An efficient preconditioner for the Riemannian trust-region method on the manifold of fixed-rank matrices

Marco Sutti

National Center for Theoretical Sciences
Taipei, Taiwan

ICIAM 2023, Waseda University, Tokyo
August 25, 2023

Overview

Preprint: Implicit low-rank Riemannian schemes for the time integration of stiff partial differential equations, M. Sutti and B. Vandereycken, submitted, arXiv preprint arXiv:2305.11532.

Contributions:

- Preconditioner for the RTR method on the manifold of fixed-rank matrices.
- Applications within implicit numerical integration schemes to solve stiff, time-dependent PDEs.

This talk:

I. Optimization on matrix manifolds.

II. The manifold of fixed-rank matrices.

III. Preconditioner.

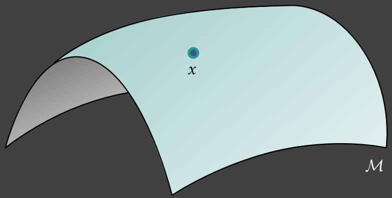
IV. Numerical application.

Optimization problems on matrix manifolds

- We can state the optimization problem as

$$\min_{x \in \mathcal{M}} f(x),$$

where $f : \mathcal{M} \to \mathbb{R}$ is the objective function and $\mathcal{M}$ is some matrix manifold.



- Matrix manifold: any manifold that is constructed from $\mathbb{R}^{n \times p}$ by taking either embedded submanifolds or quotient manifolds.
  - Examples of embedded submanifolds: orthogonal Stiefel manifold, manifold of symplectic matrices, manifold of fixed-rank matrices, …
  - Example of quotient manifold: the Grassmann manifold.
- Motivation: by exploiting the underlying geometric structure, only feasible points are considered!

Manifold optimization: [Edelman et al. 1998, Absil et al. 2008, Boumal 2023], …

Problems considered: variational problems

- Variational problem, called "LYAP" herein,

$$\begin{cases} \min_w \mathcal{F}(w(x,y)) = \int_\Omega \frac{1}{2}\|\nabla w(x,y)\|^2 - \gamma(x,y)\, w(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ \quad \text{such that} \quad w = 0 \text{ on } \partial\Omega, \end{cases}$$

  where $\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)$, $\Omega = [0,1]^2$ and γ is the source term.

- Discretization on a uniform grid: regardless of the specific form of $\mathcal{F}$, we obtain the general formulation

$$\min_W F(W) \quad \text{s.t.} \quad W \in \{X \in \mathbb{R}^{n\times n} \colon \operatorname{rank}(X) = r\},$$

  where F denotes the discretization of the functional $\mathcal{F}$.

"LYAP" variational problem: [Henson 2003, Gratton/Sartenaer/Toint 2008, Wen/Goldfarb 2009, S./Vandereycken 2021, …]

Riemannian manifold and gradient

A manifold $\mathcal{M}$ endowed with a smoothly-varying inner product (called Riemannian metric g) is called Riemannian manifold.

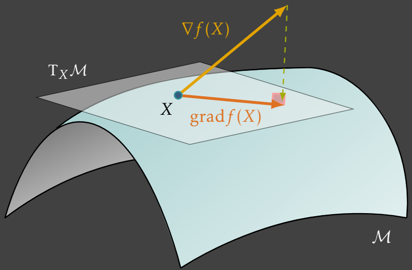
$\rightsquigarrow$ A couple $(\mathcal{M}, g)$, i.e., a manifold with a Riemannian metric on it.

Let $f : \mathcal{M} \to \mathbb{R}$. E.g., the objective function in an optimization problem.

$\rightsquigarrow$ For <u>any</u> embedded submanifold:

- Riemannian gradient: projection onto $\mathrm{T}_X\mathcal{M}$ of the Euclidean gradient

$$\operatorname{grad} f(X) = \mathrm{P}_{\mathrm{T}_X\mathcal{M}}(\nabla f(X)).$$



$\rightsquigarrow$ $\nabla f(X)$ is the Euclidean gradient of $f(X)$.

Matrix and vector calculus: The Matrix Cookbook, www.matrixcalculus.org, …
Automatic differentiation on low-rank manifolds: [Novikov/Rakhuba/Oseledets 2022]

The manifold of fixed-rank matrices

- Our optimization problem is defined over

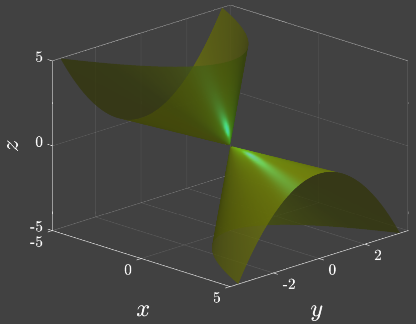
$$\mathcal{M}_r = \{X \in \mathbb{R}^{n\times n} : \operatorname{rank}(X) = r\}.$$

$\rightsquigarrow$ $\mathcal{M}_r$ has a smooth structure …

2 × 2 example:

$$X = \begin{bmatrix} x & -2y \\ y & z \end{bmatrix}.$$

Parametrization:
$\operatorname{rank}(X) = 1 \Leftrightarrow xz = -2y^2$ and $x, z \neq 0$.



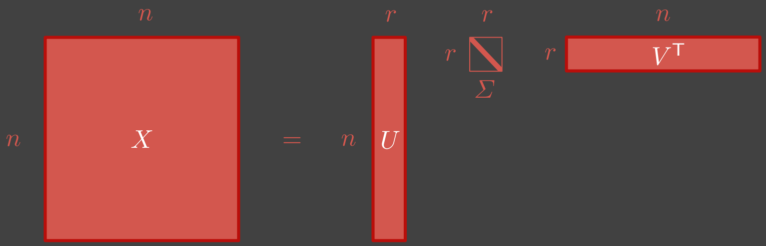
- Theorem: $\mathcal{M}_r$ is a smooth Riemannian submanifold embedded in $\mathbb{R}^{n\times n}$ of dimension $r(2n - r)$.

Optimizing on submanifold $\mathcal{M}_r$: [Vandereycken 2013]

Alternative characterization

- Using the singular value decomposition (SVD), we have the equivalent characterization

$$\mathcal{M}_r = \{U\Sigma V^\top : \ U^\top U = I_r,\ V^\top V = I_r,\ \Sigma = \mathrm{diag}(\sigma_i),\ \sigma_1 \geqslant \cdots \geqslant \sigma_r > 0\}.$$



- Only $2nr + r$ coefficients instead of n^2. If $r \ll n$, then big memory savings.
- Perform the calculations directly in the factorized format.

Riemannian Hessian and preconditioning/1

- In the case of Riemannian submanifolds, the full Riemannian Hessian of f at $x \in \mathcal{M}$ is given by the projected Euclidean Hessian plus the curvature part

$$\operatorname{Hess} f(x)[\xi] = P_x \nabla^2 f(x) P_x + P_x (\text{“curvature terms”}) P_x.$$

$\rightsquigarrow$ Use $P_x \nabla^2 f(x) P_x$ as a preconditioner in RTR.

- For LYAP, we can get the symmetric n^2-by-n^2 matrix

$$H_X = P_X (A \otimes I + I \otimes A) P_X.$$

- Inverse of H_X $\rightsquigarrow$ good candidate for a preconditioner.

  ⚠ Not inverted directly, since this would cost $\mathcal{O}(n^6)$.

- A good preconditioner should reduce the number of iterations of the inner trust-region solver. It has to be effective and cheap to compute.

Symmetric positive semidefinite matrices with fixed rank: [Vandereycken/Vandewalle 2010]

Riemannian Hessian and preconditioning/2

- Applying the preconditioner in $X \in \mathcal{M}_r$ means solving for $\xi \in \mathrm{T}_X\mathcal{M}$ the system
$$H_X \operatorname{vec}(\xi) = \operatorname{vec}(\eta),$$
where $\eta \in \mathrm{T}_X\mathcal{M}$ is a known tangent vector.
- This is equivalent to
$$\mathrm{P}_X(A\xi + \xi A) = \eta.$$
- Using the definition of the orthogonal projector onto $\mathrm{T}_X\mathcal{M}_r$, we obtain
$$P_U(A\xi + \xi A)P_V + P_U^{\perp}(A\xi + \xi A)P_V + P_U(A\xi + \xi A)P_V^{\perp} = \eta,$$
which is equivalent to the system
$$\begin{cases} P_U(A\xi + \xi A)P_V = P_U \eta P_V, \\ P_U^{\perp}(A\xi + \xi A)P_V = P_U^{\perp} \eta P_V, \\ P_U(A\xi + \xi A)P_V^{\perp} = P_U \eta P_V^{\perp}. \end{cases}$$

$$\vdots$$

$\rightsquigarrow$ Many (tedious) calculations, but the numerical results are pretty striking!

"LYAP" variational problem

Table: Effect of preconditioning: dependence on size for LYAP.

		Rank 5						Rank 10					
Prec.	size	10	11	12	13	14	15	10	11	12	13	14	15
No	n_{outer}	51	54	61	59	162	92	300	103	61	63	62	59
	$\sum n_{\text{inner}}$	4561	9431	21066	36556	30069	30096	27867	30025	33818	45760	44467	38392
	$\max n_{\text{inner}}$	1801	3191	7055	9404	1194	1851	2974	3385	8894	24367	24537	25013
Yes	n_{outer}	41	45	50	52	56	60	44	64	62	53	56	56
	$\sum n_{\text{inner}}$	44	45	50	52	56	60	69	104	82	60	69	56
	$\max n_{\text{inner}}$	4	1	1	1	1	1	9	9	8	8	8	1

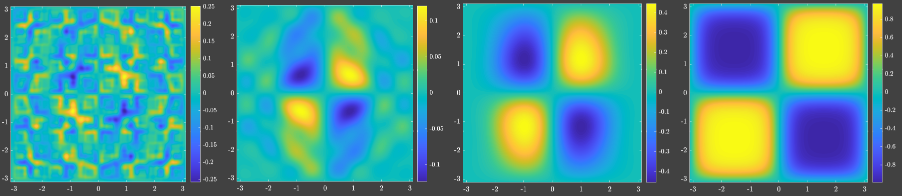
- Stopping criterion: maximum number of outer iterations $n_{\text{max outer}} = 300$. The inner solver is stopped when $\sum n_{\text{inner}}$ first exceeds $30\,000$.
- Impressive reduction in the number of iterations of the inner solver.
- n_{outer} and $\sum n_{\text{inner}}$ depend (quite mildly) on size, while $\max n_{\text{inner}}$ is basically constant.

Allen–Cahn equation/1

- Reaction-diffusion equation that models the process of phase separation in multi-component alloy systems.
  - Other applications include: mean curvature flows, two-phase incompressible fluids, complex dynamics of dendritic growth, and image segmentation …
- In its simplest form, it reads

$$\frac{\partial w}{\partial t} = \varepsilon \Delta w + w - w^3.$$

- It is a stiff, time-dependent PDE.



(a) $t = 0$ (b) $t = 0.5$ (c) $t = 2$ (f) $t = 15$

Figure: Time evolution of the solution w to the Allen–Cahn equation, with ERK4, $h = 10^{-4}$.

Allen–Cahn equation: [Allen/Cahn 1972, Allen/Cahn 1973]

Allen–Cahn equation/2 - low-rank evolution

- We build the functional

$$\min_{w} \mathcal{F}(w) := \int_{\Omega} \frac{\varepsilon h}{2}\|\nabla w\|^2 + \frac{(1-h)}{2} w^2 + \frac{h}{4} w^4 - \widetilde{w} \cdot w \,\mathrm{d}x\,\mathrm{d}y.$$

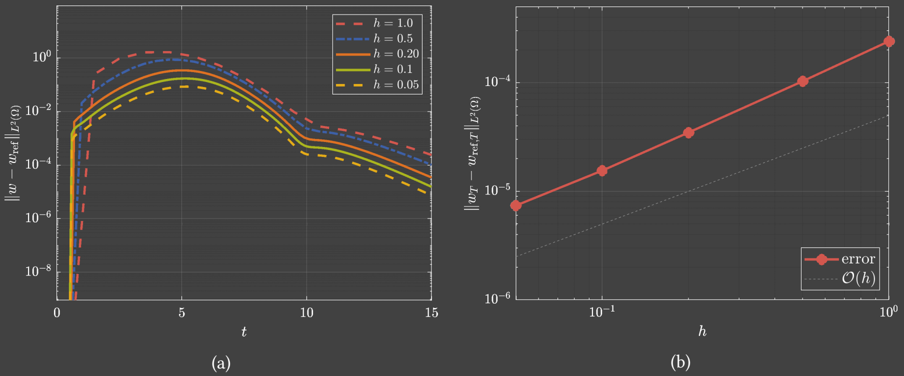


Figure: Panel (a): error versus time for the preconditioned low-rank evolution of the Allen-Cahn equation. Panel (b): error at $T = 15$ versus time step h.

Conclusions

Pros and cons:

⊕ Efficient preconditioner on the manifold of fixed-rank matrices.

⊕ Solid, quite well-understood mathematical theory behind.

⊖ If the problem does not admit a low-rank representation, then there is no advantage over using dense matrices.

Outlook:

- Go to higher-order numerical integration methods.
- Other applications in mind, e.g., diffusion problems in mathematical biology or problems with low-rank tensor structure.

Thank you for your attention!

Questions?

Bonus material

Metric, projection, gradient, retraction

- The Riemannian metric is

$$g_X(\xi,\eta) = \langle \xi,\eta\rangle = \mathrm{Tr}(\xi^\top \eta), \quad \text{with} \quad X \in \mathcal{M}_r \quad \text{and} \quad \xi,\eta \in \mathrm{T}_X\mathcal{M}_r,$$

  where ξ, η are seen as matrices in the ambient space $\mathbb{R}^{n\times n}$.

- Orthogonal projection onto the tangent space at X is

$$\mathrm{P}_{\mathrm{T}_X\mathcal{M}_r} : \mathbb{R}^{n\times n} \to \mathrm{T}_X\mathcal{M}_r, \qquad Z \to \mathrm{P}_U\, Z\, \mathrm{P}_V + \mathrm{P}_U^\perp\, Z\, \mathrm{P}_V + \mathrm{P}_U\, Z\, \mathrm{P}_V^\perp.$$

- Riemannian gradient: projection onto $\mathrm{T}_X\mathcal{M}_r$ of the Euclidean gradient

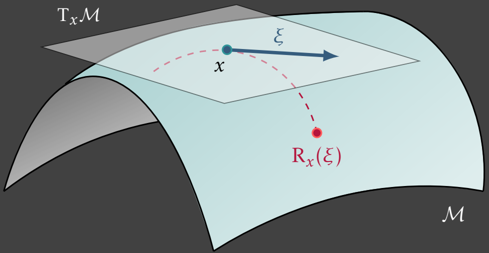
$$\mathrm{grad}\, f(X) = \mathrm{P}_{\mathrm{T}_X\mathcal{M}_r}(\nabla f(X)).$$

- Retraction $\mathrm{R}_X : \mathrm{T}_X\mathcal{M}_r \to \mathcal{M}_r$. Typical: truncated SVD.

Many retractions for $\mathcal{M}_r$: [Absil/Oseledets 2015]

Retractions

- Move in the direction of ξ while remaining constrained to $\mathcal{M}$.
- Smooth mapping $\mathrm{R}_x : \mathrm{T}_x\mathcal{M} \to \mathcal{M}$ with a local condition that preserves gradients at x.



- The Riemannian exponential mapping is also a retraction, but it is not computationally efficient.
- Retractions: first-order approximation of the Riemannian exponential!

Constructing retractions: [Absil/Malick 2012]

Riemannian trust-region (RTR) method

Algorithm 1: Riemannian trust-region (RTR)

Given $\bar{\Delta} > 0$, $\Delta_1 \in (0, \bar{\Delta})$
for $i = 1, 2, \ldots$ **do**
  Define the second-order model

$$m_i \colon \mathrm{T}_{x_i}\mathcal{M} \to \mathbb{R},\ \xi \mapsto f(x_i) + \langle \operatorname{grad} f(x_i), \xi \rangle + \frac{1}{2} \langle \operatorname{Hess} f(x_i)[\xi], \xi \rangle.$$

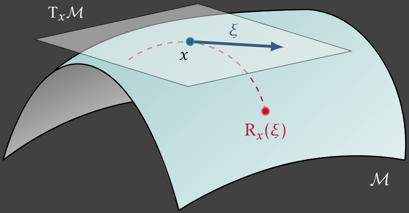
  Trust-region subproblem: compute η_i by solving

$$\eta_i = \operatorname{argmin} m_i(\xi) \quad \text{s.t.} \quad \|\xi\| \le \Delta_i.$$

  Compute $\rho_i = (\widehat{f}(0) - \widehat{f}_i(\eta_i))/(m_i(0) - m_i(\eta_i))$.
  if $\rho_i \ge 0.05$ **then**
    Accept step and set $x_{i+1} = \mathrm{R}_{x_i}(\eta_i)$.
  else
    Reject step and set $x_{i+1} = x_i$.
  end if
  Radius update: set

$$\Delta_{i+1} = \begin{cases} \min(2\Delta_i, \bar{\Delta}) & \text{if } \rho_i \ge 0.75 \text{ and } \|\eta_i\| = \Delta_i, \\ 0.25\|\eta_i\| & \text{if } \rho_i \le 0.25, \\ \Delta_i & \text{otherwise.} \end{cases}$$

end for



TR method: [Goldfeld/Quandt/Trotter 1966, Sorensen 1982, Fletcher 1980/1987 …]
RTR method: [Absil/Baker/Gallivan 2007]

An example of factorized gradient

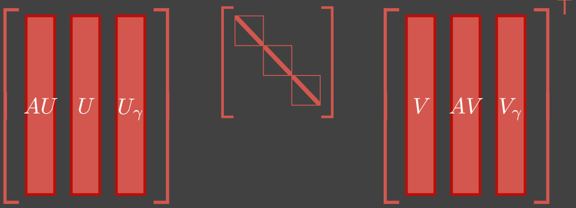
- "LYAP" functional: $\mathcal{F}(w(x,y)) = \int_\Omega \frac{1}{2}\|\nabla w(x,y)\|^2 - \gamma(x,y)\, w(x,y)\,\mathrm{d}x\,\mathrm{d}y$.
- The gradient of $\mathcal{F}$ is the variational derivative $\frac{\delta \mathcal{F}}{\delta w} = -\Delta w - \gamma$.
- The discretized Euclidean gradient in matrix form is given by

$$G = AW + WA - \Gamma.$$

  with A is the second-order periodic finite difference differentiation matrix.
- The first-order optimality condition $G = AW + WA - \Gamma = 0$ is a Lyapunov (or Sylvester) equation.

$\rightsquigarrow$ Factorized Euclidean gradient:

$$G = \begin{bmatrix} AU & U & U_\gamma \end{bmatrix} \operatorname{blkdiag}\left(\Sigma,\ \Sigma,\ \Sigma_\gamma\right) \begin{bmatrix} V & AV & V_\gamma \end{bmatrix}^\top.$$



Tangent vectors

- A tangent vector ξ at $X = U\Sigma V^\top$ is represented as

$$\xi = UMV^\top + U_pV^\top + UV_p^\top,$$

$$M \in \mathbb{R}^{r\times r}, \quad U_p \in \mathbb{R}^{n\times r}, \quad U_p^\top U = 0, \quad V_p \in \mathbb{R}^{n\times r}, \quad V_p^\top V = 0.$$

- We can rewrite it as

$$\xi = (UM + U_p)V^\top + UV_p^\top.$$

$\rightsquigarrow$ ξ is a rank-$2r$ bounded matrix. Useful in implementation.