10-707: Recitation on Probability Distributions

Slides adapted from Ruslan Salakhutdinov's 10-707 lecture on probability distributions

Why do we have this recitation?

- The distributions we talk about today will be the basics of homeworks
- When we talk about generative models (GAN, VAE, Diffusion models)
  - We will generate data distributions
  - By first sampling from a probability distribution we are familiar with

Basic intro stuff

A random variable = outcome of some uncertain or random trial

RV can be discrete:

Rolling a dice –>{1, 2, 3, 4, 5, 6} are the only possible outcomes

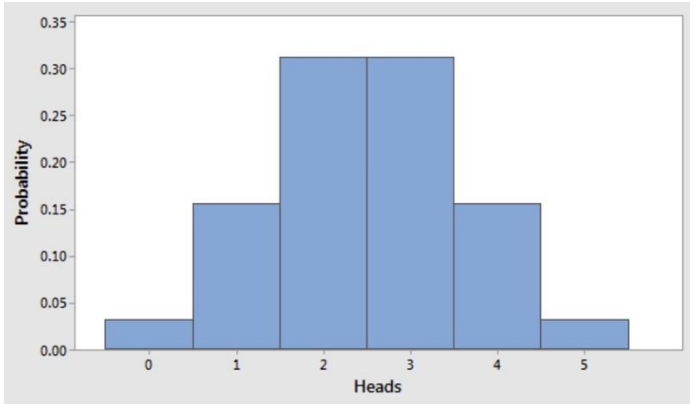
RV can be continuous (any real number in some interval):

Height of a person -> 170.6475 cm, 145.874cm, etc.

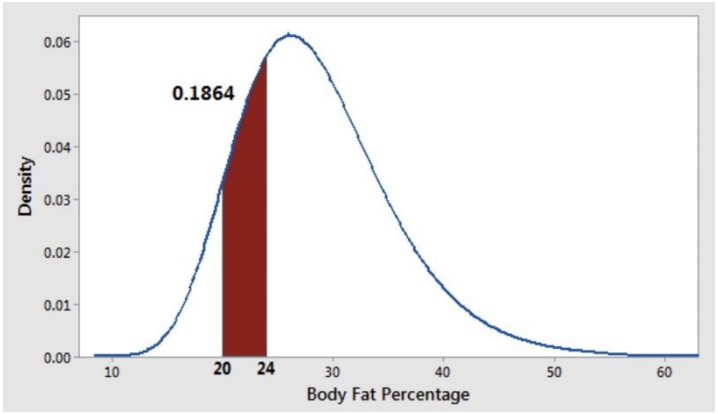
Basic intro stuff

We have tools to describe the probability that a RV will take on a particular value

Discrete: Probability Mass Function



Continuous: Probability Density Function



Bernoulli and Binomial Distributions

Bernoulli Distribution

A Bernoulli trial models the simplest random experiment that can have **two outcomes**: success and failure

When would we use it?

- Tossing a coin and getting a heads
- Passing an exam
- Customer clicking on an ad

Bernoulli Distribution

Probability Mass Function: p is probability of ‘success'

$$P(X = x) = p^x(1-p)^{1-x}, \quad x \in \{0, 1\}$$

$$\mathbb{E}[X] = p \qquad \mathrm{Var}(X) = p(1-p)$$

Bernoulli Distribution

What if we don't have the value of p beforehand?

Create data from an experiment -> use it to estimate p using log likelihood!

Bernoulli Distribution –Parameter Estimation

Collect data: $x_1, x_2, \ldots, x_n$, where each $x_i \in \{0, 1\}$.

Likelihood function: $$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

Log-likelihood: $$\ell(p) = \sum_{i=1}^{n} \left[x_i \ln(p) + (1-x_i) \ln(1-p) \right]$$

Take derivative, set to 0: $$\hat{p}_{\mathrm{MLE}} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Binomial Distribution

Bernoulli: models a **single** experiment

Binomial: models a series of **n independent** trials with the **same probability of success**

When would we use it?

- Tossing a coin and getting a heads in 10 trials
- Number of customers clicking on an ad with 1000 impressions

Binomial Distribution

Probability Mass Function: p is probability of ‘success'

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n$$

$$\mathbb{E}[X] = np \qquad \mathrm{Var}(X) = np(1-p) \qquad \hat{p} = \frac{\sum_j X_j}{mn}$$

MLE

Multinomial Distribution

Binomial: 2 outcomes per trial

Multinomial: generalizes, K outcomes per trial

$$\mu = (\mu_1, \mu_2, \ldots, \mu_K), \quad \sum_{k=1}^{K} \mu_k = 1 \qquad p(x \mid \mu) = \prod_{k=1}^{K} \mu_k^{x_k}$$

Multinomial Distribution

One-hot encoding

I.e. K= 4, correct category = 3

$$x = (0, 0, 1, 0)$$

Beta Distributions

Beta Distribution

- Bernoulli and Binomial **model data** (the probability of a random variable taking on a value
- Beta models uncertainty about p (**the probability of p taking on a certain value**)

Beta Distribution

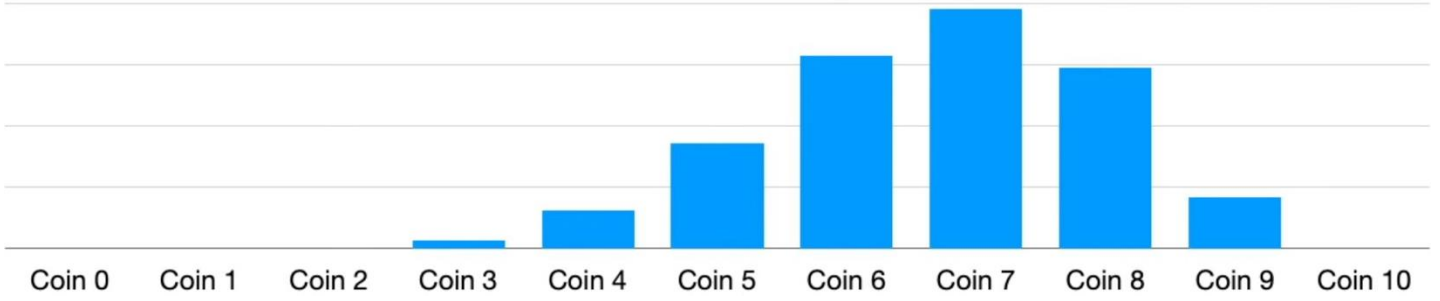
Critical Idea:

We have some data that we collected after running an experiment but we don’t know what the probability p (success) is. So, we use the beta distribution to find the probability that p takes on a certain value between 0 and 1

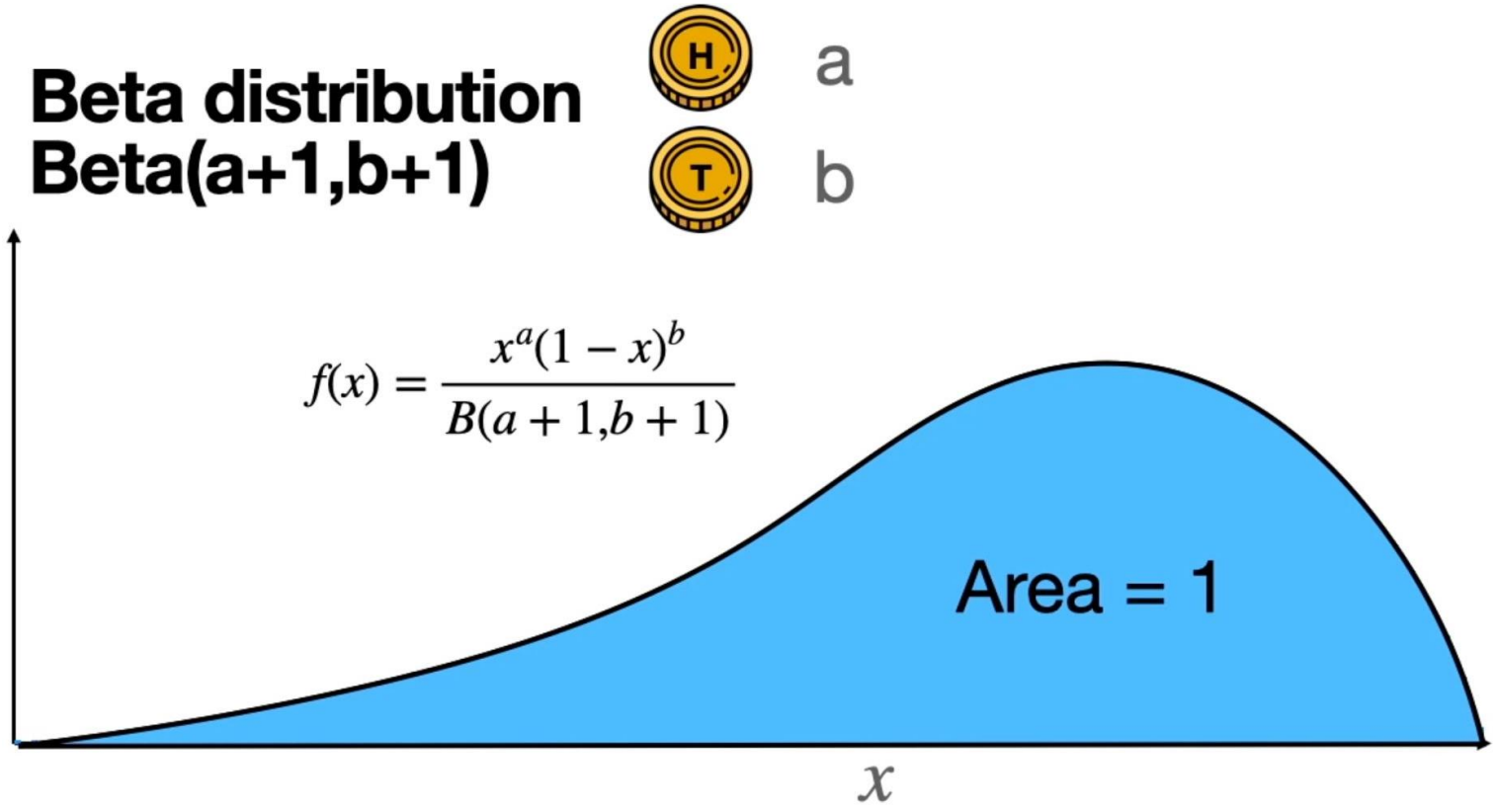
Beta Distribution

7 heads, 3 tails

Coin 0	Coin 1	Coin 2	Coin 3	Coin 4	Coin 5	Coin 6	Coin 7	Coin 8	Coin 9	Coin 10
P(H) = 0	P(H) = 0.1	P(H) = 0.2	P(H) = 0.3	P(H) = 0.4	P(H) = 0.5	P(H) = 0.6	P(H) = 0.7	P(H) = 0.8	P(H) = 0.9	P(H) = 1
$0^7 \cdot 1^3$	$0.1^7 \cdot 0.9^3$	$0.2^7 \cdot 0.8^3$	$0.3^7 \cdot 0.7^3$	$0.4^7 \cdot 0.6^3$	$0.5^7 \cdot 0.5^3$	$0.6^7 \cdot 0.4^3$	$0.7^7 \cdot 0.3^3$	$0.8^7 \cdot 0.2^3$	$0.9^7 \cdot 0.1^3$	$1^7 \cdot 0^3$
0	0.00001	0.001	0.01	0.04	0.13	0.24	0.29	0.21	0.06	0



Beta Distribution



Beta Distribution

If $X \sim \mathrm{Beta}(\alpha, \beta)$:

$$f(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1$$

where

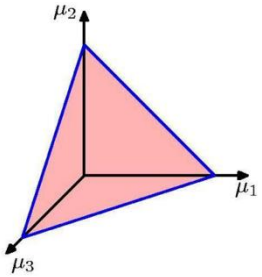
$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta} \qquad \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \qquad \Gamma(x) = \int_0^\infty u^{x-1}e^{-u}\,du$$

Dirichlet Distribution

- Consider a distribution over μ_k ,subject to constraints:

$$\forall k : \mu_k \geqslant 0 \quad \text{and} \quad \sum_{k=1}^{K} \mu_k = 1$$



- The Dirichlet Distribution is defined as:

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

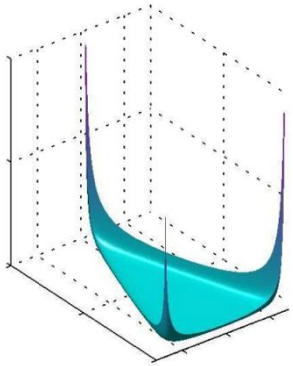
$$\alpha_0 = \sum_{k=1}^{K} \alpha_k$$

Where $\alpha_1, \alpha_2, \ldots, \alpha_k$ are the parameters of the distribution and $\Gamma(x)$ is the gamma function.

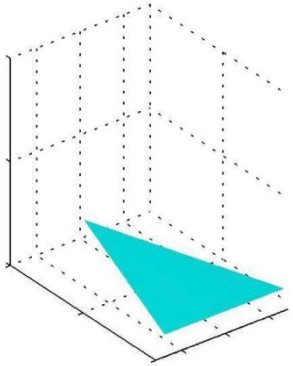
- The Dirichlet distribution is confined to a simplex as a consequence of the constraints.

Dirichlet Distribution

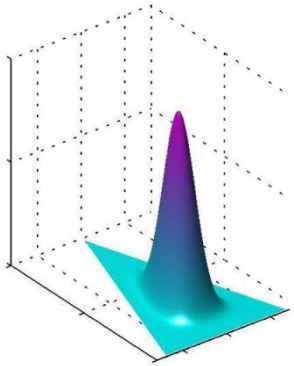
- Plots of the Dirichlet distribution over three variables.



$\alpha_k = 10^{-1}$



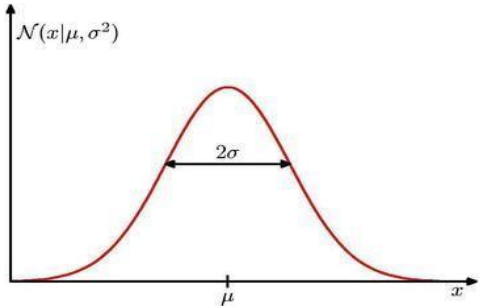
$\alpha_k = 10^{0}$



$\alpha_k = 10^{1}$

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

The Gaussian distribution satisfies:

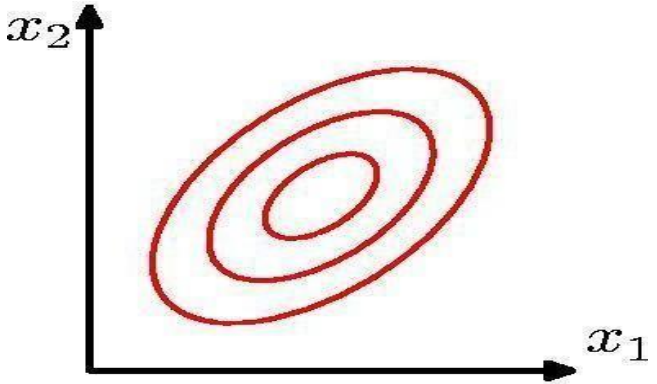
$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

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

Multivariate Gaussian Distribution

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

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



μ is a D–dimensional mean vector.

and |Σ| denotes the determinant of Σ.

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

Moments of the Gaussian Distribution

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

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

The term in z in the factor (z+μ) will vanish by symmetry (positive vs negative).

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

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 Σ governs the covariance of x under the Gaussian distribution, it is called the covariance matrix.

- Lookup table for 1d−Gaussian moments: Normal distribution - Wikipedia

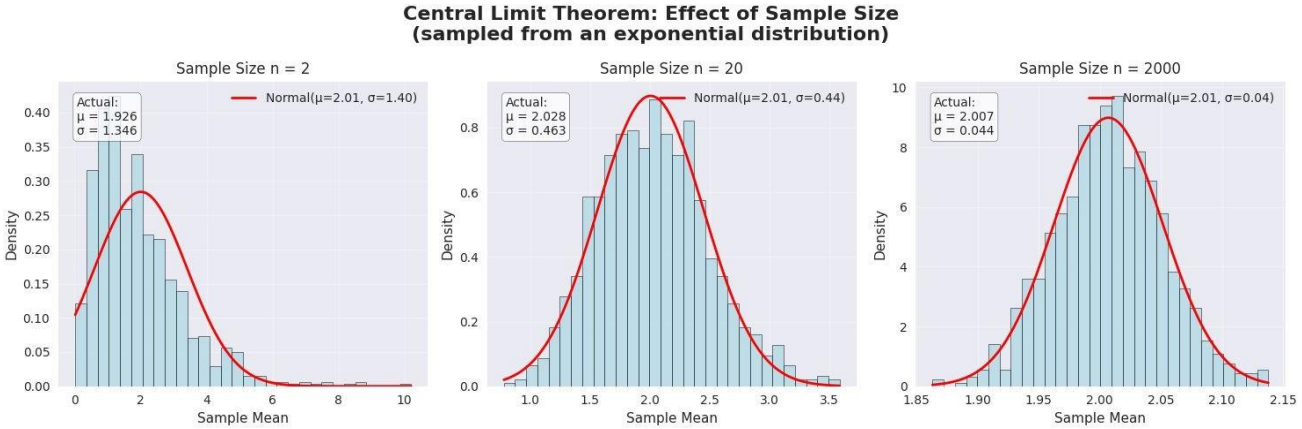
Central Limit Theorem

- The distribution of mean of N i.i.d. random variables tends to a Gaussian distribution as N increases
  - sample mean approximate normality
- It allows us to make statistical inferences, validate performance, and use algorithms that assume normality
  - i.e., assume data come from a gaussian

$$\bar{X}_n \equiv \frac{X_1 + \cdots + X_n}{n}. \qquad \sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right).$$

Central Limit Theorem

- The distribution of mean of N i.i.d. random variables tends to a Gaussian distribution as N increases
  - sample mean approximate normality
- It allows us to make statistical inferences, validate performance, and use algorithms that assume normality
  - i.e., assume data come from a gaussian



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 xa and xb:

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

- 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 Λ_{aa} is not given by the inverse of Σaa

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 xb.

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

Linear function of xb.

Marginal Distribution

- 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.

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

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 μ and Σ:

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

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

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

Sufficient Statistics

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 maximum likelihood estimate of Σ

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

Maximum Likelihood Estimation

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

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

Unbiased estimate

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

Biased estimate

Note that the maximum likelihood estimate of Σ is biased.

We can correct the bias by defining a different estimator:

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