# On the Generalization Performance of Kernel Approximation Methods

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

VERSION 1: Our story begins with an intriguing observation: across a variety of datasets and tasks, we consistently observe that random Fourier features (RFFs) [1] outperform the Nyström method [2] under a fixed memory budget. This holds true in spite of the fact that the Nyström method often has smaller kernel approximation error than the RFF approximation, in terms of both the spectral norm and the Frobenius norm. In order to explain this phenomenon, we introduce a novel metric measuring the distance between a kernel approximation matrix $\tilde{K}$ and the exact kernel matrix $K$, and bound the generalization performance in terms of this metric. Letting $\sigma_i$ and $\tilde{\sigma}_i$ be the eigenvalues of $K$ and $\tilde{K}$, and $\epsilon > 0$, we define the "$\epsilon$-log-spectral distance" $D_\epsilon(K, \tilde{K}) = \max_i |\log(\sigma_i + \epsilon) - \log(\tilde{\sigma}_i + \epsilon)|$. In order for $\tilde{K}$ to be close to $K$ under this metric, all its eigenvalues $\tilde{\sigma}_i > \epsilon$ must have similar order of magnitude to those of $K$; in particular, the rank $m$ of $\tilde{K}$ must be large enough such that $\log(\sigma_{m+1} + \epsilon) \approx \log(\epsilon)$. Because RFFs are more memory efficient than Nyström features, the RFF approximation will be higher rank than the Nyström approximation, when under a memory constraint. As a result, under the proposed metric, the RFF kernel approximation is *closer* to the exact kernel matrix than the Nyström approximation is. Empirically, we demonstrate across a number of classification and regression tasks that this metric is much more predictive of generalization performance than the spectral or Frobenius norms of the kernel approximation error. Taking inspiration from these results, we propose using *low-precision* random Fourier features in order to generate an even higher-rank approximation under the same memory budget. We demonstrate that we are able to match the performance of full-precision RFF models, using low-precision features and low-precision training algorithms; this allows us to both reduce the memory footprint of the training algorithm, and learn more compact models.

VERSION 2: In this work, we bound the generalization performance of kernel approximation methods in terms of a novel metric measuring the distance between a kernel approximation matrix $\tilde{K}$ and the exact kernel matrix $K$. Letting $\sigma_i$ and $\tilde{\sigma}_i$ be the eigenvalues of $K$ and $\tilde{K}$, and $\epsilon > 0$, we define the "$\epsilon$-log-spectral distance" $D_\epsilon(K, \tilde{K}) = \|\log(\sigma + \epsilon) - \log(\tilde{\sigma} + \epsilon)\|_\infty$. We use the insight provided by this metric to (1) explain our intriguing observation that under a fixed memory budget, random Fourier features (RFFs) [1] outperform Nyström features [2], and to (2) propose using *low-precision* random Fourier features in order to get even better performance under the same memory budget. We demonstrate that we are able to match the performance of full-precision RFF models, using low-precision features and low-precision training algorithms. Across all of these experiments, we show a very strong correspondence between the "$\epsilon$-log-spectral distance", and the generalization performance of a model.

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

We investigate the problem of training kernel approximation models under memory constraints. First, we compare the performance of two popular kernel approximation methods: the Nyström method [2], and random Fourier features Our first step toward this goal is to compare the performance of two leading kernel approximation methods, the Nystrom method and random Fourier features (RFFs), in this memory constrained setting. We find that RFFs are able to outperform Nystrom features, and explain this through the theory of fixed design linear regression. Intuitively, by using more features, the RFF method is able to approximate a larger portion of the true spectrum of the kernel matrix, and this gives it more expressive power. Motivated by this insight, we propose using low-precision RFFs to cover even more of the spectrum, under a fixed budget. Through the same regression theory, as well as through experiments, we show that there are important regimes in which lowering precision does not degrade performance. Make it upfront that why we don't play with low precision Nystrom. We demonstrate that we can attain strong empirical performance by using these low-precision features in low-precision training algorithms. This allows us to (1) perform training with a smaller memory footprint (HALP), and (2) learn more compact models (LP-SGD).

# 1 Introduction

# 2 Related Work

# 3 Preliminaries and Notation

# References


[1] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pages 1177–1184, 2007.

[2] C. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 682–688. MIT Press, 2001.