Worksheet 4: Hilbert Projection and Estimation with Polynomials

Name: Due October 9, 2023

The Hilbert Projection Theorem provides a way of concretely interpreting conditional expectation. In this worksheet, you will be working out a computational method for using Hilbert Projection, and this exercise will serve as foundation for implementation in the next programming assignment.

Consider joint probability space $(\mathcal{X} \times \mathcal{Y} = \mathbb{R}^2, \mathbb{P}_{\mathcal{X}\times\mathcal{Y}})$ and suppose that we would like to estimate random variable $\pi_{\mathcal{Y}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ by some polynomial in x of degree *at most n*, namely $\tilde{y}(x) = \sum_{j=0}^{n} a_j x^j$ with coefficients $a_j \in \mathbb{R}$.

1. (Prepping for Regression) Let $\mathcal{H} = \left\{ \sum_{j=0}^{n} a_j x^j : a_j \in \mathbb{R} \right\}$ be $n+1$ dimensional space of degree (at most) n polynomials. Verify that $\mathcal{H} \subset \{\tilde{y} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}\}$ is a *linear* subspace. Can you also show that $\mathcal{H}$ is also closed?

2. (Linear Regression) Replicating computation in class for the linear case, enumerate the separate conditions which allow us to find optimal $y^*(x) = \sum_{j=0}^{n} a_j^* x^j$, the point in $\mathcal{H}$ closest to $\pi_{\mathcal{Y}}$. If notation is cumbersome, try your hand first at the quadratic case (a pattern should emerge for generalizing to arbitrary degree). This problem is still called *linear* regression even when $n > 1$. Why?

3. In problem #2, you should have written $n+1$ equations in $n+1$ unknowns $a_0^*, \ldots, a_n^*$. Using matrix notation, express the solution for a^*.

$$a^* = \tag{1}$$

4. Now suppose you are provided data $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subset \mathcal{X} \times \mathcal{Y}$. Express an approximation for the terms in eq. (1) in terms of S. You do not need to write them all out individually; write an approximation for the general term and the corresponding approximation equation for a_5 of a^*.

$$a_5 = \tag{2}$$

5. Problem #4 (eq. (2) in particular) allows you to find an a_S providing *approximate* solution of eq. (1). Under what condition(s) does—and which (probabilistic) principle justifies that—$a_S \to a^*$?

6. (Bias-Variance) Let us write $y^*_{\mathcal{H}} = \mathbb{E}_{\mathcal{H}}(y|x)$ even when $\mathcal{H} \neq \{\tilde{y} : \mathcal{X} \to \mathcal{Y}\}$ (and ensure that we are clear on what $\mathcal{H}$ is!). For arbitrary $\tilde{y} \in \mathcal{H}$, we can decompose the *mean squared error* $\mathbb{E}\big((\tilde{y} - y)^2\big)$* as

$$\mathbb{E}\big((\tilde{y}-y)^2\big) = \underbrace{\mathbb{E}\big((\tilde{y}-y^*_{\mathcal{H}})^2\big)}_{\text{variance}} + \underbrace{\mathbb{E}\big((y-y^*_{\mathcal{H}})^2\big)}_{\text{bias}}.$$

Rederive this decomposition and for $\mathcal{H}_0 \subsetneq \mathcal{H}_1 \subsetneq \cdots \subsetneq \mathcal{H}_n \subsetneq \cdots$, plot a heuristic of each term against $\mathbb{N}$ (indexing $\mathcal{H}_j$). Be careful: $\tilde{y} \in \mathcal{H}_j$ for each j; therefore explain why you expect the variance term to behave as you've drawn it.

7. (RKHS: Recall HP) Let $(\mathcal{X}, \mathbb{P}_{\mathcal{X}})$ be a probability space, and $\mathcal{V} \subset \{f : \mathcal{X} \to \mathbb{R} : \|f^2\| < \infty\}$ a Hilbert space of squared integrable random variables. We say that a subspace $\mathcal{H} \subset \mathcal{V}$ is a *reproducing kernel Hilbert space* if there is function $k : \mathcal{X}^2 \to \mathbb{R}$ satisfying:
   (a) $k(\cdot, x) \in \mathcal{H}$ for each $x \in \mathcal{X}$ and
   (b) $f(x) = \langle f, k(\cdot, x)\rangle$ for each $f \in \mathcal{H}$.

   Suppose that $\mathcal{H} \subset \mathcal{V}$ is a *closed* subspace. Show that $\langle f, k\rangle = \pi_{\mathcal{H}}(f)$ for $f \in \mathcal{V}$, where $\pi_{\mathcal{H}}(f) := \arg\min_{h \in \mathcal{H}} \|f - h\|^2$.†

8. (Symmetrization) For random variable $(\mathcal{X}, \mathbb{P}_{\mathcal{X}})$, we have seen that $\mathcal{X} \times \mathcal{S} \to \mathbb{R}$ mapping $(x, s) \mapsto sx$—for $\mathcal{S} = \{1, -1\}$ with $\mathbb{P}_{\mathcal{S}}(1) = 1/2$—is symmetric. Show the same for $\mathcal{X}^2 \times \mathcal{S}^2 \to \mathbb{R}$ sending $(x_1, x_2, s_1, s_2) \mapsto s_1x_1 + s_2x_2$, namely that

$$\mathbb{P}_{\mathbb{R}}(s_1x_1 + s_2x_2 > t) = \mathbb{P}_{\mathbb{R}}(s_1x_1 + s_2x_2 < -t).$$

You may use the single symmetry version, and may have to cite more than one application of LTP/LTE.

*Of course, we are still operating with standard diagram: y may be read as evaluation of $\pi_{\mathcal{Y}}$ at (x, y) and $\tilde{y}$ as the composition $\tilde{y} \circ \pi_{\mathcal{X}}$.

†Part of this exercise is to carefully define domains / maps: $\langle \cdot, \cdot \rangle : \mathcal{V} \to \mathbb{R}$ takes inputs in $\mathcal{V}$ and values in $\mathbb{R}$. Apparently there is some looseness with notation, which part of this exercise is to tighten up.