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

William Weijia Zheng
Department of Information Engineering, CUHK
wjzheng@link.cuhk.edu.hk

Nov. 27, 2024

Forewords

This 4-page short report is for MATH6221: ”Topics in Numerical Analysis - 2024/25”. Thanks for your reading. Due to space limit, I refrain myself from plotting the figures on the simulation results (which should be self-explanatory enough in the original paper). I'd mainly focus on the narrative logic of the original paper, probably with some humble attempts of explanations.

1 Preliminaries

Consider a problem of calculating a (regularized) solution to 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 exact (or, ground-truth) $y^\dagger$. It is always safe to assume $\|A\| \leq 1$, since we can always scale the original problem.

To tackle the above problem, the famous Landweber method aims to minimize (not in the strict sense, since we want to avoid overfitting by early stopping)

$$J(x) \triangleq \frac{1}{2}\|y - Ax\|^2,$$

via the following procedure:

$$x_{k+1}^\delta \leftarrow x_k^\delta - \eta A^*(Ax_k^\delta - y^\delta) = x_k^\delta - A^*(Ax_k^\delta - y^\delta). \tag{1}$$

Where $\eta \leftarrow 1$ servers as a ”learning rate”, which can be absorbed and hence ignored. This iterative algorithm has long been recognized to be slow in a lot of cases. To speed up the convergence procedure, Nesterov acceleration schemes can be applied, which leads to:

$$\begin{gathered} x_{k+1}^\delta \leftarrow z_k^\delta + A^*(y^\delta - Az_k^\delta), \\ z_k^\delta \leftarrow x_k^\delta + \alpha_k(x_k^\delta - x_{k-1}^\delta), \;\; x_0^\delta \leftarrow 0, \;\; x_1^\delta \leftarrow A^*y^\delta, \end{gathered} \tag{2}$$

where $\alpha_k = \frac{k-1}{k+\beta}, \;\; k \geq 1, \;\; \beta > -1$. Note that if $\alpha_k \leftarrow 0$, or equivalently choosing $\beta \leftarrow \infty$, such iteration degenerate back to Landweber.

2 Residual polynomials

The basis of the reviewed paper was constructed on a concept called ”residual polynomial”, whose role is to give an easier-to-analyze description of the residual of 2. By the initialization of x_0^δ, x_1^δ in 2, we see

$y^\delta - Ax_0^\delta = y^\delta$, and $y^\delta - Ax_1^\delta = (1 - AA^*)y^\delta$. For the k-th iteration of 2, its residual polynomial is termed $r_k(\cdot)$, that satisfies $\forall k$:

$$\underbrace{y^\delta - Ax_k^\delta}_{\text{residual}} = r_k(AA^*)y^\delta. \tag{3}$$

By studying the residual for larger k's, the authors of [1] and [2] define a sequence of polynomials $\{r_k\}_{k\in\mathbb{N}}$ specified as:

$$\begin{gathered} r_0(\lambda) = 1,\ \ r_1(\lambda) = 1 - \lambda, \\ r_{k+1}(\lambda) = (1 - \lambda)[r_k(\lambda) + \alpha_k(r_k(\lambda) - r_{k-1}(\lambda))],\ k \geq 1. \end{gathered} \tag{4}$$

With the above recursion, 3 should be self-explanatory. Define $g_k(\lambda) \triangleq \frac{1-r_k(\lambda)}{\lambda}$, we will have $x_k^\delta = g_k(A^*A)A^*y^\delta$, deduced by 2.

2.1 Residual polynomials of Nesterov

Although complicated enough, the author of [1] managed to derive the residual polynomial (Theorem 1 and 2 in [1]) for 2. And it is:

$$r_k(\lambda) = (1 - \lambda)^{\frac{k+1}{2}} \frac{C_{k-1}^{(\frac{\beta+1}{2})}(\sqrt{1-\lambda})}{C_{k-1}^{(\frac{\beta+1}{2})}(1)},\ \ k \geq 1, \tag{5}$$

where $C_n^{(\alpha)}$ denotes the Gegenbauer polynomials. With this, one does not need to bother with the recursion 4, and this enables one to apply analysis to $r_k(\cdot)$ more easily, as we will see later.

2.2 Residual polynomials for other methods

We can derive residual polynomials for other iterative methods as well. For the Landweber method, one can derive

$$r_k^{(LW)}(\lambda) = (1 - \lambda)^k. \tag{6}$$

And, for another method called "ν-method", one can derive its residual as:

$$r_k^{(\nu)}(\lambda) = \frac{C_{2k}^{(2\nu)}(\sqrt{1-\lambda})}{C_{2k}^{2\nu}(1)}. \tag{7}$$

Remark 1 *Note that the residual of Nesterov r_k can be expressed as the product of that of $\frac{k}{2}$ Landweber iterations and that of $\frac{k}{2}$ iterations of a ν-method, with $\nu \leftarrow \frac{\beta+1}{4}$.*

3 Convergence analysis

As a standard manner in the literature, the author imposes the following smoothness source condition:

$$x^\dagger = (A^*A)^\mu\omega, \quad \mu > 0, \quad \|\omega\| < \infty. \tag{8}$$

And it is proved in [3] that under 8, the optimal convergence rate is of the form:

$$\|x_{k(\delta)}^\delta - x^\dagger\| \in \mathcal{O}\left(\delta^{\frac{2\mu}{2\mu+1}}\right), \tag{9}$$

and a scheme achieves such bound – aka, the "power" of δ should attain $\frac{2\mu}{2\mu+1}$ – is called of optimal order.

3.1 Saturation, and semi-saturation

The terminology "saturation" refers the effect that for certain regularization method, the convergence rate does not improve even when the smoothness $\mu > 0$ is larger. It is known that saturation happens for the ν-method at $\mu = \nu$. And as a counter-example, Landweber iteration does not show saturation. On the other hand, the term "semi-saturation" means that the convergence rate improves anyway but in a suboptimal way, namely, the "power" is less than the optimal $\frac{2\mu}{2\mu+1}$.

To be precise, suppose some method that gives

$$\|x^{\delta}_{k(\delta)} - x^{\dagger}\| = \begin{cases} \mathcal{O}\left(\delta^{\frac{2\mu}{2\mu+1}}\right) & \mu \leq \frac{1}{2} \\ \mathcal{O}\left(\delta^{\frac{2\mu+1}{2\mu+3}}\right) & \mu > \frac{1}{2} \end{cases} \tag{10}$$

Then there is "semi-saturation" effect for $\mu > \frac{1}{2}$. The 10 is not some cook-up example, but a published result on 2 under source condition 8 with a prior stopping rule in [4]. And a main contribution of [1] is to improve this result 10.

The author explained the semi-saturation phenomenon of Nesterov by noticing that the residual of Nesterov is a "combination" of Landweber (no saturation) and ν-method (saturation), as mentioned in Remark 1.

3.2 Improved convergence rate of Nesterov

As mentioned before, the improved convergence analysis is obtained via studing the residual polynomial(s) of Nesterov, namely 5. 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* $\left|\frac{C_{k-1}^{(\frac{\beta+1}{2})}(\sqrt{1-\lambda})}{C_{k-1}^{(\frac{\beta+1}{2})}(1)}\right| \leq 1$.

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

$$\lim_{k\to\infty} r_k(\lambda) \to 0, \ \ \forall \lambda \in (0,1).$$

Lemma 3 (Proposition 1 of [1]) *Let x^δ_k be defined by 2, with $\beta > -1$, then $\|x^\delta_k - x_k\| \leq \sqrt{2}k\delta$.*

Proof sketch.

One need to use Theorem 4.1 and 4.2 in [3] – with the constant $C \leftarrow 2$ in the textbook [3] – to get the $\sqrt{2}$ in the above lemma. Recall that by x_k we denote the noise-free iteration version solution(s).

Note that $g_k(\lambda) \underbrace{=}_{\text{by def.}} \frac{1-r_k(\lambda)}{\lambda} = \frac{r_k(0)-r_k(\lambda)}{\lambda}$. Hence $|g_k(\lambda)| = |r_k'(\tilde{\lambda})|$. Via brute force, one can derive

$$r_k'(\lambda) = \frac{k+1}{2}(1-\lambda)^{\frac{k-1}{2}} \underbrace{\frac{C_{k-1}^{(\frac{\beta+1}{2})}(\sqrt{1-\lambda})}{C_{k-1}^{(\frac{\beta+1}{2})}(1)}}_{\leq 1 \text{ (using Lemma 2)}} - \frac{1}{2}(1-\lambda)^{\frac{k}{2}} \underbrace{\frac{\frac{\partial}{\partial\lambda}[C_{k-1}^{(\frac{\beta+1}{2})}(\sqrt{1-\lambda})]}{C_{k-1}^{(\frac{\beta+1}{2})}(1)}}_{\leq 2(k-1)^2\text{, a little bit tedious.}} .$$

With the above, we see

$$|r_k'(\lambda)| \leq \frac{k+1}{2}(1-\lambda)^{\frac{k-1}{2}} + (1-\lambda)^{\frac{k}{2}}(k-1)^2 \leq \max_{\lambda\in[0,1]} \frac{k+1}{2}(1-\lambda)^{\frac{k-1}{2}} + (1-\lambda)^{\frac{k}{2}}(k-1)^2 = \frac{k+1}{2} + (k-1)^2.$$

By the Theorem 4.1 and 4.2 in [3], we obtain (not very straightforwardly)

$$\|x^\delta_k - x_k\|^2 \leq \underbrace{\|Ax_k - Ax^\delta_k\|}_{\leq C\cdot\delta\text{, with } C\leq 2} \cdot \underbrace{\|g_k(AA^*)\|}_{\leq \frac{k+1}{2}+(k-1)^2} \cdot\delta \leq 2\delta^2(\frac{k+1}{2} + (k-1)^2).$$

Hence finally, $\|x_k^\delta - x_k\| \le \sqrt{2}\delta\sqrt{\frac{k+1}{2} + (k-1)^2} \le \sqrt{2}k\delta$. ■

Because we will eventually rely on the usual technique of expanding $\|x_k^\delta - x^\dagger\| \le \|x_k^\delta - x_k\| + \|x_k - x^\dagger\|$, and then make each of them goes to zero fast enough. And the above lemma made it clear that $\|x_k^\delta - x_k\| \to 0$, if $k(\delta)\delta \to 0 \impliedby \delta \to 0$. And that step is heavily based on our study of the residual polynomial $r_k(\cdot)$.

The author of [1] proved – for both *a priori* and *discrepancy principle* stopping rules (which are different ways to determine k, 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 similar, I will focus on the *a priori* version, since it comes first in [1]. I emphasis here that the "threshold" $\frac{\beta+1}{4}$ come from the following Lemma 4, which is a literature bound.

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

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

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:*

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

- *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:*

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

Proof (the optimal one only).

As said, $\|x_{k(\delta)}^\delta - x^\dagger\| \le \underbrace{\|x_{k(\delta)}^\delta - x_{k(\delta)}\|}_{\le\sqrt{2}k(\delta)\delta,\text{ Lemma 3}} + \|x_{k(\delta)} - x^\dagger\|$. By equation (4.8) in [3], we see that $x_k - x^\dagger = r_k(A^*A)x^\dagger =_{(8)} r_k(A^*A)(A^*A)^\mu\omega$. 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. ■

For completeness, we also put the discrepancy principle theorem here, but the proof will be omitted.

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,*

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

- *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:*

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

4 Some final remarks

Note that it is not always a good choice to make β overly big, since this will make $\alpha_k \approx 0$, and our Nesterov acceleration has too tiny effect (degenerate to Landweber). So, one should choose his $\beta \sim 4\mu \pm 1$.

Also, the author did not prove the converse results for the (deemed) suboptimal cases, which may be left for future investigation.

References

[1] Stefan Kindermann, “Optimal-order convergence of nesterov acceleration for linear ill-posed problems,” *Inverse Problems*, vol. 37, no. 6, pp. 065002, May 2021.

[2] M. Hanke, “Accelerated landweber iterations for the solution of ill-posed equations,” *Numer. Math.*, 1991.

[3] Heinz W. Engl, Martin Hanke, and Andreas Neubauer, *Regularization of inverse problems*, Kluwer Academic Publishers, 1996.

[4] Andreas Neubauer, “On nesterov acceleration for landweber iteration of linear ill-posed problems,” *Journal of Inverse and Ill-posed Problems*, 2017.