機械学習特論
~理論とアルゴリズム~

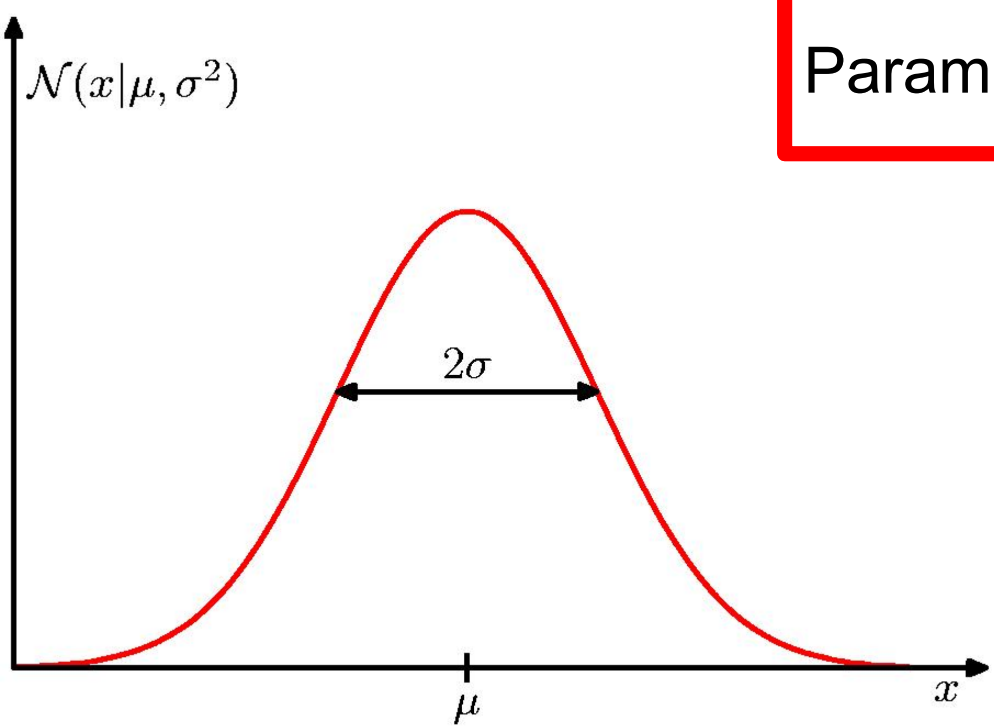
(Gaussian distribution and Maximum Likelihood Estimate)

講師: 西郷浩人

Gaussian (Normal) distribution

The Gaussian Distribution

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



Parameters: mean μ and variance σ2

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

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

Preparation

•

gauss1d_plot.m

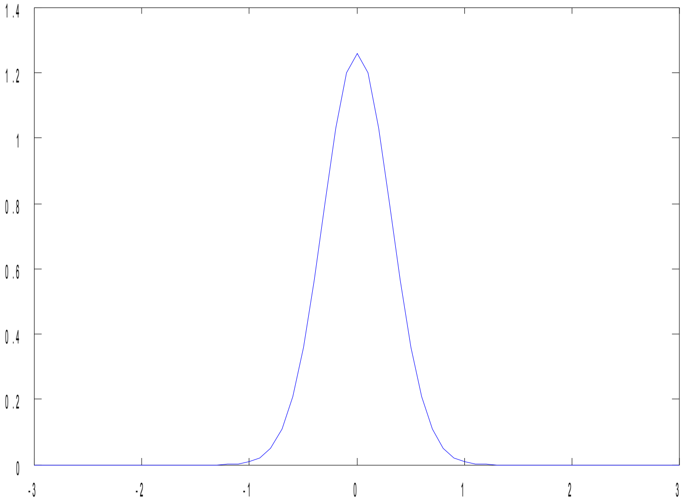
```
function []=gauss1d_plot(x,mu,sigma)

for i=1:length(x)
  z(i)=g1_pdf(x(i),mu,sigma);
end

figure(1); clf
plot(x,z);
%print -deps gauss1d_pdf.eps

function [z]=g1_pdf(x,mu,sigma)
z=(2*pi*sigma.^2)^(-1/2)*exp(-(x-mu).^2./(2*sigma.^2));
```

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

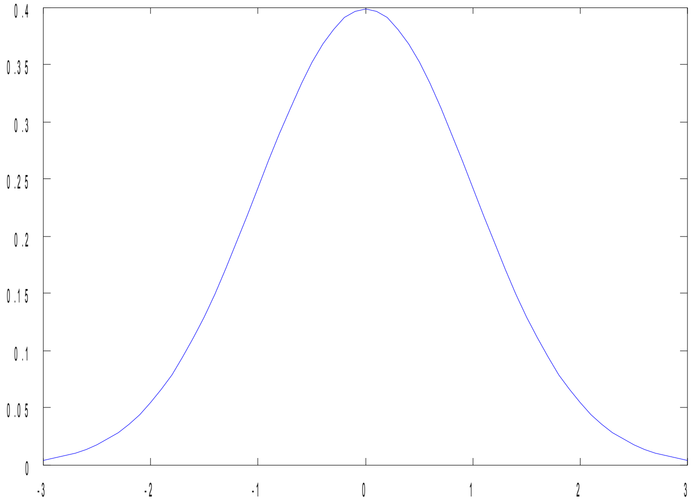
Plotting 1D Gaussian distribution

-

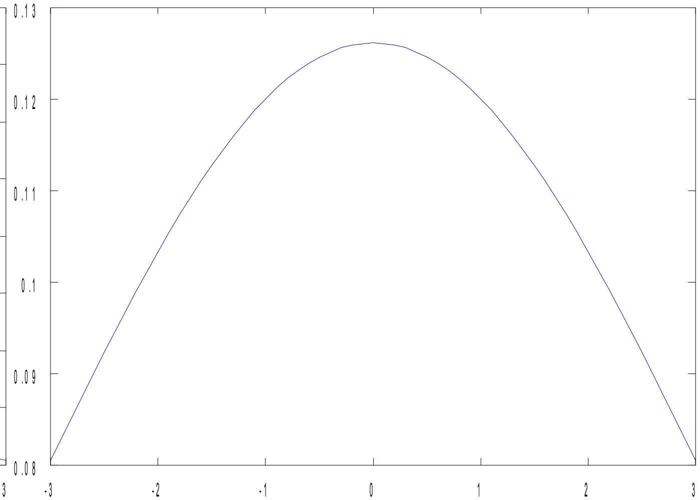
```
mean=0;
sigma=1;
x=[-3:0.1:3];
gauss1d_plot(x,mean,sigma);
```



σ=0.1



σ=1



σ=10

$$N(x;\mu,\sigma^2)=\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(\frac{-1}{2\sigma^2}(x-\mu)^2\right)$$

Gaussian Mean and Variance

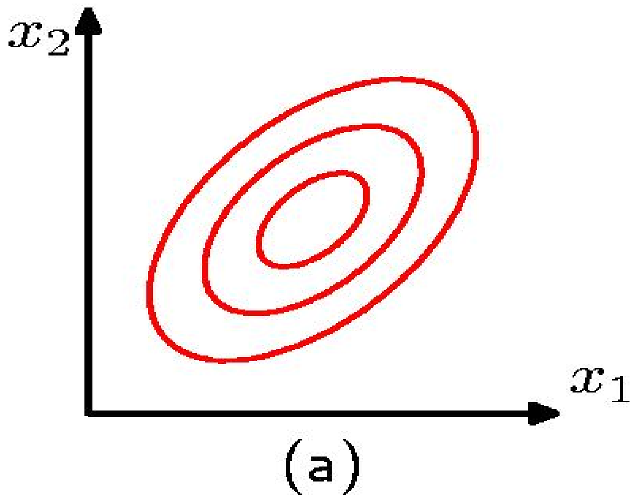
$$E[x]=\int_{-\infty}^{\infty} N(x|\mu,\sigma^2)\,x\,dx=\mu$$

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

$$var[x]=E[x^2]-E[x]^2=\sigma^2$$

The Multivariate Gaussian

$$N(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(\frac{-1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$



Parameters:

Mean vector $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$

Covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$

σ_1^2, σ_2^2 are variance, σ_12 is covariance

preparation

•

gauss2d_plot.m

```
function []=gauss2d_plot(x,y,Mu,Sigma)

for i=1:length(x)
 for j=1:length(y)
  z(i,j)=g2_pdf(x(i),y(j),Mu,Sigma);
 end
end

figure(1); clf
surf(x,y,z); view(45,60);
%print -deps gauss2d_pdf_surf.eps

figure(2);clf
contour(x,y,z);
%print -deps gauss2d_pdf_contour.eps
```

g2_pdf.m

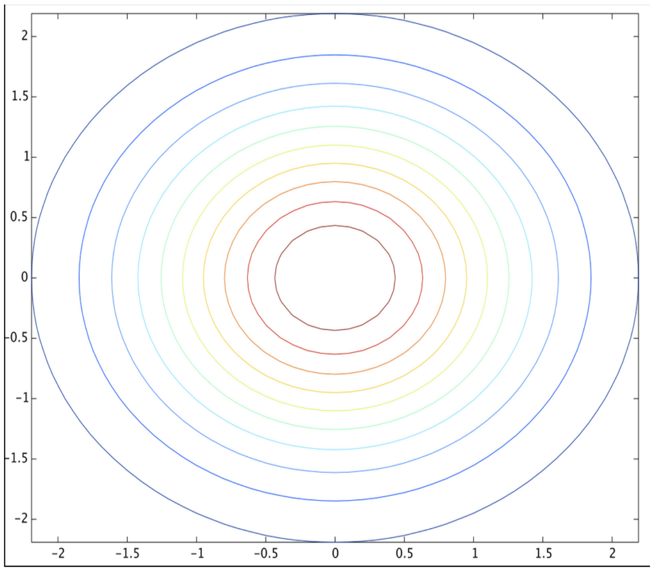
```
function [z] = g2_pdf(x,y,Mu,Sigma)
d=sqrt(det(Sigma));
v=[x;y]-Mu;
z=1/(2*pi*d)*exp(-1/2*v'*inv(Sigma)*v);
```

$$N(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(\frac{-1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

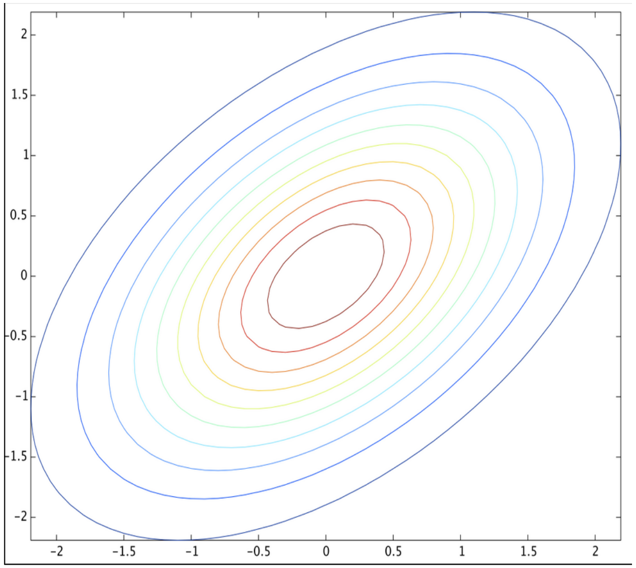
Plotting 2D Gaussian distribution

-

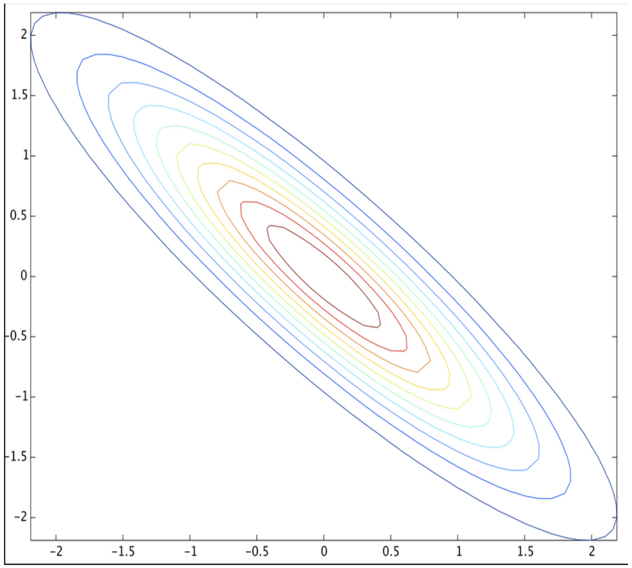
```
mean=[0 0]';
sigma=[1 0; 0 1];
x=[-3:0.1:3];
y=[-3:0.1:3];
gauss2d_plot(x,y,mean,sigma);
```



$$\boldsymbol{\Sigma}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$$



$$\boldsymbol{\Sigma}=\begin{bmatrix}1 & 0.5\\ 0.5 & 1\end{bmatrix}$$



$$\boldsymbol{\Sigma}=\begin{bmatrix}1 & -0.9\\ -0.9 & 1\end{bmatrix}$$

$$N(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(\frac{-1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Ex. 1

- Explain how the distribution changes with respect to the change in covariance matrix.

Parameter estimation via Maximum Likelihood

Likelihood and Maximum Likelihood

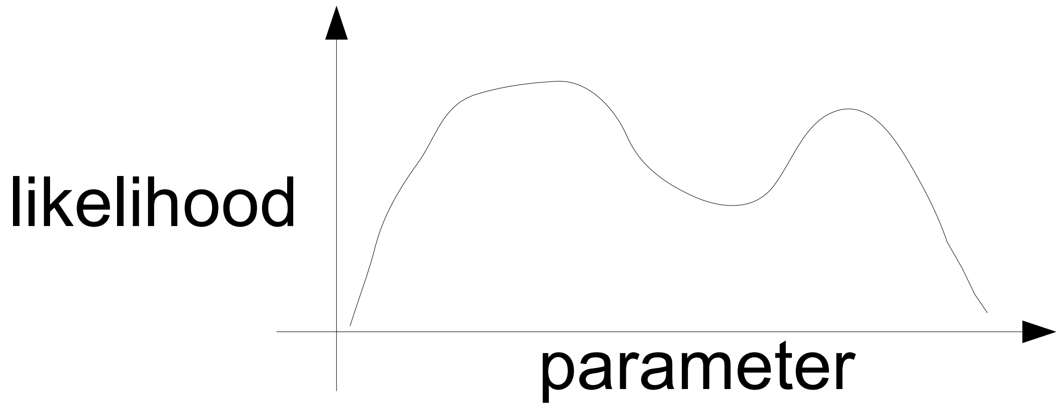
- Given a set of parameters $\boldsymbol{\theta}=\{\theta_{1,}\theta_{2,..}..,\theta_n\}$, we have data distribution $\boldsymbol{D}=\{x_{1,}x_{2,..}..,x_n\}$. The probability of the data generated (likelihood) is defined as

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

  - Higher likelihood suggests that the data are likely to be generated. We look for such parameter sets **θ.**
  - A method to look for **θ** that maximizes likelihood is called Maximum Likelihood (最尤法), and represented as

$$\hat{\boldsymbol{\theta}}=argmax_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta})$$

Maximizing Likelihood

- 

- Even if we do not know the shape of likelihood space, local maximum occurs at a point where a derivative with respect to parameter is zero (necessary condition).

$$\frac{\partial L(\hat{\boldsymbol{\theta}}_{ML})}{\partial \hat{\boldsymbol{\theta}}_{ML}} = \mathbf{0}$$

- Sufficient conditions should be checked in each case.

Maximum Likelihood continued.

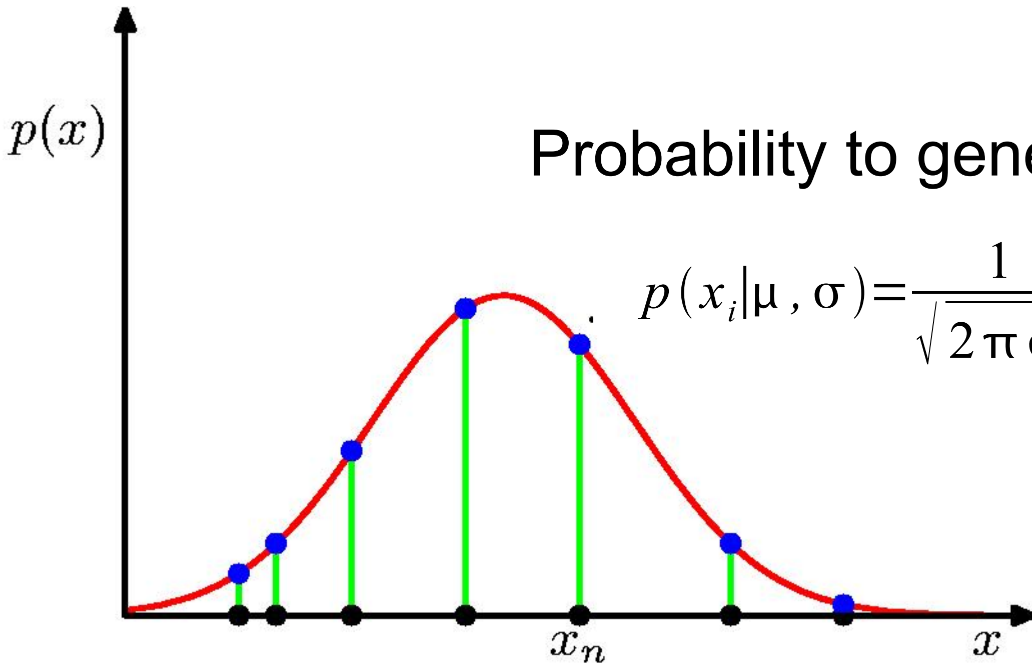
- Necessary condition

$$\frac{\partial L(\hat{\boldsymbol{\theta}}_{ML})}{\partial \hat{\boldsymbol{\theta}}_{ML}} = \mathbf{0}$$

- In practice, log L is maximized instead of L, because
  - parameter to max L = parameter to max log L
  - Addition is faster than multiplication, and can avoid underflow.

$$\frac{\partial \log L(\hat{\boldsymbol{\theta}}_{ML})}{\partial \hat{\boldsymbol{\theta}}_{ML}} = \mathbf{0}$$

Example on Gaussian distribution



Probability to generate each data point

$$p(x_i|\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp(\frac{-(x-\mu)^2}{2\sigma^2})$$

Likelihood (probability to generate the whole data points)

$$L(\mu,\sigma^2)=\prod_{i=1}^{n} p(x_i|\mu,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp(\frac{-(x_i-\mu)^2}{2\sigma^2})$$

Maximum Likelihood Estimate

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

By taking log

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

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

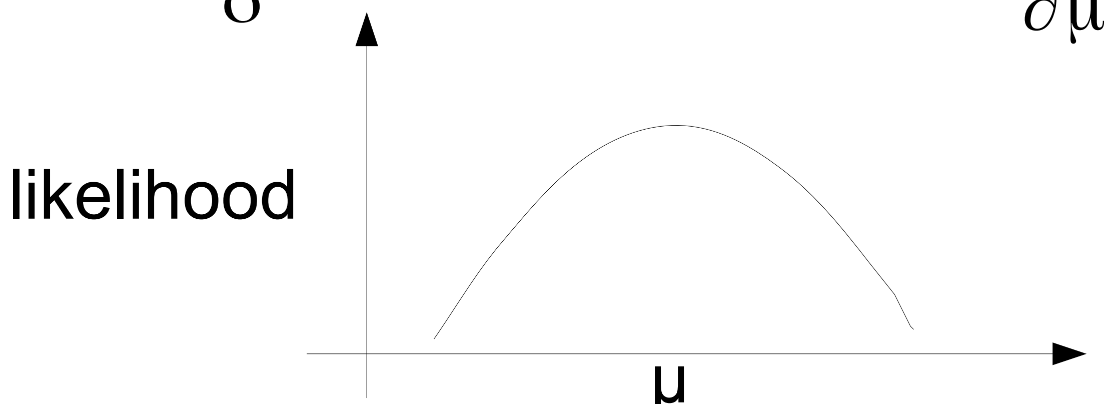
$$=\frac{-n}{2}\log(2\pi)-\frac{n}{2}\log(\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

logL is an upper convex function w.r.t. μ

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

It can be proven by showing that the second derivative is negative.

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



In this case, the parameters to maximize -logL are found uniquely, and the local maximum is equal to the global maximum.

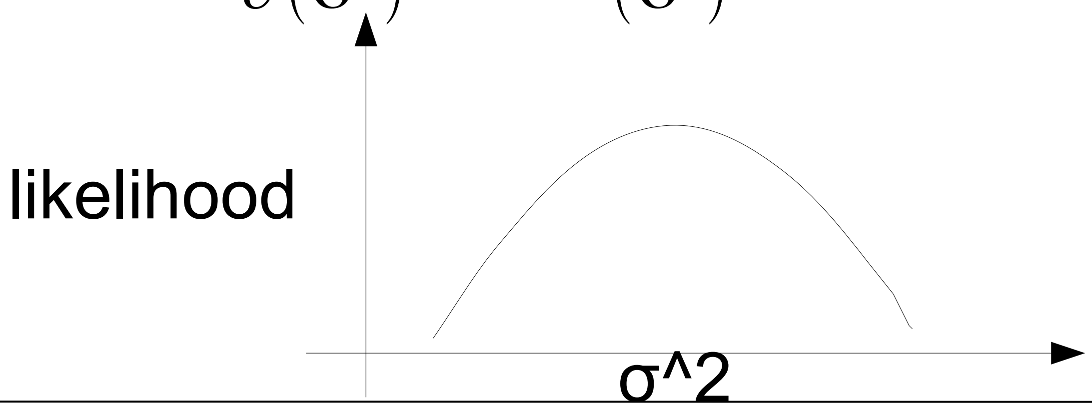
logL is an upper convex function w.r.t. σ^2

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

It can be proven by showing that the second derivative is negative.

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

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



Maximum Likelihood Estimate (MLE)

maximum likelihood estimate: parameters that maximize likelihood

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

$$\leftarrow\rightarrow \hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

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

$$\leftarrow\rightarrow \hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

MLE of mean and variance turned out to be equivalent to data (sample) mean and data (sample) variance, respectively.

Ex. 2: Estimating parameters of 1D Gaussian distribution

- gauss1d_MLE.m(on the next slide) generates data that follow gaussian distribution, then estimates parameters: mean and variance. Complete functions mean_MLE and var_MLE.
- Usage of gauss1d_MLE for generating 10 data points from N(0,1)
  - gauss1d_MLE(10,0,1)
- How does MLE of mean and variance change with respect to the increase in the number of data points?

- gauss1d_MLE.m

```
function [] = gauss1d_MLE(n,mu,sigma)
X = sigma*randn(n,1)+mu;
mu_MLE = mean_MLE(X); sigma_MLE = var_MLE(X);
X = -3:0.5:3;
Y = g1_pdf(X,mu,sigma);
disp(['true:',num2str(Y)]);
Y_MLE = g1_pdf(X,mu_MLE,sigma_MLE);
disp(['estimated:',num2str(Y_MLE)]);
plot(X,Y,'ro-',X,Y_MLE,'gx-');
legend('true','estimated');
endfunction

function [z]=g1_pdf(x,mu,sigma)
z=(2*pi*sigma.^2)^(-1/2)*exp(-(x-mu).^2./(2*sigma.^2));
endfunction

function [my_mean] = mean_MLE(X)

endfunction

function [my_var] = var_MLE(X)

endfunction
```

Ex. 3: MLE of multi-dimensional Gaussian distribution

$$p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left(\frac{-1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

- Describe the mean MLE and variance MLE of the above multi-dimensional Gaussian distribution.
  - Hints:

$$\frac{\partial}{\partial\boldsymbol{\mu}}\boldsymbol{\mu}'\boldsymbol{\Sigma}\boldsymbol{\mu}=2\boldsymbol{\Sigma}\boldsymbol{\mu}$$

$$\frac{\partial}{\partial\boldsymbol{\Sigma}}\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}=-\boldsymbol{\Sigma}^{-1}\boldsymbol{x}'\boldsymbol{x}\boldsymbol{\Sigma}^{-1}$$

$$\frac{\partial}{\partial\boldsymbol{\mu}}\boldsymbol{\mu}'\boldsymbol{\Sigma}\boldsymbol{x}=\boldsymbol{\Sigma}\boldsymbol{x}$$

$$\frac{\partial}{\partial\boldsymbol{\Sigma}}\log(|\boldsymbol{\Sigma}|)=\boldsymbol{\Sigma}^{-1}$$

(Ex .4): Estimating parameters of 2D Gaussian distribution

- Try the same as Ex.2 for 2D data.
- Usage
  - gauss2d_MLE(10,[0 0]',[1 -0.5;-0.5 1])

MLE of 2D Gaussian

- gauss2d_MLE.m

```
function [] = gauss2d_MLE(n,Mu,Sigma)
%usage: gauss2d_MLE(10,[0 0]',[1 -0.5;-0.5 1])
X = randn(n,length(Mu))*Sigma+ones(n,1)*Mu';
Mu_MLE = mean(X)';
Sigma_MLE = cov(X);

X = -3:0.1:3;
Y = -3:0.1:3;
for i=1:length(X)
  for j=1:length(Y)
    Z(i,j) = g2_pdf(X(i),Y(j),Mu,Sigma);
    Z_MLE(i,j) = g2_pdf(X(i),Y(j),Mu_MLE,Sigma_MLE);
  end
end
figure(1);clf;
surface(Z);
title('true distribution')
figure(2);clf;
surface(Z_MLE);
title('estimated distribution')
%print -deps gauss2d_MLE.eps
endfunction
```