



HAL
open science

Low-rank optimization methods based on projected-projected gradient descent that accumulate at Bouligand stationary points

Guillaume Olikier, Kyle A. Gallivan, P. -A. Absil

► **To cite this version:**

Guillaume Olikier, Kyle A. Gallivan, P. -A. Absil. Low-rank optimization methods based on projected-projected gradient descent that accumulate at Bouligand stationary points. 2024. hal-04865950

HAL Id: hal-04865950

<https://inria.hal.science/hal-04865950v1>

Preprint submitted on 6 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Low-rank optimization methods based on projected-projected gradient descent that accumulate at Bouligand stationary points*

Guillaume Olikier[†]Kyle A. Gallivan[‡]P.-A. Absil[§]

July 1, 2024

Abstract

This paper considers the problem of minimizing a differentiable function with locally Lipschitz continuous gradient on the algebraic variety of real matrices of upper-bounded rank. This problem is known to enable the formulation of several machine learning and signal processing tasks such as collaborative filtering and signal recovery. Several definitions of stationarity exist for this nonconvex problem. Among them, Bouligand stationarity is the strongest first-order necessary condition for local optimality. This paper proposes a first-order algorithm that combines the well-known projected-projected gradient descent map with a rank reduction mechanism and generates a sequence in the variety whose accumulation points are Bouligand stationary. This algorithm compares favorably with the three other algorithms known in the literature to enjoy this stationarity property, regarding both the typical computational cost per iteration and empirically observed numerical performance. A framework to design hybrid algorithms enjoying the same property is proposed and illustrated through an example.

Keywords: convergence analysis, stationary point, critical point, low-rank optimization, determinantal variety, steepest descent, tangent and normal cones, Clarke regularity.

Mathematics Subject Classification: 14M12, 65K10, 90C26, 90C30, 40A05.

1 Introduction

Low-rank optimization concerns the problem of minimizing a real-valued function over a space of matrices subject to an upper bound on their rank. Applications abound in various fields of science and engineering. The present work addresses specifically the problem

$$\min_{X \in \mathbb{R}_{\leq r}^{m \times n}} f(X) \quad (1.1)$$

of minimizing a differentiable function $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ with locally Lipschitz continuous gradient on the determinantal variety

$$\mathbb{R}_{\leq r}^{m \times n} := \{X \in \mathbb{R}^{m \times n} \mid \text{rank } X \leq r\}, \quad (1.2)$$

m , n , and r being positive integers such that $r < \min\{m, n\}$.

*This work was supported by the Fonds de la Recherche Scientifique – FNRS and the Fonds Wetenschappelijk Onderzoek – Vlaanderen under EOS Project no 30468160, by the Fonds de la Recherche Scientifique – FNRS under Grant no T.0001.23, by the ERC grant #786854 G-Statistics from the European Research Council under the European Union’s Horizon 2020 research and innovation program, and by the French government through the 3IA Côte d’Azur Investments ANR-19-P3IA-0002 managed by the National Research Agency. K. A. Gallivan was partially supported by the U.S. National Science Foundation under grant CIBR 1934157. This work was done in part while K. A. Gallivan was visiting UCLouvain, supported by the “Professeurs et chercheurs visiteurs” budget of the Science and Technology Sector.

[†]Université Côte d’Azur and Inria, Epione Project Team, 2004 route des Lucioles - BP 93, 06902 Sophia Antipolis Cedex, France (guillaume.olikier@inria.fr).

[‡]Department of Mathematics, Florida State University, 1017 Academic Way, Tallahassee, FL 32306-4510, USA (kgallivan@fsu.edu).

[§]ICTEAM Institute, UCLouvain, Avenue Georges Lemaître 4, 1348 Louvain-la-Neuve, Belgium (pa.absil@uclouvain.be).

Determinantal varieties constitute an important topic in algebraic geometry, especially over algebraically closed fields such as the field of complex numbers; see, e.g., [3, Chapter II], [5, §1C], [20, Lecture 9], and [27, Chapter 10]. Their geometry is well understood, which is particularly useful in the context of problem (1.1). For example, the smooth and singular parts of $\mathbb{R}_{\leq r}^{m \times n}$ are known to be

$$\mathbb{R}_r^{m \times n} := \{X \in \mathbb{R}^{m \times n} \mid \text{rank } X = r\}$$

and

$$\mathbb{R}_{< r}^{m \times n} := \{X \in \mathbb{R}^{m \times n} \mid \text{rank } X < r\},$$

respectively, and explicit formulas for the tangent cone to $\mathbb{R}_{\leq r}^{m \times n}$ are available.

Problem (1.1) includes as special cases many important problems such as weighted low-rank approximation [53, 15], robust affine rank minimization [24, 16, 25, 61, 55, 6], various formulations of matrix completion [54, 65, 13, 63, 38, 14], multi-task learning [1], matrix sensing [7, 66], robust PCA [19], and logistic PCA [46]. These problems are known to enable the formulation of several machine learning and signal processing tasks such as collaborative filtering and signal recovery.

1.1 Stationarity notions on the determinantal variety

In general, problem (1.1) is intractable [15] and algorithms are only expected to find a stationary point of this problem [52, 56, 64, 22, 34, 19, 35, 58, 26, 29, 33, 40, 43, 30, 47, 44, 49]. A point $X \in \mathbb{R}_{\leq r}^{m \times n}$ is said to be stationary for (1.1) if $-\nabla f(X)$ is normal to $\mathbb{R}_{\leq r}^{m \times n}$ at X . Several definitions of normality exist. Each one yields a definition of stationarity such that, possibly under mild regularity assumptions on f , every local minimizer of $f|_{\mathbb{R}_{\leq r}^{m \times n}}$ is stationary for (1.1). In particular, each of the three notions of normality in [50, Definition 6.3 and Example 6.16], namely normality in the general sense, in the regular sense, and in the proximal sense, yields an important definition of stationarity. On $\mathbb{R}_{\leq r}^{m \times n}$, normality in the proximal sense is equivalent to normality in the regular sense [44, Proposition 7.1]. The sets of general and regular normals to $\mathbb{R}_{\leq r}^{m \times n}$ at $X \in \mathbb{R}_{\leq r}^{m \times n}$ are closed cones called the *normal cone* to $\mathbb{R}_{\leq r}^{m \times n}$ at X and the *regular normal cone* to $\mathbb{R}_{\leq r}^{m \times n}$ at X and denoted by $N_{\mathbb{R}_{\leq r}^{m \times n}}(X)$ and $\widehat{N}_{\mathbb{R}_{\leq r}^{m \times n}}(X)$, respectively. These normal cones are reviewed in Section 2.3. Importantly, they are nested as follows: for every $X \in \mathbb{R}_{\leq r}^{m \times n}$,

$$\widehat{N}_{\mathbb{R}_{\leq r}^{m \times n}}(X) \subseteq N_{\mathbb{R}_{\leq r}^{m \times n}}(X), \quad (1.3)$$

and $\mathbb{R}_{\leq r}^{m \times n}$ is said to be *Clarke regular* at X if the inclusion is an equality [50, Definition 6.4]. As observed in [52, Remark 4], $\mathbb{R}_{\leq r}^{m \times n}$ is Clarke regular at $X \in \mathbb{R}_{\leq r}^{m \times n}$ if and only if $\text{rank } X = r$.

The definitions of stationarity based on the normal cone and the regular normal cone are given in Definition 1.1; comments on the terminology can be found, e.g., in [44, §3].

Definition 1.1. For problem (1.1), a point $X \in \mathbb{R}_{\leq r}^{m \times n}$ is said to be:

- *Mordukhovich stationary (M-stationary)* if $-\nabla f(X) \in N_{\mathbb{R}_{\leq r}^{m \times n}}(X)$;
- *Bouligand stationary (B-stationary)* if $-\nabla f(X) \in \widehat{N}_{\mathbb{R}_{\leq r}^{m \times n}}(X)$.

These two notions of stationarity depend on f only through its restriction $f|_{\mathbb{R}_{\leq r}^{m \times n}}$ (see [29] for B-stationarity and Proposition 2.9 for M-stationarity); this is a desirable property for a notion of stationarity. They are equivalent only on the smooth part $\mathbb{R}_r^{m \times n}$ of $\mathbb{R}_{\leq r}^{m \times n}$. For problem (1.1), B-stationarity is the strongest first-order necessary condition for local optimality, as shown by the following gradient characterization that holds at every $X \in \mathbb{R}_{\leq r}^{m \times n}$ [44, Theorem 2.5 and Proposition 7.1]:

$$\widehat{N}_{\mathbb{R}_{\leq r}^{m \times n}}(X) = \left\{ -\nabla h(X) \mid \begin{array}{l} h : \mathbb{R}^{m \times n} \rightarrow \mathbb{R} \text{ is differentiable,} \\ \nabla h \text{ is locally Lipschitz continuous,} \\ X \text{ is a local minimizer of } h|_{\mathbb{R}_{\leq r}^{m \times n}} \end{array} \right\}. \quad (1.4)$$

B-stationarity is the desired notion of stationarity when problem (1.1) is the main problem to be solved, and not merely a subproblem to be solved at every iteration of an algorithm like the augmented Lagrangian method proposed in [26, Algorithm 4.1].

In comparison, M-stationarity is a weaker notion of stationarity which is considered unsatisfactory in [22, §4], [29, §1], and [47, §2.1]. This is illustrated in Example 1.2.

Example 1.2 (M-stationarity without B-stationarity [47, Example 1]). Consider the instance of (1.1) where $m := n := 2$, $r := 1$, and $f(X) := \frac{1}{2}\|X - \text{diag}(1, 0)\|^2$ for all $X \in \mathbb{R}^{2 \times 2}$. As detailed in Section 2.3, $\{0_{2 \times 2}\} = \widehat{N}_{\mathbb{R}_{\leq 1}^{2 \times 2}}(0_{2 \times 2}) \subsetneq N_{\mathbb{R}_{\leq 1}^{2 \times 2}}(0_{2 \times 2}) = \mathbb{R}_{\leq 1}^{2 \times 2}$. Hence, $\mathbb{R}_{\leq 1}^{2 \times 2}$ is not Clarke regular at $0_{2 \times 2}$. Furthermore, $\nabla f(0_{2 \times 2}) = -\text{diag}(1, 0)$. Thus, since $-\nabla f(0_{2 \times 2}) \in N_{\mathbb{R}_{\leq 1}^{2 \times 2}}(0_{2 \times 2}) \setminus \widehat{N}_{\mathbb{R}_{\leq 1}^{2 \times 2}}(0_{2 \times 2})$, $0_{2 \times 2}$ is M-stationary but not B-stationary. Moreover, for all $\alpha \in (0, \infty)$, $0_{2 \times 2} - \alpha \nabla f(0_{2 \times 2}) = \text{diag}(\alpha, 0) \in \mathbb{R}_{\leq 1}^{2 \times 2}$. Therefore, moving from $0_{2 \times 2}$ in the direction of $-\nabla f(0_{2 \times 2})$ enables to decrease f while staying in $\mathbb{R}_{\leq 1}^{2 \times 2}$. This is illustrated in Figure 1.1.

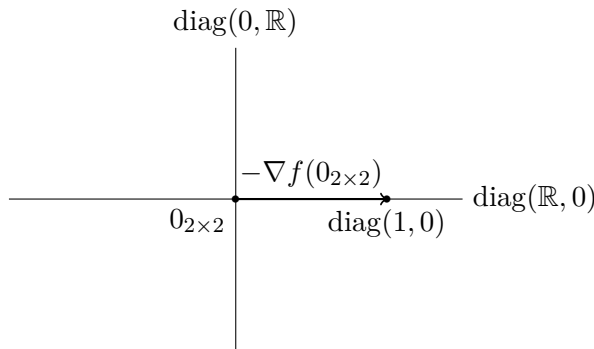


Figure 1.1: Illustration of Example 1.2 in the plane of diagonal matrices in $\mathbb{R}^{2 \times 2}$.

Furthermore, as explained in [29], distinguishing convergence to a B-stationary point from convergence to an M-stationary point is difficult in the sense that it cannot be done based on standard measures of B-stationarity, which are zero only at B-stationary points, such as

$$\mathbb{R}_{\leq r}^{m \times n} \rightarrow \mathbb{R} : X \mapsto \|P_{T_{\mathbb{R}_{\leq r}^{m \times n}}(X)}(-\nabla f(X))\|, \quad (1.5)$$

which returns the norm of any projection of $-\nabla f(X)$ onto the tangent cone to $\mathbb{R}_{\leq r}^{m \times n}$ at X (see (2.7)), and

$$\mathbb{R}_{\leq r}^{m \times n} \rightarrow \mathbb{R} : X \mapsto d(-\nabla f(X), \widehat{N}_{\mathbb{R}_{\leq r}^{m \times n}}(X)), \quad (1.6)$$

which returns the distance from $-\nabla f(X)$ to $\widehat{N}_{\mathbb{R}_{\leq r}^{m \times n}}(X)$. Indeed, these stationarity measures can fail to be lower semicontinuous at a point where $\mathbb{R}_{\leq r}^{m \times n}$ is not Clarke regular (Proposition 2.10). As a matter of fact, when applied to some instances of problem (1.1), optimization algorithms such as [52, Algorithms 3 and 4], which are reviewed in Section 1.2, generate a convergent sequence in the smooth part $\mathbb{R}_r^{m \times n}$ of $\mathbb{R}_{\leq r}^{m \times n}$ along which the measures of B-stationarity defined in (1.5) and (1.6) converge to zero and whose limit, which is in the singular part $\mathbb{R}_{< r}^{m \times n}$ of $\mathbb{R}_{\leq r}^{m \times n}$, is M-stationary but not B-stationary, a phenomenon formalized by the notion of *apocalypse* in [29]. In contrast, the measure of M-stationarity

$$\mathbb{R}_{\leq r}^{m \times n} \rightarrow \mathbb{R} : X \mapsto d(-\nabla f(X), N_{\mathbb{R}_{\leq r}^{m \times n}}(X)) \quad (1.7)$$

is lower semicontinuous (Proposition 2.12). Therefore, if it goes to zero along a convergent sequence in $\mathbb{R}_{\leq r}^{m \times n}$, then it is zero at the limit.

The coexistence of several notions of stationarity for an optimization problem is common. It arises notably in nonsmooth optimization, where several notions of subdifferential yield different notions of stationarity [32, 45, 9], and in mathematical programs with complementarity constraints, for which several notions of stationarity are presented and compared in [51].

1.2 State of the art

Five state-of-the-art algorithms aiming at solving problem (1.1) are listed in Table 1.1 (Section 1.3) and described below. Among them, PGD, RFDR, and [29, Algorithm 1]—which we term Hooked Riemannian Trust-Region (HRTR)—are the only three algorithms in the literature guaranteed to accumulate at B-stationary points of (1.1), which means that the accumulation points of every generated sequence are B-stationary for (1.1); the property of generating a sequence whose accumulation points are stationary (in some sense) is called “(global) subsequential convergence” to stationary points in [31, Theorem 2(i)], [60, Theorem 1], [11, Theorem 1], [9, Theorem 7.2.4], [4, Theorem 4.5], and [8, Lemma 4.4]. The other two, namely P²GD and RFD, can be used as subroutines to design first-order algorithms provably accumulating at B-stationary points of (1.1).

HRTR is a second-order method that deals with problem (1.1) by *lifting* it to a Riemannian manifold \mathcal{M} : given a smooth map $\varphi : \mathcal{M} \rightarrow \mathbb{R}^{m \times n}$ such that $\varphi(\mathcal{M}) = \mathbb{R}_{\leq r}^{m \times n}$, it minimizes $f \circ \varphi$ on \mathcal{M} ; φ is called a *smooth lift* of $\mathbb{R}_{\leq r}^{m \times n}$. This method takes advantage of the increasingly understood relations between the desirable points of the original problem and those of the lifted problem [19, 29, 30, 49]. This approach is appealing for its generality and elegance: it is applicable to every feasible set for which a suitable smooth lift is available and it generates iterates that depend on the cost function only through its restriction to the feasible set. However, on $\mathbb{R}_{\leq r}^{m \times n}$, it suffers from a prohibitively high computational cost in the frequently encountered case where $r \ll \min\{m, n\}$. Indeed, as explained in Section 8.1, every iteration of HRTR, used with the rank factorization lift $\varphi : \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r} \rightarrow \mathbb{R}^{m \times n} : (L, R) \mapsto LR^\top$, requires computing the smallest eigenvalue of an order- $(m+n)r$ matrix representing the Hessian of the lifted cost function $f \circ \varphi$. Furthermore, in the numerical experiments reported in Section 8.2 and based on basic implementations of PGD, RFDR, and HRTR used with the aforementioned lift, the former outperform the latter.

PGD, which is short for *projected gradient descent*, is appealing for its simplicity: given $X \in \mathbb{R}_{\leq r}^{m \times n}$ as input, the PGD map—“map” is added to the name of an algorithm to refer to its iteration map—performs a projected line search along the direction of $-\nabla f(X)$, i.e., computes a projection Y of $X - \alpha \nabla f(X)$ onto $\mathbb{R}_{\leq r}^{m \times n}$ for decreasing values of the step size $\alpha \in (0, \infty)$ until Y satisfies an Armijo condition. In the simplest version of PGD, called *monotone*, the Armijo condition ensures that the value of f at the next iterate is smaller by a specified amount than the value at the current iterate. Following the general setting proposed in [18], the value at the current iterate can be replaced with the maximum value of f over a prefixed number of the previous iterates. This version of PGD is called *nonmonotone*. This description of PGD brings to light its main drawback, already identified in [61, §2.1] and [57, §9.4.4], namely its computational cost per iteration which can be prohibitively high if $r \ll \min\{m, n\}$: if the gradient does not have a special structure such as sparse or low rank, which is the case, e.g., in weighted low-rank approximation [15], then every iteration requires projecting a possibly full-rank matrix onto $\mathbb{R}_{\leq r}^{m \times n}$ which amounts to computing a truncated singular value decomposition (SVD) of this matrix. Furthermore, as pointed out in [29], the iterates generated by PGD depend on the values taken by f outside $\mathbb{R}_{\leq r}^{m \times n}$ which are irrelevant to problem (1.1).

P²GD, which is short for *projected-projected gradient descent*, works like PGD except that it involves an additional projection: given $X \in \mathbb{R}_{\leq r}^{m \times n}$ as input, the P²GD map performs a projected line search along a projection G of $-\nabla f(X)$ onto $T_{\mathbb{R}_{\leq r}^{m \times n}}(X)$, i.e., computes a projection Y of $X + \alpha G$ onto $\mathbb{R}_{\leq r}^{m \times n}$ for decreasing values of $\alpha \in (0, \infty)$ until Y satisfies an Armijo condition. It follows from the structure of $T_{\mathbb{R}_{\leq r}^{m \times n}}(X)$ that $X + \alpha G \in \mathbb{R}_{\leq 2r}^{m \times n}$ for all $\alpha \in (0, \infty)$ [52, §3.4]. This makes the projection onto $\mathbb{R}_{\leq r}^{m \times n}$ easier to compute. Moreover, as pointed out in [29], the iterates generated by P²GD depend on f only through its restriction $f|_{\mathbb{R}_{\leq r}^{m \times n}}$.

RFD, which is short for *retraction-free descent*, relies on the *restricted tangent cone* to $\mathbb{R}_{\leq r}^{m \times n}$ (see [40, §3] and [52, §3.4]): there exists a map that associates with every $X \in \mathbb{R}_{\leq r}^{m \times n}$ a closed cone $\check{T}_{\mathbb{R}_{\leq r}^{m \times n}}(X)$ contained in $T_{\mathbb{R}_{\leq r}^{m \times n}}(X)$ such that:

1. for all $Z \in \check{T}_{\mathbb{R}_{\leq r}^{m \times n}}(X)$, $X + Z \in \mathbb{R}_{\leq r}^{m \times n}$;

2. for all $Z \in \mathbb{R}^{m \times n}$, $\|P_{\tilde{T}_{\mathbb{R}_{\leq r}^{m \times n}}(X)}(Z)\| \geq \frac{1}{\sqrt{2}} \|P_{T_{\mathbb{R}_{\leq r}^{m \times n}}(X)}(Z)\|$.

Given $X \in \mathbb{R}_{\leq r}^{m \times n}$ as input, the RFD map performs a line search along a projection G of $-\nabla f(X)$ onto $\tilde{T}_{\mathbb{R}_{\leq r}^{m \times n}}(X)$, which requires no retraction on $\mathbb{R}_{\leq r}^{m \times n}$ (in the sense of [23, §3.1.2]) since $X + \alpha G \in \mathbb{R}_{\leq r}^{m \times n}$ for all $\alpha \in (0, \infty)$; the RFD map is thus retraction-free, i.e., it performs updates along straight lines. However, as pointed out in [52, §3.4], saving the cost of computing a retraction does not confer to RFD a significant advantage over P²GD if $r \ll \min\{m, n\}$ since the latter requires projecting a point in $\mathbb{R}_{\leq 2r}^{m \times n}$ onto $\mathbb{R}_{\leq r}^{m \times n}$, and reducing the rank from $2r$ to r is typically much less expensive than evaluating f or ∇f . Furthermore, the directions in $P_{T_{\mathbb{R}_{\leq r}^{m \times n}}(X)}(-\nabla f(X))$ are more closely related to $-\nabla f(X)$ than those in $P_{\tilde{T}_{\mathbb{R}_{\leq r}^{m \times n}}(X)}(-\nabla f(X))$; while this does not imply that P²GD converges faster than RFD, such an observation was made experimentally in [52, §3.4].

RFDR is RFD equipped with a rank reduction mechanism, hence the ‘‘R’’ in the name. Given $X \in \mathbb{R}_{\leq r}^{m \times n}$ as input, the RFDR map proceeds as follows: (i) it applies the RFD map to X , thereby producing a point \tilde{X} , (ii) if the r th singular value of X is positive but smaller than some threshold $\Delta \in (0, \infty)$, it applies the RFD map to a projection \hat{X} of X onto $\mathbb{R}_{r-1}^{m \times n}$, then producing a point \tilde{X}^R , and (iii) it outputs a point among \tilde{X} and \tilde{X}^R that maximally decreases f . In particular, it has a lower typical cost per iteration than the two other methods known to accumulate at B-stationary points of (1.1), namely PGD and HRTR. Furthermore, RFDR offers a performance comparable to that of PGD in the numerical experiment described in Section 8.2.

1.3 Contribution

This paper proposes a first-order algorithm, called P²GDR, that generates a sequence in $\mathbb{R}_{\leq r}^{m \times n}$ whose accumulation points are B-stationary for (1.1) in the sense of Definition 1.1. P²GDR is P²GD equipped with a rank reduction mechanism, hence the ‘‘R’’ in the name, that is based on the Δ -rank of a matrix: for every $\Delta \in (0, \infty)$, the Δ -rank of $X \in \mathbb{R}^{m \times n}$, denoted by $\text{rank}_{\Delta} X$, is defined as the number of singular values of X that are larger than Δ [17, (5.4.5)]. Given $X \in \mathbb{R}_{\leq r}^{m \times n}$ as input, the P²GDR map (Algorithm 6.1) proceeds as follows: (i) for every $i \in \{0, \dots, \text{rank} X - \text{rank}_{\Delta} X\}$, it applies the P²GD map (Algorithm 5.1) to a projection \hat{X}^i of X onto $\mathbb{R}_{\text{rank} X - i}^{m \times n}$, thereby producing a point \tilde{X}^i , and (ii) it outputs a point among $\tilde{X}^0, \dots, \tilde{X}^{\text{rank} X - \text{rank}_{\Delta} X}$ that maximally decreases f . P²GDR is compared in Table 1.1 with the five algorithms reviewed in Section 1.2.

<i>Algorithm</i>	<i>Order</i>	$s(\cdot; f, \mathbb{R}_{\leq r}^{m \times n}) \rightarrow 0$	B-stationary
PGD [26, Algorithm 3.1]	1	✓	✓ [44, Theorem 1.2]
P ² GD [52, Algorithm 3]	1	?	✗ [29, §2.2]
RFD [52, Algorithm 4]	1	✓ [52, Theorem 3.10]	✗ [40, §8]
P ² GDR (Algorithm 6.2)	1	✓	✓ (Theorem 6.2)
RFDR [40, Algorithm 3]	1	✓	✓ [40, Theorem 6.2]
HRTR [29, Algorithm 1]	2	✓	✓ [29, Theorem 1.1]

Table 1.1: Comparison of P²GDR with five state-of-the-art algorithms aiming at solving problem (1.1). The property ‘‘ $s(\cdot; f, \mathbb{R}_{\leq r}^{m \times n}) \rightarrow 0$ ’’ means that the measure of B-stationarity defined in (1.5) goes to zero along every convergent subsequence of the generated sequence. The property ‘‘B-stationary’’ means that the algorithm accumulates at B-stationary points of (1.1), which is stronger (Proposition 2.13). The symbols ‘‘✓’’, ‘‘✗’’, and ‘‘?’’ respectively mean ‘‘yes’’, ‘‘no’’, and ‘‘open question’’.

Preliminary work leading to the current paper can be found in the technical report [41] where the P²GDR algorithm was first made public. The main novelties are: Propositions 2.9, 2.10, and 2.12 which are mentioned in Section 1.1; Section 3 which presents the theoretical framework used to analyze the proposed P²GDR; Section 7 which builds on the framework from Section 3 to propose a hybrid algorithm based on P²GD and PGD and accumulating at B-stationary points of (1.1); Section 8 which

compares P²GDR, the hybrid algorithm, and the state-of-the-art algorithms listed in Table 1.1, and contains elements from the technical report [42].

When it became public in [41], P²GDR was the first method to answer positively the following question raised in the preliminary version arXiv:2107.03877v1 of [29]: “Is there an algorithm running directly on $\mathbb{R}_{\leq r}^{m \times n}$ that only uses first-order information about the cost function and which is guaranteed to converge to a B-stationary point?” Since then, as mentioned above, RFDR was proposed and PGD was proven to also positively answer the question. However, P²GDR still compares favorably with those methods (and thus the second-order HRTR, as explained in Section 1.2), in the following sense.

P²GDR has two advantages over RFDR which correspond to those that P²GD has over RFD. First, it can be defined on feasible sets for which no restricted tangent cone is known such as the subset of $\mathbb{R}_{\leq r}^{n \times n}$ containing the symmetric positive-semidefinite matrices. Second, it can be expected to converge faster than RFDR on the instances of (1.1) on which P²GD is known to converge faster than RFD such as those studied in [52, §3.4]. The price to pay to enjoy these advantages is the number of rank reduction attempts that can be required by the rank reduction mechanism. While this number is at most one for RFDR, it can be as large as r for P²GDR in the worst case. This worst case, however, should be rather infrequent in practice if Δ is chosen reasonably small (see Section 6.2).

The main advantage that P²GDR has over PGD is the typical cost per iteration. While, as pointed out in Section 1.2, PGD requires projecting a possibly full-rank matrix onto $\mathbb{R}_{\leq r}^{m \times n}$ at every iteration, an iteration of P²GDR where the rank reduction mechanism does not take action (which is typical in practice if Δ is chosen reasonably small) amounts to an iteration of P²GD. Furthermore, as explained in [59, Algorithm 6], [52, §3.4], [61, §2.3], and [57, §9.2.4], if the input has rank r , the most expensive rank reduction required by the P²GD map is the projection onto $\mathbb{R}_{\leq r}^{m \times n}$ of a matrix whose smallest dimension is at most $2r$. The algorithms listed in Table 1.1 are compared based on the computational cost per iteration in Section 8.1.

1.4 Organization of the paper

This paper is organized as follows. Section 2 reviews elements of variational analysis that are used throughout the paper. Section 3 presents the theoretical framework used to analyze P²GDR in Section 6 and a hybrid algorithm in Section 7. In Section 4, an upper bound on the distance to $\mathbb{R}_{\leq r}^{m \times n}$ from each of its tangent lines is derived. This upper bound is used in Section 5 to obtain a lower bound on the step size required by the P²GD map to ensure an Armijo condition. This lower bound plays an instrumental role in the analysis of P²GDR in Section 6. Section 7 defines a hybrid algorithm based on P²GD and PGD and guaranteed to accumulate at B-stationary points of (1.1). P²GDR, the hybrid algorithm, and the state-of-the-art algorithms listed in Table 1.1 are compared in Section 8. Section 9 gathers concluding remarks. Appendix A provides implementation details for HRTR.

2 Elements of variational analysis

For the convenience of the reader, this section reviews background material from variational analysis that is used throughout the paper. The only new results are Proposition 2.6, Lemma 2.8, Propositions 2.9 and 2.10, Corollary 2.11, and Propositions 2.12 and 2.13. Propositions 2.1 and 2.5 state basic facts and are proven for completeness. Section 2.1 is about cones, polarity, and projection. Section 2.2 concerns the concepts of outer limit and outer semicontinuity of a correspondence. Section 2.3 reviews tangent and normal cones. Section 2.4 focuses on the two notions of stationarity given in Definition 1.1 and the stationarity measures defined in (1.5)–(1.7).

The material is presented for a general Euclidean vector space \mathcal{E} with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$. For every $x \in \mathcal{E}$ and $\rho \in (0, \infty)$, $B(x, \rho) := \{y \in \mathcal{E} \mid \|x - y\| < \rho\}$ and $B[x, \rho] := \{y \in \mathcal{E} \mid \|x - y\| \leq \rho\}$ are the open and closed balls of center x and radius ρ in \mathcal{E} . A subset S of \mathcal{E} is said to be *locally closed* at $x \in \mathcal{E}$ if there exists $\rho \in (0, \infty)$ such that $S \cap B[x, \rho]$ is closed [50, §1I].

In Sections 2.1, 2.3, and 2.4, specific facts are also provided about the case of interest in this work, namely $\mathcal{E} = \mathbb{R}^{m \times n}$ with the Frobenius inner product $(X, Y) \mapsto \text{tr } Y^\top X$. For every $p, q \in \mathbb{N}$, $0_{p \times q}$ is

the zero matrix in $\mathbb{R}^{p \times q}$, $0_p := 0_{p \times p}$, I_p is the identity matrix in $\mathbb{R}^{p \times p}$, if $q \geq p$, then

$$\text{St}(p, q) := \{U \in \mathbb{R}^{q \times p} \mid U^\top U = I_p\}$$

is a Stiefel manifold, and $O(p) := \text{St}(p, p)$ is an orthogonal group [2, §3.3.2].

2.1 Cone, polarity, and projection

A nonempty subset K of \mathcal{E} is called a *cone* if $x \in K$ implies $\alpha x \in K$ for all $\alpha \in [0, \infty)$ [50, §3B]. For every cone $K \subseteq \mathcal{E}$, the set

$$K^* := \{w \in \mathcal{E} \mid \langle v, w \rangle \leq 0 \forall v \in K\}$$

is a closed convex cone called the *polar* of K [50, 6(14)].

For every nonempty subset S of \mathcal{E} and every $x \in \mathcal{E}$, $d(x, S) := \inf_{y \in S} \|x - y\|$ is the distance from x to S , and $P_S(x) := \text{argmin}_{y \in S} \|x - y\|$ is the projection of x onto S ; the set $P_S(x)$ is nonempty and compact if S is closed [50, Example 1.20]. In what follows, a singleton is sometimes identified with its unique element.

Proposition 2.1 is invoked in the proof of Lemma 6.1.

Proposition 2.1. *Let S be a nonempty subset of \mathcal{E} . For all $x \in S$ and $y \in \mathcal{E}$, $P_S(y) \subseteq B[x, 2\|x - y\|]$.*

Proof. Let $x \in S$ and $y \in \mathcal{E}$. The inclusion is trivially true if $P_S(y)$ is empty. If $P_S(y)$ is nonempty, then, for all $z \in P_S(y)$, $\|x - z\| \leq \|x - y\| + \|y - z\| = \|x - y\| + d(y, S) \leq 2\|x - y\|$. \square

Proposition 2.2 shows that the projection onto a closed cone enjoys an orthogonality property.

Proposition 2.2 ([29, Proposition A.6]). *Let $K \subseteq \mathcal{E}$ be a closed cone. For all $x \in \mathcal{E}$ and $y \in P_K(x)$,*

$$\langle x, y \rangle = \|y\|^2$$

and thus

$$\|y\|^2 = \|x\|^2 - d(x, K)^2.$$

The proposed P²GDR exploits that, according to the Eckart–Young theorem [10], projecting onto the closed cone $\mathbb{R}_{\leq r}^{m \times n}$ amounts to computing a truncated SVD. The rest of this section reviews related results that are used in Sections 4 and 6.

The singular values of $X \in \mathbb{R}^{m \times n}$ are denoted by $\sigma_1(X) \geq \dots \geq \sigma_{\min\{m, n\}}(X) \geq 0$. Moreover, if $X \neq 0_{m \times n}$, then $\sigma_{\text{rank } X}(X)$ is denoted by $\sigma_{\min}(X)$. The spectral norm of $X \in \mathbb{R}^{m \times n}$, denoted by $\|X\|_2$, equals $\sigma_1(X)$ [21, Example 5.6.6]. Proposition 2.3 shows that the singular values are Lipschitz continuous with Lipschitz constant 1, a property that is used to prove Proposition 2.5 and Lemma 6.1.

Proposition 2.3 ([17, Corollary 8.6.2]). *For all $X, Y \in \mathbb{R}^{m \times n}$ and $i \in \{1, \dots, \min\{m, n\}\}$,*

$$|\sigma_i(X) - \sigma_i(Y)| \leq \|X - Y\|_2 \leq \|X - Y\|.$$

Proposition 2.4 is invoked in the proof of Proposition 4.1.¹

Proposition 2.4. *For all real matrices X and Y such that the number of columns of X equals the number of rows of Y ,*

$$\|XY\| \leq \|X\|_2 \|Y\|.$$

By reducing the rank of $X \in \mathbb{R}^{m \times n} \setminus \{0_{m \times n}\}$, we mean computing an element of $P_{\mathbb{R}_{\leq r}^{m \times n}}(X)$ for some nonnegative integer $r < \text{rank } X$, with the convention $\mathbb{R}_{\leq 0}^{m \times n} := \mathbb{R}_0^{m \times n} := \{0_{m \times n}\}$. The Eckart–Young theorem implies that for every nonnegative integer $r < \min\{m, n\}$ and $X \in \mathbb{R}^{m \times n}$:

1. if $\text{rank } X \leq r$, then $P_{\mathbb{R}_{\leq r}^{m \times n}}(X) = \{X\}$;

¹Proofs of Proposition 2.4 can be found at:

<https://mathoverflow.net/questions/59918/kind-of-submultiplicativity-of-the-frobenius-norm-ab-f-leq-a-2-b-f>.

2. if $\text{rank } X > r$, then $d(X, \mathbb{R}_{\leq r}^{m \times n}) = d(X, \mathbb{R}_r^{m \times n}) = \sqrt{\sum_{i=r+1}^{\text{rank } X} \sigma_i^2(X)}$ and $P_{\mathbb{R}_{\leq r}^{m \times n}}(X) = P_{\mathbb{R}_r^{m \times n}}(X)$.

Proposition 2.5 is invoked in the proof of Lemma 6.1.

Proposition 2.5. *Let $\underline{X} \in \mathbb{R}^{m \times n}$ and $r := \text{rank } \underline{X}$. For all $\Delta \in (0, \infty)$ and $X \in B[\underline{X}, \Delta]$, it holds that $\text{rank}_\Delta X \leq r$. If $\underline{X} \neq 0_{m \times n}$, then, for all $X \in B(\underline{X}, \sigma_r(\underline{X}))$, it holds that $r \leq \text{rank } X$.*

Proof. The first inequality follows from Proposition 2.3:

$$\sigma_{r+1}(X) = |\sigma_{r+1}(X) - \sigma_{r+1}(\underline{X})| \leq \|X - \underline{X}\| \leq \Delta.$$

The second inequality holds because $B(\underline{X}, \sigma_r(\underline{X})) \subseteq \mathbb{R}^{m \times n} \setminus \mathbb{R}_{< r}^{m \times n}$ since $\sigma_r(\underline{X}) = d(\underline{X}, \mathbb{R}_{< r}^{m \times n})$. \square

2.2 Outer limit and outer semicontinuity of a correspondence

The material reviewed in this section is used in Sections 2.3 and 2.4.

Let $(S_i)_{i \in \mathbb{N}}$ be a sequence of nonempty subsets of \mathcal{E} . The *outer limit* of $(S_i)_{i \in \mathbb{N}}$, denoted by $\overline{\text{Lim}}_{i \rightarrow \infty} S_i$, is the set of all possible accumulation points of sequences $(x_i)_{i \in \mathbb{N}}$ such that $x_i \in S_i$ for all $i \in \mathbb{N}$ [50, §4A].

A *correspondence*, or a *set-valued map*, is a triplet $F := (A_1, A_2, G)$ where A_1 and A_2 are sets respectively called the *set of departure* and the *set of destination* of F , and G is a subset of $A_1 \times A_2$ called the *graph* of F . If $F := (A_1, A_2, G)$ is a correspondence, written $F : A_1 \multimap A_2$, then the *image* of $x \in A_1$ by F is $F(x) := \{y \in A_2 \mid (x, y) \in G\}$ and the *domain* of F is $\text{dom } F := \{x \in A_1 \mid F(x) \neq \emptyset\}$.

Let \mathcal{E}_1 and \mathcal{E}_2 be two Euclidean vector spaces, $F : \mathcal{E}_1 \multimap \mathcal{E}_2$ be a correspondence, S be a nonempty subset of $\text{dom } F$, and $\underline{x} \in S$. The *outer limit* of F at \underline{x} relative to S , denoted by $\overline{\text{Lim}}_{S \ni x \rightarrow \underline{x}} F(x)$, is the set of all $y \in \mathcal{E}_2$ such that there exist sequences $(x_i)_{i \in \mathbb{N}}$ in S converging to \underline{x} and $(y_i)_{i \in \mathbb{N}}$ in \mathcal{E}_2 converging to y such that $y_i \in F(x_i)$ for all $i \in \mathbb{N}$ [50, 5(1)]. The correspondence F is said to be *outer semicontinuous* at \underline{x} relative to S if $\overline{\text{Lim}}_{S \ni x \rightarrow \underline{x}} F(x) \subseteq F(\underline{x})$ or, equivalently, $\overline{\text{Lim}}_{S \ni x \rightarrow \underline{x}} F(x) = F(\underline{x})$ [50, Definition 5.4].

Proposition 2.6 enables to prove Proposition 2.12.

Proposition 2.6. *Let $g : \mathcal{E} \rightarrow \mathcal{E}$ be continuous, $F : \mathcal{E} \multimap \mathcal{E}$ be closed-valued, and S be a nonempty subset of $\text{dom } F$. If F is outer semicontinuous at $\underline{x} \in S$ relative to S , then the function*

$$\text{dom } F \rightarrow \mathbb{R} : x \mapsto d(g(x), F(x))$$

is lower semicontinuous at \underline{x} relative to S .

Proof. Let F be outer semicontinuous at $\underline{x} \in S$ relative to S . For all $x \in S$,

$$\begin{aligned} |d(g(x), F(x)) - d(g(\underline{x}), F(\underline{x}))| &\leq |d(g(x), F(x)) - d(g(\underline{x}), F(x))| + |d(g(\underline{x}), F(x)) - d(g(\underline{x}), F(\underline{x}))| \\ &\leq \|g(x) - g(\underline{x})\| + |d(g(\underline{x}), F(x)) - d(g(\underline{x}), F(\underline{x}))|, \end{aligned}$$

where the first inequality follows from [62, Proposition 1.3.17]. Since g is continuous at \underline{x} , $\lim_{x \rightarrow \underline{x}} \|g(x) - g(\underline{x})\| = 0$. By [50, Proposition 5.11(a)], the function

$$\text{dom } F \rightarrow \mathbb{R} : x \mapsto d(g(\underline{x}), F(x))$$

is lower semicontinuous at \underline{x} relative to S . Thus,

$$\liminf_{S \ni x \rightarrow \underline{x}} d(g(\underline{x}), F(x)) = d(g(\underline{x}), F(\underline{x})).$$

Combining the two limits with the last inequality yields

$$\liminf_{S \ni x \rightarrow \underline{x}} d(g(x), F(x)) = d(g(\underline{x}), F(\underline{x})). \quad \square$$

2.3 Tangent and normal cones

Based on [50, Chapter 6], this section reviews the concept of tangent cone and the two notions of normality on which the two notions of stationarity given in Definition 1.1 are based. Let S be a nonempty subset of \mathcal{E} .

A vector $v \in \mathcal{E}$ is said to be *tangent* to S at $x \in S$ if there exist sequences $(x_i)_{i \in \mathbb{N}}$ in S converging to x and $(t_i)_{i \in \mathbb{N}}$ in $(0, \infty)$ such that the sequence $(\frac{x_i - x}{t_i})_{i \in \mathbb{N}}$ converges to v [50, Definition 6.1]. The set of all tangent vectors to S at $x \in S$ is a closed cone [50, Proposition 6.2] called the (*Bouligand*) *tangent cone* to S at x and denoted by $T_S(x)$. The *regular normal cone* to S at $x \in S$ is the polar

$$\widehat{N}_S(x) := T_S(x)^*,$$

and the *normal cone* to S at x is the outer limit

$$N_S(x) := \overline{\text{Lim}}_{S \ni z \rightarrow x} \widehat{N}_S(z)$$

which is a closed cone [50, Definition 6.3 and Proposition 6.5]. For every $x \in S$,

$$\widehat{N}_S(x) \subseteq N_S(x), \quad (2.1)$$

and S is said to be *Clarke regular* at x if this inclusion is an equality and S is locally closed at x [50, Definition 6.4].

If S is an embedded submanifold of \mathcal{E} around $x \in S$ [2, §3.3.1], then S is Clarke regular at x , $T_S(x)$ and $N_S(x)$ coincide respectively with the tangent and normal spaces to S at x which are the orthogonal complements of each other [50, Example 6.8], and, for every differentiable function $\varphi : \mathcal{E} \rightarrow \mathbb{R}$, the Riemannian gradient of $\varphi|_S$ at x equals $P_{T_S(x)}(\nabla\varphi(x))$ [2, §3.6.1]. (The set C in [50, Example 6.8] is an embedded submanifold of \mathbb{R}^n locally around \bar{x} , as can be seen by applying the submersion theorem [2, Proposition 3.3.3] with $\mathcal{M}_1 := O$, $d_1 := n$, $\mathcal{M}_2 := \mathbb{R}^m$, $d_2 := m$, and $y := 0$.)

For every $X \in \mathbb{R}_{\leq r}^{m \times n}$, the closed cones $T_{\mathbb{R}_{\leq r}^{m \times n}}(X)$, $\widehat{N}_{\mathbb{R}_{\leq r}^{m \times n}}(X)$, and $N_{\mathbb{R}_{\leq r}^{m \times n}}(X)$ can be described explicitly based on orthonormal bases of $\text{im } X$, $\text{im } X^\top$, and their orthogonal complements. Let $\underline{r} := \text{rank } X$, $U \in \text{St}(\underline{r}, m)$, $U_\perp \in \text{St}(m - \underline{r}, m)$, $V \in \text{St}(\underline{r}, n)$, $V_\perp \in \text{St}(n - \underline{r}, n)$, $\text{im } U = \text{im } X$, $\text{im } U_\perp = (\text{im } X)^\perp$, $\text{im } V = \text{im } X^\top$, and $\text{im } V_\perp = (\text{im } X^\top)^\perp$. Then:

- by [52, Theorem 3.2],

$$T_{\mathbb{R}_{\leq r}^{m \times n}}(X) = [U \ U_\perp] \begin{bmatrix} \mathbb{R}^{\underline{r} \times \underline{r}} & \mathbb{R}^{\underline{r} \times n - \underline{r}} \\ \mathbb{R}^{m - \underline{r} \times \underline{r}} & \mathbb{R}_{\leq r - \underline{r}}^{m - \underline{r} \times n - \underline{r}} \end{bmatrix} [V \ V_\perp]^\top \quad (2.2)$$

and, by [52, Corollary 3.3], for every $Z \in \mathbb{R}^{m \times n}$ written as

$$Z = [U \ U_\perp] \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix} [V \ V_\perp]^\top$$

with $A_{1,1} = U^\top Z V$, $A_{1,2} = U^\top Z V_\perp$, $A_{2,1} = U_\perp^\top Z V$, and $A_{2,2} = U_\perp^\top Z V_\perp$,

$$P_{T_{\mathbb{R}_{\leq r}^{m \times n}}(X)}(Z) = [U \ U_\perp] \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & P_{\mathbb{R}_{\leq r - \underline{r}}^{m - \underline{r} \times n - \underline{r}}}(A_{2,2}) \end{bmatrix} [V \ V_\perp]^\top, \quad (2.3)$$

$$\|P_{T_{\mathbb{R}_{\leq r}^{m \times n}}(X)}(Z)\| \geq \sqrt{\frac{r - \underline{r}}{\min\{m, n\} - \underline{r}}} \|Z\|; \quad (2.4)$$

- by [22, Theorem 3.1],

$$N_{\mathbb{R}_{\leq r}^{m \times n}}(X) = U_\perp \mathbb{R}_{\leq \min\{m, n\} - r}^{m - \underline{r} \times n - \underline{r}} V_\perp^\top; \quad (2.5)$$

- by [22, Corollary 2.3], if $\underline{r} < r$, then

$$\widehat{N}_{\mathbb{R}_{\leq r}^{m \times n}}(X) = \{0_{m \times n}\}. \quad (2.6)$$

As pointed out in Section 1, some of these equations are known in algebraic geometry since the last century. To exploit (2.2) and (2.3) efficiently, note the following.

1. In practice, the projection onto $T_{\mathbb{R}_{\leq r}^{m \times n}}(X)$ can be computed by [52, Algorithm 2]. This does not rely on U_{\perp} and V_{\perp} , which are huge in the frequently encountered case where $r \ll \min\{m, n\}$.
2. As pointed out in Section 1.2, for all $X \in \mathbb{R}_{\leq r}^{m \times n}$ and $Z \in T_{\mathbb{R}_{\leq r}^{m \times n}}(X)$, $X + Z \in \mathbb{R}_{\leq 2r}^{m \times n}$. (In what follows, the diag operator returns a block diagonal matrix [17, §1.3.1] whose diagonal blocks are the arguments given to the operator.) Indeed, if

$$X = [U \ U_{\perp}] \text{diag}(\Sigma, 0_{m-r \times n-r}) [V \ V_{\perp}]^{\top}$$

is an SVD and $Z \in T_{\mathbb{R}_{\leq r}^{m \times n}}(X)$ is written as

$$Z = [U \ U_{\perp}] \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix} [V \ V_{\perp}]^{\top}$$

with $A_{1,1} \in \mathbb{R}^{r \times r}$, $A_{1,2} \in \mathbb{R}^{r \times n-r}$, $A_{2,1} \in \mathbb{R}^{m-r \times r}$, and $A_{2,2} \in \mathbb{R}_{\leq r-r}^{m-r \times n-r}$, then, by [39, Proposition 3.1],

$$X + Z = [U \ U_{\perp}] \begin{bmatrix} \Sigma + A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix} [V \ V_{\perp}]^{\top} \in \mathbb{R}_{\leq r+r}^{m \times n} \subseteq \mathbb{R}_{\leq 2r}^{m \times n}.$$

This is exploited in the analysis of the computational cost per iteration of P²GD and P²GDR conducted in [40, §7].

2.4 Stationarity and measures of stationarity

This section completes Section 1.1 by establishing several results mentioned therein.

For a set that is not Clarke regular, M-stationarity is weaker than B-stationarity. For problem (1.1), this is quantified in Remark 2.7.

Remark 2.7. The formulas (2.5) and (2.6) show precisely that, for problem (1.1), M-stationarity is weaker than B-stationarity at every $X \in \mathbb{R}_{\leq r}^{m \times n}$ with $r \in \{0, \dots, r-1\}$. Indeed, while the latter amounts to $\nabla f(X) = 0_{m \times n}$, the former only amounts to the B-stationarity of X for the problem of minimizing f on $\mathbb{R}_{\leq r}^{m \times n}$ together with the inequality $\text{rank } \nabla f(X) \leq \min\{m, n\} - r$.

Proposition 2.9 states that, for problem (1.1), the M-stationarity of a point depends on f only through its restriction $f|_{\mathbb{R}_{\leq r}^{m \times n}}$. It is based on Lemma 2.8.

Lemma 2.8. *For every differentiable function $g : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ such that $g(\mathbb{R}_{\leq r}^{m \times n}) = \{0_{m \times n}\}$, it holds that $\nabla g(\mathbb{R}_{\leq r}^{m \times n}) = \{0_{m \times n}\}$.*

Proof. Let $g : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ be differentiable and such that $g(\mathbb{R}_{\leq r}^{m \times n}) = \{0_{m \times n}\}$. Let $X \in \mathbb{R}_{\leq r}^{m \times n}$. For every $(i, j) \in \{1, \dots, m\} \times \{1, \dots, n\}$, let $E_{i,j} := [\delta_{i,k} \delta_{j,l}]_{k,l=1}^{m,n}$ and $\partial_{i,j} g(X) := \lim_{t \rightarrow 0} \frac{g(X + tE_{i,j}) - g(X)}{t}$. The numerator in the limit is zero since, for every $t \in \mathbb{R}$, $\text{rank}(X + tE_{i,j}) \leq \text{rank } X + \text{rank } tE_{i,j} \leq r$. The result follows since $\nabla g(X) = [\partial_{i,j} g(X)]_{i,j=1}^{m,n}$. \square

Proposition 2.9. *Let $h : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ be differentiable and such that $h|_{\mathbb{R}_{\leq r}^{m \times n}} = f|_{\mathbb{R}_{\leq r}^{m \times n}}$. Then, $X \in \mathbb{R}_{\leq r}^{m \times n}$ is M-stationary for the minimization of $h|_{\mathbb{R}_{\leq r}^{m \times n}}$ if and only if X is M-stationary for (1.1).*

Proof. Let $X \in \mathbb{R}_{\leq r}^{m \times n}$. If $X \in \mathbb{R}_r^{m \times n}$, the result follows from the equality $N_{\mathbb{R}_{\leq r}^{m \times n}}(X) = \widehat{N}_{\mathbb{R}_{\leq r}^{m \times n}}(X)$ and the fact that the B-stationarity of X depends on f only through $f|_{\mathbb{R}_{\leq r}^{m \times n}}$. If $X \in \mathbb{R}_{< r}^{m \times n}$, Lemma 2.8 implies that $\nabla h(X) = \nabla f(X)$, and the result follows. \square

The rest of this section concerns the measures of stationarity in (1.5)–(1.7). Let C be a nonempty closed subset of \mathcal{E} .

Proposition 2.2 implies that the measure of B-stationarity in (1.5) is well defined with, for every differentiable function $\varphi : \mathcal{E} \rightarrow C$ and every $x \in C$,

$$\|P_{T_C(x)}(-\nabla\varphi(x))\| = \sqrt{\|\nabla\varphi(x)\|^2 - d(-\nabla\varphi(x), T_C(x))^2} =: s(x; \varphi, C). \quad (2.7)$$

The stationarity measure $s(\cdot; \varphi, C)$ is used in [52] to define and analyze P²GD and RFD, and in this paper to define and analyze P²GDR. It possesses the following desirable properties:

1. the map that associates with every $x \in C$ the set $P_{T_C(x)}(-\nabla\varphi(x))$ depends on φ only through its restriction $\varphi|_C$ [29, Lemmas A.7 and A.8];
2. for every $x \in C$, $-\nabla\varphi(x) \in \widehat{N}_C(x)$ if and only if $s(x; \varphi, C) = 0$ [29, Proposition 2.5];
3. for every $x \in C$, $s(x; \varphi, C)$ is the best linear rate of decrease that can be achieved in φ by moving away from x along a direction in $T_C(x)$ [29, Theorem A.9(b) in arXiv:2107.03877v3].

Proposition 2.10 implies that the measures of B-stationarity in (1.5) and (1.6) can fail to be lower semicontinuous at every point where Clarke regularity is absent.

Proposition 2.10. *For every $x \in C$ where C is not Clarke regular, there exist a sequence $(x_i)_{i \in \mathbb{N}}$ in C converging to x and a continuously differentiable function $\varphi : \mathcal{E} \rightarrow \mathbb{R}$ such that the measures of B-stationarity $s(\cdot; \varphi, C)$ and*

$$C \rightarrow \mathbb{R} : x \mapsto d(-\nabla\varphi(x), \widehat{N}_C(x))$$

converge to zero along $(x_i)_{i \in \mathbb{N}}$, yet $-\nabla\varphi(x) \in N_C(x) \setminus \widehat{N}_C(x)$.²

Proof. Assume that C is not Clarke regular at $x \in C$. Let $v \in N_C(x) \setminus \widehat{N}_C(x)$ and $\varphi : \mathcal{E} \rightarrow \mathbb{R} : z \mapsto -\langle z, v \rangle$. Then, φ is continuously differentiable and $-\nabla\varphi(z) = v$ for all $z \in \mathcal{E}$. In particular, $s(x; \varphi, C) \neq 0$. Furthermore, by definition of $N_C(x)$, there exist sequences $(x_i)_{i \in \mathbb{N}}$ in C converging to x and $(v_i)_{i \in \mathbb{N}}$ converging to v such that $v_i \in \widehat{N}_C(x_i)$ for all $i \in \mathbb{N}$. Therefore, for all $i \in \mathbb{N}$,

$$s(x_i; \varphi, C) \leq d(-\nabla\varphi(x_i), \widehat{N}_C(x_i)) \leq \|v - v_i\| \rightarrow 0 \text{ when } i \rightarrow \infty,$$

where the first inequality follows from [29, Theorem A.9(c) in arXiv:2107.03877v3]. □

The stationarity measure $s(\cdot; \varphi, C)$ is used in [29] to define the concept of apocalypse. A point $x \in C$ is said to be *apocalyptic* if there exist a sequence $(x_i)_{i \in \mathbb{N}}$ in C converging to x and a continuously differentiable function $\varphi : \mathcal{E} \rightarrow \mathbb{R}$ such that $\lim_{i \rightarrow \infty} s(x_i; \varphi, C) = 0$ whereas $s(x; \varphi, C) > 0$ [29, Definition 2.7]. Such a triplet $(x, (x_i)_{i \in \mathbb{N}}, \varphi)$ is called an *apocalypse*.

Corollary 2.11 answers a question raised in [29, §4].

Corollary 2.11. *A point $x \in C$ is apocalyptic if and only if C is not Clarke regular at x .*

Proof. This follows from [29, Corollary 2.15] and Proposition 2.10. □

On a set that is not Clarke regular, no measure of B-stationarity in the literature is known to be lower semicontinuous. In contrast, the measure of M-stationarity in (1.7) is lower semicontinuous.

Proposition 2.12. *For every continuously differentiable function $\varphi : \mathcal{E} \rightarrow \mathbb{R}$, the function*

$$C \rightarrow \mathbb{R} : x \mapsto d(-\nabla\varphi(x), N_C(x))$$

is lower semicontinuous.

Proof. This follows from Proposition 2.6 since N_C is outer semicontinuous relative to C [50, Proposition 6.6]. □

²We are grateful to an anonymous referee of arXiv:2303.16040v1 for pointing this out.

The notion of serendipity is complementary to that of apocalypse. A point $x \in C$ is said to be *serendipitous* if there exist a sequence $(x_i)_{i \in \mathbb{N}}$ in C converging to x , a continuously differentiable function $\varphi : \mathcal{E} \rightarrow \mathbb{R}$, and $\varepsilon \in (0, \infty)$ such that $s(x_i; \varphi, C) > \varepsilon$ for all $i \in \mathbb{N}$, yet $s(x; \varphi, C) = 0$ [29, Definition 2.8]. Thus, if $x \in C$ is not serendipitous and $s(x; \varphi, C) = 0$ for some continuously differentiable function $\varphi : \mathcal{E} \rightarrow \mathbb{R}$, then $\lim_{i \rightarrow \infty} s(x_i; \varphi, C) = 0$ for every sequence $(x_i)_{i \in \mathbb{N}}$ in C converging to x . A point $x \in C$ is serendipitous if and only if there exists a sequence $(x_i)_{i \in \mathbb{N}}$ in C converging to x such that $\widehat{N}_C(x) \not\subseteq \left(\overline{\lim}_{i \rightarrow \infty} T_C(x_i)\right)^*$ [29, Theorem 2.17]. Proposition 2.13 enables to deduce Corollary 6.3 from Theorem 6.2.

Proposition 2.13. *No point of $\mathbb{R}_{\leq r}^{m \times n}$ is serendipitous.*

Proof. We use [29, Theorem 2.17]. Let $X \in \mathbb{R}_{\leq r}^{m \times n}$. Let us prove that X is not serendipitous. Let $(X_i)_{i \in \mathbb{N}}$ be a sequence in $\mathbb{R}_{\leq r}^{m \times n}$ converging to X . If $\text{rank } X = r$, then, $\overline{\lim}_{i \rightarrow \infty} T_{\mathbb{R}_{\leq r}^{m \times n}}(X_i) = T_{\mathbb{R}_{\leq r}^{m \times n}}(X)$ [39, Proposition 4.3], and thus $(\overline{\lim}_{i \rightarrow \infty} T_{\mathbb{R}_{\leq r}^{m \times n}}(X_i))^* = \widehat{N}_{\mathbb{R}_{\leq r}^{m \times n}}(X)$. If $\text{rank } X < r$, then

$$\widehat{N}_{\mathbb{R}_{\leq r}^{m \times n}}(X) = \{0_{m \times n}\} \subseteq \left(\overline{\lim}_{i \rightarrow \infty} T_{\mathbb{R}_{\leq r}^{m \times n}}(X_i)\right)^*,$$

where the equality follows from (2.6). □

Proposition 2.13 ensures that, for every sequence $(X_i)_{i \in \mathbb{N}}$ in $\mathbb{R}_{\leq r}^{m \times n}$ that has at least one accumulation point that is B-stationary for (1.1) and for every $\varepsilon \in (0, \infty)$, there exists $i \in \mathbb{N}$ such that $s(X_i; f, \mathbb{R}_{\leq r}^{m \times n}) \leq \varepsilon$. This provides the following stopping criterion for every algorithm that generates such a sequence: given $\varepsilon \in (0, \infty)$, stop the algorithm at iteration

$$i_\varepsilon := \min\{i \in \mathbb{N} \mid s(X_i; f, \mathbb{R}_{\leq r}^{m \times n}) \leq \varepsilon\}. \quad (2.8)$$

3 Sufficient-descent maps

This section, based on [48, §1.3] and [36, §7.6], provides the theoretical framework used to analyze P²GDR in Section 6 and a hybrid algorithm in Section 7. Given a Euclidean vector space \mathcal{E} , a nonempty closed subset C of \mathcal{E} , and a function $\varphi : \mathcal{E} \rightarrow \mathbb{R}$ that is continuous on C , it concerns the problem

$$\min_{x \in C} \varphi(x) \quad (3.1)$$

of minimizing φ on C . Specifically, the accumulation points of every infinite sequence generated by Algorithm 3.1 are proven to lie in the nonempty set $S \subsetneq C$ (Proposition 3.2) provided that the set-valued map A , called the *iteration map* of Algorithm 3.1, satisfies a sufficient-descent condition (Definition 3.1). The example of interest in this work, where φ is assumed to be differentiable on C , is the set of B-stationary points, $S := \{x \in C \mid -\nabla\varphi(x) \in \widehat{N}_C(x)\}$. Moreover, the sufficient-descent condition is preserved under hybridization of maps (Proposition 3.3).

Algorithm 3.1 Model of an iterative algorithm aiming at solving problem (3.1)

Require: $(\mathcal{E}, C, S, \varphi, A)$ where \mathcal{E} is a Euclidean vector space, C is a nonempty closed subset of \mathcal{E} , S is a nonempty proper subset of C , $\varphi : \mathcal{E} \rightarrow \mathbb{R}$ is continuous on C , and A is a set-valued map from C to C such that $A(x) \neq \emptyset$ for all $x \in C$.

Input: $x_0 \in C$.

Output: a sequence in C .

- 1: $i \leftarrow 0$;
 - 2: **while** $x_i \notin S$ **do**
 - 3: Choose $x_{i+1} \in A(x_i)$;
 - 4: $i \leftarrow i + 1$;
 - 5: **end while**
-

Definition 3.1 (sufficient-descent map). A (φ, C, S) -sufficient-descent map is a set-valued map A from C to C that satisfies the following two conditions:

1. for all $x \in C$, $A(x) \neq \emptyset$;
2. for every $\underline{x} \in C \setminus S$, there exist $\varepsilon(\underline{x}), \delta(\underline{x}) \in (0, \infty)$ such that, for all $x \in B[\underline{x}, \varepsilon(\underline{x})] \cap C$, it holds that $\varphi(x) - \sup \varphi(A(x)) \geq \delta(\underline{x})$.

A (φ, C) -sufficient-descent map is a (φ, C, S) -sufficient-descent map where $S := \{x \in C \mid -\nabla\varphi(x) \in \widehat{N}_C(x)\}$.

In other words, every $\underline{x} \in C \setminus S$ has a neighborhood where the decrease in φ produced by the map is bounded away from zero. Examples of $(f, \mathbb{R}_{\leq r}^{m \times n})$ -sufficient-descent maps include the RFDR map [40, Proposition 6.1], the monotone PGD map (as can be deduced from [44]), and the P²GDR map (Lemma 6.1).

Following [48, Remark 14], which states that it is usually better to determine whether an algorithm generates a sequence having at least one accumulation point by examining the algorithm in the light of the specific problem to which one wishes to apply it, no condition ensuring the existence of a convergent subsequence is imposed in Proposition 3.2. As a reminder, a sequence $(x_i)_{i \in \mathbb{N}}$ in \mathcal{E} has at least one accumulation point if and only if $\liminf_{i \rightarrow \infty} \|x_i\| < \infty$.

Proposition 3.2 (accumulation points in a set). *Let A be a (φ, C, S) -sufficient-descent map and $(x_i)_{i \in \mathbb{N}}$ be a sequence in $C \setminus S$ such that $x_{i+1} \in A(x_i)$ for all $i \in \mathbb{N}$. Then, all accumulation points of $(x_i)_{i \in \mathbb{N}}$ are in S .*

Proof. Let $(x_{i_k})_{k \in \mathbb{N}}$ be a subsequence converging to $x \in C$. For the sake of contradiction, assume that $x \notin S$. Let $\varepsilon(x)$ and $\delta(x)$ be given by Definition 3.1. There exists $K \in \mathbb{N}$ such that, for all integers $k \geq K$, $x_{i_k} \in B[x, \varepsilon(x)]$ and thus $\varphi(x_{i_k}) - \varphi(x_{i_{k+1}}) \geq \delta(x)$. Thus, since $(\varphi(x_i))_{i \in \mathbb{N}}$ is decreasing, for all integers $k \geq K$, $\varphi(x_{i_k}) - \varphi(x_{i_{k+1}}) \geq \delta(x)$. Since φ is continuous, $(\varphi(x_{i_k}))_{k \in \mathbb{N}}$ converges to $\varphi(x)$. Therefore, letting k tend to infinity in the last inequality yields a contradiction. \square

Proposition 3.3 (finite union of sufficient-descent maps). *Let s be a positive integer. For every $i \in \{1, \dots, s\}$, let A_i be a (φ, C, S) -sufficient-descent map. Then, every map that associates a nonempty subset of $\bigcup_{i=1}^s A_i(x)$ with every $x \in C$ is a (φ, C, S) -sufficient-descent map.*

Proof. For every $i \in \{1, \dots, s\}$ and $\underline{x} \in C$, there exist $\varepsilon_i(\underline{x}), \delta_i(\underline{x}) \in (0, \infty)$ such that, for all $x \in B[\underline{x}, \varepsilon_i(\underline{x})] \cap C$, it holds that $\varphi(x) - \sup \varphi(A_i(x)) \geq \delta_i(\underline{x})$. Define $\varepsilon(\underline{x}) := \min_{i \in \{1, \dots, s\}} \varepsilon_i(\underline{x})$ and $\delta(\underline{x}) := \min_{i \in \{1, \dots, s\}} \delta_i(\underline{x})$. Then, for all $x \in B[\underline{x}, \varepsilon(\underline{x})] \cap C$, it holds that $\varphi(x) - \sup \varphi(\bigcup_{i=1}^s A_i(x)) \geq \delta(\underline{x})$. \square

For example, Proposition 3.3 implies that every sequence $(X_i)_{i \in \mathbb{N}}$ in $\mathbb{R}_{\leq r}^{m \times n}$ such that, for every $i \in \mathbb{N}$, X_{i+1} is an image of X_i under the RFDR, monotone PGD, or P²GDR map accumulates at B-stationary points of (1.1). This is exploited in Section 7.

4 An upper bound on the distance to the determinantal variety from a tangent line

The analysis of P²GD in Section 5 is based on Proposition 4.1 which offers an upper bound on the distance to $\mathbb{R}_{\leq r}^{m \times n}$ from each of its tangent lines. A trivial upper bound on $d(X + Z, \mathbb{R}_{\leq r}^{m \times n})$ holding for all $X \in \mathbb{R}_{\leq r}^{m \times n}$ and $Z \in T_{\mathbb{R}_{\leq r}^{m \times n}}(X)$ is $\|Z\|$, and [52, Proposition 3.6] tightens it to $\frac{1}{\sqrt{2}}\|Z\|$. However, an upper bound proportional to $\|Z\|^2$ is needed in the proof of Proposition 5.1; such a bound is given and the proportionality factor is shown to be $1/\sigma_{\min}(X)$ up to a constant in Proposition 4.1.

Proposition 4.1. *For all $X \in \mathbb{R}_{\leq r}^{m \times n} \setminus \{0_{m \times n}\}$,*

$$\frac{\sqrt{5} - 1}{r + 1} \frac{1}{2\sigma_{\min}(X)} \leq \sup_{Z \in T_{\mathbb{R}_{\leq r}^{m \times n}}(X) \setminus \{0_{m \times n}\}} \frac{d(X + Z, \mathbb{R}_{\leq r}^{m \times n})}{\|Z\|^2} \leq \frac{1}{2\sigma_{\min}(X)}.$$

Proof. Let $\underline{r} := \text{rank } X$. We first establish the upper bound. Let

$$X = [U \ U_\perp] \text{diag}(\Sigma, 0_{m-\underline{r} \times n-\underline{r}})[V \ V_\perp]^\top$$

be an SVD, and $Z \in T_{\mathbb{R}_{\leq \underline{r}}^{m \times n}}(X) \setminus \{0_{m \times n}\}$. By (2.2), there are $A_{1,1} \in \mathbb{R}^{\underline{r} \times \underline{r}}$, $A_{1,2} \in \mathbb{R}^{\underline{r} \times n-\underline{r}}$, $A_{2,1} \in \mathbb{R}^{m-\underline{r} \times \underline{r}}$, and $A_{2,2} \in \mathbb{R}_{\leq r-\underline{r}}^{m-\underline{r} \times n-\underline{r}}$ such that

$$Z = [U \ U_\perp] \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix} [V \ V_\perp]^\top.$$

Define the function

$\gamma : [0, \infty) \rightarrow \mathbb{R}_{\leq \underline{r}}^{m \times n} : t \mapsto (U + t(U_\perp A_{2,1} + \frac{1}{2}U A_{1,1})\Sigma^{-1})\Sigma(V + t(V_\perp A_{1,2}^\top + \frac{1}{2}V A_{1,1}^\top)\Sigma^{-1})^\top + tU_\perp A_{2,2}V_\perp^\top$, where the first term is inspired from [64, (13)]; γ is well defined since the ranks of the two terms are respectively upper bounded by \underline{r} and $r - \underline{r}$. For all $t \in [0, \infty)$,

$$\gamma(t) = X + tZ + \frac{t^2}{4}[U \ U_\perp] \begin{bmatrix} A_{1,1} \\ 2A_{2,1} \end{bmatrix} \Sigma^{-1} \begin{bmatrix} A_{1,1} & 2A_{1,2} \end{bmatrix} [V \ V_\perp]^\top$$

thus

$$d(X + tZ, \mathbb{R}_{\leq \underline{r}}^{m \times n}) \leq \|(X + tZ) - \gamma(t)\| = \frac{t^2}{4} \left\| \begin{bmatrix} A_{1,1}\Sigma^{-1}A_{1,1} & 2A_{1,1}\Sigma^{-1}A_{1,2} \\ 2A_{2,1}\Sigma^{-1}A_{1,1} & 4A_{2,1}\Sigma^{-1}A_{1,2} \end{bmatrix} \right\|.$$

Observe that

$$\begin{aligned} & \left\| \begin{bmatrix} A_{1,1}\Sigma^{-1}A_{1,1} & 2A_{1,1}\Sigma^{-1}A_{1,2} \\ 2A_{2,1}\Sigma^{-1}A_{1,1} & 4A_{2,1}\Sigma^{-1}A_{1,2} \end{bmatrix} \right\|^2 \\ &= \|A_{1,1}\Sigma^{-1}A_{1,1}\|^2 + 4\|A_{1,1}\Sigma^{-1}A_{1,2}\|^2 + 4\|A_{2,1}\Sigma^{-1}A_{1,1}\|^2 + 16\|A_{2,1}\Sigma^{-1}A_{1,2}\|^2 \\ &\leq \|\Sigma^{-1}\|_2^2 \left(\|A_{1,1}\|^4 + 4\|A_{1,1}\|^2\|A_{1,2}\|^2 + 4\|A_{1,1}\|^2\|A_{2,1}\|^2 + 16\|A_{1,2}\|^2\|A_{2,1}\|^2 \right) \\ &\leq \|\Sigma^{-1}\|_2^2 \|Z\|^4 \max_{\substack{x,y,z \in \mathbb{R} \\ x^2+y^2+z^2 \leq 1}} x^4 + 4x^2y^2 + 4x^2z^2 + 16y^2z^2 \\ &= 4\|\Sigma^{-1}\|_2^2 \|Z\|^4 \\ &= \frac{4}{\sigma_{\underline{r}}^2(X)} \|Z\|^4, \end{aligned}$$

where the first inequality follows from Proposition 2.4. Therefore, for all $t \in [0, \infty)$,

$$d(X + tZ, \mathbb{R}_{\leq \underline{r}}^{m \times n}) \leq t^2 \frac{1}{2\sigma_{\underline{r}}(X)} \|Z\|^2.$$

Choosing $t = 1$ yields the upper bound.

We now establish the lower bound. Let

$$X = [U \ U_\perp] \text{diag}(\sigma_1, \dots, \sigma_{\underline{r}}, 0_{m-\underline{r} \times n-\underline{r}})[V \ V_\perp]^\top$$

be an SVD, and observe that

$$Z := [U \ U_\perp] \sigma_{\underline{r}} \text{diag}(0_{\underline{r}-1}, \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, I_{r-\underline{r}}, 0_{m-r-1 \times n-r-1})[V \ V_\perp]^\top \in T_{\mathbb{R}_{\leq \underline{r}}^{m \times n}}(X).$$

The nonzero singular values of

$$X + Z = [U \ U_\perp] \text{diag}(\sigma_1, \dots, \sigma_{\underline{r}-1}, \sigma_{\underline{r}} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \sigma_{\underline{r}} I_{r-\underline{r}}, 0_{m-r-1 \times n-r-1})[V \ V_\perp]^\top$$

are $\sigma_1, \dots, \sigma_{\underline{r}}$ and the absolute values of the eigenvalues of $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ multiplied by $\sigma_{\underline{r}}$, i.e., $\frac{\sqrt{5}+1}{2}\sigma_{\underline{r}}$ and $\frac{\sqrt{5}-1}{2}\sigma_{\underline{r}}$. Thus, $d(X + Z, \mathbb{R}_{\leq \underline{r}}^{m \times n}) = \frac{\sqrt{5}-1}{2}\sigma_{\underline{r}}$, $\|Z\|^2 = (r - \underline{r} + 2)\sigma_{\underline{r}}^2 \leq (r + 1)\sigma_{\underline{r}}^2$, and the lower bound follows. \square

Proposition 4.1 can be related to geometric principles. Let γ be a curve on the submanifold $\mathbb{R}_r^{m \times n}$ of $\mathbb{R}^{m \times n}$. In view of the Gauss formula along a curve [28, Corollary 8.3], the normal part of the acceleration of γ is given by $\mathbb{I}(\gamma', \gamma')$, where \mathbb{I} denotes the second fundamental form. In view of [12, §4], the largest principal curvature of $\mathbb{R}_r^{m \times n}$ at X is $1/\sigma_r(X)$; hence $\|\mathbb{I}(\gamma'(t), \gamma'(t))\| \leq \|\gamma'(t)\|^2/\sigma_r(\gamma(t))$, and the bound is attained when $\gamma'(t)$ is along the corresponding principal direction.

5 A lower bound on the step size of the P²GD map

The P²GD map, already briefly described in Section 1.2, is defined as Algorithm 5.1. Corollary 5.2, based on Proposition 5.1, provides an upper bound on the number of iterations performed in the backtracking procedure (while loop of Algorithm 5.1). This serves as a basis for the convergence analysis conducted in Section 6 since the P²GD map is used as a subroutine by the P²GDR map (Algorithm 6.1). Algorithm 5.1 corresponds to the iteration map of [52, Algorithm 3] except that the initial step size for the backtracking procedure is chosen in a given bounded interval.

Algorithm 5.1 P²GD map on $\mathbb{R}_{\leq r}^{m \times n}$ (based on [52, Algorithm 3])

Require: $(f, r, \underline{\alpha}, \bar{\alpha}, \beta, c)$ where $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ is differentiable with ∇f locally Lipschitz continuous, $r < \min\{m, n\}$ is a positive integer, $0 < \underline{\alpha} \leq \bar{\alpha} < \infty$, and $\beta, c \in (0, 1)$.

Input: $X \in \mathbb{R}_{\leq r}^{m \times n}$ such that $s(X; f, \mathbb{R}_{\leq r}^{m \times n}) > 0$.

Output: $Y \in \text{P}^2\text{GD}(X; f, r, \underline{\alpha}, \bar{\alpha}, \beta, c)$.

- 1: Choose $G \in P_{T_{\mathbb{R}_{\leq r}^{m \times n}}(X)}(-\nabla f(X))$, $\alpha \in [\underline{\alpha}, \bar{\alpha}]$, and $Y \in P_{\mathbb{R}_{\leq r}^{m \times n}}(X + \alpha G)$;
 - 2: **while** $f(Y) > f(X) - c\alpha s(X; f, \mathbb{R}_{\leq r}^{m \times n})^2$ **do**
 - 3: $\alpha \leftarrow \alpha\beta$;
 - 4: Choose $Y \in P_{\mathbb{R}_{\leq r}^{m \times n}}(X + \alpha G)$;
 - 5: **end while**
 - 6: Return Y .
-

By [37, Lemma 1.2.3], since ∇f is locally Lipschitz continuous, i.e., for every closed ball $\mathcal{B} \subsetneq \mathbb{R}^{m \times n}$,

$$\text{Lip}_{\mathcal{B}}(\nabla f) := \sup_{\substack{X, Y \in \mathcal{B} \\ X \neq Y}} \frac{\|\nabla f(X) - \nabla f(Y)\|}{\|X - Y\|} < \infty,$$

it holds for all $X, Y \in \mathcal{B}$ that

$$|f(Y) - f(X) - \langle \nabla f(X), Y - X \rangle| \leq \frac{\text{Lip}_{\mathcal{B}}(\nabla f)}{2} \|Y - X\|^2. \quad (5.1)$$

Proposition 5.1. *Let $X \in \mathbb{R}_{\leq r}^{m \times n}$ and $\bar{\alpha} \in (0, \infty)$. Let $\mathcal{B} \subsetneq \mathbb{R}^{m \times n}$ be a closed ball such that, for all $G \in P_{T_{\mathbb{R}_{\leq r}^{m \times n}}(X)}(-\nabla f(X))$ and $\alpha \in [0, \bar{\alpha}]$, $P_{\mathbb{R}_{\leq r}^{m \times n}}(X + \alpha G) \subseteq \mathcal{B}$; an example of such a ball is $B[X, 2\bar{\alpha} s(X; f, \mathbb{R}_{\leq r}^{m \times n})]$. Then, for all $G \in P_{T_{\mathbb{R}_{\leq r}^{m \times n}}(X)}(-\nabla f(X))$ and $\alpha \in [0, \bar{\alpha}]$, it holds that*

$$\sup f(P_{\mathbb{R}_{\leq r}^{m \times n}}(X + \alpha G)) \leq f(X) + s(X; f, \mathbb{R}_{\leq r}^{m \times n})^2 \alpha (-1 + \kappa_{\mathcal{B}}(X; f, \bar{\alpha}) \alpha), \quad (5.2)$$

where

$$\kappa_{\mathcal{B}}(X; f, \bar{\alpha}) := \begin{cases} \frac{1}{2} \text{Lip}_{\mathcal{B}}(\nabla f) & \text{if } X = 0_{m \times n}, \\ \frac{\|\nabla f(X)\|}{2\sigma_{\min}(X)} + \frac{1}{2} \text{Lip}_{\mathcal{B}}(\nabla f) \left(\bar{\alpha} \frac{s(X; f, \mathbb{R}_{\leq r}^{m \times n})}{2\sigma_{\min}(X)} + 1 \right)^2 & \text{otherwise.} \end{cases}$$

Proof. The example $B[X, 2\bar{\alpha} s(X; f, \mathbb{R}_{\leq r}^{m \times n})]$ is correct by Proposition 2.1. Let $G \in P_{T_{\mathbb{R}_{\leq r}^{m \times n}}(X)}(-\nabla f(X))$. The proof of (5.2) is based on (5.1) and the equality $\langle \nabla f(X), G \rangle = -s(X; f, \mathbb{R}_{\leq r}^{m \times n})^2$ which holds by Proposition 2.2 since $T_{\mathbb{R}_{\leq r}^{m \times n}}(X)$ is a closed cone. The result follows readily from (5.1) if $X = 0_{m \times n}$ since $T_{\mathbb{R}_{\leq r}^{m \times n}}(0_{m \times n}) = \mathbb{R}_{\leq r}^{m \times n}$. Let us therefore consider the case where $X \neq 0_{m \times n}$. Let $L := \text{Lip}_{\mathcal{B}}(\nabla f)$.

For all $\alpha \in [0, \bar{\alpha}]$ and $Y \in P_{\mathbb{R}_{\leq r}^{m \times n}}(X + \alpha G)$,

$$\begin{aligned}
& f(Y) - f(X) \\
& \leq \langle \nabla f(X), Y - X \rangle + \frac{L}{2} \|Y - X\|^2 \\
& = \langle \nabla f(X), Y - (X + \alpha G) + \alpha G \rangle + \frac{L}{2} \|Y - (X + \alpha G) + \alpha G\|^2 \\
& = -\alpha s(X; f, \mathbb{R}_{\leq r}^{m \times n})^2 + \langle \nabla f(X), Y - (X + \alpha G) \rangle + \frac{L}{2} \|Y - (X + \alpha G) + \alpha G\|^2 \\
& \leq -\alpha s(X; f, \mathbb{R}_{\leq r}^{m \times n})^2 + \|\nabla f(X)\| d(X + \alpha G, \mathbb{R}_{\leq r}^{m \times n}) + \frac{L}{2} \left(d(X + \alpha G, \mathbb{R}_{\leq r}^{m \times n}) + \alpha s(X; f, \mathbb{R}_{\leq r}^{m \times n}) \right)^2 \\
& \leq -\alpha s(X; f, \mathbb{R}_{\leq r}^{m \times n})^2 + \alpha^2 \frac{\|\nabla f(X)\|}{2\sigma_{\min}(X)} s(X; f, \mathbb{R}_{\leq r}^{m \times n})^2 + \frac{L}{2} \left(\alpha^2 \frac{s(X; f, \mathbb{R}_{\leq r}^{m \times n})^2}{2\sigma_{\min}(X)} + \alpha s(X; f, \mathbb{R}_{\leq r}^{m \times n}) \right)^2 \\
& = \alpha s(X; f, \mathbb{R}_{\leq r}^{m \times n})^2 \left(-1 + \alpha \left(\frac{\|\nabla f(X)\|}{2\sigma_{\min}(X)} + \frac{L}{2} \left(\alpha \frac{s(X; f, \mathbb{R}_{\leq r}^{m \times n})}{2\sigma_{\min}(X)} + 1 \right)^2 \right) \right) \\
& \leq \alpha s(X; f, \mathbb{R}_{\leq r}^{m \times n})^2 (-1 + \alpha \kappa_{\mathcal{B}}(X; f, \bar{\alpha})),
\end{aligned}$$

where the third inequality follows from Proposition 4.1. \square

Corollary 5.2. *Let \mathcal{B} be a closed ball as in Proposition 5.1. The while loop in Algorithm 5.1 terminates after at most*

$$\max \left\{ 0, \left\lceil \ln \left(\frac{1-c}{\alpha_0 \kappa_{\mathcal{B}}(X; f, \bar{\alpha})} \right) / \ln \beta \right\rceil \right\} \quad (5.3)$$

iterations, where α_0 is the initial step size chosen in line 1, and every $Y \in P^2\text{GD}(X; f, r, \underline{\alpha}, \bar{\alpha}, \beta, c)$ satisfies the Armijo condition

$$f(Y) \leq f(X) - c \alpha s(X; f, \mathbb{R}_{\leq r}^{m \times n})^2 \quad (5.4)$$

with a step size $\alpha \in [\min\{\underline{\alpha}, \frac{\beta(1-c)}{\kappa_{\mathcal{B}}(X; f, \bar{\alpha})}\}, \bar{\alpha}]$.

Proof. For all $\alpha \in (0, \infty)$,

$$f(X) + s(X; f, \mathbb{R}_{\leq r}^{m \times n})^2 \alpha (-1 + \kappa_{\mathcal{B}}(X; f, \bar{\alpha}) \alpha) \leq f(X) - c s(X; f, \mathbb{R}_{\leq r}^{m \times n})^2 \alpha \quad \text{iff} \quad \alpha \leq \frac{1-c}{\kappa_{\mathcal{B}}(X; f, \bar{\alpha})}.$$

Since the left-hand side of the first inequality is an upper bound on $f(P_{\mathbb{R}_{\leq r}^{m \times n}}(X + \alpha G))$ for all $\alpha \in (0, \bar{\alpha}]$, the Armijo condition is necessarily satisfied if $\alpha \in (0, \min\{\bar{\alpha}, \frac{1-c}{\kappa_{\mathcal{B}}(X; f, \bar{\alpha})}\}]$. Therefore, either the initial step size α_0 chosen in $[\underline{\alpha}, \bar{\alpha}]$ satisfies the Armijo condition or the while loop ends after iteration $i \in \mathbb{N} \setminus \{0\}$ with $\alpha = \alpha_0 \beta^i$ such that $\frac{\alpha}{\beta} > \frac{1-c}{\kappa_{\mathcal{B}}(X; f, \bar{\alpha})}$. In the second case, $i < 1 + \ln(\frac{1-c}{\alpha_0 \kappa_{\mathcal{B}}(X; f, \bar{\alpha})}) / \ln(\beta)$ and thus $i \leq \lceil \ln(\frac{1-c}{\alpha_0 \kappa_{\mathcal{B}}(X; f, \bar{\alpha})}) / \ln(\beta) \rceil$. \square

6 P²GDR and its convergence analysis

The P²GDR map, already briefly described in Section 1.3, is defined as Algorithm 6.1.

Algorithm 6.1 P²GDR map on $\mathbb{R}_{\leq r}^{m \times n}$

Require: $(f, r, \underline{\alpha}, \bar{\alpha}, \beta, c, \Delta)$ where $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ is differentiable with ∇f locally Lipschitz continuous, $r < \min\{m, n\}$ is a positive integer, $0 < \underline{\alpha} \leq \bar{\alpha} < \infty$, $\beta, c \in (0, 1)$, and $\Delta \in (0, \infty)$.

Input: $X \in \mathbb{R}_{\leq r}^{m \times n}$ such that $s(X; f, \mathbb{R}_{\leq r}^{m \times n}) > 0$.

Output: $Y \in \text{P}^2\text{GDR}(X; f, r, \underline{\alpha}, \bar{\alpha}, \beta, c, \Delta)$.

- 1: **for** $i \in \{0, \dots, \text{rank } X - \text{rank}_\Delta X\}$ **do**
 - 2: Choose $\hat{X}^i \in P_{\mathbb{R}_{\text{rank } X - i}^{m \times n}}(X)$;
 - 3: Choose $\tilde{X}^i \in \text{P}^2\text{GD}(\hat{X}^i; f, r, \underline{\alpha}, \bar{\alpha}, \beta, c)$;
 - 4: **end for**
 - 5: Return $Y \in \text{argmin}_{\{\tilde{X}^i | i \in \{0, \dots, \text{rank } X - \text{rank}_\Delta X\}\}} f$.
-

P²GDR is defined as Algorithm 6.2. It corresponds to Algorithm 3.1 applied to problem (1.1) with S the set of B-stationary points.

Algorithm 6.2 P²GDR on $\mathbb{R}_{\leq r}^{m \times n}$

Require: $(f, r, \underline{\alpha}, \bar{\alpha}, \beta, c, \Delta)$ where $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ is differentiable with ∇f locally Lipschitz continuous, $r < \min\{m, n\}$ is a positive integer, $0 < \underline{\alpha} \leq \bar{\alpha} < \infty$, $\beta, c \in (0, 1)$, and $\Delta \in (0, \infty)$.

Input: $X_0 \in \mathbb{R}_{\leq r}^{m \times n}$.

Output: a sequence in $\mathbb{R}_{\leq r}^{m \times n}$.

- 1: $i \leftarrow 0$;
 - 2: **while** $s(X_i; f, \mathbb{R}_{\leq r}^{m \times n}) > 0$ **do**
 - 3: Choose $X_{i+1} \in \text{P}^2\text{GDR}(X_i; f, r, \underline{\alpha}, \bar{\alpha}, \beta, c, \Delta)$;
 - 4: $i \leftarrow i + 1$;
 - 5: **end while**
-

Algorithms 6.1 and 6.2 are analyzed in Section 6.1 based on the theoretical framework from Section 3. The choice of the parameter Δ is briefly discussed in Section 6.2.

6.1 Convergence analysis

Lemma 6.1 states that the P²GDR map is an $(f, \mathbb{R}_{\leq r}^{m \times n})$ -sufficient-descent map in the sense of Definition 3.1: for every $\underline{X} \in \mathbb{R}_{\leq r}^{m \times n}$ that is not B-stationary for (1.1), the decrease in f obtained by applying the P²GDR map to every $X \in \mathbb{R}_{\leq r}^{m \times n}$ sufficiently close to \underline{X} is bounded away from zero.

Lemma 6.1. *The P²GDR map is an $(f, \mathbb{R}_{\leq r}^{m \times n})$ -sufficient-descent map in the sense of Definition 3.1, i.e., for every $\underline{X} \in \mathbb{R}_{\leq r}^{m \times n}$ such that $s(\underline{X}; f, \mathbb{R}_{\leq r}^{m \times n}) > 0$, there exist $\varepsilon(\underline{X}), \delta(\underline{X}) \in (0, \infty)$ such that, for all $X \in B[\underline{X}, \varepsilon(\underline{X})] \cap \mathbb{R}_{\leq r}^{m \times n}$ and $Y \in \text{P}^2\text{GDR}(X; f, r, \underline{\alpha}, \bar{\alpha}, \beta, c, \Delta)$,*

$$f(Y) - f(X) \leq -\delta(\underline{X}). \quad (6.1)$$

Proof. Let $\underline{X} \in \mathbb{R}_{\leq r}^{m \times n}$ be such that $s(\underline{X}; f, \mathbb{R}_{\leq r}^{m \times n}) > 0$. Define $\underline{r} := \text{rank } \underline{X}$. This proof constructs $\varepsilon(\underline{X})$ and $\delta(\underline{X})$ based on the Armijo condition (5.4) given in Corollary 5.2. The first step of the proof ends with the definition of $\delta(\underline{X})$ in (6.7). Then, the second and third steps define $\varepsilon(\underline{X})$ for the cases where $\underline{r} = r$ and $\underline{r} < r$, respectively. The rank reduction mechanism of P²GDR plays a role only in the third step.

By Proposition 2.2, $\|\nabla f(\underline{X})\| \geq s(\underline{X}; f, \mathbb{R}_{\leq r}^{m \times n})$. Since ∇f is continuous at \underline{X} , there exists $\rho_1(\underline{X}) \in (0, \infty)$ such that, for all $X \in B[\underline{X}, \rho_1(\underline{X})]$, $\|\nabla f(X) - \nabla f(\underline{X})\| \leq \frac{1}{2}\|\nabla f(\underline{X})\|$ and hence, as $\|\nabla f(X)\| - \|\nabla f(\underline{X})\| \leq \|\nabla f(X) - \nabla f(\underline{X})\|$,

$$\frac{1}{2}\|\nabla f(\underline{X})\| \leq \|\nabla f(X)\| \leq \frac{3}{2}\|\nabla f(\underline{X})\|. \quad (6.2)$$

Define

$$\bar{\rho}(\underline{X}) := 3\bar{\alpha}\|\nabla f(\underline{X})\| + \rho_1(\underline{X}).$$

Then, for every $X \in B[\underline{X}, \rho_1(\underline{X})] \cap \mathbb{R}_{\leq r}^{m \times n}$, the inclusion $B[X, 2\bar{\alpha} s(X; f, \mathbb{R}_{\leq r}^{m \times n})] \subseteq B[\underline{X}, \bar{\rho}(\underline{X})]$ holds since, for all $Z \in B[X, 2\bar{\alpha} s(X; f, \mathbb{R}_{\leq r}^{m \times n})]$,

$$\|Z - \underline{X}\| \leq \|Z - X\| + \|X - \underline{X}\| \leq 2\bar{\alpha} s(X; f, \mathbb{R}_{\leq r}^{m \times n}) + \rho_1(\underline{X}) \leq 2\bar{\alpha} \|\nabla f(X)\| + \rho_1(\underline{X}) \leq \bar{\rho}(\underline{X}),$$

where the third and fourth inequalities follow respectively from Proposition 2.2 and the second inequality in (6.2). Thus, in view of the example given in Proposition 5.1, for all $X \in B[\underline{X}, \rho_1(\underline{X})] \cap \mathbb{R}_{\leq r}^{m \times n}$ and $\tilde{X} \in \text{P}^2\text{GD}(X; f, r, \underline{\alpha}, \bar{\alpha}, \beta, c)$, Corollary 5.2 applies with the ball $B[\underline{X}, \bar{\rho}(\underline{X})]$, and (5.4) yields

$$f(\tilde{X}) \leq f(X) - cs(X; f, \mathbb{R}_{\leq r}^{m \times n})^2 \min \left\{ \underline{\alpha}, \frac{\beta(1-c)}{\kappa_{B[\underline{X}, \bar{\rho}(\underline{X})]}(X; f, \bar{\alpha})} \right\}. \quad (6.3)$$

Define

$$\bar{\kappa}(\underline{X}) := \begin{cases} \frac{1}{2} \text{Lip}_{B[\underline{X}, \bar{\rho}(\underline{X})]}(\nabla f) & \text{if } \underline{X} = 0_{m \times n}, \\ \frac{3\|\nabla f(\underline{X})\|}{2\sigma_{\underline{r}}(\underline{X})} + \frac{1}{2} \text{Lip}_{B[\underline{X}, \bar{\rho}(\underline{X})]}(\nabla f) \left(\bar{\alpha} \frac{3\|\nabla f(\underline{X})\|}{2\sigma_{\underline{r}}(\underline{X})} + 1 \right)^2 & \text{if } \underline{X} \neq 0_{m \times n}. \end{cases} \quad (6.4)$$

If $\underline{X} \neq 0_{m \times n}$, then for all $X \in B[\underline{X}, \frac{1}{2}\sigma_{\underline{r}}(\underline{X})]$ it holds that

$$|\sigma_{\underline{r}}(X) - \sigma_{\underline{r}}(\underline{X})| \leq \|X - \underline{X}\| \leq \frac{1}{2}\sigma_{\underline{r}}(\underline{X}),$$

where the first inequality follows from Proposition 2.3, and thus $\sigma_{\underline{r}}(X) \geq \frac{1}{2}\sigma_{\underline{r}}(\underline{X})$. Hence, by Proposition 2.2 and the second inequality in (6.2), if $\underline{X} \neq 0_{m \times n}$, then for all $X \in B[\underline{X}, \min\{\rho_1(\underline{X}), \frac{1}{2}\sigma_{\underline{r}}(\underline{X})\}] \cap \mathbb{R}_{\underline{r}}^{m \times n}$ it holds that

$$\kappa_{B[\underline{X}, \bar{\rho}(\underline{X})]}(X; f, \bar{\alpha}) \leq \bar{\kappa}(\underline{X}). \quad (6.5)$$

Therefore, by (6.3), for all $X \in B[\underline{X}, \min\{\rho_1(\underline{X}), \frac{1}{2}\sigma_{\underline{r}}(\underline{X})\}] \cap \mathbb{R}_{\underline{r}}^{m \times n}$ and $\tilde{X} \in \text{P}^2\text{GD}(X; f, r, \underline{\alpha}, \bar{\alpha}, \beta, c)$,

$$f(\tilde{X}) \leq f(X) - cs(X; f, \mathbb{R}_{\leq r}^{m \times n})^2 \min \left\{ \underline{\alpha}, \frac{\beta(1-c)}{\bar{\kappa}(\underline{X})} \right\}. \quad (6.6)$$

The inequality (6.5) also holds if $\underline{X} = X = 0_{m \times n}$, hence so does (6.6). Define

$$\delta(\underline{X}) := \begin{cases} \frac{c}{4} s(\underline{X}; f, \mathbb{R}_{\leq r}^{m \times n})^2 \min \left\{ \underline{\alpha}, \frac{\beta(1-c)}{\bar{\kappa}(\underline{X})} \right\} & \text{if } \underline{r} = r, \\ \frac{c\|\nabla f(\underline{X})\|^2}{12(\min\{m, n\} - r + 1)} \min \left\{ \underline{\alpha}, \frac{\beta(1-c)}{\bar{\kappa}(\underline{X})} \right\} & \text{if } \underline{r} < r. \end{cases} \quad (6.7)$$

Let us consider the case where $\underline{r} = r$. On $B(\underline{X}, \sigma_r(\underline{X})) \cap \mathbb{R}_{\leq r}^{m \times n} = B(\underline{X}, \sigma_r(\underline{X})) \cap \mathbb{R}_r^{m \times n}$, $s(\cdot; f, \mathbb{R}_{\leq r}^{m \times n})$ coincides with the norm of the Riemannian gradient of the restriction of f to the smooth manifold $\mathbb{R}_r^{m \times n}$ (see Section 2.3), which is continuous [52, §2.1]. Thus, there exists $\rho_2(\underline{X}) \in (0, \sigma_r(\underline{X}))$ such that $s(B[\underline{X}, \rho_2(\underline{X})] \cap \mathbb{R}_r^{m \times n}; f, \mathbb{R}_{\leq r}^{m \times n}) \subseteq [\frac{1}{2} s(\underline{X}; f, \mathbb{R}_{\leq r}^{m \times n}), \frac{3}{2} s(\underline{X}; f, \mathbb{R}_{\leq r}^{m \times n})]$. Define

$$\varepsilon(\underline{X}) := \min \left\{ \rho_1(\underline{X}), \rho_2(\underline{X}), \frac{1}{2}\sigma_r(\underline{X}) \right\}.$$

Let $X \in B[\underline{X}, \varepsilon(\underline{X})] \cap \mathbb{R}_{\leq r}^{m \times n}$ and $Y \in \text{P}^2\text{GDR}(X; f, r, \underline{\alpha}, \bar{\alpha}, \beta, c, \Delta)$. There exists $\tilde{X} \in \text{P}^2\text{GD}(X; f, r, \underline{\alpha}, \bar{\alpha}, \beta, c)$ such that $f(Y) \leq f(\tilde{X})$. Therefore, by (6.6) and (6.7),

$$f(Y) \leq f(\tilde{X}) \leq f(X) - cs(X; f, \mathbb{R}_{\leq r}^{m \times n})^2 \min \left\{ \underline{\alpha}, \frac{\beta(1-c)}{\bar{\kappa}(\underline{X})} \right\} \leq f(X) - \delta(\underline{X}).$$

Let us now consider the case where $\underline{r} < r$. Since f is continuous at \underline{X} , there exists $\rho_0(\underline{X}) \in (0, \infty)$ such that $f(B[\underline{X}, \rho_0(\underline{X})]) \subseteq [f(\underline{X}) - \delta(\underline{X}), f(\underline{X}) + \delta(\underline{X})]$. Define

$$\varepsilon(\underline{X}) := \begin{cases} \min\{\Delta, \frac{1}{2}\rho_0(\underline{X}), \frac{1}{2}\rho_1(\underline{X})\} & \text{if } \underline{X} = 0_{m \times n}, \\ \min\{\Delta, \frac{1}{2}\rho_0(\underline{X}), \frac{1}{2}\rho_1(\underline{X}), \frac{1}{4}\sigma_{\underline{r}}(\underline{X})\} & \text{if } \underline{X} \neq 0_{m \times n}. \end{cases}$$

Let $X \in B[\underline{X}, \varepsilon(\underline{X})] \cap \mathbb{R}_{\leq r}^{m \times n}$ and $Y \in \text{P}^2\text{GDR}(X; f, r, \underline{\alpha}, \bar{\alpha}, \beta, c, \Delta)$. By Proposition 2.5, $\text{rank}_\Delta X \leq r \leq \text{rank} X$. Thus, $0 \leq \text{rank} X - r \leq \text{rank} X - \text{rank}_\Delta X$ and there exist $\hat{X}^{\text{rank} X - r} \in P_{\mathbb{R}_r^{m \times n}}(X)$ and $\tilde{X}^{\text{rank} X - r} \in \text{P}^2\text{GD}(\hat{X}^{\text{rank} X - r}; f, r, \underline{\alpha}, \bar{\alpha}, \beta, c)$ such that $f(Y) \leq f(\tilde{X}^{\text{rank} X - r})$. By Proposition 2.1, $\hat{X}^{\text{rank} X - r} \in B[\underline{X}, 2\varepsilon(\underline{X})]$. Therefore,

$$f(\hat{X}^{\text{rank} X - r}) \leq f(X) + 2\delta(\underline{X}). \quad (6.8)$$

Furthermore, Proposition 2.2 and (2.4) successively yield

$$\|\nabla f(\hat{X}^{\text{rank} X - r})\| \geq s(\hat{X}^{\text{rank} X - r}; f, \mathbb{R}_{\leq r}^{m \times n}) \geq \sqrt{\frac{r - r}{\min\{m, n\} - r}} \|\nabla f(\hat{X}^{\text{rank} X - r})\|.$$

Thus, combining

$$\frac{r - r}{\min\{m, n\} - r} \geq \frac{1}{\min\{m, n\} - r + 1}$$

with (6.2) yields

$$s(\hat{X}^{\text{rank} X - r}; f, \mathbb{R}_{\leq r}^{m \times n}) \in \left[\frac{1}{2\sqrt{\min\{m, n\} - r + 1}} \|\nabla f(\underline{X})\|, \frac{3}{2} \|\nabla f(\underline{X})\| \right]. \quad (6.9)$$

Hence,

$$\begin{aligned} f(Y) &\leq f(\tilde{X}^{\text{rank} X - r}) \\ &\leq f(\hat{X}^{\text{rank} X - r}) - c s(\hat{X}^{\text{rank} X - r}; f, \mathbb{R}_{\leq r}^{m \times n})^2 \min \left\{ \underline{\alpha}, \frac{\beta(1 - c)}{\bar{\kappa}(\underline{X})} \right\} \\ &\leq f(X) + 2\delta(\underline{X}) - 3\delta(\underline{X}) \\ &= f(X) - \delta(\underline{X}), \end{aligned}$$

where the second inequality follows from (6.6) and the third from (6.8), (6.9), and (6.7). \square

Theorem 6.2. *Consider a sequence generated by P^2GDR (Algorithm 6.2). If this sequence is finite, then its last element is B -stationary for (1.1) in the sense of Definition 1.1. If it is infinite, then all of its accumulation points are B -stationary for (1.1).*

Proof. This follows from Proposition 3.2 and Lemma 6.1. \square

Corollary 6.3. *Let $(X_i)_{i \in \mathbb{N}}$ be a sequence generated by P^2GDR (Algorithm 6.2). For every convergent subsequence $(X_{i_k})_{k \in \mathbb{N}}$, $\lim_{k \rightarrow \infty} s(X_{i_k}; f, \mathbb{R}_{\leq r}^{m \times n}) = 0$. If $(X_i)_{i \in \mathbb{N}}$ is bounded, which is the case if the sublevel set $\{X \in \mathbb{R}_{\leq r}^{m \times n} \mid f(X) \leq f(X_0)\}$ is bounded, then $\lim_{i \rightarrow \infty} s(X_i; f, \mathbb{R}_{\leq r}^{m \times n}) = 0$, and all accumulation points have the same image by f .*

Proof. The two limits follow from Proposition 2.13. The final claim follows from the argument given in the proof of [48, Theorem 65]. Specifically, if $(X_i)_{i \in \mathbb{N}}$ is bounded, then it contains at least one convergent subsequence. Assume that $(X_{i_k})_{k \in \mathbb{N}}$ and $(X_{j_k})_{k \in \mathbb{N}}$ converge respectively to \underline{X} and \bar{X} . The sequence $(f(X_i))_{i \in \mathbb{N}}$ is decreasing and, since $(X_i)_{i \in \mathbb{N}}$ is bounded and f is continuous, it converges to $\inf_{i \in \mathbb{N}} f(X_i)$. Therefore, $f(\underline{X}) = \lim_{k \rightarrow \infty} f(X_{i_k}) = \lim_{i \rightarrow \infty} f(X_i) = \lim_{k \rightarrow \infty} f(X_{j_k}) = f(\bar{X})$. \square

The analysis conducted in Sections 5 and 6.1 can be extended straightforwardly to the case where f is only defined on an open subset of $\mathbb{R}^{m \times n}$ containing $\mathbb{R}_{\leq r}^{m \times n}$.

6.2 On the choice of Δ

In many practical situations, when Δ is chosen reasonably small, all the iterates of P²GDR (Algorithm 6.2) satisfy $\text{rank}_\Delta X_i = \text{rank} X_i$. In this case, the range of values in the for-loop of P²GDR always reduces to $\{0\}$, and P²GDR generates the same sequence $(X_i)_{i \in \mathbb{N}}$ as P²GD. In this scenario, the only computational overhead in P²GDR is the computation of $\text{rank} X_i$ and $\text{rank}_\Delta X_i$. However, for all $i \geq 1$, in view of line 4 of Algorithm 5.1, it is reasonable to assume that X_i has been obtained by a truncated SVD, in which case $\text{rank} X_i$ and $\text{rank}_\Delta X_i$ are immediately available, making the overhead insignificant. In summary, P²GDR offers stronger convergence properties than P²GD, and while incurring an insignificant overhead in many practical situations.

However, the range of values in the for-loop of P²GDR can also be as large as $\{0, \dots, r\}$ (as recorded in Table 8.1), and there are situations where this occurs each time the for-loop is reached (e.g., in the case of a bounded sublevel set, when Δ is chosen so large that $\text{rank}_\Delta X = 0$ for all X in the sublevel set). One can thus wonder whether it is possible to restrict (conditionally or not) the range of values in the for-loop while preserving the B-stationarity of the accumulation points. This is an open question. Nevertheless, in view of Proposition 4.1, since in the neighborhood of every $X \in \mathbb{R}_{<r}^{m \times n}$, σ_i can be arbitrarily small for every $i \in \{\text{rank} X + 1, \dots, r\}$, it seems unlikely that P²GDR with a restricted for-loop can be analyzed along the lines of Section 6.1. On the other hand, should the answer to the open question be negative, a counterexample other than the one from [29, §2.2] would be required in view of [29, Remark 2.11].

7 A P²GD–PGD hybrid

This section takes advantage of Lemma 6.1 and the theoretical framework from Section 3 to define a hybrid algorithm based on the P²GD map (Algorithm 5.1) and the monotone PGD map (reviewed as Algorithm 7.1) which accumulates at B-stationary points of (1.1). This hybrid algorithm is called P²GD–PGD and its iteration map is defined as Algorithm 7.2. Other hybridizations are possible, but we emphasize this P²GD–PGD hybrid because it combines two classic algorithms and does not require rank reductions. It can be viewed as a way of speeding up the well-known PGD by offering opportunities to resort to the computationally cheaper P²GD (see Section 1.2 and Table 8.1) without relinquishing the property of accumulating at B-stationary points.

Algorithm 7.1 monotone PGD map on $\mathbb{R}_{\leq r}^{m \times n}$ ([44, Algorithm 4.1] for problem (1.1) with $\mu := 0$)

Require: $(f, r, \underline{\alpha}, \bar{\alpha}, \beta, c)$ where $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ is differentiable with ∇f locally Lipschitz continuous, $r < \min\{m, n\}$ is a positive integer, $0 < \underline{\alpha} \leq \bar{\alpha} < \infty$, and $\beta, c \in (0, 1)$.

Input: $X \in \mathbb{R}_{\leq r}^{m \times n}$ such that $s(X; f, \mathbb{R}_{\leq r}^{m \times n}) > 0$.

Output: $Y \in \bar{\text{PGD}}(X; f, r, \underline{\alpha}, \bar{\alpha}, \beta, c)$.

- 1: Choose $\alpha \in [\underline{\alpha}, \bar{\alpha}]$ and $Y \in P_{\mathbb{R}_{\leq r}^{m \times n}}(X - \alpha \nabla f(X))$;
 - 2: **while** $f(Y) > f(X) + c \langle \nabla f(\bar{X}), Y - X \rangle$ **do**
 - 3: $\alpha \leftarrow \alpha \beta$;
 - 4: Choose $Y \in P_{\mathbb{R}_{\leq r}^{m \times n}}(X - \alpha \nabla f(X))$;
 - 5: **end while**
 - 6: Return Y .
-

Algorithm 7.2 P²GD–PGD map on $\mathbb{R}_{\leq r}^{m \times n}$

Require: $(f, r, \underline{\alpha}, \bar{\alpha}, \beta, c, \Delta)$ where $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ is differentiable with ∇f locally Lipschitz continuous, $r < \min\{m, n\}$ is a positive integer, $0 < \underline{\alpha} \leq \bar{\alpha} < \infty$, $\beta, c \in (0, 1)$, and $\Delta \in (0, \infty)$.

Input: $X \in \mathbb{R}_{\leq r}^{m \times n}$ such that $s(X; f, \mathbb{R}_{\leq r}^{m \times n}) > 0$.

Output: $Y \in \text{P}^2\text{GD–PGD}(X; f, r, \underline{\alpha}, \bar{\alpha}, \beta, c, \Delta)$.

- 1: **if** $\text{rank } X = \text{rank}_{\Delta} X$ **then**
 - 2: Choose $Y \in \text{P}^2\text{GD}(X; f, r, \underline{\alpha}, \bar{\alpha}, \beta, c) \cup \text{PGD}(X; f, r, \underline{\alpha}, \bar{\alpha}, \beta, c)$;
 - 3: **else**
 - 4: Choose $Y \in \text{PGD}(X; f, r, \underline{\alpha}, \bar{\alpha}, \beta, c)$;
 - 5: **end if**
 - 6: Return Y .
-

Lemma 7.1. *The P²GD–PGD map is an $(f, \mathbb{R}_{\leq r}^{m \times n})$ -sufficient-descent map in the sense of Definition 3.1.*

Proof. If $\text{rank } X = \text{rank}_{\Delta} X$, then $\text{P}^2\text{GD}(X; f, r, \underline{\alpha}, \bar{\alpha}, \beta, c) = \text{P}^2\text{GDR}(X; f, r, \underline{\alpha}, \bar{\alpha}, \beta, c, \Delta)$, hence Algorithm 7.2 remains unchanged if P²GD is replaced with P²GDR in line 2. Since the monotone PGD map and the P²GDR map are $(f, \mathbb{R}_{\leq r}^{m \times n})$ -sufficient-descent maps, the P²GD–PGD map is an $(f, \mathbb{R}_{\leq r}^{m \times n})$ -sufficient-descent map by Proposition 3.3. \square

P²GD–PGD corresponds to Algorithm 3.1 applied to problem (1.1) with A the P²GD–PGD map and S the set of B-stationary points.

Theorem 7.2. *Consider a sequence generated by P²GD–PGD. If this sequence is finite, then its last element is B-stationary for (1.1) in the sense of Definition 1.1. If it is infinite, then all of its accumulation points are B-stationary for (1.1).*

Proof. This follows from Proposition 3.2 and Lemma 7.1. \square

8 Comparison of P²GDR with state-of-the-art methods

In this section, P²GDR and P²GD–PGD are compared with the five state-of-the-art algorithms reviewed in Section 1.2 and listed in Table 1.1; an overview of this comparison is provided in Section 1.3. The comparison is based on the computational cost per iteration (Section 8.1) and the numerical performance on the instance of (1.1) presented in [29, §2.2] (Section 8.2). As in Section 1.2, HRTR is studied with the rank factorization lift

$$\varphi : \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r} \rightarrow \mathbb{R}^{m \times n} : (L, R) \mapsto LR^{\top} \quad (8.1)$$

from [29, (1.1)], the hook from [29, Example 3.11], and the Cauchy step at every iteration. This algorithm minimizes the lifted cost function $g := f \circ \varphi$.

8.1 Comparison based on the computational cost per iteration

The respective computational costs per iteration of P²GD, RFD, P²GDR, and RFDR are compared in [40, §7] based on detailed implementations of these algorithms involving only evaluations of f and ∇f and some operations from linear algebra:

1. matrix multiplication;
2. thin QR factorization with column pivoting (see, e.g., [17, Algorithm 5.4.1]);
3. *small-scale* (truncated) SVD, i.e., the smallest dimension of the matrix to decompose is at most $2r$;
4. *large-scale* (truncated) SVD, i.e., (truncated) SVD that is not small scale.

In this list, only the (truncated) SVD relies on a convergent algorithm rather than an algorithm with finite complexity such as those for QR factorization. In this section, we include PGD and HRTR in the comparison which is summarized in Table 8.1 for the frequently encountered case where $3r \leq \min\{m, n\}$. This inequality ensures that the projection in the right-hand side of (2.3) requires a large-scale (truncated) SVD since the smallest dimension of the matrix to decompose is $\min\{m, n\} - r$ which is at least $\min\{m, n\} - r + 1$ and thus larger than $2r$.

For PGD, P²GD, and RFD, the computational cost per iteration depends on the number of iterations performed in the backtracking loop. By [44, Corollary 6.2], (5.3), and [40, (11)], given $X \in \mathbb{R}_{\leq r}^{m \times n}$ as input and $\alpha \in (0, \infty)$ as initial step size, this number is upper bounded by

$$\max \left\{ 0, \left\lceil \ln \left(\frac{1-c}{\alpha \text{Lip}_{B[X, 2\alpha \|\nabla f(X)\|]}(\nabla f)} \right) / \ln \beta \right\rceil \right\} \quad (\text{PGD}), \quad (8.2)$$

$$\max \left\{ 0, \left\lceil \ln \left(\frac{1-c}{\alpha \kappa_{B[X, 2\alpha s(X; f, \mathbb{R}_{\leq r}^{m \times n})]}(X; f, \alpha)} \right) / \ln \beta \right\rceil \right\} \quad (\text{P}^2\text{GD}), \quad (8.3)$$

$$\max \left\{ 0, \left\lceil \ln \left(\frac{2(1-c)}{\alpha \text{Lip}_{B[X, \alpha s(X; f, \mathbb{R}_{\leq r}^{m \times n})]}(\nabla f)} \right) / \ln \beta \right\rceil \right\} \quad (\text{RFD}). \quad (8.4)$$

In view of the definition of the PGD map in [44, Algorithm 4.1], (8.2) justifies the PGD line in Table 8.1 for the general case where the gradient is not assumed to have a special structure such as sparse or low rank. Similarly, the P²GD and RFD lines in Table 8.1 follow from (8.3) and (8.4), as already observed in [40, §7].

It remains to justify the HRTR line. According to the analysis conducted in Appendix A, every iteration of HRTR requires computing the smallest eigenvalue of the order- $(m+n)r$ matrix representing $\nabla^2 g(L, R)$ in a given basis of $\mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}$, where (L, R) is the current iterate, which involves computing a large-scale SVD. Moreover, every iteration that updates the current iterate requires a hook. The hook from [29, Example 3.11] involves computing a thin QR factorization of LR^\top and an SVD of the R factor, which is a small-scale SVD.

<i>Algorithm</i>	<i>f</i>	∇f	$\nabla^2 f$	QR	small SVD	large SVD
PGD	$\frac{1}{1 + (8.2)}$	1	0	0	0	$\frac{1}{1 + (8.2)}$
P ² GD	$\frac{1}{1 + (8.3)}$	1	0	$\frac{2}{4}$	$\frac{1}{1 + (8.3)}$	$\frac{0}{1}$
RFD	$\frac{1}{1 + (8.4)}$	1	0	1	0	$\frac{0}{1}$
P ² GDR	$\frac{1}{(8.4) + r \cdot (8.3)}$	$\frac{1}{r+1}$	0	$\frac{2}{4r-2}$	$\frac{1}{r \cdot (8.3)}$	$\frac{0}{r}$
RFDR	$\frac{1}{2 \cdot (8.4)}$	$\frac{1}{2}$	0	0	$\frac{1}{2}$	$\frac{0}{1}$
HRTR	1	1	1	$\frac{0}{1}$	$\frac{0}{1}$	1

Table 8.1: Operations required by six algorithms aiming at solving problem (1.1) with $3r \leq \min\{m, n\}$ to perform one iteration, matrix multiplication excluded. The fields “*f*”, “ ∇f ”, “ $\nabla^2 f$ ”, “QR”, “small SVD”, and “large SVD” respectively correspond to “evaluation of *f*”, “evaluation of ∇f ”, “evaluation of $\nabla^2 f$ ”, “QR factorization with column pivoting”, “small-scale (truncated) SVD”, and “large-scale (truncated) SVD”. When two subrows appear in a row, the upper entry corresponds to the best case and the lower one to the worst case. An iteration of HRTR performs no QR and no small SVD if and only if it does not change the iterate, i.e., the algorithm does not progress.

8.2 Comparison on the problem from [29, §2.2]

In this section, P²GDR, P²GD–PGD, and the five state-of-the-art algorithms listed in Table 1.1 are compared numerically on the instance of (1.1) presented in [29, §2.2], where $m := n := 3$, $r := 2$,

$$f : \mathbb{R}^{3 \times 3} \rightarrow \mathbb{R} : X \mapsto Q(X(1:2, 1:2)) + \psi(X(3, 3)),$$

$X(1:2, 1:2)$ is the upper-left 2×2 submatrix of X , $X(3, 3)$ its bottom-right entry, $\psi : \mathbb{R} \rightarrow \mathbb{R} : x \mapsto \frac{x^4}{4} - \frac{(x+1)^2}{2}$, $Q : \mathbb{R}^{2 \times 2} \rightarrow \mathbb{R} : Y \mapsto \frac{1}{2} \|D(Y - Y^*)\|^2$, $D := \text{diag}(1, \frac{1}{2})$, and $Y^* := \text{diag}(1, 0)$. It holds that $\text{argmin } \psi = x_0 \approx 1.32471795724475$,

$$\text{argmin } f = \left\{ \left[\begin{array}{ccc} 1 & 0 & a_1 \\ 0 & 0 & a_2 \\ a_3 & a_4 & x_0 \end{array} \right] \mid a_1, a_2, a_3, a_4 \in \mathbb{R} \right\},$$

$$\text{argmin}_{\substack{\mathbb{R}^{3 \times 3} \\ \mathbb{R}^{\leq 2}}} f = \left\{ \left[\begin{array}{ccc} 1 & 0 & a_1 \\ 0 & 0 & a_2 \\ a_3 & a_4 & x_0 \end{array} \right] \mid a_1, a_2, a_3, a_4 \in \mathbb{R}, a_2 a_4 = 0 \right\},$$

and $f^* := \min f = \min_{\mathbb{R}^{\leq 2}} f = \psi(x_0) \approx -1.932257884495233$. Furthermore, P²GD follows the apocalypse $(\text{diag}(1, 0, 0), (\text{diag}(1 + (-\frac{3}{5})^i, (\frac{3}{5})^i, 0))_{i \in \mathbb{N}}, f)$ if used on this problem with $X_0 := \text{diag}(2, 1, 0)$, $\underline{\alpha} := \bar{\alpha} := \frac{8}{5}$, $\beta \in (0, 1)$, and $c := \frac{1}{5}$.

The numerical experiment uses Matlab implementations of the algorithms³ and the parameters described in Table 8.2. In this experiment, RFD generates the same sequence as P²GD, and RFDR and P²GD–PGD generate the same sequence as P²GDR (see [40, (19)] for RFD and RFDR). The number of iterations in the plots is chosen large enough to see when the algorithms stagnate. This yields the plots in Figure 8.1a where it can be observed that PGD, P²GD, and P²GDR start to stagnate after less than 90 iterations while HRTR start to stagnate after about $5.5 \cdot 10^5$ iterations.

PGD [26, Algorithm 3.1]	$X_0 := \text{diag}(2, 1, 0)$, $\gamma_{\min} := \gamma_{\max} := \frac{5}{8}$, $\tau := 2$, $\sigma := \frac{1}{5}$, $m := 0$
P ² GD and P ² GDR	$X_0 := \text{diag}(2, 1, 0)$, $\underline{\alpha} := \bar{\alpha} := \frac{8}{5}$, $\beta := \frac{1}{2}$, $c := \frac{1}{5}$, $\Delta := \frac{1}{10}$
HRTR	$L_0 := R_0 := \begin{bmatrix} \sqrt{2} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$, $\underline{\gamma} := \bar{\gamma} := 1$, $\gamma_c := \frac{1}{2}$, $\eta := \frac{1}{10}$

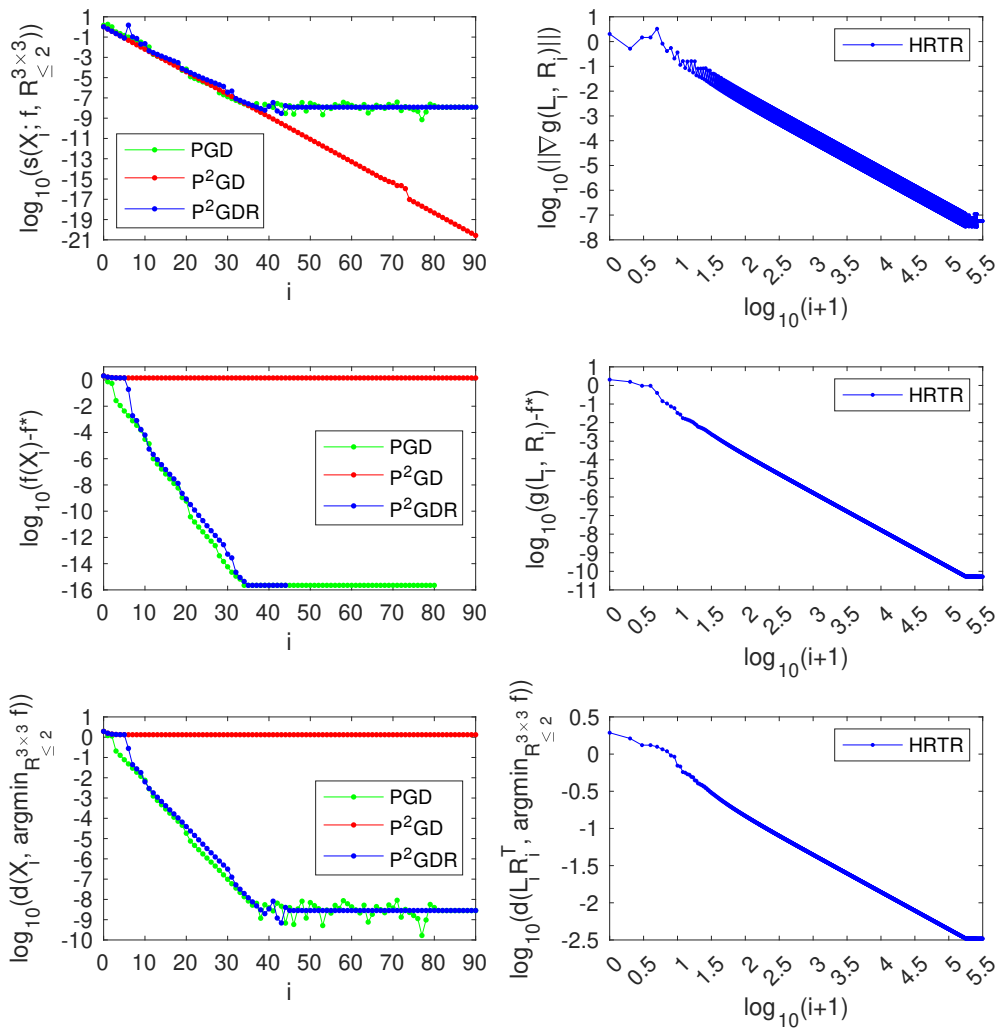
Table 8.2: Parameters of the algorithms for the numerical comparison of Section 8.2.

P²GD behaves as predicted in [29, §2.2] and is the only algorithm among the four to follow an apocalypse. PGD, P²GDR, and HRTR behave in agreement with [44, Theorem 1.2], Theorem 6.2 and Corollary 6.3, and [29, Theorem 1.1]; these results hold since all sublevel sets of f are bounded. It appears that HRTR converges much more slowly than PGD and P²GDR. This slow convergence does not contradict [29, Theorem 3.4]. Indeed, for the parameters of Table 8.2, $k(\varepsilon_1, \varepsilon_2) := \lfloor 160 \max\{\varepsilon_1^{-2}, \varepsilon_2^{-3}\} \rfloor$ is a lower bound on the upper bound on the number of iterations required by HRTR to produce an $(\varepsilon_1, \varepsilon_2)$ -approximate second-order stationary point given in [29, (3.11)], and Figure 8.1b shows that every iteration number i satisfies $i \leq k_i := k(\|\nabla g(L_i, R_i)\|, -\lambda_i)$, where λ_i denotes the smallest eigenvalue of $\nabla^2 g(L_i, R_i)$, which implies that the upper bound given in [29, (3.11)] is respected. Figure 8.1b also shows that the upper bound given in [29, (3.11)] is very pessimistic in this experiment.

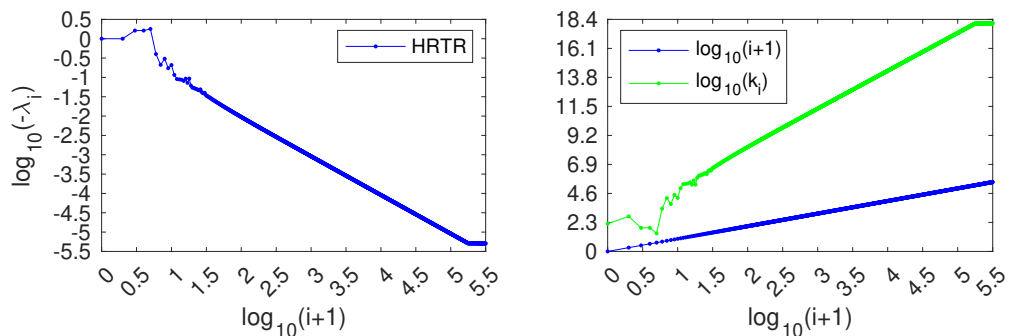
The only iteration of P²GDR that differs from a P²GD iteration is the fifth one, where $\text{rank}_\Delta X_5 = 1$, $\hat{X}_5^1 = \text{diag}(1 - (\frac{3}{5})^5, 0, 0)$, and X_6 is selected in P²GD($\hat{X}_5^1; f, r, \underline{\alpha}, \bar{\alpha}, \beta, c$). This unique intervention of the rank reduction mechanism prevents P²GDR from following the apocalypse.

If the behavior of P²GDR on this example seems satisfying, it should however be noted that, if it is stopped before the rank reduction mechanism takes action (for $\Delta = \frac{1}{10}$, before the fifth iteration),

³The implementations are available at <https://github.com/golikier/BouligandStationarityLowRankOptimization>.



(a) Evolutions of the stationarity measure, the cost function, and the distance to a global minimizer along the sequences generated by PGD, P²GD, P²GDR, and HRTR.



(b) Evolutions of the smallest eigenvalue of the Hessian and the lower bound on the right-hand side of [29, (3.11)] along the sequence generated by HRTR.

Figure 8.1: Numerical comparison of PGD, P²GD, P²GDR, and HRTR on the problem from [29, §2.2] with the parameters from Table 8.2.

then P²GDR generates the exact same (finite) sequence of iterates as P²GD. This shows that, in a practical implementation of P²GDR where it is stopped after a finite number of iterations, it is important to choose Δ in such a way that the algorithm does not stop while it is heading towards an M-stationary point that is not B-stationary (which is $\text{diag}(1, 0, 0)$ in this experiment), in the sense that, if we had continued with $\Delta := 0$, an apocalypse would have occurred.

9 Conclusion

This paper proposes a first-order algorithm, called P²GDR, that generates a sequence in $\mathbb{R}_{\leq r}^{m \times n}$ whose accumulation points are B-stationary for (1.1) in the sense of Definition 1.1. For problem (1.1), B-stationarity is the strongest first-order necessary condition for local optimality (see Section 1.1), and only a handful of algorithms in the literature provably enjoy this property (see Table 1.1). A unique aspect of the proposed P²GDR is that it achieves this feat by only resorting to the classic and computationally cheap P²GD map combined with conditional rank reductions.

Moreover, the framework proposed in Section 3 makes it possible to produce hybrid algorithms that also accumulate at B-stationary points of (1.1). Among them, the P²GD–PGD algorithm introduced in Section 7 requires no rank reduction mechanism and combines two classic iteration maps, exploiting the lower computational cost of P²GD in comparison with PGD while preserving PGD’s property of accumulating at B-stationary points.

Furthermore, it is known (see Table 1.1) that popular methods such as P²GD and RFD fail to accumulate at B-stationary points on some instances of (1.1). A loosely formulated, yet important, open question is how often such failures arise in practice. Regardless of this, algorithms such as the proposed P²GDR and P²GD–PGD provably avoid this pitfall while often incurring a minor overhead in practice.

A Practical implementation of HRTR

This section details the implementation of HRTR with the rank factorization lift (8.1) from [29, (1.1)]. The lifted cost function is $g := f \circ \varphi$. It is assumed that $\nabla f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{m \times n}$ is continuously differentiable, and $\nabla^2 f$ denotes the Hessian of f , defined as the derivative of ∇f [2, §5.5], i.e., $\nabla^2 f : \mathbb{R}^{m \times n} \rightarrow \mathcal{L}(\mathbb{R}^{m \times n}) : X \mapsto (\nabla f)'(X)$, where $\mathcal{L}(\mathbb{R}^{m \times n})$ denotes the Banach space of all (continuous) linear operators on $\mathbb{R}^{m \times n}$.

Proposition A.1. *For all $(L, R), (\dot{L}, \dot{R}) \in \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}$,*

$$\begin{aligned} \nabla g(L, R) &= (\nabla f(LR^\top)R, \nabla f(LR^\top)^\top L), \\ \nabla^2 g(L, R)(\dot{L}, \dot{R}) &= (\nabla^2 f(LR^\top)(\dot{L}R^\top + L\dot{R}^\top)R + \nabla f(LR^\top)\dot{R}, \\ &\quad \nabla^2 f(LR^\top)(\dot{L}R^\top + L\dot{R}^\top)^\top L + \nabla f(LR^\top)^\top \dot{L}). \end{aligned}$$

Proof. It holds that

$$\varphi'(L, R)(\dot{L}, \dot{R}) = \dot{L}R^\top + L\dot{R}^\top.$$

By the chain rule,

$$g'(L, R) = f'(\varphi(L, R)) \circ \varphi'(L, R).$$

Thus,

$$\begin{aligned} g'(L, R)(\dot{L}, \dot{R}) &= f'(\varphi(L, R))(\varphi'(L, R)(\dot{L}, \dot{R})) \\ &= \left\langle \nabla f(\varphi(L, R)), \varphi'(L, R)(\dot{L}, \dot{R}) \right\rangle \\ &= \left\langle \nabla f(\varphi(L, R)), \dot{L}R^\top \right\rangle + \left\langle \nabla f(\varphi(L, R)), L\dot{R}^\top \right\rangle \\ &= \left\langle \nabla f(\varphi(L, R))R, \dot{L} \right\rangle + \left\langle \nabla f(\varphi(L, R))^\top L, \dot{R} \right\rangle \\ &= \left\langle (((\nabla f) \circ \varphi)(L, R)R, ((\nabla f) \circ \varphi)(L, R)^\top L), (\dot{L}, \dot{R}) \right\rangle \end{aligned}$$

and hence

$$\nabla g(L, R) = (((\nabla f) \circ \varphi)(L, R)R, ((\nabla f) \circ \varphi)(L, R)^\top L).$$

The formula for $\nabla^2 g(L, R)(\dot{L}, \dot{R})$ follows from the product rule and the chain rule. \square

Given $(L, R) \in \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}$, the eigenvalues of $\nabla^2 g(L, R)$, which is a self-adjoint linear operator on $\mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}$, are the eigenvalues of the matrix representing it in any basis of $\mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}$. Here, we use the orthonormal basis formed by the concatenation of the sequences $((\delta_{p,l})_{i,k=1}^{m,r}, 0_{n \times r})_{p,l=1}^{m,r}$ and $(0_{m \times r}, (\delta_{q,l})_{j,k=1}^{n,r})_{q,l=1}^{n,r}$. The matrix $H(L, R)$ representing $\nabla^2 g(L, R)$ in that basis can be formed as follows. For every $(i, k) \in \{1, \dots, m\} \times \{1, \dots, r\}$, the $(i-1)r+k$ th column of $H(L, R)$ is the vector formed by concatenating the rows of $\nabla^2 f(LR^\top)((\delta_{i,k})_{p,l=1}^{m,r} R^\top)R$ and those of $\nabla^2 f(LR^\top)((\delta_{i,k})_{p,l=1}^{m,r} R^\top)^\top L + \nabla f(LR^\top)^\top (\delta_{i,k})_{p,l=1}^{m,r}$. Then, for every $(j, k) \in \{1, \dots, n\} \times \{1, \dots, r\}$, the $mr + (j-1)r + k$ th column of $H(L, R)$ is the vector formed by concatenating the rows of $\nabla^2 f(LR^\top)(L(\delta_{j,k})_{q,l=1}^{n,r})^\top R + \nabla f(LR^\top)(\delta_{j,k})_{q,l=1}^{n,r}$ and those of $\nabla^2 f(LR^\top)(L(\delta_{j,k})_{q,l=1}^{n,r})^\top L$.

References

- [1] J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *Journal of Machine Learning Research*, 10(29):803–826, 2009. URL: <http://jmlr.org/papers/v10/abernethy09a.html>.
- [2] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, USA, 2008.
- [3] E. Arbarello, M. Cornalba, P. A. Griffiths, and J. Harris. *Geometry of Algebraic Curves*, volume 267 of *Grundlehren der mathematischen Wissenschaften*. Springer New York, NY, 1985. doi:10.1007/978-1-4757-5323-3.
- [4] R. I. Boş, M. N. Dao, and G. Li. Extrapolated proximal subgradient algorithms for nonconvex and nonsmooth fractional programs. *Mathematics of Operations Research*, 47(3):2415–2443, August 2022. doi:10.1287/moor.2021.1214.
- [5] W. Bruns and U. Vetter. *Determinantal Rings*, volume 1327 of *Lecture Notes in Mathematics*. Springer Berlin, Heidelberg, 1988. doi:10.1007/BFb0080378.
- [6] M. Carlsson, D. Gerosa, and C. Olsson. An unbiased approach to low rank recovery. *SIAM Journal on Matrix Analysis and Applications*, 32(4):2969–2996, 2022. doi:10.1137/19M1294800.
- [7] Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, October 2019. doi:10.1109/TSP.2019.2937282.
- [8] E. Cohen, N. Hallak, and M. Teboulle. A dynamic alternating direction of multipliers for nonconvex minimization with nonlinear functional equality constraints. *Journal of Optimization Theory and Applications*, 193:324–353, June 2022. doi:10.1007/s10957-021-01929-5.
- [9] Y. Cui and J.-S. Pang. *Modern Nonconvex Nondifferentiable Optimization*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2021. doi:10.1137/1.9781611976748.
- [10] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, September 1936. doi:10.1007/BF02288367.
- [11] F. Facchinei, V. Kungurtsev, L. Lampariello, and G. Scutari. Ghost penalties in nonconvex constrained optimization: Diminishing stepsizes and iteration complexity. *Mathematics of Operations Research*, 46(2):595–627, May 2021. doi:10.1287/moor.2020.1079.

- [12] F. Feppon and P. F. J. Lermusiaux. A geometric approach to dynamical model order reduction. *SIAM Journal on Matrix Analysis and Applications*, 39(1):510–538, 2018. doi:10.1137/16M1095202.
- [13] R. M. Freund, P. Grigas, and R. Mazumder. An extended Frank–Wolfe method with “in-face” directions, and its application to low-rank matrix completion. *SIAM Journal on Optimization*, 27(1):319–346, 2017. doi:10.1137/15M104726X.
- [14] B. Gao and P.-A. Absil. A Riemannian rank-adaptive method for low-rank matrix completion. *Computational Optimization and Applications*, 81(1):67–90, January 2022. doi:10.1007/s10589-021-00328-w.
- [15] N. Gillis and F. Glineur. Low-rank matrix approximation with weights or missing data is NP-hard. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1149–1165, 2011. doi:10.1137/110820361.
- [16] D. Goldfarb and S. Ma. Convergence of fixed-point continuation algorithms for matrix rank minimization. *Foundations of Computational Mathematics*, 11(2):183–210, April 2011. doi:10.1007/s10208-011-9084-6.
- [17] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, USA, 4th edition, 2013. doi:10.56021/9781421407944.
- [18] L. Grippo, F. Lampariello, and S. Lucidi. A nonmonotone line search technique for Newton’s method. *SIAM Journal on Numerical Analysis*, 23(4):707–716, August 1986. doi:10.1137/0723046.
- [19] W. Ha, H. Liu, and R. F. Barber. An equivalence between critical points for rank constraints versus low-rank factorizations. *SIAM Journal on Optimization*, 30(4):2927–2955, 2020. doi:10.1137/18M1231675.
- [20] J. Harris. *Algebraic Geometry*, volume 133 of *Graduate Texts in Mathematics*. Springer-Verlag New York, 1992. doi:10.1007/978-1-4757-2189-8.
- [21] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, New-York, NY, USA, 2nd edition, 2013.
- [22] S. Hosseini, D. R. Luke, and A. Uschmajew. *Nonsmooth Optimization and Its Applications*, volume 170 of *International Series of Numerical Mathematics*, chapter Tangent and Normal Cones for Low-Rank Matrices, pages 45–53. Birkhäuser Cham, 2019. doi:10.1007/978-3-030-11370-4_3.
- [23] S. Hosseini and A. Uschmajew. A gradient sampling method on algebraic varieties and application to nonsmooth low-rank optimization. *SIAM Journal on Optimization*, 29(4):2853–2880, 2019. doi:10.1137/17M1153571.
- [24] P. Jain, R. Meka, and I. Dhillon. Guaranteed rank minimization via singular value projection. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23, pages 937–945. Curran Associates, Inc., 2010. URL: <https://proceedings.neurips.cc/paper/2010/hash/08d98638c6fcd194a4b1e6992063e944-Abstract.html>
- [25] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of Computing*, STOC ’13, pages 665–674, New York, NY, USA, June 2013. Association for Computing Machinery. doi:10.1145/2488608.2488693.
- [26] X. Jia, C. Kanzow, P. Mehlitz, and G. Wachsmuth. An augmented Lagrangian method for optimization problems with structured geometric constraints. *Mathematical Programming*, 199:1365–1415, May 2023. doi:10.1007/s10107-022-01870-z.

- [27] V. Lakshmibai and J. Brown. *The Grassmannian Variety*, volume 42 of *Developments in Mathematics*. Springer New York, NY, 2015. doi:10.1007/978-1-4939-3082-1.
- [28] J. M. Lee. *Introduction to Riemannian Manifolds*, volume 176 of *Graduate Texts in Mathematics*. Springer, Cham, 2nd edition, 2018. doi:10.1007/978-3-319-91755-9.
- [29] E. Levin, J. Kileel, and N. Boumal. Finding stationary points on bounded-rank matrices: a geometric hurdle and a smooth remedy. *Mathematical Programming*, 199:831–864, May 2023. doi:10.1007/s10107-022-01851-2.
- [30] E. Levin, J. Kileel, and N. Boumal. The effect of smooth parametrizations on nonconvex optimization landscapes. *Mathematical Programming*, 2024. doi:10.1007/s10107-024-02058-3.
- [31] G. Li and T. K. Pong. Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization*, 25(4):2434–2460, 2015. doi:10.1137/140998135.
- [32] J. Li, A. M.-C. So, and W.-K. Ma. Understanding notions of stationarity in nonsmooth optimization: A guided tour of various constructions of subdifferential for nonsmooth functions. *IEEE Signal Processing Magazine*, 37(5):18–31, September 2020. doi:10.1109/MSP.2020.3003845.
- [33] X. Li and Z. Luo. Normal cones intersection rule and optimality analysis for low-rank matrix optimization with affine manifolds. *SIAM Journal on Optimization*, 33(3):1333–1360, 2023. doi:10.1137/22M147863X.
- [34] X. Li, W. Song, and N. Xiu. Optimality conditions for rank-constrained matrix optimization. *Journal of the Operations Research Society of China*, 7(2):285–301, 2019. doi:10.1007/s40305-019-00245-0.
- [35] H. Liu and R. F. Barber. Between hard and soft thresholding: optimal iterative thresholding algorithms. *Information and Inference: A Journal of the IMA*, 9(4):899–933, December 2020. doi:10.1093/imaiai/iaz027.
- [36] D. G. Luenberger and Y. Ye. *Linear and Nonlinear Programming*, volume 228 of *International Series in Operations Research & Management Science*. Springer Cham, 5th edition, 2021. doi:10.1007/978-3-030-85450-8.
- [37] Y. Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, Cham, 2nd edition, 2018. doi:10.1007/978-3-319-91578-4.
- [38] L. T. Nguyen, J. Kim, and B. Shim. Low-rank matrix completion: A contemporary survey. *IEEE Access*, 7:94215–94237, 2019. doi:10.1109/ACCESS.2019.2928130.
- [39] G. Olikier and P.-A. Absil. On the continuity of the tangent cone to the determinantal variety. *Set-Valued and Variational Analysis*, 30:769–788, 2022. doi:10.1007/s11228-021-00618-9.
- [40] G. Olikier and P.-A. Absil. An apocalypse-free first-order low-rank optimization algorithm with at most one rank reduction attempt per iteration. *SIAM Journal on Matrix Analysis and Applications*, 44(3):1421–1435, 2023. doi:10.1137/22M1518256.
- [41] G. Olikier, K. A. Gallivan, and P.-A. Absil. An apocalypse-free first-order low-rank optimization algorithm. Technical Report UCL-INMA-2022.01, January 2022. arXiv:2201.03962v1.
- [42] G. Olikier, K. A. Gallivan, and P.-A. Absil. Comparison of an apocalypse-free and an apocalypse-prone first-order low-rank optimization algorithm. Technical Report UCL-INMA-2022.02, February 2022. arXiv:2202.09107.
- [43] G. Olikier, A. Uschmajew, and B. Vandereycken. Gauss–Southwell type descent methods for low-rank matrix optimization. July 2023. arXiv:2306.00897v2.

- [44] G. Oliker and I. Waldspurger. Projected gradient descent accumulates at Bouligand stationary points. March 2024. [arXiv:2403.02530v1](https://arxiv.org/abs/2403.02530v1).
- [45] J.-S. Pang, M. Razaviyayn, and A. Alvarado. Computing B-stationary points of nonsmooth DC programs. *Mathematics of Operations Research*, 42(1):95–118, February 2017. doi:10.1287/moor.2016.0795.
- [46] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi. Finding low-rank solutions via nonconvex matrix factorization, efficiently and provably. *SIAM Journal on Imaging Sciences*, 11(4):2165–2204, 2018. doi:10.1137/17M1150189.
- [47] E. Pauwels. Generic Fréchet stationarity in constrained optimization. February 2024. [arXiv:2402.09831v1](https://arxiv.org/abs/2402.09831v1).
- [48] E. Polak. *Computational Methods in Optimization*, volume 77 of *Mathematics in Science and Engineering*. Academic Press, 1971.
- [49] Q. Rebjock and N. Boumal. Optimization over bounded-rank matrices through a desingularization enables joint global and local guarantees. June 2024. [arXiv:2406.14211](https://arxiv.org/abs/2406.14211).
- [50] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag Berlin Heidelberg, 1998. Corrected 3rd printing 2009. doi:10.1007/978-3-642-02431-3.
- [51] H. Scheel and S. Scholtes. Mathematical programs with complementarity constraints: Stationarity, optimality, and sensitivity. *Mathematics of Operations Research*, 25(1):1–22, February 2000. doi:10.1287/moor.25.1.1.15213.
- [52] R. Schneider and A. Uschmajew. Convergence results for projected line-search methods on varieties of low-rank matrices via Łojasiewicz inequality. *SIAM Journal on Optimization*, 25(1):622–646, 2015. doi:10.1137/140957822.
- [53] N. Srebro and T. Jaakkola. Weighted low-rank approximations. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML’03, pages 720–727. AAAI Press, 2003.
- [54] J. Tanner and K. Wei. Normalized iterative hard thresholding for matrix completion. *SIAM Journal on Scientific Computing*, 35(5):S104–S125, 2013. doi:10.1137/120876459.
- [55] T. Tong, C. Ma, and Y. Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *Journal of Machine Learning Research*, 22(150):1–63, 2021. URL: <https://jmlr.org/papers/v22/20-1067.html>.
- [56] A. Uschmajew and B. Vandereycken. Greedy rank updates combined with Riemannian descent methods for low-rank optimization. In *2015 International Conference on Sampling Theory and Applications (SampTA)*, pages 420–424, 2015. doi:10.1109/SAMPPTA.2015.7148925.
- [57] A. Uschmajew and B. Vandereycken. *Handbook of Variational Methods for Nonlinear Geometric Data*, chapter Geometric Methods on Low-Rank Matrix and Tensor Manifolds, pages 261–313. Springer International Publishing, Cham, 2020. doi:10.1007/978-3-030-31351-7_9.
- [58] A. Uschmajew and B. Vandereycken. On critical points of quadratic low-rank matrix optimization problems. *IMA Journal of Numerical Analysis*, 40(4):2626–2651, October 2020. doi:10.1093/imanum/drz061.
- [59] B. Vandereycken. Low-rank matrix completion by Riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013. doi:10.1137/110845768.

- [60] Y. Wang, W. Yin, and J. Zeng. Global convergence of ADMM in nonconvex nonsmooth optimization. *Journal of Scientific Computing*, 78:29–63, January 2019. doi:10.1007/s10915-018-0757-z.
- [61] K. Wei, J.-F. Cai, T. F. Chan, and S. Leung. Guarantees of Riemannian optimization for low rank matrix recovery. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1198–1222, 2016. doi:10.1137/15M1050525.
- [62] M. Willem. *Functional Analysis: Fundamentals and Applications*. Cornerstones. Birkhäuser New York, 2013. doi:10.1007/978-1-4614-7004-5.
- [63] Y. Yang, Y. Feng, and J. A. K. Suykens. Correntropy based matrix completion. *Entropy*, 20(3), March 2018. doi:10.3390/e20030171.
- [64] G. Zhou, W. Huang, K. A. Gallivan, P. Van Dooren, and P.-A. Absil. A Riemannian rank-adaptive method for low-rank optimization. *Neurocomputing*, 192:72–80, June 2016. doi:10.1016/j.neucom.2016.02.030.
- [65] X. Zhou, C. Yang, H. Zhao, and W. Yu. Low-rank modeling and its applications in image analysis. *ACM Computing Surveys*, 47(2), January 2015. doi:10.1145/2674559.
- [66] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin. The global optimization geometry of low-rank matrix optimization. *IEEE Transactions on Information Theory*, 67(2):1308–1331, February 2021. doi:10.1109/TIT.2021.3049171.