Lecture 4: infinite width over $\mathbb{R}^q$; Fourier transforms & Barron norms.

Announcements:

* Picnic OH (more social than technical)
* HW1 out; have 20 days. Later HW shorter.

Recap.

* "Universal approximation"; worst-case, non-constructive.
* Infinite-width: univariate case, by FTC

$$\forall x \in [0,1], \quad g(x) - g(0) = \int_0^1 \mathbb{1}[x \geq b]\, g'(b)\, db$$

* We'll say more about sampling today.

Also today: Fourier & infinite-width over $\mathbb{R}^d$.

* Why infinite-width
  * Used in many optimization analyses.
  * Removes/simplifies many error terms.
  * In early phase of training, weights near <u>random</u> initialization; behaves like infinite-width due to concentration.

Sampling

Given $\int_0^1 \sigma(x-b)\, g(b)\, db = g(x)$

1. Sample $(b_1, \ldots, b_m) \sim \dfrac{|g(b)|}{\int_0^1 |g(b)|\, db}$.

2. For each j, set $a_j := \operatorname{sgn}(g(b_j)) \cdot \int_0^1 |g(b)|\, db \cdot \frac{1}{m}$

3. Output $f(x) := \sum_j a_j \sigma(x - b_j)$

Unbiased $\mathop{\mathbb{E}}_{b_1 \cdots b_m} f(x) = g(x)$

Error estimate:

$$\int_0^1 (f(x) - g(x))^2 dx \le \frac{1}{m} \left(\int_0^1 |g(b)|\, db \right)^2 \max_{b \in [0,1]} \int_0^1 \sigma(x-b)^2 dx$$

- $\frac{1}{m}$: $\sqrt{0}$ w/ number nodes
- $\int_0^1 |g(b)|\, db$: mass of weights
- $\max_{b} \int_0^1 \sigma(x-b)^2 dx$: penalty for activation

Remark ($\mathbb{R}^d$)

* in notes; $\frac{1}{m}$ still there (not $\frac{1}{m^{1/d}}$),
"mass" harder to estimate, might $\Omega(2^d)$
* More general version in notes.

Fourier & infinite width (Barron '93)

Goals (for multivariate infinite-width representation)

* $f(w) = \int \sigma(w^T x - b)\, q(b,w)\, d(b,w)$
* Mass $\int |q(b,w)|\, d(b,w)$ to not be worst case; e.g., $\int |q(b,w)|\, d(b,w) = \text{poly}(d)$ for some reasonable cases.

Recall: Fourier transform

$$\hat{f}(w) = \int \underbrace{\exp(-2\pi i w^T x)}_{\text{activation}} f(x)\, dx.$$

Fourier inversion: $f, \hat{f} \in L_1$ [Folland]

$$f(x) = \int \exp(2\pi i w^T x)\, \hat{f}(w)\, dw,$$

an infinite-width network!

Problem: want typical activation, not complex sinusoids.

Remarks
① Different convention.
② already inf-width network

Remark:

* This infinite-width form is not in (Barron '93), but first few steps of derivation are.

To start deriving threshold form, since $\text{Re}[\cdot] = f$,

$$f(x) = \text{Re}[f(x)] = \text{Re}\left[\int \exp(2\pi i w^T x)\, \hat{f}(w)\, dw\right]$$

$$= \int \text{Re}\left[\exp(2\pi i w^T x)\, \hat{f}(w)\right] dw$$

polar form $\hat{f}(w) = \underbrace{|\hat{f}(w)|}_{\text{magnitude}} \underbrace{\exp(2\pi i \theta(w))}_{\text{orientation}}$

$$= \int \text{Re}\left[\exp(2\pi i (w^T x + \theta(w)))\, |\hat{f}(w)|\right] dw$$

($|\hat{f}(w)|$ real)

$$= \int \underbrace{\text{Re}\left[\exp(2\pi i (w^T x + \theta(w)))\right]}_{\text{real-valued activation!}} \cdot |\hat{f}(w)|\, dw$$

$e^{iz} = \cos(z) + i\sin(z)$

$\cos(2\pi(w^T x + \theta(w)))$; need to fix notes

note

$$\cos(2\pi(w^T x + \theta(w))) - \cos(2\pi\theta(w)) = -\int_0^{w^T x} 2\pi \sin(2\pi(b + \theta(w)))\, db$$

$$= -2\pi \left[\int_0^{\|w\|} \mathbb{1}[w^T x \geq b] \sin(2\pi(b+\theta(w)))\, db - \int_{-\|w\|}^{0} \mathbb{1}[-w^T x \geq -b] \sin(2\pi(b+\theta(w)))\, db\right]$$

$$f(x) - f(0) = -2\pi \left[\int_0^{\|w\|} \mathbb{1}[w^T x \geq b] \sin(2\pi(b+\theta(w)))\, |\hat{f}(w)|\, db\, dw - \int_{-\|w\|}^{0} \mathbb{1}[-w^T x \geq -b] \sin(2\pi(b+\theta(w)))\, |\hat{f}(w)|\, db\, dw\right]$$

$$\iint \mathbb{1}[w^T x \geq b]\, q(b,w)\, db\, dw$$

Mass:

$$2 \cdot 2\pi \int \int_0^{\|w\|} |\sin(2\pi(b+\theta(w)))| \cdot |\hat{f}(w)|\, db\, dw$$

$$\leq 2 \cdot 2\pi \int \left(\int_0^{\|w\|} db\right) |\hat{f}(w)|\, dw$$

$$= 2 \cdot 2\pi \cdot \int \|w\| \cdot |\hat{f}(w)|\, dw$$

$$= 2 \cdot \int \|\widehat{\nabla f}(w)\|\, dw.$$

[Fact: $\|\widehat{\nabla f}(w)\| = 2\pi \|w\| \cdot |\hat{f}(w)|$]

Definition.

"Barron norm"

"Barron class"

$$\left\{ f : \int \|\widehat{\nabla f}(w)\|\, dw \leq C \right\}$$

Question: goal of not-worst-case representation (small weight mass) another in polar form?

No: (1) "mass" ends up as easily interpretable fourier quantity.
(2) get equality

Theorem. If $f, \hat{f} \in L_1$

$$f(x) - f(0) = -2\pi \left[\int_0^{\|w\|} \mathbb{1}[w^T x \geq b] \sin(2\pi(b+\theta(w)))\, |\hat{f}(w)|\, db\, dw - \int_{-\|w\|}^{0} \mathbb{1}[-w^T x \geq -b] \sin(2\pi(b+\theta(w)))\, |\hat{f}(w)|\, db\, dw\right]$$

weight mass $\leq 2 \int \|\widehat{\nabla f}(w)\|\, dw.$

Remarks.

* Some fourier mass estimates.
  * $\Omega(2^d)$ for many radial functions.
  * "almost flat" Gaussians have $\int \|\widehat{\nabla f}(w)\| = O(\sqrt{d})$
* Bound NTK representation size using Barron norms.