Introduction to nonconvex optimization

Cong Ma

University of Chicago, Autumn 2021

Unconstrained optimization

Consider an unconstrained optimization problem

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

- For simplicity, we assume $f(\boldsymbol{x})$ is twice differentiable
- We assume the minimizer $\boldsymbol{x}_{\mathsf{opt}}$ exists, i.e.,

$$\boldsymbol{x}_{\mathsf{opt}} := \arg\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

Critical/stationary points

Definition 7.1

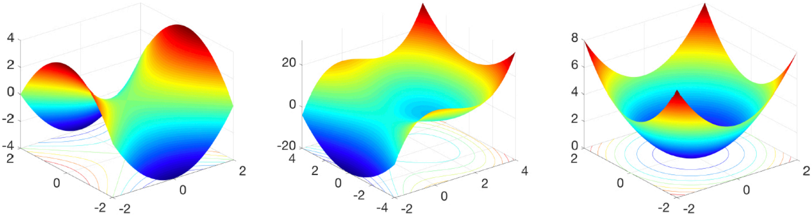
A first-order critical point of f satisfies

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

- If f is convex, any 1st-order critical point is a global minimizer
- Finding 1st-order stationary point is sufficient for convex optimization
- Example: gradient descent (GD)

How about nonconvex optimization?

First-order critical points could be global min, local min, local max, saddle points...



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

figure credit: Li et al. '16

Simple algorithms like GD could stuck at undesired stationary points

Types of critical points

Definition 7.2

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

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

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

When are nonconvex problems solvable?

(Local) strong convexity and smoothness

Definition 7.3

A twice differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is said to be α-strongly convex in a set $\mathcal{B}$ if for all $\boldsymbol{x} \in \mathcal{B}$

$$\nabla^2 f(\boldsymbol{x}) \succeq \alpha \boldsymbol{I}_n.$$

Definition 7.4

A twice differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is said to be β-smooth in a set $\mathcal{B}$ if for all $\boldsymbol{x} \in \mathcal{B}$

$$\|\nabla^2 f(\boldsymbol{x})\| \leq \beta.$$

Gradient descent theory revisited

Gradient descent method with step size $\eta > 0$

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t)$$

Lemma 7.5

Suppose f is α-strongly convex and β-smooth in the local ball $\mathcal{B}_\delta(\boldsymbol{x}_{\mathsf{opt}}) := \{\boldsymbol{x} \mid \|\boldsymbol{x} - \boldsymbol{x}_{\mathsf{opt}}\|_2 \le \delta\}$. Running gradient descent from $\boldsymbol{x}^0 \in \mathcal{B}_\delta(\boldsymbol{x}_{\mathsf{opt}})$ with $\eta = 1/\beta$ achieves linear convergence

$$\|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2 \le \left(1 - \frac{\alpha}{\beta}\right)^t \|\boldsymbol{x}^0 - \boldsymbol{x}_{\mathsf{opt}}\|_2, \quad t = 0, 1, 2, \ldots$$

Implications

- Condition number β/α determines rate of convergence
- Attains ε-accuracy (i.e., $\|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2 \leq \varepsilon \|\boldsymbol{x}_{\mathsf{opt}}\|_2$) within

$$O\left(\frac{\beta}{\alpha}\log\frac{1}{\varepsilon}\right)$$

  iterations
- Needs initialization $\boldsymbol{x}^0 \in \mathcal{B}_\delta(\boldsymbol{x}_{\mathsf{opt}})$: basin of attraction

Proof of Lemma 7.5

Since $\nabla f(\boldsymbol{x}_{\mathsf{opt}}) = \mathbf{0}$, we can rewrite GD as

$$\begin{aligned}\boldsymbol{x}^{t+1} - \boldsymbol{x}_{\mathsf{opt}} &= \boldsymbol{x}^{t} - \eta \nabla f(\boldsymbol{x}^{t}) - [\boldsymbol{x}_{\mathsf{opt}} - \eta \nabla f(\boldsymbol{x}_{\mathsf{opt}})] \\ &= \left[\boldsymbol{I}_n - \eta \int_0^1 \nabla^2 f(\boldsymbol{x}(\tau)) \mathsf{d}\tau\right] (\boldsymbol{x}^{t} - \boldsymbol{x}_{\mathsf{opt}}),\end{aligned}$$

where $\boldsymbol{x}(\tau) := \boldsymbol{x}_{\mathsf{opt}} + \tau(\boldsymbol{x}^{t} - \boldsymbol{x}_{\mathsf{opt}})$. By local strong convexity and smoothness, one has

$$\alpha \boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{x}(\tau)) \preceq \beta \boldsymbol{I}_n, \qquad \text{for all } 0 \leq \tau \leq 1$$

Therefore $\eta = 1/\beta$ yields

$$\mathbf{0} \preceq \boldsymbol{I}_n - \eta \int_0^1 \nabla^2 f(\boldsymbol{x}(\tau)) \mathsf{d}\tau \preceq (1 - \frac{\alpha}{\beta}) \boldsymbol{I}_n,$$

which further implies

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

Regularity condition

More generally, for update rule

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^{t} - \eta \boldsymbol{g}(\boldsymbol{x}^{t}),$$

where $g(\cdot) : \mathbb{R}^n \mapsto \mathbb{R}^n$

Definition 7.6

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

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

- Negative search direction $\boldsymbol{g}$ is positively correlated with error $\boldsymbol{x} - \boldsymbol{x}_{\mathsf{opt}} \Longrightarrow$ one-step improvement
- $\mu\lambda \leq 1$ by Cauchy-Schwarz

RC = one-point strong convexity + smoothness

- One-point α-strong convexity:

$$f(\boldsymbol{x}_{\mathsf{opt}}) - f(\boldsymbol{x}) \geq \langle \nabla f(\boldsymbol{x}), \boldsymbol{x}_{\mathsf{opt}} - \boldsymbol{x} \rangle + \frac{\alpha}{2} \|\boldsymbol{x} - \boldsymbol{x}_{\mathsf{opt}}\|_2^2 \qquad (7.1)$$

- β-smoothness:

$$\begin{aligned} f(\boldsymbol{x}_{\mathsf{opt}}) - f(\boldsymbol{x}) &\leq f\Big(\boldsymbol{x} - \frac{1}{\beta}\nabla f(\boldsymbol{x})\Big) - f(\boldsymbol{x}) \\ &\leq \Big\langle \nabla f(\boldsymbol{x}), -\frac{1}{\beta}\nabla f(\boldsymbol{x}) \Big\rangle + \frac{\beta}{2}\Big\| \frac{1}{\beta}\nabla f(\boldsymbol{x}) \Big\|_2^2 \\ &= -\frac{1}{2\beta}\left\|\nabla f(\boldsymbol{x})\right\|_2^2 \qquad (7.2) \end{aligned}$$

RC = one-point strong convexity + smoothness

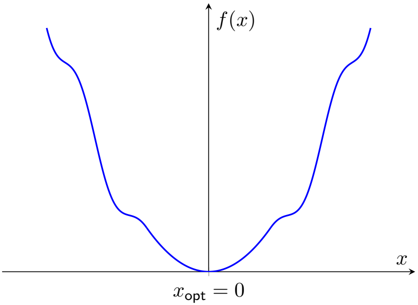
Combining relations (7.1) and (7.2) yields

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

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

Example of nonconvex functions

When $\boldsymbol{g}(\boldsymbol{x}) = \nabla f(\boldsymbol{x})$, f is not necessarily convex



$$f(x) = \begin{cases} x^2, & |x| \le 6, \\ x^2 + 1.5|x|(\cos(|x| - 6) - 1), & |x| > 6 \end{cases}$$

Convergence under RC

Lemma 7.7

Suppose $\boldsymbol{g}(\cdot)$ *obeys* $\mathsf{RC}(\mu, \lambda, \delta)$*. The update rule* $(\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \boldsymbol{g}(\boldsymbol{x}^t))$ *with* $\eta = \mu$ *and* $\boldsymbol{x}^0 \in \mathcal{B}_\delta(\boldsymbol{x}_{\mathsf{opt}})$ *obeys*

$$\|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2^2 \leq (1 - \mu\lambda)^t \, \|\boldsymbol{x}^0 - \boldsymbol{x}_{\mathsf{opt}}\|_2^2$$

- $\boldsymbol{g}(\cdot)$: 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

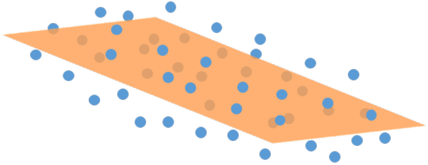
Proof of Lemma 7.7

By definition, one has

$$\begin{aligned}
&\|\boldsymbol{x}^{t+1} - \boldsymbol{x}_{\mathsf{opt}}\|_2^2 = \|\boldsymbol{x}^t - \eta \boldsymbol{g}(\boldsymbol{x}^t) - \boldsymbol{x}_{\mathsf{opt}}\|_2^2 \\
&\quad = \|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2^2 + \eta^2 \|\boldsymbol{g}(\boldsymbol{x}^t)\|_2^2 - 2\eta \left\langle \boldsymbol{g}(\boldsymbol{x}^t), \boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}} \right\rangle \\
&\quad \le \|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2^2 + \eta^2 \|\boldsymbol{g}(\boldsymbol{x}^t)\|_2^2 - \eta \left(\lambda \|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2^2 + \mu \|\boldsymbol{g}(\boldsymbol{x}^t)\|_2^2 \right) \\
&\quad = (1 - \eta\lambda) \|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2^2 + \eta(\eta - \mu) \|\boldsymbol{g}(\boldsymbol{x}^t)\|_2^2 \\
&\quad \le (1 - \mu\lambda) \|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2^2
\end{aligned}$$

A toy example: rank-1 matrix factorization

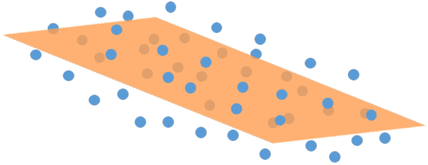
Principal component analysis



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

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

Principal component analysis



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 $\boldsymbol{M}$ $(\lambda_1 \geq \cdots \lambda_r > \lambda_{r+1} \geq \lambda_n)$, 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 $r = 1$:

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

Interesting questions

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

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

Local linear convergence of GD

Theorem 7.8

Suppose that $\|\boldsymbol{x}_0 - \sqrt{\lambda_1}\boldsymbol{u}_1\|_2 \leq \frac{\lambda_1 - \lambda_2}{15\sqrt{\lambda_1}}$ *and set* $\eta = \frac{1}{4.5\lambda_1}$*, GD obeys*

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

- condition number/eigengap determines rate of convergence
- Requires initialization: use spectral method?

Proof of Theorem 7.8

It suffices to show that for all $\boldsymbol{x}$ obeying $\underbrace{\|\boldsymbol{x} - \sqrt{\lambda_1}\boldsymbol{u}_1\|_2 \le \frac{\lambda_1 - \lambda_2}{15\sqrt{\lambda_1}}}_{\text{basin of attraction}}$,

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

Express gradient and Hessian as

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

Preliminary facts

Let $\boldsymbol{\Delta} := \boldsymbol{x} - \sqrt{\lambda_1}\boldsymbol{u}_1$. It is seen that when $\|\boldsymbol{\Delta}\|_2 \leq \frac{\lambda_1 - \lambda_2}{15\sqrt{\lambda_1}}$, one has

$$\lambda_1 - 0.25(\lambda_1 - \lambda_2) \leq \|\boldsymbol{x}\|_2^2 \leq 1.15\lambda_1;$$
$$\|\boldsymbol{\Delta}\|_2 \leq \|\boldsymbol{x}\|_2;$$
$$\|\boldsymbol{\Delta}\|_2 \|\boldsymbol{x}\|_2 \leq (\lambda_1 - \lambda_2)/12$$

Local smoothness

Triangle inequality gives

$$\begin{aligned}\|\nabla^2 f(\boldsymbol{x})\| &\leq \|2\boldsymbol{x}\boldsymbol{x}^\top\| + \|\boldsymbol{x}\|_2^2 + \|\boldsymbol{M}\| \\ &\leq 3\|\boldsymbol{x}\|_2^2 + \lambda_1 < 4.5\lambda_1,\end{aligned}$$

where the last relation follows from $\|\boldsymbol{x}\|_2^2 \leq 1.15\lambda_1$

Local strong convexity

Recall that $\boldsymbol{\Delta} = \boldsymbol{x} - \sqrt{\lambda_1}\boldsymbol{u}_1$

$$
\begin{aligned}
\boldsymbol{x}\boldsymbol{x}^\top &= \lambda_1 \boldsymbol{u}_1\boldsymbol{u}_1^\top + \boldsymbol{\Delta}\boldsymbol{x}^\top + \boldsymbol{x}\boldsymbol{\Delta}^\top - \boldsymbol{\Delta}\boldsymbol{\Delta}^\top \\
&\succeq \lambda_1 \boldsymbol{u}_1\boldsymbol{u}_1^\top - 3\|\boldsymbol{\Delta}\|_2\|\boldsymbol{x}\|_2 \boldsymbol{I}_n \qquad\qquad (\|\boldsymbol{\Delta}\|_2 \leq \|\boldsymbol{x}\|_2) \\
&\succeq \lambda_1 \boldsymbol{u}_1\boldsymbol{u}_1^\top - 0.25(\lambda_1 - \lambda_2)\boldsymbol{I}_n,
\end{aligned}
$$

where last line relies on $\|\boldsymbol{\Delta}\|_2\|\boldsymbol{x}\|_2 \leq (\lambda_1 - \lambda_2)/12$. Consequently,

$$
\begin{aligned}
\nabla^2 f(\boldsymbol{x}) &= 2\boldsymbol{x}\boldsymbol{x}^\top + \|\boldsymbol{x}\|_2^2 \boldsymbol{I}_n - \lambda_1 \boldsymbol{u}_1\boldsymbol{u}_1^\top - \textstyle\sum_{i=2}^n \lambda_i \boldsymbol{u}_i\boldsymbol{u}_i^\top \\
&\succeq 2\lambda_1 \boldsymbol{u}_1\boldsymbol{u}_1^\top + (\|\boldsymbol{x}\|_2^2 - 0.5)(\lambda_1 - \lambda_2)\boldsymbol{I}_n - \lambda_1 \boldsymbol{u}_1\boldsymbol{u}_1^\top - \textstyle\sum_{i=2}^n \lambda_i \boldsymbol{u}_i\boldsymbol{u}_i^\top \\
&\succeq (\|\boldsymbol{x}\|_2^2 - 0.5(\lambda_1 - \lambda_2) + \lambda_1)\boldsymbol{u}_1\boldsymbol{u}_1^\top \\
&\quad + \textstyle\sum_{i=2}^n (\|\boldsymbol{x}\|_2^2 - 0.5(\lambda_1 - \lambda_2) - \lambda_i)\boldsymbol{u}_i\boldsymbol{u}_i^\top \\
&\succeq (\|\boldsymbol{x}\|_2^2 - 0.5(\lambda_1 - \lambda_2) - \lambda_2)\boldsymbol{I}_n \\
&\succeq 0.25(\lambda_1 - \lambda_2)\boldsymbol{I}_n \qquad\qquad (\lambda_1 - 0.25(\lambda_1 - \lambda_2) \leq \|\boldsymbol{x}\|_2^2)
\end{aligned}
$$

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

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

Categorization of critical points

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

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

- 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 (cont.)

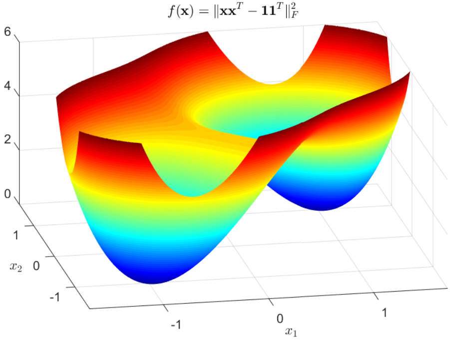
If $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_n \geq 0$, then

- $\nabla^2 f(\boldsymbol{x}_1) \succ \mathbf{0}$ $\rightarrow$ 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, strict saddle

all local are global; all saddle are strict

A pictorial example

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

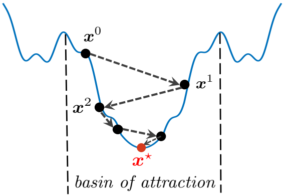


global minima: $\boldsymbol{x} = \pm \begin{bmatrix} 1 \\ 1 \end{bmatrix}$; strict saddles: $\boldsymbol{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, and $\pm \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

— *No "spurious" local minima!*

Two vignettes

Two-stage approach:



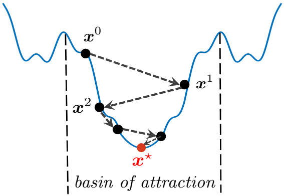
smart initialization

$+$

local refinement

Two vignettes

Two-stage approach:

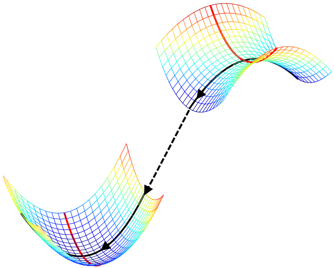


smart initialization

$+$

local refinement

Global landscape:



benign landscape

$+$

saddle-point escaping

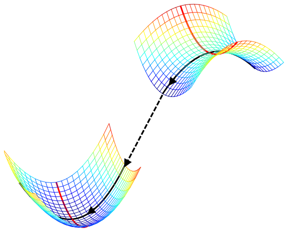
Global landscape

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
- ...



Next steps

- Generic local analysis of (regularized) gradient descent
- Refined local analysis for gradient descent
- Global landscape analysis
- Gradient descent with random initialization
- (Maybe) Gradient descent with arbitrary initialization