This set of notes is a brief tutorial on what are often called "kernel methods" in estimation and in machine learning.

1. From linear to nonlinear

Much of the second half of this class has been about applications of and algorithms for solving *linear inverse problems*, in which we are given observations $y_1, \dots, y_M$ (denoted by a vector $y \in \mathbf{R}^M$) and a model $y = A\theta + \epsilon$, where A is an $M \times N$ matrix, ϵ is random noise, and $\theta \in \mathbf{R}^N$ is some object of interest that we want to estimate.[1]

To help motivate our topic, we will recast the problem slightly as a *function estimation* problem (which relates to the theme of the first half of the class). We can think of our observations y_i as (noisy) samples $f(a_i) + \epsilon_i$ of the linear function

$$f(x) = \langle x, \theta \rangle.$$

Under this model, finding θ is equivalent to finding f.

These *linear* function models are very powerful and widely used in part because they are, by now, very well-studied and well-understood. However, it is an inconvenient fact of life that many important real-world processes are nonlinear.[2]

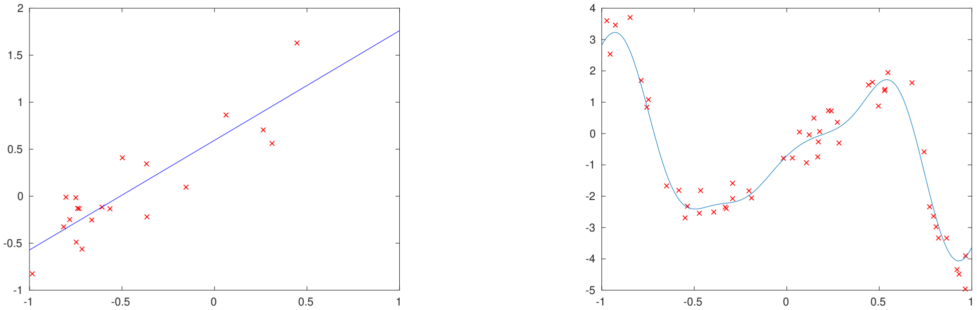

Figure 1: A good linear regression problem and a not-so-linear problem.

We would like to be able to study *nonlinear* models while still being able to use some of the powerful machinery that we have developed for linear models. The key notion that allows us to begin to do this is that our function evaluations in a linear model are *inner products*.

2. Kernels

Our first step up in complexity is to consider a model of the form

$$f(x) = \langle \Phi(x), \theta \rangle_{\mathcal{H}},$$

[1] We do not use the usual notation $x \in \mathbf{R}^N$ to avoid confusion later, when x is a point at which we evaluate a function.

[2] This is a very good thing. For example, those with an electrical/electronics engineering background may recall that semiconductor devices are useful precisely because they are nonlinear.

where now the inner product is happening in some Hilbert space $\mathcal{H}$, and Φ is a nonlinear transformation from the domain of the function to $\mathcal{H}$.

An object that will (perhaps surprisingly) be crucial is the *kernel function*

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}.$$

From the basic properties of the inner product, two important properties of k emerge:

- *Symmetry*: $k(x, y) = k(y, x)$.
- *Positive definiteness*: for any k points $x_1, \dots, x_k$, and any vector $z \in \mathbf{R}^k$,

  $$\sum_{i,j=1}^{k} z_i z_j k(x_i, x_j) \geq 0.$$

  Another way to interpret this is that the "kernel matrix" K given by $K_{ij} = k(x_i, x_j)$ is positive semidefinite.

At this point, we have little idea of what the functions Φ and k should be. However, it turns out that we can actually reverse the process described above: given a symmetric, positive definite kernel k, we can define its *reproducing kernel Hilbert space* (often abbreviated RKHS) $\mathcal{H}$ to be a Hilbert space of functions with a very special property: for any $f \in \mathcal{H}$, $f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}$. Here, $k(\cdot, x)$ is shorthand for treating $k(y, x)$ as a function of y with x fixed.

In this framework, the map Φ from above is given by $\Phi(x) = k(\cdot, x)$, and the parameter θ is simply the function f itself.

The assumption $f \in \mathcal{H}$, although much looser than assuming f is linear, is still fairly restrictive. Loosely speaking, a function $f \in \mathcal{H}$ is one that can be built up as a linear combination of kernel functions:

$$f(z) = \sum_{i=1}^{k} a_i k(z, x).$$

There are many commonly-used positive definite kernels; there is a very close connection between the kernel function k and what functions lie in its corresponding RKHS $\mathcal{H}$, so the choice of kernel often depends on what kind of function we are expecting to estimate. Two common examples are the following:

- *Gaussian radial basis function*: $k(x, y) = \exp\left(-\frac{\|x-y\|_2^2}{\ell^2}\right)$, where ℓ is a scaling parameter. These are good for representing very smooth functions.
- *Polynomial kernel*: $k(x, y) = (1 + \langle x, y \rangle)^n$. All degree-$n$ polynomials can be represented by this kernel.

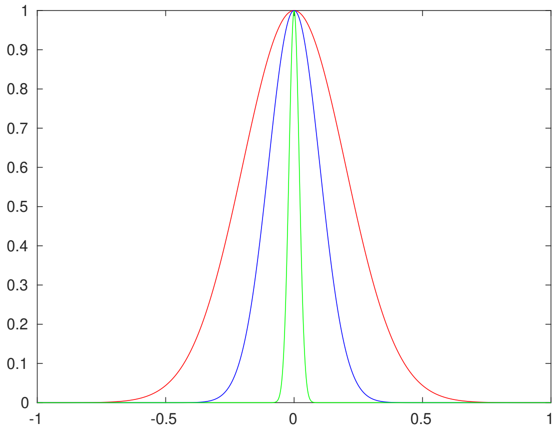

Figure 2: Gaussian RBF on $\mathbf{R}$, centered at 0, with different scale parameters.

3. Regression with Kernels

Now that we have built up a rather abstract framework for nonlinear function models, how is it useful in practice? Given $f \in \mathcal{H}$ we will consider how to estimate it given (potentially noisy) observations $y_i = f(x_i) + \epsilon_i$.

We can write $f(x_i) = \langle g_i, f \rangle_{\mathcal{H}}$, where we have abbreviated $g_i = k(\cdot, x_i)$. We aggregate these observations into the measurement operator $\mathcal{A} \colon \mathcal{H} \to \mathbf{R}^M$ given by

$$\mathcal{A}\, f = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_M) \end{bmatrix} = \begin{bmatrix} \langle g_1, f \rangle \\ \vdots \\ \langle g_M, f \rangle \end{bmatrix}$$

We try to estimate f by the following Tikhonov regularization problem:

$$\min_{f' \in H} \sum_{i=1}^{N} (y_i - \langle g_i, f' \rangle)^2 + \delta \|f'\|_{\mathcal{H}}^2 = \min_{f' \in H} \|y - \mathcal{A}\, f'\|_2^2 + \alpha \|f'\|_{\mathcal{H}}^2,$$

where $\alpha \geq 0$ is a regularization parameter (large α means a smoother function estimate). We can solve this optimization problem by setting the gradient of the objective function F equal to 0:

$$\nabla F = 2\alpha f - 2\, \mathcal{A}^*(y - \mathcal{A}\, f) = 0. \tag{1}$$

If you've seen a little bit of matrix calculus, this is the expected formula if $\mathcal{A}$ were a matrix. The *adjoint* operator to $\mathcal{A}$, which we have denoted $\mathcal{A}^* \colon \ : \mathbf{R}^M \to \mathcal{H}$, is analogous to the transpose of a matrix: we have $\langle z, \mathcal{A}\, h \rangle = \langle \mathcal{A}^*\, z, h \rangle_{\mathcal{H}}$ for any $z \in \mathbf{R}^M$ and $h \in \mathcal{H}$. For our choice of $\mathcal{A}$, it turns out that

$$\mathcal{A}^*\, z = \sum_{i=1}^{M} z_i k(\cdot, x_i).$$

There are two standard ways to solve (1). One way is to gather all the terms involving f on one side and solving, which results in the common ridge regression formula

$$\hat{f} = (\alpha \mathcal{I} + \mathcal{A}^* \mathcal{A})^{-1} \mathcal{A}^* y. \tag{2}$$

This formula can be useful for theoretical analysis, but, since it involves the inversion of an infinite-dimensional operator, it is not usually very tractable to compute directly. Furthermore, it is, in general, not even well-defined for $\alpha = 0$, since $\mathcal{A}^* \mathcal{A}$ cannot have full rank unless $\mathcal{H}$ is finite-dimensional.

Instead, note that the solution $\hat{f}$ to (1) must have the form

$$\hat{f}(x) = (\mathcal{A}^* a)(x) = \sum_{i=1}^{N} a_i k(x, x_i)$$

for some $a \in \mathbf{R}^N$. For any solution $\hat{a}$ to the equation

$$\alpha a - (y - \mathcal{A}\mathcal{A}^* a) = 0,$$

$\hat{f} = \mathcal{A}^* \hat{a}$ solves (1). We can solve this in terms of a as

$$\hat{a} = (\alpha I_N + \mathcal{A}\mathcal{A}^*)^{-1} y, \tag{3}$$

where I_N denotes the $N \times N$ identity matrix.[3] Solving for $\hat{a}$ simply (or not, if N is large) involves inverting an $N \times N$ matrix. Note that $\mathcal{A}\mathcal{A}^*$ is just the familiar kernel matrix of the set $\{g_i\}$; its (i, j)-th entry is the inner product $\langle g_i, g_j \rangle_{\mathcal{H}} = k(x_i, x_j)$.

Thus we have a practical algorithm for estimating a function f from a few of its samples; the resulting estimate is a linear combination of the shifted kernel functions $k(\cdot, x_i)$.

Finding theoretical guarantees for how good $\hat{f}$ is as an estimate of f (similarly to when we analyzed the stability of least-squares solutions) is a very active topic of research. It is beyond the scope of this single lecture, but let me know if you are interested in learning more.

[3] Note that (2) and the formula $\hat{f} = \mathcal{A}^*(\alpha I_N + \mathcal{A}\mathcal{A}^*)^{-1} y$ which we have just derived both appear in the set of notes on Tikhonov regularization. In addition, the operator $\mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}$, which we would use if we set $\alpha = 0$, is precisely the pseudoinverse of the measurement operator $\mathcal{A}$.