CS189: Discussion 12

Stephen Tu

December 1, 2016

1 Random Fourier Features

In this discussion, we will study the random lifting trick from HW2. This technique was first introduced in the context of machine learning by [2]. Recall that given a set of points $\{x_i\}_{i=1}^n$ with $x_i \in \mathbb{R}^d$, we constructed a random feature map $\phi : \mathbb{R}^d \longrightarrow \mathbb{R}^D$ as follows

$$\phi(x) = \sqrt{\frac{2}{D}} \begin{bmatrix} \cos(w_1^\mathsf{T} x + b_1) \\ \cos(w_2^\mathsf{T} x + b_2) \\ \vdots \\ \cos(w_D^\mathsf{T} x + b_D) \end{bmatrix} \;, \;\; w_i \overset{\text{iid}}{\sim} N(0, \sigma^2 I) \;, \; b_i \overset{\text{iid}}{\sim} \mathrm{Unif}([0, 2\pi]) \;. \tag{1}$$

We then used the map ϕ to do learning with the dataset $\{\phi(x_i)\}_{i=1}^n$. The question now is, why does this work? What is the principle behind such a map? Our eventual goal will be to show that ϕ is an approximation to the feature map induced by a kernel. We first start by reviewing basic kernel definitions.

1.1 Positive definite kernel functions

Recall the following definition of a positive definite kernel.

Definition 1. Let X be a set and $k : X \times X \longrightarrow \mathbb{R}$ be a symmetric function. We say that k is a *positive definite kernel* if for all $n \geqslant 1$, $x_1, ..., x_n \in X$, and $\alpha_1, ..., \alpha_n \in \mathbb{R}$,

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(x_i, x_j) \geqslant 0 \;.$$

This definition means that for any dataset $\{x_i\}_{i=1}^n$, the $n \times n$ kernel gram matrix K defined as $K_{ij} = k(x_i, x_j)$ is a positive semi-definite matrix.

We now consider a special class of kernel functions, which are called translation-invariant kernels. For the rest of this note, we will only consider kernels on $\mathbb{R}^d$, but we note that these ideas can be extended more generally to locally compact abelian groups. We now state another definition.

Definition 2. Let $k : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$ be a positive definite kernel. We say that k is *translation-invariant* if $k(x_1, x_2) = g(x_1 - x_2)$ for some function $g : \mathbb{R}^d \longrightarrow \mathbb{R}$.

Note that since k is symmetric, it must be the case that $g(x) = g(-x)$, i.e. g is symmetric around the origin. Let us now consider several examples of translation-invariant kernels on $\mathbb{R}^d$. The most classical example, which we have already seen, is the Gaussian kernel

$$k(x_1, x_2) = g(x_1 - x_2) \;, \;\; g(\Delta) = \exp(-\|\Delta\|_2^2/2\sigma^2) \;.$$

A few more examples include the Laplacian kernel

$$k(x_1, x_2) = g(x_1 - x_2) \;, \;\; g(\Delta) = \exp(-\|\Delta\|_1/\sigma) \;,$$

and the sinc kernel

$$k(x_1, x_2) = g(x_1 - x_2) \ , \ \ g(\Delta) = \frac{\sin(a(x-y))}{\pi(x-y)} \ .$$

1.2 Fourier features of the Gaussian kernel

We will momentarily focus on the Gaussian kernel and derive a nice property of its Fourier transform. We first recall the definition of the Fourier transform of a function.

Definition 3. Let $f : \mathbb{R}^d \longrightarrow \mathbb{C}$ be an $L^1(\mathbb{R}^d)$ function. The Fourier transform of f, which we denote as $\widehat{f}$, is defined as

$$\widehat{f}(\xi) = \int_{\mathbb{R}^d} f(x) e^{-j\xi \cdot x} \, dx \ .$$

Some remarks are in order. First, the operator $f \mapsto \widehat{f}$ can be extended uniquely to map L^2 functions to L^2; this is the content of Plancherel's theorem. Second, in the above expression $j = \sqrt{-1}$ and the notation $\xi \cdot x$ refers to the inner product $\xi^{\mathsf{T}} x$. Third, the Fourier transform is often defined with different constants (so the definition above may or may not be the one you are used to).

We now show a very fundamental fact about Fourier transforms.

Lemma 1. Let $f(x) = e^{-zx^2/2}$ for some positive $z > 0$. Then

$$\widehat{f}(\xi) = (2\pi)^{1/2} z^{-1/2} e^{-\xi^2/2z} \ .$$

Proof. Define the function $g(\xi) = \widehat{f}(\xi)$. We exhibit an ordinary differential equation (ODE) which g satisfies. Observe that by differentiating under the integral,

$$\begin{aligned} \frac{d}{d\xi} g(\xi) &= \frac{d}{d\xi} \int_{-\infty}^{\infty} e^{-zx^2/2} e^{-j\xi x} \, dx \\ &= \int_{-\infty}^{\infty} e^{-zx^2/2} (-jx) e^{-j\xi x} \, dx = (j/z) \int_{-\infty}^{\infty} e^{-j\xi x} \frac{d}{dx} \left(e^{-zx^2/2} \right) \, dx \ . \end{aligned}$$

Integrating by parts,

$$\begin{aligned} (j/z) \int_{-\infty}^{\infty} e^{-j\xi x} \frac{d}{dx} \left(e^{-zx^2/2} \right) \, dx &= (j/z) \left[e^{-j\xi x} e^{-zx^2/2} \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} (-j\xi) e^{-j\xi x} e^{-zx^2/2} \right] \\ &= -\frac{\xi}{z} g(\xi) \ . \end{aligned}$$

Therefore,

$$\frac{d}{d\xi} g(\xi) = -\frac{\xi}{z} g(\xi) \ . \tag{2}$$

It is straightforward to check that the function $g(\xi) = Ce^{-\xi^2/2z}$ satisfies the ODE in (2) for any constant C. We now derive the correct C by satisfying the boundary condition $C = g(0) = \int e^{-zx^2/2} \, dx$. But $g(0)$ is the normalization constant of a $N(0, 1/z)$ distribution. Hence, $C = (2\pi)^{1/2} z^{-1/2}$, which is the desired result. □

Now let us consider the d-dimensional Gaussian function $f(x) = e^{-z\|x\|_2^2/2}$ for $z > 0$. Taking the Fourier transform and using the previous result,

$$\begin{aligned}\widehat{f}(\xi) &= \int_{\mathbb{R}^d} e^{-z\|x\|_2^2/2} e^{-j\xi\cdot x}\, dx = \int_{\mathbb{R}^d} \prod_{i=1}^{d} e^{-zx_i^2/2} e^{-j\xi_i x_i}\, dx = \prod_{i=1}^{d} \int_{-\infty}^{\infty} e^{-zx_i^2/2} e^{-j\xi_i x_i}\, dx_i \\ &= \prod_{i=1}^{d} \widehat{f}(\xi_i) = (2\pi)^{d/2} z^{-d/2} e^{-\|\xi\|_2^2/2z} .\end{aligned}$$

Now, let us set $z = \sigma^2$ and $\gamma = 1/\sigma$ and study this equation. We have shown that

$$e^{-\|\xi\|_2^2/2\sigma^2} = \frac{1}{(2\pi\gamma^2)^{-d/2}} \int_{\mathbb{R}^d} e^{-j\xi\cdot x} e^{-\|x\|_2^2/2\gamma^2}\, dx .$$

If we look at the right hand side for a moment, we will notice that it is simply the expectation over a multivariate Gaussian. That is, if $w \sim N(0, \gamma^2)$, then

$$e^{-\|\Delta\|_2^2/2\sigma^2} = \mathbb{E}_w[e^{-j\Delta\cdot w}] .$$

The left hand side, however, simply the Gaussian kernel. Hence, if we define $\varphi_w(x) = e^{jx\cdot w}$,

$$k(x_1, x_2) = \mathbb{E}_w[e^{-j(x_2 - x_1)\cdot w}] = \mathbb{E}_w[e^{-jx_2\cdot w} e^{jx_1\cdot w}] = \mathbb{E}_w[\varphi_w(x_1)\overline{\varphi_w(x_2)}] . \tag{3}$$

We now take advantage of the fact that $k(x_1, x_2)$ is real valued to simplify the expectation on the right hand side.

Lemma 2. Let w be a random vector such that for every pair $x_1, x_2 \in \mathbb{R}^d$ the expectation

$$\mathbb{E}_w[\varphi_w(x_1)\overline{\varphi_w(x_2)}]$$

exists and is real valued. Define $\phi_{w,b}(x) = \sqrt{2}\cos(w^\mathsf{T} x + b)$. Then

$$\mathbb{E}_w[\varphi_w(x_1)\overline{\varphi_w(x_2)}] = \mathbb{E}_{w,b}[\phi_{w,b}(x_1)\phi_{w,b}(x_2)] ,$$

where $b \sim \mathrm{Unif}([0, 2\pi])$ and is independent of w.

Proof. First, recall the following identity,

$$\cos(\alpha - \beta) = \cos\alpha\cos\beta + \sin\alpha\sin\beta .$$

By Euler's identity and this cosine identity,

$$\begin{aligned}\varphi_w(x_1)\overline{\varphi_w(x_2)} &= (\cos(w^\mathsf{T} x_1) + j\sin(w^\mathsf{T} x_1))(\cos(w^\mathsf{T} x_2) - j\sin(w^\mathsf{T} x_2)) \\ &= \cos(w^\mathsf{T} x_1)\cos(w^\mathsf{T} x_2) + \sin(w^\mathsf{T} x_1)\sin(w^\mathsf{T} x_2) \\ &\qquad + j(\sin(w^\mathsf{T} x_1)\cos(w^\mathsf{T} x_2) - \cos(w^\mathsf{T} x_1)\sin(w^\mathsf{T} x_2)) \\ &= \cos(w^\mathsf{T}(x_1 - x_2)) \\ &\qquad + j(\sin(w^\mathsf{T} x_1)\cos(w^\mathsf{T} x_2) - \cos(w^\mathsf{T} x_1)\sin(w^\mathsf{T} x_2)) .\end{aligned}$$

Hence, since $\mathbb{E}_w[\varphi_w(x_1)\overline{\varphi_w(x_2)}] \in \mathbb{R}$, we have that

$$\mathbb{E}_w[\varphi_w(x_1)\overline{\varphi_w(x_2)}] = \mathbb{E}_w[\cos(w^\mathsf{T}(x_1 - x_2))] . \tag{4}$$

On the other hand, using another identity

$$2\cos\alpha\cos\beta = \cos(\alpha-\beta) + \cos(\alpha+\beta) \ ,$$

we have that

$$\begin{aligned}\phi_{w,b}(x_1)\phi_{w,b}(x_2) &= 2\cos(w^\mathsf{T}x_1 + b)\cos(w^\mathsf{T}x_2 + b) \\ &= \cos(w^\mathsf{T}(x_1-x_2)) + \cos(w^\mathsf{T}(x_1+x_2)+2b) \ .\end{aligned}$$

Taking expectations we conclude that

$$\begin{aligned}\mathbb{E}_{w,b}[\phi_{w,b}(x_1)\phi_{w,b}(x_2)] &= \mathbb{E}_w[\cos(w^\mathsf{T}(x_1-x_2))] + \mathbb{E}_{w,b}[\cos(w^\mathsf{T}(x_1+x_2)+2b)] \\ &= \mathbb{E}_w[\cos(w^\mathsf{T}(x_1-x_2))] + \mathbb{E}_w[\mathbb{E}_b[\cos(w^\mathsf{T}(x_1+x_2)+2b)]] \\ &\overset{(a)}{=} \mathbb{E}_w[\cos(w^\mathsf{T}(x_1-x_2))] \\ &\overset{(b)}{=} \mathbb{E}_w[\varphi_w(x_1)\overline{\varphi_w(x_2)}] \ .\end{aligned}$$

The identity (a) holds since $\int_0^{2\pi}\cos(a+2x)\,dx = 0$ for any fixed $a \in \mathbb{R}$, and (b) holds by (4). □

Combining (3) and Lemma 2, we have that

$$k(x_1,x_2) = \mathbb{E}_{w,b}[\phi_{w,b}(x_1)\phi_{w,b}(x_2)] \ . \tag{5}$$

In light of (5), the feature map we defined in (1) now makes sense. For a fixed x_1, x_2, we have that

$$\langle\phi(x_1),\phi(x_2)\rangle = \frac{1}{D}\sum_{i=1}^{D}\phi_{w_i,b_i}(x_1)\phi_{w_i,b_i}(x_2) \approx \mathbb{E}_{w,b}[\phi_{w,b}(x_1)\phi_{w,b}(x_2)] = \exp(-\sigma^2\|x_1-x_2\|_2^2/2) \ .$$

We will make the middle approximate equality precise in a little bit, but the intuition is clear. Note that the variance from which we draw our random vectors w_i is the *inverse* of the variance which shows up in the kernel– this is a consequence of the way the variance scales with the Fourier transform of a Gaussian.

1.3 Beyond Gaussian kernels: Bochner's theorem

The calculations of the preceding section showed that we can do random Fourier features for the Gaussian kernel. But it turns out this construction above extends to all translation-invariant kernels via a nice theorem in harmonic analysis called Bochner's theorem.

Theorem 1. (Bochner's theorem, informally stated and specialized to $\mathbb{R}^d$) Let k be a translation-invariant kernel on $\mathbb{R}^d$, i.e. $k(x_1,x_2) = g(x_1 - x_2)$. Then g is the Fourier transform of a non-negative Radon measure, i.e.

$$g(\xi) = \int_{\mathbb{R}^d} e^{-j\xi\cdot x}\,d\mu \ .$$

Furthermore, the Fourier transform of any non-negative Radon measure gives rise to a translation-invariant kernel.

See Section 1.4.3 of [3] for a proof of Bochner's theorem. Bochner's theorem is important because it says the following in the context of random features: one can always (after scaling the kernel by some irrelevant constant factor) write any translation-invariant kernel in the same way we wrote the Gaussian kernel as an expectation in (3), except for a general translation-invariant kernel the distribution will be different (given by the inverse Fourier transform). Then one can apply Lemma 2, as we did for Gaussian, and recover a very similar random feature map construction.

For example, let us look at the sinc kernel. We know from standard Fourier calculations that for $\gamma > 0$,

$$\widehat{\mathrm{rect}(\gamma x)}(\xi) = \frac{2}{\xi}\sin(\xi/2\gamma) \; .$$

On the other hand,

$$\begin{aligned}\widehat{\mathrm{rect}(\gamma x)}(\xi) &= \int_{-\infty}^{\infty} e^{-j\xi x}\mathrm{rect}(\gamma x)\; dx = \int_{-\infty}^{\infty} e^{-j\xi x}\mathbf{1}_{\{|x| \leqslant 1/2\gamma\}}\; dx \\ &= \frac{1}{\gamma}\int_{-\infty}^{\infty} e^{-j\xi x}(\gamma \mathbf{1}_{\{|x|\leqslant 1/2\gamma\}})\; dx = \frac{1}{\gamma}\mathbb{E}_{w\sim \mathrm{Unif}([-1/2\gamma, 1/2\gamma])}[e^{-j\xi w}] \; .\end{aligned}$$

Hence,

$$\frac{\pi}{(2\gamma)^{-1}}\frac{\sin((2\gamma)^{-1}\xi)}{\pi\xi} = \frac{2\gamma}{\xi}\sin(\xi/2\gamma) = \mathbb{E}_{w\sim \mathrm{Unif}([-1/2\gamma,1/2\gamma])}[e^{-j\xi w}] \; .$$

Therefore, for the sinc kernel $k(x_1, x_2) = \frac{\sin(a(x_1-x_2))}{\pi(x_1-x_2)}$, we have

$$\frac{\pi}{a}k(x_1,x_2) = \mathbb{E}_{w\sim \mathrm{Unif}([-a,a])}[e^{-j(x_1-x_2)w}] \; .$$

The sampling distribution on the RHS is not surprising once we note that the sinc kernel is the reproducing kernel of the Paley-Wiener space of band-limited functions (see e.g. [1] for background on reproducing kernel Hilbert spaces),

$$H = \{f \in L^2(\mathbb{R}) : \mathrm{supp}(\widehat{f}) \in [-a,a]\} = \overline{\mathrm{span}}\{y \mapsto \frac{\sin(a(y-x))}{\pi(y-x)} : x \in \mathbb{R}\} \; .$$

Above, the notation supp refers to the support of a function (i.e. $\mathrm{supp}(g) = \{x : g(x) \neq 0\}$[1]) and $\overline{\mathrm{span}}$ refers to the closure of the span of the input. Thus, functions that are represented as linear combinations of the sinc kernel have band-limited Fourier transform within the range $[-a, a]$, and the random feature construction samples frequencies uniformly within this range.

1.4 Approximation error

The next question we consider is how much error is introduced in our random feature construction (1) compared to directly using the kernel? We will first state some results concerning approximation over finite sets, and then state a more general theorem over compact sets.

1.4.1 Finite sets

The main probabilistic tool we will use is Hoeffding's inequality. The version we will use is stated below.

Theorem 2. (Hoeffding's inequality) Let $X_1, ..., X_n$ be independent random variables such that $X_i \in [a_i, b_i]$ almost surely. Put $\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i$. Then,

$$\mathbb{P}\left(|\overline{X_n} - \mathbb{E}[\overline{X_n}]| \geqslant t\right) \leqslant 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \; .$$

[1] This is not technically correct since we want to allow the function to be non-zero on a null set outside the support.

We now fix x_1, x_2 and apply Hoeffding's inequality to the random variables $\phi_{w_i,b_i}(x_1)\phi_{w_i,b_i}(x_2)$, $i = 1, ..., D$. Clearly, $|\phi_{w_i,b_i}(x_1)\phi_{w_i,b_i}(x_2)| \leqslant 2$ for every w_i, b_i. Hence, applying Hoeffding's inequality we conclude that,

$$\mathbb{P}(|\langle \phi(x_1), \phi(x_2) \rangle - k(x_1, x_2)| \geqslant t) \leqslant 2\exp(-Dt^2/8) .$$

Now this only holds for a fixed x_1, x_2. Supposing we have n data points $\{x_i\}_{i=1}^n$, a simple union bound over all $\binom{n}{2}$ pairs yields

$$\mathbb{P}\left(\max_{1\leqslant i,j\leqslant n} |\langle \phi(x_i), \phi(x_j) \rangle - k(x_i, x_j)| \geqslant t\right) \leqslant n^2 \exp(-Dt^2/8) .$$

We can now fix an $\varepsilon > 0$ and $\delta \in (0, 1)$, and ask how many random features D we need such that with probability at least $1 - \delta$,

$$\max_{1\leqslant i,j\leqslant n} |\langle \phi(x_i), \phi(x_j) \rangle - k(x_i, x_j)| \leqslant \varepsilon .$$

A little bit of arithmetic yields that a sufficient condition on D is

$$D \geqslant \frac{16}{\varepsilon^2} \log\left(\frac{n}{\delta}\right) .$$

Next, we can ask how faithful of an approximation is the $n \times n$ gram matrix $\widetilde{K}$ defined as $\widetilde{K}_{ij} = \langle \phi(x_i), \phi(x_j) \rangle$ to the original $n \times n$ kernel matrix K.

Theorem 3. (Equation 6.5.7, [5]) Fix a $\varepsilon > 0$. For a fixed dataset $\{x_i\}_{i=1}^n$, we have that as long as D satisfies

$$D \geqslant \frac{4}{\varepsilon^2} \frac{n}{\|K\|} \log(2n) ,$$

then the following operator norm error holds in expectation,

$$\mathbb{E}\|\widetilde{K} - K\| \leqslant (\frac{1}{\varepsilon} + \frac{1}{\varepsilon^2})\|K\|$$

Similar results also hold with high probability.

Proof. This is a standard application of the matrix Bernstein inequality. See [5] for more details on matrix concentration inequalities. □

1.4.2 Compact sets

We now turn to results concerning the approximation error of random features over a compact set. The following result is an informal statement of Theorem 1 from [4], which improves the uniform convergence result (Claim 1) from [2].

Theorem 4. (Informal statement of Theorem 1, [4]) Let $S \subseteq \mathbb{R}^d$ be a compact set, and let $|S|$ denote its diameter (i.e. $|S| = \sup_{x,y\in S}\|x - y\|_2$). Fix any $\varepsilon > 0$ and $\delta \in (0, 1)$. Then as long as

$$D \geqslant O\left(\frac{d}{\varepsilon^2} \log\left(\frac{|S| + 1}{\delta}\right)\right) ,$$

we have that with probability at least $1 - \delta$,

$$\sup_{x_1,x_2\in S} |\langle \phi(x_1), \phi(x_2) \rangle - k(x_1, x_2)| \leqslant \varepsilon .$$

Note that the proof of Theorem 1 is quite technical. On the other hand, the proof of Claim 1 of [2] is more approachable, at the expense of a sub-optimal rate.

References

[1] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, 2004.

[2] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *NIPS*, 2007.

[3] W. Rudin. *Fourier Analysis on Groups*. Wiley-Interscience, 1990.

[4] B. K. Sriperumbudur and Z. Szabó. Optimal rates for random fourier features. In *NIPS*, 2015.

[5] J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 2015.