Unit 6: regularization, priors and Bayesian inference

Ethan Levien

November 6, 2025

1.1 Introduction to Bayesian inference: Simple examples

So far, we think of models encode our assumptions about how data was generate. Our models depend on parameters for which the true values are fixed. When use data to fit or infer parameters, obtain a sample distribution which roughly speaking quantifies uncertainty in our estimates. This way of performing statistics, where the uncertainty is thought of as a distribution over repeated experiments, is frequentist approach to statistics. In this formulation of statistical inference we are pretending to have complete ignorance of the parameters before we see the data, but in reality, this is never true. For example, if we flip a coin $N = 3$ times and get $Y = 3$ heads, our estimate of the probability this coin will land on heads in the next flip is $\hat{q} = 1$, which is clearly not in line with out understanding that both sides are at least possible. It is also worth noting that the usual estimate standard error, $\text{se}(\hat{q}) = \sqrt{q(1-q)/N} \approx \sqrt{\hat{q}(1-\hat{q})/N}$ gives zero, but this is obviously not a good estimate of the uncertainty.

Similar issues emerge in the context of more machine learning-style data analysis, sometimes we want to incorporate vague information, such as "the function describing my data is very smooth and changes roughly on a time-scale of 5 hours". Or, we might want to include many predictors, but to avoid overfitting penalize high values of the predictors. For example, we might believe there is an interaction term, but suspect it is much smaller than the additive terms.

While there are ways handle these problems within the frequentist framework, they are handled more naturally by taking an entirely different treatment of parameters and statistical inference. This is called Bayesian statistics. In the Bayesian formulation of statistics, we will think of parameters as themselves random variables which are given a distribution before we have seen the data. Mathematically, this means instead of having a model with fixed parameter:

$$X \sim \text{ModelDistribution}(\theta) \tag{1}$$

we think of our model as a conditional on θ

$$X|\theta \sim \text{ModelDistribution}(\theta). \tag{2}$$

We call $X|\theta$ the likelihood (this term is used in both frequentists and Bayesian approaches). Then, we add a new distribution for θ, called the prior. The goal is then to condition on the observed data $X_1, \ldots, X_N$ to find a new distribution $\theta|X_1, \ldots, X_N$, called the posterior. The posterior plays a similar role to the sample distribution, and naturally we often use its mean as an estimate of θ. However, there are both philosophical and mathematical differences between these two approaches which make the comparison imperfect. We will focus less on the philosophical aspect and see what is practically different about the approaches. This is best illustrated with the following example.

1.1.1 Bayesian inference for Bernoulli trails

Consider the model

$$X \sim \text{Bernoullli}(q)$$

for the outcome of a coin flip. If we perform N flips, then recall $Y = \sum_{i=1}^{N} X_i$ is binomial. In a Bayesian model, we would specify the prior distribution of q. This should reflect what we know about the coin, which in reality

is that it is very likely to be fair. However, to make calculations simpler and better compare the frequentist approach we will take a prior where any value of q is equally likely. This seems to be more similar to what we are doing in the frequentist approach. Our Bayesian model would then me

$$q \sim \text{Uniform}(0,1)$$
$$X|q \sim \text{Bernoullli}(q)$$

Our goal is to find the posterior distribution

$$q|X_1, \ldots, X_N$$

Since the flips out independent, this distribution depends only on Y, so we can alternative obtain the posterior as the distribution of $q|Y$. Recall that the density of the uniform is simply $f(q) = 1$ (I'll use f for a density and P for probabilities, but sometimes we have to mix them). Using Bayes' theorem (or simply using the definition of conditional probability twice), we have

$$f(q|Y = y) = \frac{f(q, \{Y = y\})}{P(Y = y)} = \frac{P(Y = y|q)f(q)}{P(Y = y)} \tag{3}$$

where $f(q, \{Y = y\})$ means the joint density/probability of q and Y. Recall the binomial distribution is

$$P(Y = y|q) = \binom{N}{y} q^y (1-q)^{N-y} \tag{4}$$

The important thing to realize is that all the q dependence comes from $P(Y = y|q)$ and since our goal is to find a distribution of q, we can write

$$f(q|Y = y) = C(N, y)q^y(1-q)^{N-y} \tag{5}$$

where $C(N, y)$ is a constant which ensures $\int f(q|Y = y)dq = 1$. That is,

$$C(N, y)^{-1} = \int_0^1 q^y(1-q)^{N-y}dq \tag{6}$$

$f(q|Y = y)$ is an example of Beta distribution. More generally, we say a variable q is

$$q \sim \text{Beta}(a, b) \tag{7}$$

if q has the density

$$f(q) = q^{a-1}(1-q)^{b-1} \tag{8}$$

hence $q|Y \sim \text{Beta}(y+1, N-y+1)$. You don't need to memorize this, but you should look on Wikipedia or in the colab notebook to get a sense of what this function looks like.

Example 1 (Laplace Rule of Succession of Bayesian view)**.** Consider Bayesian inference for Bernoulli trials with uniform priors as described above. Recall that Laplace's Rule of Succession gives the estimator

$$\hat{q}_L = \frac{Y+1}{N+2} \tag{9}$$

Question: Show that when $Y = N$

$$E[q|Y] = \frac{N+1}{N+2} = \hat{q}_L$$

Solution: In this case

$$C(N, y)^{-1} = C(N, N)^{-1} = \int_0^1 q^N dq = \frac{1}{N+1} \tag{10}$$

and

$$E[q|Y] = (N+1)\int_0^1 q \times q^N dq = (N+1)\int_0^1 q^{N+1} dq = \frac{N+1}{N+2}. \tag{11}$$

To summarize the idea of Bayesian inference, If Y is a our data and θ is our parameter(s) (q in the example above) Bayes' theorem tells us that

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)} \tag{12}$$

The distributions appearing in Equation (12) have the following names and interpretations:

- $P(\theta|Y)$ is the posterior (a beta distribution in Example 1).
- $P(Y|\theta)$ is the likelihood (Binomial distribution in Example 1).
- $P(\theta)$ is the prior distribution (Uniform distribution in Example 1)
- $P(Y)$ is the evidence. It represents the chance we observe the data *unconditional* on the parameters. This can be obtained by marginalizing over the priors.

While in classical statistics, the objective is to determine (estimate) a parameter θ and measure its uncertainty (with the sample distribution), in Bayesian statistics our objective is to compute the posteior distribution. Once we have this, we can obtain so-called point estimates, for example, by taking the average of θ or maximum of $P(\theta|Y)$ (maximum likelihood). The point estimates are however less central in Bayesian inference than the posterior.

1.1.2 Selecting priors with a beta distribution (optional)

If

$$q \sim \text{Beta}(a, b)$$

then (using calculus) it can be shown that

$$E[q] = \frac{a}{a+b} \tag{13}$$

$$\text{var}(q) = \frac{ab}{(a+b)^2(a+b+1)}. \tag{14}$$

Moreover, you can check (by plotting the density in Python – see exercises) that when a and b become large the beta distribution approaches a Normal distribution.

Example 2 (Selecting priors)**.** Suppose we are given a coin. We are trying to determine whether it is biased based on the outcome of N flips. We are pretty sure that, like most coins, it is not biased. To be precise, let's say you are 95% confident that the bias of the coin is less than 10% biased towards heads or tails.

Question:

(a) Select a prior distribution for q.

(b) What is the posterior expectation if we flip the coin 5 times and see 5 heads?

Solution:

(a) We know that a β-distribution is approximately Normal if a and b are large enough. Assuming we can make a Normal approximation, we can use the formulas for the mean and variance of the b

distribution to select parameters such that

$$P(|q - 0.5| < 0.1) = P(0.4 < q < 0.6) \approx 0.95 \tag{15}$$

Since we would like the mean of our prior distribution to be $1/2$, we select

$$q \sim \text{Beta}(a, a).$$

If q is approximately Normal, then Equation 15 will hold when

$$\begin{aligned} 1.96\sqrt{\text{var}(q)} &= 0.1 \\ &\implies \sqrt{\frac{a^2}{4a^2(2a+1)}} = \frac{1}{2}\sqrt{\frac{1}{2a+1}} = \frac{0.1}{1.6} \\ &\implies a \approx 31.5 \end{aligned}$$

(b) In order to compute the posterior, we use Bayes' theorem

$$f(q|X_1, \ldots, X_N) \propto q^Y(1-q)^{N-Y} \times q^{a-1}(1-q)^{a-1} = q^{Y+a-1}(1-q)^{N-Y+a-1}$$

where $Y = \sum_{i=1}^{N} X_i$. Again, the q dependence uniquely determines the distribution, since the area under this curve must be one. We can therefore say that

$$q|X_1, \ldots, X_N \sim \text{Beta}(Y + a, N - Y + a)$$

For the mean and variance we have

$$E[q|X_1, \ldots, X_N] = \frac{Y + a}{N + 2a} = \frac{5 + 31.5}{5 + 2 \times 31.5} \approx 0.54$$

Note that if we used usual estimator we would have found $\hat{q} = 1$, while if we used uniform priors and taken the posterior expectation $\hat{q}_L = 0.85$. Yet another approach would be to take the posterior maximum (the mode) of the posterior in using the β-distribution priors.

1.1.3 Bayesian inference for Normal mean

Suppose we have a Normal model for X, but treat the noise variance as a constant and focus on Bayesian inference of μ.

$$X|\mu \sim \text{Normal}(\mu, \sigma^2/N).$$

and

$$\bar{X}|\mu \sim \text{Normal}(\mu, \sigma^2/N).$$

It is natural to take Normal priors on μ, since then the marginal of $\bar{X}$ is Normal (you should check this!) making our calculations much easier. Hence we set

$$\mu \sim \text{Normal}(m, s^2)$$

Notice that this is nothing but a linear regression model for $\bar{X}$ with μ as a predictor (and regression slope $\beta = 1$). Therefore, the computation of the posterior $\mu|\bar{X}$ amounts to "inverting" a linear regression model, something we are familiar with from a number of examples and exercises in previous units.

In particular, since $\mu|\bar{X}$ is Normal, we only need to calculate the mean and variance.

$$m = E[\mu] = b_1 E[\bar{X}] + b_0 \tag{16}$$

$$s^2 = \text{var}(\mu) = b_1^2 \text{var}(\bar{X}) + \text{var}(\mu|\bar{X}) \tag{17}$$

where b_1 and b_0 are the regression slope and intercept when Y is treated as the predictor. In particular, $b_0 = E[\mu|\bar{X} = 0]$.

The variance and expected value of Y are

$$E[\bar{X}] = m, \quad \text{var}(\bar{X}) = s^2 + \sigma^2/N. \tag{18}$$

The regression slope b appearing in the above formula can be found from the covariance. Using the relation $\text{cov}(X, Y) = \beta_1 \sigma_X^2$ (here X is μ and $\beta_1 = 1$), we have

$$\text{cov}(\mu, \bar{X}) = s^2 \tag{19}$$

hence the "inverse" regression slope is obtained by the usual regression slope relation:

$$s^2 = \text{cov}(\mu, \bar{X}) = b_1(s^2 + \sigma^2/N) \implies b_1 = \frac{s^2}{s^2 + \sigma^2/N} \tag{20}$$

Coming these facts yields

$$b_0 = m\left(1 - \frac{s^2}{s^2 + \sigma^2/N}\right) = m\frac{\sigma^2/N}{s^2 + \sigma^2/N} \tag{21}$$

Finally

$$\begin{aligned} \text{var}(\mu|\bar{X}) = \frac{s^2}{N} - b_1^2 \text{var}(Y) &= \frac{s^2}{N} - \frac{s^2 + \sigma^2/N}{(1 + \left(\frac{\sigma}{s}\right)^2 \frac{1}{N})^2} \\ &\vdots \quad \text{(some algebra)} \\ &= \frac{s^2\sigma^2}{s^2 N + \sigma^2} \end{aligned}$$

In summary, the posterior distribution is

$$\mu|\bar{X} \sim \text{Normal}\left(b_1\bar{X} + m(1 - b_1), \frac{s^2\sigma^2}{s^2 N + \sigma^2}\right). \tag{22}$$

Take note of what happens when σ/s is very large/small.

1.2 Bayesian inference for linear regression model with a single-predictor

Now we consider a linear regression model with a single predictor. To make things simpler, we assume σ_ϵ^2, $\mu_X = \mu_Y = 0$ (that is, zero marginal expectations for the predictor and response variables). We will take mean zero Normal priors on the regression coefficient, so the model is

$$\beta \sim \text{Normal}(0, \tau^2) \tag{23}$$

$$Y|X, \beta \sim \text{Normal}(\beta X, \sigma_\epsilon^2) \tag{24}$$

The assumption that the mean of the prior distribution is zero can be relaxed without too much effort, but it will make the calculations a bit simpler. The motivation for taking Normal priors on β is that the posterior will be Normal, and therefore we only need to calculate the mean and variance. In this case, the posterior distribution is obtained by conditioning β on all the X and Y points

$$\beta|(X_1, Y_1), \ldots, (X_N, Y_N) \tag{25}$$

but under the assumption we have made, this is the same as conditioning on the OLS estimator $\hat{\beta}$ and the X points.

Note that because we have assumed X and Y are mean zero, $\beta = \text{cov}(X,Y)/\text{var}(X) = E[XY]/E[X^2]$ and hence

$$\hat{\beta} = \frac{\sum_{i=1}^N X_i Y_i}{\sum_{i=1} X_i^2}. \tag{26}$$

Since the Y_i are Normal, $\hat{\beta}|(\beta, X_1, \ldots, X_N)$ is Normal and has mean and variance

$$E[\hat{\beta}|\beta X_1, \ldots, X_N] = E\left[\frac{\sum_{i=1}^N X_i Y_i}{\sum_{i=1} X_i^2}\right] = \frac{\sum_{i=1}^N X_i E[Y_i|\beta, X_i]}{\sum_{i=1} X_i^2} \tag{27}$$

$$= \frac{\sum_{i=1}^N X_i E[Y_i|\beta, X_i]}{\sum_{i=1} X_i^2} = \frac{\sum_{i=1}^N X_i^2 \beta}{\sum_{i=1} X_i^2} = \beta \tag{28}$$

$$\text{var}(\hat{\beta}|\beta X_1, \ldots, X_N) = \frac{\sum_{i=1}^N X_i^2 \text{var}(Y_i|\beta, X_i)}{\left(\sum_{i=1} X_i^2\right)^2} \tag{29}$$

$$= \frac{\text{var}(Y_i|\beta, X_i)}{\sum_{i=1} X_i^2} = \frac{\sigma_\epsilon^2}{N\hat{\sigma}_X^2} \tag{30}$$

Therefore,

$$\hat{\beta}|\beta, X_1, \ldots, X_N \sim \text{Normal}(\beta, \sigma_\epsilon^2/(N\hat{\sigma}_X^2)). \tag{31}$$

You might recognize the mean and variance from the sample distribution of the OLS estimator. Indeed, they are the same! But now we can use this to find the posterior distribution of β conditioned on the OLS estimator $\hat{\beta}$. The key is that this is a linear regression model with β as the predictor and $\hat{\beta}$ as the response variable! If we "invert" the linear regression model as we have done before (you should fill in the details for yourself – see exercises), we get

$$\beta|(\hat{\beta}, X_1, \ldots, X_N) \sim \text{Normal}\left(\hat{\beta}\frac{\tau^2}{\tau^2 + \sigma_\epsilon^2/(N\hat{\sigma}_X^2)}, \frac{\sigma_\epsilon^2/(N\hat{\sigma}_X^2)\tau^2}{\tau^2 + \sigma_\epsilon^2/(N\hat{\sigma}_X^2)}\right) \tag{32}$$

1.3 Bayesian inference for nonlinear model

We now discuss Bayesian inference for the general linear regression model with features. Let

$$f(x) = \sum_{i=1}^K \beta_i \phi_i(x)$$

and suppose that our priors are

$$\beta_i \sim \text{Normal}(0, \tau_i^2).$$

We will assume that $E[\phi_i(X)] = 0$. We will also assume, for simplicity, that σ **is known. Hence, we do not need to consider estimating it.** This simplifies the conceptual picture considerably.

In this case, the Posterior distribution of the β_i can be understood analytically. It turns out that the marginal posterior distribution of each β_i is again Normal – thus Normal priors on β_i are conjugate priors!. The marginal mean of each β_i are determined by the system of questions given in the following Theorem.

Theorem 1 (Posterior mean for regression coefficients)**.** *Define a $K \times K$ matrix Ω called the empirical covariance matrix which has entries*

$$\Omega_{i,j} = N\overline{\phi_i(X)\phi_j(X)} = \sum_{k=1}^N \sum_{z=1}^N \phi_i(X_k)\phi_j(X_z)$$

and let $\bar{\beta}_i$ denote the posterior mean of β_i; that is,

$$\bar{\beta}_i = E[\beta_i|y, X].$$

Then $\bar{\beta}_1, \ldots, \bar{\beta}_K$ satisfy the K equations

$$\left(\frac{\sigma}{\tau_i}\right)^2 \bar{\beta}_i + \sum_{j=1}^{K} \Omega_{i,j}\bar{\beta}_j = \sum_{j=1}^{N} \phi_i(X_j)Y_j \tag{33}$$

We make a few remarks on this Theorem:

- If $N > K$, then $\bar{\beta}_i$ actually have a solution. In-fact, depending on Ω, there may not be a solution. For our discussion, we will assume there is however.
- Under the assumption that $E[\phi_i(X)] = 0$, the entries of $\Omega_{i,j}$ approximate the covariances between the predictor features. Moreover, if we take the predictors to be drawn from some distribution and average over the data (both X and Y), we get

  $$\frac{E[\Omega_{i,j}]}{N} = \text{cov}(\phi_j(X), \phi_i(X)).$$

  and

  $$\frac{1}{N}\sum_{j=1}^{N} E[\phi_i(X_j)Y_j] = \text{cov}(Y, \phi_i(X)).$$

  Hence, we have the "math world" version of Equations 33, which is obtained by averaging over the data:

  $$\frac{1}{N}\left(\frac{\sigma}{\tau_i}\right)^2 E[\bar{\beta}_i] + \sum_{j=1}^{K} \text{cov}(\phi_i(X), \phi_j(X))E[\bar{\beta}_j] = \text{cov}(Y, \phi_i(X)).$$

  In the limit $N \to \infty$ and with $K = 2$, we obtain a system of equations recognizable from Week 5 notes for the regression coefficients in the two predictor model in terms of the covariances. In this context, the role of the regression coefficients (which we previously took to be fixed numbers), is now played by the averages $E[\bar{\beta}_j]$, which are the average values of β_j with respect to both the posterior and the data.
- Notice that the first term in Equation 33 vanishes as either $\tau_i \to \infty$ (very broad priors) or $\sigma \to 0$ (no variance in Y conditional on the predictors). In both cases, the vales of $\bar{\beta}_i$ are entirely determined by the data, and the priors play no role. When $\tau_i \to 0$ or $\sigma \to \infty$, the priors entirely determine determine $\bar{\beta}_i$ and in fact $\bar{\beta}_i = 0$.
- The first term in Equation 33 is an example regularization.

Theorem 1 can be elegantly stated using matrices (see Linear algebra review note for intro to matrix multiplication). This is import if we want to implement these computations in Python. Observe that one way to construct $\Omega_{i,j}$ is to define the matrices

$$A = \begin{bmatrix} | & | & & | \\ \phi_1(X) & \phi_2(X) & \cdots & \phi_K(X) \\ | & | & & | \end{bmatrix}, \quad \Lambda_0 = \begin{bmatrix} \tau_1 & 0 & \ldots & 0 \\ 0 & \tau_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \tau_K \end{bmatrix}$$

Then $\Omega = A^T A$. Now define another matrix $K \times K$ matrix

$$M = \Omega + \sigma^2 \Lambda_0^{-1}$$

Then Equations 33 can be written as

$$M\bar{\beta} = A^T y$$

where $\bar{\beta}$ and y are respectively a K-vector and N-vector

$$\bar{\beta} = \begin{bmatrix} \bar{\beta}_1 \\ \vdots \\ \bar{\beta}_K \end{bmatrix}, \quad y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix}.$$

Then $\bar{\beta}$ satisfies

$$M\bar{\beta} = A^T y \implies \bar{\beta} = \left(A^T A + \sigma^2 \Lambda_0^{-1}\right)^{-1} A^T y$$

Example 3 (Calculating posterior mean in python)**.** Consider the model with two predictors:

$$f(x) = \beta_1 X_1 + \beta_2 X_2$$

Question: Generate some simulated data and compute the posterior means of β for different choices of τ.

Solution: See colab notebook.

1.3.1 Posterior prediction

We will mostly focus on how we can use the posterior estimates to make predictions, although there is an entirely different set of questions we could ask concerning model assessment and hypothesis testing.Recall that $\hat{y}(x, D)$ is defined as our prediction, using the regression model fitted to the data D, of $E[Y|X = x]$. In the Bayesian context, we can define the posterior predictive variable $Y|X, D$ as the random representing the distribution of our response variable conditioned on the data. We can obtain this distribution by replacing the β_is in

$$Y = \sum_{i=1}^{K} \beta_i \phi_i(X)$$

with samples from the posterior distribution. In the Bayesian framework we set

$$\hat{y}(x, D) = E[Y|D, X] = \sum_{i=1}^{K} E[\beta_i|D]\phi_i(X) = \sum_{i=1}^{K} \bar{\beta}_i \phi_i(X)$$

We can write this in matrix notation by defining the vector

$$a = \begin{bmatrix} \phi_1(X) \\ \vdots \\ \phi_K(X) \end{bmatrix}.$$

Then

$$\hat{y}(x, D) = a^T \bar{\beta} = a^T M^{-1} A^T y \tag{34}$$

Example 4 (Posterior predictions for a regression model)**.** Take the Fourier model

$$f(x) = \beta_0 + \sum_{i=1}^{K} \beta_i \sin\left(\frac{2\pi i x}{L}\right) + \alpha_i \cos\left(\frac{2\pi i x}{L}\right).$$

Implement a function to compute posterior predictions with different priors τ on the coefficients. In this example, we will fit this model to data generated by adding noise to the function

$$f_{\text{true}}(x) = \sin(2\pi 3x) + 5\sin(2\pi 6x^2) + 20x^2$$

with $x \in [0, 1]$. Note that for $K < \infty$, we can not pick any values of the β_is and α_is such that $f = f_{\text{true}}$.

We will assume that β_i and α_i have the same prior distribution:

$$\beta_i \sim \text{Normal}(0, \tau_i)$$
$$\alpha_i \sim \text{Normal}(0, \tau_i)$$

Question: Compute the posterior mean $\hat{y}(x, D)$ for the following choices of priors and $K = 50$.

(a) All τ_i are the same, $\tau_i = \tau_0$.

(b) The τ_i decay with i as $\tau_i = \tau_0/i$

In each case, experiment with different choices of τ_0 and pay attention to whether the model is overfitting (high variance, low bias) or under fitting (low variance, high bias).

Solution: See colab notebook.

1.4 Regularization view of priors

- Recall that in standard least squares linear regression, the regression coefficients come from minimizing the sum of squared residuals. That is, they are the values that minimize the function

$$L(\beta) = \sum_{j=1}^{N} (\hat{y}(X_j, D) - Y_j)^2$$

This is an example of a loss function, which is simply something we want to minimize to solve our inference problem. This means that the derivative of L with respect to each β_i is zero.

$$\begin{aligned}
\frac{\partial}{\partial \beta_i} L(\beta) &= \sum_{z=1}^{N} 2(\hat{y}(X_z, D) - Y_z) \frac{\partial}{\partial \beta_i} \hat{y}(X_j, D) \\
&= \sum_{z=1}^{N} 2(\hat{y}(X_z, D) - Y_z) \phi_i(X_z) \\
&= 2 \sum_{z=1}^{N} \sum_{j=1}^{K} \beta_j \phi_i(X_z) \phi_j(X_z) - 2 \sum_{z=1}^{N} Y_z \phi_i(X_z) \\
&= 2 \sum_{z=1}^{K} \beta_j \sum_{z=1}^{N} \phi_i(X_z) \phi_j(X_z) - 2 \sum_{z=1}^{N} Y_z \phi_i(X_z) \\
&= 2 \sum_{j=1}^{K} \beta_j \Omega_{i,j} - 2 \sum_{z=1}^{N} Y_z \phi_i(X_z)
\end{aligned}$$

Setting

$$\frac{\partial}{\partial \beta_i} L(\beta) = 0$$

leads to Equation 33 without the prior term.

- The regularization view of Equation 33 is that, instead of thinking of the additional term as coming from priors, we think about adding a term to our loss function which penalizes models with too much flexibility.

A common choice is

$$L(\beta) = \sum_{j=1}^{N} (\hat{y}(X_j, D) - Y_j)^2 + \lambda \sum_{j=1}^{K} \beta_j^2$$

This is called ridge regression and λ is a parameter controlling how large a penalty we place on large values of β_j. If you compute the partial derivatives and equate it to zero, you will find that the equations β_j satisfy are exactly Equation 33 with $\sigma/\tau_i = \sqrt{\lambda}$. More generally, we could use the loss function

$$L(\beta) = \sum_{j=1}^{N} (\hat{y}(X_j, D) - Y_j)^2 + \sum_{j=1}^{K} \left(\frac{\sigma}{\tau_i}\right)^2 \beta_j^2$$

we would obtain exactly Equation 33.

Exercise 1 (Bayesian linear regression for Bernoulli model)**:** Suppose you are doing a poll of Dartmouth students political party affiliation and for simplicity suppose there are only two options, democrat or republican.

(a) Select a prior distribution for q is in Example 2. You will need to use Equations 13 (but you don't need to memorize these). This step is subjective and the goal is to see how your own prior knowledge interacts with the data. Plot the prior distribution in Python.

(b) Now assume N students are surveyed. They are all democrats. Compute and plot the posterior distribution for different values of N on the same plot as the prior distribution. Also plot the posterior mean and standard deviation as a function of N.

(c) How large does N need to be in order for you to be 95% confident that 95% of Dartmouth students are democrats? You can either estimate this by hand (using Equations 13 again) or with simulations (by sampling the posterior).

Exercise 2 (Normal model and MSE)**:** Consider the Bayesian Normal model with known σ. Suppose that the true value of μ is μ_0. Bayesian inference leads to the posterior mean

$$\hat{\mu}_P = E[\mu|\bar{X}] = m(1-b_1) + \bar{X}b_1 = m(1-b_1) + \hat{\mu}b_1 \tag{35}$$

which can be seen as an estimator of μ.

In class we calculated the MSE in the usual sense, meaning

$$\text{MSE}_{\hat{\mu}_P} = E[(\hat{\mu}_P - \mu_0)^2] \tag{36}$$

where the expectation is taken over the distribution of our data.

An alternative way to defined MSE is as the expected squared difference between the true value and a sample μ from the posterior averaged over all $\bar{X}$; that is

$$\text{MSE}'_{\hat{\mu}_P} = E[E[(\mu - \mu_0)^2|\bar{X}]]. \tag{37}$$

In particular, the first expectation is over $\mu|\bar{X}$ (μ_0 is fixed) and the second is over the data distribution. Compute this quantity and interpret the terms. Compare the bias and variance to what we found for the other case.

Exercise 3 (Bayesian linear regression ❑)**:** (a) Finish the calculation leading to Eq. 32.

(b) Generalize Eq. 32 to the case when the posterior mean is not zero

Exercise 4 (Normal model with unknown variance ❑)**:** Consider a Bayesian model of a Normal distribution with unknown variance. In this case we place priors on σ^2 as well, but we cannot use a Normal distribution since σ needs to be positive. A common approach is to take $\ln\sigma$ be Normal, leading to the Bayesian model

$$\ln\sigma \sim \text{Normal}(l, \gamma^2) \tag{38}$$
$$\mu \sim \text{Normal}(m, s^2) \tag{39}$$
$$\bar{X}|\mu, \sigma \sim \text{Normal}(\mu, \sigma^2) \tag{40}$$

where $\ln\sigma$ and μ are independent under the prior distribution. Pick some values for ν and γ^2 and write code to generate samples from the marginal distribution of $\bar{X}$. Is this a Normal distribution? Compare the to a Normal density with the same mean and variance.