Topics in Deep Learning Theory (Spring 2025)

Lecture 6: Random Feature Models

Instructor: Lei Wu Date: April 21, 2025

Abstract

In this lecture, we discuss the approximation and statistical properties of random feature models (RFMs). Originally introduced to accelerate kernel methods, RFMs have since revealed deep connections with neural networks. Understanding RFMs is therefore crucial for bridging the gap between kernel methods and neural network models.

1 Introduction

Kernel ridge regression (KRR). Given training data $\{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X}\times\mathbb{R}$ and a positive-definite kernel $k\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ with reproducing-kernel Hilbert space (RKHS) $\mathcal{H}_k$, KRR fitting the training data via

$$\hat{f}_\lambda \;=\; \operatorname*{argmin}_{f\in\mathcal{H}_k}\Big\{\frac{1}{n}\sum_{i=1}^n \big(f(x_i)-y_i\big)^2 + \;\lambda\|f\|_{\mathcal{H}_k}^2\Big\},$$

where $\lambda > 0$ is the regularisation parameter. By the representer theorem,

$$\hat{f}_\lambda(x) \;=\; \sum_{i=1}^n (\hat{\alpha}_\lambda)_i\, k(x_i, x), \qquad \hat{\alpha}_\lambda \;=\; (K+\lambda I_n)^{-1} y,$$

with kernel matrix $K = (k(x_i, x_j))_{1\le i,j\le n} \in \mathbb{R}^{n\times n}$. Naïvely computing $\hat{\alpha}_\lambda$ costs $\Omega(n^3)$ compute and $\Omega(n^2)$ memory—prohibitive when n is, for example, 10^6.

In the literature, there are many approximation methods to speed up the computation of $(K+\lambda I_n)^{-1}$, such as sketching techniques and Nystrom sampling. Among them, random feature approximation has emerged as one of the most popular ones in practice [Rahimi and Recht, 2007].

Specifically, assume the kernel k admits a *random-feature representation*

$$k(x, x') \;=\; \mathbb{E}_{w\sim\pi}\big[\varphi(x, w)\,\varphi(x', w)\big],$$

for some feature map $\varphi\colon \mathcal{X}\times\Omega \to \mathbb{R}$ and distribution π on Ω. Drawing m i.i.d. samples $\{w_j\}_{j=1}^m \sim \pi$ yields the Monte-Carlo approximation

$$k(x, x') \approx \hat{k}_m(x, x') \;=\; \frac{1}{m}\sum_{j=1}^m \varphi(x, w_j)\,\varphi(x', w_j).$$

Define the feature matrix by

$$\Phi_m = \Big(m^{-1/2}\varphi(x_i, w_j)\Big)_{1\le i\le n, 1\le j\le m} \in \mathbb{R}^{n\times m}.$$

Then, applying the Sherman–Morrison–Woodbury identity gives

$$(K+\lambda I_n)^{-1} \;\approx\; (\Phi_m\Phi_m^\top + \lambda I_n)^{-1}$$

$$= \lambda^{-1} I_n - \lambda^{-1}\Phi_m\big(\Phi_m^\top \Phi_m + \lambda I_m\big)^{-1}\Phi_m^\top, \tag{1}$$

reducing the computational complexity to $\Omega(m^3+m^2n)$ and the memory to $O(mn)$. Obviously, the improvement is substantial if we can choose $m \ll n$.

Random feature models (RFMs). It is easy to show that the above approximation procedure is equivalent to fit the following random feature model (RFM)

$$f_m(x;a) = \frac{1}{m}\sum_{j=1}^{m} a_j\,\varphi(x,w_j), \tag{2}$$

by solving the ridge-regularised least-squares problem

$$a_\lambda = \operatorname*{argmin}_{a\in\mathbb{R}^m}\Big\{\frac{1}{n}\sum_{i=1}^{n}\big(f_m(x_i;a)-y_i\big)^2 + \frac{\lambda}{m}\|a\|_2^2\Big\}. \tag{3}$$

Thus RFMs provide a scalable surrogate for KRR.

Note that RFM (2) has a natural connection with neural networks: if $\{w_j\}_{j=1}^m$ are also learnable, then it recovers two-layer neural network.

Two core questions.

Q1: Can the random feature approximation retain KRR's statistical guarantees, while allowing $m \ll n$ without sacrificing accuracy?

Q2: In what sense do RFMs bridge kernel methods and neural networks?

In this lecture, we focus on addressing **Q1** by studying the sample and parameter complexities required for RFMs to learn functions in $\mathcal{H}_k$. Question **Q2** will be the discussed in the next lecture.

2 Capacity-Controlled Approximation

We introduce the following natural approximation space for the RFM (2). As $m \to \infty$, we have

$$\lim_{m\to\infty}\frac{1}{m}\sum_{j=1}^{m} a_j\varphi(x,w_j) \to \int a(w)\varphi(x,w)\,\mathrm{d}\pi(w) =: f_a(x).$$

Definition 2.1 ($\mathcal{F}_{p,\pi}$ space)**.** For $p \in [1,+\infty)$, define

$$\mathcal{F}_{p,\pi} := \left\{ f_a : \inf_{a\in A_f}\|a\|_{L^p(\pi)} < \infty \right\},$$

where $A_f = \big\{a : f = \int a(w)\varphi(\cdot;w)\mathrm{d}\pi(w)\big\}$. The associated norm $\|\cdot\|_{\mathcal{F}_{p,\pi}}$ is given by

$$\|f\|_{\mathcal{F}_{p,\pi}} := \inf_{a\in A_f}\|a\|_{L^p(\pi)}.$$

Intuitively, $\mathcal{F}_{p,\pi}$ is the approximation space of $f_m(x;\theta)$ with the ℓ^p norm of parameters uniformly bounded. In this lecture, we restrict our attention to the case $p = 2$, namely $\mathcal{F}_{2,\pi}$, which coincides with the RKHS $\mathcal{H}_{k_\pi}$.

Lemma 2.2. $\mathcal{F}_2 = \mathcal{H}_{k_\pi}$.

Note that this is a straightforward conclusion of the feature perspective of RKHS discussed in Lecture 3. Here, we still provide a proof for completeness.

Proof. First, $k_\pi(\cdot, x) = \int \varphi(\cdot, w)\varphi(x, w)\mathrm{d}\pi(w)$. Thus, by definition, we have

$$\|k(\cdot, x)\|_{\mathcal{F}_2} \leq \int_\Omega \varphi^2(x, w)\mathrm{d}\pi(w) < \infty,$$

implying $k(\cdot, x) \in \mathcal{F}_2$.

Second, for $f \in \mathcal{F}_2$, assume $a_f \in L^2(\pi)$ such that $f = \int a(w)\varphi(\cdot; w)\mathrm{d}\pi(w)$, we have

$$\langle f, k(\cdot, x)\rangle_{\mathcal{F}_2} = \int a_f(w)\varphi(\cdot, x)\mathrm{d}\pi(w) = f(x).$$

Thus, by the uniqueness of RKHS, we must have $\mathcal{F}_2 = \mathcal{H}_{k_\pi}$. □

By Holder inequality, it holds trivially that

$$\mathcal{F}_{\infty,\pi} \subset \mathcal{F}_{p,\pi} \subset \mathcal{F}_{q,\pi} \subset \mathcal{F}_{1,\pi} \text{ for } 1 \leq q \leq p \leq \infty.$$

We impose the following assumption for the feature $\varphi(\cdot, w)$.

Assumption 2.3. $\sup_{x\in\mathcal{X}, w\in\Omega} \varphi(x; w) \leq 1$

Under Assumption 2.3, it follows that $|k_\pi(x, x')| \leq \int |\varphi(x, w)\varphi(x', w)|\mathrm{d}\pi(w) \leq 1$, and an analogous bound holds for $\hat{k}_m(x, x')$.

Throughout this lecture, we make the following assumption on the target function:

Assumption 2.4. *Assume* $f^*(x) = \mathbb{E}_{w\sim\pi}[a(w)\varphi(x; w)]$ *with* $Q = \|f^*\|_{\mathcal{F}_{\infty,\pi}} < \infty$.

Note that the $\mathcal{F}_{\infty,\pi}$ is a subset of the RKHS $\mathcal{F}_{2,\pi} = \mathcal{H}_k$. All subsequent results can be extended to the RKHS $\mathcal{F}_{2,\pi}$ by invoking either the duality framework of [Chen et al., 2023] or the integral-operator techniques of [Bach, 2017]. We work with Assumption 2.4 purely for expositional clarity, sidestepping those additional technical complications while keeping the main ideas transparent.

Theorem 2.5 (Norm-controlled approximation)**.** *Suppose Assumption 2.4 hold. Let* $W = (w_1, \ldots, w_m)$ *with* $w_j \overset{iid}{\sim} \pi$, *and* $a(W) = (a(w_1), \ldots, a(w_m))^\top$. *Then, for any* $\delta \in (0, 1)$, *with probability* $1 - \delta$ *over the sampling of* W, *we have*

$$\|f_m(\cdot; a(W)) - f^*\|_{L^2(\rho)} \lesssim \frac{Q}{\sqrt{m}}(1 + \sqrt{\log(2/\delta)})$$

Moreover, $\max_{j\in[m]} |a_j| \leq Q$.

Proof. (1) Let $W = (w_1, \ldots, w_m)$, and $S_m(w_1, \ldots, w_m) = \|f_m(\cdot; a(W)) - f^*\|_{L^2(\rho)}$ Let $\tilde{W}$ be a copy of W but with i -th element different. Then,

$$\begin{aligned}\left|S_m(W) - S_m(\tilde{W})\right| &\leq \left\|f_m(\cdot; a(W)) - f_m(\cdot; a(\tilde{W}))\right\|_{L^2(\rho)} \\ &= \left\|\frac{1}{m} a(w_i)\,\varphi(\cdot; w_i) - \frac{1}{m} a(\tilde{w}_i)\,\varphi(\cdot; \tilde{w}_i)\right\|_{L^2(\rho)} \leq \frac{2Q}{m}.\end{aligned}$$

(2) By McDiarmid's inequality, with probability $1 - \delta$, we have

$$S_m(W) \lesssim \mathbb{E}_W[S_m(W)] + \sqrt{\frac{\log(2/\delta)}{m}} Q$$

(3) Next, we evaluate $\mathbb{E}S_m(W)$. Since

$$\mathbb{E}_W[\frac{1}{m}\sum_{j=1}^{m} a(w_i)\varphi(x; w_i)] = f^*(x)$$

we obtain

$$\begin{aligned}\mathbb{E}_W[S_m^2(W)] &= \mathbb{E}_W\mathbb{E}_x\left|\frac{1}{m}\sum_{j=1}^{m} a(w_j)\,\varphi(x; w_j) - f^*(x)\right|^2 \\ &= \mathbb{E}_x \text{Var}\left[\frac{1}{m}\sum_{j=1}^{m} a(w_j)\,\varphi(x; w_j)\right] \\ &\overset{(i)}{=} \frac{1}{m}\mathbb{E}_x\mathbb{E}_w(a(w)\varphi(x, w) - \mathbb{E}_w[a(w)\varphi(x, w)])^2 \\ &\overset{(ii)}{\leq} \frac{1}{m}\mathbb{E}_x\mathbb{E}_w[a^2(w)\varphi^2(x; w)] \\ &\overset{(iii)}{\leq} \frac{1}{m}\mathbb{E}_{w\sim\pi}[a^2(w)] \leq \frac{Q^2}{m}.\end{aligned}$$

where (i) uses the independence of $w_1, \cdots, w_m$, (ii) uses $\text{Var}[X] \leq \mathbb{E}(X - c)^2$, (iii) follows from Assumption 2.3. By Jensen's inequality,

$$\mathbb{E}_W\left[S_m(W)\right] \leq \sqrt{\mathbb{E}_W\left[S_m(W)^2\right]} \lesssim \frac{Q}{\sqrt{m}}$$

(4) Combining the above, we conclude

$$S_m(W) \lesssim \frac{Q}{\sqrt{m}} + \sqrt{\frac{\log(2/\delta)}{m}} Q,$$

as stated. □

3 Generalization Analysis

We express the *training error* and *generalization error* in terms of $\hat{\mathcal{R}}(a)$ and $\mathcal{R}(a)$ as follows:

$$\hat{\mathcal{R}}(a) = \frac{1}{n}\sum_{i=1}^{n}(f_m(x_i; a) - f^*(x))^2 \quad \mathcal{R}(a) = \mathbb{E}_{x\sim\rho}(f_m(x; a) - f^*(x))^2.$$

For the simplicity of analysis, we assume $\hat{a}$ is the solution of the following optimization problem

$$\hat{a} = \operatorname*{argmin}_{a \in \mathbb{R}^m} \left(\hat{\mathcal{R}}(a) + \frac{1}{\sqrt{nm}} \|a\| \right).$$

In contrast with (3), we regularise using the linear seminorm $\|a\|$ rather than the quadratic term $\|a\|^2$. For clarity of exposition we also fix the regularisation parameter to $\lambda = (nm)^{-1/2}$. These particular choices merely streamline the analysis; every result extends verbatim to the original setting of (3), with a squared-norm penalty and an arbitrary $\lambda > 0$.

Theorem 3.1. *Suppose Assumption 2.4 hold with $\|f^*\|_{\mathcal{F}_{\infty,\pi}} = Q \geq 1$. Then, for any $\delta_1, \delta_2 \in (0,1)$, with probability $1 - \delta_1 - \delta_2$, we have*

$$\mathcal{R}(\hat{a}) \lesssim \frac{Q}{m} \left(1 + \sqrt{\log(1/\delta_1)}\right) + \frac{Q^2}{\sqrt{n}} + \frac{Q^2\sqrt{n}}{m^2} \log(1/\delta_1) + \sqrt{\frac{\log(1/\delta_2)}{n}}$$

Proof. (1) By Theorem 2.5, for any $\delta_1 \in (0,1)$, with probability $1 - \delta_1$ over the sampling of random feature, there exists $\tilde{a} \in \mathbb{R}^m$ such that

$$\hat{\mathcal{R}}(\tilde{a}) \leq \frac{Q\left(1 + \sqrt{\log(1/\delta_1}\right))}{m}, \quad \frac{\|\tilde{a}\|}{\sqrt{m}} \leq Q.$$

(2) By the definition of $\hat{a}$, we have

$$\hat{\mathcal{R}}(\hat{a}) + \frac{1}{\sqrt{nm}}\|\hat{a}\| \leq \hat{\mathcal{R}}(\tilde{a}) + \frac{1}{\sqrt{nm}}\|\tilde{a}\| \leq \frac{Q}{m}\left(1 + \sqrt{\log(1/\delta_1)}\right) + \frac{Q}{\sqrt{n}}.$$

Hence, $\frac{1}{\sqrt{m}}\|\hat{a}\| \leq Q\left(1 + \frac{\sqrt{n}\left(1+\sqrt{\log(1/\delta_1)}\right)}{m}\right) =: C(m,n,Q)$.

(3) Let $\mathcal{H}_C = \left\{ f_m(\cdot;a) : \frac{\|a\|}{\sqrt{m}} \leq C \right\}$ and let $\mathcal{F}_C = \left\{ x \mapsto (f(x) - f^*(x))^2 \mid f \in \mathcal{H}_C \right\}$, then similar to the calculation of Rademacher complexity in the empirical process analysis, we obtain

$$\widehat{\mathrm{Rad}}_n(\mathcal{F}_C) \leq \frac{4C^2}{\sqrt{n}}$$

(4) By the Rademacher complexity-based generalization bound, for any $\delta_2 \in (0,1)$, with probability $1 - \delta_2$ over the sampling of training set, we have

$$\begin{aligned}
\mathcal{R}(\hat{a}) &\leq \hat{\mathcal{R}}(\hat{a}) + 2\widehat{\mathrm{Rad}}_n\left(\mathcal{F}_{C(m,n,Q)}\right) + \sqrt{\frac{\log(1/\delta_2)}{n}} \\
&\lesssim \frac{Q}{m}\left(1 + \sqrt{\log(1/\delta_1)}\right) + \frac{Q}{\sqrt{n}} + \frac{C^2(m,n,Q)}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta_2)}{n}}.
\end{aligned}$$

Inserting the expression of $C(m,n,Q)$, we completes the proof. □

Remark 3.2. *The $\log(1/\delta_1)$ term comes from the random sampling of features. The $\log(1/\delta_2)$ term comes from the random sampling of training set.*

According to Theorem 3.1, when $m > \sqrt{n}$, the RFM preserves the classical convergence rate $O(n^{-1/2})$. This shows that random feature approximation can preserve the statistical efficiency while reducing the computational cost from $O(n^3)$ to $O(n^2)$.

4 Summary

The foregoing analysis is based on [Rahimi and Recht, 2008], which provides a clear illustration of how random feature approximations can accelerate kernel methods without compromising statistical efficiency. More recent work leverages the integral-operator technique to sharpen these guarantees. In particular, [Bach, 2017] and [Carratino et al., 2018] employ source and capacity conditions to obtain excess-risk bounds that depend explicitly on the kernel's eigenvalue decay and the relative smoothness of the target function, thereby providing a fine-grained characterisation of how much acceleration random feature approximation can deliver in different spectral regimes.

References

[Bach, 2017] Bach, F. (2017). On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751.

[Carratino et al., 2018] Carratino, L., Rudi, A., and Rosasco, L. (2018). Learning with SGD and random features. In *Advances in Neural Information Processing Systems*, pages 10213–10224.

[Chen et al., 2023] Chen, H., Long, J., and Wu, L. (2023). A duality framework for analyzing random feature and two-layer neural networks. *arXiv preprint arXiv:2305.05642*.

[Rahimi and Recht, 2007] Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184.

[Rahimi and Recht, 2008] Rahimi, A. and Recht, B. (2008). Uniform approximation of functions with random bases. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pages 555–561. IEEE.