Worksheet 4: Hilbert Projection and Estimation with Polynomials

Due October 9, 2023

Name: Aaron Zoll

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 $(\mathcal{X} \times \mathcal{Y} = \mathbf{R}^2, \mathbb{P}_{\mathcal{X}\times\mathcal{Y}})$ 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 $\bar{y}(x) = \sum_{j=0}^{n} a_j x^j$ with coefficients $a_j \in \mathbf{R}$.

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

   1) $f \equiv 0 \in \mathcal{H}$ since $\deg(f) = -\infty < n \ \forall\, n \in \mathbb{N}$

   2) $f, g \in \mathcal{H} \Rightarrow f + g = \sum_{j=0}^{n} a_j^{(f)} x^j + \sum_{j=0}^{n} a_j^{(g)} x^j = \sum_{j=0}^{n} \left(a_j^{(f)} + a_j^{(g)}\right) x^j \in \mathcal{H}$

   3) $\lambda \in \mathbb{R}, f \in \mathcal{H} \Rightarrow \lambda f = \lambda \sum_{j=0}^{n} a_j x^j = \sum_{j=0}^{n} (\lambda a_j) x^j \in \mathcal{H}$

   $\mathcal{H}$ is closed by completeness of $\mathbb{R}$ and finite dimensionality. If $\{f^n\}_{n\in\mathbb{N}} \subset \mathcal{H}$, $f^n \to f$ then $a_j^{(n)} \to a_j \ \forall\, j$ and $f = \sum_{j=0}^{n} a_j x^j \in \mathcal{H}$

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?

   Want $\forall\, k = 0, 1, \dots n \quad \mathbb{E}\left((y^* - y)x^k\right) = 0$

   $= \mathbb{E}\left(\left(\sum_{j=0}^{n} a_j^* x^j - y\right) x^k\right) = 0 \quad \Rightarrow \quad \sum_{j=0}^{n} a_j^* \mathbb{E}\left(x^{j+k}\right) = \mathbb{E}\left(yx^k\right)$

   i.e
   $a_0^* + a_1^* \mathbb{E}(x) + \dots + a_n^* \mathbb{E}(x^n) = \mathbb{E}(y)$
   $a_0^* \mathbb{E}(x) + a_1^* \mathbb{E}(x^2) + \dots + a_n^* \mathbb{E}(x^{n+1}) = \mathbb{E}(xy)$
   $a_0^* \mathbb{E}(x^2) + a_1^* \mathbb{E}(x^3) + \dots + a_n^* \mathbb{E}(x^{n+2}) = \mathbb{E}(x^2 y)$
   $\vdots$
   $a_0^* \mathbb{E}(x^n) + a_1^* \mathbb{E}(x^{n+1}) + \dots + a_n^* \mathbb{E}(x^{2n}) = \mathbb{E}(x^n y)$

   note: linear b/c problem/solution is linear. I.e. matrices

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

$$\underbrace{\begin{bmatrix} 1 & \mathbb{E}(x) & \mathbb{E}(x^2) & \dots & \mathbb{E}(x^n) \\ \mathbb{E}(x) & \mathbb{E}(x^2) & \mathbb{E}(x^3) & \dots & \mathbb{E}(x^{n+1}) \\ \vdots & & & & \\ \mathbb{E}(x^n) & \mathbb{E}(x^{n+1}) & \mathbb{E}(x^{n+2}) & \dots & \mathbb{E}(x^{2n}) \end{bmatrix}}_{A} \begin{bmatrix} a_0^* \\ a_1^* \\ \vdots \\ a_n^* \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbb{E}(y) \\ \mathbb{E}(xy) \\ \vdots \\ \mathbb{E}(x^n y) \end{bmatrix}}_{b} \qquad a^* = A^{-1} b \tag{1}$$

4. Now suppose you are provided data $S = \{(x_1, y_1), \dots, (x_m, y_m)\} \subset \mathcal{X} \times \mathcal{Y}$. 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_S of a^*.

$$\mathbb{E}(x^j y^l) \approx \frac{1}{m} \sum_{k=1}^{m} x_k^j y^l =: a_{j,l}$$

$j = 0, 1, \dots 2n$
$l = 0, 1$

$$a_S = \begin{bmatrix} a_{0,0} & a_{1,0} & \dots & a_{n,0} \\ a_{1,0} & a_{2,0} & \dots & a_{n+1,0} \\ \vdots & & \ddots & \\ a_{n,0} & & \dots & a_{2n,0} \end{bmatrix}^{-1} \begin{bmatrix} a_{0,1} \\ a_{1,1} \\ \vdots \\ a_{n,1} \end{bmatrix} \tag{2}$$

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—$a_S \to a^*$?

Finite moments + LLN
(up to $2n$)

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

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

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 $\mathbf{N}$ (indexing $\mathcal{H}_j$). Be careful: $\hat{g} \in \mathcal{H}_j$ for each j; therefore explain why you expect the variance term to behave as you've drawn it.

$$\mathbf{E}((\tilde{g}-y)^2) = \mathbf{E}\left((\tilde{g}-y^*_{\mathcal{H}}+y^*_{\mathcal{H}}-y)^2\right) = \mathbf{E}\left((\tilde{g}-y^*_{\mathcal{H}})^2 - 2\underbrace{(\tilde{g}-y^*_{\mathcal{H}})}_{:=h\in\mathcal{H}}\underbrace{(y-y^*_{\mathcal{H}})}_{:=h^\perp\in\mathcal{H}^\perp} + (y-y^*_{\mathcal{H}})^2\right)$$

$$= \mathbf{E}((\tilde{g}-y^*_{\mathcal{H}})^2) - 2\cancel{\mathbf{E}(hh^\perp)} + \mathbf{E}((y-y^*_{\mathcal{H}})^2)$$

Variance does this b/c adding more moments/terms to sum, and the square make it more positive

7. (RKHS: Recall HP) Let $(\mathcal{X}, \mathbf{P}_{\mathcal{X}})$ be a probability space, and $\mathcal{V} \subset \{f : \mathcal{X} \to \mathbf{R} : \|f^2\| < \infty\}$ 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 $k : \mathcal{X}^2 \to \mathbf{R}$ 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 $f \in \mathcal{H}$. $\Rightarrow f = \langle f, k\rangle$

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 $\pi_{\mathcal{H}}(f) := \arg\min_{h\in\mathcal{H}} \|f-h\|^2$.†

$$\forall\, h \in \mathcal{H},\ \langle f - \langle f,k\rangle, h\rangle = \int_{\mathcal{X}} f(x)\int_{\mathcal{X}} h(y)\,k(x,y)\,d\mathbf{P}_{\mathcal{X}}(y)\,d\mathbf{P}_{\mathcal{X}}(x) - \int_{\mathcal{X}} f(x)\,k(x,y)\int_{\mathcal{X}} h(y)\,d\mathbf{P}_{\mathcal{X}}(y)\,d\mathbf{P}_{\mathcal{X}}(x)$$

$$= \langle f, h\rangle - \langle\langle f,k\rangle, h\rangle = \int_{\mathcal{X}^2} f(x)h(y)\,k(x,y)\,d\mathbf{P}_{\mathcal{X}}(y)\,d\mathbf{P}_{\mathcal{X}}(x) - \int_{\mathcal{X}^2} f(x)\,k(x,y)\,h(y)\,d\mathbf{P}_{\mathcal{X}}(y)\,d\mathbf{P}_{\mathcal{X}}(x)$$

$$= \langle f, \langle h,k\rangle\rangle - \langle\langle f,k\rangle, h\rangle = 0 \Rightarrow f - \langle f,k\rangle \in \mathcal{H}^\perp \Rightarrow \langle f,\lambda\rangle = \operatorname*{argmin}_{h\in\mathcal{H}}\|f-h\|^2 =: \pi_{\mathcal{H}}(f)$$

def $\langle\cdot,\cdot\rangle : \mathcal{V}^2 \to \mathbf{R}$
$\langle f, g\rangle \mapsto \int_{\mathcal{X}} f g\, d\mathbf{P}_{\mathcal{X}}(x)$

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

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

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

$$\mathbf{P}_{\mathcal{X}^2\times\mathcal{S}^2}(s_1x_1+s_2x_2 > t) = \int_{\mathcal{S}^2}\int_{s_1x_1+s_2x_2>t} d\mathbf{P}_{\mathcal{X}^2|\mathcal{S}^2}(x_1,x_2|s_1,s_2)\,d\mathbf{P}_{\mathcal{S}^2}(s_1,s_2)$$

$$= \frac{1}{4}\mathbf{P}_{\mathcal{X}^2|\mathcal{S}^2}((1)x_1+(1)x_2 > t \mid s_1=s_2=1) + \frac{1}{4}\mathbf{P}_{\mathcal{X}^2|\mathcal{S}^2}(x_1-x_2>t \mid s_1=1, s_2=-1)$$
$$+ \frac{1}{4}\mathbf{P}_{\mathcal{X}^2|\mathcal{S}^2}(-x_1+x_2>t \mid s_1=-1, s_2=1) + \frac{1}{4}\mathbf{P}_{\mathcal{X}^2|\mathcal{S}^2}(-x_1-x_2>t \mid s_1=-1, s_2=-1)$$

$$= \frac{1}{4}\mathbf{P}_{\mathcal{X}^2|\mathcal{S}^2}(-x_1-x_2<-t \mid s_1=-1, s_2=-1) + \frac{1}{4}\mathbf{P}_{\mathcal{X}^2|\mathcal{S}^2}(-x_1+x_2<-t \mid s_1=-1, s_2=1)$$
$$+ \frac{1}{4}\mathbf{P}_{\mathcal{X}^2|\mathcal{S}^2}(x_1-x_2<-t \mid s_1=1, s_2=-1) + \frac{1}{4}\mathbf{P}_{\mathcal{X}^2|\mathcal{S}^2}(x_1+x_2<-t \mid s_1=1, s_2=1)$$

$$= \mathbf{P}_{\mathcal{X}^2\times\mathcal{S}^2}(s_1x_1+s_2x_2 < -t)$$

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

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