Polynomial Preconditioning for Gradient Methods

Nikita Doikov (EPFL, Switzerland)

Joint work with **Anton Rodomanov** (CISPA, Germany)

FGS Conference on Optimization, Gijón

June 20, 2024

Outline

I. Introduction: preconditioning of gradient methods

II. Symmetric polynomial preconditioning

III. Krylov subspace preconditioning

IV. Experiments and conclusions

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

- $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable

Assume that f is (strongly) convex and has Lipschitz gradient $\Rightarrow$ there exist $0 \leq \lambda_{\min} \leq \lambda_{\max}$ s.t.

$$\lambda_{\min}\boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq \lambda_{\max}\boldsymbol{I}, \qquad \forall \boldsymbol{x} \in \mathbb{R}^n$$

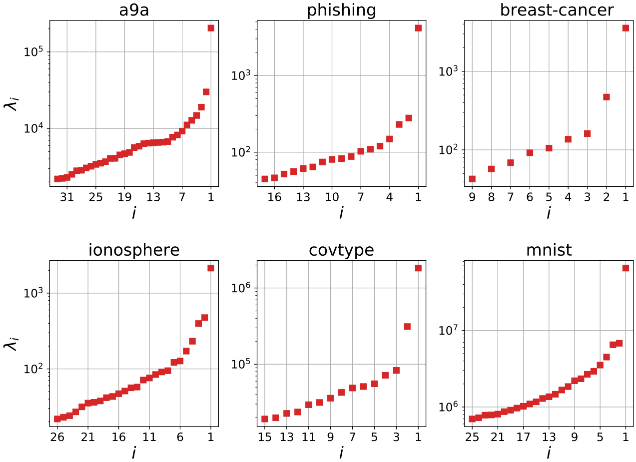
Gradient Method. Iterate, for $k \geq 0$:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \gamma\nabla f(\boldsymbol{x}_k)$$

The rate of convergence depends on the extremal characteristics of the spectrum. To find $f(\boldsymbol{x}_k) - f^\star \leq \varepsilon$ we need

- $\mathcal{O}(\frac{\lambda_{\max}}{\lambda_{\min}} \ln \frac{1}{\varepsilon})$ gradient steps **(strongly convex functions)**
- $\mathcal{O}\left(\frac{\lambda_{\max}\|\boldsymbol{x}_0-\boldsymbol{x}^\star\|^2}{\varepsilon}\right)$ gradient steps **(convex functions)**

- Distribution of the top eigenvalues:



- There are large gaps between top eigenvalues $\Rightarrow$ slow convergence

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

Fix matrix $\boldsymbol{B} = \boldsymbol{B}^\top \succ 0$ (curvature matrix)

Our assumption: for some $0 \le \mu \le L$, we have

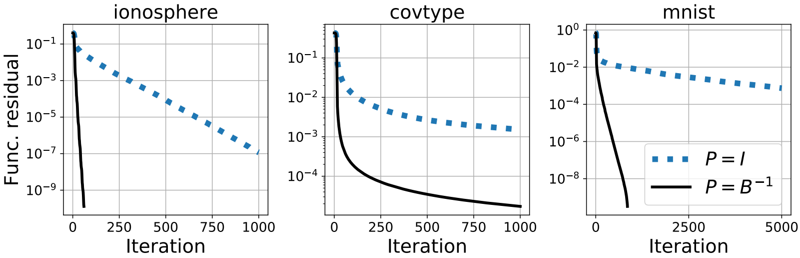
$$\mu \boldsymbol{B} \preceq \nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{B}, \qquad \forall \boldsymbol{x} \in \mathbb{R}^n$$

i.e. the function f is (strongly) convex and has Lipschitz gradient w.r.t. the induced norm $\|\boldsymbol{x}\|_{\boldsymbol{B}} := \langle \boldsymbol{B}\boldsymbol{x}, \boldsymbol{x}\rangle^{1/2}$

Example 1. Let $f(\boldsymbol{x}) = \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}\rangle - \langle \boldsymbol{b}, \boldsymbol{x}\rangle$. Then $\boldsymbol{B} := \boldsymbol{A}$ and $\mu = L = 1$.

Example 2. Let $f(\boldsymbol{x}) = g(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$. Assume that $g(\cdot)$ is μ-strongly convex and L-smooth. Then $\boldsymbol{B} := \boldsymbol{A}^\top \boldsymbol{A}$.

- Intuitively, $\boldsymbol{B}$ is the best uniform approximation of the Hessian



Gradient method: $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \gamma \nabla f(\boldsymbol{x}_k)$

Newton-type method: $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \gamma \boldsymbol{B}^{-1} \nabla f(\boldsymbol{x}_k)$

+ much faster convergence

− expensive to use $\boldsymbol{B}^{-1}$

This work: $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \gamma \boldsymbol{P} \nabla f(\boldsymbol{x}_k)$, where $\boldsymbol{P} \approx \boldsymbol{B}^{-1}$

Composite optimization problem: $\boxed{\min_{\boldsymbol{x}} F(\boldsymbol{x}) = f(\boldsymbol{x}) + \psi(\boldsymbol{x})}$

- ψ is a simple component (e.g. indicator of a convex set)

Define, for some $M > 0$ and preconditioner $\boldsymbol{P} = \boldsymbol{P}^\top \succ 0$:

$$\mathrm{GradStep}_{M,\boldsymbol{P}}(\boldsymbol{x}, \boldsymbol{g}) \stackrel{\text{def}}{=} \operatorname*{argmin}_{y} \Big\{ \langle \boldsymbol{g}, \boldsymbol{y} \rangle + \frac{M}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2_{\boldsymbol{P}^{-1}} + \psi(\boldsymbol{y}) \Big\}$$

Preconditioned Gradient Method. Iterate, $k \geq 0$:

$$\boldsymbol{x}_{k+1} \quad = \quad \mathrm{GradStep}_{M,\boldsymbol{P}}(\boldsymbol{x}_k, \nabla f(\boldsymbol{x}_k))$$

Theorem. Let $\alpha \boldsymbol{B}^{-1} \preceq \boldsymbol{P} \preceq \beta \boldsymbol{B}^{-1}$ and set $M := \beta L$. Then

$$F(\boldsymbol{x}_k) - F^\star \quad \leq \quad \Big(1 - \frac{1}{4} \frac{\boldsymbol{\alpha}}{\boldsymbol{\beta}} \frac{\mu}{L}\Big)^k (F(\boldsymbol{x}_0) - F^\star) \qquad \textbf{(strongly convex)}$$

$$F(\boldsymbol{x}_k) - F^\star \quad \leq \quad \frac{\boldsymbol{\beta}}{\boldsymbol{\alpha}} \frac{L\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2_{\boldsymbol{B}}}{k} \qquad \textbf{(convex functions)}$$

- We can accelerate the gradient steps! [Nesterov, 1983]

Preconditioned Fast Gradient Method. Set $\boldsymbol{v}_0 = \boldsymbol{x}_0$, $A_0 = 0$. Iterate, $k \geq 0$:

1. Find a_{k+1} from eq. $\frac{Ma_{k+1}^2}{A_{k+1}} = 1 + \alpha\mu A_{k+1}$, $A_{k+1} = A_k + a_{k+1}$
2. Choose $H_k = \frac{1+\alpha\mu A_{k+1}}{a_{k+1}}$, $\theta_k = \frac{a_{k+1}}{A_{k+1}}$, $\omega_k = \frac{\rho}{H_k}$, $\gamma_k = \frac{\omega_k(1-\theta_k)}{1-\omega_k\theta_k}$
3. Set $\bar{\boldsymbol{v}}_k = (1-\gamma_k)\boldsymbol{v}_k + \gamma_k \boldsymbol{x}_k$
4. Set $\boldsymbol{y}_k = (1-\theta_k)\boldsymbol{x}_k + \theta_k \bar{\boldsymbol{v}}_k$
5. Compute $\boldsymbol{v}_{k+1} = \mathrm{GradStep}_{M,\boldsymbol{P}}(\bar{\boldsymbol{v}}_k, \nabla f(\boldsymbol{y}_k))$
6. $\boldsymbol{x}_{k+1} = (1-\theta_k)\boldsymbol{x}_k + \theta_k \boldsymbol{v}_{k+1}$

Theorem. Let $\alpha \boldsymbol{B}^{-1} \preceq \boldsymbol{P} \preceq \beta \boldsymbol{B}^{-1}$ and set $M := \beta L$. Then

$$F(\boldsymbol{x}_k) - F^\star \quad \leq \quad \left(1 - \sqrt{\frac{\alpha}{\beta}\frac{\mu}{L}}\right)^k \frac{\beta}{\alpha} \frac{L\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_{\boldsymbol{B}}^2}{2} \qquad \textbf{(strongly convex)}$$

$$F(\boldsymbol{x}_k) - F^\star \quad \leq \quad \frac{\beta}{\alpha} \frac{2L\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_{\boldsymbol{B}}^2}{k^2} \qquad \textbf{(convex functions)}$$

- Standard gradient methods:

$$\boldsymbol{P} \quad := \quad \boldsymbol{I}$$

$\Rightarrow$ the condition number $\frac{\beta}{\alpha}$ is the largest: $\boxed{\frac{\beta}{\alpha} = \frac{\lambda_1}{\lambda_n}}$ where

$$\lambda_1 \quad \geq \quad \dots \quad \geq \quad \lambda_n \quad \text{are eigenvalues of } \boldsymbol{B}$$

- NB: for $\boldsymbol{P} := \boldsymbol{B}^{-1}$ we have $\frac{\beta}{\alpha} = 1$ (but too expensive)

This work: new family of preconditioners that provably improves β/α for non-uniform spectrum.

- **Example:** set $\boxed{\boldsymbol{P} := \mathrm{tr}\,(\boldsymbol{B})\boldsymbol{I} - \boldsymbol{B}}$

Then

$$\frac{\beta}{\alpha} \quad \approx \quad \frac{\lambda_2}{\lambda_n}, \qquad \text{when} \qquad \lambda_1 \gg \lambda_2$$

Outline

I. Introduction: preconditioning of gradient methods

II. Symmetric polynomial preconditioning

III. Krylov subspace preconditioning

IV. Experiments and conclusions

Symmetric polynomial preconditioner

- Family of symmetric matrices $\{\boldsymbol{P}_\tau\}_{0\leq\tau\leq n-1}$
- Set $\boldsymbol{P}_0 := \boldsymbol{I}$

Define $\boldsymbol{U}_\tau := \mathrm{tr}\,(\boldsymbol{B}^\tau)\boldsymbol{I} - \boldsymbol{B}^\tau$ and set recursively

$$\boxed{\boldsymbol{P}_\tau \quad := \quad \frac{1}{\tau}\sum_{i=1}^{\tau}(-1)^{i-1}\boldsymbol{P}_{\tau-i}\boldsymbol{U}_i}$$

We have

- $\boldsymbol{P}_1 = \mathrm{tr}\,(\boldsymbol{B})\boldsymbol{I} - \boldsymbol{B}$
- $\boldsymbol{P}_2 = \frac{1}{2}\mathrm{tr}\,(\boldsymbol{P}_1\boldsymbol{B})\boldsymbol{I} - \boldsymbol{P}_1\boldsymbol{B} = \frac{1}{2}[\mathrm{tr}\,(\boldsymbol{B})^2 - \mathrm{tr}\,(\boldsymbol{B}^2)]\boldsymbol{I} - \mathrm{tr}\,(\boldsymbol{B})\boldsymbol{B} + \boldsymbol{B}^2$
- ...
- $\boldsymbol{P}_\tau = p_\tau(\boldsymbol{B})$ where $p_\tau(\cdot)$ is a polynomial of degree τ
- ...
- $\boldsymbol{P}_{n-1} \propto \boldsymbol{B}^{-1}$

Main lemma

For $\boldsymbol{a} \in \mathbb{R}^{n-1}$ denote by $\sigma_0(\boldsymbol{a}), \ldots, \sigma_{n-1}(\boldsymbol{a})$ the elementary symmetric polynomials in $n-1$ variables. Thus,

$$\sigma_\tau(\boldsymbol{a}) \quad := \quad \sum_{1 \leq i_1 < \ldots < i_\tau \leq n-1} a_{i_1} \ldots a_{i_\tau}$$

Fix the spectral decomposition, with $\boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{I}$:

$$\boldsymbol{B} \quad = \quad \boldsymbol{Q} \mathrm{Diag}\,(\lambda_1, \ldots, \lambda_n) \boldsymbol{Q}^\top$$

Lemma. It holds:

$$\boldsymbol{P}_\tau \quad = \quad \boldsymbol{Q} \mathrm{Diag}\,(\sigma_\tau(\boldsymbol{\lambda}_{-1}), \ldots, \sigma_\tau(\boldsymbol{\lambda}_{-n})) \boldsymbol{Q}^\top$$

where $\boldsymbol{\lambda}_{-i} \in \mathbb{R}^{n-1}$ contains all eigenvalues except λ_i

- In particular, $\boldsymbol{P}_{n-1} = \det(\boldsymbol{B})\boldsymbol{B}^{-1}$

Approximation quality

Theorem. For any τ, we have

$$\lambda_n \sigma_\tau(\boldsymbol{\lambda}_{-n}) \boldsymbol{B}^{-1} \quad \preceq \quad \boldsymbol{P}_\tau \quad \preceq \quad \lambda_1 \sigma_\tau(\boldsymbol{\lambda}_{-1}) \boldsymbol{B}^{-1}.$$

$\Rightarrow$ the condition number $\frac{\beta}{\alpha}$ is bounded as

$$\frac{\beta}{\alpha} \quad = \quad \frac{\lambda_1}{\lambda_n} \cdot \xi_\tau(\boldsymbol{\lambda}), \quad \text{where} \quad \xi_\tau(\boldsymbol{\lambda}) \quad := \quad \frac{\sigma_\tau(\boldsymbol{\lambda}_{-1})}{\sigma_\tau(\boldsymbol{\lambda}_{-n})} \quad \leq \quad 1$$

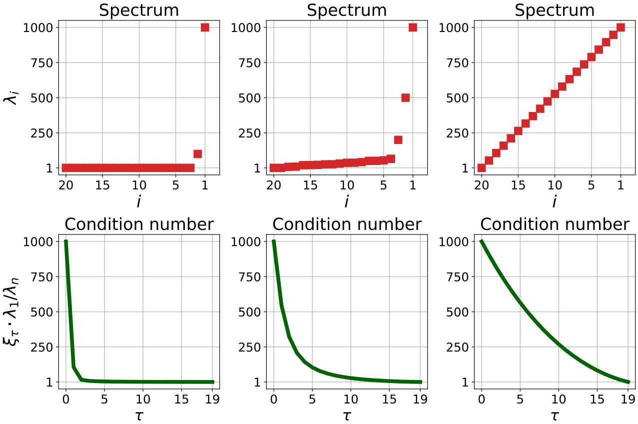
- $\xi_0(\boldsymbol{\lambda}) = 1$, $\xi_{n-1}(\boldsymbol{\lambda}) = \frac{\lambda_n}{\lambda_1}$
- $\xi_\tau(\boldsymbol{\lambda})$ monotonically decreases with τ
- $\xi_\tau(\boldsymbol{\lambda}) \to 0$ when $\frac{\lambda_1}{\lambda_{\tau+1}} \to \infty$

More precisely,

$$\xi_\tau(\boldsymbol{\lambda}) \quad \leq \quad \frac{\lambda_n + \sum_{i=\tau+1}^{n-1} \lambda_i}{\lambda_1 + \sum_{i=\tau+1}^{n-1} \lambda_i}$$

Improvement of the spectrum

- Top: different distributions of eigenvalues of $\boldsymbol{B}$



- Bottom: improvement of the condition number when using the preconditioner $\boldsymbol{P}_\tau$ of higher order $0 \leq \tau < n$

Let $S \subseteq \{1, \ldots, n\}$ be random subset of coordinates

Denote $\boldsymbol{I}_S \in \mathbb{R}^{n \times (\tau+1)}$ — the matrix obtained from $\boldsymbol{I} \in \mathbb{R}^{n \times n}$ by keeping only the columns from S

- $\boldsymbol{B}_{S \times S} := \boldsymbol{I}_S \boldsymbol{B} \boldsymbol{I}_S \in \mathbb{R}^{(\tau+1) \times (\tau+1)}$
- Then $\boldsymbol{I}_S (\boldsymbol{B}_{S \times S})^{-1} \boldsymbol{I}_S \approx \boldsymbol{B}^{-1}$

Theorem.

$$\boldsymbol{P}_\tau \quad \propto \quad \mathbb{E}_{S \sim \mathrm{Vol}_{\tau+1}(\boldsymbol{B})} \Big[\, \boldsymbol{I}_S (\boldsymbol{B}_{S \times S})^{-1} \boldsymbol{I}_S \, \Big]$$

where $\mathrm{Vol}_{\tau+1}(\boldsymbol{B})$ is the volume sampling (choose S with probability $\propto \det(\boldsymbol{B}_{S \times S})$)

[Rodomanov-Kropotov, 2020]

- Coordinate method with volume sampling:
  $\boldsymbol{x}^+ = \boldsymbol{x} - \gamma \boldsymbol{I}_S (\boldsymbol{B}_{S \times S})^{-1} \boldsymbol{I}_S \nabla f(\boldsymbol{x})$

Outline

I. Introduction: preconditioning of gradient methods

II. Symmetric polynomial preconditioning

III. Krylov subspace preconditioning

IV. Experiments and conclusions

We know that $\boldsymbol{P}_\tau = p_\tau(\boldsymbol{B})$ for some polynomial p_τ

- Can we find a better polynomial?

Set

$$\boldsymbol{P}_{\boldsymbol{a}} \;=\; a_0 \boldsymbol{I} + a_1 \boldsymbol{B} + \ldots + a_\tau \boldsymbol{B}^\tau, \qquad \boldsymbol{a} \;\in\; \mathbb{R}^{\tau+1}$$

- Preconditioned gradient step: $\boldsymbol{x}^+ = \boldsymbol{x} - \boldsymbol{P}_a \nabla f(\boldsymbol{x})$

By our assumption, we have

$$f(\boldsymbol{x}^+) \;\leq\; f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{x}^+ - \boldsymbol{x} \rangle + \tfrac{L}{2} \|\boldsymbol{x}^+ - \boldsymbol{x}\|_{\boldsymbol{B}}^2 \qquad (*)$$

Idea: minimize $(*)$ with respect to $\boldsymbol{a}$ $\Leftrightarrow$ project $\frac{1}{L}\boldsymbol{B}^{-1}\nabla f(\boldsymbol{x})$ onto the *Krylov subspace*:

$$\boldsymbol{x}^+ - \boldsymbol{x} \;=\; \operatorname*{argmin}_{\boldsymbol{h} \in \mathcal{K}_\tau} \|\boldsymbol{h} + \tfrac{1}{L}\boldsymbol{B}^{-1}\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2,$$

where $\mathcal{K}_\tau = \mathrm{span}\{\nabla f(\boldsymbol{x}), \boldsymbol{B}\nabla f(\boldsymbol{x}), \ldots, \boldsymbol{B}^\tau \nabla f(\boldsymbol{x})\}$

Gradient method with Krylov preconditioning

Iterate, $k \geq 0$:

1. Form the Gram matrix $\boldsymbol{A}_k \in \mathbb{R}^{(\tau+1)\times(\tau+1)}$:

$$\left[\boldsymbol{A}_k\right]^{(i,j)} \;=\; L \cdot \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{B}^{i+j+1} \nabla f(\boldsymbol{x}_k) \rangle$$

2. Form the vector $\boldsymbol{g}_k \in \mathbb{R}^{\tau+1}$:

$$\left[\boldsymbol{g}_k\right]^{(i)} \;=\; \langle \nabla f(\boldsymbol{x}), \boldsymbol{B}^{i} \nabla f(\boldsymbol{x}) \rangle$$

3. Compute $\boldsymbol{a}_k = \boldsymbol{A}_k^{-1} \boldsymbol{g}_k$
4. Set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{P}_{\boldsymbol{a}_k} \nabla f(\boldsymbol{x}_k)$

Theorem. Let $\boldsymbol{P} \succ 0$ be any preconditioner that is given by a polynomial of degree τ: $\boldsymbol{P} = p_\tau(\boldsymbol{B})$, and $\alpha \boldsymbol{B}^{-1} \preceq \boldsymbol{P} \preceq \beta \boldsymbol{B}^{-1}$.

Then, the method achieves the corresponding rate of GM with $\frac{\boldsymbol{\beta}}{\boldsymbol{\alpha}}$

Bounds on the condition number

- The method automatically chooses the optimal polynomial

Example 1. Set

$$q_\tau(s) \quad = \quad \left(1 - \tfrac{s}{\lambda_1}\right)\left(1 - \tfrac{s}{\lambda_2}\right) \cdot \ldots \cdot \left(1 - \tfrac{s}{\lambda_\tau}\right)$$

and $p_\tau(s) := \frac{1+q_\tau(s)\cdot(\alpha s-1)}{s}$ with $\alpha := \frac{2}{\lambda_{\tau+1}+\lambda_n}$. Then,

$$\boxed{\frac{\beta}{\alpha} \quad \leq \quad \frac{\lambda_{\tau+1}}{\lambda_n}}$$

Example 2. Fix $0 < \epsilon < 1$, let $\tau := \left\lceil \sqrt{\frac{\lambda_1}{\lambda_n}} \ln \frac{8}{\epsilon} \right\rceil$ and set $p_\tau(s) := \frac{1-Q_\tau(s)}{s}$, where $Q_\tau(\cdot)$ is a normalized Chebyshev polynomial of the first kind of degree τ. Then,

$$\boxed{\frac{\beta}{\alpha} \quad \leq \quad 1 + \epsilon}$$

Polynomial preconditioning: summary

Symmetric polynomial preconditioning

- Family of fixed preconditioners $\boldsymbol{P}_\tau$, $0 \leq \tau \leq n-1$
- Improve the condition number when $\lambda_\tau \gg \lambda_{\tau+1}$
- Can be used both in **gradient method** and **fast gradient methods**
- Stochastic interpretation through volume sampling

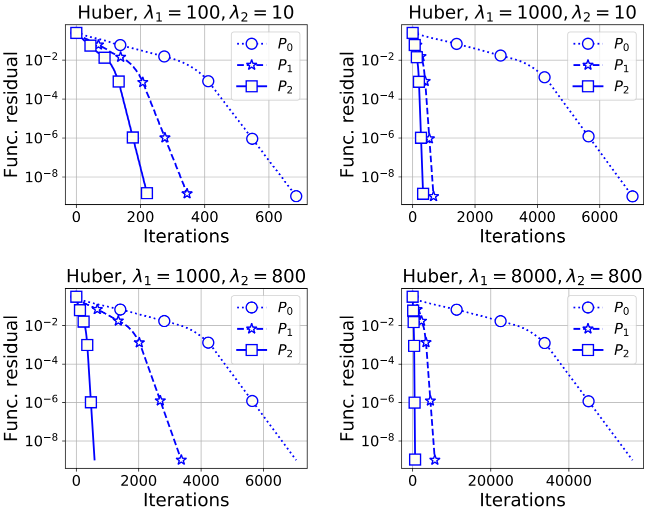
Krylov preconditioning

- Achieves the best possible polynomial preconditioning
- The preconditioner changes with iterations
- Works only with **gradient method** (unconstrained minimization)

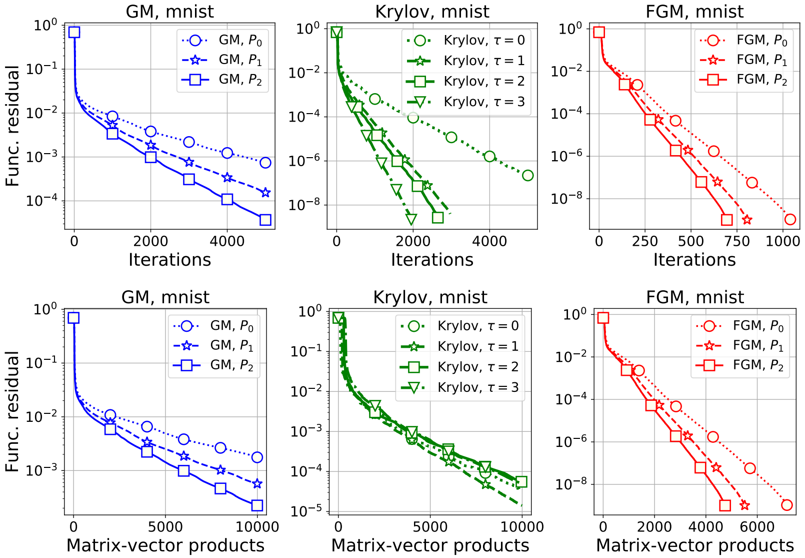
Outline

Experiments: regression with Huber loss

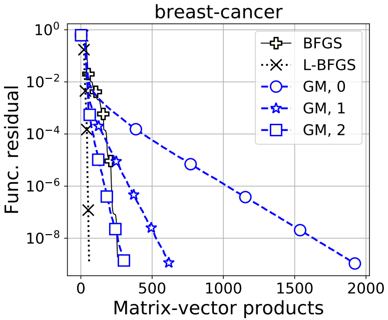
- synthetic data: control of the leading eigenvalues λ_1, λ_2



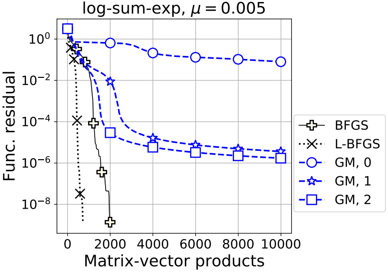
- real data (MNIST)

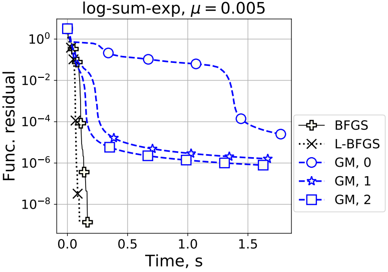


Experiments: quasi-Newton methods



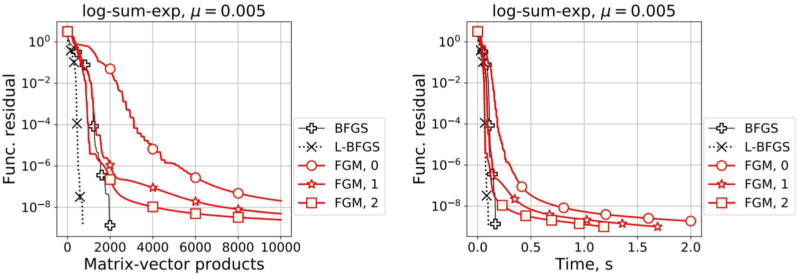
$$\min_{\boldsymbol{x}} \left\{ f_\mu(\boldsymbol{x}) \;=\; \mu \ln\left(\sum_{i=1}^{m} \exp\left(\frac{\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle - b_i}{\mu} \right) \right) \;\approx\; \max_{1 \le i \le m} \left[\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle - b_i \right] \right\}$$





- **Gradient method** vs. BFGS

$$\min_{\boldsymbol{x}} \left\{ f_\mu(\boldsymbol{x}) \;=\; \mu \ln\left(\sum_{i=1}^{m} \exp\left(\frac{\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle - b_i}{\mu} \right) \right) \;\approx\; \max_{1 \le i \le m} \left[\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle - b_i \right] \right\}$$



- **Fast gradient method** vs. BFGS

Conclusions

In practice, the spectrum of the Hessian is non-uniform

- We want the methods to exploit this information
- This work: fixed curvature matrix $\boldsymbol{B}$ $\Rightarrow$ polynomial approximation of $\boldsymbol{B}^{-1}$
- **Symmetric polynomial preconditioning** of degree τ, two operations:

$$\boldsymbol{B}^{\tau}\boldsymbol{h} \qquad \text{and} \qquad \mathrm{tr}\,(\boldsymbol{B}^{\tau}) \;=\; n \cdot \mathbb{E}_{\boldsymbol{u}\sim S^{n-1}}\big[\langle \boldsymbol{B}^{\tau}\boldsymbol{u}, \boldsymbol{u}\rangle\big]$$

- Instead of $\boldsymbol{B}$, we can use $\nabla^2 f(\boldsymbol{x})$
- Non-convex optimization $\Rightarrow$ **spectral preconditioning**

References:

1. Doikov, N., Rodomanov A., ICML 2023 (*International Conference on Machine Learning*) Polynomial Preconditioning for Gradient Methods
2. 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

- **Stochastic optimization** (the product of two random variables $\boldsymbol{P}_\xi \nabla f_\xi(\boldsymbol{x})$)
- Relations to classic **quasi-Newton methods**
- Local superlinear convergence
- **Complexity theory** for non-uniform spectrum (lower bounds and optimal methods)

Thank you very much for your attention!