Nonconvex Optimization for High-Dimensional Estimation (Part 1)

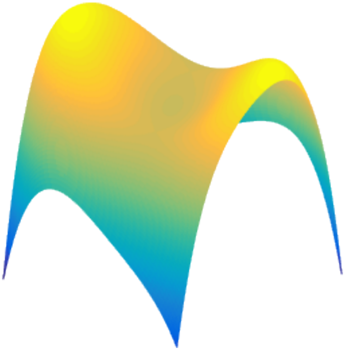
Yuxin Chen

Wharton Statistics & Data Science, Fall 2023

Nonconvex estimation problems are everywhere

Empirical risk minimization is usually nonconvex

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}; \text{data}) \quad \rightarrow \quad \text{loss function may be nonconvex}$$

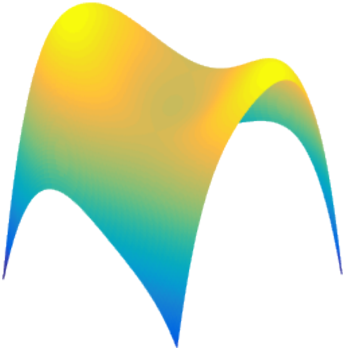


Nonconvex estimation problems are everywhere

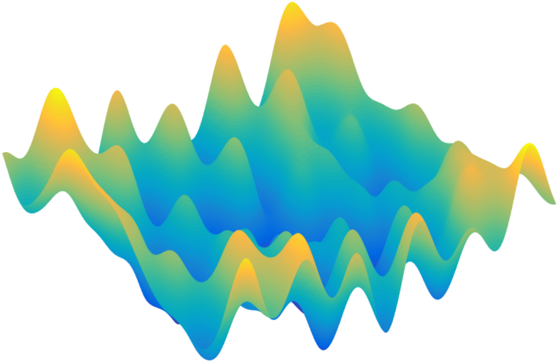
Empirical risk minimization is usually nonconvex

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}; \text{data}) \quad \rightarrow \quad \text{loss function may be nonconvex}$$

- low-rank matrix completion
- blind deconvolution
- dictionary learning
- mixture models
- deep learning
- ...



Nonconvex optimization may be super scary



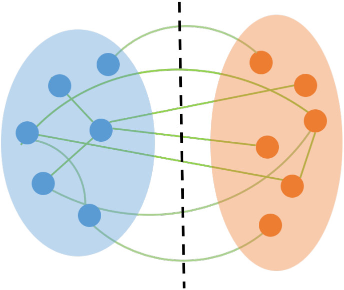
There may be bumps everywhere and exponentially many local optima

e.g. 1-layer neural net (Auer, Herbster, Warmuth '96; Vu '98)

Example: solving quadratic programs is hard

Finding maximum cut in a graph is about solving a quadratic program

$$\begin{array}{rl} \text{maximize}_{\boldsymbol{x}} & \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \\ \text{subj. to} & x_i^2 = 1, \quad i = 1, \cdots, n \end{array}$$



Example: solving quadratic programs is hard

"I can't find an efficient algorithm, but neither can all these people."

figure credit: coding horror

$1,000,000 question

One strategy: convex relaxation

Can relax into convex problems by

- finding convex surrogates (e.g. matrix completion)
- lifting into higher dimensions (e.g. Max-Cut)

Example of convex surrogate: matrix completion

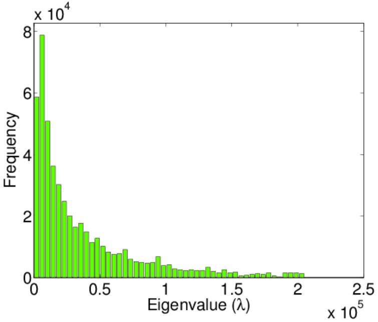
figure credit: Candès et al.

Netflix challenge

Predict unseen ratings

Low-rank modeling

figure credit: E. Candès



A few factors explain most of the data

Low-rank modeling

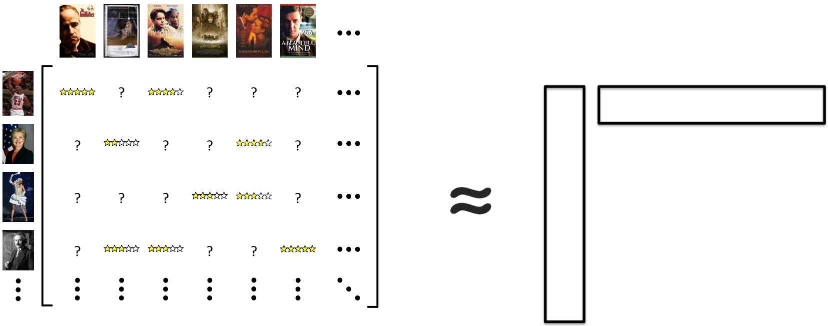


figure credit: E. Candès

A few factors explain most of the data $\longrightarrow$ low-rank approximation

How to exploit (approx.) low-rank structure in prediction?

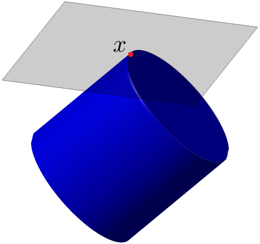
Example of convex surrogate: matrix completion

— Fazel '02, Recht, Parrilo, Fazel '10, Candès, Recht '09

$$\text{minimize}_{\boldsymbol{M}} \ \text{rank}(\boldsymbol{M}) \ \text{subj. to data constraints}$$

$\Downarrow$ cvx surrogate

$$\text{minimize}_{\boldsymbol{M}} \ \text{nuc-norm}(\boldsymbol{M}) \quad \text{subj. to data constraints}$$



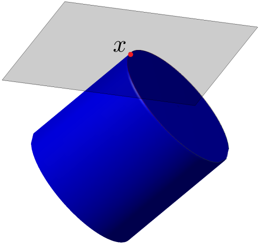
Example of convex surrogate: matrix completion

— Fazel '02, Recht, Parrilo, Fazel '10, Candès, Recht '09

$$\text{minimize}_{\boldsymbol{M}} \ \text{rank}(\boldsymbol{M}) \ \text{subj. to data constraints}$$

$$\Downarrow \ \text{cvx surrogate}$$

$$\text{minimize}_{\boldsymbol{M}} \ \text{nuc-norm}(\boldsymbol{M}) \qquad \text{subj. to data constraints}$$



robust variation used by Netflix

— Candès, Li, Ma, Wright '10

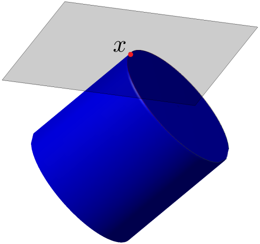
Example of convex surrogate: matrix completion

— Fazel '02, Recht, Parrilo, Fazel '10, Candès, Recht '09

minimize$_{\boldsymbol{M}}$ rank$(\boldsymbol{M})$ subj. to data constraints

$\Downarrow$ cvx surrogate

minimize$_{\boldsymbol{M}}$ nuc-norm$(\boldsymbol{M})$ subj. to data constraints



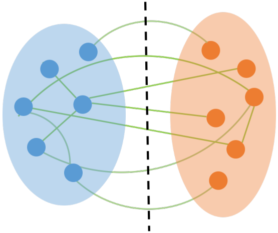
robust variation used by Netflix

— Candès, Li, Ma, Wright '10

Problem: operate in *full* matrix space even though $\boldsymbol{X}$ is low-rank

Example of lifting: Max-Cut

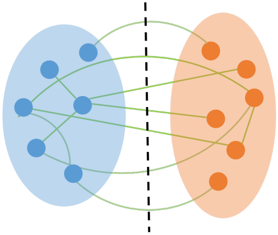
— Goemans, Williamson '95



$$\begin{aligned} \text{maximize}_{\boldsymbol{x}} \quad & \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \\ \text{subj. to} \quad & x_i^2 = 1, \quad i = 1, \cdots, n \end{aligned}$$

Example of lifting: Max-Cut

— Goemans, Williamson '95



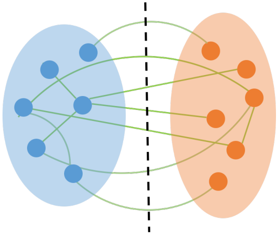
$$\text{maximize}_{\boldsymbol{x}} \quad \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x}$$
$$\text{subj. to} \quad x_i^2 = 1, \quad i = 1, \cdots, n$$

$\Downarrow$ let $\boldsymbol{X}$ be $\boldsymbol{x}\boldsymbol{x}^\top$

$$\begin{aligned} \text{maximize}_{\boldsymbol{X}} \quad & \langle \boldsymbol{X}, \boldsymbol{W} \rangle \\ \text{subj. to} \quad & \boldsymbol{X}_{i,i} = 1, \quad i = 1, \cdots, n \\ & \boldsymbol{X} \succeq \boldsymbol{0} \\ & \text{rank}(\boldsymbol{X}) = 1 \end{aligned}$$

Example of lifting: Max-Cut

— Goemans, Williamson '95



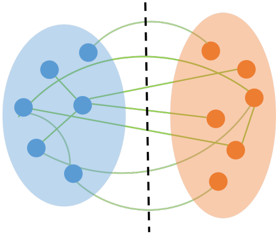
$$\begin{array}{rl} \text{maximize}_{\boldsymbol{x}} & \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \\ \text{subj. to} & x_i^2 = 1, \quad i = 1, \cdots, n \end{array}$$

$\Downarrow$ let $\boldsymbol{X}$ be $\boldsymbol{x}\boldsymbol{x}^\top$

$$\begin{array}{rl} \text{maximize}_{\boldsymbol{X}} & \langle \boldsymbol{X}, \boldsymbol{W} \rangle \\ \text{subj. to} & \boldsymbol{X}_{i,i} = 1, \quad i = 1, \cdots, n \\ & \boldsymbol{X} \succeq \boldsymbol{0} \\ & \cancel{\text{rank}(\boldsymbol{X}) = 1} \end{array}$$

Example of lifting: Max-Cut

— Goemans, Williamson '95



$$\begin{array}{ll} \text{maximize}_{\boldsymbol{x}} & \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \\ \text{subj. to} & x_i^2 = 1, \quad i = 1, \cdots, n \end{array}$$

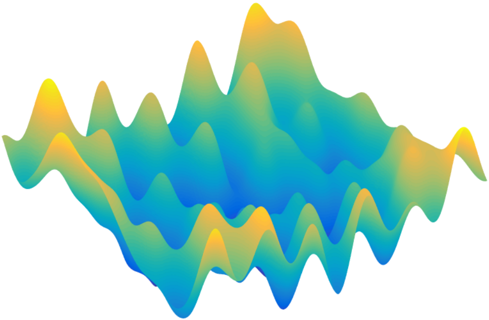
$\Downarrow$ let $\boldsymbol{X}$ be $\boldsymbol{x}\boldsymbol{x}^\top$

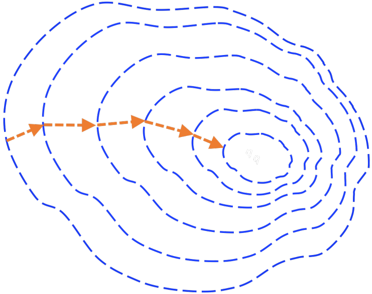
$$\begin{array}{ll} \text{maximize}_{\boldsymbol{X}} & \langle \boldsymbol{X}, \boldsymbol{W} \rangle \\ \text{subj. to} & \boldsymbol{X}_{i,i} = 1, \quad i = 1, \cdots, n \\ & \boldsymbol{X} \succeq \boldsymbol{0} \\ & \cancel{\text{rank}(\boldsymbol{X}) = 1} \end{array}$$

Problem: explosion in dimensions ($\mathbb{R}^n \to \mathbb{R}^{n \times n}$)

How about optimizing nonconvex problems directly without lifting?

Nonconvex problems are solved on a daily basis via simple algorithms like *(stochastic) gradient descent*

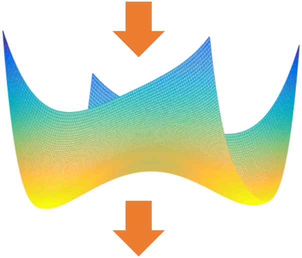




How come simple nonconvex algorithms work so well in practice?

Statistical models come to rescue

statistical models



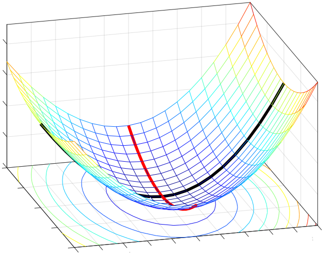
benign landscape

tractable algorithms

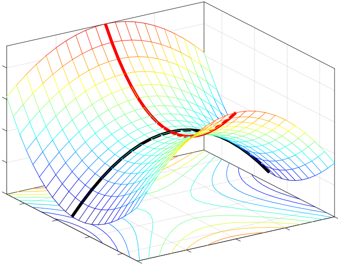
When data are generated by certain statistical models, problems are often much nicer than worst-case instances

Sometimes they are much nicer than we think

Under certain statistical models,
we see benign global geometry: no spurious local optima



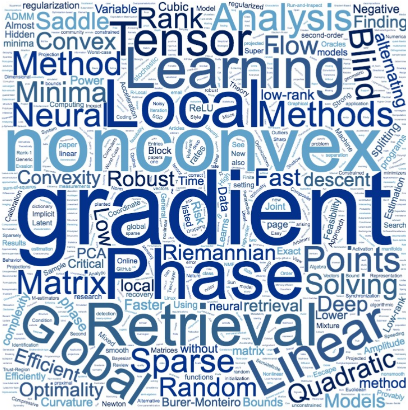
global minimum



saddle point

Even the simplest possible nonconvex methods might be remarkably efficient under suitable statistical models

Nonconvex optimization with guarantees



Phase retrieval: Gerchberg-Saxton '72, Netrapalli et al. '13, Candès, Li, Soltanolkotabi '14, Chen, Candès '15, Cai, Li, Ma '15, Zhang et al. 16, Wang et al. '16, Sun et al. '16, Ma et al. '17, Chen et al. '18, ...

Matrix completion: Keshavan et al. '09, Jain et al. '09, Hardt '13, Sun, Luo '15, Chen, Wainwright '15, Zheng, Lafferty '16, Ge et al. '16, Jin et al. '16, Ma et al. '17, ...

Matrix sensing: Jain et al. '13, Tu et al. '15, Zheng, Lafferty '15, Bhojanapalli et al. 16, Li, Zhu, Tang '18, ...

Blind deconvolution / demixing: Li et al. '16, Lee et al. '16, Ling, Strohmer '16, Huang, Hand '16, Ma et al. '17, Zhang et al. '18, Li, Bresler '18, Dong, Shi '18, ...

Dictionary learning: Arora et al. '14, Sun et al. '15, Chatterji, Bartlett '17, ...

Robust principal component analysis: Netrapalli et al. '14, Yi et al. '16, Gu et al. '16, Ge et al. '17, Cherapanamjeri et al. '17, ...

"Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview," Y. Chi, Y. M. Lu, and Y. Chen, IEEE Trans. on Signal Processing, vol. 67, no. 20, pp. 5239-5269, 2019.

Some preliminaries of optimization

Unconstrained optimization

Consider an unconstrained optimization problem

$$\text{minimize}_{\boldsymbol{x}} \qquad f(\boldsymbol{x})$$

Definition 1 (first-order critical points)

A first-order critical point of f satisfies

$$\nabla f(\boldsymbol{x}) = \mathbf{0}$$

Unconstrained optimization

Consider an unconstrained optimization problem

$$\text{minimize}_{\boldsymbol{x}} \qquad f(\boldsymbol{x})$$

Definition 2 (second-order critical points)

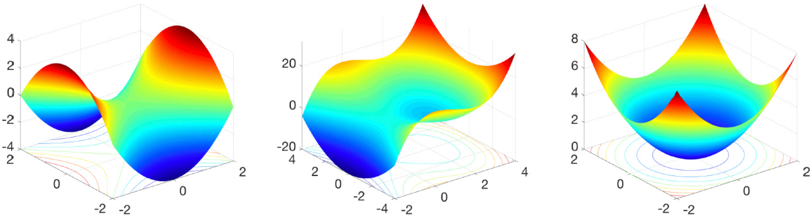
A second-order critical point $\boldsymbol{x}$ satisfies

$$\nabla f(\boldsymbol{x}) = \mathbf{0} \quad \text{and} \quad \nabla^2 f(\boldsymbol{x}) \succeq \mathbf{0}$$

Several types of critical points

For any first-order critical point $\boldsymbol{x}$:

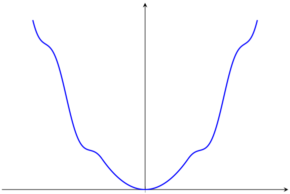
- $\nabla^2 f(\boldsymbol{x}) \prec \mathbf{0}$ $\rightarrow$ local maximum
- $\nabla^2 f(\boldsymbol{x}) \succ \mathbf{0}$ $\rightarrow$ local minimum
- $\lambda_{\min}(\nabla^2 f(\boldsymbol{x})) < 0$ $\rightarrow$ *strict* saddle point



(a) strict saddle (b) local minimum (c) global minimum

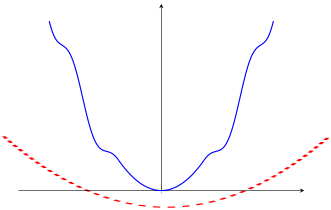
figure credit: Li et al. '16

Gradient descent theory



Two standard conditions that enable geometric convergence of GD

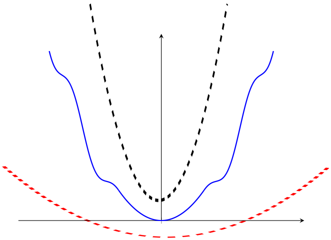
Gradient descent theory



Two standard conditions that enable geometric convergence of GD

- (local) restricted strong convexity (or regularity condition)

Gradient descent theory



Two standard conditions that enable geometric convergence of GD

- (local) restricted strong convexity (or regularity condition)
- (local) smoothness

$$\nabla^2 f(\boldsymbol{x}) \succ \mathbf{0} \quad \text{and} \quad \text{is well-conditioned}$$

Gradient descent theory revisited

f is said to be α-strongly convex and β-smooth if

$$\mathbf{0} \ \preceq \ \alpha \boldsymbol{I} \ \preceq \ \nabla^2 f(\boldsymbol{x}) \ \preceq \ \beta \boldsymbol{I}, \qquad \forall \boldsymbol{x}$$

ℓ_2 **error contraction:** GD ($\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t)$) with $\eta = 1/\beta$ obeys

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}_{\mathsf{opt}}\|_2 \leq \left(1 - \frac{\alpha}{\beta}\right) \|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2$$

Gradient descent theory revisited

f is said to be α-strongly convex and β-smooth if

$$\mathbf{0} \preceq \alpha \boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq \beta \boldsymbol{I}, \qquad \forall \boldsymbol{x}$$

ℓ_2 **error contraction:** GD ($\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t)$) with $\eta = 1/\beta$ obeys

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}_{\mathsf{opt}}\|_2 \leq \left(1 - \frac{\alpha}{\beta}\right) \|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2$$

- Condition number β/α determines rate of convergence

Gradient descent theory revisited

f is said to be α-strongly convex and β-smooth if

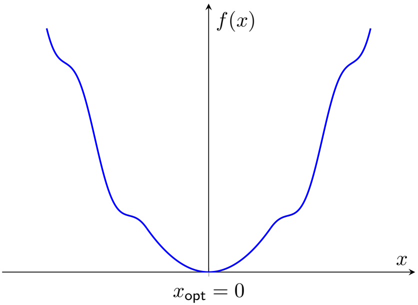
$$\mathbf{0} \preceq \alpha \boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq \beta \boldsymbol{I}, \qquad \forall \boldsymbol{x}$$

ℓ_2 **error contraction:** GD ($\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t)$) with $\eta = 1/\beta$ obeys

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}_{\mathsf{opt}}\|_2 \leq \left(1 - \frac{\alpha}{\beta}\right) \|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2$$

- Condition number β/α determines rate of convergence
- Attains ε-accuracy within $O(\frac{\beta}{\alpha} \log \frac{1}{\varepsilon})$ iterations

Regularity Condition (RC)



Definition 3 (Regularity Condition (RC))

$\boldsymbol{g}(\cdot)$ is said to obey $\text{RC}(\mu, \lambda, \zeta)$ for some $\mu, \lambda, \zeta > 0$ if

$$2\langle\, \boldsymbol{g}(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}_{\text{opt}} \rangle \geq \mu \|\boldsymbol{g}(\boldsymbol{x})\|_2^2 + \lambda \, \|\boldsymbol{x} - \boldsymbol{x}_{\text{opt}}\|_2^2 \quad \forall \boldsymbol{x}$$

Convergence under RC

ℓ_2 **error contraction:** The update rule ($\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \boldsymbol{g}(\boldsymbol{x}^t)$) with $\eta = \mu$ obeys

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}_{\mathsf{opt}}\|_2 \leq (1 - \mu\lambda) \|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2$$

- $\boldsymbol{g}(\cdot)$: more general search directions
  - example: in vanilla GD, $\boldsymbol{g}(\boldsymbol{x}) = \nabla f(\boldsymbol{x})$

Convergence under RC

ℓ_2 error contraction: The update rule ($\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \boldsymbol{g}(\boldsymbol{x}^t)$) with $\eta = \mu$ obeys

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}_{\mathsf{opt}}\|_2 \le (1 - \mu\lambda)\,\|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2$$

- $\boldsymbol{g}(\cdot)$: more general search directions
  - example: in vanilla GD, $\boldsymbol{g}(\boldsymbol{x}) = \nabla f(\boldsymbol{x})$
- The product $\mu\lambda$ determines the rate of convergence

Convergence under RC

ℓ_2 error contraction: The update rule ($\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \boldsymbol{g}(\boldsymbol{x}^t)$) with $\eta = \mu$ obeys

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}_{\mathsf{opt}}\|_2 \le (1 - \mu\lambda)\, \|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2$$

- $\boldsymbol{g}(\cdot)$: more general search directions
  - example: in vanilla GD, $\boldsymbol{g}(\boldsymbol{x}) = \nabla f(\boldsymbol{x})$
- The product $\mu\lambda$ determines the rate of convergence
- Attains ε-accuracy within $O(\frac{1}{\mu\lambda} \log \frac{1}{\varepsilon})$ iterations

RC = one-point strong convexity + smoothness

- One-point α-strong convexity:

$$f(\boldsymbol{x}_{\mathsf{opt}}) - f(\boldsymbol{x}) \geq \langle \nabla f(\boldsymbol{x}), \boldsymbol{x}_{\mathsf{opt}} - \boldsymbol{x} \rangle + \frac{\alpha}{2} \|\boldsymbol{x} - \boldsymbol{x}_{\mathsf{opt}}\|_2^2 \tag{1}$$

- β-smoothness:

$$\begin{aligned} f(\boldsymbol{x}_{\mathsf{opt}}) - f(\boldsymbol{x}) &\leq f\Big(\boldsymbol{x} - \frac{1}{\beta} \nabla f(\boldsymbol{x})\Big) - f(\boldsymbol{x}) \\ &\leq \Big\langle \nabla f(\boldsymbol{x}), -\frac{1}{\beta} \nabla f(\boldsymbol{x}) \Big\rangle + \frac{\beta}{2} \Big\| \frac{1}{\beta} \nabla f(\boldsymbol{x}) \Big\|_2^2 \\ &= -\frac{1}{2\beta} \|\nabla f(\boldsymbol{x})\|_2^2 \end{aligned} \tag{2}$$

RC = one-point strong convexity + smoothness

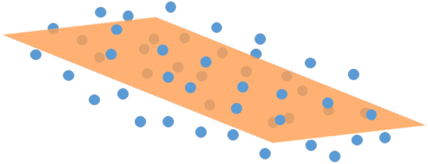
Combining (1) and (2) yields

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}_{\mathsf{opt}} \rangle \geq \frac{\alpha}{2} \|\boldsymbol{x} - \boldsymbol{x}_{\mathsf{opt}}\|_2^2 + \frac{1}{2\beta} \|\nabla f(\boldsymbol{x})\|_2^2 \qquad (3)$$

— *RC holds with $\mu = 1/\beta$ and $\lambda = \alpha$*

A toy example: rank-1 matrix factorization

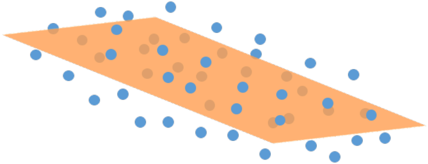
Revisiting PCA



Given $\boldsymbol{M} \succeq \boldsymbol{0} \in \mathbb{R}^{n \times n}$ (not necessarily low-rank), find its best rank-r approximation:

$$\underbrace{\widehat{\boldsymbol{M}} = \mathrm{argmin}_{\boldsymbol{Z}} \; \|\boldsymbol{Z} - \boldsymbol{M}\|_{\mathrm{F}}^2 \quad \text{s.t.} \quad \mathrm{rank}(\boldsymbol{Z}) \leq r}_{\text{nonconvex optimization!}}$$

Revisiting PCA



This problem admits a closed-form solution

- let $\boldsymbol{M} = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$ be eigen-decomposition of $\boldsymbol{M}$ ($\lambda_1 \geq \cdots \geq \lambda_n$), then

$$\widehat{\boldsymbol{M}} = \sum_{i=1}^{r} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$$

— *nonconvex, but tractable*

Optimization viewpoint

If we factorize $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{X}^\top$ with $\boldsymbol{X} \in \mathbb{R}^{n \times r}$, then it leads to a nonconvex problem:

$$\text{minimize}_{\boldsymbol{X} \in \mathbb{R}^{n \times r}} \quad f(\boldsymbol{X}) = \frac{1}{4}\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}\|_{\mathrm{F}}^2$$

To simplify exposition, set $r = 1$:

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{4}\|\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{M}\|_{\mathrm{F}}^2$$

Questions

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{4}\|\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{M}\|_{\mathrm{F}}^2$$

- Where / what are the critical points?
- What does the curvature behave like, at least locally around the global minimizer?

Critical points of $f(\cdot)$

$$\boldsymbol{x} \text{ is a critical point, i.e. } \nabla f(\boldsymbol{x}) = (\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{M})\boldsymbol{x} = \boldsymbol{0}$$

$$\Updownarrow$$

$$\boldsymbol{M}\boldsymbol{x} = \|\boldsymbol{x}\|_2^2 \boldsymbol{x}$$

$$\Updownarrow$$

$$\boldsymbol{x} \text{ aligns with an eigenvector of } \boldsymbol{M} \quad \text{or} \quad \boldsymbol{x} = \boldsymbol{0}$$

Since $\boldsymbol{M}\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i$, the set of critical points is given by

$$\{\boldsymbol{0}\} \cup \{\pm\sqrt{\lambda_i}\boldsymbol{u}_i,\ i = 1, \ldots, n\}$$

Categorization of critical points

The critical points can be further categorized based on the **Hessians**:

$$\nabla^2 f(\boldsymbol{x}) := 2\boldsymbol{x}\boldsymbol{x}^\top + \|\boldsymbol{x}\|_2^2 \boldsymbol{I} - \boldsymbol{M}$$

- For any non-zero critical point $\boldsymbol{x}_k = \pm\sqrt{\lambda_k}\boldsymbol{u}_k$:

$$\begin{aligned}
\nabla^2 f(\boldsymbol{x}_k) &= 2\lambda_k \boldsymbol{u}_k \boldsymbol{u}_k^\top + \lambda_k \boldsymbol{I} - \boldsymbol{M} \\
&= 2\lambda_k \boldsymbol{u}_k \boldsymbol{u}_k^\top + \lambda_k \left(\sum_{i=1}^n \boldsymbol{u}_i \boldsymbol{u}_i^\top \right) - \sum_{i=1}^n \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\top \\
&= \sum_{i:i \neq k} (\lambda_k - \lambda_i) \boldsymbol{u}_i \boldsymbol{u}_i^\top + 2\lambda_k \boldsymbol{u}_k \boldsymbol{u}_k^\top
\end{aligned}$$

Categorization of critical points

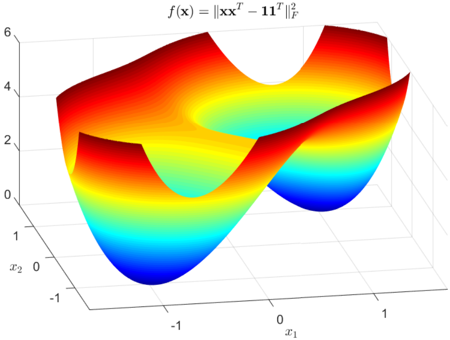
The critical points can be further categorized based on the **Hessians**:

$$\nabla^2 f(\boldsymbol{x}) := 2\boldsymbol{x}\boldsymbol{x}^\top + \|\boldsymbol{x}\|_2^2 \boldsymbol{I} - \boldsymbol{M}$$

- If $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_n \geq 0$, then
  - $\nabla^2 f(\boldsymbol{x}_1) \succ \mathbf{0}$ $\rightarrow$ local minima
  - $1 < k \leq n$: $\lambda_{\min}(\nabla^2 f(\boldsymbol{x}_k)) < 0$, $\lambda_{\max}(\nabla^2 f(\boldsymbol{x}_k)) > 0$ $\rightarrow$ strict saddle
  - $\boldsymbol{x} = \mathbf{0}$: $\nabla^2 f(\mathbf{0}) = -\boldsymbol{M} \preceq \mathbf{0}$ $\rightarrow$ local maxima

Good news: benign landscape

For example, for 2-dimensional case $f(\boldsymbol{x}) = \left\| \boldsymbol{x}\boldsymbol{x}^\top - \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \right\|_{\mathrm{F}}^2$



global minima: $\boldsymbol{x} = \pm \begin{bmatrix} 1 \\ 1 \end{bmatrix}$; strict saddles: $\boldsymbol{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, and $\pm \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

— *No "spurious" local minima!*

Local strong convexity and local linear convergence

- The global minimizers: $\boldsymbol{x}_{\mathsf{opt}} = \pm\sqrt{\lambda_1}\boldsymbol{u}_1$
- For all $\boldsymbol{x}$ obeying $\underbrace{\|\boldsymbol{x} - \boldsymbol{x}_{\mathsf{opt}}\|_2 \le \frac{\lambda_1 - \lambda_2}{15\sqrt{\lambda_1}}}_{\text{basin of attraction}}$, one has

$$0.25(\lambda_1 - \lambda_2)\boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{x}) \preceq 4.5\lambda_1 \boldsymbol{I}_n$$

Local strong convexity and local linear convergence

- The global minimizers: $\boldsymbol{x}_{\mathsf{opt}} = \pm\sqrt{\lambda_1}\boldsymbol{u}_1$
- For all $\boldsymbol{x}$ obeying $\underbrace{\|\boldsymbol{x} - \boldsymbol{x}_{\mathsf{opt}}\|_2 \le \frac{\lambda_1 - \lambda_2}{15\sqrt{\lambda_1}}}_{\text{basin of attraction}}$, one has

$$0.25(\lambda_1 - \lambda_2)\boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{x}) \preceq 4.5\lambda_1 \boldsymbol{I}_n$$

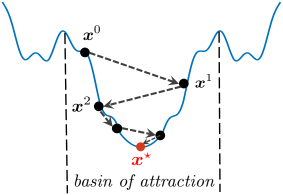
ℓ_2 **error contraction:** The GD iterates obey

$$\left\|\boldsymbol{x}^t - \sqrt{\lambda_1}\boldsymbol{u}_1\right\|_2 \le \left(1 - \frac{\lambda_1 - \lambda_2}{18\lambda_1}\right)^t \left\|\boldsymbol{x}^0 - \sqrt{\lambda_1}\boldsymbol{u}_1\right\|_2, \ t \ge 0,$$

as long as $\|\boldsymbol{x}^0 - \sqrt{\lambda_1}\boldsymbol{u}_1\|_2 \le \frac{\lambda_1 - \lambda_2}{15\sqrt{\lambda_1}}$

Two vignettes

Two-stage approach:



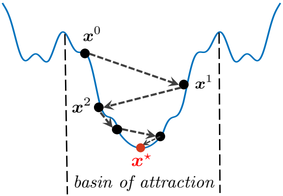
smart initialization

$+$

local refinement

Two vignettes

Two-stage approach:

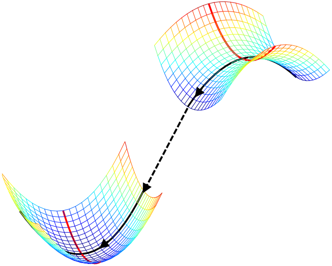


smart initialization

$+$

local refinement

Global landscape:



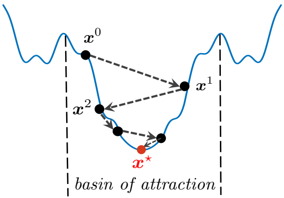
benign landscape

$+$

saddle-point escaping

Two vignettes

Two-stage approach:

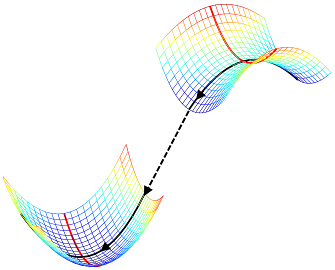


smart initialization

$+$

local refinement

Global landscape:



benign landscape

$+$

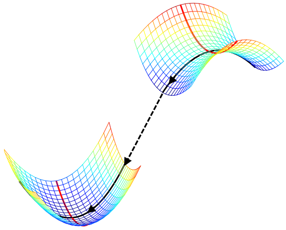
saddle-point escaping

This lecture focuses mainly on the two-stage approach

Global landscape

Benign landscape:

- all local minima = global minima
- other critical points = strict saddle points



Saddle-point escaping algorithms:

- trust-region methods
- perturbed gradient descent
- perturbed SGD
- cubic-regularization
- ...

Check the recent overview: *Zhang, Qu, Wright "From Symmetry to Geometry: Tractable Nonconvex Problems"*