Data Analytics
EEE 4774 & 6777

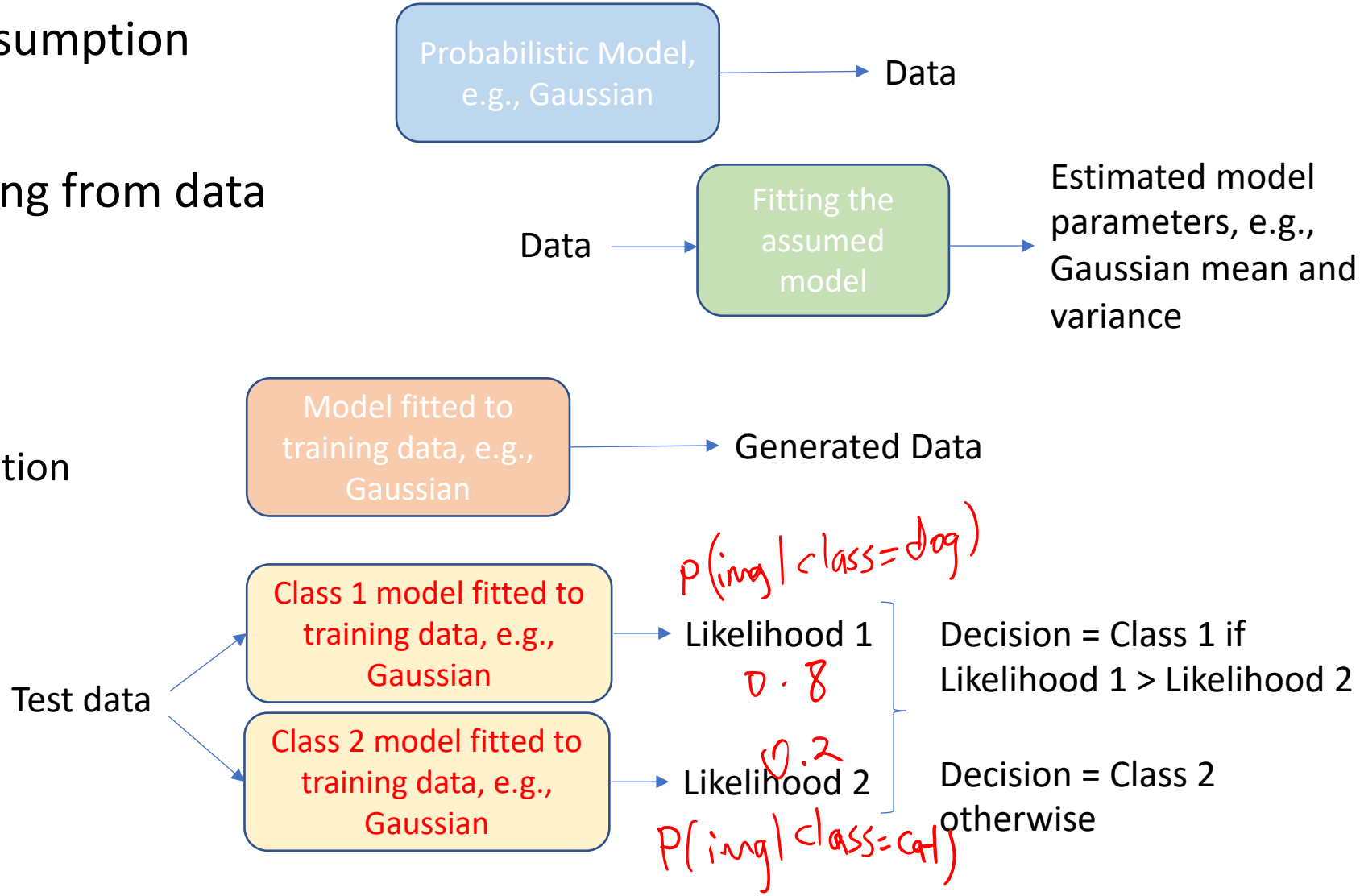
Module 2

Parameter Estimation

Spring 2022

Parameter Estimation for Model Fitting

- Generative model assumption
- Model inference/fitting from data
- Used for
  - Data generation
  - Missing value estimation
  - Classification



1-dimensional
Univariate

Gaussian Distribution

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

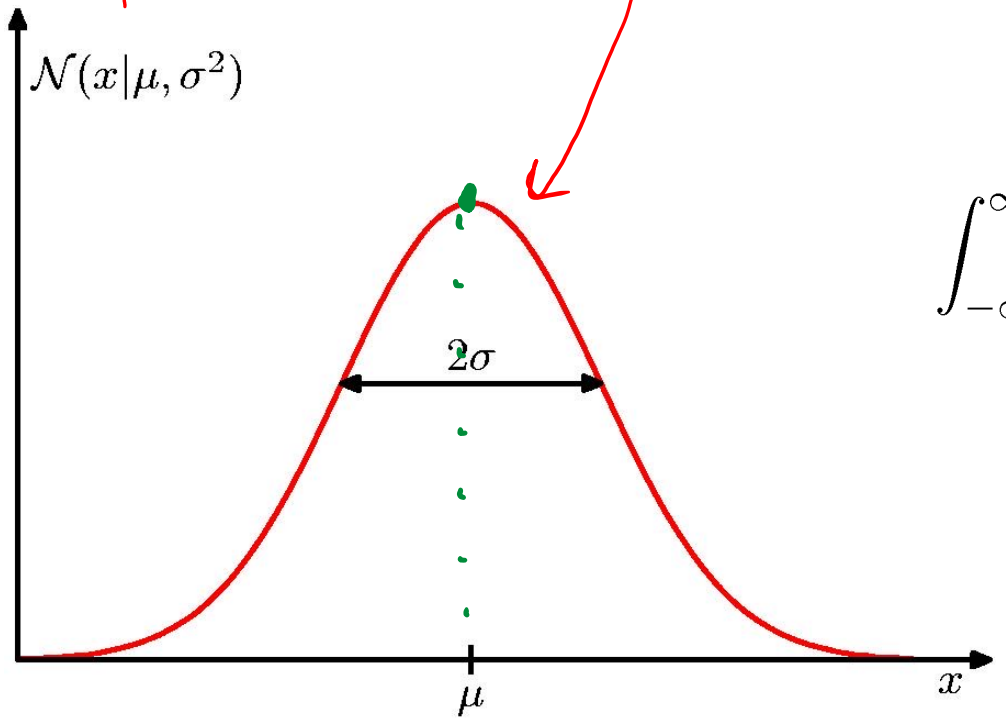
pdf

parameters

data

x_i : instance / observation / data point

e.g. : height of ith person in dataset



$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$

$p(\mu) \neq p(x=\mu) = 0$

$p(x \leq \mu)$

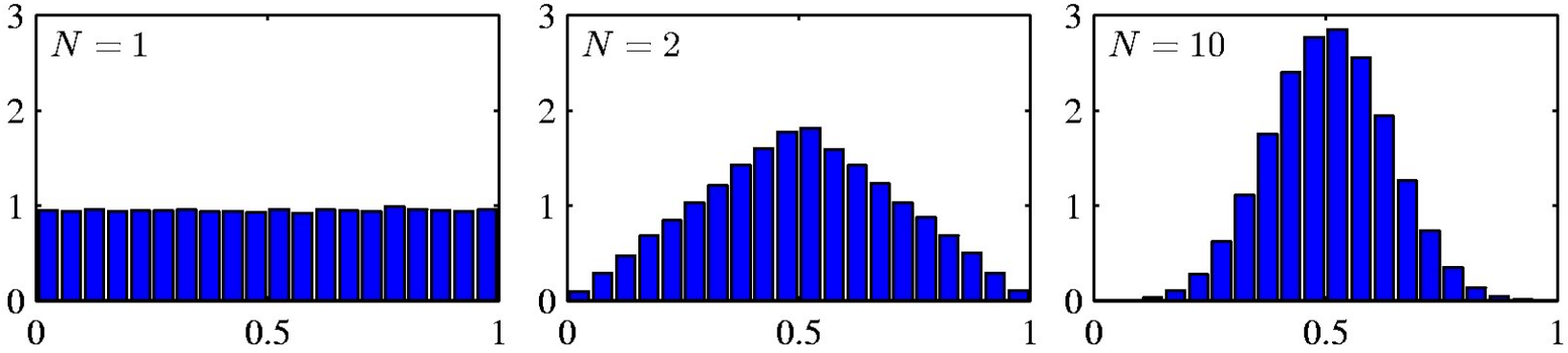
Central Limit Theorem

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

$$y = \frac{1}{N}\sum_{i=1}^{N} x_i \sim \mathcal{N}(\mu, \sigma^2)$$

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

$$x_i \sim \text{Uniform}(0,1)$$



Multivariate Gaussian

$\sigma_{12} = E[(x_1-\mu_1)(x_2-\mu_2)]$

$\neq 0$

$\neq E[x-\mu_1]\,E[x_2-\mu_2]$

$x_1 \& x_2$ correlated

$x_1 \& x_2$ dependent

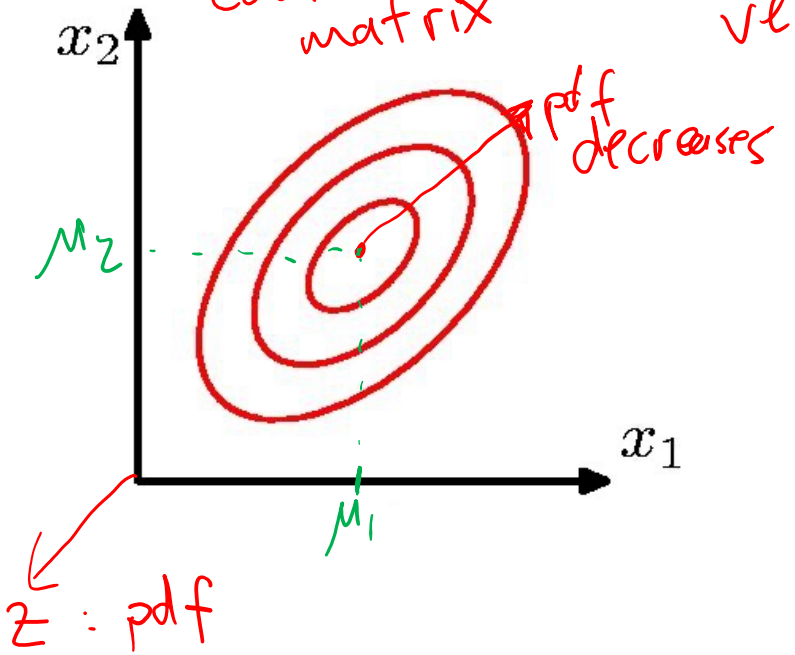
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

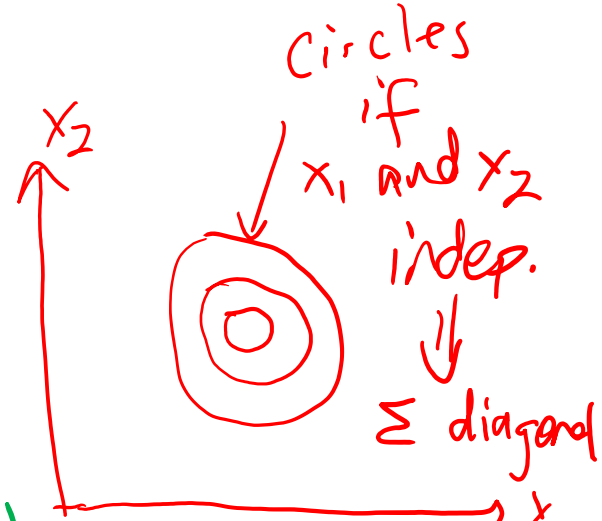
vector, matrix, covariance matrix, mean vector

Σ in general not diagonal

$\Sigma = E[(x-\mu)(x-\mu)^T]$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{12} & \sigma_2^2 & & \\ \vdots & & \ddots & \\ \sigma_{1d} & & & \sigma_d^2 \end{bmatrix}$$





dxd

$\Sigma =$ diagonal $\begin{bmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_d^2 \end{bmatrix}$

Gaussian Mean and Variance

$$E[x] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} x \, dx$$

1st moment: mean

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x \, \mathrm{d}x = \mu$$

$E[x^3]$: 3rd moment

2nd moment

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2$$

$$E\left[(x-\mu)^2\right] = \mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2 = E\left[(x - E[x])^2\right]$$

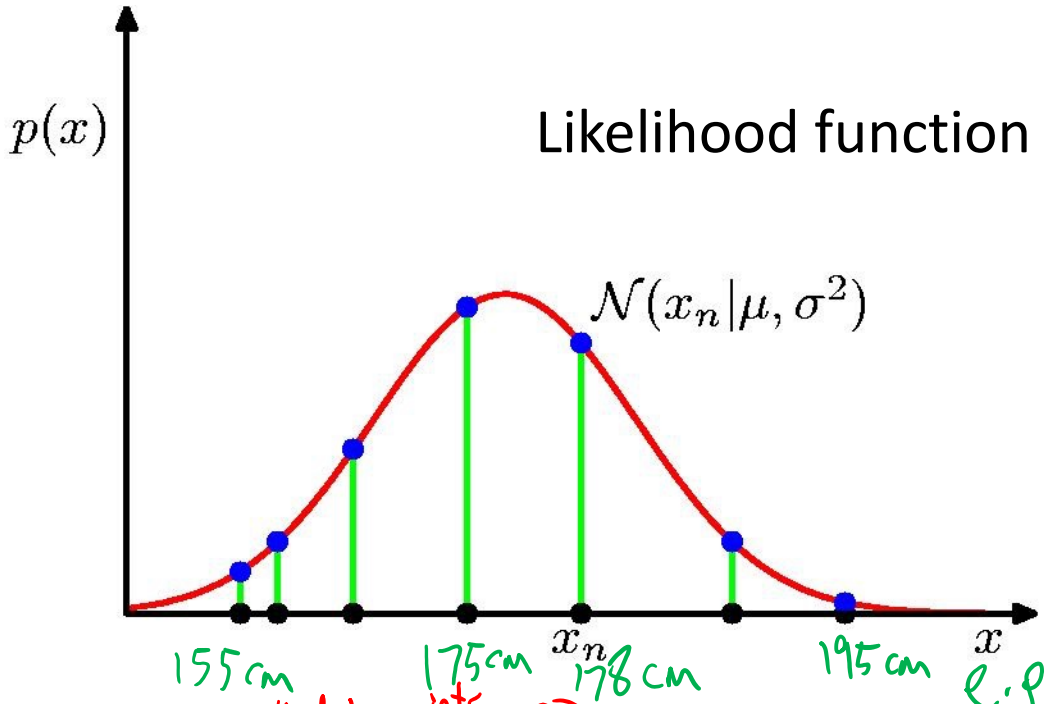
$$= E\left[x^2 - 2x\mu + \mu^2\right]$$

$$= E[x^2] - 2\underbrace{E[x]}_{\mu}\mu + \mu^2 = E[x^2] - \mu^2$$

Gaussian Parameter Estimation

n : data index

$n = 1, 2, 3, \ldots, N$



Likelihood function

$$p(\vec{\mathbf{x}}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)$$

data points → N

likelihood of observing dataset $\vec{x}$ for given μ and σ^2

data instance

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$$

e.g. height of students

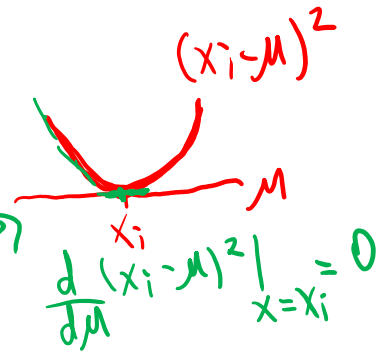
$x_1, x_2, \ldots, x_N$ indep

$$p(\vec{x}|\mu, \sigma^2) = \prod_{i=1}^{N} p(x_i|\mu, \sigma^2)$$

$p(x_1, x_2, \ldots, x_N|\mu, \sigma^2)$

$\frac{d}{d\mu}(x_i-\mu)^2 = 2(x_i-\mu)(-1) = 0$

Maximum Likelihood (ML) Estimation



$p(x_i|\mu,\sigma^2) = \frac{e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$

$p(\vec{x}|\mu,\sigma^2) = \prod_{i=1}^{N} \frac{e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} = \frac{e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2}}{(2\pi\sigma^2)^{N/2}}$

$p(x_1) > p(x_2)$

$\log p(x_1) > \log p(x_2)$

$\max_{\mu,\sigma^2} p(x|\mu,\sigma^2)$

$= \max_{\mu,\sigma^2} \log p(x|\mu,\sigma^2)$

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

$\frac{\partial}{\partial\sigma^2}\ln\sigma^2 = \frac{1}{\sigma^2}$

$\mu_{ML} = \arg\max_{\mu} \ln p(\vec{x}|\mu,\sigma^2)$

$\mu_{ML} = \arg\min_{\mu} \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$

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

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2 = E\left[(x_i-\mu)^2\right]$$

$\frac{\partial}{\partial\mu}\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\Big|_{\mu=\mu_{ML}} = 0$

$\frac{1}{2\sigma^2}\sum_{i=1}^{N}\frac{\partial}{\partial\mu}(x_i-\mu)^2\Big|_{\mu=\mu_{ML}} = 0$

$-\sum_{i=1}^{N} 2(x_i-\mu)\Big|_{\mu=\mu_{ML}} = 0$

$\sum_{i=1}^{N} x_i - \sum_{i=1}^{N}\mu_{ML} = 0$

$N\mu_{ML} = \sum_{i=1}^{N} x_i \Rightarrow$

$$\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i = E[x]$$

Properties of μ_{ML} and σ^2_{ML}

Unbiased estimator

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[x_i] = \frac{1}{N}\cdot N\cdot\mu$$

The original ML estimater for σ^2 is biased

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \left(\frac{N-1}{N}\right)\sigma^2 \neq \sigma^2$$

→ so we have one less degree of freedom

Unbiased σ^2 ML estimator

$$\widetilde{\sigma}^2 = \frac{N}{N-1}\sigma^2_{\mathrm{ML}} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2$$

by default the functions in Matlab and Python uses unbiased estimate

Parameter estimation: Bayesian for Gaussian

- Gaussian prior $p(\mu) = \mathcal{N}\left(\mu|\mu_0, \sigma_0^2\right).$

$$p(\mu|x) = \frac{p(x|\mu)p(\mu)}{p(x)}$$

likelihood prior

- posterior $p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu).$ $p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$

μ_{MAP}

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathrm{ML}},$$

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

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

$\mu_{MAP} = \arg\max_{\mu} \log p(x|\mu) + \log p(\mu)$

additional term

Maximum a posteriori estimator

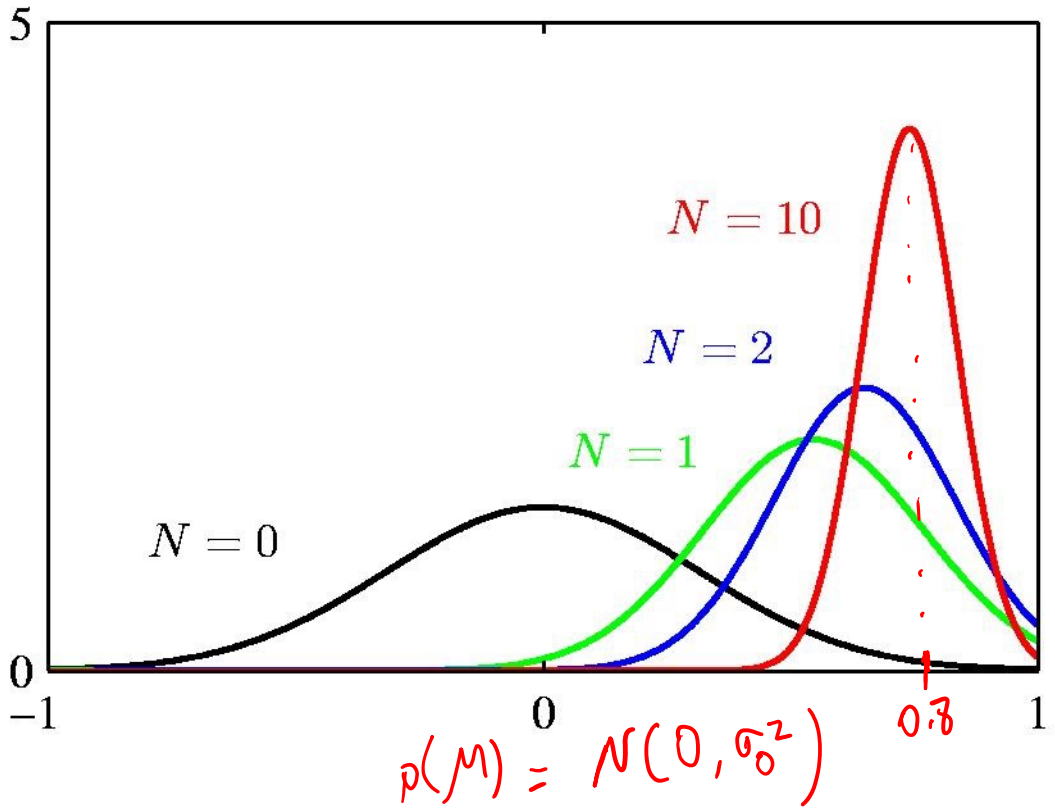
no data ↓ infinitely many data ↓

	$N = 0$	$N \to \infty$
μ_N	μ_0	μ_{ML}
σ_N^2	σ_0^2	0

Gaussian distribution provides the conjugate for Gaussian prior

Parameter estimation: Bayesian for Gaussian

- Example: $p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$ for N = 0, 1, 2 and 10.
- True mean = 0.8



Python Example

- Dataset “adult.csv” includes information (attributes/features) of a number of people such as age, gender, occupation, and whether income less than or greater than 50K per year.
- Compute the mean and standard deviation of age for two groups of people:
  - Income <= 50K
  - Income > 50K
- Use the mean and standard deviation statistics to fit a Gaussian model to each group
- Compute the histogram to check if the Gaussian model assumptions are suitable.

Python Example

- Dataset “iris.csv” includes 4 measurements (attributes/features) from a number of iris plants.
- Each plant is from one of 3 classes.
- Fit a 4-dimensional multivariate Gaussian to data from each class.
- Compute the likelihood of each test instance under the 3 Gaussian models to make a classification decision.
- Choose the class (i.e., model) with the highest likelihood value as the class prediction for each test instance.
- Compute the accuracy by comparing the predicted class labels with the ground truth.

Bernoulli Distribution

- Coin flipping: heads=1, tails=0

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

$\mu \in (0,1)$ → continuous range

$x \in \{0, 1\}$ → discrete set

- Bernoulli Distribution

pmf →

$$\begin{aligned} \text{Bern}(x|\mu) &= \mu^x(1-\mu)^{1-x} \\ \mathbb{E}[x] &= \mu \\ \text{var}[x] &= \mu(1-\mu) \end{aligned}$$

$\mu = 0.1$, $1-\mu = 0.9$

$\text{var} = 0.09$

$\mu = 0.5$, $1-\mu = 0.5$

$\text{var} = 0.25$

Binomial Distribution

x discrete r.v.
↘ # of a particular outcome in N experiments

- N coin flips:

$x \in \{0, 1, \ldots, N\}$

- Binomial Distribution

#experiments

Binary outcome

Multiple experiments

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

$\mu \in (0, 1)$

pmf

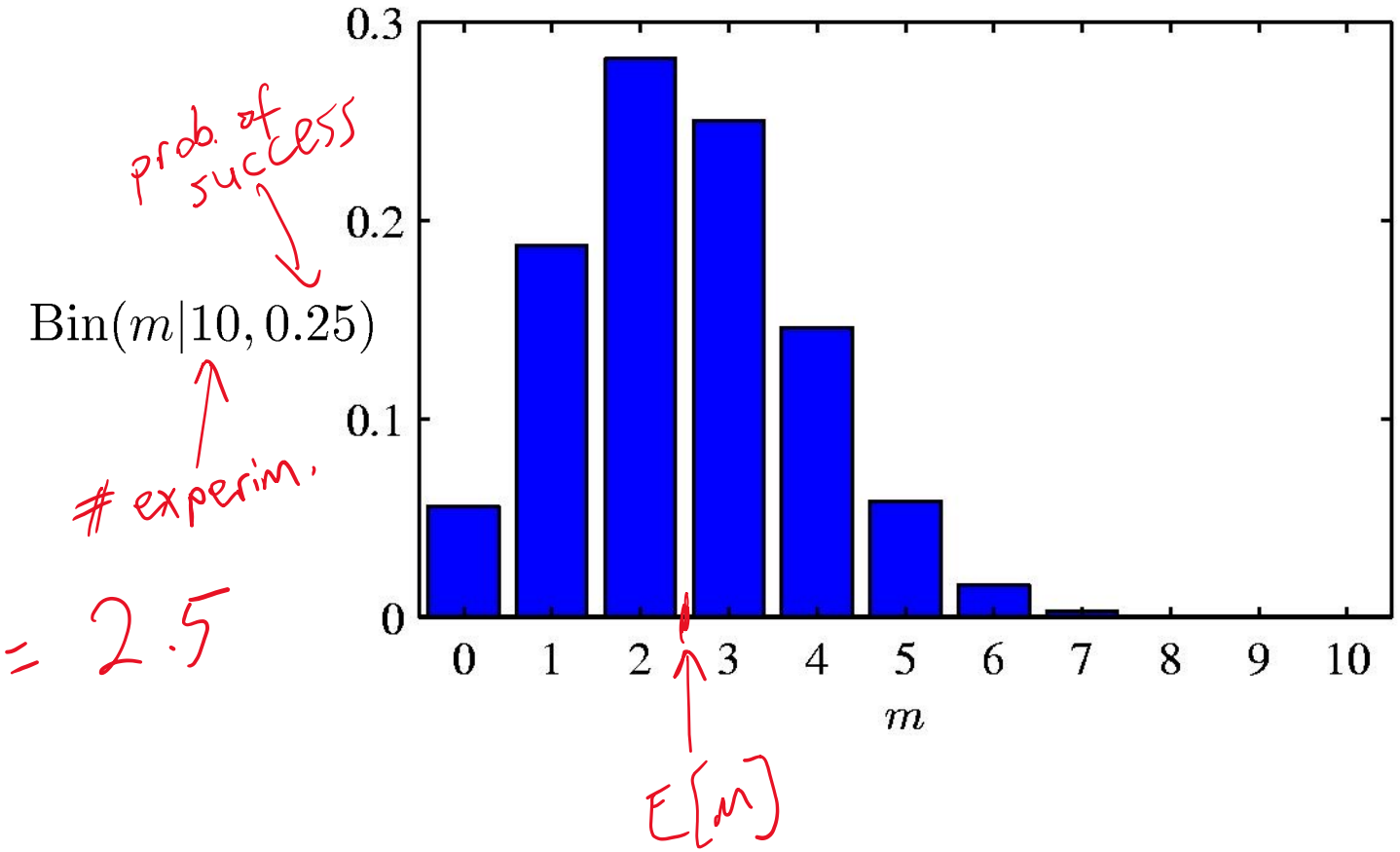
N choose m

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

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

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

Binomial Distribution



$E[m] = 2.5$

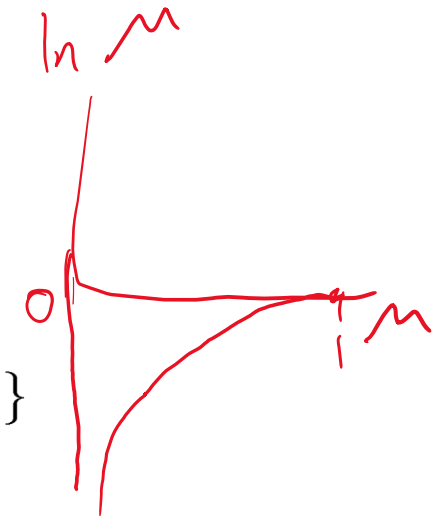
Parameter Estimation: ML for Bernoulli

dataset

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

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

Bernoulli pmf



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

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

$$\frac{\partial}{\partial \mu} \ln p(\mathcal{D}|\mu)\Big|_{\mu=\mu_{ML}} = 0$$

$$\sum_{n=1}^{N} \left\{ x_n \frac{\partial}{\partial \mu} \ln \mu + (1-x_n)\frac{\partial}{\partial \mu} \ln(1-\mu) \right\}\Big|_{\mu=\mu_{ML}} = 0 \Rightarrow \sum_{n=1}^{N} \left\{ \frac{x_n}{\mu} + \frac{1-x_n}{1-\mu}(-1) \right\}\Big|_{\mu=\mu_{ML}} = 0$$

Parameter Estimation: ML for Bernoulli

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

- Prediction: all future tosses will land heads up
- Overfitting to D

Beta Distribution

Freq. ML
Data (μ)

Bay. MAP
Data (μ)
μ (a, b)
param. hyper param.

- Distribution over $\mu \in [0, 1]$ — continuous r.v.

Prob. model

Prob. model

parameters of Beta dist.

gamma func.

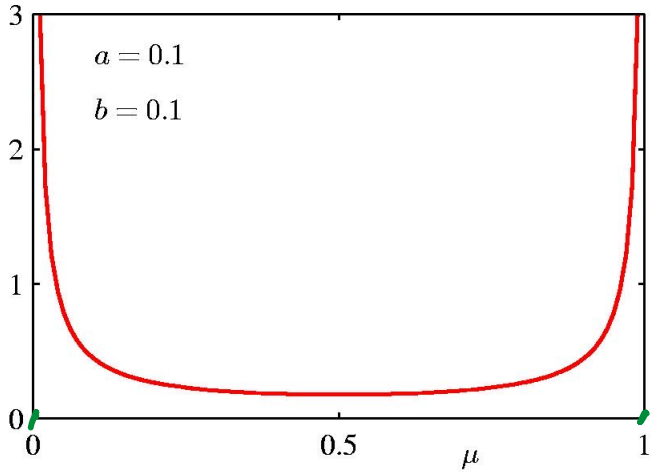
pdf

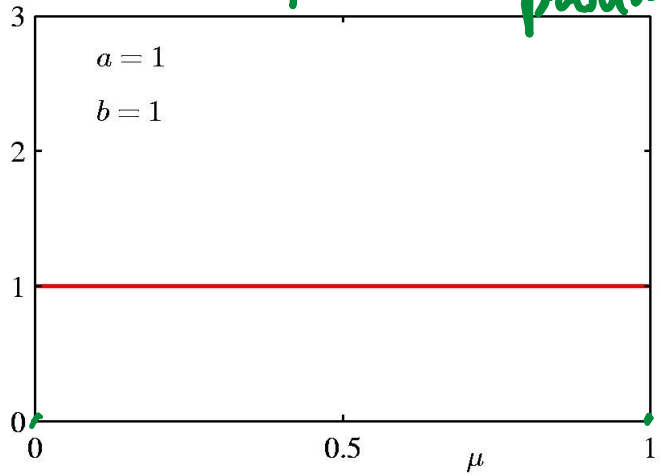
$$\text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

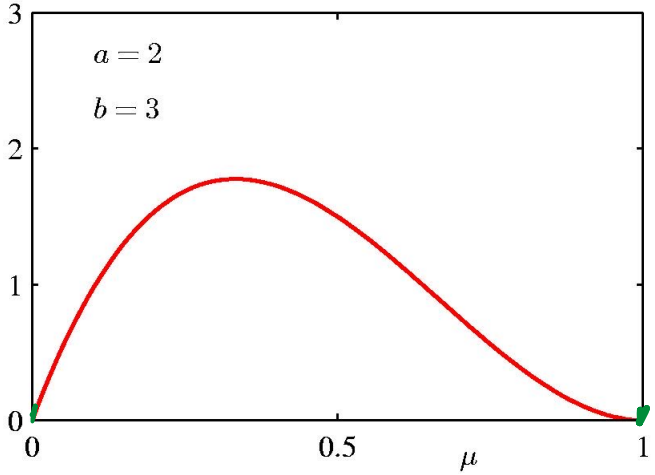
prob. param.

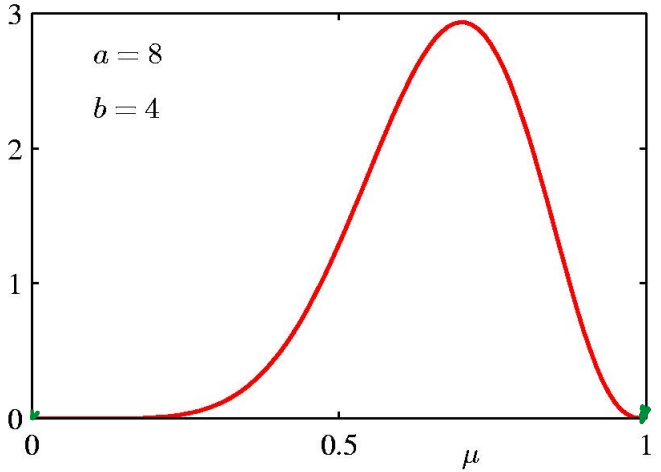
$$\mathbb{E}[\mu] = \frac{a}{a+b}$$

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









Parameter Estimation: Bayesian for Bernoulli

posterior likelihood prior

MAP Estimation

max

$$p(\mu|a_0, b_0, \mathcal{D}) \propto p(\mathcal{D}|\mu)p(\mu|a_0, b_0)$$

$$= \left(\prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}\right) \text{Beta}(\mu|a_0, b_0)$$

proportional

$$\propto \mu^{m+a_0-1}(1-\mu)^{(N-m)+b_0-1}$$

$$\propto \text{Beta}(\mu|a_N, b_N)$$

updated # success

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

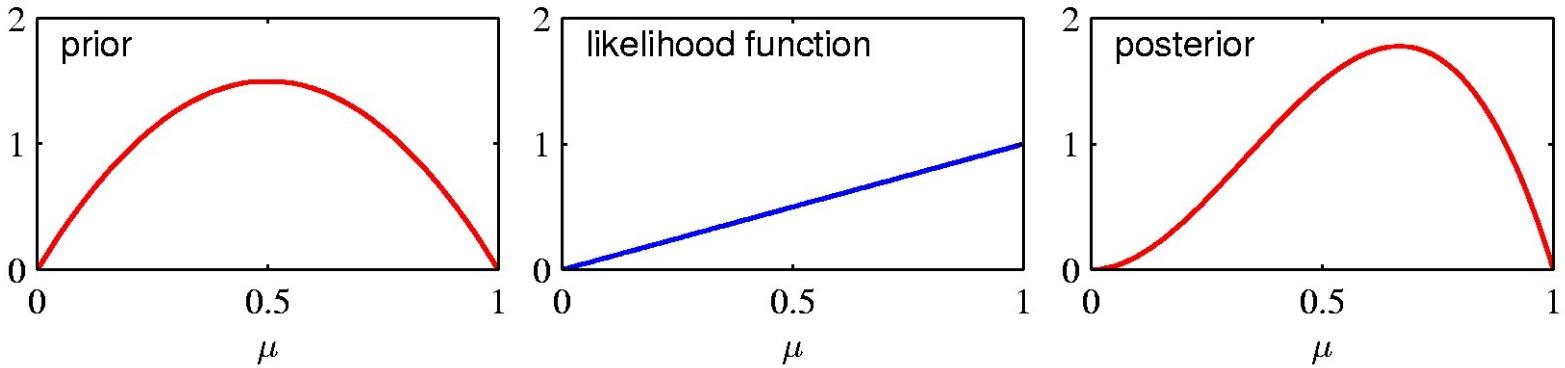
a_0 = # initial success
b_0 = # initial failures

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

Beta prior → (conjugate prior for Bernoulli likelihood) Beta posterior

Gaus. prior $N(\mu|\mu_0, \sigma_0^2)$ → (conjugate prior for Gaus. mean $N(\mu, \sigma^2)$) Gaus. posterior $N(\mu|\mu_N, \sigma_N^2)$

Prior · Likelihood = Posterior



a=2, b=2
Beta

Beta(2,2)

N=m=1

↑

Data

a=3, b=2
Beta

$a = a_0 + m = 2+1 = 3$

$b = b_0 + (N-m) = 2$

Properties of the Beta Posterior

As the size of the data set, N, increases

$$
\begin{aligned}
a_N &\rightarrow m \\
b_N &\rightarrow N - m \\
\mathbb{E}[\mu] &= \frac{a_N}{a_N + b_N} \rightarrow \frac{m}{N} = \mu_{\mathrm{ML}} \\
\mathrm{var}[\mu] &= \frac{a_N b_N}{(a_N + b_N)^2 (a_N + b_N + 1)} \rightarrow 0
\end{aligned}
$$

$a_N = a_0 + m$

$\lim_{m \to \infty} a_N = m$

as $N \to \infty$

MAP est. $\longrightarrow$ ML est.

Empirical Bayes

- In full Bayes a probability distribution is assumed/known for hyperparameters (a_0, b_0)

- In empirical Bayes, hyperparameters (a_0, b_0) are estimated from data
- Example:

Consider a baseball dataset with the total number of batting attempts A_i and the total number of hits H_i given for each player i.

The ML estimate for the probability of hit is given by the batting average H_i/A_i

To obtain a MAP estimate if you assume a Beta prior on the probability of hit, then you will either need to assume values for hyperparameters (a_0, b_0) such as $(100, 300)$ or estimate these values from the dataset.

Categorical Distribution

Bernoulli → multi-choice (Not Binary)

choices → 2

choices

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

one-hot encoding

K-1 prob. parameters

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

degree of freedom is k-1
P(the last one is): 1 - P(the rest)

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

1 experiment

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

↑ K-dim. vector

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

Binomial → Multi-choice

Multinomial Distribution

Multi-choice
Multi-exper.

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

N: # total exper.

m_1: # exper. for choice 1

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

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

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

Parameter estimation: ML for Multinomial

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

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

$\max_{\mu} \log p(\mathcal{D}|\mu)$ s.t. constraint $\sum_{k=1}^{K}\mu_k = 1$

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

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

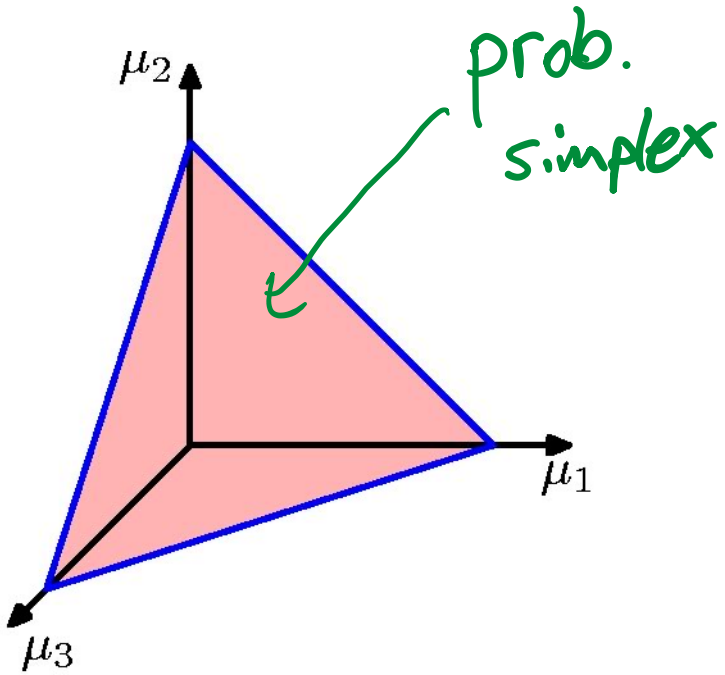
$$\mu_k = -m_k/\lambda \qquad \mu_k^{\mathrm{ML}} = \frac{m_k}{N}$$

Beta Dist. → # Dirichlet Distribution

hyper-param.

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

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



Conjugate prior for the **multinomial** distribution.

Dirichlet prior → multinomial likelihood → Dirichlet posterior

Parameter estimation: Bayesian for Multinomial

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

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

$\alpha_{k,N} = \alpha_k + m_k$

m_k: # exp. for choice k in data

α_k: hyper param. in Dirichlet for choice k

Exponential Family

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

- where η is the *natural parameter* and

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

- so g(η) can be interpreted as a normalization coefficient.

- For any member of the exponential family, there exists a ***conjugate prior,*** which makes the posterior the same distribution as itself

Nonparametric Methods

- 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.

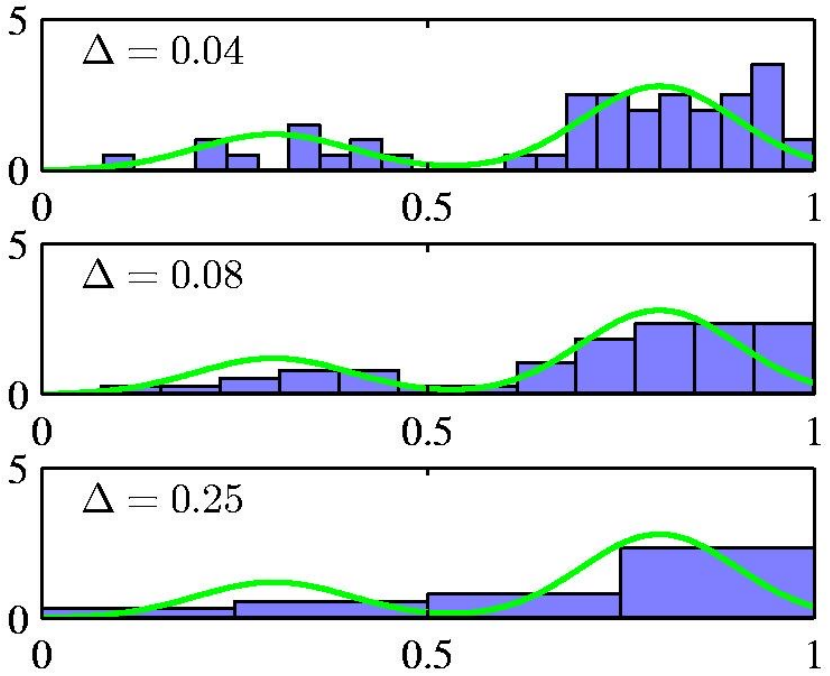
Histogram

Histogram methods partition the data space into distinct bins with widths Δ_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, $\Delta_i = \Delta$.

- Δ acts as a smoothing parameter.



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

Curse of dimensionality!

Kernel Methods

we are using Gaussian PDF or Gaussian curve not a Gaussian distribution.
We are using Gaussian curve/PDF as a mathematical tool here.

To avoid discontinuities in p(x), use a smooth kernel, e.g. a Gaussian

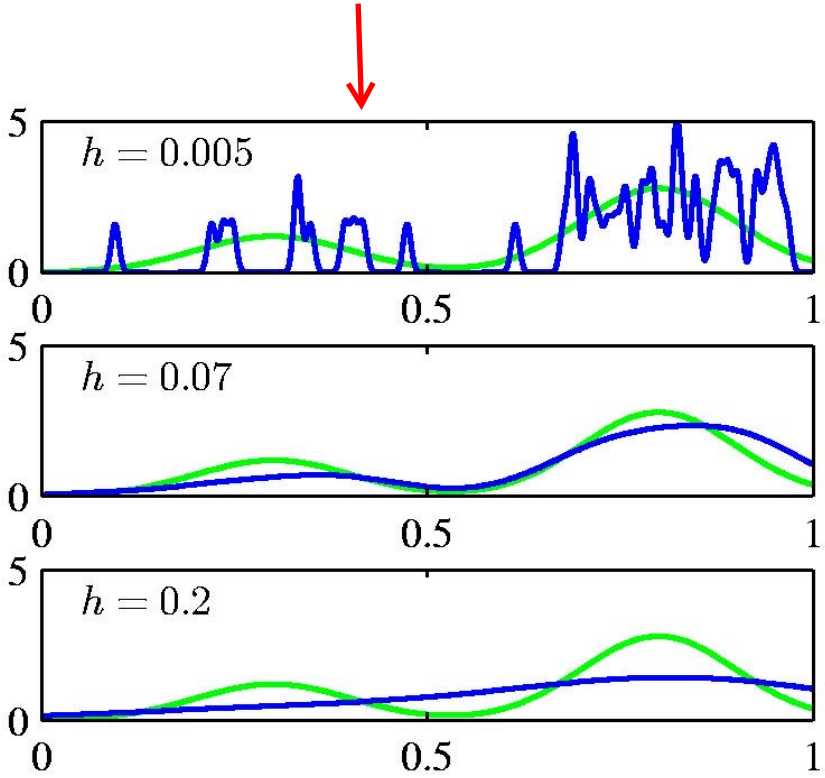
$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{(2\pi h^2)^{D/2}}\exp\left\{-\frac{\|\mathbf{x}-\mathbf{x}_n\|^2}{2h^2}\right\}$$

size of Dataset

Any kernel such that

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

will work.



h acts as a smoother.