Random processes in neuroscience and machine learning

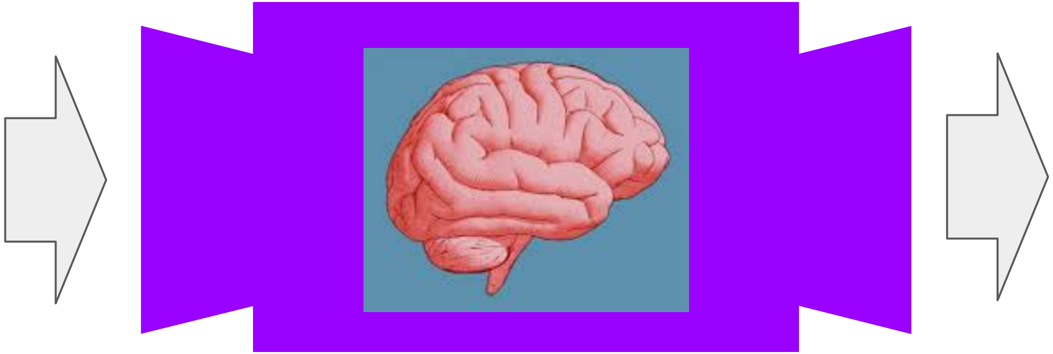
Or: How I learned to stop worrying and love the unknown

Rich Pang

#isiCNI2020

2020-01-08

There’s almost always something you don’t know



input **x**

output **y**

internal
variables
z

Any or all of **x**, **y**, or **z** can be unknown.

How do we handle this problem in a principled way?

Today’s goals

Develop intuition for working with random objects.

Practice working with probabilities mathematically in neuro/ML contexts.

Plan

Part 1

Meaning of randomness

Discrete vs continuous random variables

Probability distributions

Sampling random variables in Python

Mean and variance

Sample mean and variance

Means of functions of random variables

Part 2

Multivariate random objects

Covariance and correlation

Joint and marginal distributions

Conditional distributions and Bayes' Law

Probabilistic models

Time-series

Entropy and information theory

Part 1

What does it mean to be random?

Repeated experiments or measurements yield *different* outcomes.

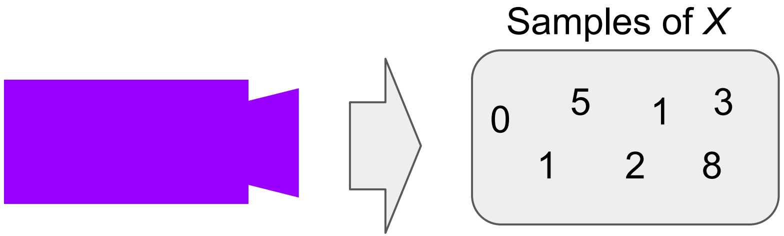
You can't predict these outcomes with perfectly.
(Maybe someone else can, though.)

We describe **random variables** using
probability distributions.

Discrete and continuous random variables

Discrete

Integers, categories
(e.g. number of spikes, cell type, image class)



Probability
mass function

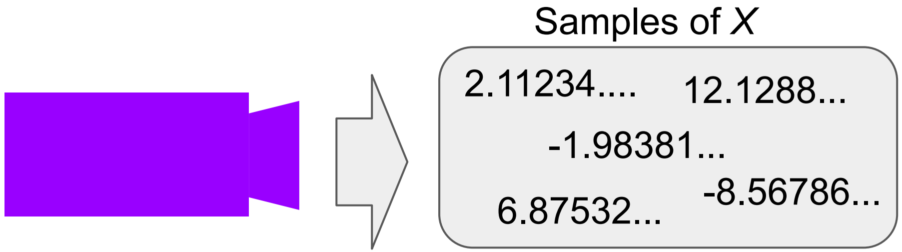
$$P(X = x_i) \equiv P(x_i)$$

$P(x_i)$ gives fraction of samples where x_i occurs

Normalization: $$\sum_i P(x_i) = 1$$

Continuous

Real-valued numbers
(e.g. firing rate, time difference)



$$P(X = x) \equiv P(x)$$

Probability
density function

$\int_a^b P(x)dx$ gives fraction of times x is between a and b

Normalization: $$\int_{-\infty}^{\infty} P(x)dx = 1$$

Problems

1.1: If a stimulus **x** makes a neuron generate a Poisson-distributed spike count with an average of 2.5 spikes, what are (a) the probability it spikes once, (b) the probability it doesn't spike, (c) the probability it spikes twice or more?

1.2: If the time interval between two spikes is exponentially distributed with mean 100 ms, what is the probability of observing (a) a time interval less than 50 ms, (b) a time interval between 100 and 200 ms, (c) a time interval greater than 200 ms (d) a time interval of exactly 150 ms?

1.3: What distributions would be good models for... ?

The number of spikes emitted by neuron in 1 second.
The absolute time interval between two spikes emitted by a neuron
Whether or not spike was emitted in 1 ms time bin.
The total synaptic input current to a neuron at one timepoint.
The number of vesicles released at a synapse after an action potential.
The time it takes to walk from Stoked to CCFM.
Which of 100 images was presented to neural network.
The number of white pixels in B&W image.
The sound intensity level at a random time during the day.
The change in firing rate of a neuron in response to a stimulus.
The difference between two neurons' firing rates.
Whether a neuron is inhibitory or excitatory.
Which of 10 cell types a neuron is.

1.4: T/F: the value $P(x_i)$ in a probability mass function can be greater than 1.

1.5: T/F: the value $P(x)$ in a probability density function can be greater than 1.

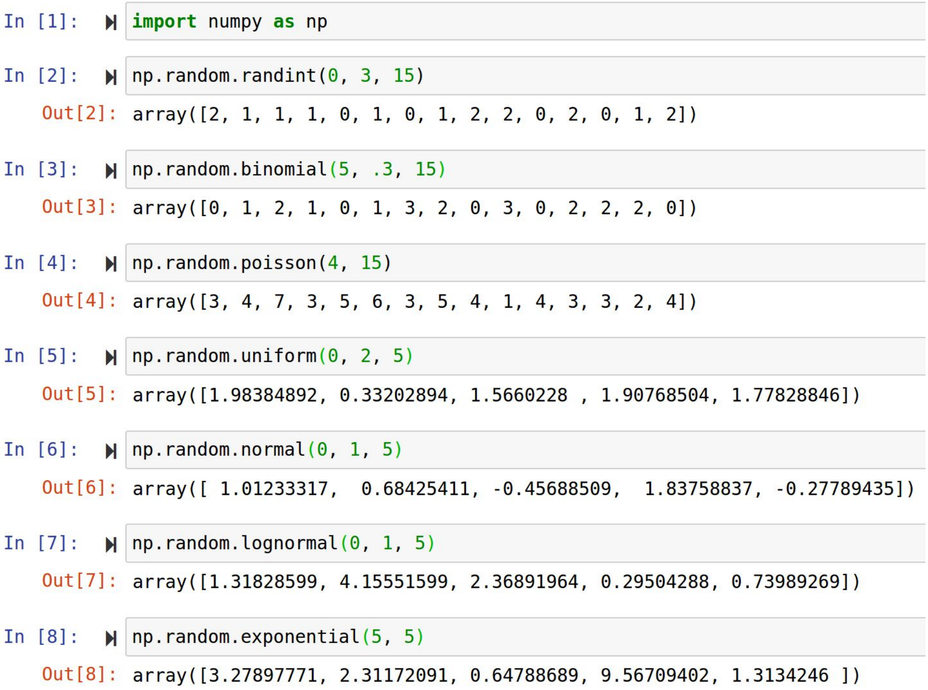
1.6: What units is probability mass in?

1.7: What units is probability density in?

1.8: Show that the exponential distribution with scale parameter β normalizes to 1.

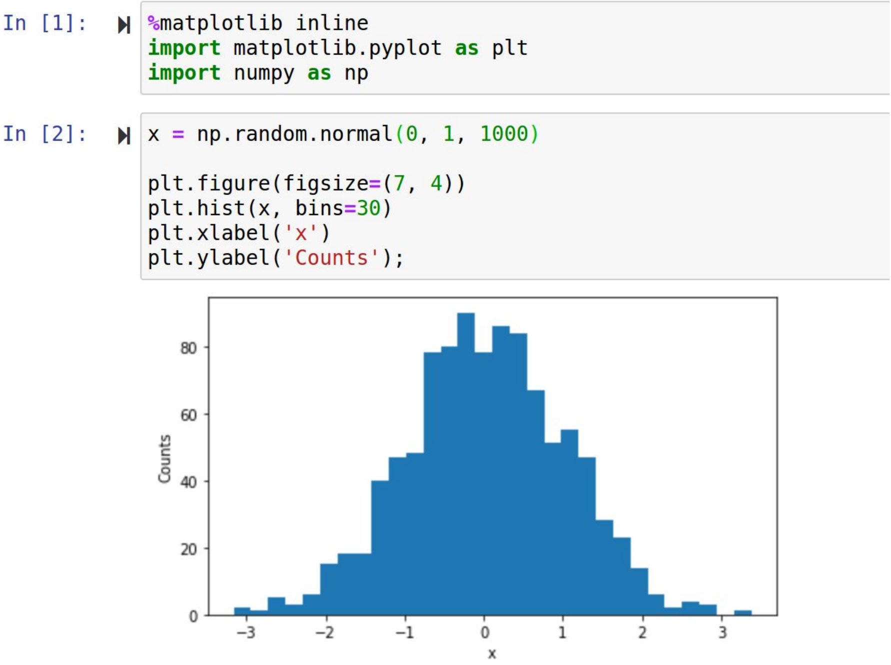
Sampling and plotting random variables in Python

Sampling



```
In [1]:   import numpy as np

In [2]:   np.random.randint(0, 3, 15)
  Out[2]: array([2, 1, 1, 1, 0, 1, 0, 1, 2, 2, 0, 2, 0, 1, 2])

In [3]:   np.random.binomial(5, .3, 15)
  Out[3]: array([0, 1, 2, 1, 0, 1, 3, 2, 0, 3, 0, 2, 2, 2, 0])

In [4]:   np.random.poisson(4, 15)
  Out[4]: array([3, 4, 7, 3, 5, 6, 3, 5, 4, 1, 4, 3, 3, 2, 4])

In [5]:   np.random.uniform(0, 2, 5)
  Out[5]: array([1.98384892, 0.33202894, 1.5660228 , 1.90768504, 1.77828846])

In [6]:   np.random.normal(0, 1, 5)
  Out[6]: array([ 1.01233317,  0.68425411, -0.45688509,  1.83758837, -0.27789435])

In [7]:   np.random.lognormal(0, 1, 5)
  Out[7]: array([1.31828599, 4.15551599, 2.36891964, 0.29504288, 0.73989269])

In [8]:   np.random.exponential(5, 5)
  Out[8]: array([3.27897771, 2.31172091, 0.64788689, 9.56709402, 1.3134246 ])
```

Plotting histograms



```
In [1]:   %matplotlib inline
          import matplotlib.pyplot as plt
          import numpy as np

In [2]:   x = np.random.normal(0, 1, 1000)

          plt.figure(figsize=(7, 4))
          plt.hist(x, bins=30)
          plt.xlabel('x')
          plt.ylabel('Counts');
```

Problems

2.1: Generate 20 samples of a random variable that is 1 with probability .5 and 0 with probability .5.

2.2: Generate 20 samples of a random variable that is 1 with probability .25 and 0 with probability .75.

2.3: Plot histograms of the following random variables, using 1000 samples each with 30 bins:

The square of a standard (mean 0, stdev 1) Gaussian random variable.

The difference between two standard Gaussian random variables.

The difference between two exponential random variables with $\beta = 2$.

The sum of two standard Gaussian random variables.

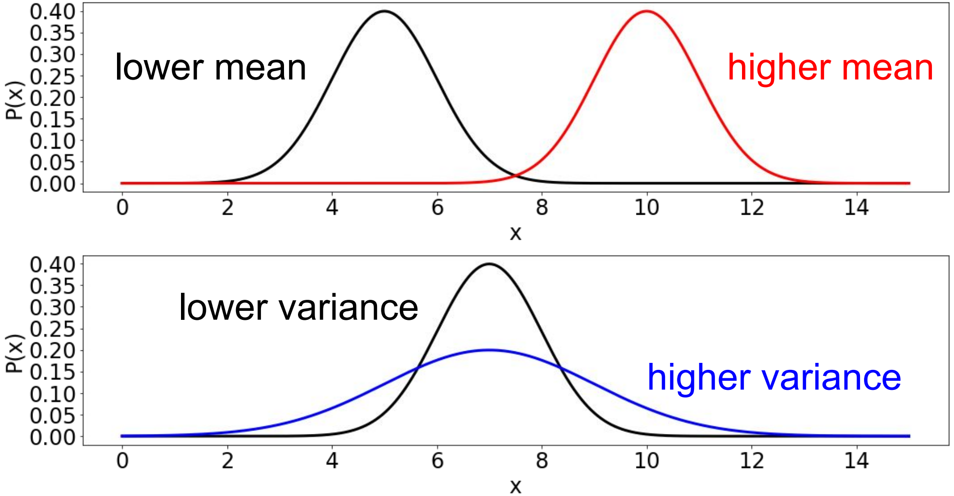
The sum of two exponential random variables with $\beta = 2$.

The sum of 100 standard Gaussian random variables.

The sum of 100 exponential random variables with $\beta = 2$.

Expectation (mean) and variance

We often want to characterize a probability distribution's key features.



			Discrete	**Continuous**
Mean	Location	$E[X]$	$\sum_i x_i P(x_i)$	$\int_{-\infty}^{\infty} x P(x) dx$
Variance	Spread^2	$\mathrm{Var}[X]$	$\sum_i (x_i - E[X])^2 P(x_i)$	$\int_{-\infty}^{\infty} (x - E[X])^2 P(x) dx$
Standard deviation	Spread	$\mathrm{Std}[X]$	$\sqrt{\sum_i (x_i - E[X])^2 P(x_i)}$	$\sqrt{\int_{-\infty}^{\infty} (x - E[X])^2 P(x) dx}$

Problems

Example: exponential distribution.

$$\boxed{P(x) = \frac{1}{\beta}\exp\left(\frac{-x}{\beta}\right) \qquad \text{for } x > 0}$$

2.4: Write out the equation for the mean of the exponential distribution with scale parameter β.

2.5: Write out the equation for the variance of the exponential distribution with scale parameter β.

2.6: Calculate the mean of the exponential distribution with scale parameter β.

Sample mean and variance

In experiments we usually have a collection of *samples*, not the true distribution.

We can ***estimate*** the mean and variance of the true distribution from these samples **using similar (but not identical!)** formulas:

Sample mean:

$$\hat{\mu} = \frac{1}{N}\sum_{a=1}^{N} x_a$$

Sample variance:

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{a=1}^{N}(x_a - \hat{\mu})^2$$

weird, I know

Note: these work for both discrete *and* continuous variables.

Problems

3.1: Run an “experiment” by drawing N = 30 samples from a Gaussian distribution with mean 0, standard deviation 2. Plot a histogram of the samples using 10 bins. What are the sample mean and sample variance of your results?

3.2: Re-run the experiment 100 times (i.e. 100 experiments of drawing N = 30 samples each). Plot a histogram of the *sample means*.

3.3: Repeat the 100 experiments for N = 10, 30, 50, 100, 1000, 5000. Plot histograms of the sample means for each value of N, with x limits from -3 to 3. How do they change as N increases?

Expectations of a function of a random variable

It is also easy to compute averages of *functions* of random variables.

$$E[f(X)] = \sum_i f(x_i)P(x_i) \qquad E[f(X)] = \int_{-\infty}^{\infty} f(x)P(x)dx$$

Remember **variance**?

$$E[(X - E[X])^2] = \sum_i \underbrace{(x_i - E[X])^2}_{f(x_i)} P(x_i) \quad \text{or} \quad \int_{-\infty}^{\infty} \underbrace{(x - E[X])^2}_{f(x)} P(x)dx$$

Problems

3.4: Suppose X, the signed time difference between two spikes is sampled from a continuous uniform distribution from -1 to 1. What is the expected value of X’s absolute value, |X|?

3.5: Verify your calculation using 1000 samples of X and computing the sample mean of |X| in Python.

Part 1 Recap

Meaning of randomness

Discrete vs continuous random variables

Probability distributions

Sampling random variables in Python

Mean and variance

Sample mean and variance

Means of functions of random variables

Part 2

Multivariate randomness (***random processes***)

Often, we must deal with *multiple* random variables sampled together.

Multivariate randomness (***random processes***)

Random variable

2.11234 12.1288

6.87532 -1.98381 -8.56786

SINGLE spike rate,

SINGLE stimulus value,

SINGLE behavior value,

Etc.

2-D Random process

[1.31443, -2.5324]

[2.1532, 5.1345]

[5.9388, -1.5833]

[6.123, -9.1234]

Spike rate AND stim value,

Stim AND behav value,

TWO behav values,

TWO spike rates,

Etc.

N-D Random process

N

[2.2381, 9.3214, ..., -1.6983]

[-1.0273, 1.2888, ..., 3.0382]

[5.9362, -2.5721, ..., 8.3702]

[-2.1174, 2.9733, ..., 4.0012]

SEVERAL spike rates,

SEVERAL stim values,

SEVERAL spike rates AND stim values

Etc.

The number one way to avoid confusion

ALWAYS make sure you know what the structure of your random object is.

Experiment 1: Present random **2-D** stimulus **x**, measure 4 neural firing rates **z**:

6-D random object: $\{\underset{\text{2-D}}{\underset{\nearrow}{\mathbf{x}}}, \underset{\text{4-D}}{\underset{\nwarrow}{\mathbf{z}}}\}$

6-D vector representation: $[x_1, x_2, z_1, z_2, z_3, z_4]$

Experiment 2: Present same stimulus **x**, measure 4 neural firing rates **z**:

4-D random vector: $[z_1, z_2, z_3, z_4]$

Problems

What is the dimension of (number of elements in) samples from the following random processes?

4.1: An experiment where a random 32 x 32 pixel black-and-white image is shown and 20 neural firing rates measured.

4.2: An experiment where a fixed stimulus **x** is shown, and one measures whether or not each of 3 neurons spikes in each of 1000 1 ms time-windows.

4.3: An experiment like 4.2, except where 2 time-varying behavioral outputs **y**(t) are also measured in 1 ms time-windows over 1 second.

4.4: An experiment where a fixed movie is shown to a person while 10 EEG channels record signals for 5 seconds at 2 KHz.

4.5: An experiment like 4.4 except where only the time-averaged signals are kept.

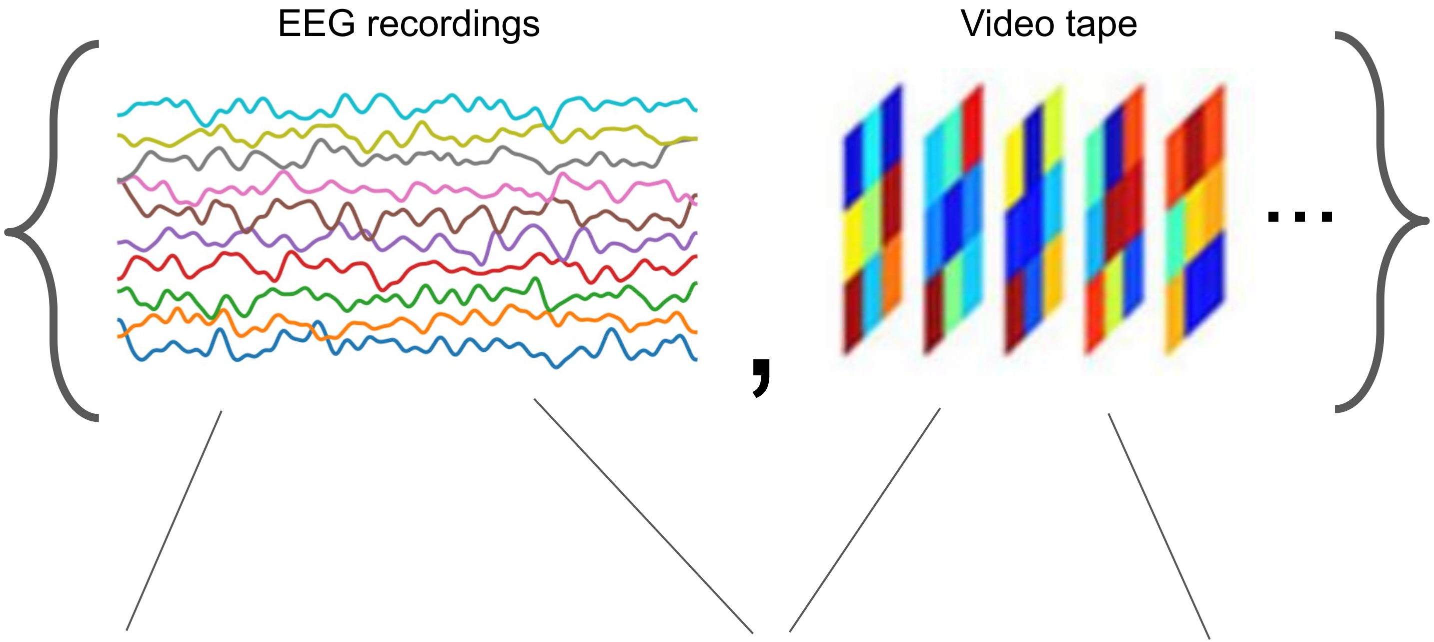
4.6: An experiment where 10 EEG channels are recorded for 5 seconds at 2 KHz along with a 5-second video tape of the person (1280 x 720 pixels, 30 fps).

4.7: A network of N neurons with random, directional connections (including self-connections).

4.8: An experiment that measures the *time-averaged* firing rate of N neurons over T seconds.

Visualizing problem 4.6

4.6: An experiment where 10 EEG channels are recorded for 5 seconds at 2 KHz along with a 5-second video tape of the person (1280 x 720 pixels, 30 fps).



Complete random object

EEG recordings

Video tape

...

Vector representation: [1.3512, 3.134125, ..., 6.9784, 1.6328, ..., 8.23166]

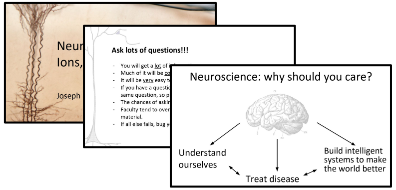
A powerful perspective

EVERYTHING comes from random processes.

Number of open ion channels, spike times, visual stimulus, behavior sequence.

~ P(image)

Book, movie, scientific lecture, seating arrangement.



Neur
Ions,

Joseph

Ask lots of questions!!!

- You will get a lot of i
- Much of it will be co
- It will be very easy t
- If you have a questio
  same question, so p
- The chances of askin
- Faculty tend to over
  material.
- If all else fails, bug y

Neuroscience: why should you care?

Understand ourselves

Treat disease

Build intelligent systems to make the world better

~ P(lecture)

Brain, animal, life, society, political turbulence.

~ P(political meeting)

Joint distributions

Distribution over *all* possible outcomes of random process.

2-D discrete joint distribution

x_i

y_j

	1	2	3	4
1	.07	.1	.04	.04
2	.04	.04	.07	.04
3	.07	.04	.04	.1
4	.1	.1	.07	.04

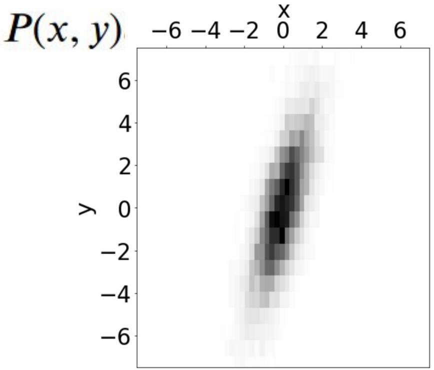
$P(x_i, y_j)$

N-D discrete joint distribution

$$P(x_i, y_j, \dots) \equiv P(\mathbf{r}_i)$$

$$\sum_i P(\mathbf{r}_i) = 1$$

2-D continuous joint distribution



N-D continuous joint distribution

$$P(x, y, \dots) \equiv P(\mathbf{r})$$

$$\int P(\mathbf{r}) d\mathbf{r} = 1$$

Multivariate expectation

Collection of samples = ***ensemble***.

[2.2381..., 9.3214..., ..., -1.6983...]

[5.9362..., -2.5721..., ..., 8.3702...]

⋮

[-1.0273..., 1.2888..., ..., 3.0382...]

[-2.1174..., 2.9733..., ..., 4.0012...]

Mean (expectation)

[3.7134..., 1.3321..., ..., 5.3107...]

Expectation of a random process = mean over *ensemble*.

$$E[\mathbf{r}] = \sum_i \mathbf{r}_i P(\mathbf{r}_i) \qquad E[\mathbf{r}] = \int \mathbf{r} P(\mathbf{r}) d\mathbf{r}$$

Dimension of expectation = dimension of a sample.

Covariance and correlation

Ensemble

[2.2381..., 9.3214..., ..., -1.6983...]

[5.9362..., -2.5721..., ..., 8.3702...]

⋮

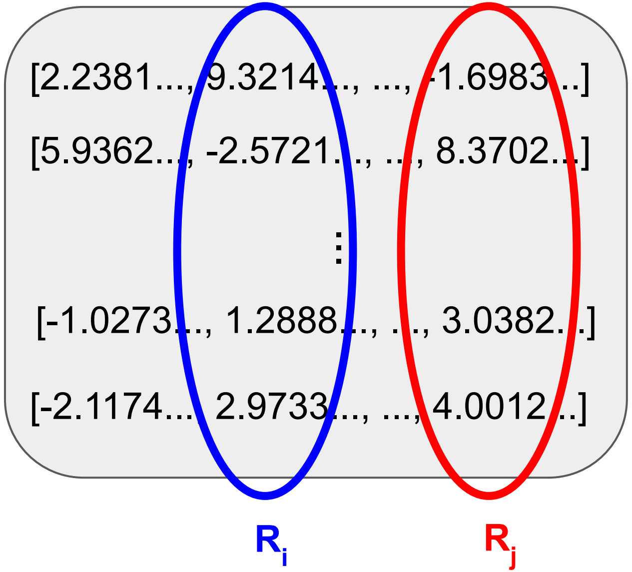
[-1.0273..., 1.2888..., ..., 3.0382...]

[-2.1174..., 2.9733..., ..., 4.0012...]

Covariance tells us how much two elements R_i and R_j of a random vector vary *together*.

Covariance and correlation

Ensemble



[2.2381..., 9.3214..., ..., -1.6983...]

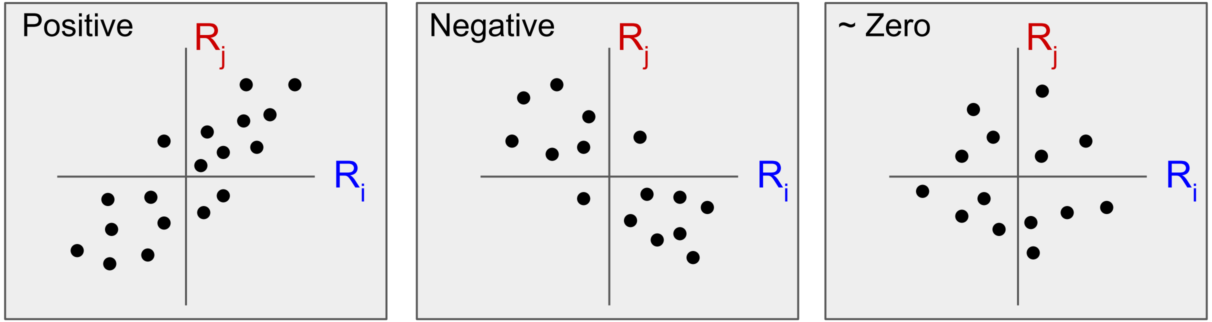
[5.9362..., -2.5721..., ..., 8.3702...]

⋮

[-1.0273..., 1.2888..., ..., 3.0382...]

[-2.1174..., 2.9733..., ..., 4.0012...]

Covariance tells us how much two elements R_i and R_j of a random vector vary *together*.

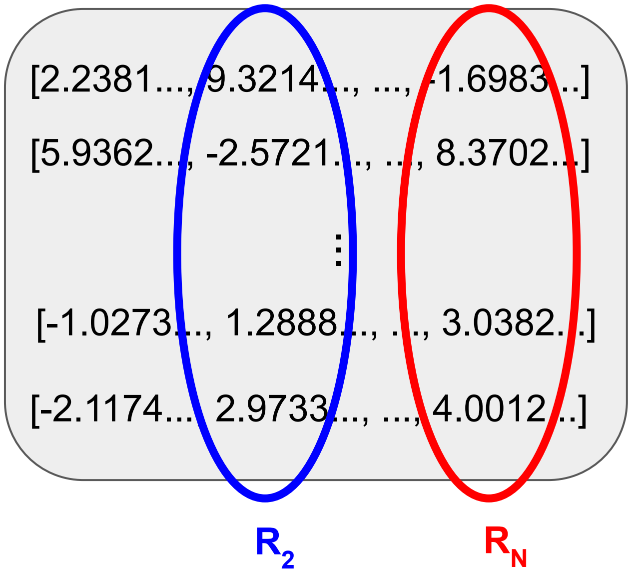


$$\mathrm{Cov}[R_i, R_j] = \mathrm{E}[(R_i - \mathrm{E}[R_i])(R_j - \mathrm{E}[R_j])]$$

Why is this a good definition of covariance?

Covariance and correlation

Ensemble



[2.2381..., 9.3214..., ..., -1.6983...]

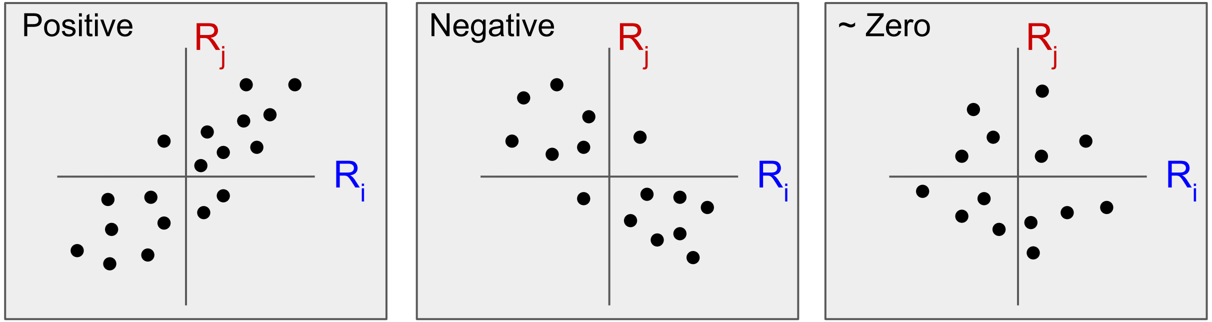
[5.9362..., -2.5721..., ..., 8.3702...]

⋮

[-1.0273..., 1.2888..., ..., 3.0382...]

[-2.1174..., 2.9733..., ..., 4.0012...]

Covariance tells us how much two elements R_i and R_j of a random vector vary *together*.



$$\mathrm{Cov}[R_i, R_j] = \mathrm{E}[(R_i - \mathrm{E}[R_i])(R_j - \mathrm{E}[R_j])]$$

Correlation is normalized covariance:

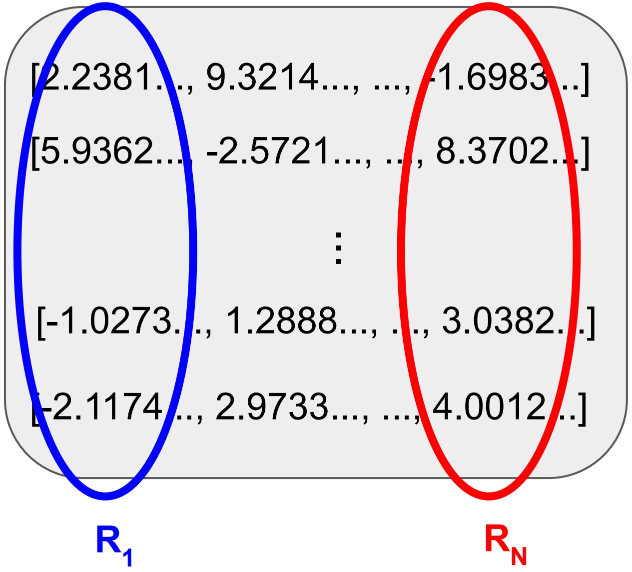
$$\mathrm{Corr}[R_i, R_j] = \frac{\mathrm{Cov}[R_i, R_j]}{\sqrt{\mathrm{Var}[R_i]\mathrm{Var}[R_j]}}$$

REMEMBER: Covariance is always between ***two*** elements of a random vector.

Covariances across all pairs of elements of an N-D vector are summarized in an N x N ***covariance matrix***.

Covariance and correlation

Ensemble



[2.2381..., 9.3214..., ..., -1.6983...]

[5.9362..., -2.5721..., ..., 8.3702...]

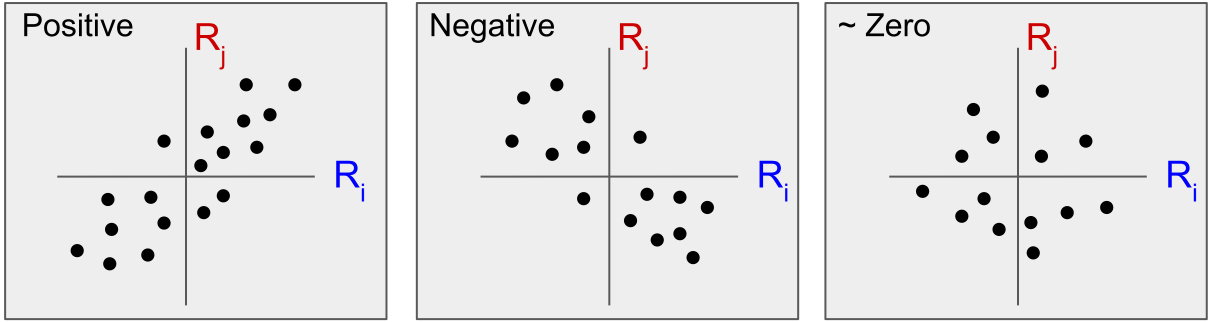
⋮

[-1.0273..., 1.2888..., ..., 3.0382...]

[-2.1174..., 2.9733..., ..., 4.0012...]

R_1 R_N

Covariance tells us how much two elements R_i and R_j of a random vector vary *together*.



$$\mathrm{Cov}[R_i, R_j] = \mathrm{E}[(R_i - \mathrm{E}[R_i])(R_j - \mathrm{E}[R_j])]$$

Correlation is normalized covariance:

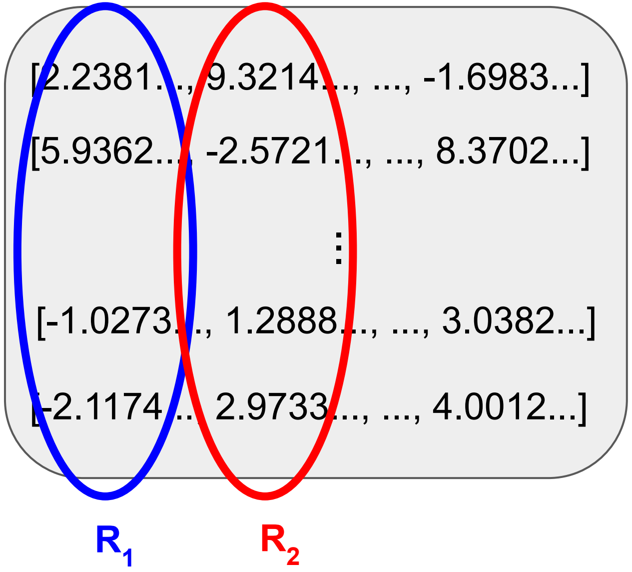
$$\mathrm{Corr}[R_i, R_j] = \frac{\mathrm{Cov}[R_i, R_j]}{\sqrt{\mathrm{Var}[R_i]\mathrm{Var}[R_j]}}$$

REMEMBER: Covariance is always between ***two*** elements of a random vector.

Covariances across all pairs of elements of an N-D vector are summarized in an N x N ***covariance matrix***.

Covariance and correlation

Ensemble



[2.2381..., 9.3214..., ..., -1.6983...]

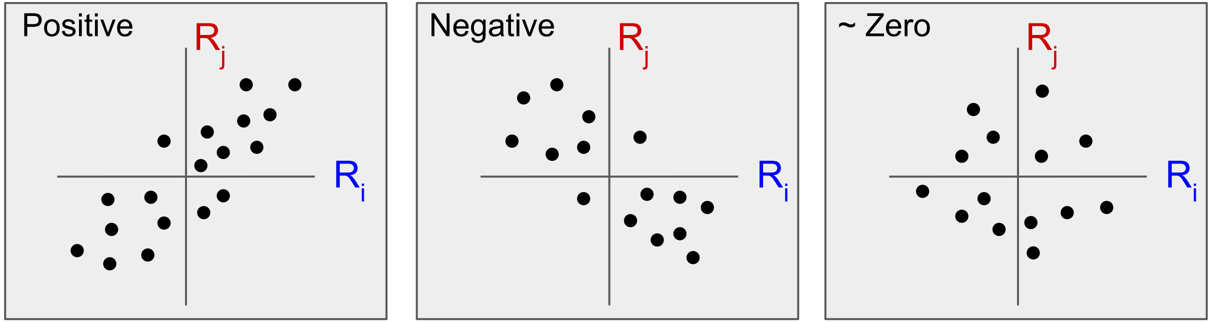
[5.9362..., -2.5721..., ..., 8.3702...]

⋮

[-1.0273..., 1.2888..., ..., 3.0382...]

[-2.1174..., 2.9733..., ..., 4.0012...]

Covariance tells us how much two elements R_i and R_j of a random vector vary *together*.



$$\mathrm{Cov}[R_i, R_j] = \mathrm{E}[(R_i - \mathrm{E}[R_i])(R_j - \mathrm{E}[R_j])]$$

Correlation is normalized covariance:

$$\mathrm{Corr}[R_i, R_j] = \frac{\mathrm{Cov}[R_i, R_j]}{\sqrt{\mathrm{Var}[R_i]\mathrm{Var}[R_j]}}$$

REMEMBER: Covariance is always between ***two*** elements of a random vector.

Covariances across all pairs of elements of an N-D vector are summarized in an N x N ***covariance matrix***.

Marginal distributions

Distribution over *subset* of variables from multivariate random process

$$P(x_i) = \sum_j P(x_i, y_j)$$

$$P(x) = \int_{-\infty}^{\infty} P(x, y)dy$$

$P(x_i, y_j)$ joint

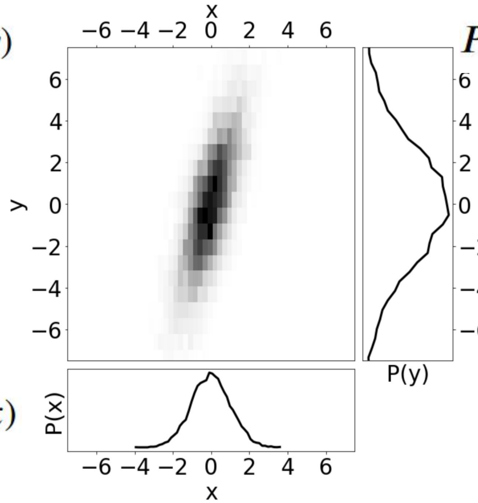
x_i

y_j

	1	2	3	4	$P(y_j)$ marginal
1	.07	.1	.04	.04	.25
2	.04	.04	.07	.04	.19
3	.07	.04	.04	.1	.25
4	.1	.1	.07	.04	.31
marginal $P(x_i)$	.28	.28	.22	.22	

$P(x, y)$ joint

$P(y)$ marginal



$P(x)$ marginal

$$P(x_i) = \sum_{j,k,\ldots} P(x_i, y_j, z_k, \ldots)$$

$$P(x_i, y_j) = \sum_{k,l\ldots} P(x_i, y_j, z_k, w_l, \ldots)$$

$$P(x) = \int_{-\infty}^{\infty} P(x, y, z \ldots)dydz\ldots$$

$$P(x, y) = \int_{-\infty}^{\infty} P(x, y, z, w \ldots)dzdw\ldots$$

Independence

N random variables are *independent* if their joint distribution is the product of their marginals:

x, y, z independent:

$$P(x, y, z) = P(x)P(y)P(z)$$

x, y, z not independent:

$$P(x, y, z) \neq P(x)P(y)P(z)$$

Problems

Given the following joint distribution over the spike counts of two neurons in 1 second window in response to stimulus **x**:

		Spike count A			
		0	1	2	3
Spike count B	0	.04	.1	.04	.04
	1	.04	.07	.07	.04
	2	.1	.04	.07	.1
	3	.07	.1	.04	.04

$P(a, b)$

5.1: What is the probability that neuron A spikes 2 times and neuron B spikes 3 times?

5.2: What is the probability that A spikes only once?

5.3: What is the expected spike count of A?

5.4: What is the expected spike count of B?

5.5: What is the expected spike count of A and B (a 2-D vector)?

5.6: Write out the full equation for the covariance of A and B.

5.7: What is the covariance of A’s and B’s spike counts?

5.8: What is the correlation of A’s and B’s spike counts?

5.9: Do A and B spike independently? Why or why not?

Problems

6.1: Suppose P(**r**) is the distribution of a firing rate vector **r** describing the firing rates of 50 neurons. How many elements does E[**r**] have?

6.2: Write out the explicit covariance formula for two continuous random variables.

6.3: What is the size of the covariance matrix of a random image of size 32 px by 32 px?

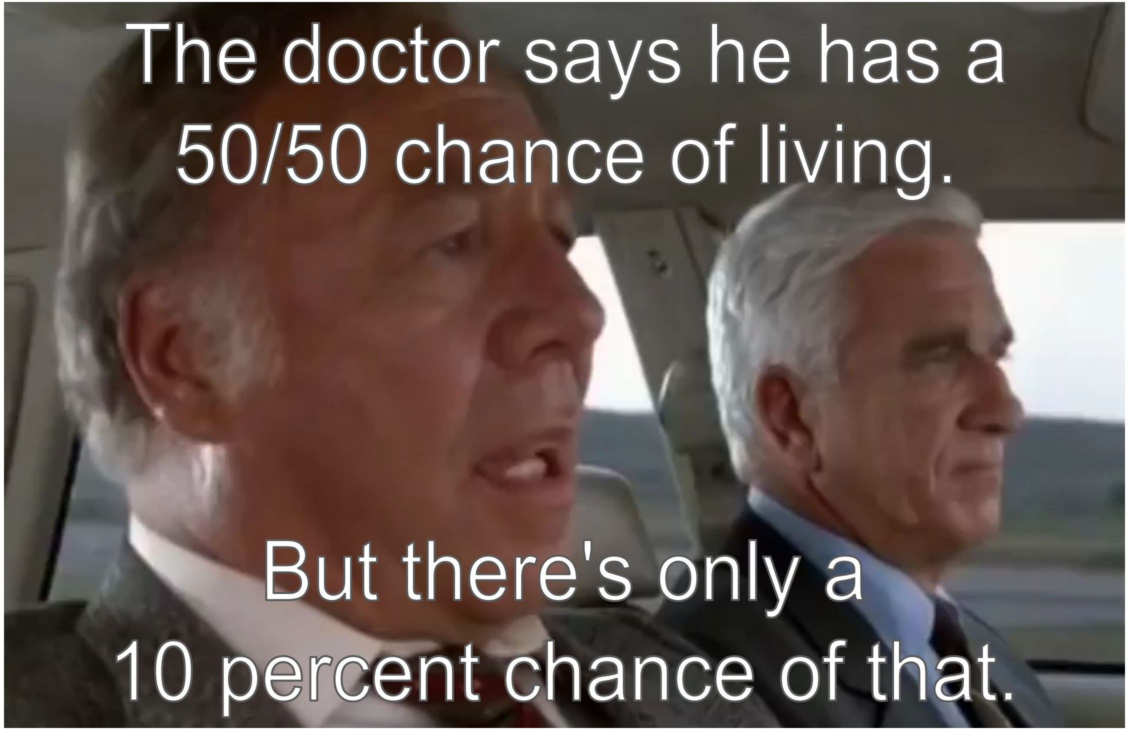
6.4: Each row in the dataset firing_rates.npy contains a 50-D sample of neural firing rates. There are 300 rows, corresponding to 300 trials in the experiment. Calculate and plot μ, the mean firing rate for each neuron across trials, and the covariance matrix K between every pair of neurons across all repetitions of the experiment using the following estimators:

$$\hat{\mu}_i = \frac{1}{N}\sum_{a=1}^{N} x_a^i \qquad \hat{K}_{i,j} = \frac{1}{N-1}\sum_{a=1}^{N}(x_a^i - E[X^i])(x_a^j - E[Y^j])$$

You can load the data into the variable X using:

```
X = np.load('firing_rates.npy', allow_pickle=True)[0]['data']
```

Conditional probabilities



Conditional probabilities

What is the probability of X when Y is *given* (fixed at specific value)?

$$P(\underset{\text{unknown}}{x}|\underset{\text{given}}{y})$$

$$\boxed{P(x|y) \neq P(y|x)}$$

$$\sum_i P(X = x_i|Y) = 1$$

$$\int_{-\infty}^{\infty} P(X = x|Y)dx = 1$$

Neurons have correlated firing rates. What is probability of neuron A's spike count X given neuron B's spike count Y?

$P(x, y)$ joint

x

y	.07	.1	.04	.04	Y = 1
	.04	.04	.07	.04	Y = 2
	.07	.04	.04	.1	Y = 3
	.1	.1	.07	.04	Y = 4
$P(x)$	.28	.28	.22	.22	Marginal

Conditional probabilities

What is the probability of X when Y is *given* (fixed at specific value)?

$$P(x|y)$$

unknown

given

$$\boxed{P(x|y) \neq P(y|x)}$$

$$\sum_i P(X = x_i|Y) = 1$$

$$\int_{-\infty}^{\infty} P(X = x|Y)dx = 1$$

Neurons have correlated firing rates. What is probability of neuron A’s spike count X given neuron B’s spike count Y?

	x				
$P(x, y)$ joint	.07	.1	.04	.04	Y = 1
y	.04	.04	.07	.04	Y = 2
	.07	.04	.04	.1	Y = 3
	.1	.1	.07	.04	Y = 4
$P(x)$	.28	.28	.22	.22	Marginal
$P(x \mid Y = 1)$	.28	.4	.16	.16	X given Y=1

Conditional probabilities

What is the probability of X when Y is *given* (fixed at specific value)?

$$P(\underset{\text{unknown}}{x}|\underset{\text{given}}{y})$$

$$\boxed{P(x|y) \neq P(y|x)}$$

$$\sum_i P(X = x_i|Y) = 1$$

$$\int_{-\infty}^{\infty} P(X = x|Y)dx = 1$$

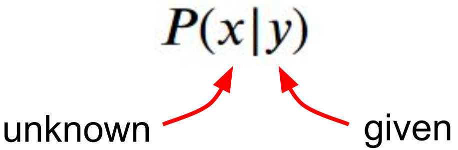
Neurons have correlated firing rates. What is probability of neuron A’s spike count X given neuron B’s spike count Y?

x

$P(x, y)$ joint	.07	.1	.04	.04	Y = 1
	.04	.04	.07	.04	Y = 2
y	.07	.04	.04	.1	Y = 3
	.1	.1	.07	.04	Y = 4
$P(x)$	.28	.28	.22	.22	Marginal
$P(x \mid Y = 1)$	.28	.4	.16	.16	X given Y=1
$P(x \mid Y = 3)$	.28	.16	.16	.4	X given Y=3

Conditional probabilities

What is the probability of X when Y is *given* (fixed at specific value)?



$$P(x|y)$$

$$\boxed{P(x|y) \neq P(y|x)}$$

$$\sum_i P(X = x_i|Y) = 1$$

$$\int_{-\infty}^{\infty} P(X = x|Y)dx = 1$$

Neurons have correlated firing rates. What is probability of neuron A's spike count X given neuron B's spike count Y?

	x				
$P(x, y)$ joint	.07	.1	.04	.04	Y = 1
	.04	.04	.07	.04	Y = 2
y	.07	.04	.04	.1	Y = 3
	.1	.1	.07	.04	Y = 4
$P(x)$	.28	.28	.22	.22	Marginal
$P(x\|Y = 1)$	.28	.4	.16	.16	X given Y=1
$P(x\|Y = 3)$	.28	.16	.16	.4	X given Y=3

$$P(x, y) = P(x|y)P(y) = P(y|x)P(x)$$

The number two way to prevent confusion

Always make sure you know what variable(s) your distribution is over.

$$\overbrace{P(x, y)}^{\text{Over X and Y together.}} = \underbrace{P(x|y)}_{\text{Over X.}}\overbrace{P(y)}^{\text{Over Y.}} = \underbrace{P(y|x)}_{\text{Over Y.}}\overbrace{P(x)}^{\text{Over X.}}$$

Useful things about conditional probabilities

$$P(\underbrace{x_1, x_2, \ldots}_{\text{unknown}} | \underbrace{y_1, y_2, \ldots}_{\text{given (fixed)}})$$

The chain rule for 2 variables:

$$P(x, y) = P(x|y)P(y) = P(y|x)P(x)$$

Marginal probabilities from conditional probabilities

$$P(x) = \sum_j P(x, Y = y_j) = \sum_j P(x|Y = y_j)P(Y = y_j)$$

Correct and incorrect normalization

$$\sum_i P(X = x_i|y_j) = 1$$

RIGHT

$$\int_{-\infty}^{\infty} P(X = x|y)dx = 1$$

$$\sum_j P(x|Y = y_j) = 1$$

WRONG

$$\int_{-\infty}^{\infty} P(x|Y = y)dy = 1$$

The chain rule for N variables:

$$P(x_N, x_{N-1}, \ldots, x_1) = P(x_N|x_{N-1}, \ldots, x_1)P(x_{N-1}|x_{N-2}, \ldots, x_1)\ldots$$

$$\ldots P(x_2|x_1)P(x_1)$$

Bayes' Theorem

Gives relationship between conditional distributions.

$$\boxed{P(x|y) \neq P(y|x)}$$

$$P(x, y) = P(x|y)P(y) = P(y|x)P(x)$$

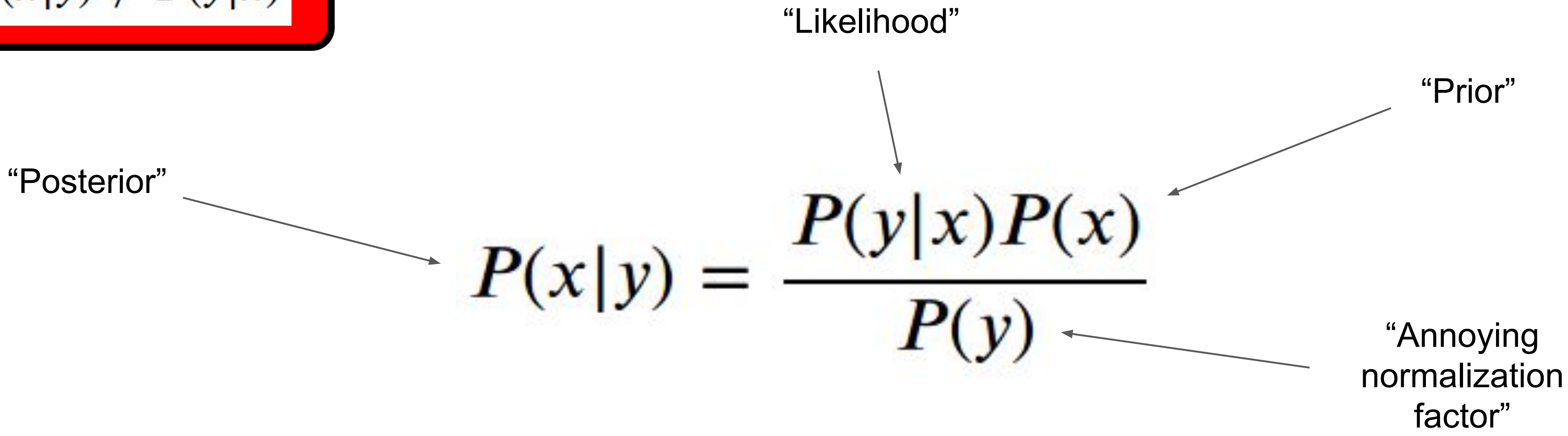
$$\Downarrow$$

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

Bayes’ Theorem

Gives relationship between conditional distributions.

$$\boxed{P(x|y) \neq P(y|x)}$$



“Likelihood”

“Prior”

“Posterior”

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

“Annoying normalization factor”

Tells us how to combine *new observation* with *prior knowledge*.

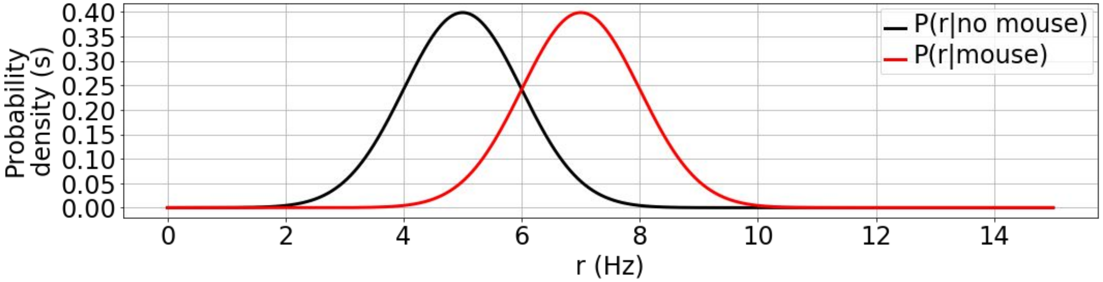
$$P(x) \ \& \ y \Longrightarrow P(x|y)$$

Bayes' Theorem

$$\boxed{P(x|y) \neq P(y|x)}$$

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

Suppose neuron in cat visual cortex is sensitive to presence of mouse:



What is the probability there is a mouse if the neuron fires at 7 Hz?

P(R = 7 Hz|mouse) = 0.4

P(R = 7 Hz|no mouse) = 0.05

P(mouse) = 0.01

P(no mouse) = 0.99

$$P(mouse|R = 7Hz) = \frac{P(R = 7Hz|mouse)P(mouse)}{P(R = 7Hz)} = \frac{0.4(0.01)}{P(R = 7Hz)}$$

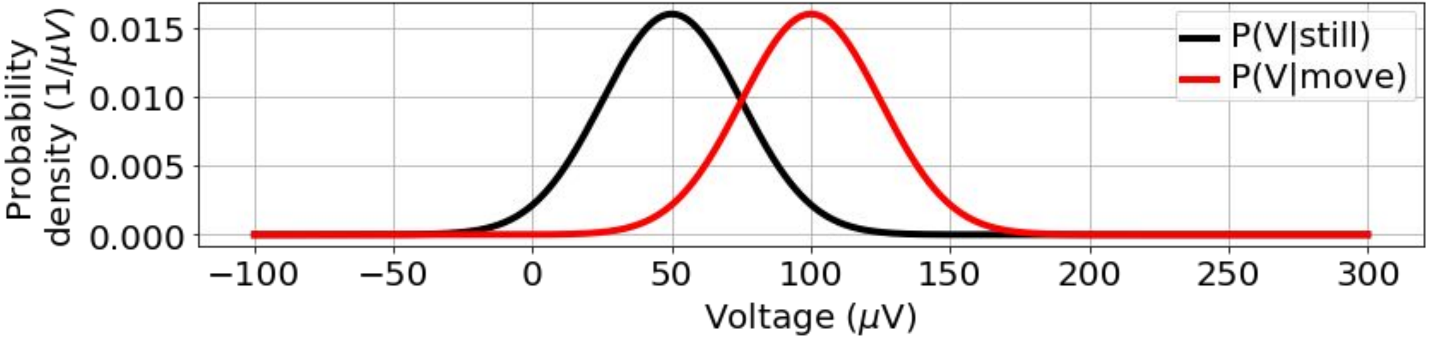
Annoying normalization constant

$$P(R = 7Hz) = \sum_i P(R = 7Hz|M = m_i)P(M = m_i)$$
$$= P(R = 7Hz|mouse)P(mouse) + P(R = 7Hz|no\ mouse)P(no\ mouse)$$
$$= .4(.01) + .05(.99) = .0535$$

$$P(mouse|R = 7Hz) = \frac{0.4(0.01)}{0.0535} \approx \boxed{.075} \ll 0.4$$

Problems

7.1: **Motor signals**: You want to decode whether someone moves their arm based on a 1-D time-averaged EEG signal. When the person moves their arm, the EEG signal follows a Gaussian distribution with mean 100 µV and stdev 25 µV. When they don't move their arm the EEG signal follows a Gaussian with mean 50 µV and stdev 25 µV. Without knowing the EEG signal, there is a 10% chance the person moves their arm. If the EEG signal is 75 µV, what is the probability they moved their arm? What if the EEG signal is 100 µV?



7.2: **Image detection**: Suppose a neuron in IT cortex in a monkey spikes with probability 0.8 in response to a picture of a cat and with probability 0.2 in response to a picture of a dog. If 25% of images shown are cats and 75% are dogs, what is the probability the image was a cat if the neuron spikes? What is the probability the image was a cat if the neuron doesn't spike? What if 50% of images shown are cats and 50% are dogs?

Bayes' Theorem

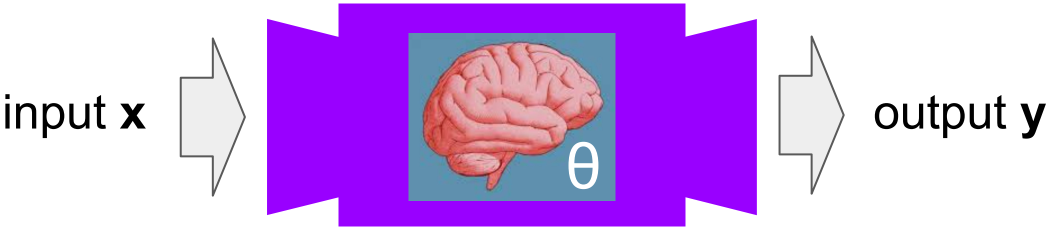
When three variables are involved:

$$P(x|y,z) = \frac{P(y|x,z)P(x|z)}{P(y|z)}$$

More generally, many ways to apply Bayes Law:

$$P(x,y,z,w) \left\{ \begin{array}{l} P(x|y,z,w) = \dfrac{P(y|x,z,w)P(x|z,w)}{P(y|z,w)} \\[2em] P(x,w|z,y) = \dfrac{P(z|x,w,y)P(x,w|y)}{P(z|y)} \\[2em] \qquad\qquad\qquad\vdots \end{array} \right.$$

Probabilistic models



input **x**

output **y**

Choose model for conditional probability distribution specifying output **y** given **x** and parameters θ.

$$\mathbf{y} \sim P(\mathbf{y}|\mathbf{x};\theta)$$

Goal: given example(s) of {**x, y**} measured from data, find θ.

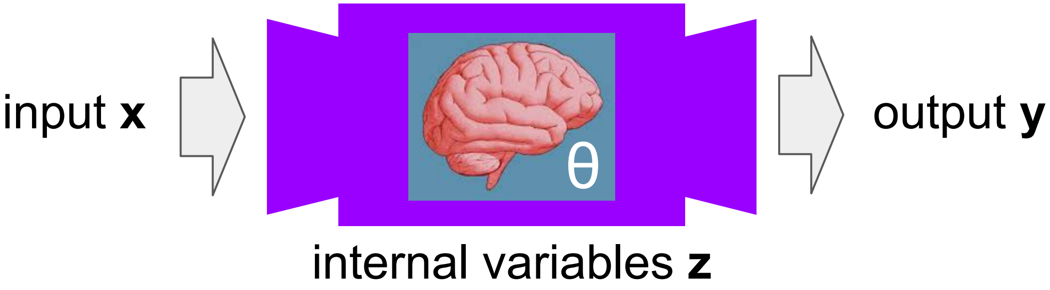
A common approach is to maximize the *likelihood* P({**y**}|{**x**}, θ):

$$\hat{\theta} = \mathrm{argmax}_{\theta} P(\{\mathbf{y}\}|\{\mathbf{x}\},\theta)$$

$$P(\{\mathbf{y}\}|\{\mathbf{x}\},\theta) =$$

$$P(\mathbf{y}_1|\mathbf{x}_1,\theta)P(\mathbf{y}_2|\mathbf{x}_2,\theta)\ldots P(\mathbf{y}_N|\mathbf{x}_N,\theta)$$

Latent variable models



Assume there are also hidden (latent) variables **z**.

Goal: learn both θ (*same* for all trials) and **z** (*different* in each trial).

Solution: iterate between learning θ and infering {**x**} via *expectation-maximization* algorithm.
https://en.wikipedia.org/wiki/Expectation-maximization_algorithm

---

Common example: Hidden Markov Model

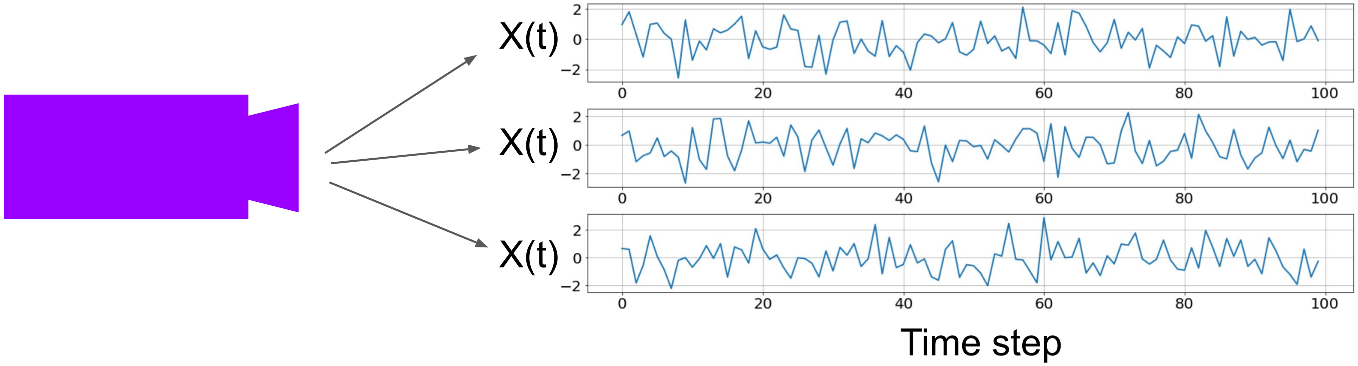
Hidden state: $\mathbf{z}_\mathbf{t} \sim P(\mathbf{z}_t|\mathbf{z}_{t-1}, \theta)$

θ = {transition probabilities among **z**, probability of **y** given **z**}

Observation: $\mathbf{y}_\mathbf{t} \sim P(\mathbf{y}_t|\mathbf{z}_t, \theta)$

Time series

We often model time-series as samples from a random process.



Random time-series are just random T-length vectors

(T = # time steps)

Mean = average across ensemble *as function of time.*

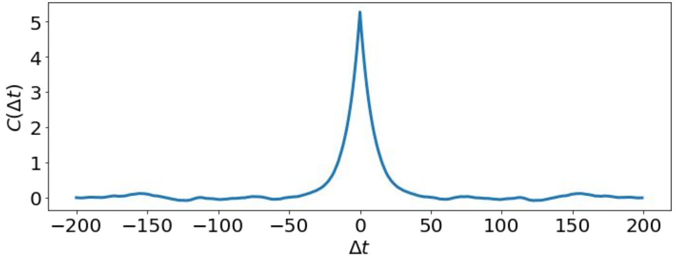
$$E[X_t] = \sum_i x_t^i P(\mathbf{x}^i)$$

Time average ≠ ensemble average.

Time average of sample = 1 number.
Ensemble average = 1 time series of length T.

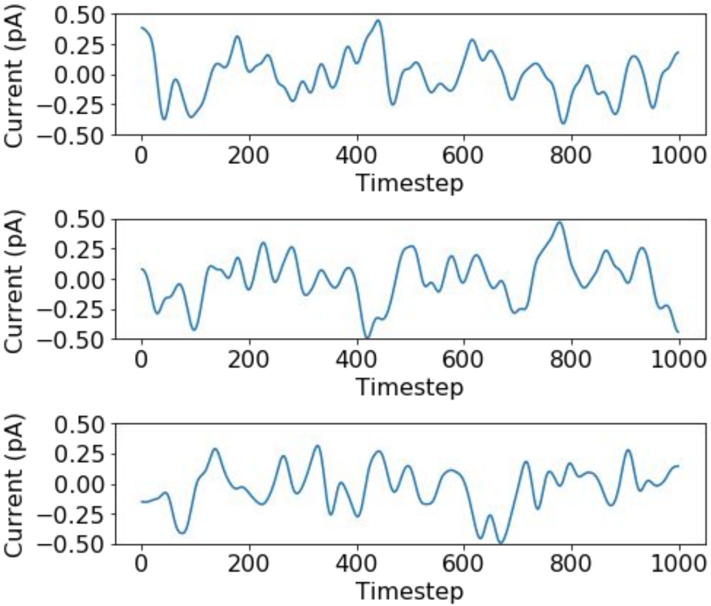
Autocovariance = covariance between values at two time points separated by time lag Δt:

$$C(\Delta t) = \mathrm{Cov}[X(t), X(t + \Delta t\Delta t)]$$

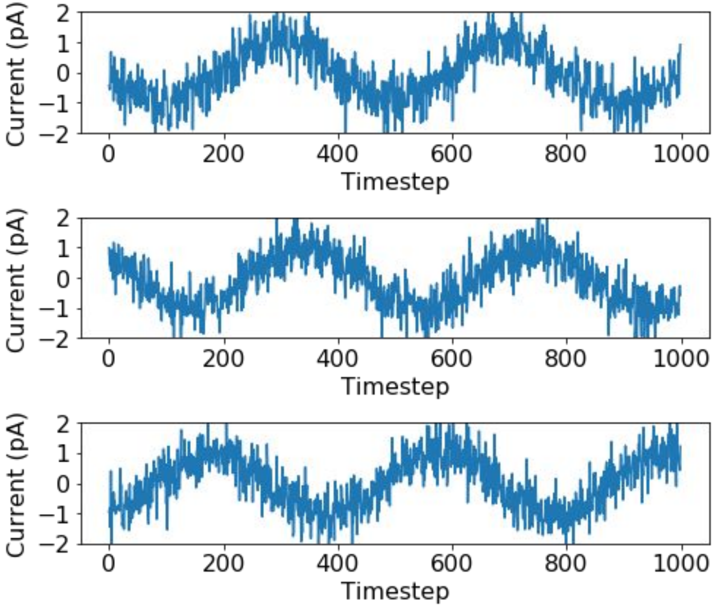


Problems

8.1: If a random process yields the following samples, is C(Δt) positive, negative, or near zero for (A) Δt = 5, (B) Δt = 200, (C) Δt = 400?



8.2: If a random process yields the following samples, is C(Δt) positive, negative, or near zero for (A) Δt = 5, (B) Δt = 200, (C) Δt = 400?

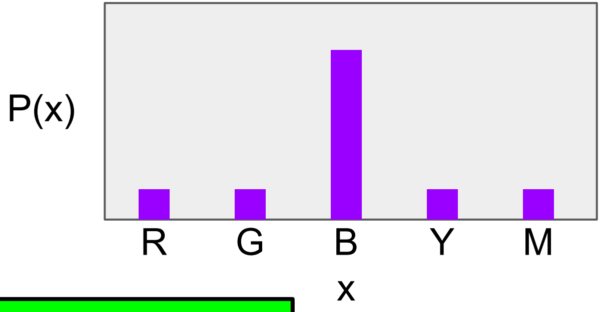


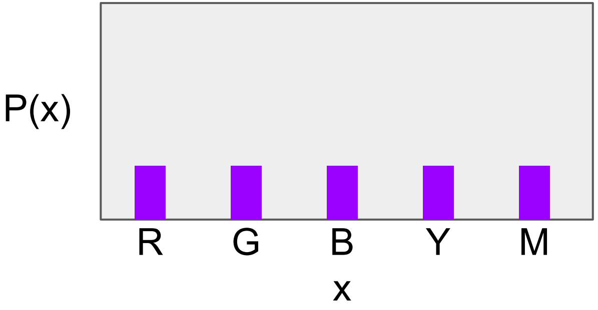
8.3 Generate 100 1000-timestep samples from the following random process and compute and plot its autocovariance function for -200 < Δt < 200.

$$X_t = .9X_{t-1} + \eta \qquad \eta \sim Gaussian(0, 1); X_0 = 0$$

Entropy and mutual information

Entropy (H) specifies how spread out or uncertain a probability distribution is.


Low entropy


High entropy

If discrete, doesn't depend on units or ordering of possible values of x.

$$H[X] = -\mathrm{E}[\log P(x)] = -\sum_i p(x_i) \log p(x_i)$$

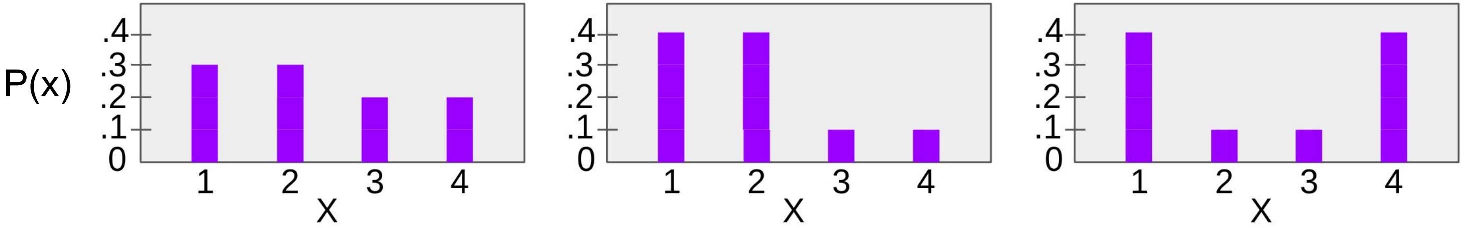
Mutual information between X and Y quantifies, on average, how much knowing value of Y decreases entropy of X.

$$MI[X;Y] = H[X] - \mathrm{E}[H[X|Y]] = H[X] - \sum_j H[X|Y = y_j]P(y_j)$$

Problems

9.1: What is the entropy of a probability mass function over N equally probable outcomes?

9.2: Calculate the entropy of the following distributions over a spike count X:



9.3: Rank the following distributions from lowest to highest entropy:

(1) Outcome of six-sided die roll, (2) Unbiased coin flip, (3) random sequence of 20 English letters, (4) random selection of word from the dictionary.

9.4: Rank the following distributions from lowest to highest entropy:

(1) Temperature on random day of the year in Muizenberg, (2) Temperature on random day in the summer in Muizenberg, (3) Temperature in random city on random day of the year.

9.5: Rank the following distributions from lowest to highest entropy:

(1) Spike count of 1 neuron in a 1-second window, (2) Spike counts of N identical independent neurons in a 1-second window, (3) Spike counts of N identical but independent neurons in a 1-second window.

9.6: Calculate the mutual information between X_1 and X_2 from the following joint probability distributions.

A

		X_1	
		0	1
X_2	0	.15	.4
	1	.4	.05

B

		X_1	
		0	1
X_2	0	.08	.12
	1	.32	.48

Part 2 Recap

Multivariate random objects

Covariance and correlation

Joint and marginal distributions

Conditional distributions and Bayes' Law

Probabilistic models

Time-series

Entropy and information theory

British Judge Rules Against Bayes’s Theorem

BY BRIGGS OCTOBER 14, 2011

A British judge has thrown a use of Bayes’s rule out of his court. Not only that, his honor (Lordship?) ruled “against using similar statistical analysis in the courts in future.”

A ruling to which this dedicated Bayesian says, “Hear, hear!”

https://wmbriggs.com/post/4468/