Greedy Quasi-Newton Method with Explicit Superlinear Convergence

Anton Rodomanov Yurii Nesterov

Catholic University of Louvain, Belgium

6 August 2019 (ICCOPT, Berlin)

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? $O(c^{k^2})$, $O(c^{k^3})$, $O(k^{-k})$, $\ldots$?

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 $f''(x)$.
- We want to decrease the distance between H and A^{-1}.

Question: How to measure the distance between H and A^{-1}?

Main property of BFGS update

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

Lemma (Progress in matrix for BFGS update)

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

Greedy BFGS update

Definition (Greedy BFGS update)

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

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

define

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

- Makes the maximal progress keeping the update cost relatively small.
- Computation of $i_{\max}(H, A)$ will be addressed later.

Main property of greedy BFGS update

Lemma (Linear convergence in matrix)

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

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

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

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

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

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

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

- From the previous slide,

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

- Thus,

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

Conclusion: If $\sigma_0 \leq 1$, we have the $O((1-\rho)^{k^2})$ superlinear rate.

$$\boxed{\text{Can we expect similar results when } f \text{ is general nonlinear?}}$$

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 $H_{k+1} := \text{GreedyBFGS}(H_k, f''(x_{k+1}))$.

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

General nonlinear functions

Lipschitz continuity of f'':

$$\|f''(x) - f''(x^*)\|_{f''(x^*)^{-1}} \le L\|x - x^*\|_{f''(x^*)}.$$

Lemma (Progress of one step of GreedyBFGS)

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

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

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

Simplification: Assuming r_k is sufficiently small and $\sigma_0 \le 1$, we get

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

Convergence of GreedyBFGS

Theorem (Local superlinear convergence of GreedyBFGS)

If $r_0 \leq \bar{r}$ *and* $\sigma_0 \leq 0.5$, *where* $\bar{r} := \frac{2c\rho}{\sqrt{n}}$ *for* $c := 0.02$, *then*

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

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

Reminder: For quadratic f, we had

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

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

Bad initial matrix

What to do if $\sigma_0 := \|H_0 - f''(x_0)^{-1}\|_{\mathrm{Fr}(f''(x_0))} > 0.5$? (Usually $H_0 := I$.)

GreedyBFGS–II

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

1. Find smallest integer $j_k \geq 0$ such that $f(x_k - 2^{-j_k} H_k f'(x_k)) \leq f(x_k)$.
2. Set $x_{k+1} := x_k - 2^{-j_k} H_k f'(x_k)$.
3. Set $H_{k+1} := \mathrm{GreedyBFGS}(H_k, f''(x_{k+1}))$.

Convergence of GreedyBFGS–II

Theorem (Local superlinear convergence of GreedyBFGS–II)

Suppose $\frac{L}{2}\|x - x^*\|_{f''(x^*)} \le \bar{r}$ *for all* $L_f(x_0) := \{x : f(x) \le f(x_0)\}$, *and let*

$$T_0 := \begin{cases} 0 & \text{if } \sigma_0 \le 0.5 \\ 2\rho^{-1}\ln(5\sigma_0) & \text{otherwise.} \end{cases}$$

Then for $\delta := \frac{8c}{1-10c} = 0.2$ *and* $b := 1 - \frac{8c}{1-2c} = 0.8333\ldots$, *we have*

$$\left.\begin{aligned} r_k &\le \bar{r}, \\ \sigma_k &\le \delta + (1 - b\rho)^k(\sigma_0 - \delta) \end{aligned}\right\} \quad 0 \le k < T_0$$

and

$$\left.\begin{aligned} r_k &\le \left(1 - \frac{\rho}{2}\right)^{\frac{k(k+1)}{2}} \bar{r}, \\ \sigma_k &\le \left(1 - \frac{\rho}{2}\right)^k \frac{1}{2} \end{aligned}\right\} \quad k \ge T_0.$$

Computing the update

For doing the GreedyBFGS update, we need to compute

$$i_{\max}(H, A) = \underset{1 \le i \le n}{\operatorname{argmax}} \frac{\|(H - A^{-1})Ae_i\|_A^2}{\|e_i\|_A^2}$$

$$= \underset{1 \le i \le n}{\operatorname{argmax}} \frac{\langle A(H - A^{-1})A(H - A^{-1})Ae_i, e_i \rangle}{\langle Ae_i, e_i \rangle}$$

- Need to compute the diagonal of A and

$$A(H - A^{-1})A(H - A^{-1})A = AHAHA - 2AHA + A.$$

Fact: For $M_1, M_2 \in \mathbb{R}^{n \times n}$, diagonal of $M_1 M_2$ can be computed in $O(n^2)$:

$$\langle M_1 M_2 e_i, e_i \rangle = \langle M_1^T e_i, M_2 e_i \rangle, \qquad 1 \le i \le n.$$

Conclusion: It suffices to keep track of 3 matrices: A, AH and AHA.
(Note that $AHAHA = AHA(AH)^T$.)

Updating auxiliary matrices

Auxiliary matrices: A, AH, AHA.

- **Rank-1 update of H:** If $H_+ := H + \gamma v v^T$, then for $z := Av$,
$$AH_+ = AH + \gamma z v^T,$$
$$AH_+ A = AHA + \gamma z z^T.$$
- **Addition of identity to A:** If $A_+ := A + \gamma I$, then
$$A_+ H = AH + \gamma H,$$
$$A_+ H A_+ = AHA + \gamma(AH + (AH)^T) + \gamma^2 H.$$
- **Rank-1 update of A:** If $A_+ := A + \gamma v v^T$, then for $z := Hv$, $q := Az$,
$$A_+ H = AH + \gamma v z^T,$$
$$A_+ H A_+ = AHA + \gamma(v q^T + q v^T) + \gamma^2 \langle v, z\rangle v v^T.$$

Complexity of each update: $O(n^2)$.

Example 1: Sparse quadratic

Let f be a strictly convex quadratic function

$$f(x) := \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle,$$

where $A \in \mathbb{S}^n_{++}$ has at most p non-zeros in each column.

Auxiliary matrices: A, AH, AHA.

Initialization: $H_0 := I \quad \Rightarrow \quad$ need to compute $AH_0A = A^2$.

Fact: A^2 contains $\leq np^2$ non-zeros and can be computed in $O(np^2 + n^2)$.

Example 2: Sparse cubically regularized quadratic

A more complicated example:

$$f(x) := \frac{1}{2}\langle Qx, x\rangle + \langle b, x\rangle + \frac{\beta}{3}\|x\|^3,$$

where $\beta > 0$, Q is sparse with at most p non-zeros in each column. Here

$$A = f''(x) = Q + \beta\|x\|I + \frac{\beta}{\|x\|}xx^T.$$

Initialization (cost $O(np^2 + n^2)$):

1. Set $H_0 := I$, $A := Q$ and compute $AH_0A = Q^2$ (previous slide).
2. Apply $A := A + \beta\|x_0\|I$ and $A := A + \frac{\beta}{\|x_0\|}x_0x_0^T$.

Update (cost $O(n^2)$):

1. Apply two rank-1 updates for H (BFGS update).
2. Apply $A := A + \beta(\|x_{k+1}\| - \|x_k\|)$.
3. Apply $A := A + \frac{\beta}{\|x_{k+1}\|}x_{k+1}x_{k+1}^T$ and $A := A - \frac{\beta}{\|x_k\|}x_kx_k^T$.

Conclusion

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

Thank you!