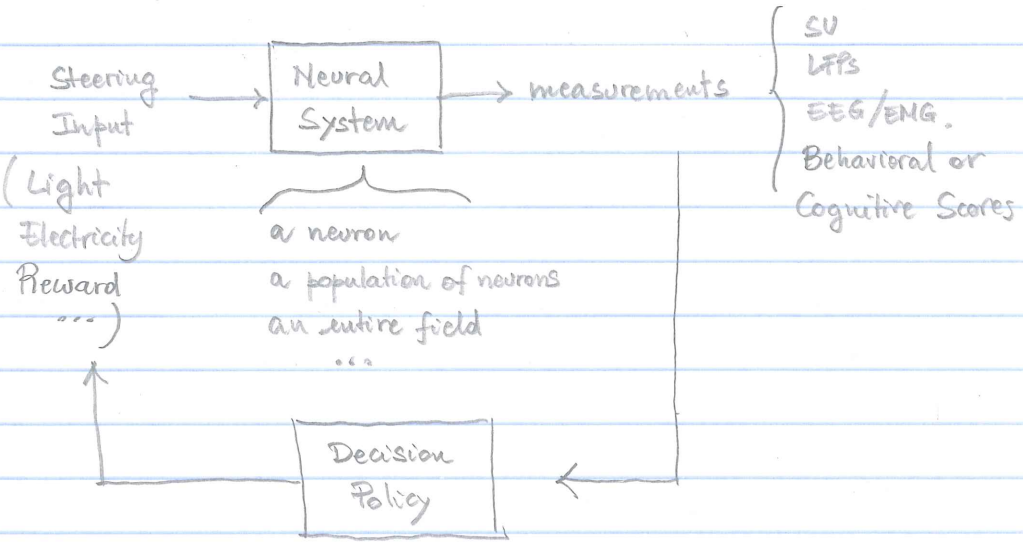
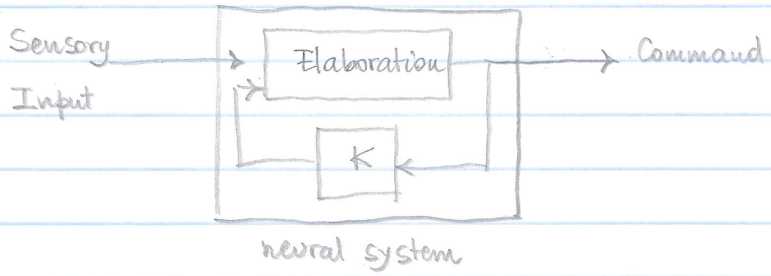
LECTURE 1

* Neural Control : what do we mean ?

1) Control = Exogenous Intervention

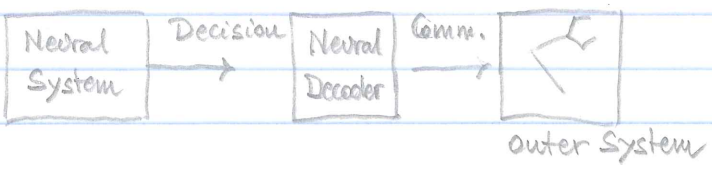


2) Control = Internal Regulation



Ex: Motor Commands
Learning, etc.

3) Control = Actuation of Outer System



In all three cases we need to provide:

- A representation of what the neural system does in response to an input

- A characterization of the uncertainty that affects such response

⇓

We need to introduce models (to represent) and estimation tools (to characterize) - We also need objectives to constrain the models

⇓

The topics of the course are: Modeling (Input - State - Output)
Estimation (State, Parameters)

Applications to:
- Regulation of Neural Activity
- Decoding of Movements
- Execution of Movements

□

* Tools that we need to carry on the estimation

1) Least - Squares

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$ - measurements (e.g., firing rates in numerous repetitions)

$$X \triangleq \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$ - explicative variables ($m < n$)

For instance, $m=1$ and x_1 is the actual firing rate (which is

supposed constant) - We have:

$$\begin{aligned} \text{error}_1 &= y_1 - x_1 \\ \text{error}_2 &= y_2 - x_1 \\ &\vdots \\ \underbrace{\text{error}_n}_{E} &= y_n - x_1 \end{aligned} \quad \leftrightarrow \quad E = Y - \underbrace{\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}}_{\triangleq A} x_1 = Y - A x_1$$

(n times)

In general, if $m > 1$, then A is $n \times m$. Moreover coefficients of A can be arbitrary:

$$E = Y - AX$$

Problem: What is the estimation $\hat{X}$ that minimizes the norm of E?

Solution:

$$\|E\|^2 \triangleq \sum_{i=1}^{n} \text{error}_i^2 = \left(Y - AX\right)^T \left(Y - AX\right) =$$

$$= Y^T Y - X^T A^T Y - Y^T A X + X^T A^T A X = f(X) \qquad (*)$$

The min of $f(X)$ (which is a scalar function) has derivatives equal to zero $\Rightarrow$ We impose:

$$\begin{cases} \frac{\partial f}{\partial x_1} = 0 \\ \frac{\partial f}{\partial x_2} = 0 \\ \vdots \\ \frac{\partial f}{\partial x_m} = 0 \end{cases}$$

If X were scalar, we would derive from $(*)$:

$$-A^T Y - Y^T A + 2 A^T A X = 0$$

$\Updownarrow$ Since it's scalar

$$2 A^T A X - 2 A^T Y = 0$$

$\Updownarrow$

$$\hat{X} = \left(A^T A\right)^{-1} A^T Y$$

The same result holds if X is a $m \times 1$ vector:

$$\hat{X} = (A^T A)^{-1} A^T Y \qquad (**)$$

Note: $(**)$ is a projection from the n-dim space of Y to the m-dim space of X ($\Rightarrow$ We are linearly combining all measurements in Y to extract all information about X)

$\downarrow$

What if not all the measurements in Y are equally good?

We introduce a matrix of weights W, e.g., a diagonal matrix:

$$W \triangleq \begin{bmatrix} \sigma_1 & & \underline{0} \\ & \sigma_2 & \\ \underline{0} & & \sigma_n \end{bmatrix}$$

and we assign each entry to one specific error:

$$\|WE\|^2 = (WY - WAX)^T (WY - WAX)$$

By repeating the argument above, we have:

$$2A^T W^T W Y - 2A^T W^T W A X = 0$$

$\Updownarrow$

$$\hat{X} = (A^T W^T W A)^{-1} A^T W^T W Y$$

Note: $W^T W = \begin{bmatrix} \sigma_1^2 & & \underline{0} \\ & \sigma_2^2 & \\ \underline{0} & & \sigma_n^2 \end{bmatrix}$ $\Rightarrow$ We are weighting each measurement a quadratic value

Note: E is a $n \times 1$ vector $\Rightarrow R = \text{cov}(E)$ is a $n \times n$ matrix $\Rightarrow$ It can be

shown that, by choosing $W^TW = R^{-1}$, the solution $\hat{X}$ is the best unbiased estimator of X:

$$\hat{X} = \left(A^T R^{-1} A\right)^{-1} A^T R^{-1} Y$$

This choice of W^TW is convenient also for another reason:

$\hat{X}$ is an estimation $\Rightarrow$ There is an error between $\hat{X}$ and X of the actual X

$\Rightarrow$ The covariance of this error is given by:

$$P \triangleq \underbrace{E}_{\text{expected value}}\left[(X-\hat{X})(X-\hat{X})^T\right] = \left(A^T R^{-1} A\right)^{-1}$$

Example: $R \triangleq \begin{bmatrix} \sigma^2 & & 0 \\ & \sigma^2 & \\ 0 & & \sigma^2 \end{bmatrix}$ $n \times n$

$A \triangleq \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$ $n \times 1$

$$\Rightarrow A^T R^{-1} A = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1/\sigma^2 & & \\ & \ddots & \\ & & 1/\sigma^2 \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$= \sum_{i=1}^{n} \frac{1}{\sigma^2} = \frac{n}{\sigma^2}$$

$\Rightarrow P = \left(A^T R^{-1} A\right)^{-1} = \sigma^2/n$ - The larger n, the lower the covariance of the error

Note: The estimation of P and $\hat{X}$ proceeds in "batch" mode ($\Rightarrow$ All measurements must be available in Y) and assumes that the A is constant $\Rightarrow$ What if A changes in time?

Assume that there are consecutive batches of measurements

$$\underbrace{Y_1}_{n_1 \times 1} \quad \underbrace{Y_2}_{n_2 \times 1} \quad \cdots \quad \underbrace{Y_K}_{n_K \times 1} \qquad n \triangleq \sum_{i=1}^{K} n_i$$

model: $A_1 X \quad A_2 X \quad A_K X$

$A_1, A_2, \ldots, A_K$ are known

$Y_1, Y_2, \ldots, Y_K$ are independent

$$\Downarrow$$

$$E = \begin{bmatrix} Y_1 \\ \hline Y_2 \\ \hline \vdots \\ Y_K \end{bmatrix} - \begin{bmatrix} A_1 \\ \hline A_2 \\ \hline \vdots \\ A_K \end{bmatrix} X \quad \text{- Block partition}$$

If we consider the first batch, we can write:

$P_1^{-1} = A_1^T R_1^{-1} A_1$ where R_1 is the covariance matrix of the error computed for the first batch

$$\hat{X}_1 = P_1 A_1^T R_1^{-1} Y_1$$

If we consider the first two batches, we have:

$$P_2^{-1} = \begin{bmatrix} A_1^T & A_2^T \end{bmatrix} \begin{bmatrix} R_1^{-1} & 0 \\ 0 & R_2^{-1} \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$$

$$= A_1^T R_1^{-1} A_1 + A_2^T R_2^{-1} A_2 = P_1^{-1} + A_2^T R_2^{-1} A_2 \qquad (a)$$

$$\hat{X}_2 = P_2 \begin{bmatrix} A_1^T & A_2^T \end{bmatrix} \begin{bmatrix} R_1^{-1} & 0 \\ 0 & R_2^{-1} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} =$$

$$= P_2 \left(A_1^T R_1^{-1} Y_1 + A_2^T R_2^{-1} Y_2 \right)$$

$$= P_2 \left(P_1^{-1} \hat{X}_1 + A_2^T R_2^{-1} Y_2 \right)$$

$$= P_2 \left(P_2^{-1} \hat{X}_1 - A_2^T R_2^{-1} A_2 \hat{X}_1 + A_2^T R_2^{-1} Y_2 \right)$$

Replace P_1^{-1} with (a)

$$= \hat{X}_1 + \underbrace{P_2 A_2^T R_2^{-1}}_{\triangleq K_2 \text{ - correction factor}} \underbrace{\left(Y_2 - A_2 \hat{X}_1 \right)}_{\text{error (innovation)}}$$

$\Rightarrow \hat{X}_2$ and P_2^{-1} are obtained recursively by using $\hat{X}$ and P^{-1} at the previous step $\Rightarrow$ The procedure can be generalized:

$$P_i^{-1} = P_{i-1}^{-1} + A_i^T R_i^{-1} A_i$$

$$K_i = P_i A_i^T R_i^{-1}$$

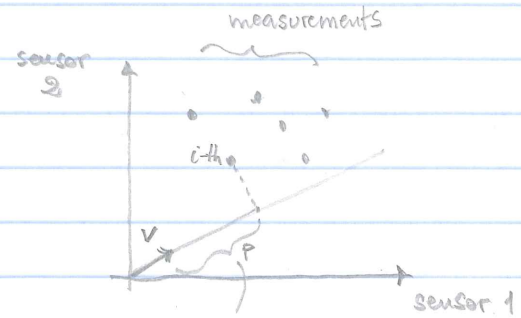
$$\hat{X}_i = X_{i-1} + K_i \left(Y_i - A_i \hat{X}_{i-1} \right)$$

□

2) Singular Value Decomposition

$$A = \begin{bmatrix} \\ n \times m \\ \\ \end{bmatrix} \Big\} \text{ e.g., measurements}$$

e.g., sensors $\qquad n \gg m$



projection of the i-th row of A on the direction of vector v

$p = \underline{a}_i \, v \Rightarrow$ For all rows: $\begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} = Av$

($\underline{a}_i$: i-th row)

$v_1 \triangleq$ first singular vector is the vector that maximizes $\|Av\|^2$

Analogously, one can define:

$v_2 \triangleq$ second singular vector is the vector that maximizes $\|Av\|^2$ conditioned to being perpendicular to v_1

The procedure can be iterated and I can find m orthogonal unit vectors that form a base for the space of rows of A

$V \triangleq [v_1 \mid v_2 \mid \dots \mid v_m]$

For each v_i, let us call: $\sigma_i \triangleq \|Av_i\|$ - i-th singular value
Now, let us consider the transformation that A performs on the vectors v_i

$u_i \triangleq \frac{1}{\sigma_i} A v_i$ - $n \times 1$ vector $\Rightarrow$ It is the i-th ~~LEFT~~ singular vector of A (it has length $=1$)

$U = [u_1 \mid u_2 \mid \dots \mid u_m]$

Now note this:

$\|Av_i\|^2 = v_i^T A^T A v_i = \sigma_i^2 \Rightarrow v_i$ is an eigenvector of A^TA

$$\|A^T u_i\|^2 = u_i^T A A^T u_i = \frac{1}{\sigma_i^2} v_i^T A^T A \underbrace{A^T A v_i}_{\sigma_i^2 v_i} = \frac{1}{\sigma_i^2} \sigma_i^2 v_i^T \cdot \sigma_i^2 v_i = \sigma_i^2$$

($\frac{1}{\sigma_i^2}$: because of the definition of u_i)

$\Rightarrow u_i$ is an eigenvector of AA^T

Therefore, the following happens:

$$[\sigma_1 u_1 \quad \sigma_2 u_2 \dots \quad \sigma_m u_m] = AV$$

and by defining $S \triangleq \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_m \\ - & - & - & - \end{bmatrix}$, one can write:

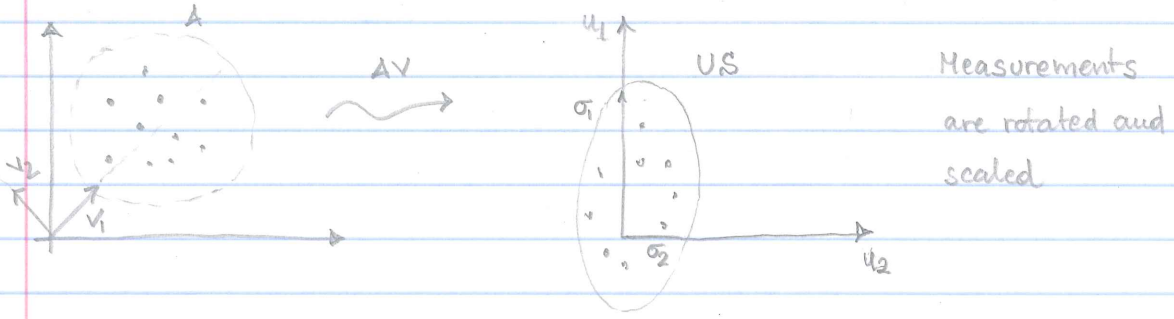
diagonal matrix ($m \times m$)

$US = AV \Rightarrow A = USV^T$ - Singular Value Decomposition of A

Note the construction process $\Rightarrow$ Decomposition is not unique

$\Rightarrow$ V is a set of orthonormal eigenvectors of A^TA (It always exists because A^TA is symmetric and with real entries), i.e.: $V^TV = I_m$

~~Example~~:



Measurements are rotated and scaled

Note: AA^T is an $n \times n$ matrix $\Rightarrow$ It may have n eigenvectors

Therefore, a typical SVD computes:

U: $n \times n$

V: $m \times m$

$S = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \sigma_m \\ - & - & - & - \\ & & \underline{0} & \end{bmatrix}$: $n \times m$ with only up to m nonnegative singular values

Let us return to the case of A being a matrix of measurements:

$A = U S V^T$

U is $n \times n \Rightarrow$ It is a collection of n temporal "modes"

V is $m \times m \Rightarrow$ It is a collection of m spatial "modes"

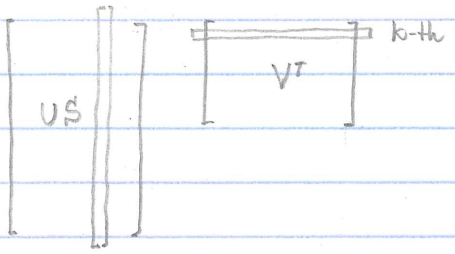
$n \times m$

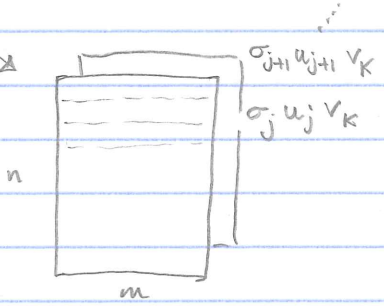
- n: temporal information
- m: spatial information

US is $n \times m \Rightarrow$ Only m temporal modes are retained and each one is weighted by the correspondent singular value

It is called "matrix of scores"

According to this interpretation, we are saying that the k-th right singular vector contributes to the measurements collected by every sensor but, at each time point, the contribution is different across temporal modes





Note: we are not adding or removing information. We are simply rearranging information via linear operations

What is the advantage of doing this?

Data compression: Since σ_i are ordered from the largest to the smallest, one can approximate A with the summation of just a few elements in the stack of matrices $\sigma_j u_j v_k$

A criterion to stop: Choose h such that $\sum_{i=1}^{h} \sigma_i \Big/ \sum_{i=1}^{m} \sigma_i$ is above a

given threshold and use the first h singular vectors $\Rightarrow$ That ratio indicates the probability of the first h components accounting for the energy (or variance) of A

Data separation: Different spatial modes may correspond to different players (e.g., sources) $\Rightarrow$ Data points mainly contributed by one mode are likely related to that source

Noise cancellation: The first singular value (if mean is not removed from the data) and the smallest singular values may be associated with non-physiologically-meaningfull modes $\Rightarrow$ Removal allows to increase the SNR of the data points of interest

□