ON THE ROLE OF DIFFERENTIAL GEOMETRY IN SIGNAL PROCESSING

Jonathan H. Manton

Department of Electrical and Electronic Engineering
The University of Melbourne, Victoria 3010, Australia
j.manton@ee.mu.oz.au

ABSTRACT

Traditionally, the majority of non-linear signal processing problems were tackled either by appropriate linearisations or by some ad hoc technique. By contrast, linear signal processing problems are routinely solved systematically by astute application of results from linear algebra. This paper shows by example how differential geometry provides the necessary tools and mindset for systematically solving certain non-linear problems commonly encountered in signal processing.

1. INTRODUCTION

Linear functions on vector spaces (the most notable vector space being $\mathbb{R}^n$) commonly arise in signal processing and are invariably dealt with using the machinery of linear algebra. Nonlinear differentiable functions on $\mathbb{R}^n$ are amenable to differential calculus and real analysis. (Sometimes though, there are more suitable tools such as algebraic geometry.) A third type of function which can arise in signal processing is a function defined on a more general space than $\mathbb{R}^n$, such as on a sphere or a torus (i.e., doughnut) in the simplest of cases. Differential geometry is the study of such functions.

Indeed, one way to motivate the development of differential geometry is to ask for what spaces other than $\mathbb{R}^n$ can the machinery of differential calculus be extended to. Mathematicians concluded the answer is manifolds. Roughly speaking, a manifold consists of open sets of $\mathbb{R}^n$ glued together. (For a formal definition, see [4].) To visualise this definition, take a circle. The top hemisphere consisting of all points with an angle strictly between 0 and π can be thought of as an open interval in $\mathbb{R}$. The same goes for the left, right and bottom hemispheres. The union of these four "open sets of $\mathbb{R}$" form a circle. Note that by necessity, some of these open sets overlap; the left and top hemispheres overlap for instance. Although this may sound meaningless on a first reading, and is indeed totally inadequate as a formal definition of a manifold, the key point to remember is that *locally* a manifold looks like $\mathbb{R}^n$ whereas *globally* it doesn't. For example, although in a neighbourhood of a point on a circle, the circle looks like a segment of $\mathbb{R}$, if you travel in a "straight line" (a geodesic) on a circle then you will soon end up where you started, thus concluding that globally the circle does not look like $\mathbb{R}$ since if you travel in a straight line on $\mathbb{R}$ you never come back to where you started. (Mathematically, the circle cannot look like an open subset of $\mathbb{R}$ because the former is compact but the latter is not; in particular, there does not exist

a homeomorphism mapping an open subset of $\mathbb{R}$ onto the circle. This is why hemispheres were glued together to form the circle.)

A more pragmatic motivation for the introduction of manifolds and differential geometry is afforded by the following three problems arising in signal processing.

1. The parameter $\theta \in \mathbb{R}^n$ is to be estimated from observed data, and it is known that θ satisfies the constraint $f(\theta) = 0$.

2. The complex channel vector $h \in \mathbb{C}^n$ can at best be estimated up to an unknown scaling factor. How then should the distance between the true channel h and an estimate $\widehat{h}$ be measured?

3. Let $A(t) \in \mathbb{R}^{5 \times 5}$ denote a time-varying matrix. It is required to track the subspace spanned by the eigenvectors associated with the two largest eigenvalues of $A(t)$.

Given a differentiable function $f : \mathbb{R}^n \to \mathbb{R}^m$, it is a standard result that the set $\{x \in \mathbb{R}^n | f(x) = 0\}$ is a manifold[1] provided the Jacobian matrix $f'(x)$ has full row rank for every x satisfying $f(x) = 0$. To understand this definition, first consider $f(x, y) = x^2 + y^2 - 1$. Its Jacobian matrix is $[2x \; 2y]$ and has full row rank at all points on the circle $x^2 + y^2 = 1$, hence the circle is a manifold. In contrast, the function $f(x, y) = x^2 + y^3$ fails the requirement since the Jacobian matrix $[2x \; 3y^2]$ does not have full row rank at the origin $(0, 0)$. Indeed, plotting this function shows that the origin is a cusp, hence $x^2 + y^3 = 0$ is not a "smooth surface" and should not be thought of as a (smooth) manifold.

Thus, differential geometry is directly relevant to Problem 1 if the set $\{\theta | f(\theta) = 0\}$ is a manifold [28]. This is discussed further in Section 4.

Focusing now on Problem 2, although the signal processing community has introduced several ad hoc ways of defining the estimation error in the presence of ambiguity, it wasn't until recently [15] that it was pointed out that the natural way of thinking of this problem is to work not in $\mathbb{C}^n$ but in complex projective space. First though, it is noted that differential geometry readily extends to complex manifolds, either by treating $\mathbb{C}^n$ as equivalent to $\mathbb{R}^{2n}$ or by defining manifolds as open subsets of $\mathbb{C}^n$ glued together. (See, for example, [16].)

Complex projective space is an example of how a new space is obtained by starting with an existing space, $\mathbb{C}^n$ minus the origin in this case, and then *quotienting out* the ambiguity, in this case the equivalence relation stating that $x, y \in \mathbb{C}^n - \{0\}$ are equivalent if there exists a scalar $\alpha \in \mathbb{C}$ such that $x = \alpha y$. In other words, $\mathbb{C}^n - \{0\}$ is partitioned so that x and y are in the same partition if

---

[1]Strictly speaking, it can be made into a smooth manifold by giving it a topology and a differentiable structure in a natural way.

and only if they differ by a scalar multiple, and complex projective space is the space formed by treating each partition as merely a point. (See [5] for a detailed introduction to projective space.)

Treating the true channel h and its estimate $\widehat{h}$ as points in complex projective space, it becomes obvious that what Problem 2 calls for is a metric $d(h, \widehat{h})$ to be defined on complex projective space.

Problem 1 exemplifies how a manifold can arise from a constraint while Problem 2 shows how a manifold can arise by quotienting out an ambiguity. Problem 3 is an example of when a manifold is essentially given to us directly. Specifically, the set of all two dimensional subspaces of $\mathbb{R}^5$ is, in a natural way, a manifold. Indeed, it is called the $(5, 2)$-Grassmann manifold. Therefore, even though the subspace tracking problem has a long history in signal processing, its "correct" formulation as a tracking problem on the Grassmann manifold is only recently being realised. The Grassmann manifold will be discussed further in Section 2.

Interesting research questions in the area of differential geometry applied to signal processing include, but are by no means limited to, optimisation on manifolds (Section 3), statistics on manifolds (Section 4), tracking on manifolds (Section 5) and quantisation and interpolation on manifolds [23].

Differential geometry is not always the most relevant tool for dealing with non-linearities. As but one example, when dealing with polynomial equations, algebraic geometry should be used instead [14, 21, 22]. The message is the same though; when faced with a non-linear signal processing problem, first check to see if geometry will provide a systematic framework for tackling it.

2. GRASSMANN AND STIEFEL MANIFOLDS

The (n, p)-Grassmann manifold is the set of all p-dimensional subspaces in $\mathbb{R}^n$. (The complex Grassmann manifold consisting of all p-dimensional subspaces in $\mathbb{C}^n$ exists too.) Since subspaces play a large part in signal processing, it is not surprising that the Grassmann manifold has already turned up in a number of signal processing applications [11, 12, 16, 17, 20, 23, 29].

Describing the Grassmann manifold as a collection of subspaces obscures the way calculations on the Grassmann manifold take place in practice. An equivalent definition of the Grassmann manifold is given below to remedy this. First, it is required to introduce another manifold, the Stiefel manifold, which is also relevant to signal processing.

Consider the set of all matrices $X \in \mathbb{R}^{n \times p}$ satisfying $X^T X = I$. (For the complex-valued version, see [16].) This set naturally forms a manifold and is called the (n, p)-Stiefel manifold. Note that when $p = 1$, the set is a sphere. A point on the Stiefel manifold is simply a matrix X satisfying $X^T X = I$. One way to think of this though is that a point on the Stiefel manifold represents an ordered basis for a p-dimensional subspace of $\mathbb{R}^n$, the columns of X being the basis vectors.

Given a point X on the Stiefel manifold, there corresponds a point on the Grassmann manifold given by the subspace spanned by the columns of X. Mathematically, there exists a projection π from the Stiefel to the Grassmann manifold taking X to $\pi(X)$. This projection is many-to-one, since if Q is any orthogonal matrix then $\pi(XQ) = \pi(X)$ because the subspace spanned by the columns of XQ is the same subspace spanned by the columns of X. Therefore, a point p on the Grassmann manifold can be stored numerically on a computer by storing the matrix X, where X is any point on the Stiefel manifold satisfying $\pi(X) = p$. Indeed, this is simply saying that a subspace is represented by writing down a basis for it, namely the columns of X. (An alternative is to store the projection matrix whose range space corresponds to the point on the Grassmann manifold one wishes to represent.)

In fact, the Stiefel manifold can be used to define the Grassmann manifold from scratch. As hinted at in the Introduction, a new manifold can be created from an existing one by quotienting out an equivalence relation. The Grassmann manifold is the quotient space obtained by starting with the Stiefel manifold and calling two points X, Y equivalent if there exists an orthogonal matrix Q such that $X = YQ$.

3. OPTIMISATION ON MANIFOLDS

Although minimising a cost function defined on a manifold has a relatively long history [6], it has only been studied in the signal processing community recently [7, 13, 15, 16, 17, 18, 20]. A fundamental question people have addressed is, how can the well-known Newton algorithm for optimising a cost function $f : \mathbb{R}^n \to \mathbb{R}$ be generalised to an algorithm for optimising a cost function $f : M \to \mathbb{R}$ defined on some manifold M?

Prior to the work in [16], the solution invariably involved endowing the manifold M with a Riemannian structure. (This is a generalisation of endowing $\mathbb{R}^n$ with an inner product and enables one to measure distances and define geodesics on M.) For example, the Newton method $x_{k+1} = x_k - [f''(x_k)]^{-1} f'(x_k)$ was generalised to a manifold M by first endowing M with a Riemannian structure, thus allowing $f'(x_k)$ and $f''(x_k)$ to be replaced by the gradient and Hessian of f at x_k (without a Riemannian structure the gradient and Hessian are undefined), and then replacing the vector addition $x_{k+1} = x_k + \delta_k$, $\delta_k = -[f''(x_k)]^{-1} f'(x_k)$ by setting x_{k+1} to be the point reached by starting at x_k and moving along the geodesic trajectory defined by δ_k for a distance of one unit. In the Euclidean case, when geodesics become straight lines, this reduces back to the ordinary Newton method.

Although this Riemannian-based approach is sensible, there are at least four disadvantages. 1) Moving along geodesics can be a computationally expensive operation. 2) Usually the cost function f has nothing to do with the Riemannian geometry so using this artificial Riemannian geometry may actually slow down the convergence. An example is given presently. 3) It does not follow immediately (i.e. it has to be proved) that the Riemannian-based algorithm has the same rate of convergence as it does in Euclidean space. 4) There exist interesting manifolds for which no *explicit* expression for a Riemannian metric is known. The Riemannian-based algorithms are thus inapplicable in practice to such manifolds.

A new paradigm for optimisation on manifolds was introduced in [16], aimed at avoiding these disadvantages. The idea is relatively straightforward; since a manifold M locally looks like $\mathbb{R}^n$, about any point $p \in M$, there always exists a parametrisation (in fact, there exist infinitely many) $\phi_p : \mathbb{R}^n \to M$ mapping (diffeomorphically) $\mathbb{R}^n$ into a subset of M containing p. (For convenience, assume $\phi_p(0) = p$.) This means that the cost function $f : M \to \mathbb{R}$ can be pulled back to the *local cost function* $f \circ \phi_p : \mathbb{R}^n \to \mathbb{R}$. Moreover, the point $p \in M$ corresponds to the origin in $\mathbb{R}^n$ by the assumption $\phi_p(0) = p$. One step of the Newton algorithm (or any other optimisation algorithm) can be applied to this local cost function to move from the origin to some point $z \in \mathbb{R}^n$. This point is now mapped back to the manifold, namely

to the point $\phi_p(z)$. This completes one step of the algorithm. Note that at the new point $\phi_p(z)$ a different parametrisation is used, and this is what allows the algorithm to move around the whole manifold rather than just be confined to part of the manifold.

It is important to note that this approach is a generalisation of the Riemannian approach; a particular choice of ϕ_p (namely, the Riemannian exponential map centred at p) will make the above algorithm identical to the Riemannian-based algorithm.

By choosing ϕ_p differently, the computational cost of each iteration can often be reduced significantly, thus overcoming Disadvantage 1 above.

How Disadvantage 2 can be overcome is explained by way of a trivial example. Consider the cost function $f(x, y) = (x \cos y - 2)^2 + (x \sin y - 3)^2$. The ordinary Newton algorithm can be thought of as the Riemannian-based algorithm with respect to the usual Euclidean inner product on $\mathbb{R}^2$. Clearly, it will take an infinite number of iterations for this algorithm to converge to the minimum if it is started away from the minimum. By comparison, if the parametrisation ϕ is chosen to map Cartesian coordinates into polar coordinates, then the local cost function $(f \circ \phi)(x, y)$ becomes quadratic and the new algorithm converges in a single iteration.

In general, a strategic choice of parametrisations can be used to reduce the contribution of the third and higher order terms in the Taylor series expansion of the local cost function, resulting in faster convergence. (Technically, the actual rate of convergence is usually the same, e.g., quadratic, but it is the constant out the front which can be decreased.)

It is remarked that the one time when it is very appropriate to consider the Riemannian-based version is when the cost function is related to the Riemannian geometry. A prime example is when the cost function involves the Riemannian distance function, such as in "centre of mass" type problems [18].

Disadvantage 3 is overcome in a paper to be submitted soon, which proves that under very mild conditions on the parametrisations ϕ_p, the local convergence properties of the generalised algorithm are equivalent to the local convergence properties of the underlying algorithm. For example, since the Newton algorithm is locally quadratically convergent, so too is the generalised algorithm for all "sensible" choices of parametrisations ϕ_p. In fact, this forthcoming paper will also point out that if the Euclidean Newton algorithm is written as $N_g(x) = x - [g''(x)]^{-1} g'(x)$ then the generalised algorithm takes the form $E(p) = (\phi_p \circ N_{f \circ \phi_p} \circ \phi_p^{-1})(p)$, showing explicitly that what is happening is that the generalised algorithm is essentially the underlying algorithm but with respect to a different coordinate system at each iteration.

Disadvantage 4 is overcome because manifolds always come with parametrisations (or coordinate charts).

When told that the Riemannian structure of a manifold won't be used in an optimisation problem, people often ask where the geometry comes into play then. The answer is that geometry helps in choosing the parametrisations ϕ_p; note that a non-Riemannian manifold can still have "geometry" or "structure" which can be exploited. For example, the parametrisations in [16] were defined by using Euclidean projections from the tangent spaces onto the manifold. The cost per iteration is lower than a Riemannian-based approach, and moreover, for certain cost functions relevant to signal processing, empirical evidence suggests fewer iterations are required to converge to within a specific tolerance of the minimum.

An important research area is the design of parametrisations ϕ_p for specific classes of cost functions and specific manifolds.

4. STATISTICS ON MANIFOLDS

Manifolds can enter statistics in a number of ways. In parameter estimation problems, it might be known that the parameter belongs to a curved surface. Similarly, the observations might lie on a manifold. It is natural to endeavour to extend statistical concepts and procedures to these cases: [3, 25, 26, 28] and references therein.

For example, [8] derives a Cramér-Rao type lower bound for manifold-valued parameters. However, it does not directly provide a quantitative measure of how accurate an unbiased estimator can be because it chose not to introduce a distance function on the parameter space. The main contribution of [28] is to provide such a quantitative measure, and moreover, a feature of the Cramér-Rao type lower bound in [28] is that it reduces to the ordinary Cramér-Rao lower bound when the manifold is Euclidean space.

Another example is that in Euclidean space, if the usual Fisher information matrix is singular, it is common for people to derive a Cramér-Rao type lower bound by taking the pseudo-inverse of the Fisher information matrix. In [27], a geometric meaning is attributed to this. Specifically, it is shown how the Cramér-Rao type lower bound can be obtained by the geometric path in [8].

Differential geometry also interacts with statistics on the following, deeper footing. Certain sets of probability density functions can be made naturally into manifolds; that is, each point on the manifold corresponds to a probability density function [1, 2]. Quantities such as Kullback-Leibler divergence can then be given a geometric interpretation. As but one application of this, a geometric interpretation of certain decoding algorithms is given in [9, 10]. This interpretation allows one to picture what is going on and thus better understand the properties of these decoding algorithms.

5. TRACKING ON MANIFOLDS

Although the myriad papers on subspace tracking can be interpreted as tracking on the Grassmann manifold, few papers have approached the subspace tracking problem in this way [25]. Moreover, even when manifolds are used [24], often the tracking is done simply by iteratively minimising a cost function, as opposed to estimating optimally the subspace based on not only the observations but also a stochastic model of the time-varying subspace.

A potential research area is to extend filtering algorithms, such as the Kalman filter, to non-linear settings. Specifically, the linear state-space model ($x_{k+1} = Ax_k + Bw_k$, $y_k = Cx_k + Dn_k$) can be generalised by replacing the linear operations by group actions. An example of a group action is a rotation applied to a point on a sphere; if x_k is a point on the sphere and R denotes a certain rotation then it makes sense to set $x_{k+1} = R \cdot x_k$ where $\cdot$ denotes the group action. The noise can similarly be modelled by another group action, this time a random one. That is, the state equation becomes $x_{k+1} = W \cdot R \cdot x_k$ where W is a random group action and R a known one. The observation equation can be modelled by a random group action for the noise and a projection operation in place of the C matrix.

It is remarked that the potential disadvantage of algorithms on manifolds is that the computational cost might be relatively high. This is not always the case (and often there is no choice but to work on a manifold albeit implicitly), but it was the motivation behind the subspace flows in [19] which were intentionally defined over $\mathbb{R}^n$ so that numerical implementations of them wouldn't have to compute projections onto Stiefel manifolds.

6. CONCLUSION

Differential geometry is an indispensable tool for solving certain non-linear problems in signal processing. The advantages of differential geometry include i) being able to better understand existing algorithms, often leading to the derivation of even better algorithms; ii) providing an appropriate mindset for solving non-linear problems; iii) allowing existing results in mathematics to be applied directly.

7. REFERENCES

[1] S.-I. Amari and H. Nagaoka. *Methods of Information Geometry*. Oxford University Press, 2000.

[2] O. E. Barndorff-Nielsen, R. D. Cox, and N. Reid. The role of differential geometry in statistical theory. *International Statistics Review*, 54:83–96, 1986.

[3] R. Bhattacharyq and V. Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds I. *The Annals of Statistics*, 31(1):1–29, 2003.

[4] W. M. Boothby. *An Introduction to Differentiable Manifolds and Riemannian Geometry*. Academic Press, second edition, 1986.

[5] D. A. Cox, J. B. Little, and D. O'Shea. *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer-Verlag, 2nd edition, 1996.

[6] D. Gabay. Minimizing a differentiable function over a differentiable manifold. *Journal of Optimization Theory and Applications*, 37(2):177–219, 1982.

[7] U. Helmke, K. Hüper, and J. B. Moore. Quadratically convergent algorithms for optimal dextrous hand grasping. *IEEE Transactions on Robotics and Automation*, 18(2):138–146, 2002.

[8] H. Hendricks. A Cramér-Rao type lower bound for estimators with values in a manifold. *Journal of Multivariate Analysis*, 38:245–261, 1991.

[9] S. Ikeda. Information geometry of turbo and LDPC codes. In *International Conference on Acoustics, Speech and Signal Processing*, Philadelphia, March 2005.

[10] S. Ikeda, T. Tanaka, and S. Amari. Information geometry of Turbo and Low-Density Parity Check codes. *IEEE Transactions on Information Theory*, 50(6):1097–1114, 2004.

[11] I. Kammoun and J.-C. Belfiore. A new family of Grassmann space-time codes for non-coherent MIMO systems. *IEEE Communications Letters*, 7(11):528–530, 2003.

[12] X. Liu, A. Srivastava, and K. Gallivan. Optimal linear representations of images for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):662–666, 2004.

[13] R. E. Mahony and J. H. Manton. The geometry of the Newton method on non-compact Lie groups. *Journal of Global Optimization*, 23(3):309–327, August 2002.

[14] J. H. Manton. Finite alphabet source recovery in polynomial systems. *Systems and Control Letters*, 47(4):279–286, November 2002.

[15] J. H. Manton. An improved least squares blind channel identification algorithm for linearly and affinely precoded communication systems. *IEEE Signal Processing Letters*, 9(9):282–285, September 2002.

[16] J. H. Manton. Optimisation algorithms exploiting unitary constraints. *IEEE Transactions on Signal Processing*, 50(3):635–650, March 2002.

[17] J. H. Manton. Design and analysis of linear precoders under a mean square error criterion, Part II: MMSE designs and conclusions. *Systems and Control Letters*, 49(2):131–140, 2003.

[18] J. H. Manton. A globally convergent numerical algorithm for computing the centre of mass on compact Lie groups. In *Eighth International Conference on Control, Automation, Robotics and Vision*, Kunming, China, December 2004.

[19] J. H. Manton, U. Helmke, and I. M. Y. Mareels. A dual purpose principal and minor component flow. *Systems and Control Letters*, 2005. In Press.

[20] J. H. Manton, R. Mahony, and Y. Hua. The geometry of weighted low rank approximations. *IEEE Transactions on Signal Processing*, 51(2):500–514, 2003.

[21] J. H. Manton and W. D. Neumann. Totally blind channel identification by exploiting guard intervals. *Systems and Control Letters*, 48(2):113–119, 2003.

[22] J. H. Manton, W. D. Neumann, and P. T. Norbury. On the algebraic identifiability of finite impulse response channels driven by linearly precoded signals. *Systems and Control Letters*, 54(2):125–134, 2005.

[23] B. Mondal, R. W. Heath, Jr., and L. W. Hanlen. Quantization on the Grassmann manifold: Applications to precoded MIMO wireless systems. In *International Conference on Acoustics, Speech and Signal Processing*, Philadelphia, March 2005.

[24] M. Nikpour, K. Hüper, and J. H. Manton. Generalizations of the Rayleigh Quotient iteration for the iterative refinement of the eigenvectors of real symmetric matrices. In *International Conference on Acoustics, Speech and Signal Processing*, Philadelphia, March 2005.

[25] A. Srivastava. A Bayesian approach to geometric subspace estimation. *IEEE Transactions on Signal Processing*, 48(5):1390–1400, 2000.

[26] A. Srivastava and E. Klassen. Monte Carlo extrinsic estimators of manifold-valued parameters. *IEEE Transactions on Signal Processing*, 50(2):299–308, 2002.

[27] J. Xavier and V. Barroso. The Riemannian geometry of certain parameter estimation problems with singular Fisher information matrices. In *International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 1021–1024, Montreal, Canada, May 2004.

[28] J. Xavier and V. Barroso. Intrinsic variance lower bound (IVLB): An extension of the Cramér-Rao Bound to Riemannian manifolds. In *International Conference on Acoustics, Speech and Signal Processing*, Philadelphia, March 2005.

[29] L. Zheng and D. N. C. Tse. Communication on the Grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel. *IEEE Transactions on Information Theory*, 48(2):359–383, 2002.