New Results on Superlinear Convergence of Classical Quasi-Newton Methods

A. Rodomanov (UCLouvain)

March 4, 2021
XIII Symposium on Numerical Analysis and Optimization
UFPR, Brazil (online)

Gradient Method

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

Gradient method: $x_{k+1} = x_k - h_k \nabla f(x_k), \quad h_k > 0,\ k \geq 0.$

Assumptions: f is 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: $Q \coloneqq L/\mu \geq 1$.

Theorem. Let $h_k \equiv 1/L$. Then, for all $k \geq 0$, we have

$$\|\nabla f(x_k)\| \leq (1 - Q^{-1})^k \|\nabla f(x_0)\|.$$

Corollary: Since $1 - Q^{-1} \leq \exp(-Q^{-1})$, we get $\|\nabla f(x_k)\| \leq \epsilon \|\nabla f(x_0)\|$ in

$$\boxed{Q \ln \frac{1}{\epsilon}} \text{ iterations.}$$

Newton's Method

Newton's method: $x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k), \quad k \geq 0.$

Interpretation: Minimization of Taylor's second-order model:

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \Big[f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \tfrac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle \Big]$$

Assumptions: f is μ-strongly convex with L_2-Lipschitz Hessian:

$$\nabla^2 f(x) \succeq \mu I, \qquad \|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L_2 \|x - y\|, \qquad \forall x, y \in \mathbb{R}^n.$$

Theorem. $\|\nabla f(x_k)\| \leq \dfrac{2\mu^2}{L_2} \left(\dfrac{L_2}{2\mu^2} \|\nabla f(x_0)\| \right)^{2^k}, \quad k \geq 0.$

Corollary: $\|\nabla f(x_0)\| \leq \dfrac{\mu^2}{L_2} \implies \|\nabla f(x_k)\| \leq \left(\dfrac{1}{2} \right)^{2^k - 1} \|\nabla f(x_0)\|, \ k \geq 0.$

Thus, we get $\|\nabla f(x_k)\| \leq \epsilon \|\nabla f(x_0)\|$ in $\boxed{\log_2 \log_2 \frac{2}{\epsilon}}$ iterations.

Comparison of Gradient and Newton Methods

Gradient method: $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k), \quad k \geq 0.$

+ Very simple. Only requires computing $\nabla f(x_k)$.
+ Iteration cost: $O(n)$.
+ Global linear convergence.
− Very sensitive to condition number Q.

Newton's method: $x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1}\nabla f(x_k), \quad k \geq 0.$

+ Extremely fast quadratic convergence.
− Requires additionally computing and inverting $\nabla^2 f(x_k)$.
− Iteration cost: $O(n^3)$.
− Convergence is only local.

Can we have something in between?

Quasi-Newton Methods. General Idea

General Quasi-Newton Method

Start with $H_0 = L^{-1}I$ and iterate for $k \geq 0$:

1. Set $x_{k+1} = x_k - H_k \nabla f(x_k)$.
2. Update H_k into H_{k+1}.

Main idea: Make $H_k \approx [\nabla^2 f(x_k)]^{-1}$ by using only the gradients of f and spending at most $O(n^2)$ operations for updating H_k into H_{k+1}.

Updating Hessian Aproximation

Goal: Improve $H \approx A^{-1}$ into $H_+ \approx A^{-1}$.

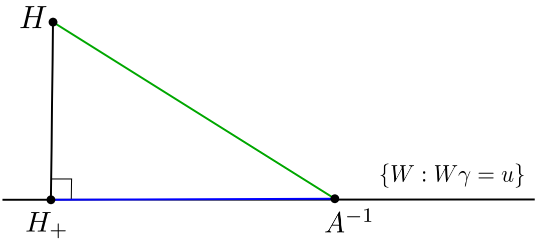
Approximation along direction

Select $u \neq 0$ and compute $\gamma = Au$. Make sure that H_+ satisfies

$$H_+^{-1} u = \gamma \iff H_+ \gamma = u.$$

Note: H_+ is not uniquely defined.

Main idea: Let H_+ be the projection of H onto $\{W : W\gamma = u\}$.



Bregman divergence

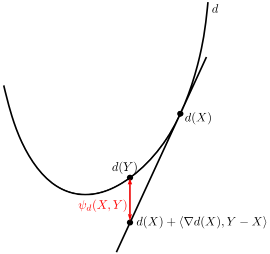
Bregman divergence

For a smooth strictly convex function d, define

$$\psi_d(X, Y) := d(Y) - d(X) - \langle \nabla d(X), Y - X \rangle$$

Properties:

- $\psi_d(X, Y) \geq 0$.
- $\psi_d(X, Y) = 0 \iff X = Y$.
- In general, $\psi_d(X, Y) \neq \psi_d(Y, X)$.



Main example: $d(X) := -\ln \det X$, defined on the set of $n \times n$ symmetric positive definite matrices:

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

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

BFGS and DFP Updates

Option 1: $H_+ = \operatorname{argmin}_{H_+}\{\psi(H_+, H) : H_+\gamma = u\}$.

Broyden–Fletcher–Goldfarb–Shanno (BFGS) update

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

Option 2: $H_+ = \operatorname{argmin}_{H_+}\{\psi(H, H_+) : H_+\gamma = u\}$.

Davidon–Fletcher–Powell (DFP) update

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

Remark: When we want to highlight that $\gamma = Au$, we prefer to use notation $\mathrm{BFGS}^{-1}(A, H, u)$ and $\mathrm{DFP}^{-1}(A, H, u)$.

Classical Quasi-Newton Methods

Classical BFGS and DFP Methods

Start with $H_0 = L^{-1} I$ and iterate for $k \geq 0$:

1. Set $x_{k+1} = x_k - H_k \nabla f(x_k)$.
2. Compute $u_k = x_{k+1} - x_k$, $\gamma_k = \nabla f(x_{k+1}) - \nabla f(x_k)$.
3. Set $H_{k+1} = \mathrm{BFGS}^{-1}(H_k, u_k, \gamma_k)$ or $H_{k+1} = \mathrm{DFP}^{-1}(H_k, u_k, \gamma_k)$.

Remarks:

- $\gamma_k = A_k u_k$, where $A_k := \int_0^1 \nabla^2 f(x_k + t u_k) dt$.
- If f is quadratic, $f(x) = \frac{1}{2}\langle Ax, x \rangle + \langle b, x \rangle$, then $A_k = \nabla^2 f(x_k) = A$.
- In practice, BFGS is much more efficient than DFP.

Superlinear Convergence. Historical Remarks

Main result: $\dfrac{\|\nabla f(x_{k+1})\|}{\|\nabla f(x_k)\|} \to 0$ 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 quasi-Newton methods.
5. ...

Open question

<u>Rate</u> of superlinear convergence?
(explicit nonasymptotic estimates)

Convex Broyden Class

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

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

Classical Quasi-Newton method ($\tau \in [0, 1]$)

Set $H_0 = L^{-1} I$ and iterate for $k \geq 0$:

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

Remark: For the analysis, it is more convenient to work in terms of the primal matrices $G \equiv H^{-1}$ and to highlight that $\gamma = Au$. We denote the corresponding update by $G_+ = \mathrm{Broyd}_\tau(A, G, u)$.

Eigenvalue Property

Eigenvalue property

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

$$\xi^{-1} A \preceq G \preceq \eta A \quad \Longrightarrow \quad \xi^{-1} A \preceq \mathrm{Broyd}_\tau(A, G, u) \preceq \eta A.$$

Corollary: If f is quadratic with Hessian A, then, for all $k \geq 0$,

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

(Recall that $G_0 = LI$, $Q = L/\mu$.)

Note: This implies 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

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

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

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

where $\delta := \frac{1}{1+\xi}(1 - \tau + \tau \frac{1}{\xi\eta})$, ψ is the log-det Bregman divergence.

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

Main assumptions

Assume the function f is:

1. μ-strongly convex with L-Lipschitz gradient ($\mu, L > 0$):
$$\mu I \preceq \nabla^2 f(x) \preceq LI, \qquad \forall x \in \mathbb{R}^n,$$
Denote by $Q := L/\mu \geq 1$ the condition number.
2. M-strongly self-concordant ($M \geq 0$):
$$\nabla^2 f(x) - \nabla^2 f(y) \preceq M\|x - y\|_{\nabla^2 f(z)} \nabla^2 f(w), \qquad \forall x, y, z, w \in \mathbb{R}^n.$$

Remark: This is the same class as that of all μ-strongly convex functions with L-Lipschitz gradient and L_2-Lipschitz Hessian for some $L_2 > 0$. In particular, we can take $M = L_2/\mu^{3/2}$.

Main property: For any $x, y \in \mathbb{R}^n$, $z \in \{x, y\}$ and $r = \|y - x\|_x$:
$$(1 + Mr)^{-1} \nabla^2 f(x) \preceq \nabla^2 f(y) \preceq (1 + Mr)\nabla^2 f(x),$$
$$(1 + \tfrac{1}{2}Mr)^{-1} \nabla^2 f(z) \preceq \int_0^1 \nabla^2 f(x + t(y - x))dt \preceq (1 + \tfrac{1}{2}Mr)\nabla^2 f(z).$$

Efficiency Estimates

Local gradient norm: $\lambda_k := \|\nabla f(x_k)\|^*_{\nabla^2 f(x_k)}$.

Theorem. Suppose x_0 is sufficiently close to the solution:

$$M\lambda_0 \le \frac{\ln \frac{3}{2}}{(\frac{3}{2})^{\frac{3}{2}}} \max\left\{\frac{1}{2Q}, \frac{1}{K_0 + 9}\right\}, \quad K_0 := \left\lceil \left(1 - \tau + \tau \frac{4}{9Q}\right)^{-1} 8n \ln(eQ) \right\rceil.$$

Then, for all $k \ge 0$, we have

$$\frac{2}{3}\nabla^2 f(x_k) \preceq G_k \preceq \frac{3Q}{2}\nabla^2 f(x_k),$$

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

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

$$\lambda_k \le \left[\frac{5}{2}\left(1 - \tau + \tau\frac{4}{9Q}\right)^{-1}\left(\exp\left\{\frac{13n\ln(eQ)}{6k}\right\} - 1\right)\right]^{k/2} \sqrt{\frac{3Q}{2}}\,\lambda_0.$$

Discussion

BFGS ($\tau = 0$):

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

DFP ($\tau = 1$):

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

Note:

- BFGS has <u>logarithmic</u> dependence on the condition number.
- DFP is much slower (very sensitive to the condition number).

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.

Paper

A. Rodomanov, Y. Nesterov. New Results on Superlinear Convergence of Classical Quasi-Newton Methods (2020), arXiv:2004.14866.

Thank you!