# Doing Bayesian Data Analysis
# Chapter 6

Chao-En Yu

National Tsing Hua University

# Exact Analysis

- This chapter presents an example of how to do Bayesian inference using analytical mathematics without any approximation.

- Though not useful for complex applications, but:

1. Nicely Reveal the underlying concepts of Bayesian inference on a continuous parameter.

2. Introduce some repeatedly used distributions.

# Assumptions

- The space of possibilities for each datum has just two nominal and mutually exclusive values.
- Each observed datum is independent of the others.
- The underlying probability is stationary through time.
- Except coin flipping, the methods can be applied to many other real-world situations.

# Distribution vs. Likelihood

- When tossing $n$ coins, the probability of observing $y$ heads is (sp.factorial(n) helps!):

$$p(y|p) = \binom{n}{y} p^y (1-p)^{n-y}$$

- This formula expresses the binomial distribution.

- However, when $y$ is known and fixed, it becomes a function of a continuous parameter $p$, and it is called the Binomial likelihood function. (not prob?)

# For Tractability

1. The prior and posterior beliefs follow the same class of distributions. That is, the numerator of Bayes' rule has the same form as the posterior.

2. The denominator of Bayes' rule can be solved analytically.

- When $p(\theta)$ and $p(y|\theta)$ combine to have the same form as the prior, $p(\theta)$ is called a conjugate prior for $p(y|\theta)$.

# Conjugacy

- The prior is conjugate only with a specific likelihood function.
- It is just a convenient desideratum.
- With conjugacy, prior and posterior belongs to the same distribution with different parameters.
- A proper prior needs to have the form of $\theta^a(1-\theta)^b$

# Beta Distribution

- A probability density of that form is called a beta distribution:

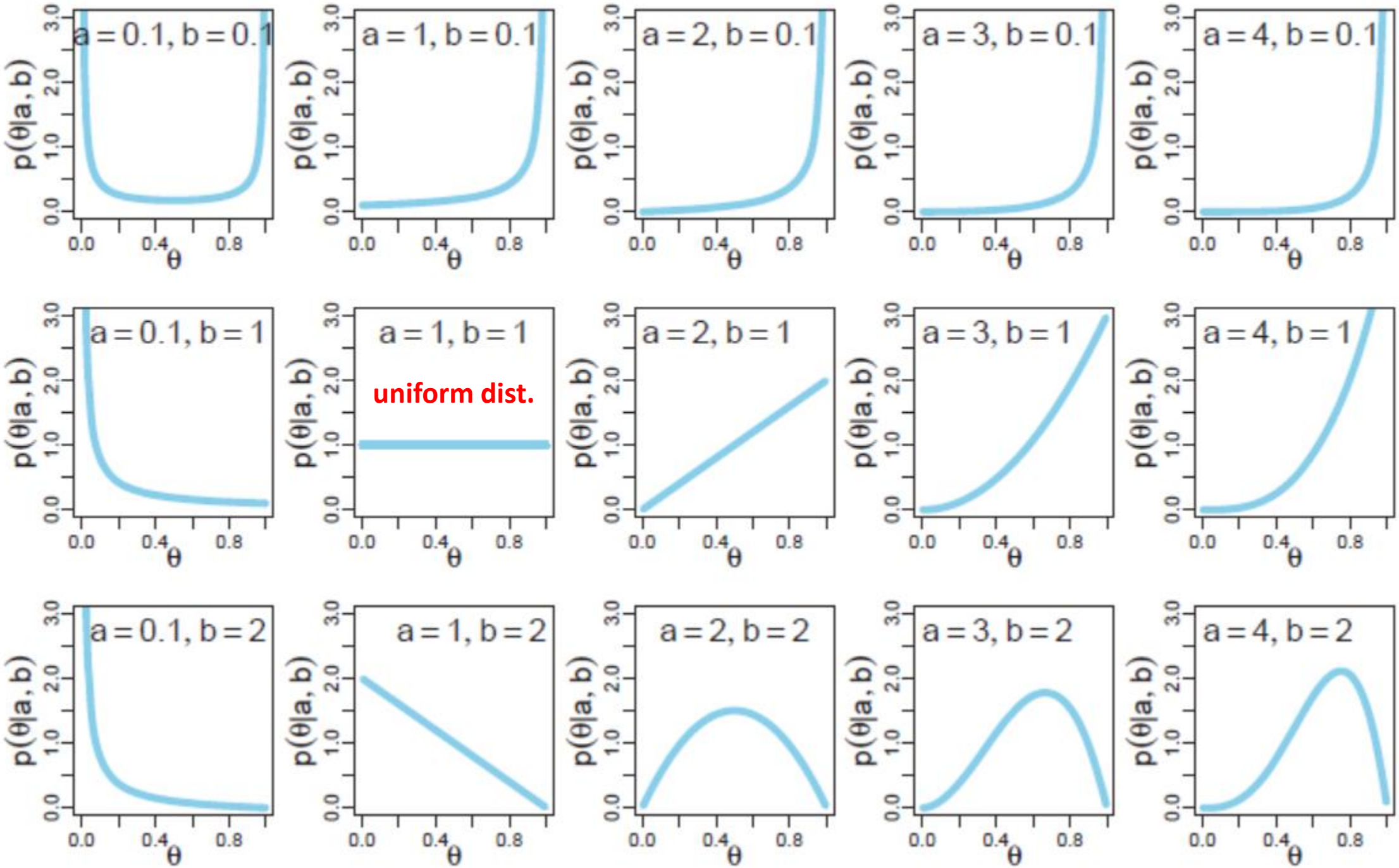
$$p(\theta|a,b) = \theta^{a-1}(1-\theta)^{b-1}/\mathrm{B}(a,b)$$

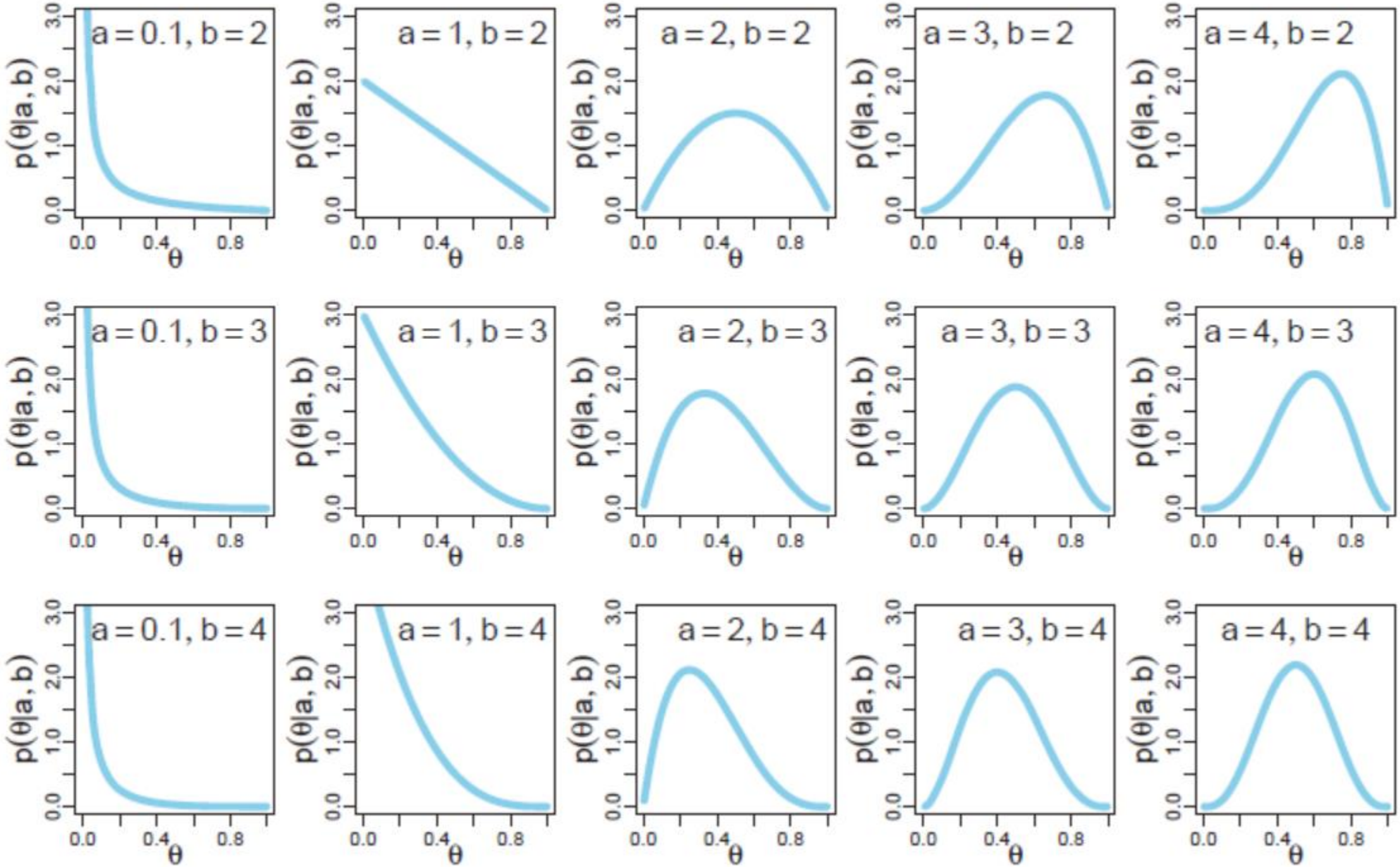
- $\mathrm{B}(a,b) \equiv \int_0^1 d\theta\ \theta^{a-1}(1-\theta)^{b-1}$ is a beta function.

- Beta distribution is defined only for $\theta \in [0,1]$.
  $a$ and $b$ (shape parameters) must be positive.
  The power of $\theta$ is $a-1$ rather than $a$.

# Beta Function

- Recall that $\mathrm{B}(a,b) \equiv \int_0^1 d\theta\ \theta^{a-1}(1-\theta)^{b-1}$ is the beta function. Is it a function of $\theta$?
- We will see the term “integrate out” a lot this semester.
- The difficulty of solving this integral is the reason why Bayesian overlaps CS (see p.116).
- Relationship between Bayesian and ML?


a = 0.1, b = 0.1
a = 1, b = 0.1
a = 2, b = 0.1
a = 3, b = 0.1
a = 4, b = 0.1
a = 0.1, b = 1
a = 1, b = 1
uniform dist.
a = 2, b = 1
a = 3, b = 1
a = 4, b = 1
a = 0.1, b = 2
a = 1, b = 2
a = 2, b = 2
a = 3, b = 2
a = 4, b = 2
p(θ|a, b)
θ
0.0
0.4
0.8
1.0
2.0
3.0


a = 0.1, b = 2
a = 1, b = 2
a = 2, b = 2
a = 3, b = 2
a = 4, b = 2
a = 0.1, b = 3
a = 1, b = 3
a = 2, b = 3
a = 3, b = 3
a = 4, b = 3
a = 0.1, b = 4
a = 1, b = 4
a = 2, b = 4
a = 3, b = 4
a = 4, b = 4
p(θ|a, b)
θ
0.0
1.0
2.0
3.0
0.4
0.8

# Interpretation of a Beta Prior

- Often we think of our prior beliefs in terms of a central tendency and certainty about that central tendency.

- You can think of $a$ and $b$ in the prior as if they were previously observed data.

- The certainty is expressed by the "concentration" $\kappa = a + b$, where $\mu = a/\kappa$ and $\omega = (a - 1)/(\kappa - 2)$.

# The Posterior Beta

- By Bayes' rule

$$p(\theta|z,N) = \frac{p(z,N|\theta)p(\theta)}{p(z,N)}$$

$$= \theta^{z}(1-\theta)^{N-z}\frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}\frac{1}{p(z,N)}$$

$$= \frac{\theta^{(z+a)-1}(1-\theta)^{(N-z+b)-1}}{[B(a,b)p(z,N)]}$$

$$= \frac{\theta^{(z+a)-1}(1-\theta)^{(N-z+b)-1}}{B(z+a,N-z+b)}$$

Observing that the numerator is the numerator of a beta$(\theta|z+a,N-z+b)$ and the posterior is a probability density.

# Compromise of Prior and LKH

- The posterior is always a *compromise* between the prior and the likelihood.

$$\underbrace{\frac{z+a}{N+a+b}}_{\text{posterior}} = \underbrace{\frac{z}{N}}_{\text{data}} \underbrace{\frac{N}{N+a+b}}_{\text{weight}} + \underbrace{\frac{a}{a+b}}_{\text{prior}} \underbrace{\frac{a+b}{N+a+b}}_{\text{weight}}.$$

- The mixing weight on the data proportion increases as $N$ increases, so the posterior mean gets closer to the data proportion.

# Prior that cannot be a Beta

- The beauty of using a beta distribution to express prior knowledge is that the posterior distribution is again exactly a beta distribution.
- Not all prior knowledge can be expressed by a beta distribution.
- Then we must use a different method to derive the posterior. Ex. grid approximation.

# Limitations

- Only simple likelihood functions have conjugate priors. In realistic applications, the complex models have no conjugate priors.
- Even if a conjugate prior exists, not all prior knowledge can be expressed in the mathematical form of the conjugate prior.
- We have to abandon exact solutions, and instead use Markov chain Monte Carlo (MCMC) methods, or …