Risco empírico X risco esperado

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

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

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? 2

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

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

↳ idealmente, podemos pensar "x contínuo".

3

Podemos pensar em termos de probabilidade: cada dado (ou par (x,y)) "contínuo" 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 <u>minimizar</u> o <u>valor esperado</u> de má classificação.

↳ em outras palavras, queremos que <u>em média</u> l se aproxime do menor valor possível (p. ex., zero se $l(z,y) = (y-z)^2$)

<u>Esperança ou valor esperado</u>

<u>Notação</u>:

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

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

x	0	100	200	300	400	
$\underbrace{P(X=x)}$	0,2	0,1	0,4	0,2	0,1	$(\Sigma=1)$

↳ probabilidade de $X=x$.

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

Resposta: ao amostrar X N vezes, 20% delas serão 0, 10% serão 100 etc. A média será

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

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

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

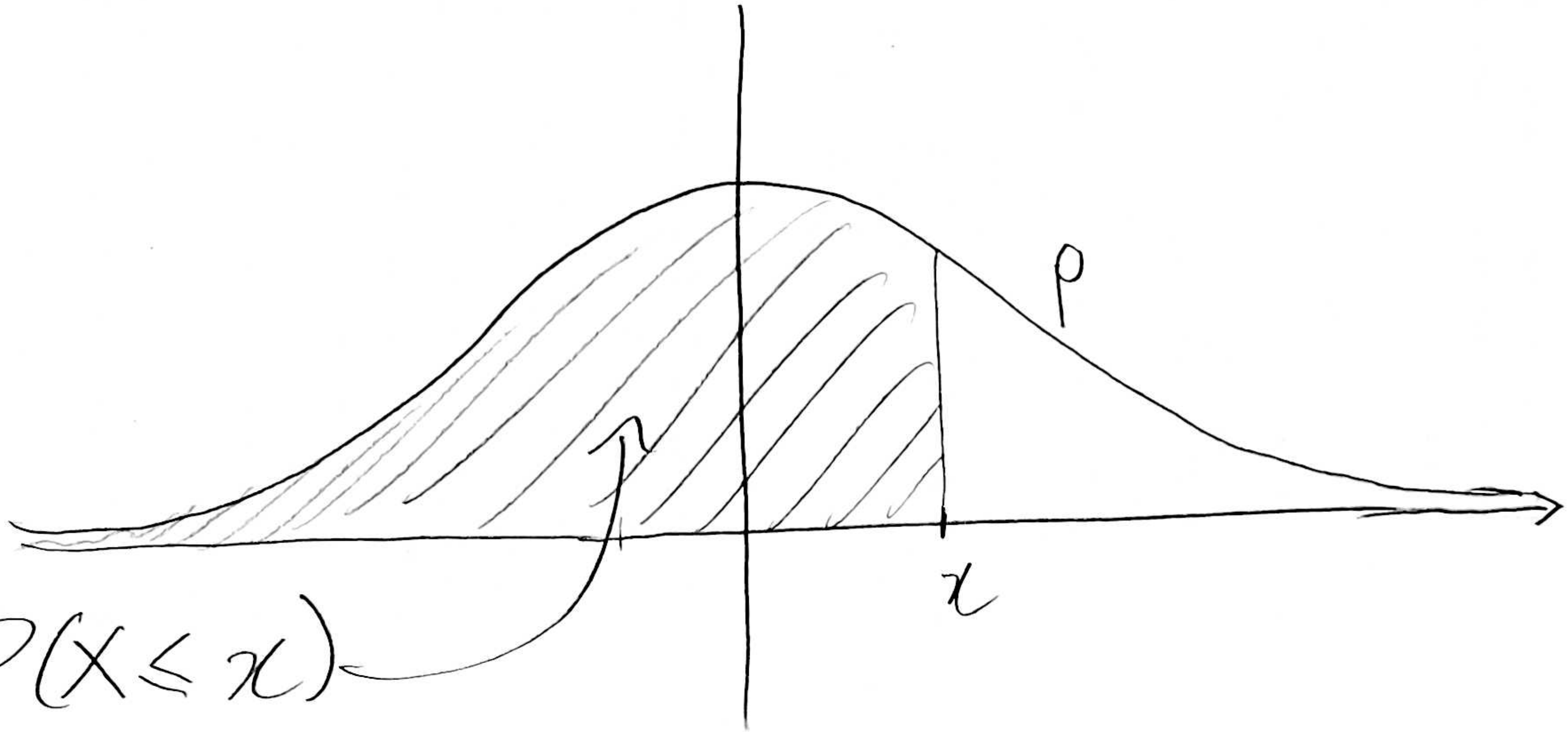
E se X for variável aleatória contínua?

Se ρ for a função densidade de probabilidade de X, então

$$E(X) = \int_{-\infty}^{\infty} x\,\rho(x)\,dx = \int_{\mathbb{R}} x\,\rho(x)\,dx .$$

Lembre-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 $\ell:\mathbb{R}\to\mathbb{R}$ é uma função, podemos
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 <u>vetores</u> de n coordenadas, e

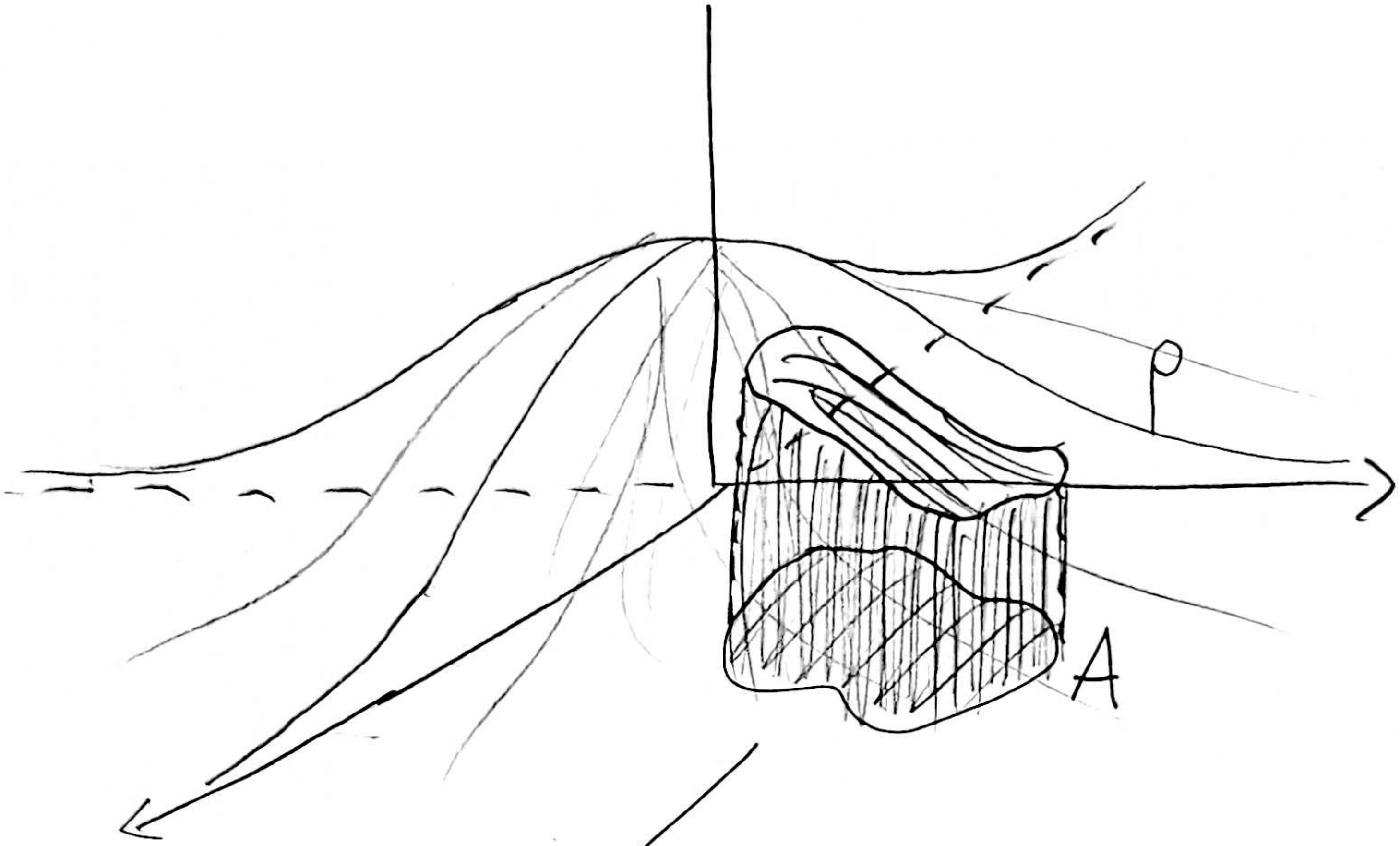
$$\ell:\mathbb{R}^n\longrightarrow\mathbb{R}.$$

Assim, podemos falar em valor esperado das imagens $\ell(X)$: 10

$$E(\ell(X)) = \int_{\mathbb{R}^n} \ell(x)\, \rho(x)\, dx$$

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

(integrais múltiplas).



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

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

$$E\left(\ell(h(X;w,b);Y)\right) = \int_{\mathbb{R}^{n_x} \times \mathbb{R}^{n_y}} \ell(h(x;w,b);y)\, p(x,y)\, dx\, dy$$

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?

- não sabemos quem é ρ ...
- só temos em mãos amostras (x_i, y_i), $i = 1, \ldots, m$, de (X, Y).
- O que pode ser dito acerca de $E(\ldots)$ a partir das amostras (x_i, y_i)?

  ↳ ou, qual a relação entre risco esperado $R(w,b)$ e empírico $R_m(w,b)$?

<u>Fixados</u> w, b, sabemos que

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

(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).

Vale o seguinte:

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

(aumentar as amostras aproxima R <u>uniformemente</u>)

↳ resultado 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 :(

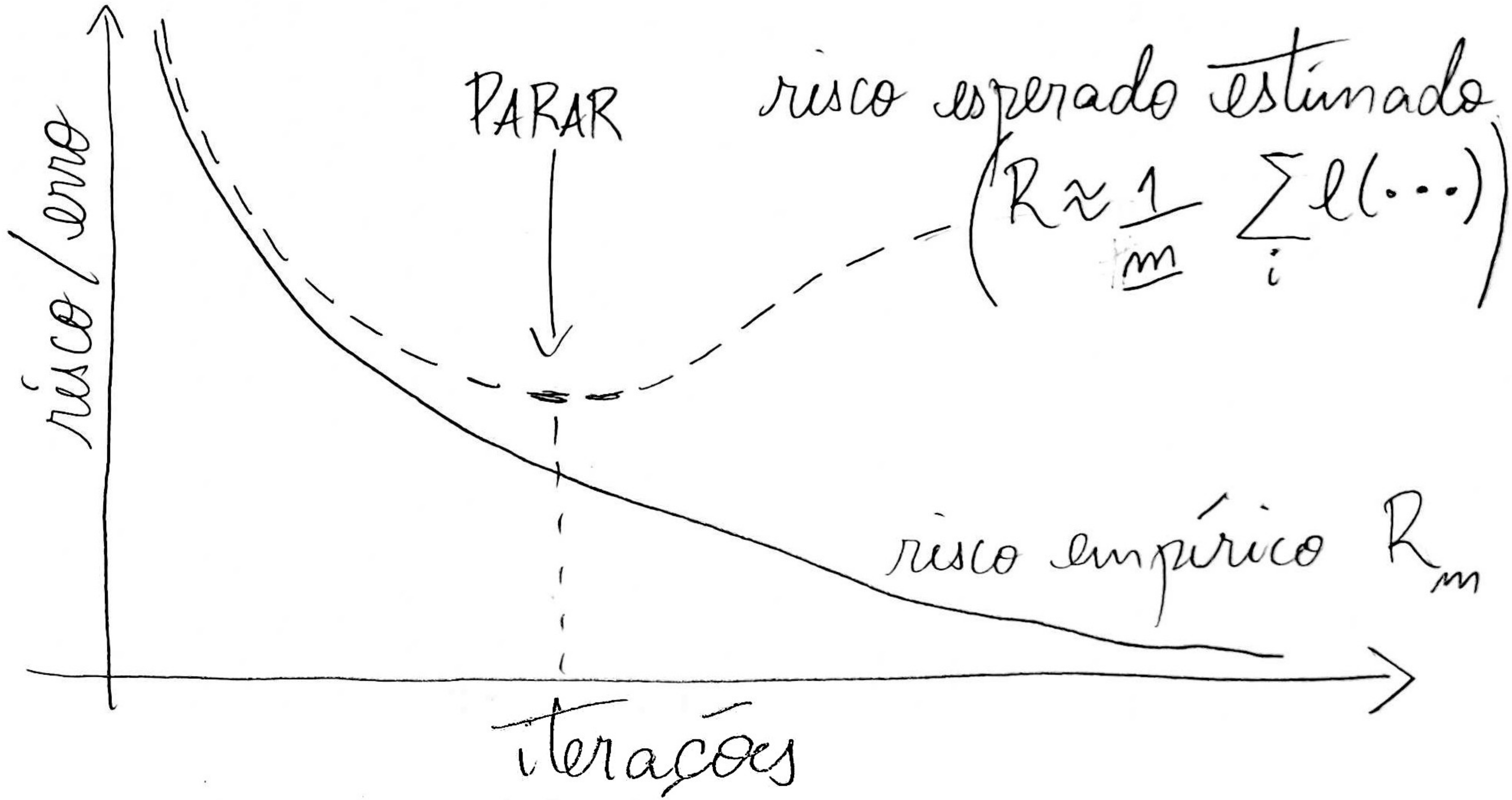
↳ de qualquer forma, queremos dizer algo sobre a minimização de R porque este é o objetivo real!

<u>Pergunta:</u>

Conseguimos estimar se minimizar R_m está minimizando R?

Resposta : parada prematura.

- separamos parte dos dados, que não serão usados no treinamento. (geralmente são poucos dados — $\underline{m} \ll m$)
- usamos esses dados para estimar R "on-line", durante a minimização de R_m
- Se R estimado começa a aumentar, paramos a minimização (caso $R'_m \approx 0$)



PARAR
risco esperado estimado
$\left(R \approx \frac{1}{m} \sum_i \ell(\dots)\right)$
risco empírico R_m
risco / erro
iterações

Propriedades da esperança

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) $E(W - E(W)) = 0$ (exercício)

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

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

5) $\underbrace{X \leq Y}_{}$ $\Rightarrow E(X) \leq E(Y)$.

"$x \leq y$ sempre"

6) $|E(W)| \leq E(|W|)$ (exercício)

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

Variância

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

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

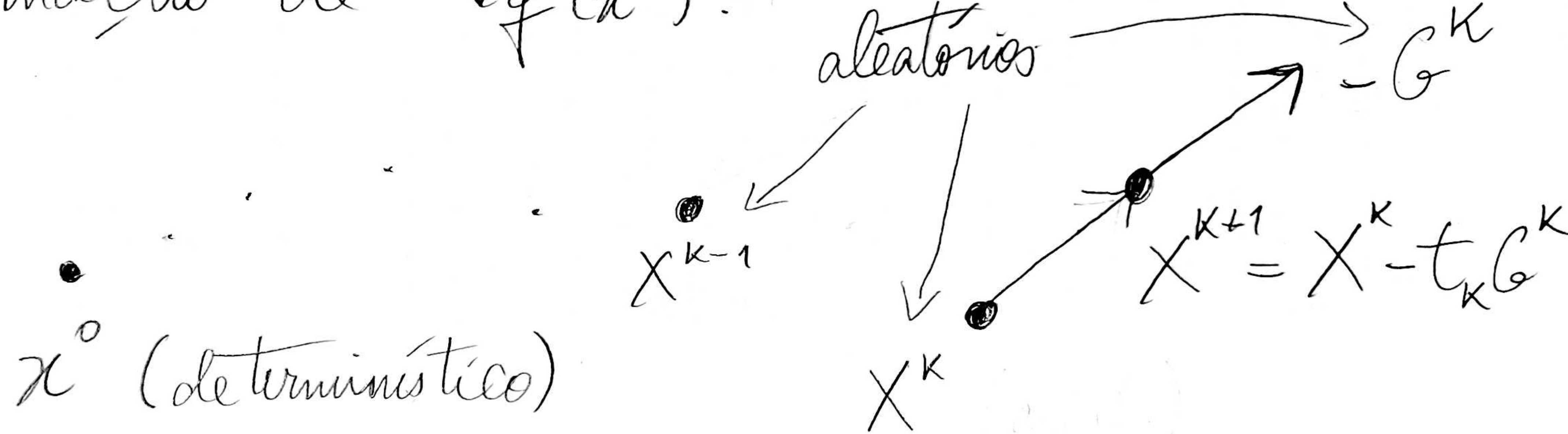
Propriedades:

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

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

Esperança condicional

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} estão condicionadas à escolha $X^K = x^K$. Portanto seu valor esperado também está!

Probabilidade condicional (variáveis discretas)	Distribuição condicional (variáveis contínuas)
$P(w\|z) = \dfrac{P(W=w \text{ e } Z=z)}{P(Z=z)}$	$p_{W\|Z}(w\|z) = \dfrac{p(w,z)}{p_Z(z)}$

distrib. conjunta

A esperança condicional de W dado $\underline{Z = z}$ é definida de maneira análoga à E:

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

(análogo para $\mathbb{R}^n$).

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

$$\rho_Z(z) = \int_{\mathbb{R}} \rho(w, z) \, dw.$$

Agora, 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 <u>esperança condicional de W dada Z</u>

Propriedades da esperança condicional

1) $E(aX + cY \mid Z) = aE(X \mid Z) + cE(Y \mid Z)$
(X, Y, Z aleatórias ; a, c constantes)

2) $E(c \mid Z) = c$ (c constante)

3) $X \leq Y \Rightarrow E(X \mid Z) \leq E(Y \mid Z)$

4) $|E(W \mid Z)| \leq E(|W| \mid Z)$

5) $E(\min\{X, Y\} \mid Z) \leq \min\{E(X \mid Z), E(Y \mid Z)\}$

6) $E\big(E(W \mid Z)\big) = E(W)$

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