Processamento Estatístico de Sinais

Charles Casimiro Cavalcante

charles@gtel.ufc.br

Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará – UFC
http://www.ppgeti.ufc.br/charles

UNIVERSIDADE FEDERAL DO CEARÁ

"O processamento de sinais mudou! Não estamos mais na era na qual a informação na forma de sinais elétricos é processada por meio de tradicionais dispositivos analógicos. Nós estamos solidamente, e, para o futuro previsível, irrevogavelmente, no âmago do processamento de sinais digitais (amostrados ou discretos no tempo) aleatórios."

Charles W. Therrien, 1992

Discrete Random Signals and Statistical Signal Processing

Conteúdo do curso

1. Revisão de modelos probabilísticos
2. Análise de momentos de segunda ordem
3. Teoria da estimação
4. Filtragem ótima
5. Predição de sinais estacionários
6. Teoria da detecção
7. Métodos recursivos no tempo
8. Filtragem adaptativa

Parte I

Revisão de Modelos Probabilísticos

Evento

Definição: Qualquer subconjunto do espaço amostral $\mathcal{S}$ que constitui um campo de Borel $\mathcal{F}$

Eventos mutuamente exclusivos

Quando a ocorrência de um impossibilita a ocorrência do outro
Exemplo: Dado

$$\left.\begin{array}{c} A = \{\text{par}\} \\ B = \{\text{impar}\} \end{array}\right\rangle A \cdot B = \emptyset \quad (\textbf{eventos mutuamente exclusivos})$$

Probabilidade (Definição Axiomática)

É qualquer função real definida na classe $\mathcal{F}$ tal que

1. $\Pr(A) \geq 0$
2. $\Pr(\mathcal{S}) = 1$
3. Se $A \cdot B = \emptyset \Rightarrow \Pr(A + B) = \Pr(A) + \Pr(B)$ (eventos mutuamente exclusivos)

Assim,

$$\Pr(\cdot) : \mathcal{F} \to \mathbb{R}$$

Probabilidade condicional

Probabilidade de ocorrência de A dado que ocorreu B

$$\Pr(A|B) \triangleq \frac{\Pr(AB)}{\Pr(B)}, \quad \Pr(B) > 0 \tag{1}$$

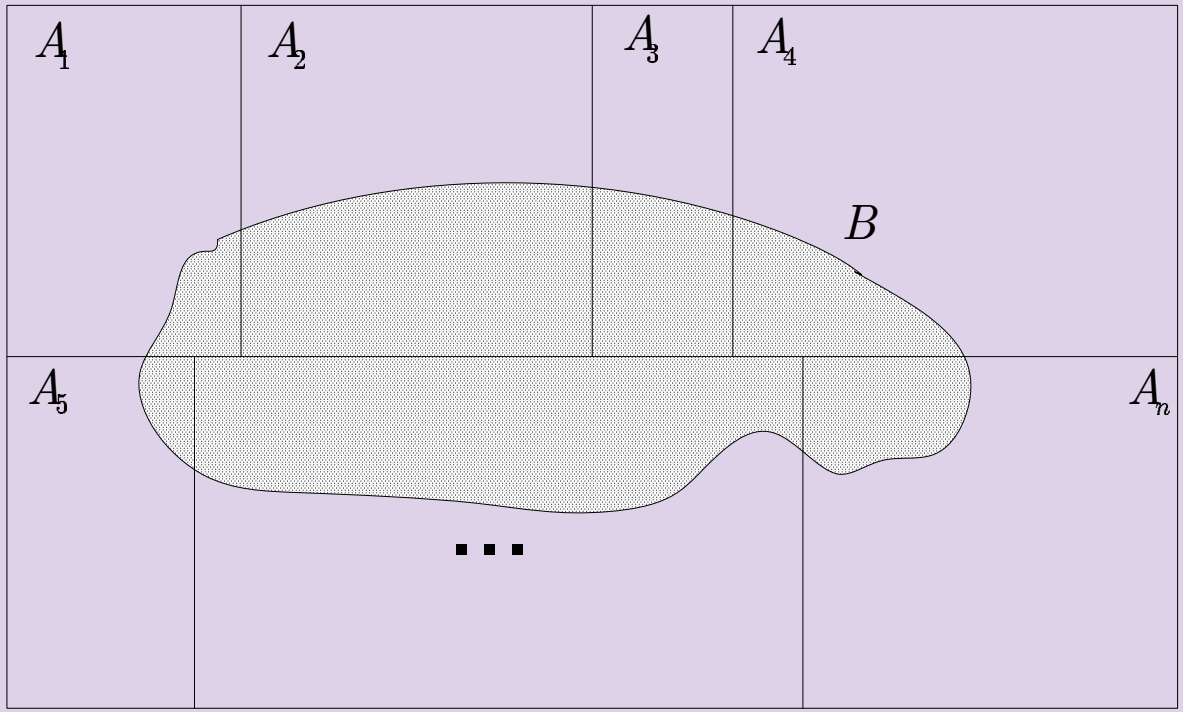
$\Pr(A|B)$ é probabilidade, pois

- $\Pr(A|B) = \frac{\Pr(AB)}{\Pr(B)} \geq 0$
- $\Pr(\mathcal{S}|B) = 1$
- Para $A \cdot C = \emptyset \Rightarrow \Pr\left[(A + C)|B\right] = \Pr(A|B) + \Pr(C|B)$

Teorema da probabilidade total

Sejam $A_1, A_2, A_3, \ldots, A_n$ eventos mutuamente exclusivos

- $\Pr(A_i) > 0, \quad i = 1, 2, \ldots, n$
- $B \subset \{A_1 + A_2 + \cdots + A_n\}$



Teorema da probabilidade total - cont.

- $\Pr(B) = \Pr(BA_1) + \Pr(BA_2) + \cdots + \Pr(BA_n)$ pois $B = \underbrace{BA_1 + BA_2 + \cdots + BA_n}_{\text{mutuamente exclusivos}}$
- $\Pr(B) = \Pr(B|A_1)\Pr(A_1) + \cdots + \Pr(B|A_n)\Pr(A_n)$

Probabilidade total

$$\Pr(B) = \sum_{i=1}^{n} \Pr(B|A_i)\Pr(A_i) \tag{2}$$

Regra de Bayes

Inverso do conceito da probabilidade total

$$\Pr(A_j|B) = \frac{\Pr(B|A_j) \cdot \Pr(A_j)}{\sum_{i=1}^{n} \Pr(B|A_i) \cdot \Pr(A_i)} \tag{3}$$

Também chamada de *probabilidade a posteriori*

Eventos independentes

Dois eventos A e B são independentes se

$$\Pr(A \cdot B) = \Pr(A) \cdot \Pr(B)$$

Generalizando (para três eventos): A, B e C

$$\left.\begin{array}{l} \Pr(AB) = \Pr(A)\Pr(B) \\ \Pr(AC) = \Pr(A)\Pr(C) \\ \Pr(BC) = \Pr(B)\Pr(C) \end{array}\right\} \Pr(ABC) = \Pr(A)\Pr(B)\Pr(C)$$

Propriedades de eventos independentes

1. $\Pr(A|B) = \Pr(A)$
2. $\Pr(\overline{A}B) = \Pr(\overline{A}) \cdot \Pr(B)$
3. $\Pr(A\overline{B}) = \Pr(A) \cdot \Pr(\overline{B})$ e $\Pr(\overline{AB}) = \Pr(\overline{A}) \cdot \Pr(\overline{B})$

Ou seja Se A e B são independentes, A e $\overline{B}$ são independentes e $\overline{A}$ e $\overline{B}$ também o são

Eventos conjuntos

Dado $\mathcal{S}$ (espaço amostral), podemos atribuir diferentes atributos aos eventos pertencentes a diferentes classes de Borel

$\mathcal{S} = \{x_1, x_2, \ldots, x_n\}$

$$\begin{cases} A_1, A_2, \cdots, A_n \in \mathcal{F}_1 \\ B_1, B_2, \cdots, B_n \in \mathcal{F}_2 \end{cases}$$

Exemplo:

$\mathcal{S} = \{$João, José, Maria $\}$

(idade,altura)

Rescrevendo:

$\mathcal{S} = \{(10, 1.50), (30, 1.80), (32, 1.65)\}$

Probabilidade marginal

$$A_1 + A_2 + \cdots + A_n = \mathcal{S}_1$$
$$B_1 + B_2 + \cdots + B_n = \mathcal{S}_2$$

$$\left.\begin{array}{l} \Pr(A_i) = \sum_{j=1}^{n} \Pr(A_i, B_j) \\ \Pr(B_j) = \sum_{i=1}^{n} \Pr(A_i, B_j) \end{array}\right\} \sum_{i=1}^{n}\sum_{j=1}^{n} \Pr(A_i, B_j) = 1 \quad (4)$$

Definição

Variável aleatória (v.a.) é qualquer função definida no espaço amostral $\mathcal{S}$ tal que:

$$\{X : \mathcal{S} \to \mathbb{R}, X(w) \in (-\infty, x], w \in \mathcal{S}\} \in \mathcal{F} \tag{5}$$

Exemplo

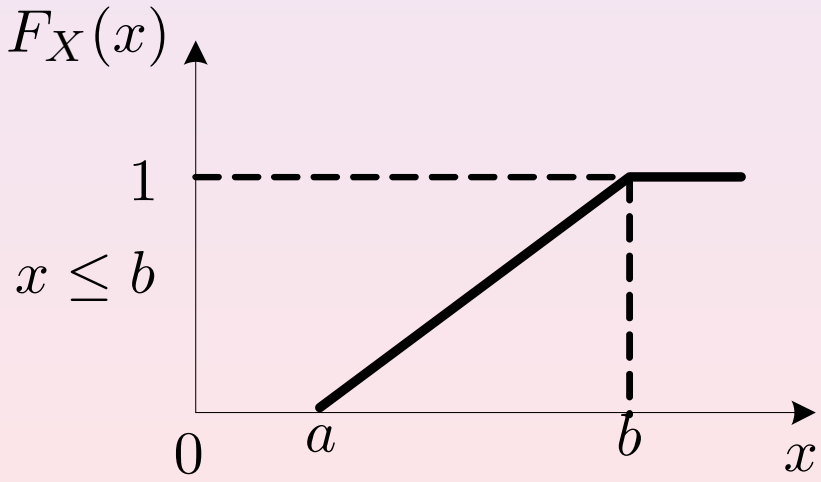
- Moeda:
  $\mathcal{S} = \{\text{cara}, \text{coroa}\}$
  $X(\text{cara}) = 0$
  $X(\text{coroa}) = 1$

Definição

$$\boxed{F_X(x) \triangleq \Pr\{X \leq x\}} \tag{6}$$

Exemplo: *Distribuição uniforme*

$$F_X(x) = \begin{cases} 0, & \text{se } x \leq a \\ x - a, & \text{se } a < x \leq b \\ 1, & \text{se } x > b \end{cases}$$



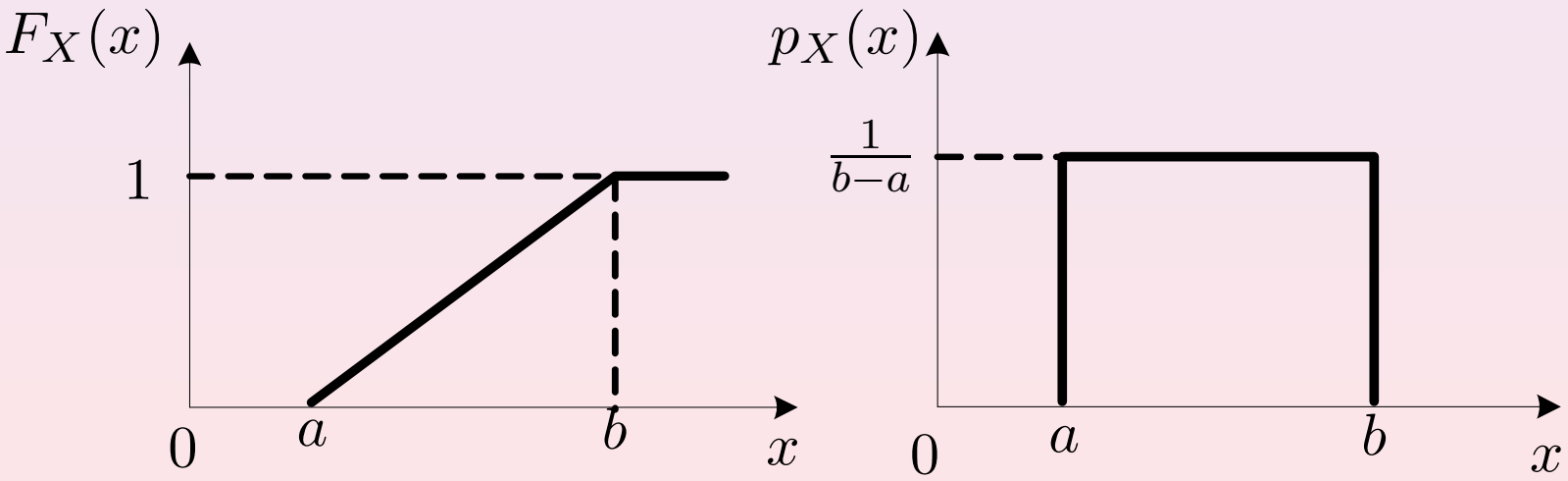
Propriedades da fdc

1. $F_X(-\infty) = 0$
2. $F_X(\infty) = 1$
3. $\Pr\{x_1 \leq X \leq x_2\} = F_X(x_2) - F_X(x_1)$
4. Se $x_1 \leq x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)$, ou seja, $F_X(x)$ é *monotônico não-decrescente*

Definição

$$p_X(x) \triangleq \frac{d}{dx} F_X(x) \tag{7}$$

Exemplo: *Distribuição uniforme*

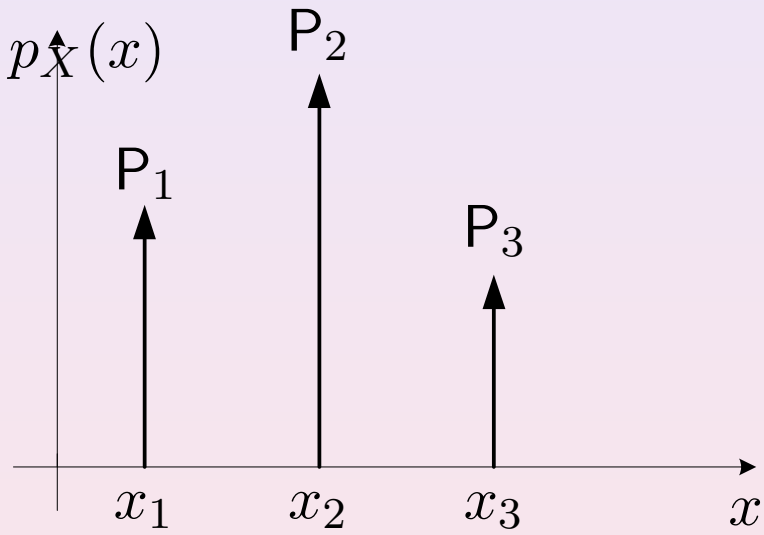


Propriedades

1. $F_X(x) = \int_{-\infty}^{x} p_X(\xi)\, d\xi$
2. $F_X(x)$ é monotônico não-decrescente $\Rightarrow p_X(x) \geq 0$
3. $\Pr\{X > x\} = 1 - F_X(x) = 1 - \Pr\{X \leq x\}$
4. $\Pr\{x_1 < X \leq x_2\} = F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} p_X(\xi)\, d\xi$

Dificuldade

Neste caso, as variáveis admitem valores somente em determinados instantes de tempo. O que ocorre com as probabilidades?



Funções $\delta(\cdot)$ de Dirac (impulsos)

$$\delta(t) \begin{cases} \int_{-\infty}^{\infty} f(t)\delta(t-t_0)\,dt = f(t_0) \\ \int_{-\infty}^{\infty} \delta(t)\,dt = 1 \end{cases}$$

$\delta(t) =$ função impulsiva de Dirac

$$= \frac{d}{dt}u(t), \text{em que } u(t) \text{é a função degrau unitário}$$

Logo, teremos

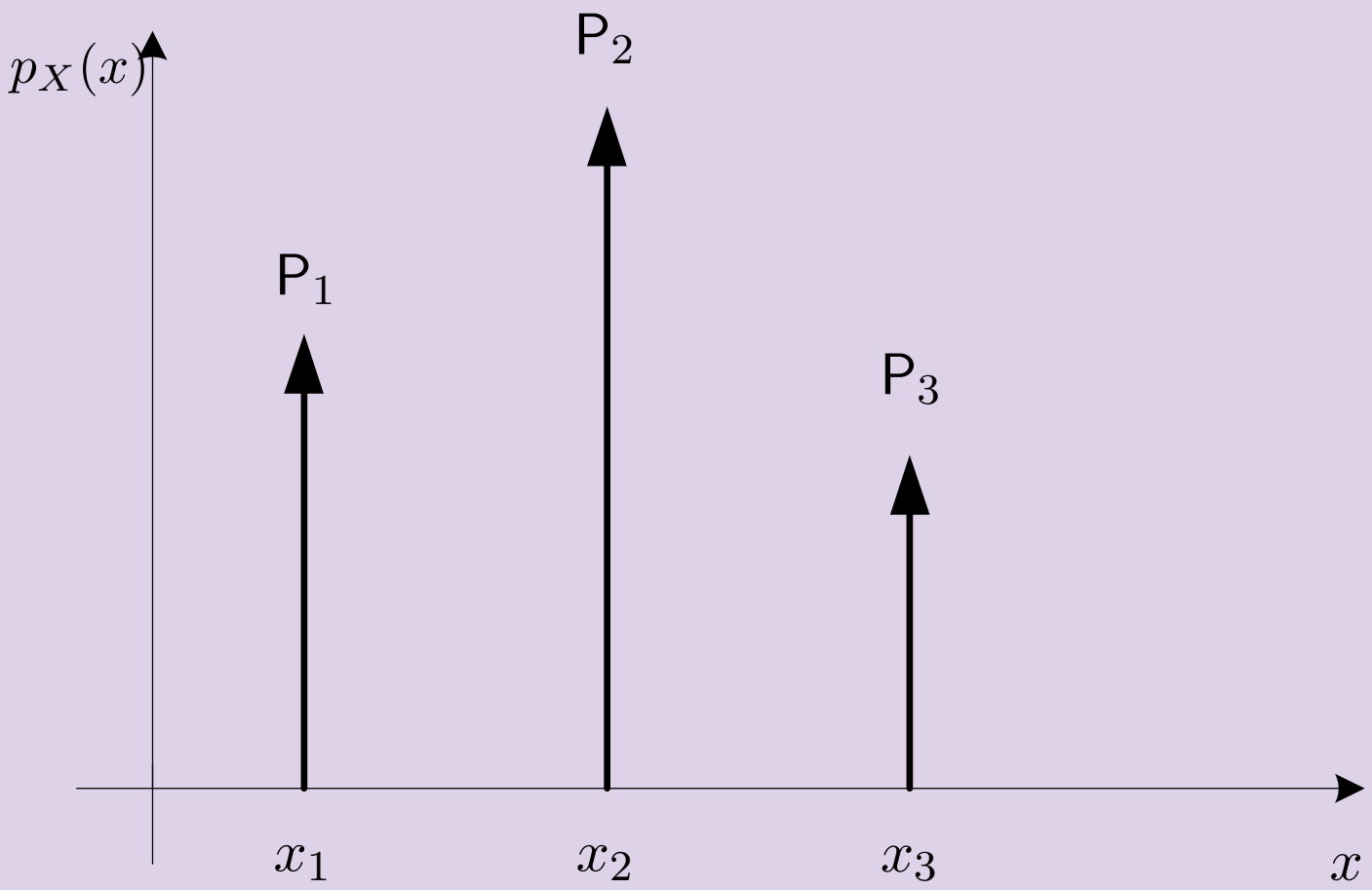
$$p_X(x) = \sum_{i=1}^{N} \Pr\{X = x_i\} \cdot \delta(x - x_i)$$

$$F_X(x) = \int_{-\infty}^{x} p_X(\xi)\, d\xi = \sum_{i=1}^{N} \Pr\{X = x_i\} \cdot \int_{-\infty}^{x} \delta(\xi - x_i)\, d\xi$$

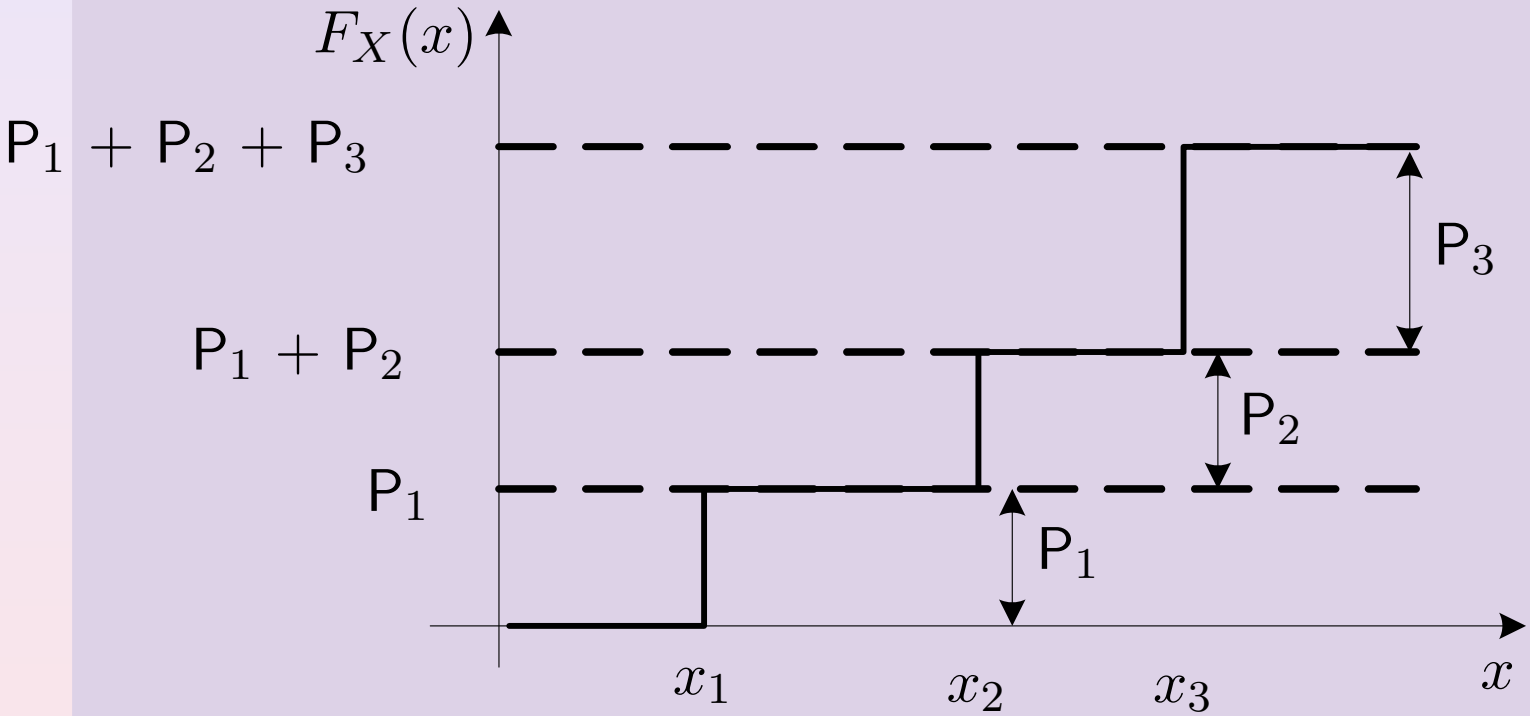
Mas sabe-se que

$$\int_{-\infty}^{x} \delta(\xi - x_i)\, d\xi = \begin{cases} 0, & \text{se } x < x_i \\ 1, & \text{se } x \geq x_i \end{cases}$$

pdf



fdc



Definição

Seja X uma v.a., X é dito ter distribuição de probabilidade gaussiana, ou *normal*, se sua densidade de probabilidade pode ser escrita da seguinte forma

$$p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{8}$$

Notação usual:

$$X \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

Parâmetros $\begin{cases} \mu \rightarrow \text{ média} \\ \sigma^2 \rightarrow \text{ variância} \end{cases}$

Normalização

$$Z \sim \mathcal{N}(0,1)$$

$$\boxed{p_Z(z) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{z^2}{2}\right)}$$

Função erro

Função de distribuição cumulativa da função gaussiana

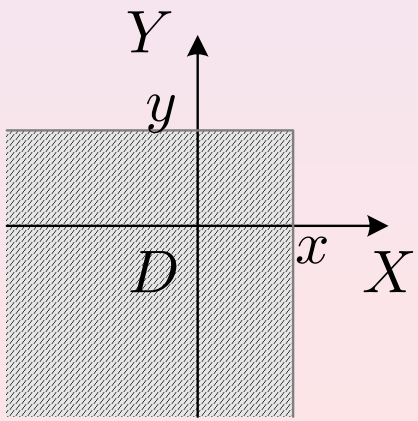
$$\begin{aligned} \text{erf}(x) = F_Z(x) &= \Pr\{Z \leq x\} \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{z^2}{2}\right) \, dz \end{aligned} \tag{9}$$

Função distribuição de probabilidade

$$F_{X,Y}(x,y) = \Pr\underbrace{\{X \leq x, \quad Y \leq y\}}_{\text{intersecção}} \tag{10}$$

$$\underbrace{\{X \leq x, \quad Y \leq y\}}_{\text{evento}} = \{\ w \in \mathcal{S} | [X(w), Y(w)] \in D\ \}$$

em que $D = \{\ (X,Y) | X \in (-\infty, x], \quad Y \in (-\infty, y]\ \}$



$$\boxed{p_{X,Y}(x,y) = \frac{\partial^2}{\partial x\ \partial y} F_{X,Y}(x,y)} \tag{11}$$

Propriedades

1. $F_{X,Y}(-\infty, y) = 0$
2. $F_{X,Y}(x, -\infty) = 0$
3. $F_{X,Y}(\infty, \infty) = 1$
4. $\left.\begin{array}{l} F_{X,Y}(x, \infty) = F_X(x) \\ F_{X,Y}(\infty, y) = F_Y(y) \end{array}\right\}$ distribuições marginais
5. $F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} p_{X,Y}(\alpha, \beta)\, d\alpha\, d\beta$
6. $p_X(x) = \int_{-\infty}^{\infty} p_{X,Y}(x, y)\, dy$

   $p_Y(y) = \int_{-\infty}^{\infty} p_{X,Y}(x, y)\, dx$

Definição

Sejam X e Y v.a.'s. Elas são independentes se:

$$\boxed{F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y)} \tag{12}$$

$$\begin{aligned} F_{X,Y}(x,y) &= \Pr\{\ X \leq x, \quad Y \leq y\ \} \\ &= \Pr\{X \leq x\} \cdot \Pr\{Y \leq y\} \end{aligned}$$

Logo:

$$\boxed{p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y)} \quad \text{pois}$$

$$p_{X,Y}(x,y) = \frac{\partial^2}{\partial x\ \partial y} F_{X,Y}(x,y)$$

$$= \frac{\partial^2}{\partial x\ \partial y} \left(F_X(x) \cdot F_Y(y)\right) \overbrace{=}^{?} p_X(x) \cdot\ p_Y(y)$$

Momentos

São *estatísticas* de uma variável aleatória capazes de representar seu comportamento probabilístico. Os infinitos momentos estatísticos definem a função de densidade de probabilidade.

Média

Também chamada de esperança matemática, valor esperado, momento de 1^a ordem, é definido como

$$\mu = \mathbb{E}\{X\} \triangleq \int_{-\infty}^{\infty} x \cdot p_X(x)\, dx \tag{13}$$

para $X \sim$ v.a. contínua

Média

Para X discreta

$$\mu = \mathbb{E}\{X\} = \sum_{i=-\infty}^{\infty} x_i \cdot \Pr\{X = x_i\} \tag{14}$$

OBS: Se $p_X(x)$ for simétrico em relação a um valor $x = a \Rightarrow \mathbb{E}\{X\} = a$

Pergunta: Num jogo de moeda, qual a média? E no caso de uma distribuição uniforme entre $[0, 1]$?

Propriedades da média

1. Linearidade - $\mathbb{E}\{X+Y\} = \mathbb{E}\{X\} + \mathbb{E}\{Y\}$ ou ainda, $\mathbb{E}\left\{\sum_{i=1}^{m} a_i X_i\right\} = \sum_{i=1}^{m} a_i \mathbb{E}\{X_i\}$
2. $\mathbb{E}\{X \cdot Y\} = \mathbb{E}\{X\} \cdot \mathbb{E}\{Y\}$ se X e Y são v.a.'s independentes
3. Transformação linear - $\mathbb{E}\{\mathbf{A}X\} = \mathbf{A}\mathbb{E}\{X\}$, em que $\mathbf{A}$ é uma matriz qualquer
4. $\mathbb{E}\{f(X,Y)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y) p_{X,Y}(x,y)\, dxdy$

Propriedades da média - cont.

5. Invariância à transformação - Seja $Y = g(X)$, então

$$\int_{-\infty}^{\infty} y p_Y(y)\ dy = \int_{-\infty}^{\infty} g(x) p_X(x)\ dx$$

6. Se X e Y são independentes
$\mathbb{E}\{f(X) \cdot g(Y)\} = \mathbb{E}\{f(X)\} \cdot \mathbb{E}\{g(Y)\}$

Momentos de ordem k

$$\mu_k = \mathbb{E}\left\{X^k\right\} = \int_{-\infty}^{\infty} x^k \cdot p_X(x)\, dx \tag{15}$$

- Os momentos de uma variável aleatória são uma representação da pdf da variável
- A coletânea dos *infinitos* momentos da v.a. definem sua pdf
- Algumas distribuições possuem alguns momentos nulos
- A estimativa de momentos cresce em complexidade e decresce em precisão com o aumento direto de k

Momentos centrados

Uma importante medida estatística é avaliar o comportamento da v.a. *em torno* da média. Assim, define-se o *momento centrado de ordem* k como sendo

$$c_k = \mathbb{E}\left\{(X-\mu)^k\right\} = \int_{-\infty}^{\infty} (x-\mu)^k \cdot p_X(x)\, dx \qquad (16)$$

De particular interesse: variância $\left(\sigma^2 = \mathbb{E}\left\{(X-\mu)^2\right\} \geq 0\right)$

OBS: Se $p_X(x)$ é simétrica em relação a média $c_k = 0$ para $\forall k$ ímpar!

Meta

Caracterização da distribuição de probabilidade de uma variável aleatória condicionada a ocorrência de outra variável aleatória ou evento

Definição distribuição cumulativa condicionada

$$\begin{aligned} F_X(x|A) &\triangleq \Pr\{X \leq x|A\} \\ &= \frac{\Pr\{X \leq x, A\}}{\Pr\{A\}} \end{aligned} \tag{17}$$

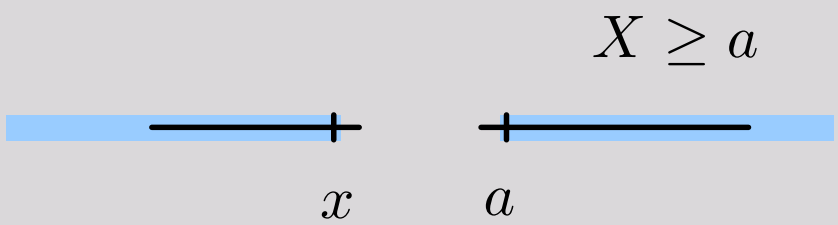
- $p_X(x|A) \triangleq \dfrac{d}{dx} F_X(x|A)$

Analisando...

Seja $A = \{x \geq a\}$ (evento)

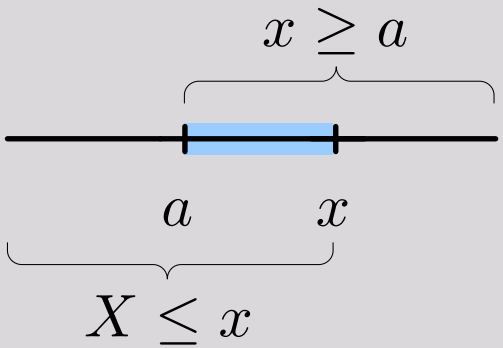
$$\begin{aligned} F_X(x|A) &= F_X(x|X \geq a) \\ &= \Pr\{X \leq x | X \geq a\} \end{aligned}$$

(a) $x < a$



$$\Rightarrow F_X(x|x \geq a) = 0$$

(a) $x \geq a$



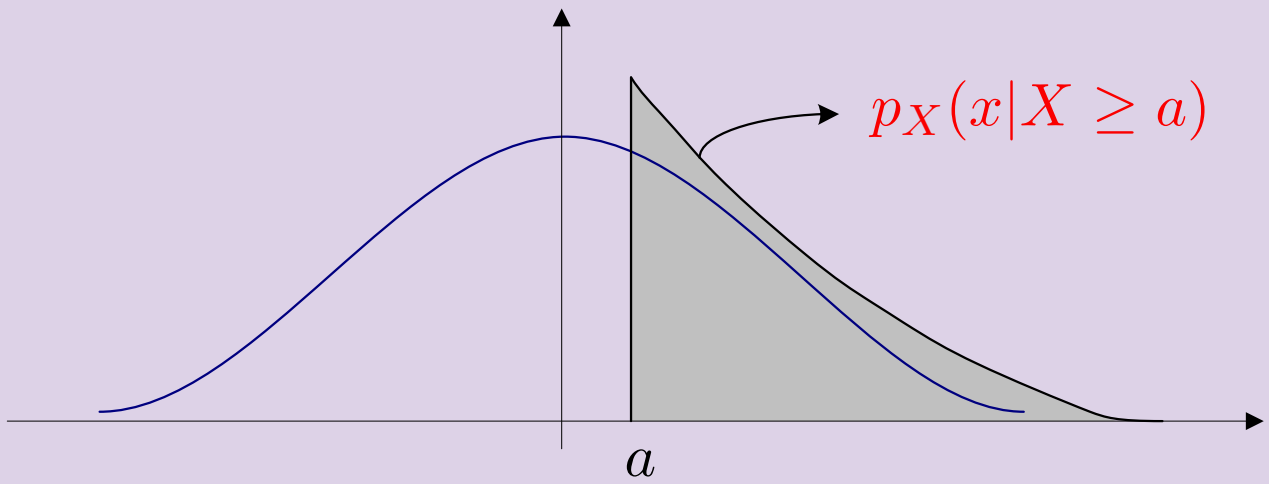
$$F_X(x|x \geq a) = \frac{\Pr\{X \leq x, X \geq a\}}{\Pr\{X \geq a\}} = \frac{\int_a^x p_X(\alpha)\, d\alpha}{\int_a^\infty p_X(\beta)\, d\beta}$$

$$= \frac{F_X(x) - F_X(a)}{1 - F_X(a)}$$

$$p_X(x|X \geq a) = \frac{d}{dx} F_X(x|X \geq a) = \frac{\frac{d}{dx} F_X(x)}{1 - F_X(a)} = \frac{p_X(x)}{1 - F_X(a)}$$

Resumindo

$$p_X(x|X \geq a) = \frac{p_X(x)}{\int_0^{\infty} p_X(\beta)\, d\beta} U(x-a)$$



Observações

1. $$\int_{-\infty}^{\infty} p_X(x|X \geq a)\ dx = \int_{-\infty}^{\infty} \left[\frac{p_X(x)}{\int_a^{\infty} p_X(\beta)\ d\beta} \right] \cdot U(x-a)\ dx$$
$$= \frac{\int_a^{\infty} p_X(x)\ dx}{\int_a^{\infty} p_X(\beta)\ d\beta} = 1$$

Observações - cont.

2. $\mathbb{E}\{X|A\} = ?$

$$A = \{X \geq a\}$$

$$\mathbb{E}\{X|X \geq a\} = \int_{-\infty}^{\infty} x \cdot p_X(x|X \geq a)\ dx$$

$$= \int_{\infty}^{\infty} \frac{x p_X(x)}{\int_a^{\infty} p_X(\beta)\ d\beta} \cdot U(x-a)\ dx$$

$$\boxed{\mathbb{E}\{X|X \geq a\} = \frac{\int_a^{\infty} x \cdot p_X(x)\ dx}{\int_a^{\infty} p_X(\beta)\ d\beta}}$$

Observações - cont.

3. Caso em que $A = \{X = a\}$

$$F_X(x|X=a) = \frac{\Pr\{X \leq x, X = a\}}{\underbrace{\Pr\{X = a\}}_{\text{v.a. contínua} \Rightarrow \Pr\{X=a\}=0}}$$

Relaxando "um pouco" $X = a$ para

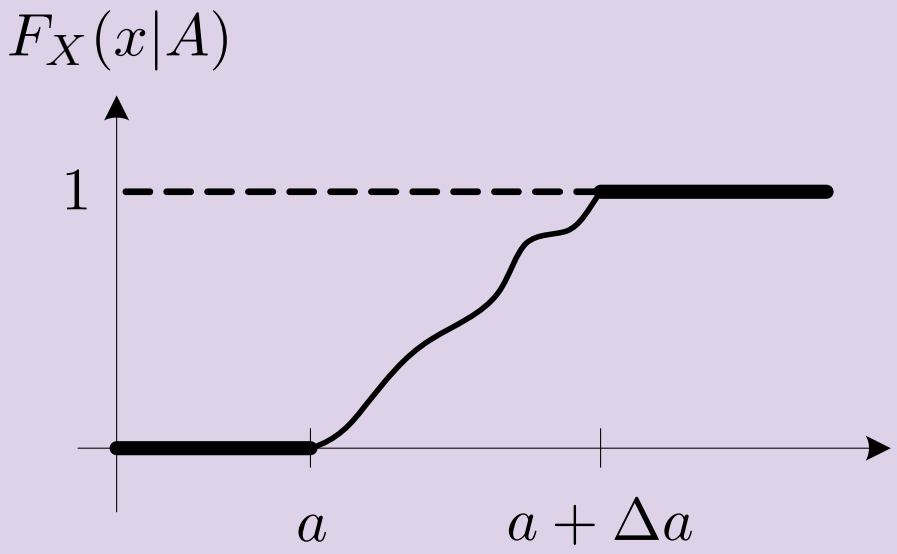
$$a \leq X \leq a + \Delta a, \quad \Delta a \to 0 \qquad \text{(depois)}$$

$$F_X(x|A) = F_X(x|a \leq X \leq a + \Delta a) = \frac{\Pr\{X \leq x, a \leq X \leq a + \Delta a\}}{\Pr\{a \leq X \leq a + \Delta a\}}$$

Observações - cont.

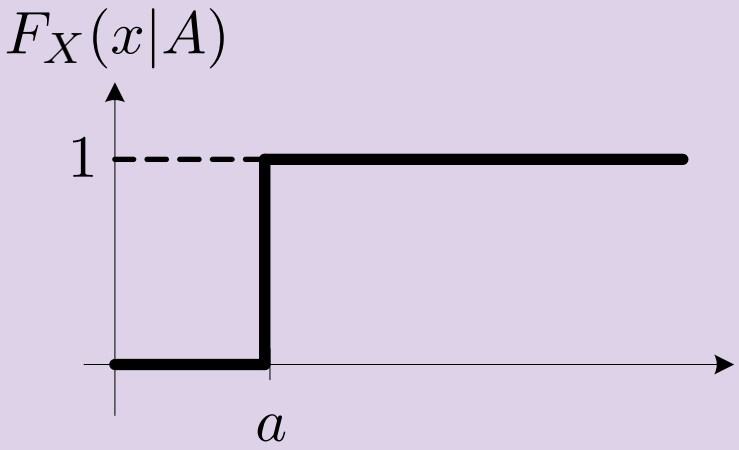
Temos 3 situações

- $X < a \Rightarrow F_X(x|A) = 0$
- $a \leq X \leq a + \Delta a \Rightarrow F_X(x|A) = \frac{\Pr\{a \leq X \leq x\}}{\Pr\{a \leq X \leq a+\Delta a\}}$
- $X > a + \Delta a \Rightarrow F_X(x|A) = 1$



Observações - cont.

Fazendo $\Delta \to 0$:



$$F_X(x|A) = \underbrace{U(x-a)}_{\text{função degrau unitário}}$$

$$\Rightarrow F_X(x|X=a) = U(x-a)$$

$$\boxed{p_X(x|X=a) = \frac{d}{dx} F_X(x|X=a) = \delta(x-a)}$$

Função densidade condicional de duas variáveis aleatórias

$$p_Y(y|x) = ?$$

$$F_Y(y|x \leq X \leq x+\Delta x) = \frac{\Pr\{Y \leq y, x \leq X \leq x+\Delta x\}}{\Pr\{x \leq X \leq x+\Delta x\}}$$

Ainda

$$\Pr\{Y \leq y, x \leq X \leq x+\Delta x\} = \int_{-\infty}^{y} \int_{x}^{x+\Delta x} p_{X,Y}(\alpha,\beta)\, d\alpha d\beta$$

$$\cong p_{X,Y}(x,\beta)\Delta x \quad \text{método de Euler}$$

$$= \int_{-\infty}^{y} p_{X,Y}(x,\beta)\, d\beta \cdot \Delta x$$

Função densidade condicional de duas variáveis aleatórias - cont.

$$\Pr\{x \leq X \leq x + \Delta x\} = \int_{x}^{x+\Delta x} p_X(\gamma)\, d\gamma \cong p_X(x) \cdot \Delta x$$

$$F_Y(y|x) \cong \frac{\int_{-\infty}^{y} p_{X,Y}(x,\beta)\, d\beta \cdot \Delta x}{p_X(x) \cdot \Delta x} = \frac{\int_{-\infty}^{y} p_{X,Y}(x,\beta)\, d\beta}{p_X(x)}$$

$$p_Y(y|x) = \frac{dF_Y(y|x)}{dy} = \frac{p_{X,Y}(x,y)}{p_X(x)}$$

Se X e Y forem independentes

$$\begin{aligned} p_{X,Y}(x,y) &= p_X(x) \cdot p_Y(y) \\ \Rightarrow p_Y(y|x) &= p_Y(y) \end{aligned} \tag{18}$$

- $p_{X|Y}(x|y) = \dfrac{p_{X,Y}(x,y)}{p_Y(y)}$
- $p_{Y|X}(y|x) = \dfrac{p_{X,Y}(x,y)}{p_X(x)}$
- $p_{Y|X}(y|x) = \dfrac{p_{X|Y}(x|y)p_Y(y)}{p_X(x)}$

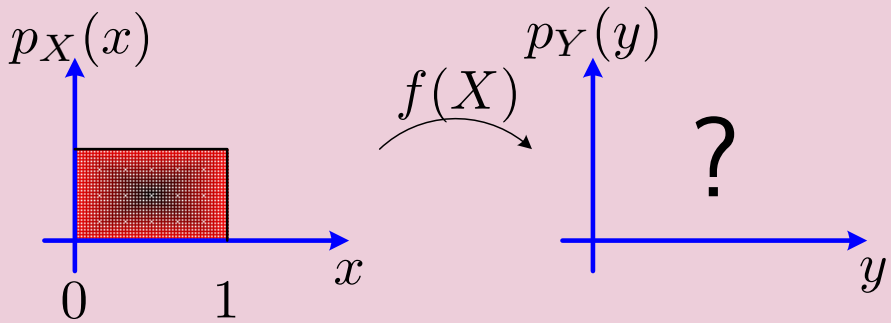
Função de uma variável aleatória

Dada $X \sim$ v.a. e uma função

$$Y = f(X)$$

a questão é como determinar $p_Y(y)$ conhecendo-se $p_X(x)$.
Por exemplo, seja X uma v.a. uniforme em $[0, 1]$

$$Y = \frac{1}{\lambda} \cdot \ln\left(\frac{1}{X}\right)$$



Vamos analisar dois casos

1o. caso

$$X \sim \text{v.a.} \quad p_X(x) \text{conhecido}$$

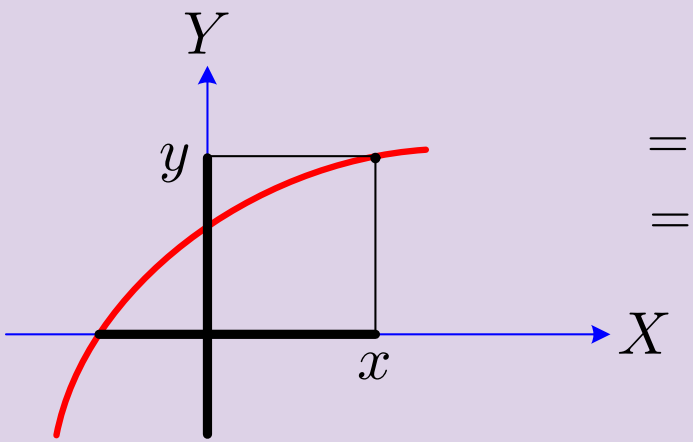
$$\begin{cases} y = f(x) \\ \frac{df(x)}{dx} > 0 \Rightarrow f(x) \text{ é monotônico crescente} \end{cases}$$

$$\rightarrow f \text{ é biunívoca} [p_Y(y), F_Y(y)]$$

$$F_Y(y) = \Pr\{Y \leq y\} = \Pr\{f(X) \leq y\}$$

Como

$$\frac{df(x)}{dx} > 0 \Rightarrow$$



$$= \Pr\{X \leq f^{-1}(y)\}$$
$$= F_X\left[f^{-1}(y)\right]$$

1o. caso - cont.

$$p_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X\left[f^{-1}(y)\right]}{dy}$$

$$= \frac{d}{dy}\int_{-\infty}^{f^{-1}(y)} p_X(x)\, dx$$

$$= p_X\left[f^{-1}(y)\right] \cdot \frac{df^{-1}(y)}{dy}$$

Mas: $\frac{df^{-1}(y)}{dy} = \frac{dx}{dy} = \frac{1}{\frac{dy}{dx}} = \frac{1}{\frac{df(x)}{dx}}$ Logo,

$$\boxed{p_Y(y) = p_X(x) \cdot \left.\frac{1}{\frac{df(x)}{dx}}\right|_{x=f^{-1}(y)} \quad \text{para} \quad \frac{df(x)}{dx} > 0}$$

2o. caso

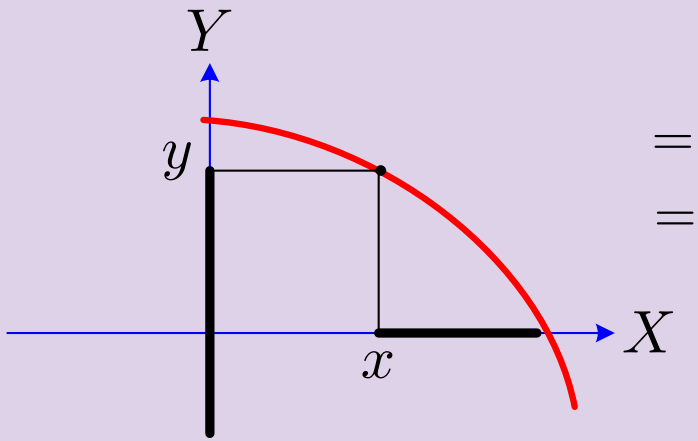
$$X \sim \text{v.a.} \quad p_X(x)\text{conhecido}$$

$$\begin{cases} y = f(x) \\ \frac{df(x)}{dx} < 0 \Rightarrow f(x) \text{ é monotônico decrescente} \end{cases}$$

$$\rightarrow f \text{ é biunívoca} [p_Y(y), F_Y(y)]$$

$$F_Y(y) = \Pr\{Y \leq y\}$$

$$F_Y(y) = \Pr\{f(X) \leq y\}$$



$$= \Pr\{X > f^{-1}(y)\}$$
$$= 1 - F_X\left[f^{-1}(y)\right]$$

2o. caso - cont.

$$p_Y(y) = \frac{F_Y(y)}{dy} = \left.\frac{-p_X(x)}{\frac{df(x)}{dx}}\right|_{x=f^{-1}(y)} \quad \text{com} \frac{df(x)}{dx} < 0$$

$$= \frac{p_X(x)}{\frac{df(x)}{dx}}$$

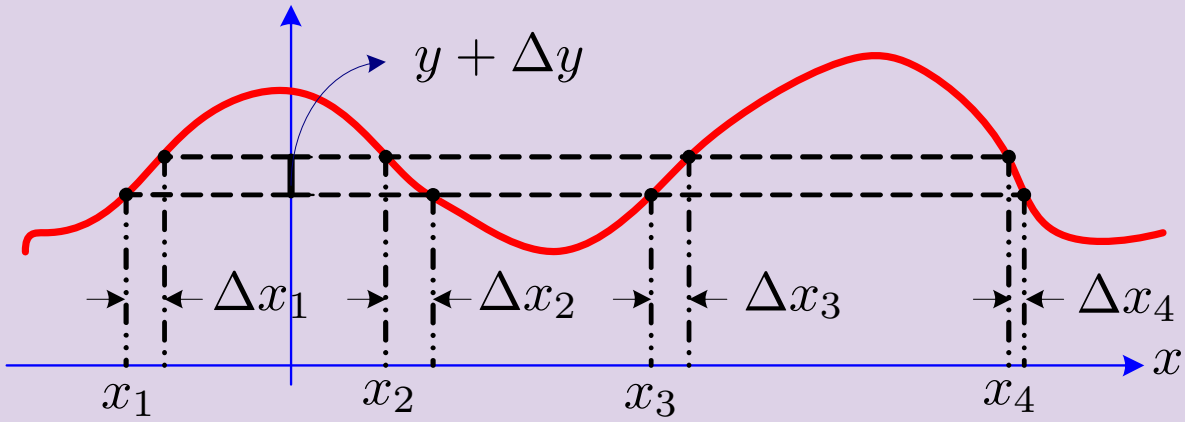
O sinal desaparece pois $\frac{df(x)}{dx}$ também é negativo.

Resumindo, para funções de uma variável aleatória temos:

Para encontrar $p_Y(y)$

$$\boxed{p_Y(y) = \left.\frac{p_X(x)}{\left|\frac{df(x)}{dx}\right|}\right|_{x=f^{-1}(y)}} \tag{19}$$

Funções não-biunívocas



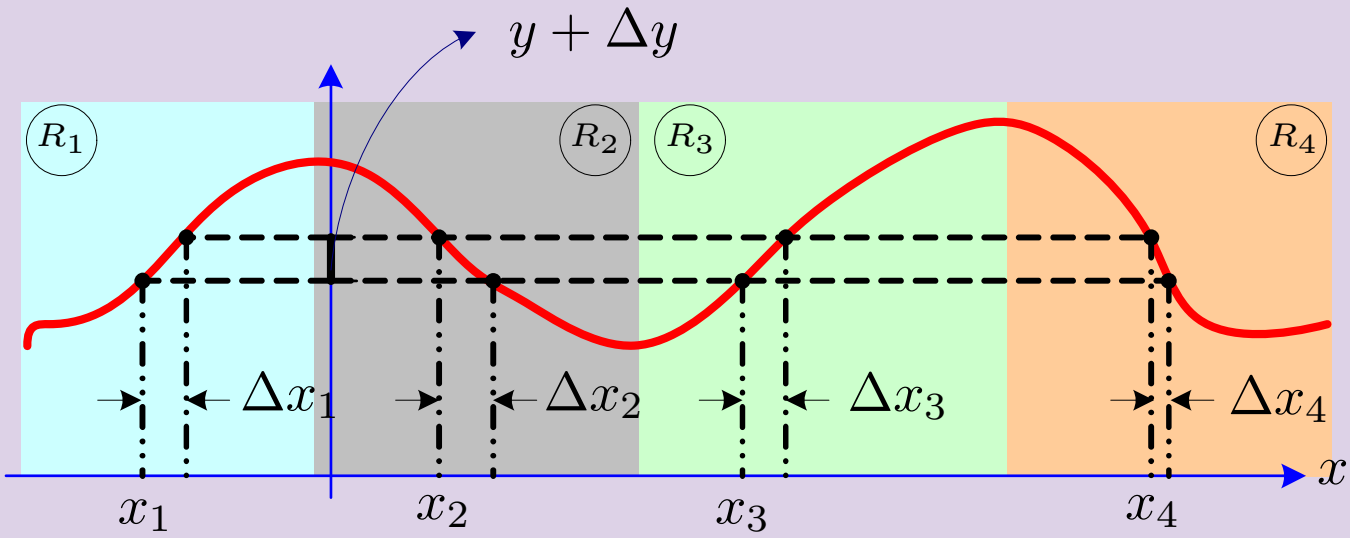
$$F_Y(y) = \Pr\{Y \leq y\}$$
$$= \Pr\{y \leq Y \leq y + \Delta y\}$$

Como há contribuições de diferentes intervalos de x, então

$$F_Y(y) = \Pr\{x_1 \leq X_1 \leq x_1 + \Delta x_1\} + \Pr\{x_2 - \Delta x_2 \leq X_2 \leq x_2\}$$
$$+ \Pr\{x_3 \leq X_3 \leq x_3 + \Delta x_3\} + \Pr\{x_4 - \Delta x_4 \leq X_4 \leq x_4\}$$

Funções não-biunívocas - cont.

Nestes casos, dividimos o espaço amostral em regiões biunívocas. Ou seja, teremos



Funções não-biunívocas - cont.

Nestes casos, chamando de

$$y = f_i(x_i) \quad \text{em cada região } R_i$$

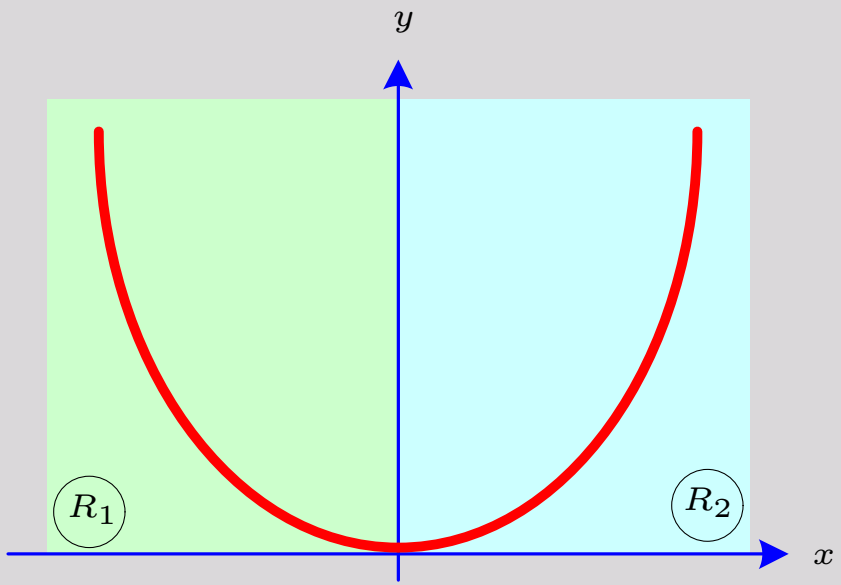
$p_Y(y)$

$$\boxed{p_Y(y) = \sum_{R_i} \left. \frac{p_X(x_i)}{\left| \frac{df_i(x_i)}{dx_i} \right|} \right|_{x_i = f^{-1}(y)}} \tag{20}$$

Exemplo

$$X \sim \text{v.a.} \quad N(0, \sigma^2)$$

$$Y = X^2$$



$$p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

Exemplo - cont.

$$p_Y(y) = \underbrace{\left.\frac{p_X(x)}{\left|\frac{df_1(x)}{dx}\right|}\right|_{x=f_1^{-1}(y)}}_{R_1} + \underbrace{\left.\frac{p_X(x)}{\left|\frac{df_2(x)}{dx}\right|}\right|_{x=f_2^{-1}(y)}}_{R_2}$$

$$\frac{df_1(x)}{dx} = \frac{dx^2}{dx} = 2x \quad x = -\sqrt{y}$$

$$\frac{df_2(x)}{dx} = \frac{dx^2}{dx} = 2x \quad x = \sqrt{y}$$

$$p_Y(y) = p_X(x)|_{x=-\sqrt{y}} \cdot \frac{1}{2\sqrt{y}} + p_X(x)|_{x=\sqrt{y}} \cdot \frac{1}{2\sqrt{y}}$$

$$= \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{y}{2\sigma^2}\right) \cdot \frac{1}{\sqrt{y}}, & y \geq 0 \\ 0, & y < 0 \end{cases}$$

Problema

$$X \sim \text{v.a.} \qquad Y \sim \text{v.a.}$$

$$\begin{cases} U = f(X,Y) \\ V = g(X,Y) \end{cases}$$

Conhecido $p_{X,Y}(x,y)$, como achar $p_{U,V}(u,v)$?

$p_{U,V}(u,v)$

Em regiões biunívocas:

$$\boxed{p_{U,V}(u,v) = \frac{p_{X,Y}(x,y)}{\left|J\left(\frac{u,v}{x,y}\right)\right|}\Bigg|_{\substack{x=\mathfrak{f}(u,v)\\ y=\mathfrak{g}(u,v)}}} \tag{21}$$

em que $\mathfrak{f}(\cdot)$ e $\mathfrak{g}(\cdot)$ são funções inversas. E

$$J\left(\frac{u,v}{x,y}\right) = \det\begin{bmatrix} \frac{du}{dx} & \frac{du}{dy} \\ \frac{dv}{dx} & \frac{dv}{dy} \end{bmatrix} = \frac{1}{\det\begin{bmatrix} \frac{d\mathfrak{f}}{du} & \frac{d\mathfrak{f}}{dv} \\ \frac{d\mathfrak{g}}{du} & \frac{d\mathfrak{g}}{dv} \end{bmatrix}}$$

é chamado *Jacobiano de u,v em relação a x,y*

Caso particular

$$Z = X + Y, \qquad X \sim \text{v.a.}, Y \sim \text{v.a.}$$

$$p_{X,Y}(x,y) \text{ conhecido}$$

$$p_Z(z) = ?$$

Definir então

$$\begin{cases} z = x + y \\ w = x \end{cases} \Rightarrow p_{Z,W}(z,w)$$

$$J\left(\frac{z,w}{x,y}\right) = \det\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} = -1$$

Logo,

$$p_{Z,W}(z,w) = \left.\frac{p_{X,Y}(x,y)}{|-1|}\right|_{\substack{x = \mathfrak{f}(z,w) \\ y = \mathfrak{g}(z,w)}}$$

Caso particular - cont.

$$x = w, \qquad y = z - w$$

$$p_{Z,W}(z,w) = p_{X,Y}(w, z-w)$$

Então, para achar a densidade marginal $p_Z(z)$, temos

$$p_Z(z) = \int_{-\infty}^{\infty} p_{X,Y}(w, z-w)\, dw$$

Caso particular - cont.

Se supormos que X e Y são independentes

$$p_{X,Y}(w, z-w) = p_X(w) \cdot p_Y(z-w)$$
$$\Rightarrow p_{Z,W}(z,w) = p_X(w) \cdot p_Y(z-w)$$
$$p_Z(z) = \int_{-\infty}^{\infty} p_{Z,W}(z,w)\, dw = \underbrace{\int_{-\infty}^{\infty} p_X(w) \cdot p_Y(z-w)\, dw}_{\text{convolução}}$$

Assim:

$$\boxed{p_Z(z) = p_x(x) \star p_Y(y)} \quad (22)$$

Em geral, podemos escrever a descrição de um modelo de entrada e saída de um modelo estocástico como

$$\left(\begin{array}{c}\text{valor atual}\\ \text{da saída}\\ \text{do modelo}\end{array}\right)+\left(\begin{array}{c}\text{combinação linear}\\ \text{dos valores passados}\\ \text{da saída do modelo}\end{array}\right)=\left(\begin{array}{c}\text{combinação linear}\\ \text{dos valores}\\ \text{presente e passados}\\ \text{da entrada do modelo}\end{array}\right) \quad (23)$$

Os processos que obedecem o comportamento acima são ditos **processos lineares**.

A estrutura do filtro linear que processará os dados, é determinada pela maneira que as duas combinações lineares indicadas na Eq. (23) são formuladas. Podemos então identificar três tipos usuais de modelos estocásticos lineares:

1. **Modelo auto-regressivo (AR)** - no qual não são utilizadas valores passados da entrada.
2. **Modelo moving average (MA)** - no qual não são usados valores passados da saída. Também chamado de modelo de média móvel.
3. **Modelo ARMA** - junção dos modelos AR e MA.

Dizemos que uma série temporal $x(n), x(n-1), \ldots, x(n-M)$ representa uma realização de um processo AR de ordem M se ela satisfaz a equação diferença seguinte:

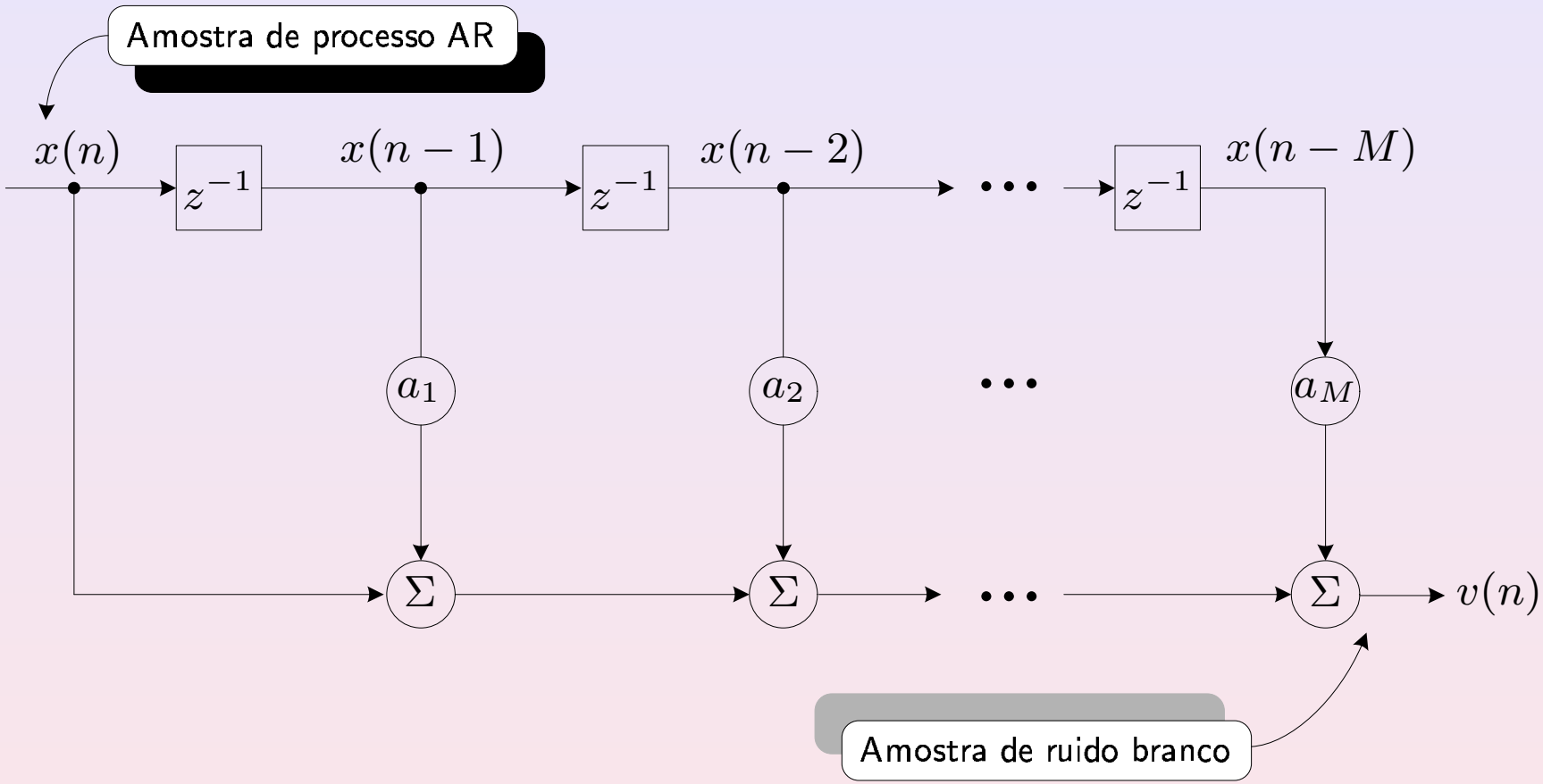
$$x(n) + a_1 x(n-1) + \cdots + a_M x(n-M) = v(n) \qquad (24)$$

em que $a_1, \ldots, a_M$ são chamados parâmetros AR e $v(n)$ é um processo de ruído branco.

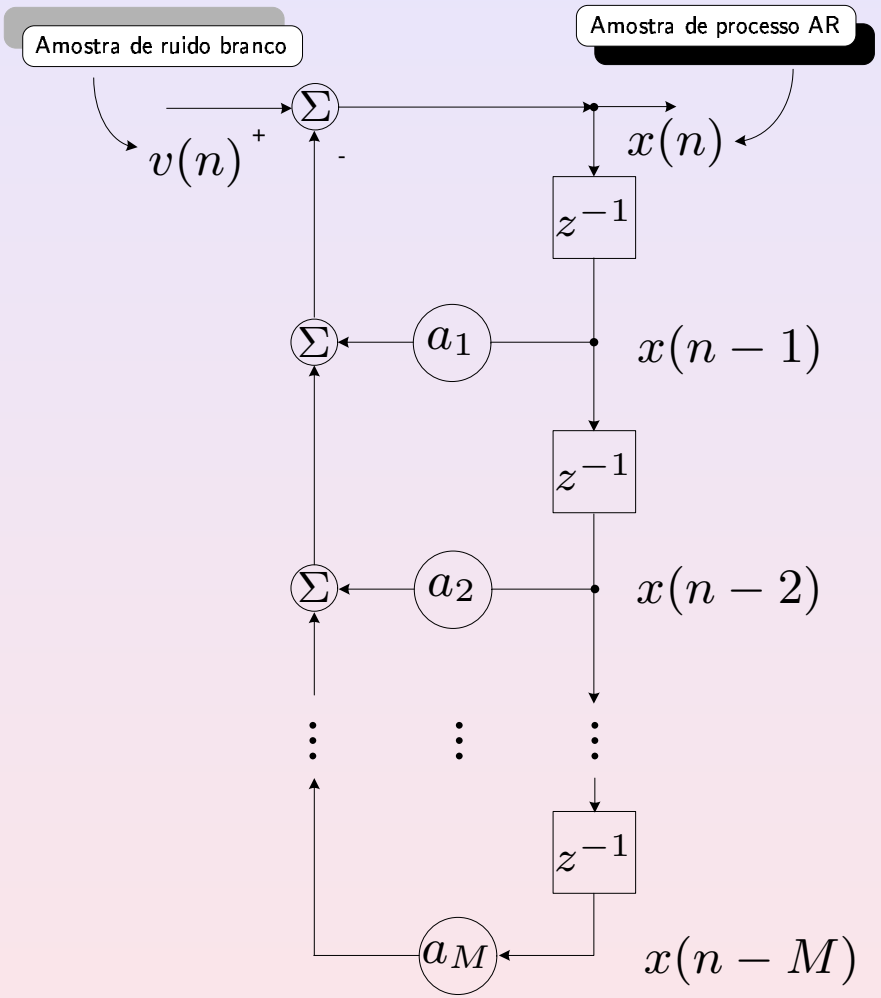
É mais simples enxergar o motivo de tal processo se chamar AR se escrevermos a Eq. (24) da seguinte forma:

$$x(n) = b_1 x(n-1) + b_2 x(n-2) + \cdots + b_M x(n-M) + v(n) \quad (25)$$

em que $b_k = -a_k$. Desta maneira vê-se facilmente que o instante atual do processo, ou seja $x(n)$ é igual a uma combinação dos valores passados do processo mais um termo de erro $v(n)$.



Analisador de processo AR



Gerador de processo AR

Em um modelo de média móvel (MA, *moving average*), o sistema é um filtro apenas com zeros e com ruído branco como entrada. O processo resultante $x(n)$ produzido é então dado pela seguinte equação diferença

$$x(n) = v(n) + b_1 v(n-1) + b_2 x(n-2) + \ldots + b_K x(n-K) \quad (26)$$

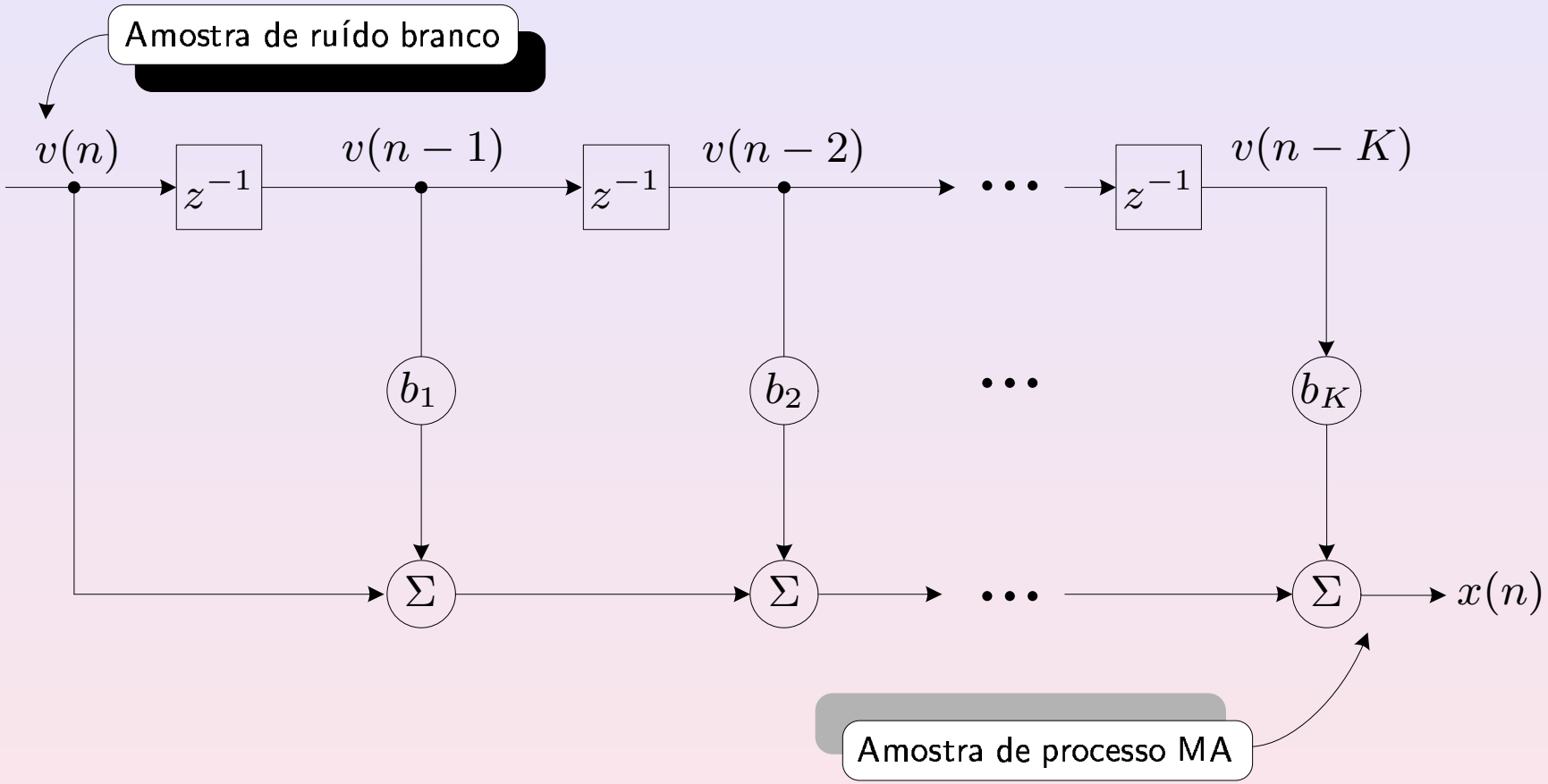
em que $b_1, \ldots, b_K$ são constantes chamadas de *parâmetros MA* e $v(n)$ é um processo de ruído branco de média zero e variância σ_v^2.

A Equação 26, representa uma versão escalar de um produto interno. Com isso, podemos representá-la como:

$$x(n) = \sum_{i=0}^{M} b_i v(n-i) = \mathbf{v}\mathbf{b}^T \tag{27}$$

em que $\mathbf{v} = [\ v(n)\ v(n-1)\ \ldots\ v(n-M)\ ]$ e
$\mathbf{b} = [\ 1\ b_1\ b_2\ \ldots\ b_K\ ]$.

A ordem do processo MA é dada por K. O termo média móvel surge pois constrói-se uma estimativa do processo x a partir de uma média ponderada das amostras do processo v.



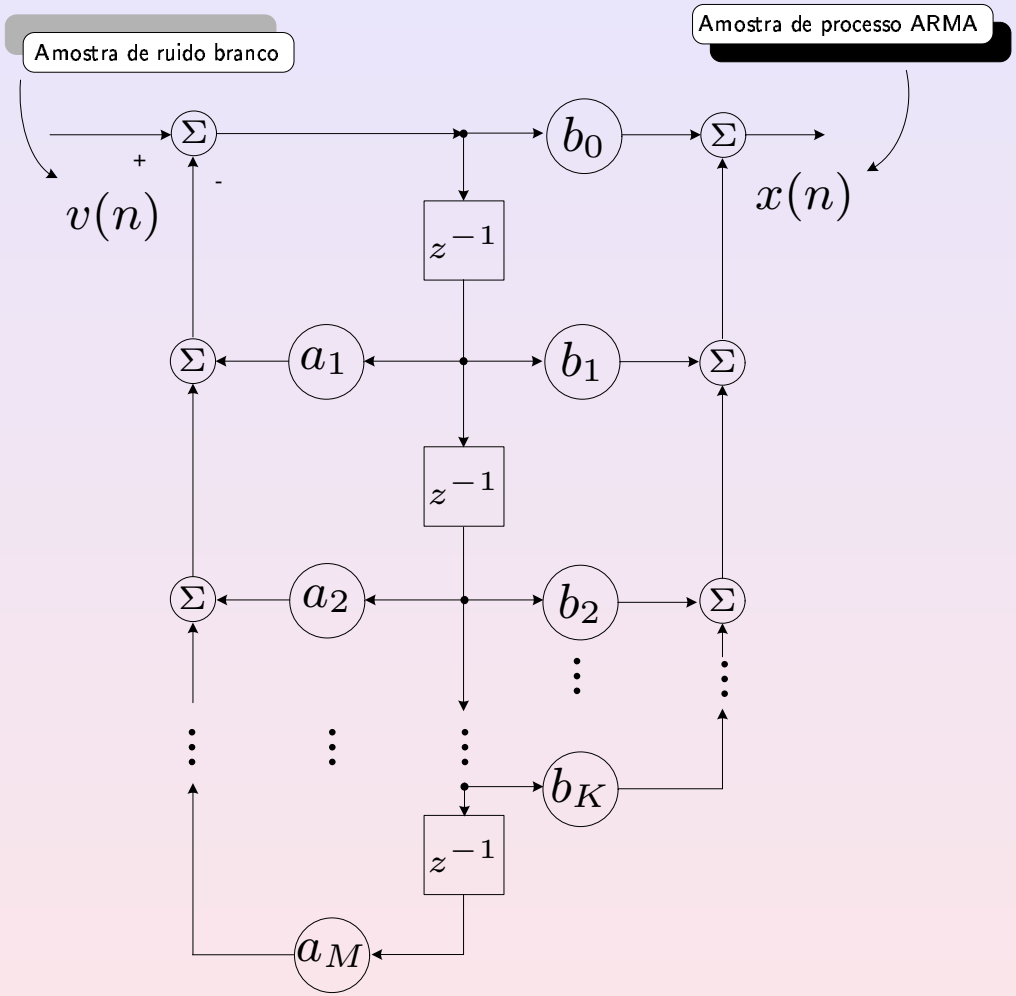
Modelo gerador de um processo de média móvel.

Para gerar um modelo **auto-regressivo-média móvel**, utilizando um processo de ruído branco como entrada, temos a seguinte equação diferença

$$\begin{aligned} x(n) + a_1 x(n-1) + \ldots + a_M x(n-M) = v(n) + b_1 v(n-1) \\ + b_2 v(n-2) + \ldots + b_K v(n-K) \end{aligned} \tag{28}$$

em que $a_1, \ldots, a_M$ e $b_1, \ldots, b_K$ são os parâmetros ARMA.

A ordem do modelo ARMA é dada por (M, K).



Modelo gerador de um processo ARMA de ordem (M, K), supondo $M > K$.

Parte II

Análise de momentos de segunda ordem

Correlação

Seja um vetor de v.a. $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix}$, temos que a *correlação* entre dois elementos, i e j, como

$$r_{ij} = \mathbb{E}\{x_i x_j\} = \int_{-\infty}^{\infty} x_i x_j^* p_{\mathbf{x}}(\mathbf{x})\, d\mathbf{x} = \int_{-\infty}^{\infty} x_i x_j^* p_{x_i,x_j}(x_i, x_j)\, dx_i\, dx_j \tag{29}$$

De uma maneira mais intuitiva, a correlação pode ser definida como a medida de relação entre as duas variáveis aleatórias. Note ainda que a mesma pode assumir valores positivos ou negativos.

Matrix de correlação

Podemos unificar uma notação para avaliar todos os pares de variáveis na correlação. Desta forma, definimos a *matriz de correlação* definida como

$$\begin{aligned}\mathbf{R_x} &= \mathbb{E}\left\{\mathbf{x}\mathbf{x}^H\right\} \\ &= \begin{bmatrix} \mathbb{E}\{|x_1|^2\} & \mathbb{E}\{x_1 x_2^*\} & \cdots & \mathbb{E}\{x_1 x_N^*\} \\ \mathbb{E}\{x_2 x_1^*\} & \mathbb{E}\{|x_2|^2\} & \cdots & \mathbb{E}\{x_2 x_N^* *\} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}\{x_N x_1^*\} & \mathbb{E}\{x_N x_2^*\} & \cdots & \mathbb{E}\{|x_N|^2\} \end{bmatrix}_{N\times N}\end{aligned} \tag{30}$$

Se escrevermos o vetor $\mathbf{x} = \begin{bmatrix} \mathbf{x}_r \\ \mathbf{x}_i \end{bmatrix}$ explicitando suas partes real $(\mathbf{x}_r)$ e imaginária $(\mathbf{x}_i)$, temos

$$\boxed{\begin{aligned} \mathbb{E}\left\{\mathbf{x}\mathbf{x}^H\right\} &= \begin{bmatrix} \mathbf{x}_r \\ \mathbf{x}_i \end{bmatrix} \begin{bmatrix} \mathbf{x}_r^H & \mathbf{x}_i^H \end{bmatrix} \\ &= \begin{bmatrix} \mathbb{E}\{\mathbf{x}_r\mathbf{x}_r^H\} & \mathbb{E}\{\mathbf{x}_r\mathbf{x}_i^H\} \\ \mathbb{E}\{\mathbf{x}_i\mathbf{x}_r^H\} & \mathbb{E}\{\mathbf{x}_i\mathbf{x}_i^H\} \end{bmatrix} \end{aligned}} \tag{31}$$

Desta forma, denotamos

$$\begin{aligned} \mathbb{E}\{\mathbf{x}_r\mathbf{x}_r^H\} = \mathbb{E}\{\mathbf{x}_i\mathbf{x}_i^H\} &= \mathbf{R}_{\mathbf{x}}^{\mathsf{E}} \\ \mathbb{E}\{\mathbf{x}_i\mathbf{x}_r^H\} = -\mathbb{E}\{\mathbf{x}_r\mathbf{x}_i^H\} &= \mathbf{R}_{\mathbf{x}}^{\mathsf{O}} \end{aligned} \tag{32}$$

em que E e O denotam as partes par (even) e ímpar (odd).

$\Rightarrow$ **Propriedades de $\mathbf{R_x}$**

1. É uma matriz *simétrica*: $\mathbf{R_x} = \mathbf{R_x}^H$
2. É uma matriz *semi-definida positiva*

$$\mathbf{a}^H \mathbf{R_x} \mathbf{a} \geq 0 \tag{33}$$

   para qualquer vetor N-dimensional $\mathbf{a} \neq \mathbf{0}$. Na prática, geralmente $\mathbf{R_x}$ é definida positiva para qualquer vetor N-dimensional $\mathbf{a} \neq \mathbf{0}$
3. Todos os autovalores de $\mathbf{R_x}$ são reais e *não-negativos* (positivos se $\mathbf{R_x}$ for definida positiva). Além disso, os autovetores de $\mathbf{R_x}$ são reais e podem sempre ser escolhidos tal que sejam *mutuamente ortogonais*.
4. $\mathbf{R_x} = 2\mathbf{R_x}^{\mathsf{E}} + j2\mathbf{R_x}^{\mathsf{O}}$

- Relembrando: momentos centrados são definidos de maneira similar aos momentos usuais, apenas a média é envolvida no cálculo da esperança.

Definimos então a **matriz de covariância** de $\mathbf{x}$ como

$$\boxed{\mathbf{C}_{\mathbf{x}} = \mathbb{E}\left\{(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^H\right\}} \tag{34}$$

e os elementos

$$c_{ij} = \mathbb{E}\left\{(x_i - \mu_i)(x_j - \mu_j)^H\right\} \tag{35}$$

da matriz $\mathbf{C}_{\mathbf{x}}$ de dimensão $N \times N$, são chamados de *covariâncias* e eles são os momentos centrados correspondentes às correlações r_{ij}.

- A matriz de covariância $\mathbf{C_x}$ possui as mesmas propriedades de simetria que a matriz de correlação $\mathbf{R_x}$.
- Usando as propriedades do operador esperança, é fácil mostrar que

$$\mathbf{R_x} = \mathbf{C_x} + \boldsymbol{\mu_x}\boldsymbol{\mu_x}^H \tag{36}$$

- Se o vetor de médias for $\boldsymbol{\mu_x} = \mathbf{0}$, as matrizes de correlação e covariância são as mesmas

Para esperanças conjuntas, ou seja, envolvendo mais de uma variável aleatória, podemos definir a **matriz de correlação cruzada**

$$\mathbf{R_{xy}} = \mathbb{E}\left\{\mathbf{x}\mathbf{y}^H\right\} \tag{37}$$

e a **matriz de covariância cruzada**

$$\mathbf{C_{xy}} = \mathbb{E}\left\{(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})(\mathbf{y}-\boldsymbol{\mu}_\mathbf{y})^H\right\} \tag{38}$$

Note que as dimensões dos vetores $\mathbf{x}$ e $\mathbf{y}$ podem ser diferentes. Assim, as matriz de correlação e covariância cruzadas não são necessariamente quadradas e são, em geral, não-simétricas. Entretanto, de suas definições segue que:

$$\begin{aligned}\mathbf{R_{xy}} &= \mathbf{R}_{\mathbf{yx}}^H \\ \mathbf{C_{xy}} &= \mathbf{C}_{\mathbf{yx}}^H\end{aligned} \tag{39}$$

Quando temos uma soma de dois vetores $\mathbf{x}$ e $\mathbf{y}$, temos as seguintes relações:

1. Correlação

$$\mathbf{R_{x+y}} = \mathbf{R_x} + \mathbf{R_{xy}} + \mathbf{R_{yx}} + \mathbf{R_y} \tag{40}$$

2. Covariância

$$\mathbf{C_{x+y}} = \mathbf{C_x} + \mathbf{C_{xy}} + \mathbf{C_{yx}} + \mathbf{C_y} \tag{41}$$

Vale relembrar que:

- Variáveis ortogonais implica em correlação zero $(\mathbf{R_{xy}} = \mathbf{0})$
- Variáveis descorrelacionadas implica em covariância zero, $(\mathbf{C_{xy}} = \mathbf{0})$

Então, temos

1. $\mathbf{R_{x+y}} = \mathbf{R_x} + \mathbf{R_y}$ para $\mathbf{x}$ e $\mathbf{y}$ ortogonais
2. $\mathbf{C_{x+y}} = \mathbf{C_x} + \mathbf{C_y}$ para $\mathbf{x}$ e $\mathbf{y}$ descorrelacionados

Transformações tempo-freqüência

- Transformação tempo-freqüência diz respeito à transformada de Fourier quando estudamos sinais determinísticos
- A transformação requer o conhecimento da função do sinal a ser avaliada na freqüência

$$\begin{aligned} S(\omega) &= \mathfrak{F}\{s(t)\} \\ &= \int_{-\infty}^{\infty} s(t) \exp(-j\omega t)\ dt \end{aligned} \tag{42}$$

- **Pergunta**: como fazer no caso de sinais aleatórios? Não se conhece os valores do sinal com precisão e a integral pode não existir.
- **Resposta**: definir de outra forma a transformada de Fourier de um processo aleatório

Classificação de processos estocásticos

1. Estacionários no *sentido estrito* (SSS)
   - TODAS as estatísticas (momentos) são independentes do tempo, ou seja,
   $$p_{X(t)}[x(t)] = p_{X(t)}(x)$$
2. Estacionários no *sentido amplo* (WSS)
   - Apenas as estatísticas de primeira e segunda ordem (média e correlação) são independentes do tempo, ou seja,

$$\mathbb{E}\{X(t)\} = \mu$$
$$\mathbb{E}\left\{(X(t)-\mu)^2\right\} = \sigma_x^2$$

Definição

Para a classe de processos estocásticos WSS definimos a **função de densidade espectral** como a transformada de Fourier da correlação na forma

$$\begin{aligned} S_x(\omega) &= \mathfrak{F}\{r_x(\tau)\} \\ &= \int_{-\infty}^{\infty} r_x(\tau) \exp(-j\omega\tau)\, d\tau \end{aligned} \tag{43}$$

A função de densidade espectral de potência é um indicador da distribuição da potência do sinal como uma função da freqüência.

Pode-se ainda calcular a correlação a partir da função de densidade espectral, uma vez que a transformada de Fourier é única, como

$$r_x(\tau) = \mathfrak{F}^{-1}\{S_x(\omega)\} = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_x(\omega) \exp(j\omega\tau)\ d\omega \qquad (44)$$

NOTA: Valem ressaltar que τ significa a diferença entre os tempos que os processos WSS requerem como parâmetro para caracterizar a correlação. Neste caso, seria o equivalente à diferença entre as amostras i e j que temos na correlação definida na Equação (29)

Propriedades

1. O valor médio quadrático de um processo WSS é dado por

$$\mathbb{E}\{X^2(t)\} = r_x(0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_x(\omega)\, d\omega \tag{45}$$

2. A densidade espectral de potência de um processo WSS é sempre não-negativa

$$S_x(\omega) \geq 0 \quad \text{para todo } \omega \tag{46}$$

Propriedades - cont.

3. A densidade espectral de potência de um processo WSS real é uma função par de ω

$$S_x(\omega) = S_x(-\omega) \tag{47}$$

4. O valor da densidade espectral de potência em $\omega = 0$ é

$$S_x(0) = \int_{-\infty}^{\infty} r_x(\tau)\, d\tau \tag{48}$$

Expansão de Karhunen-Loève

- A expansão de Karhunen-Loève (KL) é um tipo de expansão em série de um processo estocástico
- A meta é representar o processo como uma soma de funções ortonormais
- É possível mostrar para o caso contínuo (ver Papoulis, 1991 - pp. 412), mas aqui abordaremos apenas o caso discreto

Definição

Seja o vetor $\mathbf{x}(n)$ de dimensão $N \times 1$ que denota uma seqüência escolhida de um processo WSS de média zero e matriz de correlação $\mathbf{R_x}$. Sejam $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_N$ os autovetores associados com os N autovalores da matriz $\mathbf{R_x}$. O vetor $\mathbf{x}(n)$ pode ser expandido como uma combinação linear destes autovetores como

$$\boxed{\mathbf{x}(n) = \sum_{i=1}^{N} c_i(n)\mathbf{q}_i} \tag{49}$$

Os coeficientes da expansão são v.a. de média zero e descorrelacionadas definidas pelo produto interno

$$c_i(n) = \mathbf{q}_i^H \mathbf{x}(n)$$

Expansão de Karhunen-Loève - cont.

- De um ponto de vista físico, podemos interpretar os autovetores $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_N$ como sendo as coordenadas de um espaço N-dimensional e a expansão KL será uma projeção do vetor $\mathbf{x}(n)$no conjunto de suas projeções $c_1(n), c_2(n), \ldots, c_N(n)$.
- Além disso, deduz-se que

$$\sum_{i=1}^{N} |c_i(n)|^2 = \|\mathbf{x}(n)\|^2 \tag{50}$$

em que $\| \bullet \|$ é a norma euclideana.

- A equação acima implica que o coeficiente $c_i(n)$ tem a mesma energia que o vetor $\mathbf{x}(n)$ observado na i-ésima coordenada
- Tal energia é, lógicamente, uma v.a. cuja média é igual ao i-ésimo autovalor, ou seja

$$E\{|c_i(n)|^2\} = \lambda_i, \quad i = 1, 2, \ldots, N \tag{51}$$

Transformada de Karhunen-Loève (KLT)

- É uma transformação matemática que determina a combinação linear que maximiza a variância dos dados em um certo número de dimensões
- Com esta transformação a maior variância é encontrada na primeira dimensão, a segunda maior variância na segunda dimensão e assim por diante
- É também chamada de **Principal Component Analysis (PCA)**
- A meta é então encontrar uma dimensão $K < N$ de tal forma que os dados sejam adequadamente representados com um menor número de parâmetros
- Cálculo das direções principais é feito por meio da matriz de correlação $\mathbf{R_x}$

Transformada de Karhunen-Loève - cont.

Podemos escrever, através do uso da *Singular Value Decomposition (SVD)* a matriz $\mathbf{R_x}$ como

$$\mathbf{R_x} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^H \tag{52}$$

em que $\mathbf{U}$ e $\mathbf{V}$ são matrizes retangulares de ordem $K \times N$, tais que $\mathbf{U}\mathbf{U}^T = \mathbf{V}\mathbf{V}^T = \mathbf{I}_N$ e $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_K$
Assim, a transformada de Karhunen-Loève do vetor $\mathbf{x}(n)$ é dada por

$$\boxed{\overline{\mathbf{x}}(n) = \mathbf{\Lambda}\mathbf{V}^H\mathbf{x}(n)} \tag{53}$$

Observação 1: PCA/KLT não tem uma base única de vetores, a base depende dos dados
Observação 2: Em estatística multivariável é geralmente chamada de **Transformação de Mahalanobis**

Transformada de Karhunen-Loève - cont.

Na verdade, a Equação (53) projeta os dados em **todas** direções ortogonais formadas pela matriz de correlação.
Para usarmos $K < N$ dimensões, devemos usar somente os K autovetores associados aos K maiores autovalores. Desta forma teremos então

$$\overline{\mathbf{x}}(n) = \mathbf{\Lambda}_K \mathbf{V}_K^H \mathbf{x}(n) \tag{54}$$

em que as matrizes $\mathbf{\Lambda}_K$ e $\mathbf{V}_K$ são, respectivamente, a matriz diagonal com os K maiores autovalores e os seus K autovetores associados.
Uma informação importante é notar que a KLT/PCA torna os dados *brancos*, ou seja, $\mathbf{R}_{\overline{\mathbf{x}}} = \mathbf{I}_K$. Isto será de importância em vários problemas de processamento estatístico de sinais.

Momentos de ordem superior

- São estatísticas de **ordem superior a 2**
- Descrevem o comportamento estatístico dos dados com alguma informação a mais
- São importante por portarem informação da fase dos dados
- O momento de ordem 3 recebe o nome de obliquidade (*skewness* em inglês) e mede a assimetria da distribuição
- O momento de ordem 4 recebe o nome de **kurtosis** e é denotada por $\mathcal{K}$
- Podemos também classificar as distribuições através da kurtosis
  1. Para $\mathcal{K} = 0$: distribuição *mesocúrtica*
  2. Para $\mathcal{K} > 0$: distribuição *leptocúrtica*
  3. Para $\mathcal{K} < 0$: distribuição *platicurtica*

Primeira função característica

$$C(\omega) \triangleq \int_{-\infty}^{\infty} p_X(x) \cdot \exp(j\omega x)\, dx \tag{55}$$

Importante

- $C(-\omega) \triangleq \int_{-\infty}^{\infty} p_X(x) \cdot \exp(-j\omega x)\, dx$
- $C(-\omega) = \mathfrak{F}\{p_X(x)\}$ (transformada de Fourier de $p_X(x)$)

Importante - cont.

- Notação

$$P_X(\omega) = \mathfrak{F}\{p_X(x)\} = \int_{-\infty}^{\infty} p_X(x) \cdot \exp(-j\omega x)\ dx$$

$$\mathbb{E}\{\exp(-j\omega X)\}$$

- Assim

$$\boxed{p_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} P_X(\omega) \cdot \exp(j\omega x)\ d\omega}$$

Primeira função característica - Variáveis discretas

$$P_X(w) = \sum_{i=-\infty}^{\infty} \Pr\{X = x_i\} \cdot \exp\left(-j\omega x_i\right)$$
$$= \mathbb{E}\{\exp(-j\omega X)\}$$

Propriedades $P_X(\omega)$

1. $|P_X(\omega)| \leq 1$
   Prova:

$$|P_X(\omega)| = \left| \int_{-\infty}^{\infty} p_X(x) \cdot \exp(-j\omega x)\ dx \right|$$

$$\leq \int_{-\infty}^{\infty} |p_X(x)| \cdot |\exp(-j\omega x)|\ dx = \int_{-\infty}^{\infty} p_X(x)\ dx$$

desigualdade de Schwartz

2. $P_X(0) = 1$

Função geradora de momentos

$$P_X(\omega) = \mathbb{E}\{\exp(-j\omega X)\}$$

$$\exp(-j\omega x) = \sum_{k=-\infty}^{\infty} \frac{(-j\omega X)^k}{k!} \quad \text{(Série de Taylor em torno de } X = 0)$$

$$\mathbb{E}\{\exp(-j\omega x)\} = \mathbb{E}\left\{ \sum_{k=-\infty}^{\infty} \frac{(-j\omega)^k \cdot X^k}{k!} \right\}$$

$$\mathbb{E}\{\exp(-j\omega x)\} = \sum_{k=-\infty}^{\infty} \frac{(-j\omega)^k}{k!} \cdot \underbrace{\mathbb{E}\left\{X^k\right\}}_{\mu_k}$$

$$\Rightarrow \boxed{P_X(\omega) = \sum_{k=-\infty}^{\infty} \frac{(-j\omega)^k}{k!} \cdot \mu_k}$$

Função geradora de momentos - cont.

Ainda

$$P_X(\omega) = \sum_{k=0}^{\infty} \left. \frac{d^k P_X(\omega)}{d\omega^k} \right|_{\omega=0} \cdot \frac{1}{k!} \cdot \omega^k$$

Logo

$$\sum_{k=0}^{\infty} \left. \frac{d^k P_X(\omega)}{dw^k} \right|_{\omega=0} \cdot \frac{\omega^k}{k!} = \sum_{k=0}^{\infty} \mu_k \cdot (-j)^k \cdot \frac{\omega^k}{k!}$$

$$\boxed{\mu_k \cdot (-j)^k = \left. \frac{d^k P_X(\omega)}{dw^k} \right|_{\omega=0}}$$

Exemplo

Sejam X e Y v.a.s independentes com $p_X(x)$ e $p_Y(y)$ conhecidas. Se $Z = X + Y$, $p_Z(z) = ?$

Solução

$$P_Z(\omega) = \mathbb{E}\{\exp(-j\omega Z)\} = \mathbb{E}\{\exp[-j\omega(X+Y)]\}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(-j\omega X)\cdot\exp(-j\omega Y)\cdot p_{X,Y}(x,y)\,dx\,dy$$

$$\underbrace{\int_{-\infty}^{\infty} \exp(-j\omega X)\cdot p_X(x)\,dx}_{\mathbb{E}\{\exp(-j\omega X)\}}\cdot\underbrace{\int_{-\infty}^{\infty} \exp(-j\omega Y)\cdot p_Y(y)\,dy}_{\mathbb{E}\{\exp(-j\omega Y)\}}$$

$$P_Z(\omega) = P_X(\omega)\cdot P_Y(\omega)$$

$$\boxed{p_Z(z) = p_X(x) \star p_Y(y)}$$

Segunda função característica

$$\Psi(\omega) = \ln[P_X(\omega)] \tag{56}$$

Importante

- A segunda função característica é também chamada de **função geradora de cumulantes**
- Os cumulantes são de extrema importância na caracterização estatística de uma v.a.

História

Os cumulantes foram inicialmente introduzidos pelo astrônomo, contador, matemático e estaticista dinamarquês Thorvald N. Thiele (1838-1910) que os denominou *semi-invariantes*.

O termo *cumulante* surgiu pela primeira vez em 1931 no artigo "The Derivation of the Pattern Formulæ of Two-Way Partitions from Those of Simpler Patterns", Proceedings of the London Mathematical Society, Series 2, vol. 33, pp. 195-208, publicado pelo geneticista e estaticista Sir Ronald Fisher e o estaticista John Wishart, epônimo da distribuição de Wishart.

O historiador Stephen Stigler comenta que o termo cumulante foi sugerido a Fisher numa carta de Harold Hotelling. Em um outro artigo publicado em 1929, Fisher chamou-os de funções de momentos cumulativos.

Definição

O cumulante de ordem k é definido como

$$\boxed{\kappa_k = \frac{\partial^k \Psi(\omega)}{\partial \omega^k}} \tag{57}$$

Propriedades dos cumulantes

1. **Invariância e equivariância**

$$\kappa_1(Y+\alpha) = \kappa_1(Y) + \alpha$$
$$\kappa_k(Y+\alpha) = \kappa_k(Y)$$

para α uma constante qualquer.

2. **Homogeneidade** (ou multilinearidade)

$$\kappa_k(\alpha Y) = \alpha^k \cdot \kappa_k(Y)$$

3. **Aditividade**

$$\kappa_k(X+Y) = \kappa_k(X) + \kappa_k(Y)$$

se X e Y são v.a.s independentes

Cumulantes e momentos

Os cumulantes são relacionados com os momentos através da seguinte recursão:

$$\kappa_k = \mu_k - \sum_{i=1}^{k-1} \binom{k-1}{i-1} \kappa_i \cdot \mu_{k-i} \tag{58}$$

Cumulantes e momentos - cont.

Desta forma, o $k-$ésimo momento é um polinômio de grau k dos k primeiros cumulantes, dados, para o caso em que $k = 6$, na seguinte forma:

$$\mu_1 = \kappa_1$$
$$\mu_2 = \kappa_2 + \kappa_1^2$$
$$\mu_3 = \kappa_3 + 3\kappa_2\kappa_1 + \kappa_1^3$$
$$\mu_4 = \kappa_4 + 4\kappa_3\kappa_1 + 3\kappa_2^2 + 6\kappa_2\kappa_1^2 + \kappa_1^4$$
$$\mu_5 = \kappa_5 + 5\kappa_4\kappa_1 + 10\kappa_3\kappa_2 + 10\kappa_3\kappa_1^2 + 15\kappa_2^2\kappa_1 + 10\kappa_2\kappa_1^3$$
$$\mu_6 = \kappa_6 + 6\kappa_5\kappa_1 + 15\kappa_4\kappa_2 + 15\kappa_4\kappa_1^2 + 10\kappa_3^2 + 60\kappa_3\kappa_2\kappa_1 + 20\kappa_3\kappa_1^3 + 15\kappa_2^3 + 45\kappa_2^2\kappa_1^2 + 15\kappa_2\kappa_1^4 + \kappa_1^6.$$

Parte III

Teoria da Estimação

Teoria da estimação

- **Pergunta:** O que é estimação?
- **Resposta:** Encontrar quantidades de interesse num dado conjunto de medidas ruidosas (com incerteza).
- **Tipos de quantidades:** determinísticas ou aleatórias, linear ou não-linear, variantes ou invariantes no tempo.
- **Possíveis técnicas:** grande variedade com diferentes características - ótimas ou subótimas
- **Custo computacional:** associado ao tipo de estratégia, mas geralmente as soluções ótimas são mais complexas e as subótimas menos complexas.

Sejam N medidas escalares $x(0), x(1), \ldots, x(N-1)$ contendo informações sobre K quantidades $\theta_1, \theta_2, \ldots, \theta_K$ que deseja-se estimar. Tais quantidades são denominadas **parâmetros**.

Em uma notação mais compacta podemos escrever o **vetor de dados** ou **vetor de medidas**,

$$\mathbf{x}_N = \left[\begin{array}{cccc} x(0) & x(1) & \ldots & x(N-1) \end{array} \right]^T \tag{59}$$

e o **vetor dos parâmetros** como

$$\boldsymbol{\theta} = \left[\begin{array}{cccc} \theta_1 & \theta_2 & \ldots & \theta_K \end{array} \right]^T \tag{60}$$

Definição

De maneira bastante geral, um **estimador** $\widehat{\boldsymbol{\theta}}$ do vetor de parâmetros $\boldsymbol{\theta}$ é uma função matemática pela qual os parâmetros podem ser estimados por meio das medidas:

$$\boxed{\widehat{\boldsymbol{\theta}} = \mathbf{h}\left[\mathbf{x}_N\right] = \mathbf{h}\left[x(0), x(2), \ldots, x(N-1)\right]} \tag{61}$$

Para os parâmetros individuais temos

$$\widehat{\theta}_i = h_i\left[\mathbf{x}_N\right], \quad i = 1, 2, \ldots, K \tag{62}$$

Se os parâmetros forem de tipos diferentes, a equação acima pode ser bastante diferente para diferentes i. Ou seja, os componentes h_i da função vetorial $\mathbf{h}$ podem ter diferentes formas funcionais. O valor numérico de um estimador $\widehat{\theta}_i$ é chamado de **estimativa** do parâmetro θ_i.

Exemplo

Dois parâmetros geralmente necessários são a média μ e variância σ^2 de uma v.a. x. Dado um vetor de dados, eles podem ser estimados por meio das seguintes fórmulas bem conhecidas:

$$\widehat{\mu} = \frac{1}{N}\sum_{j=0}^{N-1} x(j) \tag{63a}$$

$$\widehat{\sigma}_2 = \frac{1}{N}\sum_{j=0}^{N-1} [x(j) - \widehat{\mu}]^2 \tag{63b}$$

- Um bom estimador deve respeitar algumas propriedades que facilitam a determinação das quantidades de interesse
- A medida de qualidade de um estimador é baseada no erro de estimação, que é definido como

$$\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}} = \boldsymbol{\theta} - \mathbf{h}\left[\mathbf{x}_N\right] \tag{64}$$

- Idealmente, o erro de estimação $\widetilde{\boldsymbol{\theta}}$ deve ser zero, mas é impossível atingir tal critério tão restrito para um conjunto finito de dados. Desta forma, devemos considerar um critério de desempenho menos restritivo.

Polarização e consistência

O primeiro requisito é que o valor médio do erro $E\{\widetilde{\boldsymbol{\theta}}\}$ deve ser zero. Assim, usando o operador esperança em ambos os lados da Eq. (64), temos a seguinte condição

$$\mathbb{E}\{\widehat{\boldsymbol{\theta}}\} = \mathbb{E}\{\boldsymbol{\theta}\} \tag{65}$$

Estimadores que satisfazem a Eq. (65) são chamados de *não-polarizados*. A definição acima é aplicável para parâmetros aleatórios. Para parâmetros não aleatórios a definição é dada como

$$\mathbb{E}\{\widehat{\boldsymbol{\theta}}|\boldsymbol{\theta}\} = \boldsymbol{\theta} \tag{66}$$

Geralmente densidade e esperanças condicionais, condicionadas pelo vetor de parâmetros $\boldsymbol{\theta}$ são usadas para tratar com parâmetros determinísticos. Neste caso, as esperanças são tomadas apenas sobre o vetor de dados.

Polarização e consistência - cont.

Se um estimador não obedece as condições (65) ou (66) ele é dito ser polarizado. Em particular, a polarização $\mathbf{b}$ é definida como o valor médio do erro de estimação:

$$\mathbf{b} = \mathbb{E}\{\widetilde{\boldsymbol{\theta}}\}, \quad \text{ou} \quad \mathbf{b} = \mathbb{E}\{\widetilde{\boldsymbol{\theta}}|\boldsymbol{\theta}\} \tag{67}$$

Se a polarização se aproxima de zero quando o número de medidas tende para infinito, o estimador é então dito ser **assintoticamente não-polarizado**.
Uma medida razoável para um bom estimador $\widehat{\boldsymbol{\theta}}$ é que ele deve convergir para o valor verdadeiro do vetor de parâmetros θ quando o número de medidas tende para infinito. Estimadores que satisfazem a esta propriedade assintótica são chamados de **consistentes**. Estimadores consistentes **não** são necessariamente **não-polarizados**.

Polarização e consistência - cont.

Exemplo: Média e variância

O valor esperado da média amostral é

$$\mathbb{E}\{\widehat{\mu}\} = \frac{1}{N}\sum_{j=0}^{N-1}\mathbb{E}\{x(j)\} = \frac{1}{N}N\mu = \mu \tag{68}$$

Logo, o estimador é não-polarizado. É também consistente pois

$$\mathbb{E}\{(\widehat{\mu}-\mu)^2\} = \frac{1}{N^2}\sum_{j=0}^{N-1}\mathbb{E}\{[x(j)-\widehat{\mu}]^2\} = \frac{1}{N^2}N\sigma^2 = \frac{\sigma^2}{N} \tag{69}$$

A variância se aproxima de zero quando $N \to \infty$, implicando que a média amostral converge, em probabilidade, para o valor correto μ.

Erro médio quadrático

É útil introduzir uma *função de perda* $L(\widetilde{\boldsymbol{\theta}})$ para descrever a importância relativa de específicos erros de estimação. Uma função de perda popular é o **erro quadrático de estimação** $L(\widetilde{\boldsymbol{\theta}}) = \|\widetilde{\boldsymbol{\theta}}\|^2 = \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|^2$ devido sua tratabilidade matemática.

O erro de estimação $\widehat{\boldsymbol{\theta}}$ é uma v.a. que depende do vetor de dados $\mathbf{x}_N$, logo, $L(\widetilde{\boldsymbol{\theta}})$ é também uma v.a. Para obter um uma medida de erro não-aleatória é útil definir o *índice de desempenho* ou *critério de erro* $\mathcal{E}$ como a esperança da função de perda. Assim,

$$\underbrace{\mathcal{E} = \mathbb{E}\{L(\widetilde{\boldsymbol{\theta}})\}}_{\text{aleatório}} \quad \text{ou} \quad \underbrace{\mathcal{E} = \mathbb{E}\{L(\widetilde{\boldsymbol{\theta}})|\boldsymbol{\theta}\}}_{\text{determinístico}} \tag{70}$$

Erro médio quadrático - cont.

Uma critério de erro amplamente usado é o erro médio quadrático (MSE, do inglês)

$$\mathcal{E}_{\text{MSE}} = \mathbb{E}\left\{\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|^2\right\} \tag{71}$$

Se o MSE tende assintoticamente para zero com o aumento do número de medidas, o respectivo estimador é consistente. Outra importante propriedade é que o MSE pode ser decomposto como

$$\mathcal{E}_{\text{MSE}} = \underbrace{\mathbb{E}\left\{\|\widetilde{\boldsymbol{\theta}} - \mathbf{b}\|^2\right\}}_{\text{variância de } \widetilde{\boldsymbol{\theta}}} + \underbrace{\|\mathbf{b}\|^2}_{\text{polarização}} \tag{72}$$

Se o estimador é não-polarizado então o MSE coincide com a variância do estimador.

Erro médio quadrático - cont.

Uma outra medida útil da qualidade do estimador é dada pela matriz de covariância do erro de estimação

$$\mathbf{C}_{\widetilde{\boldsymbol{\theta}}} = \mathbb{E}\left\{\widetilde{\boldsymbol{\theta}}\widetilde{\boldsymbol{\theta}}^H\right\} = \mathbb{E}\left\{(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^H\right\} \tag{73}$$

Ela mede os erros das estimativas individuais dos parâmetros enquanto o MSE mede a média dos erros de **todos** os parâmetros. De fato, o MSE pode ser obtido somando os termos da diagonal da matriz $\mathbf{C}_{\widetilde{\boldsymbol{\theta}}}$.

Robustez

- Uma importante medida de um estimador é a sua robustez. De uma maneira muito grosseira, robustez significa a insensibilidade a grandes erros de medida, na especificação dos modelos dos parâmetros.
- Uma problema típico com estimadores é que eles são sensíveis a *outliers*
- Consideração de critérios que cresçam de maneira não-quadrática (menor) que o erro geralmente são usadas.
- Questão de *arte* e conhecimento do problema.

Eficiência

- Um estimador é dito ser **eficiente** se ele fornece a menor matriz de covariância do erro de estimação entre todos os estimadores não-polarizados otimizados sob algum critério.
- Ele usa de forma ótima a informação contida nas medidas.
- Uma matriz $\mathbf{A}$ é dita ser menor que outra matriz simétrica $\mathbf{B}$, ou seja, $\mathbf{A} < \mathbf{B}$, se a matriz $\mathbf{B} - \mathbf{A}$ é definida positiva.
- **Resultado importante:** existe um limite inferior para a matriz de covariância. Este é o chamado **Limite de Cramér-Rao**.

Teorema

Se $\widehat{\boldsymbol{\theta}}$ é um estimador qualquer não-polarizado de $\boldsymbol{\theta}$ baseado nos dados medidos $\mathbf{x}_N$, então a matriz de covariância do erro de estimação é limitada inferiormente pela inversa da **matriz de informação de Fisher** $\mathbf{J}$, ou seja,

$$\boxed{E\left\{\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\right)^{H}|\boldsymbol{\theta}\right\}\geq\mathbf{J}^{-1}} \tag{74}$$

em que

$$\mathbf{J}=\mathbb{E}\left\{\left[\frac{\partial}{\partial\boldsymbol{\theta}}\ln\left[p_{\mathbf{x}_N|\boldsymbol{\theta}}(\mathbf{x}_N|\boldsymbol{\theta})\right]\right]\left[\frac{\partial}{\partial\boldsymbol{\theta}}\ln\left[p_{\mathbf{x}_N|\boldsymbol{\theta}}(\mathbf{x}_N|\boldsymbol{\theta})\right]\right]^{H}\middle|\boldsymbol{\theta}\right\} \tag{75}$$

- Estamos assumindo que $\mathbf{J}^{-1}$ existe
- O termo $\frac{\partial}{\partial \boldsymbol{\theta}} \ln \left[p_{\mathbf{x}_N|\boldsymbol{\theta}}(\mathbf{x}_N|\boldsymbol{\theta}) \right]$ é o vetor gradiente do logaritmo natural da distribuição $p_{\mathbf{x}_N|\boldsymbol{\theta}}(\mathbf{x}_N|\boldsymbol{\theta})$
- As derivadas parciais devem existir e ser absolutamente integráveis
- O estimador $\widehat{\boldsymbol{\theta}}$ deve ser não-polarizado, c.c. o teorema não é válido
- O teorema não pode ser aplicado a todas as distribuições (por exemplo, a distribuição uniforme) devido ao requisito da integrabilidade das derivadas parciais
- Pode-se ter também um estimador que não exista nenhuma limitante inferior
- Mas o CRB (*Cramér-Rao Bound*) é uma medida importante da eficiência do estimador pois pode ser calculado para um grande número de problemas

Prova: Deseja-se provar o limitante dos estimadores *não-polarizados* com *mínima variância*.
Sabe-se que não-polarização é relacionada com

$$\mathbb{E}\left\{\widehat{\theta}|\theta\right\} = \theta, \tag{76}$$

e então a Eq. (76) implica em

$$\mathbb{E}\left\{\left(\widehat{\theta}-\theta\right)|\theta\right\} = \int_{-\infty}^{\infty}\left(\widehat{\theta}-\theta\right)p_{X|\theta}(x|\theta)\,dx = 0. \tag{77}$$

Se derivarmos (77) em relação à θ nós obtemos

$$\begin{aligned}&\frac{\partial}{\partial\theta}\left[\int_{-\infty}^{\infty}\left(\widehat{\theta}-\theta\right)p_{X|\theta}(x|\theta)\,dx\right] = 0\\ &\int_{-\infty}^{\infty}\widehat{\theta}\cdot\frac{\partial p_{X|\theta}(x|\theta)}{\partial\theta}\,dx - \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}\left[\theta\cdot p_{X|\theta}(x|\theta)\right]\,dx = 0.\end{aligned} \tag{78}$$

Prova - cont.

Lembrando que

$$\frac{\partial f(x)g(x)}{\partial x} = \frac{\partial f(x)}{\partial x} g(x) + f(x) \frac{\partial g(x)}{\partial x}, \tag{79}$$

podemos escrever

$$\int_{-\infty}^{\infty} \widehat{\theta} \cdot \frac{\partial p_{X|\theta}(x|\theta)}{\partial \theta} \, dx - \left[\underbrace{\int_{-\infty}^{\infty} p_{X|\theta}(x|\theta) \, dx}_{=1} + \int_{-\infty}^{\infty} \theta \cdot \frac{\partial p_{X|\theta}(x|\theta)}{\partial \theta} \, dx \right] = 0$$

$$\int_{-\infty}^{\infty} \left(\widehat{\theta} - \theta\right) \cdot \frac{\partial p_{X|\theta}(x|\theta)}{\partial \theta} \, dx = 1. \tag{80}$$

Prova - cont.

Também sabemos que

$$\frac{\partial \ln[g(x)]}{\partial x} = \frac{1}{g(x)} \cdot \frac{\partial g(x)}{\partial x}, \tag{81}$$

e, para $g(x) = p_{X|\theta}(x|\theta)$, nós temos

$$\frac{\partial \ln[p_{X|\theta}(x|\theta)]}{\partial \theta} = \frac{1}{p_{X|\theta}(x|\theta)} \cdot \frac{\partial p_{X|\theta}(x|\theta)}{\partial \theta} \tag{82}$$

ou de forma equivalente

$$\frac{\partial p_{X|\theta}(x|\theta)}{\partial \theta} = p_{X|\theta}(x|\theta) \cdot \frac{\partial \ln[p_{X|\theta}(x|\theta)]}{\partial \theta}. \tag{83}$$

Prova - cont.

Aplicando (83) em (80) nós temos

$$\int_{-\infty}^{\infty} \left(\widehat{\theta} - \theta\right) \cdot \frac{\partial \ln[p_{X|\theta}(x|\theta)]}{\partial \theta} \cdot p_{X|\theta}(x|\theta)\ dx = 1, \qquad (84)$$

e também

$$\left[\int_{-\infty}^{\infty} \left(\widehat{\theta} - \theta\right) \cdot \frac{\partial \ln[p_{X|\theta}(x|\theta)]}{\partial \theta} \cdot p_{X|\theta}(x|\theta)\ dx\right]^2 = 1. \qquad (85)$$

Relembrando a inequação de Cauchy-Scharwz, que diz

$$\left[\int \alpha^2(x)\ dx\right] \cdot \left[\int \beta^2(x)\ dx\right] \geq \left[\int \alpha(t) \cdot \beta(x)\ dx\right]^2, \qquad (86)$$

Prova - cont.

Podemos então reorganizar os termos para

$$\begin{aligned}\alpha(x) &= \left(\widehat{\theta} - \theta\right) \cdot \sqrt{p_{X|\theta}(x|\theta)} \\ \beta(x) &= \frac{\partial \ln[p_{X|\theta}(x|\theta)]}{\partial \theta} \cdot \sqrt{p_{X|\theta}(x|\theta)}\end{aligned} \tag{87}$$

e usando (86), obtemos

$$\underbrace{\left[\int_{-\infty}^{\infty} \left(\widehat{\theta} - \theta\right)^2 \cdot p_{X|\theta}(x|\theta)\ dx\right]}_{\text{var}(\widehat{\theta})} \cdot \left[\int_{-\infty}^{\infty} \left[\frac{\partial \ln[p_{X|\theta}(x|\theta)]}{\partial \theta}\right]^2 \cdot p_{X|\theta}(x|\theta)\ dx\right] \geq 1. \tag{88}$$

Prova - cont.

Com isso, obtém-se

$$\text{var}\left(\widehat{\theta}\right) \geq 1 \Big/ \mathbb{E}\left\{\left[\frac{\partial \ln[p_{X|\theta}(x|\theta)]}{\partial \theta}\right]^2\right\} \tag{89}$$

Mas o termo $\mathbb{E}\left\{\left[\frac{\partial \ln[p_{X|\theta}(x|\theta)]}{\partial \theta}\right]^2\right\}$ não possui significado físico e outra expression é necessária. Nós sabemos também que

$$\int_{-\infty}^{\infty} p_{X|\theta}(x|\theta)\, dx = 1 \tag{90}$$

Ao calcular a derivada de (90) temos

$$\int_{-\infty}^{\infty} \frac{\partial p_{X|\theta}(x|\theta)}{\partial \theta}\, dx = 0 \tag{91}$$

Prova - cont.
Utilizando (83) em (91) resulta

$$\int_{-\infty}^{\infty} \frac{\partial \ln[p_{X|\theta}(x|\theta)]}{\partial \theta} \cdot p_{X|\theta}(x|\theta)\ dx = 0. \tag{92}$$

Tomando-se a segunda derivada com relação à θ, e usando (79), nós obtemos

$$\int_{-\infty}^{\infty} \frac{\partial^2 \ln[p_{X|\theta}(x|\theta)]}{\partial \theta^2} \cdot p_{X|\theta}(x|\theta)\ dx + \int_{-\infty}^{\infty} \frac{\partial \ln[p_{X|\theta}(x|\theta)]}{\partial \theta} \cdot \frac{\partial p_{X|\theta}(x|\theta)}{\partial \theta}\ dx = 0 \tag{93}$$

Prova - cont.

Usando (83) na segunda integral de (93) temos

$$\int_{-\infty}^{\infty} \frac{\partial^2 \ln[p_{X|\theta}(x|\theta)]}{\partial \theta^2} \cdot p_{X|\theta}(x|\theta)\, dx + \int_{-\infty}^{\infty} \frac{\partial \ln[p_{X|\theta}(x|\theta)]}{\partial \theta} \cdot \frac{\partial \ln[p_{X|\theta}(x|\theta)]}{\partial \theta} \cdot p_{X|\theta}(x|\theta)\, dx = 0 \tag{94}$$

o que implica em

$$\mathbb{E}\left\{ \frac{\partial^2 \ln[p_{X|\theta}(x|\theta)]}{\partial \theta^2} \right\} = -\mathbb{E}\left\{ \left[\frac{\partial \ln[p_{X|\theta}(x|\theta)]}{\partial \theta} \right]^2 \right\}. \tag{95}$$

Daí, podemos escrever o limitante como

$$\boxed{\text{var}\left(\widehat{\theta}\right) \geq -1 \Big/ \mathbb{E}\left\{ \left[\frac{\partial \ln[p_{X|\theta}(x|\theta)]}{\partial \theta} \right]^2 \right\}} \tag{96}$$

- Um dos métodos mais simples e antigos de estimação é o **método dos momentos**
- Intuitivamente simples (compreensível) e estimadores computacionalmente simples
- **Fraquezas teóricas**
- Forte relação com os momentos de ordem superior
- Relembrando
  - N amostras estatisticamente independentes $x(0), x(1), \ldots, x(N-1)$ com uma distribuição comum $p(x|\boldsymbol{\theta})$
  - pdf caraterizada pelo vetor de parâmetros $\boldsymbol{\theta}$
  - O momento de k-ésima ordem de X é

$$\mu_k = \mathbb{E}\left\{ X^k \middle| \boldsymbol{\theta} \right\} = \int_{-\infty}^{\infty} x^k p(x|\boldsymbol{\theta})\, dx, \qquad k = 1, 2, \ldots$$

  - As densidades condicionais são usadas para indicar que os parâmetros $\boldsymbol{\theta}$ são constantes (desconhecidas). Claramente, os momentos μ_k são função dos parâmetros $\boldsymbol{\theta}$.

- Por outro lado, podemos estimar os momentos facilmente a partir das medidas
- Seja d_k a estimativa do k-ésimo momento, chamado de k-**ésimo momento amostral**, que podemos escrever como

$$d_k = \frac{1}{N} \sum_{i=0}^{N-1} [(x(i)]^k \tag{97}$$

- A idéia básica do método dos momentos é equacionar os momentos teóricos μ_k com os estimados d_k como:

$$\mu_k(\boldsymbol{\theta}) = \mu_k \left(\ \theta_1, \quad \theta_2, \quad \ldots, \quad \theta_M \ \right) = d_k \tag{98}$$

- Em geral, M equações para os primeiros M momentos são suficientes para resolver os M parâmetros desconhecidos $\theta_1, \theta_2, \ldots, \theta_M$
- Se as Equações definidas por (98) possuem solução aceitável, o estimador é chamado de **estimador de momentos** e é denotado por $\widehat{\boldsymbol{\theta}}_{\text{MM}}$

- Alternativamente, pode-se utilizar os momentos centrados

$$c_k = \mathbb{E}\left\{ (X-\mu)^k \middle| \boldsymbol{\theta} \right\}$$

e os respectivos **momentos centrados amostrais**

$$s_k = \frac{1}{N-1} \sum_{i=1}^{N} [x(i) - d_1]^k$$

para formar as M equações

$$c_k(\boldsymbol{\theta}) = c_k \left(\ \theta_1, \quad \theta_2, \quad \ldots, \quad \theta_M \ \right) = s_k \qquad (99)$$

Exemplo

Assuma que as amostras $x(0), x(1), \ldots, x(N-1)$ são independentes e identicamente distribuídas tomadas de uma v.a. X com a seguinte pdf

$$p_{X|\boldsymbol{\theta}}(x|\boldsymbol{\theta}) = \frac{1}{\theta_2} \exp\left[-\frac{(x-\theta_1)}{\theta_2}\right] \tag{100}$$

em que $\theta_1 < x < \infty$ e $\theta_2 > 0$. Deseja-se estimar o vetor de parâmetros $\boldsymbol{\theta} = \left[\ \theta_1, \quad \theta_2\ \right]^T$ usando o método dos momentos.

Exemplo - cont.

Para isto, primeiro é necessário calcular os momentos teóricos μ_1 e μ_2:

$$\mu_1 = \mathbb{E}\{X|\boldsymbol{\theta}\} = \int_{\theta_1}^{\infty} \frac{x}{\theta_2} \exp\left[-\frac{(x-\theta_1)}{\theta_2}\right] dx = \theta_1 + \theta_2 \quad (101)$$

$$\mu_2 = \mathbb{E}\{X^2|\boldsymbol{\theta}\} = \int_{\theta_1}^{\infty} \frac{x^2}{\theta_2} \exp\left[-\frac{(x-\theta_1)}{\theta_2}\right] dx = (\theta_1+\theta_2)^2 + \theta_2^2 \quad (102)$$

Os estimadores de momentos são obtidos equacionado as expressões dos momentos teóricos com os dois primeiros momentos amostrais d_1 e d_2, como

$$\theta_1 + \theta_2 = d_1 \quad (103a)$$

$$(\theta_1+\theta_2)^2 + \theta_2^2 = d_2 \quad (103b)$$

Exemplo - cont.

Resolvendo o sistema de equações temos

$$\widehat{\theta}_{1,\text{MM}} = d_1 - \sqrt{(d_2 - d_1^2)} \tag{104a}$$

$$\widehat{\theta}_{2,\text{MM}} = \sqrt{(d_2 - d_1^2)} \tag{104b}$$

A outra solução possível $\widehat{\theta}_{2,\text{MM}} = -\sqrt{(d_2 - d_2^2)}$ deve ser rejeitada porque θ_2 tem de ser positivo.
De fato, pode ser observado que $\widehat{\theta}_{2,\text{MM}}$ é igual à estimativa do desvio padrão e que $\widehat{\theta}_{1,\text{MM}}$ pode ser interpretado como a média menos o desvio padrão da distribuição, ambos estimados à partir das amostras.

- A justificativa teórica para o método dos momentos é que os momentos amostrais d_k são estimadores consistentes dos momentos teóricos μ_k, bem como os momentos centrados e suas estimativas
- O problema é a eficiência dos estimadores. Geralmente a estimativa dos momentos de ordem elevada são bastante sensíveis à "*outliers*"
- Além disso, não se pode fazer afirmativas sobre a consistência e polarização do estimador de momentos o que dificulta sua aplicabilidade
- Quando o número de momentos a ser estimado é grande outros métodos devem ser usados para garantir um estimador eficiente

- O método de mínimos quadrados pode ser visto como uma abordagem determinística para o problema de estimação quando não são necessárias hipóteses sobre as distribuições de probabilidade
- Entretanto, argumentos estatísticos podem ser usados para justificar o método dos mínimos quadrados, e eles fornecem maiores *insights* sobre suas propriedades
- Podemos classificar o método dos mínimos quadrados em
  1. Estimadores lineares
  2. Estimadores não-lineares e estimadores generalizados

Seja o modelo linear básico, o vetor de dados (medidas) $\mathbf{x}_N$ é assumido ter o seguinte modelo

$$\mathbf{x}_N = \mathbf{H}\boldsymbol{\theta} + \mathbf{v}_N \tag{105}$$

Novamente, $\boldsymbol{\theta}$ é assumido ser o vetor de parâmetros e $\mathbf{v}_N$ é um vetor cujas componentes são erros de medida desconhecidos (apenas as estatísticas são conhecidas).

Hipóteses

- A **matriz de observação** $\mathbf{H}_{N\times M}$ é assumida ser completamente conhecida.
- O número de medidas é assumido ser pelo menos igual ao número de parâmetros desconhecidos, ou seja, $N \geq M$
- A matriz $\mathbf{H}$ tem posto (rank) máximo igual a M

Pode ser notado que, se $N = M$, podemos escolher $\mathbf{v}_N = \mathbf{0}$, e ter uma solução única, ou seja $\boldsymbol{\theta} = \mathbf{H}^{-1}\mathbf{x}_N$.

Se há mais parâmetros desconhecidos que medidas $(M > N)$ o sistema é indeterminado, pois há infinitas soluções que satisfazem $\mathbf{v}_N = \mathbf{0}$.

Entretanto, se as medidas são ruidosas ou contém erros, é geralmente desejado termos mais medidas (equações) que parâmetros para termos uma estimativa confiável. Desta forma, nos concentraremos na situação que $N > M$.

Quando $N > M$, não há solução para a qual $\mathbf{v}_N = \mathbf{0}$. Como os erros são desconhecidos, o melhor que se pode fazer é escolher um estimador $\widehat{\boldsymbol{\theta}}$ que minimize, de alguma forma, o efeito dos erros.

Uma escolha natural, por simplicidade matemática, é considerar o **critério de mínimos quadrados** dado por

$$\mathcal{E}_{\text{LS}} = \frac{1}{2}\|\mathbf{v}_N\|^2 = \frac{1}{2}\left(\mathbf{x}_N - \mathbf{H}\boldsymbol{\theta}\right)^H \left(\mathbf{x}_N - \mathbf{H}\boldsymbol{\theta}\right) \tag{106}$$

Note que o critério LS difere do critério MSE pois não há esperança matemática. Além disso, o critério $\mathcal{E}_{\text{LS}}$ tenta **minimizar os erros de medida** $\mathbf{v}_N$ e não o erro de estimação $\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}$.

Assim, minimizar o critério na Equação (106) em relação aos parâmetros desconhecidos $\boldsymbol{\theta}$, ou seja, derivar a equação em relação a $\boldsymbol{\theta}$ fornece a chamada **equação normal**

$$\left(\mathbf{H}^H\mathbf{H}\right)\widehat{\boldsymbol{\theta}}_{\text{LS}} = \mathbf{H}^H\mathbf{x}_N \tag{107}$$

para determinar a estimativa LS $\widehat{\boldsymbol{\theta}}_{\text{LS}}$ de $\boldsymbol{\theta}$.

É geralmente mais conveniente resolver $\widehat{\boldsymbol{\theta}}_{\text{LS}}$ da equação linear. Logo, uma vez que assumimos que a matriz $\mathbf{H}$ tem posto completo, podemos explicitar

$$\boxed{\widehat{\boldsymbol{\theta}}_{\text{LS}} = \left(\mathbf{H}^H\mathbf{H}\right)^{-1}\mathbf{H}^H\mathbf{x}_N = \mathbf{H}^{\dagger}\mathbf{x}_N} \tag{108}$$

em que $\mathbf{H}^{\dagger} = \left(\mathbf{H}^H\mathbf{H}\right)^{-1}\mathbf{H}^H$ é a chamada **pseudoinversa** de $\mathbf{H}$ (assumindo $N > M$ e posto completo da matriz)

Podemos analisar estatisticamente o estimador LS assumindo que os erros de medida têm média zero, ou seja, $\mathbb{E}\{\mathbf{v}_N\} = \mathbf{0}$.

É também fácil ver que o estimador LS é não polarizado, ou seja, $\mathbb{E}\left\{\widehat{\boldsymbol{\theta}}|\boldsymbol{\theta}\right\} = \boldsymbol{\theta}$.

Além disso, se a matriz de covariância do ruído $\mathbf{C}_{\mathbf{v}} = \mathbb{E}\{\mathbf{v}_N\mathbf{v}_N^H\}$ é conhecida, podemos calcular a matriz de covariância do erro de estimação $\mathbf{C}_{\widetilde{\boldsymbol{\theta}}}$ dada na Equação (73).

Mínimos quadrados generalizado

O problema de mínimos quadrados linear pode ser generalizado adicionando-se uma matriz de ponderação definida positiva $\mathbf{W}$ ao critério LS definido na Eq. (106)

$$\boxed{\mathcal{E}_{\text{WLS}} = \frac{1}{2}\|\mathbf{v}_N\|^2 = \frac{1}{2}\left(\mathbf{x}_N - \mathbf{H}\boldsymbol{\theta}\right)^H \mathbf{W}\left(\mathbf{x}_N - \mathbf{H}\boldsymbol{\theta}\right)} \qquad (109)$$

Mínimos quadrados generalizado - cont.

Nota-se que uma escolha natural para a solução ótima é que a matriz $\mathbf{W}$ seja a inversa da matriz de covariância dos erros de medida (ruído) $\mathbf{W} = \mathbf{C}_{\mathbf{v}}^{-1}$. Isto é devido esta escolha fornecer o seguinte estimador generalizado de mínimos quadrados

$$\widehat{\boldsymbol{\theta}}_{\text{WLS}} = \left(\mathbf{H}^H \mathbf{C}_{\mathbf{v}}^{-1} \mathbf{H}\right)^{-1} \mathbf{H}^H \mathbf{C}_{\mathbf{v}}^{-1} \mathbf{x}_n \tag{110}$$

também minimiza o erro quadrático de estimação $\mathcal{E}_{\text{MSE}} = \mathbb{E}\{\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|^2 | \boldsymbol{\theta}$. Assume-se que $\widehat{\boldsymbol{\theta}}$ é um estimador linear e não polarizado.

O estimador dado na Eq. (110) é chamado frequentemente de *Best Linear Unbiased Estimator* (BLUE) ou Estimador de Gauss-Markov.

Mínimos quadrados generalizado - cont.

Note que o critério generalizado se reduz ao LS linear se $\mathbf{C_v} = \sigma^2 \mathbf{I}$. Isto acontece quando as medidas têm média zero e são mutuamente independentes e identicamente distribuídas com uma variância comum σ^2.

A escolha de $\mathbf{C_v} = \sigma^2 \mathbf{I}$ também se aplica quando não se tem conhecimento *a priori* sobre a matriz de covariância do ruído. Nestes casos, o BLUE coincide com o estimador de mínimos quadrados linear. Esta conexão fornece um forte argumento estatístico para o uso do método dos mínimos quadrados, já que o critério MSE mede diretamente os erros de estimação $\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}$.

Mínimos quadrados não-linear

Nem sempre, o modelo linear descreve adequadamente a dependência entre os parâmetros $\boldsymbol{\theta}$ e as medidas $\mathbf{x}_N$. Desta maneira, uma extensão natural é considerar o seguinte modelo não-linear

$$\mathbf{x}_N = \boldsymbol{f}(\boldsymbol{\theta}) + \mathbf{v}_N \tag{111}$$

Aqui, $\boldsymbol{f}$ é um vetor de funções não-lineares e continuamente diferenciáveis sobre o parâmetro $\boldsymbol{\theta}$. Cada componente $f_i(\boldsymbol{\theta})$ ou $\boldsymbol{f}(\boldsymbol{\theta})$ é assumido ser uma função escalar conhecida dos componentes de $\boldsymbol{\theta}$

Mínimos quadrados não-linear - cont.

Similarmente ao modelo linear, o critério de mínimos quadrados não-linear $\mathcal{E}_{\text{NLS}}$ é definido como uma soma dos quadrados dos erros de medida $\|\mathbf{v}_N\|^2 = \sum_j [v(j)]^2$.

Assim, para o modelo na Eq. (111) temos

$$\mathcal{E}_{\text{NLS}} = [\mathbf{x}_N - \boldsymbol{f}(\boldsymbol{\theta})]^H [\mathbf{x}_N - \boldsymbol{f}(\boldsymbol{\theta})] \tag{112}$$

E o estimador não-linear $\widehat{\boldsymbol{\theta}}_{\text{NLS}}$ é o valor de $\boldsymbol{\theta}$ que minimiza o $\mathcal{E}_{\text{NLS}}$

Mínimos quadrados não-linear - cont.

Assim, o problema da Eq. (112) é um problema de otimização não-linear para encontrar o mínimo da função $\mathcal{E}_{\text{NLS}}$.

Tais problemas geralmente não apresentam uma solução analítica, mas várias métodos/ferramentas de otimização podem ser empregados tais como

1. Redes neurais artificiais
2. Filtros de Volterra
3. Métodos heurísticos

- Como, o nome sugere, procura maximizar a semelhança entre os parâmetros e o modelo assumido como sendo verdadeiro.
- Estimador de máxima verossimilhança (MV) assume que o vetor de parâmetros $\boldsymbol{\theta}$ é constante ou não se tem informação *a priori* sobre o mesmo
- O estimador MV tem várias propriedades de otimalidade assintótica que o tornam uma escolha desejável quando o número de amostras é grande
- Aplicação em várias áreas, como por exemplo em comunicações: algoritmo de Viterbi

Método

A estimativa de máxima verossimilhança (MV) $\widehat{\boldsymbol{\theta}}_{\text{MV}}$ dos parâmetros $\boldsymbol{\theta}$ é escolhida de tal forma que $\widehat{\boldsymbol{\theta}}_{\text{MV}}$ maximize a seguinte **função de máxima verossimilhança** (distribuição conjunta)

$$p_X(\mathbf{x}_N|\boldsymbol{\theta}) = p_X\left(x(0), \ldots, x(N-1)|\boldsymbol{\theta}\right) \tag{113}$$

das medidas $x(1), \ldots, x(N)$. O estimador MV então corresponde ao valor de $\widehat{\boldsymbol{\theta}}_{\text{MV}}$ que torna as medidas obtidas as *mais semelhantes*.

- Uma vez que muitas densidades contém uma função exponencial, é geralmente útil utilizar a função de log-máxima verossimilhança $\ln\left[p_X(\mathbf{x}_N|\boldsymbol{\theta})\right]$, uma vez que o máximo de $p_X(\mathbf{x}_N|\boldsymbol{\theta})$ também é o máximo de $\ln\left[p_X(\mathbf{x}_N|\boldsymbol{\theta})\right]$
- Como encontrar o estimador? Através das soluções da *equação de semelhança*

$$\boxed{\left.\frac{\partial}{\partial \boldsymbol{\theta}} \ln\left[p_X(\mathbf{x}_N|\boldsymbol{\theta})\right]\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{MV}}} = \mathbf{0}} \tag{114}$$

- A solução da Eq. (114) fornece os valores de $\boldsymbol{\theta}$ que maximizam (ou minimizam) a função de semelhança. Se a função possui vários máximos máximos e/ou mínimos, deve-se escolher o valor de $\widehat{\boldsymbol{\theta}}_{\text{MV}}$ que corresponde ao máximo absoluto (global)

A construção da função de MV pode ser uma tarefa bastante árdua se as amostras (medidas) forem dependentes umas das outras. Por isso, é quase sempre assumido que as amostras são **independentes** umas das outras, fato que, felizmente, se verifica freqüentemente na prática. Assim, a função de MV torna-se

$$p_X(\mathbf{x}_N|\boldsymbol{\theta}) = \prod_{i=0}^{N-1} p_X(x(i)|\boldsymbol{\theta}) \tag{115}$$

em que $p_X(x(i)|\boldsymbol{\theta})$ é a pdf condicional de uma medida escalar $x(i)$. Note que se tomarmos o logaritmo na Eq. (115) o produto torna-se um somatório de logaritmos $\sum_i \ln\left[p_X(x(i)|\boldsymbol{\theta})\right]$

Assim, o vetor de verossimilhança da Eq. (114) consiste de M equações escalares da forma

$$\frac{\partial}{\partial \theta_i} \ln \left[p_X(\mathbf{x}_N | \widehat{\boldsymbol{\theta}}_{\text{MV}}) \right] \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{MV}}} = 0, \qquad i = 1, \ldots, M \qquad (116)$$

para os M parâmetros $\widehat{\theta}_{i,\text{MV}}$, $i = 1, \ldots, M$.

Estas equações são em geral acopladas e não-lineares, o que torna sua solução possível apenas por técnicas numéricas, exceto para casos simples. Em aplicações práticas a complexidade computacional do método MV pode ser proibitiva e aproximações são necessariamente empregadas para simplificar a função de verossimilhança e/ou uso de métodos sub-ótimos.

Exemplo

Sejam N amostras independentes tomadas de uma v.a. X com distribuição gaussiana de média μ e variância σ^2. Podemos escrever a função de verossimilhança como

$$p_X(\mathbf{x}_N|\mu,\sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^N}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=0}^{N-1}[x(i)-\mu]^2\right] \quad (117)$$

Então, a função de log-MV torna-se

$$\ln[p_X(\mathbf{x}_N|\mu,\sigma^2)] = -\frac{N}{2}\ln\left[2\pi\sigma^2\right] - \frac{1}{2\sigma^2}\sum_{i=0}^{N-1}[x(i)-\mu]^2 \quad (118)$$

Exemplo - cont.

A primeira equação que temos é a seguinte

$$\frac{\partial}{\partial \mu} \ln \left[p_X \left(\mathbf{x}_N | \widehat{\mu}_{\text{MV}}, \widehat{\sigma}^2_{\text{MV}} \right) \right] = \frac{1}{\widehat{\sigma}^2_{\text{MV}}} \sum_{i=0}^{N-1} \left[x(i) - \widehat{\mu}_{\text{MV}} \right]^2 = 0 \quad (119)$$

Resolvendo a equação acima temos a seguinte estimativa MV para a média

$$\widehat{\mu}_{\text{MV}} = \frac{1}{N} \sum_{i=0}^{N-1} x(i) \quad (120)$$

Exemplo - cont.

A segunda equação é obtida derivando a função de verossimilhança em relação à variância, ou seja,

$$\frac{\partial}{\partial \sigma^2} \ln \left[p_X \left(\mathbf{x}_N | \widehat{\mu}_{\text{MV}}, \widehat{\sigma}^2_{\text{MV}} \right) \right] = -\frac{N}{2\widehat{\sigma}^2_{\text{MV}}} + \frac{N}{2\widehat{\sigma}^4_{\text{MV}}} \sum_{i=0}^{N-1} \left[x(i) - \widehat{\mu}_{\text{MV}} \right]^2 = 0 \tag{121}$$

Temos então o seguinte estimador MV para a variância

$$\widehat{\sigma}^2_{\text{MV}} = \frac{1}{N} \sum_{i=0}^{N-1} \left[x(i) - \widehat{\mu}_{\text{MV}} \right]^2 \tag{122}$$

Exemplo - cont.

Temos então as seguintes observações:

- O estimador da média é não-polarizado
- O estimador da variância é polarizado pois usa a estimativa da média. Entretanto, tal polarização é geralmente pequena e assintoticamente não-polarizado

Algumas propriedades teóricas importantes dos estimadores MV

1. Se existe um estimador que satisfaz o limite inferior de Cramér-Rao como a igualdade, ele pode ser obtido usando o método da máxima verossimilhança
2. O estimador de máxima verossimilhança $\widehat{\boldsymbol{\theta}}_{\text{MV}}$ é consistente
3. O estimador MV é *assintoticamente eficiente*. Isto significa que ele atinge o CRB para o erro de estimação.

Exemplo

Cálculo do CRB para a média de uma v.a. gaussiana unidimensional. Sabe-se que a derivada da função de log-verossimilhança em relação à média μ é dada por

$$\frac{\partial}{\partial \mu} \ln \left[p_X \left(\mathbf{x}_N | \mu \right) \right] = \frac{1}{\sigma^2} \sum_{i=0}^{N-1} \left[x(i) - \mu \right] \tag{123}$$

Como estamos interessados em apenas um parâmetro, a matriz de informação de Fisher se reduz a um escalar da forma

$$\begin{aligned} J &= \mathbb{E} \left\{ \left[\frac{\partial}{\partial \mu} \ln \left[p_X \left(\mathbf{x}_N | \mu \right) \right] \right]^2 \middle| \mu \right\} \\ &= \mathbb{E} \left\{ \left[\frac{1}{\sigma^2} \sum_{i=0}^{N-1} \left[x(i) - \mu \right] \right]^2 \middle| \mu \right\} \end{aligned} \tag{124}$$

Exemplo - cont.

Uma vez que as amostras são assumidas independentes todos os termos de correlação cruzada se anulam e temos então que

$$J = \frac{1}{\sigma^4} \sum_{i=0}^{N-1} \mathbb{E}\left\{ [x(i) - \mu]^2 \middle| \mu, \sigma^2 \right\} = \frac{N\sigma^2}{\sigma^4} = \frac{N}{\sigma^2} \tag{125}$$

Com isso, o CRB para o erro médio quadrático de qualquer estimador não-polarizado $\widehat{\mu}$ da média de uma densidade gaussiana é

$$\mathbb{E}\left\{ (\mu - \widehat{\mu})^2 | \mu \right\} \geq J^{-1} = \frac{\sigma^2}{N} \tag{126}$$

Algoritmo *Expectation-Maximization* (EM)

O algoritmo EM fornece uma abordagem geral de implementação iterativa da estimativa MV. Uma das principais vantagens do algoritmo EM é que ele, geralmente, fornece tratamento de problemas difíceis pela abordagem MV, tais como problemas com múltiplos parâmetros e funções de verossimilhança altamente não-lineares.

Entretanto, o uso do EM requer cautela uma vez que ele pode levar a mínimos/máximos locais já que possui implementação adaptativa

Há ainda uma conexão do estimador MV com o estimador LS. Se assumirmos que os parâmetros $\boldsymbol{\theta}$ são constantes desconhecidas e independentes do ruído aditivo $\mathbf{v}_N$, a distribuição condicional $p_X(\mathbf{x}_N|\boldsymbol{\theta})$ de $\mathbf{x}_N$ é a mesma que a distribuição de $\mathbf{v}_N$ no ponto $\mathbf{v}_N = \mathbf{x}_N - \boldsymbol{f}(\boldsymbol{\theta})$, assim

$$p_X(\mathbf{x}_N|\boldsymbol{\theta}) = p_X(\mathbf{x}_N - \boldsymbol{f}(\boldsymbol{\theta})|\boldsymbol{\theta}) \tag{127}$$

Se assumirmos que $\mathbf{v}_N$ é gaussiano de média nula e covariância $\sigma^2\mathbf{I}$ então teremos

$$p_X(\mathbf{x}_N|\boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi\sigma)^N}} \exp\left\{-\frac{1}{2\sigma^2}[\mathbf{x}_N - \boldsymbol{f}(\boldsymbol{\theta})]^H[\mathbf{x}_N - \boldsymbol{f}(\boldsymbol{\theta})]\right\} \tag{128}$$

Claramente, a Eq. (128) é maximizada quando o argumento da exponencial é minimizado, uma vez que o termo fora dele não é dependente dos parâmetros $\boldsymbol{\theta}$. Assim, encontrar o estimador MV neste problema corresponde a minimizar

$$[\mathbf{x}_N - \boldsymbol{f}(\boldsymbol{\theta})]^H [\mathbf{x}_N - \boldsymbol{f}(\boldsymbol{\theta})] = \|\mathbf{x}_N - \boldsymbol{f}(\boldsymbol{\theta})\|^2 \qquad (129)$$

Em resumo, mesmo para o caso não-linear, se o ruído for gaussiano, média nula, com matriz de covariância $\mathbf{C}_\mathbf{v} = \sigma^2 \mathbf{I}$ e independente dos parâmetros $\boldsymbol{\theta}$ o estimador MV e o LS fornecem a mesma solução.

- Até então, os métodos de estimação estudados assumiam que o vetor de parâmetros $\boldsymbol{\theta}$ era constante desconhecida
- Na abordagem bayesiana, os parâmetros $\boldsymbol{\theta}$ são assumidos **aleatórios**
- Isso modifica o tipo de processamento e assume uma questão importante sobre o conhecimento *a priori* sobre os parâmetros
- Nos métodos bayesianos, isto é informado no problema pelo conhecimento da densidade *a priori* $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$
- Na prática, tal conhecimento pleno é bastante raro e somente algumas hipóteses *a priori* dos parâmetros são disponíveis
- Entretanto, podemos assumir **algum** conhecimento sobre os parâmetros, e.g. se eles são gaussianos, se há um valor limite para sua média, etc

- A essência dos métodos de estimação bayesianos é a **densidade a posteriori** $p_{\boldsymbol{\Theta}|X}(\boldsymbol{\theta}|\mathbf{x}_N)$ dos parâmetros $\boldsymbol{\theta}$ dadas as medidas $\mathbf{x}_N$
- Isto é de interesse, porquê, basicamente, a densidade a posteriori contém todas as informações relavantes sobre os parâmetros $\boldsymbol{\theta}$
- Escolhendo uma estimativa $\widehat{\boldsymbol{\theta}}$ para os parâmetros $\boldsymbol{\theta}$ sobre todos os valores de $\boldsymbol{\theta}$ para os quais a densidade a posteriori é alta (máxima) é uma escolha arbitrária
- Os dois mais populares métodos de estimação bayesiana
  1. Minimização do erro médio quadrático (MMSE)
  2. Maximização da probabilidade *a posteriori*

Minimização do erro médio quadrático

No método MSE para parâmetros aleatórios $\boldsymbol{\theta}$, o estimador ótimo $\widehat{\boldsymbol{\theta}}_{\text{MSE}}$ (no sentido de mínimo erro médio quadrático) é dado pela minimização do MSE, ou seja,

$$\mathcal{E}_{\text{MSE}} = \mathbb{E}\left\{\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|^2\right\} \tag{130}$$

em relação ao estimador $\widehat{\boldsymbol{\theta}}$. Como especificar o estimador $\widehat{\boldsymbol{\theta}}_{\text{MSE}}$? O teorema a seguir responde esta questão.

Minimização do erro médio quadrático - cont.

Teorema: Assumindo que os parâmetros $\boldsymbol{\theta}$ e as medidas $\mathbf{x}_N$ possuem a densidade de probabilidade conjunta dada por $p_{\boldsymbol{\Theta},X}(\boldsymbol{\theta}, \mathbf{x}_N)$, o estimador $\widehat{\boldsymbol{\theta}}_{\text{MSE}}$ que fornece o erro quadrático médio mínimo é dado pela esperança condicional

$$\widehat{\boldsymbol{\theta}}_{\text{MSE}} = \mathbb{E}\left\{\boldsymbol{\theta}|\mathbf{x}_N\right\} \tag{131}$$

Para entender o resultado, vamos primeiro, provar o teorema.

Minimização do erro médio quadrático - cont.

Deve-se notar que podemos escrever a esperança da Eq. (130) em dois estágios, como

$$\mathcal{E}_{\text{MSE}} = \mathbb{E}\left\{\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|^2\right\} = \mathbb{E}_{\mathbf{x}}\left\{\mathbb{E}\left\{\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|^2 | \mathbf{x}_N\right\}\right\} \quad (132)$$

Uma vez que minimizar o MSE corresponde minimizar a esperança condicional na equação acima, tem-se

$$\mathbb{E}\left\{\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|^2 | \mathbf{x}_N\right\} = \widehat{\boldsymbol{\theta}}^H\widehat{\boldsymbol{\theta}} - 2\widehat{\boldsymbol{\theta}}^H\mathbb{E}\{\boldsymbol{\theta}|\mathbf{x}_N\} + \mathbb{E}\left\{\boldsymbol{\theta}^H\boldsymbol{\theta}|\mathbf{x}_N\right\} \quad (133)$$

Minimização do erro médio quadrático - cont.

Então, para encontrar o mínimo da Eq. (133) derivamos em relação aos parâmetros $\boldsymbol{\theta}$ e igualamos à zero obtendo

$$\frac{\partial \mathcal{E}_{\text{MSE}}}{\partial \boldsymbol{\theta}^H} = 2\widehat{\boldsymbol{\theta}} - 2\mathbb{E}\left\{\boldsymbol{\theta}|\mathbf{x}_N\right\} = 0 \tag{134}$$

Assim, verificamos o teorema que nos fornece o estimador MSE para a abordagem bayesiana.

Um outro resultado importante é que o estimador é não-polarizado, pois

$$\mathbb{E}\left\{\widehat{\boldsymbol{\theta}}_{\text{MSE}}\right\} = \mathbb{E}_{\mathbf{x}}\left\{\mathbb{E}\left\{\boldsymbol{\theta}|\mathbf{x}_N\right\}\right\} = \mathbb{E}\{\boldsymbol{\theta}\} \tag{135}$$

Minimização do erro médio quadrático - cont.

Entretanto, para se calcular a esperança da Eq. (131) necessita-se calcular a densidade condicional abaixo, a qual é obtida pela fórmula de Bayes

$$p_{\boldsymbol{\Theta}|\mathbf{X}}(\boldsymbol{\theta}|\mathbf{x}_N) = \frac{p_{\mathbf{X}|\boldsymbol{\Theta}}(\mathbf{x}_N|\boldsymbol{\theta})p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}{p_{\mathbf{X}}(\mathbf{x}_N)} \tag{136}$$

em que o denominador é dado pela seguinte integral (regra da probabilidade total)

$$p_{\mathbf{X}}(\mathbf{x}_N) = \int_{-\infty}^{\infty} p_{\mathbf{X}|\boldsymbol{\Theta}}(\mathbf{x}_N|\boldsymbol{\theta})p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\, d\boldsymbol{\theta} \tag{137}$$

Minimização do erro médio quadrático - cont.

Assim, a solução MMSE para método bayesiano, implica na necessidade de resolver duas integrais (a da esperança condicional e a da densidade das medidas). Infelizmente, tais integrais são, geralmente, impossíveis de avaliar/resolver analiticamente o que torna a complexidade (implementação) de tais métodos também bastante elevada.

Dois casos especiais permitem solução fácil e merecem ser mencionados

Minimização do erro médio quadrático - cont.

(1) Se o estimador $\widehat{\boldsymbol{\theta}}$ é uma função linear dos dados, $\widehat{\boldsymbol{\theta}} = \mathbf{L}\mathbf{x}_N$, é possível mostrar que o estimador ótimo que minimiza o MSE é dado por

$$\widehat{\boldsymbol{\theta}}_{\text{LMSE}} = \mathbf{m}_{\boldsymbol{\theta}} + \mathbf{C}_{\boldsymbol{\theta}\mathbf{x}}\mathbf{C}_{\mathbf{x}}^{-1}(\mathbf{x}_N - \mathbf{m}_{\mathbf{x}}) \tag{138}$$

em que $\mathbf{m}_{\boldsymbol{\theta}}$ e $\mathbf{m}_{\mathbf{x}}$ são os vetores das médias de $\boldsymbol{\theta}$ e $\mathbf{x}_N$, respectivamente, $\mathbf{C}_{\mathbf{x}}$ é a matriz de covariância de $\mathbf{x}_N$ e $\mathbf{C}_{\boldsymbol{\theta}\mathbf{x}}$ é a matriz de covariância de $\boldsymbol{\theta}$ e $\mathbf{x}_N$.

Minimização do erro médio quadrático - cont.

A matriz do erro de covariância, correspondente ao estimador $\widehat{\boldsymbol{\theta}}_{\text{LMSE}}$ é

$$\mathbb{E}\left\{\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{\text{LMSE}}\right)\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{\text{LMSE}}\right)^{H}\right\}=\mathbf{C}_{\boldsymbol{\theta}}-\mathbf{C}_{\boldsymbol{\theta}\mathbf{x}}\mathbf{C}_{\mathbf{x}}^{-1}\mathbf{C}_{\mathbf{x}\boldsymbol{\theta}} \quad (139)$$

em que $\mathbf{C}_{\boldsymbol{\theta}}$ é a matriz de covariância dos parâmetros $\boldsymbol{\theta}$.

Conclusão: Se o estimador MMSE é restrito a ser linear, para calculá-lo é suficiente conhecer as estatísticas de primeira e segunda ordem dos parâmetros $\boldsymbol{\theta}$ e dados $\mathbf{x}_N$.

Minimização do erro médio quadrático - cont.

(2) Se a densidade conjunta $p_{\boldsymbol{\Theta},\mathbf{X}}(\boldsymbol{\theta}, \mathbf{x}_N)$ for *gaussiana* o resultado do estimador linear é o ótimo geral. Isto é devido ao fato que a densidade $p_{\mathbf{X}|\boldsymbol{\Theta}}(\mathbf{x}_N|\boldsymbol{\theta})$ é também gaussiana com a média condicional dada pela Eq. (138) e matriz de covariância dada pela Eq. (139).

Isto reafirma o fato que para distribuições gaussianas, apenas as estatísticas de primeira e segunda ordem são suficientes para projetar o estimador.

Maximum a Posteriori (MAP)

Uma alternativa ao emprego da minimização do erro médio quadrático é a aplicação do método bayesiano no mesmo princípio do método da máxima verossimilhança.

Isto leva ao **estimador de máxima a posteriori (MAP)** $\widehat{\boldsymbol{\theta}}_{\text{MAP}}$, que é definido como o valor que maximiza a densidade a posteriori $p_{\boldsymbol{\Theta}|\mathbf{X}}(\boldsymbol{\theta}|\mathbf{x}_N)$ de $\boldsymbol{\theta}$ dada as medidas $\mathbf{x}_N$.

O estimador MAP pode ser entendido como o valor mais provável dos parâmetros $\boldsymbol{\theta}$ para os dados $\mathbf{x}_N$ disponíveis.

Maximum a Posteriori (MAP) - cont.

No método MMSE nós notamos que a densidade *a posteriori* pode ser obtida a partir da fórmula de Bayes dada na Eq. (136). Uma vez que apenas o numerador daquela equação depende dos parâmetros $\boldsymbol{\theta}$ o estimador MAP também pode ser obtido apenas maximizando o numerador da Eq. (136).

Assim, o estimador MAP busca encontrar o $\boldsymbol{\theta}$ que maximiza

$$p_{\mathbf{X}|\boldsymbol{\Theta}}(\mathbf{x}_N|\boldsymbol{\theta})p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \tag{140}$$

Maximum a Posteriori (MAP) - cont.

Então, de maneira análoga ao método MV, podemos encontrar o estimador MAP $\widehat{\boldsymbol{\theta}}_{\text{MAP}}$ resolvendo a seguinte equação logarítmica

$$\frac{\partial \ln\left[p_{\boldsymbol{\Theta}|\mathbf{X}}(\boldsymbol{\theta}|\mathbf{x}_N)\right]}{\partial \boldsymbol{\theta}} = \frac{\partial \ln\left[p_{\mathbf{X}|\boldsymbol{\Theta}}(\mathbf{x}_N|\boldsymbol{\theta})\right]}{\partial \boldsymbol{\theta}} + \frac{\partial \ln\left[p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\right]}{\partial \boldsymbol{\theta}} \quad (141)$$

Se compararmos a equação acima e a do método MV notamos que as duas são bastante similares, com a diferença que o estimador MAP leva em conta a informação adicional contida no termo $\frac{\partial p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$.

Maximum a Posteriori (MAP) - cont.

Se a densidade a priori $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ é uniforme, o estimador MV e o MAP tornam-se os mesmos. Este é então o caso em que nenhuma informação a priori é disponível.

Com isso, quando $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ não é uniforme, os estimadores obtidos pelos métodos MV e MAP tornam-se diferentes.

Parte IV

Filtragem Ótima

Filtragem

Processamento de sinais para extração ou modificação de certas características do sinal

Clássico *versus* moderno

- Problema de filtragem "moderna": Wiener e Kolmogorov nos anos 40
- Processar, de maneira ótima, sinais aleatórios que ocupam (geralmente) mesma faixa de freqüência
- Técnicas clássicas de PDS não satisfazem
- Otimização: escolha de critério que ressalte as características de interesse do sinal segundo uma estrutura de processamento

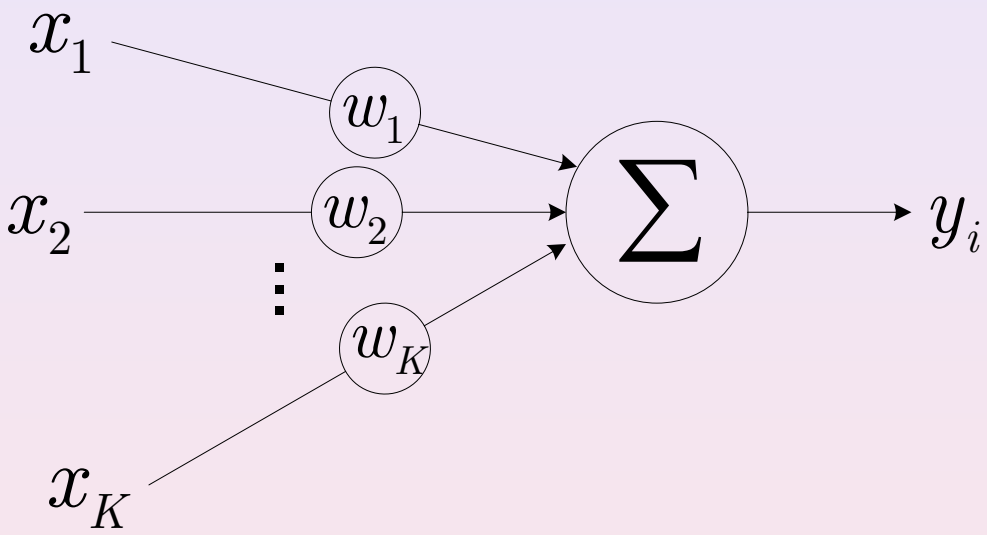
Escolhas

Estrutura: **linear** - grande potencial de aplicação e simplicidade de análise

Critério: minimização do erro médio quadrático (**MMSE**)

Cenário: sinal de treinamento disponível (**supervisionado**)

Estrutura básica



combinação linear dos dados e parâmetros

$$\overbrace{\boxed{y_i = \sum_{j=1}^{K} w_i x_i}}$$

$$y_i = \mathbf{w}^T \mathbf{x}$$

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \dots & x_K \end{bmatrix}^T$$

$$\mathbf{w} = \begin{bmatrix} w_1 & w_2 & \dots & w_K \end{bmatrix}^T$$

Meta

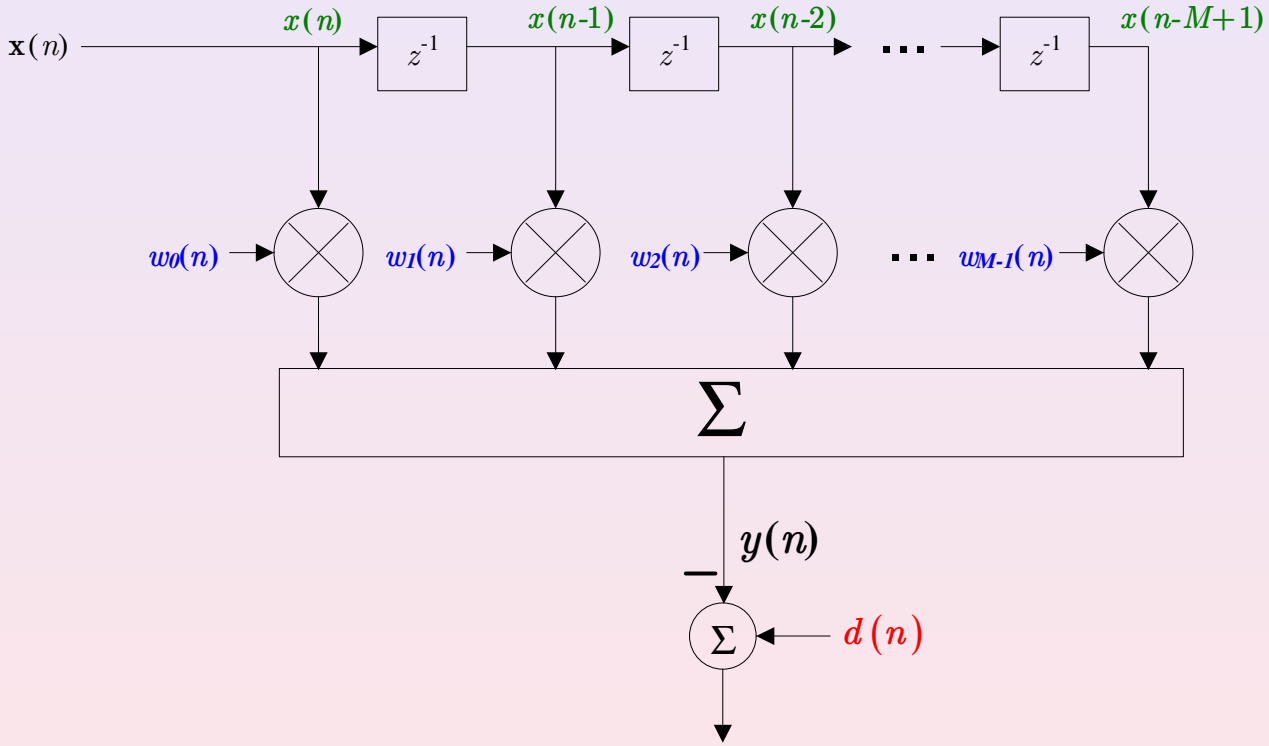
Problema

Otimizar $\mathbf{w}$, isto é, calcular $\mathbf{w}$ para que $y_i = d_i$ (sinal desejado)

- Solução para $\mathbf{w}$ ótimo: filtragem de Wiener
- Esse problema dá origem a duas configurações fundamentais que nos interessam

Primeira estrutura: filtragem temporal, em que $x_1, x_2, \ldots, x_K$ são amostras temporais de um sinal $x(n)$

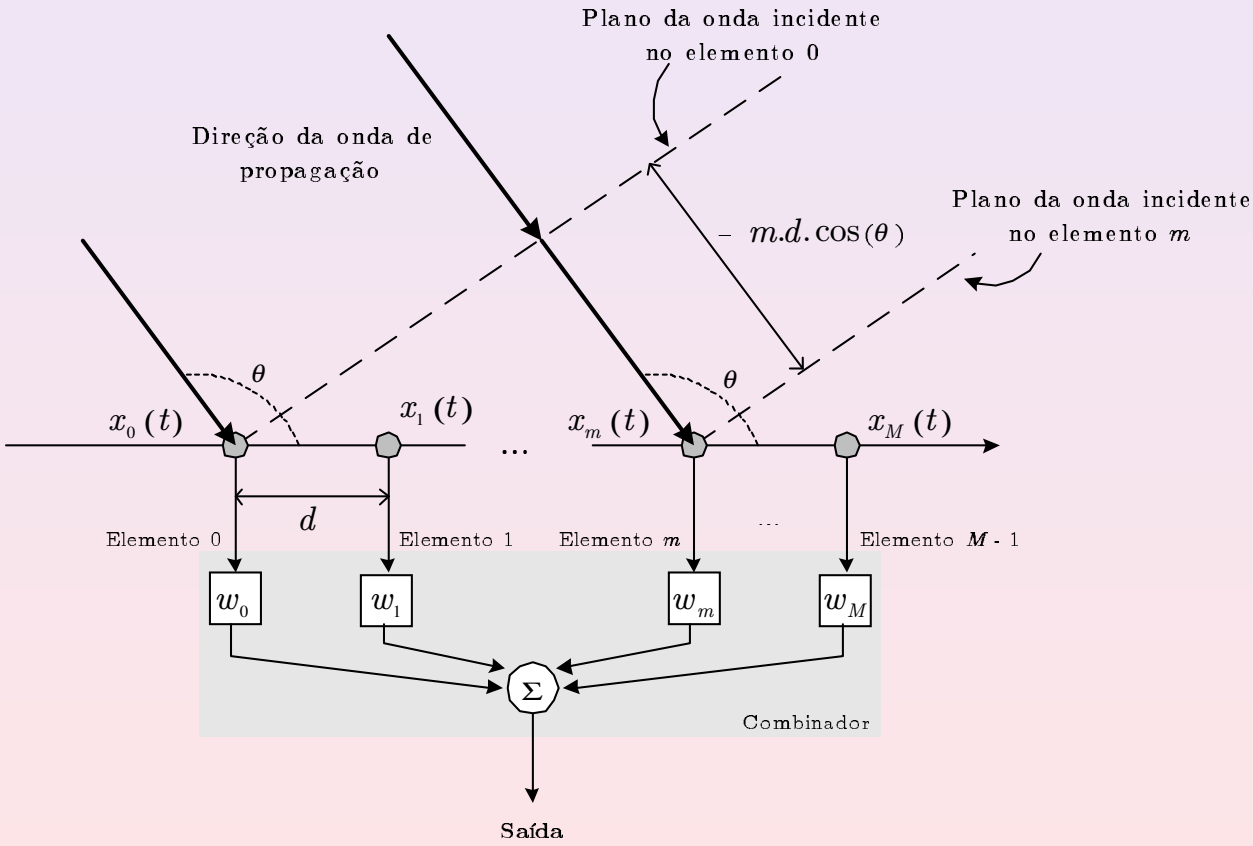
Problema: equalização de canais (Lucky, 1965)



$x(n)$ e $d(n)$ são processos estocásticos estacionários discretos

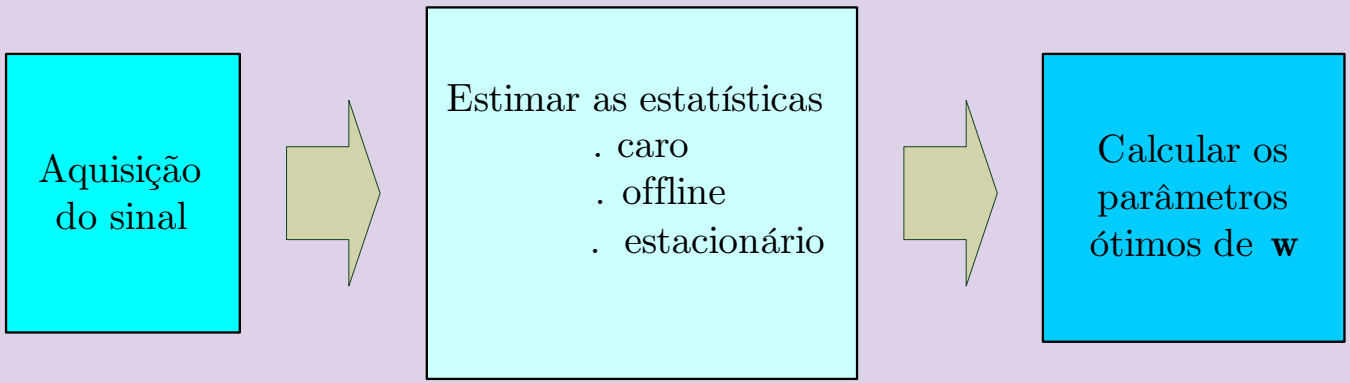
Segunda estrutura: filtragem espacial, em que $x_1, x_2, \ldots, x_K$ são amostras espaciais de um sinal $x(n)$ incidindo no conjunto de sensores (arranjo)

Problema: antenas adaptativas (Widrow, 1960)

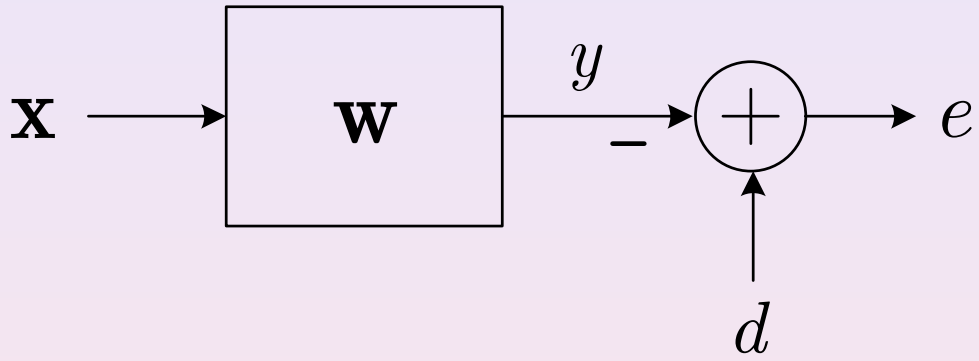


Filtragem ótima

- Boa base teórica
- Calcular $\mathbf{w}$ conhecendo as estatísticas dos sinais envolvidos
- Tipicamente o modelo abaixo



Sistema geral



1. Combinador linear: $\mathbf{x}$ composto de amostras espaciais
2. Filtro FIR transversal: $\mathbf{x}$ composto de amostras temporais

Considerando a filtragem temporal, temos então

Meta

Minimizar $\mathbb{E}\left\{e^2(n)\right\}$

- Filtro de comprimento M
- $e(n) = d(n) - y(n)$
- $\mathbf{x}(n) = \left[\begin{array}{ccc} x(n) & \cdots & x(n-M+1) \end{array} \right]^T$
- $\mathbf{w} = \left[\begin{array}{cccc} w_0 & w_1 & \cdots & w_{M-1} \end{array} \right]^T$

Considerações:

- Sinal $x(n)$ é estacionário e de média nula

$$\Rightarrow r(i,j) = \mathbb{E}\left\{x(n-i)x(n-j)\right\} = r(i-j)$$

- Sinal $d(n)$ é estacionário, de média nula e com variância igual a σ_d^2

Derivando filtro ótimo

Do modelo, temos então:

$$\begin{aligned} e(n) &= d(n) - y(n) \\ &= d(n) - \mathbf{w}^T \mathbf{x}(n) \end{aligned} \tag{142}$$

e

$$\begin{aligned} e^2(n) &= \left(d(n) - \mathbf{w}^T \mathbf{x}(n)\right) \cdot \left(d(n) - \mathbf{w}^T \mathbf{x}(n)\right)^T \\ &= d^2(n) - 2\mathbf{w}^T \mathbf{x}(n) d(n) + \mathbf{w}^T \mathbf{x}(n) \mathbf{x}^T(n) \mathbf{w} \end{aligned} \tag{143}$$

Derivando filtro ótimo - cont.

Aplicando o operando esperança...

$$\mathbb{E}\left\{e^2(n)\right\} = \underbrace{\mathbb{E}\left\{d^2(n)\right\}}_{\substack{\text{variância} \\ \text{de } d(n)}} - 2\mathbf{w}^T \underbrace{\mathbb{E}\left\{\mathbf{x}(n)d(n)\right\}}_{\substack{\text{correlação} \\ \text{cruzada}}} + \mathbf{w}^T \underbrace{\mathbb{E}\left\{\mathbf{x}(n)\mathbf{x}^T(n)\right\}}_{\substack{\text{matrix de} \\ \text{autocorrelação}}} \mathbf{w} \tag{144}$$

Assim, temos

$$\mathbb{E}\left\{e^2(n)\right\} = \sigma_d^2 - 2\mathbf{w}^T\mathbf{p}_{\mathbf{x}d} + \mathbf{w}^T\mathbf{R}_{\mathbf{x}}\mathbf{w} \tag{145}$$

Como a equação é quadrática em relação aos parâmetros $\mathbf{w}$, existe somente um ponto de mínimo (máximo)

Derivando filtro ótimo - cont.

Achar o ponto ótimo é então equivalente a encontrar o ponto onde a função tem o seu mínimo, ou seja

$$\nabla_{\mathbf{w}} \mathbb{E}\left\{e^2(n)\right\} = 0 \tag{146}$$

Para simplificar a notação, podemos chamar $\mathbb{E}\left\{e^2(n)\right\} = \varepsilon$. Logo, a Eq. (146) nos leva à derivação da Eq. (145) em relação aos parâmetros $\mathbf{w}$, ou seja

$$\nabla_{\mathbf{w}} \varepsilon = \frac{\partial \varepsilon}{\partial \mathbf{w}} = \left[\begin{array}{cccc} \dfrac{\partial \varepsilon}{\partial w_0} & \dfrac{\partial \varepsilon}{\partial w_1} & \cdots & \dfrac{\partial \varepsilon}{\partial w_{M-1}} \end{array} \right]^T \tag{147}$$

Derivando filtro ótimo - cont.

Então, temos

$$-2\mathbf{p}_{\mathbf{x}d} + 2\mathbf{R}_{\mathbf{x}}\mathbf{w} = 0 \tag{148}$$

Da qual, após multiplicação à esquerda por $\mathbf{R}_{\mathbf{x}}^{-1}$, obtém a equação do filtro ótimo:

$$\boxed{\mathbf{w}_{\text{opt}} = \mathbf{R}_{\mathbf{x}}^{-1} \cdot \mathbf{p}_{\mathbf{x}d}} \tag{149}$$

A Equação (149) é chamada então de *Equação de Wiener-Hopf*, ou conjunto de equações normais, e é por vezes escrita como

$$\sum_{i=0}^{M-1} w_{\text{opt},i} r_x(i-k) = p_{xd}(k), \quad k = 0, 1, \ldots, M-1 \tag{150}$$

em que $r(i-k)$ é a correlação para os instantes i e j e $p_{xd}(k)$ é a correlação cruzada entre $x(n-k)$ e $d(n)$.

Valor do erro quadrático mínimo

De posse do filtro ótimo, podemos então calcular o valor mínimo do erro médio quadrático, ou seja,

$$\mathbf{w} = \mathbf{w}_{\text{opt}} \Rightarrow \mathbb{E}\left\{e^2(n)\right\}\Big|_{\text{mínimo}}$$

Substituindo a Eq. (149) em (145), obtém-se

$$\begin{aligned}
\varepsilon_{\text{min}} &= \sigma_d^2 - 2\mathbf{w}^T\mathbf{p}_{\mathbf{x}d} + \mathbf{w}^T\mathbf{R}_{\mathbf{x}}\mathbf{w} \\
&= \sigma_d^2 - 2\mathbf{p}_{\mathbf{x}d}^T\mathbf{R}_{\mathbf{x}}^{-1}\mathbf{p}_{\mathbf{x}d} + \mathbf{p}_{\mathbf{x}d}^T\mathbf{R}_{\mathbf{x}}^{-1}\cancelto{\mathbf{I}}{\mathbf{R}_{\mathbf{x}}\mathbf{R}_{\mathbf{x}}^{-1}}\mathbf{p}_{\mathbf{x}d} \qquad (151) \\
&\boxed{= \sigma_d^2 - \mathbf{p}_{\mathbf{x}d}^T\mathbf{R}_{\mathbf{x}}^{-1}\mathbf{p}_{\mathbf{x}d}}
\end{aligned}$$

Características do cálculo filtro ótimo

- Depende **somente das estatísticas de segunda ordem** dos sinais envolvidos
- Em geral, estimativas precisas de $\mathbf{R}_{\mathbf{x}}$ e $\mathbf{p}_{\mathbf{x}d}$ não são disponíveis na prática
- Considerando-se a ergodicidade, é possível utilizar médias temporais para estima-las
- Supõe-se que a inversa da matriz $\mathbf{R}_{\mathbf{x}}$ existe. Na prática, resolve-se o sistema linear $\mathbf{R}_{\mathbf{x}}\mathbf{w}_{\text{opt}} = \mathbf{p}_{\mathbf{x}d}$.
- Caso do combinador linear: mesma solução. A diferença reside no cálculo das correlações
- Extensivo ao caso complexo. Definição do gradiente em relação a parâmetros complexos. Mesma solução!

Princípio da ortogonalidade

Uma questão interessante é verificada por meio da Eq. (146). Ela implica que o gradiente deve ser nulo em relação a **todos** os parâmetros w_i, ou seja

$$\nabla_{\mathbf{w}} \mathbb{E}\left\{e^2(n)\right\} = 0 \Rightarrow \mathbb{E}\left\{\frac{\partial e^2(n)}{\partial w_i}\right\} = 0 \; \forall i = 0, \ldots, M-1 \quad (152)$$

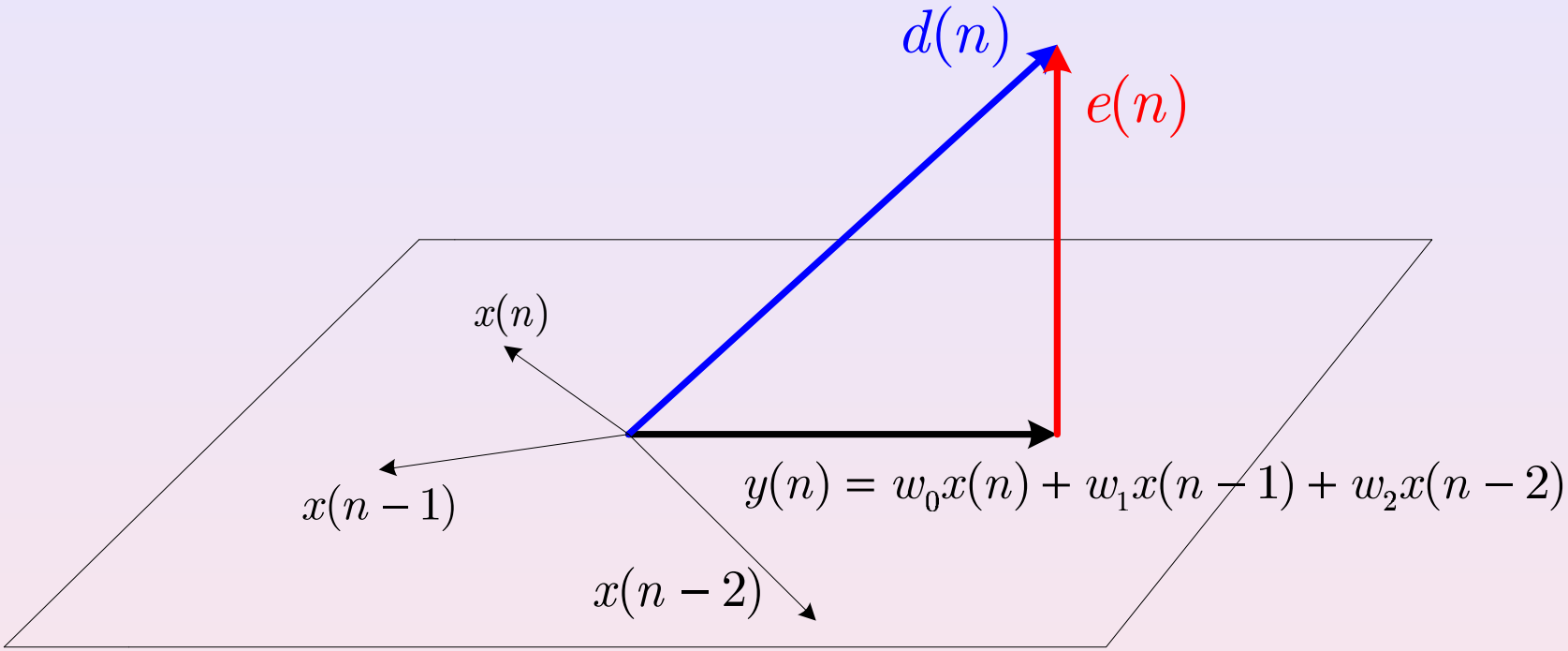
Lembrando que $e(n) = \left[d(n) - \sum_{i=0}^{M-1} w_i x(n-i)\right]$, então temos,

$$\mathbb{E}\left\{2 \cdot \frac{\partial e(n)}{\partial w_i} \cdot e(n)\right\} = 0 \quad (153)$$

$$\boxed{\mathbb{E}\left\{x(n-i) \cdot e(n)\right\} = 0 \; \forall i}$$

⚠ Ou seja, $x(n-i)$ e $e(n)$ são **ortogonais**!

Princípio da ortogonalidade - cont.

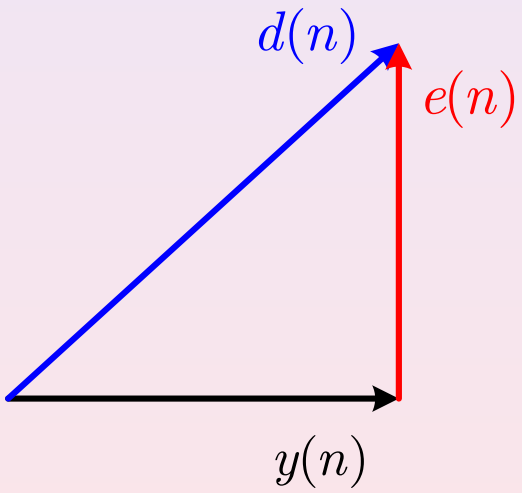


Critério de minimização do erro quadrático médio equivale a um critério de ortogonalização!

Princípio da ortogonalidade - cont.

Desta forma, estamos interessados em $d(n)$ colinear a $y(n)$ pois nesta condição

$$\exists\ \mathbf{w} \mid e(n) = 0 \Rightarrow \mathbb{E}\left\{e^2(n)\right\} = 0$$



Minimizar erro quadrático médio

$\Updownarrow$

Tornar erro ortogonal à saída do filtro

Comportamento da curva MSE (*mean square error*)

Meta: observar como se comporta o filtro quando está em torno da solução ótima.

Tomando-se a curva fornecida pelo erro médio quadrado temos a seguinte expressão:

$$\varepsilon = \mathbb{E}\left\{e^2(n)\right\} = \sigma_d^2 - 2\mathbf{w}^T\mathbf{p}_{\mathbf{x}d} + \mathbf{w}^T\mathbf{R}_{\mathbf{x}}\mathbf{w}$$

Definindo $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_{\text{opt}}$, e substituindo em (145) temos

$$\begin{aligned}\varepsilon &= \sigma_d^2 - 2(\Delta\mathbf{w} + \mathbf{w}_{\text{opt}})^T\mathbf{p}_{\mathbf{x}d} + (\Delta\mathbf{w} + \mathbf{w}_{\text{opt}})^T\mathbf{R}_{\mathbf{x}}(\Delta\mathbf{w} + \mathbf{w}_{\text{opt}}) \\ &= \underbrace{\sigma_d^2 - 2\mathbf{w}_{\text{opt}}^T\mathbf{p}_{\mathbf{x}d} + \mathbf{w}_{\text{opt}}^T\mathbf{R}_{\mathbf{x}}\mathbf{w}_{\text{opt}}}_{\varepsilon_{\text{min}}} - 2\Delta\mathbf{w}^T\mathbf{p}_{\mathbf{x}d} + \Delta\mathbf{w}^T\mathbf{R}_{\mathbf{x}}\Delta\mathbf{w} \\ &\quad + \Delta\mathbf{w}^T\mathbf{R}_{\mathbf{x}}\mathbf{w}_{\text{opt}} + \mathbf{w}_{\text{opt}}\mathbf{R}_{\mathbf{x}}\Delta\mathbf{w}\end{aligned} \tag{154}$$

continuando...

$$\begin{aligned}\varepsilon &= \varepsilon_{\min} - 2\Delta\mathbf{w}^T\mathbf{p}_{\mathbf{x}d} + \Delta\mathbf{w}^T\mathbf{R}_{\mathbf{x}}\Delta\mathbf{w} + \Delta\mathbf{w}^T\mathbf{R}_{\mathbf{x}}\mathbf{R}_{\mathbf{x}}^{-1}\mathbf{p}_{\mathbf{x}d} \\ &\quad + \mathbf{p}_{\mathbf{x}d}^T\mathbf{R}_{\mathbf{x}}^{-1}\mathbf{R}_{\mathbf{x}}\Delta\mathbf{w} \\ &= \varepsilon_{\min} - 2\Delta\mathbf{w}^T\mathbf{p}_{\mathbf{x}d} + \Delta\mathbf{w}^T\mathbf{R}_{\mathbf{x}}\Delta\mathbf{w} + \Delta\mathbf{w}^T\mathbf{p}_{\mathbf{x}d} + \mathbf{p}_{\mathbf{x}d}^T\Delta\mathbf{w} \\ &= \varepsilon_{\min} + \Delta\mathbf{w}^T\mathbf{R}_{\mathbf{x}}\Delta\mathbf{w}\end{aligned} \tag{155}$$

Podemos ainda diagonalizar a matriz $\mathbf{R_x} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$ em que $\mathbf{Q}$ é a matriz (ortogonal) dos autovetores de $\mathbf{R_x}$ dada por

$$\mathbf{Q} = \left[\begin{array}{cccc} \mathbf{q}_1 & \mathbf{q}_2 & \cdots & \mathbf{q}_M \end{array} \right]$$

e

$$\mathbf{\Lambda} = \left[\begin{array}{ccc} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_M \end{array} \right]$$

daí, pode-se escrever

$$\varepsilon = \varepsilon_{\text{min}} + \Delta\mathbf{w}^T\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T\Delta\mathbf{w} \tag{156}$$

Define-se ainda o vetor $\mathbf{v}$ de parâmetros v_i tal que

$$\mathbf{v} = \mathbf{Q}^T \Delta \mathbf{w} \tag{157}$$

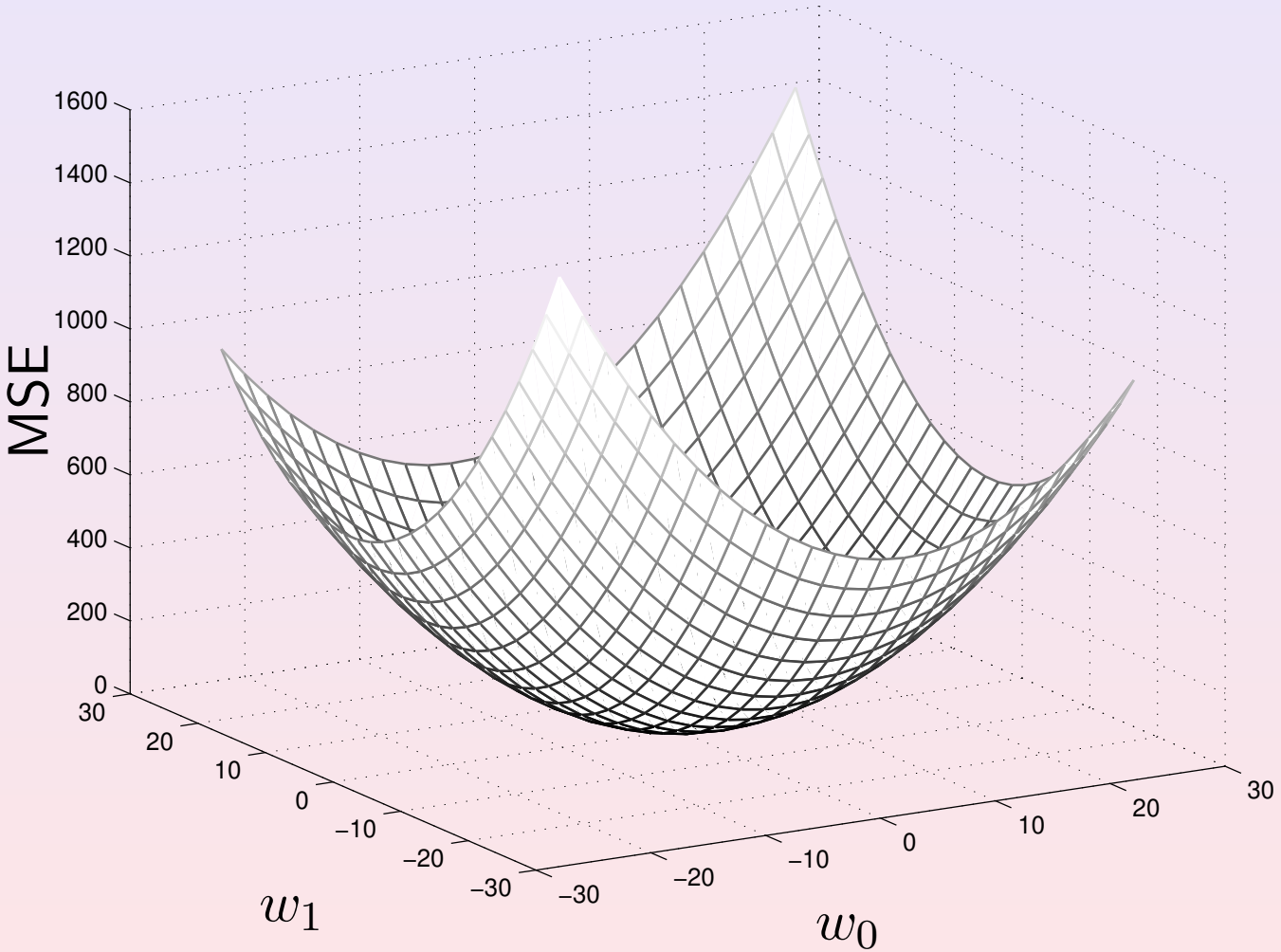
o que leva a

$$\varepsilon = \varepsilon_{\text{min}} + \mathbf{v}^T \mathbf{\Lambda} \mathbf{v} \tag{158}$$

Vantagem: $\mathbf{\Lambda}$ é uma matriz diagonal ao passo que $\mathbf{R_x}$ não.

Comportamento da curva MSE (*mean square error*) - cont.

Filtro com dois coeficientes

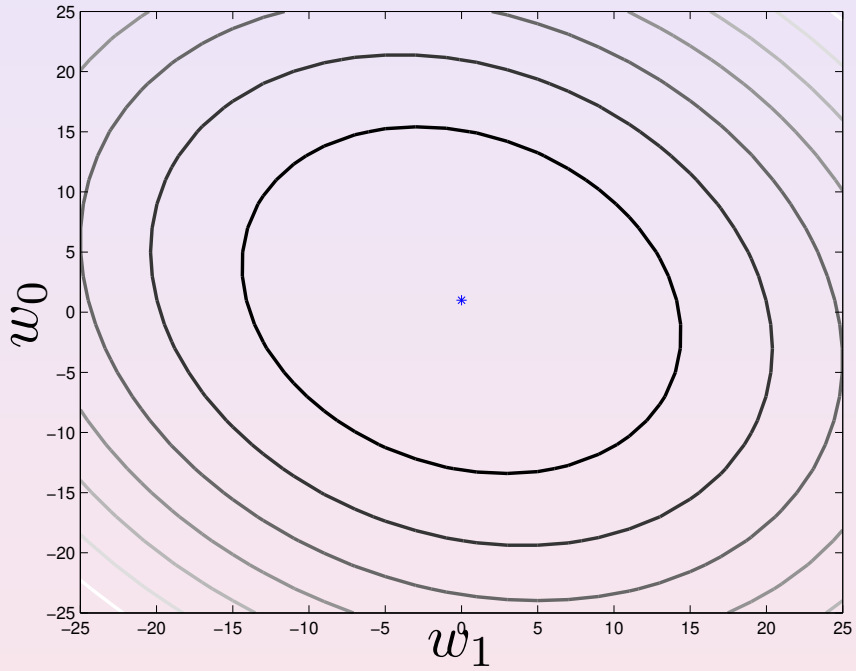


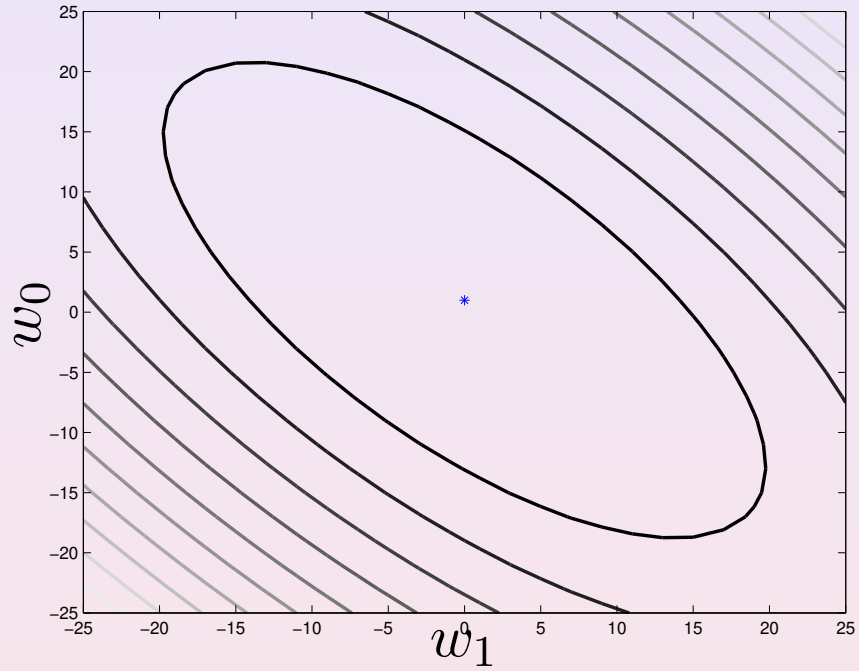
$$\varepsilon(v_0, v_1) = \varepsilon_{\text{min}} + \lambda_1 v_1^2 + \lambda_2 v_2^2 \tag{159}$$

Lembrando que: $\mathcal{C}(\mathbf{R_x}) = \frac{\lambda_{\max}}{\lambda_{\min}}$ (número de condicionamento)

1. $\mathcal{C}(\mathbf{R_x}) \approx 1 \Leftrightarrow$ curvas MSE mais circulares $\Leftrightarrow x(n)$ tem espectro mais plano
2. $\mathcal{C}(\mathbf{R_x}) \gg 1 \Leftrightarrow$ curvas MSE mais "elípticas" $\Leftrightarrow x(n)$ tem espectro com picos

Curvas de nível em função do número de condicionamento



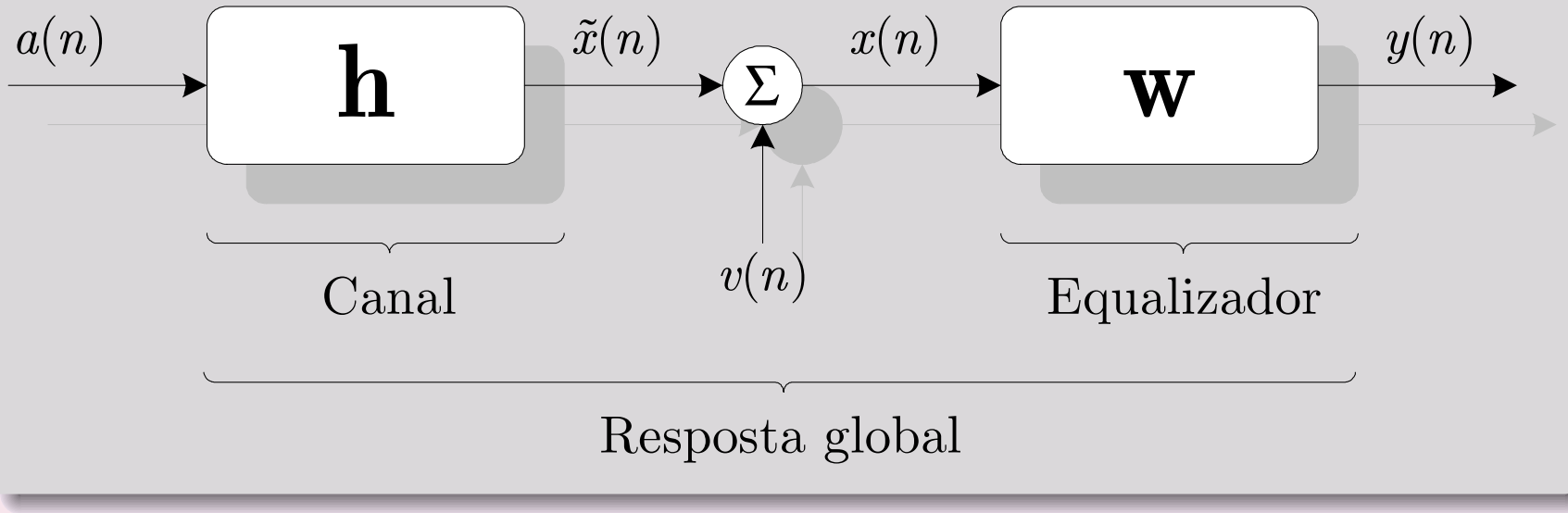


$$\mathbb{C}\left(\mathbf{R_x}\right) = 1.5$$

$$\mathbb{C}\left(\mathbf{R_x}\right) = 5.6667$$

Aplicação - equalização de canais

Exemplo



Exemplo - cont.

Buscar o filtro linear ótimo, no sentido da minimização do erro quadrático médio, que inverta o seguinte canal:

$$H(z) = 1 + 0.7z^{-1}$$

Assume-se que o alfabeto de transmissão é BPSK, ou seja, $a(n) \in \{-1, +1\}$ com igual probabilidade e que o sinal desejado será o do instante atual, ou seja, $d(n) = a(n)$.

Para este problema temos que calcular $\mathbf{R}_{\widetilde{\mathbf{x}}}$ e $\mathbf{p}_{\mathbf{x}d}$ dadas para o problema em questão por

$$\mathbf{R}_{\widetilde{\mathbf{x}}} = \begin{bmatrix} 1.49 & 0.7 \\ 0.7 & 1.49 \end{bmatrix} \quad \text{e} \quad \mathbf{p}_{\mathbf{x}d} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{160}$$

Exemplo - cont.

Supondo ainda um ruído aditivo branco na entrada do filtro (receptor) de SNR $= 20$ dB, teremos então alguma perturbação no sinal recebido e em conseqüência na sua correlação. Assim, teremos $\mathbf{R_x} = (\mathbf{R}_{\tilde{\mathbf{x}}} + \sigma_v \mathbf{I})$.

De posse de tais quantidades, podemos então calcular

$$\mathbf{w}_{\text{opt}} = \mathbf{R}_{\mathbf{x}}^{-1}\mathbf{p}_{\mathbf{x}d} = \begin{bmatrix} 0.8579 \\ -0.4058 \end{bmatrix} \tag{161}$$

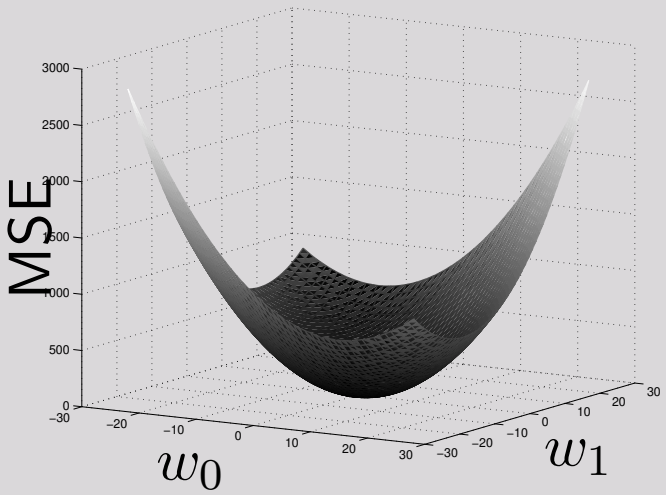
E podemos ainda calcular a resposta global, fruto da convolução do canal com o equalizador:

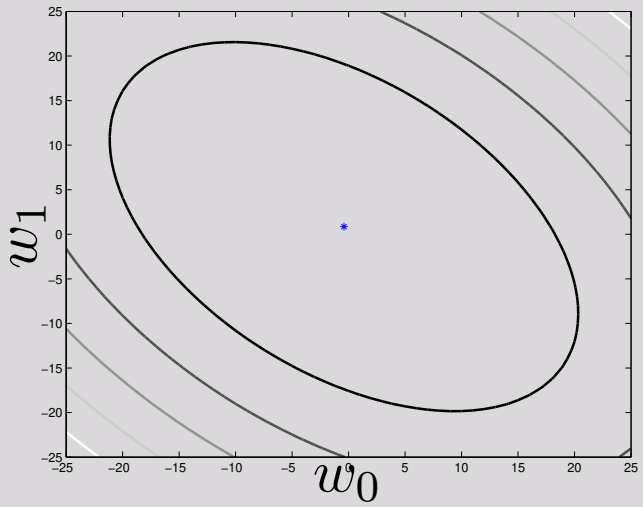
$$\mathbf{g} = \begin{bmatrix} 0.8579 & 0.1948 & -0.2840 \end{bmatrix}^T$$

Exemplo - cont.

Ainda temos,

$$\varepsilon_{\text{min}} = 0.1421$$





$$\mathcal{C}\left(\mathbf{R_x}\right) = 2.7890$$

Parte V

Predição de Sinais Estacionários

O problema de predição

Definição

Predição: estimar uma amostra $x(n)$ a partir de um conjunto de valores conhecidos deste sinal.

É, em essência, um processo de filtragem no qual o sinal desejado é uma nova amostra da seqüência de entrada do filtro.

Linear: $\widehat{x}(n)$ é uma combinação linear dos valores passados.

Forward ou **backward**:

1. *Forward*: prever uma amostra futura a partir de uma coleção de amostras passadas
2. *Backward*: prever uma amostra no passado (desconhecida) a partir de um conjunto de amostras, inclusive presente

De passo k

1. *Forward:*

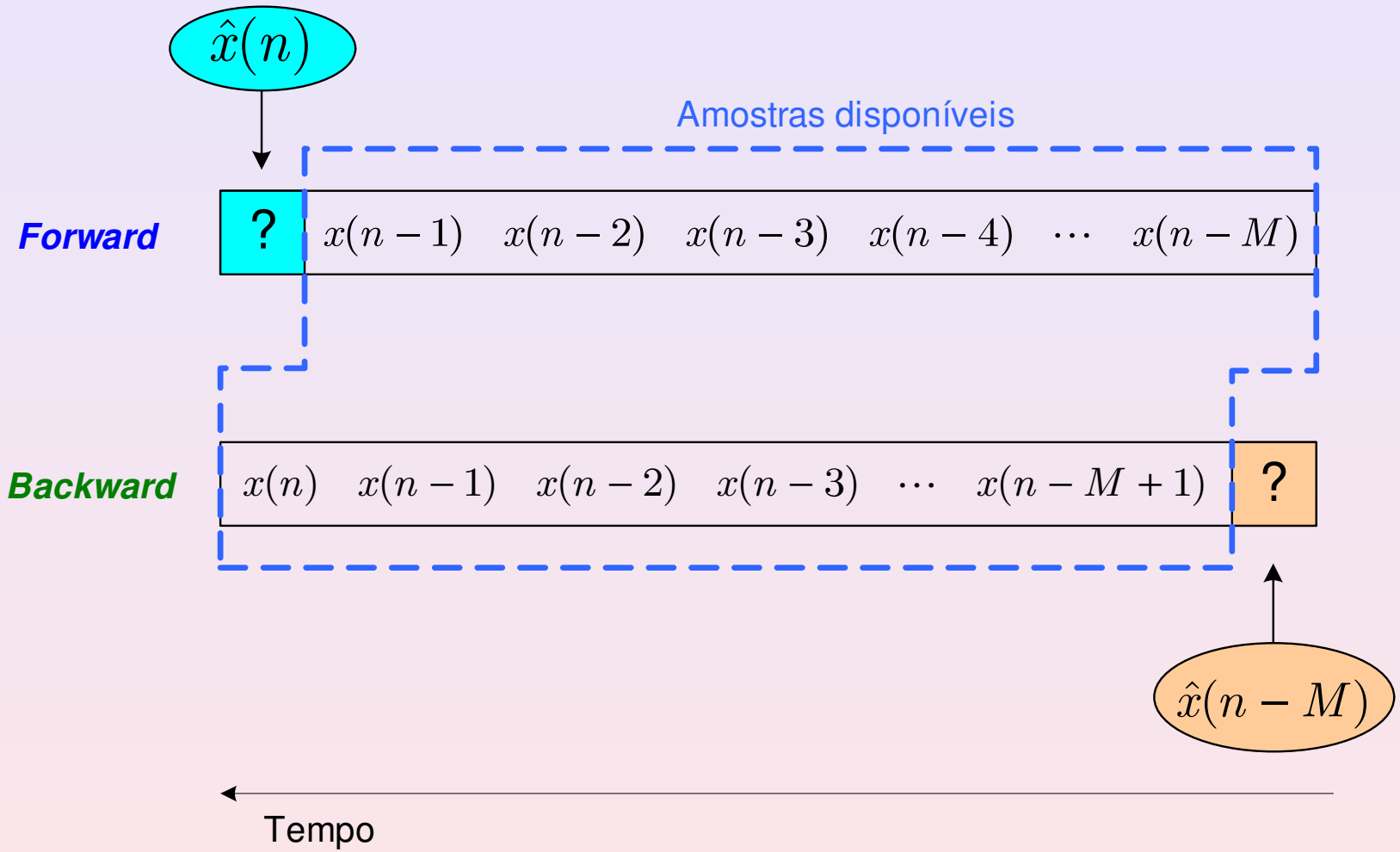
$$\widehat{x}(n) = \sum_{i=k}^{M+k-1} w_{f,i} \cdot x(n-i) \tag{162}$$

2. *Backward:*

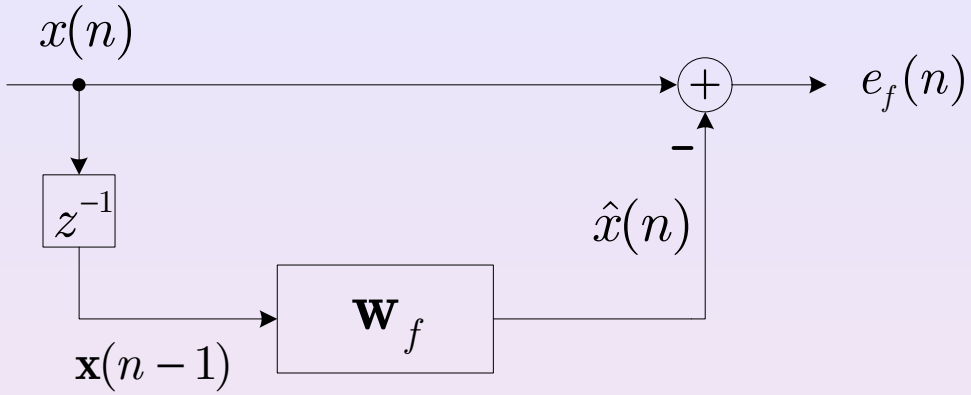
$$\widehat{x}(n-M-k+1) = \sum_{i=1}^{M-1} w_{b,i} \cdot x(n-i+1) \tag{163}$$

Inicialmente, nos deteremos nos processos de predição de passo unitário.

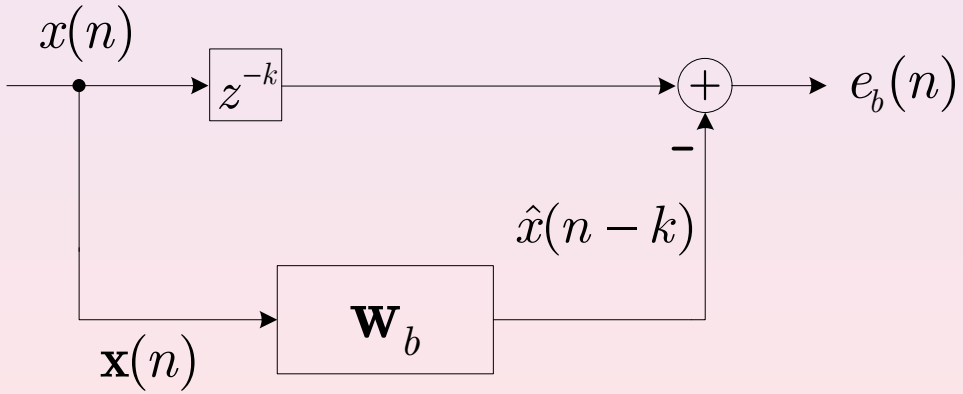
O problema de predição - cont.



O problema de predição - cont.



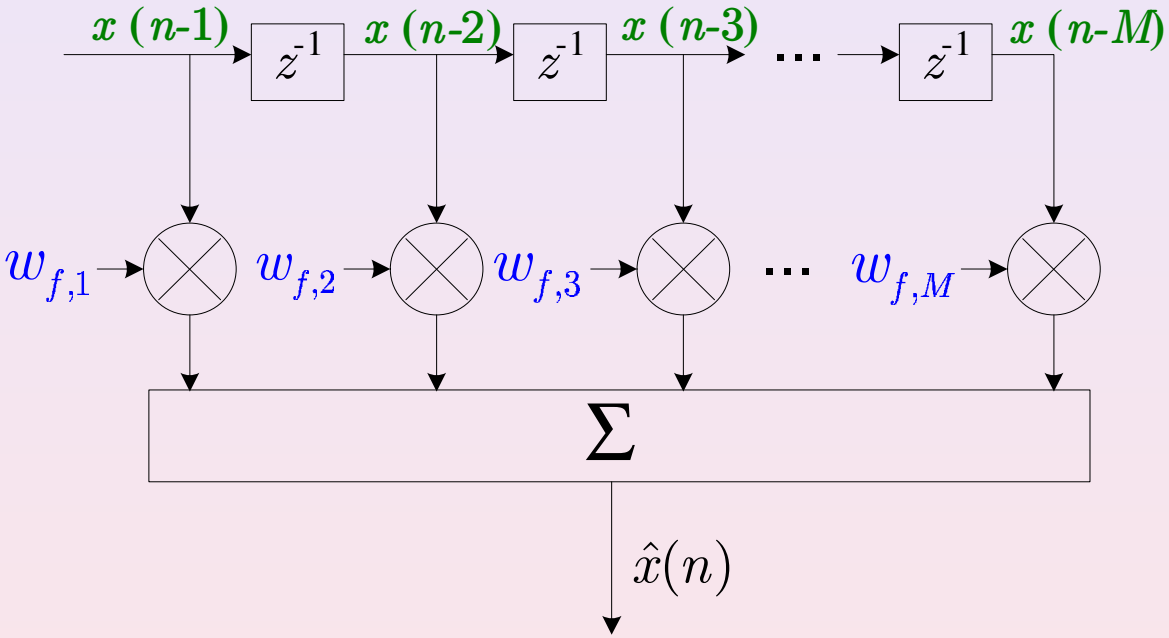
(a) Forward



(b) Backward

Estrutura do filtro

Para o preditor, usamos um filtro de linha de atraso (filtro FIR)



Para o caso de predição de um passo *forward* temos então

$$\widehat{x}(n) = \mathbf{w}_f^T \cdot \mathbf{x}(n-1) \tag{164}$$

em que $\mathbf{x}(n-1) = \left[\begin{array}{ccc} x(n-1) & \cdots & x(n-M) \end{array} \right]^T$ e
$\mathbf{w}_f = \left[\begin{array}{ccc} w_{f,1} & \cdots & w_{f,M} \end{array} \right]^T$

E também

$$e_f(n) = x(n) - \widehat{x}(n) \tag{165}$$

Meta: tornar $\widehat{x}(n)$ o mais similar possível de $x(n)$.

Critério

Minimização do erro de predição quadrático!

$$\min_{\mathbf{w}_f} \mathbb{E}\left\{e_f^2(n)\right\} \tag{166}$$

Filtragem ótima × predição

Analogia

O problema é um caso particular da filtragem de Wiener em que

1. $d(n) \leftrightarrow x(n)$,
2. $\mathbf{x}(n) \leftrightarrow \mathbf{x}(n-1)$,
3. $w_i \ (i = 0, \ldots, M-1) \ \leftrightarrow w_{f,i} \ (i = 1, \ldots, M-1)$ e
4. $y(n) \leftrightarrow \widehat{x}(n)$

Hipóteses

1. $x(n)$ é um sinal estacionário no sentido amplo, que implica

$$r_x(i,j) = \mathbb{E}\left\{x(n-i)x(n-j)\right\} = r(i-j) \tag{167}$$

2. $x(n)$ tem média nula e variância σ_x^2

Sabendo que

$$\begin{aligned} e_f(n) &= x(n) - \widehat{x}(n) \\ &= x(n) - \mathbf{w}_f^T \cdot \mathbf{x}(n-1) \\ e_f^2(n) &= x^2(n) - 2\mathbf{w}_f^T \cdot \mathbf{x}(n-1)x(n) + \mathbf{w}_f^T \cdot \mathbf{x}(n-1)\mathbf{x}(n-1)^T \mathbf{w}_f \end{aligned} \tag{168}$$

Temos então

$$\begin{aligned} \mathbb{E}\left\{e_f^2(n)\right\} &= \mathbb{E}\left\{x^2(n)\right\} - 2\mathbf{w}_f^T \mathbb{E}\left\{\mathbf{x}(n-1)x(n)\right\} \\ &\quad + \mathbf{w}_f^T \mathbb{E}\left\{\mathbf{x}(n-1)\mathbf{x}(n-1)^T\right\} \mathbf{w}_f \\ &= r_x(0) - 2\mathbf{w}_f^T \mathbf{r}_{x,f} + \mathbf{w}_f^T \mathbf{R}_{\mathbf{x}} \mathbf{w}_f \end{aligned} \tag{169}$$

em que $\mathbf{R}_{\mathbf{x}}$ e $\mathbf{r}_{x,f}$ são a matriz de autocorrelação de $\mathbf{x}(n-1)$ e o vetor de correlação entre $\mathbf{x}(n-1)$ e $x(n)$, respectivamente.

Filtro preditor ótimo - cont.

Para achar o valor ótimo, temos

$$\frac{\partial \mathbb{E}\left\{e_f^2(n)\right\}}{\partial \mathbf{w}_f} = 0 \tag{170}$$
$$-2\mathbf{r}_{x,f} + 2\mathbf{R}_{\mathbf{x}}\mathbf{w}_f = 0$$

O que nos fornece

Preditor *forward* ótimo

$$\mathbf{w}_{f,\text{opt}} = \mathbf{R}_{\mathbf{x}}^{-1} \cdot \mathbf{r}_{x,f} \tag{171}$$

em que

$$\mathbf{R}_{\mathbf{x}} = \begin{bmatrix} r_x(0) & \cdots & r_x(M-1) \\ \vdots & \ddots & \vdots \\ r_x(M-1) & \cdots & r_x(0) \end{bmatrix} \quad \text{e} \quad \mathbf{r}_{x,f} = \begin{bmatrix} r_x(1) \\ \vdots \\ r_x(M) \end{bmatrix}$$

Filtro preditor ótimo - cont.

- Se $x(n)$ é branco $\Rightarrow r_x(i) = 0$
- $\mathbf{w}_f$ é ótimo quando $\mathbb{E}\left\{e_f^2(n)\right\}$ é mínimo

Ou seja, podemos fazer

$$\frac{\partial \mathbb{E}\left\{e_f^2(n)\right\}}{\partial \mathbf{w}_f} = 0 \quad \Rightarrow \mathbb{E}\left\{\frac{\partial e_f^2(n)}{\partial \mathbf{w}_f}\right\} = 0$$
$$\mathbb{E}\left\{2e_f(n)\frac{\partial e_f(n)}{\partial \mathbf{w}_f}\right\} = 0 \tag{172}$$

Filtro preditor ótimo - cont.

Lembrando que $e_f(n) = x(n) - \sum_{i=i}^{M} w_{f,i} \cdot x(n-i)$, temos então

$$\boxed{\mathbb{E}\left\{e_f(n)x(n-i)\right\} = 0 \quad \forall\, i = 1, \ldots, M} \tag{173}$$

Mas sabe-se ainda que $e_f(n-k) = x(n-k) - \mathbf{w}_f^T \mathbf{x}(n-k-1)$, e substituindo-se em (173) temos

$$\mathbb{E}\left\{e_f(n) \cdot \left[e_f(n-i) + \mathbf{w}_f^T \mathbf{x}(n-i-1)\right]\right\} = 0$$

$$\mathbb{E}\left\{e_f(n)e_f(n-i)\right\} + \mathbf{w}_f^T \underbrace{\mathbb{E}\left\{\mathbf{x}(n-i-1)\right\}}_{\mathbf{0}} = 0 \tag{174}$$

$$\boxed{\mathbb{E}\left\{e_f(n)e_f(n-i)\right\} = 0 \quad \forall\, i = 1, \ldots, M}$$

Filtro de branqueamento!

Filtro preditor ótimo - cont.

Conseqüência

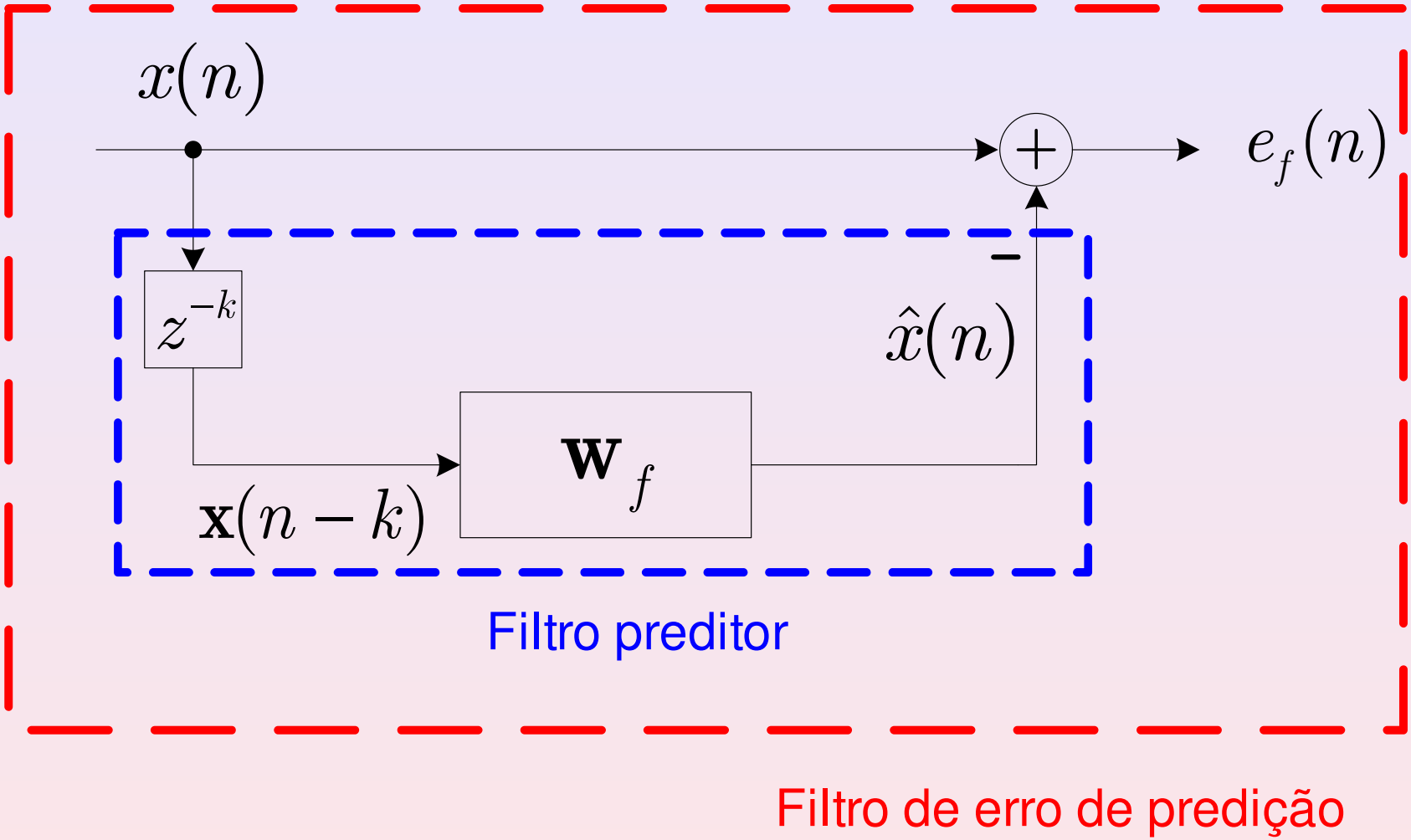
Preditor de ordem infinita $\leftrightarrow$ Ruído branco no erro de saída

Meta do filtro

O preditor busca então fornecer um sinal $e_f(n)$ ortogonal entre si (branco): filtro de erro de predição é um "filtro branqueador"

$$M \to \infty \quad \Rightarrow \begin{cases} \mathbb{E}\{e_f(n)e_f(n-i)\} = 0 \quad \forall i \geq 1 \\ e(n) = 0 \\ e(n) = \text{ruído branco} \end{cases}$$

Filtro preditor ótimo - cont.



Filtro preditor *backward*

De forma análoga ao caso *forward*, podemos derivar a equação do filtro de predição *backward*. Definindo

$$e_b(n) = x(n-M) - \mathbf{w}_b^T \mathbf{x}(n) \tag{175}$$

em que $\mathbf{x}(n) = \left[\begin{array}{ccc} x(n) & \cdots & x(n-M+1) \end{array} \right]^T$

Teremos então

$$\begin{aligned} e_b^2(n) &= x^2(n-M) - 2\mathbf{w}_b^T \mathbf{x}(n) x(n-M) + \mathbf{w}_b^T \mathbf{x}(n) \mathbf{x}(n)^T \mathbf{w}_b \\ \Rightarrow \mathbb{E}\left\{e_b^2(n)\right\} &= r_x(0) - 2\mathbf{w}_b^T \mathbf{r}_{x,b} + \mathbf{w}_b^T \mathbf{R}_{\mathbf{x}} \mathbf{w}_b \end{aligned} \tag{176}$$

Assim, fazendo $\frac{\partial \mathbb{E}\{e_b^2(n)\}}{\partial \mathbf{w}_b}$, teremos

$$\boxed{\mathbf{w}_{b,\text{opt}} = \mathbf{R}_{\mathbf{x}}^{-1}\mathbf{r}_{x,b}} \tag{177}$$

em que

$$\mathbf{R}_{\mathbf{x}} = \begin{bmatrix} r_x(0) & \cdots & r_x(M-1) \\ \vdots & \ddots & \vdots \\ r_x(M-1) & \cdots & r_x(0) \end{bmatrix} \quad \text{e} \quad \mathbf{r}_{x,b} = \begin{bmatrix} r_x(M) \\ \vdots \\ r_x(1) \end{bmatrix}$$

Com isso, pode-se visualizar que o preditor ótimo *backward* tem a mesma estrutura do preditor forward, a menos da organização das correlações nos vetores $\mathbf{r}_{x,b}$ e $\mathbf{r}_{x,f}$.

Filtro preditor *backward* - cont.

Se escrevermos as equações dos filtros ótimos, *forward* e *backward*, na sua forma direta, ou seja

$$\mathbf{R_x}\mathbf{w}_f = \mathbf{r}_{x,f} \tag{178a}$$

$$\mathbf{R_x}\mathbf{w}_b = \mathbf{r}_{x,b} \tag{178b}$$

e notarmos que

$$\mathbf{r}_{x,f} = (\mathbf{r}_{x,b})^R \tag{179}$$

em que $(\cdot)^R$ é a operação de reversão no tempo, podemos deduzir que

$$\boxed{\mathbf{w}_f = (\mathbf{w}_b)^R} \tag{180}$$

Ou seja, é possível encontrar facilmente o preditor *backward* à partir de sua versão *forward* e vice-versa.

Resumo

- Preditores direto e reverso ótimos têm a mesma estrutura da filtragem de Wiener
- Possibilidade de encontrar o preditor direto a partir do reverso e vice-versa

Parte VI

Teoria da Detecção

Teoria da detecção

1. **Definição:** Aplicação da teoria da decisão para detecção de sinais, também por vezes chamada *teoria estatística da detecção* ou ainda *teoria estatística da decisão*.
2. **Objetivo:** Descobrir qual o mais provável sinal a partir de uma observação ou conjunto de observações.
3. **Ferramentas:** Critérios que fornecem métricas para calcular qual hipótese de detecção é a mais provável dentre um conjunto.

Testes de hipóteses

- Objetiva testar algumas hipóteses e descobrir qual a mais provável
- Elaboração de um **limiar**, segundo algum critério
- Diferentes critérios possuem diferentes limiares e regras de decisão (veremos a seguir)
- **Requisitos**: cada critério com um certo conjunto de requisitos
- Aplicações: variam dependendo da adequação dos requisitos necessários.
- Na seqüência, descreveremos os critérios clássicos em teoria da estimação

Classificação

Teste binário: duas hipóteses são testadas e uma delas é escolhida (decidida) como sendo a verdadeira (ou mais provável)

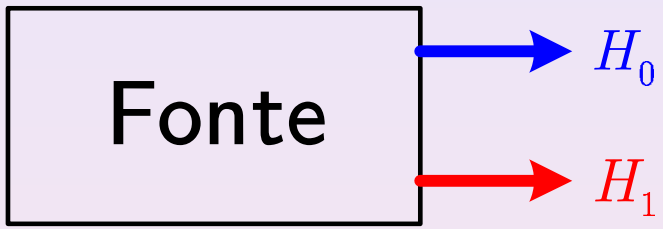
Teste M-ário: neste caso, temos M hipóteses entre as quais escolhemos (decisão) uma delas como sendo a verdadeira. É um teste mais complexo pois para cada decisão temos M possibilidades

Teste simples: neste tipo de teste, as hipóteses são caracterizadas por apenas um parâmetro.

Teste composto: teste no qual cada hipótese é caracterizada por um conjunto de parâmetros.

Inicialmente, iremos tratar dos casos de **testes binários simples**.

Modelo: fonte capaz de emitir dois valores distintos

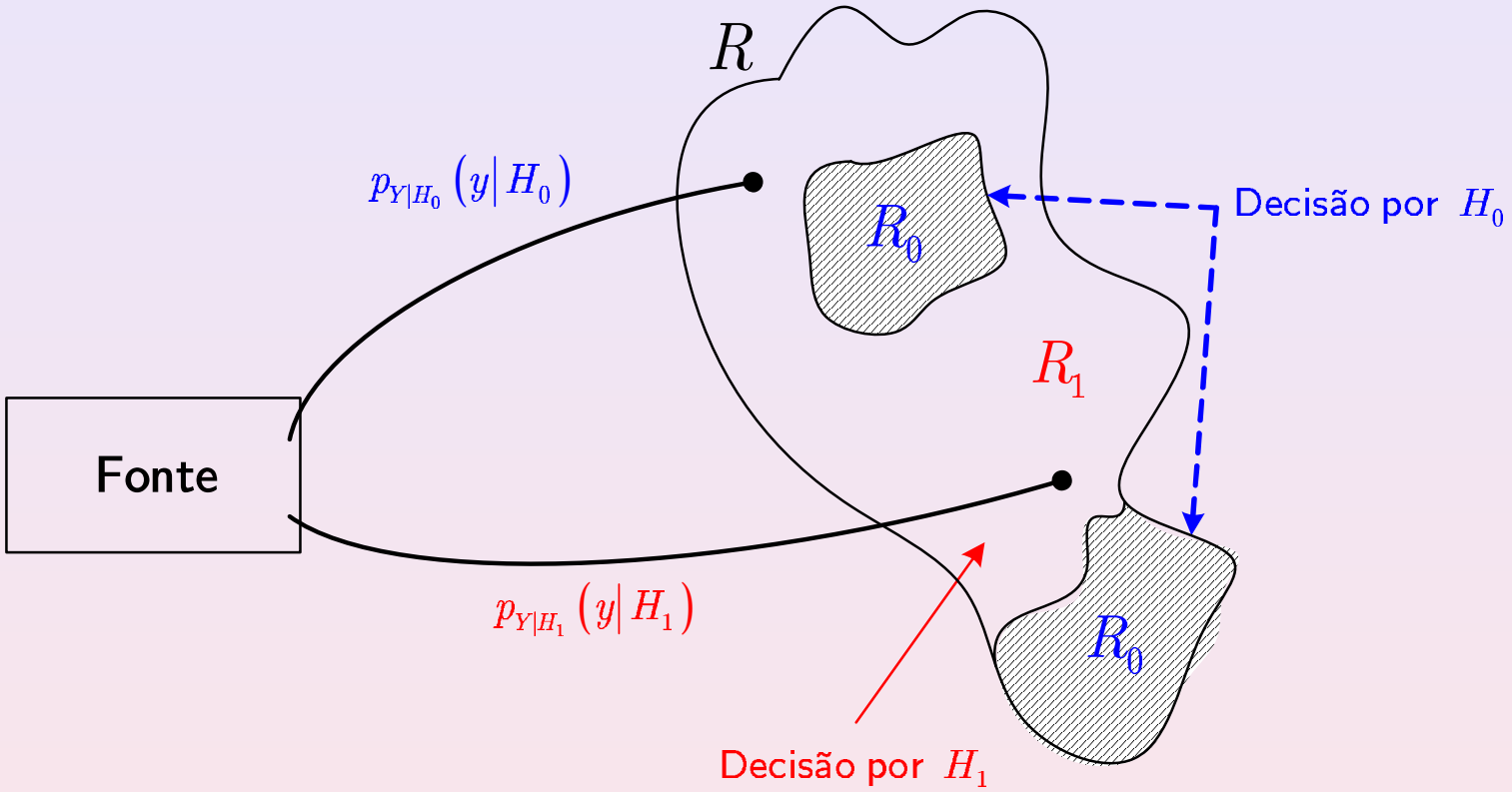


Fonte para hipóteses binárias.

- Cada hipótese a uma das possíveis saídas da fonte
- Representação por variáveis aleatórias (v.a.)
- **Decisão**: H_0 ou H_1 é verdadeira?

Assumindo-se que a decisão é tomada em uma única amostra:

1. o domínio dos valores da v.a. Y constitui o espaço de observação R;
2. particiona-se R em duas regiões (duas hipóteses);
3. se Y está na região R_0 decisão em favor de H_0;
4. se Y está na região R_1 decisão em favor de H_1.



Regiões de decisão.

Decisão deve ser tomada com base em alguma medida: **probabilidade!**

Necessidade de modelar a ocorrência de um valor da variável Y com as hipóteses (condicional)

- $p_{Y|H_0}(y|H_0)$ - relativo à hipótese H_0
- $p_{Y|H_1}(y|H_1)$ - relativo à hipótese H_1

Além disso, o espaço de observação é dado pela união dos espaços das hipóteses, ou seja,

$$R = R_0 \cup R_1 \tag{181}$$

Testes de hipóteses - cont.

Deste modo, teremos para o caso de hipóteses binárias, quatro possibilidades:

1. decidir por H_0 quando H_0 é verdadeira;
2. decidir por H_0 quando H_1 é verdadeira;
3. decidir por H_1 quando H_0 é verdadeira;
4. decidir por H_1 quando H_1 é verdadeira.

Notação

- D_0 implica em dizer que foi escolhida (decidida por) H_0 como verdadeira e
- D_1 implica em dizer que foi escolhida (decidida por) H_1 como verdadeira.

Discutiremos quais os critérios que podem ser escolhidos para guiar a tomada de decisão

Problema

Dadas duas hipóteses H_0 e H_1, a decisão consiste em determinar qual é a mais provável, a partir de uma observação ou conjunto de observações.

Se tivermos

$$p(H_0|y) > p(H_1|y), \tag{182}$$

dizemos que a hipótese H_0 é a mais provável, e o inverso, ou seja, H_1 é a mais provável caso

$$p(H_1|y) > p(H_0|y). \tag{183}$$

Usando o Teorema de Bayes, temos

$$p(H_0|y) = \frac{p(y|H_0) \cdot p(H_0)}{p(y)}, \tag{184}$$

em que $p(H_0)$ e $p(H_1)$, as probabilidades das hipóteses, são denominadas probabilidades *a priori* e denotaremos aqui $p(H_0) = \pi_0$ e $p(H_1) = \pi_1$.

Ainda, pelo Teorema da Probabilidade Total sabemos que

$$p(y) = p(y|H_0) \cdot \pi_0 + p(y|H_1) \cdot \pi_1 \tag{185}$$

Aplicando a regra de Bayes na Equação (183) temos

$$p(y|H_1) \cdot \pi_1 > p(y|H_0) \cdot \pi_0, \tag{186}$$

que implica que H_1 será assumida como verdadeira implicando em D_1, e H_0 só será assumida verdadeira, implicando em D_0,caso a inequação seja revertida.

Logo, podemos escrever:

$$p(y|H_1) \cdot \pi_1 \underset{D_0}{\overset{D_1}{\gtrless}} p(y|H_0) \cdot \pi_0. \tag{187}$$

Entretanto, é mais usual encontrar a Equação (187) na seguinte forma:

$$\boxed{L(y) = \frac{p(y|H_1)}{p(y|H_0)} \underset{D_0}{\overset{D_1}{\gtrless}} \frac{\pi_0}{\pi_1}.} \tag{188}$$

A função $L(y)$ é chamada de **razão de verossimilhança**.

O valor $\frac{\pi_0}{\pi_1} = \eta$ é geralmente chamado de limiar e é um valor fixo associado à detecção MAP.

Critério MAP - cont.

Deve-se notar que na Equação (188) não há decisão para $L(y) = \eta$. Desta forma, para evitar problemas com singularidades, geralmente se escreve a Eq. (188) como:

$$L(y) \underset{D_0}{\overset{D_1}{\gtrless}} \eta, \tag{189}$$

implicando que a decidiremos por D_1 caso $L(y) \geq \eta$.

Ainda, como a função $\ln(\cdot)$ é monotônica, uma decisão equivalente pode ser tomada se usarmos

$$\mathcal{L}(y) = \ln[L(y)] \underset{D_0}{\overset{D_1}{\gtrless}} \ln[\eta], \tag{190}$$

que é chamada de *log-verossimilhança*.

No caso de $\pi_1 = \pi_0$, temos $\eta = 1 \Rightarrow \ln[\eta] = 0$.

Adicionalmente, é interessante notar que em engenharia, os processos de famílias de exponenciais são bastante usuais sendo normalmente usada a log-verossimilhança.

Exemplo: decisão de sinal binário em AWGN

Problema: Decidir, a partir de medidas ruidosas, qual sinal foi enviado.

Assim, se tivermos $s_0 = -b$ e $s_1 = b$, para $b > 0$ e considerando $\pi_1 = \pi_0 = 0.5$, temos para o sinal observado a seguinte expressão

$$y = s_i + n, \tag{191}$$

em que $n \sim \mathcal{N}(0, \sigma^2)$.

Uma forma natural é construir as hipóteses como:

$H0$: s_0 é detectado;

$H1$: s_1 é detectado.

Exemplo: decisão de sinal binário em AWGN - cont.

Para o ruído gaussiano em questão temos então as seguintes pdfs condicionais

$$p(y|H_0) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{|y+b|^2}{2\sigma^2}\right) \tag{192}$$

e

$$p(y|H_1) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{|y-b|^2}{2\sigma^2}\right). \tag{193}$$

Assim, temos que

$$L(y) = \exp\left(\frac{2yb}{\sigma^2}\right) \underset{D_0}{\overset{D_1}{\gtrless}} 1 \tag{194}$$

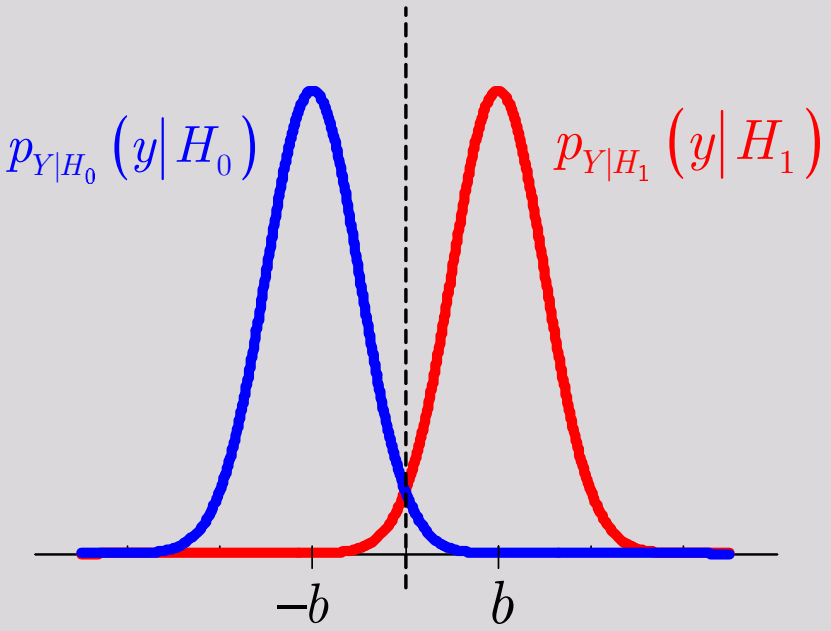
Exemplo: decisão de sinal binário em AWGN - cont.

e tomando a log-verossimilhança

$$\mathcal{L}(y) = \frac{2yb}{\sigma^2} \underset{D_0}{\overset{D_1}{\gtrless}} 0 \qquad \Rightarrow \qquad y \underset{D_0}{\overset{D_1}{\gtrless}} 0. \tag{195}$$

Exemplo: decisão de sinal binário em AWGN - cont.

A figura abaixo ilustra as densidades envolvidas neste processo, bem como o limiar.



Densidades no processo de detecção de um sinal corrompido por ruído gaussiano.

- Além do critério de decisão, saber qual a regra para decisão (considerando, ainda, duas hipóteses)
- D_0 é tomada quando y está no espaço de observação tal que $L(y) < \eta$ e
- D_1 quando y está no espaço associado à $L(y) \geq \eta$
- Usando a notação dos espaços de observação associados à D_0 e D_1 temos

$$\delta(y) = \begin{cases} 1, & y \in R_1 \\ 0, & y \in R_0 \end{cases} \tag{196}$$

E podemos então escrever a regra de decisão em função do limiar e da função de verossimilhança como

$$\delta_{\text{MAP}}(y) = \begin{cases} 1, & \forall y \in L(y) \geq \eta \\ 0, & \forall y \in L(y) < \eta \end{cases}, \tag{197}$$

que é a regra de decisão MAP.

Tipos de erro

Uma aplicação clássica da teria da detecção é na área de **radar**, da qual extraímos muito da nomenclatura usada em teoria da detecção.

Uma vez que um radar "ilumina" (jargão da área) um certo volume do espaço, o sinal de retorno pode ser modelado em termos de duas hipóteses:

(1) H_0 corresponde à "sem sinal" e

(2) H_1 corresponde ao "sinal presente" (alvo).

Neste caso, H_0 é referenciada como **hipótese nula** e H_1 como **hipótese alternativa**.

Tipos de erro - cont.

Define-se como P_{ij} a probabilidade de decidir por D_i, quando de fato a hipótese H_j é correta. Em termos da pdf temos:

$$P_{ij} = \int_{R_i} p(y|H_j)\, dy \tag{198}$$

Considerando hipóteses binárias, a terminologia histórica da área de radar é usada para três das probabilidades existentes.

Tipos de erro - cont.

Nomenclatura de erros

P_{01} probabilidade de perda (P_M - *miss* em inglês). Ou seja, o sinal está presente (H_1 está correta) e perdemos por escolher D_0;

P_{11} probabilidade de detecção (P_D). Decide por um sinal que de fato está presente;

P_{10} probabilidade de "falso alarme" (P_F). Decidir que há sinal quando de fato não há;

Não há nenhum nome específico para P_{00}.

Tipos de erro - cont.

Apenas duas destas probabilidades, P_{01} e P_{10} correspondem a erro. Em várias aplicações, no entanto, estamos interessados na probabilidade de erro média dada por

$$P_E = P_{10} \cdot \pi_0 + P_{01} \cdot \pi_1. \tag{199}$$

Curiosidade: a mesma notação de radar é empregada em testes biológicos.

Importante!

O desenvolvimento para minimizar a probabilidade de erro também leva ao detector MAP.

Critério de Bayes

- Algumas aplicações: diferentes decisões $\Rightarrow$ diferentes impactos (custos)
- Exemplo:testes médicos para doenças fatais.

Probabilidades

1. P_{00}: um correto diagnóstico da ausência de uma doença. Sem custos adicionais.
2. P_{01}: um incorreto diagnóstico de uma doença realmente existente. Custo alto (morte).
3. P_{10} e P_{11}: decisão de existir uma doença realizando o tratamento (possivelmente). Custos: medicamentos, internação, cirurgia, efeitos colaterais, etc.

O critério de Bayes trata de questões como esta introduzindo o conceito de **custo**.

Critério de Bayes - cont.

Seja C_{ij} o custo de decidir D_i quando a hipótese H_j é correta.

O risco médio $\mathcal{R}$, chamado de **risco de Bayes**, é dado por:

$$\mathcal{R} = [P_{00}C_{00} + P_{10}C_{10}] \cdot \pi_0 + [P_{01}C_{01} + P_{11}C_{11}] \cdot \pi_1 \qquad (200)$$

Deste modo, a meta agora é organizar a decisão para minimizar $\mathcal{R}$.

Critério de Bayes - cont.

Sabe-se ainda que

$$P_{11} = 1 - P_{01} \tag{201}$$

e

$$P_{10} = 1 - P_{00} \tag{202}$$

Substituindo as Eqs (201) e (202) na Eq. (200), temos:

$$\mathcal{R} = \pi_0 C_{10} + \pi_1 C_{11} - \pi_0 (C_{10} - C_{00}) P_{00} + \pi_1 (C_{01} - C_{11}) P_{01} \tag{203}$$

Mas ainda podemos escrever

$$P_{00} = \int_{R_0} p(y|H_0)\, dy \quad \text{e} \quad P_{01} = \int_{R_0} p(y|H_1)\, dy \tag{204}$$

Critério de Bayes - cont.

Assim,

$$\mathcal{R} = \pi_0 C_{10} + \pi_1 C_{11} + \int_{R_0} [\pi_0(C_{01} - C_{11})p(y|H_1) - \pi_0(C_{10} - C_{00})p(y|H_0)] \; dy \tag{205}$$

- Os dois primeiros termos são constantes;
- Minimizar $\mathcal{R}$ é equivalente a minimizar o integrando

Logo, seja a seguinte função

$$g(y) = \pi_1(C_{01} - C_{11}) \cdot p(y|H_1) - \pi_0(C_{10} - C_{00}) \cdot p(y|H_0) \quad (206)$$

Para minimizar $g(y)$, devemos ter

$$\pi_0(C_{10} - C_{00}) \cdot p(y|H_0) \geq \pi_1(C_{01} - C_{11}) \cdot p(y|H_1), \quad (207)$$

e daí

$$\boxed{L(y) = \frac{p(y|H_1)}{p(y|H_0)} \underset{D_0}{\overset{D_1}{\gtrless}} \frac{\pi_0(C_{10} - C_{00})}{\pi_1(C_{01} - C_{11})} = \eta_{\mathsf{B}}} \quad (208)$$

ou seja, D_0 é escolhida para $L(y) < \eta_{\mathsf{B}}$ (limiar de Bayes) e D_1 é escolhida para $L(y) \geq \eta_{\mathsf{B}}$.

Critério de Bayes - cont.

- Para aplicar o teste de Bayes, é necessário o conhecimento dos custos de decisão: C_{00}, C_{01}, C_{10} e C_{11}.
- No caso, se $(C_{10} - C_{00}) = (C_{01} - C_{11})$, temos que o limiar de Bayes é igual ao do critério MAP.
- Logo, o critério MAP é um caso especial do critério de Bayes.
- A regra de decisão de Bayes será então

$$\delta_{\mathsf{B}}(y) = \begin{cases} 1, & \forall y \in L(y) \geq \eta_{\mathsf{B}} \\ 0, & \forall y \in L(y) < \eta_{\mathsf{B}} \end{cases} . \tag{209}$$

Critério MiniMax

Até o momento foi discutido o desenvolvimento de critérios que necessitam de:

(a) probabilidades a priori, somente - MAP;

(b) probabilidades a priori e custos de decisão - Bayes.

O critério MiniMax é um critério que utiliza como requisito **somente o conhecimento dos custos de decisão**.

Idéia

Escolher (assumir) um valor a para a probabilidade a priori π_0. Com isso, o limiar de Bayes ficaria

$$\eta_{\mathsf{B}} = \frac{a(C_{10} - C_{00})}{(1-a)(C_{01} - C_{11})} \tag{210}$$

1. Se tomarmos a Equação (200), e tivermos $\pi_0 = 0$, então H_1 é correta.
2. $\mathcal{R} = C_{11}$, pois $\pi_0 = 0 \Rightarrow \pi_1 = 1 \Rightarrow P_{01} = 0 \Rightarrow P_{11} = 1$.
3. Se $\pi_0 = 1 \Rightarrow \pi_1 = 0 \Rightarrow P_{10} = 0 \Rightarrow P_{00}$ e $\mathcal{R} = C_{00}$.

Critério MiniMax - cont.

É conhecido que $\mathcal{R}$ é uma função côncava em π_0. Daí buscamos os valores de $\mathcal{R}$ para π_0 entre 0 e 1.

Podemos ainda escrever o custo em função de a, como:

$$\mathcal{R}(\pi_0, a) = \pi_0 \cdot \mathcal{R}_0(a) + (1 - \pi_0) \cdot \mathcal{R}_1(a), \tag{211}$$

em que $\mathcal{R}_0(a)$ e $\mathcal{R}_1(a)$ são os riscos condicionais para cada uma das hipóteses com limiar e regra de decisão controlados por a.

Se a escolha de a for correta $\mathcal{R}(\pi_0, a) = \mathcal{R}$ (de Bayes). Senão, teremos para os casos $\pi_0 = 1 \Rightarrow \mathcal{R}(\pi_0, a) = \mathcal{R}_0(a)$ e $\pi_1 = 1 \Rightarrow \mathcal{R}(\pi_0, a) = \mathcal{R}_1(a)$.

Assim, o critério minimax **minimiza o máximo risco**.

Se a função $\mathcal{R}(\pi_0, a)$ é decrescente ou crescente para valores de π_0 entre 0 e 1.

A solução minimax corresponde a encontrar o máximo entre C_{00} e C_{11}, que são os valores obtidos quando $\pi_0 = 0$ e $\pi_0 = 1$, respectivamente.

$\mathcal{R}(\pi_0, a)$ é uma função diferenciável - solução entre $\pi_0 = 0$ e $\pi_0 = 1$.

Busca-se o mínimo que é obtido derivando a Equação (211) em relação à π_0. Assim, pode-se ver que a solução minimax corresponde ao ponto

$$\mathcal{R}_0(a) = \mathcal{R}_1(a) \tag{212}$$

Com isso, deve-se buscar o valor de a que fornece as Equações (210) e (212).

$\mathcal{R}(\pi_0, a)$ é uma função não-diferenciável - solução entre $\pi_0 = 0$ e $\pi_0 = 1$.

Quando as probabilidades condicionais $p(y|H_0)$ e $p(y|H_1)$ são discretas ou híbridas, não é possível diferenciar a verossimilhança em todos os pontos.

Neste caso, consideramos duas regras de decisão. Inicialmente façamos π_0 se aproximar de a como

$$a^- = \lim_{\epsilon \to 0} a - \epsilon \tag{213}$$

e temos sua região de decisão associada para H_1 como

$$R_1^- = \forall y \ni L(y) > \frac{a^-(C_{10} - C_{00})}{(1 - a^-)(C_{01} - C_{11})} \tag{214}$$

A segunda regra de decisão é então feita para aproximar π_0 de a como

$$a^+ = \lim_{\epsilon \to 0} a + \epsilon. \tag{215}$$

E a região de decisão associada é dada por

$$R_1^+ = \forall y \ni L(y) > \frac{a^+(C_{10} - C_{00})}{(1-a^+)(C_{01} - C_{11})} \tag{216}$$

E então a região crítica ocorre quando $L(y)$ é igual ao limiar. A regra de decisão pode então ser escrita como

$$\delta_{\text{MM}}(y) = \begin{cases} 1, & L(y) > \eta_{\text{MM}} \\ \alpha, & L(y) = \eta_{\text{MM}} \\ 0, & L(y) < \eta_{\text{MM}} \end{cases}, \tag{217}$$

Se $L(y) > \eta_{\text{MM}}$ decide D_1, $L(y) < \eta_{\text{MM}}$ decide D_0 e se $L(y) = \eta_{\text{MM}}$ decide D_1 com probabilidade α.

Deseja-se então escolher a variável α tal que o risco condicional associado com a decisão D_1 e a regra $\delta_{\text{MM}}(y)$ seja o mesmo que o risco condicional associado com a decisão D_0 e $\delta_{\text{MM}}(y)$.

Para isso, denota-se estes riscos condicionais por ${}_\delta\mathcal{R}_j, j = 0, 1$. Assume-se que a região R_1^- ocorre com probabilidade α e R_1^+ ocorre com probabilidade $(1-\alpha)$. Com isso os riscos condicionais são

$$ {}_\delta\mathcal{R}_j = \alpha \cdot \mathcal{R}_j(a^-) + (1-\alpha)\mathcal{R}_j(a^+), \qquad j = 0, 1. \tag{218} $$

E a igualdade entre os riscos condicionais é obtida se

$$ \alpha = \frac{\mathcal{R}_0(a^+) - \mathcal{R}_1(a^+)}{\mathcal{R}_0(a^+) - \mathcal{R}_1(a^+) + \mathcal{R}_1(a^-) - \mathcal{R}_0(a^-)} \tag{219} $$

Critério Neyman-Pearson

Este tipo de critério não requer nem o conhecimento das probabilidades a priori nem dos custos de decisão para sua formulação.

Meta

Fixar a probabilidade de falso alarme e maximizar a probabilidade de detecção

Ou seja, fazer uma otimização com restrições do tipo:

$$\begin{cases} \text{maximizar } P_D \\ \text{sujeito a } P_F \leq \beta_F \end{cases} \tag{220}$$

A probabilidade de falso alarme é definida como

$$P_F = \Pr(D_1|H_0) = P_{10} \tag{221}$$

Ao considerarmos que a variável y tem distribuição contínua sob ambas as hipóteses, uma derivação simples seria

$$P_F = \beta \leq \beta_F. \tag{222}$$

Critério Neyman-Pearson - cont.

Usando multiplicadores de Lagrange, define-se a função a ser otimizada J_{NP} como:

$$\begin{aligned} J_{\text{NP}} &= P_D - \lambda[P_F - \beta] \\ &= \Pr(D_1|H_1) - \lambda[\Pr(D_1|H_0) - \beta] \end{aligned} \tag{223}$$

Critério Neyman-Pearson - cont.

E maximizar P_D também maximiza J_{NP}. Daí temos que:

$$\begin{aligned} J_{\text{NP}} &= \int_{R_1} p(y|H_1)\ dy - \lambda \int_{R_1} p(y|H_0)\ dy + \lambda\beta \\ &= \int_{R_1} \left[p(y|H_1) - \lambda p(y|H_0)\right]\ dy + \lambda\beta \end{aligned} \tag{224}$$

Maximizando J_{NP} para $\lambda > 0$, nós desejamos escolher R_1 tal que o integrando seja sempre positivo, ou seja

$$p(y|H_1) - \lambda p(y|H_0) > 0 \tag{225}$$

decide por D_1 (hipótese H_1). Daí podemos construir um teste da razão de verossimilhança. Logo

$$L(y) = \frac{p(y|H_1)}{p(y|H_0)} \underset{D_0}{\overset{D_1}{\gtrless}} \lambda. \tag{226}$$

Por definição temos

$$P_F = \Pr(D_1|H_0) = \int_{R_1(\lambda)} p(y|H_0)\, dy = \beta \tag{227}$$

e também

$$P_D = \Pr(D_1|H_1) = \int_{R_1(\lambda)} p(y|H_1)\, dy. \tag{228}$$

Como ambas as integrais estão no mesmo espaço $R_1(\lambda)$ então aumentar λ para aumentar (maximizar) P_D também implica em aumentar P_F. Deste modo, modifica-se o valor de λ até encontrar o maior valor tal que $\beta \leq \beta_F$.

Este valor de P_F irá corresponder ao maior valor de P_D, respeitando a restrição e o limiar (λ) será denominado de η_{NP}. A regra de decisão então será um pouco mais complicada, dada por,

$$\delta_{\text{NP}}(y) = \begin{cases} 1, & L(y) > \eta_{\text{NP}} \\ \alpha, & L(y) = \eta_{\text{NP}} \\ 0, & L(y) < \eta_{\text{NP}} \end{cases}, \tag{229}$$

D_1 é decidido com probabilidade α para $L(y) = \eta_{\text{NP}}$. Assim, η_{NP} e α são escolhidos tal que $P_F = \beta$.

Critério	**Requisitos**	
	Prob. a priori	*Custos*
MAP	Sim	Não
Bayes	Sim	Sim
Minimax	Não	Sim
Neyman-Pearson	Não	Não

Tabela: Resumo dos requisitos de cada um dos critérios.

Parte VII

Algoritmos Recursivos no Tempo

- Estruturas adaptativas: atualização dos parâmetros para se adequar às características dos sinais de interesse
- Regras de atualização: **algoritmos de recursão temporal**
- Escolha
  1. **Critério:** o que maximizar/minimizar?
  2. **Método de busca:** como minimizar?
  3. **Complexidade:** quanto se pode "pagar" pelo desempenho?
  4. **Velocidade de convergência:** qual o tempo desejado/disponível?

Introdução - cont.

- Estruturas de filtragem: FIR $\times$ IIR
- Escolha afeta complexidade e número de interações para atingir desempenho desejado
- **FIR**
  - Estabilidade garantida
  - Critério unimodal
  - Condições de estabilidade para algoritmo de atualização mais fácil
  - Problemas reais: maior complexidade de modelagem
- **IIR**
  - Requer verificação de estabilidade
  - Critério multimodal
  - Possibilidade de garantir estabilidade do algoritmo de adaptação
  - Menor complexidade para modelar problemas reais

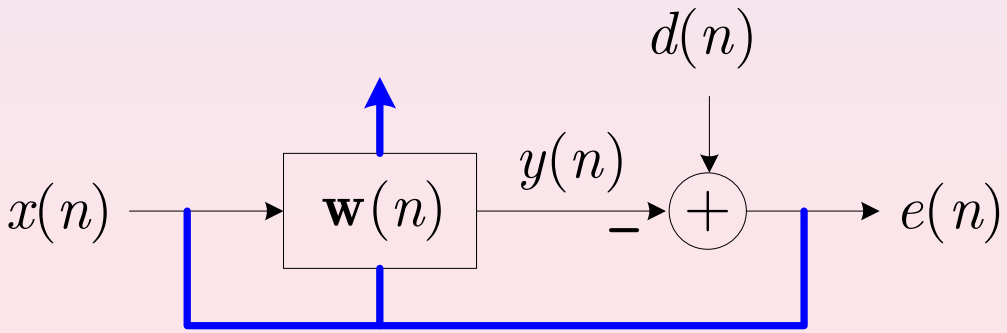
Filtragem ótima (Wiener)

1. Aquisição dos dados: obtenção de $\mathbf{R_x}$ e $\mathbf{p}_{xd}$
2. Determinação dos filtro ótimo: $\mathbf{w}_{\text{opt}} = \mathbf{R}_{\mathbf{x}}^{-1}\mathbf{p}_{xd}$
3. Complexidade computacional $\sim M^3$

Filtragem adaptativa

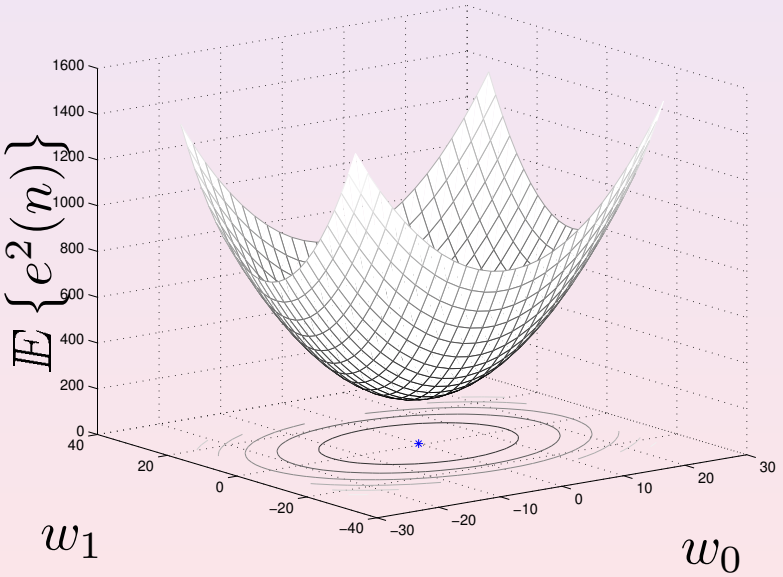
- Adquirir os dados e otimizar o sistema adaptativo ao mesmo tempo (complexidade computacional menor!)

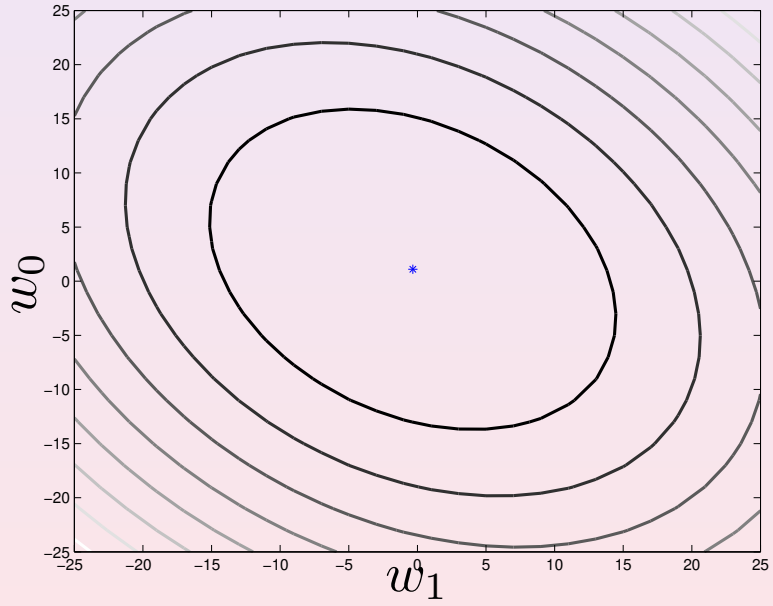
Idéia geral



Critério: atualizar $\mathbf{w}(n)$ de forma a minimizar $\mathbb{E}\left\{e^2(n)\right\}$.

Para o caso de $\mathbf{w}(n)$ com dois coeficientes $\mathbf{w}(n) = [\ w_0\ w_1\ ]^T$, temos





Métodos de busca - cont.

Otimização interativa: a partir de uma condição inicial $\mathbf{w}(0)$ chegar a $\mathbf{w}_{\text{opt}}$ para $0 < n \leq N$ iterações

Métodos de busca: baseados nos métodos clássicos de otimização de 1a (gradiente) e 2a (Newton) ordens.

Dada a função $J(\mathbf{w}) = \mathbb{E}\left\{e^2(n)\right\}$ deseja-se que

$$\mathbf{w}(n) \to \mathbf{w}(n+1) \Rightarrow J_{n+1} < J_n$$

Considerando a função $J(\mathbf{w})$ expandida em série de Taylor em torno do ponto $\mathbf{w}(n)$ tem-se

$$J(\mathbf{w})|_{\mathbf{w}(n+1)} = J(\mathbf{w})|_{\mathbf{w}(n)} + \left.\frac{\partial J}{\partial \mathbf{w}^T}\right|_{\mathbf{w}(n)} \Delta\mathbf{w}(n+1) + \frac{1}{2}\Delta\mathbf{w}^T(n+1) \left.\frac{\partial^2 J}{\partial \mathbf{w}\partial \mathbf{w}^T}\right|_{\mathbf{w}(n)} \Delta\mathbf{w}(n+1) \tag{230}$$

em que $\Delta\mathbf{w}(n+1) = \mathbf{w}(n+1) - \mathbf{w}(n)$

Dois algoritmos importantes

1. baseado na expansão de 1a ordem
2. baseado na expansão de 2a ordem

Meta: Gerar $\Delta \mathbf{w}(n+1)$ tal que $J(\mathbf{w})|_{\mathbf{w}(n+1)} < J(\mathbf{w})|_{\mathbf{w}(n)}$

1a ordem: $\frac{\partial J}{\partial \mathbf{w}^T}\Big|_{\mathbf{w}(n)} \Delta\mathbf{w}(n+1)$

Algoritmo *steepest descent* ("descida mais íngreme")

$$\mathbb{E}\left\{e^2(n)\right\} = \sigma_d^2 - 2\mathbf{w}^T\mathbf{p}_{xd} + \mathbf{w}^T\mathbf{R}_\mathbf{x}\mathbf{w}$$

$$\frac{\partial \mathbb{E}\left\{e^2(n)\right\}}{\partial \mathbf{w}} = -2\mathbf{p}_{xd} + 2\mathbf{R}_\mathbf{x}\mathbf{w}$$

Algoritmo *steepest descent* (gradiente determinístico)

$$\mathbf{w}(n+1) = \mathbf{w}(n) - 2\mu\left[\mathbf{R}_\mathbf{x}\mathbf{w}(n) - \mathbf{p}_{xd}\right] \tag{231}$$

Notar que ainda se faz necessário conhecer $\mathbf{R}_\mathbf{x}$ e $\mathbf{p}_{xd}$!

2a ordem: Método de Newton

Temos que

$$\begin{aligned}\mathbf{H}(\mathbf{w}) &= \frac{\partial^2 J}{\partial \mathbf{w} \partial \mathbf{w}^T} \\ &= 2\mathbf{R}_{\mathbf{x}}\end{aligned} \tag{232}$$

é a matriz Hessiana de $J(\mathbf{w})$.

$\mathbf{H}(\mathbf{w})$ é uma matriz definida positiva (autocorrelação), logo a aproximação quadrática tem um único e bem definido ponto de mínimo

Meta: obter $\Delta\mathbf{w}(n+1)$ tal que $\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = 0$.

Assim, temos na regra de Newton que

$$\begin{aligned}\mathbf{w}(n+1) &= \mathbf{w}(n) + \mu \mathbf{H}^{-1}(\mathbf{w}) \frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} \\ &= \mathbf{w}(n) - 2\mu \mathbf{H}^{-1}(\mathbf{w}) \left[\mathbf{R}_{\mathbf{x}} \mathbf{w}(n) - \mathbf{p}_{xd}\right] \\ \mathbf{w}(n+1) &= \mathbf{w}(n) - \mu \left[\mathbf{w}(n) - \underbrace{\mathbf{R}_{\mathbf{x}}^{-1} \mathbf{p}_{xd}}_{\mathbf{w}_{\text{opt}}}\right]\end{aligned} \tag{233}$$

Algoritmo de Newton

$$\mathbf{w}(n+1) = \mathbf{w}(n) - \mu \left[\mathbf{w}(n) - \mathbf{R}_{\mathbf{x}}^{-1} \mathbf{p}_{xd}\right] \tag{234}$$

Ainda no algoritmo de Newton, suponha

- $\mu = 1$
- $\mathbf{w}(0) = \mathbf{0}$

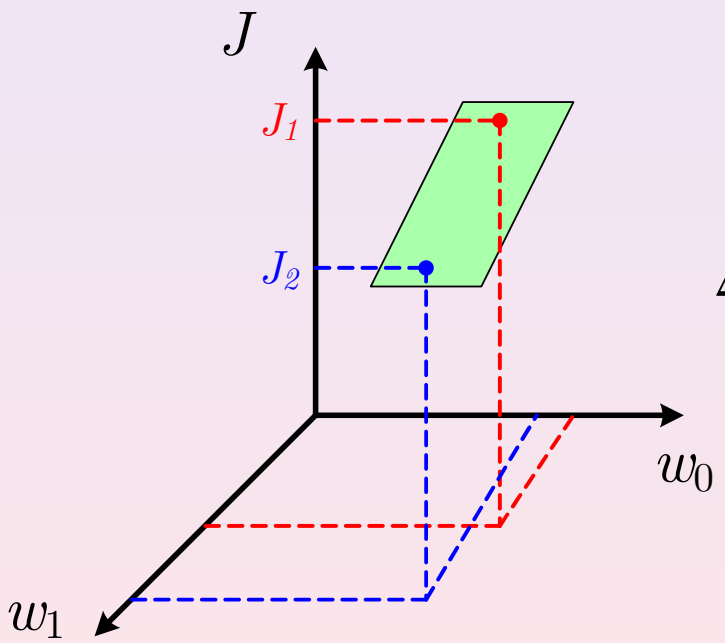
Na primeira iteração temos

$$\mathbf{w}(1) = \mathbf{R}_{\mathbf{x}}^{-1}\mathbf{p}_{xd}$$

⚠ **Solução ótima em uma iteração!**

Características do *Steepest descent*

- O método de 1a ordem aproxima $J(\mathbf{w})$ como uma função linear e "caminha" nessa função com o maior "passo" (maior declividade) possível



$$\Delta \mathbf{w}_{i+1} = -\mu \nabla_{\mathbf{w}} J(\mathbf{w})$$

Características do *Steepest descent* - cont.

Método linear $\rightarrow$ algoritmo do gradiente determinístico

$$\mathbf{w}(n+1) = \mathbf{w}(n) - \mu \nabla_{\mathbf{w}} J(\mathbf{w})$$
$$\mathbf{w}(n+1) = \mathbf{w}(n) - 2\mu \left[\mathbf{R}_{\mathbf{x}} \mathbf{w}(n) - \mathbf{p}_{xd}\right]$$

Características do algoritmo de Newton

- O método de 2a ordem "aproxima" $J(\mathbf{w})$ como uma função quadrática e procura o mínimo desta função
- Encontrar um $\Delta\mathbf{w}$ que nos leve ao mínimo, ou seja, a uma condição de gradiente nulo, no passo $i+1$.
- Obter $\Delta\mathbf{w}_{i+1}$ tal que $\left.\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}}\right|_{\mathbf{w}_{i+1}} = 0$

$$\nabla_{\mathbf{w}} J(\mathbf{w}) + \mathbf{H}(\mathbf{w})\Delta\mathbf{w} = 0$$
$$\Delta\mathbf{w} = -\mathbf{H}^{-1}(\mathbf{w}) \cdot \nabla_{\mathbf{w}} J(\mathbf{w})$$

Como $\mathbf{H}(\mathbf{w}) = 2\mathbf{R}_{\mathbf{x}}$ tem-se

Características do algoritmo de Newton - cont.

$$\mathbf{w}_{i+1} = \mathbf{w}_i - \frac{1}{2}\mathbf{R}_{\mathbf{x}}^{-1} \cdot \nabla_{\mathbf{w}} J(\mathbf{w})$$

$$\mathbf{w}_{i+1} = \mathbf{w}_i + \mathbf{R}_{\mathbf{x}}^{-1}\mathbf{p}_{xd} - \mathbf{w}_i$$

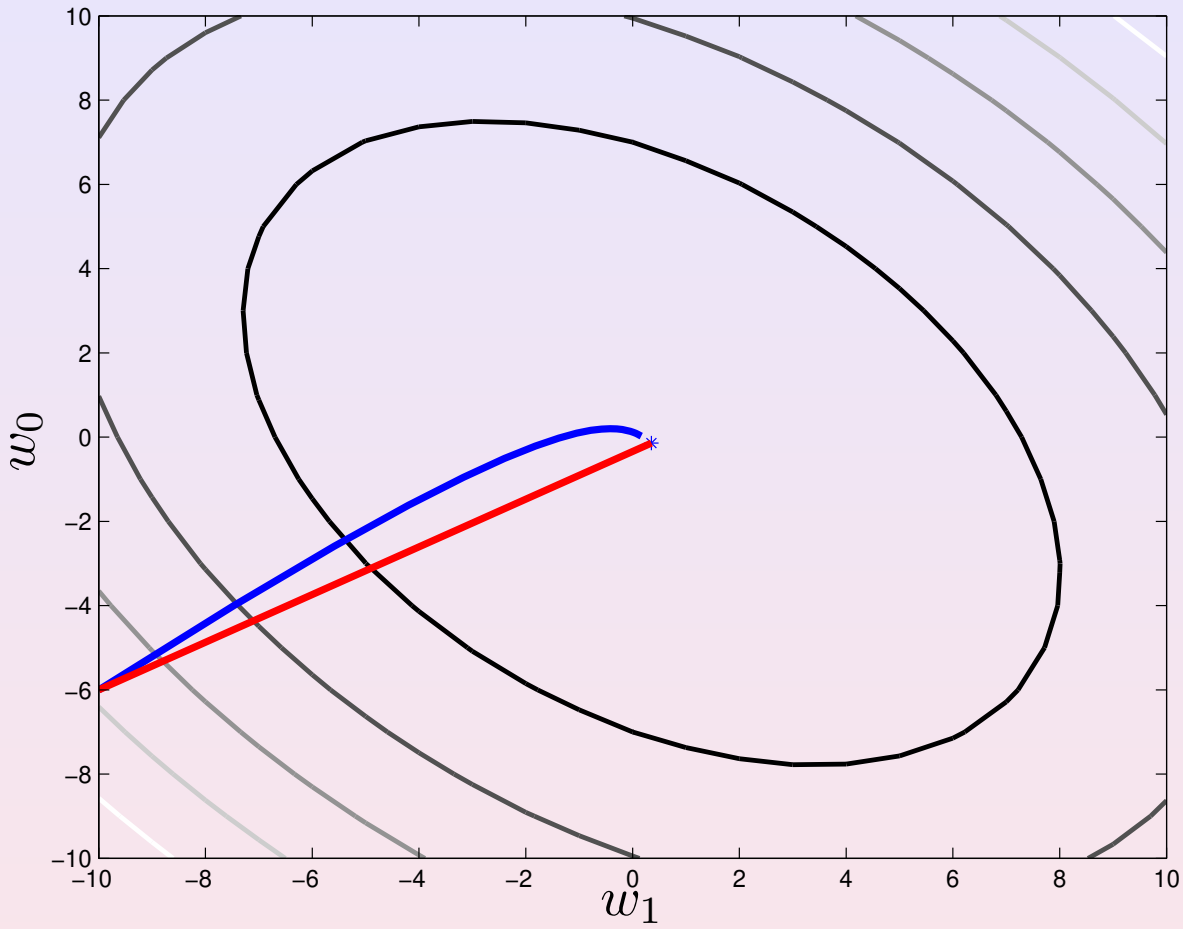
$$\boxed{\mathbf{w}_{i+1} = \mathbf{R}_{\mathbf{x}}^{-1}\mathbf{p}_{xd}}$$

que é a própria solução ótima.

Steepest descent $\times$ Algoritmo de Newton

- O algoritmo *steepest decent* busca encontrar, à cada iteração, em qual direção a função decresce mais rapidamente (gradiente descendente)
- O método de Newton, calcula para a função, qual a direção, a partir do ponto inicial, que chega mais rapidamente ao ponto ótimo.
- Método de Newton é mais complexo (inversão de matriz) e mais rápido. Para $J(\mathbf{w})$ quadrático, uma iteração é suficiente se $\mu = 1$.
- *Steepest descent* é mais simples, mas tem uma latência maior para convergir ao ponto ótimo.

Steepest decent × Newton - cont.



Convergência dos algoritmos *steepest descent* (azul) e de Newton (vermelho). Passos $\mu_{sd} = 0.1$ e $\mu_n = 1$.

Problemas

- Equações necessitam conhecimento ou estimativa das estatísticas $\mathbf{R_x}$ e $\mathbf{p}_{xd}$
- Processamento caro e não garante uma convergência ao valor desejado
- **Alternativa**: aproximações estocásticas

Origem

H. Robbins and S. Monro, "A Stochastic Approximation Method", *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400-407, 1951

Idéia: estimação recursiva de um determinado número de parâmetros θ, de forma:

$$\theta(n) = \theta(n-1) - \mu(n) \cdot f\left[\theta(n-1), x(n)\right] \tag{235}$$

em que
$x(n) =$ dados observados no tempo
$\mu(n) =$ seqüência decrescente
$f(\cdot) =$ função dos dados e parâmetros

Exemplo

Sejam

$$\theta(0) = 0$$

$$\mu = \frac{1}{n}$$

$$f[\theta(n-1), x(n)] = \theta(n-1) - x(n)$$

daí decorre que

$$\theta(n) = \frac{x(1) + x(2) + \ldots + x(n)}{n}$$

1a observação: O algoritmo de Robbins-Monro converge para $f[\theta(n-1), x(n)] = 0$.

Supondo várias realizações do algoritmo
θ_{opt} é tal que $\mathbb{E}\{f[\theta(n-1), x(n)]\} = 0$

No nosso caso (filtragem): quem é $\mathbb{E}\{f[\theta(n-1), x(n)]\}$?

Sabemos que $\nabla_{\mathbf{w}} \mathbb{E}\{f[\theta(n-1), x(n)]\} = 0$ para $\mathbf{w} = \mathbf{w}_{\text{opt}}$ então

$$f[\theta(n-1), x(n)] = \frac{\partial e^2(n)}{\partial \mathbf{w}} \tag{236}$$

Aproximação estocástica - cont.

Partindo da aproximação da Eq. (236), a recursão temporal para aproximar o critério de minimizar o erro quadrático médio seria do tipo:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + 2\mu(n+1)\mathbf{x}(n)e(n) \tag{237}$$

em que $e(n) = d(n) - \mathbf{w}^T(n)\mathbf{x}(n)$.

Se considerarmos $\mu(n+1) = \mu$ teremos então o **algoritmo do gradiente estocástico** dado por

$$\boxed{\mathbf{w}(n+1) = \mathbf{w}(n) + 2\mu\mathbf{x}(n)e(n)}$$

Gradiente determinístico: Busca na direção negativa do gradiente

$$\mathbf{w}(n+1) = \mathbf{w}(n) - 2\mu\left[\mathbf{R}_{\mathbf{x}}\mathbf{w}(n) - \mathbf{p}_{xd}\right]$$

Algoritmo de Newton: Mais rápido e mais complexo

$$\mathbf{w}(n+1) = \mathbf{w}(n) - \mu\left[\mathbf{w}(n) - \mathbf{R}_{\mathbf{x}}^{-1}\mathbf{p}_{xd}\right]$$

Gradiente estocástico: Mais simples, menos requisitos, desempenho pior

$$\mathbf{w}(n+1) = \mathbf{w}(n) + 2\mu\mathbf{x}(n)e(n)$$

Algoritmos estocásticos

- Diferentes aproximações podem ser realizadas
- Meta é reduzir a complexidade dos algoritmos provendo uma convergência para o ponto ótimo
- Algumas técnicas são discutidas a seguir

O algoritmo LMS (*Least Mean Square*) é um algoritmo de busca que utiliza uma simplificação do vetor gradiente por meio de uma modificação na função custo (objetivo)

Propriedades

- Simplicidade computacional
- Prova de convergência em ambiente estacionário
- Convergência não-polarizada, em média, para a solução ótima (Wiener)

Se tomarmos o gradiente estocástico temos então

$$\mathbf{w}(n+1) = \mathbf{w}(n) - 2\mu \left[\mathbf{R}_{\mathbf{x}}\mathbf{w}(n) - \mathbf{p}_{xd}\right]$$

mas, deseja-se trabalhar com estimativas das estatísticas no instante n, uma vez que as mesmas podem não estar disponíveis completamente, então teremos algo como

$$\mathbf{w}(n+1) = \mathbf{w}(n) - 2\mu \left[\widehat{\mathbf{R}}_{\mathbf{x}}(n)\mathbf{w}(n) - \widehat{\mathbf{p}}_{xd}(n)\right] \tag{238}$$

Então, uma solução possível é fazer uma aproximação das estatísticas por seus valores instantâneos, ou seja

$$\begin{aligned} \mathbf{R}_{\mathbf{x}} = \mathbb{E}\left\{\mathbf{x}(n)\mathbf{x}^T(n)\right\} &\approx \widehat{\mathbf{R}}_{\mathbf{x}}(n) = \mathbf{x}(n)\mathbf{x}^T(n) \\ \mathbf{p}_{xd} = \mathbb{E}\left\{\mathbf{x}(n)d(n)\right\} &\approx \widehat{\mathbf{p}}_{xd}(n) = \mathbf{x}(n)d(n) \end{aligned} \tag{239}$$

Desta forma, teremos

$$\begin{aligned}\mathbf{w}(n+1) &= \mathbf{w}(n) - 2\mu\left[\widehat{\mathbf{R}}_{\mathbf{x}}(n) - \widehat{\mathbf{p}}_{xd}(n)\right] \\ &= \mathbf{w}(n) - 2\mu\left[\mathbf{x}(n)\mathbf{x}^T(n)\mathbf{w}(n) - \mathbf{x}(n)d(n)\right] \\ &= \mathbf{w}(n) - 2\mu\mathbf{x}(n)\left[\mathbf{x}^T(n)\mathbf{w}(n) - d(n)\right] \\ &= \mathbf{w}(n) - 2\mu\mathbf{x}(n)\left[y(n) - d(n)\right]\end{aligned} \tag{240}$$

Então, a equação de recursão do LMS é dada por:

Algoritmo LMS

$$\boxed{\mathbf{w}(n+1) = \mathbf{w}(n) + 2\mu\mathbf{x}(n)e(n)} \tag{241}$$

Note que o algoritmo LMS possui a mesma regra de recursão que a aproximação do gradiente estocástico, por isto é comum usar a mesma notação para ambos.

Uma questão importante reside na garantia da convergência do algoritmo para os parâmetros ótimos. Bem como observar se esta convergência é não-polarizada.

Gradiente: o gradiente do algoritmo converge para algum valor?

Tomando as expressões do gradiente para o algoritmo determinístico e do LMS

$$\begin{aligned} \nabla_{\text{det}} &= 2\left[\mathbf{R}_{\mathbf{x}}\mathbf{w}(n) - \mathbf{p}_{xd}\right] \\ \nabla_{\text{LMS}} &= 2\left[\mathbf{x}(n)\mathbf{x}^T(n)\mathbf{w}(n) - \mathbf{x}(n)d(n)\right] \end{aligned} \tag{242}$$

podemos ver que as direções determinadas por ambos os algoritmos são diferentes (como esperado). Entretanto, se tomarmos o valor médio no caso do LMS temos

$$\begin{aligned} \mathbb{E}\left\{\nabla_{\text{LMS}}\right\} &= \mathbb{E}\left\{2\left[\mathbf{x}(n)\mathbf{x}^T(n)\mathbf{w}(n) - \mathbf{x}(n)d(n)\right]\right\} \\ &= 2\left[\mathbb{E}\left\{\mathbf{x}(n)\mathbf{x}^T(n)\right\}\mathbf{w}(n) - \mathbb{E}\left\{\mathbf{x}(n)d(n)\right\}\right] \\ &= 2\left[\mathbf{R}_{\mathbf{x}}\mathbf{w}(n) - \mathbf{p}_{xd}\right] = \nabla_{\text{det}} \end{aligned} \tag{243}$$

Estabilidade: quais os valores de μ para os quais o algoritmo converge?

Vamos considerar uma perturbação do vetor de coeficientes em torno do filtro ótimo, assim temos

$$\Delta\mathbf{w}(n) = \mathbf{w}(n) - \mathbf{w}_{\text{opt}} \tag{244}$$

Utilizando esta definição, podemos escrever o LMS como

$$\begin{aligned}
\Delta\mathbf{w}(n+1) &= \Delta\mathbf{w}(n) + 2\mu e(n)\mathbf{x}(n) \\
&= \Delta\mathbf{w}(n) + 2\mu\mathbf{x}(n)\left[\mathbf{x}(n)^T\mathbf{w}_{\text{opt}} + b(n) - \mathbf{x}(n)^T\mathbf{w}(n)\right] \\
&= \Delta\mathbf{w}(n) + 2\mu\mathbf{x}(n)\left[e_{\text{opt}}(n) - \mathbf{x}(n)^T\mathbf{w}(n)\right] \\
&= \left[\mathbf{I} - 2\mu\mathbf{x}(n)\mathbf{x}(n)^T\right]\Delta\mathbf{w}(n) + 2\mu e_{\text{opt}}(n)\mathbf{x}(n)
\end{aligned} \tag{245}$$

Sabendo que $e_{\text{opt}}(n) = d(n) - \mathbf{x}(n)^T \mathbf{w}_{\text{opt}} = b(n)$ temos então, o valor esperado de

$$\begin{aligned}\mathbb{E}\left\{\Delta\mathbf{w}(n+1)\right\} = \mathbb{E}\left\{\left[\mathbf{I} - 2\mu\mathbf{x}(n)\mathbf{x}(n)^T\right]\Delta\mathbf{w}(n)\right\} \\ + 2\mu\mathbb{E}\left\{e_{\text{opt}}(n)\mathbf{x}(n)\right\}\end{aligned} \tag{246}$$

Assumindo independência entre $\mathbf{x}(n)$, $\Delta\mathbf{w}(n)$ e $e_{\text{opt}}(n)$, temos então

$$\begin{aligned}\mathbb{E}\left\{\Delta\mathbf{w}(n+1)\right\} &= \left[\mathbf{I} - 2\mu\mathbb{E}\left\{\mathbf{x}(n)\mathbf{x}(n)^T\right\}\right]\mathbb{E}\left\{\Delta\mathbf{w}(n)\right\} \\ &= (\mathbf{I} - 2\mu\mathbf{R}_{\mathbf{x}})\mathbb{E}\left\{\Delta\mathbf{w}(n)\right\}\end{aligned} \tag{247}$$

Um fator que nos ajuda é saber que podemos decompor a matriz $\mathbf{R}_{\mathbf{x}}$ como

$$\mathbf{R}_{\mathbf{x}} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T \tag{248}$$

em que $\mathbf{Q}$ é a matriz (ortogonal) dos autovetores de $\mathbf{R}_{\mathbf{x}}$

Pré-multiplicando então a Eq. (247) por $\mathbf{Q}^T$ temos

$$\mathbb{E}\left\{\mathbf{Q}^T\Delta\mathbf{w}(n+1)\right\} = (\mathbf{I} - 2\mu\mathbf{Q}^T\mathbf{R_x}\mathbf{Q})\mathbb{E}\left\{\mathbf{Q}^T\Delta\mathbf{w}(n)\right\} \quad (249)$$

Mas sabe-se ainda que

$$\begin{aligned}\mathbf{R_x} &= \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T \\ \mathbf{Q}^T\mathbf{R_x} &= \mathbf{Q}^T\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T \\ &= \mathbf{\Lambda}\mathbf{Q}^T\end{aligned}$$

E podemos então definir

$$\mathbf{v}(n+1) = \mathbb{E}\left\{\mathbf{Q}^T\Delta\mathbf{w}(n+1)\right\} \quad (250)$$

que são versões rotacionadas dos erros dos coeficientes.

Daí, temos então

$$\begin{aligned}\mathbf{v}(n+1) &= \mathbf{v}(n) - 2\mu\mathbf{\Lambda}\mathbf{v}(n) \\ &= (I - 2\mu\mathbf{\Lambda})\mathbf{V}(n)\end{aligned} \tag{251}$$

Ou seja, para cada elemento $v_i(n+1)$ do vetor $\mathbf{v}(n+1)$ temos

$$v_i(n+1) = (1 - 2\mu\lambda_k)v_i(n) \tag{252}$$

Condição de estabilidade:

$$\begin{aligned}|1 - 2\mu\lambda_k| < 1 \quad &\rightarrow \quad -1 < 1 - 2\mu\lambda_k < 1 \\ 0 < 2\mu\lambda_k < 2 \quad &\rightarrow \quad 0 < \mu < \frac{1}{\lambda_k}\end{aligned} \tag{253}$$

$$\boxed{\text{Estabilidade: } 0 < \mu < \frac{1}{\lambda_{\text{max}}}}$$

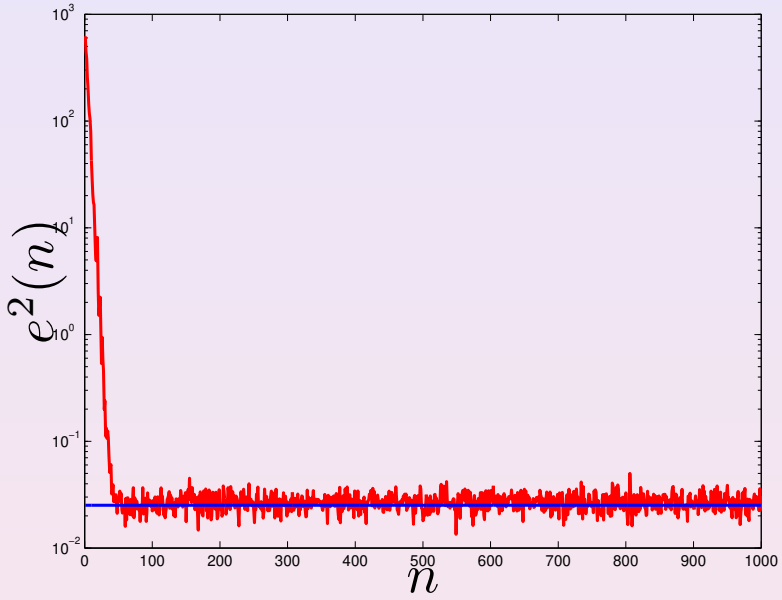
Misadjustment (desajuste): quanto a solução do LMS difere da solução ótima?

Tomando $\mathbf{w}(n) = \mathbf{w}_{\text{opt}}$, teremos

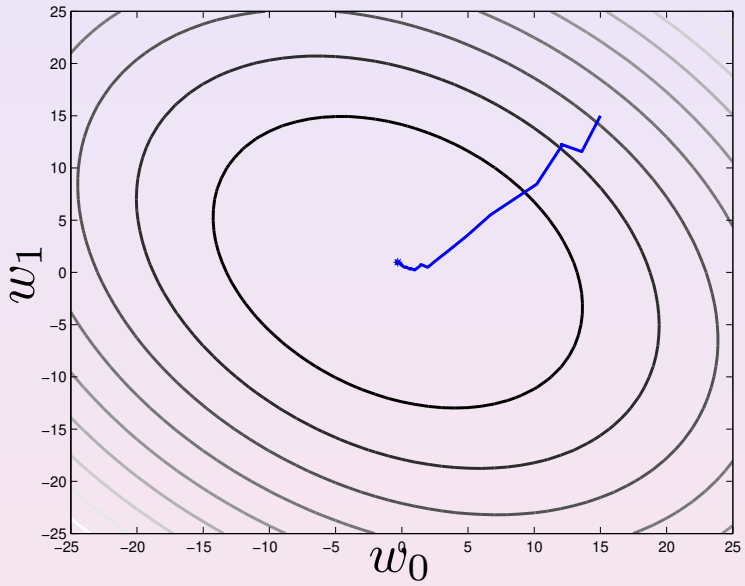
$$\begin{aligned}
\mathbf{w}(n+1) &= \mathbf{w}_{\text{opt}} + 2\mu\mathbf{x}(n)e(n) \\
&= \mathbf{w}_{\text{opt}} + 2\mu\mathbf{x}(n)[d(n) - \mathbf{x}^T(n)\mathbf{w}_{\text{opt}}] \\
&= \mathbf{w}_{\text{opt}} + 2\mu\mathbf{x}(n)[\mathbf{x}^T(n)\mathbf{w}_{\text{opt}} + b(n) - \mathbf{x}^T(n)\mathbf{w}_{\text{opt}}] \\
\mathbf{w}(n+1) - \mathbf{w}_{\text{opt}} &= 2\mu\mathbf{x}(n)b(n)
\end{aligned} \tag{254}$$

Desta forma, podemos ver que $\mathbb{E}\{\mathbf{w}(n+1) - \mathbf{w}_{\text{opt}}\} = 0$ mas que a variância não é zero devido ao termo $b(n)$.

- μ impacta na variância do erro de ajuste
- $\mathbf{w}(n+1)$ ao final da convergência fica "em torno" de $\mathbf{w}_{\text{opt}}$



Convergência do LMS (vermelho) para J_{min} (azul) usando $\mu = 0.1$



Trajetória do LMS (azul) para o ponto ótimo (solução de Wiener) usando $\mu = 0.1$

Resumo:

- Algoritmo com baixa complexidade
- Converge, em média, para o filtro ótimo
- Fator de passo influencia na velocidade de convergência
- Compromisso com o erro de desajuste

Motivação

- O algoritmo LMS apresenta o fator de passo dependente das características da correlação
- Para aumentar a velocidade de convergência, aumenta-se o fator de passo, mas o mesmo fornece um erro residual maior
- Idéia: colocar os dados para servirem de regulação ao desajuste

Sabendo que

$$\mathbf{w}(n+1) = \mathbf{w}(n) + 2\mu e(n)\mathbf{x}(n) = \mathbf{w}(n) + \Delta\widetilde{\mathbf{w}}(n) \tag{255}$$

temos então que

$$e^2(n) = d^2(n) + \mathbf{w}^T(n)\mathbf{x}(n)\mathbf{x}^T(n)\mathbf{w}(n) - 2d(n)\mathbf{w}^T(n)\mathbf{x}(n) \tag{256}$$

Se usarmos uma troca de $\widetilde{\mathbf{w}}(n) = \mathbf{w}(n) + \Delta\widetilde{\mathbf{w}}(n)$, teremos então:

$$\begin{aligned}\widetilde{e}^2(n) = e^2(n) &+ 2\Delta\widetilde{\mathbf{w}}^T(n)\mathbf{x}(n)\mathbf{x}^T(n)\mathbf{w}(n) \\ &+ \Delta\widetilde{\mathbf{w}}^T(n)\mathbf{x}(n)\mathbf{x}^T(n)\Delta\widetilde{\mathbf{w}}(n) - 2d(n)\Delta\widetilde{\mathbf{w}}^T(n)\mathbf{x}(n)\end{aligned} \tag{257}$$

Então, definindo

$$\begin{aligned}\Delta e^2(n) &= \widetilde{e}^2(n) - e^2(n) \\ &= -2\Delta\widetilde{\mathbf{w}}^T(n)\mathbf{x}(n)e(n) + \Delta\widetilde{\mathbf{w}}^T(n)\mathbf{x}(n)\mathbf{x}^T(n)\Delta\widetilde{\mathbf{w}}(n)\end{aligned} \tag{258}$$

Meta: tornar $\Delta e^2(n)$ negativo e mínimo pela escolha apropriada de μ

Substituindo $\Delta\widetilde{\mathbf{w}}(n) = 2\mu e(n)\mathbf{x}(n)$ na Eq. (258) tem-se

$$\Delta e^2(n) = -4\mu e^2(n)\mathbf{x}^T(n)\mathbf{x}(n) + 4\mu^2 e^2(n)[\mathbf{x}^T(n)\mathbf{x}(n)]^2 \tag{259}$$

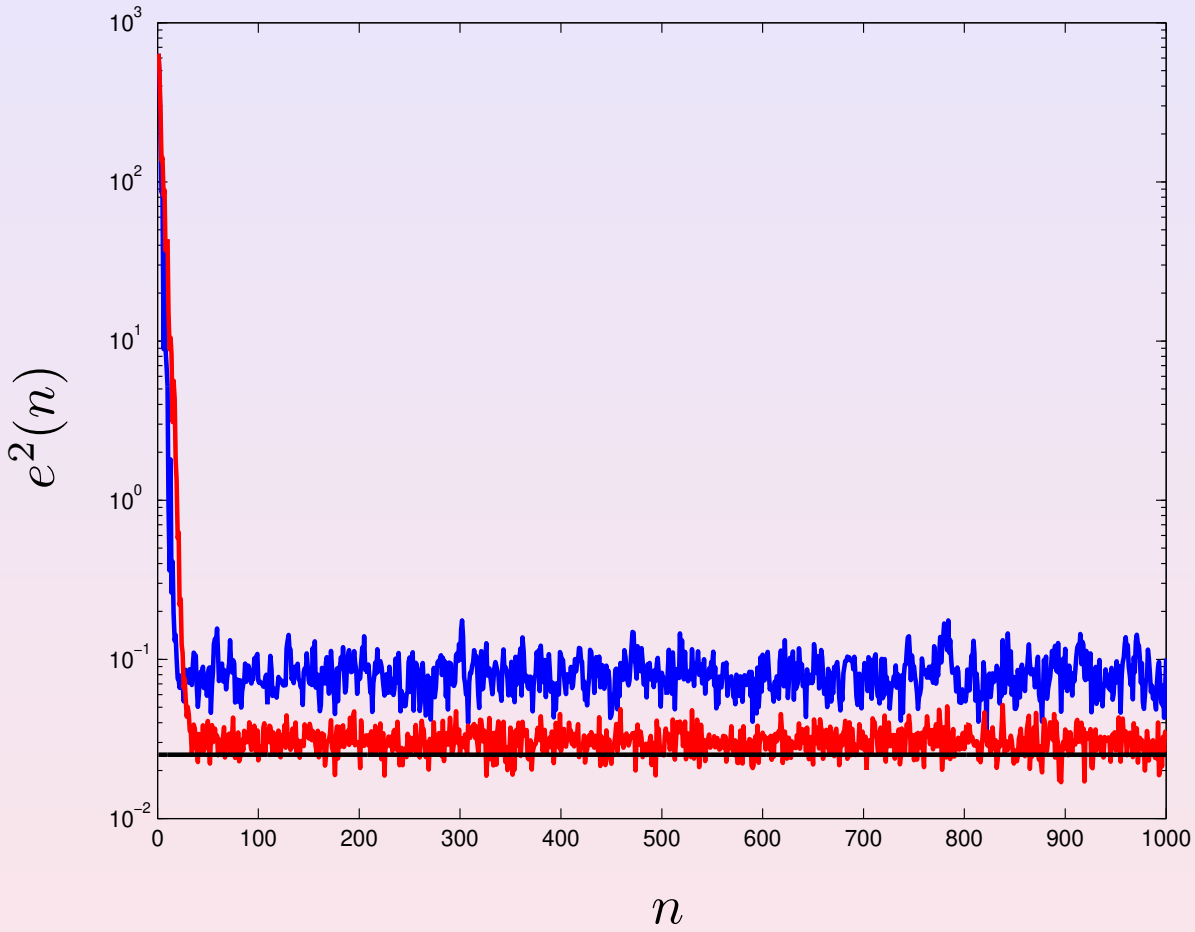
Valor de μ é dado por $\frac{\partial \Delta e^2(n)}{\partial \mu} = 0$, de onde tem-se

$$\mu = \frac{1}{2\mathbf{x}^T(n)\mathbf{x}(n)} \tag{260}$$

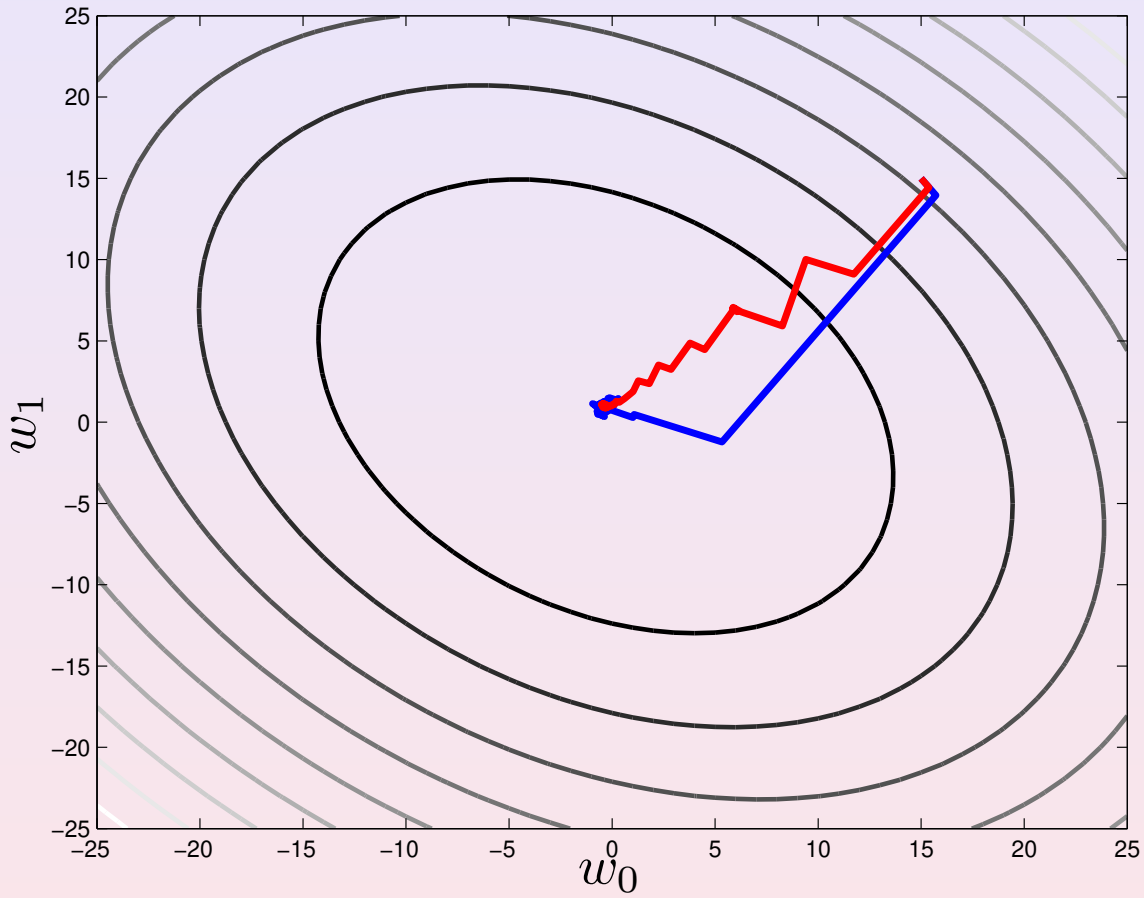
Com isso, o algoritmo do LMS normalizado é então dado por

Algoritmo LMS Normalizado

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \frac{\mu}{\gamma + \mathbf{x}^T(n)\mathbf{x}(n)}\mathbf{x}(n)e(n) \tag{261}$$



Algoritmos LMS (azul) e LMS-Normalizado (vermelho) com mesmo fator de passo $\mu_{\text{LMS}} = \mu_{\text{LMS-Norm}} = 0.5$, comparados com J_{min} (preto)



Trajetórias dos algoritmos LMS (azul) e LMS-Normalizado (vermelho) para o ponto ótimo