Note on the paper “Optimal-order convergence of Nesterov acceleration for linear ill-posed problems”

William Weijia Zheng
IE dept, CUHK

Mov 28, 2024

Content

- Preliminaries
  - *Some basic stuff, notations, backgrounds*
- Residual polynomials
  - *A useful tool to analyze performance of different methods*
- Convergence analysis and discussions
  - *Nesterov acceleration is a powerful method, in multiple aspects*
- Simulation results

Preliminaries

- We want to tackle an ill-posed linear problem
- $$y^\delta = Ax,$$
- where $A : X \to Y$ with X, Y being Hilbert spaces, and y^δ denotes a noisy observation of the ground-truth $y^\dagger$ with $\|y^\dagger - y^\delta\| = \delta$.
- Safe to assume $\|A\| \leq 1$, since we can always scale the original problem.
- ***Landweber method***, (which may be slow, as we will see later):
- $$x_{k+1}^\delta \leftarrow x_k^\delta - A^*(Ax_k^\delta - y^\delta)\,.$$
- ***Nesterov acceleration method:***
- $$x_{k+1}^\delta \leftarrow z_k^\delta + A^*(y^\delta - Az_k^\delta), \quad z_k^\delta \leftarrow x_k^\delta + \alpha_k(x_k^\delta - x_{k-1}^\delta), \quad x_0^\delta \leftarrow 0, \quad x_1^\delta \leftarrow A^*y^\delta,$$
- where $\alpha_k = \dfrac{k-1}{k+\beta}$, with $\beta > -1$.

$$x_{k+1}^\delta \leftarrow z_k^\delta + A^*(y^\delta - Az_k^\delta),$$
$$z_k^\delta \leftarrow x_k^\delta + \alpha_k(x_k^\delta - x_{k-1}^\delta), \quad x_0^\delta \leftarrow 0, \quad x_1^\delta \leftarrow A^*y^\delta,$$

Residual polynomials

- What is a "residual"? It is just the difference between the observation and our "reconstruction" based on the estimation $\hat{x}$ (of x):
- $$y^\delta - A\hat{x}$$ ⬅️ ***we call this "residual"***
- According to Nesterov method, for $k = 0{,}1$ we have:
- $$y^\delta - Ax_0^\delta = y^\delta, \quad y^\delta - Ax_1^\delta = (1 - AA^*)y^\delta, \quad y^\delta - Ax_2^\delta = \text{recusion of LHS terms}$$
- ***Previous works have noticed that***: we can express the (general) $y^\delta - Ax_k^\delta$ quite light-weighted via a auxiliary polynomial, denoted by $r_k(\cdot)$. In detail:
- $$\begin{aligned} r_k(\lambda) &= (1-\lambda)[r_k(\lambda) + \alpha_k(r_k(\lambda) - r_{k-1}(\lambda))],\ k \geq 1 \\ r_0(\lambda) &= 1, \quad r_1(\lambda) = 1 - \lambda \end{aligned}.$$
- With this, one can prove: $y^\delta - Ax_k^\delta = r_k(AA^*)y^\delta$. *[Use $k = 0{,}1$ to convince]*
- Also, define $g_k(\lambda) \triangleq \dfrac{1 - r_k(\lambda)}{\lambda}$, we will have $x_k^\delta = g_k(A^*A)A^*y^\delta$.

Residual polynomials

$$x_{k+1}^\delta \leftarrow z_k^\delta + A^*(y^\delta - Az_k^\delta),$$
$$z_k^\delta \leftarrow x_k^\delta + \alpha_k(x_k^\delta - x_{k-1}^\delta), \quad x_0^\delta \leftarrow 0, \quad x_1^\delta \leftarrow A^*y^\delta,$$

- The recursion on last slide is discovered in literature. The author managed to solve its "general expression" (for Nesterov method:)
- $$r_k(\lambda) = (1-\lambda)^{\frac{k+1}{2}} \frac{C_{k-1}^{(\frac{\beta+1}{2})}(\sqrt{1-\lambda})}{C_{k-1}^{(\frac{\beta+1}{2})}(1)}, \quad k \geq 1$$

Recall: β appears in $\alpha_k = \dfrac{k-1}{k+\beta}$

- where $C_n^{(\alpha)}$ denotes the Gegenbauer polynomials. The good news is we can forget about the recursion.
- Other methods also admit a residual polynomial:
- ***Landweber***:
- $$r_k^{(LW)}(\lambda) = (1-\lambda)^k.$$
- **ν-method:**
- $$r_k^{(\nu)}(\lambda) = \frac{C_{2k}^{(2\nu)}(\sqrt{1-\lambda})}{C_{2k}^{2\nu}(1)}.$$

Note that there is some relation between the three method, in terms of their residual polynomials!

The Nesterov is a "product" of the bottom two methods.

Convergence analysis: source condition & optimal to hope

- As a standard manner in the literature, the author imposes the following smoothness source condition:
- $$x^\dagger = (A^*A)^\mu \omega, \quad \mu > 0, \quad \|\omega\| < \infty\,.$$
- And it is proved in (a classical textbook) *"**Regularized inverse problem (1996)**"* that under the above source condition, the optimal convergence rate ***(the best we can hople for)*** is of the form:
- $$\|x^\delta_{k(\delta)} - x^\dagger\| \in \mathcal{O}\left(\delta^{\frac{2\mu}{2\mu+1}}\right),$$
- and a scheme achieves such bound —— aka, the "power" of δ should attain $\dfrac{2\mu}{2\mu+1}$ —— is called of optimal order.

Convergence analysis: saturation

- ***Saturation***:

$$\|x_{k(\delta)}^{\delta} - x^{\dagger}\| \in \mathcal{O}\left(\delta^{\frac{2\mu}{2\mu+1}}\right),$$

  - the phenomenon that for certain regularization method, the convergence rate does not improve even when the smoothness $\mu > 0$ is larger. Saturation happens for the ν-method at $\mu = \nu$. And as a counter-example, Landweber iteration does not show saturation.

- ***Semi-saturation:***
  - the convergence rate improves anyway but in a suboptimal way, namely, the "power" is less than the optimal $2\mu/(2\mu + 1)$.

- To be precise, suppose some method that gives ⬇️, then there is "semi-saturation" effect for $\mu > \dfrac{1}{2}$.

$$\|x_{k(\delta)}^{\delta} - x^{\dagger}\| = \begin{cases} \mathcal{O}\left(\delta^{\frac{2\mu}{2\mu+1}}\right) & \mu \leq \frac{1}{2} \\ \mathcal{O}\left(\delta^{\frac{2\mu+1}{2\mu+3}}\right) & \mu > \frac{1}{2} \end{cases}$$

The LHS is not some random cook-up example, but a published result on Nesterov method with a prior stopping rule in 2017.

A main contribution of the focused paper is to improve the LHS result.

Convergence analysis: some preparation 1

- The improved convergence analysis is obtained via studing the residual polynomial(s) of Nesterov method. To achieve that, some estimates involving the Gegenbauer polynomials are needed:

Lemma 2 (Eq. (13) of [1]) *Let* $\lambda \in [0,1]$ *and* $\beta > -1$, *then* $\left| \frac{C_{k-1}^{(\frac{\beta+1}{2})}(\sqrt{1-\lambda})}{C_{k-1}^{(\frac{\beta+1}{2})}(1)} \right| \leq 1.$

Recall: $$r_k(\lambda) = (1-\lambda)^{\frac{k+1}{2}} \frac{C_{k-1}^{(\frac{\beta+1}{2})}(\sqrt{1-\lambda})}{C_{k-1}^{(\frac{\beta+1}{2})}(1)}, \quad k \geq 1$$

- The above lemma immediately deduces that $\forall \lambda \in [0,1), \forall \beta > -1, \ |r_k(\lambda)| \leq 1$. Furthermore, since
- $$r_k(\lambda) = (1-\lambda)^{\frac{k+1}{2}} \times (\text{something w/ abs. value bounded by } 1),$$
- we can deduce $\lim_{k\to\infty} r_k(\lambda) \to 0, \ \ \forall \lambda \in (0,1)$.

Convergence analysis: converge to noiseless version

- Let x_k^δ be defined by Nesterov method, with $\beta > -1$, then $\|x_k^\delta - x_k\| \le \sqrt{2}k\delta$.
- We use x_k to denote the noiseless version of x_k^δ, or $x_k = g_k(A^*A)A^*y^\dagger$. Recall: $x_k^\delta = g_k(A^*A)A^*y^\delta$
- Note that $g_k(\lambda) \underbrace{=}_{\text{by definition}} \dfrac{1-r_k(\lambda)}{\lambda} = \dfrac{r_k(0)-r_k(\lambda)}{\lambda}$. Hence $|g_k(\lambda)| = |r_k'(\tilde{\lambda})|$ by mean value theorem. Via brute force, one can derive:
- $$r_k'(\lambda) = \boxed{\frac{k+1}{2}(1-\lambda)^{\frac{k-1}{2}}} \underbrace{\frac{C_{k-1}^{(\frac{\beta+1}{2})}(\sqrt{1-\lambda})}{C_{k-1}^{(\frac{\beta+1}{2})}(1)}}_{\le 1 \text{ (using last slide)}} - \frac{1}{2}\boxed{(1-\lambda)^{\frac{k}{2}}} \underbrace{\frac{\frac{\partial}{\partial \lambda}[C_{k-1}^{(\frac{\beta+1}{2})}(\sqrt{1-\lambda})]}{C_{k-1}^{(\frac{\beta+1}{2})}(1)}}_{\le 2(k-1)^2\text{, a little bit tedious.}} .$$
- With the above, we see
- $$|r_k'(\lambda)| \le \boxed{\frac{k+1}{2}(1-\lambda)^{\frac{k-1}{2}}} + \boxed{(1-\lambda)^{\frac{k}{2}}}(k-1)^2 \le \max_{\lambda\in[0,1]} \frac{k+1}{2}(1-\lambda)^{\frac{k-1}{2}} + (1-\lambda)^{\frac{k}{2}}(k-1)^2 \overset{\text{choosing } \lambda \leftarrow 0}{=} \frac{k+1}{2} + (k-1)^2 .$$
- By the Theorem 4.1 and 4.2 in the 1996 textbook, we obtain (not very straightforwardly)
- $$\|x_k^\delta - x_k\|^2 \le \underbrace{\|Ax_k - Ax_k^\delta\|}_{\le C\cdot\delta\text{, with } C\le 2} \cdot \underbrace{\|g_k(AA^*)\|}_{\le\frac{k+1}{2}+(k-1)^2} \cdot \delta \le 2\delta^2\left(\frac{k+1}{2} + (k-1)^2\right).$$
- Hence finally, $\|x_k^\delta - x_k\| \le \sqrt{2}\delta\sqrt{\dfrac{k+1}{2} + (k-1)^2} \le \sqrt{2}k\delta$.

Screenshot I copied from the book ⬇️. BTW, this book is kinda hard to find on the Internet……

$$\begin{aligned} \|x_\alpha - x_\alpha^\delta\|^2 &= \langle x_\alpha - x_\alpha^\delta, T^* g_\alpha(TT^*)(y - y^\delta)\rangle \\ &= \langle Tx_\alpha - Tx_\alpha^\delta, g_\alpha(TT^*)(y-y^\delta)\rangle \\ &\le \|Tx_\alpha - Tx_\alpha^\delta\|\,\|g_\alpha(TT^*)\|\delta \end{aligned}$$

Proceed to final results

- Because we will eventually rely on the usual technique of expanding $\|x_k^\delta - x^\dagger\| \leq \|x_k^\delta - x_k\| + \|x_k - x^\dagger\|$, and then make each of them goes to zero fast enough.

- Previous slide made it clear that $\|x_k^\delta - x_k\| \to 0$ if $(\delta \to 0 \implies k(\delta)\delta \to 0)$. And that step is heavily based on our study of the residual polynomial $r_k(\,\cdot\,)$. *[Something has not been done in the literature]*

- The author proved —— for both ***a priori*** and ***discrepancy principle*** stopping rules (which are different ways to determine the number of iterations) —— that Nesterov can be optimal-ordered, provided that β is not chosen too small.

- As the implication of the two versions of the optimality theorems are somewhat similar, I will focus on the ***a priori*** version, since it comes first in the paper.

Optimality of Nesterov (a priori)

香港中文大學
The Chinese University of Hong Kong

Lemma 4 (Proposition 2 in [1]) *Let $\beta > -1$. Then there exists some constant c_β such that*

$$\left|\frac{r_k(\lambda)\lambda^{\frac{\beta+1}{4}}}{(1-\lambda)^{\frac{k+1}{2}}}\right| \le c_\beta k^{-2\frac{\beta+1}{4}} \quad \left(\iff r_k(\lambda)\lambda^{\frac{\beta+1}{4}} \le (1-\lambda)^{\frac{k+1}{2}} c_\beta k^{-2\frac{\beta+1}{4}}\right)$$

This is a great improvement comparing to the 2017 result.

Theorem 5 (Theorem 4 in [1]) *Let $\|A^*A\| \le 1$ and $\beta > -1$, and the smoothness source condition holds for some $\mu > 0$. Then,*

- *if $\mu \le \frac{\beta+1}{4}$, choose $k(\delta) \in \mathcal{O}(\delta^{-\frac{1}{2\mu+1}})$, then the optimal order convergence is achieved, namely:*

$$\|x^\delta_{k(\delta)} - x^\dagger\| \in \mathcal{O}(\delta^{\frac{2\mu}{2\mu+1}}).$$

It seems: one should choose β large (but not too large.)

- *if $\mu > \frac{\beta+1}{4}$, choose $k(\delta) \in \mathcal{O}(\delta^{-\frac{1}{\mu+\frac{\beta+1}{4}+1}})$, then a suboptimal order convergence is obtained:*

$$\|x^\delta_{k(\delta)} - x^\dagger\| \in \mathcal{O}(\delta^{\frac{\mu+\frac{\beta+1}{4}}{\mu+\frac{\beta+1}{4}+1}}).$$

Semi-saturation !!!

Recall that x_k is the exact version of iterative solution, with
$x_k = g_k(A^*A)A^*y^\dagger = g_k(A^*A)A^*Ax^\dagger$

Hence
$x^\dagger - x_k = x^\dagger - g_k(A^*A)A^*Ax^\dagger = r_k(A^*A)x^\dagger$

The last equation holds by source condition.

- **Proof sketch** *(optimal case only)*
- $\|x^\delta_{k(\delta)} - x^\dagger\| \le \underbrace{\|x^\delta_{k(\delta)} - x_{k(\delta)}\|}_{\le\sqrt{2}k(\delta)\delta,\ \text{2 slides ago}} + \boxed{\|x_{k(\delta)} - x^\dagger\|}$. Also, we have: $\boxed{x_k - x^\dagger = r_k(A^*A)x^\dagger = r_k(A^*A)(A^*A)^\mu\omega\,.}$
- Now consider the function $r_k(\lambda)\lambda^\mu$. By Lemma 4, we see that when $\mu \le \frac{\beta+1}{4}$, we do have $r_k(\lambda)\lambda^\mu \le (1-\lambda)^{\frac{k+1}{2}} c_\beta k^{-2\mu} \le C_1 k^{-2\mu}$.
Then the prove is done, by rewriting the constant with ω, and choosing the $k(\delta)$ as the premised one.

Just repeat…

Theorem 5 (Theorem 4 in [1]) *Let $\|A^*A\| \leq 1$ and $\beta > -1$, and the smoothness source condition holds for some $\mu > 0$. Then,*

- *if $\mu \leq \frac{\beta+1}{4}$, choose $k(\delta) \in \mathcal{O}(\delta^{-\frac{1}{2\mu+1}})$, then the optimal order convergence is achieved, namely:*

$$\|x^{\delta}_{k(\delta)} - x^{\dagger}\| \in \mathcal{O}(\delta^{\frac{2\mu}{2\mu+1}}).$$

- *if $\mu > \frac{\beta+1}{4}$, choose $k(\delta) \in \mathcal{O}(\delta^{-\frac{1}{\mu+\frac{\beta+1}{4}+1}})$, then a suboptimal order convergence is obtained:*

$$\|x^{\delta}_{k(\delta)} - x^{\dagger}\| \in \mathcal{O}(\delta^{\frac{\mu+\frac{\beta+1}{4}}{\mu+\frac{\beta+1}{4}+1}}).$$

- Maybe it seems silly, but I want to stress here that the $\mu > \dfrac{\beta+1}{4}$ case is deemed suboptimal, since the "power" term
- $$\frac{\mu + \frac{\beta+1}{4}}{\mu + \frac{\beta+1}{4} + 1} < \frac{2\mu}{2\mu+1}$$

Closer to 1

The residual polynomial of Nesterov is a "product" of two methods:

(1) Landweber method (no saturation) and
(2) ν-method (saturation),

Hence, as a combination of these two, Nesterov shows "semi-saturation".

Optimality of Nesterov (discrepancy principle)

Theorem 6 (Theorem 6 in [1]) *Let $\|A^*A\| \le 1$ and $\beta > -1$. The smoothness source condition holds for some $\mu > 0$. If the iteration is stopped by the discrepancy principle, then:*

- *if $\mu \le \frac{\beta-1}{4}$ ($\iff$ $\mu + \frac{1}{2} \le \frac{\beta+1}{4}$), with posterior $k(\delta) \in \mathcal{O}(\delta^{-\frac{1}{2\mu+1}})$, then the optimal order convergence is achieved,*

$$\|x^\delta_{k(\delta)} - x^\dagger\| \in \mathcal{O}(\delta^{\frac{2\mu}{2\mu+1}}).$$

- *if $\mu \ge \frac{\beta-1}{4}$, choose $k(\delta) \in \mathcal{O}(\delta^{-\frac{1}{\mu+\frac{\beta+3}{4}}})$, then a suboptimal order convergence is obtained:*

$$\|x^\delta_{k(\delta)} - x^\dagger\| \in \mathcal{O}(\delta^{\frac{\mu+\frac{\beta+1}{4}-\frac{1}{2}}{\mu+\frac{\beta+1}{4}+\frac{1}{2}}}).$$

- Again, the "threshold" $\dfrac{\beta+1}{4}$ appears.
- Recall that: as we said, saturation happens for the ν-method at $\mu = \nu$.

Nesterov: $r_k(\lambda) = (1-\lambda)^{\frac{k+1}{2}} \dfrac{C_{k-1}^{(\frac{\beta+1}{2})}(\sqrt{1-\lambda})}{C_{k-1}^{(\frac{\beta+1}{2})}(1)}, \quad k \ge 1$

$r_k^{(LW)}(\lambda) = (1-\lambda)^k. \quad r_k^{(\nu)}(\lambda) = \dfrac{C_{2k}^{(2\nu)}(\sqrt{1-\lambda})}{C_{2k}^{2\nu}(1)}.$

Choose $2\nu = \dfrac{\beta+1}{2}$

Simulation results

- $k_{opt} = \arg\min_k \|x_k^\delta - x^\dagger\|$

Table 1. Errors compared to Nesterov iteration: $\frac{\|x^\delta_{\text{method},k} - x^\dagger\|}{\|x^\delta_{\text{Nesterov},k_{\text{opt}}} - x^\dagger\|}$.

Method	Stopping	$\delta = 10^{-5}$	10^{-4}	10^{-3}	10^{-2}	10^{-1}
Nesterov	k_{opt}	1	1	1	1	1
Landweber	k_{opt}	1.15	0.83	0.96	1.05	1.06
ν-method	k_{opt}	1.02	1.06	1.01	1.26	0.97
CGNE	k_{opt}	1.02	0.82	1.05	1.02	0.84
Nesterov	Discrepancy	1.58	1.10	1.41	2.84	1.90
Landweber	Discrepancy	2.23	1.17	1.41	2.80	1.98
ν-method	Discrepancy	1.02	1.13	1.00	1.56	1.88
CGNE	Discrepancy	1.81	1.19	1.05	2.51	1.97

Table 2. Number of iterations for various methods; setting as in table 1.

Method	Stopping	$\delta = 10^{-5}$	10^{-4}	10^{-3}	10^{-2}	10^{-1}
Nesterov	k_{opt}	371	163	65	26	15
Landweber	k_{opt}	11 000	2193	512	145	36
ν-method	k_{opt}	190	82	33	22	9
CGNE	k_{opt}	10	6	4	3	2
Nesterov	Discrepancy	260	111	39	13	1
Landweber	Discrepancy	5106	1080	220	37	1
ν-method	Discrepancy	190	96	33	10	1
CGNE	Discrepancy	8	5	4	2	1

The smaller δ is, the more # of iterations is needed. As predicted by the theory.

ν-method update rule

$$x_{k+1}^{\delta} = x_k^{\delta} + \mu_{k+1}(x_k - x_{k-1}) + \omega_{k+1}A^*(y^{\delta} - Ax_k), \quad k > 1,$$

$$\mu_{k+1} = \frac{(k-1)(2k-2)(2k+2\nu-1)}{(k+2\nu-1)(2k+4\nu-1)(2k+2\nu-3)},$$

$$\omega_{k+1} = 4\frac{(2k+2\nu-1)(k+\nu-1)}{(k+2\nu-1)(2k+4\nu-1)},$$

$$x_0 = 0, \quad x_1 = \frac{4\nu+2}{4\nu+1}A^*y^{\delta}$$ (initialization)

$$r_k^{(\nu)}(\lambda) = \frac{C_{2k}^{(2\nu)}(\sqrt{1-\lambda})}{C_{2k}^{2\nu}(1)}.$$