Prox-Regularity of Rank Constraint Sets and Implications for Algorithms

D. Russell Luke

Springer

Springer

Prox-Regularity of Rank Constraint Sets and Implications for Algorithms

D. Russell Luke

Abstract We present an analysis of sets of matrices with rank less than or equal to a specified number s. We provide a simple formula for the normal cone to such sets, and use this to show that these sets are prox-regular at all points with rank exactly equal to s. The normal cone formula appears to be new. This allows for easy application of prior results guaranteeing local linear convergence of the fundamental alternating projection algorithm between sets, one of which is a rank constraint set. We apply this to show local linear convergence of another fundamental algorithm, approximate steepest descent. Our results apply not only to linear systems with rank constraints, as has been treated extensively in the literature, but also nonconvex systems with rank constraints.

Keywords Rank optimization · Rank constraint · Sparsity · Normal cone · Prox-regular · Constraint qualification · Projection operator · Method of alternating projections · Linear convergence · Superregularity

1 Introduction

Rank optimization is a well-developed topic that has found a tremendous number of applications in recent years (see [23] and references therein). Most of the problems one encounters involve a linear data model that is underdetermined and the very poorly behaved "sparsity function", either the function determining the rank of a matrix or the function counting the number of nonzero entries in an array. A common approach to solving sparsity optimization problems is via a convex surrogate, most often the ℓ_1 or (in the case of matrices) the nuclear norm. The rational for working with such surrogates is that the original problem is NP-complete, and thus should be avoided. Inspired by earlier work proving local convergence of cyclic projections onto nonconvex sets with an application to sparse signal recovery [6], and a more recent projection-reflection algorithm for x-ray imaging [20] that appears to be very successful at working with a proximal operator of the ℓ_0 function and a nonlinear imaging model, we set out in the present note to determine whether sets with sparsity constraints have some sort of regularity that might justify working directly with sparsity rather than through convex surrogates.

Based on the work of Lewis and Sendov [15, 16], Le has obtained an explicit formula for the subdifferential of the rank function [11]. This formula shows that every point of the rank function is a critical point [8], and so reasonable algorithmic strategies should not *directly* make use of the rank function. Instead, we consider the *lower level sets* of the rank function. While sets of matrices of rank less than a specified level are not manifolds, we show here that they are quite regular, in fact *prox-regular*. While prox-regularity of these sets is not new [14], our proof of this fact established in Sect. 3 uses elementary tools, at the center of which is a particularly simple and apparently new characterization of the *normal cone* to these sets established in Proposition 3.6.

Prox-regularity of the lower level sets of the rank function immediately yields local linear convergence of fundamental algorithms for either finding the intersection of the rank constraint set with another set determined by some (nonlinear) data model, or for minimizing the distance to a rank constrained set and a data set. The result, detailed in Sect. 4, is quite general and extends to nonconvex data imaging models with rank constraints. Our results are an extension of results established recently in [3] for the vector case, however

D.R. Luke (✉)
Institut für Numerische und Angewandte Mathematik, Universität Göttingen, 37083, Göttingen, Germany
e-mail: r.luke@math.uni-goettingen.de

at the cost of additional assumptions on the regularity of the solution set. In particular, [3] establishes local linear convergence, with radius of convergence, of alternating projections between an affine constraint and the set of vectors with no more than s nonzero elements *without any assumptions* on the regularity of the intersection of these sets, beyond the assumption that it is nonempty. Our results, in contrast, are modeled after results of [14] and [13] where a stronger regularity of the intersection is assumed. We discuss the difficulties in extending the tools developed in [4] to the matrix case in the conclusion. In any case, avoiding convex surrogates is at the cost of global convergence guarantees: these results are local and offer no panacea for solving rank optimization problems. Rather, this analysis shows that certain *macro-regularity* assumptions such as restricted isometry or mutual coherence (see [23] and references therein) play no role *asymptotically* in the convergence of algorithms, but rather have bearing only on the radius of convergence. We begin this note with a review of notation and basic results and definitions upon which we build.

2 Notation

Throughout this paper $\mathcal{X}$ and $\mathcal{Y}$ are Euclidean spaces. In particular we are interested in Euclidean spaces defined on $\mathbb{R}^{m \times n}$ where we derive the norm from the trace inner product

$$\langle y, x \rangle := \mathbf{Tr}\left(y^T x\right) \quad \text{for } x, y \in \mathbb{R}^{m \times n},$$

$$\|x\| := \sqrt{\mathbf{Tr}\left(x^T x\right)}.$$

This naturally specializes to the case of $\mathbb{R}^n$ when $m = n$ above and $x \in \mathbb{R}^{n \times n}$ is restricted to the subspace of diagonal matrices. For $x \in \mathbb{R}^{m \times n}$ we denote the span of the rows of x by $\text{range}(x^T)$ and recall that this is orthogonal to the nullspace of the linear mapping $x : \mathbb{R}^n \to \mathbb{R}^m$,

$$\text{range}\left(x^T\right) = \ker(x)^\perp.$$

For $x \in \{z \in \mathbb{R}^{n \times n} \mid z_{ij} = 0 \text{ if } i \neq j\}$ (that is, when x is square diagonal) this corresponds exactly to the usual support of vectors on $\mathbb{R}^n$:

$$\text{range}\left(x^T\right) = \text{supp}\left(\text{Diag}(x)\right)$$
$$:= \left\{ y \in \mathbb{R}^n \mid y_i = 0 \text{ for all } i \in \{1, 2, \ldots, n\} \text{ with} \right.$$
$$\left. \text{Diag}(x)_i = 0 \right\}$$

where $\text{Diag}(x)$ maps the diagonal of the matrix $x \in \mathbb{R}^{m \times n}$ to a vector in $\mathbb{R}^r$ with $r = \min\{m, n\}$. In order to emphasize this connection to the support of vectors, and reduce notational clutter we will denote the span of the rows of x by

$$\text{Supp}(x) := \text{range}\left(x^T\right).$$

We denote the rank of x by $\text{rank}(x)$ and recall that $\text{rank}(x)$ is the dimension of the span of the columns— or equivalently the rows—of x which is equivalent to the number of nonzero singular values. The singular values of $x \in \mathbb{R}^{m \times n}$ are the (positive) square root of the eigenvalues of $x x^T$; these are denoted by $\sigma_j(x)$ and are assumed to be ordered so that $\sigma_i(x) \geq \sigma_j(x)$ for $i < j$. We denote by $\sigma(x) := (\sigma_1(x), \sigma_2(x), \ldots, \sigma_r(x))^T$ $(r = \min\{m, n\})$ the ordered vector of singular values of x. The corresponding diagonal matrix is denoted $\Sigma(x) := \text{diag}(\sigma(x)) \in \mathbb{R}^{m \times n}$ where $\text{diag}(\cdot)$ maps vectors in $\mathbb{R}^r$ to matrices in $\mathbb{R}^{m \times n}$. Following [12, 15, 16] we denote the (Lie) group of $n \times n$ orthogonal matrices by $O(n)$ and the product $O(m) \times O(n)$ by $O(m, n)$. A singular value decomposition of $x \in \mathbb{R}^{m \times n}$ restricted to the above ordering is then any pair of orthogonal matrices $(U, V) \in O(m, n)$ together with $\Sigma(x)$ such that $x = U \Sigma(x) V^T$. We will denote the set of pairs of orthogonal matrices that comprise singular systems for x by $\mathcal{U}(x) := \{(U, V) \in O(m, n) \mid x = U \Sigma(x) V^T\}$.

The *closed* ball centered at x with radius ρ is denoted by $\mathbb{B}(x, \rho)$; the unit ball centered at the origin is simply denoted by $\mathbb{B}$. Given a set $\Omega \subset \mathcal{X}$, we denote the distance of a point $x \in \mathcal{X}$ to Ω by $d_\Omega(x)$ where

$$d_\Omega(x) := \inf_{y \in \Omega} \|y - x\|.$$

If Ω is empty then we use the convention that the distance to this set is $+\infty$. The corresponding (multivalued) projection operator of x onto Ω, denoted $P_\Omega(x)$, is defined by

$$P_\Omega(x) := \operatorname*{argmin}_{z \in \Omega} \|z - x\|.$$

If Ω is nonempty and closed, then the projection of any point in $\mathcal{X}$ onto Ω is nonempty.

We define the *normal cone* to a closed set $\Omega \subset \mathcal{X}$ following [24, Definition 6.3]:

Definition 2.1 (normal cone) A vector $v \in \mathcal{X}$ is *normal* to a closed set $\Omega \subset \mathcal{X}$ at $\overline{x} \in \Omega$, written $v \in N_\Omega(\overline{x})$ if there are sequences $(x^k)_{k \in \mathbb{N}}$ in Ω with $x^k \underset{\Omega}{\to} \overline{x}$ and $(v^k)_{k \in \mathbb{N}}$ in X with $v^k \to v$ such that

$$\limsup_{\substack{x \underset{\Omega}{\to} x^k \\ x \neq x^k}} \frac{\langle v^k, x - x^k \rangle}{|x - x^k|} \leq 0.$$

The vectors v^k are *regular normals* to Ω at x^k and the cone of regular normals at x^k is denoted $\widehat{N}_\Omega(x^k)$.

What we are calling regular normals are called *Fréchet* normals in [19, Definition 1.1].

Here and elsewhere we use the notation $x \underset{\Omega}{\to} \overline{x}$ to mean that $x \to \overline{x}$ with $x \in \Omega$. An important example of a regular normal is a *proximal normal*, defined as any vector $v \in \mathcal{X}$

that can be written as $v = \lambda(x - \overline{x})$ for $\lambda \geq 0$ and $\overline{x} \in P_\Omega(x)$ for some $x \in \mathcal{X}$. We denote the set of proximal normals to Ω at $x \in \Omega$ by $N_\Omega^P(x)$. For Ω closed and nonempty, any normal $\overline{v} \in N_\Omega(\overline{x})$ can be approximated arbitrarily closely by a proximal normal [24, Exercise 6.18]. Thus we have the next result which is key to our analysis.

Proposition 2.2 (Theorem 1.6 of [19]) *Let $\Omega \subset \mathcal{X}$ be closed and $\overline{x} \in \Omega$. Then*

$$N_\Omega(\overline{x})$$
$$= \left\{ v \in \mathbb{R}^n \mid \exists \text{ sequences } x^k \to \overline{x} \text{ and } v^k \to v \text{ with} \right.$$
$$\left. v^k \in \mathrm{cone}\big(x^k - P_\Omega(x^k)\big) \text{ for all } k \in \mathbb{N} \right\}. \qquad (2.1)$$

Central to our results is the *regularity* of the intersection of sets, which we define in terms of a type constraint qualification formulated with the normal cones to the sets at points in the intersection.

Definition 2.3 (basic set intersection qualification) A family of closed sets $\Omega_1, \Omega_2, \ldots \Omega_m \subset \mathcal{X}$ satisfies the basic set intersection qualification at a point $\overline{x} \in \cap_i \Omega_i$, if the only solution to

$$\sum_{i=1}^m y_i = 0, \quad y_i \in N_{\Omega_i}(\overline{x}) \ (i = 1, 2, \ldots, m)$$

is $y_i = 0$ for $i = 1, 2, \ldots, m$. We say that the intersection is *strongly regular* at $\overline{x}$ if the basic set constraint qualification is satisfied there.

In the case $m = 2$, this condition can be written

$$N_{\Omega_1}(\bar{x}) \cap -N_{\Omega_2}(\bar{x}) = \{0\}.$$

The two set case is called the *basic constraint qualification for sets* in [19, Definition 3.2] and has its origins in the *generalized property of nonseparability* [18] which is the n-set case. It was later recovered as a dual characterization of what is called *strong regularity* of the intersection in [10, Theorem 3]. It is called *linear regularity* in [13].

The case of two sets also yields the following simple quantitative characterization of strong regularity.

Proposition 2.4 (Theorem 5.16 of [13]) *Suppose that Ω_1 and Ω_2 are closed subsets of $\mathcal{X}$. The intersection $\Omega_1 \cap \Omega_2$ satisfies the basic set intersection qualification at $\overline{x}$ if and only if the constant*

$$\overline{c} := \sup\big\{ \langle u, v \rangle \mid u \in N_{\Omega_1}(\overline{x}) \cap \mathbb{B}, \ v \in -N_{\Omega_2}(\overline{x}) \cap \mathbb{B} \big\} < 1. \qquad (2.2)$$

Definition 2.5 angle of regular intersections Suppose that Ω_1 and Ω_2 are closed subsets of $\mathcal{X}$. We say that the intersection $\Omega_1 \cap \Omega_2$ is *strongly regular* at $\overline{x} \in \Omega_1 \cap \Omega_2$ *with angle* $\overline{\theta} := \cos^{-1}(\overline{c}) > 0$ when the constant $\overline{c}$ given by (2.2) is less than 1.

We will also require certain regularity of the sets themselves, not just the intersection. The following definition of *prox-regularity* of sets is a modern manifestation that can be traced back to [7] and *sets of positive reach*. What we use here as a definition actually follows from the equivalence of prox-regularity of sets as defined in [22, Definition 1.1] and the single-valuedness of the projection operator on neighborhoods of the set [22, Theorem 1.3].

Definition 2.6 (prox-regularity) A nonempty closed set $\Omega \subset \mathcal{X}$ is *prox-regular* at a point $\overline{x} \in \Omega$ if $P_C(x)$ is single-valued around $\overline{x}$.

3 Properties of Lower Level Sets of the Rank Function

We collect here some facts that will be used repeatedly in what follows.

Proposition 3.1 *For any point $\overline{x} \in \mathbb{R}^{m \times n}$ and any sequence $(x^k)_{k \in \mathbb{N}}$ converging to $\overline{x}$ there is a $K \in \mathbb{N}$ such that $\mathrm{rank}(\overline{x}) \leq \mathrm{rank}(x^k)$ for all $k > K$.*

Proof This follows immediately from continuity of the singular values as a function of x. (See, for instance, [9, Appendix D].) $\qquad \square$

For the remainder of this note we will consider real $m \times n$ matrices and denote by r the minimum of $\{m, n\}$. The rank level set will be denoted by $S := \{y \in \mathbb{R}^{m \times n} \mid \mathrm{rank}(y) \leq s\}$ for $s \in \{0, 1, \ldots, r\}$. As can be found in textbooks on matrix analysis, the projection onto this set is just the truncation of the $r - s$ smallest singular vectors to zero; in the case of a tie for the s-th largest singular value, the projection is the set of all s-selections from the s-largest singular values.

Lemma 3.2 (projection onto S) *For $x \in \mathbb{R}^{m \times n}$, define*

$$\Sigma_s(x)$$
$$:= \mathrm{diag}\big((\sigma_1(x), \sigma_2(x), \ldots, \sigma_s(x), 0, \ldots, 0)^T \big) \in \mathbb{R}^{m \times n}.$$

The projection $P_S(x)$ is given by

$$P_S(x) = \bigcup_{(U,V) \in \mathcal{U}(x)} \big\{ y \mid y = U \Sigma_s(x) V^T \big\}.$$

Proof By [9, Theorem 7.4.51] any matrix $y \in S$ satisfies $\|x - y\| \geq \|\Sigma(x) - \Sigma(y)\|$. The relation holds with equality whenever $y = U\Sigma_s(x)V^T$ for some $(U, V) \in \mathcal{U}(x)$, hence $P_S(x) \supset \bigcup_{(U,V) \in \mathcal{U}(x)} \{y \mid y = U\Sigma_s(x)V^T\} \neq \emptyset$. On the other hand, if $\overline{y} \in P_S(x)$, then $\|x - \overline{y}\| \leq \|x - y\|$ for all $y \in S$. In particular, for $y = U\Sigma_s(x)V^T$ with $(U, V) \in \mathcal{U}(x)$ we have

$$\|x - \overline{y}\| \leq \|x - y\| = \left\|\Sigma(x) - \Sigma_s(x)\right\|$$
$$\leq \left\|\Sigma(x) - \Sigma_s(\overline{y})\right\| \leq \|x - \overline{y}\|,$$

hence $\Sigma_s(x) = \Sigma(\overline{y})$ and $\overline{y} \in \bigcup_{(U,V) \in \mathcal{U}(x)} \{y \mid y = U\Sigma_s(x)V^T\}$. $\qquad\square$

The next results establish that the set S is prox-regular at all points where $\mathrm{rank}(x) = s$. We make use of the following tools. For $r = \min\{m, n\}$ define the mappings $\mathbb{J}: \mathbb{R}^{m \times n} \times (\mathbb{R}_+ \cup \{+\infty\}) \to 2^{\{1,2,\dots,r\}}$ and $\alpha_s(x): \mathbb{R}^{m \times n} \to [0, +\infty]$ by

$$\mathbb{J}(x, \alpha) := \left\{j \in \{1, 2, \dots, r\} \mid \sigma_j(x) \geq \alpha\right\} \quad \text{and}$$

$$\alpha_s(x) := \sup\left\{\alpha \mid \left|\mathbb{J}(x, \alpha)\right| \geq s\right\},$$

where $|\mathbb{J}(x, \alpha)|$ denotes the cardinality of this discrete set. We define $\mathbb{J}(x, +\infty)$ to be the empty set. Before proceeding with our results, we collect some observations about these objects.

Lemma 3.3

(i) *For all $s \in \{1, 2, \dots, r\}$ the value of the supremum in the definition of $\alpha_s(x)$ is bounded and attained. If $s = 0$ then $\alpha_0(x) = +\infty$.*
(ii) *If $|\mathbb{J}(x, \alpha_s)| > s$ then $\mathrm{rank}(x) > s > 0$.*
(iii) *If $\mathrm{rank}(x) > s$ then $\alpha_s > 0$.*
(iv) *If $\mathrm{rank}(x) < s \leq r$ then $\alpha_s(x) = 0$.*

Proof (i) Since the cardinality of the empty set is zero, the supremum in the definition of α_0 is unbounded. In any case, the cardinality of $\mathbb{J}(x, \alpha)$ is monotonically decreasing with respect to α for x fixed from a value of r at $\alpha = 0$ to 0 for all $\alpha > \sigma_1(x)$. Thus for x fixed α_s is bounded for all $s \in \{1, 2, \dots, r\}$. The value of α for which the cardinality $s \geq 1$ is achieved is attained precisely when $\alpha = \sigma_j(x)$ for some j. (ii) By definition, at $s = 0$, $\alpha_0 = +\infty$ and $|\mathbb{J}(x, +\infty)| := 0$, so $|\mathbb{J}(x, \alpha_s)| > s$ implies that $s > 0$ and the implication $|\mathbb{J}(x, \alpha_s)| > \mathrm{rank}(x)$ follows immediately. (iii) If $\mathrm{rank}(x) > s$ and $s = 0$, then the result is trivial since $\alpha_0 := +\infty$. If $\mathrm{rank}(x) > s$ and $s > 0$ then $s \in \{1, \dots, r - 1\}$ (it is impossible to have rank greater than r) and there exists an $\alpha > 0$ such that $|\mathbb{J}(x, \alpha)| \geq s + 1$. As α_{s+1} is the maximum of these, $\alpha_{s+1} > 0$. By the argument in (i) $\alpha_s \geq \alpha_{s+1}$ which yields the result. (iv) In this case, only by including the zero singular values of x can the inequality $|\mathbb{J}(x, \alpha)| \geq s$ be achieved, that is by taking $\alpha = 0$. $\qquad\square$

Proposition 3.4 (properties of the projection) *The following are equivalent.*

(i) *$P_S(x)$ is multi-valued;*
(ii) *$|\mathbb{J}(x, \alpha_s)| > s$.*

Proof To show that (i) implies (ii), let y and $z \in P_S(x)$ with $y \neq z$. By Lemma 3.2 $y = U_y \Sigma_s(x) V_y^T$ and $z = U_z \Sigma_s(x) V_z^T$ for (U_y, V_y) and $(U_z, V_z) \in \mathcal{U}(x)$. Then by [9, Theorem 7.4.51]

$$0 = \left\|\Sigma(y) - \Sigma(z)\right\| < \|y - z\| \leq \|y - x\| + \|x - z\|$$
$$= 2\left\|\Sigma(x) - \Sigma_s(x)\right\|$$

hence $\mathrm{rank}(x) > s$. Since $y \neq z$ and they have the same singular values, the multiplicity of singular values $\sigma_j(x)$ with value α_s must be greater than one, hence $|\mathbb{J}(x, \alpha_s)| > s$.

Conversely, to show that (ii) implies (i) first note that by Lemma 3.3(ii) $\mathrm{rank}(x) > s > 0$. Now fix $y \in P_S(x)$ with $y = U_y \Sigma_s(x) V_y^T$ for $(U_y, V_y) \in \mathcal{U}(x)$. The corresponding decomposition for x is $U_y \Sigma(x) V_y^T$. Now construct the orthogonal matrix $\widetilde{V}$ by switching the $s + 1$'th column of V_y with the s'th column of V_y. Since $|\mathbb{J}(x, \alpha_s)| > s$ we have that $x = U_y \Sigma(x) \widetilde{V}^T$. Define $z := U_y \Sigma_s(x) \widetilde{V}^T$. By Lemma 3.2 $z \in P_S(x)$ with $\mathrm{rank}(z) = s$, but the s'th column of $\widetilde{V}$ is in the nullspace of y so $z \neq y$ and the projection is thus multi-valued. This completes the proof. $\qquad\square$

An immediate consequence of the above is the obvious observation that the projection onto the trivial sparsity sets S_0 with $s = 0$ and S_r with $s = r$ is single-valued.

Corollary 3.5 *For $x \in \mathbb{R}^{m \times n}$, if $s = 0$ or $s = r$ then $P_S(x)$ is single-valued.*

The normal cone of this set has the following simple characterization.

Proposition 3.6 (the normal cone to S) *At a point $\overline{x} \in S$*

$$N_S(\overline{x}) = \left\{v \in \mathbb{R}^{m \times n} \mid \ker(v)^{\perp} \cap \ker(\overline{x})^{\perp} = \{0\} \text{ and} \right.$$
$$\left. \mathrm{rank}(v) \leq r - s\right\}. \qquad (3.1)$$

Moreover, $N_S(\overline{x}) = N_S^P(\overline{x})$ at every $\overline{x}$ with $\mathrm{rank}(\overline{x}) = s$, while $N_S^P(\overline{x}) = \{0\}$ at every $\overline{x}$ with $\mathrm{rank}(\overline{x}) < s$.

Proof Using the definition of $\mathrm{Supp}(x) := \ker(x)^{\perp}$ define the sets

$$W := \left\{v \in \mathbb{R}^{m \times n} \mid \mathrm{Supp}(v) \cap \mathrm{Supp}(\overline{x}) = \{0\} \text{ and} \right.$$
$$\left. \mathrm{rank}(v) \leq r - s\right\} \quad \text{and}$$
$$Z(w) := \left\{z \in \mathbb{R}^{m \times n} \mid \mathrm{Supp}(z) \cap \mathrm{Supp}(w) = \{0\} \text{ and} \right.$$

$$\text{rank}(\overline{x} + z) = s\Big\}.$$

We first show that W is nonempty and hence $Z(w)$ for $w \in W$ is nonempty. For all $\overline{x} \in \mathbb{R}^{m \times n}$ and $s \in \{0, 1, 2, \ldots, r\}$ the zero matrix $0 \in W$, hence W is nonempty. Next note that for $w \in W$, $Z(w) \subset \ker(w)$ with $\dim(\ker(w)) \geq s \geq 0$, and it is always possible to find an element z of $\ker(w)$ with $\text{rank}(\overline{x} + z) = s$.

Now, choose any $w \in W$ and $z_0 \in Z(w)$ and construct the sequences $(x^k)_{k \in \mathbb{N}}$ and $(w^k)_{k \in \mathbb{N}}$ by

$$x^k = \overline{x} + \tfrac{1}{k}w + \frac{1}{\sqrt{k}}z_0 \quad \text{and}$$

$$w^k = k\big(x^k - y^k\big), \quad \text{for } y^k \in P_S\big(x^k\big) \ (k \in \mathbb{N}).$$

There is a $K \in \mathbb{N}$ such that for all $k > K$

$$\frac{1}{k}\max_j\{\sigma_j(w)\}$$

$$< \min_j\left\{\sigma_j\left(\frac{1}{\sqrt{k}}z_0 + \overline{x}\right) \,\Big|\, \sigma_j\left(\frac{1}{\sqrt{k}}z_0 + \overline{x}\right) \neq 0\right\}.$$

Thus for all $k > K$

$$P_S\big(x^k\big) = \overline{x} + \frac{1}{\sqrt{k}}z_0 \quad \text{and}$$

$$w^k = k\left(x^k - \left(\overline{x} + \frac{1}{\sqrt{k}}z_0\right)\right) = w.$$

Note that by Proposition 3.4 and Lemma 3.3(ii) the representation of the projection above holds with equality since $\text{rank}(\overline{x} + \frac{1}{\sqrt{k}}w_0) = s$. Since $x^k \to \overline{x}$, by definition, $w \in N_S(\overline{x})$. As w was arbitrary, we have $W \subset N_S(\overline{x})$.

We show next that, conversely, $N_S(\overline{x}) \subset W$ for $\overline{x} \in S$. The matrix $w = 0$ trivially belongs to W, so we assume that $w \neq 0$. By Proposition 2.2 we can write w as a limit of proximal normals, that is, the limit of sequences (x^k) and (w^k) with $x^k \notin S$ and $w^k \to w$ for $w^k = t^k(x^k - y^k)$ for $y^k \in P_S(x^k)$. We consider the corresponding singular value decompositions by $y^k = U_k \Sigma_s(x^k)V_k^T$ for $(U_k, V_k) \in \mathcal{U}(x^k)$ and $\Sigma_s(x) := \text{diag}((\sigma_1(x), \sigma_2(x), \ldots, \sigma_s(x), 0, \ldots, 0)^T) \in \mathbb{R}^{m \times n}$ (see Lemma 3.2). Note that x^k and y^k have the same left and right singular vectors with the usual ordering. The matrices U_k and V_k are also collections of left and right singular vectors for w^k, although they do not yield the usual ordering of singular values of w^k:

$$w^k = t_k U_k \widetilde{\Sigma}_s\big(x^k\big)V_k^T$$

$$\text{for } \widetilde{\Sigma}_s\big(x^k\big) := \text{diag}\big((0, 0, \ldots, 0, \sigma_{s+1}\big(x^k\big), \sigma_{s+2}\big(x^k\big),$$

$$\ldots, \sigma_r\big(x^k\big))^T\big)$$

Let $(\overline{U}, \overline{V}) \in \mathcal{U}(\overline{x})$ be the limit of left and right singular vectors of x^k, that is, $U_k \to \overline{U}$, $V_k \to \overline{V}$ where

$x^k = U_k \Sigma(x^k)V_k \to \overline{U}\Sigma(\overline{x})\overline{V} = \overline{x}$. Then $y^k = U_k \Sigma_s(x^k)V_k \to \overline{U}\Sigma(\overline{x})\overline{V}$ and $w^k = t_k U_k \widetilde{\Sigma}_s(x^k)V_k \to \overline{U}(\lim_{k \to \infty} t_k \widetilde{\Sigma}_s(x^k))\overline{V} = w$. It follows immediately that $\text{rank}(w) \leq r - s$ and $\text{Supp}(w) \perp \text{Supp}(\overline{x})$ which completes the proof of the inclusion.

To see that each normal to the set S at $\overline{x}$ with $\text{rank}(\overline{x}) = s$ is actually a proximal normal, note that if $\text{rank}(\overline{x}) = s$ then by (3.1) every point $v \in N_S(\overline{x})$ can be written as $v = \frac{1}{\tau}((\tau v + \overline{x}) - P_S(\tau v + \overline{x}))$ for $\tau > 0$ small enough. Suppose, on the other hand, that $\text{rank}(\overline{x}) < s$. Then $P_S(\tau v + \overline{x}) = \overline{x}$ for $\tau > 0$ exactly when $v = 0$: for if $\tau v + \overline{x} \in S$ then $P_S(\tau v + \overline{x}) = \tau v + \overline{x} = \overline{x}$ exactly when $v = 0$, and if $\tau v + \overline{x} \notin S$ then $\text{rank}(P_S(\tau v + \overline{x})) = s$ hence $P_S(\tau v + \overline{x}) \neq \overline{x}$. Consequently the only proximal normal at these points is $v = 0$. This completes the proof. $\square$

The normal cone condition $N_S(\overline{x}) \cap (-N_\Omega(\overline{x})) = \{0\}$ can easily be checked by determining the nullspace of matrices in $N_\Omega(\overline{x})$ as the next theorem shows.

Proposition 3.7 (strong regularity of intersections with a sparsity set) *Let $\Omega \subset \mathbb{R}^{m \times n}$ be closed. If at a point $\overline{x} \in \Omega \cap S$ all nonzero $v \in N_\Omega(\overline{x})$ have $\ker(v)^\perp \cap \ker(\overline{x})^\perp \neq \{0\}$, then the intersection is strongly regular there.*

Proof Choose any $v \in N_\Omega(\overline{x})$. Since $\text{Supp}(v) \cap \text{Supp}(\overline{x}) \neq \{0\}$ and $N_S(\overline{x})$ given by (3.1) is a subset of matrices w with $\text{Supp}(w) \cap \text{Supp}(\overline{x}) = \{0\}$ the only solution to $v - w = 0$ is $v = w = 0$. $\square$

It is known that the set of matrices with rank s is a smooth manifold [1] (although the set of matrices with rank less than or equal to s is not), from which it follows that S is prox-regular at points with rank s [14, Lemma 2.1 and Example 2.3]. We present here a simple proof of this fact based on the characterization of the normal cone.

Proposition 3.8 (prox-regularity of S) *The set S is prox-regular at all points $\overline{x} \in \mathbb{R}^{m \times n}$ with $\text{rank}(\overline{x}) = s$.*

Proof Let $(x^k)_{k \in \mathbb{N}}$ in $\mathbb{R}^{m \times n}$ be any sequence converging to $\overline{x}$ with the corresponding singular value decomposition $U_k \Sigma(x^k)V_k^T$. Decompose x^k into the sum $y^k + z^k = x^k$ where $y^k = U_k \Sigma_s(x)V_k^T$ and $z^k = U_k \widetilde{\Sigma}_s(x^k)V_k^T$ with $\Sigma_s(x^k) := (\sigma_1(x^k), \sigma_2(x^k), \ldots, \sigma_s(x^k), 0 \ldots, 0)^T$ and $\widetilde{\Sigma}_s(x^k) := (0, \ldots, 0, \sigma_{s+1}(x^k), \sigma_{s+2}(x^k), \ldots, \sigma_r(x^k))^T$ for $r = \min\{m, n\}$. Note that $y^k \to \overline{x}$ with $\text{rank}(y^k) = \text{rank}(\overline{x}) = s$ for all k large enough, while by Proposition 3.6 $z^k \to 0$ with $z^k \in N_S(y^k)$ for all k. Then for all k large enough $\max_j\{\sigma_j(z^k)\} = \sigma_{s+1}(x^k) < \sigma_s(x^k) = \min_j\{\sigma_j(y^k)\}$ and $|\mathbb{J}(x^k, \alpha_s)| = s$. By Proposition 3.4 the projection $P_S(x^k)$ is single-valued. Since the sequence was arbitrarily chosen, it follows that the projection is single-valued on a neighborhood of $\overline{x}$, hence S is prox-regular. $\square$

4 Algorithms for Optimization with a Rank Constraint

The prox-regularity of the set S has a number of important implications regarding numerical algorithms. Principal among these is local linear convergence of the elementary alternating projection and steepest descent algorithms. There has been a tremendous number of articles published in recent years about convex (and nonconvex) relaxations of the rank function, and when the solution of optimization problems with respect to these relaxations corresponds to the optimization problem with the rank function (see the review article [23] and references therein). The motivation for such relaxations is that there are polynomial-time algorithms for the solution of the relaxed problems, while the rank minimization problem is NP-complete. As we will show in this section, the above theory implies that in the neighborhood of a solution there are polynomial-time algorithms for the solution of optimization problems with rank constraints. This observation was anticipated in [2] and notably [5] where a (globally) linearly convergent projected gradient algorithm with a rank constraint was presented. Without further assumptions, however, such assurances of convergence of algorithms for problems with rank constraints is at the cost of global guarantees of convergence.

4.1 Inexact, Extrapolated Alternating Projections

To the extent that the singular value decomposition can be computed exactly, the projection of a point x onto the rank lower level set S can be calculated exactly simply by ordering the singular values of x and truncating. The above analysis immediately yields local linear convergence of exact and inexact alternating projections for finding the intersection $S \cap M$ for M closed on neighborhoods of points where the intersection is strongly regular. The following algorithm allows for inexact evaluation of the fixed point operator, and hence implementable algorithms.

Algorithm 4.1 (inexact alternating projections [17]) Fix $\gamma > 0$ and choose $x^0 \in S$ and $x^1 \in M$. For $k = 1, 2, 3, \ldots$ generate the sequence (x^{2k}) in S with $x^{2k} \in P_S(x^{2k-1})$ where the sequence (x^{2k+1}) in M satisfies

$$\left\| x^{2k+1} - x^{2k} \right\| \leq \left\| x^{2k} - x^{2k-1} \right\|, \tag{4.1a}$$

$$x^{2k+1} = x^{2k} \quad \text{if } x_*^{2k+1} = x^{2k}, \tag{4.1b}$$

and

$$d_{N_M(x_*^{2k+1})}\left(\hat{z}^k\right) \leq \gamma \tag{4.1c}$$

for

$$x_*^{2k+1} = P_{M \cap \{x^{2k} - \tau \hat{z}^k, \ \tau \geq 0\}}\left(x^{2k}\right)$$

and

$$\hat{z}^k := \begin{cases} \frac{x^{2k} - x^{2k+1}}{\| x^{2k} - x^{2k+1} \|} & \text{if } x_*^{2k+1} \neq x^{2k} \\ 0 & \text{if } x_*^{2k+1} = x^{2k}. \end{cases}$$

For $\gamma = 0$ and $x^{2k+1} = x_*^{2k+1}$ the inexact algorithm reduces to the usual alternating projections algorithm. Note that the odd iterates x^{2k+1} can lie on the interior of M. This is the major difference between Algorithm 4.1 and the one specified in [13] where all of the iterates are assumed to lie on the boundary of M. We include this feature to allow for *extrapolated* iterates in the case where M has interior.

Theorem 4.2 (inexact alternating projections with a rank lower level set) *Let* $M, S \subset \mathbb{R}^{m \times n}$ *be closed with* $S := \{y \mid \mathrm{rank}(y) \leq s\}$ *and suppose there is an* $\overline{x} \in M \cap S$ *with* $\mathrm{rank}(\overline{x}) = s$. *Suppose furthermore that* M *and* S *have strongly regular intersection at* $\overline{x}$ *with angle* $\overline{\theta}$. *Define* $\overline{c} := \cos(\overline{\theta}) < 1$ *and fix the constants* $c \in (\overline{c}, 1)$ *and* $\gamma < \sqrt{1 - c^2}$. *For* x^0 *and* x^1 *close enough to* $\overline{x}$, *the iterates in Algorithm* 4.1 *converge to a point in* $M \cap S$ *with R-linear rate*

$$\sqrt{c\sqrt{1 - \gamma^2} + \gamma\sqrt{1 - c^2}} < 1.$$

If, in addition, M *is prox-regular at* $\overline{x}$, *then the iterates converge with rate*

$$c\sqrt{1 - \gamma^2} + \gamma\sqrt{1 - c^2} < 1.$$

Proof Since by Proposition 3.8 S is prox regular at $\overline{x}$ the results follow immediately from [17, Theorem 4.4]. □

Remark 4.3 The above result requires only closedness of the set M. For example, this yields convergence for affine sets $M = \{x \mid Ax = b\}$ which are not only closed, but convex. But the above result is not restricted to such nice sets. Another important example is inverse scattering with sparsity constraints [20]. Here the set M is $M = \{x \in \mathbb{C}^n \mid |(Fx)_j|^2 = b_j, \ j = 1, 2, \ldots, n\}$ where F is a linear mapping (the discrete Fourier or Fresnel transform) and b is some measurement (a far field intensity measurement). This set is not convex, but it is certainly closed (in fact prox-regular), so again, we can apply the above results to provide local guarantees of convergence for *nonconvex* alternating projections with a sparsity set.

Paradoxically, in the vector case it is the projection onto the affine constraint that in general cannot be evaluated exactly, while the projection onto the sparsity set S *can* be implemented exactly by simply (hard) thresholding the vectors. In the matrix case, this is no longer possible since in general the singular values cannot be evaluated exactly. In order to accommodate both projections being approximately

evaluated, we explore one possible solution using a common reformulation of the problem on a product space. This is explained next.

4.2 Approximate Steepest Descent

Another fundamental approach to solving such problems is simply to minimize the sum of the (squared) distances to the sets M and S:

$$\underset{x\in\mathbb{R}^{m\times n}}{\text{minimize}}\ \frac{1}{2}\big(d^2(x,S)+d^2(x,M)\big)$$

Steepest descent without line search is: given $x_0\in\mathbb{R}^{m\times n}$ generate the sequence $(x^k)_{k\in\mathbb{N}}$ in $\mathbb{R}^{m\times n}$ via

$$x^{k+1}=x^k-\nabla\frac{1}{2}\big(d^2(x^k,S)+d^2(x^k,M)\big).$$

If S and M were convex and the distance function the Euclidean distance, it is well-known that this would be equivalent to averaged projections:

$$\begin{aligned}x^{k+1}&=x^k-\nabla\frac{1}{2}\big(d^2(x^k,S)+d^2(x^k,M)\big)\\&=\frac{1}{2}\big(P_S(x^k)+P_M(x^k)\big).\end{aligned}\tag{4.2}$$

If we assume that M is prox-regular, then, since we have already established the prox-regularity of S, the correspondence between the derivative of the sum of squared distances to these sets and the projection operators in (4.2) holds on (common) open neighborhoods of M and S [22, Theorem 1.3]. Using a common product space formulation due to [21] we can show that (4.2) is equivalent to alternating projections between the sets

$$D:=\big\{(x,y)\in\mathbb{R}^{m\times n}\times\mathbb{R}^{m\times n}\ \big|\ x=y\big\}$$

and

$$\Omega:=\big\{(x,y)\in\mathbb{R}^{m\times n}\times\mathbb{R}^{m\times n}\ \big|\ x\in S,y\in M\big\},$$

that is,

$$\big(x^{k+1},x^{k+1}\big)=P_D\big(P_\Omega\big(\big(x^k,x^k\big)\big)\big)$$

where x^{k+1} is given by (4.2). The set Ω is prox-regular if M and S are, and the set D is convex, so Theorem 4.2 guarantees local linear convergence of the sequence of iterates $(x^k)_{k\in\mathbb{N}}$ with rate depending on the angle of strong intersection of the sets D and Ω. We cannot expect to be able to compute the projection onto the set Ω exactly, but we can reasonably assume to be able to compute the projection onto the diagonal D exactly, even if the magnitudes of the elements of $P_S(x^k)$ and $P_M(x^k)$ differ in orders of magnitude beyond our numerical precision. Indeed, since the projection

operators P_S and P_M are Lipschitz continuous for S and M prox-regular [22, Theorem 1.3], we can attribute any error we in fact make in the evaluation of P_D to the evaluation of P_Ω where we compute an approximation according to Algorithm 4.1. Again, Theorem 4.2 guarantees local linear convergence with rate governed by the angle of strong regularity between D and Ω and the accuracy of the approximate projection onto Ω.

5 Conclusion

We have developed a novel characterization of the normal cone to the lower level sets of the rank function. This enables us to obtain a simple proof of the prox-regularity of such sets. This property then allows for a straight-forward application of previous results on the local linear convergence of approximate alternating projections for finding the intersection of rank constrained sets and another closed set, as long as the intersection is *strongly regular* at a reference point $\overline{x}$. Our characterization of the normal cone to rank constraint sets allows for easy characterization and verification of the strong regularity of intersections of these sets with other sets. The results are also extended to the elementary steepest descent algorithm for minimizing the sum of squared distances to sets, one of which is a rank constraint set. This implies that, in the neighborhood of a solution with sufficient regularity, there are polynomial time algorithms for directly solving rank constraint problems without resorting to convex relaxations or heuristics.

What remains to be determined is the radius of convergence of these algorithms. Using the *restricted* normal cone developed in [4] Bauschke an coauthors [3] obtained linear rates with estimates of the radius of convergence of alternating projections applied to *affine* sparsity constrained problems—that is, the vector affine case of the setting considered here—assuming only existence of solutions. The restricted normal cone is not immediately applicable here since the restrictions in [3] are over countable collections of subspaces representing all possible s-sparse vectors. For the rank function this is problematic since the collection of all possible s-rank matrices is not countable. Extending the tools of [4] to the matrix case is the focus of future research.

The results one might obtain using the tools of [4] or similar, however, are based on the regularity near the solution, what we call *micro*-regularity. We cannot expect the estimates for the radius of convergence to extend very far using these tools, unless certain local-to-global properties like convexity are assumed. In [5] a scalable restricted isometry property is used to prove global convergence of a projected gradient algorithm to the unique solution to the problem of minimizing the distance to an affine subspace subject to a rank constraint. The (scalable) restricted isometry

property and other properties like it (mutual coherence, etc.) directly concern uniqueness of solutions and indirectly provide sufficient conditions for global convergence of algorithms for solving relaxations of the original sparsity/rank optimization problem. A natural question is whether there is a more general *macro-regularity* property than the scalable restricted isometry property, one independent of considerations of uniqueness of solutions, that guarantees global convergence.

References

1. Arnol'd, V.I.: On matrices depending on parameters. Translation from Usp. Mat. Nauk 26, No. 2(158), 101–114 (1971)
2. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Lojasiewicz inequality. Math. Oper. Res. **35**(2), 438–457 (2010)
3. Bauschke, H.H., Luke, D.R., Phan, H.M., Wang, X.: Restricted normal cones and sparsity optimization with affine constraints (2012). arXiv:1205.0320
4. Bauschke, H.H., Luke, D.R., Phan, H.M., Wang, X.: Restricted normal cones and the method of alternating projections (2012). arXiv:1205.0318
5. Beck, A., Teboulle, M.: A linearly convergent algorithm for solving a class of nonconvex/affine feasibility problems. In: Bauschke, H., Burachick, R., Combettes, P., Elser, V., Luke, D.R., Wolkowicz, H. (eds.) Fixed-Point Algorithms for Inverse Problems in Science and Engineering, pp. 33–48. Springer, Berlin (2011)
6. Combettes, P.L., Trussell, H.J.: Method of successive projections for finding a common point of sets in metric spaces. J. Optim. Theory Appl. **67**(3), 487–507 (1990)
7. Federer, H.: Curvature measures. Trans. Am. Math. Soc. **93**, 418–491 (1959)
8. Hiriart-Urruty, J.-B.: When only global optimization matters. J. Glob. Optim. (2012). doi:10.1007/s10898-011-9826-7
9. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, New York (1985)
10. Kruger, A.Y.: Stationarity and regularity of set systems. Pac. J. Optim. **1**(1), 101–126 (2005)
11. Le, H.Y.: Generalized subdifferentials of the rank function. Optim. Lett. (2012). doi:10.1007/s11590-012-0456-x
12. Lewis, A.S.: Eigenvalues and nonsmooth optimization. In: Pardo, L.M., et al. (eds.) Foundations of Computational Mathematics, Santander (2005). Selected papers based on the presentations at the International Conference of the Foundations of Computational Mathematics (FoCM), Santander, Spain, June 30–July 9, 2005. Cambridge: Cambridge University Press. London Mathematical Society Lecture Note Series, vol. 331, pp. 208–229 (2006)
13. Lewis, A.S., Luke, D.R., Malick, J.: Local linear convergence of alternating and averaged projections. Found. Comput. Math. **9**(4), 485–513 (2009)
14. Lewis, A.S., Malick, J.: Alternating projections on manifolds. Math. Oper. Res. **33**, 216–234 (2008)
15. Lewis, A.S., Sendov, H.S.: Nonsmooth analysis of singular values. I: Theory. Set-Valued Anal. **13**(3), 213–241 (2005)
16. Lewis, A.S., Sendov, H.S.: Nonsmooth analysis of singular values. II: Applications. Set-Valued Anal. **13**(3), 243–264 (2005)
17. Luke, D.R.: Local linear convergence of approximate projections onto regularized sets. Nonlinear Anal. **75**, 1531–1546 (2012)
18. Mordukhovich, B.S.: Nonsmooth analysis with nonconvex generalized differentials and adjoint mappings. Dokl. Akad. Nauk BSSR **28**, 976–979 (1984). Russian
19. Mordukhovich, B.S.: Variational Analysis and Generalized Differentiation, I: Basic Theory; II: Applications. Grundlehren der mathematischen Wissenschaften. Springer, New York (2006)
20. Oszlányi, G., Sütó, A.: The charge flipping algorithm. Acta Crystalogr. A **64**, 123–134 (2008)
21. Pierra, G.: Eclatement de Contraintes en Parallèle Pour la Minimisation D'une Forme Quadratique. Lecture Notes in Computer Science, vol. 41, pp. 200–218 (1976)
22. Poliquin, R.A., Rockafellar, R.T., Thibault, L.: Local differentiability of distance functions. Trans. Am. Math. Soc. **352**(11), 5231–5249 (2000)
23. Recht, B., Fazel, M., Parrilo, P.: Guaranteed minimum rank solutions of matrix equations via nuclear norm minimization. SIAM Rev. **52**(3), 471–501 (2010)
24. Rockafellar, R.T., Wets, R.J.: Variational Analysis. Grundlehren der mathematischen Wissenschaften. Springer, Berlin (1998)

D. Russell Luke received a B.A. cum laude in 1991 from the University of California, Berkeley and M.Sc. (1997) and Ph.D. (2001) from the University of Washington in Seattle. He was a postdoctoral researcher at Universität Göttingen (2001–2003) and a PIMS Fellow at Simon Fraser University (2002–2004) before joining the faculty at the University of Delaware where he received tenure and promotion to Associate Professor (2009). He was appointed Professor of Continuous Optimization at the Universität Göttingen in 2010.