CS 2750 Machine Learning
Lecture 6

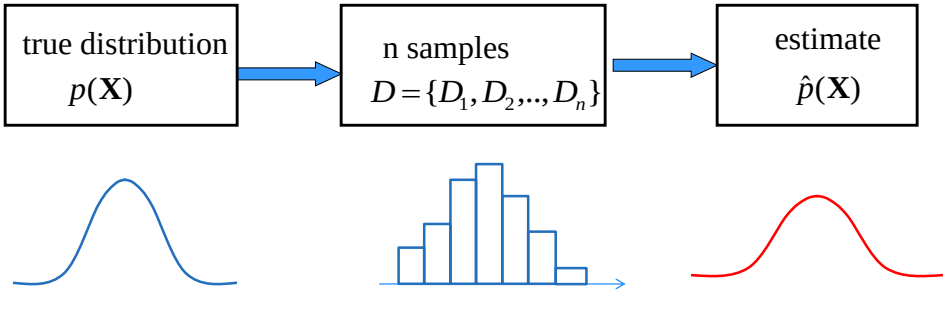
Density estimation II

Milos Hauskrecht
milos@pitt.edu
5329 Sennott Square

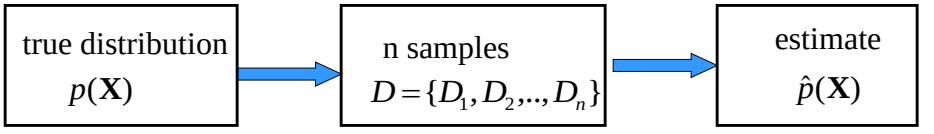
Density estimation

Data: $D = \{D_1, D_2, .., D_n\}$
$D_i = \mathbf{x}_i$ a vector of attribute values

Objective: estimate the model of the underlying probability distribution over variables $\mathbf{X}$, $p(\mathbf{X})$, using examples in D



Density estimation: iid assumptions



Standard (iid) assumptions: Samples

- **are independent of each other**
- **come from the same (identical) distribution (fixed $p(\mathbf{X})$)**

Independently drawn instances from the same fixed distribution

Density estimation

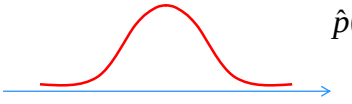
Types of density estimation:

(1) Parametric

- the distribution is modeled using a set of parameters Θ

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \Theta)$$

- **Estimation:** find parameters Θ fitting the data D
- **Example:** estimate the mean and covariance of a normal distribution



$$\hat{p}(x) = N(x \mid \mu, \sigma)$$

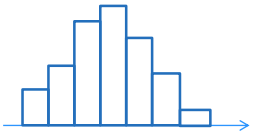
Density estimation

Types of density estimation:

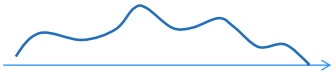
(2) Non-parametric

- The model of the distribution utilizes all examples in D
- As if all examples were parameters of the distribution
- $\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid D)$
- **Examples:**

histogram



Kernel density estimation



ML Parameter estimation

Model $\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \Theta_{ML})$ **Data** $D = \{D_1, D_2, .., D_n\}$

- **Maximum likelihood (ML)**

$$\Theta_{ML} = \arg\max_{\Theta} P(D \mid \Theta, \xi)$$

$$p(D \mid \Theta, \xi) = P(D_1, D_2, .., D_n \mid \Theta, \xi)$$

$$= P(D_1 \mid \Theta, \xi) P(D_2 \mid \Theta, \xi) \ldots P(D_n \mid \Theta, \xi)$$

$$= \prod_{i=1}^{n} P(D_i \mid \Theta, \xi)$$

Independent examples

Log likelihood – has the same maximum as likelihood

$$\Theta_{ML} = \arg\max_{\Theta} P(D \mid \Theta, \xi) = \arg\max_{\Theta} \log P(D \mid \Theta, \xi)$$

Bernoulli distribution

Model for random variable with two outcomes

- **Random variable:** x
- **Two outcomes:** 0 or 1
- **Bernoulli Distribution:**

$$P(x \mid \theta) = \theta^{x}(1-\theta)^{(1-x)}$$

where θ is the probability of x=1

Example: Coin toss

Outcomes:

- **Head → x=1**
- **Tail → x=0**
- θ **→ probability of a Head**

Maximum likelihood (ML) estimate.

Likelihood of data:

$$P(D \mid \theta, \xi) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{(1-x_i)}$$

Maximum likelihood estimate

$$\theta_{ML} = \arg\max_{\theta} P(D \mid \theta, \xi)$$

Optimize log-likelihood (the same as maximizing likelihood)

$$l(D,\theta) = N_1 \log\theta + N_2 \log(1-\theta)$$

N_1 - number of heads seen N_2 - number of tails seen

ML Solution: $\theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2}$

Maximum likelihood estimate. Example

- **Assume** the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10

What is the ML estimate of the probability of a head and a tail?

Maximum likelihood estimate. Example

- Assume the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10

What is the ML estimate of the probability of head and tail ?

Head: $\theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2} = \frac{15}{25} = 0.6$

Tail: $(1 - \theta_{ML}) = \frac{N_2}{N} = \frac{N_2}{N_1 + N_2} = \frac{10}{25} = 0.4$

Bayesian parameter estimation

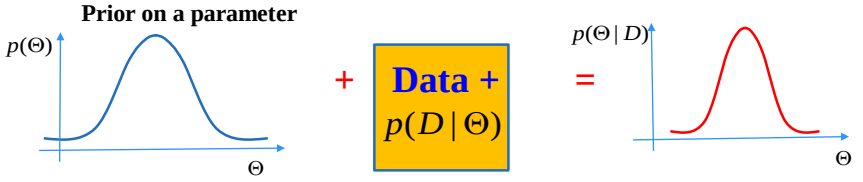
Bayesian parameter estimation

- Uses the posterior distribution of parameters
- Posterior 'covers' all possible parameter values (& their "weights")

Data Likelihood

Parameter posterior

$$p(\Theta \mid D,\xi) = \frac{p(D \mid \Theta,\xi)\,p(\Theta \mid \xi)}{p(D \mid \xi)}$$ ← Parameter prior



Bayesian parameter estimation

Uses the distributions (prior and posterior) over all possible values of the parameter θ of the sampling distribution $p(x \mid \theta)$ (Bernoulli):

Likelihood of data

Posterior **Prior**

$$p(\theta \mid D,\xi) = \frac{P(D \mid \theta,\xi)\,p(\theta \mid \xi)}{P(D \mid \xi)}$$ **(via Bayes theorem)**

Normalizing factor

We know that the likelihood is:

$$P(D \mid \theta,\xi) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{(1-x_i)} = \theta^{N_1}(1-\theta)^{N_2}$$

How to choose the prior probability?

$p(\theta \mid \xi)$ - is the prior probability on θ

Prior distribution

Choice of prior: Beta distribution

$$p(\theta \mid \xi) = Beta(\theta \mid \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1}$$

$\Gamma(x)$ - a Gamma function $\Gamma(x) = (x-1)\Gamma(x-1)$

For integer values of x $\Gamma(n) = (n-1)!$

Why to use Beta distribution?

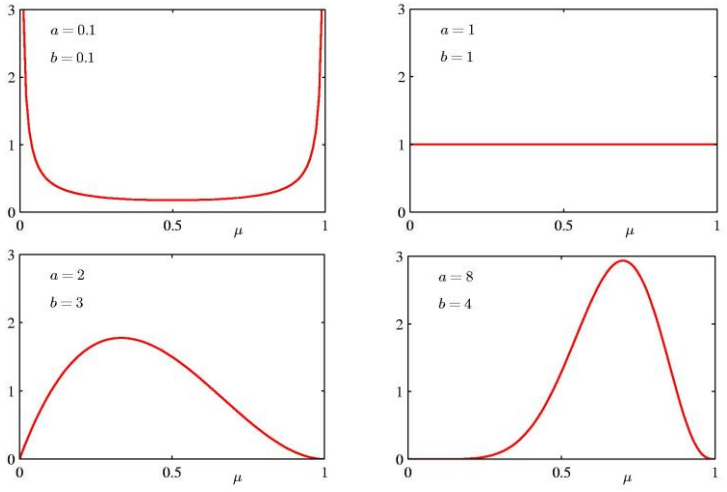
Beta distribution "**fits**" Bernoulli sample - **conjugate choices**

$$P(D \mid \theta, \xi) = \theta^{N_1}(1-\theta)^{N_2}$$

Posterior distribution is again a Beta distribution

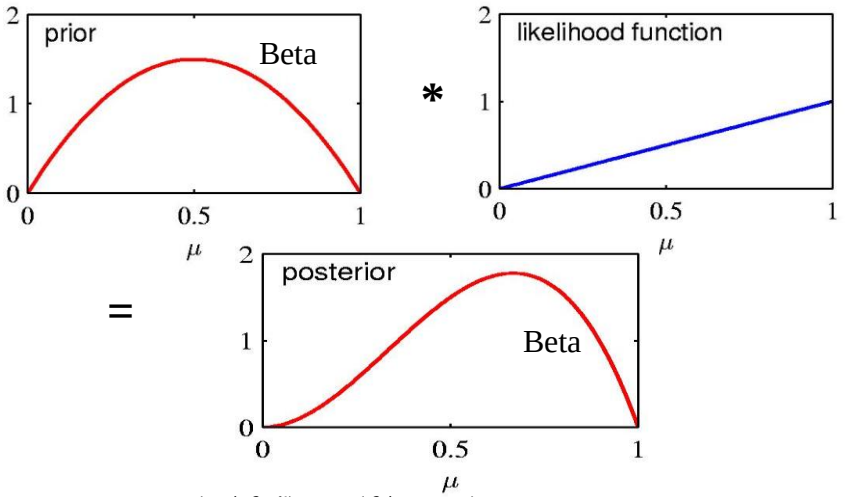
$$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi) Beta(\theta \mid \alpha_1, \alpha_2)}{P(D \mid \xi)} = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

Beta distribution



$$p(\theta \mid \xi) = Beta(\theta \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}$$

Posterior distribution



$$p(\theta \mid D,\xi) = \frac{P(D \mid \theta,\xi)Beta(\theta \mid \alpha_1,\alpha_2)}{P(D \mid \xi)} = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

Posterior distribution

Beta posterior

– **A conjugate prior to Bernoulli sample**

$$p(\theta \mid D,\xi) = \frac{P(D \mid \theta,\xi)Beta(\theta \mid \alpha_1,\alpha_2)}{P(D \mid \xi)} = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

$$= \frac{\Gamma(\alpha_1 + \alpha_2 + N_1 + N_2)}{\Gamma(\alpha_1 + N_1)\Gamma(\alpha_2 + N_2)} \theta^{N_1+\alpha_1-1}(1-\theta)^{N_2+\alpha_2-1}$$

Notice that parameters of the prior act like counts of heads and tails (sometimes they are also referred to as **prior counts**)

Posterior distribution

- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10

Example 1:

- **Assume** $p(\theta \mid \xi) = Beta(\theta \mid 5,5)$
- **Then** $p(\theta \mid D, \xi) = Beta(\theta \mid ?,?)$

Posterior distribution

- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10

Example 1:

- **Assume** $p(\theta \mid \xi) = Beta(\theta \mid 5,5)$
- **Then** $p(\theta \mid D, \xi) = Beta(\theta \mid 20,15)$

Posterior distribution

- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10

Example 1:

- **Assume** $p(\theta \mid \xi) = Beta(\theta \mid 5,5)$
- **Then** $p(\theta \mid D, \xi) = Beta(\theta \mid 20,15)$

Example 2:

- **Assume** $p(\theta \mid \xi) = Beta(\theta \mid 3,1)$
- **Then** $p(\theta \mid D, \xi) = Beta(\theta \mid ?,?)$

Posterior distribution

- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10

Example 1:

- **Assume** $p(\theta \mid \xi) = Beta(\theta \mid 5,5)$
- **Then** $p(\theta \mid D, \xi) = Beta(\theta \mid 20,15)$

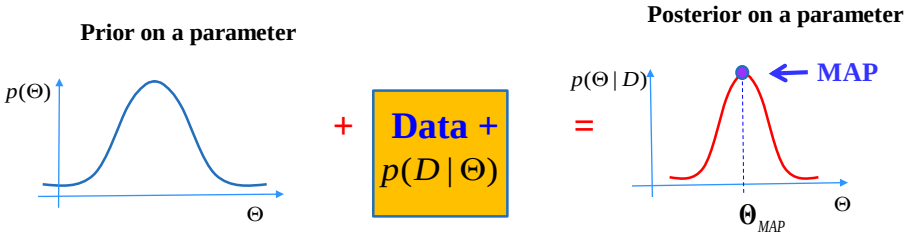
Example 2:

- **Assume** $p(\theta \mid \xi) = Beta(\theta \mid 3,1)$
- **Then** $p(\theta \mid D, \xi) = Beta(\theta \mid 18,11)$

Parameter estimation: MAP

- **Maximum a posteriori probability (MAP)**

  maximize $p(\boldsymbol{\Theta} \mid D, \xi)$



- **MAP**
  - Yields: one set of parameters $\boldsymbol{\Theta}_{MAP}$ (mode of the posterior)
  - Approximation:

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \boldsymbol{\Theta}_{MAP})$$

Maximum a posteriori estimate: MAP

Maximum a posteriori estimate

- Selects the mode of the **posterior distribution**

$$\theta_{MAP} = \arg\max_{\theta} p(\theta \mid D, \xi)$$

Likelihood of data **prior**

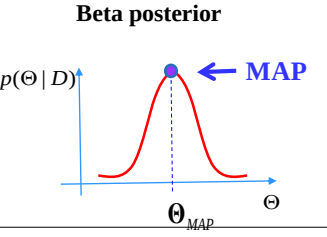
$$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi)\, p(\theta \mid \xi)}{P(D \mid \xi)}$$

Normalizing factor

By using Beta prior:

- **We get Beta posterior**
- **And MAP solution is:**

$$\theta_{MAP} = \frac{\alpha_1 + N_1 - 1}{\alpha_1 + \alpha_2 + N_1 + N_2 - 2}$$



MAP estimate example

- Assume the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10
- Assume $p(\theta \mid \xi) = Beta(\theta \mid 5,5)$

What is the MAP estimate?

MAP estimate example

- Assume the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10
- Assume $p(\theta \mid \xi) = Beta(\theta \mid 5,5)$

What is the MAP estimate ?

$$\theta_{MAP} = \frac{N_1 + \alpha_1 - 1}{N - 2} = \frac{N_1 + \alpha_1 - 1}{N_1 + N_2 + \alpha_1 + \alpha_2 - 2} = \frac{19}{33}$$

MAP estimate example

- Note that the prior and data fit (data likelihood) are combined
- **The MAP can be biased with large prior counts**
- **It is hard to overturn it with a smaller sample size**
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10
- Assume

$$p(\theta \mid \xi) = Beta(\theta \mid 5,5) \qquad \theta_{MAP} = \frac{19}{33}$$

$$p(\theta \mid \xi) = Beta(\theta \mid 5,20) \qquad \theta_{MAP} = \frac{19}{48}$$

Bayesian framework

- **Predictive probability of an outcome** $x = 1$ **in the next trial**

  $P(x = 1 \mid D, \xi)$

Posterior density

$$P(x = 1 \mid D, \xi) = \int_0^1 P(x = 1 \mid \theta, \xi)\overbrace{p(\theta \mid D, \xi)}d\theta$$

$$= \int_0^1 \theta p(\theta \mid D, \xi)d\theta = E(\theta)$$

- **Equivalent to the expected value of the parameter**
  - expectation is taken with respect to the posterior distribution

$$p(\theta \mid D, \xi) = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

Expected value of the parameter

How to calculate the expected value of Beta?

$$E(\theta) = \int_0^1 \theta Beta(\theta \mid \eta_1, \eta_2) d\theta = \int_0^1 \theta \frac{\Gamma(\eta_1 + \eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)} \theta^{\eta_1 - 1}(1-\theta)^{\eta_2 - 1} d\theta$$

$$= \frac{\Gamma(\eta_1 + \eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)} \int_0^1 \theta^{\eta_1}(1-\theta)^{\eta_2 - 1} d\theta$$

$$= \frac{\Gamma(\eta_1 + \eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)} \frac{\Gamma(\eta_1 + 1)\Gamma(\eta_2)}{\Gamma(\eta_1 + \eta_2 + 1)} \underbrace{\int_0^1 Beta(\eta_1 + 1, \eta_2) d\theta}_{1}$$

$$= \frac{\eta_1}{\eta_1 + \eta_2}$$

Note: $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$ for integer values of α

Expected value of the parameter

- **Substituting the results for the posterior:**

$$p(\theta \mid D, \xi) = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

- **We get** $E(\theta) = \frac{\alpha_1 + N_1}{\alpha_1 + N_1 + \alpha_2 + N_2}$

- **Note that the mean of the posterior is yet** another "reasonable" parameter choice:

$$\hat{\theta} = E(\theta)$$

$$\Theta_{EV} = E_{p(\Theta \mid D, \xi)}(\Theta) = \int \Theta p(\Theta \mid D, \xi) d\Theta$$

Binomial distribution

Example problem: *N* coin flips, where each coin flip can have two results: head or tail

Outcome: N_1 - number of heads seen in N trials N_2 - number of tails seen

Model: probability of a head θ
probability of a tail $(1-\theta)$

Probability of an outcome:

$$P(N_1 \mid N, \theta) = \binom{N}{N_1} \theta^{N_1} (1-\theta)^{N-N_1}$$ **Binomial distribution**

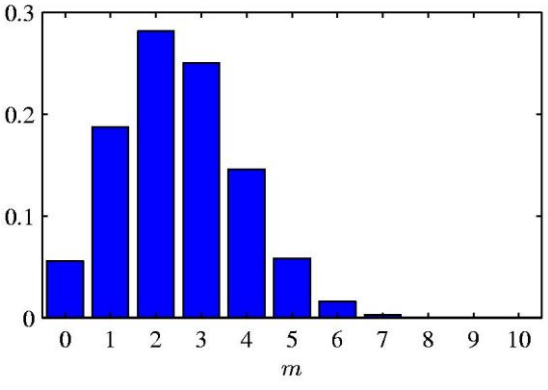
Binomial distribution:

- **models order independent sequence of Bernoulli trials**

Binomial distribution

Binomial distribution:

Bin(m|10, 0.25)



Matching prior: Beta distribution

Maximum likelihood (ML) estimate.

Likelihood of data:

$$P(D \mid \theta) = \binom{N}{N_1} \theta^{N_1} (1-\theta)^{N_2} = \frac{N!}{N_1! N_2!} \theta^{N_1} (1-\theta)^{N_2}$$

Log-likelihood

$$l(D,\theta) = \log \binom{N}{N_1} \theta^{N_1} (1-\theta)^{N_2} = \log \frac{N!}{N_1! N_2!} + N_1 \log \theta + N_2 \log(1-\theta)$$

Constant from the point of optimization !!!

ML Solution: $\theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2}$

The same as for a sequence of iid Bernoulli trials

Posterior density

Posterior density

$$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi) p(\theta \mid \xi)}{P(D \mid \xi)}$$ **(via Bayes rule)**

Prior choice

$$p(\theta \mid \xi) = Beta(\theta \mid \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1 - 1} (1-\theta)^{\alpha_2 - 1}$$

Likelihood

$$P(D \mid \theta) = \frac{\Gamma(N_1 + N_2)}{\Gamma(N_1)\Gamma(N_2)} \theta^{N_1} (1-\theta)^{N_2}$$

Posterior $p(\theta \mid D, \xi) = Beta(\alpha_1 + N_1, \alpha_2 + N_2)$

MAP estimate $\theta_{MAP} = \arg\max_{\theta} p(\theta \mid D, \xi)$

$$\theta_{MAP} = \frac{\alpha_1 + N_1 - 1}{\alpha_1 + \alpha_2 + N_1 + N_2 - 2}$$

Multinomial distribution

Example: multiple rolls of a die with 6 results

Outcome: counts of occurrences of k possible outcomes of N trials: N_i - a number of times an outcome i has been seen

$$\sum_{i=1}^{k} N_i = N$$

Model parameters: $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots \theta_k)$ **s.t.** $\sum_{i=1}^{k} \theta_i = 1$

θ_i - probability of an outcome i

Probability distribution:

$$P(N_1, N_2, \ldots N_k \mid \boldsymbol{\theta}, \xi) = \frac{N!}{N_1! N_2! \ldots N_k!} \theta_1^{N_1} \theta_2^{N_2} \ldots \theta_k^{N_k}$$

Multinomial distribution

ML estimate:

$$\theta_{i,ML} = \frac{N_i}{N}$$

Posterior and MAP estimate

Choice of the prior: Dirichlet distribution

$$Dir(\boldsymbol{\theta} \mid \alpha_1, .., \alpha_k) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \ldots \theta_k^{\alpha_k - 1}$$

Dirichlet is the conjugate choice for the multinomial sampling

$$P(D \mid \boldsymbol{\theta}, \xi) = P(N_1, N_2, \ldots N_k \mid \boldsymbol{\theta}, \xi) = \frac{N!}{N_1! N_2! \ldots N_k!} \theta_1^{N_1} \theta_2^{N_2} \ldots \theta_k^{N_k}$$

Posterior density

$$p(\boldsymbol{\theta} \mid D, \xi) = \frac{P(D \mid \boldsymbol{\theta}, \xi) Dir(\boldsymbol{\theta} \mid \alpha_1, \alpha_2, .. \alpha_k)}{P(D \mid \xi)} = Dir(\boldsymbol{\theta} \mid \alpha_1 + N_1, .., \alpha_k + N_k)$$

MAP estimate:

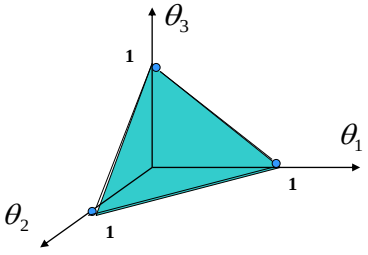
$$\theta_{i,MAP} = \frac{\alpha_i + N_i - 1}{\sum_{i=1,..k} (\alpha_i + N_i) - k}$$

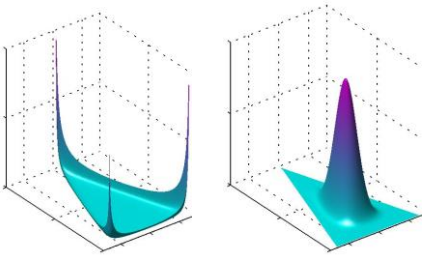
Dirichlet distribution

Dirichlet distribution:

$$Dir(\boldsymbol{\theta} \mid \alpha_1,..,\alpha_k) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \ldots \theta_k^{\alpha_k - 1}$$

Assume: k=3





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

Distribution models for random variables

Distribution models covered so far:

- **Bernoulli distribution**
  - **Model for binary random variables**

$$P(x \mid \theta) = \theta^x (1-\theta)^{(1-x)}$$

- **Binomial distribution**
  - **Model for order independent sets of binary outcomes**

$$P(N_1 \mid N, \theta) = \binom{N}{N_1} \theta^{N_1} (1-\theta)^{N-N_1}$$

- **Multinomial distribution**
  - **Model for order independent sets of k-nary outcomes**

$$P(N_1, N_2, \ldots N_k \mid \boldsymbol{\theta}, \xi) = \frac{N!}{N_1! N_2! \ldots N_k!} \theta_1^{N_1} \theta_2^{N_2} \ldots \theta_k^{N_k}$$

Distribution models for random variables

Models for other types of random variables:

- **Gaussian distribution**
  - **Models of real-valued random variable**
- **Gamma distribution:**
  - **Models of random variables for positive real numbers**
- **Exponential distribution**
  - **Models of random variables for positive real numbers**
- **Poisson distribution**
  - **Models of random variables for nonegative integers**

Conjugate choices of priors for some these distributions:

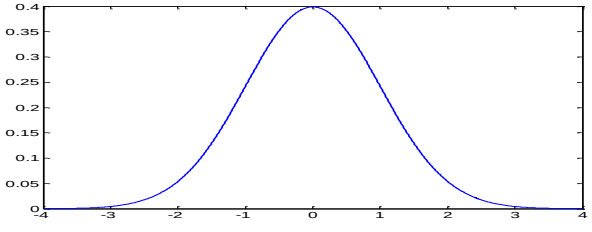
- **Exponential – Gamma**
- **Poisson – Inverse Gamma**
- **Gaussian - Gaussian (mean) and Wishart (covariance)**

Gaussian (normal) distribution

- **Gaussian:** $x \sim N(\mu, \sigma)$
- **Parameters:** μ - mean
  σ - standard deviation
- **Density function:**

$$p(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2\sigma^2}(x-\mu)^2]$$

- **Example:**



$N(0,1)$

Parameter estimates

- **Loglikelihood** $l(D,\mu,\sigma) = \log \prod_{i=1}^{n} p(x_i \mid \mu,\sigma)$
- **ML estimates of the mean and variance:**

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

  – ML variance estimate is biased

$$E_n(\sigma^2) = E_n\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2\right) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

- **Unbiased estimate:**

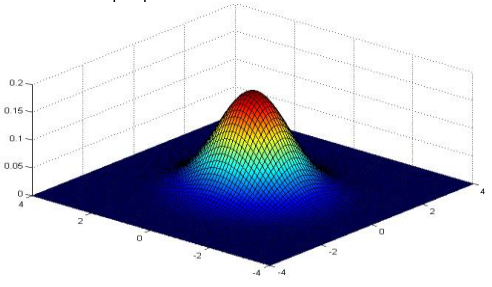
$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Multivariate normal distribution

- **Multivariate normal:** $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- **Parameters:** $\boldsymbol{\mu}$- mean
  $\boldsymbol{\Sigma}$- covariance matrix
- **Density function:**

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

- **Example:**



Partitioned Gaussian Distributions

- **Multivariate Gaussian:**

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

- **Example:**

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

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

Precision matrix

- **What are the distributions for marginals and conditionals?**

$$p(x_a) \qquad p(x_a \mid x_b)$$

Conditionals and Marginals

- **Conditional density:**

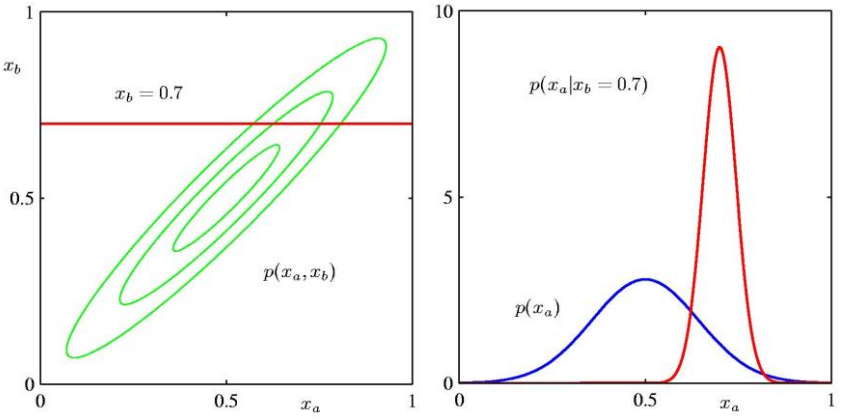
$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$$

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

- **Marginal Density:**

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

Conditionals and Marginals



Parameter estimates

- **Loglikelihood** $l(D, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log \prod_{i=1}^{n} p(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$
- **ML estimates of the mean and covariances:**

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \qquad \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$$

  – Covariance estimate is biased

$$E_n(\hat{\boldsymbol{\Sigma}}) = E_n\left(\frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T\right) = \frac{n-1}{n}\boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}$$

- **Unbiased estimate:**

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$$

Posterior of the mean of a multivariate normal

- **Assume a prior on the mean $\boldsymbol{\mu}$ that is normally distributed:**

$$p(\boldsymbol{\mu}) = N(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$$

- **Then the posterior of $\boldsymbol{\mu}$ is normally distributed**

$$p(\boldsymbol{\mu} \mid D) \approx \left(\prod_{i=1}^{n} \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right] \right)$$

$$* \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_p|^{1/2}} \exp\left[-\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_p)^T \boldsymbol{\Sigma}_p^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_p) \right]$$

$$= \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_n|^{1/2}} \exp\left[-\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_n)^T \boldsymbol{\Sigma}_n^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_n) \right]$$

Posterior of the mean of a multivariate normal

- **Then the posterior of $\boldsymbol{\mu}$ is normally distributed**

$$p(\boldsymbol{\mu} \mid D) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_n|^{1/2}} \exp\left[-\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_n)^T \boldsymbol{\Sigma}_n^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_n) \right]$$

$$\boldsymbol{\Sigma}_n^{-1} = n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_p^{-1}$$

$$\boldsymbol{\mu}_n = \boldsymbol{\Sigma}_p \left(\boldsymbol{\Sigma}_p + \frac{1}{n}\boldsymbol{\Sigma} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right) \quad + \frac{1}{n}\boldsymbol{\Sigma} \left(\boldsymbol{\Sigma}_p + \frac{1}{n}\boldsymbol{\Sigma} \right)^{-1} \boldsymbol{\mu}_p$$

$$\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_p \left(\boldsymbol{\Sigma}_p + \frac{1}{n}\boldsymbol{\Sigma} \right)^{-1} \frac{1}{n}\boldsymbol{\Sigma}$$

Other distributions

Gamma distribution:

$$p(x \mid a,b) = \frac{1}{\Gamma(a)b^a} x^{a-1} e^{-\frac{x}{b}} \quad \text{for} \quad x \in [0,\infty]$$

Exponential distribution:

- A special case of Gamma for a=1

$$p(x \mid b) = \left(\frac{1}{b}\right) e^{-\frac{x}{b}} \quad \text{for} \quad x \in [0,\infty]$$

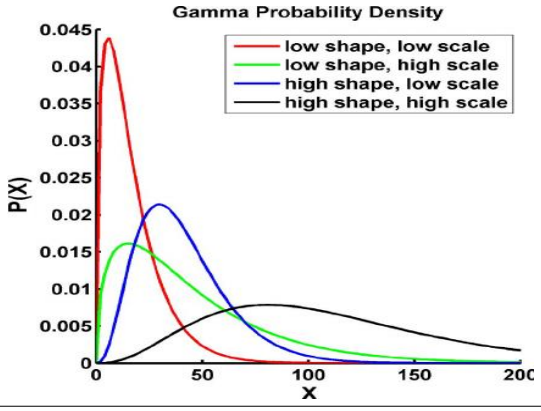
Poisson distribution:

$$p(x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{for} \quad x \in \{0,1,2,\ldots\}$$

Gamma distribution

$$p(\lambda \mid a,b) = \frac{1}{\Gamma(a)b^a} \lambda^{a-1} e^{-\frac{\lambda}{b}} \quad \text{for} \quad \lambda \in [0,\infty]$$

where *a* is the shape and *b* is a scale parameter

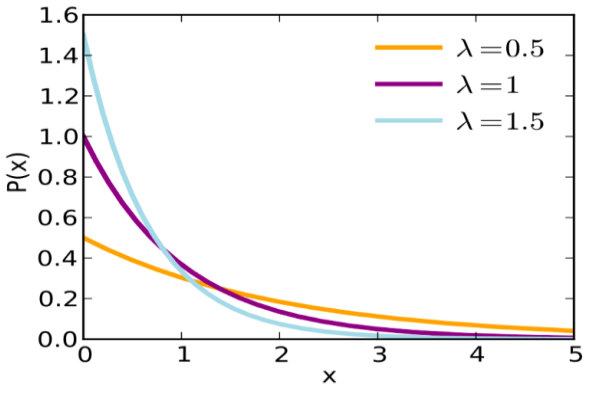


Exponential distribution

$$p(x \mid b) = \left(\frac{1}{b}\right) e^{-\frac{x}{b}} \qquad \text{for} \quad x \in [0, \infty]$$

Alternative parameterization: $p(x \mid \lambda) = \lambda e^{-\lambda x}$

where $\lambda = 1/b$



Poisson distribution

Poisson distribution:

$$p(x \mid \lambda) = \frac{e^{-\lambda} \lambda^{x}}{x!} \qquad \text{for} \quad x \in \{0,1,2,\ldots\}$$