

# Watermarking

Plan

Review

Motivation

Red/green lists

Distortion-free

Exponential Minimum Sampling

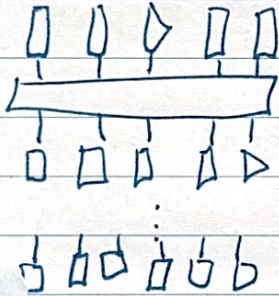
Logistics

Gone today :-

Zoom!

## Review

Mid test is horse-drawn



$W = W^{orig} + BA$

← freeze ← train

## Challenge 1: Finetune

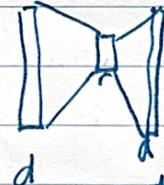


$W \in \mathbb{R}^{d \times d}$

Complexity:  $O(d^2)$

$$\sum_{i=1}^d v_i v_i^T$$

vs



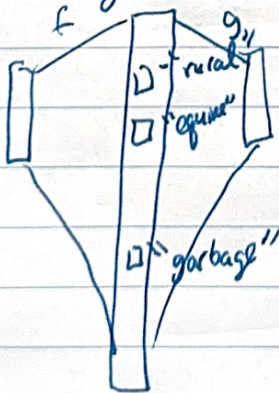
$A \in \mathbb{R}^{d \times d}$

$B \in \mathbb{R}^{d \times d}$

$O(dn)$

$$\sum_{i=1}^d u_i u_i^T$$

## Challenge 2: Understand



$$\mathcal{L}(w) = \|f(x) - x\|_2^2 + \lambda H(f(x), \frac{1}{N})$$






Motivation:

↳ check use e.g., human written?

↳ model owner control e.g., model generated?

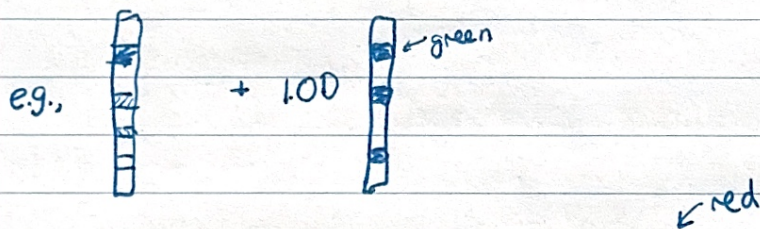
"Turmeric lemon cookies are \_\_\_\_\_"

tasty →   
yellow →   
gross → 

Activity: How can we watermark text?

Red/green Lists

new-logits = logits + c ·  $\mathbb{I}[\text{green}]$  then next ~ softmax (new-logits)



Hi! I am a human for sure. No doubt about it at all. Etes my goats.

num-green

$$\Pr(\text{num-green or more in } m \text{ trials w/ } 1/2)$$
$$= \sum_{k=\text{num-green}}^n \frac{1}{2^n} \binom{n}{k}$$



Motivation: Watermark *without* changing distribution

Attempt #1: "Vermont kale in the winter! —"

convert to number  
and use as seed

Then  $\text{next} \sim p = \text{softmax}(\log \text{its})$  when randomness seeded

But how do we detect??

Approach #2: Sample  $\text{next} \sim p$  with correlated RV

$$\text{next} = \arg \min_i \frac{-\log(x_i)}{p_i} \quad \text{where } x \sim \text{Unif}([0,1]^d)$$

$$\Pr\left(\frac{-\log x_i}{p_i} \geq t\right) = \Pr(x_i \leq \exp(-p_i t)) = \exp(-p_i t)$$

$$\Rightarrow \Pr\left(\frac{-\log x_i}{p_i} = u\right) = p_i \exp(-p_i u)$$

$$\begin{aligned} & \Pr\left(i^* = \arg \min_i \frac{-\log x_i}{p_i}, \frac{-\log x_{i^*}}{p_{i^*}} \geq t\right) \\ &= \int_{u \geq t} \underbrace{\Pr\left(\frac{-\log x_{i^*}}{p_{i^*}} = u\right)}_{p_{i^*} \exp(-p_{i^*} u)} \prod_{j \neq i^*} \underbrace{\Pr\left(\frac{-\log x_j}{p_j} > t\right)}_{\exp(-p_j u)} \\ &= p_{i^*} \int_{u \geq t} \exp(-u) = p_{i^*} \cdot [-\exp(-u)]_{u=t}^{\infty} = p_{i^*} \exp(-t) \end{aligned}$$

$$\Pr\left(i^* = \arg \min_i \frac{-\log x_i}{p_i}\right) = \sum_{t=0}^{\infty} p_{i^*} \exp(-t) = p_{i^*}$$

seed x



## Exponential Minimum Sampling (continued)

detect:

record

$$\text{cost} = \frac{1}{\text{len}(\text{tokens})} \sum_{i=1}^{\text{len}(\text{tokens})} -\log(X_i[\text{tokens}[i]])$$