Inference of Probability Distributions for Trust and Security applications

Vladimiro Sassone

Based on joint work with

Mogens Nielsen & Catuscia Palamidessi

Outline

Outline

- Motivations

Outline

- Motivations
- Bayesian vs Frequentist approach

Outline

- Motivations
- Bayesian vs Frequentist approach
- A class of functions to estimate the distribution

Outline

- Motivations
- Bayesian vs Frequentist approach
- A class of functions to estimate the distribution
- Measuring the precision of an estimation function

Motivations

- Inferring the probability distribution of a random variable
- Examples of applications in Trust & Security
  - How much we can trust an individual or a set of individuals
  - Input distribution in a noisy channel to compute the Bayes risk
  - Application of the Bayesian approach to hypothesis testing (anonymity, information flow)
  - ...

Setting and assumptions

- For simplicity we consider only binary random variables
  - honest/dishonest, secure/insecure, ...
- Goal: infer (an approximation of) the probability of success
- Means: Sequence of *n* trials. Observation (*Evidence*) : *s* , *f*

$$X = \{succ, fail\}$$

$$Pr(succ) = \theta$$

$$s = \#succ$$
$$f = \#fail = n - s$$

Using the evidence to infer θ

- The Frequentist method:

$$F(n, s) = \frac{s}{n}$$

- The Bayesian method:

Assume an *a priori* probability distribution for θ (representing your partial knowledge about θ, whatever the source may be) and combine it with the *evidence*, using Bayes' theorem, to obtain the *a posteriori* distribution

Bayesian vs Frequentist

- Criticisms to the frequentist approach
  - *Limited applicability:* sometimes it is not possible to measure the frequencies (in this talk we consider the case in which this is possible)
    - Eg: what is the probability that my submitted paper will be accepted?
  - *Misleading evidence:* For small samples (small n) we can be unlucky, i.e. get unlikely results
    - This is less dramatic for the Bayesian approach because the a priori distribution reduces the effect of a misleading evidence, provided it is close enough to the real distribution

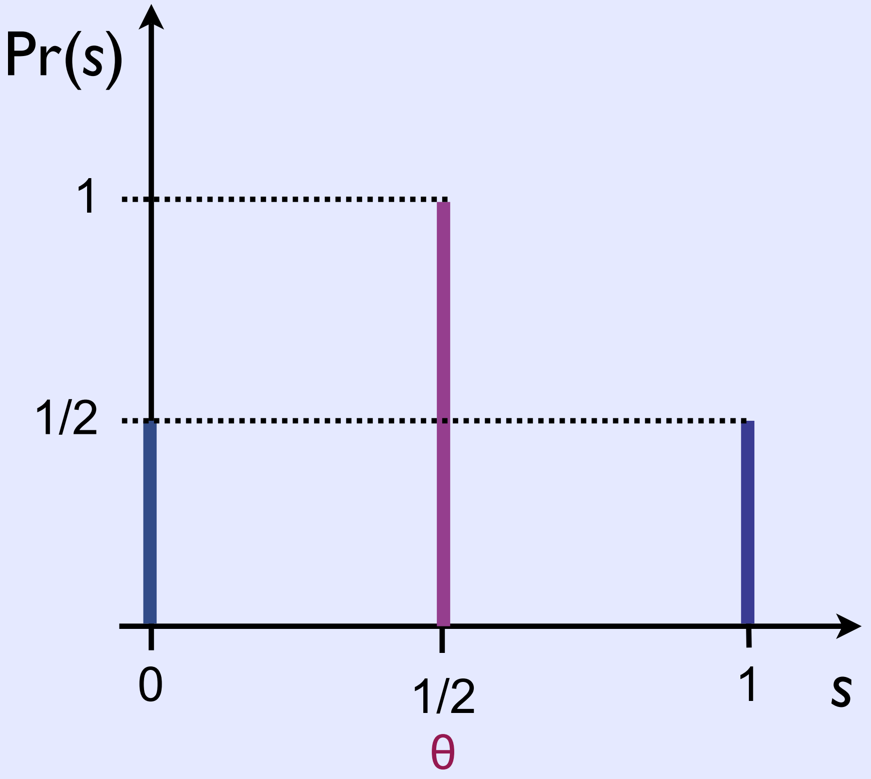
Bayesian vs Frequentist

- Criticisms to the Bayesian approach
  - We need to assume an a priori probability distribution; as we usually do not know the real distribution, the assumption can be somehow arbitrary and differ significantly from reality
- Observe that the two approaches give the same result as *n* tends to infinity: the “true” distribution
  - Frequentist approach: because of the law of large numbers
  - Bayes approach: because the a priori “washes out” for large values of *n*.

Bayesian vs Frequentist

The surprising thing is that the Frequentist approach can be worse than the Bayesian approach even when the trials give a "good" result, or when we consider the average difference (from the "true" θ) wrt all possible results

Example: "true θ" = 1/2, n = 1



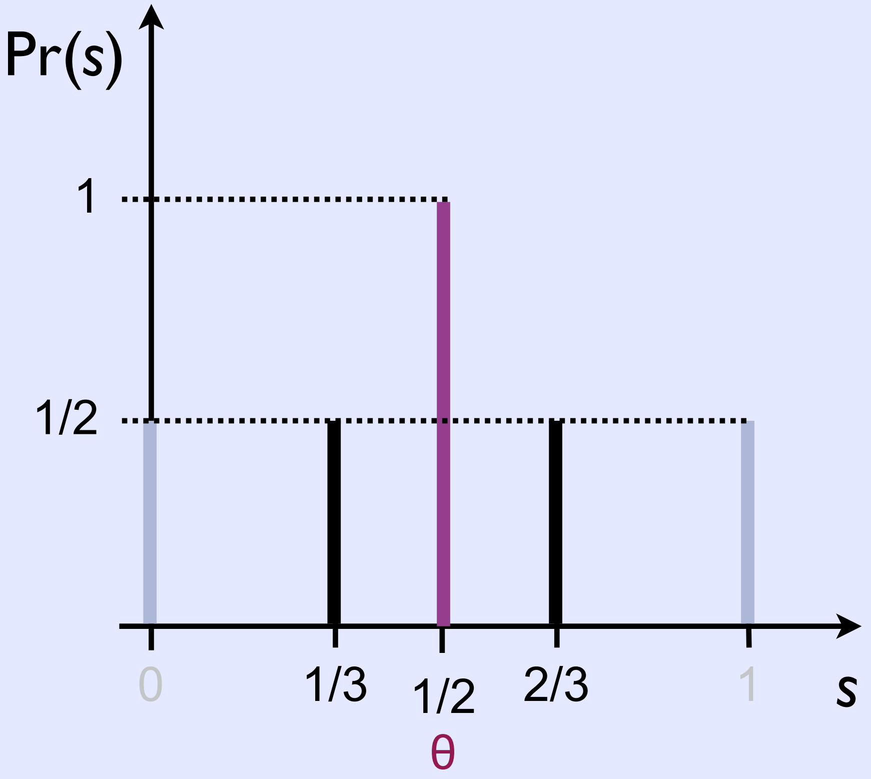
$$F(n,s) = \frac{s}{n} = \begin{cases} 0 & s = 0 \\ 1 & s = 1 \end{cases}$$

The difference from the true distribution is 1/2

Bayesian vs Frequentist

The surprising thing is that the Frequentist approach can be worse than the Bayesian approach even when the trials give a "good" result, or when we consider the average difference (from the "true" θ) wrt all possible results

Example: "true θ" = 1/2, n = 1



$$F(n,s) = \frac{s}{n} = \begin{cases} 0 & s = 0 \\ 1 & s = 1 \end{cases}$$

The difference from the true distribution is 1/2

A better function would be

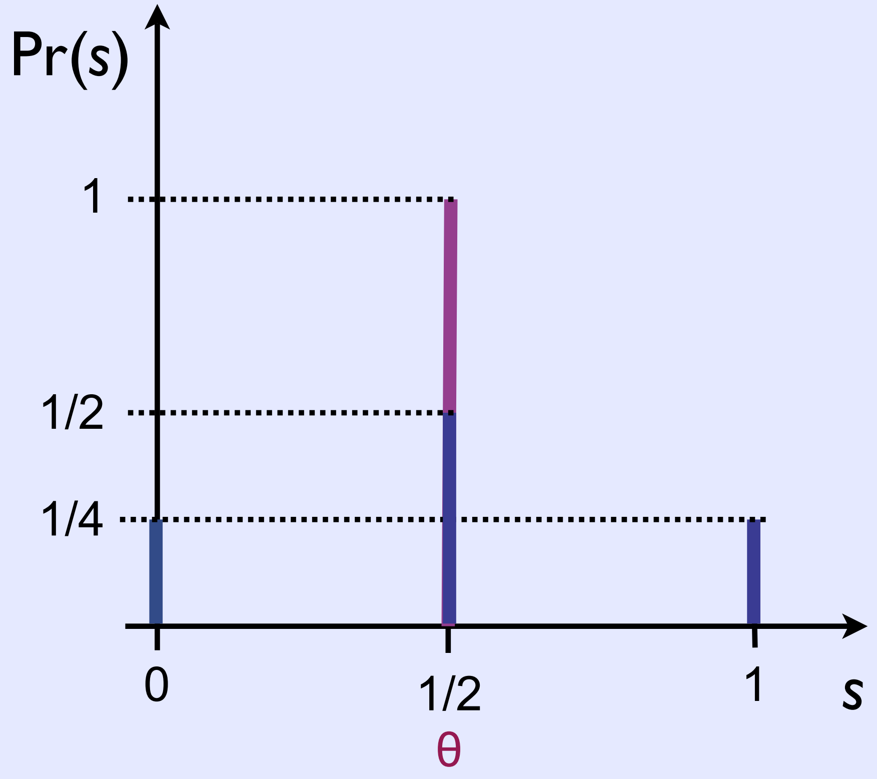
$$F_c(n,s) = \frac{s+1}{n+2} = \begin{cases} \frac{1}{3} & s = 0 \\ \frac{2}{3} & s = 1 \end{cases}$$

The difference from the true distribution is 1/6

Bayesian vs Frequentist

The surprising thing is that the Frequentist approach can be worse than the Bayesian approach even when the trials give a "good" result, or when we consider the average difference (from the "true" θ) wrt all possible results

Example: "true θ" = 1/2, n = 2



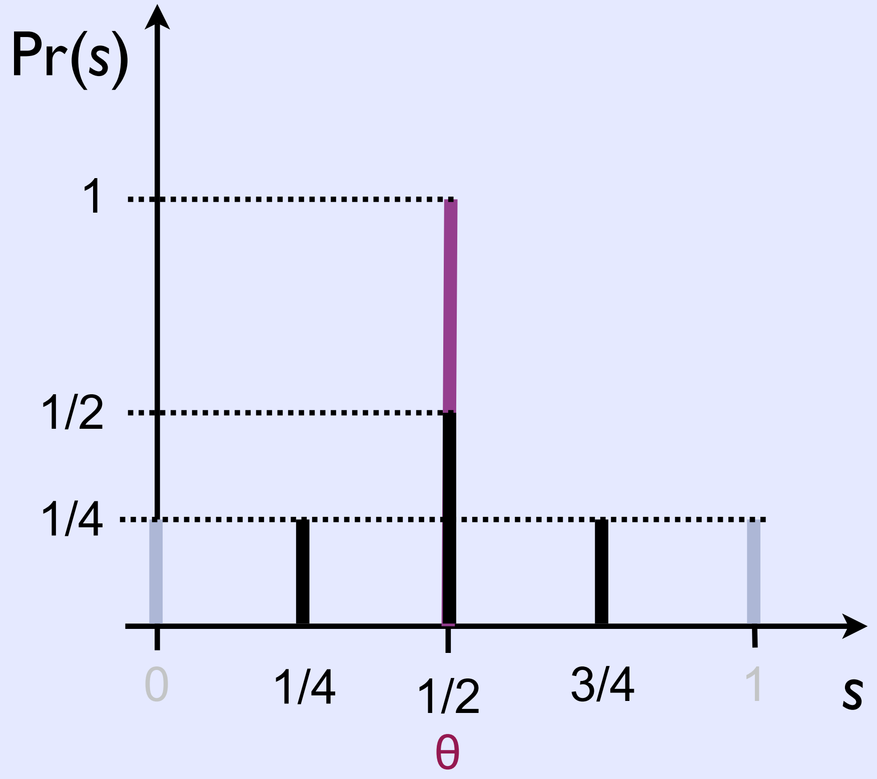
$$F(n,s) = \frac{s}{n} = \begin{cases} 0 & s = 0 \\ \frac{1}{2} & s = 1 \\ 1 & s = 2 \end{cases}$$

The average difference from the true distribution is 1/4

Bayesian vs Frequentist

The surprising thing is that the Frequentist approach can be worse than the Bayesian approach even when the trials give a "good" result, or when we consider the average difference (from the "true" θ) w.r.t. all possible results

Example: "true θ" = 1/2, n = 2



$$F(n,s) = \frac{s}{n} = \begin{cases} 0 & s=0 \\ \frac{1}{2} & s=1 \\ 1 & s=2 \end{cases}$$

The average distance from the true distribution is 1/4

Again, a better function would be

$$F_c(n,s) = \frac{s+1}{n+2} = \begin{cases} \frac{1}{4} & s=0 \\ \frac{1}{2} & s=1 \\ \frac{3}{4} & s=2 \end{cases}$$

The average distance from the true distribution is 1/8

Bayesian vs Frequentist

Bayesian vs Frequentist

- We will see that $F_c(s,n) = (s+1)/(n+2)$ corresponds to one of the possible Bayesian approaches.

Bayesian vs Frequentist

- We will see that $F_c(s,n) = (s+1)/(n+2)$ corresponds to one of the possible Bayesian approaches.
- Of course, if the "true" θ is different from 1/2 then F_c can be worse than F

Bayesian vs Frequentist

- We will see that $F_c(s,n) = (s+1)/(n+2)$ corresponds to one of the possible Bayesian approaches.
- Of course, if the "true" θ is different from 1/2 then F_c can be worse than F
- And, of course, the problem is that we don't know what θ is (the value θ is exactly what we are trying to find out!).

Bayesian vs Frequentist

- We will see that $F_c(s,n) = (s+1)/(n+2)$ corresponds to one of the possible Bayesian approaches.
- Of course, if the "true" θ is different from 1/2 then F_c can be worse than F
- And, of course, the problem is that we don't know what θ is (the value θ is exactly what we are trying to find out!).
- However, F_c is still better than F if we consider the average distance wrt all possible $\theta \in [0,1]$, assuming that they are all equally likely (i.e. that θ has a uniform distribution)

Bayesian vs Frequentist

- We will see that $F_c(s,n) = (s+1)/(n+2)$ corresponds to one of the possible Bayesian approaches.
- Of course, if the "true" θ is different from 1/2 then F_c can be worse than F
- And, of course, the problem is that we don't know what θ is (the value θ is exactly what we are trying to find out!).
- However, F_c is still better than F if we consider the average distance wrt all possible $\theta \in [0,1]$, assuming that they are all equally likely (i.e. that θ has a uniform distribution)
- In fact we can prove that, under a suitable notion of "difference", and for θ uniformly distributed, F_c is the best function of the kind $G(s,n) = (s+t)/(n+m)$

A Bayesian approach

- **Assumption**: θ is the generic value of a continuous random variable Θ whose probability density is a *Beta distribution* with (unknown) parameters σ , φ

$$B(\sigma,\varphi)(\theta) = \frac{\Gamma(\sigma+\varphi)}{\Gamma(\sigma)\Gamma(\varphi)} \ \theta^{\sigma-1}(1-\theta)^{\varphi-1}$$

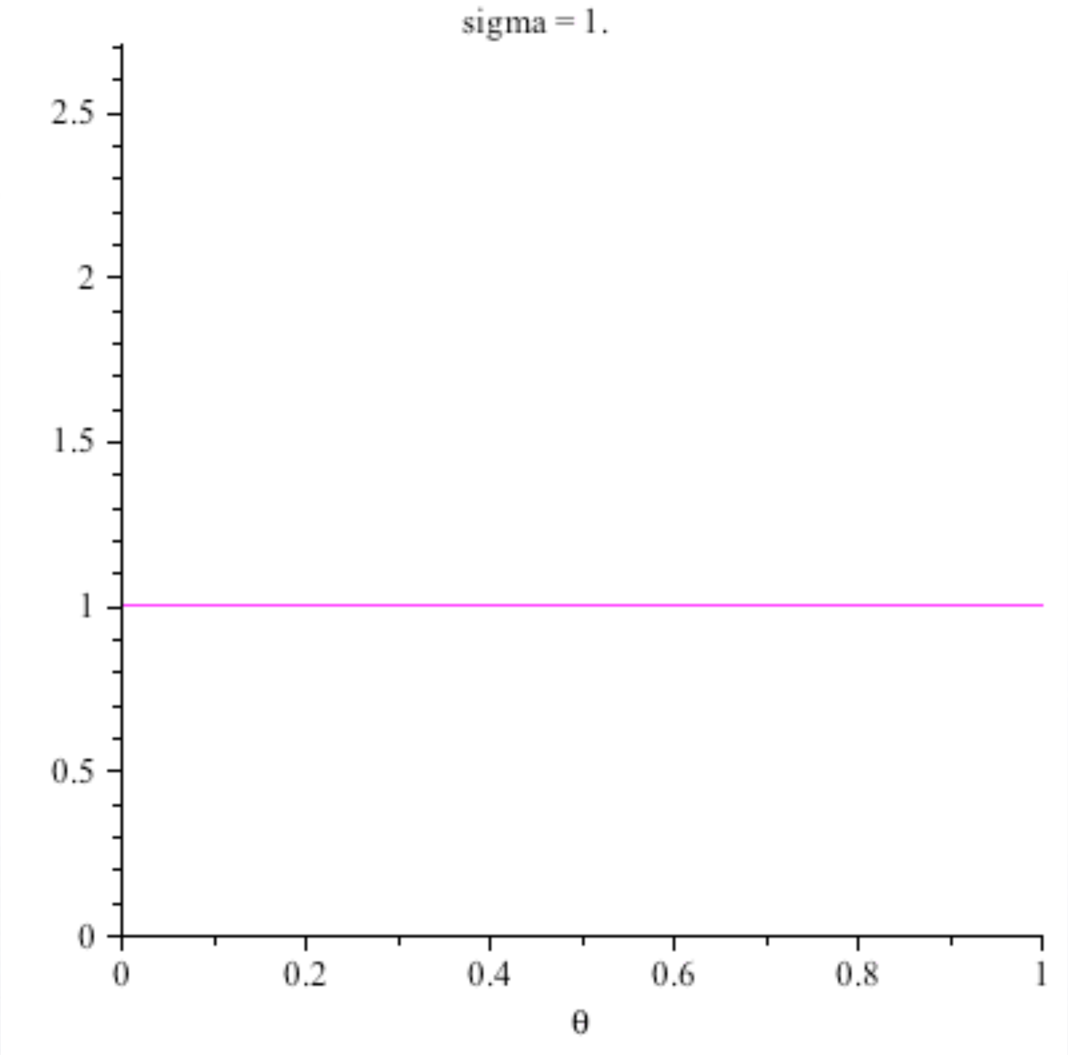
where Γ is the extension of the factorial function
i.e. $\quad \Gamma(n) = (n-1)! \qquad$ for n natural number

- Note that the uniform distribution is a particular case of Beta distribution, with $\sigma = 1$, $\varphi = 1$

- B(σ , φ) can be seen as the a posteriori probability density of Θ given by a uniform a priori (principle of maximum entropy) and a trial sequence resulting in σ -1 successes and φ -1 failures.

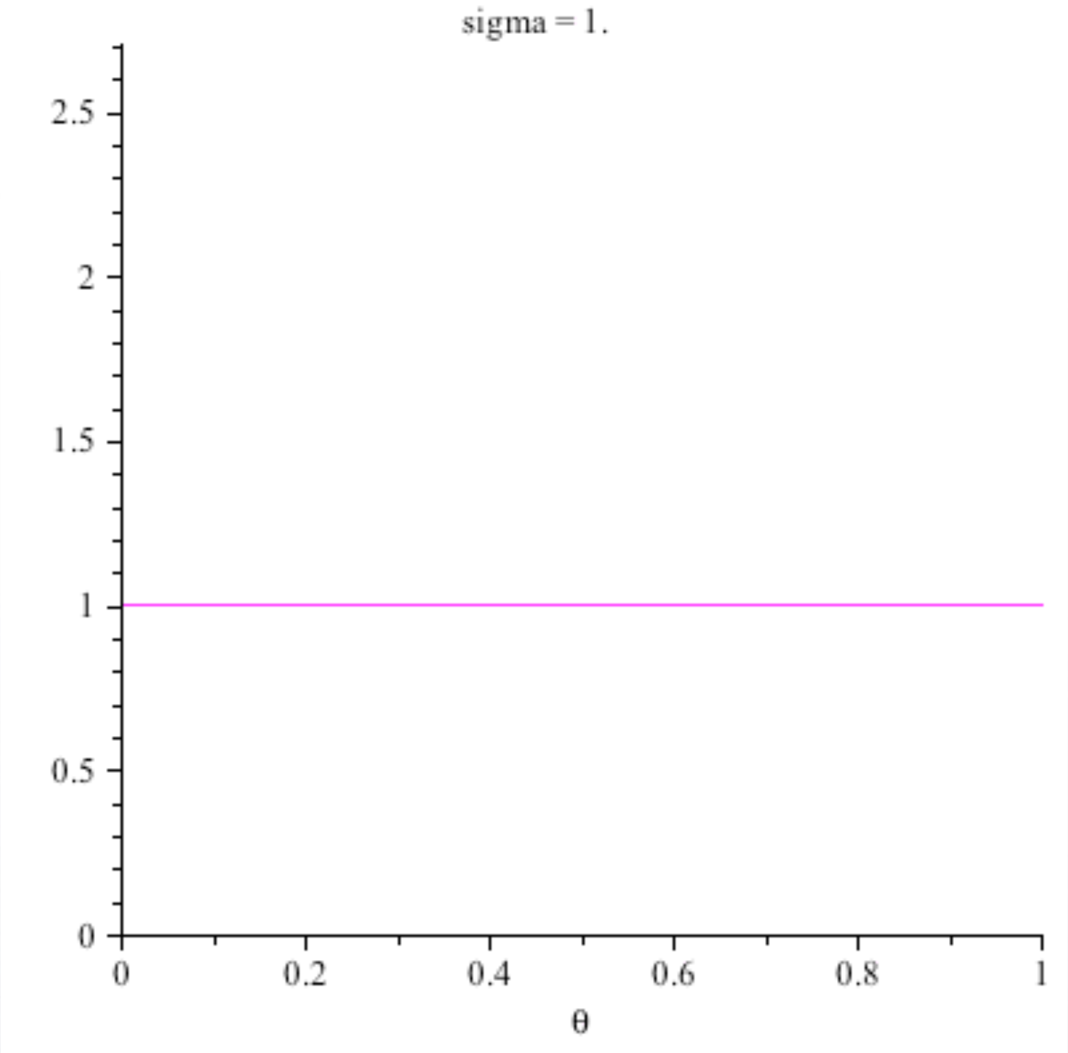
Examples of Beta Distribution

Examples of Beta Distribution

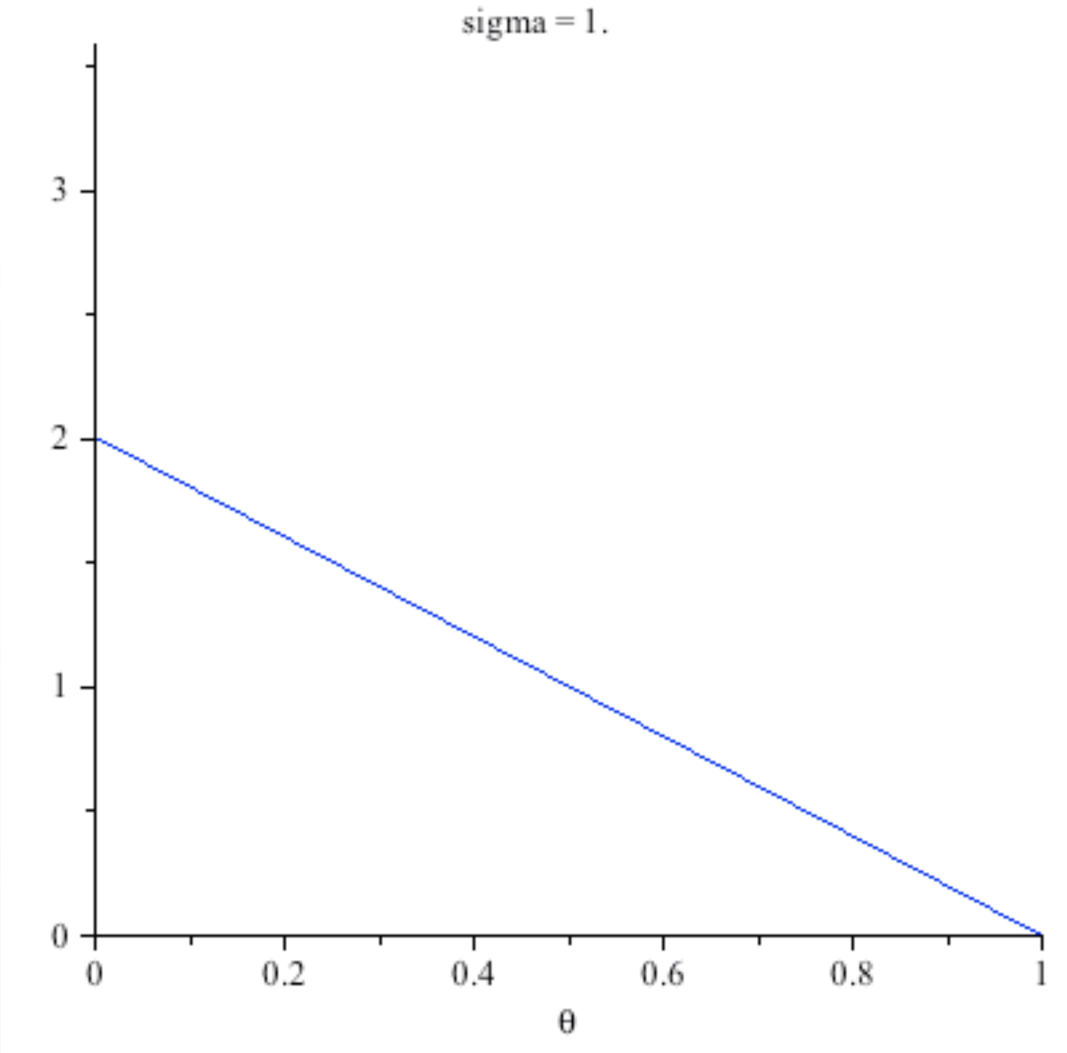


$\sigma = \ \varphi = 1 \ .. \ 6$

Examples of Beta Distribution

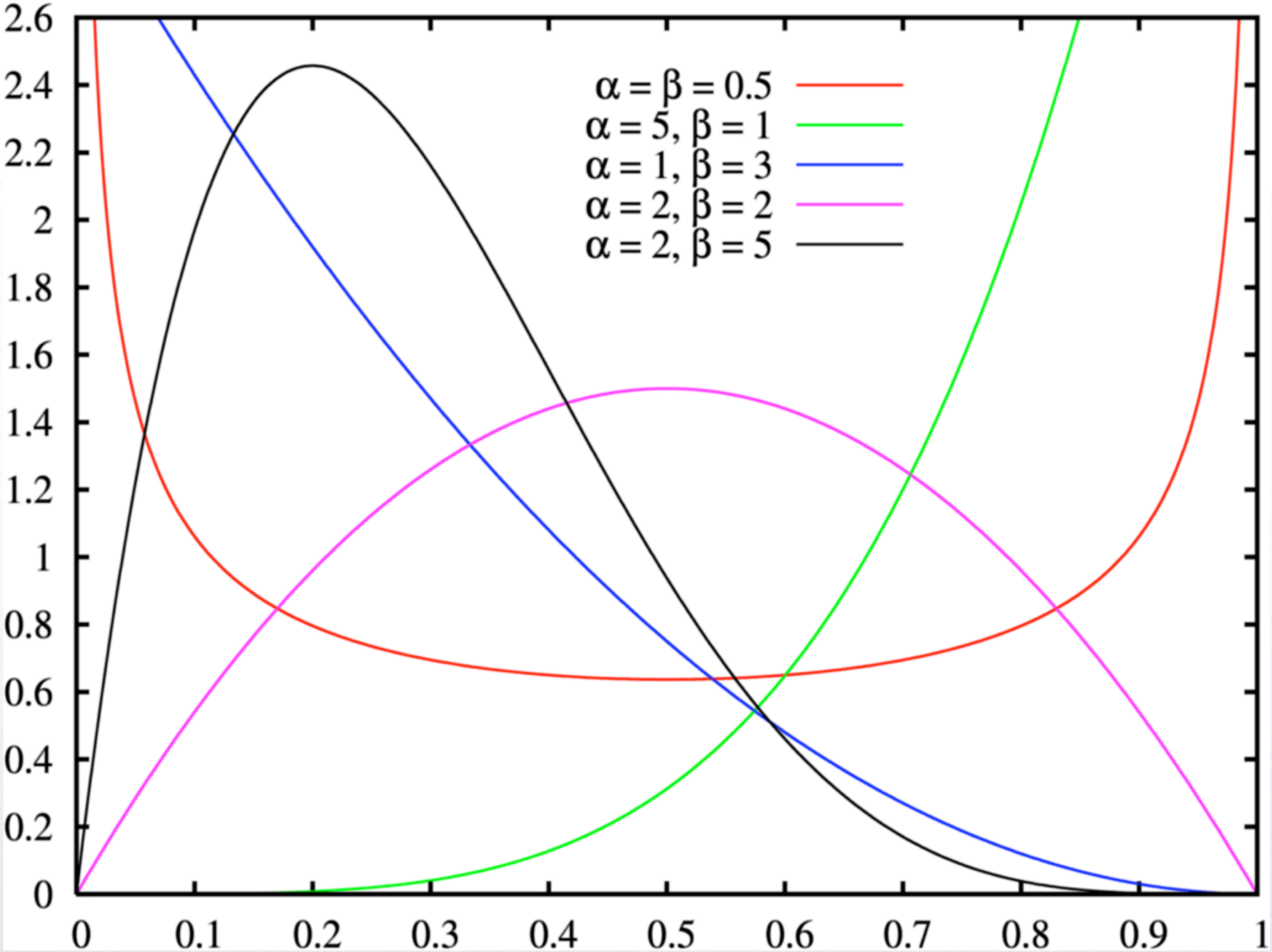


$\sigma = \ \varphi = 1 \ .. \ 6$



$\sigma = 1 \ .. \ 6 \quad \varphi = 2\sigma$

Other examples of Beta Distribution



The Bayesian Approach

- Assume an *a priori* probability distribution for *Θ* (representing our partial knowledge about *Θ*, whatever the source may be) and combine it with the *evidence*, using Bayes' theorem, to obtain the *a posteriori* probability distribution

$$\underset{\substack{\uparrow \\ \text{a posteriori}}}{Pd(\theta\,|\,s)} = \frac{\overset{\substack{\text{likelihood} \\ \downarrow}}{Pr(s\,|\,\theta)}\ \overset{\substack{\text{a priori} \\ \downarrow}}{Pd(\theta)}}{\underset{\substack{\nearrow \\ \text{evidence}}}{Pr(s)}}$$

- One possible definition for the estimation function (*algorithm*) is the mean of the a posteriori distribution

$$A(n,s) = E_{Pd(\theta|s)}(\Theta) = \int_0^1 \theta\ Pd(\theta|s)\ d\theta$$

The Bayesian Approach

- Since the distribution of Θ is assumed to be a beta distribution *B*(σ , φ), it is natural to take as *a priori* a function of the same class, i.e. *B*(α , β).

  - In general we don't know the "real parameters" σ , φ , hence α , β may be different from σ , φ

- The likelihood *Pr*(*s* | θ) is a binomial, i.e.

$$Pr(s \mid \theta) = \left(\begin{array}{c} s+f \\ s \end{array} \right) \theta^s \, (1-\theta)^f$$

- The Beta distribution is a conjugate of the binomial, which means that the application of Bayes theorem gives as *a posteriori* a function of the same class, and more precisely

$$Pd(\theta \mid s) = B(\alpha + s, \beta + f)$$

The Bayesian Approach

- Summarizing, we are considering three probability density functions for Θ:
  - $B(\sigma,\varphi)$: the "real" distribution of Θ
  - $B(\alpha,\beta)$: the *a priori* (the distribution of Θ up to our best knowledge)
  - $B(s+\alpha, f+\beta)$: the *a posteriori*

- The result of the mean-based algorithm is :

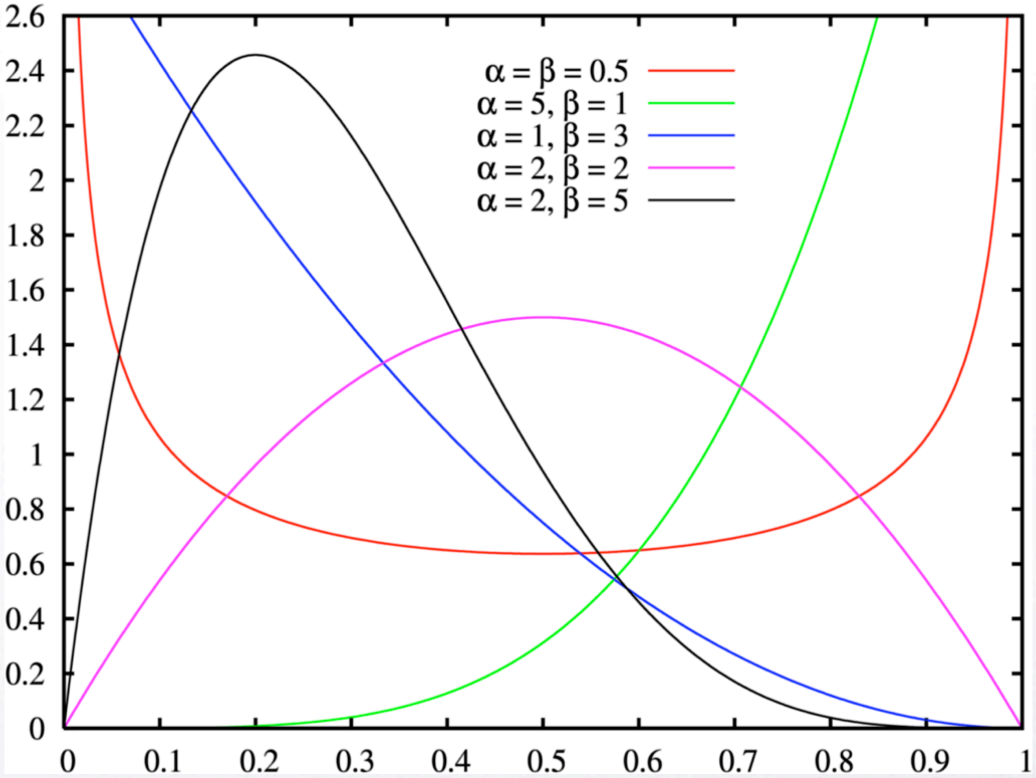
$$A_{\alpha,\beta}(n,s) = E_{B(s+\alpha,f+\beta)}(\Theta) = \frac{s+\alpha}{s+f+\alpha+\beta} = \frac{s+\alpha}{n+\alpha+\beta}$$

The Bayesian Approach

- The frequentist method can be seen as the limit of the Bayesian mean-based algorithms, for $\alpha, \beta \to 0$

- Intuitively, the Bayesian mean-based algorithms give the best result for $\alpha/(\alpha + \beta) = \theta$ and $\alpha, \beta \to \infty$

- How can we compare two Bayesian algorithms in general, i.e. independently of θ?



Measuring the precision of Bayesian algorithms

- Define a "difference" $D(A(n,s), \theta)$ (possibly a distance, but not necessarily. It does not need to be symmetric)
  - non-negative
  - zero iff $A(n,s) = \theta$
  - what else?
- Consider the expected value $D_E(A,n, \theta)$ of $D(A(n,s), \theta)$ with respect to the likelihood (the conditional probability of s given θ)

$$D_E(A,n,\theta) = \sum_{s=0}^{n} Pr(s \mid \theta)\ D(A(n,s),\theta)$$

- **Risk of *A*** : the expected value $R(A,n)$ of $D_E(A,n, \theta)$ with respect to the "true" distribution of Θ

$$R(A,n) = \int_0^1 Pd(\theta)\ D_E(A,n,\theta)\ d\theta$$

Measuring the precision of Bayesian Algorithms

- Note that the definition of "Risk of *A*" is general, i.e. it is a natural definition for any estimation algorithm (not necessarily Bayesian or mean-based)

- What other conditions should *D* satisfy?

- It seems natural to require that *D* be such that $R(A,n)$ has a minimum (for all n's) when the a priori distribution coincides with the "true" distribution

- It is not obvious that such *D* exists

Measuring the precision of Bayesian Algorithms

We have considered the following candidates for $D(x,y)$ (all of which can be extended to the n-ary case):

- The norms:
  - $|x - y|$
  - $|x - y|^2$
  - ...
  - $|x - y|^k$
  - ...
- The Kullback-Leibler divergence

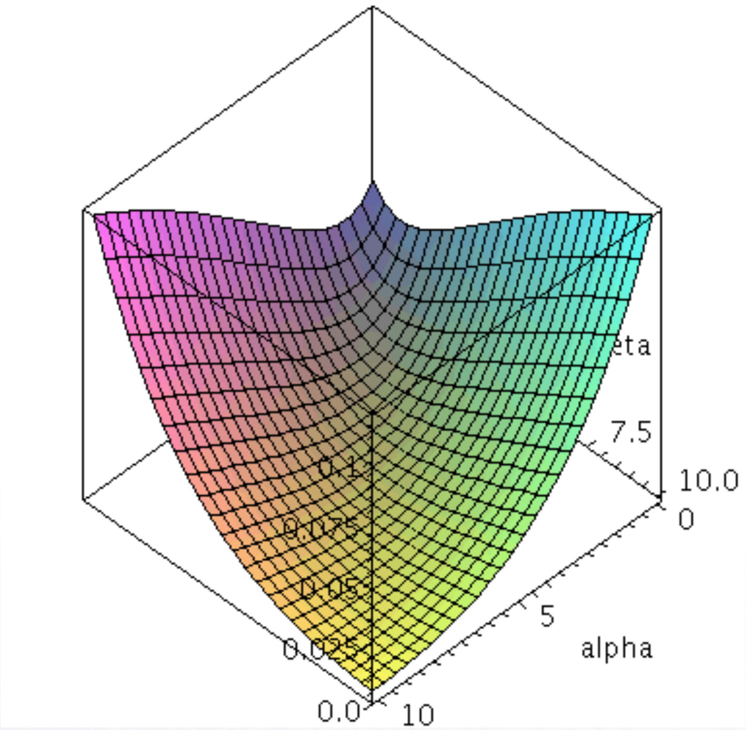
$$D_{KL}((y, 1-y) \parallel (x, 1-x)) = y\ \log_2 \frac{y}{x} + (1-y)\log_2 \frac{1-y}{1-x}$$

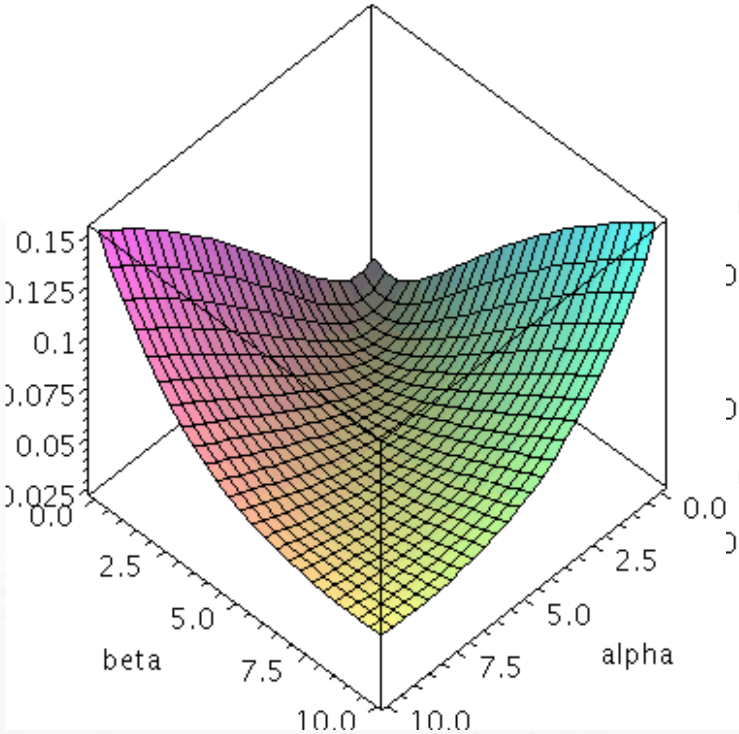
Measuring the precision of Bayesian algorithms

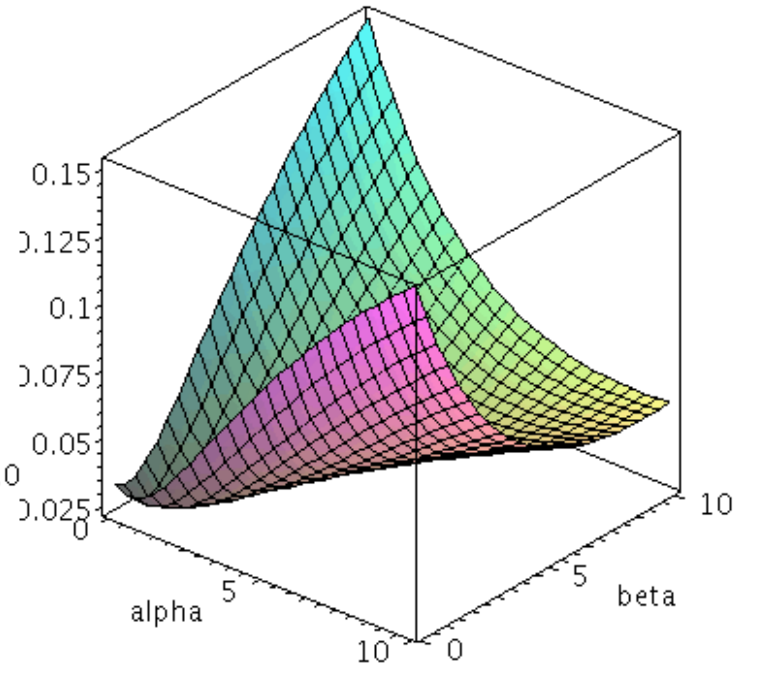
- **Theorem.** For the mean-based Bayesian algorithms, with *a priori* $B(\alpha, \beta)$, we have that the condition is satisfied (i.e. the Risk is minimum when α, β coincide with the parameters σ, φ of the "true" distribution), by the following functions:
  - The 2nd norm $(x - y)^2$
  - The Kullback-Leibler divergence
- We find it very surprising that the condition is satisfied by these two very different functions, and not by any of the other norms $|x - y|^k$ for $k \neq 2$

$$D(x,y) = (x-y)^2$$

$$\sigma = 1, \varphi = 1$$







$D_E(A_{\alpha,\beta}, 5, 1/2)$

$n = 5, \theta = 1/2$

$R(A_{\alpha,\beta}, 5)$

$n = 5$

For the Kullback-Leibler divergence the plots are similar, but much more steep, and they diverge for $\alpha \rightarrow 0$ or $\beta \rightarrow 0$

Work in progress

- Note that for the 2nd norm $D(x,y) = (x\text{-}y)^2$ the average D_E is a distance. This contrasts with the case of $D(x,y) = D_{KL}(y||x)$ and makes the first more appealing.
- How robust is the theorem that "certifies" that the 2nd-norm-based D_E is a "good" distance? In particular:
  - Does it extend to the case of multi-valued random variables?
  - Note that in the multi-valued case the likelihood is a *multinomial*, the conjugate a priori is a *Dirichelet* and the D is the *Euclidian distance* (squared)
- What are the possible applications?

Possible applications (work in progress)

- We can use D_E to compare two different estimation algorithms.
  - Mean-based vs other ways of selecting a θ
  - Bayesian vs non-Bayesian
  - In more complicated scenarios there may be different Bayesian mean-based algorithms. Example: noisy channel.
- D_E induces a metric on distributions. Bayes' equations define transformations on this metric space from the a priori to the a posteriori. We intend to study the properties of such transformations in the hope that they will reveal interesting properties of the corresponding Bayesian methods, independent of the a priori.