

**10417/10617**  
**Intermediate Deep Learning:**  
**Fall2020**

Russ Salakhutdinov

Machine Learning Department  
[rsalakhu@cs.cmu.edu](mailto:rsalakhu@cs.cmu.edu)

Lecture 3

# Bernoulli Distribution

- Consider a single binary random variable  $x \in \{0, 1\}$ . For example,  $x$  can describe the outcome of flipping a coin:

Coin flipping: heads = 1, tails = 0.

- The probability of  $x=1$  will be denoted by the parameter  $\mu$ , so that:

$$p(x = 1|\mu) = \mu \quad 0 \leq \mu \leq 1.$$

- The probability distribution, known as Bernoulli distribution, can be written as:

$$\text{Bern}(x|\mu) = \mu^x (1 - \mu)^{1-x}$$

$$\mathbb{E}[x] = \mu$$

$$\text{var}[x] = \mu(1 - \mu)$$


# Parameter Estimation

- Suppose we observed a dataset  $\mathcal{D} = \{x_1, \dots, x_N\}$
- We can construct the likelihood function, which is a function of  $\mu$ .

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

- Equivalently, we can maximize the log of the likelihood function:

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^N \ln p(x_n|\mu) = \sum_{n=1}^N \{x_n \ln \mu + (1 - x_n) \ln(1 - \mu)\}$$

- Note that the likelihood function depends on the N observations  $x_n$  only through the sum  $\sum_n x_n$   Sufficient Statistic

# Parameter Estimation

- Suppose we observed a dataset  $\mathcal{D} = \{x_1, \dots, x_N\}$

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^N \ln p(x_n|\mu) = \sum_{n=1}^N \{x_n \ln \mu + (1 - x_n) \ln(1 - \mu)\}$$

- Setting the derivative of the log-likelihood function w.r.t  $\mu$  to zero, we obtain:

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

where  $m$  is the number of heads.

# Multinomial Variables

- Consider a random variable that can take on one of  $K$  possible mutually exclusive states (e.g. roll of a dice).
- We will use so-called 1-of- $K$  encoding scheme.
- If a random variable can take on  $K=6$  states, and a particular observation of the variable corresponds to the state  $x_3=1$ , then  $\mathbf{x}$  will be resented as:

1-of- $K$  coding scheme:  $\mathbf{x} = (0, 0, 1, 0, 0, 0)^T$

- If we denote the probability of  $x_k=1$  by the parameter  $\mu_k$ , then the distribution over  $\mathbf{x}$  is defined as:

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^K \mu_k^{x_k} \quad \forall k : \mu_k \geq 0 \quad \text{and} \quad \sum_{k=1}^K \mu_k = 1$$

# Multinomial Variables

- Multinomial distribution can be viewed as a generalization of Bernoulli distribution to more than two outcomes.

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^K \mu_k^{x_k}$$

- It is easy to see that the distribution is normalized:

$$\sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu}) = \sum_{k=1}^K \mu_k = 1$$

and

$$\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \dots, \mu_K)^T = \boldsymbol{\mu}$$

# Maximum Likelihood Estimation

- Suppose we observed a dataset  $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$
- We can construct the likelihood function, which is a function of  $\mu$ .

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^N \prod_{k=1}^K \mu_k^{x_{nk}} = \prod_{k=1}^K \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^K \mu_k^{m_k}$$

- Note that the likelihood function depends on the N data points only though the following K quantities:

$$m_k = \sum_n x_{nk}, \quad k = 1, \dots, K.$$

which represents the number of observations of  $x_k=1$ .

- These are called the sufficient statistics for this distribution.

# Maximum Likelihood Estimation

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^N \prod_{k=1}^K \mu_k^{x_{nk}} = \prod_{k=1}^K \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^K \mu_k^{m_k}$$

- To find a maximum likelihood solution for  $\mu$ , we need to maximize the log-likelihood taking into account the constraint that  $\sum_k \mu_k = 1$
- Forming the Lagrangian:

$$\sum_{k=1}^K m_k \ln \mu_k + \lambda \left( \sum_{k=1}^K \mu_k - 1 \right)$$

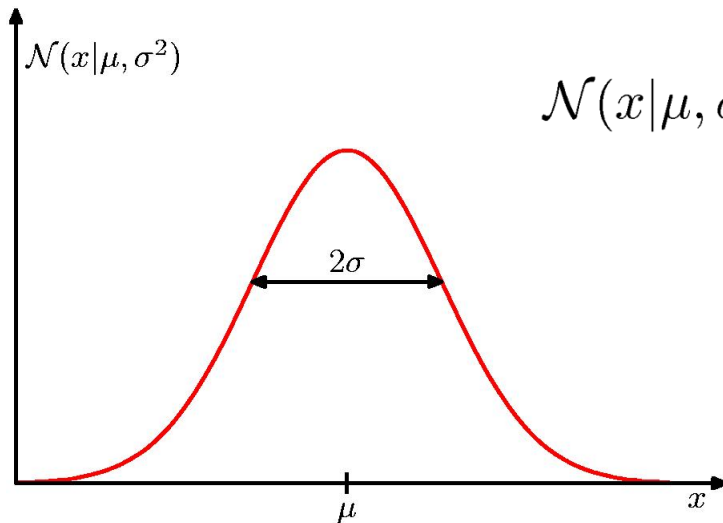
$$\mu_k = -m_k/\lambda \quad \mu_k^{\text{ML}} = \frac{m_k}{N} \quad \lambda = -N$$

which is the fraction of observations for which  $x_k=1$ .



# Gaussian Univariate Distribution

- In the case of a single variable  $x$ , the Gaussian distribution takes form:



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

which is governed by two parameters:

- $\mu$  (mean)
- $\sigma^2$  (variance)

- The Gaussian distribution satisfies:

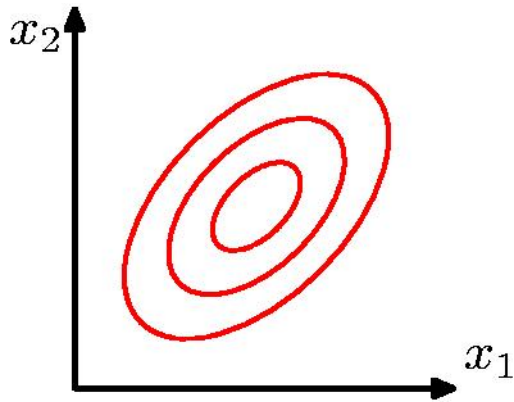
$$\mathcal{N}(x|\mu, \sigma^2) > 0$$

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

# Multivariate Gaussian Distribution

- For a D-dimensional vector  $\mathbf{x}$ , the Gaussian distribution takes form:

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



which is governed by two parameters:

- $\boldsymbol{\mu}$  is a D-dimensional mean vector.
- $\boldsymbol{\Sigma}$  is a D by D covariance matrix.

and  $|\boldsymbol{\Sigma}|$  denotes the determinant of  $\boldsymbol{\Sigma}$ .

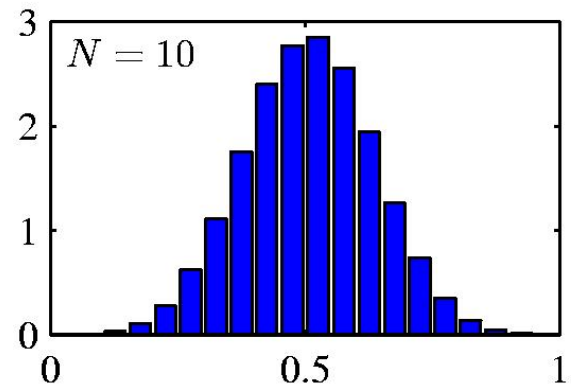
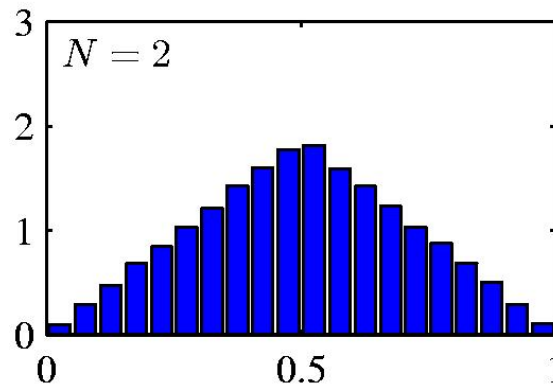
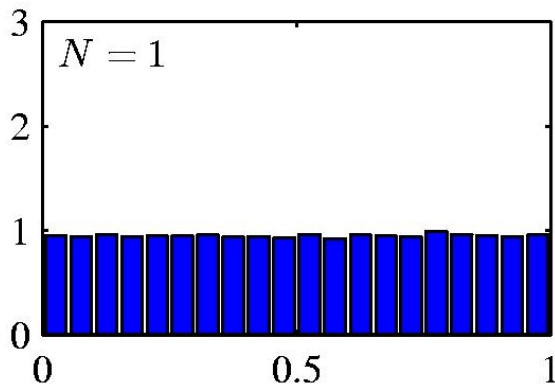
- Note that the covariance matrix is a symmetric positive definite matrix.

# Central Limit Theorem

- The distribution of the sum of  $N$  i.i.d. random variables becomes increasingly Gaussian as  $N$  grows.
- Consider  $N$  variables, each of which has a uniform distribution over the interval  $[0,1]$ .
- Let us look at the distribution over the mean:

$$\frac{x_1 + x_2 + \dots + x_N}{N}.$$

- As  $N$  increases, the distribution tends towards a Gaussian distribution.



# Moments of the Gaussian Distribution

- The expectation of  $\mathbf{x}$  under the Gaussian distribution:

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right\} \mathbf{x} \, d\mathbf{x} \\ &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int \underbrace{\exp \left\{ -\frac{1}{2}\mathbf{z}^T \boldsymbol{\Sigma}^{-1}\mathbf{z} \right\}}_{\text{symmetric}} (\mathbf{z} + \boldsymbol{\mu}) \, d\mathbf{z}\end{aligned}$$

The term in  $\mathbf{z}$  in the factor  $(\mathbf{z} + \boldsymbol{\mu})$   
will vanish by symmetry.

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$$

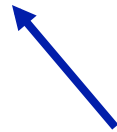
# Moments of the Gaussian Distribution

- The second order moments of the Gaussian distribution:

$$\mathbb{E}[\mathbf{x}\mathbf{x}^T] = \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma}$$

- The covariance is given by:

$$\text{cov}[\mathbf{x}] = \mathbb{E} [(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T] = \boldsymbol{\Sigma}$$

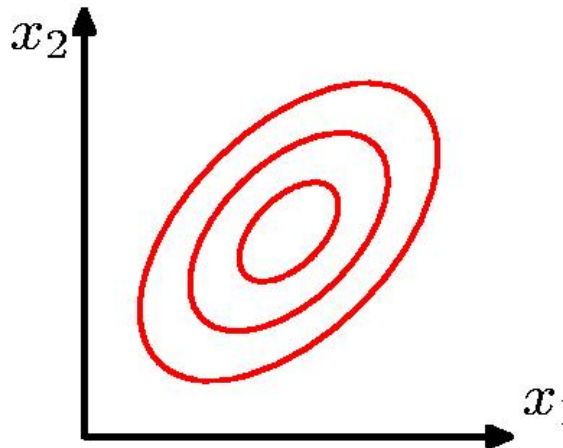


$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$$

- Because the parameter matrix  $\boldsymbol{\Sigma}$  governs the covariance of  $\mathbf{x}$  under the Gaussian distribution, it is called the covariance matrix.

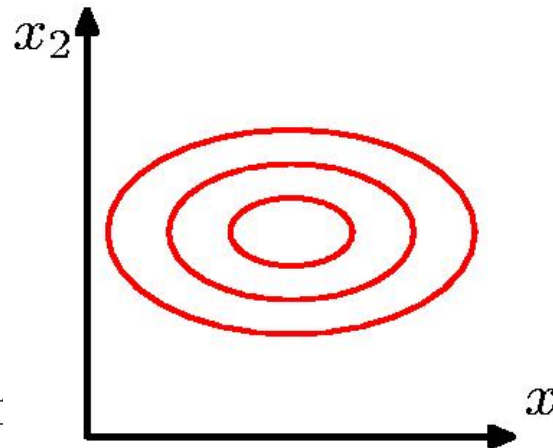
# Moments of the Gaussian Distribution

- Contours of constant probability density:



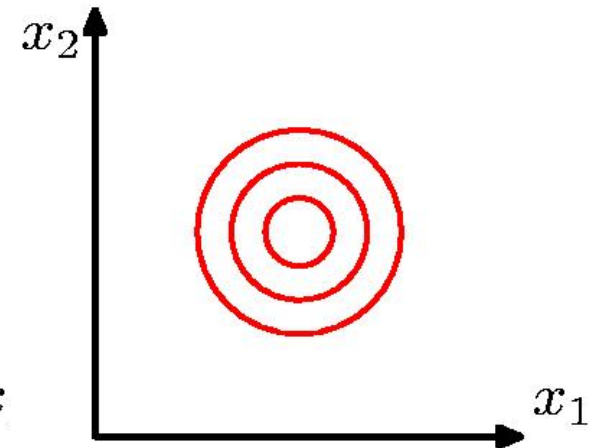
(a)

Covariance matrix is of general form.



(b)

Diagonal, axis-aligned covariance matrix.



(c)

Spherical (proportional to identity) covariance matrix.

# Partitioned Gaussian Distribution

- Consider a D-dimensional Gaussian distribution:  $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Let us partition  $\mathbf{x}$  into two disjoint subsets  $\mathbf{x}_a$  and  $\mathbf{x}_b$ :

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

- In many situations, it will be more convenient to work with the precision matrix (inverse of the covariance matrix):

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1} \quad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

- Note that  $\boldsymbol{\Lambda}_{aa}$  is not given by the inverse of  $\boldsymbol{\Sigma}_{aa}$ .

# Conditional Distribution

- It turns out that the conditional distribution is also a Gaussian distribution:

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$$

Covariance does not  
depend on  $\mathbf{x}_b$ .



$$\begin{aligned}\boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b} \{ \boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \} \\ &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\ &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)\end{aligned}$$

Linear function  
of  $\mathbf{x}_b$ .





# Marginal Distribution

- It turns out that the marginal distribution is also a Gaussian distribution:

$$\begin{aligned} p(\mathbf{x}_a) &= \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_b \\ &= \mathcal{N}(\mathbf{x}_a | \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}) \end{aligned}$$

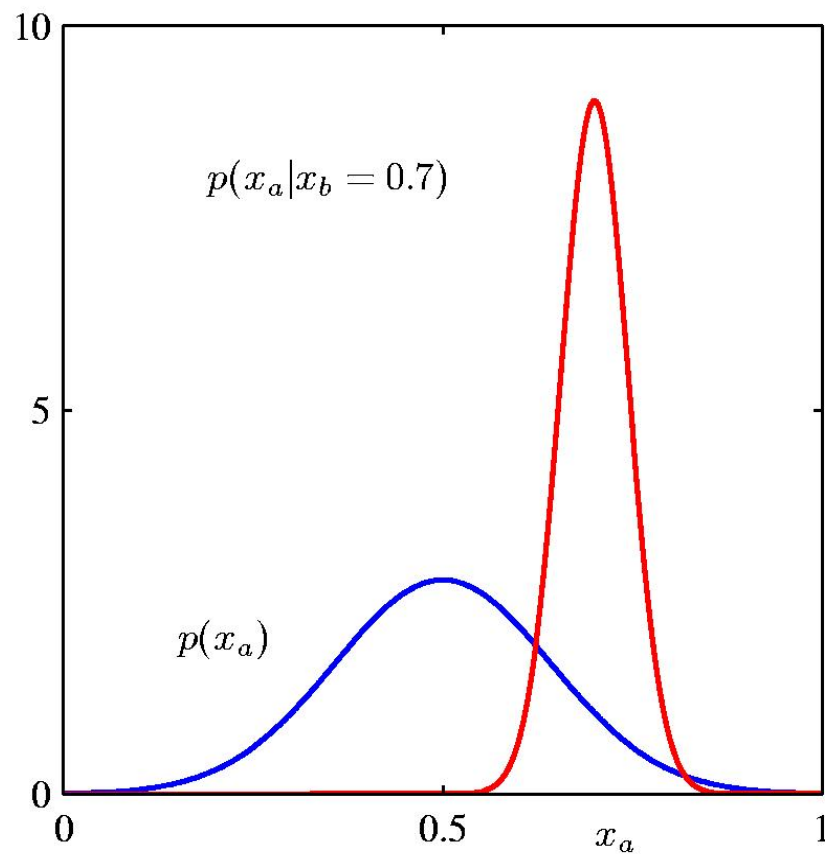
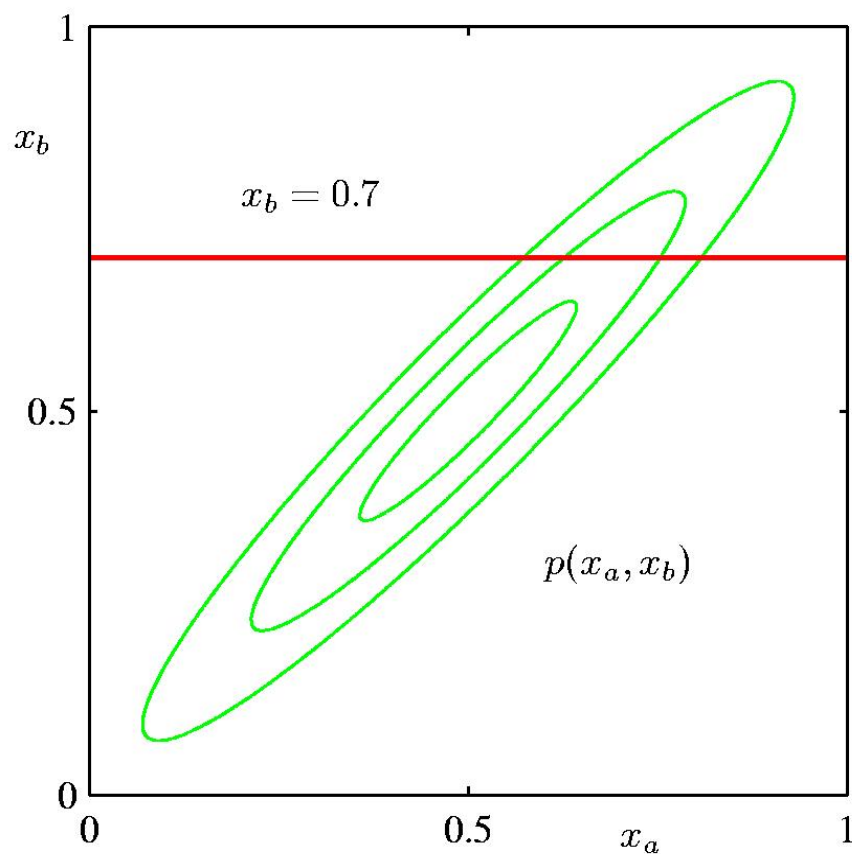
- For a marginal distribution, the mean and covariance are most simply expressed in terms of partitioned covariance matrix.

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}$$

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

# Conditional and Marginal Distributions



# Maximum Likelihood Estimation

- Suppose we observed i.i.d data  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ .
- We can construct the log-likelihood function, which is a function of  $\mu$  and  $\Sigma$ :

$$\ln p(\mathbf{X}|\mu, \Sigma) = -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln |\Sigma| - \frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \mu)^T \Sigma^{-1} (\mathbf{x}_n - \mu)$$

- Note that the likelihood function depends on the N data points only though the following sums:

## Sufficient Statistics

$$\sum_{n=1}^N \mathbf{x}_n$$

$$\sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^T$$

# Maximum Likelihood Estimation

- To find a maximum likelihood estimate of the mean, we set the derivative of the log-likelihood function to zero:

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) = 0$$

and solve to obtain:

$$\boldsymbol{\mu}_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n.$$

- Similarly, we can find the ML estimate of  $\boldsymbol{\Sigma}$ :

$$\boldsymbol{\Sigma}_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^{\text{T}}.$$

# Maximum Likelihood Estimation

- Evaluating the expectation of the ML estimates under the true distribution, we obtain:

$$\begin{aligned}\mathbb{E}[\boldsymbol{\mu}_{\text{ML}}] &= \boldsymbol{\mu} && \swarrow \text{Unbiased estimate} \\ \mathbb{E}[\boldsymbol{\Sigma}_{\text{ML}}] &= \frac{N-1}{N} \boldsymbol{\Sigma}. && \swarrow \text{Biased estimate}\end{aligned}$$

- Note that the maximum likelihood estimate of  $\boldsymbol{\Sigma}$  is biased.
- We can correct the bias by defining a different estimator:

$$\tilde{\boldsymbol{\Sigma}} = \frac{1}{N-1} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^{\text{T}}.$$

# Student's t-Distribution

- Consider Student's t-Distribution

$$\begin{aligned} p(x|\mu, a, b) &= \int_0^\infty \mathcal{N}(x|\mu, \tau^{-1}) \text{Gam}(\tau|a, b) d\tau \\ &= \int_0^\infty \mathcal{N}(x|\mu, (\eta\lambda)^{-1}) \text{Gam}(\eta|\nu/2, \nu/2) d\eta \\ &= \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \left( \frac{\lambda}{\pi\nu} \right)^{1/2} \left[ 1 + \frac{\lambda(x - \mu)^2}{\nu} \right]^{-\nu/2 - 1/2} \\ &= \text{St}(x|\mu, \lambda, \nu) \end{aligned}$$

Infinite mixture  
of Gaussians

where

$$\lambda = a/b$$

$$\eta = \tau b/a$$

$$\nu = 2a.$$

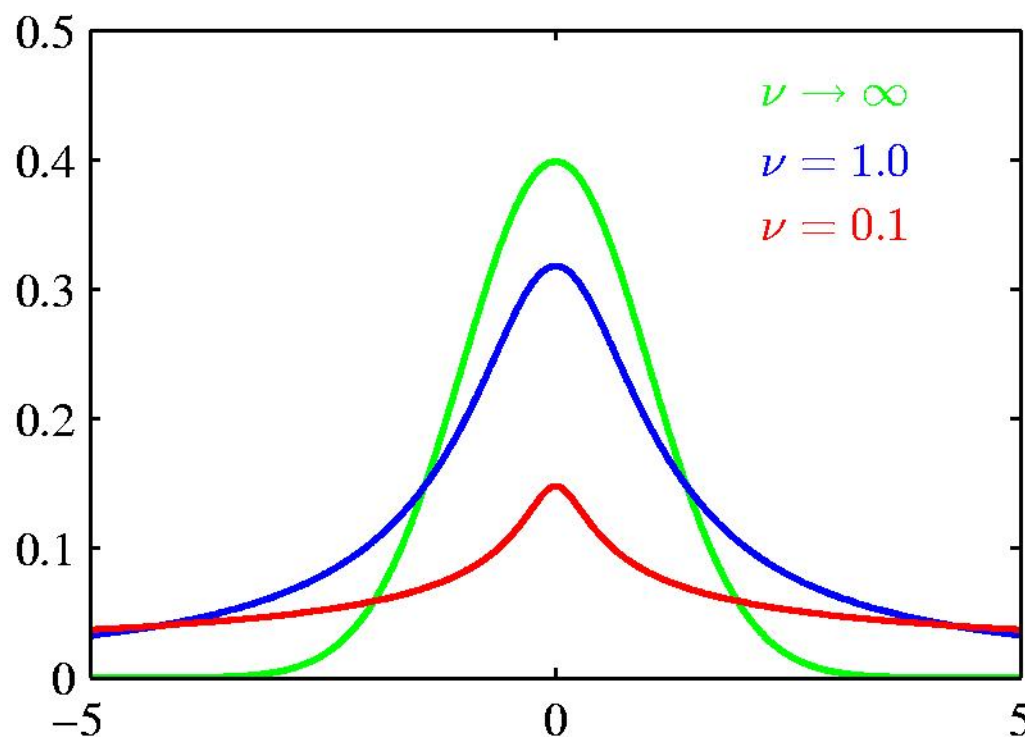
Sometimes called  
the precision  
parameter.

Degrees of freedom

# Student's t-Distribution

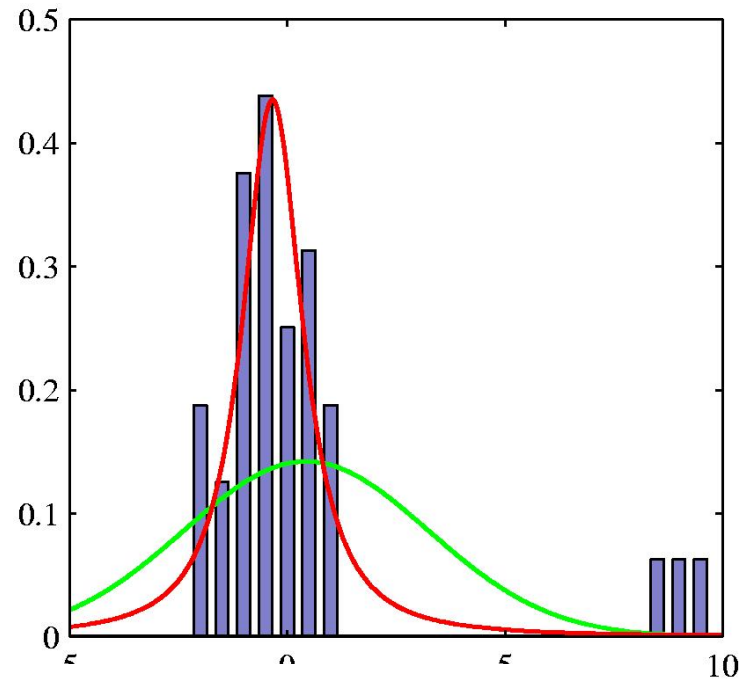
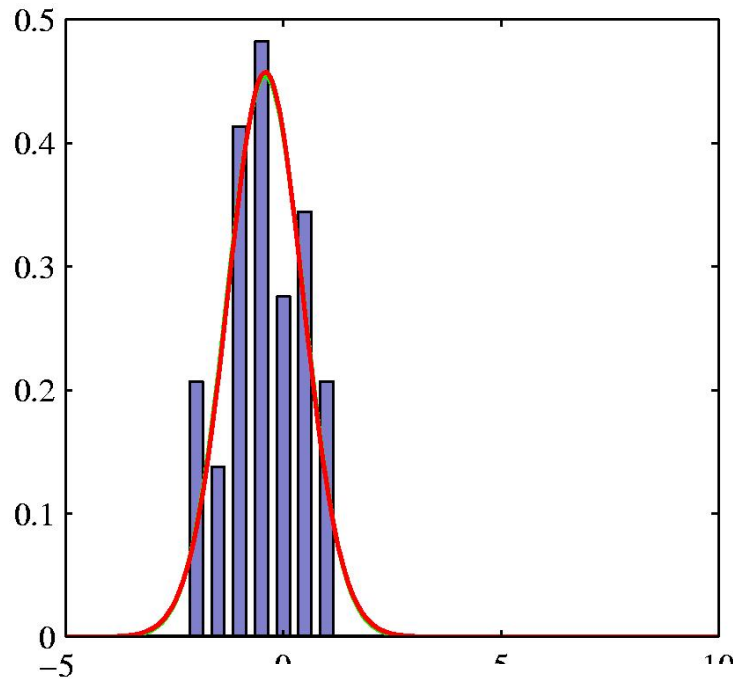
- Setting  $\nu = 1$  recovers Cauchy distribution
- The limit  $\nu \rightarrow \infty$  corresponds to a Gaussian distribution.

	$\nu = 1$	$\nu \rightarrow \infty$
$\text{St}(x \mu, \lambda, \nu)$	Cauchy	$\mathcal{N}(x \mu, \lambda^{-1})$



# Student's t-Distribution

- Robustness to outliers: Gaussian vs. t-Distribution.





# Student's t-Distribution

- The multivariate extension of the t-Distribution:

$$\begin{aligned}\text{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) &= \int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) \text{Gam}(\eta|\nu/2, \nu/2) d\eta \\ &= \frac{\Gamma(D/2 + \nu/2)}{\Gamma(\nu/2)} \frac{|\boldsymbol{\Lambda}|^{1/2}}{(\pi\nu)^{D/2}} \left[1 + \frac{\Delta^2}{\nu}\right]^{-D/2-\nu/2}\end{aligned}$$

where  $\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Lambda} (\mathbf{x} - \boldsymbol{\mu})$

- Properties:

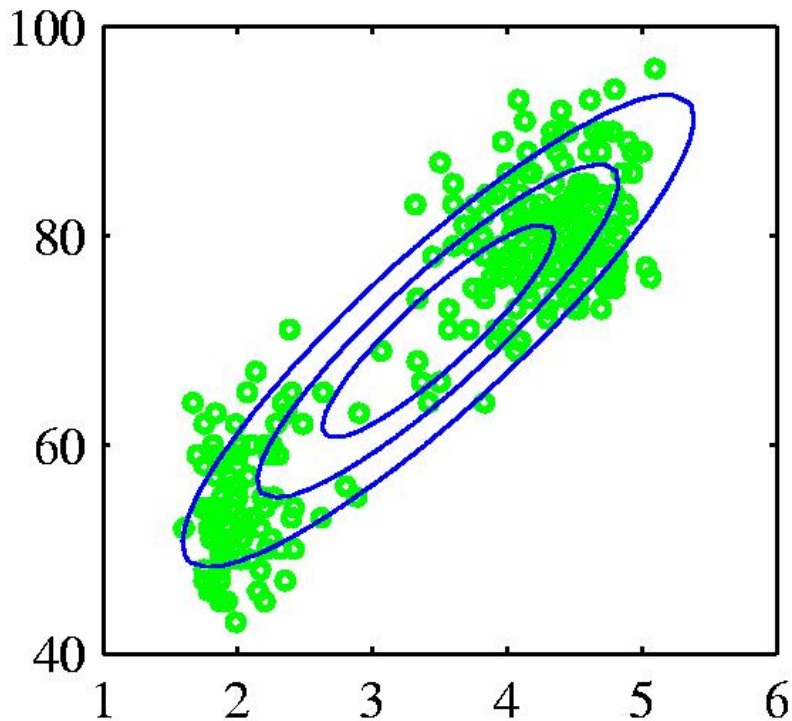
$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}, \quad \text{if } \nu > 1$$

$$\text{cov}[\mathbf{x}] = \frac{\nu}{(\nu - 2)} \boldsymbol{\Lambda}^{-1}, \quad \text{if } \nu > 2$$

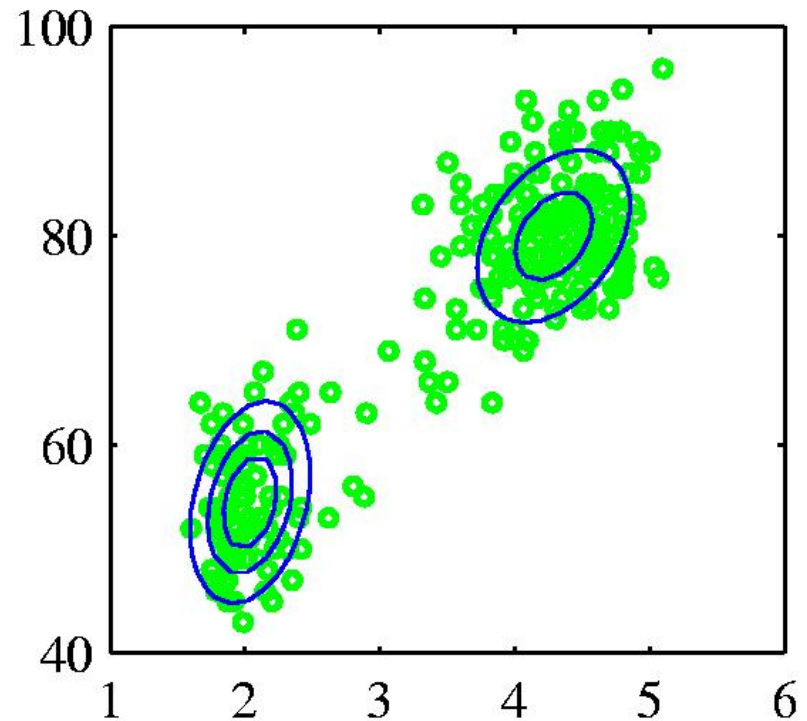
$$\text{mode}[\mathbf{x}] = \boldsymbol{\mu}$$

# Mixture of Gaussians

- When modeling real-world data, Gaussian assumption may not be appropriate.
- Consider the following example: Old Faithful Dataset



Single Gaussian



Mixture of two  
Gaussians

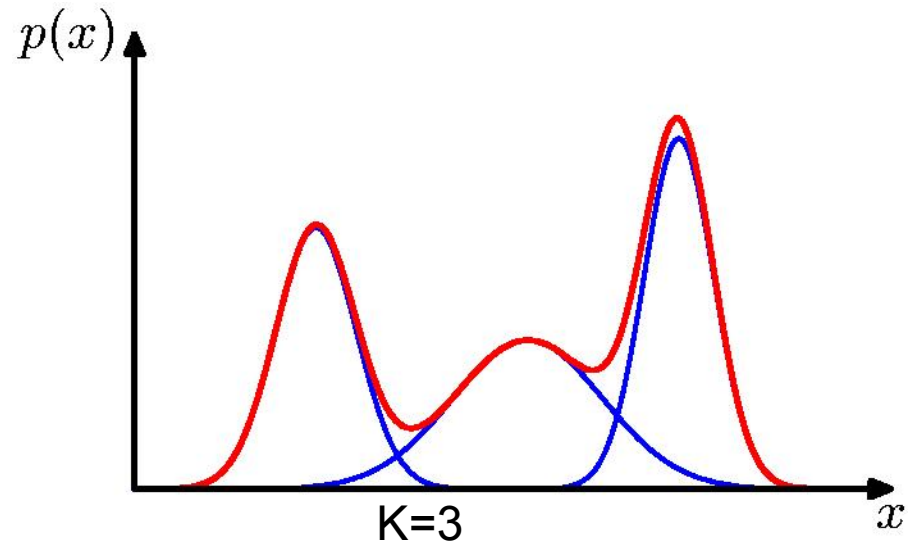
# Mixture of Gaussians

- We can combine simple models into a complex model by defining a superposition of  $K$  Gaussian densities of the form:

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \underbrace{\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{\text{Component}}$$

Mixing coefficient

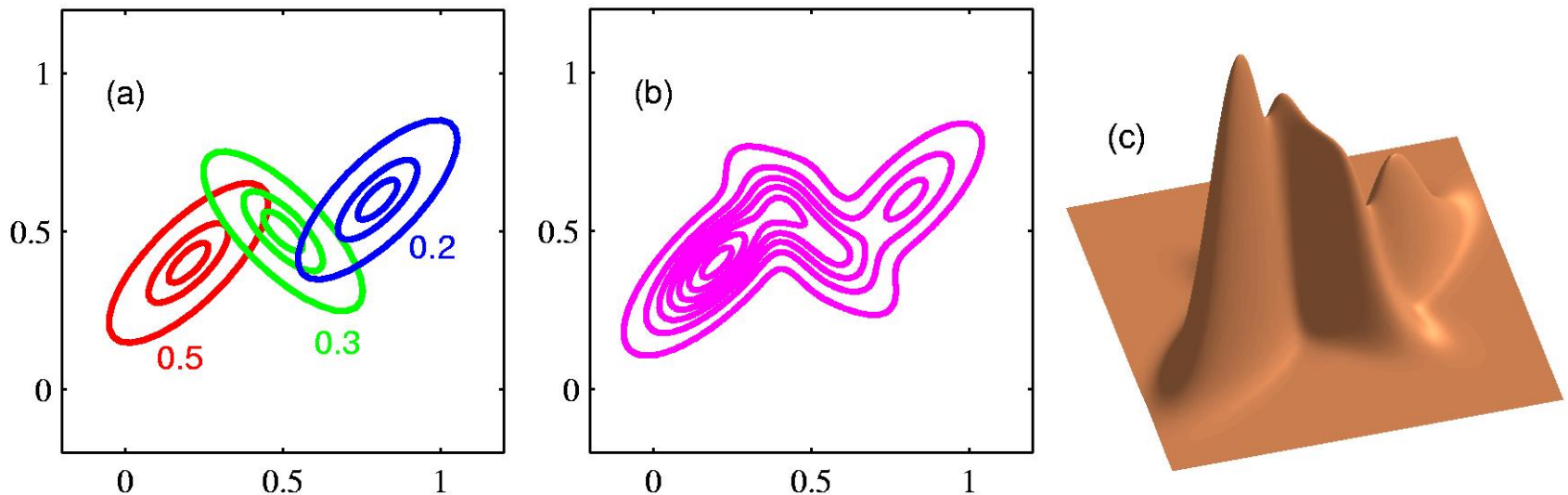
$$\forall k : \pi_k \geq 0 \quad \sum_{k=1}^K \pi_k = 1$$



- Note that each Gaussian component has its own mean  $\mu_k$  and covariance  $\Sigma_k$ . The parameters  $\pi_k$  are called mixing coefficients.
- Note generally, mixture models can comprise linear combinations of other distributions.

# Mixture of Gaussians

- Illustration of a mixture of 3 Gaussians in a 2-dimensional space:



(a) Contours of constant density of each of the mixture components, along with the mixing coefficients

(b) Contours of marginal probability density  $p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

(c) A surface plot of the distribution  $p(\mathbf{x})$ .

# Maximum Likelihood Estimation

- Given a dataset  $D$ , we can determine model parameters  $\mu_k, \Sigma_k, \pi_k$  by maximizing the log-likelihood function:

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Log of a sum: no closed form solution

- Solution:** use standard, iterative, numeric optimization methods or the Expectation Maximization algorithm.

# The Exponential Family

- The exponential family of distributions over  $\mathbf{x}$  is defined to be a set of distributions of the form:

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \}$$

where

- $\boldsymbol{\eta}$  is the vector of natural parameters
  - $\mathbf{u}(\mathbf{x})$  is the vector of sufficient statistics
- The function  $g(\boldsymbol{\eta})$  can be interpreted as the coefficient that ensures that the distribution  $p(\mathbf{x}|\boldsymbol{\eta})$  is normalized:

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \} d\mathbf{x} = 1$$

# ML for the Exponential Family

- Remember the Exponential Family:

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \}$$

- From the definition of the normalizer  $g(\boldsymbol{\eta})$ :

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \} d\mathbf{x} = 1$$

- We can take a derivative w.r.t  $\boldsymbol{\eta}$ :

$$\nabla g(\boldsymbol{\eta}) \underbrace{\int h(\mathbf{x}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \} d\mathbf{x}}_{1/g(\boldsymbol{\eta})} + g(\boldsymbol{\eta}) \underbrace{\int h(\mathbf{x}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \} \mathbf{u}(\mathbf{x}) d\mathbf{x}}_{\mathbb{E}[\mathbf{u}(\mathbf{x})]} = 0$$

- Thus

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

# ML for the Exponential Family

- Remember the Exponential Family:

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \}$$

- We can take a derivative w.r.t  $\eta$ :

$$\nabla g(\boldsymbol{\eta}) \underbrace{\int h(\mathbf{x}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \} \, d\mathbf{x}}_{1/g(\boldsymbol{\eta})} + g(\boldsymbol{\eta}) \underbrace{\int h(\mathbf{x}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \} \mathbf{u}(\mathbf{x}) \, d\mathbf{x}}_{\mathbb{E}[\mathbf{u}(\mathbf{x})]} = 0$$

- Thus

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

- Note that the covariance of  $\mathbf{u}(\mathbf{x})$  can be expressed in terms of the second derivative of  $g(\boldsymbol{\eta})$ , and similarly for the higher moments.



# ML for the Exponential Family

- Suppose we observed i.i.d data  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ .
- We can construct the log-likelihood function, which is a function of the natural parameter  $\eta$ .

$$p(\mathbf{x}|\eta) = h(\mathbf{x})g(\eta) \exp \{ \eta^T \mathbf{u}(\mathbf{x}) \}$$

$$p(\mathbf{X}|\eta) = \left( \prod_{n=1}^N h(\mathbf{x}_n) \right) g(\eta)^N \exp \left\{ \eta^T \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n) \right\}.$$

- Therefore we have

$$-\nabla \ln g(\eta_{\text{ML}}) = \underbrace{\frac{1}{N} \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n)}$$

Sufficient Statistic