A representer theorem for deep kernel learning

Why deep kernel learning?

- Many successful machine learning and regression algorithms are based on empirical risk minimization in an RKHS \( \mathcal{H} \), e.g. SVM and least squares.
- If the real solution \( h \notin \mathcal{H} \), kernel-based methods might be a bad choice.
- Combining the ideas of kernel learning with multi-layer (deep) neural networks can remedy this drawback and delivers promising results \([1, 2, 3, 4]\).

Example: Least squares regression of \( h(x, y) = (0.1 + |x - y|)^{-1} \).

A representer theorem for standard kernel learning

- Let \( \mathcal{H} \) be an RKHS of real-valued functions with domain \( \Omega \subset \mathbb{R}^d \) and let the data \( X = \{x_1, \ldots, x_N\} \subset \Omega \) and \( Y = \{y_1, \ldots, y_N\} \subset \mathbb{R} \) be given.
- The main reason for success of kernel algorithms is the representer theorem:
  \[ f = \arg \min_{f \in \mathcal{H}} \sum_{i=1}^N \alpha_i (f(x_i), y_i) + \|f\|_{\mathcal{H}}^2 \]
  with loss function \( c : (\mathbb{R}^d)^d \to [0, \infty) \) and strictly monotonically increasing \( \Phi : [0, \infty) \to [0, \infty) \) solved by
  \[ f = \sum_{i=1}^N \alpha_i K(\cdot, x_i), \]
  where \( K : \Omega^2 \to \mathbb{R} \) is the kernel of \( \mathcal{H} \) and \( \alpha_i \in \mathbb{R}, i = 1, \ldots, N \) are the corresponding coefficients.

A representer theorem for deep kernel learning

- We derived a representer theorem for multi-layer (chained) regression: Let \( \mathcal{H}_1, \ldots, \mathcal{H}_L \) be RKHS of functions with \( d_j \)-dimensional images, \( j = 1, \ldots, L \), such that a concatenation is well-defined. For a loss function \( c \) and strictly monotonically increasing \( \Phi_k, k = 1, \ldots, L \), the minimizer \( \{f_l, \ldots, f_L\} \) of
  \[ \min_{f \in \mathcal{H}_L} \sum_{j=1}^L \sum_{k=1}^j \left( \sum_{i=1}^N \alpha_i K_i(\cdot, x_i) \right) \]
  fulfills \( f_l \in V_l \subset \mathcal{H}_l \) for all \( j = 1, \ldots, L \) with
  \[ V_l = \{ f_j \in \mathcal{H}_j \mid \sum_{i=1}^N \alpha_i K_j(\cdot, x_i)^{\oplus} \text{ s.t. } e_k \mid j = 1, \ldots, N \text{ and } k = 1, \ldots, d \}, \]
  where \( K_j \) denotes the reproducing kernel of \( \mathcal{H}_j \) and \( e_k \) is the \( k \)-th unit vector.

  \( \Rightarrow \) We can consider a finite-dimensional optimization problem instead of an infinite-dimensional one. The representer theorem applies to multi-layer least squares regression or multi-layer SVM for instance.

Example: Application to two-layer least squares regression

- We consider the least-squares problem
  \[ \min_{f \in \mathcal{H}_1 \cap \mathcal{H}_2} J(f_1, f_2) := \sum_{j=1}^N \|f_1 \circ f_2(x_j) - y_j\|^2 + \lambda_1 \|f_1\|_{\mathcal{H}_1}^2 + \lambda_2 \|f_2\|_{\mathcal{H}_2}^2 \]
  with \( f_2 : [-1, 1]^2 \to \mathbb{R}^2 \) and \( f_1 : \mathbb{R}^2 \to \mathbb{R} \) for the two example functions
  \[ a(x, y) := (0.1 + |x - y|)^{-1} \]
  \[ b(x, y) := \chi_{xy}^{-\frac{1}{2}}(x, y). \]
- To determine \( \lambda_1, \lambda_2 \), we perform a 5-fold cv on 100 noisy random samples.

  \( \Rightarrow \) The minimization of \( J \) is done by a BFGS algorithm.

  Comparison to (optimal) single-layer least squares: Pointwise reconstruction errors (ranging from 0% (blue) to > 10% (red) of \( \|h\|_{\mathcal{H}_j}, j \in \{1, 2\} \).

Conclusion

- Our representer theorem allows to recast concatenated learning problems into finite-dimensional optimization tasks.
- It is applicable to modern deep kernel learning algorithms in special cases.
- Even for naive least-squares regression the two-layer variant is able to produce significantly better results than the single-layer algorithm if the original RKHS is not suitable for the problem at hand.

References


Acknowledgments

- The results presented here are based on joint work with Michael Griebel (U Bonn) and Christian Rieger (U Bonn).
- The author was partially supported by the SFB 1060 – The Mathematics of Emergent Effects funded by the Deutsche Forschungsgemeinschaft.