Geometry and symmetry in non-convex optimization

Nicolas Boumal – OPTIM, EPFL
Chair of Continuous Optimization
Institute of Mathematics
March 2021
Modeling with continuous optimization

minimize $f(x)$ for $x$ in some set

Inverse problems, statistical estimation, machine learning...

The problem and its data define $f$ and the set.

Depending on structure, can try various algorithms.

General question in this field: which structures are tractable?
“... in fact, the great watershed in optimization isn't between linearity and nonlinearity, but **convexity** and **non-convexity**.”

R. T. Rockafellar, in SIAM Review, 1993
Convex
Non-convex
Non-convex just means *not* convex.
Non-convexity can be **benign**

This can mean various things. **Theorem templates** are on a spectrum:

> “If \{conditions\}, necessary optimality conditions are sufficient.”

> “If \{conditions\}, we can initialize a specific algorithm well.”

The conditions (often on data) may be generous (e.g., *genericity*) or less so (e.g., *high-probability* event for non-adversarial distribution.)
“... in fact, the great watershed in optimization isn't between linearity and nonlinearity, but **convexity** and **non-convexity**.”

R. T. Rockafellar, in SIAM Review, 1993
Pockets of benign non-convexity: Ju Sun’s list

https://sunju.org/research/nonconvex, ~900 papers in March 2021; categories:

- Matrix Completion/Sensing
- Tensor Recovery/Decomposition & Hidden Variable Models
- Phase Retrieval
- Dictionary Learning
- Deep Learning
- Sparse Vectors in Linear Subspaces
- Nonnegative/Sparse Principal Component Analysis
- Mixed Linear Regression
- Blind Deconvolution/Calibration
- Super Resolution
- Synchronization Problems
- Community Detection
- Joint Alignment
- Numerical Linear Algebra
- Bayesian Inference
- Empirical Risk Minimization & Shallow Networks
- System Identification
- Burer-Monteiro Style Decomposition Algorithms
- Generic Structured Problems
- Nonconvex Feasibility Problems
- Separable Nonnegative Factorization (NMF)
Common thread I: geometry

We go over a first set of examples with benign non-convexity.

Their search spaces are smooth manifolds.

The manifolds are not convex, but the problems are fine.
Pockets of benign non-convexity: Ju Sun’s list

https://sunju.org/research/nonconvex, ~900 papers in March 2021; categories:

Matrix Completion/Sensing
Tensor Recovery/Decomposition & Hidden Variable Models
Phase Retrieval
Dictionary Learning
Deep Learning
Sparse Vectors in Linear Subspaces
Nonnegative/Sparse
Principal Component Analysis
Mixed Linear Regression
Blind Deconvolution/Calibration
Super Resolution

Synchronization Problems
Community Detection
Joint Alignment
Numerical Linear Algebra
Bayesian Inference
Empirical Risk Minimization & Shallow Networks
System Identification
Burer-Monteiro Style Decomposition Algorithms
Generic Structured Problems
Nonconvex Feasibility Problems
Separable Nonnegative Factorization (NMF)
Principal component analysis

Two views of subspace analysis:

• Find the subspace $S$ of dimension $d$ which best fits the data points. This is optimization over the Grassmann manifold.

• Find the matrix of rank $d$ closest to the data matrix. This is optimization over the manifold of rank-$d$ matrices.

Non-convexity is entirely benign: second-order points are optimal.

Picture: https://365datascience.com/tutorials/python-tutorials/principal-components-analysis/
Low-rank matrix completion

Each row is a user. Each column is a movie. Each entry in the matrix is the rating of that movie by that user. Most entries are unknown. Complete the matrix. Common assumption: the matrix has (approximately) rank $d$.

Find the best-fit matrix over the manifold of rank-$d$ matrices.

Assuming incoherence and sufficient samples, benign with high prob. (second-order points are optimal.)

See for example Keshavan, Montanari & Oh, IEEE Inf. Th.’10 and Ge, Jin, Zheng, ICML’17
Community detection: relax a bit

Two communities.
Nodes in the same community connected with higher probability.
Given the graph, find the communities.

Combinatorial: \{±1\}.
Can relax to optimize over circles.

In some regimes, benign in that second-order points yield good estimators (after rounding) with high probability.

See for example Abbé, Bandeira & Hall IEEE Inf. Th.’17, Bandeira, Boumal, Voroninski COLT’16
Synchronization of rotations

Optim. over rotation matrices and orthonormal frames (Stiefel).
Also benign (similar ingredients).

Common thread II: symmetry

Symmetry can preclude convexity, without necessarily making the problem hard.

Stated differently:
Symmetry complexifies the landscape, not necessarily the problem.
Pockets of benign non-convexity: Ju Sun’s list

https://sunju.org/research/nonconvex, ~900 papers in March 2021; categories:

Matrix Completion/Sensing
Tensor Recovery/Decomposition & Hidden Variable Models
Phase Retrieval
Dictionary Learning
Deep Learning *(don’t get too excited)*
Sparse Vectors in Linear Subspaces
Nonnegative/Sparse Principal Component Analysis
Mixed Linear Regression
Blind Deconvolution/Calibration
Super Resolution

Synchronization Problems
Community Detection
Joint Alignment
Numerical Linear Algebra
Bayesian Inference
Empirical Risk Minimization & Shallow Networks
System Identification
Burer-Monteiro Style Decomposition Algorithms
Generic Structured Problems
Nonconvex Feasibility Problems
Separable Nonnegative Factorization (NMF)
Orthogonal dictionary learning

Goal: find an orthonormal basis $Q = [q_1, ..., q_n]$ such that a dataset $X$ admits a sparse representation.

Modeled as: $\min_{Q,A} \frac{1}{2} \|QA - X\|_F^2 + \lambda \|A\|_1$ for $Q$ in orthogonal group.

Discrete symmetry: can permute columns of $Q$ and change signs.

Benign w.h.p. in a planted model (second-order points optimal).

See for example Sun, Qu & Wright IEEE Inf. Th.’17 and Zhang, Qu, Wright, arXiv:2007.06753
Linear neural networks (not the real deal)

Layer weights as matrices $W_1, \ldots, W_L$.
Loss: $f(W_1, \ldots, W_L) = \|W_L \cdots W_1 X - Y\|^2$

Three observations:
1. The map $\varphi(W_1, \ldots, W_L) = W_L \cdots W_1$ is onto bounded rank matrices.
2. The loss is just $f = g \circ \varphi$ with $g(W) = \|WX - Y\|^2$ (convex).
3. Minimizing $g$ over the bounded rank matrices is easy for whitened data.

The symmetry induced by $\varphi$ breaks the convexity of $g$, but the optimization problem is still benign (some technicalities).

See for example Arora, Cohen, Golowich & Hu ICLR’19 and Bah, Rauhut, Terstiege, Westdickenberg ‘19
Picture: https://www.jeremyjordan.me/autoencoders/
This brings us to the algorithmic questions

Given $f : \mathcal{M} \to \mathbb{R}$ on a smooth manifold,

How do we compute a second-order stationary point?

That’s where Riemann comes in.
The Riemannian structure gives us gradients and Hessians

The essential tools of smooth optimization are defined generally on Riemannian manifolds.

→Riemannian optimization.
A far-reaching extension of unconstrained optim.

First ideas from the ’70s.
First practical in the ’90s via bridge with NLA.
These tools are quite efficient numerically

Feel free to reach out to discuss possible applications in your field!

If you are optimizing/estimating rotations/orientations, rigid motions, subspaces, positive definite matrices, low-rank matrices/tensors, points on circles/spheres, complex phases, orthonormal frames, ...
Back in Göttingen...

If Riemann didn’t invent his geometry to pick Netflix movies, then why did he?

His motivation was to extend Gauss’ work (his advisor), to understand *curvature* in spaces or arbitrary dimension.

We still do not understand the effect curvature may or may not have in optimization. A story for another time.
Welcome to Manopt!

Toolboxes for optimization on manifolds and matrices

Optimization on manifolds is a powerful paradigm to address nonlinear optimization problems. With Manopt, it is easy to deal with various types of constraints and symmetries which arise naturally in applications, such as orthonormality, low rank, positivity and invariance under group actions. These tools are also perfectly suited for unconstrained optimization with vectors and matrices.

With Bamdev Mishra, P.-A. Absil & R. Sepulchre

Lead by J. Townsend, N. Koep & S. Weichwald

Lead by Ronny Bergmann

MATH-512 @ EPFL