ISTA 421 + INFO 521
Introduction to Machine Learning

Lecture 13:
Marginal Likelihood and Hyperpriors

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

4 October 2017 1

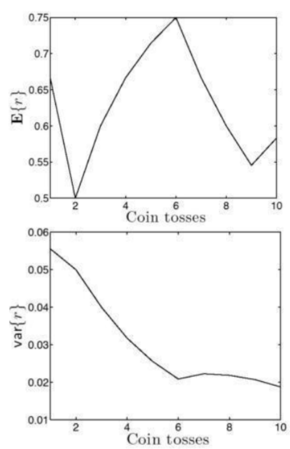
Scenario 1: Unknown Coin

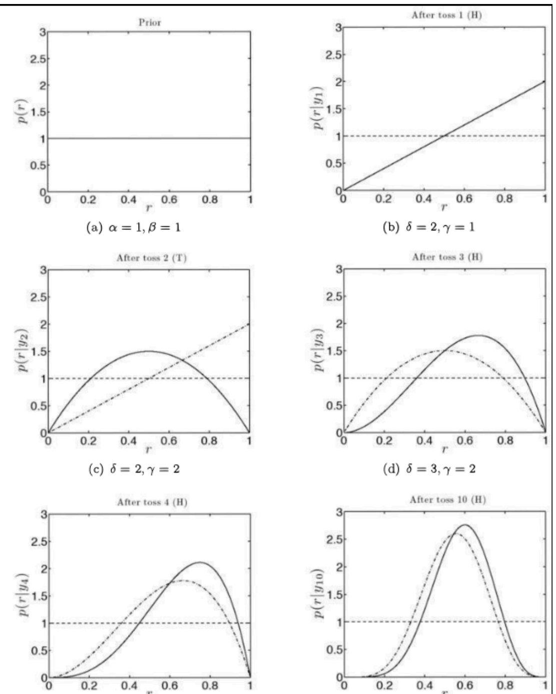
$$\alpha = \beta = 1$$

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

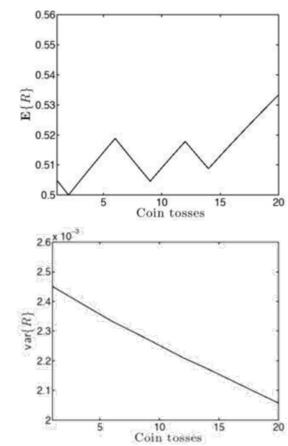
Scenario 2: Fair Coin

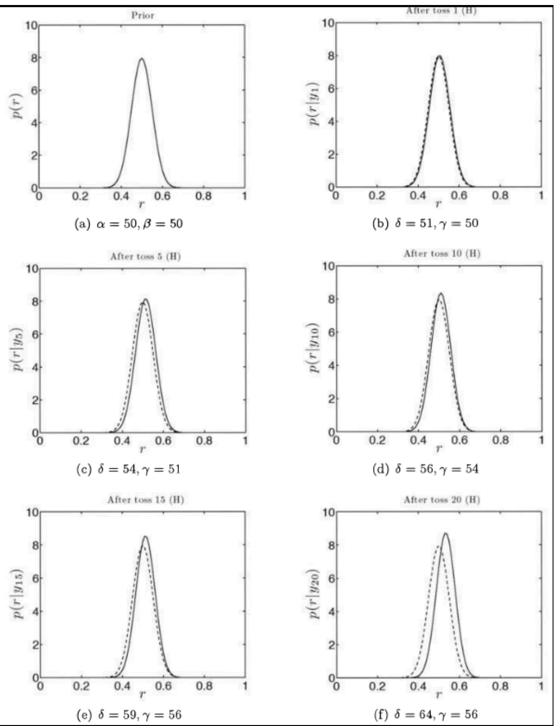
$\alpha = \beta = 50$

Observations:

H T H H H H T T T H

H H T T H H H H H H





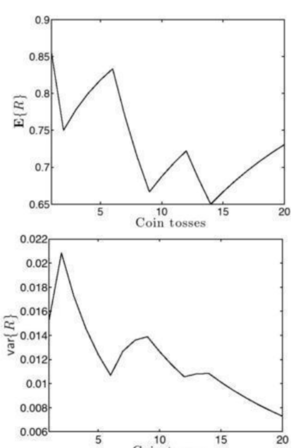
Scenario 3: Biased Coin

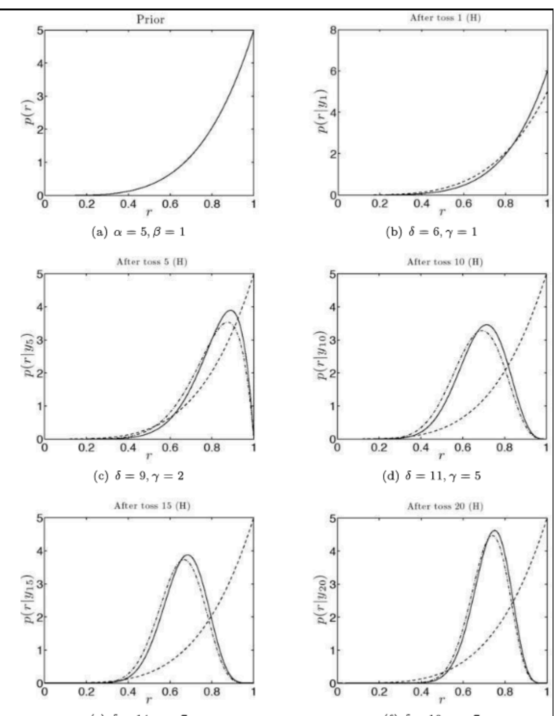
$\alpha = 5, \beta = 1$

Observations:

H T H H H H T T T H

H H T T H H H H H H



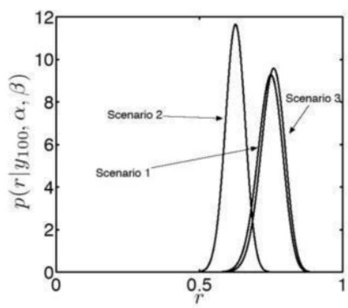


Summary

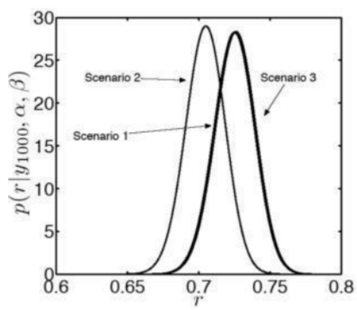
1. No prior knowledge: $\mathbf{E}_{p(r|y_N)}\left\{P(Y_{\text{new}} \leq 6|r)\right\} = 0.4045$ 0.4047

2. Fair coin: $\mathbf{E}_{p(r|y_N)}\left\{P(Y_{\text{new}} \leq 6|r)\right\} = 0.7579$

3. Biased coin: $\mathbf{E}_{p(r|y_N)}\left\{P(Y_{\text{new}} \leq 6|r)\right\} = 0.2915$.



After 100 tosses



After 1000 tosses

5

Which Prior is the Correct One?

- Well, it depends. Sometimes it is justified by background knowledge and context.
- As we get new data, the effect of the prior diminishes.
- Another approach: Look at the marginal likelihoods

6

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

- $P(y_N)$ is the marginal probability of the data. It can be related to r:

$$p(y_N) = \int_{r=0}^{r=1} p(r, y_N)\, dr = \int_{r=0}^{r=1} p(y_N|r)p(r)\, dr$$

- Need to be explicit about the parameters of r:

$$p(y_N|\alpha, \beta) = \int_{r=0}^{r=1} p(y_N|r)p(r|\alpha, \beta)\, dr$$

- This tells us how likely the data is given our choice of prior parameters α and β.
- The higher $p(y_N|\alpha,\beta)$, the better our evidence agrees with the prior specification.
- Can use $p(y_N|\alpha,\beta)$ to help choose best scenario: choose scenario with highest $p(y_N|\alpha,\beta)$.

7

Evaluate the Marginal Likelihood Integral

$$\begin{aligned} p(y_N|\alpha, \beta) &= \int_{r=0}^{r=1} p(y_N|r)p(r|\alpha, \beta)\, dr \\ &= \int_{r=0}^{r=1} \binom{N}{y_N} r^{y_N}(1-r)^{N-y_N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1}\, dr \\ &= \binom{N}{y_N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_{r=0}^{r=1} r^{\alpha+y_N-1}(1-r)^{\beta+N-y_N-1}\, dr. \end{aligned}$$

We've dealt with this integration problem before:

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

$$p(y_N|\alpha, \beta) = \binom{N}{y_N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_N)\Gamma(\beta+N-y_N)}{\Gamma(\alpha+\beta+N)}$$

Contrast with: (what's the difference?)

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

8

Evaluate the Marginal Likelihood for the different scenarios

$$p(y_N|\alpha,\beta) = \binom{N}{y_N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_N)\Gamma(\beta+N-y_N)}{\Gamma(\alpha+\beta+N)}$$

In our example, $N = 20$ and $y_N = 14$

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

1. No prior knowledge, $\alpha = \beta = 1$, $p(y_N \mid \alpha, \beta) = 0.0476$
2. Fair coin, $\alpha = \beta = 50$, $p(y_N \mid \alpha, \beta) = 0.0441$ ← lowest
3. Biased coin, $\alpha = 5$, $\beta = 1$, $p(y_N \mid \alpha, \beta) = 0.0576$ ← highest

Caution: Choosing this way makes the prior **no longer** correspond to our beliefs ***before*** we observe any data.

What's the danger?? overfitting

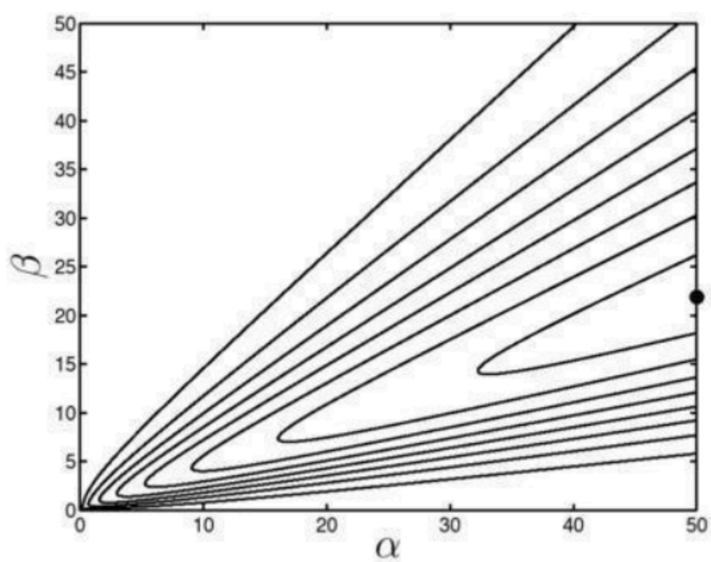
AKA: Type II Maximum Likelihood

9

Optimize α and β using Marginal Likelihood

$0 \leq \alpha \leq 50$

$0 \leq \beta \leq 30$



Treating Parameters as R.V.s

- In some cases we have good reason to select particular parameter values based on knowledge.
- Other times, we don't know the exact value, so treat as random variables themselves.
- Often useful and appropriate to treat as independent, and capitalize on conditional independence
- E.g., prior density over all parameter random variables

$$p(r, \alpha, \beta) = p(r|\alpha, \beta)p(\alpha, \beta)$$

- For our model, we want the posterior over all parameters in our model

$$\begin{aligned} p(r, \alpha, \beta | y_N) &= \frac{p(y_N | r, \alpha, \beta) p(r, \alpha, \beta)}{p(y_N)} \\ &= \frac{p(y_N|r)p(r, \alpha, \beta)}{p(y_N)} \quad \text{Conditional independence} \\ &= \frac{p(y_N|r)p(r|\alpha, \beta)p(\alpha, \beta)}{p(y_N)} \end{aligned}$$

11

R.V. Parameters may have… Hyperparameters!

- *κ* controls the density over *α* and β in the same way that *α* and β control the density for *r*

$$p(\alpha, \beta | \kappa)$$

- When computing marginal likelihood: integrate over all random variables, leaves us with the data conditioned on the hyper-parameters:

$$p(y_N|\kappa) = \int \int \int p(y_N|r)p(r|\alpha, \beta)p(\alpha, \beta|\kappa)\ dr\ d\alpha\ d\beta$$

… can keep going: hierarchical models

12

Probabilistic Graphical Models

… are a notation to compactly describe a complex relation of random variables:

nodes = RVs, edges = RV dependencies (parameterization)

$p(y_N|r)p(r|\alpha,\beta)p(\alpha|\kappa)p(\beta|\kappa)$

Toss of coin

How I say coin landed

$P(Y = y|\ X = x)$

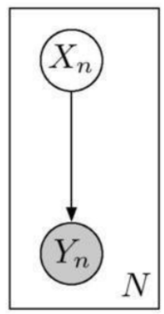
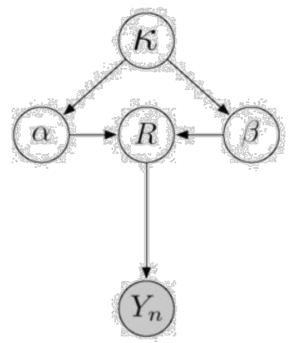


Plate Notation



Review

Let θ be some unobserved (population) parameter.
The function $\theta \mapsto f(x|\theta)$ is the likelihood function.
The *maximum likelihood* (ML) estimate of θ is then:

$$\hat{\theta}_{\mathrm{ML}}(x) = \arg\max_{\theta} f(x|\theta)$$

Now let's treat θ as a random variable itself.
Let g be a prior distribution over θ.
The posterior distribution of θ is defined, using Bayes' Theorem:

$$\theta \mapsto f(\theta|x) = \frac{f(x|\theta)g(\theta)}{\int_{\vartheta\in\Theta} f(x|\vartheta)g(\vartheta)\mathrm{d}\vartheta}$$

The *maximum a posteriori* (MAP) estimate of θ is the **mode** of the posterior distribution.

$$\hat{\theta}_{\mathrm{MAP}} x = \arg\max_{\theta} \frac{f(x|\theta)g(\theta)}{\int_{\vartheta\in\Theta} f(x|\vartheta)g(\vartheta)\mathrm{d}\vartheta} = \arg\max_{\theta} f(x|\theta)g(\theta)$$

Note: The denominator (the **marginal likelihood**, **probability of the evidence**, **partition function**) does not depend on θ, so it plays no role in the optimization!

Note: When the prior g is uniform (a constant function), then the MAP estimate of θ is the same as the ML estimate.

Review

We can select our parameters so that they maximize the marginal likelihood (type II maximum likelihood)

$$p(y_N|\alpha,\beta) = \int_{r=0}^{r=1} p(y_N|r)p(r|\alpha,\beta)\ dr$$

Finally, the "full" Bayesian approach: keep the full posterior over your parameters and when you must make a decision, integrate over the uncertainty in the parameters.

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

15

Return (again) to the Olympics 100m

The Bayesian treatment…

- First, the model:

$$t_n = w_0 + w_1 x_n + w_2 x_n^2 + \ldots + w_k x_n^k + \epsilon_n$$

k^th-order polynomial (Ch 1)
Gaussian distributed noise (Ch 2) $\epsilon \sim \mathcal{N}(0, \sigma^2)$

$$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n \qquad \mathbf{t} = \mathbf{X}^\top \mathbf{w} + \epsilon$$

$$\begin{aligned} p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2,\Delta) &= \frac{p(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2,\Delta)p(\mathbf{w}|\Delta)}{p(\mathbf{t}|\mathbf{X},\sigma^2,\Delta)} \\ &= \frac{p(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2)p(\mathbf{w}|\Delta)}{p(\mathbf{t}|\mathbf{X},\sigma^2,\Delta)}. \\ &= \frac{p(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2)p(\mathbf{w}|\Delta)}{\int p(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2)p(\mathbf{w}|\Delta)\ d\mathbf{w}} \end{aligned}$$