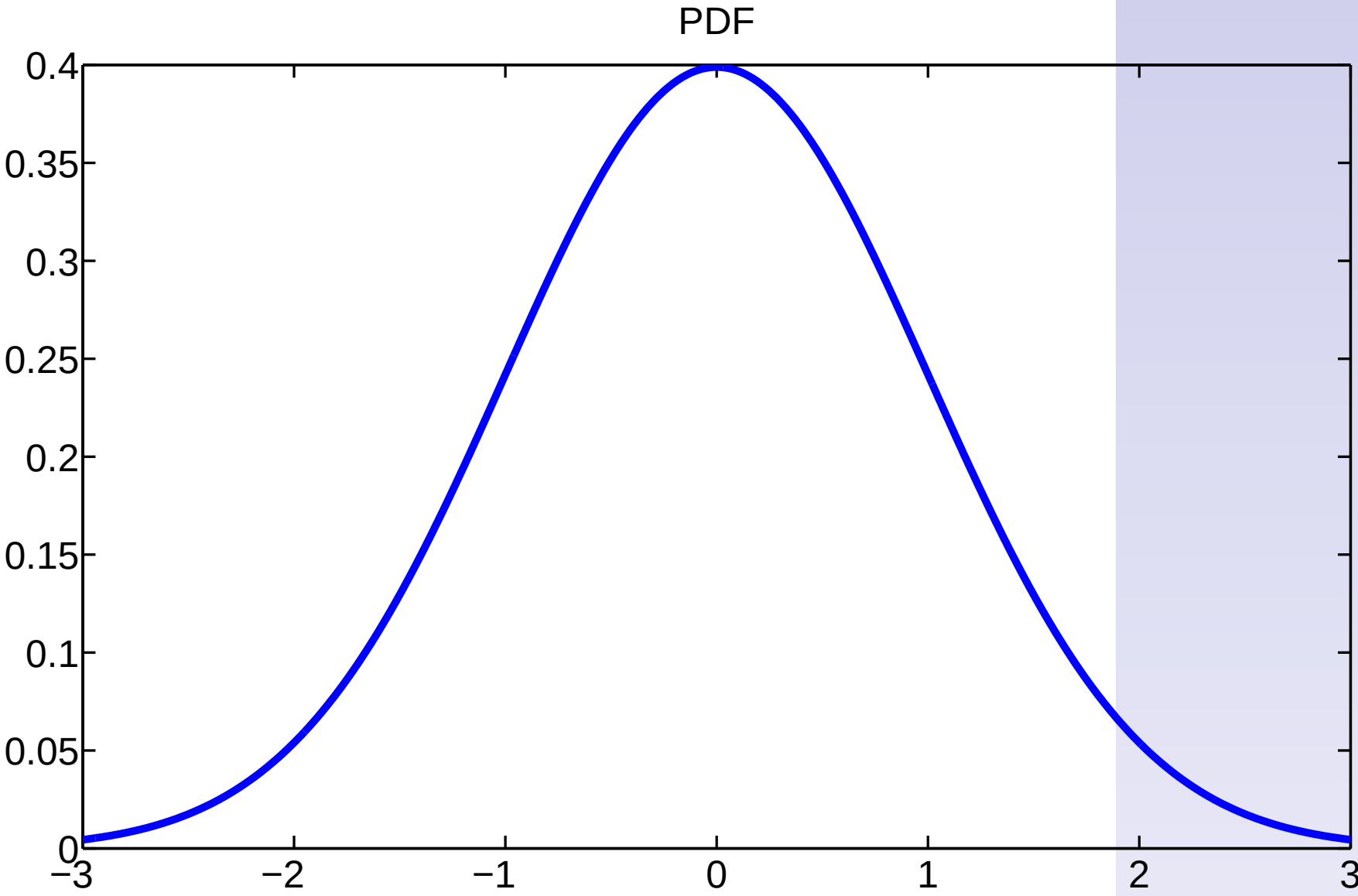
# Machine Learning
# Lecture 3: Gaussian Distributions

Devdatt Dubhashi
dubhashi@chalmers.se

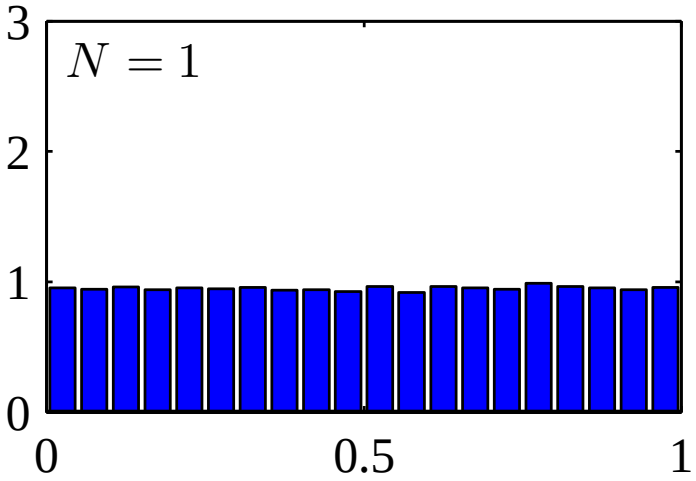
Department of Computer Science and Engineering
Chalmers University

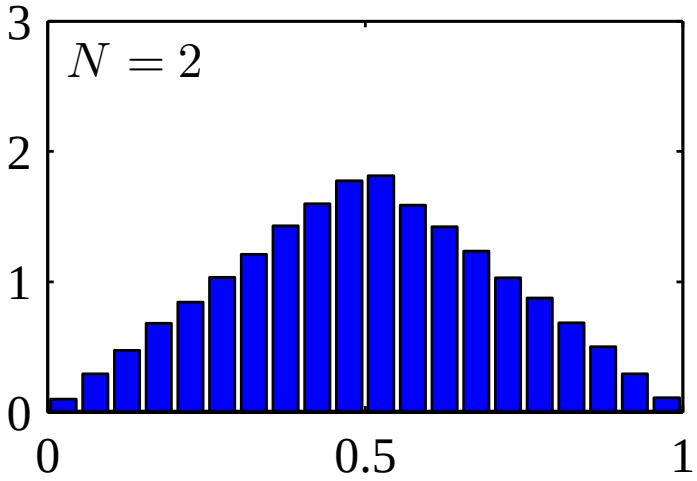
January 25, 2016

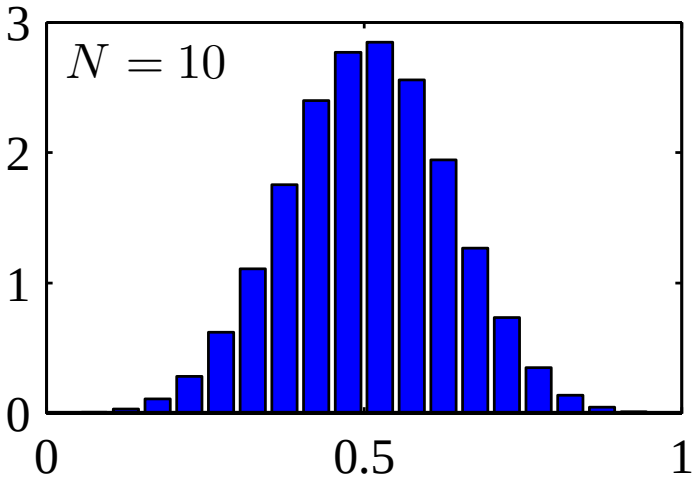
# Gaussian/Normal Distribution

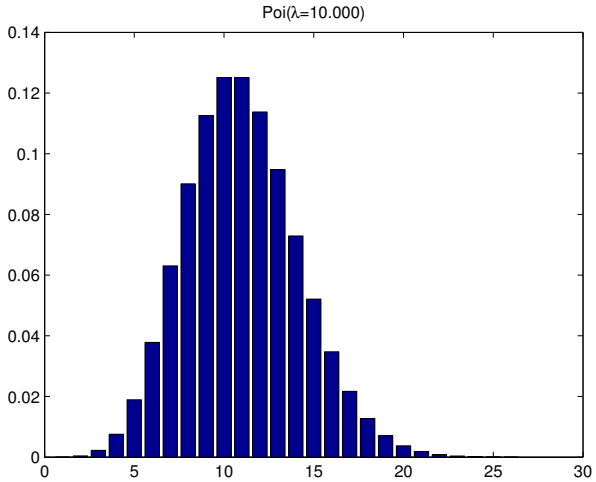


# Gaussian as a Limit



# Univariate Gaussian

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

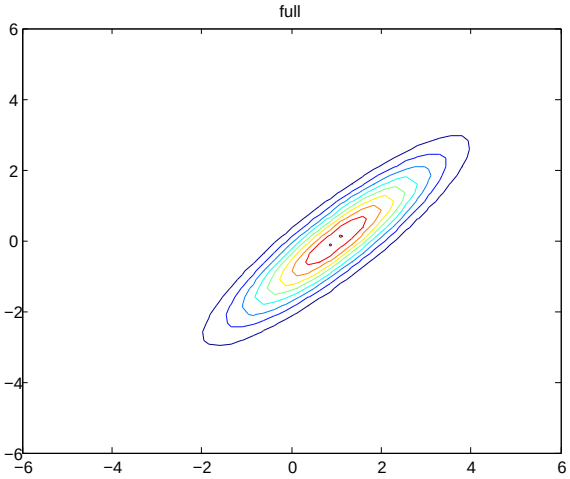
Density $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) = 1$

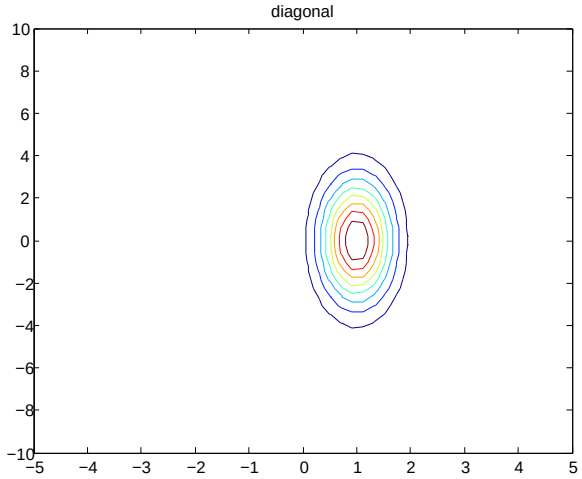
Mean $\int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) = \mu$

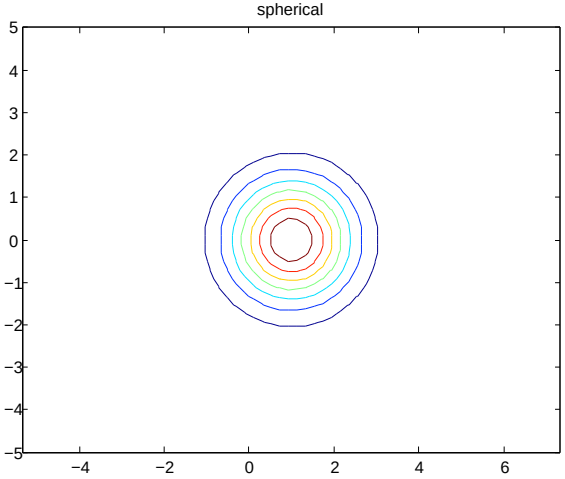
Variance $\int_{-\infty}^{\infty} x^2\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) == \mu^2 + \sigma^2$

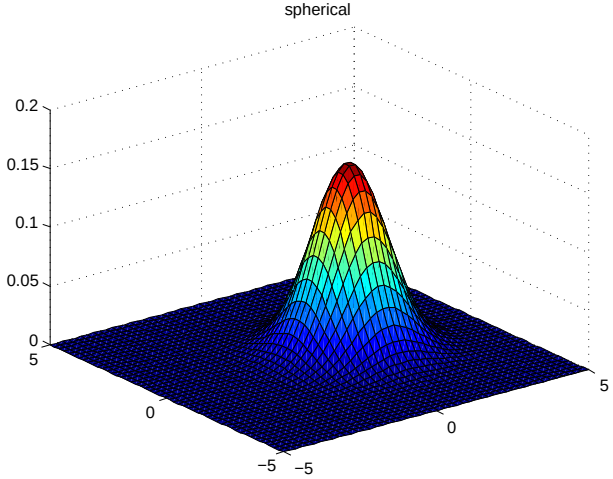
# Multivariate Gaussian





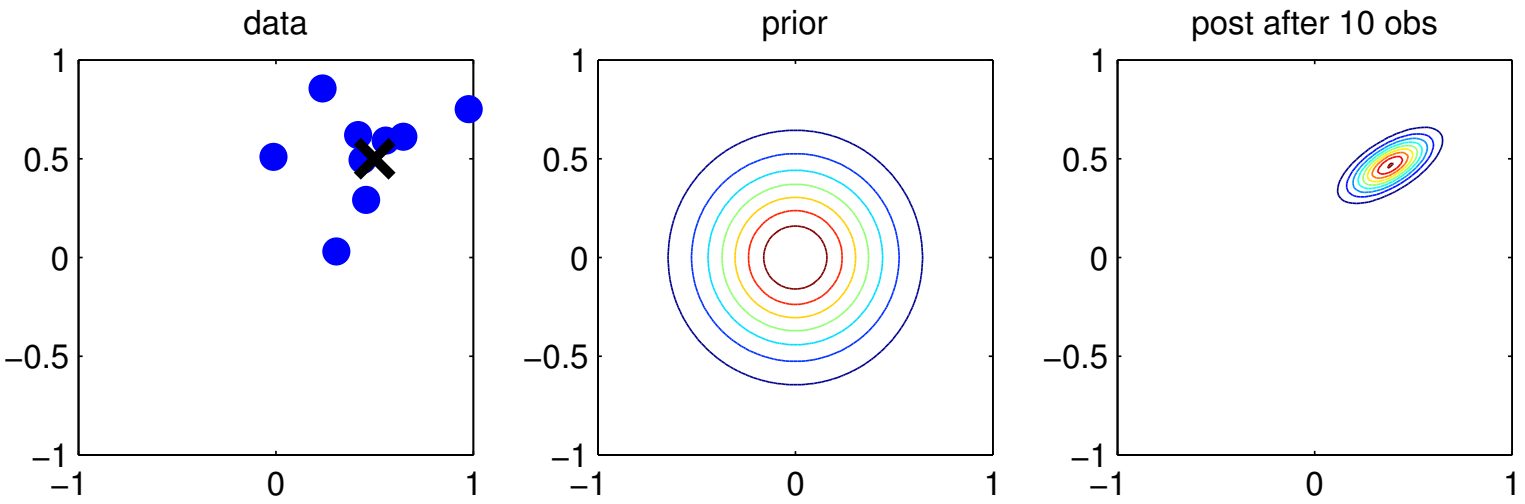
# Multivariate Gaussian



## Multivariate Gaussian pdf

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

# Inference for Gaussian: Localization



# Understanding the Gaussian: 2 D diagonal case

With $n = 2, \mu = (\mu_1, \mu_2)$ and $\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$, the distribution is:

$$\begin{aligned} \mathcal{N}(\mathbf{x} \mid \mu, \mathbf{\Sigma}) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{\sigma_1^2}(x_1 - \mu_1)^2 - \frac{1}{\sigma_2^2}(x_2 - \mu_2)^2\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{\sigma_1^2}(x_1 - \mu_1)^2\right) \cdot \\ &\quad \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{1}{\sigma_2^2}(x_2 - \mu_2)^2\right) \\ &= \mathcal{N}(x_1 \mid \mu_1, \sigma_1^2) \cdot \mathcal{N}(x_2 \mid \mu_2, \sigma_2^2) \end{aligned}$$

In general, $n$ dimensional Gaussian with mean $\mu = (\mu_1, \cdots, \mu_n)$ and $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_n^2)$ is the product of the $n$ one dimensional Gaussians with parameters $(\mu_1, \sigma_1^2), \cdots, (\mu_n, \sigma_n^2)$.

# Shape of the iso-contours

Setting

$$c = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{\sigma_1^2}(x_1-\mu_1)^2 - \frac{1}{\sigma_2^2}(x_2-\mu_2)^2\right),$$

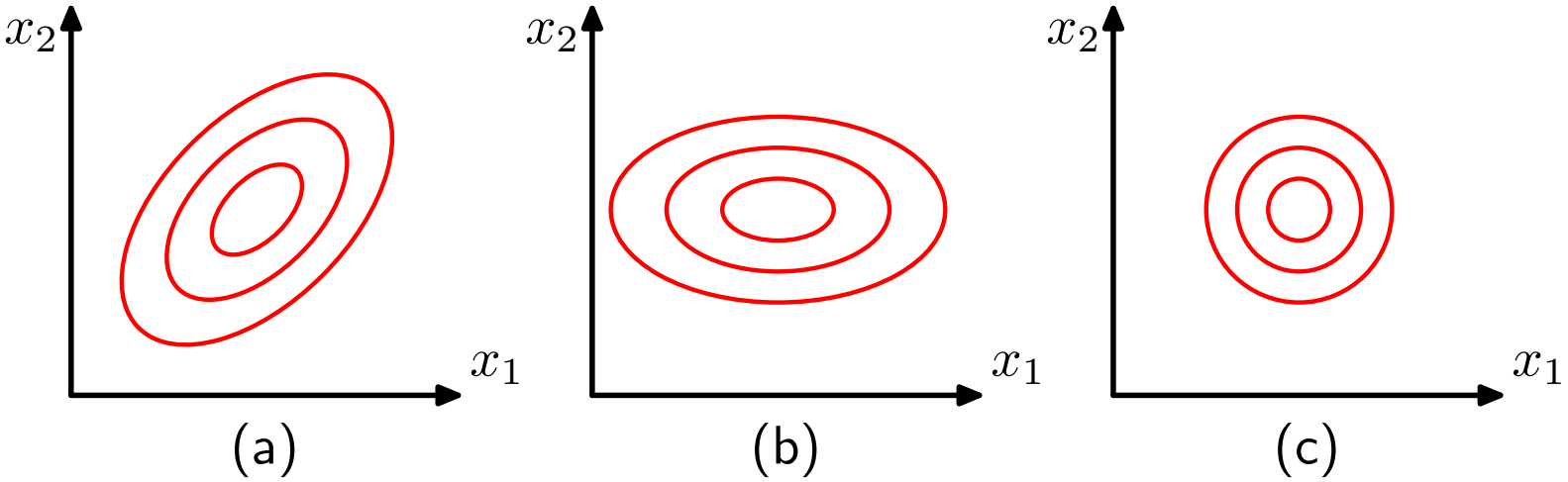
the iso–contours are ellipses given by:

$$\left(\frac{x_1-\mu_1}{r_1}\right)^2 + \left(\frac{x_2-\mu_2}{r_2}\right)^2 = 1,$$

where

$$r_1 = \sqrt{2\sigma_1^2 \log(\frac{1}{2\pi c\sigma_1\sigma_2})}, r_2 = \sqrt{2\sigma_2^2 \log(\frac{1}{2\pi c\sigma_1\sigma_2})}$$

# Iso–contours



# Multivariate Gaussian: Eigendecomposiition

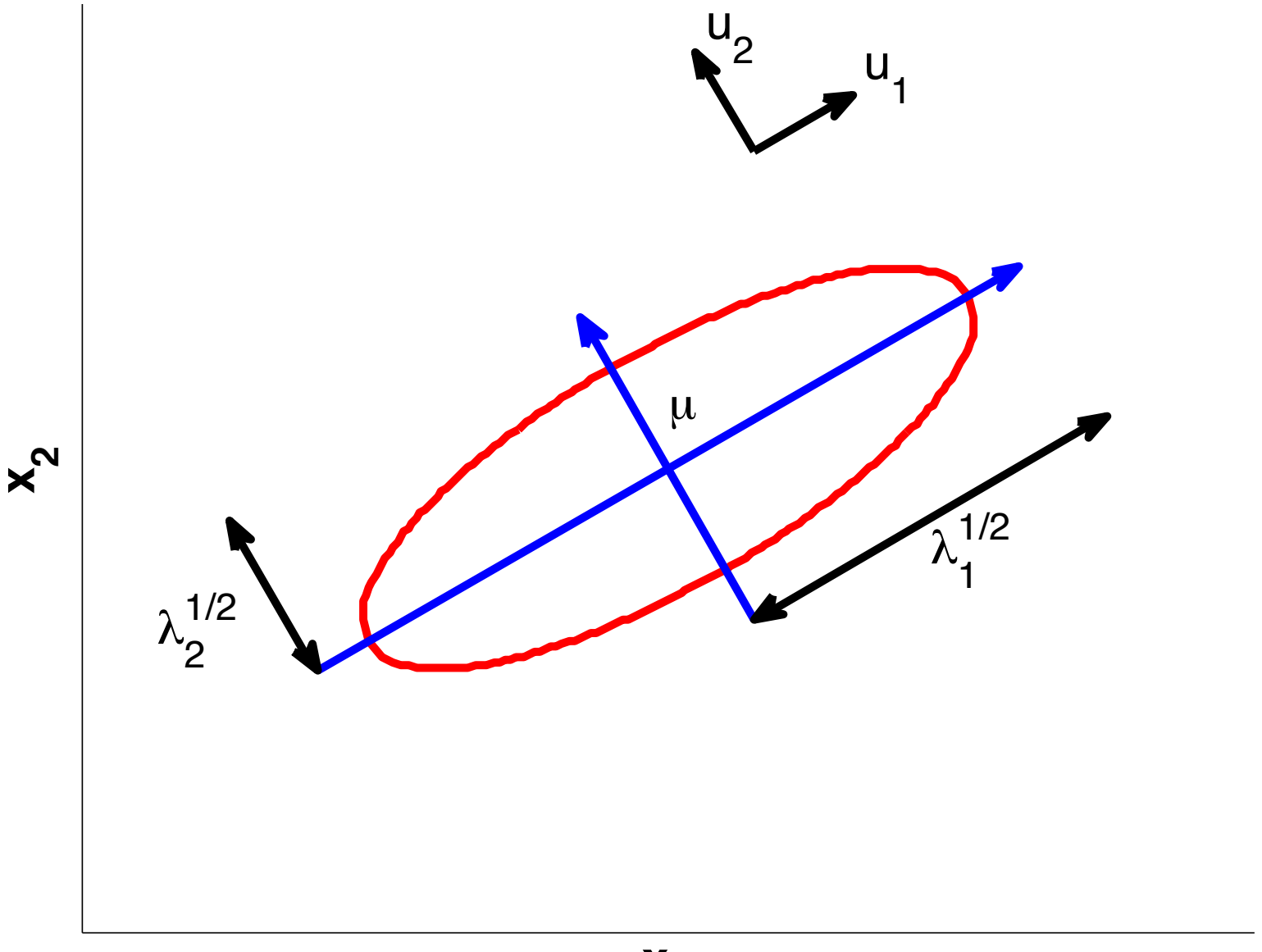
- If $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ is an eigendecomposiiton of $\mathbf{\Sigma}$, then $\mathbf{\Sigma}^{-1} = \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^T$
- 

$$\begin{aligned}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu) &= (\mathbf{x}-\mu^T)\left(\sum_{i=1}^{D}\frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^T\right)(\mathbf{x}-\mu) \\ &= \sum_{i=1}^{D}\frac{1}{\lambda_i}(\mathbf{x}-\mu^T)\mathbf{u}_i\mathbf{u}_i^T(\mathbf{x}-\mu) \\ &= \sum_{i=1}^{D}\frac{y_i^2}{\lambda_i},\end{aligned}$$

  with $y_i := \mathbf{u}_i^T(\mathbf{x}-\mu)$.

- Contours in 2D are ellipses $\frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} = 1$

# Gaussian Ellipses



# MLE for Gaussian

## MLE for Gaussian

If we have $N$ i.i.d. samples for $\mathbf{x}_i \sim \mathcal{N}(\mu, \mathbf{\Sigma})$, then the MLE estimates are

$$\hat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i =: \bar{\mathbf{x}}$$

$$\hat{\mathbf{\Sigma}}_{MLE} = \frac{1}{N}\sum_{i=1}^{N})(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T = \frac{1}{N}\left(\sum_{i+1}^{N}\mathbf{x}_i\mathbf{x}_I^T\right) - \bar{\mathbf{x}}\bar{\mathbf{x}}^T.$$

In single dimension,

$$\hat{\mu} = \frac{1}{N}\sum_i x_i = \bar{x},$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_i (x_i - \bar{x})^2 = \frac{1}{N}\sum_i x_i^2 - \bar{x}^2.$$

# Bayesian Inference: Known variance

- Single dimension
- Variance $\sigma^2$ known.
- Estimate mean $\mu$ frm $n$ independent observations $\mathbf{x} = (x_1, \cdots, x_n)$.

## Likelihood

$$P(\mathbf{x} \mid \mu) = \prod_{i=1}^{n} P(x_i \mid \mu) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right).$$

## MLE estimate

$$\mu_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

# Prior and Posterior: Known variance

We should choose conjugate prior

$$P(\mu) := \mathcal{N}(\mu \mid \mu_0, \sigma_0^2).$$

and we get:

## Posterior

$$\begin{aligned} P(\mu \mid \mathbf{x}) &\propto P(\mathbf{x} \mid \mu) P(\mu) \\ &= \mathcal{N}(\mu \mid \mu_n, \sigma_m^2), \end{aligned}$$

where

$$\mu_n = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\mu_{ML},$$

and

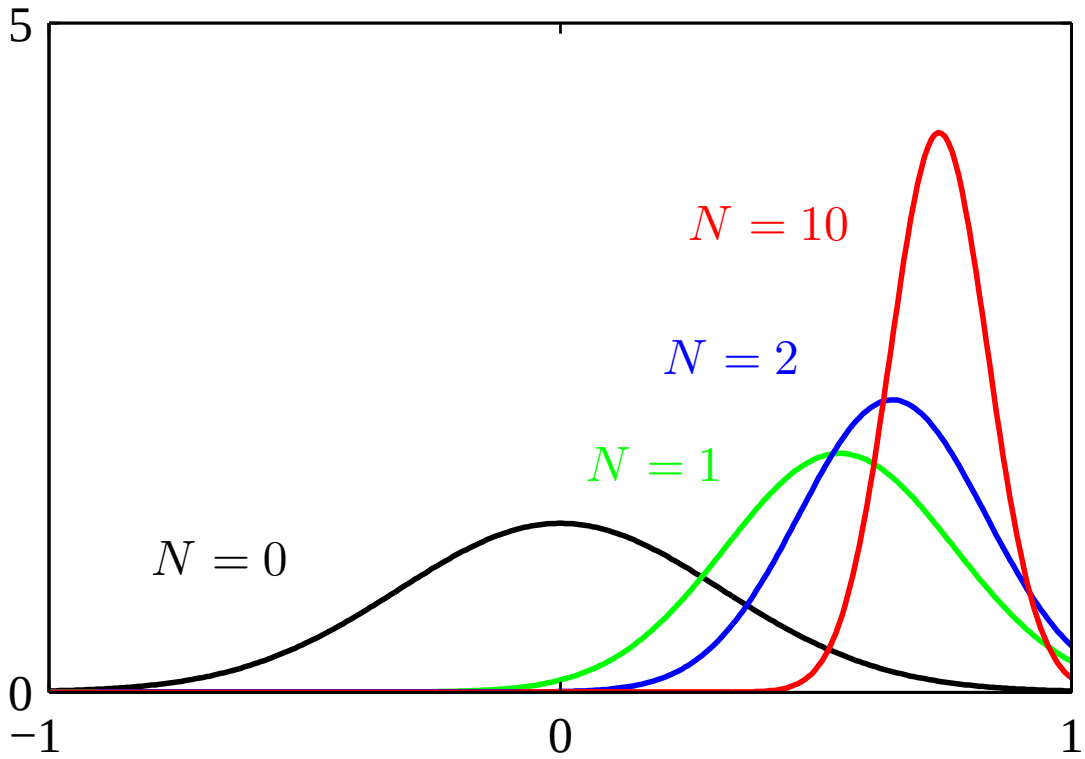
$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}.$$

# Posterior: comments

- Mean is a compromise between the the prior mean $\mu_0$ and the MLE estimate $\mu_{ML}$.
- If $n = 0$, we have the prior mean and as $n \longrightarrow \infty$, the posterior $\mu_n \longrightarrow \mu_{ML}$.
- Variance expressed more naturally in terms of the inverse called the precision which is additive.
- With more data, the precision steadily increases and
- ... as $n \longrightarrow \infty$, $\sigma_n^2 \longrightarrow 0$ and the posterior is peaked at the MLE solution.

# Posterior

Data generated by Gaussian with mean 0.8 and variance 0.1



# Sequential update view

Write the posterior as:

$$P(\mu \mid \mathcal{D}) \propto \left[ P(\mu) \prod_{i=1}^{n-1} P(\mathbf{x}_i \mid \mu) \right] P(\mathbf{x}_n \mid \mu),$$

then the first term is the posterior after observing $n-1$ data points , which can then be used as a prior for the $n$th data point.

# Bayesian inference: unknown variance



- Mean is known but variance is unknown - Homework!
- Both mean and variance unknown ...

# Bayes Factors for Model Selection

- Suppose we have two models $M_0$ and $M_1$, which one is better?
- Use the Bayes Factor which is the ratio of the likelihoods:

$$BF_{1,0} := \frac{P(\mathcal{D} \mid M_1)}{P(\mathcal{D} \mid M_0)}.$$

- Prefer model 1 if $BF_{1,0} > 1$.

# Testing if a coin is fair

**Fair** $P(\mathcal{D} \mid M_0) = 1/2^N$.

**Biased** $P(\mathcal{D} \mid M_1) = \int P(\mathcal{D}\theta)P(\theta)d\theta =$ $B(\alpha_1 + N_1, \alpha_0 + N_0)/B(\alpha_1, \alpha_0)$, if coin has a Beta prior.