

---

# **PATTERN RECOGNITION AND MACHINE LEARNING**

## **CHAPTER 2: PROBABILITY DISTRIBUTIONS**

---

# Parametric Distributions

---

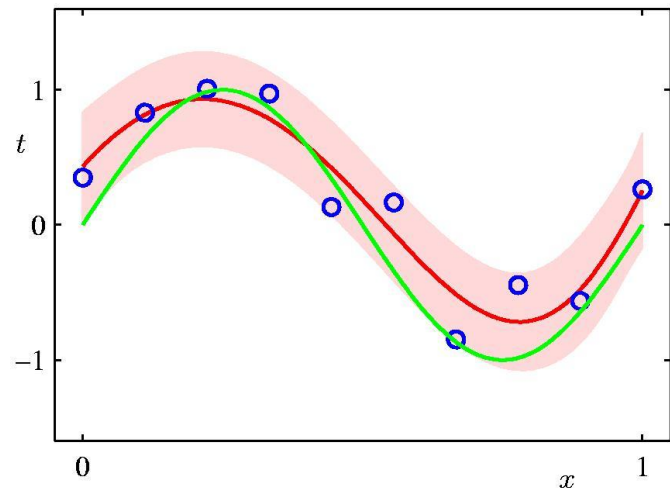
Basic building blocks:  $p(\mathbf{x}|\boldsymbol{\theta})$

Need to determine  $\boldsymbol{\theta}$  given  $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$

Representation:  $\boldsymbol{\theta}^*$  or  $p(\boldsymbol{\theta})$  ?

## Recall Curve Fitting

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w})p(\mathbf{w}|\mathbf{x}, \mathbf{t}) d\mathbf{w}$$



# Binary Variables (1)

---

Coin flipping: heads=1, tails=0

$$p(x = 1|\mu) = \mu$$

Bernoulli Distribution

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

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

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

---

# Binary Variables (2)

---

$N$  coin flips:

$$p(m \text{ heads} | N, \mu)$$

## Binomial Distribution

$$\text{Bin}(m | N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}$$

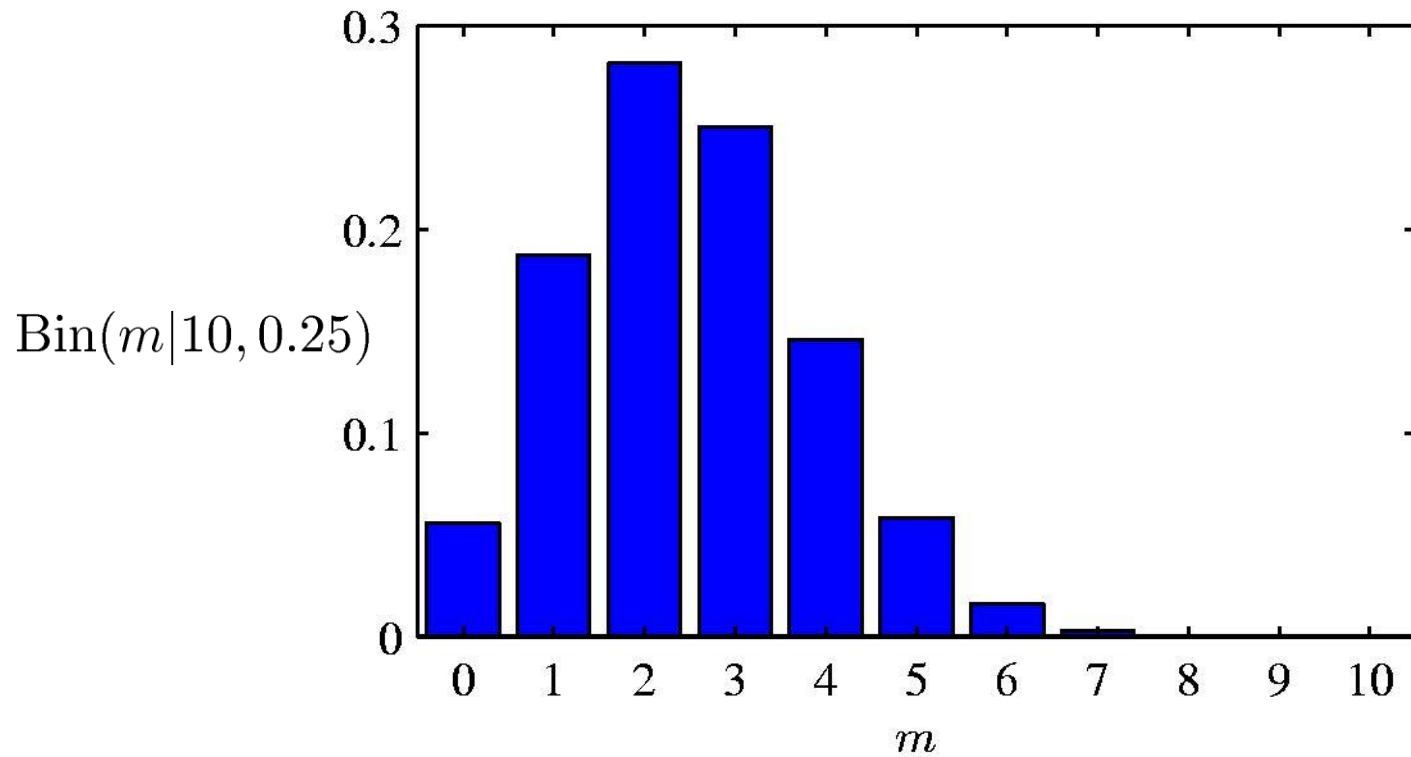
$$\mathbb{E}[m] \equiv \sum_{m=0}^N m \text{Bin}(m | N, \mu) = N\mu$$

$$\text{var}[m] \equiv \sum_{m=0}^N (m - \mathbb{E}[m])^2 \text{Bin}(m | N, \mu) = N\mu(1 - \mu)$$

---

# Binomial Distribution

---



# Parameter Estimation (1)

---

## ML for Bernoulli

Given:  $\mathcal{D} = \{x_1, \dots, x_N\}$ ,  $m$  heads (1),  $N - m$  tails (0)

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

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

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

---

# Parameter Estimation (2)

---

Example:  $\mathcal{D} = \{1, 1, 1\} \rightarrow \mu_{\text{ML}} = \frac{3}{3} = 1$

Prediction: *all* future tosses will land heads up

Overfitting to  $\mathcal{D}$

---

# Beta Distribution

---

Distribution over  $\mu \in [0, 1]$ .

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1}$$

$$\mathbb{E}[\mu] = \frac{a}{a+b}$$

$$\text{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$$



# Bayesian Bernoulli

---

$$\begin{aligned} p(\mu|a_0, b_0, \mathcal{D}) &\propto p(\mathcal{D}|\mu)p(\mu|a_0, b_0) \\ &= \left( \prod_{n=1}^N \mu^{x_n} (1 - \mu)^{1-x_n} \right) \text{Beta}(\mu|a_0, b_0) \\ &\propto \mu^{m+a_0-1} (1 - \mu)^{(N-m)+b_0-1} \\ &\propto \text{Beta}(\mu|a_N, b_N) \end{aligned}$$

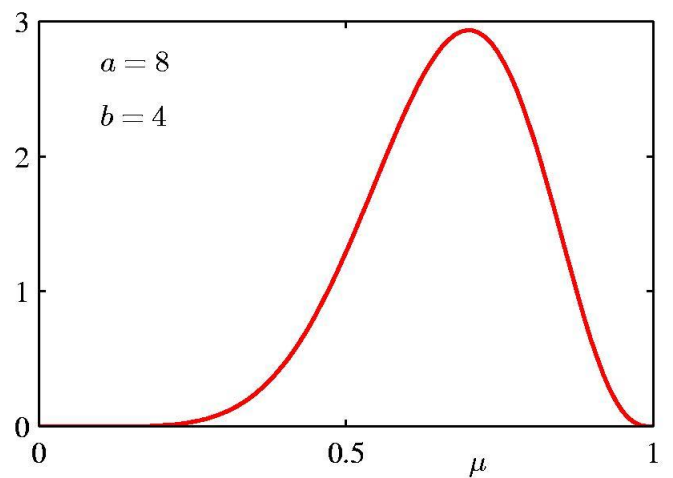
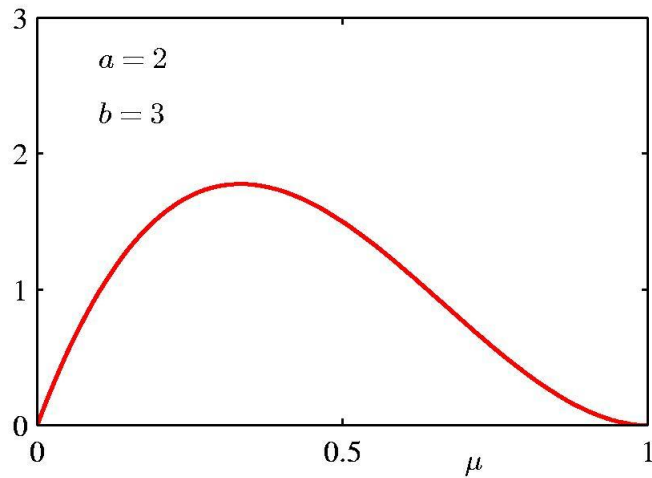
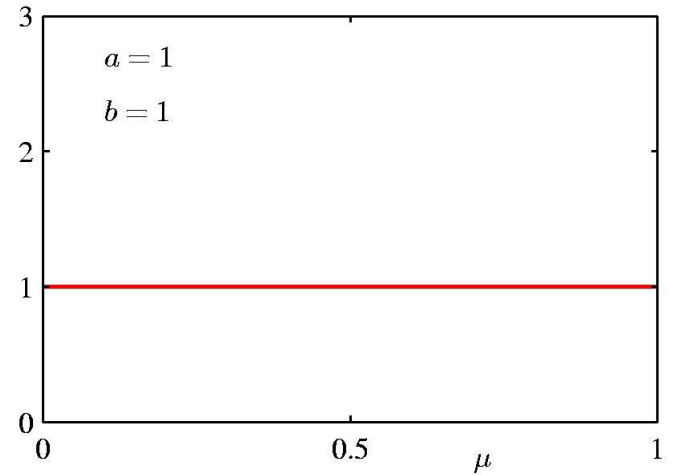
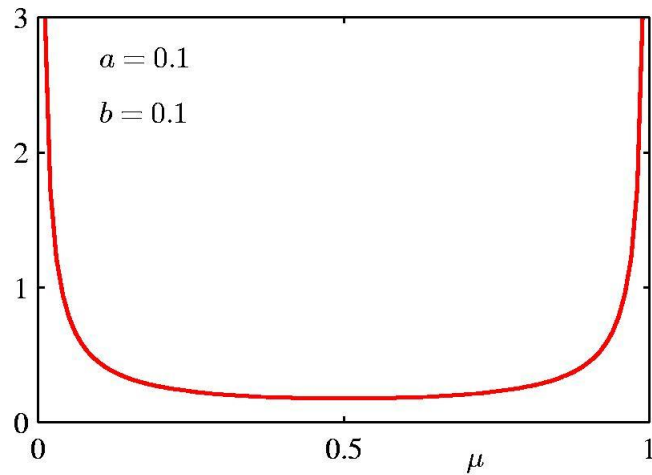
$$a_N = a_0 + m \quad b_N = b_0 + (N - m)$$

The Beta distribution provides the *conjugate* prior for the Bernoulli distribution.

---

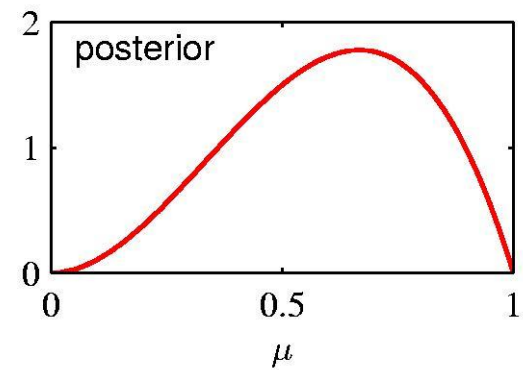
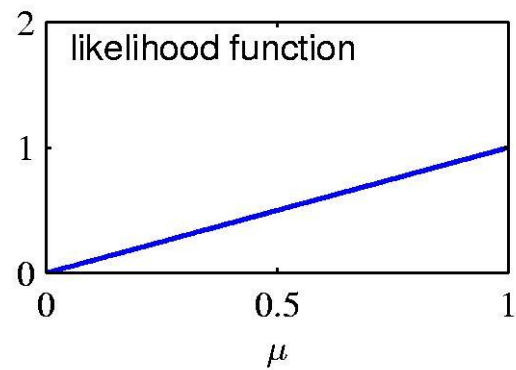
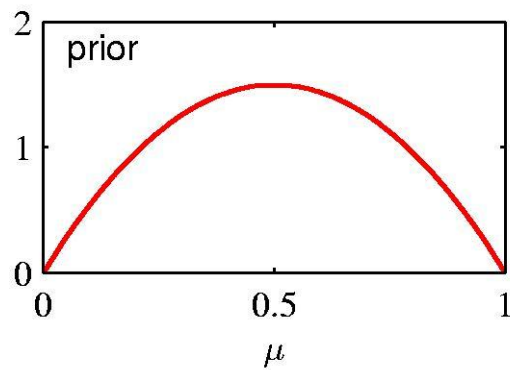
# Beta Distribution

---



# Prior · Likelihood = Posterior

---



# Properties of the Posterior

---

As the size of the data set,  $N$ , increase

$$a_N \rightarrow m$$

$$b_N \rightarrow N - m$$

$$\mathbb{E}[\mu] = \frac{a_N}{a_N + b_N} \rightarrow \frac{m}{N} = \mu_{\text{ML}}$$

$$\text{var}[\mu] = \frac{a_N b_N}{(a_N + b_N)^2 (a_N + b_N + 1)} \rightarrow 0$$

---

# Prediction under the Posterior

---

What is the probability that the next coin toss will land heads up?

$$\begin{aligned} p(x = 1|a_0, b_0, \mathcal{D}) &= \int_0^1 p(x = 1|\mu)p(\mu|a_0, b_0, \mathcal{D}) d\mu \\ &= \int_0^1 \mu p(\mu|a_0, b_0, \mathcal{D}) d\mu \\ &= \mathbb{E}[\mu|a_0, b_0, \mathcal{D}] = \frac{a_N}{b_N} \end{aligned}$$



# Multinomial Variables

---

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

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

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

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

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

---

# ML Parameter estimation

---

Given:  $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$

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

Ensure  $\sum_k \mu_k = 1$ , use a Lagrange multiplier,  $\lambda$ .

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

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

---

# The Multinomial Distribution

---

$$\text{Mult}(m_1, m_2, \dots, m_K | \boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \dots m_K} \prod_{k=1}^K \mu_k^{m_k}$$

$$\mathbb{E}[m_k] = N \mu_k$$

$$\text{var}[m_k] = N \mu_k (1 - \mu_k)$$

$$\text{cov}[m_j, m_k] = -N \mu_j \mu_k$$





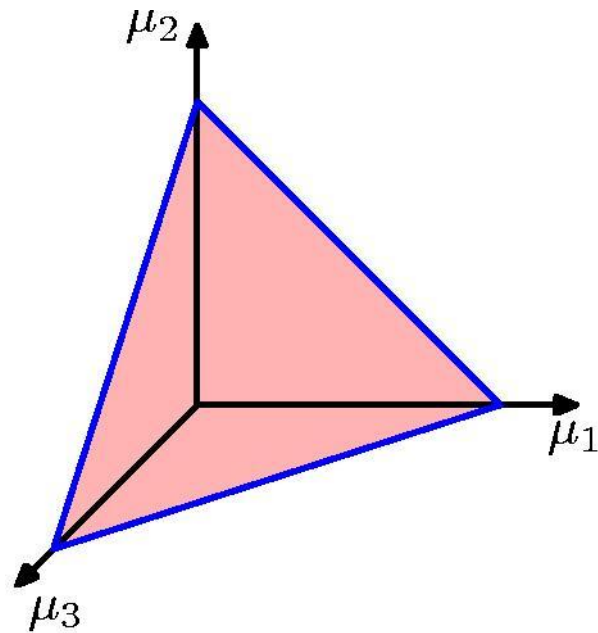
# The Dirichlet Distribution

---

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

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

Conjugate prior for the multinomial distribution.



# Bayesian Multinomial (1)

---

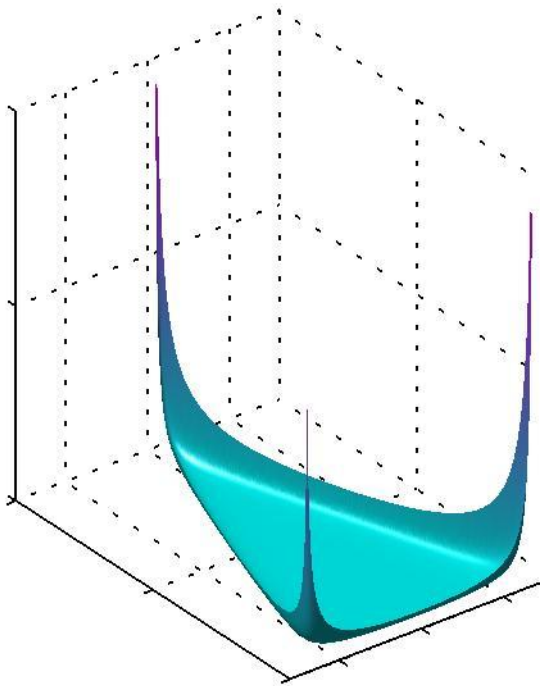
$$p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) \propto p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^K \mu_k^{\alpha_k + m_k - 1}$$

$$\begin{aligned} p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) &= \text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha} + \mathbf{m}) \\ &= \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1) \cdots \Gamma(\alpha_K + m_K)} \prod_{k=1}^K \mu_k^{\alpha_k + m_k - 1} \end{aligned}$$

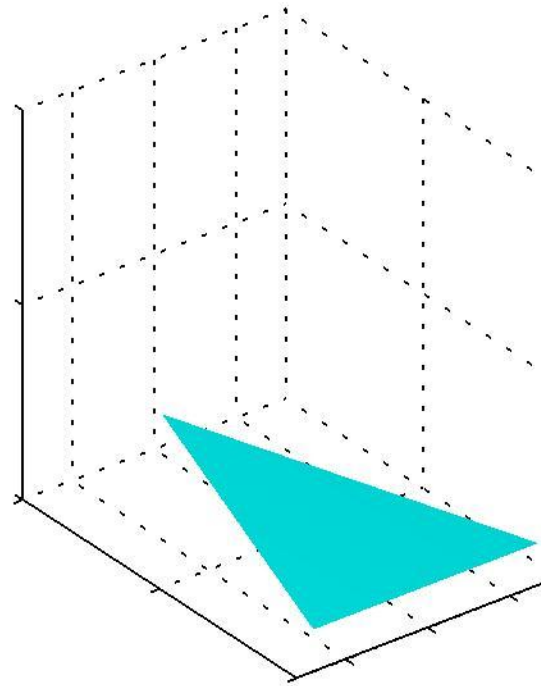
---

# Bayesian Multinomial (2)

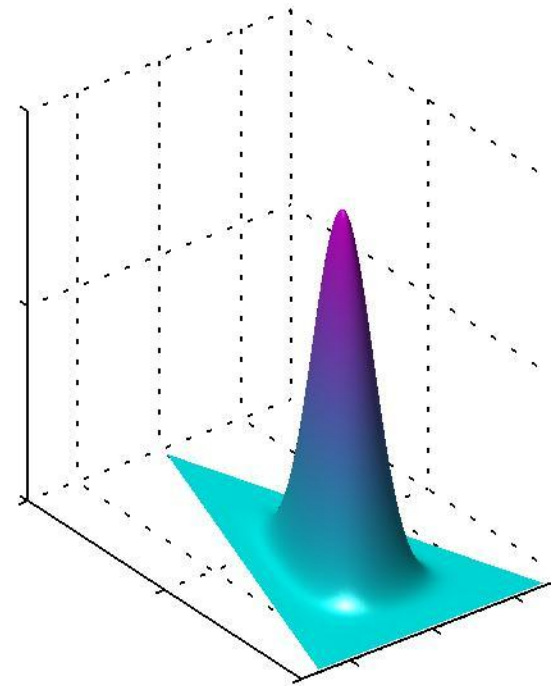
---



$$\alpha_k = 10^{-1}$$



$$\alpha_k = 10^0$$

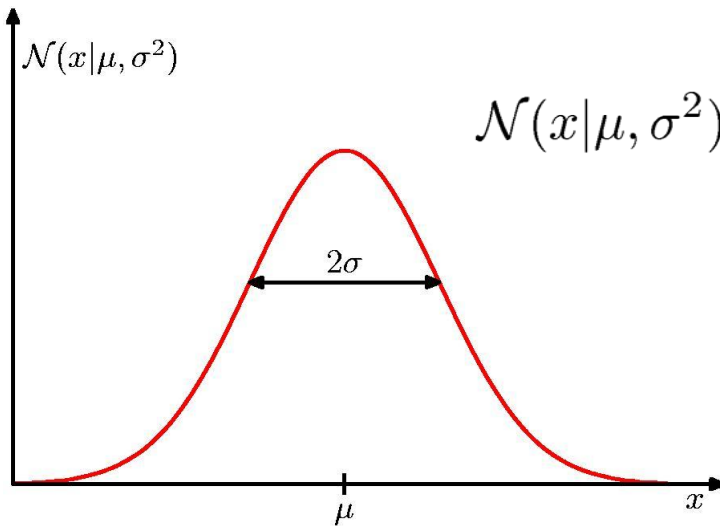


$$\alpha_k = 10^1$$

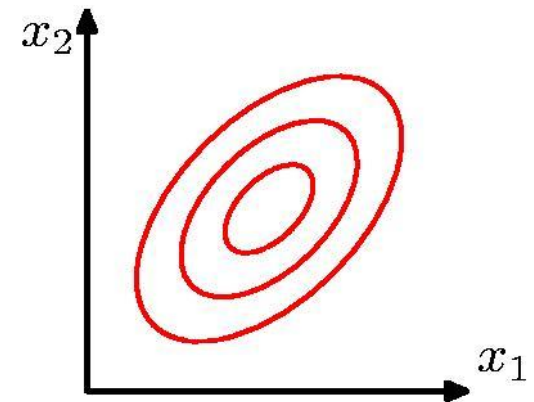
---

# The Gaussian Distribution

---



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



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

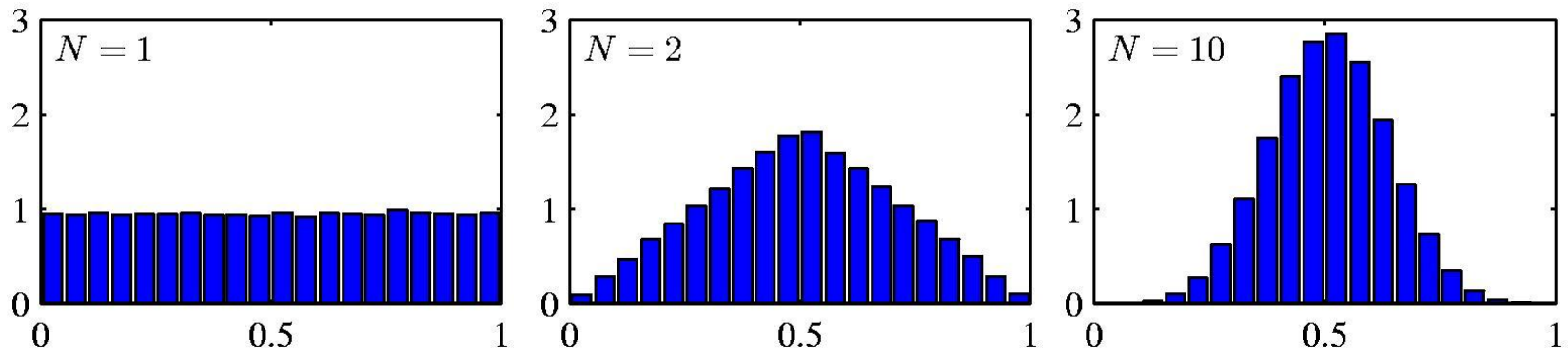
---

# Central Limit Theorem

---

The distribution of the sum of  $N$  i.i.d. random variables becomes increasingly Gaussian as  $N$  grows.

Example:  $N$  uniform  $[0,1]$  random variables.



# Geometry of the Multivariate Gaussian

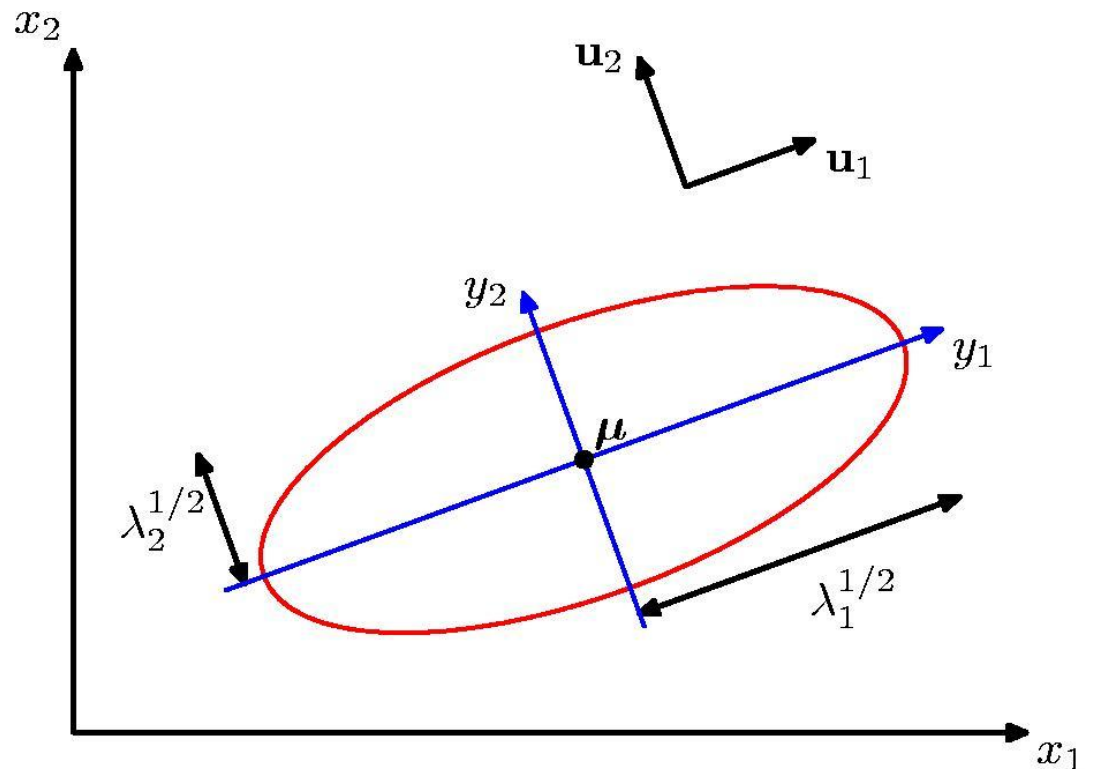
---

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^D \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T$$

$$\Delta^2 = \sum_{i=1}^D \frac{y_i^2}{\lambda_i}$$

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



# Moments of the Multivariate Gaussian (1)

---

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

thanks to anti-symmetry of  $\mathbf{z}$

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

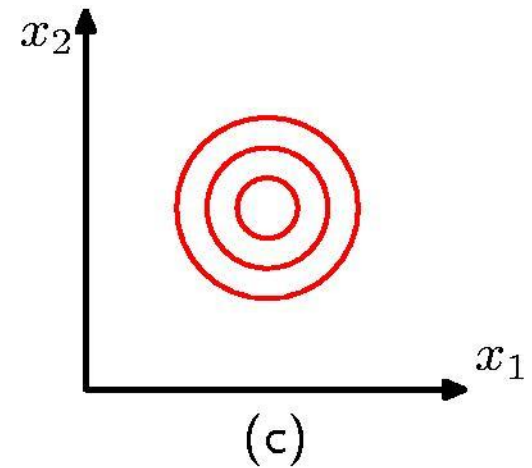
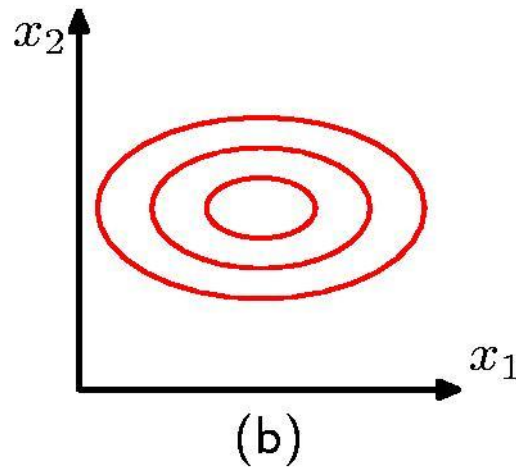
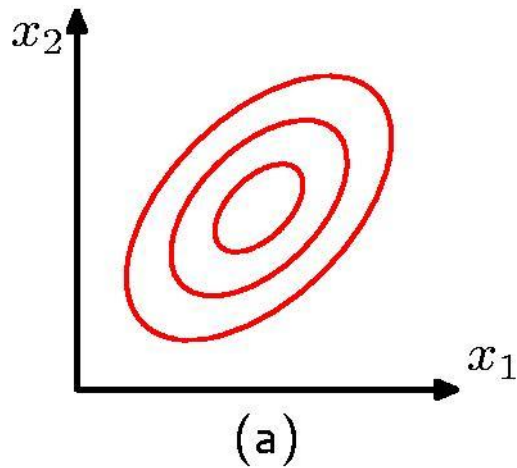
---

# Moments of the Multivariate Gaussian (2)

---

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

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





# Partitioned Gaussian Distributions

---

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

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

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

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

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$$

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

---

# Partitioned Conditionals and Marginals

---

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

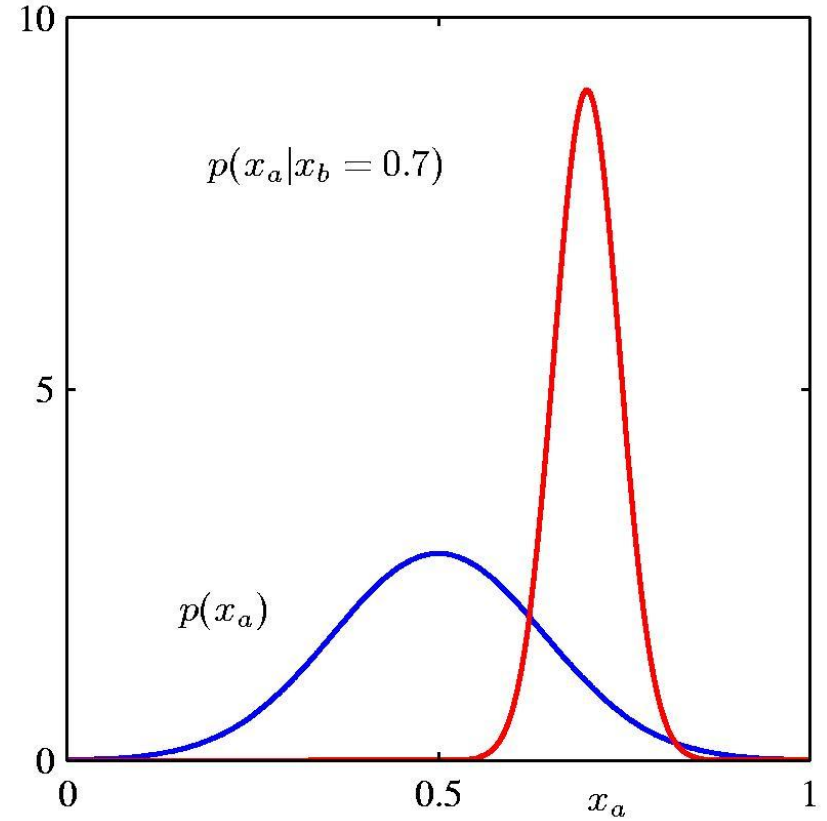
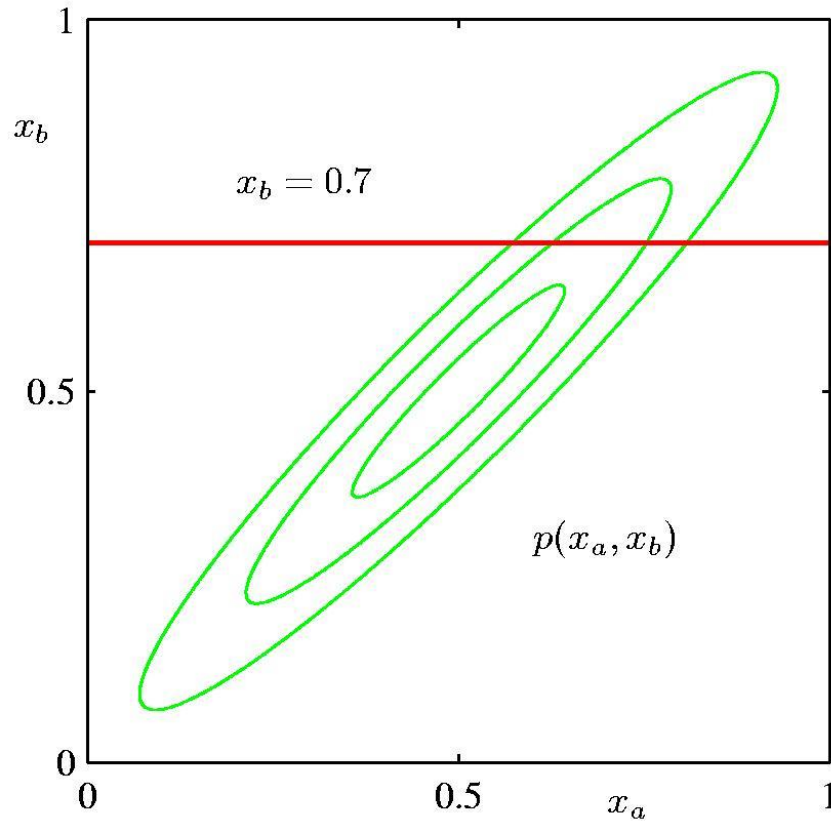
$$\begin{aligned}\boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab} \boldsymbol{\Sigma}_{bb}^{-1} \boldsymbol{\Sigma}_{ba} \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b} \{ \boldsymbol{\Lambda}_{aa} \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}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

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

---

# Partitioned Conditionals and Marginals

---



# Bayes' Theorem for Gaussian Variables

---

Given

$$\begin{aligned}p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \\p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1})\end{aligned}$$

we have

$$\begin{aligned}p(\mathbf{y}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T) \\p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma})\end{aligned}$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}$$

---

# Maximum Likelihood for the Gaussian (1)

---

Given i.i.d. data  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^T$ , the log likelihood function is given by

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

Sufficient statistics

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

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

---

# Maximum Likelihood for the Gaussian (2)

---

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

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

---

# Maximum Likelihood for the Gaussian (3)

---

Under the true distribution

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

Hence define

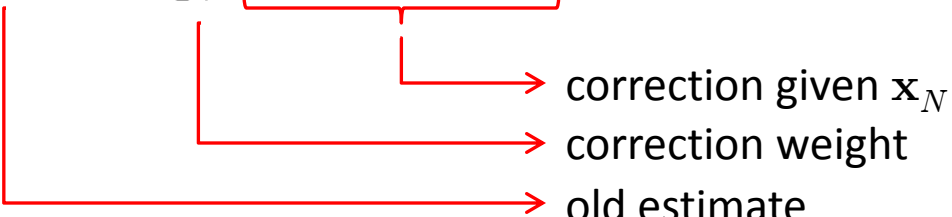
$$\tilde{\boldsymbol{\Sigma}} = \frac{1}{N-1} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^{\text{T}}.$$

---

# Sequential Estimation

---

Contribution of the  $N^{\text{th}}$  data point,  $\mathbf{x}_N$

$$\begin{aligned}\boldsymbol{\mu}_{\text{ML}}^{(N)} &= \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n \\ &= \frac{1}{N} \mathbf{x}_N + \frac{1}{N} \sum_{n=1}^{N-1} \mathbf{x}_n \\ &= \frac{1}{N} \mathbf{x}_N + \frac{N-1}{N} \boldsymbol{\mu}_{\text{ML}}^{(N-1)} \\ &= \boldsymbol{\mu}_{\text{ML}}^{(N-1)} + \frac{1}{N} (\mathbf{x}_N - \boldsymbol{\mu}_{\text{ML}}^{(N-1)})\end{aligned}$$


correction given  $\mathbf{x}_N$   
correction weight  
old estimate

---



# The Robbins-Monro Algorithm (1)

---

Consider  $\theta$  and  $z$  governed by  $p(z, \theta)$  and define the *regression function*

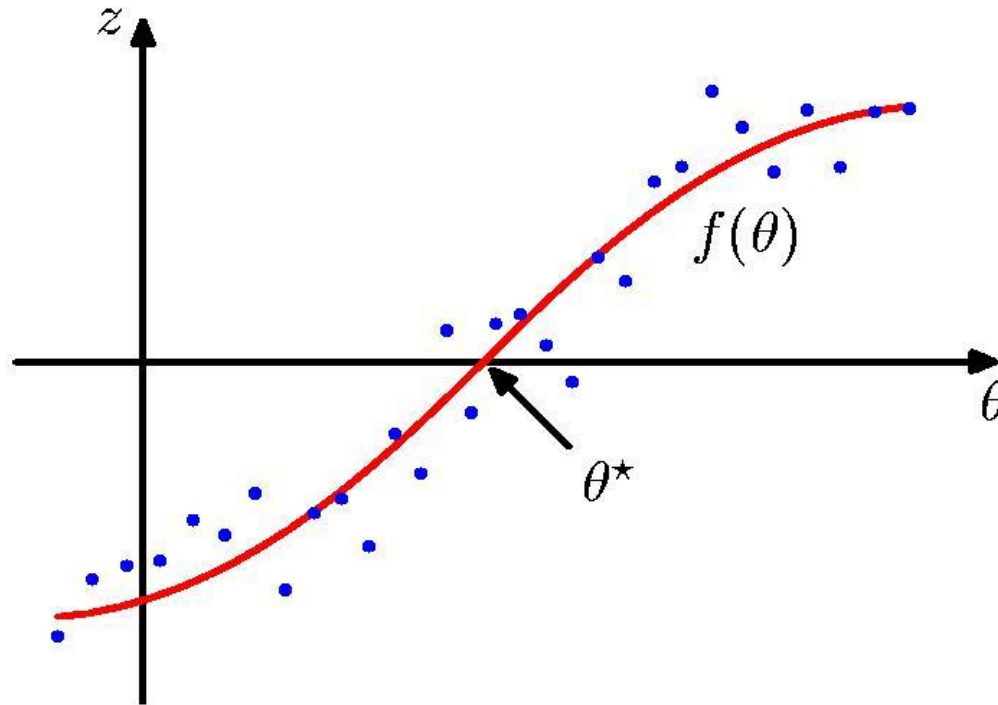
$$f(\theta) \equiv \mathbb{E}[z|\theta] = \int zp(z|\theta) dz$$

Seek  $\theta^*$  such that  $f(\theta^*) = 0$ .

---

# The Robbins-Monro Algorithm (2)

---



Assume we are given samples from  $p(z, \theta)$ , one at the time.

---

# The Robbins-Monro Algorithm (3)

---

Successive estimates of  $\theta^*$  are then given by

$$\theta^{(N)} = \theta^{(N-1)} - a_{N-1} z(\theta^{(N-1)}).$$

Conditions on  $a_N$  for convergence :

$$\lim_{N \rightarrow \infty} a_N = 0 \quad \sum_{N=1}^{\infty} a_N = \infty \quad \sum_{N=1}^{\infty} a_N^2 < \infty$$



# Robbins-Monro for Maximum Likelihood (1)

---

Regarding

$$-\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \frac{\partial}{\partial \theta} \ln p(x_n | \theta) = \mathbb{E}_x \left[ -\frac{\partial}{\partial \theta} \ln p(x | \theta) \right]$$

as a regression function, finding its root is equivalent to finding the maximum likelihood solution  $\theta_{\text{ML}}$ . Thus

$$\theta^{(N)} = \theta^{(N-1)} - a_{N-1} \frac{\partial}{\partial \theta^{(N-1)}} \left[ -\ln p(x_N | \theta^{(N-1)}) \right].$$

---

# Robbins-Monro for Maximum Likelihood (2)

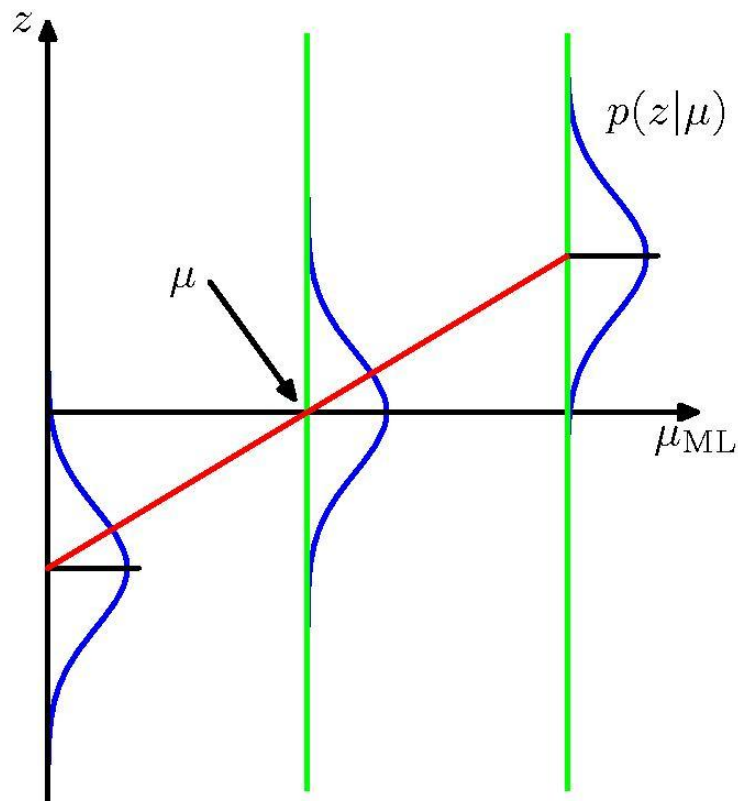
---

Example: estimate the mean of a Gaussian.

$$\begin{aligned} z &= \frac{\partial}{\partial \mu_{\text{ML}}} [-\ln p(x|\mu_{\text{ML}}, \sigma^2)] \\ &= -\frac{1}{\sigma^2}(x - \mu_{\text{ML}}) \end{aligned}$$

The distribution of  $z$  is Gaussian with mean  $\mu - \mu_{\text{ML}}$ .

For the Robbins-Monro update equation,  $a_N = \sigma^2/N$ .



# Bayesian Inference for the Gaussian (1)

---

Assume  $\sigma^2$  is known. Given i.i.d. data

$\mathbf{x} = \{x_1, \dots, x_N\}$ , the likelihood function for  $\mu$  is given by

$$p(\mathbf{x}|\mu) = \prod_{n=1}^N p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 \right\}.$$

This has a Gaussian shape as a function of  $\mu$  (but it is *not* a distribution over  $\mu$ ).

---

# Bayesian Inference for the Gaussian (2)

---

Combined with a Gaussian prior over  $\mu$ ,

$$p(\mu) = \mathcal{N}(\mu | \mu_0, \sigma_0^2).$$

this gives the posterior

$$p(\mu | \mathbf{x}) \propto p(\mathbf{x} | \mu) p(\mu).$$

Completing the square over  $\mu$ , we see that

$$p(\mu | \mathbf{x}) = \mathcal{N}(\mu | \mu_N, \sigma_N^2)$$

---

# Bayesian Inference for the Gaussian (3)

---

... where

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\text{ML}}, \quad \mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}.$$

Note:

	$N = 0$	$N \rightarrow \infty$
$\mu_N$	$\mu_0$	$\mu_{\text{ML}}$
$\sigma_N^2$	$\sigma_0^2$	0

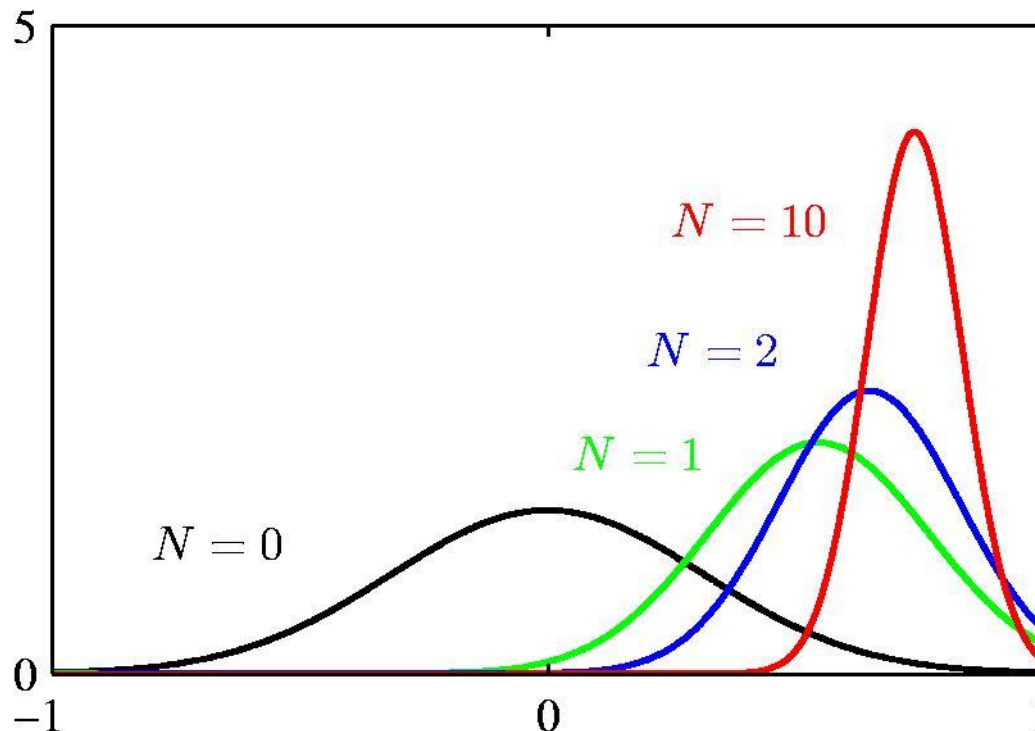
---



# Bayesian Inference for the Gaussian (4)

---

Example:  $p(\mu|\mathbf{x}) = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$  for  $N = 0, 1, 2$  and 10.



# Bayesian Inference for the Gaussian (5)

---

## Sequential Estimation

$$\begin{aligned} p(\mu|\mathbf{x}) &\propto p(\mu)p(\mathbf{x}|\mu) \\ &= \left[ p(\mu) \prod_{n=1}^{N-1} p(x_n|\mu) \right] p(x_N|\mu) \\ &\propto \mathcal{N}(\mu|\mu_{N-1}, \sigma_{N-1}^2) p(x_N|\mu) \end{aligned}$$

The posterior obtained after observing  $N - 1$  data points becomes the prior when we observe the  $N^{\text{th}}$  data point.

---

# Bayesian Inference for the Gaussian (6)

---

Now assume  $\mu$  is known. The likelihood function for  $\lambda = 1/\sigma^2$  is given by

$$p(\mathbf{x}|\lambda) = \prod_{n=1}^N \mathcal{N}(x_n|\mu, \lambda^{-1}) \propto \lambda^{N/2} \exp \left\{ -\frac{\lambda}{2} \sum_{n=1}^N (x_n - \mu)^2 \right\}.$$

This has a Gamma shape as a function of  $\lambda$ .

---

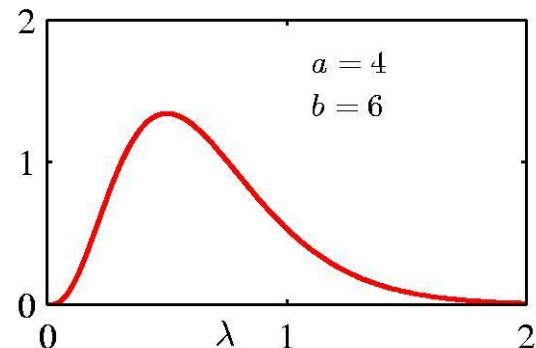
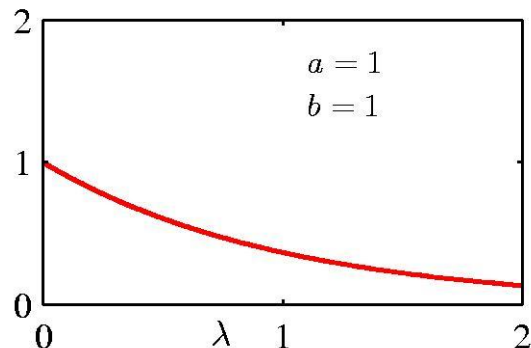
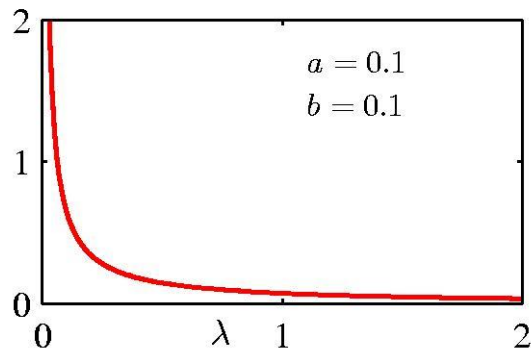
# Bayesian Inference for the Gaussian (7)

---

## The Gamma distribution

$$\text{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

$$\mathbb{E}[\lambda] = \frac{a}{b} \qquad \text{var}[\lambda] = \frac{a}{b^2}$$



# Bayesian Inference for the Gaussian (8)

---

Now we combine a Gamma prior,  $\text{Gam}(\lambda|a_0, b_0)$ , with the likelihood function for  $\lambda$  to obtain

$$p(\lambda|\mathbf{x}) \propto \lambda^{a_0-1} \lambda^{N/2} \exp \left\{ -b_0 \lambda - \frac{\lambda}{2} \sum_{n=1}^N (x_n - \mu)^2 \right\}$$

which we recognize as  $\text{Gam}(\lambda|a_N, b_N)$  with

$$\begin{aligned} a_N &= a_0 + \frac{N}{2} \\ b_N &= b_0 + \frac{1}{2} \sum_{n=1}^N (x_n - \mu)^2 = b_0 + \frac{N}{2} \sigma_{\text{ML}}^2. \end{aligned}$$

---

# Bayesian Inference for the Gaussian (9)

---

If both  $\mu$  and  $\lambda$  are unknown, the joint likelihood function is given by

$$p(\mathbf{x}|\mu, \lambda) = \prod_{n=1}^N \left( \frac{\lambda}{2\pi} \right)^{1/2} \exp \left\{ -\frac{\lambda}{2} (x_n - \mu)^2 \right\}$$
$$\propto \left[ \lambda^{1/2} \exp \left( -\frac{\lambda \mu^2}{2} \right) \right]^N \exp \left\{ \lambda \mu \sum_{n=1}^N x_n - \frac{\lambda}{2} \sum_{n=1}^N x_n^2 \right\}.$$

We need a prior with the same functional dependence on  $\mu$  and  $\lambda$ .

---

# Bayesian Inference for the Gaussian (10)

---

## The Gaussian-gamma distribution

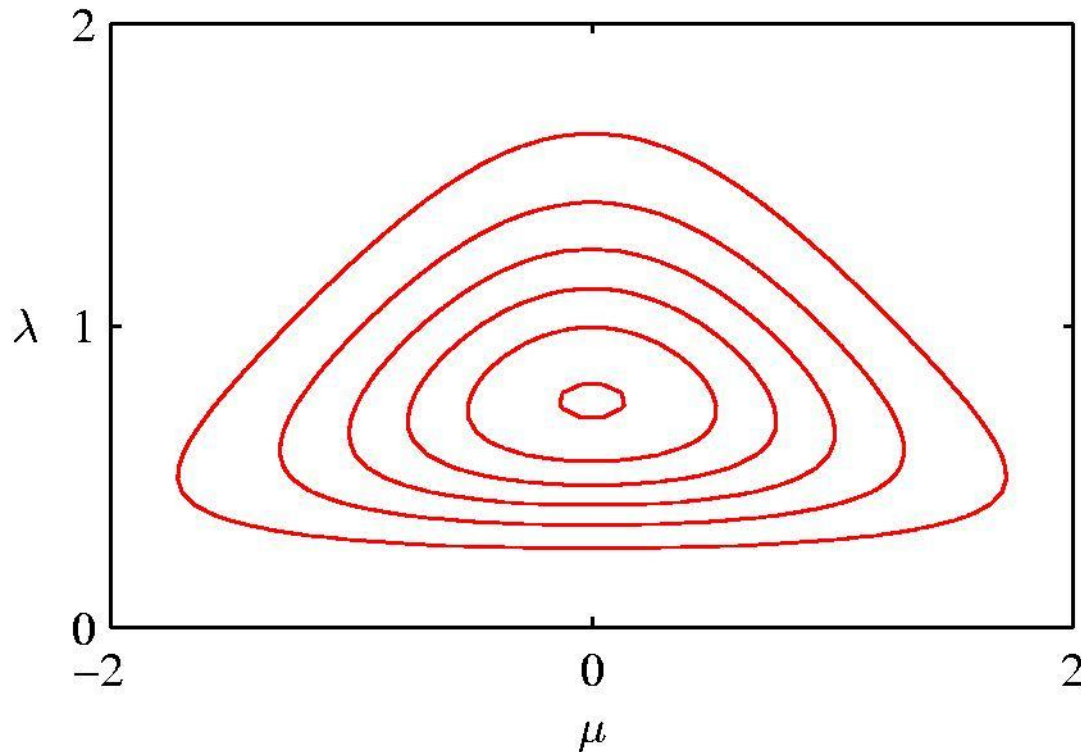
$$p(\mu, \lambda) = \mathcal{N}(\mu | \mu_0, (\beta\lambda)^{-1}) \text{Gam}(\lambda | a, b)$$
$$\propto \underbrace{\exp\left\{-\frac{\beta\lambda}{2}(\mu - \mu_0)^2\right\}}_{\text{Quadratic in } \mu} \underbrace{\lambda^{a-1} \exp\{-b\lambda\}}_{\text{Gamma distribution over } \lambda}$$

- Quadratic in  $\mu$ .
  - Linear in  $\lambda$ .
  - Gamma distribution over  $\lambda$ .
  - Independent of  $\mu$ .
-

# Bayesian Inference for the Gaussian (11)

---

## The Gaussian-gamma distribution





# Bayesian Inference for the Gaussian (12)

---

Multivariate conjugate priors

- $\boldsymbol{\mu}$  unknown,  $\boldsymbol{\Lambda}$  known:  $p(\boldsymbol{\mu})$  Gaussian.
- $\boldsymbol{\Lambda}$  unknown,  $\boldsymbol{\mu}$  known:  $p(\boldsymbol{\Lambda})$  Wishart,

$$\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}, \nu) = B|\boldsymbol{\Lambda}|^{(\nu-D-1)/2} \exp\left(-\frac{1}{2}\text{Tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right).$$

- $\boldsymbol{\Lambda}$  and  $\boldsymbol{\mu}$  unknown:  $p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$  Gaussian-Wishart,  $p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{\mu}_0, \beta, \mathbf{W}, \nu) =$

$$\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (\beta\boldsymbol{\Lambda})^{-1}) \mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}, \nu)$$

---

# Student's t-Distribution

---

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

where

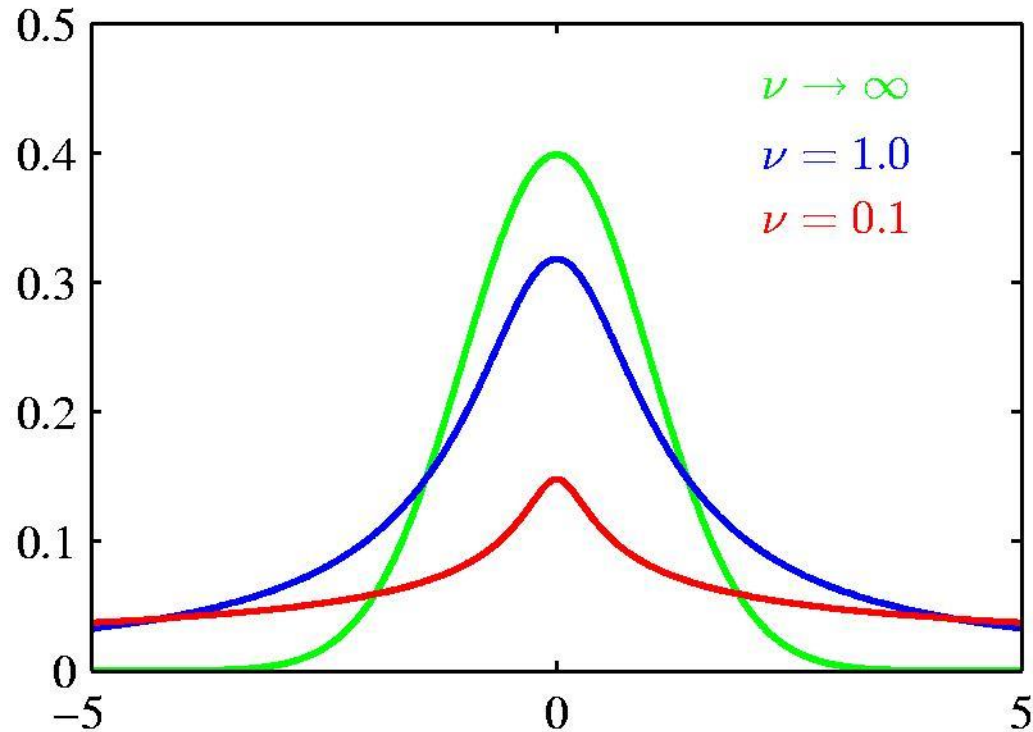
$$\lambda = a/b \quad \eta = \tau b/a \quad \nu = 2a.$$

Infinite mixture of Gaussians.

---

# Student's t-Distribution

---



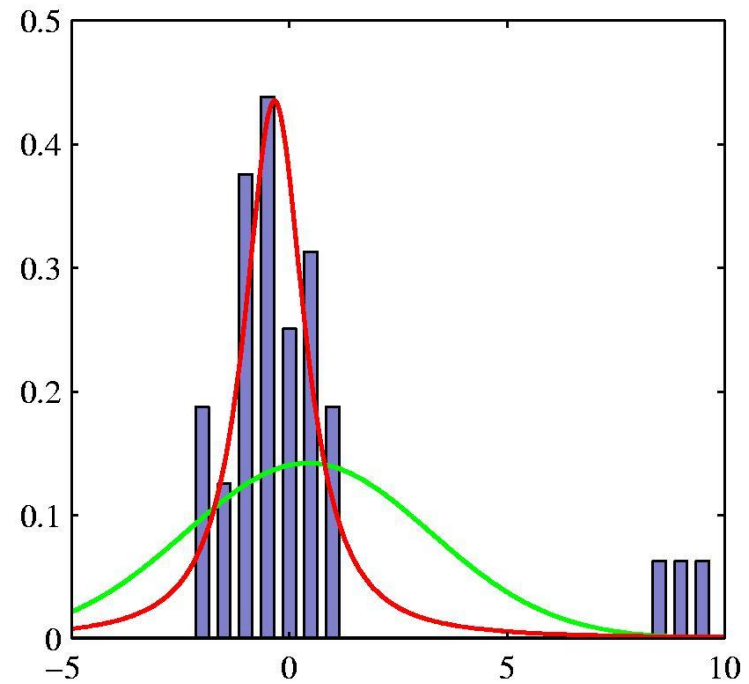
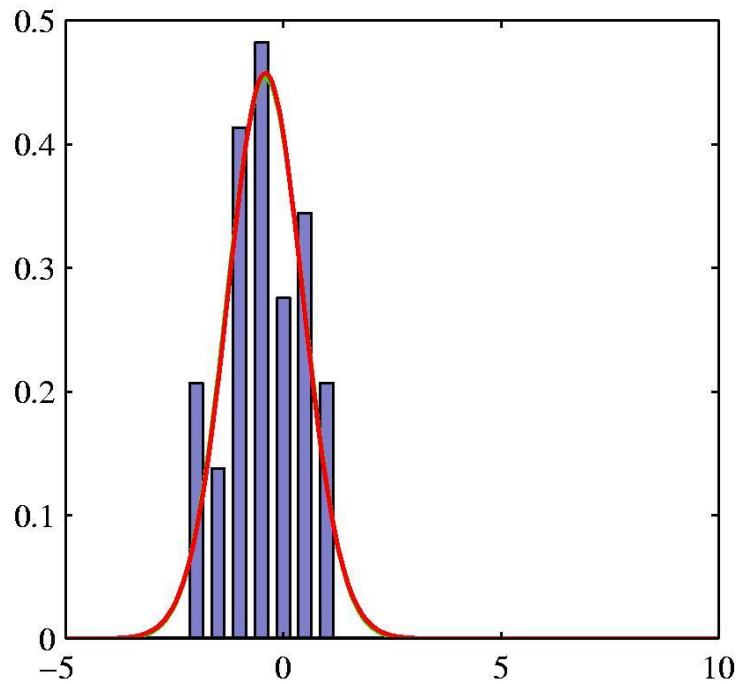
	$\nu = 1$	$\nu \rightarrow \infty$
$\text{St}(x \mu, \lambda, \nu)$	Cauchy	$\mathcal{N}(x \mu, \lambda^{-1})$

---

# Student's t-Distribution

---

Robustness to outliers: **Gaussian** vs **t-distribution**.



# Student's t-Distribution

---

The  $D$ -variate case:

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

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

Properties:

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

---

# Periodic variables

---

- Examples: calendar time, direction, ...
- We require

$$\begin{aligned}p(\theta) &\geq 0 \\ \int_0^{2\pi} p(\theta) \, d\theta &= 1 \\ p(\theta + 2\pi) &= p(\theta).\end{aligned}$$

# von Mises Distribution (1)

---

This requirement is satisfied by

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp \{m \cos(\theta - \theta_0)\}$$

where

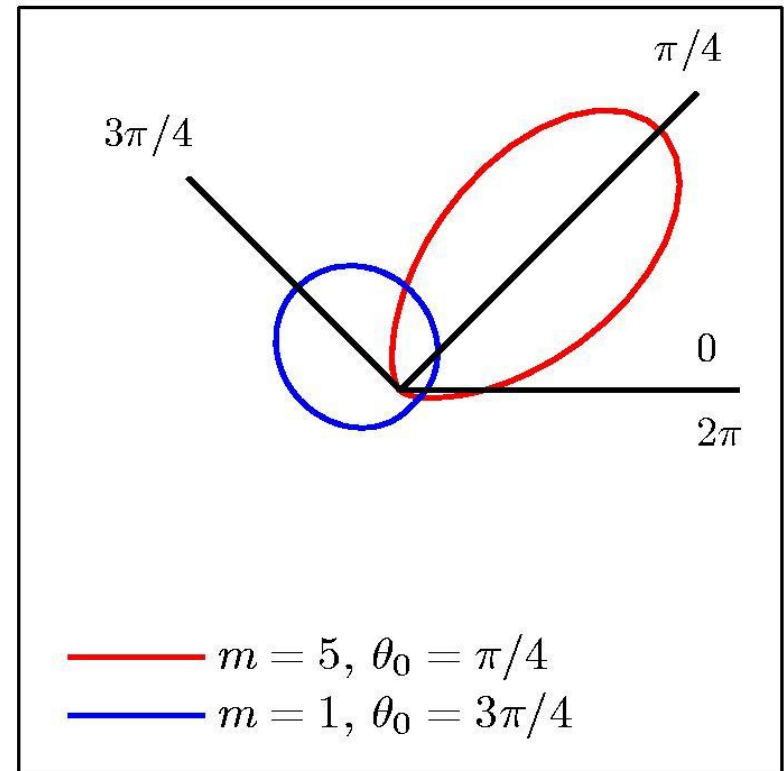
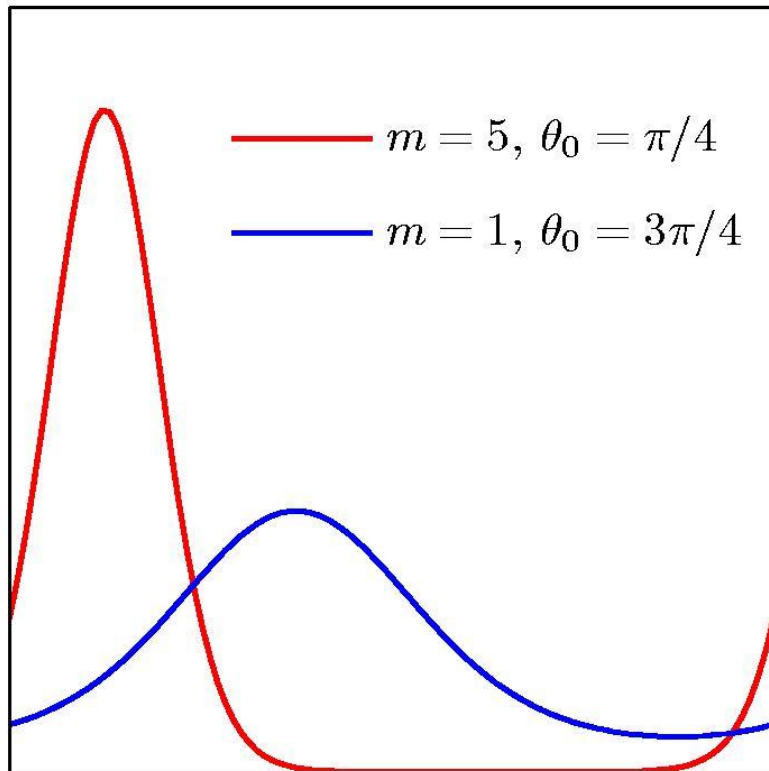
$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp \{m \cos \theta\} d\theta$$

is the 0<sup>th</sup> order modified Bessel function of the 1<sup>st</sup> kind.

---

# von Mises Distribution (4)

---





# Maximum Likelihood for von Mises

---

Given a data set,  $\mathcal{D} = \{\theta_1, \dots, \theta_N\}$ , the log likelihood function is given by

$$\ln p(\mathcal{D}|\theta_0, m) = -N \ln(2\pi) - N \ln I_0(m) + m \sum_{n=1}^N \cos(\theta_n - \theta_0).$$

Maximizing with respect to  $\theta_0$  we directly obtain

$$\theta_0^{\text{ML}} = \tan^{-1} \left\{ \frac{\sum_n \sin \theta_n}{\sum_n \cos \theta_n} \right\}.$$

Similarly, maximizing with respect to  $m$  we get

$$\frac{I_1(m_{\text{ML}})}{I_0(m_{\text{ML}})} = \frac{1}{N} \sum_{n=1}^N \cos(\theta_n - \theta_0^{\text{ML}})$$

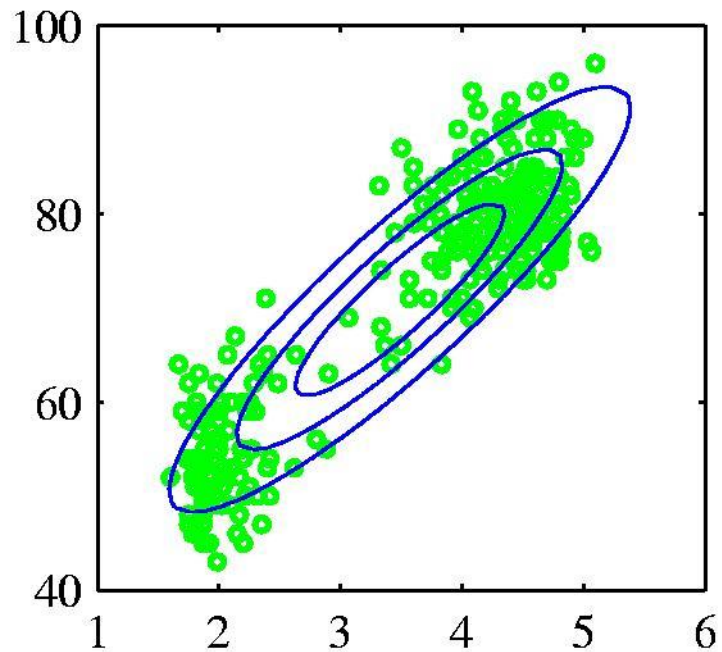
which can be solved numerically for  $m_{\text{ML}}$ .

---

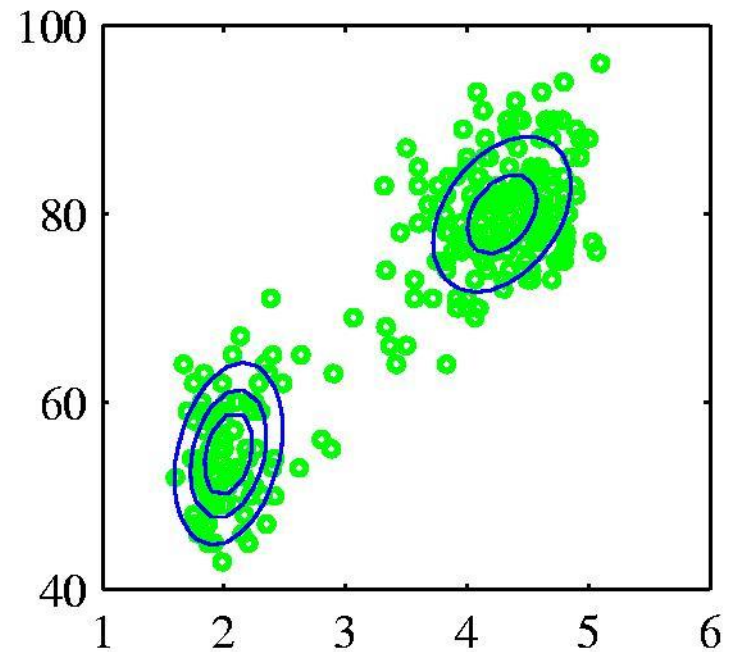
# Mixtures of Gaussians (1)

---

Old Faithful data set



Single Gaussian



Mixture of two Gaussians

---

# Mixtures of Gaussians (2)

---

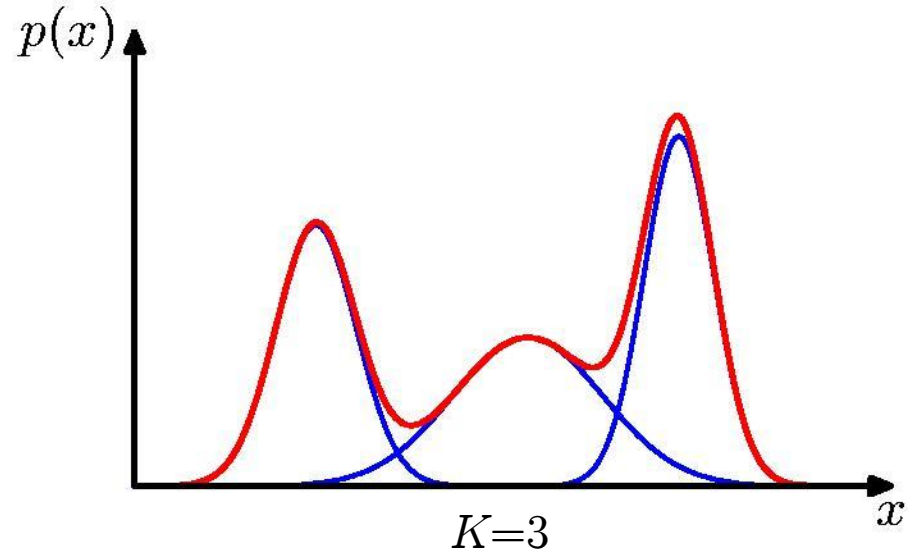
Combine simple models  
into a complex model:

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

↑  
Mixing coefficient

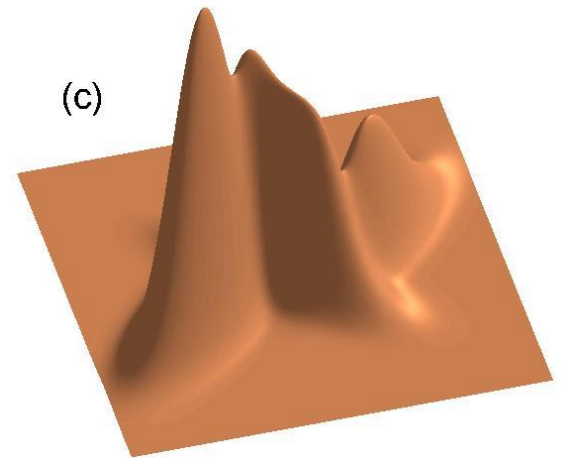
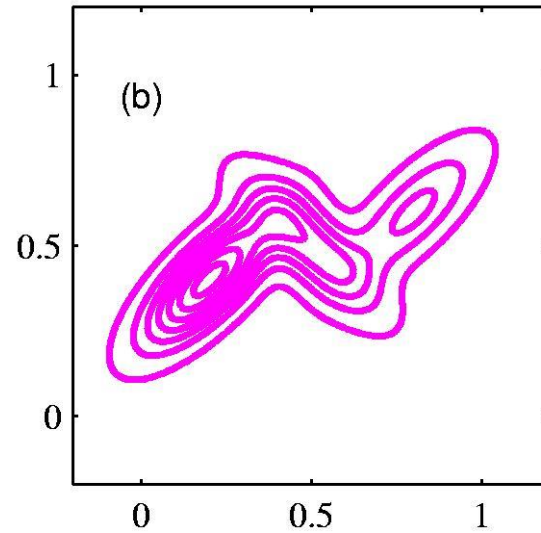
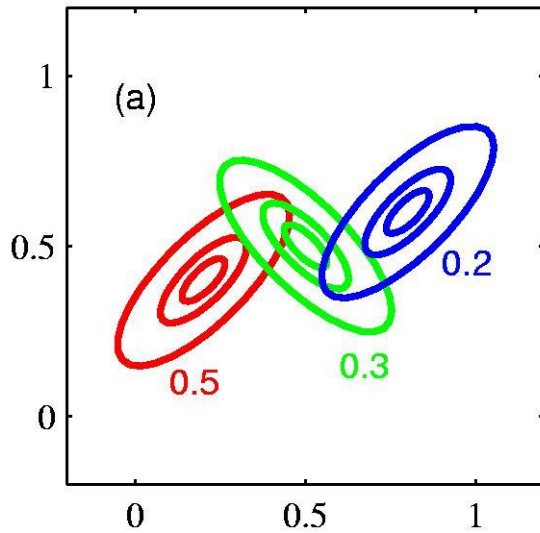
└──────────┘  
Component

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



# Mixtures of Gaussians (3)

---



# Mixtures of Gaussians (4)

---

Determining parameters  $\mu$ ,  $\Sigma$ , and  $\pi$  using maximum log likelihood

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

Log of a sum; no closed form maximum.

Solution: use standard, iterative, numeric optimization methods or the *expectation maximization* algorithm (Chapter 9).

---

# The Exponential Family (1)

---

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

where  $\boldsymbol{\eta}$  is the *natural parameter* and

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

so  $g(\boldsymbol{\eta})$  can be interpreted as a normalization coefficient.

---

# The Exponential Family (2.1)

---

## The Bernoulli Distribution

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

Comparing with the general form we see that

$$\eta = \ln \left( \frac{\mu}{1 - \mu} \right) \quad \text{and so} \quad \mu = \underbrace{\sigma(\eta)}_{\text{Logistic sigmoid}} = \frac{1}{1 + \exp(-\eta)}.$$

# The Exponential Family (2.2)

---

The Bernoulli distribution can hence be written as

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

where

$$u(x) = x$$

$$h(x) = 1$$

$$g(\eta) = 1 - \sigma(\eta) = \sigma(-\eta).$$

---



# The Exponential Family (3.1)

---

## The Multinomial Distribution

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

where,  $\mathbf{x} = (x_1, \dots, x_M)^T$ ,  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_M)^T$  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  $\eta_k$  parameters are not independent since the corresponding  $\mu_k$  must satisfy

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

# The Exponential Family (3.2)

---

Let  $\mu_M = 1 - \sum_{k=1}^{M-1} \mu_k$ . This leads to

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

Here the  $\eta_k$  parameters are independent. Note that

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

---

# The Exponential Family (3.3)

---

The Multinomial distribution can then be written as

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

where

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

---

# The Exponential Family (4)

---

## The Gaussian Distribution

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

where

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

---

# ML for the Exponential Family (1)

---

From the definition of  $g(\boldsymbol{\eta})$  we get

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

Thus

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

---

# ML for the Exponential Family (2)

---

Give a data set,  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ , the likelihood function is given by

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

Thus we have

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

---

Sufficient statistic

# Conjugate priors

---

For any member of the exponential family,  
there exists a prior

$$p(\boldsymbol{\eta}|\boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu)g(\boldsymbol{\eta})^\nu \exp \{ \nu \boldsymbol{\eta}^\text{T} \boldsymbol{\chi} \} .$$

Combining with the likelihood function, we get

$$p(\boldsymbol{\eta}|\mathbf{X}, \boldsymbol{\chi}, \nu) \propto g(\boldsymbol{\eta})^{\nu+N} \exp \left\{ \boldsymbol{\eta}^\text{T} \left( \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n) + \nu \boldsymbol{\chi} \right) \right\} .$$

Prior corresponds to  $\nu$  pseudo-observations with value  $\boldsymbol{\chi}$ .

---

# Noninformative Priors (1)

---

With little or no information available a-priori, we might choose a non-informative prior.

- $\lambda$  discrete,  $K$ -nomial :  $p(\lambda) = 1/K$ .
- $\lambda \in [a, b]$  real and bounded:  $p(\lambda) = 1/b - a$ .
- $\lambda$  real and unbounded: **improper!**

A constant prior may no longer be constant after a change of variable; consider  $p(\lambda)$  constant and  $\lambda = \eta^2$ :

$$p_{\eta}(\eta) = p_{\lambda}(\lambda) \left| \frac{d\lambda}{d\eta} \right| = p_{\lambda}(\eta^2) 2\eta \propto \eta$$

---



# Noninformative Priors (2)

---

Translation invariant priors. Consider

$$p(x|\mu) = f(x - \mu) = f((x + c) - (\mu + c)) = f(\hat{x} - \hat{\mu}) = p(\hat{x}|\hat{\mu}).$$

For a corresponding prior over  $\mu$ , we have

$$\int_A^B p(\mu) d\mu = \int_{A-c}^{B-c} p(\mu) d\mu = \int_A^B p(\mu - c) d\mu$$

for any  $A$  and  $B$ . Thus  $p(\mu) = p(\mu - c)$  and  $p(\mu)$  must be constant.

---

# Noninformative Priors (3)

---

Example: The mean of a Gaussian,  $\mu$ ; the conjugate prior is also a Gaussian,

$$p(\mu|\mu_0, \sigma_0^2) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$$

As  $\sigma_0^2 \rightarrow \infty$ , this will become constant over  $\mu$ .

---

# Noninformative Priors (4)

---

Scale invariant priors. Consider  $p(x|\sigma) = (1/\sigma)f(x/\sigma)$  and make the change of variable  $\hat{x} = cx$

$$p_{\hat{x}}(\hat{x}) = p_x(x) \left| \frac{dx}{d\hat{x}} \right| = p_x\left(\frac{\hat{x}}{c}\right) \frac{1}{c} = \frac{1}{c\sigma} f\left(\frac{\hat{x}}{c\sigma}\right) = p_x(\hat{x}|\hat{\sigma}).$$

For a corresponding prior over  $\sigma$ , we have

$$\int_A^B p(\sigma) d\sigma = \int_{A/c}^{B/c} p(\sigma) d\sigma = \int_A^B p\left(\frac{1}{c}\sigma\right) \frac{1}{c} d\sigma$$

for any  $A$  and  $B$ . Thus  $p(\sigma) \propto 1/\sigma$  and so this prior is improper too. Note that this corresponds to  $p(\ln \sigma)$  being constant.

---

# Noninformative Priors (5)

---

Example: For the variance of a Gaussian,  $\sigma^2$ , we have

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

If  $\lambda = 1/\sigma^2$  and  $p(\sigma) \propto 1/\sigma$ , then  $p(\lambda) \propto 1/\lambda$ .

We know that the conjugate distribution for  $\lambda$  is the Gamma distribution,

$$\text{Gam}(\lambda|a_0, b_0) \propto \lambda^{a_0-1} \exp(-b_0\lambda).$$

A noninformative prior is obtained when  $a_0 = 0$  and  $b_0 = 0$ .

---

# Nonparametric Methods (1)

---

Parametric distribution models are restricted to specific forms, which may not always be suitable; for example, consider modelling a multimodal distribution with a single, unimodal model.

Nonparametric approaches make few assumptions about the overall shape of the distribution being modelled.

---

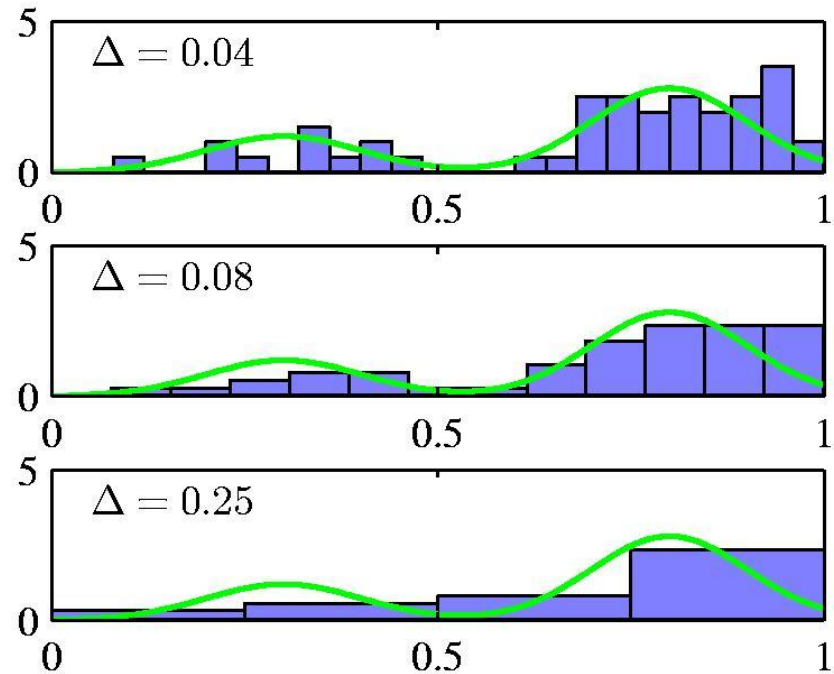
# Nonparametric Methods (2)

---

**Histogram methods** partition the data space into distinct bins with widths  $\Delta_i$  and count the number of observations,  $n_i$ , in each bin.

$$p_i = \frac{n_i}{N \Delta_i}$$

- Often, the same width is used for all bins,  $\Delta_i = \Delta$ .
- $\Delta$  acts as a smoothing parameter.



- In a  $D$ -dimensional space, using  $M$  bins in each dimension will require  $M^D$  bins!
-

# Nonparametric Methods (3)

---

Assume observations drawn from a density  $p(\mathbf{x})$  and consider a small region  $\mathcal{R}$  containing  $\mathbf{x}$  such that

$$P = \int_{\mathcal{R}} p(\mathbf{x}) d\mathbf{x}.$$

The probability that  $K$  out of  $N$  observations lie inside  $\mathcal{R}$  is  $\text{Bin}(K|N, P)$  and if  $N$  is large

$$K \simeq NP.$$

If the volume of  $\mathcal{R}$ ,  $V$ , is sufficiently small,  $p(\mathbf{x})$  is approximately constant over  $\mathcal{R}$  and

$$P \simeq p(\mathbf{x})V$$

Thus

$$p(\mathbf{x}) = \frac{K}{NV}.$$

$V$  small, yet  $K > 0$ , therefore  $N$  large?

# Nonparametric Methods (4)

---

**Kernel Density Estimation:** fix  $V$ , estimate  $K$  from the data. Let  $\mathcal{R}$  be a hypercube centred on  $\mathbf{x}$  and define the kernel function (Parzen window)

$$k((\mathbf{x} - \mathbf{x}_n)/h) = \begin{cases} 1, & |(x_i - x_{ni})/h| \leq 1/2, \quad i = 1, \dots, D, \\ 0, & \text{otherwise.} \end{cases}$$

It follows that

$$K = \sum_{n=1}^N k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right) \quad \text{and hence} \quad p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N \frac{1}{h^D} k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right).$$

---



# Nonparametric Methods (5)

---

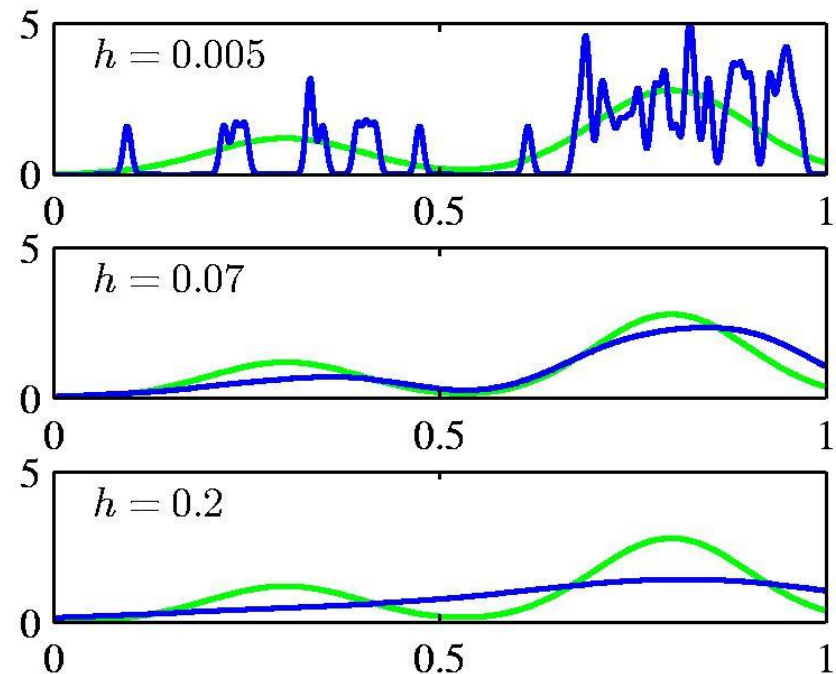
To avoid discontinuities in  $p(\mathbf{x})$ ,  
use a smooth kernel, e.g. a  
Gaussian

$$p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N \frac{1}{(2\pi h^2)^{D/2}} \exp \left\{ -\frac{\|\mathbf{x} - \mathbf{x}_n\|^2}{2h^2} \right\}$$

Any kernel such that

$$\begin{aligned} k(\mathbf{u}) &\geq 0, \\ \int k(\mathbf{u}) \, d\mathbf{u} &= 1 \end{aligned}$$

will work.



$h$  acts as a smoother.

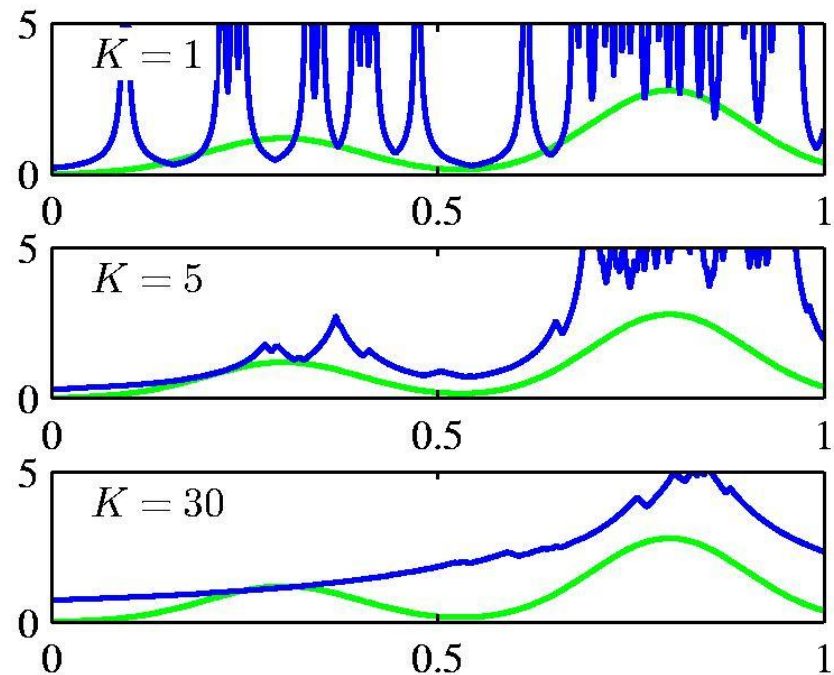
# Nonparametric Methods (6)

---

## Nearest Neighbour

**Density Estimation:** fix  $K$ , estimate  $V$  from the data. Consider a hypersphere centred on  $\mathbf{x}$  and let it grow to a volume,  $V^*$ , that includes  $K$  of the given  $N$  data points. Then

$$p(\mathbf{x}) \simeq \frac{K}{NV^*}.$$



$K$  acts as a smoother.

# Nonparametric Methods (7)

---

Nonparametric models (not histograms) requires storing and computing with the entire data set.

Parametric models, once fitted, are much more efficient in terms of storage and computation.

---

# $K$ -Nearest-Neighbours for Classification (1)

---

Given a data set with  $N_k$  data points from class  $\mathcal{C}_k$   
and  $\sum_k N_k = N$ , we have

$$p(\mathbf{x}) = \frac{K}{NV}$$

and correspondingly

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{K_k}{N_k V}.$$

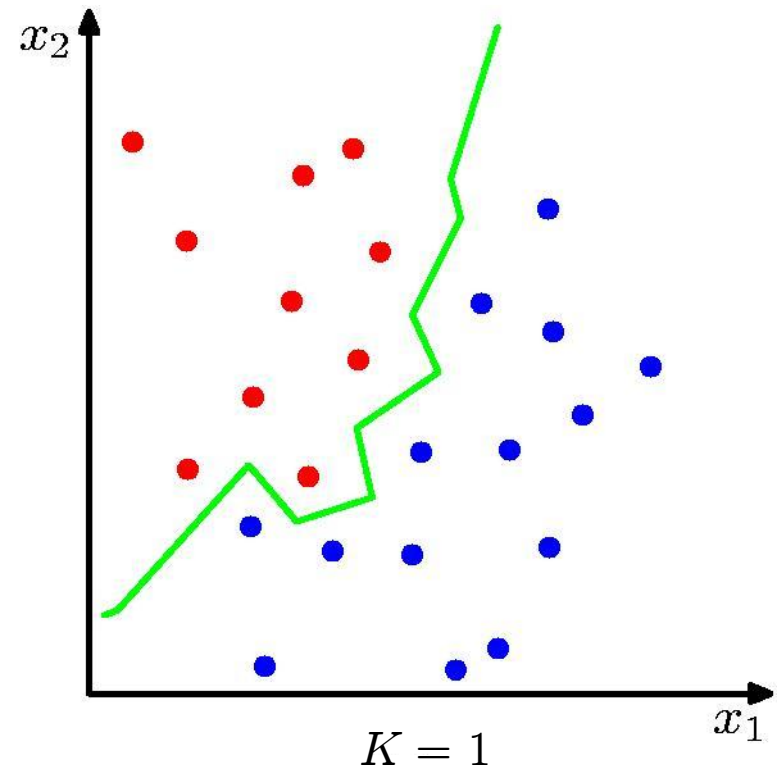
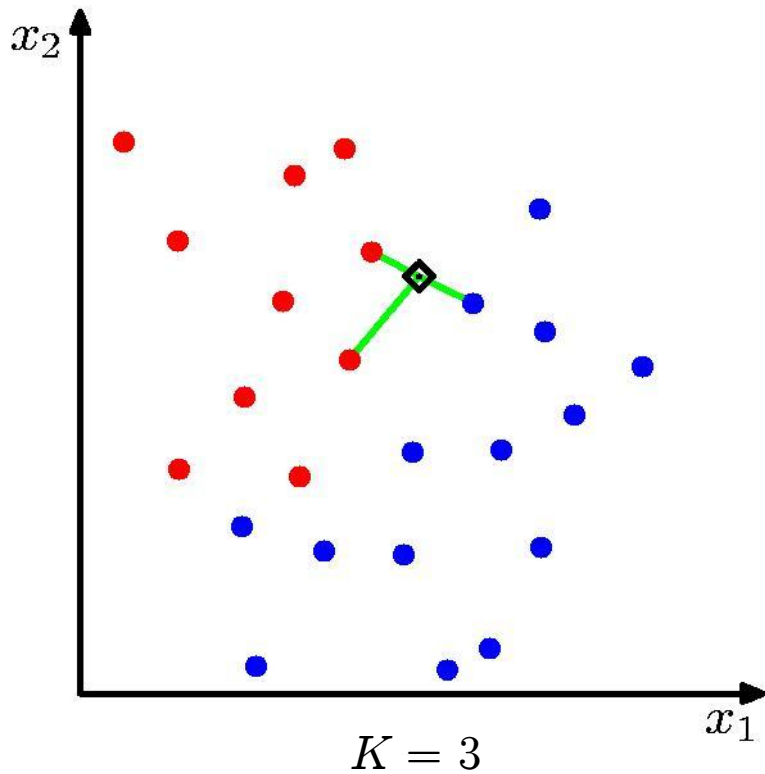
Since  $p(\mathcal{C}_k) = N_k/N$ , Bayes' theorem gives

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} = \frac{K_k}{K}.$$

---

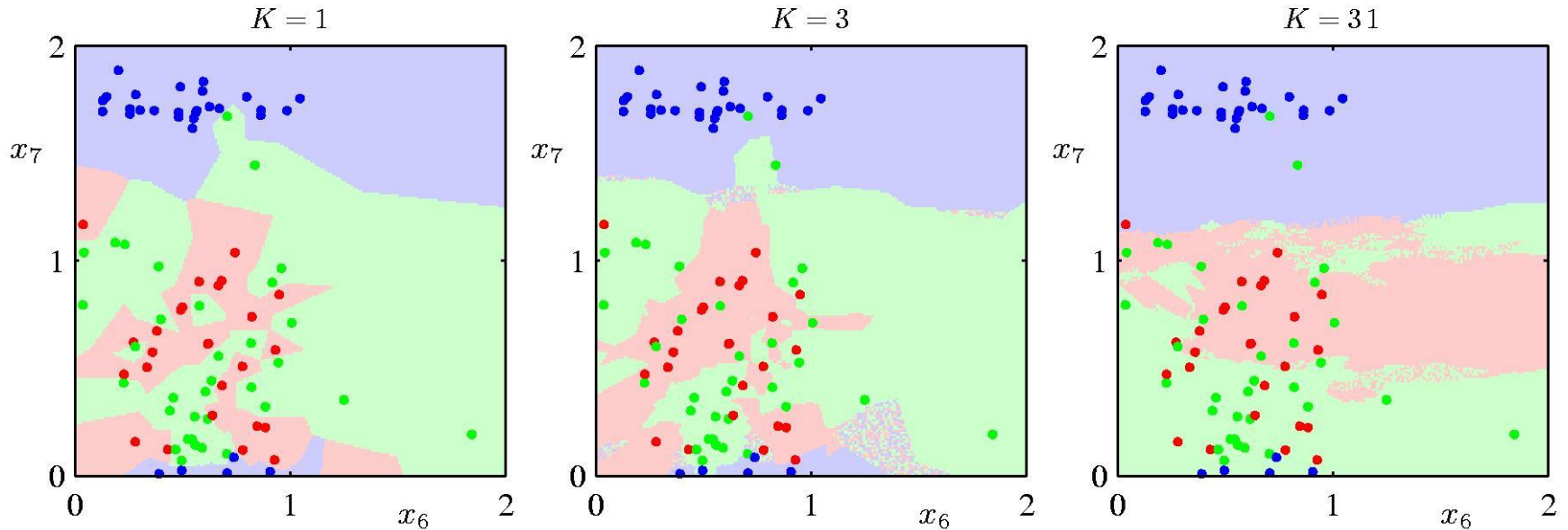
# $K$ -Nearest-Neighbours for Classification (2)

---



# $K$ -Nearest-Neighbours for Classification (3)

---



- $K$  acts as a smother
  - For  $N \rightarrow \infty$ , the error rate of the 1-nearest-neighbour classifier is never more than twice the optimal error (obtained from the true conditional class distributions).
-