Two Envelopes Revisited

• The “two envelopes” problem set-up
  ▪ Two envelopes: one contains $X$, other contains $2X$
  ▪ You select an envelope and open it
    ○ Let $Y = \$$ in envelope you selected
    ○ Let $Z = \$$ in other envelope
    \[
    E[Z \mid Y] = \frac{1}{2} \cdot \frac{Y}{2} + \frac{1}{2} \cdot 2Y = \frac{5}{4} Y
    \]
  ▪ Before opening envelope, think either equally good
    ○ So, what happened by opening envelope?
  ▪ $E[Z \mid Y]$ above assumes all values $X$ (where $0 < X < \infty$) are equally likely
    ○ Note: there are infinitely many values of $X$
    ○ So, not true probability distribution over $X$ (doesn’t integrate to 1)
Subjectivity of Probability

• Belief about contents of envelopes
  • Since implied distribution over X is not a true probability distribution, what is our distribution over X?
    • Frequentist: play game infinitely many times and see how often different values come up.
    • Problem: I only allow you to play the game once
  • Bayesian probability
    • Have prior belief of distribution for X (or anything for that matter)
    • Prior belief is a subjective probability
      • By extension, all probabilities are subjective
    • Allows us to answer question when we have no/limited data
      • E.g., probability a coin you’ve never flipped lands on heads
    • As we get more data, prior belief is “swamped” by data
The Envelope, Please

- **Bayesian**: have prior distribution over $X$, $P(X)$
  - Let $Y =$ $ in envelope you selected
  - Let $Z =$ $ in other envelope
  - Open your envelope to determine $Y$
  - If $Y > E[Z | Y]$, keep your envelope, otherwise switch
    - No inconsistency!
  - Opening envelope provides data to compute $P(X | Y)$ and thereby compute $E[Z | Y]$
  - Of course, there’s the issue of how you determined your prior distribution over $X$…
    - Bayesian: Doesn’t matter how you determined prior, but you *must* have one (whatever it is)
    - Imagine if envelope you opened contained $10.01$
The Dreaded Half Cent
Revisiting Bayes’ Theorem

Bayes’ Theorem ($\theta = \text{model parameters}, D = \text{data}$):

- **Likelihood**: you’ve seen this before (in context of MLE)
  - Probability of data given probability model (parameter $\theta$)
- **Prior**: before seeing any data, what is belief about model
  - I.e., what is distribution over parameters $\theta$
- **Posterior**: after seeing data, what is belief about model
  - After data $D$ observed, have posterior distribution $p(\theta | D)$ over parameters $\theta$ conditioned on data. Use this to predict new data.
  - Here, we assume prior and posterior distribution have same parametric form (we call them “conjugate”)

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

- Bayes’ Theorem ($\theta$ = model parameters, $D$ = data):

$$P(\theta \mid D) = \frac{P(D \mid \theta) P(\theta)}{P(D)}$$

- We have prior $P(\theta)$ and can compute $P(D \mid \theta)$
- But how do we calculate $P(D)$?
  - Complicated answer: $P(D) = \int P(D \mid \theta)P(\theta) \, d\theta$
  - Easy answer: It does not depend on $\theta$, so ignore it
    - Just a constant that forces $P(\theta \mid D)$ to integrate to 1
P(θ | D) for Beta and Bernoulli

- Prior: θ ~ Beta(a, b); D = \{n heads, m tails\}

\[ f_{\theta|D}(\theta = p | D) = \frac{f_D(\theta = p) f_{\theta}(\theta = p)}{f_D(D)} \]

\[ = \frac{\binom{n+m}{n} p^n (1-p)^m \cdot \frac{p^{a-1}(1-p)^{b-1}}{C_1}}{C_2} = \frac{n+m}{C_1 C_2} p^n (1-p)^m \cdot p^{a-1}(1-p)^{b-1} \]

\[ = C_3 p^{n+a-1} (1-p)^{m+b-1} \]

- By definition, this is Beta(a + n, b + m)
  - All constant factors combine into a single constant
  - Could just ignore constant factors along the way
Where’d Ya Get Them P(θ)?

- θ is the probability a coin turns up heads
- Model θ with 2 different priors:
  - $P_1(\theta)$ is Beta(3,8) (blue)
  - $P_2(\theta)$ is Beta(7,4) (red)
- They look pretty different!

- Now flip 100 coins; get 58 heads and 42 tails
  - What do posteriors look like?
It’s Like Having Twins

- As long as we collect enough data, posteriors will converge to the true value!
From MLE to Maximum A Posteriori

- Recall Maximum Likelihood Estimator (MLE) of $\theta$
  \[ \theta_{MLE} = \arg \max_{\theta} \prod_{i=1}^{n} f(X_i \mid \theta) \]

- Maximum A Posteriori (MAP) estimator of $\theta$:
  \[
  \theta_{MAP} = \arg \max_{\theta} f(\theta \mid X_1, X_2, \ldots, X_n) = \arg \max_{\theta} \frac{f(X_1, X_2, \ldots, X_n \mid \theta) g(\theta)}{h(X_1, X_2, \ldots, X_n)} \\
  = \arg \max_{\theta} \left( \prod_{i=1}^{n} f(X_i \mid \theta) \right) g(\theta) \\
  = \arg \max_{\theta} g(\theta) \prod_{i=1}^{n} f(X_i \mid \theta)
  \]
  where $g(\theta)$ is prior distribution of $\theta$.

- As before, can often be more convenient to use log:
  \[ \theta_{MAP} = \arg \max_{\theta} \left( \log(g(\theta)) + \sum_{i=1}^{n} \log(f(X_i \mid \theta)) \right) \]

- MAP estimate is the mode of the posterior distribution
Conjugate Distributions Without Tears

• Just for review…
• Have coin with unknown probability \( \theta \) of heads
  - Our prior (subjective) belief is that \( \theta \sim \text{Beta}(a, b) \)
  - Now flip coin \( k = n + m \) times, getting \( n \) heads, \( m \) tails
  - Posterior density: \( (\theta | n \text{ heads, } m \text{ tails}) \sim \text{Beta}(a+n, b+m) \)
    - Beta is conjugate for Bernoulli, Binomial, Geometric, and Negative Binomial
  - \( a \) and \( b \) are called “hyperparameters”
    - Saw \( (a + b – 2) \) imaginary trials, of those \( (a – 1) \) are “successes”
  - For a coin you never flipped before, use \( \text{Beta}(x, x) \) to denote you think coin likely to be fair
    - How strongly you feel coin is fair is a function of \( x \)
Mo’ Beta
Multinomial is Multiple Times the Fun

- Dirichlet($a_1, a_2, ..., a_m$) distribution
  - Conjugate for Multinomial
    - Dirichlet generalizes Beta in same way Multinomial generalizes Bernoulli/Binomial
    - Intuitive understanding of hyperparameters:
      - Saw $\sum_{i=1}^{m} a_i - m$ imaginary trials, with $(a_i - 1)$ of outcome $i$
  - Updating to get the posterior distribution
    - After observing $n_1 + n_2 + ... + n_m$, new trials with $n_i$ of outcome $i$...
    - ... posterior distribution is Dirichlet($a_1 + n_1, a_2 + n_2, ..., a_m + n_m$)
Best Short Film in the Dirichlet Category

- And now a cool animation of Dirichlet($a, a, a$)
  - This is actually log density (but you get the idea…)

Thanks Wikipedia!
Getting Back to your Happy Laplace

• Recall example of 6-sides die rolls:
  
  ▪ $X \sim \text{Multinomial}(p_1, p_2, p_3, p_4, p_5, p_6)$
  
  ▪ Roll $n = 12$ times
  
  ▪ Result: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes
    
    □ MLE: $p_1=3/12, p_2=2/12, p_3=0/12, p_4=3/12, p_5=1/12, p_6=3/12$
  
  ▪ Dirichlet prior allows us to pretend we saw each outcome $k$ times before. MAP estimate:  
    $p_i = \frac{X_i + k}{n + mk}$
    
    □ Laplace’s “law of succession”: idea above with $k = 1$
    
    □ Laplace estimate: $p_i = \frac{X_i + 1}{n + m}$
    
    □ Laplace: $p_1=4/18, p_2=3/18, p_3=1/18, p_4=4/18, p_5=2/18, p_6=4/18$
    
    □ No longer have 0 probability of rolling a three!
Good Times With Gamma

• Gamma(\(\alpha, \lambda\)) distribution
  ▪ Conjugate for Poisson
    ◦ Also conjugate for Exponential, but we won’t delve into that
  ▪ Intuitive understanding of hyperparameters:
    ◦ Saw \(\alpha\) total imaginary events during \(\lambda\) prior time periods
  ▪ Updating to get the posterior distribution
    ◦ After observing \(n\) events during next \(k\) time periods...
    ◦ ... posterior distribution is Gamma(\(\alpha + n, \lambda + k\))
    ◦ Example: Gamma(10, 5)
      ◦ Saw 10 events in 5 time periods. Like observing at rate = 2
      ◦ Now see 11 events in next 2 time periods \(\rightarrow\) Gamma(21, 7)
      ◦ Equivalent to updated rate = 3