

# STAT 447 Assignment 5

Caden Hewlett

2024-02-09

## Question 1: Sequential Updating

Consider a joint probabilistic model given by

$$\theta \sim \rho, \text{ and } (x_i | \theta) \stackrel{iid}{\sim} \nu_\theta, \text{ where } i \in \{1, 2, \dots, n\}$$

where  $\rho$  is a prior distribution for the unknown parameter  $\theta$ , and  $\{x_i\}_{i=1}^n$  is a sequence of observations with conditional distribution  $\nu_\theta$ .

### Part 1

Write down the posterior distribution of  $\theta$  given  $\{x_i\}_{i=1}^n$ .

In a more verbose sense, let  $\Theta$  be the random variable for the unknown parameter, and  $\theta$  be the realization of this random variable under the proposed prior distribution. We can then write the following (*purely for nomenclature reasons*)

$\rho = p_\Theta(\theta)$ , where  $p_\Theta(\theta)$  is the PMF/PDF given by  $\rho$

$P(X = x | \theta) = p_{X|\Theta}(x, \theta)$ , where  $p_{X|\Theta}(x, \theta)$  is the PMF given by  $\nu_\theta$

With this in mind, we can write the posterior of  $\theta$  given  $\{x_i\}_{i=1}^n$ , where we describe the event that  $\{X_i\}_{i=1}^n = \{x_i\}_{i=1}^n$  using Bayes' Rule.

We will begin by using the most verbose notation possible, for complete clarity.

$$P(\Theta = \theta | \{X_i\}_{i=1}^n = \{x_i\}_{i=1}^n) = \frac{p_\Theta(\theta)P(\{X_i\}_{i=1}^n = \{x_i\}_{i=1}^n | \Theta = \theta)}{P(\{X_i\}_{i=1}^n = \{x_i\}_{i=1}^n)}$$

We start by considering the joint likelihood function.

$$P(\{X_i\}_{i=1}^n = \{x_i\}_{i=1}^n | \Theta = \theta) = P((X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_1 = x_1) | \Theta = \theta)$$

Then, using intersections, we can write the likelihood as:

$$P(\{X_i\}_{i=1}^n = \{x_i\}_{i=1}^n | \Theta = \theta) = P\left(\bigcap_{i=1}^n (X_i = x_i) | \Theta = \theta\right)$$

Now, we use the following property of *iid* random variables

$$\text{I.I.D} \implies \forall (i \neq j) \in [1, n], (X_i | \Theta) \perp (X_j | \Theta) \implies \forall (i \neq j) \in [1, n], P(X_i \cap X_j | \Theta) = P(X_i | \Theta)P(X_j | \Theta)$$

Hence, we can write the likelihood as:

$$P(\{X_i\}_{i=1}^n = \{x_i\}_{i=1}^n \mid \Theta = \theta) = \prod_{i=1}^n \left( P(X_i = x_i \mid \Theta = \theta) \right) = \prod_{i=1}^n (\nu_\theta) = (\nu_\theta)^n$$

Which, as you can see, simplifies nicely to  $(\nu_\theta)^n$ .

Now, our expression simplifies a little bit to the following:

$$P(\Theta = \theta \mid \{X_i\}_{i=1}^n = \{x_i\}_{i=1}^n) = \frac{p_\Theta(\theta) \nu_\theta^n}{P(\{X_i\}_{i=1}^n = \{x_i\}_{i=1}^n)}$$

If we wished to, we could write the following proportionality directly to conclude:

$$p_{\Theta|X_{1:n}}(\theta, \{x_i\}_{i=1}^n) = \pi_n \propto \rho \cdot \nu_\theta^n$$

Or, letting normalizing constant  $\mathcal{Z}_{1:n} = P(\{X_i\}_{i=1}^n = \{x_i\}_{i=1}^n)$ , we can write:

$$\pi_n(\theta) = \frac{\rho \cdot \nu_\theta^n}{P(\{X_i\}_{i=1}^n = \{x_i\}_{i=1}^n)} = (\mathcal{Z}_{1:n}^{-1}) \cdot \rho \cdot \nu_\theta^n$$

Giving us a nice expression for the posterior, both in proportionality and equality according to a normalizing constant.

## Part 2

Suppose now we get an additional data point  $x_{n+1}$  with the same conditional distribution  $\nu_\theta$ . Show that using the posterior from part 1 as the *prior* and data equal to just  $x_{n+1}$  gives the same posterior distribution as redoing part 1 with the  $n + 1$  data points.

We wish to show that:

$$\pi_{(n+1)}(\theta) = \frac{p_\Theta(\theta) P(\{X_i\}_{i=1}^{n+1} = \{x_i\}_{i=1}^{n+1} \mid \Theta = \theta)}{P(\{X_i\}_{i=1}^{n+1} = \{x_i\}_{i=1}^{n+1})} = \frac{\pi_n(\theta) P(X_{n+1} = x_{n+1} \mid \Theta = \theta)}{P(X_{n+1} = x_{n+1})}$$

We'll evaluate each expression in turn.

With the first term, we can say directly that:

$$\text{LHS} = \frac{p_\Theta(\theta) P(\{X_i\}_{i=1}^{n+1} = \{x_i\}_{i=1}^{n+1} \mid \Theta = \theta)}{P(\{X_i\}_{i=1}^{n+1} = \{x_i\}_{i=1}^{n+1})} = (\mathcal{Z}_{1:(n+1)}^{-1}) \cdot \rho \cdot \nu_\theta^{n+1} \propto \rho \cdot \nu_\theta^{n+1}$$

By the exact process used in **Part 1**.

More interestingly, we can expand the second term to as follows:

$$\pi_{(n+1)}(\theta) = \frac{\pi_n(\theta) P(X_{n+1} = x_{n+1} \mid \Theta = \theta)}{P(X_{n+1} = x_{n+1})} = (\mathcal{Z}_{n+1}^{-1}) \pi_n(\theta) \nu_\theta$$

Then, substituting our expression for  $\pi_n(\theta)$  from **Part 1**:

$$\text{RHS} = (\mathcal{Z}_{n+1}^{-1}) \left( (\mathcal{Z}_{1:n}^{-1}) \cdot \rho \cdot \nu_\theta^n \right) \nu_\theta = (\mathcal{Z}_{n+1}^{-1} \cdot \mathcal{Z}_{1:n}^{-1}) \rho (\nu^n \cdot \nu^1) = (\mathcal{Z}_{n+1} \cdot \mathcal{Z}_{1:n})^{-1} \rho \cdot \nu_\theta^{n+1} \propto \rho \cdot \nu_\theta^{n+1}$$

Now, consider the following Lemma which we will use without proof.

### Lemma

Let  $f(y)$  and  $g(y)$  be probability distributions acting on the same space  $y \in Y$ . Let

$$\exists c \in \mathbb{R}^+ \text{ s.t. } \forall y \in Y, f(y) = c \cdot g(y) \implies f \text{ and } g \text{ describe equivalent distributions.}$$

In our specific case, we have:

$$(\mathcal{Z}_{1:(n+1)})^{-1} \text{LHS} = (\mathcal{Z}_{n+1} \cdot \mathcal{Z}_{1:n})^{-1} \text{RHS}$$

So, we can write:

$$\text{Let } c = \frac{(\mathcal{Z}_{n+1} \cdot \mathcal{Z}_{1:n})^{-1}}{(\mathcal{Z}_{1:(n+1)})^{-1}} = \frac{\mathcal{Z}_{1:(n+1)}}{\mathcal{Z}_{n+1} \cdot \mathcal{Z}_{1:n}}, \text{ then LHS} = c \cdot \text{RHS}$$

Which would imply that the two are equivalent under proportionality, and, assuming  $c \in \mathbb{R}^+$ , are equivalent under normalization.

## Question 2: Bayesian Inference in the Limit of Increasing Data

We will use the tractability of the coin bag example to explore the behavior of the posterior distribution as the number of observations goes to infinity. Recall that its joint distribution is

$$p \sim \text{discrete}(\{0, 1/K, 2/K, \dots, 1\}, \rho)$$

$$y_i | p \stackrel{\text{iid}}{\sim} \text{bern}(p), \quad i \in \{1, \dots, n\}$$

Where  $\rho = (\rho_0, \rho_1, \dots, \rho_K)$  is the prior, where  $\forall k \in [1, K]$  the proportion of coins of type  $k$  in the bag is  $\rho_k$ . For the sake of notation (and to avoid confusion with capital  $P$ ), we will let  $p_{\text{obs}}$  be a realization of the random variable  $p$ .

### Part 1

The following simulates the posterior for the above equation.

```
posterior_distribution = function(rho, n_heads, n_observations) {
  K = length(rho) - 1
  gamma = rho * dbinom(n_heads, n_observations, (0:K)/K)
  normalizing_constant = sum(gamma)
  gamma/normalizing_constant
}
```

Note, we can verify that passing the following gives our familiar 1/17 result from Assignment 1.

```
assign_1 = posterior_distribution(c(0, 0.5, 1), 3, 3)
# check
all.equal(1/17, assign_1[2], tolerance = 2e-10)
```

```
## [1] TRUE
```

### Part 2

Write a function `posterior_mean` that computes the posterior mean given the output of `posterior_distribution`.

We recall that the posterior mean is computed as follows:

$$\mathbb{E}(p \mid Y = y) = \sum_{\{p: \pi(p) > 0\}} p \pi(p)$$

From our “original” coin flip example, we’d have  $K = 2$  and

$$\mathbb{E}(p \mid Y = y) = \left(\frac{1}{K}\right)\pi(1) + \left(\frac{2}{K}\right)\pi(2) = \left(\frac{1}{2}\right) \cdot \frac{1}{17} + \left(\frac{2}{2}\right)\frac{16}{17} = \frac{33}{34}$$

Which we will use to test our function.

```
posterior_mean <- function(pi){  
  K = length(pi)-1  
  return(sum((0:K)/K*pi))  
}  
all.equal(posterior_mean(assign_1), 33/34)
```

```
## [1] TRUE
```

Further, we'd expect a posterior mean of 1 in the case where all the weight was put on the 1 ("always heads") coin.

```
posterior_mean(posterior_distribution(c(0,0,1), 5, 5))
```

```
## [1] 1
```

### Part 3

Write another function called `simulate_posterior_mean_error`.

It will perform the following tasks:

a.) Simulate  $p_{\text{true}} \sim \text{discrete}(\{0, 1/K, \dots, K/K\}, \rho_{\text{true}})$ , where  $\rho_{\text{true}}$  is supplied to the function.

This is done as follows (using the Assignment 1 coin flip example)

```
rho_true = c(0, 0.5, 1)/sum(c(0, 0.5, 1)); K = length(rho_true)-1  
p_true <- DiscreteDistribution(supp = (0:K)/K, rho_true)  
p_true_obs = simulate(p_true)  
p_true_obs
```

```
## [1] 1
```

b.) Simulates  $\{y_i\}_{i=1}^{n_{\text{obs}}} = \{y_1, y_2, \dots, y_{n_{\text{obs}}}\} = y_{1:n_{\text{obs}}}$  conditional on the simulated  $p_{\text{true}}$ . We recall that  $\forall i \in [1, n_{\text{obs}}], (y_i \mid p_{\text{true}}) \sim \text{bern}(p_{\text{true}})$ , where it should be noted that  $\$n_{\text{obs}} = \$n_{\text{observations}}$  is supplied to the function.

```
n_observations = 3  
n = n_observations  
y_obs = rbern(n, p_true_obs)  
y_obs
```

```
## [1] 1 1 1
```

c.) Calls `posterior_distribution` using and the simulated data. Note that  $\rho_{\text{prior}}$  is the final parameter to the function.

```
# in the coin flip, we assumed all were equally likely  
rho_prior = c(0, 1/2, 1)  
pi_obs = posterior_distribution(  
  rho_prior, sum(y_obs), 3  
)  
pi_obs
```

```
## [1] 0.00000000 0.05882353 0.94117647
```

d.) Computes the posterior mean  $\mathbb{E}(p \mid y_{\text{obs}})$ , given the observed posterior  $\pi_{\text{obs}}(p)$ .

```
post_mean = posterior_mean(pi_obs)
post_mean
```

```
## [1] 0.9705882
```

e.) Compute the absolute error  $\varepsilon_{\text{obs}}$

$$\varepsilon_{\text{obs}} = |p_{\text{true}} - \mathbb{E}(p \mid y_{1:n_{\text{obs}}})|$$

.

```
epsilon_obs = abs(p_true_obs - post_mean)
epsilon_obs
```

```
## [1] 0.02941176
```

Then we combine this all into one process

```
simulate_posterior_mean_error = function(rho_true, rho_prior, n_observations){
  # step 1
  p_true = DiscreteDistribution(supp = (0:K)/K, rho_true)
  p_true_obs = simulate(p_true)
  # step 2
  n = n_observations
  y_obs = rbern(n, p_true_obs)
  # step 3
  pi_obs = posterior_distribution(rho_prior, sum(y_obs), n)
  # step 4
  epsilon_obs = abs(p_true_obs - post_mean)
  # returns error
  return(epsilon_obs)
}
```