Ethical Algo Design in the Generative Era

Generative AI

- What is it?
- E.g. LLMs
- Next-word prediction
- Training & annotation
- History / precursors
- Embeddings
- Autoencoders/decoders

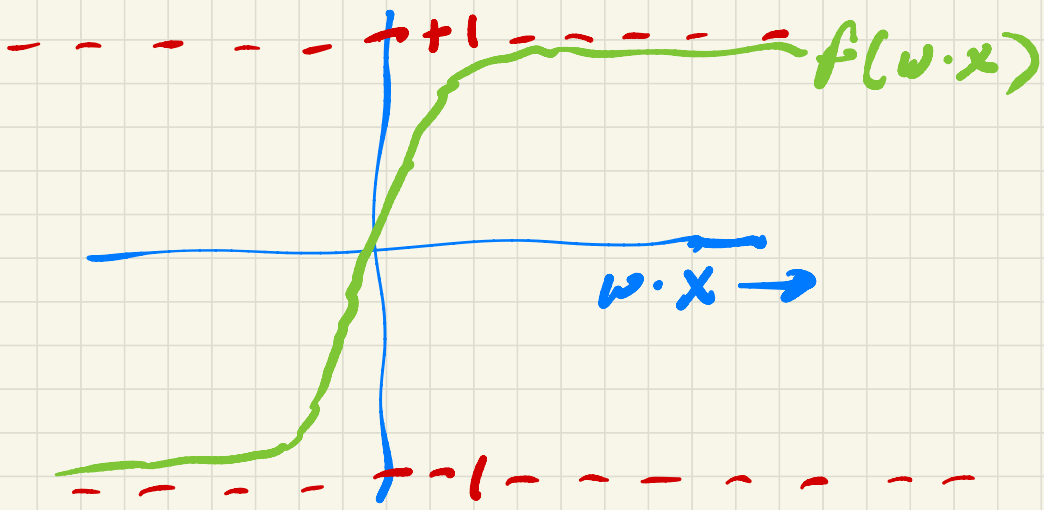
Neural Net Background

- single "neuron": on input $x \in \mathbb{R}^d$,

  output $= f(w \cdot x)$

  $= f(w_1 x_1 + w_2 x_2 + \cdots + w_d x_d)$

- typical choice for f: sigmoid

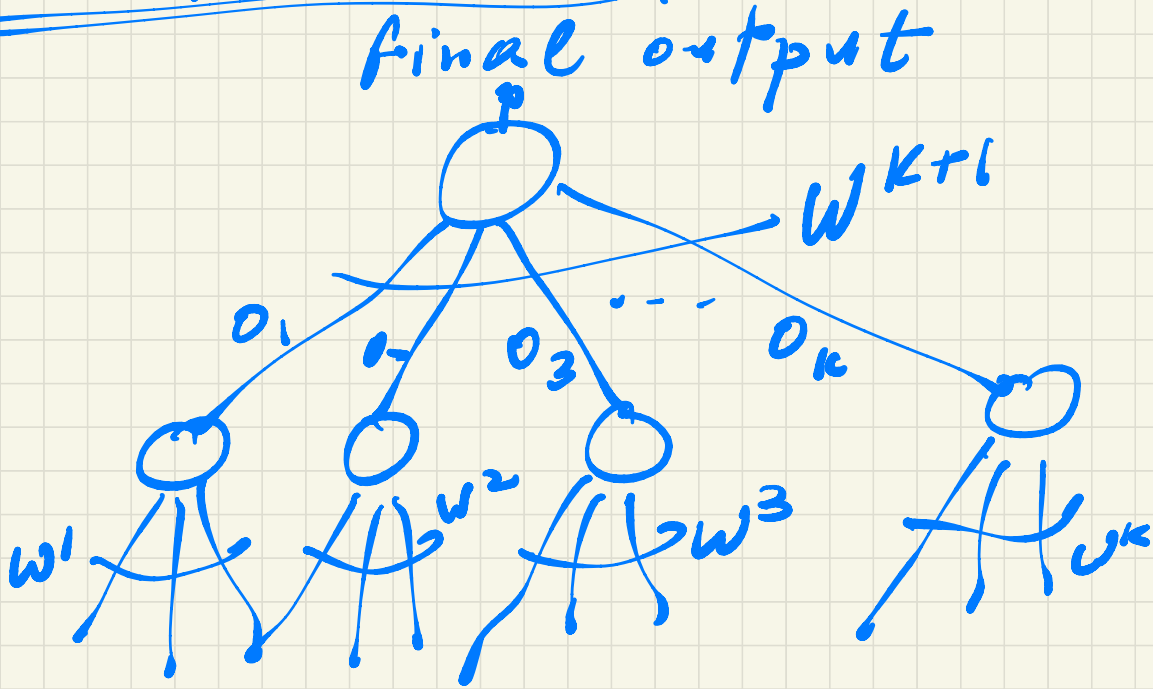


- training: weights w are parameters, loss fn is

$$\ell(w) = \sum_{\langle x,y \rangle \in S} (f(w \cdot x) - y)^2$$

differentiable

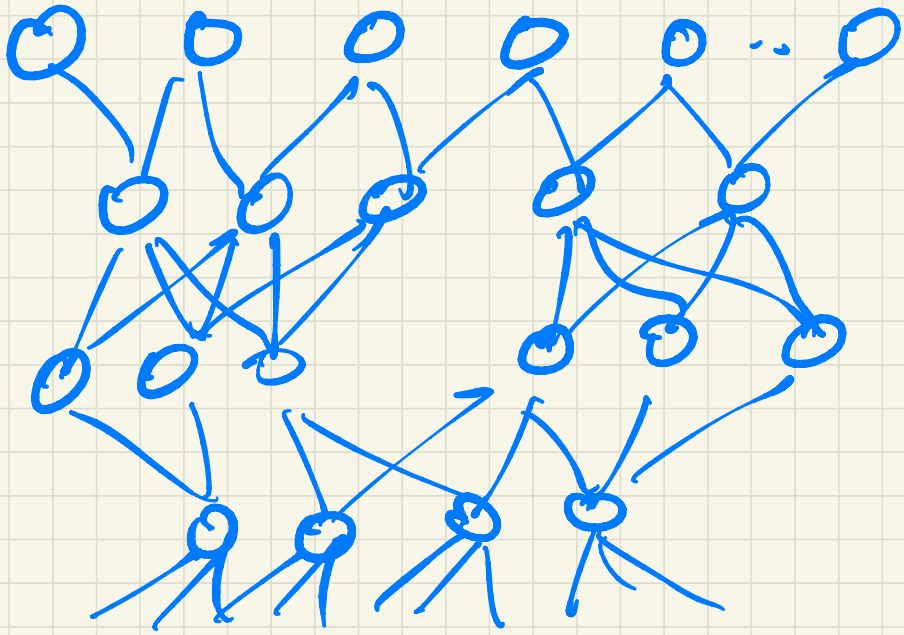
Multiple Layers



- params now $W^1, W^2, \ldots W^{K+1}$
- still differentiable (chain rule → backprop)

Key idea: intermediate neurons can learn new/better features

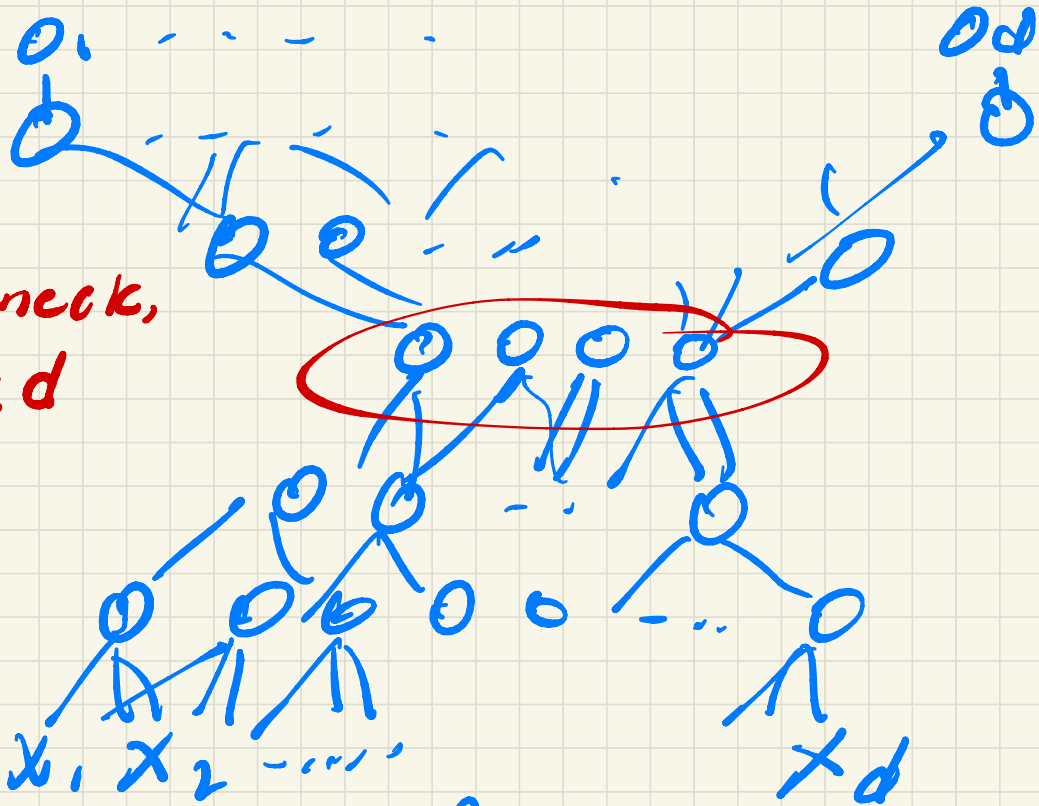
Multiple Outputs



$x_1, x_2 \cdots \cdots x_d$

- now outputs are high dimensional, still differentiable

Autoencoders



(image of a cat)

goal: minimize

$$\sum_{i=1}^{d} (o_i - x_i)^2$$

Fairness in GAI

- Compare e.g. group fairness in consumer lending
- Fairness in prompts?
- In tone?
- Biases vs. correlations, use cases

Toxicity

- Subjectivity & context
- Quotations & censorship
- Offensive opinions
- Nuance & indirection

Toxicity & Fairness

- Training data curation
- Must automate
  - → human annotation (unpleasant)
- Train "guardrail" models
- Apply to training, prompts & output

Privacy in GAI

- Regurgitation
- Virtual copying (e.g. code)
- Stylistic mimicry (more later)
- Could do DP model training, but impractical
- Robustness to membership inference attacks

Copyright & IP Concerns

- Stylistic mimicry
- Threat to creative livelihoods?

Copyright & IP

- Tech, policy & legal mechanisms
- Content attribution & compensation
- Machine Unlearning & model disgorgement
  - DP
  - Sharding

Adversarial Attacks

- E.g. image manipulation, prompt injection, data poisoning, spam filter attacks, ...
- Robust adversarial training:
  - instead of minimizing e.g.

$$\sum_{\langle x_i, y_i \rangle} (h(x_i) - y_i)^2$$

  - minimize something like

$$\max_{\delta : \|\delta\| < a} \sum_{\langle x_i, y_i \rangle} (h(x_i + \delta_i) - y_i)^2$$

Plagiarism & Cheating

- College essays, writing samples, etc.
- Tool vs. cheat?
- Author verification

Author Verification

- Train model to detect human/LLM (arms race)
- Redlist/greenlist (watermarking)

Hallucinations

- Plausible but verifiably wrong
- E.g. citations

Hallucinations

- User education
- Independent verification
- External sources (e.g. citation DBs)
- Disclaimers
- Content attribution

A bit more on hallucinations... (Kalai & Vempala, MM & MK, HW2)

- stylized facts, e.g. IMDB records - unambiguous true/false (also citations)
- schema or tuples: ⟨author$_1$..., title, year, journal⟩
- huge space of "factoids", only a fraction are facts
- training data contains only facts
- safest: only generate facts from training
- But in ML we want to generalize...

Missing Mass Estimation

- How many facts are missing in data?
- Good-Turing: let

$$\mathcal{L} = \frac{\#\text{facts occuring only once in training}}{n}$$

- Then

$$\left| \text{prob. next sample is a new fact} - \mathcal{L} \right| \leq \frac{1}{\sqrt{n}}$$

- Kalai-Vempala: for any model

$$\text{hallucination rate} \geq \mathcal{L} - \text{model "miscalibration"}$$

(one measure of overfitting)

Disruption to Work

- Passing bar exams, MBA courses, etc.
- IP concerns again

Job disruption/creation

- prompt engineer
- "English is the new programming language"

Strongest defense:

Use case specialization