Minimax Estimation of Kernel Mean Embeddings*

Bharath Sriperumbudur

Abstract

The notion of embedding probability measures in a reproducing kernel Hilbert space (RKHS) has gained lot of attention in machine learning and statistics communities due to the wide variety of applications it has been employed in. Some of these applications include kernel two-sample testing, kernel independence and conditional independence testing, density estimation, feature selection, causal inference and distribution regression. Formally, given a probability measure P and a positive definite kernel k (associated with an RKHS, H), the embedding of P in H is defined as $\int k(.,x)dP(x)$, which in words is called the mean element or the kernel mean embedding of P. In all the above mentioned statistical and machine learning applications that deal with the mean embedding, P is usually unknown and the only knowledge of P is through random samples (say of size n) drawn i.i.d. from it. Therefore, in practice, an estimator of the mean element is employed.

The simplest and most popular is the empirical estimator, which is constructed by replacing P by its empirical counterpart, i.e., the empirical measure. In fact, all the above mentioned applications deal with the empirical estimator of the mean element because of its simplicity. The question of interest is: How well does the empirical estimator approximate the mean element? It is well understood that the empirical estimator approximates the mean element very well and the error (in the RKHS norm) goes to zero as n goes infinity and the rate of this convergence is $n^{-1/2}$. Recently, various estimators of the mean element (e.g., shrinkage estimator, kernel density based estimator) have been studied and all of them are shown to have a similar asymptotic behavior to that of the empirical estimator. This raises the question: "Are there estimators that have a rate of convergence faster than $n^{-1/2}$?

In this work, we investigate the above question and show that there are no estimators that can attain a rate that is faster than $n^{-1/2}$ irrespective of the smoothness of k and P, assuming the kernel to be translation-invariant on R^d. This result therefore establishes the optimality of the empirical estimator in the minimax sense. The result is obtained by using the classical tools of statistical minimax theory.

Joint work with Ilya Tolstikhin (MPI, Tuebingen) and Krikamol Muandet (MPI, Tuebingen)

*Machine Learning External Seminar, Gatsby Unit, May 4, 2016.