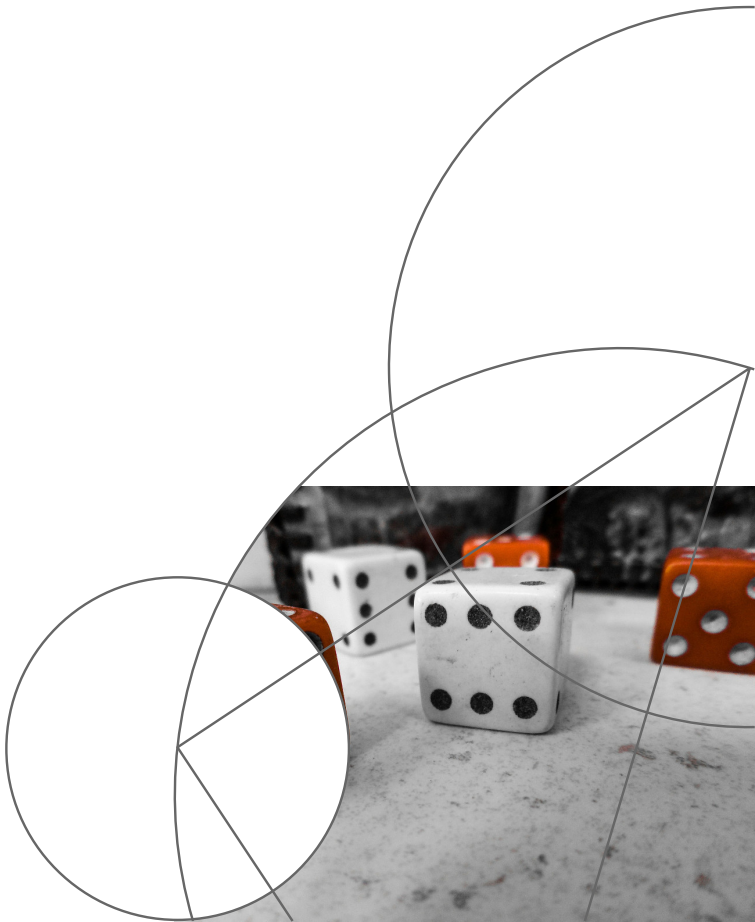
Bayesian Learning

Modelling and Analysis of Data (MAD)

Kim Steenstrup Pedersen
Based on Fabian Gieseke's slides



December 2021

Outline

1. Coin Game: Bayesian Perspective
   Read R&G Ch. 3.1 - 3.4

2. Bayesian Approach to Linear Regression
   Read R&G Ch. (3.5 - 3.7), 3.8, (3.9 - 3.10)

3. Summary & Outlook

Outline

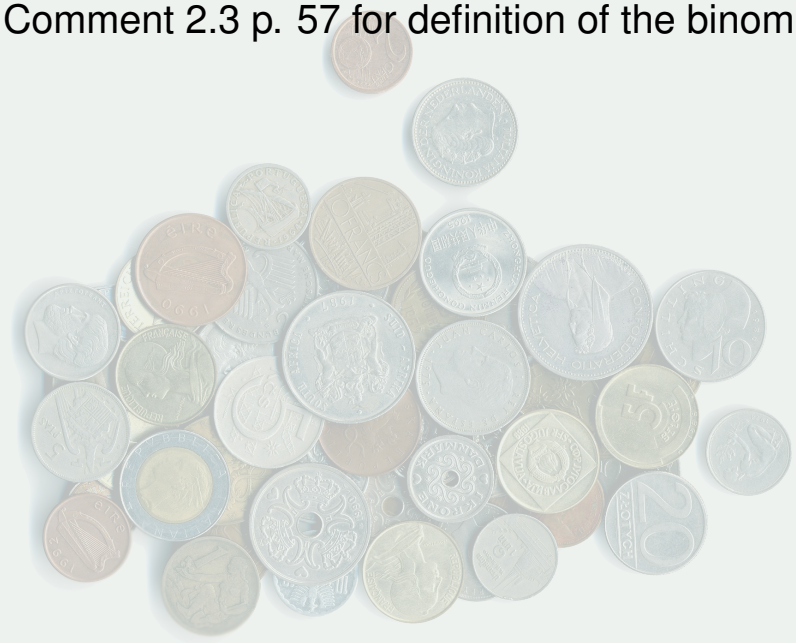
Coin Game

Binomial Distribution (Rogers & Girolami)

The binomial distribution describes the probability of a certain number of successes (heads or 1's) in N binary events. The probability of y heads from N tosses where each toss lands heads with probability r is given by

$$P(Y = y) = \binom{N}{y} r^y (1-r)^{N-y}$$

See Rogers & Girolami Comment 2.3 p. 57 for definition of the binomial coefficient.



Coin Game

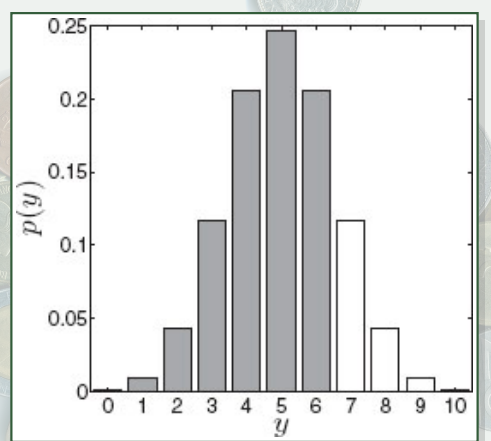
Binomial Distribution (Rogers & Girolami)

The binomial distribution describes the probability of a certain number of successes (heads or 1's) in N binary events. The probability of y heads from N tosses where each toss lands heads with probability r is given by

$$P(Y = y) = \binom{N}{y} r^y (1-r)^{N-y}$$

See Rogers & Girolami Comment 2.3 p. 57 for definition of the binomial coefficient.

Example: Given $r = 0.5$ and $N = 10$, what is the probability of $P(Y \leq 6)$?



$$P(Y \leq 6) = 1 - P(Y > 6) = 1 - (0.1172 + 0.0439 + 0.0098 + 0.0010) = 0.8281$$

Fair Coin: Expected Win

Coin Tossing Game

The stall owner tosses a coin $N = 10$ times. If the coin lands heads on six or fewer occasions, you get your stake of 1฿ plus an additional 1฿. Otherwise, you'll loose your stake of 1฿. Also assume that we have a fair coin ($r = 0.5$).

- Let X be a random variable with $X = 1$ in case you win and $X = 0$ in case you loose.

Fair Coin: Expected Win

Coin Tossing Game

The stall owner tosses a coin $N = 10$ times. If the coin lands heads on six or fewer occasions, you get your stake of 1฿ plus an additional 1฿. Otherwise, you'll loose your stake of 1฿. Also assume that we have a fair coin ($r = 0.5$).

- Let X be a random variable with $X = 1$ in case you win and $X = 0$ in case you loose.
- Let f be a function defining your winnings with $f(1) = 2$ and $f(0) = 0$.

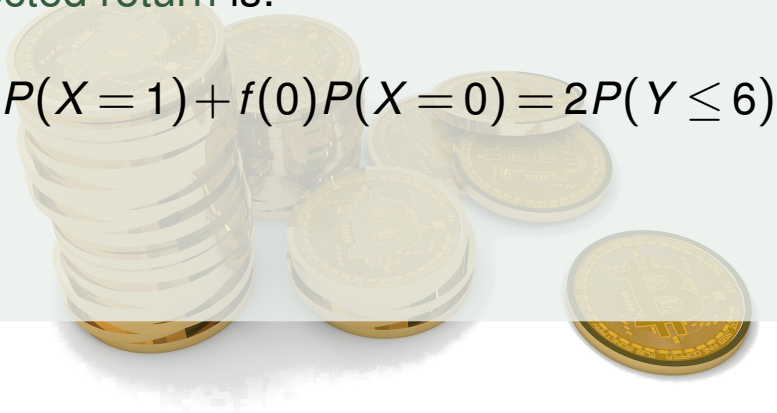
Fair Coin: Expected Win

Coin Tossing Game

The stall owner tosses a coin $N = 10$ times. If the coin lands heads on six or fewer occasions, you get your stake of 1฿ plus an additional 1฿. Otherwise, you'll loose your stake of 1฿. Also assume that we have a fair coin ($r = 0.5$).

- Let X be a random variable with $X = 1$ in case you win and $X = 0$ in case you loose.
- Let f be a function defining your winnings with $f(1) = 2$ and $f(0) = 0$.
- Then, our expected return is:

$$E[f(X)] = f(1)P(X = 1) + f(0)P(X = 0) = 2P(Y \leq 6) + 0P(Y > 6) = 1.6562$$



Fair Coin: Expected Win

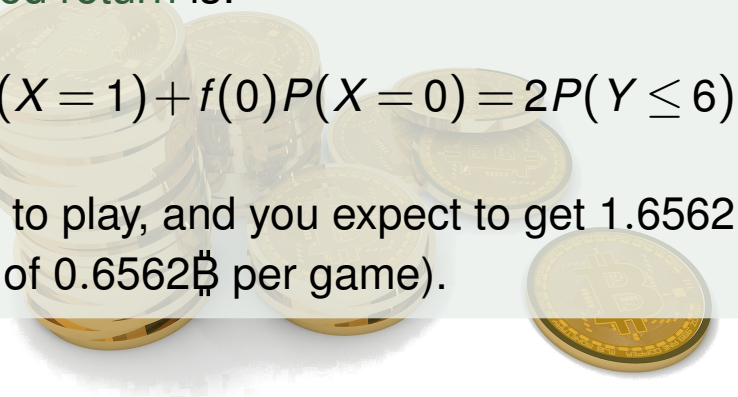
Coin Tossing Game

The stall owner tosses a coin $N = 10$ times. If the coin lands heads on six or fewer occasions, you get your stake of 1฿ plus an additional 1฿. Otherwise, you'll loose your stake of 1฿. Also assume that we have a fair coin ($r = 0.5$).

- Let X be a random variable with $X = 1$ in case you win and $X = 0$ in case you loose.
- Let f be a function defining your winnings with $f(1) = 2$ and $f(0) = 0$.
- Then, our expected return is:

$$E[f(X)] = f(1)P(X = 1) + f(0)P(X = 0) = 2P(Y \leq 6) + 0P(Y > 6) = 1.6562$$

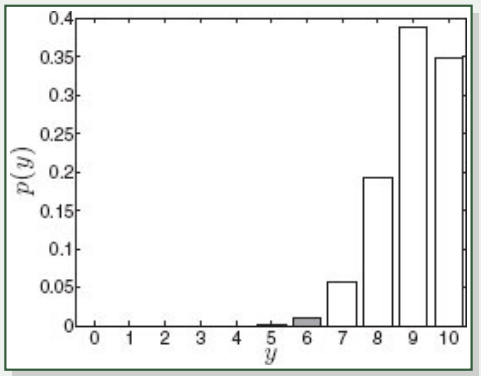
- Thus: It costs 1฿ to play, and you expect to get 1.6562฿ on average per game (i.e., a win of 0.6562฿ per game).



Coin Game: A First Run

Sequence coin tosses

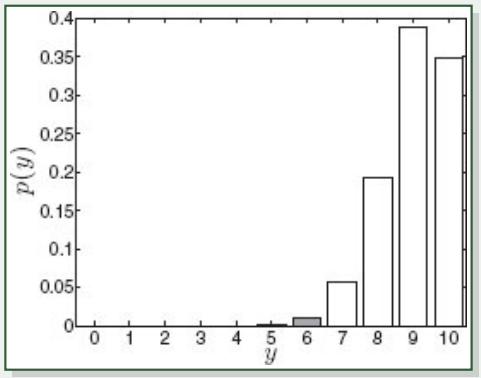
You play and get the following result: H,T,H,H,H,H,H,H,H,H. Let's use our MLE for r of a Bernoulli distribution (the probability of one coin toss) from last lecture: $\hat{r} = \frac{\sum_{n=1}^{N} x_n}{N} = \frac{9}{10}$. This induces the following distribution:



Coin Game: A First Run

Sequence coin tosses

You play and get the following result: H,T,H,H,H,H,H,H,H,H. Let's use our MLE for r of a Bernoulli distribution (the probability of one coin toss) from last lecture: $\hat{r} = \frac{\sum_{n=1}^{N} x_n}{N} = \frac{9}{10}$. This induces the following distribution:



And based on this MLE estimate $\hat{r}$ our expected return is:

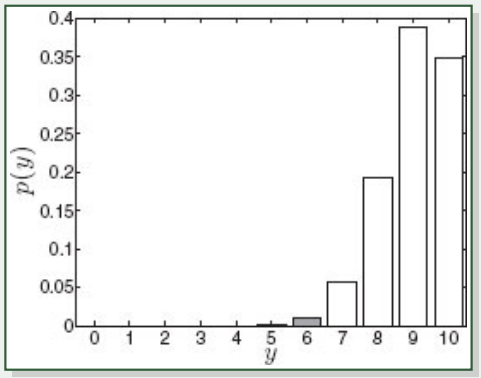
$$E[f(X)] = f(1)P(X=1) + f(0)P(X=0) = 2P(Y \leq 6) + 0P(Y > 6) = 0.0256$$

Thus: It costs 1฿ to play, and you expect to get 0.0256฿ on average per game (i.e., a loss of about -0.9744 ฿ per game).

Coin Game: A First Run

Sequence coin tosses

You play and get the following result: H,T,H,H,H,H,H,H,H,H. Let's use our MLE for r of a Bernoulli distribution (the probability of one coin toss) from last lecture: $\hat{r} = \frac{\sum_{n=1}^{N} x_n}{N} = \frac{9}{10}$. This induces the following distribution:



And based on this MLE estimate $\hat{r}$ our expected return is:

$$E[f(X)] = f(1)P(X=1) + f(0)P(X=0) = 2P(Y \leq 6) + 0P(Y > 6) = 0.0256$$

Thus: It costs 1฿ to play, and you expect to get 0.0256฿ on average per game (i.e., a loss of about -0.9744 ฿ per game).

Question: How certain are we about r? Maybe this was just bad luck ...

The Bayesian Way

- Each time we play, we might get a new (different) sequence, which induces a new $\hat{r}$. What is the uncertainty in r?

The Bayesian Way

- Each time we play, we might get a new (different) sequence, which induces a new $\hat{r}$. What is the uncertainty in r?
- Let's consider r to be the output of a random variable R. Further, let Y_N be a random variable denoting the number of heads in N tosses.

The Bayesian Way

- Each time we play, we might get a new (different) sequence, which induces a new $\hat{r}$. What is the uncertainty in r?
- Let's consider r to be the output of a random variable R. Further, let Y_N be a random variable denoting the number of heads in N tosses.
- Goal: We are interested in $p(R = r|Y_N = y_N) = p(r|y_N)$, i.e., in the probability density of R given the value of Y_N.

The Bayesian Way

- Each time we play, we might get a new (different) sequence, which induces a new $\hat{r}$. What is the uncertainty in r?
- Let's consider r to be the output of a random variable R. Further, let Y_N be a random variable denoting the number of heads in N tosses.
- Goal: We are interested in $p(R = r|Y_N = y_N) = p(r|y_N)$, i.e., in the probability density of R given the value of Y_N.
- Given this distribution $p(r|y_N)$, we could compute

$$P(Y_{new} \leq 6|y_N) \quad = \quad \int P(Y_{new} \leq 6, r|y_N)dr$$

where Y_{new} is another random variable describing the outcome of heads in a future game.

The Bayesian Way

- Each time we play, we might get a new (different) sequence, which induces a new $\hat{r}$. What is the uncertainty in r?
- Let's consider r to be the output of a random variable R. Further, let Y_N be a random variable denoting the number of heads in N tosses.
- Goal: We are interested in $p(R = r|Y_N = y_N) = p(r|y_N)$, i.e., in the probability density of R given the value of Y_N.
- Given this distribution $p(r|y_N)$, we could compute

$$\begin{aligned} P(Y_{new} \leq 6|y_N) &= \int P(Y_{new} \leq 6, r|y_N)dr \\ &= \int P(Y_{new} \leq 6|r, y_N)p(r|y_N)dr \end{aligned}$$

where Y_{new} is another random variable describing the outcome of heads in a future game.

The Bayesian Way

- Each time we play, we might get a new (different) sequence, which induces a new $\hat{r}$. What is the uncertainty in r?
- Let's consider r to be the output of a random variable R. Further, let Y_N be a random variable denoting the number of heads in N tosses.
- Goal: We are interested in $p(R = r|Y_N = y_N) = p(r|y_N)$, i.e., in the probability density of R given the value of Y_N.
- Given this distribution $p(r|y_N)$, we could compute

$$\begin{aligned} P(Y_{new} \leq 6|y_N) &= \int P(Y_{new} \leq 6, r|y_N)dr \\ &= \int P(Y_{new} \leq 6|r, y_N)p(r|y_N)dr \\ &= \int P(Y_{new} \leq 6|r)p(r|y_N)dr \end{aligned}$$

where Y_{new} is another random variable describing the outcome of heads in a future game.

(1) marginalisation, (2) chain rule, (3) in case we know r, then knowing y_N provides no further information on the prob. for Y_{new}

The Bayesian Way

Bayes' Rule

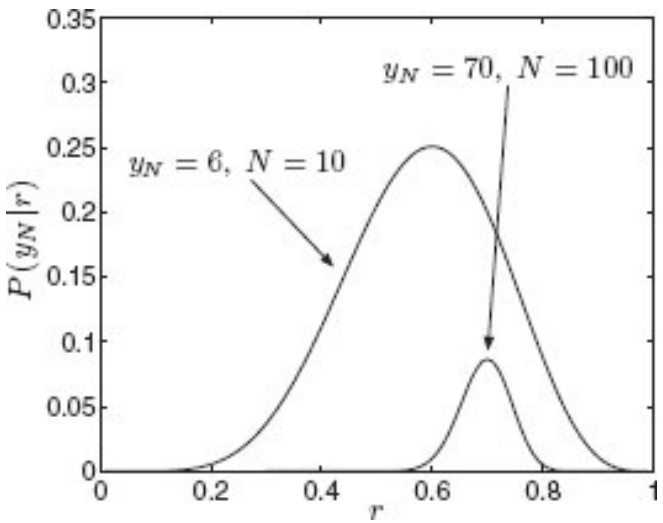
$$p(r|y_N) = \frac{p(y_N|r)p(r)}{p(y_N)}$$

The Bayesian Way

Bayes' Rule

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

The likelihood $p(y_N|r)$: How likely is it that we would observe the data y_N given a particular value for r?



The Bayesian Way

Bayes' Rule

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

The prior distribution $p(r)$: Allows us to express any prior belief we have in the value of r before we see any data.

The Bayesian Way

Bayes' Rule

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

The prior distribution $p(r)$: Allows us to express any prior belief we have in the value of r before we see any data. Three examples:

1. We do not know anything about the coin/stall owner.
2. We think the coin is fair.
3. We think the coin is biased to give more heads than tails.

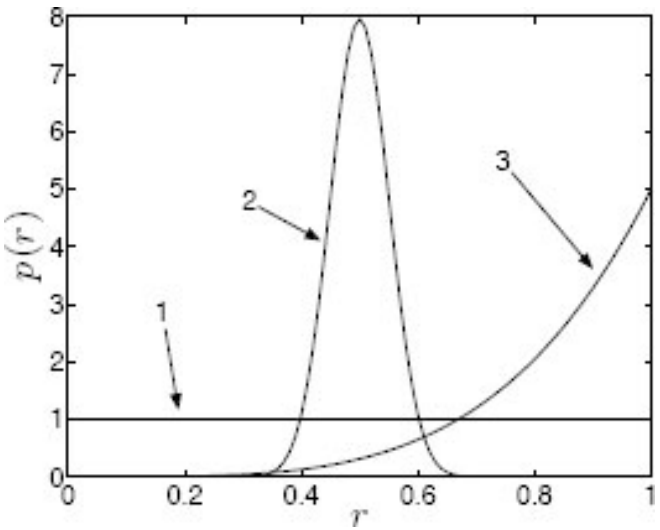
The Bayesian Way

Bayes' Rule

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

The prior distribution $p(r)$: Allows us to express any prior belief we have in the value of r before we see any data. Three examples:

1. We do not know anything about the coin/stall owner.
2. We think the coin is fair.
3. We think the coin is biased to give more heads than tails.



The Bayesian Way

Bayes' Rule

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

The marginal distribution $p(y_N)$ also known as the evidence: Acts as a normalization constant and is the likelihood of the data y_N averaged over all parameter values.

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

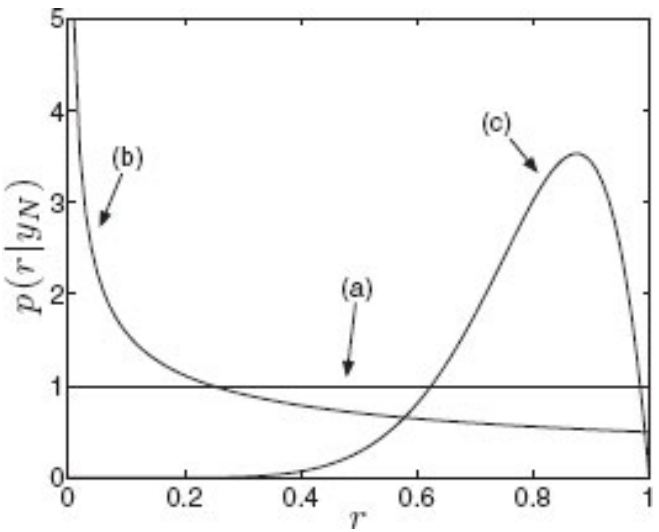
(1) marginalisation/integrating out, (2) factorisation/chain rule

The Bayesian Way

Bayes' Rule

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

The posterior distribution $p(r|y_N)$: The distribution we are interested in. It is the result of updating our prior belief $p(r)$ given new data y_N.



Computing the Posterior

Bayes' Rule

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

Comment 3.1 – Conjugate priors (Rogers & Girolami)

A likelihood-prior pair is said to be conjugate if they result in a posterior which is of the same form as the prior. This enables us to compute the posterior density analytically without having to worry about computing the denominator in Bayes' rule, the marginal likelihood.

Prior	Likelihood
Gaussian	Gaussian
Beta	Binomial
Gamma	Gaussian
Dirichlet	Multinomial

Computing the Posterior

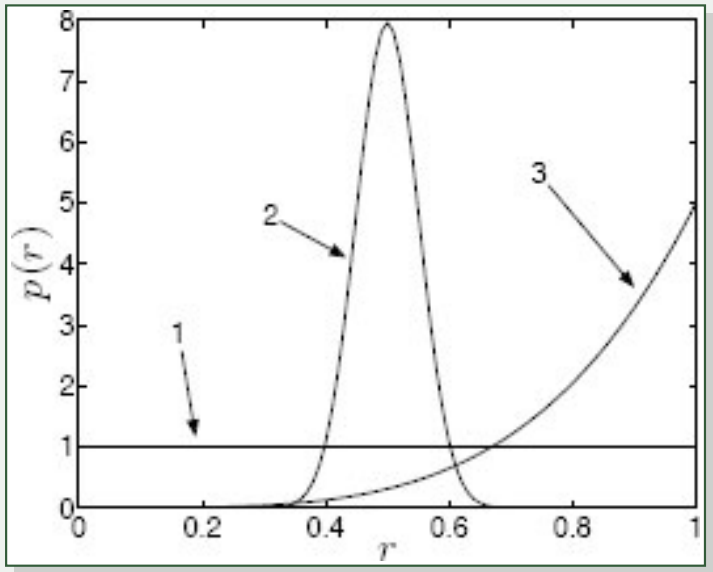
Beta Distribution (Rogers & Girolami)

The beta density function can be used for continuous random variables that are restricted to values between 0 and 1. The beta density function is defined as

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

where $\alpha \geq 0$ and $\beta \geq 0$ are the parameters that control the shape of the density function. $\Gamma(z)$ is the so-called gamma function.

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



Computing the Posterior

Bayes' Rule

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

- Let's use a beta distribution for the prior $p(r)$. In this case, since the prior and the likelihood, which is a binomial distribution, are conjugate, we know that the posterior $p(r|y_N)$ has to be a beta distribution.

Computing the Posterior

Bayes' Rule

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

- Let's use a beta distribution for the prior $p(r)$. In this case, since the prior and the likelihood, which is a binomial distribution, are conjugate, we know that the posterior $p(r|y_N)$ has to be a beta distribution.
- Let's omit $p(y_N)$ from the above equation, i.e.,

$$p(r|y_N) \propto p(y_N|r)p(r)$$

($\propto$ means proportional, i.e., $p(r|y_N) = c \cdot p(y_N|r)p(r)$ for a constant c)

Computing the Posterior

Bayes' Rule

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

- Let's use a beta distribution for the prior $p(r)$. In this case, since the prior and the likelihood, which is a binomial distribution, are conjugate, we know that the posterior $p(r|y_N)$ has to be a beta distribution.
- Let's omit $p(y_N)$ from the above equation, i.e.,

$$p(r|y_N) \propto p(y_N|r)p(r)$$

($\propto$ means proportional, i.e., $p(r|y_N) = c \cdot p(y_N|r)p(r)$ for a constant c)

- Plugging in the binomial and the beta distribution leads to

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

Computing the Posterior

- Plugging in the binomial and the beta distribution leads to

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

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

Computing the Posterior

- Plugging in the binomial and the beta distribution leads to

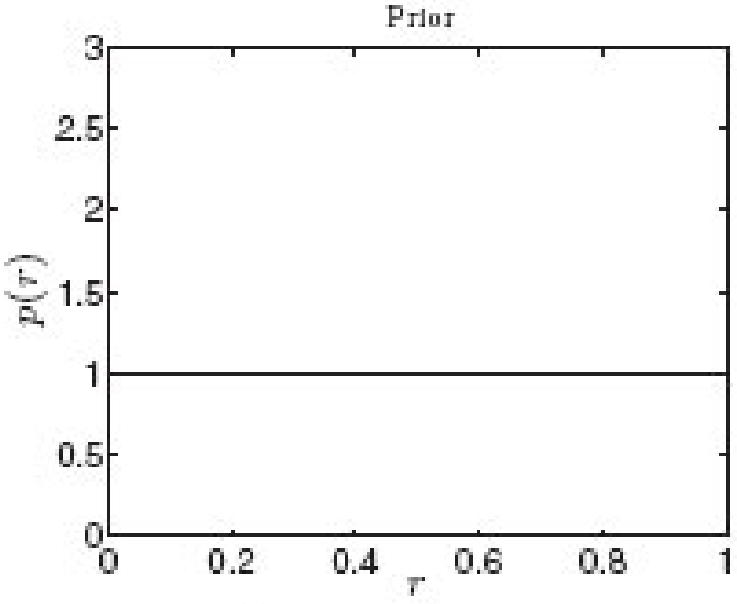
$$\begin{aligned} p(r|y_N) &\propto \left[\binom{N}{y_N} r^{y_N}(1-r)^{N-y_N}\right] \times \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1}\right] \\ &= \left[\binom{N}{y_N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right] \times \left[r^{y_N}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

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

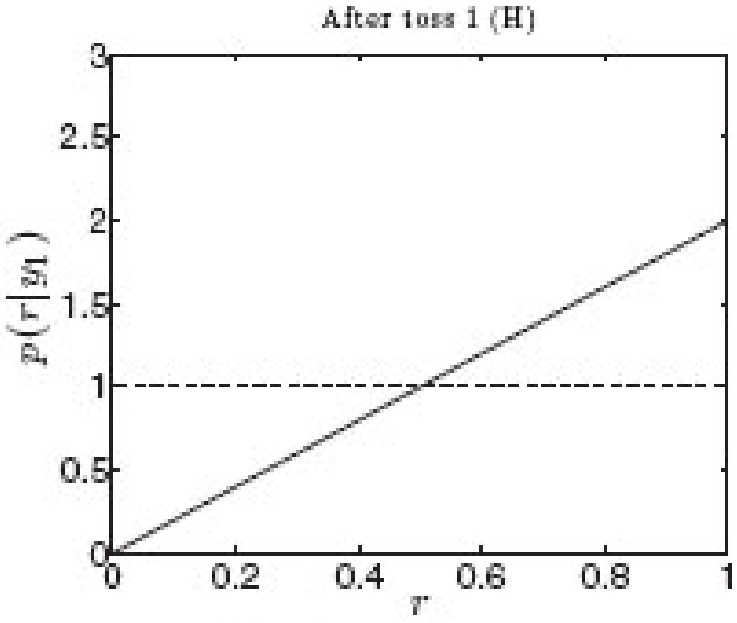
- Since we know that $p(r|y_N)$ must be a beta distribution as well (as the prior distribution!), we directly know the normalization constant. Thus:

$$p(r|y_N) = \frac{\Gamma(\delta+\gamma)}{\Gamma(\delta)\Gamma(\gamma)} r^{\delta-1}(1-r)^{\gamma-1}$$

Example: No Prior Knowledge



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

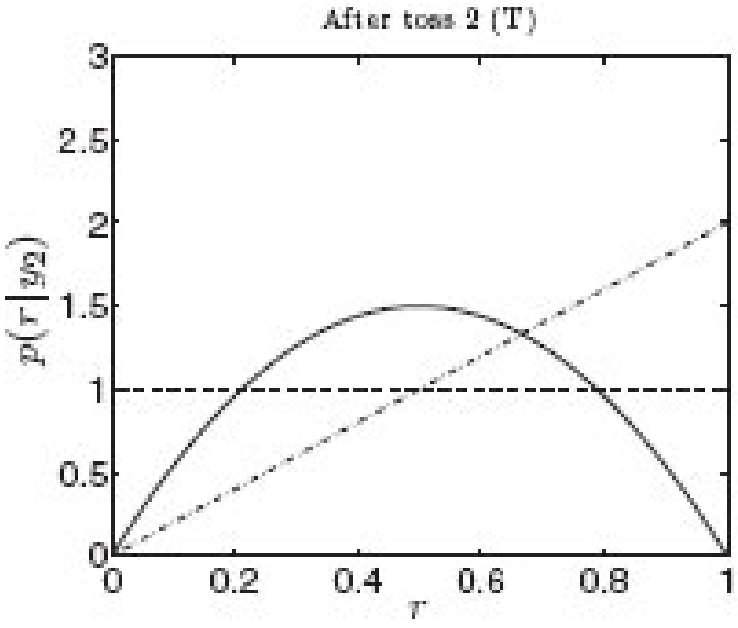


(b) $\delta = 2, \gamma = 1$

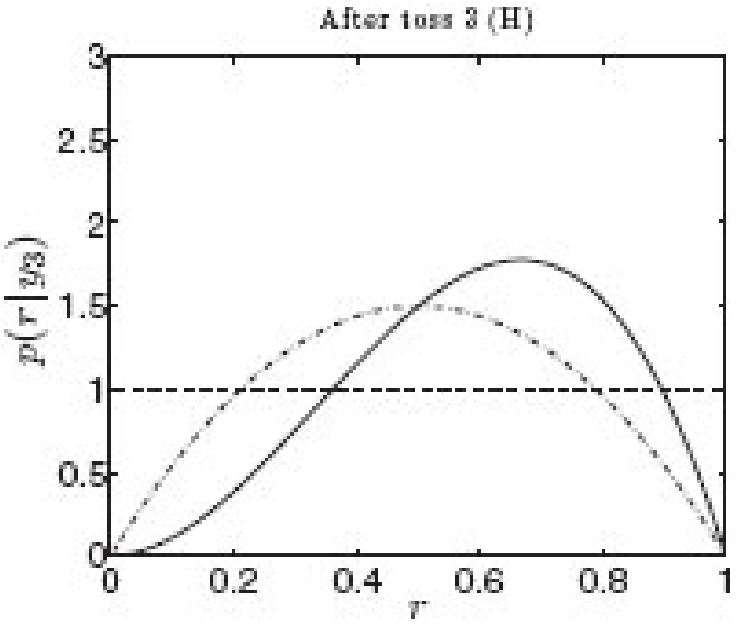
- Left: $\mathbf{E}_{p(r)}\{R\} = \frac{1}{2}$ and $\mathbf{var}\{R\} = \frac{1}{12}$
- Right: $\mathbf{E}_{p(r|y_1)}\{R\} = \frac{2}{3}$ and $\mathbf{var}\{R\} = \frac{1}{18}$

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

Example: No Prior Knowledge



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

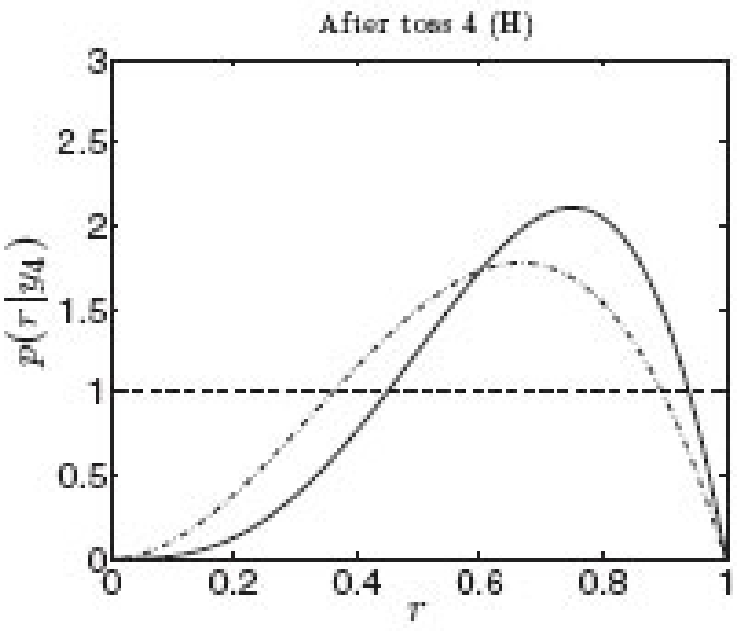


(d) $\delta = 3, \gamma = 2$

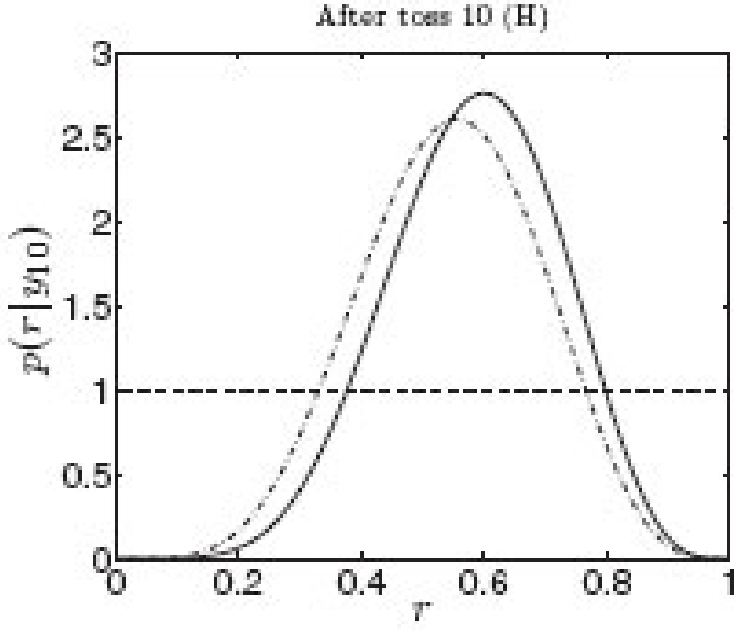
- Left: $\mathbf{E}_{p(r|y_2)}\{R\} = \frac{1}{2}$ and $\mathbf{var}\{R\} = \frac{1}{20}$
- Right: $\mathbf{E}_{p(r|y_3)}\{R\} = \frac{3}{5}$ and $\mathbf{var}\{R\} = \frac{1}{25}$

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

Example: No Prior Knowledge



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

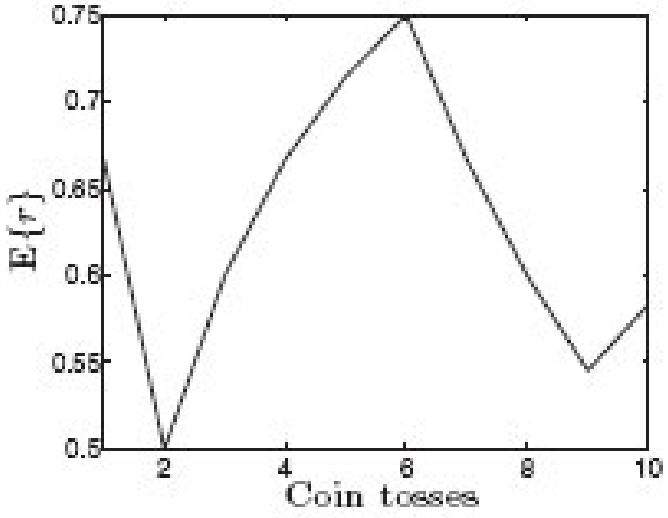


(f) $\delta = 7, \gamma = 5$

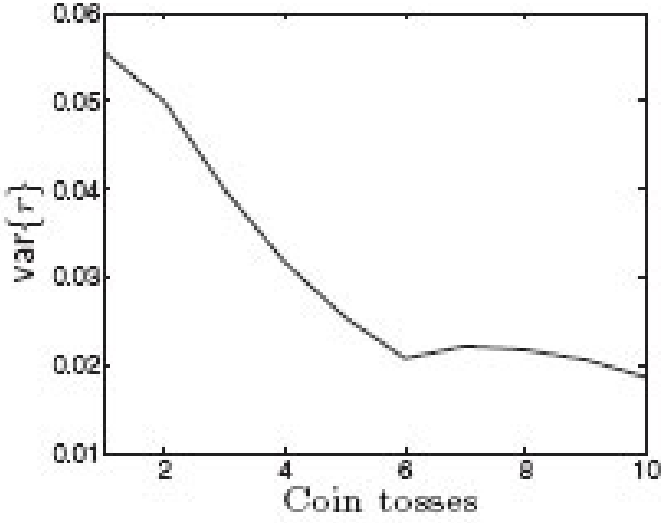
- The complete sequence is H, T, H, H, H, H, T, T, T, H
- Left: $\mathbf{E}_{p(r|y_4)}\{R\} = \frac{2}{3}$ and $\mathbf{var}\{R\} = \frac{2}{63}$
- Right: $\mathbf{E}_{p(r|y_{10})}\{R\} = \frac{7}{12}$ and $\mathbf{var}\{R\} = 0.0187$

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

Example: No Prior Knowledge



(a) Expected value

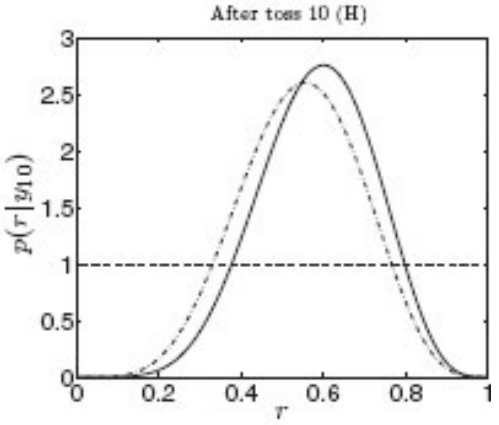


(b) Variance

- The complete sequence is H, T, H, H, H, H, T, T, T, H

Example: No Prior Knowledge

- Let's revisit the idea of using a point estimate $\hat{r}$.

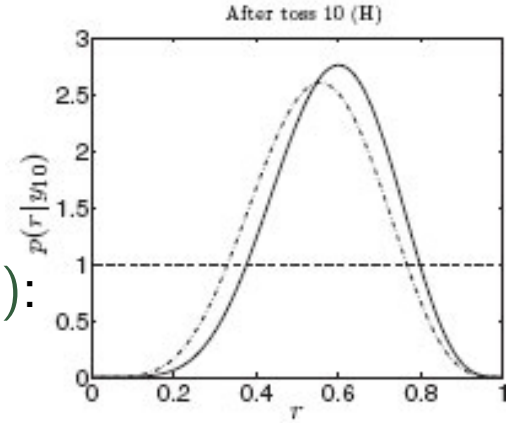


(f) $\delta = 7, \gamma = 5$

Example: No Prior Knowledge

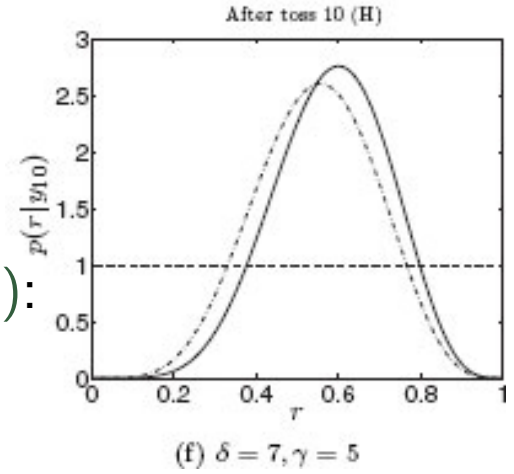
- Let's revisit the idea of using a point estimate $\hat{r}$.
- A sensible choice is the expected value w.r.t. $p(r|y_N)$:

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



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

Example: No Prior Knowledge



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

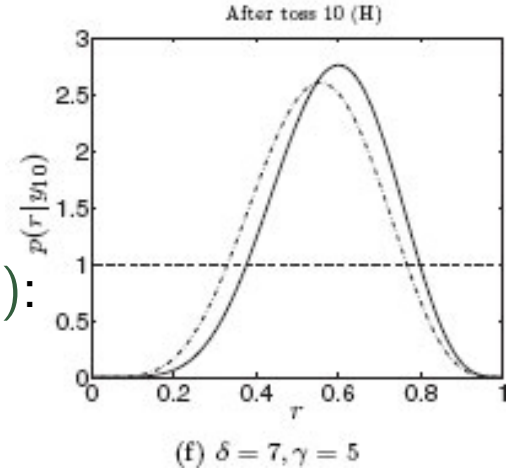
- Let's revisit the idea of using a point estimate $\hat{r}$.
- A sensible choice is the expected value w.r.t. $p(r|y_N)$:

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

- Given this point estimate, we can compute

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

Example: No Prior Knowledge



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

- Let's revisit the idea of using a point estimate $\hat{r}$.
- A sensible choice is the expected value w.r.t. $p(r|y_N)$:

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

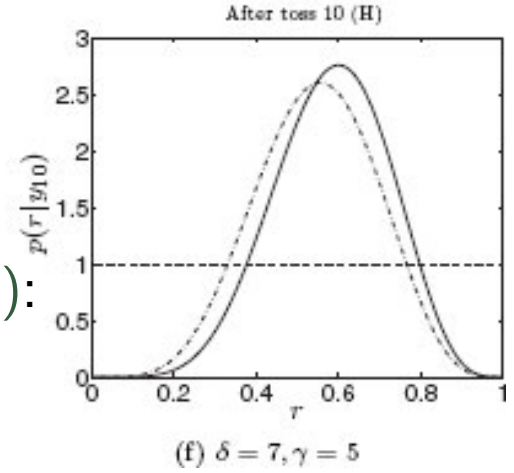
- Given this point estimate, we can compute

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

- Ideally, we would like to use all the posterior information that is given in $p(r|y_N)$, which requires computing

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

Example: No Prior Knowledge



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

- Let's revisit the idea of using a point estimate $\hat{r}$.
- A sensible choice is the expected value w.r.t. $p(r|y_N)$:

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

- Given this point estimate, we can compute

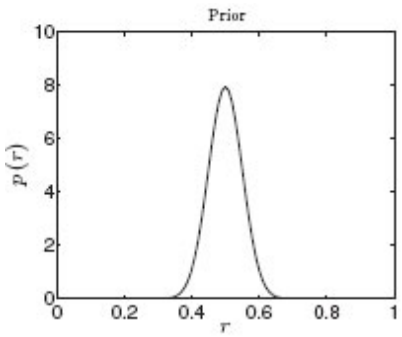
$$P(Y_{new} \leq 6|\hat{r}) = 1 - \sum_{y_{new}=7}^{10} P(Y_{new} = y_{new}|\hat{r}) = 1 - 0.3414 = 0.6586$$

- Ideally, we would like to use all the posterior information that is given in $p(r|y_N)$, which requires computing

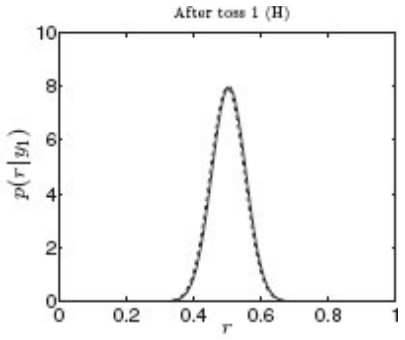
$$\mathbf{E}_{p(r|y_N)}\{P(Y_{new} \leq 6|r)\}$$

- One can show that $\mathbf{E}_{p(r|y_N)}\{P(Y_{new} \leq 6|r)\} = 0.6055$ (see textbook)

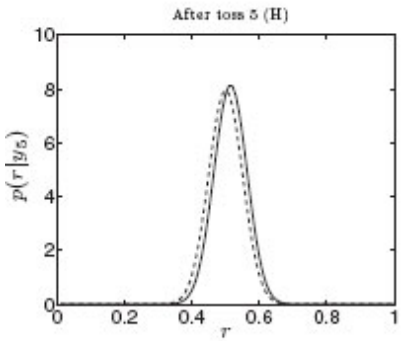
Example: Fair Coin



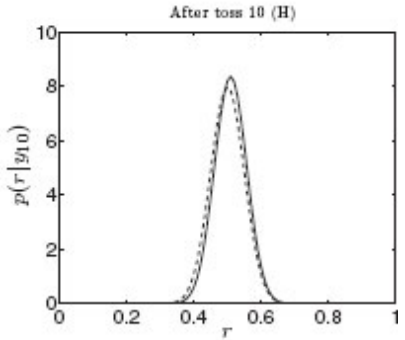
(a) $\alpha = 50, \beta = 50$



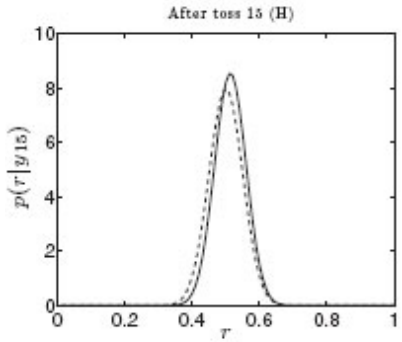
(b) $\delta = 51, \gamma = 50$



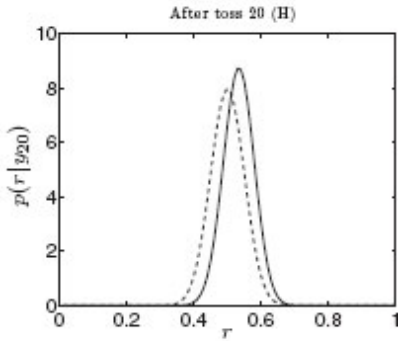
(c) $\delta = 54, \gamma = 51$



(d) $\delta = 56, \gamma = 54$

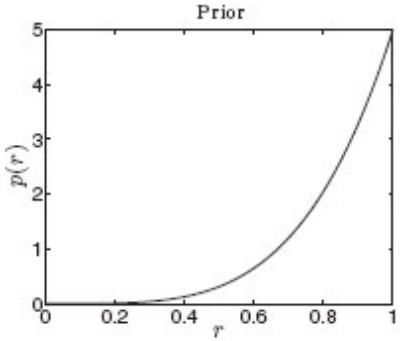


(e) $\delta = 59, \gamma = 56$

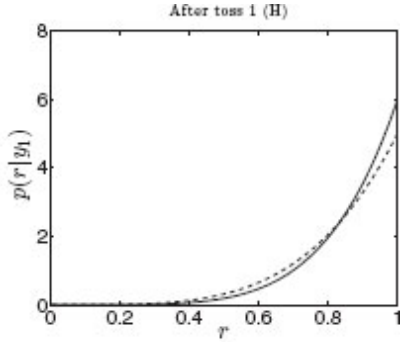


(f) $\delta = 64, \gamma = 56$

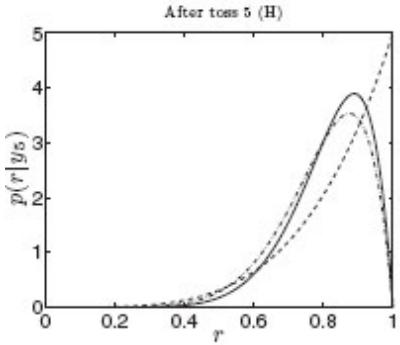
Example: Biased Coin



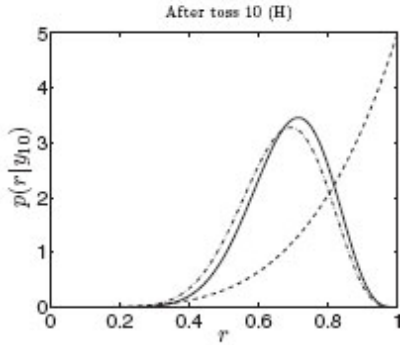
(a) $\alpha = 5, \beta = 1$



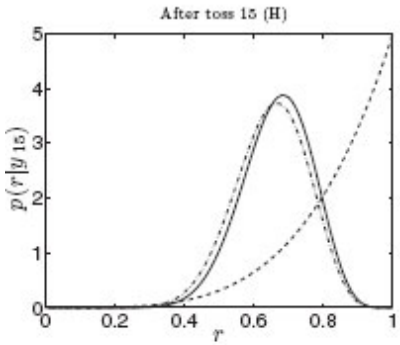
(b) $\delta = 6, \gamma = 1$



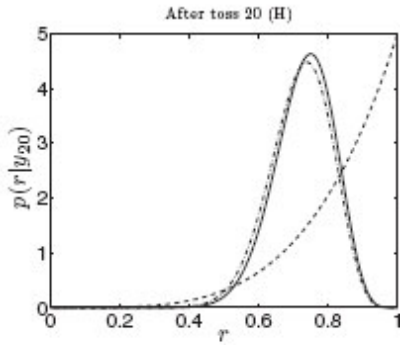
(c) $\delta = 9, \gamma = 2$



(d) $\delta = 11, \gamma = 5$

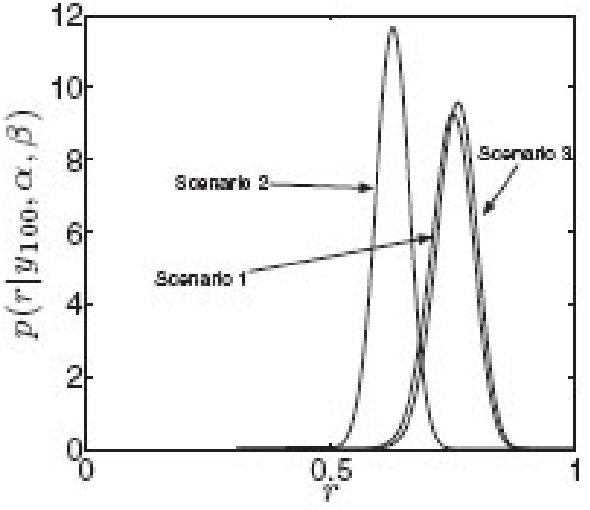


(e) $\delta = 14, \gamma = 7$

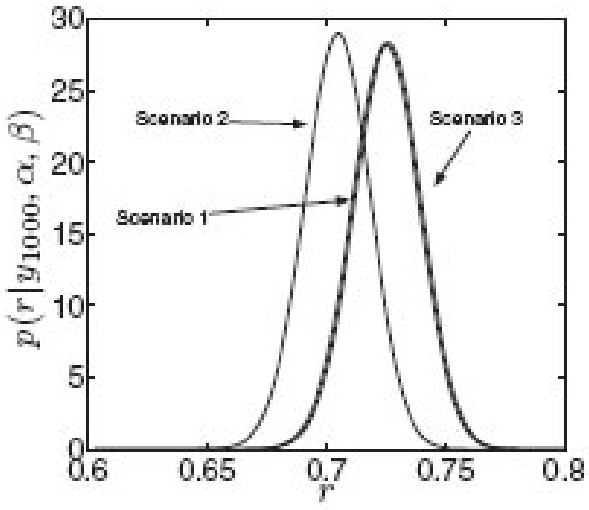


(f) $\delta = 19, \gamma = 7$

Examples: More Data



(a) The three posteriors after 100 tosses



(b) The three posteriors after 1000 tosses

- The posterior distribution depends on the choice of prior:
  - As we observer more and more coin tosses the posterior distributions approaches the same disitrbution irrespective of choice of prior.
  - The less informative (uncomitted) priors converges faster to a more proper posterior distribution over r.

Outline

The likelihood of the observed dataset

- We need to simultaneously maximize the likelihood of all the data points
- $p(t_1, \ldots, t_N | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{w}, \sigma^2) = ?$
- A joint density – we make another assumption: Our noise ε_n is i.i.d. (independently and identically distributed)
- $L(\boldsymbol{w}, \sigma^2) = p(t_1, \ldots, t_N | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n | \boldsymbol{x}_n, \boldsymbol{w}, \sigma^2)$
- Why did I not assume that the t_n were i.i.d.?
- They are not! There is a deterministic part of the model, which means the t_n are very much dependent ...
- New task: $\text{argmax}_{(\boldsymbol{w}, \sigma^2)} L(\boldsymbol{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n | \boldsymbol{x}_n, \boldsymbol{w}, \sigma^2)$

Maximum Likelihood Estimate

$$\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{t}$$

$$\widehat{\sigma^2} = \frac{1}{N} (\mathbf{t}^T \mathbf{t} - \mathbf{t}^T \mathbf{X} \hat{\mathbf{w}})$$

Single Point Estimate

Regression: Statistical Point of View

- Again, let's consider the generative linear regression model of the form

$$t_n = f(\boldsymbol{x}_n, \boldsymbol{w}) + \varepsilon_n,$$

where the random noise $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$ for some variance σ^2.

Regression: Statistical Point of View

- Again, let's consider the generative linear regression model of the form

$$t_n = f(\boldsymbol{x}_n, \boldsymbol{w}) + \varepsilon_n,$$

  where the random noise $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$ for some variance σ^2.
- We can write this for a data set of N observations using matrix-vector notation:

$$\mathbf{t} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}$$

  with $\mathbf{t} = [t_1, \ldots, t_N]^T$, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times (D+1)}$, $\mathbf{w} \in \mathbb{R}^{D+1}$, and $\boldsymbol{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_N]^T$.

Regression: Statistical Point of View

- Again, let's consider the generative linear regression model of the form

$$t_n = f(\boldsymbol{x}_n, \boldsymbol{w}) + \varepsilon_n,$$

  where the random noise $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$ for some variance σ^2.
- We can write this for a data set of N observations using matrix-vector notation:

$$\mathbf{t} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}$$

  with $\mathbf{t} = [t_1, \ldots, t_N]^T$, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times (D+1)}$, $\mathbf{w} \in \mathbb{R}^{D+1}$, and $\boldsymbol{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_N]^T$.
- In the following, we will consider $\mathbf{w}$ as a sample from a random vector $\mathbf{W}$. However, we will assume that we know the true value for the noise σ^2.

Regression: Statistical Point of View

- Again, let's consider the generative linear regression model of the form

$$t_n = f(\boldsymbol{x}_n, \boldsymbol{w}) + \varepsilon_n,$$

  where the random noise $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$ for some variance σ^2.
- We can write this for a data set of N observations using matrix-vector notation:

$$\mathbf{t} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}$$

  with $\mathbf{t} = [t_1, \ldots, t_N]^T$, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times (D+1)}$, $\mathbf{w} \in \mathbb{R}^{D+1}$, and $\boldsymbol{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_N]^T$.
- In the following, we will consider $\mathbf{w}$ as a sample from a random vector $\mathbf{W}$. However, we will assume that we know the true value for the noise σ^2.
- Goal: Based on the data and our model assumptions, find the posterior distribution $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta)$ for $\mathbf{w}$ (Δ will be defined later). Given this distribution for $\mathbf{w}$ (that captures all the involved uncertainties!), we can compute predictions by "averaging" over all possible $\mathbf{w}$.

Regression: Bayesian Approach

- Based on the data and our model assumptions, find the posterior distribution $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta)$ for $\mathbf{w}$. Using Bayes' rule, we obtain

where Δ are some parameters that define the prior distribution over $\mathbf{w}$.

Regression: Bayesian Approach

- Based on the data and our model assumptions, find the posterior distribution $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta)$ for $\mathbf{w}$. Using Bayes' rule, we obtain

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta) = \frac{p(\mathbf{t}, \mathbf{w}|\mathbf{X}, \sigma^2, \Delta)}{p(\mathbf{t}|\mathbf{X}, \sigma^2, \Delta)} = \frac{p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2, \Delta)p(\mathbf{w}|\mathbf{X}, \sigma^2, \Delta)}{p(\mathbf{t}|\mathbf{X}, \sigma^2, \Delta)}$$

where Δ are some parameters that define the prior distribution over $\mathbf{w}$.

Regression: Bayesian Approach

- Based on the data and our model assumptions, find the posterior distribution $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta)$ for $\mathbf{w}$. Using Bayes' rule, we obtain

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta) = \frac{p(\mathbf{t}, \mathbf{w}|\mathbf{X}, \sigma^2, \Delta)}{p(\mathbf{t}|\mathbf{X}, \sigma^2, \Delta)} = \frac{p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2, \Delta)p(\mathbf{w}|\mathbf{X}, \sigma^2, \Delta)}{p(\mathbf{t}|\mathbf{X}, \sigma^2, \Delta)}$$
$$= \frac{p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2)p(\mathbf{w}|\Delta)}{p(\mathbf{t}|\mathbf{X}, \sigma^2, \Delta)}$$

where Δ are some parameters that define the prior distribution over $\mathbf{w}$.

Regression: Bayesian Approach

- Based on the data and our model assumptions, find the posterior distribution $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta)$ for $\mathbf{w}$. Using Bayes' rule, we obtain

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta) = \frac{p(\mathbf{t}, \mathbf{w}|\mathbf{X}, \sigma^2, \Delta)}{p(\mathbf{t}|\mathbf{X}, \sigma^2, \Delta)} = \frac{p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2, \Delta)p(\mathbf{w}|\mathbf{X}, \sigma^2, \Delta)}{p(\mathbf{t}|\mathbf{X}, \sigma^2, \Delta)}$$
$$= \frac{p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2)p(\mathbf{w}|\Delta)}{p(\mathbf{t}|\mathbf{X}, \sigma^2, \Delta)}$$

  where Δ are some parameters that define the prior distribution over $\mathbf{w}$.

- We can further expand the marginal likelihood:

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

Regression: Bayesian Approach

- Based on the data and our model assumptions, find the posterior distribution $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta)$ for $\mathbf{w}$. Using Bayes' rule, we obtain

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta) = \frac{p(\mathbf{t}, \mathbf{w}|\mathbf{X}, \sigma^2, \Delta)}{p(\mathbf{t}|\mathbf{X}, \sigma^2, \Delta)} = \frac{p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2, \Delta)p(\mathbf{w}|\mathbf{X}, \sigma^2, \Delta)}{p(\mathbf{t}|\mathbf{X}, \sigma^2, \Delta)}$$
$$= \frac{p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2)p(\mathbf{w}|\Delta)}{p(\mathbf{t}|\mathbf{X}, \sigma^2, \Delta)}$$

  where Δ are some parameters that define the prior distribution over $\mathbf{w}$.

- We can further expand the marginal likelihood:

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

Regression: Likelihood

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

We have already derived the likelihood (in the MLE lecture):

$$p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^T\mathbf{x}_n, \sigma^2)$$

Regression: Likelihood

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

We have already derived the likelihood (in the MLE lecture):

$$p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^T\mathbf{x}_n, \sigma^2)$$

This can be expressed via a single N-dimensional Gaussian density with mean $\mathbf{Xw}$ and variance $\sigma^2\mathbf{I}$:

$$p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2) = \mathcal{N}(\mathbf{Xw}, \sigma^2\mathbf{I})$$

To see this, remember that the product of exponential functions can be written as the exponential function of a sum of terms.

Regression: Prior

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

We would like to choose a prior distribution $p(\mathbf{w}|\Delta)$ that is conjugate to the Gaussian likelihood such that we can derivate the exact posterior.

Comment 3.1 – Conjugate priors (Rogers & Girolami)

A likelihood-prior pair is said to be conjugate if they result in a posterior which is of the same form as the prior. This enables us to compute the posterior density analytically without having to worry about computing the denominator in Bayes' rule, the marginal likelihood.

Prior	Likelihood
Gaussian	Gaussian
Beta	Binomial
Gamma	Gaussian
Dirichlet	Multinomial

Regression: Prior

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

We would like to choose a prior distribution $p(\mathbf{w}|\Delta)$ that is conjugate to the Gaussian likelihood such that we can derivate the exact posterior. Thus, we will use a Gaussian prior for $\mathbf{w}$:

$$p(\mathbf{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$$

Regression: Prior

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

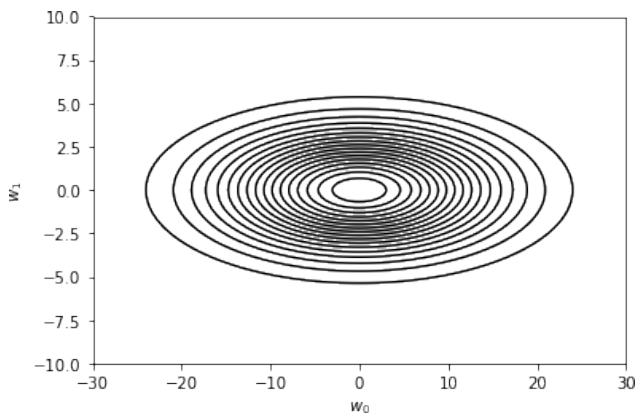
We would like to choose a prior distribution $p(\mathbf{w}|\Delta)$ that is conjugate to the Gaussian likelihood such that we can derivate the exact posterior. Thus, we will use a Gaussian prior for $\mathbf{w}$:

$$p(\mathbf{w}|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) = \mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$$

One possible choice:

1. We know nothing about the mean: $\boldsymbol{\mu}_0 = [0,0]^T$
2. We expect that w_0 might have large values; we expect w_1 to have smaller values. We assume that these two parameters are independent:

$$\boldsymbol{\Sigma}_0 = \begin{bmatrix} 100 & 0 \\ 0 & 5 \end{bmatrix}$$



Regression: Posterior

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

Since both the likelihood and the posterior are Gaussian (conjugate likelihood and prior), we can derive the exact posterior.

Regression: Posterior

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

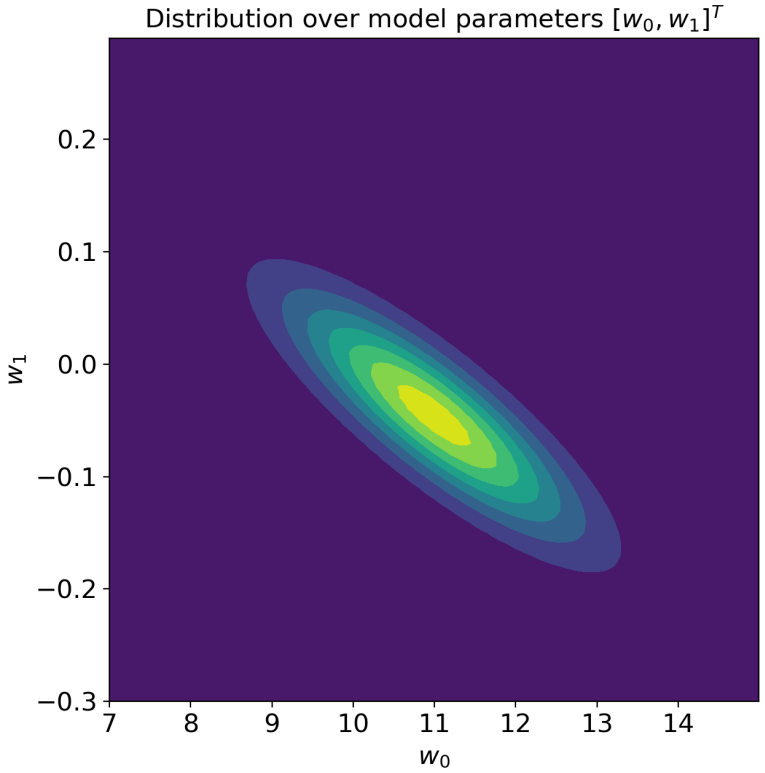
Since both the likelihood and the posterior are Gaussian (conjugate likelihood and prior), we can derive the exact posterior. This leads to

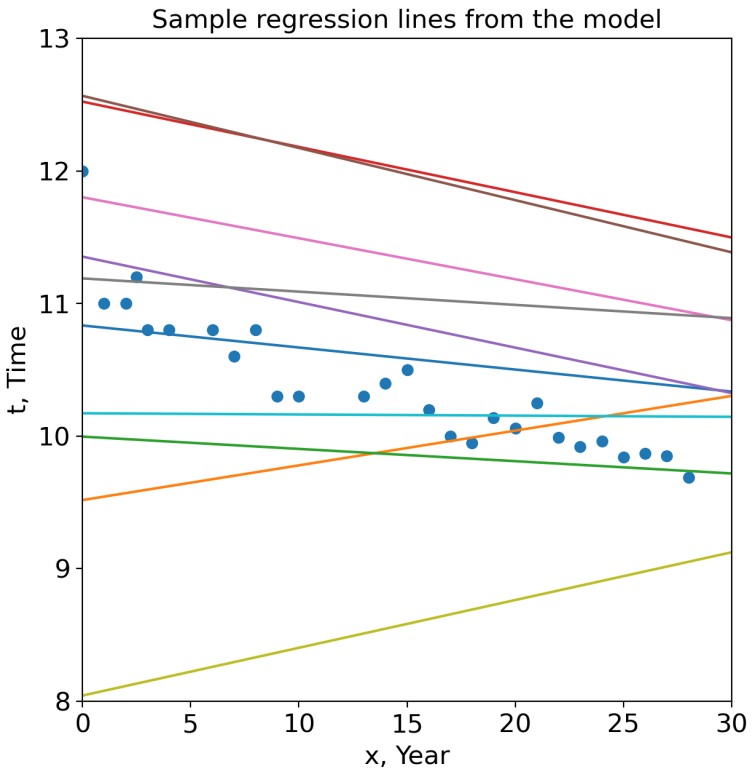
$$p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2,\Delta) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{w},\boldsymbol{\Sigma}_\mathbf{w})$$

with

$$\boldsymbol{\Sigma}_\mathbf{w} = \left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1} \quad \text{and} \quad \boldsymbol{\mu}_\mathbf{w} = \boldsymbol{\Sigma}_\mathbf{w}\left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{t} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)$$

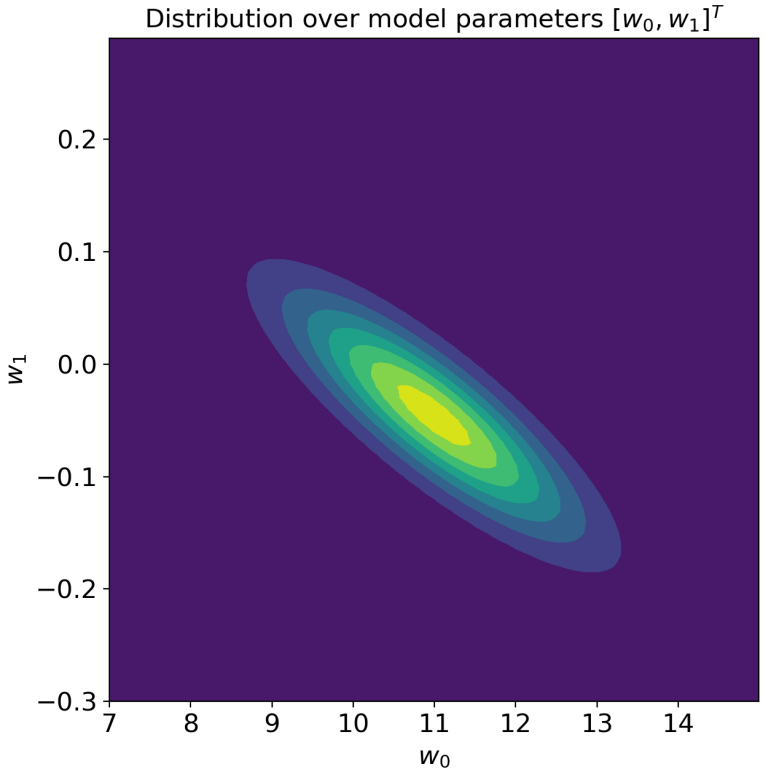
The posterior for the Olympic 100m data set

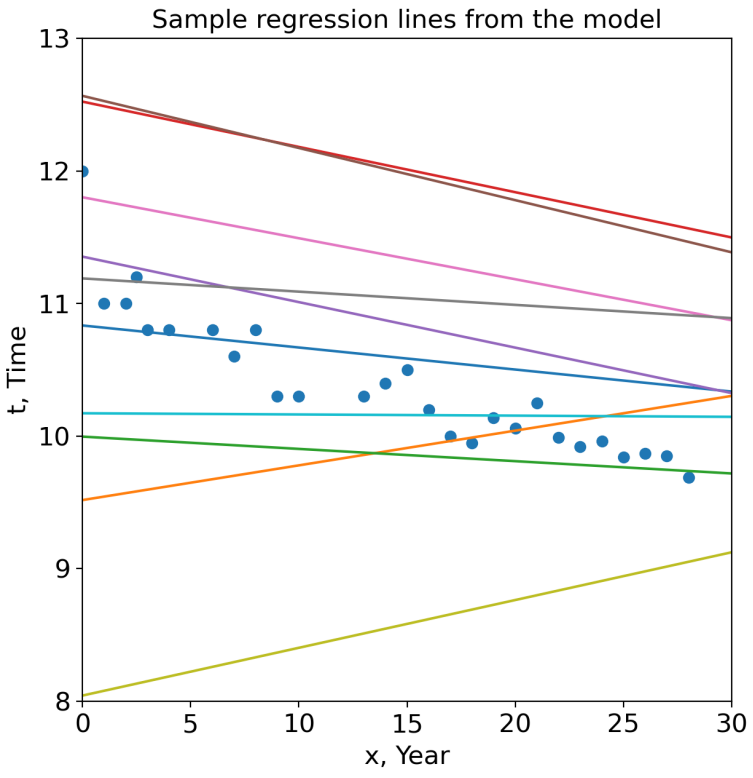




Regression with a linear generative model $t_n = f(\boldsymbol{x}_n, \boldsymbol{w}) + \varepsilon_n$ of polynomial order 1, that is $\boldsymbol{w} = (w_0, w_1)^T$ and $\boldsymbol{x}_n = (1, x_n)^T$.

The posterior for the Olympic 100m data set





Regression with a linear generative model $t_n = f(\boldsymbol{x}_n, \boldsymbol{w}) + \varepsilon_n$ of polynomial order 1, that is $\boldsymbol{w} = (w_0, w_1)^T$ and $\boldsymbol{x}_n = (1, x_n)^T$.

This is actually the illustration we ask for in Assignment 4, Exercise 4 d) (now you just have to write the code).

Outlook: Bayesian Inference ...

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

Let's consider a new set of features $\mathbf{x}_{new}$. Then, the predictive density of the target variable t_{new} is given by

$$p(t_{new}|\mathbf{x}_{new}, \mathbf{t}, \mathbf{X}, \sigma^2, \Delta) = \int p(t_{new}, \mathbf{w}|\mathbf{x}_{new}, \mathbf{t}, \mathbf{X}, \sigma^2, \Delta)d\mathbf{w}$$

Outlook: Bayesian Inference ...

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

Let's consider a new set of features $\mathbf{x}_{new}$. Then, the predictive density of the target variable t_{new} is given by

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

Outlook: Bayesian Inference ...

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

Let's consider a new set of features $\mathbf{x}_{new}$. Then, the predictive density of the target variable t_{new} is given by

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

Outlook: Bayesian Inference ...

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

Let's consider a new set of features $\mathbf{x}_{new}$. Then, the predictive density of the target variable t_{new} is given by

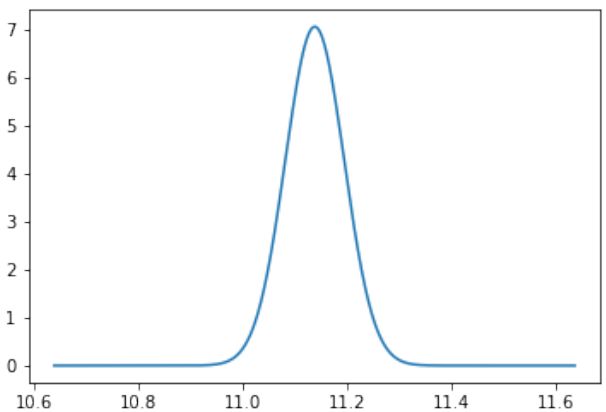
$$\begin{aligned}
p(t_{new}|\mathbf{x}_{new},\mathbf{t},\mathbf{X},\sigma^2,\Delta) &= \int p(t_{new},\mathbf{w}|\mathbf{x}_{new},\mathbf{t},\mathbf{X},\sigma^2,\Delta)d\mathbf{w} \\
&= \int p(t_{new}|\mathbf{w},\mathbf{x}_{new},\mathbf{t},\mathbf{X},\sigma^2,\Delta)p(\mathbf{w}|\mathbf{x}_{new},\mathbf{t},\mathbf{X},\sigma^2,\Delta)d\mathbf{w} \\
&= \int p(t_{new}|\mathbf{x}_{new},\mathbf{w},\sigma^2)p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2,\Delta)d\mathbf{w} \\
&= \mathbf{E}_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2,\Delta)}\left\{p(t_{new}|\mathbf{x}_{new},\mathbf{w},\sigma^2)\right\}
\end{aligned}$$

Outlook: Bayesian Inference ...

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

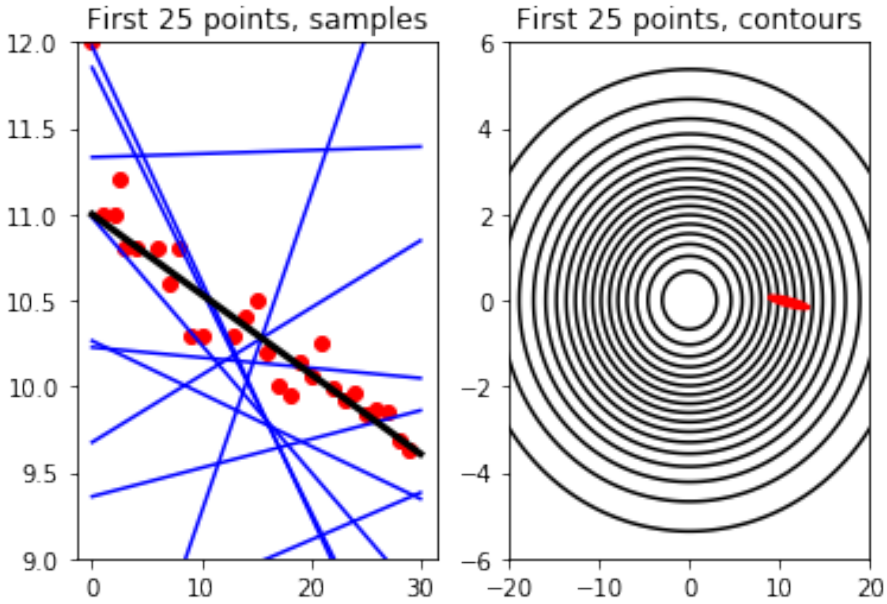
Let's consider a new set of features $\mathbf{x}_{new}$. Then, the predictive density of the target variable t_{new} is given by

$$\begin{aligned} p(t_{new}|\mathbf{x}_{new},\mathbf{t},\mathbf{X},\sigma^2,\Delta) &= \int p(t_{new},\mathbf{w}|\mathbf{x}_{new},\mathbf{t},\mathbf{X},\sigma^2,\Delta)d\mathbf{w} \\ &= \int p(t_{new}|\mathbf{w},\mathbf{x}_{new},\mathbf{t},\mathbf{X},\sigma^2,\Delta)p(\mathbf{w}|\mathbf{x}_{new},\mathbf{t},\mathbf{X},\sigma^2,\Delta)d\mathbf{w} \\ &= \int p(t_{new}|\mathbf{x}_{new},\mathbf{w},\sigma^2)p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2,\Delta)d\mathbf{w} \\ &= \mathbf{E}_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2,\Delta)}\left\{p(t_{new}|\mathbf{x}_{new},\mathbf{w},\sigma^2)\right\} \end{aligned}$$



Outline

Summary & Outlook



This lecture

- Coin Game
- Bayesian Learning & Regression

Outlook

- Classification (after Christmas)