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
    - $E[Z | Y] = \frac{1}{2} Y + \frac{1}{2} 2Y = \frac{3}{2} Y$
  - Before opening envelope, think either equally good
    - So, what happened by opening envelope?
  - $E[Z | Y]$ 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 envelop 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)

Revisting Bayes Theorem

- Bayes Theorem ($\theta = model parameters, D = data$):
  - “Prior” $P(\theta)$
  - “Likelihood” $P(D | \theta)$
  - “Posterior” $P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)}$

Computing $P(\theta | D)$

- Bayes Theorem ($\theta = model parameters, D = data$):
  - $P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)}$
  - We have prior $P(\theta)$ and can compute $P(D | \theta)$
  - But how do we calculate $P(D)$?
    - Complicated answer: $P(D) = \int P(D | \theta)P(\theta) d\theta$
    - Easy answer: It is does not depend on $\theta$, so ignore it
      - Just a constant that forces $P(\theta | D)$ to integrate to 1

$P(\theta | D)$ for Beta and Bernoulli

- Prior: $\theta \sim Beta(a, b)$
  - $D = (n$ heads, $m$ tails)
  - $f_{\text{post}}(\theta = p | D) = f_{\text{post}}(\theta = p | D, \theta = p) f_{\text{prior}}(\theta = p)
    = \frac{(n+a)p^n(1-p)^m}{C_n^m} \cdot \frac{(n+b)}{C_n^m} = \frac{(n+a)}{C_n^m} p^n(1-p)^m$
  - 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)?

- 0 is the probability a coin turns up heads
- Model θ with 2 different priors:
  - \( P_1(θ) = \text{Beta}(3, 8) \) (blue)
  - \( P_2(θ) = \text{Beta}(7, 4) \) (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 correct value!

From MLE to Maximum A Posteriori

- Recall Maximum Likelihood Estimator (MLE) of θ:
  \[ \hat{θ}_{\text{MLE}} = \arg \max_θ \prod f(X_i | θ) \]
- Maximum A Posteriori (MAP) estimator of θ:
  \[ \hat{θ}_{\text{MAP}} = \arg \max_θ \frac{\prod f(X_i | θ) g(θ)}{\int \prod f(X_i | θ) g(θ) \, dθ} \]
  where \( g(θ) \) is prior distribution of θ.
- As before, can often be more convenient to use log:
  \[ \hat{θ}_{\text{MAP}} = \arg \max_θ \left( \log g(θ) + \sum \log f(X_i | θ) \right) \]
- MAP estimate is the mode of the posterior distribution

Conjugate Distributions Without Tears

- Just for review…
- Have coin with unknown probability θ of heads
  - Our prior (subjective) belief is that θ ~ Beta(a, b)
  - Now flip coin \( k = n + m \) times, getting \( n \) heads, \( m \) tails
  - Posterior density: \( (θ | n \text{ heads, } m \text{ tails}) \sim \text{Beta}(a + n, b + m) \)
    - 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 Beta(1, 1) to denote you think coin likely to be fair
    - How strongly you feel coin is fair is a function of \( x \)

Mo’ Beta

- Dirichlet(a₁, a₂, ..., aₘ) distribution
  - Conjugate for Multinomial
    - Dirichlet generalizes Beta in same way Multinomial generalizes Bernoulli/Binomial
    \[ f(x₁, x₂, ..., xₘ) = \frac{1}{B(a₁, a₂, ..., aₘ)} \prod x_i^{a_i - 1} \]
  - Intuitive understanding of hyperparameters:
    - Saw \( \sum x_i = m \) 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(a₁ + n₁, a₂ + n₂, ..., aₘ + nₘ)
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…)

Getting Back to your Happy Laplace

- Recall example of 6-sides die rolls:
  - X ~ Multinomial(p₁, p₂, p₃, p₄, p₅, p₆)
  - Roll n = 12 times
  - Result: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes
    - MLE: p₁=3/12, p₂=2/12, p₃=0/12, p₄=3/12, p₅=1/12, p₆=3/12
    - Dirichlet prior allows us to pretend we saw each outcome k times before. MAP estimate:
      - Laplace’s “law of succession”: idea above with k = 1
      - Laplace estimate:
        - p₁=4/18, p₂=3/18, p₃=1/18, p₄=4/18, p₅=2/18, p₆=4/18
        - No longer have 0 probability of rolling a three!

Good Times With Gamma

- Gamma(α, λ) distribution
  - Conjugate for Poisson
    - Also conjugate for Exponential, but we won’t delve into that
  - Intuitive understanding of hyperparameters:
    - Saw α total imaginary events during λ prior time periods
  - Updating to get the posterior distribution
    - After observing n events during next k time periods...
    - ... posterior distribution is Gamma(α+n, λ+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

It’s Normal to Be Normal

- Normal(μ₀, σ₀²) distribution
  - Conjugate for Normal (with unknown μ, known σ²)
  - Intuitive understanding of hyperparameters:
    - A priori, believe true μ distributed ~ N(μ₀, σ₀²)
  - Updating to get the posterior distribution
    - After observing n data points...
      - ... posterior distribution is:
        \[
        N \left( \frac{\mu_0 + \sum_{i=1}^{n} x_i}{\sigma_0^2 + n}, \frac{1}{\sigma_0^2 + n} \left( \frac{\mu_0}{\sigma_0^2} + \frac{n}{\sigma_0^2} \right) \right)
        \]