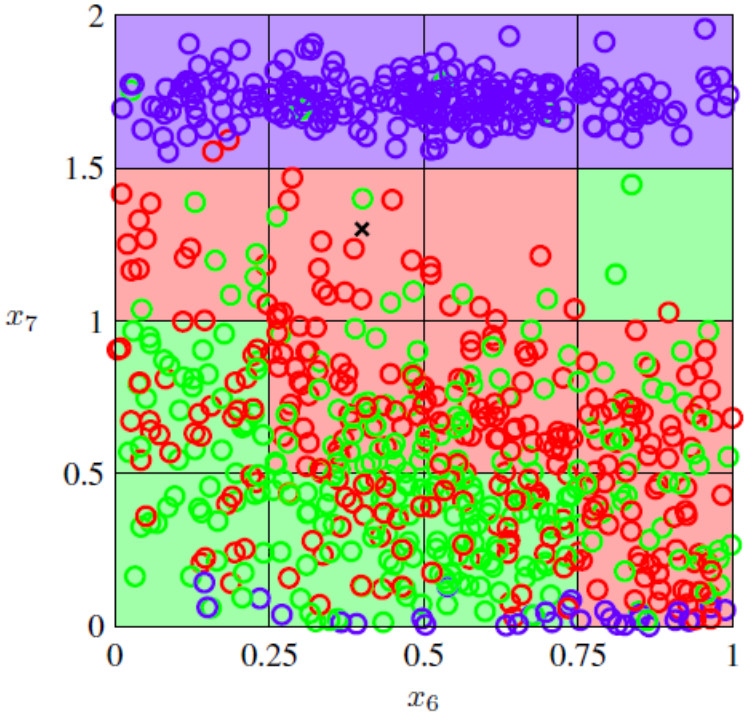
LUND UNIVERSITY

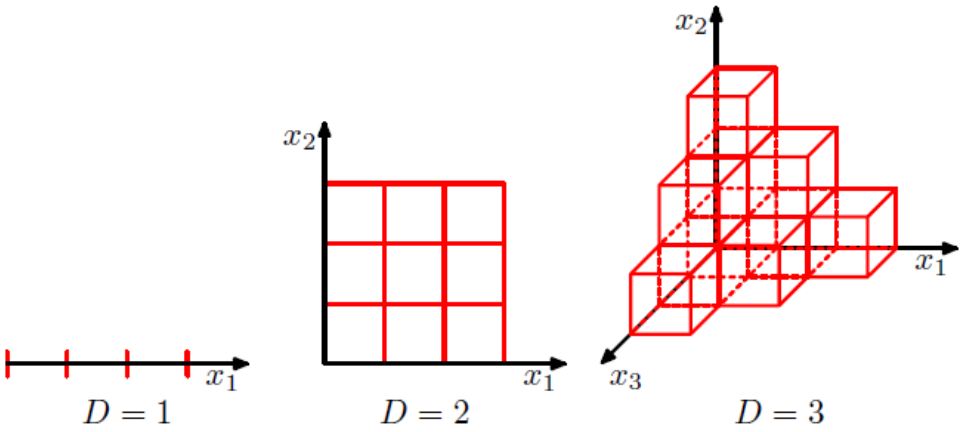
FMAN-45: Machine Learning

Lecture 2: Probability Distributions and Bayesian Modeling

Cristian Sminchisescu

Curse of Dimensionality





$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j + \sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{D} w_{ijk} x_i x_j x_k.$$

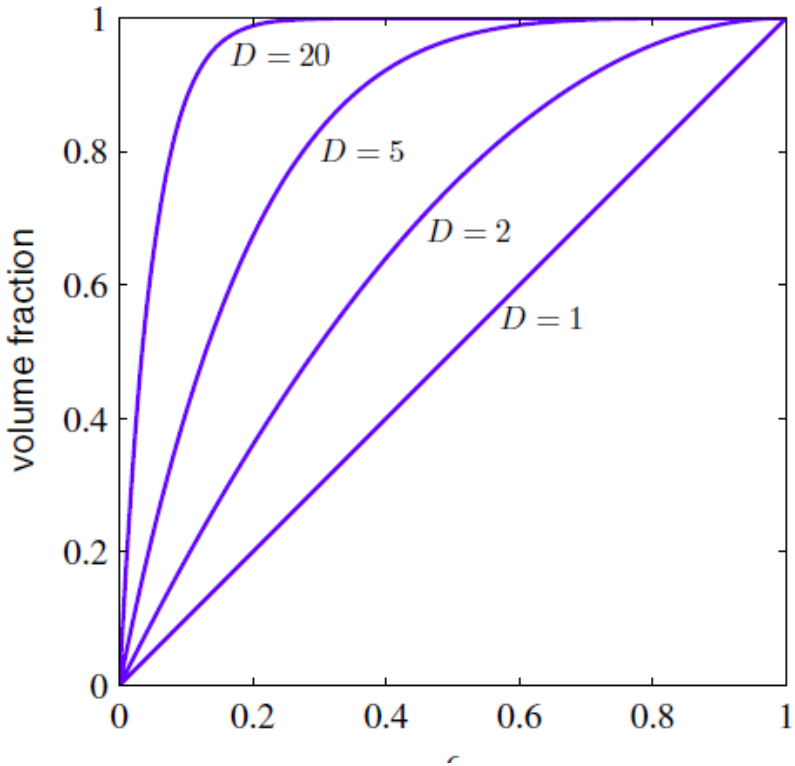
The severe computational and modeling difficulty that can arise when estimating models in spaces of many dimensions is sometimes called the *curse of dimensionality* (Bellman, 1961).

Curse of Dimensionality

$$V_D(r) = K_D r^D$$

$$\frac{V_D(1) - V_D(1-\epsilon)}{V_D(1)} = 1 - (1-\epsilon)^D$$

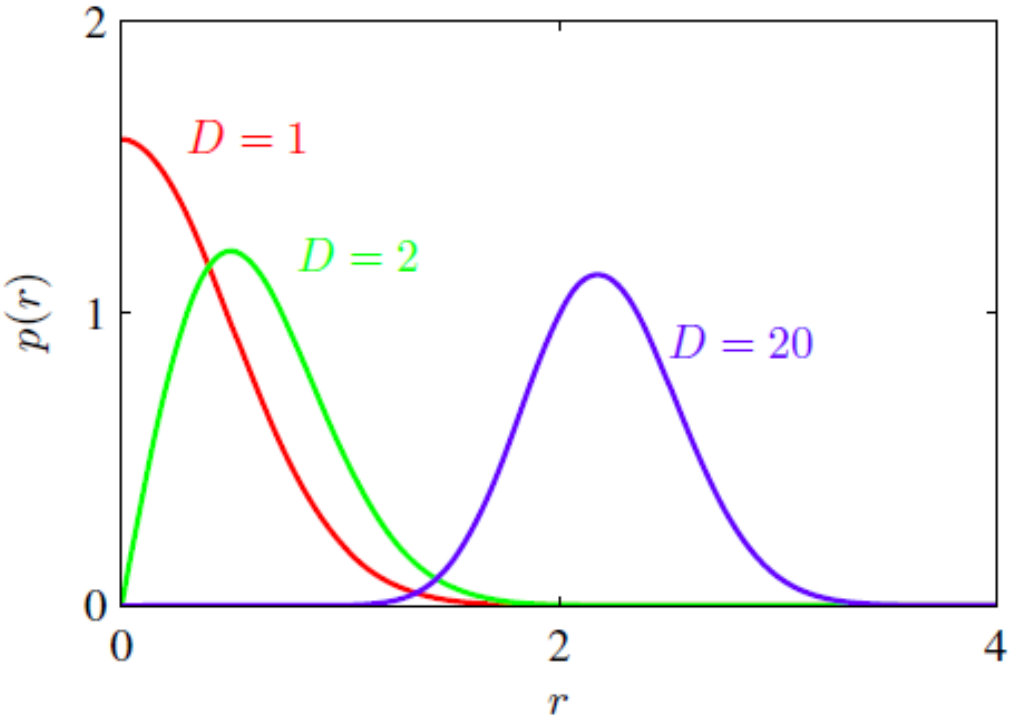
Plot of the fraction of the volume of a sphere lying in the range $r = 1-\epsilon$ to $r = 1$ for various values of the dimensionality D.



In spaces of high dimensionality, most of the volume of a sphere is concentrated in a thin shell near the surface!

Curse of Dimensionality

Plot of the probability density with respect to radius r of a Gaussian distribution for various values of the dimensionality D. In a high-dimensional space, most of the probability mass of a Gaussian is located within a thin shell at a specific radius.



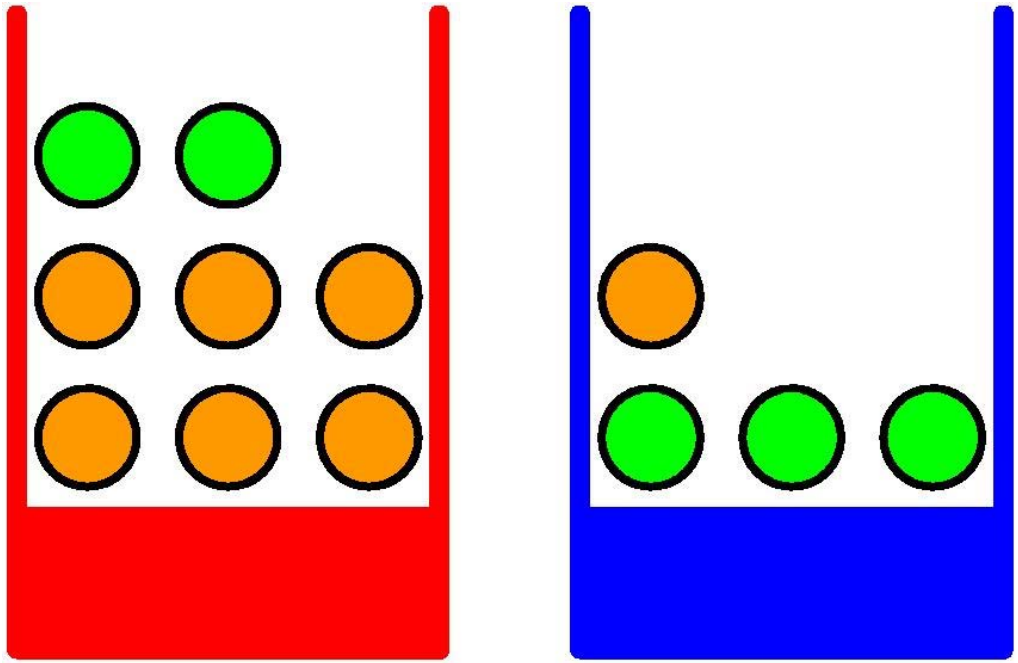
Probability and Decision Theory

- **Uncertainty** is a key concept in pattern recognition and machine learning
- It arises both from *measurement noise* and from *finite size datasets*
- **Probability theory** provides consistent framework for the quantification and manipulation of uncertainty
- When combined with **decision theory**, it allows us to make optimal predictions, given all the information available, even when that information is incomplete or ambiguous

Example of World with Probabilities

We have two boxes of Apples and Oranges (distributions shown)
We randomly pick box, then fruit, with replacement

Random variables:
- Box *B= {r,b}*
- Fruit *F={a,o}*



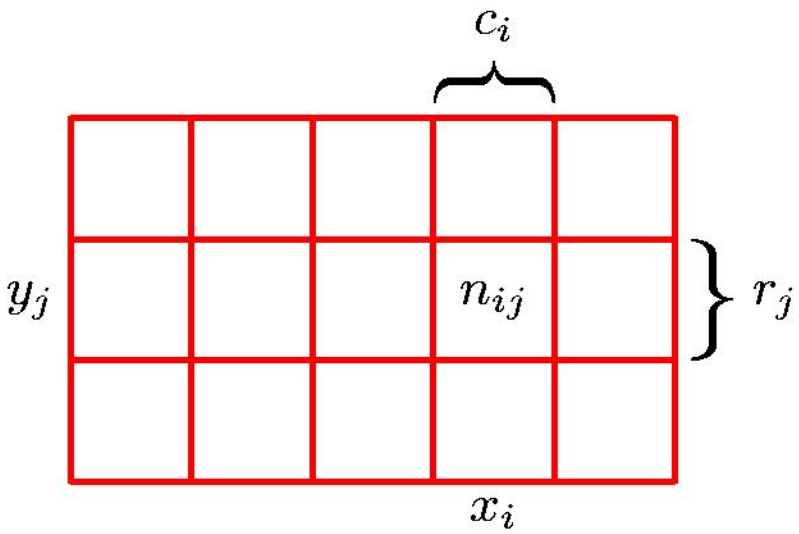
Modeling with Probabilities

- Quantities of interest in our problem are modeled as random variables

- To start with, we will define the probability of an event to be the fraction of times that event occurs, in the limit that the total number of trials goes to infinity

- Using elementary *sum and product rules of probability*, we can ask fairly sophisticated questions in our problem domain
  - Given that we chose an orange, what is the probability that the box we chose was a blue one?
  - What is the overall probability that the selection procedure will pick an apple?

Probability Theory



Marginal Probability

$$p(X = x_i) = \frac{c_i}{N}.$$

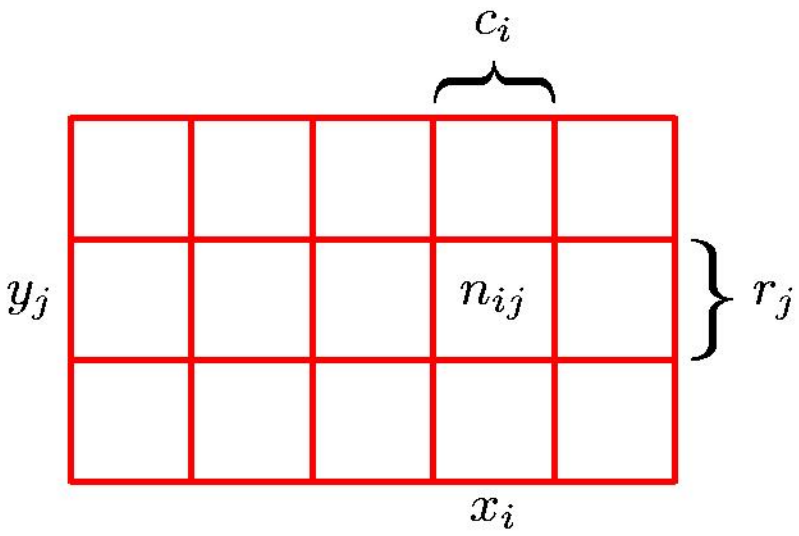
Joint Probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Conditional Probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Probability Theory



Sum Rule

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N} \sum_{j=1}^{L} n_{ij}$$

$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

Product Rule

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$

$$= p(Y = y_j | X = x_i) p(X = x_i)$$

The Rules of Probability

Sum Rule $$p(X) = \sum_{Y} p(X, Y)$$

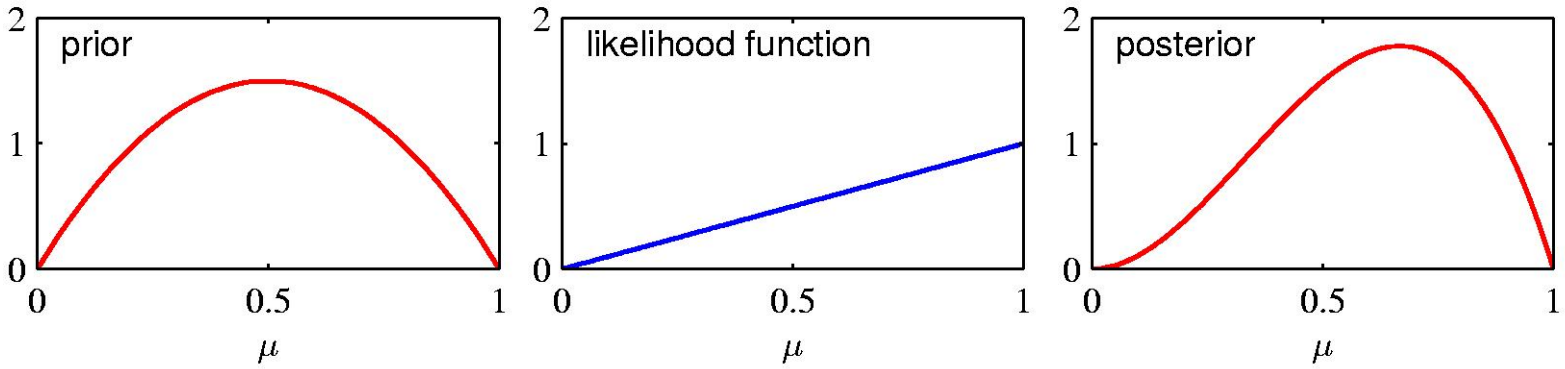
Product Rule $$p(X, Y) = p(Y|X)p(X)$$

Bayes' Theorem

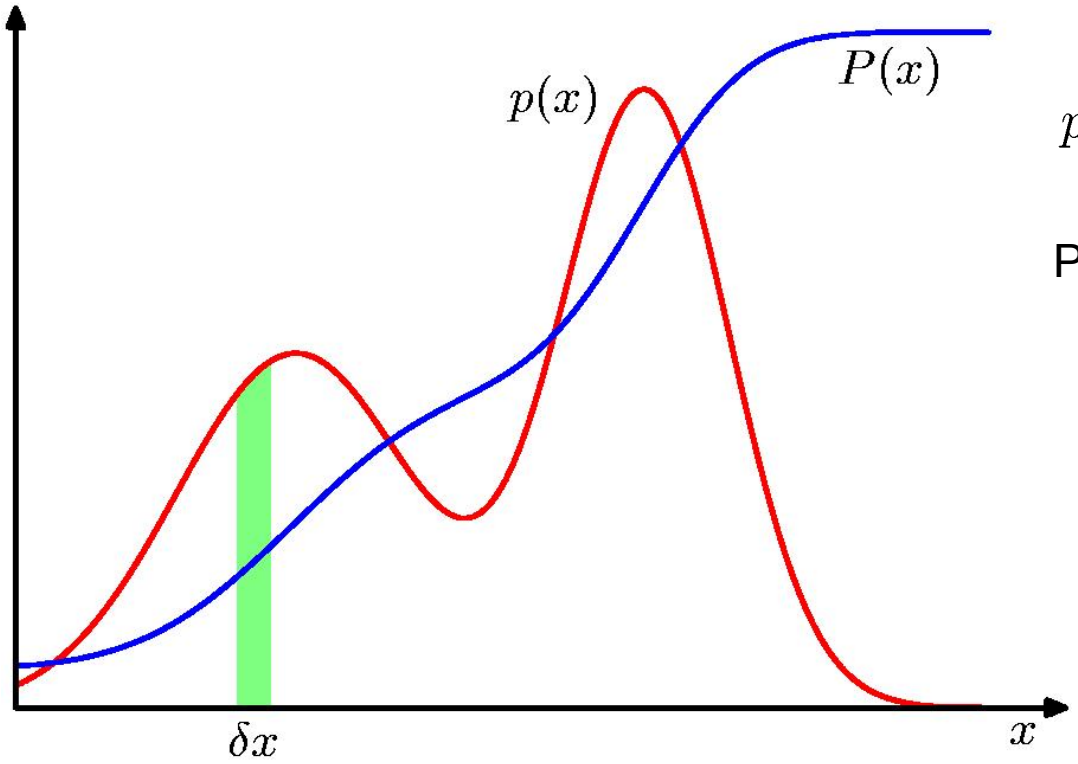
$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

$$p(X) = \sum_{Y} p(X|Y)p(Y)$$

posterior $\propto$ likelihood × prior



Probability Densities



$$p(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x$$

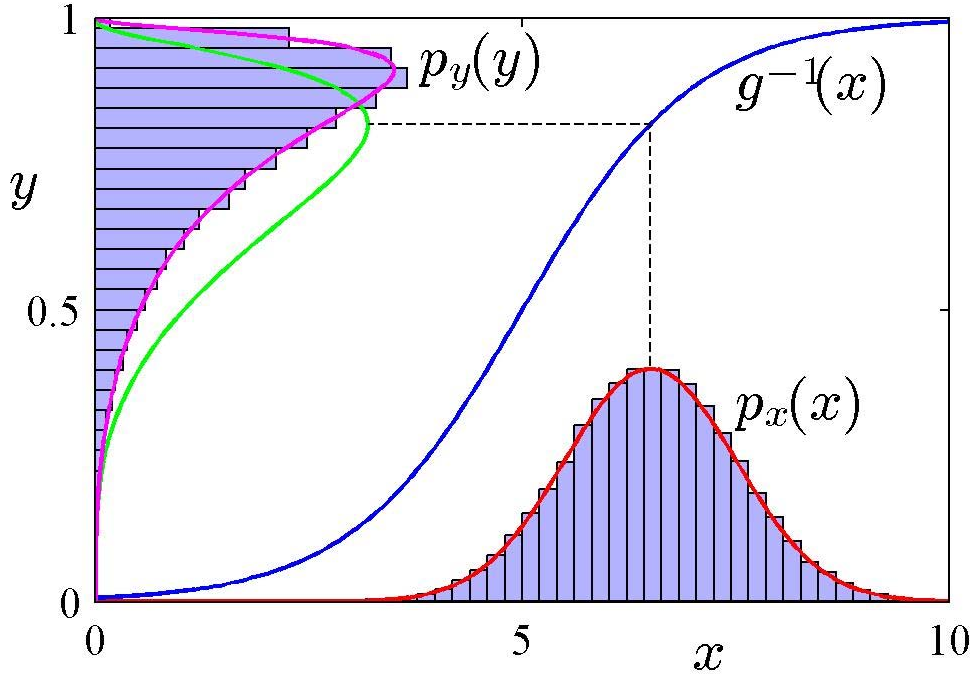
Probability density, $\delta x \to 0$

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

Cumulative distribution function, $P'(x) = p(x)$

$$p(x) \geqslant 0 \qquad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

Transformed Densities



$$p_x(x)\delta x \simeq p_y(y)\delta y,$$

$$\begin{aligned} p_y(y) &= p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| \\ &= p_x(g(y))\,|g'(y)| \end{aligned}$$

The maximum of a probability density is dependent of the choice of variable
The maxima will change, in general, under a non-linear change of variable

Expectations

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

$$\mathbb{E}[f] = \int p(x) f(x)\, \mathrm{d}x$$

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$$

Conditional Expectation (discrete)

average of the function $f(x)$ with respect to the distribution of x

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

Approximate Expectation (discrete and continuous)

Variances and Covariances

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

$$\begin{aligned}\mathrm{cov}[x, y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

$$\begin{aligned}\mathrm{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}]\end{aligned}$$

$$\begin{aligned}\mathrm{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] \\ &= \mathbb{E}[xy - x\mathbb{E}[y] - \mathbb{E}[x]y + \mathbb{E}[x]\mathbb{E}[y]] \\ &= \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y] - \mathbb{E}[x]\mathbb{E}[y] + \mathbb{E}[x]\mathbb{E}[y] \\ &= \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

Frequentist vs. Bayesian

- *Frequentists* view probabilities in terms of frequencies of random, repeatable events (classical of frequentist interpretation of probability)

- In contrast, for *Bayesians*, probabilities provide a quantification of uncertainty

- Using probability to represent uncertainty is not ad-hoc. Cox (1946) showed that if numerical values are used to represent degrees of belief, a simple set of axioms encoding such beliefs leads uniquely to a set of rules that are equivalent with the sum and product rules of probability

- Probability can be regarded as an extension of Boolean logic to situations involving uncertainty (Jaynes, 2003)

Learning: General Objective Functions

- The general structure of our learning objective function is

$$f(x, t, w) = L(x, t, w) + R(w)$$

  L is the loss function, and R is a regularizer (penalty, or prior over functions), which discourages overly complex models

- **Intuition**
  - It is good to fit the data well and achieve low training loss
  - But it is also good to bias the machine towards simpler models, in order to avoid overfitting

- Setup allows to decouple optimization from choice of training loss

Maximize Probabilities or their Log?

- We have a statistical model of outputs
- Assume the output errors on different training cases, i, are independent
- We will consider the product of the probabilities of the outputs on the training cases

$$p(\mathrm{t} \mid \mathrm{x}, w) = \prod_i p(t_i \mid x_i, w)$$

- Because the log function is monotonic, it does not change where the maxima are. Therefore, we can maximize the sum of log probabilities, or minimize negative log probabilities

$$L(\mathrm{x}, \mathrm{t}, w) = -\log p(\mathrm{t}, \mathrm{x} \mid w) = -\sum_i \log p(t_i \mid x_i, w)$$

Frequentist: General Objective Functions

- The general structure of our learning objective function is

$$f(x, t, w) = L(x, t, w) + R(w)$$

L is the loss function, and R is a regularizer (penalty, or prior over functions), which discourages overly complex models

- If we assume that all parameter configurations are equally likely (otherwise said, the prior over w is uniform, in the selected parameterization), we obtain Maximum Likelihood (ML). This selects model parameters that assign the highest probability to observed data
- If we assume a prior over models or parameters, we obtain a regularized ML estimate, the Maximum a-posteriori estimate (MAP)

Learning Setup: Bayes' Theorem

joint probability

conditional probability

$$p(d)\,p(w \mid d) = p(d, w) = p(w)\,p(d \mid w)$$

Prior probability of weight vector w

Probability of observed data given w

$$p(w \mid d) = \frac{p(w)\,p(d \mid w)}{p(d)}$$

Posterior probability of weight vector w given training data

$$\int_w p(w)\,p(d \mid w)$$

$d = \{(x_i, t_i), i = 1 \ldots N\}$

Learning in a Bayesian Framework

The Bayesian framework assumes that we have a prior distribution over all variables of interest, and our goals is to compute probability distributions, not point estimates

- The prior may be vague
- We combine our prior distribution with the data likelihood term to construct the posterior distribution
- The likelihood accounts for how probable the observed data is given the model parameters
  - It favors parameter that make the data likely
  - It counteracts the prior
  - With enough data, the likelihood dominates

Machine Learning: Frequentist vs. Bayesian

- In the frequentist setting, *w* is considered a fixed parameter, with valued determined by an estimator (ML, MAP, etc.)
- Uncertainty (error bars) for the estimate is obtained by considering the distribution of possible datasets, e.g. by bootstrap - sampling the original data with replacement
- In Bayesian perspective, there is a single dataset, the one actually observed, and uncertainty in parameters is expressed through a probability distribution over *w*
- Issues with Bayesian methods
  - Priors often selected based on mathematical convenience rather than prior beliefs
  - Models with poor choices of priors can give inferior results with high confidence
  - Reducing dependence on priors is one motivation for *non-informative priors*

Decision Theory

- We have seen that probability theory provides a consistent mathematical framework for quantifying and manipulating uncertainty
- We will now discuss decision theory
- When combined with probability theory, decision theory allows us to make optimal decisions in situations involving uncertainty

Decision Theory

Inference step

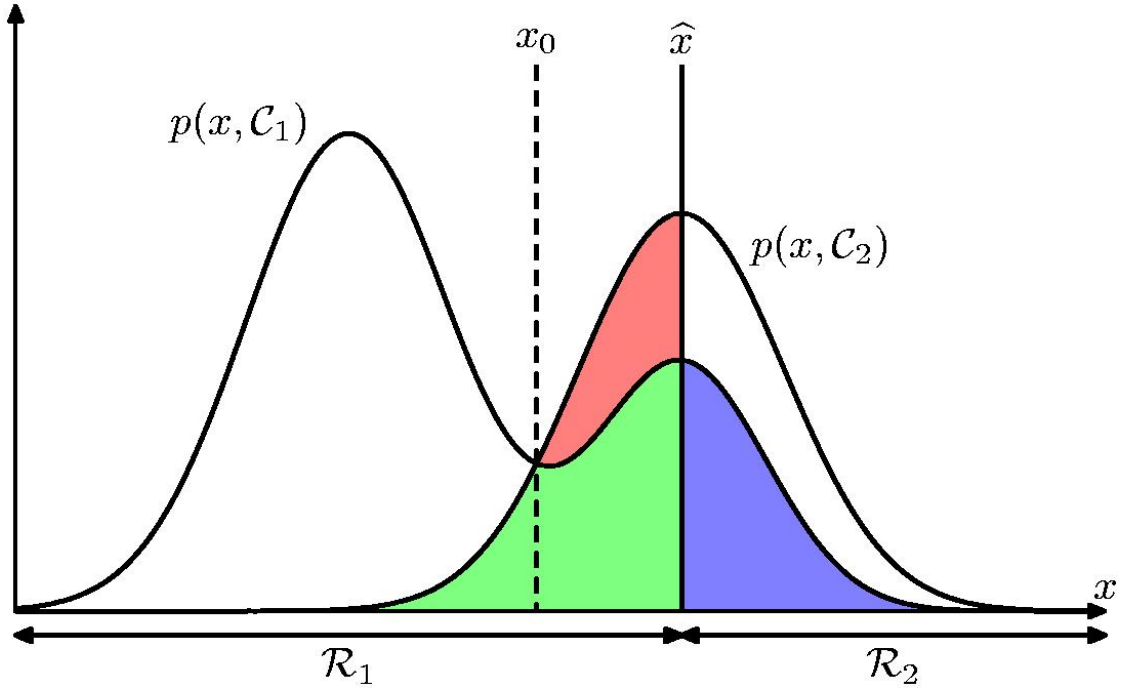
Determine $p(t, \boldsymbol{x})$ or $p(t|\boldsymbol{x})$ from a set of training data

Decision step

For given x, determine optimal t

Decision is often easy, once we solved the inference problem

Minimum Misclassification Rate



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

- *Decision regions* R_i (not necessarily contiguous) are associated to each class
- Boundaries between decision regions are called *decision surfaces*

$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\, \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\, \mathrm{d}\mathbf{x}. \end{aligned}$$

x should be assigned to class having the largest posterior $p(C_k|x)$

Minimum Expected Loss

Example: classify medical images as ‘cancer’ or ‘normal’

$$Matrix\ L_{kj} \qquad \begin{array}{c} \\ \text{Truth} \end{array} \begin{array}{l} \\ \text{cancer} \\ \text{normal} \end{array} \overset{\text{Decision}}{\begin{pmatrix} \text{cancer} & \text{normal} \\ 0 & 1000 \\ 1 & 0 \end{pmatrix}}$$

Sometimes we don’t just want to minimize the number of miss-classifications, but also take into account how important they are

Missing a cancer diagnostic is more consequential than a false positive

Minimum Expected Loss

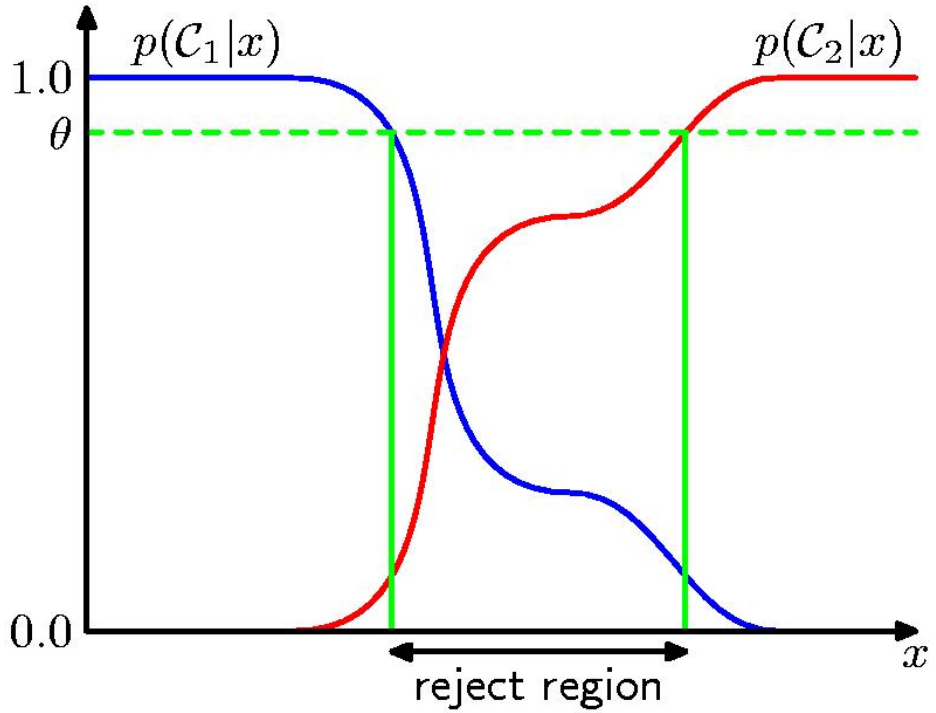
$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k)\, \mathrm{d}\mathbf{x}$$

Regions $\mathcal{R}_j$ are chosen to minimize

$$E_j[L] = \mathbb{E}[L] = \sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x})$$

Given new $\boldsymbol{x}$, we pick the class j for which the expected loss $E_j[L]$ is the smallest. Trivial once we know the posterior class probabilities.

Reject Option



E.g. use an automatic system to classify X-ray images for which there is little doubt as to the correct class, while leaving a human expert to classify the more ambiguous cases

Avoid making decisions on difficult cases

Reject inputs x, for which

$\max_k p(C_k|x) \leq \epsilon$; Notice that for $\epsilon = 1$ (all reject), $\epsilon < \frac{1}{K}$ (no reject)

Generative vs. Discriminative

Generative approach:

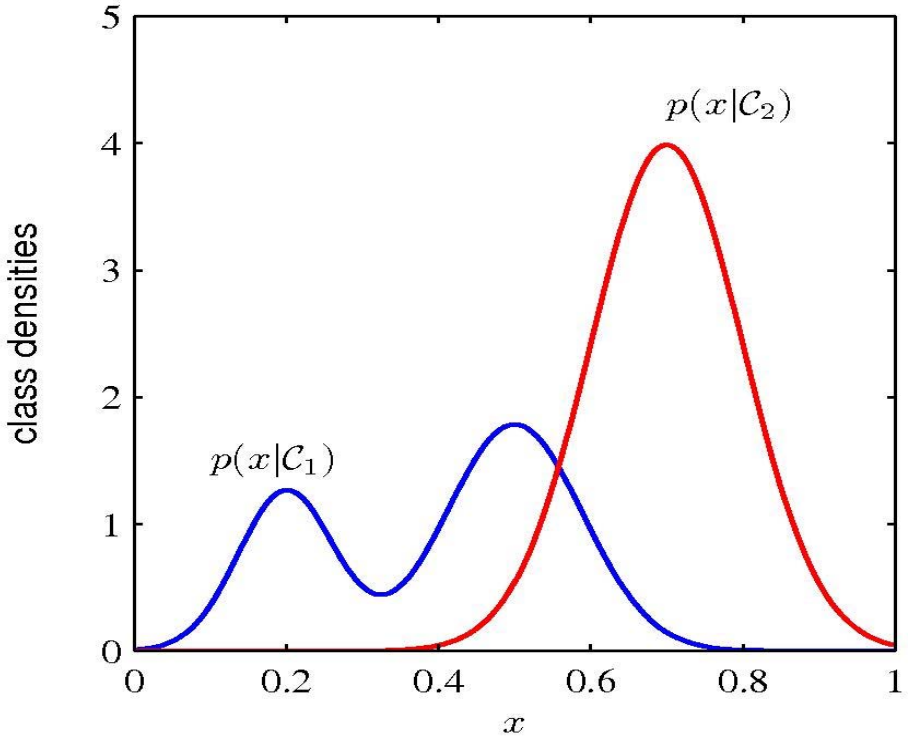
Model $p(t, \mathbf{x}) = p(\mathbf{x}|t)p(t)$

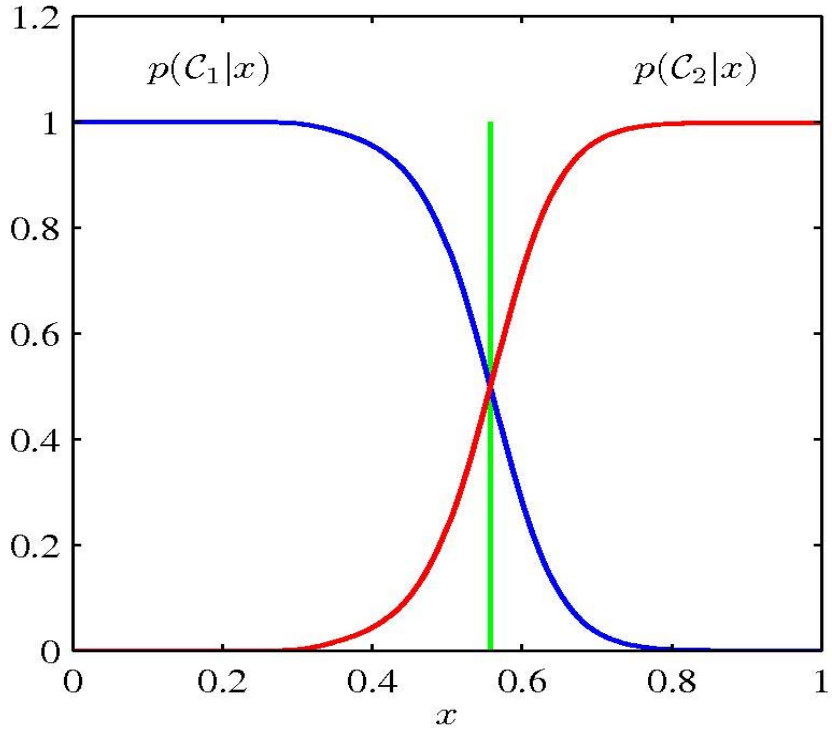
Use Bayes' theorem $p(t|\mathbf{x}) = \frac{p(\mathbf{x}|t)p(t)}{p(\mathbf{x})}$

Discriminative approach:

Model $p(t|\mathbf{x})$ directly, or $f(t|\boldsymbol{x})$

Generative vs. Discriminative





Assume $p(C_1) = p(C_2)$

Why Separate Inference and Decision?

- Minimizing risk (loss matrix may change over time)
- Reject option
- Unbalanced class priors
- Combining models

Decision Theory for Regression

Inference step

Determine $p(\mathbf{x}, t)$

Decision step

For given x, make optimal prediction $y(x)$ for t

Loss function:

$$\mathbb{E}[L] = \iint L(t, y(\mathbf{x}))p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

The Squared Loss Function

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t$$

$$\begin{aligned}\{y(\mathbf{x}) - t\}^2 &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}] + \mathbb{E}[t|\mathbf{x}] - t\}^2 \\ &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 + 2\{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} + \{\mathbb{E}[t|\mathbf{x}] - t\}^2\end{aligned}$$

substitute and integrate over t, w. r. t $p(t|x)$

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x})\, \mathrm{d}\mathbf{x} + \int \mathrm{var}\,[t|\mathbf{x}]\, p(\mathbf{x})\, \mathrm{d}\mathbf{x}$$

$$y(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] \quad \text{where } E[t|\boldsymbol{x}] \equiv E_t[t|\boldsymbol{x}] = \int tp(t|\boldsymbol{x})dt$$

regression function

- Optimal least squares predictor given by the conditional average
- First term in $E[L]$ gives prediction error according to conditional mean estimates
- Second term is independent of $y(x)$. It measures the intrinsic variability of the target data, hence it represents the irreducible minimum value of loss, independent of $y(x)$

$$\frac{\delta \mathbb{E}[L]}{\delta y(\mathbf{x})} = 2\int \{y(\mathbf{x}) - t\} p(\mathbf{x}, t)\, \mathrm{d}t = 0. \quad y(\mathrm{x}) = \frac{\int tp(\mathrm{x}, t)\, \mathrm{d}t}{p(\mathrm{x})} = \int tp(t|\mathrm{x})\, \mathrm{d}t = \mathbb{E}_t[t|\mathrm{x}]$$

Parametric Distributions

We will study several important distributions governed by a small number of adaptive parameters

The Exponential Family (1)

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

where η is the *natural parameter* and

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

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

ML for the Exponential Family (1)

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

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

Thus

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

Covariance of $\mathbf{u}(\mathbf{x})$ can be expressed in terms of the second derivative of $g(\boldsymbol{\eta})$, and similarly for higher order moments. Provided we can normalize a distribution from the exponential family, we can always find its moments by simple differentiation.

ML for the Exponential Family (2)

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

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

Taking the log of p and setting the derivative w.r.t. η to 0, we get

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

Sufficient statistic

We do not need to store the entire dataset itself, but only the value of the sufficient statistic

Sufficient Statistics

- A statistic satisfies the criterion of **sufficiency** when no other statistic which can be calculated from the same sample provides any additional information as to the value of the parameter to be estimated (R. Fisher)
- This is equivalent to the view that the distribution of a sample is independent of the underlying parameter(s) the statistic is sufficient for, conditional on the value of the sufficient statistic
- Both the statistic and the underlying parameters can be vectors
- In our case: a statistic $\mathbf{u}(\mathbf{x})$ is **sufficient for the underlying parameter $\boldsymbol{\eta}$** if the conditional probability distribution of the data $\mathbf{x}$, given the statistic $\mathbf{u}(\mathbf{x})$, is independent of the parameter $\boldsymbol{\eta}$

$$p(\mathbf{x}|\mathbf{u}(\mathbf{x}), \boldsymbol{\eta}) = p(\mathbf{x}|\mathbf{u}(\mathbf{x}))$$

Conjugate Priors

- If the posterior distributions are in the same family as the prior distributions, the prior and posterior are then called **conjugate distributions**
- The prior is called a **conjugate prior** for the likelihood
- Such priors lead to greatly simplified Bayesian analysis
- It can be useful to think of the hyper-parameters of a conjugate prior distribution as corresponding to having observed a certain number of **pseudo-observations** with properties specified by those parameters

Conjugate priors

For any member of the exponential family,

there exists a prior

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

Combining with the likelihood function, we get

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

Prior corresponds to ν pseudo-observations with value χ

Noninformative Priors

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

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

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

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

Bernoulli Distribution: Binary Variables

Coin flipping: heads=1, tails=0

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

Bernoulli Distribution

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

$$\begin{aligned} \sum_{x\in\{0,1\}} xp(x|\mu) &= 0.p(x=0|\mu) + 1.p(x=1|\mu) = \mu \\ \sum_{x\in\{0,1\}} (x-\mu)^2 p(x|\mu) &= \mu^2 p(x=0|\mu) + (1-\mu)^2 p(x=1|\mu) \\ &= \mu^2(1-\mu) + (1-\mu)^2\mu = \mu(1-\mu). \end{aligned}$$

Binomial Distribution: Binary Variables

For Ncoin flips (out of which only m heads):

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

Binomial Distribution

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

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

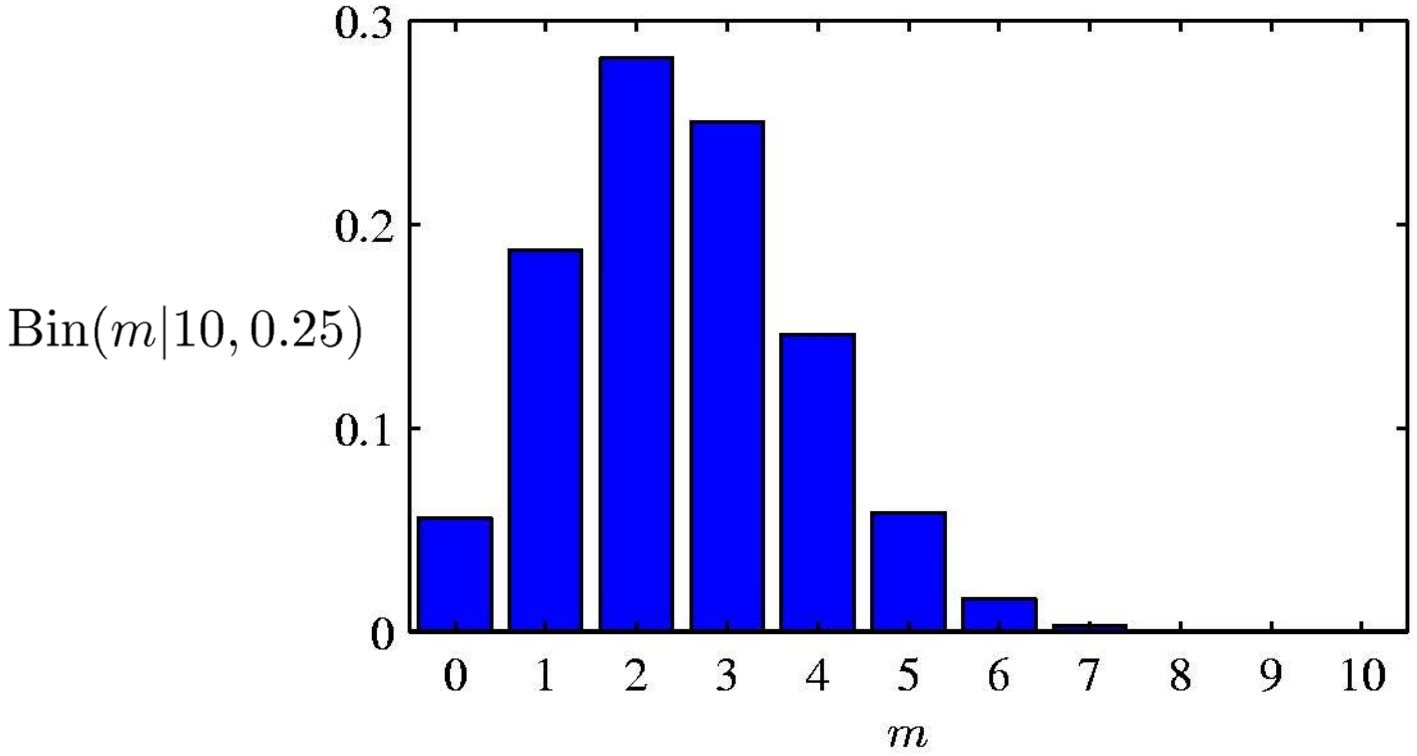
$$\text{var}[m] \equiv \sum_{m=0}^{N} (m - \mathbb{E}[m])^2 \text{Bin}(m|N, \mu) = N\mu(1-\mu)$$

For independent events

- the mean of the sum is the sum of the means
- the variance of the sum is the sum of variances

So we can re-use estimates derived for Bernoulli

Binomial Distribution



The Exponential Family (Bernoulli)

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

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

The Exponential Family (Bernoulli)

The Bernoulli distribution can hence be written as

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

where

$$\begin{aligned} u(x) &= x \\ h(x) &= 1 \\ g(\eta) &= 1 - \sigma(\eta) = \sigma(-\eta). \end{aligned}$$

$$\sigma(\eta) = \frac{1}{1 + \exp(-\eta)}$$

Parameter Estimation (ML Bernoulli)

ML for Bernoulli

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

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

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

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

Notice that log-likelihood depends on the N observations x_n only through their sum $\sum_{n=1}^{N} x_n$ (sufficient statistic)

Beta Distribution

Distribution over $\mu \in [0, 1]$, uses powers of μ and $1 - \mu$

$$\begin{aligned} \text{Beta}(\mu|a, b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\ \mathbb{E}[\mu] &= \frac{a}{a+b} \\ \text{var}[\mu] &= \frac{ab}{(a+b)^2(a+b+1)} \end{aligned}$$

where $\Gamma(x) = \int_0^\infty u^{x-1}e^{-x}du$, $\Gamma(x+1) = \text{x}\Gamma(x)$,

$\Gamma(1) = 1$, and $\Gamma(x+1) = x!$ for x integer

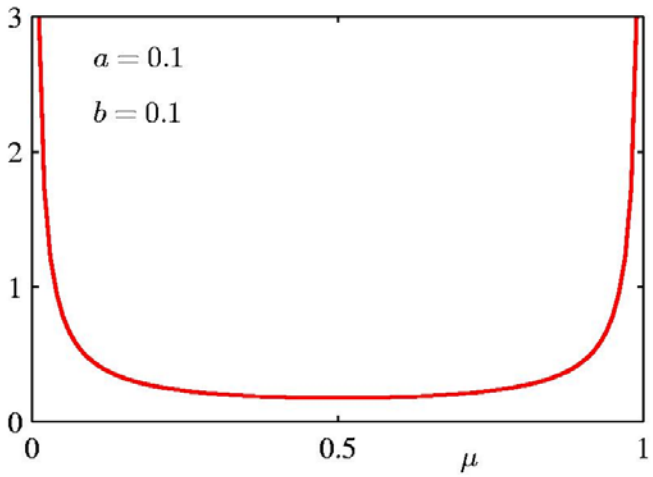
Bayesian Bernoulli

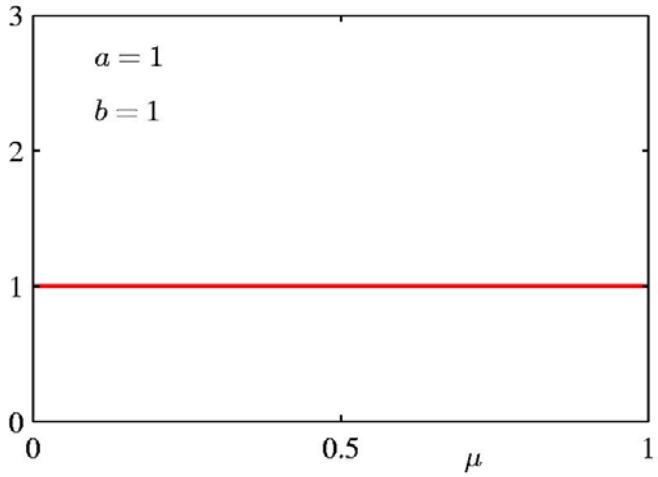
$$
\begin{aligned}
p(\mu|a_0, b_0, \mathcal{D}) &\propto p(\mathcal{D}|\mu)p(\mu|a_0, b_0) \\
&= \left(\prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n} \right) \mathrm{Beta}(\mu|a_0, b_0) \\
&\propto \mu^{m+a_0-1}(1-\mu)^{(N-m)+b_0-1} \\
&\propto \mathrm{Beta}(\mu|a_N, b_N)
\end{aligned}
$$

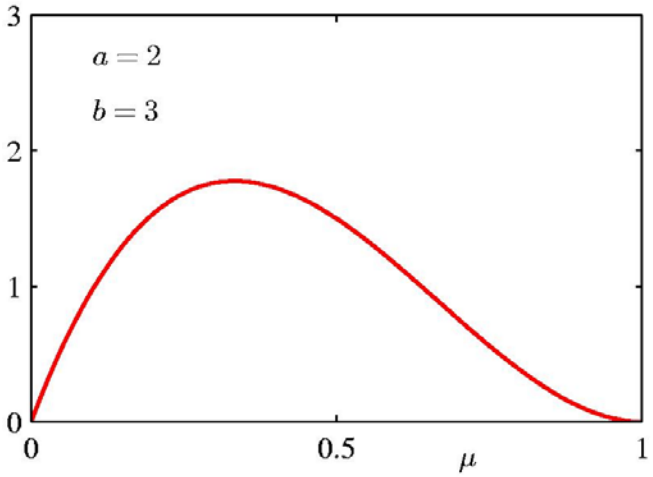
$$
a_N = a_0 + m \qquad b_N = b_0 + (N - m)
$$

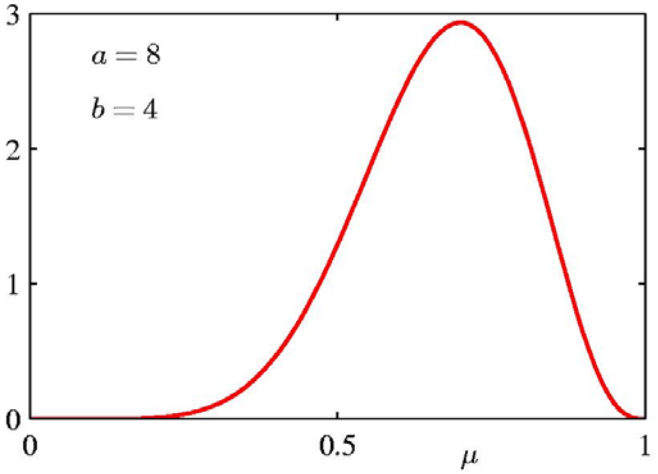
The Beta distribution provides the *conjugate* prior for the Bernoulli distribution.

Beta Distribution









Posterior more sharply peaked as the effective number of observations increase

Properties of the Posterior

As the size of the data set, N, increases

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

Prediction under the Posterior

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

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

Interpretation as total fraction of observations (both real and pseudo-observations induced by the prior) that correspond to $x = 1$

For a finite dataset, the posterior always lies between the prior mean μ and the maximum likelihood estimate for μ corresponding to the relative frequencies of events

Parameter Estimation (ML overfitting)

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

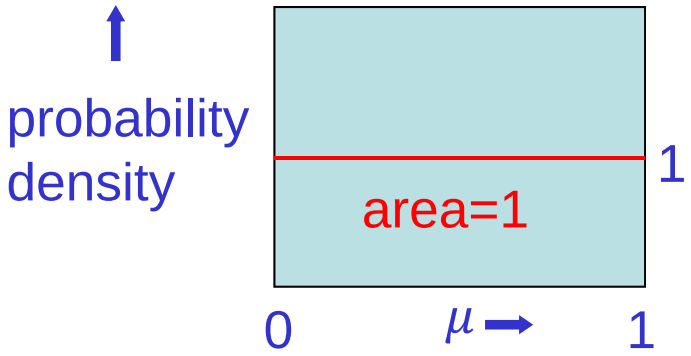
Prediction: *all* future tosses will land heads up

Overfitting to D

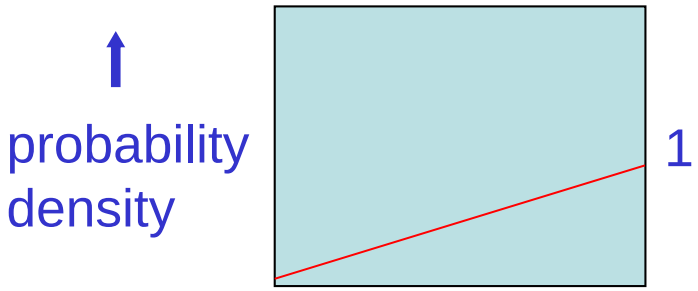
μ=0.5 could be a better answer. If we don't have much data, we are unsure about μ. Our computations of probabilities will work much better if we take uncertainty into account.

The Bayesian approach, using a prior distribution over parameter values

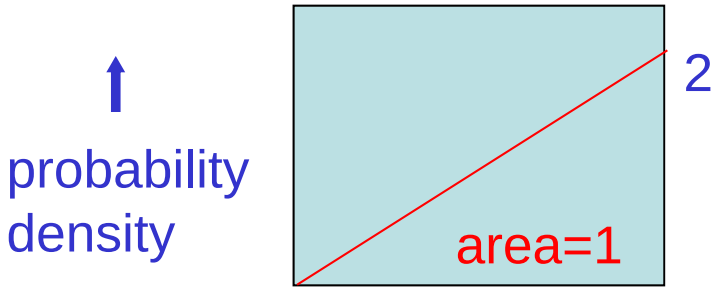
- Start with a prior distribution over μ. In this case we used a uniform distribution.



- Multiply the prior probability of each parameter value by the likelihood of observing a **head** given that value, Bin(1|1, μ)



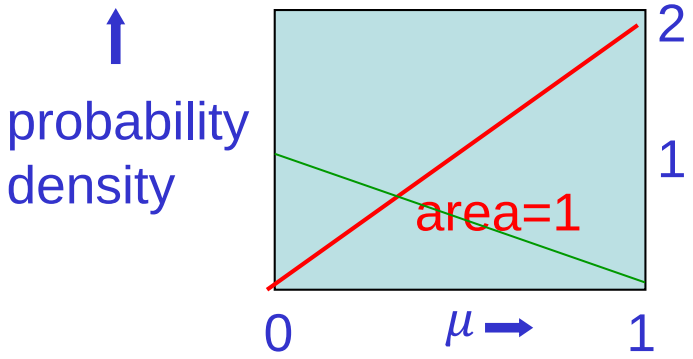
- Then scale up all of the probability densities so that their integral comes to 1. This gives the posterior distribution.



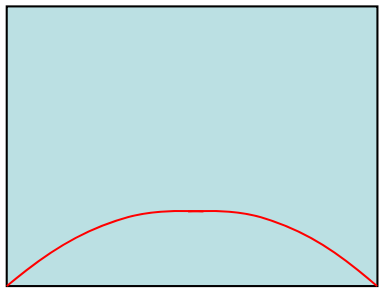
Adapted from MLG@UofT

Lets do it again: Suppose we get a tail

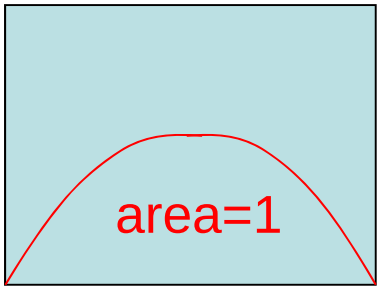
- Start with a prior distribution over μ



- Multiply the prior probability of each parameter value by the probability of observing a **tail** given that value, Bin(1|2, μ)

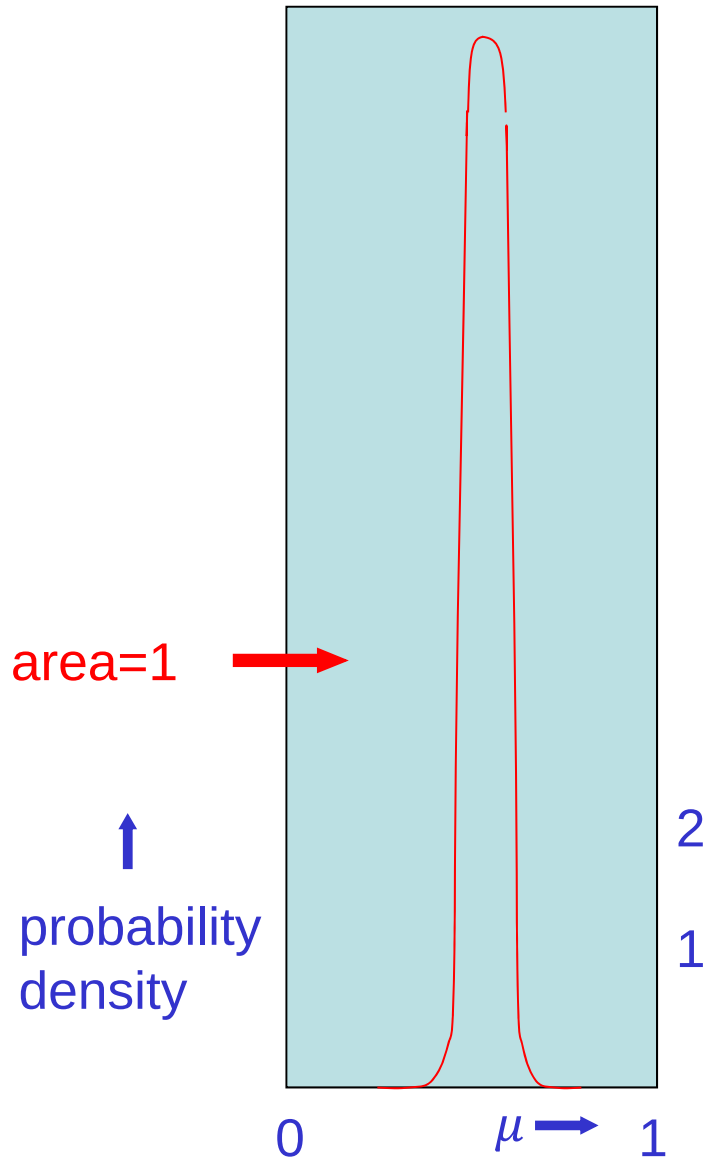


- Then renormalize to get the posterior distribution.



Lets do it another 98 times

After 53 heads and 47 tails we get a sensible posterior distribution that has its peak at 0.53 (assuming a uniform prior)



Multinomial Variables

Generalization of Bernoulli to K outcomes (not just 2)

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

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

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

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

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

ML Parameter estimation

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

Sufficient statistics

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

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

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

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

$$\lambda = -N, \text{from constaint} \sum_k \mu_k = 1$$

The Multinomial Distribution

$$\frac{N!}{m_1!\, m_2! \ldots m_K!} \text{ and } \sum_{k=1}^{K} m_k = N$$

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

The Dirichlet Distribution

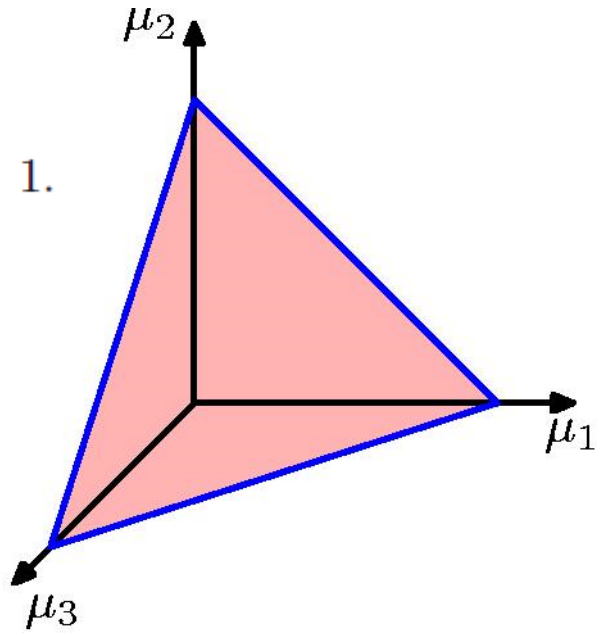
Multivariate generalization of the Beta distribution: its probability density function returns the belief that the probabilities of K rival events are μ_k given that each event has been observed α_k times.

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

$$\alpha_0 = \sum_{k=1}^{K} \alpha_k \qquad 0 \leqslant \mu_k \leqslant 1 \text{ and } \textstyle\sum_k \mu_k = 1.$$

Conjugate prior for the multinomial distribution

Confined to simplex (bounded linear manifold) of dimensionality $K - 1$ due to summation constraint



Bayesian Multinomial (1)

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

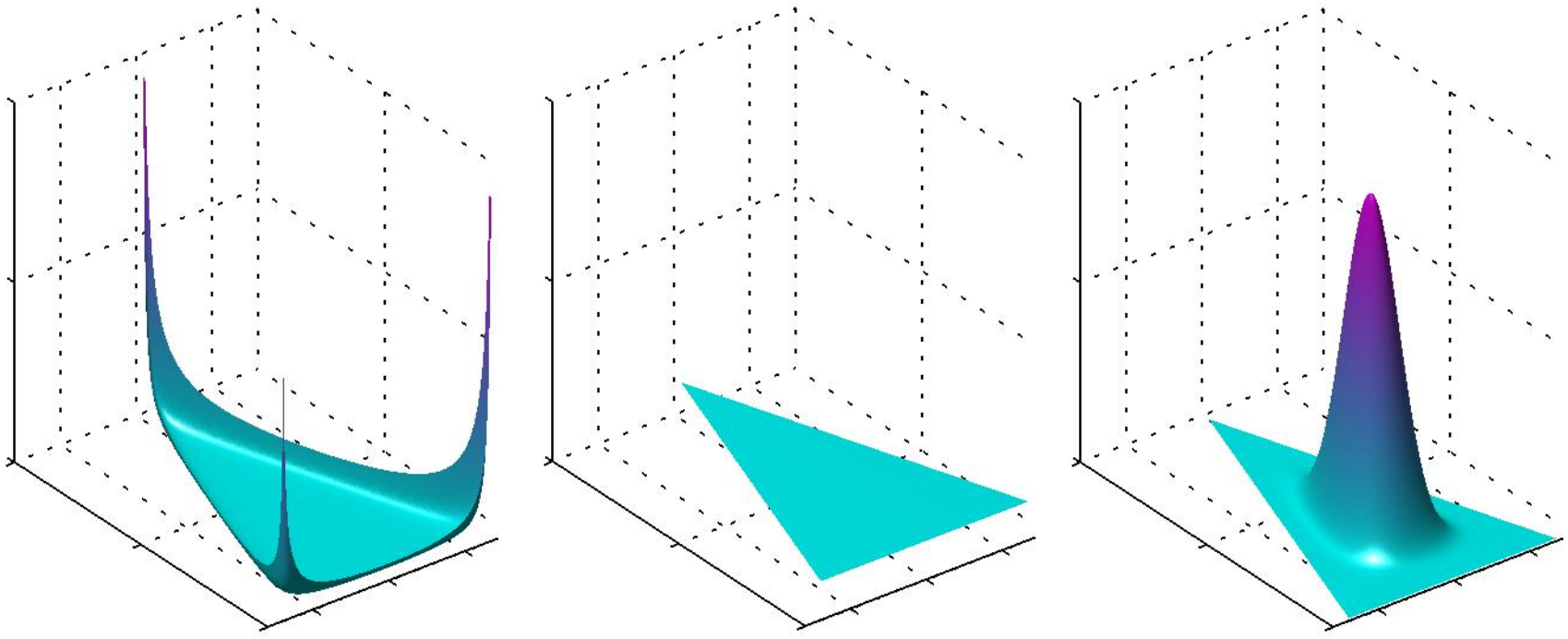
$$\begin{aligned} p(\boldsymbol{\mu}|\mathcal{D},\boldsymbol{\alpha}) &= \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}+\mathbf{m}) \\ &= \frac{\Gamma(\alpha_0+N)}{\Gamma(\alpha_1+m_1)\cdots\Gamma(\alpha_K+m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k-1} \end{aligned}$$

Can interpret α_k as the effective number of observations for x_k=1

Notice that 2-state quantities can be represented *either as*:

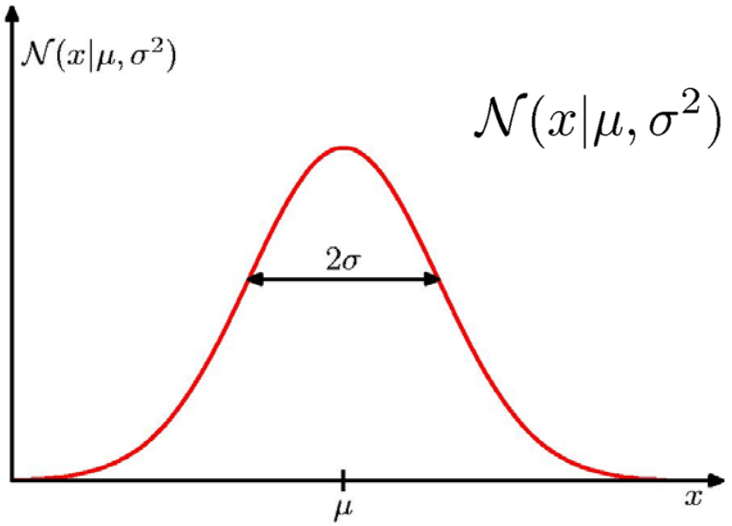
- binary variables and modeled using a binomial distribution, *or as*
- 1-of-2 variables and modeled using the multinomial distribution

Bayesian Multinomial (2)



$\alpha_k = 10^{-1}$ $\alpha_k = 10^{0}$ $\alpha_k = 10^{1}$

The Gaussian Distribution

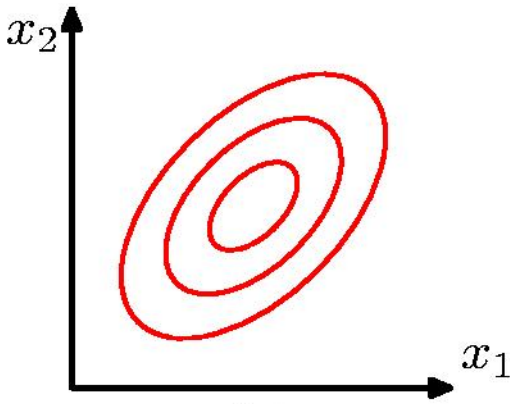


$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

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

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

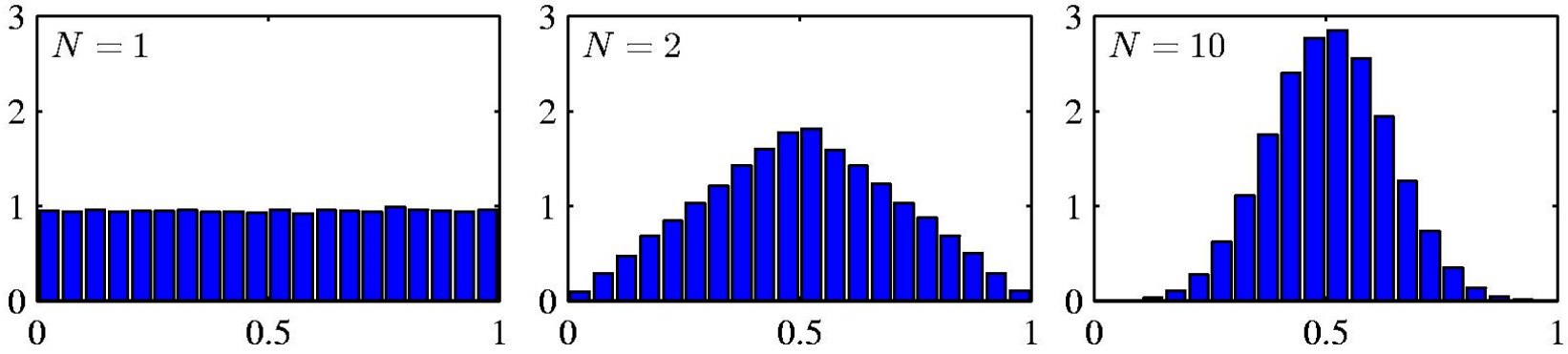
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$



Central Limit Theorem (Laplace)

The mean of N i.i.d. random variables, each with finite mean and variance, becomes increasingly Gaussian as N grows.

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



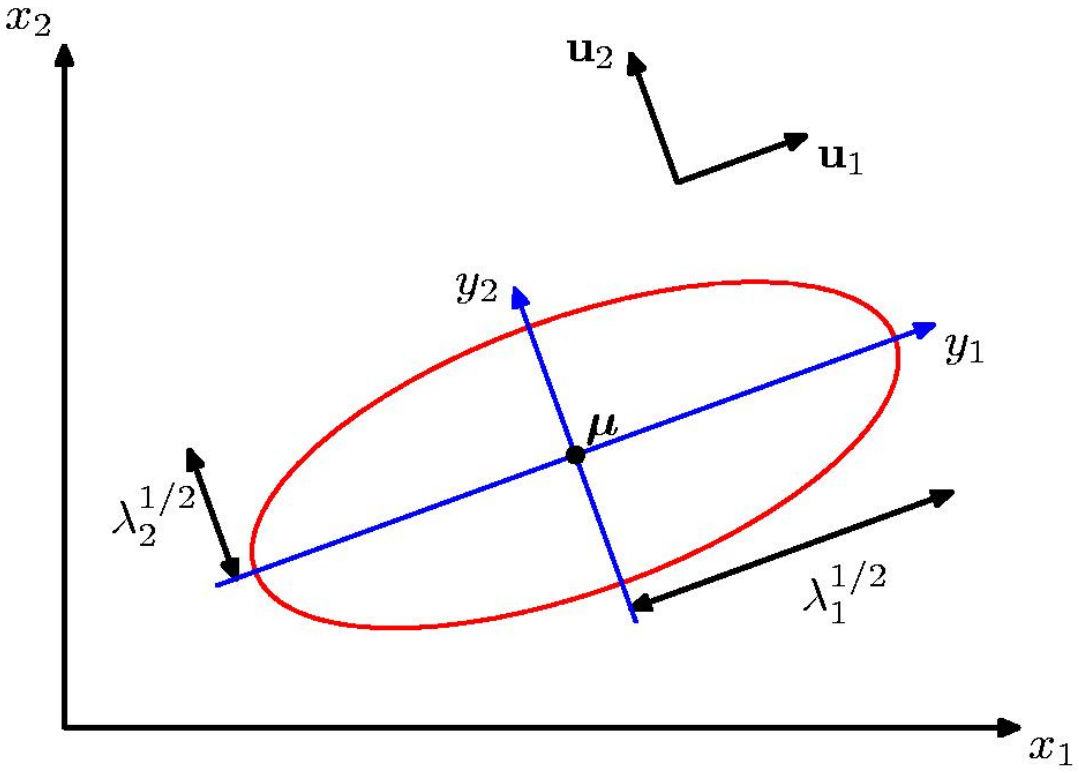
Geometry of the Multivariate Gaussian

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

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

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

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



Moments of the Multivariate Gaussian (1)

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

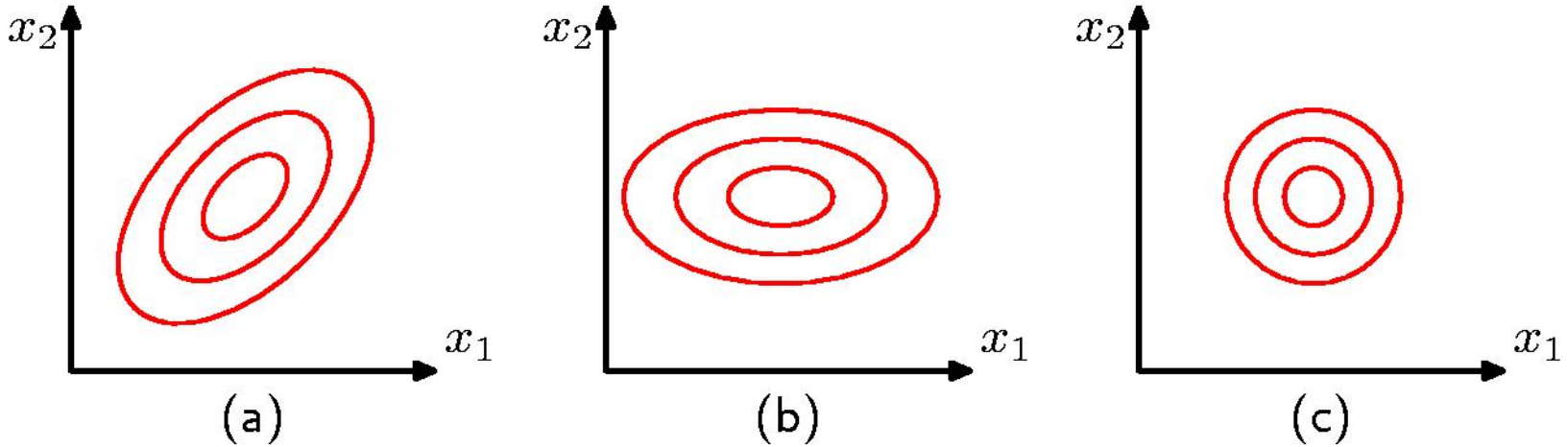
given the anti-symmetry of z, and the even function exponent

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

Moments of the Multivariate Gaussian (2)

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

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



Maximum Likelihood for the Gaussian (1)

Given i.i.d. data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^{\mathrm{T}}$,

the log-likelihood function is given by

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

Sufficient statistics

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

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

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n.$$

Similarly

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

Maximum Likelihood for the Gaussian (3)

Under the true distribution

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

Hence define

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

Partitioned Gaussian Distributions

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

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

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

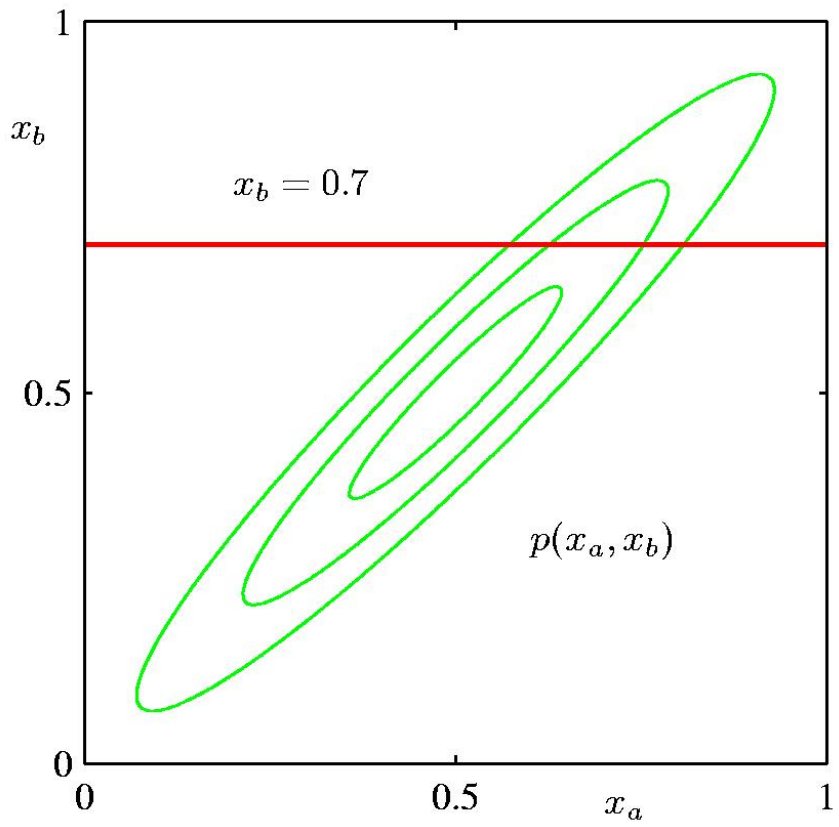
Partitioned Conditionals and Marginals

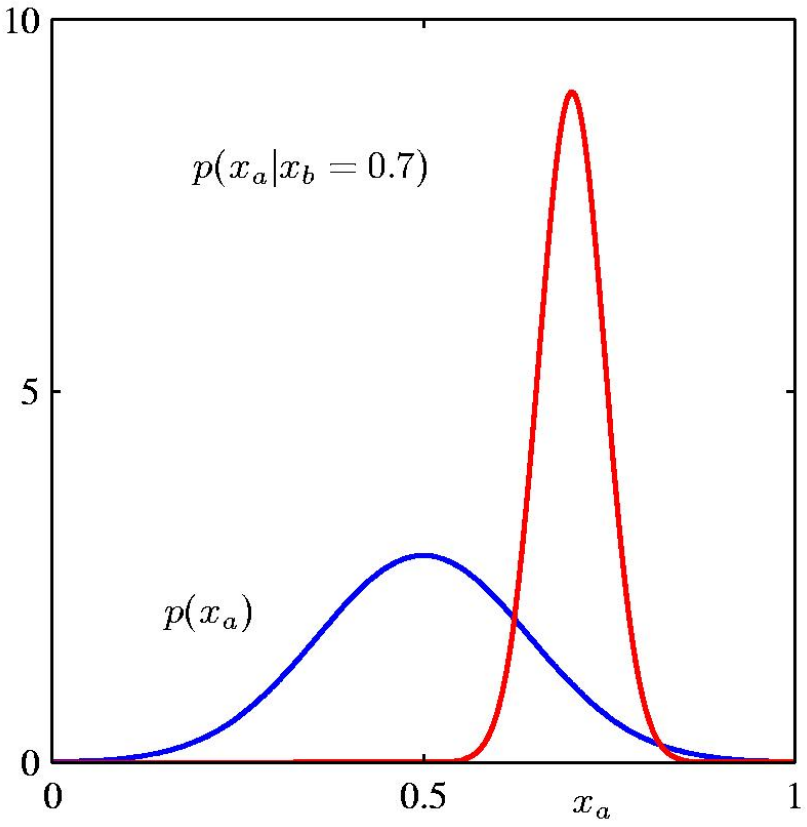
$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$$

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

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

Partitioned Conditionals and Marginals





Linear Gaussian Models

Given

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

we have

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

where

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

Many applications for linear Gaussian models,
time series models - linear dynamical systems (Kalman filtering)

Student's t-Distribution

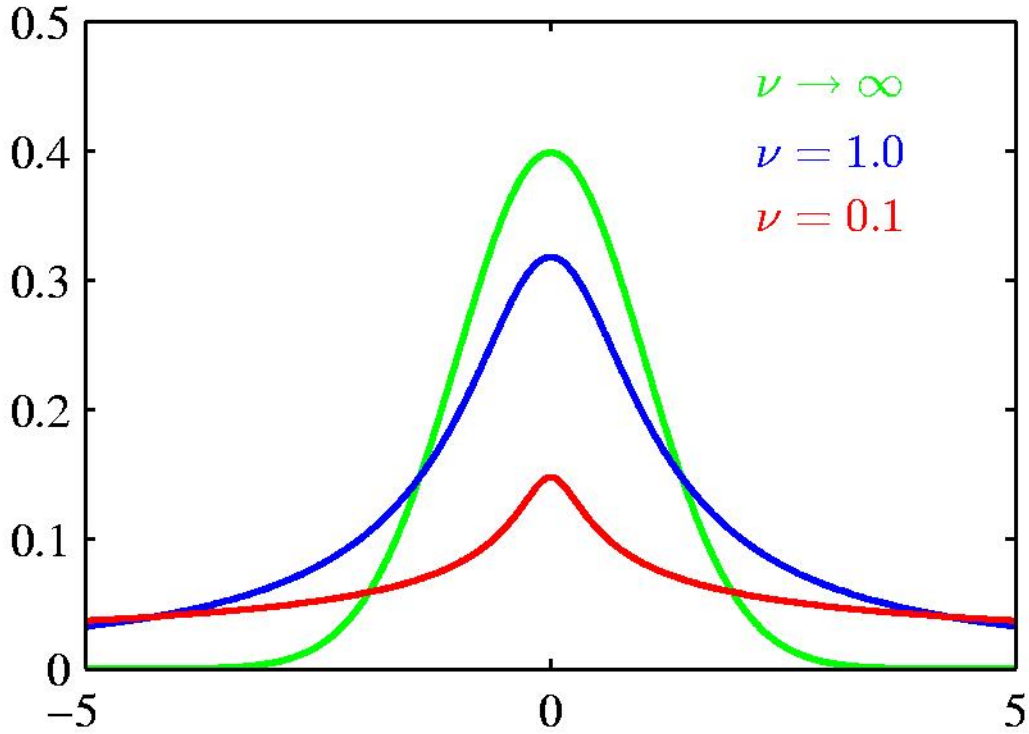
$$
\begin{aligned}
p(x|\mu,a,b) &= \int_0^\infty \mathcal{N}(x|\mu,\tau^{-1})\mathrm{Gam}(\tau|a,b)\,\mathrm{d}\tau \\
&= \int_0^\infty \mathcal{N}\left(x|\mu,(\eta\lambda)^{-1}\right)\mathrm{Gam}(\eta|\nu/2,\nu/2)\,\mathrm{d}\eta \\
&= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left[1+\frac{\lambda(x-\mu)^2}{\nu}\right]^{-\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

where

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

Infinite mixture of Gaussians.

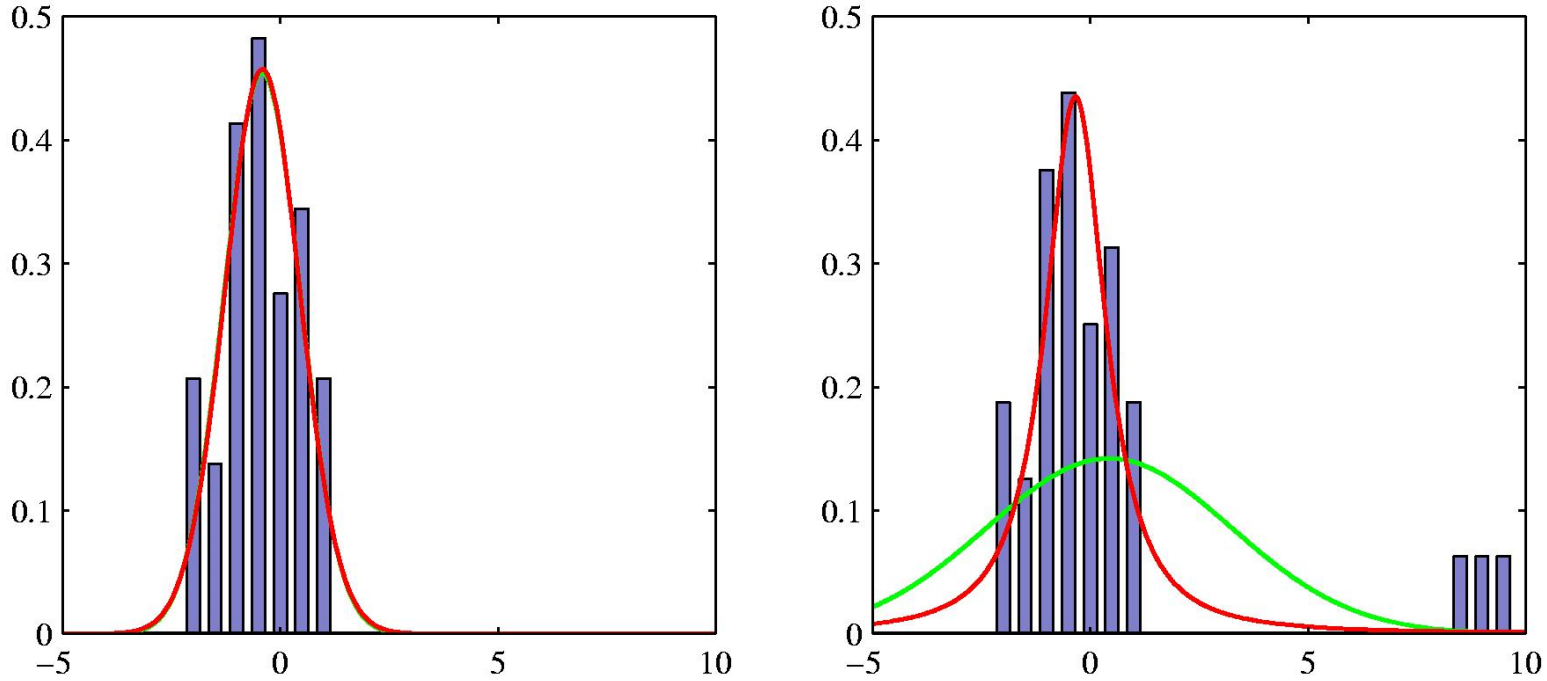
Student's t-Distribution



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

Student's t-Distribution

Robustness to outliers: Gaussian vs t-distribution.



Readings

Bishop

Ch. 1, section 1.5

Ch. 2, sections 2.1 - 2.4