GRAPH LAPLACIAN AND LINEAR SMOOTHER REPORT 2.26

Beyond Empirical Risk Minimization: the lessons of deep learning
https://www.youtube.com/watch?v=JS-Bl36aVPs&feature=share

Get perfectly on the training set, so the empirical risk is zero, and the empirical risk minimization framework is problematic.

$$\mathbb{E}(L(f^*, y)) \leq \frac{1}{n}\sum L(f^*(x_i), y_i) + bound$$

with the first term to be 0 and the bound need to be exact. The classical uniform bounds do not work. However, thinking of 1-NN, interpolation (perfect fit on training points) is allowed, and the smoothing methods can be applied for such analysis.

weighted interpolated k-NN schemes with certain singular kernels are consistent (converge to Bayes optimal) for classification in any dimension. Interpolation is feasible with deep learning because of the phenomena of *double descent risk*: When the raining error hits zero and keep adding parameters for over-parametrization, the risk actually again descents.

Risk Consistency of Kernel Based Regression Methods
https://arxiv.org/pdf/0709.0626.pdf
http://web.eecs.umich.edu/~cscott/past_courses/eecs598w14/notes/17_svm_consistency.pdf

For a given loss function $L : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$

$$R_L(f) = \mathbb{E}_{XY}[L(Y, f(X))]$$
$$R_L^* = inf\{R_L(f)|f : \mathcal{X} \to \mathbb{R}\}$$
$$R_{L,\mathcal{F}}^* = inf\{R_L(f)|f \in \mathcal{F}\}$$

Definition. A loss L is called Lipschitz if for every $y \in \mathcal{Y}$, $L(y, .)$ is C-Lipschitz in $\mathbb{R}$ where C does not depend on y.

Definition. $\mathcal{X}$ compact metric space. A kernel k on $\mathcal{X}$ is universal if its corresponding RKHS $\mathcal{F}$ is dense in $\mathcal{C}(\mathcal{X})$ with respect to the supremum norm.

If k is universal kernel, then $R_{L,\mathcal{F}}^* = R_L^*$ for any Lipschitz loss L. However, the condition that $\mathcal{X}$ being compact is restrictive. Fortunately we do have the nice result for Gaussian kernel:

Theorem. *If k is a Gaussian kernel on $\mathcal{X} = \mathbb{R}^d$ and the loss L is Lipschitz, then $R_{L,\mathcal{F}}^* = R_L^*$*

Date: April 18, 2021.

Theorem. *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be n iid samples from $\mathcal{X} \times \mathcal{Y}$. Let k be a kernel on $\mathcal{X}$ with RKHS $\mathcal{F}$ such that $R^*_{L,\mathcal{F}} = R^*_L$. Let $L : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a Lipschitz loss for which $L_0 \equiv sup_{y \in \mathcal{Y}} L(y, 0) < \infty$. For kernel method find*

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(Y_I, f(X_i)) + \lambda ||f||^2_{\mathcal{F}}$$

Assume $sup_{x \in \mathcal{X}} \sqrt{k(x,x)} = B < \infty$. Let $\lambda = \lambda_n \to 0$ such that $n\lambda_n \to \infty$ as $n \to \infty$. Then

$$R_L(\hat{f}_n) - R^*_L \to_{a.s.} 0$$

for every distribution P_{XY} on $\mathcal{X} \times \mathcal{Y}$.

Theorem. *Let $X \subset \mathbb{R}^d$ be compact, let L be an invariant, convex loss of lower and upper order $p \geq 1$ and let H be a RKHS of a universal kernel on X. Define $p^* = \max\{2p, p^2\}$ and fix a sequence $(\lambda_n) \subset (0, \infty)$ with $\lambda_n \to 0$ and $\lambda_n^{p^*} n \to 0$. Then $\hat{f}_{n,\lambda_n}$ the minimizer of the empirical risk in RKHS H is L-risk consistent for all probability distribution P with $|P|_p < \infty$.*

For such L-risk consistency result, emphasize is on the behavior of the loss function L:

- invariant
- convex
- well-controlled growth

Then about the risk, required conditions are:

- finite risk: need $\int_{X \times Y} |f|^p(y) d|P|(x, y) < \infty$
- unique minimizer
- approximation error is controlled: universal error

Relating RKHS and Graph Laplacian through Integral Operators
Reference: *On Learning with Integral Operators*
`http://www.jmlr.org/papers/volume11/rosasco10a/rosasco10a.pdf`

Integral operators play roles both in graph Laplacian and RKHS methods: What's the intuition/big picture analogy of integral operators? Think of integral as the dual operator of differential operators (in graph Laplacian case)?

From Laplacian Eigenmap:
$\hat{L}_{t,n}$ as the **point cloud Laplacian**
For data points $x_1, \ldots, x_n \in \mathcal{M} \subseteq \mathbb{R}^N$:

$$\hat{L}_{t,n}(f)(p) \equiv \frac{1}{t(4\pi t)^{k/2}} \left(\frac{1}{n} \sum_{i=1}^{n} e^{-\frac{||p - x_i||^2}{4t}} \left[f(p) - f(x_i) \right] \right)$$

The (continuous) **integral operator**, which serves as a functional approximation to the LBO $\Delta_{\mathcal{M}}$ is defined as $L_{t,n} : L^2(\mathcal{M}) \to L^2(\mathcal{M})$

$$L_{t,n}(f)(p) \equiv \frac{1}{t(4\pi t)^{k/2}} \left(\int_{\mathcal{M}} e^{-\frac{||p - y||^2}{4t}} \left[f(p) - f(y) \right] d\mu_y \right)$$

where μ is the uniform measure on $\mathcal{M}$ obtained from the volume form. The uniform distribution is an essential assumption for this result in Belkin2008 and essentially assumes that there is no information with the density and hence can ignore the coupling of density and geometric structure.
In showing that the integral operator $L_{t,n}$ converges to $\Delta_{\mathcal{M}}$, the authors employs the **heat operator** $\mathcal{H}_t : L^2 \to L^2$, which is the convolution with the heat kernel, and hence also **a form of integral operator**. It has been shown that $\frac{1-\mathcal{H}}{t}$ and $\Delta_{\mathcal{M}}$ share an eigenbasis, and the proof for $EigL_t \to Eig\Delta_{\mathcal{M}}$ goes by showing the difference $R_t = \frac{1-\mathcal{H}}{t} - L_t$ is relatively bounded.

Hence we can see that integral operator can help understand graph Laplacians. To build on this, Rosasco et. al 2010 defined the following framework for graph Laplacians:

Let $W : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric continuous weight function such that $0 < c \leq W(x, s) \leq C, x, s \in \mathcal{X}$. Note here W is not required to be positive definite, but merely has positive entries. Let $\boldsymbol{W}$ be the gram matrix for data points in $\mathcal{X}$. Let $\boldsymbol{D}$ be the degree matrix, then the sample random walk graph Laplacian is defined as $\boldsymbol{L}_n : \mathbb{C}^n \to \mathbb{C}^n$

$$\boldsymbol{L}_n = \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{W}$$

Let P be a probability measure on $\mathcal{X}$, define the degree function as:

$$m(x) = \int_{\mathcal{X}} W(x, s)dP(s)$$

and the operator $\mathbb{L} : L2(\mathcal{X}, P) \to L2(\mathcal{X}, P)$

$$(\mathbb{L}f)(x) = f(x) - \int_{\mathcal{X}} \frac{W(x, s)}{m(x)} f(s)dP(s)$$

Integral operators are also present in kernel based methods:

RKHS
Let K be a symmetric PSD kernel such that $\sup_{x\in\mathcal{X}} K(x, x) = k_0 < \infty$.

$$\begin{aligned}
\mu_{\mathcal{X}} &= \mathbb{E}_{\mathcal{X}}[k(X, .)] = \int_{\mathcal{X}} k(X, .)dP(x) \\
\hat{\mu}_{\mathcal{X}} &= \frac{1}{n}\sum_{i=1}^{n}[k(X_i, .)] = \int_X k(X, .)d\hat{P}(x) \\
C_{XY} &= \mathbb{E}_{\mathcal{X},\mathcal{Y}}[(k(X, .) - \mu_{\mathcal{X}}) \otimes (h(Y, .) - \mu_{\mathcal{Y}})] \\
\hat{C}_{XY} &= \frac{1}{n-1}\sum_{i=1}^{n}[(k(X_i, .) - \hat{\mu}_{\mathcal{X}}) \otimes (h(Y_i, .) - \hat{\mu}_{\mathcal{Y}})]
\end{aligned}$$

For RKHS we have $\hat{\mu}_{\mathcal{X}} \to_p \mu_{\mathcal{X}}$ $\sum_{k=1}^{\infty} |\lambda_k(\hat{C}_X - C_X)| = O_p(n^{-1/2})$

Moreover, when we can define the integral operator corresponding to a centered Gram matrix as Let $\boldsymbol{K}$ be the gram matrix and define the corresponding integral operator $L_K : L^2(\mathcal{X}, P) \to L^2(\mathcal{X}, P)$.

$$(L_K f)(x) = \int_{\mathcal{X}} K(x, s)f(s)dP(s)$$

Then it has been shown that L_K is a Hilbert-Schmidt, positive self-adjoint operator and

$$\lambda(L_K) = \lambda(C_X)$$

and in particular L_K is a trace-class operator and $Tr(L_K) = \sum_{i\geq 1} \lambda_i(L_K) = \mathbb{E}||\phi(X)||^2_{\mathcal{H}}$
Moreover, define a linear (evaluation/sampling/restriction) operator $T : \mathcal{H} \to \mathcal{H}$,

$$(Tf)(x) = \langle f, \phi(x)\rangle_{\mathcal{H}} = f(x)$$

and let T^* be the continuous adjoint of T, then we can have

$$C_X = T^*T, L_K = TT^*$$

Therefore we see the integral operators help build the tools for graph Laplacian and for RKHS. For the estimation from finite sample, We want to assess

- to which extent can we use the gram matrix $\boldsymbol{K}$ to estimate L_K for RKHS methods
- to which extent can we use the graph Laplacian $\boldsymbol{L}_n$ to estimate the integral operator $\mathbb{L}$

A major challenge of these goal is that **L_k and $\boldsymbol{K}$, $\boldsymbol{L}_n$ and $\mathbb{L}$ operate on different spaces!**
$\boldsymbol{L}_n$ and $\boldsymbol{K}$ are finite dimensional matrices and sends $\mathbb{C}^n \to \mathbb{C}^n$ ($\mathbb{R}^n \to \mathbb{R}^n$), but $\mathbb{L}$ and L_k sends $L^2(\mathcal{X}, P) \to L^2(\mathcal{X}, P)$.
To overcome the difficulty, work on some intermediate spaces, RKHS $\mathcal{H}$:
Integral operators for RKHS:
Let $\mathcal{H}$ be the RKHS associated with the given kernel K, and define the operators $T_{\mathcal{H}}, T_n : \mathcal{H} \to \mathcal{H}$ given by

$$T_{\mathcal{H}} = \int_{\mathcal{X}} \langle \phi(x), .\rangle_{\mathcal{H}} \phi(x) dP(x)$$

$$T_n = \frac{1}{n}\sum_{i=1}^{n} \langle \phi(X_i), .\rangle_{\mathcal{H}} \phi(X_i)$$

where $\phi(x) = k(x, .)$ is the feature map associated with a point $x \in \mathcal{X}$. Similarly as written in the previous section, define the restriction/evaluation operator $R_n : \mathcal{H} \to \mathbb{C}^n$:

$$R_n f = (f(x_1), \ldots, f(x_n))$$

and the adjoint operator $R_n^* : \mathbb{C}^n \to \mathcal{H}$ that for $(y_1, \ldots, y_n) \in \mathbb{C}^n$,

$$R_n^*(y_1, \ldots, y_n) = \frac{1}{n}\sum_{i=1}^{n} y_i \phi(X_i)$$

Hence $T_n = R_n^* R_n$ and $\boldsymbol{K} = R_n R_n^*$. Similarly let $R_{\mathcal{H}}$ be the inclusion $\mathcal{H} \to L^2(\mathcal{X}, P)$ then $T_{\mathcal{H}} = R_{\mathcal{H}}^* R_{\mathcal{H}}$ and $L_K = R_{\mathcal{H}} R_{\mathcal{H}}^*$.

$T_{\mathcal{H}}$ and L_K have the same spectra. possibly up to the zero, and have eigenvectors related by a simple equation (Proposition 8 in Rosasco et. al 2010).

T_n and $\boldsymbol{K}$ share eigenvalues up to some zeros and have eigenvectors related by simple equations (Proposition 9).

To characterize the connection between $\boldsymbol{K}$ and L_K, it suffices to connect T_n to $T_{\mathcal{H}}$. For this purpose, the convergence in Hilber-Schmidt operator norm of T_n to $T_{\mathcal{H}}$ is presented in Theorem 7.

Combining the pieces, the l_2 distance between the spectrum of L_K and $\boldsymbol{K}$ is bounded in Proposition 10 and Proposition 11. And the bound on spectral projection is shown in Theorem 12.

In sum, by studying $\boldsymbol{K}$ and L_K all in the common RKHS $\mathcal{H}$, the spectral convergence is clearly shown through the framework of integral operators. This framework can be easily adopted in this framework since the RKHS structure is already given (with which we defined $\boldsymbol{K}$ and L_K), and the given RKHS fits nicely with the framework of integral operators. The requirements for the framework to work are

- The kernel k is a reproducing kernel (symmetric, PSD)
- k continuous, and $\sup_{x\in\mathcal{X}} k(x,x) = k_{max} < \infty$
- $\mathcal{X}$ is locally compact separable metric space

To want to apply the integral operator framework to graph Laplacians. However, for graph Laplacians, the provided "kernel" $\frac{W(x,s)}{m(x)}$ is not naturally a reproducing kernel (not required to be positive semi-definite in the framework presented above), and hence more work is required to construct an auxiliary RKHS to accommodate the operators we interested in.

Integral Operators for Graph Laplacian
The definitions are restated here for ease of accessing:

Let $W : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric continuous weight function such that $0 < c \leq W(x,s) \leq C, x, s \in \mathcal{X}$. Note here W is not required to be positive definite, but merely has positive entries. Let $\boldsymbol{W}$ be the gram matrix for data points in $\mathcal{X}$. Let $\boldsymbol{D}$ be the degree matrix, then the sample random walk graph Laplacian is defined as $\boldsymbol{L} : \mathbb{C}^n \to \mathbb{C}^n$

$$\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{W}$$

Let P be a probability measure on $\mathcal{X}$, define the degree function as:

$$m(x) = \int_{\mathcal{X}} W(x,s)dP(s)$$

and the operator $\mathbb{L} : L2(\mathcal{X}, P) \to L2(\mathcal{X}, P)$

$$(\mathbb{L}f)(x) = f(x) - \int_{\mathcal{X}} \frac{W(x,s)}{m(x)} f(s)dP(s)$$

To make the quantities clear, define the following functions:

$$W_x : \mathcal{X} \to \mathbb{R} \quad W_x(t) = W(x,t)$$
$$m_n : \mathcal{X} \to \mathbb{R} \quad m_n = \frac{1}{n}\sum_{i=1}^{n} W_{x_i}$$

Recall the integral operators for RKHS are defined with in a RKHS, so for now assume there exists a RKHS $\mathcal{H}$ with bounded continuous kernel k such that

$$W_x, \frac{W_x}{m}, \frac{W_x}{m_n} \in \mathcal{H}$$
$$||\frac{W_x}{m}||_{\mathcal{H}} \leq C \quad \forall x \in \mathcal{X}$$

Under such assumptions, we can define the bounded integral operators $\mathbb{L}_{\mathcal{H}}, A_{\mathcal{H}} : \mathcal{H} \to \mathcal{H}$

$$A_{\mathcal{H}} = \int_{\mathcal{X}} \langle \phi(x_i), . \rangle_{\mathcal{H}} \frac{W_x}{m} dP(x)$$
$$\mathbb{L}_{\mathcal{H}} = I - A_{\mathcal{H}}$$
$$A_n = \frac{1}{n} \sum_{i=1}^{n} \langle \phi(x_i), . \rangle_{\mathcal{H}} \frac{W_{x_i}}{m_n} \mathbb{L}_n = I - A_n$$

In similar fashion to the RKHS case we can define restriction/evaluation operator $R_n : \mathcal{H} \to \mathbb{C}^n$

$$R_n f = (f(x_1), \ldots, f(x_n))$$

Different from the RKHS case, since an auxiliary $\mathcal{H}$ is used with the reproducing kernel not the "natural kernel counterpart" $\frac{W_x}{m}$, we need to define the adjoint (extension) operator $E_n : \mathbb{C}^n \to \mathcal{H}$ a little differently as:

$$E_n(y_1, \ldots, y_n) = \frac{1}{n} \sum_{i=1}^{n} y_i \frac{W_{x_i}}{m_n}, (y_1, \ldots, y_n) \in \mathbb{C}^n$$

Then we have

$$A_n = E_n R_n, \boldsymbol{D}^{-1}\boldsymbol{W} = R_n E_n$$

Similarly define the infinite sample restriction and extension operator $R_{\mathcal{H}}$ and $E_{\mathcal{H}}$ and can have $A_{\mathcal{H}} = E_{\mathcal{H}} R_{\mathcal{H}}$, $I - \mathbb{L} = R_{\mathcal{H}} E_{\mathcal{H}}$.

$A_{\mathcal{H}}, \mathbb{L}_{\mathcal{H}}$, and $\mathbb{L}$ have simply related eigenvalues, up to zeros, and have eigenfunctions related by simple equations (proposition 13).

$A_n, \mathbb{L}_n$, and $\boldsymbol{L}$ have simply related eigenvalues, up to zeros and have eigenfunctions related by simple equations (Proposition 14).

It then remains to bound $A_{\mathcal{H}} - A_n$. However, for such convergence, the structure of the RKHS we assumed to exist earlier need to be specified.

If the weight function W is sufficiently differentiable, can choose the auxiliary RKHS to be a suitable Sobolev Space. For the sake of simplicity, $\mathcal{X}$ is assumed to be a bounded open subset of $\mathbb{R}^d$ with a nice boundary (precise assumptions use quasi-resolved boundary open set).

For given index $s \in \mathbb{N}$, the Sobolev space $\mathcal{H}^s$ is

$$\mathcal{H}^s \equiv \{f \in L^2(\mathcal{X}, dx) | D^{\boldsymbol{\alpha}} f \in L^2(\mathcal{X}, dx), \forall |\boldsymbol{\alpha}| = s\}$$

where dx stands for the Lebesgue measure, and $\boldsymbol{\alpha} \in \mathbb{N}^d$ is multi-index. $\mathcal{H}^s$ is a separable Hilbert space with respect to the scalar product

$$\langle f, g \rangle_{\mathcal{H}^s} = \langle f, g \rangle_{L^2(\mathcal{X}, dx)} + \sum_{|\boldsymbol{\alpha}| = s} \langle D^{\boldsymbol{\alpha}} f, D^{\boldsymbol{\alpha}} g \rangle_{L^2(\mathcal{X}, dx)}$$

Let $C_b^m(\mathcal{X})$ be the set of continuous bounded functions such that all the standard derivatives of order m exists and are continuous bounded functions, and this is a Banach space with respect to the norm

$$||f||_{C_b^m} = sup_{x\in\mathcal{X}}|f(x)| + \sum_{|\boldsymbol{\alpha}|=m} sup_{x\in\mathcal{X}}|(D^{\boldsymbol{\alpha}}f)(x)|$$

Recall that we assume $\mathcal{X}$ is a bounded open subset of $\mathbb{R}^d$ with a nice boundary, and for such dimension d we let the order for Sobolev space be $s = \lfloor d/2 \rfloor + 1$, and the order for C_b^m be $m = 0$ so we are essentially looking at bounded functions. Then we have

$$\mathcal{H}^s \subset C_b^m(\mathcal{X}), ||f||_{C_b^m} \leq C_{s,m}||f||_{\mathcal{H}^s}$$

This Sobolev space $\mathcal{H}^s$ with $s = \lfloor d/2 \rfloor + 1$ is a RKHS with a continuous real valued bounded kernel K^s. Then modify the conditions on weight function W:

$$W(x,t) \geq c > 0 \quad \forall x,t \in \mathcal{X} \tag{0.1}$$

$$W \in \mathcal{H}^{d+1}(\mathcal{X}\times\mathcal{X}) \tag{0.2}$$

With the defined Sobolev space as RKHS and weight function satisfying above, the previous assumptions

$$W_x, \frac{W_x}{m}, \frac{W_x}{m_n} \in \mathcal{H}$$

$$||\frac{W_x}{m}||_{\mathcal{H}} \leq C \quad \forall x \in \mathcal{X}$$

are satisfied. (The choice of $d+1$ here help control $m - m_n$ in the Hilbert space in order to use Hoeffding's inequality, explained in Remark 19)

$$A_{\mathcal{H}}, \mathbb{L}_{\mathcal{H}}, A_n, \mathbb{L}_n \in \mathcal{H}^s, \quad s = \lfloor d/2 \rfloor + 1$$

With such constructed RKHS structure, we can get

$$||\mathbb{L}_n - \mathbb{L}_{\mathcal{H}}||_{HS} = ||A_n - A_{\mathcal{H}}||_{HS} \leq C\sqrt{\frac{t}{n}}$$

with probability $1 - 2e^{-t}$ in the Hilbert-Schimdt norm of operator in the Sobolev space of $\mathcal{H}^s$ (Theorem 15).

The connections between the eigenvalues and eigenfunctions of $\mathbb{L}$ and $\boldsymbol{L}$ is given in Proposition 21, and the relations between spectral projections in Proposition 22. For these spectral results, since the operators are no longer self-adjoint, some more spectral theory results are utilized for the proof.

In sum, when adapting the integral operator framework to graph Laplacians, the convergence results are only shown under some restrictive conditions on the weight function W and conditions on the input space $\mathcal{X}$. (0.1) essentially rule out k-NN and ε-neighborhood construction. Can try to generalize such framework.

Also can try apply such framework to some linear smoothers.

More essentially, the composition of restriction/evaluation operator with the extension operator can have some intuition how those operators really do the work.