Foundations of Trustworthy AI

Jonathan Ullman

CS7180/7880 · Fall 2026

Class 14: Necessity of Memorization

Oct 27, 2026

Label Memorization

Model

Distribution D over (X, Y) — binary labels

Data $S = \{(X_1, Y_1) \dots (X_n, Y_n)\} \sim D^n$

Algorithm A, $A(S) \to \hat{\theta}$ model $f_{\hat{\theta}}$

Definition:

Fix a sample S and index $i \in \{1, \dots, n\}$

$$\text{mem}(A, S, i) := \mathbb{P}_{\hat{\theta} \sim A(S)} [f_{\hat{\theta}}(X_i) = Y_i] - \mathbb{P}_{\hat{\theta} \sim A(S^{-i})} [f_{\hat{\theta}}(X_i) = Y_i]$$

memorization score

S^{-i}: S but with (X_i, Y_i) removed

Theorem (Informal Version):

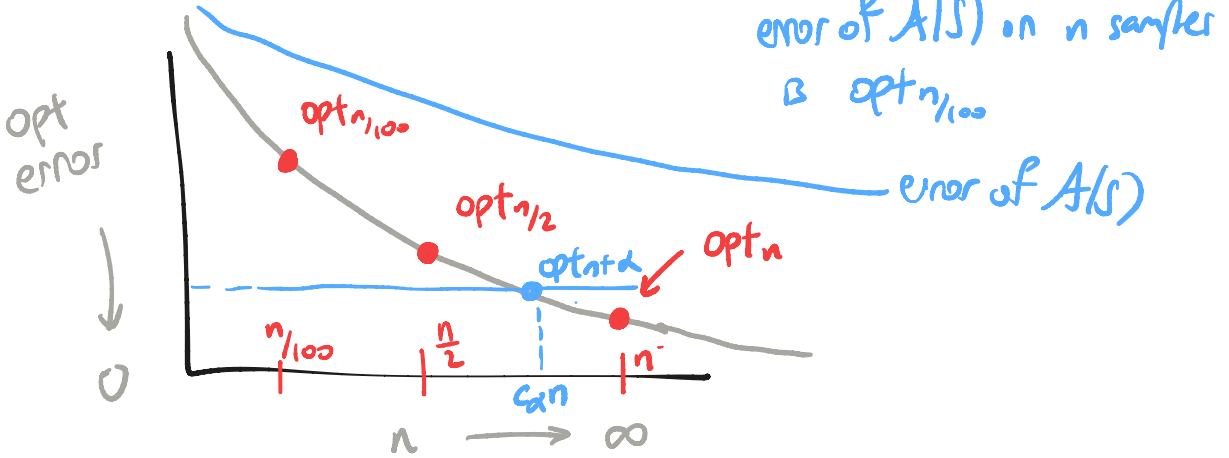
There is a family of learning problems such that any algorithm A with error sufficiently close to optimal must have

$$\mathbb{E}\left[\sum_{i=1}^{n} \text{mem}(A, S, i) \geq c \cdot n\right]$$

choice of true labels, choice of S from D

constant fraction of labels memorized

"Sufficiently Close to Optimal"



A(S): Let S' be 1/100 sample from S

Output A'(S') where A' is some opt algorithm

Long Tails

Distribution on Features

$\mathcal{X} = \{0, 1, \ldots, n\}$ n+1 "types"

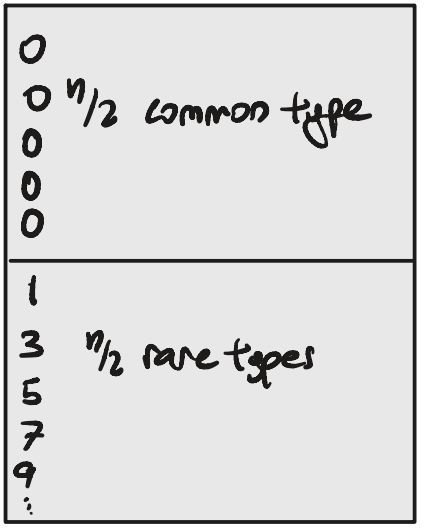
Define D_X as follows:

$\mathbb{P}[X = 0] = 1/2$ common type

$x \neq 0$ $\mathbb{P}[X = x] = 1/2n$ rare types

S

n examples



$S_{sing} = \{ x \neq 0 \text{ that appear exactly once in } S \}$

$\mathbb{E}[|S_{sing}|] = \frac{1}{\sqrt{e}}$

Long Tails

Distribution on Features

$\mathcal{X} = \{0, 1, \ldots, n\}$ $n+1$ "types"

Define D_X as follows:

$\mathbb{P}[X = 0] = 1/2$ common type

$x \neq 0$ $\mathbb{P}[X = x] = 1/2n$ rare types

S

n examples	labels
0 0 0 0 0 $n/2$ common type	1 1 1 1 1
1 3 5 7 9 : $n/2$ rare types	0 1 0 0 1

$S_{sing} = \{x \neq 0 \text{ that appear exactly once in } S\}$

$\mathbb{E}[|S_{sing}|] = \frac{1}{\sqrt{e}}$

Hey Claim: The optimal learner memorizes all the singleton labels

Distribution on Labels

Learning problem D_{f^*}:

$X \sim D_X$

$Y|X=x \sim f^*(x)$

$f^*: \mathcal{X} \to \{0,1\}$

Distribution over learning problems Q:

choose f^* randomly

Analyzing Memorization

true label is random so you have to guess

$$\text{opt}_n = \left(\text{expected fraction of types you've never seen}\right) \cdot \frac{1}{2}$$

$$= \left(2^{-n} + \left(1 - \frac{1}{2n}\right)^n\right) \cdot \frac{1}{2}$$

$$\overset{n \to \infty}{=} \frac{1}{\sqrt{e}} \cdot \frac{1}{2}$$

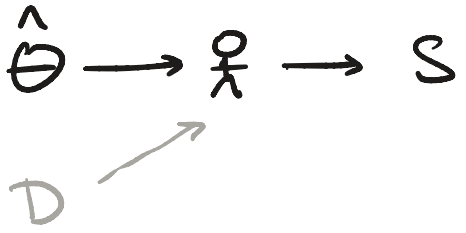
$$\mathbb{E}[|S_{sing}|] = \frac{n}{2}\left(1 - \frac{1}{2n}\right)^{n-1} \approx \frac{n}{2\sqrt{e}}$$

For any singleton $i \in S_{sing}$ $\mathbb{E}_{f^\star}[\text{mem}(A, S, i)] = \mathbb{P}_{f^\star, A(S)}[f_{\hat{\theta}}(x_i) = y_i] - \frac{1}{2}$

accuracy on singletons directly proportional to memorization of S_{sing}

Beyond label Memorization

- Want to study models that are "close" to extraction in LLMs
- In label memorization, everything we memorize is "relevant" to D



We're going to prove that (in some cases) the model must reveal information about the training data that is <u>not</u> relevant to D

Information Theory 101

Ex. Choose $x \in \{0,1\}^d$ uniformly $d = 1000$

x has 1000 random bits of uncertainty

probability of guessing x is 2^{-1000}

Suppose I tell you $x_1 \dots x_{950}$ (95% of the bits)

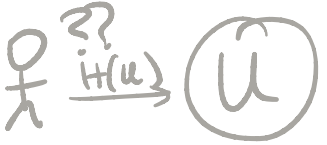
$x \mid x_1 \dots x_{950}$ has 50 random bits of uncertainty

probability of guessing x is 2^{-50}

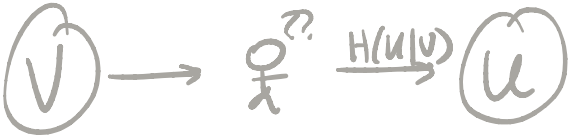
Information Theory: Measuring uncertainty in the units of coin tosses

Information Theory 101

Conditional Entropy: $H(U|V) = \mathop{\mathbb{E}}_{v \sim V}[H(U|V=v)]$



Uncertainty from the perspective of someone who knows V



Mutual Information: $I(U;V) = H(U) - H(U|V)$

How much does knowing V decrease uncertainty about U

Information Theory 101

Entropy: For a discrete random variable U

$$p_U(u) = \mathbb{P}[U = u]$$

$$H(U) = \sum_u p_U(u) \cdot \log_2\left(1/p_U(u)\right)$$

$$= \mathop{\mathbb{E}}_{u \sim U}\left[\log_2\left(1/p_U(u)\right)\right]$$

Ex. U = uniform string of 1000 bits

$$H(U) = \sum_{u \in \{0,1\}^{1000}} 2^{-1000} \cdot \log_2\left(2^{1000}\right) = 2^{1000} \cdot 2^{-1000} \cdot 1000$$

$$= 1000 = H(U)$$