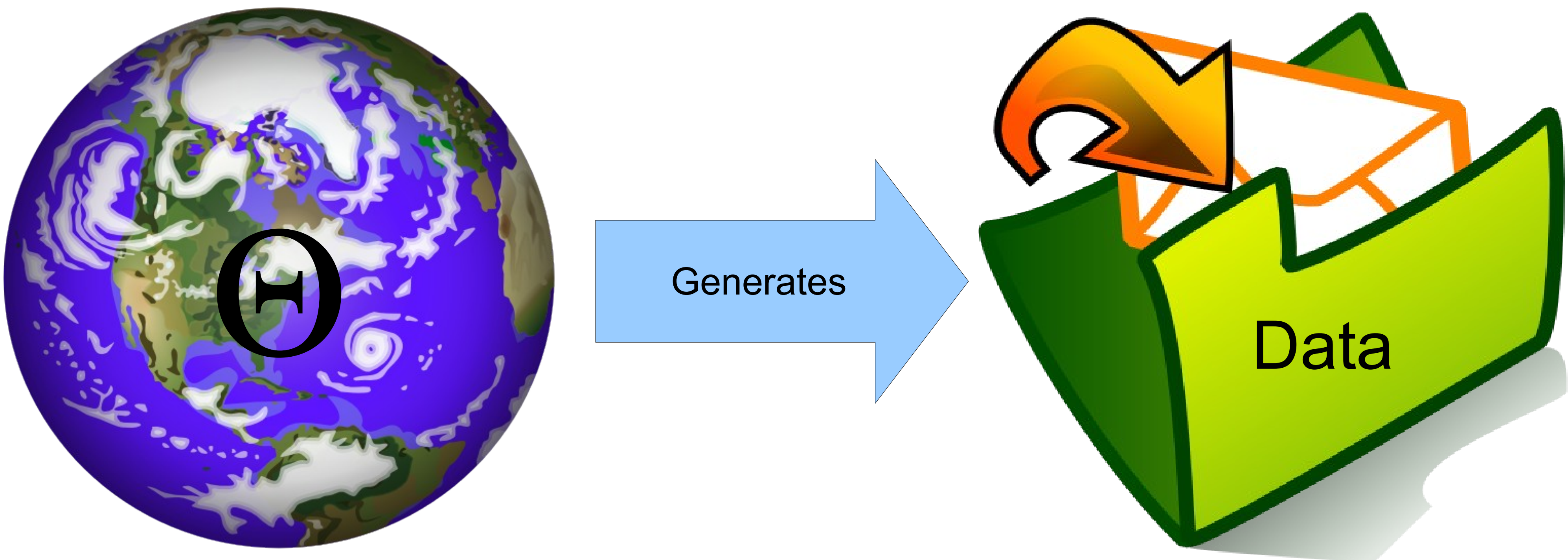
Lecture 5
Data-analysis

- Data
- Generative model
- Likelihood
- Maximum likelihood
- Bayesian learning

Generative model

- The world is described by a model that governs the probabilities of observing different kinds of data.



Data

- We will mostly handle tabular format discrete variables.

D	V0	V1	V2	V3
d_0	0	1	0	2
d_1	1	1	1	2
d_2	1	1	1	0
d_3	0	0	0	1

Likelihood $P(d|\Theta)$

- Data item d is generated by a mechanism (model) parameters Θ of which determine how probably different values of d are generated, i.e., the distribution of d.
- An example:
  - Mechanism is drawing with replacement from a bucket of black and white balls, and the parameter θ_b is the number of black balls, and the θ_w is the number of white balls in a bucket:
    - $P(b|\theta_b,\theta_w) = \theta_b/(\theta_b+\theta_w)$ and $P(w|\theta_b,\theta_w) = \theta_w/(\theta_b+\theta_w)$.
- In orthodox statistics, likelihood $P(D|\theta)$ is often seen as a function of θ, a kind of $L_D(\theta)$. Whatever.

i.i.d.

- If the data generating mechanism depends on Θ only (and not on what has been generated before), the sequence of data data is called independent and identically distributed.
- Then $P(d_{1,} d_{2,} \ldots, d_n | \theta) = \prod_{i=1}^{n} P(d_i | \theta)$
- And
  - order of d_i does not matter.
  - $P(b, w, b, b, w | \theta_b, \theta_w) = P(b, b, w, w, w | \theta_b, \theta_w)$
    $= \frac{{\theta_b}^2 {\theta_w}^3}{(\theta_b + \theta_w)^5}$

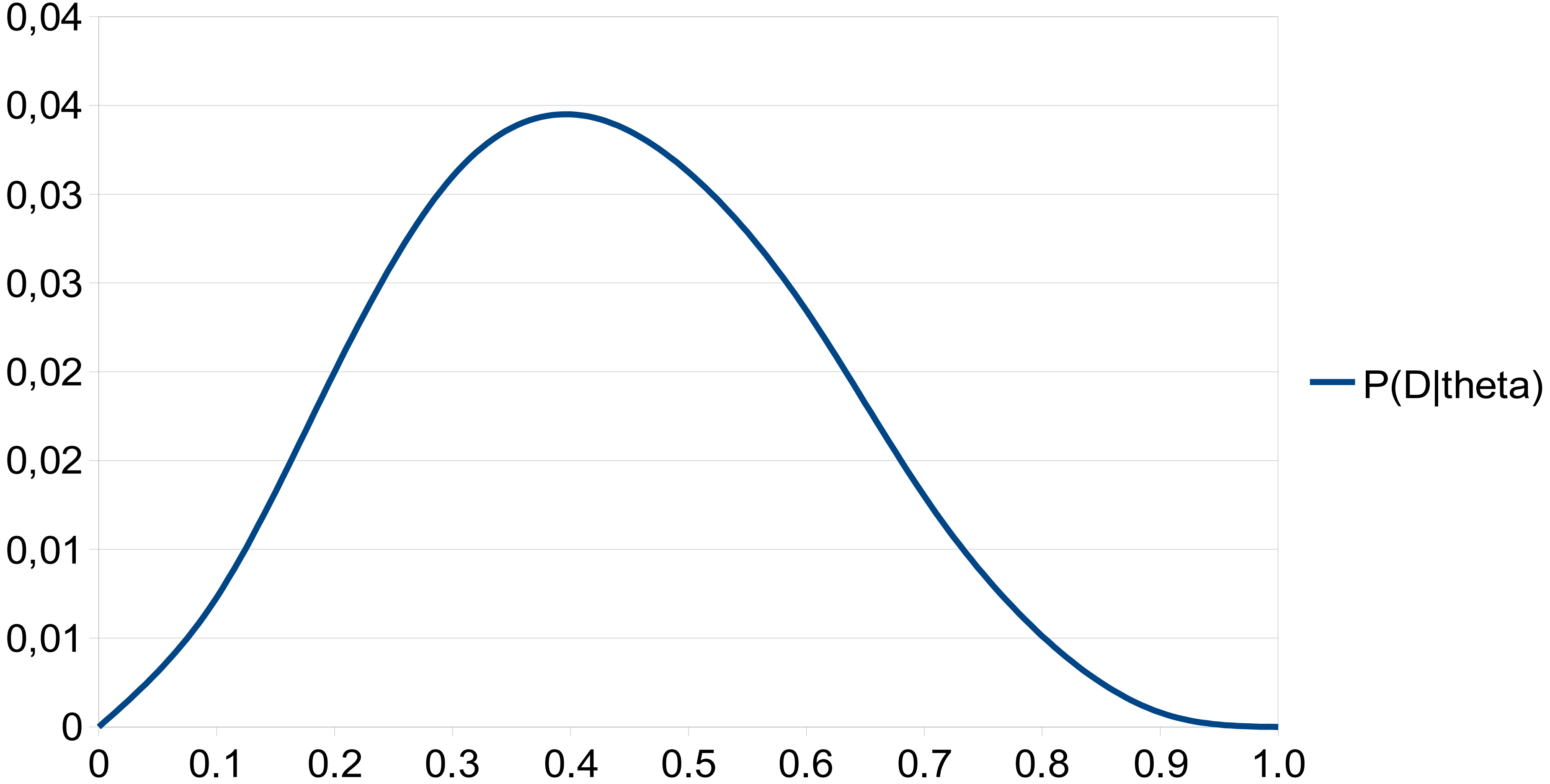
Bernoulli model

- A model for i.i.d. binary outcomes (heads,tails) (0,1), (black, white), (true, false).
- One parameter: $\theta \in [0,1]$.
  - For example:
    - $P(d=true \mid \theta) = \theta$, $P(d=false \mid \theta) = 1-\theta$.
    - NB! The probabilities of d being true are **defined by** the parameter θ. Parameters are not probabilities.
    - Black and white bucket as a Bernoulli model:
      - θ is the proportion of black balls in a bucket $P(b \mid \theta) = \theta$.
      - $P(D|\theta) = \theta^{Nb} (1-\theta)^{Nw}$, where N_b and N_w are numbers of black and white balls in the data D.
      - NB! $P(D|\theta)$ depends on data D through N_b and N_w only.

Maximum likelihood

- Given a data D, different values of θ yield different probabilities $P(D|\theta)$. The parameters that yield the largest probability of $P(D|\theta)$ are called maximum likelihood parameters for the data D.
  - $P(b,b,w,w,w|\Theta=0.7) = 0.7^2 0.3^3 = 0.1323$
  - $P(b,b,w,w,w|\Theta=0.1) = 0.1^2 0.9^3 = 0.00729$
  - $\mathrm{argmax}_\theta\, P(b,b,w,w,w|\Theta=\theta) = \mathrm{argmax}_\theta\, \theta^2(1-\theta)^3 = ?$

Likelihood P(b,b,w,w,w|Θ)



- NB! Not a distribution, but a function of Θ.

ML-parameters for the Bernoulli model.

(High school math refresher)

- So let us find ML-parameters for the Bernoulli model for the data with N_b black balls and N_w white ones.

$$P(D|\theta) = \theta^{N_b}(1-\theta)^{N_w},$$

so let us check when $P'(D|\theta)=0, \theta \in]0,1[$.

$$P'(D|\theta) = N_b \theta^{N_b - 1}(1-\theta)^{N_w} + \theta^{N_b} N_w (1-\theta)^{N_w - 1} \cdot -1$$

$$= \theta^{N_b-1}(1-\theta)^{N_w-1}[N_b(1-\theta) - \theta N_w]$$

$$= \theta^{N_b-1}(1-\theta)^{N_w-1}[N_b - (N_b+N_w)\theta] = 0$$

$$\Leftrightarrow N_b - (N_b+N_w)\theta = 0 \Leftrightarrow \theta = \frac{N_b}{N_b+N_w}$$

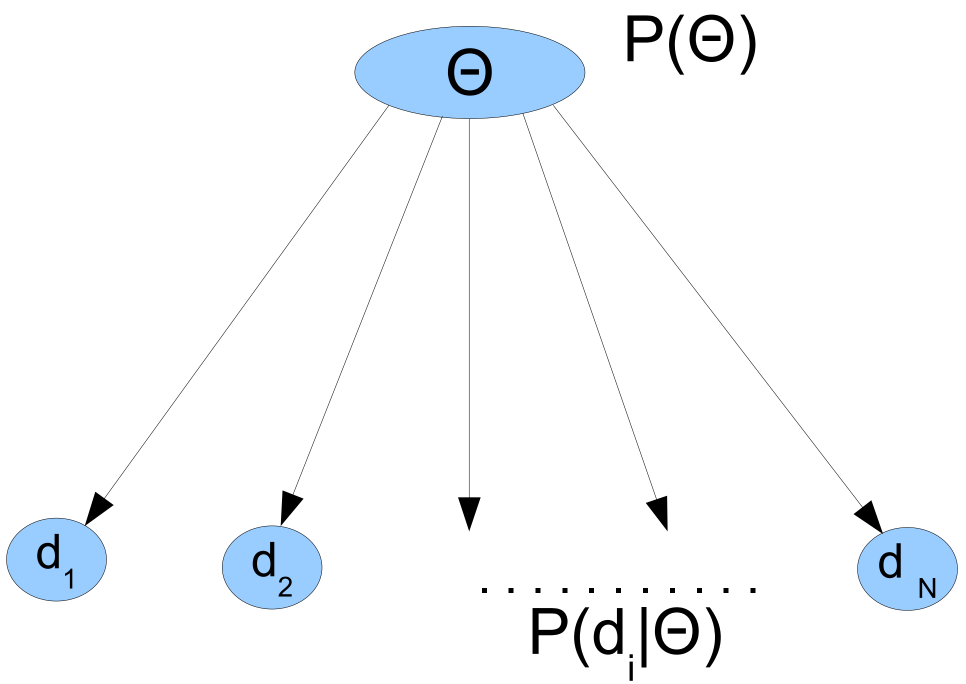
But ML-parameters are too gullible

- Assume D=(b,b), i.e., two black balls.
  - ML-parameter is $\Theta=1$.
  - Now P(next ball is white | $\Theta=1$)= 0.
  - Selecting ML parameters do not appear to be a rational choice.
- Be Bayesian:
  - Parameters are exactly the things you do not know for sure, so they have a (prior and posterior) distribution.
  - **Posterior distribution of the model is the goal of the Bayesian data-analysis.**

Good old Bayes rule

- Nothing special since Θ is just a random variable.
- And if i.i.d, we get a kind of Naïve Bayes structure.
- NB. Not a typical Bayesian network since parameter(s) also drawn as node(s).

$$P(\theta|D) = \frac{P(\theta)\, P(D|\theta)}{P(D)}$$



Predicting with posterior distribution

- Not a two phase process like in ML-case
  - first find parameters Θ.
  - then use them to calculate P(d|Θ).
- Instead: $$P(d|D) = \sum_{\theta \in \Theta} P(\theta, d|D)$$
$$= \sum_{\theta \in \Theta} P(d|\theta, D)\, P(\theta|D)$$
$$= \sum_{\theta \in \Theta} P(d|\theta)\, P(\theta|D)$$
- Bayesian prediction uses predictions P(d|θ) from all the models θ, and weighs them by the posterior probability P(θ|D) of the models.

Posterior for Bernoulli parameter

- So likelihood P(D|θ) we can calculate.
- How about the prior P(θ)?
  - We should give a real number for each θ.
    - One way out: use discrete set of parameters instead of continuous θ. Works, is flexible, but does not scale up well.
    - Another way: Study calculus.
- And how about $P(D) = \int_0^1 P(\theta)\, P(D|\theta)\, d\theta$
  - P(D) contains P(θ), so let us care about the prior first.

Prior for Bernoulli model

- The form of the likelihood gives us a hint for a comfortable prior
  - $P(D|\theta) = \theta^{Nb} (1-\theta)^{Nw}$
  - If we define the $P(\theta) = C\, \theta^{\alpha-1} (1-\theta)^{\beta-1}$,
    - C taking care that $\int P(\theta) d\theta = 1$, then
  - $P(\theta)P(D|\theta) = C\, \theta^{Nb+\alpha-1} (1-\theta)^{Nw+\beta-1}$
- Thus updating from prior to posterior is easy. Just use the formula for the prior, and update exponents α-1 and β-1.

P(Θ) of a form C $\theta^{\alpha-1}(1-\theta)^{\beta-1}$ is called Beta(α,β) distribution

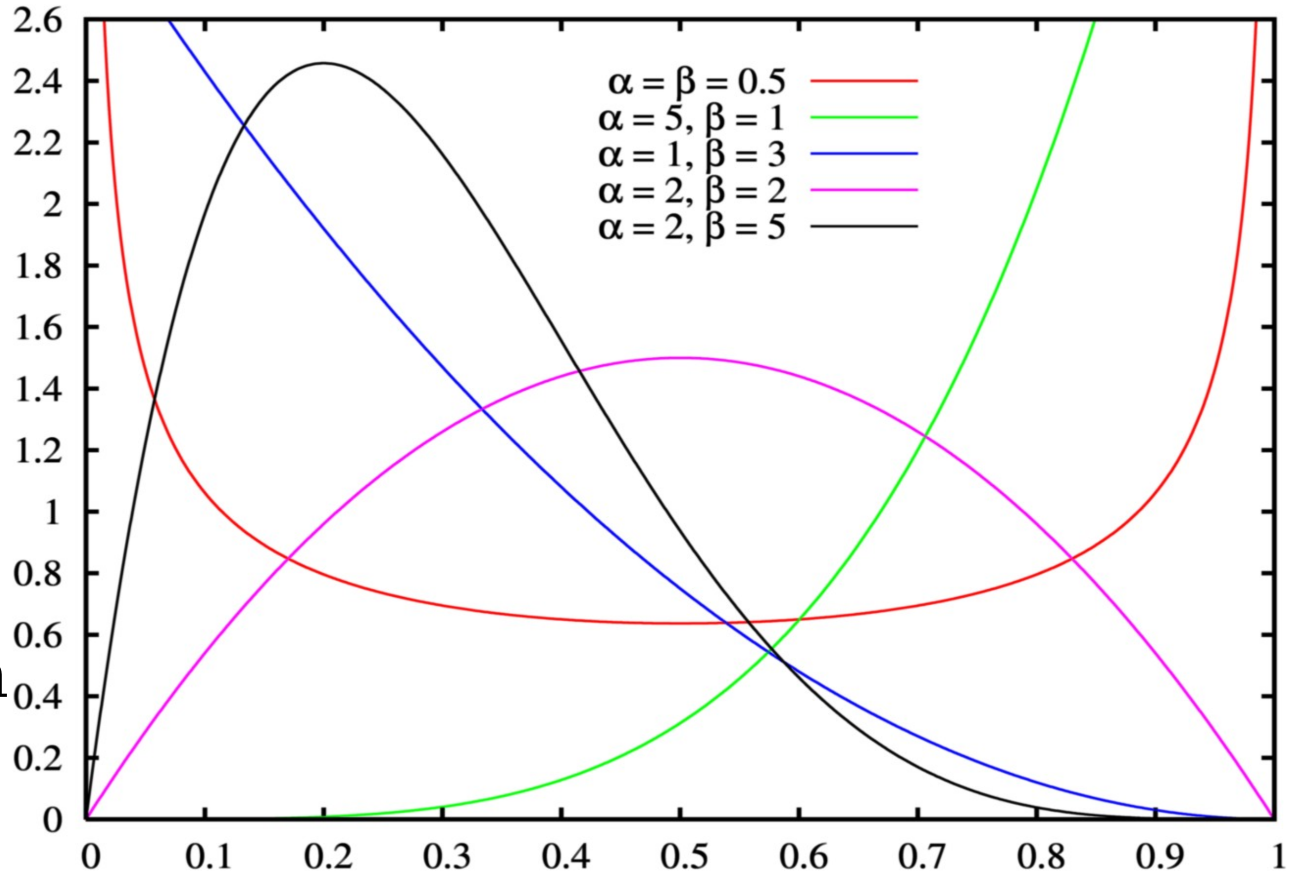
- The expected value of Θ is $\alpha/(\alpha+\beta)$.
- The normalizing constant

$$C=\frac{1}{\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\,d\theta}$$

$$=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)},$$

where Γ is the gamma-function, a continuous version of the factorial:

$$\Gamma(n)=(n-1)!$$



Posterior of the Bernoulli model

$$P(\theta|D,\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha+N_b-1}(1-\theta)^{\beta+N_w-1}$$

- Thus, a posteriori, Θ is distributed by Beta($\alpha+N_b,\beta+N_w$).
- And prediction:

$$P(b|D,\alpha,\beta)=\int_0^1 P(b|\theta,D,\alpha,\beta)P(\theta|D,\alpha,\beta)\,d\theta$$

$$=\int_0^1 P(b|\theta)P(\theta|D,\alpha,\beta)\,d\theta=\int_0^1 \theta\, P(\theta|D,\alpha,\beta)\,d\theta$$

$$=E_P(\theta)=\frac{\alpha+N_b}{\alpha+N_b+\beta+N_w}.$$

Bernoulli prediction example

$$P(b|D,\alpha,\beta)=\frac{\alpha+N_b}{\alpha+N_b+\beta+N_w}.$$

- So P(b|w,w,α=1,β=1) = (1+0) / (1+0+1+2) = 1/4.
  - sounds more rational.
  - Notice how α and β act like extra counts.
    - That's why α + β is often called “equivalent sample size”. The prior acts like seeing α black balls and β white balls before seeing data.

One variable, more than two values

- Variable X with possible values 1,2,...,n.
- Parameter vector $\theta=(\theta_1, \theta_2, ..., \theta_n)$ with $\Sigma\theta_i=1$.
- $P(X=i|\theta)=\theta_i$.
- Prior $P(\theta)=Dir(\theta; \alpha_1, \alpha_2, ..., \alpha_n) = \dfrac{\Gamma(\sum_{i=1}^{n}\alpha_i)}{\prod_{i=1}^{n}\Gamma(\alpha_i)}\prod_{i=1}^{n}\theta_i^{\alpha_i-1}$
- Posterior $P(\theta)=Dir(\theta; \alpha_1+N_1, \alpha_2+N_2, ..., \alpha_n+N_n)$
- Prediction $P(x_i \mid D, \alpha) = \dfrac{\alpha_i+N_i}{\sum_{j=1}^{n}\alpha_j+N_j}$.