# Filtros Adaptativos do tipo LMS

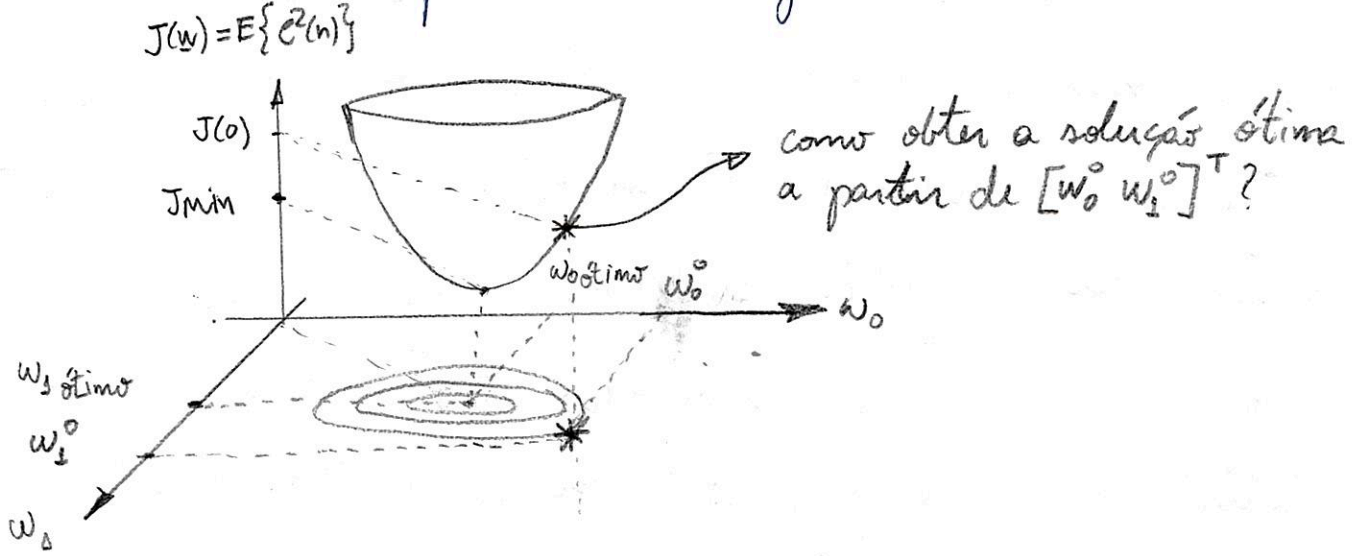
O uso do critério supervisionado MMSE em conjunto com uma estrutura linear caracteriza o paradigma de filtragem de Wiener. É importante ter em mente, porém, que o filtro de Wiener baseia-se em duas hipóteses: (i) os sinais envolvidos são WSS e (ii) as médias estatísticas $R_x$ e $p_{xd}$ são conhecidas. Em muitos casos práticos, dois aspectos violam estas suposições:

- a necessidade de operar em tempo real: requer métodos capazes de conjuntamente efetuar a aquisição e a otimização.
- a presença de sinais não-estacionários: inibe o uso de uma solução fechada, uma vez que não há mais valores fixos de correlação estatística.

Este novo cenário nos leva à fronteira entre filtragem ótima e adaptativa. Neste último caso, as soluções para o problema de filtragem (linear) serão determinadas de maneira iterativa/recursiva enquanto os dados são obtidos.

## * Algoritmo do gradiente Determinístico.

- conhecido como steepest-descent algorithm




- em vez de obter diretamente $\underline{w}^{ótimo} = R_x^{-1} p_{xd}$, partiremos de uma condição inicial $\underline{w}(0)$ e iterativamente atualizaremos o vetor de parâmetros na forma $\underline{w}(n+1) = f\{\underline{w}(n)\}$ até um certo critério de parada ser atingido.
- o método do gradiente expressa a função $J(\underline{w})$ apenas em termos da 1ª derivada (vetor gradiente)

$$J(\underline{w})\Big|_{\underline{w}_{i+1}} = J(\underline{w})\Big|_{\underline{w}_i} + \frac{\partial J(\underline{w})}{\partial \underline{w}^T}\Big|_{\underline{w}_i} \cdot \Delta\underline{w} + \frac{1}{2}\Delta\underline{w}^T \frac{\partial^2 J(\underline{w})}{\partial \underline{w}_i \partial \underline{w}^T}\Big|_{\underline{w}_i} \cdot \Delta\underline{w} + \dots \quad \text{(Taylor)}$$

$$\Delta w = \underline{w}_{i+1} - \underline{w}_i$$

- para obtermos o maior decrescimento no valor de $J(\underline{w})$, a variação $\Delta\underline{w}$ deve ser colinear ao vetor gradiente, mas com sentido oposto (se forem ortogonais, o produto escalar é nulo), (havendo um ângulo entre $\Delta\underline{w}$ e o vetor gradiente, a redução acontece cada vez mais conforme ele tende a $0^\circ$).

$$\Rightarrow \Delta \underline{w} = - \text{constante} \cdot \frac{\partial J(\underline{w})}{\partial \underline{w}^T}$$

Ora, $\nabla J(\underline{w}) = -2 p_{xd} + 2 R_x \underline{w}$, como deduzido durante a discussão sobre filtragem ótima. Com isto,

$$\underline{w}_{i+1} = \underline{w}_i - \text{constante} \frac{\partial J(\underline{w})}{\partial \underline{w}^T}$$

$$= \underline{w}_i - 2 \,\text{constante}\, [R_x \underline{w}_i - p_{xd}]$$

Substituindo a constante por $\mu/2$, chegamos a:

$$\boxed{\underline{w}_{i+1} = \underline{w}_i - \mu (R_x \underline{w}_i - p_{xd})}$$

$\mu$ = tamanho do passo a ser dado na direção definida pelo gradiente da função custo $J(\underline{w})$.

Análise de Convergência:

- os pontos de equilíbrio de um sistema dinâmico são invariantes com relação ao processo iterativo. Em outras palavras, eles são as soluções da equação $\underline{w}_{i+1} = \underline{w}_i$. Logo,

$\mu [R_x \underline{w} - p_{xd}] = 0 \Rightarrow \underline{w} = R_x^{-1} p_{xd}$

- o sistema possui um único ponto de equilíbrio e que corresponde à própria solução de Wiener.

O sistema em questão é estável?

• para verificarmos a estabilidade do sistema dinâmico, primeiro reescrevemos a equação de atualização do vetor de coeficientes da seguinte forma:

$$\underline{w}_{i+1} = [I - \mu R_x]\,\underline{w}_i + \mu p_{xd}$$

Para que o sistema seja estável, é preciso que os autovalores de $[I - \mu R_x]$ estejam dentro da circunferência de raio unitário

Seja $B = [I - \mu R_x]$, então, $\lambda_B = 1 - \mu \lambda_{R_x}$

Para que $|\lambda_B| < 1$, $|1 - \mu \lambda_{R_x}| < 1$.

O caso mais restritivo está associado ao autovalor de $R_x$ de maior módulo

Assim, $|1 - \mu \lambda_{R_x}^{max}| < 1 \Rightarrow -1 < 1 - \mu \lambda_{R_x}^{max} < 1$

Como $\lambda_{R_x}^{max}$ e $\mu$ são números <u>não-negativos</u>, chegamos a $\boxed{\mu < \frac{2}{\lambda_{R_x}^{max}}}$.

$\Rightarrow$ O algoritmo do gradiente determinístico não possui um sentido prático para aplicação em filtragem, pois pressupõe conhecimento total das informações estatísticas em $R_x$ e $p_{xd}$, além de estacionariedade.

Sua derivação e análise, porém, serve de base para os algoritmos estocásticos que veremos.

## ⊛ Método de Newton

- baseado na aproximação de 2ª ordem de $J(\underline{w})$
- obtemos um $\underline{\Delta w}$ que nos leva diretamente ao ponto de mínimo, ou seja, no ponto no qual $\frac{\partial J}{\partial \underline{w}^T}(\underline{w}_{i+1}) = 0$.

Derivando a expansão de Taylor de $J(\underline{w})$:

$$0 = \underbrace{\frac{\partial J(\underline{w})}{\partial \underline{w}^T}\bigg|_{\underline{w}_i}}_{\text{vetor gradiente em } \underline{w}_i} + \underbrace{\frac{\partial^2 J(\underline{w})}{\partial \underline{w}^T \partial \underline{w}}\bigg|_{\underline{w}_i}}_{\text{matriz hessiana em } \underline{w}_i : H(J(\underline{w}))} \cdot \underline{\Delta w} \quad , \quad \underline{\Delta w} = \underline{w}_{i+1} - \underline{w}_i$$

$$\Rightarrow \underline{w}_{i+1} = \underline{w}_i - H(J(\underline{w}))^{-1} \nabla J(\underline{w})$$

$$= \underline{w}_i - H(J(\underline{w}))^{-1} \left[ -2 p_{xd} + 2 R_x \underline{w}_i \right]$$

Quem é $H(J(\underline{w}))$?

$\frac{\partial E\{e^2(n)\}}{\partial \underline{w}^T \partial \underline{w}}$ - cada elemento $(i,j)$ desta matriz corresponde a $H_{ij}(J(\underline{w})) = \frac{\partial E\{e^2(n)\}}{\partial w_i \partial w_j}$

$$\frac{\partial E\{e^2(n)\}}{\partial w_i} = E\left\{ 2 e(n) \frac{\partial e(n)}{\partial w_i} \right\} = 2 E\{ e(n) x(n-i) \}$$

Logo, $\frac{\partial E\{e^2(n)\}}{\partial w_i w_j} = \frac{\partial [2E\{e(n)x(n-i)\}]}{\partial w_j} = 2E\{x(n-i)x(n-j)\} = 2r_x(i-j)$

Portanto, $\boxed{H(J(\underline{w})) = 2.R_x}$

Retornando à equação de atualização do vetor de coeficientes:

$$\underline{w}_{i+1} = \underline{w}_i - \frac{1}{2} R_x^{-1}(-2\not{p}_{xd} + 2R_x\underline{w}_i)$$

Note que podemos escrever:

$$\underline{w}_{i+1} = I\underline{w}_i + R_x^{-1}\not{p}_{xd} - I\underline{w}_i$$

$$= \underline{\underline{R_x^{-1}\not{p}_{xd}}}$$

Ou seja, para qualquer ponto de partida inicial, chegamos ao ponto ótimo em um único passo.

## ⊛ Algoritmo do gradiente estocástico (LMS)

• 1959/60: Widrow e Hoff propuseram um algoritmo iterativo de busca da solução ótima MSE (Wiener) combinado com a ideia de aproximação estocástica (Robbins e Monro, 1951).

Características:

* aquisição dos sinais e otimização do filtro realizadas concomitantemente.
* possibilidade de se trabalhar em contextos não-estacionários.
* simplicidade computacional
* boas propriedades de convergência.

Least-Mean-Square (LMS): realiza a adaptação do vetor de coeficientes com base no gradiente estocástico.

$$\left.\begin{aligned}\hat{R}_x &= \underline{x}(n)\,\underline{x}^T(n)\\ \hat{p}_{xd} &= \underline{x}(n)\,d(n)\end{aligned}\right\}$$ os valores esperados (médios) são substituídos pelos valores instantâneos a partir de estimativas não-polarizadas.

Substituindo as estimativas na equação do gradiente determinístico:

$$\underline{w}_{i+1} = \underline{w}_i - \mu(\hat{R}_x \underline{w}_i - \hat{p}_{xd})$$
$$= \underline{w}_i - \mu\,\underline{x}(i)\,\underline{x}^T(i)\,\underline{w}_i + \mu\,\underline{x}(i)\,d(i)$$

Mas, $\underline{x}^T(i)\,\underline{w}(i) = \underline{w}^T(i)\,\underline{x}(i) = \hat{d}(i)$

Então, $\underline{w}_{i+1} = \underline{w}_i + \mu\,\underline{x}(i)\underbrace{(d(i)-\hat{d}(i))}_{e(n)}$

Finalmente, mudando a notação para o índice n, obtemos:

$$\boxed{\underline{W}(n+1) = \underline{W}(n) + \mu\, \underline{x}(n)\, e(n)}$$

- Este algoritmo pode ser igualmente obtido minimizando-se diretamente, através do método do gradiente, a função custo $J(n) = e^2(n)$.

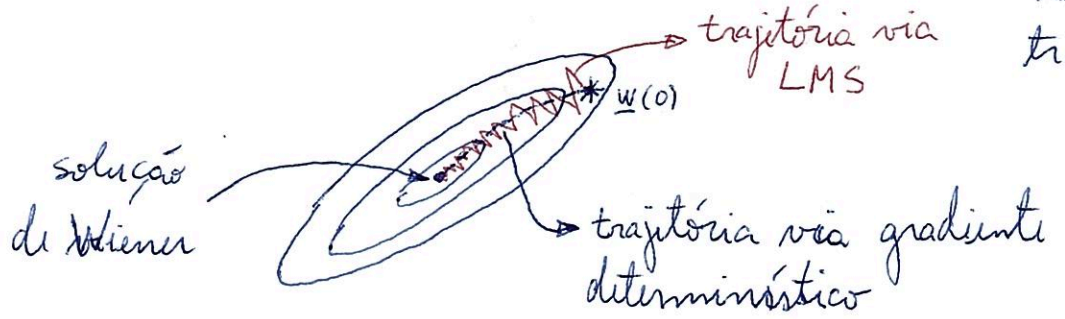
Início: $\underline{W}(0) = [0\ 0\ \dots\ 0]^T$

$\underline{x}(0) = [0\ 0\ \dots\ 0]^T$

$x(n)$ entra → $x(n-M+1)$ sai

A cada observação $x(n)$, obtenho $e(n)$ (conhecendo a saída desejada $d(n)$) e atualizo o vetor de coeficientes.

$\underline{W}(n+1) = f\{x(n), x(n-1), x(n-2), \dots, x(0)\}$ → possui uma memória de todas as amostras recebidas.




Propriedades do LMS:

- o gradiente estocástico pode ser visto como uma estimativa não-polarizada do gradiente ideal, dado um vetor fixo de coeficientes $\underline{W}$.

• <u>convergência na média</u>:

Seja $\underline{\Delta w}(n) = \underline{w}(n) - \underline{w}_{ótimo}$. Então, $E\{\underline{\Delta w}(n)\} \to 0$ para $n \to \infty$

Isso quer dizer que a tendência é ser observada – quando o tamanho do passo é adequadamente escolhido – é um comportamento oscilatório em torno de $\underline{w}_{ótimo}$.

Contudo, podemos ter situações distintas quanto à amplitude das oscilações:



Por isso, também é interessante observar o tamanho do desvio em relação ao coeficiente (ou, semelhantemente, ao erro) ótimo.

⇒ <u>Teoria da independência</u>: hipótese de que os vetores $\underline{x}(n)$, para todo $n = 0, \ldots, k$, são estatisticamente independentes

→ usada na análise de convergência do vetor de coeficientes

→ não é rigorosamente válida quando $\underline{x}(n)$ consiste de elementos de uma linha de atraso, como é o caso de um filtro FIR.

→ possibilita a obtenção de limitantes para o valor do passo de adaptação que evitem a divergência do algoritmo:

- $0 < \mu < 2/\lambda_{MAX}$, em que $\lambda_{MAX}$ é o maior autovalor de $R_x$
- $0 < \mu < 2/tr(R_x)$

$tr(R_x) = M\, r_x(0)$, em que $r_x(0) = E\{x(n)\,x(n)\} = E\{x^2(n)\} = \sigma_x^2$ (média nula)

- <u>erro em excesso e desajuste</u>

- valor esperado de $E\{e^2(n)\}$ à medida que $n \to \infty$ (situação de convergência)

$$e(n) = d(n) - \underline{w}^T \underline{x}(n)$$

$$= d(n) - \{\underline{w}_o + \underline{\Delta w}(n)\}^T \underline{x}(n) = d(n) - \underline{w}_o^T \underline{x}(n) - \underline{\Delta w}^T(n)\,\underline{x}(n)$$

$$= e_o(n) - \underline{\Delta w}^T(n)\,\underline{x}(n)$$

Então; $e^2(n) = e_o^2(n) - 2 e_o(n)\,\underline{\Delta w}^T(n)\,\underline{x}(n) + \underline{\Delta w}^T(n)\,\underline{x}(n)\,\underline{x}^T(n)\,\underline{\Delta w}(n)$

($e_o(n)$ → ortogonais $E\{e_o(n)\underline{x}(n)\} = \underline{0}$)

Aplicando o operador de esperança, chega-se a:

$$E\{e^2(n)\} = J_{MIN} + \underbrace{E\{\underline{\Delta w}^T(n)\, R_x\, \underline{\Delta w}(n)\}}$$

↳ erro residual devido à natureza estocástica do algoritmo.

A partir de $J_{EXC} = E\{\underline{\Delta w}^T(n)\, R_x\, \underline{\Delta w}(n)\}$, é mais comum avaliar o desempenho do algoritmo por meio de uma grandeza chamada de desajuste (misadjustment):

$$\mathcal{M} \triangleq \frac{J_{EXC}}{J_{MIN}} \approx \frac{1}{2}\,\mu\, M\, \sigma_x^2$$

• Velocidade de convergência (constante de tempo)

$\tau \approx \frac{1}{2\mu\lambda_{AVG}}$, em que $\lambda_{AVG} = \frac{1}{M}\sum_{i=1}^{M}\lambda_i$ = média dos autovalores de $R_x$

Observe que:

* ao aumentarmos o tamanho do passo ($\mu$), é possível reduzir a constante de tempo $\tau$, o que significa que o tempo de acomodação do algoritmo LMS foi reduzido. Ou seja, o aumento do passo proporciona um aumento na velocidade de convergência do algoritmo.

* por outro lado, quanto maior o valor de $\mu$, maior o desajuste, ou, equivalentemente, maior o erro em excesso, o que significa que perdemos qualidade ou precisão na aproximação dos coeficientes ótimos.

Constatação: um passo variável $\mu(n)$ seria ideal

→ valor de $\mu$ elevado para erros altos - enquanto estamos longe da solução ótima, permitimos valores mais elevados do passo para chegar mais rápido ao alvo.

→ valor de $\mu$ pequeno para erros baixos - quando já aconteceu a "convergência" para a região próxima à solução ótima, reduzo $\mu$ para diminuir o desajuste.

→ no caso estacionário, inicio o algoritmo com um valor de passo e gradativamente o reduzo, proporcionando maior rapidez no início e precisão no final.