Spectral Preconditioning for Gradient Methods on Graded Non-convex Functions

Nikita Doikov (EPFL, Switzerland)

Joint work with **Sebastian U. Stich** (CISPA, Germany) and **Martin Jaggi** (EPFL, Switzerland)

Workshop on Algorithmic Optimization (ALGOPT2024)
Louvain-la-Neuve
August 27, 2024

Outline

Optimization Problem

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

- $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable; can be non-convex

The goal: find $\bar{\boldsymbol{x}}$ s.t. $\|\nabla f(\bar{\boldsymbol{x}})\| \leq \varepsilon$, for a small given $\varepsilon > 0$

Gradient Method:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \gamma_k \nabla f(\boldsymbol{x}_k), \qquad k \geq 0$$

+ cheap iterations
− slow convergence rates

This work. Improve the convergence by a special $\boldsymbol{P}_k = \boldsymbol{P}_k^\top \succ 0$ called *Spectral Preconditioning*:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{P}_k \nabla f(\boldsymbol{x}_k), \qquad k \geq 0$$

Main Complexity Parameters

The Hessian of the objective $\nabla^2 f(\boldsymbol{x}) \in \mathbb{R}^{n\times n}$ is a symmetric matrix $\Rightarrow$ all eigenvalues are real. The spectrum:

$$\lambda_1(\boldsymbol{x}) \quad \geq \quad \lambda_2(\boldsymbol{x}) \quad \geq \quad \ldots \quad \geq \quad \lambda_n(\boldsymbol{x})$$

describes the complexity of our problem.

1. **Non-convex problems.** Define $L_1 := \max_{\boldsymbol{x}} \lambda_1(\boldsymbol{x})$. Then the GM needs to do

   $$k \quad = \quad \frac{2L_1(f(\boldsymbol{x}_0) - f^\star)}{\varepsilon^2} \quad \text{iterations}$$

   to find $\|\nabla f(\bar{\boldsymbol{x}}_k)\| \leq \varepsilon$.

2. **Convex problems:** $\lambda_n(\boldsymbol{x}) \geq 0$.

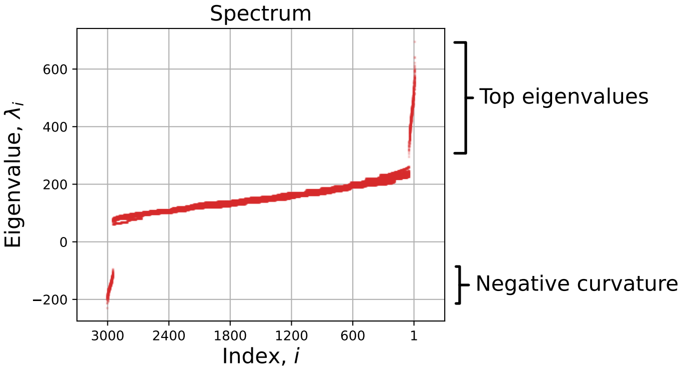
   Strongly convex: $\lambda_n(\boldsymbol{x}) \geq \mu > 0$.

   $\Rightarrow$ Better rates, e.g.: $k = \frac{L_1}{\mu} \log \frac{2L(f(\boldsymbol{x}_0) - f^\star)}{\varepsilon^2}$.

$$f(\boldsymbol{X}, \boldsymbol{Y}) \;=\; \tfrac{1}{2}\|\boldsymbol{X}\boldsymbol{Y} - \boldsymbol{C}\|^2, \qquad \boldsymbol{X} \in \mathbb{R}^{n\times r}, \boldsymbol{Y} \in \mathbb{R}^{r\times m},$$

where $\boldsymbol{C} \in \mathbb{R}^{n\times m}$ is a given data matrix.

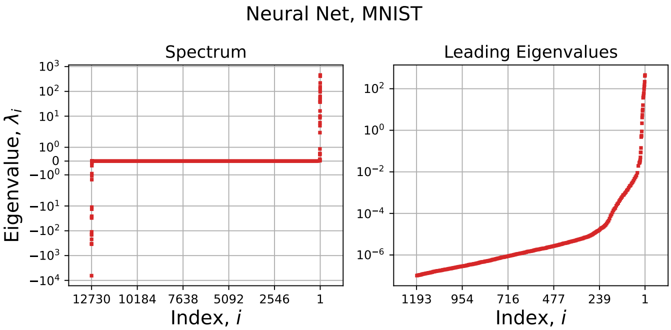
Thus, $f : \mathbb{R}^{(n+m)r} \to \mathbb{R}$, $\nabla^2 f(\boldsymbol{X}, \boldsymbol{Y}) \in \mathbb{R}^{(n+m)r\,\times\,(n+m)r}$.



Example: 2-layer NN

$$f(\boldsymbol{W}_1, \boldsymbol{b}_1, \boldsymbol{w}_2, b_2) \quad = \quad \frac{1}{m}\sum_{i=1}^{m} \ell_i(\langle \boldsymbol{w}_2, \sigma(\boldsymbol{W}_1 \boldsymbol{a}_i + \boldsymbol{b}_1)\rangle + b_2),$$

where $\{\boldsymbol{a}_i\}_{i=1}^{m}$ is a given dataset of features, $\sigma(\cdot)$ is an activation function, $\ell_i(\cdot)$ are losses.



Newton's Method

Let the Hessian be Lipschitz, $\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$.
$\Rightarrow$ Cubic regularization of Newton's method:

$$\mathcal{O}\Big(\tfrac{L^{1/2}(f(\boldsymbol{x}_0) - f^\star)}{\varepsilon^{3/2}} \Big)$$

second-order oracle calls to find $\|\nabla f(\bar{\boldsymbol{x}}_k)\| \leq \varepsilon$.

[Griewank, 1981; Nesterov-Polyak, 2006; Cartis-Gould-Toint, 2011]

- Computing $\nabla^2 f(\boldsymbol{x}_k)$ at every iteration
- Solving the cubic subproblem

Instead, we can use the **Gradient Regularization** technique:

$$\boldsymbol{x}_{k+1} \quad = \quad \boldsymbol{x}_k - \Big(\nabla^2 f(\boldsymbol{x}_k) + \sqrt{L\|\nabla f(\boldsymbol{x}_k)\|}\boldsymbol{I}\Big)^{-1} \nabla f(\boldsymbol{x}_k)$$

[Polyak, 2019; Ueda-Yamashita, 2014; Mishchenko, 2023; D-Nesterov, 2023]

+ **one linear system**
+ **fast global rates as for the Cubic Newton**
− **requires** $\nabla^2 f(\boldsymbol{x}_k) \succeq 0$ (see [Gratton-Jerad-Toint, 2023] for employing negative curvature)

Outline

I. Motivation

II. Grade of non-convexity

III. Spectral preconditioning

IV. Quasi-self-concordant functions

V. Experiments and conclusions

Spectral decomposition of the Hessian:

$$\nabla^2 f(\boldsymbol{x}) \quad \equiv \quad \sum_{i=1}^{n} \lambda_i(\boldsymbol{x}) \boldsymbol{u}_i(\boldsymbol{x}) \boldsymbol{u}_i(\boldsymbol{x})^\top,$$

where $\lambda_1(\boldsymbol{x}) \geq \ldots \geq \lambda_n(\boldsymbol{x})$ and $\boldsymbol{u}_1(\boldsymbol{x}), \ldots, \boldsymbol{u}_n(\boldsymbol{x}) \in \mathbb{R}^n$ are orthonormal eigenvectors (**NB**: several decompositions are possible, we can use any).

For a fixed $1 \leq \tau \leq n$, denote the Hessian of spectral order τ:

$$\nabla^2_\tau f(\boldsymbol{x}) \quad := \quad \sum_{i=1}^{\tau} \lambda_i(\boldsymbol{x}) \boldsymbol{u}_i(\boldsymbol{x}) \boldsymbol{u}_i(\boldsymbol{x})^\top \quad \in \mathbb{R}^{n \times n}$$

Definition. f is *non-convex of grade* τ if

$$\boxed{\nabla^2_\tau f(\boldsymbol{x}) \quad \succeq \quad 0, \qquad \forall \boldsymbol{x}}$$

$\Leftrightarrow$ top τ eigenvalues are non-negative everywhere:

$$\lambda_\tau(\boldsymbol{x}) \quad \geq \quad 0.$$

Main Properties

$$\boxed{f \in \mathcal{F}_\tau \;\overset{\text{def}}{\Leftrightarrow}\; \nabla_\tau^2 f(\boldsymbol{x}) \succeq 0.}$$

Nested family of functional cones:

$$\underset{\textbf{all functions}}{\mathcal{F}_0} \supset \mathcal{F}_1 \supset \dots \supset \mathcal{F}_{n-1} \supset \underset{\textbf{convex functions}}{\mathcal{F}_n}$$

- If $f \in \mathcal{F}_\tau$ then $\alpha f \in \mathcal{F}_\tau$ for any $\alpha \geq 0$.
- If $f \in \mathcal{F}_i$ and $g \in \mathcal{F}_j$ then

$$\begin{aligned} f + g &\in \mathcal{F}_{i+j-n}, \\ \mathrm{smax}\,(f, g) &\in \mathcal{F}_{i+j-n}, \end{aligned}$$

where $\mathrm{smax}\,(f,g)(\boldsymbol{x}) \overset{\text{def}}{=} \ln(e^{f(\boldsymbol{x})} + e^{g(\boldsymbol{x})})$.

$\Rightarrow$ summation with a convex function cannot decrease the grade.

- If $f \in \mathcal{F}_\tau(\mathbb{R}^n)$ and $g(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$ for $\boldsymbol{A} \in \mathbb{R}^{n \times m}$, $\boldsymbol{b} \in \mathbb{R}^n$,

$$g \in \mathcal{F}_{m-n+\tau}(\mathbb{R}^m).$$

Geometric Interpretation

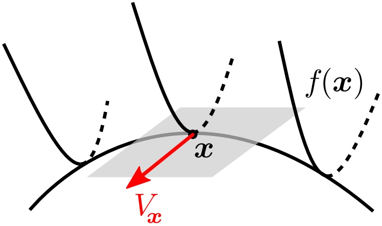
- $f \in \mathcal{F}_\tau$ cannot have strict local maxima for $\tau \geq 1$:

$$\max_{\boldsymbol{x} \in K} f(\boldsymbol{x}) \quad = \quad \max_{\boldsymbol{x} \in \partial K} f(\boldsymbol{x}).$$

- Sufficient condition for grade τ: Let for any $\boldsymbol{x}$ there exists a vector subspace $V_{\boldsymbol{x}} \subseteq \mathbb{R}^n$ with $\dim(V_{\boldsymbol{x}}) \geq \tau$ s.t.

$$f(\boldsymbol{x} + \boldsymbol{h}) \quad \geq \quad f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{h} \rangle, \qquad \forall \boldsymbol{h} \in V_{\boldsymbol{x}}.$$

Then $f \in \mathcal{F}_\tau$.



Example: Quadratics

Example. Let $f(\boldsymbol{x}) = \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}\rangle - \langle \boldsymbol{b}, \boldsymbol{x}\rangle$, for some $\boldsymbol{A} = \boldsymbol{A}^\top \in \mathbb{R}^{n\times n}$ with the top τ positive eigenvalues:

$$\lambda_1(\boldsymbol{A}) \;\geq\; \ldots \;\geq\; \lambda_\tau(\boldsymbol{A}) \;\geq\; 0.$$

Then $f \in \mathcal{F}_\tau$.

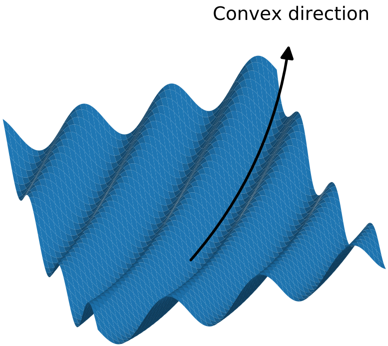
Example. Take $f(\boldsymbol{x}) = \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}\rangle - \langle \boldsymbol{b}, \boldsymbol{x}\rangle + \frac{\sigma}{p}\|\boldsymbol{x}\|^p$. Then $f \in \mathcal{F}_\tau$.

- For $p > 2$, a global solution $\boldsymbol{x}^\star = \operatorname*{argmin}_{\boldsymbol{x}} f(\boldsymbol{x})$ always exists.
- Important in applications to regularized second-order and high-order methods.

Example: Low-rank Vector Fields

Example. Let $f(\boldsymbol{x}) = \varphi(\langle \boldsymbol{u}(\boldsymbol{x}), \boldsymbol{x} \rangle)$, where $\varphi : \mathbb{R} \to \mathbb{R}$ is arbitrary, and $\boldsymbol{u} : \mathbb{R}^n \to \mathbb{R}^n$ is a low-rank differential mapping.

E.g., a constant vector field: $f(\boldsymbol{x}) = \varphi(\langle \boldsymbol{u}, \boldsymbol{x} \rangle)$ is non-convex of grade $n - 1$.



Function $f(x, y) = \sin(x + y) + q(x, y)$, *where* q *is convex.*

Example: Partial Convexity

- Let $f(\boldsymbol{x}, \boldsymbol{y}) : \mathbb{R}^{n+m} \to \mathbb{R}$ is such that for any fixed $\boldsymbol{y} \in \mathbb{R}^m$

$$f(\cdot, \boldsymbol{y}) \quad : \quad \mathbb{R}^n \to \mathbb{R}$$

is convex. Then f is non-convex of grade n.

Example. Diagonal NN: $f(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{x} \circ \boldsymbol{y} - \boldsymbol{c}\|^2 \; : \; \mathbb{R}^{2n} \to \mathbb{R}$ is non-convex of grade n.

Example. Matrix factorizations:

$$f(\boldsymbol{X}_1, \ldots \boldsymbol{X}_d) \quad = \quad \frac{1}{2}\|\boldsymbol{X}_1 \boldsymbol{X}_2 \cdots \boldsymbol{X}_d - \boldsymbol{C}\|_F^2, \quad \boldsymbol{X}_i \in \mathbb{R}^{n_i \times m_i},$$

is non-convex of grade $\tau = \max_{1 \le i \le d}[n_i \times m_i]$.

Example. Any deep model with convex losses.
$\tau =$ size of last layer.

Open question: better estimates of τ when the deepness is increasing?

Outline

I. Motivation

II. Grade of non-convexity

III. Spectral preconditioning

IV. Quasi-self-concordant functions

V. Experiments and conclusions

Spectral Preconditioning

Problem: $\boxed{\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x})}$.

Preconditioned Gradient Method:

$$\boldsymbol{x}_{k+1} \quad = \quad \boldsymbol{x}_k - \boldsymbol{P}_k \nabla f(\boldsymbol{x}_k), \qquad k \geq 0.$$

Main idea. Use $\boldsymbol{P}_k := (\boldsymbol{H}_k + \alpha_k \boldsymbol{I})^{-1}$ where $\boldsymbol{H}_k \approx \nabla_\tau^2 f(\boldsymbol{x}_k)$.

- $\alpha_k \geq 0$ is a regularization parameter (stepsize)
- $\tau = 0$: $\boldsymbol{H}_k = 0 \Rightarrow$ the gradient descent
- $\tau = n$: $\boldsymbol{H}_k = \nabla^2 f(\boldsymbol{x}_k) \Rightarrow$ regularized Newton
- Let $\tau = 1$. Take

$$\boldsymbol{H}_k \quad = \quad \lambda_1(\boldsymbol{x}_k)\boldsymbol{u}_1(\boldsymbol{x}_k)\boldsymbol{u}_1(\boldsymbol{x}_k)^\top$$

is a rank-1 matrix where $\boldsymbol{u}_1$ is the top eigenvector of $\nabla^2 f(\boldsymbol{x}_k)$.

- $\tau = 2$: top 2 eigenvectors, ...

Gradient Method with Spectral Preconditioning

Choose $\boldsymbol{x}_0 \in \mathbb{R}^n$ and $0 \leq \tau \leq n$.

For $k \geq 0$ **iterate:**

1. Estimate $\boldsymbol{H}_k \approx \nabla^2_\tau f(\boldsymbol{x}_k) \in \mathbb{R}^{n \times n}$
2. Perform the gradient step, for some $\alpha_k \geq 0$:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - (\boldsymbol{H}_k + \alpha_k \boldsymbol{I})^{-1} \nabla f(\boldsymbol{x}_k)$$

- Computing eigenvectors is difficult. For us, it is enough to use *inexact approximation*

$$\boldsymbol{H}_k \;\approx\; \nabla^2 f(\boldsymbol{x}_k).$$

- Power Method. Fast linear rate of convergence. The arithmetic complexity of each step is $\mathcal{O}(\tau^2 n)$

  More advanced: *Oja's* and *Lanczos iterations*.

- Low-rank representation $\boldsymbol{H}_k = \boldsymbol{V}_k \mathrm{Diag}\,(\boldsymbol{a}_k) \boldsymbol{V}_k^\top$, where $\boldsymbol{V}_k \in \mathbb{R}^{n \times \tau}$ has orthonormal columns. Then

$$(\boldsymbol{H}_k + \alpha_k \boldsymbol{I})^{-1} = \tfrac{1}{\alpha_k}\big[\boldsymbol{I} - \boldsymbol{V}_k\big(\boldsymbol{I} + \alpha_k \mathrm{Diag}\,(\boldsymbol{a}_k)^{-1}\big)^{-1}\boldsymbol{V}_k^\top\big].$$

  The Woodbury matrix identity: $\mathcal{O}(\tau n)$ cost.

Global Rate

- Let the Hessian be Lipschitz continuous,

$$\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\| \;\; \leq \;\; \boldsymbol{L}\|\boldsymbol{x} - \boldsymbol{y}\|, \qquad \forall \boldsymbol{x}, \boldsymbol{y}$$

- Denote $\boldsymbol{\sigma_\tau} \geq \max\{\lambda_{\tau+1}(\boldsymbol{x}), -\lambda_n(\boldsymbol{x})\}$
- Let $\delta \geq \|\boldsymbol{H}_k - \nabla^2_\tau f(\boldsymbol{x}_k)\|$

Theorem. Let $f \in \mathcal{F}_\tau$ for some fixed $0 \leq \tau \leq n$. Set

$$\alpha_k \;\; := \;\; \sqrt{\tfrac{L\|\nabla f(\boldsymbol{x}_k)\|}{2}} + \sigma_\tau + \delta.$$

Then, to ensure $\min_{1 \leq i \leq k} \|\nabla f(\boldsymbol{x}_i)\| \leq \varepsilon$ it is enough to choose

$$k \;\; = \;\; \Big\lceil 8(f(\boldsymbol{x}_0) - f^\star) \cdot \Big(\sqrt{\tfrac{\boldsymbol{L}}{2}} \tfrac{1}{\varepsilon^{3/2}} + \tfrac{\boldsymbol{\sigma_\tau} + \boldsymbol{\delta}}{\varepsilon^2} \Big) \; + \; 2 \ln \tfrac{\|\nabla f(\boldsymbol{x}_0)\|}{\varepsilon} \Big\rceil$$

- Increasing τ we improve the parameter $\boldsymbol{\sigma_\tau}$. E.g. for $\tau := 1$ the method does not depend on λ_1.
- $\tau := n$ gives the global rate of the Cubic Newton

Outline

I. Motivation

II. Grade of non-convexity

III. Spectral preconditioning

IV. Quasi-self-concordant functions

V. Experiments and conclusions

Convex Problems

Let $\nabla^2 f(\boldsymbol{x}) \succeq 0$ (i.e. $f \in \mathcal{F}_n$).

Can we improve convergence rate? Yes.

- Previous smoothness condition – the Hessian is Lipschitz:

$$\nabla^3 f(\boldsymbol{x})[\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{v}] \quad \leq \quad L\|\boldsymbol{u}\|^2\|\boldsymbol{v}\|, \qquad \forall \boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n.$$

- *Refined smoothness condition.* Assume that f is quasi-self-concordant [Bach, 2010]:

$$\nabla^3 f(\boldsymbol{x})[\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{v}] \quad \leq \quad M\|\boldsymbol{u}\|_{\boldsymbol{x}}^2\|\boldsymbol{v}\|, \qquad \forall \boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n,$$

  where $\|\boldsymbol{u}\|_{\boldsymbol{x}}^2 := \langle \nabla^2 f(\boldsymbol{x})\boldsymbol{u}, \boldsymbol{u} \rangle$ is the local norm.

A direct consequence – the Hessian is stable [Karimireddy-Stich-Jaggi, 2018], $\forall \boldsymbol{x}, \boldsymbol{y}$:

$$\nabla^2 f(\boldsymbol{y}) e^{-M\|\boldsymbol{x}-\boldsymbol{y}\|} \quad \preceq \quad \nabla^2 f(\boldsymbol{x}) \quad \preceq \quad \nabla^2 f(\boldsymbol{y}) e^{M\|\boldsymbol{y}-\boldsymbol{x}\|}.$$

Examples

Example 0. Quadratic functions: $\boxed{M = 0}$.

Example 1. $\varphi(x) = e^x$. Then $\varphi^{(p)}(x) = e^x$. Hence, $\boxed{M = 1}$.

Example 2. $\varphi(x) = \ln(1 + e^x)$, we have $\boxed{M = 1}$.

Therefore, the **logistic** and **exponential** regressions

$$f(\boldsymbol{x}) \quad = \quad \tfrac{1}{m} \sum_{i=1}^{m} \varphi(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle)$$

are quasi-SC.

Example 3. Soft maximum: $f(\boldsymbol{x}) = \mu \ln\Big(\sum_{i=1}^{m} \exp\big(\frac{\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle - b_i}{\mu} \big) \Big)$ is quasi-SC with $\boxed{M = \frac{2}{\mu}}$.

Example 4. Matrix scaling, $\boldsymbol{A} \in \mathbb{R}^{n \times n}_{+}$: $f(\boldsymbol{x}, \boldsymbol{y}) = \sum_{1 \leq i,j \leq n} A_{ij} e^{x_i - x_j}$ is quasi-SC with $\boxed{M = \sqrt{2}}$.

Gradient Regularization of Newton's Method

Consider the following full Newton method with regularization:

$$\boxed{\mathbf{x}_{k+1} \;=\; \mathbf{x}_k - \left(\nabla^2 f(\mathbf{x}_k) + \alpha_k \mathbf{I}\right)^{-1} \nabla f(\mathbf{x}_k)}$$

Theorem. [D, 2023] Set

$$\boxed{\alpha_k \;:=\; M\|\nabla f(\mathbf{x}_k)\|}$$

Then, we have the global linear rate:

$$f(\mathbf{x}_k) - f^\star \;\leq\; \exp\left(-\frac{k}{8MD}\right)\left(f(\mathbf{x}_0) - f^\star\right) + \exp\left(-\tfrac{k}{4}\right)\|\nabla f(\mathbf{x}_0)\| D,$$

where $D := \max\{\|\mathbf{x} - \mathbf{x}^\star\| \;:\; f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$.

$\Rightarrow$ **the global complexity:** $\mathcal{O}\left(MD \ln \frac{1}{\varepsilon}\right)$ to find $f(\mathbf{x}_k) - f^\star \leq \varepsilon$

NB: compare with

- The Gradient Method: $\mathcal{O}(\frac{\lambda_1 D^2}{\varepsilon})$
- The Cubic Newton: $\mathcal{O}(\sqrt{\frac{LD^3}{\varepsilon}})$

Convergence on Convex Problems

- Spectral Preconditioning:

$$\boxed{\mathbf{x}_{k+1} \;=\; \mathbf{x}_k - \big(\mathbf{H}_k + \alpha_k \mathbf{I}\big)^{-1} \nabla f(\mathbf{x}_k), \qquad \mathbf{H}_k \approx \nabla^2_\tau f(\mathbf{x}_k)}$$

Theorem. Let f be strongly convex and quasi-SC with $M > 0$. Set

$$\alpha_k \;:=\; M\|\nabla f(\mathbf{x}_k)\| + \lambda_{\tau+1} + \delta.$$

Then, to ensure $f(\mathbf{x}_k) - f^\star \leq \varepsilon$ it is enough to choose

$$k \;=\; 4\left\lceil \left(\boldsymbol{MD} + \frac{\boldsymbol{\lambda_{\tau+1}+\delta}}{\boldsymbol{2\lambda_n}} \right) \ln \frac{f(\mathbf{x}_0) - f^\star}{\varepsilon} \;+\; \ln \frac{\|\nabla f(\mathbf{x}_0)\| D}{\varepsilon} \right\rceil$$

- The rate is linear and the condition number is $\boxed{\frac{\boldsymbol{\lambda_{\tau+1}}}{\boldsymbol{\lambda_n}}}$
- Increasing τ we cut off the top τ eigenvalues of the spectrum

Example: Quadratic Problem

- Let $f(\boldsymbol{x}) = \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}\rangle - \langle \boldsymbol{b}, \boldsymbol{x}\rangle$ with $\boldsymbol{A} = \boldsymbol{A}^\top \succ 0$. $(M = 0)$

1. The classical gradient descent: $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \gamma \nabla f(\boldsymbol{x}_k)$. The number of iterations (matrix-vector products):

$$\mathcal{O}(\tfrac{\lambda_1}{\lambda_n} \log \tfrac{1}{\varepsilon}) \qquad \textbf{(*)}$$

2. Our Spectral Preconditioning for $\tau = 1$:
$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \left(a_1 \boldsymbol{v}_1 \boldsymbol{v}_1^\top + \alpha \boldsymbol{I}\right)^{-1} \nabla f(\boldsymbol{x}_k)$. The number of iterations:

$$\mathcal{O}(\tfrac{\lambda_2}{\lambda_n} \log \tfrac{1}{\varepsilon}).$$

The cost of computing $\boldsymbol{v}_1 \approx \boldsymbol{u}_1(\boldsymbol{A})$ by the Power Method is $\tilde{\mathcal{O}}(\frac{\lambda_1}{\lambda_1 - \lambda_2})$. Hence, the total complexity

$$\tilde{\mathcal{O}}\left(\frac{\lambda_1}{\lambda_n} \cdot \frac{\lambda_2}{\lambda_1 - \lambda_2}\right)$$

can be much better than in **(*)**, when $\lambda_1 \gg \lambda_2$.

Outline

I. Motivation

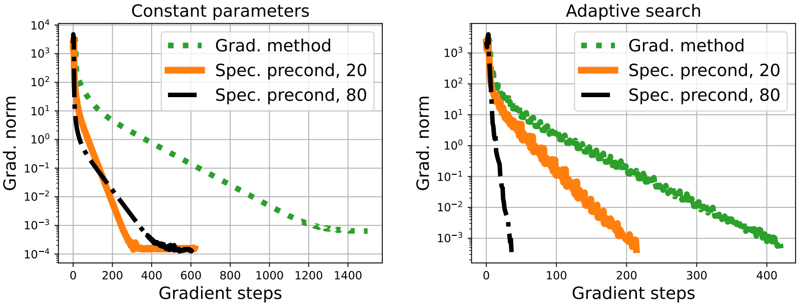
II. Grade of non-convexity

III. Spectral preconditioning

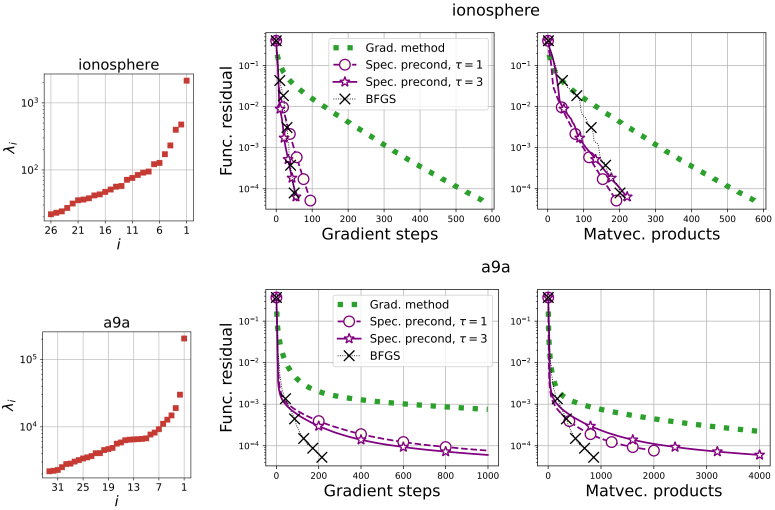
IV. Quasi-self-concordant functions

V. Experiments and conclusions

Matrix Factorization



Logistic Regression



Conclusions

- The spectrum of the Hessian $\nabla^2 f(\boldsymbol{x})$ often determines the complexity of the problem.
- In practice – a specific distribution of eigenvalues
  $\Rightarrow$ refined problem classes and more advanced methods
- **Graded non-convexity:** most of the eigenvalues are usually positive
- **Spectral Preconditioning:** we can cut the large gaps between the top eigenvalues

Reference:

- Doikov, N., Stich, S.U., Jaggi, M., ICML 2024 (*International Conference on Machine Learning*) Spectral Preconditioning for Gradient Methods on Graded Non-convex Functions

Open problems

- Other Notions of Convexity (weak, plurisubharmonic functions, convexity with respect to a linear group, ...)
- Grade of non-convexity for deep models
- Practical performance (efficient implementation of Hessian-vector products)
- Refined specification of the problem / different methods
  - **Kernel ridge regression** [Ma-Belkin, 2017]
  - **Matrix factorizations** [Zhang-Fattahi-Zhang, 2021, 2023; Ma-Xu-Tong-Chi, 2023]
  - **Heavy-ball method** [Scieur-Pedregosa, 2020]
  - **Polynomial preconditioning** [D-Rodomanov, 2023]
- Local superlinear convergence $\Rightarrow$ a bridge to quasi-Newton methods

Thank you for your attention!