MODEL-BASED REGRESSION

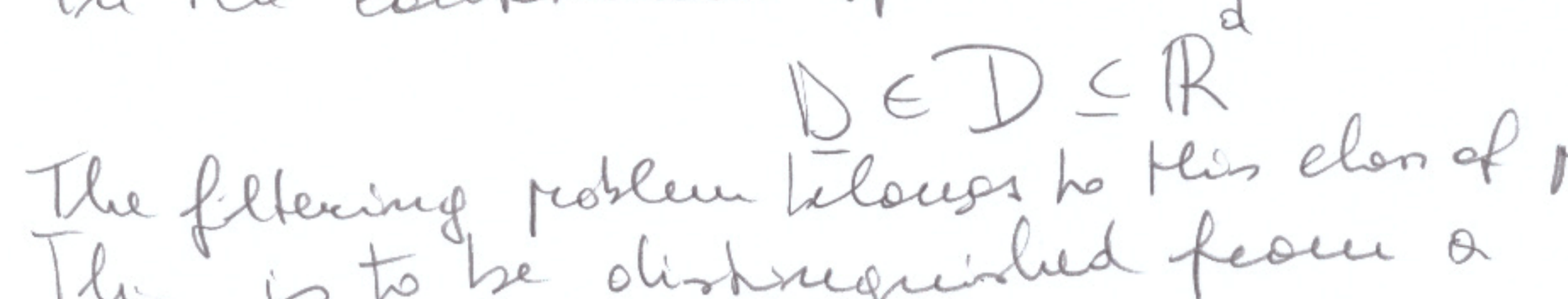
lezioni di Telecomunicazioni

Prof. FRANCESCO A.N. PALMIERI
(Università della Campania)

NOTE DAL CORSO DI "SIGNAL PROCESSING AND DATA FUSION"
A.A. 2022-23

R.D

In general estimation problems, we assume that our observations are the available data to estimate hidden parameters.



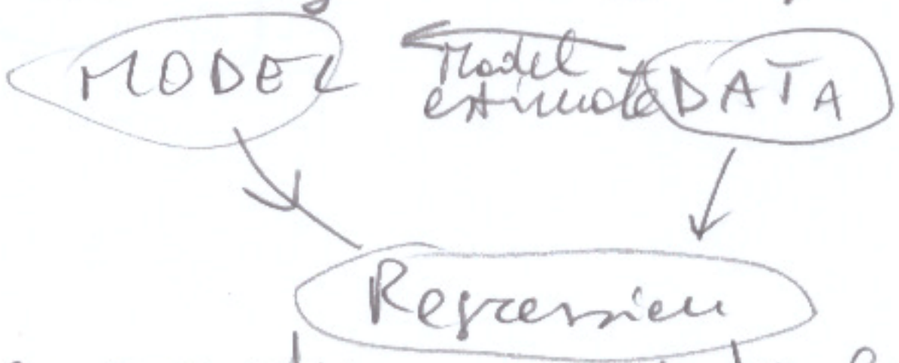
A "regression" problem is posed when the parameters to be estimated are defined in the continuous space

$$\underline{D} \in \mathcal{D} \subseteq \mathbb{R}^d$$

The following problem belongs to this class of problems. This is to be distinguished from a "classification" problem in which $\underline{D}$ belongs to a discrete finite set.

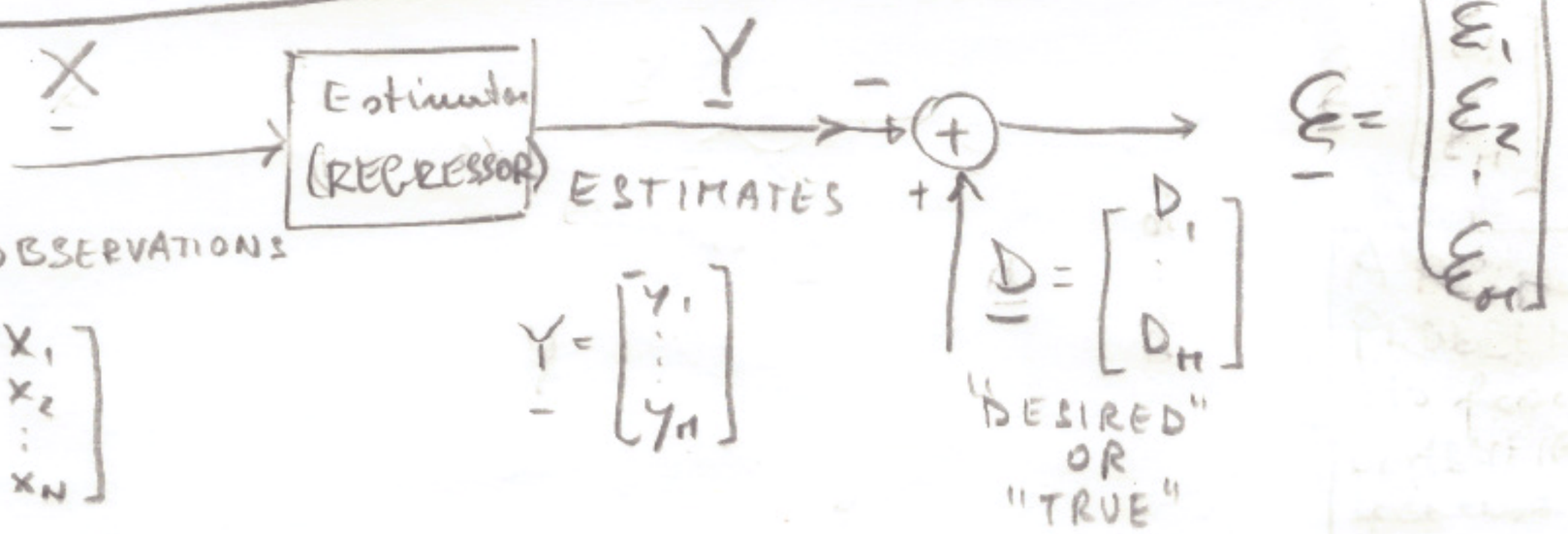
The classification problem will be treated separately.

In model-based regression we rely on knowledge of the "generative model" following the scheme, to be distinguished from Data-driven methods in which we only have examples $\{\underline{X}(n), \underline{D}(n)\}$ (TRAINING SET)



Data can be sometimes used to estimate the model.

THE GENERAL MODEL-BASED MMSE REGRESSION



$$\underline{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix} \qquad \underline{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_M \end{bmatrix} \qquad \underline{D} = \begin{bmatrix} D_1 \\ \vdots \\ D_M \end{bmatrix} \qquad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_M \end{bmatrix}$$

$\underline{X} \in \mathcal{X}$ and $\underline{D} \in \mathcal{D}$ are two statistically dependent multidimensional random variables with joint density $p_{\underline{X},\underline{D}}(\underline{x},\underline{d})$. Also $Y \in \mathcal{Y}$

What is the function $f(\underline{x})$ that minimizes a measure of the error $\underline{e}$?

Let us define a "RISK" as a function of the error that has to measure the "badness" of the estimate.

$$R = E[\mathcal{C}(\underline{D}-\underline{Y})] = E[\mathcal{C}(\underline{D}-f(\underline{X}))] = \int_{\mathcal{X}\times\mathcal{D}} \mathcal{C}(\underline{d}-f(\underline{x}))\, p_{\underline{X},\underline{D}}(\underline{x},\underline{d})\, d\underline{x}\, d\underline{d} \qquad (1)$$

The most typical choice for $\mathcal{C}$ is the square of the euclidean norm.

$$\mathcal{C}_{MS}(\underline{D}-\underline{Y}) = (\underline{D}-\underline{Y})^T(\underline{D}-\underline{Y}) \qquad (2)$$

If different components at the output are to be considered differently, a more general measure may be

$$\mathcal{C}_{MS}(\underline{Y},\underline{D}) = (\underline{D}-\underline{Y})^T Q (\underline{D}-\underline{Y}), \qquad (3)$$

where Q is an $M\times M$ positive definite matrix.
In the sequel we will assume that $Q=I$.
All the results can be easily generalized

including Q.

Using the latter measure the risk (mean squared error) is:

$$R_{ms} = \int_{X \times D} (\underline{d}-\underline{y})^T(\underline{d}-\underline{y})\, p_{\underline{X},\underline{D}}(\underline{x},\underline{d})\, d\underline{x}\, d\underline{d}$$

$$= \int_X p_{\underline{X}}(\underline{x}) \int_D (\underline{d}-\underline{y})^T(\underline{d}-\underline{y})\, p_{\underline{D}|\underline{X}}(\underline{d}|\underline{x})\, d\underline{d}\; d\underline{x} \qquad (4)$$

the result can be obtained in two steps observing that the inner integral is the conditional risk

$$R_{ms}(\underline{x}) = E[\mathcal{C}(\underline{D}-\underline{Y})|X=x] = \int_D (\underline{d}-\underline{y})^T(\underline{d}-\underline{y})\, p_{\underline{D}|\underline{x}}(\underline{d}|\underline{x})\, d\underline{d}, \qquad (5)$$

which is a function of $\underline{x}$ (not a number!).

Expanding the product we have

$$R_{ms}(\underline{x}) = \int_D \underline{d}^T\underline{d}\, p_{\underline{D}|\underline{x}}(\underline{d}|\underline{x})\, d\underline{d} - 2\,\underline{y}^T \int_D \underline{d}\, p_{\underline{D}|\underline{x}}(\underline{d}|\underline{x})\, d\underline{d}$$

$$+ \underline{y}^T\underline{y} \int_D p_{\underline{D}|\underline{x}}(\underline{d}|\underline{x})\, d\underline{d} \qquad (6)$$

$\int_D p_{\underline{D}|\underline{x}}(\underline{d}|\underline{x})\, d\underline{d} \to 1$

which can be re-written as (CANONICAL FORM)

$$R_{ms}(\underline{x}) = \left(\underline{y}^T - \int_D \underline{d}^T p_{\underline{D}|\underline{x}}(\underline{d}|\underline{x})\, d\underline{d}\right)\left(\underline{y} - \int_D \underline{d}\, p_{\underline{D}|\underline{x}}(\underline{d}|\underline{x})\, d\underline{d}\right)$$

$$+ \int_D \underline{d}^T\underline{d}\, p_{\underline{D}|\underline{x}}(\underline{d}|\underline{x})\, d\underline{d} - \left\| \int_D \underline{d}\, p_{\underline{D}|\underline{x}}(\underline{d}|\underline{x})\, d\underline{d} \right\|^2 \qquad (7)$$

(Verifiable by expanding the last expression)

Since the last two terms do not depend on $\underline{y}$, the only minimum is achieved when the first term goes to zero, namely when

$$\underline{y}_o = \underline{f}_o(\underline{x}) = \int_D \underline{d}^T p_{D|X}(\underline{d}|\underline{x})\, d\underline{d} = E[\underline{D}|\underline{X}=\underline{x}] \quad (8)$$

The above expression is the conditional mean of $\underline{D}$ and minimizes the conditional risk $R_{ms}(\underline{x})$ which in optimal condition is

$$R_{ms_o}(\underline{x}) = E[\|\underline{d}\|^2|\underline{X}=\underline{x}] - \|\underline{y}_o\|^2 . \quad (9)$$

(Notice how this quantity still depends on $\underline{x}$.)

Now we want to prove that the conditional mean estimator, not only minimizes the conditional risk $R_{ms}(\underline{x})$, but also $R_{ms} = E[R_{ms}(\underline{x})]$.

Recall the following properties of the conditional expectation operator:

$$(a)\quad E_{X|Y}[h(\underline{X},\underline{Y})|\underline{Y}=\underline{y}] = E_{\underline{X}|\underline{Y}}[h(\underline{X},\underline{y})|\underline{Y}=\underline{y}] \quad (10)$$

$$(b)\quad E_Y[E_{X|Y}[h(\underline{X},\underline{Y})|Y=y]] = E_{\underline{X}\underline{Y}}[h(\underline{X},\underline{Y})] \quad (11)$$

(The proof is straightforward through direct substitution of the density functions)

Now for the result proved above on the conditional risk, we have that

$$E_{\underline{D}|\underline{X}}[\|\underline{D}-\underline{f}_o(\underline{x})\|^2|\underline{X}=\underline{x}] \le E_{\underline{D}|\underline{X}}[\|\underline{D}-\underline{f}_1(\underline{x})\|^2|\underline{X}=\underline{x}] \quad (12)$$

for any other function of $\underline{x}$, $\underline{f}_1(\underline{x})$
the inequality is rewritten using (a) as

$$E_{\underline{D}|\underline{X}}\left[\|\underline{D} - \underline{f}_0(\underline{X})\|^2 \mid \underline{X}=\underline{x}\right] \leq E_{\underline{D}|\underline{X}}\left[\|\underline{D} - \underline{f}_1(\underline{X})\|^2 \mid \underline{X}=\underline{x}\right] \quad . (13)$$

Taking the expected value with respect to $\underline{X}$ on both sides and using (b) we have

$$E_{\underline{D},\underline{X}}\left[\|\underline{D} - \underline{f}_0(\underline{X})\|^2\right] \leq E_{\underline{D},\underline{X}}\left[\|\underline{D} - \underline{f}_1(\underline{X})\|\right] \quad , (14)$$

which proves the claim.
We can now say that the conditional expected value

$$\boxed{\underline{f}_0(\underline{x}) = E\left[\underline{D} \mid \underline{X}=\underline{x}\right]} \qquad (15)$$

is the estimator that minimizes the risk

$$R_{ms} = E\left[\|\underline{D} - \underline{f}(\underline{X})\|^2\right] \quad . \qquad (16)$$

The minimum value of the risk is

$$R_{ms0} = E_{\underline{D}}\left[\underline{D}^T\underline{D}\right] + E_{\underline{X}}\left[\underline{f}_0^T(\underline{X})\,\underline{f}_0(\underline{X})\right] - 2E_{\underline{D},\underline{X}}\left[\underline{D}^T\underline{f}_0(\underline{X})\right]$$

$$= E_{\underline{D}}\left[\underline{D}^T\underline{D}\right] + E_{\underline{X}}\left[E_{\underline{D}|\underline{X}}\left[\underline{D}^T|\underline{X}\right]\underline{f}_0(X)\right] - 2E_{\underline{D},\underline{X}}\left[\underline{D}^T\underline{f}_0(\underline{X})\right]$$

$$= E_{\underline{D}}\left[\underline{D}^T\underline{D}\right] + E_{\underline{X}}\left[E_{\underline{D}|\underline{X}}\left[\underline{D}^T\underline{f}_0(\underline{X})|\underline{X}\right]\right] - 2E_{\underline{D},\underline{X}}\left[\underline{D}^T\underline{f}_0(\underline{X})\right]$$

$$= E\left[\underline{D}^T\underline{D}\right] - E\left[\underline{D}^T\underline{f}_0(\underline{X})\right] \qquad (17)$$

ALTERNATIVE PROOF

the minimization of R_{ms} in 14) could be achieved by computing the gradient with respect to $\underline{y}$ and setting it to zero.

$$\nabla_{\underline{y}} R_{ms} = \int_{X} p_{\underline{X}}(\underline{x}) \, \nabla_{\underline{y}} \int_{D} (\underline{d}-\underline{y})^T (\underline{d}-\underline{y}) \, p_{\underline{D}|\underline{X}}(\underline{d}|\underline{x}) \, d\underline{d} \, d\underline{x}$$

Since in the outer integral $p_{\underline{X}}(\underline{x})$ is a positive function, $\nabla_{\underline{y}} R_{ms} = \underline{0}$ needs

$$\nabla_{\underline{y}} \int_{D} (\underline{d}-\underline{y})^T (\underline{d}-\underline{y}) \, p_{\underline{D}|\underline{X}}(\underline{d}|\underline{x}) \, d\underline{d} = \underline{0}$$

Expanding and taking the gradient inside the integral we get

$$-2 \int_{D} \underline{d} \, p_{\underline{D}|\underline{X}}(\underline{d}|\underline{x}) \, d\underline{d} + 2\,\underline{y} \int_{D} p_{\underline{D}|\underline{X}}(\underline{d}|\underline{x}) \, d\underline{d} = \underline{0}$$

($\int_{D} p_{\underline{D}|\underline{X}}(\underline{d}|\underline{x}) \, d\underline{d} = 1$)

and

$$\underline{y}_o = \underline{f}_o(\underline{x}) = E[\underline{D}|\underline{X}=\underline{x}] \, .$$

To prove that this is the minimum point we could take the second derivates, in the multidimensional case, the Hessian which is (*)

$$\underline{\underline{H}}_{\underline{y}} = \left(\nabla_{\underline{y}} \nabla_{\underline{y}}^T\right) R_{ms}(\underline{x}) = 2\,\underline{\underline{I}}$$

where $\underline{\underline{I}}$ is the identity matrix. The Hessian is obviously positive definite, therefore the solution is a minimum.

ERROR COVARIANCE

Since the estimation is multi-dimensional it is also useful to look at the error covariance, matrix, namely to

$$\underline{\underline{\Sigma}}_{\underline{\varepsilon}} = E_{\underline{x},\underline{D}}\left[\left((\underline{D} - \underline{f}(\underline{x})) - E[\underline{D} - \underline{f}(\underline{x})]\right)\left((\underline{D} - \underline{f}(\underline{x})) - E[\underline{D} - \underline{f}(\underline{x})]\right)^T\right]$$

which in optimal conditions is

$$\underline{\underline{\Sigma}}_{\underline{\varepsilon}0} = E\left[(\underline{D} - \underline{f}_0(\underline{x}))(\underline{D} - \underline{f}_0(\underline{x}))^T\right],$$

since $E[\underline{D} - \underline{f}_0(\underline{x})] = E[\underline{D}] - E_x[E[D|X]] = \underline{0}$

Clearly the mean square error is

$$R_{ms} = tr\, \underline{\underline{\Sigma}}_{\underline{\varepsilon}},$$

which in optimal conditions is

$$R_{ms0} = tr\, \underline{\underline{\Sigma}}_{\underline{\varepsilon}0}.$$

(*) Note that the gradient operator $\nabla_{\underline{x}} z$ applied to the scalar z is $\left[\frac{\partial z}{\partial x_1}\ \frac{\partial z}{\partial x_2} \cdots \frac{\partial z}{\partial x_N}\right]^T$. The Hessian matrix is obtained applying the operator $\nabla_{\underline{x}} \nabla_{\underline{x}}^T = \begin{bmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_N} \end{bmatrix} \left[\frac{\partial}{\partial x_1} \cdots \frac{\partial}{\partial x_N}\right] = \begin{bmatrix} \frac{\partial^2}{\partial x_1^2} & \cdots & \frac{\partial^2}{\partial x_1 \partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_1 \partial x_N} & \cdots & \frac{\partial^2}{\partial x_N^2} \end{bmatrix}$ to the scalar z.

EXAMPLE 1: $\underline{X}$ and $\underline{D}$ jointly gaussian

If the observation $\underline{X}$ and the "desired" output $\underline{D}$ are jointly gaussian with

$$E[\underline{X}] = \underline{\bar{x}}\,, \quad E[\underline{D}] = \underline{\bar{d}}\,, \quad \text{cov}[\underline{X}] = \underline{\underline{\Sigma}}_x\,, \quad \text{cov}[\underline{D}] = \underline{\underline{\Sigma}}_D$$

$$E\left[(\underline{X} - \underline{\bar{x}})(\underline{D} - \underline{\bar{d}})^T\right] = \underline{\underline{\Sigma}}_{xD}\,,$$

we have, using the results on joint gaussian variables that the minimum mean square error filter is given by

$$f_o(\underline{x}) = E[\underline{D} \mid \underline{X} = \underline{x}] = \underline{\bar{d}} + \underline{\underline{\Sigma}}_{Dx} \underline{\underline{\Sigma}}_x^{-1} (\underline{x} - \underline{\bar{x}})\,.$$

the error covariance matrix is

$$\underline{\underline{\Sigma}}_{\underline{\varepsilon} o} = E\left[(D - f_o(X))(D - f_o(X))^T\right] = \Sigma_D - \Sigma_{Dx} \Sigma_x^{-1} \Sigma_{xD}\,,$$

with minimum mean square error given by

$$R_{mso} = tr\left[\Sigma_D - \Sigma_{Dx} \Sigma_x^{-1} \Sigma_{xD}\right]\,.$$

Note how the optimum estimator $f_o(\underline{x})$ is an affine estimator (linear + constant).

If both $\underline{X}$ and $\underline{D}$ have zero mean, the estimator becomes purely linear. I would like to emphasize that even if one of the two variables, $\underline{X}$ or $\underline{D}$ has non zero mean, the estimator remains affine.

the estimator can be re-written as

$$\underline{f}_o(\underline{x}) = \underbrace{\Sigma_{DX}\Sigma_X^{-1}\underline{x}}_{\text{Linear part}} + \underbrace{\left(\bar{d} - \Sigma_{DX}\Sigma_X^{-1}\bar{\underline{x}}\right)}_{\text{biases.}} \qquad (*)$$

to emphasize the linear part and the biases.

Note that in a practical estimation problem the means are usually removed beforehand

Let us consider here in some detail some special cases of the above solutions.

EXAMPLE 1.1

$N=1$, $L=1$, $E[X]=0$, $E[D]=0$.

The solution is

$$f_o(x) = \frac{E[DX]}{E[X^2]}x = c_o x \quad , c_o = \frac{E[DX]}{E[X^2]}$$

$$R_{mso} = E[D^2] - \frac{E[DX]^2}{E[X^2]}$$

EXAMPLE 1.2

$N>1$, $L=1$, $E[\underline{X}]=\underline{0}$, $E[D]=0$

$$f_o(\underline{x}) = E[D\underline{X}^T]E[\underline{X}\underline{X}^T]^{-1}\underline{x} = \underline{c}_o^T\underline{x} \quad \text{(linear)}$$

with $\underline{c}_o = \underbrace{E[\underline{X}\underline{X}^T]^{-1}}_{\text{CORR. MATRIX}}\underbrace{E[D\underline{X}]}_{\substack{\text{CROSS-CORR.}\\ \text{VECTOR}}}$

$$R_{mso} = E[D^2] - E[D\underline{X}^T]E[\underline{X}\underline{X}^T]^{-1}E[D\underline{X}]$$

EXAMPLE 1.3

$N>1$, $L>1$ $\quad E[\underline{X}]=\underline{0}$, $E[\underline{D}]=\underline{0}$

$$\underline{f}_o(\underline{x}) = E[\underline{D}\underline{X}^T]E[\underline{X}\underline{X}^T]^{-1}\underline{x} = \underline{\underline{C}}_o^T\underline{x}$$

[N × L] [L × N]

$$\underline{\underline{C}}_o = E[\underline{X}\underline{X}^T]^{-1}E[\underline{X}\underline{D}^T]$$

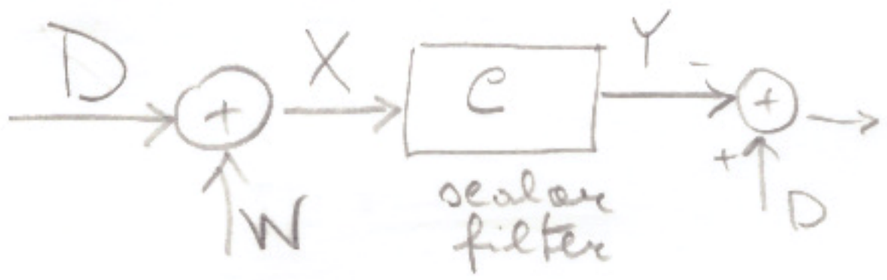
[L × N]

$$E[(\underline{D}-\underline{f}_o(\underline{X}))(\underline{D}-\underline{f}_o(\underline{x}))^T] = E[\underline{D}\underline{D}^T] - E[\underline{D}\underline{X}^T]E[\underline{X}\underline{X}^T]^{-1}E[\underline{X}\underline{D}^T]$$

EXAMPLE 1.4

Suppose the observation X is a scalar and is the result of superposition of two independent gaussian zero mean random variable D and W. We want to estimate D and assume that W is just a noise.

the statistics are $E[D]=0$, $E[W]=0$,
$E[D^2]=\sigma_D^2$ $\quad E[W^2]=\sigma_W^2$, $E[DW]=0$.



this case follows in the category of Example 1.1 since X and D are jointly gaussian, with

$$E[DX] = E[D(D+W)] = E[D^2]$$

$$E[X^2] = E[(D+W)^2] = E[D^2]+E[W^2].$$

the optimum estimator is linear, scalar and has coefficient

$$c_o = \frac{E[D^2]}{E[D^2]+E[W^2]} = \frac{\sigma_D^2}{\sigma_D^2+\sigma_W^2}$$

and minimum error

$$R_{mso} = E[D^2] - \frac{E[D^2]^2}{E[D^2]+E[W^2]}$$

$$= \frac{\cancel{E[D^2]^2} + E[W^2]E[D^2] - \cancel{E[D^2]^2}}{E[D^2]+E[W^2]}$$

$$= \frac{E[W^2]E[D^2]}{E[D^2]+E[W^2]}$$

EXAMPLE 1.5

Consider the multi-dimensional case of EXAMPLE 1.4, namely $\underline{X} = \underline{D} + \underline{W}$,

where $\underline{D}$ and $\underline{W}$ are two gaussian N-dimensional independent random variables with zero means. the statistics are $E[\underline{D}] = E[\underline{W}] = \underline{0}$, $E[\underline{W}] = \sigma_w^2 \underline{\underline{I}}$, $E[\underline{D}\underline{D}^T] = \underline{\underline{R}}_D$.
$E[\underline{D}\underline{W}^T] = \underline{\underline{0}}$.

$$E[\underline{D}\underline{X}^T] = E[\underline{D}(\underline{D}^T + \underline{W}^T)] = E[\underline{D}\underline{D}^T]$$

$$E[\underline{X}\underline{X}^T] = E[(\underline{D}+\underline{W})(\underline{D}^T+\underline{W}^T)] = E[\underline{D}\underline{D}^T] + E[\underline{W}\underline{W}^T]$$

$$= \underline{\underline{R}}_D + \sigma_w^2 \underline{\underline{I}}$$

the optimum estimator for the whole vector $\underline{D}$ is

$$\underline{f}_o(\underline{x}) = \underline{\underline{R}}_D \left(\underline{\underline{R}}_D + \sigma_w^2 \underline{\underline{I}}\right)^{-1} \underline{x} = \underline{\underline{C}}_o^T \underline{x}$$

$$\underline{\underline{C}}_o = \underline{\underline{R}}_D \left(\underline{\underline{R}}_D + \sigma_w^2 \underline{\underline{I}}\right)^{-1}$$

the error covariance in optimal conditions is

$$\underline{\underline{\Sigma}}_{\varepsilon o} = \underline{\underline{R}}_D - \underline{\underline{R}}_D \left(\underline{\underline{R}}_D + \sigma_w^2 \underline{\underline{I}}\right)^{-1} \underline{\underline{R}}_D^T .$$

In the special case of $E[DD^T] = \sigma_D^2 \underline{\underline{I}}$,

we have

$$C_0 = \frac{\sigma_D^2}{\sigma_D^2 + \sigma_W^2} \underline{\underline{I}} ,$$

which is the same as the case in example 1.4.

It is not necessary in this case to use a matrix filter, but a scalar estimator on all the components would suffice.

In the general hypothesis it is often desired to estimate just one element of $\underline{D}$ from $\underline{X}$, say D_j ($j \in \{1, \ldots, N\}$).

The cross correlations necessary are only those in

$$E[D_j \underline{X}^T] = E[D_j (\underline{D}^T + \underline{W}^T)] = E[D_j \underline{D}^T]$$

$$\hat{d}_j = f_0(\underline{x}) = E[D_j \underline{D}^T] \left(R_D + \sigma_W^2 I\right)^{-1} \underline{x} = \underline{C}_0^T \underline{x}$$

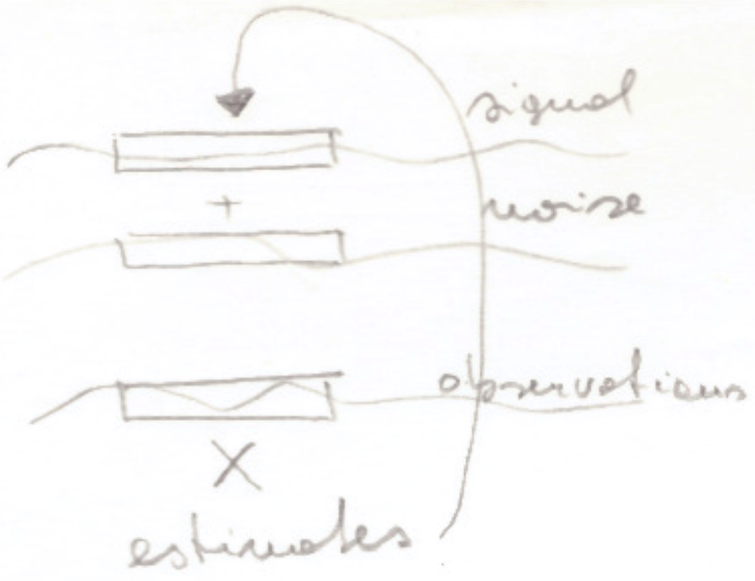
$$\underline{C}_0 = \left(R_D + \sigma_W^2 I\right)^{-1} E[D_j \underline{D}]$$

the minimum mean square error is

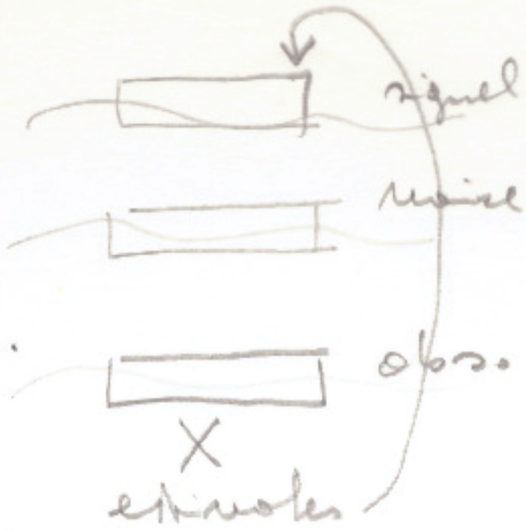
$$E[(d_j - f_0(x))^2] = E[D_j^2] - E[D_j D^T] \left(R_D + \sigma_W^2 I\right)^{-1} E[D_j \underline{D}]$$

This last case is common in signal processing when we have a stationary gaussian signal buried in gaussian noise. A filter estimates the central value of the signal from an observation window $\underline{X}$

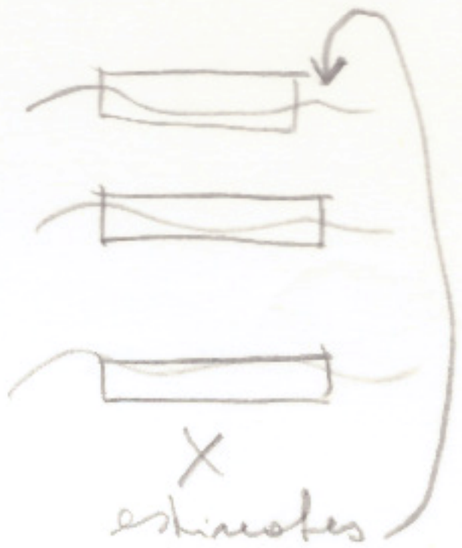
The usual nomenclature is the following one:



SMOOTHER



FILTER



PREDICTOR
(forward)

EXAMPLE 1.6

Suppose that the observation vector $\underline{X}$ is the result of the superposition of a scalar random variable (gaussian) and an independent gaussian random vector.

$$\underline{X} = D\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix} + \underline{W} = D\underline{e} + \underline{W}$$

We want to estimate D. This is a so-called location estimation problem.

$$E[D\underline{X}^T] = E[D(D\underline{e}^T + \underline{W}^T)] = E[D^2]\underline{e}^T$$

$$E[\underline{X}\underline{X}^T] = E[(D\underline{e} + \underline{W})(D\underline{e}^T + \underline{W}^T)] = E[D^2]\underline{e}\underline{e}^T + \sigma_W^2 I$$

$$f_0(\underline{x}) = E[D^2]\,\underline{e}^T\left(E[D^2]\underline{e}\underline{e}^T + \sigma_W^2 I\right)^{-1}\underline{x}$$

Use the matrix inversion lemma reviewed in one of the appendices to get a more explicit result.

↘

$$\left(\underbrace{\sigma_w^2 I}_{B^{-1}} + \underbrace{\underline{e}}_{C}\underbrace{\sigma_d^2}_{D^{-1}}\underbrace{\underline{e}^T}_{C^T}\right)^{-1} = \sigma_w^{-2} I - \sigma_w^{-2} I \underline{e}\left(\frac{1}{\sigma_d^2} + \underline{e}^T \sigma_w^{-2} I \underline{e}\right)^{-1} \underline{e}^T \sigma_w^{-2} I$$

R.13

$$= \sigma_w^{-2} I - \sigma_w^{-4} \underline{e}\left(\frac{1}{\sigma_d^2} + \frac{N}{\sigma_w^2}\right)^{-1} \underline{e}^T$$

$$f_0(\underline{x}) = \left(\sigma_d^2 \underline{e}^T \sigma_w^{-2} - \sigma_d^2 \sigma_w^{-4} N \left(\frac{1}{\sigma_d^2} + \frac{N}{\sigma_w^2}\right)^{-1} \underline{e}^T\right) \underline{x}$$

$$= \left(\frac{\sigma_d^2}{\sigma_w^2} - \frac{\sigma_d^2}{\sigma_w^2} N \frac{\sigma_d^2 \sigma_w^2}{\sigma_w^2 + N\sigma_d^2}\right) \underline{e}^T \underline{x}$$

$$= \frac{\sigma_d^2}{\sigma_w^2}\left(1 - \frac{N\sigma_d^2}{\sigma_w^2 + N\sigma_d^2}\right) \underline{e}^T \underline{x}$$

$$= \frac{\sigma_d^2}{\sigma_w^2} \frac{\sigma_w^2 + N\sigma_d^2 - N\sigma_d^2}{\sigma_w^2 + N\sigma_d^2} \underline{e}^T \underline{x}$$

$$= \frac{\sigma_d^2}{\sigma_w^2 + N\sigma_d^2} \underline{e}^T \underline{x}$$

This is a modified sample mean

Rewriting $f_0(\underline{x}) = \frac{1}{N + \frac{\sigma_w^2}{\sigma_d^2}}$, for large N it tends to the sample mean.

As we will verify in the future this is also the MAP estimator as expected since symmetry of the density function.

$$R_0 = E[D^2] - E[D f_0(\underline{x})] = \sigma_d^2 - E\left[D \frac{\sigma_d^2}{\sigma_w^2 + N\sigma_d^2} \underline{e}^T (D\underline{e} + \underline{W})\right]$$

$$= \sigma_d^2 - \frac{\sigma_d^2}{\sigma_w^2 + N\sigma_d^2} N\sigma_d^2 = \frac{\sigma_d^2\sigma_w^2 + N\sigma_d^4 - \sigma_d^4 N}{\sigma_w^2 + N\sigma_d^2}$$

$$= \frac{\sigma_d^2 \sigma_w^2}{\sigma_w^2 + N\sigma_d^2} .$$

It is in general difficult to derive the exact R.14 expression for the conditional expectation estimator for a generic situation.
Let us look just at a scalar problem.

EXAMPLE 2 (A Non-gaussian example)

Assume scalar observation $X = D + W$ with D and W independent and both distributed exponentially. Namely

$$p_D(d) = \begin{cases} \lambda_D e^{-\lambda_D d} & d \geq 0 \\ 0 & \text{else} \end{cases} \qquad p_W(w) = \begin{cases} \lambda_W e^{-\lambda_W w} & w \geq 0 \\ 0 & \text{else} \end{cases}$$

to compute the conditional expectation $E[D \mid X=x]$ we need to compute the conditional density $p_{D|X}(d|x)$.

$$p_{D|X}(d|x) = \frac{p_{X|D}(x|d)\, p_D(d)}{p_X(x)}$$

Let us compute $p_X(x)$ first:

$$p_X(x) = p_D(x) * p_W(x) = \int_{-\infty}^{+\infty} \lambda_D \lambda_W e^{-\lambda_W z} u(z)\, e^{-\lambda_D(x-z)} u(x-z)\, dz$$

$$p_X(x) = 0 \qquad x < 0$$

$$p_X(x) = \int_0^x \lambda_D \lambda_W e^{-\lambda_W z} e^{\lambda_D(z-x)}\, dz \qquad x \geq 0$$

$$= \lambda_D \lambda_W e^{-\lambda_D x} \int_0^x e^{-(\lambda_W - \lambda_D) z}\, dz$$

$$= \lambda_D \lambda_w e^{-\lambda_D x} \left[\frac{e^{-(\lambda_w - \lambda_D)z}}{-(\lambda_w - \lambda_D)} \right]_0^x = \frac{\lambda_D \lambda_w}{\lambda_w - \lambda_D} e^{-\lambda_D x} \left(1 - e^{-(\lambda_w - \lambda_D)x} \right)$$

$$p_x(x) = \frac{\lambda_D \lambda_w}{\lambda_w - \lambda_D} \left(e^{-\lambda_D x} - e^{-\lambda_w x} \right) \qquad x \geq 0$$

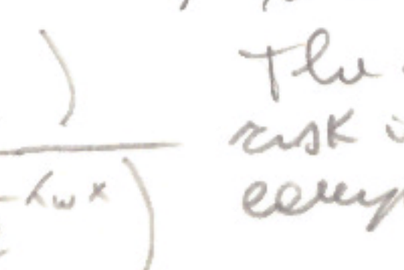
$$p_{x|D}(x|d) = p_w(x-d)$$

$$p_{D|x}(D|x) = \frac{\lambda_w e^{-\lambda_w (x-d)} u(x-d) \lambda_D e^{-\lambda_D d} u(d)}{\frac{\lambda_D \lambda_w}{\lambda_w - \lambda_D} \left(e^{-\lambda_D x} - e^{-\lambda_w x} \right) u(x)}$$

$$= (\lambda_w - \lambda_D) \frac{e^{-\lambda_w x} e^{-(\lambda_D - \lambda_w)d} u(d) u(x-d)}{\left(e^{-\lambda_D x} - e^{-\lambda_w x} \right) u(x)}$$

$$E[D|x] = \frac{(\lambda_w - \lambda_D)}{\left(e^{-\lambda_D x} - e^{-\lambda_w x} \right) u(x)} \int_0^\infty u(z)\, u(x-z)\, e^{-(\lambda_D - \lambda_w)z}\, z\, dz$$

$u(-(z-x))$



$$= \int_0^x z\, e^{-(\lambda_D - \lambda_w)z} dz = \frac{e^{-(\lambda_D - \lambda_w)z}}{-(\lambda_D - \lambda_w)} z \Bigg]_0^x - \int_0^x \frac{e^{-(\lambda_D - \lambda_w)z}}{-(\lambda_D - \lambda_w)} dz$$

$$= \frac{e^{-(\lambda_D - \lambda_w)x}}{(\lambda_D - \lambda_w)} - \frac{\left(e^{-(\lambda_D - \lambda_w)x} - 1 \right)}{(\lambda_D - \lambda_w)^2}$$

Since x is always > 0, the estimator is nonlinear and memoryless. The minimum risk is easily computed

$$\delta_0(x) = \frac{e^{-(\lambda_D - \lambda_w)x}}{e^{-\lambda_D x} - e^{-\lambda_w x}} - \frac{\left(e^{-(\lambda_D - \lambda_w)x} - 1 \right)}{(\lambda_D - \lambda_w)\left(e^{-\lambda_D x} - e^{-\lambda_w x} \right)}$$

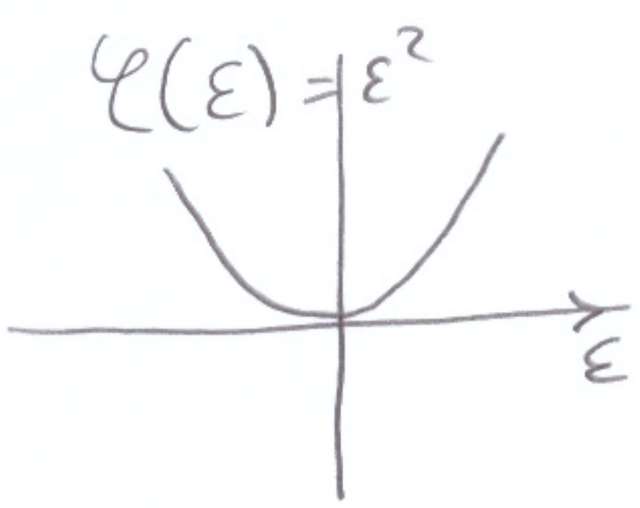
OTHER COST FUNCTIONS

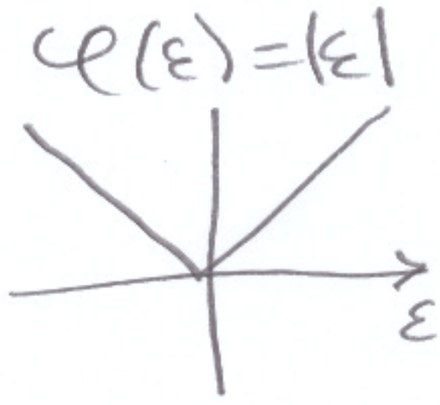
Minimum Mean Squared Error (MMSE) estimators are based on the cost function

$$\mathcal{E}_{MS} = E[\varepsilon^2[m]]$$

It is possible to use different measures (LOSS FUNCTIONS) of deviations other than the quadratic.

$$\mathcal{E} = E[\mathcal{C}(\varepsilon[m])]$$

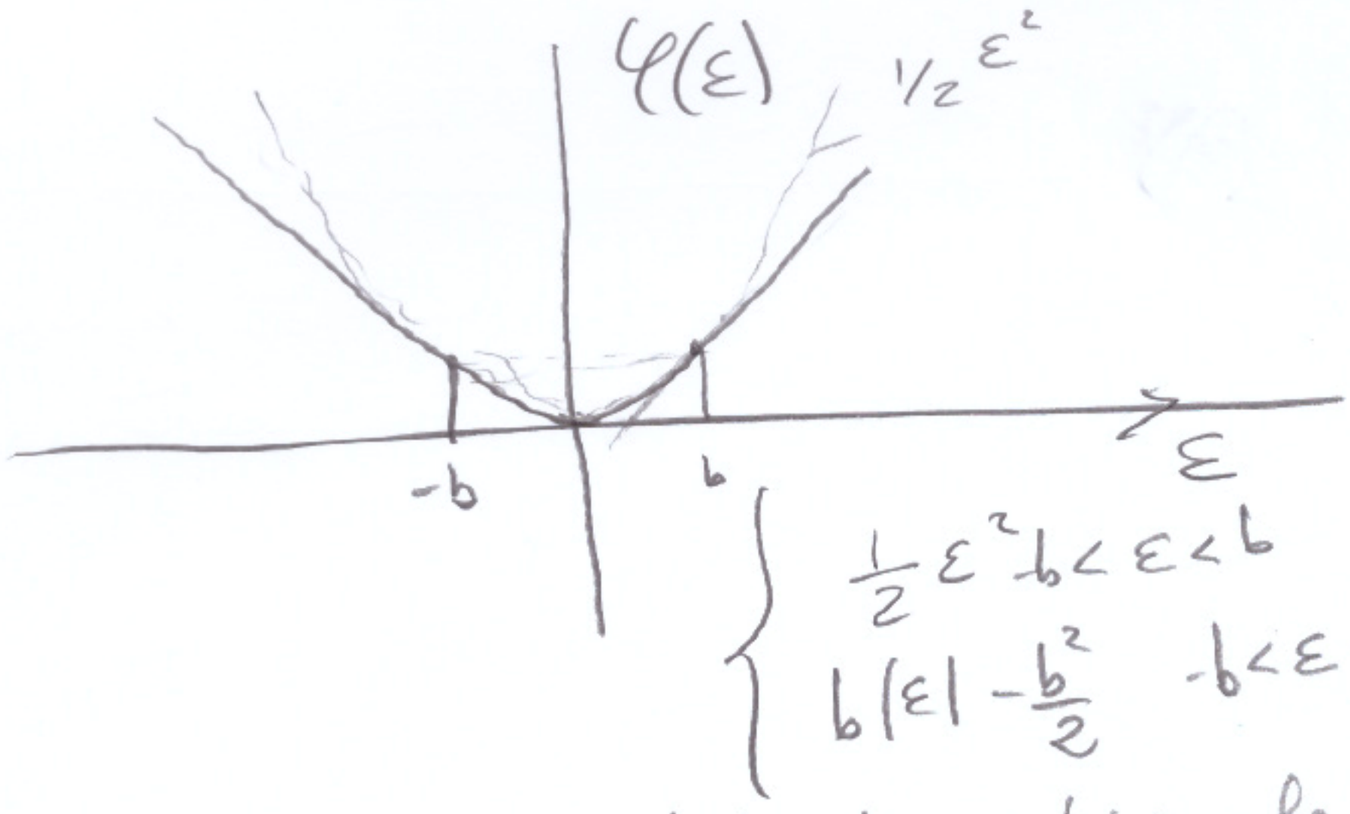




Mean Absolute Error (MAE)

MAE sometimes is preferable to MMSE because it grows linearly versus the quadratic function. The quadratic function weights more strongly large error values. Unfortunately the function $|\varepsilon|$ is discontinuous in its derivative in $\varepsilon = 0$ and does not lead to analytical solutions. In any case, sometimes it is used in search algorithms.

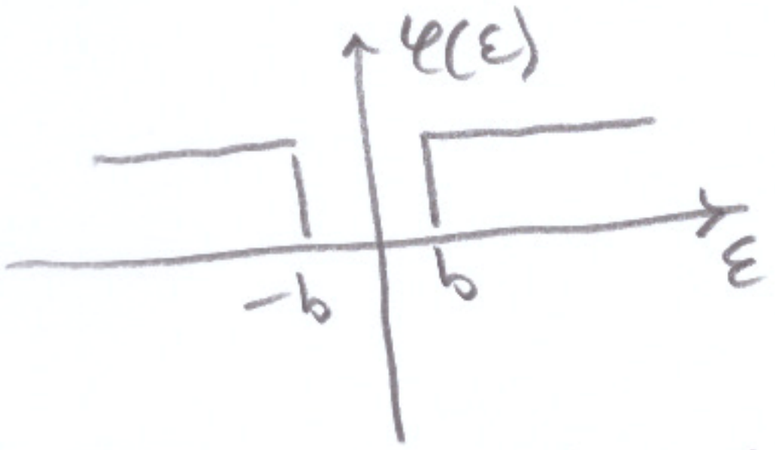
To combine the benefits of MAE and MMSE in estimation problems sometimes we use Huber loss composed by a quadratic loss for small values and an absolute deviation for large values

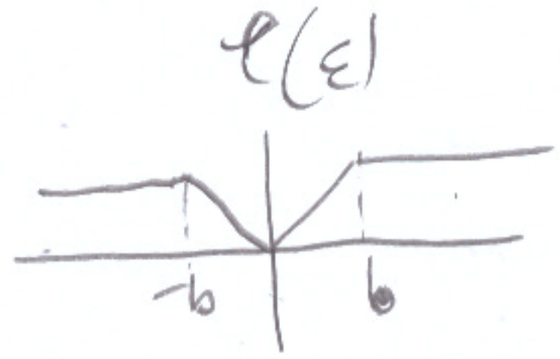


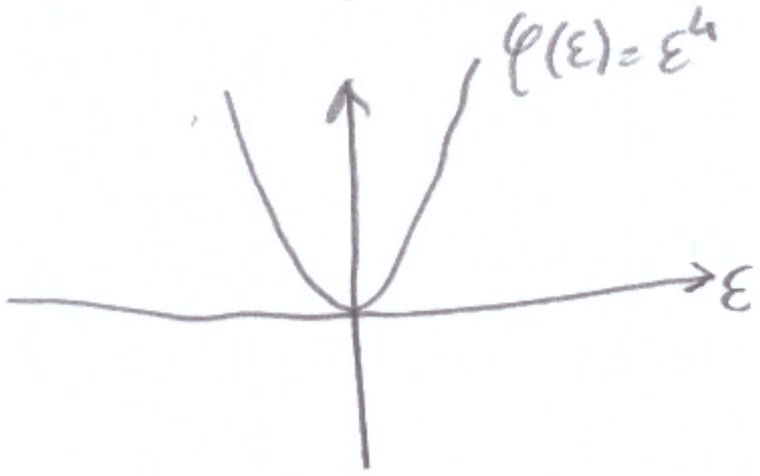
$$\begin{cases} \frac{1}{2}\varepsilon^2 & |\varepsilon| < b \\ b|\varepsilon| - \frac{b^2}{2} & |\varepsilon| > b \end{cases}$$

Note the continuity of the derivative for $\varepsilon = \pm b$.

Other loss functions are

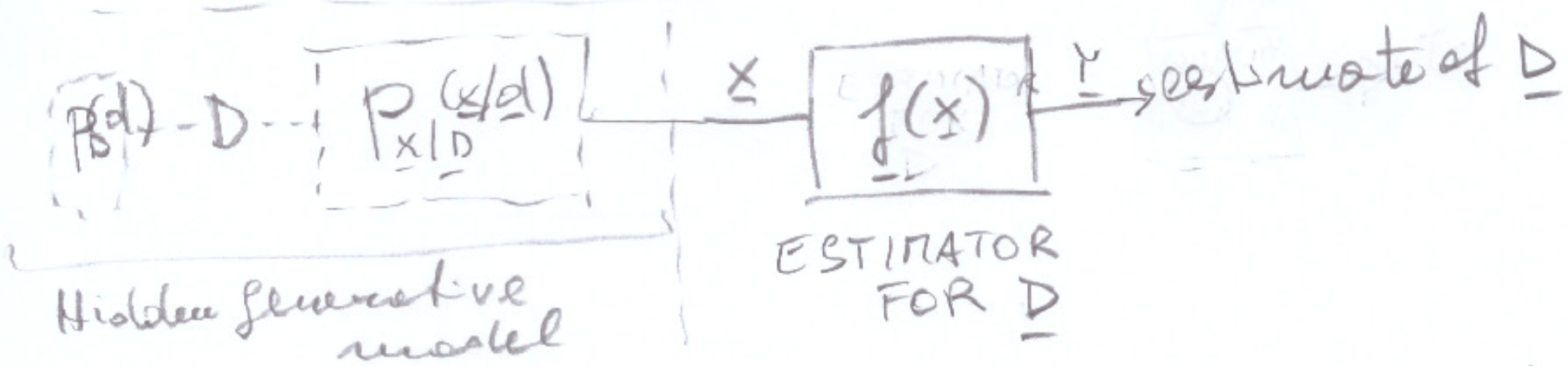






$\circ \circ \circ \circ$

MAP AND ML ESTIMATORS



When the generative model is known, ie the joint density $P_{\underline{X},\underline{D}}(\underline{x},\underline{d}) = P_{\underline{X}|\underline{D}}(\underline{x}|\underline{d}) P_{\underline{D}}(\underline{d})$ is known

↑ likelihood function ↑ prior density

An estimator for $\underline{D}$ can be found taking the value of $\underline{D}$ which maximizes the posterior probability

$$\hat{\underline{D}} = \underline{f}^{MAP}(\underline{x}) = \arg\max_{\underline{d}} P_{\underline{D}|\underline{X}}(\underline{d}|\underline{x})$$

In other words, given the observation $\underline{x}$, we take the most probable $\underline{D}$:

MAXIMUM A POSTERIORI ESTIMATE (MAP)

Note that using Bayes theorem, the posterior can be written as

$$P_{\underline{D}|\underline{X}}(\underline{d}|\underline{x}) = \frac{P_{\underline{X}|\underline{D}}(\underline{x}|\underline{d}) P_{\underline{D}}(\underline{d})}{P_{\underline{X}}(\underline{x})}$$

and the MAP estimator