Learning from Non-Random Data in Hilbert Spaces: An Optimal Recovery Perspective

Chunyang Liao
Department of Mathematics, Texas A&M University
(Joint work with Simon Foucart, Shahin Shahrampour, Yinsong Wang)

Main Results

Compute Local Worst Case Error:
- When \( \mathcal{H} \) is a finite dimensional Hilbert space \( (\dim(\mathcal{H}) = N) \), computing local worst-case error can be cast as a semi-definite programming by using S-lemma.

Optimal Recovery Map:
- Let \( \mathcal{H} \) be the unitary map \( x \in \mathbb{C}^n \rightarrow \sum_{k=1}^{m} x_k h_k \in \mathcal{H} \).

Theorem 1: The local worst-case error of an optimal learning map \( R : \mathbb{C}^n \rightarrow \mathcal{H} \), is the minimal value of the following program, in which \( w := R(y) \):

\[
\begin{align*}
\min_{c \in \mathbb{C}, d \geq 0} & \quad c + d \parallel \mathcal{H} \parallel \left( \parallel w - f_{\text{loc}}(y) \parallel \mathcal{H} \right) \\
\text{subject to} & \quad \parallel (w - f_{\text{loc}}(y))_{\mathcal{H}} \parallel \mathcal{H} \\ & \quad \mathcal{H} = \{ f \in \mathbb{C}^n : \parallel f - P_{\mathcal{H}}(y) \parallel \mathcal{H} \leq \epsilon \}.
\end{align*}
\]

Optimal Recovery Map:
- Let \( u_i \in \mathcal{H} \) be the Riesz Representers of the linear functional \( f_i \in \mathcal{H}^* \).
- Define the Gramian matrix \( G \in \mathbb{C}^{m \times m} \) with \( G_{ij} = \langle u_i, u_j \rangle \) for all \( i, j \in \{1 : m\} \).
- Define the cross-Gramian \( C \in \mathbb{C}^{m \times n} \) with \( C_{ij} = \langle u_i, v_j \rangle \) for all \( i \in \{1 : m\}, j \in \{1 : n\} \).

Theorem 2: The local optimal learning/recognition map \( R_{\text{opt}}(y) \) is linear and has closed form for each \( y \in \mathbb{C}^n \):

\[
R_{\text{opt}}(y) = \sum_{j=1}^{m} a_j u_j + \sum_{j=1}^{n} b_j v_j,
\]

where the coefficient vectors \( a \in \mathbb{C}^m \) and \( b \in \mathbb{C}^n \) are computed as

\[
\begin{align*}
b &= (C^* A^{-1} C)^{-1} C^* G^{-1} y, \\
a &= G^{-1}(y - Cb).
\end{align*}
\]

Numerical Experiments

Worst-case Error Comparison:
- Assume \( f \in \mathbb{C}^n \) with 50 observations and \( \dim(V) = 20 \).
- Compare worst-case errors produced by optimal recovery and an empirical risk minimization with \( \rho = 1 \) and \( \rho = 2 \).

Generalization Error Comparison:
- Assume \( y \in \mathcal{H}_K \), where \( \mathcal{H}_K \) is the RKHS associated with Gaussian kernel.
- \( V \) is spanned by a subset of Taylor features of order \( k = 1 \).
- To make data non-IID, we sort data according to their 5-th feature in a descending order and select the top 70% as training set and bottom 30% as test set.
- Compare optimal recovery map to kernel ridgeless regression (in \( \mathcal{H}_K \)) and Taylor features regression (in \( V \)).

Performance: Optimal recovery map also has the potential to generalize well when data points are non-IID random variables.

Future Directions

Our findings revealed an interesting connection with current machine learning methods. There are many directions to consider in the future, including:
- learning approximation space \( V \) from data;
- developing optimal recovery with noise/error in the observations;
- studying the overparameterized regime \( \dim(V) > m \);
- investigating the case where the approximation space \( V \) is not a linear space.

Acknowledgement: C. L. and Y. W. are supported by the Texas A&M Triads for Transformation (T3) Program. S. F. is partially supported by NSF grants DMS-1622134 and DMS-1664893. S. F. and S. S. also acknowledge NSF grant CCF-1934904.