Minimum Eigenvalue Routines and Nonconvex Optimization

Clément W. Royer

SIAM Conference on Applied Linear Algebra - May 16, 2024

Negative eigenvalues and nonconvex optimization

- Motivation: Interest for nonconvex problems in data science.
- Tool: Second-order derivatives (matrices).
- Question: Use of eigenvalues.

Welcome to this mini-symposium!

Negative eigenvalues and nonconvex optimization

- Motivation: Interest for nonconvex problems in data science.
- Tool: Second-order derivatives (matrices).
- Question: Use of eigenvalues.

This talk

- Quick introduction to the topic;
- One result on minimum eigenvalue estimation.

Setup

Optimization problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

$f \in \mathcal{C}^2$, bounded below, nonconvex.

Key property

If $x^* \in \operatorname{argmin}_x f(x)$, then

$$\nabla f(x^*) = 0, \quad \nabla^2 f(x^*) \succeq 0.$$

Setup

Optimization problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

$f \in \mathcal{C}^2$, bounded below, nonconvex.

Key property

If $x^* \in \operatorname{argmin}_x f(x)$, then

$$\nabla f(x^*) = 0, \quad \nabla^2 f(x^*) \succeq 0.$$

Setup

Optimization problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

$f \in \mathcal{C}^2$, bounded below, nonconvex.

Key property

If $x^* \in \operatorname{argmin}_x f(x)$, then

$$\nabla f(x^*) = 0, \quad \nabla^2 f(x^*) \succeq 0.$$

Corollary

If $\exists\ d$ such that $d^{\mathrm{T}} \nabla^2 f(x) d < 0$ (negative curvature direction), then x cannot be a minimum.

Nonconvexity and minima

Solutions of $\min_{x \in \mathbb{R}^n} f(x)$

- For convex functions, $\nabla f(x^*) = 0 \Rightarrow x$ global minimum of f.
- Not true for general nonconvex functions.
- True if $\nabla^2 f(x^*) \succeq 0$ for some problems.

Nonconvexity and minima

Solutions of $\min_{x \in \mathbb{R}^n} f(x)$

- For convex functions, $\nabla f(x^*) = 0 \Rightarrow x$ global minimum of f.
- Not true for general nonconvex functions.
- True if $\nabla^2 f(x^*) \succeq 0$ for some problems.

A tool: Landscape analysis

- Look at points for which $\nabla f(x) = 0$;
- Use (especially) Hessian eigenvalues to assess the nature of these points!

Iskander's talk will focus on landscape!

From convergence to complexity

Worst-case complexity

Given $\epsilon \in (0,1)$, bound the **worst-case cost** of an algorithm to find x such that

$$\|\nabla f(x)\| \leq \epsilon, \qquad \lambda_{\min}(\nabla^2 f(x)) \geq -\epsilon.$$

Cost: Number of iterations, derivative evaluations, etc.

From convergence to complexity

Worst-case complexity

Given $\epsilon \in (0,1)$, bound the worst-case cost of an algorithm to find x such that

$$\|\nabla f(x)\| \leq \epsilon, \qquad \lambda_{\min}(\nabla^2 f(x)) \geq -\epsilon.$$

Cost: Number of iterations, derivative evaluations, etc.

Good complexity results

- Small dependencies in ϵ.
- Few accesses to $\nabla^2 f(x)$ or $\nabla^2 f(x) v$.

Such bounds will appear in Sadok's talk!

Hessian eigenvalues and convergence

Local convergence

- Close enough to a solution;
- For gradient-based methods, can be slowed on ill-conditioned problems.

Hessian eigenvalues and convergence

Local convergence

- Close enough to a solution;
- For gradient-based methods, can be slowed on ill-conditioned problems.

What about Hessians and eigenvalues?

- Using Hessians accounts for conditioning.
- Small eigenvalues make analysis more tricky.

See Irène's talk for more!

What I'd like to talk about

- **Estimating** a minimum (Hessian) matrix eigenvalue...

What I'd like to talk about

- **Estimating** a minimum (Hessian) matrix eigenvalue...
- ...using **randomized** techniques...

What I'd like to talk about

- **Estimating** a minimum (Hessian) matrix eigenvalue...
- ...using **randomized** techniques...
- ...for the **indefinite** setting.

What I'd like to talk about

- **Estimating** a minimum (Hessian) matrix eigenvalue...
- ...using **randomized** techniques...
- ...for the **indefinite** setting.

- ***Disclaimer:*** *Guarantees are in exact arithmetic.*
- ***Nice part:*** *Randomness is pretty mild.*
- ***Takeaway:*** *You can use conjugate gradient for that!*

Problem setup

In the background: $\min_{x \in \mathbb{R}^n} f(x)$

- Optimization procedure: $\{x_k\}_{k \in \mathbb{N}}$
- Would like to know if $\nabla^2 f(x_k)$ has negative eigenvalues.
- For complexity: **Sufficiently** negative eigenvalues matter!

Problem setup

In the background: $\min_{x \in \mathbb{R}^n} f(x)$

- Optimization procedure: $\{x_k\}_{k \in \mathbb{N}}$
- Would like to know if $\nabla^2 f(x_k)$ has negative eigenvalues.
- For complexity: **Sufficiently** negative eigenvalues matter!

A first problem

Given $A = A^{\mathrm{T}} \in \mathbb{R}^{n \times n}$ and $\epsilon > 0$,

1. Either find a d such that $d^{\mathrm{T}} A d \leq -\epsilon \|d\|^2$,
2. Or determine that $\lambda_{\min}(A) > -\epsilon$.

Problem setup

In the background: $\min_{x\in\mathbb{R}^n} f(x)$

- Optimization procedure: $\{x_k\}_{k\in\mathbb{N}}$
- Would like to know if $\nabla^2 f(x_k)$ has negative eigenvalues.
- For complexity: **Sufficiently** negative eigenvalues matter!

A first problem

Given $A = A^{\mathrm{T}} \in \mathbb{R}^{n\times n}$ and $\epsilon > 0$,

1. Either find a d such that $d^{\mathrm{T}} A d \leq -\epsilon \|d\|^2$,
2. Or determine that $\lambda_{\min}(A) > -\epsilon$.

So...computing $\lambda_{\min}(A)$?

An approximate problem

Given $A = A^{\mathrm{T}} \in \mathbb{R}^{n \times n}$ and $\epsilon > 0$,

1. Either find a d such that $d^{\mathrm{T}} A d \leq -\frac{1}{2}\epsilon \|d\|^2$,
2. Or determine that $\lambda_{\min}(A) > -\epsilon$.

The real problem

An approximate problem

Given $A = A^{\mathrm{T}} \in \mathbb{R}^{n \times n}$ and $\epsilon > 0$,

1. Either find a d such that $d^{\mathrm{T}} A d \leq -\frac{1}{2}\epsilon\|d\|^2$,
2. Or determine that $\lambda_{\min}(A) > -\epsilon$.

- No need for exact calculation of $\lambda_{\min}(A)$.
- Enough for optimization purposes.
- Probabilistic guarantee $\Rightarrow$ cheaper algorithms.

Minimum eigenvalue oracle

Definition

- Inputs: $A \in \mathbb{R}^{n \times n}$ symmetric, $\epsilon > 0$.
- Outputs:
  1. Either $(d, d^{\mathrm{T}} A d)$ such that $d^{\mathrm{T}} A d \le -\frac{\epsilon}{2} \|d\|^2$
  2. Or certificate that $\lambda_{\min}(A) > -\epsilon$.

Minimum eigenvalue oracle

Definition

- Inputs: $A \in \mathbb{R}^{n \times n}$ symmetric, $\epsilon > 0$.
- Outputs:
  1. Either $(d, d^\mathrm{T} A d)$ such that $d^\mathrm{T} A d \le -\frac{\epsilon}{2} \|d\|^2$
  2. Or certificate that $\lambda_{\min}(A) > -\epsilon$.

Basic example: Exact eigenvalue calculation

- Output: $\lambda_{\min}(A)$ and $d_{\min}$ such that $A d_{\min} = \lambda_{\min} d_{\min}$ if $\lambda_{\min}(A) \le -\epsilon$.
- Certificate: Deterministic.
- Cost: Exact eigenvalue/Full matrix calculation.

Krylov-based methods (1/2)

- **Krylov subspaces**

$$\mathcal{K}_j\left(A, b\right) = \mathrm{span}\left(b, A\, b, \dots, A^{j-1} b\right).$$

Krylov-based methods (1/2)

- **Krylov subspaces**

$$\mathcal{K}_j(A, b) = \mathrm{span}\left(b, A\,b, \ldots, A^{j-1}b\right).$$

- **Power method:**

$$d_{j+1} = Ad_j, \quad d_0 = b.$$

Krylov-based methods (1/2)

- **Krylov subspaces**

$$\mathcal{K}_j\left(A, b\right) = \mathrm{span}\left(b, A\, b, \dots, A^{j-1} b\right).$$

- **Power method:**

$$d_{j+1} = A d_j, \quad d_0 = b.$$

- **Lanczos method:**

$$d_{j+1} \in \underset{d \in \mathbb{R}^n}{\mathrm{argmin}} \left\{ \ \frac{1}{2} d^{\mathrm{T}} A d \quad \text{s.t.} \quad \|d\| = 1, \ d \in \mathcal{K}_j\left(A, b\right) \ \right\}, \quad d_0 = b.$$

Krylov-based methods (1/2)

- **Krylov subspaces**

$$\mathcal{K}_j\left(A, b\right) = \mathrm{span}\left(b, A\,b, \dots, A^{j-1}b\right).$$

- **Power method:**

$$d_{j+1} = Ad_j, \quad d_0 = b.$$

- **Lanczos method:**

$$d_{j+1} \in \underset{d \in \mathbb{R}^n}{\mathrm{argmin}} \left\{ \ \tfrac{1}{2} d^{\mathrm{T}} A d \quad \text{s.t.} \quad \|d\| = 1,\ d \in \mathcal{K}_j\left(A, b\right) \ \right\}, \quad d_0 = b.$$

- If $A \succ 0$ and $b \sim \mathcal{U}(\mathbb{S}^{n-1})$ can provide probabilistic guarantees for Power and Lanczos methods (1990s papers).

Krylov-based methods (1/2)

- **Krylov subspaces**

$$\mathcal{K}_j\left(A, b\right) = \mathrm{span}\left(b, A\,b, \dots, A^{j-1} b\right).$$

- **Power method:**

$$d_{j+1} = A d_j, \quad d_0 = b.$$

- **Lanczos method:**

$$d_{j+1} \in \operatorname*{argmin}_{d \in \mathbb{R}^n} \left\{ \ \tfrac{1}{2} d^{\mathrm{T}} A d \quad \text{s.t.} \quad \|d\| = 1, \ d \in \mathcal{K}_j\left(A, b\right) \ \right\}, \quad d_0 = b.$$

- If $A \succ 0$ and $b \sim \mathcal{U}(\mathbb{S}^{n-1})$ can provide probabilistic guarantees for Power and Lanczos methods (1990s papers).
- **Actually true when A is indefinite!**

Krylov-based methods (2/2)

Theorem (From Kuczyński & Woźniakowski '92)

Let $A \in \mathbb{R}^{n \times n}$ symmetric with $\|A\| \le M$, $\delta, \epsilon \in [0,1)$. Apply Lanczos to A and $b \sim \mathcal{U}(\mathbb{S}^{n-1})$. Then, after

$$J = \min\left\{ n, \left\lceil \frac{\ln(3n/\delta^2)}{2} \sqrt{\frac{M}{\epsilon}} \right\rceil \right\} \quad \text{iterations,}$$

- Either $d_{J+1}^{\mathrm{T}} A d_{J+1} \le -\frac{\epsilon}{2}$
- Or Lanczos certifies with probability at least $1-\delta$ that $A \succ -\epsilon I$.

Krylov-based methods (2/2)

Theorem (From Kuczyński & Woźniakowski '92)

Let $A \in \mathbb{R}^{n \times n}$ symmetric with $\|A\| \leq M$, $\delta, \epsilon \in [0,1)$. Apply Lanczos to A and $b \sim \mathcal{U}(\mathbb{S}^{n-1})$. Then, after

$$J = \min \left\{ n, \left\lceil \frac{\ln(3n/\delta^2)}{2} \sqrt{\frac{M}{\epsilon}} \right\rceil \right\} \quad \text{iterations,}$$

- Either $d_{J+1}^{\mathrm{T}} A d_{J+1} \leq -\frac{\epsilon}{2}$
- Or Lanczos certifies with probability at least $1-\delta$ that $A \succ -\epsilon I$.

- **Proof:** Apply 1992 result to $MI - A \succ 0$+use Krylov subspace invariance
$$\mathcal{K}_j(A, b) = \mathcal{K}_j(A + \gamma I, b) \quad \forall \gamma \in \mathbb{R}.$$
- For power method, bound worsens to $\frac{M}{\epsilon}$.

Conjugate gradient

Goal: Solve $Ax = b$, where $A = A^{\mathrm{T}} \succ 0$.

Conjugate gradient method

Init: Set $x_0 = 0_{\mathbb{R}^n}$, $r_0 = -b$, $p_0 = b$.

For $j = 0, 1, 2, \dots$

- if $p_j^{\mathrm{T}} A p_j \le 0$ terminate.
- Compute $x_{j+1} = x_j + \frac{\|r_j\|^2}{p_j^{\mathrm{T}} A p_j} p_j$ and $r_{j+1} = A x_{j+1} + b$.
- Set $p_{j+1} = -r_{j+1} + \frac{\|r_{j+1}\|^2}{\|r_j\|^2} p_j$.

- Only requires $v \mapsto Av$ ("matrix-free").
- Terminate in $\le n$ iterations in exact arithmetic when $H \succ 0$.
- Iteration j performed **as long as** $p_j^{\mathrm{T}} A p_j > 0$.

From Lanczos to CG

- CG and Lanczos work on the same **Krylov subspaces**.
- Negative curvature detected at the same iteration.

- CG and Lanczos work on the same **Krylov subspaces**.
- Negative curvature detected at the same iteration.

Theorem (R., O'Neill, Wright '20)

Given $\bar{A}$, b, let j be the smallest integer such that $\bar{A}|_{\mathcal{K}_j(\bar{A},b)} \not\succ 0$. Then,

- $d_{j+1}^{\mathrm{T}} \bar{A} d_{j+1} \le 0$ *(d_{j+1} Lanczos iterate)*;
- CG terminates due to $p_j^{\mathrm{T}} \bar{A} p_j \le 0$ *(p_j CG direction)*.

CG and minimum eigenvalues

Theorem (R., O'Neill, Wright 2020)

Let $A \in \mathbb{R}^{n \times n}$ symmetric with $\|A\| \leq M$, $\delta, \epsilon \in [0,1)$, and CG be applied to

$$\left(A + \tfrac{\epsilon}{2} I\right) y = b \quad \text{with} \quad b \sim \mathcal{U}(\mathbb{S}^{n-1}).$$

Then, after

$$J = \min \left\{ n, \left\lceil \frac{\ln(3n/\delta^2)}{2} \sqrt{\frac{M}{\epsilon}} \right\rceil \right\} \quad \text{iterations,}$$

- Either CG finds negative curvature explicitly: $p_J^{\mathrm{T}} \left(A + \frac{\epsilon}{2} I\right) p_J \leq 0$;
- Or it certifies with probability at least $1 - \delta$ that $A \succ -\epsilon I$.

Implications

What we have: CG routine to compute negative curvature directions.

What it brings us in optimization:

- Probabilistic certificate of second-order stationarity:

$$\|\nabla f(x_k)\| \leq \epsilon, \quad \lambda_{\min}(\nabla^2 f(x_k)) \geq -\epsilon.$$

Implications

What we have: CG routine to compute negative curvature directions.

What it brings us in optimization:

- Probabilistic certificate of second-order stationarity:

$$\|\nabla f(x_k)\| \leq \epsilon, \quad \lambda_{\min}(\nabla^2 f(x_k)) \geq -\epsilon.$$

- High-probability complexity bound $\mathcal{O}(\epsilon^{-7/2})$.

Implications

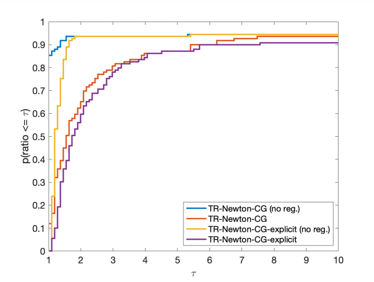
What we have: CG routine to compute negative curvature directions.

What it brings us in optimization:

- Probabilistic certificate of second-order stationarity:

$$\|\nabla f(x_k)\| \leq \epsilon, \quad \lambda_{\min}(\nabla^2 f(x_k)) \geq -\epsilon.$$

- High-probability complexity bound $\mathcal{O}(\epsilon^{-7/2})$.
- **In practice:** Only called once per algorithmic run.



Concluding with references

- Y. Carmon, J. C. Duchi, O. Hinder and A. Sidford, *Accelerated methods for nonconvex optimization*, SIAM Journal on Optimization, 2018.
- F. E. Curtis, D. P. Robinson, C. W. Royer and S. J. Wright, *Trust-region Newton-CG with strong second-order complexity guarantees for nonconvex optimization*, SIAM Journal on Optimization, 2021.
- J. Kuczyński and H. Woźniakowski, *Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start*, SIAM Journal on Matrix Analysis and Applications, 1992.
- C. W. Royer, M. O'Neill and S. J. Wright, *A Newton-CG algorithm with complexity guarantees for smooth unconstrained optimization*, Mathematical Programming, 2020.
- J. A. Tropp, *Randomized block Krylov methods for approximating extreme eigenvalues*, Numerische Mathematik, 2022.

Concluding with references

- Y. Carmon, J. C. Duchi, O. Hinder and A. Sidford, *Accelerated methods for nonconvex optimization*, SIAM Journal on Optimization, 2018.
- F. E. Curtis, D. P. Robinson, C. W. Royer and S. J. Wright, *Trust-region Newton-CG with strong second-order complexity guarantees for nonconvex optimization*, SIAM Journal on Optimization, 2021.
- J. Kuczyński and H. Woźniakowski, *Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start*, SIAM Journal on Matrix Analysis and Applications, 1992.
- C. W. Royer, M. O'Neill and S. J. Wright, *A Newton-CG algorithm with complexity guarantees for smooth unconstrained optimization*, Mathematical Programming, 2020.
- J. A. Tropp, *Randomized block Krylov methods for approximating extreme eigenvalues*, Numerische Mathematik, 2022.

Thank you!

`clement.royer@lamsade.dauphine.fr`