# 10417/10617
# Intermediate Deep Learning: Fall2019

Russ Salakhutdinov

Machine Learning Department
rsalakhu@cs.cmu.edu

https://deeplearning-cmu-10417.github.io/

## 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:

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

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

$$\begin{aligned} \mathrm{Bern}(x|\mu) &= \mu^x (1-\mu)^{1-x} \\ \mathbb{E}[x] &= \mu \\ \mathrm{var}[x] &= \mu(1-\mu) \end{aligned}$$

# Parameter Estimation

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

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

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

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

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 **x** will be resented as:

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

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

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

# 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

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

# Maximum Likelihood Estimation

- Suppose we observed a dataset $\mathcal{D} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$
- 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:

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

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:

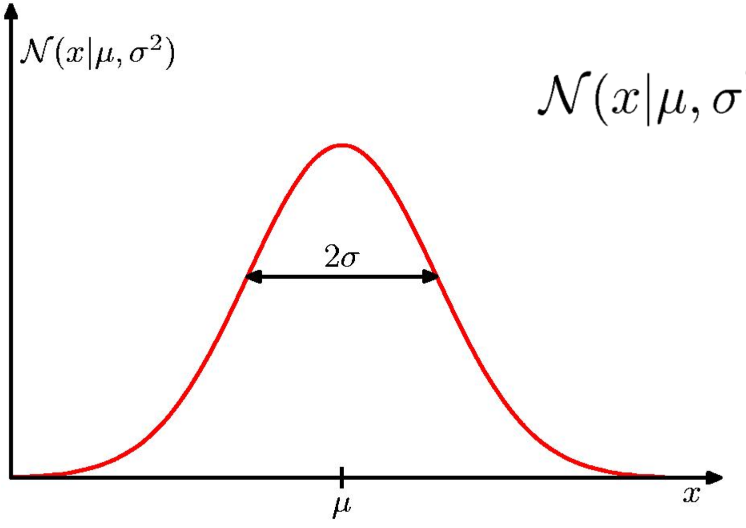
$$\sum_{k=1}^{K} m_k \ln \mu_k + \lambda \left( \sum_{k=1}^{K} \mu_k - 1 \right)$$

$$\mu_k = -m_k/\lambda \qquad \mu_k^{\mathrm{ML}} = \frac{m_k}{N} \qquad \lambda = -N$$

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:

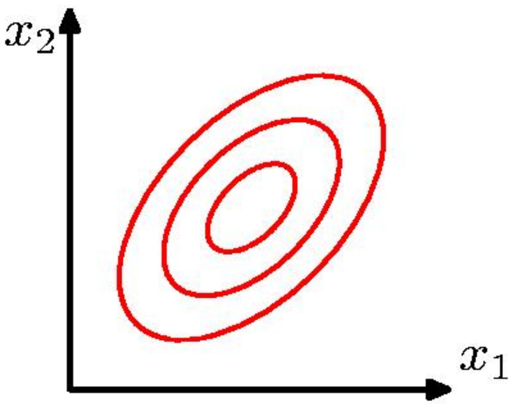
$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

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

# Multivariate Gaussian Distribution

- For a D-dimensional vector **x**, the Gaussian distribution takes form:

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




which is governed by two parameters:

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

and $|\Sigma|$ denotes the determinant of $\Sigma$.

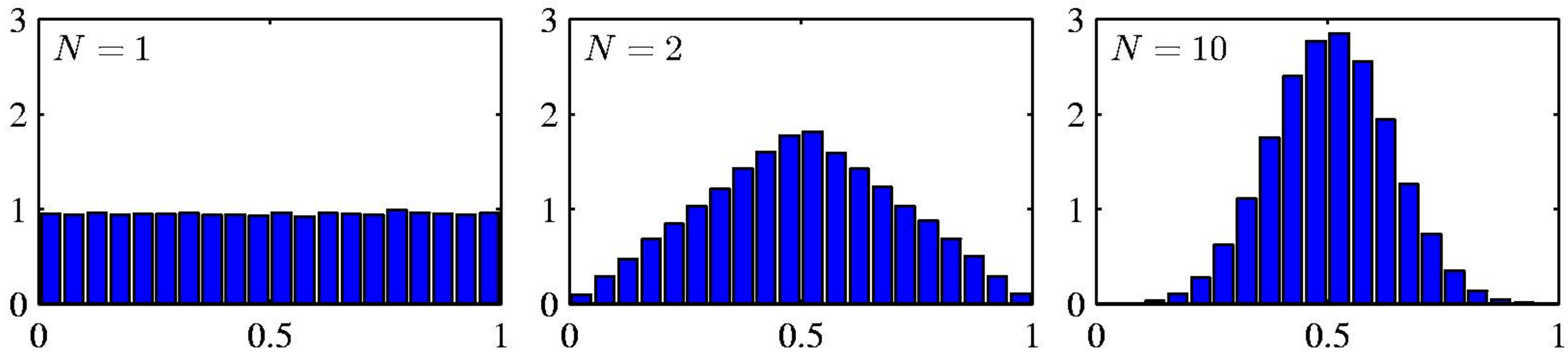
- Note that the covariance matrix is a symmetric positive definite matrix.

# 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:

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

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



# Moments of the Gaussian Distribution

- The expectation of **x** under the Gaussian distribution:

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

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

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

# Moments of the Gaussian Distribution

- The second order moments of the Gaussian distribution:

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

- The covariance is given by:

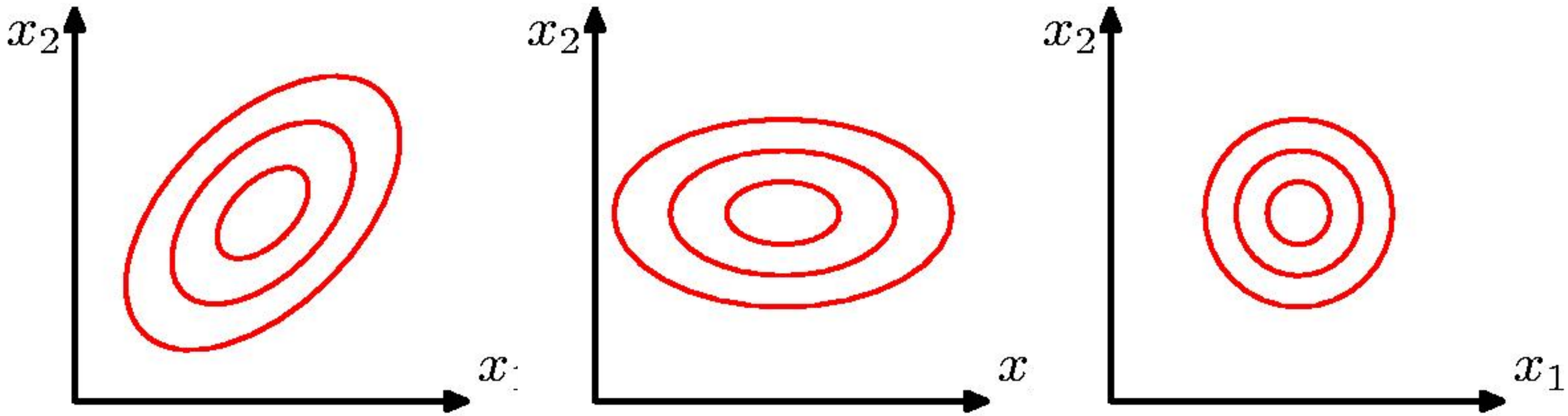
$$\mathrm{cov}[\mathbf{x}] = \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^{\mathrm{T}}\right] = \boldsymbol{\Sigma}$$

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

- Because the parameter matrix $\Sigma$ governs the covariance of 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 **x** into two disjoint subsets $x_a$ and $x_b$:

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

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

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

- Note that $\Lambda_{aa}$ is not given by the inverse of $\Sigma_{aa}$.

# 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 $x_b$.

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

Linear function of $x_b$.

# Marginal Distribution

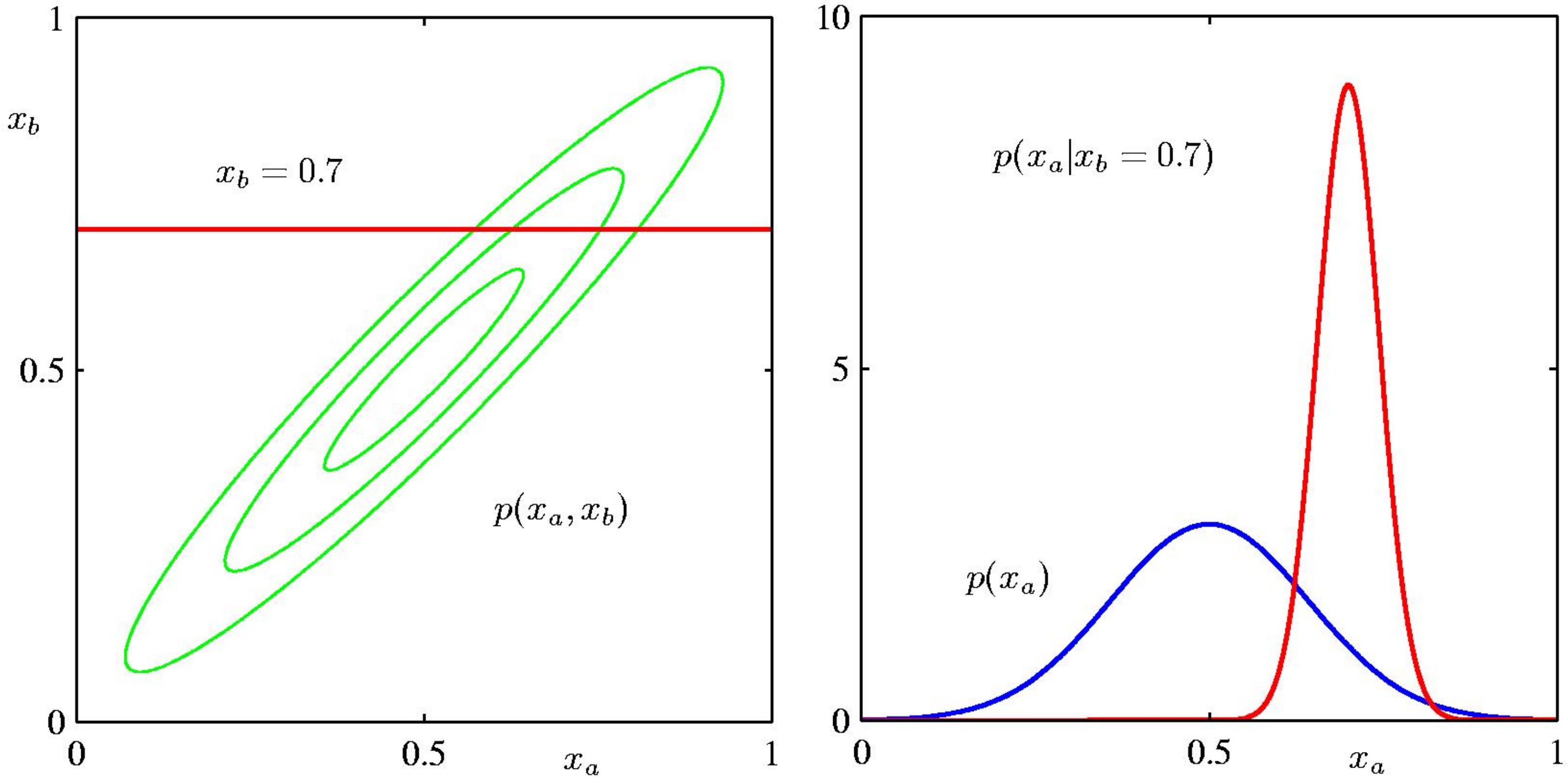
- It turns out that the marginal distribution is also a Gaussian distribution:

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

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

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

# Conditional and Marginal Distributions



# Maximum Likelihood Estimation

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

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

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

**Sufficient Statistics**

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

# 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:

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

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

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

# Maximum Likelihood Estimation

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

Unbiased estimate

$$\mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}] = \boldsymbol{\mu}$$

$$\mathbb{E}[\boldsymbol{\Sigma}_{\mathrm{ML}}] = \frac{N-1}{N}\boldsymbol{\Sigma}.$$

Biased estimate

- Note that the maximum likelihood estimate of $\Sigma$ is biased.

- We can correct the bias by defining a different estimator:

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

# Student's t-Distribution

- Consider Student's t-Distribution

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

Infinite mixture of Gaussians

where

$$\lambda = a/b \qquad \eta = \tau b/a \qquad \nu = 2a.$$

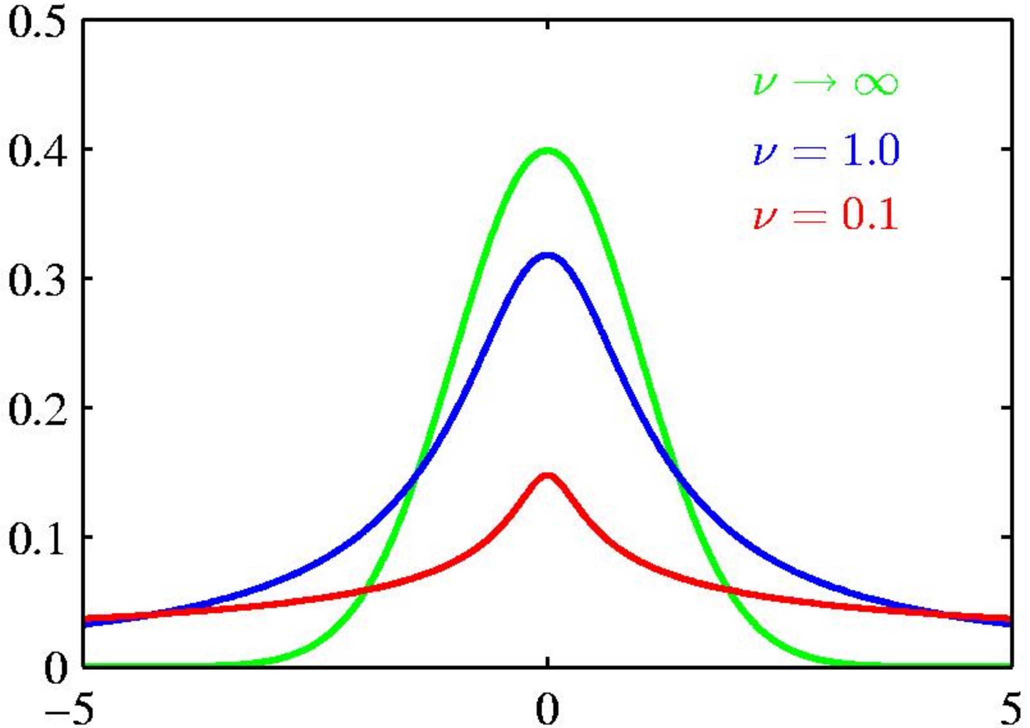
Sometimes called the precision parameter.

Degrees of freedom

# Student's t-Distribution

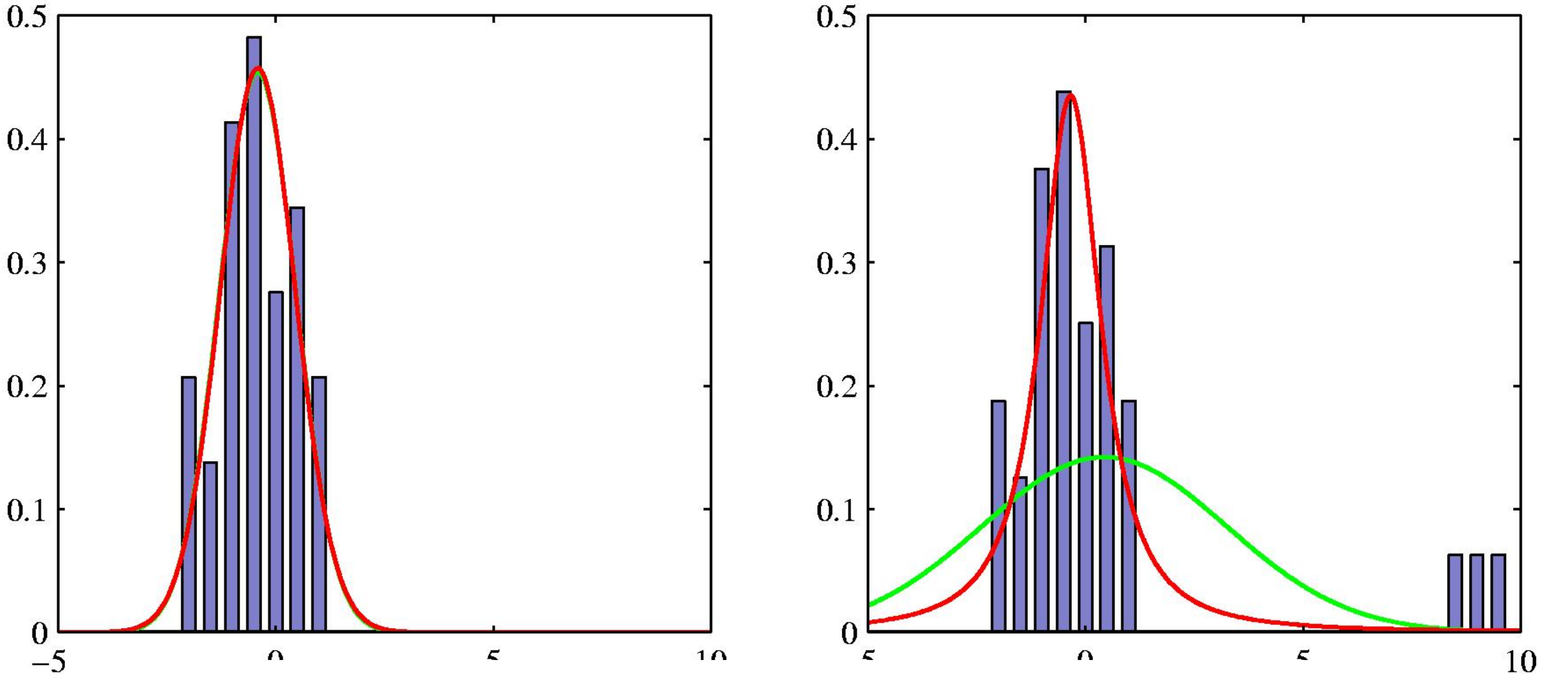
- Setting $\nu$ = 1 recovers Cauchy distribution
- The limit $\nu \to \infty$ corresponds to a Gaussian distribution.

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



# Student's t-Distribution

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



# Student's t-Distribution

- The multivariate extension of the t-Distribution:

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

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

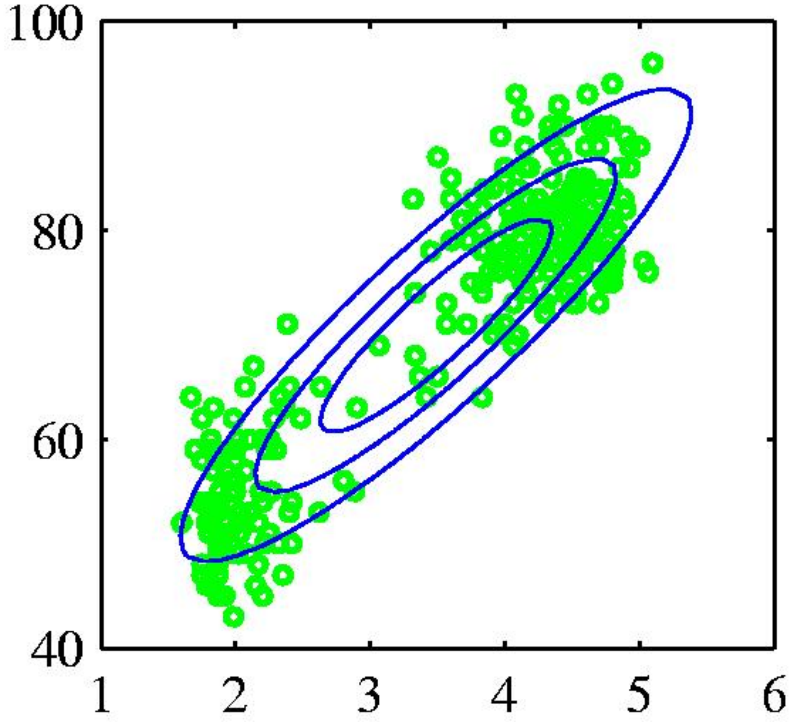
- Properties:

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \boldsymbol{\mu}, && \text{if } \nu > 1 \\ \mathrm{cov}[\mathbf{x}] &= \frac{\nu}{(\nu-2)}\boldsymbol{\Lambda}^{-1}, && \text{if } \nu > 2 \\ \mathrm{mode}[\mathbf{x}] &= \boldsymbol{\mu}\end{aligned}$$

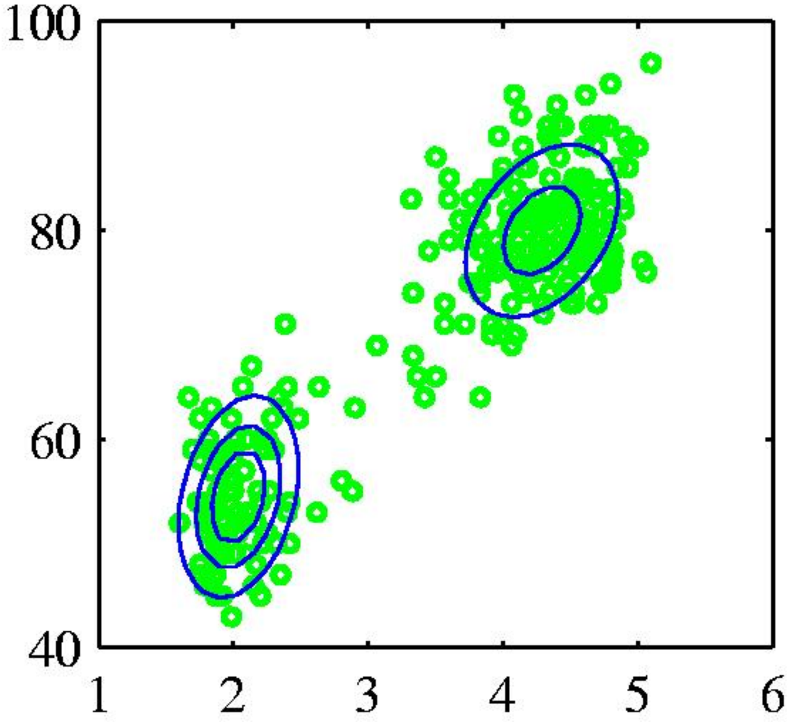
# Mixture of Gaussians

- 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

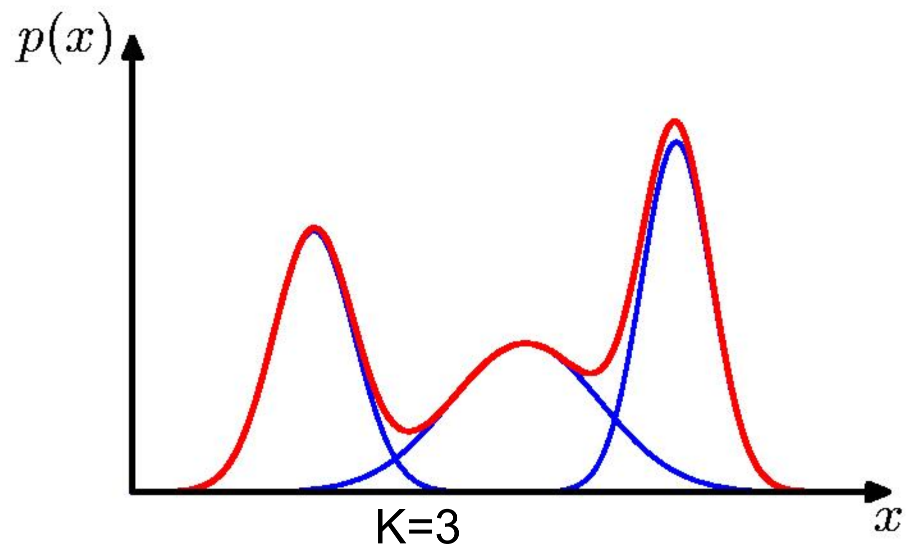
- We can combine simple models into a complex model by defining a superposition of K Gaussian densities of the form:

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

Component

Mixing coefficient

$$\forall k : \pi_k \geqslant 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$

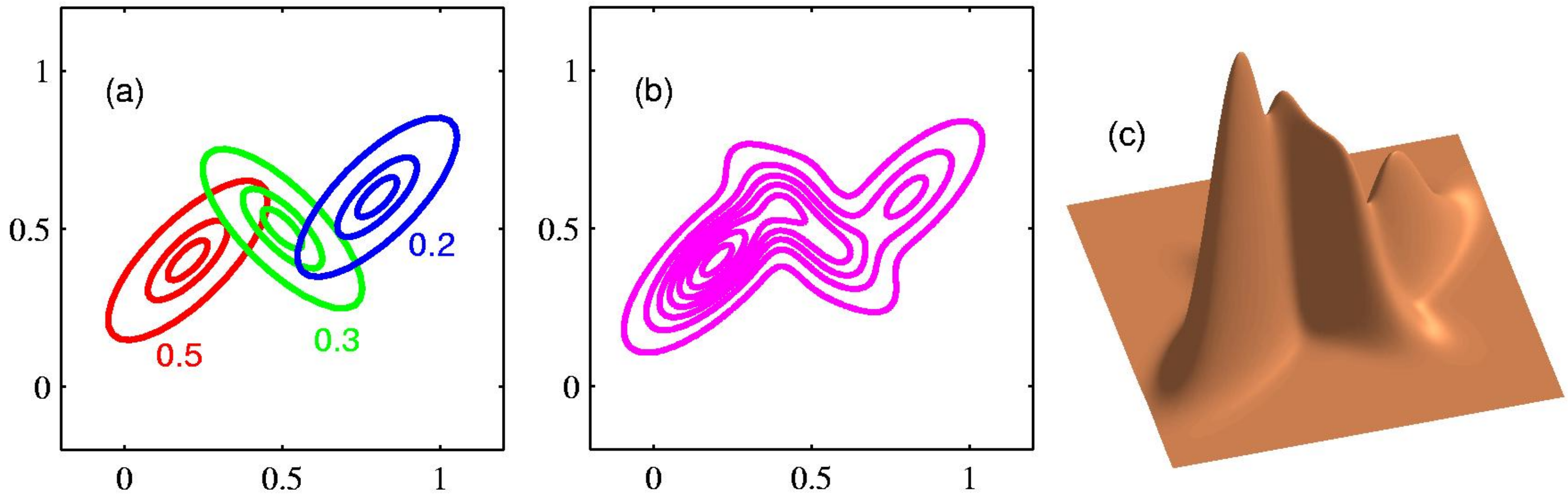



- Note that each Gaussian component has its own mean $\mu_k$ and covariance $\Sigma_k$. The parameters $\pi_k$ are called mixing coefficients.

- Mote 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(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 **x** is defined to be a set of distributions of the form:

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

where

- $\eta$ is the vector of natural parameters
- **u**(**x**) is the vector of sufficient statistics

- The function g($\eta$) can be interpreted as the coefficient that ensures that the distribution p(**x**|$\eta$) is normalized:

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

# ML for the Exponential Family

- Remember the Exponential Family:

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

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

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

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

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

- Thus

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

# ML for the Exponential Family

- Remember the Exponential Family:

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

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

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

- Thus

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

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

# ML for the Exponential Family

- Suppose we observed i.i.d data $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$.

- We can construct the log-likelihood function, which is a function of the natural parameter $\eta$.

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

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

- Therefore we have

$$-\nabla \ln g(\boldsymbol{\eta}_{\mathrm{ML}}) = \underbrace{\frac{1}{N}\sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)}_{\text{Sufficient Statistic}}$$

Sufficient Statistic

# End

# Bernoulli Distribution

- The Bernoulli distribution is a member of the exponential family:

$$
\begin{aligned}
p(x|\mu) &= \mathrm{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x} \\
&= \exp\left\{x \ln \mu + (1-x)\ln(1-\mu)\right\} \\
&= (1-\mu)\exp\left\{\ln\left(\frac{\mu}{1-\mu}\right)x\right\}
\end{aligned}
$$

- Comparing with the general form of the exponential family:

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

we see that

$$
\eta = \ln\left(\frac{\mu}{1-\mu}\right)
$$

and so

$$
\mu = \underbrace{\sigma(\eta) = \frac{1}{1+\exp(-\eta)}}_{\text{Logistic sigmoid}}.
$$

# Bernoulli Distribution

- The Bernoulli distribution is a member of the exponential family:

$$\begin{aligned} p(x|\mu) &= \mathrm{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x} \\ &= \exp\left\{x \ln \mu + (1-x)\ln(1-\mu)\right\} \\ &= (1-\mu)\exp\left\{\ln\left(\frac{\mu}{1-\mu}\right)x\right\} \end{aligned}$$

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

- The Bernoulli distribution can therefore be written as:

$$p(x|\eta) = \sigma(-\eta)\exp(\eta x)$$

where

$$\begin{aligned} u(x) &= x \\ h(x) &= 1 \\ g(\eta) &= 1-\sigma(\eta) = \sigma(-\eta). \end{aligned}$$

# Multinomial Distribution

- The Multinomial distribution is a member of the exponential family:

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{M} \mu_k^{x_k} = \exp\left\{\sum_{k=1}^{M} x_k \ln \mu_k\right\} = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left(\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right)$$

where $\mathbf{x} = (x_1, \ldots, x_M)^{\mathrm{T}}$ $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_M)^{\mathrm{T}}$

and

$$\begin{aligned} \eta_k &= \ln \mu_k \\ \mathbf{u}(\mathbf{x}) &= \mathbf{x} \\ h(\mathbf{x}) &= 1 \\ g(\boldsymbol{\eta}) &= 1. \end{aligned}$$

NOTE: The parameters $\eta_k$ are not independent since the corresponding $\mu_k$ must satisfy

$$\sum_{k=1}^{M} \mu_k = 1.$$

- In some cases it will be convenient to remove the constraint by expressing the distribution over the M-1 parameters.

# Multinomial Distribution

- The Multinomial distribution is a member of the exponential family:

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{M} \mu_k^{x_k} = \exp\left\{\sum_{k=1}^{M} x_k \ln \mu_k\right\} = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left(\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right)$$

- Let $\mu_M = 1 - \sum_{k=1}^{M-1} \mu_k$

- This leads to:

$$\eta_k = \ln\left(\frac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j}\right) \quad \text{and} \quad \mu_k = \underbrace{\frac{\exp(\eta_k)}{1 + \sum_{j=1}^{M-1} \exp(\eta_j)}}_{\text{Softmax function}}.$$

- Here the parameters $\eta_k$ are independent.

- Note that:

$$0 \leqslant \mu_k \leqslant 1 \quad \text{and} \quad \sum_{k=1}^{M-1} \mu_k \leqslant 1.$$

# Multinomial Distribution

- The Multinomial distribution is a member of the exponential family:

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{M} \mu_k^{x_k} = \exp\left\{\sum_{k=1}^{M} x_k \ln \mu_k\right\} = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left(\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right)$$

- The Multinomial distribution can therefore be written as:

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

where

$$\begin{aligned}
\boldsymbol{\eta} &= (\eta_1, \ldots, \eta_{M-1}, 0)^{\mathrm{T}} \\
\mathbf{u}(\mathbf{x}) &= \mathbf{x} \\
h(\mathbf{x}) &= 1 \\
g(\boldsymbol{\eta}) &= \left(1 + \sum_{k=1}^{M-1} \exp(\eta_k)\right)^{-1}.
\end{aligned}$$

# Gaussian Distribution

- The Gaussian distribution can be written as:

$$\begin{aligned} p(x|\mu, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2\right\} \\ &= h(x)g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(x)\right\} \end{aligned}$$

where

$$\boldsymbol{\eta} = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix} \qquad h(\mathbf{x}) = (2\pi)^{-1/2}$$

$$\mathbf{u}(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \qquad g(\boldsymbol{\eta}) = (-2\eta_2)^{1/2}\exp\left(\frac{\eta_1^2}{4\eta_2}\right).$$

# ML for the Exponential Family

- Remember the Exponential Family:

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

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

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

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

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

- Thus

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

# ML for the Exponential Family

- Remember the Exponential Family:

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

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

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

- Thus

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

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

# ML for the Exponential Family

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

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

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

- Therefore we have

$$-\nabla \ln g(\boldsymbol{\eta}_{\mathrm{ML}}) = \underbrace{\frac{1}{N}\sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)}_{\text{Sufficient Statistic}}$$