

Online Learning to Precondition II.

Quasi-Newton methods

Wenzhi Gao



gwz@stanford.edu ICME, Stanford University

Abstract

This blog continues the previous discussion on “online learning to precondition”, and I’ll show that the renowned quasi-Newton (QN) family carries the same “online learning” insight, albeit with a more intricate analysis. In particular, the secant equation used in quasi-Newton methods can be interpreted as a loss function that learns the (inverse) Hessian at the optimum.

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.

1 Introduction and background

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 $\mu I \leq A \leq LI$. For unconstrained optimization of medium sizes, arguably the most widely used algorithm is quasi-Newton method [NW06, Fle00]. The first quasi-Newton method, now known as DFP update, was invented by William C. Davidon in 1959 [Dav91]. After years of development, many QN variants have been proposed, and the most effective variants, such as BFGS and SR1, stand the test of time and are deployed in modern numerical optimization packages.

As most textbooks teach it, locally QN mimics the behavior of Newton’s method by learning the (inverse) Hessian $\nabla^2 f(x^*)$. There is a central equation behind most QN methods, known as secant equation: given two points x and x^+ , for quadratic f , the following two equations hold

$$\nabla^2 f(x^*)(x^+ - x) = \nabla f(x^+) - \nabla f(x) \quad \text{and} \quad [\nabla^2 f(x^*)]^{-1}(\nabla f(x^+) - \nabla f(x)) = x^+ - x$$

Define the step vector $s = x^+ - x$ and the corresponding yield vector $y = \nabla f(x^+) - \nabla f(x)$, the equations can be equivalently written as

$$P^* y - s = B^* s - y = 0,$$

where $P^* = [\nabla^2 f(x^*)]^{-1}$ and $B^* = \nabla^2 f(x^*)$. Given an existing P (or B) and the secant equation P^* (or B^*) should satisfy, QN looks for $P_+ \in \mathcal{P}$ following the “least change” principle [Fle91]

$$\begin{aligned} P_+ &= \min_{Q \in \mathcal{P}} V_d(Q, P) \\ \text{subject to} \quad & Qy = s, \end{aligned}$$

where $V_d(Q, P)$ is a matrix Bregman divergence induced by a strictly convex kernel d . Different choices of d , learning object P^*, B^* , and the set give different variants of QN.

Learning object & secant equation	Kernel d	Matrix space	Quasi-Newton variant
$P^* = [\nabla^2 f(x^*)]^{-1}, \quad Py = s$	$\frac{1}{2}\ \cdot\ _F^2$	$\mathbb{S}^n$	Powell-symmetric-Broyden
		$\mathbb{R}^{n \times n}$	Broyden's bad update
	$-\log \det$	$\mathbb{S}_{++}^n$	DFP
$B^* = \nabla^2 f(x^*), \quad Bs = y$	$\frac{1}{2}\ \cdot\ _F^2$	$\mathbb{S}^n$	Powell-symmetric-Broyden
		$\mathbb{R}^{n \times n}$	Broyden's good update
	$-\log \det$	$\mathbb{S}_{++}^n$	BFGS

Table 1. Instantiations of different quasi-Newton variants.

1.1 Quasi-Newton and OL2P

To see how QN fits into the OL2P idea. I'll use Broyden's update

$$B_+ = B - \frac{(Bs - y)s^\top}{\|s\|^2}$$

as a demo and introduce a useful tool from the literature.

Proximal point and Moreau envelope descent. The QN subproblem can be written more compactly as proximal minimization with an indicator $\mathbb{I}\{Qs = y\}$:

$$B_+ = \arg \min_Q \{ \mathbb{I}\{Qs = y\} + \frac{\rho}{2} \|Q - B\|_F^2 \}, \quad (1)$$

where $\rho > 0$ is the proximal parameter. Furthermore, a proximal point update as in (1) can be further written as gradient descent on the Moreau envelope function [PB14]

$$\ell_{y,s}^\rho(B) = \min_Q \{ \mathbb{I}\{Qs = y\} + \frac{\rho}{2} \|Q - B\|_F^2 \},$$

where $\ell_{y,s}^\rho(B)$ satisfies the following properties

Proposition 1. *The function $\ell_{y,s}^\rho(B) = \frac{\rho}{2\|s\|^2} \|Bs - y\|^2$ with $s \neq 0$ has the following properties:*

1. it is ρ -smooth and convex;
2. its gradient is $\nabla \ell_{y,s}^\rho(B) = \rho(B - B_+) = \rho \frac{(Bs - y)s^\top}{\|s\|^2}$;
3. it satisfies $\ell_{y,s}^\rho(B^*) = 0$.

Therefore, we can write

$$B_+ = B - \frac{1}{\rho} \nabla \ell^\rho(B).$$

In other words, QN can be viewed as online gradient descent on the smoothed loss $\ell_{y,s}^\rho$.

Now we can formally write Broyden's update in this format. Note that we do not consider any globalization procedure for brevity. Also we will resort to the shorthand notation $\ell_k^\rho = \ell_{y^k, s^k}^\rho$.

Algorithm 1. Broyden's (good) update with null step

input $x^1, B_1, \eta = \frac{1}{\rho} > 0$.

for $k = 1, \dots, K$

$$x^{k+1/2} = x^k - B_k^{-1} \nabla f(x^k)$$

choose x^{k+1} to be the better among $\{x^k, x^{k+1/2}\}$ in terms of $\|\nabla f(x)\|$

compute $s^k = x^{k+1/2} - x^k$ and $y^k = \nabla f(x^{k+1/2}) - \nabla f(x^k)$

$$B_{k+1} = B_k - \eta \nabla \ell_k^\rho(B_k) = \arg \min_Q \{ \mathbb{I}\{Qs^k = y^k\} + \frac{\rho}{2} \|Q - B_k\|_F^2 \}$$

end for

Note that the parameter choice $\eta = \rho^{-1}$ recovers the proximal point update, yet theoretically the analysis would hold for fixed value of $\eta \leq \rho^{-1}$. I'll also assume that B_k^{-1} is well-defined at each iteration.

To establish the convergence behavior, let's recall the “either-or” discussion from Blog #3: at each iteration,

- we either achieve good progress in convergence (e.g., good contraction $r^k = \frac{f(x^{k+1}) - f(x^*)}{f(x^k) - f(x^*)}$)
- or we reduce the distance between B and B^* (e.g., $\|B - B^*\|_F^2$).

The “or” relation is easy for QN. Since B^* minimizes $\{\ell_k^\rho\}_{k=1}^K$, we have reduction in the Euclidean distance:

$$\begin{aligned} \|B_{k+1} - B^*\|_F^2 &= \|B_{k+1} - B^*\|_F^2 - 2\eta \left[\ell_k^\rho(B_k) - \ell_k^\rho(B^*) + \frac{1}{2\rho} \|\nabla \ell_k^\rho(B_k)\|_F^2 \right] + \eta^2 \|\nabla \ell_k^\rho(B_k)\|_F^2 \\ &\leq \|B_{k+1} - B^*\|_F^2 - 2\eta [\ell_k^\rho(B_k) - \ell_k^\rho(B^*)] + 2\rho\eta^2 [\ell_k^\rho(B_k) - \ell_k^\rho(B^*)] - \frac{\eta}{\rho} \|\nabla \ell_k^\rho(B_k)\|_F^2 \\ &= \|B_{k+1} - B^*\|_F^2 - 2\eta(1 - \rho\eta) [\ell_k^\rho(B_k) - \ell_k^\rho(B^*)] - \frac{\eta}{\rho} \|\nabla \ell_k^\rho(B_k)\|_F^2 \\ &= \|B_{k+1} - B^*\|_F^2 - 2\eta(1 - \rho\eta) \ell_k^\rho(B_k) - 2\eta \ell_k^\rho(B_k) \\ (\eta = \rho^{-1}) &= \|B_{k+1} - B^*\|_F^2 - 2\rho^{-1} \ell_k^\rho(B_k), \end{aligned}$$

where the first inequality uses cocoercivity of convex functions [GHU26] and the last one uses $\|\nabla \ell_k^\rho(B_k)\|_F^2 = 2\rho \ell_k^\rho(B_k)$. Therefore, $\|B - B^*\|_F^2$ is reduced by an amount that is proportional to $\ell_k^\rho(B_k)$.

However, there is one issue left: the quantity $\ell_k^\rho(B_k) = \frac{\rho}{2\|s^k\|^2} \|Bs^k - y^k\|^2$ is simply a weighted residual of secant equation and seems irrelevant to any convergence metric. How can we connect it to a desired quantity such as contraction? This is a nontrivial question. But fortunately, we can stand on the shoulders of the giants. The famous (variant of) the Dennis-Moré condition provides what we need.

Lemma 1. (Dennis-Moré variant) Suppose $x^+ = x - B^{-1} \nabla f(x)$. Then with $s = x^+ - x, y = \nabla f(x^+) - \nabla f(x)$,

$$\frac{\|Bs - y\|^2}{\|s\|^2} \geq \frac{\mu^2}{2} \frac{\beta}{\beta + 1},$$

where $\beta := \frac{\|\nabla f(x^+)\|^2}{\|\nabla f(x)\|^2}$. Hence $\ell_{y,s}(B) = \frac{\rho}{2\|s\|^2} \|Bs - y\|^2 \geq \frac{\rho\mu^2}{4} \frac{\beta}{\beta + 1}$.

With **Lemma 1**, we have the desired “either-or” relation: at each iteration, we have

- either $\frac{\|\nabla f(x^{k+1/2})\|^2}{\|\nabla f(x^k)\|^2} = \beta_k$ is good
- or $\|B_{k+1} - B^*\|_F^2 \leq \|B_{k+1} - B^*\|_F^2 - 2\rho^{-1} \ell_k^\rho(B_k) \leq \|B_{k+1} - B^*\|_F^2 - \frac{\mu^2}{2} \frac{\beta_k}{\beta_k + 1}$.

Note that it is possible that $\beta_k \rightarrow \infty$ while $\ell_k^\rho(B_k)$ remains bounded, and null step guarantees that

$$r_k = \frac{\|\nabla f(x^{k+1})\|^2}{\|\nabla f(x^k)\|^2} = \min \{\beta_k, 1\}$$

so that the convergence metric will not blow up.

Potential reduction. As we did in Blog #3, define the learning potential:

$$\phi(x, B) := \log \|\nabla f(x)\|^2 + \frac{2}{\mu^2} \|B - B^*\|_F^2.$$

and it is easy to show that this potential reduces by at least a constant in each iteration:

Proposition 2. We have $\phi(x^{k+1}, B_{k+1}) \leq \phi(x^k, B_k) - \frac{1}{2}$ for all $k \geq 1$.

And the total iteration complexity follows immediately.

Theorem 1. *The total complexity for finding an ε -stationary point of f is*

$$\mathcal{O}\left(\frac{2}{\mu^2}\|B_1 - B^*\|_F^2 + \log\left(\frac{\|\nabla f(x^1)\|^2}{\varepsilon}\right)\right).$$

Superlinear convergence. We can also obtain superlinear convergence result.

Theorem 2. *Algorithm 1 also satisfies the following superlinear convergence rate*

$$\frac{\|\nabla f(x^{K+1})\|^2}{\|\nabla f(x^1)\|^2} \leq \left(\frac{4\|B_1 - B^*\|_F^2}{\mu^2 K}\right)^K.$$

1.2 Final remarks

The perspective of viewing quasi-Newton methods through the lens of online learning is not new. For example, in [JM26, JJM23], the authors consider the Dennis-Moré condition as an online loss function and designs an elegant online learning procedure to learn the Hessian. My own work (OSGM) also falls into this category, and I'll cover them in the next posts.

References

- [Dav91] William C Davidon. Variable metric method for minimization. *SIAM Journal on optimization*, 1(1):1–17, 1991.
- [Fle91] Roger Fletcher. A new variational result for quasi-newton formulae. *SIAM Journal on Optimization*, 1(1):18–21, 1991.
- [Fle00] Roger Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2000.
- [GHU26] Wenzhi Gao, Chang He, and Madeleine Udell. Small gradient norm regret for online convex optimization. *ArXiv preprint arXiv:2601.13519*, 2026.
- [JJM23] Ruichen Jiang, Qiujiang Jin, and Aryan Mokhtari. Online learning guided curvature approximation: a quasi-newton method with global non-asymptotic superlinear convergence. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1962–1992. PMLR, 2023.
- [JM26] Ruichen Jiang and Aryan Mokhtari. Online learning guided quasi-newton methods with global non-asymptotic convergence: r. jiang and a. mokhtari. *Mathematical Programming*, pages 1–64, 2026.
- [NW06] J Nocedal and S Wright. Numerical optimization. 2006.
- [PB14] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

A Proof of Lemma 1

For quadratic f , we have $x^+ = x - B^{-1}\nabla f(x)$ and $B(x^+ - x) = -\nabla f(x)$, giving

$$Bs - y = -\nabla f(x) - (\nabla f(x^+) - \nabla f(x)) = -\nabla f(x^+)$$

and

$$\begin{aligned} \frac{\|Bs - y\|^2}{\|s\|^2} &= \frac{\|\nabla f(x^+)\|^2}{\|x^+ - x^* - (x - x^*)\|^2} \\ &\geq \frac{1}{2} \frac{\|\nabla f(x^+)\|^2}{\|x^+ - x^*\|^2 + \|x - x^*\|^2} \\ &\geq \frac{\mu^2}{2} \frac{\|\nabla f(x^+)\|^2}{\|\nabla f(x^+)\|^2 + \|\nabla f(x)\|^2} = \frac{\mu^2}{2} \frac{\frac{\|\nabla f(x^+)\|^2}{\|\nabla f(x)\|^2}}{\frac{\|\nabla f(x^+)\|^2}{\|\nabla f(x)\|^2} + 1}. \end{aligned}$$

This completes the proof.

A.1 Proof of Proposition 2

By definition, we have

$$\log \|\nabla f(x^{k+1})\|^2 - \log \|\nabla f(x^k)\|^2 = \log \frac{\|\nabla f(x^{k+1})\|^2}{\|\nabla f(x^k)\|^2} = \log \min \{\beta_k, 1\}$$

and $\|B_{k+1} - B^*\|_F^2 \leq \|B_k - B^*\|_F^2 - \frac{\mu^2}{2} \frac{\beta_k}{\beta_k + 1}$. Hence the potential function is reduced by

$$\begin{aligned} \phi(x^{k+1}, B_{k+1}) &\leq \phi(x^k, B_k) + \log \min \{\beta_k, 1\} - \frac{\beta_k}{\beta_k + 1} \\ &\leq \phi(x^k, B_k) + \max_{\beta} \left\{ \log \min \{\beta, 1\} - \frac{\beta}{\beta + 1} \right\} \\ &= \phi(x^k, B_k) - \frac{1}{2} \end{aligned}$$

and this completes the proof.

A.2 Proof of Theorem 2

If $\beta \leq 1$, we have $r = \beta$ and $\frac{\beta}{\beta+1} \geq \frac{\beta}{2} = \frac{r}{2}$. Otherwise $\beta \geq 1$, and we have $r = 1$ and $\frac{\beta}{\beta+1} \geq \frac{1}{2} = \frac{r}{2}$. Hence

$$\|B_{k+1} - B^*\|_F^2 \leq \|B_k - B^*\|_F^2 - \frac{\mu^2}{2} \frac{\beta_k}{\beta_k + 1} \leq \|B_k - B^*\|_F^2 - \frac{\mu^2 r_k}{4}$$

and $\sum_{k=1}^K r_k \leq \frac{4\|B_1 - B^*\|_F^2}{\mu^2}$, giving

$$\frac{\|\nabla f(x^{K+1})\|^2}{\|\nabla f(x^1)\|^2} = \prod_{k=1}^K r_k \leq \left(\frac{1}{K} \sum_{k=1}^K r_k \right)^K \leq \left(\frac{4\|B_1 - B^*\|_F^2}{\mu^2 K} \right)^K$$

and completes the proof.