Exponential Families and Kernels

Lecture 1

Alexander J. Smola
Alex.Smola@nicta.com.au

Machine Learning Program
National ICT Australia
RSISE, The Australian National University

Outline

Exponential Families

- Maximum likelihood and Fisher information
- Priors (conjugate and normal)

Conditioning and Feature Spaces

- Conditional distributions and inner products
- Clifford Hammersley Decomposition

Applications

- Classification and novelty detection
- Regression

Applications

- Conditional random fields
- Intractable models and semidefinite approximations

Lecture 1

Model

- Log partition function
- Expectations and derivatives
- Maximum entropy formulation

Examples

- Normal distribution
- Discrete events
- Laplacian distribution
- Poisson distribution
- Beta distribution

Estimation

- Maximum Likelihood Estimator
- Fisher Information Matrix and Cramer Rao Theorem
- Normal Priors and Conjugate Priors

The Exponential Family

Definition

A family of probability distributions which satisfy

$$p(x;\theta) = \exp(\langle \phi(x), \theta \rangle - g(\theta))$$

Details

- $\phi(x)$ is called the sufficient statistics of x.
- $\mathcal{X}$ is the domain out of which x is drawn ($x \in \mathcal{X}$).
- $g(\theta)$ is the log-partition function and it ensures that the distribution integrates out to 1.

$$g(\theta) = \log \int_{\mathcal{X}} \exp(\langle \phi(x), \theta \rangle) dx$$

Example: Binomial Distribution

Tossing coins

With probability p we have heads and with probability $1-p$ we see tails. So we have

$$p(x) = p^x(1-p)^{1-x} \text{ where } x \in \{0,1\} =: \mathcal{X}$$

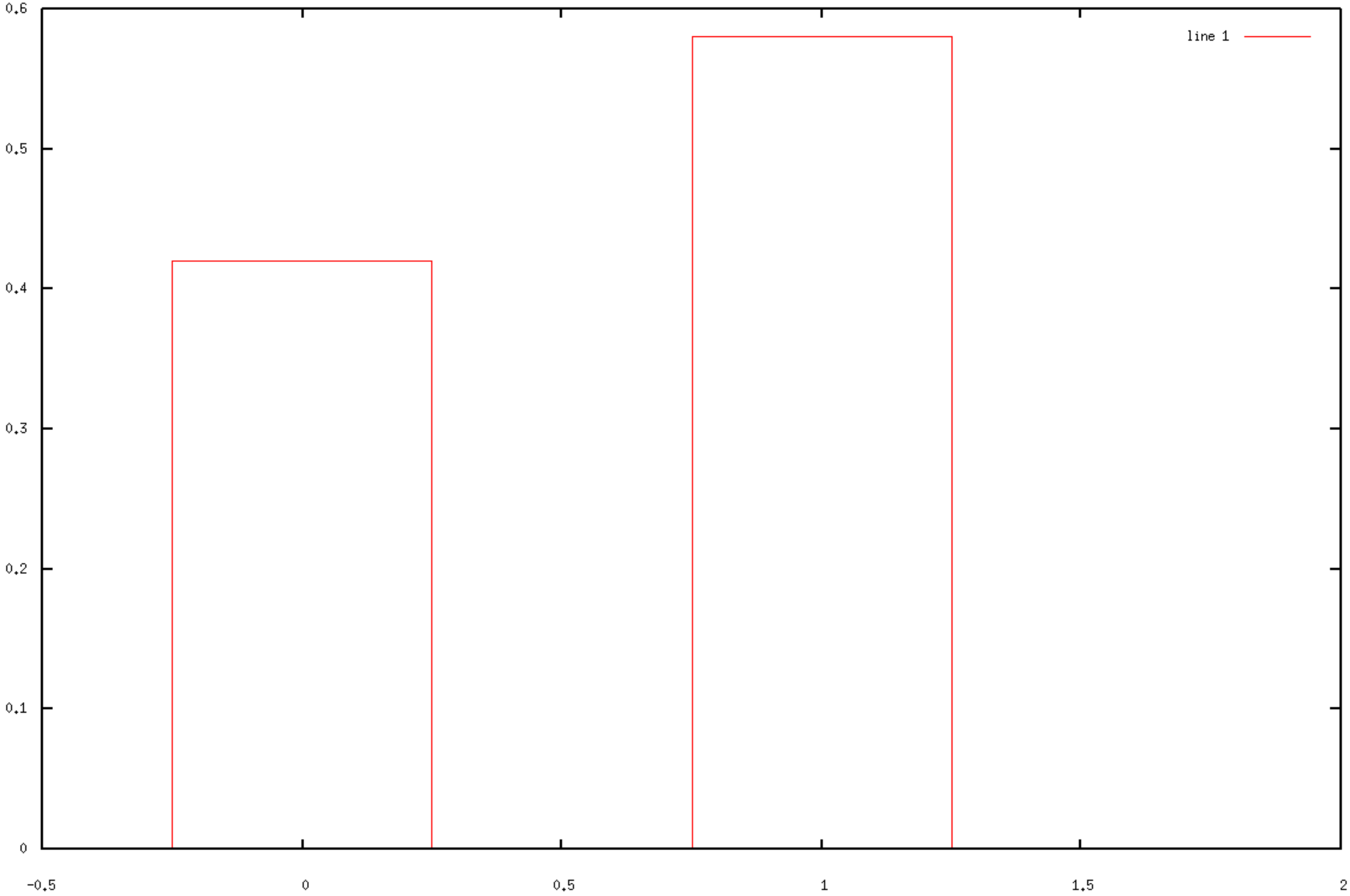
Massaging the math

$$\begin{aligned} p(x) &= \exp \log p(x) \\ &= \exp\left(x \log p + (1-x)\log(1-p)\right) \\ &= \exp\Big(\langle \underbrace{(x, 1-x)}_{\phi(x)}, \underbrace{(\log p, \log(1-p))}_{\theta} \rangle\Big) \end{aligned}$$

The Normalization Once we relax the restriction on $\theta \in \mathbb{R}^2$ we need $g(\theta)$ which yields

$$g(\theta) = \log\left(e^{\theta_1} + e^{\theta_2}\right)$$

Example: Binomial Distribution



Example: Laplace Distribution

Atomic decay

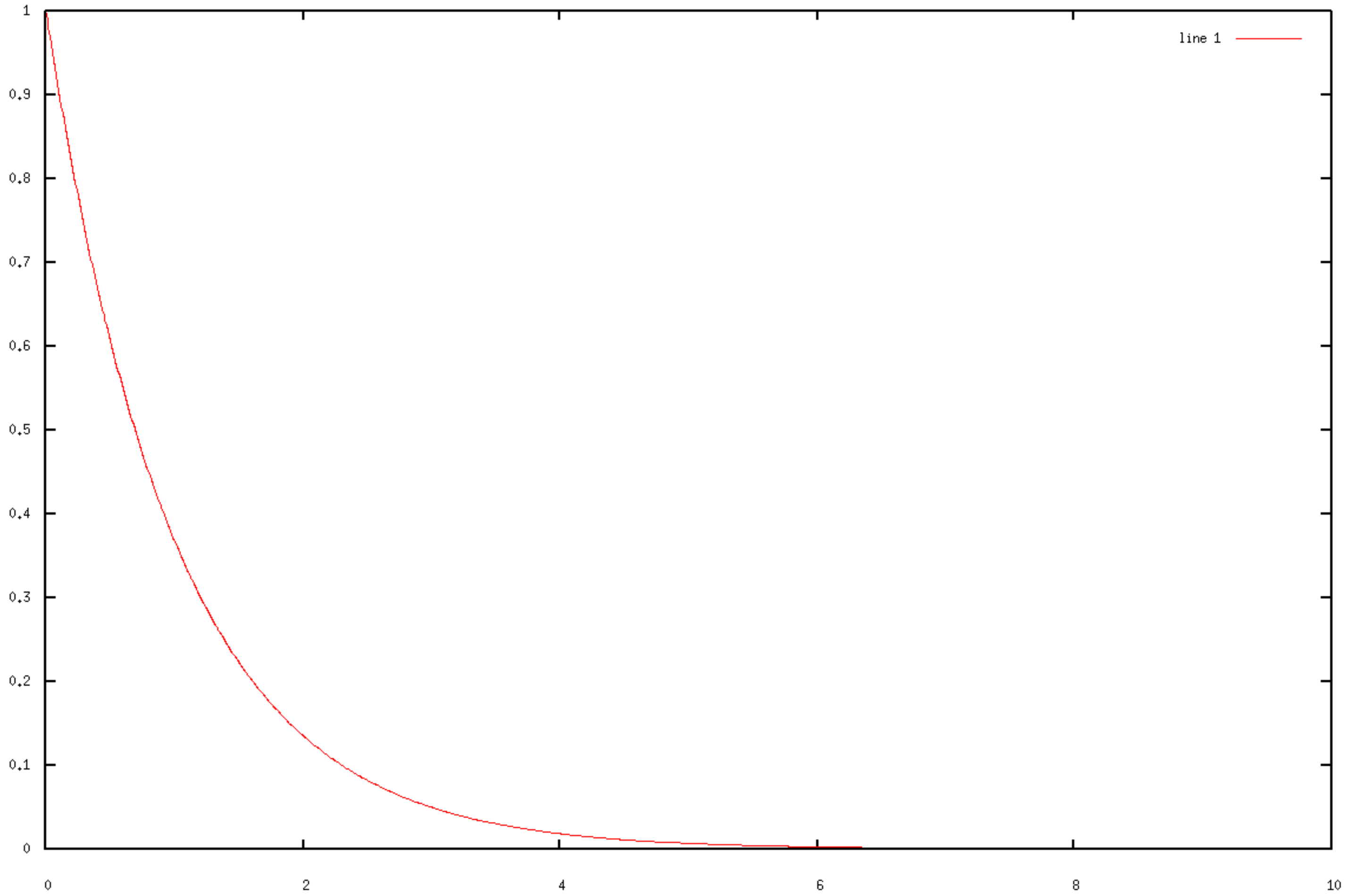
At any time, with probability θdx an atom will decay in the time interval $[x, x + dx]$ if it still exists. Consulting your physics book tells us that this gives us the density

$$p(x) = \theta \exp(\theta x) \text{ where } x \in [0, \infty) =: \mathcal{X}$$

Massaging the math

$$p(x) = \exp\Big(\langle \underbrace{-x}_{\phi(x)}, \theta \rangle - \underbrace{- \log \theta}_{g(\theta)}\Big)$$

Example: Laplace Distribution



Example: Normal Distribution

Engineer's favorite

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \text{ where } x \in \mathbb{R} =: \mathcal{X}$$

Massaging the math

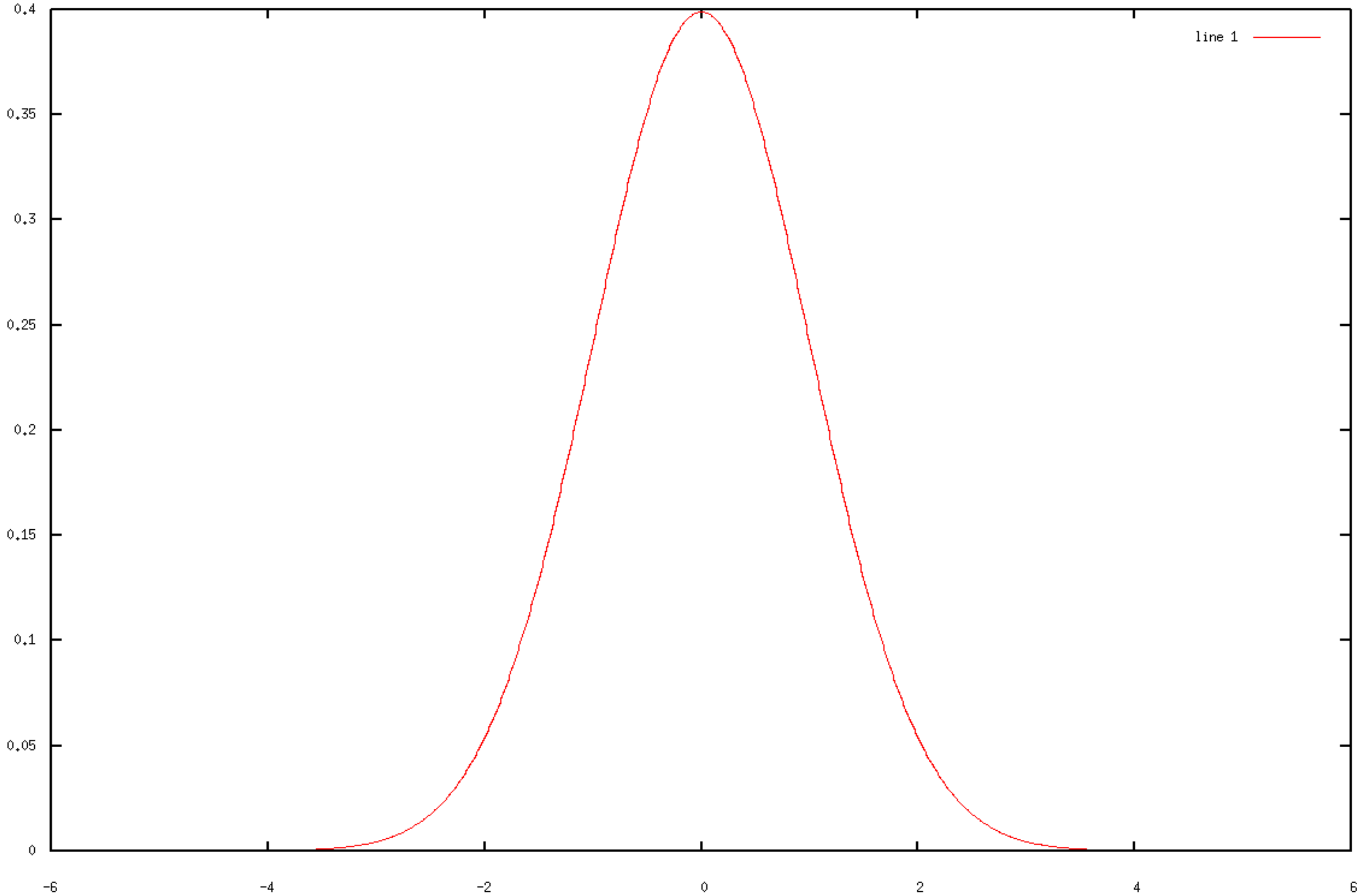
$$p(x) = \exp\left(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right)$$

$$= \exp\Big(\langle \underbrace{(x, x^2)}_{\phi(x)}, \theta\rangle - \underbrace{\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)}_{g(\theta)}\Big)$$

Finally we need to solve (μ, σ^2) for θ. Tedious algebra yields $\theta_2 := -\frac{1}{2}\sigma^{-2}$ and $\theta_1 := \mu\sigma^{-2}$. We have

$$g(\theta) = -\frac{1}{4}\theta_1^2\theta_2^{-1} + \frac{1}{2}\log 2\pi - \frac{1}{2}\log -2\theta_2$$

Example: Normal Distribution



Example: Multinomial Distribution

Many discrete events

Assume that we have disjoint events $[1..n] =: \mathcal{X}$ which all may occur with a certain probability p_x.

Guessing the answer

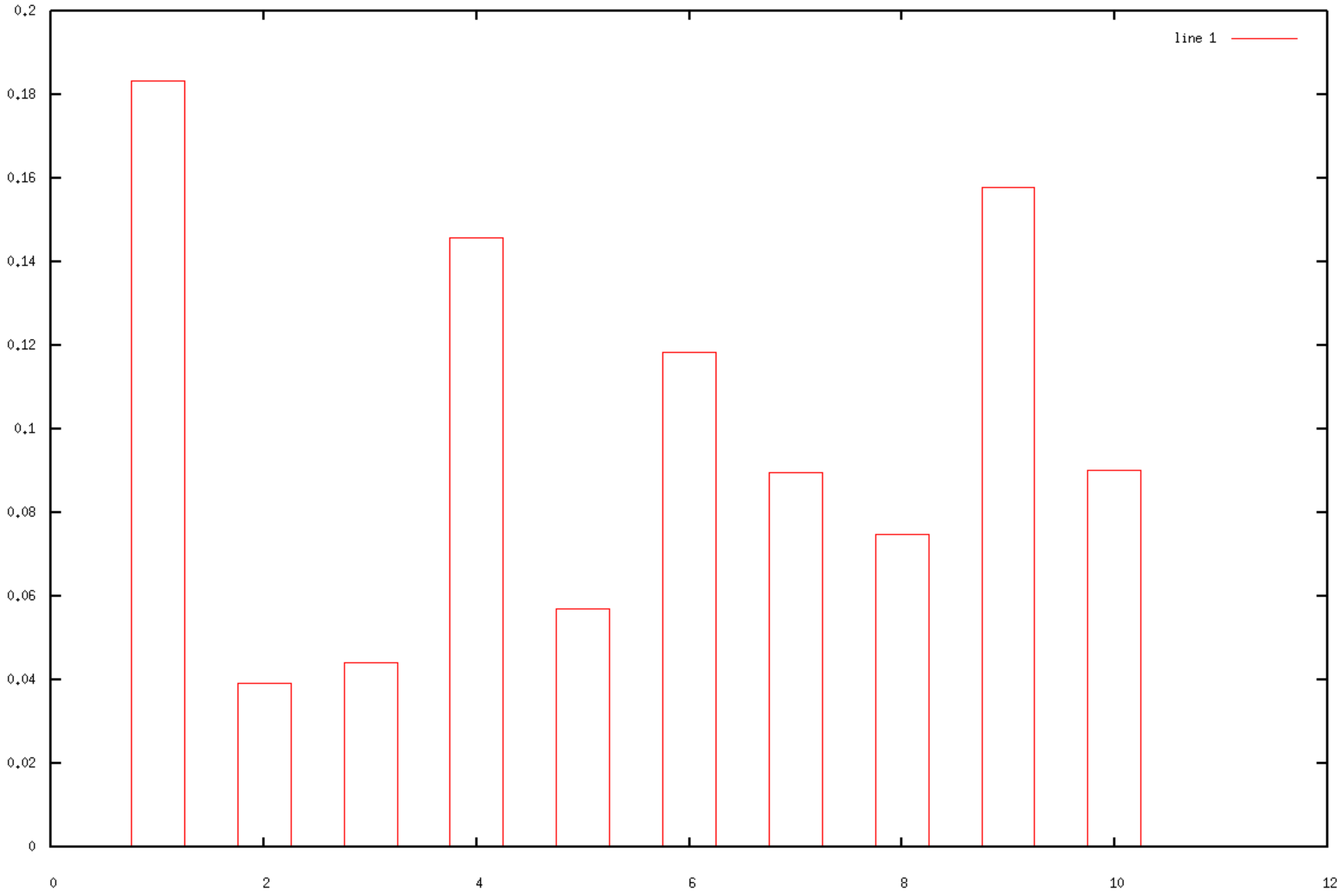
Use the map $\phi : x \to e_x$, that is, e_x is an element of the canonical basis $(0, \ldots, 0, 1, 0, \ldots)$. This gives

$$p(x) = \exp(\langle e_x, \theta \rangle - g(\theta))$$

where the normalization is

$$g(\theta) = \log \sum_{i=1}^{n} \exp(\theta_i)$$

Example: Multinomial Distribution



Example: Poisson Distribution

Limit of Binomial distribution

Probability of observing $x \in \mathbb{N}$ events which are all independent (e.g. raindrops per square meter, crimes per day, cancer incidents)

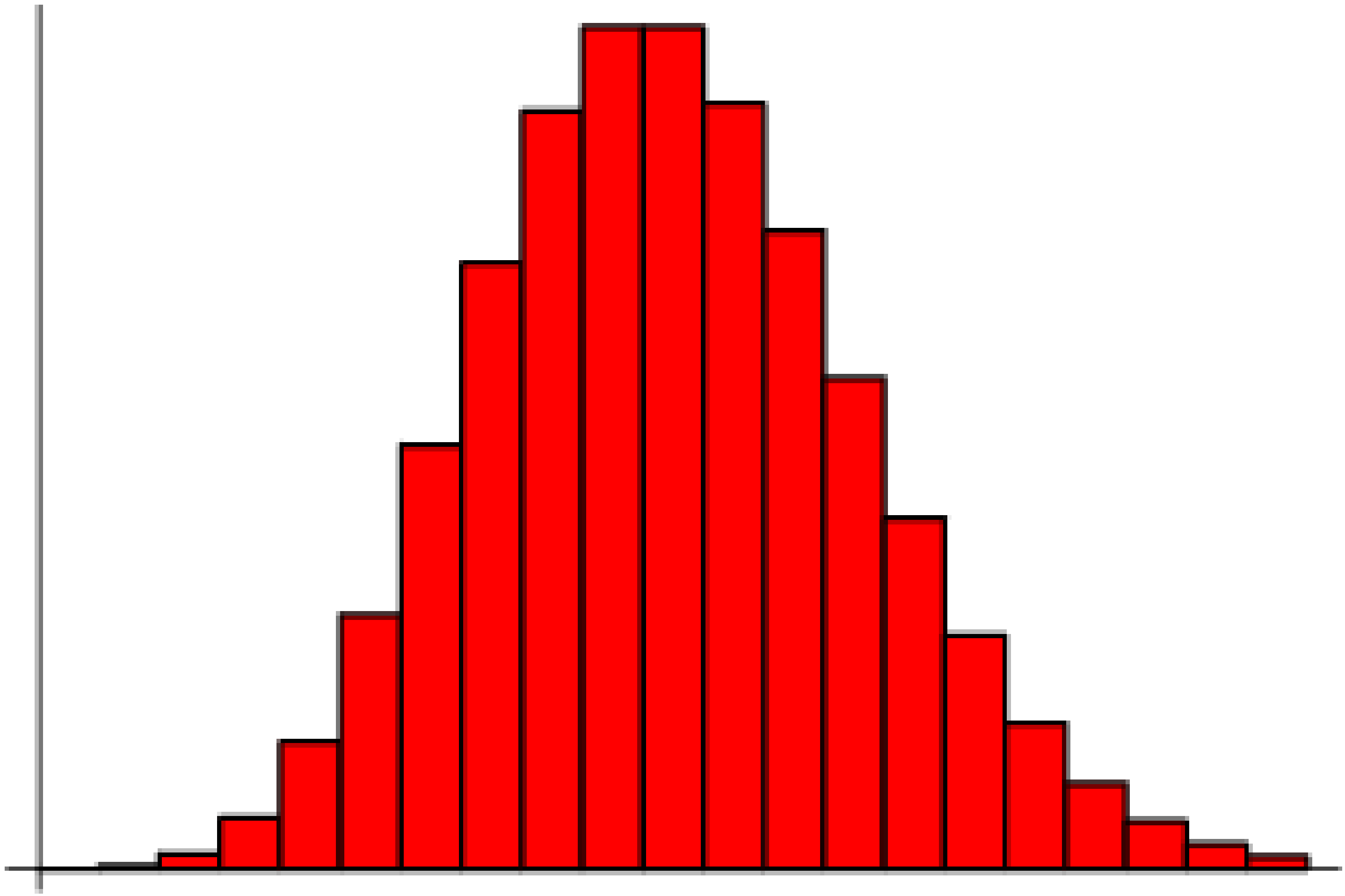
$$p(x) = \exp\left(x \cdot \theta - \log \Gamma(x+1) - \exp(\theta)\right).$$

Hence $\phi(x) = x$ and $g(\theta) = e^\theta$.

Differences

- We have a normalization dependent on x alone, namely $\Gamma(x+1)$. This leaves the rest of the theory unchanged.
- The domain is countably infinite.

Example: Poisson Distribution



Example: Beta Distribution

Usage

Often used as prior on Binomial distributions (it is a conjugate prior as we will see later).

Mathematical Form

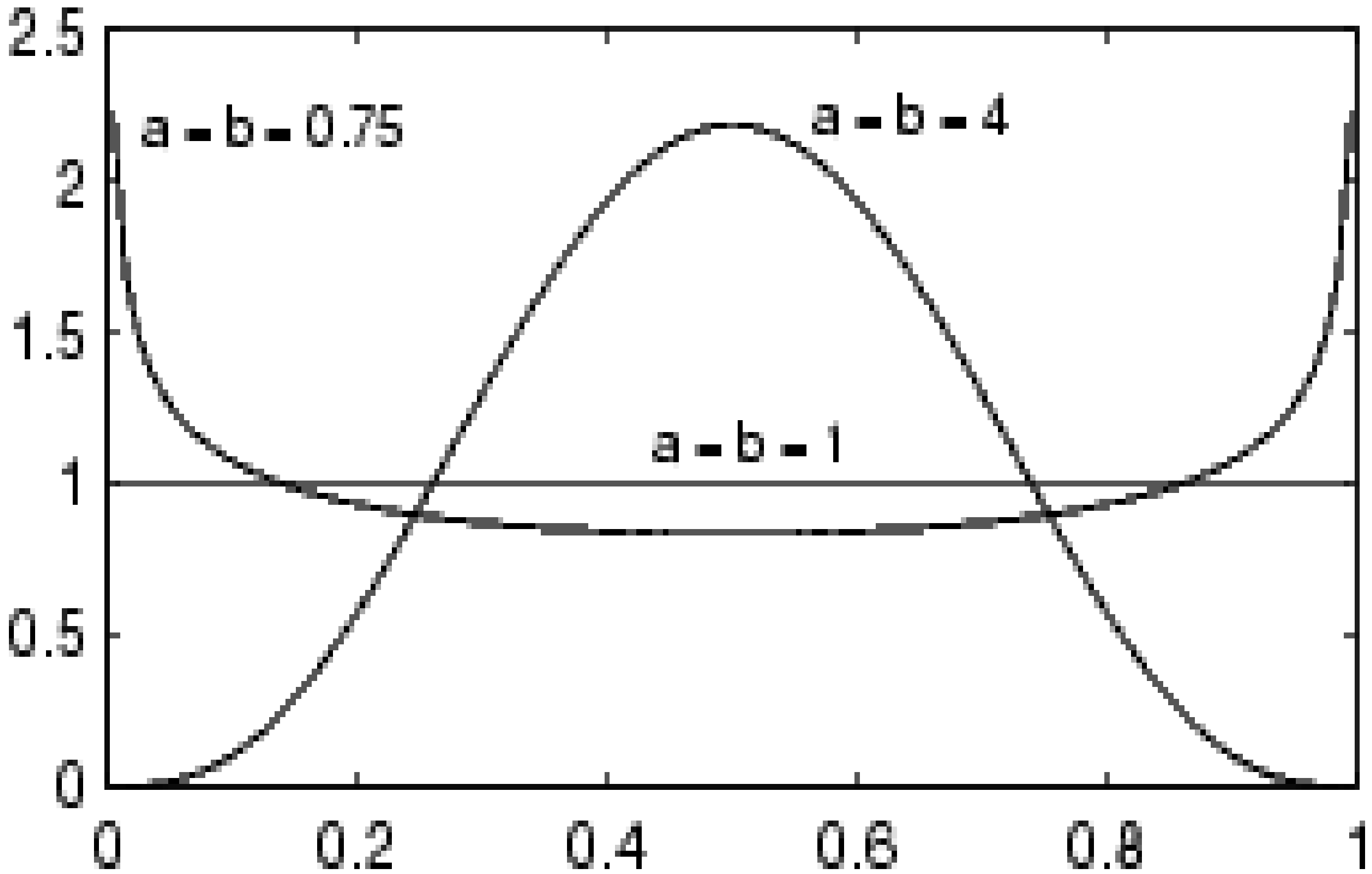
$$p(x) = \exp(\langle(\log x, \log(1-x)), (\theta_1, \theta_2)\rangle - \log B(\theta_1+1, \theta_2+1))$$

where the domain is $x \in [0, 1]$ and

$$\begin{aligned} g(\theta) &= \log B(\theta_1 + 1, \theta_2 + 1) \\ &= \log \Gamma(\theta_1 + 1) + \log \Gamma(\theta_2 + 1) - \log \Gamma(\theta_1 + \theta_2 + 2) \end{aligned}$$

Here $B(\alpha, \beta)$ is the *Beta* function.

Example: Beta Distribution



Example: Gamma Distribution

Usage

- Popular as a prior on coefficients
- Obtained from integral over waiting times in Poisson distribution

Mathematical Form

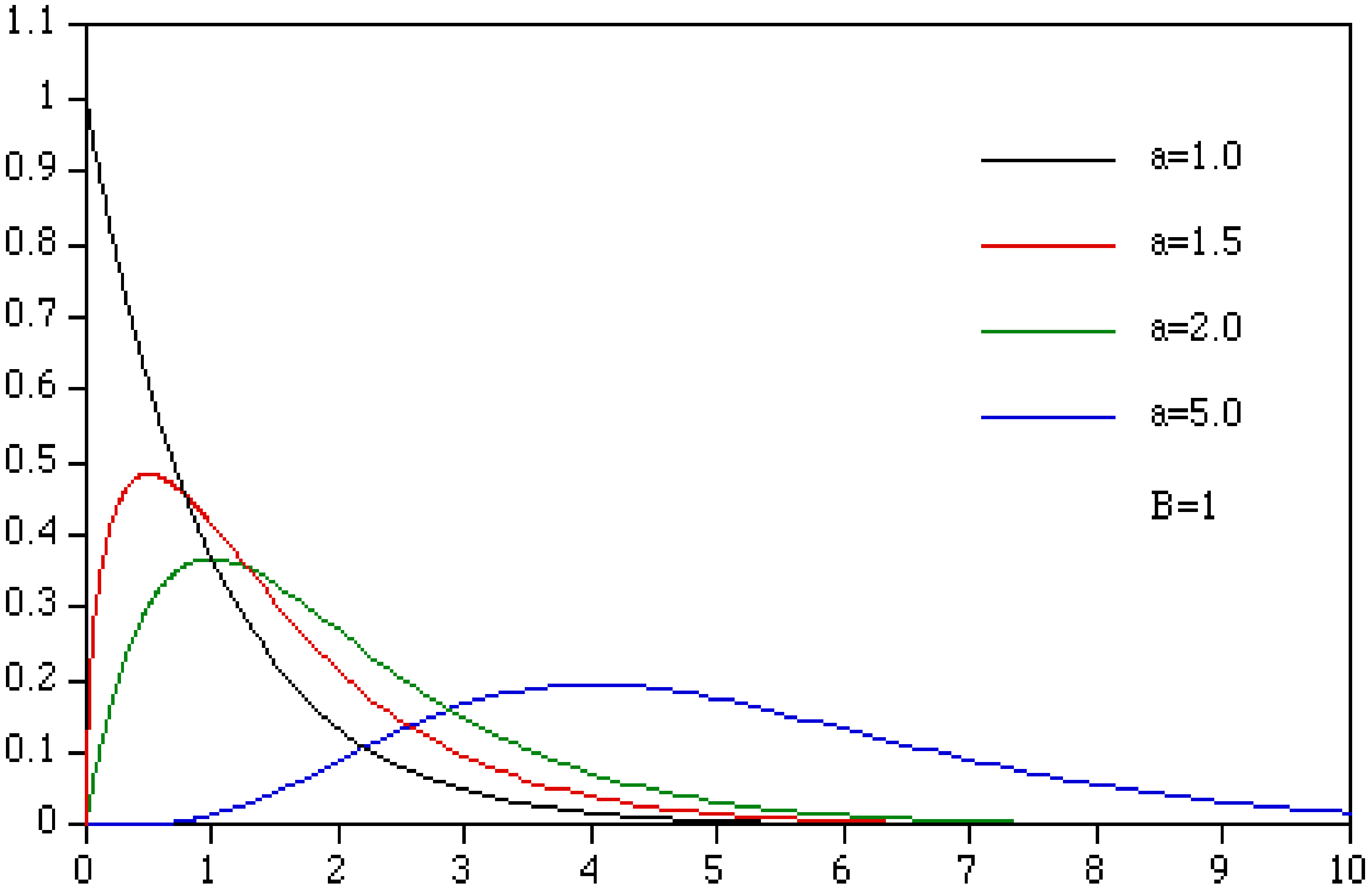
$$p(x) = \exp(\langle(\log x, x), (\theta_1, \theta_2)\rangle - \log\Gamma(\theta_1+1) + (\theta_1+1)\log -\theta_2)$$

where the domain is $x \in [0, \infty]$ and

$$g(\theta) = \log\Gamma(\theta_1 + 1) + (\theta_1 + 1)\log -\theta_2)$$

Note that $\theta \in [0, \infty) \times (-\infty, 0)$.

Example: Gamma Distribution



Zoology of Exponential Families

Name	$\phi(x)$	Domain	Measure
Binomial	$(x, 1-x)$	$\{0, 1\}$	discrete
Multinomial	e_x	$\{1, \ldots, n\}$	discrete
Poisson	x	$\mathbb{N}_0$	discrete
Laplace	x	$[0, \infty)$	Lebesgue
Normal	(x, x^2)	$\mathbb{R}$	Lebesgue
Beta	$(\log x, \log(1-x))$	$[0, 1]$	Lebesgue
Gamma	$(\log x, x)$	$[0, \infty)$	Lebesgue
Wishart	$(\log \lvert X \rvert, X)$	$X \succeq 0$	Lebesgue
Dirichlet	$\log x$	$x \in \mathbb{R}_+^n, \lVert x \rVert_1 = 1$	Lebesgue

Recall

Definition

A family of probability distributions which satisfy

$$p(x;\theta) = \exp(\langle \phi(x), \theta \rangle - g(\theta))$$

Details

- $\phi(x)$ is called the sufficient statistics of x.
- $\mathfrak{X}$ is the domain out of which x is drawn ($x \in \mathfrak{X}$).
- $g(\theta)$ is the log-partition function and it ensures that the distribution integrates out to 1.

$$g(\theta) = \log \int_{\mathfrak{X}} \exp(\langle \phi(x), \theta \rangle) dx$$

NATIONAL ICT AUSTRALIA

Benefits: Log-partition function is nice

$g(\theta)$ **generates moments:**

$$g(\theta) = \log \int \exp(\langle \phi(x), \theta \rangle) dx$$

Taking the derivative wrt. θ we can see that

$$\partial_\theta g(\theta) = \frac{\int \phi(x) \exp(\langle \phi(x), \theta \rangle) dx}{\int \exp(\langle \phi(x), \theta \rangle) dx} = \mathbf{E}_{x \sim p(x;\theta)} \left[\phi(x) \right]$$

$$\partial_\theta^2 g(\theta) = \mathbf{Cov}_{x \sim p(x;\theta)} \left[\phi(x) \right]$$

... and so on for higher order moments ...

Corollary:

$g(\theta)$ is convex

Benefits: Simple Estimation

Likelihood of a set: Given $X := \{x_1, \ldots, x_m\}$ we get

$$p(X;\theta) = \prod_{i=1}^{m} p(x_i;\theta) = \exp\left(\sum_{i=1}^{m} \langle \phi(x_i), \theta\rangle - mg(\theta)\right)$$

Maximum Likelihood

We want to minimize the negative log-likelihood, i.e.

$$\operatorname*{minimize}_{\theta} \quad g(\theta) - \left\langle \frac{1}{m}\sum_{i=1}^{m} \phi(x_i), \theta \right\rangle$$

$$\Longrightarrow \quad \mathbf{E}[\phi(x)] = \frac{1}{m}\sum_{i=1}^{m} \phi(x_i) =: \mu$$

Solving the maximum likelihood problem is **easy**.

Application: Laplace distribution

Estimate the decay constant of an atom:

We use exponential family notation where

$$p(x;\theta) = \exp(\langle(-x), \theta\rangle - (-\log\theta))$$

Computing μ

Since $\phi(x) = -x$ all we need to do is **average over all decay times** that we observe.

Solving for Maximum Likelihood

The maximum likelihood condition yields

$$\mu = \partial_\theta g(\theta) = \partial_\theta(-\log\theta) = -\frac{1}{\theta}$$

This leads to $\theta = -\frac{1}{\mu}$.

Benefits: Maximum Entropy Estimate

Entropy

Basically it's the number of bits needed to encode a random variable. It is defined as

$$H(p) = \int -p(x) \log p(x) dx \text{ where we set } 0 \log 0 := 0$$

Maximum Entropy Density

The density $p(x)$ satisfying $\mathbf{E}[\phi(x)] \geq \eta$ with maximum entropy is $\exp(\langle \phi(x), \theta \rangle - g(\theta))$.

Corollary

The most vague density with a given variance is the Gaussian distribution.

Corollary

The most vague density with a given mean is the Laplacian distribution.

Using it

Observe Data

$x_1, \ldots, x_m$ drawn from distribution $p(x|\theta)$

Compute Likelihood

$$p(X|\theta) = \prod_{i=1}^{m} \exp(\langle \phi(x_i), \theta \rangle - g(\theta))$$

Maximize it

Take the negative log and minimize, which leads to

$$\partial_\theta g(\theta) = \frac{1}{m} \sum_{i=1}^{m} \phi(x_i)$$

This can be solved analytically or (whenever this is impossible or we are lazy) by Newton's method.

Application: Discrete Events

Simple Data

Discrete random variables (e.g. tossing a dice).

Outcome	1	2	3	4	5	6
Counts	3	6	2	1	4	4
Probabilities	0.15	0.30	0.10	0.05	0.20	0.20

Maximum Likelihood Solution

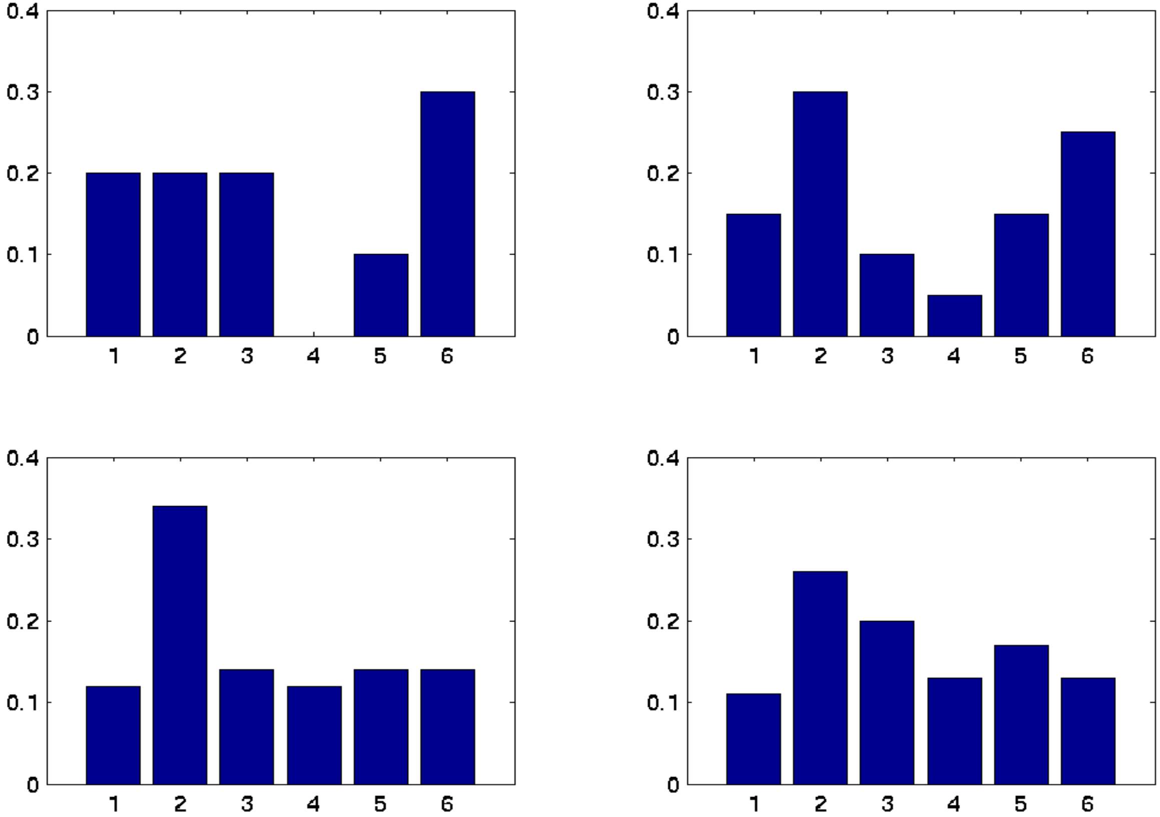
Count the number of outcomes and use the relative frequency of occurrence as estimates for the probability:

$$p_{\mathrm{emp}}(x) = \frac{\#x}{m}$$

Problems

- Bad idea if we have few data.
- Bad idea if we have continuous random variables.

Tossing a dice



Fisher Information and Efficiency

Fisher Score

$$V_\theta(x) := \partial_\theta \log p(x;\theta)$$

This tells us the influence of x on estimating θ. Its expected value vanishes, since

$$\begin{aligned}\mathbf{E}\left[\partial_\theta \log p(X;\theta)\right] &= \int p(X;\theta)\partial_\theta \log p(X;\theta) dX \\ &= \partial_\theta \int p(X;\theta) dX = 0.\end{aligned}$$

Fisher Information Matrix

It is the covariance matrix of the Fisher scores, that is

$$I := \mathrm{Cov}[V_\theta(x)]$$

NATIONAL ICT AUSTRALIA

Cramer Rao Theorem

Efficiency

Covariance of estimator $\hat{\theta}(X)$ rescaled by I:

$$e := \det \mathrm{Cov}[\hat{\theta}(X)] \mathrm{Cov}[\partial_\theta \log p(X;\theta)]$$

Theorem

The efficiency for unbiased estimators is never better (i.e. smaller) than 1. Equality is achieved for MLEs.

Proof (scalar case only)

By Cauchy-Schwartz we have

$$\left(\mathbf{E}_\theta\left[(V_\theta(X) - \mathbf{E}_\theta\left[V_\theta(X)\right])\left(\hat{\theta}(X) - \mathbf{E}_\theta\left[\hat{\theta}(X)\right]\right)\right]\right)^2$$

$$\leq \mathbf{E}_\theta\left[(V_\theta(X) - \mathbf{E}_\theta\left[V_\theta(X)\right])^2\right]\mathbf{E}_\theta\left[\left(\hat{\theta}(X) - \mathbf{E}_\theta\left[\hat{\theta}(X)\right]\right)^2\right] = IB$$

Cramer Rao Theorem

Proof

At the same time, $\mathbf{E}_\theta\left[V_\theta(X)\right] = 0$ implies that

$$\begin{aligned}
&\mathbf{E}_\theta\left[\left(V_\theta(X) - \mathbf{E}_\theta\left[V_\theta(X)\right]\right)\left(\hat{\theta}(X) - \mathbf{E}_\theta\left[\hat{\theta}(X)\right]\right)\right]^2 \\
=&\mathbf{E}_\theta\left[V_\theta(X)\hat{\theta}(X)\right]^2 \\
=&\left(\int p(X|\theta)\partial_\theta p(X|\theta)\hat{\theta}(X)dX\right) \\
=&\partial_\theta \int p(X|\theta)\hat{\theta}(X)dX = \partial_\theta\theta = 1.
\end{aligned}$$

Cautionary Note

This does not imply that a biased estimator might not have lower variance.

Fisher and Exponential Families

Fisher Score

$$\begin{aligned} V_\theta(x) &= \partial_\theta \log p(x;\theta) \\ &= \phi(x) - \partial_\theta g(\theta) \end{aligned}$$

Fisher Information

$$\begin{aligned} I &= \mathrm{Cov}[V_\theta(x)] \\ &= \mathrm{Cov}[\phi(x) - \partial_\theta g(\theta)] \\ &= \partial_\theta^2 g(\theta) \end{aligned}$$

Efficiency of estimator can be obtained directly from log-partition function.

Outer Product Matrix

It is given (up to an offset) by $\langle \phi(x), \phi(x')\rangle$. This leads to Kernel-PCA . . .

Priors

Problems with Maximum Likelihood

With not enough data, parameter estimates will be bad.

Prior to the rescue

Often we know where the solution should be. So we encode the latter by means of a prior $p(\theta)$.

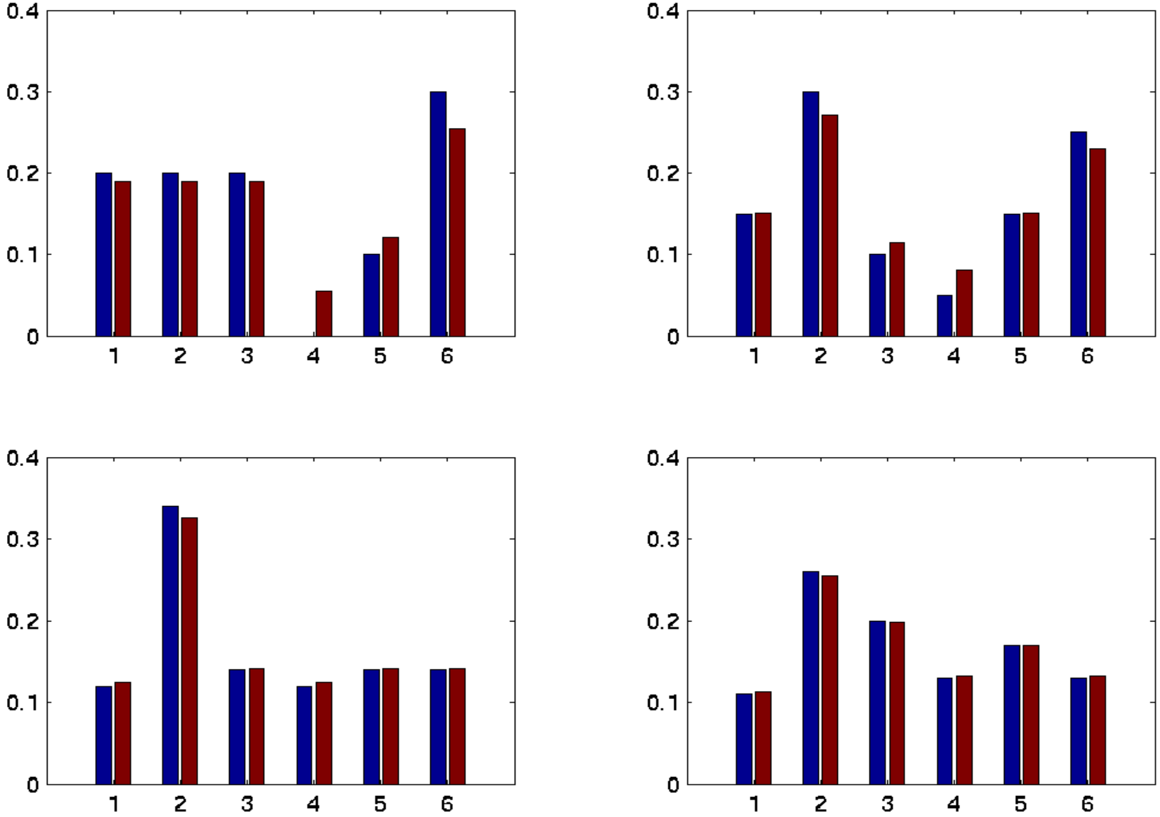
Normal Prior

Simply set $p(\theta) \propto \exp(-\frac{1}{2\sigma^2}\|\theta\|^2)$.

Posterior

$$p(\theta|X) \propto \exp\left(\sum_{i=1}^{m} \langle \phi(x_i), \theta \rangle - g(\theta) - \frac{1}{2\sigma^2}\|\theta\|^2\right)$$

Tossing a dice with priors



Conjugate Priors

Problem with Normal Prior

The posterior looks different from the likelihood. So many of the Maximum Likelihood optimization algorithms may not work ...

Idea

What if we had a prior which looked like additional data, that is

$$p(\theta|X) \sim p(X|\theta)$$

For exponential families this is easy. Simply set

$$p(\theta|a) \propto \exp(\langle \theta, m_0 a \rangle - m_0 g(\theta))$$

Posterior

$$p(\theta|X) \propto \exp\left((m + m_0)\left(\left\langle \frac{m\mu + m_0 a}{m + m_0}, \theta \right\rangle - g(\theta)\right)\right)$$

Example: Multinomial Distribution

Laplace Rule

A conjugate prior with parameters (a, m_0) in the multinomial family could be to set $a = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$. This is often also called the **Dirichlet prior**. It leads to

$$p(x) = \frac{\#x + m_0/n}{m + m_0} \text{ instead of } p(x) = \frac{\#x}{m}$$

Example

Outcome	1	2	3	4	5	6
Counts	3	6	2	1	4	4
MLE	0.15	0.30	0.10	0.05	0.20	0.20
MAP ($m_0 = 6$)	0.25	0.27	0.12	0.08	0.19	0.19
MAP ($m_0 = 100$)	0.16	0.19	0.16	0.15	0.17	0.17

Optimization Problems

Maximum Likelihood

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{m} g(\theta) - \langle \phi(x_i), \theta \rangle \implies \partial_\theta g(\theta) = \frac{1}{m} \sum_{i=1}^{m} \phi(x_i)$$

Normal Prior

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{m} g(\theta) - \langle \phi(x_i), \theta \rangle + \frac{1}{2\sigma^2} \|\theta\|^2$$

Conjugate Prior

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{m} g(\theta) - \langle \phi(x_i), \theta \rangle + m_0 g(\theta) - m_0 \langle \tilde{\mu}, \theta \rangle$$

equivalently solve $\partial_\theta g(\theta) = \frac{1}{m + m_0} \sum_{i=1}^{m} \phi(x_i) + \frac{m_0}{m + m_0} \tilde{\mu}$

Summary

Model

- Log partition function
- Expectations and derivatives
- Maximum entropy formulation

A Zoo of Densities

Estimation

- Maximum Likelihood Estimator
- Fisher Information Matrix and Cramer Rao Theorem
- Normal Priors and Conjugate Priors
- Fisher information and log-partition function