LARGE-SCALE RANDOM FEATURES FOR KERNEL REGRESSION

Valero Laparra[1], Diego Marcos Gonzalez[2], Devis Tuia[2] and Gustau Camps-Valls[1]

[1]Image Processing Lab (IPL), Universitat de València, Spain, {valero.laparra,gustau.camps}@uv.es
[2] University of Zurich, Switzerland, {diego.marcos, devis.tuia}@geo.uzh.ch

ABSTRACT

Kernel methods constitute a family of powerful machine learning algorithms, which have found wide use in remote sensing and geosciences. However, kernel methods are still not widely adopted because of the high computational cost when dealing with large scale problems, such as the inversion of radiative transfer models. This paper introduces the method of random kitchen sinks (RKS) for fast statistical retrieval of bio-geo-physical parameters. The RKS method allows to approximate a kernel matrix with a set of random bases sampled from the Fourier domain. We extend their use to other bases, such as wavelets, stumps, and Walsh expansions. We show that kernel regression is now possible for datasets with millions of examples and high dimensionality. Examples on atmospheric parameter retrieval from infrared sounders and biophysical parameter retrieval by inverting PROSAIL radiative transfer models with simulated Sentinel-2 data show the effectiveness of the technique.

1. INTRODUCTION

Kernel methods constitute an appropriate framework to approach many statistical inference problems [1]. In the last decade these methods have replaced other techniques in many fields of science and engineering, and have become the new standard in remote sensing data analysis [2, 3]. Kernel methods allow treating in the very same framework different problems, from feature extraction [4] to classification [5] and regression [6]. The fundamental building block of the theory of kernel learning is the *kernel function*, which compares (possibly complex) multidimensional data objects. In a nutshell, given n data points, all kernel methods have to operate with a squared (eventually huge) matrix of size $n \times n$, which contains all pairwise sample similarities. Designing an appropriate kernel function that captures data dependencies is, nevertheless, not easy in general. Many approaches have been followed so far to tackle this problem: from learning the metric implicit in the kernel [7], to learning compositions of simpler kernels [8]. Selecting and optimizing a kernel function is very challenging even with moderate amounts of data. Many efforts have been done to deliver large-scale versions of kernel machines able to work with several thousands of examples. They typically resort to reduce the dimensionality of the problem by decomposing the kernel matrix using a subset of bases: for instance using Nyström eigendecompositions [9], sparse and low rank approximations [10, 4], or smart sample selection [11]. However, there is no clear evidence that these approaches work in general, given that they are mere approximations to the kernel computed with all (possibly millions of) samples.

In this paper, we explore an alternative pathway: rather than *optimization* we will follow *randomization*. While odd at a first glance, the approach has surprisingly yielded competitive results in the last years, being able to exploit many samples at a fraction of the computational cost. Besides its practical convenience, the approximation of the kernel with random bases is also theoretically consistent. The seminal work in [12] presented the randomization framework. Given a sample set $\{\mathbf{x}_i \in \mathbb{R}^d | i = 1, \dots, n\}$, the idea is to approximate the kernel function with an empirical kernel mapping of the form:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle \approx \mathbf{z}(\mathbf{x}_i)^\top \mathbf{z}(\mathbf{x}_j),$$

where the implicit mapping $\boldsymbol{\phi}(\cdot)$ is replaced with an *explicit* (low-dimensional) feature mapping $\mathbf{z}(\cdot)$ of dimension D. Consequently, one can simply transform the input with $\mathbf{z}$, and then apply fast linear learning methods to approximate the corresponding nonlinear kernel machine. This approach not only provides extremely fast learning algorithms, but also good performance in the test phase: for a given test point $\mathbf{x}$, instead of $f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x})$, which requires $\mathcal{O}(nd)$ operations, one simply does a linear projection $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{z}$, which requires $\mathcal{O}(D + d)$ operations. The question now is how to construct efficient and sensible $\mathbf{z}$ mappings. The work in [12] also introduced a particular technique to do so, called *random kitchen sinks* (RKS).

The remainder of the paper is organized as follows. Section 2 reviews the RKS method, and introduces the different expansions used in this work. Section 3 presents and discusses the experimental results in two challenging problems of bio-geo-physical parameter retrieval: atmospheric parameter retrieval from infrared sounders such as IASI and inversion of PROSAIL radiative transfer models with Sentinel-2 simulated data. Section 4 concludes the paper.

We wish to thank Dr. Jochem Verrelst at the Image Processing Lab (IPL) in the Universitat de València (Spain) for preparing the data used in the second experiment.

This paper has been partially supported by the Spanish Ministry of Economy and Competitiveness under project TIN2012-38102-C03-01, and by the Swiss National Science Foundation under the grant PP00P2-150593.

2. KERNEL APPROXIMATION WITH RANDOM KITCHEN SINKS

2.1. Random kitchen sinks

Random kitchen sinks exploit a classical definition in harmonic analysis [12], by which a continuous kernel $k(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y})$ on $\mathbb{R}^d$ is positive definite if and only if k is the Fourier transform of a non-negative measure. If a shift-invariant kernel k is properly scaled, its Fourier transform $p(\omega)$ is a proper probability distribution. Defining the function $C_\omega(\mathbf{x}) = e^{j\omega^\top \mathbf{x}}$, we obtain

$$k(\mathbf{x} - \mathbf{y}) = \int_{\mathbb{R}^d} p(\omega) e^{j\omega^\top (\mathbf{x}-\mathbf{y})} d\omega = \mathbb{E}_\omega[C_\omega(\mathbf{x}) C_\omega(\mathbf{y})^*],$$

so $C_\omega(\mathbf{x}) C_\omega(\mathbf{y})^*$ is an unbiased estimate of $k(\mathbf{x} - \mathbf{y})$ when ω is drawn from p. In our case, both $p(\omega)$ and $k(\mathbf{x} - \mathbf{y})$ are real valued, what allows us to substitute the complex exponentials by cosines and to use $z_\omega(\mathbf{x})^\top z_\omega(\mathbf{y})$, where $z_\omega(\mathbf{x}) = \sqrt{2} cos(\omega^\top \mathbf{x} + b)$, as an estimator of $k(\mathbf{x} - \mathbf{y})$ as long as ω is drawn from $p(\omega)$ and b is drawn uniformly from $[0, 2\pi]$. Also note that $z_\omega(\mathbf{x})^\top z_\omega(\mathbf{y})$ has expected value $k(\mathbf{x}, \mathbf{y})$ because of the sum of angles formula. Now, one can lower the variance of the estimate of the kernel by concatenating D randomly chosen $\mathbf{z}_\omega$ into one D-dimensional vector $\mathbf{z}$ and normalizing each component by $\sqrt{D}$. An illustrative example of how RKS approximates k with random bases is given in Fig. 1.

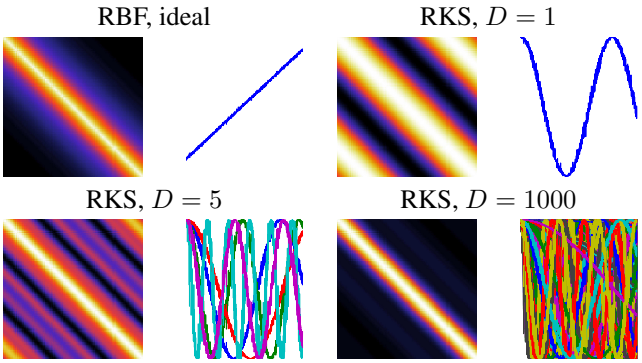

Fig. 1: Illustration of the effect of randomly sampling D bases from the Fourier domain on the kernel matrix. With sufficiently large D, the kernel matrix generated by RKS approximates that of the RBF kernel, at a fraction of the time.

2.2. RKS in practice

The RKS algorithm reduces to the following steps:

1. Draw D i.i.d. samples $\omega_1, \ldots, \omega_D \in \mathbb{R}^d$ from p, and $b_1, \ldots, b_D \in \mathbb{R}$ from the uniform distribution $[0, 2\pi]$
2. Construct the low-dimensional feature map: $\mathbf{z} = \sqrt{\frac{2}{D}} [\cos(\omega_1^\top \mathbf{x} + b_1), \ldots, \cos(\omega_D^\top \mathbf{x} + b_D)]$
3. Approximate the kernel function: $k \approx \mathbf{z}^\top \mathbf{z}$, and associated kernel matrix, $\mathbf{K} = \mathbf{Z}\mathbf{Z}^\top$

The method is very efficient in both speed and memory requirements, as shown in Table 1.

Table 1: Computational and memory costs for different approximate kernel methods in problems with d dimensions, D features, n samples.

Method	Train time	Test time	Train mem	Test mem
Naive [1]	$\mathcal{O}(n^2 d)$	$\mathcal{O}(nd)$	$\mathcal{O}(nd)$	$\mathcal{O}(nd)$
Low Rank [10]	$\mathcal{O}(nDd)$	$\mathcal{O}(Dd)$	$\mathcal{O}(Dd)$	$\mathcal{O}(Dd)$
RKS [12]	$\mathcal{O}(nDd)$	$\mathcal{O}(Dd)$	$\mathcal{O}(Dd)$	$\mathcal{O}(Dd)$

2.3. RKS beyond Fourier bases

The RKS algorithm can actually exploit other approximating functions besides Fourier expansions. Note that actually any shift-invariant kernel, i.e. $k(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y})$, can be represented using random cosine features. Randomly sampling distribution functions impacts the definition of the corresponding reproducing kernel Hilbert space (rkHs): sampling the Fourier bases with $z_\omega(\mathbf{x}) = \sqrt{2}\cos(\omega_o^\top \mathbf{x} + b)$ actually leads to the Gaussian RBF kernel $k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2/(2\sigma^2))$, while a random stump (i.e. sigmoid-shaped functions) sampling defined by $z_\omega(\mathbf{x}) = \text{sign}(\mathbf{x} - \omega)$ leads to the kernel $k(\mathbf{x}, \mathbf{y}) = 1 - \frac{1}{a}\|\mathbf{x} - \mathbf{y}\|_1$. Another possibility is to resort to binning bases functions, which partition the input space using an axis-aligned grid, and assign a binary indicator to each partition, which is shown to approximate a Laplacian kernel, $k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|_1/(2\sigma^2))$ [12]. In this paper we will also explore the possibility of Walsh and the Gabor basis functions widely used in signal and image processing.

3. EXPERIMENTS

This paper presents experimental results on the use in RKS in two remote sensing applications. We exploit several kitchen sinks: the standard random Fourier basis functions, along with the Walsh, Haar, wavelets and stumps.

3.1. Experiment 1: Atmospheric parameter retrieval

In this first experiment, we exploit random kernels in a challenging regression problem in remote sensing: the estimation of atmospheric profiles from large scale hyperspectral infrared sounders. Temperature and water vapor are atmospheric parameters of high importance for weather forecast and atmospheric chemistry studies [13]. Observations from spaceborne high spectral resolution infrared sounding instruments can be used to calculate the profiles of such atmospheric parameters with unprecedented accuracy and vertical resolution [14]. In this work we focus on the data coming from the Infrared Atmospheric Sounding Interferometer (IASI), that provides radiances in 8461 spectral channels, between 3.62 and 15.5 μm with a spectral resolution of

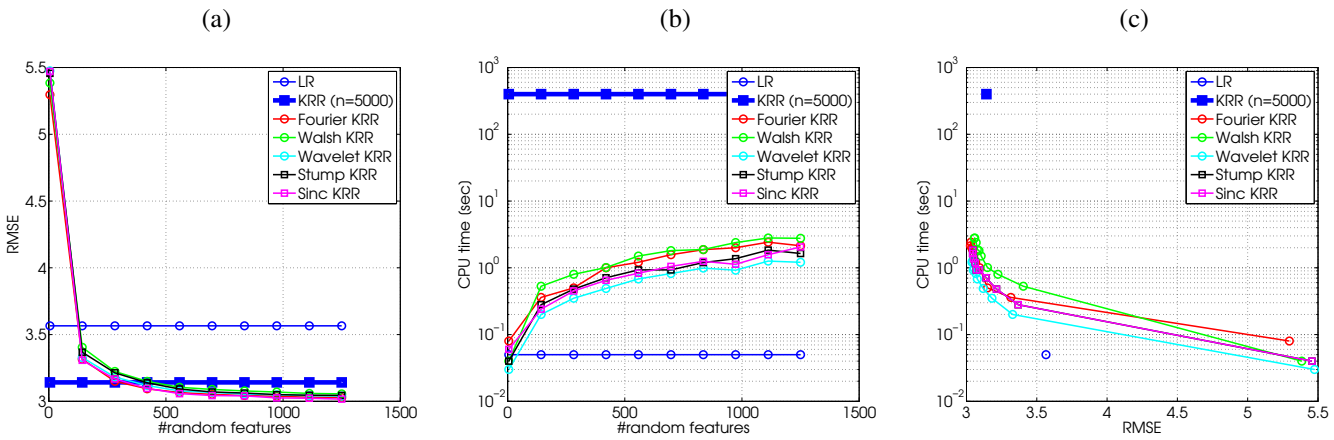

Fig. 2: Results of the RKS approach for different random sinks: (a) RMSE [K] and (b) training time [sec] versus the number of random features drawn; and (c) RMSE [K] versus training time [sec].

0.5 cm^{-1} after apodization [15]. Its spatial resolution is 25 km at nadir with an Instantaneous Field of View (IFOV) size of 12 km at an altitude of 819 km. This huge data dimensionality typically requires simple and computationally efficient processing techniques that can exploit the wealth of available observations provided by ECMWF re-analysis.

Figure 2 shows results for the prediction of the temperature atmospheric profile. We trained linear regression (LR) and kernel ridge regression (KRR) using the first 100 principal components of an IASI orbit (2008-07-17), both using 5000 samples. The RKS approximations were all trained with the ensemble of 100,000 examples. All models were then tested on the same independent test set of 20,000 examples. Experiments were performed using Matlab on an Intel 3.3 GHz processor with 8 GB RAM memory under Ubuntu 14.4. Figure 2(a) shows that LR cannot cope with the nonlinearity of the problem, which can be adressed by using the kernel least squares regression method, KRR. However, training the KRR with more than 5000 samples turns out to be hard in regular machines. Using RKS instead is beneficial. It is actually observed that a sufficiently large number of randomly sampled bases for kernel approximation can improve the results in terms of accuracy and computational efficiency: in this case > 600 random features were enough to beat the full 5000-samples KRR. The big leap in computational cost is observed in Fig. 2(b) (note the log-scale). A trade-off comparison in Fig. 2(c) reveals that the best accuracy-cost compromise in this particular example is to sample from the traditional squared-shaped Haar wavelet.

3.2. Experiment 2: Inversion of PROSAIL data

The second experiment deals with the inversion of PROSAIL radiative transfer model[1]. PROSAIL is the combination of the PROSPECT leaf optical properties model and the SAIL canopy bidirectional reflectance model. PROSAIL has been used to develop new methods for retrieval of vegetation biophysical properties. Essentially, PROSAIL links the spectral variation of canopy reflectance, which is mainly related to leaf biochemical contents, with its directional variation, which is primarily related to canopy architecture and soil/vegetation contrast. This link is key to simultaneous estimation of canopy biophysical/structural variables for applications in agriculture, plant physiology, and ecology at different scales. PROSAIL has become one of the most popular radiative transfer tool due to its ease of use, robustness, and consistent validation by lab/field/space experiments over the years.

Table 2: Configuration parameters of the simulated data.

Parameter	Sampling	Min	Max
RTM model: Prospect 4			
Leaf Structural Parameter	Fixed	1.50	1.50
C_{ab}, chlorophyll a+b [μg/cm^2]	$\mathcal{U}(14, 49)$	0.067	79.97
C_w, equivalent water thickness [mg/cm^2]	$\mathcal{U}(10, 31)$	2	50
C_m, dry matter [mg/cm^2]	$\mathcal{U}(5.9, 19)$	1.0	3.0
RTM model: 4SAIL			
Diffuse/direct light	Fixed	10	10
Soil Coefficient	Fixed	0	0
Hot spot	Fixed	0.01	0.01
Observer zenit angle	Fixed	0	0
LAI, Leaf Area Index	$\mathcal{U}(1.2, 4.3)$	0.01	6.99
LAD, Leaf Angle Distribution	$\mathcal{U}(28, 51)$	20.04	69.93
SZA, Solar Zenit Angle	$\mathcal{U}(8.5, 31)$	0.082	49.96
PSI, Azimut Angle	$\mathcal{U}(30, 100)$	0.099	179.83

In this section, we used PROSAIL to generate 1,000,000 pairs of Sentinel-2 spectral (13 spectral channels) and 7 associated parameters: Total Leaf Area Index (LAI), Leaf angle distribution (LAD), Solar Zenit Angle (SZA), Azimut Angle (PSI), chlorophyll a+b content C_{ab} [μg/cm^2], equivalent water thickness C_w [g/cm^2] and dry matter content, C_m [g/cm^2]. See Table 2 for some configuration details of the simulation. This constitutes a challenging multi-output regression problem.

[1] http://teledetection.ipgp.jussieu.fr/prosail/

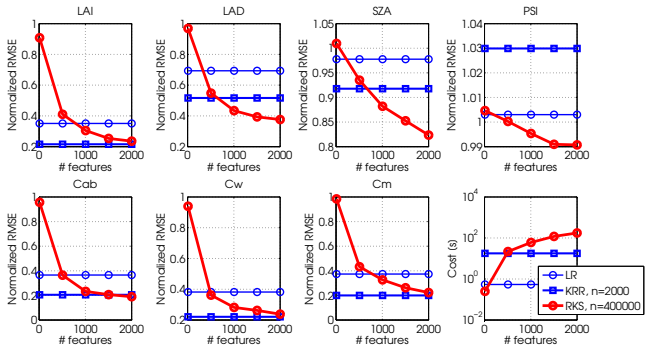

Fig. 3: RMSE results in the PROSAIL inversion experiment for the seven parameters and the computational cost (bottom right).

Figure 3 shows the obtained results for the inversion of PROSAIL. We show both the normalized RMSE and the computational cost of a regularized linear regression, KRR and RKS. In all cases we predict the seven parameters with a single multiple-output regression model. We trained KRR with 2,000 samples, and consequently trained RKS for a maximum of $D = 2000$. RKS employed 400,000 samples and cosine basis. Several conclusions can be derived: 1) RKS yields in general competitive performance versus KRR; and 2) RKS largely improves predictions for LAD, SZA, and PSI estimation, while similar in accuracy to KRR for the rest of parameters.

4. CONCLUSIONS

This paper explored the use of randomly-generated bases for large-scale kernel regression in remote sensing biophysical parameter estimation. We exploited the approximation of the kernel function via random sampling from Fourier, wavelets, Walsh and stump functions. We showed results in two problems. First we tackled a high-dimensional large scale problem very common in remote sensing: the estimation of atmospheric profiles from large scale hyperspectral infrared sounding IASI radiances. Second, we explored RKS for the inversion of the widely used PROSAIL radiative transfer model for which we used 400,000 pairs of Sentinel-2 simulations. Both are multi-output problems. Results showed that we can train kernel regression models with several thousands of data points, which is not possible in standard kernel optimization strategies. The RKS model produced big gains in accuracy and computational efficiency.

We noted however that RKS has two shortcomings. First, the memory bottleneck is still present as one has to store the $\mathbf{Z}$ matrix, which is $D \times n$, to compute the approximate kernel matrix, $\mathbf{K} = \mathbf{Z}\mathbf{Z}^\top$. This will be addressed in the future through low-rank and blocky approximation of $\mathbf{Z}$. And second, other (sparser) bases can be more appropriate. In this work, we used the Walsh basis but results did not improve those of standard Fourier bases. Alternatives to Hadamard expansions, much in line of Fastfood [16], could eventually improve further the results and efficiency.

5. REFERENCES

[1] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.

[2] G. Camps-Valls and L. Bruzzone, Eds., *Kernel methods for Remote Sensing Data Analysis*, Wiley & Sons, UK, Dec 2009.

[3] G. Camps-Valls, D. Tuia, L. Gómez-Chova, S. Jiménez, and J. Malo, Eds., *Remote Sensing Image Processing*, Morgan & Claypool Publishers, LaPorte, CO, USA, Sep 2011.

[4] J. Arenas-Garcia, K. Petersen, G. Camps-Valls, and L.K. Hansen, "Kernel multivariate analysis framework for supervised subspace learning: A tutorial on linear and kernel multivariate methods," *Signal Processing Magazine, IEEE*, vol. 30, no. 4, pp. 16–29, July 2013.

[5] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. Atli Benediktsson, "Advances in hyperspectral image classification: Earth monitoring with statistical learning methods," *Signal Processing Magazine, IEEE*, vol. 31, no. 1, pp. 45–54, Jan 2014.

[6] G. Camps-Valls, J. Muñoz and, L. Gómez, L. Guanter, and X. Calbet, "Nonlinear statistical retrieval of atmospheric profiles from MetOp-IASI and MTG-IRS infrared sounding data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 5, pp. 1759–1769, 2012.

[7] K.Q. Weinberger and G. Tesauro, "Metric learning for kernel regression," in *International Conference on Artificial Intelligence and Statistics*, 2007, pp. 608–615.

[8] A. Rakotomamonjy, F.R. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, Nov. 2008.

[9] S. Kumar, M. Mohri, and A. Talwalkar, "Sampling methods for the Nyström method," *Journal of Machine Learning Research*, vol. 13, pp. 981–1006, 2012.

[10] S. Fine and K. Scheinberg, "Efficient SVM training using low-rank kernel representations," *Journal of Machine Learning Research*, vol. 2, pp. 243–264, 2001.

[11] A. Bordes, S. Ertekin, J. Weston, and L. Bottou, "Fast kernel classifiers with online and active learning," *Journal of Machine Learning Research*, vol. 6, pp. 1579–1619, 2005.

[12] Ali Rahimi and Benjamin Recht, "Random features for large-scale kernel machines," in *Neural Information Processing Systems*, 2007.

[13] K. N. Liou, *An Introduction to Atmospheric Radiation*, Academic Press, Hampton, USA, 2nd edition, 2002.

[14] H. L. Huang, W. L. Smith, and H. M. Woolf, "Vertical resolution and accuracy of atmospheric infrared sounding spectrometers," *Journal of Applied Meteorology*, vol. 31, pp. 265–274, 1992.

[15] D. Siméoni, C. Singer, and G. Chalon, "Infrared atmospheric sounding interferometer," *Acta Astronautica*, vol. 40, pp. 113–118, 1997.

[16] Quoc Le, Tamás Sarlós, and Alex Smola, "Fastfood – approximating kernel expansions in loglinear time," in *International Conference on Machine Learning*, 2013.