Summary of "Random features for large-scale kernel machines" by Ali Rahimi and Benjamin Recht

Summary by Ryan Farell

Contents

1 Full Citation

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

2 Paper Summary

- **Dual Representation of Gaussian Processes**: A stationary Gaussian process GP is considered, which can be characterized in two equivalent forms:
  - Spatial Domain Representation: Through its covariance function $K(\mathbf{x}, \mathbf{y})$, where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, describing correlations in the spatial or input domain.
  - Frequency Domain Representation: Via its power spectral density $S(\boldsymbol{\omega})$ in the frequency domain $\boldsymbol{\omega}$.

  The paper focuses on utilizing the power spectral representation to approximate the kernel function.
- **Kernel Trick and Positive Definite Functions**: For a positive definite function $k(\mathbf{x}, \mathbf{y})$, an inner product and a lifting function ϕ are defined such that $k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$.
- **Randomized Feature Map**: The paper proposes an explicit mapping of data to a lower-dimensional Euclidean space using a randomized feature map $\mathbf{z} \colon \mathbb{R}^d \to \mathbb{R}^D$, aiming to approximate the kernel evaluation:
$$k(\mathbf{x}, \mathbf{y}) \approx \mathbf{z}(\mathbf{x})^T \mathbf{z}(\mathbf{y}).$$
- **Dimensionality Reduction**: Unlike the lifting ϕ, the proposed map $\mathbf{z}$ is low-dimensional, facilitating the use of fast linear learning methods to approximate the outcomes of nonlinear kernel machines.
- **Approximation of Shift-Invariant Kernels**: Random Fourier Features (RFF) uniformly approximate shift-invariant kernels $k(\mathbf{x} - \mathbf{y})$ within an error bound ϵ using only $D = O\left(\frac{d}{\epsilon^2} \log \frac{1}{\epsilon^2}\right)$ dimensions.

3 Concept Review

3.1 Kernel Trick

The kernel trick is a method used in machine learning to enable algorithms that depend only on the inner product between pairs of input points to operate in a higher-dimensional feature space without explicitly computing the coordinates in that space. It utilizes the property that any positive definite function $k(\mathbf{x}, \mathbf{y})$, where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, defines an inner product in some feature space. This feature space is associated with a lifting function ϕ, such that the inner product between the transformed data points is given by $\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle = k(\mathbf{x}, \mathbf{y})$. This approach allows algorithms to operate in this implicitly defined feature space without directly computing the high-dimensional representations of the data. The kernel matrix, which consists of the kernel function k applied to all pairs of data points, becomes a central component of such algorithms, representing the inner products in the feature space.

3.2 Bochner's Theorem

Theorem 3.1 (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.*

3.3 Random Fourier Features

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

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

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

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

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

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

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

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

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 $k(\mathbf{x}, \mathbf{y})$:

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

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

3.4 Hoeffding's Inequality

Given independent random variables $X_1, X_2, \ldots, X_n$ where each X_i takes values in $[a_i, b_i]$, let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ be their average and let $E[\bar{X}] = \mu$ be the expected value of the average. Then, for any $t > 0$, Hoeffding's inequality states that

$$
P(|\bar{X}-\mu| \geq t) \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i-a_i)^2}\right).
$$