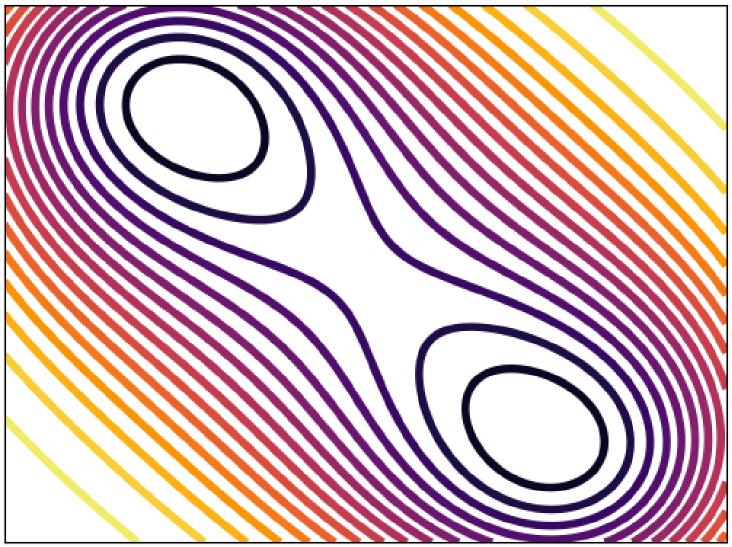
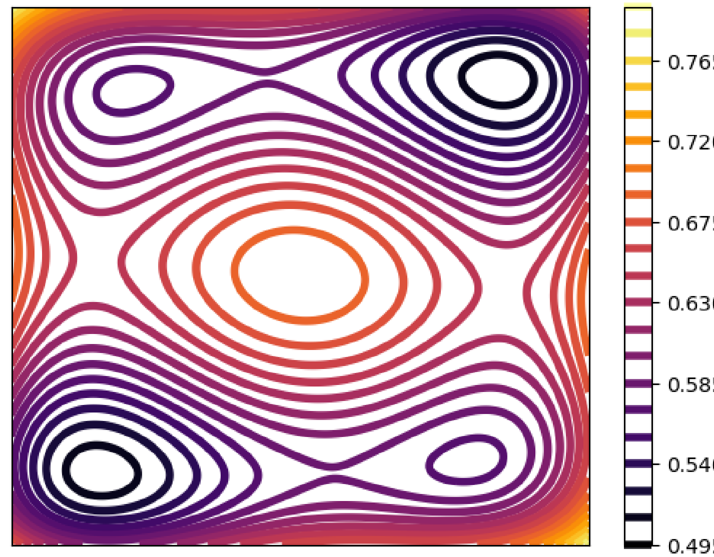
Quartic optimization problems





Radu-Alexandru Dragomir,
Post-doctoral researcher, EPFL

Joint work with Yurii Nesterov

CREST Statistics Seminar, March 27, 2023

Motivation: quadratic inverse problems

$$\min_{x \in E} \sum_{i=1}^{m} \left(\langle x, H_i x \rangle - b_i \right)^2$$

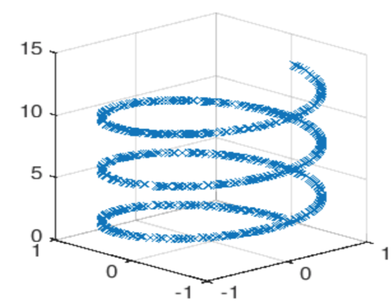
where $H_1 \dots H_m$ are linear operators.

- Phase retrieval: for $a_1 \dots a_m \in \mathbb{C}^n$,

$$\min_{x \in \mathbb{C}^n} \sum_{i=1}^{m} \left(|\langle a_i, x \rangle|_2^2 - b_i \right)^2 .$$

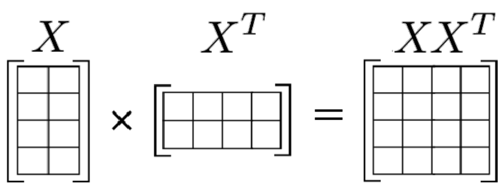
- Distance matrix completion:

$$\min_{X \in \mathbb{R}^{n \times r}} \sum_{(i,j) \in \Omega} \left(\|X_i - X_j\|^2 - d_{ij} \right)^2$$



- Low-rank matrix sensing in **factorized** form

$$\min_{X \in \mathbb{R}^{n \times r}} \sum_{i=1}^{m} \left(\langle A_i, X X^T \rangle - b_i \right)^2$$

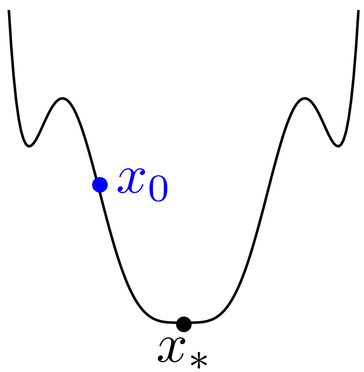


(Burer-Monteiro factorization)

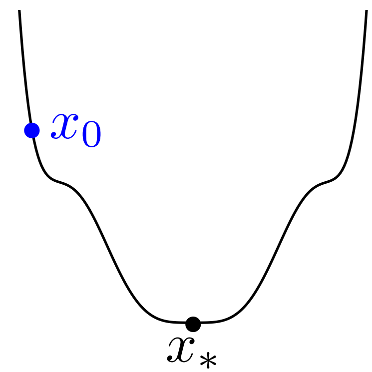
Quartic problems: challenges

1. Non-convexity When can we find the global minimum?

Good initialization + local convergence



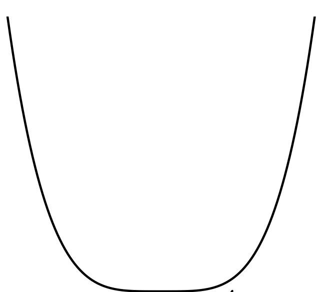
Benign landscape: no *spurious* minima



Guarantees under restrictive assumptions about problem and data.

[Chi et al. Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview. 2019]

2. Ill-conditioning



$x \mapsto x^4$

Curvature is unbounded:
no (local) **strong convexity** and no **L-smoothness**.

~~$\sigma I \preceq \nabla^2 f \preceq LI$~~

How to design efficient gradient methods for quartics?

Outline

- **Fast gradient methods for convex quartics**
- Optimal preconditoning

A simple convex quartic problem

$$\boxed{\min_{x \in E} Q[x]^4 - \langle c, x \rangle}$$

$Q : E^4 \to \mathbb{R}$ is a 4-linear symmetric map.

Assumption: the function $x \mapsto Q[x]^4$ is **convex**.

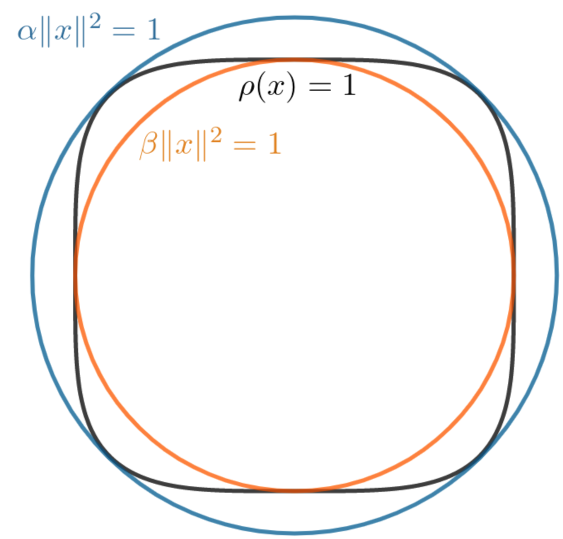
Quartic conditioning:

$$\alpha^2 \|x\|_2^4 \leq Q[x]^4 \leq \beta^2 \|x\|_2^4, \quad \kappa \triangleq \frac{\beta}{\alpha}.$$

We call κ the **quartic condition number**.

Example: for $Q[x]^4 = \sum_{i=1}^n x_i^4$,

$$\frac{1}{n}\|x\|_2^4 \leq Q[x]^4 \leq \|x\|_2^4.$$



Motivation : DC algorithm for quadratic inverse problems

$$\sum_{i=1}^{m} (\langle x, H_i x\rangle - b_i)^2 = \underbrace{\sum_{i=1}^{m} \langle x, H_i x\rangle^2}_{\rho(x)} - \underbrace{\sum_{i=1}^{m} \left(2b_i \langle x, H_i x\rangle - b_i^2\right)}_{\phi(x)}$$

In most examples, ρ and ϕ are convex! (not true in general)

Difference-of-convex optimization

$$\min_{x \in E} F(x) = \rho(x) - \phi(x), \quad \text{with } \rho, \phi \text{ convex functions.}$$

By convexity of ϕ,

$$F(x) \leq \rho(x) - \phi(\overline{x}) - \langle \nabla \phi(\overline{x}), x - \overline{x}\rangle$$

DC algorithm:

$$\boxed{x_{t+1} = \operatorname*{argmin}_{x \in E} Q[x]^4 - \langle \nabla \phi(x_t), x\rangle}$$

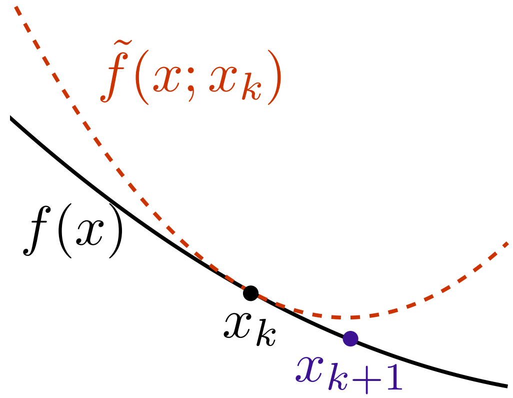
Here, ρ is a **quartic form:** $\rho(x) = Q[x]^4$ for some map Q.

Requires solving convex quartic subproblems!

Gradient methods

Iterative methods:

$$x_{k+1} = \operatorname*{argmin}_x \tilde{f}(x; x_k)$$



Quadratic approximation

$$\tilde{f}(x; x_k) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\lambda}\|x - x_k\|^2$$

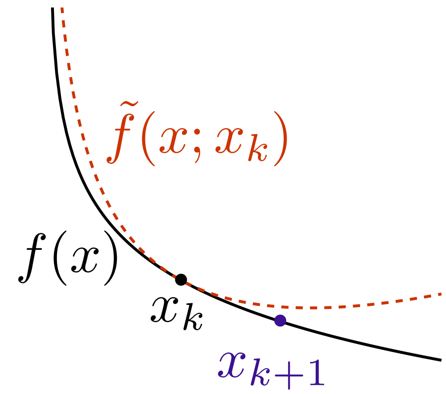
$\rightarrow$ **gradient descent**

$$x_{k+1} = x_k - \lambda \nabla f(x_k)$$

Bregman gradient methods

Iterative methods:

$$x_{k+1} = \underset{x}{\operatorname{argmin}}\, \tilde{f}(x; x_k)$$



Bregman approximation:

$$\tilde{f}(x; x_k) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{\lambda} D_h(x, x_k)$$

→ **Bregman gradient descent** **(a.k.a *mirror descent*)**

where

$$D_h(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle$$

is the **Bregman divergence** induced by convex function h.

Application to quartic problems [Bolte et al., 2018, Dragomir et al. 2021]

The reference function

$$h(x) = \frac{1}{4}\|x\|_2^4 + \frac{1}{2}\|x\|_2^2$$

is **well adapted** for quartic problems.

Which gradient method for convex quartics?

$$\boxed{\min_{x \in E} Q[x]^4 - \langle c, x \rangle.}$$

$$\alpha^2 \|x\|_2^4 \leq Q[x]^4 \leq \beta^2 \|x\|_2^4, \quad \kappa = \frac{\beta}{\alpha}$$

ρ is not L-**smooth** nor **strongly convex**.

- **Gradient descent:**

$$f(x_k) - f_* \leq \mathcal{O}\left(\frac{\beta^2 D^4}{k}\right) \quad \text{(standard)},$$

$$f(x_k) - f_* \leq \mathcal{O}\left(\frac{\beta^2 D^4}{k^2}\right) \quad \text{(accelerated)}.$$

  Requires **line search** (geometry **not adapted**).

- **Bregman gradient/mirror descent with quartic geometry:**

$$f(x_k) - f_* \leq \mathcal{O}\left(\frac{\beta^2 D^4}{k}\right)$$

  No line search needed, but no **acceleration** possible [Dragomir et al., 2021] .
  We can do better using the **polynomial structure**.

Homogenized gradient descent

$$\min_{x \in E} f(x) = Q[x]^4 - \langle c, x \rangle,$$

can be equivalently solved as the **homogenized** problem

$$\min_{y \in E} \sqrt{Q[y]^4} \quad \text{subject to } \langle c, y \rangle = 1. \qquad (P_{\text{hom}})$$

- $\sqrt{Q[\cdot]^4}$ **is convex and L-smooth:** apply projected gradient method to (P_{hom}),
- $\rho = Q[\cdot]^4$ **is uniformly convex of degree 4**:

$$\rho(x) - \rho(y) - \langle \nabla \rho(y), x - y \rangle \geq \frac{\alpha^2}{3} \|x - y\|^4.$$

These properties allow to prove

$$f(x_k) - f_* \leq \mathcal{O}\left(f_* \frac{\kappa^2}{k^2}\right) \qquad \text{(standard projected gradient)}$$

$$f(x_k) - f_* \leq \mathcal{O}\left(f_* \frac{\kappa^2}{k^4}\right) \qquad \text{(accelerated version)}$$

Outline

- Fast gradient methods for convex quartics
- **Optimal preconditoning**

Setup

Change of norm:

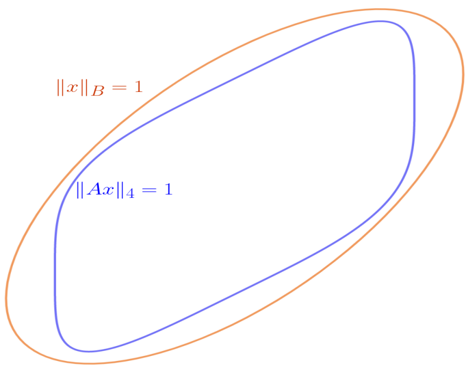
$$\alpha^2 \|x\|_B^4 \le Q[x]^4 \le \beta^2 \|x\|_B^4, \quad \kappa_B = \frac{\beta}{\alpha}$$

with $\|x\|_B^2 = \langle Bx, x\rangle$, for $B \succ 0$.

How to choose a good preconditioner B?

Assume Q is of the form

$$\boxed{Q[x]^4 = \sum_{i=1}^m \langle a_i, x\rangle^4 = \|Ax\|_4^4,} \quad A = \begin{bmatrix} a_1^T \\ \vdots \\ a_m^T \end{bmatrix} \in \mathbb{R}^{m\times d},$$



with $m \ge d$.

Goal: find $B \in \mathbb{S}_{++}^d$ such that $\|x\|_B \approx \|Ax\|_4$.

- **"Uniform choice":** $B^{(0)} = A^T A = \sum_{i=1}^m a_i a_i^T$, $\kappa_{B^{(0)}} \in [\sqrt{d}, \sqrt{m}]$
- **"Optimal choice":** $B^* = A^T D_{\tau^*} A = \sum_{i=1}^m \tau_i^* a_i a_i^T$, $\kappa_{B^*} = \sqrt{d}$
- John's theorem: generic convex bodies $\kappa = d$

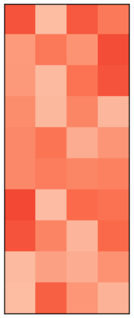
Condition number for uniform choice $B^{(0)}$

Row leverage scores of $A = [a_1, \dots a_m]^T \in \mathbb{R}^{m\times d}$

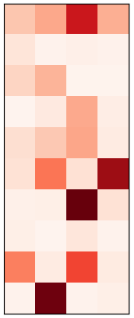
$$\ell_i(A) = \langle (A^T A)^{-1} a_i, a_i \rangle, \quad i = 1 \dots m.$$

Coherence: $$\gamma(A) = \max_{i=1\dots m} \ell_i(A)$$ $$\frac{d}{m} \le \gamma(A) \le 1$$

Low coherence
$\gamma(A) \approx \frac{d}{m}$



High coherence
$\gamma(A) \approx 1$



Condition number for uniform choice

With $B^{(0)} = A^T A$,

$$\frac{1}{m}\|x\|_{B^{(0)}}^4 \le \|Ax\|_4^4 \le \gamma(A)\|x\|_{B^{(0)}}^4$$

$$\kappa_{B^{(0)}} = \sqrt{m\gamma(A)} \in [\sqrt{d}, \sqrt{m}]$$

Condition number for optimal choice B^*

We search for a better B of the form

$$B(\tau) = \sum_{i=1}^{m} \tau_i a_i a_i^T, \quad \tau \in \mathbb{R}^m.$$

Lewis weights of order 2: there is a unique τ^* satisfying

$$\boxed{\langle B(\tau^*)^{-1} a_i, a_i \rangle = \tau_i^*, \quad i = 1 \ldots m.}$$

[Lewis 1978]

Can be computed with fixed-point iteration in $\mathcal{O}(md^2)$ time.

Theorem: the operator $B^* = B(\tau^*)$ satisfies $\boxed{\frac{1}{d}\|x\|_{B^*}^4 \leq \|Ax\|_4^4 \leq \|x\|_{B^*}^4}$

$$\kappa_{B^*} = \sqrt{d}$$

Recall that

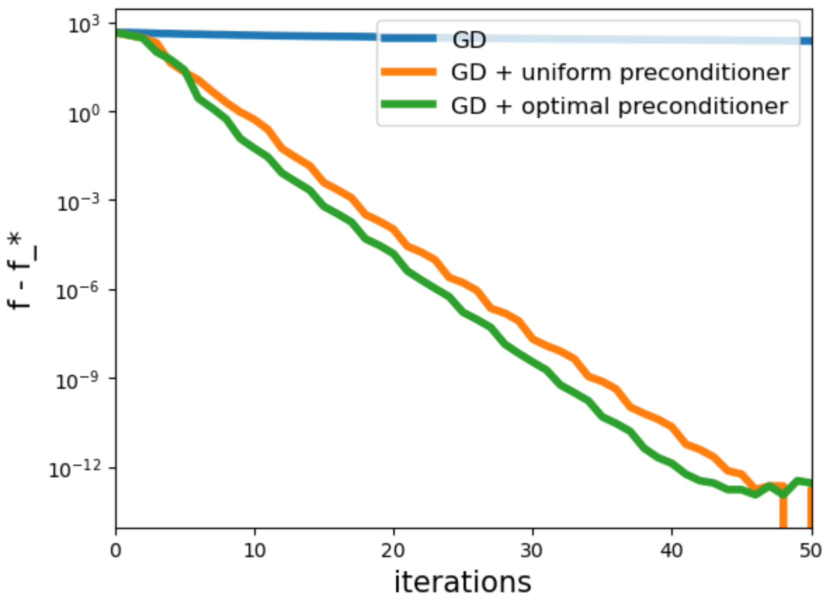
$$\kappa_{B(0)} = \sqrt{m\gamma(A)} \in [\sqrt{d}, \sqrt{m}]$$

Great improvement of B^* over $B^{(0)}$ for **tall** ($m \gg d$) and **coherent** ($\gamma(A) \approx 1$) matrices.

Numerical experiments

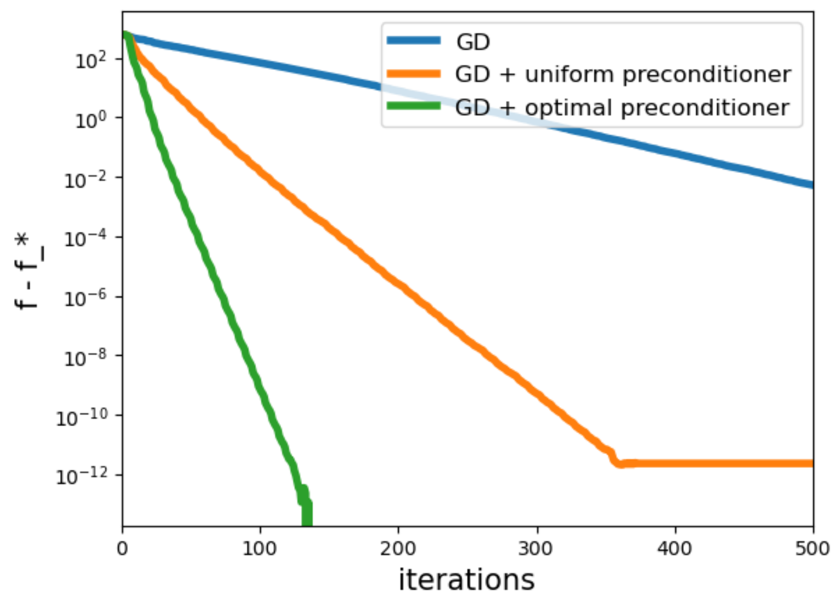
Toy problem with $d = 50, m = 1000$.

$$\min_{x \in \mathbb{R}^d} \sum_{i=1}^{m} \langle a_i, x \rangle^4 - \langle c, x \rangle$$



Low coherence $\gamma(A) \approx 5 \cdot 10^{-2}$

No improvement from B^* over $B^{(0)}$



High coherence $\gamma(A) \approx 0.8$

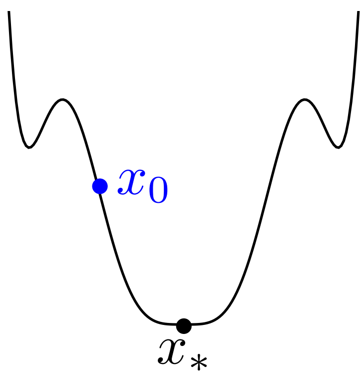
B^* improves performance

Perspectives: non-convex quartics

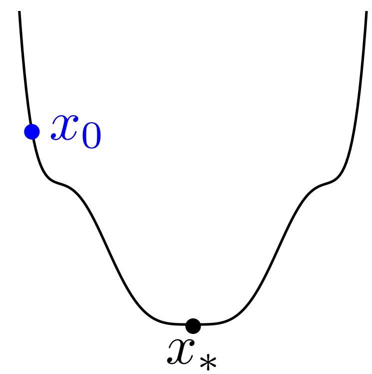
$$\min_{x \in E} \sum_{i=1}^{m} \left(\langle x, H_i x \rangle - b_i \right)^2$$

Problem is **non-convex:** can we find the global minimum x_*?

Good initialization + local convergence



Benign landscape: no *spurious* minima



Typical assumption: "restricted isometry"-like property

$$(1-\delta)\|x\|^4 \leq \sum_{i=1}^{m} \langle H_i x, x \rangle^2 \leq (1+\delta)\|x\|^4$$

Requires assumption on **distribution** of $H_1 \ldots H_m$

Ex: for phase retrieval, $H_i = a_i a_i^T$, and $a_1 \ldots a_m$ are **Gaussian i.i.d**

What about non-uniform $\{H_i\}$? Analysis of quartic conditioning?

Summary and perspectives

$$\min_{x\in E} Q[x]^4 - \langle c, x\rangle$$

$$\alpha^2\|x\|^4 \leq Q[x]^4 \leq \beta^2\|x\|^4, \quad \kappa = \frac{\beta}{\alpha}.$$

Contributions

- **Fast gradient methods** using **polynomial** structure: $\mathcal{O}(\kappa^2/k^4)$ rate.
- **Optimal preconditioner** for $Q[x]^4 = \|Ax\|_4^4$:

find $B \in \mathbb{S}^d_{++}$ such that $\|x\|_B \approx \|Ax\|_4$.

Perspectives

- Extension to **general** convex quartics polynomials ?

$$\min_{x\in E} Q[x]^4 + P[x]^3 + \langle Ax, x\rangle + \langle c, x\rangle.$$

- Practical preconditioning scheme based on **randomization**
- Lanscape analysis of quadratic inverse problems **beyond random design**

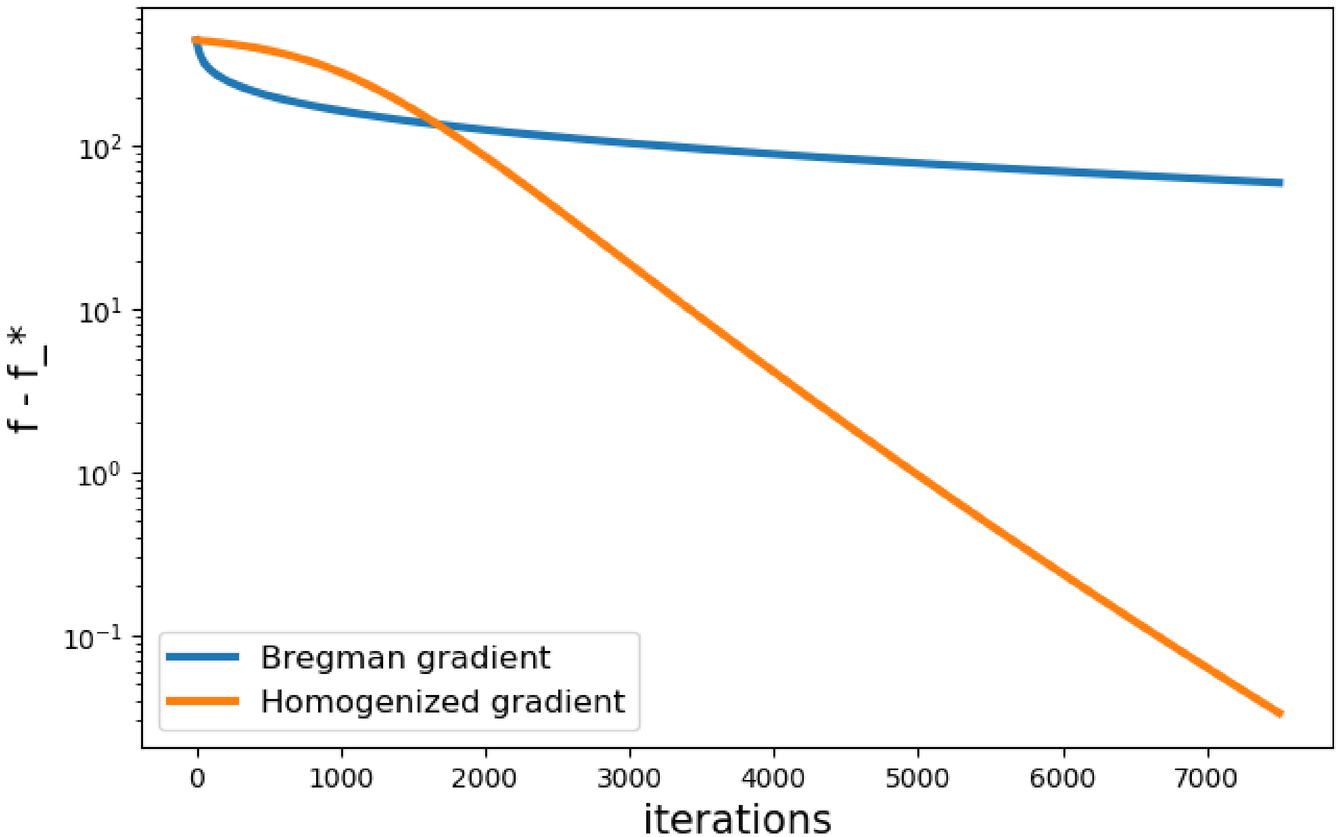
Thank you ! (paper on arXiv soon)

Appendix

Experiments

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{m} \langle a_i, x\rangle^4 - \langle c, x\rangle.$$

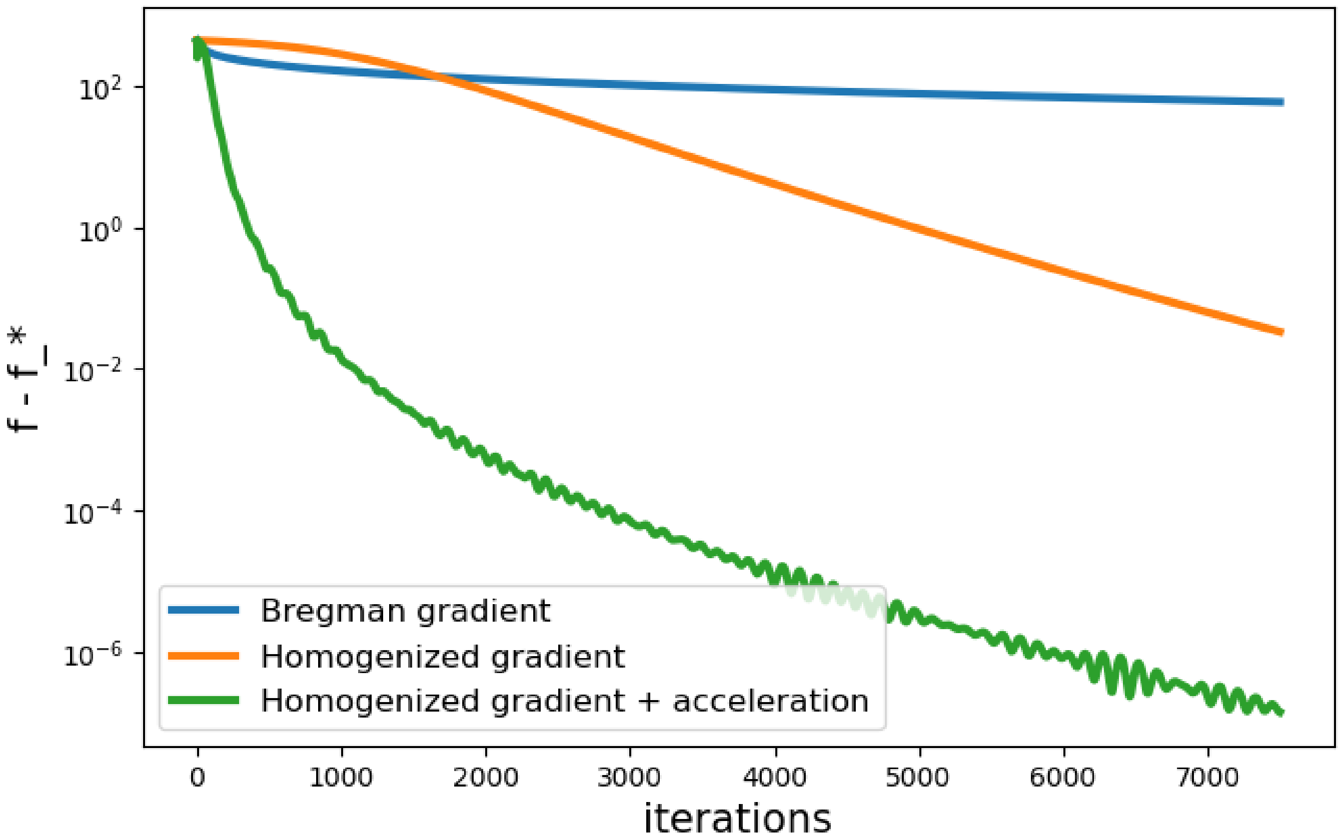
$m = 2000, n = 1000$, $a_1, \ldots, a_m$ Gaussian i.i.d.



Experiments

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{m} \langle a_i, x\rangle^4 - \langle c, x\rangle.$$

$m = 2000, n = 1000,$ $a_1, \ldots, a_m$ Gaussian i.i.d.



Experiments

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} \langle a_i, x \rangle^4 - \langle c, x \rangle.$$

$m = 2000, n = 1000$, $a_1, \ldots, a_m$ Gaussian i.i.d.