Aspects of Convex, Nonconvex, and Geometric Optimization
(Lecture 3)

Suvrit Sra

Massachusetts Institute of Technology

Hausdorff Institute for Mathematics (HIM)
Trimester: Mathematics of Signal Processing
January 2016

Outline

– Convex analysis, optimality
– First-order methods
– Proximal methods, operator splitting
– Stochastic optimization, incremental methods
– Nonconvex models, algorithms
– Geometric optimization

Nonconvex problems

- SVD, PCA
- Other eigenvalue problems
- Matrix & tensor factorization, clustering
- Deep neural networks
- Topic models, Bayesian nonparametrics
- Probabilistic mixture models
- Combinatorial optimization
- Linear, nonlinear mixed integer programming
- Optimization on manifolds
- Optimization in metric spaces
- ...

Introduction

Nonlinear program

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0, \quad i = 1, \ldots, m. \end{aligned}$$

Introduction

Nonlinear program

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0, \quad i = 1, \ldots, m. \end{aligned}$$

Claim: If f and g_i are convex, then under some "constraint qualifications" (e.g., there exists an x for which $g_i(x) < 0$ holds), *necessary and sufficient* conditions characterizing global optimality are known (e.g., *Karush-Kuhn-Tucker*)

Introduction

Nonlinear program

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & g_i(x) \leq 0, \quad i = 1, \ldots, m. \end{array}$$

$0 \in \partial f(x^*)$ necessary and sufficient ($m = 0$, cvx)

Introduction

Nonlinear program

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0, \quad i = 1, \ldots, m. \end{aligned}$$

Nonconvex: Under some constraint qualification, *necessary* conditions known. But **no known** simple conditions that are both necessary and sufficient.

Introduction

Nonlinear program

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0, \quad i = 1, \ldots, m. \end{aligned}$$

- ♠ Is alleged solution a local min? – often skipped question
- ♠ **Myth:** Algorithms converge to global minima for convex local minima for nonconvex

NP-Hardness of nonconvex opt.

Recall **subset-sum** – well-known NP-Complete problem

Given a set of integers $\{a_1, \ldots, a_n\}$, is there a non-empty subset whose sum is zero?

NP-Hardness of nonconvex opt.

Recall **subset-sum** – well-known NP-Complete problem

> Given a set of integers $\{a_1, \ldots, a_n\}$, is there a non-empty subset whose sum is zero?

In other words, is there a solution z to

$$\sum_i a_i z_i = 0 \quad z_i \in \{0, 1\} \quad \text{for } i = 1, \ldots, n.$$

NP-Hardness of nonconvex opt.

Recall **subset-sum** – well-known NP-Complete problem

> Given a set of integers $\{a_1, \ldots, a_n\}$, is there a non-empty subset whose sum is zero?

In other words, is there a solution z to

$$\sum_i a_i z_i = 0 \quad z_i \in \{0, 1\} \quad \text{for } i = 1, \ldots, n.$$

Optimization version

$$\begin{aligned} \min \quad & \Big(\sum_i a_i z_i\Big)^2 + \sum_i z_i(1 - z_i) \\ \text{s.t.} \quad & 0 \leq z_i \leq 1, \quad i = 1, \ldots, n. \end{aligned}$$

Subset-sum has feasible solution, **iff** global min objval is zero.
But subset-sum is NP-Complete; so above problem also NPC.

Nonconvex quadratic optimization

Let A be a symmetric matrix (not necessarily positive definite).

$$\min \quad x^T A x \quad \text{s.t. } x \geq 0.$$

Is $x = 0$ not a local minimum?

Nonconvex quadratic optimization

Let A be a symmetric matrix (not necessarily positive definite).

$$\min \quad x^T A x \quad \text{s.t. } x \geq 0.$$

Is $x = 0$ not a local minimum?

This is NP-Hard!

Generally, even for unconstrained nonconvex problems testing local minimality or objective boundedness (below) are NP-Hard.

In "convex" words

Copositive cone

Def. Let $CP_n := \{A \in \mathbb{S}^{n\times n} \mid x^TAx \geq 0,\ \forall x \geq 0\}$.

Exercise: Verify that CP_n is a convex cone.

- Testing membership in CP_n is co-NP complete.
  (Deciding whether given matrix is **not** copositive is NP-complete.)
- Copositive cone programming: NP-Hard

Exercise: Verify that the following matrix is copositive:

$$A := \begin{bmatrix} 1 & -1 & 1 & 1 & -1 \\ -1 & 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & 1 & -1 \\ -1 & 1 & 1 & -1 & 1 \end{bmatrix}.$$

Solvable nonconvex QP

Let A be a symmetric matrix; b some vector.

$$\min \quad x^T A x + 2b^T x, \quad \text{s.t. } x^T x \leq 1.$$

Solvable nonconvex QP

Let A be a symmetric matrix; b some vector.

$$\min \quad x^T A x + 2b^T x, \quad \text{s.t. } x^T x \leq 1.$$

When $A \not\succeq 0$, above problem is nonconvex.

Also known as, *trust-region subproblem* (TRS).

Solvable nonconvex QP

Let A be a symmetric matrix; b some vector.

$$\min \quad x^T A x + 2b^T x, \quad \text{s.t. } x^T x \leq 1.$$

When $A \not\succeq 0$, above problem is nonconvex.

Also known as, *trust-region subproblem* (TRS).

$$\text{Lagrangian} \qquad L(x,\theta) = x^T A x + 2b^T x + \theta(x^T x - 1)$$
$$L(x,\theta) = x^T (A + \theta I)x + 2b^T x - \theta.$$

Solvable nonconvex QP

Let A be a symmetric matrix; b some vector.

$$\min \quad x^T Ax + 2b^T x, \quad \text{s.t. } x^T x \leq 1.$$

When $A \not\succeq 0$, above problem is nonconvex.

Also known as, *trust-region subproblem* (TRS).

$$\text{Lagrangian} \qquad L(x,\theta) = x^T Ax + 2b^T x + \theta(x^T x - 1)$$
$$L(x,\theta) = x^T(A + \theta I)x + 2b^T x - \theta.$$

If $b \notin \mathcal{R}(A + \theta I)$, then we can choose an $x \in \mathcal{N}(A + \theta I)$ that drives $\inf_x L(x,\theta)$ to $-\infty$.

Solvable nonconvex QP

Let A be a symmetric matrix; b some vector.

$$\min \quad x^T A x + 2b^T x, \quad \text{s.t. } x^T x \leq 1.$$

When $A \not\succeq 0$, above problem is nonconvex.

Also known as, *trust-region subproblem* (TRS).

$$\text{Lagrangian} \qquad \begin{aligned} L(x,\theta) &= x^T A x + 2b^T x + \theta(x^T x - 1) \\ L(x,\theta) &= x^T (A + \theta I) x + 2b^T x - \theta. \end{aligned}$$

If $b \notin \mathcal{R}(A + \theta I)$, then we can choose an $x \in \mathcal{N}(A + \theta I)$ that drives $\inf_x L(x, \theta)$ to $-\infty$.

$$g(\theta) := \begin{cases} -b^T (A + \theta I)^\dagger b - \theta & A + \theta I \succeq 0, \quad b \in \mathcal{R}(A + \theta I) \\ -\infty & \text{otherwise.} \end{cases}$$

A nice nonconvex problem

Dual optimization problem

$$
\begin{aligned}
\max \quad & -b^T(A+\theta I)^{\dagger} b - \theta \\
\text{s.t.} \quad & A+\theta I \succeq 0,\ b \in \mathcal{R}(A+\theta I).
\end{aligned}
$$

A nice nonconvex problem

Dual optimization problem

$$\begin{aligned} \max \quad & -b^T(A+\theta I)^\dagger b - \theta \\ \text{s.t.} \quad & A+\theta I \succeq 0,\; b \in \mathcal{R}(A+\theta I). \end{aligned}$$

Consider eigendecomposition of $A = U\Lambda U^T$. Then,

$$(A+\theta I)^\dagger = U\,\mathrm{Diag}(1+\lambda_i)^{-1}U^T.$$

A nice nonconvex problem

Dual optimization problem

$$\begin{aligned} \max \quad & -b^T(A+\theta I)^\dagger b - \theta \\ \text{s.t.} \quad & A+\theta I \succeq 0,\ b \in \mathcal{R}(A+\theta I). \end{aligned}$$

Consider eigendecomposition of $A = U\Lambda U^T$. Then,

$$(A+\theta I)^\dagger = U\,\mathrm{Diag}(1+\lambda_i)^{-1}U^T.$$

Thus, above problem can be written as

$$\begin{aligned} \max \quad & -\sum\nolimits_{i=1}^{n} \frac{(u_i^T b)^2}{\lambda_i+\theta} - \theta \\ \text{s.t.} \quad & \theta \geq -\lambda_{min}(A). \end{aligned}$$

🙂 Convex optimization problem!

Matrix Factorization

The SVD

Singular Value Decomposition

Theorem SVD (Thm. 2.5.2 [GoLo96]). Let $A \in \mathbb{R}^{m\times n}$. There exist *orthogonal* matrices U and V

$$U^T AV = \mathrm{Diag}(\sigma_1, \ldots, \sigma_p), \quad p = \min(m, n),$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$.

Truncated SVD

Theorem Let A have the SVD $U\Sigma V^T$. If $k < \text{rank}(A)$ and

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T, \qquad \text{then,}$$

$$\|A - A_k\|_2 \leq \|A - B\|_2, \quad \text{s.t. } \text{rank}(B) \leq k$$
$$\|A - A_k\|_F \leq \|A - B\|_F, \quad \text{s.t. } \text{rank}(B) \leq k.$$

Truncated SVD

Theorem Let A have the SVD $U\Sigma V^T$. If $k < \text{rank}(A)$ and

$$A_k = \sum\nolimits_{i=1}^{k} \sigma_i u_i v_i^T, \qquad \text{then,}$$

$$\|A - A_k\|_2 \le \|A - B\|_2, \;\; \text{s.t. } \text{rank}(B) \le k$$
$$\|A - A_k\|_F \le \|A - B\|_F, \;\; \text{s.t. } \text{rank}(B) \le k.$$

SVD gives **globally optimal** solution to the nonconvex problem

$$\min \|X - A\|_F, \quad \text{s.t.} \quad \text{rank}(X) \le k.$$

Truncated SVD – proof

Prove: TSVD yields "best" rank-k approximation to matrix A

Proof.

1. First verify that $\|A - A_k\|_2 = \sigma_{k+1}$
2. Let B be any rank-k matrix
3. Prove that $\|A - B\|_2 \geq \sigma_{k+1}$

Since $\text{rank}(B) = k$, there are $n - k$ vectors that span the null-space $\mathcal{N}(B)$. But $\mathcal{N}(B) \cap V_{k+1} \neq \{0\}$ (??), so we can pick a unit-norm vector $z \in \mathcal{N}(B) \cap V_{k+1}$. Now $\boldsymbol{Bz} = 0$, so

$$\|A - B\|_2^2 \geq \|(A - B)z\|_2^2 = \|Az\|_2^2 = \sum_i^{k+1} \sigma_i^2 (v_i^T z)^2 \geq \sigma_{k+1}^2$$

We used: $\|\boldsymbol{Az}\|_2 \leq \|A\|_2 \|z\|_2$

Nonnegative matrix factorization

Say we want a *low-rank approximation* $A \approx BC$

Nonnegative matrix factorization

Say we want a *low-rank approximation* $A \approx BC$

- SVD yields dense B and C
- B and C contain negative entries, even if $A \geq 0$
- SVD factors may not be that easy to interpret

Nonnegative matrix factorization

Say we want a *low-rank approximation* $A \approx BC$

- SVD yields dense B and C
- B and C contain negative entries, even if $A \geq 0$
- SVD factors may not be that easy to interpret

NMF imposes $B \geq 0$, $C \geq 0$

Algorithms

$$A \approx BC \quad \text{s.t. } B, C \geq 0$$

Least-squares NMF

$$\min \quad \tfrac{1}{2}\|A - BC\|_F^2 \quad \text{s.t. } B, C \geq 0.$$

KL-Divergence NMF

$$\min \quad \sum\nolimits_{ij} a_{ij} \log \frac{(BC)_{ij}}{a_{ij}} - a_{ij} + (BC)_{ij} \quad \text{s.t. } B, C \geq 0.$$

Algorithms

$$A \approx BC \quad \text{s.t. } B, C \geq 0$$

Least-squares NMF

$$\min \quad \tfrac{1}{2}\|A - BC\|_F^2 \quad \text{s.t. } B, C \geq 0.$$

KL-Divergence NMF

$$\min \quad \sum\nolimits_{ij} a_{ij} \log \frac{(BC)_{ij}}{a_{ij}} - a_{ij} + (BC)_{ij} \quad \text{s.t. } B, C \geq 0.$$

- ♣ NP-Hard (Vavasis 2007) – no surprise
- ♣ Recently, Arora et al. showed that if the matrix A has a special "separable" structure, then actually globally optimal NMF is approximately solvable. More recent progress too!
- ♣ We look at only basic methods in this lecture

NMF Algorithms

- Hack: Compute TSVD; “zero-out” negative entries
- Alternating minimization (AM)
- Majorize-Minimize (MM)
- Global optimization (not covered)
- Incremental gradient algorithms

Alternating Descent

$$\min \quad F(B, C)$$

Alternating Descent

1. Initialize B^0, $k \leftarrow 0$
2. Compute C^{k+1} s.t. $F(A, B^k C^{k+1}) \le F(A, B^k C^k)$
3. Compute B^{k+1} s.t. $F(A, B^{k+1} C^{k+1}) \le F(A, B^k C^{k+1})$
4. $k \leftarrow k + 1$, and repeat until stopping criteria met.

Alternating Minimization

Alternating Least Squares

$$C = \underset{C}{\operatorname{argmin}} \quad \|A - B^k C\|_F^2;$$

Alternating Minimization

Alternating Least Squares

$$C = \underset{C}{\operatorname{argmin}} \quad \|A - B^k C\|_F^2; \qquad C^{k+1} \leftarrow \max(0, C)$$

Alternating Minimization

Alternating Least Squares

$$C = \operatorname*{argmin}_{C} \quad \|A - B^k C\|_F^2; \qquad C^{k+1} \leftarrow \max(0, C)$$

$$B = \operatorname*{argmin}_{B} \quad \|A - BC^{k+1}\|_F^2; \qquad B^{k+1} \leftarrow \max(0, B)$$

Alternating Minimization

Alternating Least Squares

$$C = \underset{C}{\operatorname{argmin}} \quad \|A - B^k C\|_F^2; \qquad C^{k+1} \leftarrow \max(0, C)$$

$$B = \underset{B}{\operatorname{argmin}} \quad \|A - BC^{k+1}\|_F^2; \qquad B^{k+1} \leftarrow \max(0, B)$$

ALS is fast, simple, often effective, but ...

Alternating Minimization

Alternating Least Squares

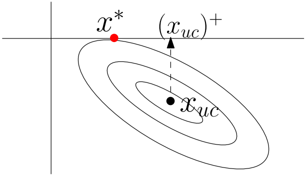
$$C = \underset{C}{\operatorname{argmin}} \quad \|A - B^k C\|_F^2; \qquad C^{k+1} \leftarrow \max(0, C)$$

$$B = \underset{B}{\operatorname{argmin}} \quad \|A - BC^{k+1}\|_F^2; \qquad B^{k+1} \leftarrow \max(0, B)$$

ALS is fast, simple, often effective, but ...

$$\|A - B^{k+1}C^{k+1}\|_F^2 \le \|A - B^k C^{k+1}\|_F^2 \le \|A - B^k C^k\|_F^2$$

descent **need not** hold



Alternating Minimization: correctly

Use alternating **nonnegative least-squares**

$$C^{k+1} = \operatorname*{argmin}_{C} \quad \|A - B^k C\|_F^2 \quad \text{s.t.} \quad C \geq 0$$

$$B^{k+1} = \operatorname*{argmin}_{B} \quad \|A - BC^{k+1}\|_F^2 \quad \text{s.t.} \quad B \geq 0$$

Advantages: Guaranteed descent. Theory of block-coordinate descent guarantees convergence to *stationary point*.

Disadvantages: more complex; slower than ALS

Convergence

AM / two block CD

$$\min \quad F(\boldsymbol{x}) = F(\boldsymbol{x}_1, \boldsymbol{x}_2) \quad \boldsymbol{x} \in \mathcal{X}_1 \times \mathcal{X}_2.$$

Theorem (Grippo & Sciandrone (2000)). Let F be continuously differentiable, and the sets $\mathcal{X}_1$, $\mathcal{X}_2$ be closed and convex. Assume that the both BCD subproblems have solutions, and that the sequence $\{\boldsymbol{x}^k\}$ has limit points. Then, every limit point of $\{\boldsymbol{x}^k\}$ is stationary.

- No need of **unique solutions** to subproblems
- BCD for 2 blocks aka **Alternating Minimization**

Alternating Proximal Method

$$\min \quad L(\boldsymbol{x}, \boldsymbol{y}) := F(\boldsymbol{x}, \boldsymbol{y}) + G(\boldsymbol{x}) + H(\boldsymbol{y}).$$

Assume: ∇F Lipschitz cont. on bounded subsets of $\mathbb{R}^m \times \mathbb{R}^n$
G: lower semicontinuous on $\mathbb{R}^m$
H: lower semicontinuous on $\mathbb{R}^n$.
Example: $F(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$

Alternating Proximal Method

$$\min \quad L(\boldsymbol{x}, \boldsymbol{y}) := F(\boldsymbol{x}, \boldsymbol{y}) + G(\boldsymbol{x}) + H(\boldsymbol{y}).$$

Assume: ∇F Lipschitz cont. on bounded subsets of $\mathbb{R}^m \times \mathbb{R}^n$
G: lower semicontinuous on $\mathbb{R}^m$
H: lower semicontinuous on $\mathbb{R}^n$.
Example: $F(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$

Alternating Proximal Method

$$\boldsymbol{x}_{k+1} \in \operatorname{argmin}\left\{L(x, y_k) + \tfrac{1}{2}c_k\|x - x_k\|^2\right\}$$
$$\boldsymbol{y}_{k+1} \in \operatorname{argmin}\left\{L(x_{k+1}, y) + \tfrac{1}{2}c_k'\|y - y_k\|^2\right\},$$

here c_k, c_k' are suitable sequences of positive scalars.

[arXiv:0801.1780. Attouch, Bolte, Redont, Soubeyran. *Proximal alternating minimization and projection methods for nonconvex problems.*]

Descent Techniques

Majorize-Minimize (MM)

Consider $F(B, C) = \frac{1}{2}\|A - BC\|_F^2$: convex separately in B and C

We use $F(C)$ to denote function restricted to C.

Since $F(C)$ *separable*, suffices to illustrate for

$$\min_{c \geq 0} \quad f(c) = \frac{1}{2}\|a - Bc\|_2^2$$

Recall, our aim is: find C_{k+1} such that $F(B_k, C_{k+1}) \leq F(B_k, C_k)$

Descent technique

$$\min_{c \geq 0} \quad f(c) = \frac{1}{2}\|a - Bc\|_2^2$$

1. Find a function $g(c, \tilde{c})$ that satisfies:

$$g(c, c) = f(c), \quad \text{for all} \quad c,$$
$$g(c, \tilde{c}) \geq f(c), \quad \text{for all} \quad c, \tilde{c}.$$

Descent technique

$$\min_{c \geq 0} \quad f(c) = \tfrac{1}{2}\|a - Bc\|_2^2$$

1. Find a function $g(c, \tilde{c})$ that satisfies:

$$\begin{aligned} g(c, c) &= f(c), && \text{for all} \quad c, \\ g(c, \tilde{c}) &\geq f(c), && \text{for all} \quad c, \tilde{c}. \end{aligned}$$

2. Compute $c^{t+1} = \operatorname{argmin}_{c \geq 0} g(c, c^t)$

Descent technique

$$\min_{c \geq 0} \quad f(c) = \frac{1}{2}\|a - Bc\|_2^2$$

1. Find a function $g(c, \tilde{c})$ that satisfies:

$$\begin{aligned} g(c, c) &= f(c), \quad \text{for all} \quad c, \\ g(c, \tilde{c}) &\geq f(c), \quad \text{for all} \quad c, \tilde{c}. \end{aligned}$$

2. Compute $c^{t+1} = \operatorname{argmin}_{c \geq 0} g(c, c^t)$
3. Then we have descent

Descent technique

$$\min_{c \geq 0} \quad f(c) = \tfrac{1}{2}\|a - Bc\|_2^2$$

1. Find a function $g(c, \tilde{c})$ that satisfies:

$$\begin{aligned} g(c, c) &= f(c), \quad \text{for all} \quad c, \\ g(c, \tilde{c}) &\geq f(c), \quad \text{for all} \quad c, \tilde{c}. \end{aligned}$$

2. Compute $c^{t+1} = \operatorname{argmin}_{c \geq 0} g(c, c^t)$
3. Then we have descent

$$f(c^{t+1})$$

Descent technique

$$\min_{c \geq 0} \quad f(c) = \tfrac{1}{2}\|a - Bc\|_2^2$$

1. Find a function $g(c, \tilde{c})$ that satisfies:

$$\begin{aligned} g(c, c) &= f(c), \quad \text{for all} \quad c, \\ g(c, \tilde{c}) &\geq f(c), \quad \text{for all} \quad c, \tilde{c}. \end{aligned}$$

2. Compute $c^{t+1} = \operatorname{argmin}_{c \geq 0} g(c, c^t)$
3. Then we have descent

$$f(c^{t+1}) \overset{\text{def}}{\leq} g(c^{t+1}, c^t)$$

Descent technique

$$\min_{c\geq 0} \quad f(c) = \tfrac{1}{2}\|a - Bc\|_2^2$$

1. Find a function $g(c, \tilde{c})$ that satisfies:

$$\begin{aligned} g(c, c) &= f(c), \quad \text{for all} \quad c, \\ g(c, \tilde{c}) &\geq f(c), \quad \text{for all} \quad c, \tilde{c}. \end{aligned}$$

2. Compute $c^{t+1} = \operatorname{argmin}_{c\geq 0} g(c, c^t)$
3. Then we have descent

$$f(c^{t+1}) \overset{\text{def}}{\leq} g(c^{t+1}, c^t) \overset{\text{argmin}}{\leq} g(c^t, c^t)$$

Descent technique

$$\min_{c \geq 0} \quad f(c) = \tfrac{1}{2}\|a - Bc\|_2^2$$

1. Find a function $g(c, \tilde{c})$ that satisfies:

$$\begin{aligned} g(c, c) &= f(c), \quad \text{for all} \quad c, \\ g(c, \tilde{c}) &\geq f(c), \quad \text{for all} \quad c, \tilde{c}. \end{aligned}$$

2. Compute $c^{t+1} = \operatorname{argmin}_{c \geq 0} g(c, c^t)$
3. Then we have descent

$$f(c^{t+1}) \overset{\text{def}}{\leq} g(c^{t+1}, c^t) \overset{\text{argmin}}{\leq} g(c^t, c^t) \overset{\text{def}}{=} f(c^t).$$

Descent technique – constructing g

We exploit that $h(x) = \frac{1}{2}x^2$ is a *convex function*

$$h(\sum_i \lambda_i x_i) \leq \sum_i \lambda_i h(x_i), \text{ where } \lambda_i \geq 0, \sum_i \lambda_i = 1$$

Descent technique – constructing g

We exploit that $h(x) = \frac{1}{2}x^2$ is a *convex function*

$h(\sum_i \lambda_i x_i) \leq \sum_i \lambda_i h(x_i)$, where $\lambda_i \geq 0$, $\sum_i \lambda_i = 1$

$$f(c) = \frac{1}{2}\sum_i (a_i - b_i^T c)^2 =$$

Descent technique – constructing g

We exploit that $h(x) = \frac{1}{2}x^2$ is a *convex function*

$$h(\sum_i \lambda_i x_i) \leq \sum_i \lambda_i h(x_i), \text{ where } \lambda_i \geq 0, \sum_i \lambda_i = 1$$

$$f(c) = \frac{1}{2}\sum_i (a_i - b_i^T c)^2 = \frac{1}{2}\sum_i a_i^2 - 2a_i b_i^T c + (b_i^T c)^2$$

Descent technique – constructing g

We exploit that $h(x) = \frac{1}{2}x^2$ is a *convex function*

$$h(\sum_i \lambda_i x_i) \leq \sum_i \lambda_i h(x_i), \text{ where } \lambda_i \geq 0, \sum_i \lambda_i = 1$$

$$\begin{aligned} f(c) &= \tfrac{1}{2}\sum_i (a_i - b_i^T c)^2 = \tfrac{1}{2}\sum_i a_i^2 - 2a_i b_i^T c + (b_i^T c)^2 \\ &= \tfrac{1}{2}\sum_i a_i^2 - 2a_i b_i^T c + \tfrac{1}{2}\sum_i \Big(\sum_j b_{ij} c_j\Big)^2 \end{aligned}$$

Descent technique – constructing g

We exploit that $h(x) = \frac{1}{2}x^2$ is a *convex function*

$h(\sum_i \lambda_i x_i) \leq \sum_i \lambda_i h(x_i)$, where $\lambda_i \geq 0$, $\sum_i \lambda_i = 1$

$$\begin{aligned} f(c) &= \tfrac{1}{2}\sum\nolimits_i (a_i - b_i^T c)^2 = \tfrac{1}{2}\sum\nolimits_i a_i^2 - 2a_i b_i^T c + (b_i^T c)^2 \\ &= \tfrac{1}{2}\sum\nolimits_i a_i^2 - 2a_i b_i^T c + \tfrac{1}{2}\sum\nolimits_i \big(\sum\nolimits_j b_{ij} c_j\big)^2 \\ &= \tfrac{1}{2}\sum\nolimits_i a_i^2 - 2a_i b_i^T c \end{aligned}$$

Descent technique – constructing g

We exploit that $h(x) = \frac{1}{2}x^2$ is a *convex function*

$h(\sum_i \lambda_i x_i) \le \sum_i \lambda_i h(x_i)$, where $\lambda_i \ge 0, \sum_i \lambda_i = 1$

$$
\begin{aligned}
f(c) &= \tfrac{1}{2}\sum_i (a_i - b_i^T c)^2 = \tfrac{1}{2}\sum_i a_i^2 - 2a_i b_i^T c + (b_i^T c)^2 \\
&= \tfrac{1}{2}\sum_i a_i^2 - 2a_i b_i^T c + \tfrac{1}{2}\sum_i \big(\sum_j b_{ij} c_j\big)^2 \\
&= \tfrac{1}{2}\sum_i a_i^2 - 2a_i b_i^T c + \tfrac{1}{2}\sum_i \big(\sum_j \lambda_{ij} b_{ij} c_j / \lambda_{ij}\big)^2
\end{aligned}
$$

Descent technique – constructing g

We exploit that $h(x) = \frac{1}{2}x^2$ is a *convex function*

$$h(\sum_i \lambda_i x_i) \le \sum_i \lambda_i h(x_i), \text{ where } \lambda_i \ge 0, \sum_i \lambda_i = 1$$

$$\begin{aligned}
f(c) &= \tfrac{1}{2}\sum_i (a_i - b_i^T c)^2 = \tfrac{1}{2}\sum_i a_i^2 - 2a_i b_i^T c + (b_i^T c)^2 \\
&= \tfrac{1}{2}\sum_i a_i^2 - 2a_i b_i^T c + \tfrac{1}{2}\sum_i \big(\sum_j b_{ij} c_j\big)^2 \\
&= \tfrac{1}{2}\sum_i a_i^2 - 2a_i b_i^T c + \tfrac{1}{2}\sum_i \big(\sum_j \lambda_{ij} b_{ij} c_j / \lambda_{ij}\big)^2 \\
&\overset{\text{cvx}}{\le} \tfrac{1}{2}\sum_i a_i^2 - 2a_i b_i^T c + \tfrac{1}{2}\sum_{ij} \lambda_{ij} \big(b_{ij} c_j / \lambda_{ij}\big)^2
\end{aligned}$$

Descent technique – constructing g

We exploit that $h(x) = \frac{1}{2}x^2$ is a *convex function*

$$h(\textstyle\sum_i \lambda_i x_i) \le \sum_i \lambda_i h(x_i), \text{ where } \lambda_i \ge 0, \sum_i \lambda_i = 1$$

$$\begin{aligned}
f(c) &= \tfrac{1}{2}\textstyle\sum_i (a_i - b_i^T c)^2 = \tfrac{1}{2}\sum_i a_i^2 - 2a_i b_i^T c + (b_i^T c)^2 \\
&= \tfrac{1}{2}\textstyle\sum_i a_i^2 - 2a_i b_i^T c + \tfrac{1}{2}\sum_i \big(\sum_j b_{ij} c_j\big)^2 \\
&= \tfrac{1}{2}\textstyle\sum_i a_i^2 - 2a_i b_i^T c + \tfrac{1}{2}\sum_i \big(\sum_j \lambda_{ij} b_{ij} c_j / \lambda_{ij}\big)^2 \\
&\overset{\text{cvx}}{\le} \tfrac{1}{2}\textstyle\sum_i a_i^2 - 2a_i b_i^T c + \tfrac{1}{2}\sum_{ij} \lambda_{ij} \big(b_{ij} c_j / \lambda_{ij}\big)^2 \\
&=: g(c, \tilde{c}), \quad \text{where} \quad \lambda_{ij} \quad \text{are convex coeffts}
\end{aligned}$$

Descent technique – constructing g

$$f(c) = \tfrac{1}{2}\|a - Bc\|_2^2$$
$$g(c, \tilde{c}) = \tfrac{1}{2}\|a\|_2^2 - \sum\nolimits_i a_i b_i^T c + \tfrac{1}{2}\sum\nolimits_{ij} \lambda_{ij}\big(b_{ij}c_j/\lambda_{ij}\big)^2.$$

Only remains to *pick* λ_{ij} as functions of $\tilde{c}$

Descent technique – constructing g

$$f(c) = \tfrac{1}{2}\|a - Bc\|_2^2$$

$$g(c, \tilde{c}) = \tfrac{1}{2}\|a\|_2^2 - \sum_i a_i b_i^T c + \tfrac{1}{2}\sum_{ij} \lambda_{ij}\left(b_{ij}c_j/\lambda_{ij}\right)^2.$$

Only remains to *pick* λ_{ij} as functions of $\tilde{c}$

$$\lambda_{ij} = \frac{b_{ij}\tilde{c}_j}{\sum_k b_{ik}\tilde{c}_k} = \frac{b_{ij}\tilde{c}_j}{b_i^T\tilde{c}}$$

Descent technique – constructing g

$$f(c) = \tfrac{1}{2}\|a - Bc\|_2^2$$
$$g(c, \tilde{c}) = \tfrac{1}{2}\|a\|_2^2 - \sum\nolimits_i a_i b_i^T c + \tfrac{1}{2}\sum\nolimits_{ij} \lambda_{ij}\left(b_{ij}c_j/\lambda_{ij}\right)^2.$$

Only remains to *pick* λ_{ij} as functions of $\tilde{c}$

$$\lambda_{ij} = \frac{b_{ij}\tilde{c}_j}{\sum_k b_{ik}\tilde{c}_k} = \frac{b_{ij}\tilde{c}_j}{b_i^T\tilde{c}}$$

Exercise: Verify that $g(c, c) = f(c)$;
Exercise: Let $f(c) = \sum_i a_i \log(a_i/(Bc)_i) - a_i + (Bc)_i$. Derive an auxiliary function $g(c, \tilde{c})$ for this $f(c)$.

Descent technique – Exercise

Key step

$$c^{t+1} = \operatorname*{argmin}_{c \geq 0} g(c, c^t).$$

Exercise: Solve $\partial g(c, c^t)/\partial c_p = 0$ to obtain

$$c_p = c_p^t \frac{[B^T a]_p}{[B^T B c^t]_p}$$

This yields the "multiplicative update" algorithm of Lee/Seung (1999).

MM algorithms

- We exploited convexity of x^2
- Expectation Maximization (EM) algorithm exploits convexity of $-\log x$
- Other choices possible, e.g., by varying λ_{ij}
- Our technique one variant of repertoire of *Majorization-Minimization* (MM) algorithms
- gradient-descent also an MM algorithm
- Related to *d.c. programming*
- MM algorithms subject of a separate lecture!

Generic descent method

Nonsmooth, nonconvex min

$$\min \quad f(x)$$

Methods that generate (x_k, w_k) such that

$$f(x_{k+1}) + a\|x_{k+1} - x_k\|^2 \leq f(x_k)$$

there exists $w_{k+1} \in \partial f(x_{k+1})$ s.t. $\|x_{k+1} - x_k\| \geq b\|w_{k+1}\|$.

Generic descent method

Nonsmooth, nonconvex min

$$\min \quad f(x)$$

Methods that generate (x_k, w_k) such that

$$f(x_{k+1}) + a\|x_{k+1} - x_k\|^2 \leq f(x_k)$$

there exists $w_{k+1} \in \partial f(x_{k+1})$ s.t. $\|x_{k+1} - x_k\| \geq b\|w_{k+1}\|$.

Condition 1: Sufficient descent from x_k to x_{k+1}
Condition 2: Captures inexactness (approx. optimality)
Example: captures nonconvex proximal gradient method.

[Attouch, Bolte, Svaiter (2011). *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods*. Math Prog.]

Other Alternating methods

- Nonconvex ADMM (e.g., arXiv:1410.1390)
- Nonconvex Douglas-Rachford (e.g., Borwein's webpage!)
- Alternating minimization for global optimization

  e.g., [Jain, Netrapalli, Sanghavi (2013). *Low-rank matrix completion using alternating minimization.* STOC 2013.]
- BCD with more than 2 blocks
- Several others...

Large-scale methods

Stochastic optimization

Assumption 1: Possible to generate iid samples $\xi_1, \xi_2, \ldots$
Assumption 2: Oracle yields stochastic gradient $g(x, \xi)$, i.e.,

$$G(x) := \mathbb{E}[g(x, \xi)] \quad \text{s.t.} \quad G(x) \in \partial F(x).$$

Stochastic optimization

Assumption 1: Possible to generate iid samples $\xi_1, \xi_2, \ldots$
Assumption 2: Oracle yields stochastic gradient $g(x, \xi)$, i.e.,

$$G(x) := \mathbb{E}[g(x, \xi)] \quad \text{s.t.} \quad G(x) \in \partial F(x).$$

Theorem Let $\xi \in \Omega$; If $f(\cdot, \xi)$ is convex, and $F(\cdot)$ is finite valued in a neighborhood of x, then

$$\partial F(x) = \mathbb{E}[\partial_x f(x, \xi)].$$

Stochastic optimization

Assumption 1: Possible to generate iid samples $\xi_1, \xi_2, \ldots$
Assumption 2: Oracle yields stochastic gradient $g(x, \xi)$, i.e.,

$$G(x) := \mathbb{E}[g(x, \xi)] \quad \text{s.t.} \quad G(x) \in \partial F(x).$$

Theorem Let $\xi \in \Omega$; If $f(\cdot, \xi)$ is convex, and $F(\cdot)$ is finite valued in a neighborhood of x, then

$$\partial F(x) = \mathbb{E}[\partial_x f(x, \xi)].$$

- So $g(x, \omega) \in \partial_x f(x, \omega)$ is a stochastic subgradient.

Stochastic gradient

- Let $x_0 \in \mathcal{X}$
- For $k \geq 0$
  - Sample ξ_k; compute $g(x_k, \xi_k)$ using oracle
  - Update $x_{k+1} = P_{\mathcal{X}}(x_k - \alpha_k g(x_k, \xi_k))$, where $\alpha_k > 0$

Simply write

$$x_{k+1} = P_{\mathcal{X}}\left(x_k - \alpha_k g_k\right)$$

Incremental Gradient Methods

$$\min F(x) = \tfrac{1}{n}\sum_{i=1}^{n} f_i(x)$$

The incremental gradient method (IGM)

- Let $x_0 \in \mathbb{R}^n$
- For $k \geq 0$
  1. Pick $i(k) \in \{1, 2, \ldots, n\}$ uniformly at random
  2. $x_{k+1} = x_k - \eta_k \nabla f_{i(k)}(x_k)$

Incremental Gradient Methods

$$\min F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

The incremental gradient method (IGM)

- Let $x_0 \in \mathbb{R}^n$
- For $k \geq 0$
  1. Pick $i(k) \in \{1, 2, \ldots, n\}$ uniformly at random
  2. $x_{k+1} = x_k - \eta_k \nabla f_{i(k)}(x_k)$

$g \equiv \nabla f_{i(k)}$ may be viewed as a **stochastic gradient**

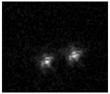
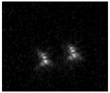
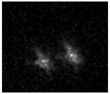
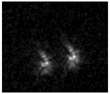
$g := g^{\text{true}} + e$, where e is mean-zero noise: $\mathbb{E}[e] = 0$

Example application

Multiframe blind deconvolution

(video)

Problem setup

time t	y_t	$=$	a_t	$*$	x	$+$	n_t
0		$=$		$*$		$+$	n_0
1		$=$		$*$		$+$	n_1
2		$=$		$*$		$+$	n_2
k		$=$		$*$		$+$	n_k

Formulation as matrix factorization

$$\begin{bmatrix} | & \vdots & | \\ y_1 & | & y_n \\ | & \vdots & | \end{bmatrix} \approx \begin{bmatrix} | & \vdots & | \\ a_1 & | & a_t \\ | & \vdots & | \end{bmatrix} * x$$

Rewrite: $a * x = Ax = Xa$

$$\begin{bmatrix} y_1 & y_2 & \cdots & y_t \end{bmatrix} \approx X \begin{bmatrix} a_1 & a_2 & \cdots & a_t \end{bmatrix}$$

$$Y \approx XA$$

Large-scale problem

Example, 5000 frames of size 512×512

$Y_{262144 \times 5000} \approx X_{262144 \times 262144} A_{262144 \times 5000}$

Without structure $\approx$ 70 billion parameters!

With structure, $\approx$ 4.8 **million parameters!**

Large-scale problem

Example, 5000 frames of size 512×512
$Y_{262144 \times 5000} \approx X_{262144 \times 262144} A_{262144 \times 5000}$

Without structure $\approx$ 70 billion parameters!
With structure, $\approx$ 4.8 **million parameters!**

Despite structure, alternating minimization **impractical**
Fix X, solve for A, requires updating $\approx$ 4.5 million params

Solving the problem

$$\min_{A_t, x} \quad \sum_{t=1}^{T} \tfrac{1}{2}\|y_t - A_t x\|^2 + \Omega(x) + \Gamma(A_t)$$

Solving the problem

$$\min_{A_t, x} \quad \sum_{t=1}^{T} \tfrac{1}{2}\|y_t - A_t x\|^2 + \Omega(x) + \Gamma(A_t)$$

Initialize guess $\boldsymbol{x}_0$
For $t = 1, 2, \ldots$
1. Observe image $\boldsymbol{y}_t$;

Solving the problem

$$\min_{A_t, x} \quad \sum\nolimits_{t=1}^{T} \tfrac{1}{2}\|y_t - A_t x\|^2 + \Omega(x) + \Gamma(A_t)$$

Initialize guess $\boldsymbol{x}_0$
For $t = 1, 2, \ldots$
1. Observe image $\boldsymbol{y}_t$;
2. Use $\boldsymbol{x}_{t-1}$ to **estimate** $\boldsymbol{A}_t$

Solving the problem

$$\min_{A_t, x} \quad \sum_{t=1}^{T} \tfrac{1}{2}\|y_t - A_t x\|^2 + \Omega(x) + \Gamma(A_t)$$

Initialize guess $\boldsymbol{x}_0$
For $t = 1, 2, \ldots$
1. Observe image $\boldsymbol{y}_t$;
2. Use $\boldsymbol{x}_{t-1}$ to **estimate** $\boldsymbol{A}_t$
3. Solve **optimization subproblem** to obtain $\boldsymbol{x}_t$

Solving the problem

$$\min_{A_t, x} \quad \sum_{t=1}^{T} \tfrac{1}{2}\|y_t - A_t x\|^2 + \Omega(x) + \Gamma(A_t)$$

Initialize guess $\boldsymbol{x}_0$
For $t = 1, 2, \ldots$
1. Observe image $\boldsymbol{y}_t$;
2. Use $\boldsymbol{x}_{t-1}$ to **estimate** $\boldsymbol{A}_t$
3. Solve **optimization subproblem** to obtain $\boldsymbol{x}_t$

Step 2. Model, estimate blur A_t — separate talk

Step 3. convex subproblem — reuse convex building blocks

Do Steps 2, 3 **inexactly** $\Longrightarrow$ realtime processing!

[Harmeling, Hirsch, Sra, Schölkopf (ICCP'09); Hirsch, Sra, Schölkopf, Harmeling (CVPR'10); Hirsch, Harmeling, Sra, Schölkopf (Astron. & Astrophy. (AA) 2011); Harmeling, Hirsch, Sra, Schölkopf, Schuler (Patent 2012); Sra (NIPS'12)]

Solving the problem: rewriting

Key idea

$$\min_{X,A} \Phi(X, A) \quad \equiv \quad \min_X \left(\min_A \Phi(X, A) \right) =$$

Solving the problem: rewriting

Key idea

$$\min_{X,A} \Phi(X,A) \equiv \min_{X} \left(\min_{A} \Phi(X,A) \right) = \min_{X} F(X)$$

$$F(X) := \min_{A} \Phi(X,A)$$

Solving the problem: rewriting

Key idea

$$\min_{X,A} \Phi(X,A) \quad \equiv \quad \min_X \left(\min_A \Phi(X,A) \right) = \min_X F(X)$$

$$F(X) \quad := \quad \min_A \Phi(X,A)$$

$$\Phi(X,A) = \|Y - XA\|^2 + \Omega(X) + \Gamma(A)$$

$$\hookrightarrow \min_X \; F(X) + \Omega(X)$$

but now F is **nonconvex**

Key to scalability

$$X^{\text{new}} \leftarrow \text{prox}_{\alpha\Omega}(X - \alpha \nabla F(X))$$

Key to scalability

$$X^{\text{new}} \leftarrow \text{prox}_{\alpha\Omega}(X - \alpha\nabla F(X) + e\ ) + p$$

If gradient is **inexactly** computed → e

If prox_{Ω} **inexactly** computed → p

Key to scalability

$$X^{\text{new}} \leftarrow \text{prox}_{\alpha\Omega}(X - \alpha\nabla F(X) + e\ ) + p$$

If gradient is **inexactly** computed → e

If prox_{Ω} **inexactly** computed → p

Example: Say $F(X) = \sum_{i=1}^{m} f_i(X)$

Instead of $\nabla F(X)$, use $\nabla f_k(x)$—**incremental**

m times cheaper (m can be in the millions or more)

Inexactness: key to scalability

incremental prox-method for **large-scale nonconvex**

[Sra (NIPS 12)]; (also *arXiv: [math.OC-1109.0258]*)

Theorem Limits points are approximately stationary.

Non-asymptotic convergence

$$\min \quad \frac{1}{n}\sum_i f_i(x)$$

SGD

1. For $t = 0$ to $T - 1$:
   1. Pick i_t from $\{1, \ldots, n\}$
   2. Update $x_{t+1} \leftarrow x_t - \eta_t \nabla f_{i_t}(x_t)$

Non-asymptotic convergence

$$\min \quad \frac{1}{n}\sum_i f_i(x)$$

SGD

1. For $t = 0$ to $T-1$:
   1. Pick i_t from $\{1, \ldots, n\}$
   2. Update $x_{t+1} \leftarrow x_t - \eta_t \nabla f_{i_t}(x_t)$

Theorem (Ghadimi, Lan). Suppose $\|\nabla f_i(x)\| \leq G$ for all i, $\eta_t = c/\sqrt{T}$, and $f \in C_L^1$. Then,

$$\mathbb{E}[\|\nabla f\|^2] \leq \frac{1}{c\sqrt{T}}\left(f(x_0) - f(x^*) + \tfrac{1}{2}Lc^2G^2\right)$$

[Ghadimi, Lan (2013). Stochastic first and zeroth-order methods for nonconvex stochastic programming. SIOPT.]

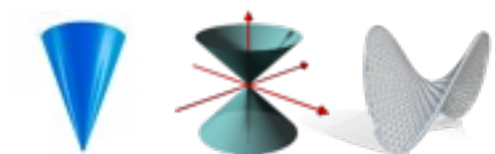
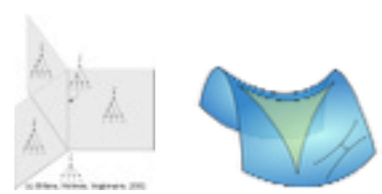
Other ncvx incremental methods

1. Incremental MM algo: [Mairal (2015). *Incremental majorization-minimization optimization with application to large-scale machine learning.* SIOPT]
2. ADMM: [Hong (2014). *A Distributed, Asynchronous and Incremental Algorithm for Nonconvex Optimization: An ADMM Based Approach*. arXiv]
3. SGD: [Lian, Huang, Li, Liu (2015). *Asynchronous parallel stochastic gradient for nonconvex optimization*. NIPS 2015]

First two do not prove rates; third one builds on Ghadimi & Lan's analysis to provide rate on $\mathbb{E}[\|\nabla f\|^2]$

Geometric Optimization

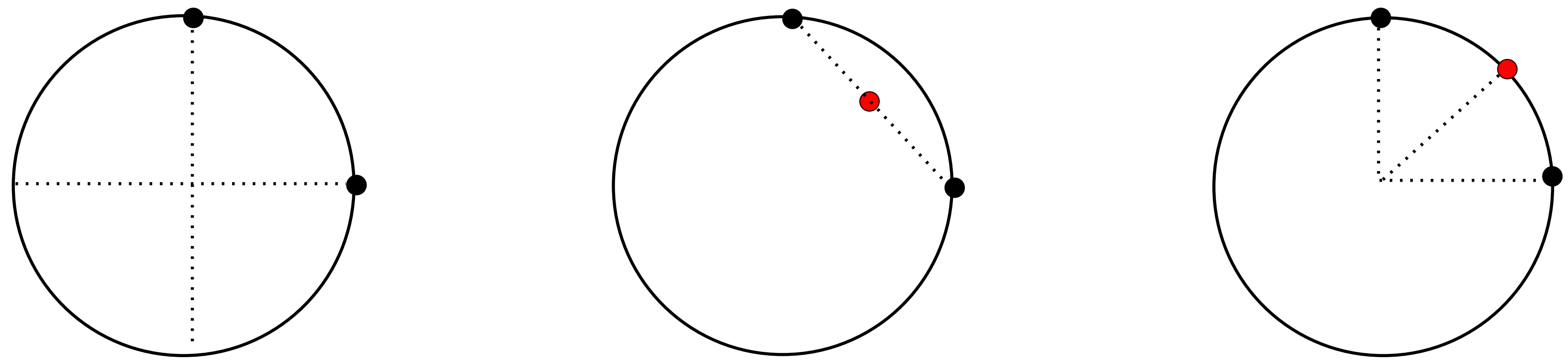
Geometry Everywhere

- **The usual vector space** 
- **Manifolds** 
  (hypersphere, orthogonal matrices, complicated surfaces)
- **Convex sets** 
  (probability simplex, semidefinite cone, polyhedra)
- **Metric spaces** 
  (tree space, Wasserstein metric, negatively curved spaces)

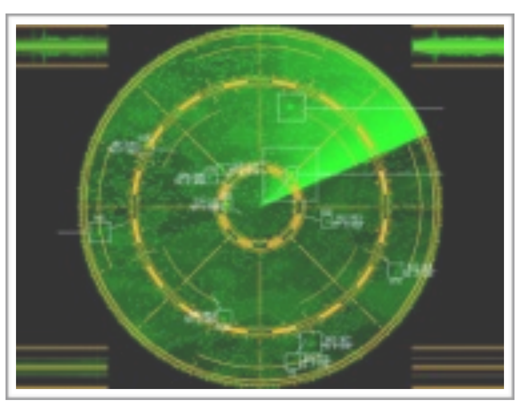
Machine Learning
Graphics
Robotics
Vision
BCI
NLP
Statistics

Geometric Data

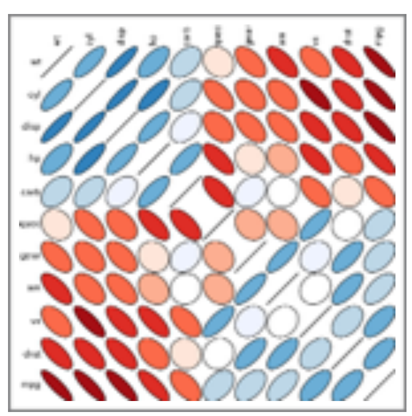
Rotations



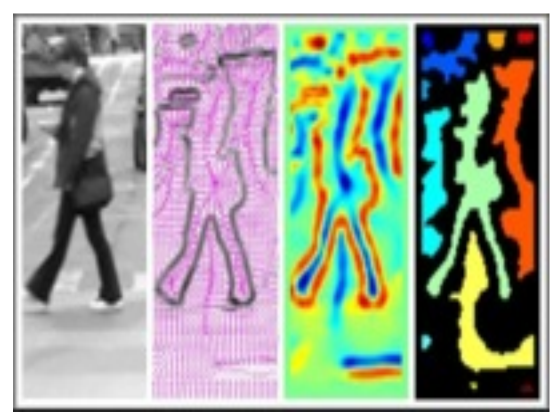
Covariances as data / features / params: $X_1, X_2, \ldots, X_n \succeq 0$



Radar

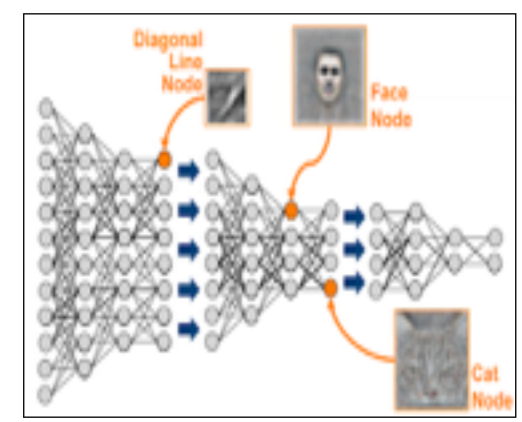


DTI



CV

BCI



DeepLrn

[Cherian, Sra, Papanikolopoulos (2012); Cherian, Sra (2015)]

Averaging Matrices

$$\min_{M \succ 0} \quad \sum_i \delta_R^2(M, A_i)$$

$$\min_{M \succ 0} \quad \sum_i \delta_S^2(M, A_i)$$

$$\delta_R^2(X, Y) := \|\log \mathrm{Eig}(X^{-1}Y)\|^2$$

$$\delta_S^2(X, Y) := \log\det\left(\tfrac{X+Y}{2}\right) - \tfrac{1}{2}\log\det(XY)$$

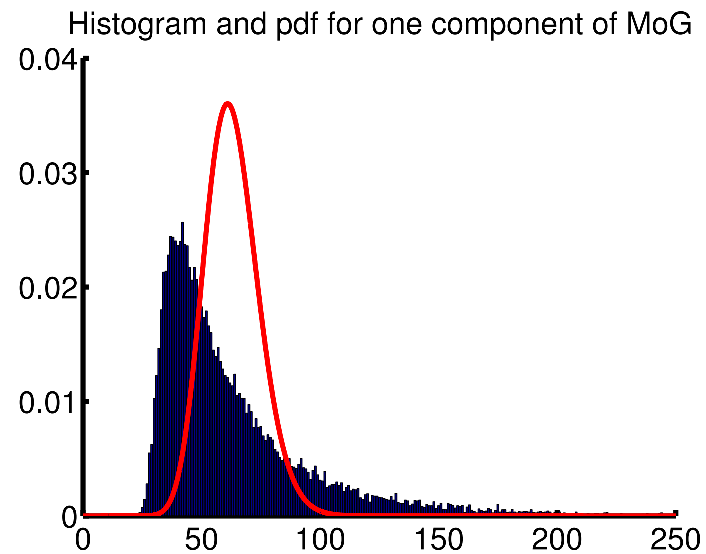
nonconvex but globally solvable!

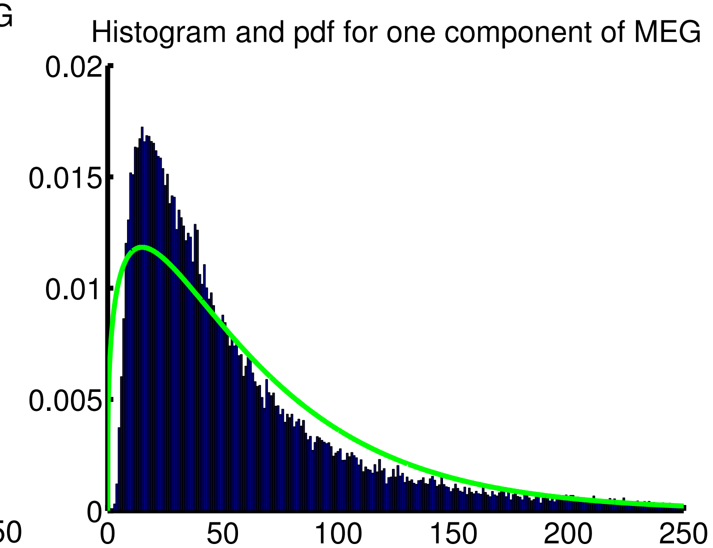
[Sra (2012, 2014)]

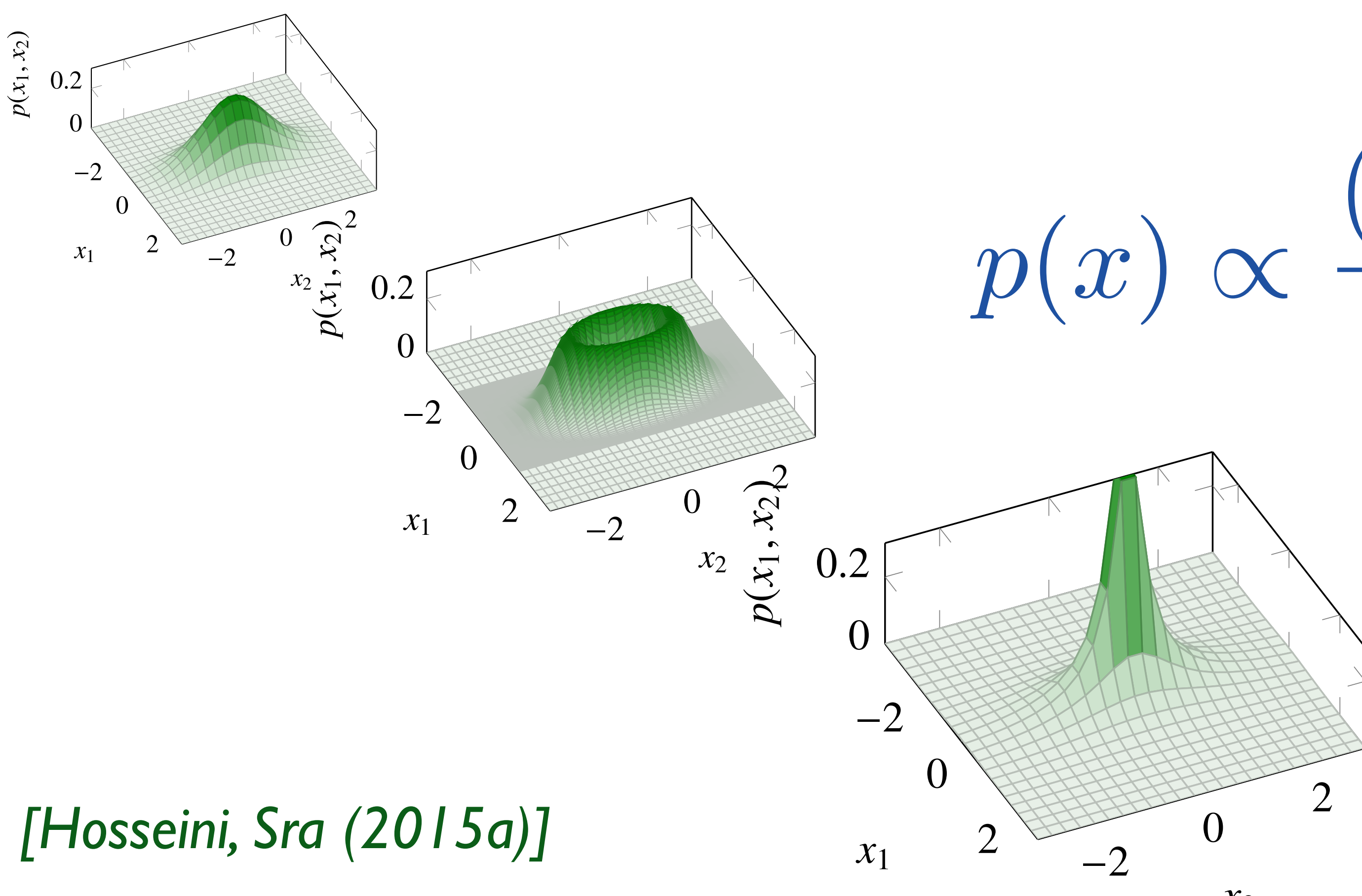
Non-Gaussian Models

Natural Image Statistics

- Extract 200,000 training patches from 4167 images
- 10 sets of 100,000 test patches
- Log-transform intensities; add small amount of white noise







$$p(x) \propto \frac{(x^T\Sigma^{-1}x)^{a-\frac{d}{2}} e^{-\frac{a}{d}x^T\Sigma^{-1}x}}{\det(\Sigma)^{1/2}}$$

Elliptically Contoured Distributions (ECD)

[Hosseini, Sra (2015a)]

Optimization Problem

Likelihood maximization

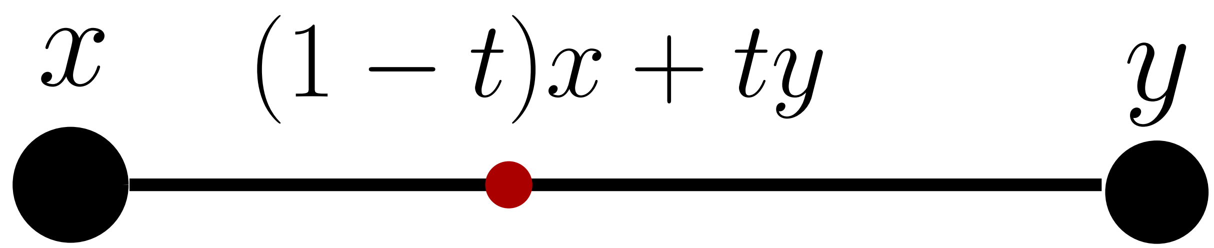
Given observations $x_1, x_2, \ldots, x_n$ find m.l.e. by solving

$$\frac{n}{2}\log\det(\Sigma) - \left(a - \frac{d}{2}\right)\sum_{i=1}^{n}\log(x_i^T\Sigma^{-1}x_i) + \frac{a}{d}\operatorname{trace}\left(\Sigma^{-1}\sum_i x_i x_i^T\right)$$

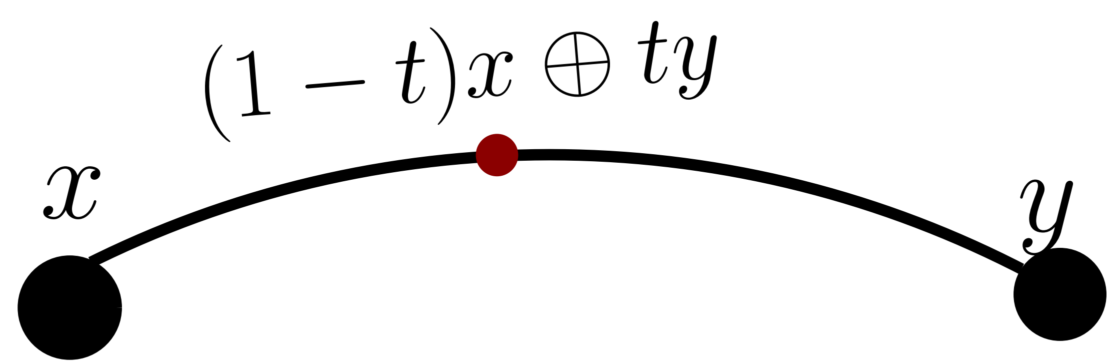
convex or nonconvex: often globally solvable!

Geometric Convexity

Convexity



Geodesic convexity



$$f((1-t)x \oplus ty) \leq (1-t)f(x) + tf(y)$$

Metric spaces & curvature: [Alexandrov; Busemann; Cartan; Bridson, Häflinger; Gromov; Perelman]

Geometric Optimization

[Sra, Hosseini (2013)]

[Sra, Hosseini (2015)]

Recognizing, constructing, and optimizing geodesically convex functions

$$X\#_t Y := X^{\frac{1}{2}}(X^{-\frac{1}{2}}YX^{-\frac{1}{2}})^t X^{\frac{1}{2}}$$

$$f(X\#_t Y) \leq (1-t)f(X) + tf(Y)$$

Corollaries

$$X \mapsto \log\det(B + \sum_i A_i^* X A_i)$$

$$X \mapsto \log\operatorname{per}(B + \sum_i A_i^* X A_i)$$

$$\delta_R^2(X,Y), \quad \delta_S^2(X,Y)$$

(jointly g-convex)

Many more theorems and corollaries

One-D version known as: **Geometric Programming**
www.stanford.edu/~boyd/papers/gp_tutorial.html
[Boyd, Kim, Vandenberghe, Hassibi (2007). 61pp.]

Averaging Matrices

$$\min_{M \succ 0} \Phi(M) = \sum\nolimits_i \delta_S^2(M, A_i)$$

$$\nabla \Phi(M) = 0$$

$$M^{-1} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{M + A_i}{2} \right)^{-1}$$

Averaging Matrices

$$\min_{M \succ 0} \Phi(M) = \sum_i \delta_S^2(M, A_i)$$

$$\nabla \Phi(M) = 0$$

$$M_{k+1}^{-1} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{M_k + A_i}{2} \right)^{-1}$$

Plug-and-play!

[Sra (2012)]

Nonlinear Perron-Frobenius fixed-point theory

Key Object

Theorem: Iteration is a **contraction** in a suitable metric space

$$\delta_T(X,Y) := \|\log(X^{-1/2} Y X^{-1/2}\|_\infty$$

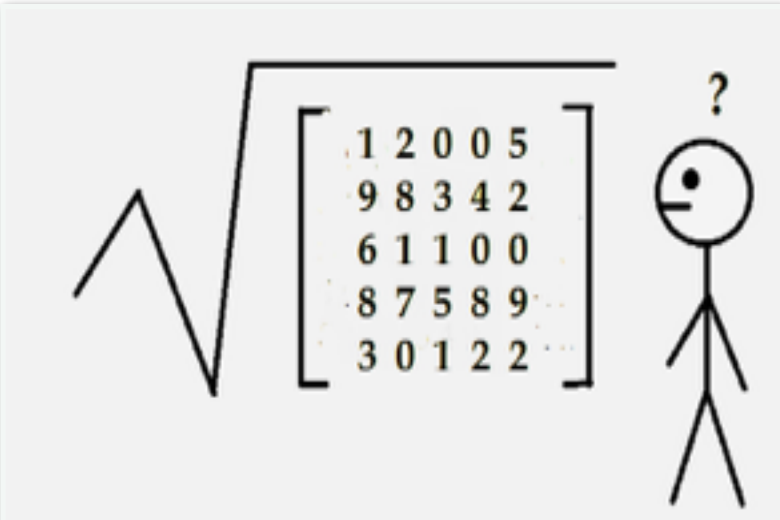
Key properties of this metric

(see [Sra, Hosseini SIOPT'15] for details)

$$\begin{aligned}
\delta_T(X^{-1}, Y^{-1}) &= \delta_T(X,Y) \\
\delta_T(B^* X B, B^* Y B) &= \delta_T(X,Y), \qquad B \in \mathrm{GL}_n(\mathbb{C}) \\
\delta_T(X^t, Y^t) &\le |t|\delta_T(X,Y), \qquad \text{for } t \in [-1,1] \\
\delta_T\Big(\sum_i w_i X_i, \sum_i w_i Y_i\Big) &\le \max_{1\le i\le m} \delta_T(X_i, Y_i), \qquad w_i \ge 0, w \ne 0 \\
\delta_T(X+A, Y+A) &\le \frac{\alpha}{\alpha+\beta}\delta_T(X,Y), \qquad A \succeq 0,
\end{aligned}$$

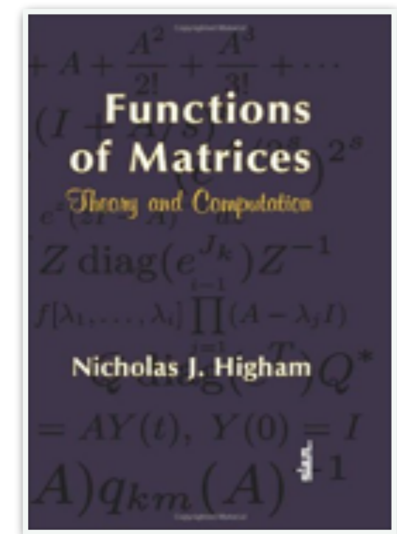
Note: Contraction does not depend on geodesic convexity

Matrix Square Root



Broadly applicable

Key to '`expm`', '`logm`'



Matrix Square Root

Nonconvex optimization through the Euclidean lens

$$\min_{X \in \mathbb{R}^{n \times n}} \|M - X^2\|_F^2$$

Gradient descent

$$X_{t+1} \leftarrow X_t - \eta(X_t^2 - M)X_t - \eta X_t(X_t^2 - M)$$

Simple(ish) algo; linear convergence; **nontrivial** analysis

[Jain, Jin, Kakade, Netrapalli; Jul. 2015]

Matrix Square Root

Nonconvex optimization thorough non-Euclidean lens

$$\min_{X \succ 0} \quad \delta_S^2(X, A) + \delta_S^2(X, I)$$

Fixed-point iteration

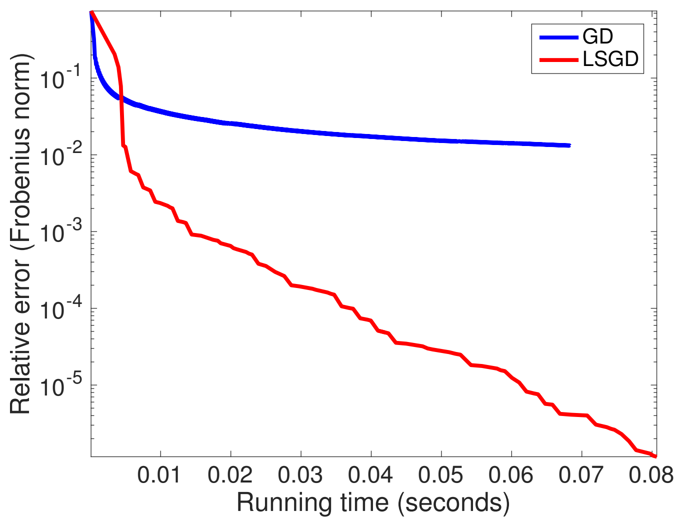
$$X_{k+1} \leftarrow [(X_k + A)^{-1} + (X_k + I)^{-1}]^{-1}$$

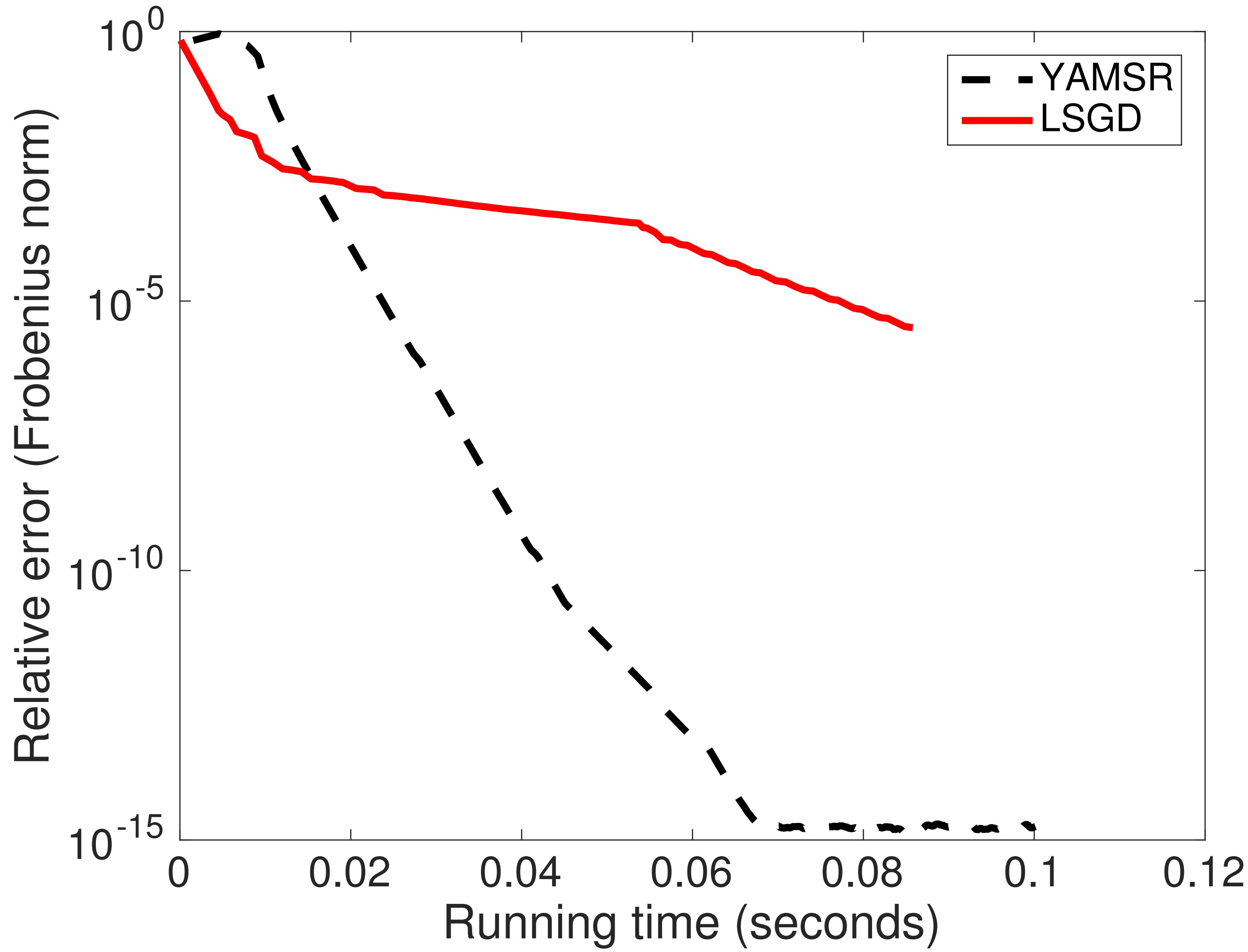
Simple method; linear convergence; 1/2 page analysis!

Global optimality thanks to geodesic convexity

[Sra; Jul. 2015] $\delta_S^2(X, Y) := \frac{1}{2} \log\det\left(\frac{X+Y}{2}\right) - \frac{1}{2}\log\det(XY)$

Matrix Square Root



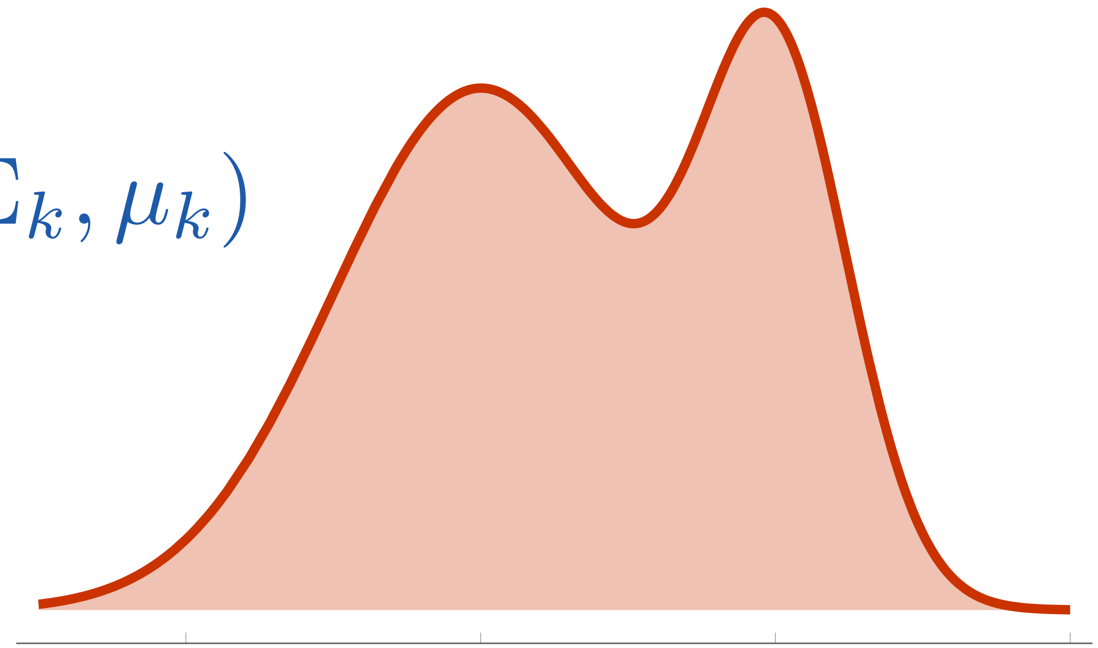


$$50 \times 50 \text{ matrix } I + \beta UU^T$$
$$\kappa \approx 64$$

Gaussian Mixture Models

$$p_{\text{mix}}(x) := \sum_{k=1}^{K} \pi_k p_{\mathcal{N}}(x; \Sigma_k, \mu_k)$$

$$\max \prod_i p_{\text{mix}}(x_i)$$



Expectation maximization (EM): default choice

$$p_{\mathcal{N}}(x; \Sigma, \mu) \propto \frac{1}{\sqrt{\det(\Sigma)}} \exp\left(-\tfrac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

[Hosseini, Sra (2015b)]

Gaussian Mixture Models

- **Nonconvex**: both via Euclidean and manifold view
- Recent surge of theoretical results in TCS
- Numerically: EM as default choice

(Newton, quasi-Newton, other optim. often inferior to EM for GMMs — Xu, Jordan '96)

Difficulty: Positive definiteness constraint on Σ

Gaussian Mixture Models

K	EM	Manifold-CG
2	17s / 29.28	947s / 29.28
5	202s / 32.07	5262s / 32.07
10	2159s / 33.05	17712 / 33.03

GMM for d=35

Off-the-shelf manifold optim. fails!

www.manopt.org

How To Fix: Intuition

log-likelihood for 1 component

$$\max_{\mu,\Sigma\succ 0} \mathcal{L}(\mu,\Sigma) := \sum_{i=1}^{n} \log p_{\mathcal{N}}(x_i;\mu,\Sigma).$$

Euclidean convex
not geodesically convex

Geodesic Convexity

$$y_i = [x_i; 1] \quad S = \begin{bmatrix} \Sigma + \mu\mu^T & \mu \\ \mu^T & 1 \end{bmatrix}$$

$$\max_{S \succ 0} \ \widehat{\mathcal{L}}(S) := \sum_{i=1}^{n} \log q_{\mathcal{N}}(y_i; S),$$

Theorem. The modified log-likelihood is g-convex. Local max of modified LL is local max of original.

$$f(X \#_t Y) \leq (1-t) f(X) + t f(Y)$$

$$X \#_t Y := X^{\frac{1}{2}} (X^{-\frac{1}{2}} Y X^{-\frac{1}{2}})^t X^{\frac{1}{2}}$$

[Hosseini, Sra (2015b)] *[Sra, Hosseini (2015)]*

Numerical Results

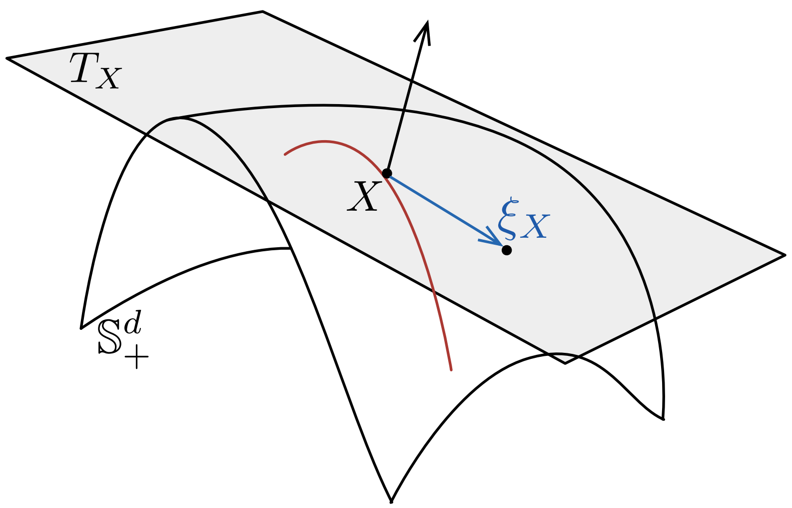
K	EM	Manifold-CG	Reparam-LBFGS
2	17s / 29.28	947s / 29.28	**14s** / 29.28
5	202s / 32.07	5262s / 32.07	**117s** / 32.07
10	2159s / 33.05	17712 / 33.03	**658s** / 33.06

GMM, d=35; convergence tol = 1E-5

Many more results in: *[Hosseini, Sra (2015b); arXiv: 1506.07677]*

Gaussian Mixture Models

Key ingredients



1. L-BFGS on the manifold
2. Careful line-search procedure

Toolboxes at:

suvrit.de/work/soft/gopt.html
github.com/utvisionlab/mixest

*[**Sra,** Hosseini (2015); Hosseini, Sra (2015b)]*

Many More Connections!

- Fundamental theory, duality, etc.
- Machine learning
- Deep learning
- Signal processing
- Engineering (EE, Aero, etc.)
- Brain-Computer interfaces
- Quantum Information Theory
- Geometry of tree-space
- Hyperbolic cones, graphs, spaces
- Nonlinear Perron-Frobenius Theory
- Matrix analysis, algebra

http://suvrit.de/gopt.html

SEVERAL OMITTED ITEMS

- See Springer Encyclopedia on Optimization (over 4500 pages!)
- Convex relaxations of nonconvex problems (SDP relaxations, SOS, etc.)
- Algorithms (trust-region methods, cutting plane techniques, bundle methods, active-set methods, and 100s of others)
- Applications
- Software, Systems
- Parallel and distributed algorithms
- Theory: convex analysis, geometry, probability
- Polynomials, sums-of-squares, noncommutative polynomials
- Infinite dimensional optimization
- Discrete optimization, including submodular minimization and maximization
- Multi-stage stochastic programming,
- Optimizing with probabilistic (chance) constraints
- Robust optimization
- Algorithms and theory details for optimization on manifolds
- Optimization in geodesic metric spaces
- And 100s of other things!

ngiyabonga
рахмат
danke
謝謝
teşekkür ederim
Баярлалаа
спасибо
mersi
barka
welalin
tack
spas
kia ora
faafetai lava
yinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
mahalo
tapadh leat
kiitos
dankie
dhanyavad
nanni
nandri
thank you
gracias
хвала
asante
manana
obrigada
tenki
bedankt
bayarlalaa
gracje
hvala
mauruuru
koszonom
dankon
aciú
akun
djiere dieuf
tau
mochchakkeram
дякую
mamnun
murakoze
chokrane
enkosi
dziękuję
chnorakaloutioun
grazias
gracies
sulpay
taiku
go raibh maith agat
sobodi
dėkuji
obrigado
sagolun
sukriya
kop khun krap
arigatô
takk
dakujem
trugarez
mèsi
didi madloba
najis tuke
ありがとう
tanemirt
grazie
terima kasih
rahmet
dhanyavadagalu
shukriya
merce
мерси
kam sah hamnida
rahmat
diolch
감사합니다
xièxie
ευχαριστώ
merci