Review of "Random Features for Large-Scale Kernel Machines"

Paper authored by Ali Rahimi and Benjamin Recht

Presented by Ryan Farell

Computer Visualization Center
Oden Institute for Computational Engineering and Sciences
University of Texas

February 19, 2024

Agenda

1. Overview of the paper and the talk
2. Bochner's Theorem
3. Derivation of Random Fourier Features
4. Convergence Guarantee of RFF
5. Random Fourier Feature (RFF) Algorithm
6. Example
7. References

Paper Overview

Paper Title: Random features for large-scale kernel machines
Authors: Ali Rahimi and Benjamin Recht
Published in: Advances in Neural Information Processing Systems
Volume: 20
Year: 2007
Citation: Ali Rahimi and Benjamin Recht. "Random features for large-scale kernel machines". In: *Advances in neural information processing systems* 20 (2007)
Objective: This paper proposes an approach to approximate the feature map of a kernel function efficiently, enabling the application of kernel methods to large-scale problems. The method is particularly well-suited for shift-invariant kernels and allows the application of linear learning algorithms to nonlinear problems using kernel methods.

Bochner's Theorem

A continuous kernel $k(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y})$ on $\mathbb{R}^d$ is positive definite if and only if $k(\boldsymbol{\tau})$ is the Fourier transform of a non-negative measure.

Derivation of Random Fourier Features: Part 1

Bochner's Theorem tells us that

$$k(\boldsymbol{\tau}) = \int_{\mathbb{R}^D} S(\boldsymbol{\omega}) e^{i\boldsymbol{\omega}^\top \boldsymbol{\tau}} d\boldsymbol{\omega}, \quad S(\boldsymbol{\omega}) = \frac{1}{2\pi} \int_{\mathbb{R}^D} k(\boldsymbol{\tau}) e^{-i\boldsymbol{\omega}^\top \boldsymbol{\tau}} d\boldsymbol{\tau}$$

where $k(\boldsymbol{\tau})$ is the shift-invariant kernel and $S(\boldsymbol{\omega})$ is its power spectral density. The relation between the power spectrum and a probability distribution is given by

$$S(\boldsymbol{\omega}) = k(\mathbf{0}) p_S(\boldsymbol{\omega}) = \sigma_0^2 p_S(\boldsymbol{\omega}).$$

Derivation of Random Fourier Features: Part 2

Defining $\zeta_\omega(\mathbf{x}) = e^{i\omega^\top \mathbf{x}}$, we have

$$\begin{aligned} k(\tau) &= k(\mathbf{x}-\mathbf{y}) \\ &= \int_{\mathbb{R}^d} S(\omega) e^{i\omega^\top(\mathbf{x}-\mathbf{y})} d\omega \\ &= \sigma_0^2 \int_{\mathbb{R}^d} p(\omega) e^{i\omega^\top(\mathbf{x}-\mathbf{y})} d\omega \\ &= \sigma_0^2 E_\omega \left[\zeta_\omega(\mathbf{x})\zeta_\omega(\mathbf{y})^*\right] \end{aligned}$$

so $\sigma_0^2 \zeta_\omega(\mathbf{x})\zeta_\omega(\mathbf{y})^*$ is an unbiased estimate of $k(\mathbf{x}, \mathbf{y})$ when ω is drawn from p_S.

Derivation of Random Fourier Features: Part 3

To obtain a real-valued random feature for k, note that both the probability distribution $p(\boldsymbol{\omega})$ and the kernel $k(\boldsymbol{\tau})$ are real, so the integrand $e^{i\boldsymbol{\omega}^\top(\mathbf{x}-\mathbf{y})}$ may be replaced with $\cos(\boldsymbol{\omega}^\top(\mathbf{x}-\mathbf{y}))$. Defining $\mathbf{z}_{\boldsymbol{\omega}}(\mathbf{x}) = [\cos(\boldsymbol{\omega}^\top\mathbf{x}), \sin(\boldsymbol{\omega}^\top\mathbf{x})]^\top$ gives a real-valued mapping that satisfies the condition $E[\mathbf{z}_{\boldsymbol{\omega}}(\mathbf{x})^\top\mathbf{z}_{\boldsymbol{\omega}}(\mathbf{y})] = k(\mathbf{x},\mathbf{y})$, since

$$\begin{aligned}\mathbf{z}_{\boldsymbol{\omega}}(\mathbf{x})^\top\mathbf{z}_{\boldsymbol{\omega}}(\mathbf{y}) &= \cos(\boldsymbol{\omega}\boldsymbol{x})\cos(\boldsymbol{\omega}\boldsymbol{y}) + \sin(\boldsymbol{\omega}\boldsymbol{x})\sin(\boldsymbol{\omega}\boldsymbol{y}) \\ &= \cos(\boldsymbol{\omega}^\top(\mathbf{x}-\mathbf{y})).\end{aligned}$$

Derivation of Random Fourier Features: Part 4

We can lower the variance of $\mathbf{z}_\omega(\mathbf{x})^\top \mathbf{z}_\omega(\mathbf{y})$ by concatenating D randomly chosen $\mathbf{z}_\omega$ into a column vector $\mathbf{z}$ and normalizing each component by $\sqrt{D}$. The inner product of points is

$$\begin{aligned}\mathbf{z}(\mathbf{x})^\top \mathbf{z}(\mathbf{y}) = \frac{1}{D}\Big[&\cos(\omega_1^T\mathbf{x}), \dots \cos(\omega_D^T\mathbf{x}), \sin(\omega_1^T\mathbf{x}), \dots \sin(\omega_D^T\mathbf{x})\Big] \\ &\Big[\cos(\omega_1^T\mathbf{y}), \dots \cos(\omega_D^T\mathbf{y}), \sin(\omega_1^T\mathbf{y}), \dots \sin(\omega_D^T\mathbf{y})\Big]^T \\ &= \frac{1}{D}\sum_{k=1}^{D} \cos(\boldsymbol{\omega}_k^\top(\mathbf{x}-\mathbf{y}))\end{aligned}$$

Convergence Guarantee of RFF

Since $\mathbf{z}_{\boldsymbol{\omega}}(\mathbf{x})^{\top} z_{\boldsymbol{\omega}}(\mathbf{y})$ is bounded between -1 and 1 , for a fixed pair of points $\mathbf{x}$ and $\mathbf{y}$, Hoeffding's inequality guarantees exponentially fast convergence in D between $\mathbf{z}(\mathbf{x})^{\top}\mathbf{z}(\mathbf{y})$ and $k(\mathbf{x}, \mathbf{y})$:

$$\Pr\left[\left|\mathbf{z}(\mathbf{x})^{\top}\mathbf{z}(\mathbf{y}) - k(\mathbf{x}, \mathbf{y})\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{D\epsilon^2}{2}\right).$$

Random Fourier Feature (RFF) Algorithm

Require: A positive definite shift-invariant kernel $k(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y})$.

Ensure: A randomized feature map $\mathbf{z}(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^{2D}$ so that $\mathbf{z}(\mathbf{x})^\top \mathbf{z}(\mathbf{y}) \approx k(\mathbf{x} - \mathbf{y})$.

1: Compute the Fourier transform S of the kernel k: $S(\boldsymbol{\omega}) = \frac{1}{2\pi} \int_{\mathbb{R}^D} k(\boldsymbol{\tau}) e^{-i\boldsymbol{\omega}^\top \boldsymbol{\tau}} d\boldsymbol{\tau}$.

2: Compute the propability p_S of the power spectrum $p_S = \frac{1}{\sigma_0^2} S(\boldsymbol{\omega})$.

3: Draw D iid samples $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_D \in \mathbb{R}^d$ from p_S.

4: Let $\mathbf{z}(\mathbf{x}) = \frac{1}{\sqrt{D}} \left[\cos(\boldsymbol{\omega}_1^\top \mathbf{x}), \ldots \cos(\boldsymbol{\omega}_D^\top \mathbf{x}), \sin(\boldsymbol{\omega}_1^\top \mathbf{x}), \ldots \sin(\boldsymbol{\omega}_D^\top \mathbf{x})\right]^\top$.

Example with Gaussian Kernel

- Now, let's explore an example using the Gaussian kernel.
- Follow along with the demonstration on GitHub for a hands-on experience.
- Repository:
  https://github.com/rfarell/2007_RahimiRecht_RandomFeaturesKernel

Example with Gaussian Kernel

The Gaussian kernel is a popular choice in various machine learning applications, particularly because of its properties of smoothness and locality. The kernel function is defined as:

$$k(\tau) = e^{-\frac{\|\tau\|^2}{2}}$$

where τ is the difference between two points in the input space. The power spectrum of the Gaussian kernel, which is its Fourier transform, is also Gaussian:

$$p(\omega) = (2\pi)^{-\frac{D}{2}} e^{-\frac{\|\omega\|^2}{2}}$$

where ω represents frequency components and D is the dimensionality of the input space.

Kernel Approximation Visualization

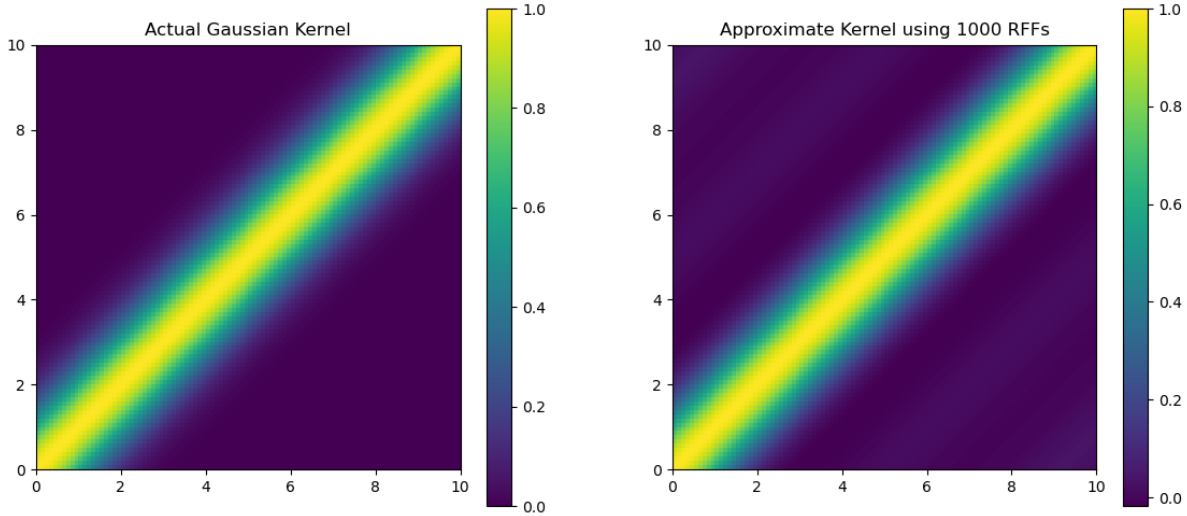

Figure: Comparison of the actual Gaussian kernel (left) with the approximate kernel using Random Fourier Features (right).

Error Analysis - Total Error

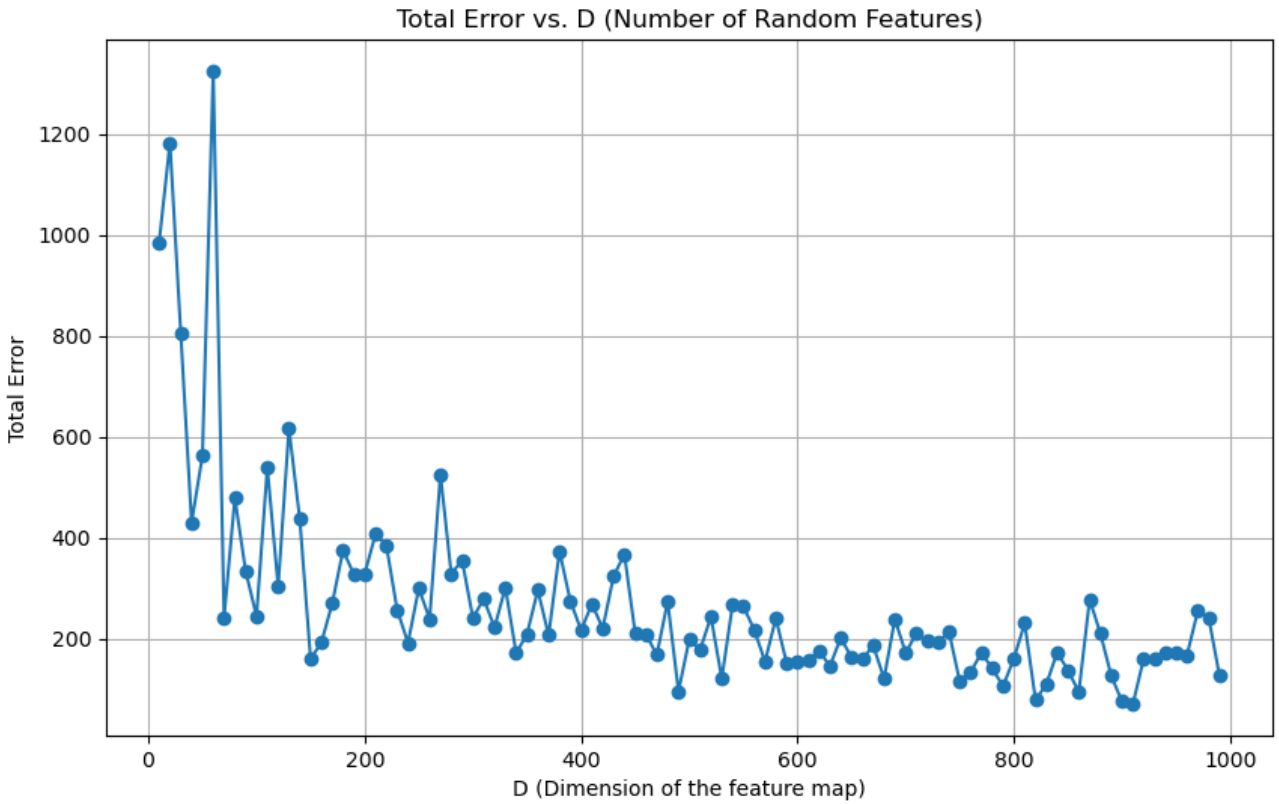

Figure: Total error plot showing the sum of absolute differences between the actual and approximate kernels over a range of feature map dimensions D.

Error Analysis - Worst-Case Error

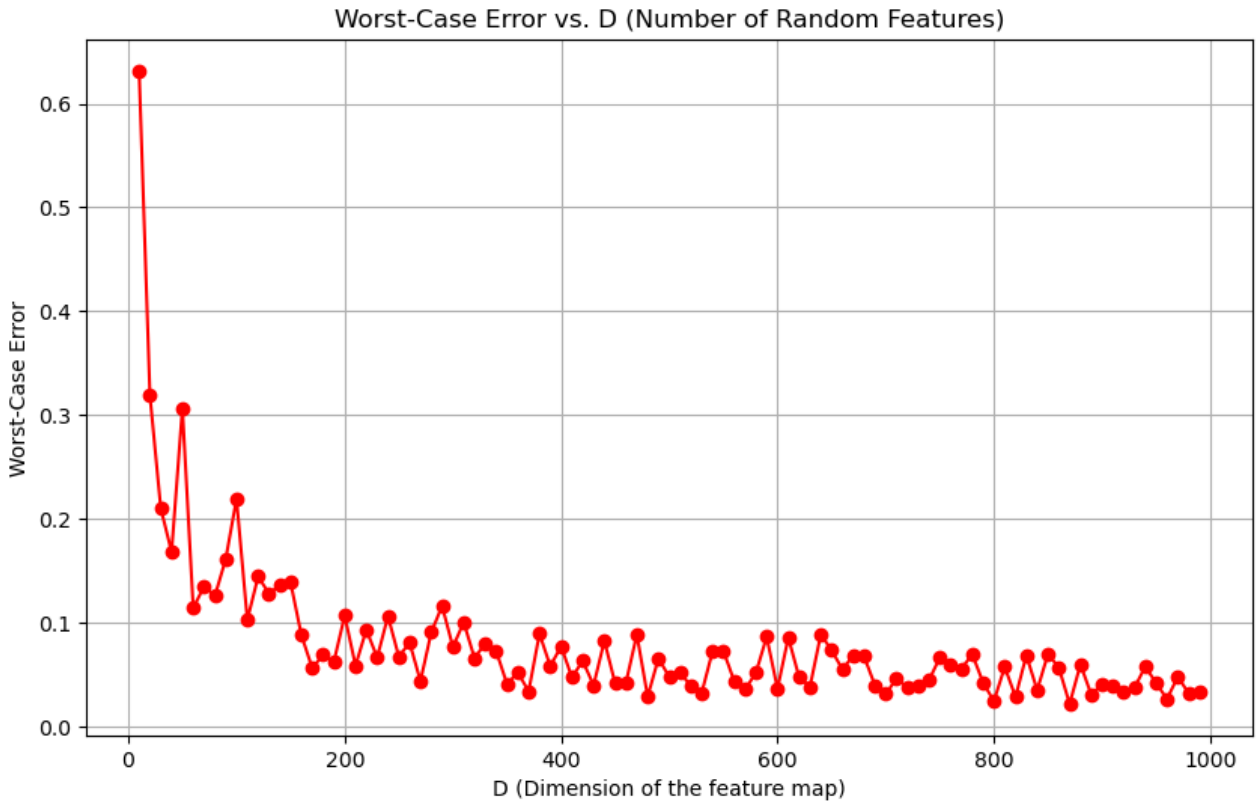

Figure: Worst-case error plot illustrating the maximum absolute difference between the actual and approximate kernels as a function of D.

References

[1] Ali Rahimi and Benjamin Recht. "Random features for large-scale kernel machines". In: *Advances in neural information processing systems* 20 (2007).