

# Método do gradiente estocástico

(1)

Problema:  $\min f(x)$ ,  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  convexa

Não temos acesso à  $\nabla f(x)$  (por exemplo, porque é caro computar).

Iteração:  $x^{k+1} = x^k - t_k g^k$ ,  $t_k > 0$

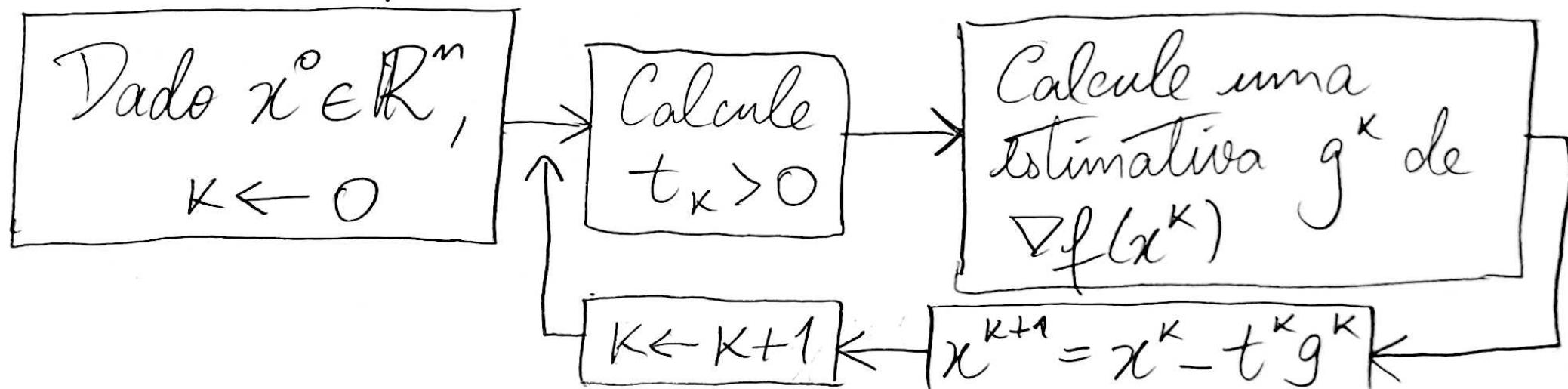
•  $g^k$  é uma estimativa aleatória de  $\nabla f(x^k)$  (passo)

• Assim,  $x^{k+1}$  depende da amostra de pontos.

anteriores (na verdade, está condicionado (2 apenas à escolha de  $x^k$  e  $g^k$ )

- ponto inicial  $x^0$  é determinístico (fornecido pelo usuário);  $x^k$ ,  $k \geq 1$ , aleatórios.

### Método do gradiente estocástico



Como deve ser a escolha de  $g^*$ ? B

$$H1) E(g^* | x^*) = \nabla f(x^*)$$

(a rigor, deveria escrever  $G^*$ , mas vou manter  $g^*$  para refletir a escolha no método)

Essa hipótese diz que, em média,  $g^*$  deve ser o gradiente  $\nabla f(x^*)$ , mas não necessariamente  $g^* = \nabla f(x^*)$ .

H1 é conhecida como "escolha sem viés".

Note que a esperança em H1 é condicionada<sup>14</sup> à escolha  $X = x^*$  (corrente com gradiente clássico, onde  $\hat{g}^* = \nabla f(x^*)$  só depende de  $x^*$ ).

Hipótese Técnica:

H2) Existe uma constante  $L > 0$  tal que

$$E(\|g^*\|^2 | x^*) \leq L^2.$$

(compare com os métodos do gradiente incremental e do subgradiente).

Exemplo:  $f(x) = \frac{1}{m} \sum_{j=1}^m f_j(x)$ ,  $f_j$  convexa [5]

Considere o seguinte estimador para  $g^*$ ,  
dado  $x^*$ :

"escolha  $i_k \in \{1, \dots, m\}$  uniformemente  
e tome  $g^* = \nabla f_{i_k}(x^*)$ ".

Qual a esperança de  $g^*$  dado  $x^*$ ?

$$E(g^k | x^k) = \sum_{i=1}^m \nabla f_i(x^k) \cdot P(i) \quad (6)$$

prob. escolha  
 de  $i$

$$= \sum_{i=1}^m \nabla f_i(x^k) \frac{1}{m} = \frac{1}{m} \sum_{i=1}^m \nabla f_i(x^k)$$

$$= \nabla f(x^k).$$

Note que é importante que a escolha de  $i_k$  seja uniforme, de modo que  $P(i) = \frac{1}{m}, \forall i.$

Este exemplo indica uma escolha possível, (7)  
simples e sem viés, para nosso problema  
de interesse!

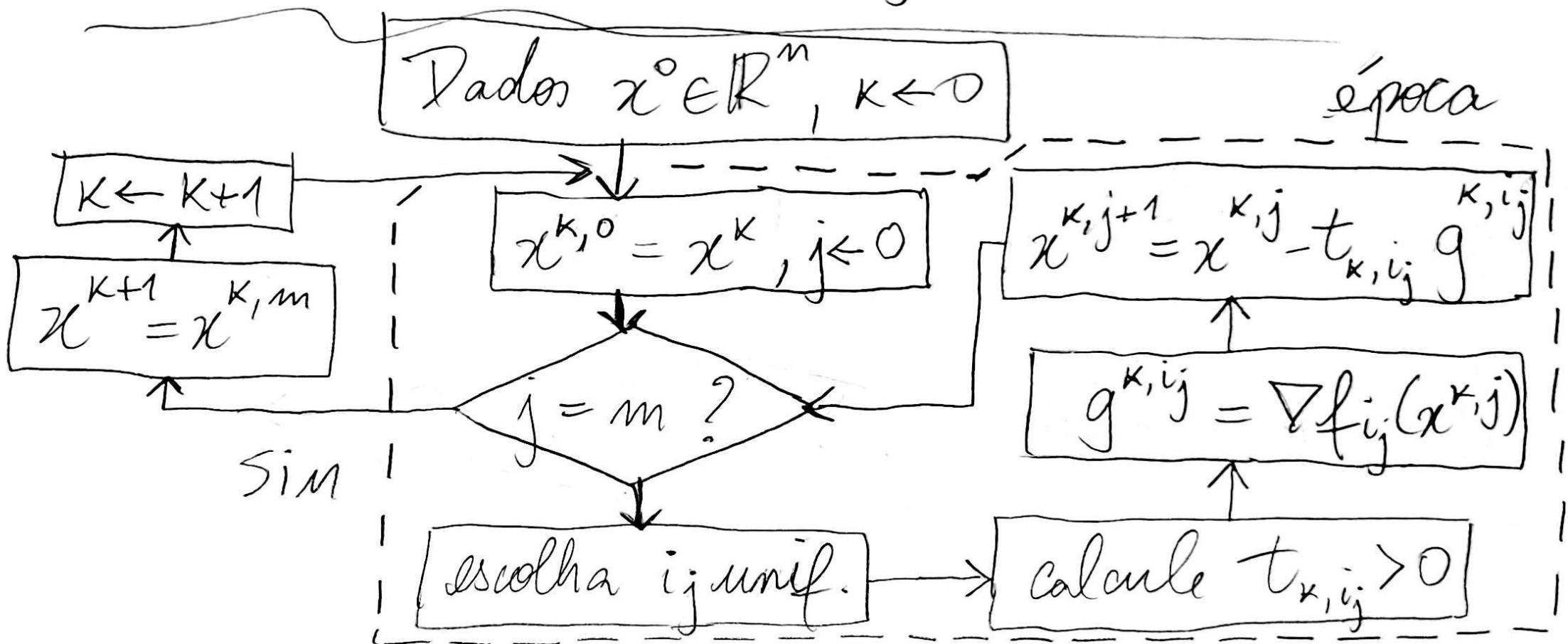
Com esta escolha, o método do gradiente  
estocástico toma-se um "gradiente incre-  
mental com  $i$  aleatório" (com  $t_k$  diferente...)

Um ciclo de  $m$  escolhas de  $i$  é chamado  
de época (note que aqui não é obrigatória  
a escolha dos  $m$  gradientes  $\nabla f_i(x^*)$  como

no ciclo do gradiente incremental) (8)

Pesta forma podemos dividir as iterações

no nosso caso  $f(x) = \frac{1}{m} \sum_j f_j(x)$  :



Obs:

19

- 1) note que há um passo para cada subiteração.
- 2) se  $t_{k,ij} = t$  cte  $\forall x_{ij}$ , então recainos no gradiente incremental, com a diferença que  $i_j$  é aleatório (aqui,  $t$  é determinístico).
- 3) não há garantia que todos  $Df_j(x^*)$ ,  $j=1, \dots, m$  sejam escolhidos a cada época.

4) Geralmente, é assim que as implementações são feitas ...

Outra forma de estimar  $\hat{g}^k$  é fazer

$$\hat{g}^k = \frac{1}{|S_k|} \sum_{j \in S_k} \nabla f_j(x^k),$$

onde  $S_k \subset \{1, \dots, m\}$ ,  $|S_k| \ll m$ .  $S_k$ : "mini-batch" ("mini-lote")

É importante que o mini-lote seja escolhido de maneira uniforme.

Uma maneira de proceder:

(11)

- 1) Temos  $|S| \ll m$  o tamanho dos mini-lotes  
(vamos supor que todos eles têm mesmo tamanho)
  - 2) Divida os dados de treinamento em  $m_B$  mini-lotes  $S_1, \dots, S_{m_B}$  (supor  $m_B = \frac{m}{|S|} \in \mathbb{N}$ ).
  - 3) Escolha um mini-lote  $S_{i_k}$  onde  
 $i_k$  é escolhido uniformemente em  $1, \dots, m_B$ .
- Note que  $E(g^k | x^k) = \sum_{i=1}^{m_B} \left( \frac{1}{|S|} \sum_{j \in S_i} \nabla f_j(x^k) \right) P(i)$

$$P(i) = \frac{1}{m_B} = \sum_{i=1}^{m_B} \left( \frac{m_B}{m} \sum_{j \in S_i} \nabla f_j(x^*) \right) \frac{1}{m_B} \quad (12)$$

$$= \frac{m_B}{m} \cdot \frac{1}{m_B} \sum_{j=1}^m \nabla f_j(x^*) = \nabla f(x^*)$$

(não há viés).

Note que aqui a época consiste em  $m_B$  escolhas de mini-lotes.

Uma outra maneira de implementar é a <sup>(13)</sup> seguinte (usada na prática):

- 1) temos  $151 \ll m$  o tamanho dos mini-lotes  
(vamos supor que todos eles têm mesmo tamanho);
- 2) a cada início de época, embaralhe os dados de treinamento;
- 3) divida os dados embaralhados em pedaços consecutivos de tamanho 151;

4) percorra sequencialmente todos os mini-lotes gerados. [14]

Isto evita viés ??? ...

De qualquer forma, valem as observações:

1)  $|S| = m$  (1 lote) corresponde ao método de gradiente, pois calculamos  $Df(x^k)$  inteiro por iteração;

- 2)  $|S| = 1$  corresponde ao método do gradiente estocástico enunciado anteriormente (15)
- 3) É imperativo  $|S| \ll m$  (iteração barata), mas pode ser interessante  $|S| > 1$  pois usar  $> 1$  gradiente  $\nabla f_j$  fornece uma iteração mais eficaz.
- ↳ o balanço de  $|S|$  é empírico!

Escolha dos passos (variantes do SG). [16]

- Método do gradiente estocástico "básico"  
(SGD)

$$t_{k,ij} = \eta > 0 \quad \text{ete.}$$

≠ da ideia  
original de  
SG (1951)

- Adagrad (2011)

$$t_{k,ij} = \frac{\eta}{\sqrt{G_{k,ij} + \epsilon}}, \quad \eta > 0, \quad \epsilon \in (0,1) \text{ parâmetros,}$$

$$G_{k,ij} = \sum_{\tilde{k} \leq k, \tilde{i}, \tilde{j}} \|\tilde{g}_{\tilde{k},\tilde{i}j}\|^2.$$

Ideia: diminuir  $t$  à medida que o (17)  
método avança, inversamente proporcio-  
nal ao acúmulo das normas dos  
gradientes escolhidos.

(Lembra a ideia de passo decrescente  
nos métodos do gradiente incremental  
e de subgradientes).

- Passos com "momento".

18

Imagine que um corpo viaje de  $x^{k-1}$  até  $x^k$  com uma velocidade  $v^k$ . No ponto  $x^k$ , o corpo não fará uma "virada busca" pois há uma tendência a permanecer na direção  $x^k - x^{k-1}$  (momento linear = massa  $\times v$ )

Por outro lado, a iteração  $x^{k+1} = x^k - t g^k$  representa uma "virada busca" na direção  $-g^k$ .

Métodos com momento buscam incorporar a ideia física na direção de busca.

A ideia é trocar  $g^*$  por um "vetor velocidade", que combina  $g^*$  com a velocidade anterior:

$$x^{k+1} = x^k - t_k v^k \quad \text{onde}$$

$$v^k = \beta v^{k-1} + (1-\beta) g^k \quad (\beta \in [0,1] \text{ parâmetro})$$

Iniciamos a velocidade nula  $v^{-1} = 0$ .

## • SGD com momento (1999)

(20)

$$t_{k,i,j} = \eta > 0 \text{ é } t_e,$$

$$\chi^{k,j+1} = \chi^{k,j} - \eta v^{k,j} \text{ onde}$$

$$v^{k,j} = \beta v^{k,j-1} + (1-\beta) g^{k,i,j}$$

,  $\beta \in [0,1]$   
parâmetro

## • RMSProp (2012?)

Melhoramento do Adagrad.

- AdaDelta (2012)

L21

Outro melhoramento do Adagrad.

---

- Adam (2014/15)

Combina passo tipo Adagrad e momento

Obs: todas variantes podem implementar mini-lotes.