

data significantly better than the preferred solution.

For example, we can modify the training criterion for linear regression to include **weight decay**. To perform linear regression with weight decay, we minimize a sum comprising both the mean squared error on the training and a criterion  $J(\mathbf{w})$  that expresses a preference for the weights to have smaller squared  $L^2$  norm. Specifically,

$$J(\mathbf{w}) = \text{MSE}_{\text{train}} + \lambda \mathbf{w}^\top \mathbf{w}, \quad (5.18)$$

where  $\lambda$  is a value chosen ahead of time that controls the strength of our preference for smaller weights. When  $\lambda = 0$ , we impose no preference, and larger  $\lambda$  forces the weights to become smaller. Minimizing  $J(\mathbf{w})$  results in a choice of weights that make a tradeoff between fitting the training data and being small. This gives us solutions that have a smaller slope, or put weight on fewer of the features. As an example of how we can control a model's tendency to overfit or underfit via weight decay, we can train a high-degree polynomial regression model with different values of  $\lambda$ . See figure 5.5 for the results.

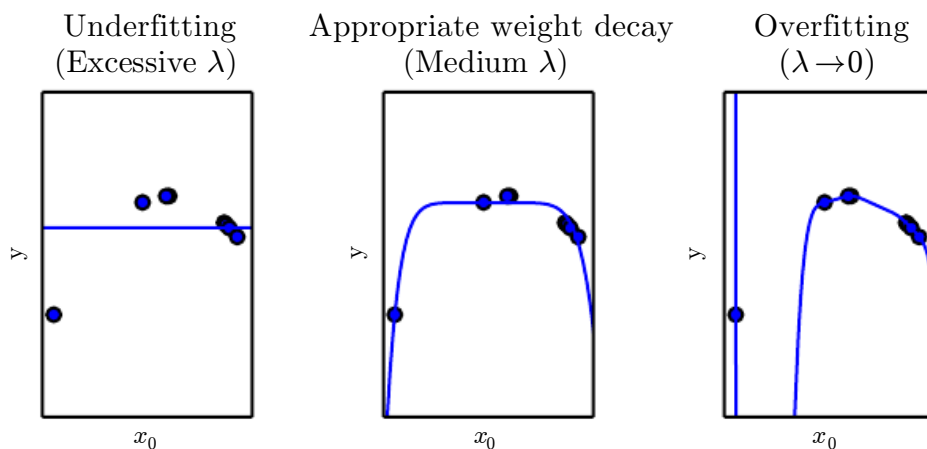


Figure 5.5: We fit a high-degree polynomial regression model to our example training set from figure 5.2. The true function is quadratic, but here we use only models with degree 9. We vary the amount of weight decay to prevent these high-degree models from overfitting. (Left) With very large  $\lambda$ , we can force the model to learn a function with no slope at all. This underfits because it can only represent a constant function. (Center) With a medium value of  $\lambda$ , the learning algorithm recovers a curve with the right general shape. Even though the model is capable of representing functions with much more complicated shape, weight decay has encouraged it to use a simpler function described by smaller coefficients. (Right) With weight decay approaching zero (i.e., using the Moore-Penrose pseudoinverse to solve the underdetermined problem with minimal regularization), the degree-9 polynomial overfits significantly, as we saw in figure 5.2.