Optimal Uniform and L^p Rates for Random Fourier Features

Bharath K. Sriperumbudur[*]
Department of Statistics
Pennsylvania State University
University Park, PA 16802, USA
bks18@psu.edu

Zoltán Szabó[*]
Gatsby Unit
University College London
London - WC1N 3AR, UK
zoltan.szabo@gatsby.ucl.ac.uk

Kernel techniques are among the most powerful approaches in machine learning and Bayesian modelling due to their capability to represent and model complex relations. However, this flexibility and richness of kernels has a price: by resorting to implicit construction of feature maps these methods operate on the Gram matrix of the data, which raises serious computational challenges while dealing with large scale data. In order to mitigate this severe numerical limitation, recently randomized constructions have been proposed in the literature, which allow the application of fast linear algorithms.

Random Fourier features (RFF) represent one of the most popular and widely used approaches for generating such low-dimensional, easily computable feature representations for shift-invariant kernels: as it has been demonstrated numerically in several applications kernel machines relying on the RFF approximation not only scale benignly and well-fitted for online settings, but also show graceful performance degradation compared to the exact solution. Another advantage with the RFFs is that unlike the low-rank matrix approximation approach, which also speeds up kernel machines, it approximates the entire kernel function and not just the Gram matrix. This property is particularly useful when one is faced with out of sample data problems.

Despite the empirical success and popularity of RFFs, very little is understood theoretically about their approximation quality and existing theoretical results focus solely on the approximation of *kernel values*. However, there are numerous real-world problems where the usage of *kernel derivatives* is of central importance. For example, as it has been shown recently in the Bayesian literature, one can construct efficient, gradient-free adaptive MCMC algorithms relying on infinite-dimensional exponential family (IDEF) distributions, where the fitting problem of IDEFs boils down to a linear equation system with entries containing kernel values and kernel derivatives. Further applications based on kernel derivatives include semi-supervised or Hermite learning with gradient information, nonlinear variable selection, or (multi-task) gradient learning.

In this work, we present detailed finite-sample theoretical analysis on the approximation quality of RFFs for kernels and kernel derivatives. Our first result shows that the RFF based kernel estimator achieves almost sure convergence uniformly on compact sets with (essentially) exponentially growing diameter as a function of the RFF dimension in contrast to existing guarantees which only allowed sublinearly increasing set sizes. This asymptotic special case of our result can be shown to be optimal, similarly to the dependence of our novel guarantee in terms of the RFF dimension. In addition to the convergence of kernel approximations in uniform norm, we also provide guarantees in L^p ($1 \leq p < \infty$) norm, and propose an RFF approximation to derivatives of a kernel with a theoretical study on its approximation quality in both uniform and L^p sense.

Acknowledgments: This work was supported by the Gatsby Charitable Foundation.

[*] Contributed equally. Theory of Big Data Conference, London, UK, 6-8 January 2016; abstract.