New Results on Superlinear Convergence of Classical Quasi-Newton Methods

Anton Rodomanov (UCLouvain)
(joint work with Y. Nesterov)

July 7, 2021
EUROPT 2021 (virtual)
18th Workshop on Advances in Continuous Optimization

Classical Quasi-Newton (QN) Methods

Problem: $\min_{x \in \mathbb{R}^n} f(x)$, where $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function.

General scheme of a QN method

Choose some $x_0 \in \mathbb{R}^n$, $H_0 \succ 0$ and iterate for $k \geq 0$:

1. Set $x_{k+1} := x_k - \alpha_k H_k \nabla f(x_k)$ for some $\alpha_k \geq 0$.
2. Update H_k into H_{k+1}.

Main idea: Ensure that $H_k \approx [\nabla^2 f(x_k)]^{-1}$.

Standard updating rules: $H_{k+1} := \mathrm{DFP}^{-1}(H_k, u_k, \gamma_k)$ and $H_{k+1} := \mathrm{BFGS}^{-1}(H_k, u_k, \gamma_k)$, where $u_k := x_{k+1} - x_k$, $\gamma_k := \nabla f(x_{k+1}) - \nabla f(x_k)$.

- $\mathrm{DFP}^{-1}(H, u, \gamma) := H - \frac{H\gamma\gamma^T H}{\langle \gamma, H\gamma \rangle} + \frac{uu^T}{\langle \gamma, u \rangle}$,
- $\mathrm{BFGS}^{-1}(H, u, \gamma) := H - \frac{H\gamma u^T + u\gamma^T H}{\langle \gamma, u \rangle} + \left(\frac{\langle \gamma, H\gamma \rangle}{\langle \gamma, u \rangle} + 1 \right) \frac{uu^T}{\langle \gamma, u \rangle}$,

Superlinear Convergence. Historical Remarks

Main result: $\frac{\|\nabla f(x_{k+1})\|}{\|\nabla f(x_k)\|} \to 0$ as $k \to \infty$.

Historical remarks:

1. [Powell, 1971] Superlinear convergence of DFP with exact line search.
2. [Dixon, 1972] Under exact line search, all methods from Broyden's class (SR1, DFP, BFGS, ...) coincide.
3. [Broyden, Dennis, Moré, 1973] Superlinear convergence of DFP, BFGS (and others) without line search (unit step size).
4. [Dennis, Moré, 1974] Characterization of superlinear convergence for QN methods.
5. ...

Open question

Rate of superlinear convergence?
(explicit nonasymptotic estimates)

QN Methods from Convex Broyden Class

Convex Broyden class ($\tau \in [0,1]$):

$$\mathrm{Broyd}_\tau^{-1}(H, u, \gamma) := (1-\tau)\,\mathrm{BFGS}^{-1}(H, u, \gamma) + \tau\,\mathrm{DFP}^{-1}(H, u, \gamma).$$

Main instances:

- $\tau = 0 \implies$ BFGS.
- $\tau = 1 \implies$ DFP.

Classical QN scheme ($\tau \in [0,1]$)

Choose $x_0 \in \mathbb{R}^n$, $H_0 \succ 0$ and iterate for $k \geq 0$:

1. Compute $x_{k+1} := x_k - H_k \nabla f(x_k)$.
2. Compute $u_k := x_{k+1} - x_k$, $\gamma_k := \nabla f(x_{k+1}) - \nabla f(x_k)$.
3. Update $H_{k+1} := \mathrm{Broyd}_\tau^{-1}(H_k, u_k, \gamma_k)$.

Main Assumptions

Assume the function f is:

1. μ-strongly convex with L-Lipschitz gradient ($\mu, L > 0$):

$$\mu I \preceq \nabla^2 f(x) \preceq LI, \qquad \forall x \in \mathbb{R}^n,$$

Condition number: $Q := L/\mu \ (\geq 1)$.

2. M-strongly self-concordant ($M \geq 0$):

$$\nabla^2 f(x) - \nabla^2 f(y) \preceq M\|x - y\|_z \nabla^2 f(w), \qquad \forall x, y, z, w \in \mathbb{R}^n,$$

where $\|h\|_z := \langle \nabla^2 f(z)h, h\rangle^{1/2}$.

Remarks:

- For quadratic functions $M = 0$.
- 1 + 2 $\iff$ 1 + L_2-Lipschitz Hessian.
- 2 is an affine invariant property.

Main property: For any $x, y \in \mathbb{R}^n$ and $r := \|y - x\|_x$:

$$(1 + Mr)^{-1} \nabla^2 f(x) \preceq \nabla^2 f(y) \preceq (1 + Mr)\nabla^2 f(x)$$

Efficiency Estimates

Local gradient norm: $\lambda_k := \|\nabla f(x_k)\|_{x_k}^*$.

Theorem. Suppose $H_0 = \frac{1}{L} I$ and x_0 is sufficiently close to the solution:

$$M\lambda_0 \leq \frac{\ln \frac{3}{2}}{(\frac{3}{2})^{\frac{3}{2}}} \max\left\{\frac{1}{2Q}, \frac{1}{K_0 + 9}\right\}, \quad K_0 := \lceil 8nQ_\tau \ln(2Q) \rceil,$$

where $Q_\tau := (1 - \tau + \tau \frac{4}{9} Q^{-1})^{-1}$. Then, for all $k \geq 0$, we have

$$\lambda_k \leq \left(1 - \frac{1}{2Q}\right)^k \sqrt{\frac{3}{2}}\, \lambda_0,$$

and, for all $k \geq 1$, we have

$$\lambda_k \leq \left[\frac{5}{2} Q_\tau \left(\exp\left\{\frac{13n\ln(2Q)}{6k}\right\} - 1\right)\right]^{k/2} \sqrt{\frac{3Q}{2}}\, \lambda_0.$$

Remark: For quadratic functions $M = 0$ and this is global convergence.

Discussion

BFGS ($\tau = 0$):

- Region of local convergence: $M\lambda_0 \lesssim \max\{Q^{-1}, [n \ln Q]^{-1}\}$.
- Rate:

$$\left[\exp\left\{\frac{n \ln Q}{k}\right\} - 1\right]^k \lesssim \left(\frac{n \ln Q}{k}\right)^k, \qquad k \gtrsim n \ln Q.$$

DFP ($\tau = 1$):

- Region of local convergence: $M\lambda_0 \lesssim Q^{-1}$.
- Rate:

$$\left[Q\left(\exp\left\{\frac{n \ln Q}{k}\right\} - 1\right)\right]^k \lesssim \left(\frac{nQ \ln Q}{k}\right)^k, \qquad k \gtrsim nQ \ln Q.$$

Note:

- BFGS has logarithmic dependence on the condition number Q.
- DFP is much slower.

Notation

Let $A \succ 0$ and $u \in \mathbb{R}^n \setminus \{0\}$ be arbitrary. Define

$$H_+ := \mathrm{Broyd}_\tau^{-1}(H, u, \gamma), \qquad \gamma := Au.$$

Classical QN update:

$$u := x_+ - x, \; A := \int_0^1 \nabla^2 f(x + tu) dt \implies Au = \nabla f(x_+) - \nabla f(x).$$

Remark: For the analysis, it is more convenient to work in terms of the primal matrices

$$G := H^{-1}, \quad G_+ := H_+^{-1}.$$

Eigenvalue Property

Eigenvalue property

For any $u \in \mathbb{R}^n$, $\tau \in [0, 1]$ and $\xi, \eta \geq 1$:

$$\xi^{-1} A \preceq G \preceq \eta A \quad \Longrightarrow \quad \xi^{-1} A \preceq G_+ \preceq \eta A.$$

Corollary: For a quadratic function f with Hessian A, we have

$$A \preceq G_0 \preceq QA \qquad [\text{since } G_0 = LI]$$

(recall that $Q := L/\mu$). Therefore, for all $k \geq 0$:

$$A \preceq G_k \preceq QA.$$

$\Longrightarrow$ The method has the linear convergence with constant $1 - Q^{-1}$.

Quality of Approximation

Directional measure of closeness: $\nu(A, G, u) := \frac{\|(G-A)u\|^*_{G_+}}{\|u\|_G}$.

Here $\|u\|_G := \langle Gu, u\rangle^{1/2}$, $\|s\|^*_{G_+} := \langle s, G_+^{-1}s\rangle^{1/2}$.

Note: If $u = x_+ - x = -G^{-1}\nabla f(x)$ and $A = \int_0^1 \nabla^2 f(x + tu)dt$, then

$$\nu(A, G, u) = \frac{\|\nabla f(x_+)\|^*_{G_+}}{\|\nabla f(x)\|^*_G}$$

because $Au = \nabla f(x_+) - \nabla f(x)$.

Corollary: $\nu_k \to 0 \iff \nabla f(x_k) \to 0$ superlinearly.

Potential Function

Augmented Log-Det Barrier

For $X, Y \succ 0$, define

$$\psi(X, Y) := -\ln\det Y + \ln\det X + \langle X^{-1}, Y - X \rangle,$$

where $\langle U, V \rangle := \mathrm{tr}(UV)$ is the Frobenius inner product.

Remarks:

- This is the Bregman distance generated by $d(X) := -\ln\det X$:

$$\psi(X, Y) = d(Y) - d(X) - \langle \nabla d(X), Y - X \rangle \geq 0.$$

- First used in [Byrd, Nocedal, 1989] for the analysis of QN methods.

Main Result

Main result: If $\xi^{-1} A \preceq G \preceq \eta A$ for some $\xi, \eta \geq 1$, then

$$\psi(G_+, A) \leq \psi(G, A) - \frac{6}{13} \ln(1 + \delta\nu^2),$$

where $\delta := \frac{1}{1+\xi}(1 - \tau + \tau \frac{1}{\xi\eta})$.

Corollary: If f is quadratic, then $\nu \to 0$.

Note: $\psi(G_0, A) \leq \ln \det G_0 - \ln \det A \leq n \ln Q$ (since $A \preceq G_0 \preceq QA$).

Remark: In the nonlinear case, we need to additionally bound the "error" term $\psi(G_+, A_+) - \psi(G_+, A)$.

Conclusion

- We have obtained explicit and nonasymptotic rates of local superlinear convergence for classical BFGS and DFP quasi-Newton methods.
- The main factor in these estimates is the starting moment of superlinear convergence: $O(n \ln Q)$ for BFGS and $O(nQ \ln Q)$ for DFP, where n is the problem dimension and Q is its condition number.

Open questions:

- Is it possible to remove the $\ln Q$ factor?
- Choice of initial matrix ($H_0 = \frac{1}{L} I$)?

Paper

A. Rodomanov, Y. Nesterov. New Results on Superlinear Convergence of Classical Quasi-Newton Methods. *Journal of Optimization Theory and Applications* **188**, 744–769 (2021).

Thank you!