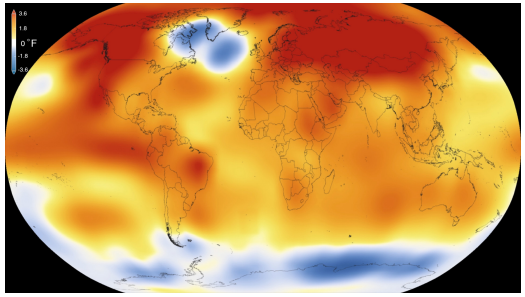
3F8: Introduction to inference

Rich Turner and José Miguel Hernández-Lobato
{ret26, jmh233}@cam.ac.uk

$$p(\mathbf{X}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})}{p(\mathbf{Y})}$$

Examples of inference problems

climate science



inputs
solar
human

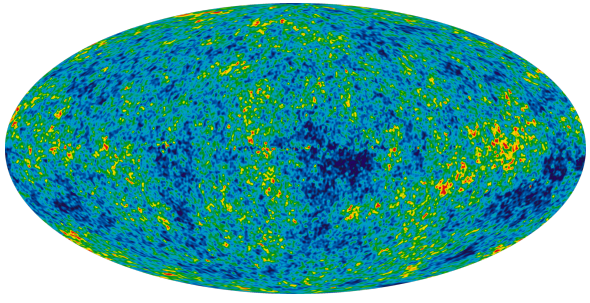
→

physics
ocean
atmosphere
land

→

observations
temperatures
sea and ice levels
CO2

astronomy



physical constants → **physics** → **observations**
CMB
supernovae

object recognition

Predicted Tags

architecture travel
no person river
building outdoors
water castle
tourism city

www.clarifai.com

speech recognition

The quick bro...

automated drug discovery
genomics
collaborative filtering
error correcting codes
...

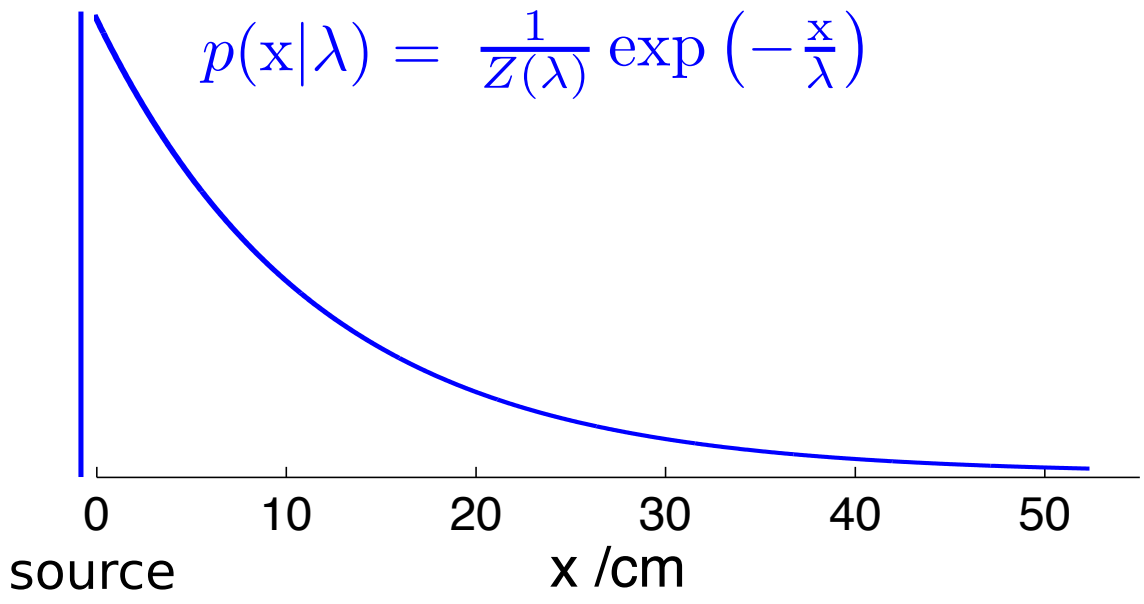
Definition of an inference problem

A problem where you have to **estimate unknown variables** from **known variables**.

known variables are sometimes called **observed variables**
unknown variables are sometimes called **unobserved variables**

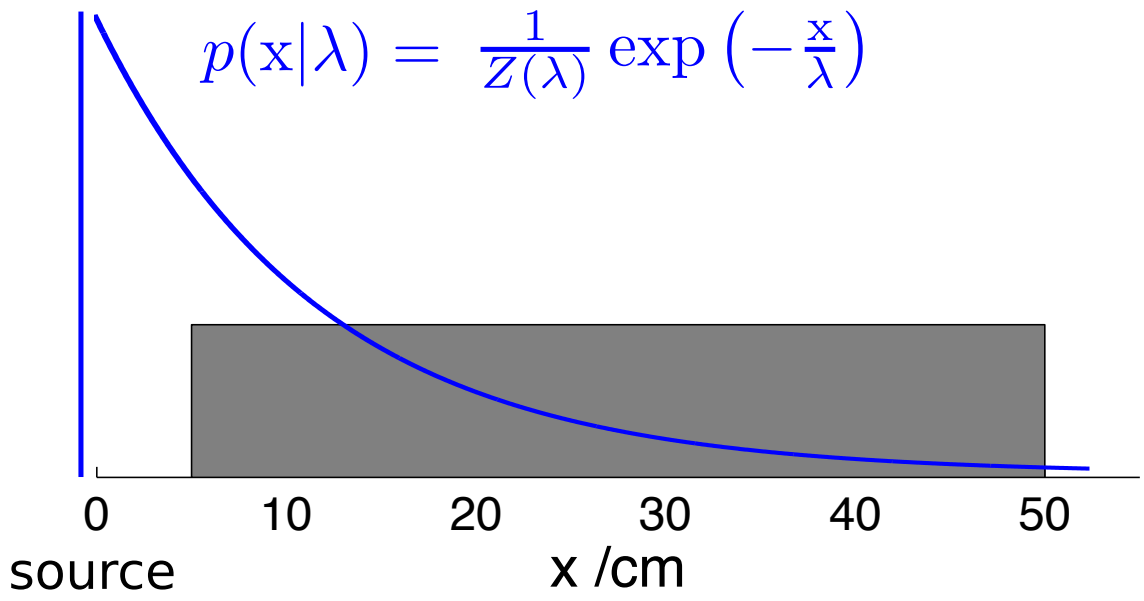
A first inference problem: radioactive decay

- unstable particles emitted from a source, decay at a distance x
- x follows an exponential distribution with characteristic length-scale λ



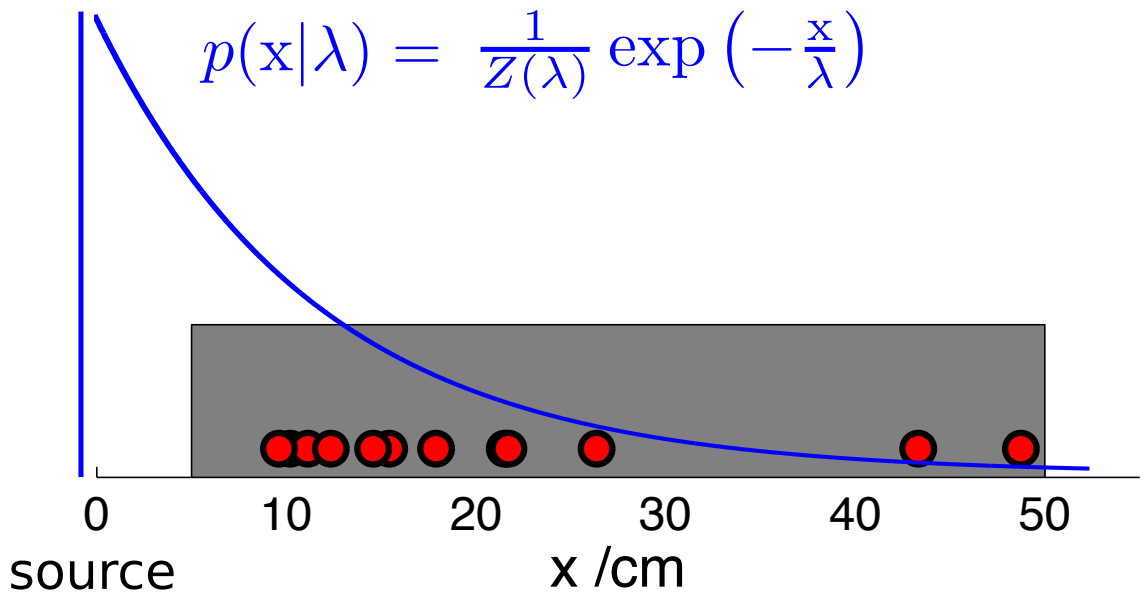
A first inference problem: radioactive decay

- unstable particles emitted from a source, decay at a distance x
- x follows an exponential distribution with characteristic length-scale λ
- decay events can only be observed in a window 5cm from source to 50cm



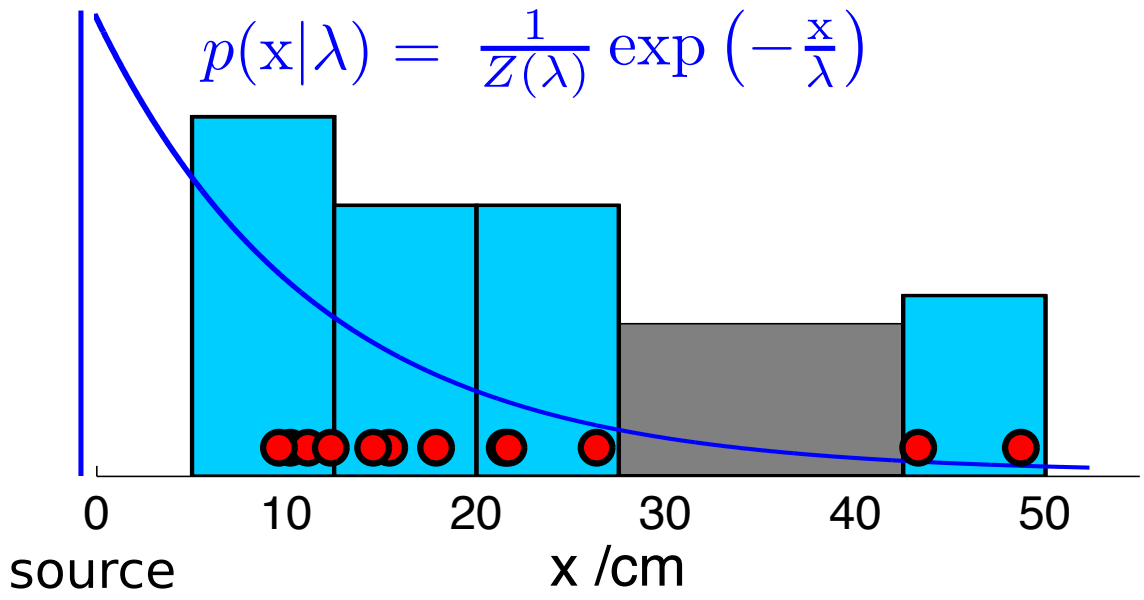
A first inference problem: radioactive decay

- unstable particles emitted from a source, decay at a distance x
- x follows an exponential distribution with characteristic length-scale λ
- decay events can only be observed in a window 5cm from source to 50cm
- N events observed at $\{\mathrm{x}_1, \ldots \mathrm{x}_N\}$. What is λ?



A first inference problem: radioactive decay

- unstable particles emitted from a source, decay at a distance x
- x follows an exponential distribution with characteristic length-scale λ
- decay events can only be observed in a window 5cm from source to 50cm
- N events observed at $\{\mathrm{x}_1, \ldots \mathrm{x}_N\}$. What is λ?



Ad hoc classes of approaches

Approach 1

- bin up into a histogram
  - where do we place the bins
- fit to density
  - what error measure do we minimise?

Approach 2

- construct an estimator e.g. the sample mean $\mu = \frac{1}{N} \sum_{n=1}^{N} \mathrm{x}_n$
  - which estimator should we choose? mean, variance, higher moments?
- relate to parameters via expectation of estimator e.g. $\mu \approx \langle x \rangle = f(\lambda)$
  - small sample effects can be problematic e.g. if $\mu > \frac{1}{2}(50 + 5)$cm

A principled method: the probabilistic approach

A principled method: the probabilistic approach

1. "right answer" to any inference problem is a probability distribution

plausibility of
unobserved variables
given observed variables

$$p(\mathrm{u}|\mathrm{o})$$

A principled method: the probabilistic approach

1. "right answer" to any inference problem is a probability distribution

plausibility of unobserved variables given observed variables

$$p(\mathrm{u}|\mathrm{o}) = p(\lambda|\{x_n\}_{n=1}^N)$$

unobserved variables $\mathrm{u} = \lambda$

observed variables $\mathrm{o} = \{x_n\}_{n=1}^N$

A principled method: the probabilistic approach

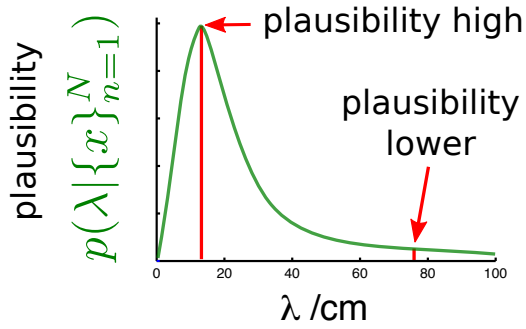
1. "right answer" to any inference problem is a probability distribution

plausibility of unobserved variables given observed variables

$$p(\mathrm{u}|\mathrm{o}) = p(\lambda|\{x_n\}_{n=1}^N)$$

unobserved variables $\mathrm{u} = \lambda$

observed variables $\mathrm{o} = \{x_n\}_{n=1}^N$



A principled method: the probabilistic approach

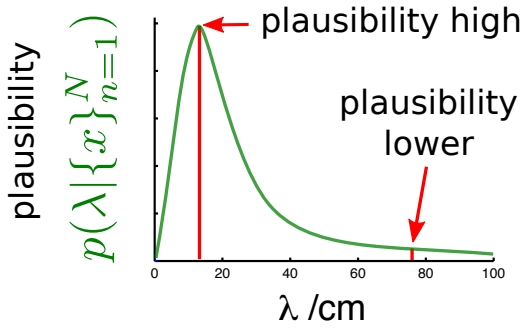
1. "right answer" to any inference problem is a probability distribution

plausibility of unobserved variables given observed variables

$$p(\mathrm{u}|\mathrm{o}) = p(\lambda|\{x_n\}_{n=1}^N)$$

unobserved variables $\mathrm{u} = \lambda$

observed variables $\mathrm{o} = \{x_n\}_{n=1}^N$



2. plausibility computed using the sum and product rules of probability

sum rule: $$p(y) = \int p(y,z)\mathrm{d}z$$

product rule: $$p(y,z) = p(z)p(y|z) = p(y)p(z|y)$$

A principled method: the probabilistic approach

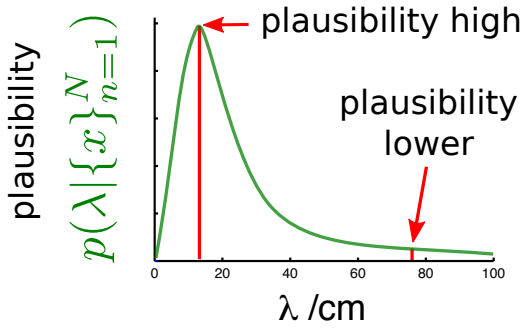
1. "right answer" to any inference problem is a probability distribution

plausibility of unobserved variables given observed variables

$$p(\mathrm{u}|\mathrm{o}) = p(\lambda|\{x_n\}_{n=1}^N)$$

unobserved variables $\mathrm{u} = \lambda$

observed variables $\mathrm{o} = \{x_n\}_{n=1}^N$



2. plausibility computed using the sum and product rules of probability

sum rule: $$p(y) = \int p(y,z)\mathrm{d}z$$

product rule: $$p(y,z) = p(z)p(y|z) = p(y)p(z|y)$$

Cox (1946) showed:
- only way to perform consistent inferences
- generalisation of logic to uncertain situations
- see supplemental slides

A principled method: the probabilistic approach

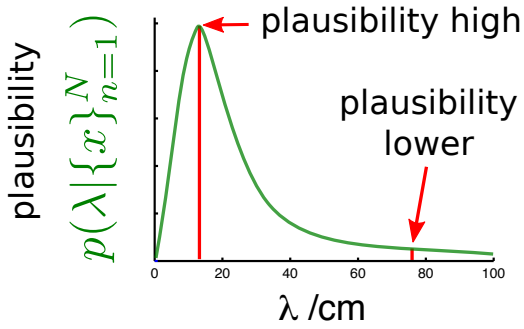
1. "right answer" to any inference problem is a probability distribution

plausibility of unobserved variables given observed variables

$$p(\mathrm{u}|\mathrm{o}) = p(\lambda|\{x_n\}_{n=1}^N)$$

unobserved variables $\mathrm{u} = \lambda$

observed variables $\mathrm{o} = \{x_n\}_{n=1}^N$



2. plausibility computed using the sum and product rules of probability

sum rule: $$p(y) = \int p(y,z)\mathrm{d}z$$

product rule: $$p(y,z) = p(z)p(y|z) = p(y)p(z|y)$$

↓

Bayes' rule: $$p(y|z) = \frac{p(y)p(z|y)}{p(z)}$$

Cox (1946) showed:
- only way to perform consistent inferences
- generalisation of logic to uncertain situations
- see supplemental slides

A principled method: the probabilistic approach

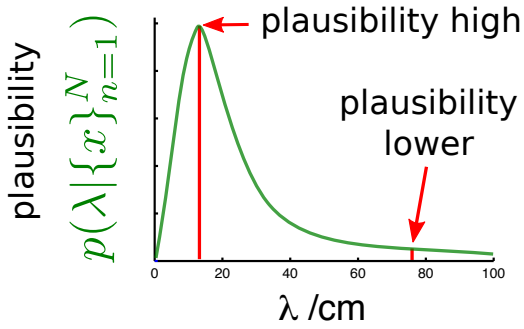
1. "right answer" to any inference problem is a probability distribution

plausibility of unobserved variables given observed variables

$$p(\mathrm{u}|\mathrm{o}) = p(\lambda|\{x_n\}_{n=1}^N)$$

unobserved variables $\mathrm{u} = \lambda$

observed variables $\mathrm{o} = \{x_n\}_{n=1}^N$



2. plausibility computed using the sum and product rules of probability

sum rule: $$p(y) = \int p(y,z)\mathrm{d}z$$

product rule: $$p(y,z) = p(z)p(y|z) = p(y)p(z|y)$$

↓

Bayes' rule: $$p(y|z) = \frac{p(y)p(z|y)}{p(z)}$$

Cox (1946) showed:
- only way to perform consistent inferences
- generalisation of logic to uncertain situations
- see supplemental slides

Apply to radioactive decay example

$$p(\lambda|\{x_n\}_{n=1}^N)$$

A principled method: the probabilistic approach

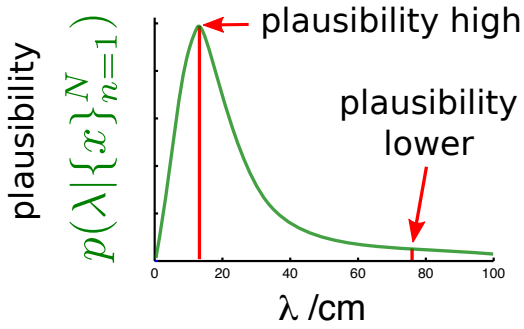
1. "right answer" to any inference problem is a probability distribution

plausibility of unobserved variables given observed variables

$$p(\mathrm{u}|\mathrm{o}) = p(\lambda|\{x_n\}_{n=1}^N)$$

unobserved variables $\mathrm{u} = \lambda$

observed variables $\mathrm{o} = \{x_n\}_{n=1}^N$



2. plausibility computed using the sum and product rules of probability

sum rule: $$p(y) = \int p(y,z)\mathrm{d}z$$

product rule: $$p(y,z) = p(z)p(y|z) = p(y)p(z|y)$$

Bayes' rule: $$p(y|z) = \frac{p(y)p(z|y)}{p(z)}$$

Cox (1946) showed:
- only way to perform consistent inferences
- generalisation of logic to uncertain situations
- see supplemental slides

$y = \lambda$

$z = \{x_n\}_{n=1}^N$

Apply to radioactive decay example

$$p(\lambda|\{x_n\}_{n=1}^N)$$

A principled method: the probabilistic approach

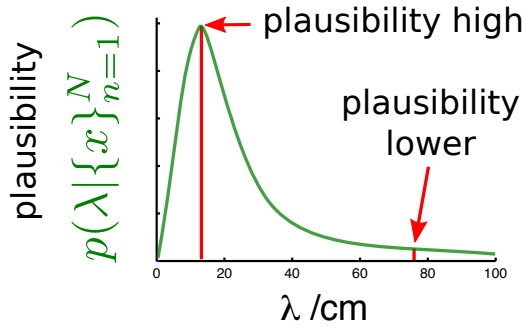
1. "right answer" to any inference problem is a probability distribution

plausibility of unobserved variables given observed variables

$$p(\mathrm{u}|\mathrm{o}) = p(\lambda|\{x_n\}_{n=1}^N)$$

unobserved variables $\mathrm{u} = \lambda$

observed variables $\mathrm{o} = \{x_n\}_{n=1}^N$



2. plausibility computed using the sum and product rules of probability

sum rule: $$p(y) = \int p(y,z)\mathrm{d}z$$

product rule: $$p(y,z) = p(z)p(y|z) = p(y)p(z|y)$$

Bayes' rule: $$p(y|z) = \frac{p(y)p(z|y)}{p(z)}$$

Cox (1946) showed:
- only way to perform consistent inferences
- generalisation of logic to uncertain situations
- see supplemental slides

Apply to radioactive decay example

$y = \lambda$

$z = \{x_n\}_{n=1}^N$

$$p(\lambda|\{x_n\}_{n=1}^N) = \frac{1}{p(\{x_n\}_{n=1}^N)} \quad p(\lambda) \quad p(\{x_n\}_{n=1}^N|\lambda)$$

A principled method: the probabilistic approach

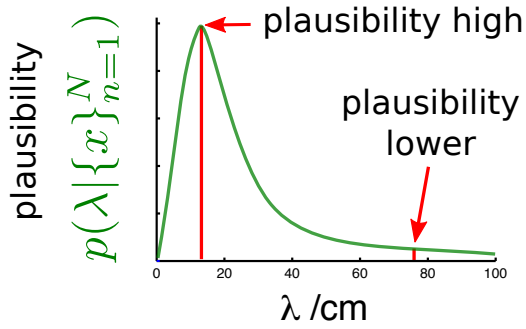
1. "right answer" to any inference problem is a probability distribution

plausibility of unobserved variables given observed variables

$$p(\mathrm{u}|\mathrm{o}) = p(\lambda|\{x_n\}_{n=1}^N)$$

unobserved variables $\mathrm{u} = \lambda$

observed variables $\mathrm{o} = \{x_n\}_{n=1}^N$



2. plausibility computed using the sum and product rules of probability

sum rule: $$p(y) = \int p(y,z)\mathrm{d}z$$

product rule: $$p(y,z) = p(z)p(y|z) = p(y)p(z|y)$$

Bayes' rule: $$p(y|z) = \frac{p(y)p(z|y)}{p(z)}$$

Cox (1946) showed:
- only way to perform consistent inferences
- generalisation of logic to uncertain situations
- see supplemental slides

Apply to radioactive decay example

$y = \lambda$

$z = \{x_n\}_{n=1}^N$

$$p(\lambda|\{x_n\}_{n=1}^N) = \frac{1}{p(\{x_n\}_{n=1}^N)} \; p(\lambda) \; p(\{x_n\}_{n=1}^N|\lambda)$$

plausibility after observing data (**posterior**) $\propto$ what we knew before hand (**prior**) x what the data tell us (**likelihood of parameters**)

A principled method: the probabilistic approach

$$p(\lambda|\{x_n\}_{n=1}^N) = \frac{1}{p(\{x_n\}_{n=1}^N)} \quad p(\lambda) \quad p(\{x_n\}_{n=1}^N|\lambda)$$

plausibility after observing data (**posterior**) $\propto$ what we knew before hand (**prior**) x what the data tell us (**likelihood of parameters**)

A principled method: the probabilistic approach

$$p(\lambda|\{x_n\}_{n=1}^N) = \frac{1}{p(\{x_n\}_{n=1}^N)} \quad p(\lambda) \quad p(\{x_n\}_{n=1}^N|\lambda)$$

plausibility after observing data (**posterior**) $\propto$ what we knew before hand (**prior**) x what the data tell us (**likelihood of parameters**)

Probability of decay events given decay constant $p(\{x_n\}_{n=1}^N|\lambda)$
(likelihood of the decay constant / what the data tell us)

A principled method: the probabilistic approach

$$p(\lambda|\{x_n\}_{n=1}^N) = \frac{1}{p(\{x_n\}_{n=1}^N)} \quad p(\lambda) \quad p(\{x_n\}_{n=1}^N|\lambda)$$

plausibility after observing data (**posterior**) $\propto$ what we knew before hand (**prior**) x what the data tell us (**likelihood of parameters**)

Probability of decay events given decay constant $p(\{x_n\}_{n=1}^N|\lambda)$
(likelihood of the decay constant / what the data tell us)

- decay events independent given decay const.

$$p(\{x_n\}_{n=1}^N|\lambda) = \prod_{n=1}^{N} p(x_n|\lambda)$$

A principled method: the probabilistic approach

$$p(\lambda|\{x_n\}_{n=1}^N) = \frac{1}{p(\{x_n\}_{n=1}^N)} \quad p(\lambda) \quad p(\{x_n\}_{n=1}^N|\lambda)$$

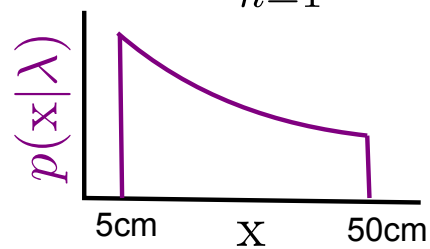
plausibility after observing data (**posterior**) $\propto$ what we knew before hand (**prior**) x what the data tell us (**likelihood of parameters**)

Probability of decay events given decay constant $p(\{x_n\}_{n=1}^N|\lambda)$
(likelihood of the decay constant / what the data tell us)

- decay events independent given decay const.
- each event follows a truncated exponential distribution

$$p(\{x_n\}_{n=1}^N|\lambda) = \prod_{n=1}^{N} p(x_n|\lambda)$$

$$p(x_n|\lambda) = \begin{cases} \frac{1}{Z(\lambda)} \exp(-x_n/\lambda) & \text{if } 5\text{cm} < x_n < 50\text{cm} \\ 0 & \text{otherwise} \end{cases}$$



A principled method: the probabilistic approach

$$p(\lambda|\{x_n\}_{n=1}^N) = \frac{1}{p(\{x_n\}_{n=1}^N)} \quad p(\lambda) \quad p(\{x_n\}_{n=1}^N|\lambda)$$

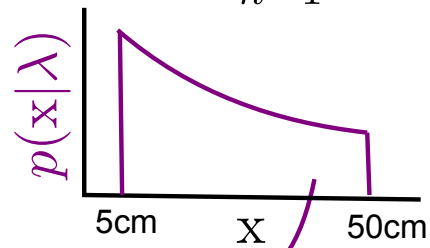
plausibility after observing data (**posterior**) $\propto$ what we knew before hand (**prior**) x what the data tell us (**likelihood of parameters**)

Probability of decay events given decay constant $p(\{x_n\}_{n=1}^N|\lambda)$
(likelihood of the decay constant / what the data tell us)

- decay events independent given decay const.
- each event follows a truncated exponential distribution

$$p(\{x_n\}_{n=1}^N|\lambda) = \prod_{n=1}^{N} p(x_n|\lambda)$$

$$p(x_n|\lambda) = \begin{cases} \frac{1}{Z(\lambda)} \exp(-x_n/\lambda) & \text{if } 5\text{cm} < x_n < 50\text{cm} \\ 0 & \text{otherwise} \end{cases}$$



$$Z(\lambda) = \int_{5\text{cm}}^{50\text{cm}} \exp(x/\lambda)\mathrm{d}x = \lambda[\exp(-5\text{cm}/\lambda) - \exp(-50\text{cm}/\lambda)]$$

area must sum to one

A principled method: the probabilistic approach

$$p(\lambda|\{x_n\}_{n=1}^N) = \frac{1}{p(\{x_n\}_{n=1}^N)} \quad p(\lambda) \quad p(\{x_n\}_{n=1}^N|\lambda)$$

plausibility after observing data (**posterior**) $\propto$ what we knew before hand (**prior**) x what the data tell us (**likelihood of parameters**)

Prior on decay constant (what we knew before seeing data) $p(\lambda)$

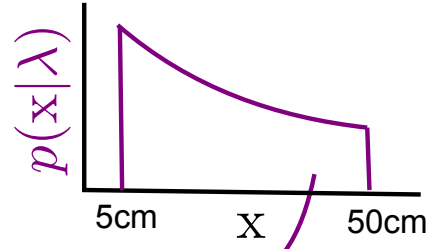
- subjective (depends on your knowledge)

Probability of decay events given decay constant $p(\{x_n\}_{n=1}^N|\lambda)$
(likelihood of the decay constant / what the data tell us)

- decay events independent given decay const.
- each event follows a truncated exponential distribution

$$p(\{x_n\}_{n=1}^N|\lambda) = \prod_{n=1}^{N} p(x_n|\lambda)$$

$$p(x_n|\lambda) = \begin{cases} \frac{1}{Z(\lambda)} \exp(-x_n/\lambda) & \text{if } 5\text{cm} < x_n < 50\text{cm} \\ 0 & \text{otherwise} \end{cases}$$



$$Z(\lambda) = \int_{5\text{cm}}^{50\text{cm}} \exp(x/\lambda)\mathrm{d}x = \lambda[\exp(-5\text{cm}/\lambda) - \exp(-50\text{cm}/\lambda)]$$

area must sum to one

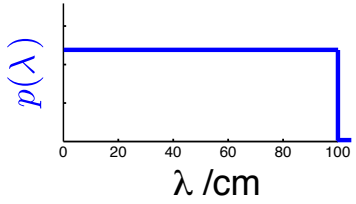
A principled method: the probabilistic approach

$$p(\lambda|\{x_n\}_{n=1}^N) = \frac{1}{p(\{x_n\}_{n=1}^N)} \quad p(\lambda) \quad p(\{x_n\}_{n=1}^N|\lambda)$$

plausibility after observing data (**posterior**) $\propto$ what we knew before hand (**prior**) x what the data tell us (**likelihood of parameters**)

Prior on decay constant (what we knew before seeing data) $p(\lambda)$

- subjective (depends on your knowledge)
- assume uniform prior between 0 cm to 100 cm here for simplicity



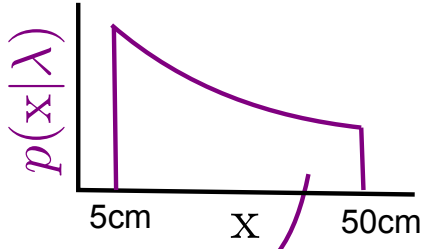
$$p(\lambda) = \begin{cases} \frac{1}{100\text{cm}} & \text{if } \lambda \leq 100\text{cm} \\ 0 & \text{otherwise} \end{cases}$$

Probability of decay events given decay constant $p(\{x_n\}_{n=1}^N|\lambda)$ (likelihood of the decay constant / what the data tell us)

- decay events independent given decay const.
- each event follows a truncated exponential distribution

$$p(\{x_n\}_{n=1}^N|\lambda) = \prod_{n=1}^N p(x_n|\lambda)$$

$$p(x_n|\lambda) = \begin{cases} \frac{1}{Z(\lambda)} \exp(-x_n/\lambda) & \text{if } 5\text{cm} < x_n < 50\text{cm} \\ 0 & \text{otherwise} \end{cases}$$



$$Z(\lambda) = \int_{5\text{cm}}^{50\text{cm}} \exp(x/\lambda)\mathrm{d}x = \lambda[\exp(-5\text{cm}/\lambda) - \exp(-50\text{cm}/\lambda)]$$

area must sum to one

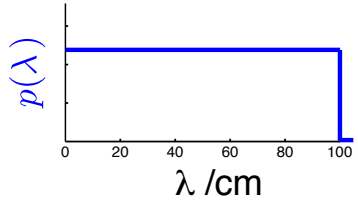
A principled method: the probabilistic approach

$$p(\lambda|\{x_n\}_{n=1}^N) = \frac{1}{p(\{x_n\}_{n=1}^N)} \quad p(\lambda) \quad p(\{x_n\}_{n=1}^N|\lambda)$$

$$\propto p(\lambda) \; \frac{1}{Z(\lambda)^N} \exp\left(-\frac{1}{\lambda}\sum_{n=1}^{N} \mathrm{x}_n\right)$$

Prior on decay constant (what we knew before seeing data) $p(\lambda)$

- subjective (depends on your knowledge)
- assume uniform prior between 0 cm to 100 cm here for simplicity



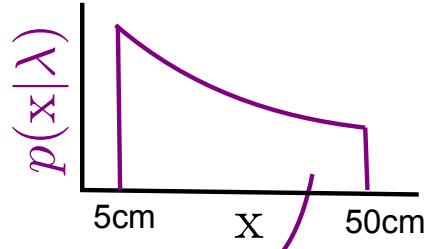
$$p(\lambda) = \begin{cases} \frac{1}{100\mathrm{cm}} & \text{if } \lambda \leq 100\mathrm{cm} \\ 0 & \text{otherwise} \end{cases}$$

Probability of decay events given decay constant $p(\{x_n\}_{n=1}^N|\lambda)$
(likelihood of the decay constant / what the data tell us)

- decay events independent given decay const.
- each event follows a truncated exponential distribution

$$p(\{x_n\}_{n=1}^N|\lambda) = \prod_{n=1}^{N} p(x_n|\lambda)$$

$$p(x_n|\lambda) = \begin{cases} \frac{1}{Z(\lambda)} \exp(-x_n/\lambda) & \text{if } 5\mathrm{cm} < x_n < 50\mathrm{cm} \\ 0 & \text{otherwise} \end{cases}$$



$$Z(\lambda) = \int_{5\mathrm{cm}}^{50\mathrm{cm}} \exp(x/\lambda)\mathrm{d}x = \lambda[\exp(-5\mathrm{cm}/\lambda) - \exp(-50\mathrm{cm}/\lambda)]$$

area must sum to one

A principled method: the probabilistic approach

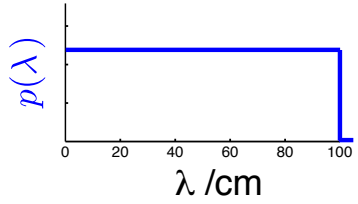
$$p(\lambda|\{x_n\}_{n=1}^N) = \frac{1}{p(\{x_n\}_{n=1}^N)} \quad p(\lambda) \quad p(\{x_n\}_{n=1}^N|\lambda)$$

$$\propto p(\lambda) \; \frac{1}{Z(\lambda)^N} \exp\left(-\frac{1}{\lambda}\sum_{n=1}^{N} \mathrm{x}_n\right)$$

Question: do we need to retain entire dataset to compute posterior?

Prior on decay constant (what we knew before seeing data) $p(\lambda)$

- subjective (depends on your knowledge)
- assume uniform prior between 0 cm to 100 cm here for simplicity



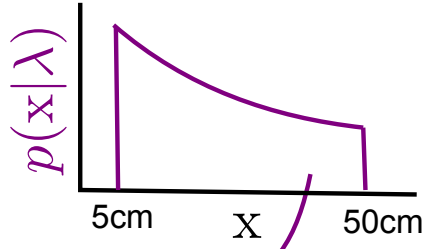
$$p(\lambda) = \begin{cases} \frac{1}{100\text{cm}} & \text{if } \lambda \leq 100\text{cm} \\ 0 & \text{otherwise} \end{cases}$$

Probability of decay events given decay constant $p(\{x_n\}_{n=1}^N|\lambda)$ (likelihood of the decay constant / what the data tell us)

- decay events independent given decay const.
- each event follows a truncated exponential distribution

$$p(\{x_n\}_{n=1}^N|\lambda) = \prod_{n=1}^{N} p(x_n|\lambda)$$

$$p(x_n|\lambda) = \begin{cases} \frac{1}{Z(\lambda)} \exp(-x_n/\lambda) & \text{if } 5\text{cm} < x_n < 50\text{cm} \\ 0 & \text{otherwise} \end{cases}$$



$$Z(\lambda) = \int_{5\text{cm}}^{50\text{cm}} \exp(x/\lambda)\mathrm{d}x = \lambda[\exp(-5\text{cm}/\lambda) - \exp(-50\text{cm}/\lambda)]$$

area must sum to one

A principled method: the probabilistic approach

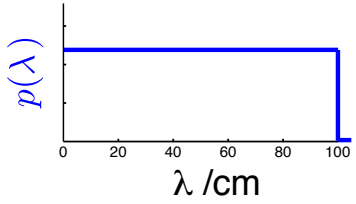
$$p(\lambda|\{x_n\}_{n=1}^N) = \frac{1}{p(\{x_n\}_{n=1}^N)} \quad p(\lambda) \quad p(\{x_n\}_{n=1}^N|\lambda)$$

$$\propto p(\lambda) \frac{1}{Z(\lambda)^N} \exp\left(-\frac{1}{\lambda}\sum_{n=1}^N \mathrm{x}_n\right)$$

Question: do we need to retain entire dataset to compute posterior?
No: only mean(x) & N

Prior on decay constant (what we knew before seeing data) $p(\lambda)$

- subjective (depends on your knowledge)
- assume uniform prior between 0 cm to 100 cm here for simplicity



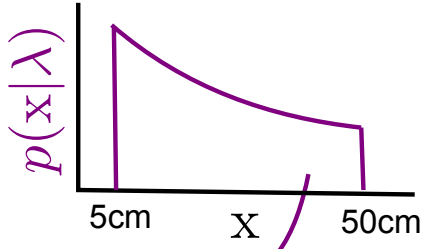
$$p(\lambda) = \begin{cases} \frac{1}{100\mathrm{cm}} & \text{if } \lambda \leq 100\mathrm{cm} \\ 0 & \text{otherwise} \end{cases}$$

Probability of decay events given decay constant $p(\{x_n\}_{n=1}^N|\lambda)$ (likelihood of the decay constant / what the data tell us)

- decay events independent given decay const.
- each event follows a truncated exponential distribution

$$p(\{x_n\}_{n=1}^N|\lambda) = \prod_{n=1}^N p(x_n|\lambda)$$

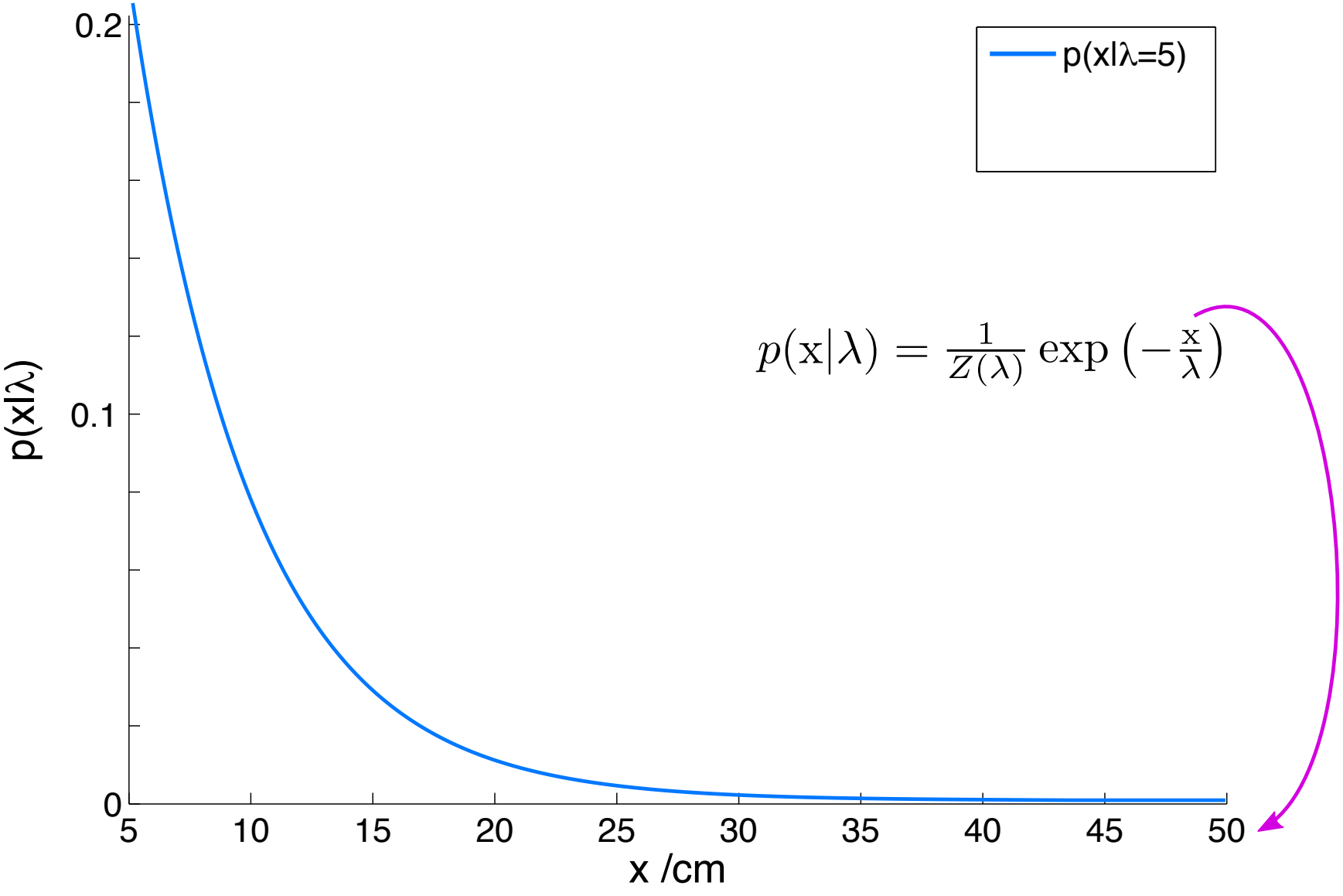
$$p(x_n|\lambda) = \begin{cases} \frac{1}{Z(\lambda)} \exp(-x_n/\lambda) & \text{if } 5\mathrm{cm} < x_n < 50\mathrm{cm} \\ 0 & \text{otherwise} \end{cases}$$



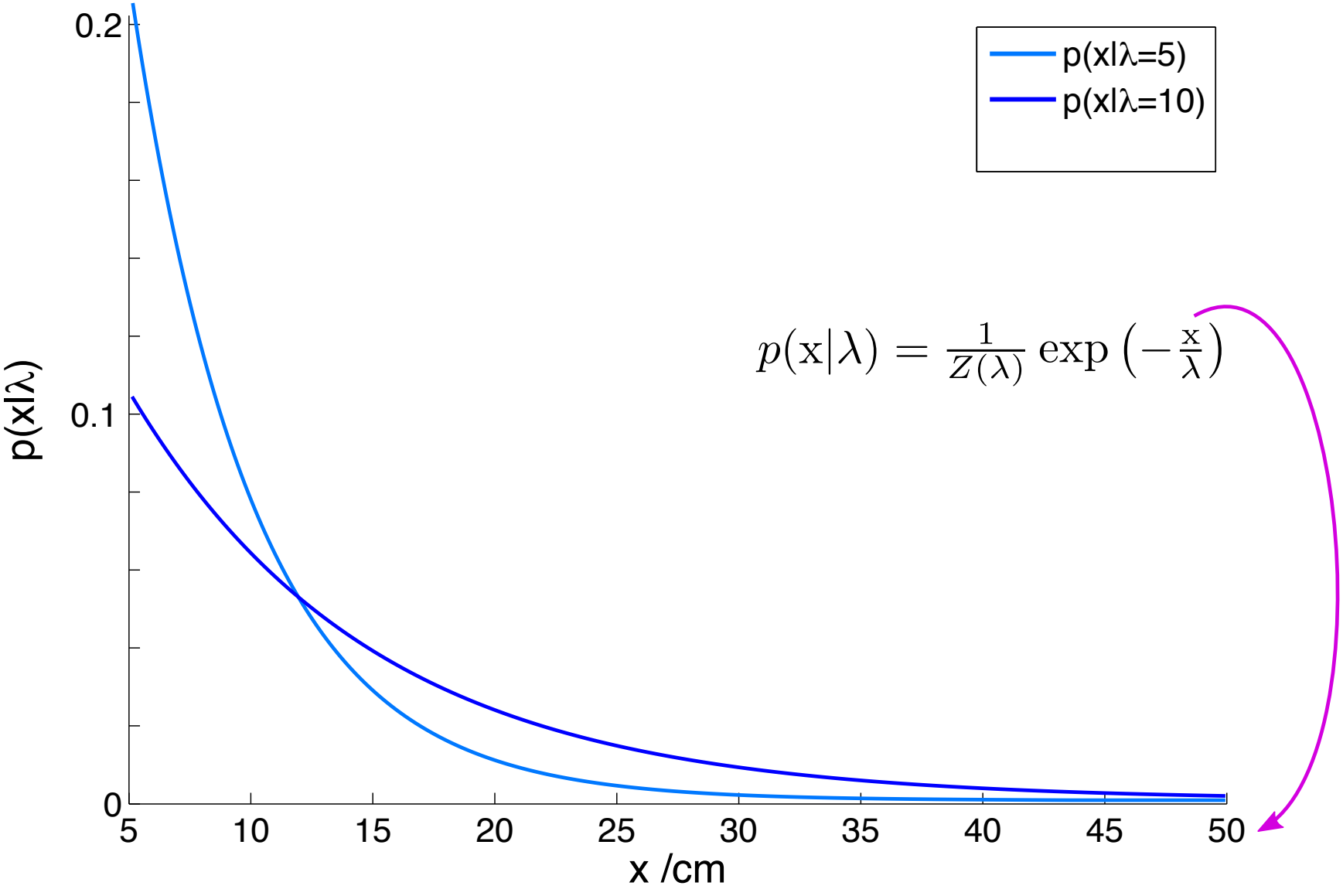
$$Z(\lambda) = \int_{5\mathrm{cm}}^{50\mathrm{cm}} \exp(x/\lambda)\mathrm{d}x = \lambda[\exp(-5\mathrm{cm}/\lambda) - \exp(-50\mathrm{cm}/\lambda)]$$

area must sum to one

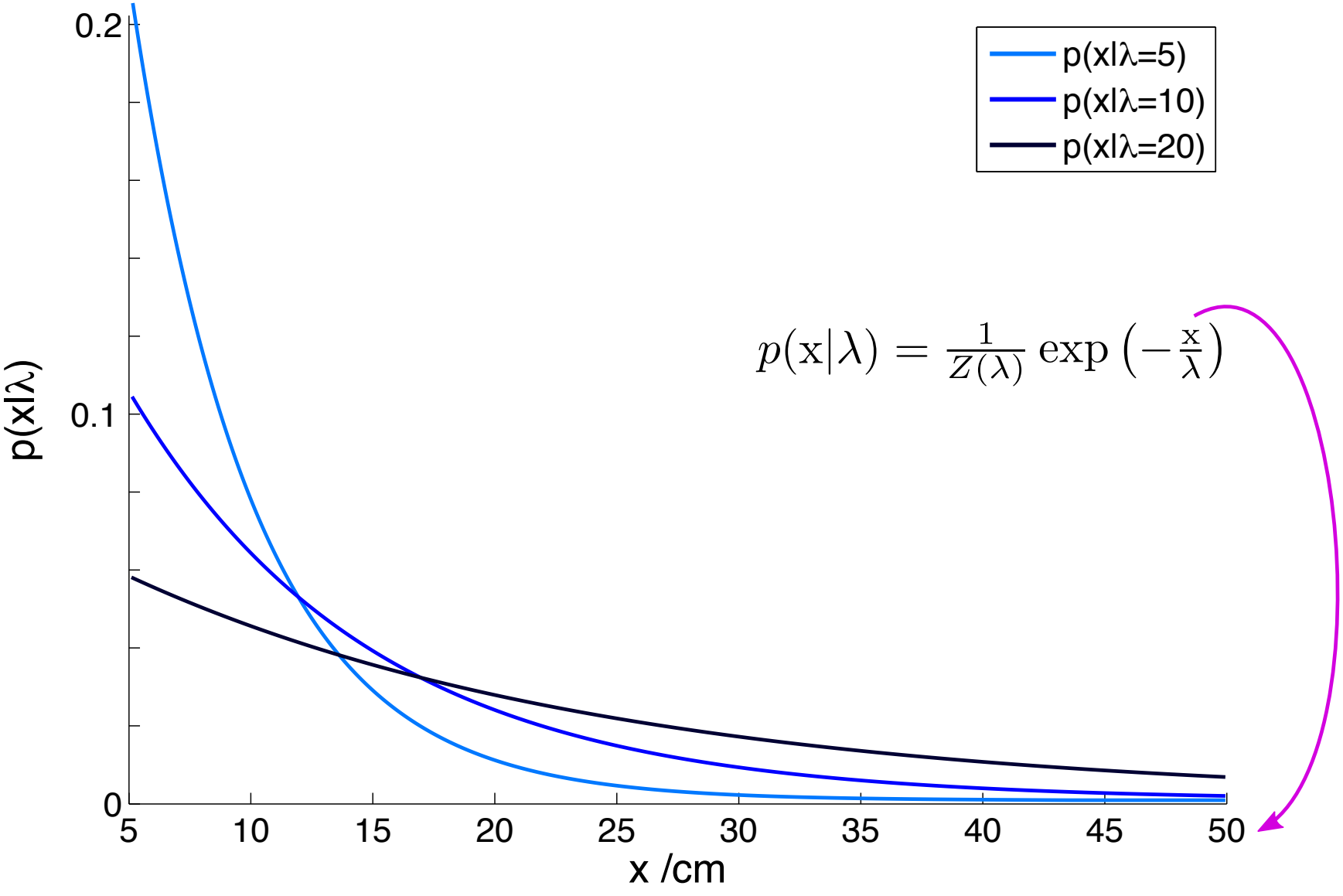
Density


p(xlλ=5)
$p(\mathrm{x}|\lambda) = \frac{1}{Z(\lambda)} \exp\left(-\frac{\mathrm{x}}{\lambda}\right)$
p(xlλ)
0.2
0.1
0
5
10
15
20
25
30
35
40
45
50
x /cm

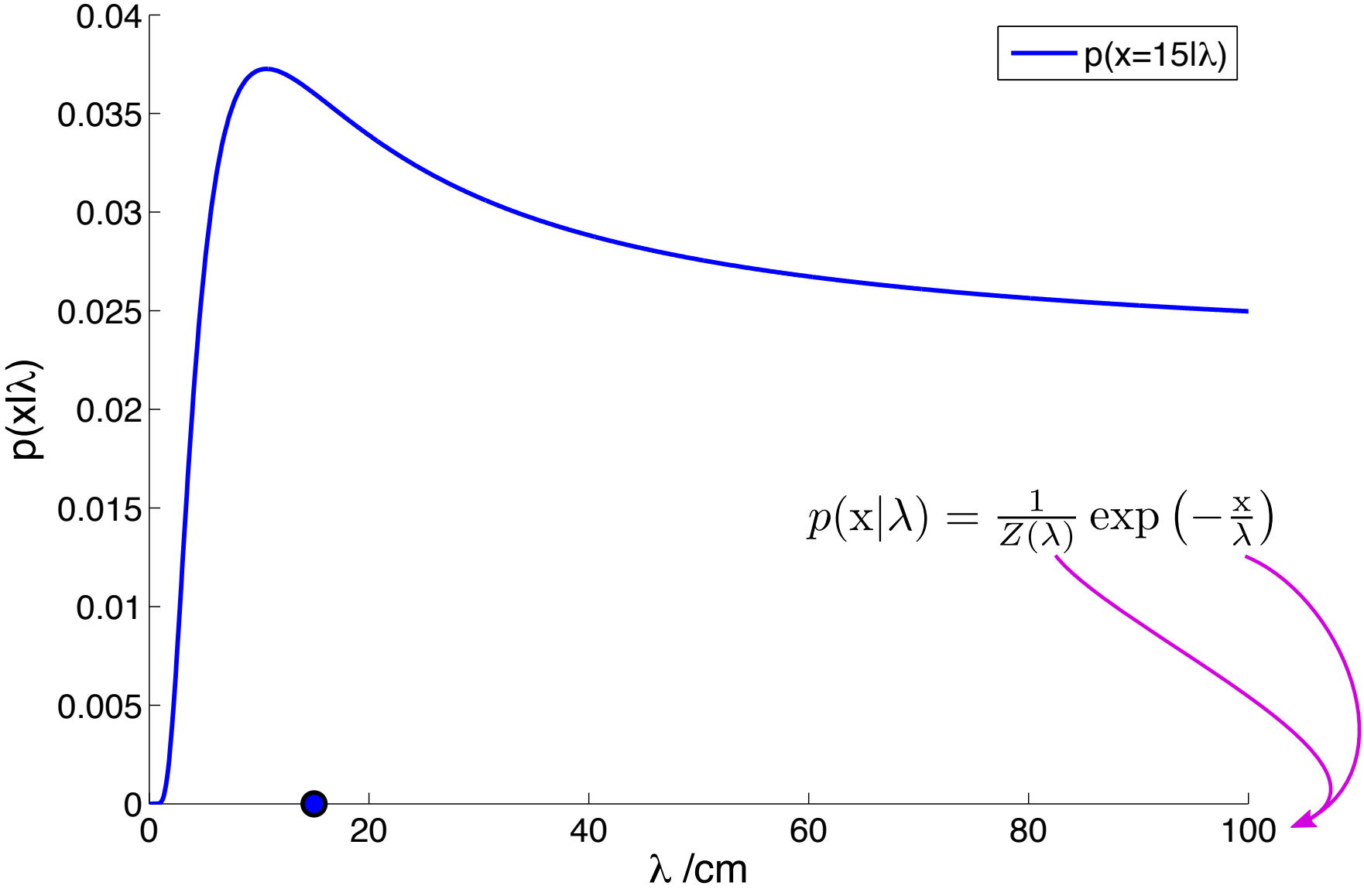
Density


0.2
0.1
0
p(x|λ)
5
10
15
20
25
30
35
40
45
50
x /cm
p(x|λ=5)
p(x|λ=10)
$p(\mathrm{x}|\lambda) = \frac{1}{Z(\lambda)} \exp\left(-\frac{\mathrm{x}}{\lambda}\right)$

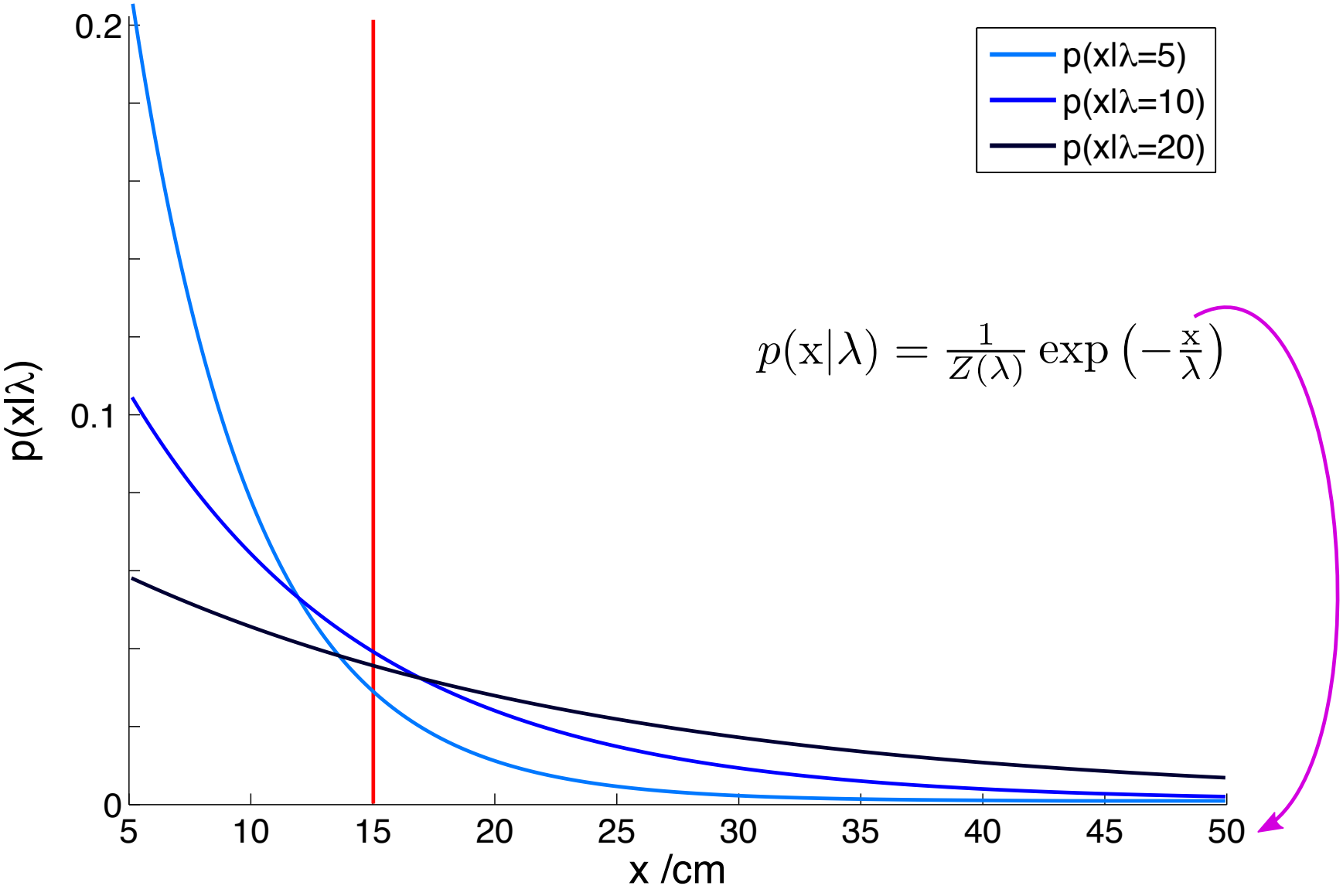
Density


p(x|λ=5)
p(x|λ=10)
p(x|λ=20)
$p(\mathrm{x}|\lambda) = \frac{1}{Z(\lambda)} \exp\left(-\frac{\mathrm{x}}{\lambda}\right)$
p(x|λ)
0.2
0.1
0
5
10
15
20
25
30
35
40
45
50
x /cm

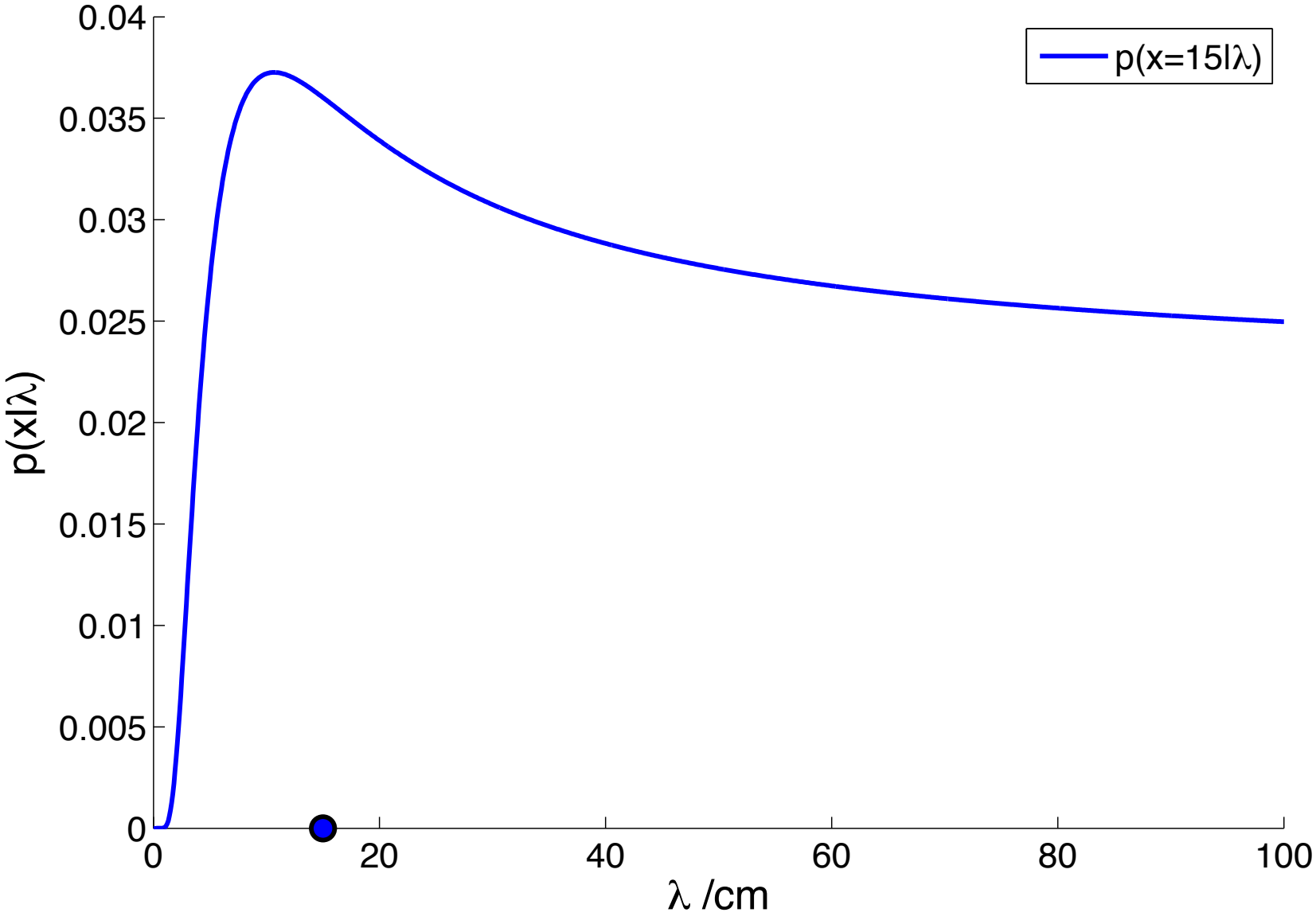
Likelihood of the parameters



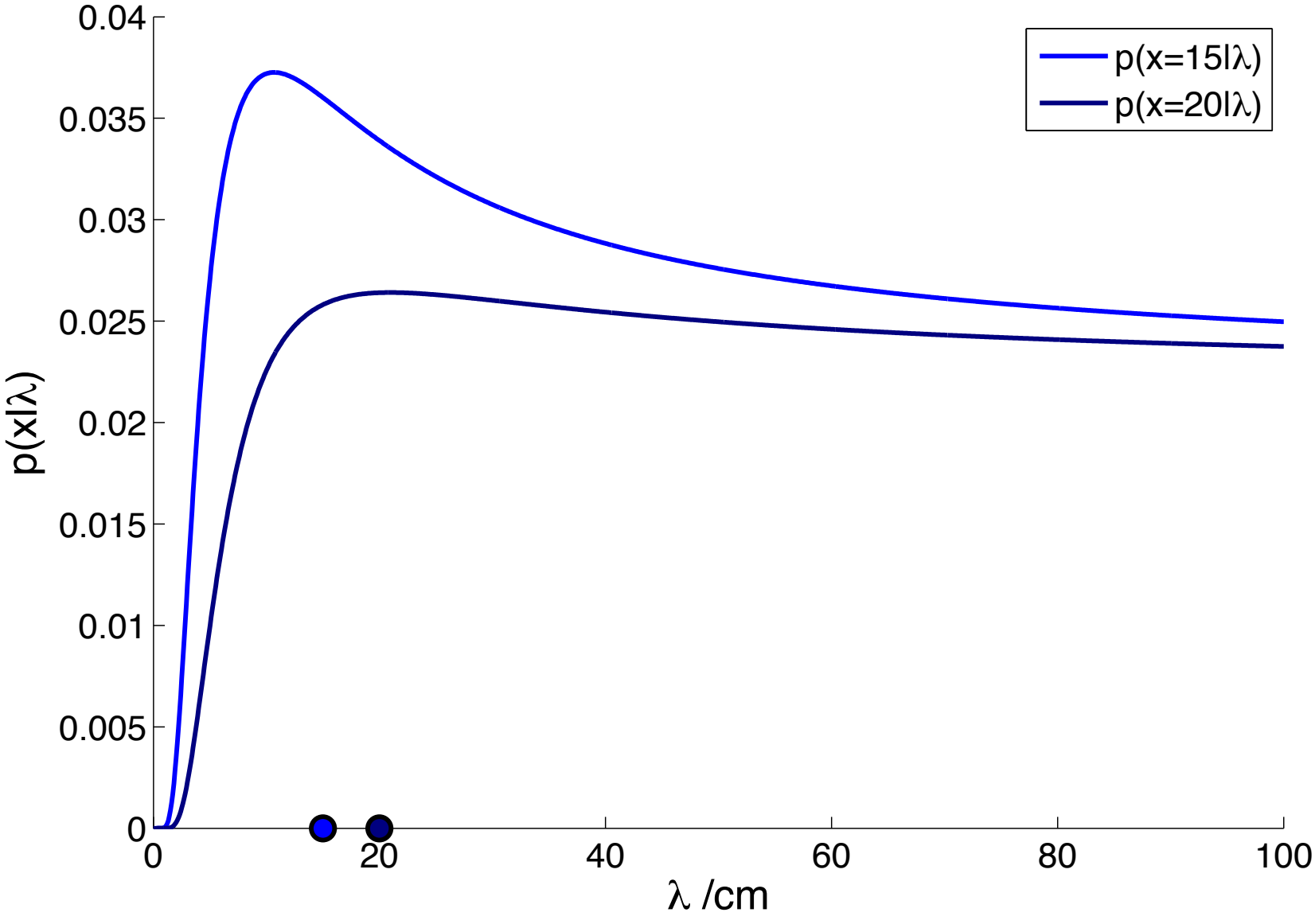
Density $p(\mathrm{x}|\lambda) = \frac{1}{Z(\lambda)} \exp(-\mathrm{x}/\lambda)$



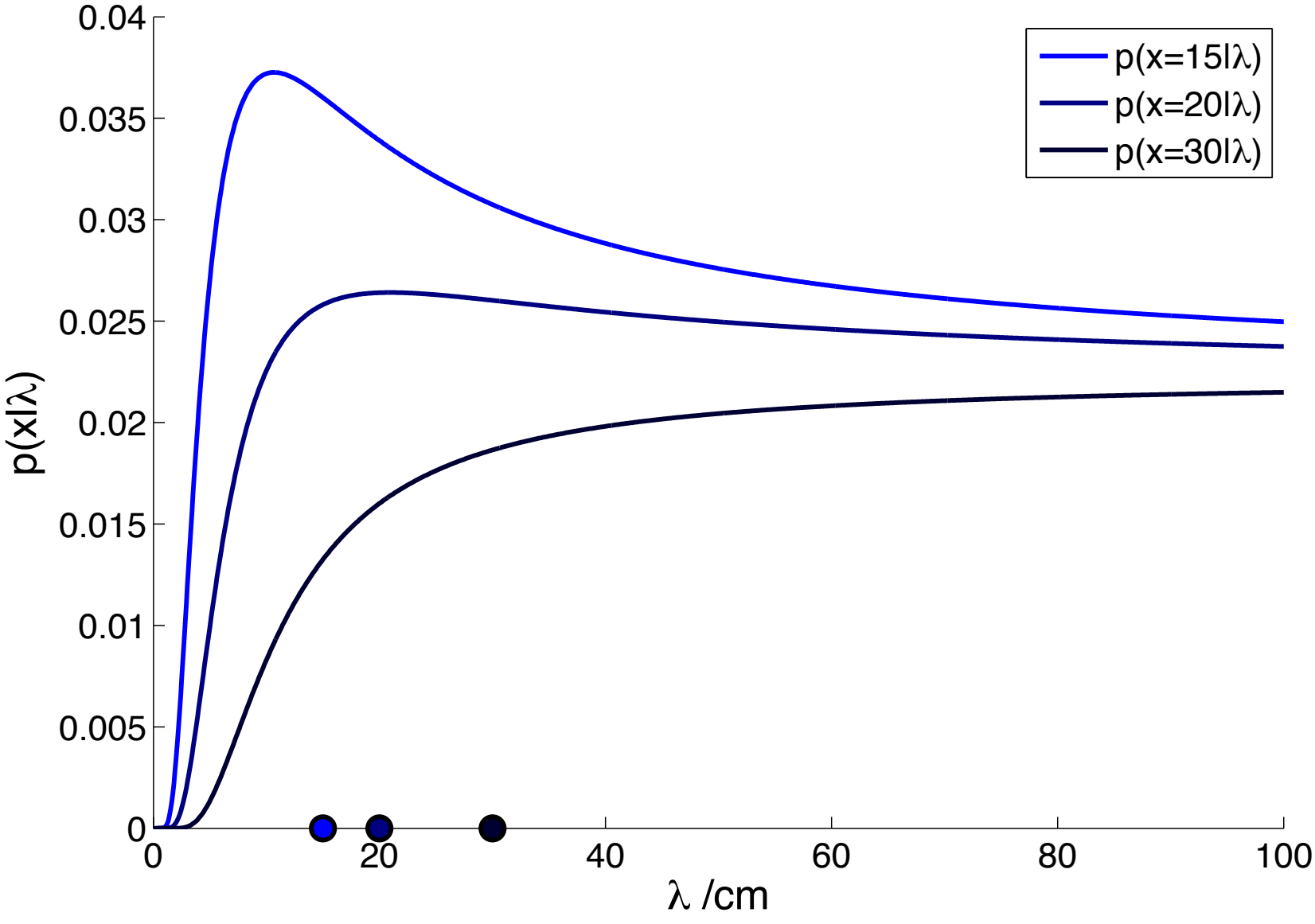
Likelihood of the parameters $p(\mathrm{x}\,|\lambda) = \frac{1}{Z(\lambda)} \exp(-\mathrm{x}\,/\lambda)$



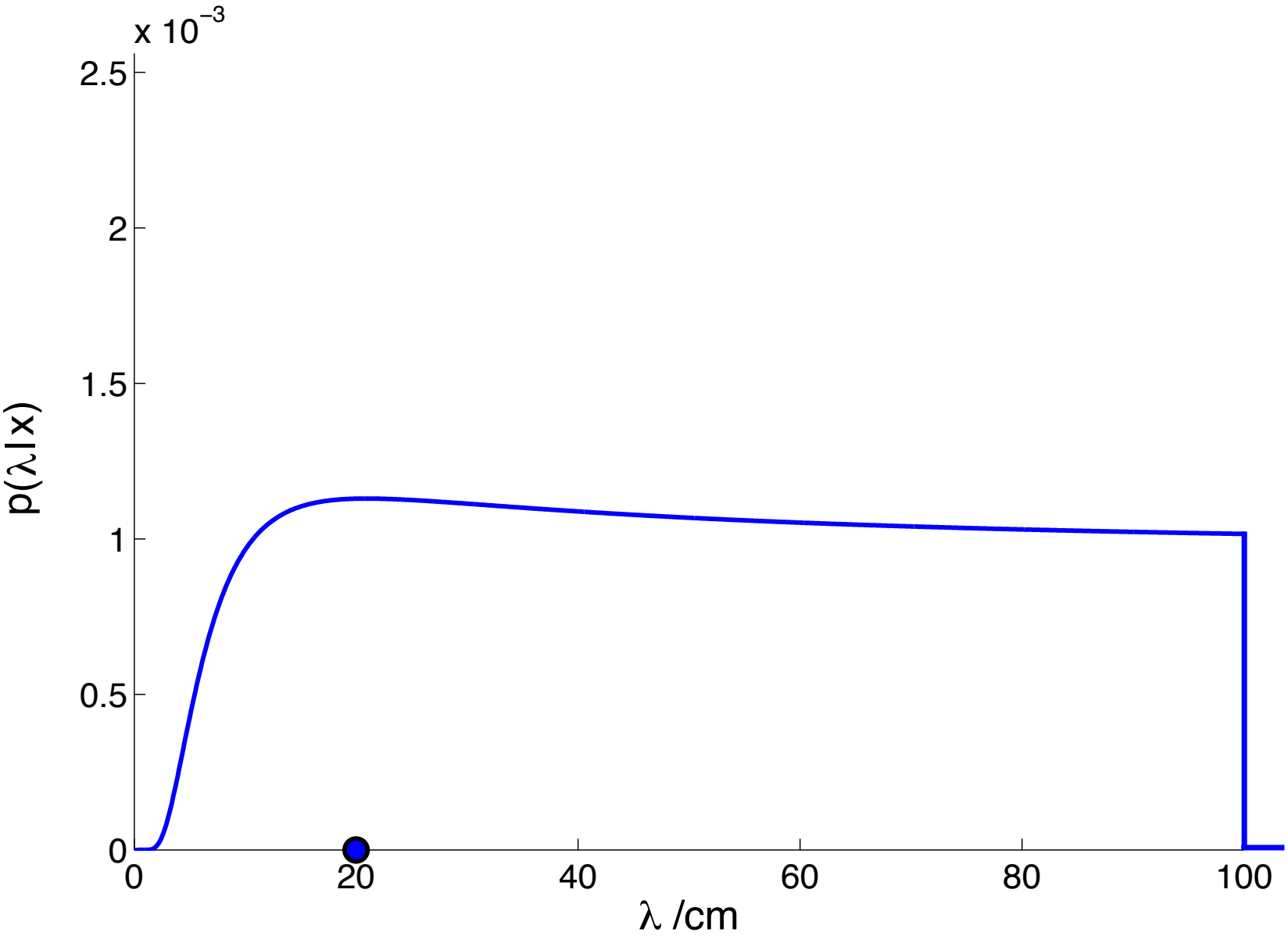
Likelihood of the parameters $p(\mathrm{x}\,|\lambda) = \frac{1}{Z(\lambda)} \exp(-\mathrm{x}\,/\lambda)$



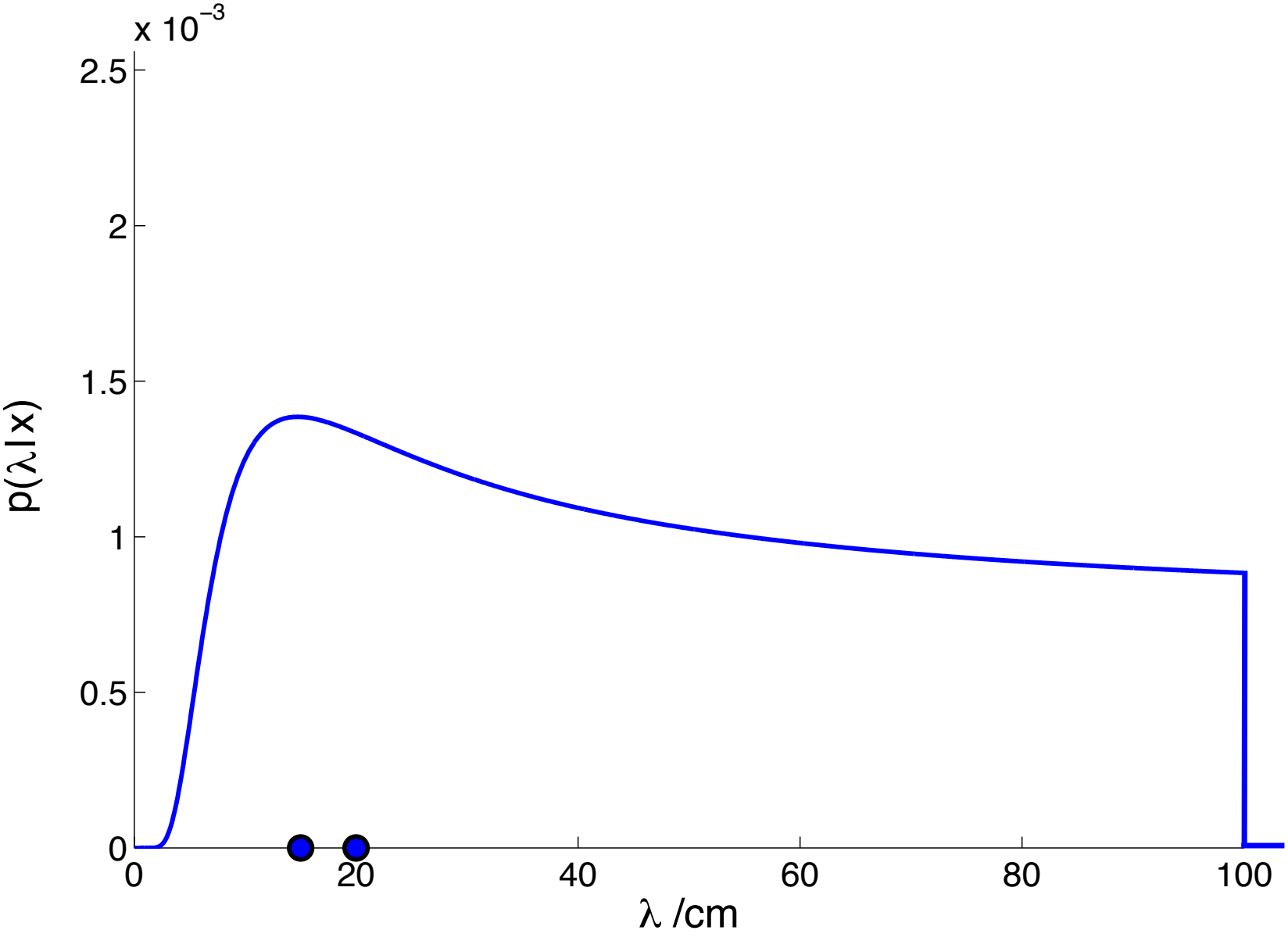
Likelihood of the parameters $p(\mathrm{x}\,|\lambda) = \frac{1}{Z(\lambda)} \exp(-\mathrm{x}\,/\lambda)$



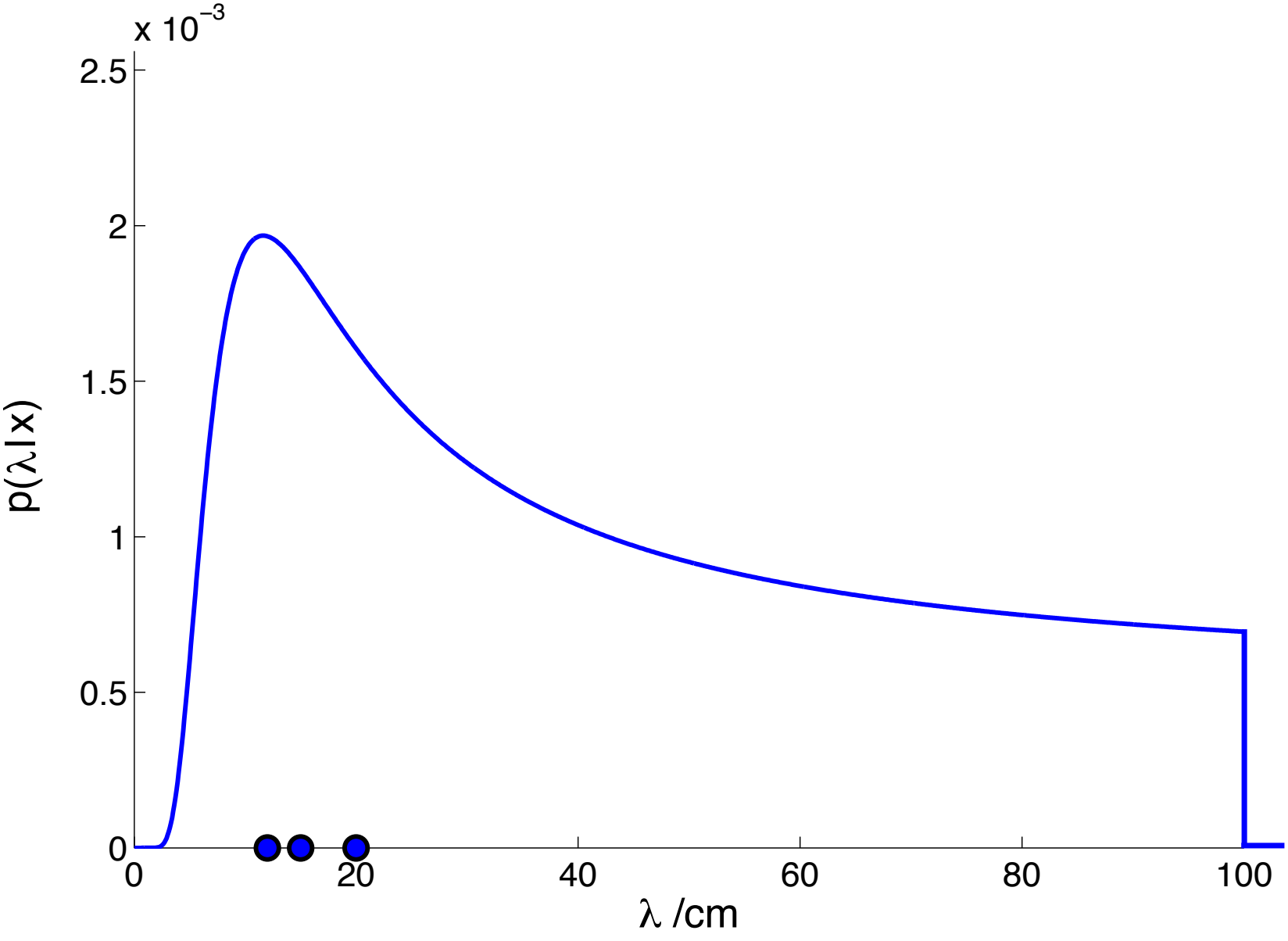
Posterior distribution: $p(\lambda|\,\mathrm{x}_1) \propto p(\lambda)p(\mathrm{x}_1\,|\lambda)$



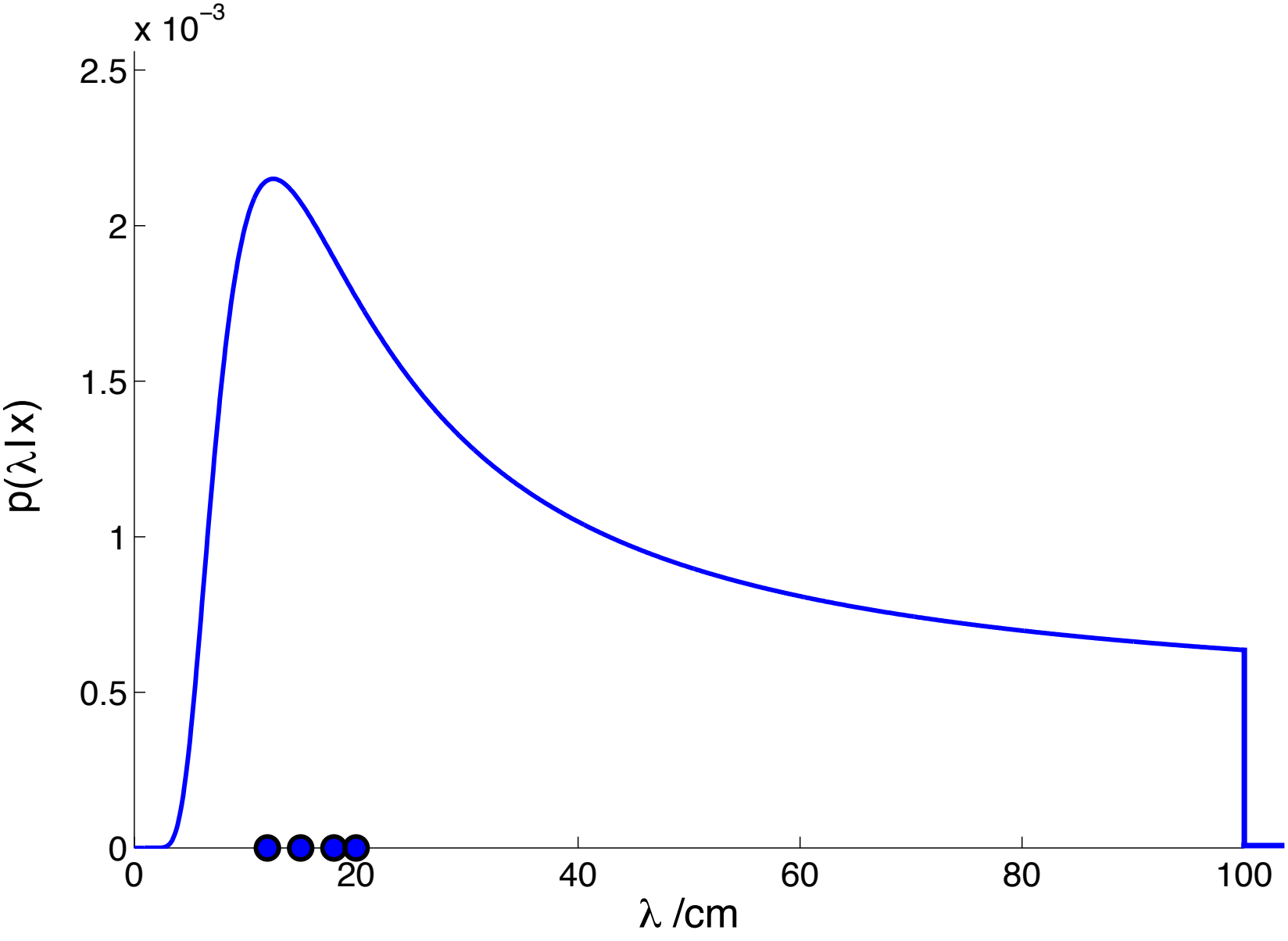
Posterior distribution: $p(\lambda|\,\mathrm{x}_1, \mathrm{x}_2) \propto p(\lambda) \prod_{n=1}^{2} p(\mathrm{x}_n\,|\lambda)$



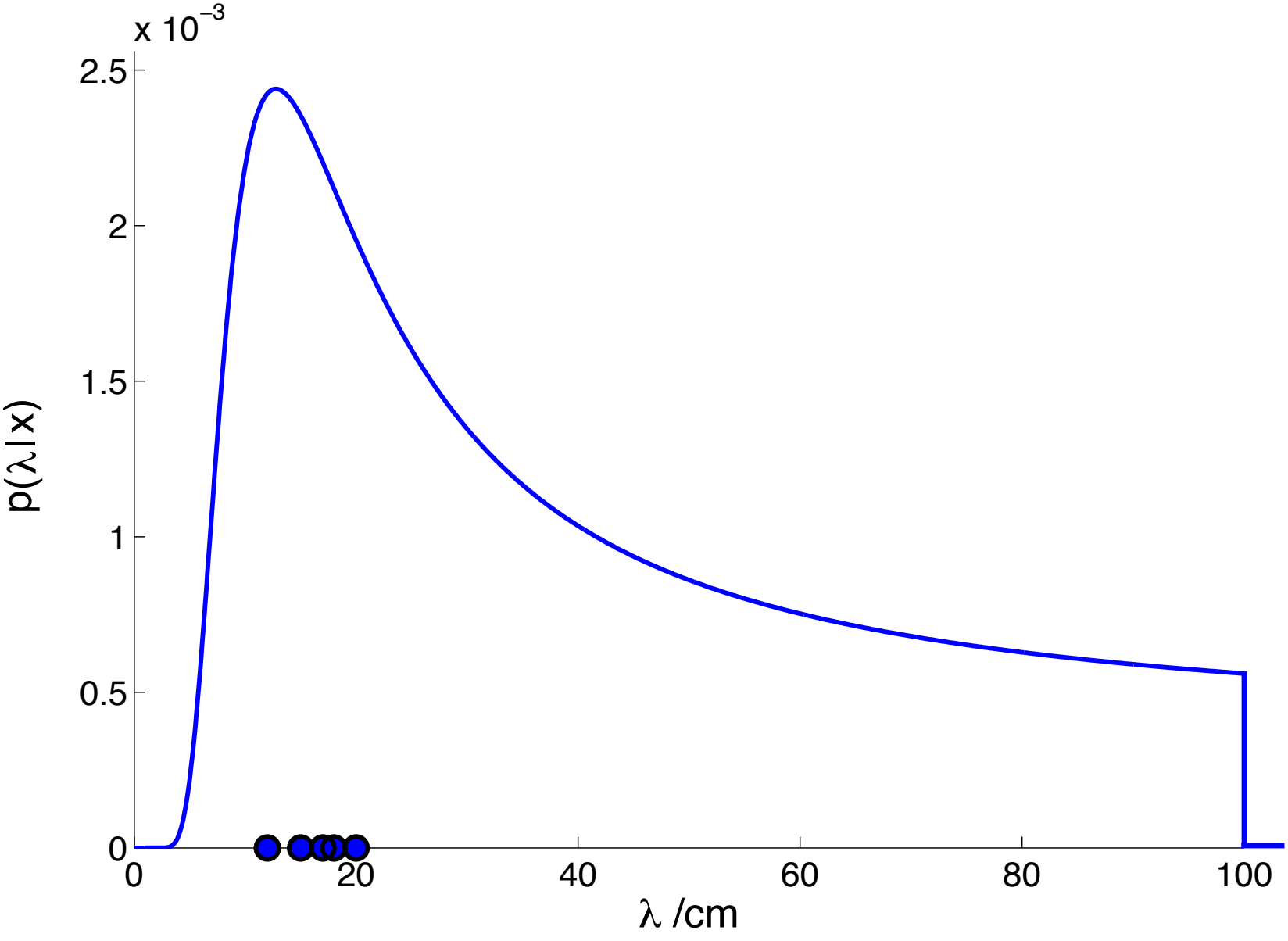
Posterior distribution: $p(\lambda|\,\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3) \propto p(\lambda) \prod_{n=1}^{3} p(\mathrm{x}_n\,|\lambda)$



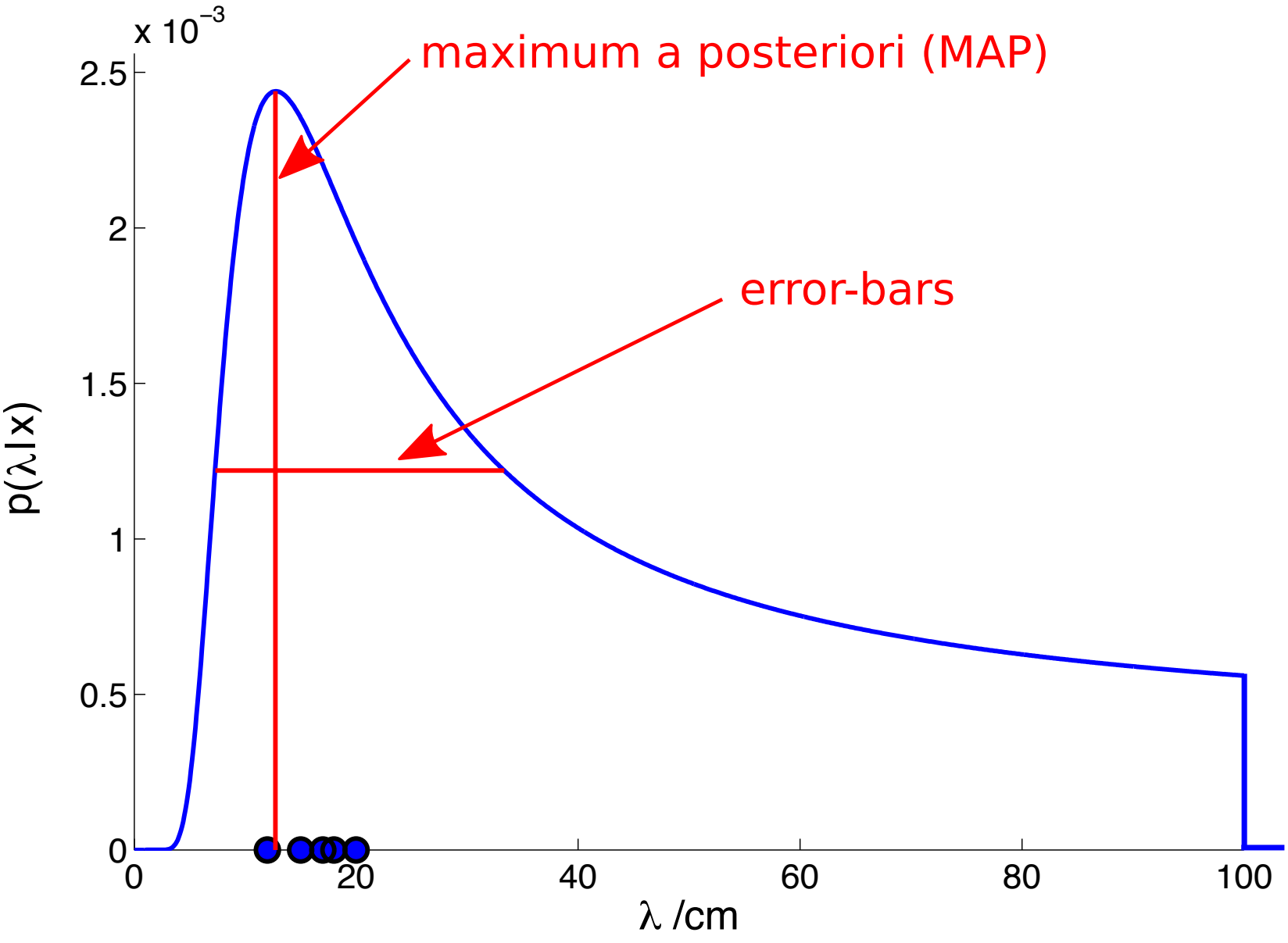
Posterior distribution: $p(\lambda|\,\mathrm{x}_1,\mathrm{x}_2,\mathrm{x}_3,\mathrm{x}_4) \propto p(\lambda)\prod_{n=1}^{4} p(\mathrm{x}_n\,|\lambda)$



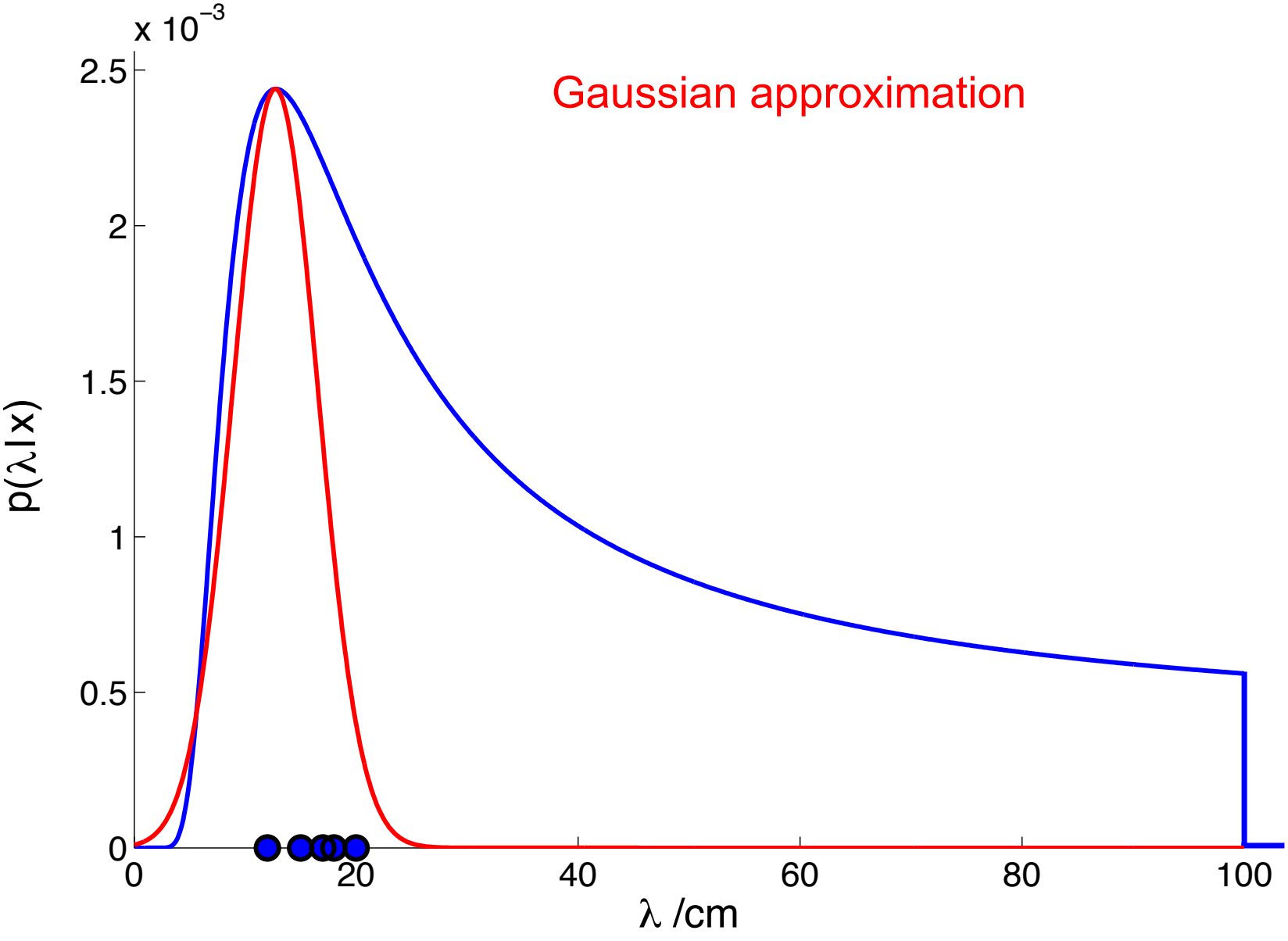
Posterior distribution: $p(\lambda|\,\mathrm{x}_1,\mathrm{x}_2,\mathrm{x}_3,\mathrm{x}_4,\mathrm{x}_5) \propto p(\lambda)\prod_{n=1}^{5} p(\mathrm{x}_n\,|\lambda)$



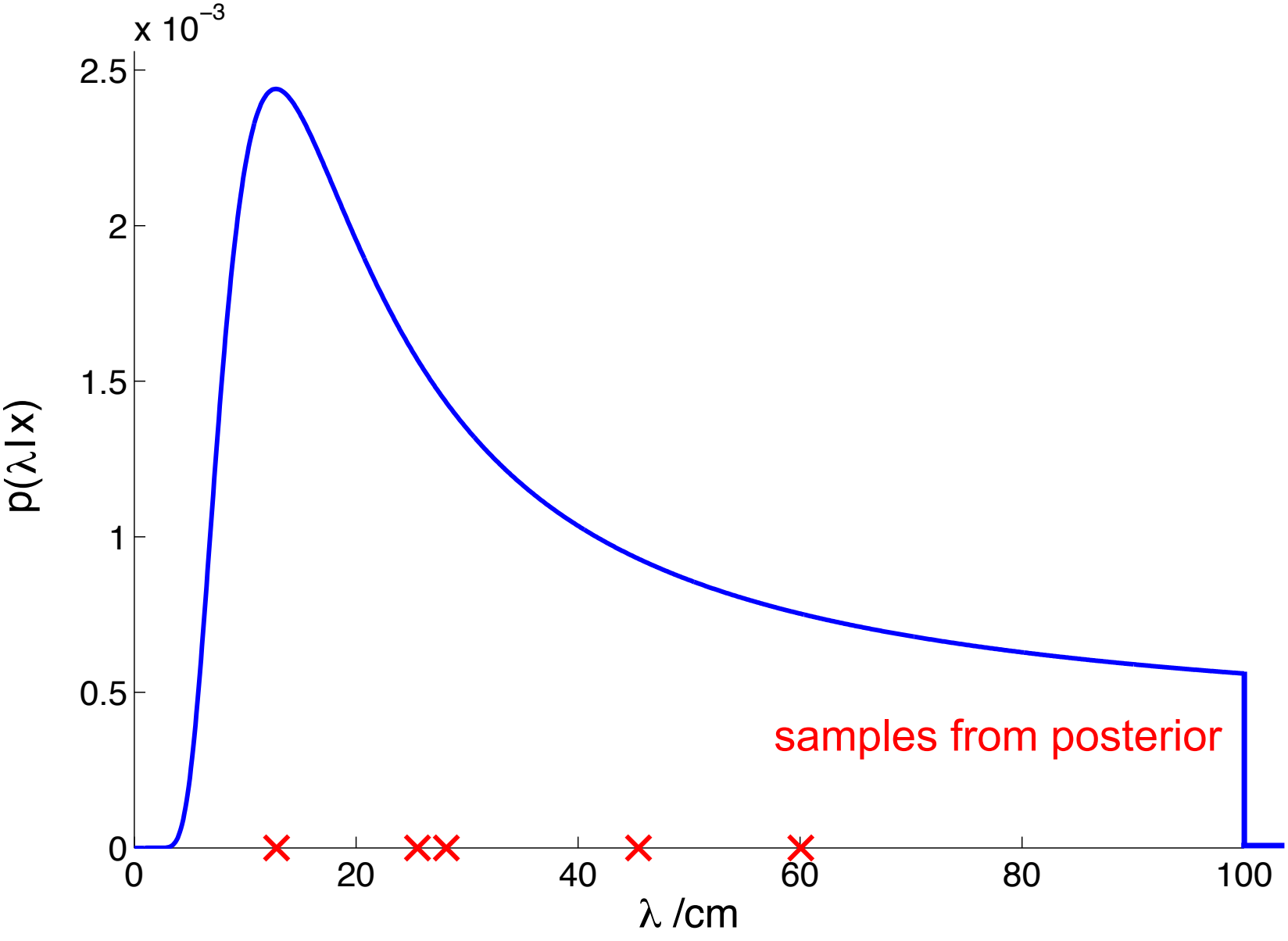
Summarising the posterior distribution



Summarising the posterior distribution



Summarising the posterior distribution



Summary of probabilistic approach

- write down the probability of everything (**joint distribution**)

$$p(\{\mathrm{x}_n\}_{n=1}^N, \lambda) = p(\lambda)p(\{\mathrm{x}_n\}_{n=1}^N|\lambda)$$

- use Bayes' rule (product rule) to form the **posterior distribution**

$$p(\lambda|\{\mathrm{x}_n\}_{n=1}^N) = \frac{1}{p(\{\mathrm{x}_n\}_{n=1}^N)}p(\lambda)p(\{\mathrm{x}_n\}_{n=1}^N|\lambda)$$

- summarise the posterior e.g. via the **maximum a posteriori** (MAP) estimate

$$\lambda^{\mathsf{MAP}} = \arg\max_{\lambda} p(\lambda|\{\mathrm{x}_n\}_{n=1}^N)$$

- **maximum likelihood** estimate is recovered when using a wide uniform prior distribution

$$\lambda^{\mathsf{ML}} = \arg\max_{\lambda} p(\{\mathrm{x}_n\}_{n=1}^N|\lambda)$$

A second inference problem: Medical example

Alice has a test for a disease:

- $a = 1$ indicates Alice has the disease, $a = 0$ indicates she does not
- $b = 1$ indicates positive test result, $b = 0$ indicates it is negative

A second inference problem: Medical example

Alice has a test for a disease:

- $a = 1$ indicates Alice has the disease, $a = 0$ indicates she does not
- $b = 1$ indicates positive test result, $b = 0$ indicates it is negative

The test is 95% reliable:

- in 95% of cases of people who really have the disease, a positive result is returned
- in 95% of cases of people who do not have the disease, a negative result is obtained

A second inference problem: Medical example

Alice has a test for a disease:

- $a = 1$ indicates Alice has the disease, $a = 0$ indicates she does not
- $b = 1$ indicates positive test result, $b = 0$ indicates it is negative

The test is 95% reliable:

- in 95% of cases of people who really have the disease, a positive result is returned
- in 95% of cases of people who do not have the disease, a negative result is obtained

5% of people of Alice's age and background have the disease

A second inference problem: Medical example

Alice has a test for a disease:

- $a = 1$ indicates Alice has the disease, $a = 0$ indicates she does not
- $b = 1$ indicates positive test result, $b = 0$ indicates it is negative

The test is 95% reliable:

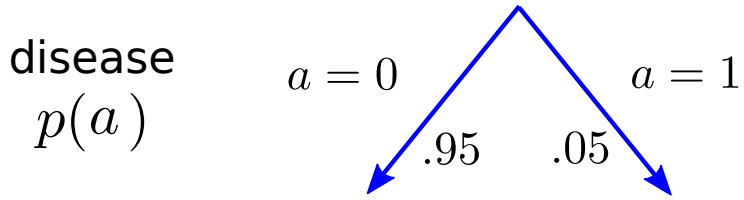
- in 95% of cases of people who really have the disease, a positive result is returned
- in 95% of cases of people who do not have the disease, a negative result is obtained

5% of people of Alice's age and background have the disease

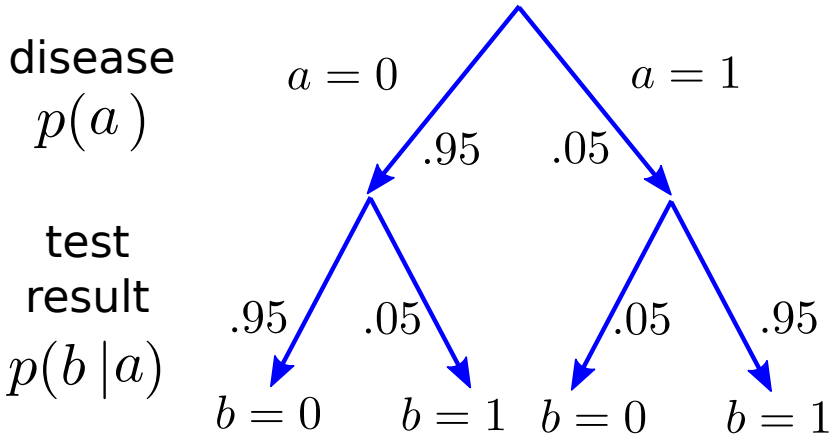
Alice has the test and the result is positive

- **Compute the probability that Alice has the disease**

A second inference problem: Medical example



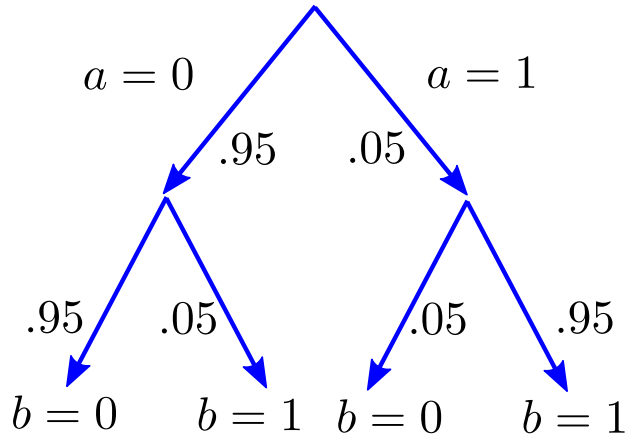
A second inference problem: Medical example



A second inference problem: Medical example

disease
$p(a)$

test
result
$p(b\,|a)$



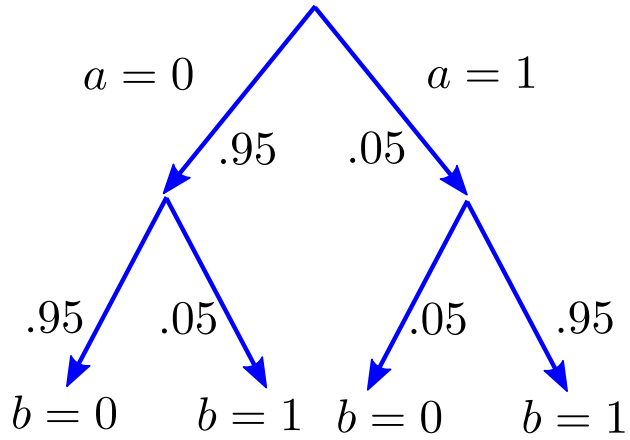
$p(b, a)$

b \ a	0	1
0	$.95 \times .95$	$.05 \times .05$
1	$.95 \times .05$	$.05 \times .95$

A second inference problem: Medical example

disease
$p(a)$

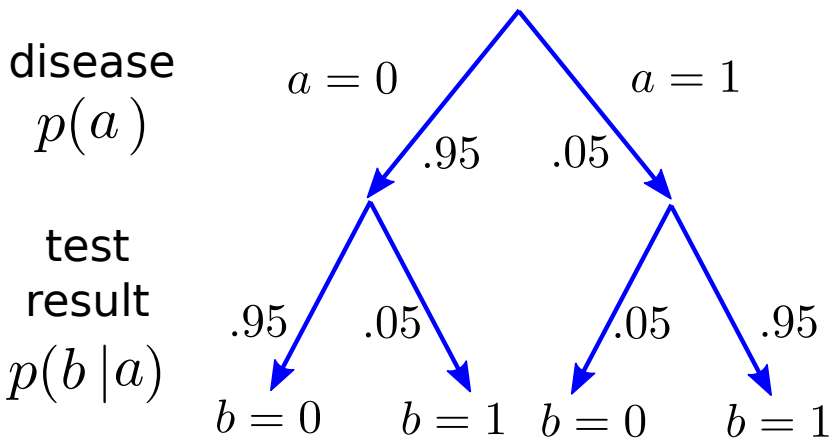
test
result
$p(b\,|a)$



$p(b, a)$

b \ a	0	1
0	$.95 \times .95$	$.05 \times .05$
1	$.95 \times .05$	$.05 \times .95$

A second inference problem: Medical example

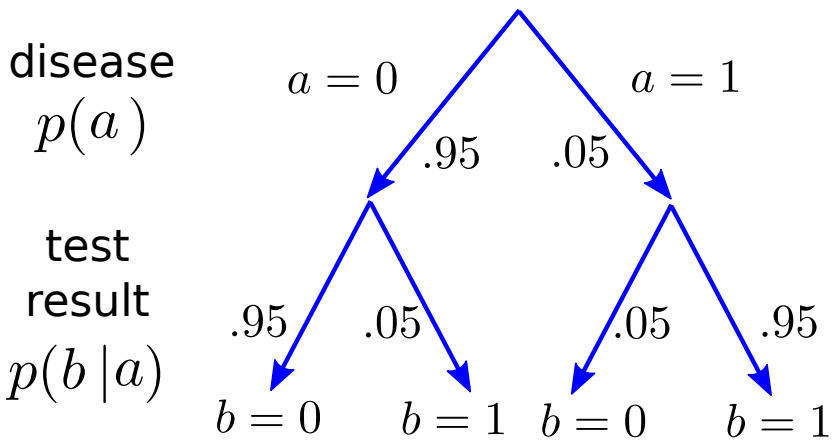


$p(b,a)$

b \ a	0	1
0	$.95\times.95$	$.05\times.05$
1	$.95\times.05$	$.05\times.95$

$$p(a=1|b=1)=\frac{p(b=1|a=1)p(a=1)}{p(b=1)}$$

A second inference problem: Medical example



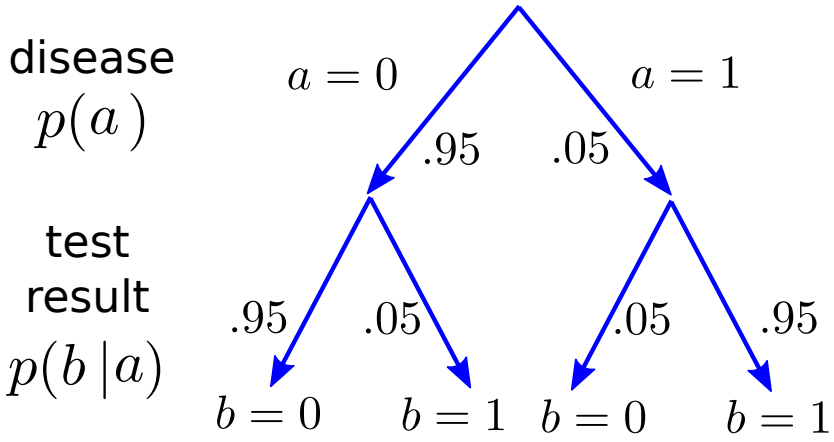
$p(b, a)$

b \ a	0	1
0	$.95 \times .95$	$.05 \times .05$
1	$.95 \times .05$	$.05 \times .95$

$$p(a=1|b=1) = \frac{p(b=1|a=1)p(a=1)}{p(b=1)}$$

$$= \frac{p(b=1|a=1)p(a=1)}{p(b=1|a=1)p(a=1) + p(b=1|a=0)p(a=0)}$$

A second inference problem: Medical example



$p(b,a)$

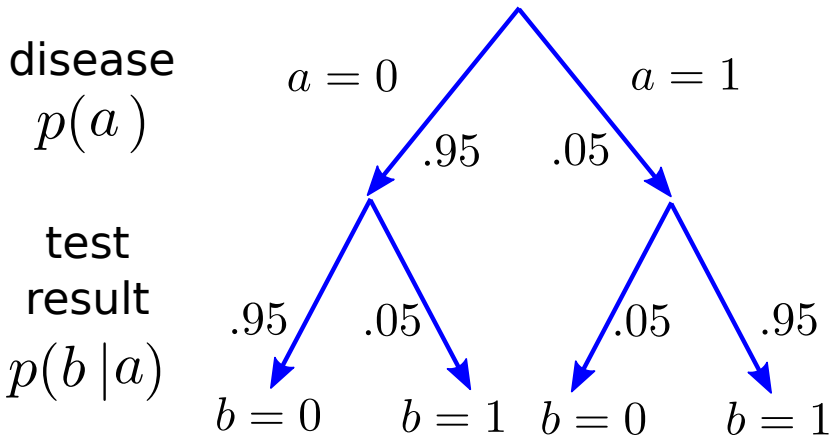
b \ a	0	1
0	$.95\times.95$	$.05\times.05$
1	$.95\times.05$	$.05\times.95$

$$p(a=1|b=1) = \frac{p(b=1|a=1)p(a=1)}{p(b=1)}$$

$$= \frac{p(b=1|a=1)p(a=1)}{p(b=1|a=1)p(a=1)+p(b=1|a=0)p(a=0)}$$

$$= \frac{.95\times 0.05}{.95\times 0.05+.05\times .95} = \frac{1}{2}$$

A second inference problem: Medical example



$p(b,a)$

b \ a	0	1
0	$.95\times.95$	$.05\times.05$
1	$.95\times.05$	$.05\times.95$

$$p(a=1|b=1) = \frac{p(b=1|a=1)p(a=1)}{p(b=1)}$$

$$= \frac{p(b=1|a=1)p(a=1)}{p(b=1|a=1)p(a=1)+p(b=1|a=0)p(a=0)}$$

$$= \frac{.95\times 0.05}{.95\times 0.05+.05\times .95} = \frac{1}{2}$$

Test is 95% reliable, but probability of having the disease is only 50%

A second inference problem: Medical example

Alice has a test for a disease:

- $a = 1$ indicates Alice has the disease, $a = 0$ indicates she does not
- $b = 1$ indicates positive test result, $b = 0$ indicates it is negative

The test is 95% reliable:

- in 95% of cases of people who really have the disease, a positive result is returned
- in 95% of cases of people who do not have the disease, a negative result is obtained

5% of people of Alice's age and background have the disease

Alice has the test and the result is positive

A second inference problem: Medical example

Alice has a test for a disease:

- $a = 1$ indicates Alice has the disease, $a = 0$ indicates she does not
- $b = 1$ indicates positive test result, $b = 0$ indicates it is negative

The test is 95% reliable:

- in 95% of cases of people who really have the disease, a positive result is returned
- in 95% of cases of people who do not have the disease, a negative result is obtained

5% of people of Alice's age and background have the disease

Alice has the test and the result is positive

Treatment for the disease:

- t =1 indicates Alice is treated, t =0 indicates that she is not

A second inference problem: Medical example

Alice has a test for a disease:

- $a = 1$ indicates Alice has the disease, $a = 0$ indicates she does not
- $b = 1$ indicates positive test result, $b = 0$ indicates it is negative

The test is 95% reliable:

- in 95% of cases of people who really have the disease, a positive result is returned
- in 95% of cases of people who do not have the disease, a negative result is obtained

5% of people of Alice's age and background have the disease

Alice has the test and the result is positive

Treatment for the disease:

- t =1 indicates Alice is treated, t =0 indicates that she is not
- Alice's quality of life R depends on whether she has the disease and whether she is treated:

$$\begin{bmatrix} R(a=0,t=0) & R(a=0,t=1) \\ R(a=1,t=0) & R(a=1,t=1) \end{bmatrix} = \begin{bmatrix} 10 & 7 \\ 3 & 5 \end{bmatrix}$$

A second inference problem: Medical example

Alice has a test for a disease:

- $a = 1$ indicates Alice has the disease, $a = 0$ indicates she does not
- $b = 1$ indicates positive test result, $b = 0$ indicates it is negative

The test is 95% reliable:

- in 95% of cases of people who really have the disease, a positive result is returned
- in 95% of cases of people who do not have the disease, a negative result is obtained

5% of people of Alice's age and background have the disease

Alice has the test and the result is positive

Treatment for the disease:

- t =1 indicates Alice is treated, t =0 indicates that she is not
- Alice's quality of life R depends on whether she has the disease and whether she is treated:

$$\begin{bmatrix} R(a=0,t=0) & R(a=0,t=1) \\ R(a=1,t=0) & R(a=1,t=1) \end{bmatrix} = \begin{bmatrix} 10 & 7 \\ 3 & 5 \end{bmatrix}$$

- **Should Alice be treated?**

Bayesian Decision Theory

posterior

$$p(a = 1|b = 1) = \frac{1}{2}$$
$$p(a = 0|b = 1) = \frac{1}{2}$$

reward

$$\begin{bmatrix} R(a = 0, t = 0) & R(a = 0, t = 1) \\ R(a = 1, t = 0) & R(a = 1, t = 1) \end{bmatrix} = \begin{bmatrix} 10 & 7 \\ 3 & 5 \end{bmatrix}$$

Bayesian Decision Theory

posterior

$$p(a=1|b=1) = \frac{1}{2}$$
$$p(a=0|b=1) = \frac{1}{2}$$

reward

$$\begin{bmatrix} R(a=0,t=0) & R(a=0,t=1) \\ R(a=1,t=0) & R(a=1,t=1) \end{bmatrix} = \begin{bmatrix} 10 & 7 \\ 3 & 5 \end{bmatrix}$$

conditional reward

$$\mathrm{R}(t) = \sum_a R(a,t)p(a|b=1)$$

reward for action in that world

action

sum over possible worlds

posterior probability of world

Bayesian Decision Theory

posterior

$$p(a=1|b=1)=\frac{1}{2}$$
$$p(a=0|b=1)=\frac{1}{2}$$

reward

$$\begin{bmatrix} R(a=0,t=0) & R(a=0,t=1) \\ R(a=1,t=0) & R(a=1,t=1) \end{bmatrix} = \begin{bmatrix} 10 & 7 \\ 3 & 5 \end{bmatrix}$$

conditional reward

$$\mathrm{R}(t)=\sum_{a} R(a,t)p(a|b=1)$$

reward for action in that world

action

sum over possible worlds

posterior probability of world

can separate inference and decision making

Bayesian Decision Theory

posterior

$$p(a=1|b=1)=\frac{1}{2}$$
$$p(a=0|b=1)=\frac{1}{2}$$

reward

$$\begin{bmatrix} R(a=0,t=0) & R(a=0,t=1) \\ R(a=1,t=0) & R(a=1,t=1) \end{bmatrix} = \begin{bmatrix} 10 & 7 \\ 3 & 5 \end{bmatrix}$$

conditional reward

$$\mathrm{R}(t) = \sum_{a} R(a,t)p(a|b=1)$$

reward for action in that world

action

sum over possible worlds

posterior probability of world

can separate inference and decision making

conditional reward for not treating

$$\mathrm{R}(t=0) = R(a=0,t=0)p(a=0|b=1) + R(a=1,t=0)p(a=1|b=1) = 6\frac{1}{2}$$

Bayesian Decision Theory

posterior

$$p(a=1|b=1)=\frac{1}{2}$$
$$p(a=0|b=1)=\frac{1}{2}$$

reward

$$\begin{bmatrix} R(a=0,t=0) & R(a=0,t=1) \\ R(a=1,t=0) & R(a=1,t=1) \end{bmatrix} = \begin{bmatrix} 10 & 7 \\ 3 & 5 \end{bmatrix}$$

conditional reward

$$\mathrm{R}(t)=\sum_{a} R(a,t)p(a|b=1)$$

reward for action in that world

action

sum over possible worlds

posterior probability of world

can separate inference and decision making

conditional reward for not treating

$$\mathrm{R}(t=0)=R(a=0,t=0)p(a=0|b=1)+R(a=1,t=0)p(a=1|b=1)=6\frac{1}{2}$$

conditional reward for treating

$$\mathrm{R}(t=1)=R(a=0,t=1)p(a=0|b=1)+R(a=1,t=1)p(a=1|b=1)=6$$

Dutch Book Theorem

Assume you are willing to accept bets with odds proportional to the strength of your beliefs

e.g. b(x) = 0.9 implies that you will accept a bet:

x is true win $\geq$ \$1
x is false lose \$9

If your beliefs do not satisfy the rules of probability theory (sum and product rules) there exists a set of simultaneous bets (called a "Dutch Book") which you are willing to accept, and for which you are **guaranteed to lose money no matter what the outcome**.

The only way to guard against Dutch Books to to ensure that your beliefs are coherent: i.e. satisfy the rules of probability.

Flavours of machine learning

Machine experiences a series of sensory inputs: $\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \mathrm{x}_4, \ldots$

Supervised learning: machine also given desired outputs $\mathrm{y}_1, \mathrm{y}_2, \ldots$ and its goal is to learn to produce the correct output given a new input.

Unsupervised learning: goal of the machine is to build a model of x that can be used for reasoning, decision making, predicting things, communicating etc.

Reinforcement learning: machine can also produce actions $a_1, a_2, \ldots$ which affect the state of the world, and receives rewards (or punishments) $r_1, r_2, \ldots$. Its goal is to learn to act in a way that maximises rewards in the long term.

Outline of the course: Regression

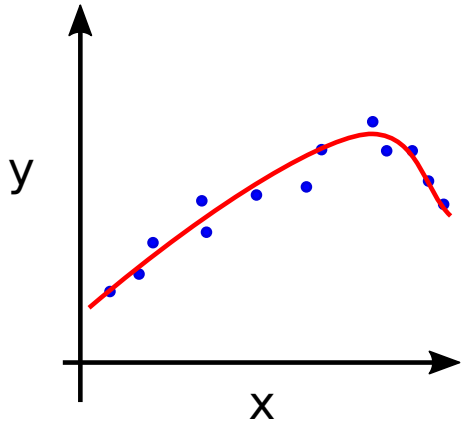
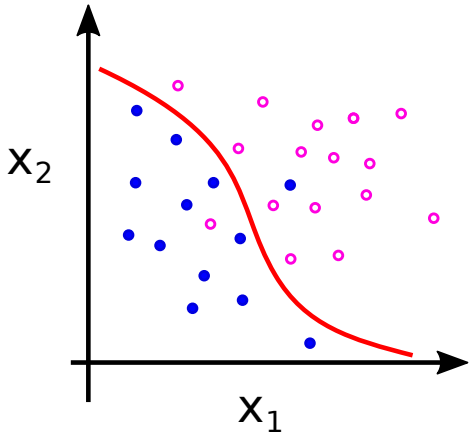


image restoration

opinion polls

pose estimation
power demand prediction
weather forecasting
...

Outline of the course: Classification



object recognition

www.clarifai.com

speech recognition

The quick brown fox jumped over the lazy ...

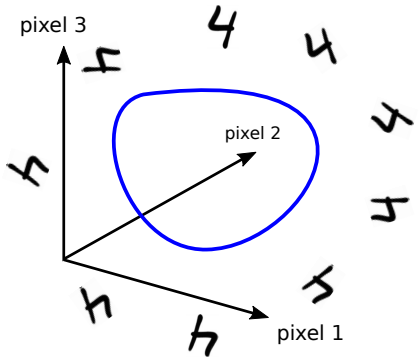
face recognition

identity

spam filtering
medical diagnosis
drug discovery

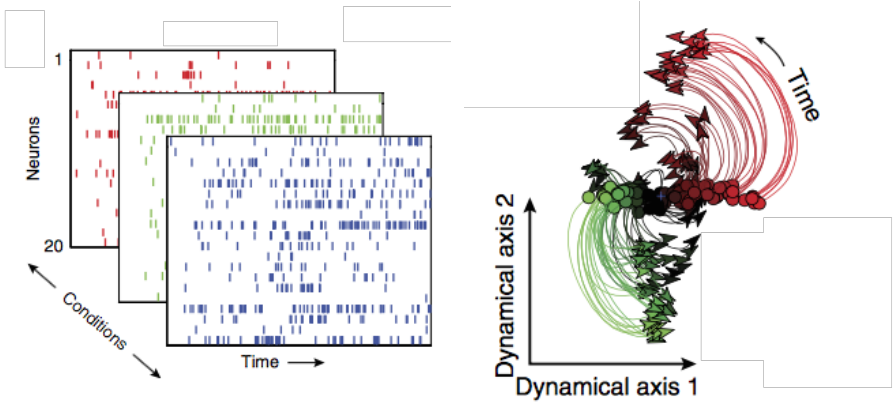
credit scoring
click stream analysis
...

Outline of the course: Dimensionality Reduction



visualisation

understanding structure
in high-dimensional data



Cunningham and Yu, Nature Neuro, 2014

visualisation

http://cs.stanford.edu/people/karpathy/cnnembed/

modelling data on/near manifolds

e.g. objects + transformations
= non-linear manifolds

preprocessing/feature learning:

reducing computational complexity
improving statistical efficiency

Outline of the course: Clustering

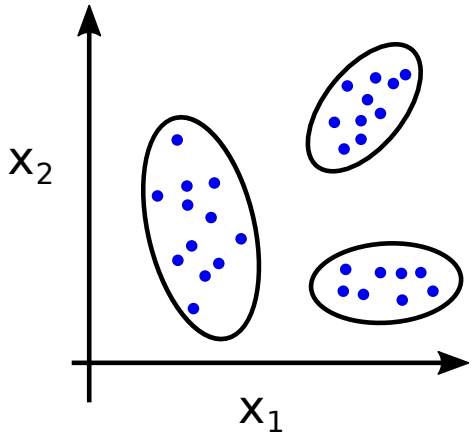
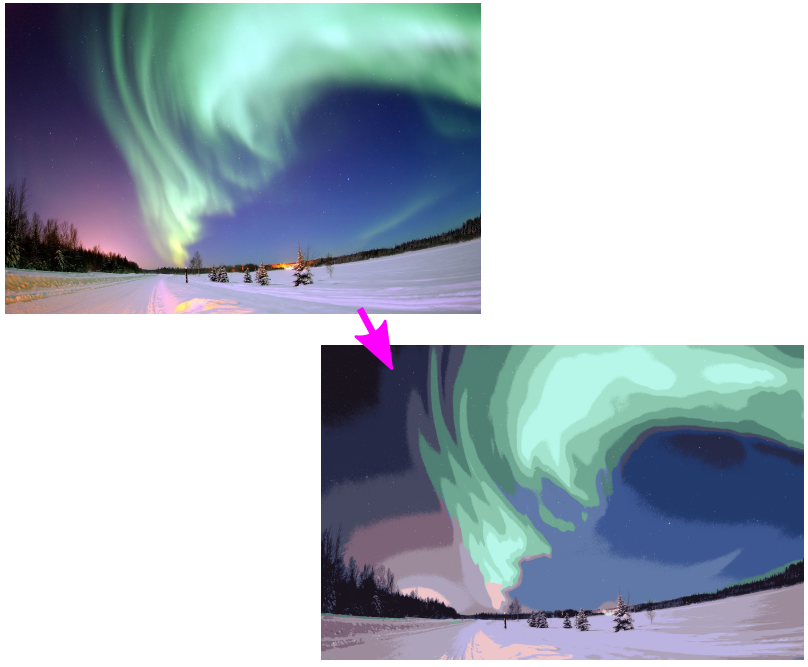
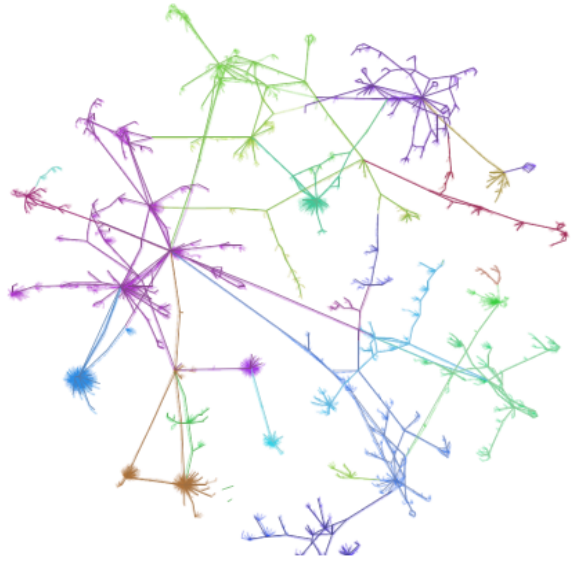


image segmentation



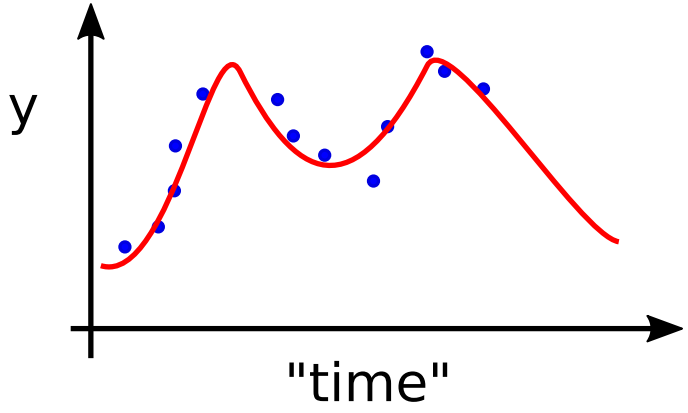
network community detection

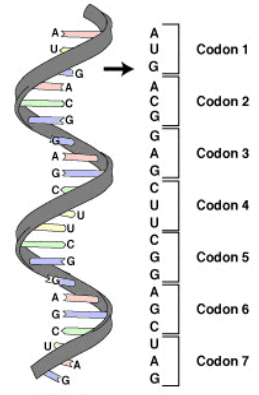


Campbell et al Social Network Analysis

vector quantisation
genetic clustering
anomaly detection
crime analysis

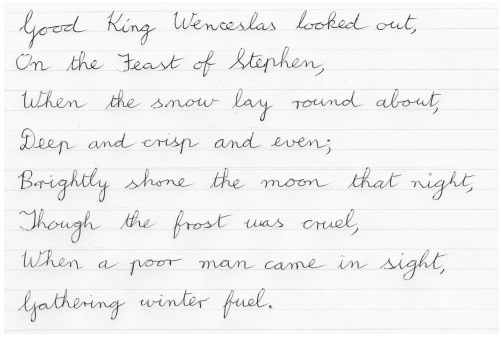
Outline of the course: Sequence modelling

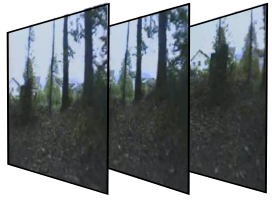


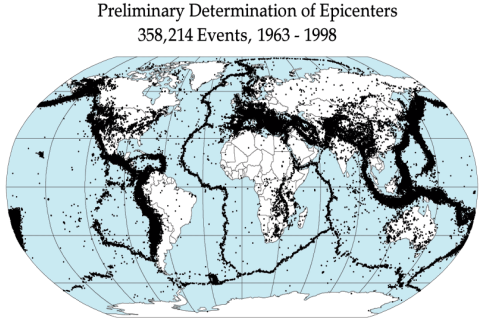


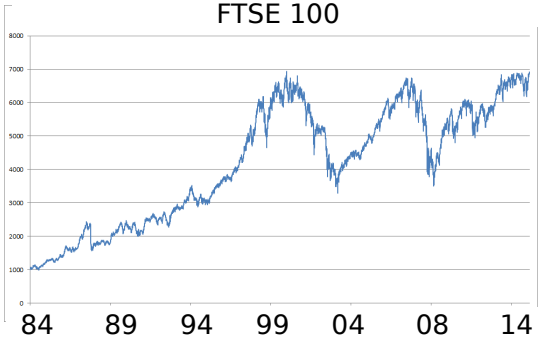
RNA

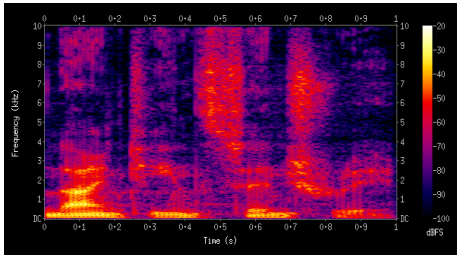
Ribonucleic acid









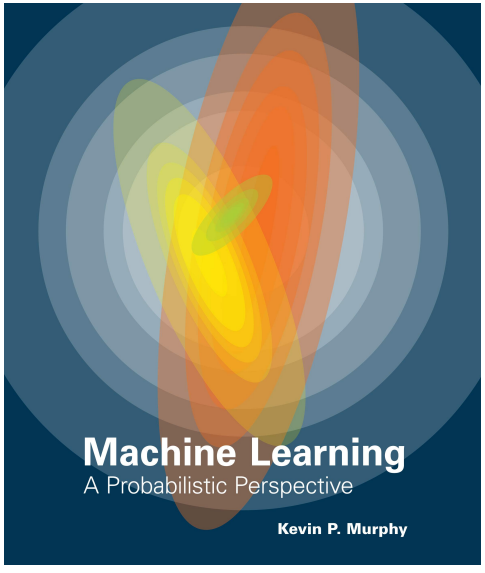


I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted. A. Turing

Course books

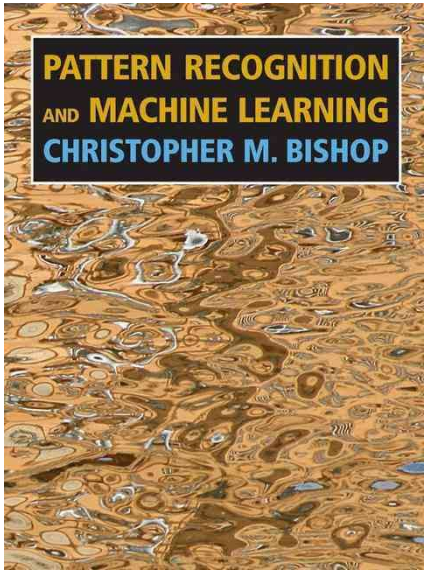
Machine Learning: a Probabilistic Perspective

Kevin Patrick Murphy



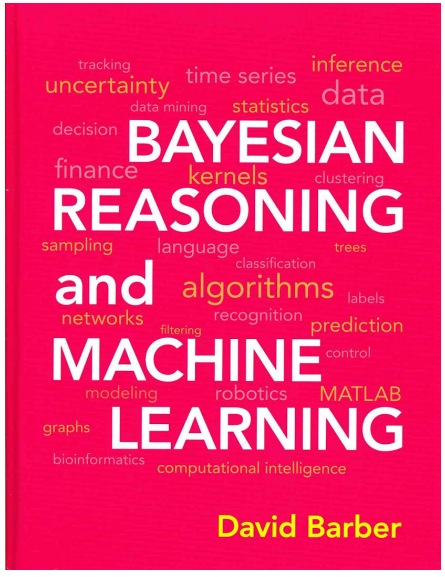
Pattern Recognition and Machine Learning

Christopher Bishop



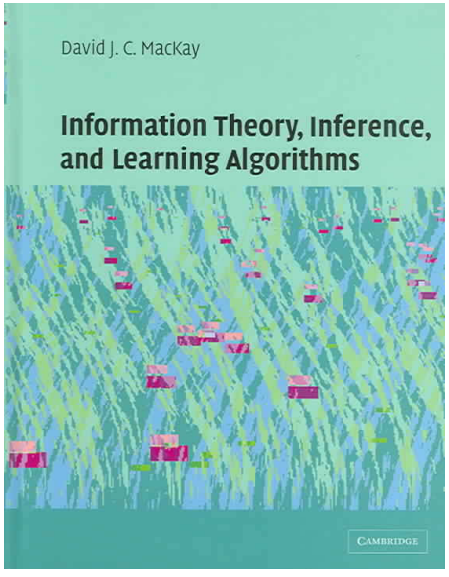
Bayesian Reasoning and Machine Learning

David Barber



Information Theory, Inference, and Learning Algorithms

David JC MacKay



and some examples from: An Introduction to Statistical Learning
Gareth James, Robert Tibshirani, and Trevor Hastie

Supervision plan

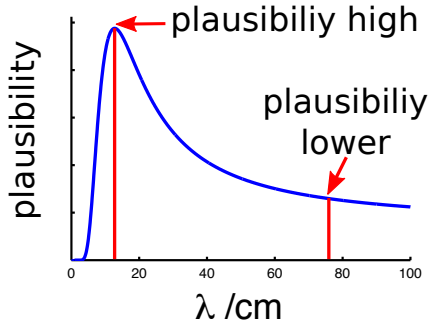
- two sets of **three examples classes** with Miguel and myself, 1hr
  - initial **triage** to catch common problems, go through hardest problems on each of the 3 examples sheets, provide hints
  - debug common questions arising from lectures
  - supervisors encouraged to attend too
  - mechanism for us to gauge level / focus of lectures
- **three small supervisions, plus a revision supervision** pairs / threes with course supervisors, 1hr
  - deal with **specific problems** you have
  - three supervisions on the three examples sheets, one for revision

Supplementary slide on Cox's axioms (non-examinable)

goal of inference: plausiblity of each parameter setting given data

or "degree of belief"

1. plausibility of each parameter setting can be represented by real numbers



2. take into account all evidence

can't leave out some data

3. consistency: if you can reason in more than one way, then each must lead to the same answer

4. equivalent states of knowledge imply same plausibility assignment

Conclusion: degrees of belief follow the rules of probability

product rule: $p(\lambda, \mathrm{x}) = p(\lambda|\mathrm{x})p(\mathrm{x}) = p(\mathrm{x}|\lambda)p(\lambda)$

sum rule: $$p(\mathrm{x}) = \sum_{\lambda} p(\lambda, \mathrm{x})$$