Projection-Like Retractions on Matrix Manifolds

Pierre-Antoine Absil, Jérôme Malick

PROJECTION-LIKE RETRACTIONS ON MATRIX MANIFOLDS*

P.-A. ABSIL† AND JÉRÔME MALICK‡

Abstract. This paper deals with constructing retractions, a key step when applying optimization algorithms on matrix manifolds. For submanifolds of Euclidean spaces, we show that the operation consisting of taking a tangent step in the embedding Euclidean space followed by a projection onto the submanifold, is a retraction. We also show that the operation remains a retraction if the projection is generalized to a projection-like procedure that consists of coming back to the submanifold along "admissible" directions, and we give a sufficient condition on the admissible directions for the generated retraction to be second order. This theory offers a framework in which previously-proposed retractions can be analyzed, as well as a toolbox for constructing new ones. Illustrations are given for projection-like procedures on some specific manifolds for which we have an explicit, easy-to-compute expression.

Key words. equality-constrained optimization, matrix manifold, feasible optimization method, retraction, projection, fixed-rank matrices, Stiefel manifold, spectral manifold

AMS subject classifications. 49Q99, 53B20, 65F30, 65K05, 90C30

1. Introduction. Computational problems abound that can be formulated as finding an optimal point of a smooth real-valued function defined on a manifold; see, e.g., [**?**, **?**] and the references therein. Following on the work of Luenberger [**?**] and Gabay [**?**], much of the early interest focused on manifold-based optimization methods that exploit the underlying geometry of those optimization problems by relying almost exclusively on mainstream differential-geometric concepts; major references include [**?**, **?**, **?**, **?**, **?**]. For example, Smith's Riemannian Newton method [**?**, Alg. 4.3] makes use of the Levi-Civita connection to define the Hessian of the cost function, and on geodesics (specifically, the Riemannian exponential) to produce the update along the computed Newton vector. In the case of Lie groups and homogeneous spaces, a similar inclination for classical differential-geometric objects can be observed in [**?**].

However, it became increasingly clear that it could be beneficial to work in a broader context that offers leeway for replacing classical differential-geometric objects with certain approximations, resulting in faster and possibly more robust algorithms. These ideas, which can be seen burgeoning in [**?**, Remark 4.9] and [**?**, §3.5.1], blossomed in the early 2000s [**?**, **?**]. In particular, relaxing the Riemannian exponential update led to the concept of retraction [**?**, §3], which can be traced back to [**?**]. The key property of a retraction R on a submanifold $\mathcal{M}$ of a Euclidean space is that, for every tangent vector u at a point x of $\mathcal{M}$, one has (see Proposition 2.2)

$$\operatorname{dist}(\Gamma(t,x,u), R(x,tu)) = O(t^2), \tag{1.1}$$

where $t \mapsto \Gamma(t,x,u)$ is the geodesic of $\mathcal{M}$ with initial position-velocity (x,u). Retractions thus generate approximations of geodesics that are first-order accurate. A retraction R can also be viewed as providing "locally rigid" mappings from the tangent spaces $\mathrm{T}_{\mathcal{M}}(x)$ into $\mathcal{M}$; such mappings come useful notably in trust-region methods [**?**] to obtain local models of the objective function that live on flat spaces.

*This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its authors.

†Department of Mathematical Engineering, ICTEAM Institute, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium (`http://www.inma.ucl.ac.be/~absil/`).

‡CNRS, Lab. J. Kuntzmann, Grenoble, France (`http://bipop.inrialpes.fr/people/malick/`).

Note that, if the right-hand side of (1.1) is replaced by $O(t)$, then R is just a *topological retraction* [**?**]. Condition (1.1) as stated characterizes *first-order retractions*, simply called *retractions* throughout this paper. A crucial reason for considering (first-order) retractions is the following: if, in Smith's Riemannian Newton method [**?**, Alg. 4.3], the exponential update is replaced by any (first-order) retraction, then local quadratic convergence is preserved [**?**, **?**].

In this paper, we pursue the effort of [**?**, **?**] and others to develop a toolbox for building retractions on manifolds. We focus on the case of d-dimensional submanifolds of n-dimensional Euclidean spaces. This is a mild restriction since, by virtue of Whitney's embedding theorem, every d-dimensional manifold can be embedded into $\mathbb{R}^n$ with $n = 2d$; however, whether this fact leads to tractable retractions depend on the tractability of the embedding. In any case, several important manifolds admit well-known expressions as submanifolds of Euclidean spaces, and more specifically, of matrix spaces; see the examples presented in this paper.

Here is the outline of this paper. Our notation and the precise definition of a retraction are presented in Section 2. Though we use the language and notions of differential geometry, the developments in Section 2 and in most of the rest of the paper do not require any prior differential-geometric background. For a submanifold $\mathcal{M}$ of a Euclidean space, we show in Section 3 that the operation that consists of moving along the tangent vector and then projecting onto $\mathcal{M}$, is a retraction. We work out easy-to-compute formulas for this projection on various specific manifolds, insisting in particular on spectral manifolds. In Section 4, we generalize this projective retraction by defining $R(x, u)$ to be the point of $\mathcal{M}$ closest to $x + u$ along an $(n - d)$-dimensional subspace $D(x, u)$ of "admissible directions", where $D(x, u)$ depends smoothly on (x, u). If the subspace $D(x, 0)$ has a trivial intersection with $\mathrm{T}_{\mathcal{M}}(x)$, a generic situation, then we show that R is a bona-fide retraction. Moreover, if the subspace $D(x, 0)$ is the orthogonal complement of $\mathrm{T}_{\mathcal{M}}(x)$, then R is a second-order retraction, which means that $\mathrm{dist}(\mathrm{Exp}(x, tu), R(x, tu)) = O(t^3)$. We show that some choices of D yield well-known retractions. In particular, the implicit definition $D(x, u) = (\mathrm{T}_{\mathcal{M}}(R(x, u))^{\perp}$ yields the projective retraction, from which it follows directly that the retraction is second-order, whereas the choice $D(x, u) = (\mathrm{T}_{\mathcal{M}}(x))^{\perp}$ yields a second-order retraction that relates to the tangential parameterization defined in [**?**, §3]. We provide examples for particular manifolds.

We conclude this introduction by mentioning that these projective and projective-like retractions relate to ideas from nonsmooth optimization. The so-called $\mathcal{U}$-Newton method for convex optimization [**?**] uses indeed a correction step to identify an underlying structure that can be interpreted as a projection to implicit smooth constraints. The picture is even clearer in [**?**] which presents such a "projected" $\mathcal{U}$-Newton for minimizing nonsmooth functions involving the maximal eigenvalue of symmetric matrix, using the projection onto fixed-rank matrix manifolds. A first attempt at connecting the scientific community of "nonsmooth optimization", the one of "constrained optimization", and the one of "optimization on manifolds" was made in [**?**, **?**], where it was explained how the $\mathcal{U}$-Newton methods relate to Riemannian Newton methods and to SQP methods. The present paper aims at clarifying the role of projections in optimization algorithms by keeping the parameter-free, geometrical viewpoint of optimization on manifolds.

2. Retractions on submanifolds. This section presents the notation, and then recalls the notion of retraction with a few of its properties. Some of these properties are well known, some are less, and all are basic. In particular, we describe how paths

defined by retractions relate to geodesics.

2.1. Background and notation. We start by recalling elementary facts about manifolds; for more information, see, e.g., [**?**].

Submanifolds. Our work space is a Euclidean space $\mathcal{E}$ of dimension n. In all the examples considered here, $\mathcal{E}$ is the matrix space $\mathbb{R}^{n\times m}$ endowed with the standard (i.e., Frobenius) inner product. Throughout the paper, $\mathcal{M}$ stands for a submanifold of $\mathcal{E}$ of class C^k ($k \geq 2$) and of dimension d, unless otherwise stated explicitly. By this, we mean that $\mathcal{M}$ is locally a *coordinate slice*, that is, for all $\bar{x} \in \mathcal{M}$, there exists a neighborhood $\mathcal{U}_\mathcal{E}$ of $\bar{x}$ in $\mathcal{E}$ and a C^k diffeomorphism ϕ on $\mathcal{U}_\mathcal{E}$ into $\mathbb{R}^n$ such that

$$\mathcal{M} \cap \mathcal{U}_\mathcal{E} = \{x \in \mathcal{U}_\mathcal{E} : \phi_{d+1}(x) = \cdots = \phi_n(x) = 0\}.$$

We denote the tangent and normal spaces of $\mathcal{M}$ at $x \in \mathcal{M}$ by $\mathrm{T}_\mathcal{M}(x)$ and $\mathrm{N}_\mathcal{M}(x)$, respectively. Since $\mathcal{M}$ is embedded in the Euclidean space $\mathcal{E}$, we can identify $\mathrm{T}_\mathcal{M}(x)$ as a linear subspace of $\mathcal{E}$, and likewise for $\mathrm{N}_\mathcal{M}(x)$; we will then consider for example sums $x + v$ with $x \in \mathcal{M}$ and $v \in \mathrm{N}_\mathcal{M}(x)$. The orthogonal projection onto $\mathrm{T}_\mathcal{M}(x)$ is denoted by $\mathrm{P}_{\mathrm{T}_\mathcal{M}(x)}$.

Bundles. We will need a few more sophisticated properties of (sub)manifolds. The tangent bundle of $\mathcal{M}$ is defined by

$$\mathrm{T}\mathcal{M} = \{(x, u) \in \mathcal{E} \times \mathcal{E} : \ x \in \mathcal{M}, u \in \mathrm{T}_\mathcal{M}(x)\};$$

it is a C^{k-1} submanifold of $\mathcal{E} \times \mathcal{E}$ of dimension $2d$. We will also consider the normal bundle

$$\mathrm{N}\mathcal{M} = \{(x, v) \in \mathcal{E} \times \mathcal{E} : \ x \in \mathcal{M}, v \in \mathrm{N}_\mathcal{M}(x)\},$$

as well as the so-called Whitney sums of such bundle manifolds, as for example, the tangent-normal bundle

$$\mathrm{B}\mathcal{M} = \mathrm{T}\mathcal{M} \oplus \mathrm{N}\mathcal{M} = \left\{(x, u, v) \in \mathcal{E}^3 : \ x \in \mathcal{M}, u \in \mathrm{T}_\mathcal{M}(x), v \in \mathrm{N}_\mathcal{M}(x)\right\}.$$

The bundles $\mathrm{N}\mathcal{M}$ and $\mathrm{B}\mathcal{M}$ are manifolds of class C^{k-1}, of dimension n and $n + d$, respectively.

Geodesics. A geodesic on $\mathcal{M}$ is a curve on $\mathcal{M}$ that locally minimizes the arc length, hence generalizing straight lines in vector spaces. Equivalently, the geodesics on $\mathcal{M}$ are the curves γ on $\mathcal{M}$ that satisfy $\gamma''(t) \in \mathrm{N}_\mathcal{M}(\gamma(t))$ for all t, where γ'' denotes the second derivative of γ viewed as a curve in the Euclidean space $\mathcal{E}$. The exponential of a tangent vector u at x, denoted by $\mathrm{Exp}(x, u)$, is defined to be $\Gamma(1, x, u)$, where $t \mapsto \Gamma(t, x, u)$ is the geodesic that satisfies

$$\Gamma(0, x, u) = x \qquad \text{and} \qquad \frac{\mathrm{d}}{\mathrm{d}t}\Gamma(0, x, u)\bigg|_{t=0} = u.$$

The exponential map is a smooth map from the tangent bundle $\mathrm{T}\mathcal{M}$ into $\mathcal{M}$. Moreover, it is easy to check that $\mathrm{Exp}(x, tu) = \Gamma(1, x, tu) = \Gamma(t, x, u)$. These results can be found, e.g., in [**?**, §VII].

2.2. Newton's method on manifolds and retractions. As mentioned in the introduction, the notion of retraction was introduced in [**?**] (see also the precursor [**?**]) in the context of Newton's method on submanifolds, which we recall briefly here to set the framework.

The Newton method on $\mathcal{M}$ to find a zero of a smooth function $F\colon \mathcal{M} \to \mathcal{E}$ proceeds as follows. At a current iterate $x \in \mathcal{M}$, the Newton equation yields the update vector

$$\eta_x := -\mathrm{D}F(x)^{-1}F(x) \in \mathrm{T}_{\mathcal{M}}(x),$$

assuming that $\mathrm{D}F(x)$ is invertible. When $\mathcal{M}$ is a linear manifold, the natural next iterate is $x_+ := x + \eta_x$. When $\mathcal{M}$ is nonlinear, it is still possible to give a meaningful definition to $x+\eta_x$ by viewing x and η_x as elements of the Euclidean space $\mathcal{E}$, but $x+\eta_x$ does in general not belong to $\mathcal{M}$. A remedy, considered, e.g., in [**?**, **?**], is to define the next iterate to be $x_+ := \mathrm{Exp}(x, \eta_x)$. However, in general, computing the exponential updates requires solving the ordinary differential equation that defines the geodesics, which is computationally expensive. Even on those manifolds where the geodesic admits a simple closed-form solution, the exponential update could advantageously be replaced (see e.g. [**?**, §3.5.1] and [**?**]) by certain approximations that reduce the computational burden per step without hurting the convergence properties of the iteration. The concept of retraction was introduced to provide a framework for these approximations to the exponential update [**?**].

In a nutshell, a retraction is any smooth map from the tangent bundle of $\mathcal{M}$ into $\mathcal{M}$ that approximates the exponential map to the first order. This idea is formalized in the next definition.

DEFINITION 2.1 (Retraction). *Let $\mathcal{M}$ be a submanifold of $\mathcal{E}$ of class C^k ($k \geq 2$). A mapping R from the tangent bundle $\mathrm{T}\mathcal{M}$ into $\mathcal{M}$ is said to be a* retraction on $\mathcal{M}$ around $\bar{x} \in \mathcal{M}$ *if there exists a neighborhood $\mathcal{U}$ of $(\bar{x}, 0)$ in $\mathrm{T}\mathcal{M}$ such that:*

1. *We have $\mathcal{U} \subseteq \mathrm{dom}(R)$ and the restriction $R\colon \mathcal{U} \to \mathcal{M}$ is of class C^{k-1};*
2. *$R(x, 0) = x$ for all $(x, 0) \in \mathcal{U}$, where 0 denotes the zero element of $\mathrm{T}_{\mathcal{M}}(x)$;*
3. *$\mathrm{D}R(x, \cdot)(0) = \mathrm{id}_{\mathrm{T}_{\mathcal{M}}(x)}$ for all $(x, 0) \in \mathcal{U}$.*

If R is a retraction on $\mathcal{M}$ around every point $x \in \mathcal{M}$, then we say that R is a retraction *on $\mathcal{M}$.*

This above definition of a retraction for a C^k-submanifold is adapted from Definition 4.1.1 in [**?**]. We insist here on two points:

- Local smoothness: Even tough the mapping R may be defined from the whole $\mathrm{T}\mathcal{M}$ to $\mathcal{M}$, its smoothness is required only locally around some point $\bar{x}$. This is sufficient to preserve important properties of algorithms built with this mapping. In particular, when $\bar{x}$ is a nondegenerate minimizer of the objective function, Newton's method on manifolds has the usual quadratic local convergence property, regardless of the retraction utilized to apply the Newton update vector [**?**, **?**].
- Accurate smoothness: We emphasize that if $\mathcal{M}$ is of class C^k, then the retraction is of class C^{k-1}. The smoothness of the retraction is thus "maximal": the tangent bundle $\mathrm{T}\mathcal{M}$ is of class C^{k-1}, so a differentiable function defined on $\mathrm{T}\mathcal{M}$ cannot be smoother than C^{k-1}.

The next result shows that the retractions are those maps that agree with the Riemannian exponential up to and including the first order. In particular, the Riemannian exponential is a retraction.

PROPOSITION 2.2 (Retractions as collections of curves). *Let $\mathcal{M}$ be a submanifold of $\mathcal{E}$ of class C^k ($k \geq 2$), and let R be a C^{k-1}-mapping on $\mathcal{U}$, a neighborhood of $(\bar{x}, 0)$ in the tangent bundle $\mathrm{T}\mathcal{M}$, into $\mathcal{M}$. For any $(x, u) \in \mathcal{U}$ fixed, we define the curve on $\mathcal{M}$*

$$t \to \gamma(x, u, t) := R(x, tu).$$

Then R is a retraction on $\mathcal{M}$ around $\bar{x}$ if and only if, for all $(x,u) \in \mathcal{U}$,

$$\gamma(x,u,0) = x \qquad \text{and} \qquad \gamma(x,u,\cdot)'(0) = u.$$

Moreover, the retractions give birth to paths on $\mathcal{M}$ that agree with geodesics up to the second-order: for any $u \in \mathrm{T}_{\mathcal{M}}(x)$,

$$R(x,tu) = \Gamma(x,u,t) + o(t) \text{ as } t \to 0. \tag{2.1}$$

If we require $k \geq 3$, then we even have

$$R(x,tu) = \Gamma(x,u,t) + O(t^2) \text{ as } t \to 0, \tag{2.2}$$

The converse is also true: if R satisfies property (2.1), then R is a retraction around $\bar{x}$.

Proof. This proof of the first point is easy since the involved objects have the right smoothness properties: we just have to use the chain rule to write

$$\gamma(x,u,\cdot)'(t) = \mathrm{D}R(x,tu)(u)$$

that allows to conclude. The last result follows from Taylor's theorem on $t \mapsto R(x,tu)$ and $t \mapsto \Gamma(x,u,t)$ viewed as curves in $\mathcal{E}$. □

We finish this preliminary section with a last definition. When $k \geq 3$, a *second-order retraction* on the submanifold $\mathcal{M}$ is a retraction R on $\mathcal{M}$ that satisfies, for all $(x,u) \in \mathrm{T}\mathcal{M}$,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} R(x,tu)|_{t=0} \in \mathrm{N}_{\mathcal{M}}(x); \tag{2.3}$$

or equivalently, $\mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}(\frac{\mathrm{d}^2}{\mathrm{d}t^2} R(x,tu)|_{t=0}) = 0$.

The next result shows that the second-order retractions are those maps that agree with the Riemannian exponential up to and including the second order. In particular, the Riemannian exponential is a second-order retraction.

Proposition 2.3 (Second-order retractions). *Let $\mathcal{M}$ be a submanifold of $\mathcal{E}$ of class C^k $(k \geq 3)$, and let R be a (first-order) C^{k-1}-retraction on $\mathcal{M}$. Then R is a second-order retraction if and only if, for all $x \in \mathcal{M}$ and all $u \in \mathrm{T}_{\mathcal{M}}(x)$,*

$$R(x,tu) = \Gamma(x,u,t) + o(t^2) \text{ as } t \to 0. \tag{2.4}$$

If we require $k \geq 4$, then we even have

$$R(x,tu) = \Gamma(x,u,t) + O(t^3) \text{ as } t \to 0. \tag{2.5}$$

Proof. The condition is sufficient. From (2.4), we have that $\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Gamma(x,u,t)\Big|_{t=0} = \frac{\mathrm{d}^2}{\mathrm{d}t^2} R(x,tu)|_{t=0}$, and it is known that $\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Gamma(x,u,t)\Big|_{t=0} \in \mathrm{N}_{\mathcal{M}}(x)$ (see definitions VII.5.1 and VII.2.2 in [**?**]).

The condition is necessary. For all t sufficiently close to zero, we can write

$$R(x,tu) = x + h(t) + v(h(t)), \tag{2.6}$$

where $h(t) \in \mathrm{T}_{\mathcal{M}}(x)$ and $v : \mathrm{T}_{\mathcal{M}}(x) \to \mathrm{N}_{\mathcal{M}}(x)$ is the function mentioned in Lemma 4.7 (or see [**?**, Th. 3.4]) satisfying $\mathrm{D}v(0) = 0$. Differentiating (2.6), we have $\frac{\mathrm{d}}{\mathrm{d}t} R(x,tu) =$

$h'(t) + \mathrm{D}v(h(t))(h'(t))$, where the first term is in $\mathrm{T}_{\mathcal{M}}(x)$ and the second in $\mathrm{N}_{\mathcal{M}}(x)$. Since $\frac{\mathrm{d}}{\mathrm{d}t}R(x,tu)|_{t=0} = u \in \mathrm{T}_{\mathcal{M}}(x)$, it follows that $h'(0) = u$. Differentiating (2.6) again, we have

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}R(x,tu) = h''(t) + \mathrm{D}^2v(h(t))(h'(t),h'(t)) + \mathrm{D}v(h(t))(h''(t))$$

and thus

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}R(x,tu)|_{t=0} = h''(0) + \mathrm{D}^2v(0)(u,u) + \mathrm{D}v(0)(h''(0)) = h''(0) + \mathrm{D}^2v(0)(u,u),$$

where again the first term is in $\mathrm{T}_{\mathcal{M}}(x)$ and the second in $\mathrm{N}_{\mathcal{M}}(x)$. Since R is assumed to be second order, i.e., (2.3), we have finally $\frac{\mathrm{d}^2}{\mathrm{d}t^2}R(x,tu)|_{t=0} = \mathrm{D}^2v(0)(u,u)$, where h no longer appears. The same reasoning applies to $\Gamma(x,u,t)$ and yields $\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Gamma(x,u,t)|_{t=0} = \mathrm{D}^2v(0)(u,u)$. Hence $\frac{\mathrm{d}^2}{\mathrm{d}t^2}R(x,tu)|_{t=0} = \frac{\mathrm{d}^2}{\mathrm{d}t^2}\Gamma(x,u,t)|_{t=0}$. Thus the functions $t \mapsto R(x,tu)$ and $t \mapsto \Gamma(x,u,t)$ coincide up to and including the second order at $t = 0$, that is, (2.4) holds. Property (2.5) follows from Taylor's theorem. □

3. Projective retractions. In this section, we analyze a particular retraction, based on the projection onto $\mathcal{M}$. This retraction will then be generalized in Section 4.

3.1. Projection, smoothness and retraction. The *projection* of a point $x \in \mathcal{E}$ onto a set $M \subset \mathcal{E}$ is the set of points of $\mathcal{E}$ that minimize the distance of x to M, that is

$$\mathrm{P}_M(x) := \operatorname{argmin}\{\|x - y\| : y \in M\}. \tag{3.1}$$

If the set M is closed, then the so-defined projection of $x \in \mathcal{E}$ exists ($\mathrm{P}_M(x) \neq \emptyset$); but it may not reduce to one element. If M is closed and convex, then the projection of any $x \in \mathcal{E}$ exists and is unique [**?**], and this property characterizes closed convex sets (this is known as Motzkin's characterization of closed convex sets, see, e.g., [**?**, §7.5]).

In our situation $M = \mathcal{M}$ is a C^k-submanifold of $\mathcal{E}$, and if we assume furthermore that $\mathcal{M}$ is the boundary of a closed convex set, then the projection is thus uniquely defined by (3.1) for x exterior to this convex set. In this case, we have moreover that the mapping $\mathrm{P}_{\mathcal{M}}$ is of class C^k (see e.g. [**?**] and [**?**]).

For a general manifold $\mathcal{M}$, existence, uniqueness and smoothness still hold locally without any further assumption. The local existence is natural since a manifold is always locally closed. The local uniqueness and the smoothness without the convexity assumption have also a geometric appeal. This is stated in the following well-known lemma (which is proved in, e.g., [**?**]); we give here a different proof which is shorter and that will be generalized later. Note that, if $p \in \mathrm{P}_{\mathcal{M}}(x)$, then p satisfies the (necessary) optimality conditions of the minimization problem of (3.1), namely

$$p \in \mathcal{M}, \qquad x - p \in \mathrm{N}_{\mathcal{M}}(p). \tag{3.2}$$

LEMMA 3.1 (Projection onto a manifold). *Let $\mathcal{M}$ be a submanifold of $\mathcal{E}$ of class C^k around $\bar{x} \in \mathcal{M}$, and let $\mathrm{P}_{\mathcal{M}}$ be the projection onto $\mathcal{M}$ (3.1). Then $\mathrm{P}_{\mathcal{M}}$ is a well-defined function $\mathrm{P}_{\mathcal{M}} \colon \mathcal{E} \to \mathcal{M}$ around $\bar{x}$. Moreover, this function $\mathrm{P}_{\mathcal{M}}$ is of class C^{k-1} around $\bar{x}$ and*

$$\mathrm{DP}_{\mathcal{M}}(\bar{x}) = \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(\bar{x})}.$$

Proof. We first notice that the tangent space at $(\bar{x}, 0)$ of the manifold $\mathrm{N}\mathcal{M}$ is

$$\mathrm{T}_{\mathrm{N}\mathcal{M}}(\bar{x}, 0) = \mathrm{T}_{\mathcal{M}}(\bar{x}) \times \mathrm{N}_{\mathcal{M}}(\bar{x}). \tag{3.3}$$

Here is a quick proof of this fact. Let $t \mapsto (x(t), v(t))$ be a smooth curve in $\mathrm{N}\mathcal{M}$ with $x(0) = \bar{x}$ and $v(0) = 0$. One has $v(t) = \mathrm{P}_{\mathrm{N}_{\mathcal{M}}(x(t))} v(t)$. Note that the orthogonal projection onto $\mathrm{N}_{\mathcal{M}}(x)$ can be viewed as a matrix-valued function of class C^{k-1}. Differentiating both sides at $t = 0$ and using the product rule yields that $v'(0) = \mathrm{P}_{\mathrm{N}_{\mathcal{M}}(\bar{x})} v'(0)$, hence $v'(0) \in \mathrm{N}_{\mathcal{M}}(\bar{x})$. The relation (3.3) follows.

So we start the core of the proof of the lemma, by considering

$$F\colon \begin{cases} \mathrm{N}\mathcal{M} & \longrightarrow \mathcal{E} \\ (x, v) & \longmapsto x + v \end{cases}$$

which is of class C^{k-1} (as is $\mathrm{N}\mathcal{M}$). Its derivative at $(\bar{x}, 0) \in \mathrm{N}\mathcal{M}$ is

$$\mathrm{D}F(\bar{x}, 0)\colon \begin{cases} \mathrm{T}_{\mathcal{M}}(\bar{x}) \times \mathrm{N}_{\mathcal{M}}(\bar{x}) & \longrightarrow \mathcal{E} \\ (u, v) & \longmapsto u + v. \end{cases}$$

This derivative is invertible with

$$\forall\, h \in \mathcal{E}, \qquad [\mathrm{D}F(\bar{x}, 0)]^{-1}(h) = \big(\mathrm{P}_{\mathrm{T}_{\mathcal{M}}(\bar{x})}(h), \mathrm{P}_{\mathrm{N}_{\mathcal{M}}(\bar{x})}(h)\big).$$

Then the local inverse theorem for manifolds yields that there are two neighborhoods ($\mathcal{U}$ of $(\bar{x}, 0)$ in $\mathrm{N}\mathcal{M}$ and $\mathcal{V}$ of $F(\bar{x}, 0) = \bar{x}$ in $\mathcal{E}$) such that $F\colon \mathcal{U} \to \mathcal{V}$ is a C^{k-1} diffeomorphism, and

$$\forall\, h \in \mathcal{E}, \qquad \mathrm{D}F^{-1}(\bar{x})(h) = \big(\mathrm{P}_{\mathrm{T}_{\mathcal{M}}(\bar{x})}(h), \mathrm{P}_{\mathrm{N}_{\mathcal{M}}(\bar{x})}(h)\big).$$

We show that the projection exists locally. Specifically, we show that, for all $\bar{x} \in \mathcal{M}$, there exists $\delta_{\mathrm{e}} > 0$ such that, for all $x \in B(\bar{x}, \delta_{\mathrm{e}})$, $\mathrm{P}_{\mathcal{M}}(x)$ is nonempty. Since $\mathcal{M}$ is a submanifold of $\mathcal{E}$, it can be shown that there exists $\delta_{\mathrm{e}} > 0$ such that $\mathcal{M} \cap \bar{B}(\bar{x}, 2\delta_{\mathrm{e}})$ is compact. Hence, for all $x \in B(\bar{x}, \delta_{\mathrm{e}})$, $\mathrm{P}_{\mathcal{M}}(x) = \mathrm{P}_{\mathcal{M} \cap \bar{B}(\bar{x}, 2\delta_{\mathrm{e}})}(x)$, which is nonempty as the projection onto a compact set.

We show that the projection is locally unique. Specifically, we show that, for all $\bar{x} \in \mathcal{M}$, there exists $\delta_{\mathrm{u}} > 0$ such that, for all $x \in B(\bar{x}, \delta_{\mathrm{u}})$, $\mathrm{P}_{\mathcal{M}}(x)$ contains no more than one point. Choose $\epsilon_1 > 0$ and $\epsilon_2 > 0$ such that $\mathcal{U}' := \{(y, u) \in \mathrm{N}_{\mathcal{M}} : y \in B(\bar{x}, \epsilon_1) \cap \mathcal{M}, \|u\| < \epsilon_2\} \subset \mathcal{U}$. Let $\mathcal{V}' = F(\mathcal{U}')$. Observe that F is a bijection from $\mathcal{U}'$ to $\mathcal{V}'$, and is thus injective on $\mathcal{U}'$. Choose $\delta_{\mathrm{u}} > 0$ such that $B(\bar{x}, 3\delta_{\mathrm{u}}) \subset \mathcal{V}'$, with $2\delta_{\mathrm{u}} < \epsilon_1$ and $3\delta_{\mathrm{u}} < \epsilon_2$. Let $x \in B(\bar{x}, \delta_{\mathrm{u}})$ and, for contradiction, assume that there exist two different points p_1 and p_2 in $\mathrm{P}_{\mathcal{M}}(x)$. Then $p_1 \in B(\bar{x}, 2\delta_{\mathrm{u}}) \subset B(\bar{x}, \epsilon_1)$, and $\|x - p_1\| \leq \|x - \bar{x}\| + \|\bar{x} - p_1\| \leq \delta_{\mathrm{u}} + 2\delta_{\mathrm{u}} = 3\delta_{\mathrm{u}} < \epsilon_2$. Thus $(p_1, x - p_1) \in \mathcal{U}'$, and moreover $F(p_1, x - p_1) = p_1 + (x - p_1) = x$. The same reasoning leads to the same conclusion for p_2. Thus, we have $F(p_1, x - p_1) = F(p_2, x - p_2)$, $(p_1, x - p_1) \in \mathcal{U}'$, $(p_2, x - p_2) \in \mathcal{U}'$, and $p_1 \neq p_2$, in contradiction with the fact that F is injective on $\mathcal{U}'$.

Finally, we show the differential properties of $\mathrm{P}_{\mathcal{M}}$. Let $\delta = \min\{\delta_{\mathrm{e}}, \delta_{\mathrm{u}}\}$. Observe that, for all $x \in B(\bar{x}, \delta)$, $\mathrm{P}_{\mathcal{M}}(x)$ is characterized by $(\mathrm{P}_{\mathcal{M}}(x), x - \mathrm{P}_{\mathcal{M}}(x)) \in \mathcal{U} \subset \mathrm{N}\mathcal{M}$. Therefore $(\mathrm{P}_{\mathcal{M}}(x), x - \mathrm{P}_{\mathcal{M}}(x)) = F^{-1}(x)$. Introducing the C^{k-1} function $\pi\colon \mathrm{N}\mathcal{M} \to \mathcal{M}, (x, v) \mapsto x$, we have that $\mathrm{P}_{\mathcal{M}} = \pi \circ F^{-1}$ is C^{k-1} on $B(\bar{x}, \delta)$, and that

$$\mathrm{D}\mathrm{P}_{\mathcal{M}}(\bar{x}) = \mathrm{D}\pi(\bar{x}, 0) \circ \mathrm{D}F^{-1}(\bar{x}) = \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(\bar{x})}\ ,$$

which completes the proof. □

We note that even if the above proof resembles the one of proposition 4.1.2 in [**?**], the formulation of that proposition is not suitable for tackling Lemma 3.1. The next proposition defines the projective retraction.

PROPOSITION 3.2 (Projective retraction). *Let $\mathcal{M}$ be a submanifold of $\mathcal{E}$ of class C^k around $\bar{x} \in \mathcal{M}$. Then the function R*

$$R\colon \begin{cases} \mathrm{T}\mathcal{M} \longrightarrow \mathcal{M} \\ (x,u) \longmapsto \mathrm{P}_{\mathcal{M}}(x+u) \end{cases} \tag{3.4}$$

is a retraction around $\bar{x}$.

Proof. Consider the C^k mapping

$$G\colon \begin{cases} \mathrm{T}\mathcal{M} \longrightarrow \mathcal{E} \\ (x,u) \longmapsto x+u. \end{cases}$$

Then $R = \mathrm{P}_{\mathcal{M}} \circ G$, and Lemma 3.1 yields that R has the desired smoothness property. Moreover using the chain rule, we get for all $u \in \mathrm{T}_{\mathcal{M}}(x)$,

$$\mathrm{D}R(x,\cdot)(0)u = \mathrm{D}\mathrm{P}_{\mathcal{M}}(x)u = \mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}u = u,$$

where the first equality comes from (3.4), the second from Lemma 3.1, and the third by the fact that u is already in $\mathrm{T}_{\mathcal{M}}(x)$. □

The retraction defined in Proposition 3.2 is moreover a second-order retraction, as it will be established from general results in Section 4.

Thus the projection onto $\mathcal{M}$ is in theory locally well-defined and smooth, and gives rise to a retraction on $\mathcal{M}$. It is also computable in many cases, especially for some matrix manifolds. We finish the section with a few examples of such matrix manifolds. Though some results of Sections 3.2 and 3.3 are known, we deliberately give some details because, first, we are not aware of precise references on them, and second, the arguments are generalized to prove the new reults of Section 3.4.

3.2. Projection onto fixed-rank matrices. Routine calculations show that the set of matrices with fixed rank r,

$$\mathcal{R}_r = \{X \in \mathbb{R}^{n\times m} : \mathrm{rank}(X) = r\},$$

is a smooth submanifold of $\mathbb{R}^{n\times m}$ (see the proof of the special case of symmetric matrices in [**?**, Ch. 5, Prop. 1.1]). Recall that a singular value decomposition of $X \in \mathbb{R}^{n\times m}$ is written

$$X = U\Sigma V^\top, \tag{3.5}$$

where the two matrices $U = [u_1, u_2, \ldots, u_n]$ and $V = [v_1, v_2, \ldots, v_m]$ are orthogonal matrices, and the only non-zeros entries in the matrix Σ are on the diagonal (the singular values of X) written in the nonincreasing order

$$\sigma_1(X) \geq \sigma_2(X) \geq \cdots \geq \sigma_{\min\{n,m\}}(X) \geq 0.$$

We now equip the space $\mathbb{R}^{n\times m}$ with the inner product associated to the canonical basis. Its associated norm is the Frobenius norm, given by

$$\|X\|^2 = \sum_{i,j} {X_{ij}}^2 = \mathrm{trace}(X^\top X) = \sum_{i=1}^{\min\{n,m\}} \sigma_i(X)^2. \tag{3.6}$$

It is well-known (see [**?**]) that the singular value decomposition gives an easy way to project a matrix $X \in \mathbb{R}^{n\times m}$ onto the (closed) set of matrices with rank less than or equal to r (which is an algebraic set). Specifically, a nearest matrix with rank no more than r is

$$\hat{X} = \sum_{i=1}^{r} \sigma_i(X) u_i v_i^\top; \tag{3.7}$$

this is the Eckart-Young result (see [**?**], or [**?**, §7.4.1]). It turns out that this expression also locally gives the expression of the projection onto the set $\mathcal{R}_r$ of the matrices with rank *equal* to r. We formalize this easy result in the next proposition. By Proposition 3.2, this gives an easy-to-compute retraction on $\mathcal{R}_r$.

PROPOSITION 3.3 (Projection onto the manifold of fixed-rank matrices). *Let $\bar{X} \in \mathcal{R}_r$; for any X such that $\|X - \bar{X}\| < \sigma_r(\bar{X})/2$, the projection of X onto $\mathcal{R}_r$ exists, is unique, and has the expression*

$$\mathrm{P}_{\mathcal{R}_r}(X) = \sum_{i=1}^{r} \sigma_i(X) u_i v_i^\top,$$

given by a singular value decomposition (3.5) *of X.*

Proof. The result comes easily from the projection (3.7) onto the set of matrices with rank lower or equal than r. Let X be such that $\|X - \bar{X}\| < \sigma_r(\bar{X})$; we just have to prove that $\hat{X} = \mathrm{P}_{\mathcal{R}_r}(X)$ in this situation, and that the projection is unique. We start by noting that [**?**, 7.3.8] yields $|\sigma_i(\bar{X}) - \sigma_i(X)| \leq \|X - \bar{X}\| < \sigma_r(\bar{X})/2$ for all i, and then in particular

$$\sigma_{r+1}(X) < \sigma_r(\bar{X})/2 < \sigma_r(X). \tag{3.8}$$

Evoking uniqueness of the singular values, observe now that

$$\sigma_i(\hat{X}) = \begin{cases} \sigma_i(X) & \text{if } \ i \leq r \\ 0 & \text{otherwise.} \end{cases}$$

From (3.8), we have $\sigma_r(X) > 0$ and therefore $\hat{X}$ is of rank r. Thus we have

$$\|\hat{X} - X\| \geq \min_{Y \in \mathcal{R}_r} \|Y - X\| \geq \min_{\mathrm{rank}(Y) \leq r} \|Y - X\| = \|\hat{X} - X\|.$$

This shows that we have equalities in the above expression, and thus $\hat{X} \in \mathrm{P}_{\mathcal{R}_r}(X)$. Finally, uniqueness comes from the uniqueness of the projection onto $\mathrm{rank}(Y) \leq r$ (see [**?**, §5.1 Cor. 1.17] or [**?**]) under the condition $\sigma_r(X) > \sigma_{r+1}(X)$ which is here guaranteed by (3.8). □

Since, as mentioned in the introduction, the concept of retraction was introduced to formalize the use of computationally efficient approximations of the Riemannian exponential, it is worth checking if the projective retraction (3.4) offers a computational advantage. In the case of the manifold $\mathcal{R}_r$, the advantage is clear: the retraction by projection admits a closed-form expression based on the truncated singular value decomposition, whereas the Riemannian exponential, with respect to the Riemannian metric inherited from the embedding of $\mathcal{R}_r$ in the Euclidean space $\mathbb{R}^{n\times m}$, involves an ODE for which we do not have an analytical solution in general, even if we restrict to symmetric matrices; see [**?**, §3]. That being said, in the symmetric case, it has recently been shown that a closed-form expression for the exponential can be obtained for at least one choice of the Riemannian metric; see [**?**] for details.

3.3. Projection onto Stiefel manifolds. The manifold of orthonormal m-frames in $\mathbb{R}^n$ ($m \leq n$), introduced by Eduard Stiefel in 1935 to solve a topological problem (see e.g. [**?**, Chap.IV]), is defined by

$$V_{n,m} := \left\{ X \in \mathbb{R}^{n \times m} : X^\top X = I_m \right\}.$$

For example, the manifold $V_{n,n}$ is simply the group of orthogonal matrices of size n.

The projection onto $V_{n,m}$ turns out to be explicit through the singular value decomposition as well. This result is mentioned without proof in [**?**, §4]; we formalize it in the following proposition. By Proposition 3.2, this gives an easy-to-compute retraction on $V_{n,m}$.

PROPOSITION 3.4 (Projection onto Stiefel manifolds). *Let $\bar{X} \in V_{n,m}$; for any X such that $\|X - \bar{X}\| < \sigma_m(\bar{X})$, the projection of X onto $V_{n,m}$ exists, is unique, and has the expression*

$$\mathrm{P}_{\mathcal{R}_r}(X) = \sum_{i=1}^{m} u_i v_i^\top,$$

given by a singular value decomposition (3.5) *of X. In other words, it is the W of the polar decomposition $X = WS$ (the product of $W \in V_{n,m}$ and a symmetric positive-definite matrix $S \in \mathbb{R}^{m \times m}$).*

Proof. The proof comes following the same lines as [**?**, 7.4.6]. For all $Y \in V_{n,m}$, we have $\|X - Y\|^2 = \|X\|^2 + m^2 - 2\,\mathrm{trace}(Y^\top X)$ and we can bound the last term, as follows. Note that $V_{n,m}$ is invariant by pre- and post-multiplication by orthogonal matrices, so we can write

$$\begin{aligned}\max_{Y \in V_{n,m}} \mathrm{trace}(Y^\top X) &= \max_{Y \in V_{n,m}} \mathrm{trace}\left(\left(U^\top Y V\right)^\top \Sigma \right) \\ &= \max_{Y \in V_{n,m}} \mathrm{trace}(Y^\top \Sigma) = \max_{Y \in V_{n,m}} \sum_{i=1}^{m} Y_{ii}\sigma_i \ \leq \sum_{i=1}^{m} \sigma_i.\end{aligned}$$

The inequality comes from the fact that the columns of Y are of norm 1 which implies $Y_{ii} \leq 1$. Moreover the bound is attained by $Y = UV^\top$, so this matrix is a projection. Finally the same arguments as in the beginning of the proof of Proposition 3.3 give that X is full rank, so the polar form is unique [**?**, 7.3.2], and so is the projection. □

Is the projective retraction a valuable alternative to the Riemannian exponential on Stiefel manifolds? Here again, the projection is based on the singular value decompositon. As for the Riemannian exponential with respect to the metric inherited from the embedding of $V_{n,m}$ in $\mathbb{R}^{n \times m}$, its expression given in [**?**, §2.2.2] (or see [**?**, example 5.4.2]) requires the computation of an $m \times m$ matrix exponential, a $2m \times 2m$ matrix exponential, and matrix multiplications, the most expensive one being an $n \times 2m$ matrix multiplied by a $2m \times 2m$ matrix. If $n \gg m \gg 1$, the dominant term in the flop count for the Riemannian exponential is $2n(2m)^2 = 8nm^2$, due to the most expensive matrix multiplication; whereas for the projective retraction, the dominant term in the flop count, which comes from the computation of the singular value decompositon of a matrix $A \in \mathbb{R}^{n \times m}$, can be as low as $2nm^2 + 2nm^2 = 4nm^2$, where one of the terms is the cost of forming $A^T A$, and the other term comes from a multiplication of the form AV to recover the U factor of the singular value decompositon. However, other algorithms, more costly but more robust, can be preferred to compute the singular value decompositon. Moreover, the assumption $n \gg m \gg 1$ may not be in force.

Finally, when comparing two retractions, one must keep in mind that the choice of the retraction in a retraction-based algorithm (such as Newton [**?**] or trust region [**?**]) may affect the number of iterates needed to reach a certain accuracy; in this respect, the Riemannian exponential need not be the best choice. Hence there is no clear and systematic winner between the exponential retraction and the projective retraction in the case of the Stiefel manifold, but it is worth keeping in mind the existence of the projective retraction as a potentially more efficient alternative to the exponential retraction.

3.4. Projection onto spectral manifolds. We study in this section a class of matrix manifolds for which the projection admits an explicit expression. Those submanifolds of the space of symmetric matrices are called "spectral" since they are defined by properties of the eigenvalues of the matrix. The projection onto spectral manifolds comes through an eigendecomposition of the matrix to project, in a same way as in previous sections, the projections came through singular value decomposition.

We start with some notation. By $\mathbf{S}_n$, $\mathbf{O}_n$, $\mathbf{\Sigma}_n$ and $\mathbb{R}^n_\downarrow$, we denote respectively the space of $n \times n$ symmetric matrices, the group of $n \times n$ orthogonal matrices, its subgroup of permutation matrices, and the subset of $\mathbb{R}^n$ such that

$$x = (x_1, \ldots, x_n) \in \mathbb{R}^n_\downarrow \iff x_1 \geq x_2 \geq \cdots \geq x_n.$$

For $X \in \mathbf{S}_n$, by $\lambda(X) \in \mathbb{R}^n_\downarrow$ we denote the vector of eigenvalues of X in nonincreasing order:

$$\lambda_1(X) \geq \lambda_2(X) \geq \cdots \geq \lambda_n(X).$$

As for (3.6), we now equip the space $\mathbf{S}_n$ with the usual Frobenius inner product $\langle X, Y \rangle = \operatorname{trace}(XY)$, whose associated norm, termed the Frobenius norm, is given by

$$\|X\|^2 = \sum_{i,j=1}^{n} {X_{ij}}^2 = \operatorname{trace}(X^2) = \sum_{i=1}^{n} \lambda_i(X)^2.$$

For simplicity, we keep the same symbol $\|\cdot\|$ for both the Frobenius norm in $\mathbf{S}_n$ and the Euclidean norm in $\mathbb{R}^n$; thus there holds

$$\|\lambda(X)\| = \|X\|. \tag{3.9}$$

The important inequality involving the two inner products is the following (see [**?**] for example): for all X and Y in $\mathbf{S}_n$

$$\langle X, Y \rangle = \operatorname{trace}(XY) \leq \lambda(X)^\top \lambda(Y). \tag{3.10}$$

This implies in particular that the mapping $\lambda \colon \mathbf{S}_n \to \mathbb{R}^n_\downarrow$ is 1-lipschitz, namely there holds

$$\|\lambda(X) - \lambda(Y)\| \leq \|X - Y\|. \tag{3.11}$$

Notice also that there are two invariance properties for the norms: for any $X \in \mathbf{S}_n$ and $U \in \mathbf{O}_n$, we have

$$\|X\| = \|UXU^\top\|; \tag{3.12}$$

besides for any $x \in \mathbb{R}^n$ and $P \in \mathbf{\Sigma}_n$, we have

$$\|x\| = \|Px\|. \tag{3.13}$$

We consider the so-called *spectral sets* of $\mathbf{S}_n$. These are the sets of symmetric matrices defined by properties of their eigenvalues: a spectral set can be written

$$\lambda^{-1}(M) = \Big\{ X \in \mathbf{S}_n : \ \lambda(X) \in M \Big\} = \Big\{ U \operatorname{Diag}(x) U^\top : \ U \in \mathbf{O}_n, x \in M \Big\} \tag{3.14}$$

for an associated subset $M \subset \mathbb{R}^n_\downarrow$. When the underlying subset of $\mathbb{R}^n$ is a smooth manifold with some local symmetry, the associated spectral set inherits this smoothness. This is the content of the main result of [**?**], that we partly recall here.

THEOREM 3.5 (spectral manifolds). *Let $\mathcal{M}$ be a submanifold of $\mathbb{R}^n$ of class C^k with $k = 2$ or ∞. Consider the spectral set $\mathcal{S} = \lambda^{-1}(\mathcal{M} \cap \mathbb{R}^n_\downarrow) \subset \mathbf{S}_n$, let $\bar{X} \in \mathcal{S}$ and set $\bar{x} = \lambda(\bar{X}) \in \mathcal{M}$. Assume that there exists $\delta > 0$ such that $\mathcal{M} \cap \mathrm{B}(\bar{x}, \delta)$ is* strongly locally symmetric*: for any $x \in \mathcal{M} \cap \mathrm{B}(\bar{x}, \delta)$ and for any permutation $P \in \mathbf{\Sigma}_n$ such that $Px = x$, there holds*

$$P\big(\mathcal{M} \cap \mathrm{B}(\bar{x}, \delta)\big) = \mathcal{M} \cap \mathrm{B}(\bar{x}, \delta). \tag{3.15}$$

Then $\mathcal{S}$ is a submanifold of $\mathbf{S}_n$ around $\bar{X}$ of class C^k, whose dimension is related to the one of $\mathcal{M}$.

EXAMPLE 3.6 (Largest eigenvalue). Let $\mathcal{M}_p = \{x \in \mathbb{R}^n : x_1 = \cdots = x_p > x_j \text{ for all } j \geq p+1\}$. Observe that $\mathcal{M}_p$ is a submanifold of $\mathbb{R}^n$ of class C^∞, and that the set $\lambda^{-1}(\mathcal{M}_p \cap \mathbb{R}^n_\downarrow)$ is the subset of symmetric matrices whose largest eigenvalue is of multiplicity p. Moreover, the whole manifold $\mathcal{M}_p$ is strongly locally symmetric (i.e., for all $x \in \mathcal{M}_p$ and all $P \in \mathbf{\Sigma}_n$ such that $Px = x$, we have $P\mathcal{M}_p = \mathcal{M}_p$). It then follows from Theorem 3.5 that $\lambda^{-1}(\mathcal{M}_p \cap \mathbb{R}^n_\downarrow)$ is a submanifold of $\mathbf{S}_n$. □

These spectral manifolds often appear in applications, in particular when using alternating projection methods (see references in the introduction of [**?**]). It turns out indeed that we have an explicit expression of the projection onto these matrix manifolds using eigendecomposition and the projection onto the underlying manifold $\mathcal{M}$. This projection property is in fact even more general for spectral sets. Below we state the projection result onto the spectral sets in Lemma 3.7, then we formalize an intermediate result about permutations (Lemma 3.8) that is used to prove a particular projection result for spectral manifolds (Theorem 3.9). Some of these results generalize or make more explicit those of the appendix of [**?**].

LEMMA 3.7 (Projection onto spectral sets). *Let M be a closed subset of $\mathbb{R}^n_\downarrow$, and $X \in \mathbf{S}_n$ with an eigendecomposition $X = U \operatorname{Diag} \lambda(X) U^\top$ with $U \in \mathbf{O}_n$. Then we have*

$$U \operatorname{Diag}(z)\, U^\top \in \mathrm{P}_{\lambda^{-1}(M)}(X) \iff z \in \mathrm{P}_M(\lambda(X)).$$

Proof. We start by proving the implication "⇐". Let $z \in \mathrm{P}_M(\lambda(X))$ and set

$$\hat{X} = U \operatorname{Diag}(z) U^\top$$

which lies in $\lambda^{-1}(M)$. We write the following inequalities:

$$\begin{aligned}
& \min_{Y\in\lambda^{-1}(M)} \|Y-X\|^2 \\
\geq\ & \min_{Y\in\lambda^{-1}(M)} \|\lambda(Y)-\lambda(X)\|^2 && \text{[by (3.11)]}\\
\geq\ & \min_{y\in M} \|y-\lambda(X)\|^2 && \text{[by (3.14)]}\\
=\ & \|z-\lambda(X)\|^2 && \text{[by definition of } z\text{]}\\
=\ & \|U\operatorname{Diag}(z)U^\top - U\operatorname{Diag}(\lambda(X))U^\top\|^2 && \text{[by (3.12)]}\\
=\ & \|\hat{X}-X\|^2 && \text{[by definition of } U \text{ and } \hat{X}\text{]}\\
\geq\ & \min_{Y\in\lambda^{-1}(M)} \|Y-X\|^2 && \text{[since } \hat{X}\in\lambda^{-1}(M)\text{]}
\end{aligned}$$

These inequalities thus turn out to be all equalities. We have in particular

$$\|\hat{X}-X\|^2 = \min_{Y\in\lambda^{-1}(M)} \|Y-X\|^2 = \min_{y\in M} \|y-\lambda(X)\|^2 = \|z-\lambda(X)\|^2 \tag{3.16}$$

which gives the implication "$\Leftarrow$". The reverse implication "$\Rightarrow$" also follows easily from (3.16). Let $\bar{z}$ such that $\bar{X} = U\operatorname{Diag}(\bar{z})U^\top \in \mathrm{P}_{\lambda^{-1}(M)}(X)$. For any $X\in\lambda^{-1}(M)$, we have by (3.11) and (3.16)

$$\|\bar{z}-\lambda(X)\|^2 \leq \|\bar{X}-X\|^2 = \min_{Y\in\lambda^{-1}(M)} \|Y-X\|^2 = \min_{y\in M} \|y-\lambda(X)\|^2$$

which proves that $\bar{z}\in \mathrm{P}_M(\lambda(X))$. ∎

LEMMA 3.8. *Let $\bar{x}\in\mathbb{R}^n_\downarrow$. Then for all $\delta>0$ small enough, we have that, for any $y\in \mathrm{B}(\bar{x},\delta)$ and $x\in\mathbb{R}^n_\downarrow\cap\mathrm{B}(\bar{x},\delta)$, the maximum of the inner product $x^\top Py$ over the permutations that fix $\bar{x}$,*

$$\max_{P\in\mathbf{\Sigma}_n,\ P\bar{x}=\bar{x}} x^\top Py,$$

is attained when $Py\in\mathbb{R}^n_\downarrow$.

Proof. We define

$$\bar{\delta} := \frac{1}{3}\min\Big\{\bar{x}_i-\bar{x}_{i+1}:\quad i=1,\dots,n\,,\ \text{ such that } \bar{x}_i-\bar{x}_{i+1}>0\Big\}.$$

and we set $0<\delta\leq\bar{\delta}$. For any $z\in\mathbb{R}^n$ such that $\|\bar{x}-z\|\leq\delta$ (and then $|\bar{x}_i-z_i|\leq\delta$ for all i), observe that

$$\bar{x}_i>\bar{x}_{i+1} \implies \forall\, i_1\leq i,\ \forall\, i_2\geq i+1, \qquad z_{i_1}>z_{i_2}. \tag{3.17}$$

Assume now that the maximum of $x^\top Py$ is attained at $Py\notin\mathbb{R}^n_\downarrow$. We just need to show that the maximum is also attained at $Py\in\mathbb{R}^n_\downarrow$. This means that there exist indexes $i\leq j$ such that

$$(Py)_i<(Py)_j \qquad (\text{ and } x_i\geq x_j).$$

Apply (3.17) with $z=Py$; we can do this since

$$\|z-\bar{x}\| = \|Py-\bar{x}\| = \|Py-P\bar{x}\| = \|y-\bar{x}\|$$

by (3.13). This yields $\bar{x}_i=\bar{x}_j$, and thus the permutation P_{ij} that permutes i and j and leaves the other indexes invariant fixes $\bar{x}$ as well. We also have $x^\top Py\leq x^\top(P_{ij}P)y$, since

$$x_i(Py)_j + x_j(Py)_i - \big(x_i(Py)_i + x_j(Py)_j\big) = (x_i-x_j)((Py)_j-(Py)_i)\geq 0$$

In other words we do not reduce $x^\top Py$ by permuting $(Py)_i$ and $(Py)_j$. A finite number of such exchanges leads to a nonincreasing order of the Py, which shows that the maximum is also attained when $Py \in \mathbb{R}^n_\downarrow$, and the proof is complete. □

THEOREM 3.9 (Projection onto spectral manifolds). *Assume the assumptions of Theorem 3.5 are in force. The projection onto the manifold $\mathcal{S}$ of a matrix X such that $\|X - \bar{X}\| \leq \delta/2$ is*

$$\mathrm{P}_{\mathcal{S}}(X) = U \operatorname{Diag}\big(\mathrm{P}_{\mathcal{M}}(\lambda(X))\big)U^\top$$

where $U \in \mathbf{O}_n$ is such that $X = U \operatorname{Diag}(\lambda(X))U^\top$.

Proof. Consider $\bar{X} \in \lambda^{-1}(\mathcal{M})$ and set $\bar{x} = \lambda(\bar{X}) \in \mathbb{R}^n_\downarrow$. Let $X \in \lambda^{-1}(\mathcal{M}) \cap \mathrm{B}(\bar{X}, \delta/2)$, write the spectral decomposition $X = U \operatorname{Diag}(x)U^\top$ with $x = \lambda(X)$. Note that we have $\|x - \bar{x}\| \leq \delta/2$ by (3.11). Restricting δ if necessary, we assume that Lemma 3.1 is enforced. Lemma 3.7 then gives:

$$\mathrm{P}_{\mathcal{S}}(X) = U \operatorname{Diag}\big(\mathrm{P}_{\mathcal{M}\cap\mathbb{R}^n_\downarrow}(\lambda(X))\big)U^\top.$$

We are going to establish

$$\mathrm{P}_{\mathcal{M}}(x) = \mathrm{P}_{\mathcal{M}\cap\mathrm{B}(\bar{x},\delta)}(x) = \mathrm{P}_{\mathcal{M}\cap\mathrm{B}(\bar{x},\delta)\cap\mathbb{R}^n_\downarrow}(x) = \mathrm{P}_{\mathcal{M}\cap\mathbb{R}^n_\downarrow}(x) \tag{3.18}$$

which will allow us to conclude. We note that the first and third equalities are straightforward. For both $M = \mathcal{M}$ and $M = \mathcal{M}\cap\mathbb{R}^n_\downarrow$, we have indeed $\mathrm{P}_M(x) \in \mathrm{B}(\bar{x}, \delta)$, which yields

$$\min_{y\in M} \|x - y\| = \min_{y\in M\cap\mathrm{B}(\bar{x},\delta)} \|x - y\|. \tag{3.19}$$

To see this, note that $\|\mathrm{P}_M(x) - x\| \leq \|\bar{x} - x\|$ by definition of $\mathrm{P}_M(x)$, which allows to write

$$\|\mathrm{P}_M(x) - \bar{x}\| \leq \|\mathrm{P}_M(x) - x\| + \|x - \bar{x}\| \leq 2\|x - \bar{x}\| \leq \delta,$$

and therefore (3.19). We now prove that

$$\min_{y\in\mathcal{M}\cap\mathrm{B}(\bar{x},\delta)} \|x - y\| = \min_{y\in\mathcal{M}\cap\mathrm{B}(\bar{x},\delta)\cap\mathbb{R}^n_\downarrow} \|x - y\|. \tag{3.20}$$

To do so, we exploit (partly) the symmetry property of $\mathcal{M}$ of Theorem 3.5: notice indeed that in particular for any $P \in \mathbf{\Sigma}_n$ such that $P\bar{x} = \bar{x}$, we have

$$P\big(\mathcal{M} \cap \mathrm{B}(\bar{x}, \delta)\big) = \mathcal{M} \cap \mathrm{B}(\bar{x}, \delta). \tag{3.21}$$

Then we can develop the following equalities

$$\begin{aligned}
& \min_{y\in\mathcal{M}\cap\mathrm{B}(\bar{x},\delta)} \|y - x\|^2 \\
= \; & \min_{y\in\mathcal{M}\cap\mathrm{B}(\bar{x},\delta),\ P\in\mathbf{\Sigma}_n,\ P\bar{x}=\bar{x}} \|Py - x\|^2 && \text{[by (3.21)]} \\
= \; & \min_{y\in\mathcal{M}\cap\mathrm{B}(\bar{x},\delta),\ P\in\mathbf{\Sigma}_n,\ P\bar{x}=\bar{x}} (\|Py\|^2 + \|x\|^2 - 2x^\top Py) \\
= \; & \min_{y\in\mathcal{M}\cap\mathrm{B}(\bar{x},\delta)} (\|y\|^2 + \|x\|^2 - 2\max_{P\in\mathbf{\Sigma}_n,\ P\bar{x}=\bar{x}} x^\top Py) && \text{[by (3.13)]} \\
= \; & \min_{z\in\mathcal{M}\cap\mathrm{B}(\bar{x},\delta)\cap\mathbb{R}^n_\downarrow} (\|P^{-1}z\|^2 + \|x\|^2 - 2x^\top z) && \text{[by Lemma 3.8 and } z = Py\text{]} \\
= \; & \min_{z\in\mathcal{M}\cap\mathrm{B}(\bar{x},\delta)\cap\mathbb{R}^n_\downarrow} \|z - x\|^2 && \text{[by (3.13)]}
\end{aligned}$$

Putting together (3.19) and (3.20) proves (3.18). Lemma 3.7 then allows us to conclude. ∎

EXAMPLE 3.10 (Projection onto largest eigenvalue). We continue Example 3.6. We can see that the projection $\hat{x}$ of a vector $x \in \mathbb{R}^n_\downarrow$ onto $\mathcal{M}_p \cap \mathbb{R}^n_\downarrow \cap \mathrm{B}(\bar{x}, \delta)$ is defined by $\hat{x}_i = x_i$ if $i > p$ and

$$\hat{x}_i = \alpha := \frac{1}{p} \sum_{\ell=1}^{p} x_\ell$$

Thus the projection of X close to $\bar{X}$ onto $\lambda^{-1}(\mathcal{M}_p)$ is

$$\mathrm{P}_{\lambda^{-1}(\mathcal{M}_p)}(X) = U \,\mathrm{Diag}(\alpha, \ldots, \alpha, x_{p+1}, \ldots, x_n) U^\top.$$

This completes the partial result of [**?**, Th. 13]. □

We are not aware of efficient ways of computing the Riemannian exponential on general spectral manifolds, hence the projective retraction comes as a valuable alternative.

4. Projection-like retractions. Inspired by the projective retractions of the previous section, we define and study here new retractions constructed from projection-like operations. These operations use a *retractor* prescribing admissible directions to get back to the manifold. We show that moving tangentially and then along these admissible directions produces a retraction, and even a second-order retraction if some orthogonality holds.

4.1. Retraction by retractors. As before, we consider a d-dimensional submanifold $\mathcal{M}$ of an n-dimensional Euclidean space $\mathcal{E}$. Let $\mathrm{Gr}(n-d, \mathcal{E})$ denote the set of all $(n-d)$-dimensional linear subspaces of $\mathcal{E}$. This set of subspaces admits a natural differentiable structure, endowed with which it is termed the *Grassmann manifold* of $(n-d)$-planes in $\mathcal{E}$; see e.g. [**?**, **?**] for details. Recall that two subspaces of dimension d and $n-d$ of $\mathcal{E}$ are *transverse* if their intersection is trivial, or equivalently, if $\mathcal{E}$ is their direct sum. The next definition is illustrated on Figure 4.1.

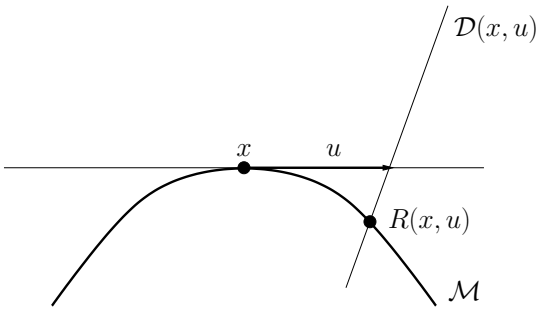

FIG. 4.1. *Illustration of the concept of a retractor.*

DEFINITION 4.1 (Retractor). *Let $\mathcal{M}$ be a d-dimensional submanifold of class C^k (with $k \geq 2$) of an n-dimensional Euclidean space $\mathcal{E}$. A* retractor *on $\mathcal{M}$ is a C^{k-1} mapping D from the tangent bundle $\mathrm{T}\mathcal{M}$ into the Grassmann manifold $\mathrm{Gr}(n-d, \mathcal{E})$, whose domain contains a neighborhood of the zero section of $\mathrm{T}\mathcal{M}$, and such that, for all $x \in \mathcal{M}$, $D(x, 0)$ is transverse to $\mathrm{T}_{\mathcal{M}}(x)$.*

The choice of the term *retractor* in the above definition is justified by the result below, whose proof is given in Section 4.2.

THEOREM 4.2 (Retractors give retractions). *Let D be a retractor as defined above and, for all $(x,u) \in \mathrm{dom}(D)$, define the affine space $\mathcal{D}(x,u) = x + u + D(x,u)$. Consider the point-to-set function $R : \mathrm{dom}(D) \rightrightarrows \mathcal{M}$ such that $R(x,u)$ is the set of points of $\mathcal{M} \cap \mathcal{D}(x,u)$ nearest to $x+u$. Then R is a retraction on $\mathcal{M}$. (In particular, for all $\bar{x} \in \mathcal{M}$, there is a neighborhood of $(\bar{x},0)$ in $\mathrm{T}\mathcal{M}$ on which R maps to singletons.) The retraction R thus defined is called the retraction* induced *by the retractor D.*

EXAMPLE 4.3 (Orthographic retractor). A simple and important example of retractor is the constant mapping defined for $(x,u) \in \mathrm{T}\mathcal{M}$ by

$$D(x,u) = \mathrm{N}_{\mathcal{M}}(x). \tag{4.1}$$

Theorem 4.2 implies that this retractor does induce a retraction. We call this retraction the *orthographic retraction*, since on a sphere it relates to the orthographic projection known in cartography. □

REMARK 4.4. In [**?**, §3], a particular family of parameterizations of $\mathcal{M}$ is introduced and called *tangential parameterization*. Tangential parameterizations yield the orthographic retraction through the procedure given in [**?**, §4.1.3]. The result [**?**, Th. 3.4] implies that properties 2 and 3 of retractions (see Definition 2.1) hold for the orthographic retraction, but it does not imply the smoothness property 1, whereas Theorem 4.2 does. More precisely, Lemma 4.7, which is Theorem 4.2 for the particular case (4.1), goes beyond [**?**, Th. 3.4] by showing that v is a C^{k-1} function, not only with respect to its u variable but with respect to (x,u). The core difficulty is that, whereas $u \mapsto v(x,u)$ is a function between two vector spaces, $(x,u) \mapsto v(x,u)$ is a function between nonlinear manifolds. □

EXAMPLE 4.5 (Projective retractor). The projective retraction (3.4), given by $R(x,u) = \mathrm{P}_{\mathcal{M}}(x+u)$, fits in the framework of retractors: it is induced by the *projective retractor* defined implicitly, for $(x,u) \in \mathrm{T}\mathcal{M}$, by

$$D(x,u) = \mathrm{N}_{\mathcal{M}}(\mathrm{P}_{\mathcal{M}}(x+u)). \tag{4.2}$$

The mapping D is a bona-fide retractor since, for small u, this mapping is smooth in view of Lemma 3.1, and moreover we have $D(x,0) = \mathrm{N}_{\mathcal{M}}(\mathrm{P}_{\mathcal{M}}(x)) = \mathrm{N}_{\mathcal{M}}(x)$ transverse to $\mathrm{T}_{\mathcal{M}}(x)$. □

EXAMPLE 4.6 (Sphere and manifolds of co-dimension 1). When $\mathcal{M}$ is a sphere, these results on retractors show that several types of cartographic projections lead to retractions: the gnomonic projection (which is the classical, nearest-point projection on the sphere), the orthographic projection, the stereographic projection. Roughly speaking, all projections into the tangent plane with a point of perspective (possibly at infinity) lead to retractions, provided that the point of perspective changes smoothly with the reference point and that it does not belong to the tangent space. In a way, retractors generalize to general manifolds these cartographic retractions, initially quite specific to the sphere and to submanifolds of co-dimension one. □

4.2. Proof of the main result. To prove Theorem 4.2, we proceed in two steps. We first prove the theorem for the normal case $D(x,u) = \mathrm{N}_{\mathcal{M}}(x)$ (Lemma 4.7), then we deduce the general case by smoothly "straightening" the space $D(x,u)$ onto $\mathrm{N}_{\mathcal{M}}(x)$ (Lemma 4.8).

LEMMA 4.7 (Special case $D(x,u) = \mathrm{N}_{\mathcal{M}}(x)$). *Let $\mathcal{M}$ be a submanifold of class C^k ($k \geq 2$) of an n-dimensional Euclidean space $\mathcal{E}$. For all $\bar{x} \in \mathcal{M}$, there exists a*

neighborhood $\mathcal{U}_{\mathrm{T}\mathcal{M}}$ *of* $(\bar{x},0)$ *in* $\mathrm{T}\mathcal{M}$ *such that, for all* $(x,u)\in\mathcal{U}_{\mathrm{T}\mathcal{M}}$*, there is one and only one smallest* $v\in\mathrm{N}_{\mathcal{M}}(x)$ *such that* $x+u+v\in\mathcal{M}$*. Call it* $v(x,u)$ *and define*

$$R(x,u)=x+u+v(x,u).$$

We have $\mathrm{D}_u v(x,0)=0$ *and thus* R *defines a retraction around* $\bar{x}$*. Since the expression of* R *does not depend on* $\bar{x}$ *or* $\mathcal{U}_{\mathrm{T}\mathcal{M}}$*,* R *defines a retraction on* $\mathcal{M}$*.*

Establishing this result consists essentially in applying the implicit function theorem, but to prove it rigorously we have to use charts and resort to technicalities. We do not skip any argument since a similar rationale will be used in the forthcoming proof of Theorem 4.2.

Proof. Let $\bar{x}\in\mathcal{M}$ and let $\Phi:\mathcal{U}_\Phi\to\mathbb{R}^{n-d}$ of class C^k be a local equation of $\mathcal{M}$ around $\bar{x}$, i.e., $\mathcal{U}_\Phi$ is a neighborhood of $\bar{x}$ in $\mathcal{E}$ and

$$\Phi(x)=0 \iff x\in\mathcal{M}\cap\mathcal{U}_\Phi,$$

with $\mathrm{D}\Phi(x)$ of full rank $n-d$ for all $x\in\mathcal{U}_\Phi$. Define

$$\Psi:\mathrm{B}\mathcal{M}\to\mathbb{R}^{n-d}:(x,u,v)\mapsto\Phi(x+u+v).$$

We have that $\mathrm{D}_v\Psi(\bar{x},0,0)=\mathrm{D}\Phi(\bar{x})$ is surjective on $\mathrm{N}_{\mathcal{M}}(\bar{x})$ onto $\mathbb{R}^{n-d}$. Thus we would be ready to apply the implicit function theorem, if only Ψ was defined on the Cartesian product of an "(x,u)" space and a "v" space. To remedy this, we read Ψ in a chart. Specifically, let $\theta:\mathcal{U}_{\mathcal{M}}\to\mathbb{R}^d$ be a chart of $\mathcal{M}$. Then a chart of $\mathrm{B}\mathcal{M}$ around $(\bar{x},0,0)$ is given by

$$\Theta:T\mathcal{U}_{\mathcal{M}}\oplus N\mathcal{U}_{\mathcal{M}}\to\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^{n-d}:(x,u,v)\mapsto(\theta(x),\mathrm{D}\theta(x)u,\mathrm{D}\Phi(x)v)=:(\hat{x},\hat{u},\hat{v}),$$

where $T\mathcal{U}_{\mathcal{M}}\oplus N\mathcal{U}_{\mathcal{M}}\subseteq\mathrm{B}\mathcal{M}$ denotes the Whitney sum $\{(x,u,v):x\in\mathcal{U}_{\mathcal{M}},u\in\mathrm{T}_{\mathcal{U}_{\mathcal{M}}}(x),v\in\mathrm{N}_{\mathcal{U}_{\mathcal{M}}}(x)\}$. Choose $\mathcal{U}_{\mathrm{T}\mathcal{M}}\subseteq T\mathcal{U}_{\mathcal{M}}\subseteq\mathrm{T}\mathcal{M}$ neighborhood of $(\bar{x},0)$ and $\mathcal{U}_3\subseteq\mathbb{R}^{n-d}$ sufficiently small that for all $(x,u)\in\mathcal{U}_{\mathrm{T}\mathcal{M}}$ and all $\hat{v}\in\mathcal{U}_3$, we have $x+u+v\in\mathcal{U}_\Phi$. Let also $\hat{\mathcal{U}}_{\mathrm{T}\mathcal{M}}=\{(\hat{x},\hat{u})\in\mathbb{R}^d\times\mathbb{R}^d:(x,u)\in\mathcal{U}_{\mathrm{T}\mathcal{M}}\}$. Then the chart expression $\hat{\Psi}$ of Ψ is given by

$$\hat{\Psi}:\hat{\mathcal{U}}_{\mathrm{T}\mathcal{M}}\times\mathcal{U}_3\to\mathbb{R}^{n-d}:(\hat{x},\hat{u},\hat{v})\mapsto\hat{\Psi}(\hat{x},\hat{u},\hat{v}):=\Psi(x,u,v),$$

where, in keeping with our notation, $(x,u,v)=\Theta^{-1}(\hat{x},\hat{u},\hat{v})$. Note that $\hat{\Psi}$ is of class C^{k-1} since $\mathrm{B}\mathcal{M}$ is of class C^{k-1}. Moreover, we have $\hat{\Psi}(\hat{\bar{x}},0,0)=0$ and $\mathrm{D}_3\hat{\Psi}(\hat{\bar{x}},0,0)=\mathrm{D}_3\Psi(\bar{x},0,0)\mathrm{D}\Phi(\bar{x})^{-1}=\mathrm{D}\Phi(\bar{x})\mathrm{D}\Phi(\bar{x})^{-1}$ which, being the identity, is invertible. By the implicit function theorem, shrinking further $\hat{\mathcal{U}}_{\mathrm{T}\mathcal{M}}$ around $(\hat{\bar{x}},0)$ if necessary, there exists a unique continuous function $\hat{v}:\hat{\mathcal{U}}_{\mathrm{T}\mathcal{M}}\to\mathbb{R}^{n-d}$ such that $\hat{v}(\hat{\bar{x}},0)=0$ and that, for all $(\hat{x},\hat{u})\in\hat{\mathcal{U}}_{\mathrm{T}\mathcal{M}}$, $\hat{\Psi}(\hat{x},\hat{u},\hat{v}(\hat{x},\hat{u}))=0$. The function $\hat{v}$ is of class C^{k-1} and, locally around $(\hat{\bar{x}},0,0)$, $\hat{\Psi}(\hat{x},\hat{u},\hat{v})=0$ if and only if $\hat{v}=\hat{v}(\hat{x},\hat{u})$. This means that there exists a neighborhood $\mathcal{U}_{\mathrm{T}\mathcal{M}}$ of $(\bar{x},0)$ in $\mathrm{T}\mathcal{M}$ for which there exists a unique continuous function $v:\mathcal{U}_{\mathrm{T}\mathcal{M}}\to N\mathcal{M}$ such that (i) $v(x,u)\in\mathrm{N}_{\mathcal{M}}(x)$ for all $(x,u)\in\mathcal{U}_{\mathrm{T}\mathcal{M}}$ and (ii) $v(\bar{x},0)=0$ and (iii) $\Psi(x,u,v(x,u))=0$ for all $(x,u)\in\mathcal{U}_{\mathrm{T}\mathcal{M}}$. The function v is of class C^{k-1} and, locally around $(\bar{x},0,0)$ in $\mathrm{B}\mathcal{M}$,

$$\Psi(x,u,v)=0 \iff v=v(x,u). \tag{4.3}$$

Note that the "only if" part of (4.3) is not longer guaranteed if "locally around $(\bar{x},0,0)$ in $\mathrm{B}\mathcal{M}$" is replaced by "locally around $(\bar{x},0)$ in $\mathrm{T}\mathcal{M}$". To relax the locality condition

on the v variable, one can observe that, locally around $(\bar{x}, 0)$ in $\mathrm{T}\mathcal{M}$, $v(x,u)$ is the unique smallest v that satisfies $x + u + v \in \mathcal{M}$. Indeed, otherwise the "only if" part of (4.3) would not hold locally around $(\bar{x}, 0, 0)$ in $\mathrm{B}\mathcal{M}$. Finally, $\mathrm{D}_u v(x,0) = 0$ follows from

$$\mathrm{D}_u v(x,0) = -[\mathrm{D}_v \Psi(x,0,0)]^{-1}[\mathrm{D}_u \Psi(x,0,0)],$$

since $\mathrm{D}_v \Psi(x,0,0)$ is invertible and we have $\mathrm{D}_u \Psi(x,0,0)u = \mathrm{D}\Phi(x)u = 0$ for all $u \in \mathrm{T}_{\mathcal{M}}(x)$ by definition. □

We now turn to the second technical result.

LEMMA 4.8 (Straightening up). *Let D be a retractor as defined above. Then there exists a neighborhood $\mathcal{U}_{\mathrm{T}\mathcal{M}}$ of the zero section in $\mathrm{T}\mathcal{M}$ and a unique C^{k-1} map*

$$\mathcal{A}\colon \begin{cases} \mathcal{U}_{\mathrm{T}\mathcal{M}} \oplus \mathrm{N}\mathcal{M} \longrightarrow \mathrm{T}\mathcal{M} \\ \quad (x,u,v) \quad \longmapsto (x, A(x,u)v) \end{cases}$$

where $A(x,u)$ is a linear mapping from $\mathrm{N}_{\mathcal{M}}(x)$ to $\mathrm{T}_{\mathcal{M}}(x)$ such that, for all $(x,u) \in \mathcal{U}_{\mathrm{T}\mathcal{M}}$,

$$D(x,u) = \{v + A(x,u)v,\ v \in \mathrm{N}_{\mathcal{M}}(x)\}.$$

Proof. We pick once and for all an orthonormal basis of $\mathcal{E}$, which turns $\mathcal{E}$ into $\mathbb{R}^n$ endowed with the standard inner product. Let $\bar{x} \in \mathcal{M}$, $(\mathcal{U}_{\mathcal{E}}, \phi)$ be a coordinate slice of $\mathcal{M}$ containing $\bar{x}$, $\mathcal{U}_{\mathcal{M}} = \mathcal{U}_{\mathcal{E}} \cap \mathcal{M}$, and $\mathbf{O}_n$ denote the group of all $n \times n$ orthogonal matrices. Then, from the coordinate slice, it is possible to construct a function $\mathtt{B} : \mathcal{U}_{\mathcal{M}} \to \mathbf{O}_n$ of class C^{k-1} such that, for all $x \in \mathcal{U}_{\mathcal{M}}$, the first d columns $\mathtt{B}_T(x)$ of $\mathtt{B}(x)$ are in $\mathrm{T}_{\mathcal{M}}(x)$ (and thus form an orthonormal basis of $\mathrm{T}_{\mathcal{M}}(x)$) and the other $n-d$ columns $\mathtt{B}_N(x)$ of $\mathtt{B}(x)$ are in $\mathrm{N}_{\mathcal{M}}(x)$ (and thus form an orthonormal basis of $\mathrm{N}_{\mathcal{M}}(x)$). Hence we have for all $z \in \mathbb{R}^n$ the decomposition

$$z = \mathtt{B}_T(x)\mathtt{B}_T(x)^\top z + \mathtt{B}_N(x)\mathtt{B}_N(x)^\top z.$$

Let $V_{n,n-d}$ denote the (Stiefel) manifold of all orthonormal $(n-d)$-frames in $\mathbb{R}^n$. From a smooth local section in the quotient $\mathrm{Gr}(n-d,n) = V_{n,n-d}/\mathbf{O}_{n-d}$ around $D(\bar{x},0)$, it is possible to find a neighborhood $\bar{\mathcal{U}}_{\mathrm{T}\mathcal{M}}$ of $(\bar{x},0)$ in $\mathrm{T}\mathcal{M}$ and a C^{k-1} function $\mathtt{D} : \bar{\mathcal{U}}_{\mathrm{T}\mathcal{M}} \to V_{n,n-d}$ such that, for all $(x,u) \in \bar{\mathcal{U}}_{\mathrm{T}\mathcal{M}}$, $\mathtt{D}(x,u)$ is an orthonormal basis of $D(x,u)$. Taking $\bar{\mathcal{U}}_{\mathrm{T}\mathcal{M}}$ sufficiently small, we have, for all $(x,u) \in \bar{\mathcal{U}}_{\mathrm{T}\mathcal{M}}$, that

$$D(x,u) \cap \mathrm{T}_{\mathcal{M}}(x) = \{0\}$$

and then by dimension reasons, the two subspaces (considered as linear subspaces of $\mathbb{R}^n$) are in direct sum. This implies that if $\eta \in \mathbb{R}^{n-d}$ is such that $\mathtt{B}_N(x)^\top \mathtt{D}(x,u)\eta = 0$, i.e., $\mathtt{D}(x,u)\eta \in \mathrm{T}_{\mathcal{M}}(x)$, then $\eta = 0$; this means that $\mathtt{B}_N(x)^\top \mathtt{D}(x,u)$ is invertible.

The idea is now to decompose $\mathtt{D}(x,u)$ into the normal and tangent spaces and to show that the tangent part can be expressed with the help of the normal part. We write

$$\begin{aligned} D(x,u) &= \left\{\mathtt{D}(x,u)\eta :\ \eta \in \mathbb{R}^{n-d}\right\} \\ &= \left\{\mathtt{B}_N(x)\mathtt{B}_N(x)^\top \mathtt{D}(x,u)\eta + \mathtt{B}_T(x)\mathtt{B}_T(x)^\top \mathtt{D}(x,u)\eta :\ \eta \in \mathbb{R}^{n-d}\right\} \\ &= \left\{\mathtt{B}_N(x)\tilde{\eta} + \mathtt{B}_T(x)\mathtt{B}_T(x)^\top \mathtt{D}(x,u)[\mathtt{B}_N(x)^\top \mathtt{D}(x,u)]^{-1}\tilde{\eta} :\ \tilde{\eta} \in \mathbb{R}^{n-d}\right\} \\ &= \left\{\mathtt{B}_N(x)\tilde{\eta} + \mathtt{B}_T(x)\mathtt{B}_T(x)^\top \mathtt{D}(x,u)[\mathtt{B}_N(x)^\top \mathtt{D}(x,u)]^{-1}\mathtt{B}_N(x)^\top \mathtt{B}_N(x)\tilde{\eta} :\ \tilde{\eta} \in \mathbb{R}^{n-d}\right\} \end{aligned}$$

where we use first the change of variable in $\mathbb{R}^{n-d}$ given by $\mathtt{B}_N(x)^\top \mathtt{D}(x,u)$, and second the fact that $\mathtt{B}_N(x)^\top \mathtt{B}_N(x)$ is the identity. We set

$$A(x,u) := \mathtt{B}_T(x)\mathtt{B}_T(x)^\top \mathtt{D}(x,u)[\mathtt{B}_N(x)^\top \mathtt{D}(x,u)]^{-1}\mathtt{B}_N(x)^\top$$

and we have

$$D(x,u) = \left\{\mathtt{B}_N(x)\tilde{\eta} + A(x,u)\mathtt{B}_N(x)\tilde{\eta} :\ \tilde{\eta} \in \mathbb{R}^{n-d}\right\},$$

which is the desired property. Thus $\mathcal{A}$ exists and is C^{k-1} on $\bar{\mathcal{U}}_{\mathrm{T}\mathcal{M}} \oplus \mathrm{N}\mathcal{M}$. Its uniqueness is straightforward. Moreover, this rationale holds for every $\bar{x} \in \mathcal{M}$. Hence $\mathcal{A}$ is well defined and C^{k-1} on a neighborhood of the zero section in $\mathrm{T}\mathcal{M}$. Finally, the smoothness of the function comes by construction. ∎

We are in position to prove Theorem 4.2.

Proof. (of Theorem 4.2) Let $\bar{x} \in \mathcal{M}$, let $\mathcal{U}_{\mathrm{T}\mathcal{M}}$ be a neighborhood of $(\bar{x}, 0)$ in $\mathrm{T}\mathcal{M}$ chosen sufficiently small that it fits in the two eponymous neighborhoods defined in Lemmas 4.7 and 4.8, and let v be defined as in Lemma 4.7. Let $\mathcal{U}_{\mathcal{M}} = \{x+u+v(x,u) : (x,u) \in \mathcal{U}_{\mathrm{T}\mathcal{M}}\}$, and observe that $\mathcal{U}_{\mathcal{M}}$ is a neighborhood of $\bar{x}$ in $\mathcal{M}$. For $(x,u) \in \mathcal{U}_{\mathrm{T}\mathcal{M}}$, consider

$$z \in \mathcal{U}_{\mathcal{M}} \cap (x + u + D(x,u)).$$

Then z is characterized by the following two equations. On the one hand, since $z \in (x + u + D(x,u))$, there exists a normal vector $\tilde{v}(x,u) \in \mathrm{N}_{\mathcal{M}}(x)$ (by Lemma 4.8) such that

$$z = x + u + \tilde{v}(x,u) + A(x,u)\tilde{v}(x,u). \tag{4.4}$$

On the other hand, since $z \in \mathcal{U}_{\mathcal{M}}$, there exists a tangent vector $\tilde{u}(x,u) \in \mathrm{T}_{\mathcal{M}}(x)$ (by Lemma 4.7) such that

$$z = x + \tilde{u}(x,u) + v(x, \tilde{u}(x,u)). \tag{4.5}$$

Combining (4.4) and (4.5) and decomposing on the tangent and normal spaces, we get

$$\begin{cases} \tilde{u}(x,u) = u + A(x,u)\tilde{v}(x,u) \\ v(x,\tilde{u}(x,u)) = \tilde{v}(x,u), \end{cases}$$

which yields $u = A(x,u)v(x,\tilde{u}(x,u)) + \tilde{u}(x,u)$. Introduce now the function

$$F\colon \begin{cases} \mathcal{U}_{\mathrm{T}\mathcal{M}} \oplus \mathcal{U}_{\mathrm{T}\mathcal{M}} \longrightarrow \mathrm{T}\mathcal{M} \\ \quad (x,u,\tilde{u}) \quad \longmapsto (x, A(x,u)v(x,\tilde{u}) + \tilde{u} - u), \end{cases}$$

such that

$$\tilde{u} = \tilde{u}(x,u) \iff F(x,u,\tilde{u}) \in 0_{\mathrm{T}\mathcal{M}},$$

where $0_{\mathrm{T}\mathcal{M}}$ stands for the zero section in $\mathrm{T}\mathcal{M}$. Much as in the proof of Lemma 4.7, we work in a chart and use the hat diacritic to denote coordinate expressions. Consider $\hat{F}_2\colon (\mathbb{R}^d)^3 \to \mathbb{R}^d$ that consists in the projection of $\hat{F}$ onto the last d elements (the subscript "2" indicates that the last d components are extracted), so that the condition

$F(x,u,\tilde{u}) \in 0_{\mathrm{T}\mathcal{M}}$ becomes $\hat{F}_2(\hat{x},\hat{u},\hat{\tilde{u}}) = 0$. Since $\mathrm{D}_u v(x,0) = 0$, we obtain that $\mathrm{D}_3\hat{F}_2(\hat{\bar{x}},0,0) = I$. Therefore, as in the proof of Lemma 4.7, by the implicit function theorem, $\tilde{u}$ is locally well defined as a C^{k-1} function of (x,u), and thus the intersection $R(x,u) := \mathcal{U}_{\mathcal{M}} \cap (x+u+D(x,u))$ is locally well defined as a smooth function of $(x,u) \in \mathrm{T}\mathcal{M}$. This shows in particular that $R(\bar{x},u)$ is a singleton for all u in a neighborhood of 0. An expression of R that does not involve the neighborhood $\mathcal{U}_{\mathcal{M}}$ is obtained by observing that locally around $(\bar{x},0)$, $R(x,u)$ is the element of $\mathcal{M} \cap (x+u+D(x,u))$ nearest to $x+u$. (In other words, $R(x,u)$ is the point of $\mathcal{M}$ where the correction—or restoration step—from $x+u$ along $D(x,u)$ is the smallest.) Since this rationale is valid for every $\bar{x} \in \mathcal{M}$, R is well-defined and C^{k-1} on a neighborhood of the zero section in $\mathrm{T}\mathcal{M}$.

It is straightforward that the consistency condition $R(x,0) = x$ holds. There is just left to show the first order rigidity condition $\mathrm{D}R(x,\cdot)(0) = \mathrm{id}_{\mathrm{T}_{\mathcal{M}}(x)}$. In view of (4.5) and since $\mathrm{D}_2 v(x,0) = 0$, it is sufficient to show that $\mathrm{D}_2\tilde{u}(x,0) = \mathrm{id}$. We have

$$\mathrm{D}_2\tilde{u}(x,u) = -(\mathrm{D}_3 F_2(x,u,\tilde{u}(x,u)))^{-1}\mathrm{D}_2 F_2(x,u,\tilde{u}(x,u)). \tag{4.6}$$

This yields that $\mathrm{D}_2\tilde{u}(x,0) = \mathrm{id}$. □

4.3. Retractors and second-order retractions. Recall from (2.3) that a second-order retraction is a retraction that satisfies $\frac{\mathrm{d}^2}{\mathrm{d}t^2} R(x,tu)|_{t=0} \in \mathrm{N}_{\mathcal{M}}(x)$. The next result gives an easy-to-check sufficient condition for a retraction induced by a retractor to be second order.

THEOREM 4.9 (Retractors and second-order retractions). *Let D be a retractor (Definition 4.1), and assume that, for all $\bar{x} \in \mathcal{M}$, there holds $D(\bar{x},0) = \mathrm{N}_{\mathcal{M}}(x)$. Then R defined in Theorem 4.2 is a second-order retraction on $\mathcal{M}$.*

Proof. With the notation of the proof of Theorem 4.2, we want to compute $\mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}\mathrm{D}_{22} z(x,0) = \mathrm{D}_{22}\tilde{u}(x,0)$, for $(x,u) \mapsto z$ as in (4.5). To lighten the notation, we omit to specify that the functions are evaluated at $u = 0$ and $\tilde{u} = 0$. We perform usual calculations with total derivatives:

$$\begin{aligned}
\mathrm{D}_{22}\tilde{u} &= (\mathrm{D}_3 F_2)^{-1}(\mathrm{D}_{23}F_2 + \mathrm{D}_{33}F_2\,\mathrm{id})(\mathrm{D}_3 F_2)^{-1}\mathrm{D}_2 F_2 - (\mathrm{D}_{22}F_2 + \mathrm{D}_{32}F_2\,\mathrm{id}),\\
\mathrm{D}_{23}F_2 &= \mathrm{D}_u(A\mathrm{D}_2 v + I) = \mathrm{D}_2 A\mathrm{D}_2 v + A\mathrm{D}_{22}v = 0,\\
\mathrm{D}_{33}F_2 &= \mathrm{D}_{\tilde{u}}(A\mathrm{D}_2 v + I) = A\mathrm{D}_{22}v = 0,\\
\mathrm{D}_{22}F_2 &= \mathrm{D}_{22}Av = 0,\\
\mathrm{D}_{32}F_2 &= \mathrm{D}_{\tilde{u}}(\mathrm{D}_2 Av - I) = \mathrm{D}_2 A\mathrm{D}_2 v = 0,
\end{aligned}$$

where we used (4.5), (4.6), $\mathrm{D}_2\tilde{u}(x,0) = \mathrm{id}$, $\mathrm{D}_2 v(x,0) = 0$, $A(x,0) = 0$ (since $D(\bar{x},0) = \mathrm{N}_{\mathcal{M}}(x)$), and $v(x,0) = 0$. In conclusion, $\mathrm{P}_{\mathrm{T}_{\mathcal{M}}(x)}\mathrm{D}_{22} z(x,0) = 0$, which means that R is a second-order retraction. □

EXAMPLE 4.10 (Orthographic and projective retractions). In view of Theorem 4.9, the orthographic retraction (Example 4.3) and the projective retraction (Example 4.5) are second-order retractions. □

4.4. Orthographic retraction on fixed-rank matrices. Let us come back to the example of Section 3.2, i.e., $\mathcal{R}_r = \{X \in \mathbb{R}^{n\times m} : \mathrm{rank}(X) = r\}$. Consider the singular-value decomposition of $X \in \mathcal{R}_r$,

$$X = U \begin{bmatrix} \Sigma_0 & 0 \\ 0 & 0 \end{bmatrix} V^T, \tag{4.7}$$

with $\Sigma_0 \in \mathbb{R}^{r\times r}$ the diagonal matrix of non-zero singular values. An element $Z \in \mathrm{T}_{\mathcal{R}_r}(X)$ can be decomposed as

$$Z = U \begin{bmatrix} A & C \\ B & 0 \end{bmatrix} V^\top. \tag{4.8}$$

In other words,

$$\mathrm{T}_{\mathcal{R}_r}(X) = \left\{ H \in \mathbb{R}^{n\times m} : \ u_i{}^\top H v_j = 0, \text{ for all } r < i \le n, \ r < j \le m \right\}.$$

Proposition 4.11 (orthographic retraction on the set of fixed-rank matrices). *The orthographic retraction R on the set $\mathcal{R}_r = \{X \in \mathbb{R}^{n\times m} : \mathrm{rank}(X) = r\}$ is given by*

$$\begin{aligned} R(X, Z) &= U \begin{bmatrix} \Sigma_0 + A & C \\ B & B(\Sigma_0 + A)^{-1}C \end{bmatrix} V^T \\ &= U \begin{bmatrix} \Sigma_0 + A \\ B \end{bmatrix} \begin{bmatrix} I & (\Sigma_0 + A)^{-1}C \end{bmatrix} V^T, \end{aligned} \tag{4.9}$$

where U, V, Σ_0, A, B, and C are obtained from (4.7) *and* (4.8).

Proof. The the normal space is

$$\mathrm{N}_{\mathcal{R}_r}(X) = \left\{ U \begin{bmatrix} 0 & 0 \\ 0 & D \end{bmatrix} V^\top : D \in \mathbb{R}^{(n-r)\times(m-r)} \right\}.$$

To obtain the orthographic retraction on $\mathcal{R}_r$, we want to find the smallest $Y = U \left[\begin{smallmatrix} 0 & 0 \\ 0 & D \end{smallmatrix}\right] V^\top \in \mathrm{N}_{\mathcal{R}_r}(X)$ such that

$$X + Z + Y = U \begin{bmatrix} \Sigma_0 + A & C \\ B & D \end{bmatrix} V^T \tag{4.10}$$

belongs to $\mathcal{R}_r$. There is a neighborhood of the origin in $\mathrm{T}_{\mathcal{R}_r}(X)$ such that, if Z belongs to that neighborhood, then A is small enough to make $\Sigma_0 + A$ invertible, which guarantees that (4.10) has at least rank r. It thus remains to choose D such that the matrix has rank exactly r. This is equivalent to demanding that each of the last $m-r$ columns of $\left[\begin{smallmatrix} \Sigma_0+A & C \\ B & D \end{smallmatrix}\right]$ be a linear combination of the first r columns. Equivalently, we need to solve $\left[\begin{smallmatrix} \Sigma_0+A & C \\ B & D \end{smallmatrix}\right] \left[\begin{smallmatrix} E \\ I \end{smallmatrix}\right] = \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right]$ for $E \in \mathbb{R}^{r\times(m-r)}$ and $D \in \mathbb{R}^{(n-r)\times(m-r)}$, which yields $D = B(\Sigma_0 + A)^{-1}C$. ∎

What is the computational cost of the orthographic retraction compared to the one of projective retraction discussed in Section 3.2? The projective retraction requires the computation of $\mathrm{P}_{\mathcal{R}_r}(X + Z)$, where the rank of $X + Z$ can be at most $2r$ in view of (4.7) and (4.8). The matrices involved reduce to small $2r \times 2r$ if one works in appropriate left and right bases. These bases are readily computed if X and Z are available in the form $X = NM^T$ and $Z = N\Delta_r^T + \Delta_l M^T$, where N, M, Δ_r and Δ_l are all matrices with r columns: the span of the left basis is the span of the columns of N and Δ_l, and the span of the right basis is the span of the columns of M and Δ_r. Using the same technique, the computations needed to compute the orthographic retraction also reduce to operations on matrices of size $2r \times 2r$ or less. The computational cost is thus comparable when $r \ll n, m$. Note however that the computation of the orthographic retraction $R(X, Z)$ only requires matrix multiplications and a small matrix inversion once the SVD of X is known.

4.5. Orthographic retraction on Stiefel manifolds and on orthogonal matrices. This section works out the examples of matrix manifolds introduced in Section 3.3. We show first that the orthographic retraction on the Stiefel manifold leads to a Riccati equation. This result is not new; it is mentioned in [**?**, (9.10)]. We present it in the formalism of this paper, and derive a closed-form expression of the orthographic projection for the special case of the orthogonal group.

The tangent and normal spaces to $V_{n,m}$, seen as subspaces of $\mathbb{R}^{n\times m}$, are given by

$$\begin{aligned}\mathrm{T}_{V_{n,m}}(X) &= \{Z \in \mathbb{R}^{n\times m} : X^\top Z + Z^\top X = 0\}\\ &= \{X\Omega + X_\perp K : \Omega \text{ skew-sym}, K \in \mathbb{R}^{(n-m)\times m}\}\\ \mathrm{N}_{V_{n,m}}(X) &= \{XS : S \text{ sym}\}\end{aligned}$$

where the columns of $X_\perp \in \mathbb{R}^{n\times(n-m)}$ complete those of X to make an orthogonal basis of $\mathbb{R}^m$. The orthographic retraction is thus given by

$$R(X, X\Omega + X_\perp K) = X + X\Omega + X_\perp K + XS \in V_{n,m}, \tag{4.11}$$

where S is the smallest possible symmetric matrix. Theorem 4.2 guarantees that this smallest S exists and is unique for all $X\Omega + X_\perp K$ sufficiently small. Condition $R(X, X\Omega + X_\perp K) \in V_{n,m}$ reads

$$(X + X\Omega + X_\perp K + XS)^T(X + X\Omega + X_\perp K + XS) = I \tag{4.12}$$

which, taking all the assumptions into account, is equivalent to

$$S^2 + (I + \Omega^T)S + S(I + \Omega) + \Omega^T\Omega + K^TK = 0, \tag{4.13}$$

where the unknown S is symmetric and Ω is skew-symmetric. This is a continuous-time algebraic Riccati equation, which can be solved by various means; see, e.g., [**?**]. When $n \gg m$, the dominant cost is in forming K^TK and XS, which is $O(nm^2)$, and thus comparable with the cost of the projective retraction discussed in Section 3.3.

The case when $m = n$ (and thus $K = 0$) leads to a closed-form solution as established in Proposition 4.12 below. This closed-form uses the square root of a symmetric positive-semidefinite matrix A, defined for an eigenvalue decomposition $A = U\,\mathrm{Diag}(\lambda_1, \ldots, \lambda_n)U^\top$ by

$$\sqrt{A} = U\,\mathrm{Diag}\big(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}\big)U^\top.$$

So again, computing the retraction essentially amounts to computing the eigenvalues of a matrix.

PROPOSITION 4.12 (orthographic retraction on $\mathbf{O}_n$). *The orthographic retraction on the orthogonal group $\mathbf{O}_n$ is given for $X \in \mathbf{O}_n$ by*

$$R(X, X\Omega) = X\big(\Omega + \sqrt{I - \Omega^T\Omega}\big).$$

Proof. For $n = m$, (4.13) gives

$$S^2 + 2S + \Omega^T\Omega = 0,$$

that we reformulate as

$$(S + I)^2 = I - \Omega^T\Omega. \tag{4.14}$$

Note that this yields that $I - \Omega^T\Omega$ is positive semidefinite. The solutions of (4.14) are then

$$S_+ = -I + \sqrt{I - \Omega^\top\Omega} \qquad \text{and} \qquad S_- = -I - \sqrt{I - \Omega^\top\Omega},$$

given a squared root of $I-\Omega^\top\Omega$. Given the eigendecomposition $\Omega^\top\Omega = U\,\mathrm{Diag}(\lambda_1, \dots, \lambda_n)U^\top$, we have

$$S_\pm = U\ \mathrm{Diag}\big(-1 \pm \sqrt{1-\lambda_1}, \dots, 1 \pm \sqrt{1-\lambda_n}\big)U^\top,$$

to get the associated norms

$$\|S_\pm\|^2 = \sum_{i=1}^{n} (-1 \pm \sqrt{1-\lambda_i})^2.$$

Between both, the one with smallest norm is S_+, and then (4.11) gives the result. □

Acknowledgements. The authors would like to thank Paul Van Dooren for several useful discussions.