

Gradiente incremental

(1)

$$\min_{x \in \mathbb{R}^m} f(x) = \sum_{j=1}^m f_j(x).$$

- No nosso caso, $m \gg 1$ e o cálculo de ∇f fica caro.
- Ao invés de avaliar ∇f (e consequentemente $\nabla f_j, V_j$), avaliamos ∇f_j para algum j por iterações (muito mais barato!)
- Vamos intuir o método do gradiente usando essa ideia de "gradientes incrementais".

• A cada iteração, escolhemos um $i \in \mathbb{Z}$
damos um passo na direção $-\nabla f_{i+1}(x^k, i)$,
onde $x^{k,i}$ é o ponto corrente.

↳ i pode ser escolhido aleatoriamente
em $\{0, \dots, m-1\}$ desde que a escolha
seja uniforme.

↳ podemos escolher um subconjunto
pequeno $I \subset \{1, \dots, m\}$ e iterar
na direção $-\sum_{i \in I} \nabla f_i(x^k)$ (descenso
por blocos)

↳ para simplificar, vamos escolher índice 3
por índice i (gradiente incremental)
em ordem crescente 0, 1, ..., m-1.

Exemplo: $f(x_1, x_2) = \underbrace{(x_1^2 + 2x_2)}_{f_1(x_1, x_2)} + \underbrace{(2x_2^2)}_{f_2(x_1, x_2)}$

Ponto inicial: $x^0 = (2, 2)$, $k=0$

Passo : $t = 1/2$

Iteração gradiente tradicional:

(4)

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

$$\nabla f(x) = \begin{bmatrix} 2x_1 \\ 4x_2 + 2 \end{bmatrix}$$

$$x^1 = (2, 2) - \frac{1}{2}(4, 10) = (0, -3)$$

$$f(x^{k+1}) = 12$$

Iteração gradiente incremental:

$$f(x) = f_1(x) + f_2(x) , \quad \nabla f_1 = \begin{bmatrix} 2x_1 \\ 2 \end{bmatrix} , \quad \nabla f_2 = \begin{bmatrix} 0 \\ 4x_2 \end{bmatrix}$$

15

$$\lambda = \frac{1}{2}$$

$$x^{k,0} = x^0 = (2, 2)$$

$$\overbrace{x^{k,1}} = x^{k,0} - t \nabla f_1(x^{k,0}) \quad (i=0)$$

$$= (2, 2) - \frac{1}{2}(4, 2) = (0, 1)$$

$$\overbrace{x^{k,2}} = x^{k,1} - t \nabla f_2(x^{k,1}) \quad (i=1)$$

$$= (0, 1) - \frac{1}{2}(0, 4) = (0, -1)$$

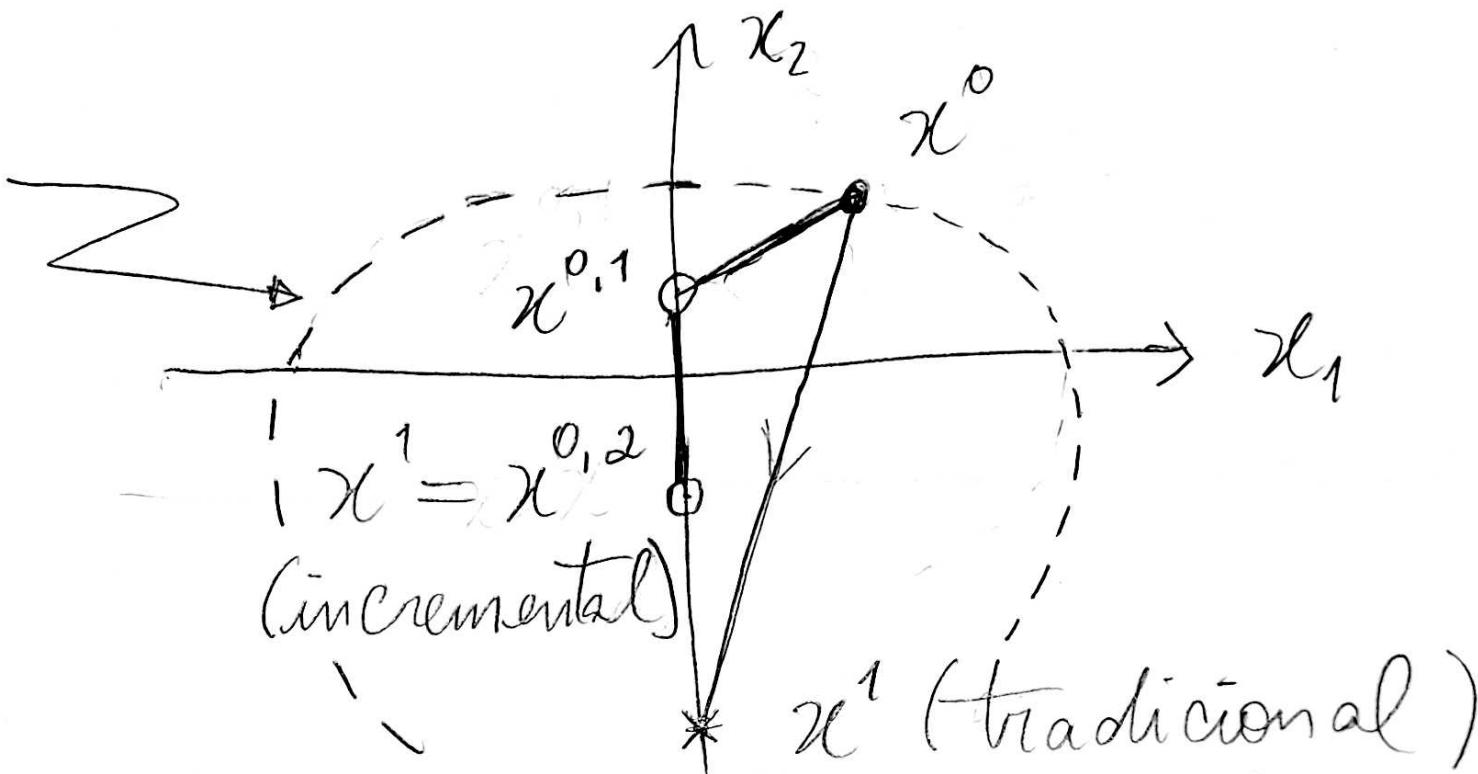
$$x^{k+1} = x^{k,2} = (0, -1).$$

$$f(x^{k+1}) = f(0, -1) = 0$$

[6]

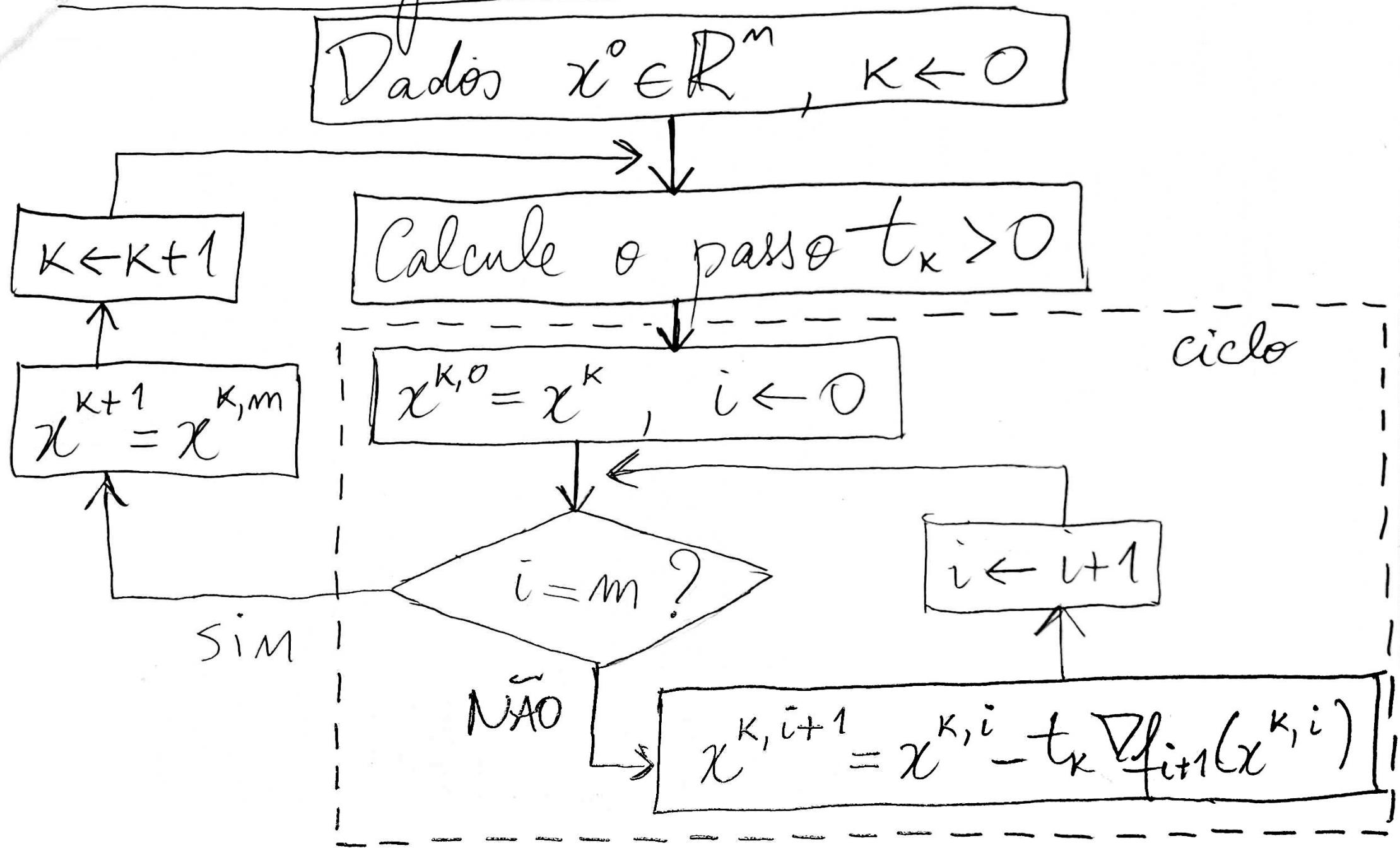
- As iterações incremental reduzem mais f e com custo igual (cálculo de ∇f_1 e ∇f_2)

minim
de f



Método do gradiente incremental

17



- O laço que percorre todos os gradientes $\nabla f_{i+1}, i=0, \dots, m-1$, é chamado de ciclo. 18
- Note que $x^{k,i+1}$ é o resultado de um passo a partir de $x^{k,i}$. Essa é a diferença com o gradiente tradicional.
↳ visa acelerar a convergência a um custo não maior.
- Note que t_k é o mesmo p/ cada ciclo.

- versões com i escolhido por blocos seguem o mesmo esquema.
- o critério de parada tipo " $\nabla f(x^k) \approx 0$ " é mais complicado, pois não calculamos ∇f inteiro. Na prática (em particular no aprendizado de máquina), paramos com número fixo de iterações externas.

- Sobre da solução, gradiente incremental ∇f costuma ser muito efetivo, muito mais que gradiente tradicional!
- Porém, perto da solução não! Isso porque próximo à solução, os passos devem ser mais refinados, e logo usar ∇f é mais efetivo.
- Assim, gradiente incremental é útil quando não precisamos de alta precisão (caso do aprendizado de máquina).

Convergência do método do gradiente

11

incremental para funções convexas

- f_j convexa, $j=1, \dots, m$.

Neste caso, temos que

$$f_j(y) \geq f_j(x) + \nabla f_j(x)^T (y - x), \forall y, x.$$

- $\{x^k\}$ sequência gerada pelo método.

Hipótese: H1) $\exists L > 0$ tal que $L \geq \|\nabla f_{j+1}(x^{k,i})\|$

$\forall i, k$ (compara com subgradiente)

Lema: Suponha f_j convexa $\forall j \in H_1$. . . (12)

Para todo $y \in \mathbb{R}^m$ temos

$$\|x^{k+1} - y\|^2 \leq \|x^k - y\|^2 - 2t_k(f(x^k) - f(y)) + t_k^2 m^2 L^2.$$

Prova: Para cada $i = 0, \dots, m-1$, temos

$$\|x^{k,i+1} - y\|^2 = \|x^{k,i} - t_k \nabla f_{i+1}(x^{k,i}) - y\|^2$$

$$= \|x^{k,i} - y\|^2 - 2t_k \nabla f_{i+1}(x^{k,i})^T (x^{k,i} - y) + t_k^2 \|\nabla f_{i+1}(x^{k,i})\|^2$$

$$\leq \|x^{k,i} - y\|^2 - 2t_k (f_{i+1}(x^{k,i}) - f_{i+1}(y)) + t_k^2 L^2$$

f_{i+1} convexa,
 H_1

Somando para $i=0, \dots, m-1$ e cancelando 13 termos de ambos os lados das desigualdades, obtemos (lembrando que $x^{k,0} = x^k$ e $x^{k,m} = x^{k+1}$)

$$\|x^{k+1} - y\|^2 \leq \|x^k - y\|^2 - 2t_k \sum_{i=0}^{m-1} (f_{i+1}(x^{k,i}) - f_{i+1}(y)) + t_k^2 m L^2$$

Como $\sum_{i=0}^{m-1} f_{i+1}(z) = f(z)$, temos

$$\|x^{k+1} - y\|^2 \leq \|x^k - y\|^2 - 2t_k [f(x^k) - f(y) + \sum_{i=0}^{m-1} (f_{i+1}(x^{k,i}) - f_{i+1}(y))]$$

$$-f_{i+1}(x^k)\Big) \Big] + t_k^2 m L^2.$$

[14]

Como f_{i+1} é convexa, temos

$$f_{i+1}(x^{k,i}) - f_{i+1}(x^k) \geq \nabla f_{i+1}(x^k)^t (x^{k,i} - x^k).$$

Dai,

$$\|x^{k+1} - y\|^2 \leq \|x^k - y\|^2 - 2t_k (f(x^k) - f(y))$$

$$+ 2t_k \sum_{i=0}^{m-1} -\nabla f_{i+1}(x^k)^t (x^{k,i} - x^k) + t_k^2 m L^2$$

$$\leq \|\nabla f_{i+1}(x^k)\| \cdot \|x^{k,i} - x^k\|$$

$$\begin{aligned}
 & \leq \|x^k - y\|^2 - 2t_k (f(x^k) - f(y)) \\
 & + 2t_k L \sum_{i=0}^{m-1} \|x^{k,i} - x^k\| + t_k^2 m L^2.
 \end{aligned} \tag{15}$$

Agora, vemos que

- $\|x^{k,0} - x^k\| = \|x^k - x^k\| = 0$
- $\|x^{k,1} - x^k\| = \|x^{k,0} - t_k \nabla f_1(x^{k,0}) - x^k\|$
 $\leq \|x^{k,0} - x^k\| + t_k \|\nabla f_1(x^{k,0})\| \leq t_k L$

$$\begin{aligned} \|\boldsymbol{x}^{k,2} - \boldsymbol{x}^k\| &= \|\boldsymbol{x}^{k,1} - t_k \nabla f_2(\boldsymbol{x}^{k,1}) - \boldsymbol{x}^k\| \\ &\leq \|\boldsymbol{x}^{k,1} - \boldsymbol{x}^k\| + t_k \|\nabla f_2(\boldsymbol{x}^{k,1})\| \leq 2t_k L. \end{aligned} \quad [16]$$

Em geral,

$$\|\boldsymbol{x}^{k,i} - \boldsymbol{x}^k\| \leq i t_k L, \quad i = 0, \dots, m-1$$

Usando estas desigualdades obtemos

$$\begin{aligned} \|\boldsymbol{x}^{k+1} - \boldsymbol{y}\|^2 &\leq \|\boldsymbol{x}^k - \boldsymbol{y}\|^2 - 2t_k (f(\boldsymbol{x}^k) - f(\boldsymbol{y})) \\ &\quad + 2t_k^2 L^2 \sum_{i=0}^{m-1} i + t_k^2 m L^2. \end{aligned}$$

Como $\sum_{i=0}^{m-1} i = \frac{m(m-1)}{2}$, obtemos . (17)

$$\|x^{k+1} - y\|^2 \leq \|x^k - y\|^2 - 2t_k(f(x^k) - f(y)) + t_k^2 m^2 L^2,$$

Como queríamos provar. 

Este resultado é parecido com o obtido para o método do subgradiente (verifique).

Aliás, é possível definir o método do subgradiente incremental (lista exercícios).

Como no método do subgradiente, é comum (18) considerar os passos nos seguintes casos:

1) passo constante: $t_k = t > 0, \forall k$.

2) passo decrescente: $\{t_k\}$ tal que

$$t_k \rightarrow 0^+, \sum_{k=0}^{\infty} t_k^2 < \infty, \sum_{k=0}^{\infty} t_k = \infty.$$

No aprendizado de máquina, ambas as ideias são empregadas. Note que o lema indica que para $t \ll 1$, o método funcionará.

Teorema: Suponha f_j convexa $\forall j$, H1 válida (19) e que $f(x) = \sum_{j=1}^m f_j(x)$ admira minimizadora x^* . Seja $\{t_k\}$ tal que

$$t \rightarrow 0^+, \quad \sum_{k=0}^{\infty} t_k^2 < \infty \quad \text{e} \quad \sum_{k=0}^{\infty} t_k = \infty.$$

Então

$$f_\infty = \liminf_{k \rightarrow \infty} f(x^k) = f^*,$$

onde $f^* = f(x^*) = \min_x f(x) (> -\infty)$.

Prova: Pelo Lema (com $y = x^*$), temos (20)

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - 2t_k(f(x^k) - f^*) + t_k^2 m^2 L^2$$

$$\begin{aligned} &\leq \|x^{k-1} - x^*\|^2 - 2t_{k-1}(f(x^{k-1}) - f^*) + t_{k-1}^2 m^2 L^2 \\ &\quad - 2t_k(f(x^k) - f^*) + t_k^2 m^2 L^2 \end{aligned}$$

$$\leq \dots \leq \|x^0 - x^*\|^2 - 2 \sum_{i=0}^k t_i(f(x^i) - f^*) + m^2 L^2 \sum_{i=0}^k t_i^2$$

$$\Rightarrow \left(2 \sum_{i=0}^k t_i \right) (f_k - f^*) \leq 2 \sum_{i=0}^k t_i (f(x^i) - f^*)$$

$$\leq \|x^0 - x^*\|^2 + m^2 L^2 \sum_{i=0}^K t_i^2, \quad (2.1)$$

onde $f_K = \min_{0 \leq i \leq K} f(x^i)$. Daí,

$$f_K - f^* \leq \frac{\|x^0 - x^*\|^2 + m^2 L^2 \sum_{i=0}^K t_i^2}{2 \sum_{i=0}^K t_i}.$$

O resultado segue das condições sobre $\{t_k\}$
fazendo $K \rightarrow \infty$

Pode-se provar resultado parecido com (22) o método de subgradiêntes para $t_k = t$ constante, usando o lema anterior.
(exercício).