

Lecture 2: Data, Variables and Distributions -

H1 20/01/20

H2 Binary Variables: Maximum Likelihood

H3 Example: Coin Tossing

Suppose we have a biased coin. At each toss we get:

- Heads with probability μ
- Tails with probability $1 - \mu$

We toss the coin 100 times and observe 53 head and 47 tails. **What is μ ?**

The Frequentist Answer: Pick the value of μ that makes the observations most probable.

H3 Mathematically Encoding

- Each toss is a **binary random variable** X
- X can take two values: 1(head)/0(tail)
- For known, μ , X follows the **Bernoulli Distribution** :

$$p(x|\mu) = p(X = x|\mu) = \mu^x (1 - \mu)^{(1-x)}$$

$p(x|\mu)$ is the plausibility of a probability μ

- Data $D = \{x_1 = 1, x_2 = 1, x_3 = 0, \dots\}$
- **Assume we knew** μ then:

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

Frequentist View: find the maximum of this

H3 Find Maximum

- Differentiate $p(D|\mu)$ is difficult
- But it is easier to find the maximum of $\ln p(D|\mu)$ and if $\mu^* = \arg \max_{\mu} f(\mu)$, then $\mu^* = \arg \max_{\mu} \ln f(\mu)$

Using $f(x) = \prod_{n=1}^N [g(x)]$, $\ln f(x) = \sum_{n=1}^N [\ln g(x)]$

$$\ln p(D|\mu) = \sum_n [x_n \ln \mu + (1 - x_n) \ln(1 - \mu)]$$

Using $f(\mu) = x \ln \mu$, $\frac{d}{d\mu} f = \frac{x}{\mu}$

$$\frac{d}{d\mu} \ln p(D|\mu) = \sum_n \left[\frac{x_n}{\mu} - \frac{1 - x_n}{1 - \mu} \right]$$

Let $\frac{d}{d\mu} \ln p(D|\mu) = 0$

$$\begin{aligned}\sum_n \left[\frac{x_n}{\mu} - \frac{1-x_n}{1-\mu} \right] &= 0 \\ \sum_n \frac{x_n}{\mu} &= \sum_n \frac{1-x_n}{1-\mu} \\ \frac{1}{\mu} \sum_n x_n &= \frac{1}{1-\mu} \sum_n [1-x_n] \\ \frac{1}{\mu} \sum_n x_n &= \frac{N}{1-\mu} - \frac{1}{1-\mu} \sum_n x_n \\ \frac{1}{\mu(1-\mu)} \sum_n x_n &= \frac{N}{1-\mu}\end{aligned}$$

Maximiser:

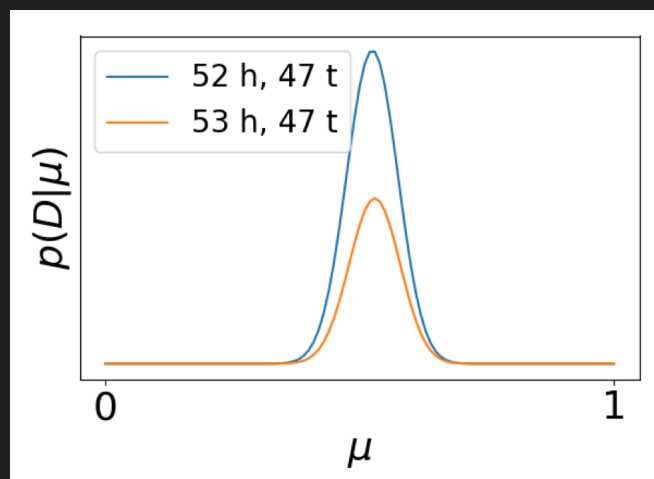
$$\mu_{ML} = \frac{\sum_n x_n}{N}$$

H3 Terminology

- $p(D|\mu)$ is the **joint probability** of D
- $p(D|\mu)$ is also the **likelihood** of μ
- $\ln p(D|\mu)$ is the **log-likelihood** of μ
- μ_{ML} is the **maximum-likelihood parameter**
- As all $x_n \in D$ are drawn independently from the same distribution we say they are **independent and identically distributed** or **i.i.d**

H3 Problems with Maximum-Likelihood

- If we got two heads, $\mu = 1$ is yielded, but not sensible
- μ might not be a single answer
- Insufficient data leads to uncertainty about μ
- Taking uncertainty into account leads to better reasoning



$p(D|\mu)$ is not a probability distribution over μ as the area under the curve isn't 1

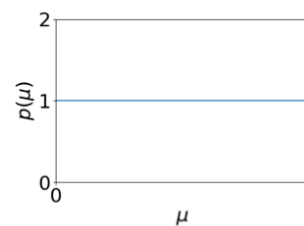
H2 Bayesian Variables: A Bayesian Approach

[如何通俗理解Beta分布 - 知乎](#)

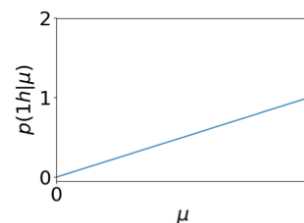
$posterior \propto likelihood \times prior$

H3 Example: first coin is head and second coin is tail

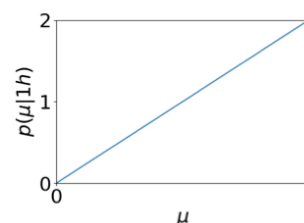
Start with a prior distribution for μ .
Here we use the uniform distribution.



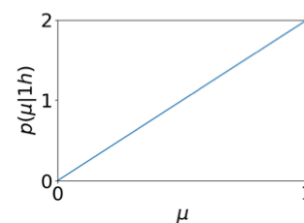
Likelihood $p(1h|\mu)$ charts probability of a head given μ .



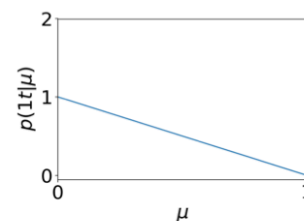
For each μ (pointwise) multiply $p(\mu)$ with $p(1h|\mu)$, then rescale (normalise) so the integral sums to 1.
This is the posterior probability density.



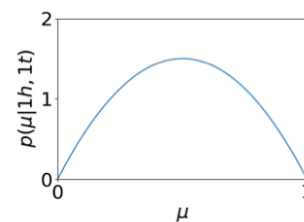
The new prior is $p(\mu|1h)$
(the posterior from the previous toss).



Likelihood $p(1t|\mu)$ charts probability of a tail given μ .

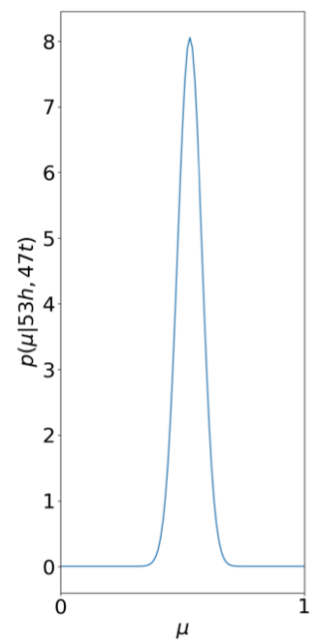


Pointwise multiply prior and likelihood then renormalise to get the posterior.



Let's do it another 98 times:

After 53 heads and 47 tails we get a very sensible looking distribution, with a peak (and expectation) both at $\mu \approx 0.53$.



H2 The Beta Distribution

The beta distribution describes continuous random-variables in the range $[0, 1]$ and has the form:

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{(a-1)} (1-\mu)^{(b-1)}$$

$$\Gamma(n) = (n-1)! \forall n \in \mathbb{N}^+$$

Ignore the normalisation:

$$\text{Beta}(\mu|a, b) \propto \mu^{(a-1)} (1-\mu)^{(b-1)}$$

Mean:

$$E(\mu|a, b) = \frac{a}{a+b}$$

Variance:

$$\text{var}(\mu|a, b) = \frac{ab}{(a+b)^2(a+b+1)}$$

H3 A Beta Prior

We choose a **Beta prior** $p(\mu) = \text{Beta}(\mu|a, b) = \mu^{(a-1)} (1-\mu)^{(b-1)}$ for our coin then observe m heads and l tails

For uncluttered notation, we sometimes write $p(\mu) = p(\mu|a, b)$

Using Bayes Theory our **posterior** looks like:

$$\begin{aligned}
p(\mu|D) &\propto p(D|\mu)p(\mu) \\
&\propto \mu^m (1-\mu)^l \mu^{(a-1)} (1-\mu)^{(b-1)} \\
&= \mu^{(m+a-1)} (1-\mu)^{(l+b-1)}
\end{aligned}$$

Once correctly **normalised**:

$$p(\mu|D) = \text{Beta}(\mu|m+a, l+b)$$

New observation added to the experience data

When our posterior takes the same form as the prior, then the prior is said to be **conjugate**

H3 Conjugate Beta Priors

With Beta prior $\text{Beta}(\mu|a, b)$ we observe m heads and l tails.

Prior Estimate: $E[\mu|a, b] = \frac{a}{a+b}$

Posterior Estimate: $E[\mu|a, b, m, l] = \frac{a+m}{a+m+b+l}$

Recall that $\text{Beta}(\mu|a, b) \propto \mu^{(a-1)} (1-\mu)^{(b-1)}$

- Can interpret a and b as **effective prior observations**
- $a = 1$ and $b = 1$ give a **flat prior** ($p(\mu)$ constant)
- a and b must be greater than 0
- a and b don't necessarily need to be integers

H2 Reak Valued Data

Another form of data we deal with regularly is unbounded reals ($x_n \in \mathbb{R}$). Can often model this with a **Gaussian**:

- the Gaussian has many nice properties (as well will see)
- reasons to expect data to be (approximately) Gaussian

a Gaussian prior can induce a Gaussian posterior (conjugacy)

- we can test data for its **Gaussianity**

H2 The Gaussian (Normal) Distribution

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

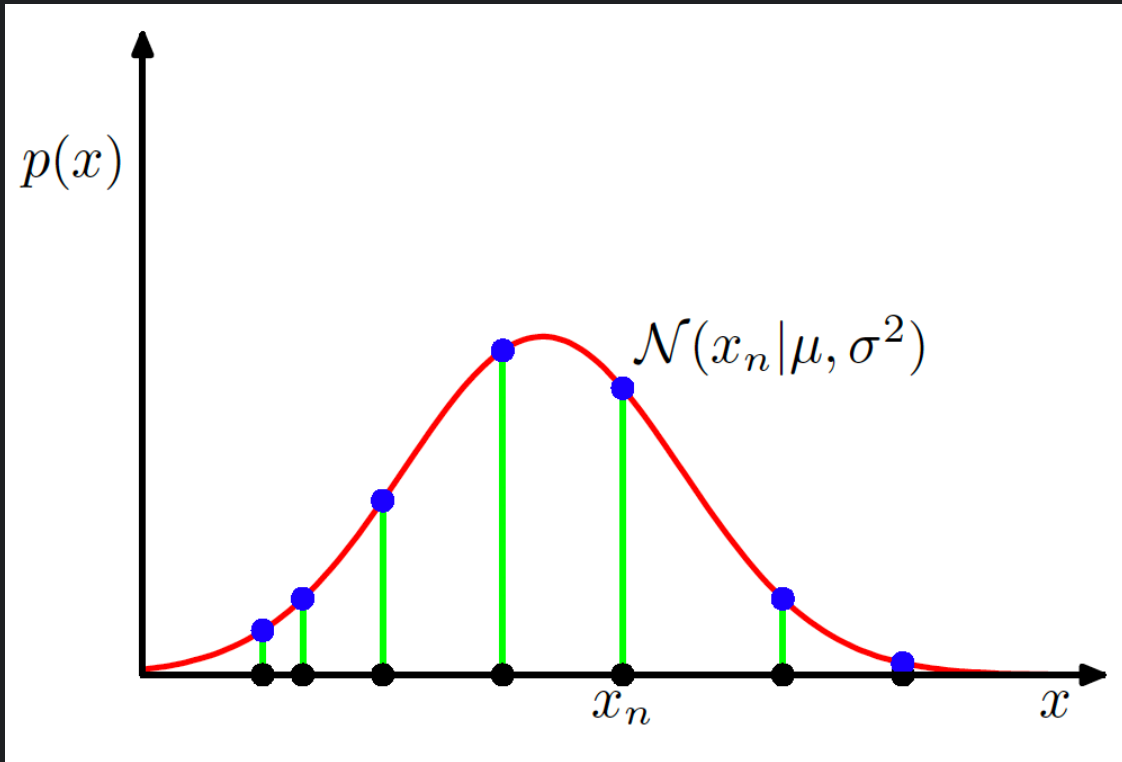
for real values random samples $x_n \in \mathbb{R}$

Properties:

- $N(x|\mu, \sigma^2) > 0$ for $x \in \mathbb{R}$
- $\int_{-\infty}^{\infty} N(x|\mu, \sigma^2) dx = 1$
- $E[x] = \mu$
- $\text{var}[x] = \sigma^2$

H2 Gaussian Likelihood

If we draw N samples $\mathbf{x} = (x_1, \dots, x_N)^T$ *i.i.d* from our Gaussian, e.g. $x_n \sim N(x|\mu, \sigma^2)$



The *likelihood* is:

$$\begin{aligned}
 p(\mathbf{x}|\mu, \sigma^2) &= \prod_{n=1}^N N(x_n|\mu, \sigma^2) \\
 &= \prod_{n=1}^N \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x_n - \mu)^2\right] \right] \\
 &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^N \exp\left[-\sum_{n=1}^N \frac{1}{2\sigma^2}(x_n - \mu)^2\right] \\
 &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^N \exp\left[-\frac{1}{2\sigma^2} \left[\sum_n x_n^2 - \sum_n 2\mu x_n + N\mu^2 \right] \right]
 \end{aligned}$$

Maximum likelihood parameters are a pair of values $\mu = \mu_{ML}, \sigma^2 = \sigma_{ML}^2$ that maximises the likelihood.

Easier to find maximisers using *log likelihood*:

$$\begin{aligned}
 \ln p(\mathbf{x}|\mu, \sigma^2) &= N \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \ln\left(\exp\left[-\frac{1}{2\sigma^2} \left[\sum_n x_n^2 - \sum_n 2\mu x_n + N\mu^2 \right] \right]\right) \\
 &= -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left[\sum_n x_n^2 - \sum_n 2\mu x_n + N\mu^2 \right] \\
 &= -\frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2
 \end{aligned}$$

Differentiate $\ln p(\mathbf{x}|\mu, \sigma^2)$ and set to zero:

$$\begin{aligned}
\frac{d}{d\mu} \ln p(\mathbf{x}|\mu, \sigma^2) &= -\frac{1}{2\sigma^2} \sum_{n=1}^N [-2x_n + 2\mu] \\
\text{let } \frac{d}{d\mu} \ln p(\mathbf{x}|\mu, \sigma^2) &= 0 \\
-\frac{1}{2\sigma^2} \sum_{n=1}^N [-2x_n + 2\mu] &= 0 \\
\sum_{n=1}^N \mu &= \sum_{n=1}^N x_n \\
\mu &= \frac{1}{N} \sum_{n=1}^N x_n \\
\frac{d}{d\sigma^2} \ln p(\mathbf{x}|\mu, \sigma^2) &= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{n=1}^N (x_n - \mu)^2 \\
\text{let } \frac{d}{d\sigma^2} \ln p(\mathbf{x}|\mu, \sigma^2) &= 0 \\
\frac{N}{2\sigma^2} &= \frac{1}{2\sigma^4} \sum_{n=1}^N (x_n - \mu)^2 \\
\sigma^2 &= \frac{1}{N} \sum_{n=1}^N (x_n - \mu)^2
\end{aligned}$$

Thus

$$\begin{aligned}
\mu_{ML} &= \frac{1}{N} \sum_{n=1}^N x_n \\
\sigma_{ML}^2 &= \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{ML})^2
\end{aligned}$$

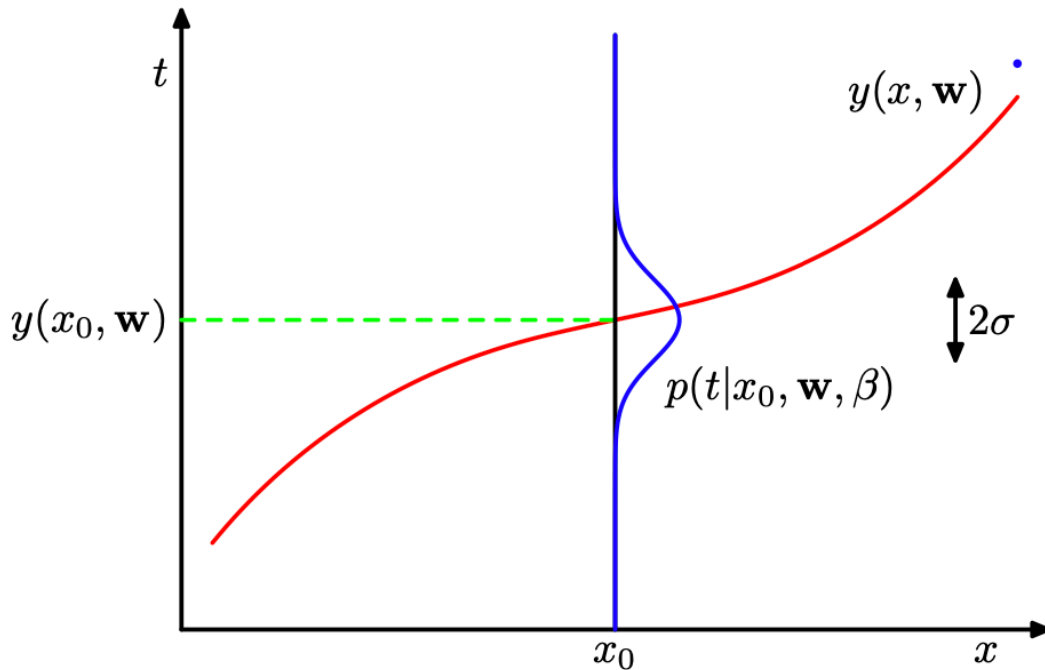
H2 Maximum Likelihood Curve Fitting

Why assume Gaussian? Gaussian Distribution is a convenient default assumption, it makes the maths easy

M - order

Like saying: $t = y(\mathbf{x}; \mathbf{w}) + \epsilon$

where: $\epsilon \sim \mathcal{N}(.|0, \beta^{-1})$.



Using the curve fitting example from **Lecture 1**:

- N inputs $\mathbf{x} = (x_1, \dots, x_N)^T$
- N targets $\mathbf{t} = (t_1, \dots, t_N)^T$

Assume given x_i , then t_i **Gaussian** with mean $y(x_i; \mathbf{w})$ (a polynomial with weights \mathbf{w}), i.e.

$$p(t_i|x_i, \mathbf{w}, \beta) = N(t_i|y(x_i; \mathbf{w}), \beta^{-1})$$

Note that: $f(x; a, b, c)$ means x is the independent variable while a, b, c are the parameters

Where

$$y(x_i; \mathbf{w}) = \sum_{j=0}^M w_j x_i^j$$

In which M is the order of the hypothesis function.

β is the **precision** (inverse variance), $\beta^{-1} = \sigma^2$

Using $\{\mathbf{x}, \mathbf{t}\}$ to find the maximum likelihood parameters for \mathbf{w} and β . The likelihood:

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}) = \prod_{n=1}^N N(t_n|y(x_n; \mathbf{w}), \beta^{-1})$$

Note that: for this Gaussian distribution, the mean is $y(x_n; \mathbf{w})$ and the variance β^{-1} , and random variable being t_n

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}) = -\frac{\beta}{2} \sum_{n=1}^N [y(x_n; \mathbf{w}) - t_n]^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$$

which, for \mathbf{w} , is the same as minimising:

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N [y(x_n, \mathbf{w}) - t_n]^2$$

H3 Insights

- maximising likelihood (minimising error) gives \mathbf{w}_{ML} (independent of β)
- can then find β_{ML}
- now have predictive distribution for new samples t given x :
 $p(t|x, \mathbf{w}_{ML}, \beta_{ML}) = N(t|y(x; \mathbf{w}_{ML}), \beta_{ML})$, this measures uncertainty over t
- doesn't measure uncertainty in \mathbf{w} or β

H2 Gaussian: A More Bayesian View

H3 Moving Towards a More Bayesian View

Assume samples $\mathbf{x} = (x_1, \dots, x_n)^T$ from $N(x_i|\mu, \sigma^2)$

- Assume we know the variance σ^2
- Define prior on $\mu, p(\mu)$
- But what should the **prior** look like?

Likelihood:

$$\begin{aligned} p(\mathbf{x}|\mu) &= \prod_{n=1}^N N(x_n|\mu, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}^N} \exp\left[-\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2\right] \end{aligned}$$

Since the prior and the posterior should be conjugate

Using **Bayes Rule**, the posterior:

$$p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu)$$

and the **likelihood** has form:

$$p(\mathbf{x}|\mu) \propto \exp\left[-\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2\right]$$

This is an exponential that is (-ve) quadratic in μ since the index is
 $-\frac{1}{2\sigma^2} [\sum_n x_n^2 - \sum_n 2\mu x_n + N\mu^2]$

So $p(\mu)$ should also have similar form:

$$\begin{aligned} p(\mu) &\propto \exp\left[-\frac{(\mu - m_0)^2}{2s_0^2}\right] \\ &\propto N(\mu|m_0, s_0^2) \end{aligned}$$

Prior is a **Gaussian**

From conjugacy, we know the **posterior** will also be Gaussian, so

$$p(\mu|\mathbf{x}) = N(\mu|m_N, s_N^2)$$

By working through the maths:

$$m_N = \frac{\sigma^2}{Ns_0^2 + \sigma^2}m_0 + \frac{Ns_0^2}{Ns_0^2 + \sigma^2}\mu_{ML}$$

$$\frac{1}{s_N^2} = \frac{1}{s_0^2} + \frac{N}{\sigma^2}$$

where μ_{ML} is the **maximum likelihood mean** from before:

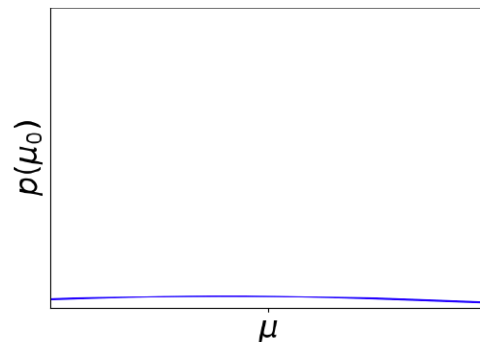
$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^N x_n$$

H3 Insight

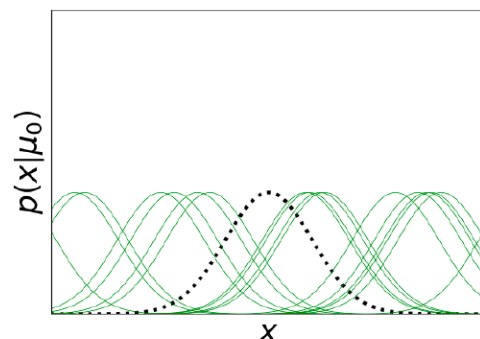
- m_N is compromise between m_0 and μ_{ML}
- For large N , $m_N \approx \mu_{ML}$ (refer to **Lecture 1**)
- Precisions, e.g. $\frac{1}{\sigma^2}$, more natural than variance
- For large N , variance of estimate vanishes, i.e. $s_N^2 \approx 0$
- As $s_0^2 \rightarrow \infty$, then $s_N^2 \rightarrow \frac{\sigma^2}{N}$

H2 Machine Learning Application

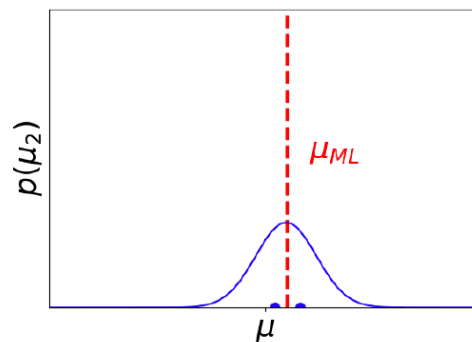
- Initially, we have a broad prior $p(\mu)$
- So we are uncertain about μ



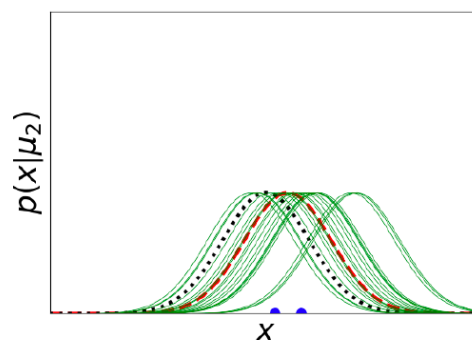
- Means we're uncertain about $p(x|\mu)$ – a new x
- Posterior samples in green
- True $p(x|\mu)$ shown in black
- Remember: we know σ^2



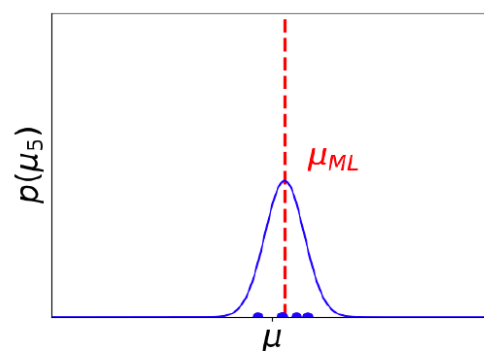
- After observing 2 samples, we know more about μ
- Captured by $p(\mu|\mathbf{x})$



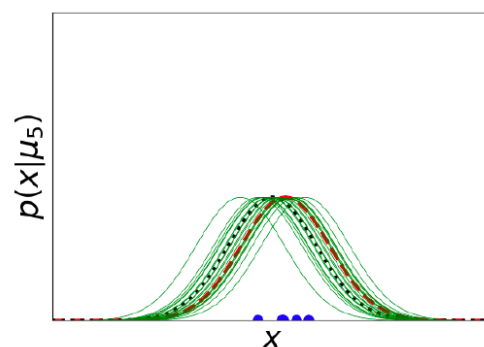
- Meaning we are a little more sure about $p(x|\mu)$
- $p(x|\mu_{ML})$ shown in red



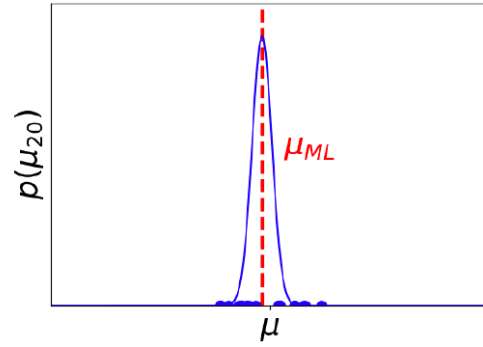
- Adding 3 more samples narrows our uncertainty about μ
- $p(\mu|\mathbf{x})$ more peaked around μ_{ML}



- This makes us more certain about $p(x|\mu)$
- But there is still uncertainty not captured by μ_{ML}



- Another 15 samples and $p(\mu|\mathbf{x})$ is sharply peaked around μ_{ML}



- More certain now about $p(x|\mu)$ (and accurate)
- μ_{ML} is closer to the true mean too

