Introduction to Probabilistic Graphical Models

Christoph Lampert

IST Austria (Institute of Science and Technology Austria)

Schedule

Refresher of Probabilities
Introduction to Probabilistic Graphical Models
Probabilistic Inference
Learning Conditional Random Fields
MAP Prediction / Energy Minimization
Learning Structured Support Vector Machines

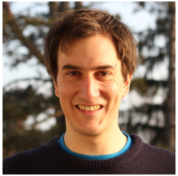
Links to slide download: `http://pub.ist.ac.at/~chl/courses/PGM_W16/`

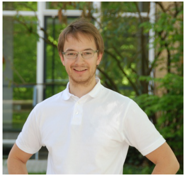
Password for ZIP files (if any): `pgm2016`

Email for questions, suggestions or typos that you found: `chl@ist.ac.at`

Thanks to . . .


Sebastian Nowozin


Peter Gehler


Andreas Geiger

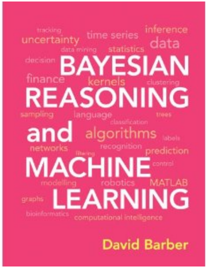

Björn Andres


Raquel Urtasun

and David Barber, author of textbook **"Bayesian Reasoning and Machine Learning"**
`http://web4.cs.ucl.ac.uk/staff/D.Barber/pmwiki/pmwiki.php?n=Brml.HomePage`

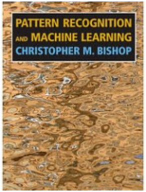
for material and slides.

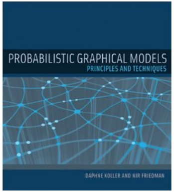
Textbooks on Graphical Models

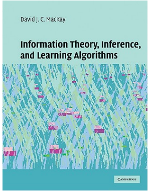


- David Barber, **Bayesian Reasoning and Machine Learning**, Cambridge University Press, 2011, ISBN-13: 978-0521518147
- Available online **for free**: `http://tinyurl.com/3flppuo`

For the curious ones...

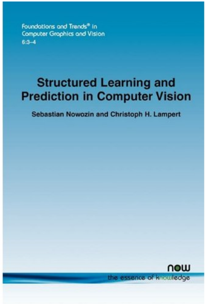






- Bishop, **Pattern Recognition and Machine Learning**, Springer New York, 2006, ISBN-13: 978-0387310732
- Koller, Friedman, **Probabilistic Graphical Models: Principles and Techniques**, The MIT Press, 2009, ISBN-13: 978-0262013192
- MacKay, **Information Theory, Inference and Learning Algorithms**, Cambridge University Press, 2003, ISBN-13: 978-0521642989

Older tutorial...



Parts published in

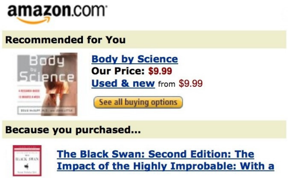
- Sebastian Nowozin, Chrsitoph H. Lampert, **"Structured Learning and Prediction in Computer Vision"**, Foundations and Trends in Computer Graphics and Vision, now publisher, `http://www.nowpublishers.com/`
- available as PDF on my homepage

Introduction

Success Stories of Machine Learning

Robotics

Time Series Prediction



Social Networks

Language Processing

Healthcare

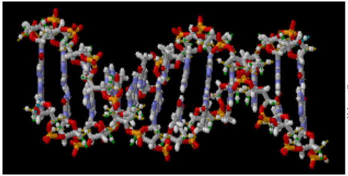
Natural Sciences

All of these require dealing with Structured Data

This course is about **modeling structured data**. . .

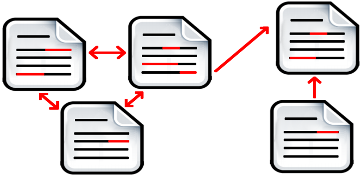
Jemand musste Josef K. verleumdet haben, denn ohne dass er etwas Böses getan hätte, wurde er eines Morgens verhaftet. »Wie ein Hund! « sagte er, es war, als sollte die Scham ihn überleben. Als Gregor Samsa eines Morgens aus unruhigen Träumen erwachte, fand er sich in seinem Bett zu einem ungeheueren Ungezieferverwandelt. Und es war ihnen wie eine Bestätigung ihrer neuen Träume und guten Absichten, als am Ziele ihrer Fahrt die Tochter als erste sich erhob und ihren jungen Körper dehnte. »Es ist ein eigentümlicher Apparat«, sagte der Offizier zu dem Forschungsreisenden und überblickte mit einem gewissermaßen

Text

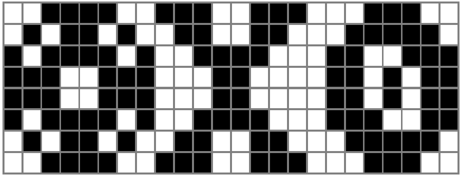


Source: wikipedia.org

Molecules / Chemical Structures



Documents/HyperText

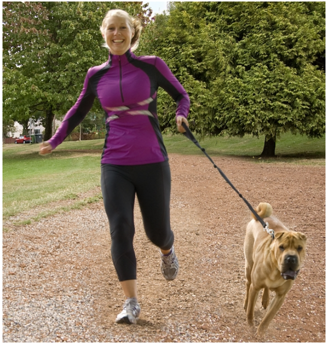


Images

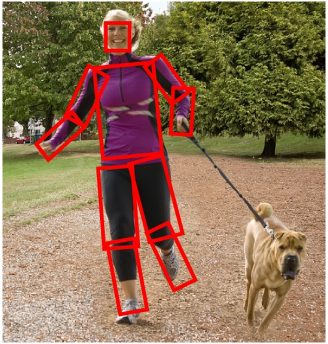
. . . and about **predicting structured data**:

- Natural Language Processing:
  - Automatic Translation (output: sentences)
  - Sentence Parsing (output: parse trees)
- Bioinformatics:
  - Secondary Structure Prediction (output: bipartite graphs)
  - Enzyme Function Prediction (output: path in a tree)
- Speech Processing:
  - Automatic Transcription (output: sentences)
  - Text-to-Speech (output: audio signal)
- Robotics:
  - Planning (output: sequence of actions)
- Computer Vision:
  - Human Pose Estimation (output: locations of body parts)
  - Image Segmentation (output: segmentation mask)

Example: Human Pose Estimation



$x \in \mathcal{X}$



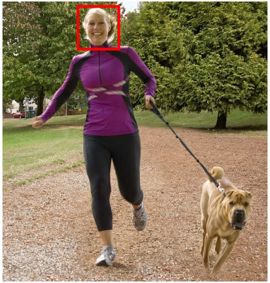
$y \in \mathcal{Y}$

- Given an image, where is a person and how is it articulated?

$$f : \mathcal{X} \to \mathcal{Y}$$

- Image x, but what is human pose $y \in \mathcal{Y}$ precisely?

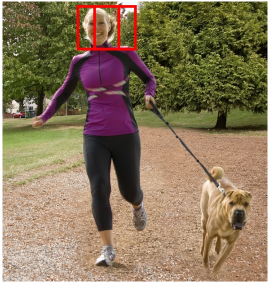
Human Pose $\mathcal{Y}$



Example y_{head}

- Body Part: $y_{head} = (u, v, \theta)$ where (u, v) center, θ rotation
  - $(u, v) \in \{1, \ldots, M\} \times \{1, \ldots, N\}, \theta \in \{0, 45^\circ, 90^\circ, \ldots\}$

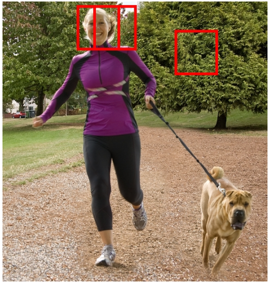
Human Pose $\mathcal{Y}$



Example y_{head}

- Body Part: $y_{head} = (u, v, \theta)$ where (u, v) center, θ rotation
  - $(u, v) \in \{1, \dots, M\} \times \{1, \dots, N\}, \theta \in \{0, 45^\circ, 90^\circ, \dots\}$

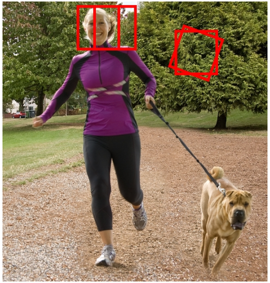
Human Pose $\mathcal{Y}$



Example y_{head}

- Body Part: $y_{head} = (u, v, \theta)$ where (u, v) center, θ rotation
  - $(u, v) \in \{1, \ldots, M\} \times \{1, \ldots, N\}, \theta \in \{0, 45^\circ, 90^\circ, \ldots\}$

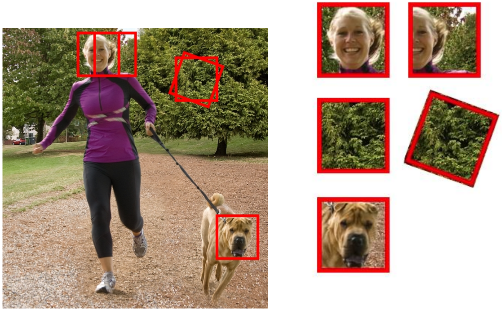
Human Pose $\mathcal{Y}$



Example y_{head}

- Body Part: $y_{head} = (u, v, \theta)$ where (u, v) center, θ rotation
  - $(u, v) \in \{1, \ldots, M\} \times \{1, \ldots, N\}, \theta \in \{0, 45^\circ, 90^\circ, \ldots\}$

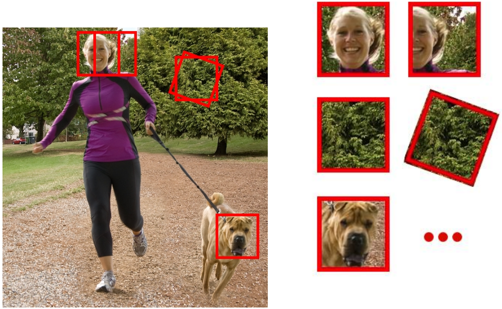
Human Pose $\mathcal{Y}$



Example y_{head}

- Body Part: $y_{head} = (u, v, \theta)$ where (u, v) center, θ rotation
  - $(u, v) \in \{1, \ldots, M\} \times \{1, \ldots, N\}, \theta \in \{0, 45^\circ, 90^\circ, \ldots\}$

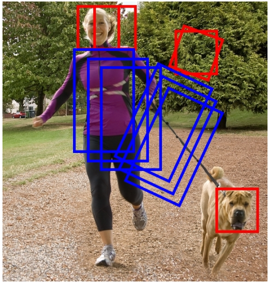
Human Pose $\mathcal{Y}$



Example y_{head}

- Body Part: $y_{head} = (u, v, \theta)$ where (u, v) center, θ rotation
  - $(u, v) \in \{1, \ldots, M\} \times \{1, \ldots, N\}, \theta \in \{0, 45^\circ, 90^\circ, \ldots\}$

Human Pose $\mathcal{Y}$



Example y_{head}

- Body Part: $y_{head} = (u, v, \theta)$ where (u, v) center, θ rotation
  - $(u, v) \in \{1, \ldots, M\} \times \{1, \ldots, N\}, \theta \in \{0, 45^\circ, 90^\circ, \ldots\}$
- Entire Body: $y = (y_{\text{head}}, y_{\text{torso}}, y_{\text{left-lower-arm}}, \ldots\} \in \mathcal{Y}$

Human Pose $\mathcal{Y}$

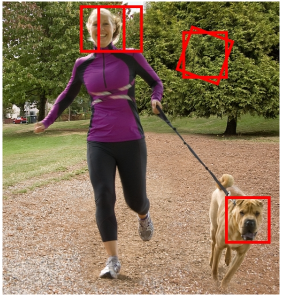
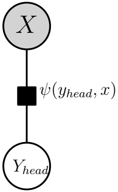


Image $x \in \mathcal{X}$

Example y_{head}



Head detector

- Idea: Have a head classifier (CNN, SVM, ...) $\psi(y_{head}, x) \in \mathbb{R}_+$

Human Pose $\mathcal{Y}$

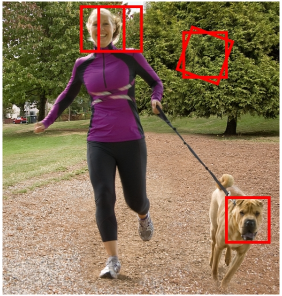
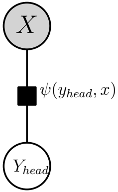


Image $x \in \mathcal{X}$

Example y_{head}



Head detector

- Idea: Have a head classifier (CNN, SVM, ...) $\psi(y_{head}, x) \in \mathbb{R}_+$
- Evaluate everywhere and record score

Human Pose $\mathcal{Y}$

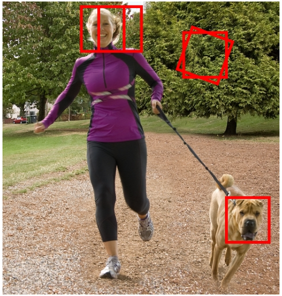
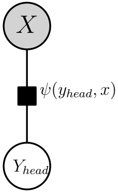


Image $x \in \mathcal{X}$

Example y_{head}



Head detector

- Idea: Have a head classifier (CNN, SVM, ...) $\quad \psi(y_{head}, x) \in \mathbb{R}_+$
- Evaluate everywhere and record score
- Repeat for all body parts

Human Pose Estimation

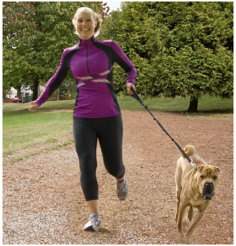
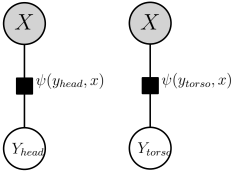


Image $x \in \mathcal{X}$



- Compute

$$y^* = (y^*_{head}, y^*_{torso}, \cdots) = \underset{y_{head}, y_{torso}, \cdots}{\operatorname{argmax}} \psi(y_{head}, x)\psi(y_{torso}, x)\cdots$$

Human Pose Estimation

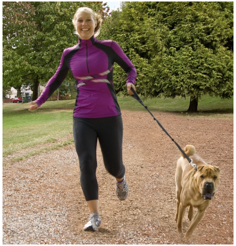
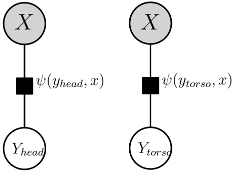


Image $x \in \mathcal{X}$



- Compute

$$
\begin{aligned}
y^* &= (y^*_{head}, y^*_{torso}, \cdots) = \underset{y_{head}, y_{torso}, \cdots}{\operatorname{argmax}} \; \psi(y_{head}, x)\psi(y_{torso}, x)\cdots \\
&= (\underset{y_{head}}{\operatorname{argmax}}\, \psi(y_{head}, x), \; \underset{y_{torso}}{\operatorname{argmax}}\, \psi(y_{torso}, x), \; \cdots)
\end{aligned}
$$

Human Pose Estimation

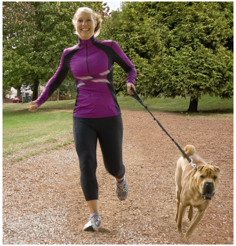
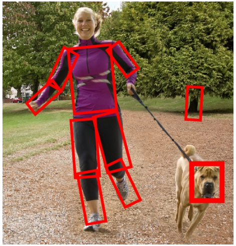


Image $x \in \mathcal{X}$



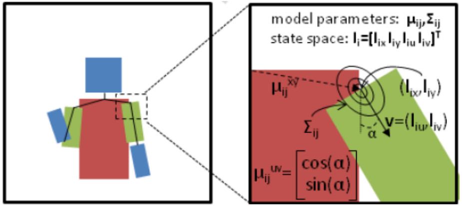
Prediction $y^* \in \mathcal{Y}$

- Compute

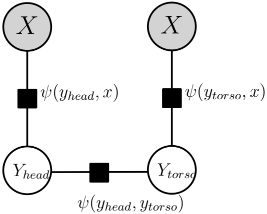
$$
\begin{aligned}
y^* &= (y^*_{head}, y^*_{torso}, \cdots) = \operatorname*{argmax}_{y_{head}, y_{torso}, \cdots} \psi(y_{head}, x)\psi(y_{torso}, x)\cdots \\
&= (\operatorname*{argmax}_{y_{head}} \psi(y_{head}, x), \ \operatorname*{argmax}_{y_{torso}} \psi(y_{torso}, x), \ \cdots)
\end{aligned}
$$

- Problem solved!?

Idea: Connect up the body



$\psi(y_{torso}, y_{arm})$



Head-Torso Model

- Ensure *head* is on top of *torso*

$$\psi(y_{head}, y_{torso}) \in \mathbb{R}_+$$

- Compute

$$y^* = \underset{y_{head}, y_{torso}, \cdots}{\operatorname{argmax}} \; \psi(y_{head}, x)\psi(y_{torso}, x)\psi(y_{head}, y_{torso}) \cdots$$

This does not decompose anymore. Easy problem has become difficult!

left image by Ben Sapp

Example: Part-of-Speech (POS) Tagging

- given an English sentence, what part-of-speech is each word?
- useful for automatic natural language processing
  - text-to-speech,
  - automatic translation,
  - question answering, etc.

They	refuse	to	permit	us	to	obtain	the	refuse	permit
`pronoun`	`verb`	`inf-to`	`verb`	`pronoun`	`inf-to`	`verb`	`article`	`noun`	`noun`

- prediction task: $f : \mathcal{X} \to \mathcal{Y}$
- $\mathcal{X}$: sequences of English words, $(x_1, \ldots, x_m)$
- $\mathcal{Y}$: sequences of tags, $(y_1, \ldots, y_m)$ with $y_i \in \{\texttt{noun}, \texttt{verb}, \texttt{participle}, \texttt{article}, \texttt{pronoun}, \texttt{preposition}, \texttt{adverb}, \texttt{conjunction}, \texttt{other}\}$

Example: Part-of-Speech (POS) Tagging

They	refuse	to	permit	us	to	obtain	the	refuse	permit
`pronoun`	`verb`	`inf-to`	`verb`	`pronoun`	`inf-to`	`verb`	`article`	`noun`	`noun`

Simplest idea: classify each word

- learn a mapping $g : \{words\} \to \{tags\}$
- problem: words are ambiguous
  - **permit** can be `verb` or `noun`
  - **refuse** can be `verb` or `noun`

  per-word prediction cannot avoid mistakes

Structured model: allow for dependencies between tags

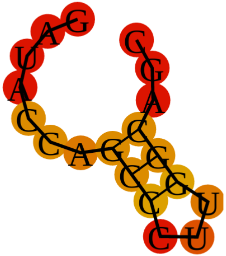
- `article` is typically followed by `noun`
- `inf-to` is typically followed by `verb`

We need to assign tags jointly for the whole sentence, not one word at a time.

Example: RNA Secondary Structure Prediction

Given an RNA sequence in text form, what's it geometric arrangement in the cell?

`GAUACCAGCCCUUGGCAGC`



Prior knowledge:

- two possible binding types: G↔C and A↔U
- big loops can form: local information is not sufficient

Structured model: combine local binding energies into globally optimal arrangement

Refresher: Probabilities

Refresher of probabilities

Most quantities in machine learning are not fully deterministic.

- true randomness of events
  - a photon reaches a camera's CCD chip, is it detected or not?
    it depends on quantum effects, which -to our knowledge- are stochastic
- incomplete knowledge
  - what will be the next email I receive?
  - who won the football match last night?
- modeling choice
  - *"For any bird, the probability that it can fly is high."*
    vs.
  - *"All birds can fly, except flightless species, or birds that are still very young, or bird which are injured in a way that prevents them..."*

In practice, there is no difference between these!

Probability theory allows us to deal with this.

Random Variables

A **random variable** is a variable that randomly takes one of its possible values:

- the number of photons reaching a CCD chip
- the text of the next email I will receive
- the position of an atom in a molecule

Some notation: we will write

- random variables with capital letters, *e.g.* X
- the set of possible values it can take with curly letters, *e.g.* $\mathcal{X}$
- any individual value it can take with lowercase letters, *e.g.* x

How likely each value $x \in \mathcal{X}$ is specified by a *probability distribution*.
There are, slightly different, possibilities:

- $\mathcal{X}$ is discrete (typically finite),
- $\mathcal{X}$ is continuous.

Discrete Random Variables

For discrete $\mathcal{X}$ (*e.g.* $\mathcal{X} = \{0, 1\}$:

- $p(X = x)$ is the probability that X takes the value $x \in \mathcal{X}$.
  If it's clear which variable we mean, we'll just write $p(x)$.
- for example, rolling a die, $p(X = 3) = p(3) = 1/6$
- we write $x \sim p(x)$ to indicate that the distribution of X is $p(x)$

For things to make sense, we need

$$0 \leq p(x) \leq 1 \quad \text{for all } x \in \mathcal{X} \qquad \text{(positivity)}$$

$$\sum_{x \in \mathcal{X}} p(x) = 1 \qquad \text{(normalization)}$$

Example: English words

- X_{word}: pick a word randomly from an English text. Is it *"word"*?
- $\mathcal{X}_{word} = \{\texttt{true}, \texttt{false}\}$

$$p(X_{the} = \texttt{true}) = 0.05 \qquad p(X_{the} = \texttt{false}) = 0.95$$

$$p(X_{horse} = \texttt{true}) = 0.004 \qquad p(X_{horse} = \texttt{false}) = 0.996$$

Continuous Random Variables

For continuous $\mathcal{X}$ (*e.g.* $\mathcal{X} = \mathbb{R}$):

- probability that X takes a value in the set M is

$$\Pr(X \in A) = \int_M p(x)\mathrm{d}x$$

- we call $p(x)$ the probability density over x

For things to make sense, we need:

$$p(x) \geq 0 \quad \text{for all } x \in \mathcal{X} \qquad \text{(positivity)}$$

$$\int_{\mathcal{X}} p(x) = 1 \qquad \text{(normalization)}$$

Note: for convenience of notation, we use the notation of discrete random variable everywhere.

Joint probabilities

Probabilities can be assigned to more than one random variable at a time:

- $p(X = x, Y = y)$ is the probability that $X = x$ and $Y = y$ (at the same time)

joint probability

Example: English words

Pick three consecutive English words: X_{word}, Y_{word}, Z_{word}

- $p(X_{the} = \texttt{true}, Y_{horse} = \texttt{true}) = 0.00080$
- $p(X_{horse} = \texttt{true}, Y_{the} = \texttt{true}) = 0.00001$
- $p(X_{probabilitistic} = \texttt{true}, Y_{graphical} = \texttt{true}, Y_{model} = \texttt{true}) = 0.000000045$

Marginalization

We can recover the probabilities of individual variables from the joint probability by summing over all variables we are not interested in.

- $p(X = x) = \sum_{y \in \mathcal{Y}} p(X = x, Y = y)$
- $p(X_2 = z) = \sum_{x_1 \in \mathcal{X}_1} \sum_{x_2 \in \mathcal{X}_2} \sum_{x_4 \in \mathcal{X}_4} p(X_1 = x_1, X_2 = z, X_3 = x_3, X_4 = x_4)$

marginalization

Marginalization

We can recover the probabilities of individual variables from the joint probability by summing over all variables we are not interested in.

- $p(X = x) = \sum_{y \in \mathcal{Y}} p(X = x, Y = y)$
- $p(X_2 = z) = \sum_{x_1 \in \mathcal{X}_1} \sum_{x_2 \in \mathcal{X}_2} \sum_{x_4 \in \mathcal{X}_4} p(X_1 = x_1, X_2 = z, X_3 = x_3, X_4 = x_4)$

marginalization

Example: English text

- $p(X_{the} = \texttt{true}, Y_{horse} = \texttt{true}) = 0.0008$
- $p(X_{the} = \texttt{true}, Y_{horse} = \texttt{false}) = 0.0492$
- $p(X_{the} = \texttt{false}, Y_{horse} = \texttt{true}) = 0.0032$
- $p(X_{the} = \texttt{false}, Y_{horse} = \texttt{false}) = 0.9468$

- $p(X_{the} = \texttt{true}) = 0.0008 + 0.0492 = 0.05$, etc.

Conditional probabilities

One random variable can contain information about another one:

- $p(X = x \mid Y = y)$: **conditional probability**
  what is the probability of $X = x$, if we already know that $Y = y$?
- $p(X = x)$: **marginal probability**
  what is the probability of $X = x$, without any additional information?
- conditional probabilities can be computed from joint and marginal:

$$p(X = x | Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} \qquad \text{(not defined if } p(Y = y) = 0\text{)}$$

Conditional probabilities

One random variable can contain information about another one:

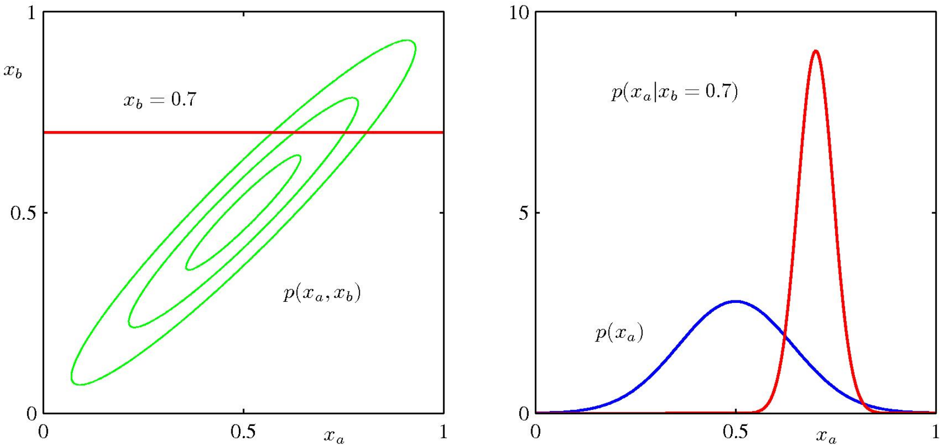
- $p(X = x \mid Y = y)$: **conditional probability**
  what is the probability of $X = x$, if we already know that $Y = y$?
- $p(X = x)$: **marginal probability**
  what is the probability of $X = x$, without any additional information?
- conditional probabilities can be computed from joint and marginal:

$$p(X = x|Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} \qquad \text{(not defined if } p(Y = y) = 0)$$

Example: English text

- $p(X_{the} = \texttt{true}) = 0.05$
- $p(X_{the} = \texttt{true}|Y_{horse} = \texttt{true}) = \frac{0.0008}{0.004} = 0.20$
- $p(X_{the} = \texttt{true}|Y_{the} = \texttt{true}) = \frac{0.0003}{0.05} = 0.006$

Illustration



joint (level sets), marginal, conditional probability

Bayes rule (Bayes theorem)

Bayes rule

Most famous formula in probability: $p(A|B) = \dfrac{p(B|A)p(A)}{p(B)}$

Bayes rule (Bayes theorem)

Bayes rule

Most famous formula in probability: $p(A|B) = \frac{p(B|A)p(A)}{p(B)}$

Formally, nothing spectacular: direct consequence of definition of conditional probability.

$$p(A|B) = \frac{p(A, B)}{p(B)} = \frac{p(B|A)p(A)}{p(B)}$$

Bayes rule (Bayes theorem)

Bayes rule

Most famous formula in probability: $p(A|B) = \frac{p(B|A)p(A)}{p(B)}$

Formally, nothing spectacular: direct consequence of definition of conditional probability.

$$p(A|B) = \frac{p(A, B)}{p(B)} = \frac{p(B|A)p(A)}{p(B)}$$

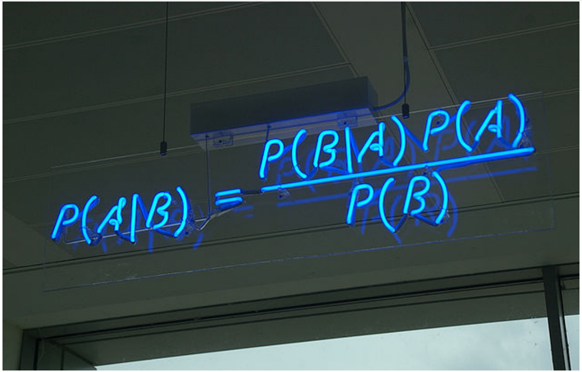
Nevertheless very useful at least for two situations:

- when A and B have different role, so $p(A|B)$ is intuitive but $p(B|A)$ is not
  - $A = \text{age}$, $B = \{\text{smoker}, \text{nonsmoker}\}$
    $p(A|B)$ is the age distribution amongst smokers and nonsmokers
    $p(B|A)$ is the probability that a person of a certain age smokes
- the information in B help us to update our knowledge about A: $p(A) \mapsto p(A|B)$

Bayes rule (Bayes theorem)

Bayes rule

Most famous formula in probability: $p(A|B) = \dfrac{p(B|A)p(A)}{p(B)}$



Dependence/Independence

Not every random variable is informative about every other.

- We say X **is independent of** Y **(write:** $X \perp\!\!\!\perp Y$**)** if
$$p(X=x, Y=y) = p(X=x)p(Y=y) \qquad \text{for all } x \in \mathcal{X} \text{ and } y \in \mathcal{Y}$$
- equivalent (if defined):
$$p(X=x|Y=y) = p(X=x), \qquad p(Y=y|X=x) = p(Y=y)$$

Dependence/Independence

Not every random variable is informative about every other.

- We say X **is independent of** Y **(write: $X \perp\!\!\!\perp Y$)** if

$$p(X = x, Y = y) = p(X = x)p(Y = y) \qquad \text{for all } x \in \mathcal{X} \text{ and } y \in \mathcal{Y}$$

- equivalent (if defined):

$$p(X = x|Y = y) = p(X = x), \qquad p(Y = y|X = x) = p(Y = y)$$

Other random variables can influence the independence:

- X **and** Y **are conditionally independent given** Z **(write $X \perp\!\!\!\perp Y|Z$)** if

$$p(X = x, Y = y|Z = z) = p(X = x|Z = z)p(Y = y|Z = z)$$

- equivalent (if defined):

$$p(x|y, z) = p(x|z), \qquad p(y|x, z) = p(y|z)$$

Example: rolling dice

Let X and Y be the outcome of independently rolling two dice and let $Z = X + Y$ be their sum.

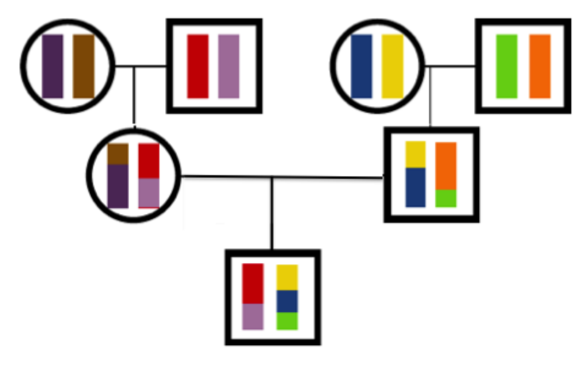
- X and Y are independent
- X and Z are not independent, Y and Z are not independent
- conditioned on Z, X and Y are not independent anymore
  (for fixed $Z = z$, X and Y can only take certain value combinations)

Example: toddlers

Let X be the height of a toddler, Y the number of words in its vocabulary and Z its age.

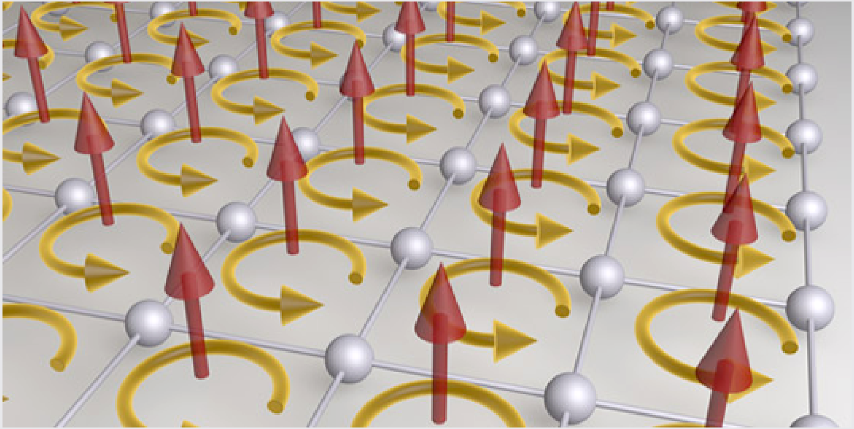
- X and Y are not independent: overall, toddlers who are taller know more words
- however, X and Y are conditionally independent given Z:
  at a fixed age, toddlers' growth and vocabulary develop independently

Example



- X = your genome
- Y_1, Y_2 = your parents' genomes
- Z_1, Z_2, Z_3, Z_4 = your grantparents' genomes

Discrete Random Fields



Magnetic spin in each atoms of a crystal: $X_{i,j}$ for $i, j \in \mathbb{Z}$

Image: https://www.en.uni-muenchen.de/news/newsarchiv/2013/f-m-81-13.html

Continuous Random Fields

Distribution of matter in the universe: X_p for $p \in \mathbb{R}^3$

Expected value

We apply a function to (the values of) one or more random variables:

- $f(x) = x^2$ or $f(x_1, x_2, \ldots, x_k) = \frac{x_1 + x_2 + \cdots + x_k}{k}$

The **expected value** or **expectation** of a function f with respect to a probability distribution is the weighted average of the possible values:

$$\mathbb{E}_{x \sim p(x)}[f(x)] := \sum_{x \in \mathcal{X}} p(x) f(x)$$

In short, we just write $\mathbb{E}_x[f(x)]$ or $\mathbb{E}[f(x)]$ or $\mathbb{E}[f]$ or $\mathbb{E}f$.

Example: rolling dice

Let X be the outcome of rolling a die and let $f(x) = x$

$$\mathbb{E}_{x \sim p(x)}[f(x)] = \mathbb{E}_{x \sim p(x)}[x] = \frac{1}{6}1 + \frac{1}{6}2 + \frac{1}{6}3 + \frac{1}{6}4 + \frac{1}{6}5 + \frac{1}{6}6 = 3.5$$

Expected value

Example: rolling dice

X_1, X_2: the outcome of rolling two dice independently, $f(x_1, x_2) = x_1 + x_2$

$$\mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[f(x_1, x_2)] =$$

Expected value

Example: rolling dice

X_1, X_2: the outcome of rolling two dice independently, $f(x_1, x_2) = x_1 + x_2$

$$\mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[f(x_1, x_2)] =$$

Straight-forward computation: 36 options for (x_1, x_2), each has probability $\frac{1}{36}$

Expected value

Example: rolling dice

X_1, X_2: the outcome of rolling two dice independently, $f(x_1, x_2) = x_1 + x_2$

$$\mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[f(x_1, x_2)] =$$

Straight-forward computation: 36 options for (x_1, x_2), each has probability $\frac{1}{36}$

$$\begin{aligned}
\mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[f(x_1, x_2)] &= \sum_{x_1,x_2} p(x_1, x_2)(x_1 + x_2) \\
&= \frac{1}{36}(1+1) + \frac{1}{36}(1+2) + \frac{1}{36}(1+3) + \ldots \\
&\quad + \frac{1}{36}(2+1) + \frac{1}{36}(2+2) + \cdots + \frac{1}{36}(6+6) \\
&= \frac{252}{36} = 7
\end{aligned}$$

Expected value

Example: rolling dice

X_1, X_2: the outcome of rolling two dice independently, $f(x_1, x_2) = x_1 + x_2$

$$\mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[f(x_1, x_2)] = \mathbf{7}$$

Straight-forward computation: 36 options for (x_1, x_2), each has probability $\frac{1}{36}$

$$\begin{aligned}
\mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[f(x_1, x_2)] &= \sum_{x_1,x_2} p(x_1, x_2)(x_1 + x_2) \\
&= \frac{1}{36}(1+1) + \frac{1}{36}(1+2) + \frac{1}{36}(1+3) + \ldots \\
&\quad + \frac{1}{36}(2+1) + \frac{1}{36}(2+2) + \cdots + \frac{1}{36}(6+6) \\
&= \frac{252}{36} = 7
\end{aligned}$$

Expected value

Example: rolling dice

X_1, X_2: the outcome of rolling two dice independently, $f(x_1, x_2) = x_1 + x_2$

$$\mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[f(x_1, x_2)] =$$

Sometimes a good strategy: count how often each value occurs and sum over values

$s = (x_1 + x_2)$	2	3	4	5	6	7	8	9	10	11	12
count n_s	1	2	3	4	5	6	5	4	3	2	1

Expected value

Example: rolling dice

X_1, X_2: the outcome of rolling two dice independently, $f(x_1, x_2) = x_1 + x_2$

$$\mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[f(x_1, x_2)] = \mathbf{7}$$

Sometimes a good strategy: count how often each value occurs and sum over values

$s = (x_1 + x_2)$	2	3	4	5	6	7	8	9	10	11	12
count n_s	1	2	3	4	5	6	5	4	3	2	1

$$\begin{aligned}\mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[f(x_1, x_2)] &= \sum_{x_1,x_2} p(x_1, x_2)(x_1 + x_2) = \sum_s \frac{n_s}{n} s \\ &= \frac{1}{36}2 + \frac{2}{36}3 + \frac{3}{36}4 + \frac{4}{36}5 + \cdots + \frac{2}{36}11 + \frac{1}{36}12 = \frac{252}{36} = 7\end{aligned}$$

Properties of expected values

Example: rolling dice

X_1, X_2: the outcome of rolling two dice independently, $f(x_1, x_2) = x_1 + x_2$

$$\mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[f(x_1, x_2)] =$$

Properties of expected values

Example: rolling dice

X_1, X_2: the outcome of rolling two dice independently, $f(x_1, x_2) = x_1 + x_2$

$$\mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[f(x_1, x_2)] =$$

The expected value has a useful property: it is *linear* in its argument.

- $\mathbb{E}_{x\sim p(x)}[f(x) + g(x)] = \mathbb{E}_{x\sim p(x)}[f(x)] + \mathbb{E}_{x\sim p(x)}[g(x)]$
- $\mathbb{E}_{x\sim p(x)}[\lambda f(x)] = \lambda \mathbb{E}_{x\sim p(x)}[f(x)]$

If a random variables does not show up in a function, we can ignore the expectation operation with respect to it

- $\mathbb{E}_{(x,y)\sim p(x,y)}[f(x)] = \mathbb{E}_{x\sim p(x)}[f(x)]$

Properties of expected values

Example: rolling dice

X_1, X_2: the outcome of rolling two dice independently, $f(x_1, x_2) = x_1 + x_2$

$$\begin{aligned}\mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[f(x_1,x_2)] &= \mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[x_1 + x_2] \\ &= \mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[x_1] + \mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[x_2] \\ &= \mathbb{E}_{x_1\sim p(x_1)}[x_1] + \mathbb{E}_{x_2\sim p(x_2)}[x_2] = 3.5 + 3.5 = \mathbf{7}\end{aligned}$$

The expected value has a useful property: it is *linear* in its argument.

- $\mathbb{E}_{x\sim p(x)}[f(x) + g(x)] = \mathbb{E}_{x\sim p(x)}[f(x)] + \mathbb{E}_{x\sim p(x)}[g(x)]$
- $\mathbb{E}_{x\sim p(x)}[\lambda f(x)] = \lambda\mathbb{E}_{x\sim p(x)}[f(x)]$

If a random variables does not show up in a function, we can ignore the expectation operation with respect to it

- $\mathbb{E}_{(x,y)\sim p(x,y)}[f(x)] = \mathbb{E}_{x\sim p(x)}[f(x)]$

Example: rolling dice

- we roll one die
- X_1: number facing up, X_2: number facing down
- $f(x_1, x_2) = x_1 + x_2$

$$\mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[f(x_1, x_2)] =$$

Example: rolling dice

- we roll one die
- X_1: number facing up, X_2: number facing down
- $f(x_1, x_2) = x_1 + x_2$

$$\mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[f(x_1, x_2)] = \mathbf{7}$$

Answer 1: explicit calculation with dependent X_1 and X_2

$$p(x_1, x_2) = \begin{cases} \frac{1}{6} & \text{for combinations } (1,6),\ (2,5),\ (3,4),\ (4,3),\ (5,2),\ (6,1) \\ 0 & \text{for all other combinations.} \end{cases}$$

$$\mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[f(x_1, x_2)] = \sum\nolimits_{(x_1,x_2)} p(x_1, x_2)(x_1 + x_2)$$

$$= 0(1+1) + 0(1+2) + \cdots + \frac{1}{6}(1+6) + 0(2+1) + \cdots = 6 \cdot \frac{7}{6} = 7$$

Example: rolling dice

- we roll one die
- X_1: number facing up, X_2: number facing down
- $f(x_1, x_2) = x_1 + x_2$

$$\mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[f(x_1, x_2)] = \mathbf{7}$$

Answer 2: use properties of expectation as earlier

$$\begin{aligned}
\mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[f(x_1, x_2)] &= \mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[x_1 + x_2] \\
&= \mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[x_1] + \mathbb{E}_{(x_1,x_2)\sim p(x_1,x_2)}[x_2] \\
&= \mathbb{E}_{x_1\sim p(x_1)}[x_1] + \mathbb{E}_{x_2\sim p(x_2)}[x_2] = 3.5 + 3.5 = 7
\end{aligned}$$

The rules of probability take care of dependence, etc.

Some expected values show up so often that they have special names.

Variance

The **variance** of a random variable X is the expected squared deviation from its mean

$$\mathrm{Var}(X) = \mathbb{E}_x[(x - \mathbb{E}_x[x])^2]$$

also

$$\mathrm{Var}(X) = \mathbb{E}_x[x^2] - (\mathbb{E}_x[x])^2 \qquad \text{(exercise)}$$

The variance

- measures how much the random variable *fluctuates* around its mean
- is invariant under addition

$$\mathrm{Var}(X + a) = \mathrm{Var}(X) \qquad \text{for } a \in \mathbb{R}.$$

- scales with the square of multiplicative factors

$$\mathrm{Var}(\lambda X) = \lambda^2 \mathrm{Var}(X) \qquad \text{for } \lambda \in \mathbb{R}.$$

More intuitive:

Standard deviation

The **standard deviation** of a random variable is the square root of the its variance.

$$\mathrm{Std}(X) = \sqrt{\mathrm{Var}(X)}$$

The standard deviation

- is invariant under addition

$$\mathrm{Std}(X + a) = \mathrm{Std}(X) \qquad \text{for } a \in \mathbb{R}.$$

- scales with the absolute value of multiplicative factors

$$\mathrm{Std}(\lambda X) = |\lambda| \, \mathrm{Std}(X) \qquad \text{for } \lambda \in \mathbb{R}.$$

For two random variables at a time, we can test if their fluctuations around the mean are consistent or not

Covariance

The **covariance** of two random variables X and Y is the expected value of the product of their deviations from their means

$$\mathrm{Cov}(X, Y) = \mathbb{E}_{(x,y)\sim p(x,y)}[(x - \mathbb{E}_x[x])(y - \mathbb{E}_y[y])]$$

The covariance

- of a random variable with itself it its variance, $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
- is invariant under addition: $\mathrm{Cov}(X + a, Y) = \mathrm{Cov}(X, Y) = \mathrm{Cov}(X, Y + a) \qquad$ for $a \in \mathbb{R}$.
- scales linearly under multiplications:
  $\mathrm{Cov}(\lambda X, Y) = \lambda\,\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, \lambda Y) \qquad$ for $\lambda \in \mathbb{R}$.
- is 0, if X and Y are independent, but can be 0 even if for dependent X and Y (exercise)

If we do not care about the scales of X and Y, we can normalize by their standard deviations:

Correlation

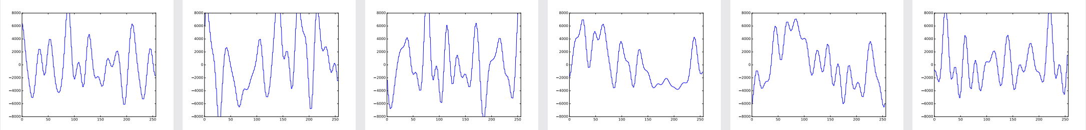
The **correlation coefficient** of two random variables X and Y is their covariance divided by their standard deviations

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Std}(X)\,\mathrm{Std}(Y)} \qquad = \frac{\mathbb{E}_{(x,y)\sim p(x,y)}[(x-\mathbb{E}_x[x])(y-\mathbb{E}_y[y])]}{\sqrt{\mathbb{E}_x(x-\mathbb{E}_x[x])^2}\sqrt{\mathbb{E}_y(y-\mathbb{E}_y[y])^2}}$$

The correlation

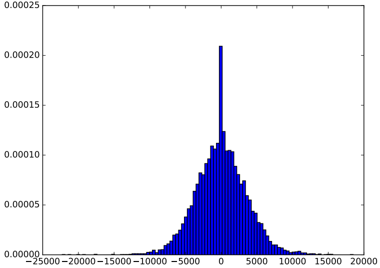
- always has values in the interval $[-1,1]$
- is invariant under addition: $\mathrm{Cov}(X+a,Y) = \mathrm{Cov}(X,Y) = \mathrm{Cov}(X,Y+a) \qquad \text{for } a \in \mathbb{R}$.
- is invariant under multiplication with positive constants
  $\mathrm{Corr}(\lambda X,Y) = \mathrm{Corr}(X,Y) = \mathrm{Corr}(X,\lambda Y) \qquad \text{for } \lambda > 0$.
- inverts its sign under multiplication with negative constants
  $\mathrm{Corr}(-\lambda X,Y) = -\mathrm{Corr}(X,Y) = \mathrm{Corr}(X,-\lambda Y) \qquad \text{for } \lambda > 0$.
- is 0, if X and Y are independent, but can be 0 even if for dependent X and Y (exercise)

Example: Audio Signals



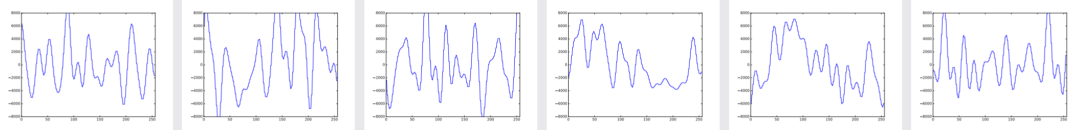
X_t: accustic pressure at any time t, uniformly over all songs on your MP3 player

For example, $t = 60s$, what's the probability distribution?



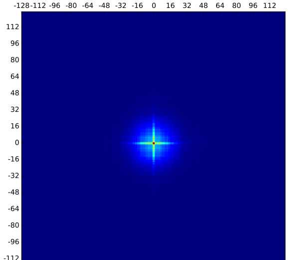
- $X_t \in \{-32768, -32767, \ldots, 32767\}$
- $\mathbb{E}_x[X_t] \approx 21$
- $\Pr(X_t = 21) \approx 0.00045$
- $\mathrm{Var}(X_t) \approx 12779141$
- $\mathrm{Std}(X_t) \approx 3575$

Example: Audio Signals



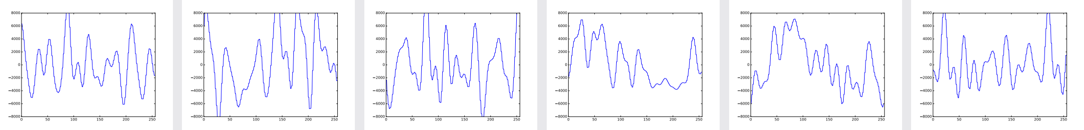
X_t, Y_t: accustic pressures at any time t for two different randomly chosen songs

Joint probability distribution for $t = 60s$:



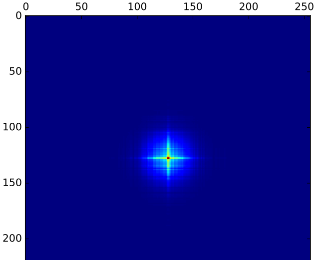
- $X_t, Y_t \in \{-32768, -32767, \ldots, 32767\}$
- $\mathbb{E}_x[X_t] = \mathbb{E}_y[Y_t] \approx 21$
- $\mathrm{Cov}(X_t, Y_t) = 0$
- $\mathrm{Corr}(X_t, Y_t) = 0$

Example: Audio Signals



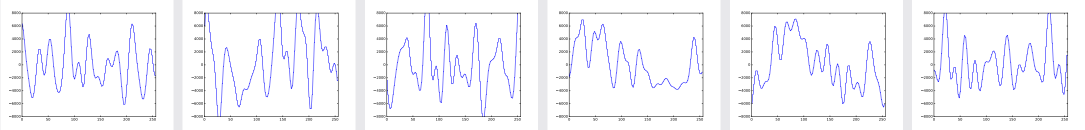
X_s, X_t: accustic pressures at times s and t for <u>one</u> randomly chosen songs

Joint probability distribution for $s = 60s$, $t = 61s$:



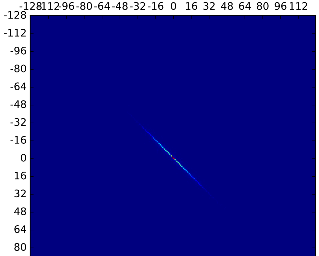
- $X_s, X_t \in \{-32768, -32767, \ldots, 32767\}$
- $\mathbb{E}_x[X_s] \approx 21, \quad \mathbb{E}_y[X_t] \approx -39$
- $\mathrm{Cov}(X_s, X_t) \approx 0$
- $\mathrm{Corr}(X_s, X_t) \approx 0$

Example: Audio Signals



X_s, X_t: accustic pressures at times s and t for <u>one</u> randomly chosen songs

Joint probability distribution for $s = 60s$, $t = (60 + \frac{1}{65536})s$ (one sampling step):



- $X_s, X_t \in \{-32768, -32767, \ldots, 32767\}$
- $\mathbb{E}_x[X_s] \approx 21, \quad \mathbb{E}_y[X_t] \approx 22$
- $\mathrm{Cov}(X_s, X_t) \approx 12613175$
- $\mathrm{Corr}(X_s, X_t) \approx 0.988$

In practice, we might know the distribution of a model, but for the real world we have to work with samples.

Random sample (in statistics)

A set $\{x_1, \ldots, x_n\}$ is a random sample for a random variable X, if each x_i is a realization of X.

Equivalently, but easier to treat formally: we create n random variables, $X_1, \ldots, X_n$, where each X_i is distributed identically to X, and we obtain one realization: $x_1, \ldots, x_n$.

Note: in machine learning, we also call each individual realization a sample, and the set of multiple samples a sample set.

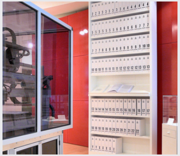
I.i.d. sample

We call the sample set independent and identically distributed (i.i.d.), if the X_i are independent of each other, *i.e.* $p(X_1, \ldots, X_n) = \prod_i p(X_i)$.

In practice, we use samples to estimate properties of the underlying probability distribution. This is easiest for i.i.d. sample sets, but we'll see other examples as well.

Example

- X = human genome
- number of samples collected:



1

deCODE genetics

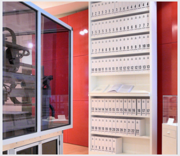
2,636

Genomics england

100,000

Example

- X = human genome
- number of samples collected:



1

2,636

100,000

1 million

>1 million

2 million

Example: Matter Distribution in the Universe

- random field: X_p for $p \in \mathbb{R}^3$
- our universe is **one realization** $\rightarrow$ how to estimate anything?

Example: Matter Distribution in the Universe

- random field: X_p for $p \in \mathbb{R}^3$
- our universe is **one realization** $\rightarrow$ how to estimate anything?
- assume homogeneity (=translation invariance):
  for any $p_1, \ldots, p_k \in \mathbb{R}^3$ and for any $t \in \mathbb{R}^3$:

$$p(X_{p_1}, X_{p_2}, \ldots, X_{p_k}) = p(X_{p_1+t}, X_{p_2+t}, \ldots, X_{p_k+t})$$

- estimate quantities (*e.g.* average matter density or correlation functions) by averaging over multiple locations instead of multiple universes