Lecture 1

[Room will change!]

Plan for today:

* Mathematical setup
* Course schedule & logistics
* Further setup, first theorems?

In practice, a deep network a set of functions $\mathcal{F}$, s.t.

[standard.]

① $\mathcal{F} : \{ x \mapsto F(x; w) : w \in \mathbb{R}^d \}$ for fixed parameter dim d.

② F has uniformly bounded runtime (in terms # elementary ops on real numbers

[less standard]

③ Efficient hardware

④ Convenient programming libraries

⑤ Amendable to GD.

$\hat{R}$ = "perf on seen data" | R = "perf on unseen data"

Examples

Ⓐ DALL-E input: english sentence | outputs $1024 \times 1024 \times 3$

$p \approx 2^{32}$ | # data $\hat{R} \approx 2^{30}$

Ⓑ Alphafold input: amino acid sequence $\mapsto$ 3-d structure

$\hat{R}$ & R hard to obtain, but well-defined

but "no common distribution"

Scope of this class: Not just simpler models, but specific focus:

① "feedforward networks", typically with 2-layers

② Goal is to show $R(\hat{f})$ small

output of alg

unseen performance

via a specific error decomposition

③ Short proofs & reusable tools

$f(x; w) \in \mathbb{R}^d$

④ Avoid looseness via adaptive complexity, esp. to the alg.

⑤ Bridge old & new.

tell me if bad

Defn. (Feed forward network)

$$x \mapsto \sigma_L(\cdots \sigma_2(W_2\, \sigma_1(W_1 x + b_1) + b_L) \cdots)$$

weights

biases

matrix

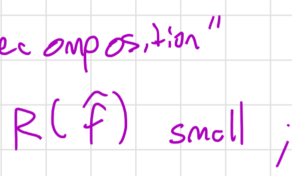
vector

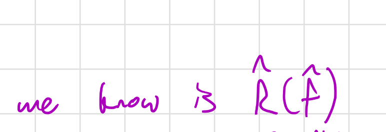
nonlinearity, transfer, activation

Esp. $x \mapsto a^T \sigma(Vx)$ (2-layer, no bias)

eg/ $\sigma_i : \mathbb{R}^{m_i} \times \mathbb{R}^{m_{i+1}}$ are coordinatewise scalar functions ($m_i = m_{i+1}$),

ReLU $z \mapsto \max\{0, z\}$, sigmoid $z \mapsto \frac{1}{1+e^{-z}}$





"Error decomposition"

Want: $R(\hat{f})$ small; all we know is $\hat{R}(\hat{f})$ smallish

helper: $\bar{f} \in \mathcal{F}$ "nearly best over R"

$$
\begin{aligned}
R(\hat{f}) = \ & R(\hat{f}) - \hat{R}(\hat{f}) && \text{"generalization"} \\
& + \hat{R}(\hat{f}) - \hat{R}(\bar{f}) && \text{"optimization"} \\
& + \hat{R}(\bar{f}) - R(\bar{f}) && \text{"generalization"} \\
& + R(\bar{f}) && \text{"approximation"}
\end{aligned}
$$

Remark "interpolation" $\hat{R}(\hat{f}) = 0 \ll R(\bar{f}) \approx R(\hat{f})$

* web page has everything
* places
  * in-person; OH outside (TR, 30min each)
  * zoom available, de-emphasized
  * gradescope
  * discussion edstem ("scheduled")
* Grading
  * 80% in 4 homeworks, typed, submit in gradescope, don't cheat
  * proj
* course notes

Apx
  shallow/constructive apx
  initialization & overparameterization
  non-shallow construction
Opt
  NTK smooth / strong convexity } 6 lectures
  (mean-field?)
  non-NTK margin analysis

?? generalization?
other topics