ISTA 421 + INFO 521
Introduction to Machine Learning

Lecture 12:
Priors and Marginal Likelihood

Clay Morrison
claytonm@email.arizona.edu
Harvill 437A
Phone 621-6609

2 October 2017

1

Recap: Coin Scenario

$$p(y_N|r) = \binom{N}{y} r^y (1-r)^{N-y}$$

$$P(Y_{10} \leq 6|r)$$

$$p(r|y_N)$$

$$p(r) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1}$$

Model how prior combines with evidence (from likelihood) to update beliefs

$$\mathbf{E}_{p(r|y_N)}\left\{P(Y_{10} \leq 6|r)\right\} = \int_{r=0}^{r=1} P(Y_{10} \leq 6|r)p(r|y_N)\, dr$$

2

Computing the Posterior Directly

We can do this with the conjugate
Beta prior and Binomial Likelihood

$$p(r|y_N) \propto \left[\binom{N}{y_N} r^{y_N}(1-r)^{N-y_N}\right] \times \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1}\right]$$

$$\begin{aligned} p(r|y_N) &\propto \left[\binom{N}{y_N}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right] \times \left[r^{y_N} r^{\alpha-1}(1-r)^{N-y_N}(1-r)^{\beta-1}\right] \\ &\propto r^{y_N+\alpha-1}(1-r)^{N-y_N+\beta-1} \\ &\propto r^{\delta-1}(1-r)^{\gamma-1} \end{aligned}$$

where $\delta = y_N + \alpha$ and $\gamma = N - y_N + \beta$.

$$p(r|y_N) = \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+y_N)\Gamma(\beta+N-y_N)} r^{\alpha+y_N-1}(1-r)^{\beta+N-y_N-1}$$

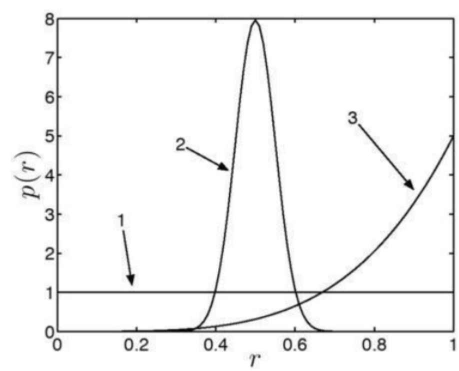
3

Affect of 3 Different Priors

$$p(r|y_N) = \frac{P(y_N|r)p(r)}{P(y_N)}$$

(2) **The Prior**: $p(r)$

“Allows us to express any belief we have in the value of *r* ***before*** we see any data.”



1) We don’t know anything about the coins or the stall owner
$\alpha = 1, \beta = 1$

2) We think the coin (and the stall owner) is fair $\alpha = 50, \beta = 50$

3) We think the coin (and the stall owner) is biased to give more heads
$\alpha = 5, \beta = 1$

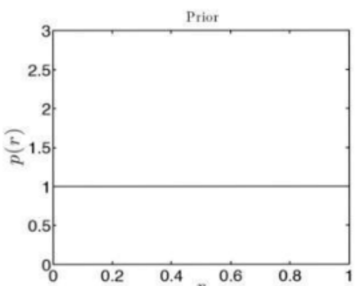
$$p(r) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1}$$

Scenario 1: Don't know anything

$$p(r) = \mathcal{B}(\alpha, \beta) \qquad \mathbf{E}_{p(r)}\{R\} = \frac{\alpha}{\alpha+\beta} \qquad \mathrm{var}\{R\} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Scenario 1 prior: $\alpha = 1,\, \beta = 1 \quad \mathbf{E}_{p(r)}\{R\} = \frac{\alpha}{\alpha+\beta} = \frac{1}{2} \quad \mathrm{var}\{R\} = \frac{1}{12}$

General posterior: $\delta = \alpha + y_N \quad \gamma = \beta + N - y_N \quad p(r|y_N) = \mathcal{B}(\delta, \gamma)$



(a) $\alpha = 1, \beta = 1$

Observations:

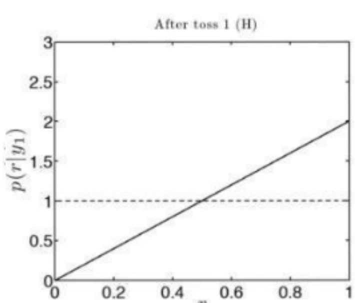
5

Scenario 1: Don't know anything

$$p(r) = \mathcal{B}(\alpha, \beta) \qquad \mathbf{E}_{p(r)}\{R\} = \frac{\alpha}{\alpha+\beta} \qquad \mathrm{var}\{R\} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Scenario 1 prior: $\alpha = 1,\, \beta = 1 \quad \mathbf{E}_{p(r)}\{R\} = \frac{\alpha}{\alpha+\beta} = \frac{1}{2} \quad \mathrm{var}\{R\} = \frac{1}{12}$

General posterior: $\delta = \alpha + y_N \quad \gamma = \beta + N - y_N \quad p(r|y_N) = \mathcal{B}(\delta, \gamma)$



(b) $\delta = 2, \gamma = 1$

Observations: H

$$\delta = 1 + 1 = 2$$
$$\gamma = 1 + 1 - 1 = 1$$

Posterior: $\mathbf{E}_{p(r|y_N)}\{R\} = \frac{2}{3}$

$\mathrm{var}\{R\} = \frac{1}{18}$

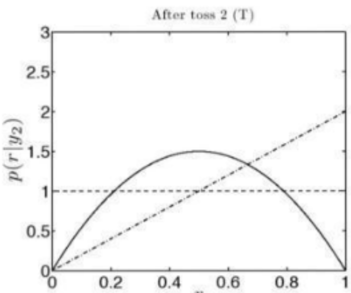
6

Scenario 1: Don't know anything

$p(r) = \mathcal{B}(\alpha, \beta)$ $\mathbf{E}_{p(r)}\{R\} = \frac{\alpha}{\alpha+\beta}$ $\text{var}\{R\} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Scenario 1 prior: $\alpha = 1, \beta = 1$ $\mathbf{E}_{p(r)}\{R\} = \frac{\alpha}{\alpha+\beta} = \frac{1}{2}$ $\text{var}\{R\} = \frac{1}{12}$

General posterior: $\delta = \alpha + y_N$ $\gamma = \beta + N - y_N$ $p(r|y_N) = \mathcal{B}(\delta, \gamma)$



(c) $\delta = 2, \gamma = 2$

Observations: H T

$$\delta = 1 + 1 = 2$$
$$\gamma = 1 + 2 - 1 = 2$$

Posterior: $\mathbf{E}_{p(r|y_N)}\{R\} = \frac{1}{2}$

$\text{var}\{R\} = \frac{1}{20}$

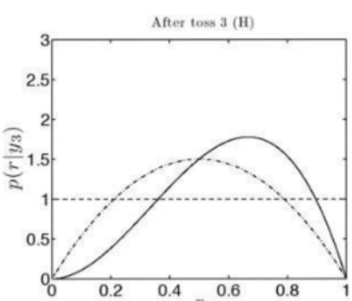
7

Scenario 1: Don't know anything

$p(r) = \mathcal{B}(\alpha, \beta)$ $\mathbf{E}_{p(r)}\{R\} = \frac{\alpha}{\alpha+\beta}$ $\text{var}\{R\} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Scenario 1 prior: $\alpha = 1, \beta = 1$ $\mathbf{E}_{p(r)}\{R\} = \frac{\alpha}{\alpha+\beta} = \frac{1}{2}$ $\text{var}\{R\} = \frac{1}{12}$

General posterior: $\delta = \alpha + y_N$ $\gamma = \beta + N - y_N$ $p(r|y_N) = \mathcal{B}(\delta, \gamma)$



(d) $\delta = 3, \gamma = 2$

Observations: H T H

$$\delta = 1 + 2 = 3$$
$$\gamma = 1 + 3 - 2 = 2$$

Posterior: $\mathbf{E}_{p(r|y_N)}\{R\} = \frac{3}{5}$

$\text{var}\{R\} = \frac{1}{25}$

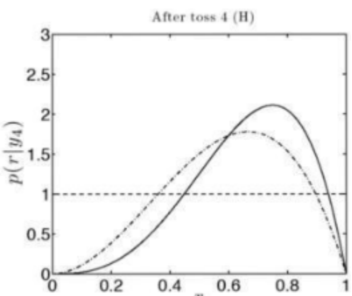
8

Scenario 1: Don’t know anything

$$p(r) = \mathcal{B}(\alpha, \beta) \qquad \mathbf{E}_{p(r)}\{R\} = \frac{\alpha}{\alpha+\beta} \qquad \text{var}\{R\} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Scenario 1 prior: $\alpha = 1, \beta = 1 \quad \mathbf{E}_{p(r)}\{R\} = \frac{\alpha}{\alpha+\beta} = \frac{1}{2} \quad \text{var}\{R\} = \frac{1}{12}$

General posterior: $\delta = \alpha + y_N \quad \gamma = \beta + N - y_N \quad p(r|y_N) = \mathcal{B}(\delta, \gamma)$



(e) $\delta = 4, \gamma = 2$

Observations: H T H H

$$\delta = 1 + 3 = 4$$
$$\gamma = 1 + 4 - 3 = 2$$

Posterior: $\mathbf{E}_{p(r|y_N)}\{R\} = \frac{2}{3}$

$$\text{var}\{R\} = \frac{2}{63} = 0.0317$$

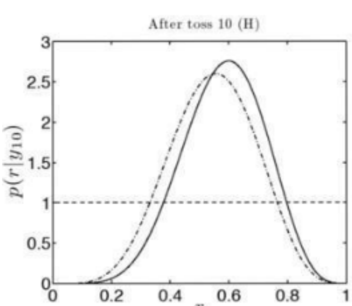
9

Scenario 1: Don’t know anything

$$p(r) = \mathcal{B}(\alpha, \beta) \qquad \mathbf{E}_{p(r)}\{R\} = \frac{\alpha}{\alpha+\beta} \qquad \text{var}\{R\} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Scenario 1 prior: $\alpha = 1, \beta = 1 \quad \mathbf{E}_{p(r)}\{R\} = \frac{\alpha}{\alpha+\beta} = \frac{1}{2} \quad \text{var}\{R\} = \frac{1}{12}$

General posterior: $\delta = \alpha + y_N \quad \gamma = \beta + N - y_N \quad p(r|y_N) = \mathcal{B}(\delta, \gamma)$



(f) $\delta = 7, \gamma = 5$

Observations: H T H H H H T T T H

$$\delta = 1 + 6 = 7$$
$$\gamma = 1 + 10 - 6 = 5$$

Posterior: $\mathbf{E}_{p(r|y_N)}\{R\} = \frac{7}{12} = 0.5833$

$$\text{var}\{R\} = 0.0187$$

10

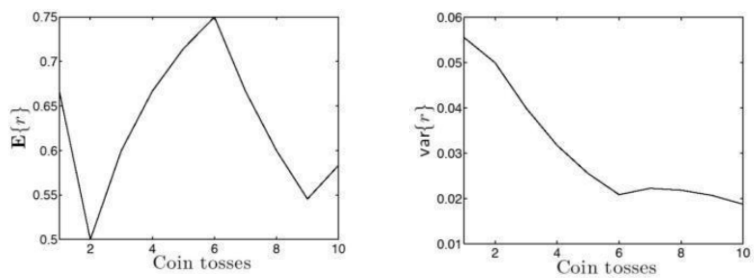
Scenario 1: Don't know anything

$$p(r) = \mathcal{B}(\alpha, \beta) \qquad \mathbf{E}_{p(r)}\{R\} = \frac{\alpha}{\alpha+\beta} \qquad \text{var}\{R\} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Scenario 1 prior: $\alpha = 1, \beta = 1 \quad \mathbf{E}_{p(r)}\{R\} = \frac{\alpha}{\alpha+\beta} = \frac{1}{2} \quad \text{var}\{R\} = \frac{1}{12}$

General posterior: $\delta = \alpha + y_N \quad \gamma = \beta + N - y_N \quad p(r|y_N) = \mathcal{B}(\delta, \gamma)$

Observations: H T H H H H T T T H



11

Uses of our Posterior Density of *r*

The posterior density encapsulates ***all*** of the information we have about *r*

$$p(r|y_N) = \mathcal{B}(\delta, \gamma)$$

We have a couple of choices about how to use this posterior belief information.

Approach 1: obtain a **point estimate** of *r* by extracting a single value $\hat{r}$ from the posterior density.

$$p(r) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1}$$

What should be our single value $\hat{r}$ chosen from the posterior distribution of *r* ?

$$\hat{r} = \mathbf{E}_{p(r|y_N)}\{R\} = \frac{\delta}{\delta+\gamma} = \frac{7}{12}$$

$$\mathbf{E}_{p(r)}\{R\} = \frac{\alpha}{\alpha+\beta}$$

$$P(Y_{\text{new}} \leq 6|\hat{r}) = 1 - \sum_{y_{\text{new}}=7}^{10} P(Y_{\text{new}} = y_{\text{new}}|\hat{r}) = 1 - 0.3414 = 0.6586$$

6 H out of 10 flips
$\boldsymbol{\delta} = 7$, $\boldsymbol{\gamma} = 5$

12

Uses of our Posterior Density of *r*

OR, we could use ***all*** of the posterior information**!**

Note the difference from what we did on the previous slide:

$P(Y_{new} \leq 6|\hat{r})$, $\hat{r} = \mathbf{E}_{p(r|y_N)}\{R\}$

$$\begin{aligned}\mathbf{E}_{p(r|y_N)}\{P(Y_{\text{new}} \leq 6|r)\} &= \mathbf{E}_{p(r|y_N)}\{1 - P(Y_{\text{new}} \geq 7|r)\} \\ &= 1 - \mathbf{E}_{p(r|y_N)}\{P(Y_{\text{new}} \geq 7|r)\} \\ &= 1 - \mathbf{E}_{p(r|y_N)}\left\{\sum_{y_{\text{new}}=7}^{y_{\text{new}}=10} P(Y_{\text{new}} = y_{\text{new}}|r)\right\} \\ &= 1 - \sum_{y_{\text{new}}=7}^{y_{\text{new}}=10} \mathbf{E}_{p(r|y_N)}\{P(Y_{\text{new}} = y_{\text{new}}|r)\}.\end{aligned}$$

$$\begin{aligned}\mathbf{E}_{p(r|y_N)}\{P(Y_{\text{new}} = y_{\text{new}}|r)\} &= \int_{r=0}^{r=1} P(Y_{\text{new}} = y_{\text{new}}|r)p(r|y_N)\,dr \\ &= \int_{r=0}^{r=1}\left[\binom{N_{\text{new}}}{y_{\text{new}}} r^{y_{\text{new}}}(1-r)^{N_{\text{new}}-y_{\text{new}}}\right]\left[\frac{\Gamma(\delta+\gamma)}{\Gamma(\delta)\Gamma(\gamma)} r^{\delta-1}(1-r)^{\gamma-1}\right] dr \\ &= \binom{N_{\text{new}}}{y_{\text{new}}} \frac{\Gamma(\delta+\gamma)}{\Gamma(\delta)\Gamma(\gamma)} \int_{r=0}^{r=1} r^{y_{\text{new}}+\delta-1}(1-r)^{N_{\text{new}}-y_{\text{new}}+\gamma-1}\,dr.\end{aligned}$$

$$\int_{r=0}^{r=1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1} dr = 1 \qquad \int_{r=0}^{r=1} r^{\alpha-1}(1-r)^{\beta-1} dr = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$\mathbf{E}_{p(r|y_N)}\{P(Y_{\text{new}} = y_{\text{new}}|r)\} = \binom{N_{\text{new}}}{y_{\text{new}}} \frac{\Gamma(\delta+\gamma)}{\Gamma(\delta)\Gamma(\gamma)} \frac{\Gamma(\delta+y_{\text{new}})\Gamma(\gamma+N_{\text{new}}-y_{\text{new}})}{\Gamma(\delta+\gamma+N_{\text{new}})}$$

13

Uses of our Posterior Density of *r*

OR, we could use ***all*** of the posterior information**!**

After 10 tosses, we have 6 heads and 4 tails; so N=10, δ=7, γ=5. Plug in!

$$\mathbf{E}_{p(r|y_N)}\{P(Y_{\text{new}} \leq 6|r)\} = 1 - \sum_{y_{\text{new}}=7}^{y_{\text{new}}=10} \mathbf{E}_{p(r|y_N)}\{P(Y_{\text{new}} = y_{\text{new}}|r)\}$$

1 – 0.3916

0.6084

Comparing this with the point estimate (0.6586), we see that both predict we will win more often than not.

This agrees with the evidence: the one person we have fully observed got 6 heads, 4 tails

The point estimate gives a higher probability; **ignoring the posterior uncertainty makes it more likely that we will win.**

$$\mathbf{E}_{p(r|y_N)}\{P(Y_{\text{new}} = y_{\text{new}}|r)\} = \binom{N_{\text{new}}}{y_{\text{new}}} \frac{\Gamma(\delta+\gamma)}{\Gamma(\delta)\Gamma(\gamma)} \frac{\Gamma(\delta+y_{\text{new}})\Gamma(\gamma+N_{\text{new}}-y_{\text{new}})}{\Gamma(\delta+\gamma+N_{\text{new}})}$$

14

Uses of our Posterior Density of *r*

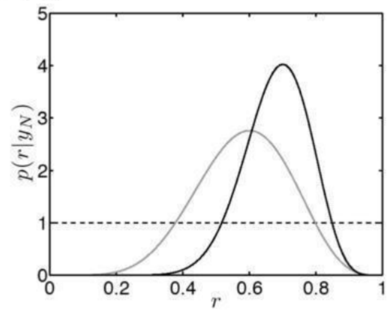
Observations: H T H H H H T T T H H H T T H H H H H H

After 10 MORE tosses, we have a total of 14 heads and 6 tails: N=20, δ=15, γ=7. Plug in!

$\mathbf{E}_{p(r|y_N)}\{R\} = 0.6818, \text{var}\{R\} = 0.0094$

$\mathbf{E}_{p(r|y_N)}\{P(Y_{\text{new}} \leq 6|r)\} = 0.4045$ 0.4047

($P(Y_{\text{new}} \leq 6|\hat{r}) = 0.3994$ The not-fully-Bayes pt. estimate)



$$\mathbf{E}_{p(r|y_N)}\{P(Y_{\text{new}} = y_{\text{new}}|r)\} = \binom{N_{\text{new}}}{y_{\text{new}}} \frac{\Gamma(\delta+\gamma)}{\Gamma(\delta)\Gamma(\gamma)} \frac{\Gamma(\delta+y_{\text{new}})\Gamma(\gamma+N_{\text{new}}-y_{\text{new}})}{\Gamma(\delta+\gamma+N_{\text{new}})}$$

15

Theoretical note: So, what is the relationship of the expected value point estimate to the full integral over r?

Imagine continuing to collect data.

The variance in the posterior belief in r will continue to decrease, to such an extent that in the **limit** there is a single value of *r* that had probability 1 of occurring with $p(r|y_N)$ zero everywhere else.

$$\mathbf{E}_{p(r|y_N)}\{P(Y_{\text{new}} \leq 6|r)\} = \int_{r=0}^{r=1} P(Y_{\text{new}} \leq 6|r)p(r|y_N)\, dr$$
$$= P(Y_{\text{new}} \leq 6|\hat{r})$$

16

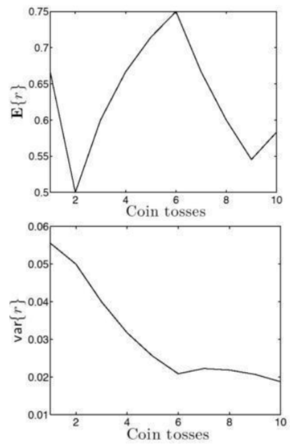
Scenario 1: Unknown Coin

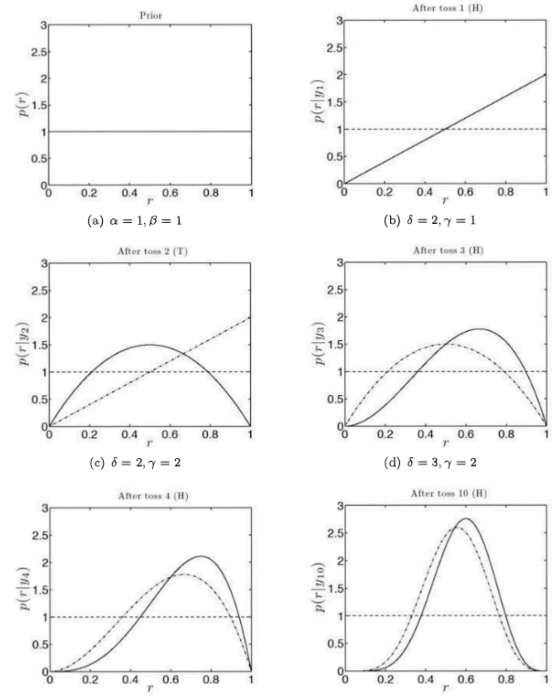
$\alpha = \beta = 1$

Observations:

H T H H H H T T T H

H H T T H H H H H H





(a) $\alpha = 1, \beta = 1$ (b) $\delta = 2, \gamma = 1$ (c) $\delta = 2, \gamma = 2$ (d) $\delta = 3, \gamma = 2$ (e) $\delta = 4, \gamma = 2$ (f) $\delta = 7, \gamma = 5$

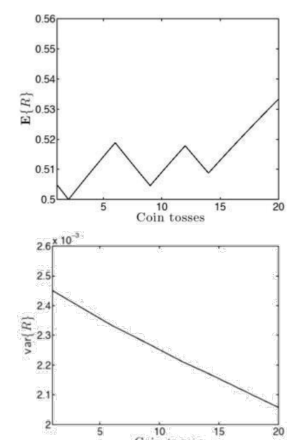
Scenario 2: Fair Coin

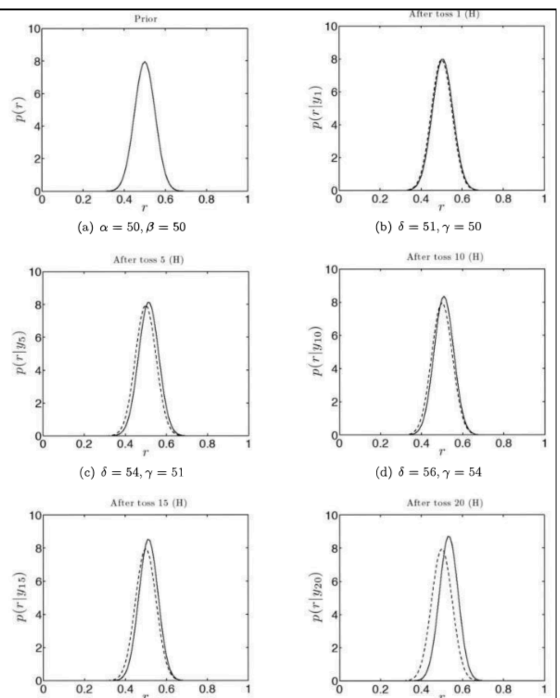
$\alpha = \beta = 50$

Observations:

H T H H H H T T T H

H H T T H H H H H H





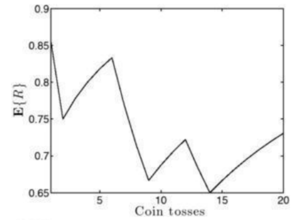
(a) $\alpha = 50, \beta = 50$ (b) $\delta = 51, \gamma = 50$ (c) $\delta = 54, \gamma = 51$ (d) $\delta = 56, \gamma = 54$ (e) $\delta = 59, \gamma = 56$ (f) $\delta = 64, \gamma = 56$

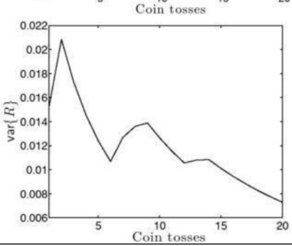
Scenario 3: Biased Coin

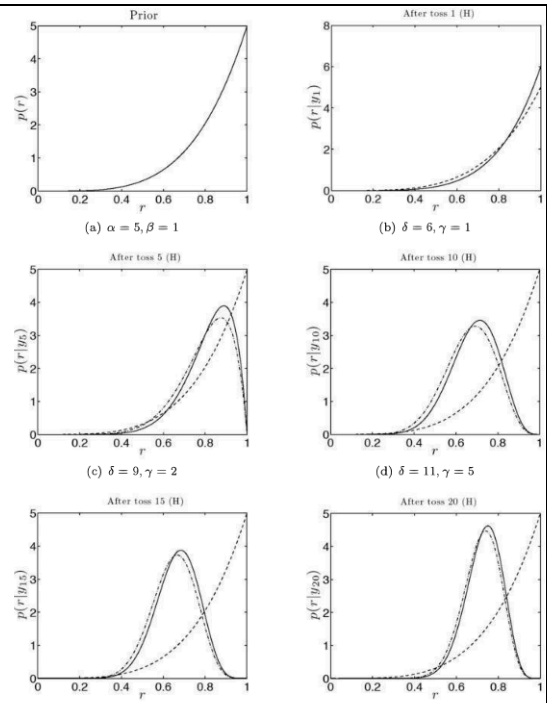
$\alpha = 5,\ \beta = 1$

Observations:

H T H H H H T T T H
H H T T H H H H H H



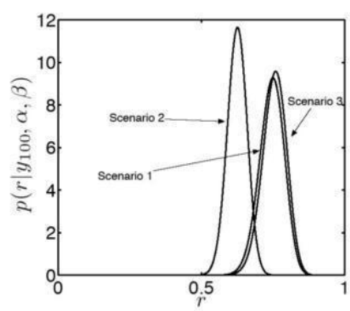




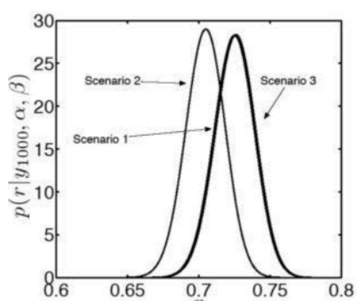
(a) $\alpha = 5, \beta = 1$ (b) $\delta = 6, \gamma = 1$ (c) $\delta = 9, \gamma = 2$ (d) $\delta = 11, \gamma = 5$ (e) $\delta = 14, \gamma = 7$ (f) $\delta = 19, \gamma = 7$

Summary

1. No prior knowledge: $\mathbf{E}_{p(r|y_N)}\left\{P(Y_{\mathsf{new}} \leq 6|r)\right\} = 0.4045$ 0.4047
2. Fair coin: $\mathbf{E}_{p(r|y_N)}\left\{P(Y_{\mathsf{new}} \leq 6|r)\right\} = 0.7579$
3. Biased coin: $\mathbf{E}_{p(r|y_N)}\left\{P(Y_{\mathsf{new}} \leq 6|r)\right\} = 0.2915.$



After 100 tosses



After 1000 tosses

20