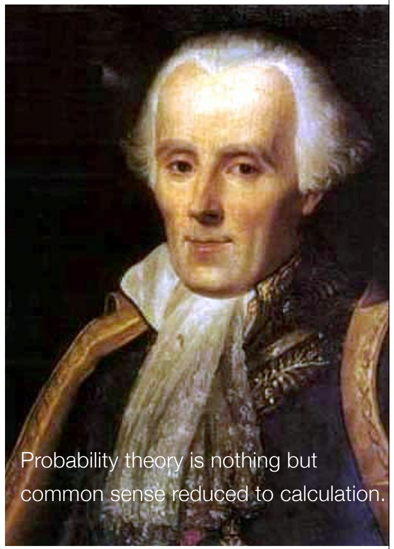
Royal Institute of Technology

DD2447 STAT. METH. IN CS FALL 2014

- ★ Lecture 2 - Probability, Bayesian
- ★ Chapter 2, 3

CHAPTER 2

- ★ Known concepts
- ★ Distributions
  - Beta, Gamma, Dirichlet
- ★ Sampling
- ★ Information theory



BERNOULLI & BINOMIAL

$$\mathrm{Ber}(x|\theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1 - \theta & \text{if } x = 0 \end{cases} \qquad \mathrm{Bin}(k|n,\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

- ★ One or several (unordered) coin tosses

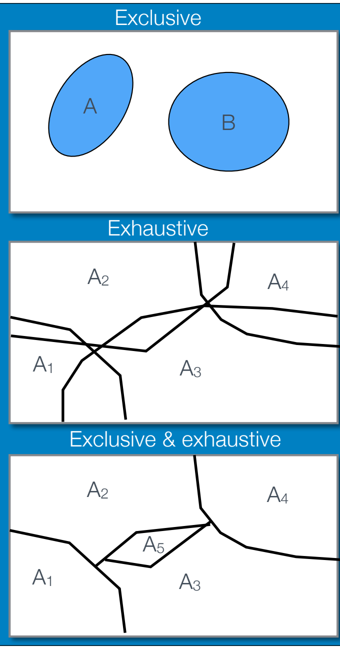
CATEGORICAL & MULTINOMIAL

$$\mathrm{Cat}(x|\boldsymbol{\theta}) = \theta_x \qquad \mathrm{Mul}(\boldsymbol{x}|n,\boldsymbol{\theta}) = \binom{n}{x_1, \ldots, x_K} \prod_{k=1}^{K} \theta_k^{x_k}$$

- ★ One or several (unordered) coin tosses

CONDITIONING

$$p(x,y) = p(y)p(x|y) \quad \text{or} \quad p(x|y) = \frac{p(x,y)}{p(y)}$$



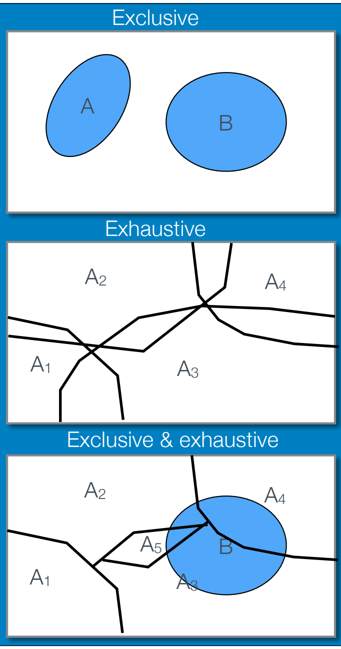
EXCLUSIVE & EXHAUSTIVE

- Exclusive

$$p(A \text{ or } B) = p(A) + P(B)$$

- Exclusive & exhaustive

$$\sum_i p(A_i) = 1$$



EXCLUSIVE & EXHAUSTIVE

- Exclusive

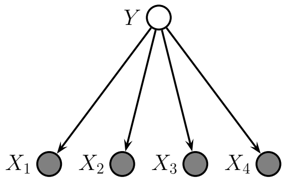
$$p(A \text{ or } B) = p(A) + P(B)$$

- Exclusive & exhaustive

$$p(B) = \sum_i p(B, A_i) = \sum_i p(A_i)p(B|A_i)$$

CHAPTER 3 - BEYOND BAYES THEOREM, CONCEPTS

$$p(X|Y) = \frac{p(X,Y)}{p(Y)} = \frac{p(X)p(Y|X)}{\sum_x p(x)p(Y|x)}$$



NAIVE BAYES CLASSIFIER

$$p(\boldsymbol{x}, y) = p(y) \prod_{t=1}^{4} p(x_t|y)$$

BAYES — GENERATIVE CLASSIFIER

- **X** given Y the natural direction
- Classifier

$$p(Y = c|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{p(Y = c|\boldsymbol{\theta})p(\boldsymbol{x}|Y = c, \boldsymbol{\theta})}{\sum_{\boldsymbol{x}} p(Y = c|\boldsymbol{\theta})p(\boldsymbol{x}|Y = c, \boldsymbol{\theta})}$$

BREAST CANCER TEST

- X test – 1 positive
- Y breast cancer – 1 cancer

THE BRAIN

★ Biology

the more data the better

★ Philosophy of the mind

John Searl

WHICH **FAIR** DICE?

- $x_1,\ldots,x_N \in [100]$
  - $[M] = \{1,\ldots,M\}$
- Ex.
  - 1,4,5,2,2,5,3,5,3,6,…
  - 14,8,28,2,36,…
  - 1,14,16,20,19,…

DATA

- Data D=14,8,28,2,36,…
- Dice H = 20-sided with even numbers in [40]
  - Likelihood $p(D|H) = (1/20)^5$
- Dice H' = 6-sided with numbers 14,8,28,2,36,7
  - Likelihood $p(D|H') = (1/6)^5$
- Least number of sides wins - Occam's razor

LIKELIHOOD

- Data D=14,8,28,2,36,
- 6-sided dice H' with numbers 14,8,28,2,36,7
- pretty unnatural
  - we give it prior probability $p(H') = 1/10^6$
- Posterior probability (after observation)
- In general $$p(H'|D) = \frac{p(D|H')p(H')}{p(D)} = \frac{p(D|H')p(H')}{\sum_{H' \in \mathcal{H}'} p(D|H')}$$
- Here $$p(H'|D) = \frac{(1/6)^5 10^{-6}}{p(D)}$$

PRIORS

★ Data D=14,8,28,2,36,

★ 20-sided Dice H with even numbers in [40]

- fairly natural
- we give it prior probability p(H) = 1/1000

★ Posterior probability

$$p(H|D) = \frac{(1/20)^5 10^{-3}}{p(D)}$$

★ So

$$\frac{p(H|D)}{p(H'|D)} = \frac{(1/20)^5 10^{-3}}{(1/6)^5 10^{-6}} \approx 10^3/411 > 1$$

PRIORS

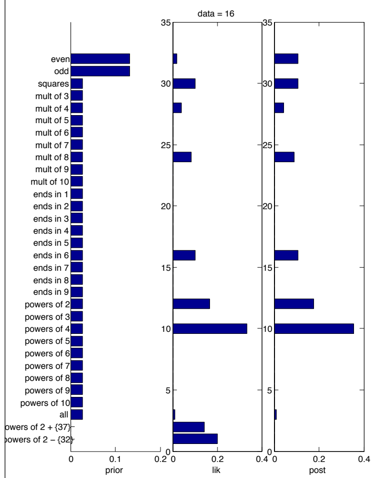
INFLUENCE OF PRIOR IN THIS MODEL

★ MAP = argmax_H p(H|D) = argmax log p(H|D) = argmax log p(D|H) + log p(H)

- ★ here p(D|H)=$(1/C)^N$ decreases exponentially in N=#throws or data points
- ★ p(H) is constant

★ Conclusions: as N → ∞ p(D|H) will dominate

★ Data overwhelms the prior



MAP ESTIMATE

★ Maximum a posteriori

- ✳ (compare a priori)
- ✳ MAP = argmax_H p(H|D)

★ Convergence

- ✳ Consistent (as is MLE)
- ✳ Uncommon in practice

POSTERIOR IS IMPORTANT (FOR THIS CHOICE)

One dataset

Another dataset

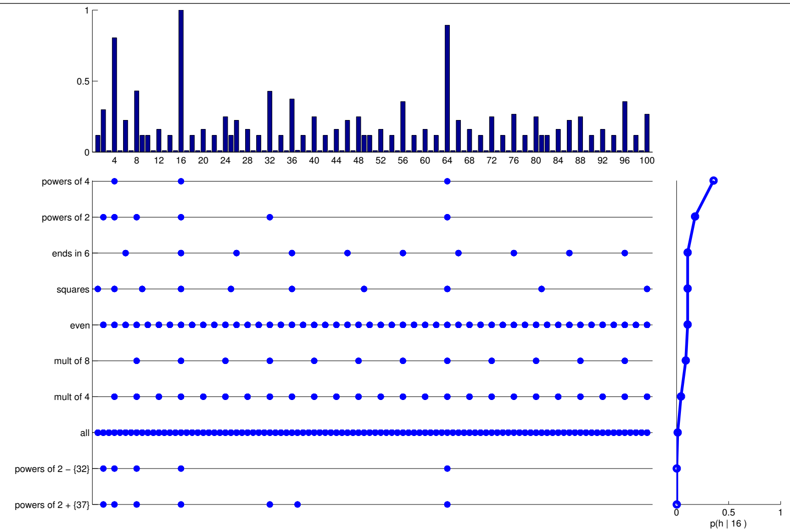
BAYESIAN MODEL AVERAGING

- Posterior - what we believe
- Using the MAP for prediction

$$p(x|D) = p(x|H^D_{\text{MAP}})$$

- Alternative model averaging

$$p(x|D) = \sum_{H \in \mathcal{H}} p(x, H|D) = \sum_{H \in \mathcal{H}} p(x|H)p(H|D)$$



POSTERIOR - OBSERVED 16

BINOMIAL - IID BERNOULLI

- Likelihood
  - N_1 heads and N_0 tails
  - p(head) = θ
  - sequence $p(D) = \theta^{N_1}(1-\theta)^{N_0}$
  - counts $p(D) = \binom{N_1 + N_0}{N_1}\theta^{N_1}(1-\theta)^{N_0}$

MLE BY FREQUENCIES

- Likelihood is up to a constant $p(D) = \theta^{N_1}(1-\theta)^{N_0}$
- Log-likelihood $l(D) = N_1 \log\theta + N_0 \log(1-\theta)$
- Optimized by same θ
- Derivation $l'(D) = \frac{N_1}{\theta} - \frac{N_0}{(1-\theta)}$
- Setting to zero $\frac{N_1}{\theta} = \frac{N_0}{(1-\theta)}$
- So

$N_1 - \theta N_1 = \theta N_0$ and $\theta = \frac{N_1}{N_1 + N_0}$

MLE FOR MULTINOMIAL AND AND CATEGORICAL

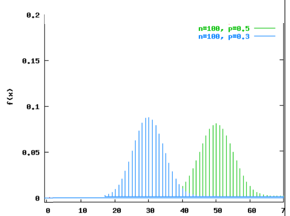
- Likelihood $p(D) = \prod_{i \in [k]} \theta_i^{N_i}$
- where $\sum_{i \in [k]} \theta_i = 1$
- as well as loglikelihood $p(D) = \sum_{i \in [k]} N_i log \theta_i$
- is maximized by $\theta_i = \frac{N_i}{\sum_{i \in [k]} N_i}$

SUFFICIENT STATISTICS

- Maximum Likelihood (ML) estimate maximize the probability of the data
  - here $\max_\theta p(N_1, N_0 | \theta)$
- The pair N_1, N_0 is a sufficient statistic for our coin model
  - i.e., given those ML estimate follows

PRIOR FOR CATEGORICAL AND BERNOULLI - FIRST BERNOULLI

- Assumption: we don't know θ
- We need a prior over the outcome probabilities
  - N_1 heads and N_0 tails
  - p(head) denoted θ
  - sequence $p(D) = \theta^{N_1}(1-\theta)^{N_0}$ iid Bernoulli
  - counts $p(D) = \binom{N_1 + N_0}{N_1} \theta^{N_1}(1-\theta)^{N_0}$ Binomial

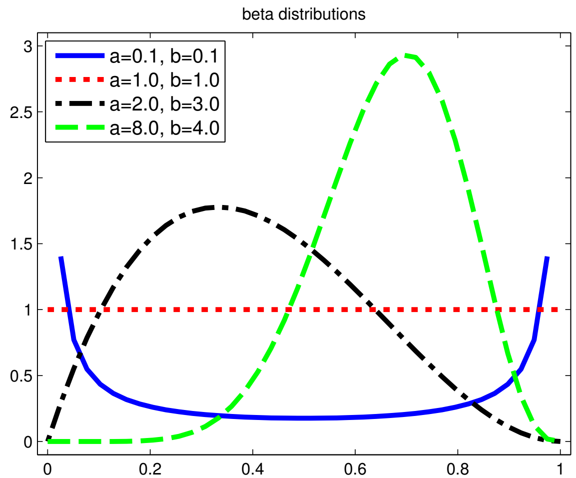


BETA DISTRIBUTION

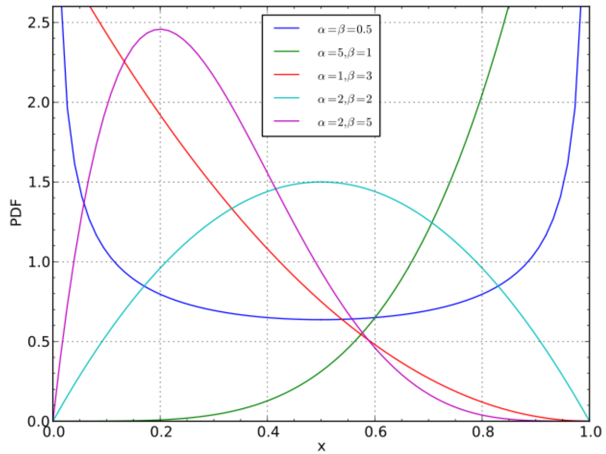
- PDF $\text{Beta}(x|a,b) = \frac{1}{B(a,b)} x^{(a-1)} x^{(b-1)}$
  - where $B(a,b) = \frac{\Gamma(a)\Gamma(a)}{\Gamma(a+b)}$
  - and $\Gamma(a) = (a-1)\Gamma(a-1)$
- In particular, for integer n

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

- Proper – the integral is 1.



BETA DISTRIBUTION

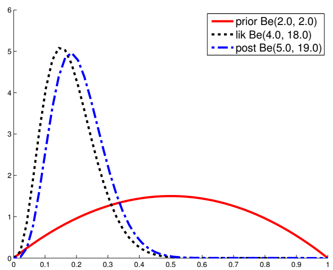


BETA DISTRIBUTION

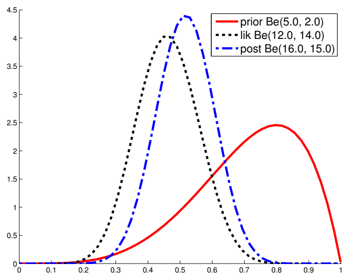
HYPER-PARAMETERS

- Parameters γ_1 and γ_0 for prior called
  - hyperparameters
  - pseudocounts
- Prior's effective sample size is $\gamma_1 + \gamma_0$
- Prior that gives posterior of the same sort is called conjunctive

BETA-BINOMIAL: PRIOR, LIKELIHOOD,POSTERIOR



$N_1=3$, $N_0= 17$



$N_1=11$, $N_0= 13$

BETA BINOMIAL

- Beta distribution up to a constant $p(\theta|\gamma_1,\gamma_0) \propto \theta^{\gamma_1-1}(1-\theta)^{\gamma_0-1}$

- Posterior
$$\begin{aligned} p(\theta|D) &= p(D|\theta)p(\theta|\gamma_1,\gamma_0) \\ &\propto \theta^{N_1}(1-\theta)^{N_0}\theta^{\gamma_1-1}(1-\theta)^{\gamma_0-1} \\ &= \theta^{N_1+\gamma_1-1}(1-\theta)^{N_0+\gamma_0-1} \\ &\propto \mathrm{Beta}(\theta|N_1+\gamma_1, N_0+\gamma_0) \end{aligned}$$

- Beta is a conjunctive prior for Binomial

$D = D_a \cup D_b$

$S(D_a) = (N_1^a, N_0^a)$ and $S(D_b) = (N_1^b, N_0^b)$

so $S(D) = (N_1^a + N_1^b, N_0^a + N_0^b) = (N_1, N_0)$

again $p(\theta|D) = \mathrm{Beta}(\theta|N_1+\gamma_1, N_0+\gamma_0)$

2 steps $p(\theta|D_a, D_b) \propto p(D_b|\theta)p(\theta|D_a) = ?$

BETA-BINOMIAL: BATCH OR IN TWO STEPS

$D = D_a \cup D_b$

$S(D_a) = (N_1^a, N_0^a)$ and $S(D_b) = (N_1^b, N_0^b)$

so $S(D) = (N_1^a + N_1^b, N_0^a + N_0^b) = (N_1, N_0)$

again $p(\theta|D) = \mathrm{Beta}(\theta|N_1+\gamma_1, N_0+\gamma_0)$

2 steps
$$\begin{aligned} p(\theta|D_a, D_b) &\propto p(D_b|\theta)p(\theta|D_a) \\ &= \mathrm{Bin}(N_1^b, N_0^b)\mathrm{Beta}(\theta|N_1^a+\gamma_1, N_0^a+\gamma_0) \\ &= \mathrm{Beta}(\theta|N_1^b+N_1^a+\gamma_1, N_0^b+N_0^a+\gamma_0) \\ &= \mathrm{Beta}(\theta|N_1+\gamma_1, N_0+\gamma_0) \end{aligned}$$

BETA-BINOMIAL: BATCH OR IN TWO STEPS

BETA-BINOMIAL - MLE, MAP, AND PM

Let $\gamma = \gamma_1 + \gamma_0$ and $N = N_1 + N_0$

Then

PM - Posterior mean

$$\theta_{\mathrm{MLE}} = \frac{N_1}{N}, \quad \theta_{\mathrm{MAP}} = \frac{N_1+\gamma_1-1}{N+\gamma-2}, \text{ and } \theta_{\mathrm{PM}} = \frac{N_1+\gamma_1}{N+\gamma}$$

Let $f_1 = \gamma_1/\gamma$

Then
$$E[\theta|D] = \frac{\gamma f_1 + N_1}{N+\gamma} = \frac{\gamma}{N+\gamma}f_1 + \frac{N}{N+\gamma}\theta_{\mathrm{MLE}}$$

Posterior to our observation what is the probability of a specific outcome?

$$p(x=1|D) = \int_0^1 p(x=1|\theta)p(\theta|D)\mathrm{d}\theta$$
$$= \int_0^1 \theta \mathrm{Beta}(\theta|N_1+\gamma_1, N_0+\gamma_0)\mathrm{d}\theta$$
$$= \frac{N_1+\gamma_1}{N+\gamma}$$

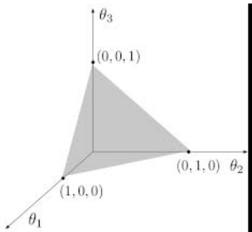
POSTERIOR PREDICTIVE

Uniform prior, i.e., $\gamma_1=\gamma_0=1$, gives

$$p(x=1|D) = \frac{N_1+1}{N+2}$$

Black Swan “paradox”

LAPLACE’S RULE OF SUCCESSION



DIRICHLET

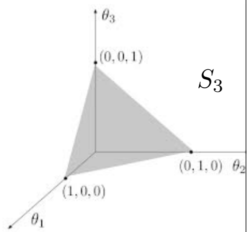
$D = \{x_1, \ldots, x_N\}$ where $x_i \in \{1, \ldots, K\}$

$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \in S_K$ the K-dim. probability simplex, i.e., pos. & $\sum_{i \in [K]} \theta_k = 1$

$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ hyperparameters

Prior $Dir(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k \in [K]} \theta_k^{\alpha_k - 1}$

Likelihood $p(\boldsymbol{\theta}|D) \propto \prod_{k \in [K]} \theta_k^{N_k}$



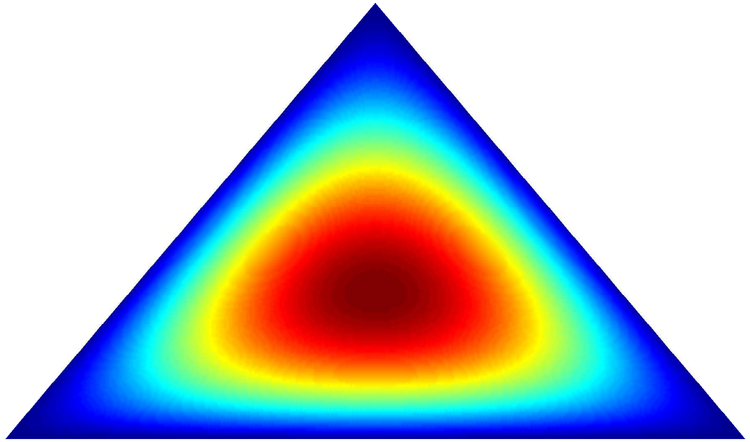
BACK TO THE DICE – DIRICHLET-MULTINOMIAL

Prior $Dir(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k \in [K]} \theta_k^{\alpha_k - 1}$

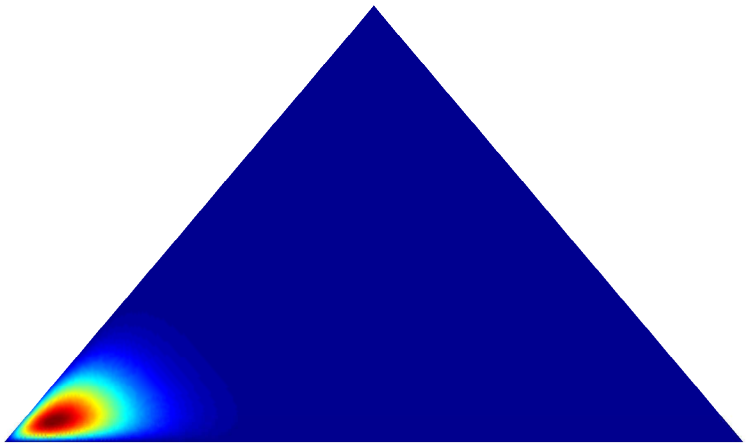
Likelihood $p(\boldsymbol{\theta}|D) \propto \prod_{k \in [K]} \theta_k^{N_k}$

Posterior
$$\begin{aligned} p(\boldsymbol{\theta}|D) &= p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &\propto \prod_{k \in [K]} \theta_k^{N_k} \prod_{k \in [K]} \theta_k^{\alpha_k - 1} \\ &= \prod_{k \in [K]} \theta_k^{N_k + \alpha_k - 1} \\ &\propto \mathrm{Dir}(\boldsymbol{\theta}|N_1 + \alpha_1, \ldots, N_K + \alpha_K) \end{aligned}$$

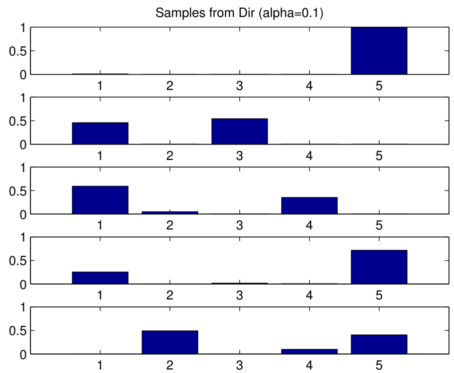
POSTERIOR FOR DIRICHLET-MULTINOMIAL



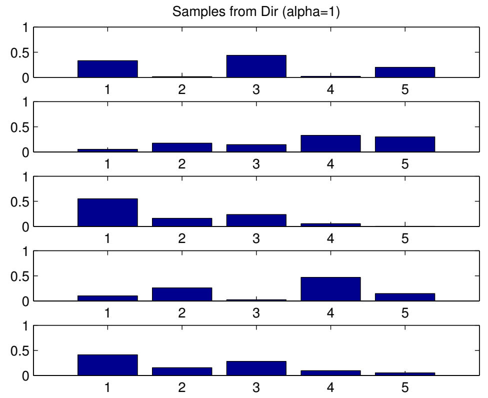
α = (2, 2, 2)



α = (20, 2, 2)



SAMPLES α = (0.1, 0.1, 0.1)



SAMPLES $\alpha = (1,1,1)$

The end