9 Regularized Regression – Taming Complexity in Machine Learning

9.1 Understanding Regularization: Why Less Can Be More

9.1.1 The Overfitting Problem

Regularization is a fundamental technique used to prevent overfitting in machine learning models. As discussed earlier, overfitting occurs when a model has too many parameters or is overly flexible, causing it to memorize noise in the training data rather than learn meaningful patterns. When this happens, the model performs well on training data but generalizes poorly to new samples.

To address this, we modify the objective function by adding a penalty term—called a regularizer—that discourages overly large parameter values. This penalty shrinks the coefficients toward zero, a phenomenon known as shrinkage. By doing so, the model becomes simpler and more robust. In this chapter, we will visualize how regularization modifies the loss surface for multivariate linear regression and how it changes the resulting regression coefficients.

The focus of this chapter is on three types of regularization: L1, L2, and a combination of L1 and L2. L1 regularization uses the L1 norm of the parameter vector, L2 uses the L2 norm, and the mixed version (Elastic Net) blends the two.

As discussed in the previous chapters, for multiple linear regression using Ordinary Least Squares (OLS), the goal is to find parameters that minimize the sum of squared errors:

$$\arg\min_{\boldsymbol{b}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 \tag{1}$$

To make the discussion concrete, consider the two-parameter case. Ignoring the bias term for simplicity, the loss surface becomes a smooth quadratic bowl, forming an elliptical paraboloid, as shown in Figure 1.

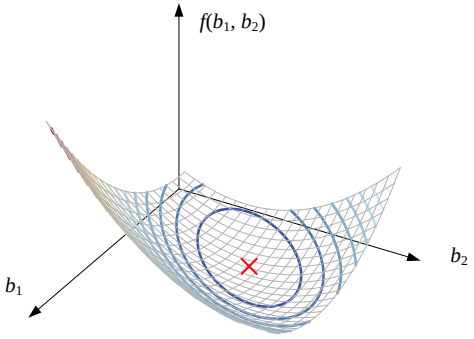

Figure 1. OLS loss surface for a two-parameter linear regression model

9.1.2 How Penalties Simplify Models

Intuitively, L2 regularization penalizes the sum of squared coefficients, gently pushing all parameters toward zero. Geometrically, it adds a circular (or spherical) constraint to the loss surface. When the OLS surface is combined with the L2 penalty, the new minimum shifts toward the origin, as illustrated in Figure 2.

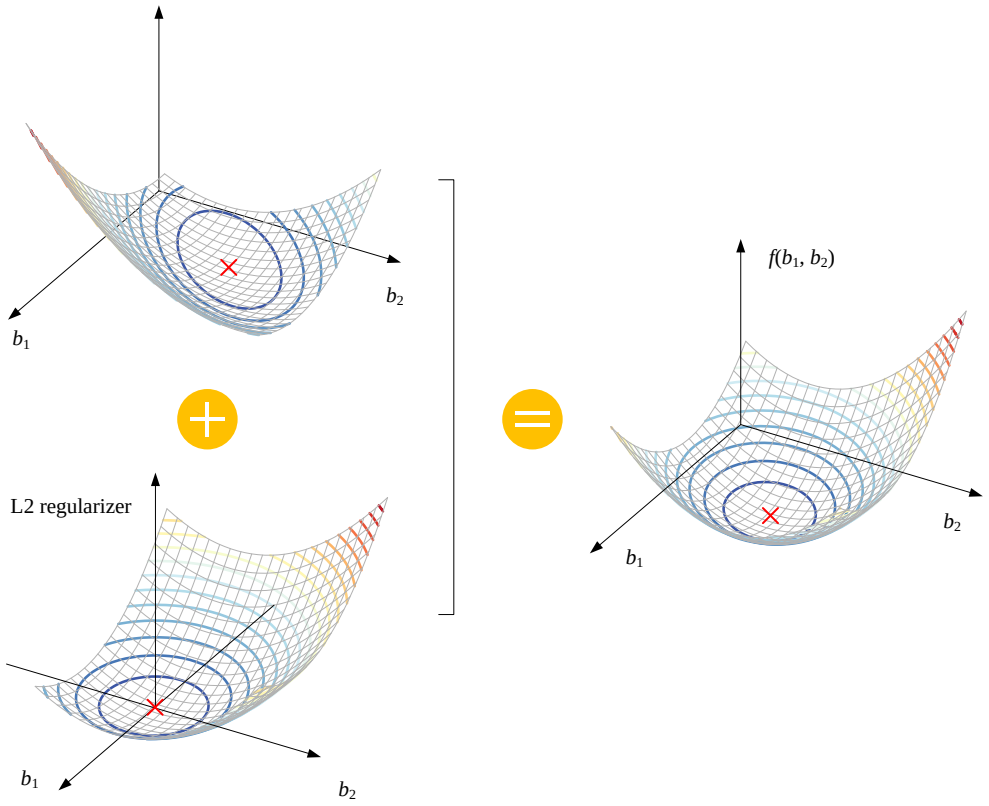

Figure 2. Ridge regression loss surface showing shrinkage toward the origin

L1 regularization penalizes the sum of the absolute values of the coefficients. Unlike L2, the L1 penalty has a diamond-shaped contour (a rotated square in 2D). When the OLS loss surface intersects this shape, the minimum often lands exactly on an axis, resulting in one or more coefficients becoming exactly zero. This gives LASSO the ability to perform feature selection, which we will explore in more depth later. Figure 3 shows the loss surface with the L1 constraint.

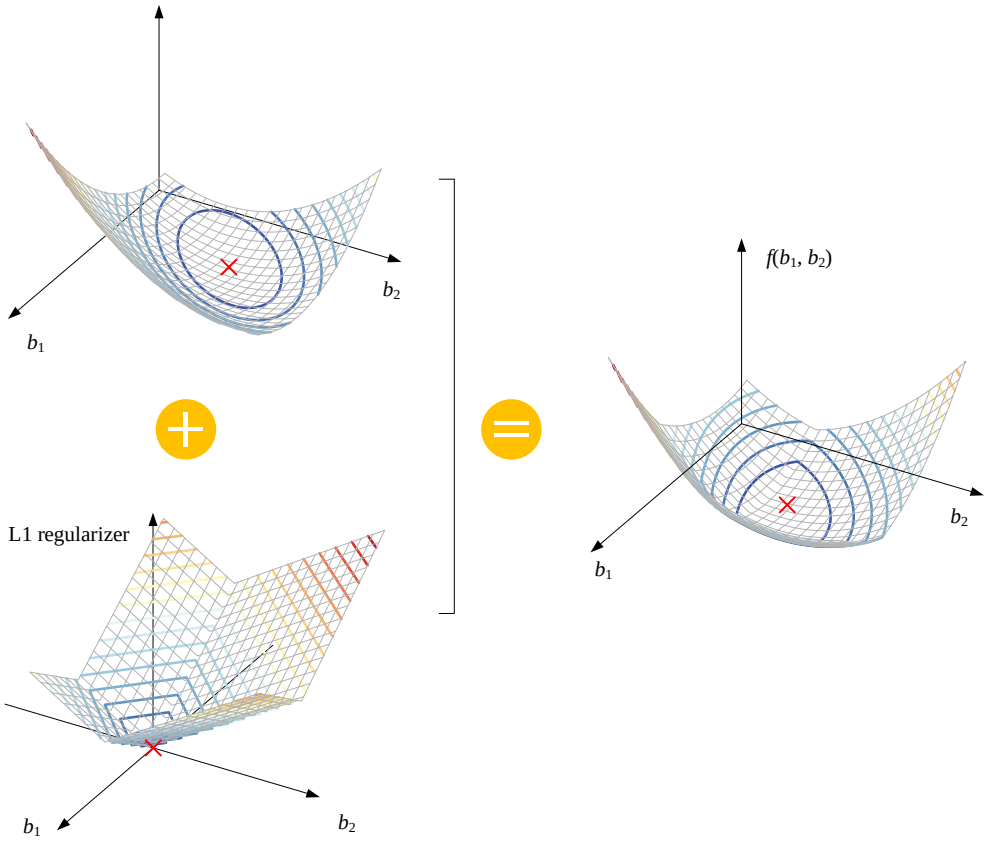

Figure 3. LASSO regression loss surface and its axis-aligned sparsity effect

Elastic Net inherits the strengths of both Ridge and LASSO: it performs shrinkage while still being capable of selecting features. Figure 4 illustrates the combined geometric effect.

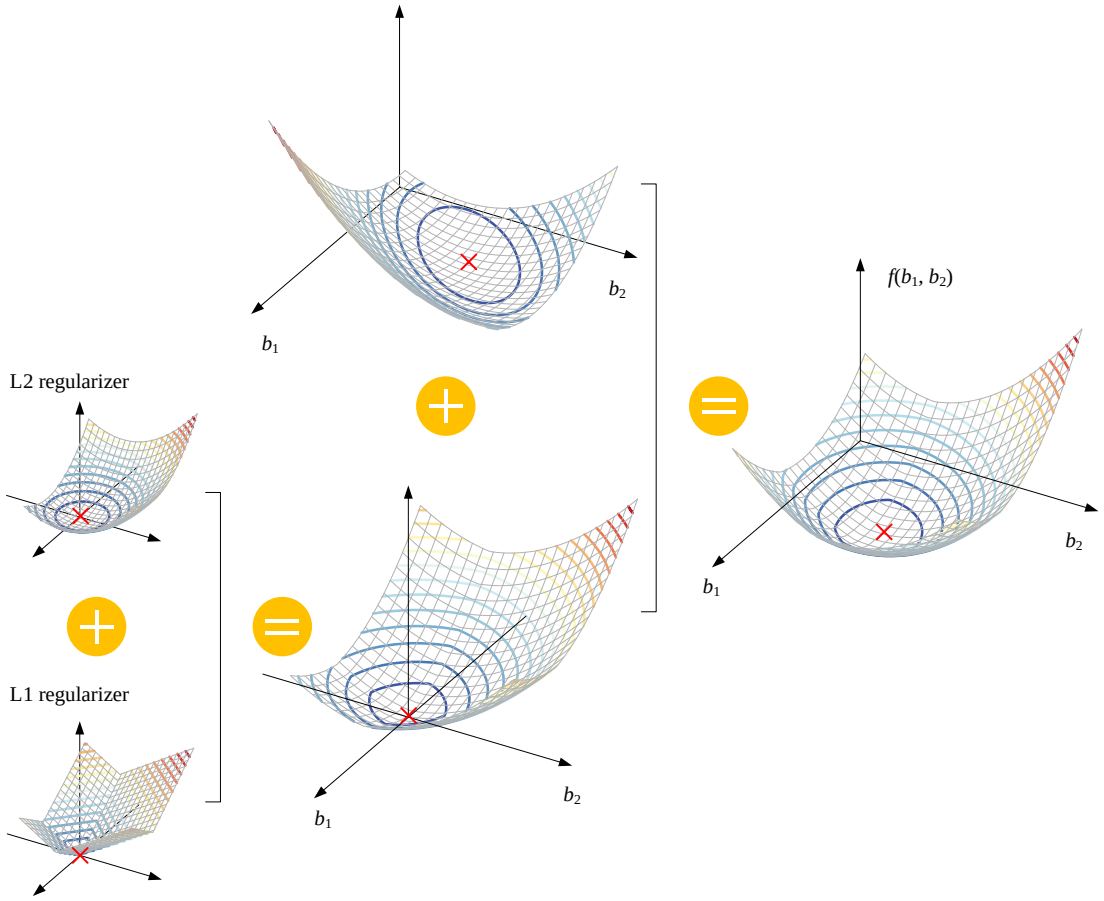

Figure 4. Elastic Net loss surface combining L1 and L2 constraints

9.1.3 Revisiting Vector Norms

In Chapter 2, we explored the concept of vector norms, particularly the Lp norm, which provides a general way to measure the length or magnitude of a vector. Before we proceed, let's briefly review what an Lp norm represents and how different values of p lead to distinct geometric interpretations.

Figure 5 illustrates how the shape of the Lp norm changes with different values of p, and why this matters in regularized regression. Each plot shows the contour of points that share the same Lp distance from the origin in a two-parameter setting.

When $p = 1$, the contour becomes a diamond, which creates sharp corners on the axes; this geometry explains why L1-based regularization (such as LASSO) tends to push coefficients to exactly zero, enabling sparsity and feature selection.

When $p = 2$, the contour is a smooth circle, matching the behavior of L2 regularization (as in Ridge regression), which shrinks parameters more uniformly without setting them to zero.

As p increases, the contours grow more square-like, and in the limit $p = \infty$, they form a perfect square. These shapes show how different norms impose different geometric constraints on the optimization process, ultimately influencing the type of solutions obtained in regularized regression.

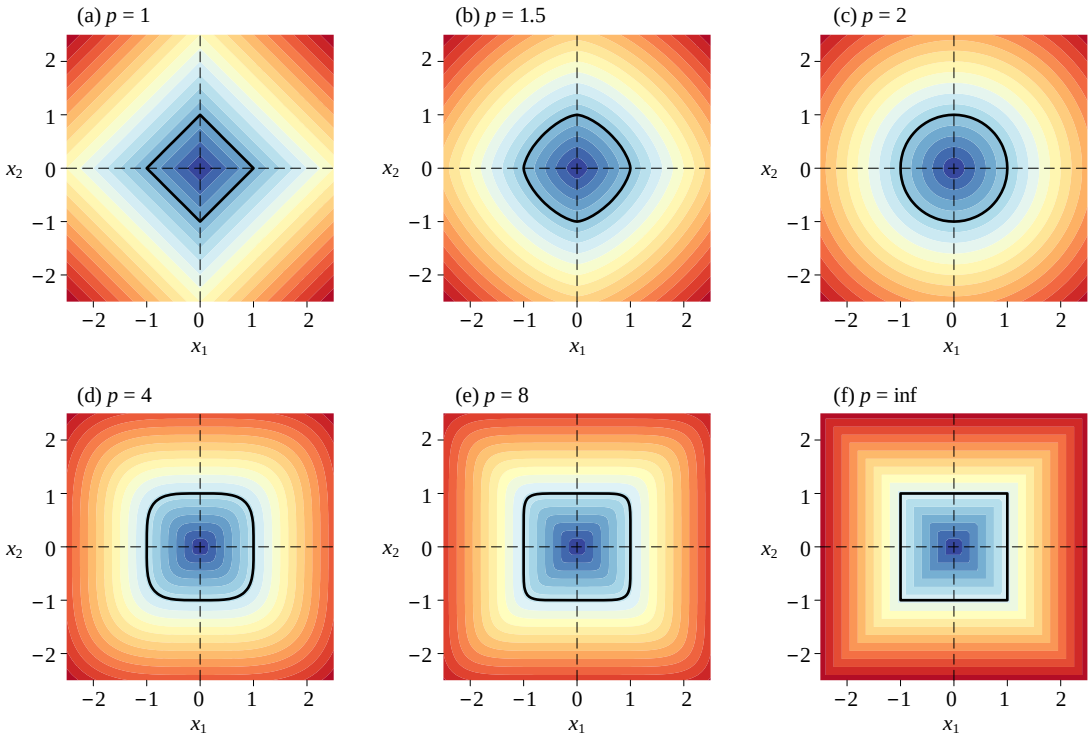

Figure 5. Contours of the Lp norm for different values of p. Note that Lp is strictly a norm only for $p \geq 1$.

9.2 Ridge Regression: Smooth Shrinkage with L2

9.2.1 Ridge Optimization Explained

As discussed earlier, Ridge Regression introduces an L2 regularization term to shrink model coefficients and reduce overfitting. The optimization objective is

$$f(\boldsymbol{b}) = \underbrace{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2}_{\text{OLS}} + \underbrace{\alpha \|\boldsymbol{b}\|_2^2}_{\text{L2 regularizer}} \tag{2}$$

To build intuition, consider the simple case with only two parameters, b b_1 and b_2, and ignore the intercept term.

Figure 6 shows the contour plot of the Ridge objective under a fixed α. The OLS loss forms an elliptical paraboloid, whose minimum is marked by the red ×.

This red point is the standard OLS solution. The L2 penalty term, on the other hand, forms a circular paraboloid with its minimum at the origin, marked by the blue ×, where both coefficients are zero. When these two quadratic surfaces are combined, the result is still a quadratic surface with a unique minimum. The Ridge solution, shown as the yellow ×, lies between the red and blue points, meaning the coefficients are shrunk toward zero but not forced to be zero.

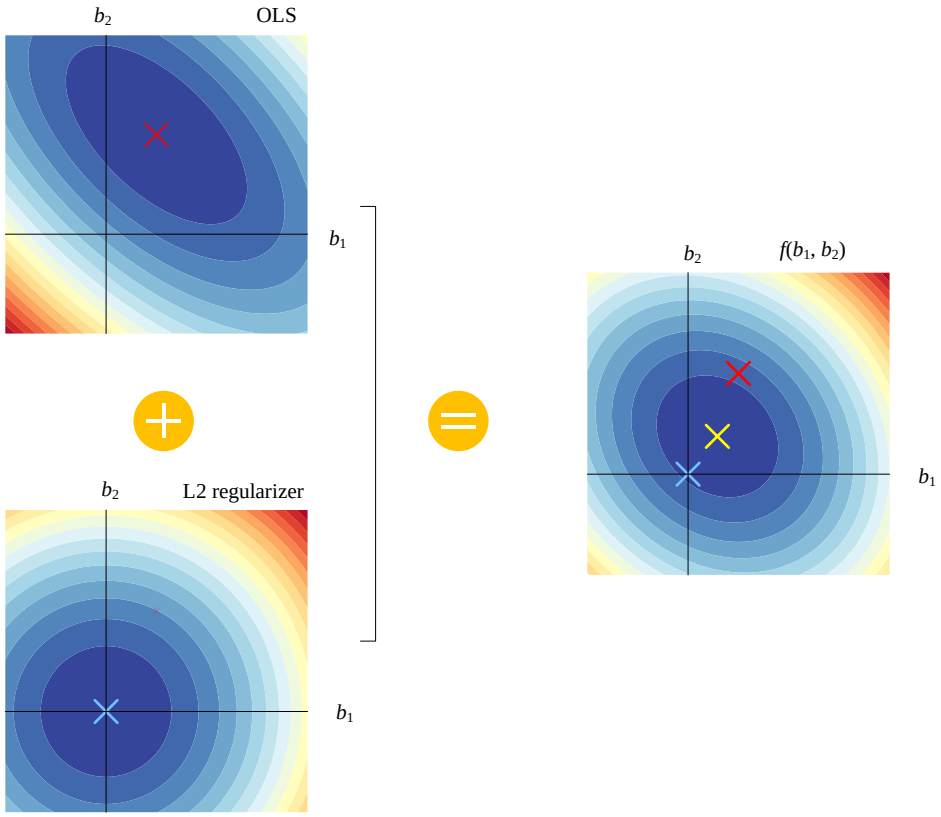

Figure 6. Ridge objective surface formed by combining the OLS loss and the L2 penalty

As the regularization strength α increases, the effect of the L2 penalty becomes more dominant. Figure 7 illustrates how the Ridge solution moves closer to the origin as α grows. The contours of the combined surface gradually transition from elongated ellipses (dominated by the OLS term) to more circular shapes (dominated by the L2 term). In the limit of very large α, the coefficients shrink heavily, but unlike LASSO, they never become exactly zero.

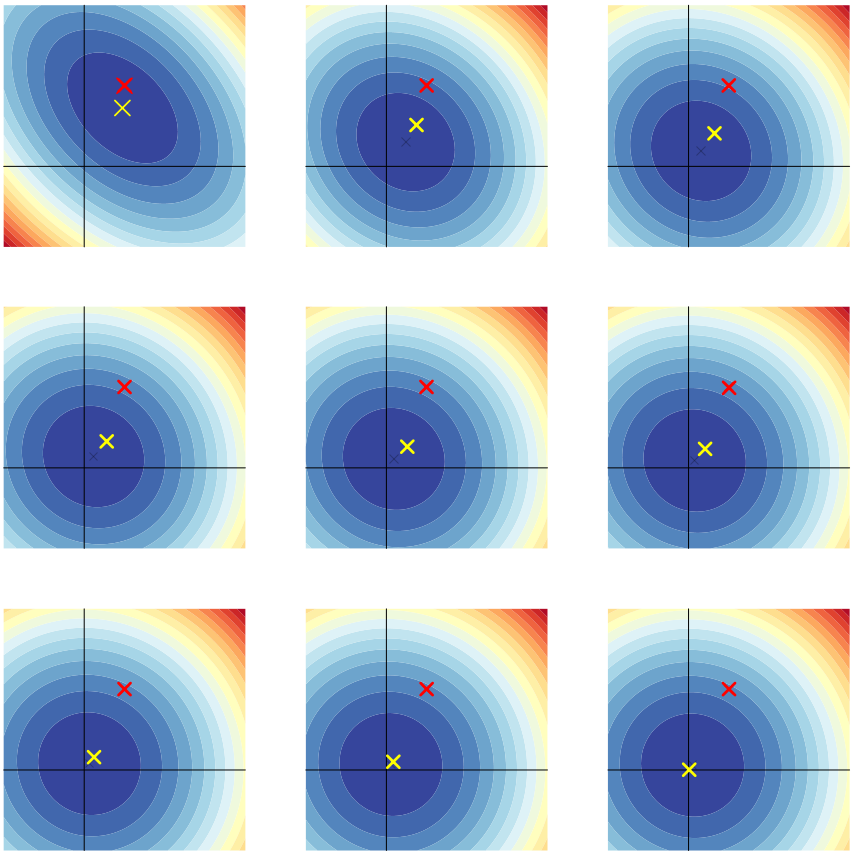

Figure 7. Movement of Ridge regression coefficients as αincreases

9.2.2 Polynomial Regression with Ridge

Figure 5 illustrates how the Ridge regularization penalty factor, denoted as α, affects the behavior of a polynomial regression model. As α increases, the fitted curve becomes noticeably smoother, indicating that the model has become less flexible and simpler. This happens because Ridge regularization introduces a penalty on large coefficient values, effectively discouraging the model from fitting noise in the data.

When the degree of the polynomial is high (in this case, fixed at 8), the model has enough capacity to overfit—capturing not only the underlying trend but also random fluctuations in the sample. Ridge regularization helps counter this problem by shrinking the magnitude of the coefficients, thus controlling model complexity. In other words, α acts as a tuning knob: small values allow a close fit to the data (potentially overfitting), while large values produce a smoother, more generalizable curve.

Table 1 lists the analytic expressions of the polynomial models under different α values, making it clear how the coefficients evolve as regularization strength changes.

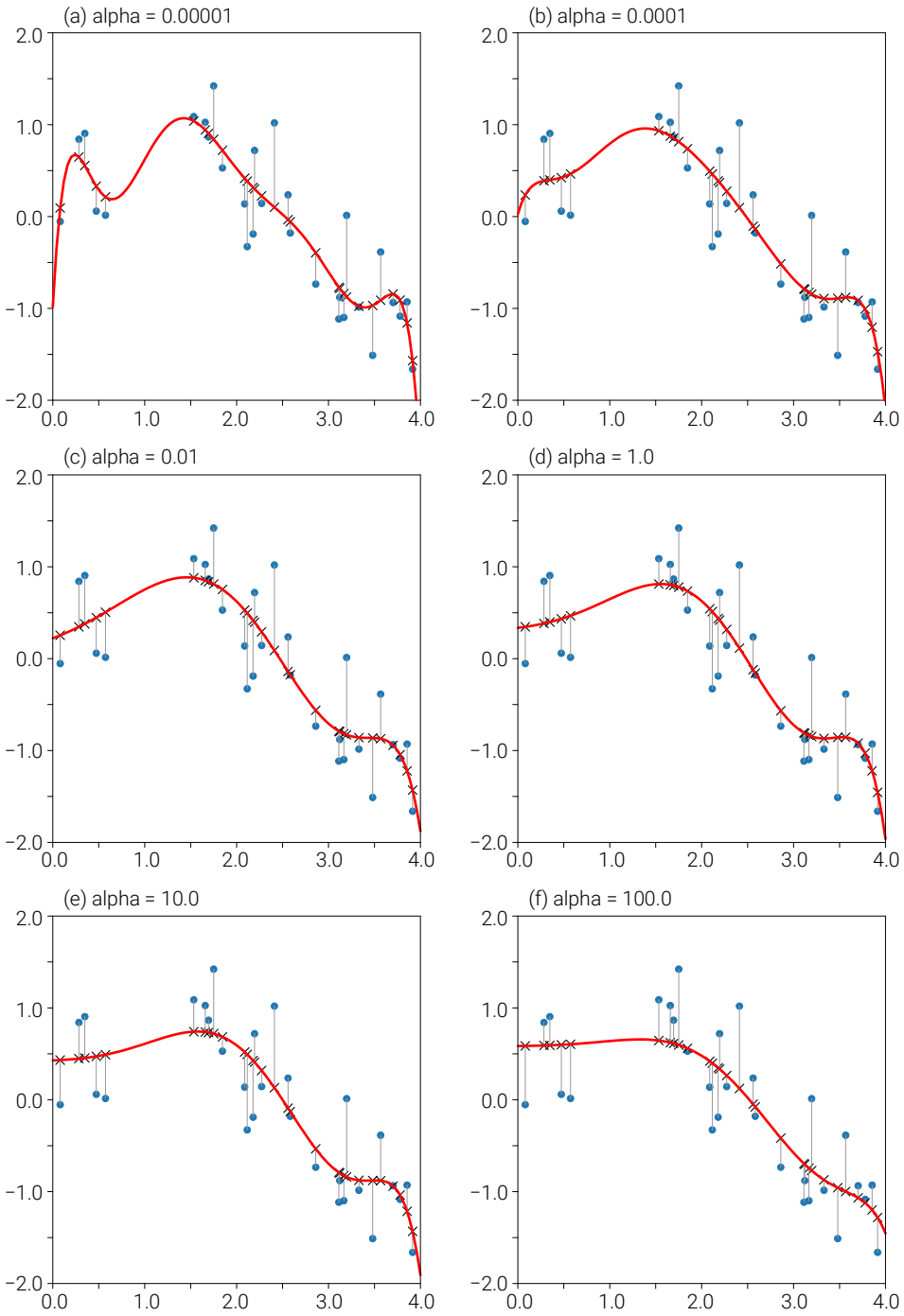

Figure 8. Effect of the Ridge penalty α on the shape of the polynomial regression curve. Figure generated by Ch09_01_Polynomial_regression_Ridge.ipynb.

Table 1. Ridge penalty α and the corresponding polynomial regression model equations.

α	Model
0.00001	$y = -0.985 + 18.400x^1 - 71.750x^2 + 122.612x^3 - 108.324x^4 + 53.620x^5 - 15.058x^6 + 2.243x^7 - 0.138x^8$
0.0001	$y = 0.026 + 3.491x^1 - 13.188x^2 + 24.668x^3 - 23.210x^4 + 12.008x^5 - 3.515x^6 + 0.547x^7 - 0.035x^8$
0.01	$y = 0.222 + 0.380x^1 + 0.149x^2 + 0.258x^3 - 0.391x^4 + 0.203x^5 - 0.093x^6 + 0.027x^7 - 0.003x^8$

1.0	$y = 0.335 + 0.125x^1 + 0.132x^2 + 0.099x^3 + 0.019x^4 - 0.048x^5 - 0.033x^6 + 0.022x^7 - 0.003x^8$
10.0	$y = 0.428 + 0.045x^1 + 0.064x^2 + 0.070x^3 + 0.049x^4 - 0.008x^5 - 0.065x^6 + 0.030x^7 - 0.004x^8$
100.0	$y = 0.585 + 0.013x^1 + 0.020x^2 + 0.024x^3 + 0.019x^4 - 0.004x^5 - 0.029x^6 + 0.013x^7 - 0.002x^8$

To better understand how the penalty factor (also called the regularization strength) affects the coefficients in a polynomial regression model, we visualize their behavior in Figure 9. The horizontal axis represents the penalty factor α on a logarithmic scale, increasing from right to left. As α becomes larger (moving leftward), most polynomial coefficients gradually shrink toward zero. This illustrates the core idea behind Ridge regularization—it penalizes large coefficients, thereby reducing model complexity and preventing overfitting.

It is important to note that the dark blue line in the figure remains fixed at zero. This line does not represent the true intercept term of the model, which is excluded from regularization by default. If you wish to see how the intercept changes with α, you can compute and plot it separately.

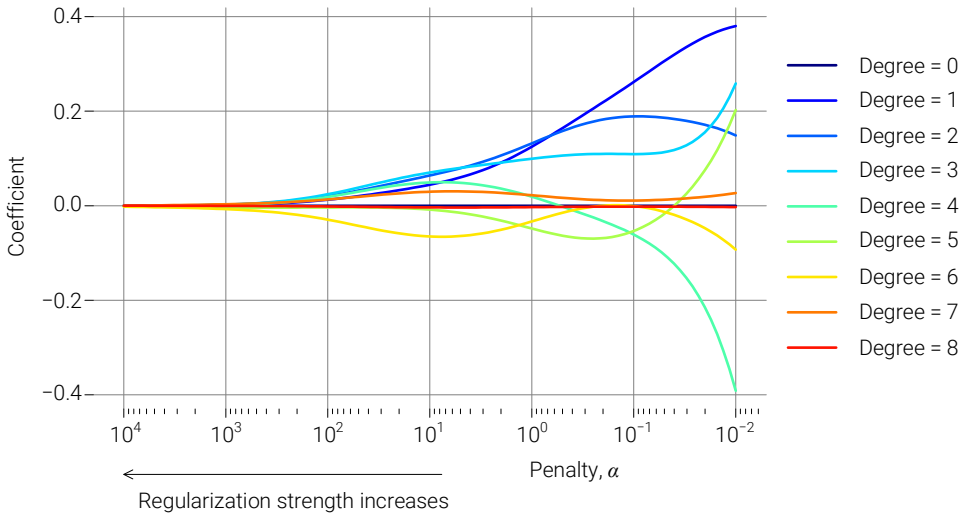

Figure 9. Variation of polynomial regression coefficients with respect to the Ridge penalty α. Figure generated by Ch09_01_Polynomial_regression_Ridge.ipynb.

9.3 LASSO Regression: Sparsity Through L1

9.3.1 LASSO Optimization Intuition

The name LASSO stands for Least Absolute Shrinkage and Selection Operator, which reflects its two key effects: shrinking coefficients and automatically selecting important features. The optimization problem for LASSO can be written as:

$$f(\boldsymbol{b}) = \underbrace{\frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2}_{\text{OLS}} + \underbrace{\alpha\|\boldsymbol{b}\|_1}_{\text{L1 regularizer}} \tag{3}$$

The first term is the familiar OLS loss, whose contour lines form ellipses. The second term is the L1 penalty, whose contour lines form a rotated diamond shape. When these two surfaces are combined, the new minimum lies closer to the origin than the OLS solution.

In the two-parameter illustration of Figure 10, the red × marks the OLS solution, the blue × marks the L1 penalty minimum at the origin, and the yellow × shows the LASSO solution, which lies between them. Because

the L1 contour has sharp corners aligned with the coordinate axes, the optimum is often pulled onto one of the axes, forcing some coefficient estimates to become exactly zero.

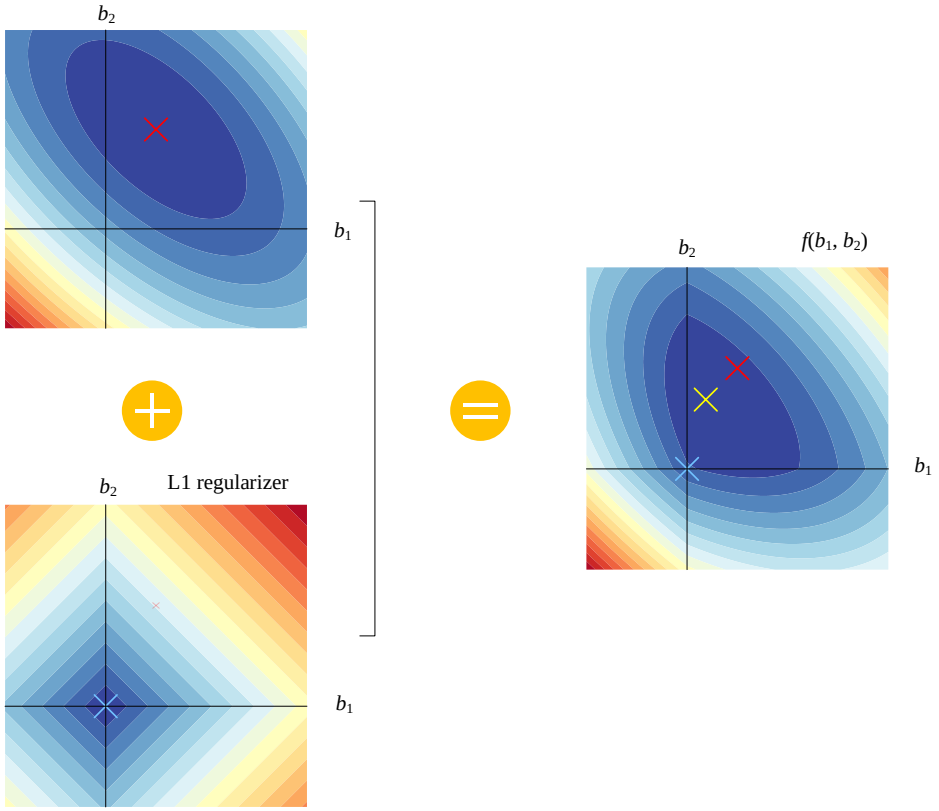

Figure 10. LASSO objective surface formed by combining the OLS loss and the L1 penalty

As the regularization strength α increases, the shrinkage effect becomes stronger. Figure 11 shows how the LASSO solution moves toward the origin as α grows. Due to the geometry of the L1 norm, LASSO can naturally drive some coefficients to zero, effectively removing those features from the model. This is why LASSO is widely used for feature selection and dimensionality reduction, especially when many features are irrelevant or redundant.

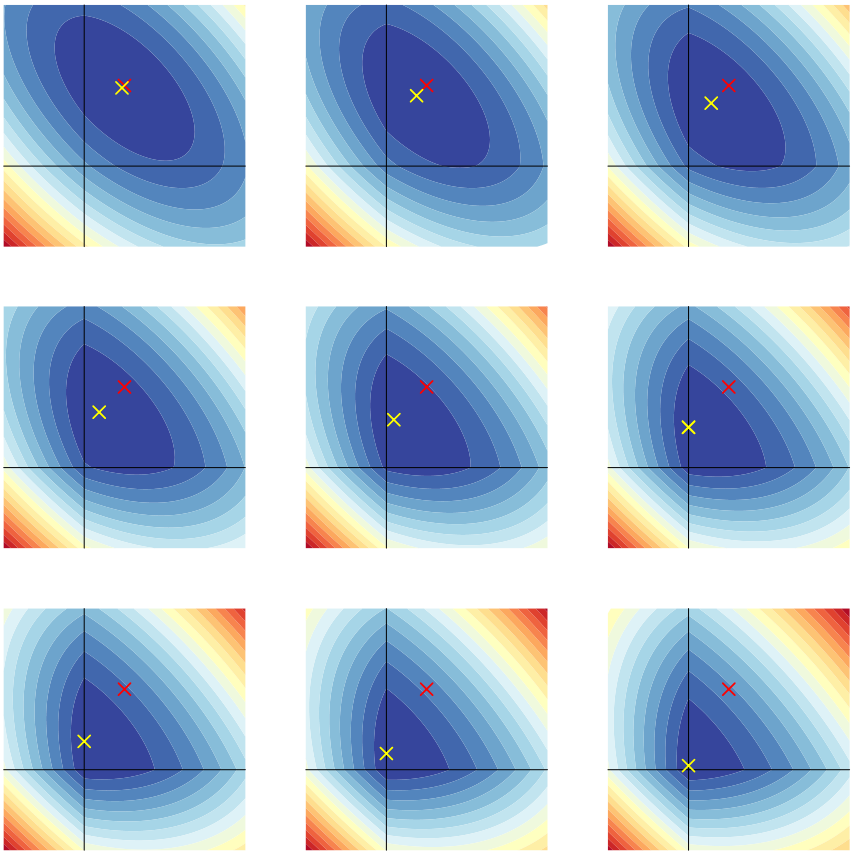

Figure 11. LASSO parameter path illustrating the effect of increasing α α

9.4 Elastic Net: Best of Both Worlds

9.4.1 Combining L1 and L2 Penalties

Elastic Net regression incorporates both L1 and L2 regularization, combining the strengths of LASSO and Ridge. The Elastic Net objective is

$$f(\boldsymbol{b}) = \underbrace{\frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}\|_2^2}_{\text{OLS}} + \underbrace{\alpha\left(\rho\|\boldsymbol{b}\|_1 + \frac{(1-\rho)}{2}\|\boldsymbol{b}\|_2^2\right)}_{\text{Elastic net regularizer}} \tag{4}$$

Here, α α controls the overall strength of regularization, while ρ determines the balance between L1 and L2 penalties. Both are user-defined hyperparameters. Figure 12 illustrates how the Elastic Net objective surface is formed by combining the two terms.

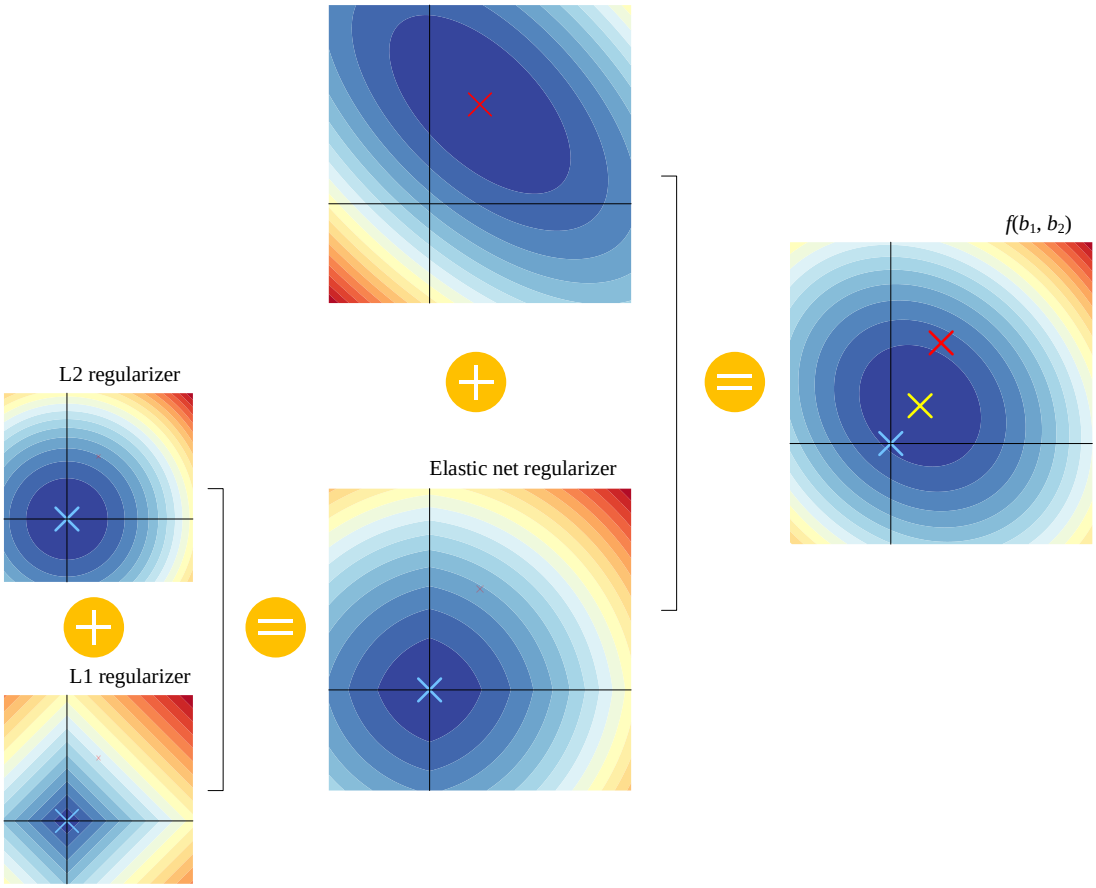

Figure 12. Elastic Net objective surface and parameter contour plot

As α increases, the Elastic Net solution shrinks toward the origin in a manner similar to LASSO and Ridge. However, the shrinkage pattern is distinct: coefficients move toward zero, and some may become exactly zero, but they generally do so more slowly than in pure LASSO. Figure 13 shows how the coefficient path changes with increasing α. Near zero, Elastic Net coefficients shrink more gradually than LASSO coefficients, reflecting the stabilizing effect of the L2 component.

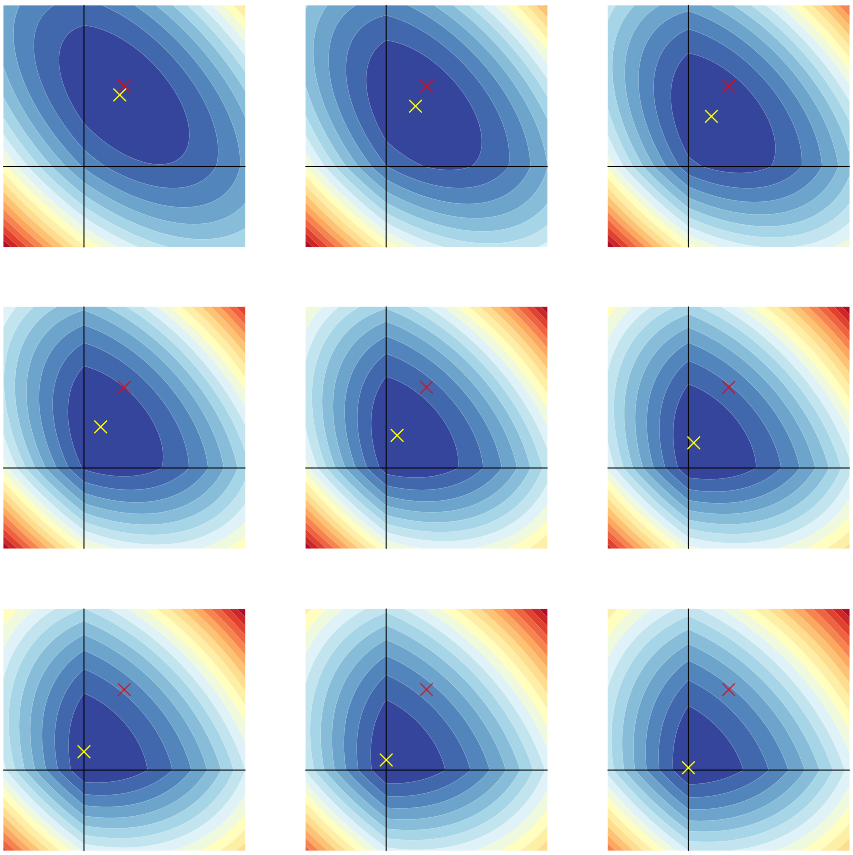

Figure 13. Movement of Elastic Net coefficients as α α increases

9.4.2 The Role of ρ in Shaping the Penalty

The parameter ρ plays a key role in shaping the regularization contour. When ρ is large, the penalty behaves more like L1, producing sharper corners in the constraint region and stronger sparsity. When ρ is small, the shape becomes smoother and more circular, behaving like L2 with weaker sparsity. Figure 14 and Figure 15 illustrate how the geometry of the Elastic Net penalty changes as ρ varies in two and three dimensions.

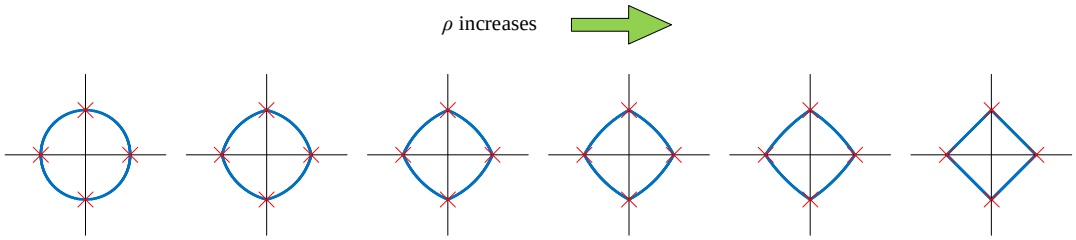

Figure 14. Effect of ρ on the 2-D Elastic Net penalty contour

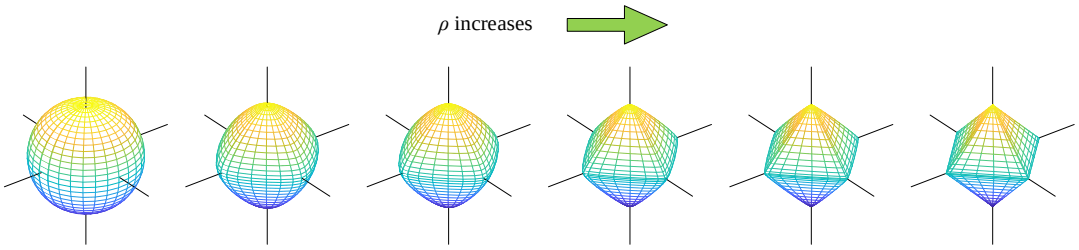

Figure 15. Effect of ρ on the 3-D Elastic Net penalty surface

9.5 Conclusion

Regularized regression is a key technique in machine learning to prevent overfitting, which occurs when a model learns noise in the training data instead of meaningful patterns. Regularization adds a penalty to the loss function, shrinking model coefficients and making the model simpler and more robust.

Ridge regression uses L2 regularization, which gently reduces all coefficients toward zero without eliminating them, while LASSO uses L1 regularization, which can push some coefficients exactly to zero, enabling feature selection.

Elastic Net combines L1 and L2 penalties, offering both shrinkage and sparsity. Geometrically, L2 creates circular constraints, L1 creates diamond-shaped constraints that encourage sparsity, and Elastic Net blends these effects. By adjusting the strength of regularization, these methods control the trade-off between fitting the training data and maintaining generalization, helping models perform well on new, unseen data while reducing complexity and focusing on the most important features.