Lecture 9: Kernel ridge regression

Announcements

- HW 2 out tonight, due 9.11. 23:59
- Proofs skipped in class / exercise for home: You are supposed to fully understand those steps, also of the exercises in class and in the homework - the oral exam will primarily test your understanding of how different proof steps fit together

Plan for today

- Another example of prediction error of square-loss minimizer: Prediction error bound for ERM of norm-bounded RKHS
- Prediction error bound for *regularized* regression

Recap: Non-parametric prediction error bound

Definition (Localized (empirical) Gaussian complexity)

The localized Gaussian complexity around $f^\star$ of scale δ is

$$\widetilde{\mathcal{G}}_n(\mathcal{F}^\star;\delta_n) := \widetilde{\mathcal{G}}_n(\mathcal{F}^\star(x_1^n)\cap\mathbb{B}_n(\delta_n)) = \mathbb{E}\sup_{\|\hat{\Delta}\|_n\le\delta_n,\hat{\Delta}\in\mathcal{F}^\star}\frac{1}{n}\sum_{i=1}^n w_i\hat{\Delta}(x_i)$$

Lemma (Critical radius, MW 13.6.)

For any star-shaped $\mathcal{F}$, it holds that $\frac{\widetilde{\mathcal{G}}_n(\mathcal{F};\delta)}{\delta}$ is non-increasing and the critical inequality

$$\frac{\widetilde{\mathcal{G}}_n(\mathcal{F};\delta)}{\delta}\le\frac{\delta}{\sigma}$$

has a smallest solution $\delta_n>0$ that we call the critical quantity/radius.

Theorem (Prediction error bound, MW Thm 13.5.)

If $\mathcal{F}^\star$ is star-shaped, we have for the square loss minimizer $\widehat{f}$ for any $t\ge 1$

$$\mathbb{P}(\|\widehat{f}-f^\star\|_n^2\ge 16t\delta_n^2)\le e^{-\frac{nt\delta_n^2}{2\sigma^2}}$$

Recap: Reproducing Kernel Hilbert Spaces (RKHS)

- Recap motivation of kernel trick and kernel spaces
- abstract definition of reproducing kernel Hilbert spaces $\to$ can be associated uniquely with a kernel $\mathcal{K}$ and equal to its induced (unique) Hilbert space which is the completion of
- $\mathcal{F}_{\text{pre}}=\{\sum_{i=1}^N\alpha_i\mathcal{K}(\cdot,x_i):N\in\mathbb{N},\alpha\in\mathbb{R}^N,x_1,\dots,x_N\in\mathcal{X}\}$ with inner product $\langle\mathcal{K}(\cdot,x),\mathcal{K}(\cdot,y)\rangle_{\mathcal{F}_\mathcal{K}}=\mathcal{K}(x,y)$

Theorem (Existence of kernel, MW Thm 12.13)

Given an RKHS $\mathcal{F}$, there is a unique psd kernel $\mathcal{K}_\mathcal{F}$ that satisfies the reproducing property

- $\mathcal{F}_{lin}=\{f:f(x)=\langle w,x\rangle,w\in\mathbb{R}^d\}$ is an RKHS with $\mathcal{K}(x,y)=\langle x,z\rangle$ as a reproducing kernel as a reproducing kernel $f=\langle w_f,\cdot,\rangle$ and $g=\langle w_g,\cdot\rangle$ the inner product $\langle f,g\rangle=w_f^\top w_g$

From function class (RKHS) to kernel: Sobolev spaces

$\mathcal{L}^2([0,1])$ is not an RKHS because convergence not point-wise

Some restrictions on $\mathcal{L}^2([0,1])$ can fix that: Sobolev space on $[0,1]$ $\mathcal{W}_2^1([0,1]) = \{f : [0,1] \to \mathbb{R} \mid f(0) = 0, f' \in \mathcal{L}^2([0,1])\}$ where derivative exists almost everywhere

- IP $\langle f, g \rangle = \int_0^1 f'(x) g'(z) \mathrm{d}z$ (interpretable)
- Sobolev kernel: $\mathcal{K}(x, y) = \min\{x, y\}$
- Reproducing prop.: $\langle f(\cdot), \min\{\cdot, z\} \rangle = \int_0^1 f'(x) \mathbb{1}_{x \leq z} \mathrm{d}x = \int_0^z f'(x) \mathrm{d}x = f(z)$
- can extend to higher order derivatives / smoothness (HW 2) $\mathcal{W}_2^\alpha([0,1]) = \{f : [0,1] \to \mathbb{R} \mid f^{(\alpha)}(0) = 0, f^{(\alpha)} \in \mathcal{L}^2([0,1])\}$

Non-parametric regression in RKHS

Setting: $f^\star \in \mathcal{F}_\mathcal{K}$ for some kernel $\mathcal{K}$ and $y_i = f^\star(x_i) + \sigma w_i$ w/ i.i.d. $w_i \sim \mathcal{N}(0,1)$

- Recall the non-parametric (unpenalized) estimate $\widehat{f}$ is defined as

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 \text{ (possibly non-unique)}$$

Today:

- compute generalization bound for $\widehat{f}$ in a particular RKHS
- Minimization of square loss in constrained space $\mathcal{F}_R = \{f \in \mathcal{F} : \|f\|_\mathcal{F} \leq R\}$ (ommitting subscript $\mathcal{K}$) or kernel ridge regression (regularized square loss) using localized complexities

Unregularized kernel regression

- Given empirical loss $\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$ and (empirical) prediction error $\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$.
- Define the *empirical kernel matrix* K with $K_{ij} := \frac{\mathcal{K}(x_i,x_j)}{n}$ (*this is the normalized kernel matrix, more interpretable since eigenvalues converge to operator eigenvalues*)
- Now assume that the empirical kernel matrix is invertible.

Neighbor-Q:
a) What is the minimum value of the empirical loss?
b) How about the prediction error?
c) How about the localized Gaussian complexity?
d) For which kernels is the kernel matrix invertible?

Remember how to rewrite the empirical loss in matrix vector notation.
Compute the localized complexity and critical radius

Regularized kernel regression

If $\mathcal{K}$ is s.t. K is pd/full-rank for all distinct inputs $\rightarrow$ can interpolate!
In that case the localized Gaussian complexity will be of order 1.

$\mathcal{F}$ too large! $\rightarrow$ require bounded norm $\mathcal{F}_R = \{f \in \mathcal{F} : \|f\|_{\mathcal{F}} \leq R\}$

So we defined the regularized estimator $\widehat{f}_R$ is defined as

$$\widehat{f}_R \in \arg\min_{f \in \mathcal{F}_R} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \text{ (possibly non-unique)}$$

By the representer theorem we can then write it as

$$\min_{f \in \mathcal{F}_R} \frac{1}{2n}\|y - f(x_1^n)\|_2^2 = \min_{\alpha \in \mathbb{R}^n} \frac{1}{2n}\|y - K\alpha\|_2^2$$

- We now see eigenvalues of the kernel matrix can be used to bound prediction error of $\widehat{f}_R$ w.h.p. via the critical inequality!

Localized G.C. for RKHS with bounded norm

Lemma (local G.C. for norm-bounded RKHS, MW Cor. 13.18)

Defining $\hat{\mu}_j$ as eigenvalues of the kernel matrix K we have

$$\widetilde{\mathcal{G}}_n(\mathcal{F}_1;\delta) \leq \sqrt{\frac{2}{n}}\sqrt{\sum_{j=1}^{n} \min\{\delta^2, \hat{\mu}_j\}}.$$

In fact, more generally $\widetilde{\mathcal{G}}_n(\mathcal{F}_r;\delta) \leq \sqrt{\frac{r^2+1}{n}}\sqrt{\sum_{j=1}^{n} \min\{\delta^2, \hat{\mu}_j\}}$.

Definition (R-modified critical quantity $\delta_{n;R}$)

We define $\delta_{n;R}$ to be the smallest $\delta > 0$ satisfying

$$\frac{4}{\sqrt{n}}\sqrt{\sum_{j=1}^{n} \min\{\delta^2, \hat{\mu}_j\}} \leq \frac{\delta^2 R}{\sigma}$$

- By Lemma it then holds that $\frac{\sigma\widetilde{\mathcal{G}}_n(\mathcal{F}_3;\delta_{n;R})}{\delta_{n;R}} \leq \delta_{n;R} R$

Prediction error bound for RKHS with bounded norm

Theorem (Prediction error of norm-bounded RKHS)

Assume $f^\star \in \mathcal{F}_R$. Then we have for least-squares estimate $\widehat{f}_R \in \mathcal{F}_R$

$$\|\widehat{f}_R - f^\star\|_n^2 \leq c_0 R^2 \delta_{n;R}^2$$

with probability $\geq 1 - c_1 e^{-c' \frac{nR^2\delta_{n;R}^2}{\sigma^2}}$.

Note: Can easily generalize to $f^\star \notin \mathcal{F}_R$ (more technical, without new core insights) with additional approx. error $\inf_{\|f\|_{\mathcal{F}} \leq R} \|f - f^\star\|_n^2$

Rates for actual kernel spaces $\mathcal{F}$

- Ex. 1: α-smooth functions w/ $\hat{\mu}_j \sim j^{-2\alpha} \to \|\widehat{f} - f^\star\|_n^2 \leq (\frac{R\sigma^2}{n})^{2/3}$
- Ex. 2: Gaussian kernel w/ $\hat{\mu}_j \sim e^{-cj\log j} \to \|\widehat{f} - f^\star\|_n^2 \leq \frac{\sigma^2 \log(\frac{Rn}{\sigma})}{n}$
- For $\mathcal{K}$ on compact $\mathcal{X}$ empirical matrix eigenvalues $\hat{\mu}_j \sim \mu_j$ for big n where μ_j are integral operator eigenvalues (Koltchinskii, Gine '00)

Proof for Theorem (prediction error of $\widehat{f} \in \mathcal{F}_R$)

- Scale basic inequality by R to obtain $\widetilde{f^\star} = \frac{f^\star}{R}$, $\widetilde{f} = \frac{\widehat{f}}{R}$, $\tilde{\sigma} = \frac{\sigma}{R}$

$$\frac{1}{nR^2}\sum_{i=1}^n (y_i - \widehat{f}(x_i))^2 \le \frac{1}{nR^2}\sum_{i=1}^n (y_i - f^\star(x_i))^2$$

$$\|\widetilde{f} - \widetilde{f^\star}\|_n^2 \le 2\frac{\tilde{\sigma}}{n}\sum_{i=1}^n w_i(\widetilde{f}(x_i) - \widetilde{f^\star}(x_i))$$

- Since $\widetilde{f^\star}, \widetilde{f} \in \mathcal{F}_1$, $\tilde{\Delta} \in \mathcal{F}_1^\star = \mathcal{F}_1 - \widetilde{f^\star} \subset \mathcal{F}_3$ ($\mathcal{F}_2$ suffices for norm-bounded RKHS, but use $\mathcal{F}_3$ for penalized later) . . .
- Now argue similar to last lecture
  - Want $\frac{\tilde{\sigma}}{n}\sum_i w_i \tilde{\Delta}(x_i) \le 2\|\tilde{\Delta}\|_n \delta_{n;R}$ for all $\|\tilde{\Delta}\|_n \ge \delta_{n;R}$ for some $\delta_{n;R}$
  - Using $\mathbb{E}_w \sup_{\tilde{\Delta}\in\mathcal{F}_3, \|\tilde{\Delta}\|_n \le \delta} \frac{\tilde{\sigma}}{n}\sum_{i=1}^n w_i \tilde{\Delta}(x_i) = \tilde{\sigma}\widetilde{\mathcal{G}}_n(\mathcal{F}_3; \delta)$
  - It's sufficient that $\sup\limits_{\|\tilde{\Delta}\|_n \le \delta_{n;R}, \tilde{\Delta}\in\mathcal{F}_3} \frac{\tilde{\sigma}}{n}\sum_{i=1}^n w_i \frac{\tilde{\Delta}(x_i)}{\delta_{n;R}} \le \delta_{n;R}$ where we ned modified critical inequality $\tilde{\sigma}\widetilde{\mathcal{G}}_n(\mathcal{F}_3; \delta_{n;R}) \le \delta_{n;R}^2$ in tail bound
- Observing $\|\widehat{f} - f^\star\|_n^2 = R^2\|\tilde{\Delta}\|_n^2$ yields the theorem. □

Proof of Lemma (local. compl. for norm-bounded RKHS)

- By representer theorem, can take sup over $\mathcal{F}_S$ by parameterizing $\Delta(\cdot) = \frac{1}{\sqrt{n}}\sum_i \alpha_i \mathcal{K}(\cdot, x_i) \in \mathcal{F}_S \subset \mathcal{F}$ and hence $\Delta(x_1^n) = \sqrt{n}K\alpha$, s.t.

$$\widetilde{\mathcal{G}}_n(\mathcal{F}_r; \delta) = \mathbb{E}_w \sup_{\|\Delta\|_{\mathcal{F}} \le r, \|\Delta\|_n \le \delta} \frac{1}{n}\sum_i w_i \Delta(x_i)$$

$$= \frac{1}{\sqrt{n}}\mathbb{E}_w \sup_{\alpha^\top K\alpha \le r^2, \alpha^\top K^2\alpha \le \delta^2} w^\top K\alpha$$

- Let $K = U^\top \Lambda U$ and $\theta := \Lambda U\alpha \rightarrow \widetilde{\mathcal{G}}_n(\mathcal{F}_r; \delta) = \frac{1}{\sqrt{n}}\mathbb{E}_w \max_{\theta\in\mathbb{T}} w^\top\theta$

$$\text{with } \mathbb{T} = \{\theta \in \mathbb{R}^n \mid \sum_i \theta_i^2 \le \delta^2, \sum_{i=1}^n \frac{\theta_i^2}{\hat{\mu}_i} \le r^2\}$$

- Let $\mathcal{E} := \{\theta \in \mathbb{R}^n \mid \sum_i \eta_i\theta_i^2 \le 1 + r^2\} \supset \mathbb{T}$ w/ $\eta_i = \max\{\delta^{-2}, \hat{\mu}_i^{-1}\}$

$$\max_{\theta\in\mathcal{E}}\langle w, \theta\rangle \iff \max_{\theta^\top \mathrm{diag}(\eta_i)\theta \le 1+r^2}\langle w, \theta\rangle \iff \max_{\|\beta\|_2 \le \sqrt{1+r^2}}\langle \mathrm{diag}^{-1/2}(\eta_i)w, \beta\rangle$$

- Hence $\widetilde{\mathcal{G}}_n(\mathcal{F}_r; \delta) \le \sqrt{\frac{1+r^2}{n}}\mathbb{E}_w\sqrt{\sum_i \frac{w_i^2}{\eta_i}} \le \sqrt{\frac{1+r^2}{n}}\sqrt{\sum_i \frac{1}{\eta_i}}$ via

Regularized regression guarantees for metric spaces

- So far looked at empirical risk minimizers for the square loss of type $\widehat{f} \in \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2$
- But often type we minimize a loss with an additive penalty such as in ridge regression

$$\widehat{f}_{\lambda_n} = \arg\min_{f\in\mathcal{F}} \frac{1}{2n}\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda_n \|f\|_{\mathcal{F}}^2$$

- With the same definition of $\delta_{n;R}$ as before

Theorem (Prediction error for reg. estimators - MW Thm 13.17.)

For any convex function class $\mathcal{F}$ with a norm and $\mathcal{F}^\star$ star-shaped, when $\lambda_n \geq 2\delta_{n;R}^2$, there is a universal constant such that for $f^\star \in \mathcal{F}_R$

$$\|\widehat{f}_{\lambda_n} - f^\star\|_n^2 \leq cR^2(\delta_{n;R}^2 + \lambda_n) \text{ w/ prob. } \geq 1 - c_0 e^{-c_1 \frac{nR^2\delta_{n;R}^2}{\sigma^2}}.$$

- Again, if $f^\star \notin \mathcal{F}_R$ yields add. approx. error $\inf_{\|f\|_{\mathcal{F}}\leq R} \|f - f^\star\|_n^2$
- if additional term $\lambda_n \sim \delta_{n;R}^2$, same order as constrained

Proof of bound for regularized regression estimate

For simplicity we write $\widehat{f}$ for $\widehat{f}_{\lambda_n}$

1. By optimality we have

$$\frac{1}{2n}\sum_{i=1}^n (f^\star(x_i) + \sigma w_i - \widehat{f}(x_i))^2 + \lambda_n\|\widehat{f}\|_{\mathcal{F}}^2 \leq \frac{\sigma^2}{2n}\sum_{i=1}^n w_i^2 + \lambda_n \|f^\star\|_{\mathcal{F}}^2$$

which yields basic inequality after rearranging terms

$$\frac{1}{2}\|\Delta\|_n^2 \leq \frac{\sigma}{n}\sum_{i=1}^n w_i\Delta(x_i) + \lambda_n(\|f^\star\|_{\mathcal{F}}^2 - \|\widehat{f}\|_{\mathcal{F}}^2)$$

2. Normalize $f^\star, \widehat{f}, \sigma$ by $\frac{1}{R}$ like for norm-bounded $\rightarrow$ $\widetilde{f^\star}, \widetilde{f}, \tilde{\sigma}, \tilde{\Delta} = \widetilde{f} - \widetilde{f^\star}$ ($\widetilde{f}$ different than in MW!)

$$\frac{1}{2}\|\tilde{\Delta}\|_n^2 \leq \underbrace{\frac{\tilde{\sigma}}{n}\sum_{i=1}^n w_i\tilde{\Delta}(x_i)}_{T_1} + \underbrace{\lambda_n(\|\widetilde{f^\star}\|_{\mathcal{F}}^2 - \|\widetilde{f}\|_{\mathcal{F}}^2)}_{T_2}$$

Note that T_2 is a new term **and** $\tilde{\Delta}, \widetilde{f}$ are not necessarily $\mathcal{F}$-norm-bounded which enters in localized G.C. for $\mathcal{F}_R$ to bound T_1

Proof of bound for regularized regression estimate

3. Either $\|\tilde{\Delta}\|_{n;R} \leq \delta_n$ and we are done, or $\|\tilde{\Delta}\|_n > \delta_{n;R}$ on which event we further analyze two events based on the $\mathcal{F}$-norm of $\tilde{\Delta}$ and show that in both events it holds that

$$c'\|\tilde{\Delta}\|_n^2 \leq c\delta_{n;R}\|\tilde{\Delta}\|_n + \lambda_n$$

for different constants c', c(details in next slide)

a) on Event 1 $\|\widetilde{f}\|_{\mathcal{F}} \leq 2$ using previous arguments on T_1 as for the prediction error for norm-bounded RKHS using the critical inequality and tail bound, as well as the fact that $T_2 \leq \|\widetilde{f^\star}\|_{\mathcal{F}}^2 \leq 1$.

b) on Event 2 $\|\widetilde{f}\|_{\mathcal{F}} > 2$ using a new (peeling) lemma for all $\|\tilde{\Delta}\|_{\mathcal{F}} \geq 1$. There we use T_2 to “cancel” large norms

4. Solving the quadratic yields $\|\tilde{\Delta}\|_n^2 \leq c(\delta_{n;R}^2 + \lambda_n)$ $\square$

Proof of 4. - regularization plays role of norm-bounding

We use the shorthand δ_n for $\delta_{n;R}$. We now show that on both events 1 & 2, $c'\|\tilde{\Delta}\|_n^2 \leq c\delta_n\|\tilde{\Delta}\|_n + \lambda_n$ for some (different) constants c', c

a) Event 1: $\|\widetilde{f}\|_{\mathcal{F}} \leq 2$, then $\|\tilde{\Delta}\|_{\mathcal{F}} \leq 3$ and we can use slide 10 and the fact that $T_2 \leq 1$: $\rightarrow$ yields $\frac{1}{2}\|\tilde{\Delta}\|_n^2 \leq c\delta_{n;R}\|\tilde{\Delta}\|_n + \lambda_n$,

b) Event 2: $\|\widetilde{f}\|_{\mathcal{F}} > 2 > 1 \geq \|\widetilde{f^\star}\|_{\mathcal{F}} \rightarrow \|\tilde{\Delta}\|_{\mathcal{F}} \geq 1$

- T_1: can still bound T_1 using similar idea as in sl. 10, but iteratively (peeling lemma) on event $\|\tilde{\Delta}\|_{\mathcal{F}} \geq 1$ (MW Lem. 13.23) yields with probability at least $\geq 1 - c_1 e^{-\frac{n\delta_n^2}{c_2\tilde{\sigma}^2}}$

$$\sup_{\tilde{\Delta}\in\mathcal{F}^\star, \|\tilde{\Delta}\|_{\mathcal{F}}\geq 1} \frac{\tilde{\sigma}}{n}\sum_i w_i\tilde{\Delta}(x_i) \leq 2\delta_n\|\tilde{\Delta}\|_n + 2\delta_n^2\|\tilde{\Delta}\|_{\mathcal{F}} + \frac{\|\tilde{\Delta}\|_n^2}{16} \quad (1)$$

- T_2: $\lambda_n(\|\widetilde{f^\star}\|_{\mathcal{F}}^2 - \|\widetilde{f}\|_{\mathcal{F}}^2) \leq 2\lambda_n - \lambda_n\|\tilde{\Delta}\|_{\mathcal{F}}$ using $\|\tilde{\Delta}\|_{\mathcal{F}} \leq \|\widetilde{f}\|_{\mathcal{F}} + \|\widetilde{f^\star}\|_{\mathcal{F}}$ and $\|\widetilde{f^\star}\|_{\mathcal{F}}^2 - \|\widetilde{f}\|_{\mathcal{F}}^2 \leq \|\widetilde{f^\star}\|_{\mathcal{F}} - \|\widetilde{f}\|_{\mathcal{F}}$
$\rightarrow$ green “swallows” red term for large enough $\lambda_n \geq 2\delta_n^2$
$\rightarrow$ regularization takes care of not having explicit norm bound!

- Putting things together yields $\frac{1}{2}\|\tilde{\Delta}\|_n^2 \leq c\delta_n\|\tilde{\Delta}\|_n + \frac{1}{16}\|\tilde{\Delta}\|_n^2 + 2\lambda_n$

Peeling lemma idea - MW Lem. 13.23 (skipped in class)

- The idea is to make T_1 depend on the $\mathcal{F}$-norm which we can then "kill" via regularization (large enough λ_n)
- By star-shapedness of $\mathcal{F}$ we only need to show inequality with sup over $\|\tilde{\Delta}\|_{\mathcal{F}} = 1$
- However then, we no longer have $\|\tilde{\Delta}\|_n \geq \delta_n$ (can essentially only use the star-shaped argument on one of the norms)
- Then we do something like in chaining - split up event where eq. 1 does not hold and $\|\tilde{\Delta}\|_{\mathcal{F}} = 1$ (without boundedness of $\|\tilde{\Delta}\|_n$) into subevents where $\|\tilde{\Delta}\|_n \in [t_m, t_{m+1}]$ with $t_m = 2^m \delta_n$ and union bound.
- Union bounding with this choice of t_m with the usual concentration bound (Lipschitz function of Gaussians in MW Thm 2.26)

For a detailed proof we refer to the book.

References

Reproducing Kernel Hilbert spaces:

- MW Chapter 12
- SC Chapter 4

Non-parametric regression:

- MW Chapter 13

Kernel eigenvalues (skipped in class)

- The empirical and population Gaussian complexities are close within constants MW Prop 14.25.
- population Gaussian compl. depends on kernel operator eigenvalues
- For $\mathcal{K}$ on compact $\mathcal{X}$ empirical matrix eigenvalues $\hat{\mu}_j \sim \mu_j$ for big n where μ_j are integral operator eigenvalues (Koltchinskii, Gine '00)

Define bounded, linear Hilbert-Schmidt integral operator $T_{\mathcal{K}} : \mathcal{L}^2 \to \mathcal{L}^2$ with $T_{\mathcal{K}} f = \int \mathcal{K}(x, y) f(y) dy$, and we call μ_j eigenvalues and ψ_j eigenfunctions if $T_{\mathcal{K}} \psi_j = \mu_j \psi_j$

Theorem (Mercer's) (SC Thm 4.49, 4.51, MW Thm 12.20)

For $\mathcal{K}$ psd with RKHS $\mathcal{F}_{\mathcal{K}}$, there exist eigenfcuntions and eigenvalues $\psi_j, \mu_j \geq 0$ of $T_{\mathcal{K}}$ that satisfy

1. *ψ_j form an ONB in $\mathcal{L}^2(\mathbb{P})$ and $\phi_j = \sqrt{\mu_j}\psi_j$ is an ONS in $\mathcal{F}_{\mathcal{K}}$.*
2. *$\mathcal{K}(x, y) = \sum_j \mu_j \psi_j(x) \psi_j(y)$ converges in $\mathcal{L}^2(\mathbb{P})$*
3. *If $\mathcal{K}$ also continuous, above sum converges absolutely and uniformly*

Crucial: μ_j, ψ_j depends on distribution $\mathbb{P}$!

Proof of Mercer's Theorem (skipped in class)

1. Main component: Hilbert-Schmidt Theorem (spectral theorem) (e.g. Knapp Thm 2.5., any functional analysis book)
   - For any kernel, $T_{\mathcal{K}}$ is compact, self-adjoint, has eigenspaces
   - decomposition of image of $T_{\mathcal{K}}$ into ψ_j (countable) ONB of $\mathcal{L}_2$ that are eigenvectors of $T_{\mathcal{K}}$
   - sum converges in $\mathcal{L}^2$.
2. Positivity by definition of the operator and kernel psd
3. Why $T_{\mathcal{K}}$ maps to $\mathcal{F}_{\mathcal{K}}$ SC 4.26.: Hoelder ineq, Bochner integrability
4. Absolute uniform convergence of sum for continuous kernel: Non-decreasing sequences of continuous functions with a continuous limit converge uniformly (e.g. Rudin 7.13).

Notes in S.C. they define it $T_{\mathcal{K}}$ more rigorously