Riemannian preconditioning

Bamdev Mishra

Joint work with Rodolphe Sepulchre

University of Liège, Belgium
University of Cambridge, UK

Workshop on low-rank optimization and applications
Hausdorff Center for Mathematics, Bonn, Germany

June 2015
Example 1: matrix completion

\begin{equation}
\begin{bmatrix}
? & ? & * & ? \\
* & * & ? & * \\
? & * & * & ? \\
* & ? & * & ? \\
\end{bmatrix}
\end{equation}

$n$ Movies

$\approx$

$m$ Users

\begin{equation}
\begin{bmatrix}
r \\
n \\
G \\
H^T \\
m \\
\end{bmatrix}
\end{equation}

Low-rank prior

$(n + m - r)r, r \ll (m, n)$

($G$ and $H$ are full column rank matrices)

[Netflix Challenge, 2006]
Example 1: matrix completion as a rank constrained error minimization problem

Minimize the error to the known ratings,

\[
\min_{\mathbf{X} \in \mathbb{R}^{n \times m}} \sum_{(i,j) \in \Omega} (\tilde{X}_{ij} - (X)_{ij})^2
\]

subject to \( \text{rank}(\mathbf{X}) = r \)

\[
\text{Fixed-rank, } \mathbf{X} = \mathbf{G} \mathbf{H}^T
\]

\[
\min_{\mathbf{G} \in \mathbb{R}_{*}^{n \times r}} \sum_{(i,j) \in \Omega} (\tilde{X}_{ij} - (\mathbf{G} \mathbf{H}^T)_{ij})^2.
\]

\( \Omega \) is the set of known ratings.

\( \mathbb{R}_{*}^{m \times r} \) denotes full rank \( m \times r \) matrices.
Example 1: fixed-rank parameterizations have structured symmetries

\[
X = \begin{pmatrix}
\mathbf{G} \\
\mathbf{M} & \mathbf{M}^{-1}
\end{pmatrix}
\begin{pmatrix}
\mathbf{H}^T \\
\mathbf{H}^T
\end{pmatrix}
\]

\[
= \begin{pmatrix}
\mathbf{G} \\
\mathbf{M} & \mathbf{M}^{-1}
\end{pmatrix}
\begin{pmatrix}
\mathbf{H}^T \\
\mathbf{H}^T
\end{pmatrix}
\]

\(\mathbf{M}\) is \(r\)-by-\(r\) non-singular.

Other fixed-rank parameterizations have symmetries too.
Example 1: matrix completion as a least-squares problem with symmetries

\[ \begin{array}{l}
\min_{G \in \mathbb{R}^{n \times r}, \ H \in \mathbb{R}^{m \times r}} \\
\sum_{(i,j) \in \Omega} (\tilde{X}_{ij} - (GH^T)_{ij})^2 \\
\Rightarrow \text{Optimization on } [G, H]
\end{array} \]

- Equivalence classes:
  \[ [G, H] := \{(GM^{-1}, HM^T) : M \text{ non-singular } \in \text{GL}(r)\}. \]

- Explicit computations: \((G, H) \in \mathbb{R}^{n \times r} \times \mathbb{R}^{m \times r}\)

- Implicit optimization: \([G, H] \in \mathbb{R}^{n \times r} \times \mathbb{R}^{m \times r} / \text{GL}(r).\)
Example 2: eigenvalue problem

\[
\max_{x \in \mathbb{R}^n} \frac{x^T A x}{x^T x}
\]

- **cost unchanged** under map \( x \mapsto \alpha x \), \( \alpha \) non-zero scalar \( \in \mathbb{R}_* \).
- **Solutions not isolated**, i.e., but are equivalence classes \([x] = \{ \alpha x : \alpha \in \mathbb{R}_* \}\) (real projective space).
- An interpretation: **direction** is important and not length.

Explicit computations: \( \mathbb{R}^n \), but
Implicit optimization: \( \mathbb{R}^n / \mathbb{R}_* \).
Example 2: eigenvalue problem, the block case

- Generalization: $r$-dominant eigenvalue-eigenvector pairs:

\[
\max_{X \in \mathbb{R}^{n \times r}} \quad \text{Trace}(X^TAX) \\
\text{subject to} \quad X^TX = I.
\]

- Symmetry under map $X \mapsto XO$ such that $OO^T = O^TO = I$, denoted by $O(r)$.

- Interpretation: subspace is important and not basis vectors.

- Search space is Grassmann manifold $St(r, n)/O(r)$ [Edelman et al., 1998; Absil et al., 2008].
Problems have two fundamental structures: least-squares and symmetries

\[
\min_{x \in \mathcal{M}} f(x) \quad \leftarrow \text{least-squares cost}
\]

subject to \([x] \in \mathcal{M}/\sim \). \leftarrow \text{equivalence classes on } \mathcal{M}

Both \(\mathcal{M}\) and \(\sim\) result from interplay of few matrix manifolds. \(\mathcal{M}/\sim\) has quotient manifold structure.

<table>
<thead>
<tr>
<th>Rank constraint</th>
<th>Orthogonality constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{St}(r, n))</td>
<td>(\text{St}(r, n))</td>
</tr>
<tr>
<td>(\mathbb{R}^n_{\times r})</td>
<td>(\mathbb{R}^n_{\times r})</td>
</tr>
<tr>
<td>(\text{O}(r))</td>
<td>(\text{O}(r))</td>
</tr>
<tr>
<td>(\text{GL}(r))</td>
<td>(\text{GL}(r))</td>
</tr>
<tr>
<td>(S_{++}(r))</td>
<td>(S_{++}(r))</td>
</tr>
</tbody>
</table>
A number of applications relate to quadratic optimization with orthogonality and rank constraints

\[
\begin{align*}
\text{min} & \quad \text{Trace}(X^TAX) \\
\text{subject to} & \quad X^TX = I
\end{align*}
\]

Power, inverse, RQI, . . .
PCA, Sparse PCA, . . .
Max-cut
Procrustes

\[
\begin{align*}
\text{min} & \quad \|X - X^*\|_F^2 \\
\text{subject to} & \quad \text{rank}(X) = r
\end{align*}
\]

Linear matrix equations, e.g., Lyapunov
Matrix completion, Tensor completion
Optimization on manifold framework has gained much attention lately

\[
\min \quad f(x) \\
\text{subject to} \quad x \in \mathcal{M}.
\]

- Nonlinear optimization methods: SQP.

- Geometric methods: Optimization on manifolds [Smith, 1994; Edelman et al., 1998; Absil et al., 2008]

  + Manopt: a Matlab toolbox at Manopt.org [Boumal et al., 2014].
Our contribution

- Exploit **cost** and **constraints with symmetries** structures to develop efficient algorithms.

- We propose a notion of **preconditioning on manifolds** with
  - **metric tuning** (identifying a good inner product)
  - exploit second-order information in first-order algorithms.
Outline

- Motivation for a geometric framework for constraints with symmetries

- Exploiting structures
  - Metric tuning
  - Quadratic optimization with orthogonality and rank constraints

- Algorithms for low-rank matrix completion with fixed-rank constraint
Outline

- Motivation for a geometric framework for constraints with symmetries
- Exploiting structures
  - Metric tuning
  - Quadratic optimization with orthogonality and rank constraints
- Algorithms for low-rank matrix completion with fixed-rank constraint
Two complementary views of optimization

Sequential quadratic programming

\[
\min_{x \in \mathbb{R}^n} f(x) \\
\text{subject to } h(x) = 0
\]

(constraints are embedded into the cost function)

\[
\max \min_{\lambda \in \mathbb{R}^p} \min_{x \in \mathbb{R}^n} f(x) - \langle \lambda, h(x) \rangle,
\]
where $\lambda$ is the Lagrange multiplier

Riemannian framework

\[
\min_{x \in M} f(x),
\]
where $M = \{x : h(x) = 0\}$ has dimension $p$

(constraints are encoded into the search space)
Two complementary views of optimization

- Lagrangian view
  - (Constrained viewpoint)

- Manifold optimization view
  - (An ant-on-manifold view: unconstrained viewpoint)
Sequential quadratic programming (SQP)

\[
\min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to} \quad h(x) = 0.
\]

1. Compute search direction \(\zeta_x^*\),

\[
\begin{align*}
\arg \min_{\zeta_x \in \mathbb{R}^n} & \quad f(x) + \langle f_x(x), \zeta_x \rangle + \frac{1}{2} \langle \zeta_x, D^2 \mathcal{L}(x, \lambda_x)[\zeta_x] \rangle \\
\text{subject to} & \quad D h(x)[\zeta_x] = 0. \\
& \quad (\langle \cdot, \cdot \rangle \text{ is scalar product})
\end{align*}
\]

2. Next \(x_+\) is obtained by projecting \(x + s\zeta_x^*\) onto \(h(x) = 0\).

Lagrangian function \(\mathcal{L}(x, \lambda) = f(x) - \langle \lambda, h(x) \rangle\).

Estimate \(\lambda_x = (h_x(x)(h_x(x))^T)^{-1}h_x(x)f_x(x)\) locally [Nocedal and Wright (2006)].
A critical assumption of SQP is not satisfied for constraints with symmetries

- Competitive algorithm near the minimum.

- A well-defined problem with unique solution when
  \[ \langle \xi_x, D^2 L(x, \lambda_x)[\xi_x] \rangle > 0 \text{ on } Dh(x)[\xi_x] = 0. \]

- Condition not satisfied for a search space with symmetries
  \[ \langle \xi_x, D^2 L(x, \lambda_x)[\xi_x] \rangle \geq 0. \]
Motivation for Riemannian framework

SQP fails on the eigenvalue problem

\[
\min_{x \in \mathbb{R}^n} - \frac{x^T A x}{x^T x}
\]

1. Compute the search direction

\[
\zeta_x^* = \arg \min_{\zeta_x \in \mathbb{R}^n} f(x) + \langle f_x(x), \zeta_x \rangle + \frac{1}{2} \langle \zeta_x, D^2 \mathcal{L}(x)[\zeta_x] \rangle
\]

\[
\Rightarrow \quad \zeta_x^* = x. \quad \text{(true for any homogenous function)}
\]

2. Consequently, the next iterate \( x_+ = x + s \zeta_x^* = (1 + s)x \), i.e., \( x_+ = (1 + s)x \).
Why SQP fails for constraints with symmetries?

\[ Dh(x)[\xi_x] = 0 \equiv T_x\mathcal{M} \]

\[ T_x\mathcal{M} = \mathcal{H}_x \oplus \mathcal{V}_x \]

\[ [x] := \{ y : y \sim x, y \in \mathcal{M} \} \]

\[ [x_+] := \{ \zeta_x \in T_x\mathcal{M} : \zeta_x^* = \arg \min_{\zeta_x \in T_x\mathcal{M}} f(x) + \langle f_x(x), \zeta_x \rangle + \frac{1}{2} \langle \zeta_x, D^2\mathcal{L}(x)[\zeta_x] \rangle \in \mathcal{V}_x \}. \]

\[ \text{Resolve: exclude } \mathcal{V}_x, \text{ and only } \mathcal{H}_x \text{ is relevant.} \]
The Riemannian optimization framework requires a Riemannian metric

The Riemannian framework enables us to separate $\mathcal{H}_x$ and $\mathcal{V}_x$.

Riemannian metric that is invariant to $[x]$. 

Motivation for Riemannian framework
The Riemannian steepest-descent algorithm

\[ \min_{x \in \mathcal{M}} f(x) \]

1. Compute the negative Riemannian gradient \( \xi_x = -\text{grad}_x f \) w.r.t Riemannian metric \( g_x \), i.e.,

\[
\text{grad}_x f = - \arg \min_{\zeta_x \in T_x \mathcal{M}} f(x) + \langle f_x(x), \zeta_x \rangle + \frac{1}{2} g_x(\zeta_x, \zeta_x).
\]

2. Next iterate \( x_+ \) is computed using the retraction, equivalent to projection.

- A well-defined scheme on the quotient manifold \( \mathcal{M}/\sim \) of \( \mathcal{M} \).
Metric profoundly affects performance of Riemannian gradient algorithm

- Conventional: metric $g$ is only motivated by symmetry search space, but ignores cost.

- As a result, performance of the steepest-descent algorithm is profoundly affected for different cost functions [Manton, 2002].
Outline

- Motivation for a geometric framework for constraints with symmetries
- Exploiting structures
  - Metric tuning
  - Quadratic optimization with orthogonality and rank constraints
- Algorithms for low-rank matrix completion with fixed-rank constraint
Connecting SQP to Riemannian gradient descent

SQP: \[ \arg \min_{\zeta_x \in \mathbb{R}^n} \quad f(x) + \langle f_x(x), \zeta_x \rangle + \frac{1}{2} \langle \zeta_x, D^2 \mathcal{L}(x, \lambda_x)[\zeta_x] \rangle \]
subject to \[ Dh(x)[\zeta_x] = 0 \]

Riemann: \[ \arg \min_{\zeta_x \in T_x \mathcal{M}} \quad f(x) + \langle f_x(x), \zeta_x \rangle + \frac{1}{2} g_x(\zeta_x, \zeta_x) \]

Theorem

If \( x^* \in \mathcal{M} \) is a local minimum of \( f : \mathcal{M} \rightarrow \mathbb{R} \) on \( \mathcal{M}/\sim \), then

(i) \[ \langle \eta_{x^*}, D^2 \mathcal{L}(x^*, \lambda_{x^*})[\eta_{x^*}] \rangle = 0 \] for all \( \eta_{x^*} \in \mathcal{V}_{x^*} \),

(ii) \[ \langle \xi_{x^*}, D^2 \mathcal{L}(x^*, \lambda_{x^*})[\xi_{x^*}] \rangle \] captures the second-order information.
The second-derivative of the Lagrangian induces a valid metric on the quotient space

Metric induced by Lagrangian, i.e., \( \langle \zeta_x, D^2 \mathcal{L}(x, \lambda_x)[\eta_x] \rangle \) is only a pseudometric in \( T_x \mathcal{M} \).

What we require:

- metric is well-defined in the entire \( T_x \mathcal{M} \).
- the metric is a global structure.

Resolve: we exploit the Lagrangian structure further.
Riemannian optimization and local convexity

Lagrangian $\mathcal{L}(x, \lambda) = f(x) - \langle \lambda, h(x) \rangle$, $\langle \cdot, \cdot \rangle$ standard inner product.

$$g_x(\xi_x, \eta_x) = \langle \xi_x, D^2\mathcal{L}(x, \lambda)[\eta_x] \rangle$$

$$= \langle \xi_x, D^2f(x)[\eta_x] \rangle + \langle \xi_x, D^2c(x, \lambda)[\eta_x] \rangle.$$
Quadratic optimization with orthogonality constraints: revisiting the eigenvalue problem

\[ \max_{X \in \mathbb{R}^{n \times r}} \text{Trace}(X^TAX) \quad \text{subject to} \quad X^TX = I \]

\[ \lambda_x = X^TAX \]

\[ \mathcal{L}(x, \lambda_x) = \text{Trace}(X^TAX)/2 - \langle \lambda_x, X^TX - I \rangle/2 \]

\[ \Rightarrow \quad \mathcal{L}_x(x, \lambda_x) = AX - X\lambda_x \]

\[ \Rightarrow \quad D^2\mathcal{L}(x, \lambda_x)[\xi_x] = A\xi_x - \xi_x\lambda_x \]

\[ g_x(\xi_x, \eta_x) = \left\langle \xi_x, A\eta_x \right\rangle - \left\langle \xi_x, \eta_xX^TAX \right\rangle. \]

- Connects to power, inverse, and Rayleigh quotient iterations [Absil et al., 2002].
Capturing both cost and constraint information in metric leads to a superior performance

\[ \langle \xi_x, A \eta_x \rangle - \omega \langle \xi_x, \eta_x X^T A X \rangle \]

Figure: \( \omega \) is updated with iterations. metric: \( \langle \xi_x, A \eta_x \rangle - \omega \langle \xi_x, \eta_x X^T A X \rangle \).
Similar story for generalized eigenvalue problem

\[
\begin{align*}
\min_{X \in \mathbb{R}^{n \times r}} & \quad \frac{1}{2} \text{Trace}(X^TAX) \\
\text{subject to} & \quad X^T BX = I,
\end{align*}
\]

\[
g_x(\xi_x, \eta_x) = \langle \xi_x, A \eta_x \rangle - \langle \xi_x, B \eta_x \lambda_x \rangle,
\]

\[
\lambda_x = \text{Sym}((X^TBBX)^{-1}(X^TBAX)).
\]
Quadratic optimization with the rank constraint

\[ \min_{X \in \mathbb{R}^{n \times m}} \ \frac{1}{2} \text{Trace}(X^TAXB) + \text{Trace}(X^TC) \]
subject to \( \text{rank}(X) = r. \)

- We use the parameterization \( X = GH^T, \ G \in \mathbb{R}^{n \times r}, \ H \in \mathbb{R}^{m \times r}. \)
- The cost is **quadratic and convex** in arguments \( G, H \) individually.

\[ g_x(\xi_x, \eta_x) = \langle \eta_G, 2AG\text{Sym}(H^TB\xi_H) + C\xi_H \rangle \]
[\[ + \langle \eta_H, 2BHS\text{Sym}(G^TA\xi_G) + C^T\xi_G \rangle \]
[\[ + \langle \eta_G, A\xi_GH^TBH \rangle + \langle \eta_H, B\xi_HG^TAG \rangle. \]

\( x = (G, H). \)

- The block-diagonal choice provides a **simpler and convenient** metric choice.

Block diagonal approximation of \( L_{xx}(x) \)
The case of symmetric fixed-rank constraint is dealt with similarly.

The case of symmetric fixed-rank constraint is handled by making \( G = H \), i.e.,

\[
X = GG^T.
\]

The metric tuning ideas follow through.
Outline

- Motivation for a geometric framework for constraints with symmetries
- Exploiting structures
  - Metric tuning
  - Quadratic optimization with orthogonality and rank constraints
- Algorithms for low-rank matrix completion with fixed-rank constraint
Revisiting the matrix completion problem: a simpler cost function to infer the metric

\[
\min_{\mathbf{X} \in \mathbb{R}^{n \times m}} \sum_{(i,j) \in \Omega} (\tilde{X}_{ij} - X_{ij})^2 \quad \text{subject to} \quad \text{rank}(\mathbf{X}) = r.
\]

- **Second-order derivative** of \( \sum_{(i,j) \in \Omega} (\tilde{X}_{ij} - X_{ij})^2 \) w.r.t \( \mathbf{X} = \mathbf{GH}^T \) is computationally cumbersome.

- Consider **simpler** cost function

\[
\min_{\mathbf{X} \in \mathbb{R}^{n \times m}} \frac{1}{2} \text{Trace}(\mathbf{X}^T \mathbf{X}) - \text{Trace}(\mathbf{X}^T \tilde{\mathbf{X}}) \equiv \| \mathbf{X} - \tilde{\mathbf{X}} \|_F^2 \quad \text{subject to} \quad \text{rank}(\mathbf{X}) = r.
\]
The matrix scaling acts a computationally efficient preconditioner

- Two parameterizations

\[ X = GH^T \quad \text{(two-factor factorization)} \]
\[ \mathbb{R}_r^{n \times r} \times \mathbb{R}_r^{m \times r} \]

\[ X = URV^T \quad \text{(SVD-type factorization).} \]
\[ \text{St}(r, n) \times \text{GL}(r) \times \text{St}(r, m) \]

- Novel metrics based on block diagonal approximation

\[ g_x(\xi_x, \eta_x) = \langle \eta_G, \xi_G H^T H \rangle + \langle \eta_H, \xi_H G^T G \rangle \]
\[ g_x(\xi_x, \eta_x) = \langle \eta_U, \xi_U R R^T \rangle + \langle \eta_R, \xi_R \rangle + \langle \eta_V, \xi_V R^T R \rangle \]

\[ x = (G, H), \quad x = (U, R, V). \]
We develop conjugate gradient algorithms

- The matrix scaling connects to various state-of-the-art algorithms, e.g., scaled ALS [Wen et al., 2012] and scaled subspace iteration [Ngo and Saad, 2012].

- Our algorithms, R2MC and R3MC, have shown competitive performance.
Under high sampling, the performance of metrics is distinctly different
Movielens-1M dataset

(Similar conclusion also obtained on the Netflix and Yahoo datasets.)
Extension to tensors: Tucker decomposition

\[ \mathcal{X} = [\mathcal{G}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3] = \mathcal{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3, \]

where

\((r_1, r_2, r_3)\) is the multilinear rank of \(\mathcal{X}\) and \(\mathbf{U}_1, \mathbf{U}_2, \) and \(\mathbf{U}_3\) are on the Stiefel manifold.

\((\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathcal{G}) \mapsto (\mathbf{U}_1 \mathbf{O}_1, \mathbf{U}_2 \mathbf{O}_2, \mathbf{U}_3 \mathbf{O}_3, \mathcal{G} \times_1 \mathbf{O}_1^T \times_2 \mathbf{O}_2^T \times_3 \mathbf{O}_3^T)\)
Riemannian preconditioning for tensor completion

Red is ours...

Figure: Small-scale instances of rank \((10, 10, 10)\). OS denotes the over sampling ratio.
Tensor completion: large-scale instances

Red is ours...
Preconditioning:


Matrix completion:


Tensor completion:

Summary

- Categorized popular problems into quadratic optimization with rank and orthogonality constraints.

- Importance of a geometric framework.

- Question of selecting a metric addressed. Least-squares with rank and/or orthogonality constraints.

- Classical Power, inverse, and Rayleigh quotient iterations interpreted.

- Concrete large-scale algorithms developed for low-rank matrix completion with novel metric.
Riemannian preconditioning

Bamdev Mishra

Joint work with Rodolphe Sepulchre

University of Liège, Belgium
University of Cambridge, UK

Workshop on low-rank optimization and applications
Hausdorff Center for Mathematics, Bonn, Germany

June 2015
References


