

## Risco empírico x risco esperado

C

Vimos anteriormente que, dados pares entrada-saída  $(x_i, y_i)$ ,  $i = 1, \dots, m$ , buscamos

$$\min_{w, b} R_m(w, b) = \frac{1}{m} \sum_{i=1}^m l(h(x_i; w, b); y_i)$$

onde  $l$  mede o erro na classificação do dado  $x_i$  ( $h(x_i; w, b) \neq y_i$ ).

( $R_m$ : risco empírico)

Mas isso é realmente o que queremos? 12

↳ na soma estão todas as possíveis entradas  $x$ ? Claro que não! Estamos querendo treinar uma rede neural justamente para responder a dados  $x$  desconhecidos ...

↳  $x$  pode assumir coordenadas em conjuntos muito grandes, não enumeráveis até ...

↳ idealmente, podemos pensar " $x$  contínuo".

Podemos pensar em termos de polaridade: cada lado (ou par  $(x, y)$ ) "continuo" tem uma chance de ocorrer. Queremos que o computador responda corretamente "na média" dos  $x$ 's. Assim, "x" mais prováveis devem ter mais peso que os menos prováveis.

Isso nos leva à querer minimizar o valor esperado de má classificação.

↳ em outras palavras, queremos que L4  
em média  $\hat{l}$  se aproxime do menor  
valor possível (p. ex., zero se  $l(z, y) = (y - z)^2$ )

Esperança ou valor esperado

Notação:

- $X$  — variável aleatória.
- $x$  — amostra de  $X$ .

Exemplo: Suponha que  $X$  seja variável aleatória discreta com a seguinte distribuição de probabilidade: 5

$x$	0	100	200	300	400
$P(X=x)$	0,2	0,1	0,4	0,2	0,1

→ probabilidade de  $X=x$ .

Pergunta: qual o valor médio de  $X$  em uma amostra de tamanho  $N$  ?

Resposta: ao amostrar  $\times N$  vezes, 20% delas serão 0, 10% serão 100 etc.

A média será

$$\frac{1}{N} (0 \cdot 0,2N + 100 \cdot 0,1N + 200 \cdot 0,4N + 300 \cdot 0,2N + 400 \cdot 0,1N) = 190.$$

Assim, "a média a longo prazo" é 190 //

Para variáveis aleatórias discretas, definimos a esperança (ou valor esperado ou valor médio)

de  $X$  como sendo

(7)

$$E(X) = \sum x \cdot P(x).$$

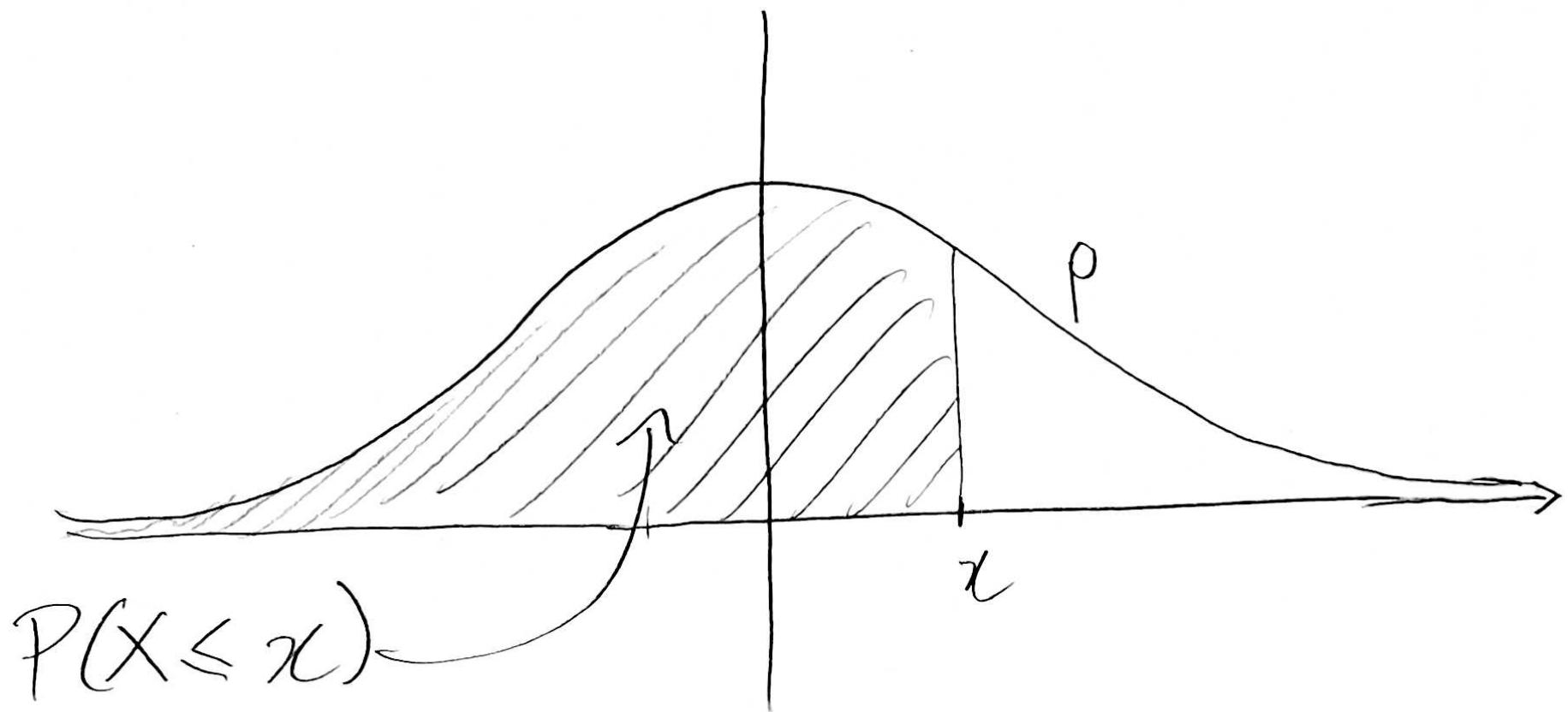
E se  $X$  for variável aleatória contínua?

Se  $p$  for a função densidade de probabilidade de  $X$ , então

$$E(X) = \int_{-\infty}^{\infty} x p(x) dx = \int_R x p(x) dx.$$

Sembra -se que  $\int_{\mathbb{R}} p(x) dx = 1$ ,  $p(x) \geq 0$  e

$$P(X \leq x) = \int_{-\infty}^x p(x) dx.$$



Se  $\ell: \mathbb{R} \rightarrow \mathbb{R}$  é uma função, podemos (a)  
falar em valor esperado da variável  
aleatória  $Y = \ell(X)$ :

$$E(Y) = E(\ell(X)) = \int_{\mathbb{R}} \ell(x) p(x) dx.$$

Agora, suponha que  $X$  (e  $x$ ) sejam  
vetores de  $n$  coordenadas, e

$$\ell: \mathbb{R}^n \rightarrow \mathbb{R}.$$

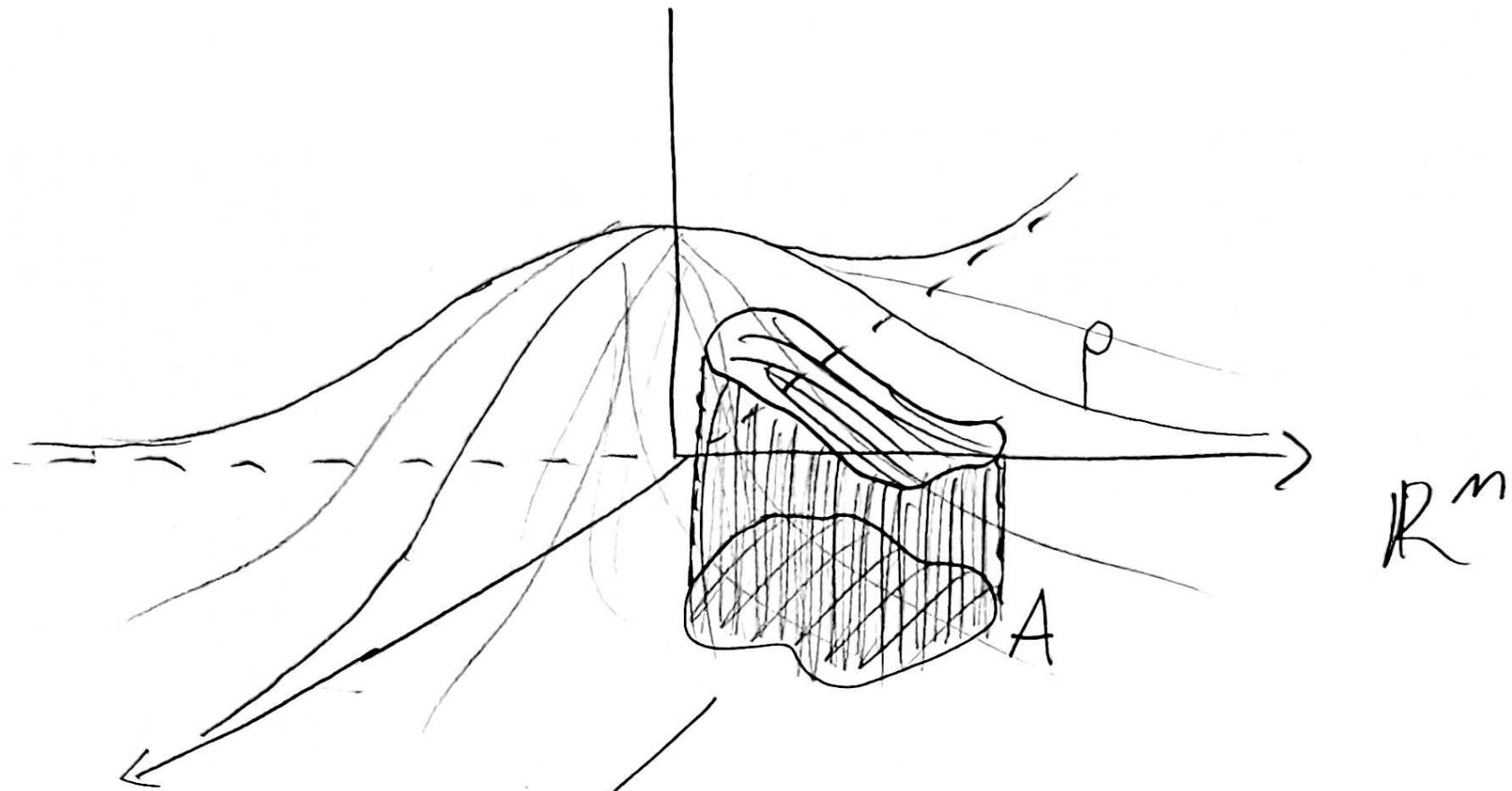
Assim, podemos falar em valor esperado  $E^{\rho}$  das imagens  $l(x)$ :

$$E(l(x)) = \int_{\mathbb{R}^m} l(x) \rho(x) dx$$

Aqui,  $\rho: \mathbb{R}^m \rightarrow \mathbb{R}$ ,  $\rho \geq 0$  e  $\int_{\mathbb{R}^m} \rho(x) dx = 1$

(integrais múltiplas).

11



$$P(X \in A) = \int_A p(x) dx.$$

Verdadeiro objetivo: minimizar o risco esperado  $R(w, b) =$  (12)

$$E(l(h(x; w, b); y)) = \int_{\mathbb{R}^m_x \times \mathbb{R}^m_y} l(h(x; w, b)) p(x, y) dx dy.$$

onde  $p(x, y)$  é a função densidade conjunta das variáveis  $X$  e  $Y$ .

→ no caso discreto temos  $P(x, y) = P(X=x \text{ e } Y=y)$

Quais os problemas?

(13)

- não sabemos quem é  $P$  ...
- só temos em mãos amostras  $(x_i, y_i)$ ,  
 $i=1, \dots, m$ , de  $(X, Y)$ .
- O que pode ser dito acerca de  $E(\dots)$   
a partir das amostras  $(x_i, y_i)$ ?  
↳ ou, qual a relação entre risco  
esperado  $R(w, b)$  e empírico  $R_m(w, b)$ ?

Fixados  $w, b$ , sabemos que [14]

$$|R(w, b) - R_m(w, b)| \xrightarrow{m \rightarrow \infty} 0$$

(lei dos grandes números). Isto é, tomando mais amostras, aproximamos melhor o risco esperado (razoável!).

Mas  $w, b$  variam!

(estamos procurando  $w^*, b^*$  — rede treinada).

Dale o seguinte:

L15

$$\sup_{w,b} |R(w,b) - R_m(w,b)| \rightarrow 0$$

(aumentar as amostras aproxima  $R$  uniformemente)

→ resultados apenas teórico, mas que pelo menos dá esperança que  $\min R_m(w,b)$  tem relação com  $\min R(w,b)$  ...

(na verdade não temos muita escolha : (

↳ de qualquer forma, queremos dizer (16) algo sobre a minimização de  $R$ , porque este é o objetivo real!

Perguntas:

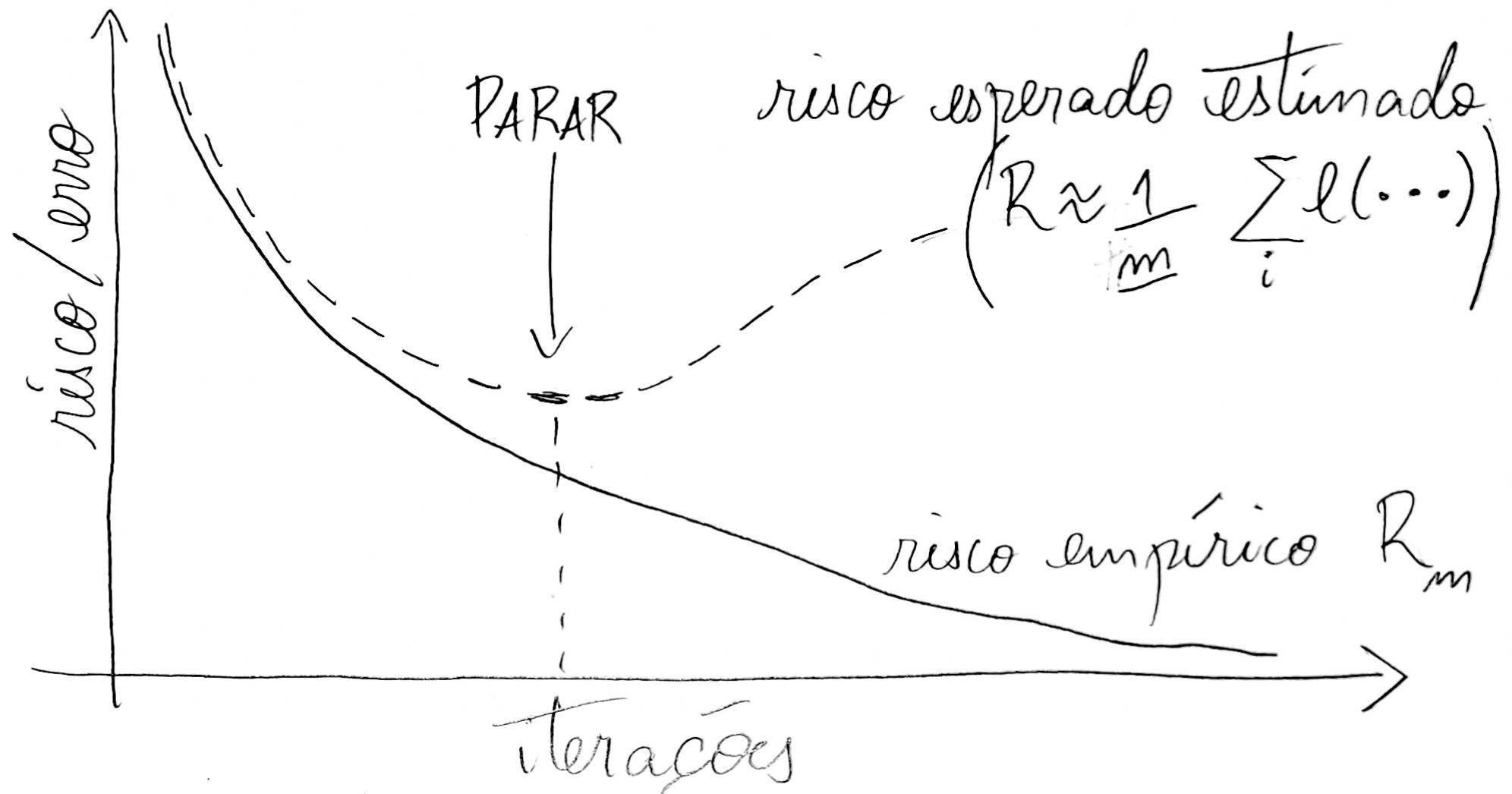
- 1) Conseguimos estimar de minimizar  $R_m$  está minimizando  $R$ ?
- 2) Conseguimos um método que minimize  $R_m$  (função acessível) "olhando"  $R$ ?

A resposta à pergunta 2 recai no método do gradiente estocástico, e será visto à frente. (12)

Resposta à pergunta 1: cross-validation.

- Separamos parte dos dados, que não serão usados no treinamento.  
(geralmente são poucos dados —  $m \ll n$ )
- Usamos esses dados para estimar  $R$  "on-line", durante a minimização de  $R_m$

- Se  $R$  estimado começa a aumentar, paramos a minimização (caso  $R_m \approx 0$ ) (18)



## Propriedades da esperança

(19)

1)  $E(aW + c) = aE(W) + c$  ( $W$  aleatória;  $a, c$ , cte)

De fato,

$$E(aW + c) = \int_{\mathbb{R}} (aw + c) p(w) dw$$

$$= a \int_{\mathbb{R}} w p(w) dw + c \int_{\mathbb{R}} p(w) dw = aE(W) + c.$$

2) Em particular,  $E(c) = c$ .

$$3) E(W - E(W)) = 0 \quad (\text{exercício}) \quad (20)$$

"o valor esperado da diferença entre W e seu valor médio é 0"

$$4) E(X+Y) = E(X) + E(Y)$$

$$5) X \leq Y \Rightarrow E(X) \leq E(Y).$$

" $X \leq Y$  sempre"

$$6) |E(W)| \leq E(|W|) \quad (\text{exercício})$$

7)  $E(\min\{X, Y\}) \leq \min\{E(X), E(Y)\}$ . (21)

### Variância

$$\text{Var}(W) = E((W - E(W))^2)$$

A variância é uma medida de dispersão da variável aleatória de sua média  $E(W)$ .

### Propriedades:

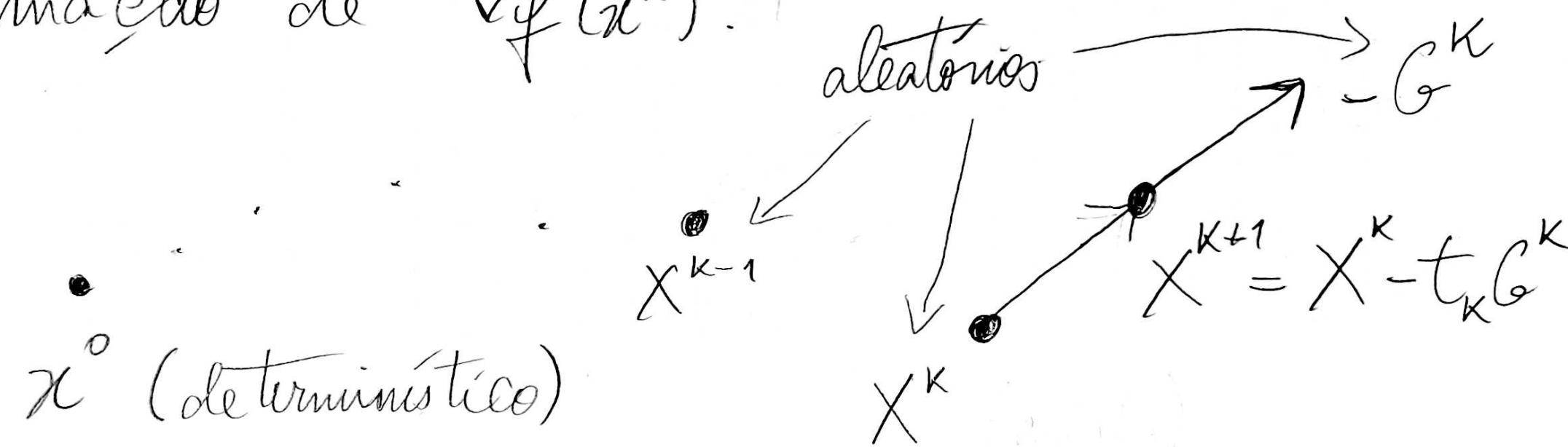
1)  $\text{Var}(aW + c) = a^2 \text{Var}(W)$  (exercício)

2)  $\text{Var}(W) = E(W^2) - (E(W))^2$  (exercício)

## Esperança condicional

22

No método do gradiente estocástico, o passo  $x^{k+1} = x^k - t_k g^k$  é feito com  $g^k$  escolhido aleatoriamente como aproximação de  $\nabla f(x^k)$ .



As variáveis aleatórias  $G^k$  e  $X^{k+1}$  L23

estão condicionadas à escolha  $X^k = x^k$ .

Portanto seu valor esperado também está!

Probabilidade condicional  
(variáveis discretas)

$$P(w|z) = \frac{P(W=w \wedge Z=z)}{P(Z=z)}$$

Distribuição condicional  
(variáveis contínuas)

$$p_{W|Z}(w|z) = \frac{f(w, z)}{p_Z(z)}$$

distrib. conjunta

A esperança condicional de  $W$  dado  $Z=z$  (24)

$Z=z$  é definida de maneira análoga à  $E$ :

$$E(W|Z=z) = \int_{\mathbb{R}} w p_{W|Z}(w|z) dw$$

(análogo para  $\mathbb{R}^n$ ).

Obs.: estamos supondo que  $p_Z(z) > 0$ , isto é,  $w$  tem chance de escolha com  $Z=z$ .

$$p_Z(z) = \int_{\mathbb{R}} p(w|z) dw.$$

Ligado, observe que para cada escolha  $Z = z$ , temos uma média de  $W$  condicionado à  $Z = z$ :  $E(W|Z=z)$ . Podemos considerar a função  $\varphi(z) = E(W|Z=z)$  e a variável aleatória  $\varphi(Z) = E(W|Z)$ .

Essa é a esperança condicional de  $W$  dada  $Z$

## Propriedades da esperança condicional

(26)

- 1)  $E(aX + cY | Z) = aE(X|Z) + cE(Y|Z)$   
 $(X, Y, Z \text{ aleatórias}; a, c \text{ constantes})$
- 2)  $E(c | Z) = c \quad (c \text{ constante})$
- 3)  $X \leq Y \Rightarrow E(X|Z) \leq E(Y|Z)$
- 4)  $|E(W|Z)| \leq E(|W| | Z)$

$$5) E(\min\{X, Y\} | Z) \leq \min\{E(X|Z), E(Y|Z)\}, \quad (27)$$

$$6) E(E(W|Z)) = E(W)$$

"O valor esperado da variável aleatória  $\varphi(Z) = E(W|Z)$  é o valor esperado para  $W$ "