CS 2750 Machine Learning
Lecture 6

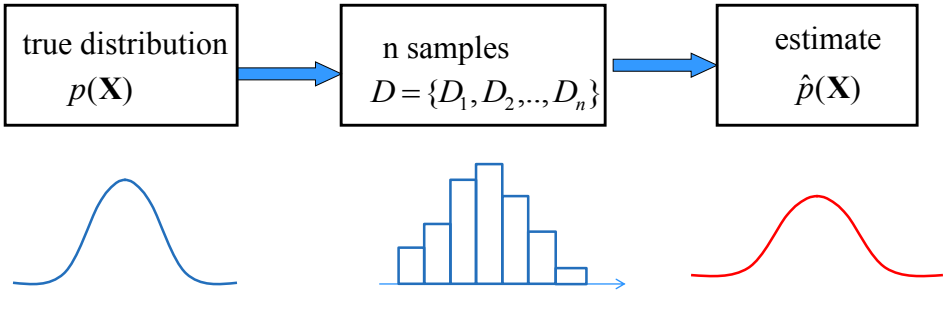
Density estimation II

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

Types of density estimation:

Parametric

- the distribution is modeled using a set of parameters Θ

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

- **Example:** mean and covariances of a multivariate normal
- **Estimation:** find parameters Θ describing data D

Non-parametric

- The model of the distribution utilizes all examples in D
- As if all examples were parameters of the distribution
- **Examples:** Nearest-neighbor

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

$$\log p(D \mid \Theta, \xi) = \sum_{i=1}^{n} \log P(D_i \mid \Theta, \xi)$$

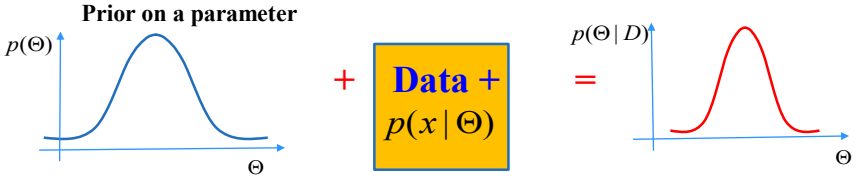
Bayesian parameter estimation

Bayesian parameter estimation

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

Parameter posterior — Data Likelihood — Parameter prior

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



Conjugate choices: Prior distribution **matches** the data distribution → Posterior is the same type

Parameter estimation

Other criteria:

- **Maximum a posteriori probability (MAP)** (mode of the posterior)

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

  $$\Theta_{MAP} = \arg\max_{\Theta} p(\boldsymbol{\Theta} \mid D,\xi)$$

- **Expected value of the parameter**

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

  – Expectation taken with regard to posterior

  – (mean of the posterior)

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

Bernoulli distribution

Model for random variable with two outcomes

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

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?

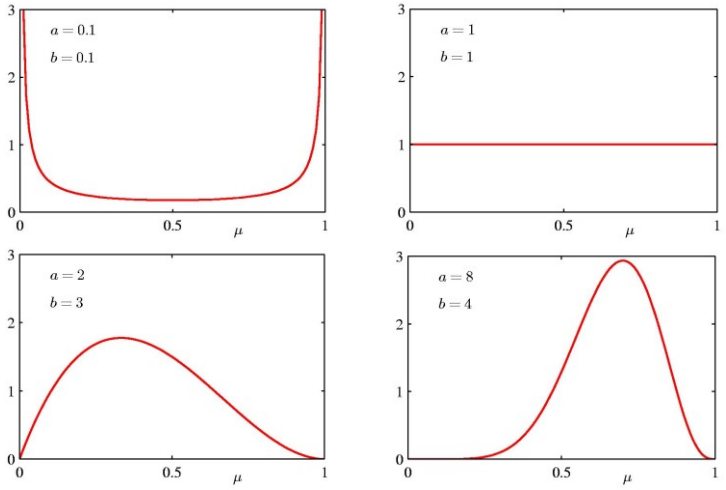
Beta distribution "**fits**" Bernoulli sample - **conjugate choices**

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

Posterior distribution is again a Beta distribution

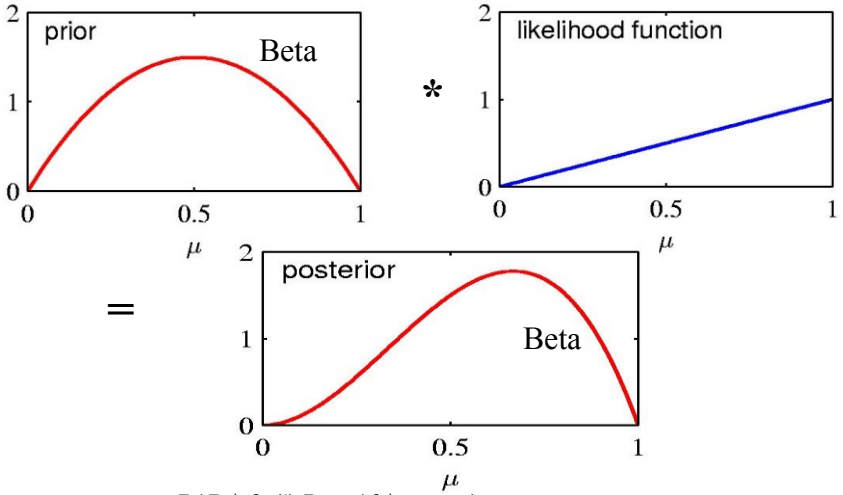
$$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi)Beta(\theta \mid \alpha_1, \alpha_2)}{P(D \mid \xi)} = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

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



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

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:

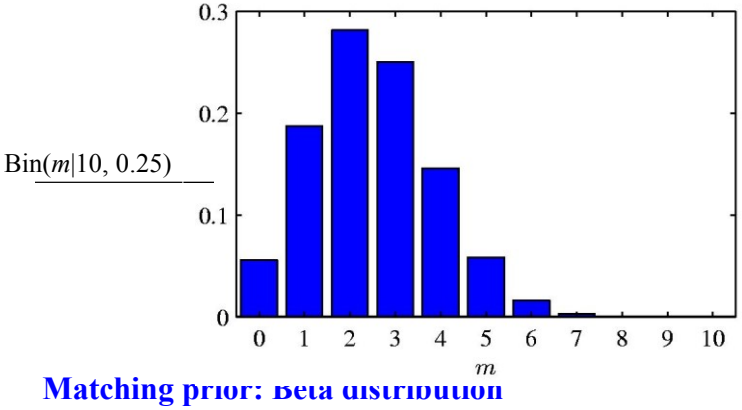
$$P(N_1 \mid N, \theta) = \binom{N}{N_1} \theta^{N_1} (1-\theta)^{N-N_1}$$ **Binomial distribution**

Binomial distribution:

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

Binomial distribution

Binomial distribution:



Matching prior: Beta distribution

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:

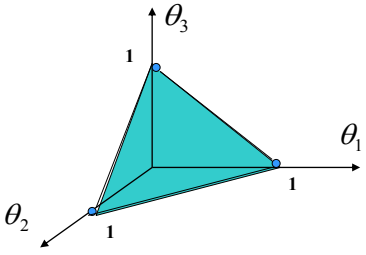
$$\theta_{i,MAP} = \frac{\alpha_i + N_i - 1}{\sum_{i=1,..k} (\alpha_i + N_i) - k}$$

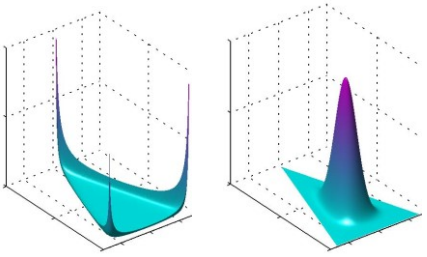
Dirichlet distribution

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





[redacted] [redacted]

Other distributions

The same ideas can be applied to other distributions

- Typically we choose distributions that behave well so that computations lead to **"nice" solutions**

- **Exponential family of distributions**

Conjugate choices for some of the distributions from the exponential family:

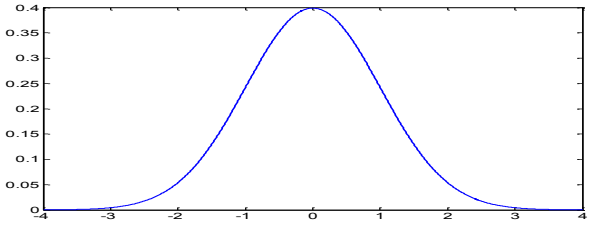
- **Binomial – Beta**
- **Multinomial - Dirichlet**
- **Exponential – Gamma**
- **Poisson – Inverse Gamma**
- **Gaussian - Gaussian (mean) and Wishart (covariance)**

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:**

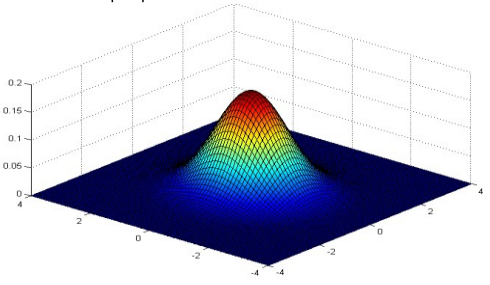
$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

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:**

- **Example:**

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

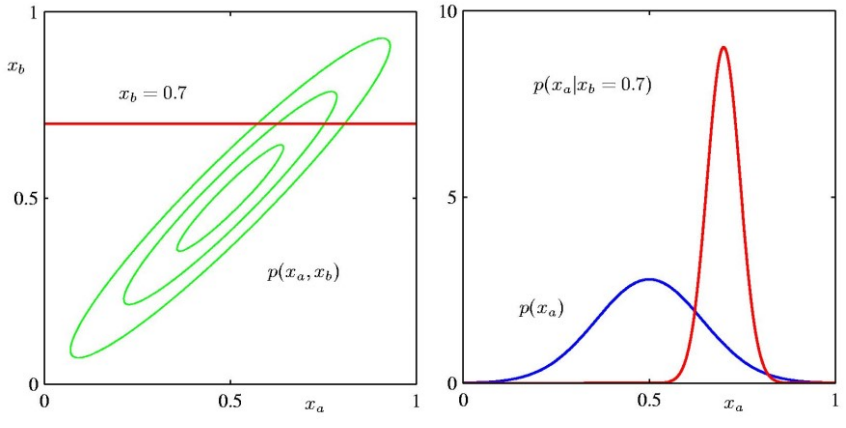
$p(x_a)$ $\quad$ $p(x_a \mid x_b)$

Conditionals and Marginals

- **Conditional density:**

- **Marginal Density:**

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:**

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

- **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**

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

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

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

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

Other distributions

Gamma distribution:

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

Exponential distribution:

- A special case of Gamma for a=1

$$p(x \mid b) = \left(\frac{1}{b}\right) e^{-\frac{x}{b}}$$

Poisson distribution:

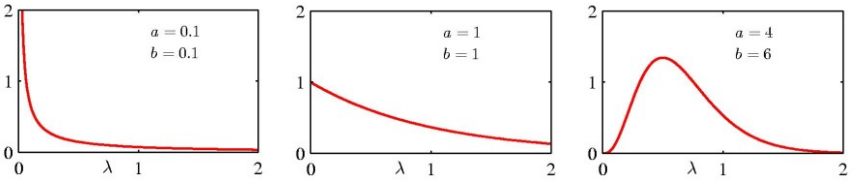
$$p(x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad \text{for} \quad x \in \{0,1,2,\ldots\}$$

Other distributions

Gamma distribution:

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

$$E(\lambda) = \frac{a}{b} \qquad \text{var}(\lambda) = \frac{a}{b^2}$$



Sequential Bayesian parameter estimation

- **Sequential Bayesian approach**
  - Under the iid the estimates of the posterior can be computed incrementally for a sequence of data points

$$p(\Theta \mid D,\xi) = \frac{p(D \mid \Theta,\xi)p(\Theta \mid \xi)}{\int_{\Theta} p(D \mid \Theta,\xi)p(\Theta \mid \xi)d\Theta}$$

- If we use a conjugate prior we get back the same posterior
- Assume we split the data D in the last element **x** and the rest

$$p(D \mid \boldsymbol{\Theta}) = P(x \mid \boldsymbol{\Theta})P(D_{n-1} \mid \boldsymbol{\Theta})$$

- **Then:**

A "new" prior

$$p(\Theta \mid D,\xi) = \frac{P(x \mid \boldsymbol{\Theta})\overbrace{P(D_{n-1} \mid \boldsymbol{\Theta})p(\Theta \mid \xi)}}{\int_{\Theta} P(x \mid \boldsymbol{\Theta})P(D_{n-1} \mid \boldsymbol{\Theta})p(\Theta \mid \xi)d\Theta}$$