Métodos quase-Newton tipo secante

Método de Newton para $\min f(x)$:

$$x^{k+1} = x^k + s_k, \quad \nabla^2 f(x^k) s_k = -\nabla f(x^k).$$

O método é bom (convergência superlinear / quadrática próximo à soluções), porém é caro (computar $\nabla^2 f$ é caro).

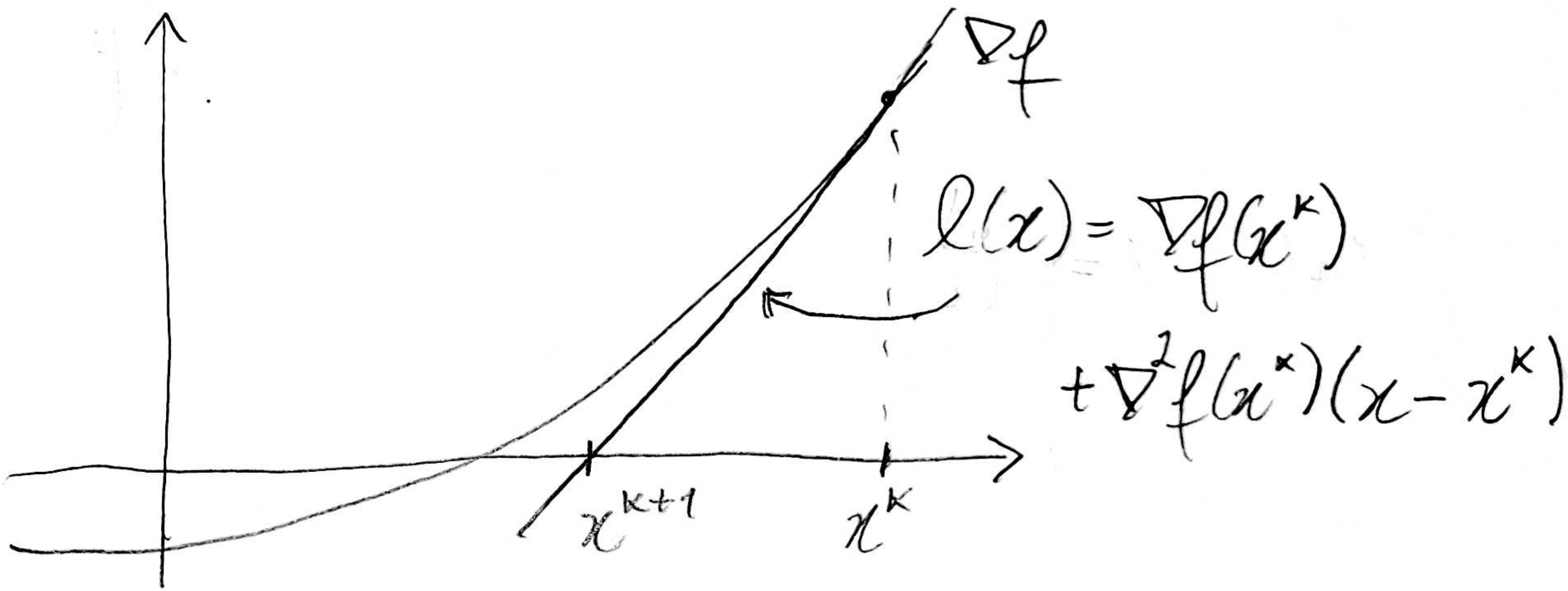
Os métodos quase-Newton tentam imitar Newton trocando $\nabla^2 f(x^k)$ por uma matriz "barata".

Aproximação linear de ∇f ao redor de x^k:

$$\nabla f(x^{k+1}) = \nabla f(x^k + s_k) \approx \nabla f(x^k) + \nabla^2 f(x^k)\, s_k.$$

Newton visa anular a aproximação linear:

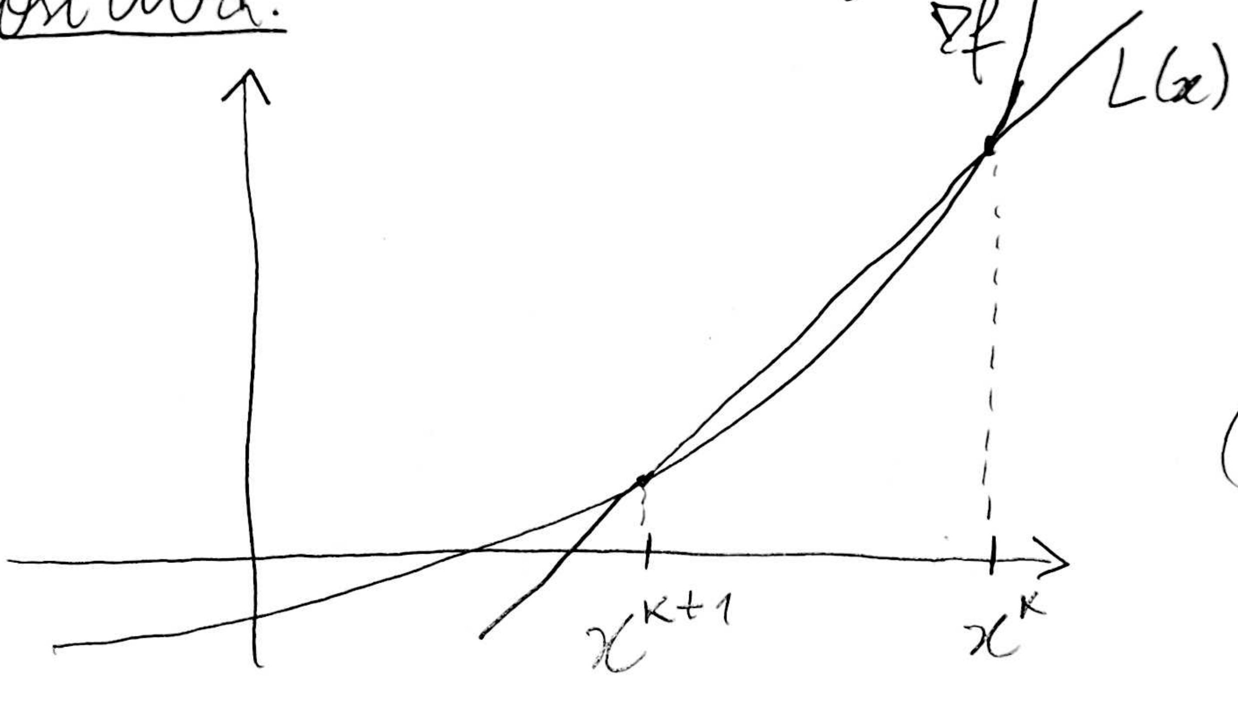
• "$\nabla f(x^{k+1}) = 0$", $\quad 0 = \nabla f(x^k) + \nabla^2 f(x^k)\, s_k$



• <u>Ideia secante</u>: interpolar $\nabla f(x)$ nos pontos x^k e x^{k+1} pela aproximação linear

$$L(x) = \nabla f(x^k) + B_{k+1}(x - x^k),$$

onde B_{k+1} é matriz <u>simétrica</u> e <u>definida positiva</u>.



Isto é,

(i) $L(x^k) = \nabla f(x^k)$

(ii) $L(x^{k+1}) = \nabla f(x^{k+1})$

A condição (i) é automática.

$$(ii) \iff \nabla f(x^k) + B_{k+1}(x^{k+1} - x^k) = \nabla f(x^{k+1})$$

$$\iff \boxed{B_{k+1} s_k = y_k}$$, onde $s_k = x^{k+1} - x^k$ e

$$y_k = \nabla f(x^{k+1}) - \nabla f(x^k).$$

(equação secante)

* Observe que B_{k+1} "faz o papel" de $\nabla^2 f(x^k)$.

* Por que B_{k+1} simétrica e definida positiva?
  ↳ Porque $\nabla^2 f(x^k)$ é simétrica;

↳ Porque Newton funciona quando $\nabla^2 f(x^k)$ é positiva definida (veja teorema de convergência)

↳ Porque $-B_{k+1}^{-1} \nabla f(x^k)$ é direção de descida para f a partir de x^k:

De fato, sendo B_{k+1} simétrica e definida positiva, admite fatoração de Cholesky $B_{k+1} = GG^t$, com G inversível. Assim $B_{k+1}^{-1} = G^{-t} G^{-1}$ é definida positiva dado

que $z^t B_{k+1}^{-1} z = z^t (G^{-1})^t G^{-1} z = \|G^{-1} z\|^2 > 0$
$\forall z \neq 0$. Logo

$$\nabla f(x^k)^t \left(-B_{k+1}^{-1} \nabla f(x^k)\right) = -\nabla f(x^k)^t B_{k+1}^{-1} \nabla f(x^k) < 0.$$

* A equação secante $B s_k = y_k$ admite infinitas soluções B caso $n > 1$ e $s_k \neq 0$.

↳ Logo há vários métodos quase-Newton!

Algumas atualizações

1) Correção de posto 1

B_k simétrica, $\quad B_{k+1} = B_k + \underbrace{uu^t}_{\text{matriz } n\times n \text{ de posto } 1}$

matriz $n \times n$ de posto 1.

- B_{k+1} é simétrica
- Obrigando $B_{k+1} s_k = y_k$:

$$(B_k + uu^t) s_k = y_k \Rightarrow uu^t s_k = y_k - B_k s_k \qquad (u)$$

$\Rightarrow \quad u = \dfrac{1}{u^t s_k}(y_k - B_k s_k)$. Daí

$u u^t = \dfrac{1}{(u^t s_k)^2}(y_k - B_k s_k)(y - B_k s_k)^t$. Por outro lado,

$(1) \Rightarrow (u^t s_k)^2 = (y_k - B_k s_k)^t s_k$. Assim

$$\boxed{B_{k+1} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^t}{(y_k - B_k s_k)^t s_k}}$$

* <u>não</u> é garantido que B_{k+1} seja definida positiva.

* B_{k+1} está bem definido somente se
$(y_k - B_k s_k)^t s_k \neq 0$.

---

2) BFGS (Broyden-Fletcher-Goldfarb-Shanno)

B_k simétrica e definida positiva,

$$B_{k+1} = B_k + \underbrace{\alpha u u^t + \beta v v^t}_{\text{correção de posto 2.}}, \quad u = y_k, \; v = B_k s_k$$

- B_{k+1} é simétrica

- $B_{k+1} s_k = y_k$

$$\Rightarrow \quad B_k s_k + \alpha y_k (y_k^t s_k) + \beta B_k s_k (s_k^t B_k s_k) = y_k.$$

Uma solução é tomar $\alpha = \dfrac{1}{y_k^t s_k}$ e $\beta = -\dfrac{1}{s_k^t B_k s_k}$,

o que fornece

$$\boxed{B_{k+1}^{BFGS} = B_k + \frac{y_k y_k^t}{y_k^t s_k} - \frac{B_k s_k s_k^t B_k}{s_k^t B_k s_k}}$$

Teorema: Seja B_k simétrica e definida positiva. Se $y_k^t s_k > 0$ então

(i) B_{k+1}^{BFGS} está bem definida;

(ii) B_{k+1}^{BFGS} é simétrica;

(iii) B_{k+1}^{BFGS} é definida positiva.

Prova: Suponha $y_k^t s_k > 0$. B_{k+1}^{BFGS} está bem definida pois $s_k^t B_k s_k > 0$ ($s_k \neq 0$) e é claramente simétrica. Basta mostrar (iii).

Considere $x \in \mathbb{R}^n$ qualquer. Temos

$$x^t B_{k+1}^{BFGS} x = x^t B_k x + \frac{(x^t y_k)^2}{y_k^t s_k} - \frac{(x^t B_k s_k)^2}{s_k^t B_k s_k}$$

$$(*) \qquad = \frac{(x^t y_k)^2}{y_k^t s_k} + \frac{(s_k^t B_k s_k)(x^t B_k x) - (x^t B_k s_k)^2}{s_k^t B_k s_k}$$

Como B_k é definida positiva, possui fatoração Cholesky, digamos, $B_k = GG^t$, G triangular inferior com diagonal positiva. Assim

- $s_k^t B_k s_k = (s_k^t G)(G^t s_k) = \|G^t s_k\|^2$,
- $x_k^t B_k x = \|G^t x\|^2$ e
- $x^t B_k s_k = (x^t G)(G^t s_k) = (G^t x)^t (G^t s_k)$.

Pela desigualdade de Cauchy-Schwartz obtemos

$$(G^t x)^t (G^t s_k) \leq \|G^t x\|^2 \cdot \|G^t s_k\|^2,$$

ou seja,

$$(s_k^t B_k s_k)(x^t B_k x) - (x^t B_k s_k) \geq 0$$

$\Longrightarrow x^t B_{k+1}^{BFGS} x \geq 0$ (B_{k+1}^{BFGS} é semi-def. positiva)

(*)

Agora, $x^t B_{k+1}^{BFGS} x = 0 \underset{(*)}{\Rightarrow} x^t y_k = 0$ e $(s_k^t B_k s_k)(x^t B_k x) - (x^t B_k s_k) = 0$. Sabemos de Cauchy-Schwartz que a última igualdade ocorre somente se $G^t x$ e $G^t s_k$ forem colineares.

Afirmamos que $G^t x = 0$. De fato, se $G^t x \neq 0$ então $G^t x = \mu G^t s_k$ para algum $\mu \in \mathbb{R}$

$\underset{G \text{ inversível}}{\Longrightarrow} x = \mu s_k \Rightarrow 0 = y_k^t x = \mu \underbrace{y_k^t s_k}_{>0} \Rightarrow \mu = 0.$

Logo $G^t x = \mu G^t s_k = 0 \Rightarrow x = G^{-t} G^t x = 0$.

Concluímos que $x^t B_{k+1}^{BFGS} x > 0 \;, \forall x \neq 0$

(B_{k+1}^{BFGS} é definida positiva) $\blacksquare$

3) DFP (Davidon–Fletcher–Powell)

B_k simétrica e definida positiva

$$B_{k+1}^{DFP} = (I - \alpha y_k s_k^t) B_k (I - \alpha y_k s_k^t) + \alpha y_k y_k^t \,,$$

$\alpha = \frac{1}{y_k^t s_k} > 0$. (ver lista de exercícios)

Observações:

1) BFGS é considerado o melhor método.

2) a convergência de BFGS e outros métodos é similar à Newton (apenas local), e possue velocidade superlinear (melhor que gradiente, pior que Newton puro)

3) BFGS pode ser globalizado como Newton,

colocando-o no esquema geral de descida

4) Um porém: a iteração

$$x^{k+1} = x^k - t_k \left(B_{k+1}^{BFGS}\right)^{-1} \nabla f(x^k)$$

com passo $t_k > 0$ apenas satisfazendo Armijo pode resultar em $y_k^t s_k \leq 0$. Neste caso a próxima matriz BFGS pode não ser definida positiva

Possíveis soluções:

(i) quando $y_k^t s_k \leq 0$, tomar $B_{k+1}^{BFGS} = B_k^{BFGS}$.

(ii) usar uma busca linear para t_k mais refinada. É comum trabalhar com condições de Wolfe (mais exigente que Armijo). Há algoritmos sofisticados para isso.

5) B_0 deve ser iniciada de maneira barata. Geralmente, $B_0 = I$ ou $= \lambda I$, $\lambda > 0$ (no início, a direção é $-\nabla f(x^k)$).

6)

6) O cálculo da direção pode ser feito como Newton, ou seja, resolvendo o sistema

$$B_{k+1}\, d = -\nabla f(x^k).$$

Pode ser interessante trabalhar diretamente com a inversa,

$$d = -B_{k+1}^{-1}\, \nabla f(x^k).$$

<u>Não devemos inverter</u> B_{k+1} numericamente. Para isso, utilizamos a fórmula de Sherman-Morrison (veja lista de exercícios).

7) É possível obter as inversas

$$H_{k+1} = B_{k+1}^{-1}$$

diretamente da equação secante:

$$B_{k+1} s_k = y_k \iff H_{k+1} y_k = s_k .$$

O livro de Ana Friedlander trabalha diretamente com H_{k+1} para BFGS e DFP. (compare B_{k+1}^{BFGS} com H_{k+1}^{DFP} - veja lista de exercícios.)