Some recent advances in SDP performance analysis

Etienne de Klerk (joint work with Hadi Abbaszadehpeivasti and Moslem Zamani)

PEP talks, UCL

The iterative methods of Jacobi, Gauss-Seidel, and Cauchy

The iterative method of Jacobi to solve $A\mathbf{x} = b$ $(A = (a_{ij}) \succ 0)$

Set N (iterations) and pick $\mathbf{x}^0 \in \mathbb{R}^n$.

Set $D = I \circ A$ (Hadamard product).

For $k = 0, 1, \ldots, N$ perform the following step:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - D^{-1}\left(A\mathbf{x}^k - b\right) \iff x_i^{k+1} = \frac{1}{a_{ii}}\left(b_i - \sum_{j \neq i}^{n} a_{ij} x_j^k\right) \ \forall i.$$

Carl Gustav Jacob Jacobi (1804–1851)

The Gauss-Seidel iterative method to solve $A\mathbf{x} = b$ ($A \succ 0$)

Set N and pick $\mathbf{x}^0 \in \mathbb{R}^n$.

For $k = 0, 1, \ldots, N-1$ perform the following for $i = 1, \ldots, n$:

$$x_i^{k+1} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij} x_j^k \right).$$

Johann Carl Friedrich Gauss (1777–1855)   Philipp Ludwig von Seidel (1821 – 1896)

The gradient descent method of Cauchy

Input: $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{x}^0 \in \mathbb{R}^n$, number of steps N and $\{t_k\}_{k=0}^N$ (step lengths).

for $k = 0, 1, \ldots, N$

$$\mathbf{x}^{k+1} = \mathbf{x}^k - t_k \nabla f(\mathbf{x}^k)$$

Augustin-Louis Cauchy (1789–1857)

(Studied the gradient descent method in 1847.)

Jacobi vs Cauchy

- Solving $A\mathbf{x} = b$ with $A \succ 0$ is the same as minimizing $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x} - b^\top \mathbf{x}$.
- We obtain the method of Jacobi from Cauchy;'s method by replacing the direction $-\nabla f(\mathbf{x})$ by $-D^{-1}\nabla f(\mathbf{x}) = -D^{-1}(A\mathbf{x} - b)$.
- This simply amounts to a change of inner product ... (next slide)

Change of inner product

- The gradient of f at $\mathbf{x}$ w.r.t. $\langle \cdot, \cdot \rangle$ is denoted by $g(\mathbf{x})$:

$$\lim_{\|\mathbf{h}\| \to 0} \frac{f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - \langle g(\mathbf{x}), \mathbf{h} \rangle}{\|\mathbf{h}\|} = 0.$$

- If $\langle \cdot, \cdot \rangle$ is the Euclidean dot product then $g(\mathbf{x}) = \nabla f(\mathbf{x}) = \left[\frac{\partial f(\mathbf{x})}{\partial x_i}\right]_{i=1,\ldots,n}.$
- If $B : \mathbb{R}^n \to \mathbb{R}^n$ self-adjoint positive definite linear operator, define $\langle \cdot, \cdot \rangle_B$ via $\langle \mathbf{x}, \mathbf{y} \rangle_B = \langle \mathbf{x}, B\mathbf{y} \rangle$ $\forall x, y \in \mathbb{R}^n$.
- Change of gradient if we change the inner product:

$$\langle \cdot, \cdot \rangle \to \langle \cdot, \cdot \rangle_B \Rightarrow g(\mathbf{x}) \to B^{-1} g(\mathbf{x}).$$

Extension to the nonlinear Jacobi's method

- Assume now f twice continuously differentiable with convex domain $D_f \subset \mathbb{R}^n$ ...
- ... with positive definite Hessian $H(\mathbf{x})$ for all $\mathbf{x} \in D_f$:

$$\lim_{\|\mathbf{h}\| \to 0} \frac{\|g(\mathbf{x}+\mathbf{h}) - g(\mathbf{x}) - H(\mathbf{x})\mathbf{h}\|}{\|\mathbf{h}\|} = 0.$$

- Key Idea: fix $\mathbf{x} \in D_f$ and change the inner product:

$$\langle \cdot, \cdot \rangle \to \langle \cdot, \cdot \rangle_{I \circ \nabla^2 f(\mathbf{x})} \Rightarrow g(\mathbf{x}) \to (I \circ \nabla^2 f(\mathbf{x}))^{-1} g(\mathbf{x}),$$

  and $-(I \circ \nabla^2 f(\mathbf{x}))^{-1} g(\mathbf{x})$ is precisely the nonlinear Jacobi direction.
- So we may analyse one step of Jacobi's method by using a suitable inner product, or N steps for quadratic f.

Functions of bounded curvature (aka hypoconvex)

Smooth, strongly convex functions

- Convex quadratic f are examples of smooth, convex functions;
- A function f has a maximum curvature $0 \leq L < \infty$ if

$$\mathbf{x} \mapsto \frac{L}{2}\|\mathbf{x}\|^2 - f(\mathbf{x}) \text{ is convex } \ldots$$

- ... and minimum curvature $-\infty < \mu \leq L$ if

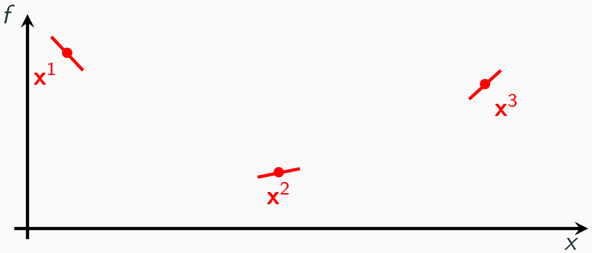
$$\mathbf{x} \mapsto f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|^2 \text{ is convex.}$$

  Notation: $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$. Note that $\langle \cdot, \cdot \rangle$ determines the class $\mathcal{F}_{\mu,L}(\mathbb{R}^n)$.
    - $\mathcal{F}_{\mu,L}(\mathbb{R}^n)$: L-smooth convex functions if $\mu \geq 0$ (strongly convex if $\mu > 0$).
    - If $f \in C^2(\mathbb{R}^n)$, then $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$ iff

$$\mu I \preceq H(\mathbf{x}) \preceq LI, \quad \forall\, \mathbf{x} \in \mathbb{R}^n.$$

Interpolation Problem

Consider an index set S, and given values $\left\{(\mathbf{x}^i, \mathbf{g}^i, f^i)\right\}_{i\in S}$ where $\mathbf{x}^i \in \mathbb{R}^n$, $\mathbf{g}^i \in \mathbb{R}^n$ and $f^i \in \mathbb{R}$.



$$?\exists \;\; f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)\text{:}\;\; f(\mathbf{x}^i) = f^i, \quad \text{and} \quad \mathbf{g}^i = \nabla f(\mathbf{x}^i), \qquad \forall i \in S.$$

If yes, we say $\left\{(\mathbf{x}^i, \mathbf{g}^i, f^i)\right\}_{i\in S}$ is $\mathcal{F}_{\mu,L}(\mathbb{R}^n)$-interpolable.

Theorem (Taylor, Hendrickx, and Glineur (2017a,b), Rotaru, Glineur, Panagiotis (2022))

Let $-\infty < \mu \leq L \leq \infty$. The following statements are equivalent:

1. $\left\{(\mathbf{x}^i, \mathbf{g}^i, f^i)\right\}_{i\in S}$ *is* $\mathcal{F}_{\mu,L}(\mathbb{R}^n)$*-interpolable;*
2. $\forall i, j \in S$*:*

$$\frac{1}{L}\left\|\mathbf{g}^i - \mathbf{g}^j\right\|^2 + \mu\left\|\mathbf{x}^i - \mathbf{x}^j\right\|^2 - \frac{2\mu}{L}\left\langle \mathbf{g}^j - \mathbf{g}^i, \mathbf{x}^j - \mathbf{x}^i\right\rangle$$
$$\leq 2(1 - \tfrac{\mu}{L})\left(f^i - f^j - \left\langle \mathbf{g}^j, \mathbf{x}^i - \mathbf{x}^j\right\rangle\right). \qquad (\clubsuit)$$

The proof is constructive.

Interpolation theorem (ctd.)

The interpolation theorem in the smooth ($L < \infty$) and strongly convex case ($\mu > 0$) is from:

A.B. Taylor, J.M. Hendrickx, and F. Glineur (2017a)

Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161:1-2, 307–345, 2017.

The L-smooth case $\mu = -L$ is from:

A.B. Taylor, J.M. Hendrickx, and F. Glineur (2017b)

Exact worst-case performance of first-order methods for composite convex optimization. *SIAM Journal on Optimization* 27(3), 1283–1313.

The general case was shown in:

Rotaru, T., Glineur, F., & Patrinos, P. (2022)

Tight convergence rates of the gradient method on hypoconvex functions. arXiv preprint arXiv:2203.00775.

Performance estimation: one iteration gradient method

Worst-case computation

Performance estimation problem for first gradient step $\mathbf{x}^0 \to \mathbf{x}^1$:

$$\max_{f, \mathbf{x}^0, \mathbf{x}^1, \mathbf{x}^\star} \; f(\mathbf{x}^1) - f(\mathbf{x}^\star) \text{ OR } \|\mathbf{x}^1 - \mathbf{x}^\star\|^2 \text{ OR } \|g(\mathbf{x}^1)\|^2$$

$$\begin{aligned} \text{s.t. } & f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n) \\ & \mathbf{x}^\star \text{ optimal for } f \\ & \mathbf{x}^1 = \mathbf{x}^0 - t_1 g(\mathbf{x}^0) \\ & f(\mathbf{x}^0) - f(\mathbf{x}^\star) \le R \text{ OR } \|\mathbf{x}^0 - \mathbf{x}^*\|^2 \le R \text{ OR } \|g(\mathbf{x}^0)\|^2 \le R \end{aligned}$$

Key idea - SDP reformulation; Seminal paper:

Y. Drori and M. Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1-2):451–482, 2014.

Extension to 'noisy' gradients

- We allow for inaccurate ('noisy') gradients, say $\hat{g}(\mathbf{x})$, in the sense

$$\|g(\mathbf{x}) - \hat{g}(\mathbf{x})\|^2 \leq \varepsilon \|g(\mathbf{x})\|^2 \quad \text{for all } \mathbf{x} \in \mathbb{R}^n,$$

  for some given $\varepsilon > 0$.
- This corresponds, e.g. to computing the gradient to a fixed number of accurate digits.

Semidefinite programming (SDP) formulation

- Parameters: $L \geq \mu > 0$, $R > 0$, $\varepsilon > 0$; $t_1 > 0$
- Variables: $\left\{(\mathbf{x}^i, \mathbf{g}^i, \hat{\mathbf{g}}^i, f^i)\right\}_{i \in S}$ $\quad (S = \{*, 0, 1\})$.

Performance estimation SDP:

$$\max f^1 - f^* \text{ OR } \|\mathbf{x}^1 - \mathbf{x}^*\|^2 \text{ OR } \|\mathbf{g}^1\|^2$$

s.t.	$\left\{(\mathbf{x}^i, \mathbf{g}^i, f^i)\right\}_{i \in S}$ satisfy (♣)	Functional class
	$\mathbf{g}^* = \mathbf{0}$	Optimal point
	$\mathbf{x}^1 = \mathbf{x}^0 - t_1 \hat{\mathbf{g}}^0$	Algorithm
	$\|\mathbf{g}^0 - \hat{\mathbf{g}}^0\|^2 \leq \varepsilon \|\mathbf{g}^0\|^2$	Noisy gradient
	$f^0 - f^* \leq R$ OR $\|\mathbf{x}^0 - \mathbf{x}^*\|^2 \leq R$ OR $\|\mathbf{g}^0\|^2 \leq R$	Initial distance

SDP formulation: simply consider the Gram matrix of $\left\{\mathbf{x}^i, \mathbf{g}^i, \hat{\mathbf{g}}^i\right\}_{i \in S}$ w.r.t. $\langle \cdot, \cdot \rangle$.

Worst-case of noisy gradient method

Theorem (De Klerk, Glineur, Taylor (2020))

Let $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n)$. *Let* $\kappa = \mu/L$. *If* $\varepsilon \leq \frac{2\mu}{L+\mu}$, *and* $t_1 = \frac{2\mu - \varepsilon(L+\mu)}{(1-\varepsilon)\mu(L+\mu)}$ $(= \frac{2}{L+\mu}$ *if* $\varepsilon = 0)$, *one has*

$$
\begin{aligned}
f(\mathbf{x}^1) - f(\mathbf{x}^*) &\leq \left(\frac{1-\kappa}{1+\kappa} + \varepsilon\right)^2 (f(\mathbf{x}^0) - f(\mathbf{x}^*)),\\
\|g(\mathbf{x}^1)\| &\leq \left(\frac{1-\kappa}{1+\kappa} + \varepsilon\right) \|\mathbf{g}^0\|,\\
\|\mathbf{x}^1 - \mathbf{x}^*\| &\leq \left(\frac{1-\kappa}{1+\kappa} + \varepsilon\right) \|\mathbf{x}^0 - \mathbf{x}^*\|.
\end{aligned}
$$

E. de Klerk, F. Glineur, A.B. Taylor (2020)

Worst-case convergence analysis of gradient and Newton methods through semidefinite programming performance estimation. *SIAM Journal on Optimization*, Volume 30, Issue 3, 2053–2082.

Implications for the Jacobi method

The iterative method of Jacobi to solve $A\mathbf{x} = b$ ($A = (a_{ij}) \succ 0$)

$$\mathbf{x}^{k+1} = \mathbf{x}^k - t_k D^{-1}\left(A\mathbf{x}^k - b\right) \quad \text{(Recall } D = I \circ A\text{)}.$$

Assume at each iteration we only compute the direction $-D^{-1}\left(A\mathbf{x}^k - b\right)$ with ε-relative accuracy (w.r.t. $\|\cdot\|_D$).

Corollary (Abbaszadehpeivasti,De Klerk, Zamani (2023))

Let $\mu = \lambda_{\min}(D^{-1}A)$, $L = \lambda_{\max}(D^{-1}A)$, and $\kappa = \mu/L$. If $\varepsilon \le \frac{2\mu}{L+\mu}$, and $t_1 = \frac{2\mu - \varepsilon(L+\mu)}{(1-\varepsilon)\mu(L+\mu)}$ ($= \frac{2}{L+\mu}$ if $\varepsilon = 0$), one has

$$\begin{aligned} \|A\mathbf{x}^1 - b\|_{D^{-1}} &\le \left(\frac{1-\kappa}{1+\kappa} + \varepsilon\right)\|A\mathbf{x}^0 - b\|_{D^{-1}}, \\ \|\mathbf{x}^1 - \mathbf{x}^*\|_D &\le \left(\frac{1-\kappa}{1+\kappa} + \varepsilon\right)\|\mathbf{x}^0 - \mathbf{x}^*\|_D. \end{aligned}$$

Discussion

- The convergence rate without noisy residual is a classical result.
- Similar results for noisy residuals in:

Gene H Golub and Michael L Overton (1988).

The convergence of inexact Chebyshev and Richardson iterative methods for solving linear systems. *Numerische Mathematik*, 53(5):571–593.

Back to the Gauss-Seidel method

Cyclic coordinate descent methods

Generic cyclic coordinate descent to minimize f

Set number of cycles K, $\{t_k\}_{k=0}^{N-1}$ (step lengths), pick $\mathbf{x}^0 \in \mathbb{R}^n$ and set $N = nK$.

For $k = 0, 1, 2, \ldots, N-1$ perform the following step:

1. Set $i = k \pmod{n} + 1$
2. $\mathbf{x}^{k+1} = \mathbf{x}^k - t_k[\nabla f(\mathbf{x}^k)]_i e_i$.

Gauss-Seidel is the special case where $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x} - b^\top \mathbf{x}$.

SDP performance estimation

SDP PEP formulated in:

Yassine Kamri, Julien M Hendrickx, and François Glineur (2022)

On the worst-case analysis of cyclic coordinate-wise algorithms on smooth convex functions. *arXiv:2211.17018*.

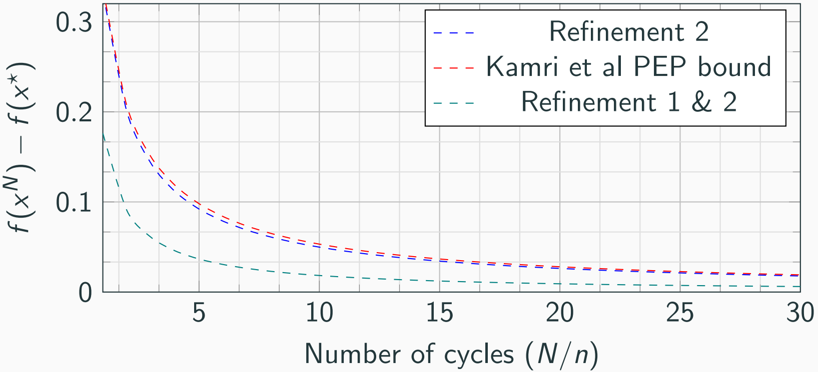
Two helpful refinements for convex quadratic f:

1. Add to PEP:

$$\frac{1}{2}\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle = f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

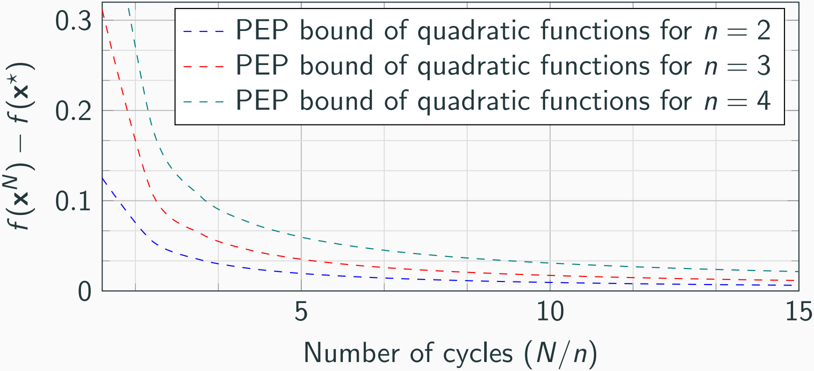
2. Use that $t \mapsto f(\mathbf{x}^k + te_i)$ is 1-smooth for $\langle \cdot, \cdot \rangle_D$.

Numerical solutions of PEPs



for $n = 2, L = 2, \ell_1 = 1, \ell_2 = 1, t = 0.5$, when $\|\mathbf{x}^0 - \mathbf{x}^\star\|_D \leq 1$.

Worst-case PEP bound for Gauss-Seidel method



when $\|x^0 - x^\star\|_D \leq 1$.

Concluding remarks

- PEP results for Gauss-Seidel from:

H. Abbaszadehpeivasti, E. de Klerk, and M. Zamani (2022)

Convergence rate analysis of randomized and cyclic coordinate descent for convex optimization through semidefinite programming. *Applied Set-Valued Analysis and Optimization*, to appear. Preprint at arXiv:2212.12384

- Only numerical results so far — no closed for expressions for PEP solutions.
- PEP for (nonlinear) Jacobi is work in progress.

The End