Método do gradiente espectral

Quase-Newton: $x^{k+1} = x^k - B_k^{-1} \nabla f(x^k)$,

$B_{k+1} s_k = y_k$ (equação secante), onde

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

Escolha simples: $B_{k+1} = \sigma_{k+1} I$, $\sigma_{k+1} > 0$.

Esta escolha não satisfaz a equação secante em geral, mas podemos procurar o σ_{k+1} que melhor aproxima a equação secante "$(\sigma I)s_k = y_k$" no seguinte sentido:

$$\min_{\sigma} \frac{1}{2}\|\sigma s_k - y_k\|^2.$$

Este problema é convexo e pode ser resolvido anulando a derivada da função objetivo:

$$0 = \frac{d}{d\sigma}\left(\frac{1}{2}\|\sigma s_k - y_k\|^2\right) = (\sigma s_k - y_k)^t s_k$$

$$\Rightarrow \boxed{\sigma_{k+1} = \frac{s_k^t y_k}{s_k^t s_k}} \quad \text{(sempre que } s_k \neq 0\text{)}$$

Sendo $B_{k+1} = \sigma_{k+1} I$, temos $B_{k+1}^{-1} = \frac{1}{\sigma_{k+1}} I$.

Agregando busca linear, temos a iteração

$$\boxed{x^{k+1} = x^k - t_k \lambda_k \nabla f(x^k)}, \quad \lambda_k = \frac{1}{\sigma_k} = \frac{s_{k-1}^t s_{k-1}}{s_{k-1}^t y_{k-1}}$$

(desde que $s_{k-1}^t y_{k-1} > 0$),

com passo $t_k > 0$. Esta é a iteração do método do <u>gradiente espectral</u>.

Observe que se trata de uma espécie de "gradiente escalado", com direção $d^k = -\lambda_k \nabla f(x^k)$.

Sobre o cálculo do passo espectral λ_k

1) Assim como nos métodos quase-Newton, o primeiro passo λ_0 deve ser fornecido (lembre-se que $B_0 = \frac{1}{\lambda_0} I$ é dada pelo usuário).

Uma escolha que funciona bem é

$$\lambda_0 = \min\left\{\lambda_{máx},\ máx\left\{\lambda_{min},\ \frac{1}{\|\nabla f(x^0)\|_\infty}\right\}\right\},$$

onde $0 < \lambda_{min} < \lambda_{máx}$ são parâmetros.

Note que se $\nabla f(x^0) = 0$, nada mais é necessário pois x^0 já é solução.

λ_{min}, $\lambda_{máx}$ são limites para λ que visam manter certa estabilidade numérica. A expressão anterior é a projeção de $\frac{1}{\|\nabla f(x^0)\|_\infty}$ sobre o intervalo $[\lambda_{min}, \lambda_{máx}]$.

Na prática, $\lambda_{min} = 10^{-30}$ e $\lambda_{máx} = 10^{30}$.

$\lambda \in [\lambda_{min}, \lambda_{máx}]$ serve ainda para convergência teórica. (não vamos fazer aqui).

2) Se $s_{k-1}^t y_{k-1} \leq 0$, $\frac{1}{\sigma_k} = \frac{s_{k-1}^t s_{k-1}}{s_{k-1}^t y_{k-1}}$ será negativo ou mesmo não definido. Neste caso tomamos

$$\lambda_k = \lambda_{máx}.$$

Ou seja, tenta-se "andar ao máximo" na direção $-\nabla f(x^k)$, deixando todo ajuste para a busca linear.

Sobre o cálculo de t_k (busca linear)

O método

$$x^{k+1} = x^k - t_k \lambda_k \nabla f(x^k) \quad , \quad \lambda_k \in [\lambda_{min}, \lambda_{máx}]$$

converge globalmente se $t_k > 0$ é calculado como no esquema geral de descida, isto é,

"Armijo + interpolação quadrática + backtracking + salvaguardas".

Porém, se $t_k < 1$ estaremos descartando o passo espectral λ_k, que contém informação valiosa da equação secante. 8

Em outras palavras, gostaríamos que $t_k = 1$ com frequência, <u>mesmo que</u> eventualmente f aumente. Ou seja, queremos "passos verdadeiros" do método espectral

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

Para tanto, $t_k > 0$ deve satisfazer a condição de Armijo "relaxada"

$$f(x^k - t\lambda_k \nabla f(x^k)) \leq f_{máx} - t\eta\lambda_k \|\nabla f(x^k)\|^2 \quad (*),$$

onde

$$f_{máx} = máx\{f(x^k), f(x^{k-1}), \ldots, f(x^{k-m})\},$$

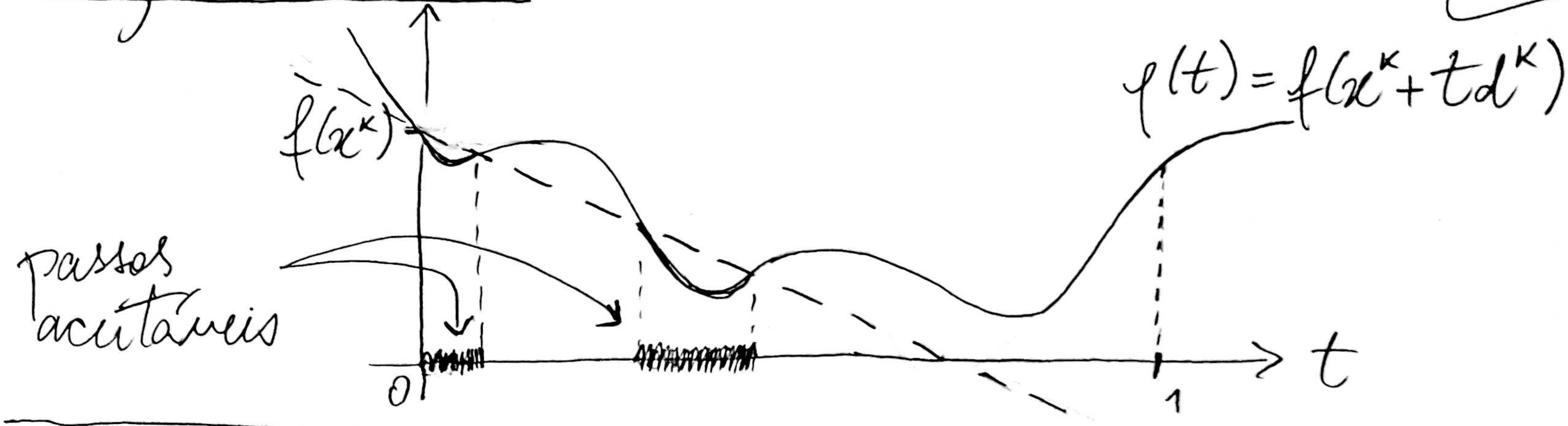
$m > 0$ parâmetro (na prática, $m = 100$).

A busca linear com $(*)$ é "não monótona" pois permite f crescer eventualmente.

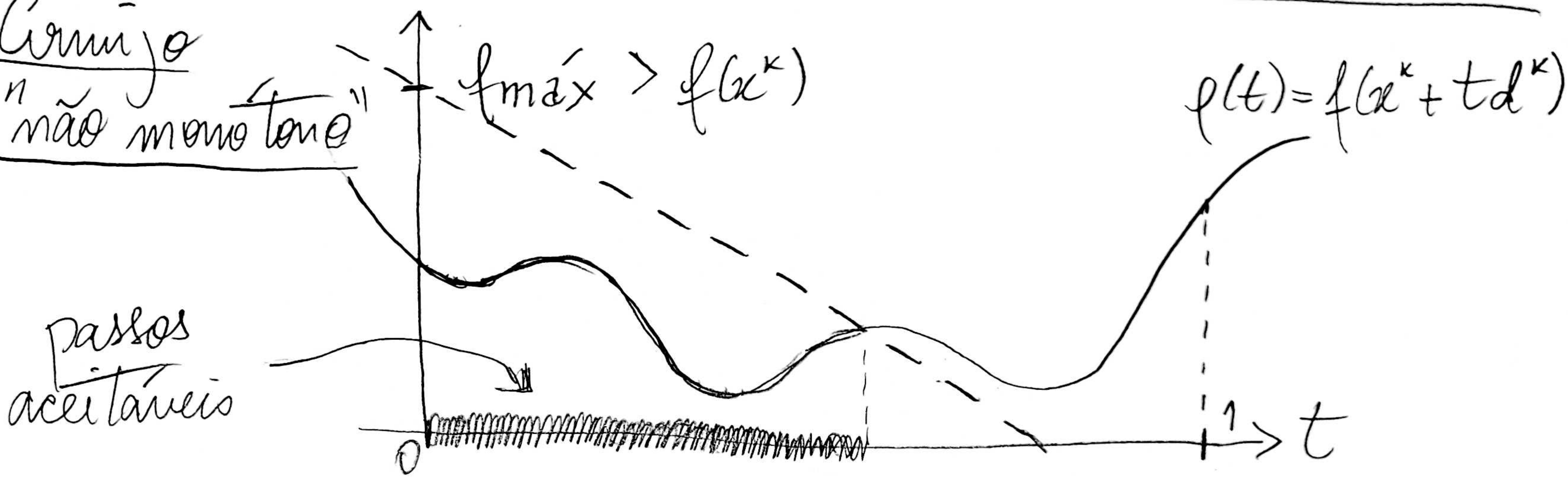
Observe que (*) é Armijo com a direção $d^k = -\lambda_k \nabla f(x^k)$ trocando $f(x^k)$ por $f_{máx}$ (verifique). Claro que "Armijo tradicional $\Rightarrow$ (*)". ((*) é mais frouxa)

Em implementações, armazenamos o vetor das M últimas f's, que é iniciado com todas as entradas $= -\infty$.

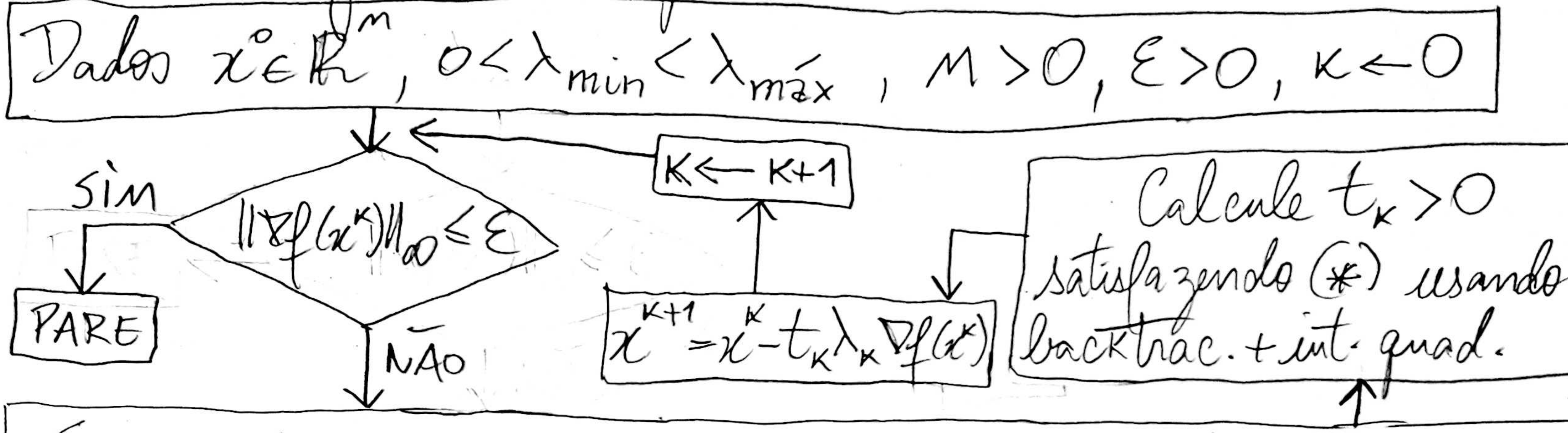
Armijo tradicional



Armijo "não monótono"



Método do gradiente espectral



Se $k=0$, $\lambda_0 = \min\{\lambda_{máx}, \max\{\lambda_{min}, 1/\|\nabla f(x^0)\|_\infty\}\}$

Se $k \geq 1$,

$$\lambda_k = \begin{cases} \min\left\{\lambda_{máx}, \max\left\{\lambda_{min}, \dfrac{s_{k-1}^t s_{k-1}}{s_{k-1}^t y_{k-1}}\right\}\right\} & \text{caso } s_{k-1}^t y_{k-1} > 0 \\ \lambda_{máx} & \text{caso contrário} \end{cases}$$

Observações:

1) o método do gradiente espectral é barato (= método gradiente) e logo é adequado à problemas com muitas variáveis.

2) numericamente é muito bom para minimizar funções não convexas quaisquer.

3) Mesmo com Armijo modificado (*), a convergência é global.

4) O nome "espectral" vem do fato de 13

$\sigma_{k+1} = \dfrac{s_k^t y_k}{s_k^t s_k}$ ser um quociente de Rayleigh (que é usado para descrever autovalores) da seguinte "hessiana média":

$$\tilde{B} = \int_0^1 \nabla^2 f(x^k + t s_k)\, dt.$$

De fato, defina

$$g(t) = \nabla f(x^k + t s_k), \qquad t \in [0,1].$$

Então

$$\int_0^1 g'(t)\,dt = g(1) - g(0)$$

$$\Rightarrow \int_0^1 \nabla^2 f(x^k + t s_k)\, s_k\, dt = \underbrace{\nabla f(x^k + s_k) - \nabla f(x^k)}_{= y_k}$$

$$\Rightarrow \left[\int_0^1 \nabla^2 f(x^k + t s_k)\, dt\right] s_k = y_k \,,$$

ou seja, $\tilde{B} s_k = y_k$. Logo, temos o quociente de Rayleigh

$$\frac{s_k^t \tilde{B} s_k}{s_k^t s_k} = \frac{s_k^t y_k}{s_k^t s_k} = \sigma_{k+1} \,.$$