

Decizie și Estimare în Prelucrarea Informației

## Capitolul III. Elemente de Teoria Estimării

### III.1 Introducere

## Ce înseamnă “estimare”?

- ▶ Un emițător transmite un semnal  $s_\Theta(t)$  care depinde de parametru **necunoscut**  $\Theta$
- ▶ Semnalul este afectat de zgomot, se recepționează

$$r(t) = s_\Theta(t) + \text{zgomot}$$

- ▶ Vrem să **găsim** valoarea parametrului  $\Theta$ 
  - ▶ pe baza eșantioanelor din semnalul recepționat, sau a întregului semnal
  - ▶ datele recepționate au zgomot  $\Rightarrow$  parametrul este “estimat”
- ▶ Valoarea găsită este  $\hat{\Theta}$ , **estimatul** lui  $\Theta$ 
  - ▶ există întotdeauna eroare de estimare  $\epsilon = \hat{\Theta} - \Theta$

## Ce înseamnă “estimare”?

- ▶ Exemple:

- ▶ Amplitudinea unui semnal constant:  $r(t) = A + \text{zgomot}$ , trebuie estimat  $A$
- ▶ Faza unui semnal sinusoidal:  $r(t) = \cos(2\pi ft + \phi) + \text{zgomot}$ , de estimat  $\phi$
- ▶ Exemple mai complicate:
  - ▶ De estimat/decis ce cuvânt este pronunțat într-un semnal vocal

- ▶ Fie următoarea problemă de estimare:

Se recepționează un semnal  $r(t) = A + \text{zgomot}$ , estimați-l pe  $A$

- ▶ La detectie: se alege între **două valori cunoscute** ale  $A$ :

- ▶ de ex.  $A$  poate fi 0 sau 5 (ipotezele  $H_0$  și  $H_1$ )

- ▶ La estimare:  $A$  poate fi oricât  $\Rightarrow$  se alege între **o infinitate de opțiuni** ale  $A$

- ▶  $A$  poate fi orice valoare din  $\mathbb{R}$ , în general

- ▶ Detectie = Estimare **restrânsă** doar la un set discret de opțiuni
- ▶ Estimare = Detectie cu un număr **infiit de opțiuni** posibile
- ▶ Metodele statistice sunt similare
  - ▶ În practică, distincția între estimare și detectie nu este strictă
  - ▶ (de ex. când trebuie să alegem între 1000 de ipoteze, este “detectie” sau “estimare”?)

- ▶ Semnalul recepționat este  $r(t) = s_{\Theta}(t) + zgomot$ 
  - ▶ este afectat de zgomot
  - ▶ depinde de parametrul necunoscut  $\Theta$
- ▶ Considerăm **N eșantioane** din  $r(t)$ , luate la momentele de timp  $t_i$

$$\mathbf{r} = [r_1, r_2, \dots r_N]$$

- ▶ Eșantioanele depind de valoarea lui  $\Theta$

- ▶ Fiecare eșantion  $r_i$  este o variabilă aleatoare ce depinde de  $\Theta$  (și de zgomot)
  - ▶ Fiecare eșantion are o distribuție care depinde de  $\Theta$

$$w_i(r_i|\Theta)$$

- ▶ Întregul vector de eșantioane  $\mathbf{r}$  este o variabilă aleatoare N-dimensională ce depinde de  $\Theta$  (și de zgomot)
  - ▶ Are o distribuție N-dimensională ce depinde de  $\Theta$
  - ▶ Egală cu produsul tuturor  $w_i(r_i|\Theta)$

$$w(\mathbf{r}|\Theta) = w_1(r_1|\Theta) \cdot w_2(r_2|\Theta) \cdot \dots \cdot w_N(r_N|\Theta)$$

► Considerăm două tipuri de estimare:

1. **Estimare de plauzibilitate maximă** (Maximum Likelihood Estimation, MLE): În afară de  $r$  nu se cunoaște nimic despre  $\Theta$ , decât cel mult vreun domeniu de existență (de ex.  $\Theta > 0$ )
2. **Estimare Bayesiană**: În afară de  $r$  se mai cunoaște o distribuție *a priori*  $w(\Theta)$  a lui  $\Theta$ , care indică ce valori ale lui  $\Theta$  sunt mai probabile / mai puțin probabile

- caz mai general decât primul

## II.2 Estimarea de plauzibilitate maximă (Maximum Likelihood)

- ▶ Dacă nu se cunoaște vreo distribuție *a priori* se folosește metoda estimării de plauzibilitate maximă (“Maximum Likelihood”, ML)
- ▶ Se definește **plauzibilitatea** unui valori  $\Theta$ , dat fiind vectorul de observații  $\mathbf{r}$ :

$$L(\Theta|\mathbf{r}) = w(\Theta|\mathbf{r})$$

- ▶  $L(\Theta|\mathbf{r})$  reprezintă funcția de plauzibilitate
- ▶ “Plauzibilitatea unei valori  $\Theta$ , date fiind măsurătorile  $\mathbf{r} = \mathbf{r}_0$  este probabilitatea de a se fi generat  $\mathbf{r}$  dacă valoarea parametrului ar fi fost  $\Theta$ ”
- ▶ A se compara cu formula din Cap. 2, slide 20
  - ▶ e aceeași
  - ▶ aici “ghicim” pe  $\Theta$ , acolo “ghiceam” pe  $H_i$

Estimarea de plauzibilitate maximă (Maximum Likelihood, ML):

- ▶ Estimatul  $\hat{\Theta}_{ML}$  este **valoarea care maximizează plauzibilitatea, dat fiind valorile observate  $r$** 
  - ▶ i.e. valoarea care maximizează  $L(\Theta|r)$ , adică maximizează  $w(r|\Theta)$

$$\hat{\Theta}_{ML} = \arg \max_{\Theta} L(\Theta|r) = \arg \max_{\Theta} w(r|\Theta)$$

- ▶ Dacă  $\Theta$  aparține doar unui anumit interval, se face maximizarea doar pe acel interval

## ► Notății matematice generale

- ▶  $\arg \max_x f(x)$  = “valoarea  $x$  care maximizează funcția  $f(x)$ ”
- ▶  $\max_x f(x)$  = “valoarea maximă a funcției  $f(x)$ ”

- ▶ Estimarea ML este foarte similară cu decizia ML!
- ▶ Criteriul de decizie ML:
  - ▶ “se alege ipoteza cu plauzibilitate mai mare”:

$$\frac{L(H_1|r)}{L(H_0|r)} = \frac{w(r|H_1)}{w(r|H_0)} \stackrel{H_1}{\gtrless} 1$$

- ▶ Estimare ML:
  - ▶ “se alege valoarea care maximizează plauzibilitatea”

$$\hat{\Theta}_{ML} = \arg \max_{\Theta} L(\Theta|r) = \arg \max_{\Theta} w(r|\Theta)$$

- ▶ Cum se rezolvă problema de maximizare?
  - ▶ adică cum se găsește estimatul  $\hat{\Theta}_{ML}$  care maximizează  $L(\Theta|\text{vec}r)$
- ▶ Maximul se găsește prin derivare și egalare cu 0

$$\frac{dL(\Theta|\mathbf{r})}{d\Theta} = 0$$

- ▶ Se poate aplica **logaritmul natural** asupra funcției  $L(\Theta|\mathbf{r})$  înainte de derivare (funcția “log-likelihood”)

$$\frac{d \ln(L(\Theta|\mathbf{r}))}{d\Theta} = 0$$

## Procedura de găsire a estimatului

Procedura de găsire a estimatului ML:

1. Se găsește expresia funcției

$$L(\Theta|\mathbf{r}) = w(\mathbf{r}|\Theta)$$

2. Se pune condiția ca derivata lui  $L(\Theta|\mathbf{r})$  sau a lui  $\ln(L(\Theta|\mathbf{r}))$  să fie 0

$$\frac{dL(\Theta)}{d\Theta} = 0, \text{ sau } \frac{d \ln(L(\Theta))}{d\Theta} = 0$$

3. Se rezolvă ecuația, se găsește valoarea  $\hat{\Theta}_{ML}$

4. Se verifică că derivata a doua în punctul  $\hat{\Theta}_{ML}$  este negativă, pentru a verifica că este un punct de maxim

- ▶ Întrucât derivata = 0 și pentru maxime și pentru minime
- ▶ uneori sărim peste această etapă

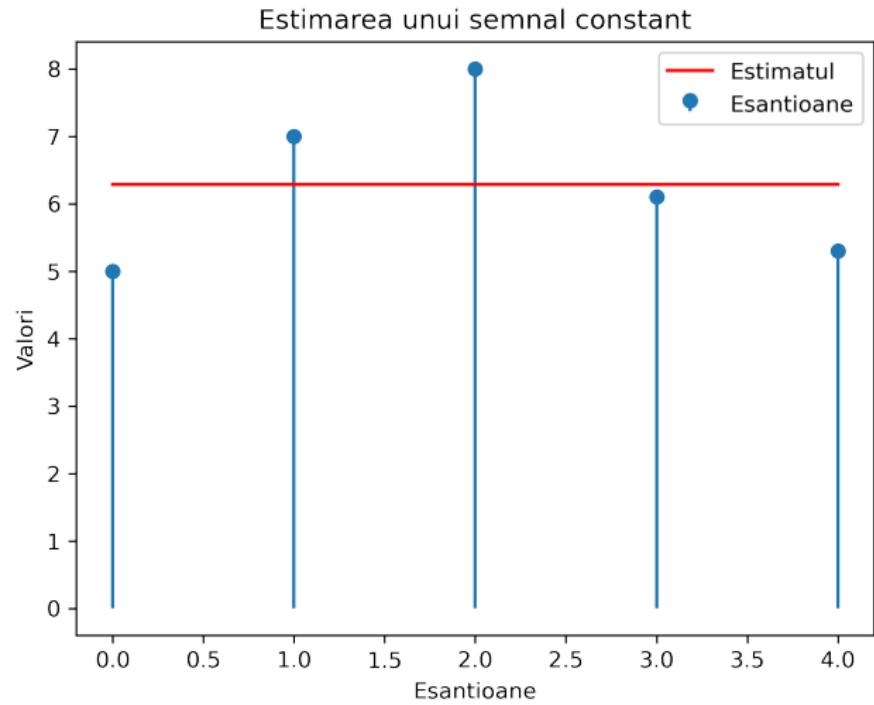
- ▶ Estimarea unui semnal constant în zgomot gaussian:

Găsiți estimatul Maximum Likelihood pentru un semnal de valoare constantă  $s_0(t) = A$  din 5 măsurători afectate de zgomot

$r_i = A + \text{zgomot}$ , cu valori egale cu [5, 7, 8, 6.1, 5.3]. Zgomotul este AWGN  $\mathcal{N}(\mu = 0, \sigma^2)$ .

- ▶ Soluție: la tablă
- ▶ Estimatul  $\hat{A}_{ML}$  este chiar valoarea medie a eșantioanelor
  - ▶ (de loc surprinzător)

# Simulare numerică



- ▶ Estimare = aproximare a unei curbe
  - ▶ se găsește cea mai bună potrivire a lui  $s_\Theta(t)$  pri datele  $\mathbf{r}$
- ▶ Din exemplul grafic anterior:
  - ▶ avem un set de date  $\mathbf{r}$
  - ▶ se cunoaște forma semnalului = o dreaptă orizontală ( $A$  constant)
  - ▶ se aproximează în mod optim dreapta prin setul de date

## Semnal oarecare în AWGN

- ▶ Fie semnalul original  $s_\Theta(t)$
- ▶ Zgomotul este AWGN  $\mathcal{N}(\mu = 0, \sigma^2)$
- ▶ Eșantioanele  $r_i$  sunt luate la momentele  $t_i$
- ▶ Eșantioanele  $r_i$  au distribuție normală, cu media  $\mu = s_\Theta(t_i)$  și varianța  $\sigma^2$
- ▶ Funcția de plauzibilitate globală = produsul plauzibilităților fiecărui eșantion  $r_i$

$$\begin{aligned}L(\Theta | \mathbf{r}) = w(\mathbf{r} | \Theta) &= \prod_{i=1}^N \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(r_i - s_\Theta(t_i))^2}{2\sigma^2}} \\&= \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^N e^{-\frac{\sum (r_i - s_\Theta(t_i))^2}{2\sigma^2}}\end{aligned}$$

- Logaritmul plauzibilității ("log-likelihood") este

$$\ln(L(\Theta|\mathbf{r})) = \underbrace{\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)}_{constant} - \frac{\sum(r_i - s_\Theta(t_i))^2}{2\sigma^2}$$

## Semnal oarecare în AWGN

- Maximul funcției = minimul exponentului

$$\hat{\Theta}_{ML} = \arg \max_{\Theta} L(\Theta | \mathbf{r}) = \arg \min \sum (r_i - s_{\Theta}(t_i))^2$$

- Termenul  $\sum (r_i - s_{\Theta}(t_i))^2$  este **distanța  $d(\mathbf{r}, s_{\Theta})$  la pătrat**

$$d(\mathbf{r}, s_{\Theta}) = \sqrt{\sum (r_i - s_{\Theta}(t_i))^2}$$

$$(d(\mathbf{r}, s_{\Theta}))^2 = \sum (r_i - s_{\Theta}(t_i))^2$$

- ▶ Estimarea ML se poate scrie sub forma:

$$\hat{\Theta}_{ML} = \arg \max_{\Theta} L(\Theta | \mathbf{r}) = \arg \min_{\Theta} d(\mathbf{r}, \mathbf{s}_{\Theta})^2$$

- ▶ Estimatul de plauzibilitate maximă (ML)  $\hat{\Theta}_{ML}$  = valoarea care face  $s_{\Theta}(t_i)$  **cel mai apropiat de vectorul recepționat r**
  - ▶ mai aproape = potrivire mai bună = mai probabil
  - ▶ cel mai aproape = cea mai bună potrivire = cel mai probabil = plauzibilitate maximă

- ▶ Estimare ML în zgomot gaussian = **minimizarea distanței**
- ▶ Aveam aceeași interpretare și la decizia ML!
  - ▶ dar la decizie alegeam minimul din 2 opțiuni
  - ▶ aici alegem minimul dintre toate opțiunile posibile
- ▶ Relația e valabilă pentru orice fel de spații vectoriale
  - ▶ vectori cu N elemente, semnale continue, etc
  - ▶ doar se înlocuiește definiția distanței Euclidiene

Procedura pentru estimarea tip ML în zgomot AWGN:

1. Se scrie expresia pentru pătratul distanței:

$$D = (d(\mathbf{r}, s_\Theta))^2 = \sum (r_i - s_\Theta(t_i))^2$$

2. Vrem minimul, deci egalem derivata cu 0:

$$\frac{dD}{d\Theta} = \sum 2(r_i - s_\Theta(t_i))(-\frac{ds_\Theta(t_i)}{d\Theta}) = 0$$

3. Se rezolvă și obținem valoarea  $\hat{\Theta}_{ML}$

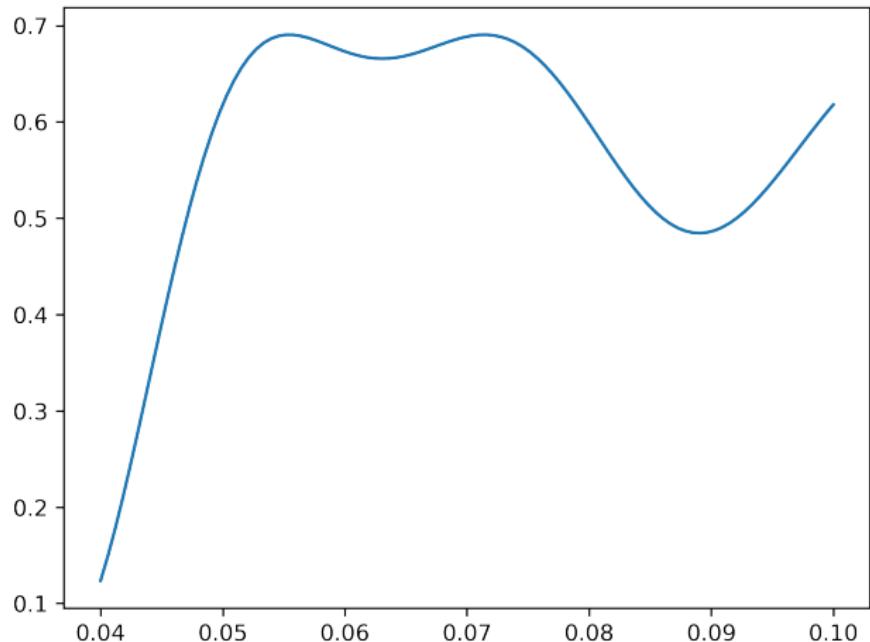
4. Se verifică că derivata a doua în punctul  $\hat{\Theta}_{ML}$  este pozitivă, pentru a se verifica că punctul este un minim

- ▶ uneori sărim peste această etapă

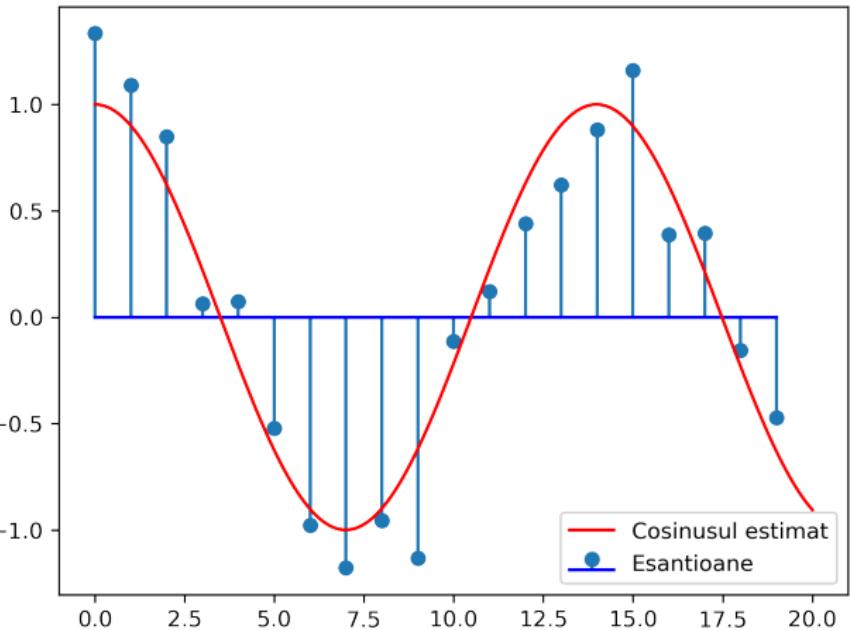
Estimarea frecvenței  $f$  a unui semnal sinusoidal

- ▶ Găsiți estimatul Maximum Likelihood pentru frecvența  $f$  a unui semnal  $s_{\Theta}(t) = \cos(2\pi ft_i)$ , din 10 măsurători afectate de zgomot  $r_i = \cos(2\pi ft_i) + \text{zgomot}$  de valori [...]. Zgomotul este AWGN  $\mathcal{N}(\mu = 0, \sigma^2)$ . Momentele de eșantionare sunt  $t_i = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9]$
- ▶ Soluție: la tablă

Funcția de plauzibilitate este



Frecvența originală = 0.070000, estimatul = 0.071515



- ▶ Estimarea ML se poate folosi și pentru a estima parametrii unor distribuții
- ▶ Avem un set de valori  $r_i$ , pe care le modelăm ca fiind eșantioane dintr-o distribuție. Cum găsim parametrii acelei distribuții?
- ▶ Momentan, considerăm un singur parametru necunoscut

## Estimarea parametrilor distribuției normale

- ▶ Presupunem că  $r_i$  sunt eșantioane dintr-o distribuție normală  $\mathcal{N}(\mu, \sigma^2)$
- ▶ Distribuția are doi parametri: media  $\mu$  și deviația standard  $\sigma$
- ▶ Estimarea lui  $\mu$ :

Este identică cu estimarea unui semnal constant în zgomot gaussian cu media 0:

$$\hat{\mu}_{ML} = \frac{1}{N} \sum_{i=1}^N r_i$$

- ▶ Estimarea lui  $\sigma^2$ :

Nu se poate formula ca estimarea unui semnal afectat, prin adunare, de zgomot gaussian, dar cu toate acestea se poate utiliza în continuare metoda ML:

$$\hat{\sigma}_{ML} = \arg \max_{\sigma} w(\mathbf{r}|\sigma)$$

## Estimarea parametrilor distribuției normale

$$\begin{aligned}\hat{\sigma}_{ML} &= \arg \max_{\sigma} w(\mathbf{r}|\sigma) \\ &= \arg \max_{\sigma} \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^N e^{-\frac{\sum(r_i - \mu)^2}{2\sigma^2}} \quad (\text{aplicăm } \ln()) \\ &= \arg \max_{\sigma} \left( -N \ln(\sigma \sqrt{2\pi}) - \frac{\sum(r_i - \mu)^2}{2\sigma^2} \right)\end{aligned}$$

Derivăm și egalăm cu 0 pentru a obține minimul:

$$\begin{aligned}-N \frac{1}{\sigma \sqrt{2\pi}} \sqrt{2\pi} - \frac{\sum(r_i - \mu)^2}{2} (-2)\sigma^{-3} &= 0 \\ -\frac{N}{\sigma} + \frac{\sum(r_i - \mu)^2}{\sigma^3} &= 0 \\ \sigma^2 &= \frac{\sum(r_i - \mu)^2}{N}\end{aligned}$$

- ▶ Estimarea parametrilor unei distribuții normale e similară cu definițiile mediei și varianței:

$$\hat{\mu}_{ML} = \frac{1}{N} \sum_{i=1}^N r_i$$
$$\hat{\sigma}_{ML} = \sqrt{\frac{\sum_{i=1}^N (r_i - \hat{\mu})^2}{N}}$$

- ▶ Notă: estimarea lui  $\sigma_{ML}$  necesită valoarea lui  $\mu$ 
  - ▶ Dacă  $\mu$  este cunoscut, totul e în regulă
  - ▶ Dacă  $\mu$  este necunoscut, se poate folosi  $\hat{\mu}_{ML}$ , dar atunci estimăm pe baza unei alte estimări, ceea ce e problematic (estimatorul este deplasat, vom vedea)

## Estimarea parametrilor distribuției uniforme

- ▶ Presupunem că  $r_i$  sunt eșantioane dintr-o distribuție uniformă  $\mathcal{U}[a, b]$
- ▶ Distribuția are doi parametri: limitele  $a$  și  $b$
- ▶ Estimarea lui  $a$  și  $b$ :

$$\hat{a}_{ML} = \arg \max_a w(\mathbf{r}|a)$$

$$\hat{b}_{ML} = \arg \max_b w(\mathbf{r}|b)$$

Prin raționament:

$$\hat{a}_{ML} = \min(r_i)$$

$$\hat{b}_{ML} = \max(r_i)$$

- ▶ Intervalul trebuie să cuprindă toate valorile (altfel, probabilitatea ar fi 0), dar nu trebuie să fie mai mare decât strict necesar (altfel, probabilitatea ar fi mai mică)

- ▶ Dacă semnalul depinde de mai mulți parametri?
  - ▶ de ex. amplitudinea, frecvența și faza inițială a unui cosinus:

$$s(t) = A \cos(2\pi ft + \phi)$$

- ▶ Se va considera  $\Theta$  ca fiind un vector:

$$\Theta = [\Theta_1, \Theta_2, \dots, \Theta_M]$$

- ▶ e.g.  $\Theta = [\Theta_1, \Theta_2, \Theta_3] = [A, f, \phi]$

- ▶ Se rezolvă cu aceeași procedură, dar în loc de o singură derivată vom avea  $M$  deriveate
- ▶ Se rezolvă sistemul:

$$\begin{cases} \frac{\partial L}{\partial \Theta_1} = 0 \\ \frac{\partial L}{\partial \Theta_2} = 0 \\ \dots \\ \frac{\partial L}{\partial \Theta_M} = 0 \end{cases}$$

- ▶ uneori este dificil/imposibil

## Coborâre după gradient (Gradient Descent)

- ▶ Cum se estimează parametrii  $\Theta$  în cazuri complicate?
  - ▶ În aplicații reale, unde pot fi foarte mulți parametri ( $\Theta$  este vector)
- ▶ De obicei nu se pot găsi valorile optime prin formule directe
- ▶ Se îmbunătățesc valorile în mod iterativ cu algoritmi tip **coborâre după gradient** (Gradient Descent)
- ▶ Gradient Descent este o metodă generală de găsire a minimului (sau a maximului) unei funcții

# Coborâre după gradient (Gradient Descent)

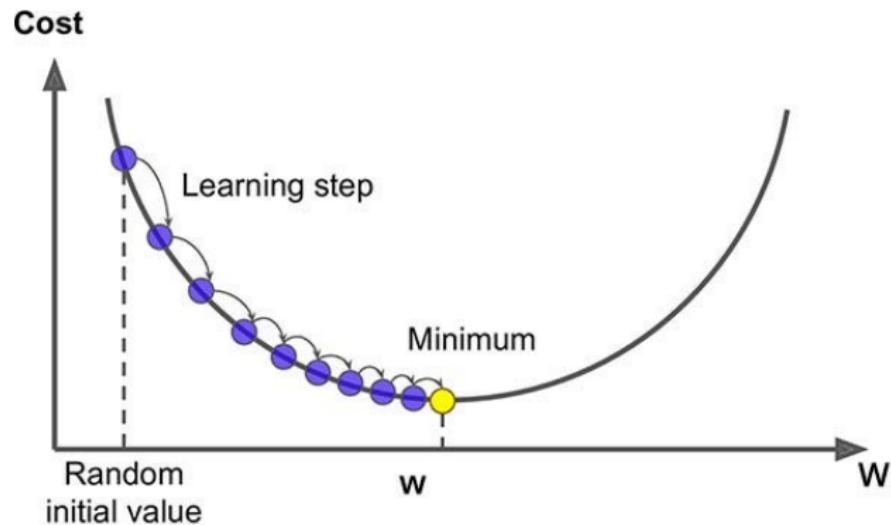


Figure 1: Coborâre după gradient<sup>1</sup>

<sup>1</sup>Imagine: Quick Guide to Gradient Descent and Its Variants, Sahdev Kansal, Towards Data Science, 2020

## Coborâre după gradient (Gradient Descent)

1. Se initializează parametrii cu valori aleatoare  $\Theta^{(0)}$
2. Repetă la fiecare iterație  $k$ :
  - 2.1 Se calculează funcția  $L(\Theta^{(k)} | \mathbf{r})$
  - 2.2 Se calculează derivatele  $\frac{\partial L}{\partial \Theta_i^{(k)}}$  pentru toți  $\Theta_i$  ("Gradient")
  - 2.3 Se actualizează toate valorile  $\Theta_i$  prin scăderea derivatei ("Descent"):

$$\Theta_i^{(k+1)} = \Theta_i^{(k)} - \mu \frac{\partial L}{\partial \Theta_i^{(k)}}$$

► sau, sub formă vectorială:

$$\Theta^{(k+1)} = \Theta^k - \mu \frac{\partial L}{\partial \Theta^{(k)}}$$

3. Până la îndeplinirea unui criteriu de terminare (de ex. parametrii nu se mai modifică mult)

## Coborâre după gradient (Gradient Descent)

- ▶ În fiecare punct, derivata ne spune în ce direcție să mergem
- ▶ Pentru găsirea minimului unei funcții, se scade derivata (coborâre după gradient)

$$\Theta^{(k+1)} = \Theta^{(k)} - \mu \frac{\partial L}{\partial \Theta^{(k)}}$$

- ▶ Pentru găsirea maximului unei funcții, se adună derivata (urcare după gradient)

$$\Theta^{(k+1)} = \Theta^{(k)} + \mu \frac{\partial L}{\partial \Theta^{(k)}}$$

- ▶ Parametrul  $\mu$  se numește **rată de învățare** (learning rate) și se alege empiric, la o valoare mică
- ▶ GD depinde de valoarea inițială, și poate ajunge la minimul local, nu global
- ▶ Alte explicații la tablă
- ▶ Exemplu practic: regresia logistică cu valori 2D

- ▶ Cel mai proeminent exemplu: **Rețele Neurale Artificiale** (a.k.a. “Rețele Neurale”, “Deep Learning”, etc.)
  - ▶ Pot fi văzute ca un exemplu de estimare ML
  - ▶ Se utilizează algoritmul *Gradient Descent* pentru găsirea parametrilor
  - ▶ Aplicații de vârf: recunoașterea de imagini, automated driving etc.
- ▶ Mai multe informații despre rețele neurale / machine learning:
  - ▶ căutați cursuri sau cărți online
  - ▶ IASI AI Meetup

- ▶ Cum caracterizăm calitatea unui estimator?
- ▶ Un estimator  $\hat{\Theta}$  este **o variabilă aleatoare**
  - ▶ poate avea diverse valori, pentru că se calculează pe baza eșantioanelor recepționate, care depind de zgomot
  - ▶ exemplu: se repetă aceeași estimare pe calculatoare diferite => valori estimate ușor diferite
- ▶ Fiind o variabilă aleatoare, se pot defini:
  - ▶ valoarea medie a estimatorului:  $E \{ \hat{\Theta} \}$
  - ▶ varianța estimatorului:  $E \{ (\hat{\Theta} - E \{ \hat{\Theta} \})^2 \}$

# Deplasarea și varianța estimatorilor

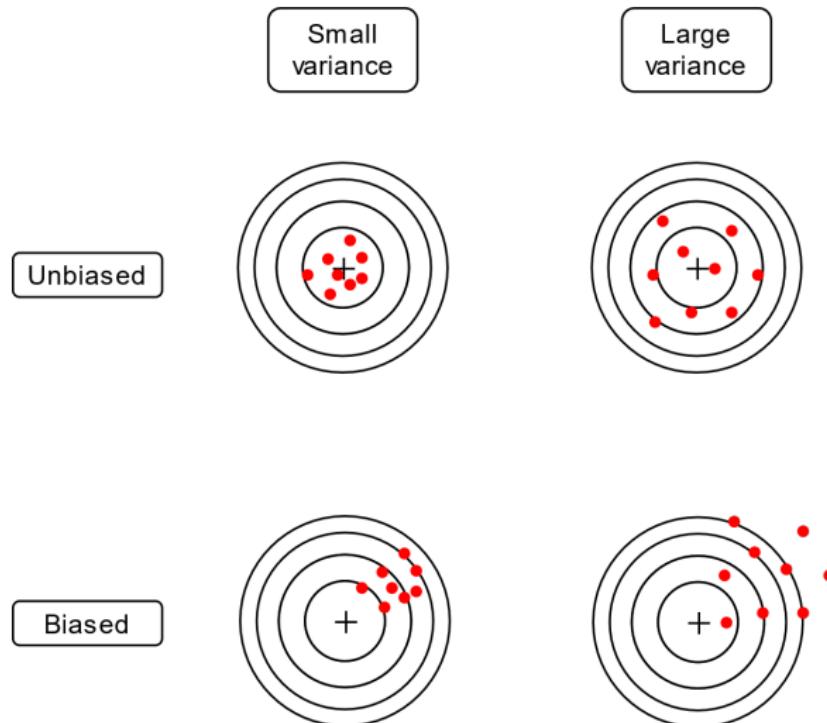


Figure 2: Deplasarea și varianța estimatorilor

- **Deplasarea** ("bias") unui estimator = diferența dintre valoarea medie a estimatorului și valoarea adevărată  $\Theta$

$$\text{Deplasare} = E\{\hat{\Theta}\} - \Theta$$

- Estimator **nedeplasat** = valoarea medie a estimatorului este egală cu valoarea adevărată a parametrului  $\Theta$

$$E\{\hat{\Theta}\} = \Theta$$

- Estimator **deplasat** = valoarea medie a estimatorului diferă de valoarea adevărată a parametrului  $\Theta$

- diferența  $E\{\hat{\Theta}\} - \Theta$  este **deplasarea** estimatorului

## Deplasarea unui estimator

- ▶ Exemplu: semnal constant  $A$ , zgomot Gaussian (cu media 0), estimatorul de plauzibilitate maximă este  $\hat{A}_{ML} = \frac{1}{N} \sum_i r_i$

- ▶ Atunci:

$$\begin{aligned} E\left\{\hat{A}_{ML}\right\} &= \frac{1}{N} E\left\{\sum_i r_i\right\} \\ &= \frac{1}{N} \sum_{i=1}^N E\{r_i\} \\ &= \frac{1}{N} \sum_{i=1}^N E\{A + \text{zgomot}\} \\ &= \frac{1}{N} \sum_{i=1}^N A \\ &= A \end{aligned}$$

- ▶ Acest estimator este nedeplasat

## Deplasarea unui estimator

- Exemplu: estimatorul varianței unei distribuții normale, când se folosește media estimată  $\hat{\mu}_{ML}$ :

$$\hat{\sigma}_{ML}^2 = \frac{\sum_{i=1}^N (r_i - \hat{\mu}_{ML})^2}{N}$$

- Acet estimator este **deplasat**:

$$\begin{aligned} E\left\{\hat{\sigma}_{ML}^2\right\} &= E\left\{\frac{\sum_{i=1}^N (r_i - \hat{\mu}_{ML})^2}{N}\right\} \\ &= \dots \\ &= \frac{N-1}{N} \sigma^2 \end{aligned}$$

unde  $\sigma^2$  este varianța reală a distribuției

- Demonstrație: Wikipedia sau “Maximum Likelihood Estimator for Variance is Biased: Proof”, Dawen Liang, Carnegie Mellon University

- ▶ Estimatorul ML al varianței este deplasat, și **subestimează** varianța reală a distribuției cu un factor  $(N-1)/N$
- ▶ Pentru a obține un estimator nedeplasat al varianței, se folosește formula:

$$\hat{\sigma}_{ML}^2 = \frac{1}{N-1} \sum_{i=1}^N (r_i - \hat{\mu}_{ML})^2$$

- ▶ Diferența: se împarte la  $N - 1$  în loc de  $N$
- ▶ Justificare intuitivă: discuție la tablă, cazul cu 2 puncte; media este la mijloc; varianța e minimizată, deci subestimează varianța reală

- ▶ **Varianța** unui estimator măsoară “abaterile” estimatorului în jurul valorii medii
  - ▶ aceasta e definiția varianței  $\sigma^2$  în general
- ▶ Dacă un estimator are **varianță** mare, valoarea estimată poate fi departe de cea reală, chiar dacă estimatorul este nedeplasat
- ▶ De obicei se preferă estimatori cu **varianță mică**, tolerându-se o eventuală mică deplasare

## II.3 Estimare Bayesiană

- ▶ **Estimarea Bayesiană** ia în calcul termeni suplimentari pe lângă  $w(\mathbf{r}|\Theta)$ :
  - ▶ o distribuție *a priori*  $w(\Theta)$
  - ▶ optional, o funcție de cost
- ▶ Se obține echivalentul din estimare pentru criteriile de decizie MPE și MR

- ▶ Conceptual, estimarea Bayesiană constă în doi pași:
  1. Găsirea distribuției **posteroare**  $w(\Theta|r)$
  2. Estimarea unei valori din această distribuție, pe baza unei **funcții de cost**

- ▶ Se definește **distribuția a posteriori** a lui  $\Theta$ , date fiind observațiile  $\mathbf{r}$ , folosind **regula lui Bayes**:

$$w(\Theta|\mathbf{r}) = \frac{w(\mathbf{r}|\Theta) \cdot w(\Theta)}{w(\mathbf{r})}$$

- ▶ Termenii:
  - ▶  $\Theta$  este parametrul necunoscut
  - ▶  $\mathbf{r}$  este vectorul de observații
  - ▶  $w(\Theta|\mathbf{r})$  este probabilitatea ca parametrul să aibă valoarea  $\Theta$ , dat fiind vectorul de observații  $\mathbf{r}$ ;
  - ▶  $w(\mathbf{r}|\Theta)$  este funcția de plauzibilitate
  - ▶  $w(\Theta)$  este distribuția **a priori** a lui  $\Theta$
  - ▶  $w(\mathbf{r})$  este o constantă (distribuția **a priori** a lui  $\mathbf{r}$ ); singurul său rol este să normalizeze expresia, astfel încât integrala lui  $w(\Theta|\mathbf{r})$  să fie 1, așa cum stă bine unei distribuții de probabilitate

Comentarii:

- ▶ La estimarea ML, avem doar termenul  $w(\mathbf{r}|\Theta)$ . Acesta este o funcție de  $\Theta$ , dar nu e chiar distribuția de probabilitate lui  $\Theta$ . E doar o mărime pe care vrem să o maximizăm.
- ▶ Estimarea Bayesiană folosește însă chiar distribuția de probabilitate a lui  $\Theta$ ,  $w(\Theta|\mathbf{r})$ , calculată cu regula lui Bayes, care ne spune riguros șansele ca  $\Theta$  să aibă o anumită valoare sau alta.

- ▶ Relația precedentă arată că, în general, estimarea lui  $\Theta$  depinde de două lucruri:
  1. De vectorul observațiilor  $r$ , prin termenul  $w(r|\Theta)$
  2. De informația "a priori" avută despre  $\Theta$ , prin termenul  $w(\Theta)$
  - ▶ (numitorul  $w(r)$  se presupune constant și are doar rol de normalizare)
- ▶ Distribuția a priori  $w(\Theta)$  reflectă cunoștințele noastre anterioare despre parametrul  $\Theta$ , înainte de a avea observațiile  $r$ .
- ▶ Distribuția a posteriori  $w(\Theta|r)$  reflectă cunoștințele noastre după ce avem și observațiile  $r$ .
- ▶ Numele este "estimare Bayesiană"
  - ▶ Thomas Bayes = matematician englez, a descoperit regula cu acest nume
  - ▶ Notiunile bazate pe regula lui Bayes poartă deseori numele de "Bayesiene"

- ▶ Presupunem că se știe de dinainte o distribuție a lui  $\Theta$ ,  $w(\Theta)$ 
  - ▶ adică, știm de dinainte care e probabilitatea de a fi a anume valoare sau alta, sau un anume interval interzis etc.
  - ▶ se numește distribuția *a priori*, adică “de dinainte de a avea observațiile”
- ▶ Care este efectul ei? Estimarea va fi “trasă” puțin înspre valorile preferate de distribuția *a priori*
  - ▶ de exemplu, dacă distribuția *a priori* este concentrată în jurul unei valori, estimarea va fi mai probabil să cadă în jurul acelei valori
- ▶ Dacă nu avem informații *a priori*, putem folosi o distribuție uniformă, adică  $w(\Theta) = \text{constant}$

- ▶ Cunoaștem distribuția *a posteriori*  $w(\Theta|\mathbf{r})$ . Care este valoarea estimată?
- ▶ Se poate alege valoarea care are probabilitate maximă
- ▶ Estimatorul **Maximum A Posteriori (MAP)** este

$$\hat{\Theta}_{MAP} = \arg \max_{\Theta} w(\Theta|\mathbf{r}) = \arg \max_{\Theta} w(\mathbf{r}|\Theta) \cdot w(\Theta)$$

- ▶ Estimatorul MAP alege acea valoare  $\Theta$  unde distribuția *a posteriori*  $w(\Theta|\mathbf{r})$  este maximă
- ▶ Estimatorul MAP maximizează **produsul** dintre plauzibilitate și **distribuția *a priori***  $w(\Theta)$

## Estimatorul MAP

Exemplu: Imagine

- ▶ Estimatorul ML:

$$\arg \max w(\mathbf{r}|\Theta)$$

- ▶ Estimatorul MAP:

$$\arg \max w(\mathbf{r}|\Theta) \cdot w(\Theta)$$

- ▶ Estimatorul ML este un caz particular de MAP pentru  $w(\Theta)$  constant

- ▶  $w(\Theta) = \text{constant}$  înseamnă că toate valorile lui  $\Theta$  sunt *a priori* echiprobabile
- ▶ i.e. nu avem extra informații despre valoarea lui  $\Theta$

- ▶ Criteriul probabilității minime de eroare:  $\frac{w(r|H_1)}{w(r|H_0)} \stackrel{H_1}{\underset{H_0}{\gtrless}} \frac{P(H_0)}{P(H_1)}$
- ▶ Se poate re scrie ca  $w(r|H_1) \cdot P(H_1) \stackrel{H_1}{\underset{H_0}{\gtrless}} w(r|H_0)P(H_0)$ 
  - ▶ adică se alege ipoteza pentru care  $w(r|H_i) \cdot P(H_i)$  este mai mare
- ▶ **Criteriul de decizie MPE:** se alege ipoteza care maximizează  $w(r|H_i) \cdot P(H_i)$ 
  - ▶ dintre cele două ipoteze  $H_0, H_1$
- ▶ **Estimarea MAP:** se alege valoarea care maximizează  $w(\mathbf{r}|\Theta) \cdot w(\Theta)$ 
  - ▶ dintre toate valorile posibile pentru  $\Theta$
- ▶ Același principiu!

- ▶ Vrem să găsim un echivalent și pentru criteriul MR
- ▶ Avem nevoie de un echivalent pentru costurile  $C_{ij}$
- ▶ **Eroarea de estimare** = diferența între estimatul  $\hat{\Theta}$  și valoarea reală  $\Theta$

$$\epsilon = \hat{\Theta} - \Theta$$

- ▶ **Funcția de cost**  $C(\epsilon) =$  atribuie un cost pentru fiecare eroare de estimare posibilă
  - ▶ când  $\epsilon = 0$ , costul  $C(0) = 0$
  - ▶ erori  $\epsilon$  mici au costuri mici
  - ▶ erori  $\epsilon$  mari au costuri mari

## Functia de cost

- ▶ Funcții de cost uzuale:

- ▶ Pătratică:

$$C(\epsilon) = \epsilon^2 = (\hat{\Theta} - \Theta)^2$$

- ▶ Uniformă:

$$C(\epsilon) = \begin{cases} 0, & \text{if } |\epsilon| = |\hat{\Theta} - \Theta| \leq E \\ 1, & \text{if } |\epsilon| = |\hat{\Theta} - \Theta| > E \end{cases}$$

- ▶ Liniară:

$$C(\epsilon) = |\epsilon| = |\hat{\Theta} - \Theta|$$

- ▶ De desenat la tablă

- ▶ Funcția de cost  $C(\epsilon)$  reprezintă echivalentul costurilor  $C_{ij}$  de la detectie
  - ▶ la detectie aveam doar 4 valori:  $C_{00}$ ,  $C_{01}$ ,  $C_{10}$ ,  $C_{11}$
  - ▶ aici avem un cost pentru fiecare eroare posibilă  $\epsilon$
- ▶ Funcția de cost dictează ce valoarea alegem din distribuția  $w(\Theta|r)$

## Importanța funcției de cost

- ▶ Fie distribuția a posteriori următoare:

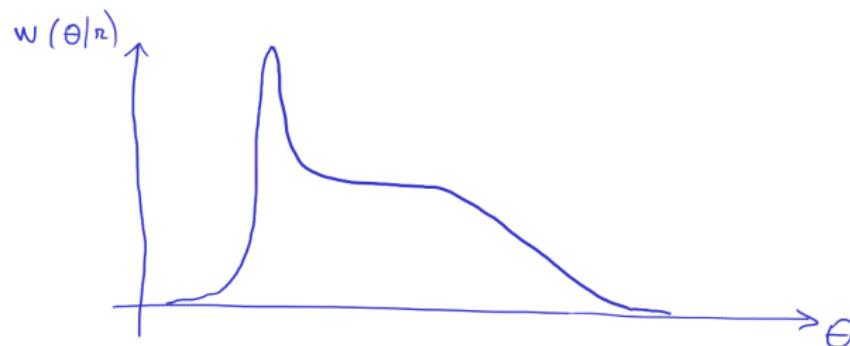


Figure 3: Asymmetrical posterior distribution

- ▶ Care este estimatorul MAP?
- ▶ Dar dacă avem funcția de cost următoare:
  - ▶ dacă estimarea  $\hat{\Theta}$  este  $<$  valoarea reală  $\Theta$ , te costă 1000 dolari
  - ▶ dacă estimarea  $\hat{\Theta}$  este  $>$  valoarea reală  $\Theta$ , platești 1 dolar
  - ▶ schimbăm valoarea estimată ? :)

- ▶ Funcția de cost este cea care impune alegerea unei anume valori estimate  $\hat{\Theta}$  de pe distribuția valorilor posibile
- ▶ Valoarea cea mai probabilă nu este întotdeauna cea mai bună!
- ▶ Valoarea cea mai bună este cea care minimizează costul (adică, valoarea medie ("expected value") a costului, întrucât acesta poate fi unoieri mai mic, alteori mai mare)

- ▶ Distribuția *a posteriori*  $w(\Theta|\mathbf{r})$  dă probabilitatea fiecărei valori  $\Theta$  de a fi cea corectă
- ▶ Alegerea unui estimat  $\hat{\Theta}$  implică o anume eroare  $\epsilon$
- ▶ Eroarea de estimare are un anumit cost  $C(\epsilon) = C(\hat{\Theta} - \Theta)$
- ▶ **Riscul** = costul mediu în raport cu toate valorile posibile ale  $\Theta$  = integrala din  $C(\epsilon) \times$  probabilitatea:

$$R = \int_{-\infty}^{\infty} C(\hat{\Theta} - \Theta) w(\Theta|\mathbf{r}) d\Theta$$

- ▶ Alegem valoarea  $\hat{\Theta}$  care **minimizează costul mediu**  $R$

$$\hat{\Theta} = \arg \min_{\hat{\Theta}} \int_{-\infty}^{\infty} C(\hat{\Theta} - \Theta) w(\Theta | \mathbf{r}) d\Theta$$

- ▶ O obținem înlocuind  $C(\epsilon) = C(\hat{\Theta} - \Theta)$  cu definiția sa, și derivând după  $\hat{\Theta}$ 
  - ▶ Atenție: se derivează după  $\hat{\Theta}$ , nu  $\Theta$ !

## Estimatorul EPMM (eroare pătratică medie minimă)

- ▶ Când funcția de cost este pătratică  $C(\epsilon) = \epsilon^2 = (\hat{\Theta} - \Theta)^2$

$$R = \int_{-\infty}^{\infty} (\hat{\Theta} - \Theta)^2 w(\Theta | \mathbf{r}) d\Theta$$

- ▶ Vrem  $\hat{\Theta}$  care minimizează  $R$ , deci derivăm

$$\frac{dR}{d\hat{\Theta}} = 2 \int_{-\infty}^{\infty} (\hat{\Theta} - \Theta) w(\Theta | \mathbf{r}) d\Theta = 0$$

- ▶ Echivalent cu

$$\underbrace{\hat{\Theta} \int_{-\infty}^{\infty} w(\Theta | \mathbf{r}) d\Theta}_{1} = \int_{-\infty}^{\infty} \Theta w(\Theta | \mathbf{r}) d\Theta$$

- ▶ Estimatorul de eroare pătratică medie minimă (EPMM)  
("Minimum Mean Squared Error, MMSE"):

$$\hat{\Theta}_{EPMM} = \int_{-\infty}^{\infty} \Theta \cdot w(\Theta | \mathbf{r}) d\Theta$$

- ▶ **Estimatorul EPMM:** estimatorul  $\hat{\Theta}$  este **valoarea medie** a distribuției *a posteriori*  $w(\Theta|\mathbf{r})$

$$\hat{\Theta}_{EPMM} = \int_{-\infty}^{\infty} \Theta \cdot w(\Theta|\mathbf{r}) d\Theta$$

- ▶ EPMM = “Eroare Pătratică Medie Minimă”
  - ▶ valoarea medie = sumă (integrală) din fiecare  $\Theta$  ori probabilitatea sa  $w(\Theta|\mathbf{r})$
- ▶ Estimatprul EPMM se obține din distribuția *a posteriori*  $w(\Theta|\mathbf{r})$ , considerând funcția de cost pătratică

- Dacă funcția de cost este uniformă

$$C(\epsilon) = \begin{cases} 0, & \text{if } |\epsilon| = |\hat{\Theta} - \Theta| \leq E \\ 1, & \text{if } |\epsilon| = |\hat{\Theta} - \Theta| > E \end{cases}$$

- Stim că  $\Theta = \hat{\Theta} - \epsilon$
- Se obține

$$R = \int_{-\infty}^{\hat{\Theta}-E} w(\Theta|\mathbf{r})d\Theta + \int_{\hat{\Theta}+E}^{\infty} w(\Theta|\mathbf{r})d\Theta$$

$$R = 1 - \int_{\hat{\Theta}-E}^{\hat{\Theta}+E} w(\Theta|\mathbf{r})d\Theta$$

- ▶ Pentru minimizarea  $R$ , trebuie să maximizăm  $\int_{\hat{\Theta}-E}^{\hat{\Theta}+E} w(\Theta|\mathbf{r})d\Theta$ , integrala din jurul punctului  $\hat{\Theta}$
- ▶ Pentru  $E$  foarte mic, funcția  $w(\Theta|\mathbf{r})$  este aproximativ constantă, deci se va alege punctul unde funcția este maximă
- ▶ **Estimatorul Maximum A Posteriori (MAP)** = valoarea  $\hat{\Theta}$  care maximizează  $w(\Theta|\mathbf{r})$

$$\hat{\Theta}_{MAP} = \arg \max_{\Theta} w(\Theta|\mathbf{r}) = \arg \max_{\Theta} \Theta w(\mathbf{r}|\Theta) \cdot w(\Theta)$$

- ▶ Estimatorul MAP:  $\hat{\Theta} = \text{valoarea care maximizează distribuția } a \text{ posteriori}$
- ▶ Estimatorul EPMM:  $\hat{\Theta} = \text{valoarea medie a distribuției } a \text{ posteriori}$

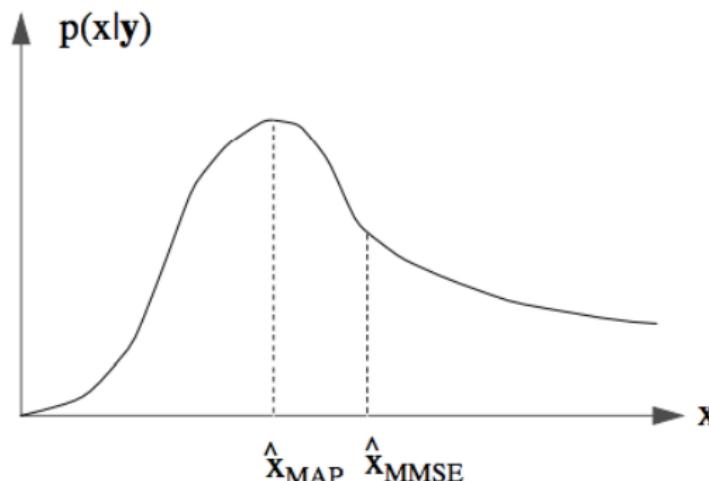


Figure 4: Estimatorul MAP vs EPMM(MMSE)

- ▶ Estimatorul MAP = minimizează costul mediu, folosind funcția de cost uniformă
  - ▶ ca le detectie: criteriul MPE = criteriul MR când costurile sunt la fel
- ▶ Estimatorul EPMM = minimizează costul mediu, folosind funcția de cost pătratică
  - ▶ similar cu criteriul MR, dar la estimare

### 1. Filtru Kalman: estimarea poziției unui obiect în mișcare

- ▶ Se cunosc poziția și viteza inițiale ale obiectului
- ▶ Se fac măsurători succesive ale poziției obiectului
- ▶ Se estimează poziția obiectului la fiecare măsurătoare nouă

La fiecare măsuratoare nouă, avem două distribuții ale poziției:

- ▶ cea dată de măsurătoare respectivă,  $w(r|\Theta)$  ("likelihood")
- ▶ cea **prezisă** pe baza poziției și vitezei de data trecută,  $w(\Theta)$  ("prior")
- ▶ ambele presupuse să fi Gaussiene, caracterizate doar prin medie și varianță

Cele două se combină prin regula lui Bayes  $\Rightarrow$  o distribuție mai precisă  $w(\Theta|r)$ , tot Gaussiană ("posterior")

- ▶ poziția exactă se estimează prin EPMM (media lui  $w(\Theta|r)$ )
- ▶  $w(\Theta|r)$  prezice poziția de la momentul următor

# Filtru Kalman pentru estimarea poziției unui obiect în mișcare

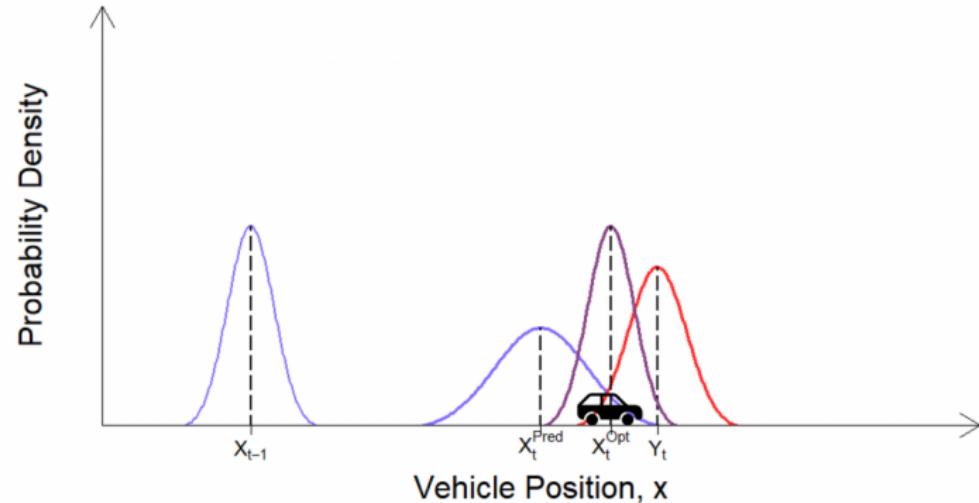


Figure 5: Estimarea poziției cu filtru Kalman <sup>2</sup>

<sup>2</sup>Image source: <https://www.lancaster.ac.uk/stor-i-student-sites/jack-trainer/how-nasa-used-the-kalman-filter-in-the-apollo-program/>

## Filtru Kalman pentru estimarea poziției unui obiect în mișcare

- ▶ Predicția poziției obiectului la pasul următor necesită cunoașterea vitezei
- ▶ În exemplul anterior, viteză este cunoscută (constantă, sau cu o distribuție cunoscută)
- ▶ În general, filtrele Kalman estimează și **viteză** unui obiect în mișcare, doar pe baza pozițiilor măsurate

## Filtru Kalman pentru estimarea poziției unui obiect în mișcare

Filtru Kalman pentru estimarea poziției și vitezei simultan (de-a lungul unei singure axe)

- ▶ Se estimează simultan poziția și viteza, adică starea  $s = \begin{bmatrix} x \\ v \end{bmatrix}$
- ▶ Se lucrează cu distribuții Gaussiene 2D (poziție, viteză) (mișcarea este pe o singură axă)

Etape:

1. La pasul  $k$ , se cunoaște distribuția stării  $s^k$
2. Se prezice distribuția stării la pasul  $k + 1$  ("prior distribution")
3. Se face o măsurătoare a poziției  $z^{k+1}$  ("likelihood")
4. Se calculează noua distribuție a stării ("posterior distribution"), pe baza regulii lui Bayes, înmulțindu-le
5. Media acestei distribuții este estimatul stării (poziției și vitezei) la pasul  $k + 1$  (estimare EPMM)

# Filtru Kalman pentru estimarea poziției unui obiect în mișcare

Ilustrare:

1. <https://www.bzarg.com/p/how-a-kalman-filter-works-in-pictures/>
2. <https://towardsdatascience.com/what-i-was-missing-while-using-the-kalman-filter-for-object-tracking-8e4c29f6b795>

Aplicație:

- ▶ Netezarea traiectoriilor: proiect TrafAlert