# DS-GA 1003: Machine Learning and Computational Statistics
# Homework 7: Bayesian Methods

**Due: Tuesday, May 10, 2016, at 6pm (Submit via NYU Classes)**

**Instructions**: Your answers to the questions below, including plots and mathematical work, should be submitted as a single file, either HTML or PDF. You may include your code inline or submit it as a separate file. You may either scan hand-written work or, preferably, write your answers using software that typesets mathematics (e.g. LaTeX, LyX, or MathJax via iPython).

## 1 Introduction

In this homework we work through several basic concepts in Bayesian statistics via one of the simplest problems there is: estimating the probability of heads in a coin flip. Later we'll extend this to the probability of estimating click-through rates in mobile advertising.

## 2 Coin Flipping: Maximum Likelihood

1. Suppose we flip a coin and get the following seqeunce of heads and tails:

$$\mathcal{D} = (H, H, T)$$

Give an expression for the probability of observing $\mathcal{D}$ given that the probability of heads is $\theta$. That is, give an expression for $p(\mathcal{D} \mid \theta)$. This is called the **likelihood of $\theta$ for the data $\mathcal{D}$**.
SOLUTION:

$$p(\mathcal{D} \mid \theta) = \theta^2 (1-\theta).$$

2. How many different sequences of 3 coin tosses have 2 heads and 1 tail? If we coss the coin 3 times, what is the probability of 2 heads and 1 tail? (Answer should be in terms of $\theta$.)
SOLUTION: Let's do the more general case of $n_h$ heads and $n_t$ tails. Let $n = n_h + n_t$ be the total number of coin flips. We simply need to choose $n_h$ slots out of $n$ to have the head flips. There are $\binom{n}{n_h}$ ways to do this. Thus the probability is $\binom{n}{n_h}\theta^{n_h}(1-\theta)^{n_t}$. For the case of 3 coins and 2 heads, there are 3 ways to do it, and the probability is $3\theta^2(1-\theta)$.

3. More generally, give an expression for the likelihood $p(\mathcal{D} \mid \theta)$ for a particular sequence of flips $\mathcal{D}$ that has $n_h$ heads and $n_t$ tails. Make sure you have expressions that make sense even for $\theta = 0$ and $n_h = 0$, and other boundary cases. You may use the convention that $0^0 = 1$, or you can break your expression into cases if needed.
SOLUTION:

$$p(\mathcal{D} \mid \theta) = \theta^{n_h} (1-\theta)^{n_t}$$

4. Prove that the maximum likelihood estimate of $\theta$ given we observed a sequence with $n_h$ heads and $n_t$ tails is

$$\hat{\theta}_{\text{MLE}} = \frac{n_h}{n_h + n_t}.$$

(Hint: Maximizing the log-likelihood is equivalent and is often easier. As usual, make sure everything make sense for the boundary cases, such as data with only heads.)
SOLUTION: The log-likelihood is

$$\ell(\theta) \quad = \quad n_h \log \theta + n_t \log(1-\theta).$$

Differentiating with respect to $\theta$ and equating to zero, we get

$$\begin{aligned} n_h \frac{1}{\theta} - n_t \frac{1}{1-\theta} &= 0 \\ \iff n_h (1-\theta) - n_t \theta &= 0 \\ \iff -\theta (n_t + n_h) + n_h &= 0 \\ \iff \theta &= \frac{n_h}{n_t + n_h}. \end{aligned}$$

Thus any local maximum must occur at $n_h / (n_t + n_h)$. If $n_h, n_t > 0$, then $l(\theta)$ approaches as $\theta \to 0$ and $\theta \to 1$. Since $\ell(\theta)$ is concave, it must attain its maximum at $n_h / (n_t + n_h)$. If $n_h = 0$, then $\ell(\theta)$ is clearly maximized at $\theta = 0$. Similarly, if $n_t = 0$ then the maximum is at $\theta = 1$. Thus $n_h / (n_t + n_h)$ gives the MLE for all values of $n_h$ and $n_t$.

# 3 Coin Flipping: Bayesian Approach with Beta Prior

We'll now take a Bayesian approach to the coin flipping problem, in which we treat $\theta$ as a random variable sampled from some prior distribution $p(\theta)$. We'll represent the $i$th coin flip by a random variable $X_i \in \{0, 1\}$, where $X_i = 1$ if the $i$th flip is heads. We assume that the $X_i$'s are conditionally indendent given $\theta$. This means that the joint distribution of the coin flips and $\theta$ factorizes as follows:

$$\begin{aligned} p(x_1, \ldots, x_n, \theta) &= p(\theta) p(x_1, \ldots, x_n \mid \theta) \text{ (always true)} \\ &= p(\theta) \prod_{i=1}^{n} p(x_i \mid \theta) \text{ (by conditional independence).} \end{aligned}$$

1. Suppose that our prior distribution on $\theta$ is $\text{Beta}(h, t)$, for some $h, t > 0$. That is, $p(\theta) \propto \theta^{h-1} (1-\theta)^{t-1}$. Suppose that our sequence of flips $\mathcal{D}$ has $n_h$ heads and $n_t$ tails. Show that the posterior distribution for $\theta$ is $\text{Beta}(h + n_h, t + n_t)$. That is, show that

$$p(\theta \mid \mathcal{D}) \propto \theta^{h-1+n_h} (1-\theta)^{t-1+n_t}.$$

We say that the Beta distribution is **conjugate** to the Bernoulli distribution since the prior and the posterior are both in the same family of distributions (i.e. both Beta distributions).
SOLUTION:

$$\begin{aligned} p(\theta \mid \mathcal{D}) &\propto p(\theta) p(\mathcal{D} \mid \theta) \\ &= \theta^{h-1} (1-\theta)^{t-1} \theta^{n_h} (1-\theta)^{n_t} \\ &= \theta^{n_h+h-1} (1-\theta)^{n_t+t-1}, \end{aligned}$$

which is the density for the Beta$(h+n_h, t+n_t)$ distribution.

2. Give expressions for the MLE, the MAP, and the posterior mean estimates of $\theta$. [Hint: You may use the fact that a Beta$(h,t)$ distribution has mean $h/(h+t)$ and has mode $(h-1)\,/\,(h+t-2)$ for $h,t>1$. For the Bayesian solutions, you should note that as $h+t$ gets very large, the posterior mean and MAP approach the prior mean $h/\,(n+h)$, while for fixed $h$ and $t$, the posterior mean approaches the MLE when $n_h+n_t\to\infty$.
SOLUTION: As shown above, the MLE is

$$\hat{\theta}_{\text{MLE}} = \frac{n_h}{n_h+n_t}.$$

The MAP estimate is the mode of the posterior distribution, which is

$$\hat{\theta}_{\text{MAP}} = \frac{h+n_h-1}{h+n_h+t+n_t-2},$$

and the posterior mean is

$$\hat{\theta}_{\text{POSTERIOR MEAN}} = \frac{h+n_h}{h+n_h+t+n_t}.$$

3. What happens to $\hat{\theta}_{\text{MLE}}$ , $\hat{\theta}_{\text{MAP}}$, and $\hat{\theta}_{\text{POSTERIOR MEAN}}$ as the number of coin flips approaches infinity?
SOLUTION: They all converge to $\theta$. When we don't have many coin flips, $\hat{\theta}_{\text{MAP}}$ and $\hat{\theta}_{\text{POSTERIOR MEAN}}$ are biased towards an estimate that has high likelihood under our prior. As we get more data, the effect of the prior gradually vanishes.

   (a) The MAP and posterior mean estimators of $\theta$ were derived from a Bayesian perspective. Let's now evaluate them from a frequentist perspective. Suppose $\theta$ is fixed and unknown. Which of the MLE, MAP, and posterior mean estimators give **unbiased** estimates of $\theta$, if any? [Hint: The answer may depend on the parameters $h$ and $t$ of the prior.]
   SOLUTION: The MLE is unbiased in this case. In general, the others are biased. The posterior mean is asymptotically unbiased as $h,t\to 0$. The MAP is unbiased unless $h=t=1$.

4. Suppose somebody gives you a coin and asks you to give an estimate of the probability of heads, but you can only toss the coin 3 times. You have no particular reason to believe this is an unfair coin. Would you prefer the MLE or the posterior mean as a point estimate of $\theta$? If the posterior mean, what would you use for your prior?

# 4 Hierarchical Bayes for Click-Through Rate Estimation

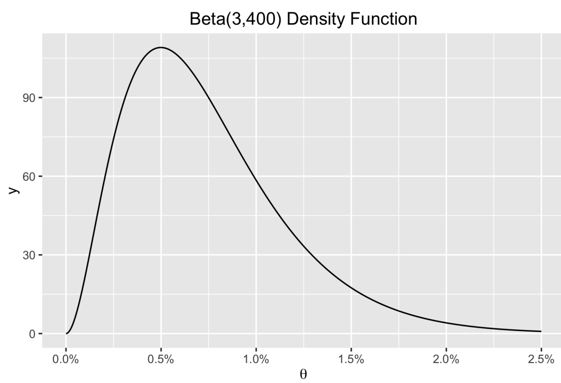
In mobile advertising, ads are often displayed inside apps on a phone or tablet device. When an ad is displayed, this is called an "impression." If the user clicks on the ad, that is called a "click." The probability that an impression leads to a click is called the "click-through rate" (CTR).

Suppose we have $d=1000$ apps. For various reasons, each app tends to have a different overall CTR. For the purposes of designing an ad campaign, we want estimates of all the app-level

CTRs, which we'll denote by $\theta_1, \ldots, \theta_d$. Of course, the particular user seeing the impression and the particular ad that is shown have an effect on the CTR, but we'll ignore these issues for now. [Because so many clicks on mobile ads are accidental, it turns out that the overall app-level CTR often dominates the effect of the particular user and the specific ad.]

If we have enough impressions for a particular app, then the empirical fraction of clicks will give a good estimate for the actual CTR. However, if we have relatively few impressions, we'll have some problems using the empirical fraction. Typical CTRs are less than 1%, so it takes a fairly large number of observations to get a good estimate of CTR. For example, even with 100 impressions, the only possible CTR estimates are $0\%, 1\%, 2\%, \ldots, 100\%$. The 0% estimate is almost certainly much too low, and anything 2% or higher is almost certainly much too high. Our goal is to come up with reasonable point estimates for $\theta_1, \ldots, \theta_{1000}$, even when we have very few observations for some apps.

If we wanted to apply the Bayesian approach worked out in the previous problem, we could come up with a prior that seemed reasonable. For example, we could use the following $\text{Beta}(3, 400)$ as a prior distribution on each $\theta_i$:




In this basic Bayesian approach, the parameters 3 and 400 would be chosen by the data scientist based on prior experience, or "best guess", but without looking at the new data. Another approach would be to use the data to help you choose the parameters $a$ and $b$ in $\text{Beta}(a, b)$. This would **not** be a Bayesian approach, though it is frequently used in practice. One method in this direction is called **empirical Bayes**. Empirical Bayes can be considered a frequentist approach, in which estimate $a$ and $b$ from the data $\mathcal{D}$ using some estimation technique, such as maximum likelihood. The proper Bayesian approach to this type of thing is called **hierarchical Bayes**, in which we put another prior distribution on $a$ and $b$.

### Mathematical Description

We'll now give a mathematical description of our model, assuming the prior parameters $a$ and $b$ are given. Let $n_1, \ldots, n_d$ be the number of impressions we observe for each of the $d$ apps. In this problem, we will not consider these to be random numbers. For the $i$th app, let $c_i^1, \ldots, c_i^{n_i} \in \{0, 1\}$ be indicator variables determining whether or not each impression was clicked. That is, $c_i^j = 1(j\text{th impression on } i\text{th app was clicked})$. We can summarize the data on the $i$th app by $\mathcal{D}_i = (x_i, n_i)$, where $x_i = \sum_{j=1}^{n_i} c_i^j$ is the total number of impressions that were clicked for app $i$. Let $\theta = (\theta_1, \ldots, \theta_d)$, where $\theta_i$ is the CTR for app $i$.

In our Bayesian approach, we act as though the data were generated as follows:

1. Sample $\theta_1, \dots, \theta_d$ i.i.d. from $\text{Beta}(a, b)$.
2. For each app $i$, sample $c_i^1, \dots, c_i^{n_i}$ i.i.d. from $\text{Bernoulli}(\theta_i)$.

### 4.1 Empirical Bayes for a single app

We start by working out some details for Bayesian inference for a single app. That is, suppose we only have the data $\mathcal{D}_i$ from app $i$, and nothing else. Mathematically, this is exactly the same setting as the coin tossing setting above, but here we push it further.

1. Give an expression for $p(\mathcal{D}_i \mid \theta_i)$, the likelihood of $\mathcal{D}_i$ given the probability of click $\theta_i$, in terms of $\theta_i$, $x_i$ and $n_i$.
   SOLUTION:

$$p(\mathcal{D}_i \mid \theta_i) \quad = \quad \theta_i^{x_i} (1 - \theta_i)^{n_i - x_i}$$

2. The probability density for the $\text{Beta}(a, b)$ distribution, evaluated at $\theta_i$, is given by

$$\text{Beta}(\theta_i; a, b) = \frac{1}{B(a, b)} \theta_i^{a-1} (1 - \theta_i)^{b-1}$$

   where $B(a, b)$ is called the Beta function. Explain why we must have

$$\int \theta_i^{a-1} (1 - \theta_i)^{b-1} \, d\theta_i = B(a, b),$$

   and give the full density function for the prior on $\theta_i$, in terms of $a$, $b$, and the normalization function $B$.
   SOLUTION: Since $\text{Beta}(\theta_i; a, b)$ is a density function in $\theta_i$, it must integrate to 1. This explains the integral expression.

3. Give an expression for the posterior distribution $p(\theta_i \mid \mathcal{D}_i)$. In this case, include the constant of proportionality. (In other words, do not use the "is proportional to" sign $\propto$ in your final expression.) [Hint: This problem is essentially a repetition of an earlier problem.]
   SOLUTION: (Same calculation as in earlier problem)

$$\begin{aligned} p(\theta_i \mid \mathcal{D}_i) &\propto p(\theta_i) p(\mathcal{D}_i \mid \theta_i) \\ &= \frac{1}{B(a, b)} \theta_i^{a-1} (1 - \theta_i)^{b-1} \theta_i^{x_i} (1 - \theta_i)^{n_i - x_i} \\ &\propto \theta_i^{a + x_i - 1} (1 - \theta_i)^{b + n_i - x_i - 1}. \end{aligned}$$

   Since this is proportional to the $\text{Beta}(a + x_i, b + n_i - x_i)$ density over $\theta_i$, and $p(\theta_i \mid \mathcal{D}_i)$ is density on $\theta_i$, we must actually have

$$\begin{aligned} p(\theta_i \mid \mathcal{D}_i) &= \text{Beta}(\theta_i; a + x_i, b + n_i - x_i) \\ &= \frac{\theta^{a + x_i - 1} (1 - \theta)^{b + n_i - x_i - 1}}{B(a + x_i, b + n_i - x_i)} \end{aligned}$$

4. Give a closed form expression for $p(\mathcal{D}_i)$, the marginal likelihood of $\mathcal{D}_i$, in terms of the $a, b, x_i$, and $n_i$. You may use the normalization function $B(\cdot,\cdot)$ for convenience, but you should not have any integrals in your solution. (Hint: $p(\mathcal{D}_i) = \int p\left(\mathcal{D}_i \mid \theta_i\right) p(\theta_i)\, d\theta_i$, and the answer will be a ratio of two beta function evaluations.)
SOLUTION: We have

$$\begin{aligned} p(\mathcal{D}_i) &= \int p\left(\mathcal{D}_i \mid \theta_i\right) p(\theta_i)\, d\theta_i \\ &= \int \frac{1}{B(a,b)} \theta_i^{a-1} \left(1-\theta_i\right)^{b-1} \theta_i^{x_i} \left(1-\theta_i\right)^{n_i - x_i}\, d\theta_i \\ &= \frac{B(a+x_i, b+n_i-x_i)}{B(a,b)}. \end{aligned}$$

5. The maximum likelihood estimate for $\theta_i$ is $x_i/n_i$. Let $p_{\mathrm{MLE}}(\mathcal{D}_i)$ be the marginal likelihood of $\mathcal{D}_i$ when we use a prior on $\theta_i$ that puts all of its probability mass at $x_i/n_i$. Note that

$$\begin{aligned} p_{\mathrm{MLE}}(\mathcal{D}_i) &= p\left(\mathcal{D}_i \mid \theta_i = \frac{x_i}{n_i}\right) p\left(\theta_i = \frac{x_i}{n_i}\right) \\ &= p\left(\mathcal{D}_i \mid \theta_i = \frac{x_i}{n_i}\right). \end{aligned}$$

Explain why, or prove, that $p_{\mathrm{MLE}}(\mathcal{D}_i)$ is larger than $p(\mathcal{D}_i)$ for any other prior we might put on $\theta_i$. If it's too hard to reason about all possible priors, it's fine to just consider all Beta priors. [Hint: This does not require much or any calculation. It may help to think about the integral $p(\mathcal{D}_i) = \int p\left(\mathcal{D}_i \mid \theta_i\right) p(\theta_i)\, d\theta_i$ as a weighted average of $p(\mathcal{D}_i \mid \theta_i)$ for different values of $\theta_i$, where the weights are $p(\theta_i)$.]
SOLUTION: Explanation is that $p(\mathcal{D}_i) = \int p\left(\mathcal{D}_i \mid \theta_i\right) p(\theta_i)\, d\theta_i$ is essentially taking a weighted average over values of $p(\mathcal{D}_i \mid \theta_i)$ over $\theta_i \in [0,1]$, where the weight is determined by $p(\theta_i)$. Since $p(\mathcal{D}_i \mid \theta_i)$ is maximized at $x_i/n_i$, the averaging can only be worse than putting all the weight at $x_i/n_i$. It's a simple proof, but it requires Lebesgue integration to account for any possible distribution. Let $P$ be any probability distribution on $[0,1]$. Then the marginal probability of $\mathcal{D}_i$ for prior $P$ is given by

$$\begin{aligned} p(\mathcal{D}_i) &= \int p\left(\mathcal{D}_i \mid \theta_i\right)\, dP(\theta_i) \\ &\leq \int p_{\mathrm{MLE}}(\mathcal{D}_i)\, dP(\theta_i) \\ &= p_{\mathrm{MLE}}(\mathcal{D}_i) \int 1\, dP(\theta_i) \\ &= p_{\mathrm{MLE}}(\mathcal{D}_i). \end{aligned}$$

6. One approach to getting an **empirical Bayes** estimate of the parameters $a$ and $b$ is to use maximum likelihood. Such an empirical Bayes estimate is often called an **ML-2** estimate, since it's maximum likelihood, but at a higher level in the Bayesian hierarchy. To emphasize the dependence of the likelihood of $\mathcal{D}_i$ on the parameters $a$ and $b$, we'll now write it as

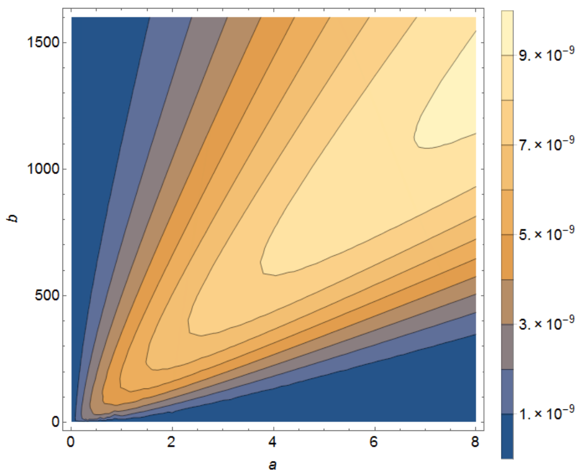



Figure 1: A plot of $p(\mathcal{D}_i \mid a, b)$ as a function of $a$ and $b$.

$p(\mathcal{D}_i \mid a, b)$[1]. The empirical Bayes estimates for $a$ and $b$ are given by

$$(\hat{a}, \hat{b}) = \underset{(a,b) \in \mathbf{R}^{>0} \times \mathbf{R}^{>0}}{\arg\max}\, p(\mathcal{D}_i \mid a, b).$$

To make things concrete, suppose we observed $x_i = 3$ clicks out of $n_i = 500$ impressions. A plot of $p(\mathcal{D}_i \mid a, b)$ as a function of $a$ and $b$ is given in Figure 1. It appears from this plot that the likelihood will keep increasing as $a$ and $b$ increase, at least if $a$ and $b$ maintain a particular ratio. Indeed, this likelihood function never attains its maximum, so we cannot use ML-2 here. Explain what's happening to the prior as we continue to increase the likelihood. [Hint: It is a property of the Beta distribution (not difficult to see), that for any $\theta \in (0, 1)$, there is a Beta distribution with expected value $\theta$ and variance less than $\varepsilon$, for any $\varepsilon > 0$. What's going in here is similar to what happens when you attempt to fit a gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ to a single data point using maximum likelihood.]
SOLUTION: Again, everything comes down to the fact that

$$p(\mathcal{D}_i) = \int p\left(\mathcal{D}_i \mid \theta_i\right)\, dP(\theta_i)$$

and $p(\mathcal{D}_i \mid \theta_i = x_i/n_i) \geq p(\mathcal{D}_i \mid \theta)$ for all $\theta \in [0, 1]$. The more the prior distribution $P$ is concentrated around the MLE, the larger $p(\mathcal{D}_i)$ will get. Since we can take a sequence of beta priors with expected value $x_i/n_i$ and variance going to 0, we can always increase the likelihood.

## 4.2 Empirical Bayes Using All App Data

In the previous section, we considered working with data from a single app. With a fixed prior, such as Beta(3,400), our Bayesian estimates for $\theta_i$ seem more reasonable (to me, the person who

[1]Note that this is a slight (though common) abuse of notation, because $a$ and $b$ are not random variables in this setting. It might be more appropriate to write this as $p(\mathcal{D}_i; a, b)$ or $p_{a,b}(\mathcal{D}_i)$. But this isn't very common.

chose the prior) than the MLE when our sample size $n_i$ is small. The fact that these estimates seem reasonable is an immediate consequence of the fact that I chose the prior to give high probability to estimates that seem reasonable to me, before ever seeing the data. Our attempt to use empirical Bayes (ML-2) to choose the prior in a data-driven way was not successful. With only a single app, we were essentially overfitting the prior to the data we have. In this section, we'll consider using the data from all the apps, in which case empirical Bayes makes more sense.

1. Let $\mathcal{D} = (\mathcal{D}_1, \ldots, \mathcal{D}_d)$ be the data from all the apps. Give an expression for $p(\mathcal{D} \mid a, b)$, the **marginal likelihood** of $\mathcal{D}$. Expression should be in terms of $a, b, x_i, n_i$ for $i = 1, \ldots, d$. (Hint: This problem should be easy, based on a problem from the previous section.)
SOLUTION: Since $\mathcal{D}_1, \ldots, \mathcal{D}_d$ are i.i.d., we have

$$\begin{aligned} p(\mathcal{D}) &= \prod_{i=1}^{d} p(\mathcal{D}_i) \\ &= \prod_{i=1}^{d} \frac{B(a + x_i, b + n_i - x_i)}{B(a, b)} \end{aligned}$$

2. Explain why $p(\theta_i \mid \mathcal{D}) = p(\theta_i \mid \mathcal{D}_i)$, according to our model. In other words, once we choose values for parameters $a$ and $b$, information about one app does not give any information about other apps.
SOLUTION: Each $\mathcal{D}_i$ depends only on $\theta_i$, and the $\theta_i$'s are independent. So $D_j$ gives no information about $\mathcal{D}_i$ for $j \neq i$. That's sufficient. But a mathy proof would be the following: Write $\mathcal{D}_{-i}$ for the data from all apps except the $i$'th. Then

$$\begin{aligned} p(\theta_i \mid \mathcal{D}) &= p(\theta_i, \mathcal{D})/p(\mathcal{D}) \\ &= p(\theta_i \mathcal{D}_i, \mathcal{D}_{-i})/p(\mathcal{D}) \\ &= p(\theta_i, \mathcal{D}_i) p(\mathcal{D}_{-i})/p(\mathcal{D} \text{ (by independence assumptions)} \\ &\propto p(\theta_i, \mathcal{D}_i) \text{ (where constant of proportionality doesn't depend on } \theta_i) \\ &\propto p\left(\theta_i \mid \mathcal{D}_i\right) \text{ (where constant of proportionality doesn't depend on } \theta_i). \end{aligned}$$

Since two distributions on $\theta_i$ are proportional, they must actually be equal. So $p(\theta_i \mid \mathcal{D}) = p(\theta_i \mid \mathcal{D}_i)$.

3. [Optional] Suppose we have data from 6 apps. 3 of the apps have a fair number of impressions, and 3 have relatively few. Suppose we observe the following:

| | Num Clicks | Num Impressions |
|---|---|---|
| App 1 | 50 | 10000 |
| App 2 | 160 | 20000 |
| App 3 | 180 | 60000 |
| App 4 | 0 | 100 |
| App 5 | 0 | 5 |
| App 6 | 1 | 2 |

Compute the empirical Bayes estimates for $a$ and $b$. The empirical Bayes prior is then $\text{Beta}(\hat{a}, \hat{b})$, where $\hat{a}$ and $\hat{b}$ are our estimates. What are the corresponding prior mean and standard deviation? [Hint: You're encouraged to use a general purpose function optimization

routine that does not require a gradient.]
SOLUTION: My calculations give $\hat{a} = 6.47$ and $\hat{b} = 1181.4$. Prior mean is .54% and prior SD is .21%.

4. [Optional] Complete the following table:

| | NumClicks | NumImpressions | MLE | MAP | PosteriorMean | PosteriorSD |
|---|---|---|---|---|---|---|
| App 1 | 50 | 10000 | 0.5% | | | |
| App 2 | 160 | 20000 | 0.8% | | | |
| App 3 | 180 | 60000 | 0.3% | | | |
| App 4 | 0 | 100 | 0% | | | |
| App 5 | 0 | 5 | 0% | | | |
| App 6 | 1 | 2 | 50% | | | |

SOLUTION:

| | NumClicks | NumImpressions | MLE | MAP | PosteriorMean | PosteriorSD |
|---|---|---|---|---|---|---|
| 1 | 50 | 10000 | 0.5% | 0.5% | 0.5% | 0.07% |
| 2 | 160 | 20000 | 0.8% | 0.78% | 0.79% | 0.06% |
| 3 | 180 | 60000 | 0.3% | 0.3% | 0.3% | 0.02% |
| 4 | 0 | 100 | 0% | 0.43% | 0.5% | 0.2% |
| 5 | 0 | 5 | 0% | 0.46% | 0.54% | 0.21% |
| 6 | 1 | 2 | 50% | 0.54% | 0.63% | 0.23% |

### 4.3 Hierarchical Bayes

In Section 4.2 we managed to get empirical Bayes ML-II estimates for $a$ and $b$ by assuming we had data from multiple apps. However, we didn't really address the issue that ML-II, as a maximum likelihood method, is prone to overfitting if we don't have enough data (in this case, enough apps). Moreover, a true Bayesian would reject this approach, since we're using our data to determine our prior. If we don't have enough confidence to choose parameters for $a$ and $b$ without looking at the data, then the only proper Bayesian approach is to put another prior on the parameters $a$ and $b$. If you are very uncertain about values for $a$ and $b$, you could put priors on them that have high variance.

1. [Optional] Suppose $P$ is the Beta$(a, b)$ distribution. Conceptually, rather than putting priors on $a$ and $b$, it easier to reason about priors on the mean $m$ and the variance $v$ of $P$. If we parameterize $P$ by its mean $m$ and the variance $v$, give an expression for the density function Beta$(\theta; m, v)$. You are free to use the internet to get this expression. [Hint: You may find it convenient to write some expression in terms of $\eta = a + b$.]
SOLUTION: We have these simultaneous equations:

$$\begin{aligned} v &= \frac{ab}{(a+b)^2 (a+b+1)} \\ m &= \frac{a}{a+b} \end{aligned}$$

Noting that $1-m=b/(a+b)$, and letting $\eta=a+b$, we see that

$$\nu \quad = \quad \frac{m(1-m)}{\eta+1}$$

and so

$$\eta=a+b=\frac{m(1-m)}{v}-1.$$

We also get

$$\begin{aligned}
v &= \left(\frac{a}{a+b}\right)\left(\frac{b}{a+b}\right)\left(\frac{1}{a+b+1}\right) \\
&= \frac{m\,(1-m)}{\frac{a}{m}+1} \\
\implies a &= m\left[\frac{m(1-m)}{v}-1\right]=m\eta \\
\implies b &= \eta-a=(1-m)\left[\frac{m(1-m)}{v}-1\right]=(1-m)\,\eta.
\end{aligned}$$

So

$$\mathrm{Beta}(\theta;m,v)=\frac{1}{B(m\eta,(1-m)\eta)}\theta_i^{m\eta-1}\,(1-\theta_i)^{(1-m)\eta-1}\,,$$

where $\eta=a+b=\frac{m(1-m)}{v}-1$.

2. [Optional] Suggest a prior distribution to put on $m$ and $v$. [Hint: You might want to use one of the distribution families given in this lecture: `https://davidrosenberg.github.io/mlcourse/Lectures/10b.conditional-probability-models.pdf#page=7`.]
SOLUTION: Since $m\in(0,1)$, a Beta prior is most obvious. Since $v\in(0,\infty)$, a Gamma distribution would work.

3. [Optional] Once we have our prior on $m$ and $v$, we can go "full Bayesian" and compute posterior distributions on $\theta_1,\ldots,\theta_d$. However, these no longer have closed forms. We would have to use approximation techniques, typically either a Monte Carlo sampling approach or a variational method, which are beyond the scope of this course[2]. After observing the data $\mathcal{D}$, $m$ and $v$ will have some posterior distribution $p(m,v\mid\mathcal{D})$. We can approximate that distribution by a point mass at the mode of that distribution $(m_{\mathrm{MAP}},v_{\mathrm{MAP}})=\arg\max_{m,v}p(m,v\mid\mathcal{D})$. **Give expressions** for the posterior distribution $p(\theta_1,\ldots,\theta_d\mid\mathcal{D})$, with and without this approximation. You do not need to give any explicit expressions here. It's fine to have expressions like $p(\theta_1,\ldots,\theta_d\mid m,v)$ in your solution. Without the approximation, you will probably need some integrals. It's these integrals that we need sampling or variational approaches to approximate. While one can see this approach as a way to approximate the proper Bayesian approach, one could also be skeptical and say this is just another way to determine your prior from the data. The estimators $(m_{\mathrm{MAP}},v_{\mathrm{MAP}})$ are often called **MAP-II estimators**, since

[2] If you're very ambitious, you could try out a package like PyStan (`https://pystan.readthedocs.io/en/latest/`) to see what happens.

they are MAP estimators at a higher level of the Bayesian hierarchy.
SOLUTION: Without the approximation, it's

$$p(\theta_1, \ldots, \theta_d \mid \mathcal{D}) \quad = \quad \int_0^\infty \int_0^1 p(\theta_1, \ldots, \theta_d \mid m, v) p(m, v \mid \mathcal{D}) \, dm \, dv$$

and with the approximation, it's

$$p(\theta_1, \ldots, \theta_d \mid \mathcal{D}) \approx p(\theta_1, \ldots, \theta_d \mid m_{\mathrm{MAP}}, v_{\mathrm{MAP}}).$$