Introduction to the Theory of Optimisation on Manifolds

Jonathan H. Manton

ARC Special Research Center for

Ultra-Broadband Information Networks

THE UNIVERSITY OF MELBOURNE

Part I: Preliminaries

- What is a manifold?
- Why would we want to minimise a cost function on a manifold?
- Constrained Optimisation
- Definition of Concrete Manifold
- Signal Processing Example — Stiefel Manifold
- Abstract Manifolds
- Signal Processing Example — Grassmann Manifold

Constrained Optimisation

$$\min_{x \in M} f(x), \qquad M = \{x \in \mathcal{R}^m | h(x) = 0\}$$

- Depending on M (and perhaps on f too), we are led to different optimisation problems: linear, convex, ...
- If M is “smooth” (no self-intersections, no cusps, uniform dimension etc), then it is a manifold.
- This is true if the Jacobian of h at all points on M has full row rank.
  - $\{(x, y) \in \mathcal{R}^2 | x^2 + y^2 = 1\}$ is a manifold. ($J = [2x\ 2y]$ doesn’t vanish.)
  - $\{(x, y) \in \mathcal{R}^2 | x^2 + y^3 = 0\}$ is not a manifold because there is a singularity at the origin. ($J = [2x\ 3y^2]$ vanishes at origin.)
- The constrained problem in $\mathcal{R}^m$ is equivalent to an unconstrained problem on M. We have lowered the dimension of the problem in general.

Manifolds as Parametrisations

Besides $\{x|h(x) = 0\}$, what other sets are manifolds?

- Intuitively, open sets in $\mathcal{R}^n$ are nice; they have dimension n. Thus, we would want any such set to be a manifold.
- Homeomorphic images of open sets in $\mathcal{R}^n$ should also have dimension n; we would also want any such set to be a manifold.
- Finally, just as we will see the circle consists of 1-D pieces patched together nicely, we choose to define a manifold as a collection of homeomorphic images of open sets patched together in a compatible way. (This patching allows us to define which functions on M are smooth, what their directional derivatives are etc. For our purposes, we get nice parametrisations which enable us to do optimisation.)
- This patching is necessary since global parametrisations do not exist in general, as now seen.

The Circle

- Because $M = \{(x, y) \in \mathcal{R}^2 | x^2 + y^2 = 1\}$ is compact but any open subset of $\mathcal{R}$ is not, there cannot exist a homeomorphism from any open subset of $\mathcal{R}$ to M; there is no global parametrisation. (Sending θ to $(\cos\theta, \sin\theta)$ doesn't work because it does not cover M if $\theta \in (-\pi, \pi)$ yet it is not injective on any larger open set.)

- However, there are local parametrisations, e.g.

$$\mu : (-\pi, \pi) \to M, \qquad \mu(\theta) = (\cos\theta, \sin\theta),$$
$$\nu : (\pi/2, 3\pi/2) \to M, \quad \nu(\theta) = (\cos\theta, \sin\theta).$$

- Moreover, μ and ν are "compatible" in that

$$\mu^{-1} \circ \nu(\theta) = \begin{cases} \theta, & \theta \in (\pi/2, \pi), \\ \theta - 2\pi, & \theta \in (\pi, 3\pi/2) \end{cases}$$

  is smooth on the domain of definition of $\mu^{-1} \circ \nu$, namely $(\pi/2, \pi) \cup (\pi, 3\pi/2)$.

Manifolds and Concrete Manifolds

- A manifold M is
  - a nice (Hausdorff and paracompact) topological space;
  - about each point $p \in M$, there exists a neighbourhood U of p and a coordinate chart $\phi : U \to \mathcal{R}^n$ which is a homeomorphism onto an open subset of $\mathcal{R}^n$.
  - In fact, there are typically many such neighbourhoods and coordinate charts about each point. These charts must be related in a smooth fashion.
  - Specifically, if ϕ, ψ are two charts, then $\phi \circ \psi^{-1}$ must be smooth when defined.

- The key property is that M looks locally like $\mathcal{R}^n$ where n is the dimension of the manifold. i.e. Locally, we can parametrise M.

- A concrete manifold is one which sits (embeds) naturally in $\mathcal{R}^m$, such as when M is defined by $M = \{x \in \mathcal{R}^m | h(x) = 0\}$.

- If $M = \{x \in \mathcal{R}^m | h(x) = 0\}$ then the topology is that induced from $\mathcal{R}^m$ and the coordinate charts are constructed using the implicit function theorem.

Stiefel Manifold

[6]

$$\mathrm{St}(p,n) = \{X \in \mathcal{R}^{n\times p} | X^T X = I\}.$$

- Define $h(X)$ to be the upper triangular part of $X^T X - I$; then $\mathrm{St}(p,n) = \{X | h(X) = 0\}$.
- There are $p+(p-1)+\cdots+1 = p(p+1)/2$ (algebraically independent) constraints, hence $\dim \mathrm{St}(p,n) = np - p(p+1)/2$.
- The topology of $\mathrm{St}(p,n)$ is that induced by the ambient space $\mathcal{R}^{n\times p}$.
- In particular, $\mathrm{St}(p,n)$ is compact.
- The differentiable structure (coordinate charts) come from the implicit function theorem. This structure is equivalent to defining what smooth functions are.
- A function $\tilde{f} : \mathcal{R}^{n\times p} \to \mathcal{R}$ restricts to a function $f : M \to \mathcal{R}$. We say f is smooth if $\tilde{f}$ is smooth at each point of M.

Stiefel Manifold — Example

- Assume we want to compute the eigenvectors associated with the p smallest eigenvalues of the symmetric matrix A.
- This is achieved by minimising $f(X) = \mathrm{tr}(X^T A X N)$ subject to the constraint $X^T X = I$. Here, N is a diagonal matrix with distinct eigenvalues.
- Equivalently, we want to minimise f as a function on $\mathrm{St}(p, n)$.
- Note that if $p = 1$ then $f(X)$ is the Rayleigh quotient.
- Can think of $\mathrm{St}(p, n)$ as the set of all ordered basis vectors.

Abstract Manifolds

- In the definition of a manifold (nice topological space with compatible coordinate charts), no mention of the ambient space $\mathcal{R}^m$ was made.
- Hence, any nice topological space M can be made into a manifold by specifying a collection of compatible coordinate charts. (Equivalently, we are told how to parametrise M locally about each point.)
- $f : M \to \mathcal{R}$ is smooth if all $f \circ \phi^{-1}$ are smooth.
- Whitney's embedding theorem says that any manifold can be embedded in $\mathcal{R}^m$ provided m is sufficiently large.
  - We can still think of an abstract manifold as a smooth surface in $\mathcal{R}^m$.
  - However, this embedding is not necessarily natural (and definitely not unique), so we usually prefer to think in terms of the coordinate charts.

Grassmann Manifold — Motivational Example [9]

- If $f(X) = \mathrm{tr}(X^T A X)$ then its minimum, subject to $X^T X = I$, is not unique. Any matrix X whose columns span the minor subspace of A minimises $f(X)$.
- Note $f(XQ) = f(X)$ for any orthogonal matrix Q.
- To compute the minor subspace of A, we want to minimise $f(X)$ subject to $X^T X = I$, but we also want to "quotient out" the ambiguity $f(XQ) = f(X)$.
- Define $\mathrm{Gr}(p, n)$ to be the quotient space of $\mathrm{St}(p, n)$ obtained by saying $X, Y \in \mathrm{St}(p, n)$ are equivalent if there exists an orthogonal Q such that $X = YQ$. (Note for example, a rectangle with its two long sides identified is a cylinder.)
- Technically, f induces a cost function on $\mathrm{Gr}(p, n)$, and the minimum of this function is unique and is the minor subspace of A (assuming A has a unique minor subspace).

Grassmann Manifold

- As a set of points, the Grassmann manifold $\mathrm{Gr}(p, n)$ is the set of all p-dimensional subspaces of $\mathcal{R}^n$.
- Since $f(X) = \mathrm{tr}(X^T A X)$ is really a function of the space spanned by the columns of X, there are interesting functions defined on $\mathrm{Gr}(p, n)$.
- What does it mean for a function on $\mathrm{Gr}(p, n)$ to be smooth?
- First give $\mathrm{Gr}(p, n)$ a topology and then a differentiable structure (i.e. coordinate charts). Recall that then $f : \mathrm{Gr}(p, n) \rightarrow \mathcal{R}$ is smooth if $f \circ \phi^{-1}$ is smooth for all ϕ.
- Without going into details, because $\mathrm{Gr}(p, n)$ is a quotient space of $\mathrm{St}(p, n)$, it inherits a topology and differentiable structure in a natural way. Smooth functions on $\mathcal{R}^{n \times p}$ which depend only on the column space induce smooth functions on $\mathrm{Gr}(p, n)$.

Summary of Part I

What is a manifold?

- A topological space which locally looks like $\mathcal{R}^n$ where n is the dimension of the manifold. (Other, more technical, conditions are required too.)
- Visually, a concrete manifold is a “smooth” surface or subset of $\mathcal{R}^m$, such as a sphere. Can think of all manifolds in this way, but better to think in terms of coordinate charts or parametrisations.

Why would we want to minimise a cost function on a manifold?

- The problem arises naturally in signal processing, such as when subspaces are involved; Grassmann and Stiefel manifolds.
- Applications include precoder design, weighted low rank approximations, source separation, joint diagonalisation and so forth.

Part II: The Optimisation Problem

- Is a particular optimisation algorithm good or bad?
- Is it better to solve constrained or unconstrained problems?
- Measures of Performance of Optimisation Algorithms
- Common Mistake #1
- Implications
- Constrained or Unconstrained?

Measures of Performance

An iterative algorithm (e.g. Newton's method) for minimising a function generates a sequence $x_0, x_1, \cdots$ which hopefully converges to a local minimum.

Its performance can be based on:

- how many iterations are required to converge to within ϵ of a local minimum;
- how computationally expensive each iteration is;
- how close to the minimum we need to initialise the algorithm for it to converge.

Common Mistake #1

- One usually thinks optimisation on manifolds is concerned with evaluating $\arg\min_{p \in M} f(p)$ where M is a manifold and $f : M \to \mathcal{R}$ a cost function.

- However, it is essential to realise the problem is not to compute $\arg\min f$ for fixed f, but rather:
  - There is a predefined set Ω of functions $f : M \to \mathcal{R}$.
  - The objective is to design an algorithm which
    - takes $f \in \Omega$ as input, and
    - returns $\arg\min_{p \in M} f(p)$.
  - The algorithm is to "perform well" for all (or most) $f \in \Omega$.

- Otherwise, if Ω consists of a single function, the best algorithm simply returns the pre-computed minimum of that single function.

- Similarly, if Ω is too large, then it can be shown that no algorithm works well.

Implications

- The properties of M alone do not suffice to determine a good algorithm.
  - If M is convex, then need the $f \in \Omega$ to be convex too for this to be useful.
  - If M has extra structure (Riemannian metric), unless the $f \in \Omega$ are somehow related to this extra structure, we may as well ignore it.

- A focus of research should be on

  1. finding practically useful but sufficiently small sets of functions Ω;
  2. coming up with novel algorithms for optimising these functions.

- Note: The eigenvalue problem is that of optimising the class of functions $f(X) = \mathrm{tr}(X^T A X N)$, indexed by A, on $\mathrm{St}(p, n)$.

Constrained or Unconstrained

- The constrained problem $\min_{(x,y)\in M} f(x,y)$, where $M = \{(x,y) \in \mathcal{R}^2 | x+y = 0\}$, can be written as the unconstrained problem $\min_{t\in\mathcal{R}} f(t,-t)$.
- We can argue that the unconstrained problem has lower dimension and so should be easier to solve. (Maybe not true in general.)
- What we have done is introduce a global parametrisation.
- We will see that optimisation on (concrete) manifolds is essentially this idea, but because global parametrisations don't exist on manifolds in general, we use local parametrisations instead.
- We thus argue that optimisation on manifolds is attractive because the dimension is usually reduced significantly.
- In general though, it really depends on the class of functions Ω.

Summary of Part II

Is a particular optimisation algorithm good or bad?

- Cannot comment until we know the set of functions Ω for which the algorithm will be used for.
- If Ω is too large, no algorithm is “good”.
- If Ω is small, some algorithms will consistently require less flops to converge to within ϵ of a local minimum (roughly speaking).

Is it better to solve constrained or unconstrained problems?

- Must be considered on a case by case basis.
- On abstract manifolds, it is not natural to re-pose it as an unconstrained problem.
- When the manifold has extra structure (Lie group, homogeneous space, symmetric space) and the elements of Ω are somehow compatible with this structure, the manifold approach is likely to be preferable.

Part III: Traditional Approach

- What is the traditional approach to optimisation on manifolds?
- Is it good or bad?
- The Riemannian Approach
- Its Advantages
- Its Disadvantages
- Common Mistake #2

The Riemannian Approach

- The Riemannian approach has been around since at least 1982 (Gabay).
- It seeks to generalise the formula $x_{k+1} = x_k - [f''(x_k)]^{-1} f'(x_k)$ so it is valid for $f : M \to \mathcal{R}$. (Iterates other than the Newton iterate can be used instead.)
- Since $f'(p)$ and $f''(p)$ are not defined unless M has a metric structure, the manifold M is first made into a Riemannian manifold. (Note that the gradient of a function on $\mathcal{R}^n$ is not defined until an inner product is given.)
- $f'(p)$ is replaced by the gradient of f at p, $f''(p)$ replaced by the Hessian, and the increment $x_{k+1} = x_k - \delta x$ replaced by a unit step along the geodesic in the direction $-\delta x$.

Its Advantages

- Intuitively, it is “well-conditioned” in that it can be expected to converge within a reasonable number of iterations for “well-behaved” cost functions.
- Guaranteed convergence if f is convex with respect to the Riemannian geometry.
- If f is related to the Riemannian geometry (e.g. f is convex, or f depends on the induced distance function such as in centre of mass type problems) then intuitively a Riemannian approach is worth considering.

Its Disadvantages

- Unless the cost function $f : M \rightarrow \mathcal{R}$ is related to the Riemannian geometry, there is no reason to introduce a Riemannian structure in the first place, and no reason why the performance should be particularly good.
- Often, computationally intensive to compute updates along geodesics.

Common Mistake #2

“We should make use of the differentiable structure!”
“A non-Riemannian approach is only an approximation of a Riemannian approach!”

- There is more than one way of generalising a Newton method to a manifold!
- Often, the only differentiable structure is the coordinate charts. We are not a priori given a Riemannian metric, even if one is naturally associated with a manifold M.
- Unless f is intimately related to a particular Riemannian geometry of M, no reason why the Riemannian approach should be best.
- Examples exist showing that using a Riemannian structure different from the “natural” Riemannian structure leads to better algorithms, and we suspect there are other practical cases when not using any Riemannian structure at all gives better performance.

Summary of Part III

What is the traditional approach to optimisation on manifolds?

- First add (usually artificial!) extra structure, namely a Riemannian metric.
- Use a generalised version of, say, Newton's iterate, obtained by replacing straight lines by geodesics and derivatives by their Riemannian versions.

Is it good or bad?

- It is sensible.
- It works well if f is convex.
- Unless the set Ω of functions is somehow related to the Riemannian geometry, can expect there to exist better methods.

Part IV: Our Approach

- What can we fiddle with to get better performance?
- Does this solve the problem?
- The Varying Parametrisation Approach
- Its Relation to the Riemannian Approach
- Choice of Parametrisations
- Current Work

The Varying Parametrisation Approach

- Beforehand, associate with every point $p \in M$ a parametrisation $\mu_p : \mathcal{R}^n \to M$ centred at p ($\mu_p(0) = p$).
- Therefore, when at point p_k, can form local cost function $f \circ \mu_{p_k}$.
- Thus, it is proposed to apply the ordinary Newton iterate to $f \circ \mu_{p_k}$ at the origin and map the result back to M using $\mu_{p_k}^{-1}$. (Can use iterates other than the Newton iterate.)
- The real trick is to choose the μ_p. Different people have proposed different choices.

Its Relation to the Riemannian Approach

- It includes the Riemannian approach as a special case. (Choose μ_p to be the inverse of the Riemannian exponential map.)
- It is a strict generalisation; some choices of μ_p lead to algorithms which cannot be written in terms of a Riemannian metric.
- Empirical evidence shows choices of μ_p other than the exponential map can give better performance in practice.
- We have a nicer (but unpublished) interpretation which makes it clear that this approach is not an approximation of a Newton method. Rather, it corresponds to a Newton iterate applied in different coordinate systems; see later slide.

Choice of Parametrisations

- The challenge is to come up with excellent choices of μ_p for specific problems (i.e. pairs (M, Ω)).
- Intuitively, for the Newton iterate, want μ_p so that the local cost functions $f \circ \mu_p$ are approximately quadratic yet the μ_p^{-1} are computationally straightforward to evaluate.
- One generic choice, for concrete manifolds, is to use the parametrisation formed by projection from the tangent plane. (e.g. Use this if Ω is very large.)

Current Work

- We have generalised the varying parametrisation approach further.
- We have a nicer interpretation of it; essentially, the problem can be mapped to one in Euclidean space.
- In Euclidean space, we can perform a Newton iterate in polar coordinates or cartesian coordinates and so forth.
- The idea is to change the coordinate system based on the point we are currently at.
- If these changes are reasonable, quadratic convergence is guaranteed.
- We also show what approximations can be made without affecting quadratic convergence.
- We conjecture all quadratically convergent algorithms for a sufficiently rich class Ω of functions are of this form.

Summary of Part IV

What can we fiddle with to get better performance?

- We can choose the coordinate system in which to perform the current iterate (e.g. Newton step).
- We can make various approximations to reduce the computational complexity.

Does this solve the problem?

- It is a framework only.
- The challenge is to take this framework and design good algorithms for pairs M, Ω.
- The framework provides intuition as well as guarantees local convergence.
- Global convergence is a much harder problem but the same ideas apply.

Conclusion

- A fundamental tool in signal processing is linear algebra.
- If we move beyond linear algebra, we enter the world of algebraic geometry (polynomial equations) or differential geometry (manifolds) and so forth.
- Both algebraic and differential geometry have been applied to signal processing problems in the past.
- It is expected the number of applications will grow.
- Finding nice optimisation on manifold algorithms for particular problems (i.e. for pairs M, Ω) is a fruitful research direction.
- We now have a framework on which to build for tackling this problem (not yet published though).

References

1. For an overview, see the paper “On the Various Generalisations of Optimisation Algorithms to Manifolds” presented at MTNS 2004 in Belgium. (Available online from the conference website.)

2. For the varying coordinate approach, see “Optimization Algorithms Exploiting Unitary Constraints”, TR-SP 2002.

3. For an application, see “The Geometry of Weighted Low-Rank Approximations”, TR-SP 2003.