Theoretical Deep Learning

Lecture 5: RKHS II

July 29, 2021

Lecturer: Lei Wu *Scribe: Lei Wu*

1 A feature perspective

We know that KRR in the feature space corresponds to

$$\frac{1}{n}\sum_{i=1}^{n}(\langle\beta,\Phi(x_i)\rangle_{\mathcal{H}}-y_i\rangle^2+\lambda\|\beta\|_{\mathcal{H}}^2.$$

Naturally, we can consider the following target function space.

Definition 1.1. Given a kernel k, let $\Phi:\mathcal{X}\mapsto\mathcal{H}$ be a feature map of k. Let

$$\mathcal{F}=\{f(x;\beta)=\langle\beta,\Phi(x)\rangle_{\mathcal{H}}\ :\ \beta\in\mathcal{H}\}.$$

For $f\in\mathcal{F}$, define

$$\|f\|_{\mathcal{F}}=\inf_{f=\langle\beta,\Phi(\cdot)\rangle_{\mathcal{H}}}\|\beta\|_{\mathcal{H}}.$$

The infimum is taken such that the norm is independent of the specific representation β.

Lemma 1.2. *$\|\cdot\|_{\mathcal{F}}$ is indeed a well-defined norm.*

Proof. Assume $f_1=\langle\beta_1,\Phi(\cdot)\rangle_{\mathcal{H}}, f_2=\langle\beta_2,\Phi(\cdot)\rangle_{\mathcal{H}}$. Then,

$$\lambda_1f_1+\lambda_2f_2=\langle\lambda_1\beta_1+\lambda_2\beta_2,\Phi(\cdot)\rangle_{\mathcal{H}}.$$

By the definition,

$$\|\lambda_1f_1+\lambda_2f_2\|_{\mathcal{F}}\leq\|\lambda_1\beta_1+\lambda_2\beta_2\|_{\mathcal{H}}\leq|\lambda_1|\|\beta_1\|_{\mathcal{H}}+|\lambda_2|\|\beta_2\|_{\mathcal{H}}.$$

Taking infimum over β_1 and β_2 yields

$$\|\lambda_1f_1+\lambda_2f_2\|_{\mathcal{F}}\leq|\lambda_1|\|f_1\|_{\mathcal{F}}+|\lambda_2|\|f_2\|_{\mathcal{F}}.$$

In addition, let $\|f\|_{\mathcal{F}}=0$. By definition, for any $\varepsilon>0$, there exist β_ε such that $f=\langle\beta_\varepsilon,\Phi(\cdot)\rangle_{\mathcal{H}}$ and $\|\beta_\varepsilon\|_{\mathcal{H}}\leq\varepsilon$. Hence, for any $x\in\mathcal{X}$,

$$|f(x)|=|\langle\beta_\varepsilon,\Phi(x)\rangle_{\mathcal{H}}|\leq\|\beta_\varepsilon\|_{\mathcal{H}}\|\Phi(x)\|_{\mathcal{H}}\leq\varepsilon\|\Phi(x)\|_{\mathcal{H}}.$$

Taking $\varepsilon\to0$, we obtain $f(x)=0$ for any $x\in\mathcal{X}$. □

Presumably, the associated function space should only depend on the kernel k instead of the specific feature map. After all, the feature map may not be uniquely defined, in particular when k is rank degenerate. Does the above definition relies on the specific choice of Φ and $\mathcal{H}$?

Definition 1.3. Given a kernel k, let $\mathcal{F}$ be the function space defined in Definition 1.1. For any $f, g \in \mathcal{F}$, define

$$\langle f, g \rangle_{\mathcal{F}} = \frac{\|f+g\|_{\mathcal{F}}^2 - \|f-g\|_{\mathcal{F}}^2}{4}.$$

Lemma 1.4. *For any $f, g \in \mathcal{F}$, there exists $\beta_f, \beta_g \in \mathcal{H}$ such that $f = \langle \beta_f, \Phi(\cdot) \rangle, g = \langle \beta_g, \Phi(\cdot) \rangle$ and*

$$\langle f, g \rangle_{\mathcal{F}} = \langle \beta_f, \beta_g \rangle.$$

Proof. Taking β_f, β_g such that

$$\|f\|_{\mathcal{F}}^2 = \|\beta_f\|_{\mathcal{H}}^2$$
$$\|g\|_{\mathcal{F}}^2 = \|\beta_g\|_{\mathcal{H}}^2.$$

Hence,

$$\langle f, g \rangle_{\mathcal{F}} = \frac{\|\beta_f + \beta_g\|_{\mathcal{H}}^2 - \|\beta_f - \beta_g\|_{\mathcal{H}}^2}{4} = \langle \beta_f, \beta_g \rangle.$$

□

The above lemma shows that the inner product of the functions are equivalent to the inner product of the corresponding coefficients.

Lemma 1.5. *For any $x \in \mathcal{X}$, $\|k(\cdot, x)\|_{\mathcal{F}} = \|k(\cdot, x)\|_{\mathcal{H}}$.*

Proof. Notice that

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}.$$

For any β_x such that

$$k(x, x') = \langle \beta_x, \Phi(x') \rangle_{\mathcal{H}},$$

we have

$$\langle \beta_x - \Phi(x), \Phi(x') \rangle_{\mathcal{H}} = 0, \quad \forall x' \in \mathcal{X}.$$

This means that $\beta_x - \Phi(x) \perp \text{span}\{\Phi(x')\}$. Hence,

$$\|\beta_x\|_{\mathcal{H}}^2 = \|\beta_x - \Phi(x) + \Phi(x)\|_{\mathcal{H}}^2 = \|\beta_x - \Phi(x)\|_{\mathcal{H}}^2 + \|\Phi(x)\|_{\mathcal{H}}^2 \geq \|\Phi(x)\|_{\mathcal{H}}^2.$$

□

Theorem 1.6. $(\mathcal{F}, \langle \cdot \rangle_{\mathcal{F}}) = (\mathcal{H}_k, \langle \cdot, \cdot \rangle_{\mathcal{H}_k})$*, $\mathcal{H}_k$ is the RKHS constructed in Moore-Aronsajn theorem.*

Proof. By the uniqueness of RKHS, we only need to verify that $(\mathcal{F}, \langle \cdot \rangle_{\mathcal{F}})$ is a RKHS, for which k is the reproducing kernel. First, by Lemma 1.5, $k(\cdot, x) \in \mathcal{F}$ for any $x \in \mathcal{X}$. For any $f \in \mathcal{F}$, assume $f(x) = \langle \beta_f, \Phi(x) \rangle$ and $\|\beta_f\|_{\mathcal{H}}^2 = \|f\|_{\mathcal{F}}^2$. Then,

$$\langle f, k(\cdot, x) \rangle_{\mathcal{F}} = \langle \beta_f, \Phi(x) \rangle_{\mathcal{H}} = f(x).$$

Combining them, we show that $\mathcal{F}$ is a RKHS. □

The above perspective is useful in understanding the random feature models (RFMs) and two-layer neural networks. Consider the RFM:

$$f_m(x;\beta) = \frac{1}{m}\sum_{j=1}^m \beta_j \psi(x;w_j),$$

where $\psi : \mathcal{X} \times \Omega \mapsto \mathbb{R}$ and $\{w_j\}$ are independently drawn from a fixed distribution π. It can be viewed as the discretization of the continuous model:

$$f(x;\beta) = \mathbb{E}_{w\sim\pi}[\beta(w)\psi(x;w)],$$

where $\beta \in L^2(\pi)$ and $\psi(x;\cdot) \in L^2(\pi)$.

Consider the ridge regularization,

$$\min_{a\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n (f_m(x;\beta) - y_i)^2 + \frac{\lambda}{m}\sum_{j=1}^m \beta_j^2, \tag{1.1}$$

As $m \to \infty$, it corresponds to

$$\min_{a\in L^2(\pi)} \frac{1}{n}\sum_{i=1}^n (f(x;\beta) - y_i)^2 + \frac{\lambda}{m}\|\beta\|^2_{L^2(\pi)}.$$

1.1 A generalization analysis of the RFM

For the random feature model, define the associate kernel

$$k_\pi(x,x') = \int \psi(x;w)\psi(x';w)\,\mathrm{d}\pi(w). \tag{1.2}$$

Theorem 1.7. *Assume $f^* \in \mathcal{H}_{k_\pi}$. Let $W = (w_1,\dots,w_m)$ with $b_j \overset{iid}{\sim} \pi$, and $\beta(W) = (\beta(w_1),\dots,\beta(w_m))^T$. Then,*

$$\mathbb{E}_W \|f_m(\cdot;\beta(W)) - f^*\|^2_{L^2(\mathbb{P}_x)} \le \frac{\|f^*\|^2_{\mathcal{H}_{k_\pi}}}{m}.$$

Proof. Note that $f^*(x) = \mathbb{E}_{w\sim\pi}[\beta(w)\psi(x;w)]$. Then,

$$\begin{aligned}
&\mathbb{E}_W\,\mathbb{E}_x\,|\frac{1}{m}\sum_{j=1}^m \beta(w_j)\phi(x;w_j) - f^*(x)|^2 \\
&\qquad = \mathbb{E}_x\,\mathbb{E}_W\,\frac{1}{m^2}\sum_{i,j=1}^m \left(\beta(w_i)\psi(x;w_i) - f^*(x)\right)\left(\beta(w_j)\psi(x;w_j) - f^*(x)\right) \\
&\qquad = \mathbb{E}_x\,\frac{1}{m^2}\sum_{i=1}^m \mathbb{E}_{w_i}\left(\beta(w_i)\psi(x;w_i) - f^*(x)\right)^2 \qquad \text{(Use the independence)} \\
&\qquad \le \frac{1}{m}\mathbb{E}_{w\sim\pi}[\beta(w)^2] \qquad (\sup_{x\in\mathcal{X},w\in\Omega} |\psi(x;w)| \le 1).
\end{aligned}$$

□

The approximation error is a nonlinear function of $w_1, \ldots, w_m$. McDiarmid's inequality needs L^∞ boundedness.

Theorem 1.8. *Assume $f^*(x) = \mathbb{E}_{w\sim\pi}[\beta(w)\phi(x;w)]$ with $\operatorname{ess\,sup}_w |\beta(w)| \leq Q$. Let $W = (w_1, \ldots, w_m)$ with $w_j \overset{iid}{\sim} \pi$, and $\beta(W) = (\beta(w_1), \ldots, \beta(w_m))^T$. Then, for any $\delta \in (0,1)$, with probability $1-\delta$ over the sampling of W, we have*

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

Moreover, $\sup_j |\beta_j| \leq Q$.

Remark.

- In comparison to Theorem 1.7, the above theorem provide a high probability guarantee.
- The extra technical condition is that $\beta(\cdot) \in L^\infty(\pi)$. In contrast, for Theorem 1.7, we only need $\beta(\cdot) \in L^2(\pi)$.

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

$$\begin{aligned}|S_m(W) - S_m(\tilde{W})| &\leq \|f_m(\cdot;\beta(W)) - f_m(\cdot;\beta(\tilde{W}))\|_{L^2(\mathbb{P}_x)} \\ &= \|\frac{1}{m}\beta(w_i)\psi(x;w_i) - \frac{1}{m}\beta(\tilde{w}_i)\psi(x;\tilde{w}_i)\|_{L^2(\mathbb{P}_x)} \leq \frac{2Q}{m}.\end{aligned}$$

(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) Similar to the proof of Theorem 1.7, we have

$$\mathbb{E}_W[S_m(W)] \leq \sqrt{\mathbb{E}_W[S_m(W)^2]} = \sqrt{\mathbb{E}_W\|f_m(\cdot;\beta(W)) - f^*\|^2_{L^2(\mathbb{P}_x)}} \lesssim \frac{Q}{\sqrt{m}}.$$

(4) Combining them, we have

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

□

Proposition 1.9. *Let $\mathcal{H}_\pi^Q = \{f(x) = \mathbb{E}_{w\sim\pi}[\beta(w)\psi(x;w)] : \mathbb{E}_{w\sim\pi}[\beta(w)^2] \leq Q^2\}$. Then, we have*

$$\widehat{\operatorname{Rad}}_n(\mathcal{H}_\pi^Q) \leq \frac{Q}{\sqrt{n}}.$$

Proof. By Theorem 1.6, $\mathcal{H}_\pi^Q = \mathcal{H}_{k_\pi}^Q$. Hence by the bound of the Rademacher complexity of RKHS and $\sup_x k_\pi(x,x) \leq 1$, we complete the proof. □

Taking $\pi = \frac{1}{m}\sum_{j=1}^m \delta(\cdot - w_j)$ yields the bound of the Rademacher complexity for random feature models. Let $\mathcal{H}_Q = \{\frac{1}{m}\sum_{j=1}^m \beta_j \psi(x; w_j) : \frac{\|\beta\|_2}{\sqrt{m}} \leq Q\}$. Then,

$$\widehat{\mathrm{Rad}}_n(\mathcal{H}_Q) \leq \frac{Q}{\sqrt{n}}.$$

Consider

$$\hat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^m} \hat{\mathcal{R}}_n(\beta) + \frac{1}{\sqrt{n}} \frac{\|\beta\|}{\sqrt{m}}. \tag{1.3}$$

Here, the loss function $\ell(y, y') = (y - y')^2$ is not globally Lipschitz . Hence, for technical simplicity, we assume $\sup_{x \in \mathcal{X}} |f^*(x)| \leq 1$ and consider the truncated random feature model:

$$\tilde{f}_m(x; \beta) = \min(\max(f_m(x; \beta), -1), 1).$$

Theorem 1.10. *Suppose $f^*(x) = \mathbb{E}_{w \sim \pi}[\beta(w)\psi(x; w)]$ with $\operatorname{ess\,sup}_w |\beta(w)| \leq Q$. Then, for any $\delta_1, \delta_2 \in (0, 1)$, with probability $1 - \delta_1 - \delta_2$, we have*

$$\mathcal{R}(\hat{\beta}) \lesssim \frac{Q}{m}(1 + \sqrt{\log(1/\delta_1)}) + \frac{Q}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta_2)}{n}}.$$

- 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.

Proof. (1) By Theorem 1.8, for any $\delta_1 \in (0, 1)$, with probability $1 - \delta_1$ over the sampling of random features, there exist a $\tilde{\beta} \in \mathbb{R}^m$ such that

$$\hat{\mathcal{R}}_n(\tilde{\beta}) \leq \frac{Q(1 + \sqrt{\log(1/\delta_1)})}{m}, \qquad \frac{\|\tilde{\beta}\|}{\sqrt{m}} \leq Q.$$

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

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

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

(3) Let $\mathcal{H}_C = \{f_m(\cdot; \beta) : \frac{\|\beta\|}{\sqrt{m}} \leq C\}$ and Let $\mathcal{F}_C = \{\ell(f, f^*) : f \in \mathcal{H}_C\}$. Since $\ell(\cdot, y)$ is 2-Lipschitz continuous, the contraction lemma implies that

$$\widehat{\mathrm{Rad}}_n(\mathcal{F}_C) \leq 2\widehat{\mathrm{Rad}}_n(\mathcal{H}_C).$$

(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

$$\mathcal{R}(\hat{\beta}) \leq \hat{\mathcal{R}}_n(\hat{\beta}) + 2\widehat{\mathrm{Rad}}_n(\mathcal{F}_{C(m,n,Q)}) + \sqrt{\frac{\log(1/\delta_2)}{n}}$$

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

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

□

2 The spectral perspective

For a kernel $k : \mathcal{X} \times \mathcal{X} \mapsto R$, define an integral operator $T_k : L^2(\mathcal{X}; \mu) \mapsto L^2(\mathcal{X}; \mu)$,

$$T_k f(x) = \int k(x, x') f(x') \, \mathrm{d}\mu(x').$$

Here, μ is a probability measure on $\mathcal{X}$.

Theorem 2.1 (Mercer's theorem)**.** *Let k be a continuous kernel on a **compact** set $\mathcal{X}$. Then, $\forall x, x' \in \mathcal{X}$,*

$$k(x, x') = \sum_{j=1}^{\infty} \lambda_j e_j(x) e_j(x').$$

The convergence is uniform on $\mathcal{X} \times \mathcal{X}$, and absolute for each $(x, x') \in \mathcal{X} \times \mathcal{X}$.

$(\lambda_j)_{j\geq 1}$ and $(e_j)_{j\geq 1}$ are the eigenvalues and eigenfunctions of the integral operator T_k, respectively. Mercer's theorem gives a feature map for the kernel k. Let

$$\Phi : \mathcal{X} \mapsto \ell^2, \qquad \Phi(x) = \left(\sqrt{\lambda_1} e_1(x), \sqrt{\lambda_2} e_2(x), \ldots, \sqrt{\lambda_j} e_j(x), \ldots\right)^T.$$

Then,

$$k(x, x') = \sum_{j=1}^{\infty} \sqrt{\lambda_j} e_j(x) \sqrt{\lambda_j} e_j(x') = \langle \Phi(x), \Phi(x') \rangle_{\ell^2}.$$

Theorem 2.2. *Let k be a continuous kernel on a compact set $\mathcal{X}$, and $\{e_j\}$ be the orthonormal basis given in Mercer's theorem. Define*

$$\mathcal{H} = \big\{ f = \sum_j a_j e_j : \sum_j \frac{a_j^2}{\lambda_j} < \infty \big\},$$

with the inner product

$$\big\langle \sum_j a_j e_j, \sum_j b_j e_j \big\rangle_{\mathcal{H}} = \sum_j \frac{a_j b_j}{\lambda_j}.$$

Then, $\mathcal{H} = \mathcal{H}_k$.

Proof. By Mercer's theorem, $k(\cdot, x) = \sum_j (\lambda_j e_j(x)) e_j$. Hence,

$$\|k(\cdot, x)\|_{\mathcal{H}}^2 = \sum_j \frac{\lambda_j^2 e_j(x)^2}{\lambda_j} = \sum_j \lambda_j e_j(x) e_j(x) = k(x, x) < \infty.$$

So, $k(\cdot, x) \in \mathcal{H}$ for any $x \in X$. Reproducing property: Let $f = \sum_j a_j e_j \in \mathcal{H}$. Then,

$$\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = \sum_j \frac{a_j \lambda_j e_j(x)}{\lambda_j} = f(x). \tag{2.1}$$

So, $\mathcal{H}$ is a RKHS with the reproducing kernel k. By the uniqueness of RKHS, we conclude that $\mathcal{H} = \mathcal{H}_k$. □

Remark 2.3. Note that the integral operator T_k and the associated eigenfunctions $\{e_j\}$ depend on the underlying distribution μ. However, $\mathcal{H}$ coincide with the RKHS $\mathcal{H}_k$. This means that $\mathcal{H}$ actually does not depend on the choice of μ at all.

Weighted L^2 space. In this way, RKHS can be viewed as a L^2 space weighted by the eigenvalues. The faster is the eigenvalue decay, the smaller is the RKHS. Consider $\lambda_j = \frac{1}{j^s}$. Then,

$$\|f\|_{\mathcal{H}}^2 = \sum_{j=1}^{\infty} j^s a_j^2 < \infty \quad \Rightarrow \quad a_j^2 = O(\frac{1}{j^{s+1+\alpha}}) \quad \text{for some } \alpha > 0.$$

The previous Rademacher complexity analysis needs the following quantity to be finite:

$$\int k(x,x) \, \mathrm{d}\mu(x) = \int \sum_{j=1}^{\infty} \lambda_j e_j(x) e_j(x) \, \mathrm{d}\mu(x) = \sum_{j=1}^{\infty} \lambda_j \|e_j\|_{L^2(\mu)}^2 = \sum_{j=1}^{\infty} \lambda_j < \infty,$$

i.e., the trace of k is finite. The previous analysis cannot distinguish the RKHSs with different eigenvalue decays. A refined analysis need to reflect the fact that the faster is the eigenvalue decay, the smaller is the RKHS, i.e., the easier is the learning. For this type of sophisticated analysis, we refer to [Bach, 2017] and references therein.

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.