

Online Learning to Precondition III

Concise analysis of a quasi-Newton type algorithm

Wenzhi Gao



gwz@stanford.edu ICME, Stanford University

Abstract

This is a short post that continues the previous discussion on “online learning to precondition”, and I’ll present a very concise quasi-Newton type method with a two-line proof. The algorithm follows exactly the OL2P intuition.

AI-disclosure. All my blogs marked by 🧠 emphasize insights behind technical results, and I write the whole blog draft by myself. Generative AI tools are only used for result verification/proofreading.

Again and again, to simplify the discussion, I’ll resort to the quadratic function convention

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \langle x, Ax \rangle - \langle b, x \rangle =: f(x)$$

where $\mu I \leq A \leq LI$, and it corresponds to solving a linear system $Ax = b$.

Sparse approximate inverse preconditioner. There are many ways to construct preconditioners for A , and one of them is the sparse approximate inverse preconditioner (SPAI) [BT99] that aims to find P such that $\|PA - I\|_F$ (or $\|AP - I\|_F$) is reduced. We will abuse the notation of distance and define

$$\text{dist}(P, A) := \|AP - I\|_F^2.$$

A SPAI preconditioner aims to solve $\min_P \|AP - I\|_F^2$. Now, suppose we are somewhat “lazy”, and we don’t want to solve the minimization problem. Say we just randomly pick some P and obtain

$$x^+ = x - P \nabla f(x).$$

Again, there are basically two cases, categorized by the contraction $r = \frac{\|\nabla f(x^+)\|}{\|\nabla f(x)\|} = \|(I - AP) \frac{\nabla f(x)}{\|\nabla f(x)\|}\|$.

Case 1. Either the contraction is good. i.e., $r = \frac{\|\nabla f(x^+)\|}{\|\nabla f(x)\|}$ is small. It implies our random choice of P is lucky.

Case 2. Or the contraction is bad. i.e., r is large. In this case, we know that P is not a good preconditioner, and we need to fix it. Suppose we want to update P with a low-rank modification:

$$P_+ = P + \eta u v^\top, \quad \|u\| = \|v\| = 1,$$

where u, v will be specified later. Given u, v , the parameter η is chosen to minimize $\|AP - I\|_F^2$:

$$\eta^* = \arg \min_{\eta} \|A(P + \eta u v^\top) - I\|_F^2 = \frac{\langle Au, (I - AP)v \rangle}{\|Au\|^2}$$

and by taking $u = \frac{(I - AP)v}{\|(I - AP)v\|}$, we have $\|AP_+ - I\|_F^2 \leq \|AP - I\|_F^2 - \frac{1}{\kappa} \|(I - AP)v\|^2$.

The question is how to choose $\|v\| = 1$ with large $\|(I - AP)v\|$. The answer is obvious given

$$r = \left\| (I - AP) \frac{\nabla f(x)}{\|\nabla f(x)\|} \right\|.$$

Taking $v = \frac{\nabla f(x)}{\|\nabla f(x)\|}$, we have $\|AP_+ - I\|_F^2 \leq \|AP - I\|_F^2 - \frac{1}{\kappa} r^2$. The algorithm is very simple.

Algorithm 1. Two-line quasi-Newton

input x^1, P_1

for $k = 1, \dots, K$

compute $v = \frac{\nabla f(x^k)}{\|\nabla f(x^k)\|}, u = \frac{(I - AP)v}{\|(I - AP)v\|}, \eta = \frac{\langle Au, (I - AP)v \rangle}{\|Au\|^2}$

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

$P_{k+1} = P_k + \eta u v^\top$

end for

Again, the proof follows from the “online learning to precondition” relation. Define the learning potential

$$\phi(x, P) := \log \|\nabla f(x)\| + \kappa \|AP - I\|_F^2.$$

We have the following reduction in the potential.

Lemma 1. *We have $\phi(x^{k+1}, P_{k+1}) \leq \phi(x^k, P_k) - \frac{1}{2} \log 2e$ for all $k \geq 1$.*

Proof. Since $\|AP_{k+1} - I\|_F^2 \leq \|AP_k - I\|_F^2 - \frac{1}{\kappa} r_k^2$, we have

$$\phi(x^{k+1}, P_{k+1}) - \phi(x^k, P_k) \leq \log r_k - r_k^2 \leq \max_r \{\log r - r^2\} = -\frac{1}{2} \log 2e. \quad \square$$

Complexity. The complexity is straightforward

$$\mathcal{O} \left(\underbrace{\kappa \|AP - I\|_F^2}_{\text{Learning}} + \log \left(\frac{1}{\varepsilon} \right) \right)$$

And superlinear convergence follows from $\sum_{k=1}^K r_k^2 \leq \kappa \|AP_1 - I\|_F^2$ and the quadratic-mean inequality

$$\frac{\|\nabla f(x^{K+1})\|}{\|\nabla f(x^1)\|} = \prod_{k=1}^K r_k \leq \left(\sqrt{\frac{1}{K} \sum_{k=1}^K r_k^2} \right)^K \leq \left(\frac{\sqrt{\kappa} \|AP_1 - I\|_F}{\sqrt{K}} \right)^K.$$

References

[BT99] Michele Benzi and Miroslav Tuma. A comparative study of sparse approximate inverse preconditioners. *Applied Numerical Mathematics*, 30(2-3):305–340, 1999.