# Contents

# List of Figures

Chapter 1

# Introduction

Felix's experiments consist these following parameters: coherence, number of dots, lifetime, and switch rate. The total dots are between 100 and 200 dots. Coherence is the percentage of signal dots and it ranges between 0.1 and 0.8. When coherence is 0.1, only 10% of dots move in the designated direction. Dots' lifetime is number of frames that dots last until disappear and be replaced by new dots. For example, this lifetime is 12 frames in Felix's experiment. If there are 120 dots, the dots will be divided into 10 sets with 12 dots per set. The first set will be replaced at frame 1, 13, 25, etc. The second set will be replaced at frame 2, 14, 26, and so on. The switch rate is the duration of stimulus before the switch. This switch rate lasts about 1 to 2.5 seconds.

In experiments, each monitor has its own picture rate. This can be used to determine the dots' lifetime. For instance, if a picture rate is 60 Hz, then each frame lasts 0.01667 s ($= 1/60$). If the dots only last for 200 ms, then the lifetime is $200/16.67 = 12$ frames.

Chapter 2

# Leaky Integrator in 2 dimensions

## 2.1 Translate experimental variables into leaky integrator parameters

In random dot motion discrimination task (RDM), we have 2 variables: number of dots and coherence. Signal dots move to veridical direction while random dots move in random direction per frame.

### 2.1.1 Compute pdf of functions of cosine and sine functions

First, we compute the pdf of $\cos(X)$ and $\sin(X)$ where random variable $X$ is uniform on $[-\pi, \pi]$.

Let $W = |X|$, then $W$ is uniform on $[0, \pi]$.

Let $Y = \cos(W)$.

We can compute the CDF of $Y$:

$$\begin{aligned}
F_Y(y) &= P(Y \leq y) \\
&= P(\cos(W) \leq y) \\
&= P(W \geq \cos^{-1}(y)) \quad \text{(note that smaller angles have greater cosine values)} \\
&= P(W \leq \cos^{-1}(-y)) \\
&= \frac{\cos^{-1}(-y)}{\pi} \quad \text{(because } W \text{ is uniform on } [0, \pi])
\end{aligned}$$

Thus, the pdf of $Y$ is:

$$\begin{aligned}
f_Y(y) &= \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) \\
&= \frac{1}{\pi} \frac{\mathrm{d}}{\mathrm{d}y} \cos^{-1}(-y) \\
&= \frac{1}{\pi} \frac{1}{\sqrt{1-y^2}} \quad \text{for } -1 < y < 1
\end{aligned}$$

Similarly, the cdf and pdf of $Y_2 = \sin(X)$ are $F_{Y_2}(y_2) = \frac{\sin^{-1}(y_2)}{\pi}$ and $f_{Y_2}(y_2) = \frac{1}{\pi} \frac{1}{\sqrt{1-y_2^2}}$.

### 2.1.2 Expected value and variance of resultant vector's norm

Signal dots move to "correct" direction while random dots move in random direction per frame. There are two ways we can define the stimulus: stimulus as a resultant vector of all dots (green vector in Fig. 2.1.1), or stimulus as a normalized resultant vector (green vector divided by total of dots). We want to map the parameters from experiment into parameters of leaky-integrator model.

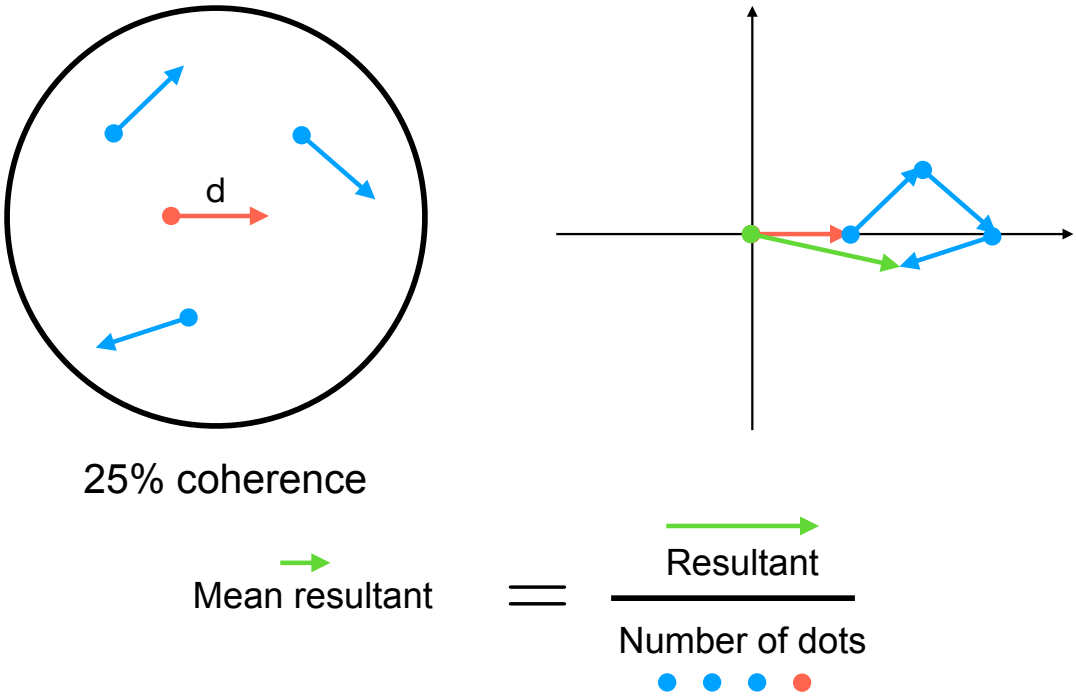


Fig. 2.1.1: Here, the number of dots $n = 4$ and coherence $= 0.25$. The blue dots are noise dots, and their directions are random. The red dot is the signal dot that points at the stimulus direction. Each vector has length $d$. The normalized resultant (or the mean resultant) is the resultant divided by the number of dots.

Let $R_X$ be the x-component of resultant vector and $r$ be norm of each individual vector. We have the following variables:

$$n_s = \text{coherence} \cdot n$$

$$n_r = 1 - n_s$$

where $n_s$ is number of signal dots, $n_r$ is number of noise dots, and $n$ is the total number of dots. Then:

$$R_X = n_s \cdot x_s + \sum_{i=1}^{n_r} x_{r,i}$$

where $x_s$ is the x-component of signal dot and $x_{r,i}$ is the x-component of $i$-th random dot. The expected norm $< R_X >$ is:

$$E[R_X] = n_s \cdot d\cos(a_s) + \sum_{i=1}^{n_r} E[x_{r,i}]$$

where $a_s$ is the direction of the stimulus, $x_{r,i} = cos(\alpha_{r,i})$ and RV $\alpha_{r,i}$ is uniform on $[-\pi, \pi]$. Thus $x_{r,i}$'s pdf is $f_{x_{r,i}} = \frac{1}{\pi\sqrt{1-x^2}}$.

The expected x-component of noise dots is:

$$E[x_r] = \int_{-1}^{1} d \cdot x \cdot \frac{1}{\pi}\frac{1}{\sqrt{1-x^2}}\mathrm{d}x = 0.$$

Since the pdf's of cosine and sine are the same, the expected y-component of noise dot is also $E[y_r] = 0$. Therefore, $E[R_X] = n_s \cdot d\cos(a_s)$ and $E[R_Y] = n_s \cdot d\sin(a_s)$.

On the other hand, the random vectors are independent and variance of a constant is zero, so the variance is:

$$\begin{aligned}
&\mathrm{Var}[R_X] = E[R_X^2] - E[R_X]^2 \\
=&E\left[\left(n_s \cdot d\cos(a_s) + \sum_{i=1}^{n_r} x_r\right)^2\right] - \left(n_s \cdot d\cos(a_s)\right)^2 \\
=&\left(n_s \cdot d\cos(a_s)\right)^2 + E\left[\sum_{i=1}^{n_r} x_{r,i}^2 + \sum_{i\neq j}^{n_r} x_{r,i}x_{r,j}\right] - \left(n_s \cdot d\cos(a_s)\right)^2 \\
=&n_r\left(E\left[x_r^2\right] + E\left[x_r\right]E\left[x_r\right]\right) \\
=&n_r \int_{-1}^{1} d^2 x_r^2 \frac{1}{\pi}\frac{1}{\sqrt{1-x_r^2}}\mathrm{d}x_r \\
=&n_r \cdot \frac{1}{\pi}\frac{\pi}{2} \\
&= \frac{1}{2}d^2 \cdot n_r
\end{aligned}$$

Similarly, the variance of y-component is $\mathrm{Var}[R_Y] = \mathrm{Var}[R_X] = \frac{1}{2}d^2 \cdot n_r$.

In the case of normalized resultant vector, the expected values of x- and y-component are:

$$E[R_X] = \text{coherence} \cdot d\cos(a_s)$$

$$E[R_Y] = \text{coherence} \cdot d\sin(a_s)$$

and the variance is:

$$\text{Var}[R_X] = \text{Var}[R_Y] = \frac{1}{2}d^2 \cdot \frac{1 - \text{coherence}}{n}$$

For simplicity, from now on we set $d = 1$. The parameters for leaky integrators are:

$$\boldsymbol{\mu} = \text{coherence} \cdot \begin{bmatrix} \cos(a_s) \\ \sin(a_s) \end{bmatrix} \tag{2.1.1}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0 & 0.5 \cdot \frac{1-\text{coherence}}{n} \\ 0.5 \cdot \frac{1-\text{coherence}}{n} & 0 \end{bmatrix} \tag{2.1.2}$$

The mean of x- and y-components, $\mu_X$ and $\mu_Y$, remain constant as number of dots increases (see Fig. 2.1.2) while the variance decreases. We simulated the random dots motion (red dots in Fig. 2.1.2). The simulation fits well with our formulae.

In general, the stimulus at time $t$ is simulated as:

$$X_t \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{2.1.3}$$

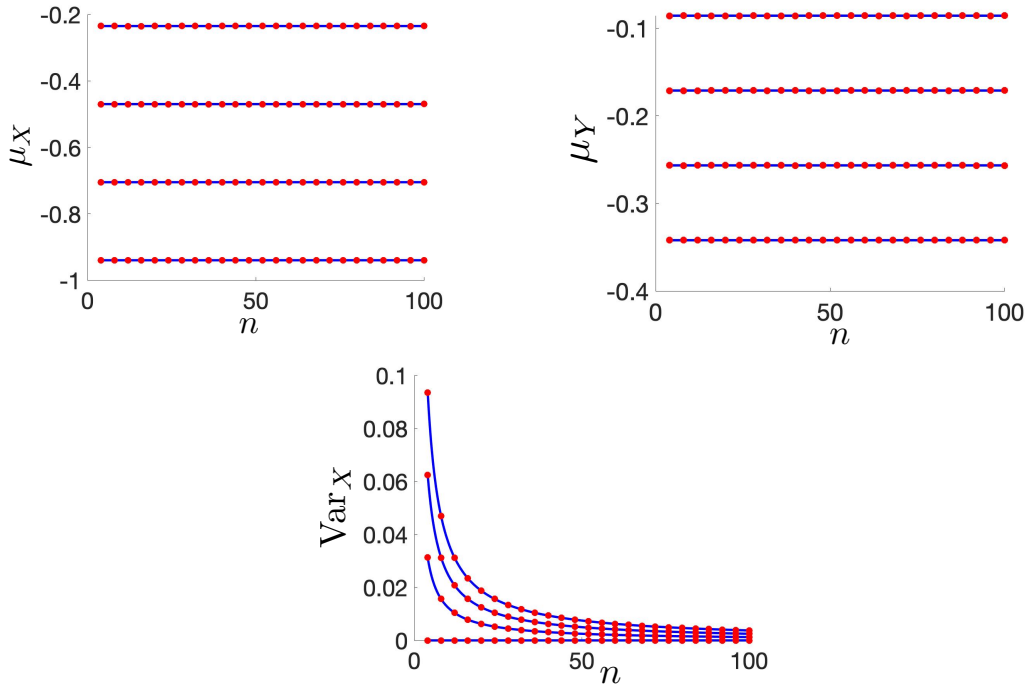


Fig. 2.1.2: Solid line is the theoretical result while red dots are simulation. Here, the angle is 200 degrees. From bottom to top, coherence is $1, 0.75, 0.5, 0.25$.

### 2.1.3 Modeling RDP with lifetime

Suppose each dot has lifetime of 25 frames, i.e. the dot moves to one direction for 25 frames before disappearing and being replaced by a new dot. The dots will be distributed uniformly to 25 groups. Then, dots in group 1 will be replaced in frame 1, dots in group 2 will be replaced in frame 2, and so on. This is more complicated as the evidence between time frame is no longer independent.

Let $X_t$ be the total x-component of noise dots at time $t$.

$$X_t = \sum_{<n_r/25>} x_{\text{noise, g1}} + \sum_{<n_r/25>} x_{\text{noise, g2}} + \cdots + \sum_{<n_r/25>} x_{\text{noise, g25}}$$

$$X_{t+1} = \sum_{<n_r/25>} x_{\text{noise, g1}} + \sum_{<n_r/25>} y_{\text{noise, g2}} + \cdots + \sum_{<n_r/25>} x_{\text{noise, g25}}$$

Here, at time $t+1$, only dots in group 2 are replaced by new dots while other groups continue following their previous direction.

$$X_{t+1} - X_t = \sum_{<n_r/25>} y_{\text{noise, g2}} - \sum_{<n_r/25>} x_{\text{noise, g2}}$$

The expected value of the difference is:

$$E(X_{t+1} - X_t) = E\left( \sum_{<n_r/25>} y_{\text{noise, g2}} - \sum_{<n_r/25>} x_{\text{noise, g2}} \right) = 0$$

since the noise dots are independent.

Moreover,

$$
\begin{aligned}
E[(X_{t+1} - X_t)^2] &= E\left[\left(\sum_{<n_r/25>} y_{\text{noise, g2}} - \sum_{<n_r/25>} x_{\text{noise, g2}}\right)^2\right] \\
&= E\left[\left(\sum_{<n_r/25>} y_{\text{noise, g2}}\right)^2 + \left(\sum_{<n_r/25>} x_{\text{noise, g2}}\right)^2\right. \\
&\left. - 2\sum_{<n_r/25>} y_{\text{noise, g2}} \sum_{<n_r/25>} x_{\text{noise, g2}}\right] \\
&= E\left[\sum_{<n_r/25>} (y_{\text{noise, g2}})^2 + \sum_{i\neq j}^{<n_r/25>} (y_{\text{noise, i, g2}})(y_{\text{noise, j, g2}})\right. \\
&\left. + \sum_{<n_r/25>} (x_{\text{noise, g2}})^2 + \sum_{i\neq j}^{<n_r/25>} (x_{\text{noise, i, g2}})(x_{\text{noise, j, g2}})\right] \\
&= E\left[\sum_{<n_r/25>} (y_{\text{noise, g2}})^2\right] + E\left[\sum_{<n_r/25>} (x_{\text{noise, g2}})^2\right] \\
&= \frac{n_r}{25} \cdot d^2 \\
&= \frac{1-\text{coh}}{25n} \cdot d^2
\end{aligned}
$$

In general, we have

$$X_{t+1} = X_t + \eta_t \tag{2.1.4}$$

where $\eta_t \sim \mathcal{N}\left(0, \frac{1-\text{coh}}{n_{\text{life}} \cdot n} \cdot d^2\right)$, $n_{\text{life}}$ is lifetime of dots, $n$ is number of dots, and $d$ is the distance that dot travels between frame.

### 2.1.4 Stimulus data

There are 2 types of stimulus. One was generated by either Formula 2.1.3 or 2.1.3, we call it simulated stimulus. The second type comes directly from data, in which we call experimental stimulus. In the CPR experiment, the positions $(x, y)$ for each dot are recorded per frame. There are 503 dots.

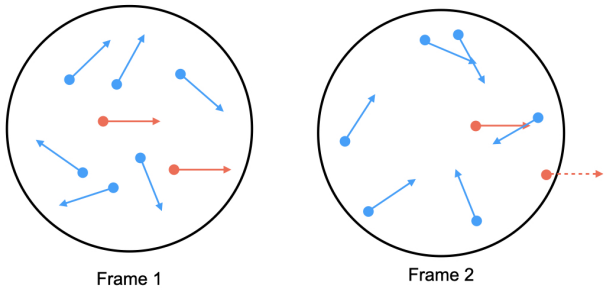


Fig. 2.1.3: Solid line is the theoretical result while red dots are simulation. Here, the angle is 200 degrees. From bottom to top, coherence is $1, 0.75, 0.5, 0.25$.

We assume that the resultant vector on each frame is the evidence that subject observes. The length of resultant vector is 0.6667, which is the distance that dot travels between frame while the direction of resultant vector represents the experiment stimulus that the subject see on the screen. We took all trials from NaK's first session and compute the difference between our experiment stimulus and veridical direction. We plotted the mean and standard deviation of this difference (see Fig. 2.1.4). The mean is 0 except coherence 0.05. The standard deviation shrinks as the coherence increases.

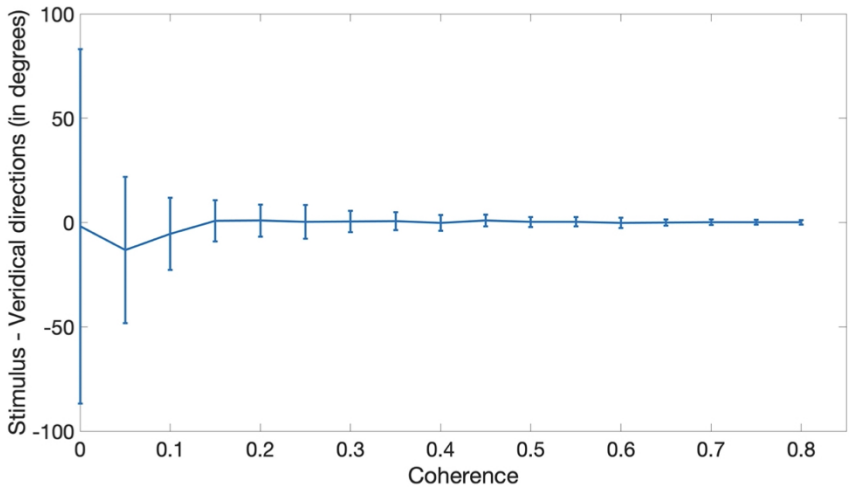


Fig. 2.1.4: Mean and standard deviation of the stimulus deviation from veridical direction (in degrees). The data is taken from NaK session 1.

### 2.1.5 Leaky integrator in 2 dimensions

Consider the direction of stimulus $a_s$, then the stimulus signal is $\boldsymbol{X}_{\text{stim}} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where

$$\boldsymbol{\mu} = \text{coherence} \cdot \begin{bmatrix} \cos(a_s) \\ \sin(a_s) \end{bmatrix}$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0 & 0.5 \cdot \frac{1-\text{coherence}}{n} \\ 0.5 \cdot \frac{1-\text{coherence}}{n} & 0 \end{bmatrix}.$$

Let $\boldsymbol{X}_{\text{leak}}(t)$ be the "belief" of subject on direction of the stimulus. The belief at

time $t$ is:

$$\boldsymbol{X}_{\text{leak}}(0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{2.1.5}$$

$$\begin{aligned} \boldsymbol{X}_{\text{leak}}(t) &= \boldsymbol{X}_{\text{leak}}(t-1) + \boldsymbol{X}_{\text{stim}}(t) \cdot \mathrm{d}t - \lambda \cdot \mathrm{d}t \cdot \boldsymbol{X}_{\text{leak}}(t-1) \\ &= (1 - \lambda \cdot \mathrm{d}t)\boldsymbol{X}_{\text{leak}}(t-1) + \boldsymbol{X}_{\text{signal}}(t) \cdot \mathrm{d}t \end{aligned} \tag{2.1.6}$$

where $\lambda$ (1/second) is leak rate (or discount rate). Old evidence is discounted more quickly when leak rate $\lambda$ is higher.

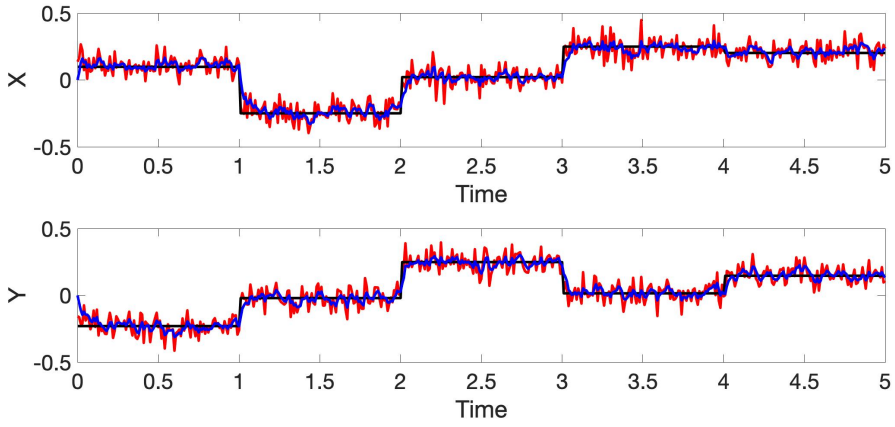


Fig. 2.1.5: Here, coherence $= 0.25$, switch rate $= 1$ second, number of dots is 100. Black color indicates the stimulus without noise; Red color indicates the stimulus with noise; and blue curves are the outputs of leaky integrator. The leaky constant $\lambda = 34.6325$ is the optimal value, which will be discussed in the next section.

### 2.1.6 Minimize mean-squared error

First, note the following inequality. Suppose we have $n$ vectors, $x_1, x_2, \cdots x_n$, then:

$$\| \sum_{i=1}^{n} x_i \| \leq \sum_{i=1}^{n} \|x_i\|. \tag{2.1.7}$$

The mean squared error (MSE) is defined as:

$$\text{MSE} = \frac{1}{2n} \sum_{i}^{n} \|\boldsymbol{\mu}_i - \lambda \boldsymbol{X}_{\text{leak},i}\|^2 \tag{2.1.8}$$

where $\boldsymbol{\mu}_i$ is the stimulus without noise as we want $\boldsymbol{X}_{\text{leak}}$ to be as close as the "true" signal.

The optimal lambda increases as coherence increases. This makes sense because high coherence means the momentary evidence is strong that it's okay to discount all of past evidence. Lower switch rate also yields higher optimal lambda.

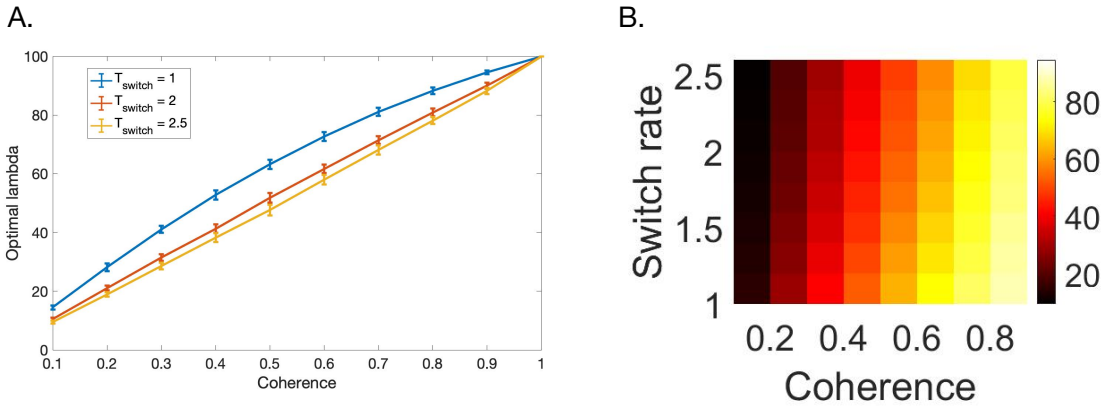


Fig. 2.1.6: **A.** The mean and error of optimal lambda as a function of coherence for 3 different switch rate: $T_{\text{switch}} = 1$ (blue), $T_{\text{switch}} = 2$ (red), and $T_{\text{switch}} = 2.5$ (yellow). **B.** Color map of optimal lambda as a function of switch rate and coherence.

### 2.1.7 Accuracy

We define accuracy as:

$$\text{Accuracy} = \left| 1 - \frac{|a_s - a_{\text{leak}}|}{\pi} \right| \tag{2.1.9}$$

where $a_s$ is the "true" direction of signal dot while $a_{\text{leak}} = \arctan\left(\frac{X_{\text{leak}}(2)}{X_{\text{leak}}(1)}\right)$.

Note that the difference between two angles is a bit tricky as its range is between 0 and $\pi$. Hence, if $a_s$ and $a_{\text{leak}}$ are the same, the accuracy is 1. If two angles are completely opposite, the accuracy is 0.

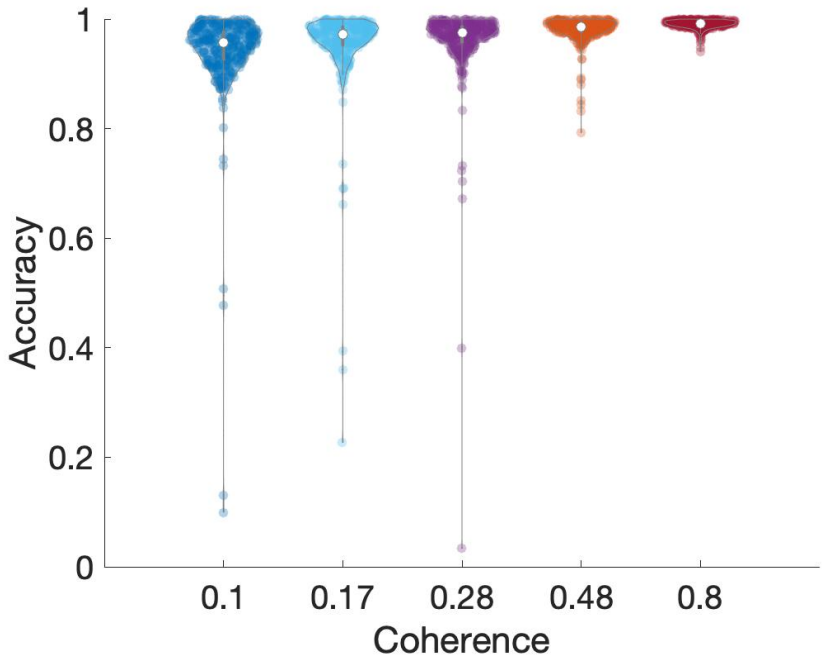


Fig. 2.1.7: Here, coherence = 0.25, switch rate = 1 second, number of dots is 100. Black color indicates the stimulus without noise; Red color indicates the stimulus with noise; and blue curves are the outputs of leaky integrator. The leaky constant $\lambda = 34.6325$ is the optimal value, which will be discussed in the next section.

## 2.2 Leaky integrator with delay

In the last section, we assume the evidence-accumulation is instantaneous. In reality, there is a delay from the moment the stimulus appears before subjects perceive that stimulus. This is called perceptual delay, and we assumed this $T_{\text{perc. del.}} = 100$ msec.

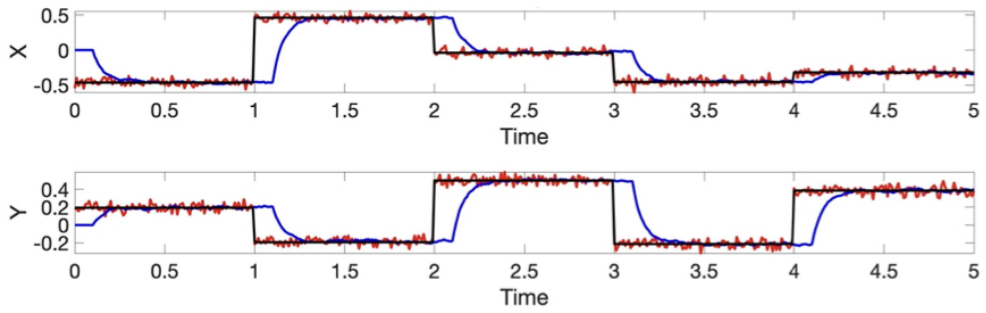


Fig. 2.2.1: Here, coherence $= 0.5$, switch rate $= 1$ second, number of dots is 160. Black color indicates the stimulus without noise; Red color indicates the stimulus with noise; and blue curves are the outputs of leaky integrator. The leaky constant $\lambda = 16$ which is close to the optimal value. The optimal lambda when there is delay is $\lambda_{\text{opt}} = 16.3198$.

We do the same analysis as the instanteneous evidence-accumulation. Interestingly, with the added delay, the optimal leaky constant decreases for the same pair of coherence level and switch rate (see Fig. 2.2.2). Even though the delay yields same patterns, i.e. higher coherence and shorter switch rate yield higher optimal leaky constant, the growth rate is different.

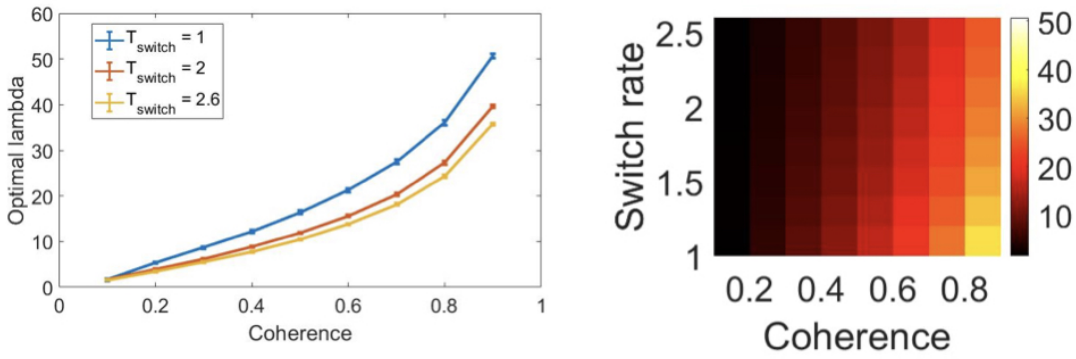


Fig. 2.2.2: **A.** The mean and error of optimal lambda as a function of coherence for 3 different switch rate: $T_{\text{switch}} = 1$ (blue), $T_{\text{switch}} = 2$ (red), and $T_{\text{switch}} = 2.5$ (yellow). **B.** Color map of optimal lambda as a function of switch rate and coherence.

## 2.3 Simulated vs. Experimental stimulus

**MATLAB: scprm_leaky_vs_data.m**

One question we need to address: Does our "real" stimulus improve the joystick direction prediction? To answer this question, we fit the model using our simulated stimulus (2.1.3) and computed the optimal leaky constant $\lambda$ and delay time $T_D$. We then replaced the stimulus input and fit the leaky integrator model again with the same optimal constants. The better model will yield smaller CMSE and AIC.

We selected two blocks (block 6 and 46) in Sol's one session. The coherence level in block 46 is 0.8. From the simulated stimulus, we found the optimal constants as $\lambda = 39.9371$ and $T_D = 0.32923$ with CMSE $= 0.0273$ and AIC $= -2.6289e4$ (see Fig. 2.3.1). When we substituted the input, the CMSE $= 0.0288$ and AIC $= -2.6169e4$. For block 6, the coherence is 0.4. We repeated the same

steps. For this block, we found optimal constants are $\lambda = 30.8665$ and $T_D = 0.40675$. The CMSE for the experimental stimulus and simulated stimulus are 0.0662 abd 0.0701, respectively. Additionally, the AIC for experimental stimulus and simulated stimulus are -2.1102e4 and -2.0984e4, respectively. In both cases, the experimental stimulus yields smaller CMSE and AIC, i.e. the experiment stimuli produces better fits.

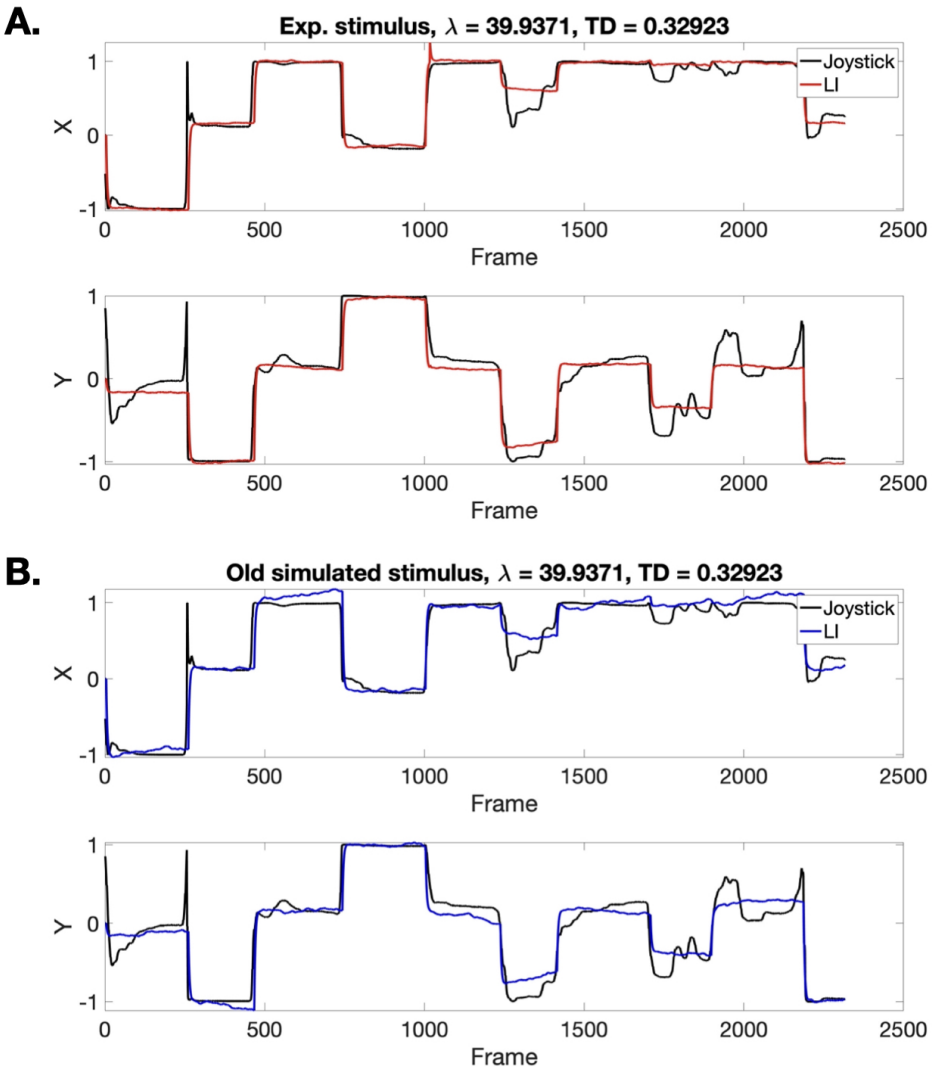


Fig. 2.3.1: The joystick direction and leaky integrator using **A.** stimulus from data and **B.** generated stimulus. We used block 46, subject Sol, and the coherence level was 0.8.

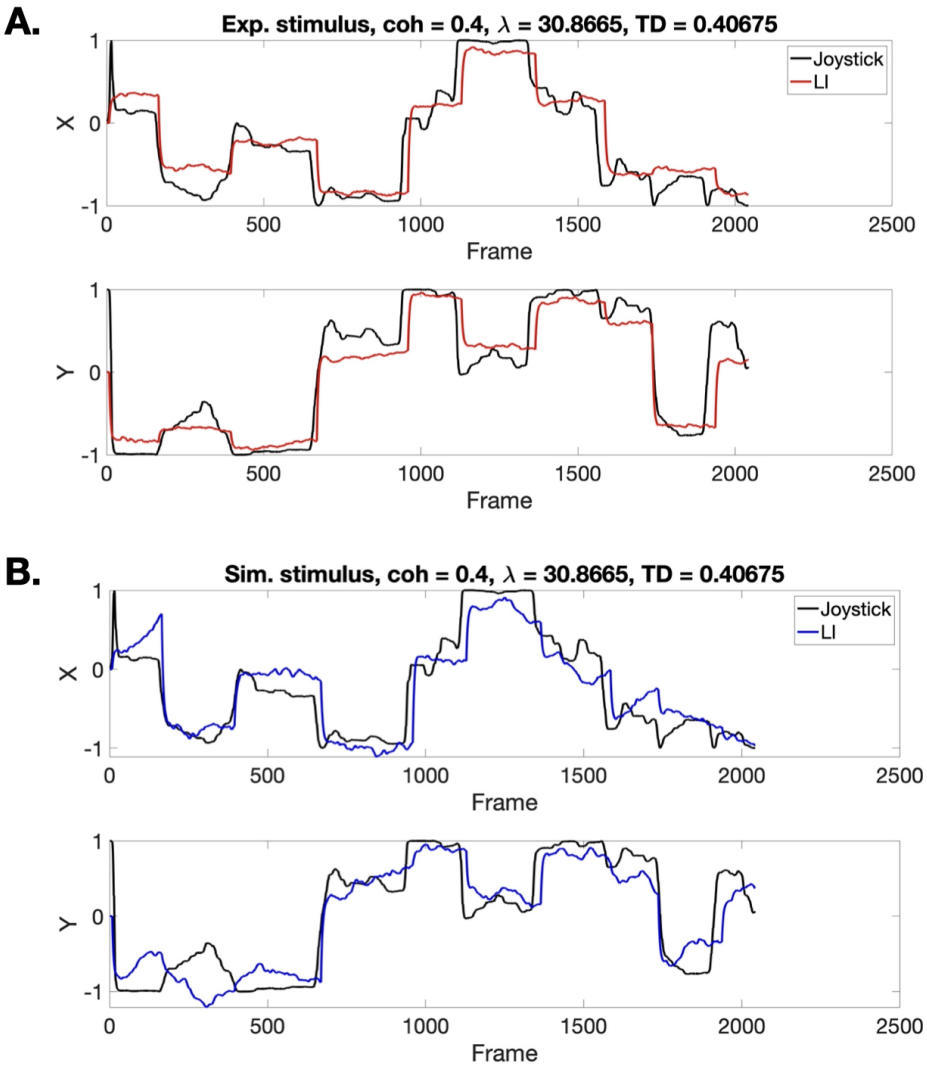


Fig. 2.3.2: The joystick direction and leaky integrator using **A.** stimulus from data and **B.** generated stimulus. We used block 6, subject Sol, and the coherence level was 0.4.

Chapter 3

# Modeling confidence

## 3.1 Confidence models in Fleming's paper

### 3.1.1 Introduction to 3 confidence models

There are 3 different confidence models in Fleming's paper (see Fig. 3.1.1). Each model consists 2 ingredients: categorical world state $d$ and response $a$. In two-alternative forced choice (2AFC) tasks, particularly RDM, the categorical state $d = 1$ (or $d = -1$) if stimulus moves right (or left). The action $a = 1$ (or $a = -1$) indicates subject's choice. The subject makes a correct choice if $a = d$. The confidence is defined as a degree of belief that a particular choice is correct given particular set of internal states $X$, model $m$ and parameter $\nu$:

$$z = P(a = d|X, a, \nu, m) \tag{3.1.1}$$

where internal states $X$ are generated from world states $d$ with Gaussian noise. On each "trial", internal states $X = [X_{\text{act}} X_{\text{conf}}]$ denote decision and confidence

variables.

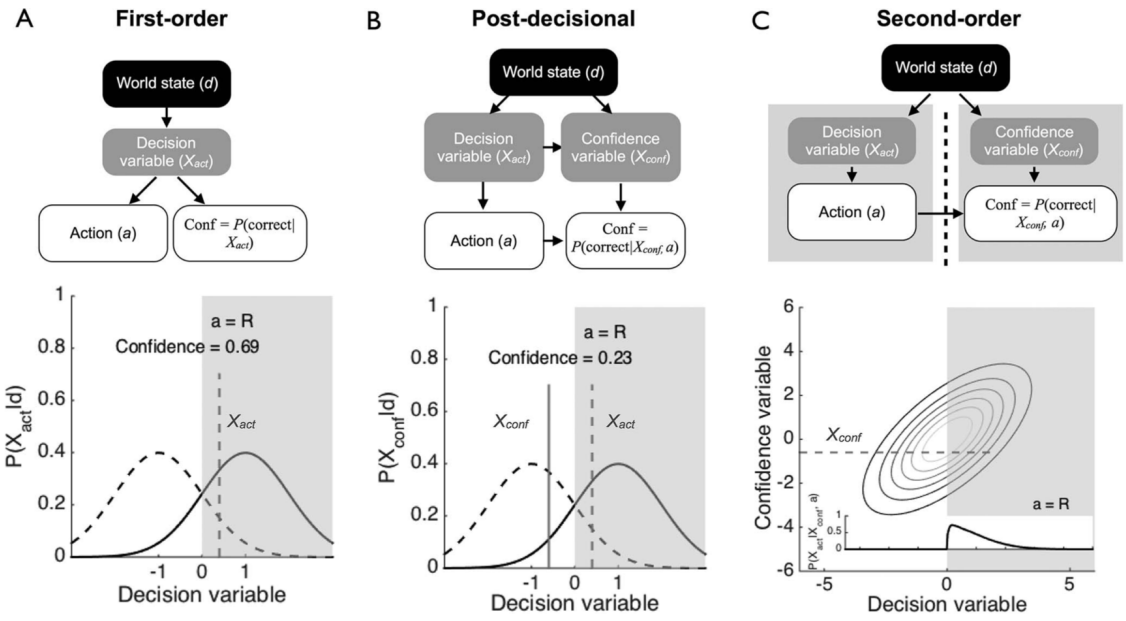


Fig. 3.1.1: Three models from Fleming (Fleming 2017): **A.** First-order model, **B.** Post-decisional model, and **C.** Second-order model.

In the "first-order" model, the decision and confidence variables are identical, i.e. same internal state supports both choices and confidence. Let $X_{\text{act}} \sim \mathcal{N}(d, \sigma^2)$. The confidence variable $X_{\text{conf}} = X_{\text{act}}$.

The confidence is defined as probability of correct decision conditioned on decision variable $X_{\text{act}}$:

$$
\begin{aligned}
z &= P(a = d|X_{\text{conf}}, \sigma^2) \\
&= P(a|X_{\text{conf}}, \sigma^2)P(d|X_{\text{conf}}, \sigma^2) \\
&= P(a|X_{\text{conf}})P(d|X_{\text{conf}}, \sigma^2) \\
&= P(a = 1|X_{\text{conf}})P(d = 1|X_{\text{conf}}, \sigma^2) + P(a = -1|X_{\text{act}})P(d = -1|X_{\text{act}}, \sigma^2)
\end{aligned}
$$

where $P(a = 1|X_{\text{conf}}) = 1$ if $X_{\text{conf}} > 0$ $P(a = 1|X_{\text{conf}}) = 0$ if $X_{\text{conf}} < 0$.

The confidence is then:

$$z = P(a = d|X_{\text{conf}}, a, \sigma^2) = \begin{cases} P(d = 1|X_{\text{conf}}, \sigma^2) & \text{if } a = 1 \\ 1 - P(d = 1|X_{\text{conf}}, \sigma^2) & \text{if } a = -1 \end{cases} \tag{3.1.2}$$

We can compute $P(d|X_{\text{conf}}, \sigma^2)$ by using Bayes' rule:

$$\begin{aligned} P(d|X_{\text{conf}}, \sigma^2) &= \frac{P(d, X_{\text{conf}}, \sigma^2)}{P(X_{\text{conf}}, \sigma^2)} \\ &= \frac{P(X_{\text{conf}}|d, \sigma^2)P(d, \sigma^2)}{P(X_{\text{conf}}, \sigma^2|d = 1)P(d = 1) + P(X_{\text{conf}}, \sigma^2|d = -1)P(d = -1)} \\ &= \frac{P(X_{\text{conf}}|d, \sigma^2)P(d, \sigma^2)}{\sum_d P(X_{\text{conf}}|d, \sigma^2)P(d, \sigma^2)} \\ &= \frac{P(X_{\text{conf}}|d, \sigma^2)}{\sum_d P(X_{\text{conf}}|d, \sigma^2)} \quad \text{since } P(d = 1, \sigma^2) = P(d = -1, \sigma^2). \end{aligned}$$

This is also equation (5) in Fleming 2017.

The post-decisional model is similar to "first-order" model. The confidence variable $X_{\text{conf}}$ is derived from $X_{\text{act}}$ plus the additional information about the world state, $X_{\text{new}}$:

$$X_{\text{conf}} = X_{\text{act}} + X_{\text{new}}$$

where $X_{\text{new}} \sim \mathcal{N}(d, \sigma^2)$ is an additional sample of evidence. The variance of $X_{\text{conf}}$ is $\text{Var}(X_{\text{conf}}) = \text{Var}(X_{\text{act}}) + \text{Var}(X_{\text{new}}) = 2\sigma^2$. Thus, the confidence of the observer is:

$$z = P(a = d|X_{\text{conf}}, a, 2\sigma^2) = \begin{cases} P(d = 1|X_{\text{conf}}, 2\sigma^2) & \text{if } a = 1 \\ 1 - P(d = 1|X_{\text{conf}}, 2\sigma^2) & \text{if } a = -1 \end{cases} \tag{3.1.3}$$

where

$$P(d = 1|X_{\text{conf}}, 2\sigma^2) = \frac{P(X_{\text{conf}}|d, 2\sigma^2)}{\sum_d P(X_{\text{conf}}|d, 2\sigma^2)}.$$

The "second-order" model considers two individuals: Actor (act) and Confidence-rater (conf). The actor carries out a two-choice discrimination task. Both Actor and Confidence-rater receive internal samples $X_{\text{act}}$ and $X_{\text{conf}}$ generated from binary world state $d$. These samples are drawn from bivariate Gaussian with covariance matrix $\mathbf{\Sigma}$:

$$\begin{bmatrix} X_{\text{act}} \\ X_{\text{conf}} \end{bmatrix} \sim \mathcal{N}(\boldsymbol{d}, \mathbf{\Sigma})$$

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma^2_{\text{act}} & \rho\sigma_{\text{act}}\sigma_{\text{conf}} \\ \rho\sigma_{\text{act}}\sigma_{\text{conf}} & \sigma^2_{\text{conf}} X_{\text{conf}} \end{bmatrix}$$

where $\sigma_{\text{act}}$ and $\sigma_{\text{conf}}$ control the noise of the signal for the Actor and Confidence-rater, respectively; $\rho$ is correlation parameter that governs the association between two samples. Confidence-rater determines how confidence they are in the Actor responding correctly, or the posterior probability that the Actor's action $a$ was appropriate for the inferred state of the world $d$, conditional on beliefs about different sources of variability. The observer infers the state of decision variable $X_{\text{act}}$ from the confidence variable $X_{\text{conf}}$. The confidence is:

$$z = P(a = d|X_{\text{conf}}, a, \mathbf{\Sigma}) = \begin{cases} P(d = 1|X_{\text{conf}}, a, \mathbf{\Sigma}) & \text{if } a = 1 \\ 1 - P(d = 1|X_{\text{conf}}, a, \mathbf{\Sigma}) & \text{if } a = -1 \end{cases} \tag{3.1.4}$$

where

$$P(d|X_{\text{conf}}, a, \boldsymbol{\Sigma}) \propto P(d|X_{\text{conf}}, \boldsymbol{\Sigma})P(a|X_{\text{conf}}, d, \Sigma)$$
$$= P(d|X_{\text{conf}}, \boldsymbol{\Sigma}) \int P(a|X_{\text{act}}, \boldsymbol{\Sigma})P(X_{\text{act}}|X_{\text{conf}}, d, \boldsymbol{\Sigma} \mathrm{d}X_{\text{act}})$$

### 3.1.2 First-order and post-decisional models simulation

The confidence is simulated as a function of stimulus stength $\theta$, i.e. coherence, which varies between 0 and 1. The decision variable $X_{\text{act}} \sim \mathcal{N}(\mu, \sigma^2)$ is drawn from normal distribution with mean $\mu = d \cdot \theta$. A given direction $d$ and stimulus strength $\theta$ leads to a range of sample $X_{\text{act}}$. If the subject's sample is $X_{\text{act}} = +0.05$ while $d = -1$, subject will errorneously respond "right". This sample may have arisen from many different objective stimulus strength $\theta$, including both correct and error trials, and occur more often with some than others. As $\theta$ increases, the likely values of $X_{\text{conf}}$ following an incorrect response therefore decrease in magnitude.

To produce Fig. 3.1.2 I simulated 10,000 trials at each of 7 levels of stimulus strength $\theta = [0, 0.032, 0.064, 0.128, 0.256, 0.512, 1.0]$ and $\sigma = 1$.

**Step 1:** On each trial, choose a world state, i.e. $d = \pm 1$ randomly.

**Step 2:** For each coherence level, define a decision variable $X_{\text{act}} = \text{coherence} \cdot d + \text{randn} \cdot \sigma$.

**Step 3:** Define a confidence variable $X_{\text{conf}}$. If it's first-order model, $X_{\text{conf}} = 0$. If it's post-decisional model, set $X_{\text{conf}} = \text{coherence} \cdot d + \text{randn} \cdot \sigma$. Note that $X_{\text{act}}$ and $X_{\text{conf}}$ do not need to have the same values.

**Step 4:** Depending on response $a$, the confidence of first-model is computed as:

$$\text{Confidence} = \frac{\text{normpdf}(X_{\text{conf}}, a, \sigma)}{\text{normpdf}(X_{\text{conf}}, a, \sigma) + \text{normpdf}(X_{\text{conf}}, -a, \sigma)}$$

The confidence of post-decisional model is:

$$\text{Confidence} = \frac{\text{normpdf}(X_{\text{conf}}, a, \sqrt{2}\sigma)}{\text{normpdf}(X_{\text{conf}}, a, \sqrt{2}\sigma) + \text{normpdf}(X_{\text{conf}}, -a, \sqrt{2}\sigma)}$$

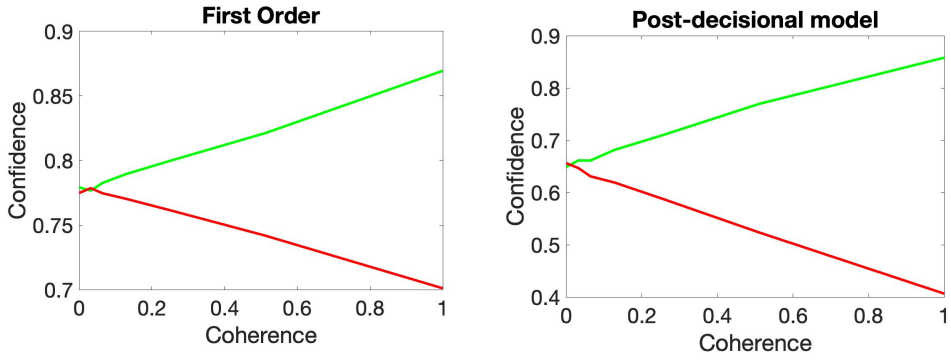


Fig. 3.1.2: Simulation of first-order model (left) and post-decisional model (right).

In a between-subjects design, participants were assigned to one of three conditions:

- Deciding if a target word was an anagram and then judging confidence.
- Judging confidence after seeing the target but before making a decision.
- Rating confidence before seeing the target word.

## 3.2 Implement post-decisional model in 2 dimensions

In our model, the response $a$ is not simply 1 or -1. Here, $a$ can be anything between 0 and $2\pi$. The $\boldsymbol{X}_{\text{act}} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a 2-dimensional decision variable. Note that $\boldsymbol{\Sigma}$ is not the same as $\boldsymbol{\Sigma}$ in the second-order model.

Confidence is probability of correct response given by confidence variable $X_{\text{conf}}$ and response $a$.

$$z = P(\text{correct}|\boldsymbol{X}_{\text{conf}}, a, \boldsymbol{\Sigma})$$

$$P(\text{correct}) = P(a = d) = P(a = 1|d = 1)(?)$$

Assuming the response is $a = 1$, by the law of total probability, we have:

$$\begin{aligned}
P(\text{correct}) &= P(\text{correct}|\boldsymbol{X}_{\text{conf}}, a = 1)P(\boldsymbol{X}_{\text{conf}}, a = 1) \\
&+ P(\text{correct}|\boldsymbol{X}_{\text{conf}}, a = -1)P(\boldsymbol{X}_{\text{conf}}, a = -1) \\
&= P(\text{correct}|\boldsymbol{X}_{\text{conf}}, a = 1)P(\boldsymbol{X}_{\text{conf}} > 0) \\
&+ P(\text{correct}|\boldsymbol{X}_{\text{conf}}, a = -1)P(\boldsymbol{X}_{\text{conf}} < 0) \\
&= P(\text{correct}|\boldsymbol{X}_{\text{conf}}, a = 1)P(\boldsymbol{X}_{\text{conf}} > 0|d = 1, \sigma^2)P(d = 1, \sigma^2) \\
&+ P(\text{correct}|\boldsymbol{X}_{\text{conf}}, a = -1)P(\boldsymbol{X}_{\text{conf}} < 0)
\end{aligned}$$

I suggest the general form of confidence is:

$$z = P(\text{correct}) \cdot P(d|\boldsymbol{X}_{\text{conf}}, \Sigma^2) + (1 - P(\text{correct})) \cdot (1 - P(d|\boldsymbol{X}_{\text{conf}}, \Sigma^2))$$

## 3.3 My attempt to define "confidence"

Assume that the observer infers the direction of motion, and finds a posterior $p(\theta)$ at the time of decision. The guess should then be $\hat{\theta}$ which could be the posterior mean. How confident is the observer that they will get a reward ($R = 1$), or no reward ($R = 0$), equals the subjective probability of hitting the target given the

posterior distribution.

$$P(\text{hit target}|\text{guess } \hat{\theta}) = P(\hat{\theta} \in \text{target region}) \tag{3.3.1}$$

$$= \int_{\hat{\theta}-\epsilon}^{\hat{\theta}+\epsilon} P(\theta)\mathrm{d}\theta \tag{3.3.2}$$

Any true direction within $\epsilon$ of the estimate, $\hat{\theta}$, gives a reward (see Fig. 3.3.1).

- As target shrinks $\epsilon \to 0$, so does confidence.
- As posterior gets narrower confidence grows.

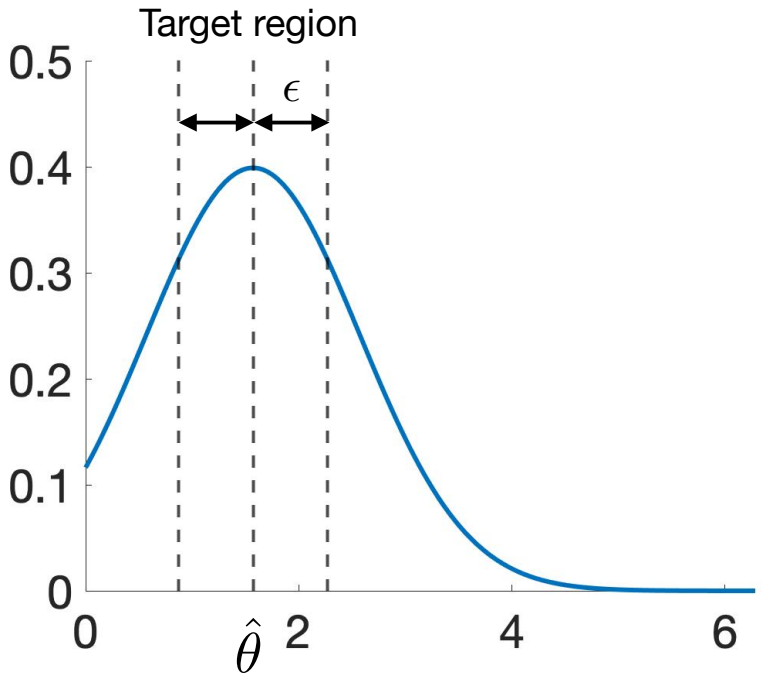


Fig. 3.3.1: This is posterior of decision where $\hat{\theta}$ is the estimate. If the decision is within the region between

Chapter 4

# Data fusion

## 4.1 Data fusion in Ernst & Banks 2002

When a person makes decision, sometimes they will have more than one type of information. For instance, both vision and touch can provide information for estimating the properties of the object. Vision frequently dominates the integrated visual-haptic percept, but in some cases, percept is affected by haptics. The authors proposed a general principle to minimize variance in final estimate and determines the degree to which vision or haptics dominates.

Let the estimate of an environmental property by sensory system be:

$$\hat{S}_i = f_i(S) \tag{4.1.1}$$

where $S$ is the physical property being estimated and $f$ is the operation by which the nervous system does the estimation, the subscript $i = \{H, V\}$ refer to the modality haptic $H$ and visual $V$. Each estimate, $\hat{S}_i$ is corrupted by independent

Gaussian noises with variance $\sigma_i^2$. The pdf's of haptic and visual estimation is:

$$f_V(S) = \frac{1}{\sigma_V\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{S-\hat{S}_V}{\sigma_V^2}\right)^2\right]$$
$$f_H(S) = \frac{1}{\sigma_H\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{S-\hat{S}_H}{\sigma_H^2}\right)^2\right]$$

.

The maximum-likelihood estimation (MLE) of the environmental property is:

$$\begin{aligned}
\text{MLE} &= f_V(S)\cdot f_H(S) \\
&= \frac{1}{2\pi\sigma_V\sigma_H} \exp\left[-\frac{1}{2}\left(\frac{S-\hat{S}_V}{\sigma_V^2}\right)^2 - \frac{1}{2}\left(\frac{S-\hat{S}_H}{\sigma_H^2}\right)^2\right] \\
\frac{\partial \text{MLE}}{\partial S} &= \exp\left[-\frac{1}{2}\left(\frac{S-\hat{S}_V}{\sigma_V^2}\right)^2 - \frac{1}{2}\left(\frac{S-\hat{S}_H}{\sigma_H^2}\right)^2\right]\left[\frac{S-\hat{S}_V}{\sigma_V^2} + \frac{S-\hat{S}_H}{\sigma_H^2}\right] = 0 \\
&\Longrightarrow \frac{S-\hat{S}_V}{\sigma_V^2} + \frac{S-\hat{S}_H}{\sigma_H^2} = 0 \\
&\Longrightarrow \left[\frac{1}{\sigma_V^2} + \frac{1}{\sigma_H^2}\right] S = \frac{1}{\sigma_V^2}\hat{S}_V + \frac{1}{\sigma_H^2}\hat{S}_H \\
&\Longrightarrow \hat{S} = \sum_i w_i \hat{S}_i \quad \text{with} \quad w_i = \frac{1/\sigma_i^2}{\sum_j 1/\sigma_j^2}
\end{aligned}$$

This is Eq. (2) in Ernst & Banks 2002.

The variance of combined haptics and vision is:

$$\begin{aligned}
\text{Var}(\hat{S}) &= \left(\frac{\sigma_V^2}{\sigma_V^2+\sigma_H^2}\right)^2 \text{Var}(\hat{S}_V) + \left(\frac{\sigma_H^2}{\sigma_V^2+\sigma_H^2}\right)^2 \text{Var}(\hat{S}_H) \\
\Longrightarrow\ \sigma_{VH}^2 &= \frac{\sigma_V^2\sigma_H^2}{\sigma_V^2+\sigma_H^2}
\end{aligned}$$

This is equation (3) in Ernst 2002. In visual-haptic tasks, the MLE integrator always uses information from both sensory systems, so the combined percept will

always reflect both sources of information. With large discrepancies between information sources, the nervous system may exhibit robust behaviour in which a discrepant source is discounted.

## 4.2 Data fusion in 2 dimensions

We can extend this method into 2 dimensions. The bivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has pdf:

$$f(\boldsymbol{X}) = \frac{1}{2\pi \left|\boldsymbol{\Sigma}\right|^{1/2}} \exp\left( -\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \right)$$

The log-likelihood function for 2 distribution is:

$$L = \log\left[ \frac{1}{2\pi \left|\boldsymbol{\Sigma_1}\right|^{1/2}} \exp\left( -\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu_1})^T {\boldsymbol{\Sigma_1}}^{-1} (\boldsymbol{X} - \boldsymbol{\mu_1}) \right) \right.$$
$$\left. \cdot \frac{1}{2\pi \left|\boldsymbol{\Sigma_2}\right|^{1/2}} \exp\left( -\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu_2})^T {\boldsymbol{\Sigma_2}}^{-1} (\boldsymbol{X} - \boldsymbol{\mu_2}) \right) \right]$$

For some vector $\boldsymbol{w}$, if $\boldsymbol{w}$ does not depend on $\boldsymbol{A}$ and $\boldsymbol{A}$ is symmetric, then:

$$\frac{\partial w^T A w}{\partial w} = 2Aw$$

Some linear algebra and calculus properties are:

- The trace is invariant under cyclic permutations of matrix products:

$$\mathrm{tr}[ABC] = \mathrm{tr}[CAB] = \mathrm{tr}[BCA]$$

- Since $x^T \boldsymbol{A} x$ is a scalar, we can take its trace and obtain the same value:

$$x^T \boldsymbol{A} x = \text{tr}[x^T \boldsymbol{A} x] = \text{tr}[x^T x \boldsymbol{A}].$$

- $\frac{\partial}{\partial A}\text{tr}[AB] = B^T$
- $\frac{\partial}{\partial A} \log |\boldsymbol{A}| = \boldsymbol{A}^{-T}$ .
- The deterinant of the inverse of an invertible matrix is the inverse of the determinant:

$$|\boldsymbol{A}| = \frac{1}{\left|\boldsymbol{A}^{-1}\right|}.$$

Combining these properties, we have:

$$\frac{\partial}{\partial \boldsymbol{A}} x^T A x = \frac{\partial}{\partial \boldsymbol{A}} \text{tr}[x^T x \boldsymbol{A}] = [x x^T]^T = (x^T)^T x^T = x x^T.$$

Chapter 5

# Standard DDM and mean decision time

The standard DDM can be derived as the continuum limit of a recursive equation for the log-likelihood ratio (LLR). When the environmental states are uncorrelated between trials and equally likely, an ideal observer has no bias at the start of each trial. We assume that on one trial the observer integrates a stream of noisy measurements of the true state, $H$ (see Fig. 5.0.1 A). If these measurements are conditionally independent, the functional central limit theorem yields the DDM for the scaled LLR, $y(t) = D \ln \frac{\mathrm{P}(H=H_+|\text{observations})}{\mathrm{P}(H=H_-|\text{observations})}$, after observation time $t$,

$$\mathrm{d}y = g\mathrm{d}t + \sqrt{2D}\mathrm{d}W. \tag{5.0.1}$$

We have written the above as a Langevin equation in which $W$ is a Wiener process, the drift $g \in g_\pm$ depends on the environmental state, and $D$ is the variance which depends on the quality of each observation. Often, the prefactor of the Weiner process is taken to be $\sigma$, rather than $\sqrt{2D}$, but the form we have chosen

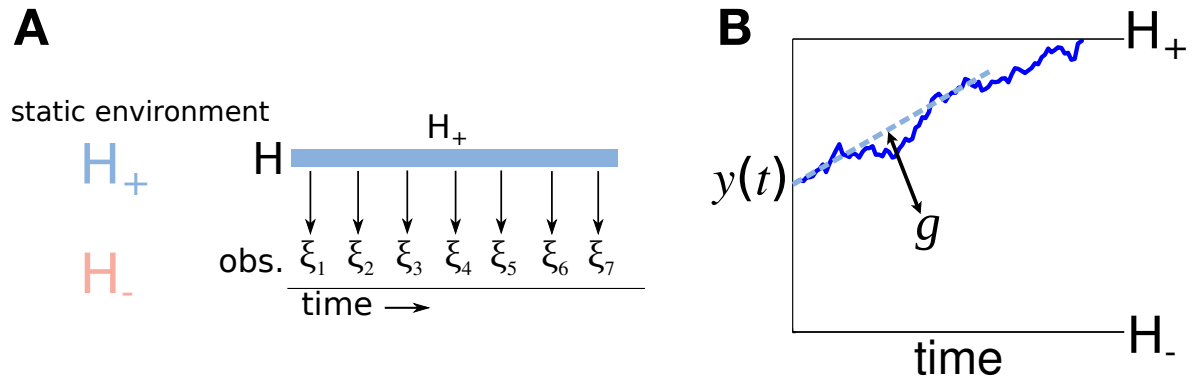


Fig. 5.0.1: **A.** In a standard two-alternative forced choice task, observers make a sequence of observations, $\xi_{1:n}$, to determine which of the two choices is correct. **B.** Decision time measures how long it takes $y(t)$ to reach a threshold. Positive/ Negative threshold indicates decisions $H_+$/ $H_-$. The correct probability is the probability that $y(t)$ reaches the correct threshold. For example, when $g > 0$, the correct probability is the probability of exiting through the positive threshold when starting at an initial position $y(0)$.

simplifies several expressions we derive subsequently. Eq. (5.0.1) may be written as a stochastic differential equation $\dot{y} = g + \sqrt{2D}\zeta$, where $\zeta$ is Gaussian white noise, and some use the form $\mathrm{d}y = g\mathrm{d}t + \zeta \cdot \sqrt{2D \cdot \mathrm{d}t}$.

For simplicity and consistency with typical random dot kinetogram tasks, we assume each drift direction is of equal unit strength: $g_+ = -g_- = 1$, and task difficulty is controlled by scaling the variance through $D$. An arbitrary drift amplitude $g$ can also be scaled out via a change of variables $y = g\tilde{y}$, so the resulting DDM for $\tilde{y}$ has unit drift, and $\tilde{D} = D/g^2$. The initial condition is determined by the observer's prior bias, $y(0) = D\ln[\mathrm{P}(H = H_+)/\mathrm{P}(H = H_-)]$. Thus for an unbiased observer, $y(0) = 0$.

There are two primary ways of obtaining a response from the DDM given by

Eq.(5.0.1) that mirror common experimental protocols. The first way is that the ideal observer interrogates at a set time, $t = T$, responds with $\text{sign}(y(T)) = \pm 1$ indicating the more likely of the two states, $H = H_{\pm}$, given the accumulated evidence. On the other hand, an observer free to choose their response time can trade speed for accuracy in making a decision. This is typically modeled in the DDM by defining a decision threshold, $\theta$, and assuming that at the first time, $T$, at which $|y(T)| \geq \theta$, the evidence accumulation process terminates, and the observer chooses $H_{\pm}$ if $\text{sign}(y(T)) = \pm 1$ (see Fig. 5.0.1 B).

## 5.1 Derivation of the Drift-Diffusion Model for a Single Trial

We assume that the optimal observer integrates a stream of noisy measurements $\xi_{1:m} = (\xi_1, \xi_2, ..., \xi_m)$ at equally spaced times $t_{1:m} = (t_1, t_2, ..., t_m)$. The likelihood functions, $f_{\pm}(\xi) := \mathrm{P}(\xi|H_{\pm})$, define the probability of each measurement, $\xi$, conditioned on the environmental state, $H_{\pm}$. Observations on each trial are combined with the prior probabilities $\mathrm{P}(H_{\pm})$. We assume a symmetric prior, $\mathrm{P}(H_{\pm}) = 1/2$. The probability ratio is then

$$R_m = \frac{\mathrm{P}(H = H_+|\xi_{1:m})}{\mathrm{P}(H = H_-|\xi_{1:m})} = \frac{f_+(\xi_1)f_+(\xi_2)\cdots f_+(\xi_m)\mathrm{P}(H = H_+)}{f_-(\xi_1)f_-(\xi_2)\cdots f_-(\xi_m)\mathrm{P}(H = H_-)}, \tag{5.1.1}$$

due to the independence of each measurement $\xi_m$. Thus, if $R_m > 1 (R_m < 1)$ then $H = H_+ (H = H_-)$ is the more likely state. Eq. (5.1.1) can be written recursively

$$R_m = \left(\frac{f_+(\xi_m)}{f_-(\xi_m)}\right) R_{m-1}, \tag{5.1.2}$$

where $R_0 = \mathrm{P}(H_+)/\mathrm{P}(H_-) = 1$ due to the equal probability of both states at the beginning of the first trial.

Taking the logarithm of Eq. (5.1.2) allows us to express the recursive relation for the log-likelihood ratio (LLR) $L_m^1 = \ln R_m$ as an iterative sum

$$L_m = L_{m-1} + \ln \frac{f_+(\xi_m)}{f_-(\xi_m)},$$

where $L_0 = \ln\left[\mathrm{P}(H_+)/\mathrm{P}(H_-)\right]$, so if $L_m \gtrless 0$ then $H = H_\pm$ is the more likely state. Taking the continuum limit $\Delta t \to 0$ of the timestep $\Delta t := t_m - t_{m-1}$, we can use the functional central limit theorem to yield the DDM

$$\mathrm{d}y = g\mathrm{d}t + \sqrt{2D}\mathrm{d}W, \tag{5.1.3}$$

where $W$ is a Wiener process, the drift $g \in g_\pm = \frac{1}{\Delta t}\mathrm{E}_\xi\left[\ln \frac{f_+(\xi)}{f_-(\xi)} | H_\pm\right]$ depends on the state, and $2D = \frac{1}{\Delta t}\mathrm{Var}_\xi\left[\ln \frac{f_+(\xi)}{f_-(\xi)} | H_\pm\right]$ is the variance which depends only on the noisiness of each observation but not the state.

With Eq. (5.0.1) in hand, we can relate $y(t)$ to the probability of either state $H_\pm$ by noting its Gaussian statistics in the case of free boundaries

$$\mathrm{P}(y(t)|H = H_\pm) = \frac{1}{\sqrt{4\pi D t}} \mathrm{e}^{-(y(t) \mp t)^2/(4Dt)},$$

so that

$$\frac{\mathrm{P}(H = H_+|y(t))}{\mathrm{P}(H = H_-|y(t))} = \frac{\mathrm{P}(y(t)|H = H_+)}{\mathrm{P}(y(t)|H = H_-)} = \mathrm{e}^{y(t)/D}. \tag{5.1.4}$$

Note that an identical relation was obtained in in the case of absorbing boundaries. Either way, it is clear that $y = D \cdot \mathrm{LLR}$. This means that before any observations have been made $\mathrm{e}^{y(0)/D} = \frac{\mathrm{P}(H=H_+)}{\mathrm{P}(H=H_-)}$, so we scale the log ratio of the prior by $D$ to yield the initial condition $y_0$ in Eq. (5.0.1).

## 5.2 Mean First Passage Time

Here we provide a brief overview of how to compute the mean first passage time in a DDM by analyzing the corresponding Kolmogorov backward equation. Assume the decision variable, $y$, evolves according to Eq. (5.0.1) with positive drift $g^1 = 1$ (a similar derivation can be carried out for $g^1 = -1$), and let $p(y, t|y_0, 0)$ denote the conditional density of state $y$ at time $t$ given initial condition $y_0$ at time $t = 0$ whose corresponding Kolmogorov equation is

$$\partial_t p(y, t|y_0, 0) = \partial_{y_0} p(y, t|y_0, 0) + D \cdot \partial^2_{y_0} p(y, t|y_0, 0), \tag{5.2.1}$$

given initial condition $p(y, 0|y_0, 0) = \delta(y - y_0)$ and boundary conditions $p(y, t| \pm \theta, 0) = 0$ for $y \in (-\theta, \theta)$, implying absorbing boundaries at $y_0 = \pm\theta$. The survival probability, that $y \in (-\theta, \theta)$ at time $t$, is defined:

$$G(y_0, t) := \int_{-\theta}^{\theta} p(y, t|y_0, 0)\mathrm{d}y = \mathrm{P}(T \geq t),$$

where $T$ is a random variable, corresponding to the time the state $y$ leaves the domain $y \in (-\theta, \theta)$. Integrating Eq. (5.2.1), we obtain

$$\partial_t G(y_0, t) = \partial_{y_0} G(y_0, t) + D \cdot \partial^2_{y_0} G(y_0, t). \tag{5.2.2}$$

Boundary conditions for Eq. (5.2.1) imply $G(\pm\theta, t) = 0$.

The mean first passage time, $T(y_0; \theta)$, for the state to depart the domain $(-\theta, \theta)$ given initial condition $y_0$ is then given by marginalizing $t$ against the flux $\partial_t G(y_0, t)$ of survival probability out of $(-\theta, \theta)$:

$$T(y_0; \theta) = -\int_0^{\infty} t\partial_t G(y_0, t)\mathrm{d}t = \int_0^{\infty} G(y_0, t)\mathrm{d}t.$$

Noting $\int_0^\infty G(y_0,t)\mathrm{d}t = \lim_{t\to\infty} G(y_0,t) - G(y_0,0) = -1$ and integrating Eq. (5.2.2), we find

$$\partial_{y_0} T(y_0;\theta) + D\cdot\partial_{y_0}^2 T(y_0;\theta) = -1,$$

with boundary condition $T(\pm\theta;\theta) = 0$, which can be solved explicitly

$$T(y_0;\theta) = \theta\left[\frac{\mathrm{e}^{\theta/D} + \mathrm{e}^{-\theta/D} - 2\mathrm{e}^{-y_0/D}}{\mathrm{e}^{\theta/D} - \mathrm{e}^{-\theta/D}}\right] - y_0, \tag{5.2.3}$$

which for $y_0 = 0$ simplifies to

$$T(0;\theta) = \theta\left[\frac{1-\mathrm{e}^{-\theta/D}}{1+\mathrm{e}^{-\theta/D}}\right] = \mathrm{DT}. \tag{5.2.4}$$

## 5.3 Probability of Exit Through a Boundary

To derive the probability of exit from either boundary $y = \pm\theta$, when the decision variable begins at $y_0$, we integrate the probability current through the boundary of interest, $\pm J(\pm\theta, t|y_0, 0)$. For instance, the probability the decision variable exits $y = +\theta$ after time $t$ is

$$g_\theta(y_0,t) = \int_t^\infty J(\theta,t'|y_0,0)\mathrm{d}t' = \int_t^\infty \left[p(\theta,t'|y_0,0) - D\partial_{y_0}p(\theta,t'|y_0,0)\right]\mathrm{d}t'. \tag{5.3.1}$$

Using the fact that $p(\theta,t|y_0,0)$ satisfies Eq. (5.2.1), we find that $g_\theta(y_0,t)$ satisfies

$$\partial_{y_0} g_\theta(y_0,t) + D\partial_{y_0}^2 g_\theta(y_0,t) = \int_t^\infty \partial_{t'} J(\theta,t'|y_0,0)\mathrm{d}t' = J(\theta,t|y_0,0) = \partial_t g_\theta(y_0,t). \tag{5.3.2}$$

Taking $t \to 0$ and defining $\pi_\theta(y) := g_\theta(y_0, 0)$ , we see that $J(\theta, 0|y_0, 0)$ vanishes if $\theta \neq y_0$, since $p(\theta, 0|y_0, 0) = \delta(y_0 - \theta)$, so the right hand side goes to zero and

$$\partial_{y_0} \pi_\theta(y_0) + D\partial_{y_0}^2 \pi_\theta(y_0) = 0, \tag{5.3.3}$$

where $\pi_\theta(\theta) = 1$, $\pi_\theta(-\theta) = 0$, and $\pi_\theta(y_0) + \pi_{-\theta}(y_0) = 1$. Solving the above equations explicitly, we find

$$\pi_\theta(y_0) = \frac{1 - \mathrm{e}^{-(y_0+\theta)/D}}{1 - \mathrm{e}^{-2\theta/D}}, \qquad \pi_{-\theta}(y_0) = \frac{\mathrm{e}^{-(y_0+\theta)/D} - \mathrm{e}^{-2\theta/D}}{1 - \mathrm{e}^{-2\theta/D}}. \tag{5.3.4}$$

An exit through the threshold $\theta$ results in a correct choice of $H = H_+$, so $c = \pi_\theta(0)$. The correct probability $c$ increases with $\theta$, since more evidence is required to reach a larger threshold. Defining the decision as $d = \pm 1$ if $y(T) = \pm\theta$, Bayes' rule implies

$$\frac{1}{1 + \mathrm{e}^{-\theta/D}} = c = \mathrm{P}(d = \pm 1 | H = H_\pm) = \mathrm{P}(H = H_\pm | d = \pm 1), \tag{5.3.5a}$$

$$\frac{\mathrm{e}^{-\theta/D}}{1 + \mathrm{e}^{-\theta/D}} = 1 - c = \mathrm{P}(d = \pm 1 | H = H_\mp) = \mathrm{P}(H = H_\mp | d = \pm 1), \tag{5.3.5b}$$

since $\mathrm{P}(H = H_\pm) = \mathrm{P}(d = \pm 1) = 1/2$. Rearranging the expressions in Eq. (5.3.5) and isolating $\theta$ relates the threshold $\theta$ to the LLR given a decision $d = \pm 1$

$$\pm\theta = D \ln \frac{\mathrm{P}(H = H_+ | d = \pm 1)}{\mathrm{P}(H = H_- | d = \pm 1)}.$$

Chapter 6

# Compared to Felix's data

## 6.1 Felix's experiments

I obtained Felix's data from "pop_tbl.mat". I'm interested in trial duration (trl_dur), trial coherence (trl_coh), RDP direction (rdp_dir), joystick direction (js_dir), and joystick strength (js_str) in column 3, 4, 7, 11, 13, respectively. "Each" trial consists 10 steady states (rdp_dir) with same coherence (trl_coh). The trial duration is total time across 10 steady states. Each frame lasts for 10 seconds, and each steady state lasts for 1to 2.5 seconds.

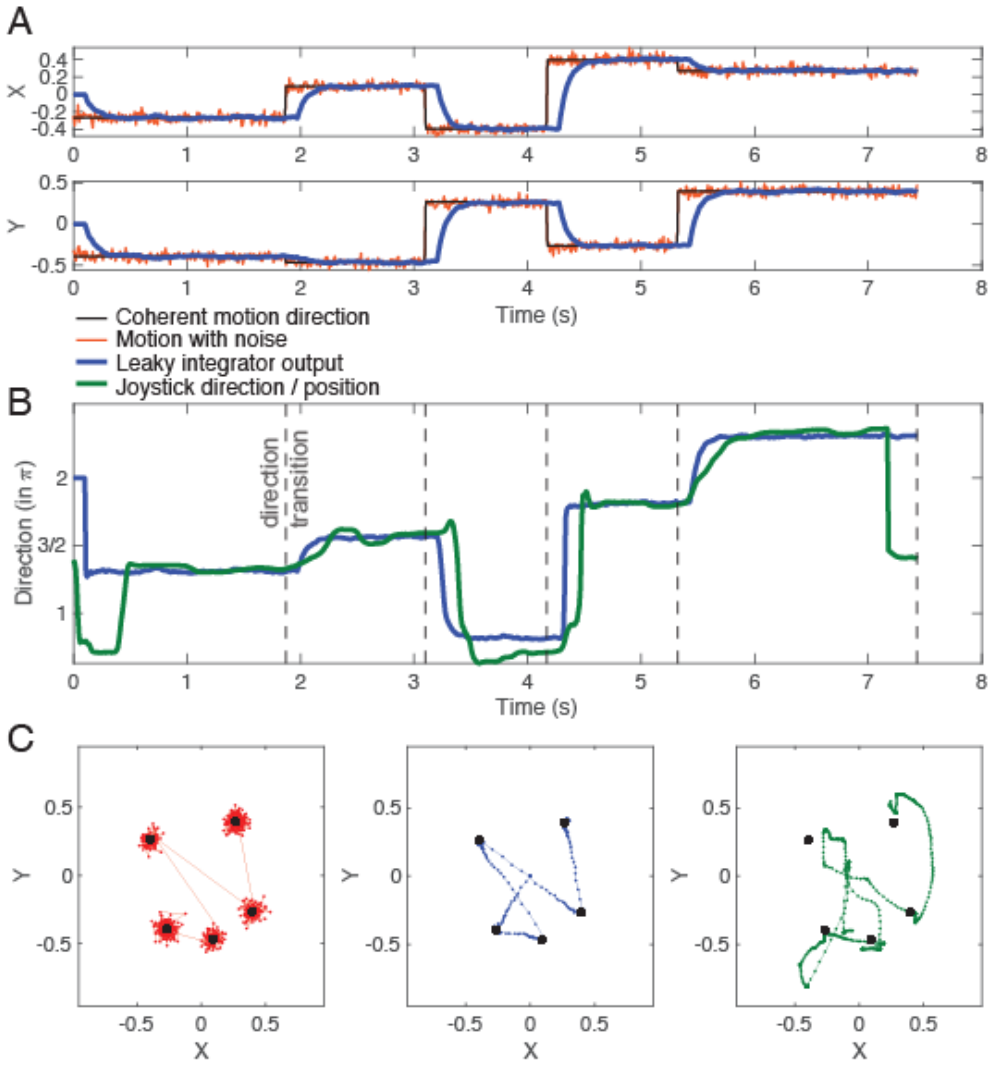


Fig. 6.1.1: Simulations of 2D leaky integrator for the solo CPR task. (A) Horizontal and vertical components of 5 cycles of stead-state random dot motion (RDM) stimulus direction at the motion coherence 0.47, and the output of the optimal leaky integrator (angular accuracy optimized), delayed by 100 ms to account for the visual response latency. (B) The direction of leaky integrator output and a human participant joystick direction, shifted back 270 ms to account for motor response delay (shift value derived from cross-correlation). (C) 2D trajectories of the above (here the green trace is the actual joystick position).

## 6.2 Frame-by-frame dot positions

To better model the evidence, Felix started collecting positions of dots frame-by-frame. Three important variables are 'STIM_RDP_dot_positions', 'IO_joystickDirection', and 'IO_joystickStrength', which indicate dots' position frame-by-frame, subject's joystick direction and joystick strength, respectively. The screen frame rate is 120 Hz, or 8.333 msec per frame. Two other variables that represent the coherence level and rdp direction are 'STIM_RDP_coherence' and 'STIM_RDP_direction'. The data is in shared folder (/Volumes/DPZ/KognitiveNeurowissenschaften/CNL/-DATA/fxs/CPR_psychophysics).

Note that the time stamp for dot positions variables is in frame, while time stamp for joystick data is in time. The joystick data is recorded every 10 msec while each frame only lasts for 8.333 msec (see Fig. 6.2.1). When we extract the data, we need to convert joystick data time stamp to frame. The "new" joystick value at the "new" current timestamp is the same value of the current "old" timestamp if both "new" and "old" timestamps are coincided. Otherwise, (or) the previous "old" timestamp.

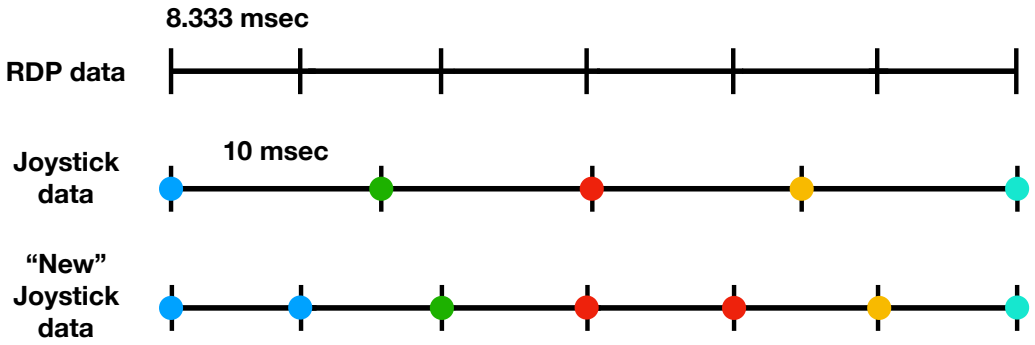


Fig. 6.2.1: Each rdp data is recorded at every frame, or every 8.333 msec while each joystick data is recorded at every 10 msec. Dots with the same color have the same value.

## 6.3 Leaky integrator of real subjects

Previously, we modeled leaky integrator as an ideal observer and find optimal parameters such as leaky constant $\lambda$, delay time $T_{\text{delay}}$, and that tracking the veridical direction of the rdp. Since we now have frame-by-frame rdp data and joystick direction data, we can find optimal parameters that predict the behaviors of joystick.

We need to pay attention is the coordinate system of the experiment because it's not standard coordinate. In the standard coordinate, reference point $0^{o}$ is at 3 o'clock . In our experiment, the reference point is at 12 o'clock. To convert the standard coordinates to our coordinates, we first reflect the standard coordinates about x-axis, then rotate by $90^{o}$ clockwise (see Fig. 6.3.1).

$$\begin{bmatrix} x \\ y \end{bmatrix} \xrightarrow[\text{reflection}]{\text{vertical}} \begin{bmatrix} -x \\ y \end{bmatrix} \xrightarrow[90^{o}\text{clockwise}]{\text{rotate}} \begin{bmatrix} \cos(-90^{o}) & -\sin(-90^{o}) \\ \sin(-90^{o}) & cos(-90^{o}) \end{bmatrix} \cdot \begin{bmatrix} -x \\ y \end{bmatrix} = \begin{bmatrix} y \\ x \end{bmatrix}$$

Therefore,

$\boldsymbol{X}_{\text{stim}} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where

$$\boldsymbol{\mu} = \text{coherence} \cdot d \cdot \begin{bmatrix} \sin(a_s) \\ \cos(a_s) \end{bmatrix}$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0 & 0.5 \cdot d^2 \cdot \frac{1-\text{coherence}}{n} \\ 0.5 \cdot d^2 \cdot \frac{1-\text{coherence}}{n} & 0 \end{bmatrix}.$$

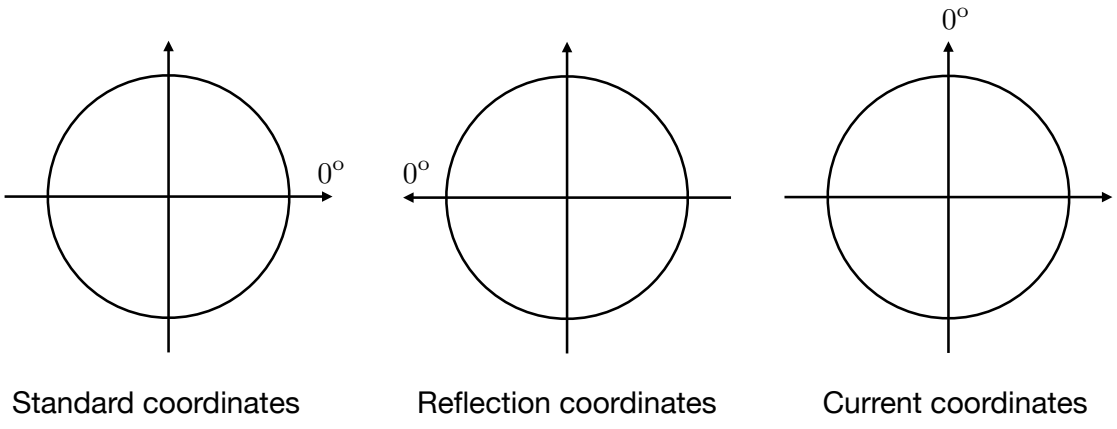


Fig. 6.3.1: Different coordinates: standard coordinate (left), vertical reflection coordinate (middle), and the coordinate we use in the experiment (right).

One problem we need to solve before optimizing mean squared errors between

joystick and leaky integrator is optimizing mean squared errors between rdp direction and joystick direction. Recalls that the rdp is scaled on coherence level while joystick range is between 0 and 1 for any sets of coherence level. The evidence at each time s

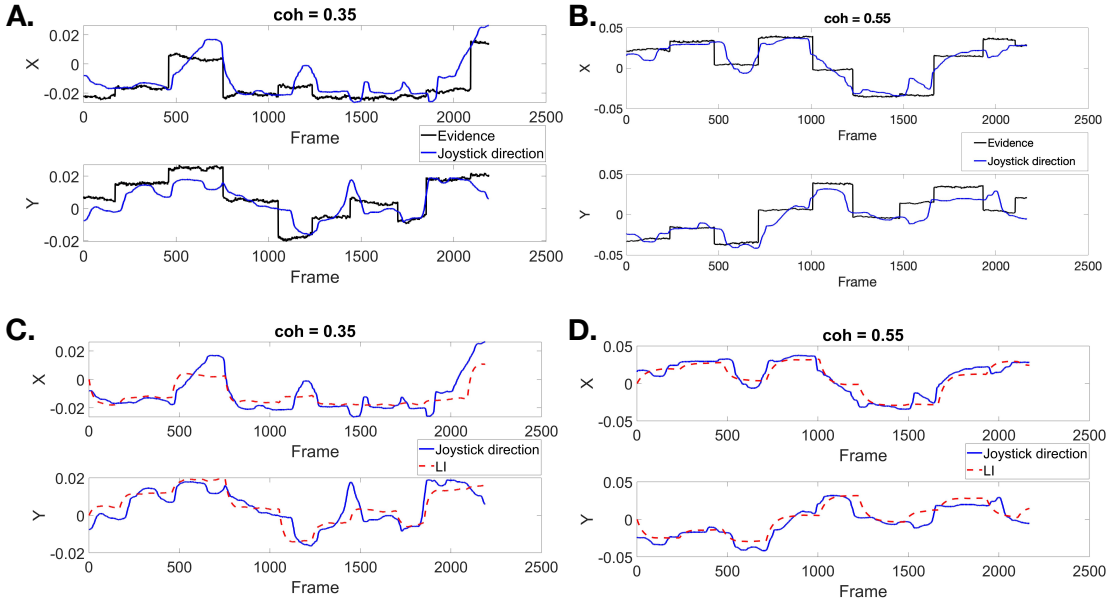


Fig. 6.3.2: Real subject's joystick direction vs. ideal subject's joystick direction (aka. leaky integrator) . **A.& B.** Subject nak

Chapter 7

# ARMA model

## 7.1 ARMA model vs. leaky integrator model

We want to compare the ARMA(1,1) to the leaky integrator.

ARMA(1,1):

$$\theta_t = a_t + b \cdot \theta_{t-1} + c \cdot \alpha_{t-\tau} \tag{7.1.1}$$

where $\theta_t$ is joystick direction, $\alpha_{t-\tau}$ is true direction of the stimulus at time $t - \tau$.

The leaky integrator:

$$\mathbf{X}_t = \mathbf{\Sigma} \mathrm{d}W_t + (1 - \lambda \cdot \mathrm{d}t)\mathbf{X}_{t-1} + \boldsymbol{\mu}_{t-T_{delay}} \mathrm{d}t \tag{7.1.2}$$

$\mathbf{X}_t = (x_t, y_t)$ is position of the joystick at time $t$, $\lambda$ is the leaky constant,

$$\boldsymbol{\mu}_{t-T_{\text{delay}}} = \begin{bmatrix} \text{coh} \cdot \cos(\alpha_{t-T_{\text{delay}}}) \\ \text{coh} \cdot \sin(\alpha_{t-T_{\text{delay}}}) \end{bmatrix},$$

and

$$\mathbf{\Sigma} = \begin{bmatrix} \frac{1}{2}\frac{1-\text{coh}}{n} & 0 \\ 0 & \frac{1}{2}\frac{1-\text{coh}}{n} \end{bmatrix}$$

where $n$ is number of dots.

The joystick direction is:

$$\theta_t = \arctan\left(\frac{y_t}{x_t}\right)$$

## 7.2 Circular Mean Square Error

The circular mean square error:

$$\text{CMSE} = \frac{1}{n}\sum_{i=1}^{n} 1 - \cos(\theta_i - \hat{\theta}_i) \tag{7.2.1}$$

where $\hat{\theta}_i$ is the estimator of $\theta_i$

Chapter 8

# Granger Causality

## 8.1 Granger Causality step-by-step

Given two time series $X(t)$, $Y(t)$.

Hypothesis: $X(t)$ causes $Y(t)$.

We then construct restricted and unrestricted models:

$$Y(t) = \sum_{i=1}^{k} \alpha_i Y(t-i) + c_1 + \epsilon_t \quad \textbf{(Restricted)}$$

$$Y(t) = \sum_{i=1}^{k} \alpha_i Y(t-i) + \sum_{j=1}^{k} \beta_j X(t-j) + c_2 + \epsilon_t^{'} \quad \textbf{(Unrestricted)}$$

where $\alpha_i$, $\beta_j$, $c_1$, $c_2$ are coefficients, $\epsilon_t$, $\epsilon_t^{'}$ are residuals, and $k$ is the order of regression.

To find the optimal order, we compare the AIC/BIC for each order. For a linear regression with $n$ observations and $k$ parameters used to fit, AIC and BIC can be

determined by

$$\text{AIC} = n \cdot \ln\left(\frac{\text{SSE}}{n}\right) + 2 \cdot (k+1) \tag{8.1.1}$$

$$\text{BIC} = n \cdot \ln\left(\frac{\text{SSE}}{n}\right) + \ln(n) \cdot (k+1) \tag{8.1.2}$$

AIC/BIC can be negative. The order that yields the least AIC/BIC is the optimal order.

Coefficients $\alpha_i$ and $\beta_j$ can be determined by minimizing the residuals, i.e. $\text{argmin}(Y_t - \sum_{i=1}^{k} \alpha_i Y(t-1) - c_1)$. We say $X(t)$ Granger causes $Y(t)$ when there is an improvement in residuals, i.e. $\epsilon_t' < \epsilon_t$.

The standard measure of G causality in the literature is defined for univariate predictor and predictee variables X and Y, and is given by the natural logarithm of the ratio of the residual variance in the restricted regression of the unrestricted regression:

$$F_{X \to Y} = \ln \frac{\text{Var}(\epsilon_t)}{\text{Var}(\epsilon_t')} \tag{8.1.3}$$

The Granger causality has the following properties:

(1) For scalar X and Y it is possible both for Y to G-cause X and for X to G-cause Y, a feedback stochastic process.

(2) $F$ can never be negative.

(3) Statistical significance can be determined via F-statistic:

$$\mathcal{F} = \frac{\frac{\text{RSS}_\text{r} - \text{RSS}_\text{ur}}{m}}{\frac{\text{RSS}_\text{ur}}{T-2m-1}} \tag{8.1.4}$$

where $\text{RSS}_\text{r} = \sum_{t=m+1}^{T} \epsilon_t^2$

Null hypothesis: $\beta_j = 0 \; \forall j$, i.e., $X(t)$ does not Granger cause $Y(t)$.

Alternate hypothesis: at least one of the lags of $X$ is significant, i.e. $X(t)$ Granger causes $Y(t)$.

From F-statistics formula 8.1.4, we can determine the p-value. If the p-value is less than 0.05 (for 95% confidence interval), we reject the null hypothesis. Therefore, $X(t)$ Granger causes $Y(t)$.

Granger Causality is not symmetric. $X(t)$ Granger causes $Y(t)$ does not mean $Y(t)$ Granger causes $X(t)$. We repeat same steps for the opposite direction.

## 8.2 Examples

### 8.2.1 Chickens, Eggs, and Causality, or Which Came First?

Thurman and Fisher examined annual U.S. time series from 1930 to 1983 of egg production and chicken population and perforemed Granger causality tests using one to four lags.

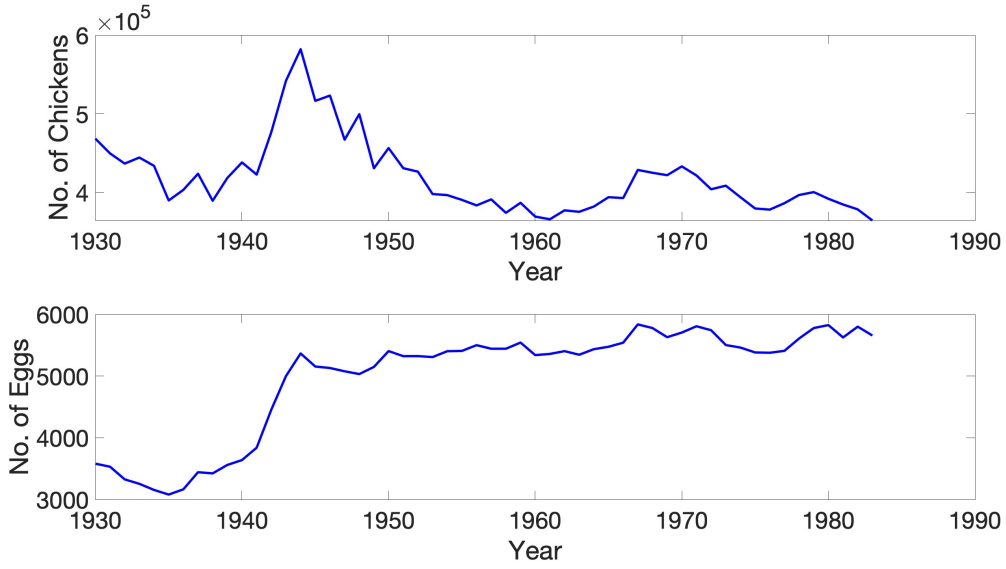


Fig. 8.2.1: Time series of chicken population (top) and egg production (bottom) from 1930 to 1983 in the United States. The data came from U.S. Department of Agriculture.

To conclude that one "came" first, they (1) reject the noncausality of the one to the other, and (2) at the same time fail to reject noncausality of the other to the one. If either both cause each other or neither causes the other, the question will remain unanswered.

First direction: did the chicken come first? We considered the restricted and non-restricted linear regressions:

$$\text{Eggs}_t = \sum_{i=1}^{L} \alpha_i \cdot \text{Eggs}_{t-1} + c + \epsilon_t \tag{8.2.1}$$

$$\text{Eggs}_t = \sum_{i=1}^{L} \alpha_i \cdot \text{Eggs}_{t-1} + \sum_{i=1}^{L} \beta_i \cdot \text{Chickens}_{t-1} + c + \epsilon_t^{'} \tag{8.2.2}$$

where $L$ is number of lags, $c$ is constant, $\epsilon_t$ and $\epsilon_t^{'}$ are residuals of the linear regressions.

| Lags | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $c$ | AIC |
|---|---|---|---|---|---|---|
| 1 | 0.9624 | N/A | N/A | N/A | 226.2566 | 546.3798 |
| 2 | 1.3287 | -0.3725 | N/A | N/A | 243.7684 | 530.3606 |
| 3 | 1.3017 | -0.3608 | 0.0042 | N/A | 306.1473 | 519.7606 |
| 4 | 1.2677 | -0.3166 | -0.0402 | 0.0281 | 341.5484 | 511.9186 |

Table 8.1: Table of coefficients for equation (8.2.1). Lag 4 yields the least AIC, so 4 is optimal lag.

| L | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $c$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.9624 | N/A | N/A | N/A | -0.0001 | N/A | N/A | N/A | 279.3413 |
| 2 | 1.4414 | -0.4896 | N/A | N/A | -0.0013 | 0.0007 | N/A | N/A | 508.2338 |
| 3 | 1.3892 | -0.4271 | -0.0221 | N/A | -0.0012 | 0.0010 | -0.0003 | N/A | 567.3588 |
| 4 | 1.3697 | -0.4194 | 0.0744 | -0.0916 | -0.0012 | 0.0008 | -0.0005 | 0.0003 | 638.4819 |

Table 8.2: Table of coefficients for equation (8.2.2). The AIC for lags 1 to 4 are 548.3300, 532.4489, 523.7438, and 518.0381, respectively. Here, lag 4 yields the least AIC, so 4 is optimal lag.

Finally, we compute the F-statistics (equation 8.1.4) and p-value:

| Lags | $m$ | $T$ | F statistics | p-value |
|---|---|---|---|---|
| 1 | 1 | 53 | 0.0470 | 0.8292 |
| 2 | 2 | 52 | 0.8800 | 0.4215 |
| 3 | 3 | 51 | 0.5916 | 0.6238 |
| 4 | 4 | 50 | 0.3929 | 0.8125 |

Table 8.3: Table of lags, degree of freedom $m$, total number of observations $T$, F statistics and p-value corresponding to those parameters to determine the statistical significance of chickens Granger causes eggs.

Recall our null hypothesis and alternate hypothesis:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_L = 0$ (chickens do not Granger cause eggs).

$H_A$ : At least one of $\beta$ is non-zero (chickens Granger cause eggs)

Table 8.3 failed to reject our null hypothesis at 5% level.

| Lags | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $c$ | AIC |
|---|---|---|---|---|---|---|
| 1 | 0.8500 | N/A | N/A | N/A | 6.1130e4 | 1,074.7 |
| 2 | 0.7576 | 0.1141 | N/A | N/A | 5.1983e4 | 1,056.6 |
| 3 | 0.7838 | 0.3134 | -0.2567 | N/A | 6.5750e+04 | 1,035.8 |
| 4 | 0.7790 | 0.3266 | -0.2387 | -0.0384 | 7.0589e+04 | 1,018.1 |

Table 8.4: Table of coefficients for equation (8.2.3). Lag 4 yields the least AIC, so 4 is optimal lag.

| L | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $c$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.8349 | N/A | N/A | N/A | -4.3214 | N/A | N/A | N/A | 8.8952e4 |
| 2 | 0.3280 | 0.4646 | N/A | N/A | 89.2693 | -94.2842 | N/A | N/A | 1.0566e+05 |
| 3 | 0.2920 | 0.4453 | 0.0041 | N/A | 76.5731 | -47.0783 | -35.9327 | N/A | 1.3354e5 |
| 4 | 0.2333 | 0.4580 | -0.0185 | 0.0257 | 87.3847 | -62.4941 | -8.2145 | -22.6355 | 1.4733e+05 |

Table 8.5: Table of coefficients for equation (8.2.4). The AIC for lags 1 to 4 are 1.0754e3, 1.0441e3, 1.0258e3, and 1.0087e3, respectively. Here, lag 4 yields the least AIC, so 4 is optimal lag.

Second direction: did the egg come first? We considered the restricted and non-restricted linear regressions:

$$\text{Chickens}_t = \sum_{i=1}^{L} \alpha_i \cdot \text{Chickens}_{t-1} + c + \epsilon_t \tag{8.2.3}$$

$$\text{Chickens}_t = \sum_{i=1}^{L} \alpha_i \cdot \text{Chickens}_{t-1} + \sum_{i=1}^{L} \beta_i \cdot \text{Eggs}_{t-1} + c + \epsilon_t^{'} \tag{8.2.4}$$

where $L$ is number of lags, $c$ is constant, $\epsilon_t$ and $\epsilon_t^{'}$ are residuals of the linear regressions.

We compute the F-statistics (equation 8.1.4) and p-value:

| Lags | $m$ | $T$ | F statistics | p-value |
|---|---|---|---|---|
| 1 | 1 | 53 | 1.2071 | 0.2772 |
| 2 | 2 | 52 | 8.8175 | 0.0006 |
| 3 | 3 | 51 | 5.40506 | 0.0030 |
| 4 | 4 | 50 | 4.2568 | 0.0057 |

Table 8.6: Table of lags, degree of freedom $m$, total number of observations $T$, F statistics and p-value corresponding to those parameters to determine the statistical significance of eggs Granger causes chickens.

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_L = 0$ (eggs do not Granger cause chickens).

$H_A$ : At least one of $\beta$ is non-zero (eggs Granger cause chickens)

F-statistics and p-values obtained from table 8.6 rejects the null hypothesis.

Therefore, we can conclude that eggs came first. This is the same conclusion as in Thurman's and Fisher's paper.

### 8.2.2 Toy models

We generate two time series

$$\begin{aligned} \mathrm{dX} &= 0.2 \cdot X(t) + 0.01 \cdot \mathrm{dW_t} \\ \mathrm{dY} &= 0.1 \cdot Y(t) + 2 \cdot X(t) + 0.02 \cdot \mathrm{dW_t} \end{aligned} \tag{8.2.5}$$

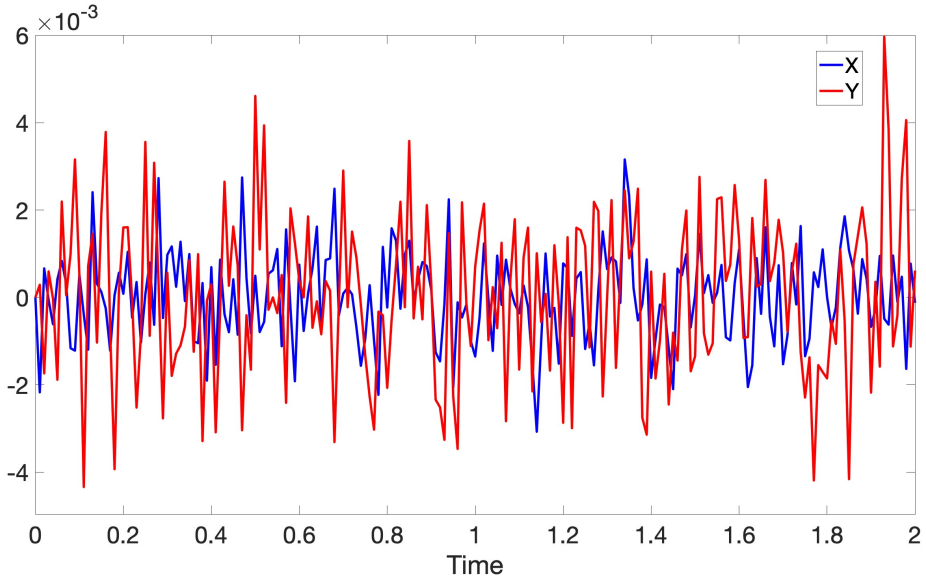


Fig. 8.2.2: Time series generated from equations (8.2.5). Here $\mathrm{d}t = 0.01$.

It is clear that $X(t)$ is independent on $Y(t)$ while $Y(t)$ depends on $X(t)$. First, we check to see if $X(t)$ Granger causes $Y(t)$.

Hypothesis: $X(t)$ Granger causes $Y(t)$. We suppose the lag to be 1, then

$$Y(t) = a_1 \cdot Y(t-1) + c_1 + \epsilon_t \quad \textbf{(Restricted)}$$

$$Y(t) = a_2 \cdot Y(t-1) + b_2 \cdot X(t-1) + c_2 + \epsilon_t' \quad \textbf{(Unrestricted)}$$

We computed AIC and BIC for restricted and unrestricted models for lags from 1 to 10:

| Lags | Restricted AIC | Unrestricted AIC |
|---|---|---|
| 1 | -2.5107e3 | -2.5088e3 |
| 2 | -2.4967e3 | -2.4936e3 |
| 3 | -2.4848e3 | -2.4810e3 |
| 4 | -2.4693e3 | -2.4640e3 |
| 5 | -2.4538e3 | -2.4466e3 |
| 6 | -2.4391e3 | -2.4300e3 |
| 7 | -2.4250e3 | -2.4143e3 |
| 8 | -2.4103e3 | -2.3986e3 |
| 9 | -2.3954e3 | -2.3818e3 |
| 10 | -2.3824e3 | -2.3666e3 |

Table 8.7: AIC of restricted and unrestricted models. Note that the optimal lag is 1, and this is consistent with how we generated our data.

Similarly, when we constructed restricted and unrestricted models for the hypothesis that $Y(t)$ Granger causes $X(t)$, the optimal lag is also 1. Though we did not show the results here.

To test our null hypothesis, we computed F-statistics and p-values. We have two null hypotheses: (1) $X(t)$ does not Granger cause $Y(t)$, and (2) $Y(t)$ does not Granger cause $X(t)$. We compare the p-values with $\alpha = 0.05$ (95% confidence interval) and find that both null hypotheses are not rejected for lags from 1 to 10 (see Fig. 8.2.3).

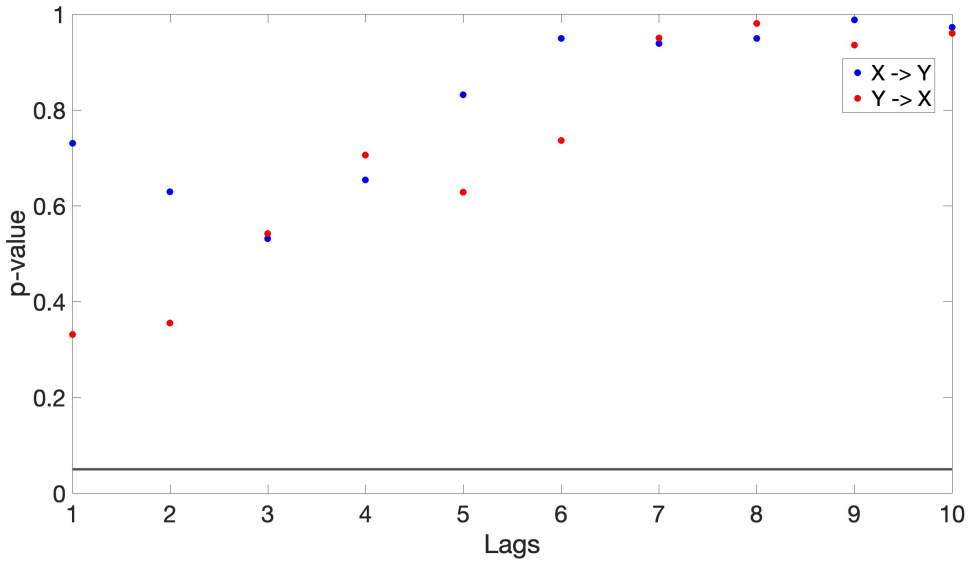


Fig. 8.2.3: Black line indicates $\alpha = 0.05$. The blue (red) dots indicate p-values with null hypothesis that $X(t)$ ($Y(t)$) does not Granger cause $Y(t)$ ($X(t)$).

Therefore, we cannot draw any conclusions about these two time series.

Thus, we generate two other time series, but this time, we increase the coupling strength between $Y(t)$ and $X(t)$.

$$\begin{aligned} \mathrm{dX} &= 0.2 \cdot X(t) + 0.01 \cdot \mathrm{dW_t} \\ \mathrm{dY} &= 0.1 \cdot Y(t) + 50 \cdot X(t) + 0.02 \cdot \mathrm{dW_t} \end{aligned} \tag{8.2.6}$$

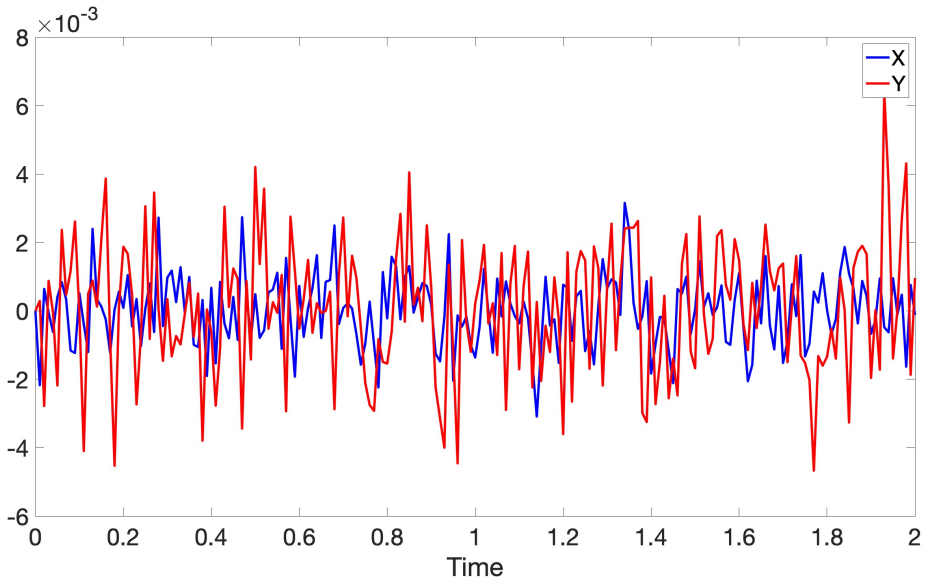


Fig. 8.2.4: Time series $X(t)$ and $Y(t)$ generated from stochastic differential equations (8.2.6)

We repeat the same computation (see Fig. 8.2.5) and find that the null hypothesis $Y(t)$ does not Granger cause $X(t)$ is not rejected for all lags. On the other hand, null hypothesis $X(t)$ does not Granger cause $Y(t)$ is rejected for lags 1 to 9. Therefore, we can conclude that $X(t)$ Granger causes $Y(t)$.

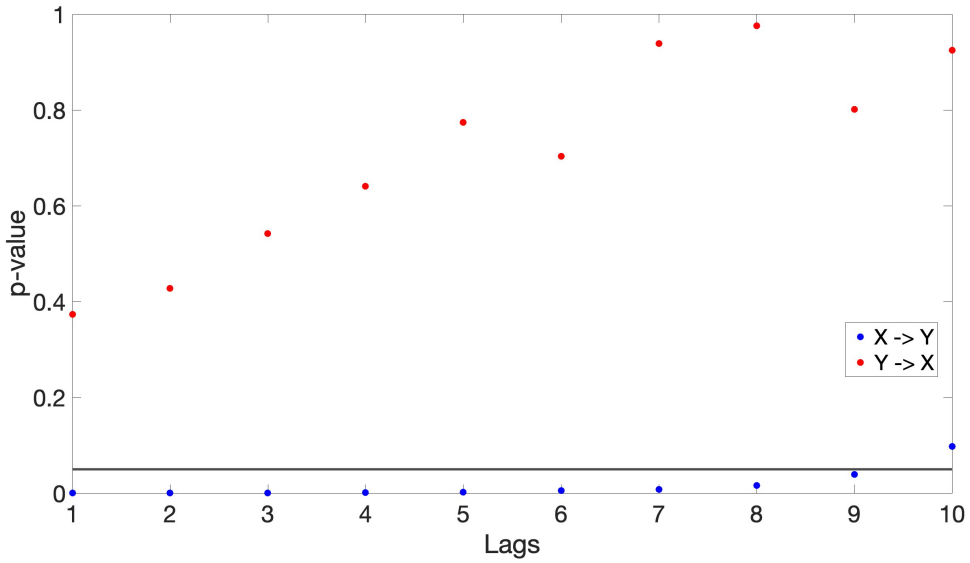


Fig. 8.2.5: Black line indicates $\alpha = 0.05$. The blue (red) dots indicate p-values with null hypothesis that $X(t)$ ($Y(t)$) does not Granger cause $Y(t)$ ($X(t)$).

## 8.3 Transfer entropy

Lionel Barnett stated that "Granger Causality and Transfer Entropy are equivalent for Gaussian variables". In the first subsection, I will derive Granger causality and transfer entropy for Gaussian variables, and in the second subsection, I will discuss a few methods to estimate transfer entropy.

### 8.3.1 Derivation

Given jointly distributed multivariate random variables $\boldsymbol{X}$, $\boldsymbol{Y}$, we denote by $\boldsymbol{\Sigma}(\boldsymbol{X})$ the $n \times n$ matrix of covariances $\mathrm{cov}(X_i, X_j)$ and by $\boldsymbol{\Sigma}(\boldsymbol{X}, \boldsymbol{Y})$ the $n \times m$ matrix of cross-covariances $\mathrm{cov}(X_i, Y_\alpha)$. The covariance matrix of the residuals of a linear

regression of $\boldsymbol{X}$ on $\boldsymbol{Y}$ is $\boldsymbol{\Sigma}(\boldsymbol{X}|\boldsymbol{Y})$

$$\boldsymbol{\Sigma}(\boldsymbol{X}|\boldsymbol{Y}) \equiv \boldsymbol{\Sigma}(\boldsymbol{X}) - \boldsymbol{\Sigma}(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{\Sigma}(\boldsymbol{Y})^{-1}\boldsymbol{\Sigma}(\boldsymbol{X},\boldsymbol{Y}))^{\mathrm{T}} \tag{8.3.1}$$

where $\boldsymbol{Y}$ is invertible. $\boldsymbol{\Sigma}(\boldsymbol{X}|\boldsymbol{Y})$ is also termed as "partial covariance" of $\boldsymbol{X}$ given $\boldsymbol{Y}$.

If $\boldsymbol{X}$ is a multivariate Gaussian random variable, then the entropy is:

$$H(\boldsymbol{X}) = \frac{1}{2}\ln(|\boldsymbol{\Sigma}(\boldsymbol{X}|\boldsymbol{Y})|) + \frac{1}{2}n\ln(2\pi e) \tag{8.3.2}$$

The conditional entropy $H(\boldsymbol{X}|\boldsymbol{Y})$ for two jointly multivariate Gaussian variables may be expressed in terms of the determinant of the corresponding partial covariance matrix:

$$H(\boldsymbol{X}|\boldsymbol{Y}) \equiv H(\boldsymbol{X}\oplus\boldsymbol{Y}) - H(\boldsymbol{Y}) \tag{8.3.3}$$

$$= \frac{1}{2}\ln(|\boldsymbol{\Sigma}(\boldsymbol{X}\oplus\boldsymbol{Y})|) - \frac{1}{2}\ln(|\boldsymbol{\Sigma}(\boldsymbol{Y})|) + \frac{1}{2}\ln(2\pi e) \tag{8.3.4}$$

If the process $\boldsymbol{X}_t$, $\boldsymbol{Y}_t$, $\boldsymbol{Z}_t$ are jointly multivariate Gaussian (i.e., any finite subset of the component variables $X_{ti}$, $Y_{s\alpha}$, $Z_{ua}$ has a joint Gaussian distribution), it follows that the transfer entropy becomes:

$$\mathcal{T}_{Y\to X|Z} \equiv \frac{1}{2}\ln\left(\frac{\left|\boldsymbol{\Sigma}(\boldsymbol{X}|\boldsymbol{X}^{-}\oplus\boldsymbol{Z}^{-})\right|}{\left|\boldsymbol{\Sigma}(\boldsymbol{X}|\boldsymbol{X}^{-}\oplus\boldsymbol{Y}^{-}\oplus\boldsymbol{Z}^{-}\right|)}\right) \tag{8.3.5}$$

On the other hand, the Granger causality is:

$$\mathcal{F}_{Y\to X|Z} \equiv \ln\left(\frac{|\boldsymbol{\Sigma}(\epsilon_t)|}{\left|\boldsymbol{\Sigma}(\epsilon'_t)\right|}\right) \tag{8.3.6}$$

$$= \ln\left(\frac{\left|\boldsymbol{\Sigma}(\boldsymbol{X}|\boldsymbol{X}^{-}\oplus\boldsymbol{Z}^{-})\right|}{\left|\boldsymbol{\Sigma}(\boldsymbol{X}|\boldsymbol{X}^{-}\oplus\boldsymbol{Y}^{-}\oplus\boldsymbol{Z}^{-})\right|}\right) \tag{8.3.7}$$

Therefore, when $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ are normally distributed, we can compute transfer entropy directly from Granger causality:

$$\mathcal{T}_{Y\to X|Z} = \frac{1}{2}\mathcal{F}_{Y\to X|Z} \tag{8.3.8}$$

The transfer entropy (in bits) quantifies information flow from one to another time series.

## 8.4 Minimum sample size to detect Granger causality at low coupling strength

### 8.4.1 Transfer entropy and p values for longer time series

As we saw in the example (see equations 8.2.5) when the coupling strength is small, we cannot detect Granger causality between two time series. Therefore, we extend the time series to see if this helps the GC detection.

In Felix's experiment, the frame rate is 120 Hz, therefore $\mathrm{d}t = 1/120 = 0.00833$ seconds. Thus, for this extension example, we changed $\mathrm{d}t = 0.00833$ seconds instead of $\mathrm{d}t = 0.01$ seconds. We increase our maximum time from 2 seconds to 200, 600, and 1000 seconds. Previously, we only have 201 observations, and now we have 24001, 72001, and 120001 observations. We can see from Fig. 8.4.1 **D**, **E**, **F** that $X(t)$ Granger-causes $Y(t)$ significantly for longer time series. Therefore, for weak couplings, the more observations we have, the better we can detect their Granger-cause relationship. Our goal is to find the minimum sample size that can detect Granger causality at low coupling strength.

Since $X(t)$ and $Y(t)$ are normally distributed, we can compute transfer entropy (TE) directly from Granger causality using formula 8.3.8. From Fig. 8.4.1 **A, B, C.**, we can see that there are more information flows from $X(t)$ to $Y(t)$ (blue dots).

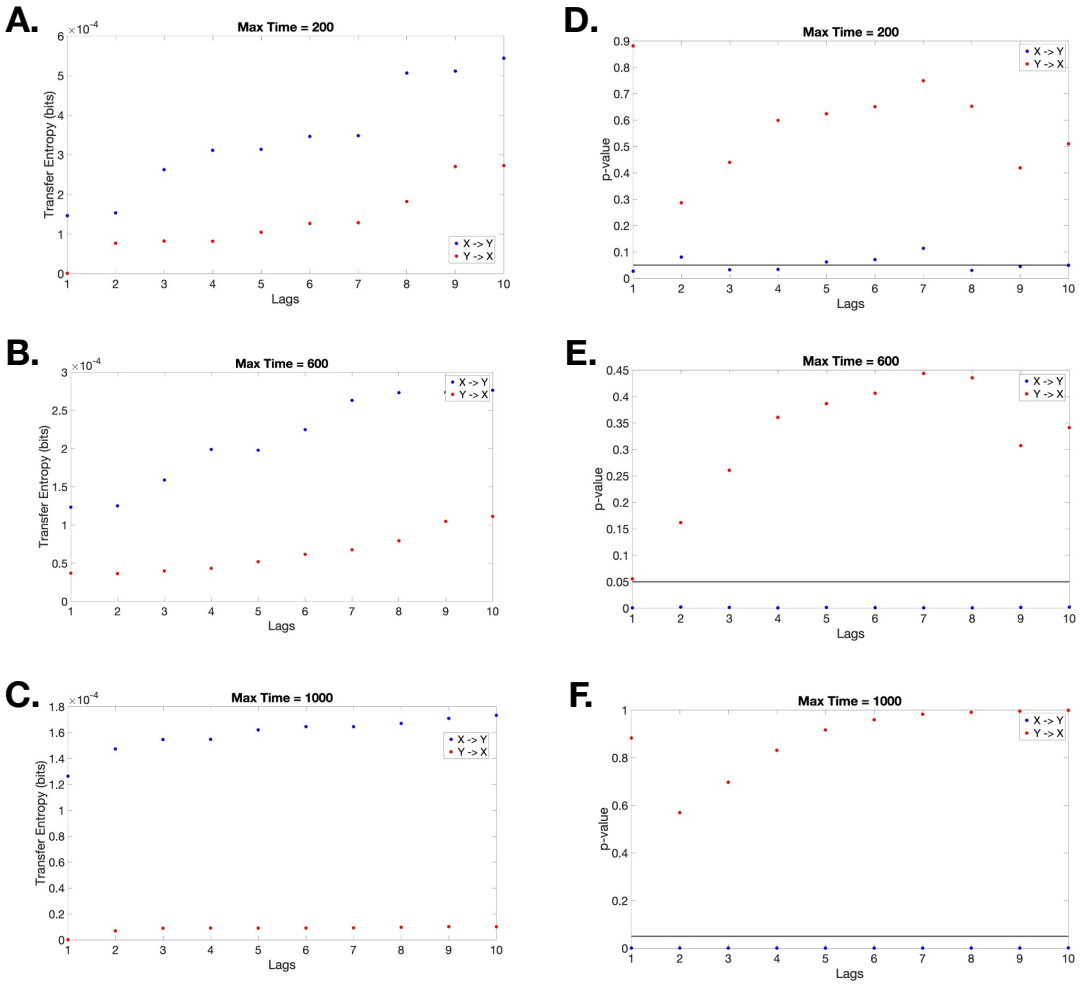


Fig. 8.4.1: **A, B, C.** Transfer entropy and **D, E, F.** p values when maximum time is $T = 200, 600, 1000$ seconds. Time series $X(t)$ and $Y(t)$ generated from stochastic differential equations (8.2.5).

### 8.4.2 Transfer entropy and p values for longer time series

In their study "Minimum Sample Size for Reliable Causal Inference Using Transfer Entropy", Ramos & Macau employed transfer entropy as a statistical test

to find minimum number of samples to produce reliable true positive.

They assume a null hypothesis: $X$ and $Y$ are independent for a particular lag $\tau$. At this lag, an ensemble of surrogate data is generated from original data. A confidence interval is defined as a percentile of surrogate's TE distribution. If TE calculated from original dataset is significantly higher than the confidence interval, then the null hypothesis is rejected and causality is detected. However, for a small dataset, the bias is considerable when compared to original TE, and this could leads to hypothesis is falsely accepted.

Their methodology is the following:

- For each sample size tested, 100 runs with different initial conditions are performed.
- For each run, an ensemble of 2000 surrogates is obtained from the original data, and the respective TE is calculated at a fixed $\tau$ (delay).
- A significance level of $\alpha = 0.0005$ is chosen, so 99.95% of the surrogate ensemble is defined as upper confidence bound. Let's call this bound $I_{\text{thr}}$.
- Whenever the orginal TE is higher than $I_{\text{thr}}$, the null hypothesis is rejected.

Another method to find the minimum sample size is power analysis. Power is defined as the probability that a statistical test will reject a false null hypothesis. Main output of power analysis is estimation of an appropriate sample size.

"If we think the effect is probably **A**, and there's **B** variation in X, and there's **C** variation in Y unrelated to X, and you want to have statistical precision of **E** or better, then you'll need a sample size of at least **D**." Given **A**, **B**, and **C**, what sample size **D** do you need to make your standard errors **E** or smaller?

We need to be able to fill in the values for 4 of those 5 pieces. Some aren't about guesses but about standards. For instance, 80% statistical power has been standard.

### 8.4.3 Significance testing process using randomized probability distributions

The results tend to be bias when we have limited data. When we estimate the probability distribution, noise in the estimation can produce a nonzero information result even though the true underlying probability distribution would produce a zero result. To account for this bias is to compare the performance between two information theory values via significance testing.

Our null hypothesis is: $X$ and $Y$ are independent.

- If $p \geq \alpha$, then null hypothesis cannot be rejected.
- If $p < \alpha$, then null hypothesis is rejected.

where $p$ is the p-value and $\alpha$ is the significance level, for example, $\alpha = 0.05$ for 95% confidence interval.

Let's consider a simple model system with two variables ($X$ and $Y$) and a measurement of their mutual information. Each variable take one of two possible

states. Then varied the interactions in the model using a parameter $a$ such that:

$$\begin{aligned} p(x=1,y=1) &= 0.25(1+a) \\ p(x=1,y=2) &= 0.25(1-a) \\ p(x=2,y=1) &= 0.25(1-a) \\ p(x=2,y=2) &= 0.25(1+a) \end{aligned} \tag{8.4.1}$$

From Law of Total Probability, for any value of $a$, we have $p(x=1) = p(x=2) = p(y=1) = p(y=2) = 0.5$. Their mutual information will be:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= \sum_{x\in X, y\in Y} p(x,y)\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right) \\ &= p(x=1,y=1)\log_2\left(\frac{p(x=1,y=1)}{p(x=1)p(y=1)}\right) \\ &+ p(x=1,y=2)\log_2\left(\frac{p(x=1,y=2)}{p(x=1)p(y=2)}\right) \\ &+ p(x=2,y=1)\log_2\left(\frac{p(x=2,y=1)}{p(x=2)p(y=1)}\right) \\ &+ p(x=2,y=2)\log_2\left(\frac{p(x=2,y=2)}{p(x=2)p(y=2)}\right) \\ &= 0.5(1+a)\log_2(1+a) + 0.5(1-a)\log_2(1-a) \end{aligned}$$

When $a = 0$, there was no relationship between $X$ and $Y$, thus in a perfect system, the mutual information $I(X;Y) = 0$.

Let's consider an example that conducted 100 observations from a model with no interactions between $X$ and $Y$. Its mutual information is nonzero due to random fluctuations in observations. We created null model data after generating

probability distributions by randomizing 100 observations in 10,000 runs. We obtained PDF and CDF from the null model (Fig. 8.4.2)

Next, we obtained PDF and CDF of mutual information distribution for 100 observations from histogram of 10,000 simulations. We set number of bins to be 50. The p-value is the intersection between the blue curve and the black vertical line (see Fig. 8.4.2 **B**.) From the CDF, we can compute the p-value of our hypothetical mutual information by doing linear interpolating between two points around the values that we want to estimate.

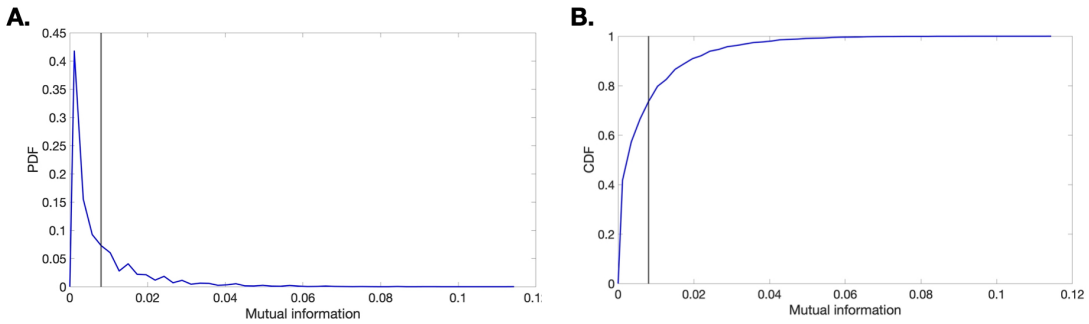


Fig. 8.4.2: **A.** PDF and **B.** CDF of mutual information of 100 observations from 10,000 simulations, where $X$ and $Y$ are discrete random variables and $p(X = 1) = p(X = 2) = p(Y = 1) = p(Y = 2) = 0.5$. The black vertical line is an example of mutual information where $a = 0$ (i.e. no interaction between $X$ and $Y$). The p-value is the difference between 1 and the intersection between the CDF (blue curve) and the real value (black line).

Chapter 9

# MATLAB

## 9.1 Programs

**circ_dist.m**: from Circular toolbox. It calculates the circular difference between two angles.

**fminsearchbnd.m**: this searches for the minimum within a bound.

**scprm_programs_ARMApq_func**: Find the constants of ARMA(p,q) model by minimizing the circular mean square error. The inputs are stimulus and joystick direction.

**scprm_programs_contour_func.m**: function that takes mean and variance matrix and compute the contours.

**scprm_programs_extract_func.m**: Extract veridical direction (a_unique), stimulus direction (a_stim), and coherence value (coh_val) from data.

**scprm_programs_leak_func.m**: leaky integrator model. The inputs are stimulus (in Cartesian coordinate $(X, Y)$), time interval, time step dt, leaky constant $\lambda$,

and delay time.

## 9.2 Codes

**scprm_ARMApq.m**: this file takes human data and computed ARMA constants.

**scprm_dots_analysis.m**: it computes the experimental stimulus per frame by computing angle of the resultant vectors.

**scprm_extract_mwkfile.m**: it extracts variables from mwk file.

**scprm_js_vs_stim.m**: it compares leaky integrator (with simulated and experimental stimuli) and joystick direction.

**scprm_l_opt_colorMap.m**: this file computes the optimal leaky constant $\lambda$ for each coherence and switch rate. It produces the color map in Fig. 2.1.6.

**scprm_lambda_opt.m**: it computes leaky integrator from optimal $\lambda$. It generates Fig. 6.1.1 C.

**scprm_leaky_vs_data.m**: it computes optimal values from simulated stimulus and uses those to compute leaky integrator from experimental stimulus. This file also computes circular mean squared error (CMSE) and AIC. The model with less CMSE and AIC fits the data better.

**scprm_leakytoymodel.m**: a toy model of leaky integrator of one steady state. Note that the numbers in this example are not from the experiment.

**scprm_stim_veridical.m**: it takes all blocks of one session to compute the experimental stimulus. This file also plots the mean and standard deviation of difference of experimental stimulus and corresponding veridical direction for each

coherence level. It takes time to run this file, so I only takes one session. If you plan to take all sessions and all subjects, please check the coherence levels for that session. Some sessions have different set of coherence levels.