Insights, resources, and application of "Large Scale (non-)Linear Models"

BY YALIN LIAO

I have read the paper - Random Features for Large-Scale Kernel Machines. The summary and understanding are detailed in the following paragraphs.

Unlike previous work, instead of considering efficiently computing the Gram matrix, this paper approximates the kernel function directly by the randomized feature mapping. The insight of the first introduced feature mapping is inherited from Bochner's theorem,

Theorem 1. *A continuous and shift-invariant kernel $k(x,y)=k(x-y)$ is positive definite if and only if it can be represented as as $k(x-y)=\int p(\omega)e^{i\omega^T(x-y)}d\omega=\mathbb{E}_\omega[e^{i\omega^T x}(e^{i\omega^T y})^*]$[1] where $p(\omega)$ is a probability density function.*

Note that $k(x,y)$ and $p(\omega)$ are real-valued functions, and

$$\begin{aligned} k(x,y) &= \int p(\omega)e^{i\omega^T(x-y)}d\omega \\ &= \int p(\omega)[\cos(\omega^T(x-y))+i\sin(\omega^T(x-y))]d\omega \\ &= \int p(\omega)\cos(\omega^T(x-y))d\omega + i\int p(\omega)\sin(\omega^T(x-y))d\omega \end{aligned}$$

Then we have $\int p(\omega)\sin\omega^T(x-y)d\omega=0$ and thus, $k(x-y)=\int p(\omega)\cos\omega^T(x-y)d\omega$. We know $p(\omega)$ is the Fourier transform of $k(\delta)$, $p(\omega)=\frac{1}{2\pi}\int k(\delta)e^{-j\omega^T\delta}d\delta$.[2] How the random feature mapping is defined? Let $z_\omega(x)=\sqrt{2}\cos(\omega^T x+b)$. Then[3]

$$\begin{aligned} \mathbb{E}_\omega[z_\omega(x)z_\omega(y)] &= \int p(\omega)\sqrt{2}\cos(\omega^T x+b)\sqrt{2}\cos(\omega^T y+b)d\omega \\ &= \int p(\omega)2\cos(\omega^T x+b)\cos(\omega^T y+b)d\omega \\ &= \int p(\omega)[\cos(\omega^T x+b+\omega^T y+b)+\cos(\omega^T x+b-\omega^T x-b)]d\omega \\ &= \int p(\omega)\cos\omega^T(x-y)d\omega + \int p(\omega)\cos 2(\omega^T(x+y)+b)d\omega \\ &? \int p(\omega)\cos\omega^T(x-y)d\omega, b\sim U[0,2\pi] \text{ such that } \int p(\omega)\cos 2(\omega^T(x+y)+b)d\omega=0 \\ &= k(x-y) \end{aligned}$$

i.e., $\mathbb{E}_\omega[z_\omega(x)z_\omega(y)]=k(x,y)$. By Monte Carlo method,

$$\frac{1}{D}\sum_{j=1}^{D} z_{\omega_j}(x)z_{\omega_j}(y)\approx k(x,y)$$

1. Inverse Fourier transform of $p(w)$
2. $\delta=x-y$
3. $\cos A\cos B=\frac{1}{2}[\cos(A+B)+\cos(A-B)]$

where $\omega_j \sim p(\omega)$ if D is sufficient large. Naturally, define the feature map (finite dimension)

$$z(x) = \frac{1}{\sqrt{D}}(z_{w_1}(x), \cdots, z_{w_D}(x))$$

Then

$$z(x)^T z(y) \approx k(x, y)$$

The story is perfect. I just have two questions:

(1) Even we know the density function $p(\omega)$, often it's not easy to draw samples from $p(\omega)$. How do we efficiently draw samples $\omega_1, \omega_2, \cdots, \omega_d$ from $p(\omega)$?

(2) Even though the claim 2 shows the inner product based on random feature mapping can approximate the kernel, I still don't think the defined inner product can really approximate the kernel when $D < d$ since the general Monte Carlo methos is not efficient.