CSE 515T (Spring 2015) Midterm solutions

1. Consider two coins with unknown bias θ_1 and θ_2, respectively. We place independent, identical beta priors on these quantities:

$$p(\theta_1) = \mathcal{B}(\theta_1; 2, 2); \qquad p(\theta_2) = \mathcal{B}(\theta_2; 2, 2).$$

Imagine someone flips both coins and tells you that *exactly one* of the outcomes (but not which) was a "head." Thus the observation was either HT or TH, but you are not told which. The below expressions are conditioned on "H" to indicate this observation.

- Give an expression for the posterior of the first coin's bias given this observation, $p(\theta_1 \mid \mathrm{H})$. Simplify the result as much as you can. Plot the prior and the posterior for θ_1 over the interval $\theta_1 \in (0, 1)$.
- Give an expression for the joint posterior $p(\theta_1, \theta_2 \mid \mathrm{H})$. Plot the joint prior, the likelihood, and the joint posterior as three separate heat maps over the unit square $(\theta_1, \theta_2) \in (0, 1)^2$. Use a grid with at least 100 values along each of the two θ axes.
- Summarize what the observation taught us about the bias of the coins.

Solution

The outcome of the unseen experiment was either HT or TH. These are mutually exhaustive and independent events, so we may use the sum rule to derive the desired posterior:

$$p(\theta_1 \mid \mathrm{H}) = \Pr(\mathrm{HT})p(\theta_1 \mid \mathrm{HT}) + \Pr(\mathrm{TH})p(\theta_1 \mid \mathrm{TH}).$$

Notice that both $p(\theta_1 \mid \mathrm{HT})$ and $p(\theta_1 \mid \mathrm{TH})$ can be computed explicitly as updated beta distributions, now that the coins in the outcomes have been identified:

$$\begin{aligned} p(\theta_1 \mid \mathrm{HT}) &= \mathcal{B}(\theta_1; 3, 2) \\ p(\theta_1 \mid \mathrm{TH}) &= \mathcal{B}(\theta_1; 2, 3). \end{aligned}$$

What is $\Pr(\mathrm{HT})$? It can be calculated explicitly, but a simpler approach is to appeal to symmetry between the coins to conclude $\Pr(\mathrm{HT}) = \Pr(\mathrm{TH}) = 1/2$. Thus

$$p(\theta_1 \mid \mathrm{H}) = \tfrac{1}{2}\big(p(\theta_1 \mid \mathrm{HT}) + p(\theta_1 \mid \mathrm{TH})\big) = \tfrac{1}{2}\big(\mathcal{B}(\theta_1; 3, 2) + \mathcal{B}(\theta_1; 2, 3)\big).$$

We can simply this expression further. The posterior is proportional to

$$p(\theta_1 \mid \mathrm{H}) \propto \theta_1^2(1 - \theta_1) + \theta_1(1 - \theta_1)^2 = \theta_1(1 - \theta_1) \propto \mathcal{B}(\theta_1; 2, 2).$$

Therefore the distribution of θ_1 has not changed given our observation! The prior (and posterior!) for θ_1 is plotted in Figure 1.

There are two ways to compute the joint posterior over (θ_1, θ_2): the easy way and the hard way. The easy way is to use the sum rule again to write

$$\begin{aligned} p(\theta_1, \theta_2 \mid \mathrm{H}) &= \Pr(\mathrm{HT})p(\theta_1, \theta_2 \mid \mathrm{HT}) + \Pr(\mathrm{TH})p(\theta_1, \theta_2 \mid \mathrm{TH}) \\ &= \tfrac{1}{2}\mathcal{B}(\theta_1; 3, 2)\mathcal{B}(\theta_2; 2, 3) + \tfrac{1}{2}\mathcal{B}(\theta_1; 2, 3)\mathcal{B}(\theta_2; 3, 2). \end{aligned}$$

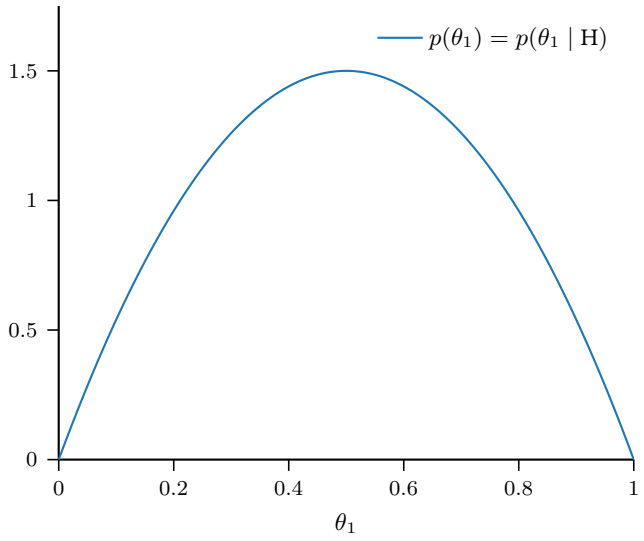

Figure 1: The prior and posterior (given the observation H) of θ_1.

If we go with the hard way, we begin by computing the joint prior. By independence, we have:

$$p(\theta_1, \theta_2) = \mathcal{B}(\theta_1; 2, 2)\mathcal{B}(\theta_2; 2, 2).$$

To derive the likelihood, we again note that our observation could have been generated by two mutually independent events: HT or TH. Given θ_1 and θ_2, the total probability of these events is:

$$\Pr(\mathrm{H} \mid \theta_1, \theta_2) = \theta_1(1 - \theta_2) + (1 - \theta_1)\theta_2;$$

the first term accounts for HT and the second term for TH. The posterior is now:

$$\begin{aligned} p(\theta_1, \theta_2 \mid \mathrm{H}) &= \tfrac{1}{Z} \Pr(\mathrm{H} \mid \theta_1, \theta_2) p(\theta_1, \theta_2) \\ &= \tfrac{1}{Z} \big(\theta_1(1 - \theta_2) + (1 - \theta_1)\theta_2\big) \mathcal{B}(\theta_1; 2, 2)\mathcal{B}(\theta_2; 2, 2). \end{aligned}$$

The normalization constant is

$$Z = \Pr(\mathrm{H}) = \int_0^1 \int_0^1 \big(\theta_1(1 - \theta_2) + (1 - \theta_1)\theta_2\big) \mathcal{B}(\theta_1; 2, 2)\mathcal{B}(\theta_2; 2, 2) \, \mathrm{d}\theta_1 \, \mathrm{d}\theta_2.$$

This integral is tractable and equals ½.[1] In fact, this is always true for any arbitrary mean-½ beta priors on θ_1, θ_2: if our best guess is that each coin is unbiased, then the outcomes HT/TH always have equal combined probability as the outcomes HH/TT.

The posterior is now

$$p(\theta_1, \theta_2 \mid \mathrm{H}) = 2\big(\theta_1(1 - \theta_2) + (1 - \theta_1)\theta_2\big) \mathcal{B}(\theta_1; 2, 2)\mathcal{B}(\theta_2; 2, 2).$$

This is equivalent to the expression we derived with "the easy way."

[1] http://goo.gl/wRofuX

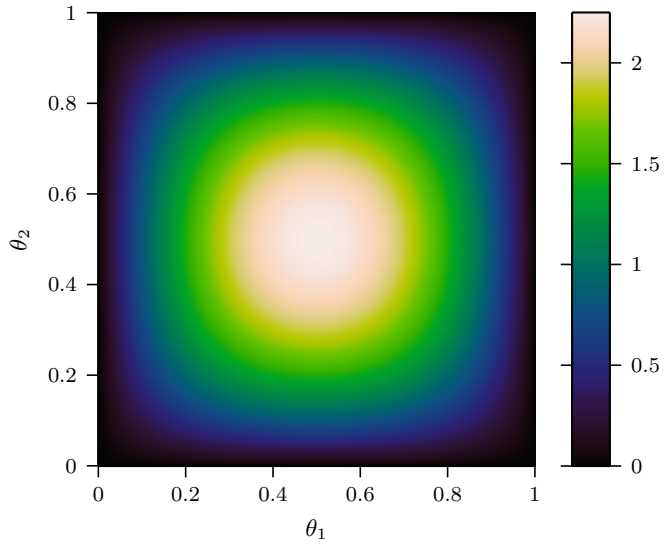

Figure 2: The joint prior $p(\theta_1, \theta_2)$.

The prior, likelihood, and posterior are plotted below. From the posterior, we can see that joint probabilities corresponding to jointly low or high values: these combinations would correspond to a higher probability of seeing either HH or TT observations. Despite the marginals for θ_1 and θ_2 remaining unchanged, the H observation has entangled the previously independent beliefs in the anticorrelated posterior.

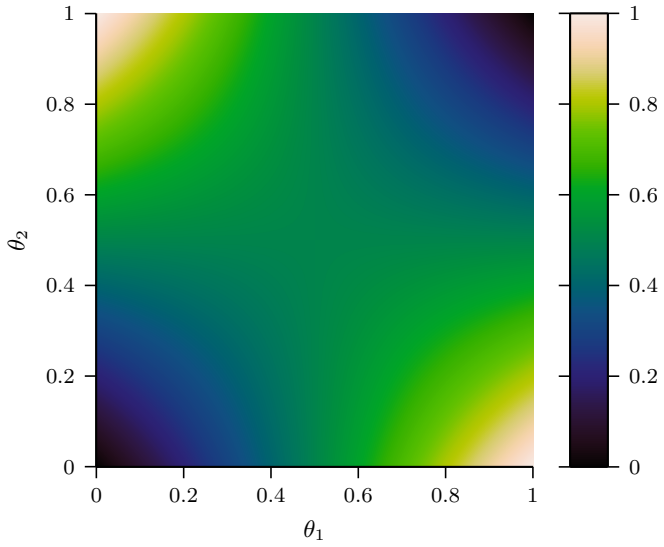

Figure 3: The joint likelihood $p(\mathrm{H} \mid \theta_1, \theta_2)$.

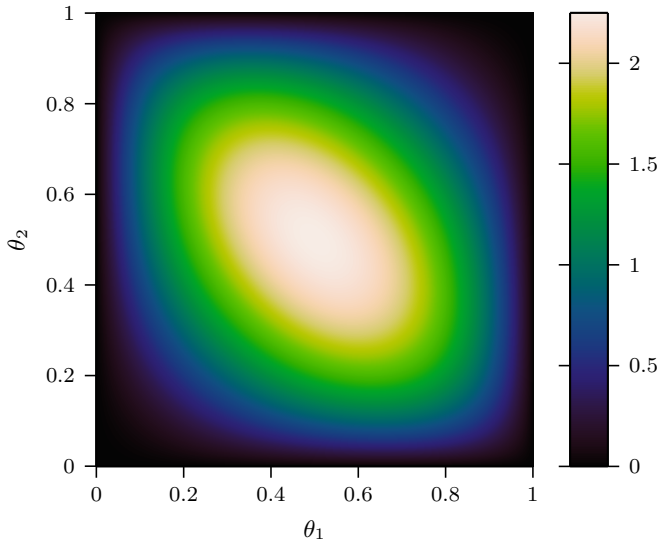

Figure 4: The joint posterior $p(\theta_1, \theta_2 \mid \mathrm{H})$.

2. Consider the three-dimensional parameter vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3]^\top$, with the following joint multivariate Gaussian prior:

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}\left(\begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} ; \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 0 \\ 2 & 9 & 0 \\ 0 & 0 & 16 \end{bmatrix} \right).$$

We are going to consider a decision problem with action space $\mathcal{A} = \{1, 2, 3\}$. The result of choosing an action $a \in \mathcal{A}$ will be to observe the exact value of θ_a, the ath element of $\boldsymbol{\theta}$.

Consider the following loss functions, ℓ_1 and ℓ_2:

$$\ell_1(\boldsymbol{\theta}, a) = \begin{cases} 1 & \theta_a > 0 \\ 0 & \theta_a \leq 0 \end{cases} \qquad \ell_2(\boldsymbol{\theta}, a) = \min(0, \theta_a).$$

For each:

- Write a generic expression for the expected loss of action a in terms of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Evaluate any integrals you encounter.
- Give a numerical value for the expected loss of each action, using the values of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ provided above.
- State the Bayes action.

Solution

First, we note that the loss functions only depend on $\boldsymbol{\theta}$ through θ_a, so we must only consider the marginal belief about θ_a when contemplating action a. By applying the marginalization formula for multivariate Gaussians, this belief is:

$$p(\theta_a) = \mathcal{N}(\theta_a; \mu_a, \Sigma_{aa}).$$

For loss ℓ_1, we may calculate the expected loss of each action:

$$\mathbb{E}\big[\ell_1(\boldsymbol{\theta}, a)\big] = \int_{-\infty}^{\infty} \ell_1(\boldsymbol{\theta}, a) p(\theta_a) \, \mathrm{d}\theta_a = \int_{0}^{\infty} \mathcal{N}(\theta_a; \mu_a, \Sigma_{aa}) \, \mathrm{d}\theta_a = 1 - \Phi(0; \mu_a, \Sigma_{aa}).$$

Using this result, we may numerically calculate the expected loss for each action:

$$\mathbb{E}\big[\ell_1(\boldsymbol{\theta}, 1)\big] = 0.5 \qquad \mathbb{E}\big[\ell_1(\boldsymbol{\theta}, 2)\big] = 0.631 \qquad \mathbb{E}\big[\ell_1(\boldsymbol{\theta}, 3)\big] = 0.691.$$

The Bayes action is $a = 1$, with the lowest expected loss.

For loss ℓ_2, we proceed in the same way:

$$\mathbb{E}\big[\ell_2(\boldsymbol{\theta}, a)\big] = \int_{-\infty}^{\infty} \ell_2(\boldsymbol{\theta}, a) p(\theta_a) \, \mathrm{d}\theta_a = \int_{-\infty}^{0} \theta_a \mathcal{N}(\theta_a; \mu_a, \Sigma_{aa}) \, \mathrm{d}\theta_a.$$

We may compute this definite integral; I used a table of Gaussian integrals[2] and the identity

$$\phi\left(\frac{a - \mu}{\sigma}\right) = \sigma \mathcal{N}(a; \mu, \sigma^2)$$

[2] http://en.wikipedia.org/wiki/List_of_integrals_of_Gaussian_functions

to derive

$$\mathbb{E}\big[\ell_2(\boldsymbol{\theta}, a)\big] = \mu_a \Phi(0; \mu_a, \Sigma_{aa}) - \Sigma_{aa} \mathcal{N}(0; \mu_a, \Sigma_{aa}).$$

Using this result, we may numerically calculate the expected loss for each action:

$$\mathbb{E}\big[\ell_2(\boldsymbol{\theta}, 1)\big] = -0.399 \qquad \mathbb{E}\big[\ell_2(\boldsymbol{\theta}, 2)\big] = -0.763 \qquad \mathbb{E}\big[\ell_2(\boldsymbol{\theta}, 3)\big] = -0.791.$$

The Bayes action is now $a = 3$, with the lowest expected loss.

3. Consider a d-dimensional vector $\boldsymbol{\theta}$ with an arbitrary multivariate Gaussian distribution:

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

   - Give a general expression for the distribution of the following (scalar) value τ.

$$\tau = \theta_1 + 2\theta_2 + \cdots d\theta_d$$

   - Consider again the specific distribution of the three-dimensional vector $\boldsymbol{\theta}$ from the last problem, as well as the action space $\mathcal{A}$ with the same observation mechanism: after choosing $a \in \mathcal{A}$, we will observe the corresponding value θ_a. Suppose we may select one action and then must predict τ under a squared loss function:

$$\ell(\tau, \hat{\tau}) = (\tau - \hat{\tau})^2.$$

     Using the distribution from the last problem. what is the expected loss of each of the three available actions? Which is the Bayes action?

Solution

Define the (row) vector $\mathbf{d}^\top = [1, 2, \ldots, d]$. We first notice that τ is simply a linear transformation of $\boldsymbol{\theta}$:

$$\tau = \mathbf{d}^\top \boldsymbol{\theta};$$

therefore τ has a multivariate Gaussian distribution:

$$p(\tau) = p(\mathbf{d}^\top \boldsymbol{\theta}) = \mathcal{N}(\tau; \mathbf{d}^\top \boldsymbol{\mu}, \mathbf{d}^\top \boldsymbol{\Sigma} \mathbf{d}).$$

In the second part of the question, we must consider estimating τ under a squared loss function $\ell(\tau, \hat{\tau}) = (\tau - \hat{\tau})^2$. A general result from Bayesian decision theory is that the Bayes action is to estimate $\hat{\tau}$ as the (posterior) mean of τ. For example, given the initial belief from the last problem, we would predict $\hat{\tau} = \mathbf{d}^\top \boldsymbol{\mu} = 8$. What is the *expected* loss when predicting the mean $\hat{\tau} = \mathbb{E}[\tau]$?

$$\mathbb{E}\Big[\ell\big(\tau, \mathbb{E}[\tau]\big)\Big] = \mathbb{E}\Big[\big(\tau - \mathbb{E}[\tau]\big)^2\Big] = \mathrm{var}[\tau] = \mathbf{d}^\top \boldsymbol{\Sigma} \mathbf{d}.$$

The expected squared loss is simply the variance of τ! Conveniently, we have a closed-form expression for this variance.

The problem asks us to consider how we would proceed with estimating τ if we could observe one of the entries of the vector $\boldsymbol{\theta}$ before making our prediction $\hat{\tau}$. If we wish to minimize our expected loss, we should minimize the variance of τ with our observation. Observing an entry of $\boldsymbol{\theta}$ is a conditioning observation of a multivariate Gaussian. We have a closed-form expression for the posterior covariance of $\boldsymbol{\theta}$ after observing any entry θ_a. Remarkably, the posterior covariance of $\boldsymbol{\theta}$ does not depend on the actual value we observe, only the index of the entry we choose, a. The

posterior covariance matrices $\mathbf{\Sigma}_{\boldsymbol{\theta}|\theta_a}$ for each available action $a \in \mathcal{A}$ are:

$$\mathbf{\Sigma}_{\boldsymbol{\theta}|\theta_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 16 \end{bmatrix}$$
$$\mathbf{\Sigma}_{\boldsymbol{\theta}|\theta_2} = \begin{bmatrix} \frac{5}{9} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 16 \end{bmatrix}$$
$$\mathbf{\Sigma}_{\boldsymbol{\theta}|\theta_3} = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 9 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

The expected losses of our final prediction of $\hat{\tau}$ given θ_a is now given by

$$\mathbb{E}\big[\ell(\tau, \hat{\tau}) \mid \theta_a\big] = \mathbf{d}^\top \mathbf{\Sigma}_{\boldsymbol{\theta}|\theta_a} \mathbf{d}.$$

For our particular problem, the expected final loss after each potential action is:

$$\mathbb{E}\big[\ell(\tau, \hat{\tau}) \mid \theta_1\big] = 164 \qquad \mathbb{E}\big[\ell(\tau, \hat{\tau}) \mid \theta_2\big] = 144\,5/9 \qquad \mathbb{E}\big[\ell(\tau, \hat{\tau}) \mid \theta_3\big] = 45.$$

The Bayes action is $a = 3$. Despite the fact that θ_3 is uncorrelated with the other two entries, collapsing its large variance from 16 to zero has the effect of reducing the variance of τ (and therefore our expected loss) by $3^2 \cdot 16 = 144$.

4. Consider the following data:

$$\mathbf{x} = [0.54, 1.84, -2.26, 0.86, 0.32]^\top;$$
$$\mathbf{y} = [-1.31, -0.43, 0.34, 3.58, 2.77]^\top.$$

Consider the Bayesian linear regression model with $\phi(x) = [1, x]^\top$. Use the prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{0}, \mathbf{I})$.

Plot the posterior probability that the slope of the regression line is positive as a function of the standard deviation of the observation noise σ (the noise variance is then σ^2). Use a grid of at least 100 points in the range $\sigma \in (0.01, 10)$.

Solution

Using the given linear regression model, we assume

$$y = \phi(x)^\top \mathbf{w} + \varepsilon = w_1 + w_2 x + \varepsilon.$$

The second entry of the weight vector $\mathbf{w}$, w_2, therefore serves as the slope of the regression line.

The Bayesian linear regression model gives the following posterior for $\mathbf{w}$ given observations $\mathcal{D}$ and a specified noise variance σ^2:

$$p(\mathbf{w} \mid \mathcal{D}, \sigma^2) = \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}_{\mathbf{w}|\mathcal{D}}, \boldsymbol{\Sigma}_{\mathbf{w}|\mathcal{D}}),$$

where

$$\boldsymbol{\mu}_{\mathbf{w}|\mathcal{D}} = \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top + \sigma^2 \mathbf{I})^{-1} \mathbf{y};$$
$$\boldsymbol{\Sigma}_{\mathbf{w}|\mathcal{D}} = \mathbf{I} - \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top + \sigma^2 \mathbf{I})^{-1} \mathbf{X},$$

where we have plugged the given prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{0}, \mathbf{I})$ into the general result.

Given a value of σ, the formulas above give the posterior over $\mathbf{w}$. To determine the probability that w_2 is positive, we simply take the marginal posterior distribution and evaluate the normal CDF:

$$\Pr(w_2 > 0 \mid \mathcal{D}, \sigma^2) = 1 - \Phi\big(0; (\boldsymbol{\mu}_{\mathbf{w}|\mathcal{D}})_2, (\boldsymbol{\Sigma}_{\mathbf{w}|\mathcal{D}})_{22}\big).$$

This quantity is plotted as a function of σ in Figure 5. The larger the noise, the less confident we become about the sign of the slope.

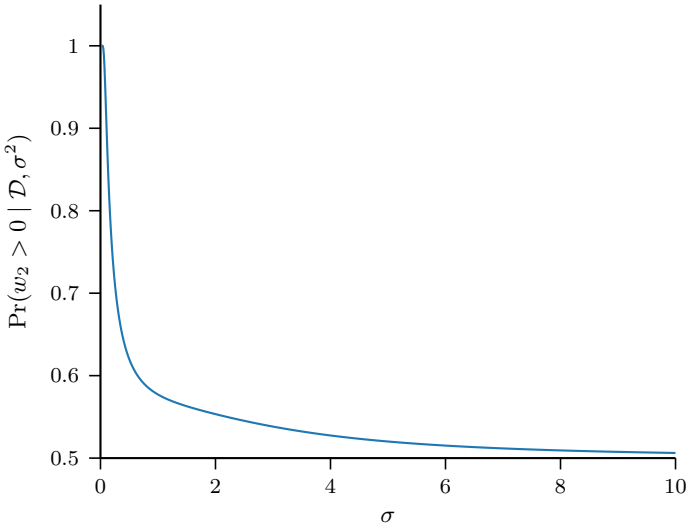

Figure 5: The posterior probability that the slope of the regression line is positive as a function of the noise standard deviation σ. Note the limits on the y-axis.