

Probability & Statistics



Seyed Abbas Hosseini
Sharif University of Technology

All slides are adopted from PRML book

Outline

- ❑ **Bayesian Inference, Conjugate Prior and Parameter Estimation**
 - ❑ **Bernoulli and beta distributions**
 - ❑ **Multinomial and Dirichlet distributions**
- ❑ **Gaussian Distribution**
 - ❑ **univariate and Multivariate Gaussian distribution**
 - ❑ **Marginal and Conditional Gaussian Distribution**
 - ❑ **ML Estimation of Gaussian parameters**
 - ❑ **Bayesian inference for the Gaussian**

Bayesian Inference, Conjugate Prior, Parameter Estimation

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:

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

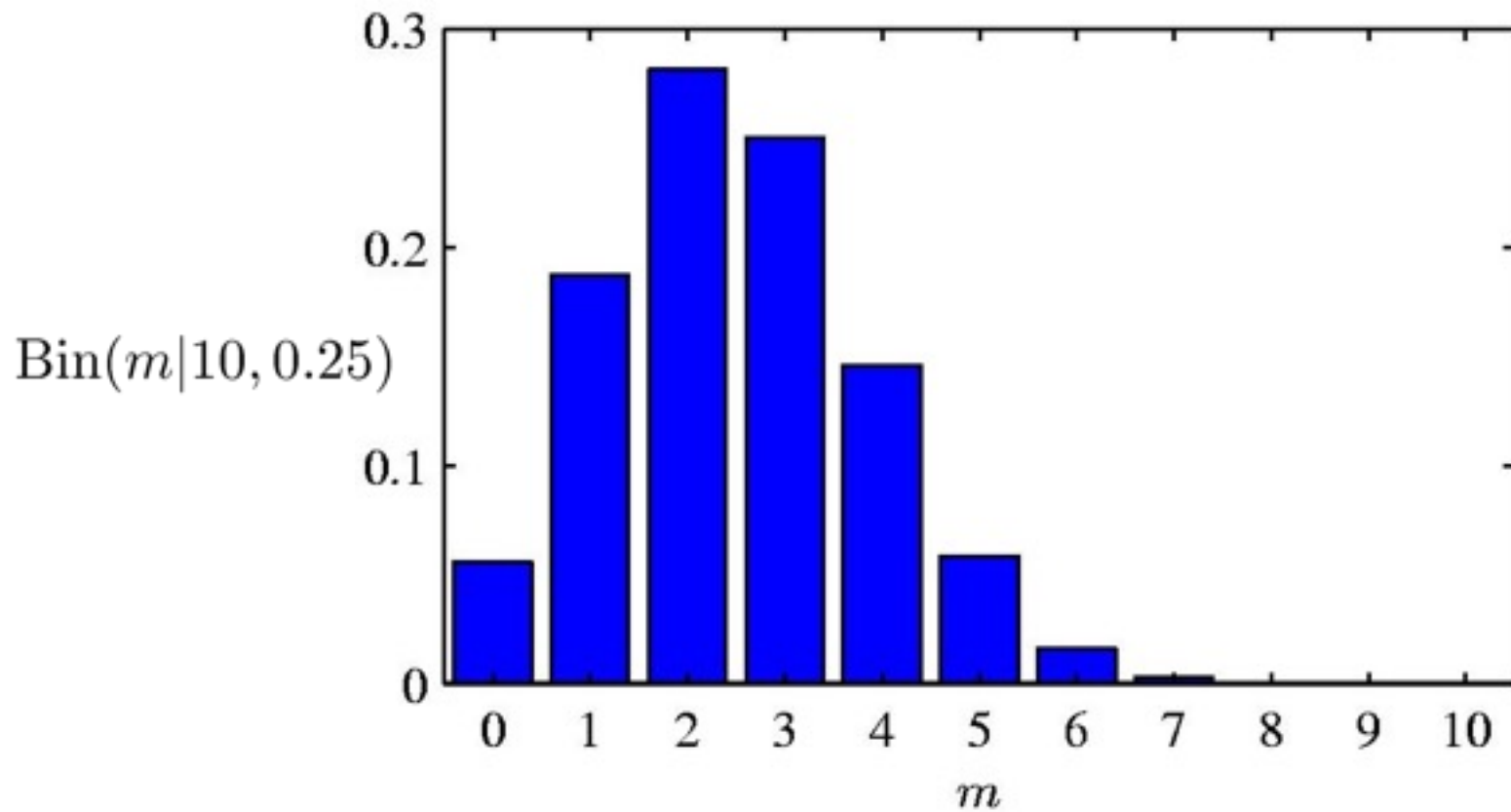
- Binomial Distribution

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

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

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

Binomial Distribution



Parameter Estimation (1)

- ML for Bernoulli

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

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

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

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

$$\mu \in [0, 1]$$

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

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

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

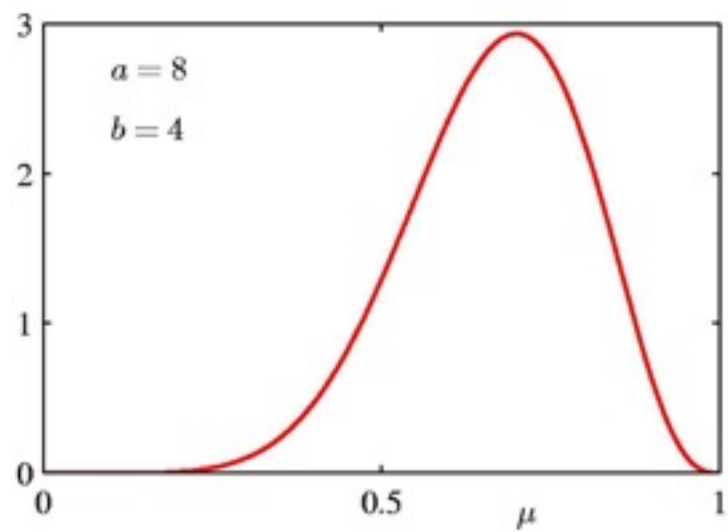
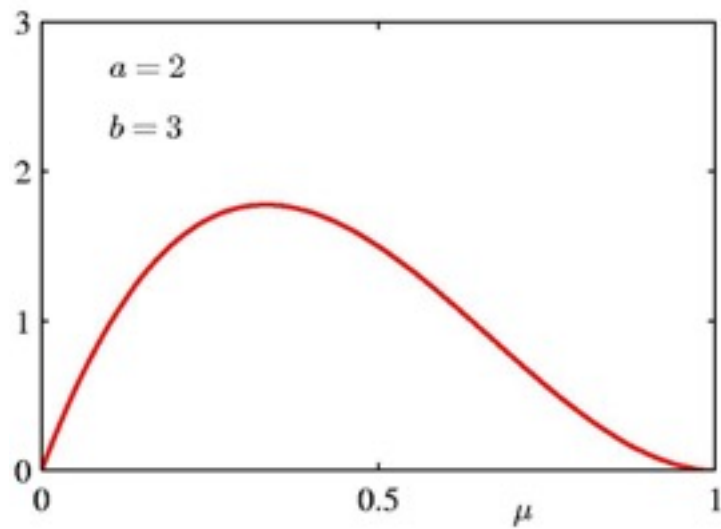
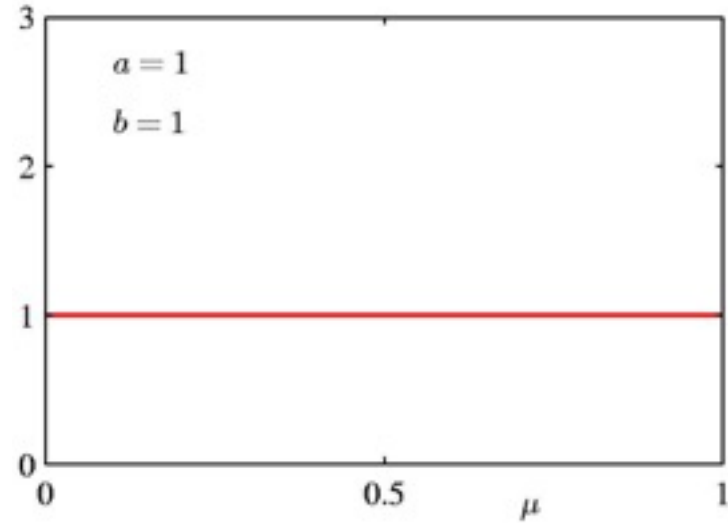
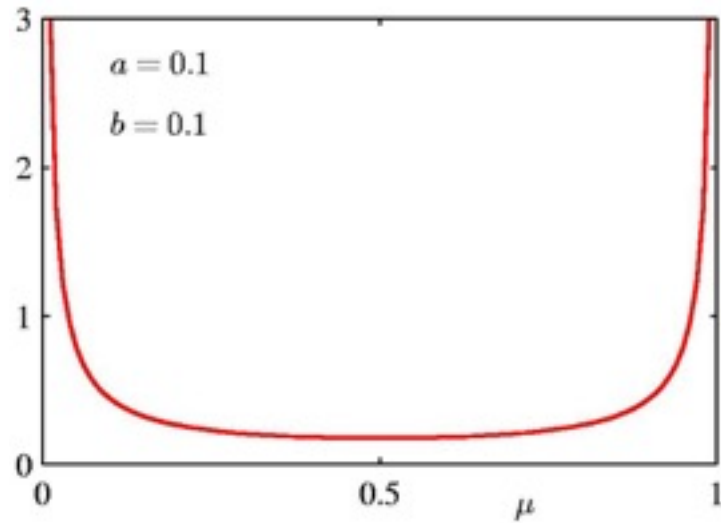
Bayesian Bernoulli

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

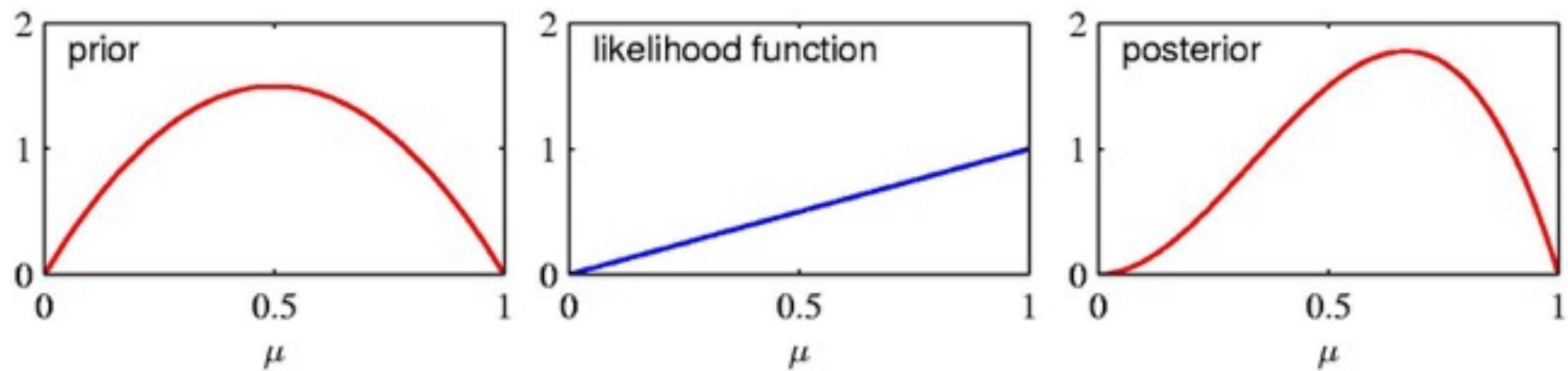
$$a_N = a_0 + m \quad b_N = b_0 + (N - m)$$

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, λ .

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

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

The Multinomial Distribution

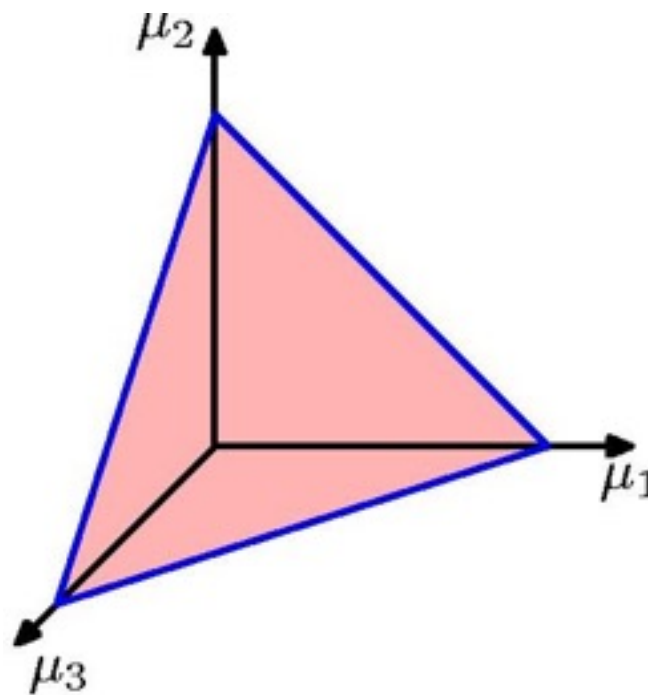
$$\begin{aligned}\text{Mult}(m_1, m_2, \dots, m_K | \boldsymbol{\mu}, N) &= \binom{N}{m_1 m_2 \dots m_K} \prod_{k=1}^K \mu_k^{m_k} \\ \mathbb{E}[m_k] &= N \mu_k \\ \text{var}[m_k] &= N \mu_k (1 - \mu_k) \\ \text{cov}[m_j, m_k] &= -N \mu_j \mu_k\end{aligned}$$

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.

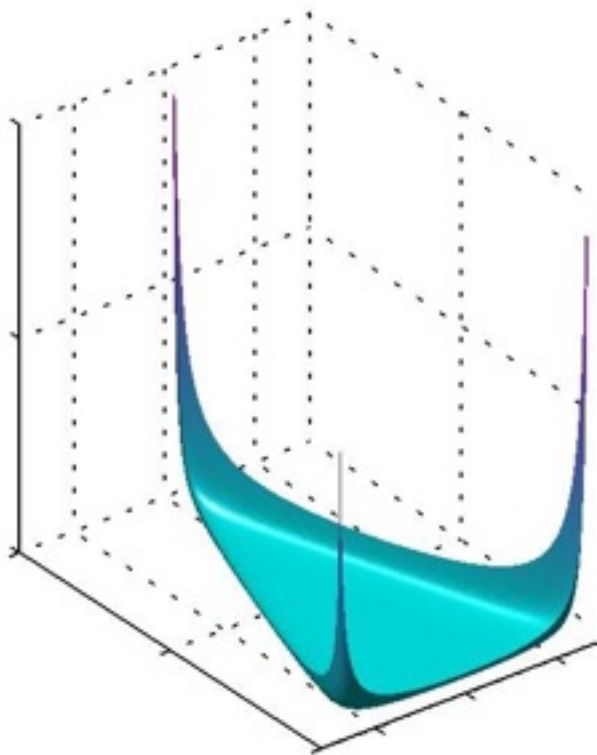


Bayesian Multinomial (1)

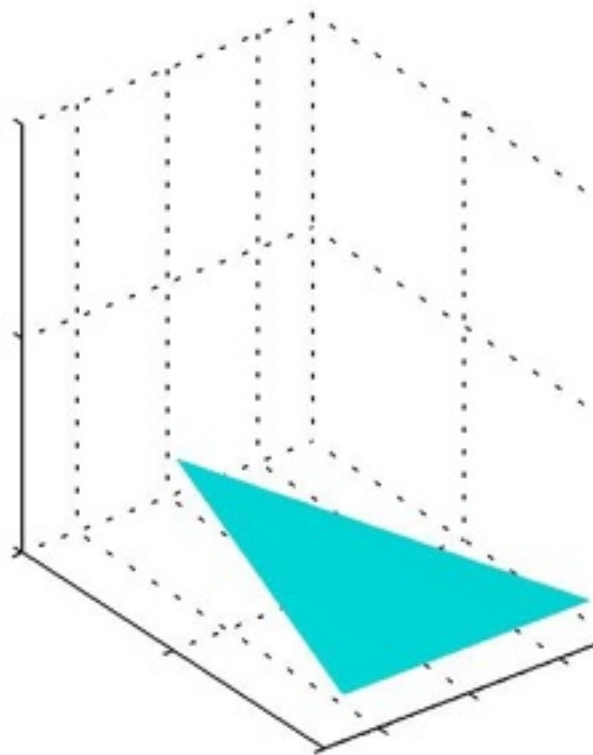
$$p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) \propto p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^K \mu_k^{\alpha_k + m_k - 1}$$

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

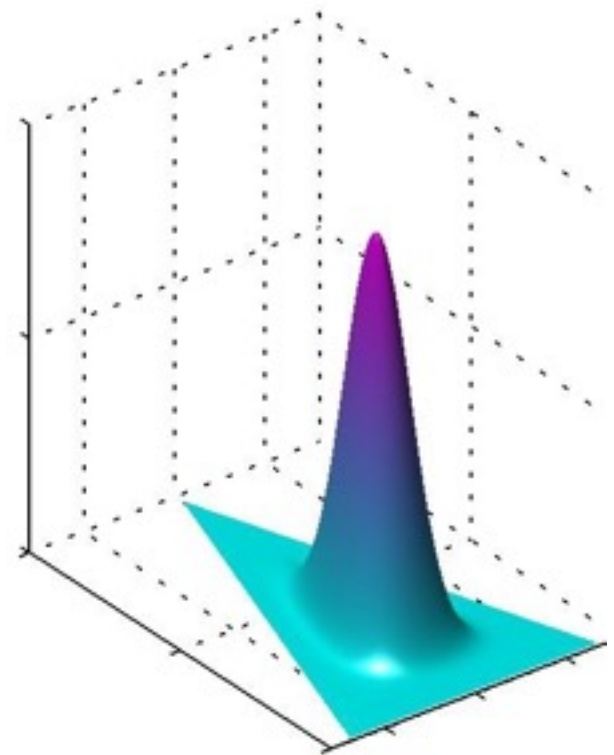
Bayesian Multinomial (2)



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



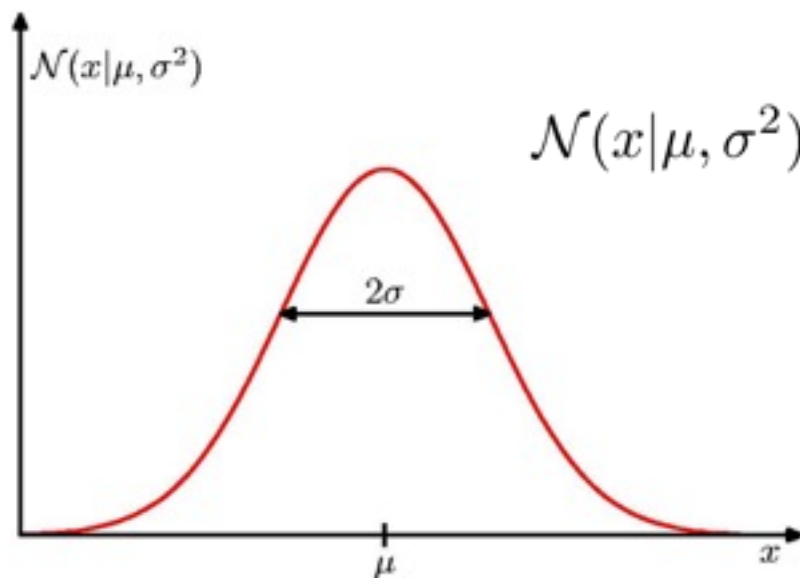
$$\alpha_k = 10^0$$



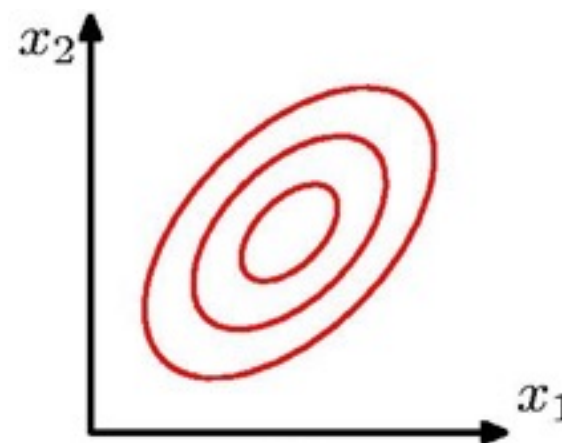
$$\alpha_k = 10^1$$

Gaussian Distribution

The Gaussian Distribution



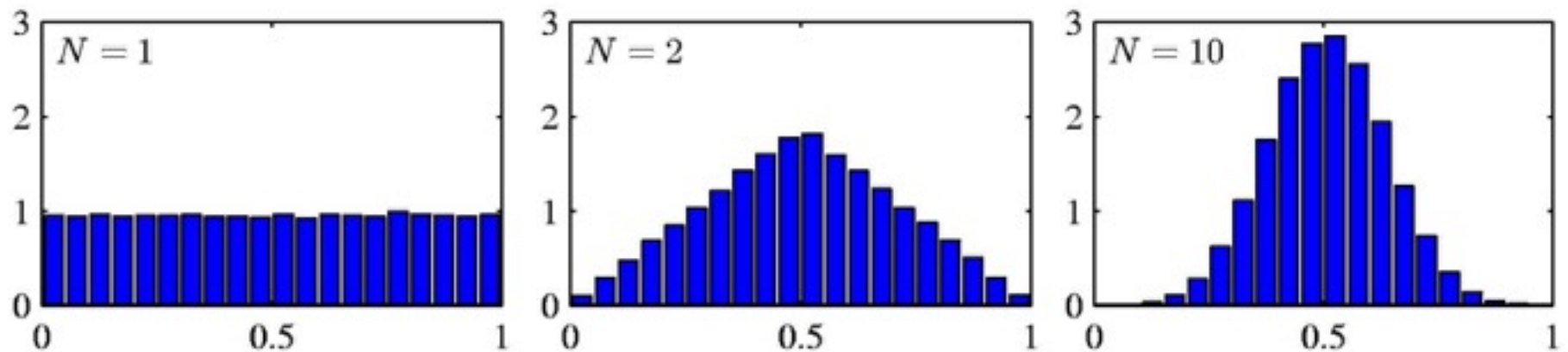
$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\}$$



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

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.



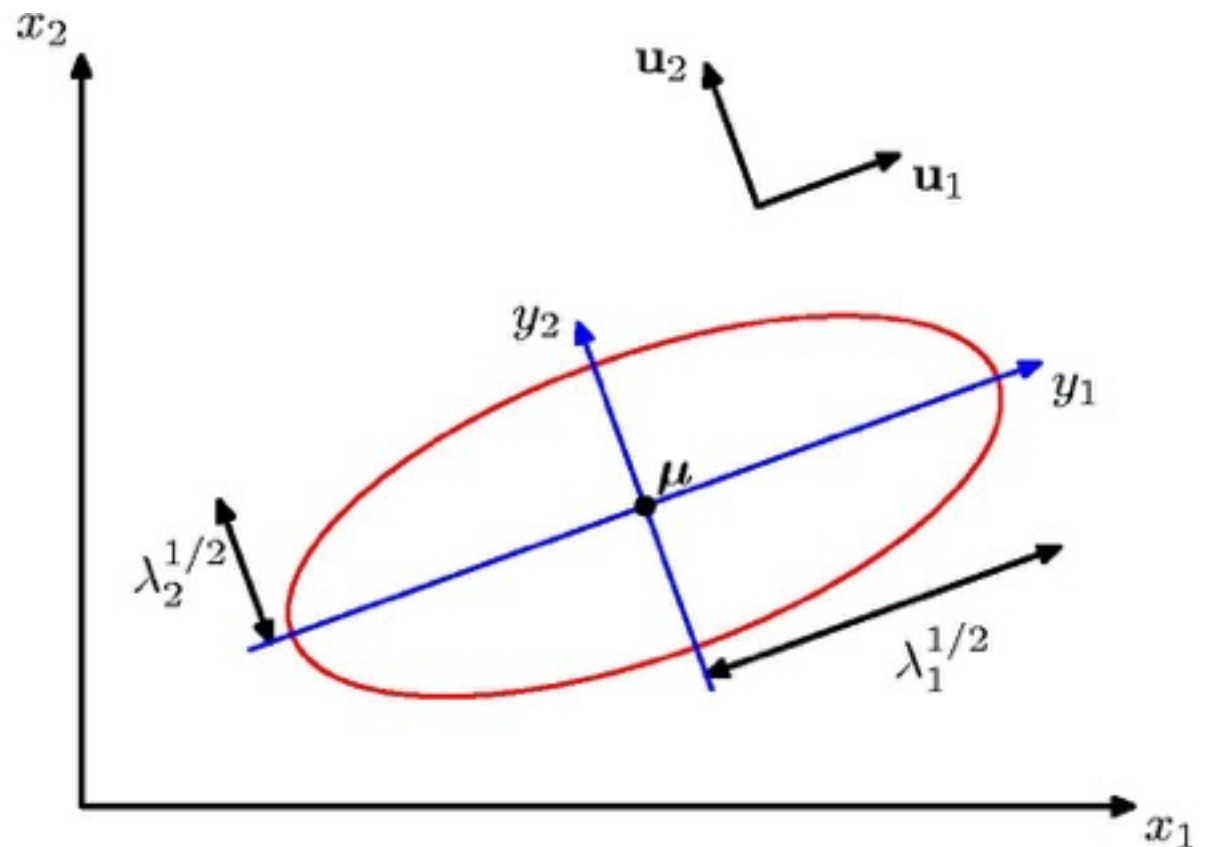
Geometry of the Multivariate Gaussian

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

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

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

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



Moments of the Multivariate Gaussian (1)

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

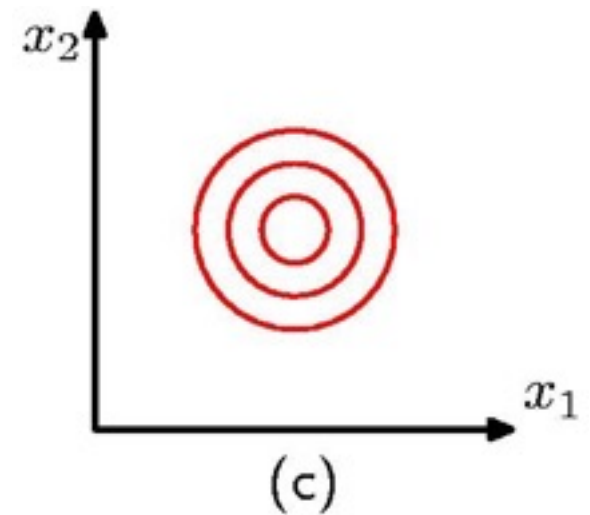
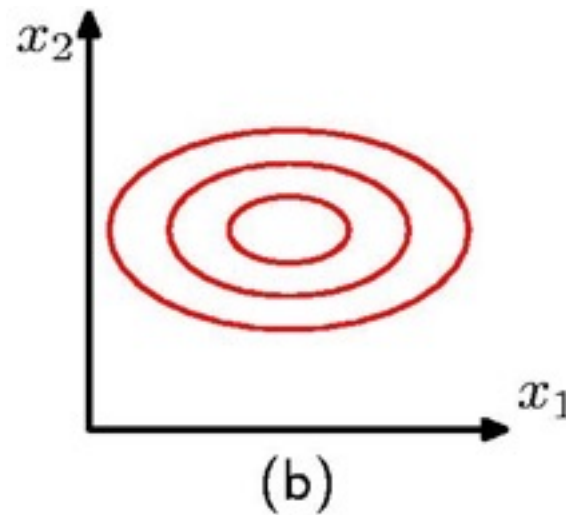
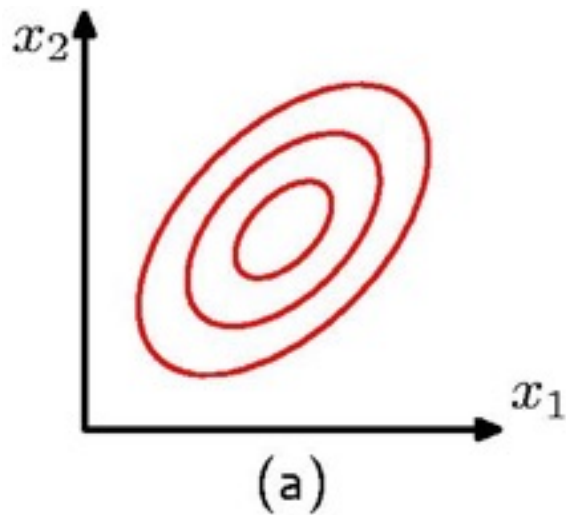
thanks to anti-symmetry of \mathbf{z}

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

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

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

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

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

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

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

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

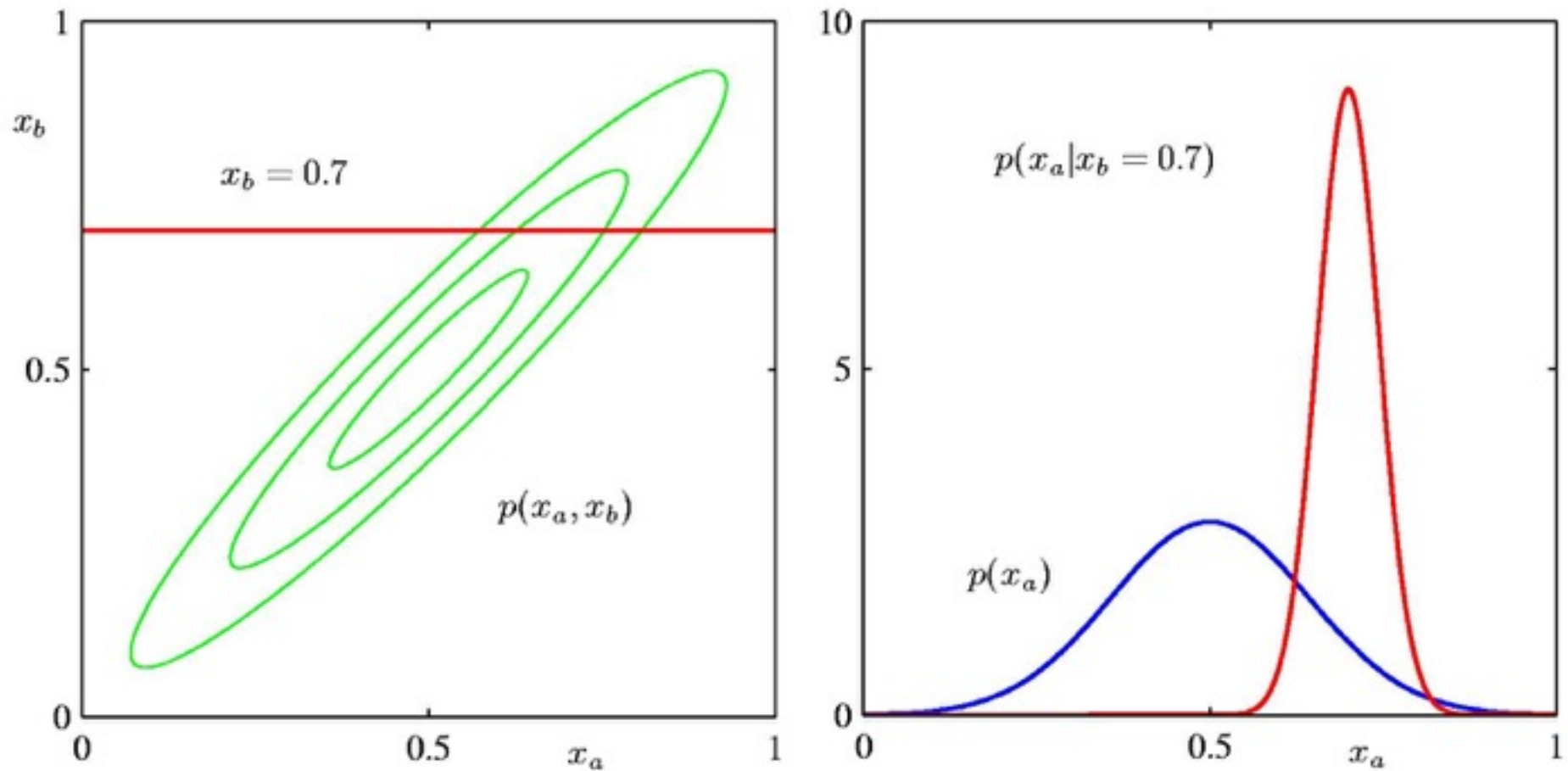
Partitioned Conditionals and Marginals

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

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

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

Partitioned Conditionals and Marginals



Bayes' Theorem for Gaussian Variables

- Given

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1})$$

- we have

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

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

- 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^{th} data point, \mathbf{x}_N

$$\begin{aligned}\mu_{\text{ML}}^{(N)} &= \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n \\ &= \frac{1}{N} \mathbf{x}_N + \frac{1}{N} \sum_{n=1}^{N-1} \mathbf{x}_n \\ &= \frac{1}{N} \mathbf{x}_N + \frac{N-1}{N} \mu_{\text{ML}}^{(N-1)} \\ &= \underbrace{\mu_{\text{ML}}^{(N-1)}}_{\text{old estimate}} + \underbrace{\frac{1}{N}}_{\text{correction weight}} \underbrace{(\mathbf{x}_N - \mu_{\text{ML}}^{(N-1)})}_{\text{correction given } \mathbf{x}_N}\end{aligned}$$

Bayesian Inference for the Gaussian (1)

- Assume variance is known. Given i.i.d. data $\mathbf{x} = \{x_1, \dots, x_N\}$, the likelihood function for mean 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 μ (but it is *not* a distribution over μ).

Bayesian Inference for the Gaussian (2)

- Combined with a Gaussian prior over μ ,

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

- this gives the posterior

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

- Completing the square over μ , we see that

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

Bayesian Inference for the Gaussian (3)

- ... where

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

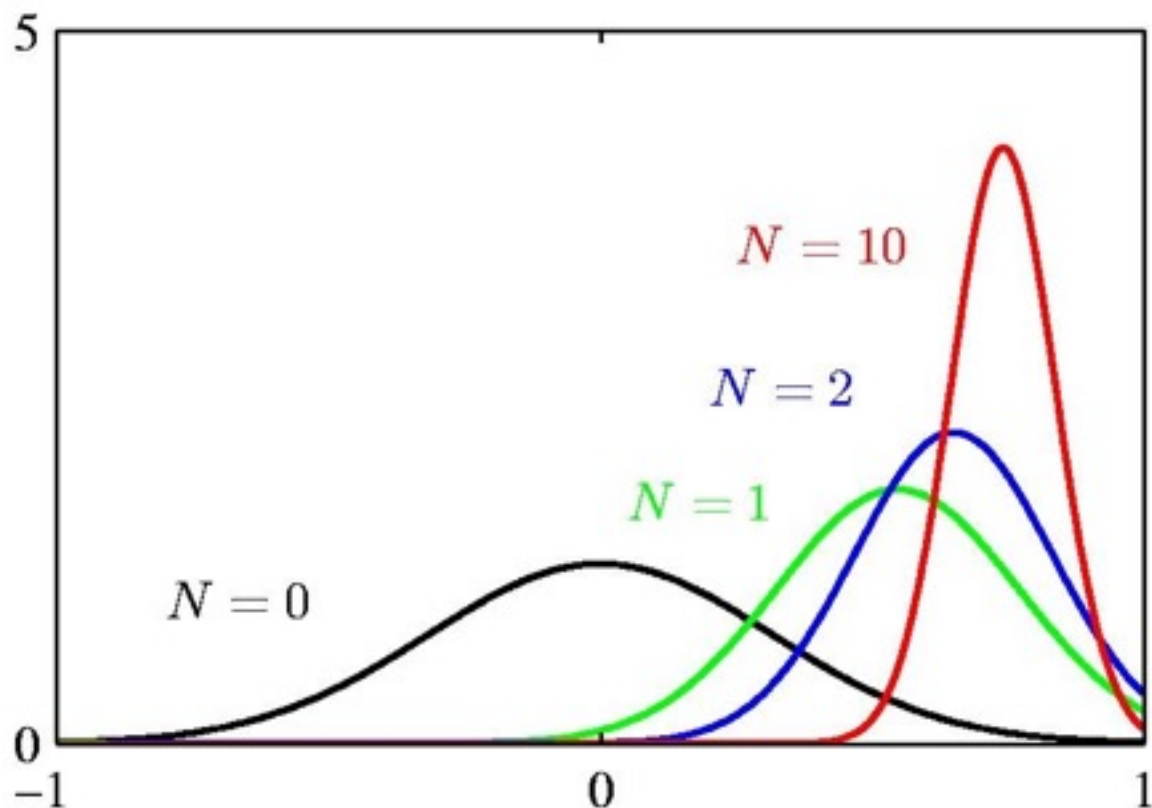
$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}.$$

- Note:

	$N = 0$	$N \rightarrow \infty$
μ_N	μ_0	μ_{ML}
σ_N^2	σ_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^{th} data point.

Bayesian Inference for the Gaussian (6)

- Now assume μ is known. The likelihood function for precision is given by

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

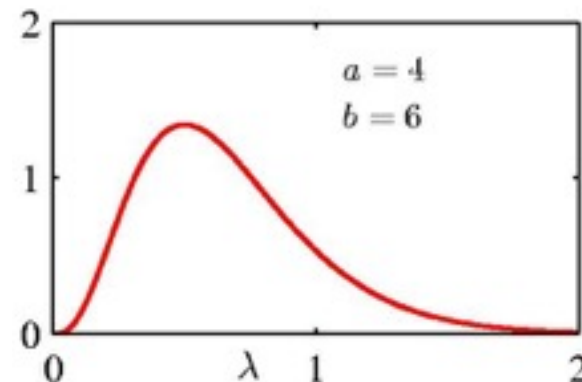
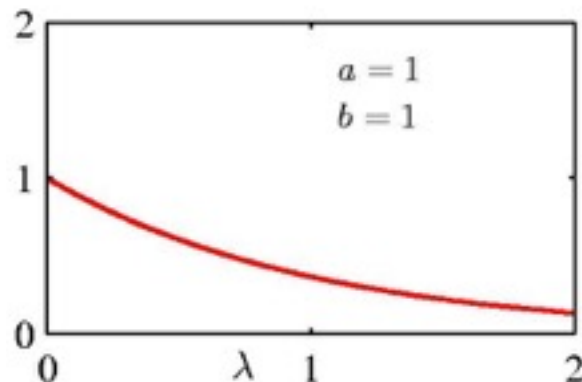
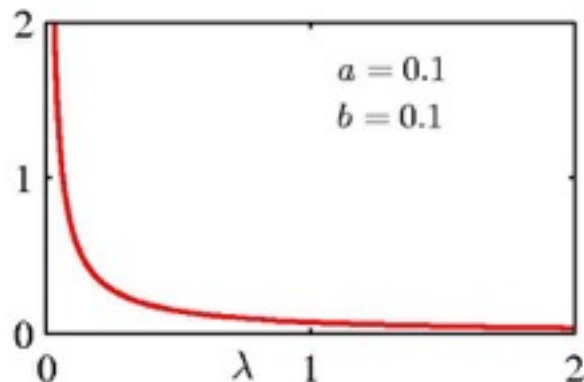
- This has a Gamma shape as a function of λ .

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 μ and λ are unknown, the joint likelihood function is given by

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

- We need a prior with the same functional dependence on μ and λ .

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 μ .
- Linear in λ .
- Gamma distribution over λ .
- Independent of μ .

Any Questions?!