INM431: Machine Learning

Bayesian Inference and Maximum Likelihood Estimation

Pranava Madhyastha (pranava.madhyastha@city.ac.uk)

Based on materials by Artur S. d’Avila Garcez + Bishop’s book + Murphy’s book

The Bayes' theorem

$$P(h \mid \mathscr{D}) = \frac{P(\mathscr{D} \mid h)P(h)}{P(\mathscr{D})}$$

posterior ∝ likelihood × prior

In general, in frequentist ML, we are interested in $P(h \mid \mathscr{D})$, i.e., probability that h is true given the data $\mathscr{D}$

In bayesian learning methods, we have the opportunity to calculate $P(h \mid \mathscr{D})$ using both $P(h)$ and $P(\mathscr{D})$, but computing this is not always tractable

Let us go back to our question in the tutorial

If we only tossed a coin once and got heads:

Is $P(\text{heads}) = 0.5$ or $P(\text{heads}) = 1$?

Is it reasonable to give a single answer?

If we don't have much data, we are unsure about the probability P.

Our computations will work much better if we take this uncertainty into account.

Bayesian framework

It assumes that a prior distribution always exists.

The prior may be very vague

As we observe data, we combine our prior distribution with a likelihood function to get a posterior distribution (and keep applying Bayes' theorem iteratively)

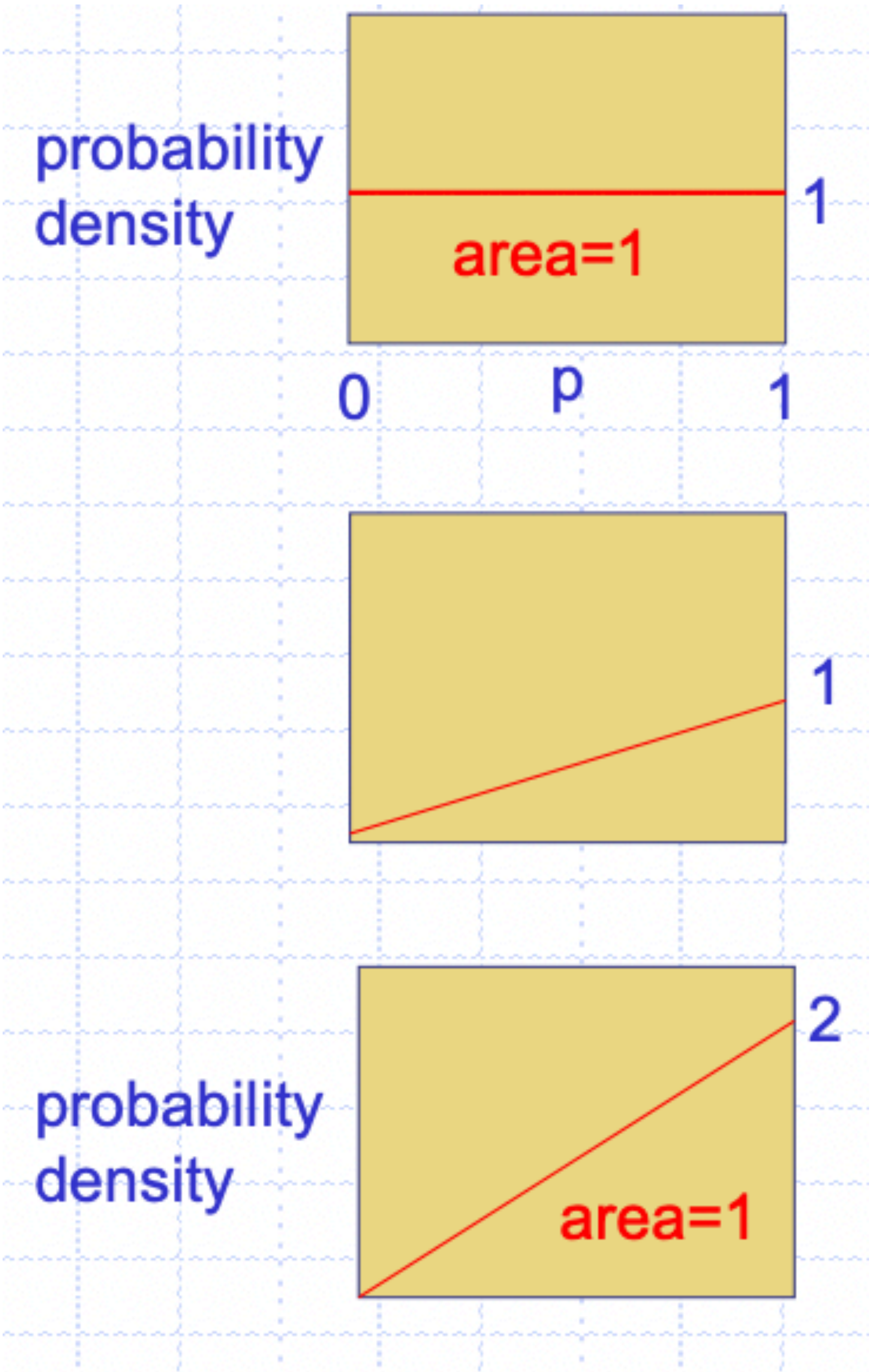
$$P(h \mid \mathscr{D}) = \frac{P(\mathscr{D} \mid h)P(h)}{P(\mathscr{D})}$$

Bayesian framework

Start with a prior distribution over P. In this case we used a uniform distribution

Multiply the prior probability of each parameter value by the probability of observing heads given that value

Then scale up all of the probability densities so that their integral comes to 1. This gives the posterior distribution.



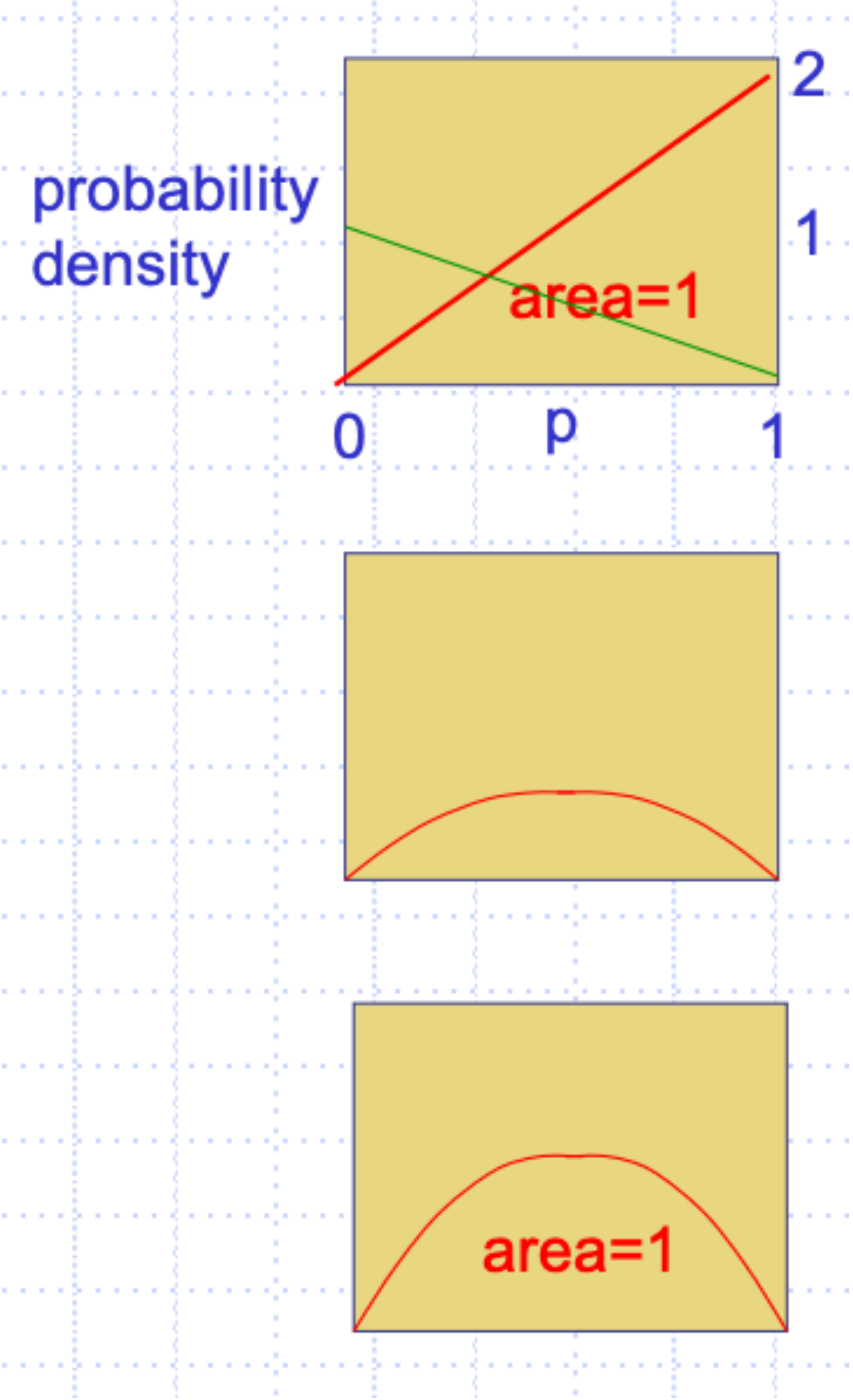
Again

Start with a prior distribution over p.

Suppose you get tails now.

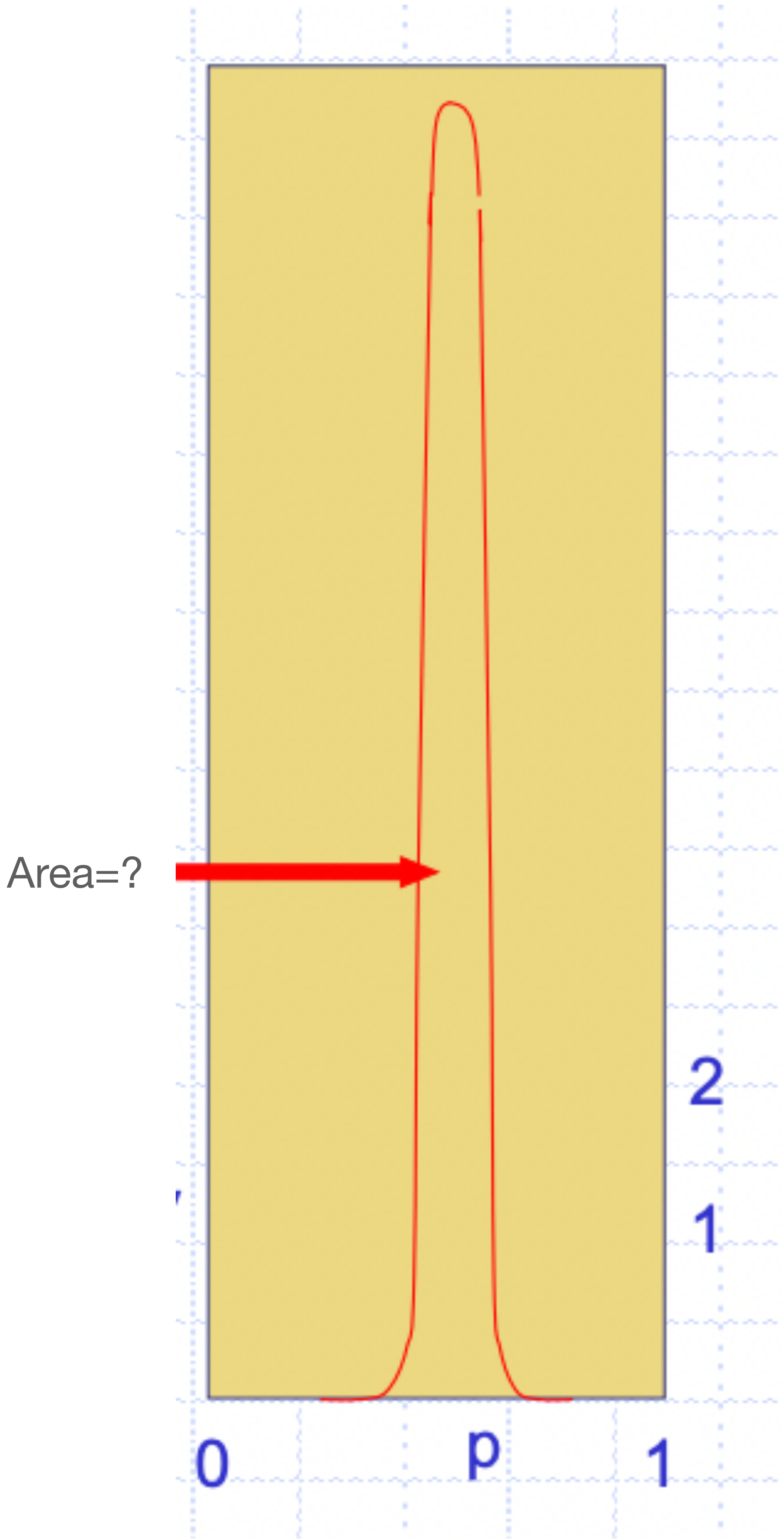
Multiply the prior of each parameter value by the probability of observing tails given that value.

Then re-normalise to get the posterior distribution.

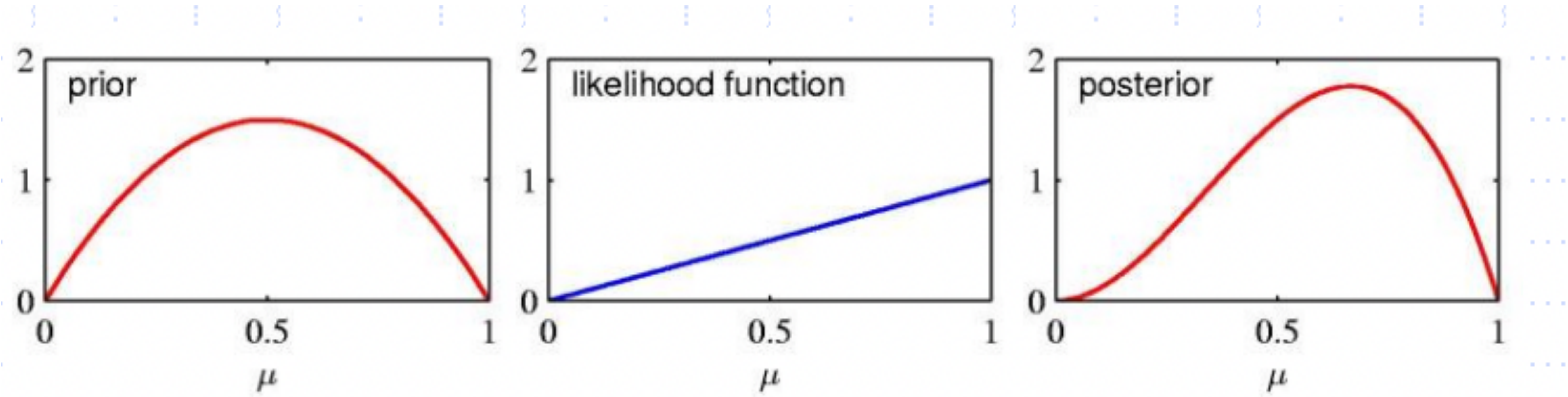


Doing it a hundred times

After 53 heads and 47 tails we get a very sensible posterior distribution that has its peak at 0.53 (having assumed a uniform prior)



The Bayesian Approach



One step of sequential Bayesian inference: the prior is given by a Beta distribution

Beta distribution: constant $\times \mu^{\alpha-1}(1-\mu)^{\beta-1}$

with parameters $\alpha = 2, \beta = 2$ in this case, and the likelihood function corresponds to a single observation of heads.

The posterior is a beta distribution with parameters $\alpha = 3, \beta = 2$.

The parameters a and b of the Beta distribution are called hyperparameters because they control the distribution of the parameter μ.

Density Estimation

We would like to model the probability distribution $P(X)$ of a random variable x given a finite set $D = \{x_1, \cdots, x_n\}$ of observations (data points)

The problem of density estimation is fundamentally ill-posed because there are infinitely many probability distributions that could have given rise to the finite data set observed

Sampling assumption

Assume that the training examples are drawn independently from the set of all possible examples.

Assume that each time a training example is drawn, it comes from an identical distribution (i.i.d)

Assume that the test examples are drawn in exactly the same way: i.i.d. and from the same distribution as the training data.

These assumptions make it very unlikely that a strong regularity in the training data will be absent in the test data.

Maximising log-likelihood

Suppose that the probability of a coin landing heads $(P(x=1)=\mu)$ is not necessarily the same as that of it landing tails $(P(x=0)=1-\mu)$

If the i^{th} observation is heads then $x_i = 1$; otherwise $x_i = 0$

We can construct a likelihood function on the assumption that the observations are drawn independently (i.i.d):

$$P(D\,|\,\mu) = \prod_{n=1}^{N} P(x_n\,|\,\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

We can estimate a value for μ by maximising the log of the likelihood function; if we set the derivative of $\log P(D\,|\,\mu)$ with respect to μ equal to zero, we obtain the maximum likelihood estimator, where m is the number of heads:

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{m}{N}$$

On priors

Without a prior, suppose we flip a coin 3 times and observe 3 heads. Thus, $\mu_{ML} = 1$

If we choose a prior to be proportional to powers of μ and $(1 - \mu)$ then the posterior distribution, which is proportional to the product of the prior by the likelihood function, will have the same functional form as the prior.

Some tricks

Assume that the errors on different training examples d_c are independent.

We want to maximise the product of the probabilities of training examples:

$$P(D \mid \theta) = \prod_c P(d_c \mid \theta)$$

Because the log function is monotonic, it does not change where the maxima are. So we can maximise sums of log probabilities as:

Recall that $\log(xy) = \log(x) + \log(y)$ and $\log(\frac{x}{y}) = \log(x) - \log(y)$, it helps keep the values low by replacing product by sum

$$\log P(D \mid \theta) = \sum_c \log P(d_c \mid \theta)$$

Recall that: $P(\theta \mid D) = \frac{P(D \mid \theta) P(\theta)}{P(D)}$

Loss function: $-\log P(\theta \mid D) = -\log P(D \mid \theta) - \log P(\theta) + \log P(D)$

Some tricks

Ignore the partition function $\log P(D)$ and focus on:

$$-\log P(\theta | D) \propto -\log P(D | \theta) - logP(\theta)$$

Ignore the prior over θ. This is equivalent to giving all possible set of parameters the same prior probability density

Then all we need to do is maximise $\log P(D | \theta)$

This is called Maximum Likelihood (ML) learning. It is widely used for fitting models in statistics

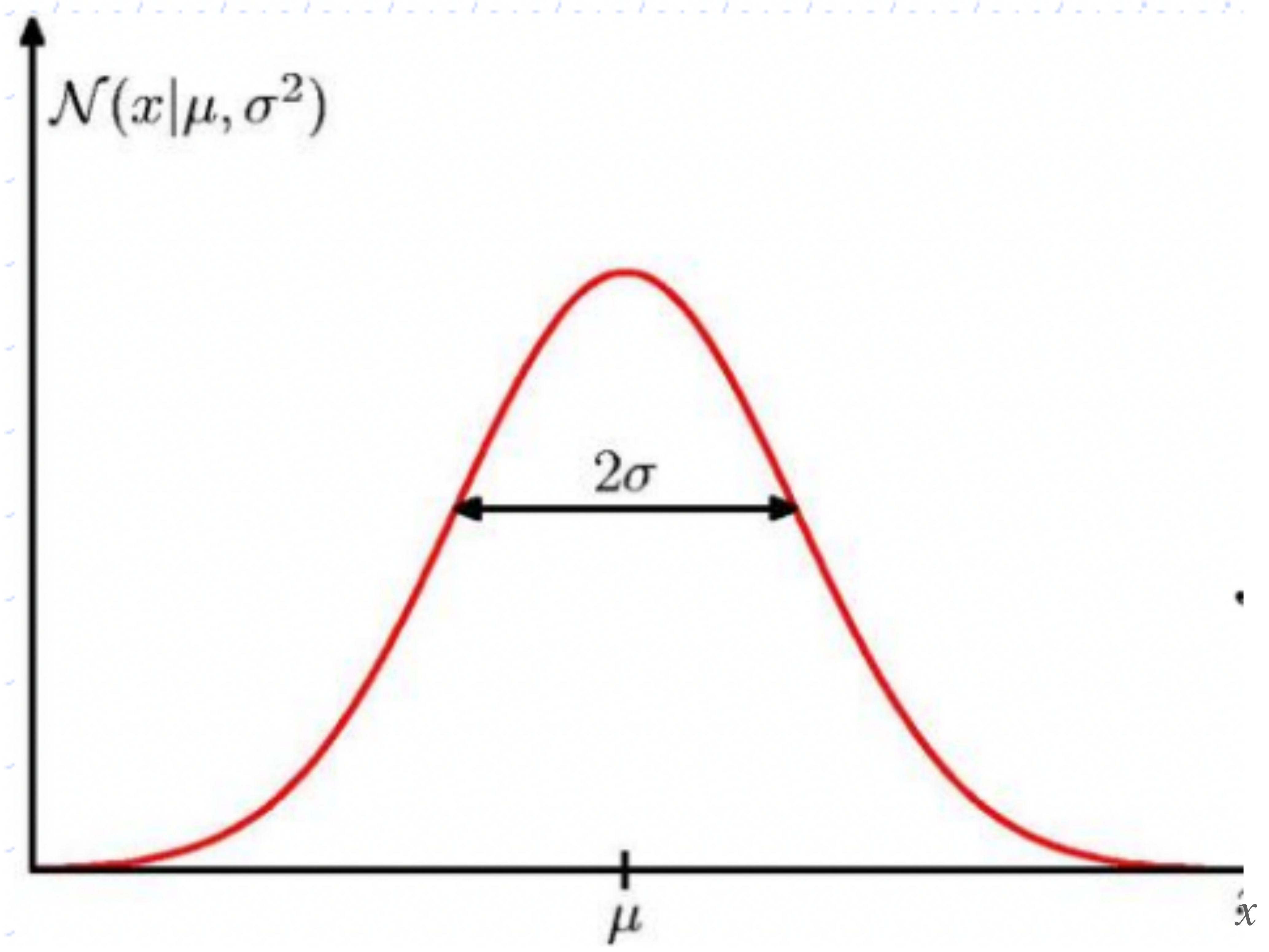
If the derivative of $\log P(D | \theta)$ cannot be calculated analytically, we approximate the computation by starting with a random θ and adjusting it successively in a direction that improves $\log P(D | \theta)$

Gaussian distribution

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

$$\mathcal{N}(x \,|\, \mu, \sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}(x \,|\, \mu, \sigma^2) dx = 1$$



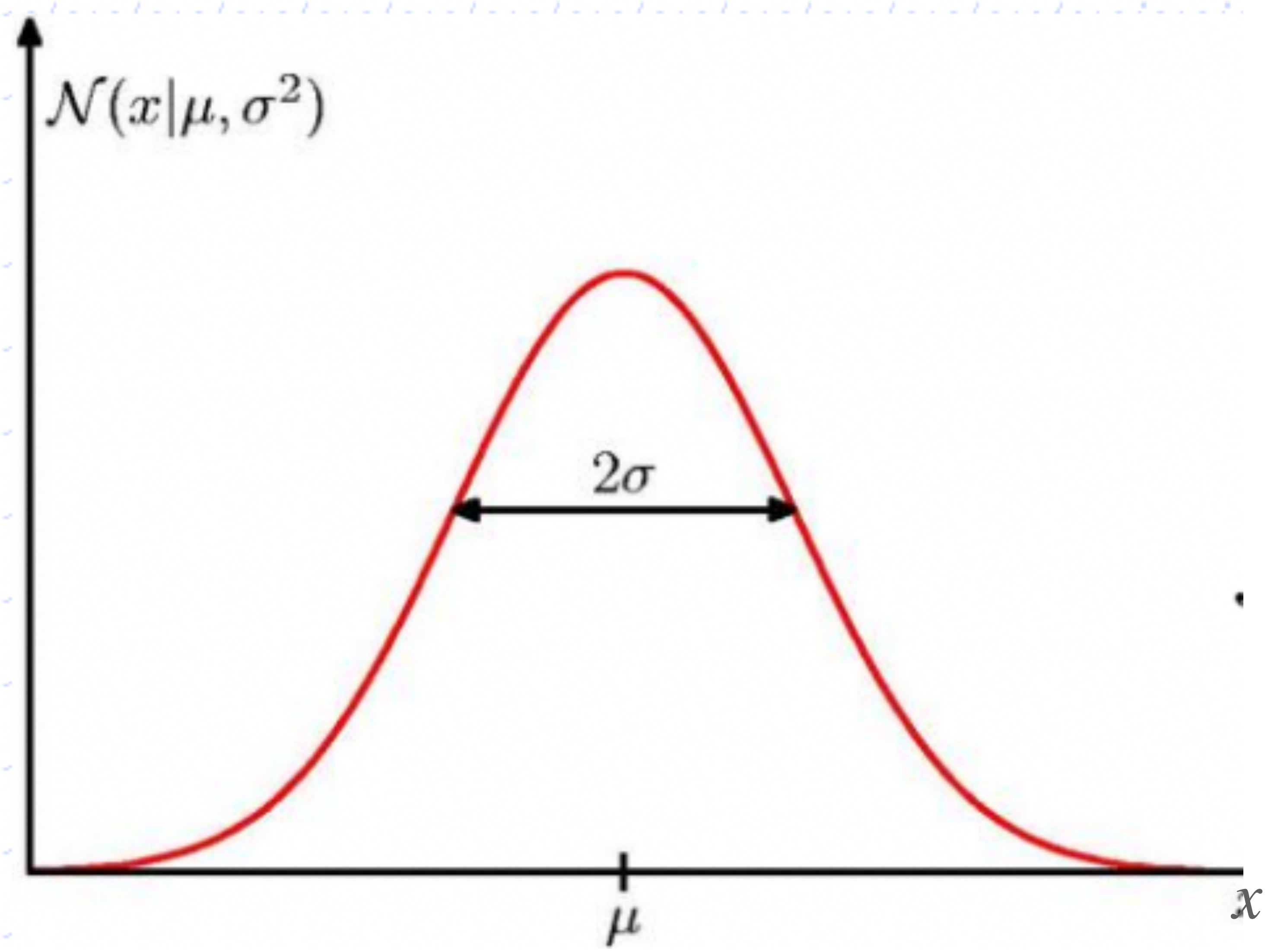
Gaussian mean and variance

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x \mid \mu, \sigma^2) x dx = \mu$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x \mid \mu, \sigma^2) x^2 dx = \mu^2 + \sigma^2$$

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

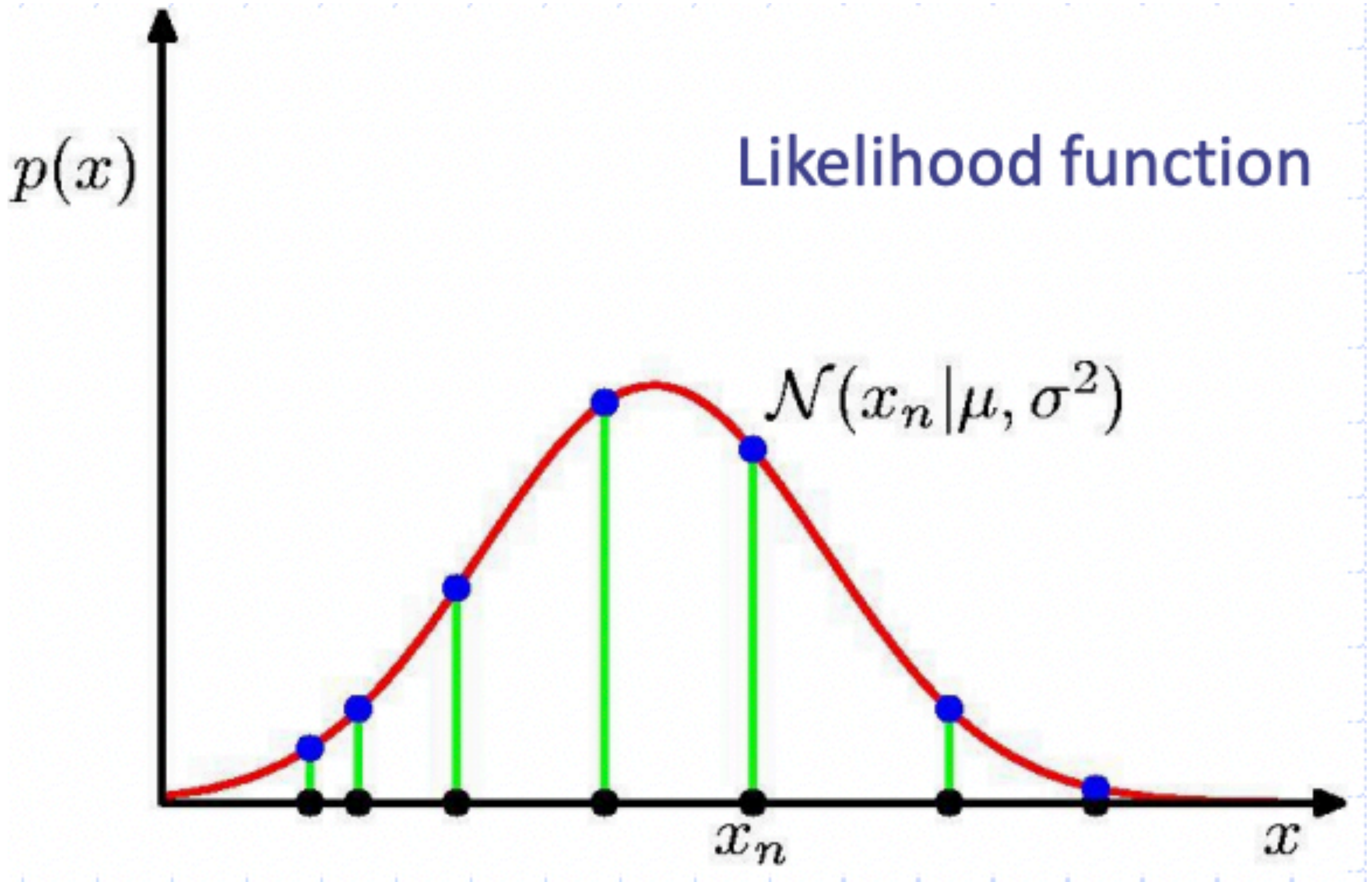
Precision parameter $\beta = \dfrac{1}{\mathrm{var}[x]}$



Gaussian mean and variance

Maximising the likelihood function:

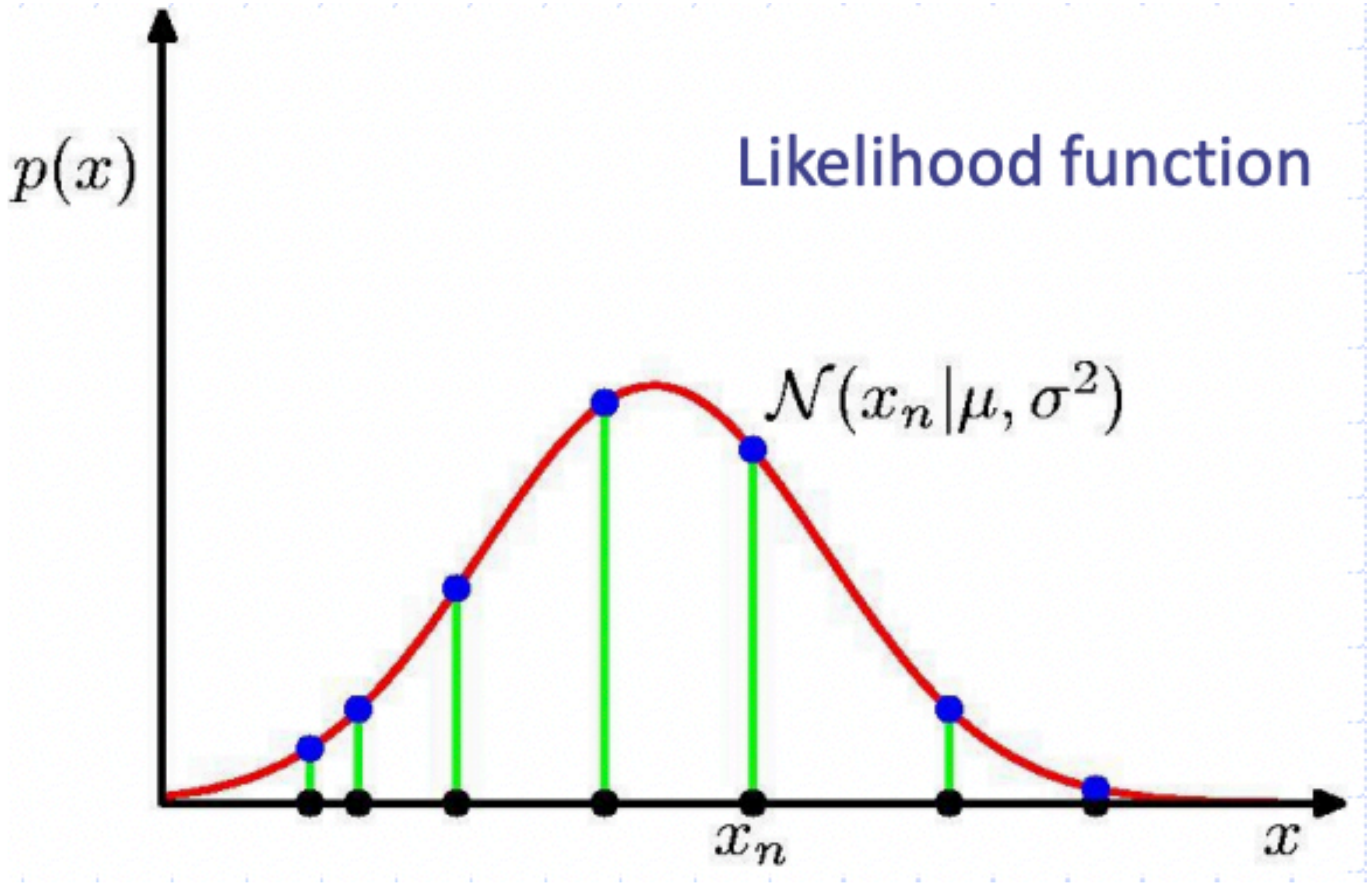
$$P(\mathbf{X} \mid \mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n \mid \mu, \sigma^2)$$



Gaussian mean and variance

We want to determine the values of μ and σ that maximise the log likelihood:

$$\log P(\mathbf{X} \,|\, \mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\log\sigma^2 - \frac{N}{2}\log(2\pi)$$



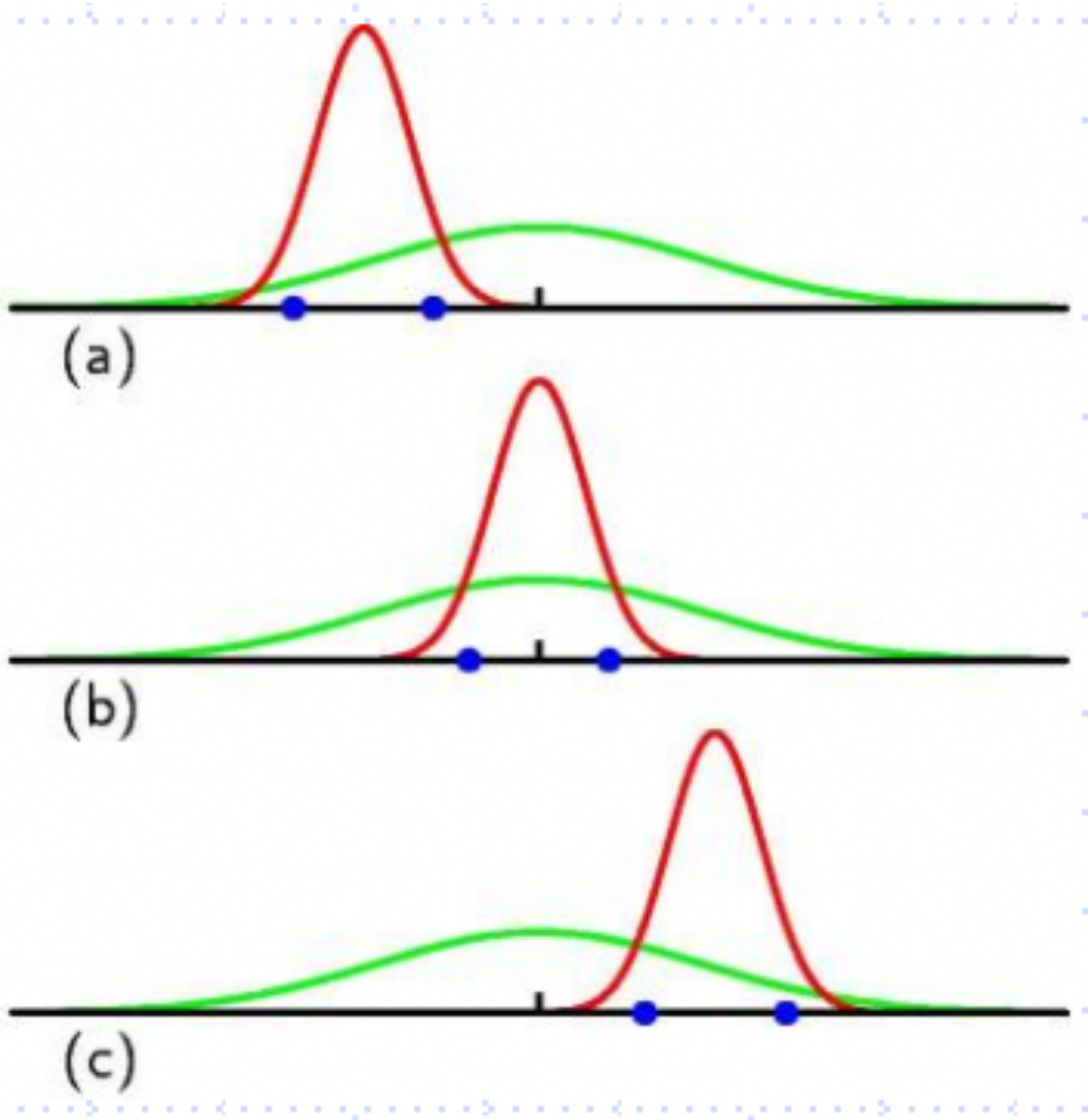
Maximising w.r.t. μ: $\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$

Maximising w.r.t. σ2 : $\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$

Properties of μ_{ML} and σ^2_{ML}

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

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



Hence, on average, maximum likelihood will obtain the correct mean, but will underestimate the variance by a factor $\frac{N-1}{N}$