From classification to regression

We now turn our attention more fully to the problem of *regression*, which corresponds to the supervised learning setting when $\mathcal{Y} = \mathbb{R}$. Said differently, we will not attempt to learn a discrete label anymore as in classification but a continuously changing one. Classification is a special case of regression, but the discrete nature of labels lends itself to specific insights and analysis, which is why we studied it separately. Looking at regression will require the introduction of new concepts and will allow us to obtain new insights into the learning problem.

Our regression model is that the relation between label and data is of the form $y = h(\boldsymbol{x}) + z$ with $h \in \mathcal{H}$, where $\mathcal{H}$ is a class of functions (polynomials, splines, kernels, etc.), and z is some random noise.

Linear regression corresponds to the situation in which $\mathcal{H}$ is the set of affine functions, i.e.,

$$h(\boldsymbol{x}) = \boldsymbol{\beta}^\mathsf{T}\boldsymbol{x} + \beta_0 \text{ with } \boldsymbol{\beta} = [\beta_1, \cdots, \beta_d]^\mathsf{T}$$

Least squares regression corresponds to the situation in which the we fit the parameters of our model using the *sum of squared errors* as our measure of empirical risk, e.g.,

$$\text{SSE}(\boldsymbol{\beta}, \beta_0) = \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}_i - \beta_0)^2$$

Linear least squares regression is a widely used technique in applied mathematics and can be traced back to the work of Legendre in *Nouvelles méthodes pour la détermination des orbites des comètes* (1805) and Gauss in *Theoria Motus* (1809, but privately discovered in 1795).

Georgia Tech ECE 6254 Spring 2024; Notes by M. Bloch, M. A. Davenport, and J. Romberg. Last updated 17:01, February 7, 2024

We will make a change of notation to simplify our analysis moving forward. We set

$$\boldsymbol{\theta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix} \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} 1 & -\boldsymbol{x}_1^\intercal - \\ 1 & -\boldsymbol{x}_2^\intercal - \\ \vdots & \vdots \\ 1 & -\boldsymbol{x}_n^\intercal - \end{bmatrix},$$

which allows us to rewrite the sum of squared errors as

$$\mathrm{SSE}(\boldsymbol{\theta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2.$$

One of the reasons that makes linear least squares regression so popular is the existence of a closed form analytical solution.

Lemma 1 *If $\boldsymbol{X}^\intercal\boldsymbol{X}$ is non-singular the minimizer of the SSE is*

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{y}$$

Proof See class slides. ∎

The existence of this solution is sometimes misleading because computing $\widehat{\boldsymbol{\theta}}$ can be extremely numerically unstable (i.e., the matrix $(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}$ could be ill-conditioned). We will return to this issue shortly, but we first emphasize that linear least squares methods are far more powerful than might initially seem. This is because one can use the same methodology for nonlinear regression by using a nonlinear feature map $\Phi : \mathbb{R}^d \to \mathbb{R}^{d'}$. The regression model becomes

$$\boldsymbol{y} = \boldsymbol{\beta}^\intercal\Phi(\boldsymbol{x}) + \beta_0 \text{ with } \boldsymbol{\beta} \in \mathbb{R}^{d'}. \tag{1}$$

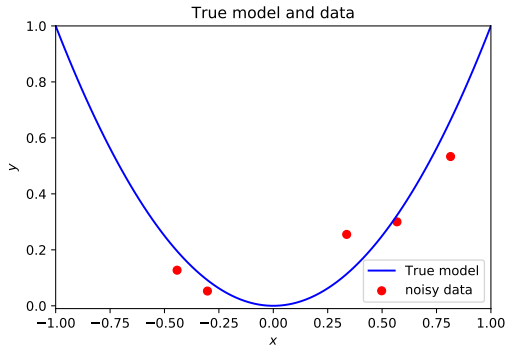
Example. To obtain a least square estimate of cubic polynomial h with $d = 1$, one can use the nonlinear map

$$\Phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \end{bmatrix}. \tag{2}$$

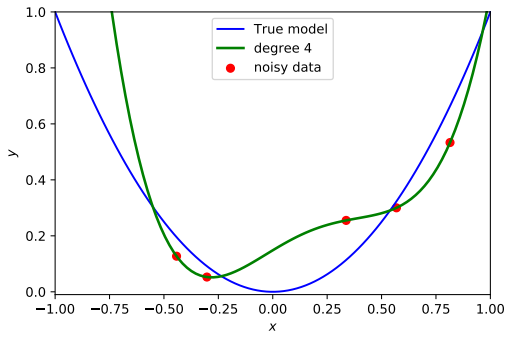
Overfitting and regularization

Overfitting is the problem that happens when fitting the data well no longer ensures that the out-of-sample error is small, i.e., the underlying model learned generalized poorly. This happens not only when there are too many degrees of freedom in model so that one "learns the noise " but also when the hypothesis set contains simpler functions than the target function h but the number of sample points n is too small. In general, *overfitting occurs as the number of features d begins to approach the number of observations n.*

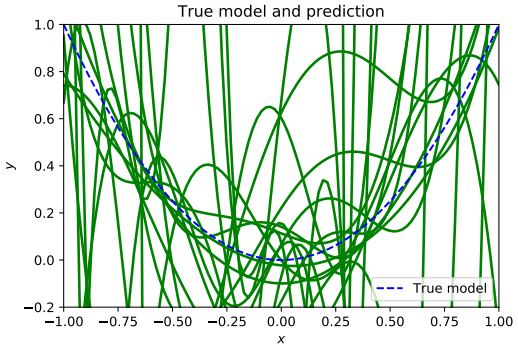
To illustrate this, consider the following example in data is generated as $y = x^2 + z$ with $x \in [-1; 1]$, where $z \sim \mathcal{N}(0, \sigma = 0.1)$. We perform regression with polynomials of degree d. Fig. 1a shows the true underlying model and five samples obtained independently and uniformly at random. Fig. 1 shows the resulting predictor obtained by fitting the data to a polynomial of degree $d = 4$. Since we only have five points, there exists a degree four polynomial that predicts exactly the value of all five training point. This is an example where our regression is effectively learning the noise in the model. To fully appreciate the consequence of overfitting, Fig. 1c shows the regression results for twenty randomly sampled sets of five points.



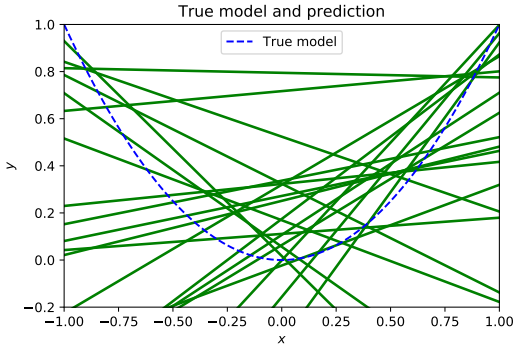
(a) True model and sample points.



(b) Regression fit with $d = 4$.



(c) Many regressions with $d = 4$ and $n = 5$.

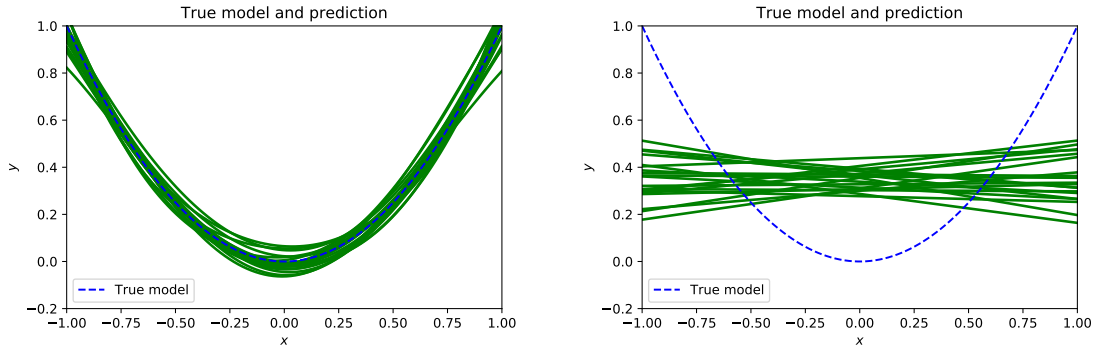


(d) Many regressions with $d = 1$ and $n = 5$.

Figure 1: Overfitting and regression with too few datapoints

Note that there is a huge *variance* in the resulting predictor, suggesting that we have an unstable prediction that does not generalize well. Note also that one observes a similar variance when trying to fig the data to a polynomial of degree $d = 1$. In the latter situation, the degree of the polynomial is one less than the true model so that the model cannot fit the noise; however, the variance stems from the fact that there are few sample points. As shown in Fig. 2a and Fig. 2b, overfitting disappears once we have enough data points.

In practice though, we are often interested in limiting overfitting even when the number of data points is small. The key solution is a



(a) Many regressions with $d = 4$ and $n = 50$.

(b) Many regressions with $d = 1$ and $n = 50$.

Figure 2: Regression with enough data points

technique called *regularization*.

Tikhonov regularization

The key idea behind regularization is to introduce a penalty term to "regularize" the vector $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \|\boldsymbol{\Gamma}\boldsymbol{\theta}\|_2^2$$

where $\boldsymbol{\Gamma} \in \mathbb{R}^{(d+1)\times(d+1)}$.

Lemma 2 *The minimizer of the least-square problem with Tikhonov regularization is*

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^\intercal \boldsymbol{X} + \boldsymbol{\Gamma}^\intercal \boldsymbol{\Gamma})^{-1} \boldsymbol{X}^\intercal \boldsymbol{y}$$

Proof See slides. ∎

For the special case $\boldsymbol{\Gamma} = \sqrt{\lambda}\mathbf{I}$ for some $\lambda > 0$, we obtain

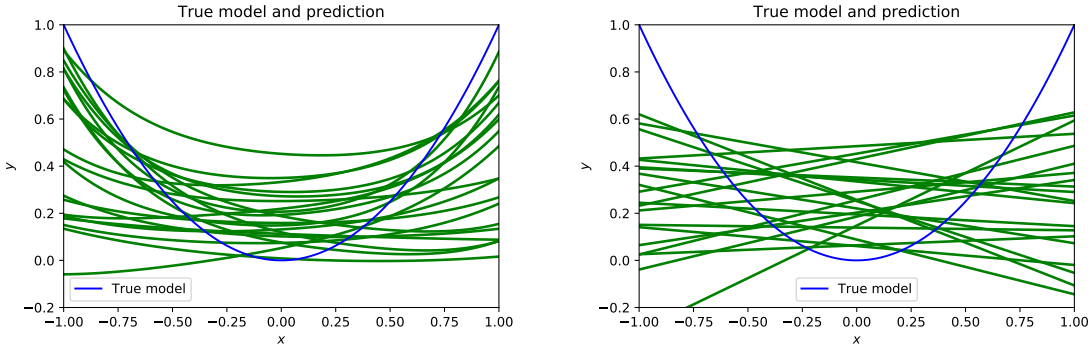
$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^\intercal \boldsymbol{X} + \lambda \mathbf{I})^{-1} \boldsymbol{X}^\intercal \boldsymbol{y}$$

This simple change has many benefits, including improving numerical stability when computing $\widehat{\boldsymbol{\theta}}$ since $\boldsymbol{X}^\intercal \boldsymbol{X} + \lambda \mathbf{I}$ is better conditioned than $\boldsymbol{X}^\intercal \boldsymbol{X}$.

Ridge regression is a slight variant of the above that does not penalize β_0 and corresponds to

$$\boldsymbol{\Gamma} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \sqrt{\lambda} \end{bmatrix}$$

To illustrate the effect of regularization, Fig. 3 shows the resulting regressions with $\lambda = 1$ in the same low-sample situation as earlier. Notice how the variance of the regression is substantially reduced.



(a) Many ridge regressions with $d = 4$ and $n = 5$.

(b) Many ridge regressions with $d = 1$ and $n = 5$.

Figure 3: Ridge regression

It is also useful to understand Tikhonov regularization as a constrained optimization problem. One can show that the minimizer of

the least-square problem with Tikhonov regularization is the solution of

$$\arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 \text{ such that } \|\boldsymbol{\Gamma}\boldsymbol{\theta}\|_2^2 \leq \tau$$

for some $\tau > 0$.

Fig. 4 illustrates the effect of Tikhonov regularization in $\mathbb{R}^2$ assuming that $\boldsymbol{\Gamma} = \boldsymbol{I}$. The Tikhonov solution is shrinked towards the zero vector to satisfy the constraint. Intuitively, the regularized solution corresponds to the point where the level set of $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2$ first intersects the feasible region $\|\boldsymbol{\theta}\|_2^2 = \tau$.

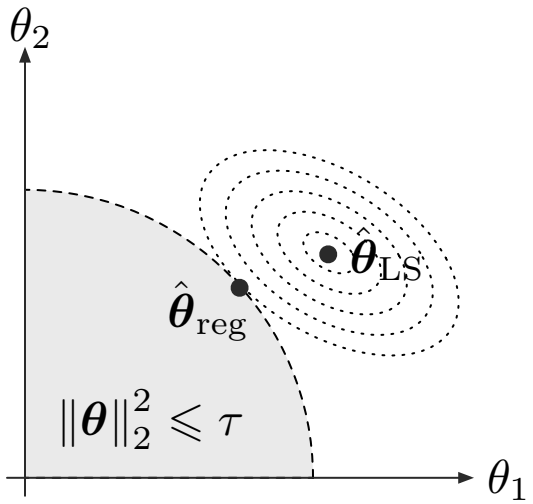


Figure 4: Illustration of Tikhonov regularization

Shrinkage estimators

The Tikhonov regularization previously introduced is a *shrinkage* estimator, in the sense that its shrinks a naive estimate towards some guess. As illustrated in Fig. 4, one can think of the regularization as shrinking the least-square estimate $\boldsymbol{\theta}_{\text{LS}}$ towards zero.

Shrinkage estimators are arguably a bit strange, especially because it may not be clear priori how biasing an estimate towards a guess would bring any benefit. The intuition you should have is that the shrinkage often leads to a lower variance of the estimator, perhaps at the expense of an increase in the bias. The example below illustrates this idea in a simple situation.

Example. Let $\{x_i\}_{i=1}^N$ be iid samples drawn according to unknown distribution with variance σ^2. Consider two estimators of the variance

$$\widehat{\sigma}^2_{\text{biased}} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \widehat{\mu})^2 \qquad \widehat{\sigma}^2_{\text{unbiased}} = \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{x}_i - \widehat{\mu})^2 \tag{3}$$

As the names suggest, it is not hard to show that

$$\mathbb{E}\,\widehat{\sigma}^2_{\text{biased}} = \frac{N-1}{N}\sigma^2 \qquad \mathbb{E}\,\widehat{\sigma}^2_{\text{unbiased}} = \sigma^2 \tag{4}$$

Perhaps surprisingly one can also show that the biased estimate has a lower variance than the unbiased one.