Optimization on manifolds and semidefinite relaxations

Nicolas Boumal
Princeton University

Based on work with Pierre-Antoine Absil, Afonso Bandeira, Coralia Cartis and Vladislav Voroninski
Max-Cut relaxation
An example of global optimality on manifolds
Max-Cut

Given a graph, split its nodes in two classes, maximizing the number of in-between edges.

One of Karp's 21 NP-complete problems.
Max-Cut semidefinite relaxation

$A$ is the adjacency matrix of the graph:

$$\min_X \text{Tr}(AX) \text{ s.t. } \text{diag}(X) = 1, X \succeq 0$$

Goemans & Williamson '95

Approximate the best cut within 87% by randomized projection of optimal $X$ to $\{\pm 1\}^n$. 
Convex, but IPM’s run out of memory (and time)

For a 2000 node graph (edge density 1%), CVX runs out of memory on my former laptop. On the new one, it returns with poor accuracy after 3 minutes.

The methods we will discuss solve the SDP in 6 seconds on old laptop, with certificate.
Max-Cut SDP has a low-rank solution

$$\min_X \text{Tr}(AX) \text{ s.t. } \text{diag}(X) = 1, X \succeq 0$$

Shapiro ‘82, Grone et al. ‘90, Pataki ‘94, Barvinok ‘95

There is an optimal $X$ whose rank $r$ satisfies

$$\frac{r(r + 1)}{2} \leq n$$

A fortiori, $r \leq \sqrt{2n}$. 
This justifies restricting the rank

\[
\min_X \text{Tr}(AX) \quad \text{s.t.} \quad \text{diag}(X) = 1, X \succeq 0, \text{rank}(X) \leq p
\]

Parameterize as \( X = YY^T \) with \( Y \) of size \( n \times p \):

\[
\min_{Y:n \times p} \text{Tr}(AYY^T) \quad \text{s.t.} \quad \text{diag}(YY^T) = 1
\]

Lower dimension and no conic constraint!
Burer & Monteiro ‘03, ‘05, Journée, Bach, Absil, Sepulchre ’10

But nonconvex...
Key feature: search space is smooth

\[
\min_{Y:n\times p} \text{Tr}(AYY^T) \quad \text{s.t.} \quad \text{diag}(YY^T) = 1
\]

Constraints → rows of \( Y \) have unit norm.

The search space is a product of spheres: smooth cost function on a smooth manifold.
Our main result for Max-Cut

\[
\min_{Y: n \times p} \text{Tr}(AYY^T) \quad \text{s.t.} \quad \text{diag}(YY^T) = 1
\]

If \( \frac{p(p+1)}{2} > n \), for almost all \( A \), all sop’s are optimal.

If \( p > n/2 \), for all \( A \), all sop’s are optimal.

sop: second-order critical point (zero gradient, psd Hessian)
Main proof ingredients

1. $X = YY^T$ is optimal iff

$$S = S(Y) = A - \text{ddiag}(AYY^T) \succeq 0$$

For all feasible $\hat{X}$,

$$0 \leq \text{Tr}(S\hat{X}) = \text{Tr}(A\hat{X}) - \text{Tr}(\text{ddiag}(AYY^T)\hat{X}) = \text{Tr}(A\hat{X}) - \text{Tr}(AYY^T).$$

2. If $Y$ is sop and rank deficient, $S(Y) \succeq 0$

3. For almost all $A$, all critical points are rank deficient (if $\frac{p(p+1)}{2} > n$).
Main result for smooth SDP’s

\[
\begin{align*}
\min_{X:n \times n} \quad & \text{Tr}(AX) \quad \text{s.t.} \quad \text{Lin}(X) = b, X \succeq 0 \\
\min_{Y:n \times p} \quad & \text{Tr}(AYY^T) \quad \text{s.t.} \quad \text{Lin}(YY^T) = b
\end{align*}
\]

If the search space in \( X \) is compact and the search space in \( Y \) is a manifold, and if \( \frac{p(p+1)}{2} > \#\text{constraints} \), then, for almost all \( A \), all sop’s are optimal.
Why the manifold assumption?

What can we compute?
→ KKT points.

When are KKT conditions necessary at $Y$?
→ When constraint qualifications hold at $Y$.

What if CQ’s hold at all $Y$’s?
→ Set of $Y$’s is a smooth manifold.
Covers a range of applications

Max-Cut
$\mathbb{Z}_2$-synchronization
Community detection in stochastic block model
Matrix cut norm
Phase-Cut for phase retrieval
Phase synchronization
Orthogonal-Cut (synchronization of rotations)
...

Optimization on manifolds
Not harder (nor easier) than unconstrained optimization
Optimization on many manifolds

Spheres, orthonormal bases (Stiefel), rotations, positive definite matrices, fixed-rank matrices, Euclidean distance matrices, semidefinite fixed-rank matrices, shapes, linear subspaces (Grassmann), phases, essential matrices, special Euclidean group, fixed-rank tensors, Euclidean spaces...

Products and quotients of all of these, real and complex versions...
Taking a close look at gradient descent
We need Riemannian geometry

At each point $x$ in the search space $M$

We linearize $M$ into a tangent space $T_xM$

And pick a metric on $T_xM$.

This gives intrinsic notions of gradient and Hessian.
An excellent book
Optimization algorithms on matrix manifolds

A Matlab toolbox

www.manopt.org

Welcome to Manopt!
A Matlab toolbox for optimization on manifolds

Optimization on manifolds is a powerful paradigm to address nonlinear optimization problems with various types of constraints that arise naturally in applications, such as orthonormality or low rank. 

Download  Get started
Example: Max-Cut relaxation

$$\min_{Y:n \times p} \text{Tr}(AYY^T) \text{ s.t. } \text{diag}(YY^T) = 1$$

Rows of $Y$ have unit norm: product of spheres.

Tangent space: $\{\dot{Y}: \text{diag}(\dot{YY}^T + YY^T) = 0\}$

Gradient: project $2AY$ to tangent space.

Retraction: normalize rows of $Y + \dot{Y}$. 
function Y = maxcut_manopt(A)

% Select an appropriate relaxation rank p.
  n = size(A, 1);
  p = ceil(sqrt(2*n));

% Select the manifold to optimize over.
  problem.M = obliquefactory(p, n, true);

% Define the cost function to be minimized.
  problem.cost = @(Y) sum(sum(Y.*(A*Y)));
  problem.egrad = @(Y) 2*(A*Y);
  problem.ehess = @(Y, Ydot) 2*(A*Ydot);

% Call a standard solver
% (random initialization, default parameters.)
  Y = truststregions(problem);

end

\[
\min_{Y:n \times p} \text{Tr} \left( AYY^T \right) \text{ s.t. } \text{diag} \left( YY^T \right) = 1
\]
\[ Y = \text{maxcut\_manopt}(A); \]

\[
\begin{array}{llllll}
\text{f:} & -1.189330e+01 & |\text{grad}|: & 3.969772e+02 \\
\text{acc TR+ k:} & 1 & \text{num\_inner:} & 1 & \text{f:} & -5.933834e+03 & |\text{grad}|: & 3.214287e+02 \\
\text{acc k:} & 2 & \text{num\_inner:} & 1 & \text{f:} & -1.092386e+04 & |\text{grad}|: & 2.744089e+02 \\
\text{acc k:} & 3 & \text{num\_inner:} & 3 & \text{f:} & -1.344741e+04 & |\text{grad}|: & 2.542660e+02 \\
\text{acc k:} & 4 & \text{num\_inner:} & 3 & \text{f:} & -1.541521e+04 & |\text{grad}|: & 1.351628e+02 \\
\text{acc k:} & 5 & \text{num\_inner:} & 5 & \text{f:} & -1.616969e+04 & |\text{grad}|: & 7.579978e+01 \\
\text{acc k:} & 6 & \text{num\_inner:} & 10 & \text{f:} & -1.641459e+04 & |\text{grad}|: & 4.638172e+01 \\
\text{REJ TR- k:} & 7 & \text{num\_inner:} & 20 & \text{f:} & -1.641459e+04 & |\text{grad}|: & 4.638172e+01 \\
\text{acc TR+ k:} & 8 & \text{num\_inner:} & 6 & \text{f:} & -1.654937e+04 & |\text{grad}|: & 1.057115e+01 \\
\text{acc k:} & 9 & \text{num\_inner:} & 25 & \text{f:} & -1.656245e+04 & |\text{grad}|: & 3.576517e+00 \\
\text{acc k:} & 10 & \text{num\_inner:} & 18 & \text{f:} & -1.656370e+04 & |\text{grad}|: & 3.951183e-01 \\
\text{acc k:} & 11 & \text{num\_inner:} & 43 & \text{f:} & -1.656377e+04 & |\text{grad}|: & 1.330375e-01 \\
\text{acc k:} & 12 & \text{num\_inner:} & 48 & \text{f:} & -1.656378e+04 & |\text{grad}|: & 5.752944e-02 \\
\text{acc k:} & 13 & \text{num\_inner:} & 67 & \text{f:} & -1.656378e+04 & |\text{grad}|: & 2.430253e-02 \\
\text{acc k:} & 14 & \text{num\_inner:} & 89 & \text{f:} & -1.656378e+04 & |\text{grad}|: & 2.475079e-03 \\
\text{acc k:} & 15 & \text{num\_inner:} & 123 & \text{f:} & -1.656378e+04 & |\text{grad}|: & 1.896680e-05 \\
\text{acc k:} & 16 & \text{num\_inner:} & 224 & \text{f:} & -1.656378e+04 & |\text{grad}|: & 1.103767e-09 \\
\end{array}
\]

Gradient norm tolerance reached; \( \text{options.tolgradnorm} = 1e-06 \).
Total time is 5.14 [s]

Optimality gap: \( n \cdot \lambda_{\min}(S(Y)) = -4.2 \cdot 10^{-6} \)
Convergence guarantees for Riemannian gradient descent

Global convergence to critical points.

Linear convergence rate locally.

Reach $\|\nabla f(x)\| \leq \varepsilon$ in $O\left(\frac{1}{\varepsilon^2}\right)$ iterations under Lipschitz assumptions.

With Cartis & Absil (arXiv 1605.08101).
Convergence guarantees for Riemannian trust regions

Global convergence to second-order critical points.

Quadratic convergence rate locally.

\[ \| \nabla f(x) \| \leq \varepsilon \text{ and } \text{Hess} f(x) \succeq -\varepsilon I \text{ in } O\left(\frac{1}{\varepsilon^3}\right) \]

iterations under Lipschitz assumptions.

With Cartis & Absil (arXiv 1605.08101).
Low-rank matrix completion
Gaussian mixture models

Matrix Manifold Optimization for Gaussian Mixture Models

Reshad Hosseini, Suvrit Sra, 2015 (NIPS)

\[ p(x) := \sum_{j=1}^{K} \alpha_j p_N(x; \mu_j, \Sigma_j), \quad x \in \mathbb{R}^d, \]

and where \( p_N \) is a (multivariate) Gaussian with mean \( \mu \in \mathbb{R}^d \) and covariance \( \Sigma > 0 \). That is,

\[ p_N(x; \mu, \Sigma) := \det(\Sigma)^{-1/2} (2\pi)^{-d/2} \exp\left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right). \]

Given i.i.d. samples \( \{x_1, \ldots, x_n\} \), we wish to estimate \( \{\hat{\mu}_j \in \mathbb{R}^d, \hat{\Sigma}_j > 0\}_{j=1}^{K} \) and weights \( \hat{\alpha} \in \Delta_K \), the \( K \)-dimensional probability simplex. This leads to the GMM optimization problem

\[ \max_{\alpha \in \Delta_K, \{\mu_j, \Sigma_j > 0\}_{j=1}^{K}} \sum_{i=1}^{n} \log \left( \sum_{j=1}^{K} \alpha_j p_N(x_i; \mu_j, \Sigma_j) \right). \]
Dictionary learning

Complete Dictionary Recovery over the Sphere I: Overview and the Geometric Picture

Ju Sun, Student Member, IEEE, Qing Qu, Student Member, IEEE, and John Wright, Member, IEEE

Abstract

We consider the problem of recovering a complete (i.e., square and invertible) matrix $A_0$, from $Y \in \mathbb{R}^{n \times p}$ with $Y = A_0 X_0$, provided $X_0$ is sufficiently sparse. This recovery problem is central to the theoretical understanding of dictionary learning, which seeks a sparse representation for a collection of input signals, and finds numerous applications in modern signal processing. We recover $A_0$ when $X_0$ satisfies suitable conditions, with results based on efficient convex optimization for any constant $\delta \in (0, 1)$.

Our algorithmic perspective, without an explicit sparsity constraint, and hence is potentially applicable to a wider range of problem settings, provides a unified framework that sheds light on the tractability of various optimization problems. This framework combines methodological and theoretical insights to design a Riemannian optimization method that can be applied to a wide range of problems, in particular those arising in dictionary learning for strongly convex objectives. Our method is efficient and provably convergent, and we provide empirical evidence through simulations to support our theoretical results.

This paper provides a unified framework for the geometric analysis of dictionary learning objective landscapes. In the following sections, we present algorithms for recovering $A_0$.

Fig. 2: Why is dictionary learning over $S^{n-1}$ tractable? Assume the target dictionary $A_0$ is orthogonal. Left: Large sample objective function $E_{X_0} [f(q)]$. The only local minimizers are the columns of $A_0$ and their negatives. Center: The same function, visualized as a height above the plane $a_1^\perp$ ($a_1$ is the first column of $A_0$, and is also a global minimizer). Right: Around $a_1$, the function exhibits a small region of positive curvature, a region of large gradient, and finally a region in which the direction away from $a_1$ is a direction of negative curvature.
Phase retrieval

A Geometric Analysis of Phase Retrieval

Ju Sun, Qing Qu, and John Wright
{js4038, qq2105, jw2966}@columbia.edu
Department of Electrical Engineering, Columbia University, New York, USA
January 31, 2016

Abstract

Can we recover a complex signal from its Fourier magnitudes? More generally, given a set of $m$ measurements, $y_k = |a_k^*x|$ for $k = 1, \ldots, m$, is it possible to recover $x \in \mathbb{C}^n$ (i.e., length-$n$ complex vector)? This is the generalized cross correlation task in various disciplines. Natural nontrivial measurements in practice, but lack clear theoretical explanation in this gap. We show that when the measurements are sparse and the number of measurements is large enough, there is a natural least-squares formulation for the problem. There are no spurious local minimizers, assuming that (1) there are equivalent copies; and (2) the objective function has a unique minimizer without special initialization. This structure allows a number of optimization algorithms to find a local second-order trust-region algorithm.

Keywords. Phase retrieval, Nonconvex optimization, Rigorous rigor, Ridable saddles, Trust-region methods

Figure 5: Function landscape of (1.1) for $x = [1; 0]$ and $m \to \infty$ for the masked Fourier transform measurements (coded diffraction model [CLS15b]). Compared to the landscape under the Gaussian model (Figure 2), the landscape here has an analogous shape qualitatively. The benign geometric structure is evident.
Phase synchronization

Nonconvex phase synchronization

Nicolas Boumal*

March 29, 2016

Abstract

We estimate $n$ phases (angles) from noisy pairwise relative phase measurements. The task is modeled as a nonconvex least-squares optimization problem. It was recently shown that this problem can be solved in polynomial time via convex relaxation, under some conditions on the noise. In this paper, under similar but more restrictive conditions, we show that a modified version of the power method converges to the global optimizer. This is simpler and (empirically) faster than convex approaches. Empirically, they both succeed in the same regime. Further analysis shows that, in the same noise regime as previously, second-order necessary optimality conditions for this quadratically constrained quadratic program are also sufficient, despite nonconvexity.

1 Introduction

We consider the problem of estimating $n$ phases (complex numbers with unit modulus) based on noisy measurements of the relative phases. The target parameter is

$$z \in \mathbb{C}^n \triangleq \{x \in \mathbb{C}^n : |x_1| = \cdots = |x_n| = 1\},$$

and the measurements $C_{ij} \approx z_i \bar{z}_j$ are stored in the Hermitian matrix

$$C = zz^* + \Delta,$$

where $\Delta$ is a Hermitian perturbation. Motivated by the scenario where $\Delta$ contains white Gaussian noise, we focus on the associated maximum likelihood estimation problem (it
Synchronization of rotations

Robust estimation of rotations, 2013
B., Singer and Absil
Sensor network localization

Noisy sensor network localization, robust facial reduction and the Pareto frontier

Cheung, Drusvyatskiy, Krislock and Wolkowicz 2014

Figure 2: Illustration of robust facial reduction with refinement applied on an instance with 1000 sensors (no anchors) on a $[-0.5, 0.5]^2$ box, with noise factor 0.05 and radio range 0.1. From left to right: (1) using Algorithm I without refinement (RMSD = 61.52%R); (2) using Algorithm I with refinement via Manopt (RMSD = 1.39%R); (3) using only Manopt (RMSD = 380.59%R). Blue: true location; red: estimated location and discrepancy.
Protein structure determination in NMR spectroscopy

Residual Dipolar Coupling, Protein Backbone Conformation and Semidefinite Programming

Yuehaw Khoo, Amit Singer and David Cowburn, 2016

Figure 1 (a) Example of an articulated structure with joints with indices $J_i$'s (Red dots) and $H_i$'s. The hinges are represented by black bars in the figure. (b) Protein backbone consists of peptide planes and CA bodies. These rigid units are chained together at the bonds (N, CA) and (C, CA).

Figure 4 The trace of protein backbone drawn using N, CA and C. The black, blue and red curves come from the X-ray model IUBQ, RDC-SDP solution and RDC-NOE-SDP respectively.
Nonsmooth with MADMM


Compressed modes

Functional correspondence

Figure 1: The first six compressed modes of a human mesh containing $n = 8K$ points computed using MADMM. Parameter $\mu = 10^{-3}$ and three manifold optimization iterations in $X$-step were used in this experiment.

Figure 4: Examples of correspondences obtained with MADMM (top) and least-squares solution (bottom). Similar colors encode corresponding points. Bottom left: examples of correspondence between a pair of shapes (outliers are shown in red).
Take home message

Optimization on manifolds has many applications and is easy to try with Manopt.

It comes with the same guarantees as unconstrained nonlinear optimization.

For some problems, we get global optimality.
Max-Cut

\( A \) is the adjacency matrix of the graph:

\[
\max_{x_1, \ldots, x_n \in \{\pm 1\}} \sum_{i,j} A_{ij} \frac{1 - x_i x_j}{2}
\]

\[
\max_{x_1, \ldots, x_n \in \{\pm 1\}} 1^T A 1 - x^T A x
\]

\[
\min_{x} x^T A x \quad \text{s.t.} \quad x_i^2 = 1 \quad \forall i
\]
Max-Cut

\[
\min_{x} x^T Ax \quad \text{s.t.} \quad x_i^2 = 1 \quad \forall i
\]

\[
\min_{x} \text{Tr}(Axx^T) \quad \text{s.t.} \quad (xx^T)_{ii} = 1 \quad \forall i
\]

\[
\min_{X} \text{Tr}(AX) \quad \text{s.t.} \quad \text{diag}(X) = 1, \quad X \succeq 0, \quad \text{rank}(X) = 1
\]
This is \textit{not} projected gradients

Optimization on manifolds is \textit{intrinsic}. There is no need for an embedding space. Works for abstract manifolds, \textit{quotient spaces}.