Proximal quasi-Newton method

Andersen Ang

Department of Combinatorics and Optimization, University of Waterloo, Canada

ms$\boldsymbol{x}$ang@uwaterloo.ca, where $\boldsymbol{x} = \lfloor \pi \rfloor$ Homepage: angms.science

First draft: September 7, 2022 Last update: October 17, 2022

Table of Contents

Minimizing composite function

- A standard class of convex optimization problem

$$(\mathcal{P}) \ : \ \min_{\boldsymbol{x}} \Big\{ F(\boldsymbol{x}) \coloneqq f(\boldsymbol{x}) + g(\boldsymbol{x}) \Big\}$$

  - $f : \mathbb{R}^n \to \mathbb{R}$ is convex and L-smooth
  - $g : \mathbb{R}^n \to \bar{\mathbb{R}} \cup \{+\infty\}$ is convex and possibly nonsmooth

- Example of $(\mathcal{P})$: Basis Pursuit Denoising

$$(\mathcal{P}_1) \ : \ \min_{\boldsymbol{x}} \frac{1}{2} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \lambda \|\boldsymbol{x}\|_1.$$

  - $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ with $m \ll n$ (standard setting)
  - $\boldsymbol{b} \in \mathbb{R}^m$ is observation
  - $\boldsymbol{x} \in \mathbb{R}^n$ is variable

Gradient descent and proximal gradient method

Gradient descent

- Problem class: $\min_{\boldsymbol{x}} f(\boldsymbol{x})$, f convex and differentiable
- Method: gradient descent: $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k)$.

Proximal gradient method

- Problem class: $\min_{\boldsymbol{x}} f(\boldsymbol{x}) + g(\boldsymbol{x})$, f, g convex, f differentiable
- Method: proximal gradient $\boldsymbol{x}_{k+1} = \operatorname{prox}_{\alpha_k g}\Big(\boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k)\Big)$
- Proximal operator: $\operatorname{prox}_{\gamma g}(\boldsymbol{x}) \coloneqq \operatorname*{argmin}_{\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{\xi} - \boldsymbol{x}\|_2^2 + \gamma g(\boldsymbol{\xi})$.

Proximal gradient method = gradient descent plus g

Both gradient descent and proximal gradient method are based on a local quadratic model of f:

Gradient descent

$$\begin{aligned}
\boldsymbol{x}_{k+1} &= \boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k) \\
&= \operatorname*{argmin}_{\boldsymbol{\xi}} f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{\xi} - \boldsymbol{x} \rangle + \frac{1}{2\alpha_k} \|\boldsymbol{\xi} - \boldsymbol{x}_k\|_2^2 \\
&= \operatorname*{argmin}_{\boldsymbol{\xi}} f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{\xi} - \boldsymbol{x} \rangle + \frac{1}{2} \langle \frac{1}{\alpha_k} \boldsymbol{I}(\boldsymbol{\xi} - \boldsymbol{x}_k), \boldsymbol{\xi} - \boldsymbol{x}_k \rangle \\
&= \operatorname*{argmin}_{\boldsymbol{\xi}} \frac{1}{2\alpha_k} \left\| \boldsymbol{\xi} - \Big(\boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k) \Big) \right\|_2^2
\end{aligned}$$

Proximal gradient method

$$\begin{aligned}
\boldsymbol{x}_{k+1} &= \operatorname{prox}_{\alpha_k g}\Big(\boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k) \Big) \\
&= \operatorname*{argmin}_{\boldsymbol{\xi}} \frac{1}{2} \left\| \boldsymbol{\xi} - \Big(\boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k) \Big) \right\|_2^2 + \alpha_k g(\boldsymbol{\xi}) \\
&= \operatorname*{argmin}_{\boldsymbol{\xi}} \frac{1}{2\alpha_k} \left\| \boldsymbol{\xi} - \Big(\boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k) \Big) \right\|_2^2 + g(\boldsymbol{\xi}) \\
&= \operatorname*{argmin}_{\boldsymbol{\xi}} f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{\xi} - \boldsymbol{x} \rangle + \frac{1}{2} \langle \frac{1}{\alpha_k} \boldsymbol{I}(\boldsymbol{\xi} - \boldsymbol{x}_k), \boldsymbol{\xi} - \boldsymbol{x}_k \rangle + g(\boldsymbol{\xi})
\end{aligned}$$

Generalizing proximal gradient method

- Proximal gradient method

$$\boldsymbol{x}_{k+1} = \operatorname*{argmin}_{\boldsymbol{\xi}} f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{\xi} - \boldsymbol{x} \rangle + \frac{1}{2} \langle \frac{1}{\alpha_k} \boldsymbol{I} (\boldsymbol{\xi} - \boldsymbol{x}_k), \boldsymbol{\xi} - \boldsymbol{x}_k \rangle + g(\boldsymbol{\xi})$$

- Generalize it by replacing $\frac{1}{\alpha_k} \boldsymbol{I}$ with a positive definite matrix $\boldsymbol{H}_k \succ \boldsymbol{0}$

$$\begin{aligned}
\boldsymbol{x}_{k+1} &= \operatorname*{argmin}_{\boldsymbol{\xi}} f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{\xi} - \boldsymbol{x} \rangle + \frac{1}{2} \langle \boldsymbol{H}_k (\boldsymbol{\xi} - \boldsymbol{x}_k), \boldsymbol{\xi} - \boldsymbol{x}_k \rangle + g(\boldsymbol{\xi}) \\
&= \operatorname*{argmin}_{\boldsymbol{\xi}} f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{\xi} - \boldsymbol{x} \rangle + \frac{1}{2} \|\boldsymbol{\xi} - \boldsymbol{x}_k\|_{\boldsymbol{H}_k}^2 + g(\boldsymbol{\xi}) \\
&= \operatorname*{argmin}_{\boldsymbol{\xi}} \frac{1}{2} \left\| \boldsymbol{\xi} - \left(\boldsymbol{x}_k - \boldsymbol{H}_k^{-1} \nabla f(\boldsymbol{x}_k) \right) \right\|_{\boldsymbol{H}_k}^2 + g(\boldsymbol{\xi})
\end{aligned}$$

where $\|\boldsymbol{x}\|_{\boldsymbol{H}} \coloneqq \sqrt{\langle \boldsymbol{H}\boldsymbol{x}, \boldsymbol{x} \rangle}$.

Proximal Newton's update

$$\begin{aligned}
\boldsymbol{x}_{k+1} &= \operatorname*{argmin}_{\boldsymbol{\xi}} f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{\xi} - \boldsymbol{x} \rangle + \frac{1}{2} \|\boldsymbol{\xi} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k} + g(\boldsymbol{\xi}) \\
&= \operatorname*{argmin}_{\boldsymbol{\xi}} \frac{1}{2} \left\| \boldsymbol{\xi} - \left(\boldsymbol{x}_k - \boldsymbol{H}_k^{-1} \nabla f(\boldsymbol{x}_k) \right) \right\|^2_{\boldsymbol{H}_k} + g(\boldsymbol{\xi}) \\
&=: \operatorname{prox}_g^{\boldsymbol{H}_k} \left(\boldsymbol{x}_k - \boldsymbol{H}_k^{-1} \nabla f(\boldsymbol{x}_k) \right)
\end{aligned}$$

- If $\boldsymbol{H}_k = \frac{1}{\alpha_k} \boldsymbol{I}$, Proximal Newton's method reduces to the classical proximal gradient method.
- If $\boldsymbol{H}_k =$ the Hessian of f at $\boldsymbol{x}_k$, such update is called Proximal Newton's method.
  - Advantage: $\boldsymbol{H}_k$ brings the curvature information at $\boldsymbol{x}_k$, improves the local quadratic model.
- If $\boldsymbol{H}_k$ is not the Hessian of f at $\boldsymbol{x}_k$ but an approximate, such update is called Proximal quasi-Newton's method.
- $\operatorname{prox}_g^{\boldsymbol{H}_k}$ itself is an optimization problem: can be hard/expensive to solve in general.
  - A trade-off: you sacrifice for having high computational complexity on solving $\operatorname{prox}_g^{\boldsymbol{H}_k}$ to gain speed up on the convergence of $\left\{\boldsymbol{x}_k\right\}_{k \in \mathbb{N}}$.

Proximal Newton's update with a line search

$$\boldsymbol{x}_{k+1} \quad = \quad \operatorname{prox}_g^{\boldsymbol{H}_k}\Big(\boldsymbol{x}_k - \boldsymbol{H}_k^{-1}\nabla f(\boldsymbol{x}_k)\Big) \qquad \text{(Proximal Newton's update)}$$

- Adding a linear combination with the previous point $\boldsymbol{x}_k$

$$\begin{aligned} \boldsymbol{x}_{k+1} &= (1-\alpha)\boldsymbol{x}_k + \alpha \underbrace{\operatorname{prox}_g^{\boldsymbol{H}_k}\Big(\boldsymbol{x}_k - \boldsymbol{H}_k^{-1}\nabla f(\boldsymbol{x}_k)\Big)}_{=:\,\boldsymbol{x}^*_{\boldsymbol{H}_k}} \\ &= (1-\alpha)\boldsymbol{x}_k + \alpha \boldsymbol{x}^*_{\boldsymbol{H}_k} \end{aligned}$$

  where $\alpha > 0$ is a stepsize that $\alpha \in (0, 1]$, to be shown later.

- Why adding a line search: sometimes it improves the convergence.

- No line search: it means $\alpha_k = 1$ for all k.

Proximal Newton's (with a line search) algorithm

$$(\mathcal{P}) \ : \ \min_{\boldsymbol{x}} \Big\{ F(\boldsymbol{x}) \coloneqq f(\boldsymbol{x}) + g(\boldsymbol{x}) \Big\}$$

- Starting with an initial point $\boldsymbol{x}_0$, let

$$m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) \coloneqq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{\xi} - \boldsymbol{x} \rangle + \frac{1}{2} \|\boldsymbol{\xi} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k},$$

Algorithm 1: Proximal Newton's algorithm

Result: An approximate solution to $(\mathcal{P})$

1 **for** $k = 1, 2, \ldots$ **do**

2 $\quad \boldsymbol{x}^*_{\boldsymbol{H}_k} = \operatorname{argmin}_{\boldsymbol{\xi}} \, m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) + g(\boldsymbol{\xi})$ solving the proximal Newton subproblem;

3 $\quad \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k(\boldsymbol{x}^*_{\boldsymbol{H}_k} - \boldsymbol{x}_k)$ convex combination (line search on α_k)

4 **end**

- If $\alpha_k = 1$ then the algorithm reduces to

$$\boldsymbol{x}_{k+1} = \operatorname*{argmin}_{\boldsymbol{\xi}} \, m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) + g(\boldsymbol{\xi}) = \operatorname{prox}_g^{\boldsymbol{H}_k}\Big(\boldsymbol{x}_k - \boldsymbol{H}_k^{-1} \nabla f(\boldsymbol{x}_k)\Big).$$

Last slide before mindfuck: summary of the theory

- The sequence $F(\boldsymbol{x}_k)$ converges to F^* with with rate $\mathcal{O}(\frac{1}{k})$.
- The next 20+ pages are just to explain this.
- Warning: attempt to understand them will drain your stamina and cause brain damage.
- I warned you, if you are ok then let's go.

Table of Contents

Fixed-point property

$$(\mathcal{P}) \ : \ \min_{\boldsymbol{x}} \Big\{ F(\boldsymbol{x}) \coloneqq f(\boldsymbol{x}) + g(\boldsymbol{x}) \Big\}$$

$$\text{algo} \ : \begin{cases} \boldsymbol{x}^*_{\boldsymbol{H}_k} = \operatorname*{argmin}_{\boldsymbol{\xi}} \ m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) + g(\boldsymbol{\xi}), \quad m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) \coloneqq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{\xi} - \boldsymbol{x}_k \rangle + \frac{1}{2} \|\boldsymbol{\xi} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k} \\ \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k (\boldsymbol{x}^*_{\boldsymbol{H}_k} - \boldsymbol{x}_k) \end{cases}$$

Lemma 0. $\boldsymbol{H}_k(\boldsymbol{x}^*_{\boldsymbol{H}_k} - \boldsymbol{x}_k) = \boldsymbol{0}$ if and only if $\boldsymbol{x}_k$ solves $(\mathcal{P})$.

Proof. $\boldsymbol{H}_k(\boldsymbol{x}^*_{\boldsymbol{H}_k} - \boldsymbol{x}_k) = \boldsymbol{0} \overset{\boldsymbol{H}_k \succ \boldsymbol{0}}{\iff} \boldsymbol{x}^*_{\boldsymbol{H}_k} - \boldsymbol{x}_k = \boldsymbol{0} \iff \boldsymbol{x}^*_{\boldsymbol{H}_k} = \boldsymbol{x}_k$.

Subgradient 1st-order optimality on $\boldsymbol{x}^*_{\boldsymbol{H}_k}$: $\boldsymbol{x}^*_{\boldsymbol{H}_k} = \operatorname*{argmin}_{\boldsymbol{\xi}} \ m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) + g(\boldsymbol{\xi}) \iff \nabla m_{\boldsymbol{H}_k}(\boldsymbol{x}^*_{\boldsymbol{H}_k}, \boldsymbol{x}_k) + \partial g(\boldsymbol{x}^*_{\boldsymbol{H}_k}) \ni \boldsymbol{0}$.

Recall

- the quadratic term in $m_{\boldsymbol{H}_k}$ is $\| \cdot \|^2_{\boldsymbol{H}_k}$, so when we take ∇ there will be a $\boldsymbol{H}_k$ matrix,
- $\boldsymbol{H}_k \succ \boldsymbol{0}$ so $m_{\boldsymbol{H}_k}$ is strongly convex so the solution set to $\operatorname*{argmin}_{\boldsymbol{\xi}} \ m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) + g(\boldsymbol{\xi})$ is singleton, which is $\boldsymbol{x}^*_{\boldsymbol{H}_k}$

i.e, we have

$$\nabla f(\boldsymbol{x}_k) + \boldsymbol{H}_k(\boldsymbol{x}^*_{\boldsymbol{H}_k} - \boldsymbol{x}_k) + \partial g(\boldsymbol{x}^*_{\boldsymbol{H}_k}) \ni \boldsymbol{0},$$

which gives $\nabla f(\boldsymbol{x}_k) + \partial g(\boldsymbol{x}_k) \ni \boldsymbol{0}$ if $\boldsymbol{x}^*_{\boldsymbol{H}_k} = \boldsymbol{x}_k$, which is when $\boldsymbol{x}_k$ is a solution of $(\mathcal{P})$. □

Descent property

$$(\mathcal{P}) \;:\; \min_{\boldsymbol{x}} \Big\{ F(\boldsymbol{x}) \coloneqq f(\boldsymbol{x}) + g(\boldsymbol{x}) \Big\}$$

$$\text{algo} \;:\; \begin{cases} \boldsymbol{x}^*_{\boldsymbol{H}_k} = \operatorname*{argmin}_{\boldsymbol{\xi}} m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) + g(\boldsymbol{\xi}), \quad m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) \coloneqq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{\xi} - \boldsymbol{x}_k \rangle + \frac{1}{2}\|\boldsymbol{\xi} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k} \\ \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k(\boldsymbol{x}^*_{\boldsymbol{H}_k} - \boldsymbol{x}_k) \end{cases}$$

Lemma 1. While $\boldsymbol{x}_k \notin \operatorname{argmin} F$, for sufficiently small stepsize $\alpha_k > 0$, we have $F(\boldsymbol{x}_{k+1}) < F(\boldsymbol{x}_k)$.

- **Proof**. By definition, $\boldsymbol{x}^*_{\boldsymbol{H}_k}$ is the minimizer of $\phi \coloneqq m_{\boldsymbol{H}_k} + g$, i.e., therefore $\phi(\boldsymbol{x}^*_{\boldsymbol{H}_k}) \leq \phi(\boldsymbol{x}_{k+1})$. Furthermore, ϕ is strongly convex since $\boldsymbol{H}_k \succ \boldsymbol{0}$, thus we have strict inequality

$$\phi(\boldsymbol{x}^*_{\boldsymbol{H}_k}) \;<\; \phi(\boldsymbol{x}^*_{\boldsymbol{H}_k})$$

$$f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}^*_{\boldsymbol{H}_k} - \boldsymbol{x}_k \rangle + \frac{1}{2}\|\boldsymbol{x}^*_{\boldsymbol{H}_k} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k} + g(\boldsymbol{x}^*_{\boldsymbol{H}_k}) \;<\; f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle + \frac{1}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k} + g(\boldsymbol{x}_{k+1}).$$

- Remove $f(\boldsymbol{x}_k)$, let $\boldsymbol{d}_k \coloneqq \boldsymbol{x}^*_{\boldsymbol{H}_k} - \boldsymbol{x}_k$ and by definition $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$ thus $\boldsymbol{x}_{k+1} - \boldsymbol{x}_k = \alpha_k \boldsymbol{d}_k$

$$\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k \rangle + \frac{1}{2}\|\boldsymbol{d}_k\|^2_{\boldsymbol{H}_k} + g(\boldsymbol{x}^*_{\boldsymbol{H}_k}) \;<\; \alpha_k \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k \rangle + \frac{\alpha_k^2}{2}\|\boldsymbol{d}_k\|^2_{\boldsymbol{H}_k} + \underline{\underline{g(\boldsymbol{x}_{k+1})}}. \tag{Δ}$$

- By assumption g is convex so it satisfies the Jensen inequality of convexity

$$g(\boldsymbol{x}_{k+1}) = g\Big(\boldsymbol{x}_k + \alpha_k(\boldsymbol{x}^*_{\boldsymbol{H}_k} - \boldsymbol{x}_k)\Big) = g\Big((1-\alpha)\boldsymbol{x}_k + \alpha_k \boldsymbol{x}^*_{\boldsymbol{H}_k}\Big) \leq (1-\alpha_k) g(\boldsymbol{x}_k) + \alpha_k g(\boldsymbol{x}^*_{\boldsymbol{H}_k}). \tag{∇}$$

  Put (∇) into (Δ) gives

$$\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k \rangle + \frac{1}{2}\|\boldsymbol{d}_k\|^2_{\boldsymbol{H}_k} + g(\boldsymbol{x}^*_{\boldsymbol{H}_k}) \;<\; \alpha_k \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k \rangle + \frac{\alpha_k^2}{2}\|\boldsymbol{d}_k\|^2_{\boldsymbol{H}_k} + \underline{\underline{(1-\alpha_k) g(\boldsymbol{x}_k) + \alpha_k g(\boldsymbol{x}^*_{\boldsymbol{H}_k})}}. \tag{$\square$}$$

- Simply ($\square$)

$$\begin{aligned}
(1-\alpha_k)\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k\rangle + \frac{1-\alpha_k^2}{2}\|\boldsymbol{d}_k\|^2_{\boldsymbol{H}_k} + (1-\alpha_k)g(\boldsymbol{x}^*_{\boldsymbol{H}_k}) &< (1-\alpha_k)g(\boldsymbol{x}_k) && \text{cancel and group terms}\\
\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k\rangle + \frac{1+\alpha_k}{2}\|\boldsymbol{d}_k\|^2_{\boldsymbol{H}_k} + g(\boldsymbol{x}^*_{\boldsymbol{H}_k}) &< g(\boldsymbol{x}_k) && \text{divide by } 1-\alpha_k \qquad (*)\\
\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k\rangle + g(\boldsymbol{x}^*_{\boldsymbol{H}_k}) - g(\boldsymbol{x}_k) &< -\frac{1+\alpha_k}{2}\|\boldsymbol{d}_k\|^2_{\boldsymbol{H}_k} && \text{rearrange}
\end{aligned}$$

- Consider $F(\boldsymbol{x}_{k+1}) - F(\boldsymbol{x}_k)$ and apply the Jansen inequality on g again

$$\begin{aligned}
F(\boldsymbol{x}_{k+1}) - F(\boldsymbol{x}_k) = f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k) + g(\boldsymbol{x}_{k+1}) - g(\boldsymbol{x}_k) &\le f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k) + \underline{\underline{(1-\alpha_k)g(\boldsymbol{x}_k) + \alpha_k g(\boldsymbol{x}^*_{\boldsymbol{H}_k})}} - g(\boldsymbol{x}_k)\\
&= f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k) - \underline{\underline{\alpha_k g(\boldsymbol{x}_k) + \alpha_k g(\boldsymbol{x}^*_{\boldsymbol{H}_k})}} \qquad (\#)
\end{aligned}$$

- Recall the Taylor expansion for f at $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$ is

$$f(\boldsymbol{x}_{k+1}) \;=\; f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k) \;=\; f(\boldsymbol{x}_k) + \alpha_k \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k\rangle + \mathcal{O}\Big((\alpha_k)^2\Big). \qquad (\dagger)$$

- Put (†) into (#) gives

$$\begin{aligned}
F(\boldsymbol{x}_{k+1}) - F(\boldsymbol{x}_k) &\le \alpha_k \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k\rangle + \mathcal{O}\Big((\alpha_k)^2\Big) - \alpha_k g(\boldsymbol{x}_k) + \alpha_k g(\boldsymbol{x}^*_{\boldsymbol{H}_k})\\
&= \alpha_k \Big(\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k\rangle - g(\boldsymbol{x}_k) + g(\boldsymbol{x}^*_{\boldsymbol{H}_k})\Big) + \mathcal{O}\Big((\alpha_k)^2\Big)\\
&\overset{(*)}{\le} \alpha_k \Big(-\frac{1+\alpha_k}{2}\|\boldsymbol{d}_k\|^2_{\boldsymbol{H}_k}\Big) + \mathcal{O}\Big((\alpha_k)^2\Big)\\
&= -\frac{\alpha_k}{2}\|\boldsymbol{d}_k\|^2_{\boldsymbol{H}_k} + -\frac{\alpha_k^2}{2}\|\boldsymbol{d}_k\|^2_{\boldsymbol{H}_k} + \mathcal{O}\Big((\alpha_k)^2\Big)\\
&< 0
\end{aligned}$$

where the last inequality is true for sufficiently small α_k. $\square$

Stepsize range $\alpha_k \in (0, 1]$

$$(\mathcal{P}) \ : \ \min_{\boldsymbol{x}} \Big\{ F(\boldsymbol{x}) \coloneqq f(\boldsymbol{x}) + g(\boldsymbol{x}) \Big\}$$

$$\text{algo} \ : \begin{cases} \boldsymbol{x}^*_{\boldsymbol{H}_k} = \operatorname*{argmin}_{\boldsymbol{\xi}} \ m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) + g(\boldsymbol{\xi}), \quad m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) \coloneqq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{\xi} - \boldsymbol{x}_k \rangle + \frac{1}{2} \|\boldsymbol{\xi} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k} \\ \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k (\boldsymbol{x}^*_{\boldsymbol{H}_k} - \boldsymbol{x}_k) \end{cases}$$

Lemma 2. If $F(\boldsymbol{x}^*_{\boldsymbol{H}_k}) \leq m_{\boldsymbol{H}_k}(\boldsymbol{x}^*_{\boldsymbol{H}_k}, \boldsymbol{x}_k) + g(\boldsymbol{x}^*_{\boldsymbol{H}_k})$ holds for $\boldsymbol{x}^*_{\boldsymbol{H}_k}$, then $F(\boldsymbol{x}_{k+1}) < F(\boldsymbol{x}_k)$ for any $\alpha_k \in (0, 1]$.

- **Proof**. First we show that $F(\boldsymbol{x}^*_{\boldsymbol{H}_k}) < F(\boldsymbol{x}_k)$

$$\begin{aligned} F(\boldsymbol{x}^*_{\boldsymbol{H}_k}) &\leq m_{\boldsymbol{H}_k}(\boldsymbol{x}^*_{\boldsymbol{H}_k}, \boldsymbol{x}_k) + g(\boldsymbol{x}^*_{\boldsymbol{H}_k}) && \text{by assumption} \\ &< m_{\boldsymbol{H}_k}(\boldsymbol{x}_k, \boldsymbol{x}_k) + g(\boldsymbol{x}_k) && \boldsymbol{x}^*_{\boldsymbol{H}_k} = \operatorname*{argmin}_{\boldsymbol{\xi}} \ m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) + g(\boldsymbol{\xi}), \ m_{\boldsymbol{H}_k} \text{ strongly convex} \\ &= F(\boldsymbol{x}_k) && F(\boldsymbol{x}_k) = m_{\boldsymbol{H}_k}(\boldsymbol{x}_k, \boldsymbol{x}_k) + g(\boldsymbol{x}_k) \end{aligned}$$

- Apply Jansen inequality on F

$$F(\boldsymbol{x}_{k+1}) \ = \ F\Big((1-\alpha)\boldsymbol{x}_k + \alpha_k \boldsymbol{x}^*_{\boldsymbol{H}_k}\Big) \ \leq \ (1-\alpha) F(\boldsymbol{x}_k) + \alpha_k F(\boldsymbol{x}^*_{\boldsymbol{H}_k}) \overset{F(\boldsymbol{x}^*_{\boldsymbol{H}_k}) < F(\boldsymbol{x}_k)}{<} F(\boldsymbol{x}_k).$$

Jansen inequality holds when $\alpha_k \in [0, 1]$, but when $\alpha_k = 0$ we have $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k \implies F(\boldsymbol{x}_{k+1}) = F(\boldsymbol{x}_k)$, so we have $F(\boldsymbol{x}_{k+1}) < F(\boldsymbol{x}_k)$ for $\alpha_k \in (0, 1]$. □

Comments ... 1/2

- Problem and algorithm

$$(\mathcal{P}) \quad : \quad \min_{\boldsymbol{x}} \Big\{ F(\boldsymbol{x}) \coloneqq f(\boldsymbol{x}) + g(\boldsymbol{x}) \Big\}.$$

$$\text{algo} \quad : \quad \begin{cases} m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) \coloneqq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{\xi} - \boldsymbol{x}_k \rangle + \dfrac{1}{2} \|\boldsymbol{\xi} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k}. \\ \boldsymbol{x}^*_{\boldsymbol{H}_k} = \operatorname{argmin}_{\boldsymbol{\xi}} \; m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) + g(\boldsymbol{\xi}). \\ \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k (\boldsymbol{x}^*_{\boldsymbol{H}_k} - \boldsymbol{x}_k). \end{cases}$$

- What we have proved
  - Fixed-point property: if $\boldsymbol{x}_k \in \Big\{ \operatorname{argmin} F \Big\} \iff \boldsymbol{H}_k(\boldsymbol{x}^*_{\boldsymbol{H}_k} - \boldsymbol{x}_k) = \boldsymbol{0}$.
  - If $\boldsymbol{x}_k \notin \Big\{ \operatorname{argmin} F \Big\}$, then for sufficiently small $\alpha_k > 0$, we have $F(\boldsymbol{x}_{k+1}) < F(\boldsymbol{x}_k)$.
  - If $F(\boldsymbol{x}^*_{\boldsymbol{H}_k}) \leq m_{\boldsymbol{H}_k}(\boldsymbol{x}^*_{\boldsymbol{H}_k}, \boldsymbol{x}_k) + g(\boldsymbol{x}^*_{\boldsymbol{H}_k})$, then $F(\boldsymbol{x}_{k+1}) < F(\boldsymbol{x}_k)$ for any $\alpha_k \in (0, 1]$.
- The "weakness" of such result: only descent but not sufficient descent. Recall that for proximal gradient update we have sufficient descent condition.

Comments ... 2/2

- Problem and algorithm

$$(\mathcal{P}) \quad : \quad \min_{\boldsymbol{x}} \Big\{ F(\boldsymbol{x}) \coloneqq f(\boldsymbol{x}) + g(\boldsymbol{x}) \Big\}.$$

$$\text{algo} \quad : \quad \begin{cases} m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) \coloneqq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{\xi} - \boldsymbol{x}_k \rangle + \dfrac{1}{2} \|\boldsymbol{\xi} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k}. \\ \boldsymbol{x}^*_{\boldsymbol{H}_k} = \operatorname{argmin}_{\boldsymbol{\xi}} \; m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) + g(\boldsymbol{\xi}). \\ \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k (\boldsymbol{x}^*_{\boldsymbol{H}_k} - \boldsymbol{x}_k). \end{cases}$$

- Lemma 2. If $F(\boldsymbol{x}^*_{\boldsymbol{H}_k}) \leq m_{\boldsymbol{H}_k}(\boldsymbol{x}^*_{\boldsymbol{H}_k}, \boldsymbol{x}_k) + g(\boldsymbol{x}^*_{\boldsymbol{H}_k})$, then $F(\boldsymbol{x}_{k+1}) < F(\boldsymbol{x}_k)$ for any $\alpha_k \in (0, 1]$.

- Lemma 2 requires $F(\boldsymbol{x}^*_{\boldsymbol{H}_k}) \leq m_{\boldsymbol{H}_k}(\boldsymbol{x}^*_{\boldsymbol{H}_k}, \boldsymbol{x}_k) + g(\boldsymbol{x}^*_{\boldsymbol{H}_k})$, this condition restricts the scope of $\boldsymbol{H}_k$: we cannot use any arbitrary $\boldsymbol{H}_k$ but one that *majorize* f at $\boldsymbol{x}^*_{\boldsymbol{H}_k}$.

$$\begin{aligned} & F(\boldsymbol{x}^*_{\boldsymbol{H}_k}) && \leq && m_{\boldsymbol{H}_k}(\boldsymbol{x}^*_{\boldsymbol{H}_k}, \boldsymbol{x}_k) + g(\boldsymbol{x}^*_{\boldsymbol{H}_k}) \\ \iff \quad & f(\boldsymbol{x}^*_{\boldsymbol{H}_k}) + g(\boldsymbol{x}^*_{\boldsymbol{H}_k}) && \leq && m_{\boldsymbol{H}_k}(\boldsymbol{x}^*_{\boldsymbol{H}_k}, \boldsymbol{x}_k) + g(\boldsymbol{x}^*_{\boldsymbol{H}_k}) \\ \iff \quad & f(\boldsymbol{x}^*_{\boldsymbol{H}_k}) && \leq && f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}^*_{\boldsymbol{H}_k} - \boldsymbol{x}_k \rangle + \frac{1}{2} \|\boldsymbol{x}^*_{\boldsymbol{H}_k} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k} \end{aligned}$$

Table of Contents

Convergence with no line search ($\alpha_k = 1$), part 1

$$(\mathcal{P}) \ : \ \min_{\boldsymbol{x}} \Big\{ F(\boldsymbol{x}) \coloneqq f(\boldsymbol{x}) + g(\boldsymbol{x}) \Big\}$$

$$\text{algo} \quad : \begin{cases} \boldsymbol{x}^*_{\boldsymbol{H}_k} = \operatorname*{argmin}_{\boldsymbol{\xi}} \ m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) + g(\boldsymbol{\xi}), \quad m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) \coloneqq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{\xi} - \boldsymbol{x}_k \rangle + \frac{1}{2} \|\boldsymbol{\xi} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k} \\ \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k (\boldsymbol{x}^*_{\boldsymbol{H}_k} - \boldsymbol{x}_k) \end{cases}$$

Lemma 3. Suppose $F(\boldsymbol{x}_{k+1}) \leq m_{\boldsymbol{H}_k}(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) + g(\boldsymbol{x}_{k+1})$, then

$$F(\boldsymbol{x}) - F(\boldsymbol{x}_{k+1}) \geq \frac{1}{2} \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k} + \langle \boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}^*_{\boldsymbol{H}_k}), \boldsymbol{x} - \boldsymbol{x}_k \rangle \ \forall \boldsymbol{x} \in \mathbb{R}^n.$$

- Before the proof, we use compact notation to shorten the proof: $F_k \coloneqq F(\boldsymbol{x}_k), \nabla f_k \coloneqq \nabla f(\boldsymbol{x}_k), g_k \coloneqq g(\boldsymbol{x}_k), \ldots$.
- **Proof**. First, based on the fact that f and g are convex,

$$\begin{array}{lll} f(\boldsymbol{x}) \geq f_k + \langle \nabla f_k, \boldsymbol{x} - \boldsymbol{x}_k \rangle & & f \text{ is convex} \\ g(\boldsymbol{x}) \geq g_{k+1} + \langle \partial g_{k+1}, \boldsymbol{x} - \boldsymbol{x}_{k+1} \rangle & & g \text{ is convex} \end{array}$$

  where $\partial g_{k+1} \coloneqq \partial g(\boldsymbol{x}_{k+1})$ is a subgradient of g at $\boldsymbol{x}_{k+1}$.
  Note: we consider f at $\boldsymbol{x}_k$ and g at $\boldsymbol{x}_{k+1}$, therefore this step is actually tricky.
- Sum the two inequalities

$$F(\boldsymbol{x}) \geq f_k + \langle \nabla f_k, \boldsymbol{x} - \boldsymbol{x}_k \rangle + g_{k+1} + \langle \partial g_{k+1}, \boldsymbol{x} - \boldsymbol{x}_{k+1} \rangle \qquad (X)$$

- By hypothesis $F_{k+1} \leq m_{\boldsymbol{H}_k}(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) + g_{k+1}$

$$\begin{array}{rcl} F(\boldsymbol{x}) - F_{k+1} & \geq & F(\boldsymbol{x}) - m_{\boldsymbol{H}_k}(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) - g_{k+1} \\ & \underset{(X)}{\geq} & \\ & \geq & f_k + \langle \nabla f_k, \boldsymbol{x} - \boldsymbol{x}_k \rangle + g_{k+1} + \langle \partial g_{k+1}, \boldsymbol{x} - \boldsymbol{x}_{k+1} \rangle - m_{\boldsymbol{H}_k}(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) - g_{k+1} \\ & = & f_k + \langle \nabla f_k, \boldsymbol{x} - \boldsymbol{x}_k \rangle + \langle \partial g_{k+1}, \boldsymbol{x} - \boldsymbol{x}_{k+1} \rangle - m_{\boldsymbol{H}_k}(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k). \end{array}$$

  The second line explains why we are considering f at $\boldsymbol{x}_k$ and g at $\boldsymbol{x}_{k+1}$ in step (X).

- By definition $m_{\boldsymbol{H}_k}(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) \coloneqq f_k + \langle \nabla f_k, \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle + \frac{1}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k}$

$$\begin{array}{rcl} F(\boldsymbol{x}) - F_{k+1} & \geq & \underline{\langle \nabla f_k, \boldsymbol{x} - \boldsymbol{x}_k \rangle} + \langle \partial g_{k+1}, \boldsymbol{x} - \boldsymbol{x}_{k+1} \rangle \underline{- \langle \nabla f_k, \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle} - \frac{1}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k} \\ & = & \underline{\langle \nabla f_k, \boldsymbol{x} - \boldsymbol{x}_{k+1} \rangle} + \langle \partial g_{k+1}, \boldsymbol{x} - \boldsymbol{x}_{k+1} \rangle - \frac{1}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k} \\ & = & \underline{\langle \nabla f_k + \partial g_{k+1}, \boldsymbol{x} - \boldsymbol{x}_{k+1} \rangle} - \frac{1}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k} \end{array} \qquad (\mathsf{O})$$

- We now deal with $\nabla f_k + \partial g_{k+1}$. By definition $\boldsymbol{x}_{k+1}$ is the minimizer of $m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) + g(\boldsymbol{\xi})$, so

$$\boldsymbol{0} \in \nabla m_{\boldsymbol{H}_k}(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) + \partial g_{k+1} \qquad \Longleftrightarrow \qquad \boldsymbol{0} \in \nabla f_k + \boldsymbol{H}_k(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) + \partial g_{k+1}$$

  in other words $\boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}_{k+1}) \in \nabla f(\boldsymbol{x}_k) + \partial g(\boldsymbol{x}_{k+1})$. Put this into (O) gives

$$F(\boldsymbol{x}) - F_{k+1} \geq \langle \boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}_{k+1}), \boldsymbol{x} - \boldsymbol{x}_{k+1} \rangle - \frac{1}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k}.$$

  The proof completes by considering $\langle \boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}_{k+1}), \boldsymbol{x} - \boldsymbol{x}_k + \boldsymbol{x}_k - \boldsymbol{x}_{k+1} \rangle$. □

Convergence with no line search ($\alpha_k = 1$), part 2

$$(\mathcal{P}) \ : \ \min_{\boldsymbol{x}} \Big\{ F(\boldsymbol{x}) \coloneqq f(\boldsymbol{x}) + g(\boldsymbol{x}) \Big\}$$

$$\text{algo} \quad : \begin{cases} \boldsymbol{x}^*_{\boldsymbol{H}_k} = \operatorname*{argmin}_{\boldsymbol{\xi}} \ m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) + g(\boldsymbol{\xi}), \quad m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) \coloneqq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{\xi} - \boldsymbol{x}_k \rangle + \frac{1}{2} \|\boldsymbol{\xi} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k} \\ \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k (\boldsymbol{x}^*_{\boldsymbol{H}_k} - \boldsymbol{x}_k) \end{cases}$$

Lemma 4. Suppose $F(\boldsymbol{x}_{k+1}) \leq m_{\boldsymbol{H}_k}(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) + g(\boldsymbol{x}_{k+1})$, then

$$F_k - F_{k+1} \geq \frac{\left(F_{k+1} - F^*\right)^2}{2\bar{\lambda}\delta^2}$$

- $F^* = F(\boldsymbol{x}^*)$ for $\boldsymbol{x}^* \in \operatorname{argmin} F$
- $\bar{\lambda} = \max_{i=1,\dots,k} \lambda(\boldsymbol{H}_i)$
- Define the sublevel set $L_{F(\boldsymbol{x}_0)} \coloneqq \Big\{ \boldsymbol{x} \in \mathbb{R}^n \ : \ F(\boldsymbol{x}) \leq F(\boldsymbol{x}_0) \Big\}$.
- $\delta \coloneqq$ diameter of $L_{F(\boldsymbol{x}_0)}$.

- To prove lemma 4 we use lemma 3 we just proved

$$F(\boldsymbol{x}) - F(\boldsymbol{x}_{k+1}) \geq \langle \boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}_{k+1}), \boldsymbol{x} - \boldsymbol{x}_k \rangle + \frac{1}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k}.$$

with $\boldsymbol{x} = \boldsymbol{x}^*$.

$$\begin{aligned}
F^* - F_{k+1} &\geq \langle \boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}_{k+1}), \boldsymbol{x}^* - \boldsymbol{x}_k \rangle + \tfrac{1}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k} \\
&= \tfrac{1}{2}\langle \boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}_{k+1}), 2\boldsymbol{x}^* - 2\boldsymbol{x}_k + \boldsymbol{x}_k - \boldsymbol{x}_{k+1} \rangle \\
&= \tfrac{1}{2}\langle \boldsymbol{H}_k(\boldsymbol{x}^* - \boldsymbol{x}_k) - \boldsymbol{H}_k(\boldsymbol{x}^* - \boldsymbol{x}_{k+1}), (\boldsymbol{x}^* - \boldsymbol{x}_k) + (\boldsymbol{x}^* - \boldsymbol{x}_{k+1}) \rangle \\
&= \tfrac{1}{2}\Big(\|\boldsymbol{x}^* - \boldsymbol{x}_{k+1}\|^2_{\boldsymbol{H}_k} - \|\boldsymbol{x}^* - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k} \Big) \\
&= \tfrac{1}{2}\Big(\|\boldsymbol{x}^* - \boldsymbol{x}_{k+1}\|_{\boldsymbol{H}_k} - \|\boldsymbol{x}^* - \boldsymbol{x}_k\|_{\boldsymbol{H}_k} \Big) \Big(\|\boldsymbol{x}^* - \boldsymbol{x}_{k+1}\|_{\boldsymbol{H}_k} + \|\boldsymbol{x}^* - \boldsymbol{x}_k\|_{\boldsymbol{H}_k} \Big)
\end{aligned}$$

- For $\|\boldsymbol{x}^* - \boldsymbol{x}_{k+1}\|_{\boldsymbol{H}_k} - \|\boldsymbol{x}^* - \boldsymbol{x}_k\|_{\boldsymbol{H}_k}$, by triangle inequality

$$\|\boldsymbol{x}^* - \boldsymbol{x}_{k+1}\|_{\boldsymbol{H}_k} + \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_{\boldsymbol{H}_k} \geq \|\boldsymbol{x}^* - \boldsymbol{x}_k\|_{\boldsymbol{H}_k} \implies \|\boldsymbol{x}^* - \boldsymbol{x}_{k+1}\|_{\boldsymbol{H}_k} - \|\boldsymbol{x}^* - \boldsymbol{x}_k\|_{\boldsymbol{H}_k} \geq -\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_{\boldsymbol{H}_k}$$

therefore

$$\begin{aligned}
& F^* - F_{k+1} \geq \tfrac{-1}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_{\boldsymbol{H}_k} \Big(\|\boldsymbol{x}^* - \boldsymbol{x}_{k+1}\|_{\boldsymbol{H}_k} + \|\boldsymbol{x}^* - \boldsymbol{x}_k\|_{\boldsymbol{H}_k} \Big) \\
\implies \quad & \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_{\boldsymbol{H}_k} \geq \frac{-2\Big(F^* - F_{k+1}\Big)}{\|\boldsymbol{x}^* - \boldsymbol{x}_{k+1}\|_{\boldsymbol{H}_k} + \|\boldsymbol{x}^* - \boldsymbol{x}_k\|_{\boldsymbol{H}_k}} = 2\frac{F_{k+1} - F^*}{\|\boldsymbol{x}^* - \boldsymbol{x}_{k+1}\|_{\boldsymbol{H}_k} + \|\boldsymbol{x}^* - \boldsymbol{x}_k\|_{\boldsymbol{H}_k}} \qquad \text{(AA)}
\end{aligned}$$

$$\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|_{\boldsymbol{H}_k} \geq 2\frac{F_{k+1}-F^*}{\|\boldsymbol{x}^*-\boldsymbol{x}_{k+1}\|_{\boldsymbol{H}_k}+\|\boldsymbol{x}^*-\boldsymbol{x}_k\|_{\boldsymbol{H}_k}} \tag{AA}$$

- Now focus on the norm $\|\boldsymbol{x}^*-\boldsymbol{x}_k\|_{\boldsymbol{H}_k}$. Based on norm inequality,

$$\|\boldsymbol{x}^*-\boldsymbol{x}_k\|_{\boldsymbol{H}_k} \;\leq\; \|\boldsymbol{H}_k\|_2\cdot\|\boldsymbol{x}^*-\boldsymbol{x}_k\|_2 \;=\; \lambda(\boldsymbol{H}_k)\|\boldsymbol{x}^*-\boldsymbol{x}_k\|_2 \;\leq\; \bar{\lambda}\|\boldsymbol{x}^*-\boldsymbol{x}_k\|_2 \tag{$\cup$}$$

  where the last inequality is due to definition that $\bar{\lambda}=\max\limits_{i=1,\ldots,k}\lambda(\boldsymbol{H}_i)$.

- ($\cup$) holds for any k, so

$$\|\boldsymbol{x}^*-\boldsymbol{x}_{k+1}\|_{\boldsymbol{H}_k}+\|\boldsymbol{x}^*-\boldsymbol{x}_k\|_{\boldsymbol{H}_k} \;\leq\; \bar{\lambda}\Big(\|\boldsymbol{x}^*-\boldsymbol{x}_{k+1}\|_2+\|\boldsymbol{x}^*-\boldsymbol{x}_k\|_2\Big)$$

  Therefore

$$\frac{1}{\|\boldsymbol{x}^*-\boldsymbol{x}_{k+1}\|_{\boldsymbol{H}_k}+\|\boldsymbol{x}^*-\boldsymbol{x}_k\|_{\boldsymbol{H}_k}} \;\geq\; \frac{\bar{\lambda}^{-1}}{\|\boldsymbol{x}^*-\boldsymbol{x}_{k+1}\|_2+\|\boldsymbol{x}^*-\boldsymbol{x}_k\|_2} \tag{BB}$$

- Put (BB) into (AA)

$$\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|_{\boldsymbol{H}_k} \geq \frac{2}{\bar{\lambda}}\frac{F_{k+1}-F^*}{\|\boldsymbol{x}^*-\boldsymbol{x}_{k+1}\|_2+\|\boldsymbol{x}^*-\boldsymbol{x}_k\|_2} \tag{CC}$$

- Too proceed, we need to make usre of sublevel set argument to bound the terms $\|\boldsymbol{x}^*-\boldsymbol{x}_{k+1}\|_2$ and $\|\boldsymbol{x}^*-\boldsymbol{x}_k\|_2$

1-page review of the diameter of a sublevel set

- Let $d(\boldsymbol{x}, \boldsymbol{y})$ be the distance between two points $\boldsymbol{x}, \boldsymbol{y}$.
- Let $\mathcal{A}$ be a compact set.
- Diameter of $\mathcal{A}$, denoted as $\operatorname{diam} \mathcal{A}$, is defined as the supremum (if it exists) of the distance $d(\boldsymbol{x}, \boldsymbol{y})$ across all pairs of points $\boldsymbol{x}, \boldsymbol{y}$ in $\mathcal{A}$

$$\operatorname{diam} \mathcal{A} \coloneqq \sup \Big\{ d(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{x}, \boldsymbol{y} \in \mathcal{A} \Big\}.$$

- Now our $\mathcal{A}$ is the sublevel set $L_{F(\boldsymbol{x}_0)} \coloneqq \Big\{ \boldsymbol{x} \in \mathbb{R}^n \ : \ F(\boldsymbol{x}) \leq F(\boldsymbol{x}_0) \Big\}$.
- Since F is a convex function, thus the sublevel set $L_{F(\boldsymbol{x}_0)}$ is a convex set and it is bounded.
- In Euclidean space $\mathbb{R}^n$, distance $d(\boldsymbol{x}, \boldsymbol{y})$ is the Euclidean norm $\|\boldsymbol{x} - \boldsymbol{y}\|_2$
- The diameter of $L_{F(\boldsymbol{x}_0)}$ is thus

$$\operatorname{diam} L_{F(\boldsymbol{x}_0)} = \sup \Big\{ \|\boldsymbol{x} - \boldsymbol{y}\|_2 \ : \ F(\boldsymbol{x}) \leq F(\boldsymbol{x}_0), F(\boldsymbol{y}) \leq F(\boldsymbol{x}_0) \Big\}$$

  i.e., $\operatorname{diam} L_{F(\boldsymbol{x}_0)}$ is the upper bound of the distance $\|\boldsymbol{x} - \boldsymbol{y}\|_2$ for any $\boldsymbol{x}, \boldsymbol{y}$ that $F(\boldsymbol{x}) \leq F(\boldsymbol{x}_0), F(\boldsymbol{y}) \leq F(\boldsymbol{x}_0)$.
- Such supremum exists for $\operatorname{diam} L_{F(\boldsymbol{x}_0)}$ because F is convex so $L_{F(\boldsymbol{x}_0)}$ is a convex and bounded.

Using the sublevel set argument

- We want to proceed with

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_{\boldsymbol{H}_k} \geq \frac{2}{\bar{\lambda}} \frac{F_{k+1} - F^*}{\|\boldsymbol{x}^* - \boldsymbol{x}_{k+1}\|_2 + \|\boldsymbol{x}^* - \boldsymbol{x}_k\|_2} \tag{CC}$$

- We have
  - sublevel set $L_{F(\boldsymbol{x}_0)} \coloneqq \left\{\boldsymbol{x} \in \mathbb{R}^n \ : \ F(\boldsymbol{x}) \leq F(\boldsymbol{x}_0)\right\}$.
  - $\delta \coloneqq \operatorname{diam} L_{F(\boldsymbol{x}_0)} = \sup\left\{\|\boldsymbol{x} - \boldsymbol{y}\|_2 \ : \ F(\boldsymbol{x}) \leq F(\boldsymbol{x}_0), F(\boldsymbol{y}) \leq F(\boldsymbol{x}_0)\right\}$
  - $F^* = F(\boldsymbol{x}^*) \leq F(\boldsymbol{x}_0)$ by definition $\boldsymbol{x}^*$ is a (global) minimizer
  - $F(\boldsymbol{x}_k) \leq F(\boldsymbol{x}_0)$ and $F(\boldsymbol{x}_{k+1}) \leq F(\boldsymbol{x}_0)$ by the descent property of the update (Lemma 2)

  all these imply the three points $\boldsymbol{x}^*, \boldsymbol{x}_{k+1}, \boldsymbol{x}_k$ are inside the sublevel set $L_{F(\boldsymbol{x}_0)}$ and thus the distances $\|\boldsymbol{x}^* - \boldsymbol{x}_k\|_2$ and $\|\boldsymbol{x}^* - \boldsymbol{x}_{k+1}\|_2$ are upper bounded by δ

- Thus we have $\|\boldsymbol{x}^* - \boldsymbol{x}_k\|_2 + \|\boldsymbol{x}^* - \boldsymbol{x}_{k+1}\|_2 \leq 2\delta \implies \dfrac{1}{\|\boldsymbol{x}^* - \boldsymbol{x}_k\|_2 + \|\boldsymbol{x}^* - \boldsymbol{x}_{k+1}\|_2} \geq \dfrac{1}{2\delta}$. Now (CC) becomes

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_{\boldsymbol{H}_k} \geq \frac{F_{k+1} - F^*}{\bar{\lambda}\delta} \tag{DD}$$

- Lastly, using Lemma 3 again on $\boldsymbol{x} = \boldsymbol{x}_k$ gives

$$F(\boldsymbol{x}_k) - F(\boldsymbol{x}_{k+1}) \geq \frac{1}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k}. \tag{EE}$$

Combine (DD) and (EE) finishes the proof. □

Convergence with no line search ($\alpha_k = 1$), part 3

$$(\mathcal{P}) \ : \ \min_{\boldsymbol{x}} \Big\{ F(\boldsymbol{x}) \coloneqq f(\boldsymbol{x}) + g(\boldsymbol{x}) \Big\}$$

$$\text{algo} \quad : \begin{cases} \boldsymbol{x}^*_{\boldsymbol{H}_k} = \operatorname*{argmin}_{\boldsymbol{\xi}} \ m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) + g(\boldsymbol{\xi}), \quad m_{\boldsymbol{H}_k}(\boldsymbol{\xi}, \boldsymbol{x}_k) \coloneqq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{\xi} - \boldsymbol{x}_k \rangle + \frac{1}{2} \|\boldsymbol{\xi} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k} \\ \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k (\boldsymbol{x}^*_{\boldsymbol{H}_k} - \boldsymbol{x}_k) \end{cases}$$

Lemma 5. For $\alpha_k \equiv 1$, $\boldsymbol{H}_k \succ \boldsymbol{0}$, suppose $F(\boldsymbol{x}_{k+1}) \leq m_{\boldsymbol{H}_k}(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) + g(\boldsymbol{x}_{k+1})$, then for all $\boldsymbol{x} \in \mathbb{R}^n$,

$$\begin{aligned} F(\boldsymbol{x}) &\geq F(\boldsymbol{x}_{k+1}) + \langle \boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}^*_{\boldsymbol{H}_k}), \boldsymbol{x} - \boldsymbol{x}_k \rangle + \tfrac{1}{2} \|\boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}^*_{\boldsymbol{H}_k})\|^2_{\boldsymbol{H}_k^{-1}} \\ &\geq F(\boldsymbol{x}_{k+1}) + \langle \boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}^*_{\boldsymbol{H}_k}), \boldsymbol{x} - \boldsymbol{x}_k \rangle + \tfrac{1}{2\lambda_1(\boldsymbol{H}_k)} \|\boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}^*_{\boldsymbol{H}_k})\|^2_2 \end{aligned}$$

- **Proof**. First, starting from Lemma 3

$$\begin{aligned} F(\boldsymbol{x}) &\geq F(\boldsymbol{x}_{k+1}) + \langle \boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}^*_{\boldsymbol{H}_k}), \boldsymbol{x} - \boldsymbol{x}_k \rangle + \tfrac{1}{2} \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2_{\boldsymbol{H}_k} \\ &= F(\boldsymbol{x}_{k+1}) + \langle \boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}^*_{\boldsymbol{H}_k}), \boldsymbol{x} - \boldsymbol{x}_k \rangle + \tfrac{1}{2} \|\boldsymbol{H}_k(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)\|^2_{\boldsymbol{H}_k^{-1}} \\ &\geq F(\boldsymbol{x}_{k+1}) + \langle \boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}^*_{\boldsymbol{H}_k}), \boldsymbol{x} - \boldsymbol{x}_k \rangle + \tfrac{1}{2\lambda_1(\boldsymbol{H}_k)} \|\boldsymbol{H}_k(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)\|^2_2 \end{aligned}$$

- For the last inequality sign, see explanation next page.

Inequality on positive definite matrices

- Let $\lambda_1(\cdot)$ be the maximum eigenvalue of a matrix, then from the fact that $\lambda_1 \geq \lambda_{i>1}$

$$\lambda_1 \geq \lambda_2,\ \lambda_1 \geq \lambda_3,\ \ldots\ \lambda_1 \geq \lambda_n \underbrace{\geq 0}_{\boldsymbol{H}_k \succ \mathbf{0}}, \quad \Longrightarrow \quad \frac{1}{\lambda_n} \geq \frac{1}{\lambda_1},\ \ldots\ \frac{1}{\lambda_2} \geq \frac{1}{\lambda_1}.$$

- Therefore, for matrices, we have the following inequality

$$\underbrace{\boldsymbol{H}_k}_{\lambda_1,\lambda_2,\lambda_3,\ldots} \preceq \underbrace{\lambda_1(\boldsymbol{H}_k)\boldsymbol{I}}_{\lambda_1,\lambda_1,\lambda_1,\ldots} \quad \Longrightarrow \quad \underbrace{(\boldsymbol{H}_k)^{-1}}_{\frac{1}{\lambda_n},\frac{1}{\lambda_{n-1}},\ldots,\frac{1}{\lambda_1}} \succeq \underbrace{\frac{1}{\lambda_1(\boldsymbol{H}_k)}\boldsymbol{I}}_{\frac{1}{\lambda_1},\ldots,\frac{1}{\lambda_1}}$$

- Therefore, for any $\boldsymbol{\xi} \in \mathbb{R}^n$,

$$\|\boldsymbol{\xi}\|^2_{(\boldsymbol{H}_k)^{-1}} \quad \geq \quad \|\boldsymbol{\xi}\|^2_{\frac{1}{\lambda_1(\boldsymbol{H}_k)}\boldsymbol{I}} \quad = \quad \frac{1}{\lambda_1(\boldsymbol{H}_k)}\|\boldsymbol{\xi}\|_2^2.$$

- Put $\boldsymbol{\xi} = \boldsymbol{H}_k(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)$gives

$$\|\boldsymbol{H}_k(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)\|^2_{\boldsymbol{H}_k^{-1}} \quad \geq \quad \frac{1}{\lambda_1(\boldsymbol{H}_k)}\|\boldsymbol{H}_k(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)\|_2^2.$$

Comment on Lemma 5

Lemma 5. For $\alpha_k \equiv 1$, $\boldsymbol{H}_k \succ \boldsymbol{0}$, suppose $F(\boldsymbol{x}_{k+1}) \leq m_{\boldsymbol{H}_k}(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) + g(\boldsymbol{x}_{k+1})$, then for all $\boldsymbol{x} \in \mathbb{R}^n$,

$$\begin{aligned} F(\boldsymbol{x}) &\geq F(\boldsymbol{x}_{k+1}) + \langle \boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}^*_{\boldsymbol{H}_k}), \boldsymbol{x} - \boldsymbol{x}_k \rangle + \tfrac{1}{2}\|\boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}^*_{\boldsymbol{H}_k})\|^2_{\boldsymbol{H}_k^{-1}} \\ &\geq F(\boldsymbol{x}_{k+1}) + \langle \boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}^*_{\boldsymbol{H}_k}), \boldsymbol{x} - \boldsymbol{x}_k \rangle + \tfrac{1}{2\lambda_1(\boldsymbol{H}_k)}\|\boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}^*_{\boldsymbol{H}_k})\|_2^2 \end{aligned}$$

- Put $\boldsymbol{x} = \boldsymbol{x}_k$, the inner product term is gone and we have

$$\begin{aligned} F_k &\geq F_{k+1} + \tfrac{1}{2}\|\boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}^*_{\boldsymbol{H}_k})\|^2_{\boldsymbol{H}_k^{-1}} \\ &\geq F_{k+1} + \tfrac{1}{2\lambda_1(\boldsymbol{H}_k)}\|\boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}^*_{\boldsymbol{H}_k})\|_2^2 \end{aligned}$$

- Re-arrange

$$F_{k+1} \leq F_k - \frac{1}{2\lambda_1(\boldsymbol{H}_k)}\|\boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}^*_{\boldsymbol{H}_k})\|_2^2$$

  Recall the discussion on Lemma 0, the term $\boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}^*_{\boldsymbol{H}_k})$ is related to the subgradient of F

- In fact, such express resembles the sufficient descent condition of gradient descent (see here)

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \frac{L}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_2^2 = f(\boldsymbol{x}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|_2^2.$$

A lemma on decreasing sequence

Lemma 6. If $\{\omega_k\}_{k\in\mathbb{N}}$

- is a decreasing sequence
- $\omega_k - \omega_{k+1} \geq \dfrac{(\omega_{k+1}-\omega^*)^2}{\mu}$ for all k
- $\omega_1 - \omega^* \leq 4\mu$

then for all k

$$\omega_k - \omega^* \leq \frac{4\mu}{k}.$$

- **Proof**. By induction. Case $k = 1$ is true by assumption.
- Case $k+1$

$$\omega_{k+1} - \omega^* \;=\; \omega_k - \omega^* + \omega_{k+1} - \omega_k \;\leq\; \omega_k - \omega^* - \frac{(\omega_{k+1}-\omega^*)^2}{\mu} \;\leq\; \frac{4\mu}{k} - \frac{(\omega_{k+1}-\omega^*)^2}{\mu}.$$

So we have

$$\omega_{k+1} - \omega^* \leq \frac{4\mu}{k} - \frac{(\omega_{k+1}-\omega^*)^2}{\mu} \iff \frac{(\omega_{k+1}-\omega^*)^2}{\mu} + \omega_{k+1} - \omega^* - \frac{4\mu}{k} \leq 0.$$

$$\frac{(\omega_{k+1}-\omega^*)^2}{\mu}+\omega_{k+1}-\omega^*-\frac{4\mu}{k}\leq 0$$

- Let $\nu=\omega_{k+1}-\omega^*$ and $p_k=\frac{4\mu}{k}$

$$\frac{\nu^2}{\mu}+\nu-p_k\leq 0$$

which has a nonnegative solution given by

$$\nu=\frac{-\mu+\sqrt{\mu^2+4p_k\mu}}{2}=\frac{2p_k}{1+\sqrt{1+\frac{4p_k}{\mu}}}$$

- $f(x)=\dfrac{1}{1+\sqrt{1+x}}$ is convex and on any interval $[0,\alpha]$ it is bounded above by its secant interpolation.

- To finish the proof, consider two separate cases, when $k=1$ to show that the lemma holds for $k+1=2$, and when $k\geq 2$ to show that the lemma holds for $k+1\geq 3$.

- For the second case we need to make use of the fact that $\frac{1}{k}-\frac{1}{k^2}\leq\frac{1}{k+1}$.

Convergence rate $\frac{1}{k}$

Theorem 1. Suppose

- $\boldsymbol{x}_{k+1} = \boldsymbol{x}^*_{\boldsymbol{H}_k}$ ($\alpha_k \equiv 1$)
- $\delta := \mathrm{diam} L_{F(\boldsymbol{x}_0)} = \sup\left\{\|\boldsymbol{x}-\boldsymbol{y}\|_2 : F(\boldsymbol{x}) \le F(\boldsymbol{x}_0), F(\boldsymbol{y}) \le F(\boldsymbol{x}_0)\right\}$ (sublevel set for Lemma 4)
- $\bar{\lambda} = \max\limits_{i=1,\dots,k} \lambda(\boldsymbol{H}_i)$ (sublevel set for Lemma 4)
- $F(\boldsymbol{x}^*_{\boldsymbol{H}_k}) \le m_{\boldsymbol{H}_k}(\boldsymbol{x}^*_{\boldsymbol{H}_k}, \boldsymbol{x}_k) + g(\boldsymbol{x}^*_{\boldsymbol{H}_k})$ (condition on $\boldsymbol{H}_k$ for Lemma 2)

then for all k

$$F(\boldsymbol{x}_k) - F^* \le \frac{\max\left\{8\bar{\lambda}^2\delta^2, F(\boldsymbol{x}_0) - \boldsymbol{F}^*\right\}}{k}.$$

- **Proof**. Combine Lemma 4 and Lemma 6. □

Table of Contents

Last page - summary

Discussed

- Proximal Newton's method.
- Convergence of Proximal Newton's method.

Not discussed

- Proximal Newton's method with Nesterov's acceleration
- Problem-specific implementation of $\boldsymbol{H}_k$ (Identity minus rank one in the paper)
- Other implementation of $\boldsymbol{H}_k$ (Identity minus multiple rank matrix)

Reference

Sahar Karimi and Stephen Vavasis, "IMRO: A Proximal Quasi-Newton Method for Solving ℓ_1-Regularized Least Squares Problems", SIAM J. Optimization, 2017.

End of document