such approximate projections are valid whenever errors are absolutely summable, and the $O(1/n)$ rate holds if they are square-root summable.

Acknowledgements

The authors thank anonymous reviewers for insightful comments that helped to greatly improve the manuscript.

Joong-Ho Won was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT, No. 2019R1A2C1007126). Kenneth Lange was supported by grants from the USA National Human Genome Research Institute (HG006139) and the USA National Institute of General Medical Sciences (GM053275).

References

Altenbuchinger, M., Rehberg, T., Zacharias, H., Stämmler, F., Dettmer, K., Weber, D., Hiergeist, A., Gessner, A., Holler, E., Oefner, P. J., et al. Reference point insensitive molecular data analysis. *Bioinformatics*, 33(2):219–226, 2017.

Attouch, H., Bolte, J., Redont, P., and Soubeyran, A. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

Balashov, M. V. and Golubev, M. O. About the Lipschitz property of the metric projection in the Hilbert space. *Journal of Mathematical Analysis and Applications*, 394 (2):545–551, 2012.

Bauschke, H. H. and Borwein, J. M. On the convergence of von Neumann's alternating projection algorithm for two sets. *Set-Valued Analysis*, 1(2):185–212, 1993.

Bierstone, E. and Milman, P. D. Semianalytic and subanalytic sets. *Publications Mathématiques de l'Institut des Hautes Études Scientifiques*, 67(1):5–42, 1988.

Bochnak, J., Coste, M., and Roy, M.-F. *Real algebraic geometry*, volume 36. Springer Science & Business Media, 2013.

Bolte, J., Daniilidis, A., and Lewis, A. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2010.

Combettes, P. L. and Wajs, V. R. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

Davis, D. and Yin, W. A three-operator splitting scheme and its optimization applications. *Set-valued and Variational Analysis*, 25(4):829–858, 2017.

Dykstra, R. L. An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78(384):837–842, 1983.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

El-Arini, K., Xu, M., Fox, E. B., and Guestrin, C. Representing documents through their readers. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 14–22. ACM, 2013.

Elsner, L., Koltracht, I., and Neumann, M. Convergence of sequential and asynchronous nonlinear paracontractions. *Numerische Mathematik*, 62(1):305–319, 1992.

Gaines, B. R., Kim, J., and Zhou, H. Algorithms for fitting the constrained lasso. *Journal of Computational and Graphical Statistics*, 27(4):861–871, 2018.

Garber, D. and Hazan, E. Faster rates for the Frank-Wolfe method over strongly-convex sets. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 541–549, 2015.

Gurobi Optimization, LLC. Gurobi optimizer reference manual, 2018. URL `http://www.gurobi.com`.

Hiriart-Urruty, J.-B. and Lemaréchal, C. *Fundamentals of Convex Analysis*. Springer Science & Business Media, 2012.

James, G. M., Paulson, C., and Rusmevichientong, P. Penalized and constrained regression. *Unpublished Manuscript, available at* `http://www-bcf.usc.edu/~gareth/research/Research.html`, 2013.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. *arXiv preprint arXiv:1608.04636 v3*, 2018.

Kruger, A. Y. On Fréchet subdifferentials. *Journal of Mathematical Sciences*, 116(3):3325–3358, 2003.

Lange, K. *Optimization*. Springer, 2nd edition, 2013.

Lange, K. *MM Optimization Algorithms*. SIAM, 2016.

Li, G. and Pong, T. K. Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of Computational Mathematics*, 18(5):1199–1232, 2018.

Lin, W., Shi, P., Feng, R., and Li, H. Variable selection in regression with compositional covariates. *Biometrika*, 101(4):785–797, 2014.

Liu, J. and Ye, J. Efficient ℓ_1/ℓ_q norm regularization. *arXiv preprint arXiv:1009.4766*, 2010.

Liu, J., Ji, S., and Ye, J. Slep: Sparse learning with efficient projections. Technical report, Arizona State University, 2011. URL https://github.com/jiayuzhou/SLEP.

Qin, X. and An, N. T. Smoothing algorithms for computing the projection onto a minkowski sum of convex sets. *arXiv preprint arXiv:1801.08285*, 2018.

Rockafellar, R. T. and Wets, R. J.-B. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

Schmidt, M., Roux, N. L., and Bach, F. R. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in Neural Information Processing Systems*, pp. 1458–1466, 2011.

Shalev-Shwartz, S. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University of Jerusalem, 7 2007.

Tibshirani, R. J. and Taylor, J. The solution path of the generalized lasso. *Annals of Statistics*, 39(3):1335–1371, 2011. ISSN 0090-5364.

von Neumann, J. *Functional Operators*, volume 2. Princeton University Press, Princeton, NJ, 1950. Reprint of mimeographed lecture notes first distributed in 1933.

Wu, L., Yang, Y., and Liu, H. Nonnegative-lasso and application in index tracking. *Computational Statistics & Data Analysis*, 70:116–126, 2014.

Yang, Y. and Zou, H. A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, 25(6):1129–1141, 2015.

Yuan, L., Liu, J., and Ye, J. Efficient methods for overlapping group lasso. In *Advances in Neural Information Processing Systems*, pp. 352–360, 2011.

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68 (1):49–67, 2006.

Zangwill, W. I. *Nonlinear Programming: A Unified Approach*. Prentice-Hall, 1969.