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)

Motivation

Generalization in classical statistical learning is often attached to the postulate that data points are independent and identically distributed (IID) random variables. But this postulate may not hold in many applications. It encourages the development of learning frameworks that are robust to non-IID data. In this work, we consider learning problem from an optimal recovery perspective, which does not rely on any probabilistic assumption on the data.

Optimal Recovery Perspective

- Unknown relations f_0 are required through $y_i = \ell_i(f_0), i \in [1:m]$, where $\ell_i \in \mathcal{H}^\star$ are point evaluation functionals.
- Assume $f_0 \in \mathcal{H}$ where $\mathcal{H}$ is a Hilbert space.
- Define observation operator $L : \mathcal{H} \to \mathbb{R}^m$ by $L(f) = [\ell_1(f); \dots; \ell_m(f)]$.
- Prior information on f_0 is $f_0 \in \mathcal{K}$, where
$$\mathcal{K} = \{f \in \mathcal{H} : \|f - P_V f\| \le \epsilon\}, \tag{1}$$
is called model set and V is a linear subspace of $\mathcal{H}$ ($n := \dim(V) \le m$).
- Recovery map $R : \mathbb{R}^m \to \mathcal{H}$ returns $R(\mathbf{y})$ as an estimator for f_0.
- Assessed by local/global worst case error
$$\mathrm{err}^{\mathrm{loc}}_{\mathcal{K}}(L, R(\mathbf{y})) := \sup_{f \in \mathcal{K}, L(f) = \mathbf{y}} \|f - R(\mathbf{y})\|_{\mathcal{H}}. \tag{2}$$
$$\mathrm{err}^{\mathrm{glo}}_{\mathcal{K}}(L, R) := \sup_{f \in \mathcal{K}} \|f - R(L(f))\|_{\mathcal{H}}. \tag{3}$$
- We can generalize previous Hilbert space setting to Banach space setting.
- Assume linear transformation $T(f_0)$ is well approximated by V. For example T is the differential operator.
- Estimate quantity of interest $Q(f_0)$ depending on f_0. For example, $Q(f_0)$ is the integral of f_0 or max/min of f_0 over an interval.
- Optimal recovery contains several mathematical models including numerical PDEs, quadrature rule, interpolation, compressive sensing and learning problem etc.

Relation to Supervised Learning

- Assume $f_0 \in \mathcal{H}_K$, where $\mathcal{H}_K$ is the Reproducing Kernel Hilbert Space(RKHS) associated with a kernel function $K(\cdot,\cdot)$.
- Let $V = \{0\}$ (same as kernel regression) or $V = \mathrm{span}\{K(\mathbf{x}_i, \cdot), i \in I\}$ for subset $I \subseteq [1:m]$, then the locally optimal recovery map is same as kernel ridgeless regression.
- Minimum-(semi)norm interpolation problems are well explained by optimal recovery framework.
- Worst-case error $\|f - R(\mathbf{y})\|$ is estimation error but empirical risk minimization $\|\mathbf{y} - L(f)\|_p^p$ is data error.

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 recast as a semi-definite programming by using S-lemma.
- $(h_1, \dots, h_N)$ be an orthonormal basis for $\mathcal{H}$ such that $(h_1, \dots, h_{N-m})$ is an orthonormal basis for $\ker(L)$.
- Let H be the unitary map $x \in \mathbb{R}^{N-m} \mapsto \sum_{k=1}^{N-m} x_k h_k \in \ker(L)$.

Theorem 1: The local worst-case error of a learning/recovery map $R : \mathbb{R}^m \to \mathcal{H}$ at $\mathbf{y} \in \mathbb{R}^m$ under model set (1) can be expressed, with $g := R(\mathbf{y})$, as
$$e^{\mathrm{loc}}_{\mathcal{K}}(L, g) = \left[\|h - P_{\ker(L)^\perp}(g)\|^2_{\mathcal{H}} + \|P_{\ker(L)}(g)\|^2_{\mathcal{H}} + c^\star\right]^{1/2}, \tag{4}$$
where h is the unique element in $\ker(L)^\perp$ satisfying $L(h) = \mathbf{y}$ and $c^\star$ is the minimal value of the following program, in which $w := P_{V^\perp}(h)$:
$$\underset{c,d \in \mathbb{R}}{\text{minimize}} \quad c \quad \text{s.to } d \ge 0 \quad \text{and} \tag{5}$$
$$\begin{bmatrix} H^*(dP_{V^\perp} - I_{\mathcal{H}})H & \mid & H^*(dw + P_{\ker(L)}(g)) \\ (dw + P_{\ker(L)}(g))^* H & \mid & c + d(\|w\|^2_{\mathcal{H}} - \epsilon^2) \end{bmatrix} \succeq 0.$$

Optimal Recovery Map:

- Let $u_i \in \mathcal{H}$ be the Riesz Representers of the linear functional $\ell_i \in \mathcal{H}^*$.
- $(v_1, \dots, v_n)$ is a basis of linear subspace V.
- Define the Gramian matrix $\mathbf{G} \in \mathbb{R}^{m \times m}$ with $\mathbf{G}_{i,i'} = \langle u_i, u_{i'} \rangle = \ell_i(u_{i'})$ for all $i, i' \in [1:m]$.
- Define the cross-Gramian $\mathbf{C} \in \mathbb{R}^{m \times n}$ with $\mathbf{C}_{i,j} = \langle u_i, v_j \rangle = \ell_i(v_j)$ for all $i \in [1:m], j \in [1:n]$.

Theorem 2: The local optimal learning/recovery map $R^{opt}(\mathbf{y})$ is linear and has closed form for each $\mathbf{y} \in \mathbb{R}^m$ by
$$R^{\mathrm{opt}}(\mathbf{y}) = \sum_{i=1}^m a_i u_i + \sum_{j=1}^n b_j v_j, \tag{6}$$
where the coefficient vectors $\mathbf{a} \in \mathbb{R}^m$ and $\mathbf{b} \in \mathbb{R}^n$ are computed as
$$\mathbf{b} = (\mathbf{C}^\top \mathbf{G}^{-1} \mathbf{C})^{-1} \mathbf{C}^\top \mathbf{G}^{-1} \mathbf{y}, \tag{7}$$
$$\mathbf{a} = \mathbf{G}^{-1}(\mathbf{y} - \mathbf{C}\mathbf{b}). \tag{8}$$

Theorem 2 is based on the observation that the optimal recovery map $R^{opt}(\mathbf{y}) \in \mathcal{H}$ is the solution to the following convex minimization program
$$\underset{f \in \mathcal{H}}{\text{minimize}} \|f - P_V f\|_{\mathcal{H}} \qquad \text{subject to } L(f) = \mathbf{y}. \tag{9}$$
Form supervised learning point of view, the minimization problem can be understood as

- a special minimum-norm interpolation problem,
- empirical risk minimization with special regularization term $\|f - P_V f\|^2_{\mathcal{H}}$.

Numerical Experiments

Worst-case Error Comparison:

- Assume $f \in \mathbb{R}^{200}$ with 50 observations and $\dim(V) = 20$.
- Compare worst-case errors produced by optimal recovery map and empirical risk minimization with $p = 1$ and $p = 2$.

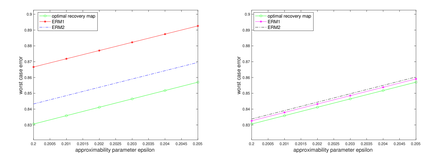

Figure: ERMs with constraints $f \in V$ are in left plot. ERMs in right plot are constructed with constraints $f \in \mathcal{K}$.

Generalization Error Comparison:

- Assume $f_0 \in \mathcal{H}_K$ 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).

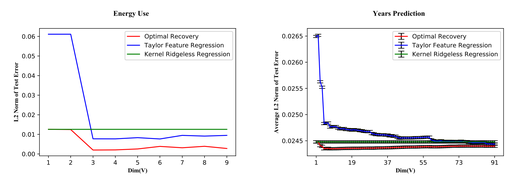

Figure: Optimal recovery and two benchmark regression algorithms on two benchmark data sets.

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 overparametrized 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-1664803. S. F and S. S. also acknowledge NSF grant CCF-1934904.