**Introduction**

We revisit Rahimi and Recht (2007)'s kernel random Fourier features (RFF) method through the lens of the PAC-Bayesian theory.

New perspective on RFF
- The Fourier transform is a prior distribution on trigonometric hypotheses.
- Then we learn a pseudo-posterior by minimizing PAC-Bayesian bounds.

Two learning approaches
- A kernel alignment algorithm.
- A landmarks-based similarity measure learning.

**Random Fourier Features (RFF)**

Let's consider a translation invariant kernel $k: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$:

$$k(x, x') = k(x - x') = k(\delta)$$

with $\delta = x - x'$. Denote $p(\omega)$ the Fourier transform of $k(\delta)$,

$$p(\omega) = \frac{1}{|2\pi|^d} \int_{\mathbb{R}^d} k(\delta) e^{i\omega \cdot \delta} d\delta$$

Thus,

$$k(x, x') = \int_{\mathbb{R}^d} p(\omega) e^{i\omega \cdot (x - x')} d\omega = E_{\omega \sim p} \left[ \cos (\omega \cdot (x - x')) \right]$$

Example: Gaussian Kernel

The Gaussian kernel is given by

$$k(x, x') = \exp \left(-\frac{1}{2\sigma^2} \|x - x'\|^2\right)$$

Its Fourier transform is

$$p(\omega) = \left(\frac{1}{2\pi}\right)^d e^{-\frac{1}{2}\|\omega\|^2} = N(\omega, 0, \frac{1}{2\pi})$$

**PAC-Bayesian Theory**

Given a data distribution $D$, a training set $S = \{(x_i, y_i), \ldots, (x_n, y_n)\} \sim D^n$, with $x_i \in \mathbb{R}^d$ and $y_i \in \{1, \ldots, n\}$, a loss $L: \{1, \ldots, n\} \rightarrow [0, 1]$, and a predictor $f \in F$:

$$L_{S/f}(f) = E_{(x, y) \sim D} L(f(x), y)$$

Given a RFF $\hat{k}$ with $\omega \sim p$, the RFF loss is

$$\hat{L}_{\omega}(f) = \frac{1}{n} \sum_{i=1}^{n} L(f(x_i), y_i)$$

**PAC-Bayesian Theorem (Lever et al. 2013)**

For $t > 0$, for any prior $p$ on $F$, with probability $1-\varepsilon$ on the choice of $S \sim D^n$, we have for all posterior distribution $\rho$ on $F$:

$$E_{f \sim \rho} L_{S/f}(f) \leq \hat{E}_{\omega} \hat{L}_{\omega}(f) + \frac{t}{2} \left(1 + \frac{1}{2} \log \frac{1}{\varepsilon}\right)$$

Thus, the optimal posterior distribution $\rho^*$ is:

$$\rho^*(f) = \frac{1}{Z} \left(1 + \frac{1}{2} \log \frac{1}{\varepsilon}\right) \hat{L}_{\omega}(f)$$

**PAC-Bayesian Theory for RFF**

Idea: See the Fourier transform of a kernel as a prior over predictors.

$$k(x, x') = E_{\omega \sim p} \left[ \cos (\omega \cdot (x - x')) \right]$$

with $h_\omega(\delta) = \cos (\omega \cdot \delta)$.

**Kernel alignment loss**

Loss of a predictor $h_\omega$ on $\delta$, $\lambda = \Delta_{\omega} \lambda$, given by two draws according to $D$:

$$\ell(h_\omega(\delta), \lambda) = \frac{1}{\lambda} - \frac{1}{\lambda} h_\omega(\delta)$$

with $\lambda = 1$ if $y = y'$, $-1$ otherwise.

**Goal:** Learn a posterior $q$ to minimize the loss of $k_{\hat{h}}$

$$L_{\omega}(k_{\hat{h}}) = E_{(x, y) \sim D, \omega \sim p} \ell(h_\omega(\delta), \lambda) = E_{\omega \sim q} L_{\omega}(h_\omega)$$

**Empirical loss estimation**

Given $S \sim D^n$, an unbiased second-order estimator of $L_{\omega}(k_{\hat{h}})$ is:

$$\hat{L}_{\omega}(k_{\hat{h}}) = \frac{1}{n} \sum_{i=1}^{n} \ell(h_{\omega}(\delta), \lambda)$$

**Approach 1: Kernel Alignment Algorithm**

For $t > 0$ and a prior distribution $\rho$ over $\mathbb{R}^d$, with probability $1-\varepsilon$, we have $\forall q$ on $\mathbb{R}^d$:

$$L_{\rho}(k_{\hat{h}}) \leq \hat{L}_{\omega}(k_{\hat{h}}) + \frac{t}{2} \left(1 + \frac{1}{2} \log \frac{1}{\varepsilon}\right)$$

Consider a uniform prior $P$ on $N$ random features drawn according to $p$,

then $Q(h_{\omega})$ for $m = 1, \ldots, N$:

$$Q(h_{\omega}) = \frac{1}{N} \exp \left(-L_{\omega}(h_{\omega})\right)$$

We then sample $D \sim N$ features according to $Q$ and execute an SVM using the learned kernel

$$\hat{k}_{\omega}(x, x') = \frac{1}{N} \sum_{i=1}^{N} h_{\omega}(x - x')$$

**Experiments**

![Figure 1: Train and test error of the kernel learning approaches according to the number of random features D.](image)

**Approach 2: Similarity Measure Learning**

For an $L$-subset of landmarks $L = \{(x_i, y_i)\} \subseteq S$ chosen such that $(x_i, y_i) \in S$:

$$L_{\omega}(k_{\hat{h}}) = E_{(x, y) \sim D} \ell(h_{\omega}(x - x'), \lambda(y, y'))$$

**PAC-Bayesian Corollary**

For $t > 0$, $i \in \{1, \ldots, L\}$, and a prior distribution $p$ over $\mathbb{R}^d$, with probability $1-\varepsilon$, we have $\forall q$ on $\mathbb{R}^d$:

$$L_{\rho}(k_{\hat{h}}) \leq \hat{L}_{\omega}(k_{\hat{h}}) + \frac{t}{2} \left(1 + \frac{1}{2} \log \frac{1}{\varepsilon}\right)$$

For all $x_i \in L$, consider a uniform prior $P$ on $D$ random features drawn according to $p$,

then learn $Q(h_{\omega})$:

$$Q(h_{\omega}) = \frac{1}{L} \exp \left(-\hat{L}_{\omega}(h_{\omega})\right)$$

Then execute an SVM on the following mapping

$$x \mapsto \left(\hat{k}_{\omega}(x, x), \hat{k}_{\omega}(x, x_1), \ldots, \hat{k}_{\omega}(x, x_L)\right)$$

**Experiments**

Intuition of the method with toy problem:

![Figure 2: From a RBF prior on 0 randomly selected landmarks (blue row), PB-Landmarks (red row) successfully finds a representation from which the linear SVM can predict well.](image)

**Real data results**

![Figure 3: Behavior of the landmarks-based approach according to the percentage of landmarks on the dataset "ads".](image)