Gradiente incremental

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

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

- A cada iteração, escolhemos um i e
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 $\{0, \dots, m-1\}$ <u>desde que</u> a escolha seja uniforme.

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

↳ para simplificar, vamos escolher índice por índice i (<u>gradiente incremental</u>) em ordem crescente $0, 1, \ldots, m-1$.

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

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

Passo: $t = 1/2$

Iteração gradiente tradicional:

$$x^{k+1} = x^k - t\,\nabla f(x^k) \quad , \quad \nabla f(x) = \begin{bmatrix} 2x_1 \\ 4x_2 + 2 \end{bmatrix}$$

$$= (2,2) - \tfrac{1}{2}(4,10) = (0,-3)$$

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

Iteração gradiente incremental:

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

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

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

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

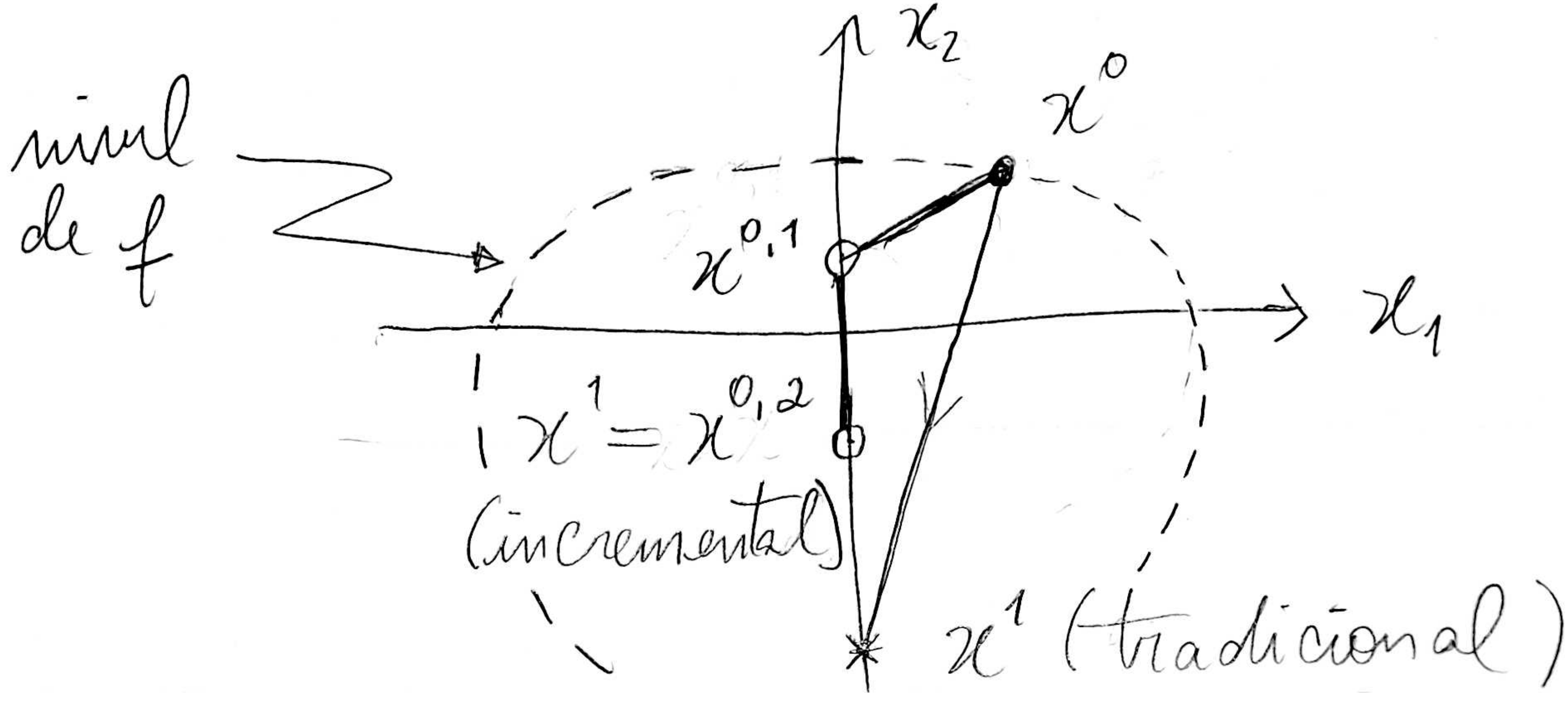
$$x^{k,2} = x^{k,1} - t\nabla f_2(x^{k,1}) \qquad (i=1)$$

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

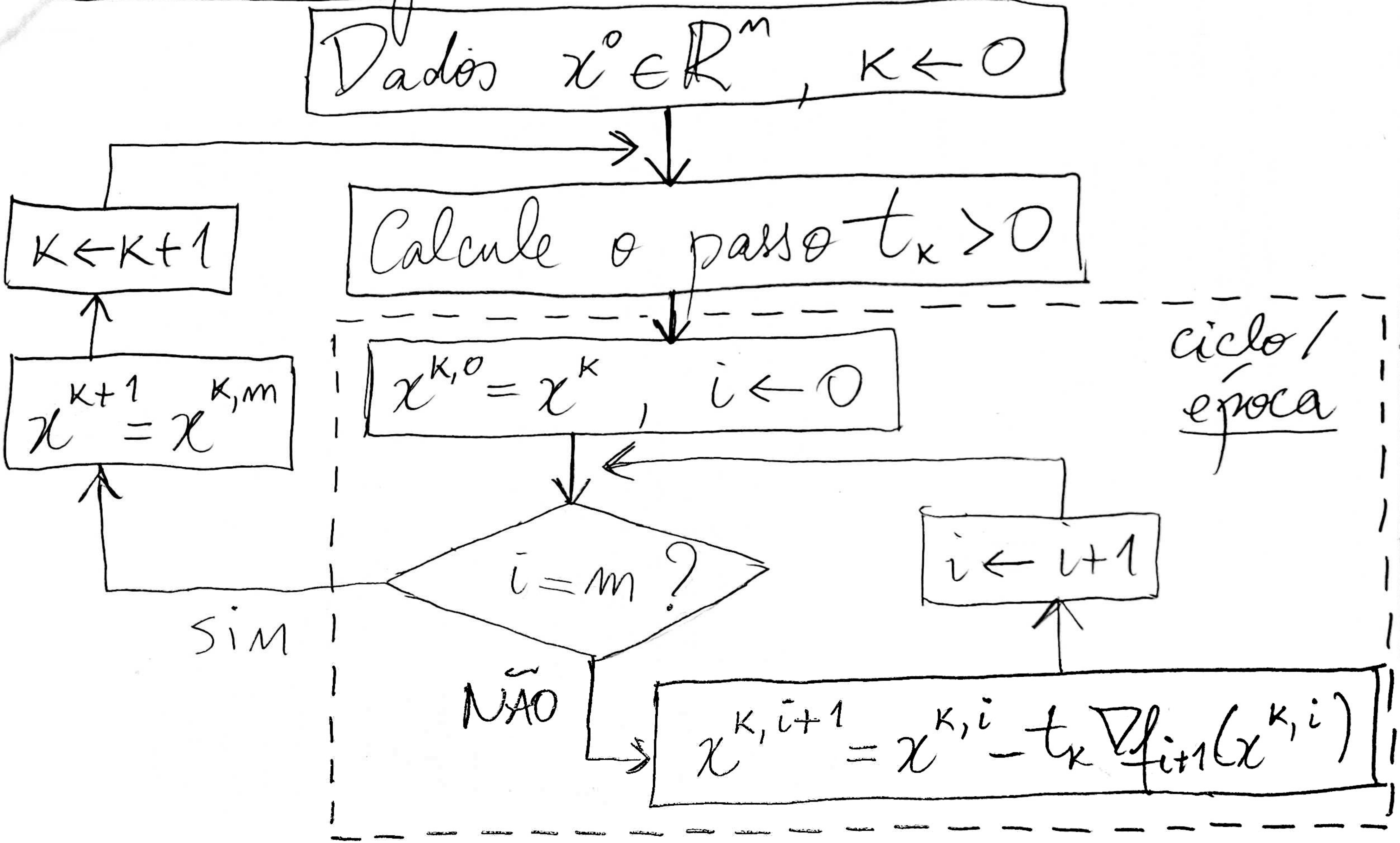
$$x^{k+1} = x^{k,2} = (0,-1).$$

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

- a iteração incremental reduziu mais f e com custo igual (cálculo de ∇f_1 e ∇f_2)



Método do gradiente incremental



- o laço que percorre <u>todos</u> os gradientes ∇f_{i+1}, $i=0,\dots,m-1$, é chamado de <u>ciclo</u> ou <u>época</u>.

- 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 aleatoriamente e 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.

- Longe da solução, gradiente incremental 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 incremental para funções convexas

- f_j convexa, $j=1,\dots,m$.
  Neste caso, lembre-se 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 $L \geq \|\nabla f_{i+1}(x^{k,i})\|$ $\forall i, k$ (compare com subgradiente)

<u>Lema</u>: Suponha f_j convexa $\forall_j$ e H1.

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

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

<u>Prova</u>: 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,
H1

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

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

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

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

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

Como f_{i+1} é convexa, temos

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

Daí,

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

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

$$\leq \|x^k - y\|^2 - 2t_k\left(f(x^k) - f(y)\right)$$

$$+ 2t_k L \sum_{i=0}^{m-1} \|x^{k,i} - x^k\| + t_k^2 m L^2 .$$

Agora, veja que

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

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

- $$\|x^{k,2} - x^k\| = \|x^{k,1} - t_k \nabla f_2(x^{k,1}) - x^k\|$$
$$\leq \|x^{k,1} - x^k\| + t_k \|\nabla f_2(x^{k,1})\| \leq 2 t_k L.$$

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

Usando estas desigualdades ~~de~~ temos
$$\|x^{k+1} - y\|^2 \leq \|x^k - y\|^2 - 2t_k (f(x^k) - f(y))$$
$$+ 2 t_k^2 L^2 \sum_{i=0}^{m-1} i + t_k^2 m L^2.$$

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

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

como queríamos provar. $\blacksquare$

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 considerar os passos nos seguintes casos:

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

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

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

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

<u>Teorema:</u> Suponha f_j convexa $\forall j$, H1 válida e que $f(x) = \sum_{j=1}^{m} f_j(x)$ admita minimizador x^*. Seja $\{t_k\}$ tal que

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

Então

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

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

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

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

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

$$- 2t_k(f(x^k) - f^*) + t_k^2 m^2 L^2$$

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

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

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

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 $k \to \infty$ ∎

Pode-se provar resultado parecido com o método de subgradientes para $t_k = t$ constante, usando o lema anterior.

(exercício)