Maximum Likelihood Estimation

Nipun Batra

August 16, 2023

IIT Gandhinagar

Agenda

Revision - Prior, Posterior, MLE, MAP

Distributions, IID

MLE

MLE for Bernoulli Distribution

MLE for Univariate Normal Distribution

MLE for Multivariate Normal Distribution

MLE for Linear Regression

MLE for Logistic Regression

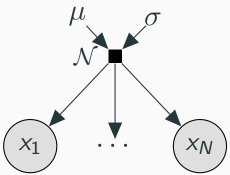
Revision - Prior, Posterior, MLE, MAP

Distributions, IID

Notebook

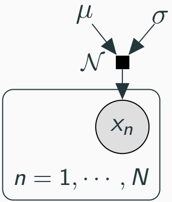
Graphical model

Assume model parameters are θ and data is D. We can write the joint probability distribution as:



Graphical model

Assume model parameters are θ and data is D. We can write the joint probability distribution as:



Factorisation

$$
\begin{aligned}
P(D|\theta) &= P(x_1, x_2, \ldots, x_n|\theta) \\
&= P(x_1|\theta) \cdot P(x_2|\theta) \cdot \ldots \cdot P(x_n|\theta)
\end{aligned}
$$

MLE

Pop Quiz

We have three courses: C1, C2, C3. Assume no student takes more than one course. The scores of students in these courses are normally distributed with the following parameters:

- C1: $\mu_1 = 80, \sigma_1 = 10$
- C2: $\mu_2 = 70, \sigma_2 = 10$
- C3: $\mu_3 = 90, \sigma_3 = 5$

Pop Quiz

We have three courses: C1, C2, C3. Assume no student takes more than one course. The scores of students in these courses are normally distributed with the following parameters:

- C1: $\mu_1 = 80, \sigma_1 = 10$
- C2: $\mu_2 = 70, \sigma_2 = 10$
- C3: $\mu_3 = 90, \sigma_3 = 5$

I randomly pick up a student and ask them their marks. They say 82. Which course do you think they are from? To keep things simple, for now assume that all three courses have equal number of students.

Pop Quiz

We have three courses: C1, C2, C3. Assume no student takes more than one course. The scores of students in these courses are normally distributed with the following parameters:

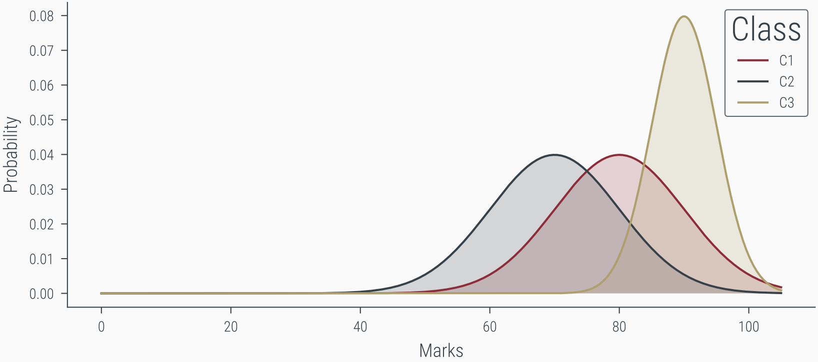
- C1: $\mu_1 = 80, \sigma_1 = 10$
- C2: $\mu_2 = 70, \sigma_2 = 10$
- C3: $\mu_3 = 90, \sigma_3 = 5$

I randomly pick up a student and ask them their marks. They say 82. Which course do you think they are from?

Most likely C1. But why?

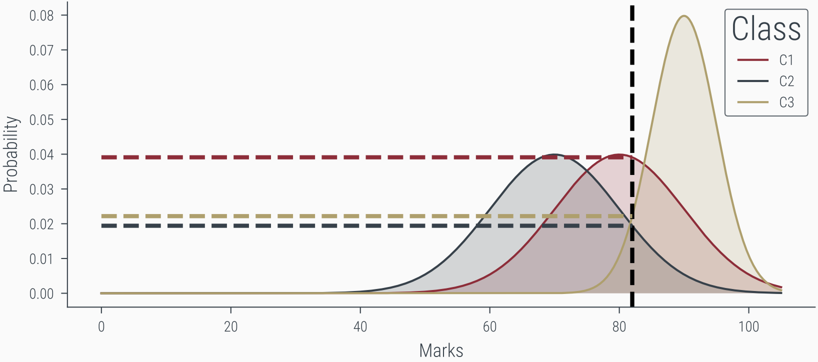
Pop Quiz

Let us plot the probability density functions of the three courses.



Pop Quiz

Let us plot the probability density functions of the three courses.



Notebook

Pop Quiz 2

Let us say we observed a value of 20. We know it came from a normal distribution with $\sigma = 1$. What is the most likely value of μ?

Pop Quiz 2

Let us say we observed a value of 20. We know it came from a normal distribution with $\sigma = 1$. What is the most likely value of μ?

20. But why?

Pop Quiz 2

Let us say we observed a value of 20. We know it came from a normal distribution with $\sigma = 1$. What is the most likely value of μ?

20. But why?

Let us evaluate probability density function at 20 for different values of μ for $\sigma = 1$, i.e., $f(x = 20|\mu, \sigma = 1)$.

Pop Quiz 2

Let us say we observed a value of 20. We know it came from a normal distribution with $\sigma = 1$. What is the most likely value of μ?

20. But why?

Let us evaluate probability density function at 20 for different values of μ for $\sigma = 1$, i.e., $f(x = 20|\mu, \sigma = 1)$.

Importantly, this is a function of μ and not x (which is fixed at 20).

Notebook

Pop Quiz 3

Let us now go back to our original problem. We have three courses: C1, C2, C3. Assume no student takes more than one course.

We ask two students their marks. The first student says 82 and the second student says 72. Which course do you think they are from? Assumption: Both are from the same course.

Let us create a table of probabilities for each course:

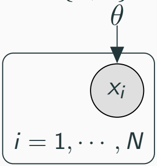
MLE for Bernoulli Distribution

MLE for Bernoulli Distribution

The probability mass function of a bernoulli distribution is given by:

$$f(x|\theta) = \theta^x (1-\theta)^{(1-x)} \tag{1}$$

Let us assume we have a dataset $D = \{x_1, x_2, \ldots, x_n\}$, where each x_i is an independent sample from the above distribution and $x_i \in \{0, 1\}$. We want to estimate the parameter θ from the data.



Log Likelihood Function

Our likelihood function is given by:

$$P(D|\theta) = \mathcal{L}(\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{2}$$

Log-likelihood function:

$$\log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta) \tag{3}$$

Simplifying the above equation, we get:

$$\begin{aligned} \log \mathcal{L}(\theta) &= \sum_{i=1}^{n} \log f(x_i|\theta) \\ &= \sum_{i=1}^{n} \log \left(\theta^{x_i} (1-\theta)^{(1-x_i)} \right) \end{aligned}$$

$$\log \mathcal{L}(\theta) = \sum_{i=1}^{n} \left(\log \left(\theta^{x_i}\right) + \log \left((1-\theta)^{(1-x_i)} \right) \right)$$
$$= \sum_{i=1}^{n} \left(x_i \log(\theta) + (1-x_i) \log(1-\theta) \right)$$

Log Likelihood Function for Bernoulli Distribution

Log-likelihood function for Bernoulli distributed data is:

$$\log \mathcal{L}(\theta) = \sum_{i=1}^{n} (x_i \log(\theta) + (1-x_i) \log(1-\theta))$$

Maximum Likelihood Estimate for θ

To find the MLE for θ, we differentiate the log-likelihood function with respect to θ and set it to zero:

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

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

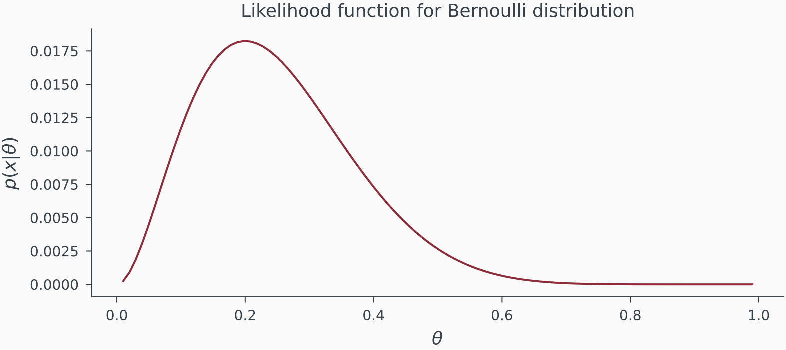
$$\theta = \frac{\sum_{i=1}^{n} x_i}{n}$$

Maximum Likelihood Estimate for θ

MLE of θ, denoted as $\hat{\theta}_{\text{MLE}}$, is given by:

$$\hat{\theta}_{\text{MLE}} = \frac{\sum_{i=1}^{n} x_i}{n}$$

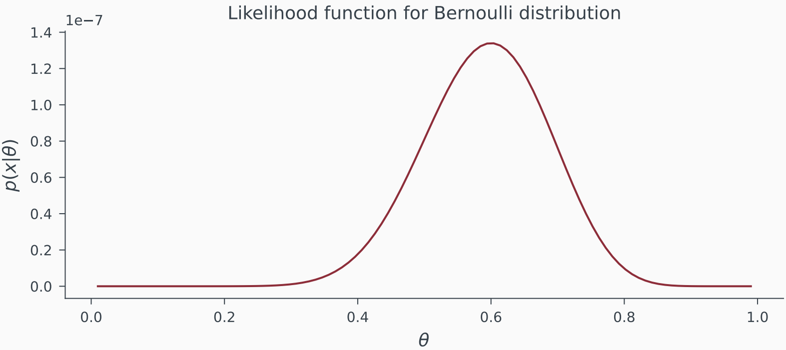
The data, D consists of the results of coin tosses which can be H/T. Let suppose $D_1 = (T, H, T, T, T, T, H, T, T, T)$. By calculating θ_{MLE}, we get its value as 0.2. We vary θ from 0 to 1 and calculate the likelihood at each value. We find that the likelihood is maximum around $\theta = 0.2$ which is our MLE estimate.



Data, $D_2 = (H, H, H, H, H, H, T, T, T, T)$.

True $\theta = 0.6$.

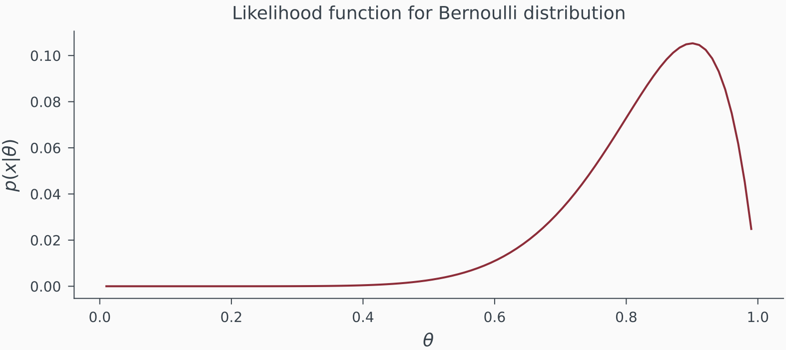
Corresponding plot of likelihood $P(D|\theta)$ V/s θ is given below:



Data, $D_3 = (H, H, H, H, H, H, T, H, H, T)$.

True $\theta = 0.9$.

Corresponding plot of likelihood $P(D|\theta)$ V/s θ is given below:



MLE for Univariate Normal Distribution

Univariate Normal Distribution

The probability density function of a univariate normal distribution is given by:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{4}$$

Let us assume we have a dataset $D = \{x_1, x_2, \ldots, x_n\}$, where each x_i is an independent sample from the above distribution. We want to estimate the parameters $\theta = \{\mu, \sigma\}$ from the data.

Our likelihood function is given by:

$$P(D|\theta) = \mathcal{L}(\mu, \sigma^2) = \prod_{i=1}^{n} f(x_i|\mu, \sigma^2) \tag{5}$$

Log Likelihood Function

Log-likelihood function:

$$\log \mathcal{L}(\mu, \sigma^2) = \sum_{i=1}^{n} \log f(x_i|\mu, \sigma^2) \tag{6}$$

Simplifying the above equation, we get:

$$\begin{aligned}
\log \mathcal{L}(\mu, \sigma^2) &= \sum_{i=1}^{n} \log f(x_i|\mu, \sigma^2) \\
&= \sum_{i=1}^{n} \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2} \right) \right) \\
&= \sum_{i=1}^{n} \left(\log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \log\left(\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2} \right) \right) \right)
\end{aligned}$$

$$
\begin{aligned}
\log \mathcal{L}(\mu, \sigma^2) &= \sum_{i=1}^{n} \left(\log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{(x_i - \mu)^2}{2\sigma^2} \right) \\
&= \sum_{i=1}^{n} \left(-\frac{1}{2} \log(2\pi\sigma^2) - \frac{(x_i - \mu)^2}{2\sigma^2} \right) \\
&= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2
\end{aligned}
$$

Log Likelihood Function for Univariate Normal Distribution

Log-likelihood function for normally distributed data is:

$$
\log \mathcal{L}(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi) - n \log(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2
$$

Maximum Likelihood Estimate for μ

To find the MLE for μ, we differentiate the log-likelihood function with respect to μ and set it to zero:

$$\frac{\partial \log \mathcal{L}(\mu, \sigma^2)}{\partial \mu} = \frac{\partial}{\partial \mu}\left(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) = 0$$

$$\frac{\partial}{\partial \mu}\left(\sum_{i=1}^{n}(x_i - \mu)^2\right) = 0$$

Maximum Likelihood Estimate for μ

MLE of μ, denoted as $\hat{\mu}_{\text{MLE}}$, is given by:

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

MLE for σ for normally distributed data

Recall that the log-likelihood function is given by:

$$\log \mathcal{L}(\mu, \sigma^2) = \sum_{i=1}^{n} \log f(x_i|\mu, \sigma^2) \tag{7}$$

Let us find the maximum likelihood estimate of σ^2 now. We can do this by taking the derivative of the log-likelihood function with respect to σ^2 and equating it to zero.

$$\frac{\partial \log \mathcal{L}(\mu, \sigma^2)}{\partial \sigma^2} = \sum_{i=1}^{n} \frac{\partial \log f(x_i|\mu, \sigma^2)}{\partial \sigma^2} = 0 \tag{8}$$

MLE for σ for normally distributed data

Log Likelihood Function for Univariate Normal Distribution

Log-likelihood function for normally distributed data is:

$$\log \mathcal{L}(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Now, we can differentiate the log-likelihood function with respect to σ and equate it to zero.

MLE for σ for normally distributed data

$$\frac{\partial}{\partial \sigma} \log \mathcal{L}(\mu, \sigma^2) = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

Multiplying through by σ^3, we have:

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

Maximum Likelihood Estimate for σ^2

MLE of σ^2, denoted as $\hat{\sigma}^2_{\text{MLE}}$, is given by:

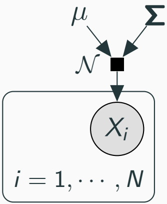
$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

MLE for Multivariate Normal Distribution

MLE for Multivariate Normal Distribution

The probability density function of a multivariate normal distribution is given by:

$$f(x|\mu, \Sigma) = (2\pi)^{-\frac{k}{2}} \det(\Sigma)^{-\frac{1}{2}} exp^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \tag{9}$$

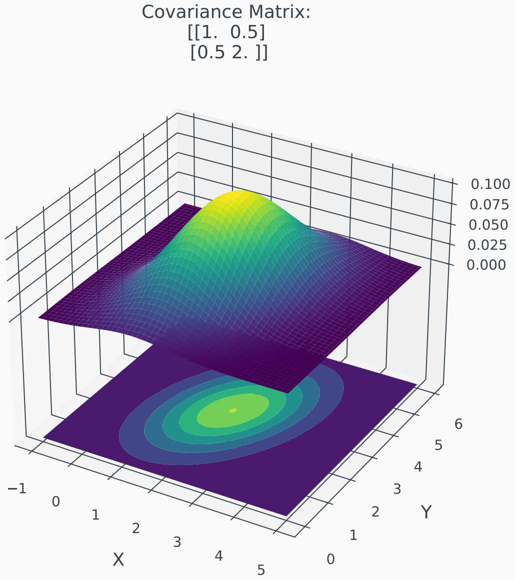


Let us assume we have a dataset $D = \{x_1, x_2, \dots, x_n\}$, where each x_i is an independent sample from the above distribution. We want to estimate the parameters $\theta = \mu, \sigma$ from the data.

Our likelihood function is given by:

$$P(D|\theta) = \mathcal{L}(\mu, \Sigma) = \prod_{i=1}^{n} f(x_i|\mu, \Sigma) \tag{10}$$

For example: A bivariate Normal distribution can be visualized as given below:



Log Likelihood Function

Log-likelihood function:

$$\log \mathcal{L}(\mu, \Sigma) = \sum_{i=1}^{n} \log f(x_i|\mu, \Sigma) \tag{11}$$

Simplifying the above equation, we get:

$$\begin{aligned}
\log \mathcal{L}(\mu, \Sigma) &= \sum_{i=1}^{n} \log f(x_i|\mu, \Sigma) \\
&= \sum_{i=1}^{n} \log \left((2\pi)^{-\frac{k}{2}} \det(\Sigma)^{-\frac{1}{2}} exp^{-\frac{1}{2}(x_i-\mu)^T \Sigma^{-1} (x_i-\mu)} \right) \\
&= \sum_{i=1}^{n} \log((2\pi)^{-\frac{k}{2}}) + \sum_{i=1}^{n} \log(\det(\Sigma)^{-\frac{1}{2}}) + \\
&\quad \sum_{i=1}^{n} \log(exp^{-\frac{1}{2}(x_i-\mu)^T \Sigma^{-1} (x_i-\mu)}))
\end{aligned}$$

Continuing, we get:

$$= -\frac{kn}{2}\log(2\pi) - \frac{n}{2}\log(\Sigma) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T\Sigma^{-1}(x_i - \mu)$$

Log Likelihood Function for Multivariate Normal Distribution

Log-likelihood function for multivariate normally distributed data is:

$$-\frac{kn}{2}\log(2\pi) - \frac{n}{2}\log(\Sigma) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T\Sigma^{-1}(x_i - \mu)$$

Maximum Likelihood Estimate for μ

To find the MLE for μ, we differentiate the log-likelihood function with respect to μ and set it to zero:

$$
\begin{aligned}
&= \frac{\partial}{\partial \mu}\left(-\frac{kn}{2}\log(2\pi) - \frac{n}{2}\log(\Sigma) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right) \\
&= \frac{\partial}{\partial \mu}\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right) \\
&= -\frac{1}{2}\sum_{i=1}^{n}\left(\Sigma^{-1}(x_i-\mu) + (x_i-\mu)^T\Sigma^{-1}\right) = 0 \\
&= -\frac{1}{2}\sum_{i=1}^{n} 2\Sigma^{-1}(x_i-\mu) = 0
\end{aligned}
$$

$$\text{as } (x_i-\mu)^T\Sigma^{-1} = \Sigma^{-1}(x_i-\mu)$$

$$
\begin{aligned}
&= \Sigma^{-1} \sum_{i=1}^{n} (x_i - \mu) = 0 \\
&= \sum_{i=1}^{n} (x_i) - n\mu = 0 \\
\mu &= \frac{\sum_{i=1}^{n} x_i}{n}
\end{aligned}
$$

Maximum Likelihood Estimate for μ

MLE of μ, denoted as $\hat{\mu}_{\text{MLE}}$, is given by:

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

MLE for Σ for multivariate normally distributed data

Recall that the log-likelihood function is given by:

$$\log \mathcal{L}(\mu, \Sigma) = \sum_{i=1}^{n} \log f(x_i|\mu, \Sigma) \tag{12}$$

Let us find the maximum likelihood estimate of Σ now. We can do this by taking the derivative of the log-likelihood function with respect to Σ and equating it to zero.

$$\frac{\partial \log \mathcal{L}(\mu, \Sigma)}{\partial \Sigma} = \sum_{i=1}^{n} \frac{\partial \log f(x_i|\mu, \Sigma)}{\partial \Sigma} = 0 \tag{13}$$

After differentiating and simplifying, we get:

$$\Sigma = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T$$

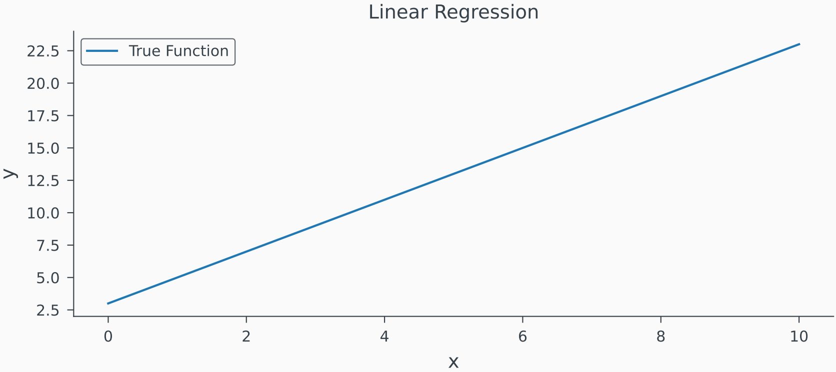
Maximum Likelihood Estimate for Σ

MLE of Σ, denoted as $\hat{\Sigma}_{\mathrm{MLE}}$, is given by:

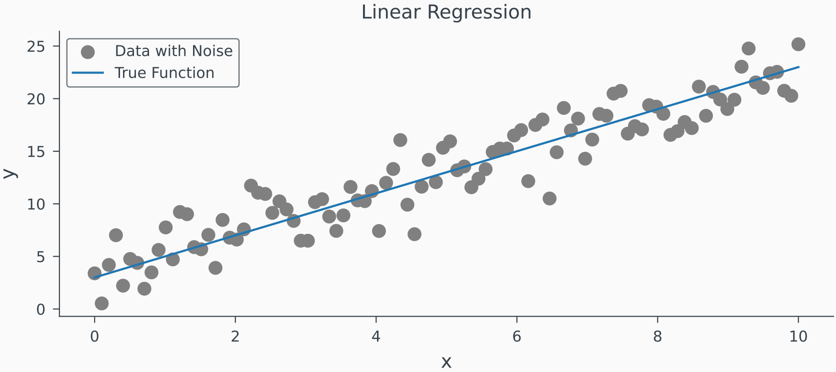
$$\hat{\Sigma}_{\mathrm{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T$$

MLE for Linear Regression

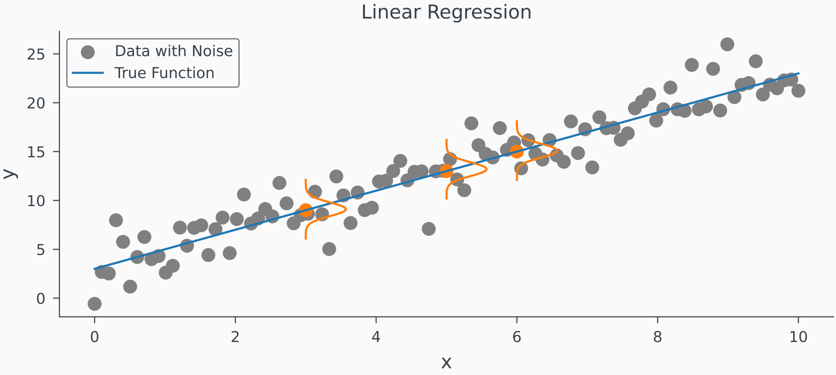
MLE for Linear Regression



MLE for Linear Regression

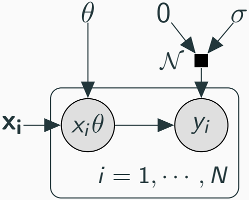


MLE for Linear Regression



We consider a regression problem with the likelihood function:
$p(y|x) = \mathbb{N}(y|f(x), \sigma^2)$.
Let us assume we have a dataset
$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where $x_i \in R^d, y_i \in R$.

The functional relationship between x and y is given as
$y = f(\mathbf{x}) + \epsilon$ where $\epsilon \sim \mathbb{N}(0, \sigma^2)$.
Thus $f(\mathbf{x}) = x^T\theta$.



Our likelihood function (Normal distribution) is given by:

$$P(\mathcal{Y}|\mathcal{X},\theta) = p(y_1,\ldots,y_n|x_1,\ldots,x_n,\theta) = \prod_{i=1}^{n} p(y_i|x_i,\theta) \tag{14}$$

The MLE equation is given by:

$$\theta_{MLE} \in \arg_\theta \max p(Y|X,\theta) \tag{15}$$

Maximizing the likelihood $\equiv$ Maximizing the log likelihood $\equiv$ Minimizing the negative log likelihood.
Taking negative log, we get:

$$-\log p(\mathcal{Y} \mid \mathcal{X},\boldsymbol{\theta}) = -\log \prod_{i=1}^{N} p\left(y_i \mid \boldsymbol{x}_i,\boldsymbol{\theta}\right) = -\sum_{i=1}^{N} \log p\left(y_i \mid \boldsymbol{x}_i,\boldsymbol{\theta}\right)$$

For a given point (x_i, y_i),

$$-\log p\left(y_i \mid \boldsymbol{x}_i,\boldsymbol{\theta}\right) = -\frac{1}{2\sigma^2}\left(y_i - \boldsymbol{x}_i^{\top}\boldsymbol{\theta}\right)^2 + const$$

Thus the negative log likelihood is simplified to:

$$
\begin{aligned}
-\mathcal{L}(\boldsymbol{\theta}) &:= -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \boldsymbol{x}_i^\top\boldsymbol{\theta}\right)^2 \\
&= -\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) = -\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2
\end{aligned}
$$

NLL Equation

NLL is equal to:

$$
-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2
$$

This is none other than squared error loss!

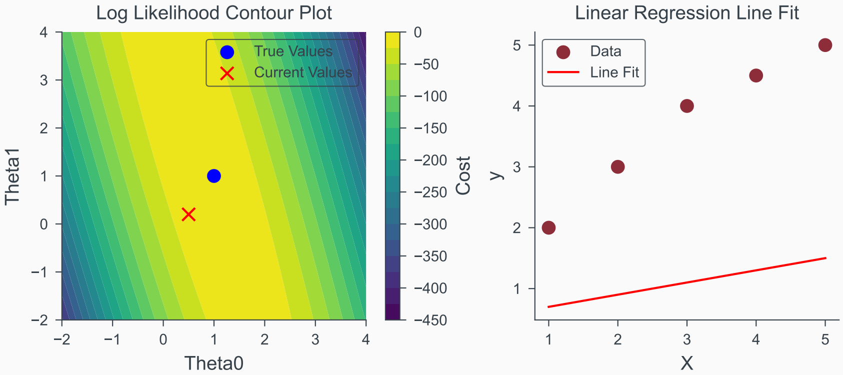
To minimize this, we differentiate wrt θ. In the end, we get:

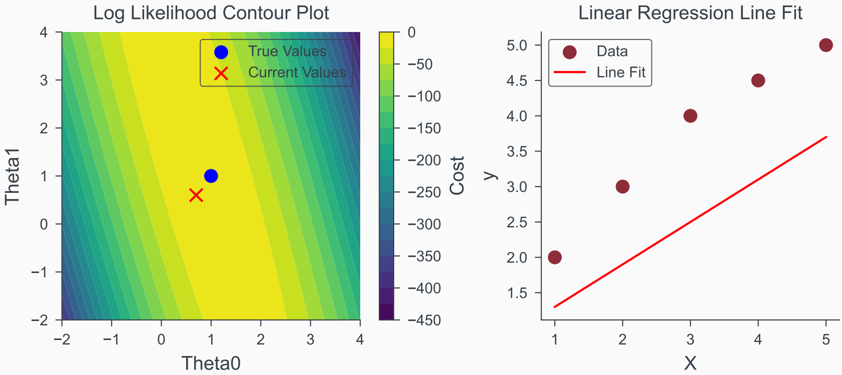
$$\theta = (X^T X)^{-1} X^T y \tag{16}$$

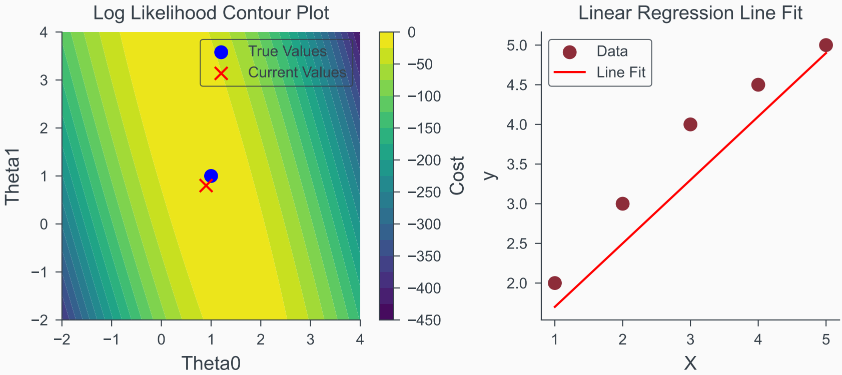
Maximum Likelihood Estimate for θ

MLE of θ, denoted as $\hat{\theta}_{\text{MLE}}$, is given by:

$$\hat{\theta}_{\text{MLE}} = (X^T X)^{-1} X^T y$$


Log Likelihood Contour Plot
True Values
Current Values
Theta1
Theta0
Cost
0
−50
−100
−150
−200
−250
−300
−350
−400
−450
Linear Regression Line Fit
Data
Line Fit
y
X

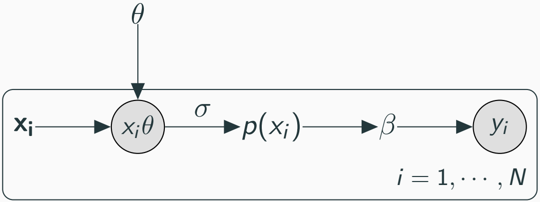

Log Likelihood Contour Plot
True Values
Current Values
Theta0
Theta1
Cost
0
−50
−100
−150
−200
−250
−300
−350
−400
−450
Linear Regression Line Fit
Data
Line Fit
X
y



Log Likelihood Contour Plot
True Values
Current Values
Theta1
Theta0
Cost
Linear Regression Line Fit
Data
Line Fit
y
X

MLE for Logistic Regression

MLE for Logistic Regression



Binary Classification:

The probability distribution in case of Logistic Regression considering two classes is Bernoulli distribution but there is a slight difference. The probability is now the output of the logistic function. Parameters are $\theta = [\theta_0, \theta_1]$.

$$p = P(Y = 1|X) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 X)}} \tag{17}$$

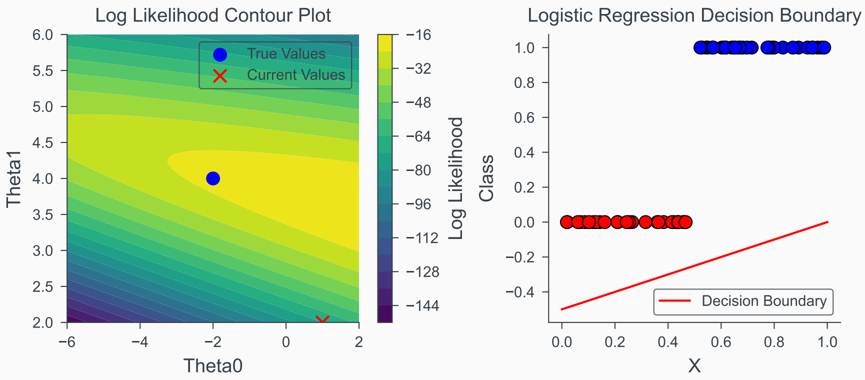
Rewriting the likelihood in this manner:

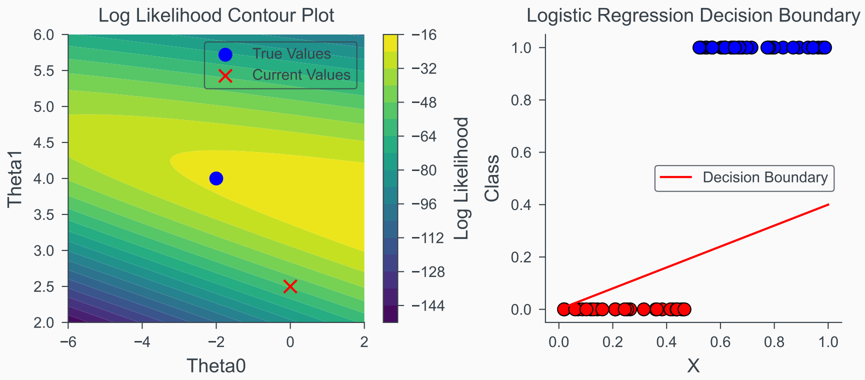
$$\begin{aligned} L(\theta) &= \prod_{y_i=1} p(x_i) \prod_{y_i=0} (1 - p(x_i)) \\ &= \prod \left(p(x_i)^{y_i} (1 - p(x_i))^{1-y_i} \right) \end{aligned}$$

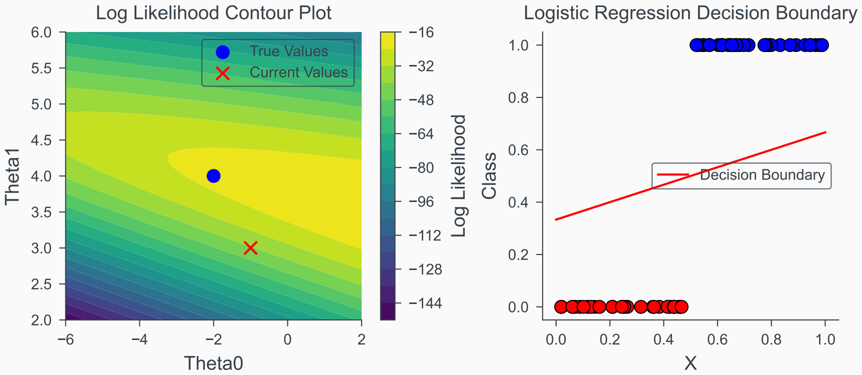
Taking log on both sides:

$$\log(L(\theta)) = \sum_{i=1}^{n} y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i))$$

If we multiply this by $-\frac{1}{n}$,this is nothing but the binary cross entropy loss function!

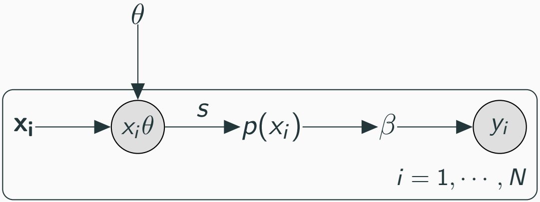

Log Likelihood Contour Plot
Theta1
Theta0
True Values
Current Values
Log Likelihood
−16
−32
−48
−64
−80
−96
−112
−128
−144
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
−6
−4
−2
0
2
Logistic Regression Decision Boundary
Class
X
Decision Boundary
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0.0
0.2
0.4
0.6
0.8
1.0


Log Likelihood Contour Plot
True Values
Current Values
Theta0
Theta1
Log Likelihood
Logistic Regression Decision Boundary
Decision Boundary
X
Class


Log Likelihood Contour Plot
True Values
Current Values
Theta1
Theta0
Log Likelihood
−16
−32
−48
−64
−80
−96
−112
−128
−144
Logistic Regression Decision Boundary
Decision Boundary
Class
X

Coin toss V/s Binary Logistic Regression

Coin toss	Binary Logistic Regression
Likelihood$=\theta^t(1-\theta)^{1-t}$	Likelihood$=(\sigma(x^T\theta))^t(1-\sigma(x^T\theta))^{1-t}$
Outcome is Head/Tail	Outcome is out of two possible classes
Parameter is scalar	Parameter is vector with two values



Multi-class Classification:
The probability distribution in case of Logistic Regression considering more than two classes is Categorical distribution. The probability is now the output of the softmax function. Parameters are $\theta = [\theta_0, \theta_1, \ldots, \theta_k]$.

$$p = P(Y = i|X) = \frac{e^{\theta x_i}}{\sum_{j=1}^{n} e^{\theta x_j}} \tag{18}$$

Now:

$$L(\theta) = \prod_{i=1}^{n} \prod_{j=1}^{K} p^j(x_i)$$

Taking log on both sides:

$$\log(L(\theta)) = \sum_{i=1}^{n} \sum_{j=1}^{K} y_i^k \log(p^k(x_i))$$

If we multiply this by $-\frac{1}{n}$,this is nothing but the cross entropy loss function!

Now if we differentiate this wrt θ, it is difficult to find a analytical solution with it. Thus in order to solve for MLE for logistic regression, methods like Gradient Descent, Newton-Raphson, etc. are used. For example through Gradient descent, the below decision boundary i.e. θ has been calculated.

Binary V/s Multiclass Logistic Regression

Binary Logistic Regression	**Multiclass Logistic Regression**
Binary Cross Entropy Loss	Cross Entropy Loss
$p(x) = \sigma(x^T\theta)$	$p(x) = s(x^T\theta)$
Bernoulli Likelihood	Categorical Likelihood

Random variable and Random sample

Random variable: $X : \Omega \to \mathbb{R}$ is a function from the sample space to the real line.

Random sample: Collection of n independent and identically distributed (i.i.d.) random variables $X_1, X_2, X_3, \ldots, X_n$. A group of experiments constitutes a sample.

For example:
Random variable: Y (possible outcomes 1 to 6)
Random sample: 4,2,6 (outcomes of three consecutive die tosses)

Bias of an Estimator

Bias of an Estimator

The bias of an estimator $\hat{\theta}$ of a parameter θ is defined as:

$$\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

where $\mathbb{E}(\hat{\theta})$ is the expected value of the estimator $\hat{\theta}$.

- An estimator is said to be unbiased if $\text{Bias}(\hat{\theta}) = 0$.
- An estimator is said to be biased if $\text{Bias}(\hat{\theta}) \neq 0$.

Bias of an Estimator: $\hat{\mu}_{MLE}$

Question: What is the expectation of $\hat{\mu}_{MLE}$ calculated over? What is the source of randomness?

If $X_i's$ are normally distributed random variables with mean μ and variance σ^2 respectively, then $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$.

Recall that if an estimator $\hat{\theta}$ of a parameter θ is unbiased then: $\mathbb{E}(\hat{\theta}) = \theta$.

Bias of an Estimator: $\hat{\mu}_{MLE}$

$$
\begin{aligned}
\mathbb{E}(\hat{\mu}_{MLE}) &= \mathbb{E}(\bar{X}) \\
&= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) \\
&= \frac{1}{n}(n\mu) = \mu
\end{aligned}
$$

Estimator $\hat{\mu}_{MLE}$ is unbiased

$\mathbb{E}(\hat{\mu}_{MLE}) = \mu$

Bias of σ^2_{MLE}

The MLE of σ^2 is given by

$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

$$
\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}x_i + \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}\frac{1}{n}\sum_{i=1}^{n} x_i + \bar{x}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2
\end{aligned}
$$

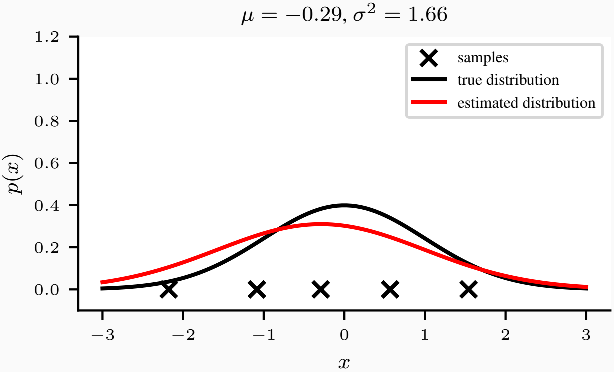
Bias of σ^2_{MLE}

$$\mathbb{E}(\hat{\sigma}^2) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right] = \left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i^2)\right] - \mathbb{E}(\bar{X}^2)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2 + \mu^2\right) - \left(\frac{\sigma^2}{n} + \mu^2\right)$$
$$= \frac{1}{n}\left(n\sigma^2 + n\mu^2\right) - \frac{\sigma^2}{n} - \mu^2$$
$$= \sigma^2 - \frac{\sigma^2}{n} = \frac{n\sigma^2 - \sigma^2}{n} = \frac{(n-1)\sigma^2}{n}$$

Estimator $\hat{\sigma}_{MLE}$ is biased

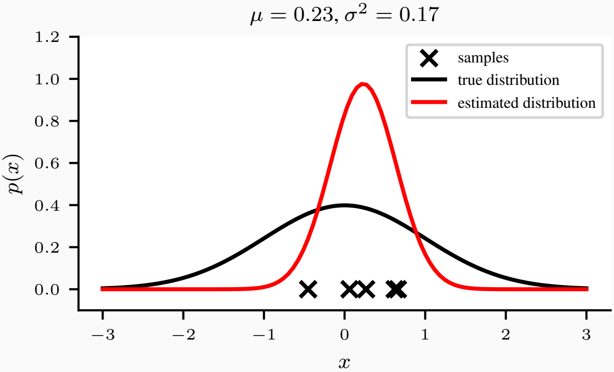
$\mathbb{E}(\hat{\sigma}_{MLE}) = \frac{(n-1)\sigma^2}{n}$

Bias



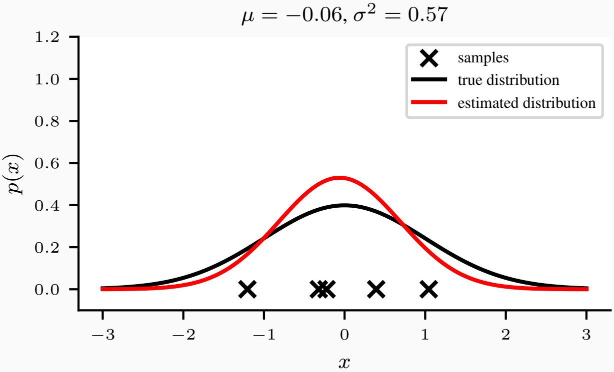
$\mu = -0.29, \sigma^2 = 1.66$
samples
true distribution
estimated distribution
$p(x)$
x
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3

Bias



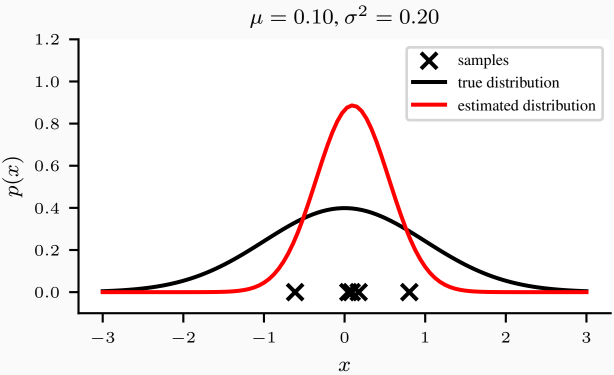
$\mu = 0.23, \sigma^2 = 0.17$
samples
true distribution
estimated distribution
$p(x)$
x
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

Bias



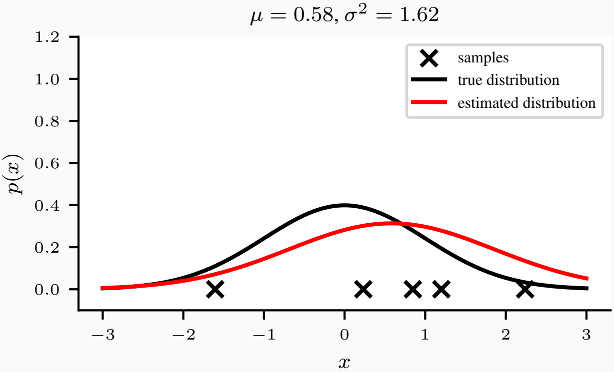
$\mu = -0.06, \sigma^2 = 0.57$
samples
true distribution
estimated distribution
$p(x)$
x
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

Bias



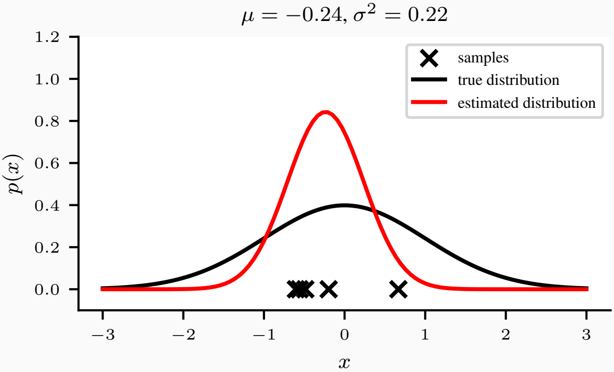
$\mu = 0.10, \sigma^2 = 0.20$
samples
true distribution
estimated distribution
$p(x)$
x
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

Bias



$\mu = 0.58, \sigma^2 = 1.62$
1.2
1.0
0.8
0.6
0.4
0.2
0.0
$p(x)$
samples
true distribution
estimated distribution
−3
−2
−1
0
1
2
3
x

Bias



$\mu = -0.24, \sigma^2 = 0.22$
samples
true distribution
estimated distribution
$p(x)$
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
x

MAP Plate Notation for Beta-Bernoulli