https://uva-iai.github.io

Apply for the Inclusive AI Mentorship program

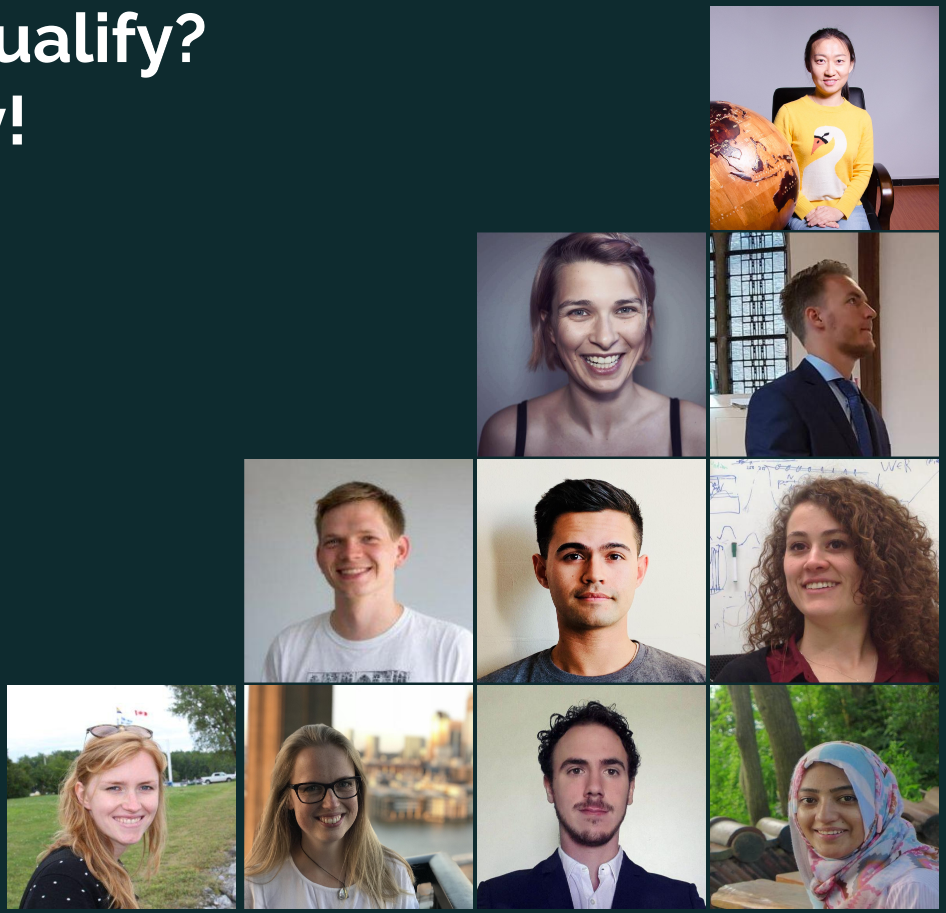
Unique opportunity to get advice from senior peers (PhDs, postdocs and assistant profs) from academia and industry!

For underrepresented groups

Unsure if you qualify? Apply anyway!

Mentees should expect to be able to:

- ask mentor for practical advice, e.g. how to write a CV or motivation letter, where to apply for jobs, when to apply for PhDs
- network with other students
- seek non-academic advice from mentor
- learn how to present their research
- connect with other researchers



Machine Learning 1

- Lecture 2 -

Supervised Learning: Linear Regression

-Patrick Forré-

Slides created by:
Rianne van den Berg

Image credit: Kirillm | Getty Images

Discussion forum

Piazza:

piazza.com/university_of_amsterdam/fall2019/ml1

If you have not found a lab partner, use Piazza to find one!

Overview

1. **Probability theory**

2. **Statistical learning principles:**

    I. Maximum likelihood

    II. Maximum a posteriori

    III. Bayesian prediction

Overview

1. **Probability theory**

2. Statistical learning principles**:**

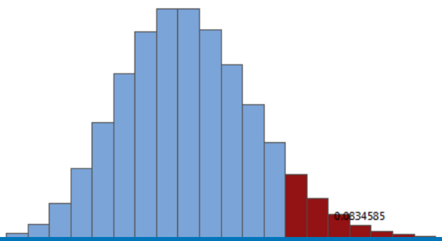
   I. Maximum likelihood

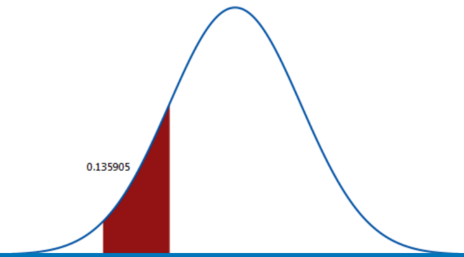
   II. Maximum a posteriori

   III. Bayesian prediction

The rules of probability theory

For random variables *X* ∊ $\mathcal{X}$ and *Y* ∊ $\mathcal{Y}$:

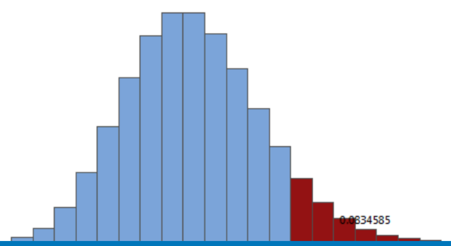


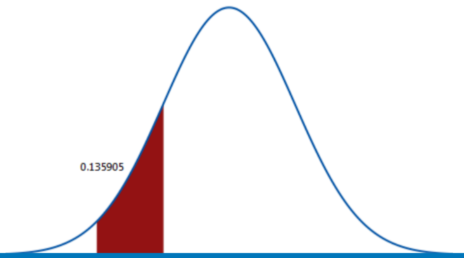


	Discrete	Continous
Additivity	$p(X \in A) = \sum_{x \in A} p(x)$	
Positivity	$p(x) \geq 0$	$p(x) \geq 0$
Normalization		$\int_{\mathcal{X}} p(x) dx = 1$
Sum Rule	$p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$	
Product Rule	$p(x, y) = p(x \mid y) p(y)$	$p(x, y) = p(x \mid y) p(y)$

The rules of probability theory

For random variables *X* ∊ $\mathcal{X}$ and *Y* ∊ $\mathcal{Y}$:





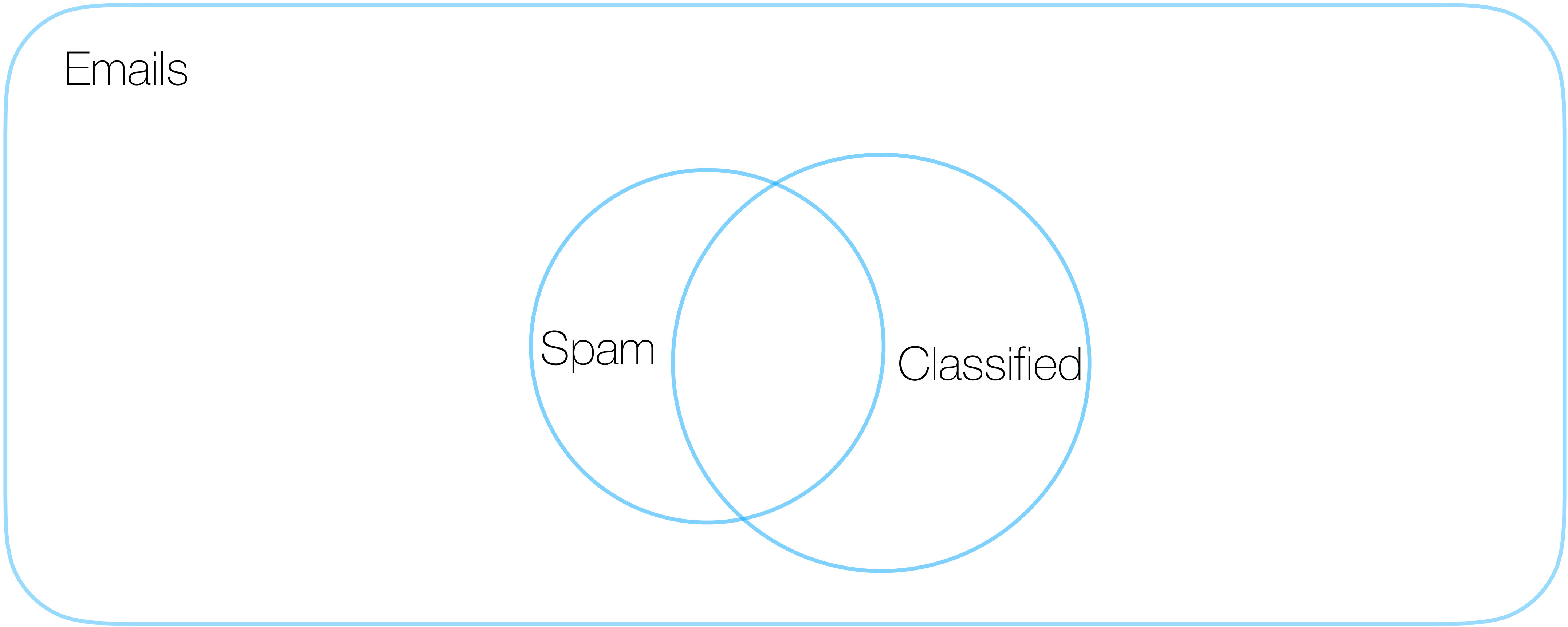
	Discrete	Continous
Additivity	$p(X \in A) = \sum_{x \in A} p(x)$	$p(X \in A) = \int_A p(x)$
Positivity	$p(x) \geq 0$	$p(x) \geq 0$
Normalization	$\sum_{x \in \mathcal{X}} p(x) = 1$	$\int_{\mathcal{X}} p(x)dx = 1$
Sum Rule	$p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$	$\int_{\mathcal{Y}} p(x, y)$
Product Rule	$p(x, y) = p(x\|y)p(y)$	$p(x, y) = p(x\|y)p(y)$

Bayes Rule

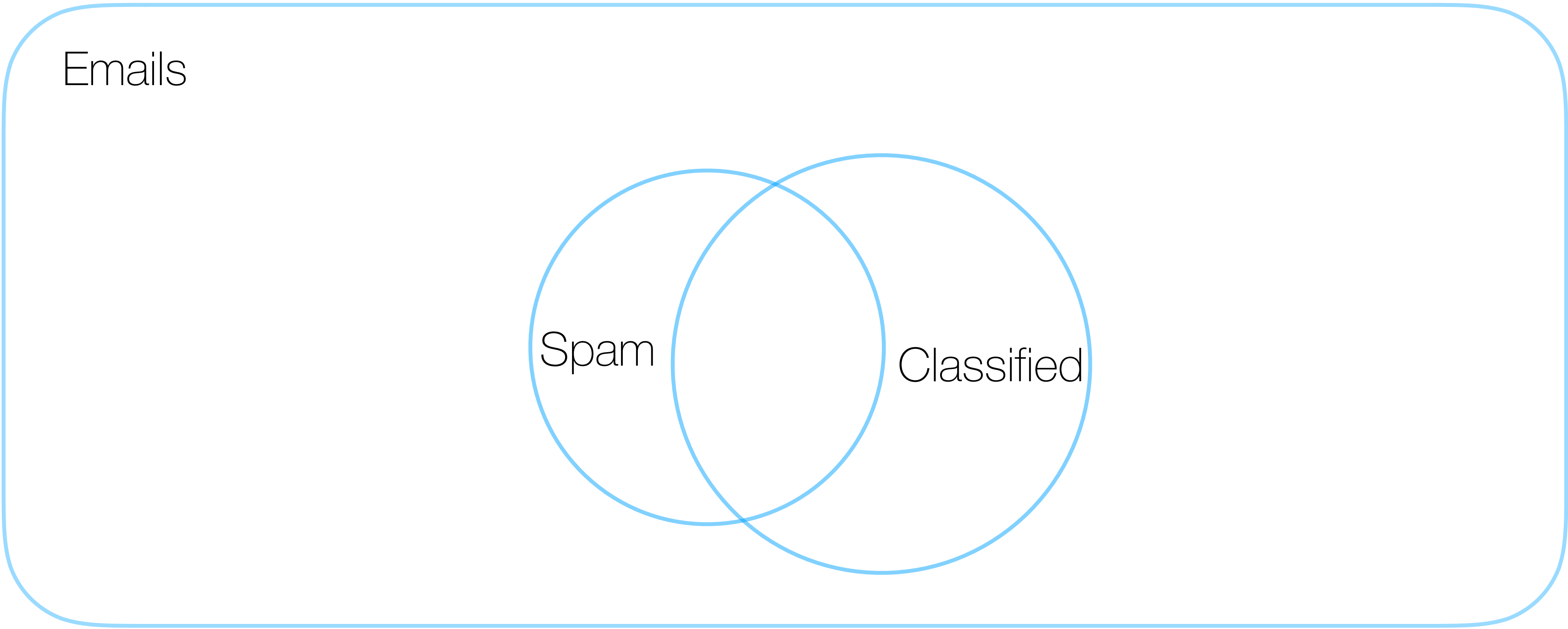
$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

- *p(y)* : the prior probability of *Y* = *y*
- *p(y | x)* : the posterior probability of *Y* = *y*
- *p(x | y)* : the likelihood of *X* = *x* given *Y* = *y*
- *p(x)* : the evidence for X = x

Bayes Rule

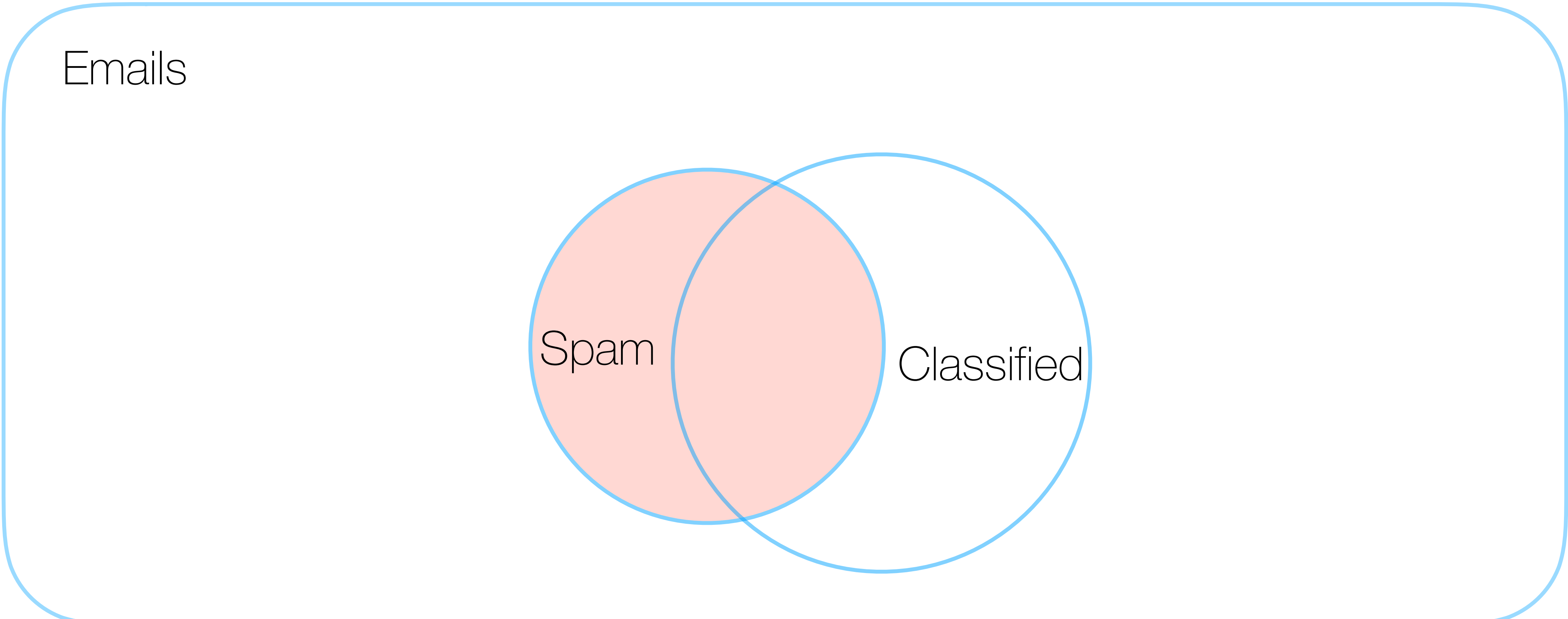


Bayes Rule



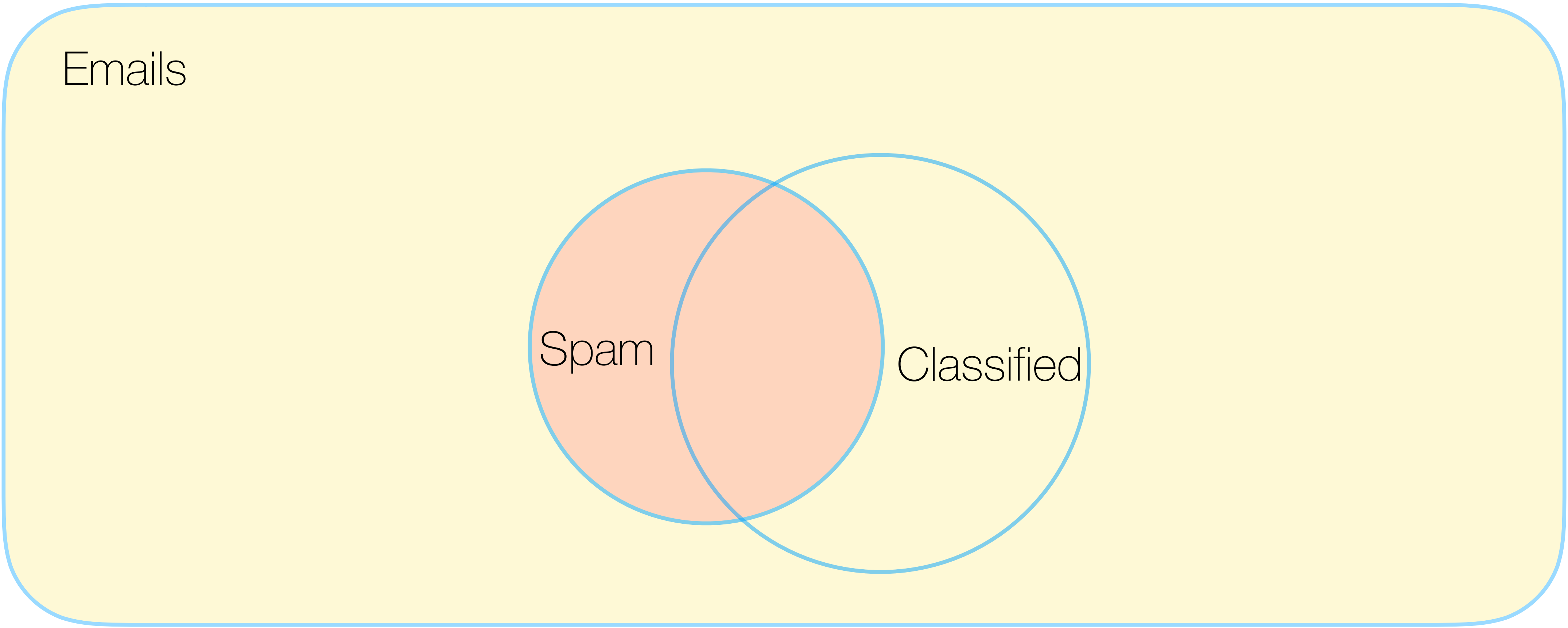
$$p(S) = \frac{S}{E}$$

Bayes Rule



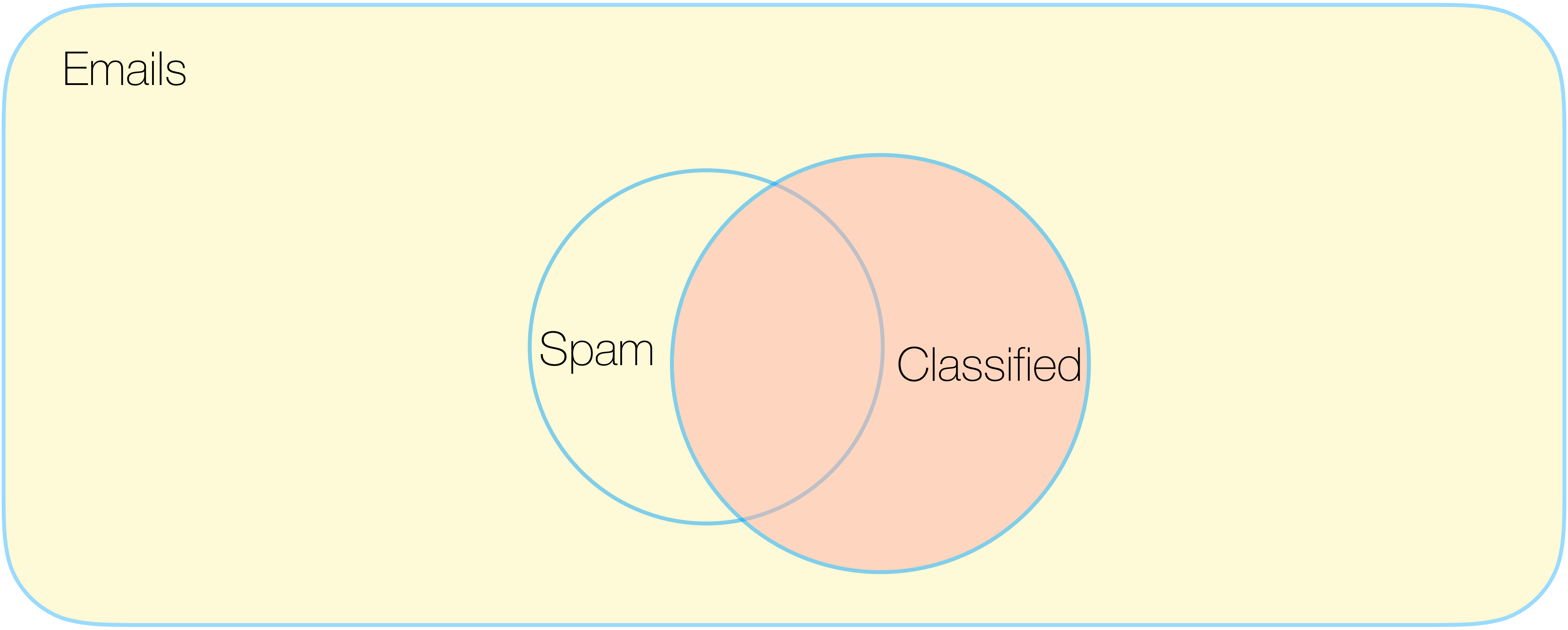
$$p(S) = \frac{S}{E}$$

Bayes Rule



$$p(S) = \frac{S}{E}$$

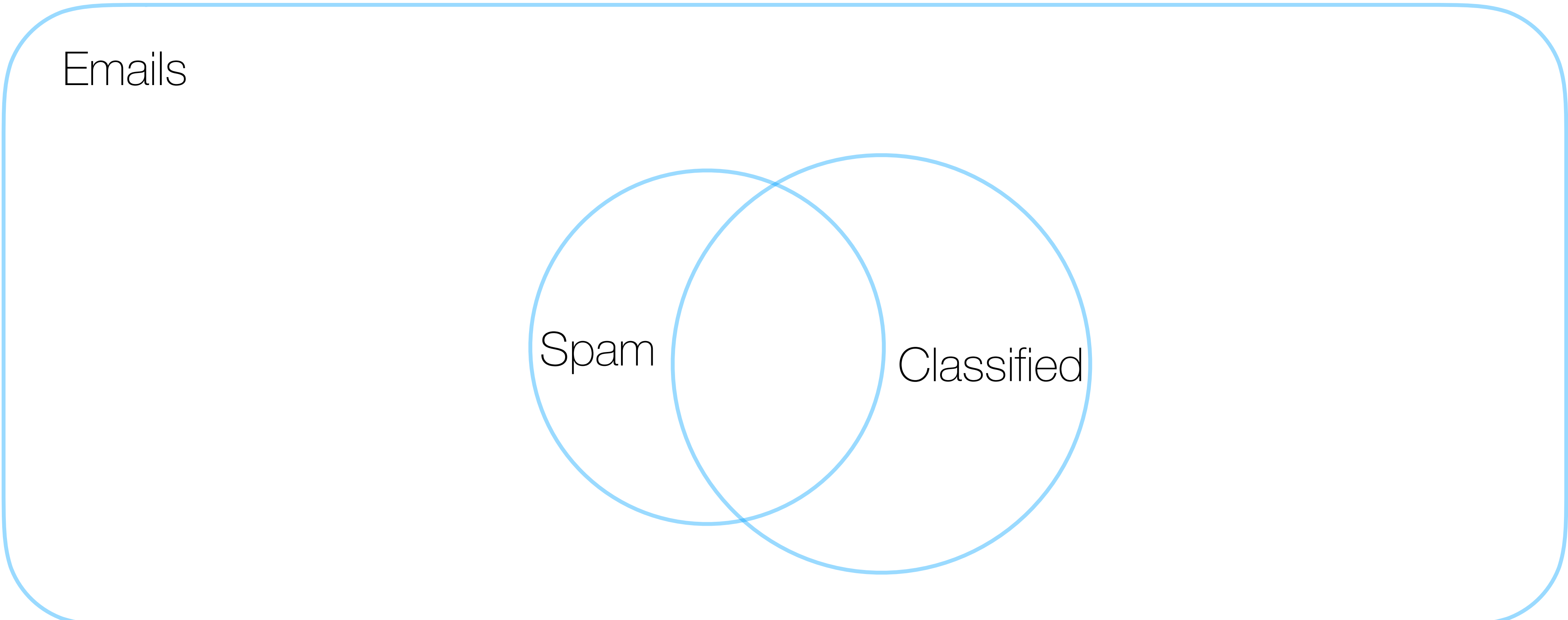
Bayes Rule



$$p(S) = \frac{S}{E}$$

$$p(C) = \frac{C}{E}$$

Bayes Rule

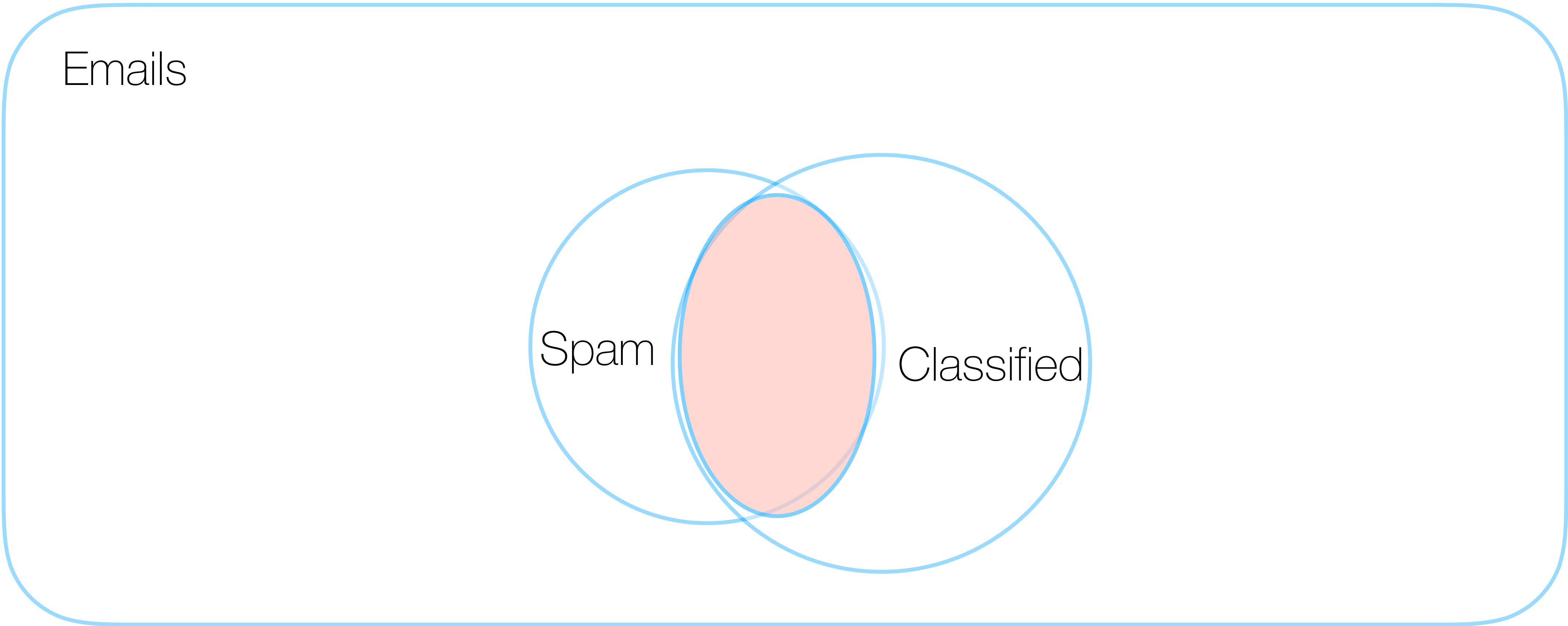


$$p(S) = \frac{S}{E}$$

$$p(S|C) = \frac{S \cap C}{C}$$

$$p(C) = \frac{C}{E}$$

Bayes Rule

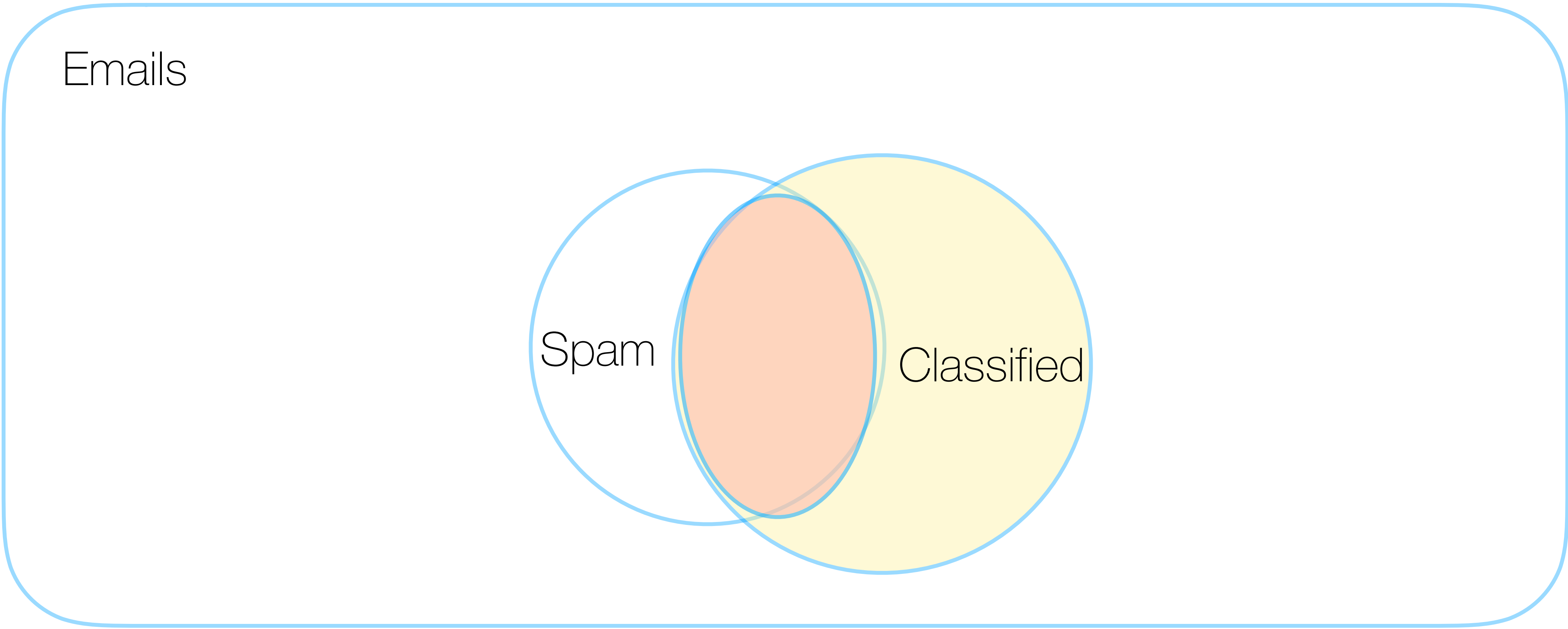


$$p(S) = \frac{S}{E}$$

$$p(S|C) = \frac{S \cap C}{C}$$

$$p(C) = \frac{C}{E}$$

Bayes Rule

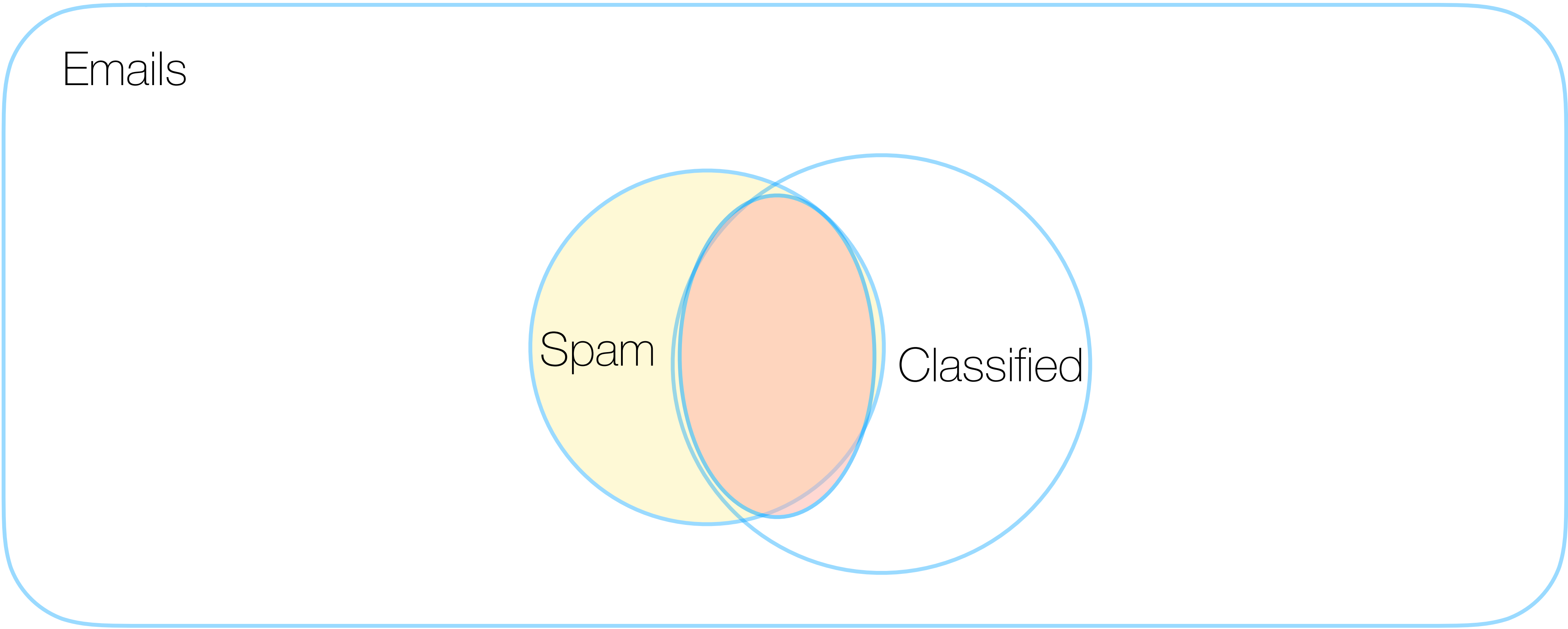


$$p(S) = \frac{S}{E}$$

$$p(S|C) = \frac{S \cap C}{C}$$

$$p(C) = \frac{C}{E}$$

Bayes Rule



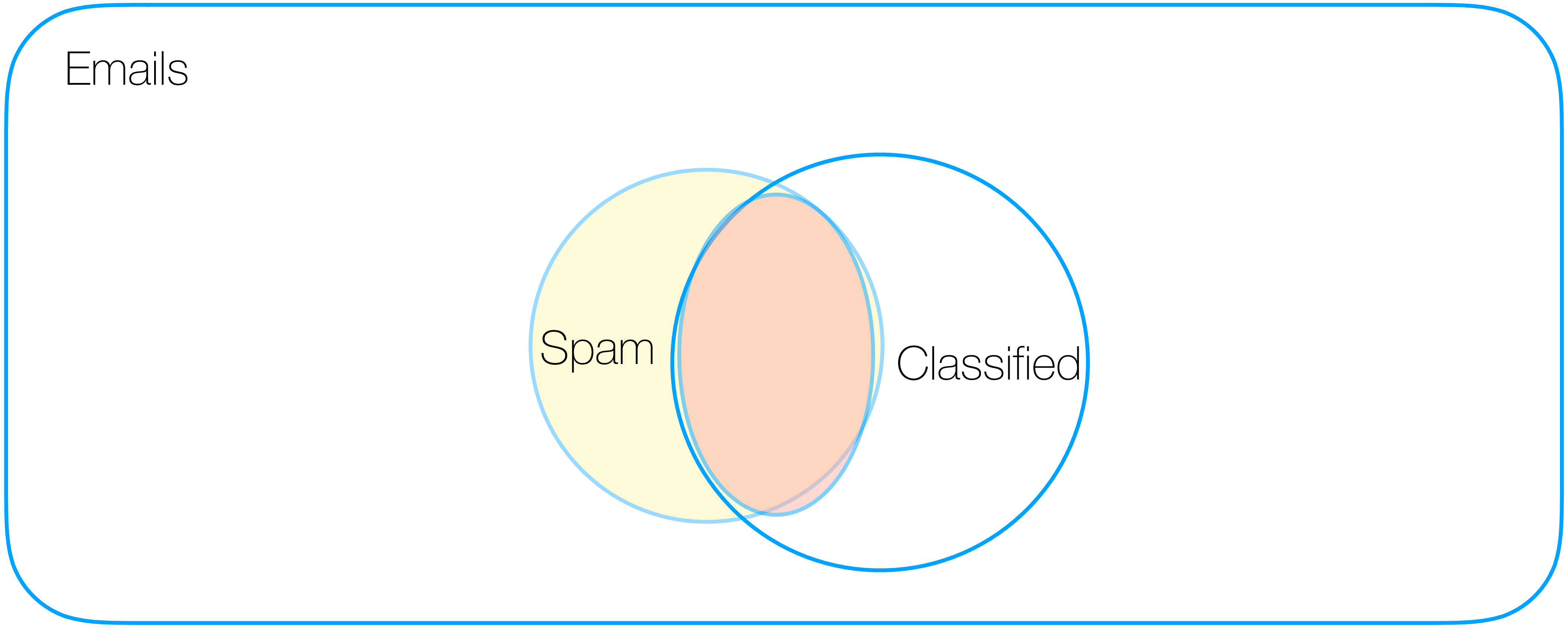
$$p(S) = \frac{S}{E}$$

$$p(S|C) = \frac{S \cap C}{C}$$

$$p(C) = \frac{C}{E}$$

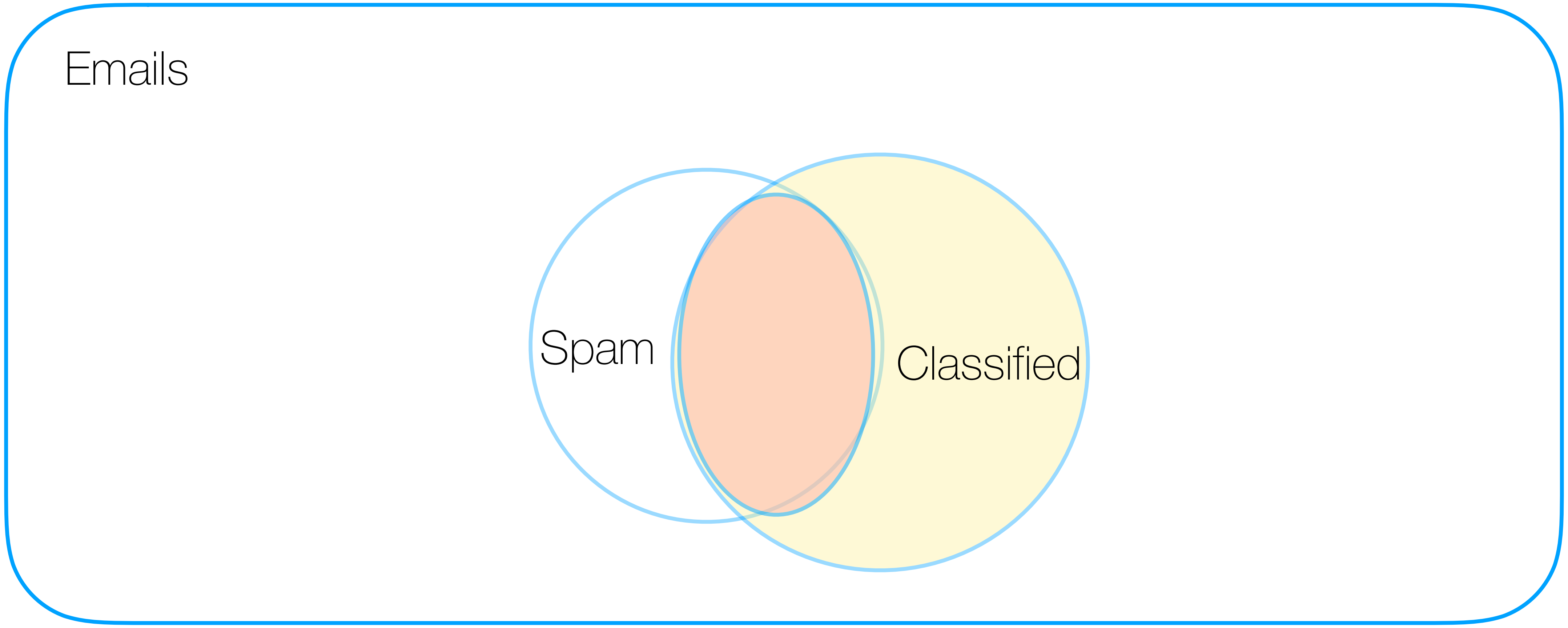
$$p(C|S) = \frac{S \cap C}{S}$$

Bayes Rule



$$p(C|S) = \frac{S \cap C}{S}$$

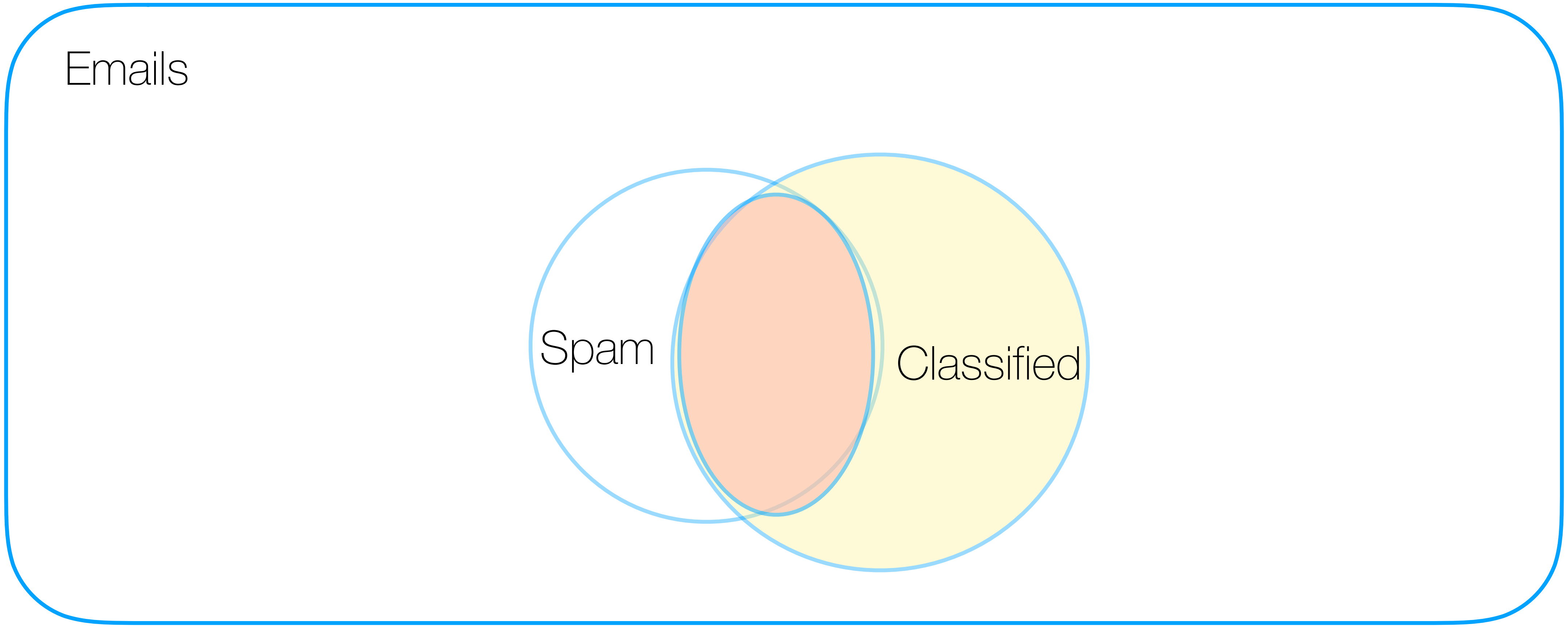
Bayes Rule



$$p(C|S) = \frac{S \cap C}{S}$$

$$p(S|C) = \frac{S \cap C}{S} \frac{S}{C}$$

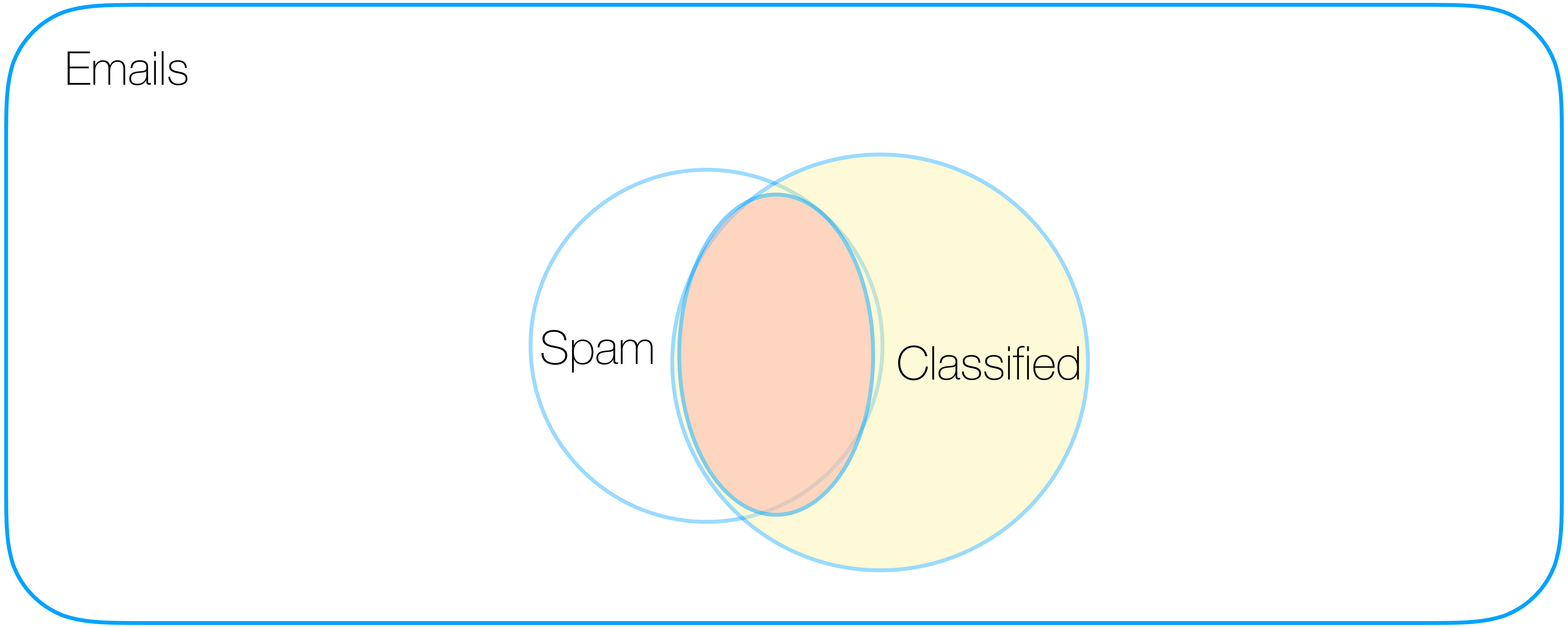
Bayes Rule



$$p(C|S) = \frac{S \cap C}{S}$$

$$p(S|C) = \frac{S \cap C}{S} \frac{S}{C}$$

Bayes Rule



$$p(C|S) = \frac{S \cap C}{S}$$

$$p(S|C) = \frac{S \cap C}{S}\frac{S}{C} = \frac{P(C|S)P(S)}{P(C)}$$

Independent Random Variables

Two random variables *X* and *Y* are *independent* iff measuring *X* gives no information on *Y*, and vice versa.

- Formally: *X* and *Y* are called independent if

  for all $x \in \mathcal{X},\ y \in \mathcal{Y}$

- Equivalent to

$$p(x|y) =$$

- Example:

Independent Random Variables

Two random variables *X* and *Y* are *independent* iff measuring *X* gives no information on *Y*, and vice versa.

- Formally: *X* and *Y* are called independent if

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

- Equivalent to

$$p(x|y) =$$

- Example:

Independent Random Variables

Two random variables *X* and *Y* are *independent* iff measuring *X* gives no information on *Y*, and vice versa.

- Formally: *X* and *Y* are called independent if

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

- Equivalent to

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

- Example:

Independent Random Variables

Two random variables *X* and *Y* are *independent* iff measuring *X* gives no information on *Y*, and vice versa.

- Formally: *X* and *Y* are called independent if

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

- Equivalent to

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

- Example:

Independent Random Variables

Two random variables *X* and *Y* are *independent* iff measuring *X* gives no information on *Y*, and vice versa.

- Formally: *X* and *Y* are called independent if

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

- Equivalent to

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

- Example:

Independent Random Variables

Two random variables *X* and *Y* are *independent* iff measuring *X* gives no information on *Y*, and vice versa.

- Formally: *X* and *Y* are called independent if

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

- Equivalent to

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

- Example:

Throwing two dices

Expectations

- random variable $X \in \mathcal{X}$ and function $f : \mathcal{X} \to \mathbb{R}$

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

- For *N* points drawn from p(X):

$$\mathbb{E}[f] =$$

- Conditional expectation:

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

Expectations

- random variable $X \in \mathcal{X}$ and function $f : \mathcal{X} \to \mathbb{R}$

$$\mathbb{E}[f] = \mathbb{E}_{x \sim p(X)}[f(x)] = \int p(x) f(x) dx$$

- For *N* points drawn from p(X):

$$\mathbb{E}[f] =$$

- Conditional expectation:

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

Expectations

- random variable $X \in \mathcal{X}$ and function $f : \mathcal{X} \to \mathbb{R}$

$$\mathbb{E}[f] = \mathbb{E}_{x \sim p(X)}[f(x)] = \int p(x) f(x) dx$$

- For *N* points drawn from p(X): $\{x_1, \ldots, x_N\}, x_n \sim p(x)$

$$\mathbb{E}[f] =$$

- Conditional expectation:

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

Expectations

- random variable $X \in \mathcal{X}$ and function $f : \mathcal{X} \to \mathbb{R}$

$$\mathbb{E}[f] = \mathbb{E}_{x \sim p(X)}[f(x)] = \int p(x) f(x) dx$$

- For N points drawn from p(X): $\{x_1, \ldots, x_N\}, x_n \sim p(x)$

$$\mathbb{E}[f] = \approx \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

- Conditional expectation:

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

Expectations

- random variable $X \in \mathcal{X}$ and function $f : \mathcal{X} \to \mathbb{R}$

$$\mathbb{E}[f] = \mathbb{E}_{x \sim p(X)}[f(x)] = \int p(x) f(x) dx$$

- For N points drawn from p(X): $\{x_1, \ldots, x_N\}, x_n \sim p(x)$

$$\mathbb{E}[f] = \approx \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

- Conditional expectation:

$$\mathbb{E}[f \,|\, y] = \mathbb{E}_{x \sim p(X|Y=y)}[f(x)] = \int p(x|y) f(x) dx$$

Variance

- The expected quadratic distance between f and its mean $\mathbb{E}[f]$

$$\mathrm{var}[f] = E_{x\sim p(x)}\left[(f(x) - E(f))^2\right]$$

Variance

- The expected quadratic distance between f and its mean $\mathbb{E}[f]$

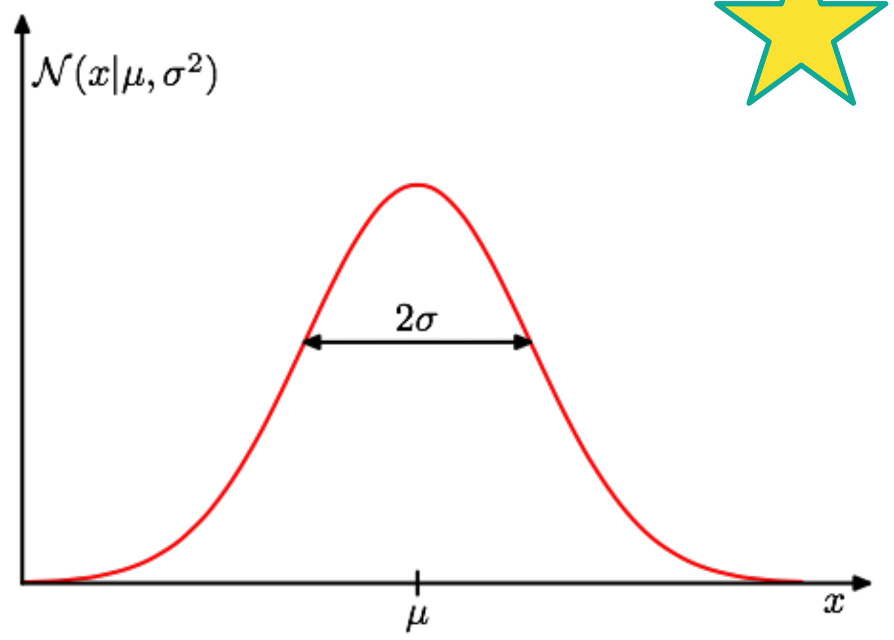
$$\begin{aligned}\operatorname{var}[f] &= E_{x\sim p(x)}\left[(f(x)-E(f))^2\right] \\ &= E\left[f(x)^2 - 2f(x)E[f(x)] + E[f]^2\right]\end{aligned}$$

Variance

- The expected quadratic distance between f and its mean $\mathbb{E}[f]$

$$
\begin{aligned}
\operatorname{var}[f] &= E_{x\sim p(x)}\left[(f(x) - E(f))^2\right] \\
&= E\left[f(x)^2 - 2f(x)E[f(x)] + E[f]^2\right] \\
&= E\left[f(x)^2\right] - 2E[f(x)]E[f(x)] + E[f]^2 \\
&= E\left[f(x)^2\right] - E[f(x)]^2
\end{aligned}
$$

Gaussian Distribution



- Real valued stochastic variable *X*

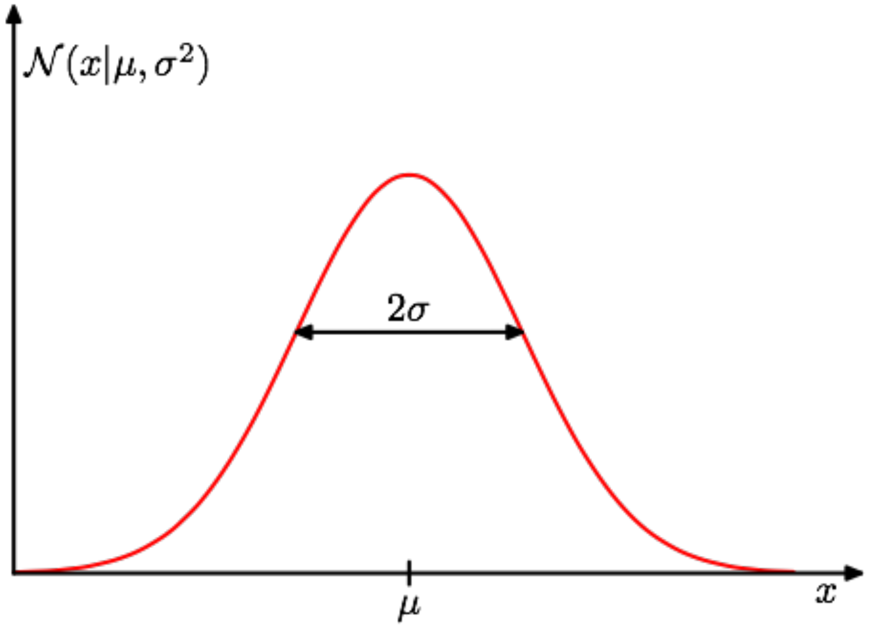
$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

- Mean: $\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2) x \, \mathrm{d}x = ?$

- Variance: $\mathrm{var}[x] = \mathbb{E}[(x - \mathbb{E}[x])^2] = ?$

Gaussian Distribution

- Real valued stochastic variable *X*



$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

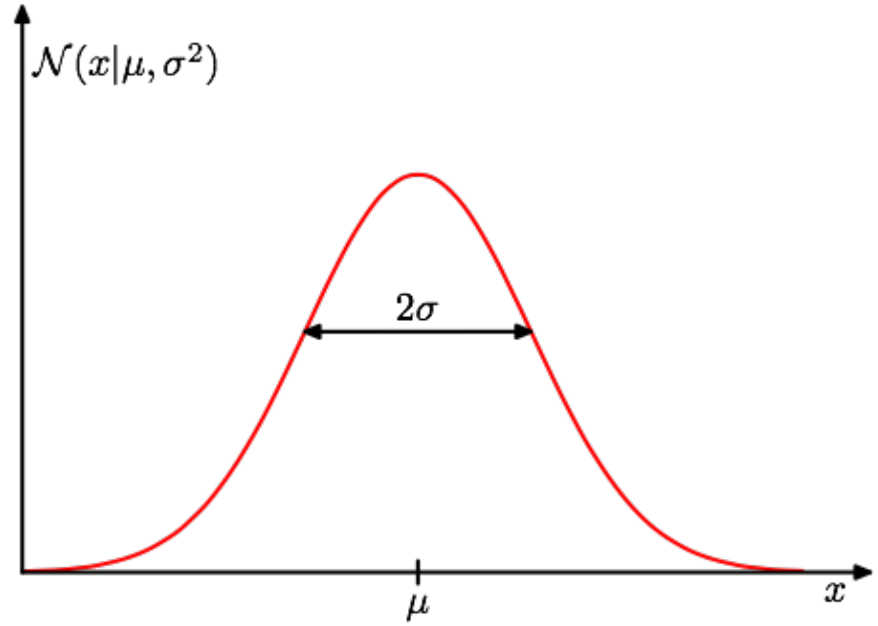
- Mean: $\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2) x \, \mathrm{d}x =$

$$\boxed{\int_{-\infty}^{\infty} y e^{-y^2} \mathrm{d}y = 0}$$

$$\boxed{\int_{-\infty}^{\infty} e^{-y^2} \mathrm{d}y = \sqrt{\pi}}$$

Gaussian Distribution

- Real valued stochastic variable *X*



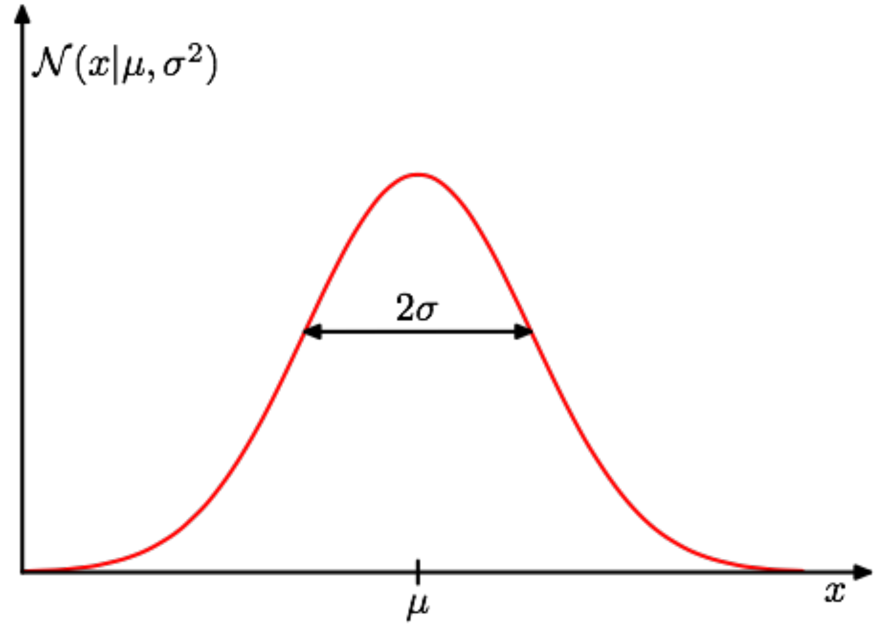
$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

- Mean: $\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)x \,\mathrm{d}x = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$

$$\boxed{\int_{-\infty}^{\infty} y e^{-y^2} \mathrm{d}y = 0} \quad \boxed{\int_{-\infty}^{\infty} e^{-y^2} \mathrm{d}y = \sqrt{\pi}}$$

Gaussian Distribution

- Real valued stochastic variable *X*



$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

- Mean: $\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2) x \, \mathrm{d}x = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$

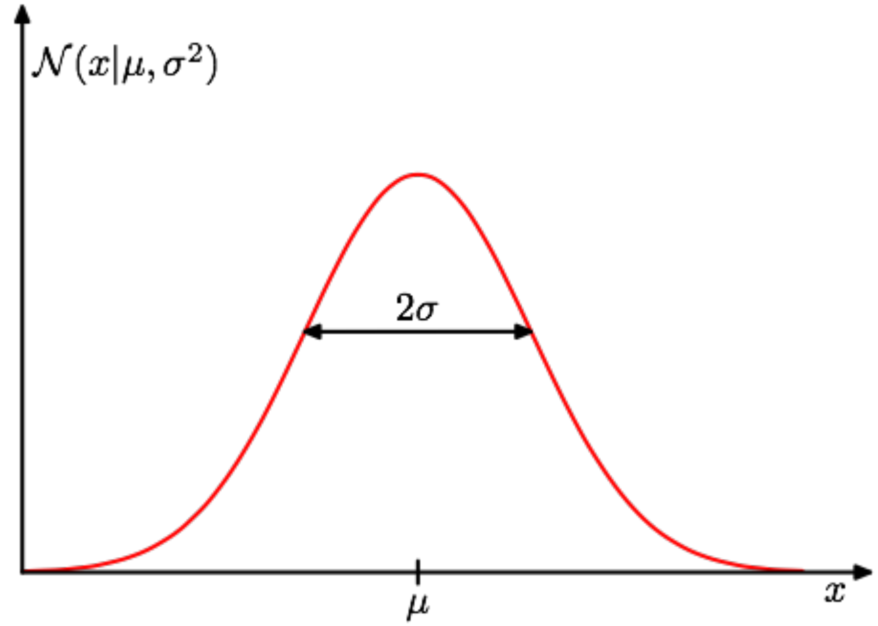
$$y = \frac{1}{\sqrt{2\sigma^2}}(x-\mu)$$

$$\boxed{\int_{-\infty}^{\infty} y e^{-y^2} \mathrm{d}y = 0} \quad \boxed{\int_{-\infty}^{\infty} e^{-y^2} \mathrm{d}y = \sqrt{\pi}}$$

$$dx = \sqrt{2\sigma^2} dy$$

Gaussian Distribution

- Real valued stochastic variable *X*



$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

- Mean: $\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)x \, \mathrm{d}x = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$

$$y = \frac{1}{\sqrt{2\sigma^2}}(x-\mu)$$

$$\downarrow$$

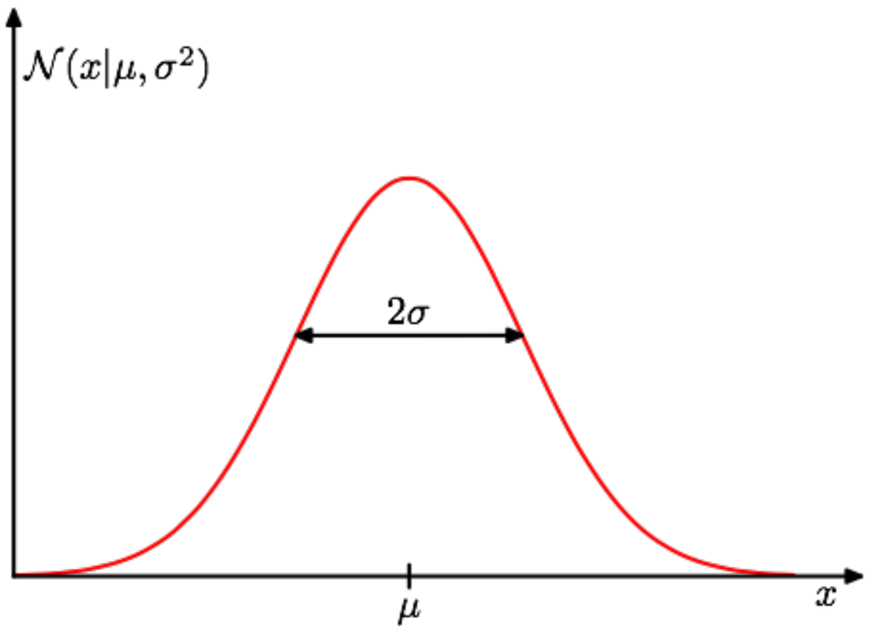
$$x = \mu + y\sqrt{2\sigma^2}$$

$$dx = \sqrt{2\sigma^2}dy$$

$$\boxed{\int_{-\infty}^{\infty} ye^{-y^2} \mathrm{d}y = 0} \quad \boxed{\int_{-\infty}^{\infty} e^{-y^2} \mathrm{d}y = \sqrt{\pi}}$$

Gaussian Distribution

- Real valued stochastic variable *X*



$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

- Mean: $\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2) x \,\mathrm{d}x = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$

$$= \int_{-\infty}^{\infty} \frac{\mu + y\sqrt{2\sigma^2}}{\sqrt{2\pi\sigma^2}} e^{-y^2} \sqrt{2\sigma^2} dy$$

$$y = \frac{1}{\sqrt{2\sigma^2}}(x-\mu)$$

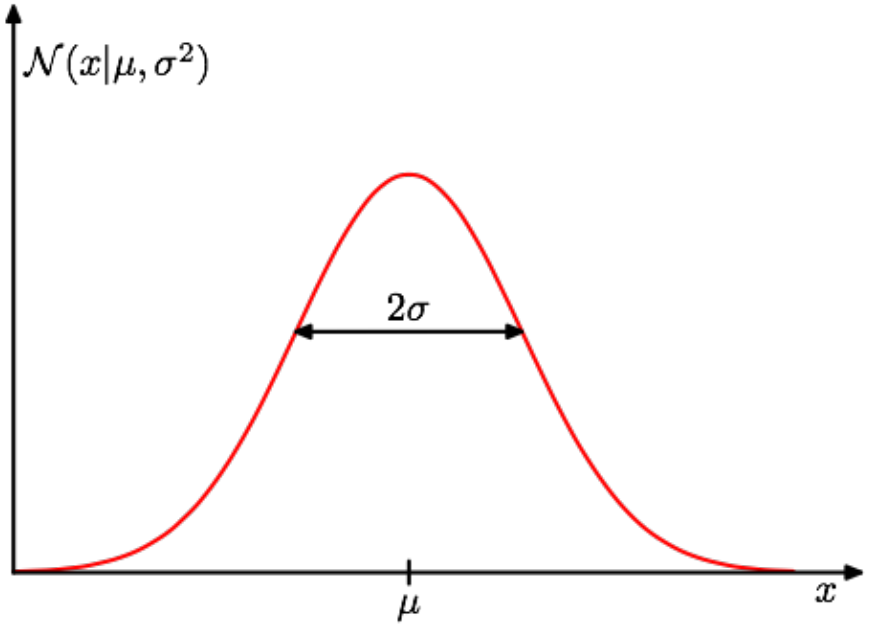
$$\downarrow$$

$$x = \mu + y\sqrt{2\sigma^2}$$

$$dx = \sqrt{2\sigma^2} dy$$

$$\boxed{\int_{-\infty}^{\infty} y e^{-y^2} \mathrm{d}y = 0} \quad \boxed{\int_{-\infty}^{\infty} e^{-y^2} \mathrm{d}y = \sqrt{\pi}}$$

Gaussian Distribution



- Real valued stochastic variable *X*

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

- Mean: $\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)x \,\mathrm{d}x = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$

$$= \int_{-\infty}^{\infty} \frac{\mu + y\sqrt{2\sigma^2}}{\sqrt{2\pi\sigma^2}} e^{-y^2}\sqrt{2\sigma^2}dy \;= \frac{\mu\sqrt{2\sigma^2}}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{-y^2}dy + \frac{2\sigma^2}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} ye^{-y^2}$$

$$y = \frac{1}{\sqrt{2\sigma^2}}(x-\mu)$$

$$\downarrow$$

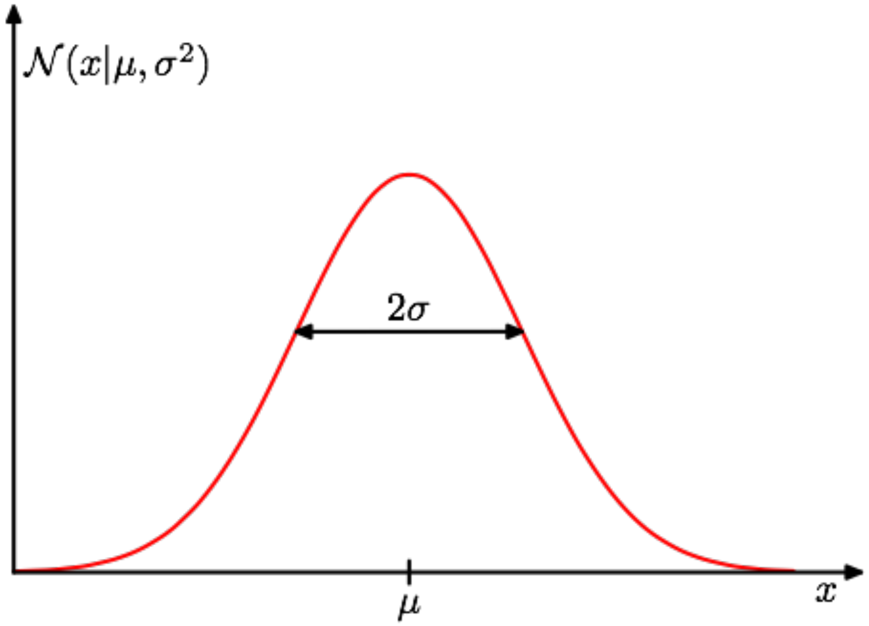
$$x = \mu + y\sqrt{2\sigma^2}$$

$$dx = \sqrt{2\sigma^2}dy$$

$$\boxed{\int_{-\infty}^{\infty} ye^{-y^2}\mathrm{d}y = 0} \quad \boxed{\int_{-\infty}^{\infty} e^{-y^2}\mathrm{d}y = \sqrt{\pi}}$$

Gaussian Distribution

- Real valued stochastic variable *X*



$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

- Mean: $\mathbb{E}[x] = \int_{-\infty}^{\infty}\mathcal{N}(x|\mu,\sigma^2)x\,\mathrm{d}x = \int_{-\infty}^{\infty}\frac{x}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}dx$

$$= \int_{-\infty}^{\infty}\frac{\mu + y\sqrt{2\sigma^2}}{\sqrt{2\pi\sigma^2}}e^{-y^2}\sqrt{2\sigma^2}dy = \frac{\mu\sqrt{2\sigma^2}}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}e^{-y^2}dy + \frac{2\sigma^2}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}ye^{-y^2}$$

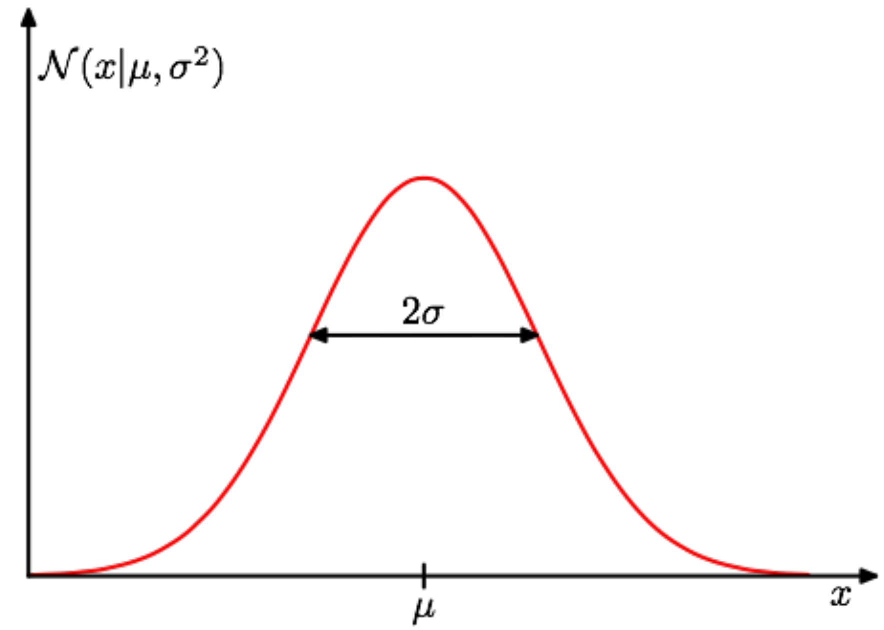
$$y = \frac{1}{\sqrt{2\sigma^2}}(x-\mu)$$

$$x = \mu + y\sqrt{2\sigma^2}$$

$$dx = \sqrt{2\sigma^2}dy$$

$$\boxed{\int_{-\infty}^{\infty}ye^{-y^2}\mathrm{d}y = 0} \qquad \boxed{\int_{-\infty}^{\infty}e^{-y^2}\mathrm{d}y = \sqrt{\pi}}$$

Gaussian Distribution



- Real valued stochastic variable *X*

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

- Mean: $\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)x \, \mathrm{d}x = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$

$$= \int_{-\infty}^{\infty} \frac{\mu + y\sqrt{2\sigma^2}}{\sqrt{2\pi\sigma^2}} e^{-y^2}\sqrt{2\sigma^2}dy = \frac{\mu\sqrt{2\sigma^2}}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{-y^2}dy + \frac{2\sigma^2}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} ye^{-y^2}$$

$$= \frac{\mu\sqrt{2\sigma^2}}{\sqrt{2\pi\sigma^2}}\sqrt{\pi} + 0 = \mu$$

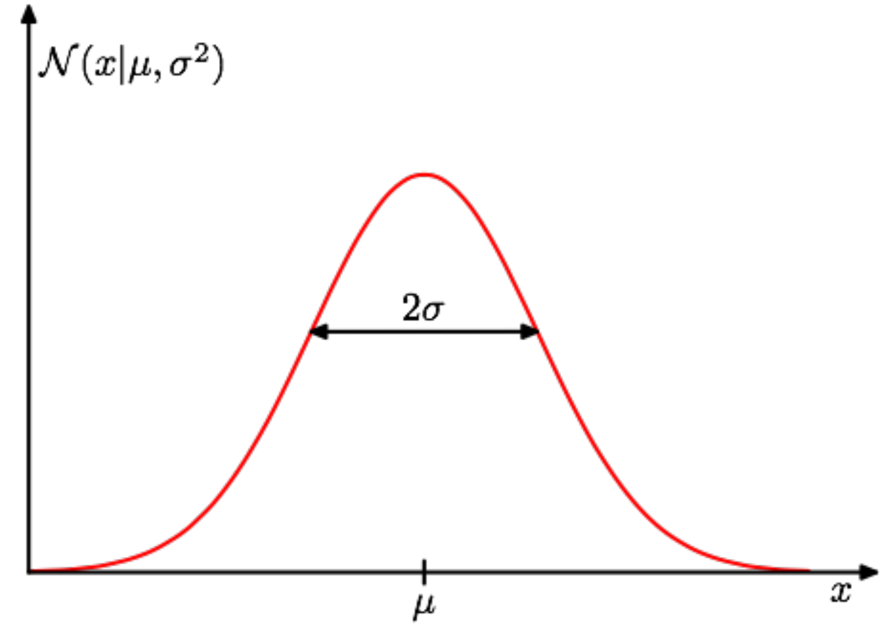
$$y = \frac{1}{\sqrt{2\sigma^2}}(x-\mu)$$

$$x = \mu + y\sqrt{2\sigma^2}$$

$$dx = \sqrt{2\sigma^2}dy$$

$$\boxed{\int_{-\infty}^{\infty} ye^{-y^2}\mathrm{d}y = 0} \qquad \boxed{\int_{-\infty}^{\infty} e^{-y^2}\mathrm{d}y = \sqrt{\pi}}$$

Gaussian Distribution



- Real valued stochastic variable *X*

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

- Mean: $\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2) x \, \mathrm{d}x = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$

$$= \int_{-\infty}^{\infty} \frac{\mu + y\sqrt{2\sigma^2}}{\sqrt{2\pi\sigma^2}} e^{-y^2} \sqrt{2\sigma^2} dy = \frac{\mu\sqrt{2\sigma^2}}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-y^2} dy + \frac{2\sigma^2}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} y e^{-y^2}$$

$$= \frac{\mu\sqrt{2\sigma^2}}{\sqrt{2\pi\sigma^2}} \sqrt{\pi} + 0 = \mu$$

$$y = \frac{1}{\sqrt{2\sigma^2}}(x-\mu)$$

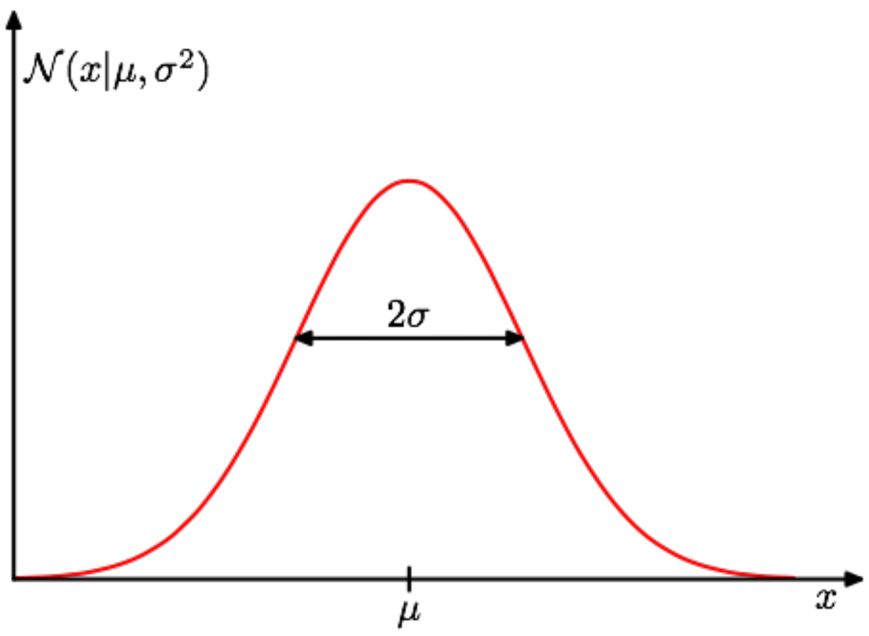
$$x = \mu + y\sqrt{2\sigma^2}$$

$$dx = \sqrt{2\sigma^2} dy$$

$$\boxed{\int_{-\infty}^{\infty} y e^{-y^2} \mathrm{d}y = 0} \qquad \boxed{\int_{-\infty}^{\infty} e^{-y^2} \mathrm{d}y = \sqrt{\pi}}$$

Gaussian Distribution

- Real valued stochastic variable *X*



$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

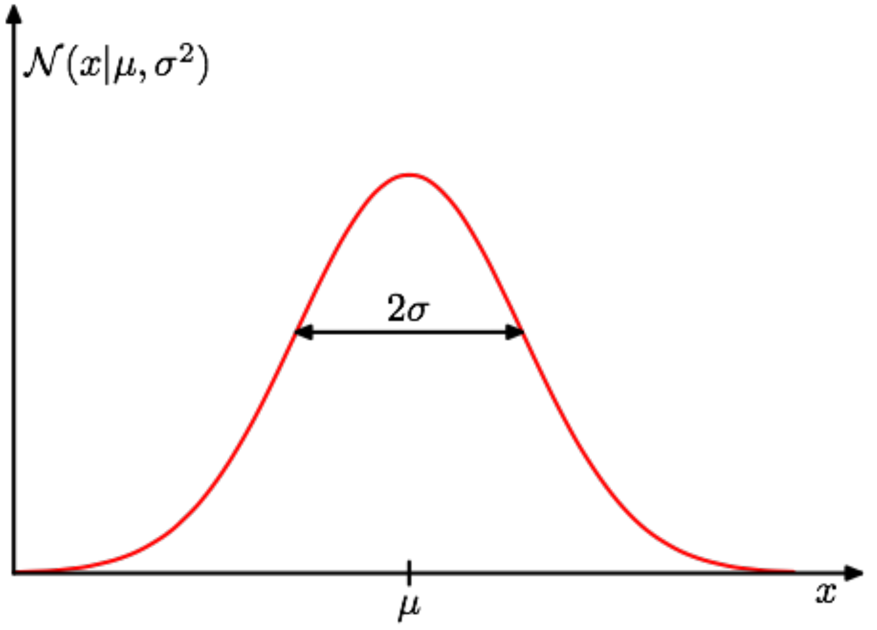
- Variance: $\mathrm{var}[x] = \mathbb{E}[(x-\mathbb{E}[x])^2]$

$$\mathrm{var}[x] =$$

$$\boxed{\int_{-\infty}^{\infty} e^{-x^2} \mathrm{d}x = \sqrt{\pi}}$$

$$\boxed{\int_{-\infty}^{\infty} x^2 e^{-ax^2} \mathrm{d}x = -\frac{\partial}{\partial a}\int_{-\infty}^{\infty} e^{-ax^2}\mathrm{d}x = -\frac{\partial}{\partial a}\sqrt{\frac{\pi}{a}} = \frac{1}{2}\sqrt{\frac{\pi}{a^3}}}$$

Gaussian Distribution



- Real valued stochastic variable *X*

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

- Variance: $\mathrm{var}[x] = \mathbb{E}[(x-\mathbb{E}[x])^2] = \int_{-\infty}^{\infty} (x-\mu)^2 \mathcal{N}(x|\mu,\sigma^2)\, x \mathrm{d}x$

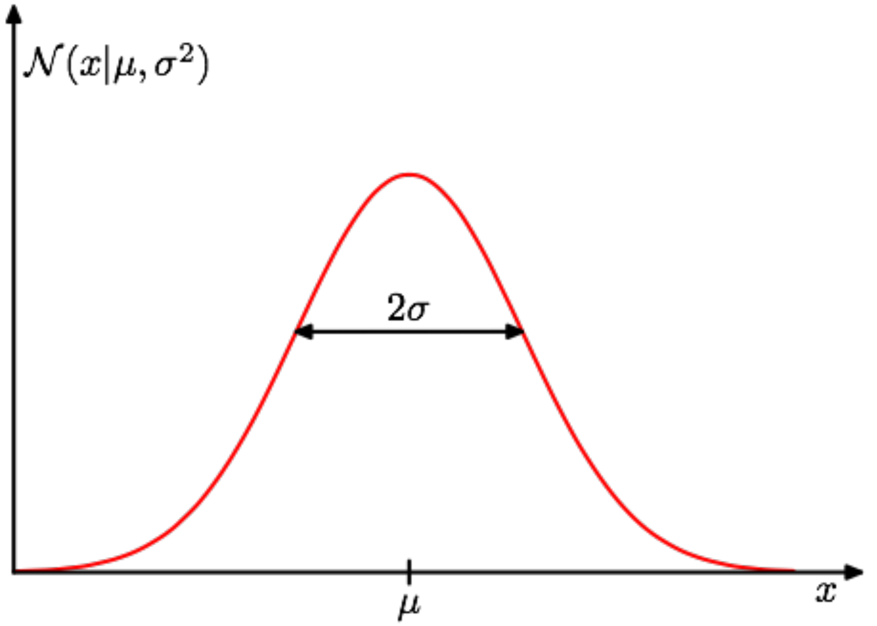
$$\mathrm{var}[x] =$$

$$\boxed{\int_{-\infty}^{\infty} e^{-x^2} \mathrm{d}x = \sqrt{\pi}}$$

$$\boxed{\int_{-\infty}^{\infty} x^2 e^{-ax^2} \mathrm{d}x = -\frac{\partial}{\partial a}\int_{-\infty}^{\infty} e^{-ax^2}\mathrm{d}x = -\frac{\partial}{\partial a}\sqrt{\frac{\pi}{a}} = \frac{1}{2}\sqrt{\frac{\pi}{a^3}}}$$

Gaussian Distribution

- Real valued stochastic variable *X*



$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

- Variance: $\mathrm{var}[x] = \mathbb{E}[(x-\mathbb{E}[x])^2] = \int_{-\infty}^{\infty} (x-\mu)^2 \mathcal{N}\left(x|\mu,\sigma^2\right) x\mathrm{d}x$

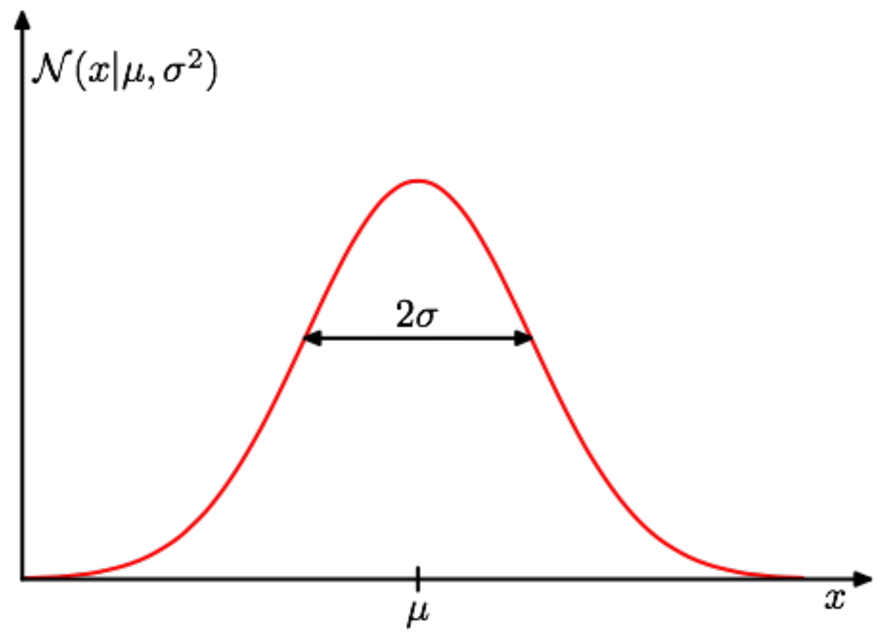
$$\mathrm{var}[x] =$$

$$y = \frac{1}{\sqrt{2\sigma^2}}(x-\mu), dx = \sqrt{2\sigma^2}dy$$

$$\boxed{\int_{-\infty}^{\infty} e^{-x^2}\mathrm{d}x = \sqrt{\pi}}$$

$$\boxed{\int_{-\infty}^{\infty} x^2 e^{-ax^2}\mathrm{d}x = -\frac{\partial}{\partial a}\int_{-\infty}^{\infty} e^{-ax^2}\mathrm{d}x = -\frac{\partial}{\partial a}\sqrt{\frac{\pi}{a}} = \frac{1}{2}\sqrt{\frac{\pi}{a^3}}}$$

Gaussian Distribution



- Real valued stochastic variable *X*

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

- Variance: $\mathrm{var}[x] = \mathbb{E}[(x-\mathbb{E}[x])^2] = \int_{-\infty}^{\infty}(x-\mu)^2\mathcal{N}(x|\mu,\sigma^2)\,x\mathrm{d}x$

$$\mathrm{var}[x] = \sigma^2$$

$$y = \frac{1}{\sqrt{2\sigma^2}}(x-\mu),\, dx = \sqrt{2\sigma^2}dy$$

$$\boxed{\int_{-\infty}^{\infty} e^{-x^2}\mathrm{d}x = \sqrt{\pi}}$$

$$\boxed{\int_{-\infty}^{\infty} x^2 e^{-ax^2}\mathrm{d}x = -\frac{\partial}{\partial a}\int_{-\infty}^{\infty} e^{-ax^2}\mathrm{d}x = -\frac{\partial}{\partial a}\sqrt{\frac{\pi}{a}} = \frac{1}{2}\sqrt{\frac{\pi}{a^3}}}$$

Multivariate Gaussian Distribution

- *D*-dimensional vector $\mathbf{x} = (x_1, x_2, ..., x_D)^T$
- $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) =$

$|\boldsymbol{\Sigma}| = \det \boldsymbol{\Sigma}$

- $\boldsymbol{\Sigma} =$

(D x D matrix)

- $\mathbb{E}[\mathbf{x}] =$

$$\boxed{\int \exp\{-\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}\}\mathrm{d}^D x = \frac{(2\pi)^{D/2}}{|\mathbf{A}|^{1/2}}}$$

Multivariate Gaussian Distribution

- *D*-dimensional vector $\mathbf{x} = (x_1, x_2, ..., x_D)^T$
- $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d^2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$

$$|\boldsymbol{\Sigma}| = \det \boldsymbol{\Sigma}$$

- $\boldsymbol{\Sigma} =$ (D x D matrix)
- $\mathbb{E}[\mathbf{x}] =$

$$\boxed{\int \exp\{-\frac{1}{2}\mathbf{x}^T \mathbf{A} \mathbf{x}\} \mathrm{d}^D x = \frac{(2\pi)^{D/2}}{|\mathbf{A}|^{1/2}}}$$

Multivariate Gaussian Distribution

- *D*-dimensional vector $\mathbf{x} = (x_1, x_2, ..., x_D)^T$
- $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d^2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right]$

$|\boldsymbol{\Sigma}| = \det \boldsymbol{\Sigma}$

- $\boldsymbol{\Sigma} = \mathrm{cov}[\mathbf{x}]$

(D x D matrix)

- $\mathbb{E}[\mathbf{x}] =$

$$\boxed{\int \exp\{-\frac{1}{2}\mathbf{x}^T \mathbf{A} \mathbf{x}\} \mathrm{d}^D x = \frac{(2\pi)^{D/2}}{|\mathbf{A}|^{1/2}}}$$

Multivariate Gaussian Distribution

- *D*-dimensional vector $\mathbf{x} = (x_1, x_2, ..., x_D)^T$
- $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d^2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right]$

$|\boldsymbol{\Sigma}| = \det \boldsymbol{\Sigma}$

- $\boldsymbol{\Sigma} = \mathrm{cov}[\mathbf{x}]$

(D x D matrix)

- $\mathbb{E}[\mathbf{x}] = E[\mathbf{x}] = \int \mathbf{x} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) d\mathbf{x} = \boldsymbol{\mu}$

$$\boxed{\int \exp\{-\frac{1}{2}\mathbf{x}^T \mathbf{A} \mathbf{x}\} \mathrm{d}^D x = \frac{(2\pi)^{D/2}}{|\mathbf{A}|^{1/2}}}$$

Exercises with Gaussians

- Compute:
- Mean of uni-/multivariate Gaussian (-> mu)
- Variance of univariate Gaussian (-> sigma)
- Covariance matrix of multivariate Gaussian (-> Sigma)
- Integral: $\int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi}$
- Multivariate Gaussian integral, normalisation factor.

Exercises with Gaussians

- Compute:
- Mean of uni-/multivariate Gaussian (-> mu)
- Variance of univariate Gaussian (-> sigma)
- Covariance matrix of multivariate Gaussian (-> Sigma)
- Integral: $\int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi}$
- Multivariate Gaussian integral, normalisation factor.

Overview

1. Probability theory

2. **Statistical learning principles:**

    I. **Maximum likelihood**

    II. Maximum a posteriori

    III. Bayesian prediction

Maximum Likelihood Estimation

- Dataset $D = (x_1, x_2, \ldots, x_N)$ of N independent observations.
- Likelihood of the dataset: $p(D|\mathbf{w})$

Maximum Likelihood Estimation

- Dataset $D = (x_1, x_2, \ldots, x_N)$ of N independent observations.
- Likelihood of the dataset: $p(D|\mathbf{w})$
- Maximum likelihood estimation: the most likely "explanation" of D is given by $\boldsymbol{w}_{\mathrm{ML}}$ which maximizes the likelihood function

$$\mathbf{w}_{\mathrm{ML}} = \arg\max_{\mathbf{w}} p(D|\mathbf{w})$$

Maximum Likelihood Estimation

- Dataset $D = (x_1, x_2, \ldots, x_N)$ of N independent observations.

- Likelihood of the dataset: $p(D|\mathbf{w})$

- Maximum likelihood estimation: the most likely "explanation" of D is given by $\mathbf{w}_{\mathrm{ML}}$ which maximizes the likelihood function

$$\mathbf{w}_{\mathrm{ML}} = \arg\max_{\mathbf{w}} p(D|\mathbf{w})$$

- i.i.d. assumption: each $x_i \in D$ is independently distributed according to the same distribution, conditioned on **w**.

$$x \sim p(x|\mathbf{w})$$

- If i.i.d., joint distribution

$$p(D|\mathbf{w}) = p(x_1, x_2, ..., x_N|\mathbf{w}) =$$

Maximum Likelihood Estimation

- Dataset $D = (x_1, x_2, \ldots, x_N)$ of N independent observations.
- Likelihood of the dataset: $p(D|\mathbf{w})$
- Maximum likelihood estimation: the most likely "explanation" of D is given by $\mathbf{w}_{ML}$ which maximizes the likelihood function

$$\mathbf{w}_{\mathrm{ML}} = \arg\max_{\mathbf{w}} p(D|\mathbf{w})$$

- i.i.d. assumption: each $x_i \in D$ is independently distributed according to the same distribution, conditioned on **w**.

$$x \sim p(x|\mathbf{w})$$

- If i.i.d., joint distribution

$$p(D|\mathbf{w}) = p(x_1, x_2, ..., x_N|\mathbf{w}) =$$

Maximum Likelihood Estimation

- Dataset $D = (x_1, x_2, \ldots, x_N)$ of N independent observations.
- Likelihood of the dataset: $p(D|\mathbf{w})$
- Maximum likelihood estimation: the most likely "explanation" of D is given by $\mathbf{w}_{ML}$ which maximizes the likelihood function

$$\mathbf{w}_{\mathrm{ML}} = \arg\max_{\mathbf{w}} p(D|\mathbf{w})$$

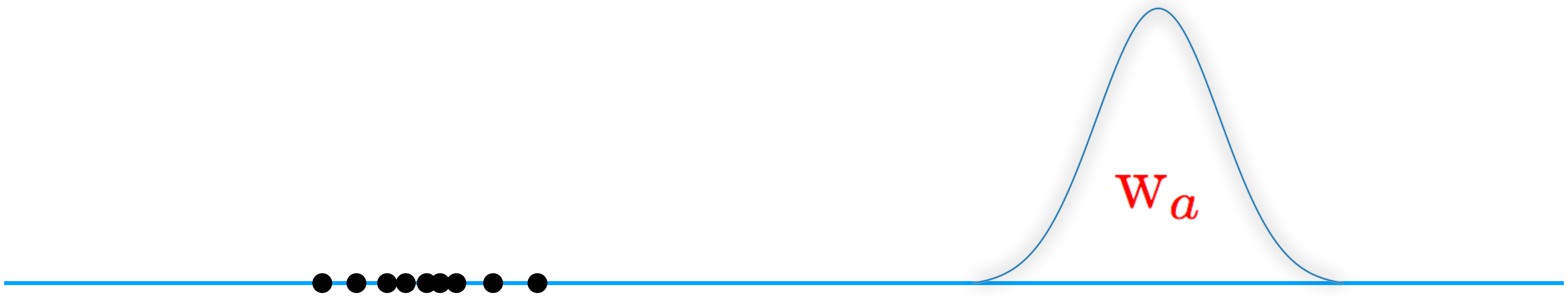
- i.i.d. assumption: each $x_i \in D$ is independently distributed according to the same distribution, conditioned on **w**.

$$x \sim p(x|\mathbf{w})$$

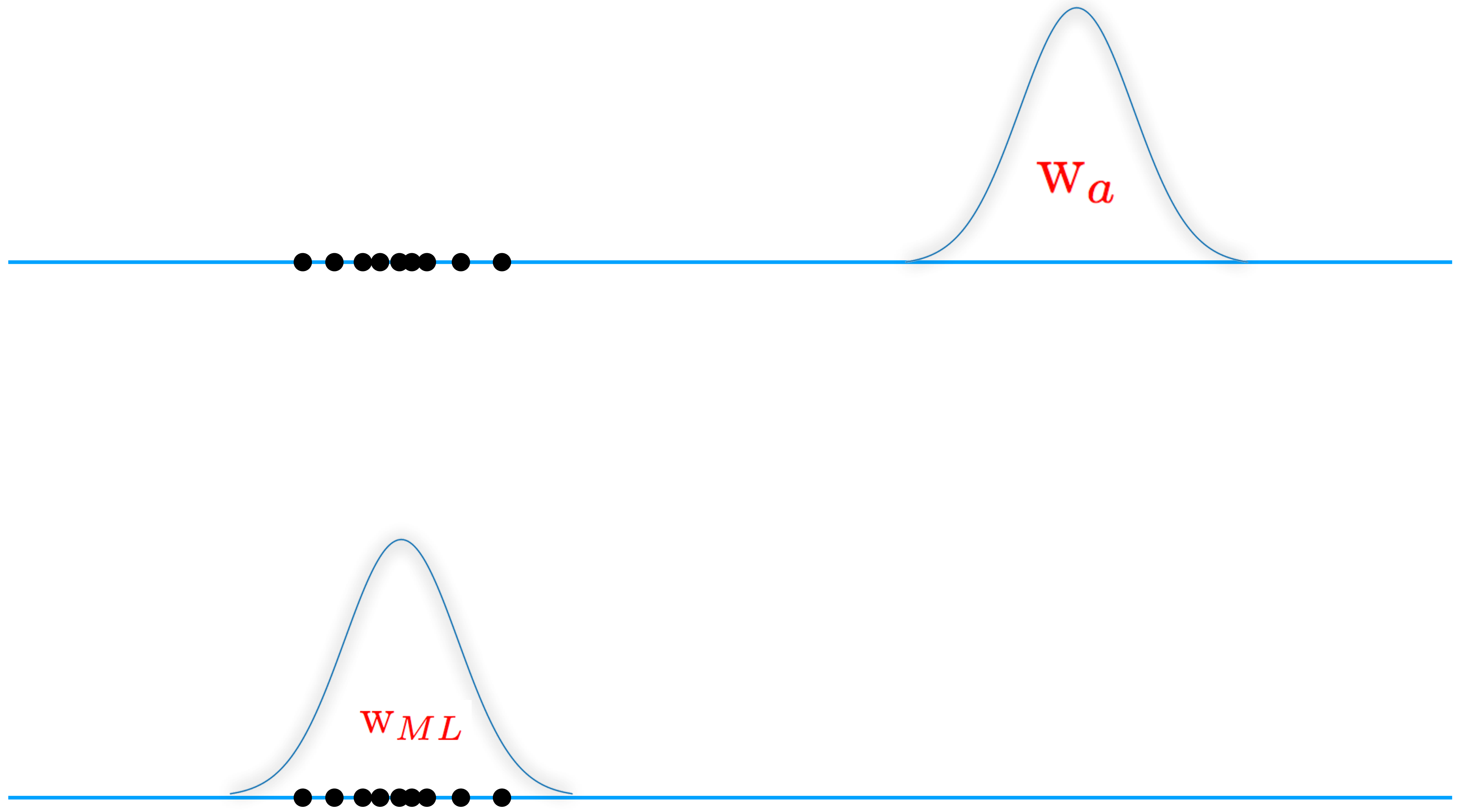
- If i.i.d., joint distribution

$$p(D|\mathbf{w}) = p(x_1, x_2, ..., x_N|\mathbf{w}) = \prod_{i=1}^{N} p(x_i|\mathbf{w})$$

Maximum Likelihood Estimation



Maximum Likelihood Estimation



Maximum Likelihood Estimation

- Maximum likelihood estimation $\boldsymbol{w}_{\mathrm{ML}}$

$$\mathbf{w}_{\mathrm{ML}} = \arg\max_{\mathbf{w}} p(D|\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i=1}^{N} p(x_i|\mathbf{w})$$

numerical underflow/overflow

Maximum Likelihood Estimation

- Maximum likelihood estimation $\boldsymbol{w}_{\mathrm{ML}}$

$$\mathbf{w}_{\mathrm{ML}} = \arg\max_{\mathbf{w}} p(D|\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i=1}^{N} p(x_i|\mathbf{w})$$

numerical underflow/overflow

- How do we maximize?

Maximum Likelihood Estimation

- Maximum likelihood estimation $\boldsymbol{w}_{\text{ML}}$

$$\mathbf{w}_{\text{ML}} = \arg\max_{\mathbf{w}} p(D|\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i=1}^{N} p(x_i|\mathbf{w})$$

numerical underflow/overflow

- How do we maximize?

- Maximize log-likelihood instead:

$$\mathbf{w}_{\text{ML}} = \text{arg}\underset{\mathbf{w}}{max} \prod_{i=1}^{N} p(x_i \,|\, \mathbf{w}) =$$

Maximum Likelihood Estimation

- Maximum likelihood estimation $\boldsymbol{w}_{\text{ML}}$

$$\mathbf{w}_{\text{ML}} = \arg\max_{\mathbf{w}} p(D|\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i=1}^{N} p(x_i|\mathbf{w})$$

numerical underflow/overflow

- How do we maximize?

- Maximize log-likelihood instead:

$$\mathbf{w}_{\text{ML}} = \arg\underset{\mathbf{w}}{max} \prod_{i=1}^{N} p(x_i \,|\, \mathbf{w}) = \arg\max_{\mathbf{w}} \sum_i \log p(x_i|\mathbf{w})$$

Maximum Likelihood Estimation

- Maximum likelihood estimation ***w***ML

$$\mathbf{w}_{\mathrm{ML}} = \arg\max_{\mathbf{w}} p(D|\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i=1}^{N} p(x_i|\mathbf{w})$$

numerical underflow/overflow

- How do we maximize?

- Maximize log-likelihood instead:

$$\mathbf{w}_{\mathrm{ML}} = \arg\max_{\mathbf{w}} \prod_{i=1}^{N} p(x_i \,|\, \mathbf{w}) = \arg\max_{\mathbf{w}} \sum_i \log p(x_i|\mathbf{w})$$

$$= \arg\min_{\mathbf{w}} - \sum_i \log p(x_i|\mathbf{w})$$

Maximum Likelihood Estimation

- Maximum likelihood estimation $\boldsymbol{w}_{\mathrm{ML}}$

$$\mathbf{w}_{\mathrm{ML}} = \arg\max_{\mathbf{w}} p(D|\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i=1}^{N} p(x_i|\mathbf{w})$$

numerical underflow/overflow

- How do we maximize?

- Maximize log-likelihood instead:

$$\mathbf{w}_{\mathrm{ML}} = \underset{\mathbf{w}}{\mathrm{arg}max} \prod_{i=1}^{N} p(x_i\,|\,\mathbf{w}) = \arg\max_{\mathbf{w}} \sum_{i} \log p(x_i|\mathbf{w})$$

$$= \arg\min_{\mathbf{w}} - \sum_{i} \log p(x_i|\mathbf{w})$$

- Error function: $E(D;\mathbf{w}) = -\log p(D|\mathbf{w}) = -\sum_{i=1}^{N} \log p(x_i|\mathbf{w})$

ML Estimator for Gaussian Distributions (I)

- i.i.d. Gaussian distributed real variables $D = (x_1, x_2, ..., x_N)$

$$p(x|\mathbf{w}) = \mathcal{N}(x|\mu, \sigma^2) \quad \Rightarrow \quad p(D|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right]$$

ML Estimator for Gaussian Distributions (I)

- i.i.d. Gaussian distributed real variables $D = (x_1, x_2, ..., x_N)$

$$p(x|\mathbf{w}) = \mathcal{N}(x|\mu, \sigma^2) \quad \Rightarrow \quad p(D|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right]$$

$$p(D|\mathbf{w}) = \prod_{i=1}^{N} p(x_i|\mathbf{w})$$

ML Estimator for Gaussian Distributions (I)

- i.i.d. Gaussian distributed real variables $D = (x_1, x_2, ..., x_N)$

$$p(x|\mathbf{w}) = \mathcal{N}(x|\mu, \sigma^2) \Rightarrow p(D|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right]$$

$$p(D|\mathbf{w}) = \prod_{i=1}^{N} p(x_i|\mathbf{w})$$

- Log likelihood

$$\log p(D|\mu, \sigma^2) =$$

ML Estimator for Gaussian Distributions (I)

- i.i.d. Gaussian distributed real variables $D = (x_1, x_2, ..., x_N)$

$$p(x|\mathbf{w}) = \mathcal{N}(x|\mu, \sigma^2) \Rightarrow p(D|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right]$$

$$p(D|\mathbf{w}) = \prod_{i=1}^{N} p(x_i|\mathbf{w})$$

- Log likelihood

$$\log p(D|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2 - \frac{N}{2} \log \sigma^2 - \frac{N}{2} \log 2\pi$$

ML Estimator for Gaussian Distributions (I)

- i.i.d. Gaussian distributed real variables $D = (x_1, x_2, ..., x_N)$

$$p(x|\mathbf{w}) = \mathcal{N}(x|\mu, \sigma^2) \quad \Rightarrow \quad p(D|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2 \right]$$

$$p(D|\mathbf{w}) = \prod_{i=1}^{N} p(x_i|\mathbf{w})$$

- Log likelihood

$$\log p(D|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2 - \frac{N}{2} \log \sigma^2 - \frac{N}{2} \log 2\pi$$

- Estimate model parameters:

ML Estimator for Gaussian Distributions (I)

- i.i.d. Gaussian distributed real variables $D = (x_1, x_2, ..., x_N)$

$$p(x|\mathbf{w}) = \mathcal{N}(x|\mu, \sigma^2) \quad \xrightarrow[p(D|\mathbf{w}) = \prod_{i=1}^{N} p(x_i|\mathbf{w})]{} \quad p(D|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right]$$

- Log likelihood

$$\log p(D|\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2 - \frac{N}{2}\log\sigma^2 - \frac{N}{2}\log 2\pi$$

- Estimate model parameters:

$$\mu_{ML} = ? \qquad \sigma_{ML} = ?$$

ML Estimator for Gaussian Distributions (II)

- log likelihood:

$$\log p(D|\mu, \sigma^2) = -\frac{N}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2$$

- Maximum Likelihood solution for μ

$$\frac{\partial}{\partial \mu} \log p(D|\mu, \sigma^2) =$$

ML Estimator for Gaussian Distributions (II)

- log likelihood:

$$\log p(D|\mu,\sigma^2) = -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$$

- Maximum Likelihood solution for μ

$$\frac{\partial}{\partial\mu}\log p(D|\mu,\sigma^2) = 0$$

ML Estimator for Gaussian Distributions (II)

- log likelihood:

$$\log p(D|\mu, \sigma^2) = -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2$$

- Maximum Likelihood solution for μ

$$\frac{\partial}{\partial \mu}\log p(D|\mu, \sigma^2) = 0 = \frac{2}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)$$

ML Estimator for Gaussian Distributions (II)

- log likelihood:

$$\log p(D|\mu,\sigma^2) = -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$$

- Maximum Likelihood solution for μ

$$\frac{\partial}{\partial\mu}\log p(D|\mu,\sigma^2) = 0 = \frac{2}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)$$

$$\rightarrow \sum_{i=1}^{N}x_i = \sum_{i=1}^{N}\mu$$

ML Estimator for Gaussian Distributions (II)

- log likelihood:

$$\log p(D|\mu,\sigma^2) = -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$$

- Maximum Likelihood solution for μ

$$\frac{\partial}{\partial\mu}\log p(D|\mu,\sigma^2) = 0 = \frac{2}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)$$

$$\rightarrow \sum_{i=1}^{N} x_i = \sum_{i=1}^{N}\mu \rightarrow \sum_{i=1}^{N} x_i = N\mu$$

ML Estimator for Gaussian Distributions (II)

- log likelihood:

$$\log p(D|\mu,\sigma^2) = -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$$

- Maximum Likelihood solution for μ

$$\frac{\partial}{\partial\mu}\log p(D|\mu,\sigma^2) = 0 = \frac{2}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)$$

$$\rightarrow \sum_{i=1}^{N}x_i = \sum_{i=1}^{N}\mu \rightarrow \sum_{i=1}^{N}x_i = N\mu$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N}x_i$$

ML Estimator for Gaussian Distributions (II)

- log likelihood:

$$\log p(D|\mu, \sigma^2) = -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2$$

- Maximum Likelihood solution for $\underline{\sigma^2}$

$$\frac{\partial}{\partial\sigma^2}\log p(D|\mu, \sigma^2) =$$

ML Estimator for Gaussian Distributions (II)

- log likelihood:

$$\log p(D|\mu,\sigma^2) = -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$$

- Maximum Likelihood solution for σ^2

$$\frac{\partial}{\partial\sigma^2}\log p(D|\mu,\sigma^2) = \textcolor{red}{0}$$

ML Estimator for Gaussian Distributions (II)

- log likelihood:

$$\log p(D|\mu, \sigma^2) = -\frac{N}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2$$

- Maximum Likelihood solution for σ^2

$$\frac{\partial}{\partial \sigma^2} \log p(D|\mu, \sigma^2) = 0$$

$$\sigma^2_{ML} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

ML Estimator for Gaussian Distributions (IV)

- How well do the ML estimators represent the true parameters? $$p(D|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}(x_i-\mu)^2\right]$$
- If I draw multiple datasets, what is the expected value of μ_{ML}?
- ML estimate of the mean:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\mu_{\mathrm{ML}}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] =$$

- Bias of estimator:

$$\mathbb{E}[\mu_{\mathrm{ML}}] - \mu$$

ML Estimator for Gaussian Distributions (IV)

- How well do the ML estimators represent the true parameters? $p(D|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right]$

- If I draw multiple datasets, what is the expected value of μ_{ML}?

- ML estimate of the mean:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\mu_{\mathrm{ML}}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\sum E[x_i] = \mu$$

- Bias of estimator:

$$\mathbb{E}[\mu_{\mathrm{ML}}] - \mu$$

ML Estimator for Gaussian Distributions (IV)

- How well do the ML estimators represent the true parameters? $p(D|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right]$
- If I draw multiple datasets, what is the expected value of μ_{ML}?
- ML estimate of the mean:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\mu_{\mathrm{ML}}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\sum E[x_i] = \mu$$

- Bias of estimator:

$$\mathbb{E}[\mu_{\mathrm{ML}}] - \mu = 0$$

ML Estimator for Gaussian Distributions (V)

- ML estimate of the variance:

$$\mathbb{E}_{D \sim p(D|\mu,\sigma^2)}[\sigma^2_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{n=1}^{N}x_n\right)^2\right] =$$

ML Estimator for Gaussian Distributions (V)

- ML estimate of the variance:

$$\mathbb{E}_{D \sim p(D|\mu,\sigma^2)}[\sigma^2_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{n=1}^{N}x_n\right)^2\right] =$$

$$\frac{1}{N}\sum_{i=1}^{N} E[(x_i - \frac{1}{N}\sum_{j=1}^{N} x_j)^2]$$

ML Estimator for Gaussian Distributions (V)

- ML estimate of the variance:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\sigma^2_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{n=1}^{N}x_n\right)^2\right] =$$

$$\frac{1}{N}\sum_{i=1}^{N}E[(x_i - \frac{1}{N}\sum_{j=1}^{N}x_j)^2] = \frac{1}{N}\sum_{i=1}^{N}E[x_i x_i - \frac{2}{N}x_i\sum_{j=1}^{N}x_j + \frac{1}{N^2}\sum_{k=1,l=1}^{N}x_k x_l]$$

ML Estimator for Gaussian Distributions (V)

- ML estimate of the variance:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\sigma^2_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{n=1}^{N}x_n\right)^2\right] =$$

$$\frac{1}{N}\sum_{i=1}^{N}E[(x_i - \frac{1}{N}\sum_{j=1}^{N}x_j)^2] = \frac{1}{N}\sum_{i=1}^{N}E[x_ix_i - \frac{2}{N}x_i\sum_{j=1}^{N}x_j + \frac{1}{N^2}\sum_{k=1,l=1}^{N}x_kx_l]$$

$$E[x_ix_j] = \begin{cases} E[x^2] = \mu^2 + \sigma^2 & \text{if} \quad i = j \\ E[x_i]E[x_j] = \mu^2 & \text{if} \quad i \neq j \end{cases}$$

ML Estimator for Gaussian Distributions (V)

- ML estimate of the variance:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\sigma^2_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{n=1}^{N}x_n\right)^2\right] =$$

$$\frac{1}{N}\sum_{i=1}^{N}E[(x_i - \frac{1}{N}\sum_{j=1}^{N}x_j)^2] = \frac{1}{N}\sum_{i=1}^{N}E[x_i x_i - \frac{2}{N}x_i\sum_{j=1}^{N}x_j + \frac{1}{N^2}\sum_{k=1,l=1}^{N}x_k x_l]$$

$$E[x_i x_j] = \begin{cases} E[x^2] = \mu^2 + \sigma^2 & \text{if} \quad i = j \\ E[x_i]E[x_j] = \mu^2 & \text{if} \quad i \neq j \end{cases}$$

$$\boxed{\sigma^2 = E[x^2] - \mu^2}$$

ML Estimator for Gaussian Distributions (V)

- ML estimate of the variance:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\sigma_{ML}^2] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{n=1}^{N}x_n\right)^2\right] =$$

$$\frac{1}{N}\sum_{i=1}^{N}E[(x_i - \frac{1}{N}\sum_{j=1}^{N}x_j)^2] = \frac{1}{N}\sum_{i=1}^{N}E[x_ix_i - \frac{2}{N}x_i\sum_{j=1}^{N}x_j + \frac{1}{N^2}\sum_{k=1,l=1}^{N}x_kx_l]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\Big(\mu^2+\sigma^2-\frac{2}{N}(\mu^2+\sigma^2)+\frac{2}{N}(N-1)\mu^2+\frac{1}{N^2}N(N-1)\mu^2+\frac{1}{N^2}N(\mu^2\sigma^2)\Big)$$

$$\boxed{\sigma^2 = E[x^2] - \mu^2}$$

$$E[x_ix_j] = \begin{cases} E[x^2] = \mu^2 + \sigma^2 & \text{if} \quad i = j \\ E[x_i]E[x_j] = \mu^2 & \text{if} \quad i \neq j \end{cases}$$

ML Estimator for Gaussian Distributions (V)

- ML estimate of the variance:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\sigma^2_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{n=1}^{N}x_n\right)^2\right] =$$

$$\frac{1}{N}\sum_{i=1}^{N}E[(x_i - \frac{1}{N}\sum_{j=1}^{N}x_j)^2] = \frac{1}{N}\sum_{i=1}^{N}E[x_i x_i - \frac{2}{N}x_i\sum_{j=1}^{N}x_j + \frac{1}{N^2}\sum_{k=1,l=1}^{N}x_k x_l]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(\mu^2+\sigma^2 - \frac{2}{N}(\mu^2+\sigma^2) + \frac{2}{N}(N-1)\mu^2 + \frac{1}{N^2}N(N-1)\mu^2 + \frac{1}{N^2}N(\mu^2\sigma^2)\right)$$

$$E[x_i x_j] = \begin{cases} E[x^2] = \mu^2 + \sigma^2 & \text{if} \quad i = j \\ E[x_i]E[x_j] = \mu^2 & \text{if} \quad i \neq j \end{cases}$$

$$\boxed{\sigma^2 = E[x^2] - \mu^2}$$

ML Estimator for Gaussian Distributions (V)

- ML estimate of the variance:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\sigma^2_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{n=1}^{N}x_n\right)^2\right] =$$

$$\frac{1}{N}\sum_{i=1}^{N}E[(x_i - \frac{1}{N}\sum_{j=1}^{N}x_j)^2] = \frac{1}{N}\sum_{i=1}^{N}E[\underline{x_i x_i} - \underline{\frac{2}{N}x_i\sum_{j=1}^{N}x_j} + \frac{1}{N^2}\sum_{k=1,l=1}^{N}x_k x_l]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\Big(\underline{\mu^2+\sigma^2} - \underline{\frac{2}{N}(\mu^2+\sigma^2)} + \frac{2}{N}(N-1)\mu^2 + \frac{1}{N^2}N(N-1)\mu^2 + \frac{1}{N^2}N(\mu^2\sigma^2)\Big)$$

$$E[x_i x_j] = \begin{cases} E[x^2] = \mu^2 + \sigma^2 & \text{if} \quad i = j \\ E[x_i]E[x_j] = \mu^2 & \text{if} \quad i \neq j \end{cases}$$

$$\boxed{\sigma^2 = E[x^2] - \mu^2}$$

ML Estimator for Gaussian Distributions (V)

- ML estimate of the variance:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\sigma^2_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{n=1}^{N} x_n\right)^2\right] =$$

$$\frac{1}{N}\sum_{i=1}^{N} E[(x_i - \frac{1}{N}\sum_{j=1}^{N} x_j)^2] = \frac{1}{N}\sum_{i=1}^{N} E[x_i x_i - \frac{2}{N}x_i\sum_{j=1}^{N} x_j + \frac{1}{N^2}\sum_{k=1,l=1}^{N} x_k x_l]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\Big(\mu^2+\sigma^2 - \frac{2}{N}(\mu^2+\sigma^2) + \frac{2}{N}(N-1)\mu^2 + \frac{1}{N^2}N(N-1)\mu^2 + \frac{1}{N^2}N(\mu^2\sigma^2)\Big)$$

$$E[x_i x_j] = \begin{cases} E[x^2] = \mu^2 + \sigma^2 & \text{if} \quad i = j \\ E[x_i]E[x_j] = \mu^2 & \text{if} \quad i \neq j \end{cases}$$

$$\boxed{\sigma^2 = E[x^2] - \mu^2}$$

ML Estimator for Gaussian Distributions (V)

- ML estimate of the variance:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\sigma_{ML}^2] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{n=1}^{N}x_n\right)^2\right] =$$

$$\frac{1}{N}\sum_{i=1}^{N}E[(x_i - \frac{1}{N}\sum_{j=1}^{N}x_j)^2] = \frac{1}{N}\sum_{i=1}^{N}E[\underline{x_i x_i} - \underline{\frac{2}{N}x_i\sum_{j=1}^{N}x_j} + \underline{\frac{1}{N^2}\sum_{k=1,l=1}^{N}x_k x_l}]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(\underline{\mu^2+\sigma^2} - \underline{\frac{2}{N}(\mu^2+\sigma^2)} + \underline{\frac{2}{N}(N-1)\mu^2} + \frac{1}{N^2}\underline{N(N-1)\mu^2} + \frac{1}{N^2}N(\mu^2\sigma^2)\right)$$

$$E[x_i x_j] = \begin{cases} E[x^2] = \mu^2 + \sigma^2 & \text{if} \quad i = j \\ E[x_i]E[x_j] = \mu^2 & \text{if} \quad i \neq j \end{cases}$$

$$\boxed{\sigma^2 = E[x^2] - \mu^2}$$

ML Estimator for Gaussian Distributions (V)

- ML estimate of the variance:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\sigma^2_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{n=1}^{N}x_n\right)^2\right] =$$

$$\frac{1}{N}\sum_{i=1}^{N}E[(x_i - \frac{1}{N}\sum_{j=1}^{N}x_j)^2] = \frac{1}{N}\sum_{i=1}^{N}E[\underline{x_i x_i} - \underline{\frac{2}{N}x_i\sum_{j=1}^{N}x_j} + \underline{\frac{1}{N^2}\sum_{k=1,l=1}^{N}x_k x_l}]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(\underline{\mu^2+\sigma^2} - \underline{\frac{2}{N}(\mu^2+\sigma^2)} + \underline{\frac{2}{N}(N-1)\mu^2} + \underline{\frac{1}{N^2}N(N-1)\mu^2} + \underline{\frac{1}{N^2}N(\mu^2\sigma^2)}\right)$$

$$\boxed{\sigma^2 = E[x^2] - \mu^2}$$

$$E[x_i x_j] = \begin{cases} E[x^2] = \mu^2 + \sigma^2 & \text{if} \quad i = j \\ E[x_i]E[x_j] = \mu^2 & \text{if} \quad i \neq j \end{cases}$$

ML Estimator for Gaussian Distributions (V)

- ML estimate of the variance:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\sigma^2_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{n=1}^{N}x_n\right)^2\right] =$$

$$\frac{1}{N}\sum_{i=1}^{N}E[(x_i-\frac{1}{N}\sum_{j=1}^{N}x_j)^2] = \frac{1}{N}\sum_{i=1}^{N}E[x_ix_i - \frac{2}{N}x_i\sum_{j=1}^{N}x_j + \frac{1}{N^2}\sum_{k=1,l=1}^{N}x_kx_l]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\Big(\mu^2+\sigma^2-\frac{2}{N}(\mu^2+\sigma^2)+\frac{2}{N}(N-1)\mu^2+\frac{1}{N^2}N(N-1)\mu^2+\frac{1}{N^2}N(\mu^2\sigma^2)\Big)$$

ML Estimator for Gaussian Distributions (V)

- ML estimate of the variance:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\sigma^2_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{n=1}^{N}x_n\right)^2\right] =$$

$$\frac{1}{N}\sum_{i=1}^{N}E[(x_i - \frac{1}{N}\sum_{j=1}^{N}x_j)^2] = \frac{1}{N}\sum_{i=1}^{N}E[x_i x_i - \frac{2}{N}x_i\sum_{j=1}^{N}x_j + \frac{1}{N^2}\sum_{k=1,l=1}^{N}x_k x_l]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\Big(\mu^2+\sigma^2-\frac{2}{N}(\mu^2+\sigma^2)+\frac{2}{N}(N-1)\mu^2+\frac{1}{N^2}N(N-1)\mu^2+\frac{1}{N^2}N(\mu^2\sigma^2)\Big)$$

$$\cdots = \sigma^2(\frac{N-1}{N})$$

ML Estimator for Gaussian Distributions (VI)

- For data generated from

$$p(D|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}(x_i-\mu)^2\right]$$

- ML gives biased estimator

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \quad \cdots = \sigma^2\left(\frac{N-1}{N}\right)$$

Variance is underestimated, because it is measured relative to the sampled mean

ML Estimator for Gaussian Distributions (VI)

- For data generated from

$$p(D|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right]$$

- ML gives biased estimator

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \quad \cdots = \sigma^2 (\frac{N-1}{N})$$

Variance is underestimated, because it is measured relative to the sampled mean

- Unbiased variance estimator:

$$\tilde{\sigma}^2 =$$

ML Estimator for Gaussian Distributions (VI)

- For data generated from

$$p(D|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}(x_i-\mu)^2\right]$$

- ML gives biased estimator

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \quad \cdots = \sigma^2\left(\frac{N-1}{N}\right)$$

Variance is underestimated, because it is measured relative to the sampled mean

- Unbiased variance estimator:

$$\tilde{\sigma}^2 \quad = \frac{N}{N-1}\sigma_{ML}$$

Biased Maximum Likelihood Estimator

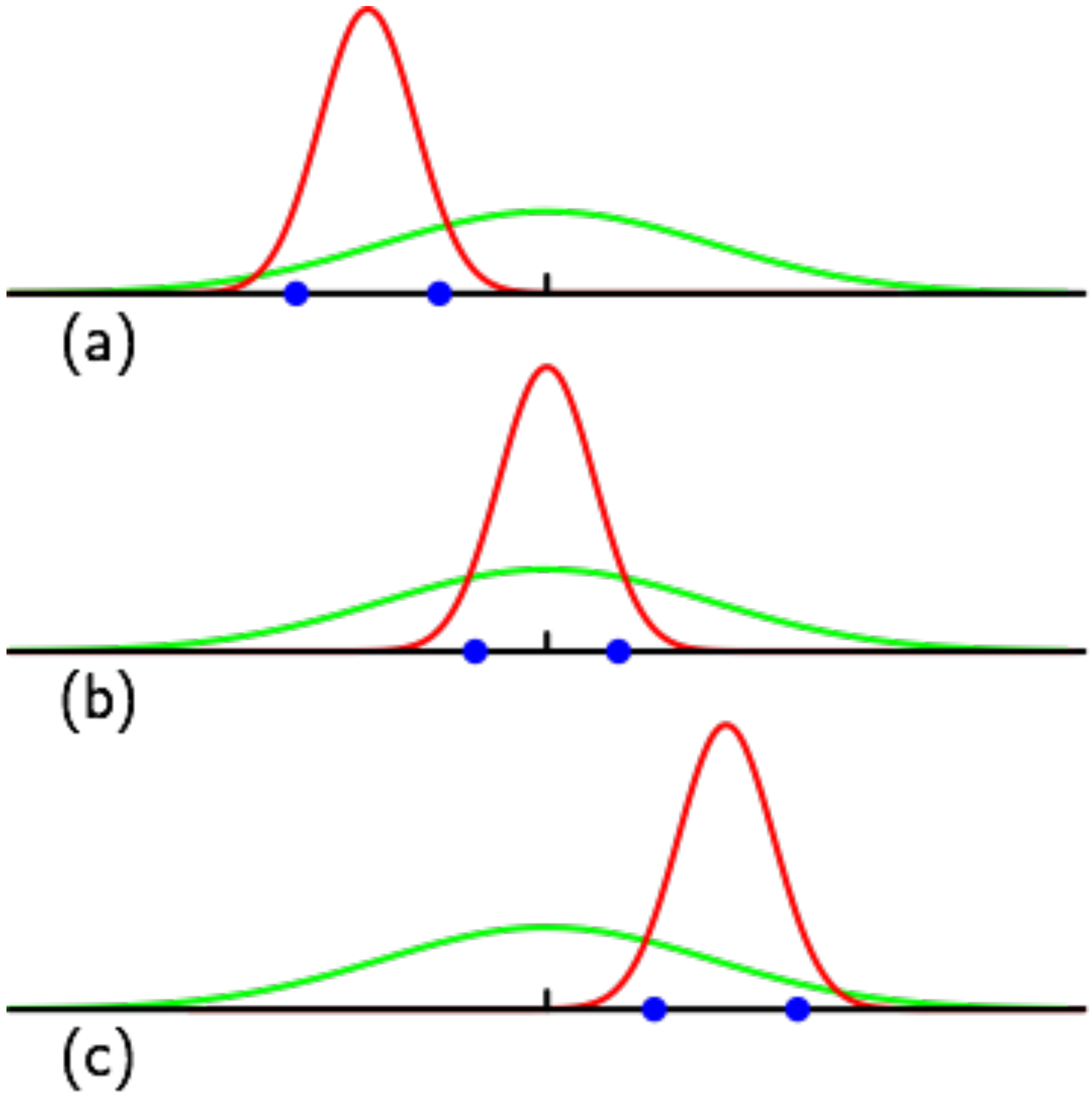


Figure: Bias in ML estimator for variance (Bishop 1.15)

Curve Fitting: Maximum Likelihood Estimates

- Data $D = \{(x_1, t_1), ..., (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$

- Assume targets are generated by

$$t = y(x, \mathbf{w}) + \sigma\varepsilon, \quad \varepsilon \sim \mathcal{N}(\varepsilon|0, 1)$$

- Target distribution:

$$p(t|x, \mathbf{w}, \beta) =$$

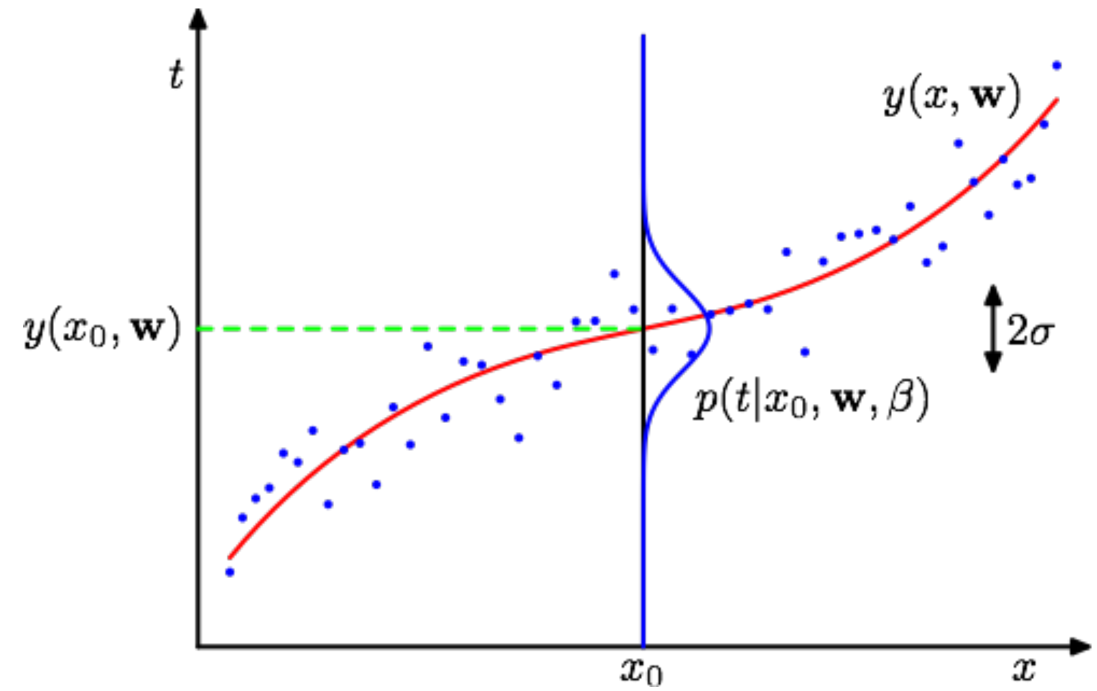


Figure: Gaussian conditional distribution (Bishop 1.16)

Curve Fitting: Maximum Likelihood Estimates

- Data $D = \{(x_1, t_1), ..., (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$

- Assume targets are generated by

$$t = y(x, \mathbf{w}) + \sigma\varepsilon, \quad \varepsilon \sim \mathcal{N}(\varepsilon|0, 1)$$

- Target distribution:

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})$$

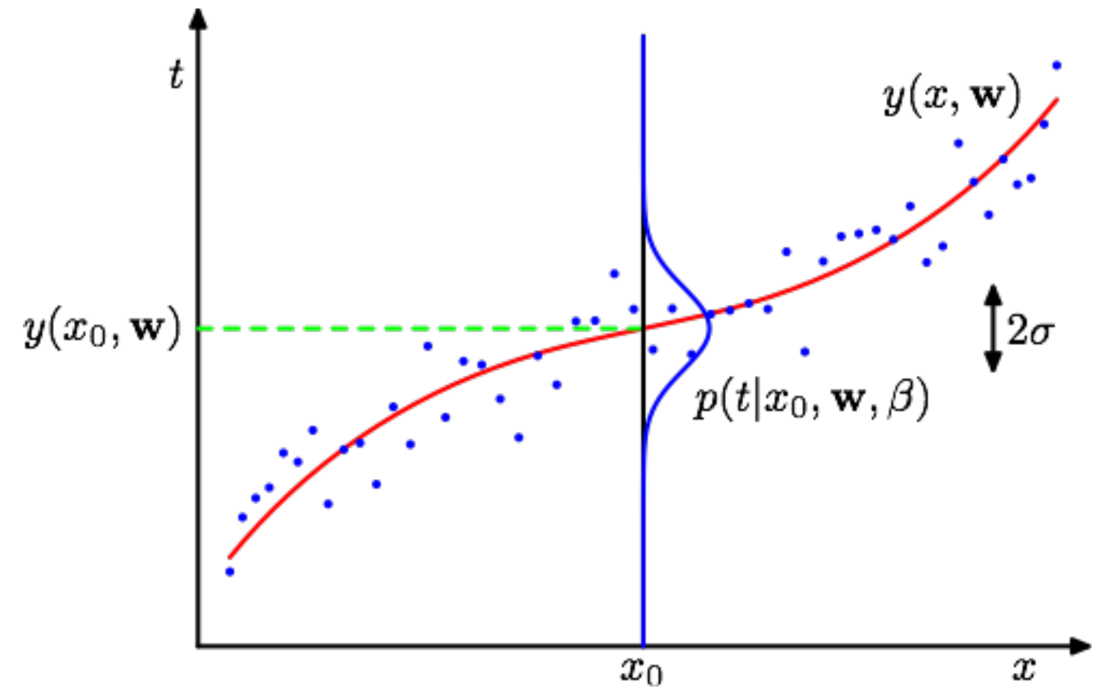


Figure: Gaussian conditional distribution (Bishop 1.16)

Curve Fitting: Maximum Likelihood Estimates

- Data $D = \{(x_1, t_1), ..., (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$

- Assume targets are generated by

$$t = y(x, \mathbf{w}) + \sigma\varepsilon, \quad \varepsilon \sim \mathcal{N}(\varepsilon|0,1)$$

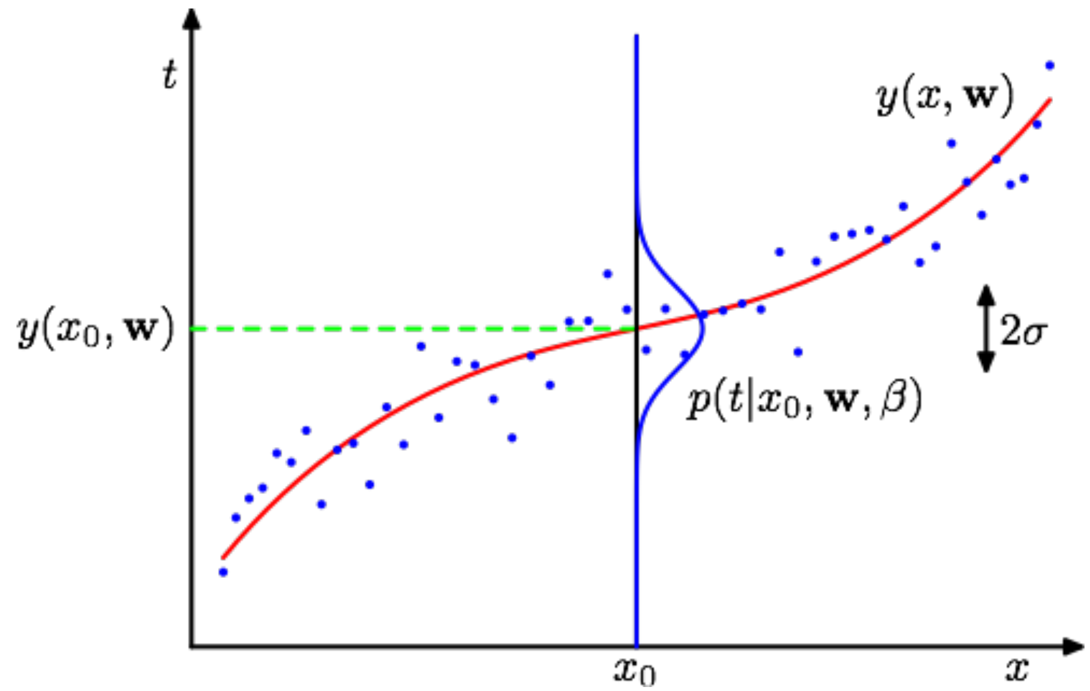


Figure: Gaussian conditional distribution (Bishop 1.16)

- Target distribution:

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) = \frac{1}{\sqrt{2\pi\beta^{-1}}} e^{-\frac{\beta}{2}(t - y(x, \mathbf{w}))^2}$$

Curve Fitting: Maximum Likelihood Estimates

- Data $D = \{(x_1, t_1), ..., (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$

- Assume targets are generated by

$$t = y(x, \mathbf{w}) + \sigma\varepsilon, \quad \varepsilon \sim \mathcal{N}(\varepsilon|0, 1)$$

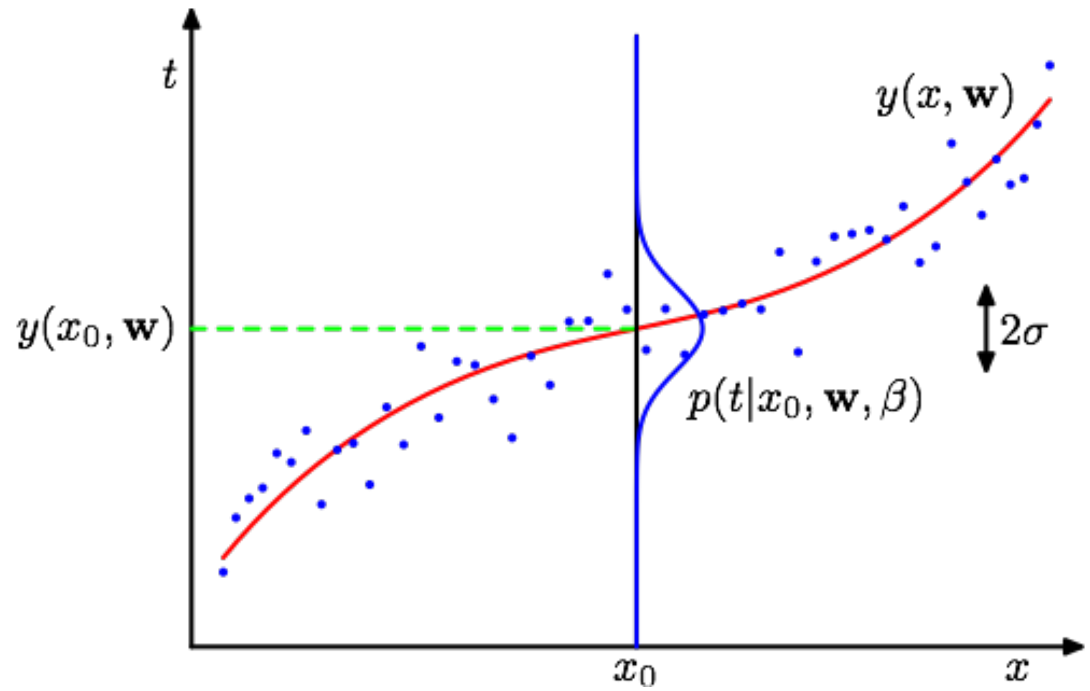


Figure: Gaussian conditional distribution (Bishop 1.16)

- Target distribution:

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) = \frac{1}{\sqrt{2\pi\beta^{-1}}} e^{-\frac{\beta}{2}(t - y(x,\mathbf{w}))^2}$$

- Log likelihood:

$$\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) =$$

Curve Fitting: Maximum Likelihood Estimates

- Data $D = \{(x_1, t_1), ..., (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$

- Assume targets are generated by

$$t = y(x, \mathbf{w}) + \sigma\varepsilon, \quad \varepsilon \sim \mathcal{N}(\varepsilon|0, 1)$$

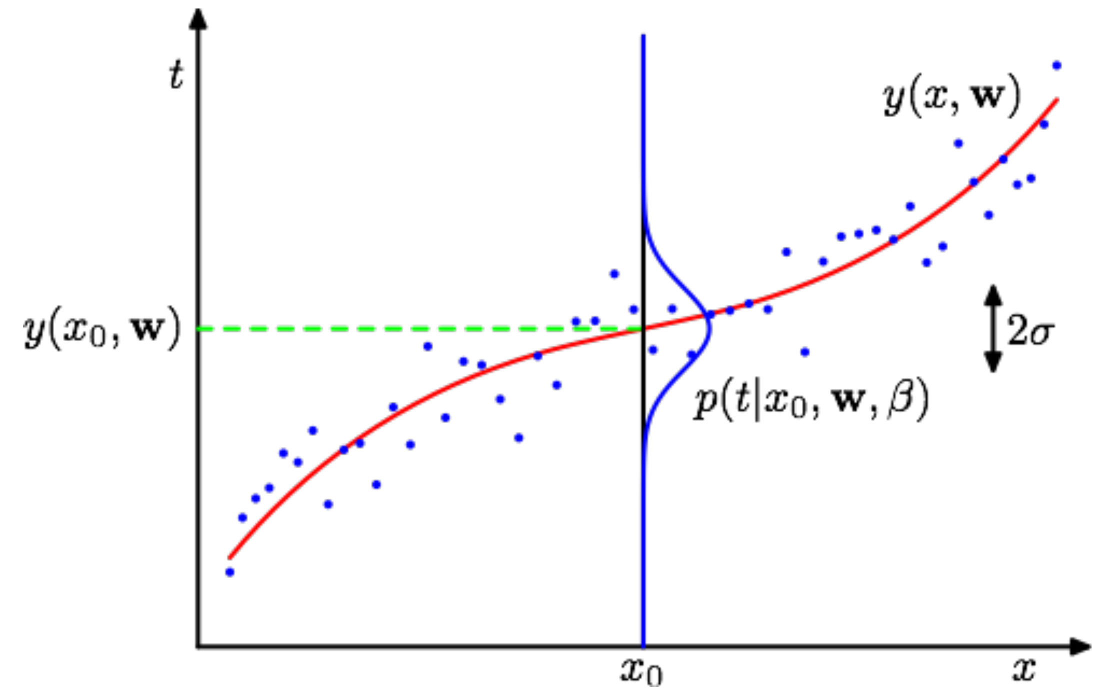


Figure: Gaussian conditional distribution (Bishop 1.16)

- Target distribution:

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) = \frac{1}{\sqrt{2\pi\beta^{-1}}} e^{-\frac{\beta}{2}(t-y(x,\mathbf{w}))^2}$$

- Log likelihood:

$$\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = log \prod_{i=1}^{N} \mathcal{N}(t_i|y(x_i, \mathbf{w}), \beta^{-1})$$

Curve Fitting: Maximum Likelihood Estimates

- Data $D = \{(x_1, t_1), ..., (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$

- Assume targets are generated by

$$t = y(x, \mathbf{w}) + \sigma\varepsilon, \quad \varepsilon \sim \mathcal{N}(\varepsilon|0, 1)$$

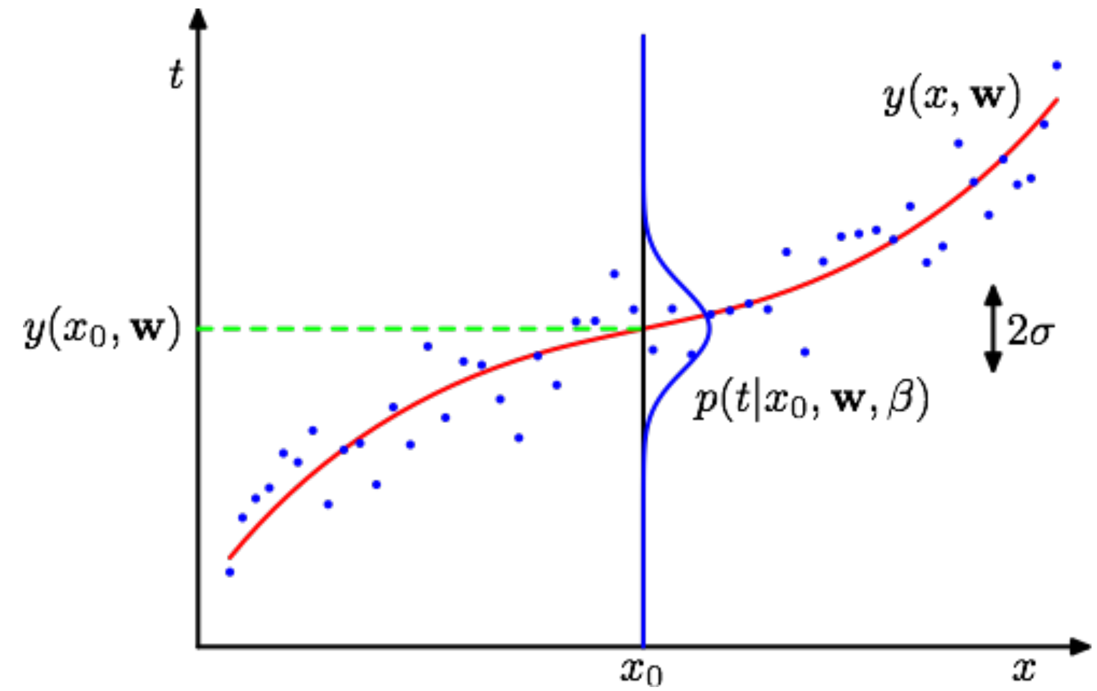


Figure: Gaussian conditional distribution (Bishop 1.16)

- Target distribution:

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) = \frac{1}{\sqrt{2\pi\beta^{-1}}} e^{-\frac{\beta}{2}(t - y(x,\mathbf{w}))^2}$$

- Log likelihood:

$$\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = log \prod_{i=1}^{N} \mathcal{N}(t_i|y(x_i, \mathbf{w}), \beta^{-1}) = -\frac{\beta}{2} \sum_{i=1}^{N} (y(x_i, \mathbf{w}) - t_i)^2 + \frac{N}{2} \log \beta - \frac{N}{2} \log 2\pi$$

Curve Fitting: Maximum Likelihood Estimates

- ML: minimize $E(\mathbf{x}, \mathbf{t}, \mathbf{w}, \beta) = -\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)$ w.r.t. ***w*** and $\boldsymbol{\beta}$

$$E(\mathbf{x}, \mathbf{t}, \mathbf{w}, \beta) = \frac{\beta}{2} \sum_{i=1}^{N} (y(x_i, \mathbf{w}) - t_i)^2 - \frac{N}{2} \log \beta + \frac{N}{2} \log 2\pi$$

- Maximum likelihood solution:

$$\mathbf{w}_{\text{ML}} = \arg\min_{\mathbf{w}} \frac{\beta}{2} \sum_{i=1}^{N} (y(x_i, \mathbf{w}) - t_i)^2$$

$$\frac{1}{\beta_{\text{ML}}} =$$

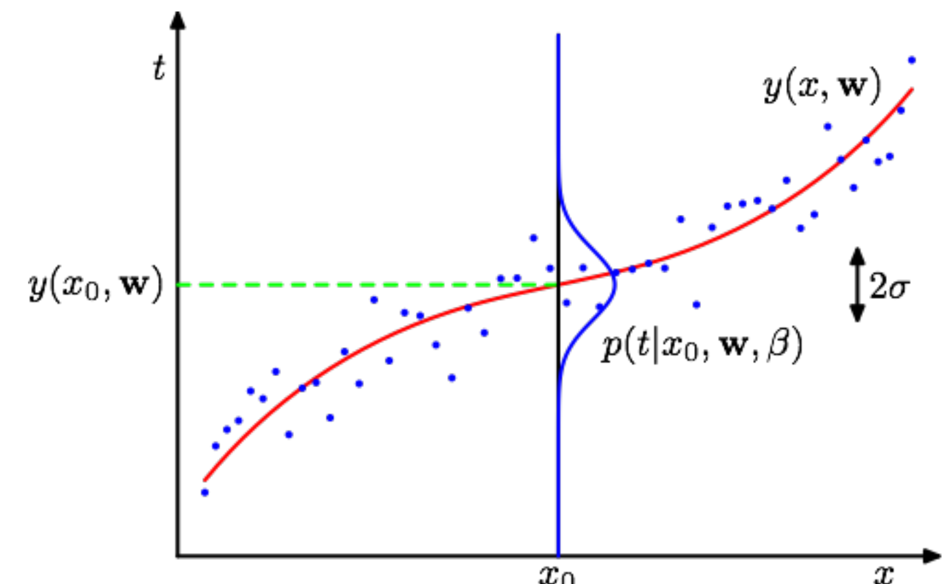


Figure: Gaussian conditional distribution (Bishop 1.16)

Curve Fitting: Maximum Likelihood Estimates

- ML: minimize $E(\mathbf{x}, \mathbf{t}, \mathbf{w}, \beta) = -\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)$ w.r.t. ***w*** and $\boldsymbol{\beta}$

$$E(\mathbf{x}, \mathbf{t}, \mathbf{w}, \beta) = \frac{\beta}{2} \sum_{i=1}^{N} (y(x_i, \mathbf{w}) - t_i)^2 - \frac{N}{2} \log \beta + \frac{N}{2} \log 2\pi$$

- Maximum likelihood solution:

$$\mathbf{w}_{\mathrm{ML}} = \arg\min_{\mathbf{w}} \frac{\beta}{2} \sum_{i=1}^{N} (y(x_i, \mathbf{w}) - t_i)^2$$

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N} \sum_{i=1}^{N} \Big(y(x_i, \mathbf{w}_{ML}) - t_i\Big)^2$$

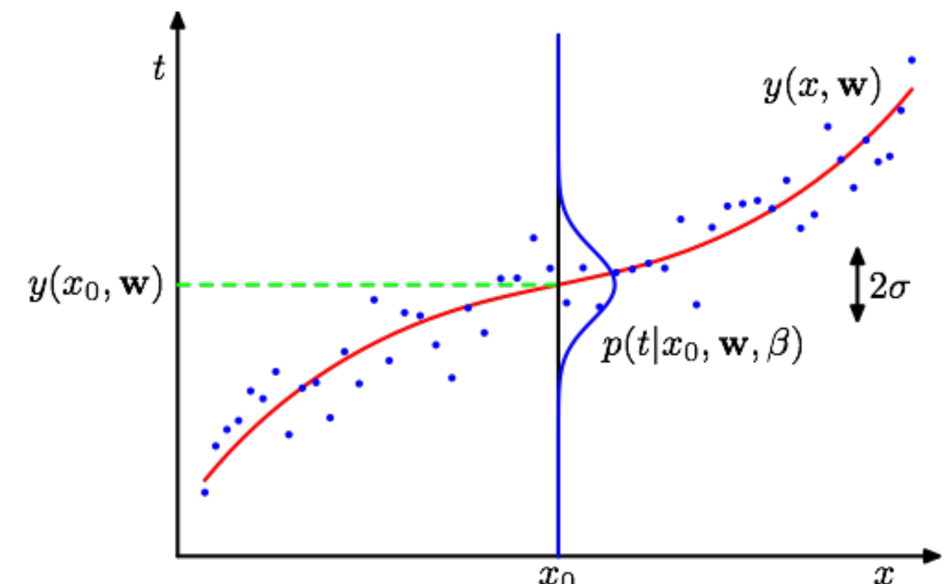


Figure: Gaussian conditional distribution (Bishop 1.16)

Curve Fitting: Maximum Likelihood Estimates

- ML: minimize $E(\mathbf{x}, \mathbf{t}, \mathbf{w}, \beta) = -\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)$ w.r.t. ***w*** and $\boldsymbol{\beta}$

$$E(\mathbf{x}, \mathbf{t}, \mathbf{w}, \beta) = \frac{\beta}{2} \sum_{i=1}^{N} (y(x_i, \mathbf{w}) - t_i)^2 - \frac{N}{2} \log \beta + \frac{N}{2} \log 2\pi$$

- Maximum likelihood solution:

$$\mathbf{w}_{\mathrm{ML}} = \arg\min_{\mathbf{w}} \frac{\beta}{2} \sum_{i=1}^{N} (y(x_i, \mathbf{w}) - t_i)^2$$

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N} \sum_{i=1}^{N} \left(y(x_i, \mathbf{w}_{ML}) - t_i \right)^2$$

- Predictive distribution:

$$p(t' \,|\, x', \mathbf{w}_{\mathrm{ML}}, \beta_{ML}) =$$

$$\mathbb{E}[t' \,|\, x', \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}] =$$

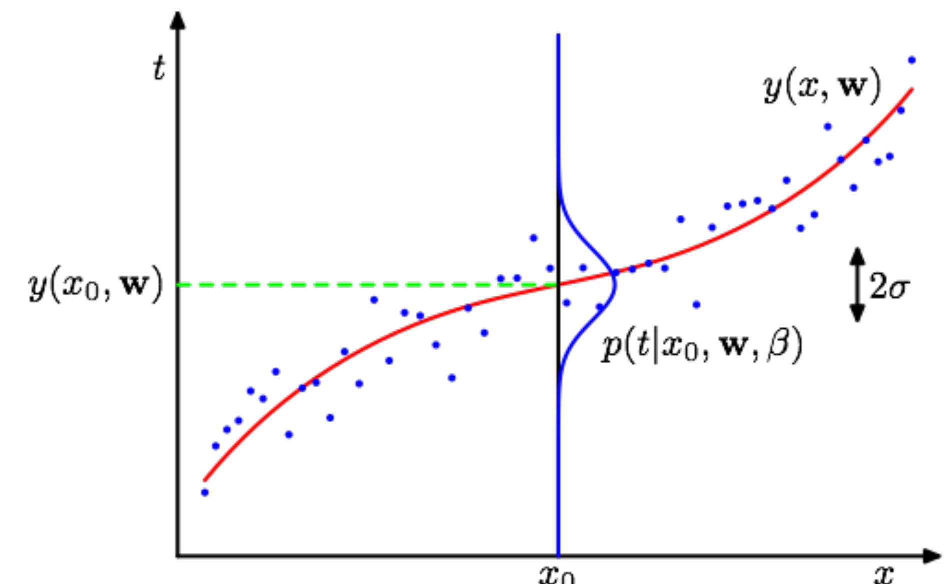


Figure: Gaussian conditional distribution (Bishop 1.16)

Curve Fitting: Maximum Likelihood Estimates

- ML: minimize $E(\mathbf{x}, \mathbf{t}, \mathbf{w}, \beta) = -\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)$ w.r.t. ***w*** and $\boldsymbol{\beta}$

$$E(\mathbf{x}, \mathbf{t}, \mathbf{w}, \beta) = \frac{\beta}{2} \sum_{i=1}^{N} (y(x_i, \mathbf{w}) - t_i)^2 - \frac{N}{2} \log \beta + \frac{N}{2} \log 2\pi$$

- Maximum likelihood solution:

$$\mathbf{w}_{\mathrm{ML}} = \arg\min_{\mathbf{w}} \frac{\beta}{2} \sum_{i=1}^{N} (y(x_i, \mathbf{w}) - t_i)^2$$

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N} \sum_{i=1}^{N} \left(y(x_i, \mathbf{w}_{ML}) - t_i \right)^2$$

- Predictive distribution:

$$p(t' \,|\, x', \mathbf{w}_{\mathrm{ML}}, \beta_{ML}) = \mathcal{N}(t' | y(x', \mathbf{w}_{ML}), \beta_{ML}^{-1})$$

$$\mathbb{E}[t' \,|\, x', \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}] =$$

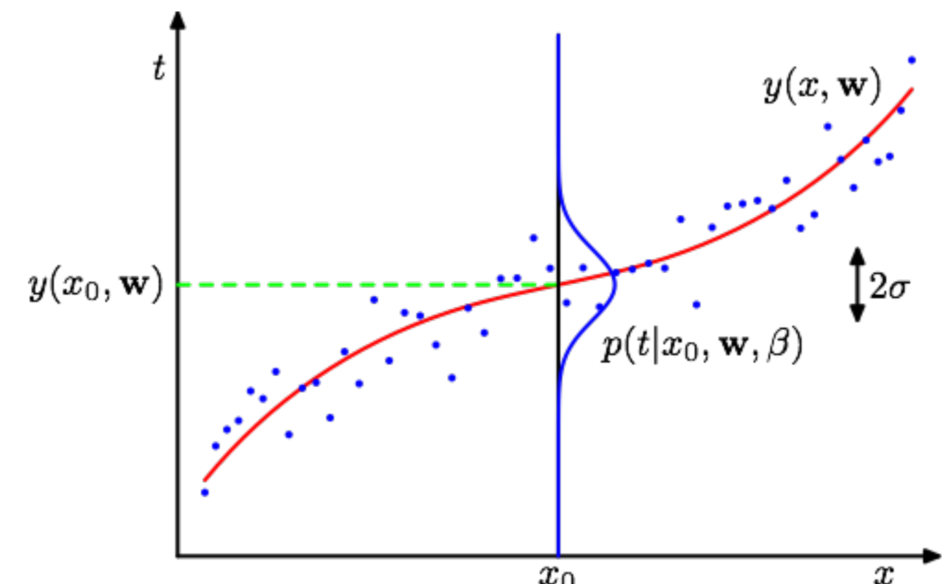


Figure: Gaussian conditional distribution (Bishop 1.16)

Curve Fitting: Maximum Likelihood Estimates

- ML: minimize $E(\mathbf{x}, \mathbf{t}, \mathbf{w}, \beta) = -\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)$ w.r.t. $\boldsymbol{w}$ and $\boldsymbol{\beta}$

$$E(\mathbf{x}, \mathbf{t}, \mathbf{w}, \beta) = \frac{\beta}{2} \sum_{i=1}^{N} (y(x_i, \mathbf{w}) - t_i)^2 - \frac{N}{2} \log \beta + \frac{N}{2} \log 2\pi$$

- Maximum likelihood solution:

$$\mathbf{w}_{\mathrm{ML}} = \arg\min_{\mathbf{w}} \frac{\beta}{2} \sum_{i=1}^{N} (y(x_i, \mathbf{w}) - t_i)^2$$

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N} \sum_{i=1}^{N} \left(y(x_i, \mathbf{w}_{ML}) - t_i\right)^2$$

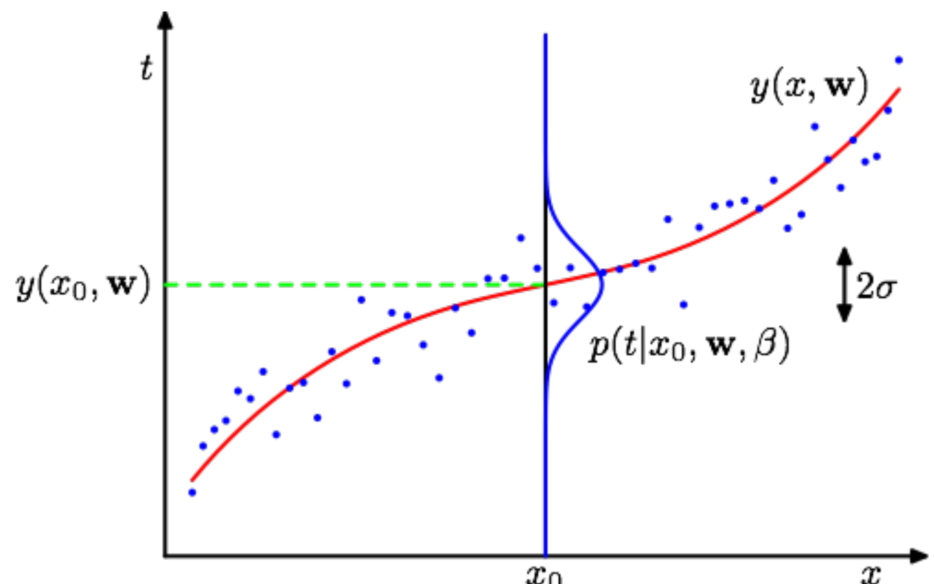


Figure: Gaussian conditional distribution (Bishop 1.16)

- Predictive distribution:

$$p(t' | x', \mathbf{w}_{\mathrm{ML}}, \beta_{ML}) = \mathcal{N}(t' | y(x', \mathbf{w}_{ML}), \beta_{ML}^{-1})$$

$$\mathbb{E}[t' | x', \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}] = y(x', \mathbf{w}_{ML})$$

Overview

1. Probability theory

2. **Statistical learning principles:**

   I. Maximum likelihood

   II. Maximum a posteriori

   III. Bayesian prediction

Maximum A Posteriori Estimates

- Dataset $D = (x_1, x_2, \ldots, x_N)$ of N independent observations.

- ML estimate: choose $\boldsymbol{w}$ such that data likelihood is maximized:

$$\mathbf{w}_{\mathrm{ML}} = \arg\max_{\mathbf{w}} p(D|\mathbf{w})$$

- MAP estimate: choose most probable $\boldsymbol{w}$ given the data.

$$\mathbf{w}_{\mathrm{MAP}} =$$

Maximum A Posteriori Estimates

- Dataset $D = (x_1, x_2, \ldots, x_N)$ of N independent observations.
- ML estimate: choose ***w*** such that data likelihood is maximized:

$$\mathbf{w}_{\mathrm{ML}} = \arg\max_{\mathbf{w}} p(D|\mathbf{w})$$

- MAP estimate: choose most probable ***w*** given the data.

$$\mathbf{w}_{\mathrm{MAP}} = \arg\max_{\mathbf{w}} p(\mathbf{w}|D)$$

Maximum A Posteriori Estimates

- Dataset $D = (x_1, x_2, \ldots, x_N)$ of N independent observations.
- ML estimate: choose $\boldsymbol{w}$ such that data likelihood is maximized:

$$\mathbf{w}_{\mathrm{ML}} = \arg\max_{\mathbf{w}} p(D|\mathbf{w})$$

- MAP estimate: choose most probable $\boldsymbol{w}$ given the data.

$$\mathbf{w}_{\mathrm{MAP}} = \arg\max_{\mathbf{w}} p(\mathbf{w}|D) = \arg\max_{\mathbf{w}} p(D|\mathbf{w})p(\mathbf{w})$$

Curve Fitting: Maximum A Posteriori Estimates

- Dataset $D = \{(x_1, t_1), ..., (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$
- Model: $p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\left[-\frac{\beta}{2}(t - y(x, \mathbf{w}))^2\right]$
- ML estimate: choose ***w*** such that data likelihood is maximized:

$$\mathbf{w}_{\mathrm{ML}} = \arg\max_{\mathbf{w}} p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \arg\min_{\mathbf{w}} - \log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)$$

Curve Fitting: Maximum A Posteriori Estimates

- Dataset $D = \{(x_1, t_1), ..., (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$
- Model: $p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\left[-\frac{\beta}{2}(t - y(x, \mathbf{w}))^2\right]$
- ML estimate: choose ***w*** such that data likelihood is maximized:

$$\mathbf{w}_{\mathrm{ML}} = \arg\max_{\mathbf{w}} p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \arg\min_{\mathbf{w}} - \log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)$$

- MAP estimate: choose most probable ***w*** given the data.

$$\mathbf{w}_{\mathrm{MAP}} = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \beta)$$

Curve Fitting: Maximum A Posteriori Estimates

- Dataset $D = \{(x_1, t_1), ..., (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$
- Model: $p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\left[-\frac{\beta}{2}(t - y(x, \mathbf{w}))^2\right]$
- Given a prior $p(\mathbf{w}|\alpha)$ the posterior distribution is

$$p(\mathbf{w}|\mathbf{t}, \mathbf{x}, \beta, \alpha) =$$

Curve Fitting: Maximum A Posteriori Estimates

- Dataset $D = \{(x_1, t_1), ..., (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$
- Model: $p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\left[-\frac{\beta}{2}(t - y(x, \mathbf{w}))^2\right]$
- Given a prior $p(\mathbf{w}|\alpha)$ the posterior distribution is

$$p(\mathbf{w}|\mathbf{t}, \mathbf{x}, \beta, \alpha) = \frac{p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)}{p(\mathbf{t}|\mathbf{x}, \beta, \alpha)}$$

Curve Fitting: Maximum A Posteriori Estimates

- Dataset $D = \{(x_1, t_1), ..., (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$
- Model: $p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\left[-\frac{\beta}{2}(t - y(x, \mathbf{w}))^2\right]$
- Given a prior $p(\mathbf{w}|\alpha)$ the posterior distribution is

$$p(\mathbf{w}|\mathbf{t}, \mathbf{x}, \beta, \alpha) = \frac{p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)}{p(\mathbf{t}|\mathbf{x}, \beta, \alpha)}$$

- Maximum A Posteriori Estimate:

$$\mathbf{w}_{\mathrm{MAP}} = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{t}, \mathbf{x}, \beta, \alpha) =$$

Curve Fitting: Maximum A Posteriori Estimates

- Dataset $D = \{(x_1, t_1), ..., (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$
- Model: $p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\left[-\frac{\beta}{2}\left(t - y(x, \mathbf{w})\right)^2\right]$
- Given a prior $p(\mathbf{w}|\alpha)$ the posterior distribution is

$$p(\mathbf{w}|\mathbf{t}, \mathbf{x}, \beta, \alpha) = \frac{p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)}{p(\mathbf{t}|\mathbf{x}, \beta, \alpha)}$$

- Maximum A Posteriori Estimate:

$$\mathbf{w}_{\mathrm{MAP}} = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{t}, \mathbf{x}, \beta, \alpha) = \arg\min_{\mathbf{w}} - \log p(\mathbf{w}|\mathbf{t}, \mathbf{x}, \beta, \alpha)$$

Curve Fitting: Maximum A Posteriori Estimates

- Dataset $D = \{(x_1, t_1), ..., (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$
- Model: $p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\left[-\frac{\beta}{2}(t - y(x, \mathbf{w}))^2\right]$
- Given a prior $p(\mathbf{w}|\alpha)$ the posterior distribution is

$$p(\mathbf{w}|\mathbf{t}, \mathbf{x}, \beta, \alpha) = \frac{p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)}{p(\mathbf{t}|\mathbf{x}, \beta, \alpha)}$$

- Maximum A Posteriori Estimate:

$$\mathbf{w}_{\mathrm{MAP}} = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{t}, \mathbf{x}, \beta, \alpha) = \arg\min_{\mathbf{w}} - \log p(\mathbf{w}|\mathbf{t}, \mathbf{x}, \beta, \alpha)$$

$$= \arg\min_{\mathbf{w}} - \log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) - \log p(\mathbf{w}|\alpha)$$

Curve Fitting: Maximum A Posteriori Estimates

- Gaussian prior: $\mathbf{w} \in \mathbb{R}^M$

$$p(\mathbf{w}|\alpha) = \prod_{i=1}^{M} \mathcal{N}(w_i|0, \alpha^{-1}) = (\frac{\alpha}{2\pi})^{M/2} e^{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}}$$

$$\mathbf{w}_{\text{MAP}} = -\arg\min_{\mathbf{w}} \log p(\mathbf{w}|\mathbf{t}, \mathbf{x}, \beta, \alpha) = \arg\min_{\mathbf{w}} -\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) - \log p(\mathbf{w}|\alpha)$$

$$=$$

Curve Fitting: Maximum A Posteriori Estimates

- Gaussian prior: $\mathbf{w} \in \mathbb{R}^M$

$$p(\mathbf{w}|\alpha) = \prod_{i=1}^{M} \mathcal{N}(w_i|0, \alpha^{-1}) = (\frac{\alpha}{2\pi})^{M/2} e^{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}}$$

$$\mathbf{w}_{\text{MAP}} = -\arg\min_{\mathbf{w}} \log p(\mathbf{w}|\mathbf{t}, \mathbf{x}, \beta, \alpha) = \arg\min_{\mathbf{w}} -\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) - \log p(\mathbf{w}|\alpha)$$

$$= \arg\min_{\mathbf{w}} -\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

Curve Fitting: Maximum A Posteriori Estimates

- Gaussian prior: $\mathbf{w} \in \mathbb{R}^M$

$$p(\mathbf{w}|\alpha) = \prod_{i=1}^{M} \mathcal{N}(w_i|0, \alpha^{-1}) = (\frac{\alpha}{2\pi})^{M/2} e^{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}}$$

$$\begin{aligned}\mathbf{w}_{\mathrm{MAP}} &= -\arg\min_{\mathbf{w}} \log p(\mathbf{w}|\mathbf{t}, \mathbf{x}, \beta, \alpha) = \arg\min_{\mathbf{w}} - \log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) - \log p(\mathbf{w}|\alpha) \\ &= \arg\min_{\mathbf{w}} - \log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\end{aligned}$$

- Curve fitting a function with Gaussian noise and Gaussian prior:

$$p(t|x, \mathbf{w}, \beta) = \sqrt{\frac{\beta}{2\pi}} \exp\left[-\frac{\beta}{2}(t - y(x, \mathbf{w}))^2\right]$$

Curve Fitting: Maximum A Posteriori Estimates

- Gaussian prior: $\mathbf{w} \in \mathbb{R}^M$

$$p(\mathbf{w}|\alpha) = \prod_{i=1}^{M} \mathcal{N}(w_i|0, \alpha^{-1}) = \left(\frac{\alpha}{2\pi}\right)^{M/2} e^{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}}$$

$$\mathbf{w}_{\text{MAP}} = -\arg\min_{\mathbf{w}} \log p(\mathbf{w}|\mathbf{t}, \mathbf{x}, \beta, \alpha) = \arg\min_{\mathbf{w}} - \log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) - \log p(\mathbf{w}|\alpha)$$

$$= \arg\min_{\mathbf{w}} - \log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

- Curve fitting a function with Gaussian noise and Gaussian prior:

$$p(t|x, \mathbf{w}, \beta) = \sqrt{\frac{\beta}{2\pi}} \exp\left[-\frac{\beta}{2}\left(t - y(x, \mathbf{w})\right)^2\right]$$

$$\mathbf{w}_{\text{MAP}} = \arg\min_{\mathbf{w}} \quad \frac{\beta}{2}\sum_{i=1}^{N}\Big(y\left(x_i, \mathbf{w}\right) - t_i\Big)^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

Curve Fitting: Maximum A Posteriori Estimates

- Gaussian prior: $\mathbf{w} \in \mathbb{R}^M$

$$p(\mathbf{w}|\alpha) = \prod_{i=1}^{M} \mathcal{N}(w_i|0, \alpha^{-1}) = (\frac{\alpha}{2\pi})^{M/2} e^{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}}$$

$$\mathbf{w}_{\text{MAP}} = -\arg\min_{\mathbf{w}} \log p(\mathbf{w}|\mathbf{t}, \mathbf{x}, \beta, \alpha) = \arg\min_{\mathbf{w}} -\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) - \log p(\mathbf{w}|\alpha)$$

$$= \arg\min_{\mathbf{w}} -\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

- Curve fitting a function with Gaussian noise and Gaussian prior:

$$p(t|x, \mathbf{w}, \beta) = \sqrt{\frac{\beta}{2\pi}} \exp\left[-\frac{\beta}{2}(t - y(x, \mathbf{w}))^2\right]$$

$$\mathbf{w}_{\text{MAP}} = \arg\min_{\mathbf{w}} \quad \frac{\beta}{2}\sum_{i=1}^{N}\Big(y(x_i, \mathbf{w}) - t_i\Big)^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

- Predictive distribution:

$$p(t'|x', \mathbf{w}_{\text{MAP}}, \beta) =$$

$$\mathbb{E}[t'|x', \mathbf{w}_{\text{MAP}}, \beta] =$$

Curve Fitting: Maximum A Posteriori Estimates

- Gaussian prior: $\mathbf{w} \in \mathbb{R}^M$

$$p(\mathbf{w}|\alpha) = \prod_{i=1}^{M} \mathcal{N}(w_i|0, \alpha^{-1}) = (\frac{\alpha}{2\pi})^{M/2} e^{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}}$$

$$\mathbf{w}_{\mathrm{MAP}} = -\arg\min_{\mathbf{w}} \log p(\mathbf{w}|\mathbf{t}, \mathbf{x}, \beta, \alpha) = \arg\min_{\mathbf{w}} -\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) - \log p(\mathbf{w}|\alpha)$$

$$= \arg\min_{\mathbf{w}} -\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

- Curve fitting a function with Gaussian noise and Gaussian prior:

$$p(t|x, \mathbf{w}, \beta) = \sqrt{\frac{\beta}{2\pi}} \exp\left[-\frac{\beta}{2}\left(t - y(x, \mathbf{w})\right)^2\right]$$

$$\mathbf{w}_{\mathrm{MAP}} = \arg\min_{\mathbf{w}} \quad \frac{\beta}{2}\sum_{i=1}^{N}\left(y\left(x_i, \mathbf{w}\right) - t_i\right)^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

- Predictive distribution:

$$p(t'|x', \mathbf{w}_{\mathrm{MAP}}, \beta) = \mathcal{N}(t'|y(x', \mathbf{w}_{MAP}), \beta^{-1})$$

$$\mathbb{E}[t'|x', \mathbf{w}_{\mathrm{MAP}}, \beta] = y(x', \mathbf{w}_{MAP})$$

Overview

1. Probability theory

2. **Statistical learning principles:**

    I. Maximum likelihood

    II. Maximum a posteriori

    III. Bayesian prediction

Bayesian Approach

- Dataset $D = (x_1, x_2, \ldots, x_N)$ of N independent observations.

- Frequentist approach: search for one optimal estimate of ***w***

$$\mathbf{w}_{\mathrm{ML}} =$$

$$\mathbf{w}_{\mathrm{MAP}} =$$

- Bayesian approach: Given a prior belief over **w**, p(**w**), and our data D, we are interested in the posterior distribution

$$p(\mathbf{w} \,|\, D) =$$

- *p(**w** |D)* reflects the plausibility of different ***w,*** given our prior knowledge and how likely our data is generated using **w**.

Bayesian Approach

- Dataset $D = (x_1, x_2, \ldots, x_N)$ of N independent observations.

- Frequentist approach: search for one optimal estimate of $\boldsymbol{w}$

$$\mathbf{w}_{\mathrm{ML}} =$$

$$\mathbf{w}_{\mathrm{MAP}} =$$

- Bayesian approach: Given a prior belief over **w**, p(**w**), and our data D, we are interested in the posterior distribution

$$p(\mathbf{w} \mid D) =$$

- $p(\boldsymbol{w} \mid D)$ reflects the plausibility of different ***w***, given our prior knowledge and how likely our data is generated using **w**.

Bayesian Approach

- Dataset $D = (x_1, x_2, \dots, x_N)$ of N independent observations.

- Frequentist approach: search for one optimal estimate of ***w***

$$\mathbf{w}_{\mathrm{ML}} = \arg\max_{\mathbf{w}} p(D|\mathbf{w})$$

$$\mathbf{w}_{\mathrm{MAP}} = \arg\max_{\mathbf{w}} p(\mathbf{w}|D) = \arg\max_{\mathbf{w}} p(D|\mathbf{w})p(\mathbf{w})$$

- Bayesian approach: Given a prior belief over **w**, p(**w**), and our data D, we are interested in the posterior distribution

$$p(\mathbf{w}\,|\,D) =$$

- *p(**w** |D)* reflects the plausibility of different ***w*,** given our prior knowledge and how likely our data is generated using **w**.

Bayesian Approach

- Dataset $D = (x_1, x_2, \ldots, x_N)$ of N independent observations.

- Frequentist approach: search for one optimal estimate of ***w***

$$\mathbf{w}_{\text{ML}} = \arg\max_{\mathbf{w}} p(D|\mathbf{w})$$

$$\mathbf{w}_{\text{MAP}} = \arg\max_{\mathbf{w}} p(\mathbf{w}|D) = \arg\max_{\mathbf{w}} p(D|\mathbf{w})p(\mathbf{w})$$

- Bayesian approach: Given a prior belief over **w**, p(**w**), and our data D, we are interested in the posterior distribution

$$p(\mathbf{w}\,|\,D) = \frac{p(D|\mathbf{w})p(\mathbf{w})}{p(D)}$$

- *p(**w** |D)* reflects the plausibility of different ***w**,* given our prior knowledge and how likely our data is generated using **w**.

Bayesian Approach

- Dataset $D = (x_1, x_2, \ldots, x_N)$ of N independent observations.

- Frequentist approach: search for one optimal estimate of ***w***

$$\mathbf{w}_{\text{ML}} = \arg\max_{\mathbf{w}} p(D|\mathbf{w})$$

$$\mathbf{w}_{\text{MAP}} = \arg\max_{\mathbf{w}} p(\mathbf{w}|D) = \arg\max_{\mathbf{w}} p(D|\mathbf{w})p(\mathbf{w})$$

- Bayesian approach: Given a prior belief over **w**, p(**w**), and our data D, we are interested in the posterior distribution

$$p(\mathbf{w}\,|\,D) = \frac{p(D|\mathbf{w})p(\mathbf{w})}{\boxed{p(D)}}$$

- $p(\mathbf{w}\,|D)$ reflects the plausibility of different ***w*,** given our prior knowledge and how likely our data is generated using **w**.

Bayesian Approach

- Prior distribution: $p(\mathbf{w})$, should represent some prior knowledge/belief of the plausibility of ***w***.

- After observing data $D = (x_1, x_2, \ldots, x_N)$, posterior distribution

$$p(\mathbf{w} \mid D) = \frac{p(D \mid \mathbf{w}) p(\mathbf{w})}{p(D)}$$

Bayesian Approach

- Prior distribution: $p(\mathbf{w})$, should represent some prior knowledge/belief of the plausibility of ***w***.

- After observing data $D = (x_1, x_2, \ldots, x_N)$, posterior distribution

$$p(\mathbf{w} \,|\, D) = \frac{p(D \,|\, \mathbf{w}) p(\mathbf{w})}{p(D)}$$

- Predictive distribution:

- $$p(x' \,|\, D) = \int \mathrm{d}\mathbf{w}\, p(x', \mathbf{w} \,|\, D) =$$

Bayesian Approach

- Prior distribution: $p(\mathbf{w})$, should represent some prior knowledge/belief of the plausibility of ***w***.

- After observing data $D = (x_1, x_2, \ldots, x_N)$, posterior distribution

$$p(\mathbf{w} \mid D) = \frac{p(D \mid \mathbf{w})p(\mathbf{w})}{p(D)}$$

- Predictive distribution:

- $$p(x' \mid D) = \int \mathrm{d}\mathbf{w} p(x', \mathbf{w} \mid D) = \int d\mathbf{w} p(x' | \mathbf{w}) p(\mathbf{w} | D)$$

Bayesian Approach

- Prior distribution: $p(\mathbf{w})$, should represent some prior knowledge/belief of the plausibility of ***w***.

- After observing data $D = (x_1, x_2, \ldots, x_N)$, posterior distribution

$$p(\mathbf{w} \mid D) = \frac{p(D \mid \mathbf{w}) p(\mathbf{w})}{p(D)}$$

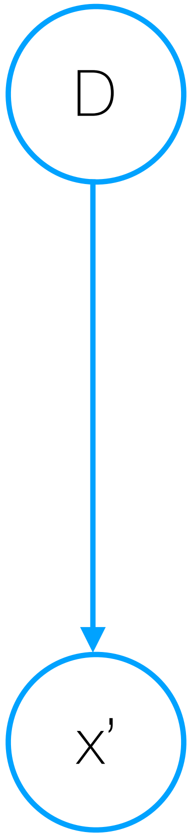
- Predictive distribution:

$$p(x' \mid D) = \int \mathrm{d}\mathbf{w}\, p(x', \mathbf{w} \mid D) = \int \mathrm{d}\mathbf{w}\, p(x' | \mathbf{w}) p(\mathbf{w} | D)$$

- **Note**: even if $p(D \mid \mathbf{w}) = \prod_{i=1}^{N} p(x_i \mid \mathbf{w})$

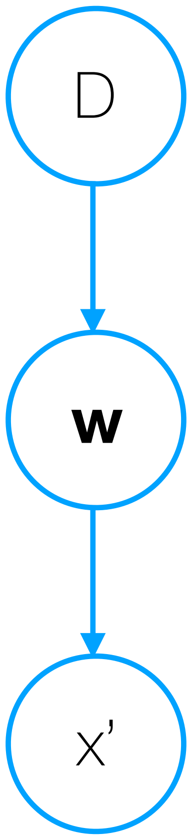
$$p(D) = \int \mathrm{d}\mathbf{w}\, p(D, \mathbf{w}) = \int \mathrm{d}\mathbf{w}\, p(D \mid \mathbf{w}) p(\mathbf{w}) \neq \prod_{i=1}^{N} p(x_i)$$

Curve Fitting: Bayesian Approach



$$p(x' \mid D)$$

Curve Fitting: Bayesian Approach



$$p(x'|D) = \int d\mathbf{w}\, p(x'|\mathbf{w})p(\mathbf{w}|D)$$

Curve Fitting: Bayesian Approach

- Dataset $D = \{(x_1, t_1), ..., (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$
- Posterior distribution after observing data:

Curve Fitting: Bayesian Approach

- Dataset $D = \{(x_1, t_1), ..., (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$

- Posterior distribution after observing data:

$$p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{x}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{t}|\mathbf{x})} \quad \text{with} \quad p(\mathbf{t}|\mathbf{x}) = \int p(\mathbf{t}|\mathbf{x}, \mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w}$$

Curve Fitting: Bayesian Approach

- Dataset $D = \{(x_1, t_1), ..., (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$

- Posterior distribution after observing data:

$$p(\mathbf{w} \,|\, \mathbf{x}, \mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{x}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{t}|\mathbf{x})} \quad \text{with} \quad p(\mathbf{t}|\mathbf{x}) = \int p(\mathbf{t}|\mathbf{x}, \mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w}$$

Curve Fitting: Bayesian Approach

- Dataset $D = \{(x_1, t_1), ..., (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$

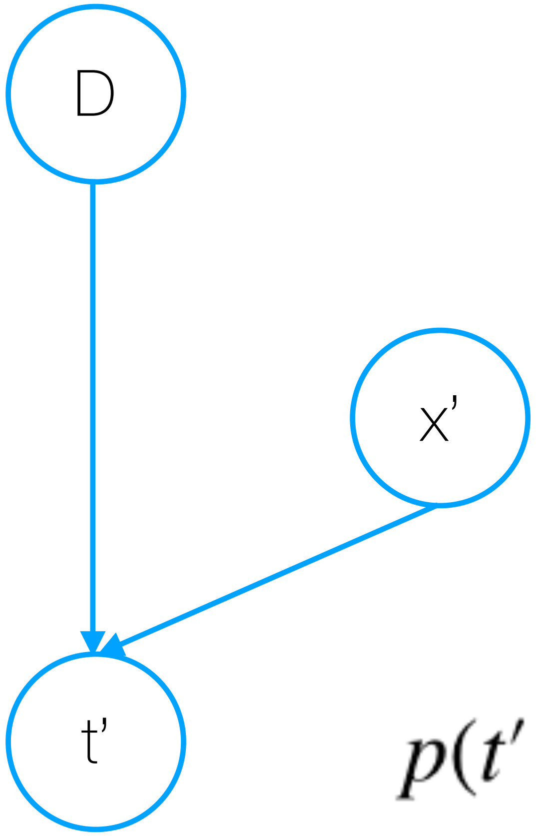
- Posterior distribution after observing data:

$$p(\mathbf{w} \,|\, \mathbf{x}, \mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{x}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{t}|\mathbf{x})} \quad \text{with} \quad p(\mathbf{t}|\mathbf{x}) = \int p(\mathbf{t}|\mathbf{x}, \mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w}$$

- Predictive distribution:

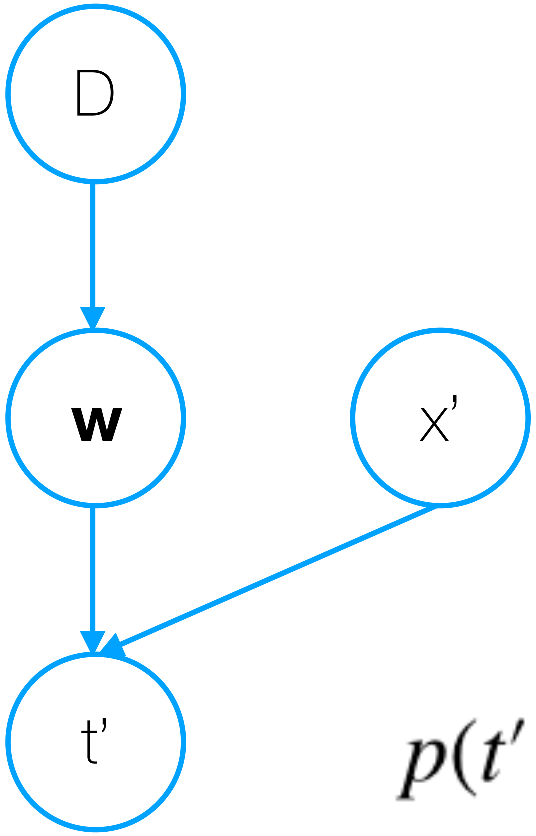
$$p(t' \,|\, x', \mathbf{x}, \mathbf{t}) = \int \mathrm{d}\mathbf{w} p(t', \mathbf{w} \,|\, x', \mathbf{x}, \mathbf{t}) = \int p(t'|x', \mathbf{w})p(\mathbf{w}|\mathbf{x}, \mathbf{t})\mathrm{d}\mathbf{w}$$

Curve Fitting: Bayesian Approach



$$p(t' \mid x', \mathbf{x}, \mathbf{t})$$

Curve Fitting: Bayesian Approach



$$p(t' \mid x', \mathbf{x}, \mathbf{t}) = \int p(t'|x', \mathbf{w})p(\mathbf{w}|\mathbf{x}, \mathbf{t})\mathrm{d}\mathbf{w}$$

Curve Fitting: Bayesian Approach

- Predictive distribution: $p(t'|x', \mathbf{x}, \mathbf{t}) = \int p(t'|x', \mathbf{w})p(\mathbf{w}|\mathbf{x}, \mathbf{t})\mathrm{d}\mathbf{w}$

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{x}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{t}|\mathbf{x})} \quad \text{with} \quad p(\mathbf{t}|\mathbf{x}) = \int p(\mathbf{t}|\mathbf{x}, \mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w}$$

Advantages:

- Inclusion of prior knowledge
- Represents uncertainty in ***t'*** both due to target noise, and uncertainty over ***w***.

Disadvantages:

- Posterior is hard to compute analytically → approximate!
- Prior is often chosen for mathematical convenience, not reflection of prior belief!