Introduction to Probabilistic Graphical Models
Lecture 4

Exponential Family Distributions, Conjugate Priors

Télécom ParisTech,

Université Paris-Saclay, Paris, France

Instructor: Umut Şimşekli

Disclaimer

- All the material that will be used within this course is adapted from the "Bayesian Statistics and Machine Learning" course that has been given by A. Taylan Cemgil at Boğaziçi University, Istanbul

- For more info, please see `http://www.cmpe.boun.edu.tr/~cemgil/`

Outline

- Common probability distributions
- Conjugacy
- Practical Guidelines for stable numerical computation

Probabilistic Modelling

Probability Distributions

- Following distributions are used often as elementary building blocks:
  - Discrete
    - Categorical, Bernoulli, Binomial, Multinomial, Poisson
  - Continuous
    - Gaussian,
    - Beta, Dirichlet
    - Gamma, Inverse Gamma, Exponential, Chi-square, Wishart
    - Student-t, von-Mises

Exponential Family

- Many of those distributions can be written as

$$p(x|\theta) = h(x)\exp\{\theta^\top \psi(x) - A(\theta)\}$$

$$A(\theta) = \log \int_{\mathcal{X}^n} dx\; h(x)\exp(\theta^\top \psi(x))$$

$A(\theta)$	log-partition function
θ	canonical parameters
$\psi(x)$	sufficient statistics
$h(x)$	weighting function, sometimes $h(x;\theta)$

Bernoulli Distribution. $\mathcal{BE}(c; w)$

Binary (Bernoulli) random variable $c = \{0, 1\}$ with probability of success w

$$p(c = 1|w) = w \qquad p(c = 0|w) = 1 - w$$

We write

$$\begin{aligned} p(c|w) &= w^c(1-w)^{1-c} \\ &= \exp\left(c \log w + (1-c)\log(1-w)\right) \\ &= \exp\left(\log(\frac{w}{1-w})c + \log(1-w)\right) \\ &\equiv \mathcal{BE}(c; w) \end{aligned}$$

Is Bernoulli an Exponential Family ?

$$\mathcal{BE}(c; w) = \exp\left(\log(\frac{w}{1-w})c + \log(1-w)\right)$$

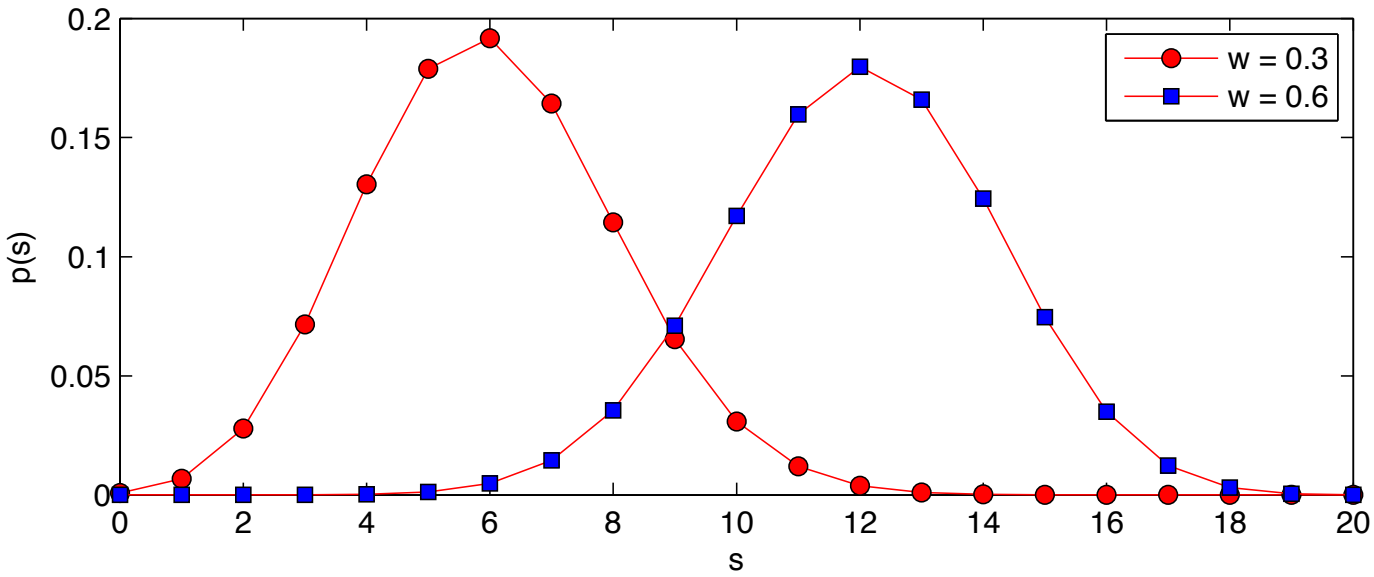
$$p(c|\theta) = h(c)\exp\{\theta^\top \psi(c) - A(\theta)\}$$

$\theta = \log(\frac{w}{1-w})$	canonical parameters
$A(\theta) = -\log(1+e^\theta)$	log-partition function
$\psi(c) = c$	sufficient statistics
$h(c) = 1$	weighting function

Binomial Distribution. $\mathcal{BI}(s; N, w)$

s is the number of successful outcomes in N independent Bernoulli trials with success probability w

$$\begin{aligned}\mathcal{BI}(s; N, w) &= \binom{N}{s} w^s (1-w)^{N-s} \\ &= \frac{N!}{s!(N-s)!} \exp(s \log w + (N-s) \log(1-w))\end{aligned}$$



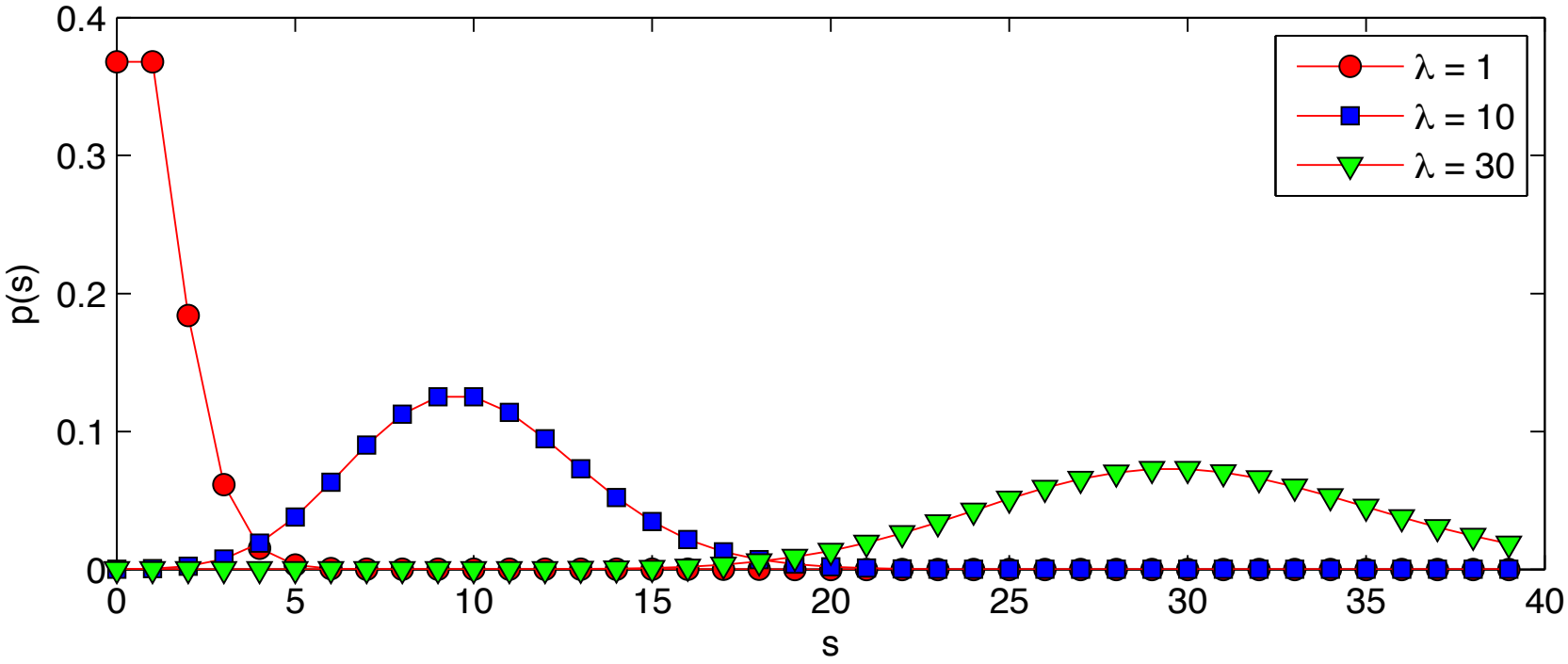
Is Binomial an Exponential Family ?

$$
\begin{aligned}
\mathcal{BI}(s; N, w) &= \binom{N}{s} w^s (1-w)^{N-s} \\
&= \binom{N}{s} \exp(s \log \frac{w}{1-w} + N \log(1-w)) \\
p(s|\theta) &= h(s;\theta) \exp\{\theta^\top \psi(s) - A(\theta)\}
\end{aligned}
$$

$\theta = \log(\frac{w}{1-w})$	canonical parameters
$A(\theta) = -N \log(1-w)$	log-partition function
$\psi(s) = s$	sufficient statistics
$h(s;\theta) = \binom{N}{s}$	weighting function

Poisson Distribution. $\mathcal{PO}(s;\lambda)$

$$\mathcal{PO}(s;\lambda) \quad = \quad \frac{e^{-\lambda}}{s!}\lambda^s = \exp(s\log\lambda - \lambda - \log(s!))$$



Is Poisson an Exponential Family ?

$$\begin{aligned} \mathcal{PO}(s;\lambda) &= \frac{e^{-\lambda}}{s!}\lambda^s = 1/s! \exp(s\log\lambda - \lambda) \\ p(s|\theta) &= h(s;\theta)\exp\{\theta^\top \psi(s) - A(\theta)\} \end{aligned}$$

$$\begin{aligned} \theta &= \log\lambda && \text{canonical parameters} \\ A(\theta) &= \lambda && \text{log-partition function} \\ \psi(s) &= s && \text{sufficient statistics} \\ h(s) &= 1/s! && \text{weighting function} \end{aligned}$$

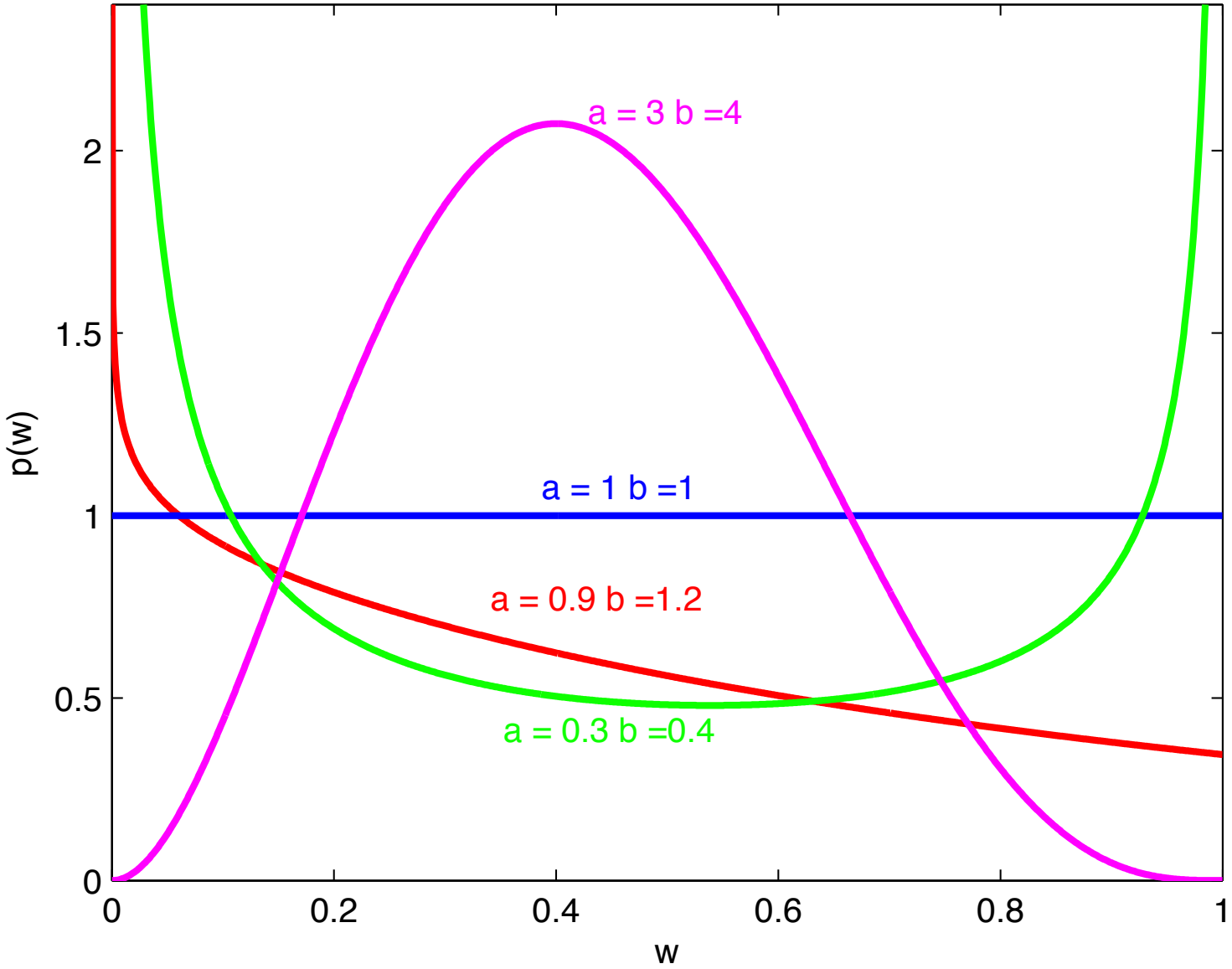
Beta Distribution. $\mathcal{B}(w; a, b)$

$$
\begin{aligned}
\mathcal{B}(w; a, b) &\equiv \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} w^{a-1}(1-w)^{b-1} \\
&= \exp\left((a-1)\log w + (b-1)\log(1-w) - A(a,b)\right) \\
&= \exp\left(\begin{pmatrix} a-1 & b-1 \end{pmatrix} \begin{pmatrix} \log w \\ \log(1-w) \end{pmatrix} - A(a,b) \right) \\
A(a,b) &= \log\Gamma(a) + \log\Gamma(b) - \log\Gamma(a+b)
\end{aligned}
$$

Mean :

$$\langle w \rangle_{\mathcal{B}} = a/(a+b)$$

Beta Distribution. $\mathcal{B}(w; a, b)$



Univariate Gaussian. $\mathcal{N}(x; m, S)$

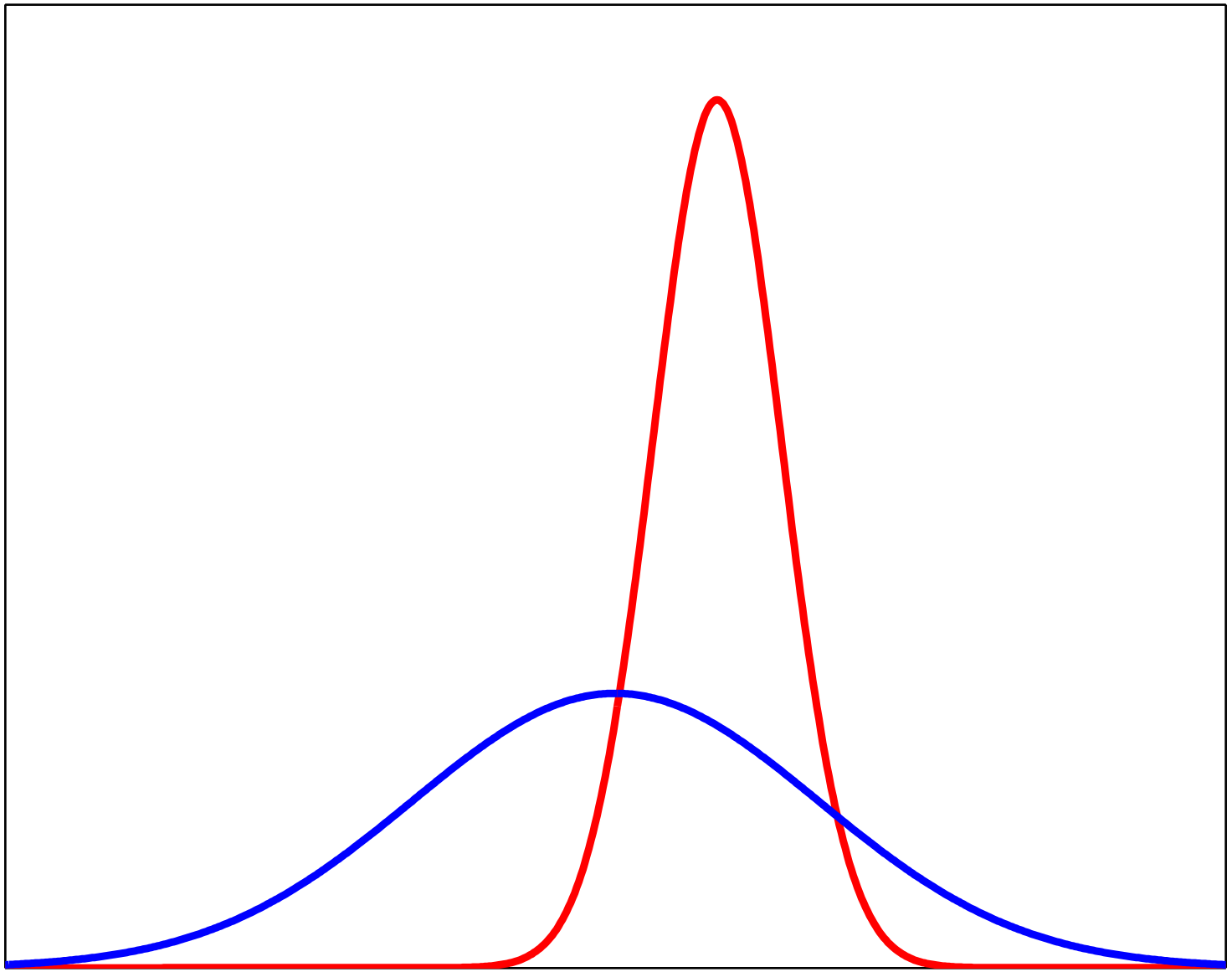
The Gaussian distribution with mean m and covariance S has the form

$$\begin{aligned}
\mathcal{N}(x; m, S) &= (2\pi S)^{-1/2} \exp\{-\frac{1}{2}(x-m)^2/S\} \\
&= \exp\{-\frac{1}{2}(x^2 + m^2 - 2xm)/S - \frac{1}{2}\log(2\pi S)\} \\
&= \exp\left\{\frac{m}{S}x - \frac{1}{2S}x^2 - \left(\frac{1}{2}\log(2\pi S) + \frac{1}{2S}m^2\right)\right\} \\
&= \exp\{\underbrace{\begin{pmatrix} m/S \\ -\frac{1}{2}/S \end{pmatrix}^\top}_{\theta} \underbrace{\begin{pmatrix} x \\ x^2 \end{pmatrix}}_{\psi(x)} - A(\theta)\}
\end{aligned}$$

Hence by matching coefficients we have

$$\exp\left\{-\tfrac{1}{2}Kx^2 + hx + g\right\} \Leftrightarrow S = K^{-1} \qquad m = K^{-1}h$$

Gaussian.



Inverse Gamma Distribution. $\mathcal{IG}(r; a, b)$

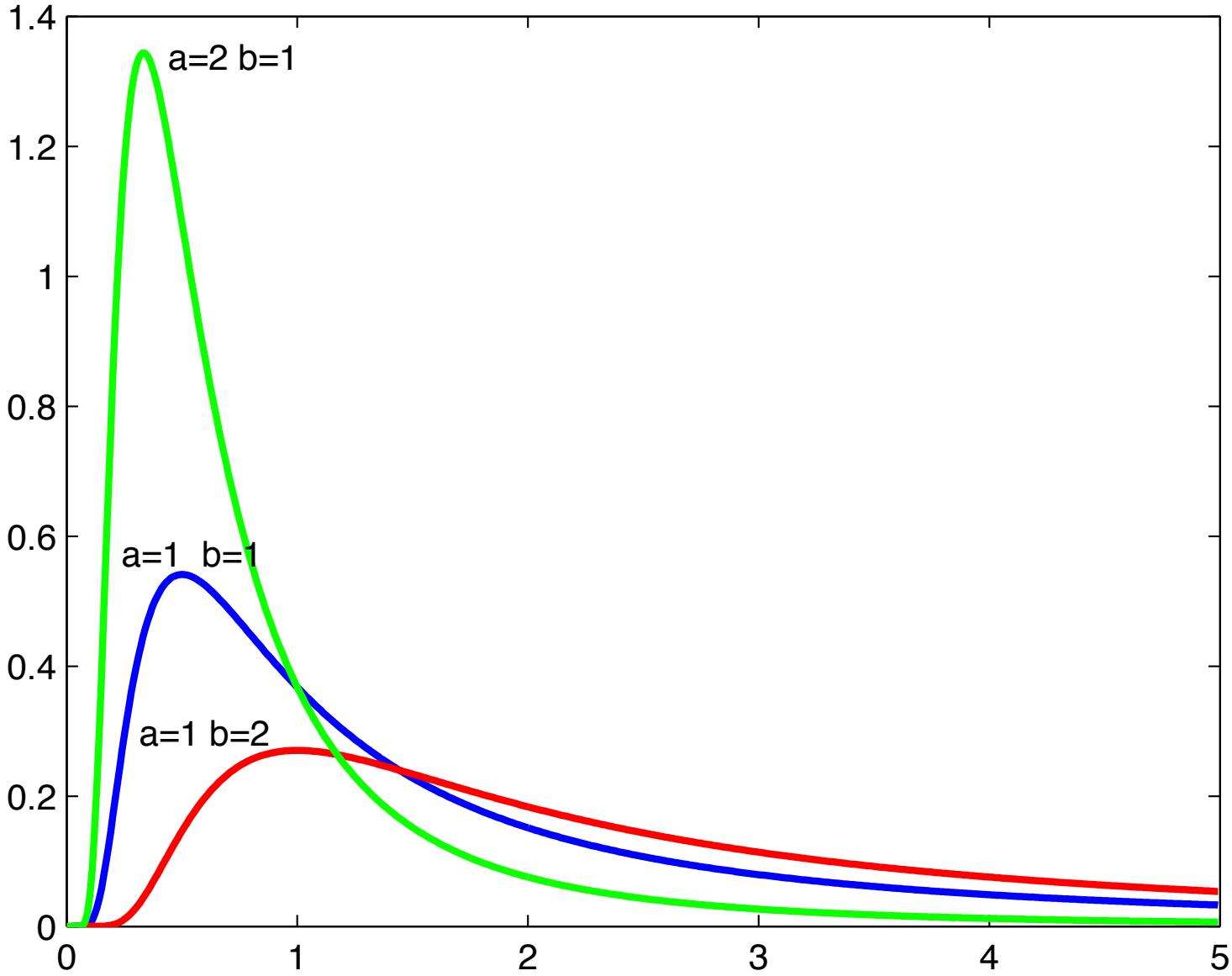
The inverse Gamma distribution with shape a and scale b

$$\begin{aligned}
\mathcal{IG}(r; a, b) &= \frac{1}{\Gamma(a)} \frac{r^{-(a+1)}}{b^{-a}} \exp(-\frac{b}{r}) \\
&= \exp\left(-(a+1)\log r - \frac{b}{r} - \log\Gamma(a) + a\log b\right) \\
&= \exp\left(\begin{pmatrix} -(a+1) \\ -b \end{pmatrix}^\top \begin{pmatrix} \log r \\ 1/r \end{pmatrix} - \log\Gamma(a) + a\log b\right)
\end{aligned}$$

Hence by matching coefficients, we have

$$\exp\left\{\alpha\log r + \beta\frac{1}{r} + c\right\} \Leftrightarrow a = -\alpha - 1 \qquad b = -\beta$$

Inverse Gamma



Gamma Distribution. $\mathcal{G}(\lambda; a, b)$

The Gamma distribution with shape a and **inverse scale** b

$$\begin{aligned}
\mathcal{G}(\lambda; a, b) &= \frac{1}{\Gamma(a)} b^a \lambda^{(a-1)} \exp(-b\lambda) \\
&= \exp\left((a-1)\log\lambda - b\lambda - \log\Gamma(a) + a\log b\right) \\
&= \exp\left(\begin{pmatrix} (a-1) \\ -b \end{pmatrix}^\top \begin{pmatrix} \log\lambda \\ \lambda \end{pmatrix} - \log\Gamma(a) + a\log b\right)
\end{aligned}$$

Hence by matching coefficients, we have

$$\exp\left\{\alpha\log r + \beta\frac{1}{r} + c\right\} \Leftrightarrow a = \alpha + 1 \qquad b = -\beta$$

Gamma and Inverse Gamma Warning

- Unlike a Gaussian, parametrisation is not standard
  - (that may even be true in the slides)
- Beware of the (shape-scale) convention. Somebody can write

$$\mathcal{G}(\cdot; a, b)$$

  but actually mean according to our (shape-inverse scale) convention

$$\mathcal{G}(\cdot; a, 1/b)$$

- Always double check which parametrisation is being used!!!
- Matlab uses the shape-scale convention!!!

Random number generation

- Bernoulli: $\mathcal{BE}(x; p)$

```
x = double(rand<p);
```

- Binomial: $\mathcal{BI}(x; p, N)$

```
x = sum(double(rand(N,1)<p));
```

  Not efficient for large N

- Poisson: $\mathcal{PO}(x; \lambda)$

```
x = poissrnd(lambda);
```

- Beta: $\mathcal{B}(x; a, b)$

```
x = betarnd(a, b);
```

- Gaussian: $\mathcal{N}(x; \mu, S)$

```
x = sqrt(S).*randn(size(S)) + mu;
```

- Gamma: $x \sim \mathcal{G}(x; a, b)$

```
x = gamrnd(a, 1./b);
```

or more securely

```
x = gamrnd(a, 1)./b;
```

which is also

```
x = gamrnd(a)./b;
```

- Inverse Gamma $x \sim \mathcal{IG}(x; a, b)$

```
x = b./gamrnd(a);
```

Conjugate priors: Posterior is in the same family as the prior.

Example: posterior inference for the probability of success w of a binary (Bernoulli) random variable c

$$
\begin{aligned}
p(c|w) &= \mathcal{BE}(c;w) = \exp\left(c\log w + (1-c)\log(1-w)\right) \\
p(w) &= \mathcal{B}(w;a,b)
\end{aligned}
$$

$$
\begin{aligned}
p(w|c) &\propto p(c|w)p(w) \\
&\propto \exp\left(c\log w + (1-c)\log(1-w)\right) \\
&\quad \times \exp\left((a-1)\log w + (b-1)\log(1-w)\right) \\
&\propto \mathcal{B}(w;a+c,b+(1-c))
\end{aligned}
$$

$$
p(w|c) = \begin{cases} \mathcal{B}(w;a+1,b) & c=1 \\ \mathcal{B}(w;a,b+1) & c=0 \end{cases}
$$

Conjugate priors: Posterior is in the same family as the prior.

Example: posterior inference for the variance R of a zero mean Gaussian.

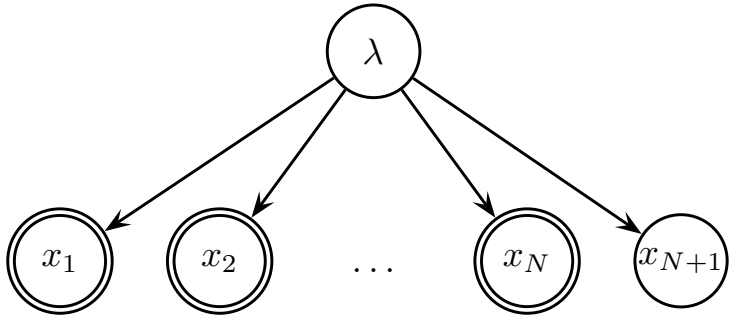
$$\begin{aligned} p(x|R) &= \mathcal{N}(x; 0, R) \\ p(R) &= \mathcal{IG}(R; a, b) \end{aligned}$$

$$\begin{aligned} p(R|x) &\propto p(R)p(x|R) \\ &\propto \exp\left(-(a+1)\log R - b\frac{1}{R}\right)\exp\left(-(x^2/2)\frac{1}{R} - \frac{1}{2}\log R\right) \\ &= \exp\left(\begin{pmatrix} -(a+1+\frac{1}{2}) \\ -(b+x^2/2) \end{pmatrix}^\top \begin{pmatrix} \log R \\ 1/R \end{pmatrix}\right) \\ &\propto \mathcal{IG}(R; a+\frac{1}{2}, b+x^2/2) \end{aligned}$$

Like the prior, this is an inverse-Gamma distribution.

Conjugate priors: Posterior is in the same family as the prior.

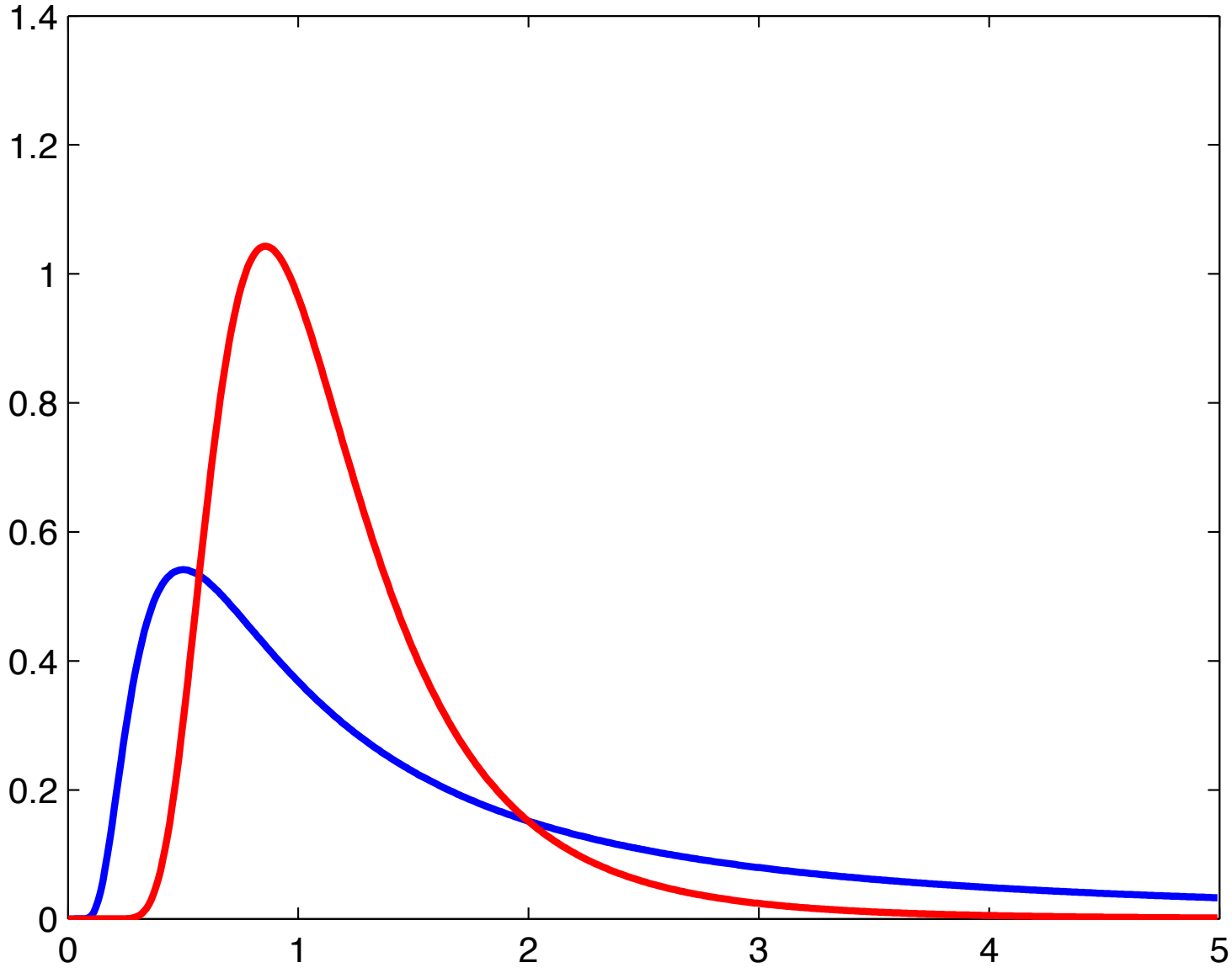
Example: posterior inference of variance R from $x_1, \ldots, x_N$.



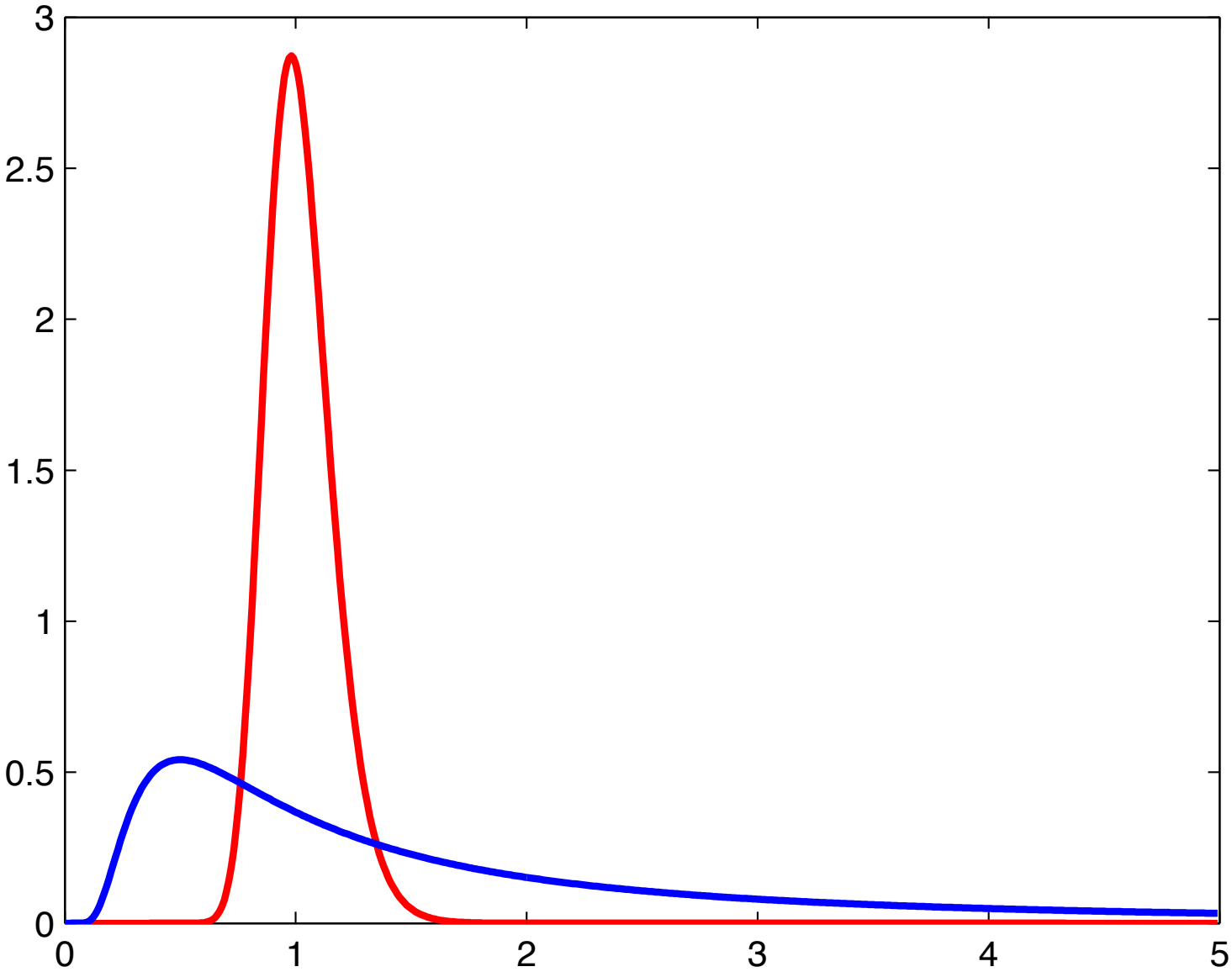
$$
\begin{aligned}
p(R|x) &\propto p(R) \prod_{i=1}^{N} p(x_i|R) \\
&\propto \exp\left(-(a+1)\log R - b\frac{1}{R}\right) \exp\left(-\left(\frac{1}{2}\sum_i x_i^2\right)\frac{1}{R} - \frac{N}{2}\log R\right) \\
&= \exp\left(\begin{pmatrix} -(a+1+\frac{N}{2}) \\ -(b+\frac{1}{2}\sum_i x_i^2) \end{pmatrix}^\top \begin{pmatrix} \log R \\ 1/R \end{pmatrix}\right) \propto \mathcal{IG}(R; a+\frac{N}{2}, b+\frac{1}{2}\sum_i x_i^2)
\end{aligned}
$$

Sufficient statistics are **additive**

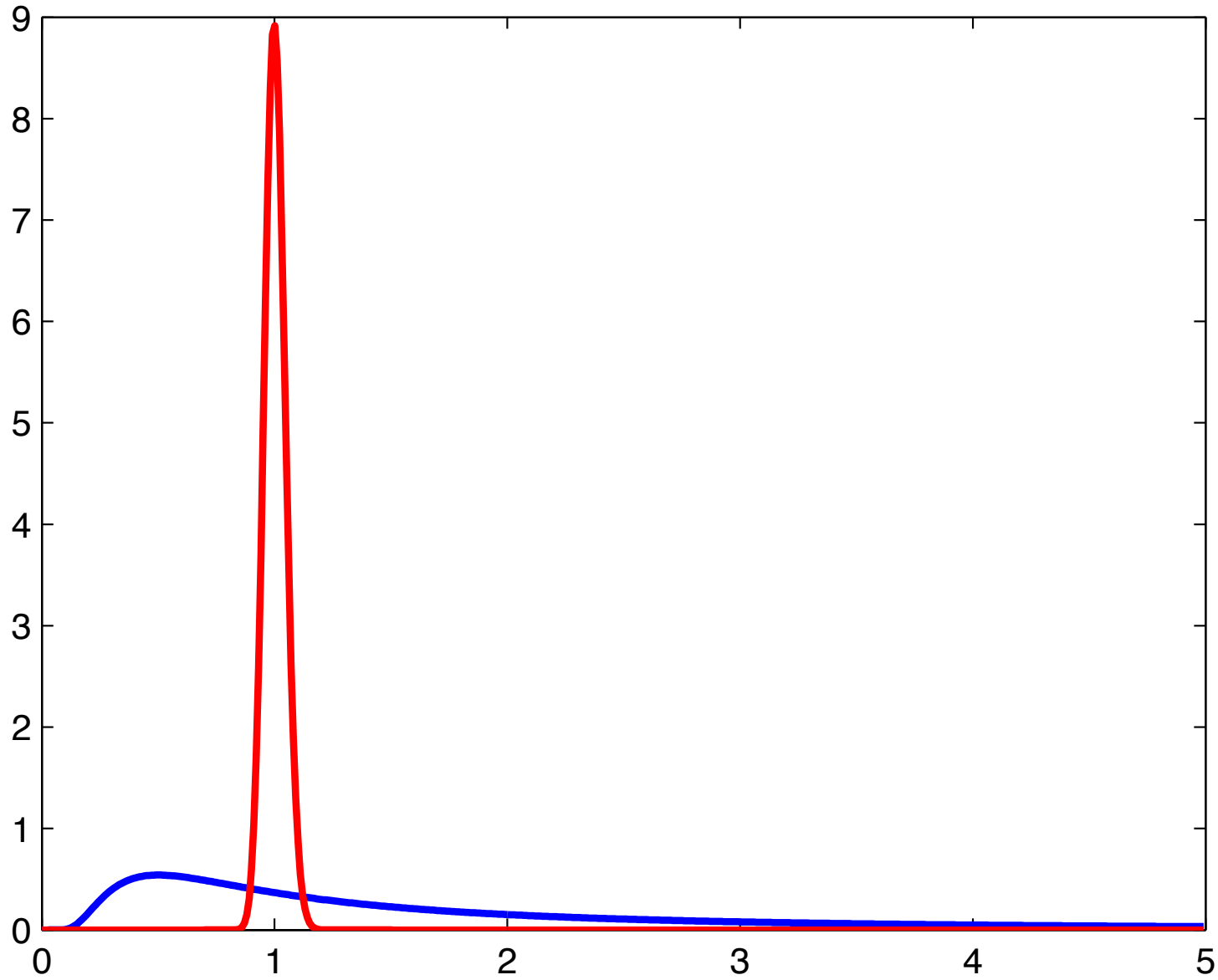
Inverse Gamma, $\sum_i x_i^2 = 10 \quad N = 10$



Inverse Gamma, $\sum_i x_i^2 = 100 \quad N = 100$



Inverse Gamma, $\sum_i x_i^2 = 1000 \quad N = 1000$



Keywords Summary

Bernoulli, Binomial, Multinomial

Gaussian, Beta, Gamma, Inverse Gamma, Poisson

Exponential Family, Canonical Parametrisation, Sufficient statistics