Week 4: Model Fitting

Bayesian Statistics for Machine Learning

Dr. Daniel Worrall

AMLab, University of Amsterdam

September 27, 2019

I: Maximum likelihood

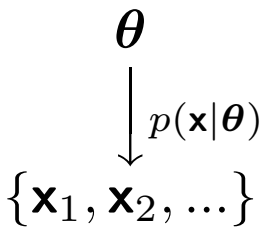
Probability and Statistics

We are mostly concerned with models which look like

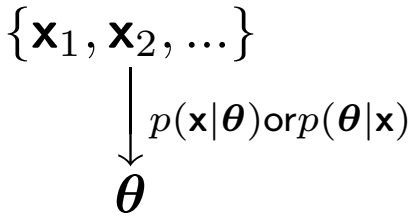
$$p(\mathbf{x}|\boldsymbol{\theta}).$$

In many case $\mathbf{x}$ refers to an *observation* and $\boldsymbol{\theta}$ refers to a set of *parameters*.

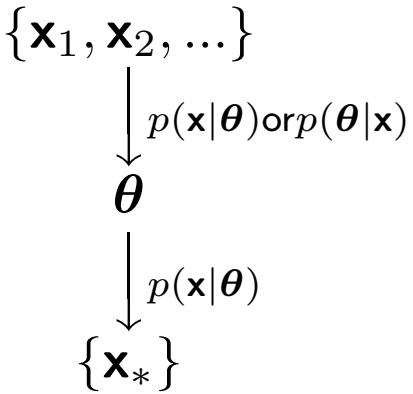
Probability



Statistics

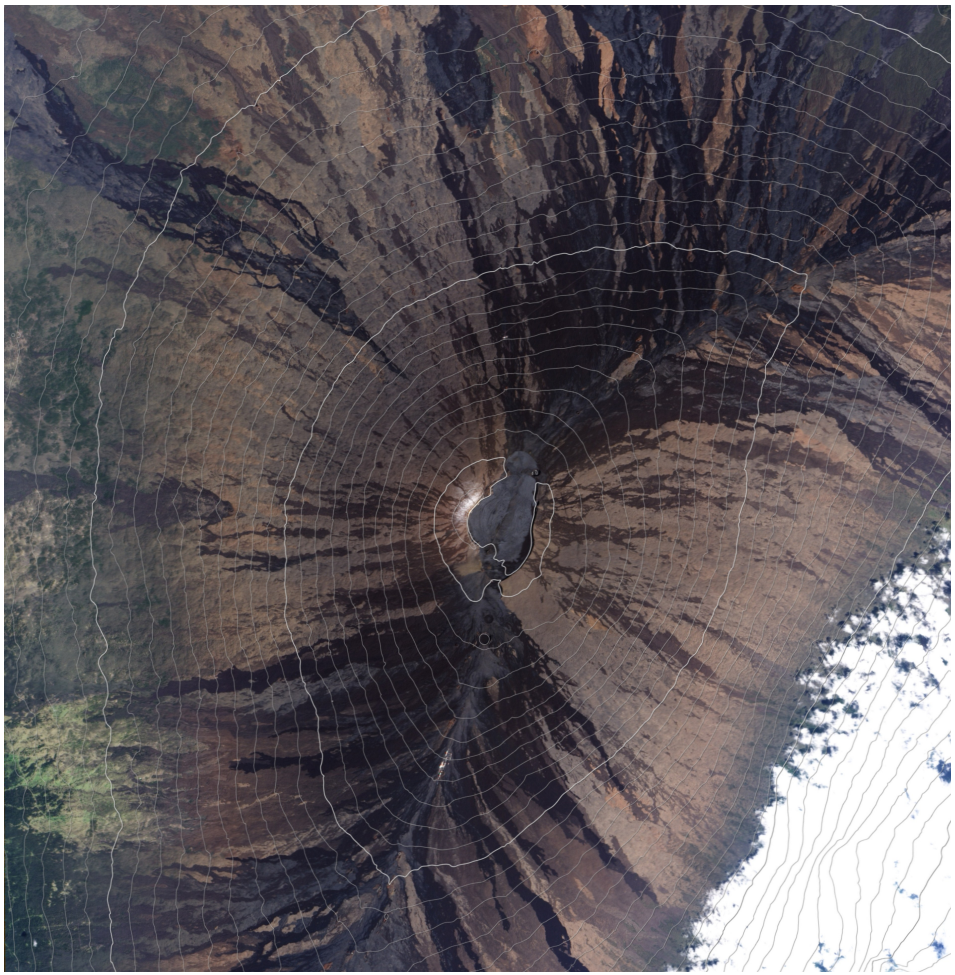


Machine Learning



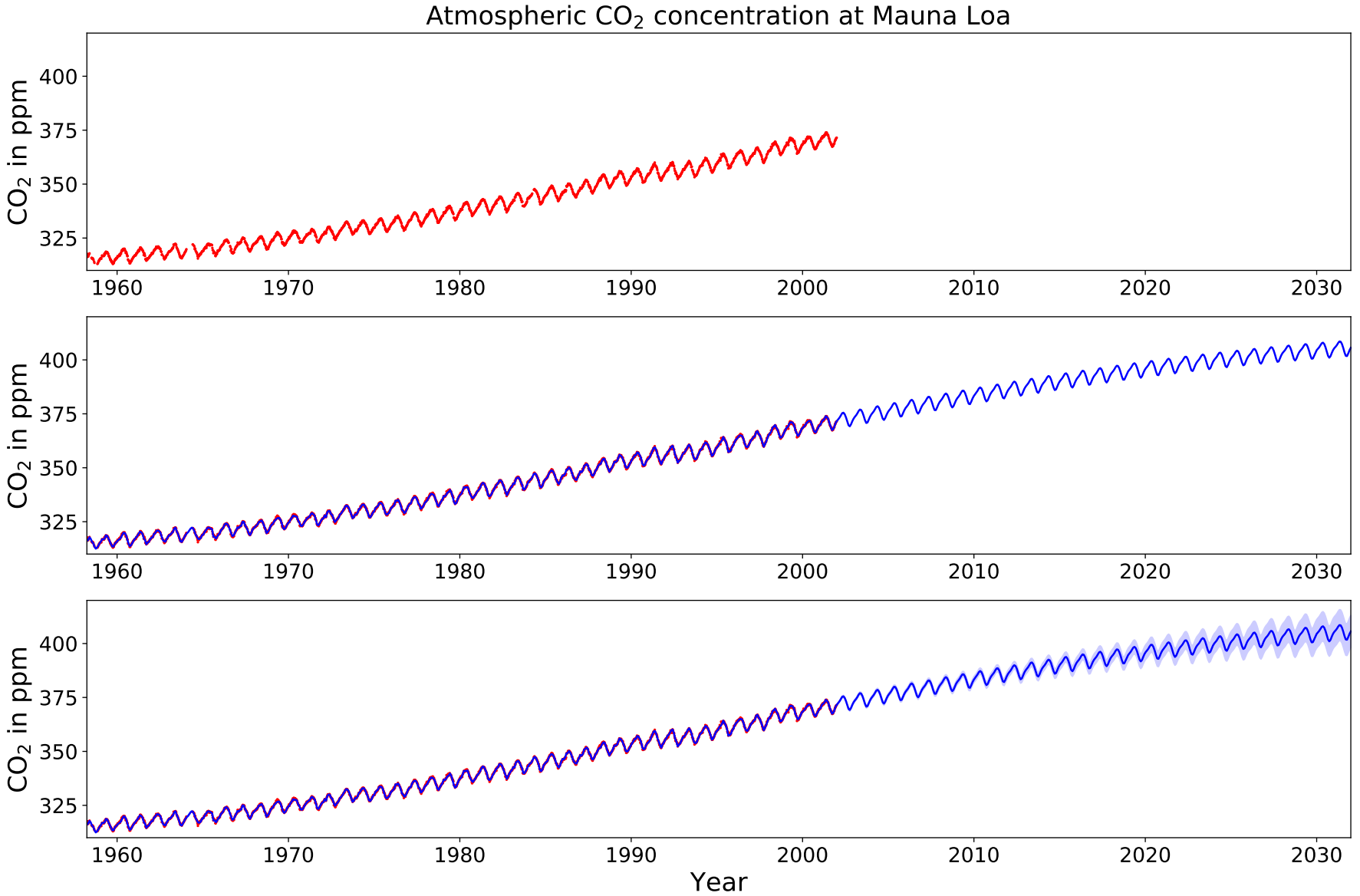
Machine learning

Mauna Loa is one of five volcanoes that form the Island of Hawaii in the U.S. state of Hawaii in the Pacific Ocean. The largest subaerial volcano in both mass and volume, Mauna Loa has historically been considered the largest volcano on Earth, dwarfed only by Tamu Massif.



Machine learning

In machine learning, we use past data to make predictions about the future.



How would we model the Mauna Loa CO_2 levels as a function of time? Perhaps

$$y = w_1 x + w_2?$$

Well this is a straight line, we need an extra periodic component, so how about

$$y = w_1 x + w_2 \cos(2\pi x + w_3) + w_4?$$

But this is no good, because y is negative for certain values of x...

The reality of modelling

All models are wrong!

George Box, 1976

- So how do we choose a model? (Week 6)
- How do I tell if one model is better than another? (Now - Week 6)

The Likelihood

Rutger goes to the data store wanting to buy some data. He decides to buy N samples of univariate Gaussian data

$$\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}.$$

When he gets home he realizes he forgot to ask for a receipt, and cannot remember exactly which Gaussian the data was drawn from.

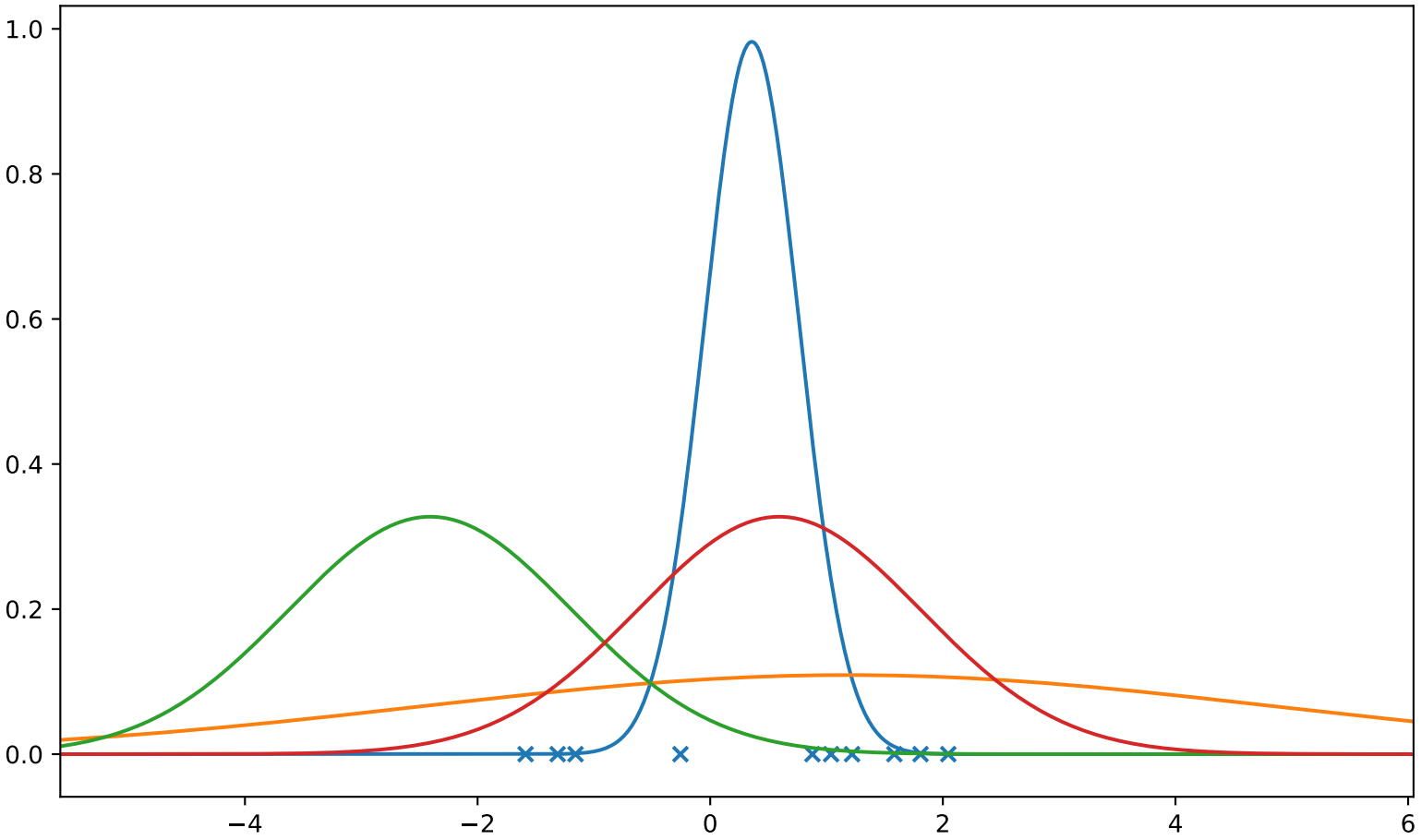
How can he figure out the parameters of the original Gaussian?

Recall

$$\boxed{p(\mathbf{x}|\boldsymbol{\theta}) = \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}}$$

The Likelihood

Rutger first makes a couple of guesses. Which of theses seems likely?



The Likelihood

Rutger surmises

- the hump of the Gaussian should be roughly centered on the data
- the hump should be wide enough so no datapoint has tiny probability.

He decides to look at the probability that the data was generated by a model with parameters $\{\mu, \sigma^2\}$. The best parameters will maximize this probability:

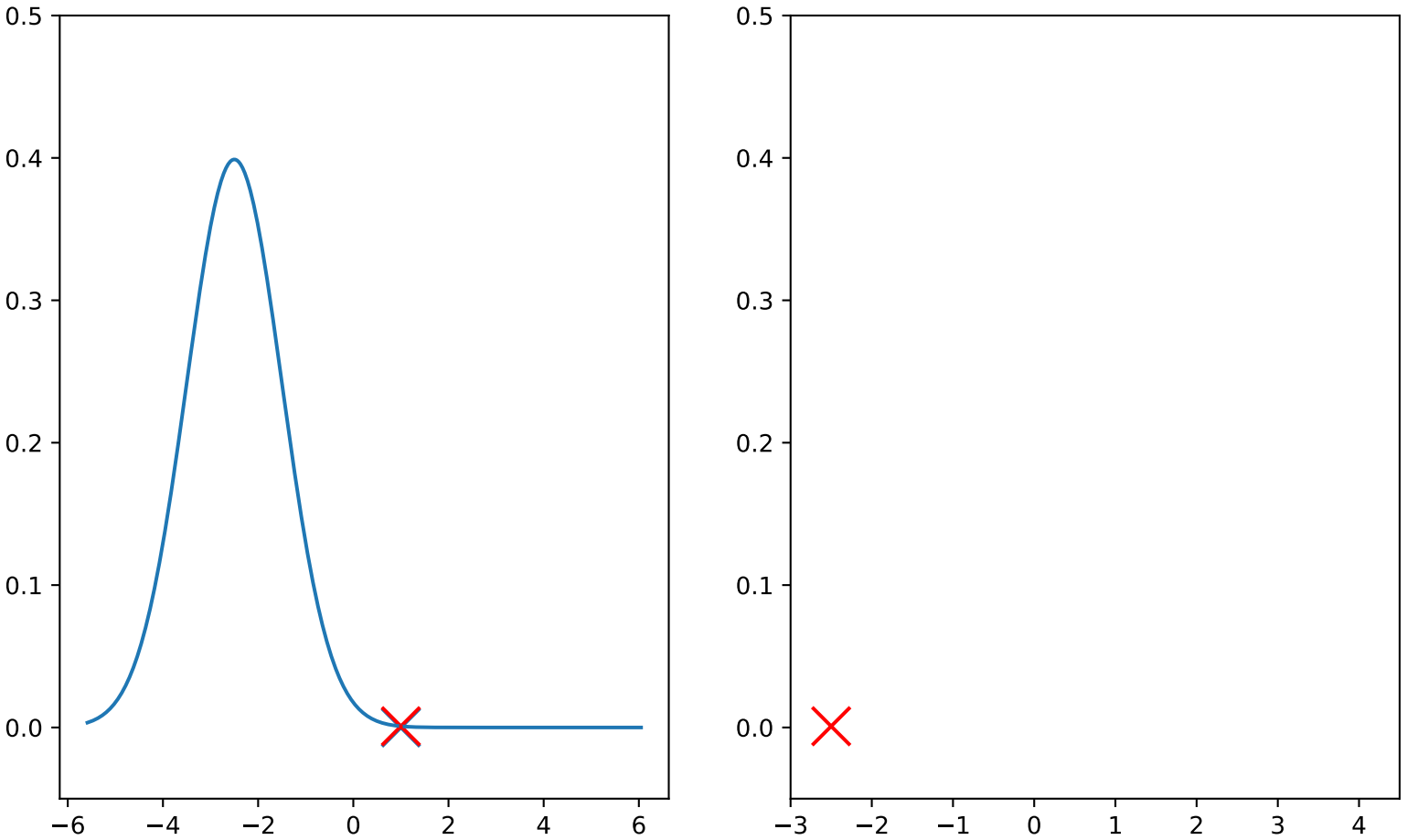
$$\hat{\mu}, \hat{\sigma} = \arg\max_{\mu,\sigma} p(\mathcal{D}|\mu, \sigma^2)$$

But what does $p(\mathcal{D}|\mu, \sigma^2)$ look like?

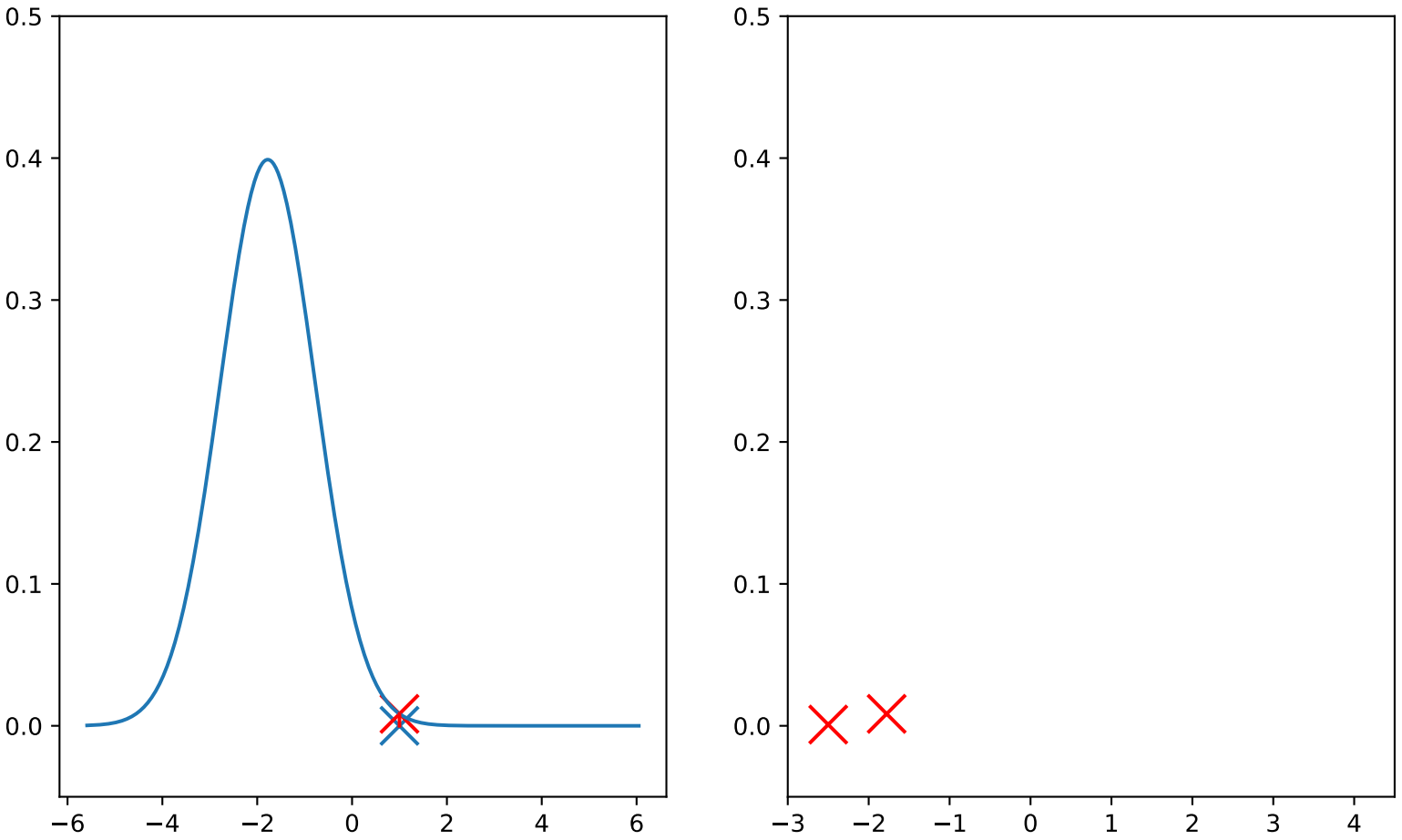
The Likelihood

Let's consider the simplest case when we have a single observation $\mathcal{D} = \{x_1\}$.

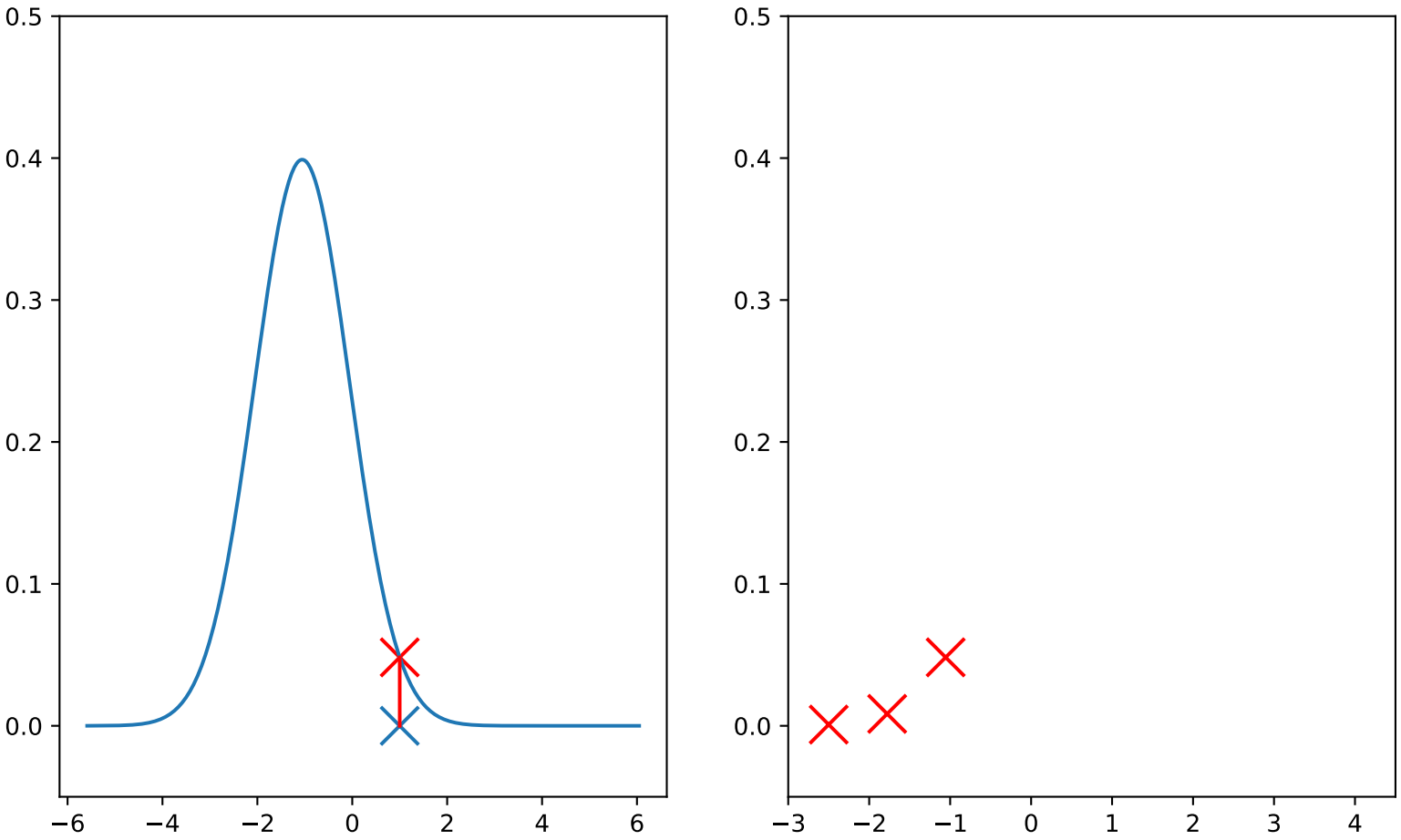
Rutger decides to fix the variance $\sigma^2 = 1$ and to vary μ.



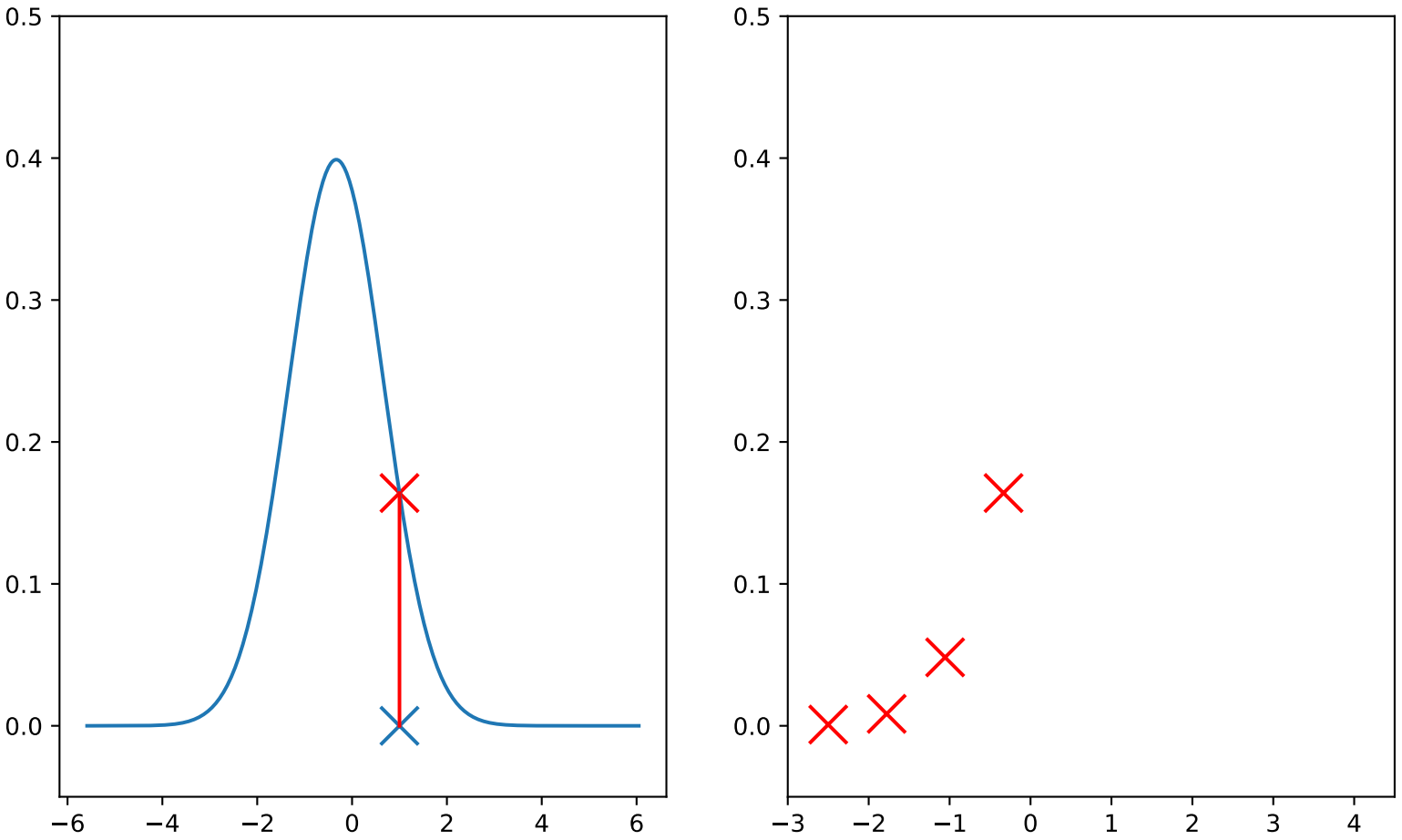
The Likelihood



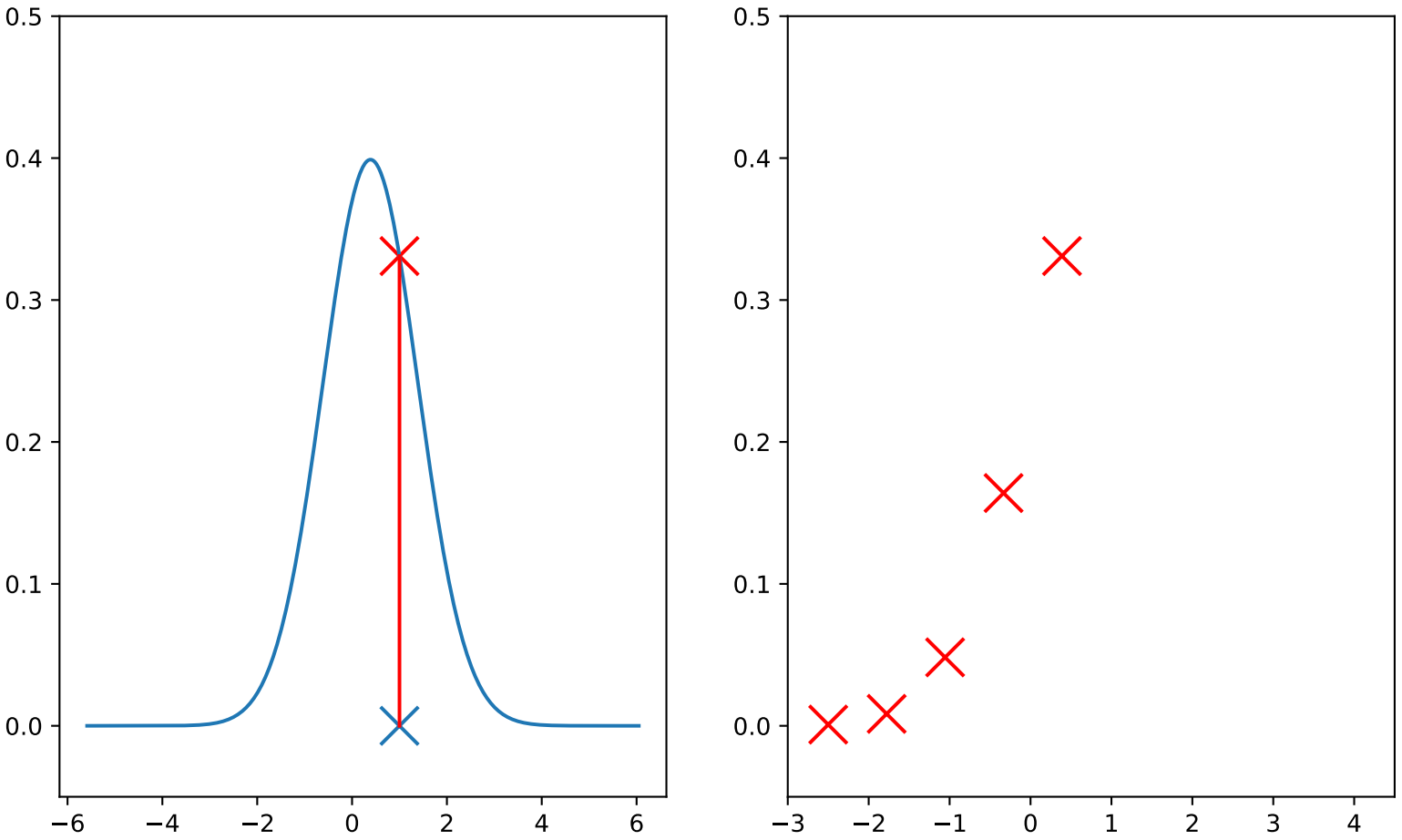
The Likelihood



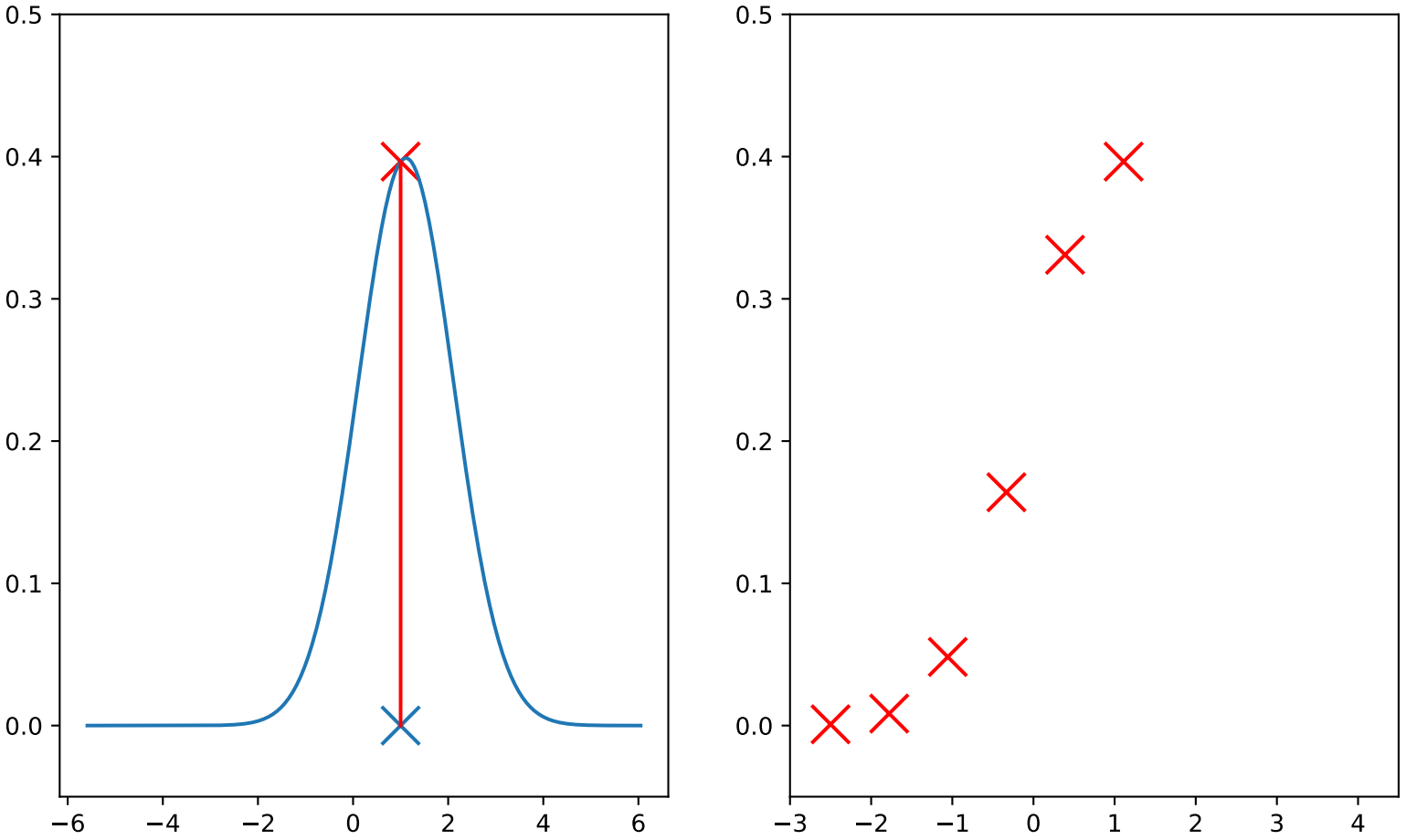
The Likelihood



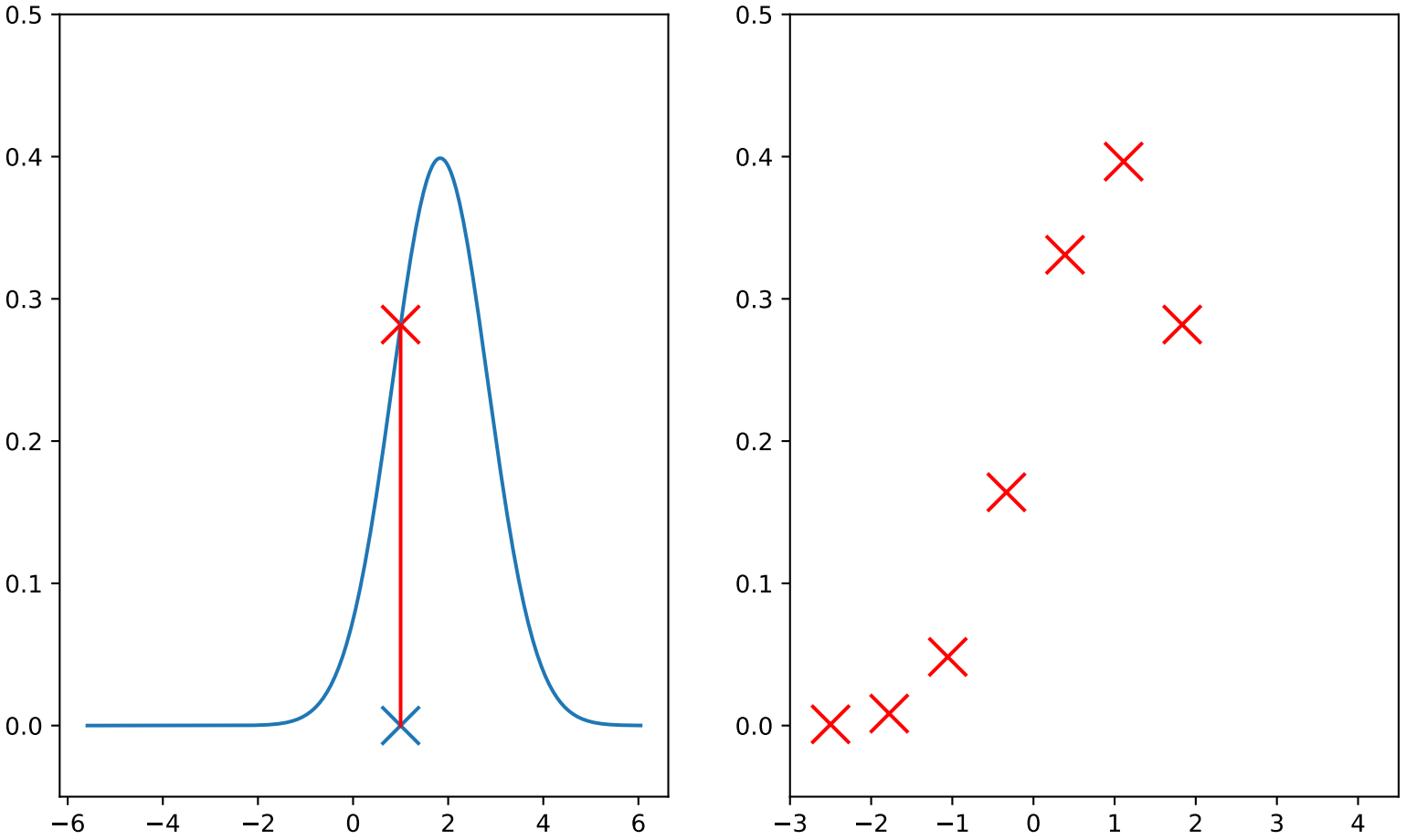
The Likelihood



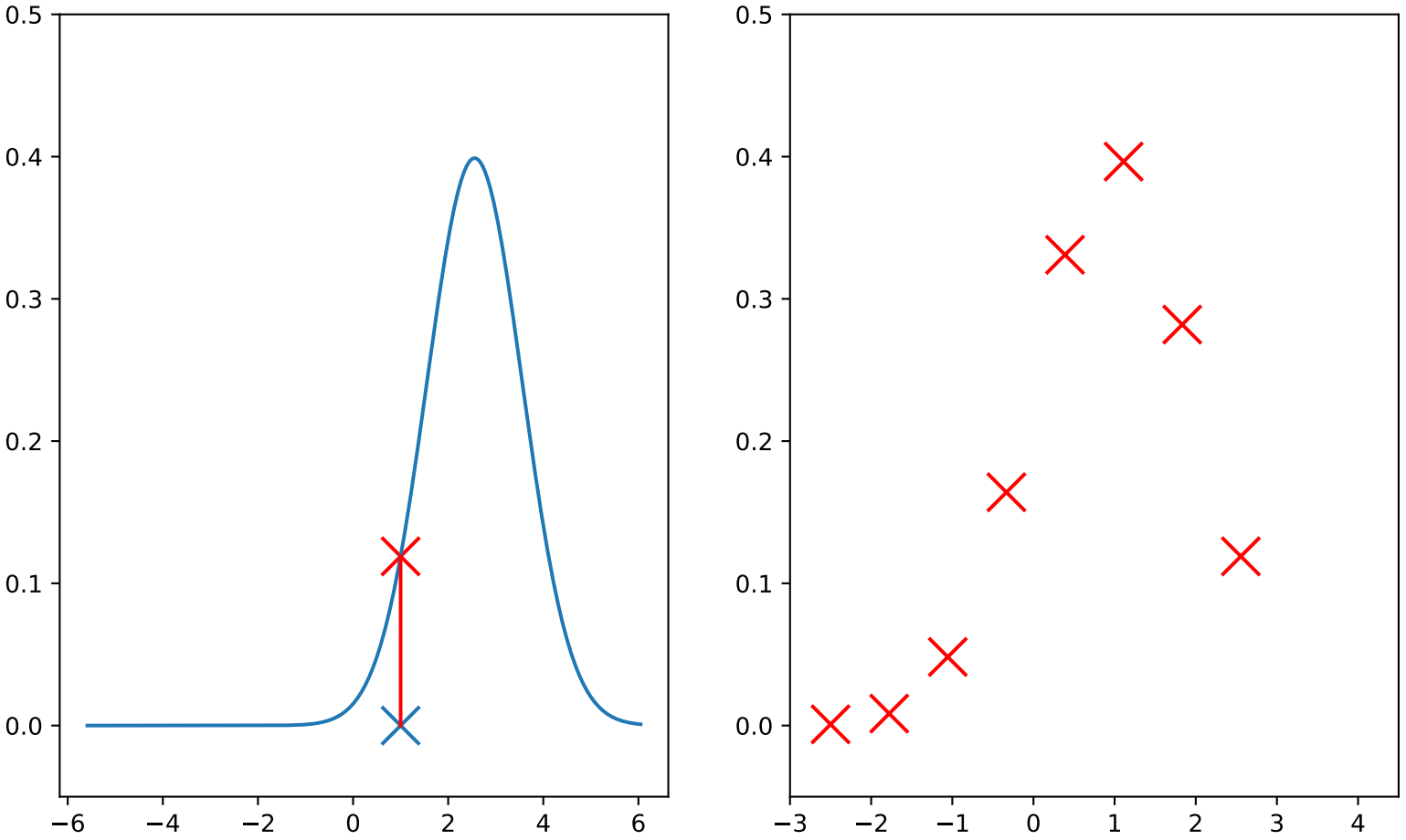
The Likelihood



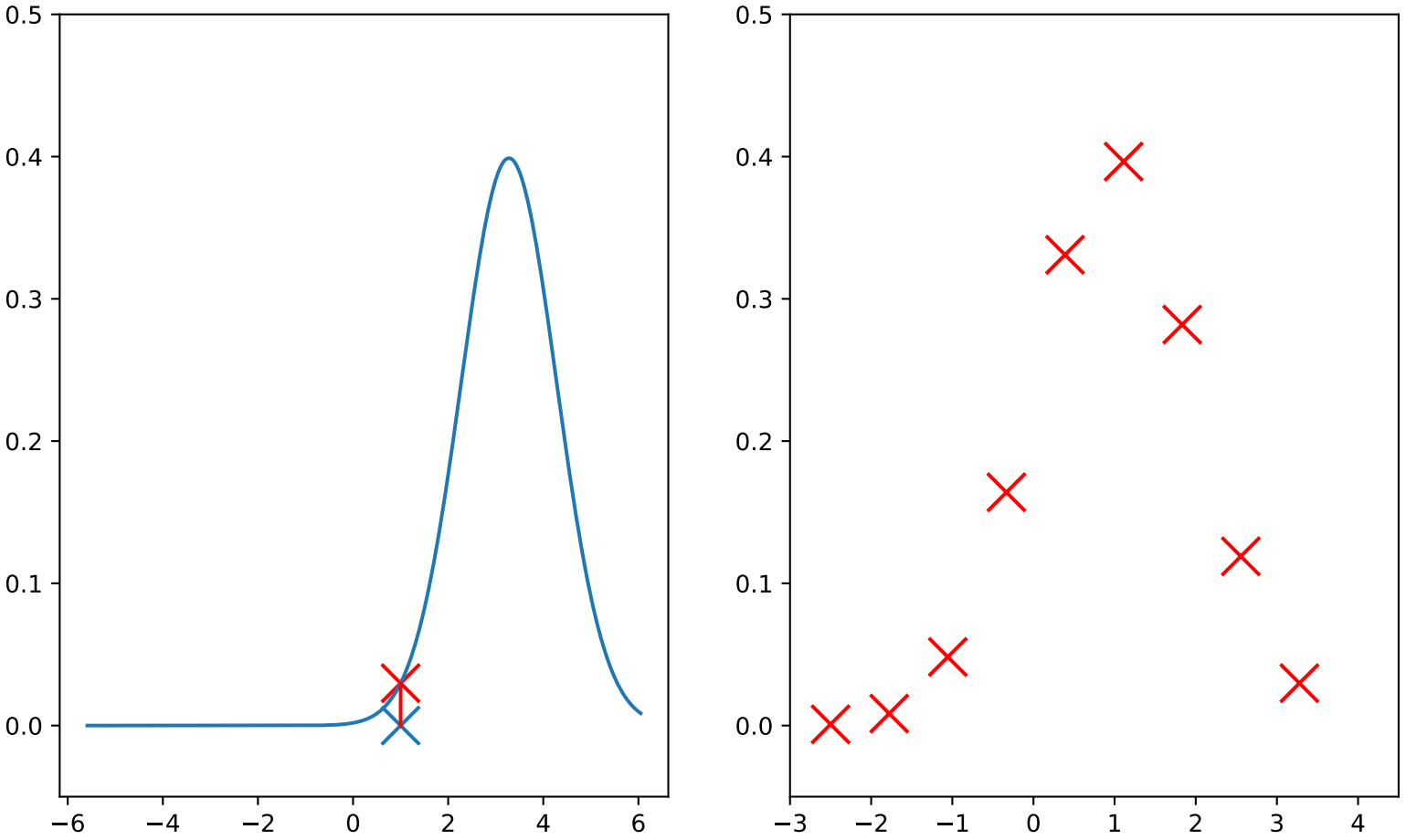
The Likelihood



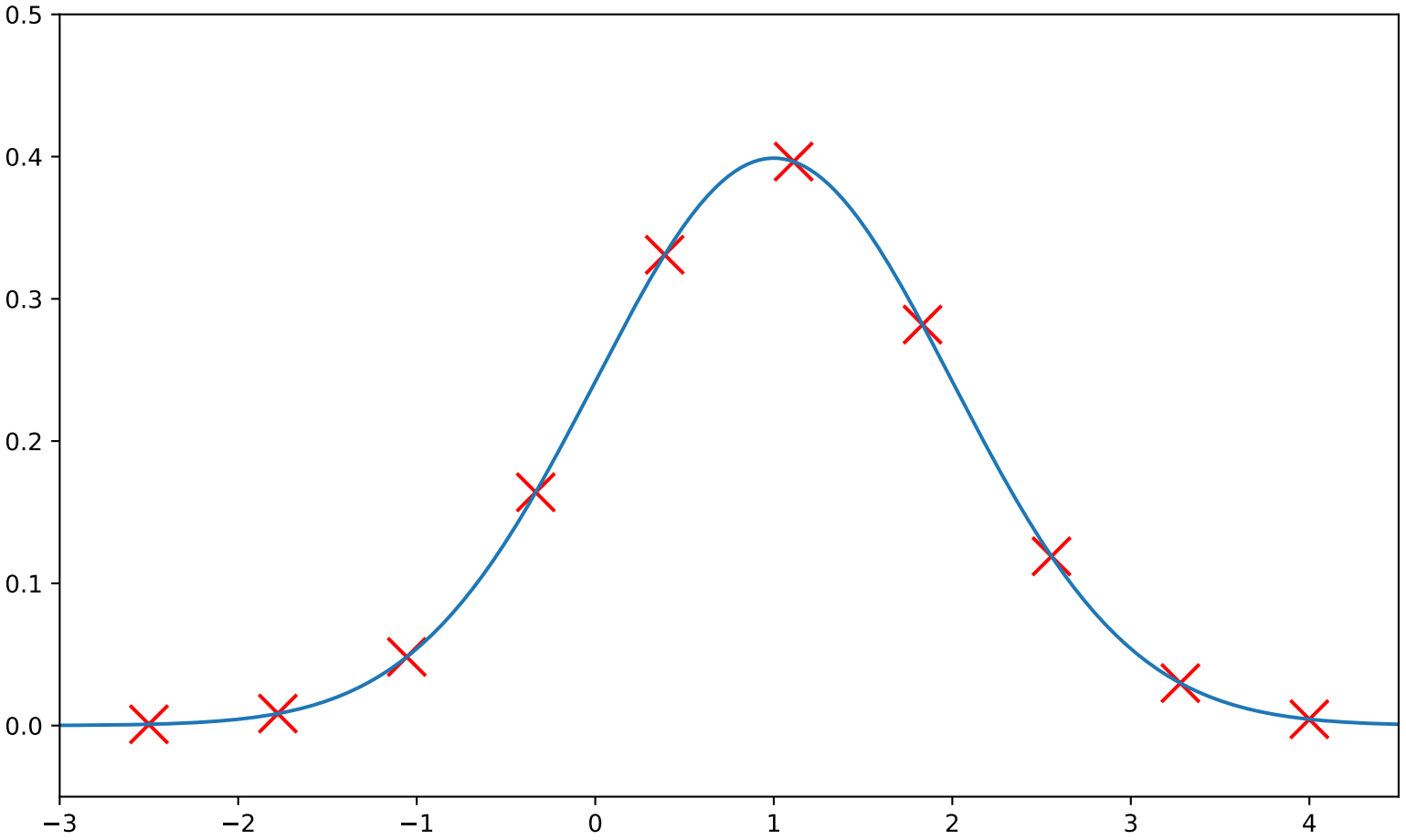
The Likelihood



The Likelihood



The Likelihood



The Likelihood

So $\mathcal{N}(1.0|\mu, 1^2)$ has a Gaussian shape in μ. We could have actually seen this by looking at the equation

$$p(\mathcal{D}|\boldsymbol{\theta}) = \mathcal{N}(1.0|\mu, 1.0^2) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(1.0-\mu)^2}{2}\right\}$$

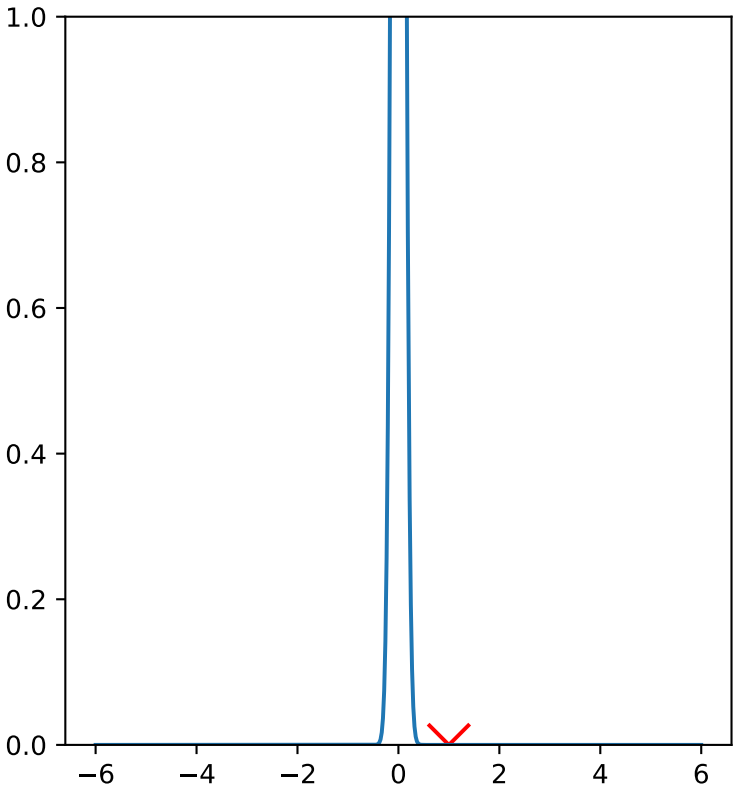
This is a function of μ, not a function of $\mathcal{D}$.

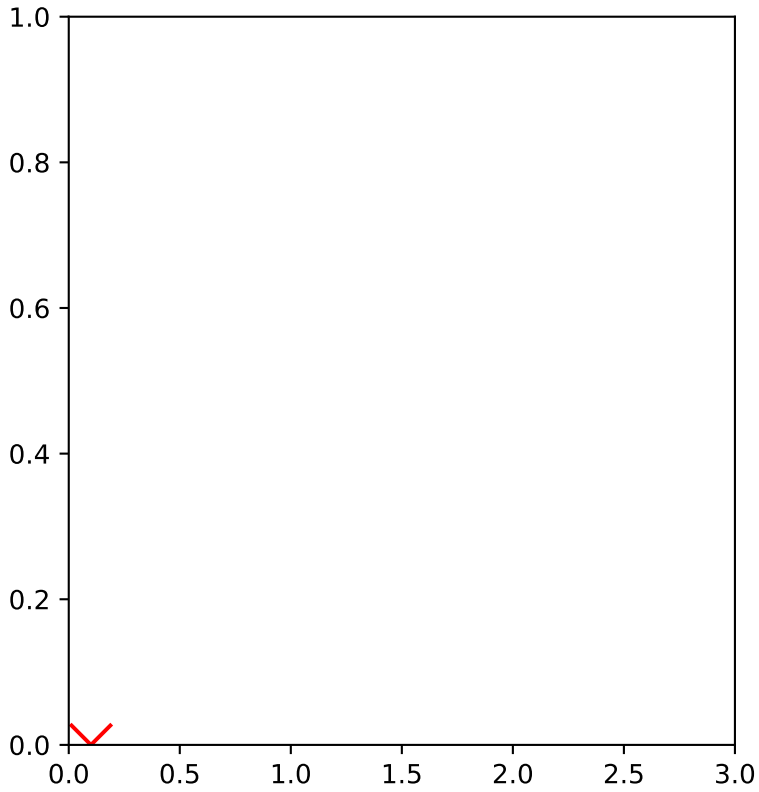
This is given the name the *likelihood of μ given $\mathcal{D}$*, or the *likelihood of the parameters given the data*, or just the *likelihood*.

It is a cardinal sin to say the likelihood of $\mathcal{D}$ given μ

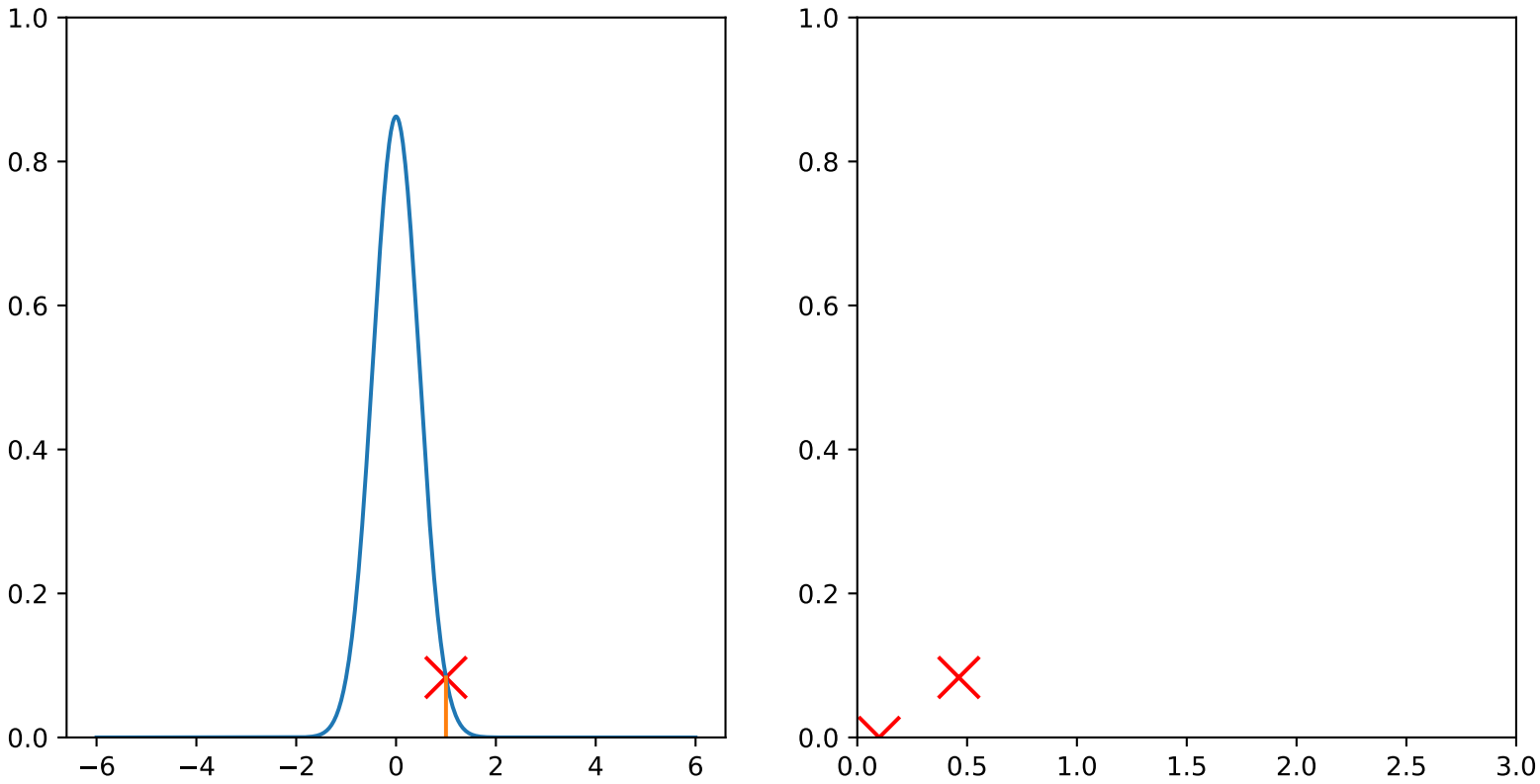
In reality, the Gaussian likelihood is a function of both μ and σ. So it should be a two dimensional plot. Let's see what happens if we search over σ instead of μ.

The Likelihood

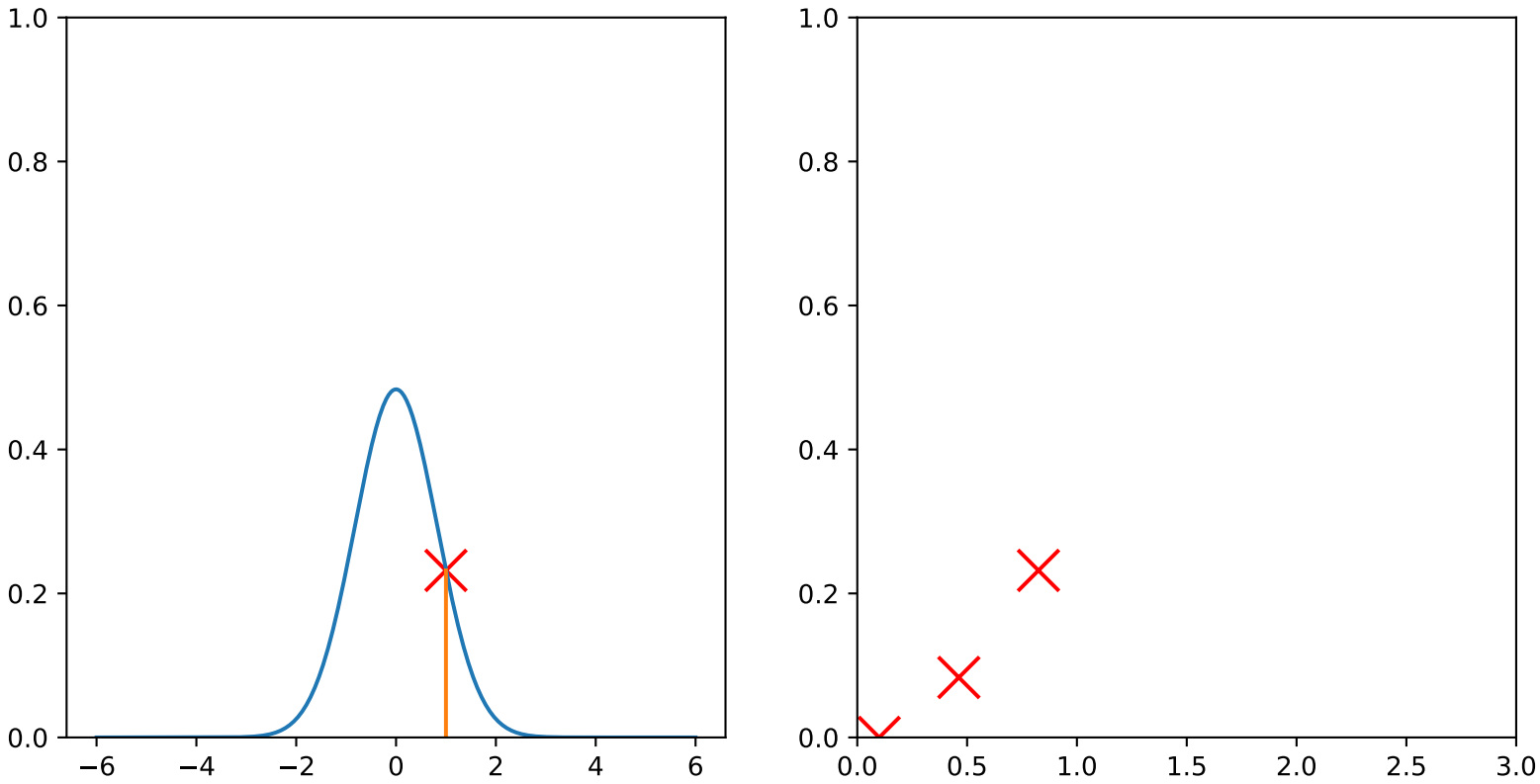




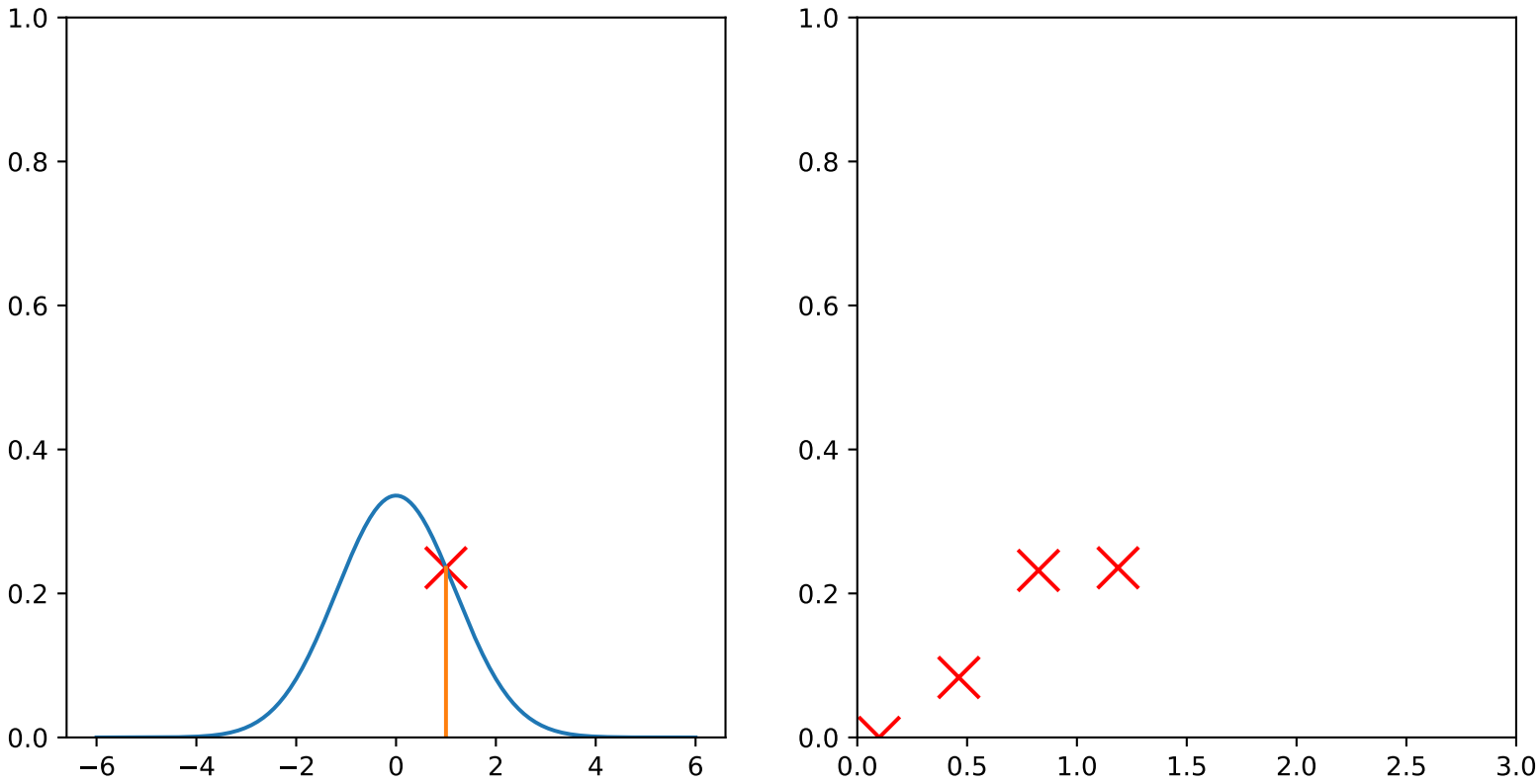
The Likelihood



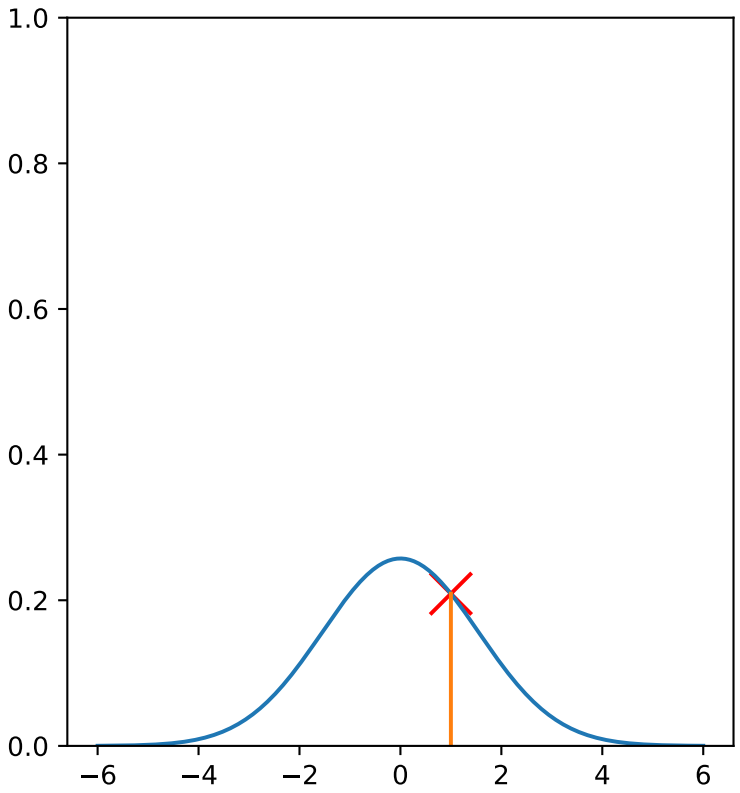
The Likelihood

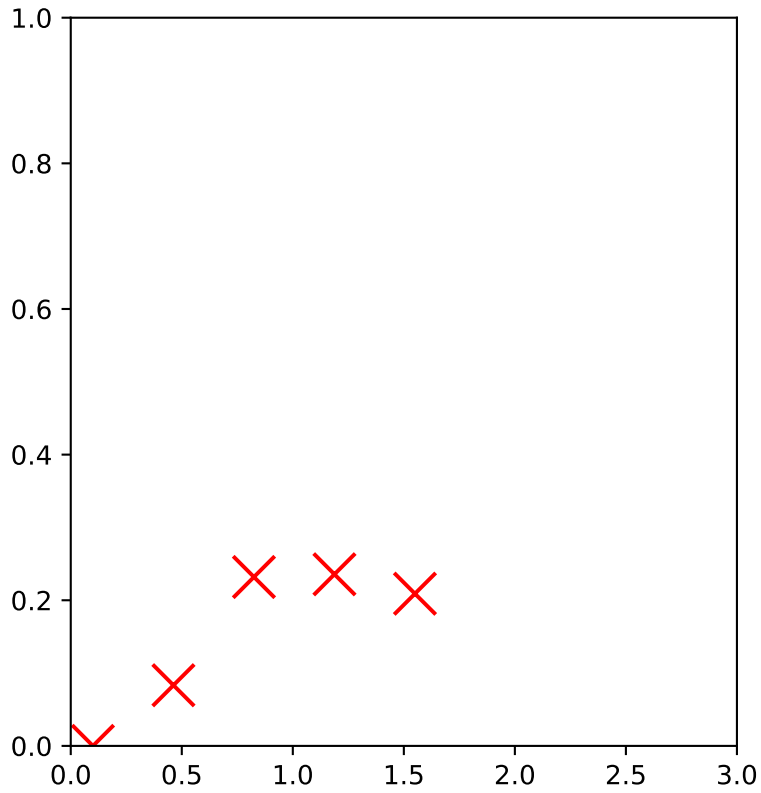


The Likelihood

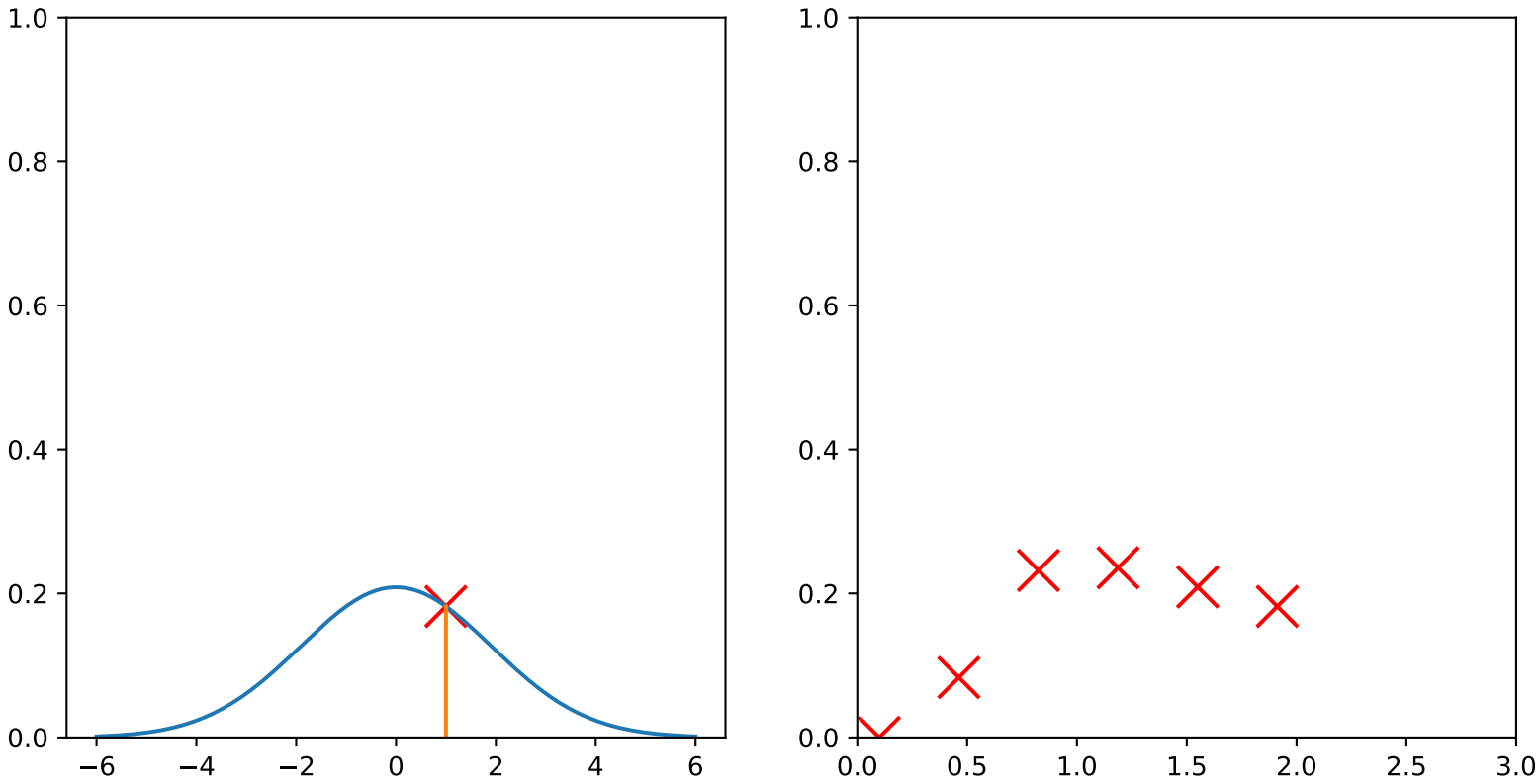


The Likelihood

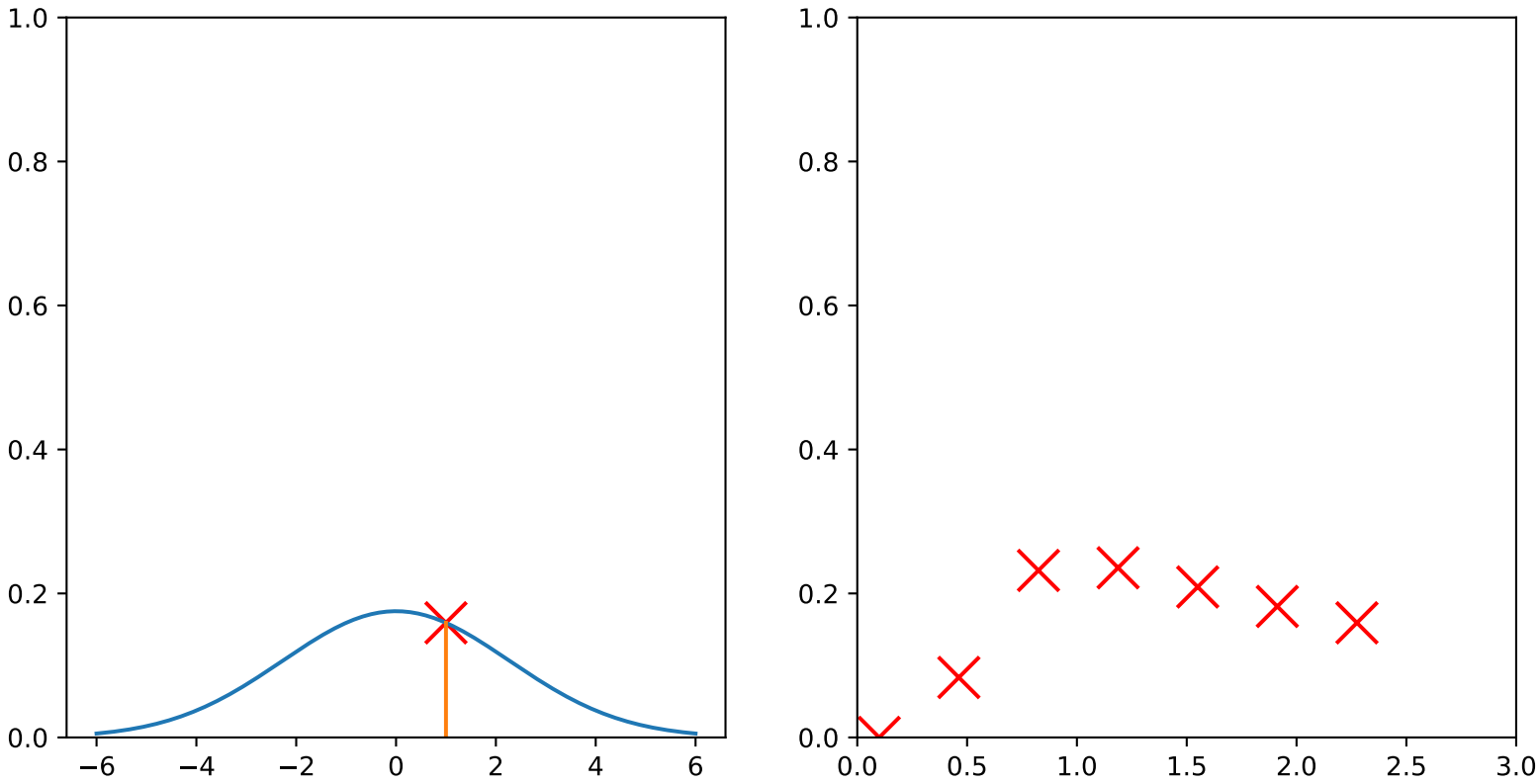




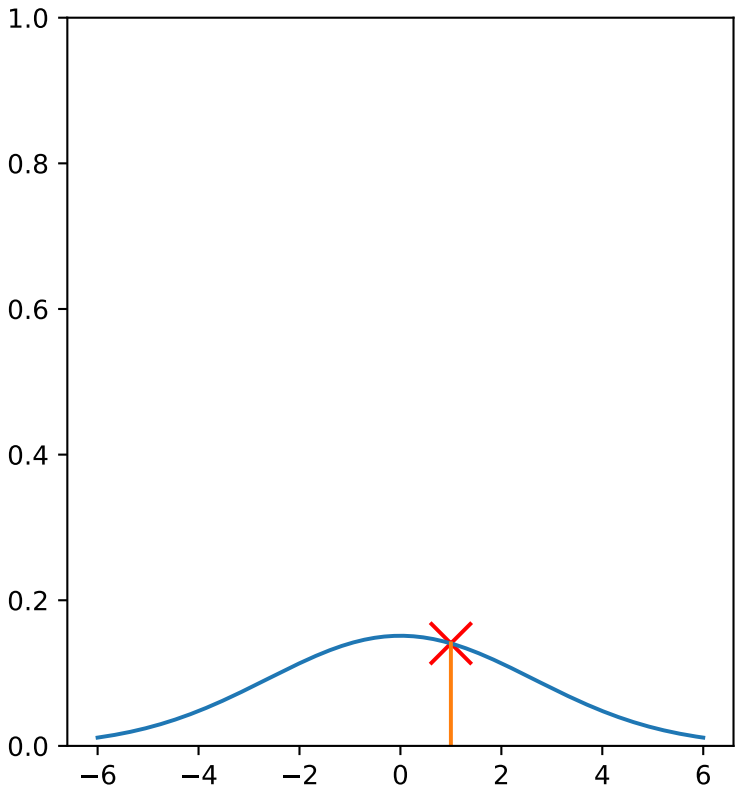
The Likelihood

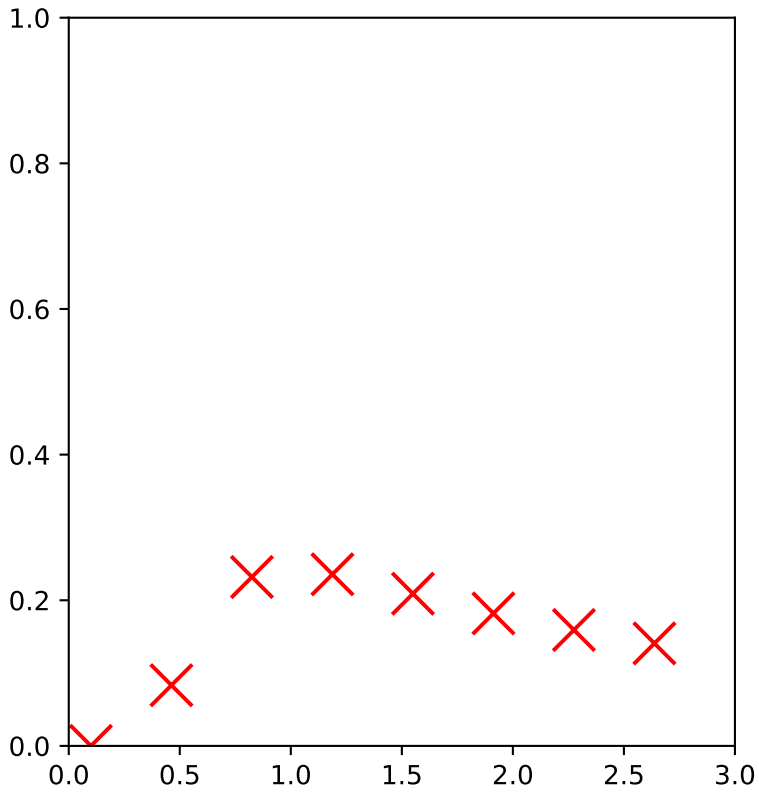


The Likelihood

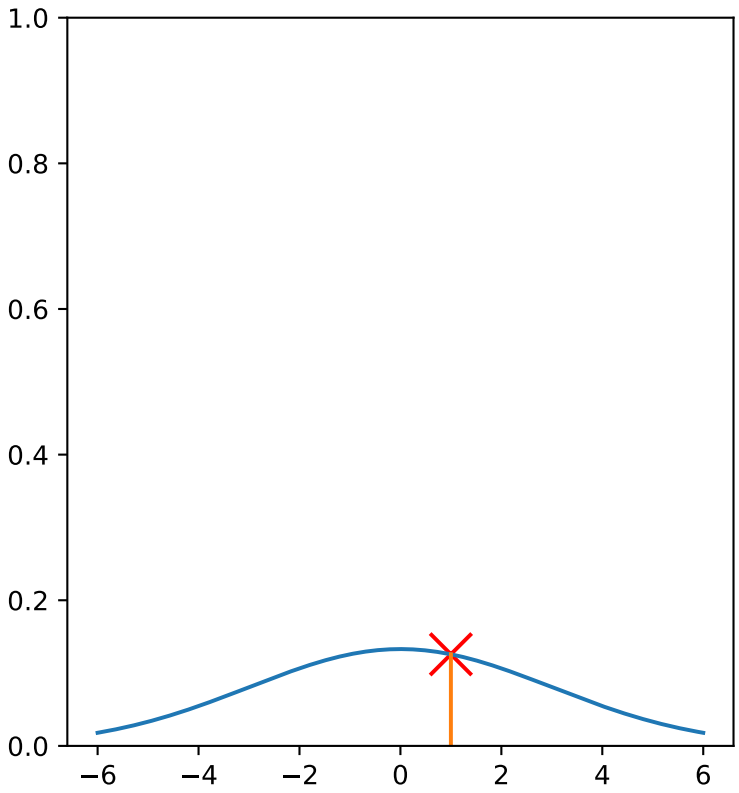


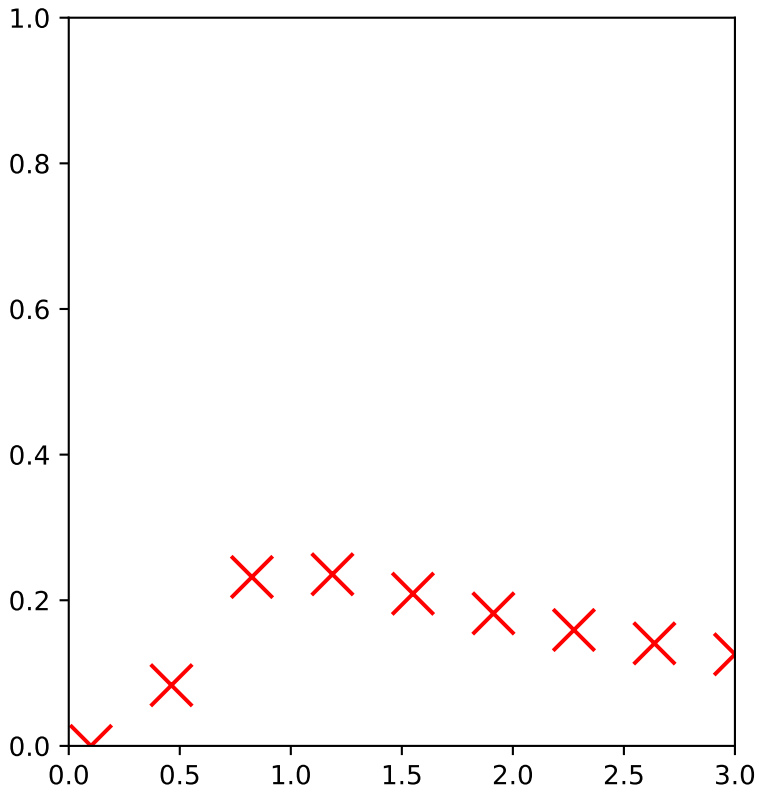
The Likelihood





The Likelihood





The Likelihood

It turn out that $\mathcal{N}(1.0|1, \sigma^2)$ has the shape of something called the inverse-Gamma distribution[1].

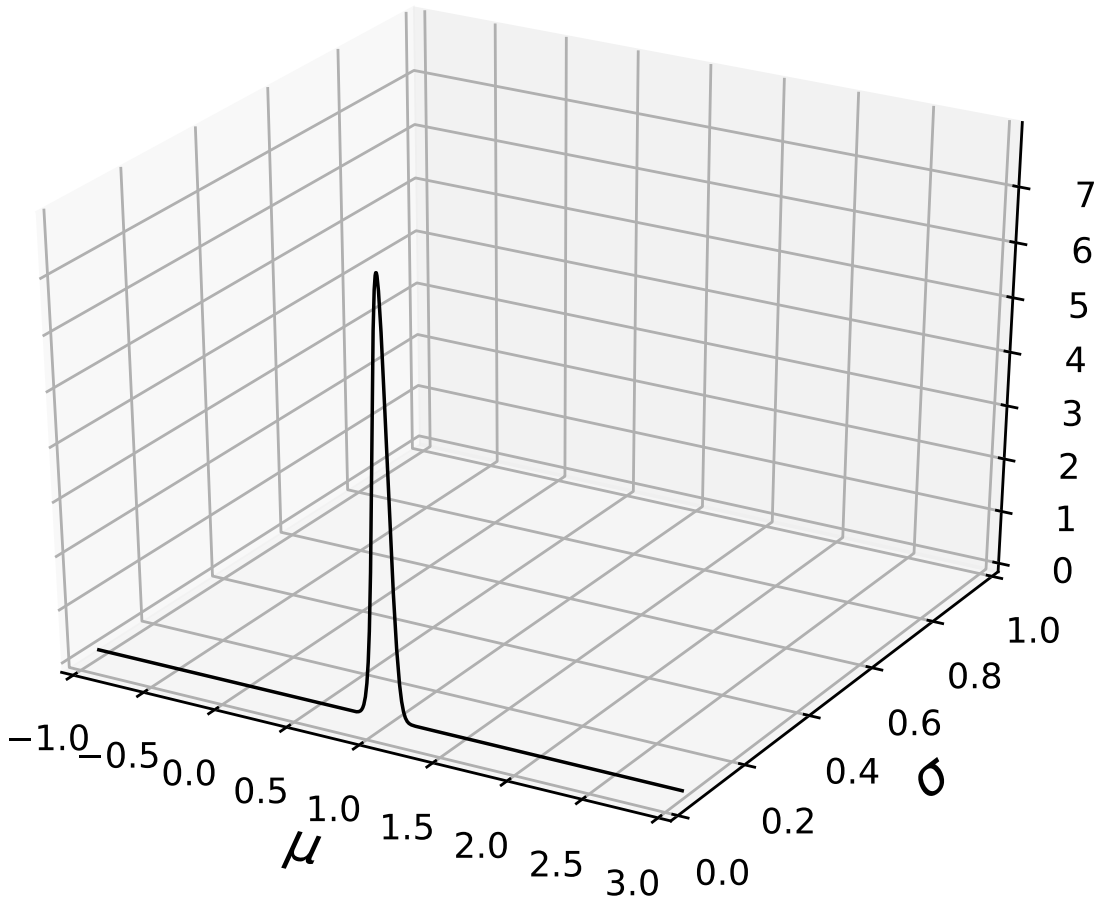
$$
\begin{aligned}
p(\mathcal{D}|\boldsymbol{\theta}) &= \mathcal{N}(1.0|0, \sigma^2) \\
&= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(1.0-0)^2}{2\sigma^2}\right\} \\
&= \frac{\sigma^{-1}}{\sqrt{2\pi}} \exp\left\{-\frac{\sigma^{-2}}{2}\right\} \\
&\propto \sigma^{-1} \exp\{-\sigma^{-2}\}
\end{aligned}
$$

Now let's explore the 2D likelihood function.

[1] But note that it isn't a distribution!

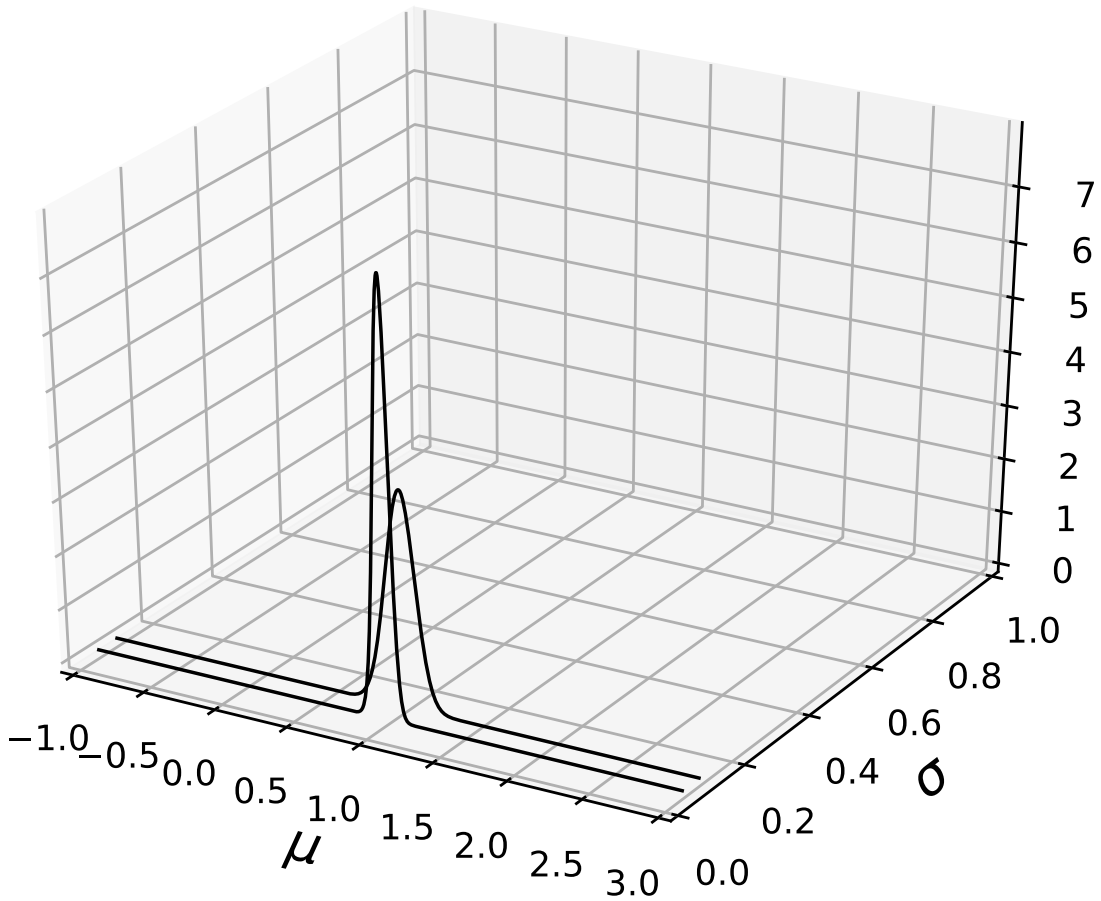
The Gaussian Likelihood

$$p(\mathcal{D}|\mu, \sigma^2 = 0.05^2) = \frac{1}{0.05\sqrt{2\pi}} \exp\left\{-\frac{(1.0-\mu)^2}{0.05^2}\right\}$$



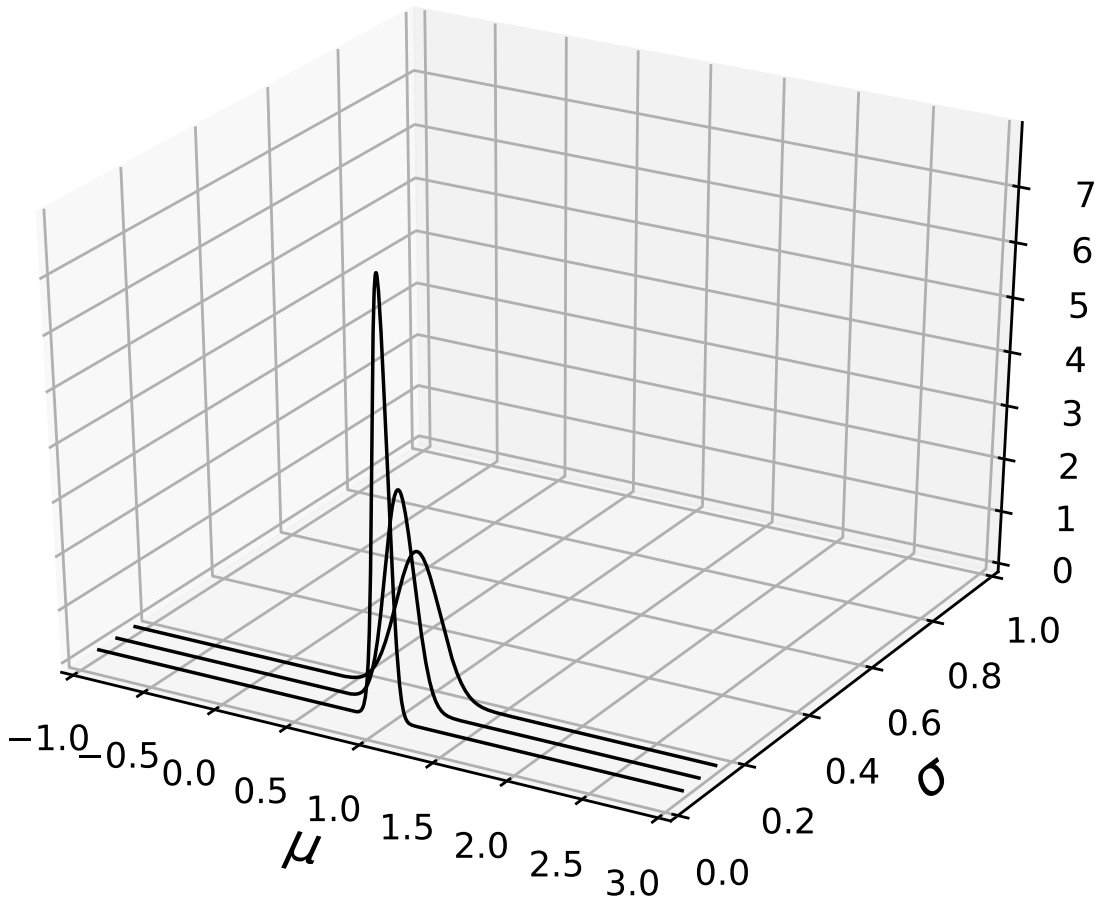
The Gaussian Likelihood

$$p(\mathcal{D}|\mu, \sigma^2 = 0.10^2) = \frac{1}{0.10\sqrt{2\pi}} \exp\left\{-\frac{(1.0-\mu)^2}{0.10^2}\right\}$$



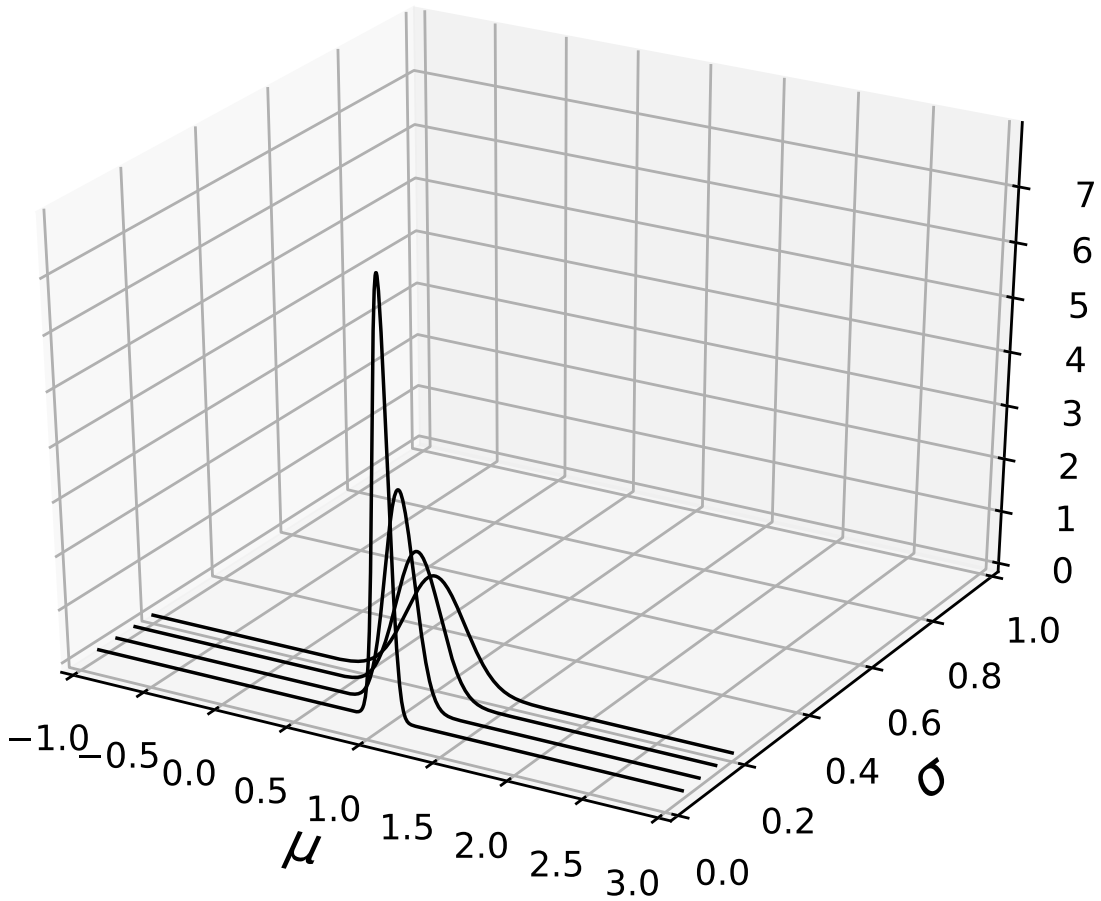
The Gaussian Likelihood

$$p(\mathcal{D}|\mu, \sigma^2 = 0.15^2) = \frac{1}{0.15\sqrt{2\pi}} \exp\left\{-\frac{(1.0-\mu)^2}{0.15^2}\right\}$$



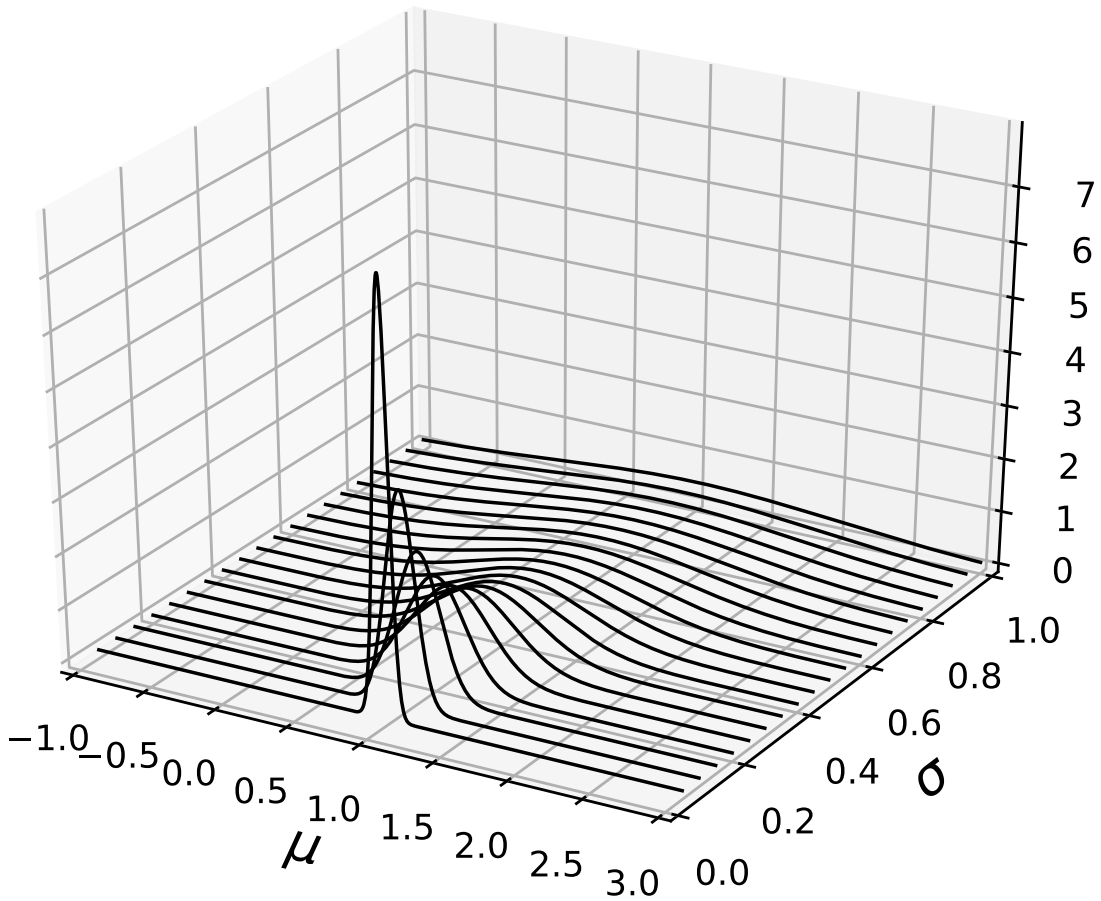
The Gaussian Likelihood

$$p(\mathcal{D}|\mu, \sigma^2 = 0.20^2) = \frac{1}{0.20\sqrt{2\pi}} \exp\left\{-\frac{(1.0-\mu)^2}{0.20^2}\right\}$$



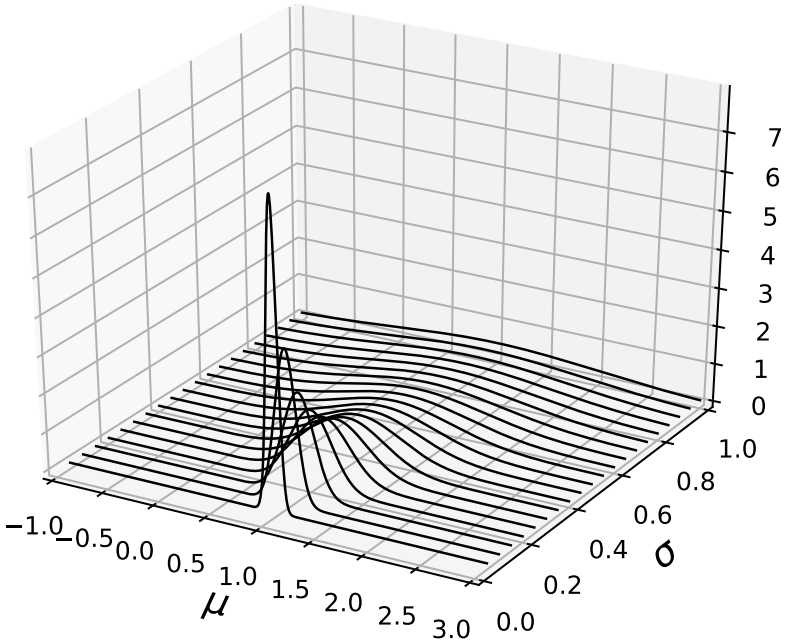
The Gaussian Likelihood

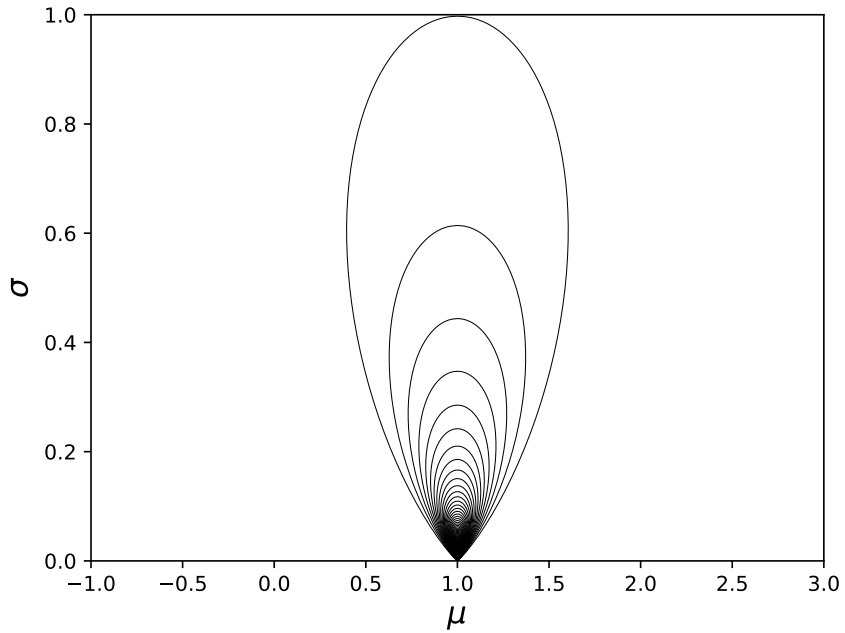
$$p(\mathcal{D}|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(1.0-\mu)^2}{\sigma^2}\right\}$$



The Likelihood

$$p(\mathcal{D}|\boldsymbol{\theta}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(1.0-\mu)^2}{2\sigma^2}\right\}$$





It is a good exercise to work out the equation for isocontours of this function. Notice how there is a delta peak when $\sigma = 0$ (not drawn).

The Likelihood

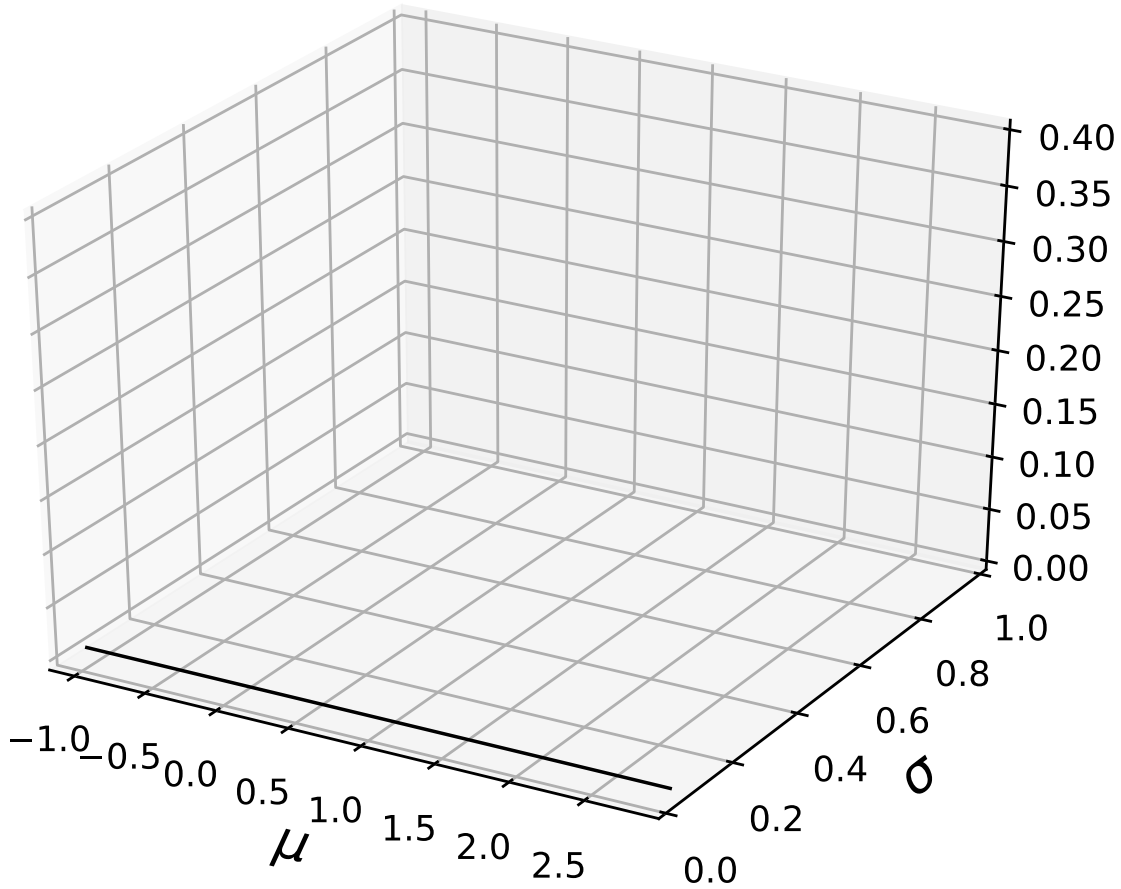
Let's consider what happens to the likelihood when Rutger has $N = 5$ samples.

We know that $\mathcal{D} = \{x_1, x_2, x_3, x_4, x_5\}$, and that each point was drawn iid, so

$$\begin{aligned} p(\mathcal{D}|\mu, \sigma^2) &= p(x_1, x_2, x_3, x_4, x_5|\mu, \sigma^2) \\ &= p(x_1|\mu, \sigma^2)p(x_2|\mu, \sigma^2)p(x_3|\mu, \sigma^2)p(x_4|\mu, \sigma^2)p(x_5|\mu, \sigma^2) \\ &= \prod_{i=1}^{N} p(x_i|\mu, \sigma^2) \\ &= \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x_i - \mu)^2}{\sigma^2}\right\} \end{aligned}$$

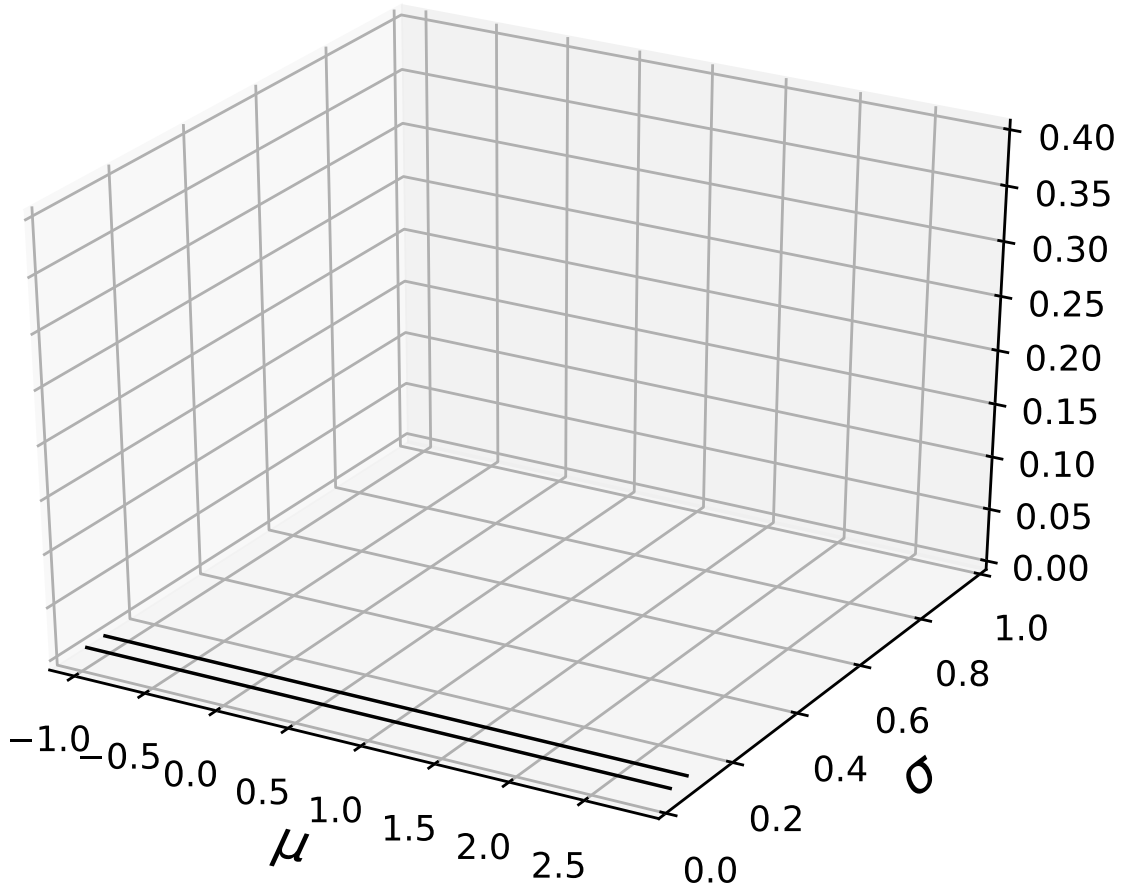
The Gaussian Likelihood

$$p(\mathcal{D}|\mu, \sigma^2 = 0.05^2) = \prod_{i=1}^{N} \frac{1}{0.05\sqrt{2\pi}} \exp\left\{-\frac{(x_i - \mu)^2}{0.05^2}\right\}$$



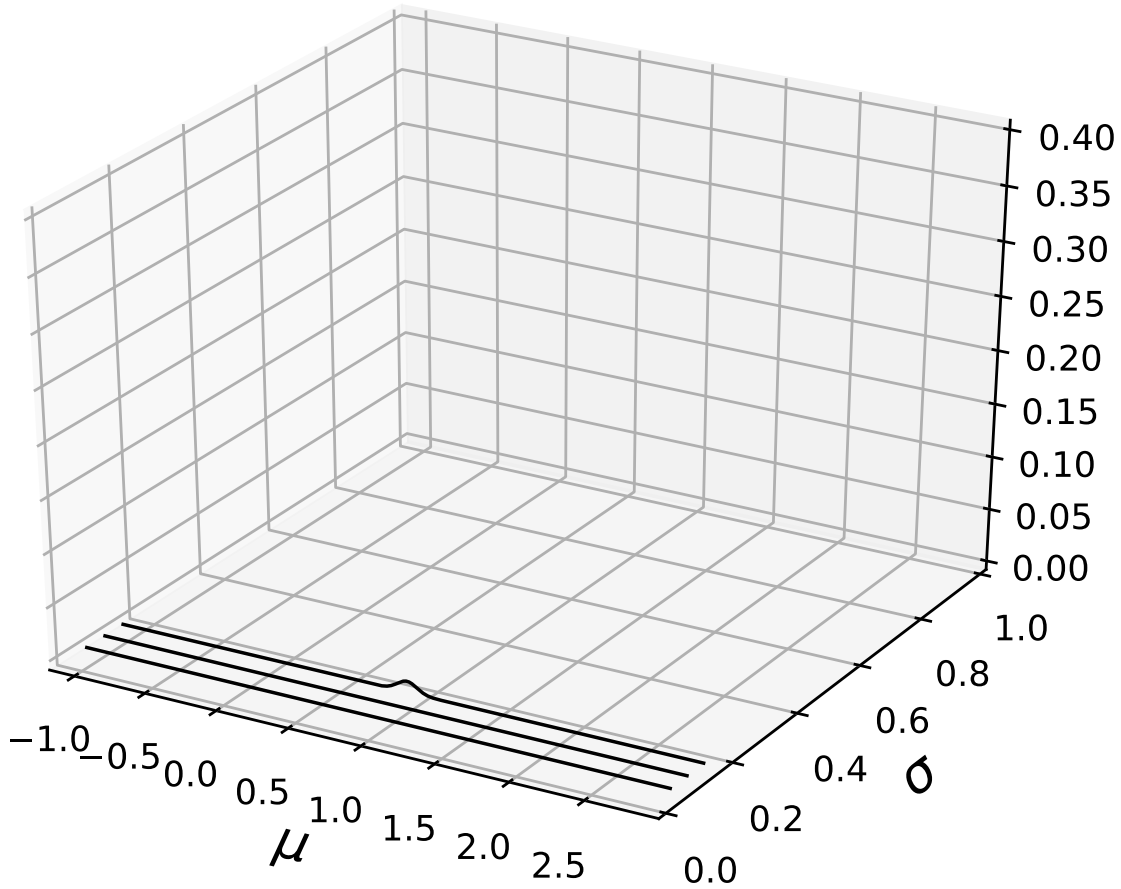
The Gaussian Likelihood

$$p(\mathcal{D}|\mu, \sigma^2 = 0.10^2) = \prod_{i=1}^{N} \frac{1}{0.10\sqrt{2\pi}} \exp\left\{-\frac{(x_i - \mu)^2}{0.10^2}\right\}$$



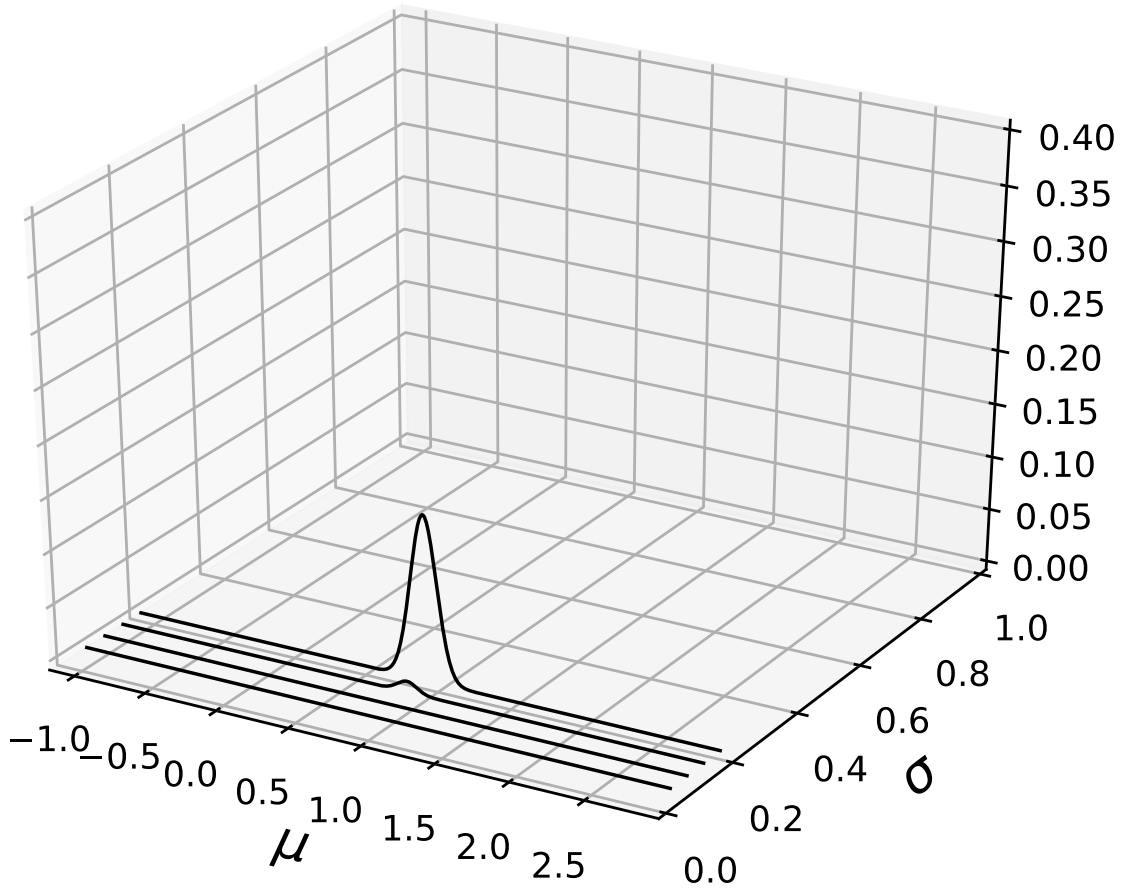
The Gaussian Likelihood

$$p(\mathcal{D}|\mu, \sigma^2 = 0.15^2) = \prod_{i=1}^{N} \frac{1}{0.15\sqrt{2\pi}} \exp\left\{-\frac{(x_i - \mu)^2}{0.15^2}\right\}$$



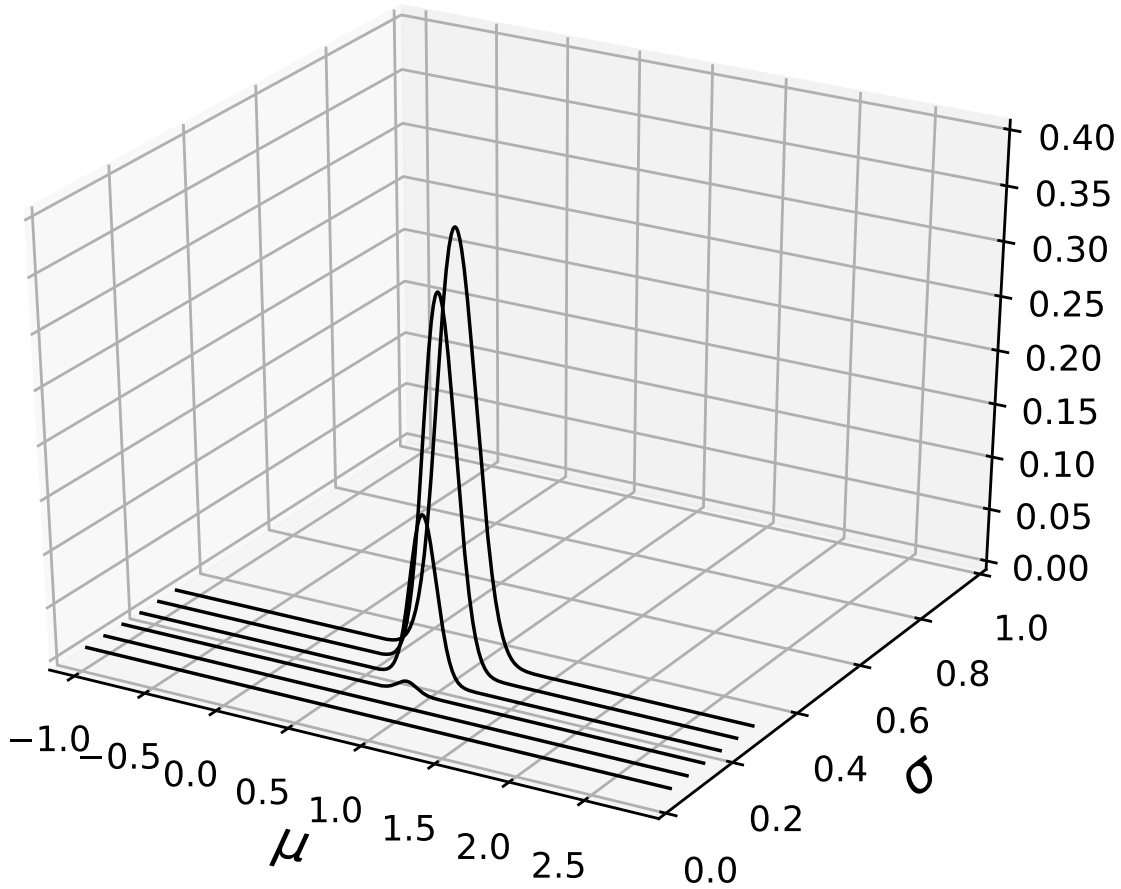
The Gaussian Likelihood

$$p(\mathcal{D}|\mu, \sigma^2 = 0.20^2) = \prod_{i=1}^{N} \frac{1}{0.20\sqrt{2\pi}} \exp\left\{-\frac{(x_i - \mu)^2}{0.20^2}\right\}$$



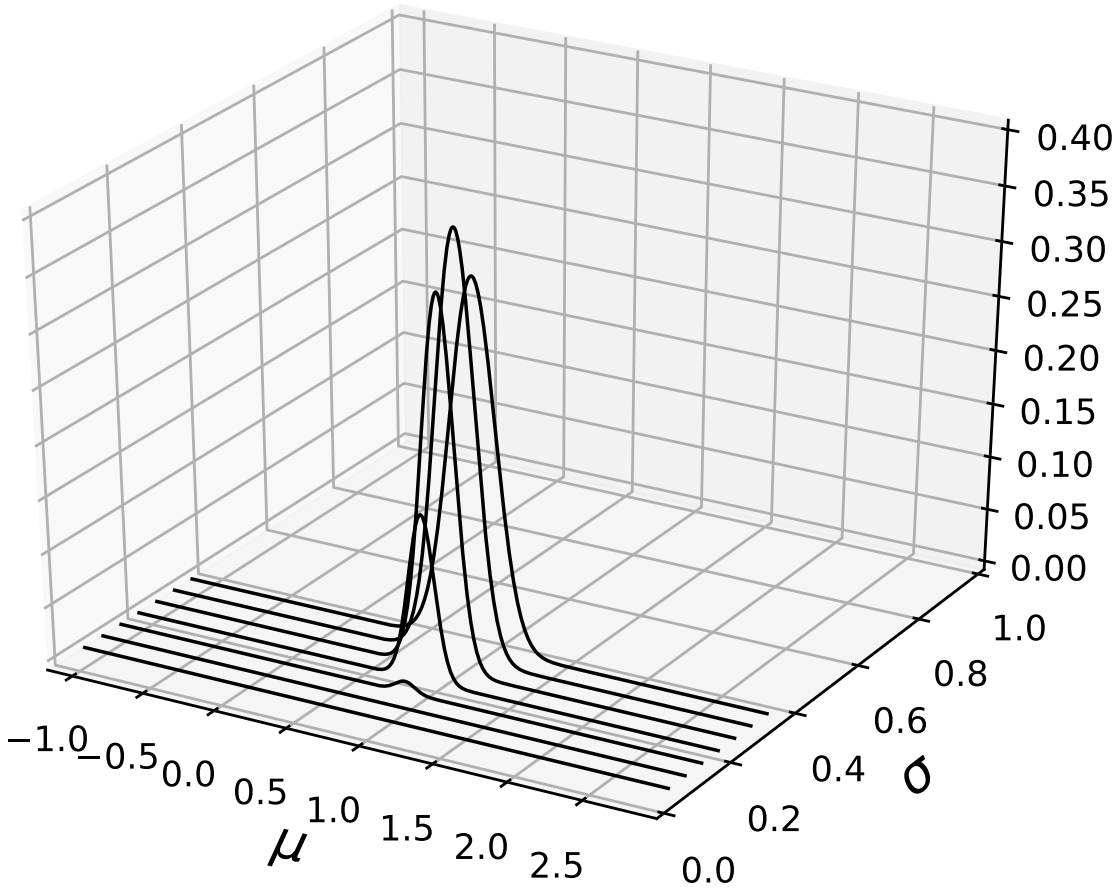
The Gaussian Likelihood

$$p(\mathcal{D}|\mu, \sigma^2 = 0.30^2) = \prod_{i=1}^{N} \frac{1}{0.30\sqrt{2\pi}} \exp\left\{-\frac{(x_i - \mu)^2}{0.30^2}\right\}$$



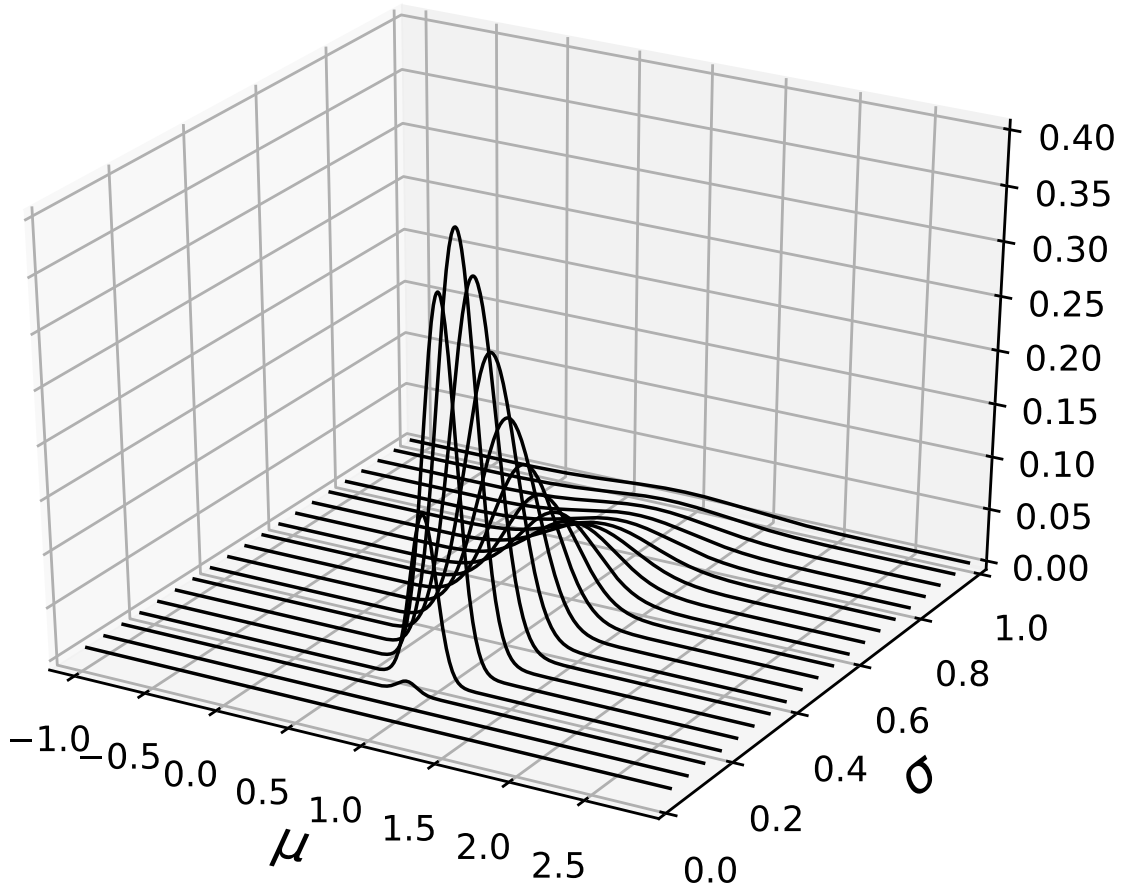
The Gaussian Likelihood

$$p(\mathcal{D}|\mu, \sigma^2 = 0.35^2) = \prod_{i=1}^{N} \frac{1}{0.35\sqrt{2\pi}} \exp\left\{-\frac{(x_i - \mu)^2}{0.35^2}\right\}$$



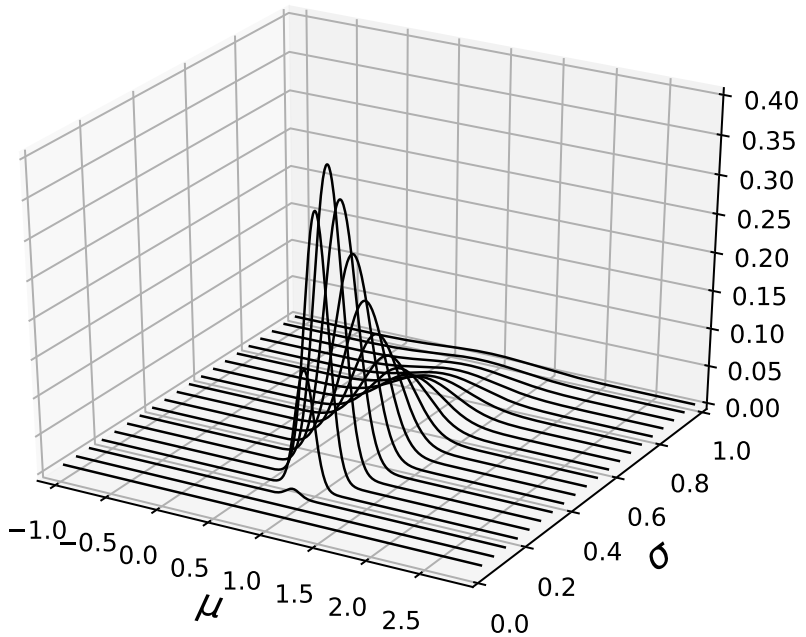
The Gaussian Likelihood

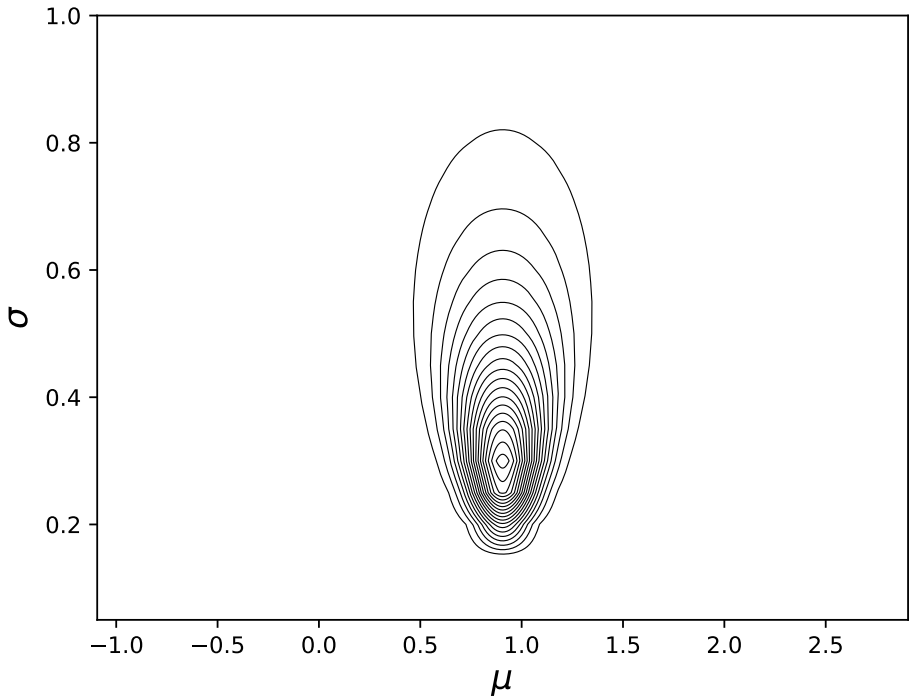
$$p(\mathcal{D}|\mu, \sigma^2 = 1.00^2) = \prod_{i=1}^{N} \frac{1}{1.00\sqrt{2\pi}} \exp\left\{-\frac{(x_i - \mu)^2}{1.00^2}\right\}$$



The Likelihood

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





Main idea

Best $\boldsymbol{\theta}$ maximizes the likelihood:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} p(\mathcal{D}|\boldsymbol{\theta})$$

We can find the maximum of $p(\mathcal{D}|\boldsymbol{\theta})$ by finding the point where

$$\frac{\partial}{\partial \boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta}) = \mathbf{0}.$$

Note that we should also check that the Hessian (second derivative) is negative (semi-)definite, but for the kinds of likelihoods we will consider, we need not bother.

Log Likelihood

We need to be able to compute the derivatives of a product of terms, for instance

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\left[f_1(\theta)f_2(\theta)f_3(\theta)\right] = \frac{\mathrm{d}f_1}{\mathrm{d}\theta}f_2(\theta)f_3(\theta) + f_1(\theta)\frac{\mathrm{d}f_2}{\mathrm{d}\theta}f_3(\theta) + f_1(\theta)f_2(\theta)\frac{\mathrm{d}f_3}{\mathrm{d}\theta}.$$

If some of the $f_i(\theta)$ are small, then we are sure to run into numerical underflow.

But notice that if we take the logarithm then we reduce numerical underflow

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\left[\log\left(f_1(\theta)f_2(\theta)f_3(\theta)\right)\right] = \frac{\mathrm{d}}{\mathrm{d}\theta}\log f_1(\theta) + \frac{\mathrm{d}}{\mathrm{d}\theta}\log f_2(\theta) + \frac{\mathrm{d}}{\mathrm{d}\theta}\log f_3(\theta).$$

Most importantly, the logarithm does not affect the placement of the maximum, since it is a monotonic, point-wise function

$$\arg\max_x f(x) = \arg\max_x \log f(x)$$

Maximum Likelihood Gaussian

Given a dataset $\mathcal{D} = \{x_1, x_2, x_3, ..., x_N\}$ and assuming the data was generated from a Gaussian, what is the maximum likelihood mean and variance?

$$\begin{aligned}
\arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta}) &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \log\left(\prod_{i=1}^{N} p(x_i|\boldsymbol{\theta})\right) \\
&= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{N} \log p(x_i|\boldsymbol{\theta}) \\
&= \arg\max_{\mu,\sigma} \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}\right) \\
&= \arg\max_{\mu,\sigma} \sum_{i=1}^{N} -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}
\end{aligned}$$

Maximum Likelihood Gaussian

$$
\begin{aligned}
\arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta}) &= \arg\max_{\mu,\sigma} \sum_{i=1}^{N} -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2} \\
&= \arg\max_{\mu,\sigma} -\frac{1}{2}\left(\sum_{i=1}^{N} \log 2\pi\sigma^2 + \frac{(x_i-\mu)^2}{\sigma^2}\right) \\
&= \arg\min_{\mu,\sigma} \sum_{i=1}^{N} \log 2\pi\sigma^2 + \frac{(x_i-\mu)^2}{\sigma^2} \\
&= \arg\min_{\mu,\sigma} N\log 2\pi\sigma^2 + \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{\sigma^2}
\end{aligned}
$$

Maximum Likelihood Gaussian

Fixed σ^2 and maximize $\mathcal{L}$ wrt μ

$$\frac{\partial \mathcal{L}}{\partial \mu} = \frac{\partial}{\partial \mu}\left(\cancel{N \log 2\pi\sigma^2} + \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^2}\right)$$

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

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

$$= -\frac{2}{\sigma^2} \sum_{i=1}^{N} (x_i - \mu) = 0$$

$$\implies \sum_{i=1}^{N} (x_i - \mu) = 0 \implies \boxed{\mu = \frac{1}{N} \sum_{i=1}^{N} x_i}$$

Maximum Likelihood Gaussian

Fixed μ and maximize $\mathcal{L}$ wrt σ^2: Rewrite $\lambda = 1/\sigma^2$ (λ is called the **precision**)

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \frac{\partial}{\partial \lambda}\left(N \log \frac{2\pi}{\lambda} + \sum_{i=1}^{N} (x_i - \mu)^2 \lambda \right)$$

$$= \cancel{\frac{\partial}{\partial \lambda} N \log 2\pi} - \frac{\partial}{\partial \lambda} N \log \lambda + \frac{\partial}{\partial \lambda}\left(\sum_{i=1}^{N} \lambda^{-1} (x_i - \mu)^2 \right)$$

$$= -\frac{N}{\lambda} + \sum_{i=1}^{N} (x_i - \mu)^2 = 0$$

$$\implies \frac{N}{\lambda} = \sum_{i=1}^{N} (x_i - \mu)^2$$

$$\implies \boxed{\sigma^2 = 1/\lambda = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$$

Maximum likelihood estimators

The maximum likelihood mean and variance are thus

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

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

Don't these expressions look familiar? They are the mean and variance estimator! The mean estimator is unbiased, but the variance estimator is only consistent.

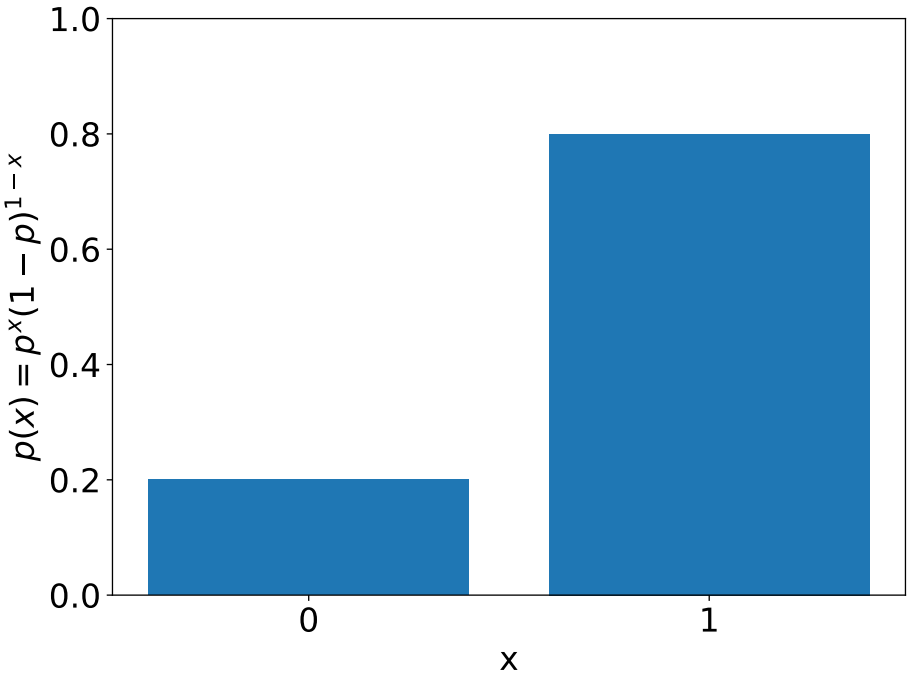
In this example, we took the $\arg\max$ wrt $\lambda = 1/\sigma^2$ instead of σ. Maximum likelihood is parameterisation invariant.

Maximum Likelihood Bernoulli

Rutger next tries the Bernoulli distribution, with $x_i \sim \text{Bern}(p) = p^x(1-p)^{1-x}$.

The log-likelihood

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \log P(x_i|\boldsymbol{\theta}) \\
&= \sum_{i=1}^{N} \log p^{x_i}(1-p)^{1-x_i} \\
&= \sum_{i=1}^{N} \log p^{x_i} + \log(1-p)^{1-x_i} \\
&= \sum_{i=1}^{N} x_i \log p + (1-x_i)\log(1-p)
\end{aligned}$$



Maximum Likelihood Bernoulli

Take derivatives

Again we form the derivative of $\mathcal{L}$ wrt $\boldsymbol{\theta}$, where $\boldsymbol{\theta} = p$

$$\frac{\partial}{\partial p}\mathcal{L} = \frac{\partial}{\partial p}\sum_{i=1}^{N} x_i \log p + (1 - x_i)\log(1 - p)$$

$$= \sum_{i=1}^{N}\left(\frac{x_i}{p} - \frac{1 - x_i}{1 - p}\right) = 0$$

Where we used the fact that $\frac{\partial}{\partial p}\log f(p) = \frac{f'(p)}{f(p)}$

$$\sum_{i=1}^{N} x_i(1 - p) - (1 - x_i)p = 0$$

$$\sum_{i=1}^{N} x_i - \cancel{x_i p} - p + \cancel{x_i p} = 0$$

$$\sum_{i=1}^{N} x_i - p = 0 \implies \boxed{p = \frac{1}{N}\sum_{i=1}^{N} x_i = 0}$$

Maximum Likelihood Poisson

A Poisson distribution takes the form

$$P(r|\lambda) = \frac{e^{-\lambda}\lambda^r}{r!}$$

where $r \in \{0, 1, 2, 3, ...\}$ and $\lambda \in \mathbb{R}_{>0}$.

Poisson distributions are commonly used to model the number of events r occurring in a fixed interval of time or space at constant average rate λ e.g.

- The number of patients walking into a hospital between 0700 and 0800
- The number of insects per square meter of forest
- The number of meteorites of size Xm striking the Earth every year

Let's do maximum likelihood on this too.

Maximum Likelihood Poisson

First write down the likelihood function

$$\mathcal{L}(\lambda) = \sum_{i=1}^{N} \log \frac{e^{-\lambda}\lambda^{r_i}}{r_i!} = \sum_{i=1}^{N} -\lambda + r_i \log \lambda - \log r_i!$$

Then find the maximum

$$\begin{aligned}
\frac{\partial}{\partial \lambda}\mathcal{L} &= \frac{\partial}{\partial \lambda} \sum_{i=1}^{N} -\lambda + r_i \log \lambda - \log r_i! \\
&= \sum_{i=1}^{N} -1 + \frac{r_i}{\lambda} = 0 \\
&\implies \sum_{i=1}^{N} -\lambda + r_i = 0 \\
&\implies \boxed{\lambda = \frac{1}{N} \sum_{i=1}^{N} r_i}
\end{aligned}$$

Maximum Likelihood Summary

Maximum likelihood is easy if you can differentiate and solve simultaneous equations.

1. Write down the likelihood $p(\mathcal{D}|\theta_1, \theta_2, ...)$
2. Form the log-likelihood $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathcal{D}|\theta_1, \theta_2, ...)$
3. Take gradient of the log-likelihood and set it to zero $\frac{\partial}{\partial \theta_i}\mathcal{L} = 0$
4. Rearrange equations in the form $\theta_i = ...$

If $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$ is generated from the model $p(\mathbf{x}|\boldsymbol{\theta})$, then as $N \to \infty$, $\boldsymbol{\theta}_{\text{ML}} \to \boldsymbol{\theta}_{\text{true}}$.

Maximum Likelihood Logarithmic

The logarithmic distribution is

$$p(y) = \frac{(1-\theta)^y}{-y \log \theta}$$

where $y \in \{0, 1, 2, 3, ...\}$, $0 < \theta < 1$.

$$\begin{aligned}
\mathcal{L}(\theta) &= \sum_{i=1}^{N} \log \frac{(1-\theta)^{y_i}}{-y_i \log \theta} \\
&= -N \log(-\log \theta) + \sum_{i=1}^{N} y_i \log(1-\theta) - \log y_i \\
\frac{\partial}{\partial \theta} \mathcal{L}(\theta) &= -\frac{N}{\theta \log \theta} - \sum_{i=1}^{N} \frac{y_i}{1-\theta} = 0
\end{aligned}$$

This equation cannot actually be solved in closed form, and we have to resort to numerical approximations, which are slow to compute.

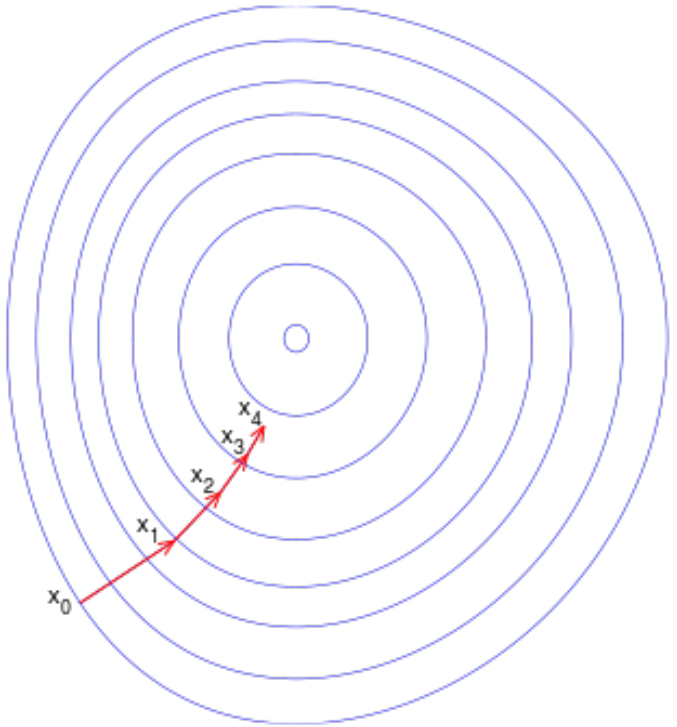
Numerical optimization: Gradient Descent

One way to find the maximum likelihood parameters for the logarithmic distribution is to use a numerical optimizer.

Gradient descent

Main idea: Find minima on $f(x)$ (called the *loss/objective function*).

1. Guess initial ML parameters $\boldsymbol{\theta}_0$ and set $t = 0$
2. Compute update $\Delta\boldsymbol{\theta}_t$
3. $\boxed{\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \eta\Delta\boldsymbol{\theta}_t}$, $(\eta \ll 1)$
4. Terminate?
5. Increment counter: $t = t + 1$
6. Goto 2



Numerical optimization: Steepest Descent

Want an update $\Delta\boldsymbol{\theta}_t$ guaranteed to improve/maintain the current loss value.

$$\boxed{\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \eta\Delta\boldsymbol{\theta}_t}$$

Many options: simplest is *steepest descent* direction (negative gradient)

$$\Delta\boldsymbol{\theta}_t := -\left.\frac{\partial f}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t}$$

Obviously if we perform maximization instead of minimization, we do steepest *ascent*.

Numerical optimization: Quadratic Example

e.g. What are the steepest descent updates for $f(x) = \frac{1}{2}x^2$?

$$\Delta x_t = -\frac{\partial}{\partial x}\frac{1}{2}x^2\Big|_{x=x_t} = -x_t$$

So

$$x_{t+1} \leftarrow x_t - \eta x_t = (1-\eta)x_t$$

Since $\eta \ll 1$ we see that regardless of initial value x_0, $\lim_{t\to\infty} x_t = 0 = x^*$.

Numerical optimization: Logistic Example

e.g. What are the steepest descent updates for the logarithmic log-likelihood

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \frac{(1-\theta)^{y_i}}{-y_i \log \theta} = -N \log(-\log \theta) + \sum_{i=1}^{N} y_i \log(1-\theta) - \log y_i ?$$

So

$$\Delta\theta_t = -\left.\frac{\partial}{\partial\theta}\mathcal{L}(\theta)\right|_{\theta=\theta_t} = \frac{N}{\theta \log \theta} + \frac{1}{1-\theta}\sum_{i=1}^{N} y_i$$

So

$$\theta_{t+1} \leftarrow \theta_t + \eta \left(\frac{N}{\theta_t \log \theta_t} + \frac{1}{1-\theta_t}\sum_{i=1}^{N} y_i \right)$$

The Exponential Family

Which distributions have analytical maximum likelihood solutions?

Most of the distributions we have looked at are fairly similar.

They have three main components:

$$p(x|\boldsymbol{\theta}) = \underbrace{\frac{1}{Z(\boldsymbol{\theta})}}_{\text{normalizer}} \cdot \overbrace{b(x)}^{\text{fnc of } x} \cdot \underbrace{\exp\left\{\boldsymbol{\theta}^\top \mathbf{t}(x)\right\}}_{\text{exp of linear fnc of } \boldsymbol{\theta}}$$

$\boldsymbol{\theta}$ is called the *natural parameters*, and $\mathbf{t}(x)$ is called the *sufficient statistics* of the distribution.

This may seem like an odd choice, but it has some very handy properties, which allow for lightning fast computation.

Counter examples of exponential family distributions are the uniform and logarithmic distributions.

The Exponential Family

First of all, let's see how some distributions belong to the exponential family.

$$\boxed{p(x|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} b(x) \exp\left\{\boldsymbol{\theta}^\top \mathbf{t}(x)\right\}}$$

Bernoulli

$$\begin{aligned} P(x|p) &= p^x (1-p)^{1-x} \\ &= e^{x \log p} e^{(1-x)\log(1-p)} \\ &= e^{x \log p + \log(1-p) - x \log(1-p)} \\ &= (1-p) e^{x \log \frac{p}{1-p}} \end{aligned}$$

The Exponential Family

First of all, let's see how some distributions belong to the exponential family.

$$p(x|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} b(x) \exp\left\{\boldsymbol{\theta}^\top \mathbf{t}(x)\right\}$$

Poisson

$$\begin{aligned} P(x|\lambda) &= \frac{1}{x!} e^{-\lambda} \lambda^x \\ &= e^{-\lambda} \frac{1}{x!} e^{x \log \lambda} \\ &= \frac{1}{e^{\lambda}} \frac{1}{x!} e^{x \log \lambda} \end{aligned}$$

First of all, let's see how some distributions belong to the exponential family.

$$\boxed{p(x|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} b(x) \exp\left\{\boldsymbol{\theta}^\top \mathbf{t}(x)\right\}}$$

Gaussian

$$\begin{aligned}
p(x|\mu, \sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\
&= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right\} \\
&= \frac{e^{-\frac{\mu^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}\right\} \\
&= \frac{e^{-\frac{\mu^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} \exp\left\{-\begin{bmatrix}\frac{\mu}{\sigma^2} & \frac{1}{2\sigma^2}\end{bmatrix}\begin{bmatrix}x \\ x^2\end{bmatrix}\right\}
\end{aligned}$$

The Exponential Family

The exponential family has the nice property that at the maximum likelihood solution

$$\mathbb{E}_{x\sim p(x|\boldsymbol{\theta})}\left[\mathbf{t}(x)\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbf{t}(x_i)$$

This will be shown to you in the werkcollege.

This means that the empirical mean of the sufficient statistics should match the theoretical expectation. This is called *moment matching*.

e.g. Bernoulli

$$\mathbb{E}_x[t(x)] = \mathbb{E}_x[x] = p = \frac{1}{N}\sum_{i=1}^{N}x_i$$

e.g. Gaussian

$$\mathbb{E}_x[\mathbf{t}(x)] = \mathbb{E}_x \begin{bmatrix} x \\ x^2 \end{bmatrix} = \begin{bmatrix} \mu \\ \mu^2 + \sigma^2 \end{bmatrix} = \frac{1}{N} \sum_{i=1}^{N} \begin{bmatrix} x_i \\ x_i^2 \end{bmatrix}$$

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

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

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

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