Lecture 10: NTK and random design

Announcements

- HW 2 released, due 9.11. 23:59
- HW 1 grades released these days via gradescope

Plan for today

- Prediction error bound for random design
- Add-on: Random features and NTK

Random design

- So far, we only controlled $\|\widehat{f} - f^\star\|_n^2$ w.h.p. over observation noise w

$$\|\widehat{f} - f^\star\|_n^2 = R(\widehat{f}) - R(f^\star) = \mathbb{E}_w \frac{1}{n}\sum_{i=1}^n (y_i - \widehat{f}(x_i))^2 - \mathbb{E}_w \frac{1}{n}\sum_{i=1}^n w_i^2$$
$$= \frac{1}{n}\sum_{i=1}^n (\widehat{f}(x_i) - f^\star(x_i))^2$$

- can be bounded using empirical Gaussian complexities via basic inequality → basic inequality

How does the error look like on the whole domain $\mathcal{X}$?

Now we view X as random and take expectation also over X, i.e. for any $f \in \mathcal{L}^2(\mathbb{P})$, we have

$$\|f - f^\star\|_2^2 = R(f) - R(f^\star) = \mathbb{E}_{X,W}(Y - f(X))^2 - \mathbb{E}W^2$$
$$= \mathbb{E}_X(f(X) - f^\star(X))^2 = \mathbb{E}_{x_1,\ldots,x_n}\|f - f^\star\|_n^2$$

and want to bound $\|\widehat{f} - f^\star\|_2^2$ for an estimator $\widehat{f}$

Prediction error bound for random design - uniform law?

Maybe use $\|\widehat{f} - f^\star\|_2^2 - \|\widehat{f} - f^\star\|_n^2 \leq \sup_{f\in\mathcal{F}} \|f - f^\star\|_2^2 - \|f - f^\star\|_n^2$ and then plug in previous bound on $\|\widehat{f} - f^\star\|_n^2$?

Definition (Rademacher complexity - recap)

Given a function class $\mathcal{H}$ and distribution $\mathbb{P}$ on its domain $\mathcal{Z}$, we define the Rademacher complexity as

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_{\epsilon,z} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \epsilon_i h(z_i)$$

Theorem (Uniform law - recap)

For b-unif. bounded $\mathcal{H}$ with $\mathcal{R}_n(\mathcal{H}) = \mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i h(z_i)$

$$\mathbb{P}(\sup_{h\in\mathcal{H}} \mathbb{E}h - \frac{1}{n}\sum_{i=1}^n h(z_i) \geq 2\mathcal{R}_n(\mathcal{H}) + t) \leq e^{-\frac{nt^2}{2b^2}}$$

w/ prob. over the training data. If $\mathcal{R}_n(\mathcal{H}) = o(1)$, then $\sup_{h\in\mathcal{H}} R(h) - R_n(h) \overset{a.s.}{\to} 0$.

Using the uniform law for (uniformly bounded) regression

Partner-Q: Using the uniform law, derive a h.p. upper bound for $\|\widehat{f} - f^\star\|_2^2$ for linear functions $f(x) = \langle w, x \rangle$ with $\|x\|_2 \le D, \|w\|_2 \le B$, bounded noise. Use Rademacher contraction

It suffices to bound $\mathbb{E}_X(Y - \widehat{f}(X))^2 = \|\widehat{f} - f^\star\|_2^2 + \sigma^2$ using a uniform law on the generalization error with the square loss

$$R(f) - R_n(f) := \mathbb{E}_X(Y - \widehat{f}(X))^2 - \|y - \widehat{f}(x_1^n)\|_2^2$$

First of all, in this setting, by assumption, the loss is uniformly bounded since $|y_i - f(x_i)| \le D'$ is bounded by some constant D'.

- Define $\widetilde{\mathcal{F}}(z_1^n) = \{(y_1 - f(x_1), \ldots, y_n - f(x_n)) : f \in \mathcal{F}\} \subset \mathbb{R}^n$
- Then for the square function $\ell_{sq}(u) = u^2$ for $|u| \le D'$ we have $|\ell_{sq}(u) - \ell_{sq}(u')| \le |u^2 - u'^2| \le |u - u'||u + u'| \le 2D'|u - u'|$, i.e. ℓ_{sq} is $2D'$-Lipschitz
- Then, analogous to the SVM example, we have $\mathcal{H}(z_1^n) = \ell_{sq} \circ \widetilde{\mathcal{F}}(z_1^n)$ and $\tilde{\mathcal{R}}_n(\mathcal{H}(z_1^n)) \le 2D' \tilde{\mathcal{R}}_n(\mathcal{F}(z_1^n))$ using Rademacher contraction, and where $\mathcal{F}$ is the space of bounded linear functions

Motivating the localized uniform law

- Analogously to the SVM excess risk bound, the uniform law yields a squared error bound of order $O(1/\sqrt{n}) \to$ highly suboptimal!

$\to$ In fact, we can *localize* the uniform law as well!

- in the sequel, we write g for $f - f^\star$ instead of $\hat{\Delta}$ for simplicity
- Indeed, for b-uniformly bounded $\mathcal{F}^\star$, we can define the critical inequality on the *population* localized Rademacher complexity

$$\mathcal{R}_n(\mathcal{F}^\star; \delta) = \frac{1}{n}\mathbb{E}_{X,\epsilon} \sup_{g \in \mathcal{F}, \|g\|_2 \le \delta} \sum_{i=1}^n \epsilon_i g(x_i) \le \frac{\delta^2}{16b}$$

- Let $\overline{\delta}_n$ be a δ that satisfies this inequality.

Now what? Can't directly use our localization / basic inequality approach, since that only holds for finite samples!

Precise statement of localized uniform law

Theorem (Localized uniform law, MW Thm 14.1)

For star-shaped and b−uniformly bounded $\mathcal{F}^\star$, let $\bar\delta_n$ as defined above. Then if $\bar\delta_n^2 > c\frac{\log[4\log(1/\bar\delta_n)]}{n}$ then w.p. at least $1 - c_1 e^{-c_2\frac{n\bar\delta_n^2}{b^2}}$ we have

$$\sup_{g\in\mathcal{F}^\star} \|g\|_2 - \|g\|_n \le c\bar\delta_n$$

- Note that the condition is not too strong: if $\bar\delta_n \asymp 1/n$, i.e. we have the best possible achievable rate, then the inequality is still true for small enough c (only slightly depending on n), since $\log\log n$ is "almost constant". For $\bar\delta_n \ge \omega(1/n)$, this condition always holds for large enough n.

Recall in the proof for empirical prediction error:

- For localization we used the basic inequality for the empirical error
- There we had LHS $\|g\|_n^2$ with $g \in \mathcal{F}^\star$ which we self-bounded by $\delta_n\|g\|_n$ when $\|g\|_n > \delta_n$

Proof idea for localized uniform law

- We can do something similar here: we choose $\|g\|_2^2 - \|g\|_n^2$ as our RHS and will also "self-upper-bound" it
- Observe that the binomial formula yields for any $g \in \mathcal{F}^\star$

$$\|g\|_2 - \|g\|_n = \frac{\|g\|_2^2 - \|g\|_n^2}{\|g\|_2 + \|g\|_n}$$

- Hence the proof goes through either with
  a) $\frac{\|g\|_2^2 - \|g\|_n^2}{\|g\|_2 + \|g\|_n} \le \bar\delta_n$ if $\|g\|_2 \le \bar\delta_n$
  b) or $\sup_{g\in\mathcal{F}^\star, \|g\|_2 \ge \bar\delta_n} \|g\|_2^2 - \|g\|_n^2 \le \|g\|_2\bar\delta_n$ w.h.p. if $\|g\|_2 \ge \bar\delta_n$ (uniformly for all $g \in \mathcal{F}^\star$) yields

We give intuition for the proof of b)

Proof of b): case $\|g\|_2 \geq \bar{\delta}_n$

For simplicity of the proof, assume $b = 1$ and hence $\|g\|_2 \leq 1$ (general case follows from scaling arguments as last time)

1. Step: For fixed $r \geq \bar{\delta}_n$, bounding $\sup_{g\in\mathcal{F}^\star,\|g\|_2\leq r} \|g\|_2^2 - \|g\|_n^2$ (MW Lemma 14.9.)

- symmetrization and Rademacher contraction for $r \geq \bar{\delta}_n$

$$\mathbb{E} \sup_{g\in\mathcal{F}^\star,\|g\|_2\leq r} \|g\|_2^2 - \|g\|_n^2 \leq 2\mathbb{E} \sup_{g\in\mathcal{F}^\star,\|g\|_2\leq r} \frac{1}{n}\sum_{i=1}^n \epsilon_i g^2(x_i)$$
$$\leq 4\mathbb{E} \sup_{g\in\mathcal{F}^\star,\|g\|_2\leq r} \frac{1}{n}\sum_{i=1}^n \epsilon_i g(x_i) \leq r\bar{\delta}_n$$

where the last inequality follows from definition of $\bar{\delta}_n$

- we then use Talagrand concentration (MW Thm 3.27) to derive that w.p $\geq 1 - \mathrm{e}^{-cn\bar{\delta}_n^2}$ we have $\sup_{g\in\mathcal{F}^\star,\|g\|_2\leq r} \|g\|_2^2 - \|g\|_n^2 \leq \frac{r\bar{\delta}_n}{2}$

Proof of b): case $\|g\|_2 \geq \bar{\delta}_n$

2. Step: If we could plug in $r = \|g\|_2$ we'd be done, but above h.p. bound only holds for fixed r!

- Use peeling argument like before and split $S := \{\sup_{g\in\mathcal{F}^\star,\|g\|_2\geq\bar{\delta}_n} \|g\|_2^2 - \|g\|_n^2 \geq \|g\|_2\bar{\delta}_n\}$ into sub-events: $S_m = \{\|g\|_2 \in [t_{m-1}, t_m]\}$ where $t_m = 2^m\bar{\delta}_n$. In particular, by uniform boundedness $\|g\|_2 \leq 1$, we have that $S \subset \bigcup_{m=1}^M \{S \cap S_m\}$ with $M = 4\log(1/\bar{\delta}_n)$

- using $\sup_{g\in\mathcal{F}^\star,\|g\|_2\leq r} \|g\|_2^2 - \|g\|_n^2 \leq \frac{r\bar{\delta}_n}{2}$ with $r = t_m$ and using union bound gives

$$\mathbb{P}(S) \leq \sum_{m=1}^M \mathbb{P}(S\cap S_m) \leq \sum_{m=1}^M \mathbb{P}(\sup_{g\in\mathcal{F}^\star,\|g\|_2\leq t^m} \|g\|_2^2 - \|g\|_n^2 \geq \frac{t_m\bar{\delta}_n}{2})$$
$$\leq \sum_{m=1}^M \mathrm{e}^{-cn\bar{\delta}_n^2} \leq \mathrm{e}^{-cn\bar{\delta}_n^2 + \log M} \leq \mathrm{e}^{-cn\bar{\delta}_n^2}$$

Kernel $\rightarrow$ feature maps (unbounded, translation-invariant)

We saw some examples for RKHS and their kernels with **compact supports** (e.g Sobolev spaces). What if domain is non-compact?

Consider RBF kernels $\mathcal{K}(x, y) = h(x - y)$

Theorem (Bochner: feature maps for translation-invariant kernels)

If $\mathcal{K}(x, y) = h(x - y)$ with h continuous and $x, y \in \mathbb{R}^d$, then there is a unique, finite, non-negative measure μ on $\mathbb{R}^d$ such that

$$h(t) = \int_{\mathbb{R}^d} e^{-i\langle t, \omega \rangle} \mu(d\omega)$$

Reminiscent of the Fourier basis, we call μ spectral measure, and if it has a density, we call $s(\omega)d\omega = \mu(d\omega)$ the spectral density

Kernels as expectations

For Gaussian kernels $\mathcal{K}(x, y) = \mathrm{e}^{-\frac{\|x-y\|_2^2}{2\sigma^2}}$ on $\mathbb{R}^d$ where Bochner holds with $s(\omega) = \left(\frac{2\pi}{\sigma^2}\right)^{-d/2} \mathrm{e}^{-\frac{\sigma^2 \|\omega\|_2^2}{2}}$ (Fourier transform)

- For feature maps $\phi(\omega; x) = \mathrm{e}^{-\mathrm{i}\langle x, \omega \rangle}$, we can rewrite the kernel as an expectation over measure $\mu(\mathrm{d}\omega) = s(\omega)\mathrm{d}\omega$, i.e.

$$\mathcal{K}(x, y) = \mathbb{E}_{\omega \sim \mu} \phi(\omega; x)\phi(\omega; y) = \langle \phi(\cdot; x), \phi(\cdot; y) \rangle_{\mathcal{L}^2(\mu)}$$

  proof by completing the square

The corresponding kernel space $\mathcal{F}_{\mathcal{K}}$ can be described as follows:

- kernel space
  $\mathcal{F}_{\mathcal{K}} = \{f : f(x) = \int \tilde{f}(\omega)\mathrm{e}^{-\mathrm{i}\langle x, \omega \rangle} \mu(\mathrm{d}x) = \langle \tilde{f}, \phi \rangle_{\mathcal{L}^2(\mu)}, \tilde{f} \in \mathcal{L}^2(\mu)\}$

Kernels as expectations → random features

- Instead of the true expectation, can approximate/unbiased estimate $\mathcal{K}$ via empirical expectation $\hat{\mathbb{P}}_m$ over m samples of ω_j from μ

$$\hat{\mathcal{K}}(x,y) = \mathbb{E}_{\omega\sim\hat{\mathbb{P}}_m}\phi(\omega;x)\phi(\omega;y) := \frac{1}{m}\sum_{j=1}^{m}\phi(\omega_j;x)\phi(\omega_j;y)$$

- w/ (approx) m-dim feature map $\widehat{\phi}(x) = \frac{1}{\sqrt{m}}(\phi(\omega_1;x),\ldots,\phi(\omega_m;x))$
- can then again define the induced RKHS $\mathcal{F}_{\hat{\mathcal{K}}} = \{f : f = \frac{1}{m}\sum_{j=1}^{m}\tilde{f}(\omega_j)\phi(\omega_j;x), \tilde{f}\in\mathcal{H}\}$

Random features 'ctd

Theorem (Approximation for random features, Rahimi Recht '08)

For $f = \mathbb{E}_{\omega\sim\mu}\tilde{f}(\omega)\phi(\omega;\cdot)\in\mathcal{F}_{\mathcal{K}}$ with $\|\tilde{f}\|_\infty \le C$, define $\widehat{f} = \mathbb{E}_{\omega\sim\hat{\mathbb{P}}_m}\tilde{f}(\omega)\phi(\omega;\cdot)\in\mathcal{F}_{\hat{\mathcal{K}}}$. Then w/ prob. $\ge 1-\delta$ we have

$$\|\widehat{f}-f\|^2_{\mathcal{L}^2(\mathbb{P})} \le \frac{C}{\sqrt{m}}(1+\sqrt{2\log 1/\delta}).$$

- Proof via McDiarmid + Jensen's (on the expectation of norms) (see Percy Liang's notes)
- ∞-dim to n-dim to m-dim problem, since we can just solve linear problem by expressing $f(x_1^n) = \Phi\alpha$ with $\alpha\in\mathbb{R}^m$

→ choosing m too small gets bad approx. error. In practice would choose $\sim n$ (statistical error), so no real computational gain if no additional structural assumptions are made on $\mathcal{F}_{\mathcal{K}}$

Example: two-layer fully-connected NN

- Taylor "linearization" around initialization of width-m 2-layer NN

$$f_{NN}(x) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_j \sigma(\langle w_j, x\rangle) \approx \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_{0,j}\sigma(\langle w_{0,j}, x\rangle)$$

$$+ \underbrace{\sum_j \frac{(a_j - a_{0,j})}{\sqrt{m}} \sigma(\langle w_{0,j}, x\rangle)}_{T_1(x)} + \underbrace{\sum_j (w_j - w_{0,j})^\top (a_{0,j} x \sigma'(\langle w_{0,j}, x\rangle))}_{T_2(x)}$$

where $w_{0,j} \overset{i.i.d.}{\sim} \mu_w$, $a_{0,j} \overset{i.i.d.}{\sim} \mu_a$ at initialization, w/ non-linearity σ

- $T_1 \in \mathcal{F}_{RF} := \{f_1 : f_1(x) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} s_j \sigma(\langle w_{0,j}, x\rangle), s \in \mathbb{R}^m\}$ with feature maps $\phi_j(x) = \sigma(\langle w_{0,j}, x\rangle) \to \mathcal{F}_{RF}$ has kernel $\hat{\mathcal{K}}(x, y) = \frac{1}{m} \sum_{j=1}^{m} \sigma(\langle w_{0,j}, x\rangle)\sigma(\langle w_{0,j}, y\rangle)$ that approximates $\mathcal{K}(x, y) = \mathbb{E}_{\mu_w} \sigma(\langle w_{0,j}, x\rangle)\sigma(\langle w_{0,j}, y\rangle)$ as the layer width $m \to \infty$

- $T_2 \in \mathcal{F}_{NTK} := \{f_2 : f_2(x) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} v_j^\top (a_{0,j} x \sigma'(\langle w_{0,j}, x\rangle)), v_j \in \mathbb{R}^d\}$ with feature maps $\phi_{ij} = x_i a_{0,j} \sigma'(\langle w_{0,j}, x\rangle)$, $i \in [d], j \in [m]$ $\to \mathcal{F}_{NTK}$ has kernel $\hat{\mathcal{K}}(x, y) = \frac{1}{m} \sum_{j=1}^{m} x^\top y \sigma'(\langle w_{0,j}, x\rangle)\sigma'(\langle w_{0,j}, y\rangle)$ that approximates $\mathcal{K}(x, y) = \mathbb{E}_{\mu_w} x^\top y \sigma'(\langle w_{0,j}, x\rangle)\sigma'(\langle w_{0,j}, y\rangle)$

Idea:

- $\mathcal{F}_{RF}$ corresponds to class where first layer stays fixed at initialized value, second layer trainable, and $\mathcal{F}_{NTK}$ vice versa

- sum of both kernels yields another kernel and hence forms a "new" RKHS $\mathcal{F} = \mathcal{F}_{RF} \oplus \mathcal{F}_{NTK}$

$\to$ You could say, optimizing 2-layer NN $\approx$ optimizing loss in RKHS ($\to$ analyzable!)

- linear expansion is only good when $\|w_j - w_{0,j}\|$ small $\to$ people show for large enough width changes are indeed small

- just showed that infinite-width limit kernels "make sense" (check out arc-cosine kernel)

- infinite width is far from what we use $\to$ people are trying to show optimization and generalization results for poly or logarithmic in n, d

References

Random design

- MW Chapter 14

Translation-invariant kernels and Random features

- *Percy Liang Lecture Notes*: Lectures 11, 12
- *Rahimi and Recht '08: Random Features for Large-Scale Kernel Machines* (Neurips)

Neural networks and kernels

- Matus Telgarsky's deep learning theory lectures: https://mjt.cs.illinois.edu/courses/dlt-f19~/files/lec5-handout.pdf
- *Cho, Saul '09: Kernel methods for deep learning* (Neurips): arc-cosine kernel
- NTK related: e.g. Jacot, Gabriel, Hongler '18, Chizat, Bach '19
- Approximation properties of $\mathcal{F}_{NTK}, \mathcal{F}_{RF}$ and the infinite width limit: Ghorbani, Misiakiewicz, Mei, Montanari '19, Mei, Montanari '19