12 Bayesian Linear Regression – Learning from Both Data and Belief

12.1 Bayesian Inference: Updating Beliefs with Data

12.1.1 Treating Parameters as Random Variables

Bayesian inference provides a systematic way to update our beliefs about unknown parameters using observed data (see Figure 1). It views model parameters as random variables, each with its own probability distribution.

Before seeing any data, we rely on prior knowledge or assumptions about these parameters, which is expressed through a prior distribution. Once sample data becomes available, we compute how likely the data is under different parameter values, which gives us the likelihood. Bayes' theorem then combines the prior and the likelihood to produce the posterior distribution, a refined and data-informed belief about the parameters.

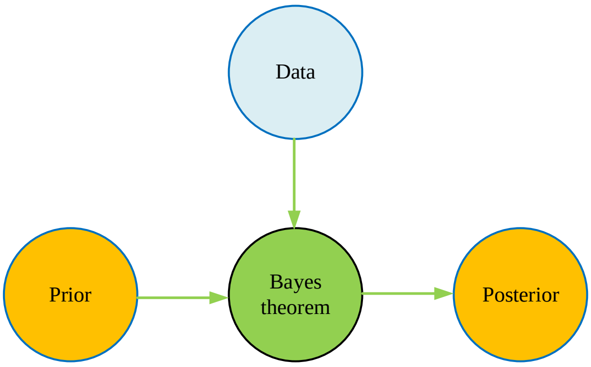

Figure 1. Bayesian inference: posterior distribution is obtained by updating the prior using the likelihood.

The fundamental formula is:

$$\overbrace{f_{\Theta|X}\left(\theta|x\right)}^{\text{Posterior}} = \frac{\overbrace{f_{X|\Theta}\left(x|\theta\right)}^{\text{Likelihood}}\overbrace{f_{\Theta}\left(\theta\right)}^{\text{Prior}}}{\int_{\vartheta} f_{X|\Theta}\left(x|\vartheta\right) f_{\Theta}\left(\vartheta\right)\mathrm{d}\vartheta} \tag{1}$$

This posterior distribution serves as the basis for statistical inference and decision-making. In practical machine learning, Bayesian inference often leads to Maximum A Posteriori (MAP) estimation, where we choose the parameter value that maximizes the posterior probability. Later in this chapter, we will apply Bayesian inference to build a Bayesian version of linear regression.

12.1.2 Prior, Likelihood, and Posterior

To connect linear regression with Bayesian inference, we write the multiple linear regression model in the form

$$\hat{y}^{(i)} = \theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \cdots + \theta_D x_D^{(i)} \tag{2}$$

When D = 1, this becomes the familiar single-variable linear model:

$$\hat{y}^{(i)} = \theta_0 + \theta_1 x_1^{(i)} \tag{3}$$

With the noise term, the equation above is re-written as

$$\hat{y}^{(i)} = \theta_0 + \theta_1 x_1^{(i)} + \varepsilon^{(i)} \tag{4}$$

We assume that the noise term $\varepsilon^{(i)}$ follows a normal distribution $N(0, \sigma^2)$. Under this assumption, the likelihood of observing the sample data is

$$f_{\gamma|\Theta}\left(\boldsymbol{y} \mid \boldsymbol{\theta}\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y^{(i)} - \left(\theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \cdots + \theta_D x_D^{(i)}\right)\right)^2}{2\sigma^2} \right) \tag{5}$$

This means that the assumed residuals obey $N(0, \sigma^2)$.

12.1.3 Maximum A Posteriori (MAP) Estimation

Using Bayes' theorem, we can get the posterior distribution:

$$f_{\Theta|\gamma}\left(\boldsymbol{\theta} \mid \boldsymbol{y}\right) = \frac{f_{\gamma|\Theta}\left(\boldsymbol{y} \mid \boldsymbol{\theta}\right) \cdot f_{\Theta}\left(\boldsymbol{\theta}\right)}{f_{\gamma}\left(\boldsymbol{y}\right)} \tag{6}$$

Maximum A Posteriori estimation chooses the parameter that maximizes this posterior

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} f_{\Theta|\gamma}\left(\boldsymbol{\theta} \mid \boldsymbol{y}\right) \tag{7}$$

Because the posterior is proportional to the likelihood multiplied by the prior, MAP optimization can also be viewed as maximizing

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} \ln f_{\gamma|\Theta}\left(\boldsymbol{y} \mid \boldsymbol{\theta}\right) + \ln f_{\Theta}\left(\boldsymbol{\theta}\right) \tag{8}$$

Taking logarithms avoids numerical underflow in computation and converts products into sums, making the optimization more stable.

As more samples are observed, the posterior distribution becomes sharper, and the MAP estimate gradually approaches the true parameter values, as illustrated in Figure 2.

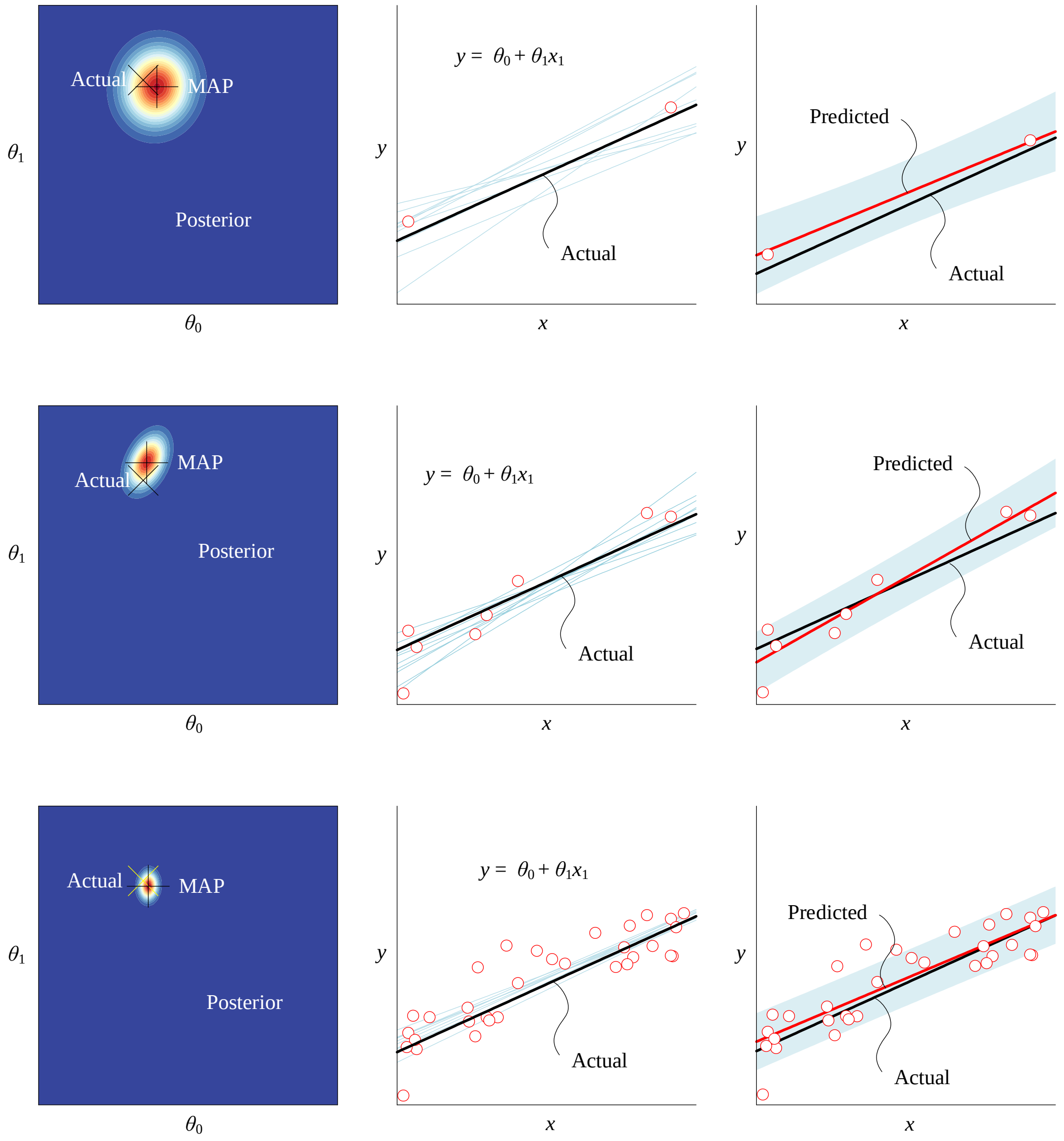

Figure 2. In Bayesian regression, the posterior distribution becomes more concentrated as additional sample data is introduced.

12.2 Bayesian Regression with Non-Informative Priors

12.2.1 Flat Priors and Maximum Likelihood

In some situations, we have no reliable prior knowledge about the model parameters. When the prior is completely unknown, a common approach is to use a non-informative prior (also called an uninformative or flat prior). A typical choice is a constant prior, which assigns equal probability to all parameter values and therefore does not bias the inference in any particular direction.

Under this assumption, the Maximum A Posteriori (MAP) estimate is obtained by maximizing the posterior distribution

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} \ln f_{\mathcal{Y}|\Theta}\left(\boldsymbol{y} \mid \boldsymbol{\theta}\right) \tag{9}$$

Because the prior is a constant, it does not affect the optimization. The MAP objective then reduces to

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} \ln f\left(\boldsymbol{y};\boldsymbol{\theta}\right) \tag{10}$$

which is exactly the same objective used in Maximum Likelihood Estimation (MLE).

12.2.2 Connection to Classical Linear Regression

Substituting the Gaussian likelihood into the objective gives

$$\begin{aligned} \ln f_{\gamma|\Theta}\left(\boldsymbol{y}\mid\boldsymbol{\theta}\right) &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y^{(i)} - \left(\theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \cdots + \theta_D x_D^{(i)}\right)\right)^2 + \underbrace{n\ln\frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{Constant}} \\ &= -\frac{\left\|\boldsymbol{y} - \boldsymbol{X\theta}\right\|_2^2}{2\sigma^2} + \underbrace{n\ln\frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{Constant}} \end{aligned} \tag{11}$$

To simplify computation, we take the negative log of the objective and ignore additive constants. The resulting optimization becomes

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}} \left\|\boldsymbol{y} - \boldsymbol{X\theta}\right\|_2^2 \tag{12}$$

This reveals an important connection: using a non-informative prior in Bayesian regression leads to the same optimization problem as Ordinary Least Squares (OLS). In other words, OLS can be viewed as a special case of Bayesian regression where no meaningful prior information is available.

12.3 Implementing Bayesian Linear Regression in PyMC

12.3.1 Setting up the Model

In this section, we use PyMC, a Python library for probabilistic programming, to implement Bayesian linear regression. PyMC provides a convenient framework for building Bayesian models by defining random variables, selecting prior distributions, and linking them to observed data. Once the model is specified, PyMC applies advanced Markov Chain Monte Carlo (MCMC) algorithms to sample from the posterior distribution, allowing us to estimate model parameters and quantify uncertainty.

Figure 3 illustrates the experiment setup. The black line represents the true underlying linear model, where the intercept is $\theta_0 = 1$ and the slope is $\theta_1 = 2$. The blue scatter points are the observed samples generated from this model, each with random noise added.

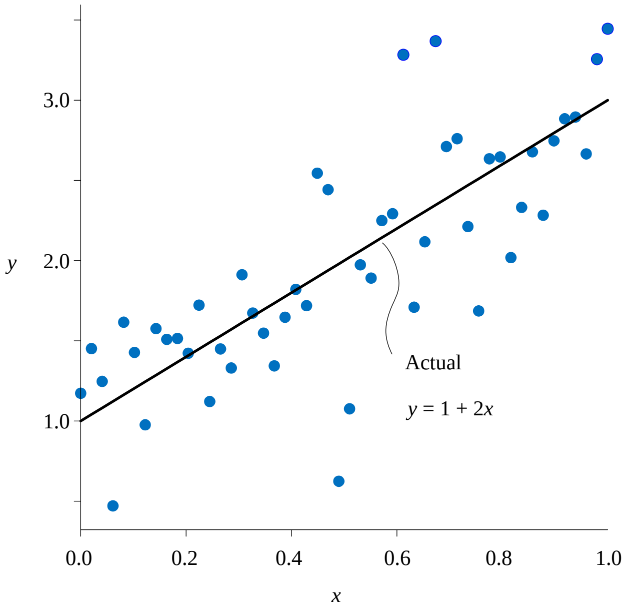

Figure 3. True linear model with observed sample points. Ch12_01_Bayesian_regression.ipynb.

12.3.2 Sampling the Posterior with MCMC

After constructing the Bayesian model in PyMC, we run MCMC to obtain posterior samples for each parameter. Figure 4 shows the MCMC trace plots. Each parameter has two chains, and only the converged portion of the trajectories is displayed. As explained earlier, the residual term ε is assumed to follow a normal distribution $N(0, \sigma^2)$, which makes σ itself a model parameter.

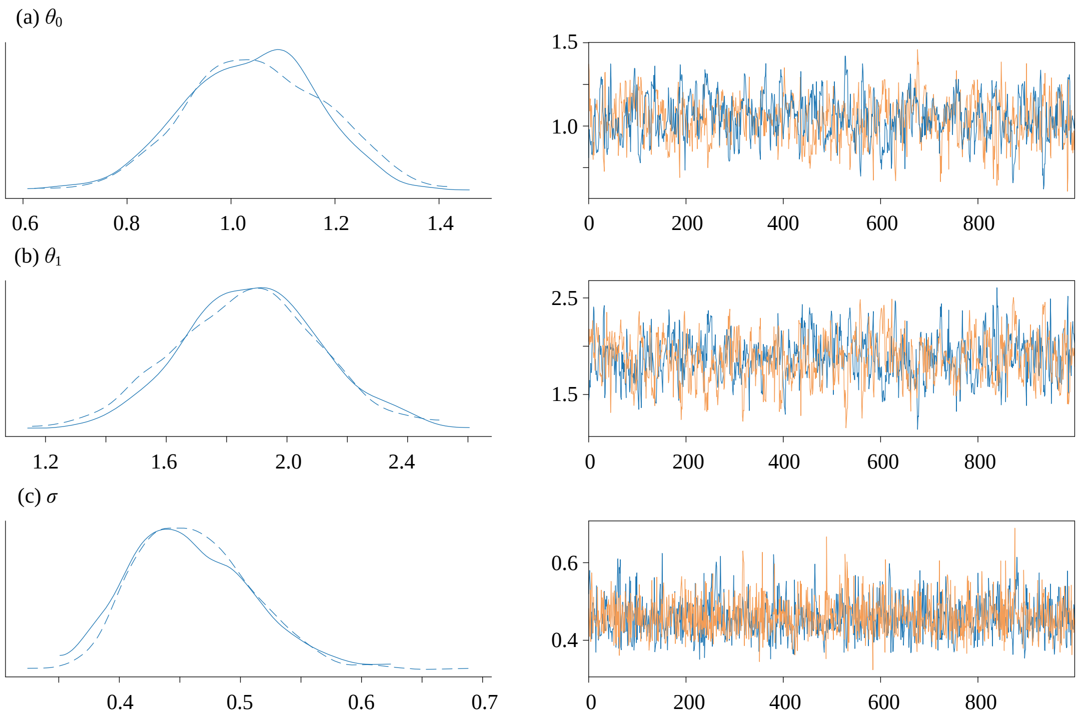

Figure 4. MCMC trace plots for θ_0, θ_1, and σ. Ch12_01_Bayesian_regression.ipynb.

In the accompanying code, the prior for σ is chosen to be a half-normal distribution, whose probability density curve is shown in Figure 5.

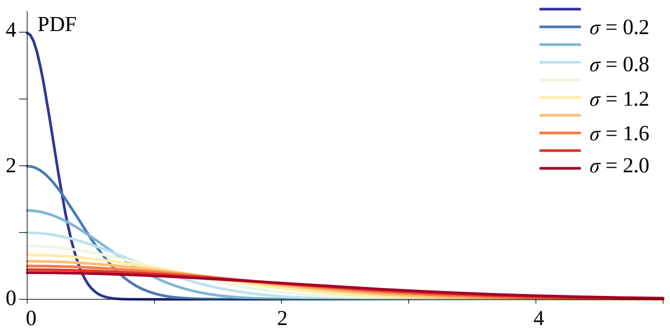

Figure 5. Half-normal prior distribution for the residual standard deviation

12.3.3 Visualizing Posterior Distributions

Figure 6 presents the posterior histograms of the parameters. The histogram merges samples from both MCMC chains. The mean of each posterior distribution can be treated as the MAP estimate, while the HDI (Highest Density Interval) provides a credible interval for the parameter. A narrower HDI indicates greater certainty in the Bayesian estimate.

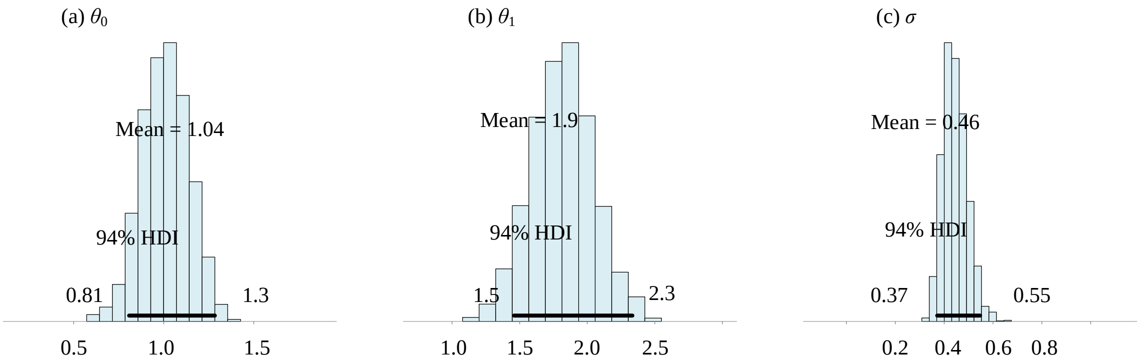

Figure 6. Posterior histograms of the model parameters. Ch12_01_Bayesian_regression.ipynb.

Figure 7 shows the joint posterior distribution of θ_0 and θ_1, visualizing their relationship and how the samples are distributed in parameter space.

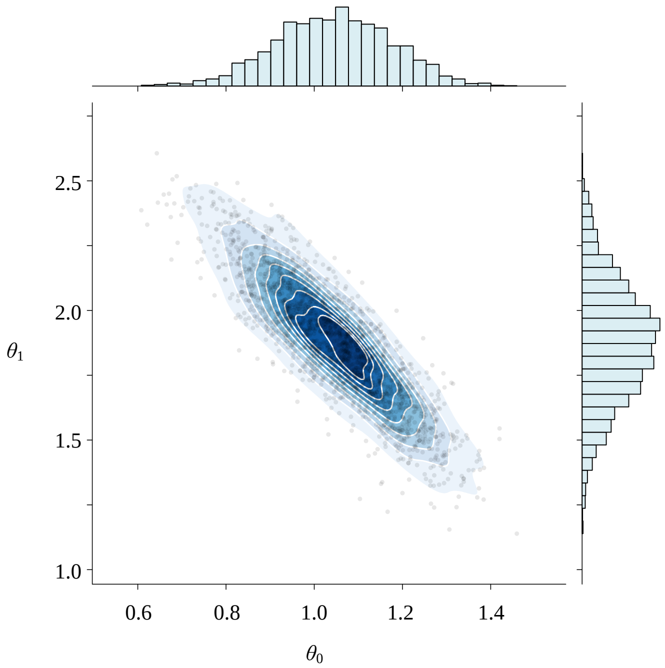

Figure 7. Joint posterior distribution of θ_0 and θ_1. Ch12_01_Bayesian_regression.ipynb.

Finally, the regression result is shown in Figure 8. The red line is the posterior mean prediction, which acts as the final regression line. The light blue lines represent fifty regression functions drawn from the posterior distribution. Each of these lines corresponds to one sampled parameter pair (θ_0, θ_1). The red line can be interpreted as the "average belief" of all these posterior samples, while the blue lines reveal the uncertainty around it, something ordinary least squares cannot express.

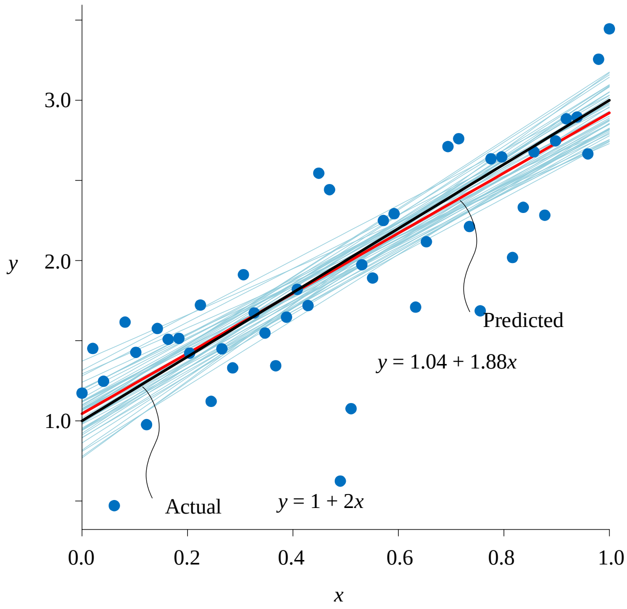

Figure 8. Bayesian linear regression fit with posterior predictive samples. Ch12_01_Bayesian_regression.ipynb.

12.4 A Bayesian Perspective on Ridge Regression

12.4.1 Gaussian Priors and Shrinkage

Ridge regression, introduced in the previous chapter, can be naturally interpreted through the lens of Bayesian inference. In the Bayesian setting, we place a prior belief on the linear regression parameters. Suppose each parameter in the weight vector $\boldsymbol{\theta}$ follows a zero-mean normal distribution,

$$f_{\Theta_j}\left(\theta_j\right) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{\theta_j^2}{2\tau^2}\right) \tag{13}$$

where τ controls how strongly we believe the weights should stay close to zero. A larger τ indicates a weaker prior (we are less certain about the parameter values), while a smaller τ represents a stronger belief that the weights should not deviate far from zero. Figure 9 shows how the shape of the prior distribution changes with different values of τ.

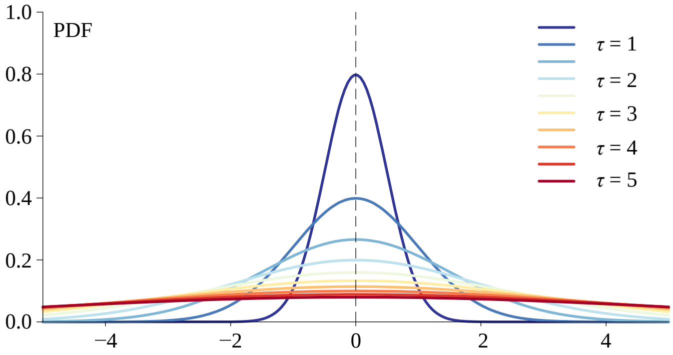

Figure 9. Effect of τ on the Gaussian prior distribution for model parameters

12.4.2 Linking MAP to Ridge Optimization

When this Gaussian prior is combined with the likelihood, the MAP optimization problem becomes

$$\arg\max_{\theta} \left[\ln \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y^{(i)} - \left(\theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \cdots + \theta_D x_D^{(i)}\right)\right)^2}{2\sigma^2} \right) + \ln \prod_{j=1}^{D} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{\theta_j^2}{2\tau^2} \right) \right] \tag{14}$$

The first term corresponds to the data-fitting objective,

$$-\frac{\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\right\|_2^2}{2\sigma^2} + n\underbrace{\ln\frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{Constant}} \tag{15}$$

while the second term represents the log of the Gaussian prior,

$$-\frac{\left\|\boldsymbol{\theta}\right\|_2^2}{2\tau^2} + D\underbrace{\ln\frac{1}{\sqrt{2\pi\tau^2}}}_{\text{Constant}} \tag{16}$$

After removing constants that do not affect optimization, equation (14) can be rewritten as

$$\underset{\boldsymbol{\theta}}{\arg\max}\left[-\frac{\left\|\boldsymbol{y}-\boldsymbol{X\theta}\right\|_2^2}{2\sigma^2}-\frac{\left\|\boldsymbol{\theta}\right\|_2^2}{2\tau^2}\right] \qquad (17)$$

The optimization above can be rewritten as

$$\underset{\boldsymbol{\theta}}{\arg\min}\,\frac{1}{2\sigma^2}\left(\left\|\boldsymbol{y}-\boldsymbol{X\theta}\right\|_2^2+\frac{\sigma^2}{\tau^2}\left\|\boldsymbol{\theta}\right\|_2^2\right) \qquad (18)$$

(18) is equivalent to:

$$\underset{\boldsymbol{\theta}}{\arg\min}\underbrace{\left\|\boldsymbol{y}-\boldsymbol{X\theta}\right\|_2^2}_{\text{OLS}}+\underbrace{\lambda\left\|\boldsymbol{\theta}\right\|_2^2}_{\text{L2 regularizer}} \qquad (19)$$

which is the familiar form of the ridge regression objective. Here, the regularization strength λ is related to the prior variance by

$$\lambda=\frac{\sigma^2}{\tau^2} \qquad (20)$$

This means that a smaller τ (a stronger prior belief and higher certainty) corresponds to a larger λ, which pushes the parameters closer to zero. In other words, the MAP solution under a Gaussian prior produces exactly the same optimization problem as ridge regression, revealing that ridge is a Bayesian estimator with a normal prior.

12.4.3 Visualizing Shrinkage Paths

Figure 10 visualizes this "shrinkage" phenomenon: as λ increases, the optimal solution moves steadily toward the origin, reflecting the influence of the prior.

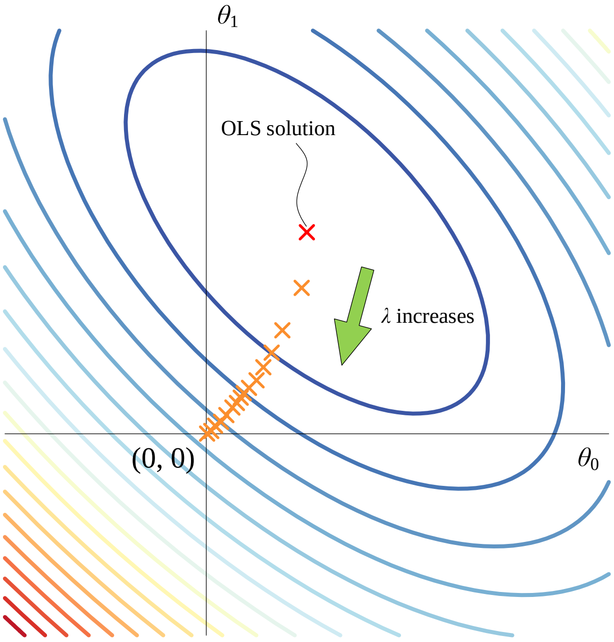

Figure 10. Shrinkage path of ridge regression as the regularization parameter λ λ increases

12.5 A Bayesian Perspective on Lasso Regression

12.5.1 Laplace Priors and Sparsity

If we place a Laplace (double exponential) prior on the regression coefficients, the Bayesian formulation leads directly to the lasso penalty. The Laplace prior for a single parameter θ can be written as

$$f_{\Theta_j}\left(\theta_j\right)=\frac{1}{2b}\exp\left(-\frac{\left|\theta_j\right|}{b}\right) \tag{21}$$

where the scale parameter $b > 0$ controls how strongly the prior concentrates mass near zero. Larger b produces a sharper peak at zero and therefore a stronger push toward sparsity; smaller b gives a flatter prior and less shrinkage. Figure 11 illustrates how the Laplace prior changes with b.

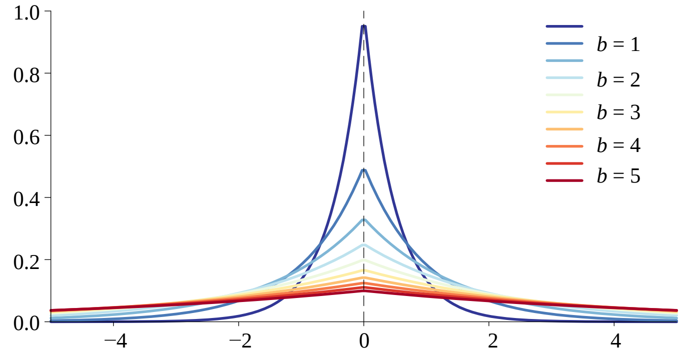

Figure 11. Laplace prior (double exponential) for different values of the scale parameter *b*.

12.5.2 MAP Optimization and Lasso Objective

Combine this prior with a Gaussian likelihood for the data. The posterior is proportional to the product of the likelihood and the Laplace prior, so taking the negative log of the posterior (and dropping constants that do not affect the optimizer) yields the objective

$$\underset{\theta}{\arg\max}\left[\ln\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(y^{(i)}-\left(\theta_0+\theta_1x_1^{(i)}+\theta_2x_2^{(i)}+\cdots+\theta_Dx_D^{(i)}\right)\right)^2}{2\sigma^2}\right)+\ln\prod_{j=1}^{D}\frac{1}{2b}\exp\left(-\frac{\left|\theta_j\right|}{b}\right)\right] \tag{22}$$

The first term also corresponds to the data-fitting objective,

$$-\frac{\left\|\boldsymbol{y}-\boldsymbol{X\theta}\right\|_2^2}{2\sigma^2}+n\ln\underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{Constant}} \tag{23}$$

while the second term represents the log of the Laplace prior,

$$-\frac{1}{b}\sum_{j=1}^{D}\left|\theta_j\right|+\underbrace{D\ln\frac{1}{2b}}_{\text{Constant}}=-\frac{1}{b}\left\|\boldsymbol{\theta}\right\|_1+\underbrace{D\ln\frac{1}{2b}}_{\text{Constant}} \tag{24}$$

The optimization can be re-written as:

$$\arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y}-\boldsymbol{X\theta}\|_2^2+\lambda\|\boldsymbol{\theta}\|_1 \tag{25}$$

In which

$$\lambda=\frac{2\sigma^2}{b} \tag{26}$$

This objective is exactly the lasso optimization problem introduced in the previous chapter.

12.5.3 Coefficient Paths under Increasing λ

The Laplace prior has a sharp peak at zero and heavier tails than a Gaussian. Intuitively, that sharp peak assigns relatively high prior mass exactly at zero while still allowing some mass away from zero. When the posterior combines this prior with data, the resulting MAP estimate tends to set many coefficients exactly to zero rather than merely shrinking them; this is why lasso performs variable selection as well as regularization.

As λ increases, coefficients are driven toward zero and some become identically zero, producing a sparse solution. Figure 12 shows a typical coefficient path under increasing λ: coefficients start near their ordinary least squares values, then move toward zero; many drop to zero at distinct points, after which they remain exactly zero as λ grows.

Because the lasso objective is not differentiable at zero, its solution path is piecewise linear and can change structure abruptly as λ varies. This contrasts with ridge regression, whose Gaussian prior yields smooth shrinkage but not exact zeros. From the Bayesian perspective, then, lasso corresponds naturally to a Laplace prior: the prior's geometry favors sparsity and produces the characteristic coefficient behavior observed in practice.

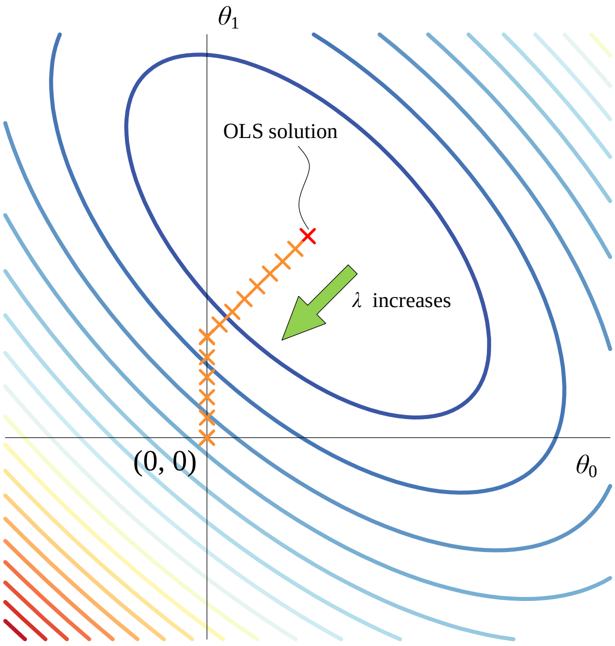

Figure 12. Lasso coefficient paths as the regularization parameter λ increases; coefficients shrink and many become exactly zero, illustrating sparsity.

12.6 Conclusion

Bayesian linear regression extends traditional regression by treating model parameters as random variables with probability distributions. Before observing data, prior beliefs describe what parameter values are considered reasonable. After data is collected, the likelihood measures how well different parameters explain the observations. Bayes inference combines the prior and the likelihood to form the posterior, which represents an updated belief about the parameters. When no useful prior knowledge is available, a flat prior makes Bayesian regression equivalent to ordinary least squares.

With PyMC, we can sample from the posterior using Markov Chain Monte Carlo and visualize both parameter uncertainty and prediction uncertainty. Ridge regression can be viewed as Bayesian regression with a normal prior, which shrinks parameters smoothly toward zero. Lasso regression corresponds to a Laplace prior, which encourages sparsity and can drive some parameters to zero. This chapter shows how Bayesian thinking unifies regression, uncertainty quantification, and regularization.