

Chapter Four

Line-Search Algorithms on Manifolds

Line-search methods in \mathbb{R}^n are based on the update formula

$$x_{k+1} = x_k + t_k \eta_k, \quad (4.1)$$

where $\eta_k \in \mathbb{R}^n$ is the *search direction* and $t_k \in \mathbb{R}$ is the *step size*. The goal of this chapter is to develop an analogous theory for optimization problems posed on nonlinear manifolds.

The proposed generalization of (4.1) to a manifold \mathcal{M} consists of selecting η_k as a tangent vector to \mathcal{M} at x_k and performing a search along a curve in \mathcal{M} whose tangent vector at $t = 0$ is η_k . The selection of the curve relies on the concept of retraction, introduced in Section 4.1. The choice of a computationally efficient retraction is an important decision in the design of high-performance numerical algorithms on nonlinear manifolds. Several practical examples are given for the matrix manifolds associated with the main examples of interest considered in this book.

This chapter also provides the convergence theory of line-search algorithms defined on Riemannian manifolds. Several example applications related to the eigenvalue problem are presented.

4.1 RETRACTIONS

Conceptually, the simplest approach to optimizing a differentiable function is to continuously translate a test point $x(t)$ in the direction of steepest descent, $-\text{grad } f(x)$, on the constraint set until one reaches a point where the gradient vanishes. Points x where $\text{grad } f(x) = 0$ are called *stationary points* or *critical points* of f . A numerical implementation of the continuous gradient descent approach requires the construction of a curve γ such that $\dot{\gamma}(t) = -\text{grad } f(\gamma(t))$ for all t . Except in very special circumstances, the construction of such a curve using numerical methods is impractical. The closest numerical analogy is the class of optimization methods that use *line-search* procedures, namely, iterative algorithms that, given a point x , compute a descent direction $\eta := -\text{grad } f(x)$ (or some approximation of the gradient) and move in the direction of η until a “reasonable” decrease in f is found. In \mathbb{R}^n , the concept of moving in the direction of a vector is straightforward. On a manifold, the notion of moving in the direction of a tangent vector, while staying on the manifold, is generalized by the notion of a retraction mapping.

1. moving along ξ to get the point $x + \xi$ in the linear embedding space;
2. “projecting” the point $x + \xi$ back to the manifold \mathcal{M} .

The issue is to define a projection that (i) turns the procedure into a well-defined retraction and (ii) is computationally efficient. In the embedded submanifolds of interest in this book, as well as in several other cases, the second step can be based on matrix decompositions. Examples of such decompositions include QR factorization and polar decomposition. The purpose of the present section is to develop a general theory of decomposition-based retractions. With this theory at hand, it will be straightforward to show that several mappings constructed along the above lines are well-defined retractions.

Let \mathcal{M} be an embedded manifold of a vector space \mathcal{E} and let \mathcal{N} be an abstract manifold such that $\dim(\mathcal{M}) + \dim(\mathcal{N}) = \dim(\mathcal{E})$. Assume that there is a diffeomorphism

$$\phi : \mathcal{M} \times \mathcal{N} \rightarrow \mathcal{E}_* : (F, G) \mapsto \phi(F, G),$$

where \mathcal{E}_* is an open subset of \mathcal{E} (thus \mathcal{E}_* is an open submanifold of \mathcal{E}), with a neutral element $I \in \mathcal{N}$ satisfying

$$\phi(F, I) = F, \quad \forall F \in \mathcal{M}.$$

(The letter I is chosen in anticipation that the neutral element will be the identity matrix of a matrix manifold \mathcal{N} in cases of interest.)

Proposition 4.1.2 *Under the above assumptions on ϕ , the mapping*

$$R_X(\xi) := \pi_1(\phi^{-1}(X + \xi)),$$

where $\pi_1 : \mathcal{M} \times \mathcal{N} \rightarrow \mathcal{M} : (F, G) \mapsto F$ is the projection onto the first component, defines a retraction on \mathcal{M} .

Proof. Since \mathcal{E}_* is open, it follows that $X + \xi$ belongs to \mathcal{E}_* for all ξ in some neighborhood of 0_X . Since ϕ^{-1} is defined on the whole \mathcal{E}_* , it follows that $R_X(\xi)$ is defined for all ξ in a neighborhood of the origin of $T_X\mathcal{M}$. Smoothness of R and the property $R_X(0_X) = X$ are direct. For the local rigidity property, first note that for all $\xi \in T_X\mathcal{M}$, we have

$$D_1\phi(X, I)[\xi] = D\phi(X, I)[(\xi, 0)] = \xi.$$

Since $\pi_1 \circ \phi^{-1}(\phi(F, I)) = F$, it follows that, for all $\xi \in T_X\mathcal{M}$,

$$\xi = D(\pi_1 \circ \phi^{-1})(\phi(X, I)) [D_1\phi(X, I)[\xi]] = D(\pi_1 \circ \phi^{-1})(X)[\xi] = DR_X(0_X)[\xi],$$

which proves the claim that R_X is a retraction. \square

Example 4.1.1 *Retraction on the sphere S^{n-1}*

Let $\mathcal{M} = S^{n-1}$, let $\mathcal{N} = \{x \in \mathbb{R} : x > 0\}$, and consider the mapping

$$\phi : \mathcal{M} \times \mathcal{N} \rightarrow \mathbb{R}_*^n : (x, r) \mapsto xr.$$

It is straightforward to verify that ϕ is a diffeomorphism. Proposition 4.1.2 yields the retraction

$$R_x(\xi) = \frac{x + \xi}{\|x + \xi\|},$$

for all x in some neighborhood of x_* .)

We now give a stability result.

Theorem 4.4.2 (capture theorem) *Let $F : \mathcal{M} \rightarrow \mathcal{M}$ be a descent mapping for a smooth cost function f and assume that, for every $x \in \mathcal{M}$, all the accumulation points of $\{F^{(k)}(x)\}_{k=1,2,\dots}$ are critical points of f . Let x_* be a local minimizer and an isolated critical point of f . Assume further that $\text{dist}(F(x), x)$ goes to zero as x goes to x_* . Then x_* is an asymptotically stable point of F .*

Proof. Let \mathcal{U} be a neighborhood of x_* . Since x_* is an isolated local minimizer of f , it follows that there exists a closed ball

$$\overline{B}_\epsilon(x_*) := \{x \in \mathcal{M} : \text{dist}(x, x_*) \leq \epsilon\}$$

such that $\overline{B}_\epsilon(x_*) \subset \mathcal{U}$ and $f(x) > f(x_*)$ for all $x \in \overline{B}_\epsilon(x_*) - \{x_*\}$. In view of the condition on $\text{dist}(F(x), x)$, there exists $\delta > 0$ such that, for all $x \in B_\delta(x_*)$, $F(x) \in \overline{B}_\epsilon(x_*)$. Let α be the minimum of f on the compact set $\overline{B}_\epsilon(x_*) - B_\delta(x_*)$. Let

$$\mathcal{V} = \{x \in \overline{B}_\epsilon(x_*) : f(x) < \alpha\}.$$

This set is included in $B_\delta(x_*)$. Hence, for every x in \mathcal{V} , it holds that $F(x) \in \overline{B}_\epsilon(x_*)$, and it also holds that $f(F(x)) \leq f(x) < \alpha$ since F is a descent mapping. It follows that $F(x) \in \mathcal{V}$ for all $x \in \mathcal{V}$, hence $F^{(n)}(x) \in \mathcal{V} \subset \mathcal{U}$ for all $x \in \mathcal{V}$ and all n . This is stability. Moreover, since by assumption x_* is the only critical point of f in \mathcal{V} , it follows that $\lim_{n \rightarrow \infty} F^{(n)}(x) = x_*$ for all $x \in \mathcal{V}$, which shows asymptotic stability. \square

The additional condition on $\text{dist}(F(x), x)$ in Theorem 4.4.2 is not satisfied by every instance of Algorithm 1 because our accelerated line-search framework does not put any restriction on the step length. The distance condition is satisfied, for example, when η_k is selected such that $\|\eta_k\| \leq c \|\text{grad } f(x_k)\|$ for some constant c and x_{k+1} is selected as the Armijo point.

In this section, we have assumed for simplicity that the next iterate depends only on the current iterate: $x_{k+1} = F(x_k)$. It is possible to generalize the above result to the case where x_{k+1} depends on x_k and on some “memory variables”: $(x_{k+1}, y_{k+1}) = F(x_k, y_k)$.

4.5 SPEED OF CONVERGENCE

We have seen that, under reasonable assumptions, if the first iterate of Algorithm 1 is sufficiently close to an isolated local minimizer x_* of f , then the generated sequence $\{x_k\}$ converges to x_* . In this section, we address the issue of how fast the sequence converges to x_* .

4.5.1 Order of convergence

A sequence $\{x_k\}_{k=0,1,\dots}$ of points of \mathbb{R}^n is said to converge linearly to a point x_* if there exists a constant $c \in (0, 1)$ and an integer $K \geq 0$ such that, for

