Greedy Quasi-Newton Method with Explicit Superlinear Convergence

Anton Rodomanov   Yurii Nesterov

Catholic University of Louvain, Belgium

8 October 2019

Seminar in Mathematical Engineering, Louvain-la-Neuve

Quasi-Newton methods for minimizing functions

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

General quasi-Newton method

Initialize $x_0 \in \mathbb{R}^n$, $H_0 \in \mathbb{S}^n_{++}$ and iterate for $k \geq 0$:

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

Denote $s_k := x_{k+1} - x_k$ and $y_k := f'(x_{k+1}) - f'(x_k)$.

- (SR1) $H_{k+1} := H_k + \frac{(s_k - H_k y_k)(s_k - H_k y_k)^T}{\langle y_k, s_k - H_k y_k \rangle}$.
- (DFP) $H_{k+1} := H_k - \frac{H_k y_k y_k^T H_k}{\langle y_k, H_k y_k \rangle} + \frac{s_k s_k^T}{\langle y_k, s_k \rangle}$.
- (BFGS) $H_{k+1} := \left(I - \frac{s_k y_k^T}{\langle y_k, s_k \rangle} \right) H_k \left(I - \frac{y_k s_k^T}{\langle y_k, s_k \rangle} \right) + \frac{s_k s_k^T}{\langle y_k, s_k \rangle}$.

Superlinear convergence of quasi-Newton methods

Theorem (Dennis-Moré 1974, 1977)

If (x_0, H_0) is sufficiently close to $(x^*, f''(x^*)^{-1})$, then both DFP and BFGS are superlinearly convergent: $\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} \to 0$.

Main question: Rate of convergence? $O(c^{k^2})$, $O(c^{k^3})$, $O(k^{-k})$, ...?

Our goal:

Present a new quasi-Newton method with an explicit superlinear rate.

BFGS update and norms

Definition (BFGS update)

For $A \in \mathbb{S}^n_{++}$, $H \in \mathbb{S}^n$ and $s \in \mathbb{R}^n$, define

$$\mathrm{BFGS}(H, A, s) := \left(I - \frac{ss^T A}{\langle As, s\rangle}\right) H \left(I - \frac{Ass^T}{\langle As, s\rangle}\right) + \frac{ss^T}{\langle As, s\rangle}.$$

- Here A plays the role of $f''(x)$ and $y := As$.

Our goal: Decrease the distance between H and A^{-1}.

Main property of BFGS update

- Introduce the Euclidean norm induced by A:
$$\|x\|_A := \langle Ax, x\rangle^{1/2}.$$
- The corresponding conjugate norm:
$$\|y\|_A^* := \max_{\|x\|_A \le 1} \langle y, x\rangle = \langle y, A^{-1}y\rangle^{1/2}.$$
- Operator norm:
$$\|W\|_A := \max_{\|y\|_A^* \le 1} \|Wy\|_A = \lambda_{\max}(WAWA)^{1/2}.$$
- Frobenius norm:
$$\|W\|_{\mathrm{Fr}(A)} := \mathrm{Tr}(WAWA)^{1/2} \quad (\ge \|W\|_A).$$

Lemma (Progress in matrix for BFGS update)

For $H_+ := \mathrm{BFGS}(H, A, s)$, *we have*
$$\|A^{-1} - H_+\|^2_{\mathrm{Fr}(A)} \le \|A^{-1} - H\|^2_{\mathrm{Fr}(A)} - \frac{\|(HA - I)s\|_A^2}{\|s\|_A^2}.$$

Greedy BFGS update

Definition (Greedy BFGS update)

Let $e_1, \ldots, e_n$ be the standard orthonormal basis in $\mathbb{R}^n$. For

$$i_{\max}(H, A) := \operatorname*{argmax}_{1 \le i \le n} \frac{\|(HA - I)e_i\|_A^2}{\|e_i\|_A^2},$$

define

$$\text{GreedyBFGS}(H, A) := \text{BFGS}(H, A, e_{i_{\max}(H,A)}).$$

- Makes the maximal progress keeping the update cost relatively small.
- **NB:** Using more sophisticated reasoning, one can instead work with
$$i_{\max}(H, A) := \operatorname*{argmax}_{1 \le i \le n} \frac{\langle Be_i, e_i \rangle}{\langle Ae_i, e_i \rangle},$$
where $B := H^{-1}$. This requires computing only the diagonal of the Hessian at each iteration.

Main property of greedy BFGS update

Lemma (Linear convergence in matrix)

For $H_+ := \mathrm{GreedyBFGS}(H, A)$, *we have*

$$\|A^{-1} - H_+\|_{\mathrm{Fr}(A)} \leq (1 - \rho)\|A^{-1} - H\|_{\mathrm{Fr}(A)},$$

where $\rho := \rho(A)$ *is the coordinate condition number of* A*:*

$$\rho(A) := \frac{\lambda_{\min}(A)}{2\,\mathrm{Tr}(A)} \geq \frac{\lambda_{\min}(A)}{2n\lambda_{\max}(A)}.$$

- Follows from lower bounding the maximum by the expectation for i chosen randomly with probability $\pi_i := \frac{\|a_i\|_A^2}{\mathrm{Tr}(A)}$.
- The randomized version was first proposed in [Gower-Richtárik 2016].

Superlinear convergence on quadratic functions

Consider a simple quadratic function

$$f(x) := \frac{1}{2}\langle Ax, x\rangle = \frac{1}{2}\|x\|_A^2.$$

- Denote $r_k := \|x_k - x^*\|_A$ and $\sigma_k := \|A^{-1} - H_k\|_{\mathrm{Fr}(A)}$.
- **Quasi-Newton step:** $x_{k+1} = x_k - H_k f'(x_k) = (A^{-1} - H_k)Ax_k$.
- Hence,

$$r_{k+1} \le \sigma_k r_k \qquad \Rightarrow \qquad r_k \le r_0 \prod_{i=0}^{k-1} \sigma_i.$$

- From the previous slide,

$$\sigma_{k+1} \le (1-\rho)\sigma_k \qquad \Rightarrow \qquad \sigma_k \le (1-\rho)^k \sigma_0.$$

- Thus,

$$r_k \le r_0 \prod_{i=0}^{k-1} ((1-\rho)^i \sigma_0) = \sigma_0^k (1-\rho)^{\frac{k(k-1)}{2}} r_0.$$

Conclusion: If $\sigma_0 \le \frac{1}{2}$, we obtain the $(\frac{1}{2})^k(1-\rho)^{k^2}$ superlinear rate.

Can we prove similar results for general nonlinear f?

GreedyBFGS method

Problem: $\min_{x \in \mathbb{R}^n} f(x)$.

GreedyBFGS method for minimizing functions

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

1. Set $x_{k+1} := x_k - H_k f'(x_k)$
2. Set $H_{k+1} :=$ GreedyBFGS$(H_k, f''(x_{k+1}))$.

NB: $A := f''(x_{k+1})$ changes at every iteration.

General nonlinear functions

Lipschitz continuity of f'':

$$\|f''(y) - f''(x)\|_{x^*} \leq L\|y - x\|_{x^*}.$$

Lemma (Progress of one step of GreedyBFGS)

For $r_k := \frac{L}{2}\|x_k - x^*\|_{x^*}$, $\sigma_k := \|f''(x_k)^{-1} - H_k\|_{\mathrm{Fr}(x_k)}$ *and* $\rho := \rho(f''(x^*))$,

$$r_{k+1} \leq \frac{(1+r_k)^{3/2}}{(1-2r_k)\sqrt{1-r_k}}\sigma_k r_k + \frac{3\sqrt{1+r_k}}{(1-2r_k)\sqrt{1-r_k}} r_k^2$$

$$\sigma_{k+1} \leq \left(1 - \frac{1-2r_{k+1}}{1+2r_{k+1}}\rho\right)\frac{1+2r_{k+1}}{1-2r_k}\sigma_k + \frac{2\sqrt{n}}{1-2r_k}(r_k + r_{k+1}).$$

Simplification: Assuming r_k is sufficiently small, we get

$$\begin{aligned} r_{k+1} &\leq \sigma_k r_k, \\ \sigma_{k+1} &\leq (1-\rho)\sigma_k \end{aligned} \quad \Rightarrow \quad \begin{aligned} r_k &\leq \sigma_0^k (1-\rho)^{k^2} r_0 \\ \sigma_k &\leq (1-\rho)^k \sigma_0. \end{aligned}$$

Local superlinear convergence of GreedyBFGS

Theorem

If $r_0 \leq \frac{\rho}{25\sqrt{n}}$ *and* $\sigma_0 \leq \frac{1}{2}$, *then*

$$r_k \leq \left(\frac{1}{2}\right)^k \left(1 - \frac{\rho}{2}\right)^{\frac{k(k-1)}{2}} r_0$$

$$\sigma_k \leq \left(1 - \frac{\rho}{2}\right)^k \frac{1}{2}.$$

Reminder: For quadratic f, we had

$$r_k \leq \left(\frac{1}{2}\right)^k (1 - \rho)^{\frac{k(k-1)}{2}} r_0$$

$$\sigma_k \leq (1 - \rho)^k \frac{1}{2}.$$

Conclusion

- New quasi-Newton method for minimizing nonlinear functions.
- It uses classic BFGS rule with greedily selected direction.
- Explicit $(\frac{1}{2})^k(1-\rho)^{k^2}$ superlinear convergence rate.

Thank you!