Mathematical Basics

Yu Yao

Translational Neuromodeling Unit

Computational Psychiatry Course 2020

Zurich | 8th September 2020

Universität Zürich[UZH]

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Introduction

Interactive example: 3 cards

1. black
2. mixed
3. white

Introduction

Interactive example: 3 cards

1. black
2. mixed
3. white

Question:

probability that other side is **white**, if visible side is **black**

A. 1/3
B. 1/2
C. 2/3
D. something else

Introduction

- Random variables
- Probability distributions
- Expectation

References:

C. Bishop “Pattern Recognition and Machine Learning”

D. MacKay “Information Theory, Inference, and Learning Algorithms”

Random Variables and Probability

- **Random variable:** a variable whose possible values are outcomes (events) of a random experiment, e.g.:
  - rolling a dice — 1, 2, 3, 4, 5, 6
  - tossing a coin — head, tail
  - measuring height — $[0, \infty)$
  - measuring voltage — $(-\infty, \infty)$

Random Variables and Probability

- **Random variable:** a variable whose possible values are outcomes (events) of a random experiment, e.g.:
  - rolling a dice
  - tossing a coin
  - measuring height
  - measuring voltage

discrete

continuous

Random Variables and Probability

Example: 1. tossing coin

Possible outcomes/events: head $x = 1$, or tail $x = 0$

Probabilities: $0 \le q \le 1$ for head, $1 - q$ for tail

Random Variables and Probability

Example: 1. tossing coin

Possible outcomes/events: head $x = 1$, or tail $x = 0$

Probabilities: $0 \leq q \leq 1$ for head, $1 - q$ for tail

Probability distribution:

$$\text{function: outcome} \rightarrow \text{probability}$$
$$prob = p(x)$$

Random Variables and Probability

Example: 1. tossing coin (discrete binary random variable)

Possible outcomes/events: head $x = 1$, or tail $x = 0$

Probabilities: $0 \leq q \leq 1$ for head, $1 - q$ for tail

Probability distribution: Bernoulli distribution

$$p(x) = q^x \cdot (1 - q)^{1-x}$$

Random Variables and Probability

Example: 1. tossing coin (discrete binary random variable)

Possible outcomes/events: head $x = 1$, or tail $x = 0$

Probabilities: $0 \leq q \leq 1$ for head, $1 - q$ for tail

Probability distribution: Bernoulli distribution

$$p(x) = q^x \cdot (1 - q)^{1-x}$$

Note:

$p(1) = q^1 \cdot (1 - q)^{1-1} = q$ while $p(0) = q^0 \cdot (1 - q)^{1-0} = 1 - q$

Random Variables and Probability

Example: 2. voltage (continuous random variable)

Possible outcomes/events: $x \in (-\infty, \infty)$

Random Variables and Probability

Example: 2. voltage (continuous random variable)

Possible outcomes/events: $x \in (-\infty, \infty)$

Probability ~~distribution~~ density function

function: $X \rightarrow \text{probability}\, X \leq x < X + \delta$

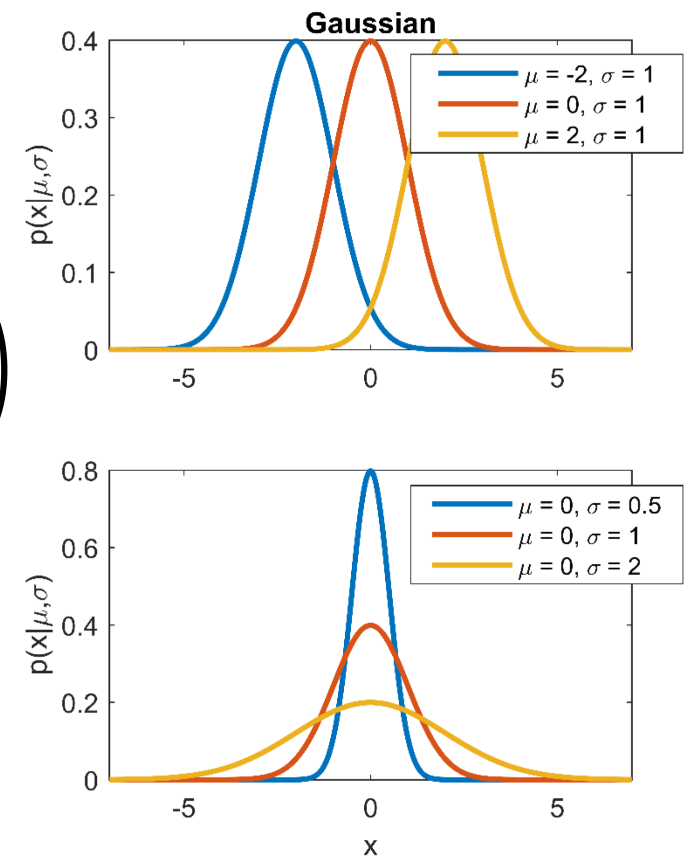
Random Variables and Probability

Example: 2. voltage (continuous random variable)

Possible outcomes/events: $x \in (-\infty, \infty)$

Probability density: Gaussian/normal density

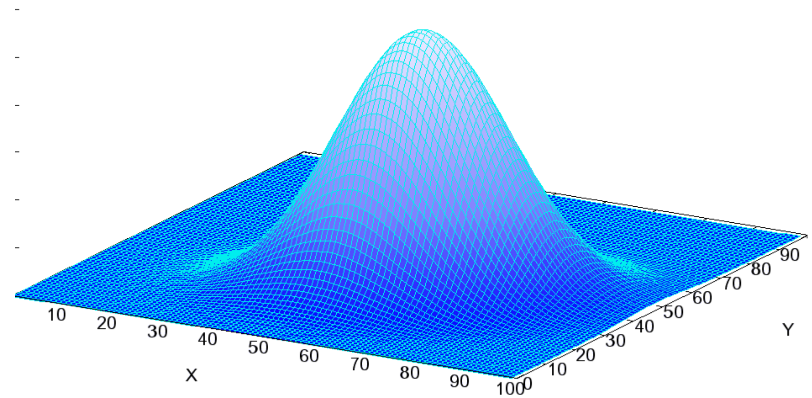
$$p(x) = N(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



Note: μ is called mean and σ^2 variance

Random Variables and Probability

Example: 2. continuous random variable



Possible outcomes/events: $x = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix}$

Probability density: Gaussian/normal density

$$p(x) = N(x) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left(-\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2}\right)$$

Note: μ is a mean vector and Σ the covariance matrix

Random Variables and Probability

Example: 3. height (continuous positive random variable)

Possible outcomes/events: $x \in [0, \infty)$

Random Variables and Probability

Example: 3. height (continuous positive random variable)

Possible outcomes/events: $x \in [0, \infty)$

Transformation: $x \in [0, \infty) \ \rightarrow y = \log(x) \in (-\infty, \infty)$

Random Variables and Probability

Example: 3. height (continuous positive random variable)

Possible outcomes/events: $x \in [0, \infty)$

Transformation: $x \in [0, \infty) \ \rightarrow y = \log(x) \in (-\infty, \infty)$

Probability density: log-normal density

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

$$p(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log(x)-\mu)^2}{2\sigma^2}\right)$$

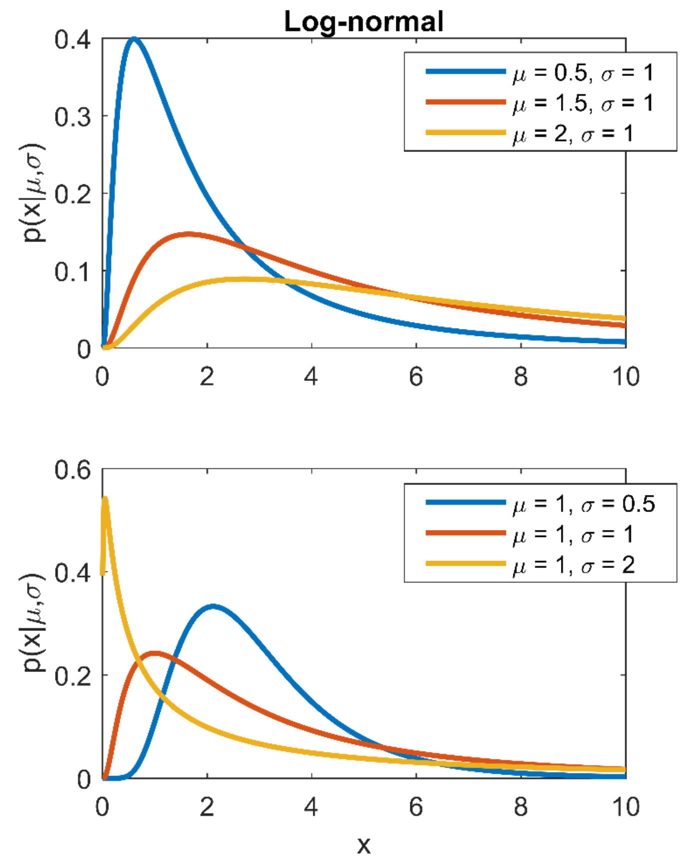
Note: when transforming continuous random variables, the gradient has to be taken into account

Random Variables and Probability

log-normal density

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

$$p(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log(x)-\mu)^2}{2\sigma^2}\right)$$

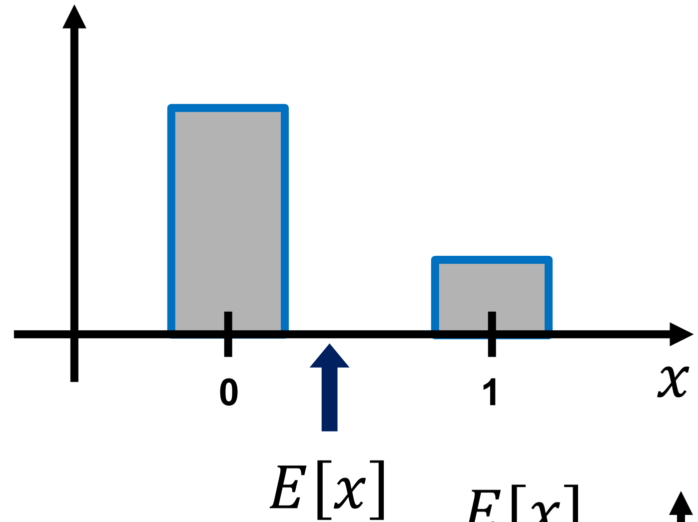


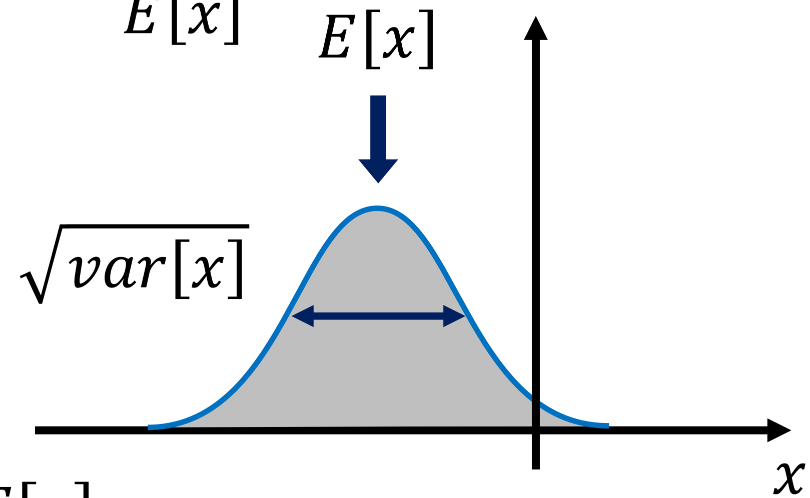
Distributions and Densities

Positivity and normalization:

$$p(x) \geq 0 \; \forall x$$

$$\sum_x p(x) = 1 \;\text{ or }\; \int p(x)dx = 1$$





expectation:

$$E[g(x)] = \sum_x g(x)p(x)$$

$$E[g(x)] = \int g(x)p(x)dx$$

mean: $\bar{x} = E[x]$

variance: $var[x] = E[(x - E[x])^2]$

Advanced Concepts

- Joint and conditional probability
- Sum and product rule

References:

C. Bishop “Pattern Recognition and Machine Learning”

D. MacKay “Information Theory, Inference, and Learning Algorithms”

Joint and Conditional Probability

In situations involving multiple random variables, it is useful to define:

Joint probability: $p(x = 1, y = 0)$

probability that random variables take a certain joint configuration

$x = \text{head}$ and $y = \text{tail}$

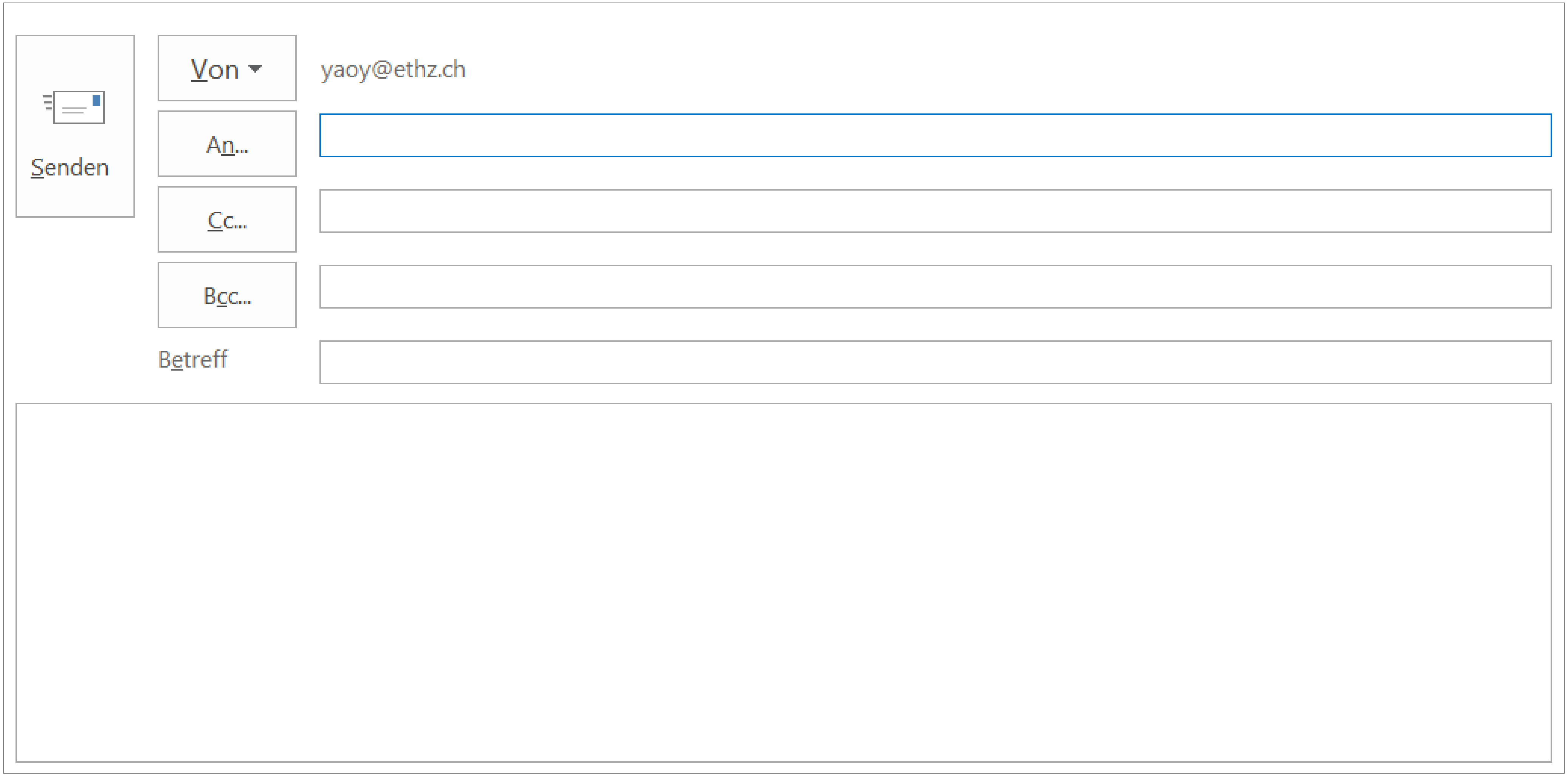
Conditional probability: $p(x = 1|y = 0)$

probability of one random variable taking a certain value, when the value of the other variables are already known

$x = \text{head}$ given $y = \text{tail}$

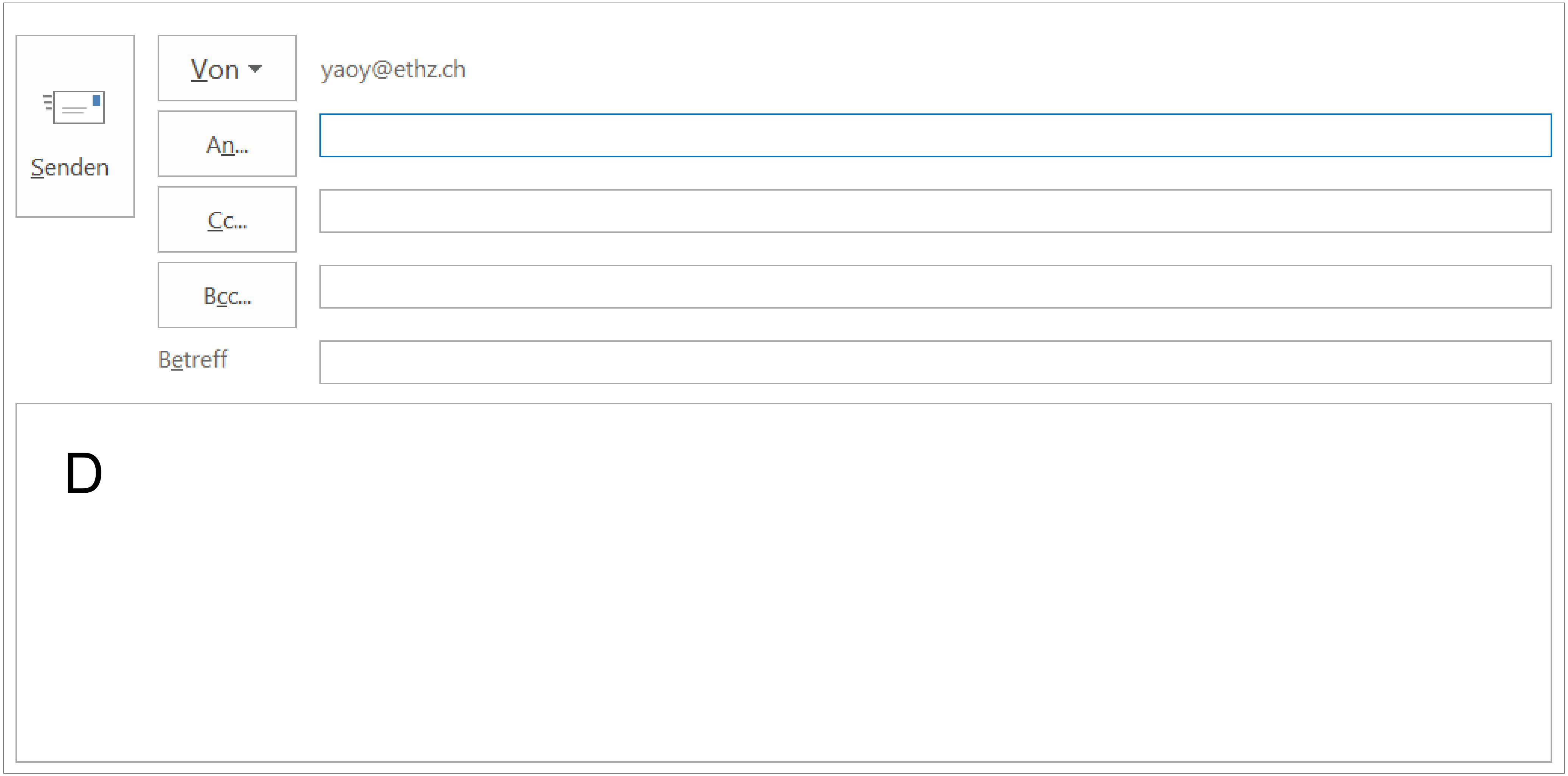
Joint and Conditional Probability

example: text prediction



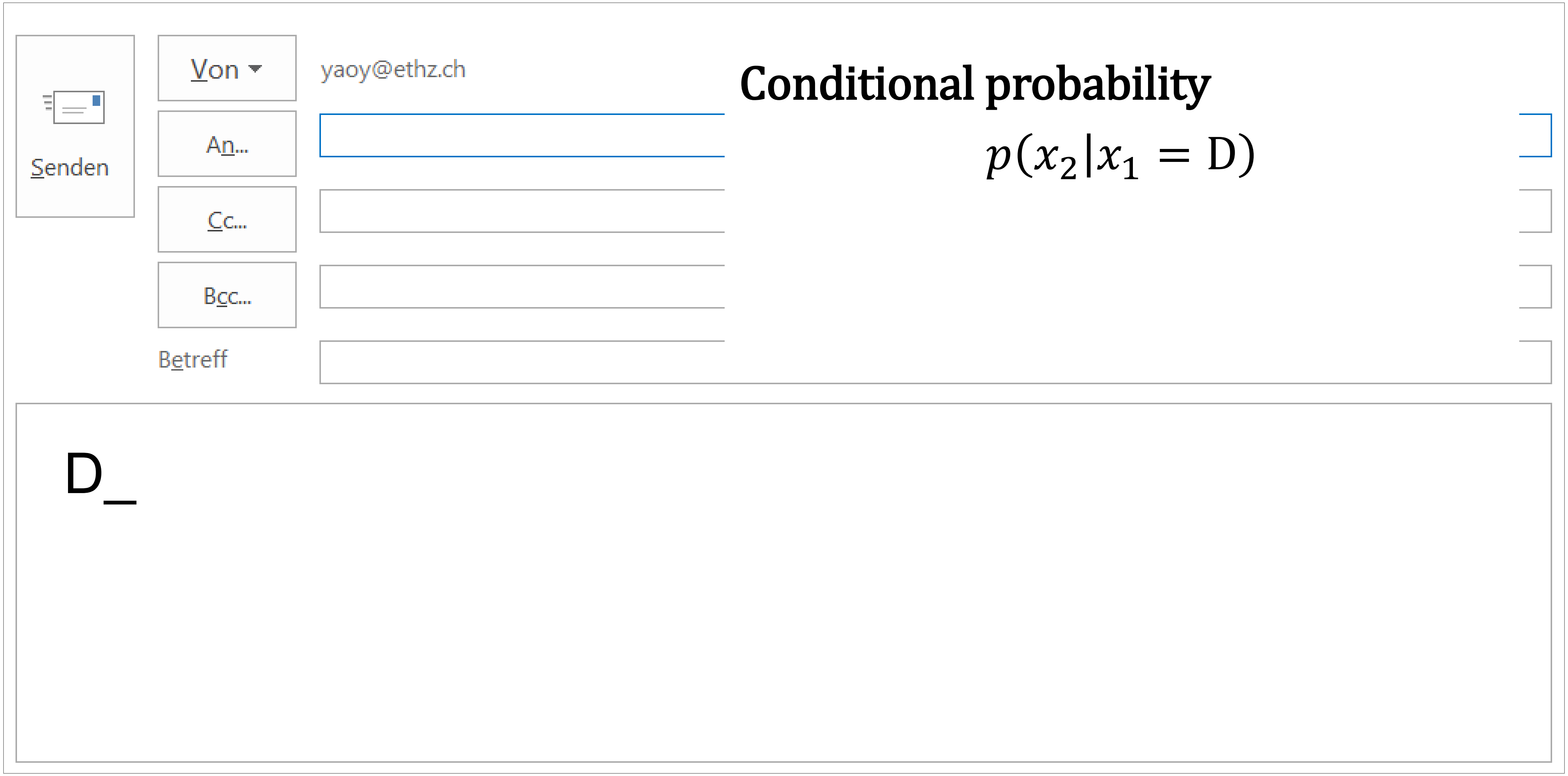
Joint and Conditional Probability

example: text prediction



Joint and Conditional Probability

example: text prediction

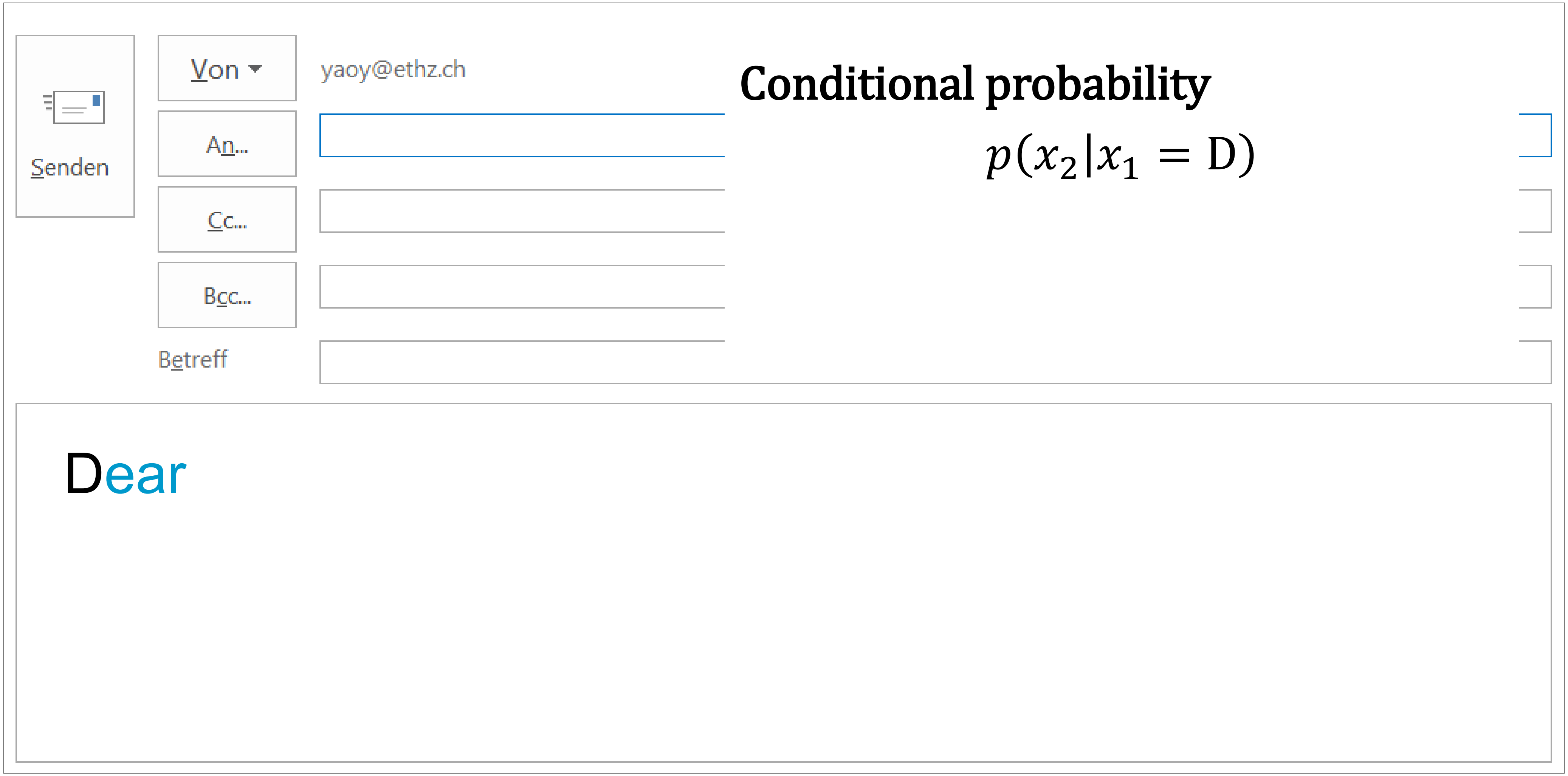


Conditional probability

$$p(x_2|x_1 = \mathrm{D})$$

Joint and Conditional Probability

example: text prediction

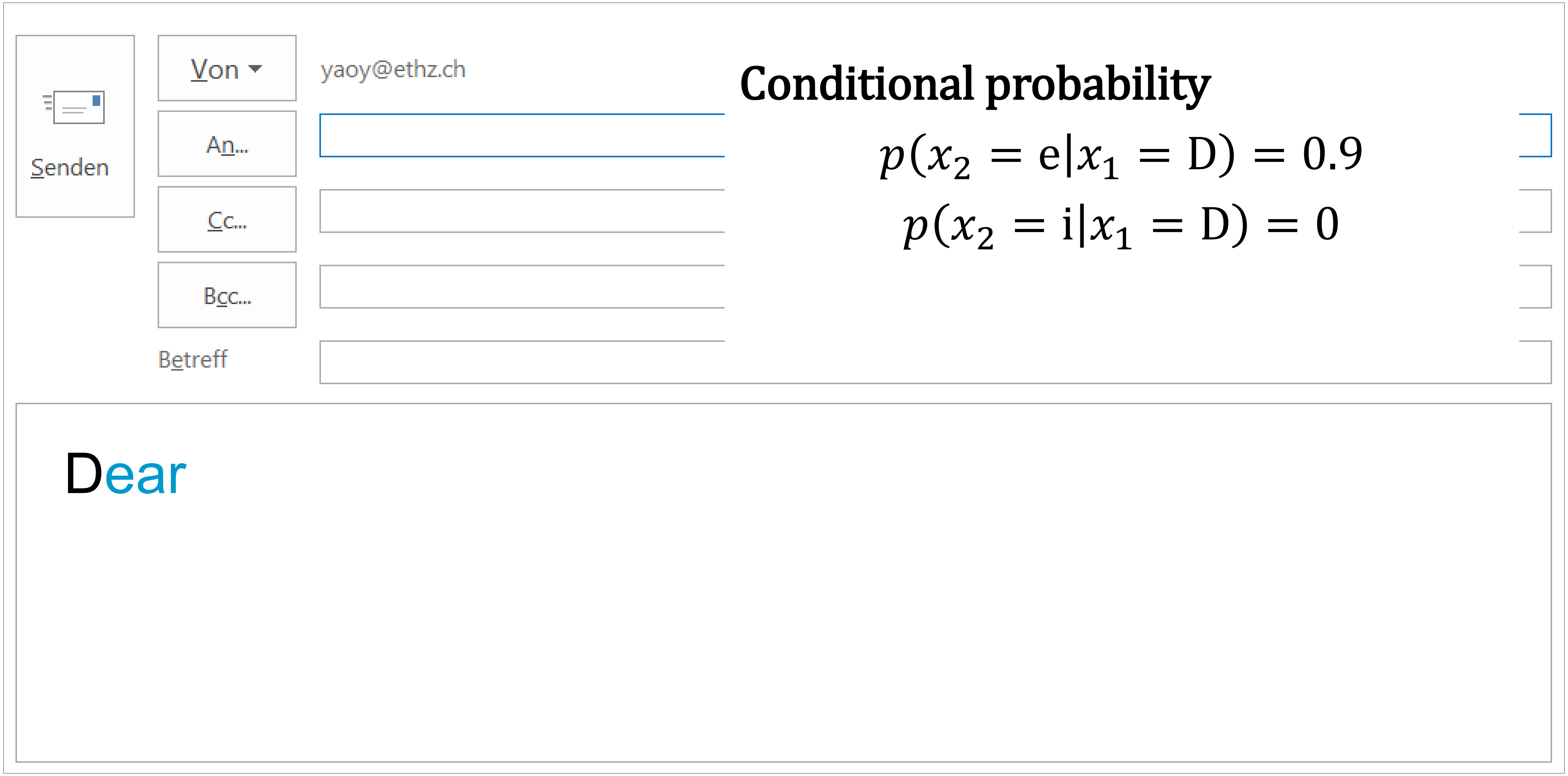


Conditional probability

$$p(x_2 | x_1 = \mathrm{D})$$

Joint and Conditional Probability

example: text prediction



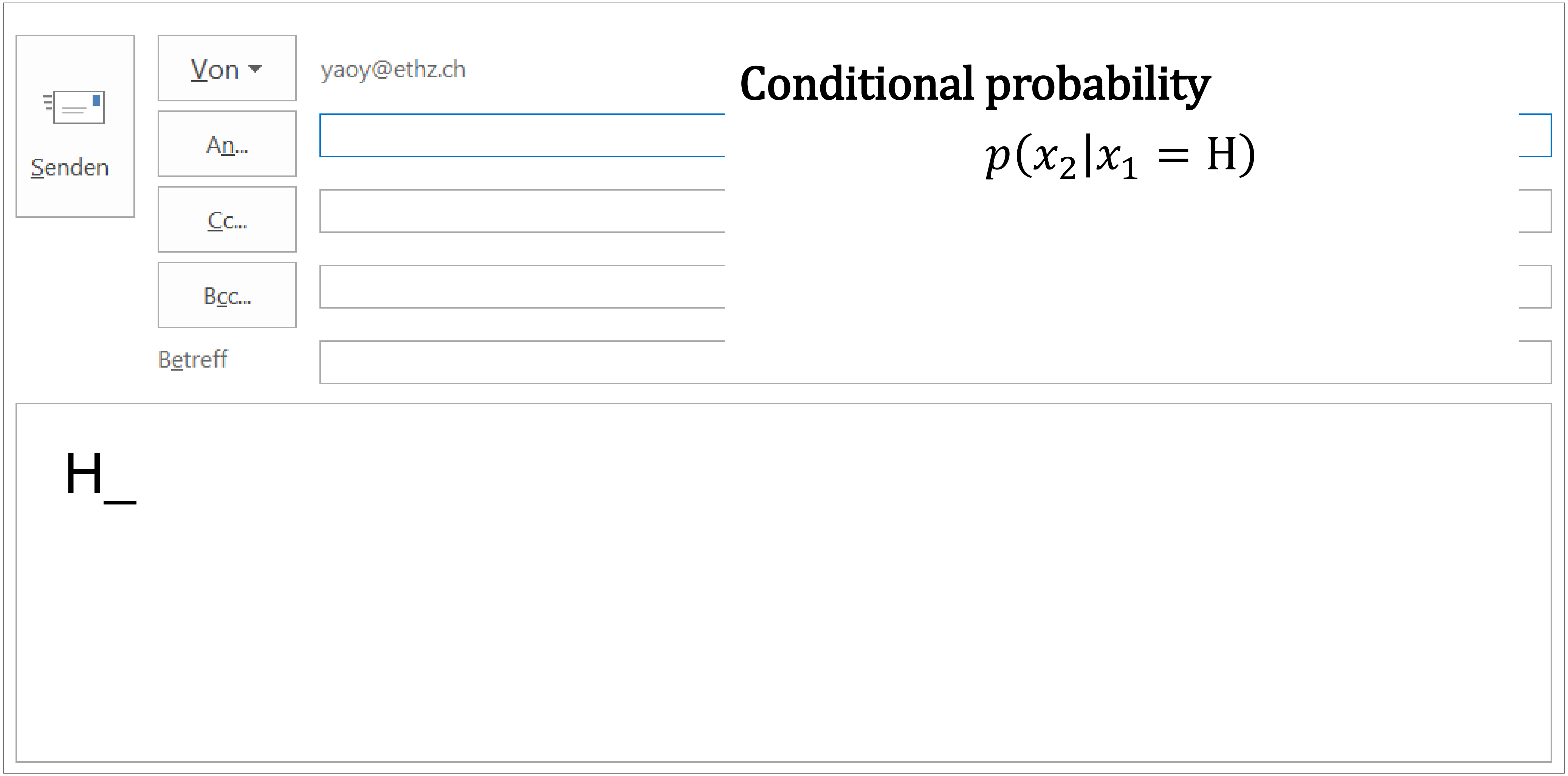
Conditional probability

$$p(x_2 = \mathrm{e} | x_1 = \mathrm{D}) = 0.9$$

$$p(x_2 = \mathrm{i} | x_1 = \mathrm{D}) = 0$$

Joint and Conditional Probability

example: text prediction

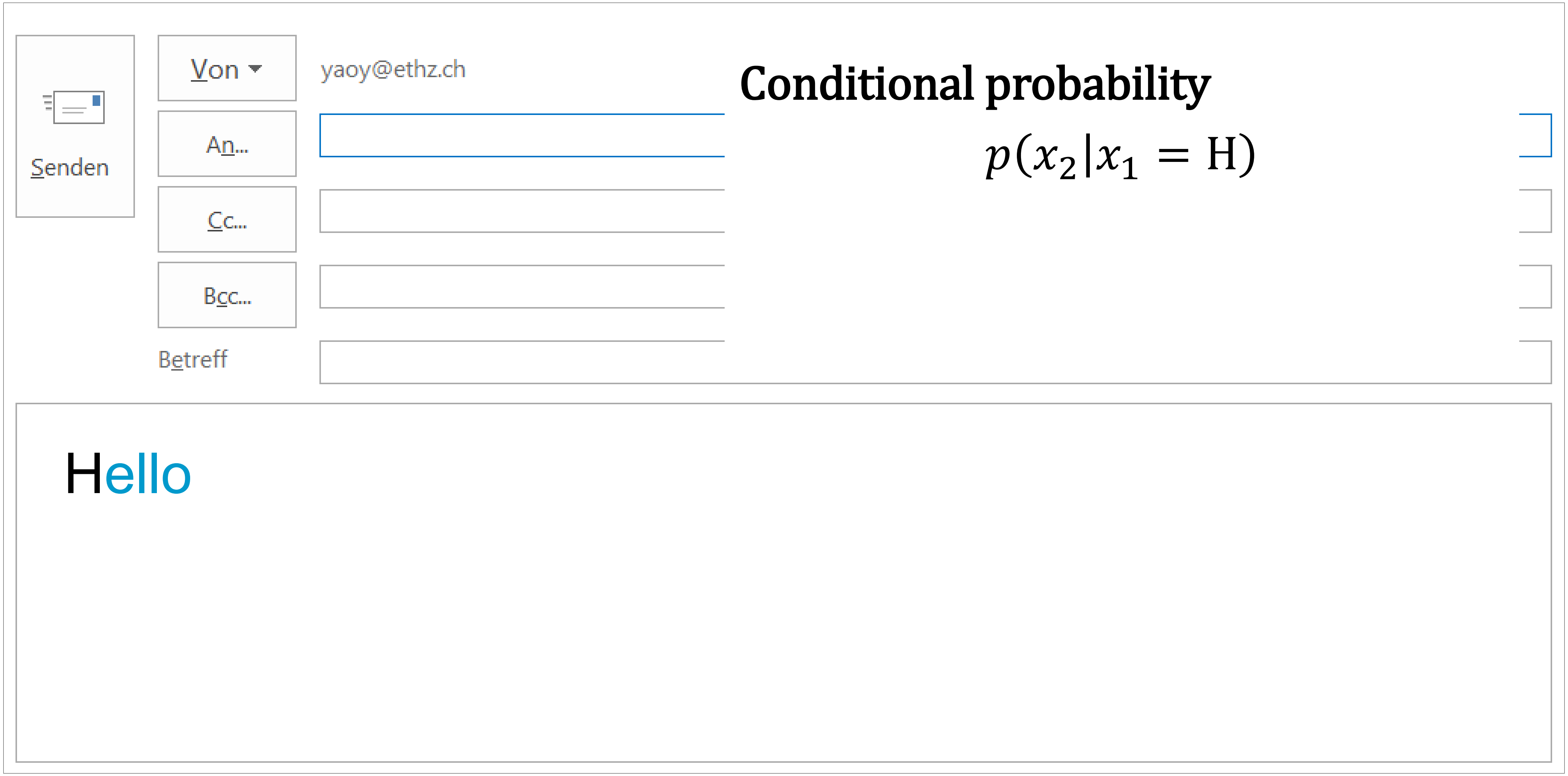


Conditional probability

$$p(x_2 | x_1 = \mathrm{H})$$

Joint and Conditional Probability

example: text prediction

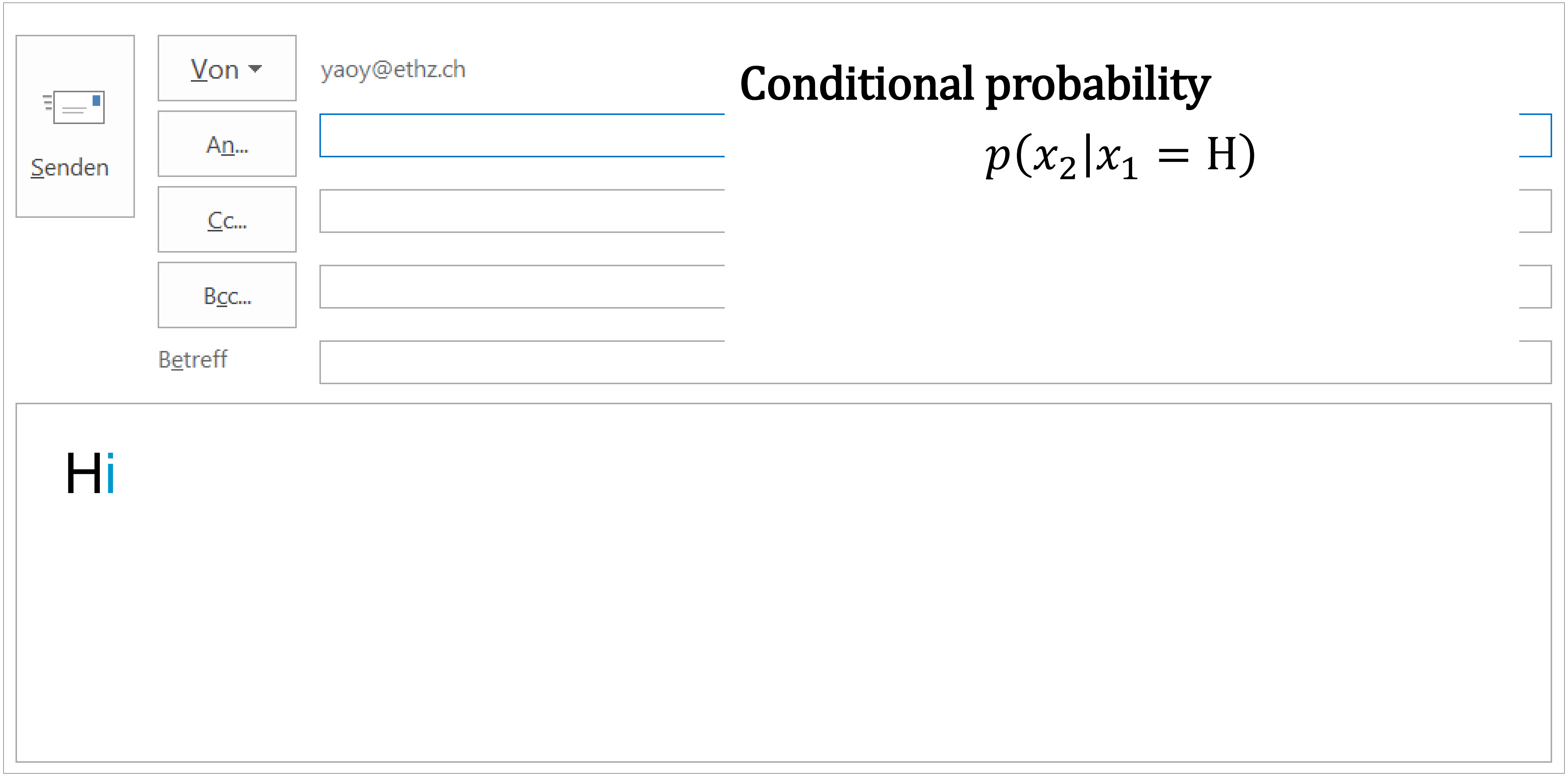


Conditional probability

$$p(x_2|x_1 = \mathrm{H})$$

Joint and Conditional Probability

example: text prediction

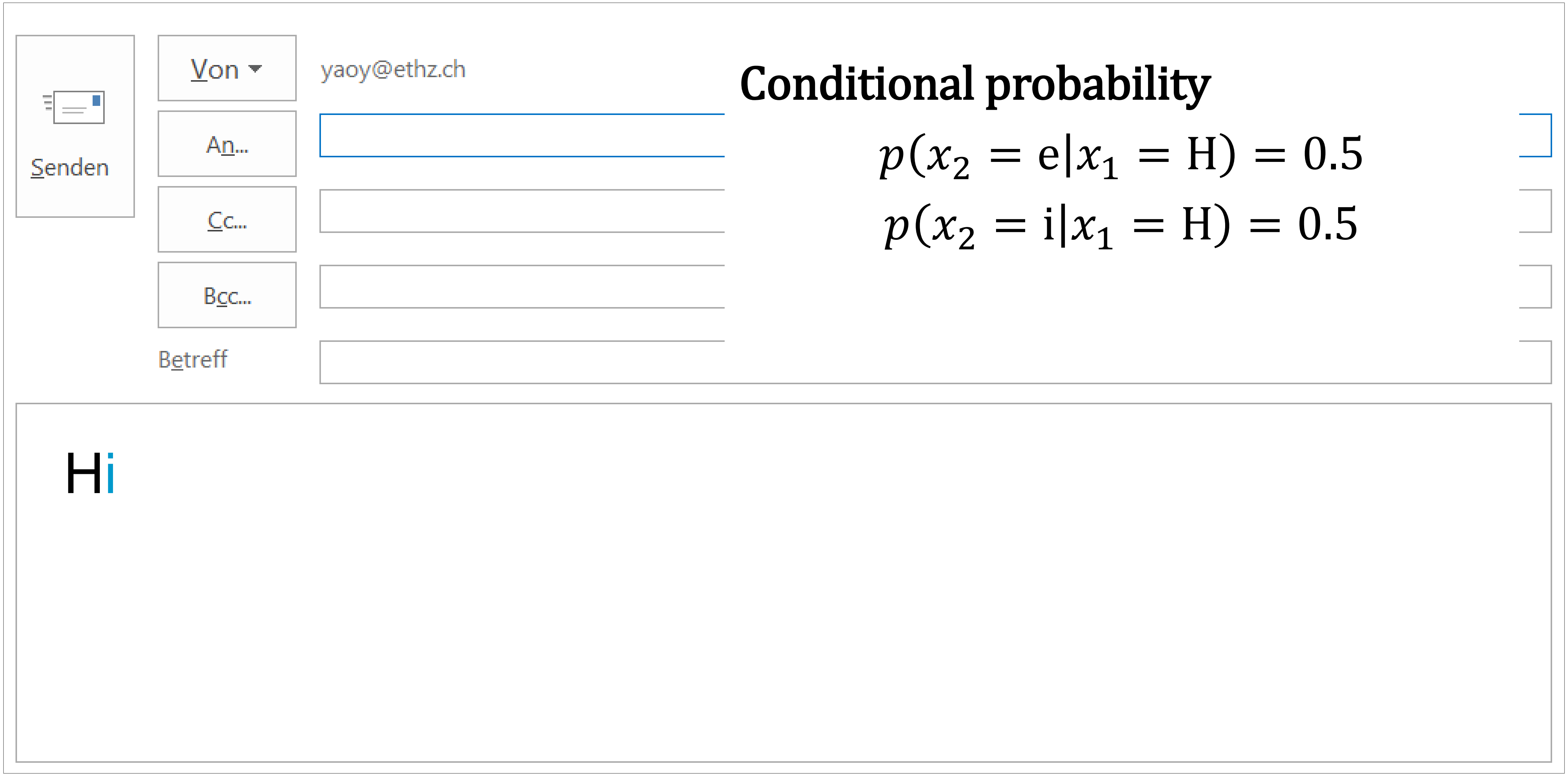


Conditional probability

$$p(x_2 | x_1 = \mathrm{H})$$

Joint and Conditional Probability

example: text prediction



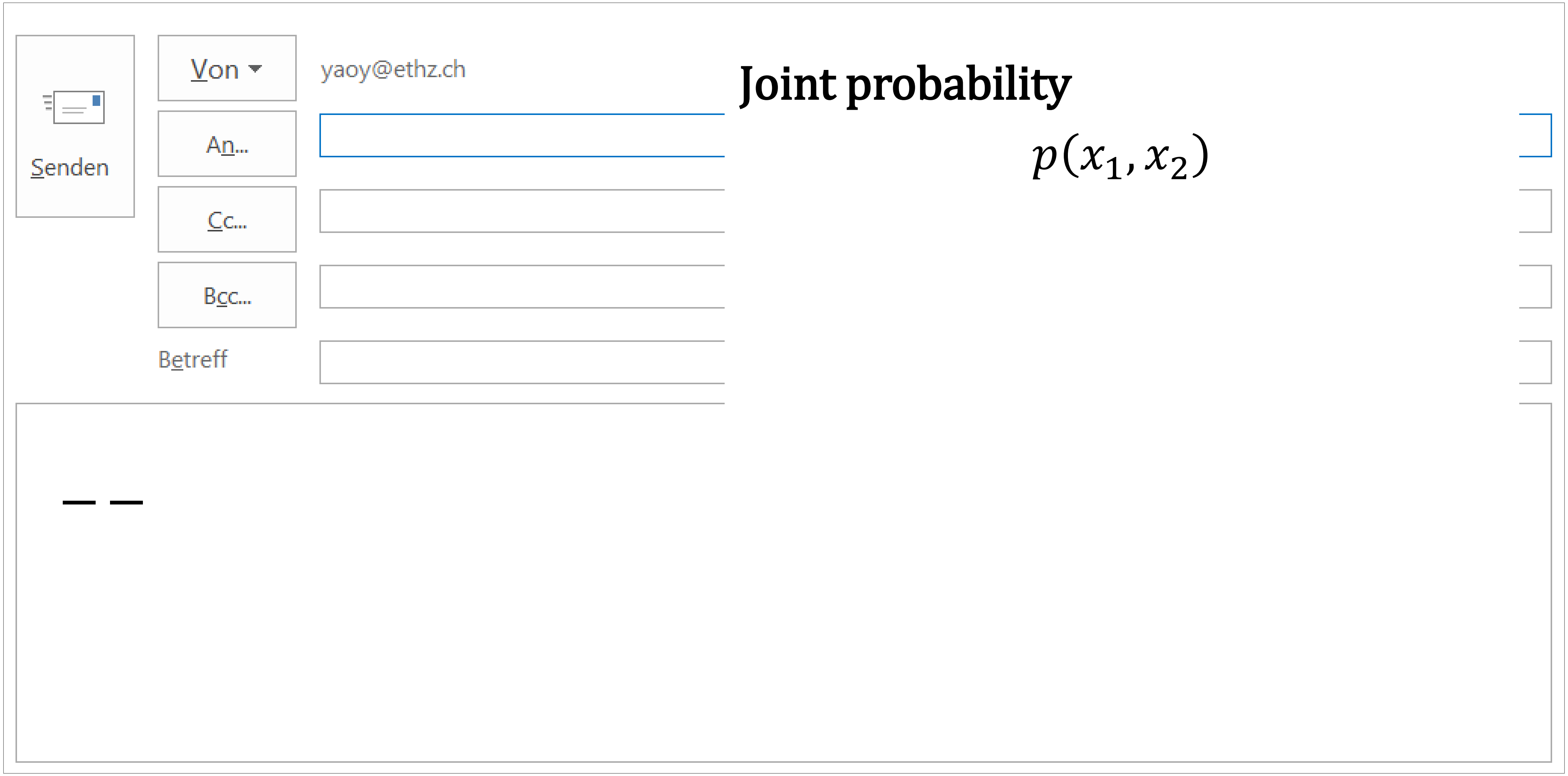
Conditional probability

$$p(x_2 = \mathrm{e} | x_1 = \mathrm{H}) = 0.5$$

$$p(x_2 = \mathrm{i} | x_1 = \mathrm{H}) = 0.5$$

Joint and Conditional Probability

example: text prediction

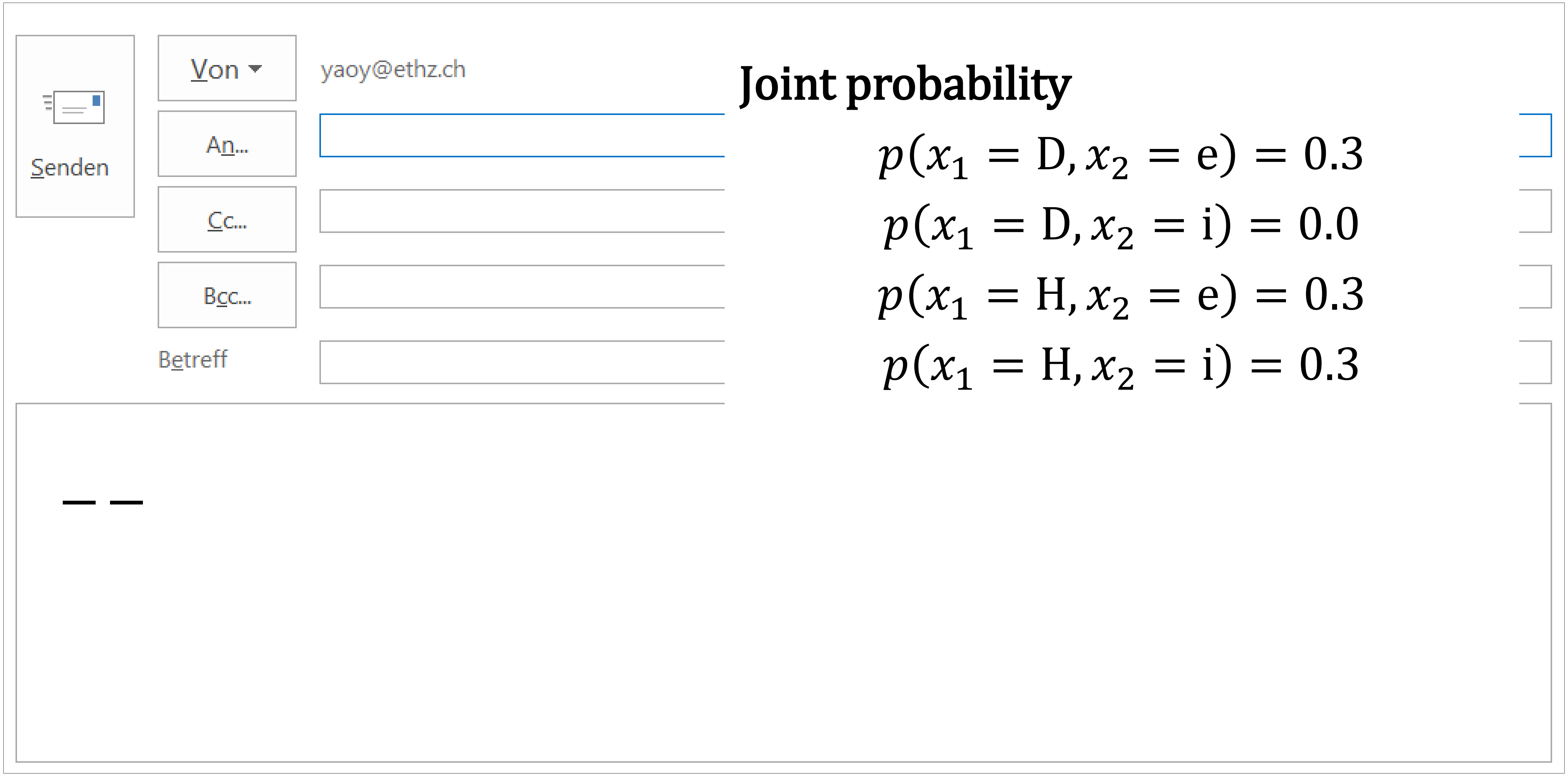


Joint probability

$$p(x_1, x_2)$$

Joint and Conditional Probability

example: text prediction



Joint probability

$$p(x_1 = \mathrm{D}, x_2 = \mathrm{e}) = 0.3$$
$$p(x_1 = \mathrm{D}, x_2 = \mathrm{i}) = 0.0$$
$$p(x_1 = \mathrm{H}, x_2 = \mathrm{e}) = 0.3$$
$$p(x_1 = \mathrm{H}, x_2 = \mathrm{i}) = 0.3$$

Joint and Conditional Probability

example: sensitivity and specificity

		True condition	
		positive	negative
Test	positive	True positive	False positive
	negative	False negative	True negative

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

"proportion of positives that are correctly identified"

Joint and Conditional Probability

example: sensitivity and specificity

		True condition	
		positive	negative
Test	positive	True positive	False positive
	negative	False negative	True negative

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

$$\approx p(test = pos | cond = pos)$$

Joint and Conditional Probability

example: sensitivity and specificity

		True condition	
		positive	negative
Test	positive	True positive	False positive
	negative	False negative	True negative

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

$$\approx p(test = pos|cond = pos)$$

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\approx p(cond = pos|test = pos)$$

Sum and Product Rule

Product Rule

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

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

$$\text{joint} = \text{conditional} \times \text{marginal}$$

Sum and Product Rule

Product Rule

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

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

$$\text{joint} \; = \; \text{conditional} \times \; \text{marginal}$$

Note: $p(x, y, z) = p(x|y, z)p(y|z)p(z)$

Sum and Product Rule

Product Rule

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

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

$$\text{joint} = \text{conditional} \times \text{marginal}$$

Sum Rule

$$p(x) = \sum_y p(x, y)$$

$$\text{marginal} = \sum_y \text{joint}$$

Note: the summation is over all possible outcomes of y (can be very large)

Sum and Product Rule

Product Rule

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

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

$$\text{joint} \; = \; \text{conditional} \times \; \text{marginal}$$

Sum Rule

$$p(x) = \sum_{y} p(x, y)$$

Also note: for continuous variables

$$p(x) = \int p(x, y)dy$$

Sum and Product Rule

Example: Bayes' rule

Product rule

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

Sum Rule

$$p(x) = \sum_{y} p(x, y)$$

Sum and Product Rule

Example: Bayes' rule

Product rule

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

Sum Rule

$$p(x) = \sum_y p(x, y)$$

$$p(x, y) = p(y|x) \sum_y p(x, y)$$

Sum and Product Rule

Example: Bayes' rule

Product rule

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

Sum Rule

$$p(x) = \sum_y p(x,y)$$

$$p(x,y) = p(y|x) \sum_y p(x,y)$$

$$p(y|x) = \frac{p(x,y)}{\sum_y p(x,y)}$$

Sum and Product Rule

Example: Bayes' rule

Product rule

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

Sum Rule

$$p(x) = \sum_y p(x, y)$$

$$p(x, y) = p(y|x) \sum_y p(x, y)$$

$$p(y|x) = \frac{p(x, y)}{\sum_y p(x, y)}$$

$$p(y|x) = \frac{p(x|y)p(y)}{\sum_y p(x, y)}$$

Bayes' rule

Model identification and fitting

- 3 cards example revisited
- Likelihood vs probability
- Overfitting

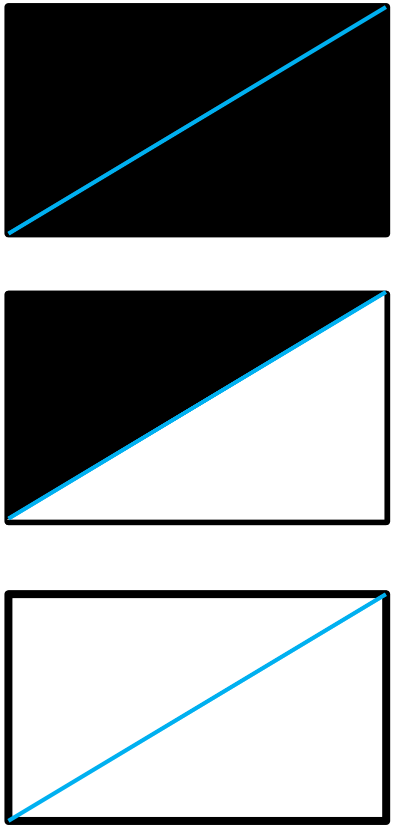
References:

C. Bishop “Pattern Recognition and Machine Learning”

D. MacKay “Information Theory, Inference, and Learning Algorithms”

Cards example revisited

see: D. MacKay *Information Theory, Inference, and Learning Algorithms*



Rules:

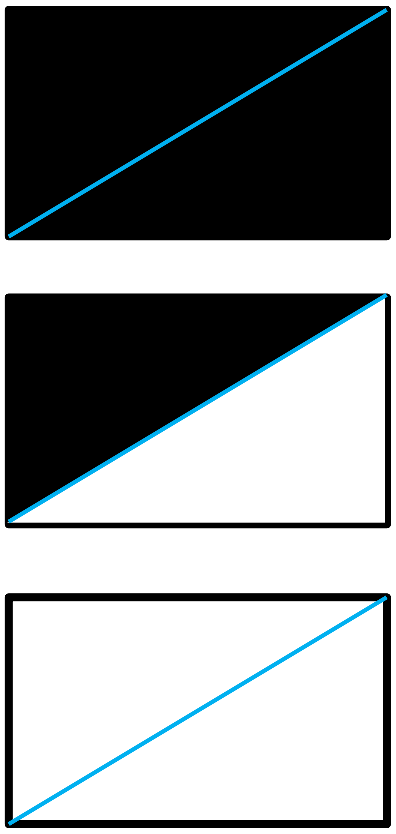
1. shuffle cards
2. draw 1 card at random
3. choose random side

… side is **black**

Question:

probability that other side is **white**

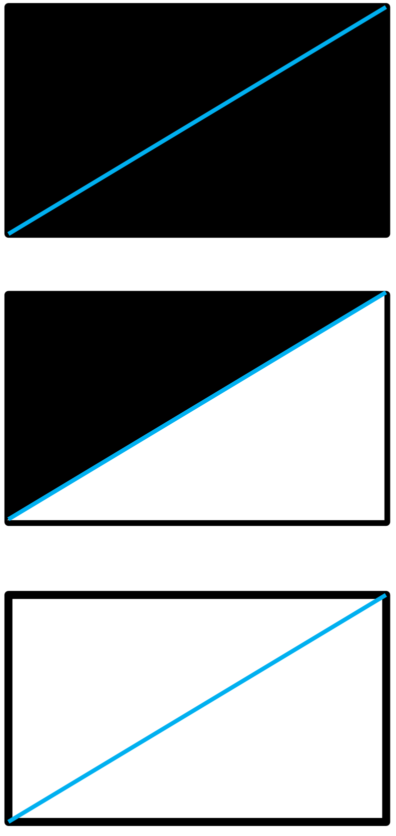
Cards example revisited



Model identification:

1. relevant variables and sample space
   a) card (c): #1, #2, #3
   b) visible side (v): white, black
   c) hidden side (h): white, black

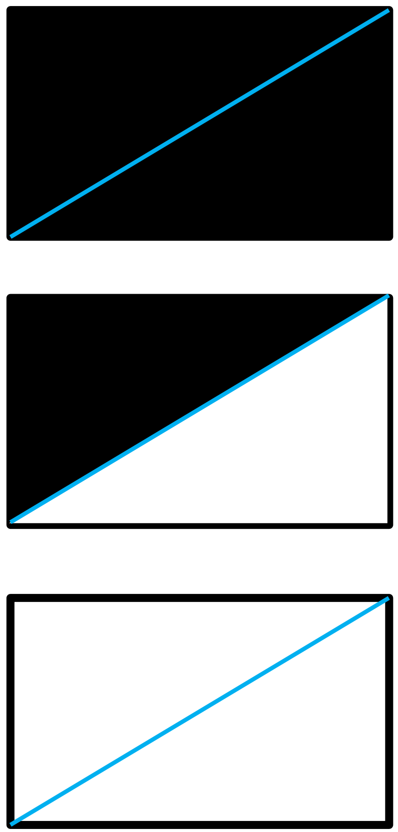
Cards example revisited



Model identification:

1. relevant variables: card, visible, hidden
2. dependency structure and probabilities
   a) card:

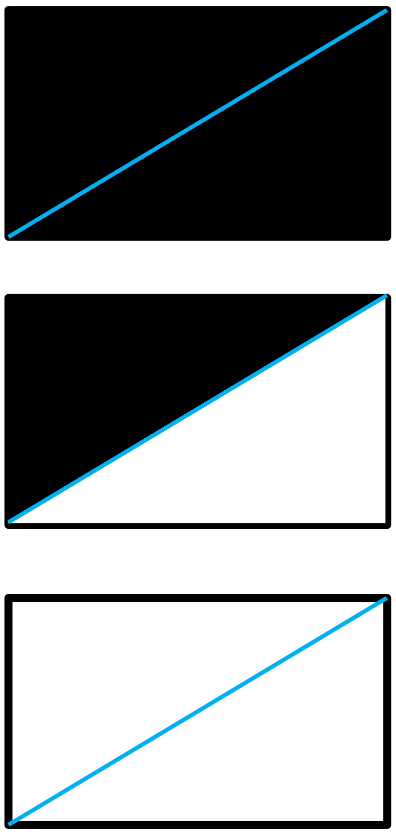
Cards example revisited



Model identification:

1. relevant variables: card, visible, hidden
2. dependency structure and probabilities
   a) card: $p(c) = \frac{1}{3}$, $c \in \{1,2,3\}$

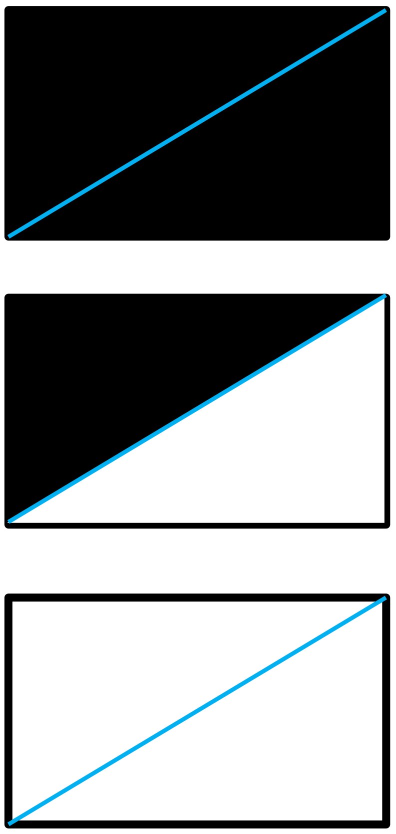
Cards example revisited



Model identification:

1. relevant variables: card, visible, hidden
2. dependency structure and probabilities
   a) card: $p(c) = \frac{1}{3}$, $c \in \{1,2,3\}$
   b) visible:

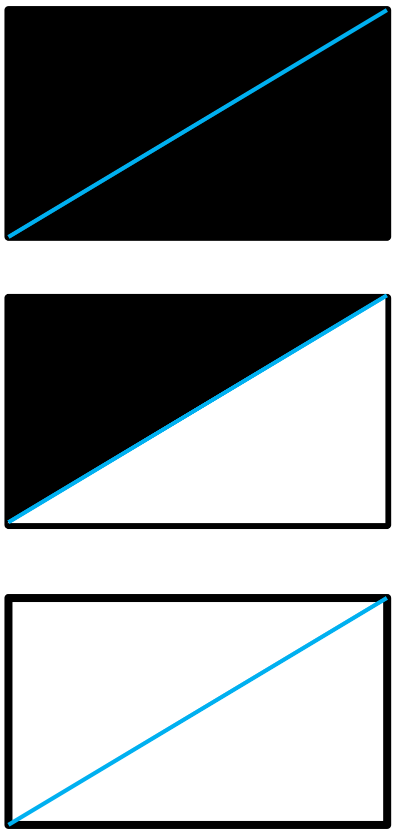
Cards example revisited



Model identification:

1. relevant variables: card, visible, hidden
2. dependency structure and probabilities
   a) card: $p(c) = \frac{1}{3}$, $c \in \{1,2,3\}$
   b) visible: e.g: $p(v = b | c = 1) = 1$

Cards example revisited

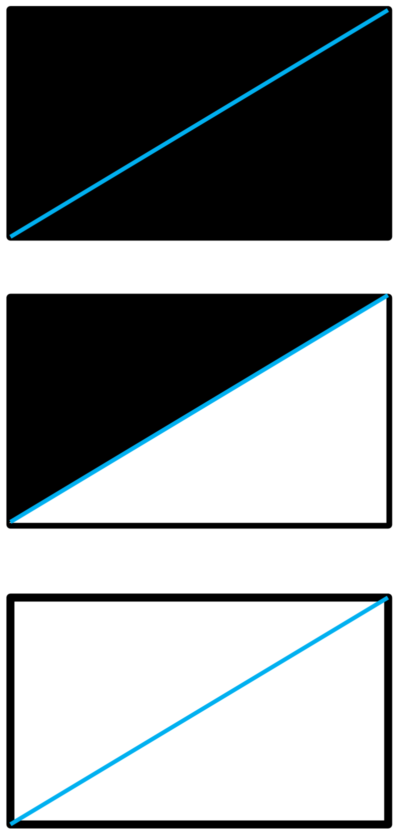


Model identification:

1. relevant variables: card, visible, hidden
2. dependency structure and probabilities
   a) card: $p(c) = \frac{1}{3}$, $c \in \{1,2,3\}$
   b) visible: e.g: $p(v = b|c = 1) = 1$

$\boldsymbol{v}$	$\boldsymbol{c}$	$\boldsymbol{p(v\|c)}$
b	1	1
b	2	0.5
b	3	0

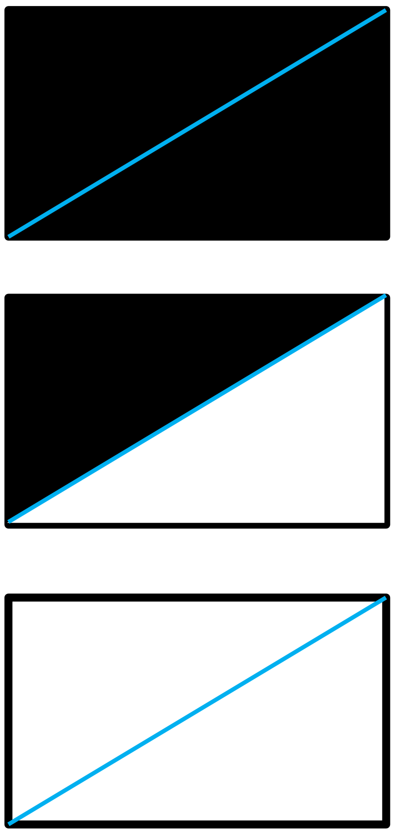
Cards example revisited



Model identification:

1. relevant variables: card, visible, hidden
2. dependency structure and probabilities
   a) card: $p(c) = \frac{1}{3}$, $c \in \{1,2,3\}$
   b) visible: e.g: $p(v = b|c = 1) = 1$
   c) hidden:

Cards example revisited

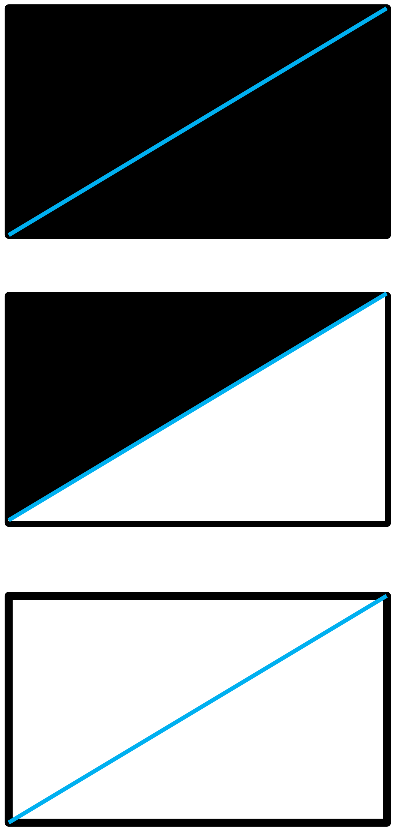


Model identification:

1. relevant variables: card, visible, hidden
2. dependency structure and probabilities
   a) card: $p(c) = \frac{1}{3}$, $c \in \{1,2,3\}$
   b) visible: e.g: $p(v = b|c = 1) = 1$
   c) hidden:

$\boldsymbol{h}$	$\boldsymbol{v}$	$\boldsymbol{c}$	$\boldsymbol{p(h\|v,c)}$
w	b	1	0
w	b	2	1
…	…	…	…

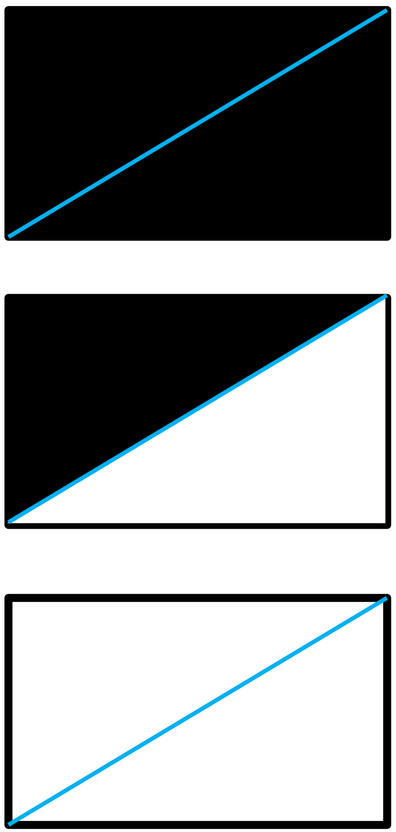
Cards example revisited



Model identification:

1. relevant variables: card, visible, hidden
2. dependency structure and probabilities
   a) card: $p(c) = \frac{1}{3}$, $c \in \{1,2,3\}$
   b) visible: e.g: $p(v = b|c = 1) = 1$
   c) hidden: $p(h|v, c)$
   d) joint: $p(h, v, c) = p(h|v, c)p(v|c)p(c)$

Cards example revisited



Model identification:

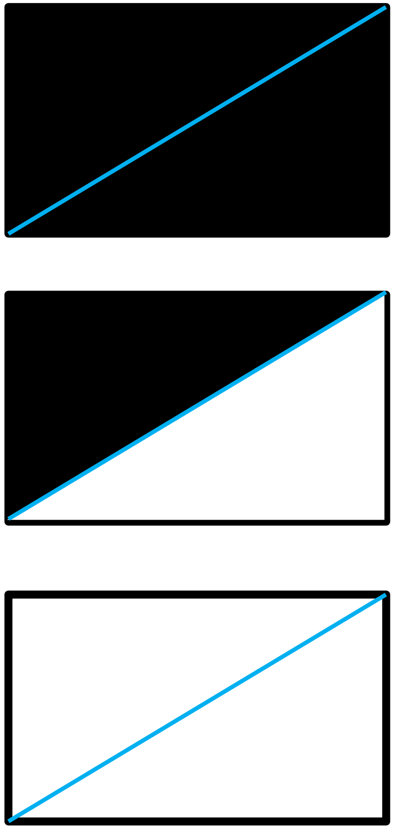
1. relevant variables: card, visible, hidden
2. dependency structure and probabilities

   $$p(h, v, c) = p(h|v, c)p(v|c)p(c)$$

3. question:

   probability that other side is **white**, if visible side is **black**

Cards example revisited



Model identification:

1. relevant variables: card, visible, hidden
2. dependency structure and probabilities

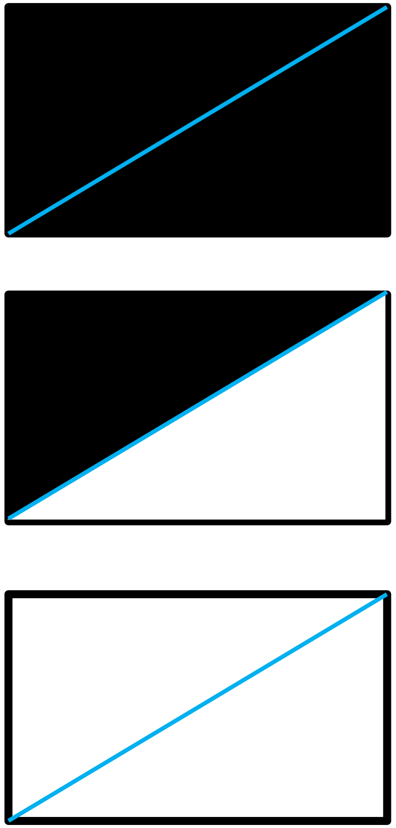
   $$p(h, v, c) = p(h|v, c)p(v|c)p(c)$$

3. question:

   probability that other side is **white**, if visible side is **black**

   $$p(h = white|v = black)$$

Cards example revisited



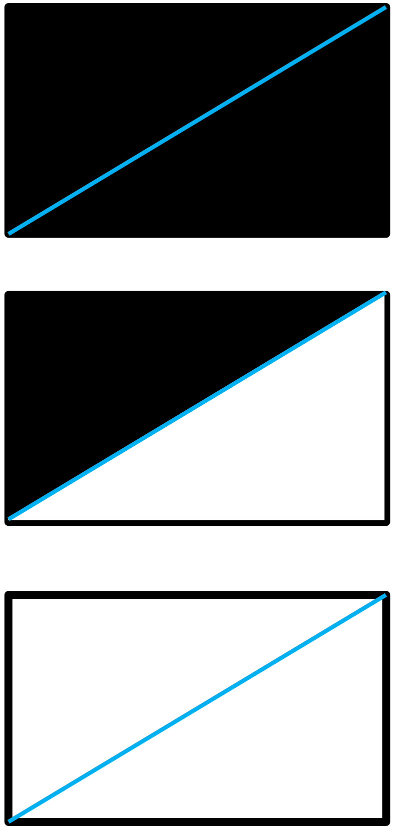
probability that other side is **white**, if visible side is **black**

$$p(h = white|v = black)$$

joint:

$$p(h, v, c) = \ p(h|v, c)p(v|c)p(c)$$

Cards example revisited



probability that other side is **white**, if visible side is **black**

$$p(h = white|v = black)$$

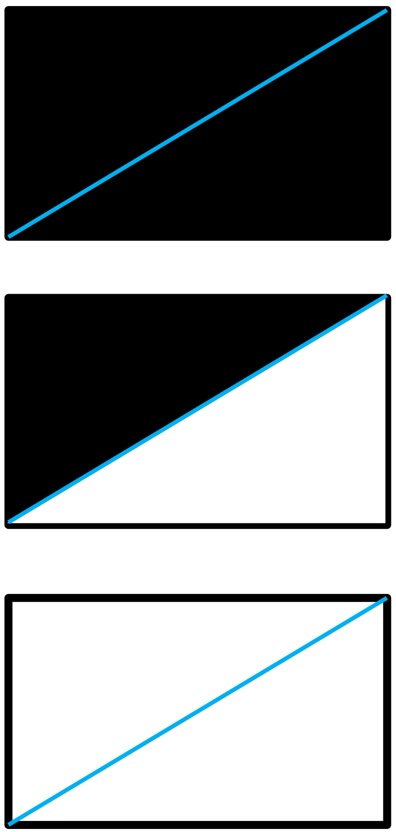
joint:

$$p(h, v, c) = p(h|v, c)p(v|c)p(c)$$

sum rule:

$$p(h, v) = \sum_c p(h, v, c)$$

Cards example revisited



probability that other side is **white**, if visible side is **black**

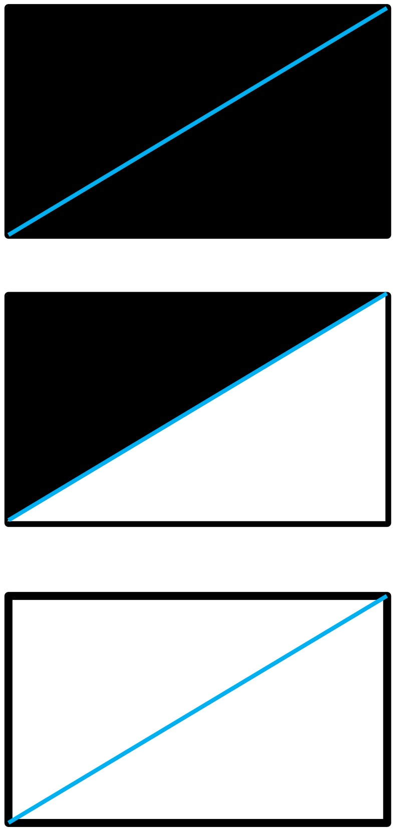
$$p(h = white|v = black)$$

joint:

$$p(h, v, c) = \ p(h|v, c)p(v|c)p(c)$$

$\boldsymbol{h}$	$\boldsymbol{v}$	$\boldsymbol{c}$	$\boldsymbol{p(c)}$	$\boldsymbol{p(v\|c)}$	$\boldsymbol{p(h\|v,c)}$	$\boldsymbol{p(h,v,c)}$
w	b	1	1/3	1	0	0
w	b	2	1/3	1/2	1	1/6
w	b	3	1/3	0	0	0
Sum rule					$\boldsymbol{p(h = w, v = b)}$	1/6

Cards example revisited



probability that other side is **white**, if visible side is **black**

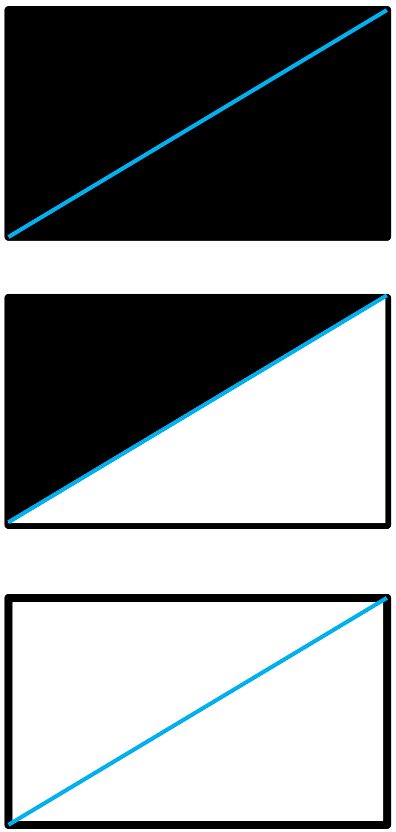
$$p(h = white|v = black)$$

joint:

$$p(h, v, c) = p(h|v, c)p(v|c)p(c)$$

$$p(h = w, v = b) = 1/6$$

Cards example revisited



probability that other side is **white**, if visible side is **black**

$$p(h = white|v = black)$$

joint:

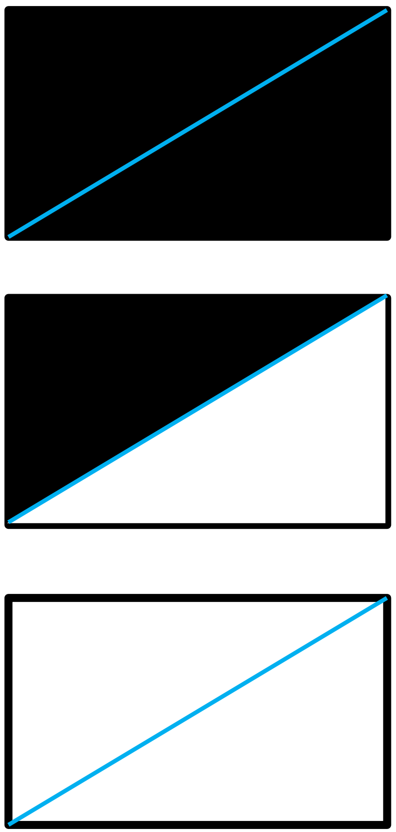
$$p(h, v, c) = \; p(h|v, c)p(v|c)p(c)$$

$$p(h = w, v = b) = 1/6$$

Bayes' rule:

$$p(h|v) = \frac{p(h, v)}{p(v)}$$

Cards example revisited



probability that other side is **white**, if visible side is **black**

$$p(h = white|v = black)$$

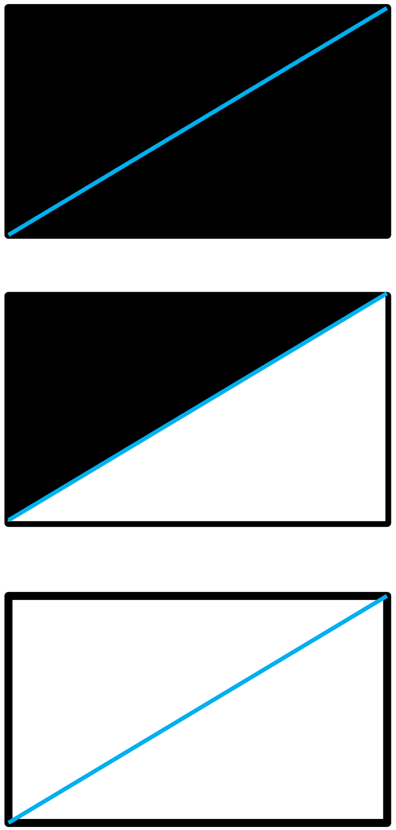
joint:

$$p(h, v, c) = p(h|v, c)p(v|c)p(c)$$

sum rule:

$$p(v) = \sum_c p(v, c) = \sum_c p(v|c)p(c)$$

Cards example revisited



probability that other side is **white**, if visible side is **black**

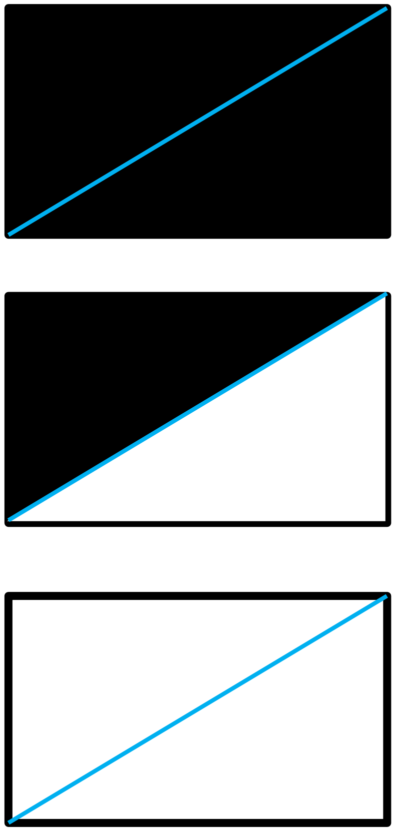
$$p(h = white | v = black)$$

joint:

$$p(h, v, c) = \ p(h|v, c)p(v|c)p(c)$$

$\boldsymbol{v}$	$\boldsymbol{c}$	$\boldsymbol{p(c)}$	$\boldsymbol{p(v\|c)}$	$\boldsymbol{p(v, c)}$
b	1	1/3	1	1/3
b	2	1/3	1/2	1/6
b	3	1/3	0	0
Sum rule			$\boldsymbol{p(v = b)}$	1/2

Cards example revisited



probability that other side is **white**, if visible side is **black**

$$p(h = white | v = black)$$

joint:

$$p(h, v, c) = p(h|v, c)p(v|c)p(c)$$

$$p(h = w, v = b) = 1/6$$

$$p(v = b) = 1/2$$

Bayes' rule: $p(h = w|v = b) = \dfrac{p(h = w, v = b)}{p(v = b)} = \dfrac{1/6}{1/2}$

$$= \frac{1}{3}$$

Likelihood vs probability

In statistics, one distinguishes between **probability** and **likelihood**

Example: 3 cards

$$X \in \{black, white\}$$

(conditional) **probability**

$$f(X) = p(v = X | card = 2)$$

likelihood function:

$$L(X) = p(v = black | card = X)$$

$$X \in \{1,2,3\}$$

Likelihood vs probability

In statistics, one distinguishes between **probability** and **likelihood**

Example: 3 cards

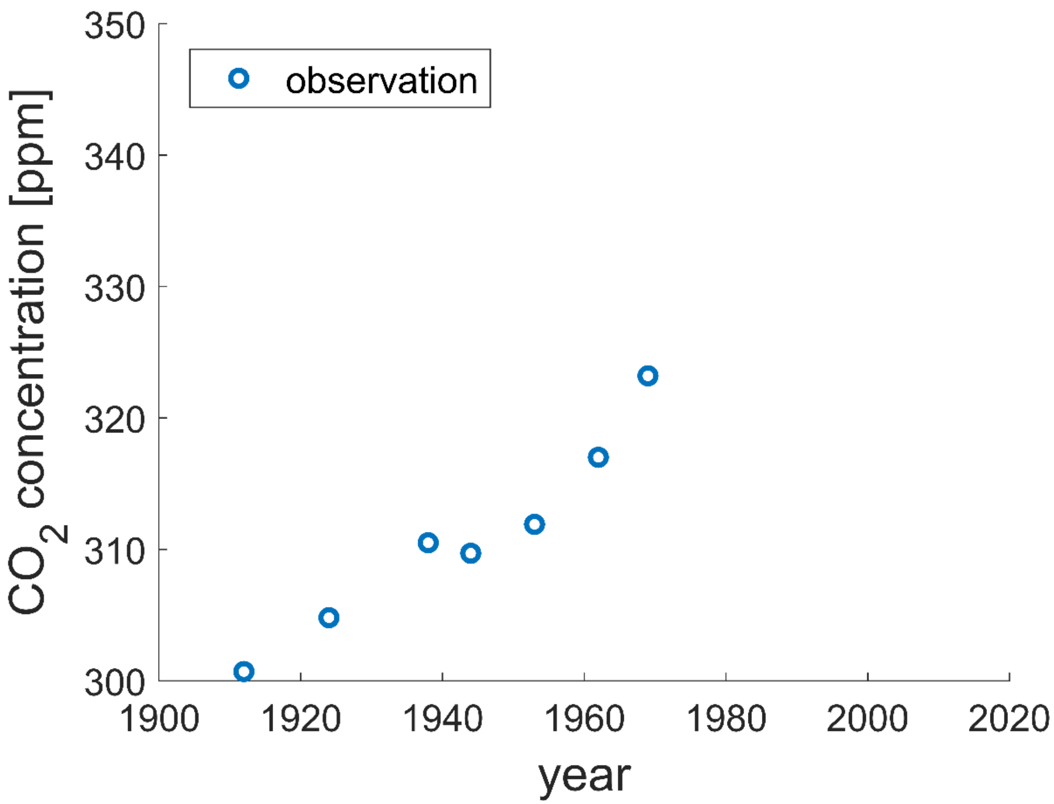
Note: unlike probabilities, the likelihood is not normalized

$$f(black) + f(white) = 1/2 + 1/2 = 1$$

$$L(1) + L(2) + L(3) = 1 + \frac{1}{2} + 0 \neq 1$$

Overfitting

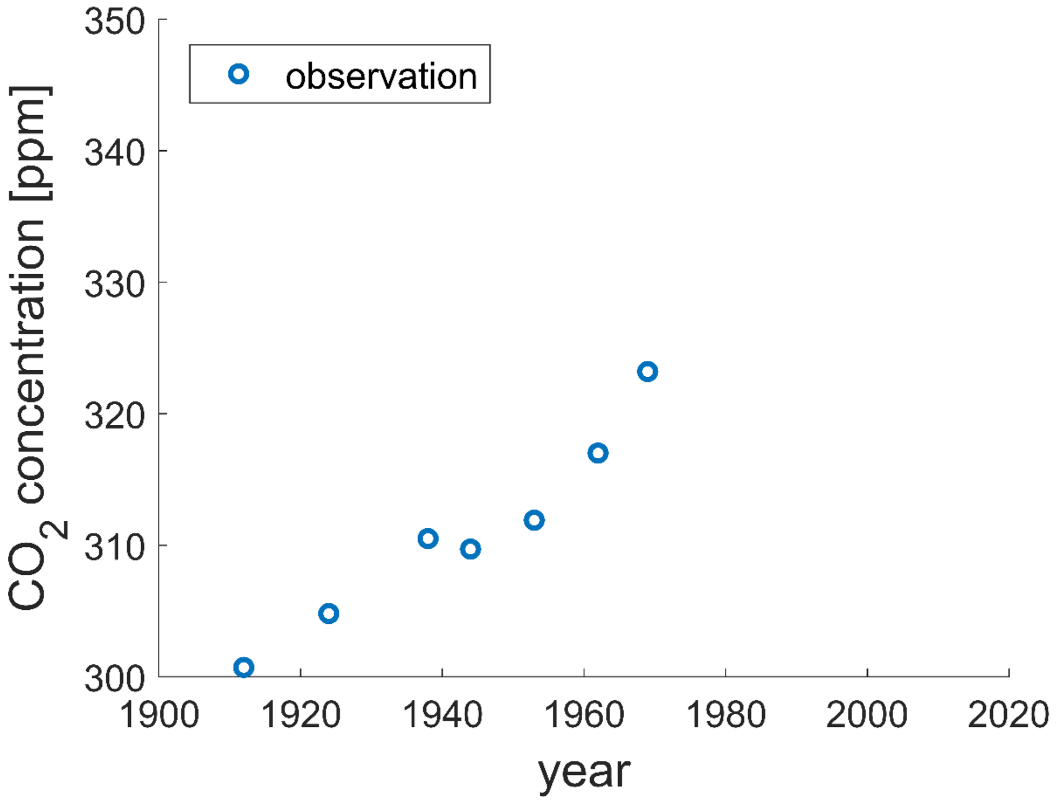
Data: Atmospheric CO_2 concentration



Data from: Etheridge et al. (1998)

Overfitting

Data: Atmospheric CO_2 concentration



Model: 1st order polynomial

$$y = \theta_1 t + \theta_0 + e$$

$$p(e) = N(e \mid 0, 1)$$

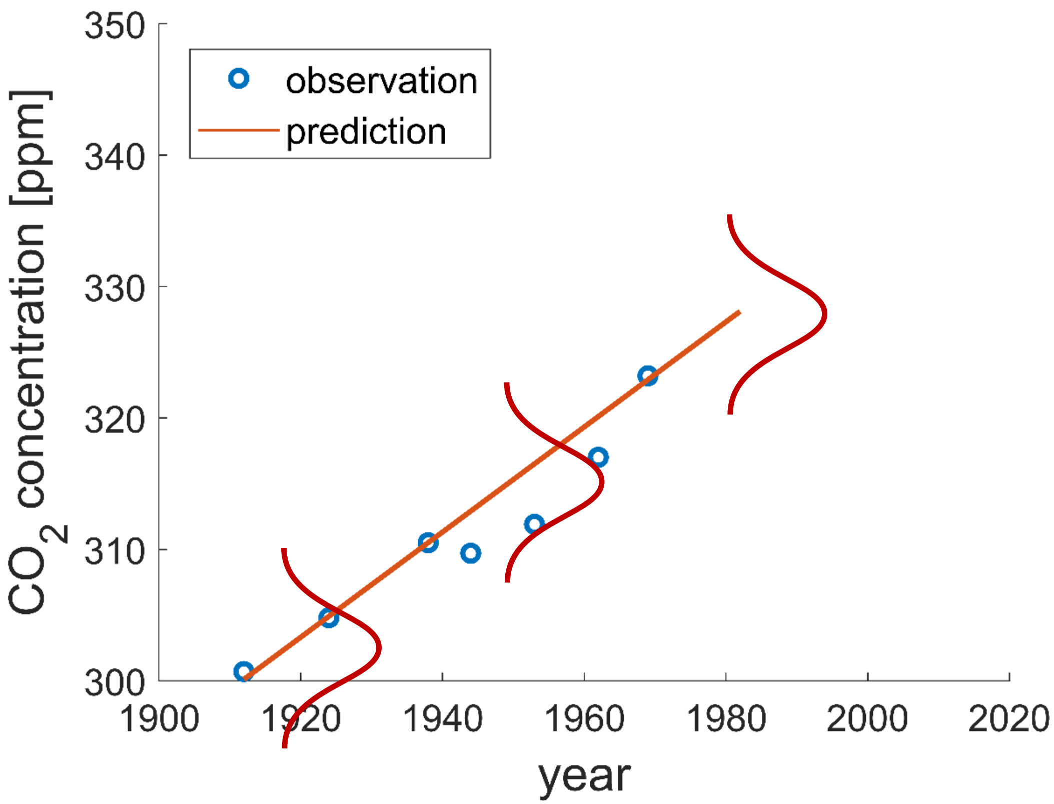
Likelihood:

$$L(\theta) = p(y \mid \theta)$$
$$= N(y \mid \theta_1 t + \theta_0, 1)$$

Data from: Etheridge et al. (1998)

Overfitting

Data: Atmospheric CO_2 concentration



Model: 1st order polynomial

$$y = \theta_1 t + \theta_0 + e$$

$$p(e) = N(e \mid 0, 1)$$

Likelihood:

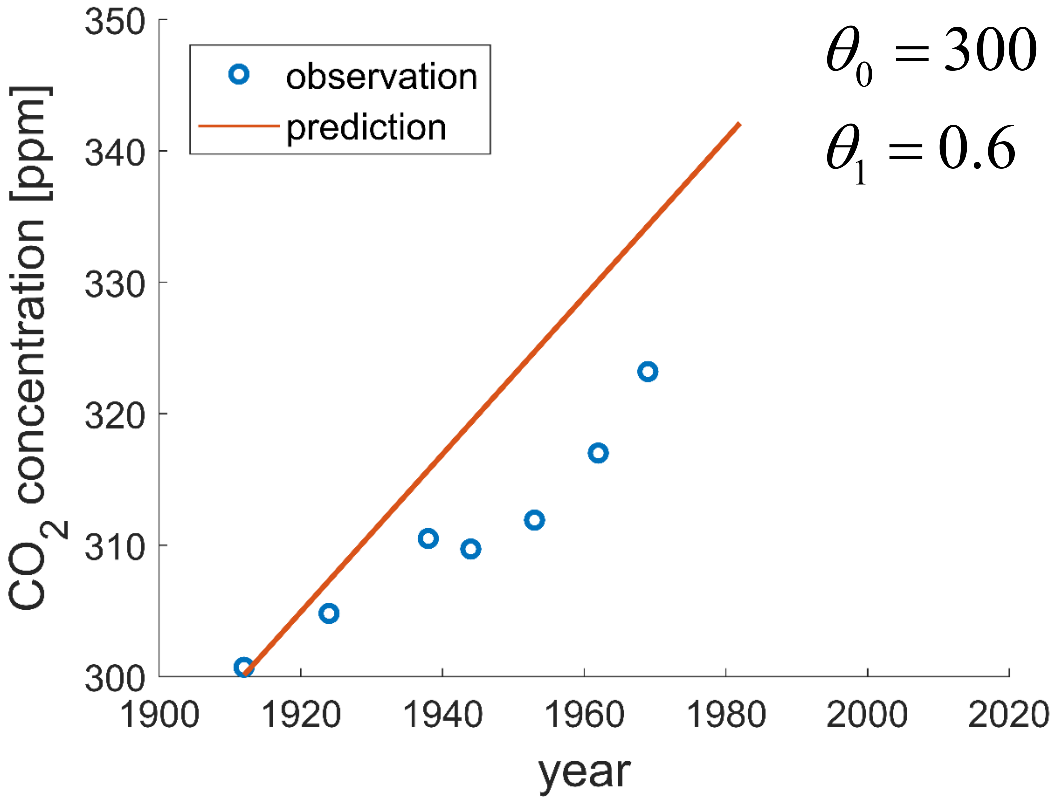
$$L(\theta) = p(y \mid \theta)$$

$$= N(y \mid \theta_1 t + \theta_0, 1)$$

Data from: Etheridge et al. (1998)

Overfitting

Data: Atmospheric CO_2 concentration

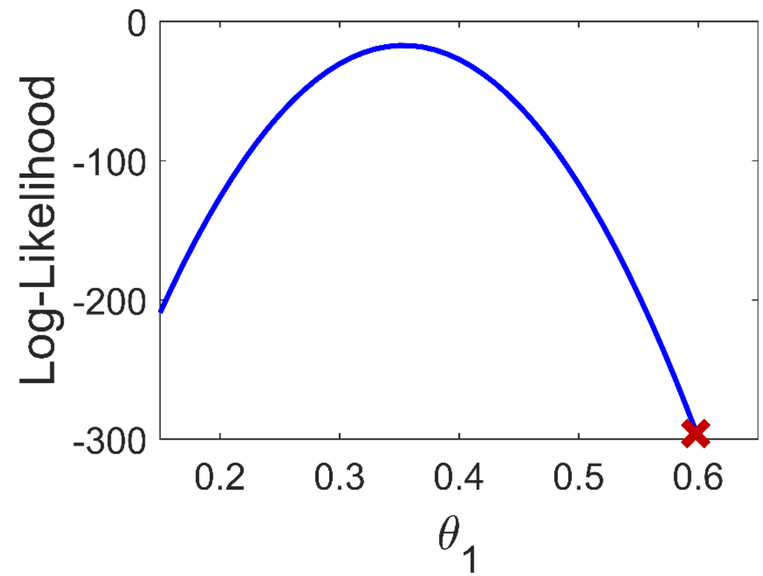


$$\theta_0 = 300$$
$$\theta_1 = 0.6$$

Model: 1st order polynomial

$$y = \theta_1 t + \theta_0 + e$$
$$p(e) = N(e \mid 0,1)$$

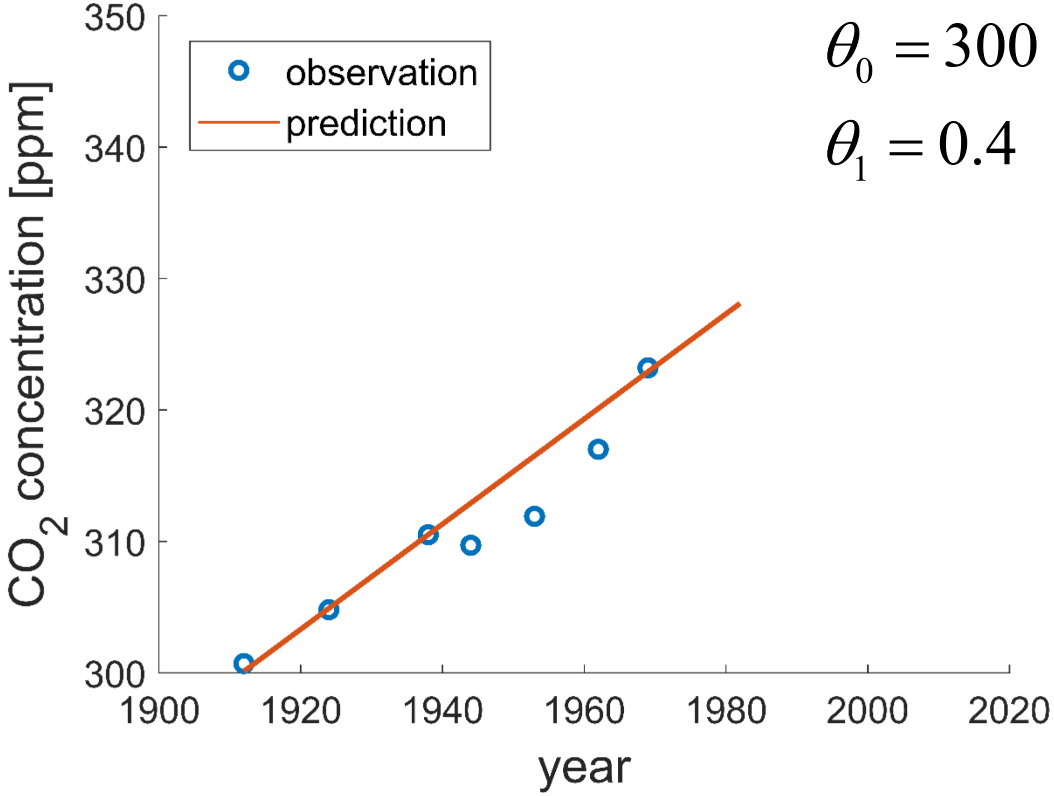
Likelihood:



Data from: Etheridge et al. (1998)

Overfitting

Data: Atmospheric CO_2 concentration

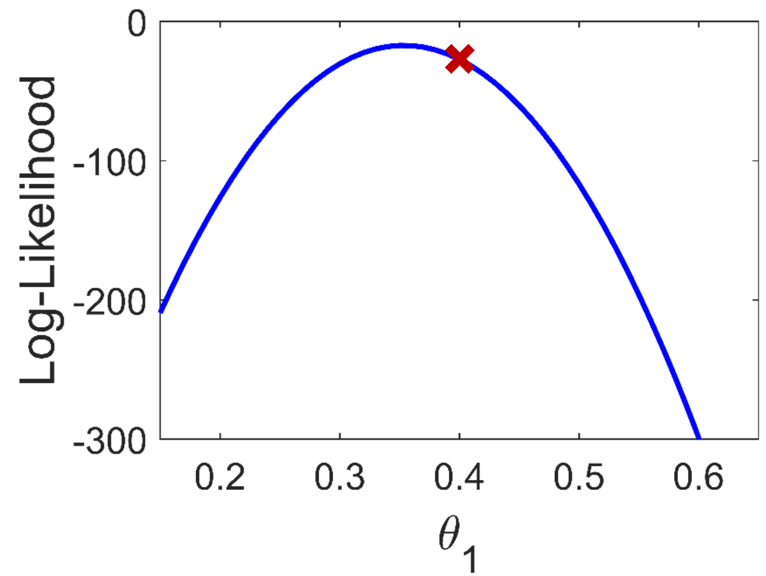


$$\theta_0 = 300$$
$$\theta_1 = 0.4$$

Model: 1st order polynomial

$$y = \theta_1 t + \theta_0 + e$$
$$p(e) = N(e \mid 0,1)$$

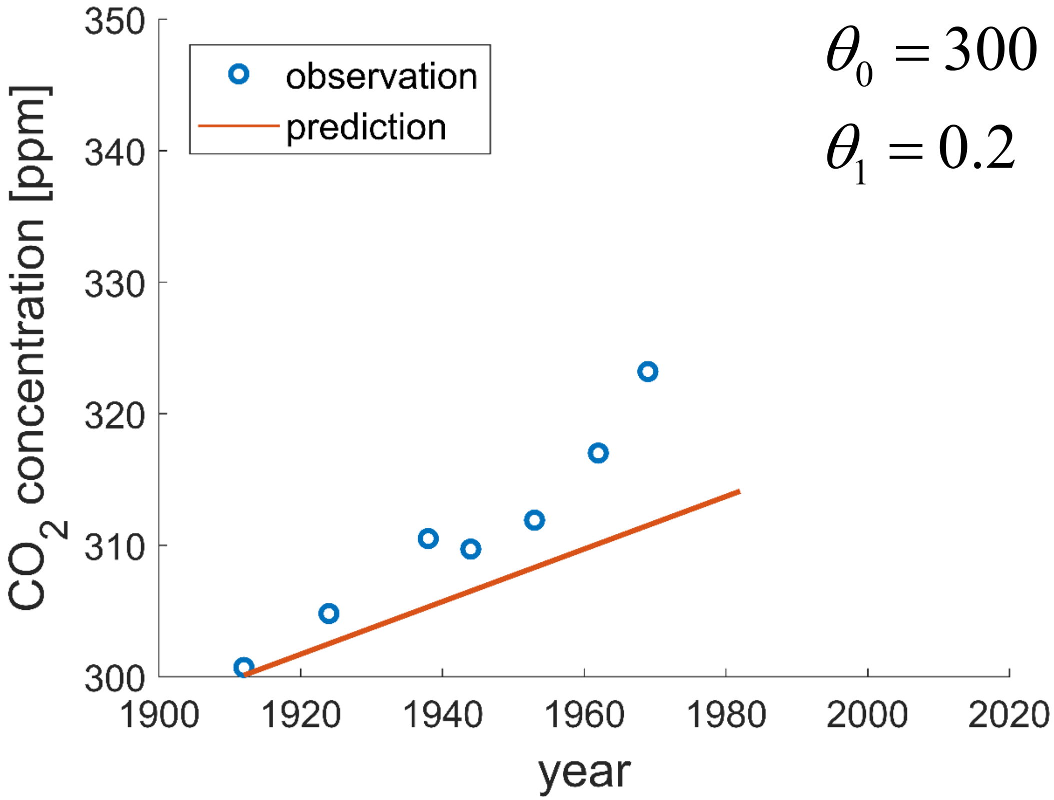
Likelihood:



Data from: Etheridge et al. (1998)

Overfitting

Data: Atmospheric CO_2 concentration

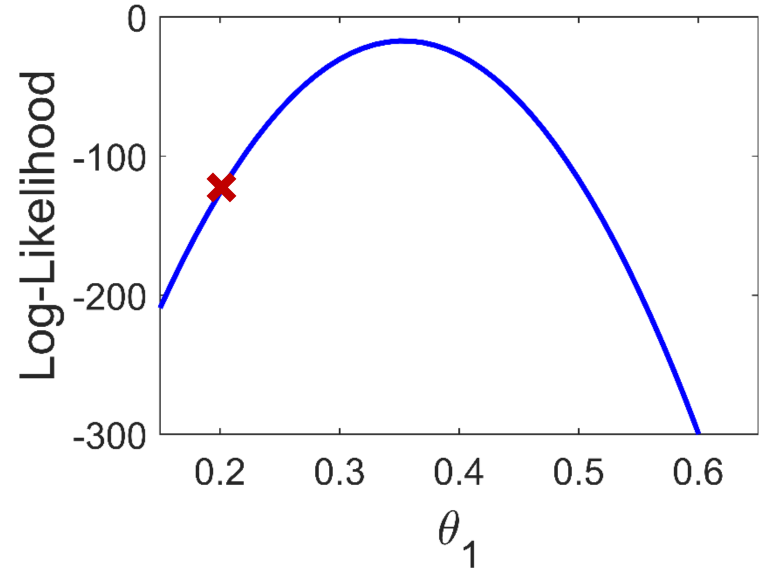


$$\theta_0 = 300$$
$$\theta_1 = 0.2$$

Model: 1st order polynomial

$$y = \theta_1 t + \theta_0 + e$$
$$p(e) = N(e \mid 0,1)$$

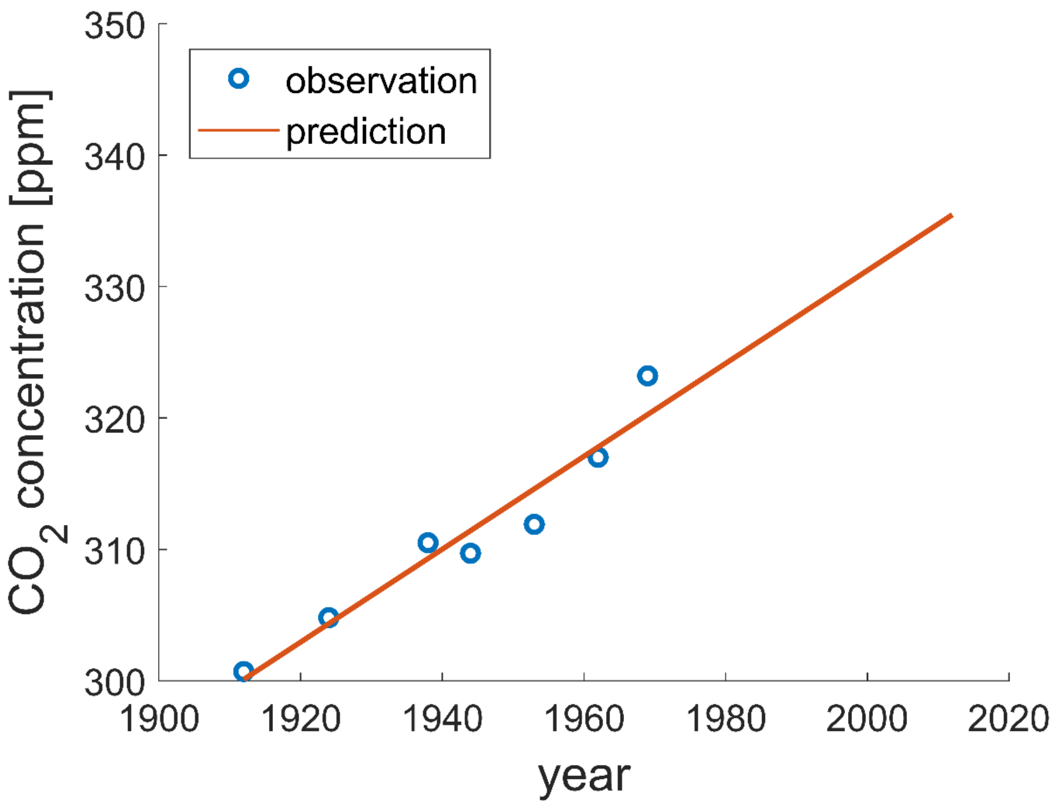
Likelihood:



Data from: Etheridge et al. (1998)

Overfitting

Data: Atmospheric CO_2 concentration

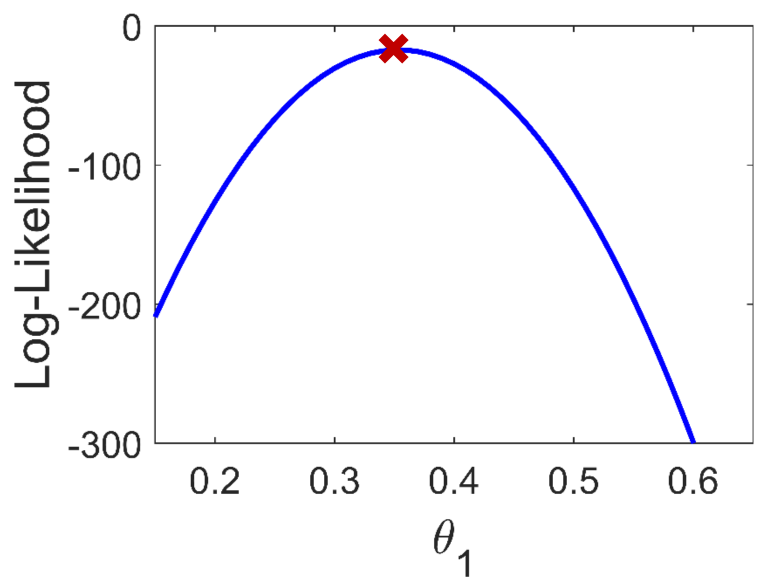


Model: 1st order polynomial

$$y = \theta_1 t + \theta_0 + e$$

$$p(e) = N(e \mid 0,1)$$

Likelihood:



Data from: Etheridge et al. (1998)

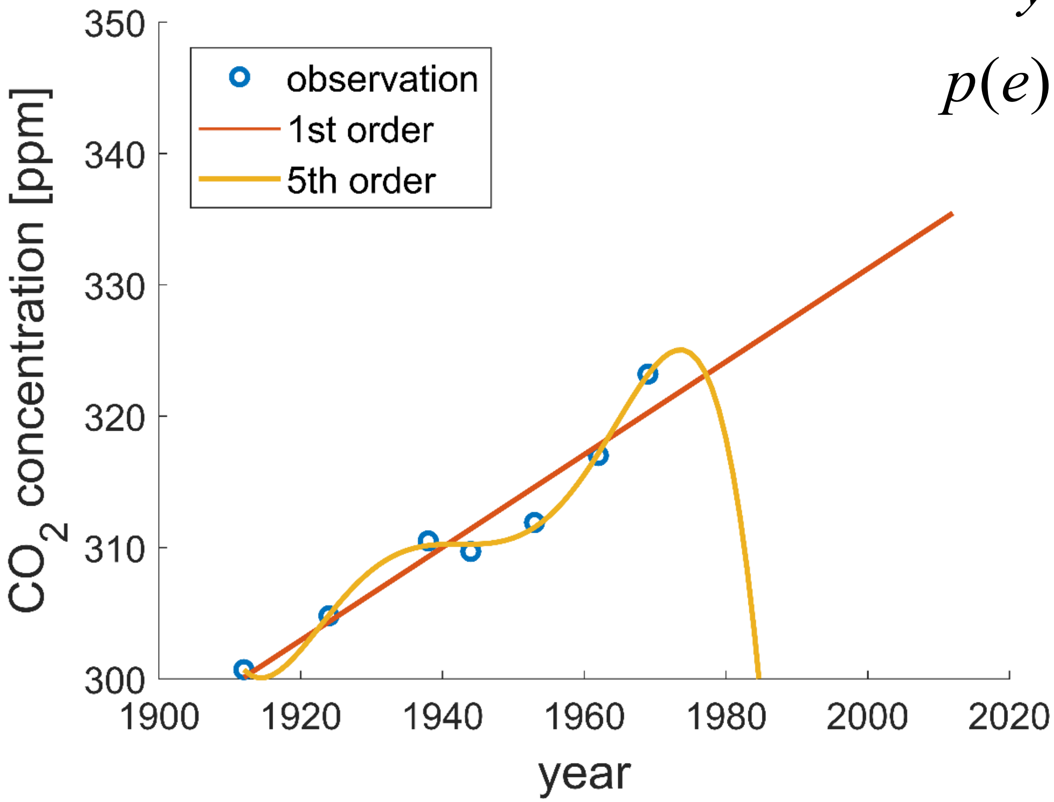
Overfitting

Data: Atmospheric CO_2 concentration

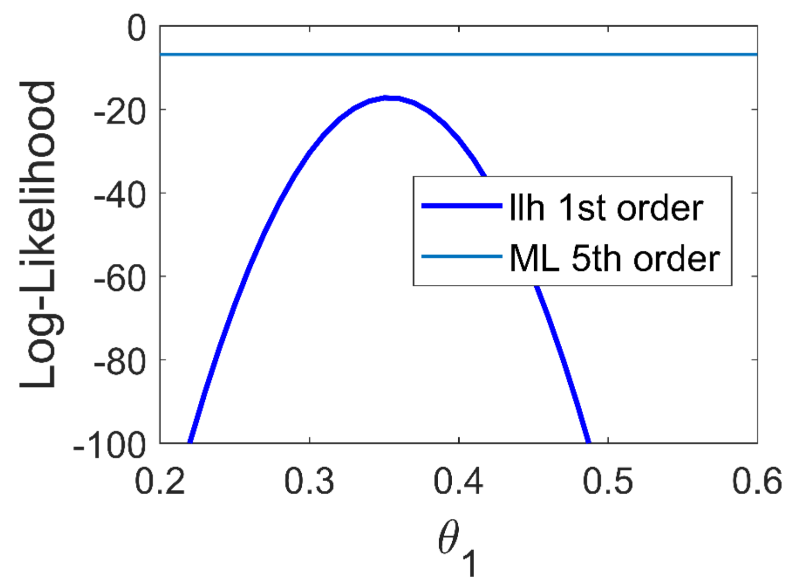
Model: 5th order polynomial

$$y = \theta_5 t^5 + \theta_4 t^4 + \theta_3 t^3 + \theta_2 t^2 + \theta_1 t + \theta_0 + e$$

$$p(e) = N(e \mid 0,1)$$



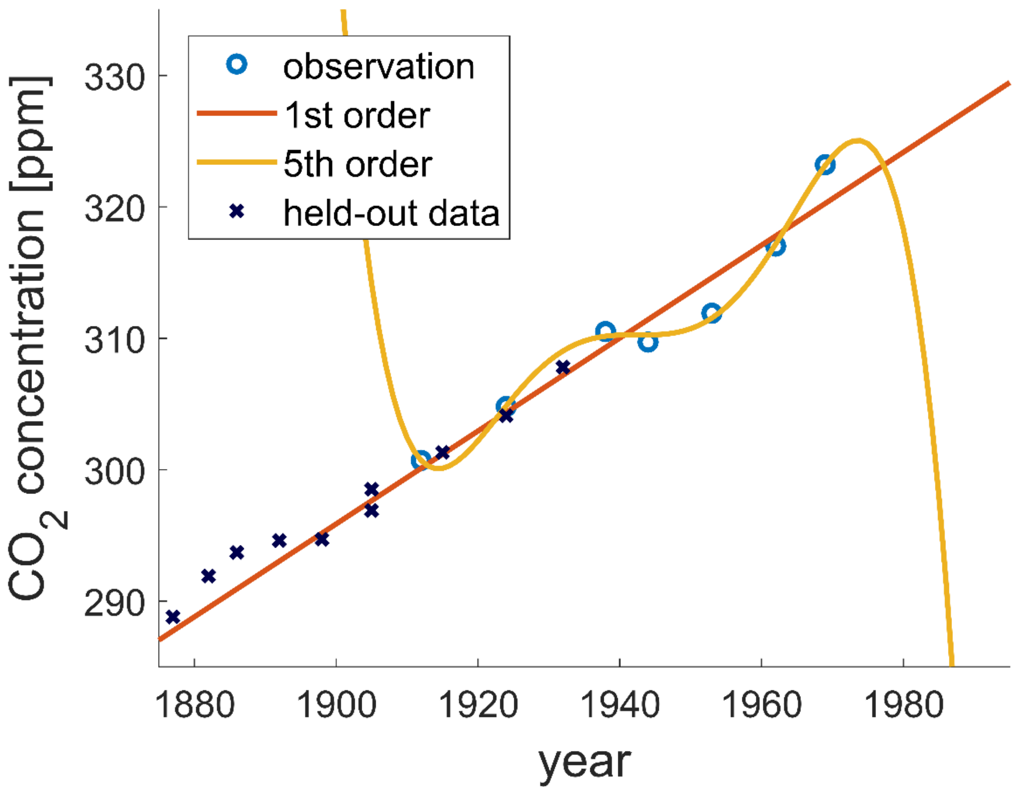
Likelihood:



Data from: Etheridge et al. (1998)

Overfitting

Data: Atmospheric CO_2 concentration

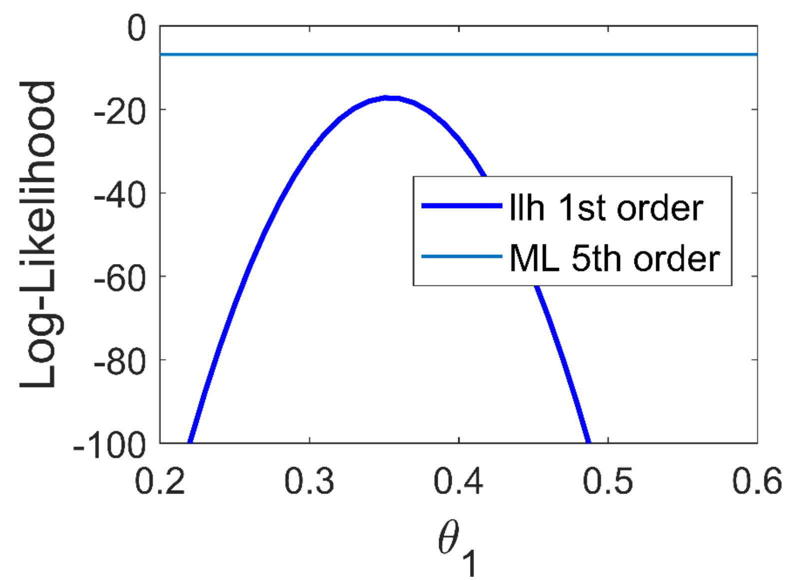


Log-likelihood on held-out data:

1st order: -18.7

5th order: -4.3×10^5

Likelihood:



Data from: Etheridge et al. (1998)

Further Reading

- Bishop: *Pattern Recognition and Machine Learning*
  - chapters 1 and 2, appendix B
- MacKay: *Information Theory, Inference, and Learning Algorithms*
  - *pages: 3 – 64, chapter 23*
  - *http://www.inference.org.uk/itprnn/book.pdf*
- Gelman: *Bayesian Data Analysis*
  - *appendix A*

Thank you