Lecture 5: General non-parametric regression prediction error bound and from feature maps to RKHS

Announcements

• HW 2 was posted, due in two weeks on Sunday
• Project proposal due this Friday
• Office hours will happen on demand (write email to TA(s))
• Re Corona:
  • Lectures will also be recorded
  • Some form of midterm assessment will happen on 29.4, most probably via a remote zoom session, but will have to be confirmed
• Re Piazza Feedback:
  • Incorporating proof details + more examples + slower pace simultaneously unfortunately won’t work :)
  • → will try to put more examples (or mark them as such)
Prediction error bound for constrained $^2$-loss minimizer

**Definition (Critical inequality, quantity)**

We say $\delta_n$ is the critical quantity/radius, if it is the smallest $\delta > 0$ to satisfy the critical inequality

$$\frac{\tilde{G}_n(F^*; \delta)}{\delta} \leq \frac{\delta}{\sigma}$$

**Theorem (Prediction error bound)**

If $F^*$ is star-shaped, $\frac{\tilde{G}_n(F^*; \delta)}{\delta}$ is non-decreasing, $\delta_n$ exists and we have for the square loss minimizer $\hat{f}$ for any $t \geq 1$

$$\mathbb{P}(\|\hat{f} - f^*\|_n^2 \geq 16t\delta_n^2) \leq e^{-\frac{nt^2\delta_n^2}{2\sigma^2}}$$
Illustration localized complexity

Figure 1: Blue solid: \( f(\delta) = \frac{\tilde{G}_n(F;\delta)}{\delta} \), Green dashed: \( f(\delta) = \delta \)

Directly helps for \( \ell_0 \)

Let’s say we’re trying to find the best sparse linear fit

\[ \hat{f} = \arg \min_{f \in \mathcal{F}_{lin, s}} \| y - X\theta \|_2^n \]

with \( \mathcal{F}_{lin, s} = \{ f(\cdot) = \langle \theta, x \rangle : \| \theta \|_0 \leq s \} \)

- In HW 2 we prove \( \tilde{G}_n(\mathcal{F}_{lin, s}; \delta) \leq O(\delta \sqrt{\frac{s \log(ed/s)}{n}}) \) when \( \lambda_{\max}(\frac{X_S^T X_S}{n}) \) bounded for all subsets \( S \) of size \( s \)

- Hence the critical radius satisfies \( \frac{\tilde{G}_n(\mathcal{F}_{lin, s}; \delta)}{\delta} = \sqrt{\frac{s \log(ed/s)}{n}} \leq \frac{\delta}{\sigma} \)

- Thus using the theorem plugging in \( \delta^2_n \) at equality, we can obtain with probability at least \( 1 - \delta \)

\[ \| \hat{f} - f^* \|_2^n \leq O \left( \frac{s \log(ed/s) \log 1/\delta}{n} \right) \]

Also see MW Example 13.16.
Dudley’s integral for critical radius

**Theorem (Dudley’s integral & critical quantity)**

If $\mathcal{F}$ is star-shaped, any $\delta \in [0, \sigma]$ such that

\[
\frac{16}{\sqrt{n}} \int_0^{\delta} \sqrt{\log \mathcal{N}(t; \mathcal{F}(x_1^n) \cap \mathbb{B}_n(\delta), \| \cdot \|_n)} dt \leq \frac{\delta^2}{4\sigma}
\]

does not satisfy the critical inequality.

Proof via chaining (see MW Cor. 13.7.)

**Examples**

1. $\mathcal{F}_L$: Lipschitz functions on $[0, 1]$ and $f(0) = 0$ has $\log \mathcal{N}(\epsilon) \leq O\left(\frac{L}{\epsilon}\right)$

\[
\frac{1}{\sqrt{n}} \int_0^{\delta} \sqrt{\log \mathcal{N}(t; \mathcal{F}_L(x_1^n), \| \cdot \|_n)} dt \leq \frac{1}{\sqrt{n}} \int_0^{\delta} \left(\frac{L}{t}\right)^{\frac{1}{4}} dt \leq \sqrt{\frac{L\delta}{n}} \leq \frac{\delta^2}{4\sigma^2}
\]

→ Rearranging terms yields $\| \hat{f} - f^*\|_n^2 \leq \delta_n(\mathcal{F}_L)^2 = O\left(\frac{L\sigma^2}{n}\right)^{\frac{2}{3}}$

**Example 2: linear regression**

2. $\mathcal{F}_{1,c}$: $f \in \mathcal{F}_1$ and convex, has $\log \mathcal{N}(\epsilon) \leq O\left(\frac{1}{\epsilon^{1/2}}\right)$

\[
\frac{1}{\sqrt{n}} \int_0^{\delta} \sqrt{\log \mathcal{N}(t; \mathcal{F}_{1,c}(x_1^n), \| \cdot \|_n)} dt \leq \frac{1}{\sqrt{n}} \int_0^{\delta} \left(\frac{1}{t}\right)^{\frac{1}{4}} dt \leq \frac{\delta^{3/4}}{\sqrt{n}} \leq \frac{\delta^2}{4\sigma^2}
\]

→ Rearranging terms yields $\delta_n(\mathcal{F}_{1,c})^2 = O\left((\frac{\sigma^2}{n})^{\frac{4}{5}}\right)$

Reflection of size via $\delta_n(\mathcal{F})$ vs. $\mathcal{R}_n(\mathcal{F})$ using Dudley (see also HW 2)

- $\delta_n(\mathcal{F})$: Critical quantity reflects difference in metric entropy (size)
- $\mathcal{R}_n(\mathcal{F})$ via Dudley: If integrals $\int_0^D \sqrt{\log \mathcal{N}(t; \mathcal{F}(x_1^n), \| \cdot \|_n)} dt$ are bounded, then best is to use that and R.C. gets $\frac{1}{\sqrt{n}}$ rate. (check)
  → For both integrals are bounded, Rademacher complexity has $\frac{1}{\sqrt{n}}$
  → does not reflect size difference compared to $\delta_n(\mathcal{F})$!
Proof of Theorem on prediction error

• Recall $\Delta = f - f^*$, $t > 1$. It suffices to show for any $\|\hat{\Delta}\|_n \geq \sqrt{t\delta_n}$,

$$\|\hat{\Delta}\|_n^2 \leq \frac{2\sigma}{n} \sum_{i=1}^{n} w_i \hat{\Delta}(x_i) \overset{(i)}{\leq} 4\|\hat{\Delta}\|_n \sqrt{t\delta_n}$$

w/ prob. $\geq 1 - e^{-\frac{nt\delta_n^2}{2\sigma^2}}$. The theorem follows from rearranging terms.

• We already established that $\neg(i)$ is upper bounded by

$$\mathbb{P}( \sup_{\|\hat{\Delta}\|_n \leq \sqrt{t\delta_n}} \|\hat{\Delta}\|_n \leq \sqrt{t\delta_n} \sigma \sum_{i=1}^{n} w_i \hat{\Delta}(x_i) \leq 2t\delta_n^2)$$

• We now upper bound the RHS by $\leq e^{-\frac{nt\delta_n^2}{2\sigma^2}}$ via tail bounding the process

$$g_n(w) = \sup_{\|\hat{\Delta}\|_n \leq \sqrt{t\delta_n}} \frac{\sigma}{n} \sum_{i=1}^{n} w_i \hat{\Delta}(x_i)$$

Proof of error bound: tail bounding $g_n(w)$

We now establish the tail bound for $g_n(w)$

1. $g_n(w)$ as a function of $w_i \sim \mathcal{N}(0,1)$ is $\frac{\sigma \sqrt{t\delta_n}}{\sqrt{n}}$-Lipschitz so that

$$\mathbb{P}(g_n(w) \geq \mathbb{E}g_n(w) + s) \leq e^{-\frac{ns^2}{2\sigma^2 t\delta_n^2}}$$

(see Lecture 1 / MW Thm 2.26)

2. Furthermore $\mathbb{E}g_n(w) = \widetilde{G}_n(F; \sqrt{t\delta_n})$

3. The map $\delta \rightarrow \frac{\widetilde{G}_n(F; \delta)}{\delta}$ is non-decreasing by MW Lemma 13.6.

4. By 2. and definition of $\delta_n$ we have $\sigma \frac{\widetilde{G}_n(F; \sqrt{t\delta_n})}{\sqrt{t\delta_n}} \leq \sigma \frac{\widetilde{G}_n(F; \delta_n)}{\delta_n} \leq \delta_n$ and setting $s = t\delta_n^2$, we obtain

$$\mathbb{P}( \sup_{\|\hat{\Delta}\|_n \leq \sqrt{t\delta_n}} \frac{\sigma}{n} \sum_{i=1}^{n} w_i \hat{\Delta}(x_i) \geq 2t\delta_n^2) \leq \mathbb{P}( \sup_{\|\hat{\Delta}\|_n \leq \sqrt{t\delta_n}} \frac{\sigma}{n} \sum_{i=1}^{n} w_i \hat{\Delta}(x_i) \geq \sigma \widetilde{G}_n(F; \sqrt{t\delta_n}) + t\delta_n^2) \leq e^{-\frac{nt\delta_n^2}{2\sigma^2}} \Box
Moving beyond linear (and parametric) classifiers: RKHS

Chat-Q: What do you remember about the keys in the kernel trick?

Handwritten: From features maps to kernels to ∞-dim Hilbert spaces

Properties of Reproducing Kernel Hilbert space \( \mathcal{F} \) defined via

- the basis (feature maps) \( \phi \) of a native space \( \mathcal{F}^0 \) and
- eigenvalues \((\mu_j)_{j=1}^\infty\) (why called eigenvalues \( \rightarrow \) next week):

1. Has basis \( \{\tilde{\phi}_j : \sqrt{\mu_j} \phi_j\}_{j=1}^\infty \), and \( \mathcal{F} = \{f = \sum_{j=1}^\infty \beta_j \phi_j : \sum_{j=1}^\infty \frac{\beta_j^2}{\mu_j} < \infty\} \)

\((\mu_j)_{j=1}^\infty\) determines “size” of \( \mathcal{F} \)! Fast decay \( \rightarrow \) small \( \mathcal{F} \)!

2. Inner product for \( f = \sum_j \beta_j \phi_j, g = \sum_j \gamma_j \phi_j \) reads \( \langle f, g \rangle_\mathcal{F} = \sum_j \frac{\beta_j \gamma_j}{\mu_j} \)

3. The kernel function \( K(x, y) = \sum_j \mu_j \phi_j(x)\phi_j(y) \)
   - as an inner product defines feature map space \( \mathcal{H} \)
   - is psd, i.e. kernel matrix \( K_{ij} := \frac{K(x_i, x_j)}{\sqrt{n}} \) for any \( x_1, \ldots, x_n \) is psd

4. \( K(\cdot, y) \) is reproducing, that means for all \( f = \sum_j \beta_j \phi_j \in \mathcal{F} \)
   \[
   \langle K(\cdot, y), f(\cdot) \rangle_\mathcal{F} = \langle \sum_j \mu_j \phi_j(y)\phi(\cdot), \sum_j \beta_j \phi_j \rangle = \sum_j \mu_j \beta_j \phi_j(y) = f(y) \]
   since \( \langle \sqrt{\mu_j} \phi_j, \sqrt{\mu_i} \phi_i \rangle_\mathcal{F} = 1_{i=j} \).

Examples for RKHS

4. \( K(\cdot, y) \) is reproducing, that means for all \( f = \sum_j \beta_j \phi_j \in \mathcal{F} \)

   \[
   \langle K(\cdot, y), f(\cdot) \rangle_\mathcal{F} = \langle \sum_j \mu_j \phi_j(y)\phi(\cdot), \sum_j \beta_j \phi_j \rangle = \sum_j \mu_j \beta_j \phi_j(y) = f(y) \]
   since \( \langle \sqrt{\mu_j} \phi_j, \sqrt{\mu_i} \phi_i \rangle_\mathcal{F} = 1_{i=j} \).

Examples for RKHS

- Desiring feature map \( \phi \): Degree \( m \) polynomial basis
  \[
  \phi_j(x) = x_1^{\alpha_1(j)} \cdots x_d^{\alpha_d(j)} \cdot 1^{\alpha_{d+1}} \cdot 1^{\alpha_{d+m}} \text{ with } \sum_{i=1}^{d+m} \alpha_{i}^{(j)} = m, \alpha^{(j)} \in \mathbb{N}^{d+m} \]
  induces kernel \( K(x, y) = (1 + \langle x, y \rangle)^m \)

- Desiring structured \( \mathcal{F} \): \( \alpha \)-differentiable functions, with square-integrable derivatives and \( \mathcal{F} \subset \mathcal{F}^0 = L^2(\mathbb{P}) \) (HW 2,3)

- Desiring \( K \) (or \( \mu \)): e.g. Gaussian kernel \( K(x, y) = e^{-\frac{||x-y||^2}{2\sigma}} \)

Can in fact always find feature map \( \phi \) given \( K, \mathcal{F}! \)
(existence via Mercer’s theorem (next week))
References

Dudley and Prediction error bound:
- MW Section 13.2.

Feature maps and RKHS
- Percy Liang: Lecture Notes: Lecture 10
- MW Chapter 12

Non-parametric regression in RKHS with bounded norm
- MW Chapter 12.5., 13.4.
Specific example for non-parametric function classes: kernel spaces

**Linear function**

\[ f(x) = \sum_{j=1}^{n} \beta_j \phi_j(x) \]

**Polynomial with degree n**

\[ f(x) = \sum_{i=0}^{n} \beta_i \phi_i(x) \]

**Infinite-dim Hilbert space**

\[ F^0 = L^2(\mathbb{R}) \]

\[ f(x) = \sum_{i=1}^{\infty} \beta_i \phi_i(x) \]

**Least-squares estimate**

\[ \hat{\beta} = \sum_{i=1}^{n} \beta_i \phi_i(x_i) \]

\[ \hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (x_i - \phi_i(x))^2 \]

**Kernel trick**

\[ K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \]

Why do we want this?

- Kernel trick: \( K(x_i, x_j) \) fast to compute
- \( K(x, y) = \exp(-\|x-y\|^2) \)

Plan from "desire" to rigor (real rigor next week)

Feature map \[ \phi \rightarrow H \] with \( \|\phi(x)\|_H < \infty \)

\[ F \rightarrow \text{functional of the form} \]

- \( a = \phi(x), \phi(y) = K(x, y) \)

Define some Hilbert space \( H \) s.t. \( \phi : X \rightarrow H \)

and \( \|\phi(x)\|_H < \infty \) for all \( x \)

and \( K(x, y) = \langle \phi(x), \phi(y) \rangle > 0 \)

\[ \left\{ \begin{array}{l}
\text{Naive} \ \rightarrow \ \langle \phi(x), \phi(y) \rangle \geq 0 \\
\text{Try} 1 \ \rightarrow \ \langle \phi(x), \phi(y) \rangle := K(x, y) \end{array} \right. \]

for some basis \( \phi_1, \phi_2, \ldots \)

\[ \text{not guaranteed for all } x \]

\[ \Rightarrow \left\{ \begin{array}{l}
\text{Try} 2 \ \rightarrow \ \langle \phi(x), \phi(y) \rangle := K(x, y) \end{array} \right. \]

many possible

\[ \checkmark \text{Note } K \text{ is } \]
2. Want new Hilbert space $\mathcal{F}^*$ with orthonormal basis $\{\phi_j\}_{j=1}^\infty$

Note: $\mathcal{F}$ includes $f: \mathcal{X} \rightarrow \mathbb{R}$ whereas $\mathcal{H}$ includes sequences $(h_j)_{j} \in (\mathbb{R}^2)^\infty$
linear combination of features

inner product should satisfy $<\phi_j, \phi_j> = 1$ and $<\phi_j, \phi_i> = 0 \quad \forall \; i \neq j$

Since $\{\phi_j\}_{j=1}^\infty$ ORB for $\mathcal{F}$, all $f, g \in \mathcal{F}^*$ can be written $f = \sum \phi_j \tilde{b}_j$

hence $<f, g>_{\mathcal{F}^*} = <\sum \phi_j \tilde{b}_j, \sum \phi_i \tilde{c}_i>_{\mathcal{F}^*} = \sum_{j=1}^\infty \tilde{b}_j \tilde{c}_j = \sum_{j=1}^\infty \frac{\tilde{b}_j \tilde{c}_j}{\tilde{\phi}_j^2}$

For $f$ to be in $\mathcal{F}^*$ w/ new inner product, we require for $f = \sum \phi_j \tilde{b}_j$

$\|f\|_{\mathcal{F}^*}^2 = \sum \frac{\tilde{b}_j^2}{\tilde{\phi}_j^2} < \infty$, i.e. $(\frac{\tilde{b}_j}{\tilde{\phi}_j})_{j=1}^\infty \in \ell^2(\mathbb{N})$

$\Rightarrow$ RKHS $\mathcal{F}$ for choice of: (i) fixed basis $\{\phi_j\}_{j=1}^\infty$ given desired $\mathcal{F}$ (ii) choice $\{\phi_j\}_{j=1}^\infty$

$f = \sum \phi_j \tilde{b}_j \phi_j(\cdot) \quad \forall \phi_j \tilde{b}_j$ satisfying $\sum \frac{\tilde{b}_j^2}{\tilde{\phi}_j^2} < \infty$

Note: $\tilde{b}_j$ determines size of space, the faster it "decays" the smaller