Towards optimization on varieties

Eitan Levin

a senior thesis submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Arts in Mathematics at

Princeton University

Adviser: Nicolas Boumal and Joe Kileel

May 4, 2020

Abstract

Many optimization problems over matrices arising in applications are believed to have low-rank solutions. We may be able to efficiently solve such problems by optimizing only over low-rank matrices. If the rank of the solution is known, we can optimize over the set of matrices having precisely that rank, which is a smooth manifold. If we only have an upper bound on the rank, we can optimize over the set of matrices whose rank is appropriately bounded, which is an algebraic variety. While there exist several algorithms for optimization over smooth manifolds possessing good theoretical guarantees, no such theory exists for optimization over (singular) varieties.

We consider the problem of optimizing over varieties in general, and over bounded-rank matrices in particular. We show that in general optimization over varieties is poorly behaved near singularities. Therefore, we consider parametrizing the variety in such a way that the parameter space, which we call a 'lift' because it typically lies in a higher dimensional space, is a smooth manifold. Unfortunately, we show that such lifts can also be poorly behaved—a point that appears to be a stationary point or even a local minimum on the lifted space may not be stationary on the variety. We show that the existence of these 'false stationary points' on lifts is unavoidable for a large class of algebraic varieties.

Specializing to the variety of bounded-rank matrices, we study three lifts and show that the above pathologies are observed in all three of them. Nevertheless, we show that second order critical points on these lifts are guaranteed to correspond to first order critical points on the variety. Moreover, if we add a natural regularization term to the cost function, we can guarantee that any stationary point of the regularized cost is a local minimum on the lift if and only if it corresponds to a local minimum on the variety. Using this lift and associated regularization scheme, we give an algorithm that converges to a stationary point on the variety of bounded-rank, which is a local minimum on the lift if and only if it corresponds to a local minimum on the variety.

We also discuss difficulties encountered when directly optimizing over the variety of bounded-rank matrices. We show that a natural extension of gradient descent to the variety may fail to converge to a stationary point. We then discuss a potential fix involving projection to the singular locus. To analyze this modified algorithm, we also explicitly bound the difference between the metric projection retraction to the bounded-rank matrix variety and its first-order Taylor expansion. Also, we show that the algorithm does converge to a stationary point for varieties which have finitely many singularities. Unfortunately, for general varieties our analysis is unsatisfactory.

Finally, we state some of the open questions and future directions suggested by this work.

Acknowledgements

I would like to thank my thesis advisors Nicolas Boumal and Joe Kileel for many interesting discussions and comments on my work. I would also like to thank Amit Singer for advising my research on cryo-EM, and Tamir Bendory for advising my research on cryo-EM and phase retrieval. I thoroughly enjoyed our work together, and the experience has been invaluable. Thanks also to the rest of our PACM gang, past and present: Ti-Yen Lan, Nick Marshall, Amit Moscovich, Ayelet Heimowitz, Nir Sharon, Will Leeb, Roy Lederman, Amit Halevi, and João Pereira. The second floor of Fine has been a home away from home to me for the past three years, and I'm very grateful for our community.

Last but not least, I want to thank my parents for supporting me all these years. I wouldn't be here without you (literally).

Declaration

I declare that I have not violated the Honor Code during the composition of this work. This paper represents my own work in accordance with University regulations.

I authorize Princeton University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purposes of scholarly research.

Contents

Chapter 1

Optimization over varieties: background, goals, and difficulties

1.1 Introduction

Suppose we want to minimize a function $f(X)$ over matrices in $\mathbb{R}^{m\times n}$ where m, n are very large. If we know that the solution has low rank $k \ll m, n$, we can dramatically reduce storage and perhaps even computation time by restricting f to the smooth manifold of matrices of fixed rank k, denoted by $\mathcal{M}_k^{m\times n}$ (we shall drop the $m \times n$ superscript when the dimensions are clear). Optimization over $\mathcal{M}_k$ can be done directly using the framework of optimization on Riemannian manifolds described e.g. in [1, 3]. If the exact rank of the solution is not known but k is an upper bound on this rank, then we want to optimize over the algebraic variety of all $m \times n$ matrices of rank at most k, denoted $\mathcal{M}_{\leq k}^{m\times n}$ (again, we shall drop $m \times n$ if possible). In contrast to smooth manifolds however, the theory of optimization over varieties is limited.

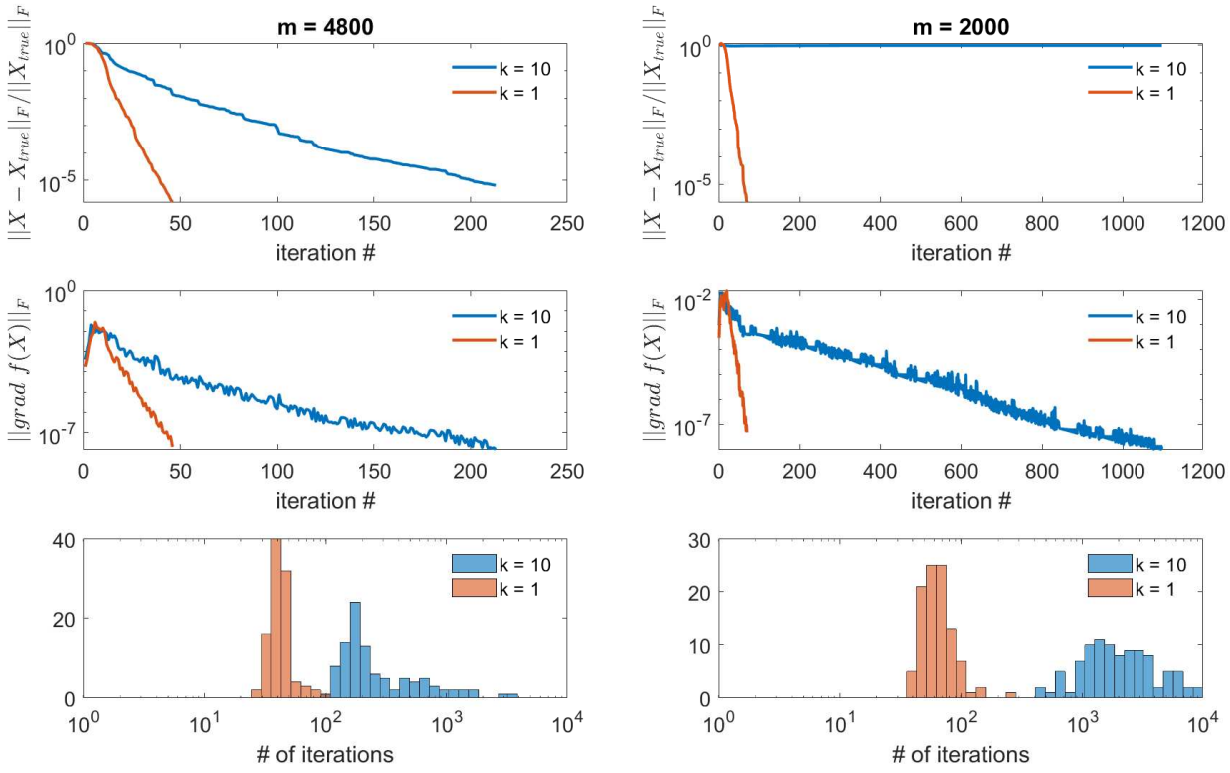
Since the smooth locus of a variety is dense, we can converge to any point on the variety by a sequence of smooth points. Therefore, we might attempt to apply the tools from optimization over smooth manifolds to the smooth locus of a variety. This approach suffers from several drawbacks when converging to a singular point. First, the smooth locus may be disconnected, in which case we may need to initialize from each of the components. Second, as we show in Sec. 3.1, standard algorithms over smooth manifolds may fail to converge to a stationary point on the variety. Third, convergence from a sequence of smooth points to a singular one may be slow. This has been observed in the literature many times, e.g. in [23, Sec. 4], [24, Sec. IV], [22, Sec. 3.4], [25, Sec. 5.5]. Finally, for some problems, overestimating the rank introduces spurious stationary points to which standard algorithms such as gradient descent may converge.

We illustrate the last two drawbacks with a simple experiment. In the so-called generalized phase retrieval problem, we obtain m quadratic measurements of an unknown matrix $X^* \in \mathcal{M}_{\leq k}^{n\times n}$ of the form $\langle X^*, A_i\rangle^2$ where $i = 1, \ldots, m$ and $\langle A, B\rangle = \mathrm{Tr}(A^T B)$ is the Frobenius inner product. If $\mathcal{Q}(X) = (\langle X, A_1\rangle^2, \ldots, \langle X, A_m\rangle^2)^T$ is the measurement operator, to recover X^* we wish to minimize

$$f(X) = \frac{1}{4||\mathcal{Q}(X^*)||^2}||\mathcal{Q}(X) - \mathcal{Q}(X^*)||^2, \tag{1.1}$$

where $\mathcal{Q}(X^*)$ is given. If $\mathrm{rank}(X^*) = r$, it is shown in [26, Sec. 3.2] that $m \geq 2nr - r^2$ measurements from generic matrices A_i of prescribed rank uniquely determine X^*. In other words, under these conditions we have $f(X) = 0$ for $X \in \mathcal{M}_{\leq k}$ iff $X = X^*$, so that X^* is the unique global minimizer of $f(X)$ over $\mathcal{M}_{\leq k}$. Even though we have uniqueness for $m \geq 2nr - r^2$, the function $f(X)$ may have spurious local minima when m is close to the lower bound so actually minimizing $f(X)$ may be difficult. To illustrate the slow convergence stated above, we took $n = 20$ and $m = 12n^2$, generated $X^* = x_l x_r^T$ (so the true rank is 1) and $A_i = a_i b_i^T$ where $x_l, x_r, a_i, b_i \sim \mathcal{N}(0, I_n)$. Here $\mathcal{N}(0, I_n)$ denotes the standard multivariate normal distribution. We then used the implementation of Riemannian gradient descent in Manopt [6] to optimize first over $\mathcal{M}_{10}^{20\times 20}$ where we overestimated the rank to be 10, and then over $\mathcal{M}_1^{20\times 20}$ where we assumed the correct rank (using the same instance of the problem for both). The initialization for the run on $\mathcal{M}_1$ was obtained by projecting the rank 10 initialization from $\mathcal{M}_{10}$ to $\mathcal{M}_1$ by truncating its SVD. The algorithm was terminated when either the cost value $f(X)$ went below 10^{-12} or the gradient norm went below 10^{-8}. The experiment was repeated 100 times with different random instances of the problem and random initialization. We plot a histogram of the number of iterations until termination in Fig. 1.1a. We also plot the error $||X_k - X^*||_F/||X^*||_F$ where $||X||_F = \sqrt{\mathrm{Tr}(X^T X)}$ and the norm of the Riemannian gradient grad $f(X)$ (see Sec. 1.2.1) vs. the iteration number for a randomly chosen instance of the problem. We can clearly see that overestimating the rank leads to significantly slower convergence. To illustrate the existence of bad stationary points to which the algorithm may converge, we took $m = 5n^2$ in the same experiment and plotted the errors in Fig. 1.1b. In this experiment, when we overestimated the rank we converged to a bad stationary point with too large a rank, which we suspect to be a local minimum (Riemannian trust-regions, a second order algorithm, converges to the same point when initialized nearby). Again, the experiment was repeated 100 times. The Matlab code used to perform these experiments is available on github: https://github.com/eitangl/eitanl_thesis_2019-20.

Since the bad cases above occur when we overestimate the rank, several schemes have been proposed in the literature to start from a small initial rank and adaptively increase it, see e.g. [23, 24, 22, 28]. This approach seems to perform well in practice, but lacks theoretical guarantees in general. For the references



(a) (b)

Figure 1.1: Generalized phase retrieval experiment with m measurements by rank 1 random gaussian matrices of a 20×20 rank 1 matrix. The optimization was done using Riemannian gradient descent over the smooth manifold of fixed rank k matrices, where k is given in the legend above. Each experiment was repeated 100 times.

cited above, there are no guarantees proved in the first two references, the reference [22] only considers a specific class of problems, and the results in [28] fall short of guaranteeing that the algorithm converges to a stationary point on the entire variety. In fact, the difficulties involved in proving guarantees for this approach seem to be the same as those outlined in Chap. 3 for optimization directly on the variety.

Another general approach for optimization of polynomial cost functions over varieties is the Sum-Of-Squares (SOS) relaxation, whereby we solve a hierarchy of semidefinite programs (SDPs) [19, 12]. Unfortunately, the size of the SDPs involved quickly becomes unwieldy, so this approach can only be used to solve small instances.

Another approach used in practice for low rank matrix recovery is to parametrize the variety $\mathcal{M}_{\leq k}$ in such a way that the parameter space is a smooth manifold, see e.g. [16, 10, 29, 18]. We call such parameter spaces 'lifts', because they typically lie in higher-dimensional spaces compared to the original variety. In principle, such smooth lifts can be obtained for any variety using resolution of singularities [11]. The most

common way of doing that for the variety of bounded-rank matrices is to note that a matrix $X \in \mathbb{R}^{m\times n}$ has rank $\leq k$ if and only if it can be factored as $X = LR^T$ for $(L, R) \in \mathbb{R}^{m\times k} \times \mathbb{R}^{n\times k}$. Instead of optimizing $X \mapsto f(X)$ over the singular variety $\mathcal{M}_{\leq k}$, we can therefore optimize $(L, R) \mapsto f(LR^T)$ over the Euclidean space $\mathbb{R}^{m\times k} \times \mathbb{R}^{n\times k}$. All the standard theory for optimization over smooth manifolds applies to the latter problem, so we can guarantee convergence to a stationary point on the lift. The difficulty with this approach lies in guaranteeing that a point which is stationary on the lifted space corresponds to a stationary point on the variety. In fact, for many varieties there is no lift to a smooth manifold that preserves first order critical points in the above sense, as we show in Sec. 2.1.2. For the variety of bounded-rank matrices, we study three possible lifts in Chap. 2. Specifically, for each of the three lifts we ask whether critical points and local minima on the lift correspond to critical points and local minima, respectively, on the variety. First order critical points on these lift do not necessarily correspond to first order critical points on the variety and in fact, no such lift to a smooth manifold exists by our more general results in Sec. 2.1.2. However, second order critical points on these lifts do correspond to first order critical points on the variety. Also, all three lifts contain points which can be local minima on the lifts but correspond to saddles on the variety. Nevertheless, for one of the lifts we can identify a set of points that cannot be such 'false local minima' and show how to guarantee convergence to such points.

For the remainder of this chapter, we state some results from the theory of optimization over smooth manifolds, and show that many standard properties relied upon in that theory fail to hold near singularities of varieties in general. In the next two chapters, we then specialize to the variety of bounded-rank matrices.

1.2 Preliminaries

In this section, we review the basic definitions and algorithms in optimization on Riemannian manifolds and their generalization to algebraic varieties. We assume that the reader is familiar with the basics of Riemannian geometry, see [1, 3] for a treatment aimed at applications to optimization, and [13, 14] for a more comprehensive and standard treatment.

1.2.1 Optimality conditions

Suppose we have a smooth function $f\colon \mathcal{M} \to \mathbb{R}$ where $\mathcal{M}$ is a smooth manifold. Recall that the Riemannian gradient of f at a point $x \in \mathcal{M}$ is defined as the unique vector in the tangent space $T_x\mathcal{M}$ of $\mathcal{M}$ at x satisfying

$$Df(x)[v] = \langle \operatorname{grad} f(x), v\rangle_x, \text{ for all } v \in T_x\mathcal{M}, \tag{1.2}$$

where $Df(x)[v]$ is the differential of f at x applied to v. The Riemannian Hessian of f at $x \in \mathcal{M}$ along $v \in T_x\mathcal{M}$ is defined by

$$\text{Hess } f(x)[v] = \nabla_v \text{grad } f, \tag{1.3}$$

where ∇ denotes the Levi-Civita connection on $\mathcal{M}$. It can be shown that this is a self-adjoint linear operator on $T_x\mathcal{M}$ [3, Sec. 5.5]. The first and second order optimality conditions then look the same as their Euclidean counterparts:

Definition 1 (First and second order optimality conditions)**.** A point $x \in \mathcal{M}$ is first order critical (abbreviated 1-critical) for f if grad $f(x) = 0$. It is second order critical (abbreviated 2-critical) for f if it is 1-critical and if Hess $f(x) \succeq 0$ on $T_x\mathcal{M}$.

1.2.2 Retractions

In order to construct optimization algorithms on a manifold, we need a way to move on the manifold from a point $x \in \mathcal{M}$ along a direction $v \in T_x\mathcal{M}$. In Euclidean space, we could simply move to $x + v$. For a general manifold however, even if the sum $x+v$ is well-defined (e.g. if $\mathcal{M}$ is embedded in Euclidean space) it may not lie on $\mathcal{M}$. More generally, the natural way to take such a step is to follow the exponential map to $\text{Exp}_x(v)$. Unfortunately, evaluating the exponential map is costly in general. We therefore define a retraction to be a smooth map that agrees with the exponential map to first order, and we define a second order retraction as a retraction that further agrees with the exponential map to second order:

Definition 2 (Retraction)**.** A retraction is a smooth map $R\colon T\mathcal{M} \to \mathcal{M}$, such that its restrictions $R_x\colon T_x\mathcal{M} \to \mathcal{M}$ satisfy: (1) $R_x(0) = x$ for all $x \in \mathcal{M}$; (2) $DR_x(0) = \text{id}_{T_x\mathcal{M}}$.

A retraction is called second order if $\frac{D^2}{dt^2}\Big|_{t=0} R_x(tv) = 0$ for all $x \in \mathcal{M}$ and $v \in T_x\mathcal{M}$, where D/dt denotes the covariant derivative along the curve $t \mapsto R_x(tv)$. In other words, the retraction curve $t \mapsto R_x(tv)$ has zero initial intrinsic acceleration.

1.2.3 Riemannian gradient descent

We now generalize gradient descent to Riemannian manifolds, and give a proof of convergence to show the key ingredients that are being used, and to later show that they fail to hold in general over singular varieties.

Riemannian gradient descent is given in Alg. 1.

Algorithm 1 Riemannian gradient descent

procedure RGD(f, x_0, τ, β, t_0) ▷ Function f, initial guess x_0, linesearch parameters τ, β, t_0.

  for $i = 0, 1, 2, \ldots$ **do**

    $v_i \leftarrow -\text{grad } f(x_i)$ ▷ descent direction

    $t_i \leftarrow$ backtrack-RM($f, x_i, v_i, \tau, \beta, t_0$) ▷ choose step size

    $x_{i+1} \leftarrow R_{x_i}(t_i v_i)$ ▷ move

  end for

end procedure

Algorithm 2 Backtracking linesearch on Riemannian manifolds

procedure BACKTRACK-RM($f, x, v, \tau, \beta, t_0$) ▷ Function f, point $x \in \mathcal{M}$, direction $v \in T_x\mathcal{M}$, suff. decrease factor $\tau \in (0, 1)$, geom. decrease rate $\beta \in (0, 1)$, and initial stepsize function t_0

  $t \leftarrow t_0(x)$

  while $f(x) - f(R_x(tv)) < \tau t \langle -\text{grad } f(x), v \rangle$ **do**

    $t \leftarrow \beta t$

  end while

  return t

end procedure

A common assumption used to analyze RGD is the following analog of having a Lipschitz gradient in Euclidean space [5]:

Definition 3 (Retraction-Lipschitz gradient)**.** A function $f \colon \mathcal{M} \to \mathbb{R}$ has retraction-Lipschitz gradient if there exists a constant $L > 0$ satisfying

$$f(R_x(v)) \leq f(x) + \langle \text{grad } f(x), v \rangle_x + \frac{L}{2} ||v||_x^2, \tag{1.4}$$

for all $x \in \mathcal{M}$ and $v \in T_x\mathcal{M}$.

Using this assumption, we can lower-bound the step size chosen by Alg. 2 away from zero. Indeed, if step size t is returned by the algorithm, then either it is the initial step size or t/β does not satisfy the sufficient decrease condition so

$$f(x) - f(R_x(-(t/\beta)\text{grad } f(x))) < \frac{\tau t}{\beta} ||\text{grad } f(x)||_x^2. \tag{1.5}$$

On the other hand, by the retraction-Lipschitz property we have

$$f(x) - f(R_x(-t\text{grad } f(x))) \geq \left(\frac{t}{\beta} - \frac{Lt^2}{2\beta^2}\right) ||\text{grad } f(x)||_x^2. \tag{1.6}$$

Together, the above two inequalities give $t \geq 2\beta(1-\tau)/L$. If the function value is bounded from below by f_{low} and we run Alg. 1 for T steps, we then have

$$f(x_0) - f_{\text{low}} \geq f(x_0) - f(x_T) = \sum_{i=0}^{T-1} \Big[f(x_i) - f(x_{i+1})\Big] \geq \sum_{i=0}^{T-1} t_i ||\text{grad } f(x_i)||^2 \geq \frac{2\beta(1-\tau)}{L} \sum_{i=0}^{T-1} ||\text{grad } f(x_i)||^2. \tag{1.7}$$

Thus, we conclude that $\sum_{i=0}^{\infty} ||\text{grad } f(x_i)||^2 < \infty$ so $||\text{grad } f(x_i)|| \to 0$, and that

$$\min_{i=0,\ldots,T-1} ||\text{grad } f(x_i)|| \leq \sqrt{\frac{(f(x_0) - f_{\text{low}})L}{2\beta(1-\tau)}} \cdot \frac{1}{\sqrt{T}}, \tag{1.8}$$

giving a $1/\sqrt{T}$ global rate of convergence. If the sequence (x_i) has a cluster point x^* in $\mathcal{M}$, then that point must have grad $f(x^*) = 0$ by continuity of the gradient vector field, so it is 1-critical for f on $\mathcal{M}$.

1.2.4 Embedded manifolds

If $\mathcal{M} \subset \mathbb{R}^n$ is embedded in Euclidean space, the cost $f\colon \mathcal{M} \to \mathbb{R}$ can be extended to a neighborhood of $\mathcal{M}$ in $\mathbb{R}^n$ [3, Sec. 3.3]. Using this extension, we get a simple expression for the Riemannian gradient as

$$\text{grad } f(x) = \Pi_x \nabla f(x), \tag{1.9}$$

where $\nabla f(x)$ is the usual Euclidean gradient of f and $\Pi_x\colon \mathbb{R}^n \to T_x\mathcal{M}$ is the orthogonal projection onto the tangent space at x. Therefore, a point $x \in \mathcal{M}$ is 1-critical for f iff $\nabla f(x) \perp T_x\mathcal{M}$.

For an embedded manifold, a natural candidate for a retraction is the metric projection $R_x(v) = \arg\min_{y \in \mathcal{M}} ||x + v - y||$. However, for a general non-convex set $\mathcal{M}$ the argmin may not exist or it may not be singleton. Nevertheless, if $\mathcal{M}$ is a smooth manifold then this expression is uniquely defined and smooth for all $x \in \mathcal{M}$ and all small enough $v \in T_x\mathcal{M}$, and moreover it is second order [3, Sec. 5.11],[2]. The caveat is that the definition of a retraction assumes R is defined on the entire tangent bundle $T\mathcal{M}$, which the metric projection above may not. For this work, we will only consider closed subsets of $\mathbb{R}^n$ for which the argmin above always exists. If it is not uniquely defined, we show that making an arbitrary choice suffices to get an optimization algorithm.

1.2.5 Quotient manifolds

In practice, we often want to optimize a function f over a smooth manifold $\overline{\mathcal{M}} \subset \mathbb{R}^n$ where f is invariant under the action of a certain group G. In that case, we may want to optimize directly over the more abstract orbit space $\overline{\mathcal{M}}/G$ while representing the iterates on a computer and doing computations on the more accessible embedded manifold $\overline{\mathcal{M}}$ where points can be viewed as vectors in $\mathbb{R}^n$.

First, we must ensure that the quotient $\overline{\mathcal{M}}/G$ is a smooth manifold. To do so, we must impose some constraints on G and its action on $\overline{\mathcal{M}}$. Specifically, we define

Definition 4. The action of a Lie group G on a manifold $\overline{\mathcal{M}}$, denoted $g \cdot x$ for $g \in G$ and $x \in \overline{\mathcal{M}}$, is called:

- *smooth* if the action map $G \times \overline{\mathcal{M}} \to \overline{\mathcal{M}}$ sending $(g, x) \mapsto g \cdot x$ is smooth;
- *free* if $g \cdot x = x$ for some $x \in \overline{\mathcal{M}}$ implies $g = \mathrm{id}$;
- *proper* if the map $(g, x) \mapsto (g \cdot x, x)$ is a proper map $G \times \overline{\mathcal{M}} \to \overline{\mathcal{M}} \times \overline{\mathcal{M}}$. Recall that a map is proper iff the inverse image of any compact set is compact.

Of the three properties, properness is the hardest to verify. Fortunately, if G is compact then it is satisfied automatically by [13, Cor. 21.6]. We are now ready to state

Theorem 1.1 ([13, Thm. 21.10])**.** *If G is a Lie group acting on a smooth manifold $\overline{\mathcal{M}}$ smoothly, freely, and properly, then $\mathcal{M} = \overline{\mathcal{M}}/G$ is a smooth manifold of dimension* $\dim \overline{\mathcal{M}} - \dim G$.

The manifold $\overline{\mathcal{M}}$ is called the *total space*.

Let $\pi \colon \overline{\mathcal{M}} \to \mathcal{M}$ denote the quotient map, and for any $x \in \overline{\mathcal{M}}$ denote its orbit by $[x] = \pi(x)$. For any $x \in \overline{\mathcal{M}}$ we can decompose the tangent space as $T_x\overline{\mathcal{M}} = V_x \oplus H_x$ where $V_x = \ker D\pi(x)$ is called the vertical space and $H_x = V_x^{\perp}$ is called the horizontal space. Then the restriction $D\pi(x)|_{H_x} \colon H_x \to T_{[x]}\mathcal{M}$ is an isomorphism whose inverse we denote by $\mathrm{lift}_x \colon T_{[x]}\mathcal{M} \to H_x$. If the metric on $\overline{\mathcal{M}}$ satisfies

$$[x] = [y] \implies \langle \mathrm{lift}_x(\xi), \mathrm{lift}_x(\zeta) \rangle_x = \langle \mathrm{lift}_y(\xi), \mathrm{lift}_y(\zeta) \rangle_y, \quad \text{for all } \xi, \zeta \in T_{[x]}\mathcal{M}, \tag{1.10}$$

then we can define a metric on $\mathcal{M}$ by

$$\langle \xi, \zeta, \rangle_{[x]} = \langle \mathrm{lift}_x(\xi), \mathrm{lift}_x(\zeta) \rangle_x. \tag{1.11}$$

With this metric, $\mathcal{M}$ is called a Riemannian quotient manifold of $\overline{\mathcal{M}}$ [3, Sec. 9.7]. We shall assume $\mathcal{M}$ is a Riemannian quotient manifold of $\overline{\mathcal{M}}$ from now on.

If we have a cost function $\bar{f}\colon \overline{\mathcal{M}} \to \mathbb{R}$ which is invariant under the group action, then it induces a cost $f\colon \mathcal{M} \to \mathbb{R}$ satisfying $\bar{f} = f \circ \pi$. The former is smooth iff the latter is smooth, and their Riemannian gradients (with respect to the Riemannian quotient metric defined in Eq. (1.11)) satisfy [3, Sec. 9.8]

$$\mathrm{lift}_x(\mathrm{grad}\ f([x])) = \mathrm{grad}\ \bar{f}(x). \tag{1.12}$$

If we denote $\Pi_x^H : \mathbb{R}^n \to H_x$ the orthogonal projection to the horizontal space at x, the Hessians of $f, \bar{f}$ are related by

$$\mathrm{lift}_x(\mathrm{Hess}\ f([x])(\xi)) = \Pi_x^H(\mathrm{Hess}\ \bar{f}(x)[\mathrm{lift}_x(\xi)]), \quad \text{for all } \xi \in T_{[x]}\mathcal{M}. \tag{1.13}$$

We are now ready to prove the equivalence of the first and second order optimality conditions on $\mathcal{M}$ and $\overline{\mathcal{M}}$:

Proposition 1.2.

- $[x] \in \mathcal{M}$ is 1-critical for f if and only if $x \in \overline{\mathcal{M}}$ is 1-critical for $\bar{f}$.
- $[x] \in \mathcal{M}$ is 2-critical for f if and only if x is 2-critical for $\bar{f}$ on $\overline{\mathcal{M}}$.

Proof. Since lift_x is an isomorphism, by Eq. (1.12) we have $\mathrm{grad}\ f([x]) = 0$ iff $\mathrm{grad}\ \bar{f}(x) = 0$, giving an equivalence between 1-criticality on the total and quotient manifolds.

By Eq. (1.13), we have

$$\langle \mathrm{Hess}\ f([x])[\xi], \xi\rangle_{[x]} = \langle \mathrm{Proj}_x^H(\mathrm{Hess}\ \bar{f}(x)[\mathrm{lift}_x(\xi)], \mathrm{lift}_x(\xi)\rangle_x = \langle \mathrm{Hess}\ \bar{f}(x)[\mathrm{lift}_x(\xi)], \mathrm{lift}_x(\xi)\rangle_x, \tag{1.14}$$

where for the last equality we used the fact that $\mathrm{lift}_x(\xi) \in H_x$ to discard the projection onto H_x. Thus, $[x]$ is 2-critical for f on $\mathcal{M}$ iff $\mathrm{Hess}\ \bar{f}(x)|_{H_x} \succeq 0$. In particular, if x is 2-critical for $\bar{f}$ on $\overline{\mathcal{M}}$ then $[x]$ is 2-critical for f on $\mathcal{M}$.

For the converse, observe that at 1-critical points of $\bar{f}$ we have $\mathrm{Hess}\ \bar{f}(x)[v] = 0$ for all $v \in V_x$, the 'vertical space'. Indeed, for any $v \in V_x$ we can find a smooth curve $c\colon (-\epsilon, \epsilon) \to \pi^{-1}([x])$ by [3, Prop. 9.3]. We then have

$$\mathrm{Hess}\ \bar{f}(x)[v] = \nabla_v \mathrm{grad}\ \bar{f}(x) = \left.\frac{D}{dt}\right|_{t=0} \mathrm{grad}\ \bar{f}(c(t)) = \left.\frac{D}{dt}\right|_{t=0} \mathrm{lift}_{c(t)}\mathrm{grad}\ f([c(t)]) = \left.\frac{D}{dt}\right|_{t=0} \mathrm{lift}_{c(t)}\mathrm{grad}\ f([x]) = 0, \tag{1.15}$$

where we used $[c(t)] = [x]$ for all t and $\mathrm{grad}\ f([x]) = 0$ by 1-criticality. For any $w \in T_x\overline{\mathcal{M}}$ decompose it as

$w = w^H + w^V$ where $w^H \in H_x$ and $w^V \in V_x$. Then

$$\begin{aligned}\langle \text{Hess } \bar{f}(x)[w], w\rangle &= \langle \text{Hess } \bar{f}(x)[w^H], w^H\rangle + \langle \text{Hess } \bar{f}(x)[w^H], w^V\rangle + \langle \underbrace{\text{Hess } \bar{f}(x)[w^V]}_{=0}, w\rangle \\ &= \langle \text{Hess } \bar{f}(x)[w^H], w^H\rangle + \langle w^H, \underbrace{\text{Hess } \bar{f}(x)[w^V]}_{=0}\rangle \\ &= \langle \text{Hess } \bar{f}(x)[w^H], w^H\rangle,\end{aligned} \tag{1.16}$$

where in going from the first to the second lines we used the fact that the Hessian is self-adjoint. Thus, at 1-critical points for $\bar{f}$ we have Hess $\bar{f}(x)|_{H_x} \succeq 0$ if and only if Hess $\bar{f}(x) \succeq 0$, so we conclude that if $[x]$ is 2-critical for f on $\mathcal{M}$ then x is 2-critical for $\bar{f}$ on $\overline{\mathcal{M}}$. □

In particular:

Corollary 1.3. *A point $x \in \overline{\mathcal{M}}$ is 2-critical for $\bar{f}$ iff grad $\bar{f}(x) = 0$ (so x is 1-critical) and $\langle$Hess $\bar{f}(x)[v], v\rangle \geq 0$ for all $v \in H_x$ (so we only need to check non-negativity on the horizontal space).*

Proof. This statement is equivalent to showing that at 1-critical points for $\bar{f}$ we have $\langle \text{Hess } \bar{f}(x)[v], v\rangle \geq 0$ for all $v \in H_x$ implies the same inequality for all $v \in T_x\overline{\mathcal{M}}$. This follows from Eq. (1.16). □

Two of the lifted spaces we consider for $\mathcal{M}_{\leq k}$ will be quotient manifolds, and it will be much simpler for us to state the optimality conditions on their total spaces than on the quotients. By Prop. 1.2, the two are equivalent.

1.2.6 Varieties

For the purpose of this thesis, we only consider real algebraic varieties. Furthermore, for the purpose of optimization we want to work over closed subsets of $\mathbb{R}^n$ to ensure that the point to which our algorithm converges is in that same set. Thus, we only consider closed varieties in $\mathbb{R}^n$, i.e. Zariski closed sets which are precisely the sets defined as zero loci of systems of real polynomial equations.

We generalize the optimality conditions introduced for smooth manifolds in Sec. 1.2.1 as well as the RGD algorithm Alg. 1, following [21]. Since the smooth locus of a variety $\mathcal{V}$ is open in $\mathcal{V}$ and is a smooth manifold, the optimality conditions from Sec. 1.2.4 are unchanged at smooth points. Namely, a smooth point $x \in \mathcal{V}$ is 1-critical for $f : \mathcal{U} \to \mathbb{R}$ defined in a Euclidean neighborhood $\mathcal{U} \supseteq \mathcal{V}$ of the variety iff $\nabla f(x) \perp T_x\mathcal{V}$ where $T_x\mathcal{V}$ is the tangent space to $\mathcal{V}$ at x and $\nabla f(x)$ is the Euclidean gradient of f at x. If x is singular on $\mathcal{V}$, instead of the tangent space we need to consider the tangent cone defined as

$$T_x\mathcal{V} = \{u \in \mathbb{R}^n : \text{there exist } (\alpha_k) \subset \mathbb{R}_{>0},\ (x_k) \subset \mathcal{V} \text{ such that } x_k \to x \text{ and } \alpha_k(x_k - x) \to u\}. \tag{1.17}$$

In other words, the tangent cone is the limit of all secant rays on $\mathcal{V}$ having x as an endpoint. It is always a closed cone, but not necessarily convex [20, Sec. 3.3.1]. If x is smooth, this definition coincides with the tangent space at x [20, Thm. 3.15]. The polar of the tangent cone is defined as

$$[T_x\mathcal{M}]^\circ = \{y \in \mathbb{R}^n : \langle y, u\rangle \leq 0 \text{ for all } u \in T_x\mathcal{M}\}. \tag{1.18}$$

If x is smooth then $T_x\mathcal{M}$ is a linear space, so $[T_x\mathcal{M}]^\circ = [T_x\mathcal{M}]^\perp$ is the normal space at x. Using this notion, we can generalize the first order optimality condition:

Definition 5 (First order optimality on a variety)**.** A point $x \in \mathcal{V}$ is first order critical for $f : \mathcal{V} \to \mathbb{R}$ iff $-\nabla f(x) \in [T_x\mathcal{M}]^\circ$.

Another issue we need to consider is retracting a tangent vector to the variety at a singular point. For a smooth manifold, we require a retraction to be smooth and agree with the exponential map to first order. At singular points however, the tangent cone is not a linear space in general, so differentiation on the cone may not be well-defined. Fortunately, for the purposes of getting first order algorithms on the variety (algorithms using only the gradient), it suffices to require [21, Defn. 2.4]:

Definition 6 (Retraction on a variety)**.** A map $R\colon \bigcup_{x\in\mathcal{V}}\{x\} \times T_x\mathcal{V} \to \mathcal{V}$ is a retraction if for any fixed $x \in \mathcal{V}$ and $u \in T_x\mathcal{V}$ the map $\alpha \mapsto R_x(\alpha u)$ is continuous on $[0, r)$ for some $r \in \mathbb{R}_{>0}$ and

$$\lim_{\alpha\to 0^+} \frac{R_x(\alpha u) - (x + \alpha u)}{\alpha} = 0. \tag{1.19}$$

In other words, $R_x(u)$ agrees with $x + u$ to first order. With this definition, metric projection remains a retraction [21, Sec. 2.4.1].

We can now extend Alg. 1 to varieties simply by replacing $\operatorname{grad} f(x) \mapsto \Pi_x \nabla f(x)$ where $\Pi_x\colon \mathbb{R}^n \to T_x\mathcal{V}$ denotes projection onto the tangent cone at x, and the retraction $R_x(v)$ is understood in the sense of the above looser definition. The result is given in Alg. 3, which coincides with Alg. 1 for embedded manifolds. Note that $T_x\mathcal{V}$ is not necessarily convex, so Π_X may not be single-valued. In this case, making an arbitrary choice for the projection suffices for the resulting algorithm to be well-defined (i.e. for a step size satisfying sufficient decrease as required by Alg. 4 to exist), see [21, Prop. 2.8]. Note that if $v \in [T_x\mathcal{V}]^\circ$ then $\Pi_x v = 0$, since if $u \in T_x\mathcal{V} \setminus \{0\}$ then

$$||v - u||^2 = ||v||^2 + \underbrace{||u||^2}_{>0} \underbrace{-2\langle u, v\rangle}_{\geq 0} > ||v||^2. \tag{1.20}$$

Therefore, Alg. 3 terminates if it reaches a 1-critical point.

Algorithm 3 Gradient descent on varieties

procedure GD-V(f, x_0, τ, β, t_0) ▷ Function f, initial guess x_0, linesearch parameters τ, β, t_0.

for $i = 0, 1, 2, \ldots$ **do**

$v_i \leftarrow \Pi_{x_i}[-\nabla f(x_i)]$ ▷ descent direction

$t_i \leftarrow \text{backtrack}(f, x_i, v_i, \tau, \beta, t_0)$ ▷ choose step size

$x_{i+1} \leftarrow R_{x_i}(t_i v_i)$ ▷ move

end for

end procedure

Algorithm 4 Backtracking linesearch on varieties

procedure BACKTRACK($f, x, v, \tau, \beta, t_0$) ▷ Function f, point $x \in \mathcal{M}$, direction $v \in T_x\mathcal{M}$, suff. decrease factor $\tau \in (0, 1)$, geom. decrease rate $\beta \in (0, 1)$, and initial stepsize function t_0

$t \leftarrow t_0(x)$

while $f(x) - f(R_x(tv)) < \tau t \langle -\nabla f(x), v \rangle$ **do**

$t \leftarrow \beta t$

end while

return t

end procedure

Finally, we describe the notion of resolution of singularities for varieties. As discussed in Sec. 1.1, optimizing directly over singular varieties can be problematic. Instead, we may attempt to find a smooth parametrization for the variety and optimize over the parameter space, which we call a *lift*. Such a smooth lift always exists for varieties. Specifically, any real variety $\mathcal{V}$ has a *resolution of singularities*, which consists of a smooth variety $\overline{\mathcal{V}}$ and a proper birational morphism of varieties $\varphi\colon \overline{\mathcal{V}} \to \mathcal{V}$. Morphism of varieties means that the coordinate functions of φ are polynomials in the entries of the inputs (in particular, it is smooth) and birational means it is 'almost invertible' — there are open dense subsets $U \subset \overline{\mathcal{V}}$ and $W \subset \mathcal{V}$ such that $\varphi\colon U \to W$ is an isomorphism of varieties. Proper means that inverse images of compact sets are compact[1]. See [11] for introduction and details.

[1]The algebraic definition of properness is equivalent to the topological one we state here, see discussion in https://mathoverflow.net/questions/20879/when-is-a-morphism-proper.

1.2.7 The variety of bounded-rank matrices and its stratification

For the variety of bounded rank matrices $\mathcal{M}_{\leq k}$, the tangent cone is given by

$$T_X\mathcal{M}_{\leq k}^{m\times n} = \{V \in \mathbb{R}^{m\times n} : \operatorname{rank}(P_{\operatorname{col}(X)^\perp} V P_{\operatorname{row}(X)^\perp}) \leq k - \operatorname{rank}(X)\}, \tag{1.21}$$

where $P_{\operatorname{col}(X)^\perp} : \mathbb{R}^m \to \operatorname{col}(X)^\perp$ is the orthogonal projection onto the orthogonal complement of the column space of X, and similarly $P_{\operatorname{row}(X)^\perp} : \mathbb{R}^n \to \operatorname{row}(X)^\perp$ is the orthogonal projection onto $\operatorname{row}(X)^\perp$. In the basis of the singular vectors of X, if $\operatorname{rank}(X) = r$ we can write

$$\begin{aligned} X &= \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix}, \quad \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_r) \succ 0, \\ T_X\mathcal{M}_{\leq k} &= \left\{ \begin{pmatrix} V_1 & V_2 \\ V_3 & V_4 \end{pmatrix} \in \mathbb{R}^{m\times n} : V_1 \in \mathbb{R}^{r\times r},\ \operatorname{rank}(V_4) \leq k - r \right\}. \end{aligned} \tag{1.22}$$

The singular points of $\mathcal{M}_{\leq k}$ are precisely the points with rank $< k$. As can be readily seen from Eq. (1.22), the tangent cones to these singular points are not convex.

We can write $\mathcal{M}_{\leq k} = \bigsqcup_{r=1}^{k} \mathcal{M}_r$, which is a disjoint union of smooth manifolds. This is a general phenomenon called Whitney stratification. Specifically, any variety $\mathcal{V}$ can be written as a finite disjoint union of smooth submanifolds $\mathcal{V} = \bigsqcup_{i=1}^{m} \mathcal{V}_i$ where each $\mathcal{V}_i$ is called a *stratum*. The tangent spaces to different strata satisfy certain regularity conditions [15, 17]. In particular, if $(X_i) \subset \mathcal{V}_k$ is a sequence such that $X_i \to X \in \mathcal{V}_r$ with $r \neq k$ and the sequence of tangent spaces $T_{X_i}\mathcal{V}_k$ converge to a subspace T (i.e. if the orthogonal projectors Π_{X_i} converge to an orthogonal projector Π_T), then $T \supseteq T_X\mathcal{V}_r$. Consequently, if we have an algorithm on $\mathcal{V}_k$ generating a sequence (X_i) along which the Riemannian gradient tends to zero, i.e. $\Pi_{X_i}\nabla f(X_i) = \operatorname{grad} f(X_i) \to 0$, then since the Grassmannian is compact after passing to a subsequence we may assume that the subspaces $T_{X_i}\mathcal{V}_k$ converge to some T. Since we must have $T \supseteq T_X\mathcal{V}_r$, we have

$$||\Pi_X \nabla f(X)|| \leq ||\Pi_T \nabla f(X)|| = \lim_{i\to\infty} ||\Pi_{X_i} \nabla f(X)|| \leq \lim_{i\to\infty} ||\Pi_{X_i} \nabla f(X_i)|| + \lim_{i\to\infty} ||\Pi_{X_i}(\nabla f(X) - \nabla f(X_i))|| = 0. \tag{1.23}$$

Thus, the limit X is 1-critical to its own stratum.

Note that $[T_X\mathcal{M}_{\leq k}]^\circ = \{0\}$ whenever $\operatorname{rank}(X) < k$. We therefore conclude that a point $X \in \mathcal{M}_{\leq k}$ is 1-critical for f iff either $\operatorname{rank}(X) = k$ and X is 1-critical on its own stratum $\mathcal{M}_k$, or $\operatorname{rank}(X) < k$ and $\nabla f(X) = 0$ (so X is 1-critical on all of $\mathbb{R}^{m\times n}$) [21, Cor. 3.4].

We now obtain a notion of 2-criticality on the variety of bounded-rank matrices $\mathcal{M}_{\leq k}$. Recall (e.g. from

[20, Thm. 3.45]) that a necessary second order condition for a point $\hat{X}$ to be optimal for $\min_{X\in\mathcal{X}} f(X)$ where $\mathcal{X} \subseteq \mathbb{R}^{m\times n}$ is any subset and f is $\mathcal{C}^2$ is that

$$\langle \nabla f(\hat{X}), W\rangle + \langle \nabla^2 f(\hat{X})[Z], Z\rangle \geq 0, \tag{1.24}$$

for every $Z \in T_{\hat{X}}\mathcal{X}$ satisfying $\langle \nabla f(\hat{X}), Z\rangle = 0$ and every $W \in T^2_{\hat{X},Z}\mathcal{X}$. Here $T^2_{\hat{X},Z}\mathcal{X}$ is the second order tangent set defined e.g. in [20, Defn. 3.41], and the tangent cone $T_{\hat{X}}\mathcal{X}$ is defined as in Eq. (1.17). Specializing to $\mathcal{X} = \mathcal{M}_{\leq k}$ and taking a point X which is 1-critical and of rank $< k$, we have $\nabla f(X) = 0$. Therefore, we call a rank-deficient point X a second order critical point for f on $\mathcal{M}_{\leq k}$ if it satisfies

$$\langle \nabla^2 f(X)[Z], Z\rangle \geq 0, \quad \text{for all } Z \in T_X\mathcal{M}_{\leq k}, \tag{1.25}$$

where now $T_X\mathcal{M}_{\leq k}$ is the tangent cone to $\mathcal{M}_{\leq k}$ at X. Note also that if $\operatorname{rank}(X) = k$ then Eq. (1.24) gives precisely $\operatorname{Hess} f(X) \succeq 0$ as shown in [27]. We define

Definition 7 (2-criticality on $\mathcal{M}_{\leq k}$)**.** A point $X \in \mathcal{M}_{\leq k}$ is 2-critical for f on all of $\mathcal{M}_{\leq k}$ if either $\operatorname{rank}(X) = k$ and X is 2-critical for f on the smooth manifold $\mathcal{M}_k$, or $\operatorname{rank}(X) < k$ and $\nabla f(X) = 0$ and $\langle \nabla^2 f(X)[Z], Z\rangle \geq 0$ for all $Z \in T_X\mathcal{M}_{\leq k}$.

1.3 What can go wrong in general

In this section, we outline some of the main properties stated for smooth manifolds in Sec. 1.2 that fail when optimizing over a singular variety.

1.3.1 Metric projection retraction is not locally single-valued, not twice-differentiable

Near a smooth point on the variety, the metric projection is uniquely defined for sufficiently small step sizes and is a second order retraction which means that the intrinsic acceleration of the retraction curve at the base point vanishes, or equivalently that $\frac{d^2}{dt^2}\Big|_{t=0} R_x(t\xi) \in [T_x\mathcal{M}]^\circ$ where $[T_x\mathcal{M}]^\circ$ is the polar of the tangent cone at x (which for a smooth point is simply the normal space), see [2].

Unfortunately, near singularities none of the above properties hold in general. For example, consider the cuspidal cubic $y^2 = x^3$ in the plane. The tangent cone at the origin is $T_0\mathcal{M} = \{\xi_v = (v, 0) : v \geq 0\}$, see Fig. 1.2. The metric projection then reads

$$R_0(\xi_v) = \left(\frac{\sqrt{1+6v}-1}{3}, \pm\left[\frac{\sqrt{1+6v}-1}{3}\right]^{3/2}\right), \tag{1.26}$$

which in particular is not single-valued even for arbitrarily small v. Also, both branches of $R_0(t\xi_v)$ are not twice-differentiable in t at the origin (and in particular is not a second-order retraction): we have $\lim_{t\to 0^+} \frac{d^2}{dt^2} R_0(t\xi_v) = (-3, \pm\infty)$.

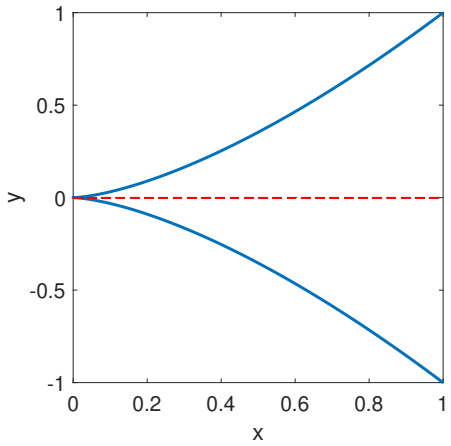


Figure 1.2: Cuspidal cubic $y^2 = x^3$ in solid blue, and its tangent cone at the origin in dashed red.

Note that the retraction may not be single-valued near a singular point on the bounded-rank matrix variety as well. Specifically, the retraction $R_X(V) = P_k(X+V)$ where $X \in \mathcal{M}_{\leq k}$ and P_k is metric projection to $\mathcal{M}_{\leq k}$ is not single-valued whenever the k and $(k+1)$th singular values of $X+V$ are equal. We claim that we can pick k, X, V such that $R_X(tV)$ is not single-valued for any $t \in \mathbb{R}$. For example, let

$$X = \begin{pmatrix} I_2 & 0 \\ 0 & 0 \end{pmatrix} \in \mathcal{M}_{\leq 3}^{4\times 4}, \; V = \begin{pmatrix} 0 & I_2 \\ I_2 & 0 \end{pmatrix} \in T_X \mathcal{M}_{\leq 3}^{4\times 4}. \tag{1.27}$$

Then the 3rd and 4th singular values of $X + tV$ are[2]

$$\sigma_3(X+tV) = \sigma_4(X+tV) = \frac{\sqrt{1+4t^2}-1}{2}, \tag{1.28}$$

so its projection to $\mathcal{M}_{\leq 3}$ is not single-valued for any t.

In the analyses in Sec. 1.2.3 and in [21, Sec. 2.4], the properties of the retraction are used to guarantee that a positive step size satisfying the sufficient decrease condition in Alg. 4 exists. For this purpose, if the retraction has multiple branches then it suffices to make an arbitrary choice of branch. Indeed, if $-\nabla f(x)$ has a nontrivial projection $\Pi_x[-\nabla f(x)] = v \in T_x\mathcal{M}$ then $\langle -\nabla f(x), v\rangle = ||v||^2 > 0$ by [21, Prop. 2.6], hence there exists a small enough step-size giving sufficient decrease by [21, Prop. 2.8].

[2]WolframAlpha link: https://bit.ly/2VD3OGO

1.3.2 Smooth functions do not have locally retraction-Lipschitz gradients

For a smooth manifold, since $f \circ R_x$ is smooth near the origin and $\nabla(f \circ R_x)(0) = \text{grad } f(x)$, the function f satisfies the retraction-Lipschitz property from Defn. 3 locally near each x on the manifold. However, that need not be true if x is a singular point on a variety. For example, for the cuspidal cubic $y^2 = x^3$ considered above (and plotted in Fig. 1.2), its metric projection-based retraction at the origin is given in Eq. (1.26), which is not single-valued. Choosing the positive branch for example, and considering the cost function $f(x, y) = y$, note that

$$f(R_0(t\xi_v)) = (tv)^{3/2} + O((tv)^{5/2}), \tag{1.29}$$

which shows that f does not satisfy the property in Defn. 3 in any neighborhood of the origin.

1.3.3 Linear functions do not have retraction-Lipschitz gradients on $\mathcal{M}_k$

Here we show that even for many linear functions over the smooth manifold of fixed-rank matrices $\mathcal{M}_k$, there is no single constant $L > 0$ satisfying Defn. 3 everywhere for any second order retraction R.

Lemma 1.4. *If there exists $L > 0$ satisfying Defn. 3 with some second order retraction R, then $\sup_{X \in \mathcal{M}_k} \lambda_{max}(\text{Hess } f(X)) < \infty$ where $\lambda_{max}(\cdot)$ is the largest eigenvalue of a matrix.*

Proof. If Defn. 3 is satisfied, we can Taylor expand (using the fact that R is second order, see [3, Sec. 5.10])

$$f(R_X(V)) = f(X) + \langle \nabla f(X), V \rangle + \frac{1}{2}\langle \text{Hess } f(X)[V], V \rangle + O(||V||_F^3) \leq f(X) + \langle \nabla f(X), V \rangle + \frac{L}{2}||V||_F^2. \tag{1.30}$$

Hence

$$\frac{\langle \text{Hess } f(X)[V], V \rangle}{||V||_F^2} \leq L + O(||V||), \tag{1.31}$$

for all $V \in T_X\mathcal{M}_k \setminus \{0\}$. By setting $V = tV_1$ where V_1 is the eigenvector corresponding to the largest eigenvalue $\lambda_{\max}$ of Hess $f(X)$ and letting $t \to 0$, we obtain $\lambda_{\max} \leq L$. □

We can now show:

Proposition 1.5. Let $f(X) = \langle A, X \rangle$ where $\text{rank}(A) \leq \min(n-k, m-k)$. Then

$$\sup_{X \in \mathcal{M}_k} \lambda_{\max}(\text{Hess } f(X)) = \infty$$

.

Proof. Let $A = U_A \Sigma_A V_A^T$ be the thin SVD of A, so U_A is $m \times r$, Σ_A is $r \times r$, and V_A^T is $n \times r$ for $r \leq \min(m-k, n-k)$. Because of the latter inequality, we can choose U, V with orthonormal columns and

of sizes $m \times k$ and $n \times k$ such that $U^T U_A = 0$ and $V^T V_A = 0$. Let $X(t) = tUV^T$ so $\text{rank}(X(t)) = k$ for all $t > 0$. Also set $Z = U_A V^T + U V_A^T \in T_{X(t)}\mathcal{M}_k$ for all $t > 0$. Observe that

$$\text{Hess } f(X(t))[Z] = \frac{1}{t}\left[(I - UU^T)AV_A V^T + UU_A^T A(I - VV^T)\right] = \frac{1}{t}\left[U_A V^T + UV_A^T\right] = \frac{1}{t}Z. \tag{1.32}$$

Thus,

$$\lambda_{\max}(\text{Hess } f(X(t))) \geq \frac{\langle \text{Hess } f(X(t))[Z], Z\rangle}{||Z||_F^2} = \frac{1}{t} \xrightarrow{t\to 0} \infty. \tag{1.33}$$

□

If one requires that f be bounded below on $\mathcal{M}_k$, we can add $\frac{1}{2}||X||^2$ to $f(X)$, which would simply add Z to the Riemannian Hessian and yield the same result. This result applies in particular to the metric projection retraction, which we shall use in Chap. 3.

1.3.4 Gradient converges to zero, but limit point is not a critical point

If $(x_n) \subset \mathcal{M}$ is a sequence converging to a smooth limit point $x^* \in \mathcal{M}$ such that $||\Pi_{x_n} \nabla f(x_n)|| \to 0$, then we must have $||\Pi_{x^*} \nabla f(x^*)|| = 0$ so x^* is 1-critical for f on $\mathcal{M}$ (because Π_x is continuous on a smooth manifold). This was used to prove the convergence of RGD (Alg. 1) to a 1-critical point in Sec. 1.2.3.

However, this need not be the case if x^* is singular. For example, consider the nodal cubic $y^2 = x^2(x+1)$ (plotted on the left panel of Fig. 1.3), whose only singularity is at the origin. Let $f(x, y) = x - y$, and consider the sequence of points $(x_n, y_n) = \left(\frac{1}{n}, \frac{1}{n}\sqrt{1 + \frac{1}{n}}\right) \to x^* = (0, 0)$. The tangent line at (x_n, y_n) is

$$T_{(x_n,y_n)}\mathcal{M} = \left\{(x, y) : y - y_n = \frac{3x_n + 2}{2\sqrt{x_n + 1}}(x - x_n)\right\}, \tag{1.34}$$

and the projection of $\nabla f = (1, -1)$ onto it is

$$\Pi_{(x_n,y_n)} \nabla f(x_n) = \frac{1 - \frac{3x_n+2}{2\sqrt{x_n+1}}}{\sqrt{1 + \frac{(3x_n+2)^2}{4(x_n+1)}}} \cdot \left(1, \frac{3x_n + 2}{2\sqrt{x_n + 1}}\right) \xrightarrow{n\to\infty} (0, 0). \tag{1.35}$$

However, the tangent cone at the origin is $T_0\mathcal{M} = \{y = x\} \cup \{y = -x\}$ so the projection onto it is

$$\Pi_{(0,0)} \nabla f(x^*) = \nabla f(x^*) = (1, -1) \neq 0. \tag{1.36}$$

1.3.5 Local minimum on lift may not be a local minimum on variety

Suppose we want to minimize $f(x, y) = -x - y$ over that same curve $y^2 = x^2(x+1)$ starting from $(1, -\sqrt{2})$. If we parametrize the curve by $t \mapsto (t^2 - 1, t(t^2 - 1))$, we want to minimize $g(t) = f(t^2 - 1, t(t^2 - 1)) = -(t+1)^2(t-1)$ over $\mathbb{R}$ starting from $t = -\sqrt{2}$, instead of optimizing over the curve directly as in the preceding paragraph. The function $g(t)$ is plotted in Fig. 1.4. Since $g(t)$ has a local minimum at $t = -1$, many reasonable optimization algorithms initialized from e.g. $t = -2$ will converge to $t = -1$, that is, $(0, 0)$. However, that point is not even 1-critical on the original curve—indeed, the point $t = 1$ also corresponding to the origin has $g'(1) \neq 0$.

Geometrically, optimizing $g(t)$ over $\mathbb{R}$ instead of $f(x, y)$ over the curve directly is equivalent to lifting the 2D curve $y^2 = x^2 + x^3$ which intersects itself to the 3D curve $t \mapsto (t^2 - 1, t(t^2 - 1), t)$: see Fig. 1.3 which is colored by the value of $g(t)$. The function $f(x, y)$ has a local minimum at the origin along one of the two intersecting branches, but not on the other. Hence the origin is not a local minimum for f on the curve, but because the two branches separate in the 3D lift one of the preimages of the origin is a local minimum on the lift.

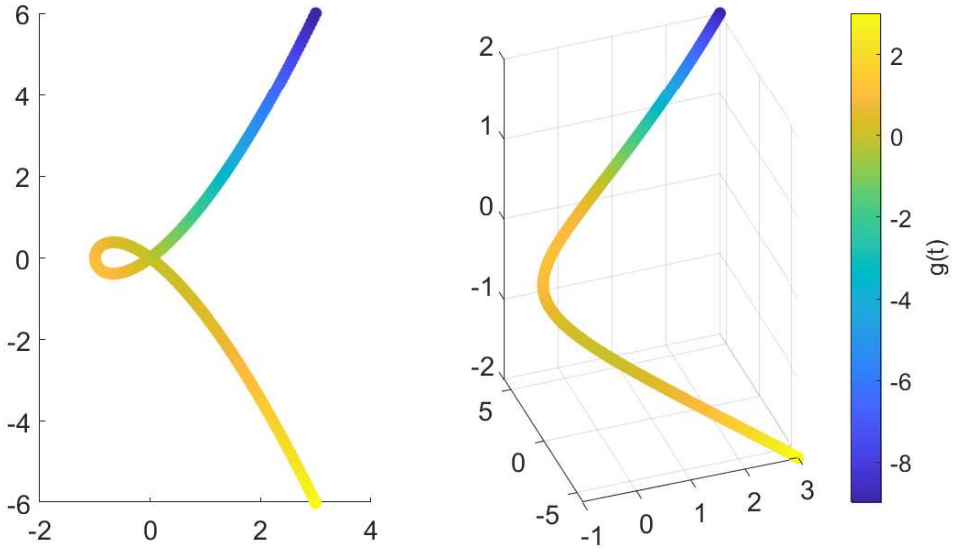


Figure 1.3: Nodal cubic in 2D and its 3D lift, colored by the value of the function $f(x, y) = -x - y$.

Thus, we see that we can lift a singular variety to a nonsingular variety in a higher dimensional space over which it should be easier to optimize. However, a point may be a local minimum on the lift while corresponding to a saddle on the original variety. We shall see that the same phenomenon is observed for lifts of the low-rank matrices.

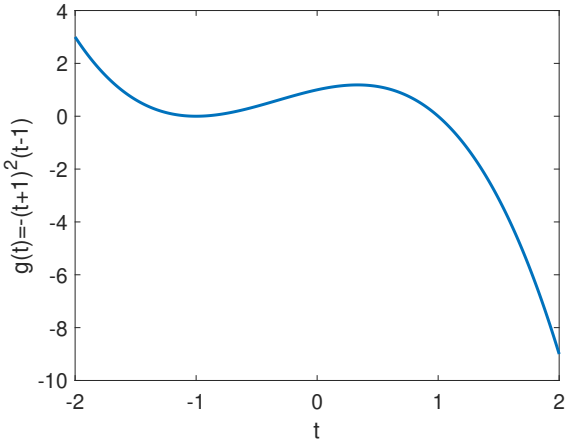


Figure 1.4: Plot of $g(t) = f(t^2 - 1, t(t^2 - 1))$ vs. t.

Chapter 2

Optimizing on lifted spaces

As we have seen in the last chapter, optimizing directly over the variety of bounded-rank matrices is challenging, and standard algorithms may fail in general. In this chapter, we consider parameterizing the variety in different ways and optimizing over the parameter spaces, which we call lifts. These we will take to be smooth manifolds, so standard convergence theorems then apply. However, we must verify that the limit points to which we converge satisfy desirable properties on the original variety. We begin by studying the correspondence between local minima and critical points on the lifts and on the original variety in general. We then restrict ourselves to studying three lifts for the bounded-rank matrix variety in detail.

2.1 General results on critical points and local minima on lifts

2.1.1 Local minima on lifted spaces

Suppose $\varphi\colon \overline{\mathcal{M}} \to \mathcal{M}$ is a continuous surjective map between topological spaces, so a continuous cost function $f\colon \mathcal{M} \to \mathbb{R}$ induces a continuous cost $\bar{f} = f \circ \varphi\colon \overline{\mathcal{M}} \to \mathbb{R}$. We say φ satisfies the *Matching Local Minima Property (MLMP) at* $\bar{x} \in \overline{\mathcal{M}}$ if for any continuous cost function f, the point $x = \varphi(\bar{x}) \in \mathcal{M}$ is a local minimum for f if and only if $\bar{x}$ is a local minimum for $\bar{f}$. If φ satisfies MLMP at all $\bar{x} \in \overline{\mathcal{M}}$, then we simply say φ satisfies MLMP. We now ask under what conditions does φ satisfy this property.

First, continuity and openness of φ give:

Proposition 2.1.

- If φ is continuous and $x \in \mathcal{M}$ is a local minimum for f, then any $\bar{x} \in \varphi^{-1}(x)$ is a local minimum for $\bar{f}$.
- If φ is open and some $\bar{x} \in \varphi^{-1}(x)$ is a local minimum for $\bar{f}$, then x is a local minimum for f.

Proof. Suppose φ is continuous and $x \in \mathcal{M}$ is a local minimum for f. Then there exists an open neighborhood $U \subset \mathcal{M}$ such that $f(y) \geq f(x)$ for all $y \in U$. Since φ is continuous, $\varphi^{-1}(U) \subset \overline{\mathcal{M}}$ is an open neighborhood of any $\bar{x} \in \varphi^{-1}(x)$ in which we have $\bar{f}(\bar{y}) \geq \bar{f}(\bar{x})$ for all $\bar{y} \in \varphi^{-1}(U)$.

Suppose φ is open and $\bar{x} \in \varphi^{-1}(x)$ is a local minimum for $\bar{f}$. Then there exists an open neighborhood $\bar{U} \subset \overline{\mathcal{M}}$ such that $f(\varphi(\bar{y})) = \bar{f}(\bar{y}) \geq \bar{f}(\bar{x}) = f(x)$ for all $\bar{y} \in \bar{U}$. Since φ is open, $\varphi(\bar{U}) \subset \mathcal{M}$ is an open neighborhood of x in which x is a local minimum. □

Corollary 2.2. *If φ is continuous and open, then it satisfies MLMP. In particular, quotient maps $\pi\colon \overline{\mathcal{M}} \to \mathcal{M} = \overline{\mathcal{M}}/G$ satisfy MLMP.*

Proof. The first statement is immediate from Prop. 2.1. The second statement follows because π is continuous by definition of the quotient topology and open by [13, Lemma 21.1]. □

Suppose $\mathcal{M}$ is embedded in a metric space with metric d. We say φ satisfies the *Approximate Subsequence Lifting Property (ASLP) at* $\bar{x} \in \overline{\mathcal{M}}$ if for any sequence $(x_n) \subset \mathcal{M}$ such that $x_n \to x = \varphi(\bar{x})$ and any $(\epsilon_n) \subset \mathbb{R}_{>0}$ such that $\epsilon_n \to 0$, there exist a subsequence $n_1 < n_2 < \dots$ and points $(\bar{x}_{n_i}) \subset \overline{\mathcal{M}}$ satisfying $\bar{x}_{n_i} \to \bar{x}$ and $d(\varphi(\bar{x}_{n_i}), x_{n_i}) \leq \epsilon_{n_i}$ for all i. If φ satisfies ASLP at all $\bar{x} \in \overline{\mathcal{M}}$, then we simply say that φ satisfies ASLP. We then have the following characterization:

Theorem 2.3 (due to Joe Kileel)**.** *Suppose φ is continuous and let $\bar{x} \in \overline{\mathcal{M}}$. Then φ satisfies MLMP at $\bar{x}$ if and only if it satisfies ASLP at $\bar{x}$.*

Proof. Suppose φ satisfies ASLP at $\bar{x} \in \overline{\mathcal{M}}$ and let $f\colon \mathcal{M} \to \mathbb{R}$ be an arbitrary continuous cost function. If x is not a local minimum for f, then there is a sequence $(x_n) \subset \mathcal{M}$ such that $x_n \to x$ and $f(x_n) < f(x)$. Moreover, by continuity of f there is a sequence $(\epsilon_n) \subset \mathbb{R}_{>0}$ such that $\epsilon_n \to 0$ and $f(y_n) < f(x)$ for all $y_n \in \mathcal{M}$ satisfying $d(y_n, x_n) \leq \epsilon_n$. We can now use ASLP to find a sequence $(\bar{x}_{n_i}) \subset \overline{\mathcal{M}}$ such that $\bar{x}_{n_i} \to \bar{x}$ and $d(\varphi(\bar{x}_{n_i}), x_{n_i}) \leq \epsilon_{n_i}$. The latter property implies $\bar{f}(\bar{x}_{n_i}) = f(\varphi(\bar{x}_{n_i})) < f(x) = \bar{f}(\bar{x})$. Thus, $\bar{x}$ is not a local minimum for $\bar{f}$. Together with the continuity of φ and Prop. 2.1, this shows φ satisfies MLMP at $\bar{x}$.

Suppose φ does not satisfy ASLP at some $\bar{x} \in \overline{\mathcal{M}}$. Then there is a sequence $(x_n) \subset \mathcal{M}$ such that $x_n \to x$ and $(\epsilon_n) \subset \mathbb{R}_{>0}$ such that $\epsilon_n \to 0$, but no subsequence can be approximately lifted. Denote $B(x, \epsilon) = \{x' \in \mathcal{M} : d(x, x') \leq \epsilon\}$. Note that $x \notin B(x_n, \epsilon_n)$ for all large n, otherwise the constant sequence $\bar{x}_n = \bar{x}$ gives an approximate lift. Since $\epsilon_n \to 0$, after passing to a subsequence we may assume that the balls $B(x_n, \epsilon_n)$ are pairwise disjoint and none contain x.

Define $f(x') = -\sqrt{\epsilon_n^2 - d(x_n, x')^2}$ if $x' \in B(x_n, \epsilon_n)$ for some n and $f(x') = 0$ otherwise. This is well-defined because the balls $B(x_n, \epsilon_n)$ are disjoint, and is easily seen to be continuous. Note that x is not a local minimum of f since $x_n \to x$ and $f(x_n) = -\epsilon_n < 0 = f(x)$. However, $\bar{x}$ is a local minimum for $\bar{f} = f \circ \varphi$:

if there is a sequence $(\bar{x}_i)$ such that $\bar{x}_i \to \bar{x}$ and $\bar{f}(\bar{x}_i) < \bar{f}(\bar{x}) = 0$ then $\varphi(\bar{x}_i) \in B(x_{n_i}, \epsilon_{n_i})$ for an infinite subsequence n_i, so $\bar{x}_i$ is an approximate lift. Thus, φ does not satisfy MLMP at $\bar{x}$ □

As we show in Sec. 2.1.1, none of the lifts that we consider for the bounded-rank matrices satisfy MLMP. Nevertheless, we use Thm. 2.3 to find a subset of points of one of the lifts at which MLMP is satisfied (see Prop. 2.34). We then regularize our cost function on that lift to ensure that MLMP is satisfied at any of the 1-critical points of the regularized cost.

2.1.2 Obstruction to lifts preserving 1-critical points

Suppose $\mathcal{M} \subset \mathbb{R}^n$ is a singular variety. We want to find a smooth lift $\varphi\colon \overline{\mathcal{M}} \to \mathcal{M}$ such that for any smooth cost function $f\colon \mathcal{M} \to \mathbb{R}$, a point $x \in \mathcal{M}$ is 1-critical for f if and only if there exists $\bar{x} \in \varphi^{-1}(x)$ which is 1-critical for $\bar{f} = f \circ \varphi$ on $\overline{\mathcal{M}}$. Here by 'smooth lift' we mean that $\overline{\mathcal{M}}$ is a smooth manifold and φ is surjective and smooth as a map of smooth manifolds $\overline{\mathcal{M}} \to \mathbb{R}^n$, and by smooth cost $f\colon \mathcal{M} \to \mathbb{R}$ we mean that f extends to a smooth function on a Euclidean neighborhood $\mathcal{U} \supseteq \mathcal{M}$. We prove that such lifts do not exist for a large class of varieties.

In general, suppose $\overline{\mathcal{M}} \subset \mathbb{R}^m, \mathcal{M} \subset \mathbb{R}^n$ are subsets of Euclidean space and $\varphi\colon \overline{\mathcal{U}} \to \mathcal{U}$ is a surjective differentiable map defined on Euclidean neighborhoods $\overline{\mathcal{U}} \supseteq \overline{\mathcal{M}}$ and $\mathcal{U} \supseteq \mathcal{M}$. Let $\varphi'(\bar{x})$ be the Jacobian of φ at $\bar{x} \in \overline{\mathcal{M}}$. Recall the definition of the tangent cone in Eq. (1.17), which applies to any subset of Euclidean space and not just varieties. The map φ induces a map between the tangent cones:

Lemma 2.4. *$\varphi'(\bar{x}) T_{\bar{x}}\overline{\mathcal{M}} \subseteq T_x\mathcal{M}$ for all $\bar{x} \in \overline{\mathcal{M}}$ and $x = \varphi(\bar{x})$.*

Proof. Suppose $\bar{s} \in T_{\bar{x}}\overline{\mathcal{M}}$, and let $\bar{x}_i, \alpha_i$ satisfy $\alpha_i(\bar{x}_i - \bar{x}) \to \bar{s}$. By differentiability of φ, we have

$$\begin{aligned} &\varphi(\bar{x}_i) = \varphi(\bar{x}) + \varphi'(\bar{x}) \cdot (\bar{x}_i - \bar{x}) + o(||\bar{x}_i - \bar{x}||) \\ &\implies \alpha_i(\varphi(\bar{x}_i) - \varphi(\bar{x})) = \varphi'(\bar{x})\alpha_i(\bar{x}_i - \bar{x}) + o\Big(||\alpha_i(\bar{x}_i - \bar{x})||\Big) \xrightarrow{n\to\infty} \varphi'(\bar{x})\bar{s}, \end{aligned} \tag{2.1}$$

so $\varphi'(\bar{x})\bar{s} \in T_x\mathcal{M}$. □

Recall that a point $x \in \mathcal{M}$ is 1-critical for a differentiable $f\colon \mathcal{U} \to \mathbb{R}$ defined on a neighborhood of $\mathcal{M}$ iff $-\nabla f(x) \in [T_x\mathcal{M}]^\circ$, where K° is the polar cone of a cone K (defined in Eq. (1.18)). Therefore, a point $\bar{x} \in \overline{\mathcal{M}}$ is 1-critical for $\bar{f} = f \circ \varphi$ iff $-\nabla \bar{f}(\bar{x}) = \varphi'(\bar{x})^T(-\nabla f(x)) \in [T_{\bar{x}}\overline{\mathcal{M}}]^\circ$, which happens iff $\langle \varphi'(\bar{x})^T(-\nabla f(x)), \bar{s}\rangle = \langle -\nabla f(x), \varphi'(\bar{x})\bar{s}\rangle \leq 0$ for all $\bar{s} \in T_{\bar{x}}\overline{\mathcal{M}}$. Thus, $\bar{x}$ is 1-critical for $\bar{f}$ iff $-\nabla f(x) \in [\varphi'(\bar{x}) T_{\bar{x}}\overline{\mathcal{M}}]^\circ$.

Corollary 2.5. *If $x \in \mathcal{M}$ is 1-critical for $f\colon \mathcal{U} \to \mathbb{R}$ on $\mathcal{M}$, then any $\bar{x} \in \varphi^{-1}(x)$ is 1-critical for $\bar{f} = f \circ \varphi$ on $\overline{\mathcal{M}}$.*

Proof. By Lemma 2.4, we have $[\varphi'(\bar{x})T_{\bar{x}}\overline{\mathcal{M}}]^\circ \supseteq [T_x\mathcal{M}]^\circ$. Hence if x is 1-critical for f, then $-\nabla f(x) \in [T_x\mathcal{M}]^\circ$ so also $-\nabla f(x) \in [\varphi'(\bar{x})T_{\bar{x}}\overline{\mathcal{M}}]^\circ$ and hence $\bar{x}$ is 1-critical for $\bar{f}$. □

We say the lift φ satisfies the *Matching 1-Critical Points Property (M1CPP)* if for any differentiable $f\colon \mathcal{U} \to \mathbb{R}$, a point $\bar{x} \in \varphi^{-1}(x)$ is 1-critical for $\bar{f} = f \circ \varphi$ if and only if x is 1-critical for f.

Theorem 2.6. *$\varphi\colon \overline{\mathcal{M}} \to \mathcal{M}$ satisfies M1CPP if and only if $[\varphi'(\bar{x})T_{\bar{x}}\overline{\mathcal{M}}]^\circ = [T_x\mathcal{M}]^\circ$ for all $x \in \mathcal{M}$ and $\bar{x} \in \varphi^{-1}(x)$.*

Proof. If $[\varphi'(\bar{x})T_{\bar{x}}\overline{\mathcal{M}}]^\circ = [T_x\mathcal{M}]^\circ$ then the two 1-criticality conditions derived above are equivalent and φ satisfies M1CPP.

Conversely, suppose $[\varphi'(\bar{x})T_{\bar{x}}\overline{\mathcal{M}}]^\circ \supsetneq [T_x\mathcal{M}]^\circ$ for some $x \in \mathcal{M}$ and $\bar{x} \in \varphi^{-1}(x)$. Let $s \in [\varphi'(\bar{x})T_{\bar{x}}\overline{\mathcal{M}}]^\circ \setminus [T_x\mathcal{M}]^\circ$ and let $f(x) = \langle x, -s\rangle$ so $-\nabla f(x) = s$. Since $s \notin [T_x\mathcal{M}]^\circ$, the point x is not 1-critial for f on $\mathcal{M}$. Since $s \in [\varphi'(\bar{x})T_{\bar{x}}\mathcal{M}]^\circ$, the point $\bar{x}$ is 1-critical for $\bar{f}$ on $\overline{\mathcal{M}}$. Therefore, φ does not satisfy M1CPP. □

Corollary 2.7. *Suppose $\overline{\mathcal{M}}$ is a smooth manifold, while $\mathcal{M}$ is a singular variety such that the tangent cone to one of its singular points is not a linear space. Then φ does not satisfy M1CPP.*

Proof. Let $x \in \mathcal{M}$ be such a singular point and let $\bar{x} \in \varphi^{-1}(x)$ be arbitrary. Since $\overline{\mathcal{M}}$ is a smooth manifold, $T_{\bar{x}}\overline{\mathcal{M}}$ is a linear space. Since $\varphi'(\bar{x})$ is a linear map, $\varphi'(\bar{x})T_{\bar{x}}\overline{\mathcal{M}} \subsetneq T_x\mathcal{M}$ is a linear subspace too. Since $T_x\mathcal{M}$ is not a linear space, the subspace $\varphi'(\bar{x})T_{\bar{x}}\overline{\mathcal{M}}$ is a proper subspace of the embedding Euclidean space, so $[\varphi'(x)T_{\bar{x}}\overline{\mathcal{M}}]^\circ = [\varphi'(x)T_{\bar{x}}\overline{\mathcal{M}}]^\perp$ is a nontrivial linear subspace. Pick $s \in T_x\mathcal{M} \setminus \varphi'(\bar{x})T_{\bar{x}}\overline{\mathcal{M}}$ and note that there exists $s' \in [\varphi'(\bar{x})T_{\bar{x}}\overline{\mathcal{M}}]^\perp$ such that $\langle s', s\rangle > 0$. This implies $s' \in [\varphi'(\bar{x})T_{\bar{x}}\overline{\mathcal{M}}]^\perp \setminus [T_x\mathcal{M}]^\circ$, showing that φ does not satisfy M1CPP. □

In particular, the variety of bounded-rank matrices $\mathcal{M}_{\leq k}$, the nodal cubic $y^2 = x^2 + x^3$, and the cuspidal cubic $y^2 = x^3$ all satisfy the above hypotheses. Therefore, none of them admit smooth lifts that preserve 1-critical points. Nevertheless, as we shall see in Sec. 2.3, we can construct lifts for $\mathcal{M}_{\leq k}$ for which 2-critical points on the lift map to 1-critical points on the variety.

2.2 Lifts for bounded-rank matrices

We study three lifts for $\mathcal{M}_{\leq k}^{m\times n}$. First, since a matrix X has rank $\leq k$ iff it can be written as $X = LR^T$ for $L \in \mathbb{R}^{m\times k}$ and $R \in \mathbb{R}^{n\times k}$, instead of solving $\min_{X\in\mathcal{M}_{\leq k}} f(X)$ we consider solving

$$\min_{(L,R)\in\mathcal{M}_k^{(L,R)}} f(LR^T), \quad \mathcal{M}_k^{(L,R)} = \mathbb{R}^{m\times k} \times \mathbb{R}^{n\times k}. \tag{2.2}$$

The lift $\varphi\colon \mathcal{M}_k^{(L,R)} \to \mathcal{M}_{\leq k}$ mapping $\varphi(L,R) = LR^T$ is the first lift we consider. We refer to it as the (L,R) factorization.

Unfortunately, the dimension of the Euclidean space above is $(m+n)k > (m+n-k)k = \dim \mathcal{M}_{\leq k}^{m\times n}$. This is therefore an over-parametrization of the original problem. There is substantial ambiguity in the factorization $X = LR^T$. For example, (L,R) and (LJ, RJ^{-T}) are factorizations of the same matrix for any $J \in \mathrm{GL}(k)$, where $\mathrm{GL}(k) = \{J \in \mathbb{R}^{k\times k} : \det(J) \neq 0\}$ is the general linear group. If $\mathrm{rank}(LR^T) = k$, this is in fact the only ambiguity. To reduce the ambiguity, we begin by restricting the first factor to have orthonormal columns and optimize

$$\min_{(U,W)\in\overline{\mathcal{M}}_k^{(U,W)}} f(UW^T), \quad \overline{\mathcal{M}}_k^{(U,W)} = \mathrm{St}(m,k)\times\mathbb{R}^{n\times k}, \tag{2.3}$$

where $\mathrm{St}(m,k) = \{U \in \mathbb{R}^{m\times k} : U^TU = I_k\}$ is the Stiefel manifold, which is smooth [3, Sec. 7.3]. Some ambiguity is still present however, because (U,W), (UQ,WQ) factorize the same matrix for any $Q \in O(k)$ where $O(k) = \mathrm{St}(k,k)$ is the $k\times k$ orthogonal group. Since the action of $O(k)$ on $\overline{\mathcal{M}}_k^{(U,W)}$ is smooth (it acts by matrix multiplication), free (because $U = UQ$ implies, after multiplying on the left by U^T, that $Q = I$), and proper (because $O(k)$ is a compact Lie group), we conclude that the quotient $\mathcal{M}_k^{(U,W)} = \overline{\mathcal{M}}_k^{(U,W)}/O(k)$ is a smooth manifold. The second lift that we consider is then

$$\min_{[(U,W)]\in\mathcal{M}_k^{(U,W)}} f(UW^T), \quad \mathcal{M}_k^{(U,W)} = \overline{\mathcal{M}}_k^{(U,W)}/O(k), \tag{2.4}$$

with the map $\varphi\colon \mathcal{M}_k^{(U,W)} \to \mathcal{M}_{\leq k}$ defined by $\varphi([(U,W)]) = UW^T$. We refer to it as the (U,W) factorization. Note that

$$\dim \mathcal{M}_k^{(U,W)} = \dim \overline{\mathcal{M}}_k^{(U,W)} - \dim O(k) = \dim \mathrm{St}(m,k) + \dim \mathbb{R}^{n\times k} - \dim O(k) = (m+n-k)k = \dim \mathcal{M}_{\leq k}^{m\times n}, \tag{2.5}$$

so there are no excess dimensions for this lift.

Finally, we consider the desingularization studied in [10]. There, the authors observe that a matrix X has rank $\leq k$ iff there is a $(n-k)$-dimensional subspace on which X is zero. We denote by $\mathrm{Gr}(n,n-k)$ the Grassmannian of $(n-k)$-dimensional subspaces of $\mathbb{R}^n$, and view it as a quotient $\mathrm{Gr}(n,n-k) = \mathrm{St}(n,n-k)/O(n-k)$ where a matrix $Y \in \mathrm{St}(n,n-k)$ corresponds to the $(n-k)$-dimensional subspace $[Y]$ spanned by its columns. The condition $X|_{[Y]} = 0$ is equivalent to $XY = 0$, so we can optimize

$$\min_{(X,[Y])\in\mathcal{M}_k^{(X,Y)}} f(X), \quad \mathcal{M}_k^{(X,Y)} = \{(X,[Y]) \in \mathbb{R}^{m\times n}\times \mathrm{Gr}(n,n-k) : XY = 0\}. \tag{2.6}$$

Together with the map $\varphi : \mathcal{M}_k^{(X,Y)} \to \mathcal{M}_{\leq k}$ sending $\varphi(X, [Y]) = X$, this is our third lift. We refer to it as the (X, Y) desingularization. Viewing it as a quotient of the total space $\overline{\mathcal{M}}_k^{(X,Y)} = \mathbb{R}^{m\times n} \times \mathrm{St}(n, n-k)$ by the group $\{\mathrm{id}\} \times O(n-k)$, note that the action of this group on $\overline{\mathcal{M}}_k^{(X,Y)}$ is smooth, free and proper. We conclude that the quotient above is a smooth manifold of dimension

$$\dim \mathcal{M}_k^{(X,Y)} = \dim \mathbb{R}^{m\times n} + \dim \mathrm{St}(n, n-k) - \dim O(n-k) = \dim \mathcal{M}_{\leq k}, \tag{2.7}$$

so once again there is no overparamterization (see [10, Thm. 2] for an alternative proof). The lift $\mathcal{M}_k^{(X,Y)}$ is in fact a resolution of singularities—the lift map $(X, [Y]) \mapsto X$ defined on $\mathcal{M}_k^{(X,Y)}$ is a morphism of varieties and an isomorphism when restricted to points with $\mathrm{rank}(X) = k$ since then $[Y] = \ker X$ is uniquely determined. If $K \subset \mathcal{M}_{\leq k}$ is a compact subset, then its inverse image is $K \times \mathrm{Gr}(n, n-k)$ which is compact as a product of compact sets, so the map is proper.

2.3 Comparison of optimality conditions

In this section, we study the first and second order necessary optimality conditions for the three lifts introduced above, and compare them to the corresponding optimality conditions on the original variety.

Our findings can be summarized by the chain of implications:

$$\text{2-critical on lifts} \implies \text{1-critical on } \mathcal{M}_{\leq k} \implies \text{1-critical on lifts} \implies \text{1-critical on its own stratum}$$

In addition, the optimality conditions on the lifts and on the variety coincide for rank k points.

The (L, R) factorization.

We start by considering the factored problem $\min_{(L,R)\in\mathcal{M}_k^{(L,R)}} f(LR^T)$. Denote by $g(L, R) = f(LR^T)$ the induced cost function. We consider four classes of pairs (L, R), namely those such that $X = LR^T$ is 1-critical on its own stratum of $\mathcal{M}_{\leq k}$ (i.e. 1-critical on the smooth manifold $\mathcal{M}_{\mathrm{rank}(X)}$), those that are 1-critical on the entire variety $\mathcal{M}_{\leq k}$ (i.e. either $\mathrm{rank}(X) = k$ and X is 1-critical on $\mathcal{M}_k$ or $\mathrm{rank}(X) < k$ and $\nabla f(X) = 0$), and those that are 1- and 2-critical for g on $\mathcal{M}_k^{(L,R)}$.

Denote by $P_{\mathrm{col}(X)} \colon \mathbb{R}^m \to \mathrm{col}(X)$ the orthogonal projection onto the column space of X, and $P_{\mathrm{row}(X)} \colon \mathbb{R}^n \to \mathrm{row}(X)$ the orthogonal projection onto $\mathrm{row}(X) = \mathrm{col}(X^T)$. The Riemannian gradient of f at $X \in \mathcal{M}_{\mathrm{rank}(X)}$ (i.e. to the stratum of $\mathcal{M}_{\leq k}$ containing X) is given by [25, Eq. (2.5)]

$$\mathrm{grad}\ f(X) = \Pi_X \nabla f(X) = P_{\mathrm{col}(X)} \nabla f(X) + \nabla f(X) P_{\mathrm{row}(X)} - P_{\mathrm{col}(X)} \nabla f(X) P_{\mathrm{row}(X)}, \tag{2.8}$$

where $\nabla f(X)$ is the Euclidean gradient of f (viewed as a function on $\mathbb{R}^{m\times n}$). The gradient of $g(L,R)$ is given by the chain rule in terms of the gradient of f as

$$\nabla g(L,R) = \Big(\nabla f(LR^T)R, \big[\nabla f(LR^T)\big]^T L\Big). \tag{2.9}$$

Proposition 2.8. If $\nabla g(L,R) = 0$, then grad $f(LR^T) = 0$.

Proof. Set $X = LR^T$. Because $\mathrm{col}(X) \subseteq \mathrm{col}(L)$ and $L^T\nabla f(LR^T) = 0$, we have $P_{\mathrm{col}(X)}\nabla f(X) = 0$, and similarly, because $\nabla f(LR^T)R = 0$ and $\mathrm{row}(X) \subseteq \mathrm{col}(R)$, we have $\nabla f(X)P_{\mathrm{row}(X)} = 0$. Thus, grad $f(X) = 0$ so X is a critical point of f on $\mathcal{M}_{\mathrm{rank}(X)}$. □

Thus, if (L,R) is 1-critical for g then LR^T is 1-critical on its own stratum for f. If $\mathrm{rank}(LR^T) = k$, the converse also holds:

Proposition 2.9. If grad $f(X) = 0$ and $\mathrm{rank}(X) = k$, then $\nabla g(L,R) = 0$ for any matrices $(L,R) \in \mathbb{R}^{m\times k} \times \mathbb{R}^{n\times k}$ satisfying $X = LR^T$.

Proof. Any such L, R must have full column rank so $\mathrm{col}(X) = \mathrm{col}(L)$ and $\mathrm{row}(X) = \mathrm{col}(R)$. We then have

$$0 = P_{\mathrm{col}(X)}\mathrm{grad}\ f(X) = P_{\mathrm{col}(X)}\nabla f(X) = L(L^TL)^{-1}L^T\nabla f(X) \implies L^T\nabla f(X) = 0, \tag{2.10}$$

where the last implication is obtained by multiplying by L^T. Similarly,

$$0 = \mathrm{grad}\ f(X)P_{\mathrm{row}(X)} = \nabla f(X)P_{\mathrm{row}(X)} = \nabla f(X)R(R^TR)^{-1}R^T \implies \nabla f(X)R = 0. \tag{2.11}$$

Therefore, $\nabla g(L,R) = 0$. □

We conclude that

Corollary 2.10. *Suppose rank$(X) = k$. If X is 1-critical for f, then any pair (L,R) satisfying $X = LR^T$ is 1-critical for g. Conversely, if there exists a pair (L,R) satisfying $X = LR^T$ that is 1-critical for g, then X is 1-critical for f on $\mathcal{M}_{\le k}$.*

Proof. Follows from Props. 2.8 and 2.9. □

Corollary 2.11. *If $X \in \mathcal{M}_{\le k}$ is 1-critical for f, then any pair $(L,R) \in \mathcal{M}_k^{(L,R)}$ satisfying $X = LR^T$ is 1-critical for g.*

Proof. Any point with $\mathrm{rank}(X) < k$ that is a critical point for f satisfies $\nabla f(X) = 0$ and hence $\nabla g(L,R) = 0$ for all L, R such that $X = LR^T$. □

The above result is true for lifts in general by Cor. 2.5.

There may be points on $\mathcal{M}_{\leq k}$ that are 1-critical with rank $< k$ that are 1-critical on their own startum but whose factorization is not 1-critical for g:

Proposition 2.12. It is possible that grad $f(X) = 0$ for $X = LR^T$ but $\nabla g(L, R) \neq 0$ when $\text{rank}(X) < k$.

Proof. Take $f(X) = \frac{1}{2}||X - X^*||_F^2$, where $\text{rank}(X^*) = k$. Since $\nabla f(X) = X - X^*$, we have in particular that $\nabla f(P_{k-1}X^*) = P_{k-1}(X^*) - X^* \perp T_{P_{k-1}(X^*)}\mathcal{M}_{k-1}$ because P_{k-1} is metric projection, hence grad $f(P_{k-1}X^*) = 0$. If $X^* = U\Sigma V^T$ is a thin SVD of X^*, then $\nabla f(P_{k-1}X^*) = -\sigma_n u_n v_n^T$. On the other hand, if we let $L = U$ and $R = [V_{:,1:k-1}, 0]\Sigma$, then $LR^T = P_{k-1}X^*$ but $\nabla f(P_{k-1}X^*)^T L = -\sigma_n v_n (u_n^T U) = [0, \ldots, 0, -\sigma_n v_n] \neq 0$ hence $\nabla g(L, R) \neq 0$. Note that $L \in \text{St}(m, k)$, a fact that will be used in the analysis of the (U, W) factorization. □

There may also be 1-critical points (L, R) for g such that $X = LR^T$ is not 1-critical for f on the entire variety:

Proposition 2.13. There exists a cost function f and point $X = LR^T$ with $\text{rank} < k$ such that (L, R) is 1-critical for g but X is not 1-critical for f.

Proof. Set $f(X) = \frac{1}{2}||X + X^*||_F^2$ for some nonzero $X^* \in \mathcal{M}_k$. Then $\nabla f(0) = X^* \neq 0$. If $m \geq 2k$ we can choose L such that $L^T X^* = 0$ (i.e. all the columns of L are orthogonal to all the columns of X^*), in which case $\nabla g(L, 0) = 0$. Note that we can choose $L \in \text{St}(m, k)$, a fact that will be used in the analysis of the (U, W) factorization. □

The existence of such f follows in general by Cor. 2.7.

Since 1-criticality for g does not suffice to get 1-criticality for f on $\mathcal{M}_{\leq k}$, we now consider 2-critical points for g. The quadratic form defined by the Hessian for g is:

$$\begin{aligned}\langle \nabla^2 g(L, R)[\dot{L}, \dot{R}], (\dot{L}, \dot{R})\rangle = \langle \nabla^2 f(LR^T)[\dot{L}R^T + L\dot{R}^T], \dot{L}R^T + L\dot{R}^T\rangle \\ + 2\langle \nabla f(LR^T), \dot{L}\dot{R}^T\rangle,\end{aligned} \tag{2.12}$$

where $(\dot{L}, \dot{R}) \in \mathbb{R}^{m\times k} \times \mathbb{R}^{n\times k}$ are arbitrary. Furthermore:

Proposition 2.14. The quadratic form defined by the Riemannian Hessian of f at X (defined with respect to the stratum of X) is:

$$\langle \text{Hess } f(X)[Z], Z\rangle = \langle \nabla^2 f(X)[Z], Z\rangle + 2\langle P^{\perp}_{\text{col}(X)} \nabla f(X) P^{\perp}_{\text{row}(X)}, ZX^{\dagger}Z\rangle, \tag{2.13}$$

where $Z \in T_X \mathcal{M}_{\text{rank}(X)}$.

Proof. The Riemannian Hessian of f at X is [3, Sec. (7.5)]

$$\begin{aligned}\text{Hess } f(X)[Z] =&\Pi_X \nabla^2 f(X)[Z] + U\Sigma^{-1}U_p^T \nabla f(X) P^\perp_{\text{row}(X)} \\ &+ P^\perp_{\text{col}(X)} \nabla f(X) V_p \Sigma^{-1} V^T,\end{aligned} \tag{2.14}$$

where any tangent vector $Z \in T_X \mathcal{M}_{\text{rank}(X)}$ can be written as $Z = UMV^T + U_p V^T + UV_p^T$ where $X = U\Sigma V^T$ is the thin SVD of X, M is arbitrary, and $U_p^T U = V_p^T V = 0$. We can re-write this in a form independent of the SVD of X as follows. Observe that $X^\dagger = V\Sigma^{-1}U^T$ is the pseudoinverse of X and that $U_p = P^\perp_{\text{col}(X)} ZV$, hence

$$U\Sigma^{-1}U_p^T = (U\Sigma^{-1}V^T)Z^T P^\perp_{\text{col}(X)} = (X^\dagger)^T Z^T P^\perp_{\text{col}(X)}. \tag{2.15}$$

Similarly, $V_p = P^\perp_{\text{row}(X)} Z^T U$ so

$$V_p \Sigma^{-1} V^T = P^\perp_{\text{row}(X)} Z^T (U\Sigma^{-1}V^T) = P^\perp_{\text{row}(X)} Z^T (X^\dagger)^T. \tag{2.16}$$

Combining these results, we have

$$\begin{aligned}\text{Hess } f(X)[Z] =&\Pi_X \nabla^2 f(X)[Z] + (X^\dagger)^T Z^T P^\perp_{\text{col}(X)} \nabla f(X) P^\perp_{\text{row}(X)} \\ &+ P^\perp_{\text{col}(X)} \nabla f(X) P^\perp_{\text{row}(X)} Z^T (X^\dagger)^T.\end{aligned} \tag{2.17}$$

Since $Z \in T_X \mathcal{M}_{\text{rank}(X)}$ which is a subspace, and Π_X is projection to that subspace, we have

$$\langle \Pi_X \nabla^2 f(X)[Z], Z\rangle = \langle \nabla^2 f(X)[Z], Z\rangle. \tag{2.18}$$

Since any matrices with compatible sizes A, B, C satisfy $\langle A^T B, C\rangle = \text{Tr}(C^T A^T B) = \text{Tr}((AC)^T B) = \langle B, AC\rangle$, we have

$$\langle (X^\dagger)^T Z^T P^\perp_{\text{col}(X)} \nabla f(X) P^\perp_{\text{row}(X)}, Z\rangle = \langle P^\perp_{\text{col}(X)} \nabla f(X) P^\perp_{\text{row}(X)}, ZX^\dagger Z\rangle. \tag{2.19}$$

Similarly, since $\langle AB^T, C\rangle = \text{Tr}(C^T AB^T) = \text{Tr}(B^T C^T A) = \text{Tr}((CB)^T A) = \langle A, CB\rangle$, we have

$$\langle P^\perp_{\text{col}(X)} \nabla f(X) P^\perp_{\text{row}(X)} Z^T (X^\dagger)^T, Z\rangle = \langle P^\perp_{\text{col}(X)} \nabla f(X) P^\perp_{\text{row}(X)}, ZX^\dagger Z\rangle. \tag{2.20}$$

Adding the above three terms together, we obtain the claimed result. □

To compare the two quadratic forms, it is useful to observe that:

Proposition 2.15. Any $Z \in T_X \mathcal{M}_{\text{rank}(X)}$ can be written as $Z = \dot{L}R^T + L\dot{R}^T$ for some $(\dot{L}, \dot{R}) \in \mathbb{R}^{m\times k} \times \mathbb{R}^{n\times k}$.

If $\operatorname{rank}(X) = \operatorname{rank}(L) = \operatorname{rank}(R)$ (which must be the case when $\operatorname{rank}(X) = k$), then the converse also holds: $\dot{L}R^T + L\dot{R}^T \in T_X\mathcal{M}_{\operatorname{rank}(X)}$ for any $(\dot{L}, \dot{R}) \in \mathbb{R}^{m\times k} \times \mathbb{R}^{n\times k}$.

Proof. Suppose $\operatorname{rank}(X) = k$, so $\operatorname{rank}(L) = \operatorname{rank}(R) = k$ and $P_{\operatorname{col}(X)} = LL^\dagger$ and $P_{\operatorname{row}(X)} = (R^\dagger)^T R^T$. Then $Z \in T_X\mathcal{M}_k$ iff

$$Z = \Pi_X Z = P_{\operatorname{col}(X)}Z + ZP_{\operatorname{row}(X)} - P_{\operatorname{col}(X)}ZP_{\operatorname{row}(X)} = LL^\dagger Z + Z(R^\dagger)^T R^T - LL^\dagger Z(R^\dagger)^T R^T. \tag{2.21}$$

Defining e.g. $\dot{R} = Z^T(L^\dagger)^T$ and $\dot{L} = Z(R^\dagger)^T - LL^\dagger Z(R^\dagger)^T$ we conclude that if $Z \in T_X\mathcal{M}_k$ then $Z = \dot{L}R^T + L\dot{R}^T$ for some $(\dot{L}, \dot{R})$.

Conversely, if $Z = \dot{L}R^T + L\dot{R}^T$ and $\operatorname{rank}(X) = \operatorname{rank}(L) = \operatorname{rank}(R)$, then $\operatorname{col}(X) = \operatorname{col}(L)$ and $\operatorname{row}(X) = \operatorname{col}(R)$ so

$$\begin{aligned}\Pi_X Z &= P_{\operatorname{col}(X)}(\dot{L}R^T + L\dot{R}^T) + (\dot{L}R^T + L\dot{R}^T)P_{\operatorname{row}(X)} - P_{\operatorname{col}(X)}(\dot{L}R^T + L\dot{R}^T)P_{\operatorname{row}(X)} \\ &= P_{\operatorname{col}(X)}\dot{L}R^T + L\dot{R}^T + \dot{L}R^T + L\dot{R}^T P_{\operatorname{row}(X)} - P_{\operatorname{col}(X)}\dot{L}R^T - L\dot{R}^T P_{\operatorname{row}(X)} \\ &= L\dot{R}^T + \dot{L}R^T = Z,\end{aligned} \tag{2.22}$$

hence $Z \in T_X\mathcal{M}_k$. □

A pair (L, R) is 2-critical for g if $\nabla g(L, R) = 0$ and $\nabla^2 g(L, R) \succeq 0$. A point X is 2-critical for f on its own stratum if grad $f(X) = 0$ and Hess $f(X) \succeq 0$. Finally, by Defn. 7, a point X is 2-critical for f on all of $\mathcal{M}_{\le k}$ if either $\operatorname{rank}(X) = k$ and X is 2-critical on $\mathcal{M}_k$, or $\operatorname{rank}(X) < k$ and we have $\nabla f(X) = 0$ and $\langle \nabla^2 f(X)[Z], Z\rangle \ge 0$ for all $Z \in T_X\mathcal{M}_{\le k}$.

Proposition 2.16. Suppose $X = LR^T$ satisfies $\operatorname{rank}(X) = k$. Then (L, R) is 2-critical for g if and only if X is 2-critical on its own stratum for f.

Proof. By Cor. 2.10 we have $\nabla g(L, R) = 0$ iff grad $f(LR^T) = 0$ when $\operatorname{rank}(LR^T) = k$. Thus, it remains to show that $\nabla^2 g(L, R) \succeq 0$ iff Hess $f(X) \succeq 0$.

Note that $Z \in T_X\mathcal{M}_k$ iff it can be written as $Z = \dot{L}R^T + L\dot{R}^T$ for some $(\dot{L}, \dot{R})$. Since L and R have full column rank, we have $X^\dagger = (R^\dagger)^T L^\dagger = R(R^T R)^{-1}(L^T L)^{-1}L^T$, and a simple computation shows

$$ZX^\dagger Z = \dot{L}\dot{R}^T + \underbrace{[\dot{L}L^\dagger\dot{L}]R^T + L[\dot{R}R^\dagger\dot{R}]^T + L[(R^\dagger\dot{R})^T L^\dagger\dot{L}]R^T}_{\in T_X\mathcal{M}_k}. \tag{2.23}$$

Since $\nabla f(X) \in T_X^\perp\mathcal{M}_k$ and $\Pi_X Z = Z$, we obtain

$$\langle \text{Hess } f(X)[Z], Z\rangle = \langle \Pi_X\nabla^2 f(X)[Z], Z\rangle + 2\langle \nabla f(X), \dot{L}\dot{R}^T\rangle = \langle \nabla^2 g(L, R)[\dot{L}, \dot{R}], (\dot{L}, \dot{R})\rangle, \tag{2.24}$$

which gives the result. □

Proposition 2.17. Suppose $X = LR^T$ is a 2-critical point of g with $\text{rank}(X) < k$. Then X is a 1-critical point of f on $\mathcal{M}_{\leq k}$ and 2-critical on $\mathcal{M}_{\text{rank}(X)}$, but not necessarily 2-critical on all of $\mathcal{M}_{\leq k}$.

Proof. If $\text{rank}(X) = k$, this is shown in the preceding proposition. Suppose $\text{rank}(X) < k$. First, we argue that $\nabla f(X) = 0$ so X is 1-critical for f on $\mathcal{M}_{\leq k}$. The argument is taken from [9, Sec. 2.1.1]. Let u_1 and v_1 be the top left and right singular vectors of $\nabla f(X)$, respectively, so $u_1^T \nabla f(X) v_1 = \sigma_1(\nabla f(X)) = ||\nabla f(X)||$. Since $\text{rank}(LR^T) < k$, we must have either $\text{rank}(L) < k$ or $\text{rank}(R) < k$. Indeed, if $\text{rank}(L) = \text{rank}(R) = k$ then $\dim(R^T\mathbb{R}^n) = k$ so $R^T\mathbb{R}^n = \mathbb{R}^k$, and $\dim L(R^T\mathbb{R}^n) = \dim L(\mathbb{R}^k) = \text{rank}(L) = k$ so $\text{rank}(X) = \dim X\mathbb{R}^n = k$. Suppose $\text{rank}(L) < k$, and choose $w \in \ker(L)$ with $||w||_2 = 1$. Set $\dot{L} = u_1 w^T$ and $\dot{R} = -\alpha v_1 w^T$ where $\alpha \geq 0$ is arbitrary. We then have

$$\dot{L}R^T + L\dot{R}^T = u_1 w^T R^T - \alpha(Lw)v_1^T = u_1 w^T R^T, \tag{2.25}$$

and

$$\langle \nabla f(X), \dot{L}\dot{R}^T \rangle = -\alpha ||\nabla f(X)|| \cdot ||w||_2 = -\alpha ||\nabla f(X)||, \tag{2.26}$$

so the second-order optimality condition on g implies

$$\langle \nabla^2 f(X)[u_1 w^T R^T], u_1 w^T R^T \rangle \geq \alpha ||\nabla f(X)||, \tag{2.27}$$

for all $\alpha \geq 0$. Since the RHS is independent of α while the RHS can be made arbitrarily large by taking $\alpha \to \infty$ if $\nabla f(X) \neq 0$, we conclude that $\nabla f(X) = 0$. Thus, any 2-critical rank-deficient point of g is 1-critical for f on $\mathcal{M}_{\leq k}$.

Since $\nabla f(X) = 0$, the quadratic form defined by the Hessians reduce to

$$\begin{aligned} &\langle \nabla^2 g(L,R)[\dot{L},\dot{R}], (\dot{L},\dot{R}) \rangle = \langle \nabla^2 f(X)[\dot{L}R^T + L\dot{R}^T], \dot{L}R^T + L\dot{R}^T \rangle \\ &\langle \text{Hess } f(X)[Z], Z \rangle = \langle \nabla^2 f(X)[Z], Z \rangle. \end{aligned} \tag{2.28}$$

Since any $Z \in T_X\mathcal{M}_{\text{rank}(X)}$ can be written as $Z = \dot{L}R^T + L\dot{R}^T$ for some $(\dot{L}, \dot{R})$, we conclude that if (L, R) is 2-critical for g (i.e. $\langle \nabla^2 f(X)[\dot{L}R^T + L\dot{R}^T], \dot{L}R^T + L\dot{R}^T \rangle \geq 0$ for all $(\dot{L}, \dot{R})$) then LR^T is 2-critical for f on its own stratum (i.e. $\langle \nabla^2 f(X)[Z], Z \rangle \geq 0$ for all $Z \in T_X\mathcal{M}_{\text{rank}(X)}$).

To see that X is not necessarily 2-critical on all of $\mathcal{M}_{\leq k}$, define

$$\begin{aligned} X^* &= \begin{pmatrix} I_{k-1} & \mathbf{0}_{(k-1)\times(n-k+1)} \\ \mathbf{0}_{(m-k+1)\times(k-1)} & \mathbf{0}_{(m-k+1)\times(n-k+1)} \end{pmatrix} \\ L &= \begin{pmatrix} I_{k-1} & \mathbf{0}_{(k-1)\times 1} \\ \mathbf{0}_{(m-k+1)\times(k-1)} & \mathbf{0}_{(m-k+1)\times 1} \end{pmatrix}, \; R = \begin{pmatrix} I_{k-1} & \mathbf{0}_{(k-1)\times 1} \\ \mathbf{0}_{(n-k+1)\times(k-1)} & \mathbf{0}_{(n-k+1)\times 1} \end{pmatrix}. \end{aligned} \tag{2.29}$$

let $f(X) = \mathrm{vec}(X - X^*)^T Q \mathrm{vec}(X - X^*)$ where $Q = \mathrm{diag}(1, \ldots, 1, -1) \in \mathbb{R}^{mn \times mn}$. In this case, $\nabla f(X) = 2Q\mathrm{vec}(X - X^*)$ so $\nabla f(X^*) = 0$ and X^* is 1-critical for f on $\mathcal{M}_{\leq k}$. Also,

$$T_{X^*}\mathcal{M}_{\leq k} = \left\{ \begin{pmatrix} A & B \\ C & D \end{pmatrix} : A \in \mathbb{R}^{(k-1)\times(k-1)}, \; \mathrm{rank}(D) = 1 \right\}, \tag{2.30}$$

so in particular, $Z = E_{m,n} \in T_{X^*}\mathcal{M}_{\leq k}$ where $E_{m,n}$ is the matrix with all zeros except for a 1 in the (m,n)th entry. In this case $\nabla^2 f(X^*)[Z] = -E_{m,n}$ so $\langle \nabla^2 f(X^*)[Z], Z \rangle = -1$ and hence X^* is not 2-critical for f on $\mathcal{M}_{\leq k}$. However, for any $(\dot{L}, \dot{R})$, if we write

$$\dot{L} = \begin{pmatrix} U_{L,1} & U_{L,2} \\ U_{L,3} & U_{L,4} \end{pmatrix}, \tag{2.31}$$

where $U_{L,1} \in \mathbb{R}^{(k-1)\times(k-1)}$, $U_{L,2} \in \mathbb{R}^{(k-1)\times 1}$, $U_{L,3} \in \mathbb{R}^{(m-k+1)\times(k-1)}$, and $U_{L,4} \in \mathbb{R}^{(m-k+1)\times 1}$, and similarly for $\dot{R}$ (with m replaced by n in the dimensions above), we get

$$\dot{L}R^T + L\dot{R}^T = \begin{pmatrix} U_{L,1} + U_{R,1}^T & U_{R,2}^T \\ U_{L,2} & 0 \end{pmatrix}, \tag{2.32}$$

in which case

$$\langle \nabla^2 f(X^*)[\dot{L}R^T + L\dot{R}^T], \dot{L}R^T + L\dot{R}^T \rangle = ||\dot{L}R^T + L\dot{R}^T||_F^2 \geq 0, \tag{2.33}$$

for all $(\dot{L}, \dot{R})$. Thus, (L, R) is 2-critical for g but not 2-critical for f on all of $\mathcal{M}_{\leq k}$. Note that $L \in \mathrm{St}(m, k)$, a fact that will be used in the analysis of the (U, W) factorization. □

However, note that:

Proposition 2.18. Suppose $f \colon \mathcal{M}_{\leq k} \to \mathbb{R}$ extends to a convex function on $\mathbb{R}^{m\times n}$. Then a point $X = LR^T$ is 2-critical for f on all of $\mathcal{M}_{\leq k}$ iff the pair (L, R) is 2-critical for g.

Proof. If $\operatorname{rank}(X) = k$ then the convexity assumption is unnecessary and this was already shown in Prop. 2.16 above. Suppose $\operatorname{rank}(X) < k$. If X is 2-critical for g, then $\nabla f(X) = 0$ so it is 1-critical for f, and the condition that it is 2-critical for f is trivially satisfied by convexity: we have $\langle \nabla^2 f(X)[Z], Z\rangle \geq 0$ for all $Z \in T_X \mathcal{M}_{\leq k}$ (in fact, for all $Z \in \mathbb{R}^{m\times n}$).

Conversely, suppose X is 2-critical for f. Then $\nabla f(X) = 0$ so $\nabla g(L, R) = 0$ as well, and again the second order condition is trivially satisfied by convexity: $\langle \nabla^2 g(L, R)[\dot{L}, \dot{R}], (\dot{L}, \dot{R})\rangle = \langle \nabla^2 f(X)[\dot{L}R^T + L\dot{R}^T], \dot{L}R^T + L\dot{R}^T\rangle \geq 0$ for all $(\dot{L}, \dot{R})$. □

Thus, we conclude that 2-criticality of (L, R) on the lift $\mathcal{M}_k^{(L,R)}$ gives 1-criticality for $X = LR^T$ on $\mathcal{M}_{\leq k}$ and 2-criticality on its own stratum $\mathcal{M}_{\operatorname{rank}(X)}$. However, it does not give 2-criticality on all of $\mathcal{M}_{\leq k}$ in general, with the exception of convex f.

The (U, W) factorization.

We now consider the second lifted problem $\min_{[(U,W)]\in\mathcal{M}_k^{(U,W)}} f(UW^T)$. First, by Prop. 1.2, a point $[(U, W)]$ is 1- or 2-critical on $\mathcal{M}_k^{(U,W)}$ if and only if (U, W) is respectively 1- or 2-critical on $\overline{\mathcal{M}}_k^{(U,W)} \subset \mathcal{M}_k^{(L,R)}$. We now claim a point (U, W) is 1- or 2-critical for $f(UW^T)$ on $\overline{\mathcal{M}}_k^{(U,W)}$ iff it is 1- or 2-critical, respectively, for $g(L, R) = f(LR^T)$ on all of $\mathcal{M}_k^{(L,R)} = \mathbb{R}^{m\times k} \times \mathbb{R}^{n\times k}$.

Proposition 2.19. $(U, W) \in \overline{\mathcal{M}}_k^{(U,W)} \subset \mathcal{M}_k^{(L,R)}$ is 1-critical for $g(U, W)$ on $\overline{\mathcal{M}}_k^{(U,W)}$ if and only if it is 1-critical for g on all of $\mathcal{M}_k^{(L,R)}$.

Proof. We can write

$$\overline{\mathcal{M}}_k^{(U,W)} = \{(U, W) \in \mathcal{M}_k^{(L,R)} : U^T U = I_k\}, \tag{2.34}$$

so the Lagrangian for the problem $\min_{(U,W)} f(UW^T)$ is

$$\mathcal{L}(U, W, \Lambda) = f(UW^T) + \frac{1}{2}\langle \Lambda, U^T U - I_k\rangle, \tag{2.35}$$

where $\Lambda \in \operatorname{Sym}(k)$ is a matrix of Lagrange multipliers. The first order optimality conditions read (these are equivalent to the Riemannian gradient vanishing by [27])

$$\begin{cases} \nabla f(UW^T)W + U\Lambda = 0, \\ \nabla f(UW^T)^T U = 0, \\ U^T U = I_k. \end{cases} \tag{2.36}$$

Multiplying the first equation by U^T on the left, we obtain $\Lambda = -U^T \nabla f(UW^T)W = 0$ where from the

second equation we have $U^T \nabla f(UW^T) = 0$. Substituting $\Lambda = 0$, we note that the first two equations become precisely $\nabla g(U, W) = 0$. This shows the equivalence of the first order conditions. □

Proposition 2.20. $(U, W) \in \overline{\mathcal{M}}_k^{(U,W)} \subset \mathcal{M}_k^{(L,R)}$ is 2-critical for $g(U, W)$ on $\overline{\mathcal{M}}_k^{(U,W)}$ if and only if it is 2-critical for g on all of $\mathcal{M}_k^{(L,R)}$.

Proof. By [20, Thm. 3.46] (and [27] which shows the equivalence of this condition to the positive-semidefiniteness of the Riemannian Hessian) a point (U, W) is 2-critical for $f(UW^T)$ if

$$\langle \nabla^2_{(U,W)} \mathcal{L}|_{\Lambda=0}[\dot{U}, \dot{W}], (\dot{U}, \dot{W})\rangle = \langle \nabla^2 g(U, W)[\dot{U}, \dot{W}], (\dot{U}, \dot{W})\rangle \geq 0, \tag{2.37}$$

for all $(\dot{U}, \dot{W}) \in T_{(U,W)}\overline{\mathcal{M}}_k^{(U,W)} = T_U \mathrm{St}(m, k) \oplus \mathbb{R}^{n\times k}$. This shows that if (U, W) is 2-critical on all of $\mathcal{M}_k^{(L,R)}$ then it is 2-critical on $\overline{\mathcal{M}}_k^{(U,W)}$.

Suppose (U, W) is 2-critical for g on $\overline{\mathcal{M}}_k^{(U,W)}$. Let $U_\perp \in \mathrm{St}(n, n-k)$ be such that the concatenation $(U,\ U_\perp) \in O(n)$ is orthogonal. Recall from [3, Eqns. (7.18), (7.23)] that

$$\begin{aligned} T_U \mathrm{St}(m, k) &= \{UA + U_\perp B : A \in \mathrm{Skew}(k),\ B \in \mathbb{R}^{(n-k)\times k}\}, \\ T_U^\perp \mathrm{St}(m, k) &= \{US : S \in \mathrm{Sym}(k)\}. \end{aligned} \tag{2.38}$$

For any $(\dot{L}, \dot{R}) \in \mathbb{R}^{m\times k} \times \mathbb{R}^{n\times k} = T_{(U,W)}\mathcal{M}_k^{(L,R)}$, there exist $A \in \mathrm{Skew}(k),\ S \in \mathrm{Sym}(k),\ B \in \mathbb{R}^{(n-k)\times k}$ such that $\dot{L} = UA + U_\perp B + US$. That is because we can write $\dot{L} = \Pi_U \dot{L} + (I - \Pi_U)\dot{L}$ where $\Pi_U : \mathbb{R}^{m\times k} \to T_U \mathrm{St}(m, k)$ is the orthogonal projection onto $T_U \mathrm{St}(m, k)$ so $I - \Pi_U$ is orthogonal projection onto $T_U^\perp \mathrm{St}(m, k)$. (In fact, these A, B, S are unique because $B = U_\perp^T \dot{L}$, and S, A are the symmetric and skew-symmetric parts of $U^T \dot{L}$, respectively.) Defining $\dot{U} = UA + U_\perp B \in T_U \mathrm{St}(m, k)$ and $\dot{W} = \dot{R} + WS$ so $(\dot{U}, \dot{W}) \in T_{(U,W)}\overline{\mathcal{M}}_k^{(U,W)}$. A simple computation shows that

$$\begin{aligned} &\dot{L}W^T + U\dot{R}^T = \dot{U}W^T + U\dot{W}^T, \\ &\dot{L}\dot{R}^T - \dot{U}\dot{W}^T = \Big(U(A - I)S + U_\perp BS\Big)W^T = \Big(U(A - I)S + U_\perp BS\Big)U^T X \in T_X \mathcal{M}_{\mathrm{rank}(X)}, \end{aligned} \tag{2.39}$$

where for the last equality we used $X = UW^T \implies U^T X = W^T$ since $U \in \mathrm{St}(m, k)$, and for the last inclusion we used $X P^\perp_{\mathrm{row}(X)} = 0$. From Eq. (2.12), we get

$$\langle \nabla^2 g(U, W)[\dot{L}, \dot{R}], (\dot{L}, \dot{R})\rangle - \langle \nabla^2 g(U, W)[\dot{U}, \dot{W}], (\dot{U}, \dot{W})\rangle = \langle \nabla f(UW^T), \Big(U(A-I)S + U_\perp BS\Big)U^T X\rangle. \tag{2.40}$$

Since (U, W) is in particular 1-critical for g on all of $\mathcal{M}_k^{(L,R)}$ by the first part of the proof, Prop. 2.8 implies

$\nabla f(UW^T) \perp T_X \mathcal{M}_{\mathrm{rank}(X)}$. Thus, we have

$$\langle \nabla^2 g(U,W)[\dot{L},\dot{R}],(\dot{L},\dot{R})\rangle = \langle \nabla^2 g(U,W)[\dot{U},\dot{W}],(\dot{U},\dot{W})\rangle \geq 0, \tag{2.41}$$

so (U,W) is 2-critical for g on all of $\mathcal{M}_k^{(L,R)}$. □

Since the first and second order optimality conditions are the same, all the conclusions of the preceding discussion of $\mathcal{M}_k^{(L,R)}$ carries through to $\mathcal{M}_k^{(U,W)}$. The only results in the analysis of $\mathcal{M}_k^{(L,R)}$ not carrying through immediately are those asserting the existence of 'bad points', namely Props. 2.12, 2.13, 2.17. However, examining the proofs we see that all the points given there satisfy $L \in \mathrm{St}(m,k)$, so those counterexamples in fact lie in $\overline{\mathcal{M}}_k^{(U,W)}$ and carry through to $\mathcal{M}_k^{(U,W)}$.

The (X,Y) desingularization.

As before, by Prop. 1.2 it suffices to characterize the critical points on the total space $\overline{\mathcal{M}}_k^{(X,Y)}$, which can be written as

$$\overline{\mathcal{M}}_k^{(X,Y)} = \{(X,Y) \in \mathbb{R}^{m\times n} \times \mathbb{R}^{n\times(n-k)} : XY = 0,\ Y^TY = I\}. \tag{2.42}$$

The Lagrangian for the problem is then

$$\mathcal{L}(X,Y,\Lambda,M) = f(X) + \langle XY,\Lambda\rangle + \frac{1}{2}\langle Y^TY - I_{n-k}, M\rangle, \tag{2.43}$$

where $\Lambda \in \mathbb{R}^{m\times(n-k)}$ and $M \in \mathrm{Sym}(n-k)$ are matrices of Lagrange multipliers. The first-order optimality condition then reads (as in [10, Eq. (17)]):

$$\begin{cases} \nabla f(X) + \Lambda Y^T = 0, \\ X^T\Lambda = 0, \\ XY = 0, \\ Y^TY = I_{n-k}. \end{cases} \tag{2.44}$$

Here Λ is a matrix of Lagrange multipliers. Multiplying the first equation on the right by Y and using $Y^TY = I$, we obtain $\Lambda = -\nabla f(X)Y$. Hence the first equation reads

$$0 = \nabla f(X) + \Lambda Y^T = \nabla f(X)(I - YY^T) = \nabla f(X) P^{\perp}_{\mathrm{col}(Y)}, \tag{2.45}$$

where for the last equality we used the fact that the columns of Y are orthonormal. The second equation then reads

$$0 = X^T\Lambda = -X^T\nabla f(X)Y \iff X^T\nabla f(X)Y = 0 \iff P_{\mathrm{col}(X)}\nabla f(X)P_{\mathrm{col}(Y)} = 0. \tag{2.46}$$

For the last equivalence, we argue as follows. Supposing $X^T\nabla f(X)Y = 0$, multiply on the left by $(X^\dagger)^T$ and on the right by Y^T to obtain $P_{\mathrm{col}(X)}\nabla f(X)P_{\mathrm{col}(Y)} = 0$. In the other direction, suppose $0 = P_{\mathrm{col}(X)}\nabla f(X)P_{\mathrm{col}(Y)} = (X^\dagger)^T X^T\nabla f(X)YY^T = 0$. Multiply on the left by X^T and on the right by Y and use $XX^\dagger X = X$ and $Y^TY = I$ to get $X^T\nabla f(X)Y = 0$. Thus, the two equations are in fact equivalent. Finally, since

$$P_{\mathrm{col}(X)}\nabla f(X) = P_{\mathrm{col}(X)}\nabla f(X)(P_{\mathrm{col}(Y)} + P^\perp_{\mathrm{col}(Y)}), \tag{2.47}$$

the two equations $\nabla f(X)P^\perp_{\mathrm{col}(Y)} = 0$ and $P_{\mathrm{col}(X)}\nabla f(X)P_{\mathrm{col}(Y)} = 0$ are equivalent to the two equations $\nabla f(X)P^\perp_{\mathrm{col}(Y)} = 0$ and $P_{\mathrm{col}(X)}\nabla f(X) = 0$. We thus obtain the equivalent set of first order optimality conditions:

$$\begin{cases} \nabla f(X)P^\perp_{\mathrm{col}(Y)} = 0, \\ P_{\mathrm{col}(X)}\nabla f(X) = 0, \\ XY = 0, \\ Y^TY = I_{n-k}. \end{cases} \tag{2.48}$$

Proposition 2.21. If (X, Y) is 1-critical on $\overline{\mathcal{M}}_k^{(X,Y)}$, then X is 1-critical on its own stratum.

Proof. Because $\mathrm{col}(Y) \subseteq \ker(X) = \mathrm{row}(X)^\perp$, we have $\mathrm{col}(Y)^\perp \supseteq \mathrm{row}(X)$ so the first optimality condition above implies $\nabla f(X)P_{\mathrm{row}(X)} = 0$. Therefore,

$$\Pi_X\nabla f(X) = P_{\mathrm{col}(X)}\nabla f(X) + \nabla f(X)P_{\mathrm{row}(X)} - P_{\mathrm{col}(X)}\nabla f(X)P_{\mathrm{row}(X)} = 0, \tag{2.49}$$

so $\nabla f(X) \perp T_X\mathcal{M}_{\mathrm{rank}(X)}$ and X is 1-critical on its own stratum. □

Proposition 2.22. Suppose $\mathrm{rank}(X) = k$. Then (X, Y) is 1-critical on $\overline{\mathcal{M}}_k^{(X,Y)}$ if and only if X is 1-critical on $\mathcal{M}_{\leq k}$.

Proof. Since $\mathrm{rank}(X) = k$, we must have $\mathrm{col}(Y) = \ker(X)$ so $\mathrm{col}(Y)^\perp = \mathrm{row}(X)$. Thus, the 1-criticality conditions on $\overline{\mathcal{M}}_k^{(X,Y)}$ reduce to $\nabla f(X)P_{\mathrm{row}(X)} = 0$ and $P_{\mathrm{col}(X)}\nabla f(X) = 0$, which are equivalent to 1-criticality of X on $\mathcal{M}_{\leq k}$ by Cor. 2.10. □

Proposition 2.23. If (X, Y) is 1-critical on $\overline{\mathcal{M}}_k^{(X,Y)}$ and $\mathrm{rank}(X) < k$, then X may not be 1-critical on all

of $\mathcal{M}_{\leq k}$.

Proof. This was shown in general in Cor. 2.7, but we can also give an explicit example here. Eq. (2.48) only gives $\nabla f(X) = P^{\perp}_{\mathrm{col}(X)} \nabla f(X) P_{\mathrm{col}(Y)}$ but this quantity is nonzero in general. For example, $X^* = \mathrm{diag}(a, b, 0)$, $X = \mathrm{diag}(a, 0, 0)$ and $f(X) = \frac{1}{2}||X - X^*||_F^2$ viewed as a function over $\mathcal{M}_{\leq 2}$. Here $m = n = 3$ and $k = 2$. Choose $Y = (0, 1, 0)^T \in \mathrm{St}(3, 1)$ which indeed satisfies $Y^T Y = 1$ and $XY = 0$. Here $\nabla f(X) = \mathrm{diag}(0, -b, 0)$ so

$$\begin{aligned} \nabla f(X) P^{\perp}_{\mathrm{col}(Y)} &= \mathrm{diag}(0, -b, 0)\mathrm{diag}(1, 0, 1) = 0, \\ P_{\mathrm{col}(X)} \nabla f(X) &= \mathrm{diag}(1, 0, 0)\mathrm{diag}(0, -b, 0) = 0, \end{aligned} \tag{2.50}$$

so (X, Y) is 1-critical on $\overline{\mathcal{M}}_k^{(X,Y)}$. However, $\nabla f(X) \neq 0$ so X is not first-order critical on $\mathcal{M}_{\leq k}$. □

Once again, this result follows more generally from Cor. 2.7.

We proceed to derive second order optimality conditions for $\bar{f}(X, Y) = f(X)$ on $\overline{\mathcal{M}}_k^{(X,Y)}$. As shown in [10, Sec. 2.3], the horizontal space is given by

$$H_{(X,Y)} = \{(\dot{X}, \dot{Y}) \in \mathbb{R}^{m \times n} \times \mathbb{R}^{n \times (n-k)} : \dot{X}Y + X\dot{Y} = 0,\ \dot{Y}^T Y = 0\}. \tag{2.51}$$

The second order condition then reads [20, Thm. 3.46]

$$\langle \nabla^2 f(X)[\dot{X}], \dot{X} \rangle - 2\langle \nabla f(X) Y, \dot{X}\dot{Y} \rangle \geq 0, \tag{2.52}$$

for all $(\dot{X}, \dot{Y}) \in H_{(X,Y)}$ where we restricted to perturbations in the horizontal space by Cor. 1.3.

Proposition 2.24. Any 2-critical point for $\bar{f}$ on $\overline{\mathcal{M}}_k^{(X,Y)}$ maps to a 1-critical point for f on $\mathcal{M}_{\leq k}$.

Proof. The proof is inspired by [9, Thm. 1]. If $r = \mathrm{rank}(X)$, let $U \in \mathrm{St}(m, r)$ and $V \in \mathrm{St}(n, r)$ satisfy

$$U^T X V = \Sigma_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r), \quad \sigma_i \neq 0 \text{ for all } i. \tag{2.53}$$

Starting from $XY = U\Sigma_r V^T Y = 0$ and multiplying by $\Sigma_r^{-1} U^T$ on the left we obtain $V^T Y = 0$. Thus, if we concatenate the columns of V and Y we get a Stiefel matrix $(V\ Y) \in \mathrm{St}(n, n-(k-r))$. Define $Y' \in \mathrm{St}(n, k-r)$ and $U_\perp \in \mathrm{St}(m, m-r)$ so that the concatenated matrices $\bar{U} = (U\ U_\perp) \in O(m)$ and $\bar{V} = (V\ Y'\ Y) \in O(n)$. In this basis, we have

$$\bar{U}^T X \bar{V} = \begin{pmatrix} \Sigma_r & \mathbf{0}_{r \times (n-r)} \\ \mathbf{0}_{(m-r) \times r} & \mathbf{0}_{(m-r) \times (n-r)} \end{pmatrix}, \quad \bar{V}^T Y = \begin{pmatrix} \mathbf{0}_{k \times (n-k)} \\ I_{n-k} \end{pmatrix}. \tag{2.54}$$

Let u_1 and v_1 be the leading left and right singular vectors for $\nabla f(X)Y$, so $u_1^T \nabla f(X) Y v_1 = ||\nabla f(X)Y||$. Let

$$\dot{X} = \begin{pmatrix} \mathbf{0}_{m\times(k-1)} & u_1 & \mathbf{0}_{m\times(n-k)} \end{pmatrix} \bar{V}^T, \quad \dot{Y} = \alpha \bar{V} \begin{pmatrix} \mathbf{0}_{(k-1)\times(n-k)} \\ v_1^T \\ \mathbf{0}_{(n-k)\times(n-k)} \end{pmatrix}, \tag{2.55}$$

where $\alpha \in \mathbb{R}$ is arbitrary. Since $r \leq k-1$ and $\bar{U}, \bar{V}$ are orthogonal matrices,

$$X\dot{Y} = 0 \iff (\bar{U}^T X \bar{V})(\bar{V}^T \dot{Y}) = \alpha \begin{pmatrix} \Sigma_r & \mathbf{0}_{r\times(n-r)} \\ \mathbf{0}_{(m-r)\times r} & \mathbf{0}_{(m-r)\times(n-r)} \end{pmatrix} \begin{pmatrix} \mathbf{0}_{(k-1)\times(n-k)} \\ v_1^T \\ \mathbf{0}_{(n-k)\times(n-k)} \end{pmatrix} = 0. \tag{2.56}$$

Similarly,

$$\dot{X}Y = 0 \iff (\dot{X}\bar{V})(\bar{V}^T Y) = \begin{pmatrix} \mathbf{0}_{m\times(k-1)} & u_1 & \mathbf{0}_{m\times(n-k)} \end{pmatrix} \begin{pmatrix} \mathbf{0}_{k\times(n-k)} \\ I_{n-k} \end{pmatrix} = 0, \tag{2.57}$$

and

$$\begin{aligned} \dot{Y}^T Y &= \alpha \begin{pmatrix} \mathbf{0}_{(n-k)\times(k-1)} & v_1 & \mathbf{0}_{(n-k)\times(n-k)} \end{pmatrix} (\bar{V}^T Y) \\ &= \alpha \begin{pmatrix} \mathbf{0}_{(n-k)\times(k-1)} & v_1 & \mathbf{0}_{(n-k)\times(n-k)} \end{pmatrix} \begin{pmatrix} \mathbf{0}_{k\times(n-k)} \\ I_{n-k} \end{pmatrix} = 0. \end{aligned} \tag{2.58}$$

Finally,

$$\dot{X}\dot{Y} = \alpha \begin{pmatrix} \mathbf{0}_{m\times(k-1)} & u_1 & \mathbf{0}_{m\times(n-k)} \end{pmatrix} (\bar{V}^T \bar{V}) \begin{pmatrix} \mathbf{0}_{(k-1)\times(n-k)} \\ v_1^T \\ \mathbf{0}_{(n-k)\times(n-k)} \end{pmatrix} = \alpha u_1 v_1^T. \tag{2.59}$$

The second order condition gives us

$$\langle \nabla^2 f(X)[\dot{X}], \dot{X} \rangle \geq 2\alpha ||\nabla f(X)Y||, \tag{2.60}$$

for all $\alpha \geq 0$. Since $\dot{X}$ is independent of α, taking $\alpha \to \infty$ implies $||\nabla f(X)Y|| = 0$. By the first order critical conditions, we then have

$$\nabla f(X) = P^{\perp}_{\mathrm{col}(X)} \nabla f(X) P_{\mathrm{col}(Y)} = P^{\perp}_{\mathrm{col}(X)} (\nabla f(X)Y) Y^T = 0. \tag{2.61}$$

Thus, X is 1-critical for f on all of $\mathcal{M}_{\leq k}$.

If $r = k$ then $\text{rank}(X) = k$ and first order criticality for $\bar{f}$ is sufficient for f on $\mathcal{M}_k$ as shown above. □

Comparison with factorization

We now compare with the approach optimizing $g(L, R) = f(LR^T)$ over $\mathcal{M}_k^{(L,R)}$. Specifically, we compare the set of points on $\mathcal{M}_{\leq k}$ that have preimages that are 1-critical on either $\mathcal{M}_k^{(X,Y)}$ or $\mathcal{M}_k^{(L,R)} = \mathbb{R}^{m\times k} \times \mathbb{R}^{n\times k}$.

First, we consider the maximal rank case:

Proposition 2.25. $X \in \mathcal{M}_k$ is 1-critical for f on $\mathcal{M}_{\leq k}$ iff (L, R) is 1-critical for g on $\mathbb{R}^{m\times k} \times \mathbb{R}^{n\times k}$ for any one factorization $X = LR^T$, iff $(X, \ker(X))$ is 1-critical for f on $\mathcal{M}_k^{(X,Y)}$.

These equivalences were already shown in Cor. 2.10 and Prop. 2.22.

The rank-deficient case is more delicate. Denote by $\bar{f} \colon \overline{\mathcal{M}}_k^{(X,Y)} \to \mathbb{R}$ the induced cost function.

Proposition 2.26. Suppose $X = LR^T$ is a factorization of X where $\text{rank}(L) = \text{rank}(R) = \text{rank}(X)$. If (X, Y) is 1-critical for $\bar{f}$ for some $Y \in \text{St}(n, n-k)$, then (L, R) is 1-critical for g.

Proof. A point (L, R) is 1-crtical for g iff $P_{\text{col}(L)}\nabla f(X) = 0$ and $\nabla f(X)P_{\text{col}(R)} = 0$.

Since (X, Y) is 1-critical for $\bar{f}$, we have $P_{\text{col}(X)}\nabla f(X) = 0$ and $\nabla f(X)P_{\text{col}(Y)^\perp} = 0$. Since $\text{rank}(L) = \text{rank}(X)$ and $X = LR^T$, we have $\text{col}(X) = \text{col}(L)$ so $P_{\text{col}(X)}\nabla f(X) = 0 \iff P_{\text{col}(L)}\nabla f(X) = 0$. Since $\text{rank}(R) = \text{rank}(X)$, we have $\text{col}(R) = \text{row}(X) = \ker(X)^\perp \subseteq \text{col}(Y)^\perp$. Therefore, $\nabla f(X)P_{\text{col}(Y)^\perp} = 0 \implies \nabla f(X)P_{\text{col}(R)} = 0$. □

Proposition 2.27. There exists f and X such that for any factorization $X = LR^T$ satisfying $\text{rank}(L) = \text{rank}(R) = \text{rank}(X) < k$ and (L, R) is 1-critical for g, there exists $Y \in \text{St}(n, n-k)$ such that $XY = 0$ and (X, Y) is not 1-critical for $\bar{f}$.

Proof. Let $f(X) = \frac{1}{2}||X + X^*||_F^2$ for $X^* \in \mathcal{M}_k$. Then let $X = L = R = 0$. Note that $(0, 0)$ is 1-critical for g. On the other hand, $\nabla f(0) = X^*$, and let $Y \in \text{St}(n, n-k)$ be such that its columns span $\ker(X^*)$. Then $P_{\text{col}(Y)^\perp} = P_{\text{row}(X^*)}$ and $\nabla f(X)P_{\text{col}(Y)^\perp} = X^*P_{\text{row}(X^*)} = X^* \neq 0$, so (X, Y) is not 1-critical for $\bar{f}$. □

To relate the second order conditions, observe:

Proposition 2.28. $(\dot{X}, \dot{Y}) \in H_{(X,Y)}$ is in the Horizontal space for $X = LR^T \in \mathcal{M}_k$ (note: $\text{rank}(X) = k$) if and only if there exist $(\dot{L}, \dot{R}) \in \mathbb{R}^{m\times k} \times \mathbb{R}^{n\times k}$ such that

$$\dot{X} = \dot{L}R^T + L\dot{R}^T, \quad \dot{Y} = -X^\dagger \dot{X} Y = -(R^\dagger)^T \dot{R}^T Y. \tag{2.62}$$

Proof. First, we verify that $(\dot{X}, \dot{Y})$ given above are in the horizontal space to (X, Y). Note that $0 = XY = LR^TY$ implies, after multiplying on the left by $(L^TL)^{-1}L^T$ that $R^TY = 0$. Therefore,

$$\begin{aligned} &\dot{X}Y = \dot{L}R^TY + L\dot{R}^TY = L\dot{R}^TY, \\ &X\dot{Y} = -(XX^\dagger)(\dot{X}Y) = -P_{\mathrm{col}(X)}L\dot{R}^TY = -P_{\mathrm{col}(L)}L\dot{R}^TY = -L\dot{R}^TY \\ &\implies \dot{X}Y + X\dot{Y} = 0. \end{aligned} \tag{2.63}$$

Similarly, note that $(X^\dagger)^TY = (L^\dagger)^TR^\dagger Y = (L^\dagger)^T(R^TR)^{-1}(R^TY) = 0$. Therefore,

$$\dot{Y}^TY = -Y^T\dot{X}(X^\dagger)^TY = 0. \tag{2.64}$$

Thus, the given pair is indeed in the horizontal space.

Conversely, suppose $(\dot{X}, \dot{Y}) \in H_{(X,Y)}$. First, multiplying the equation $\dot{X}Y + X\dot{Y} = 0$ by $P^\perp_{\mathrm{col}(X)}$ on the left and by Y^T on the right we obtain $P^\perp_{\mathrm{col}(X)}\dot{X}YY^T = P^\perp_{\mathrm{col}(X)}\dot{X}P^\perp_{\mathrm{row}(X)} = 0$, hence $\dot{X} \in T_X\mathcal{M}_k$. Therefore, there exist $(\dot{L}, \dot{R})$ satisfying $\dot{X} = \dot{L}R^T + L\dot{R}^T$. Next, differentiating the expression $R^TY = 0$ we obtain $R^T\dot{Y} + \dot{R}^TY = 0$ so $R^T\dot{Y} = -\dot{R}^TY$. Since $Y^T\dot{Y} = 0$, we conclude that $P_{\mathrm{col}(Y)^\perp}\dot{Y} = \dot{Y}$. Since $\mathrm{col}(R) = \mathrm{row}(X) = \mathrm{col}(Y)^\perp$, we have

$$\dot{Y} = P_{\mathrm{col}(R)}\dot{Y} = (R^\dagger)^TR^T\dot{Y} = -(R^\dagger)^T\dot{R}^TY. \tag{2.65}$$

Now note that

$$-X^\dagger\dot{X}Y = -(R^\dagger)^T(\underbrace{L^\dagger}_{=(L^TL)^{-1}L^T}L)\dot{R}^TY = -(R^\dagger)^T\dot{R}^TY = \dot{Y}. \tag{2.66}$$

Thus, $(\dot{X}, \dot{Y})$ are of the claimed form. □

Proposition 2.29. Fix $X \in \mathcal{M}_k$ (note: $\mathrm{rank}(X) = k$), and suppose that there exists $Y \in \mathrm{St}(n, n-k)$ such that $XY = 0$ and (X, Y) is 2-critical for $\bar{f}$. Then (L, R) is 2-critical for g for any factorization $X = LR^T$. Conversely, if there exists a factorization $X = LR^T$ such that (L, R) is 2-critical for g, then (X, Y) is 2-critical for $\bar{f}$ for any $Y \in \mathrm{St}(n, n-k)$ such that $XY = 0$.

Proof. By Prop. 2.25 we only need to check the equivalence of second order optimality conditions. Let (L, R) be any factorization of X and $Y \in \mathrm{St}(n, n-k)$ any matrix whose columns span $\ker(X)$. We show that the quadratic forms giving the second-order optimality conditions for the two approaches are equal at first-order critical points, in the sense of the 'dictionary' between the two tangent spaces given by Prop. 2.28.

The quadratic form whose non-negativity is the second-order condition for g is

$$Q_g(\dot{L},\dot{R}) = \langle \nabla^2 f(X)[\dot{L}R^T + L\dot{R}^T], \dot{L}R^T + L\dot{R}^T\rangle + 2\langle \nabla f(X), \dot{L}\dot{R}^T\rangle, \tag{2.67}$$

and the one for $\bar{f}$ is

$$Q_{\bar{f}}(\dot{X},\dot{Y}) = \langle \nabla^2 f(X)[\dot{X}], \dot{X}\rangle - 2\langle \nabla f(X)Y, \dot{X}\dot{Y}\rangle. \tag{2.68}$$

By Prop. 2.28, we have $(\dot{X},\dot{Y}) \in H_{(X,Y)}$ iff there exist $(\dot{L},\dot{R})$ such that $\dot{X} = \dot{L}R^T + L\dot{R}^T$ and $\dot{Y} = -X^\dagger \dot{X} Y$. As in Eq. (2.23), we have $\dot{X}X^\dagger\dot{X} = \dot{L}\dot{R}^T + W$ where $W \in T_X\mathcal{M}_k$. Moreover, at 1-critical points we have $\nabla f(X) \perp T_X\mathcal{M}_k$ so $\nabla f(X) = P_{\mathrm{col}(X)^\perp} \nabla f(X) P_{\mathrm{row}(X)^\perp}$. Also $YY^T = P_{\mathrm{col}(Y)} = P_{\ker(X)} = P_{\mathrm{row}(X)^\perp}$. Therefore, $\nabla f(X) YY^T = P_{\mathrm{col}(X)^\perp} \nabla f(X) P_{\mathrm{row}(X)^\perp} = \nabla f(X)$. Therefore,

$$\langle \nabla f(X) YY^T, \dot{X}X^\dagger\dot{X}\rangle = \langle \nabla f(X), \dot{L}\dot{R}^T\rangle. \tag{2.69}$$

Thus, we get

$$Q_{\bar{f}}(\dot{X}(\dot{L},\dot{R}),\dot{Y}(\dot{L},\dot{R})) = \langle \nabla^2 f(X)[\dot{L}R^T + L\dot{R}^T], \dot{L}R^T + L\dot{R}^T\rangle + 2\langle \nabla f(X), \dot{L}\dot{R}^T\rangle = Q_g(\dot{L},\dot{R}). \tag{2.70}$$

This shows the desired equivalence. □

2.4 Analysis of local minima

We ask whether a local minimum in a lifted space necessarily corresponds to a local minimum on the original variety.

Local minima on $\mathcal{M}_k^{(L,R)}$ and $\mathcal{M}_k^{(U,W)}$.

We begin by showing that there are points on both lifts that can be local minima on the lift but map to saddles on the original variety $\mathcal{M}_{\leq k}$:

Proposition 2.30. There is a cost function f and factorizations $(L,R) \in \mathrm{St}(m,k) \times \mathbb{R}^{n\times k}$ for which (L,R) is a local minimum for $g(L,R) = f(LR^T)$ but $X = LR^T$ is a saddle point for f.

Proof. Let $X = \mathbf{0}_{3\times 3}$ and

$$L = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 0 & 0 \end{pmatrix}, \quad R = \mathbf{0}_{3\times 2}, \tag{2.71}$$

and $f\colon \mathcal{M}_{\leq 2}^{3\times 3} \to \mathbb{R}$ be given by

$$f(X) = X_{13}^2 + X_{23}^2 - X_{33}^2. \tag{2.72}$$

For perturbations (P, Q) of (L, R) such that $||P||_\infty, ||Q||_\infty \leq \epsilon$, we have

$$\begin{aligned} g(L+P, R+Q) &= \left[(1/\sqrt{2} + P_{11})Q_{31} + (1/\sqrt{2} + P_{12})Q_{31}\right]^2 + \left[(1/\sqrt{2} + P_{21})Q_{31} + (-1/\sqrt{2} + P_{22})Q_{32}\right]^2 \\ &\quad - \left[P_{31}Q_{31} + P_{31}Q_{31}\right]^2 \\ &= (Q_{31} + Q_{32})^2/2 + (Q_{31} - Q_{32})^2/2 \\ &\quad + \sqrt{2}(Q_{31} + Q_{32})(P_{11}Q_{31} + P_{12}Q_{32}) + \sqrt{2}(Q_{31} - Q_{32})(P_{21}Q_{31} + P_{22}Q_{32}) \\ &\quad + (P_{11}Q_{31} + P_{12}Q_{32})^2 + (P_{21}Q_{31} + P_{22}Q_{32})^2 - (P_{31}Q_{31} + P_{32}Q_{32})^2. \end{aligned} \tag{2.73}$$

Note that the first row in the last equality contains terms of order ϵ^2 which are both positive and cannot be made simultaneously zero without making the rest of the terms zero as well. On the other hand, the second and third rows contain terms of order ϵ^3, ϵ^4, respectively. Since the first row dominates for small ϵ, we expect $g(L+P, R+Q) \geq 0 = g(L, R)$ for all small enough ϵ, which would show (L, R) is a local minimum.

We now make the above intuition precise. Define:

$$q = \begin{pmatrix} Q_{31} \\ Q_{32} \end{pmatrix},\ p = \begin{pmatrix} P_{31} \\ P_{32} \end{pmatrix},\ M = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},\ P = \begin{pmatrix} P_{11} & P_{21} \\ P_{12} & P_{22} \end{pmatrix}. \tag{2.74}$$

Also define

$$\begin{aligned} A &:= (Q_{31} + Q_{32})^2/2 + (Q_{31} - Q_{32})^2/2, \\ B &:= \sqrt{2}(Q_{31} + Q_{32})(P_{11}Q_{31} + P_{12}Q_{32}) + \sqrt{2}(Q_{31} - Q_{32})(P_{21}Q_{31} + P_{22}Q_{32}), \\ C &:= (P_{11}Q_{31} + P_{12}Q_{32})^2 + (P_{21}Q_{31} + P_{22}Q_{32})^2 - (P_{31}Q_{31} + P_{32}Q_{32})^2, \end{aligned} \tag{2.75}$$

so $g(L+P, R+Q) = A + B + C$. Note that

$$\begin{aligned} A &= ||Mq||_2^2 \geq ||M^{-1}||_{op}^{-2} ||q||_2^2, \\ B &= 2q^T PMq = 2\langle PM, qq^T \rangle \geq -2||PM||_F \cdot ||q||_2^2 \geq -2||M||_F ||P||_F \cdot ||q||_2^2 \geq -4\epsilon ||M|| \cdot ||q||_2^2, \\ C &= ||P^T q||_2^2 - \langle q, p \rangle^2 \geq -||p||_2^2 ||q||_2^2 \geq -2\epsilon^2 ||q||_2^2. \end{aligned} \tag{2.76}$$

Overall, we have

$$g(L+P, R+Q) \geq (||M^{-1}||_{op}^{-2} - 4\epsilon ||M||_F - 2\epsilon^2)||q||_2^2 \geq 0 = g(L, R), \quad \text{for all small } \epsilon. \tag{2.77}$$

(Note that M is a fixed matrix independent of P, Q, ϵ). Thus, we conclude that (L, R) is a local minimum for g on $\mathbb{R}^{m\times k} \times \mathbb{R}^{n\times k}$.

On the other hand, $X = LR^T = 0$ is a saddle point for f on $\mathcal{M}_{\leq 2}^{3\times 3}$, since $f(\text{diag}(0, 0, \epsilon)) = -\epsilon^2 < 0 = f(X)$ and $f(\epsilon E_{13}) = \epsilon^2 > 0 = f(X)$ where ϵ is arbitrary and $E_{13} = e_1 e_3^T$ has a 1 in the $(1, 3)$ entry and zero otherwise. □

If one requires f to be bounded from below, we can add to f a smooth function vanishing in a neighborhood of 0 but increasing rapidly outside of it. Alternatively, one can show that $f(X) + X_{33}^4$ is bounded from below and works as a counterexample.

Since the counterexample we gave above has $L \in \text{St}(m, k)$, we immediately get a similar result for our second lift:

Corollary 2.31. *There is a cost function f and factorization $(U, W) \in \overline{\mathcal{M}}_k^{(U,W)}$ such that $[(U, W)] \in \mathcal{M}_k^{(U,W)}$ is a local minimum on $\mathcal{M}_k^{(U,W)}$ while $X = UW^T$ is a saddle point for f on $\mathcal{M}_{\leq k}$.*

Proof. Let f and $(U, W) = (L, R)$ be as in the proof of Prop. 2.30. Since (U, W) is a local minimum on all of $\mathcal{M}_k^{(L,R)}$, it is in particular a local minimum on $\overline{\mathcal{M}}_k^{(U,W)}$. By Prop. 2.2, we conclude that $[(U, W)]$ is a local minimum on $\mathcal{M}_k^{(U,W)}$. However, as shown in the proof of Prop. 2.30, $X = UW^T$ is a saddle on $\mathcal{M}_{\leq k}$. □

Balanced factorizations

In spite of the above counterexample, we can show that a large set of points on $\mathcal{M}_k^{(L,R)}$ cannot be 'false local minima' in the above sense. Specifically, if we consider 'balanced' factorizations (L, R) of $X = LR^T$ which satisfy $\text{rank}(L) = \text{rank}(R) = \text{rank}(X)$, then we show that the lift satisfies the approximate subsequence lifting property (ASLP) at (L, R) introduced in Sec. 2.1.1. Hence by Prop. 2.3, X is a local minimum on $\mathcal{M}_{\leq k}$ if and only if (L, R) is a local minimum on $\mathcal{M}_k^{(L,R)}$. To show ASLP at (L, R), we first need to characterize the non-uniqueness in the factorization $X = LR^T$.

Lemma 2.32. *Suppose $X = L_1 R_1^T = L_2 R_2^T$ are two factorizations for X satisfying $\text{rank}(L_i) = \text{rank}(R_i) = \text{rank}(X)$ for $i = 1, 2$. Then there exists an invertible $J \in GL(k)$ such that $(L_1, R_1) = (L_2 J, R_2 J^{-T})$.*

Proof. Say $\text{rank}(X) = r \leq k$, and fix a particular factorization of X satisfying the hypotheses of the form $L = (\widetilde{L}, 0)$ and $R = (\widetilde{R}, 0)$ where $\widetilde{L} \in \mathbb{R}^{m\times r}$ and $\widetilde{R} \in \mathbb{R}^{n\times r}$ are full rank. Such a factorization can be obtained e.g. from the SVD of X. It suffices to show that $(L_1, R_1) = (LJ, RJ^{-T})$ for this fixed factorization and some invertible $J \in \text{GL}(k)$.

Since $\text{rank}(L_1) = \text{rank}(L) = \text{rank}(X)$, we must have $\text{col}(L_1) = \text{col}(L) = \text{col}(X)$. Since the columns of $\widetilde{L}, L_1$ span the same subspace, there exists $\widetilde{J} \in \mathbb{R}^{r\times k}$ of rank r satisfying $L_1 = \widetilde{L}\widetilde{J}$. Similarly, since

$\operatorname{rank}(R_1) = \operatorname{rank}(R) = \operatorname{row}(X)$, there exists $\widetilde{K} \in \mathbb{R}^{r\times k}$ of rank r such that $R_1 = RK = \widetilde{R}\widetilde{K}$. Since $LR^T = L_1R_1^T$ we conclude that $\widetilde{L}(I - \widetilde{J}\widetilde{K}^T)\widetilde{R}^T = 0$. Since $\widetilde{L}, \widetilde{R}$ have full column rank, multiplying on the left by $\widetilde{L}^\dagger$ and on the right by $\widetilde{R}^{\dagger T}$ we get $\widetilde{J}\widetilde{K}^T = I$. Extend the r linearly independent rows of $\widetilde{J}$ to a basis by adding $k-r$ vectors, and concatenate the transpose of those vectors to form $\widetilde{J}' \in \mathbb{R}^{(k-r)\times k}$ such that $J = \begin{pmatrix} \widetilde{J} \\ \widetilde{J}' \end{pmatrix}$ is invertible. Then, define a matrix $\widetilde{K}' \in \mathbb{R}^{(r-k)\times k}$ by its action on the basis of $\mathbb{R}^k$ consisting of the rows of J, namely, $\widetilde{K}'\widetilde{J}^T = 0$ and $\widetilde{K}'\widetilde{J}'^T = I_{k-r}$. In other words, $\widetilde{K}' = (0, I_{k-r})J^{-T}$ which clearly has rank $k-r$. Moreover, if we define $K = \begin{pmatrix} \widetilde{K} \\ \widetilde{K}' \end{pmatrix}$ then $JK^T = I_k$, hence $K = J^{-T}$. Finally, we have by construction that $L_1 = \widetilde{L}\widetilde{J} = LJ$ and $R_1 = \widetilde{R}\widetilde{K} = RK = RJ^{-T}$. □

A particularly nice set of balanced factorizations is given by:

Lemma 2.33. *Suppose $X = LR^T$ for $L \in \mathbb{R}^{m\times k}$ and $R \in \mathbb{R}^{n\times k}$ such that $L^TL = R^TR$. Then $\operatorname{rank}(L) = \operatorname{rank}(R) = \operatorname{rank}(X)$.*

Proof. Let $L^TL = R^TR = V\Sigma^2V^T$ be the eigendecomposition where $V \in O(k)$ and $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_k) \succeq 0$. By the polar decomposition, we can write $L = W_L\sqrt{L^TL} = (W_LV)\Sigma V^T$ where $W_L \in \operatorname{St}(m,k)$ and $\sqrt{L^TL}$ is the unique square root of a positive-semidefinite matrix, and similarly $R = W_R\sqrt{R^TR} = (W_RV)\Sigma V^T$ where $W_R \in \operatorname{St}(n,k)$. Defining $U_L = W_LV$ and $U_R = W_RV$, note that $U_L \in \operatorname{St}(m,k)$ and $U_R \in \operatorname{St}(n,k)$. Therefore, the decompositions $L = U_L\Sigma V^T$ and $R = U_R\Sigma V^T$ are the SVDs of L, R. Also, $X = LR^T = U_L\Sigma^2U_R^T$ is the SVD of X. Thus, $\operatorname{rank}(X) = \operatorname{rank}(L) = \operatorname{rank}(R) = \operatorname{rank}(\Sigma)$ as claimed. □

We are now ready to prove ASLP for balanced factorizations:

Proposition 2.34. Suppose (L, R) satisfies $\operatorname{rank}(L) = \operatorname{rank}(R) = \operatorname{rank}(LR^T)$. Then (L, R) is a local minimum for g if and only if it $X = LR^T$ is a local minimum for f.

Proof. The map $(L, R) \mapsto LR^T$ is continuous, so by Prop. 2.1 if X is a local minimum, then so is (L, R).

For the converse, we show that the lift $(L, R) \mapsto LR^T$ satisfies ASLP at such balanced factorizations and invoke Prop. 2.3. Pick any $X \in \mathcal{M}_{\leq k}$ and any (L, R) satisfying $X = LR^T$ and $\operatorname{rank}(L) = \operatorname{rank}(R) = \operatorname{rank}(X)$. Such factorizations always exist: If $X = U\Sigma V^T$ is a thin SVD of X with $U \in \operatorname{St}(m,k)$, $V \in \operatorname{St}(n,k)$ and $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_k) \succeq 0$, then we can set $L = U\sqrt{\Sigma}$ and $R = V\sqrt{\Sigma}$ which have $\operatorname{rank}(L) = \operatorname{rank}(R) = \operatorname{rank}(\Sigma) = \operatorname{rank}(X)$.

Let $(X_n) \subset \mathcal{M}_{\leq k}$ be a sequence of points such that $X_n \to X$, and let $(\epsilon_n) \subset \mathbb{R}_{>0}$ satisfy $\epsilon_n \to 0$. Let $X_n = U_n\Sigma_nV_n^T$ be a thin SVD for X_n such that $U_n \in \operatorname{St}(m,k)$, $V_n \in \operatorname{St}(n,k)$ and $\Sigma_n \in \mathbb{R}^{k\times k}$ is diagonal with nonnegative entries on the diagonal. Define $L_n = U_n\sqrt{\Sigma_n}$ and $R_n = V_n\sqrt{\Sigma_n}$. Note that $L_n^TL_n = R_n^TR_n = \Sigma_n$

and $L_n R_n^T = X_n$. Since $||L_n|| = ||R_n|| = \sqrt{||X_n||}$ is bounded, after passing to a subsequence we may assume that $\lim_n L_n = \bar{L}$ and $\lim_n R_n = \bar{R}$ exist. By continuity, we have $\bar{L}\bar{R}^T = X$ and $\bar{L}^T\bar{L} = \bar{R}^T\bar{R}$, which also implies $\operatorname{rank}(X) = \operatorname{rank}(\bar{L}) = \operatorname{rank}(\bar{R})$ by Lemma 2.33. By Lemma 2.32, there exists $J \in \mathrm{GL}(k)$ satisfying $(L, R) = (\bar{L}J, \bar{R}J^{-T}) = \lim_n (L_n J, R_n J^{-T})$. Since $||L_n R_n^T - X_n|| = 0 < \epsilon_n$, this shows that the lift satisfies ASLP at $(X, (L, R))$. □

Corollary 2.35. *A point $X \in \mathcal{M}_k$ of rank k is a local minimum on $\mathcal{M}_{\leq k}$ if and only if any of its factorizations (L, R) is a local minimum on $\mathcal{M}_k^{(L,R)}$.*

Proof. Any such factorization must satisfy $\operatorname{rank}(L) = \operatorname{rank}(R) = \operatorname{rank}(X) = k$. □

Corollary 2.36. *If the cost function f extends to a convex function on $\mathbb{R}^{m\times n}$, then a point $X \in \mathcal{M}_{\leq k}$ is a local minimum if and only if any of its factorizations $(L, R) \in \mathcal{M}_k^{(L,R)}$ is a local minimum.*

Proof. If $\operatorname{rank}(X) = k$, this follows from the preceding Corollary. Suppose $\operatorname{rank}(X) < k$ and (L, R) is a factorization of X which is a local minimum on $\mathcal{M}_k^{(L,R)}$. Then in particular, (L, R) is 2-critical on $\mathcal{M}_k^{(L,R)}$ and by Prop. 2.17 X is 1-critical on $\mathcal{M}_{\leq k}$. Since $\operatorname{rank}(X) < k$, this implies $\nabla f(X) = 0$. By convexity of f, this means X is a global minimum for f on all of $\mathbb{R}^{m\times n}$, so (L, R) is a global minimum as well. □

The set

$$\{(L, R) \in \mathbb{R}^{m\times k} \times \mathbb{R}^{n\times k} : L^T L = R^T R\}, \tag{2.78}$$

is an algebraic variety. If we could optimize over it, we would get a lift with no 'false local minima' by the above argument. Unfortunately, it is singular. Indeed, the differential of the defining equations $F(L, R) = L^T L - R^T R$ can be written

$$\begin{aligned} \operatorname{vec}\big(DF(L, R)[\dot{L}, \dot{R}]\big) &= \operatorname{vec}\big(\dot{L}^T L + L^T \dot{L} - \dot{R}^T R - R^T \dot{R}\big) \\ &= \begin{pmatrix} (I_{k^2} + \Pi_{k,k})(I_k \otimes L^T), & (I_{k^2} + \Pi_{k,k})(I_k \otimes R^T) \end{pmatrix} \begin{pmatrix} \operatorname{vec}(\dot{L}) \\ \operatorname{vec}(\dot{R}) \end{pmatrix}, \end{aligned} \tag{2.79}$$

where $\operatorname{vec}(M)$ is the vector obtained from a matrix M by stacking its columns one on another and $\Pi_{k,k}$ is a permutation matrix satisfying $\Pi_{k,k}\operatorname{vec}(M) = \operatorname{vec}(M^T)$ for all $M \in \mathbb{R}^{k\times k}$. The Jacobian is therefore

$$F'(L, R) = \begin{pmatrix} (I_{k^2} + \Pi_{k,k})(I_k \otimes L^T), & (I_{k^2} + \Pi_{k,k})(I_k \otimes R^T) \end{pmatrix}. \tag{2.80}$$

For $m = n = 3$ and $k = 2$, one can check numerically that $\operatorname{rank}(F'(L, R)) = 3$ for random L, R (the expected codimension, since $L^T L = R^T R$ gives us $\frac{k(k+1)}{2} = 3$ distinct equations), but $\operatorname{rank}(F'(E_{11}, E_{11})) = 2$ where

$E_{11} = e_1 e_1^T$ has a 1 in the $(1,1)$ entry and zero otherwise. Therefore, we cannot simply restrict ourselves to the set of such factorizations.

Local minima on $\mathcal{M}_k^{(X,Y)}$.

We show that this lift has false local minima as well:

Proposition 2.37. There exists a cost function f and point $(X,Y) \in \overline{\mathcal{M}}_k^{(X,Y)}$ such that (X,Y) is a local minimum for f on $\overline{\mathcal{M}}_k^{(X,Y)}$ but X is a saddle point on $\mathcal{M}_{\leq k}$.

Proof. Define $f(X) = \frac{1}{2}\langle \mathcal{A}(X - X^*), X - X^* \rangle$ defined over $\mathcal{M}_{\leq 2}^{3\times 3}$ where $X^* = \mathrm{diag}(1,0,0)$ and $\mathcal{A}$ is the linear operator satisfying

$$\mathcal{A}\begin{pmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & -X_{33} \end{pmatrix}. \tag{2.81}$$

We claim (X^*, e_3) is a local minimum for $\hat{f}(X,Y) = f(X)$, but X^* is a saddle point for f on $\mathcal{M}_{\leq 2}^{3\times 3}$. Here $e_3 = (0,0,1)^T$.

We parameterize $(\overline{\mathcal{M}}_{\leq 2}^{(X,Y)})^{3\times 3}$ near (X^*, e_3). Since $Y^T Y = 1$, write $Y(x,y) = (x, y, \sqrt{1-x^2-y^2})^T$ for x, y near 0 (indeed, $Y(0,0) = e_3$). Next, for all x, y sufficiently close to 0 we have $\sqrt{1-x^2-y^2} > 0$. In this case the equation $XY = 0$ can be solved in terms of X_{i1}, X_{i2} for $i = 1,2,3$:

$$X_{i3} = -\frac{xX_{i1} + yX_{i2}}{\sqrt{1-x^2-y^2}}, \quad i = 1,2,3. \tag{2.82}$$

In the neighborhood $\{(X(X_{i1}, X_{i2}), Y(x,y)) : x^2 + y^2 \leq 1/2\}$ we have:

$$\begin{aligned} \hat{f}(X,Y) &\geq \frac{1}{2}\left[X_{31}^2 + X_{32}^2 - \frac{(xX_{31} + yX_{32})^2}{1-x^2-y^2}\right] \geq \frac{(X_{31}^2 + X_{32}^2)(1-x^2-y^2) - (x^2+y^2)(X_{31}^2 + X_{32}^2)}{2(1-x^2-y^2)} \\ &= \frac{(X_{31}^2 + X_{32}^2)(1-2(x^2+y^2))}{2(1-x^2-y^2)} \geq 0 = \hat{f}(X^*, e_3), \quad \text{for all } (x,y) \text{ such that } x^2+y^2 \leq 1/2. \end{aligned} \tag{2.83}$$

where for the first inequality we neglected all the other obviously positive summands, and for the second inequality we factored a common denominator and used Cauchy-Schwarz to conclude that $(xX_{31} + yX_{32})^2 \leq (x^2+y^2)(X_{31}^2 + X_{32}^2)$. This shows (X^*, e_3) is a local minimum.

On the other hand, let $X_\epsilon = \mathrm{diag}(1,0,\epsilon) \in \mathcal{M}_{\leq 2}^{3\times 3}$ for $\epsilon > 0$. Then $||X^* - X_\epsilon|| = \epsilon$ which can be made arbitrarily small, but $f(X_\epsilon) = -\epsilon^2/2 < 0 = f(X^*)$, demonstrating that X_ϵ is not a local minimum. Of course, it is not a local maximum either, e.g. $f(\mathrm{diag}(1,\epsilon,0)) = \epsilon^2/2 > 0 = f(X^*)$. Thus, X^* is a saddle point on $\mathcal{M}_{\leq 2}^{3\times 3}$. □

If one is bothered by the fact that f is not bounded from below, consider $g(X) = f(X) + \frac{1}{2}X_{33}^4$. Then $g(X_\epsilon) = (-\epsilon^2 + \epsilon^4)/2 < 0 = g(X^*)$ for all small ϵ so X^* is still a saddle on $\mathcal{M}_{\leq 2}^{3\times 3}$. Also,

$$g_\lambda(X, Y) = \hat{f}(X, Y) + \frac{1}{2}X_{33}^4 \geq 0 = g_\lambda(X^*, e_3), \tag{2.84}$$

as above, so (X^*, e_3) is a local minimum on the desingularied variety. We also have

$$f(X) \geq \frac{-X_{33}^2 + X_{33}^4}{2} \geq -\frac{1}{8}, \tag{2.85}$$

so f is bounded from below.

Incidentally, note that (X^*, e_2) is a saddle on the desingularization, since (X_ϵ, e_2) is a valid point on the desingularization. We show in Prop. 2.39 below that this is generally true for all three lifts we consider—if *all* points in the inverse image of a point are local minima on the lift, then so is the point itself.

Note also that no convex counterexample exists:

Proposition 2.38. If the cost function f extends to a convex function on $\mathbb{R}^{m\times n}$, then a point $X \in \mathcal{M}_{\leq k}$ is a local minimum if and only if there exists $(X, [Y]) \in \mathcal{M}_k^{(X,Y)}$ that is a local minimum.

Proof. If $\text{rank}(X) = k$, this follows because the projection map $\mathcal{M}_k^{(X,Y)} \to \mathcal{M}_{\leq k}$ is a diffeomorphism when restricted to matrices of rank k, see Sec. 2.2. If $\text{rank}(X) < k$ and $(X, [Y])$ is a local minimum on $\mathcal{M}_k^{(X,Y)}$, then it is in particular 2-critical on $\mathcal{M}_k^{(X,Y)}$. By Prop. 2.24 X is 1-critical on $\mathcal{M}_{\leq k}$. Since $\text{rank}(X) < k$, this implies $\nabla f(X) = 0$. By convexity of f, this means X is a global minimum for f on all of $\mathbb{R}^{m\times n}$, so $(X, [Y])$ is a global minimum as well. □

Regularization

All of the three lifts we considered can contain local minima that correspond to saddles on the original variety. However, for the lifts we have considered, $X \in \mathcal{M}_{\leq k}$ cannot be a saddle point if *all* points in the preimage of X on the lift are local minima:

Proposition 2.39. Suppose $X \in \mathcal{M}_{\leq k}$ is not a local minimum. Then for each of the three lifts $\mathcal{M}_k^{(L,R)}$, $\mathcal{M}_k^{(U,W)}$, $\mathcal{M}_k^{(X,Y)}$, there is a point in the preimage of X on the lift that is not a local minimum.

Proof. Suppose X is not a local minimum on $\mathcal{M}_{\leq k}$, so there is a sequence $(X_n) \subset \mathcal{M}_{\leq k}$ such that $X_n \to X$ and $f(X_n) < f(X)$. Then:

- Let $X_n = U_n\Sigma_n V_n$ be the thin SVD of X_n with $U_n \in \text{St}(m, k)$, $\Sigma_n = \text{diag}(\sigma_1, \ldots, \sigma_k) \succeq 0$ and $V_n \in \text{St}(n, k)$. Let $L_n = U_n\sqrt{\Sigma_n}$ and $R_n = V_n\sqrt{\Sigma_n}$. Then $||L_n|| = ||R_n|| = \sqrt{||X_n||}$ which is bounded,

hence after passing to a subsequence we may assume $L_n \to L$ and $R_n \to R$ such that $X = LR^T$ by continuity. Since we have $g(L_n, R_n) = f(X_n) < f(X) = g(L, R)$ for all n, the pair (L, R) is a point in the preimage of X which is not a local minimum.

- In the same setup as above, let $W_n = V_n\Sigma_n$ so again after passing to a subsequence we have $(U_n, W_n) \to (U, W)$ such that $X = UW^T$. Hence $(U, W) \in \overline{\mathcal{M}}_k^{(U,W)}$ is a point in the preimage of X that is not a local minimum. By Prop. 2.1, the point $[(U, W)] \in \mathcal{M}_k^{(U,W)}$ is also not a local minimum.
- Let $(X_n, Y_n) \in \overline{\mathcal{M}}_k^{(X,Y)}$ where $Y_n \in \mathrm{St}(m, k)$ is any Stiefel matrix satisfying $X_nY_n = 0$. Since $\mathrm{St}(m, k)$ is compact and the sequence (X_n) is bounded, we can again pass to a convergent subsequence $(X_n, Y_n) \to (X, Y)$ whose limit is then not a local minumum. By Prop. 2.1, the point $(X, [Y]) \in \mathcal{M}_k^{(X,Y)}$ is then not a local minimum either.

□

Thus, in all three cases there is some point in the preimage of a saddle which is a saddle on the lift. To exploit this algorithmically, we need to characterize the points on the lift which cannot be 'false local minima' and ensure that we only converge to such points. For the (L, R) factorization, Prop. 2.34 shows that any 'balanced' factorization satisfying $\mathrm{rank}(L) = \mathrm{rank}(R)$ is a local minimum on the lift if and only if it is a factorization of a local minimum on the original variety. As we have seen above, we cannot simply restrict to balanced factorizations because they do not form a smooth manifold. However, by regularizing our cost function we can ensure that any 1-critical point will be a balanced factorization.

We can regularize our cost function to promote balanced factorizations by setting

$$g_\lambda(L, R) = f(LR^T) + \frac{\lambda}{8}||L^TL - R^TR||_F^4. \tag{2.86}$$

Using an algorithm such as trust-regions (see [1, Chap. 7], [3, Chap. 6]), we can obtain a 2-critical point for g_λ. We argue that such a point is 2-critical for $g(L, R) = f(LR^T)$ as well, and that if such a point is a local minimizer for g_λ on $\mathcal{M}_k^{(L,R)}$ then $X = LR^T \in \mathcal{M}_{\leq k}$ is a local minimizer for f.

The following argument is essentially taken from [29, Thm. 3]. The only difference is that we took the penalty $||L^TL - R^TR||_F$ to the fourth power, whereas in [29] they took the penalty to the second power. The reason we need to take the fourth power is to ensure that both the first and second derivative of $||L^TL - R^TR||_F^4$ vanish at points satisfying $L^TL = R^TR$, which is used in Prop. 2.42 to show that any 1- and 2-critical point for g_λ is also 1- and 2-critical, respectively, for g.

Proposition 2.40. Any 1-critical point for g_λ satisfies $L^TL = R^TR$ whenever $\lambda > 0$.

Proof. Note that $\nabla g_\lambda(L,R)=0$ if and only if

$$\begin{aligned}\nabla f(X)R+\lambda||L^TL-R^TR||_F^2L(L^TL-R^TR)=0,\\ \nabla f(X)^TL-\lambda||L^TL-R^TR||_F^2R(L^TL-R^TR)=0.\end{aligned}\tag{2.87}$$

Multiplying the first equation by L^T on the left, taking the tranpose of the second equation and multiplying it by R on the right, and subtracting the two resulting equations, we obtain

$$\begin{aligned}0&=\lambda||L^TL-R^TR||_F^2\Big[(L^TL)(L^TL-R^TR)+(L^TL-R^TR)(R^TR)\Big]\\ &=\lambda||L^TL-R^TR||_F^2\Big[(L^TL)^2-(R^TR)^2\Big].\end{aligned}\tag{2.88}$$

Thus, either $L^TL=R^TR$ or $(L^TL)^2=(R^TR)^2$. In the latter case, since L^TL is the unique positive semidefinite square root of $(L^TL)^2$ and similarly for R^TR and $(R^TR)^2$, we conclude that $L^TL=R^TR$ in the latter case as well. □

To prove that a local minimizer for g_λ is also a local minimizer for g, we need the following Lemma, strengthening Lemma 2.32 for factorizations (L,R) satisfying $L^TL=R^TR$:

Lemma 2.41. *Suppose (L_1,R_1) and (L_2,R_2) are two factorizations for X satisfying $L_i^TL_i=R_i^TR_i$ for $i=1,2$. Then there exists an orthogonal $Q\in O(k)$ satisfying $(L_1,R_1)=(L_2Q,R_2Q)$.*

Proof. From Lemma 2.32 and Lemma 2.33, there exists invertible $J\in\mathrm{GL}(k)$ satisfying $(L_1,R_1)=(L_2J,R_2J^{-T})$. Let $P=JJ^T\succ 0$ and $M=L_2^TL_2=R_2^TR_2$. Then the equation $L_1^TL_1=R_1^TR_1$ implies $M=PMP$. Letting $P=U\Lambda U^T$ be an eigendecomposition for P and $\bar{M}=U^TMU$, we then have $\bar{M}=\Lambda\bar{M}\Lambda$, or in terms of entries $\bar{M}_{i,j}=\lambda_i\lambda_j\bar{M}_{i,j}$. If we denote the columns of R_2U by $[r_1,\dots,r_k]$ then the $i=j$ equations say that $||r_i||^2=\lambda_i^2||r_i||^2$ and similarly for the columns of L_2U. Hence either $\lambda_i=1$ (because $\lambda_i^2=1$ and $\lambda_i>0$) or $r_i=0$. Thus, after permuting the eigendecomposition of P, we may assume that $R_2U=[r_1,\dots,r_p,0,\dots,0]$, $L_2U=[\ell_1,\dots,\ell_p,0,\dots,0]$, and $\Lambda=\mathrm{diag}(1,\dots,1,\lambda_{p+1},\dots,\lambda_k)$.

If $J=U\Sigma V^T$ is the SVD, then $\Sigma^2=\Lambda$ so the first p singular values of J (under this permutation) are all 1. Let $Q=UV^T\in O(k)$. Observe that

$$L_2J=L_2U\Sigma V^T=[\ell_1,\dots,\ell_p,0,\dots,0]\mathrm{diag}(1,\dots,1,\sigma_{p+1},\dots,\sigma_k)V^T=\sum_{i=1}^p\ell_iv_i^T=L_2Q,\tag{2.89}$$

and similarly $R_2J=R_2Q$. Thus, $(L_1,R_1)=(L_2Q,R_2Q)$ where $Q\in O(k)$. □

Proposition 2.42. Any 1-critical point for g_λ is also 1-critical for g. Any 2-critical point for g_λ is also

2-critical for g. A local minimizer for g_λ is also a local minimizer for f (hence a local minimizer for g as well).

Proof. Since at any 1-critical point for g_λ we have $L^TL = R^TR$, and since the first two derivatives of $||L^TL - R^TR||_F^4$ vanish at such points (this is why we needed to take the regularizer to the 4th power), the first two claims are immediate.

Suppose (L, R) is a local minimizer for g_λ but $X = LR^T$ is not a local minimizer for f. Since (L, R) is a local minimizer, it is in particular 1-critical so $L^TL = R^TR$. Next, since X is not a local minimizer for f there exists a sequence (X_n) such that $X_n \to X$ and $f(X_n) < f(X)$ for all n. As in the proof of Prop. 2.34, we can construct factorizations $(L_n, R_n) \to (\bar{L}, \bar{R})$ such that $X_n = L_nR_n^T$ and $L_n^TL_n = R_n^TR_n$ and by continuity $\bar{L}^T\bar{L} = \bar{R}^T\bar{R}$. By Lemma 2.41, there exists $Q \in O(k)$ satisfying $(L, R) = (\bar{L}Q, \bar{R}Q)$. Then $(L'_n, R'_n) = (L_nQ, R_nQ)$ are factorizations of X which satisfy $L_n'^TL'_n = R_n'^TR'_n$ and $(L'_n, R'_n) \to (L, R)$. Thus, we have $g_\lambda(L'_n, R'_n) = f(L'_nR_n'^T) = f(X_n) < f(X) = g_\lambda(L, R)$ and $(L'_n, R'_n) \to (L, R)$, showing that (L, R) is not a local minimum for g_λ. □

Summary

- The preimage of a local minimum on the variety is a local minimum on the lift.
- If the point has rank k, then if one its preimages on either one of the three lifts is a local minimum then so is the point itself on the variety.
- There are cost functions and points of rank $< k$ that are saddle points on the variety but that have a preimage that is a local minimum on the lifts. For the lifts we have considered, such cost functions cannot be convex.
- For any of the three lifts we considered, if a point on the variety is not a local minimum then there is a point in its preimage which is not a local minimum on the lift.
- For the factorization approach, if a 'balanced' factorization is a local minimum on the lift then it is the factorization of a local minimum on the variety.
- Any 1-critical point for the regularized cost function over the factorization gives a 'balanced' factorization. Any 1- and 2- critical points for the regularized cost are also 1- and 2-critical, respectively, for the non-regularized cost on the lift. Finally, if the point to which we converge is a local minimum on the lift then it is also a local minimum on the variety.

Thus, optimizing the regularized cost function using a second-order optimization algorithm is a good algorithmic solution to the problem posed in the first chapter.

Chapter 3

Optimizing directly over bounded-rank matrices

In this chapter, we optimize directly on the variety of bounded-rank matrices by attempting to generalize algorithms optimizing over smooth manifolds. Our efforts in this direction are unsatisfactory, but they illustrate the main difficulties that we and others in the literature encountered.

Throughout this chapter, $||\cdot||$ denotes the spectral 2-norm on matrices (i.e. the largest singular value) and $||\cdot||_F$ denotes the Frobenius norm.

3.1 Failure of gradient descent on bounded-rank matrices

In this section, we exhibit an example where the algorithm from [21] extending gradient descent to the variety of bounded-rank matrices fails to converge to a critical point. We then present a modification of the algorithm meant to prevent such failures in Sec. 3.3, but are able to prove only limited results in that direction.

We specialize the algorithm of [21] given in Alg. 3 to the variety $\mathcal{M}_{\leq k}$ by using the metric projection retraction $R_X(V) = P_k(X + V)$. This is the natural generalization of gradient descent to $\mathcal{M}_{\leq k}$. If the sequence generated by the algorithm has a rank-k cluster point, then that point is 1-critical for f on $\mathcal{M}_{\leq k}$ and convergence rates are available [21, Thm. 2.3]. However, if the cluster point has rank $< k$ then it need not be 1-critical, in which case we say that the algorithm fails. We proceed to give an example of this situation.

We construct a function $f\colon \mathcal{M}_{\leq 2}^{3\times 3} \to \mathbb{R}$ and initialization Y_0 such that Alg. 3 takes infinitely many steps

to converge to a singular saddle which is not 1-critical.

To do so, we begin by constructing a convex function $Q\colon \mathbb{R}^{2\times 2} \to \mathbb{R}$ that would take infinitely-many steps to minimize. Let

$$Q(X) = \frac{1}{2}||D(X - X^*)||_F^2, \quad D = \mathrm{diag}(1, 1/2), \quad X^* = \mathrm{diag}(1, 0). \tag{3.1}$$

With a constant step size η, GD iterates

$$X_{k+1} = X_k - \eta\nabla Q(X) = (I - \eta D^2)X_k - \eta D^2 X^*, \tag{3.2}$$

and applying this formula inductively we get

$$X_k - X^* = (I - \eta D^2)^k (X_0 - X^*). \tag{3.3}$$

Initializing from $X_0 = \mathrm{diag}(2, 1)$, we have $X_k = X^* + \mathrm{diag}((1-\eta)^k, (1-\eta/4)^k)$. Taking $\eta = 8/5$, we obtain the fastest rate of convergence of $(3/5)^k$. Nevertheless, this takes infinitely many steps to converge to the unique global minimizer X^*. Note also that $\mathrm{rank}(X_k) = \mathrm{rank}(\mathrm{diag}(1 + (-3/5)^k, (3/5)^k)) = 2$ for all $k \geq 0$, while $\mathrm{rank}(X^*) = 1$.

We now define our bad example. Let $f\colon \mathcal{M}_{\leq 2}^{3\times 3} \to \mathbb{R}$ be given by

$$f(Y) = Q(Y_{1:2,1:2}) - \frac{(Y_{3,3} + 1)^2}{2} + \frac{Y_{3,3}^4}{4}, \tag{3.4}$$

where $Y_{1:2,1:2}$ is the principal 2×2 minor of Y. Note that the unique global minimum for f is attained at $Y^* = \mathrm{blkdiag}(X^*, x_0) = \mathrm{diag}(1, 0, x_0) \in \mathcal{M}_2^{3\times 3}$ where $x_0 \approx 1.32$ is the unique global minimizer of the univariate function $x \mapsto -\frac{(x+1)^2}{2} + \frac{x^4}{4}$. In particular, f is bounded from below and Y^* is a smooth point on the variety $\mathcal{M}_{\leq k}$.

Note that any matrix of the form $Y = \mathrm{blkdiag}(X, 0)$ with $X \in \mathcal{M}_2^{2\times 2}$ is smooth on $\mathcal{M}_{\leq 2}^{3\times 3}$ (because $\mathrm{rank}(Y) = 2$) and its tangent space is

$$T_Y \mathcal{M}_{\leq 2}^{3\times 3} = \left\{ \begin{pmatrix} A & b \\ c^T & 0 \end{pmatrix} : A \in \mathbb{R}^{2\times 2},\ b, c \in \mathbb{R}^2 \right\}, \tag{3.5}$$

because $\mathrm{col}(Y) = \mathrm{row}(Y) = \mathrm{span}\{e_1, e_2\} \subset \mathbb{R}^3$. The orthogonal projector onto this subspace simply zeros out the $(3, 3)$ entry.

Initialize Riemannian Gradient Descent (RGD) from $Y_0 = \text{blkdiag}(X_0, 0) = \text{diag}(2, 1, 0)$. For any point of the form $Y = \text{blkdiag}(X, 0)$ with $\text{rank}(X) = 2$, note that

$$\nabla f(Y) = \begin{pmatrix} \nabla Q(X) & 0 \\ 0 & -1 \end{pmatrix}, \quad \Pi_Y \nabla f(Y) = \text{blkdiag}(\nabla Q(X), 0) \in T_Y \mathcal{M}_{\leq 2}^{3\times 3}, \tag{3.6}$$

where Π_Y denotes the orthogonal projection onto $T_Y \mathcal{M}_{\leq 2}^{3\times 3}$. The next iterate is then

$$Y_1 = \Pi_2(Y_0 - \eta \nabla f(X_0)) = \text{blkdiag}(X_0 - \eta \nabla Q(X_0), 0) = \text{blkdiag}(X_1, 0), \tag{3.7}$$

where we observe that the projection P_2 onto $\mathcal{M}_{\leq 2}^{3\times 3}$ is redundant because $\text{rank}(Y_0 - \eta \nabla f(X_0)) = \text{rank}(X_1) = 2$. Suppose that the kth iterate has the form $Y_k = \text{blkdiag}(X_k, 0)$. Using $\text{rank}(X_k) = 2$, the same argument applies inductively and we have $\Pi_{Y_k} \nabla f(Y_k) = \text{blkdiag}(\nabla Q(Y_k), 0)$ and $Y_{k+1} = \text{blkdiag}(X_{k+1}, 0)$. Thus, we conclude that after infinitely many steps, we have convergence $Y_k \to Y^{**} = \text{blkdiag}(X^*, 0)$.

However, Y^{**} is not even 1-critical for f on $\mathcal{M}_{\leq 2}^{3\times 3}$. Indeed, since $\text{rank}(Y^{**}) = 1 < 2$ we need to consider the projection of $\nabla f(Y^{**})$ onto the tangent *cone* at Y^{**}, which is given by

$$T_{Y^{**}} \mathcal{M}_{\leq 2}^{3\times 3} = \left\{ \begin{pmatrix} a & b \\ c & E \end{pmatrix} : a, b, c \in \mathbb{R},\ \text{rank}(E) = 1 \right\}. \tag{3.8}$$

Since $\nabla f(Y^{**}) = \text{blkdiag}(\mathbf{0}_{2\times 2}, -1) \in T_{Y^{**}} \mathcal{M}_{\leq 2}^{3\times 3}$, we have $\Pi_{Y^{**}} \nabla f(Y^{**}) = \nabla f(Y^{**}) \neq 0$. Thus, RGD generates an infinite sequence of points converging to a saddle point which is not even 1-critical (and which is a singular point on the variety $\mathcal{M}_{\leq k}$).

Note that the above step size also satisfies the sufficient decrease condition in Alg. 4, with $\tau = 1/5$. Indeed, if $Y = \text{blkdiag}(Y, 0)$, a simple computation shows

$$\begin{aligned} f(Y) - f(P_k(Y - \eta \Pi_Y \nabla f(Y))) = Q(X) - Q(X - \eta \nabla Q(X)) &= \frac{1}{2} \langle \Big[I - (I - \eta D)^2 \Big] D(X - X^*), D(X - X^*) \rangle \\ &\geq \frac{\lambda_{\min}(I - (I - \eta D)^2)}{2} ||D(X - X^*)||_F^2 = \tau \eta ||\nabla Q(X)||_F^2 = \tau \eta ||\Pi_Y \nabla f(Y)||_F^2, \end{aligned} \tag{3.9}$$

since $I - (I - \eta D)^2 = \text{diag}(16, 24)/25$. Thus, Alg. 3 fails with $\tau = 1/5$ and $t_0 = 8/5$.

Incidentally, note that we should also be able to make second-order optimization algorithms fail using

the same type of example. Indeed, the Riemannian Hessian of f at our iterates is

$$\text{Hess } f(Y_k)[Z] = \Pi_{Y_k} D\Big(\Pi_{Y_k} \nabla f(Y_k)\Big)[Z] = \text{blkdiag}(\nabla^2 Q(X_k)[Z'], 0), \quad \text{where } Z = \begin{pmatrix} Z' & b \\ c^T & 0 \end{pmatrix} \in T_{Y_k} \mathcal{M}_{\leq 2}^{3\times 3}. \tag{3.10}$$

Hence if we can find a function $Q\colon \mathbb{R}^{2\times 2} \to \mathbb{R}$ that takes infinitely-many steps to optimize using our second-order algorithm of choice, and such that all intermediate iterates have rank 2, then the same example as above applies.

Since Alg. 3 fails when we converge from a sequence of smooth points to a singular point on the variety, for the remainder of the chapter we focus on modifying Alg. 3 by projecting to the singular locus when the iterates get sufficiently close to it.

3.2 Results on metric projection to bounded-rank matrices

As we have seen in Sec. 1.2.3, an important property in analyzing RGD (Alg. 1) is the existence of a constant $L > 0$ satisfying

$$f(R_X(V)) \leq f(X) + \langle \nabla f(X), V \rangle + \frac{L}{2} ||V||_F^2. \tag{3.11}$$

Suppose f has a Lipschitz-continuous gradient in the ambient Euclidean space $\mathbb{R}^{m\times n}$, so there exists $L' > 0$ satisfying

$$f(R_X(V)) \leq f(X) + \langle \nabla f(X), R_X(V) - X \rangle + \frac{L'}{2} ||R_X(V) - X||_F^2. \tag{3.12}$$

We can rewrite Eq. (3.12) to the form of Eq. (3.11). Indeed, noting that $R_X(V) = \arg\min_{Y \in \mathcal{M}_{\leq k}} ||X + V - Y||_F$, we have

$$||R_X(V) - X||_F \leq ||R_X(V) - (X+V)||_F + ||V||_F \leq ||X - (X+V)||_F + ||V||_F = 2||V||_F. \tag{3.13}$$

Thus, Eq. (3.12) implies

$$f(R_X(V)) \leq f(X) + \langle \nabla f(X), V \rangle + ||\nabla f(X)||_F \cdot ||R_X(V) - (X+V)||_F + 2L'||V||_F^2. \tag{3.14}$$

Since $R_X(V) = X + V + O(||V||_F^2)$ by definition of a retraction at smooth points, we should be able to bound locally $||R_X(V) - (X+V)||_F \leq C||V||_F^2$ to obtain a constant $L = 4L' + 2C||\nabla f(X)||_F$ that satisfies Eq. (3.11). In this section, we therefore concern ourselves with bounding $||R_X(V) - (X+V)||_F$ by a term quadratic in $||V||_F$. Also, since it was shown in Sec. 1.3 that metric projection may not be single-valued

even locally near singularities even for $\mathcal{M}_{\leq k}$, we shall be interested in seeing when is the metric projection single-valued for $\mathcal{M}_{\leq k}$.

Our main result is:

Theorem 3.1. *Suppose $X \in \mathcal{M}_{\leq k}$ has rank$(X) = s \leq k$, and let $V \in T_X\mathcal{M}_{\leq k}$ be a tangent vector with rank$(\Pi_X^{\perp} V) = k - s - l$ for some $l \in \{0, 1, \ldots, k-s\}$. Then*

(1.) We have the estimates

$$\sigma_i(X + tV) = \begin{cases} \Theta(1) & i \leq s, \\ \Theta(t) & s < i \leq k - l \\ O(t^2) & k - l < i \leq r(m, n, k, s, l) \\ 0 & i > r(m, n, k, s, l) \end{cases}, \quad \text{as } t \to 0, \tag{3.15}$$

where σ_i is the ith largest singular value of a matrix and

$$r(m, n, k, s, l) = \min(k + s - l,\ m + k - s - l,\ n + k - s - l). \tag{3.16}$$

The constants in the O and Θ notations depend on $||V||$ and $\sigma_s(X)$.

(2.) we have

$$||R_X(V) - (X + V)|| \leq 26||X^{\dagger}|| \cdot ||\Pi_X V||^2, \quad \text{whenever } ||\Pi_X V|| \leq ||X^{\dagger}||^{-1}, \tag{3.17}$$

where $X^{\dagger}$ is the pseudo-inverse of X, and $||X^{\dagger}|| = 1/\sigma_s(X)$.

(3.) If $l = 0$ then $R_X(tV)$ is single-valued whenever $|t| < \frac{\sigma_s(X)\sigma_{k-s}(\Pi_X^{\perp} V)}{32||V||^2}$.

Note that when $m, n \geq 2s$, we have $r(m, n, k, s, l) = k + s - l$.

First, we shall need the following simple perturbation bounds. The first result is classical, the second is taken from [7, Thm. 3.1] but for completeness, we give a more direct proof below:

Lemma 3.2. *Let $A \in \mathbb{R}^{m \times n}$ be an arbitrary matrix, let $E \in \mathbb{R}^{m \times n}$ be arbitrary and $D_L \in \mathbb{R}^{m \times m}$ and $D_R \in \mathbb{R}^{n \times n}$ be invertible. Then*

$$\begin{aligned} \sigma_i(A) - ||E|| \leq & \sigma_i(A + E) \leq \sigma_i(A) + ||E|| \\ \frac{\sigma_i(A)}{||D_L^{-1}|| \cdot ||D_R^{-1}||} \leq & \sigma_i(D_L A D_R) \leq \sigma_i(A)||D_L|| \cdot ||D_R||. \end{aligned} \tag{3.18}$$

Proof. Write

$$\sigma_i(A) = \min_{\substack{U\subset\mathbb{R}^n \\ \dim(U)=n-i+1}} \max_{\substack{y\in U \\ ||y||\leq 1}} ||Ay||, \tag{3.19}$$

and suppose the min is attained at $\widetilde{U}$. Then

$$\sigma_i(A+E) \leq \max_{\substack{y\in\widetilde{U} \\ ||y||\leq 1}} ||(A+E)y|| \leq \max_{\substack{y\in\widetilde{U} \\ ||y||\leq 1}} ||Ay|| + \max_{\substack{y\in\widetilde{U} \\ ||y||\leq 1}} ||Ey|| \leq \sigma_i(A) + ||E||. \tag{3.20}$$

Therefore, we also have

$$\sigma_i(A) = \sigma_i(A+E-E) \leq \sigma_i(A+E) + ||-E|| = \sigma_i(A+E) + ||E||, \tag{3.21}$$

from which we get the first claim.

Since D_R is invertible, we have $\dim(D_R^{-1}\widetilde{U}) = \dim(\widetilde{U})$. Therefore,

$$\begin{aligned}
\sigma_i(D_L A D_R) &\leq \max_{\substack{y\in D_R^{-1}\widetilde{U} \\ ||y||\leq 1}} ||D_L A D_R y|| \leq ||D_L|| \max_{\substack{y\in D_R^{-1}\widetilde{U} \\ ||y||\leq 1}} ||A(D_R y)|| \\
&= ||D_L|| \max_{\substack{y\in\widetilde{U} \\ ||y||\leq ||D_R||}} ||Ay|| = ||D_L|| \cdot ||D_R|| \max_{\substack{y\in\widetilde{U} \\ ||y||\leq 1}} ||Ay|| \\
&= ||D_L|| \cdot ||D_R|| \sigma_i(A),
\end{aligned} \tag{3.22}$$

where in going from the first to the second line we changed variables $y' = D_R y$, noted that $y \in D_R^{-1}\widetilde{U}$ iff $y' \in \widetilde{U}$, and that $||y'|| \leq ||D_R||$ whenever $||y|| \leq 1$. Therefore, we also have

$$\sigma_i(A) = \sigma_i(D_L^{-1} D_L A D_R D_R^{-1}) \leq ||D_L^{-1}|| \cdot ||D_R^{-1}|| \sigma_i(D_L A D_R), \tag{3.23}$$

from which we obtain the second claim. □

We are now ready to prove the Prop.:

Proof. (Thm. 3.1) In the basis of the singular vectors of X, write

$$\begin{aligned}
&X = \begin{pmatrix} \Sigma_X & 0 \\ 0 & 0 \end{pmatrix}, \quad V = \begin{pmatrix} V_1 & V_2 \\ V_3 & V_4 \end{pmatrix}, \quad \Pi_X V = \begin{pmatrix} V_1 & V_2 \\ V_3 & 0 \end{pmatrix}, \\
&\Sigma_X = \mathrm{diag}(\sigma_1(X), \ldots, \sigma_s(X)) \in \mathbb{R}^{s\times s}, \\
&V_1 \in \mathbb{R}^{s\times s},\ V_2 \in \mathbb{R}^{s\times(n-s)},\ V_3 \in \mathbb{R}^{(m-s)\times s},\ V_4 \in \mathbb{R}^{(m-s)\times(n-s)}, \\
&\mathrm{rank}(V_4) = k - s - l.
\end{aligned} \tag{3.24}$$

We shall assume without loss of generality that $t > 0$ (otherwise, replace $V \mapsto -V$). From the additive perturbation bound in Lemma 3.2, we have

$$\sigma_i(X + tV) \in \sigma_i(X) \pm t||V||, \quad i = 1, \ldots, s, \tag{3.25}$$

which shows $\sigma_i(X + tV) = \Theta(1)$ for $i = 1, \ldots, s$ since $\sigma_i(X) \neq 0$ for these i (and the notation $a \in b \pm c$ means $a \in [b - c, b + c]$).

From the same additive perturbation bound, we also have $\sigma_s(\Sigma_X + tV_1) \geq \sigma_s(\Sigma_X) - ||tV_1|| \geq \sigma_s(X) - ||t\Pi_X V||$. Therefore, as long as $||t\Pi_X V|| \leq \sigma_s(X)/2$ we have $\Sigma_X + tV_1$ is invertible and

$$||(\Sigma_X + tV_1)^{-1}|| = \frac{1}{\sigma_s(\Sigma_X + tV_1)} \leq \frac{2}{\sigma_s(X)}. \tag{3.26}$$

Next, write

$$X + tV = D_L(t)\widetilde{X}(t)D_R(t), \tag{3.27}$$

where

$$\begin{aligned}
D_L(t) &= \begin{pmatrix} I_s & \mathbf{0}_{s\times(m-s)} \\ tV_3(\Sigma_X + tV_1)^{-1} & I_{m-s} \end{pmatrix} \\
\widetilde{X}(t) &= \begin{pmatrix} \Sigma_X + tV_1 & \mathbf{0}_{s\times(n-s)} \\ \mathbf{0}_{(m-s)\times s} & tV_4 - t^2V_3(\Sigma_X + tV_1)^{-1}V_2 \end{pmatrix} \\
D_R(t) &= \begin{pmatrix} I_s & t(\Sigma_X + tV_1)^{-1}V_2 \\ \mathbf{0}_{(n-s)\times s} & I_{n-s} \end{pmatrix}.
\end{aligned} \tag{3.28}$$

(This block-LDU decomposition using the Schur complement is only stated for square matrices in a number of references, but note that it applies for rectangular matrices as well. Also, when the sizes of zero blocks are clear we shall omit the bold font and subscripts above and simply write 0.) Note that $D_L(t), D_R(t)$ are

invertible, $\text{rank}(\Sigma_X + tV_1) = s$ for the values of t considered above, and

$$\begin{aligned}\text{rank}(tV_4 - t^2V_3(\Sigma_X + tV_1)^{-1}V_2) &\leq \text{rank}(V_4) + \text{rank}(V_3(\Sigma_X + tV_1)^{-1}V_2)) \\ &\leq k - s - l + \min(m - s, s, n - s) \\ &= \min(m + k - 2s - l, k - l, n + k - l - 2s).\end{aligned} \tag{3.29}$$

Therefore,

$$\text{rank}(X + tV) \leq s + \min(m + k - 2s - l, k - l, n + k - l - 2s) = r(m, n, k, s, l). \tag{3.30}$$

Hence $\sigma_i(X + tV) = 0$ for $i > r(m, n, k, s, l)$.

Since

$$||V_3(\Sigma_X + tV_1)^{-1}V_2|| \leq ||V_3|| \cdot ||V_2|| \cdot ||(\Sigma_X + tV_1)^{-1}|| \leq \frac{2||\Pi_X V||^2}{\sigma_s(X)}, \tag{3.31}$$

the additive perturbation bound in Lemma 3.2 gives

$$\sigma_i(tV_4 - t^2V_3(\Sigma_X + tV_1)^{-1}V_2) \in t\sigma_i(V_4) \pm \frac{2t^2||\Pi_X V||^2}{\sigma_s(X)} \tag{3.32}$$

Note that because $\widetilde{X}(t)$ is block-diagonal, its singular values are the union of the singular values of $\Sigma_X + tV_1$ (which are $\Theta(1)$) and the singular values of $tV_4 - t^2V_3(\Sigma_X + tV_1)^{-1}V_2$ (which the above shows are $O(t)$). Choosing t small enough so that

$$t||V|| + \frac{2t^2||V||^2}{\sigma_s(X)} \leq \sigma_s(X) - t||V||, \tag{3.33}$$

or equivalently

$$t \leq \frac{\sqrt{3} - 1}{2} \frac{\sigma_s(X)}{||V||}, \tag{3.34}$$

we then guarantee that the singular values of the lower block are smaller than all the singular values of the upper block of $\widetilde{X}(t)$. Since the top block gives s singular values for $\widetilde{X}(t)$, we conclude that the bottom $k - l$ singular values of $\widetilde{X}(t)$ come from the lower block. Next, observe that

$$D_L(t) - I_m = \begin{pmatrix} 0 & 0 \\ tV_3(\Sigma_X + tV_1)^{-1} & 0, \end{pmatrix}, \tag{3.35}$$

so $||D_L(t) - I|| = t||V_3(\Sigma_X + tV_1)^{-1}||$, and hence $||D_L(t)|| \leq 1 + \frac{2t||\Pi_X V||}{\sigma_s(X)}$ and similarly $||D_R(t)|| \leq 1 + \frac{2t||\Pi_X V||}{\sigma_s(X)}$.

Furthermore, observe that

$$D_L(t)^{-1} = \begin{pmatrix} I_s & \mathbf{0}_{s\times(m-s)} \\ -tV_3(\Sigma_X + tV_1)^{-1} & I_{m-s} \end{pmatrix}, \tag{3.36}$$

and similarly for $D_R(t)^{-1}$, so the same argument shows $||D_L(t)^{-1}||, ||D_R(t)^{-1}|| \leq 1 + \frac{2t||\Pi_X V||}{\sigma_s(X)}$. From the multiplicative perturbation bound in Lemma 3.2, we get

$$\begin{aligned} \sigma_i(X+tV) &\leq \sigma_{i-s}(tV_4 - t^2V_3(\Sigma_X+tV_1)^{-1}V_2)\left(1+\frac{2t||\Pi_X V||}{\sigma_s(X)}\right)^2 \\ &\leq \left(t\sigma_{i-s}(V_4) + \frac{2t^2||\Pi_X V||^2}{\sigma_s(X)}\right)\left(1+\frac{2t||\Pi_X V||}{\sigma_s(X)}\right)^2 \\ &\leq t\sigma_{i-s}(V_4) + \frac{26t^2||\Pi_X V||^2}{\sigma_s(X)}, \end{aligned} \tag{3.37}$$

for all $i > s$ and whenever $t \leq \sigma_s(X)/2||\Pi_X V||$. In particular, since $\text{rank}(V_4) = k-s-l$, we have

$$\sigma_i(X+tV) \leq \frac{26}{\sigma_s(X)}t^2||\Pi_X V||^2 = O(t^2), \quad i > k-l, \tag{3.38}$$

proving another part of the claim. Note moreover that to have the above bound on $\sigma_i(X+tV)$ for $i > k-l$, we only need the $(k-l+1)$th singular value of $X+tV$ to be the $(k-s-l+1)$th singular value of $tV_4 - t^2V_3(\Sigma_X + tV_1)^{-1}V_2$. For that, we need to require the $(k-s-l+1)$th singular value of $tV_4 - t^2V_3(\Sigma_X+tV_1)^{-1}V_2$ to be smaller than the sth singular value of $\Sigma_X + tV_1$. A sufficient condition for that using our perturbation bounds is

$$\frac{2t^2||\Pi_X V||^2}{\sigma_s(X)} \leq \sigma_s(X) - t||\Pi_X V||, \text{ or equivalently, } t \leq \frac{\sigma_s(X)}{||\Pi_X V||}. \tag{3.39}$$

Finally, we also have the lower bound

$$\begin{aligned} \sigma_i(X+tV) &\geq \sigma_{i-s}(tV_4 - t^2V_3(\Sigma_X+tV_1)^{-1}V_2)\left(1+\frac{2t||\Pi_X V||}{\sigma_s(X)}\right)^{-2} \\ &\geq \left(t\sigma_{i-s}(V_4) - \frac{2t^2||\Pi_X V||^2}{\sigma_s(X)}\right)\left(1-\frac{4t||\Pi_X V||}{\sigma_s(X)}\right) \\ &\geq t\sigma_{i-s}(V_4) - \frac{6t^2||\Pi_X V||^2}{\sigma_s(X)}. \end{aligned} \tag{3.40}$$

Thus, for $s+1 \leq i \leq k-l$ we have $\sigma_i(X+tV) = \Theta(t)$ as claimed. Finally, if $l = 0$ then we can guarantee $\sigma_k(X+tV) > \sigma_{k+1}(X+tV)$ (which is equivalent to the retraction $R_X(tV)$ being single-valued) by requiring

$$t\sigma_{k-s}(V_4) - \frac{6t^2||V||^2}{\sigma_s(X)} > \frac{26t^2||V||^2}{\sigma_s(X)}, \tag{3.41}$$

or equivalently

$$t < \frac{\sigma_s(X)\sigma_{k-s}(V_4)}{32||V||^2}. \tag{3.42}$$

□

Note that the dependence of the bound in Thm. 3.1(2) on $1/\sigma_s(X)$ is sharp up to constants in general. For example, if $m, n \geq k + s \geq 2s$ and in the basis of the singular vectors of X we set

$$V_1 = \mathbf{0}_{s\times s},\ V_2 = \begin{pmatrix} \mathbf{0}_{s\times(n-2s)} & -I_s \end{pmatrix},\ V_3 = \begin{pmatrix} \mathbf{0}_{(m-2s)\times s} \\ I_s \end{pmatrix},\ V_4 = \begin{pmatrix} I_{k-s} & \mathbf{0}_{(k-s)\times(n-k)} \\ \mathbf{0}_{(m-k)\times(k-s)} & \mathbf{0}_{(m-k)\times(n-k)} \end{pmatrix}, \tag{3.43}$$

then

$$tV_4 - t^2V_3(\Sigma_X + tV_1)^{-1}V_2 = \begin{pmatrix} tI_{k-s} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & t^2\Sigma_X^{-1} \end{pmatrix}, \tag{3.44}$$

whose $(k - s + 1)$th singular value is $t^2/\sigma_s(X)$ as long as $t \leq \sigma_s(X)$. Also, one can check that $V^TV = \text{blkdiag}(I_k, \mathbf{0}_{n-k-s}, I_s)$ so $||V|| = 1$. In that case, the multiplicative perturbation bound gives

$$\sigma_{k+1}(X + tV) \geq \frac{t^2}{\sigma_s(X)}\left(1 - \frac{4t}{\sigma_s(X)}\right) \geq \frac{t^2}{2\sigma_s(X)}, \tag{3.45}$$

whenever $t \leq \sigma_s(X)/8$. Thus, on the line $\{tV : t \in \mathbb{R}\} \subset T_X\mathcal{M}_{\leq k}$, we have

$$||R_X(tV) - (X + tV)|| \geq \frac{||tV||^2}{2\sigma_s(X)}. \tag{3.46}$$

We are ready to bound the difference between the retraction $R_X(V)$ and its first order approximation $X + V$:

Corollary 3.3. *We have*

$$||R_X(V) - (X + V)|| \leq 26||X^\dagger|| \cdot ||\Pi_X V||^2, \tag{3.47}$$

for all $V \in T_X\mathcal{M}_{\leq k}$. Using the Frobenius norm, we also have

$$||R_X(V) - (X + V)||_F \leq 26 \cdot \textit{rank}(X) \cdot ||X^\dagger|| \cdot ||\Pi_X V||_F^2, \tag{3.48}$$

for all $V \in T_X\mathcal{M}_{\leq k}$.

Proof. Note that $X + \Pi_X^\perp V \in \mathcal{M}_{\leq k}$ because $\text{rank}(X + \Pi_X^\perp V) = k - l \leq k$. By definition of metric projection

(the SVD truncation is also the projection in the 2-norm by Eckart–Young–Mirsky), we then have

$$||R_X(V) - (X+V)|| \leq ||X + \Pi_X^{\perp} V - (X+V)|| = ||\Pi_X V||. \tag{3.49}$$

Therefore, if $||\Pi_X V|| \geq \sigma_s(X)/26$, we have

$$||R_X(V) - (X+V)|| \leq ||\Pi_X V|| \leq \frac{26}{\sigma_s(X)} ||\Pi_X V||^2. \tag{3.50}$$

Since the above bound also holds for $||\Pi_X V|| \leq \sigma_s(X)$, we conclude that it holds for all $V \in T_X \mathcal{M}_{\leq k}$, giving the first estimate.

For the second estimate, note that $\operatorname{rank}(X+V) \leq k+s$ as shown above, while $R_X(V)$ is the projection of $X+V$ onto $\mathcal{M}_{\leq k}$. Therefore, the map $X+V \mapsto R_X(V)$ sets at most $s = \operatorname{rank}(X)$ of the bottom singular values of $X+V$ to zero. Therefore,

$$||R_X(V) - (X+V)||_F = \sqrt{\sum_{i=k+1}^{k+s} \sigma_i(X+V)^2} \leq s\sigma_{k+1}(X+V) \leq 26s||X^{\dagger}|| \cdot ||\Pi_X V||^2 \leq 26s||X^{\dagger}|| \cdot ||\Pi_X V||_F^2, \tag{3.51}$$

giving the second estimate. □

The fact that the error between the retraction and its 1st-order approximation becomes unbounded as one approaches a low rank point prevents us from controlling the step size chosen by backtracking. In turn, this prevents us from guaranteeing convergence.

3.3 Gradient descent with projection to singular locus

As the example in Sec. 3.1 shows, simply extending gradient descent to singular points on the variety $\mathcal{M}_{\leq k}^{m \times n}$ as done in Alg. 3 does not guarantee convergence to 1-critical points. However, as shown in [21], this problem only arises if the sequence generated by the algorithm converges to a matrix of rank $< k$. Similarly, as shown in Cor. 3.3, the norm (both spectral and Frobenius) of the difference between the retraction and its first order estimate increases only when the smallest nonzero singular value of the iterates goes to zero. This naturally suggests a modification of the algorithm whereby we truncate singular values below a certain threshold. In this section, we analyze such an algorithm.

Denote by $\Pi_X^s \colon \mathbb{R}^{m \times n} \to T_X \mathcal{M}_{\operatorname{rank}(X)}$ the orthogonal projection to the tangent space at X to its own stratum. Also let $T_\delta \colon \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ be truncation of the singular values of a matrix at δ, i.e. $T_\delta(X)$ has

the same singular vectors as X but

$$\sigma_i(T_\delta(X)) = \begin{cases} \sigma_i(X) & \text{if } \sigma_i(X) \geq \delta \\ 0 & \text{if } \sigma_i(X) < \delta \end{cases} \tag{3.52}$$

The issue with introducing a truncation step to our algorithm is that the cost function may increase due to the truncation. To ensure that the iterates remain within some fixed sublevel set of the function, we therefore require the decrease in the cost when we increase the rank of the iterates to compensate for the potential increase in cost due to truncations. Specifically, consider Alg. 5 (where backtrack refers to Alg. 4).

Algorithm 5 Modified Riemannian GD

procedure MRGD($f, \overline{X}_{0,0}, k, \epsilon, \delta, \tau, \beta, t_0$) ▷ Function f, initial guess $\overline{X}_{0,0}$, rank bound k, gradient norm cutoff ϵ, SVD cutoff δ, linesearch parameters τ, β, t_0.
  for $i = 1, 2, \ldots$ **do** ▷ epoch number
    $\overline{X}_{i,0} \leftarrow \overline{X}_{i-1,\text{end}}$
    $X_{i,0} \leftarrow T_\delta(\overline{X}_{i,0})$
    for $j = 0, 1, 2, \ldots$ **do**
      $V_{i,j} \leftarrow \Pi^s_{X_{i,j}}\big[-\nabla f(X_{i,j})\big]$ ▷ descent direction
      if $||V_{i,j}||_F \leq \epsilon$ **then**
        break
      end if
      $r_{i,j} \leftarrow \text{rank}(X_{i,j})$
      $t_{i,j} \leftarrow \text{backtrack}(f, X_{i,j}, V_{i,j}, r_{i,j}, \tau, \beta, t_0)$ ▷ stepsize
      $\overline{X}_{i,j+1} \leftarrow P_{r_{i,j}}(X_{i,j} + t_{i,j} V_{i,j})$ ▷ move
      $X_{i,j+1} \leftarrow T_\delta(\overline{X}_{i,j+1})$ ▷ project down if needed
    end for
    $W_i \leftarrow (I - \Pi^s_{X_{i,\text{end}}}) \circ \Pi_{X_{i,\text{end}}}\big[-\nabla f(X_{i,\text{end}})\big]$ ▷ direction for rank increase
    $t_{i,\text{end}+1} \leftarrow \text{backtrack}(f, X_{i,\text{end}}, W_i, k, \tau, \beta, t_0)$
    $\overline{X}_{i,\text{end}+1} \leftarrow X_{i,\text{end}} + t_{i,\text{end}} W_i$
    if $f(X_{i,\text{end}}) - f(\overline{X}_{i,\text{end}}) < \epsilon^2$ **then**
      return $X_{i,\text{end}}$
    end if
  end for
end procedure

In words: For each i, we optimize over a given stratum until either the projection of the gradient to the tangent space of that stratum or the smallest nonzero singular value of the iterate become sufficiently small. If some singular values become smaller than δ, we truncate them and project onto a lower stratum. We break out of the j-loop only when the projection of the gradient to the tangent space becomes sufficiently small, at which point we attempt a rank increase. We increase the rank only if the decrease in the cost function due to the rank increase compensates for the potential increase in the cost function incurred when we project down (see below).

3.3.1 Assumptions and Analysis

We shall make the following assumptions:

Assumption 0. f is $\mathcal{C}^2$.

Assumption 1. The function f is Lipschitz continuous in the ambient Euclidean space. Specifically, there exists a constant $L_f > 0$ satisfying

$$|f(X) - f(Y)| \leq L_f ||X - Y||. \tag{3.53}$$

Note that this assumption is quite restrictive. For example, quadratics are not Lipschitz continuous on the entire Euclidean space. We need this assumption to bound the potential increase in the cost function incurred when we truncate singular values and ensure that we remain in some sublevel set, see Lemma 3.5.

Assumption 2. The sublevel set $S_0 = \{X : f(X) \leq f(\overline{X}_{0,0}) + k\delta L_f\}$ is compact.

Assumption 3. The initial stepsize rule is bounded away from zero: $t_0(X) \geq t_0 > 0$ for all $X \in \mathcal{M}_{\leq k}$.

Assumption 4. ϵ, δ are chosen to satisfy

$$k\delta L_f < \epsilon^2. \tag{3.54}$$

We turn to analyzing Alg. 5. There are three things we need to verify for the algorithm to be well defined: that a stepsize giving sufficient decrease exists, that the inner loop (over j) terminates in finitely many steps for all i, and that $\operatorname{rank}(\overline{X}_{i,j+1}) \leq k$ in line 17.

The last point is easy to justify. In the basis of the singular vectors of $X_{i,j}$, we have

$$\begin{aligned} X_{i,j} &= \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix}, \quad W_i = \begin{pmatrix} 0 & 0 \\ 0 & P_{k-s}(W) \end{pmatrix}, \\ \Sigma &= \operatorname{diag}(\sigma_1(X_{i,j}), \ldots, \sigma_s(X_{i,j})) \in \mathbb{R}^{s \times s}, \\ W &\in \mathbb{R}^{(m-s)\times(n-s)}, \end{aligned} \tag{3.55}$$

where $s = \operatorname{rank}(X_{i,j})$ and W is some matrix depending on $\nabla f(X_{i,j})$. Adding the two together yields a block diagonal matrix with matrices of rank s and at most $k - s$ on the diagonal, with a total rank $\leq k$. Hence $P_k(X_{i,j} + t_{i,j}W_i) = X_{i,j} + t_{i,j}W_i$, which simplifies line 17.

We continue by arguing that the stepsizes $t_{i,j}$ exist and are uniformly bounded away from zero. Let

$S = \operatorname{conv} S_0$ which is also compact. Define

$$G = \sup_{X \in S_0} ||\nabla f(X)||_F, \ H = \sup_{X \in S} ||\nabla^2 f(X)||_{op}, \tag{3.56}$$

where we view $\nabla^2 f$ as a linear operator on $\mathbb{R}^{m \times n}$ and take the operator norm with respect to the Frobenius norm on $\mathbb{R}^{m \times n}$. Both G and H are finite because S_0, S are compact. Define

$$L \coloneqq \frac{26kG}{\delta} + 2H. \tag{3.57}$$

Then for $X \in \mathcal{M}_{\leq k}$ with $\operatorname{rank}(X) = s$ such that $||X^\dagger|| \leq \delta^{-1}$ (or equivalently, $X = T_\delta(X)$) and $V \in T_X \mathcal{M}_{\leq k}$, there exists Ξ on the line segment joining X and $P_s(X+V)$ satisfying

$$\begin{aligned} f(P_s(X+V)) &= f(X) + \langle \nabla f(X), V \rangle + \langle \nabla f(X), P_s(X+V) - (X+V) \rangle \\ &\quad + \frac{1}{2} \langle \nabla^2 f(\Xi)[P_s(X+V) - X)], P_s(X+V) - X \rangle \\ &\leq f(X) + \langle \nabla f(X), V \rangle + ||\nabla f(X)||_F \cdot 26 \cdot \operatorname{rank}(X) ||X^\dagger|| \cdot ||V||_F^2 \\ &\quad + \frac{1}{2} ||\nabla^2 f(\Xi)||_{op} ||P_s(X+V) - X||_F^2 \\ &\leq f(X) + \langle \nabla f(X), V \rangle + \frac{26kG}{\delta} ||V||_F^2 + \frac{1}{2} H \cdot 4 ||V||_F^2 \\ &= f(X) + \langle f(X), V \rangle + L||V||_F^2. \end{aligned} \tag{3.58}$$

We can now prove (recall that $V_{i,j} = \Pi^s_{X_{i,j}}[-\nabla f(X_{i,j})]$):

Lemma 3.4. *Stepsizes $t_{i,j}$ giving sufficient decrease exist and are bounded away from zero by*

$$t_{i,j} \geq T \coloneqq \min\left(t_0, \frac{\beta(1-\tau)}{L}\right) > 0, \tag{3.59}$$

for all i, j.

Proof. A stepsize satisfying sufficient decrease exists by [21, Prop. 2.8]. Observe that the smallest singular value of all $X_{i,j}$ are $\geq \delta$, so $||X_{i,j}^\dagger|| \leq \delta^{-1}$ for all i, j. Therefore, for $t \geq 0$

$$f(P_{r_{i,j}}(X_{i,j} + tV_{i,j})) - f(X_{i,j}) \leq t\langle \nabla f(X_{i,j}), V_{i,j} \rangle + Lt^2 ||V_{i,j}||_F^2 = (-t + t^2 L)||V_{i,j}||_F^2. \tag{3.60}$$

If t does not satisfy sufficient decrease, we have

$$(t - t^2 L)||V_{i,j}||_F^2 \leq f(X_{i,j}) - f(P_{r_{i,j}}(X_{i,j} + tV_{i,j})) < \tau t ||V_{i,j}||_F^2, \tag{3.61}$$

hence

$$t > \frac{1-\tau}{L}. \tag{3.62}$$

By Alg. 4, $t_{i,j}/\beta$ does not satisfy sufficient decrease. Therefore,

$$t_{i,j} \geq \frac{\beta(1-\tau)}{L}. \tag{3.63}$$

Of course, it is possible that the initial stepsize $t_0(X) \geq t_0$ already gives sufficient decrease and is smaller than the above bound, hence the min in the definition of T. □

Next, we show that the iterates never leave the sublevel set S_0:

Lemma 3.5. *$X_{i,j}, \overline{X}_{i,j} \in S_0$ for all i, j (i.e. the iterates never leave the sublevel set S_0).*

Proof. We show inductively that $f(\overline{X}_{i,0}) \leq f(\overline{X}_{0,0})$ and $f(X_{i,j}), f(\overline{X}_{i,j}) \leq f(\overline{X}_{0,0}) + k\delta L_f$ for $j \geq 1$.

Suppose $f(\overline{X}_{i,0}) \leq f(\overline{X}_{0,0})$ for some i (for $i = 0$, this is trivial). Note that if $\overline{X}_{i,j} \neq X_{i,j}$ for some j then $\overline{X}_{i,j}$ has some singular values that are $\leq \delta$ and that are being truncated. Therefore, every time we project down in rank the cost can increase by at most

$$f(X_{i,j}) - f(\overline{X}_{i,j}) \leq L_f ||X_{i,j} - \overline{X}_{i,j}|| \leq L_f \delta. \tag{3.64}$$

In particular, $f(X_{i,0}) \leq f(\overline{X}_{i,0}) + \delta L_f \leq f(\overline{X}_{0,0}) + \delta L_f$.

For any J such that the step $X_{i,J} \to X_{i,J+1}$ occurs inside the inner loop, we then have

$$\begin{aligned} f(\overline{X}_{i,0}) - f(X_{i,J+1}) &= \sum_{j=0}^{J} \left[f(X_{i,j}) - f(\overline{X}_{i,j+1}) \right] + \sum_{j=0}^{J} \left[f(\overline{X}_{i,j+1}) - f(X_{i,j+1}) \right] \\ &\geq \sum_{j=0}^{J} \left[f(\overline{X}_{i,j+1}) - f(X_{i,j+1}) \right] \\ &\geq -k\delta L_f, \end{aligned} \tag{3.65}$$

since the choice of stepsize guarantees $f(X_{i,j}) > f(\overline{X}_{i,j})$, and rank can decrease at most k times. This shows $f(X_{i,J}) \leq f(\overline{X}_{i,0}) + k\delta L_f \leq f(\overline{X}_{0,0}) + k\delta L_f$ for all such $J \geq 1$. Since $f(\overline{X}_{i,J+1}) \leq f(X_{i,J})$, we have shown that $f(X_{i,j}), f(\overline{X}_{i,j}) \leq f(\overline{X}_{0,0}) + k\delta L_f$ for all j inside the inner loop.

Finally, if the step $X_{i,J} \to \overline{X}_{i,J+1}$ occurs after the inner loop (i.e. it is the rank increase step), then by the same argument we get

$$f(\overline{X}_{i,0}) - f(\overline{X}_{i,J+1}) \geq f(X_{i,J}) - f(\overline{X}_{i,J+1}) - k\delta L_f \geq \epsilon^2 - k\delta L_f > 0, \tag{3.66}$$

where we used Assumption 4. Therefore, $f(\overline{X}_{i,J+1}) < f(\overline{X}_{i,0})$. Since $\overline{X}_{i+1,0} = \overline{X}_{i,J+1}$, this concludes the inductive step. □

Using Assumption 2, let $f_{\text{low}} = \min_{X \in S_0} f(X) > -\infty$. Lemma 3.5 implies that this is a lower bound on the function value at all the iterates.

Lemma 3.6. *For each i, the inner loop in Alg. 5 breaks after at most*

$$N_i = \frac{f(X_{i,0}) - f_{\text{low}} + k\delta L_f}{\tau T \epsilon^2}, \tag{3.67}$$

iterations.

Proof. If the while loop runs for N iterations, then

$$\begin{aligned} f(X_{i,0}) - f_{\text{low}} \geq f(X_{i,0}) - f(X_{i,N}) &= \sum_{j=0}^{N-1} \left[f(X_{i,j}) - f(\overline{X}_{i,j+1}) \right] + \sum_{j=0}^{N-1} \left[f(\overline{X}_{i,j+1}) - f(X_{i,j+1}) \right] \\ &\geq \sum_{j=0}^{N-1} \tau t_{i,j} ||V_{i,j}||_F^2 - k\delta L_f \geq \tau T N \min_{j=0,\ldots,N-1} ||V_{i,j}||_F^2 - k\delta L_f. \end{aligned} \tag{3.68}$$

This implies that if $N \geq N_i$ then $\min_{j=0,\ldots,N-1} ||V_{i,j}||_F \leq \epsilon$ and so the iterations must terminate. □

Define

$$L(X) \coloneqq 26kG||X^\dagger|| + 2H, \quad T(X) \coloneqq \min\left(t_0, \frac{\beta(1-\tau)}{L(X)}\right). \tag{3.69}$$

By the same proof as Lemma 3.4, we have $t \geq T(X)$ for any stepsize t satisfying sufficient decrease for any $V \in T_X \mathcal{M}_{\leq k}$.

Theorem 3.7. *After at most $\frac{f(\overline{X}_{0,0}) - f_{\text{low}}}{\epsilon^2 - k\delta L_f}$ executions of the outer loop in Alg. 5, the algorithm returns a point X whose smallest nonzero singular value is $\geq \delta$ such that either rank$(X) = k$ and $||\Pi_X^{\text{s}} \nabla f(X)||_F \leq \epsilon$, or rank$(X) < k$ and $||\Pi_X \nabla f(X)||_F \leq \epsilon \sqrt{1 + \frac{1}{\tau T(X)}}$.*

Proof. We begin by showing that the algorithm must return after the stated number of iterations. Suppose the outer loop in Alg. 5 executes M times, and for each i the inner loop executes N_i times. The rank increase

occurs $X_{i,N_i} \mapsto \overline{X}_{i,N_i+1}$. Then we can write

$$\begin{aligned} f(\overline{X}_{0,0}) - f_{\text{low}} &\geq \sum_{i=1}^{M} \sum_{j=0}^{N_i} \left\{ \left[f(\overline{X}_{i,j}) - f(X_{i,j}) \right] + \left[f(X_{i,j}) - f(\overline{X}_{i,j+1}) \right] \right\} \\ &\geq \sum_{i=1}^{M} \left\{ \sum_{j=0}^{N_i} \left[f(\overline{X}_{i,j}) - f(X_{i,j}) \right] + \left[f(X_{i,N_i}) - f(\overline{X}_{i,N_i+1}) \right] \right\} \\ &\geq \sum_{i=1}^{M} (-k\delta L_f + \epsilon^2) = M(\epsilon^2 - k\delta L_f), \end{aligned} \tag{3.70}$$

so

$$M \leq \frac{f(\overline{X}_{0,0}) - f_{\text{low}}}{\epsilon^2 - k\delta L_f}. \tag{3.71}$$

Therefore, after at most the claimed number of executions of the outer loop, the algorithm must return.

We have $||\Pi_X^s \nabla f(X)||_F \leq \epsilon$ by the breaking condition for the inner loop. Also, from the if condition preceding line 19, the sufficient decrease requirement, and the stepsize bound mentioned before the Thm. statement, we have

$$\tau T(X) ||(I - \Pi_X^s) \circ \Pi_X \nabla f(X)||_F^2 < \epsilon^2 \quad \text{so} \quad ||(I - \Pi_X^s) \circ \Pi_X \nabla f(X)||_F < \frac{\epsilon}{\sqrt{\tau T(X)}}. \tag{3.72}$$

Thus,

$$||\Pi_X \nabla f(X)||_F^2 = ||\Pi_X^s \nabla f(X)||_F^2 + ||(I - \Pi_X^s) \circ \Pi_X \nabla f(X)||_F^2 \leq \epsilon^2 + \frac{\epsilon^2}{\tau T(X)}, \tag{3.73}$$

as claimed. If $\text{rank}(X) = k$ then $(I - \Pi_X^s) \circ \Pi_X \nabla f(X) = 0$, hence the claimed bound in that case. □

If $\sigma_{\text{rank}(X)}(X) \gg \epsilon^2$ then the bound in Thm. 3.7 is good. However, in the worst case $\sigma_{\text{rank}(X)}(X) = \delta$. As $\epsilon, \delta \to 0$ such that Assumption 4 is satisfied, we have $\epsilon\sqrt{1 + \frac{1}{\tau T}} \sim \frac{\epsilon}{\sqrt{\delta}} > \sqrt{kL_f}$.

3.3.2 Adaptively decreasing the cutoffs

If we can choose *one* pair of ϵ, δ satisfying Assumption 4, we can adaptively decrease them during the algorithm to get an infinite sequence whose limit is a critical point of f on $\mathcal{M}_{\leq k}$. Suppose $(\epsilon_n)_{n \geq 0}$ and $(\delta_n)_{n \geq 0}$ are two sequences such that $\epsilon_n, \delta_n \to 0$ and $k\delta_n L_f < \epsilon_n^2$ for all n. Consider the following modification to Alg. 5, described in Alg. 6.

We can allow ϵ_n, δ_n to depend on $Y_0, \ldots, Y_{n-1}$, but not Y_n.

Some observations:

- By Thm. 3.7 we have

$$||\Pi_{Y_n}^s \nabla f(Y_n)||_F \leq \epsilon_n. \tag{3.74}$$

Algorithm 6 Modified Riemannian GD with decreasing cutoffs

procedure M2RGD($f, \overline{X}_{0,0,0}, k, (\epsilon_n)_{n\geq 0}, (\delta_n)_{n\geq 0}, \tau, \beta, t_0$)
  for $n = 1, 2, \ldots$ **do**
    $\overline{X}_{n,0,0} \leftarrow \overline{X}_{n-1,\text{end},\text{end}}$
    for $i = 1, 2, \ldots$ **do**
      $\overline{X}_{n,i,0} \leftarrow \overline{X}_{n,i-1,\text{end}}$
      $X_{n,i,0} \leftarrow T_{\delta_n}(\overline{X}_{n,i,0})$
      for $j = 0, 1, 2, \ldots$ **do**
        $V_{n,i,j} \leftarrow \Pi_{X_{n,i,j}}\big[-\nabla f(X_{n,i,j})\big]$ ▷ descent direction
        if $||V_{n,i,j}||_F \leq \epsilon_n$ **then**
          break ▷ break out of j-loop
        end if
        $r_{n,i,j} \leftarrow \text{rank}(X_{n,i,j})$
        $t_{n,i,j} \leftarrow \text{backtrack}(f, X_{n,i,j}, V_{n,i,j}, r_{n,i,j}, \tau, \beta, t_0)$ ▷ stepsize
        $\overline{X}_{n,i,j+1} \leftarrow P_{r_{n,i,j}}(X_{n,i,j} + t_{n,i,j}V_{n,i,j})$ ▷ move
        $X_{n,i,j+1} \leftarrow T_{\delta_n}(\overline{X}_{n,i,j+1})$ ▷ project down if needed
      end for
      $W_{n,i} \leftarrow \Pi^{\perp}_{X_{n,i,\text{end}}} \circ \Pi^{\leq k}_{X_{n,i,\text{end}}}\big[-\nabla f(X_{n,i,\text{end}})\big]$ ▷ direction for rank increase
      $t_{n,i,\text{end}+1} \leftarrow \text{backtrack}(f, X_{n,i,\text{end}}, V_{n,i,\text{end}}, k, \tau, \beta, t_0)$
      $\overline{X}_{n,i,\text{end}+1} \leftarrow X_{n,i,\text{end}} + t_{n,i,\text{end}}V_{n,i,\text{end}}$
      if $f(X_{n,i,\text{end}}) - f(\overline{X}_{n,i,\text{end}}) < \epsilon_n^2$ **then**
        $\overline{X}_{n,i,\text{end}} \leftarrow X_{n,i,\text{end}}$ ▷ Undo rank increase
        $Y_n \leftarrow X_{n,i,\text{end}}$
        break ▷ break out of i-loop
      end if
    end for
  end for
end procedure

- All nonzero singular values of Y_n are $\geq \delta_n$.

We can prove:

Theorem 3.8. *If* $\limsup_{n\to\infty} \sigma_{r_n}(Y_n) > 0$, *then the sequence* $(Y_n)_{n\geq 1}$ *has a cluster point that is a critical point of* f *on* $\mathcal{M}_{\leq k}$.

Proof. By Lemma 3.5 and Assumption 2, the sequence (Y_n) and any of its subsequences indeed has a cluster point. If $\limsup_n \sigma_{r_n}(Y_n) > 0$, then we can pick an infinite subsequence (Y_{n_i}) such that $\sigma_{r_{n_i}}(Y_{n_i}) \geq \lambda > 0$ for some λ and all i. By passing to a subsequence again, we may assume that $\lim_{i\to\infty} Y_{n_i} = Y^*$ exists. For simplicity of notation, we drop the i sub-subscript and just write $Y^* = \lim_n Y_n$ and $\sigma_{r_n}(Y_n) \geq \lambda > 0$.

Let $r = \operatorname{rank}(Y^*)$. We have $||Y^* - Y_n|| < \lambda$ for all large n. Since all the singular values of all the Y_n are $\geq \lambda$ while

$$\sigma_{r+1}(Y_n) = |\sigma_{r+1}(Y^*) - \sigma_{r+1}(Y_n)| \leq ||Y^* - Y_n|| < \lambda, \tag{3.75}$$

we have $\sigma_{r+1}(Y_n) = 0$ for all large n. On the other hand,

$$|\sigma_r(Y^*) - \sigma_r(Y_n)| \leq ||Y^* - Y_n|| \leq \sigma_r(Y^*)/2, \tag{3.76}$$

for all large n, hence $\sigma_r(Y_n) > 0$ for all large n. Thus, $\operatorname{rank}(Y_n) = r$ for all large n.

Regardless of r, since $\epsilon_n \to 0$ we have $||\Pi^s_{Y_n} \nabla f(Y_n)||_F \to 0$ so $||\Pi^s_{Y^*} \nabla f(Y^*)||_F = 0$. If $r = k$ this implies Y^* is a critical point.

Suppose $r < k$. As argued above, $\sigma_r(Y_n) \geq \sigma_r(Y^*)/2$ for all large n. Therefore,

$$L(Y_n) \leq L^* = \frac{52kG}{\sigma_r(Y^*)} + 2H \implies T(Y_n) \geq T^* = \min\left(t_0, \frac{\beta(1-\tau)}{L^*}\right). \tag{3.77}$$

By Thm. 3.7, we have

$$||\Pi_{Y_n} \nabla f(Y_n)||_F \leq \epsilon_n \sqrt{1 + \frac{\tau}{L^*}} \to 0. \tag{3.78}$$

Since $\operatorname{rank}(Y_n) = r < k$ for all large n, this implies $||\Pi_{Y^*} \nabla f(Y^*)||_F = 0$ (which also implies $\nabla f(Y^*) = 0$) and Y^* is a critical point of f. □

The crucial property that allowed us to derive the above result is the existence of a convergence subsequence such that the rank of the limit equals the rank of the iterates.

Of course, the hypothesis in Thm. 3.8 is completely unjustified. Under the same hypothesis, regular gradient descent as in Alg. 3 (taken from [21]) also converges to a critical point, by the same proof. However,

it is unclear how to avoid it — to get convergence to a critical point, we need to let $\delta \to 0$, but in doing so we lost control over the smallest nonzero singular value.

3.3.3 The case of finitely many singularities

Alg. 6 provably works and can be significantly simplified for varieties with finitely many singular points, such as $\mathcal{M}_{\leq 1}$. The idea is that whenever a sequence gets close to a singular point, we project the next iterate to the singular locus. If the resulting singular point is not 1-critical, the cost function decreases by some amount that only depends on the singular point. If there are only finitely many singular points and the function is bounded from below, either we converge to one of those singular points in finitely many steps and stop, or we converge to a smooth point. Therefore, the issues arising in the preceding section when a sequence of smooth points converges to a singular one do not occur.

Specifically, suppose $\mathcal{V} \subset \mathbb{R}^n$ is a variety with finitely many singularities, and let Π_{sing} be the projection to the singular locus. Consider Alg. 7, where 'backtrack' again refers to Alg. 4.

Algorithm 7 Gradient descent on varieties with finitely many singularities.

procedure RGD_FMS($f, x_0, \delta_0, \tau, \beta, t_0$) ▷ Function f, initial guess x_0, initial cutoff δ_0 on distance to singular locus, linesearch parameters τ, β, t_0.

  for $i = 0, 1, 2, \ldots$ **do**

    $v_i \leftarrow \Pi_{x_i}[-\nabla f(x_i)]$ ▷ descent direction

    $t_i \leftarrow \text{backtrack}(f, x_i, v_i, k, \tau, \beta, t_0)$ ▷ stepsize

    $x_{i+1} \leftarrow R_{x_i}(t_i v_i)$ ▷ move

    if $x_{i+1} = x_i$ **then**

      break

    else if $||x_{i+1} - \Pi_{\text{sing}}(x_{i+1})|| \leq \delta_i$ **then** ▷ project and halve cutoff

      $x_{i+1} \leftarrow \Pi_{\text{sing}}(x_{i+1})$

      $\delta_{i+1} \leftarrow \delta_i / 2$

    else

      $\delta_{i+1} \leftarrow \delta_i$

    end if

  end for

end procedure

We argue that:

Proposition 3.9. Suppose f is $\mathcal{C}^1$, bounded from below, and has compact sublevel sets. Then the sequence $(x_i)_{i\geq 0}$ generated by Alg. 7 is either finite and terminates at a singular 1-critical point, or has a smooth 1-critical point as a cluster point.

Proof. First, we argue that the sequence (x_i) remains in some fixed sublevel set. Note that f can only increase when we project $x_{i+1} \leftarrow \Pi_{\text{sing}}(x_{i+1})$, since when we retract $x_{i+1} \leftarrow R_{x_i}(t_i v_i)$ the cost decreases by at least $f(x_i) - f(x_{i+1}) \geq t_i \tau \langle -\nabla f(x_i), v_i \rangle > 0$ by our choice of step size t_i in Alg. 4. Therefore, if the projection step is executed finitely many times, the sequence remains in some sublevel set. Assume that the projection step is executed infinitely many times. Then the cutoff δ_i is halved infinitely many times, so $\delta_i \to 0$. Let $y_1, \ldots, y_p$ be the singular points of $\mathcal{V}$. Since f is $\mathcal{C}^1$, it is locally Lipschitz near each of those points, so there exist $\Delta_1, \ldots, \Delta_p > 0$ and $L_1, \ldots, L_p > 0$ such that $|f(x) - f(y_i)| \leq L_i ||x - y_i||$ for all x satisfying $||x - y_i|| \leq \Delta_i$. Let $\Delta = \min_i \Delta_i$ and $L = \max_i L_i$. For all $i \geq i_0$ for some large enough i_0, we have $\delta_i < \Delta$. If for such large i we project x_{i+1} to the singular locus, e.g. if $\Pi_{\text{sing}}(x_{i+1}) = y_j$, then the increase in cost is bounded by $|f(y_j) - f(x_{i+1})| \leq L||x_{i+1} - y_j|| \leq L\delta_i$. Thus, starting from some finite i_0, the increase in cost is bounded by $L \sum_{i=0}^{\infty} \delta_0 / 2^i = 2L\delta_0$. Therefore, the iterates remain inside some sublevel set.

Second, suppose $y \in \mathcal{V}$ is a singular point such that $x_i = y$ for infinitely many i. We claim that y must be 1-critical for f, so $\Pi_y[-\nabla f(y)] = 0$ and the algorithm terminates at y after finitely many steps. Indeed, if not then for infinitely many i we have $x_i = y$ and $x_{i+1} = R_y(t_y v_y)$ where $v_y = \Pi_y[-\nabla f(y)] \neq 0$ and t_y is chosen by Alg. 4 guaranteeing that $f(x_i) - f(x_{i+1}) \geq \tau t_y \langle -\nabla f(y), v_y \rangle > 0$. Therefore, the cost function decreases by at least a constant amount infinitely many times. Since the iterates remain in some sublevel set so the increase in the function is bounded, this contradicts the fact that f is bounded from below. Thus, we must have $v_y = 0$ so y is 1-critical and $x_i = y$ for all large i.

Third, if a singular point y is a cluster point of Alg. 7, then in fact y is 1-critical and the algorithm terminates at y after finitely many steps. Indeed, if a singular point is a cluster point of (x_i) then we must have $\delta_i \to 0$, otherwise if $\delta_i \geq \delta > 0$ then the distance from the sequence (x_i) and the singular locus of $\mathcal{V}$ is $\geq \delta$. Since δ_i is decreased only after projection to the singular locus and the singular locus is finite, one of the singular points must appear in the sequence (x_i) infinitely many times. The claim therefore follows by the preceding paragraph.

Finally, since the sequence remains in some sublevel set which is assumed to be compact, it must have some cluster point. Thus, either the cluster point is smooth, or it is singular in which case the point is 1-critical and the algorithm terminates there after finitely many steps. If it smooth, then the algorithm reduces to standard Riemannian gradient descent on the smooth locus of $\mathcal{V}$, in which case the argument in Sec. 1.2.3 shows that the cluster point must be 1-critical. □

In particular, the above shows that $\delta_i \geq \delta > 0$ is bounded away from zero by some problem- and initialization-dependent bound. If $\mathcal{V} = \mathcal{M}_{\leq 1}$, whose only singularity is at the origin, this gives a bound on the step sizes as in Lemma 3.4. The same argument as in Lemma 3.6 then shows that the algorithm either converges to 0 which is 1-critical, or finds a smooth point a distance $\geq \delta$ away from 0 with $||\text{grad } f(X)|| \leq \epsilon$ in $O(1/\epsilon^2)$ iterations.

Chapter 4

Conclusions and future directions

We summarize our main findings:

- Optimization over varieties in general can be complicated — the metric projection retraction may not be single-valued nor twice-differentiable at a singular point (Sec. 1.3.1), smooth functions may not have locally retraction-Lipschitz gradients near singularities (Sec. 1.3.2), and even linear functions may not have retraction-Lipschitz gradients (Sec. 1.3.3). Most importantly for optimization, there may exist sequences converging to a singular point along which the gradient tends to zero, but the limit singular point is not 1-critical (Sec. 1.3.4).

- To circumvent the above problems for the variety of bounded-rank matrices, we considered optimizing over lifted spaces which are smooth manifolds. We show that for a large class of varieties, there are no lifts to smooth manifolds such that 1-critical points on the lift map to 1-critical points on the variety (Cor. 2.7). Nevertheless, for all three lifts we considered for the bounded-rank matrix variety, we show that 2-critical point on the lift map to 1-critical points on the variety (Props. 2.17, 2.20, 2.24). Unfortunately, there may exist local minima on these lifted spaces that map to saddles on the original variety (Prop. 2.30, Cor. 2.31, and Prop. 2.37).

- For one of the lifts, we were able to identify a subset of points such that they are local minima on the lift if and only if they correspond to a local minimum on the variety (Prop. 2.34). While they do not form a smooth manifold, we can guarantee convergence to them by regularizing the cost function (Props. 2.40-2.42).

- We gave an explicit example showing that naively extending gradient descent to the variety of bounded-rank matrices can result in convergence to a saddle point (Sec. 3.1). We proposed a fix based on

projecting to the singular locus (Sec. 3.3). Our analysis is unsatisfactory, but illustrates the main difficulties. Along the way, we also bound the difference between the metric projection retraction to the bounded-rank matrix variety and its first-order Taylor expansion (Thm. 3.1). Also, we prove that the above approach does work for varieties with finitely many singular points.

We close by stating some of the open questions suggested by the above work.

- Does a good lift of the bounded-rank matrix variety exist? Specifically, does there exist a lift $\varphi : \overline{\mathcal{M}}_{\leq k} \to \mathcal{M}_{\leq k}$ such that $\overline{\mathcal{M}}_{\leq k}$ is a smooth manifold, which satisfies MLMP everywhere (any local minimum for any cost function on the lift maps to a local minimum on the variety)? Can we find a lift that preserves 2-critical points (i.e. a point on the lift is 2-critical iff it maps to a 2-critical point on the variety), or are there obstructions like Cor. 2.7 for 1-critical points? Can we identify a class of varieties that admit lifts such that 2-critical points on the lift map to 1-critical points on the variety, generalizing our results for $\mathcal{M}_{\leq k}$?

- In some cases, it is possible to *compactify* the variety of interest. For example, considering the lift based on factorizations $(U, W) \in \mathrm{St}(m, k) \times \mathbb{R}^{n \times k}$ (see Sec. 2.2), one can "project" the variable W, as follows: $\bar{g}(U) \coloneqq g(U, W_U) = f(UW_U^T)$ where $W_U \in \arg\min_W g(U, W)$. If W_U exists, then one can check that the resulting function $\bar{g}$ is well defined on the Grassmannian manifold $\mathrm{Gr}(m, k)$. Since the latter is compact, this provides a means to "lift" the variety of matrices of bounded rank to a smooth and compact manifold (at the price of making the cost function more complicated, possibly substantially so). This is the approach used in [4] to solve matrix completion. Can we generalize that approach to a more general class of varieties and cost functions? How do the necessary optimality conditions and local minima on the compactification compare to those on the original variety?

- Can we design an algorithm to directly optimize over the variety $\mathcal{M}_{\leq k}$, without lifting? Specifically, can we design an algorithm generating a sequence that is guaranteed to have a 1-critical point on the entire variety as a cluster point? Moreover, can we prove global convergence rates for such an algorithm?

- The example we gave in Sec. 3.1 is fragile. Specifically, only a zero measure set of initializations gives the pathological behavior outlined there. Is this the general behavior for such examples, or is there an example for which a positive-measure set of initializations yields the same pathology?

- Is there a principled way to optimize over more general stratified spaces, such as other varieties or spaces arising from proper but non-free group actions [8]? For example, is there a systematic way of constructing good explicit lifts that can be used in algorithms?

Bibliography

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[2] P.-A. Absil and J. Malick. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012.

[3] N. Boumal. An introduction to optimization on smooth manifolds. To appear, 2020.

[4] N. Boumal and P.-A. Absil. Low-rank matrix completion via preconditioned optimization on the grassmann manifold. *Linear Algebra and its Applications*, 475:200 – 239, 2015.

[5] N. Boumal, P.-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 02 2018.

[6] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15:1455–1459, 2014.

[7] S. C. Eisenstat and I. C. F. Ipsen. Relative perturbation techniques for singular value problems. *SIAM Journal on Numerical Analysis*, 32(6):1972–1988, 1995.

[8] J. Giacomoni. On the stratification by orbit types. *Bulletin of the London Mathematical Society*, 46(6):1167–1170, 08 2014.

[9] W. Ha, H. Liu, and R. F. Barber. An equivalence between stationary points for rank constraints versus low-rank factorizations. *arXiv preprint arXiv:1812.00404*, 2018.

[10] V. Khrulkov and I. Oseledets. Desingularization of bounded-rank matrix sets. *SIAM Journal on Matrix Analysis and Applications*, 39(1):451–471, 2018.

[11] J. Kollár. Resolution of singularities–seattle lecture. *arXiv preprint math/0508332*, 2005.

[12] J. B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

[13] J. M. Lee. *Introduction to smooth manifolds*. Springer, 2012.

[14] J. M. Lee. *Introduction to Riemannian manifolds*. Springer, 2018.

[15] Y.-C. Lu. *Thom-Whitney Stratification Theory*, pages 120–141. Springer New York, New York, NY, 1976.

[16] B. Mishra, G. Meyer, S. Bonnabel, and R. Sepulchre. Fixed-rank matrix factorizations and riemannian low-rank optimization. *Computational Statistics*, 29(3-4):591–621, 2014.

[17] N. Nguyen, S. Trivedi, and D. Trotman. A geometric proof of the existence of definable whitney stratifications. *Illinois J. Math.*, 58(2):381–389, 2014.

[18] 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.

[19] P. A. Parrilo. Semidefinite programming relaxations for semialgebraic problems. *Mathematical programming*, 96(2):293–320, 2003.

[20] A. Ruszczyński. *Nonlinear Optimization*. Princeton University Press, 2006.

[21] 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.

[22] M. Tan, I. W. Tsang, L. Wang, B. Vandereycken, and S. J. Pan. Riemannian pursuit for big matrix recovery. In *International Conference on Machine Learning*, pages 1539–1547, 2014.

[23] A. Uschmajew and B. Vandereycken. Line-search methods and rank increase on low-rank matrix varieties. In *Proceedings of the 2014 International Symposium on Nonlinear Theory and its Applications (NOLTA2014)*, pages 52–55, 2014.

[24] 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. IEEE, 2015.

[25] B. Vandereycken. Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013.

[26] Y. Wang and Z. Xu. Generalized phase retrieval: measurement number, matrix recovery and beyond. *Applied and Computational Harmonic Analysis*, 47(2):423–446, 2019.

[27] W. H. Yang, L.-H. Zhang, and R. Song. Optimality conditions for the nonlinear programming problems on riemannian manifolds. *Pacific Journal of Optimization*, 10(2):415–434, 2014.

[28] G. Zhou, W. Huang, K. A. Gallivan, P. V. Dooren], and P.-A. Absil. A riemannian rank-adaptive method for low-rank optimization. *Neurocomputing*, 192:72 – 80, 2016. Advances in artificial neural networks, machine learning and computational intelligence.

[29] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin. Global optimality in low-rank matrix optimization. *IEEE Transactions on Signal Processing*, 66(13):3614–3628, 2018.