Mathematisches Forschungsinstitut Oberwolfach

Report No. 36/2020

DOI: 10.4171/OWR/2020/36

Mini-Workshop: Computational Optimization on Manifolds (online meeting)

Organized by
Pierre-Antoine Absil, Louvain-la-Neuve
Roland Herzog, Chemnitz
Gabriele Steidl, Berlin

15 November – 21 November 2020

Abstract. The goal of the mini-workshop was to study the geometry, algorithms and applications of unconstrained and constrained optimization problems posed on Riemannian manifolds. Focus topics included the geometry of particular manifolds, the formulation and analysis of a number of application problems, as well as novel algorithms and their implementation.

Mathematics Subject Classification (2010): 90C30 nonlinear programming, 90C46 optimality conditions, 49Q99 manifolds (in the context of optimization), 65K05 math programming methods.

Introduction by the Organizers

The mini-workshop *Computational Optimization on Manifolds* was organized by Pierre-Antoine Absil (Louvain-la-Neuve), Roland Herzog (Chemnitz) and Gabriele Steidl (Berlin). Due do the ongoing pandemic, it was held as an online event with a reduced program. Nonetheless, the mini-workshop was well attended by a total of 17 participants from Europe, Asia and North America.

The topics of the event revolved around practical aspects for optimization problems on manifolds, where the nonlinear Riemannian geometry often presents difficulties not present in an Euclidean setting. A challenging example of such a problem, the bivariate fitting of manifold-valued data by minimizing a linear combination of a thin plate spline energy and a data fidelity term, was presented and analyzed in the talk by Benedikt Wirth. A second example, discussing solutions of the parabolic heat equation in the space of tensor product functions of low rank,

was presented by André Uschmajew. Here the solution of each time step was posed as an optimization problem.
Two further talks elaborated on the geometry of two particular spaces appearing frequently in applications. The presentation by Nicolas Boumal focused on the algebraic variety of bounded rank matrices, whose failure to be a smooth manifold presents further challenges to theory and algorithms. One way to overcome these challenges is by lifting the problem onto a manifold. A second presentation given by Estelle Massart considered the manifold of fixed-rank positive semidefinite matrices. A quotient geometry was proposed, which allows for an efficient evaluation of the exponential and logarithmic maps.

The remaining talks focused on particular algorithms and their realization in software. Hiroyuki Sato reviewed nonlinear conjugate gradient methods on Riemannian manifolds. Ronny Bergmann spoke about a Riemannian version of the primal-dual hybrid gradient (Chambolle–Pock) algorithm, which relies on a recent proposal of the Fenchel conjugate on manifolds. He also discussed implementation in a new toolbox for optimization on manifolds written in Julia. The presentation by Jan Lellmann introduced a semismooth Newton method for the same class of problems, whose implementation was also achieved in the Julia toolbox. In her talk, Nina Miolane introduced a Python package for computations, statistics and the solution of machine learning tasks on nonlinear manifolds. She also presented applications across various disciplines. Finally, Anton Schiela addressed a sequential quadratic programming method for equality constrained optimization problems on Riemannian manifolds with applications in optimal control.

Overall, the mini-workshop demonstrated that computational optimization on Riemannian manifolds is a very active research area with numerous applications and challenges not present in Euclidean spaces.

Acknowledgement: The organizers would like to thank MFO staff for providing the technical infrastructure to conduct this mini-workshop as an online event. They would also like to thank Estefanía Loayza-Romero for her help with the online sessions and the preparation of this report.

Mini-Workshop (online meeting): Computational Optimization on Manifolds

Table of Contents

Abstracts

Optimization of Smooth Functions on Nonsmooth Sets

Nicolas Boumal
(joint work with Eitan Levin, Joe Kileel)

We consider the problem of minimizing a smooth function f restricted to a possibly nonsmooth set $\mathcal{X} \subseteq \mathcal{E}$, where $\mathcal{E}$ is a Euclidean space:

$$\min_{x \in \mathcal{X}} f(x).$$

As an example of interest, we may consider $\mathcal{X}$ being the real algebraic variety consisting in all matrices of a given size and bounded rank:

$$\mathcal{X} = \{X \in \mathbb{R}^{m \times n} : \operatorname{rank}(X) \leq r\}.$$

If $\mathcal{X}$ were an embedded submanifold of $\mathcal{E}$, we could describe it locally using its tangent spaces. In the general case, we can describe $\mathcal{X}$ locally using its tangent cones:

$$\mathrm{T}_x\mathcal{X} = \Big\{v \in \mathcal{E} : \text{ there exist sequences } (x_n) \subset \mathcal{X} \text{ and } (\tau_n) \subset \mathbb{R}_+ \text{ satisfying } x_n \to x, \tau_n \to 0 \text{ and } v = \lim_{n\to\infty} \frac{x_n - x}{\tau_n}\Big\}.$$

Using the gradient ∇f of f, a reasonable measure of stationarity for a point $x \in \mathcal{X}$ is

$$\left\|\mathrm{Proj}_{\mathrm{T}_x\mathcal{X}}(-\nabla f(x))\right\|,$$

where $\|\cdot\|$ denotes the norm on $\mathcal{E}$ and $\mathrm{Proj}_{\mathrm{T}_x\mathcal{X}}$ denotes metric projection (in that same norm) to the tangent cone at x. When this quantity is zero, we say x is stationary.

Since local (and global) minimizers of f are stationary, it is of interest to determine whether optimization algorithms admit only stationary points as limit points. Unfortunately, the nonsmoothness of $\mathcal{X}$ can make this quite challenging. In particular, we exhibit a function f on the rank variety and an initialization X_0 for which a reasonable algorithm studied by Schneider and Uschmajew (2015) produces a sequence (X_n) which has the following properties:

(1) All matrices in the sequence have rank r,
(2) the sequence converges to a matrix X of rank less than r,
(3) the stationarity measure converges to zero along the sequence,
(4) yet X is not stationary.

We call such events *apocalypses*: they are clear liabilities for optimization purposes.

A standard workaround for optimization (and other endeavors) on sets with less than desirable properties is to resort to a *lift*. Specifically, assume we have at our disposal a smooth manifold $\mathcal{M}$ together with a smooth map $\varphi\colon \mathcal{M} \to \mathcal{E}$ such that

$\varphi(\mathcal{M}) = \mathcal{X}$. Then, with $g = f \circ \varphi$ the problem above can be stated equivalently as:

$$\min_{y \in \mathcal{M}} g(y).$$

Conveniently, both $\mathcal{M}$ (by assumption) and g (by composition) are smooth. For the bounded-rank variety, we may consider $\mathcal{M} = \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$ and $\varphi(L, R) = LR^\top$, or $\mathcal{M} = \{(X, K) \in \mathbb{R}^{m\times n} \times \mathrm{Gr}(n, n - r) : K \subseteq \ker X\}$ and $\varphi(X, K) = X$ with Gr denoting Grassmannians; the former is standard, the latter is called the *desingularization lift* Khrulkov and Oseledets (2018).

A number of natural questions ensue: how do {global optima, local optima, stationary points, approximate stationary points} of the two problems compare? And if f has certain desirable properties for the purpose of optimization, can we expect g to retain some of these properties? If not, can we compensate for those losses?

In the talk, we go over those questions one by one, methodically establishing full characterizations where we could do so, and highlighting further questions that arose along the way. This is ongoing work.

References

V. Khrulkov and I. Oseledets. Desingularization of bounded-rank matrix sets. *SIAM Journal on Matrix Analysis and Applications*, **39**(1):451–471, 2018. DOI:10.1137/16M1108194.

R. Schneider and A. Uschmajew. Convergence results for projected line-search methods on varieties of low-rank matrices via Łojasiewicz inequality. *SIAM Journal on Optimization*, **25**(1):622–646, 2015. DOI: 10.1137/140957822.

Thin Plate Spline Type Fitting to Riemannian Data

Benedikt Wirth

(joint work with Pierre-Antoine Absil, Pierre-Yves Gousenbourger)

Given coordinate-data pairs $(t_i, \mathfrak{d}_i) \in [0, 1] \times \mathbb{R}$, $i = 1, \dots, K$, it is a classical and straightforward task to find a curve $\overline{\mathbf{S}} : [0, 1] \to \mathbb{R}$ that interpolates the data or approximates it, for instance in a least squares sense. A prototypical and widespread method is approximation with a cubic spline, which is smooth, has a simple infinite-dimensional variational interpretation of minimizing accumulated squared acceleration, yet can be parameterized by a finite number of parameters (and thus is straightforward to compute), even if the coordinates t_i are nonequispaced. These properties persist in an approximate sense if the data points $\mathfrak{d}_i$ lie in a Riemannian manifold $\mathcal{M}$ with metric g and an approximating curve $\overline{\mathbf{S}} : [0, 1] \to \mathcal{M}$ is sought.

Dealing with multivariate approximation of manifold-valued data is more difficult, already in the bivariate case which we consider throughout. Letting $\Omega \subset \mathbb{R}^2$ be open and connected with smooth boundary, consider first Euclidean data

$(t_i, \mathfrak{d}_i) \in \Omega \times \mathbb{R}$, $i = 1, \dots, K$. The bivariate generalization of cubic spline fitting would be to find $\overline{\mathbf{S}} : \Omega \to \mathbb{R}$ by solving

$$\min \left\{ E[\overline{\mathbf{S}}] + \lambda \sum_{i=1}^{K} |\overline{\mathbf{S}}(t_i) - \mathfrak{d}_i|^2 \,\middle|\, \overline{\mathbf{S}} : \Omega \to \mathbb{R} \right\} \text{ with } E[\overline{\mathbf{S}}] = \int_\Omega |D^2 \overline{\mathbf{S}}(t)|^2 \,\mathrm{d}t = |\overline{\mathbf{S}}|_{H^2}^2$$

for some fixed weight $\lambda > 0$ (the limit $\lambda \to \infty$ leads to an interpolation problem, $\lambda \to 0$ leads to linear regression). The function $\overline{\mathbf{S}}$ is known as a *thin plate spline*, and the corresponding linear Euler–Lagrange equation

$$\Delta^2 \overline{\mathbf{S}} = \lambda \sum_{i=1}^{K} (\mathfrak{d}_i - \overline{\mathbf{S}}(t_i)) \delta_{t_i} \text{ in } \Omega, \quad n \cdot D^2 \overline{\mathbf{S}} n = 0, \; \partial_n \Delta \overline{\mathbf{S}} + \partial_{n^\perp} (n^\perp \cdot D^2 \overline{\mathbf{S}} n) = 0 \text{ on } \partial\Omega$$

(with δ_t the Dirac distribution in $t \subset \Omega$ and n the unit outward normal to $\partial\Omega$) has a unique solution that can readily be solved for numerically. The situation is further simplified if $\Omega = \mathbb{R}^2$, in which case the minimizer turns out to be of the form

$$\overline{\mathbf{S}}(t) = \textstyle\sum_{i=1}^{K} \alpha_i \phi(|t - t_i|) + \beta \cdot \binom{t}{1}$$

with $\phi(r) = r^2 \log r$ being the radially symmetric fundamental solution to $\Delta^2 \phi = 8\pi\delta_0$. In this case the minimization problem reduces to solving a linear system for the coefficients $\alpha_i \in \mathbb{R}$ and $\beta \in \mathbb{R}^3$.

Now consider instead a manifold-valued bivariate fitting function $\mathbf{S} : \Omega \to \mathcal{M}$. We define the *thin plate spline energy* of $\mathbf{S}$ as

$$E[\mathbf{S}] = \int_\Omega \operatorname{tr}([D^2 \mathbf{S}(t)]^* D^2 \mathbf{S}(t)) \,\mathrm{d}t,$$

where the second derivative of $\mathbf{S}$ at $t \in \mathbb{R}^2$ is the bounded bilinear form

$$D^2 \mathbf{S}(t) : \mathbb{R}^2 \otimes \mathbb{R}^2 \to T_{\mathbf{S}(t)} \mathcal{M}, \quad (v, w) \mapsto \nabla^{\mathbf{S}}_w (\partial_v \mathbf{S})(t).$$

The term $\nabla^{\mathbf{S}}_w \sigma$ essentially equals $\nabla_{D\mathbf{S}w} (\sigma \circ \mathbf{S}^{-1}) \circ \mathbf{S}$ with ∇ the classical Levi-Civita covariant derivative, however, this formula is only valid if $\mathbf{S}$ is locally invertible.

Definition 1. A *manifold thin plate spline* approximating data $(t_1, d_1), ..., (t_K, d_K) \in \Omega \times \mathcal{M}$ with fitting weight λ is a solution of the minimization problem

$$\min \left\{ E[\mathbf{S}] + \lambda \textstyle\sum_{i=1}^{K} \operatorname{dist}(\mathbf{S}(t_i), d_i)^2 \,\middle|\, \mathbf{S} : \Omega \to \mathcal{M} \right\}. \tag{1}$$

The following statements illustrate different problems with this generalization: the Euler–Lagrange equation for (1) is highly nonlinear and thus difficult to solve, there may not even exist a global solution to (1), and solutions to (1) are prone to degenerating to curves.

Theorem 2. *A solution to* (1)*, smooth away from* $t_1, \dots, t_K$*, satisfies*

$$0 = \lambda \textstyle\sum_{i=1}^{K} \delta_{t_i} \log_{\mathbf{S}(t_i)} d_i - \sum_{i,j=1,2} \left(\nabla^{\mathbf{S}}_{\partial_i} \nabla^{\mathbf{S}}_{\partial_j} \nabla^{\mathbf{S}}_{\partial_j} \partial_i \mathbf{S} + R(\nabla^{\mathbf{S}}_{\partial_i} \partial_j \mathbf{S}, \partial_i \mathbf{S}) \partial_j \mathbf{S} \right) \text{ in } \Omega,$$

$$\begin{aligned} 0 = \nabla^{\mathbf{S}}_n \partial_n \mathbf{S} = \nabla^{\mathbf{S}}_{n^\perp} (\nabla^{\mathbf{S}}_{n^\perp} \partial_n \mathbf{S}) + \nabla^{\mathbf{S}}_n \left(\nabla^{\mathbf{S}}_{\partial_1} \partial_1 \mathbf{S}(s,t) + \nabla^{\mathbf{S}}_{\partial_2} \partial_2 \mathbf{S}(s,t) \right) \\ + R(\partial_1 \mathbf{S}, \partial_n \mathbf{S}) \partial_1 \mathbf{S} + R(\partial_2 \mathbf{S}, \partial_n \mathbf{S}) \partial_2 \mathbf{S} \qquad \text{on } \partial\Omega, \end{aligned}$$

where R is the Riemann curvature tensor and $\delta_{\bar{t}}$ is defined by $\int_\Omega g_{\mathbf{S}(t)}(\delta_{\bar{t}}\phi(t), \psi(t))\,\mathrm{d}t = g_{\mathbf{S}(\bar{t})}(\phi(\bar{t}), \psi(\bar{t}))$ for any continuous liftings $\phi, \psi : \Omega \to T\mathcal{M}$ of $\mathbf{S}$.

Theorem 3. *The manifold $\mathcal{M}$ and data $d_1, \dots, d_K \in \mathcal{M}$ can be chosen such that* (1) *has no global solution.*

Theorem 4. *One has $E[\mathbf{S}] = 0$ if and only if one of the following holds.*

1. *$\mathbf{S}(t) = y \in \mathcal{M}$ is constant, or*
2. *$\mathbf{S}(t) = \gamma(t \cdot v)$ for a constant speed geodesic $\gamma : \mathbb{R} \to \mathcal{M}$ and a $v \in \mathbb{R}^2$, or*
3. *$\mathbf{S}(t) = i(At)$ for a symmetric positive definite $A \in \mathbb{R}^{2\times 2}$ and a local isometry $i : A(\Omega) \to \mathcal{M}$ such that $s \mapsto i(as + b)$ is a constant speed geodesic in $\mathcal{M}$ for any $a, b \in \mathbb{R}^2$, thus $i(\Omega)$ is a flat geodesic submanifold.*

Corollary 5. *Let $\mathbf{S}^\lambda$ denote a global minimizer of* (1) *under an additional bound on the $W^{1,4}$-norm of $\mathbf{S}$ (one can show that such an $\mathbf{S}^\lambda$ exists for any finite bound), and let $\lambda_k \searrow 0$. If $\mathcal{M}$ does not contain a flat geodesic submanifold, then any sequence $\mathbf{S}^{\lambda_k}$ contains a subsequence that converges uniformly to a degenerate surface as in cases* (1)-(2) *of the previous theorem.*

Unless $\mathcal{M}$ is a product manifold $\mathbb{R} \times \mathcal{N}$, $\mathcal{M}$ does usually not contain a flat geodesic submanifold. Thus, as one chooses λ small to achieve a bivariate approximation that behaves like linear regression in the Euclidean case or so-called geodesic regression in the manifold-valued univariate case, then instead one will typically obtain an almost degenerate fitting function $\mathbf{S}$.

Since the concept of manifold thin plate splines entails so many problems, a natural remedy is to simply linearize the manifold and work in tangent space.

Definition 6. Let $p \in \mathcal{M}$ and $\mathfrak{d}_i \in T_p\mathcal{M}$ with $\exp_p \mathfrak{d}_i = d_i$ for $i = 1, \dots, K$. The *tangent space thin plate spline* $\overline{\mathbf{S}} : \Omega \to T_p\mathcal{M}$ for data $(t_1, d_1), \dots, (t_K, d_K) \in \Omega \times \mathcal{M}$ and fitting parameter $\lambda > 0$ is the unique solution of

$$\min \left\{ \int_\Omega |D^2\overline{\mathbf{S}}|^2 \,\mathrm{d}t + \lambda \sum_{i=1}^K |\mathfrak{d}_i - \overline{\mathbf{S}}(t_i)|^2 \;\middle|\; \overline{\mathbf{S}} : \Omega \to T_p\mathcal{M} \right\}.$$

The associated *retracted thin plate spline* is the map $\mathbf{S} = \exp_p \circ \overline{\mathbf{S}}$.

Of course, the retracted thin plate spline will not have equally good smoothness properties as a manifold thin plate spline, so a natural question is how much it actually differs from the (more intrinsic and thus more natural concept of the) manifold thin plate spline. We will quantify this difference for the case of interpolation, $\lambda = \infty$, since for $\lambda < \infty$ we have seen that there is a strong bias towards degenerate fitting functions. The difference will depend on how close $\mathcal{M}$ is to Euclidean space in the following sense.

Definition 7. A Riemannian manifold $\mathcal{M}$ shall be called *ϵ-flat in C^k on $B_r(p)$* (with $B_r(p) \subset \mathcal{M}$ the closed geodesic ball around $p \in \mathcal{M}$ of radius r) if in normal coordinates at p we have $\|g - I\|_{C^k(B_r(0))} < \epsilon$ for the metric $g : B_r(0) \to \mathbb{R}^{n\times n}$.

A more geometric view on ϵ-flatness is provided by the following statement.

Theorem 8. *There exists $C > 0$ (depending on the dimension of $\mathcal{M}$) such that*

(1) *if the sectional curvatures of $\mathcal{M}$ are bounded in absolute value by ϵ and $r < \frac{\pi}{2\sqrt{\epsilon}}$, then $\mathcal{M}$ is $Cr\epsilon$-flat on $B_r(p)$ in C^0,*
(2) *if in addition $\nabla_u R(u,v)v - \nabla_v R(v,u)u \leq \epsilon|u|^2|v|^2$ for the curvature tensor R and all tangent vectors v, u, then $\mathcal{M}$ is $Cr\epsilon$-flat on $B_r(p)$ in C^1.*

Now one can estimate the difference between retracted and manifold thin plate splines. This estimate will be independent of the manifold injectivity radius, whose finiteness usually causes the nonexistence of global minimizers to (1).

Theorem 9. *Let $\lambda = \infty$ and $m \in \{0,1,2\}$, $q \in [2,\infty]$ such that $H^2(\Omega) \subset W^{m,q}(\Omega)$, and define the grid width $h = \sup_{t\in\Omega} \inf_i |t - t_i|$. Let $\overline{\mathbf{S}}$ be the tangent space thin plate spline at $p \in \mathcal{M}$ for the given data. There exist $C, h_0 > 0$ depending on Ω, m, q and $\delta, r > 0$ depending on Ω and $\overline{\mathbf{S}}$ such that (1) has a local minmizer of the form $\mathbf{S} = \exp_p \circ S$ and*

$$\|D^m(\overline{\mathbf{S}} - S)\|_{L^q(\Omega)} \leq Ch^{1+\frac{2}{q}-m}\sqrt{\epsilon}(\|\overline{\mathbf{S}}\|_{H^2} + \|\overline{\mathbf{S}}\|_{H^2}^2)$$

whenever $h < h_0$, $\epsilon < \delta$, and $\mathcal{M}$ is ϵ-flat in C^1 on $B_r(p)$.

Of course, in applications $\mathcal{M}$ will not be ϵ-flat on large enough balls, so to maintain closeness to intrinsically defined manifold thin plate splines one needs to localize. We suggest localization by blending multiple (local) retracted thin plate splines at different base points $p \in \mathcal{M}$ into one global fitting function as defined below. One can show that for appropriate base points and blending weights one achieves globally C^1 functions whose evaluation at a point only requires few Riemannian exponentials and logarithms, and one can devise appropriate weights, adaptive refinement schemes, and schemes to transfer the data $d_1, \ldots, d_K$ to the different tangent spaces.

Definition 10. Partition Ω into rectangles $\Omega_{ij} = [x_i, x_{i+1}] \times [y_j, y_{j+1}]$ and associate each (x_k, y_l) with a base point $p_{kl} \in \mathcal{M}$. Given retracted thin plate splines $\mathbf{S}_{kl}$ at p_{kl}, the *blended surface* $\mathbf{S} : \Omega \to \mathcal{M}$ is defined on each Ω_{ij} by the (well-known) Riemannian weighted averaging

$$\mathbf{S}(x,y) = \mathbf{av}\big(\mathbf{S}_{ij}(x,y), \mathbf{S}_{i+1,j}(x,y), \mathbf{S}_{i,j+1}(x,y), \mathbf{S}_{i+1,j+1}(x,y)\,;\\ w_{ij}(x,y), w_{i+1,j}(x,y), w_{i,j+1}(x,y), w_{i+1,j+1}(x,y)\big),$$

with smooth weights $w_{kl} : \Omega \to [0,1]$ of support in $[x_{k-1}, x_{k+1}] \times [y_{l-1}, y_{l+1}]$.

Dynamical Low-Rank Approximation for Parabolic Problems

André Uschmajew
(joint work with Markus Bachmayr, Henrik Eisenmann, Emil Kieri)

We show existence and uniqueness of dynamical low-rank approximations for parabolic problems in Hilbert spaces (Bachmayr et al., 2020). As a model problem for a more general setup, consider a two-dimensional parabolic partial differential equation

$$u_t(x,t) - \nabla \cdot \alpha(t)\nabla u(x,t) = f(x,t) \tag{1}$$

on a product domain $x \in \Omega = (0,1)^2$, with zero Dirichlet boundary conditions $u(x,t) = 0$ for $(x,t) \in \partial\Omega \times (0,T)$, and an initial value $u(x,0) = u_0(x)$ a.e. in Ω. For this particular problem, the dynamical low-rank approximation (DLRA) approach would ask for an approximate solution curve on a set

$$\mathcal{M}_r = \{u \in L_2(\Omega) \colon \operatorname{rank}(u) = r\}$$

of 'rank-r' functions, that is, functions that have a decomposition

$$u = \sum_{k=1}^{r} u_k^1 \otimes u_k^2$$

into a fixed number r (and not less) tensor product functions $(u^1 \otimes u^2)(x_1, x_2) = u^1(x_1)u^2(x_2)$ a.e. It can be shown that $\mathcal{M}_r$ is a locally embedded submanifold in $L_2(\Omega)$. The functions in $\mathcal{M}_r$ are analogous to infinite rank-r matrices. DLRA hence provides a separation of variables and allows for a low-parametric representation of the solution. It is particularly well studied for finite matrices, and has been used for different classes of evolution problems in scientific computing; see, e.g., Koch and Lubich (2007); Sapsis and Lermusiaux (2009); Lubich and Oseledets (2014); Musharbash et al. (2015); Einkemmer and Lubich (2018); Mena et al. (2018); Ostermann et al. (2019). Here we consider parabolic problems.

To obtain a well posed problem in the parabolic case with mild regularity assumptions one has to work with a weak formulation. Using the Hilbert spaces

$$H = L_2(\Omega) = L_2(0,1) \otimes L_2(0,1), \qquad V = H_0^1(\Omega),$$

let $a : V \times V \times [0,T] \to \mathbb{R}$ denote the induced bilinear form of the differential operator in (1), which we assume to be uniformly symmetric, bounded and coercive. A suitable weak formulation of DLRA is based on a time-dependent variational principle, also called Dirac-Frenkel principle: Given $f \in L_2(0,T;V^*)$ and $u_0 \in \mathcal{M}_r$, find

$$u \in W_2^1(0,T;V,H) = \{u \in L_2(0,T;V) \colon\ u' \in L_2(0,T;H)\}$$

such that $u(t) \in \mathcal{M}_r$ for all $t \in [0,T]$, and such that for almost all $t \in (0,T)$,

$$\begin{gathered} \langle u'(t), v\rangle + a(u(t), v; t) = \langle f(t), v\rangle \quad \text{for all } v \in T_{u(t)}\mathcal{M}_r \cap V, \\ u(0) = u_0. \end{gathered} \tag{2}$$

Here

$$T_u\mathcal{M}_r = \left\{ v = \sum_{k=1}^{r} v_k^1 \otimes u_k^2 + u_k^1 \otimes v_k^2 : \ v_k^1, v_k^2 \in L_2(0,1) \right\}$$

is the tangent space of $T_u\mathcal{M}_r$ of $\mathcal{M}_r$ at u. This space is closed in $L_2(\Omega)$. Thus, in contrast to a standard weak formulation of (1) we seek a curve $t \mapsto u(t)$ on $\mathcal{M}_r$ which for almost every $t \in (0,T)$ satisfies the weak parabolic formulation (2) on the tangent space only.

In the following we write $\mathcal{M}$ instead of $\mathcal{M}_r$. To prove existence and uniqueness of solutions to (2) we use a Rothe-type temporal discretization in Hilbert space. As a generalization of the implict Euler method, which is a classical approach for parabolic PDEs, we consider a time stepping scheme via optimization problems at time steps $t_i = ih$, $h = T/N$, as follows:

$$u_{i+1} = \operatorname*{arg\,min}_{u \in \overline{\mathcal{M}}^w \cap V} F(u) = \frac{1}{2(t_{i+1}-t_i)} \|u - u_i\|_H^2 + \frac{1}{2} a(u,u;t_{i+1}) - \langle f_{i+1}, u \rangle.$$

Here $\overline{\mathcal{M}}^w$ denotes the weak closure of $\mathcal{M}$ in H and f_{i+1} is the average of f in the interval $[t_i, t_{i+1}]$, assuming $f(t) \in H$. By standard results, this optimization problem admits at least one solution, essentially since the cost function is convex, continuous and coercive on V. Moreover, the necessary optimality condition formally matches the weak formulation (2) at the time point t_{i+1}. In this way, one obtains approximate solutions $u_1, \dots, u_N \in \overline{\mathcal{M}}^w \cap V$. Let $\hat{u}_h(t)$ and $\hat{v}_h(t)$ be the piecewise linear and piecewise constant interpolants on the interval $[0,T]$, respectively. Our main result is that these interpolants converge for $h \to 0$ in a suitable sense to solutions of the weak DLRA formulation (2). First this is shown for a small time interval.

Theorem 1. *Assume $f \in L_2(0,T;H)$ and $u_0 \in \mathcal{M} \cap V$. Let $\sigma_r > 0$ be the smallest positive singular value of u_0 (the H-distance to $\mathcal{M}_{r-1}$). Then the following hold.*

(a) *The functions $\hat{u}_h$ and $\hat{v}_h$ converge, up to subsequences, weakly in $L_2(0,T;V)$ and strongly in $L_2(0,T;H)$, to the same function $\hat{u} \in L_\infty(0,T;V)$ with $\hat{u}(0) = u_0$, while the weak derivatives $\hat{u}_h'$ converge weakly to $\hat{u}'$ in $L_2(0,T;H)$, again up to subsequences. It holds $\hat{u}(t) \in \overline{\mathcal{M}}^w \cap V$ for almost all $t \in [0,T]$.*

(b) *There exists a constant $c > 0$ independent of σ_r such that $\hat{u}$ solves (2) for almost all $t < (\sigma_r/c)^2$, and $\hat{u}(t) \in \mathcal{M}$ for all $t < (\sigma_r/c)^2$.*

Note that the assumptions $u_0 \in V$ and $f \in L_2(0,T;H)$ are stronger than usual for existence of weak solutions to parabolic PDEs, but still weaker than needed for strong solutions. We then proceed by showing existence of a solution to (2) on a maximal time interval. It can also be shown to be the unique solution.

Theorem 2. *Let the assumptions in Theorem 1 hold. There exist $T^* \in (0,T]$ and $u \in W_2^1(0,T^*;V,H) \cap L_\infty(0,T^*;V)$ such that u solves problem (2) on the time interval $[0,T^*]$, and its continuous representative $u \in C(0,T^*;H)$ satisfies*

$u(t) \in \mathcal{M}$ for all $t \in [0, T^*)$. Here T^* is maximal for the evolution on $\mathcal{M}$ in the sense that if $T^* < T$, then

$$\liminf_{t \to T^*} \sigma_r(u(t)) = 0,$$

where $\sigma_r(u(t))$ is the smallest positive singular value of $u(t)$. In either case, u is the unique solution of (2) in $W_2^1(0, T^*; V, H)$.

The results can be stated and proved in a more general framework of parabolic evolution equations on conic 'manifolds' $\mathcal{M} \subset H$ where $V \subseteq H \subseteq V^*$ is a Gelfand triplet, and under assumptions that reflect the related properties of the model problem (1). A crucial property is that the associated operator $A(t)$ in (1) can be split into a diagonal part that maps any $u \in \mathcal{M}_r \cap V$ to the tangent space $T_u\mathcal{M}_r \cap V$ (in a suitable sense), and an off-diagonal part (corresponding to mixed derivatives) which can be shown to be locally bounded from $\mathcal{M} \cap V$ to H. This is based on the mixed regularity of functions in $\mathcal{M}_r \cap V$. Another important tool in the proof are curvature bounds in the form of Lipschitz constants for the H-orthogonal tangent space projectors at different points in $\mathcal{M}$, as they are well known for the rank-r manifold $\mathcal{M}_r$. We expect that the framework developed in Bachmayr et al. (2020) for proving existence and uniqueness of solutions is also applicable to certain types of dynamical low-rank tensor approximation for higher-dimensional parabolic problems.

References

M. Bachmayr, H. Eisenmann, E. Kieri, and A. Uschmajew. Existence of dynamical low-rank approximations to parabolic problems *arXiv preprint arXiv:2002.12197*, 2020.

L. Einkemmer and C. Lubich. A low-rank projector-splitting integrator for the Vlasov-Poisson equation. *SIAM Journal on Scientific Computing*, **40**(5):B1330–B1360, 2018.

O. Koch and C. Lubich. Dynamical low-rank approximation. *SIAM Journal on Matrix Analysis and Applications*, **29**(2):434–454, 2007.

C. Lubich and I.V. Oseledets. A projector-splitting integrator for dynamical low-rank approximation. *BIT Numerical Mathematics*, **54**(1):171–188, 2014.

H. Mena, A. Ostermann, L.-M. Pfurtscheller, and C. Piazzola. Numerical low-rank approximation of matrix differential equations. *Journal of Computational and Applied Mathematics*, **340**:602–614, 2018.

E. Musharbash, F. Nobile, and T. Zhou. Error analysis of the dynamically orthogonal approximation of time dependent random PDEs. *SIAM Journal on Scientific Computing*, **37**(2):A776–A810, 2015.

A. Ostermann, C. Piazzola, and H. Walach. Convergence of a low-rank Lie-Trotter splitting for stiff matrix differential equations. *SIAM Journal on Numerical Analysis*, **57**(4):1947–1966, 2019.

T. P. Sapsis and P.F.J. Lermusiaux. Dynamically orthogonal field equations for continuous stochastic dynamical systems. *Physica D: Nonlinear Phenomena*, **238**(23-24):2347–2360, 2009.

Conjugate Gradient Methods on Riemannian Manifolds

Hiroyuki Sato

Let M be a manifold and $f\colon M \to \mathbb{R}$ be a function defined on M. We consider the following unconstrained optimization problem on manifold M:

$$\min_{x \in M} f(x).$$

In this abstract, we review Euclidean and Riemannian conjugate gradient (CG) methods, introduce several studies on Riemannian CG (R-CG) methods, and propose a general framework for R-CG methods.

The first CG method is the linear CG method (Hestenes and Stiefel, 1952), which was originally proposed for minimizing $f(x) = \frac{1}{2}x^T Ax - b^T x$ on $M = \mathbb{R}^n$ with a symmetric positive definite matrix $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$. The minimization of $f(x)$ is equivalent to solving $\nabla f(x) = Ax - b = 0$, i.e., the linear equation $Ax = b$ for x. For a given initial point $x_0 \in \mathbb{R}^n$ and search direction $\eta_0 := -\nabla f(x_0)$, the linear CG method iterates $x_{k+1} := x_k + t_k \eta_k$ and $\eta_{k+1} := -\nabla f(x_{k+1}) + \beta_{k+1}\eta_k$ for $k \geq 0$, where $t_k := \arg\min_{t \geq 0} f(x_k + t d_k) = -\frac{\nabla f(x_k)^T \eta_k}{\eta_k^T A \eta_k}$ (exact line search) and $\beta_k := \frac{\|\nabla f(x_k)\|_2^2}{\|\nabla f(x_{k-1})\|_2^2}$, which ensures $\eta_k^T A \eta_l = \delta_{kl}$, i.e., $\eta_0, \eta_1, \ldots, \eta_{n-1}$ are mutually A-conjugate. We define $g_k := \nabla f(x_k)$ and $y_k := g_k - g_{k-1}$, and β_k can be written in several mathematically equivalent forms:

$$\beta_k = \frac{g_k^T g_k}{g_{k-1}^T g_{k-1}} = \frac{g_k^T g_k}{\eta_{k-1}^T y_k} = \frac{g_k^T g_k}{-g_{k-1}^T \eta_{k-1}} = \frac{g_k^T y_k}{g_{k-1}^T g_{k-1}} = \frac{g_k^T y_k}{\eta_{k-1}^T y_k} = \frac{g_k^T y_k}{-g_{k-1}^T \eta_{k-1}}.$$

The concept of iterating $x_{k+1} := x_k + t_k \eta_k$ with $\eta_k := -g_k + \beta_k \eta_{k-1}$ can be used even when a more general smooth objective function f on $M = \mathbb{R}^n$ is considered, which leads to the Euclidean nonlinear CG methods. Step length t_k can be approximately computed such that it satisfies some conditions such as the Armijo or Wolfe conditions. As for β_k in the Euclidean nonlinear CG methods, the above six types of formulas have been individually studied. For instance,

$$\beta_k^{\mathrm{FR}} = \frac{g_k^T g_k}{g_{k-1}^T g_{k-1}},\ \beta_k^{\mathrm{DY}} = \frac{g_k^T g_k}{d_{k-1}^T y_k},\ \beta_k^{\mathrm{PRP}} = \frac{g_k^T y_k}{g_{k-1}^T g_{k-1}},$$
$$\beta_k^{\mathrm{HS}} = \frac{g_k^T y_k}{d_{k-1}^T y_k},\ \beta_k^{\mathrm{LS}} = \frac{g_k^T y_k}{-g_{k-1}^T d_{k-1}}$$

were proposed by Fletcher and Reeves; Dai and Yuan; Polak, Ribière, and Polyak; Hestenes and Stiefel; and Liu and Storey, respectively. The review in Hager and Zhang (2006) may be referred to for details.

Next, we proceed to the manifold case. In what follows, we assume that M is a Riemannian manifold with Riemannian metric $\langle \cdot, \cdot \rangle$ and that f is smooth and bounded below. In particular, for a given retraction R on M, we assume that there exists a constant $L > 0$ such that $|\mathrm{D}(f \circ R_x)(t\eta)[\eta] - \mathrm{D}(f \circ R_x)(0)[\eta]| \leq Lt$ for any $t \geq 0$, $\eta \in T_x M$ with $\|\eta\|_x = 1$, and $x \in M$. We denote the Riemannian

gradient of f on M by $\operatorname{grad} f(x)$ and again use the notation $g_k := \operatorname{grad} f(x_k)$ in R-CG methods.

In line-search-based methods used in optimization problems on manifolds, search direction η_k is chosen as a tangent vector to M at x_k, and the update formula in $\mathbb{R}^n$, i.e., $x_{k+1} := x_k + t_k\eta_k$, is generalized to $x_{k+1} := R_{x_k}(t_k\eta_k)$, where $R\colon TM \to M$ is a retraction on M (Absil et al., 2008). When computing search direction $\eta_{k+1} \in T_{x_{k+1}}M$ in R-CG methods, $-g_{k+1} \in T_{x_{k+1}}M$ and $\beta_{k+1}\eta_k \in T_{x_k}M$ belong to distinct tangent spaces and, thus, cannot be added. Therefore, careful consideration is required for computing η_{k+1}.

Smith (1994) discussed R-CG methods in which search direction η_k is computed as $\eta_{k+1} := -g_{k+1} + \beta_{k+1}P(\eta_k)$, where P is the parallel translation with respect to the geodesic from x_k to x_{k+1}. He proposed $\beta_{k+1} := \frac{\langle g_{k+1} - P(g_k),\, g_{k+1}\rangle_{x_{k+1}}}{-\langle g_k, \eta_k\rangle_{x_k}}$, which is a generalization of the Euclidean version of $\beta^{\mathrm{LS}}_{k+1}$. Moreover, Edelman et al. (1998) discussed the same framework for computing η_{k+1} with several β_{k+1} including $\beta_{k+1} := \frac{\langle g_{k+1} - P(g_k),\, g_{k+1}\rangle_{x_{k+1}}}{\langle g_k, g_k\rangle_{x_k}}$, which is a generalization of the Euclidean $\beta^{\mathrm{PRP}}_{k+1}$.

However, in some cases, the parallel translation is numerically impractical. To resolve this issue, Absil et al. (2008) proposed the notion of vector transport $\mathcal{T}\colon TM \oplus TM \to TM$ on M, which is a generalization of the parallel translation. Using $\mathcal{T}$, they proposed a framework for R-CG methods, wherein η_{k+1} is computed as $\eta_{k+1} := -\operatorname{grad} f(x_{k+1}) + \beta_{k+1}\mathcal{T}_{t_k\eta_k}(\eta_k)$. In this framework, Ring and Wirth Ring and Wirth (2012) discussed $\beta^{\text{R-FR}}_{k+1} := \frac{\|\operatorname{grad} f(x_{k+1})\|^2_{x_{k+1}}}{\|\operatorname{grad} f(x_k)\|^2_{x_k}}$, which is a generalization of the Euclidean $\beta^{\mathrm{FR}}_{k+1}$, and provided a global convergence analysis for the R-CG method with $\beta^{\text{R-FR}}_{k+1}$ under the condition

$$\|\mathcal{T}_{t_k\eta_k}(\eta_k)\|_{x_{k+1}} \le \|\eta_k\|_{x_k}. \tag{1}$$

Unfortunately, this condition is sometimes violated (see Sato and Iwai (2015) for some examples). To avoid assuming (1), Sato and Iwai (2015) defined the scaled vector transport $\mathcal{T}^0$ associated with the differentiated retraction $\mathcal{T}^R$ as $\mathcal{T}^0_\eta(\xi) := \frac{\|\xi\|_x}{\|\mathcal{T}^R_\eta(\xi)\|_{R_x(\eta)}}\mathcal{T}^R_\eta(\xi)$ for $\xi, \eta \in T_xM$, where $\mathcal{T}^R_\eta(\xi) := \mathrm{D}R_x(\eta)[\xi]$. They proposed the formula $\eta_{k+1} := -\operatorname{grad} f(x_{k+1}) + \beta_{k+1}\mathcal{T}^{(k)}_{t_k\eta_k}(\eta_k)$, where $\mathcal{T}^{(k)}$ is identical to $\mathcal{T}^R$ when (1) is satisfied and is otherwise identical to $\mathcal{T}^0$. They proved that the modified Fletcher–Reeves-type R-CG method has a global convergence property without the assumption (1). Using this scaling technique, Sato (2016) also generalized the Euclidean version of $\beta^{\mathrm{DY}}_{k+1}$ as $\beta^{\text{R-DY}}_{k+1} := \frac{\|g_{k+1}\|^2_{x_{k+1}}}{\langle g_{k+1},\, \mathcal{T}^{(k)}_{t_k\eta_k}(\eta_k)\rangle_{x_{k+1}} - \langle g_k, \eta_k\rangle_{x_k}}$ and analyzed the R-CG method with $\beta^{\text{R-DY}}_{k+1}$. By generalizing the work of Dai and Yuan (2001), Sakai and Iiduka (2020) proposed the use of β_k that satisfies $-\sigma \le r_k \le 1$, and presented a global convergence analysis, where $r_k := \beta_k/\beta^{\text{R-DY}}_k$. Here, $\sigma := (1 - c_2)/(1 + c_2)$ with c_2 in the strong Wolfe condition $|\phi'_k(t_k)| \le c_2|\phi'_k(0)|$ for $\phi_k(t) := f(R_{x_k}(t\eta_k))$. This method comprises $\beta_k :=$

$\max\{0, \min\{\beta_k^{\text{R-DY}}, \beta_k^{\text{R-HS}}\}\}$ and $\beta_k := \max\{-\sigma\beta_k^{\text{R-DY}}, \min\{\beta_k^{\text{R-DY}}, \beta_k^{\text{R-HS}}\}\}$ as examples, where $\beta_k^{\text{R-HS}} := \frac{\langle g_k, g_k - \mathcal{T}^{(k-1)}_{t_{k-1}\eta_{k-1}}(g_{k-1})\rangle}{\langle g_k, \mathcal{T}^{(k-1)}_{t_{k-1}\eta_{k-1}}(\eta_{k-1})\rangle_{x_k} - \langle g_{k-1}, \eta_{k-1}\rangle_{x_{k-1}}}$ is a generalization of β_k^{HS}.

Recently, Zhu and Sato (2020) proposed R-CG methods with an inverse retraction. It is assumed that two retractions R^{fw} and R^{bw} on M are given. Herein, R^{fw} plays the same role as R in the above discussion, and we allow the case $R^{\text{fw}} = R^{\text{bw}}$. We compute $x_{k+1} := R^{\text{fw}}_{x_k}(t_k\eta_k)$ and $\eta_{k+1} := -\operatorname{grad} f(x_{k+1}) - \beta_{k+1} s_k t_k^{-1}(R^{\text{bw}}_{x_{k+1}})^{-1}(x_k)$, where $s_k := \min\left\{1, \|\eta_k\|_{x_k} \big/ \|t_k^{-1}\left(R^{\text{bw}}_{x_{k+1}}\right)^{-1}(x_k)\|_{x_{k+1}}\right\} > 0$ ensures global convergence by scaling the norm of the obtained tangent vector, which is obtained by applying the same idea as that used for the scaled vector transport in Sato and Iwai (2015). An important feature of this framework for R-CG methods is that there is no requirement for a vector transport.

We can summarize the above discussion as follows. A natural method of moving $\eta_k \in T_{x_k}M$ to $T_{x_{k+1}}M$ is the use of the parallel translation. A more general approach is the use of a vector transport instead. A scaled vector transport may also be considered as such a method and makes the assumption (1) unnecessary. However, it does not have linearity and is thus not a vector transport. Furthermore, an inverse retraction can be exploited instead of a vector transport to compute the search directions in R-CG methods. Hence, we can say that a more general mapping than a vector transport can be used in R-CG methods and propose a new framework with a formula for computing η_{k+1} as

$$\eta_{k+1} := -\operatorname{grad} f(x_{k+1}) + \beta_{k+1}\mathcal{T}^{(k)}(\eta_k),$$

where $\mathcal{T}^{(k)}$ is a map (not necessarily a vector transport) from $T_{x_k}M$ to $T_{x_{k+1}}M$ that satisfies $\|\mathcal{T}^{(k)}(\eta_k)\|_{x_{k+1}} \leq \|\eta_k\|_{x_k}$. Further details will be addressed in future research.

References

P. -A. Absil, R. Mahony, and R. Sepulchre. Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton, 2008.

Y.-H. Dai and Y. Yuan. An efficient hybrid conjugate gradient method for unconstrained optimization. *Annals of Operations Research*, **103**(1–4), 33–47, 2001.

A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, **20**(2), 303–353, 1998.

R. Fletcher and C. M. Reeves. Function minimization by conjugate gradients. *The Computer Journal*, **7**(2), 149–154, 1964.

W. W. Hager and H. Zhang. A survey of nonlinear conjugate gradient methods. *Pacific Journal of Optimization*, **2**(1), 35–58, 2006.

M. R. Hestenes and E. Stiefel. Methods of conjugate gradient for solving linear systems. *Journal of Research of the National Bureau of Standards*, **49**(6), 409–436, 1952.

W. Ring and B. Wirth. Optimization methods on Riemannian manifolds and their application to shape space. *SIAM Journal on Optimization*, **22**(2), 596–627, 2012.

H. Sakai and H. Iiduka. Hybrid Riemannian conjugate gradient methods with global convergence properties. *Computational Optimization and Applications*, **77**(3), 811–830, 2020.

H. Sato. A Dai–Yuan-type Riemannian conjugate gradient method with the weak Wolfe conditions. *Computational Optimization and Applications*, **64**(1), 101–118, 2016.

H. Sato and T. Iwai. A new, globally convergent Riemannian conjugate gradient method. *Optimization*, **64**(4), 1011–1031, 2015.

S. T. Smith. Optimization techniques on Riemannian manifolds. *Hamiltonian and Gradient Flows, Algorithms and Control*, 113–135, American Mathematical Society, 1994.

X. Zhu and H. Sato. Riemannian conjugate gradient methods with inverse retraction. *Computational Optimization and Applications*, **77**(3), 779–810, 2020.

Geomstats: A Python Package for Riemannian Geometry in Machine Learning

Nina Miolane

(joint work with Nicolas Guigui, Alice Le Brigant, Johan Mathe, Benjamin Hou, Yann Thanwerdas, Stefan Heyder, Olivier Peltre, Niklas Koep, Hadi Zaatiti, Hatem Hajri, Yann Cabanes, Thomas Gerald, Paul Chauchat, Christian Shewmake, Daniel Brooks, Bernhard Kainz, Claire Donnat, Susan Holmes, Xavier Pennec)

Introduction. We introduce Geomstats, an open-source Python package for computations and statistics on nonlinear manifolds (Miolane et al., 2020a,b). Data on manifolds naturally arise in different fields (see Figure 1). Geomstats provides computational methods that take into account the geometry of the data space. The source code is freely available under the MIT license at geomstats.ai or on GitHub at github.com/geomstats/geomstats.

Datasets. Figure 1 shows data on manifolds available in Geomstats and visualized with the module `visualization`. Cities on the Earth are points on the sphere; social networks can be represented as elements of the hyperbolic space; brain connectomes are traditionally modelled as elements of the manifold of symmetric positive definite matrices; poses of 3D objects are elements of the Lie groups of rotations SO(3) or of rigid transformations SE(3); shapes belong to the shape space; probability distributions are elements of Riemannian manifolds in information geometry; etc.

Objectives. The objectives of the package Geomstats are three-fold:

- Teach "hands-on" Geometric Statistics — by providing coding exercises and visualizations for courses on differential geometry and statistics on manifolds (see `notebooks` repository),
- Democratize the use of Geometric Statistics — by providing an API similar to `scikit-learn`'s to increase the use of geometric methods by machine learners,
- Support research in Geometric Statistics — by providing a platform where researchers share code associated to published works following common standards, to improve reproducibility and re-usability of results.

Overview. The package Geomstats offers object-oriented and extensively unit-tested implementations. Manifolds come equipped with families of Riemannian metrics with associated exponential and logarithmic maps, geodesics, and parallel

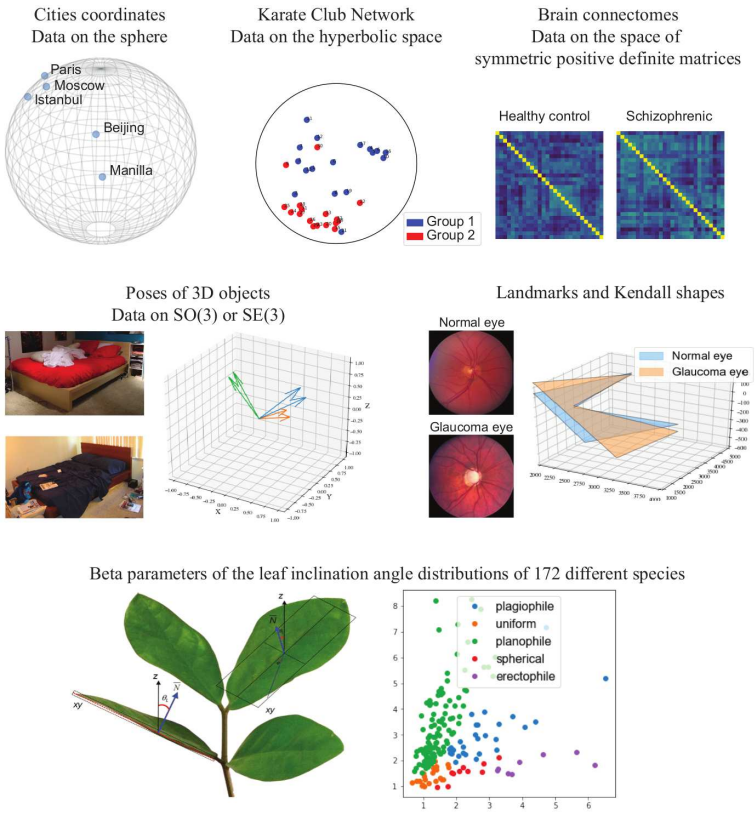

FIGURE 1. Visualizations of datasets of manifolds using GEOMSTATS.

transport. Statistics and learning algorithms provide methods for estimation, clustering, and dimension reduction on manifolds. All associated operations are vectorized for batch computation and support different execution backends—namely NumPy, PyTorch, and TensorFlow.

Usage. Three steps are needed to run learning algorithms on manifolds with GEOMSTATS: (i) instantiate the manifold of interest, (ii) instantiate the learning algorithm of interest, and (iii) run the algorithm. The following code snippet illustrates the use of online K-means on the hypersphere.

```
sphere = Hypersphere(dim=5)
data = sphere.random_uniform(n_samples=10)
clustering = OnlineKMeans(metric=sphere.metric,
                          n_clusters=4)
clustering = clustering.fit(data)
```

All geometric computations are performed behind the scenes. The user only needs a high-level understanding of Riemannian geometry. Each algorithm can be used with any of the manifolds and metrics implemented in the package. The folders `examples` and `notebooks` provide more code snippets that help users get started with Geomstats.

Comparison with related Python packages. Tables 1-2 summarize the comparison between existing Python packages related to data on manifolds. The package TheanoGeometry (Kühnel and Sommer, 2017) is the most closely related to Geomstats and provides nonlinear statistics and stochastic equations on Riemannian manifolds. The differential geometric tensors are computed with automatic differentiation, which inspired several submodules available in Geomstats. However, this library does not provide statistical learning algorithms and lacks engineering maintenance.

Several other packages focus on optimization on Riemannian manifolds. Pymanopt (Townsend et al., 2016) computes gradients and Hessian-vector products on Riemannian manifolds with automatic differentiation and provides the following solvers: steepest descent, conjugate gradient, the Nelder-Mead algorithm, particle swarm optimization, and the Riemannian trust regions. Geoopt (Kochurov et al., 2019) focuses on stochastic adaptive optimization on Riemannian manifolds, for machine learning problems. The library provides stochastic solvers, stochastic gradient descent and Adam, as well as the following samplers: Stochastic Gradient Langevin Dynamics, Hamiltonian Monte-Carlo, Stochastic Gradient Hamiltonian Monte-Carlo. Lastly, McTorch (Meghwanshi et al., 2018) provides optimization on Riemannian manifold for deep learning by adding a "Manifold" parameter to PyTorch's network layers and optimizers. The library provides the following solvers: stochastic gradient descent, AdaGrad and conjugate gradients.

As these libraries focus on optimization, they substitute potentially computationally expensive operations by practical proxies, for example, by replacing exponential maps by so-called retractions. However, they are less modular than Geomstats in terms of the Riemannian geometry and do not provide statistical learning algorithms. The optimization libraries are complementary to Geomstats and interact easily with it: an example integrating Pymanopt and Geomstats can be found in Geomstats' `examples` folder.

Conclusion. We presented the Python package Geomstats that provides the wider machine learning community with off-the-shelf geometric learning algorithms. The package offers a wide variety of manifolds, together with a flexibility in the choice of metrics, while being faithful to the mathematician's formulation of Riemannian geometry. This sometimes comes at cost of efficiency, and future contributions will be devoted to addressing this caveat.

	Manifolds	**Geometry**
Pymanopt	Euclidean manifold, symmetric matrices, sphere, complex circle, SO(n), Stiefel, Grassmannian, oblique manifold, SPD(n), elliptope, fixed-rank positive semidefinite matrices	Exponential and logarithmic maps, retraction, vector transport, `egrad2rgrad`, `ehess2rhess`, inner product, distance, norm
Geoopt	Euclidean manifold, sphere, Stiefel, Poincaré ball	Same as Pymanopt
McTorch	Stiefel, SPD(n)	Same as Pymanopt
TheanoGeometry	Sphere, ellipsoid, SPD(n), Landmarks, GL(n), SO(n), SE(n)	Inner product, exponential and logarithmic maps, parallel transport, Christoffel symbols, Riemann, Ricci and scalar curvature, geodesics, Fréchet mean
Geomstats	Euclidean, Minkowski, hyperbolic space, Poincaré polydisk, hypersphere, SO(n), SE(n), GL(n), Stiefel, Grassmannian, SPD(n), symmetric matrices, skew-symmetric matrices, discretized curves on manifolds, landmarks on manifolds	Levi-Civita connection, Christoffel symbols, parallel transport, exponential and logarithmic maps, inner product, distance, norm, geodesics, group invariant metrics, Fréchet means and learning algorithms on manifolds

TABLE 1. Comparison of libraries in terms of geometric operations (as of 2020)

	Backends	**Continuous integration and coverage**
Pymanopt	Autograd, PyTorch, TensorFlow, Theano	CI, coverage 85%
Geoopt	PyTorch	75%
McTorch	PyTorch	CI, coverage 84%
TheanoGeometry	Theano	No CI, no unit tests
Geomstats	NumPy, PyTorch, TensorFlow	CI, coverage 92% (NumPy), 76% (TensorFlow), 79% (PyTorch)

TABLE 2. Comparison of code infrastructure (as of 2020)

Acknowledgements. This work is partially supported by the National Science Foundation, grant NSF DMS RTG 1501767, the Inria-Stanford associated team GeomStats, the European Research Council (ERC) under the EU Horizon 2020 research and innovation program (grant agreement G-Statistics No. 786854) and by the French Government through the 3IA Côte d'Azur Investments in the Future project (National Research Agency ANR-19-P3IA-0002).

References

A. Barachant. PyRiemann: Python package for covariance matrices manipulation and Biosignal classification with application in Brain Computer interface, 2015.

G. Bécigneul and O. -E. Ganea. Riemannian Adaptive Optimization Methods. *Proceedings of the International Conference on Learning Representations (ICLR) 2019*, 1–16, 2018.

A. Censi. PyGeometry: Library for handling various differentiable manifolds,2012.

M. Kochurov, R. Karimov, and S. Kozlukov. Geoopt: Riemannian Adaptive Optimization Methods with pytorch optim, 2019.

L. Kühnel and S. Sommer. Computational Anatomy in Theano. *Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics*, 164–176, Cham, 2017. Springer International Publishing.

M. Meghwanshi, P. Jawanpuria, A. Kunchukuttan, H. Kasai, and B. Mishra. McTorch, a manifold optimization library for deep learning, 2018.

N. Miolane, A. L. Brigant, J. Mathe, B. Hou, N. Guigui, Y. Thanwerdas, S. Heyder, O. Peltre, N. Koep, H. Zaatiti, H. Hajri, Y. Cabanes, T. Gerald, P. Chauchat, C. Shewmake, B. Kainz, C. Donnat, S. Holmes, and X. Pennec. Geomstats: A python package for Riemannian geometry in machine learning, 2020.

N. Miolane, A. L. Brigant, J. Mathe, B. Hou, N. Guigui, Y. Thanwerdas, S. Heyder, O. Peltre, N. Koep, H. Zaatiti, H. Hajri, Y. Cabanes, T. Gerald, P. Chauchat, C. Shewmake, B. Kainz, C. Donnat, S. Holmes, and X. Pennec. Introduction to Geometric Learning in Python with Geomstats. *Proceedings of the 19th Python in Science Conference*, Meghann Agarwal, Chris Calloway, Dillon Niederhut, and David Shupe, editors, 48 – 57, 2020.

J. Townsend, N. Koep, and S. Weichwald. Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. *Journal of Machine Learning Research*, **17**(137),1–5, 2016.

K. Wynn. PyQuaternions: A fully featured, pythonic library for representing and using quaternions, 2014.

The Riemannian Chambolle–Pock Algorithm and Optimization on Manifolds in Julia

Ronny Bergmann

(joint work with Seth Axen, Mateusz Baran, Roland Herzog, Maurício Silva Louzeiro, Daniel Tenbrinck, José Vidal-Núñez)

In the recent years many algorithms for (Euclidean) nonsmooth optimization have been generalized to Riemannian manifolds. Among these are the cyclic proximal point algorithm (CPPA) by Bačák (2014) and the (parallel) Douglas-Rachford algorithm (PDRA) in Bergmann et al. (2016). The latter is known to be equivalent on Euclidean space to the Chambolle-Pock algorithm (CPA) that was introduced by Chambolle and Pock (2011).

In order to introduce a Riemannian Chambolle-Pock algorithm, this talk is first concerned with defining a suitable generalization of the Fenchel dual, see Bergmann et al. (2020). The main ingredient is a base point $m \in \mathcal{M}$ to introduce the m-Fenchel dual F_m^*. Most properties of the Fenchel conjugate can be generalized, especially the Fenchel-Young inequality as well as the Fenchel-Moreau identity.

These are then used to derive the Riemannian Chambolle-Pock algorithm in an exact and a linearized variant, and a convergence proof finishes the first part of the talk.

The second part of the talk is about the implementation of the RCPA in Manopt.jl (Bergmann, 2020), a Julia package for optimization on manifolds. It uses the ManifoldsBase.jl[1] interface, such that all algorithms implemented in this package can be used in combination with Manifolds. jl (Axen et al., 2020).

A manifold within the interface is a type inheriting from Manifold{$\mathbb{F}$}, where $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\}$ denotes the field the manifold is build upon. Then one implements for example exp(M,p,X) and log(M,p,q), while for the geodesic a default function is then automatically available. Furthermore, a sophisticated decorator pattern is available, to add additionally properties to a manifold, like a Lie structure or distinguish different metrics using Julia's dispatch mechanism. Functions unrelated to the Riemannian metric are transparently passed to the original manifold, such that common properties have to only be implemented once.

the solver framework from Manopt.jl is based on describing a Problem to solve and Options to describe and setup the solver. Then, several function, gradients, proximal maps are available as well as debug and recording features for arbitrary fields of aforementioned types. Within Manopt.jl both variants of RCPA are available.

Numerical experiments illustrate that the newly introduced RCPA performs as well as the PDRA in number of iterations but outperforms the latter in runtime.

References

S. Axen, M. Baran, and R. Bergmann. Manifolds.jl, 2020.

M. Bačák. Computing medians and means in hadamard spaces. *SIAM Journal on Optimization*, **24**(3):1542–1566.

R. Bergmann. Manopt.jl, 2020.

R. Bergmann, R. Herzog, M. Silva Louzeiro, D. Tenbrinck, and J. Vidal -Núñez. Fenchel duality theory and a primal-dual algorithm on Riemannian manifolds. *Accepted for publication in Foundations of Computational Mathematics*.

R. Bergmann, J. Persch, and G. Steidl. A parallel douglas rachford algorithm for minimizing rof-like functionals on images with values in symmetric hadamard manifolds. *SIAM Journal on Imaging Sciences*, **9**(4):901–937.

A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, **40**(1):120–145.

[1] see https://juliamanifolds.github.io/Manifolds.jl/stable/interface.html

An SQP Method for Equality Constrained Optimization on Hilbert Manifolds

Anton Schiela
(joint work with Julián Ortiz-Lopez)

1. Constrained Optimization Problems and their Pullbacks

In the following, we will consider optimization problems of the form:

$$\min_{x\in X} f(x) \;\; s.t. \;\; c(x) = y_*, \quad c : X \to Y, \; y_* \in Y \tag{1}$$

on Hilbert manifolds X and Y of class C^2, modelled over Hilbert spaces, as defined e.g. in Lang (2001). Let X be equipped with a Riemannian metric $\langle\cdot,\cdot\rangle_x$. Let $f : X \to \mathbb{R}$ be a C^2-functional and $c : X \to Y$ be a C^2 mapping. While unconstrained optimization on manifolds is a well established field, only few works are available on constrained optimization algorithms Liu and Boumal (2019); Obara et al. (2020); Schiela and Ortiz (2020) up to now.

Following the approach in Absil et al. (2008) we will use *retractions* $R_x : T_xX \to X$, which are widely used in optimization on manifolds. For the codomain Y of c we also need mappings $S_y : Y \to T_yY$ in the other direction, which we call *linearizing maps*, sometimes also called generalized logarithms (Boumal, 2010). On Riemannian manifolds, the exponential map $\exp_x : T_xX \to X$ and the logarithmic map $\log_y : Y \to T_yY$ are canonical examples, however, as discussed in Absil et al. (2008), more efficient retractions and linearizing maps can often be used.

Definition 1.1. A C^2-mapping $R : TX \to X$, where $R_x : T_xX \to X$, is called a retraction, if $R_x(0_x) = x$ and $T_{0_x}R_x = id_{T_xX}$. A C^2-mapping $S : Y \times Y \to TY$, where $S_y : Y \to T_yY$, is called a linearizing map, if $S_y(y) = 0_y$ and $T_yS_y = id_{T_yY}$.

We define pullbacks $\mathbf{f} : T_xX \to \mathbb{R}$ of f and $\mathbf{c} : T_xX \to T_{c(x)}Y$ of $c(x) = y_*$ via

$$\begin{aligned}\mathbf{f}(\delta x) &:= (f \circ R_x)(\delta x)\\ \mathbf{c}(\delta x) &:= S_{c(x)} \circ c \circ R_x(\delta x) - S_{c(x)}(y_*).\end{aligned}$$

In this way, we can now define the pullback of (1) to tangent spaces:

$$\min_{\delta x\in T_xX} \mathbf{f}(\delta x) \quad \text{s.t.} \quad \mathbf{c}(\delta x) = 0_{c(x)}, \qquad \mathbf{c} : T_xX \to T_{c(x)}Y. \tag{2}$$

Since $c(x_*) = y_*$ is equivalent to $S_{c(x_*)}(y_*) = 0_{c(x_*)}$ an element $x_* \in X$ is a local minimizer of (1) if and only if 0_{x_*} is a local minimizer of its pullback at x_*.

Thus, we have reduced (1) locally to a nonlinear optimization problem on Hilbert spaces, to which techniques of constrained optimization on linear spaces can be applied.

2. A Local SQP-Method

We will pursue the idea of SQP methods and derive a linearly constrained quadratic model of (2), which is of the following form:

$$\min_{\delta x \in T_x X} \mathbf{q}(\delta x) \quad \text{s.t.} \quad \mathbf{c}'(0_x)\delta x + \mathbf{c}(0_x) = 0_y, \qquad \mathbf{c} : T_x X \to T_y Y. \tag{3}$$

Here $\mathbf{q} : T_x X \to \mathbb{R}$ is a quadratic model for $\mathbf{f}$, which also will use second order information of the problem. For that we will need the local Lagrangian function:

$$\begin{aligned} \mathbf{L} : T_x X \times T_{c(x)} Y^* &\to \mathbb{R} \\ (\delta x, p) &\mapsto \mathbf{L}(\delta x, p) := \mathbf{f}(\delta x) + p \circ \mathbf{c}(\delta x). \end{aligned} \tag{4}$$

We will carry over the ideas of Absil et al. (2008) from unconstrained optimization on Riemannian manifolds to equality constrained optimization on Hilbert manifolds. In Absil et al. (2008) quadratic models of the objective f are computed independently of the retractions used by the optimization algorithm. First order models use $f'(x)$, which is invariant of retractions. Second order models are computed by the Riemannian hessian of f, i.e $\mathbf{f}''_\circ(0_x) := (f \circ \exp_x)''(0_x)$. This yields a second order quadratic model $\mathbf{q}_\circ$ for $\mathbf{f}_\circ$. If an algorithm is implemented via a retraction $R_x \neq \exp_x$ creating a pullback $\mathbf{f} \neq \mathbf{f}_\circ$, we see that $\mathbf{f}''(0_x) \neq \mathbf{f}''_\circ(0_x)$ in general and thus, $\mathbf{q}_\circ$ is not a second order model for $\mathbf{f}$.

From that perspective, steps are computed with the help of *two potentially different retractions*: a natural one $R^\circ_x = \exp_x$, to define a quadratic model $\mathbf{q}_\circ$ and an implemented one R_x to compute an update $R_x(\delta x)$.

In equality constrained optimization second order quadratic models employ, besides f', the second derivative of the Lagrangian function, which is $\mathbf{L}''_\circ(0_x, p_x)$, computed via R°_x, $S^\circ_{c(x)}$. Thus, for some given Lagrange multiplier $p_x \in T_{c(x)} Y^*$ our quadratic model reads:

$$\mathbf{q}_\circ(\delta x) := f(x) + f'(x)\delta x + \frac{1}{2}\mathbf{L}''_\circ(0_x, p_x)(\delta x, \delta x). \tag{5}$$

This leads to the following linearly constrained quadratic problem:

$$\min_{\delta x \in T_x X} \mathbf{q}_\circ(\delta x) \quad \text{s.t.} \quad c'(x)\delta x + \mathbf{c}(0_x) = 0. \tag{6}$$

If $c'(x)$ is surjective and $\mathbf{q}_\circ$ is elliptic on $\ker c'(x)$, a minimizer Δx of (6) exists, and we call it a *full SQP-step*. An SQP method creates a sequence of iterates by solving these quadratic problems. Observe that the computation of Δx is completely independent of the particular retraction R_x. Thus, Δx may enter into the construction of R_x.

3. Local Convergence Analysis

In Schiela and Ortiz (2020) we performed a local convergence analysis of Algorithm 1, which we will sketch here, in the framework of *affine covariant Newton methods* (Deuflhard, 2011). We will denote by x_* a local solution of (1) and impose the following assumptions:

Algorithm 1 Local SQP method

Require: initial iterate x
  repeat
    compute a Lagrange multiplier estimate $p_x \in T_{c(x)}Y^*$
    compute Δx by solving (6), using R_x°, $S_{c(x)}^\circ$, $S_{c(x)}$
    $x \leftarrow R_x(\Delta x)$
  until converged

Assumption 3.1. Let U_{x_*} be a neighborhood of x_*. Assume that there are constants $\rho_0 > 0, \omega_{\mathbf{f}'}, \omega_{\mathbf{c}}$, such that the following estimates hold for all $x \in U_{x_*}$, and all $v \in T_xX$, $\xi, \delta x \in B_{\rho_0}^x$:

$$\|\mathbf{f}'(\delta x) - f'(x)\|_{x,*} \le \omega_{\mathbf{f}'}\|\delta x\|_x, \tag{7}$$

$$\mathbf{c}'(\xi) \text{ surjective}, \quad \|\mathbf{c}'(\xi)^-(\mathbf{c}'(\delta x) - c'(x))v\|_x \le \omega_{\mathbf{c}}\|\delta x\|_x\|v\|_x \tag{8}$$

Here $\mathbf{c}'(\xi)^-$ denotes the minimal norm pseudo-inverse of $\mathbf{c}'(\xi)$ with respect to $\|\cdot\|_x$.

Further, assume that there are constants $\omega_{\mathbf{L}}$, $\alpha_{\mathbf{L}_\circ''} > 0, M_{\mathbf{L}_\circ''}$, M_ϕ such that for all $x \in U_{x_*}$ and all $\delta x \in B_{\rho_0}^x$:

$$|(\mathbf{L}''(\delta x, p_x) - \mathbf{L}''(0_x, p_x))(v, w)| \le \omega_{\mathbf{L}}\|\delta x\|_x\|v\|_x\|w\|_x \quad \forall v, w \in T_xX \tag{9}$$

$$\mathbf{L}_\circ''(0_x, p_x)(v, v) \ge \alpha_{\mathbf{L}_\circ''}\|v\|_x^2 \qquad \forall v \in \ker c'(x) \tag{10}$$

$$|\mathbf{L}_\circ''(0_x, p_x)(v, w)| \le M_{\mathbf{L}_\circ''}\|v\|_x\|w\|_x \qquad \forall v, w \in T_xX \tag{11}$$

$$\|(R_x^{-1} \circ R_x^\circ)''(0_x)(v, w)\|_x \le M_\Phi\|v\|_x\|w\|_x \qquad \forall v, w \in T_xX. \tag{12}$$

The use of the *affine covariant Lipschitz constant* $\omega_{\mathbf{c}}$ for $\mathbf{c}'$ in (8) follows the ideas of Deuflhard (2011). It avoids the use of norms on T_yY and can be estimated a-posteriori during the run of a globalized algorithm. The other assumptions are all standard for local convergence analysis. Due to (12) R_x need not be a second order retraction to obtain fast local convergence.

Theorem 3.2. Suppose that Assumption 3.1 holds at x_*. Then for initial values, sufficiently close to x_*, Algorithm 1 converges quadratically to x_*

More details, a globalization strategy, discussion of transition to local convergence, and numerical results can be found in Schiela and Ortiz (2020) and also in Ortiz (2020).

References

S. Lang. Fundamentals of Differential Geometry. Springer New York, 2001.

C. Liu and N. Boumal. Simple algorithms for optimization on Riemannian manifolds with constraints. *arXiv preprint arXiv:1901.10000*, 2019.

M. Obara, T. Okuno, and A. Takeda. Sequential quadratic optimization for nonlinear optimization problems on Riemannian manifolds. *arXiv preprint arXiv:2009.07153*, 2020.

P.-A. Absil, R. Mahony, and R. Sepulchre. Optimization algorithms on matrix manifolds. Princeton University Press, 2008.

N. Boumal. Discrete curve fitting on manifolds. Master's thesis, Université catholique de Louvain, jun 2010.

A. Schiela and J. Ortiz. An SQP method for equality constrained optimization on manifolds. *arXiv preprint arXiv:2005.06844*, 2020.

P. Deuflhard. Newton methods for nonlinear problems: affine invariance and adaptive algorithms. *Springer Science & Business Media*, volume **35**, 2011.

J. Ortiz. Constrained Optimization on Manifolds. PhD thesis, Bayreuth, December 2020.

A Quotient Geometry with Simple Geodesics on the Manifold of Fixed-Rank Positive-Semidefinite Matrices

Estelle Massart

(joint work with Pierre-Antoine Absil, Julien M. Hendrickx)

Positive-semidefinite (PSD) matrices are ubiquitous in nowadays' life, appearing, e.g., as variables in semidefinite programming, covariance matrices in statistics, diffusion tensors in brain imaging, and covariance descriptors in image set classification. In some cases (e.g., when the data points are low-rank representatives of large PSD matrices), the rank of the matrices can be assumed to be fixed, and the data belong to the set $\mathcal{S}_+(p,n)$ of PSD matrices of size n and rank p.

The set $\mathcal{S}_+(p,n)$ does obviously not have a vector space structure: the sum of two PSD matrices of rank p has usually a rank larger than p. However, this set can be turned into a Riemannian manifold. Different geometries were proposed for the manifold $\mathcal{S}_+(p,n)$, none of them having the desirable property of turning $\mathcal{S}_+(p,n)$ into a geodesically complete manifold with closed-form expressions for both the exponential and the logarithm maps Vandereycken et al. (2009, 2013); Bonnable and Sepulchre (2009).

In this talk, we follow the route of Journée et al. (2010) by identifying $\mathcal{S}_+(p,n)$ with the quotient manifold $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$, where $\mathbb{R}_*^{n\times p}$ is the set of full-rank $n\times p$ matrices and $\mathcal{O}_p$ is the orthogonal group of order p. The total space $\mathbb{R}_*^{n\times p}$ is equipped with the Euclidean metric.

There are two main reasons to consider this geometry. The first one is the fact that the computation cost of the resulting exponential and logarithm maps is low in comparison with other proposed geometries, so that this geometry is particularly suitable for numerical computations.

The second motivation is the fact that the associated distance coincides with the Wasserstein distance between degenerate centered Gaussian distributions. Indeed, any degenerate centered Gaussian distribution is parameterized by a positive-semidefinite covariance matrix. The Wasserstein metric between degenerate centered Gaussian distributions induces then a distance between positive-semidefinite matrices, that coincides with the distance on $\mathbb{R}_*^{n\times p}/\mathcal{O}_p$ computed here. The latter is also a direct generalization of the Bures–Wasserstein distance between positive-definite matrices, presented in, e.g., Takatsu (2011); Bhatia et al. (2018).

The main drawback of this geometry is that it does not turn the manifold into a complete metric space. This drawback is mitigated by two observations. First, this situation is not isolated, see, e.g., Absil and Oseledets (2014) that proposes

several retractions (first-order approximations of the exponential map) on the low-rank manifold, that are not defined everywhere. Secondly, several recent works take into account situations where the exponential map (or more generally the retraction) is not defined everywhere, see, e.g., Boumal et al. (2018).

We derive expressions for the Riemannian logarithm and the injectivity radius on this manifold, as well as tight bounds on its sectional curvature. These last two concepts play a key role in convergence guarantees of several optimization and consensus algorithms, and allow to ensure continuity of the results of some data fitting algorithms. This talk relies on the two papers Massart and Absil (2020); Massart et al. (2019).

References

P.-A. Absil, P.-Y. Gousenbourger, P. Striewski, and B. Wirth. Differentiable piecewise-Bézier surfaces on Riemannian manifolds. *SIAM Journal on Imaging Sciences*, **9**(4):1788–1828, 2016.

P.-A. Absil and I.V. Oseledets. Low-rank retractions: a survey and new results. *Computational Optimization and Applications*, **62**(1):5–29, 2014.

B. Afsari, R. Tron, and R. Vidal. On the convergence of gradient descent for finding the Riemannian center of mass. *SIAM Journal on Control and Optimization*, **51**(3):2230–2260, 2013.

R. Bhatia, T. Jain, and Y. Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 2018.

S. Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions Automatic Control*, **58**(9):2217–2229, 2013.

S. Bonnabel and R. Sepulchre. Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. *SIAM Journal on Matrix Analysis and Applications*, **31**(3):1055–1070, 2009.

N. Boumal, P.-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 2018.

P.-Y. Gousenbourger, E. Massart, and P.-A. Absil. Data fitting on manifolds with composite Bézier-like curves and blended cubic splines. *Journal of Mathematical Imaging and Vision*, 2018. Preprint: `https://sites.uclouvain.be/absil/2018.04`.

M. Journée, F. Bach, P.-A. Absil, and R. Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, **20**(5):2327–2351, 2010.

E. Massart and P.-A. Absil. Quotient geometry with simple geodesics for the manifold of fixed-rank positive-semidefinite matrices. *SIAM Journal on Matrix Analysis and Applications*, **41**(1):171–198, 2020.

E. Massart, J.M. Hendrickx, and P.-A. Absil. Curvature of the manifold of fixed-rank positive-semidefinite matrices endowed with the Bures-Wasserstein metric. *4th conference on Geometric Sciences of Information (GSI 2019)*, 2019.

A. Takatsu. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, *48*(4):1005–1026, 2011.

R. Tron, B. Afsari, and R. Vidal. Riemannian consensus for manifolds with bounded curvature. *IEEE Transactions on Automatic Control*, **58**(4):921–934, 2013.

B. Vandereycken, P.-A. Absil, and S. Vandewalle. Embedded geometry of the set of symmetric positive semidefinite matrices of fixed rank. *IEEE/SP 15th Workshop on Statistical Signal Processing*, 389–392, 2009.

B. Vandereycken, P.-A. Absil, and S. Vandewalle. A Riemannian geometry with complete geodesics for the set of positive semidefinite matrices of fixed rank. *IMA Journal of Numerical Analysis*, **33**(2):481–514, 2013.

Higher-Order Non-Smooth Optimization on Manifolds

Jan Lellmann
(joint work with Willem Diepeveen, Caterina Rust)

We consider manifold-valued optimization problems of the form

$$\inf_{p\in\mathcal{M}} \{f(p) + g(\Lambda(p))\}, \tag{1}$$

where $\mathcal{M}$ and $\mathcal{N}$ are smooth Riemannian manifolds, $\Lambda : \mathcal{M} \to \mathcal{N}$ is a general differentiable non-linear map, and $f : \mathcal{M} \to \bar{\mathbb{R}}$ and $g : \mathcal{N} \to \bar{\mathbb{R}}$ are non-smooth. In the context of mathematical image processing, such problems occur for example in variational approaches to diffusion tensor magnetic resonance imaging, interpolation on SO(3), and the analysis of interferometric synthetic aperture radar data.

Existing strategies for numerically finding a minimizer of (1) can be categorized into first-order and higher-order methods. First-order methods are based on subgradient- or proximal steps, have been found to be relatively robust in practice, but suffer from slow tail convergence Bergmann et al. (2016, 2020) or require highly specialized approaches Storath et al. (2016). In contrast, higher-order methods provide superlinear or even quadratic convergence Hintermüller (2010); Xiao et al. (2018), but so far have not been successfully applied to the non-smooth manifold-valued setting.

In order to achieve this goal, we rely on a dualization approach on manifolds developed in Bergmann et al. (2020): After fixing a base point $m \in \mathcal{M}$, we replace (1) by a linearized saddle-point problem,

$$\inf_{p\in\mathcal{M}} \sup_{\xi_n\in T_n^*} \{f(p) + \langle D_m\Lambda[\log_m p], \xi_n\rangle - g_n^*(\xi_n)\}. \tag{2}$$

Here $D_m\Lambda[v]$ denotes the differential of Λ at $m \in \mathcal{M}$ applied to the tangent vector $v \in T_p\mathcal{M}$ and we choose $n := \Lambda(m)$ as the base point of the generalized conjugate g_n^*.

In Bergmann et al. (2020), it was shown that in the Hadamard case, a saddle point can be found by solving the linearized primal-dual optimality system

$$p = \operatorname{prox}_{\sigma f}\left(\exp_p\left(\mathcal{P}_{m\to p}\left(-\sigma\left(D_m\Lambda\right)^*\left[\mathcal{P}_{n\to\Lambda(m)}\xi_n\right]\right)^\sharp\right)\right), \tag{3}$$

$$\xi_n = \operatorname{prox}_{\tau g_n^*}\left(\xi_n + \tau\left(\mathcal{P}_{\Lambda(m)\to n} D_m\Lambda\left[\log_m p\right]\right)^\flat\right), \tag{4}$$

where $\mathcal{P}_{m\to p}$ denotes parallel transport and $\sharp, \flat$ are the musical isomorphisms converting between tangent- and cotangent spaces, and $\operatorname{prox}_{\sigma f}(\cdot)$ denotes the generalized proximal map with respect to the function f and step size σ.

The recently proposed first-order lRCPA method Bergmann et al. (2020) can be understood as a fixed-point iteration on this system. In order to apply a Newton-type method to (3)–(4) instead, we rewrite the system as the problem of finding a

zero of the *vector field* $X : \mathcal{M} \times T_n^*\mathcal{M} \to T\mathcal{M} \times T_n^*\mathcal{M}$:

$$X(p,\xi_n) := \begin{pmatrix} -\log_p \operatorname{prox}_{\sigma f}\left(\exp_p\left(\mathcal{P}_{m\to p}\left(-\sigma\left(D_m\Lambda\right)^*\left[\mathcal{P}_{n\to\Lambda(m)}\xi_n\right]\right)^\sharp\right)\right) \\ \xi_n - \operatorname{prox}_{\tau g_n^*}\left(\xi_n + \tau\left(\mathcal{P}_{\Lambda(m)\to n} D_m\Lambda\left[\log_m p\right]\right)^\flat\right) \end{pmatrix} \tag{5}$$

The non-differentiability of X prohibits the use of a classical Riemannian Newton method Absil et al. (2009). Therefore we apply a superlinearly convergent Riemannian Semismooth Newton (RSSN) method de Oliveira et al. (2020), which relies on choosing an element from the Clarke generalized differential $\partial_{\mathcal{M},C}X(q^k)$ of X at the current iterate $q^k := (p^k, \xi_n^k)$, and solving a Newton system for the step.

In order to account for inexact steps and Quasi-Newton approaches, we also investigate an *inexact* version of RSSN:

Algorithm 1 Inexact Semismooth Newton

Initialization: $q^0 \in \mathcal{M} \times T_n^*\mathcal{M}, a^0 \geq 0$
for $k = 0, 1, \ldots$ **do**
  Choose $V_k(q^k) \in \partial_{\mathcal{M},C}X(q^k)$
  Solve $V_k(q^k)d^k = -X(q^k) + r^k$, where $\|r^k\|_{(q^k)} \leq a^k\|X(q^k)\|_{(q^k)}$
  $q^{k+1} := \exp_{q^k}(q^k)$
  Choose $a^{k+1} \geq 0$
end for

We can show the following theorem regarding superlinear convergence:

Theorem 1. (1) *There exist $a > 0$ and $\delta > 0$ such that for every $q^0 \in B_\delta(q^*)$ and $a^k \leq a$, the sequence $(q^k)_{k\geq 0}$ generated by Alg. 1 is well-defined, is contained in $B_\delta(q^*)$, and converges Q-linearly to the solution q^*.*

(2) *If the sequence $(q^k)_{k\geq 0}$ generated by Alg. 1 converges to the solution q^* and further $\|r^k\|_{(q^k)} \in o\left(\|X(q^k)\|_{(q^k)}\right)$, then the rate of convergence is Q-superlinear.*

(3) *If the sequence $(q^k)_{k\geq 0}$ generated by Alg. 1 converges to the solution q^*, X is μ-order semismooth at q^*, and $\|r^k\|_{(q^k)} \in O\left(\|X(q^k)\|_{(q^k)}^{1+\mu}\right)$, then the rate of convergence is Q-*order $1+\mu$.

Consequently, if the relative residual can be bounded, we obtain linear convergence. If the relative residual converges to zero, the convergence is superlinear, and even of higher order if the decrease is fast enough and the optimality system is higher-order semismooth.

An implementation based on the `manopt.jl` library exhibits superlinear local convergence as predicted. Interestingly, the method appears to work well not only on Hadamard manifolds such as the symmetric positive definite matrices $\mathcal{P}(3)$ as supported by the duality theory, but experimentally also shows rapid convergence on the positively-curved unit sphere $\mathcal{S}^2$.

References

R. Bergmann, J. Persch, and G. Steidl. A parallel Douglas–Rachford algorithm for minimizing rof-like functionals on images with values in symmetric Hadamard manifolds. *SIAM Journal on Imaging Sciences*, **9**(3):901–937, 2016.

R. Bergmann, R. Herzog, M. Silva Louzeiro, D. Tenbrinck, and J. Vidal -Núñez. Fenchel duality theory and a primal-dual algorithm on Riemannian manifolds. *Accepted for publication in Foundations of Computational Mathematics*.

M. Storath, A. Weinmann, and M. Unser. Exact algorithms for l^1-tv regularization of real-valued or circle-valued signals. *SIAM Journal on Scientific Computing*, **38**(1):A614–A630, 2016.

M. Hintermüller. Semismooth Newton methods and applications. *Department of Mathematics, Humboldt-University of Berlin*, 2010.

X. Xiao, Y. Li, Z. Wen, and L. Zhang. A regularized semi-smooth Newton method with projection steps for composite convex programs. *Journal of Scientific Computing*, **76**(1):364–389, 2018.

P.-A. Absil, R. Mahony, and R. Sepulchre. Optimization algorithms on matrix manifolds. Princeton University Press, 2009.

F.R. de Oliveira, O.P. Ferreira, et al. Newton method for finding a singularity of a special class of locally Lipschitz continuous vector fields on Riemannian manifolds. *Journal of Optimization Theory and Applications*, **185**(2):522–539, 2020.

Reporter: Estefanía Loayza-Romero

Participants

Prof. Dr. Pierre-Antoine Absil
UCLouvain
Avenue Georges Lemaître 4
1348 Louvain-la-Neuve
BELGIUM

Dr. Ronny Bergmann
Fakultät für Mathematik
TU Chemnitz
09107 Chemnitz
GERMANY

Dr. Nicolas Boumal
EPFL SB MATH
MA C2 627 (Bâtiment MA)
Station 8
1015 Lausanne
SWITZERLAND

Prof. Dr. Roland Herzog
Fakultät für Mathematik
Technische Universität Chemnitz
Reichenhainer Strasse 41
09126 Chemnitz
GERMANY

Prof. Dr. Jan Lellmann
University of Lübeck
Institute of Mathematics and Image Computing
Maria-Goeppert-Str. 3
23562 Lübeck
GERMANY

M.Sc. Estefania Loayza-Romero
Fakultät für Mathematik
Technische Universität Chemnitz
Reichenhainer Strasse 41
09126 Chemnitz
GERMANY

Estelle Massart
Andrew Wiles Building
Radcliffe Observatory Quarter
Woodstock Road
Oxford
OX2 6GG
Mathematical Institute, University of Oxford and National Physical Laboratory
OX2 6GG Oxford
UNITED KINGDOM

Sebastian Neumayer
TU Berlin
Institut für Mathematik
Straße des 17. Juni 136
10623 Berlin
GERMANY

Dr. Hiroyuki Sato
Department of Applied Mathematics and Physics
Graduate School of Informatics
Kyoto University
Yoshida-honmachi, Sakyo-ku
Kyoto 606-8501
JAPAN

Prof. Dr. Anton Schiela
Lehrstuhl für Angewandte Mathematik
Mathematisches Institut
Universität Bayreuth
95440 Bayreuth
GERMANY

Prof. Dr. Gabriele Steidl
Institut für Mathematik
Technische Universität Berlin
Sekretariat MA 4-3
Straße des 17. Juni 136
10623 Berlin
GERMANY

André Uschmajew
Max Planck Institute for Mathematics in the Sciences
Inselstr. 22
04103 Leipzig
GERMANY

Prof. Dr. Bart Vandereycken
Section of Mathematics
University of Geneva
Rue du Lièvre 2-4
1211 Genève 4
SWITZERLAND

Prof. Dr. Max Wardetzky
Institut für Numerische und Angewandte Mathematik
Universität Göttingen
Lotzestrasse 16-18
37083 Göttingen
GERMANY

Prof. Dr. Kathrin Welker
Fakultät für Maschinenbau
Helmut-Schmidt Universität / Universität der Bundeswehr Hamburg
Holstenhofweg 85
22043 Hamburg
GERMANY

Prof. Dr. Benedikt Wirth
Fachbereich Mathematik und Informatik
Universität Münster
Einsteinstrasse 62
48149 Münster
GERMANY