Exact Optimal Accelerated Complexity for Fixed-Point Iterations

Jisun Park[1] and **Ernest K. Ryu**[1]

Long Presentation
International Conference on Machine Learning, 2022

[1]Department of Mathematical Sciences, Seoul National University

Fixed-point iteration

Fixed-point iteration with $\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^n$ computes

$$x_{k+1} = \mathbb{T}x_k$$

with some starting point $x_0 \in \mathbb{R}^n$.

Rubric: Formulate solution as fixed point of an operator and perform the fixed-point iteration.

Ubiquitous throughout applied mathematics, science, engineering, and machine learning. However, the computational complexity of the abstract fixed-point iteration has not been studied extensively.

Question) What is the optimal (accelerated) iteration complexity of fixed-point iterations?

Fixed-point problem ⇔ Monotone inclusion problem

Our analysis relies on the following equivalence.

Fixed-point problem

$$\underset{y \in \mathbb{R}^n}{\text{find}} \quad y = \mathbb{T}y$$

with $1/\gamma$-Lipschitz $\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^n$ (with $\gamma \geq 1$) is equivalent to monotone inclusion problem

$$\underset{x \in \mathbb{R}^n}{\text{find}} \quad 0 \in \mathbb{A}x$$

with maximal μ-strongly monotone $\mathbb{A}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ (with $\mu \geq 0$).

Lemma

With $\gamma = \frac{1}{1+2\mu}$, there is a one-to-one correspondence

$$\mathbb{A} = \left(\mathbb{T} + \frac{1}{\gamma}\mathbb{I}\right)^{-1}\left(1 + \frac{1}{\gamma}\right) - \mathbb{I} \quad \Leftrightarrow \quad \mathbb{T} = \left(1 + \frac{1}{1+2\mu}\right)\mathbb{J}_{\mathbb{A}} - \frac{1}{1+2\mu}\mathbb{I}$$

and $x_\star$ is a zero of $\mathbb{A}$ if and only if it is a fixed point of $\mathbb{T}$.

Outline

Exact optimal methods

Optimal Contractive Halpern (OC-Halpern):

$$y_k = \left(1 - \frac{1}{\varphi_k}\right)\mathbb{T}y_{k-1} + \frac{1}{\varphi_k}y_0 \qquad \text{(OC-Halpern)}$$

where $\mathbb{T}$ is $1/\gamma$-contractive, $\varphi_k = \sum_{i=0}^{k}\gamma^{2i}$, and y_0 is a starting point.

Optimal Strongly-monotone Proximal Point Method (OS-PPM):

$$\begin{aligned} x_k &= \mathbb{J}_{\mathbb{A}}y_{k-1} \qquad \text{(OS-PPM)} \\ y_k &= x_k + \frac{\varphi_{k-1}-1}{\varphi_k}(x_k - x_{k-1}) - \frac{2\mu\varphi_{k-1}}{\varphi_k}(y_{k-1} - x_k) + \frac{(1+2\mu)\varphi_{k-2}}{\varphi_k}(y_{k-2} - x_{k-1}) \end{aligned}$$

where $\mathbb{A}$ is maximal μ-strongly monotone, $\varphi_k = \sum_{i=0}^{k}(1+2\mu)^{2i}$, $\varphi_{-1} = 0$, and $x_0 = y_0 = y_{-1}$ is a starting point.

Exact optimal methods

These two methods are equivalent:

$$y_k = \left(1 - \frac{1}{\varphi_k}\right) \mathbb{T} y_{k-1} + \frac{1}{\varphi_k} y_0 \qquad \text{(OC-Halpern)}$$

$$x_k = \mathbb{J}_{\mathbb{A}} y_{k-1} \qquad \text{(OS-PPM)}$$

$$y_k = x_k + \frac{\varphi_{k-1} - 1}{\varphi_k}(x_k - x_{k-1}) - \frac{2\mu\varphi_{k-1}}{\varphi_k}(y_{k-1} - x_k) + \frac{(1+2\mu)\varphi_{k-2}}{\varphi_k}(y_{k-2} - x_{k-1})$$

Lemma

The y_k-iterates of (OC-Halpern) *and* (OS-PPM) *are identical provided they start from the same initial point y_0.*

Accelerated rate (exact optimal)

Theorem

Let $\mathbb{A}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ *be maximal* μ*-strongly monotone with* $\mu \geq 0$*.*
(OS-PPM) *exhibits the rate*

$$\|\tilde{\mathbb{A}} x_N\|^2 \leq \left(\frac{1}{\sum_{k=0}^{N-1} (1+2\mu)^k} \right)^2 \|y_0 - x_\star\|^2,$$

where $\tilde{\mathbb{A}} x_N = x_{N-1} - x_N$*.*

This is the fastest rate. When $\mu = 0$, the rate

$$\|\tilde{\mathbb{A}} x_N\|^2 \leq \mathcal{O}(1/N^2)$$

is faster than the $\mathcal{O}(1/N)$ rate for (unaccelerated) PPM.

$\mathcal{O}(1/N^2)$ rate due to (Kim 2021). Rate for $\mu > 0$ is new.
Kim, Accelerated proximal point method for maximally monotone operators, *MPA*, 2021.

Accelerated rate (exact optimal)

Corollary

Let $\mathbb{T}\colon \mathbb{R} \to \mathbb{R}$ *be* γ^{-1}*-contractive with* $\gamma \geq 1$*.* (OC-Halpern) *exhibits the rate*

$$\|y_N - \mathbb{T}y_N\|^2 \leq \left(1 + \frac{1}{\gamma}\right)^2 \left(\frac{1}{\sum_{k=0}^{N} \gamma^k}\right)^2 \|y_0 - y_\star\|^2.$$

This is the fastest rate. When $\gamma = 1$, the rate

$$\|y_N - \mathbb{T}y_N\|^2 \leq \mathcal{O}(1/N^2)$$

is faster than the $\mathcal{O}(1/N)$ rate for plain (KM) fixed-point iteration.

$\mathcal{O}(1/N^2)$ rate due to (Lieder 2021). Rate for $\gamma < 1$ is new.

Lieder, On the convergence rate of the Halpern-iteration. *OPTL*, 2021.

Outline

Exact optimality

Theorem

For $n \geq N+1$, there exists an $1/\gamma$-Lipschitz operator $\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^n$ with a fixed point $y_\star \in \operatorname{Fix} \mathbb{T}$ such that

$$\|y_N - \mathbb{T}y_N\|^2 \geq \left(1 + \frac{1}{\gamma}\right)^2 \left(\frac{1}{\sum_{k=0}^{N} \gamma^k}\right)^2 \|y_0 - y_\star\|^2$$

for any iterates $\{y_k\}_{k=0}^N$ satisfying

$$y_k \in y_0 + \operatorname{span}\{y_0 - \mathbb{T}y_0, y_1 - \mathbb{T}y_1, \dots, y_{k-1} - \mathbb{T}y_{k-1}\}$$

for $k = 1, \dots, N$.

Lower bound matches upper bound *exactly*.

$\Theta(1/N^2)$ lower bound for $\gamma = 1$ due to (Diakonikolas 2020), but our bound improves the constant by a factor of about 80. Lower bound for $\gamma < 1$ is new.
Diakonikolas, Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities, *COLT*, 2020.

Construction of worst-case operator

Lemma

$\mathbb{T}$ is $\frac{1}{\gamma}$-contractive if and only if $\mathbb{G} = \frac{\gamma}{1+\gamma}(\mathbb{I} - \mathbb{T})$ is $\frac{1}{1+\gamma}$-averaged.

Lemma

Let $R > 0$. Define $\mathbb{N}, \mathbb{G} \colon \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$ as

$$\mathbb{N}(x_1, x_2, \dots, x_N, x_{N+1}) = (x_{N+1}, -x_1, -x_2, \dots, -x_N) - \frac{1+\gamma^{N+1}}{\sqrt{1+\gamma^2+\cdots+\gamma^{2N}}} Re_1$$

and $\mathbb{G} = \frac{1}{1+\gamma}\mathbb{N} + \frac{\gamma}{1+\gamma}\mathbb{I}$. That is,

$$\mathbb{G}x = \frac{1}{1+\gamma}\begin{bmatrix} \gamma & 0 & \cdots & 0 & 1 \\ -1 & \gamma & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \gamma & 0 \\ 0 & 0 & \cdots & -1 & \gamma \end{bmatrix} x - \frac{1}{1+\gamma}\frac{1+\gamma^{N+1}}{\sqrt{1+\gamma^2+\cdots+\gamma^{2N}}} Re_1.$$

Then $\mathbb{N}$ is nonexpansive, and $\mathbb{G}$ is $\frac{1}{1+\gamma}$-averaged.

Outline

Exact optimal methods (upper bound)

Complexity lower bound

Acceleration under Hölder-type growth condition

Experiments

Monotone $\supset$ Uniform mon. $\supset$ Strong mon.

(OS-PPM) provides an acceleration under monotonicity or strong monotonicity. In practice, these assumptions are often too weak or too strong, respectively. Uniform monotonicity is a practical middle ground.

$\mathbb{A}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is *uniformly monotone with parameters* $\mu > 0$ *and* $\alpha > 1$ if it is monotone and

$$\langle \mathbb{A}x, x - x_\star \rangle \geq \mu \|x - x_\star\|^{\alpha+1}$$

for any $x \in \mathbb{R}^n$ and $x_\star \in \operatorname{Zer} \mathbb{A}$. ($\alpha = \infty$ corresponds monotonicity and $\alpha = 1$ to strong monotonicity.)

We also refer to this as a Hölder-type growth condition, as it resembles the Hölderian error bound condition with function-value suboptimality replaced by $\langle \mathbb{A}x, x - x_\star \rangle$.

PPM under uniform monotonicity

Under uniform monotonicity, we first establish the rate the unaccelerated proximal point method (PPM)

$$x_{k+1} = \mathbb{J}_{\mathbb{A}} x_k.$$

Theorem

Let $\mathbb{A} \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ *be uniformly monotone with parameters* $\mu > 0$ *and* $\alpha > 1$. *Then*

$$\|\tilde{\mathbb{A}} x_N\|^2 \leq \mathcal{O}\left(\frac{1}{N^{\frac{\alpha+1}{\alpha-1}}}\right)$$

where $\tilde{\mathbb{A}} x_N = x_{N-1} - x_N$.

Restarted OS-PPM

We accelerate the rate using (OS-PPM) and restarting[†]. *Restarted OS-PPM*:

$$\begin{aligned} \tilde{x}_0 &= \mathbb{J}_{\mathbb{A}} x_0 \\ \tilde{x}_k &\leftarrow \textbf{OS-PPM}_0(\tilde{x}_{k-1}, t_k), \qquad k = 1, \dots, R, \end{aligned} \tag{OS-PPM$_0^{\text{res}}$}$$

where $\textbf{OS-PPM}_0(\tilde{x}_{k-1}, t_k)$ is the execution of t_k iterations of (OS-PPM) with $\mu = 0$ starting from $\tilde{x}_{k-1}$.

Theorem

Let $\mathbb{A} \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be uniformly monotone with parameters $\mu > 0$ and $\alpha > 1$. There is a restarting schedule $t_1, \dots, t_R$ such that

$$\|\tilde{\mathbb{A}} x_N\|^2 \leq \mathcal{O}\left(\frac{1}{N^{\frac{2\alpha}{\alpha-1}}}\right) = \mathcal{O}\left(\frac{1}{N^{\frac{\alpha+1}{\alpha-1}+1}}\right).$$

[†]Nesterov, Gradient methods for minimizing composite functions, *MPA*, 2013
[†]Roulet and d'Aspremont, Sharpness, restart, and acceleration, *SIOpt*, 2020.

Outline

Exact optimal methods (upper bound)

Complexity lower bound

Acceleration under Hölder-type growth condition

Experiments

Illustrative 2D toy examples

Toy examplex provide insight into acceleration mechanism.

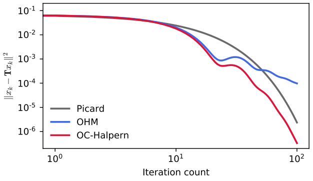
$\frac{1}{\gamma}$-contractive $\mathbb{T}_\theta \colon \mathbb{R}^2 \to \mathbb{R}^2$ and a maximal μ-strongly monotone $\mathbb{M} \colon \mathbb{R}^2 \to \mathbb{R}^2$

$$\mathbb{T}_\theta \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{\gamma} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

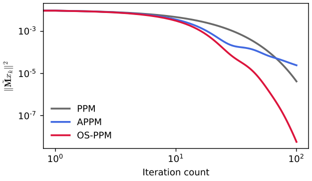
$$\mathbb{M} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \left(\frac{1}{N-1} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} + \begin{bmatrix} \mu & 0 \\ 0 & \mu \end{bmatrix} \right) \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

with $\gamma = 1/0.95 = 1.0526$, $\mu = 0.035$, and $\theta = 15^\circ$.

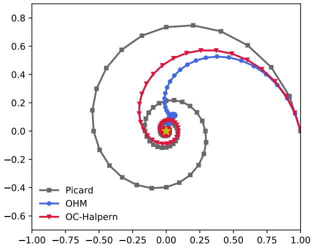
Illustrative 2D toy examples



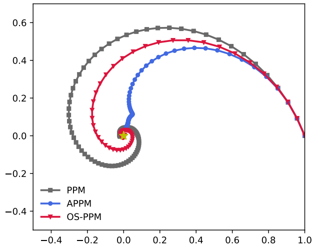
(a) Fixed-point residual of $\mathbb{T}_\theta$



(b) Resolvent residual norm of $\mathbb{M}$



(c) Trajectory of $\mathbb{T}_\theta$



(d) Trajectory of $\mathbb{M}$

Real-world problems

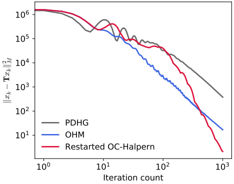
Experiment on several real world problems.

(a) $\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\|Ex-b\|^2+\lambda\|Dx\|_1,$

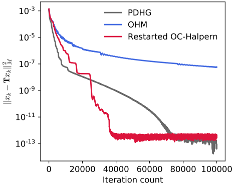
(b) $\underset{m_x,m_y}{\text{minimize}} \quad \|\mathbf{m}\|_{1,1}=\sum_{i=1}^{n}\sum_{j=1}^{n}|m_{x,ij}|+|m_{y,ij}|$

$\text{subject to} \quad \text{div}(\mathbf{m})+\rho_1-\rho_0=0,$

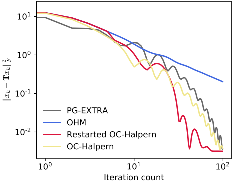
(c) $\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^{n}\|A_{(i)}x-b_{(i)}\|^2+\lambda\|x\|_1.$



(a) CT imaging



(b) Earth mover's distance



(c) Decentralized compressed sensing

In all three applications, restarting provides an acceleration.

Conclusion

(i) Classical fixed-point iteration is suboptimal.
(ii) Appropriate use of anchoring yields acceleration and is exactly optimial.
(iii) With restarting, we demonstrate a practical benefit in a wide range of setups.