

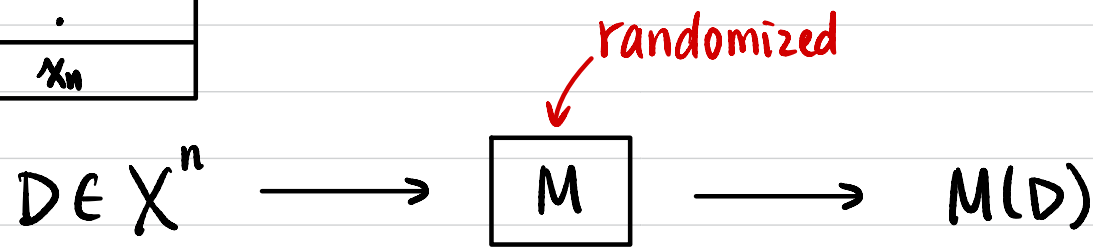
CSCI 1515 Applied Cryptography

This Lecture:

- Differential Privacy (Continued)
- AI Watermarking
- Example: Green-Red Watermark for LLM

Differential Privacy

x_1
x_2
\vdots
x_n

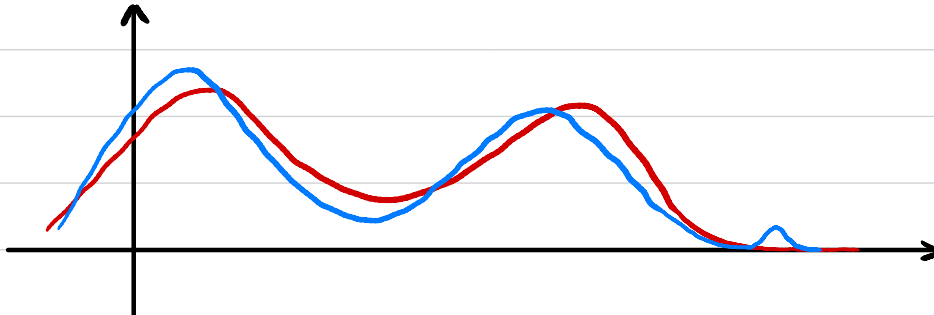


Def (ϵ, δ) - Differential Privacy for a randomized mechanism:

\forall neighboring datasets D_1 & D_2 (differing in one row),

$\forall T \subseteq \text{range}(M)$,

$$\Pr[M(D_1) \in T] \leq e^\epsilon \cdot \Pr[M(D_2) \in T] + \delta$$



Laplace Mechanism

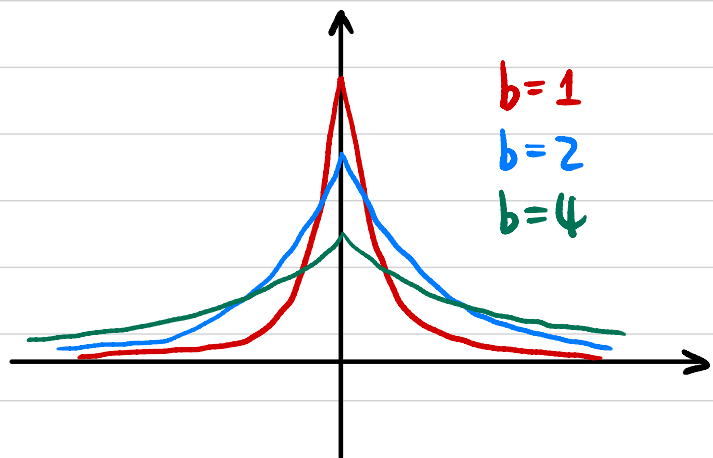
Def Sensitivity of a function $f: X^n \rightarrow \mathbb{R}$

$$\Delta f := \max_{D_1 \sim D_2} |f(D_1) - f(D_2)|$$

Laplace Mechanism: $M(D) = f(D) + \text{Lap}(\Delta f / \epsilon)$

Thm The Laplace Mechanism is ϵ -DP.

Laplace distribution:



probability distribution function

$$\text{PDF}(x) = \frac{1}{2b} \cdot \exp\left(-\frac{|x|}{b}\right)$$

For $X \sim \text{Lap}(b)$, $\Pr[|X| \geq bt] \leq \exp(-t)$

Is a bigger b better for privacy, or worse?

Better

Composition Theorems

Thm (post-processing) If $M: X^n \rightarrow Y$ is (ϵ, δ) -DP,

$f: Y \rightarrow Z$ is an arbitrary randomized function,

then $f \circ M: X^n \rightarrow Z$ is also (ϵ, δ) -DP.

Thm (group privacy) If $M: X^n \rightarrow Y$ is $(\epsilon, 0)$ -DP,

then M is $(k \cdot \epsilon, 0)$ -DP for groups of size k .

Thm (composition) If $M_i: X^n \rightarrow Y$ is (ϵ_i, δ_i) -DP $\forall i \in [k]$,

then $M(D) := (M_1(D), \dots, M_k(D))$ is $(\sum_{i \in [k]} \epsilon_i, \sum_{i \in [k]} \delta_i)$ -DP.

AI Watermarking

How to detect AI-generated content?

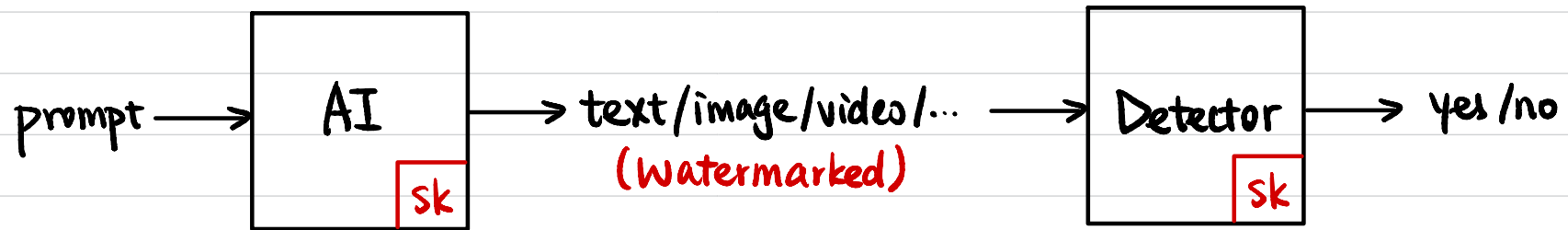


Watermarking:

Insert a statistical signal into AI-generated content.

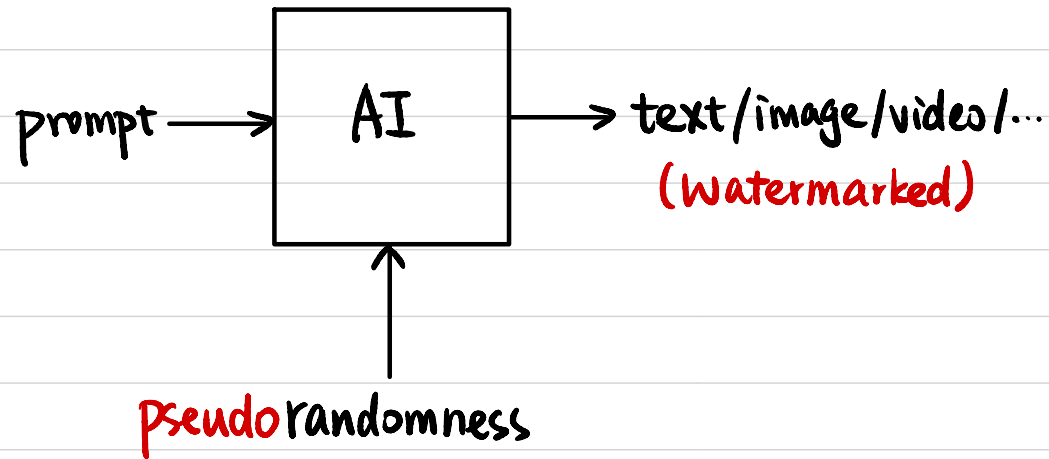
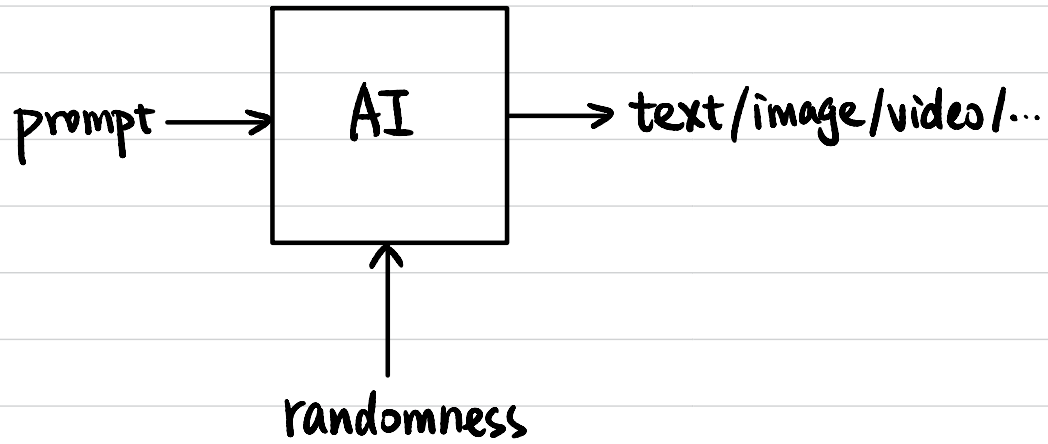
What properties do we want from AI watermarking?

AI Watermarking



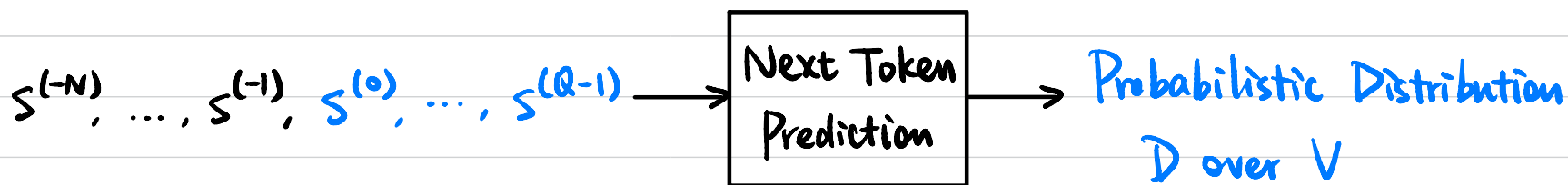
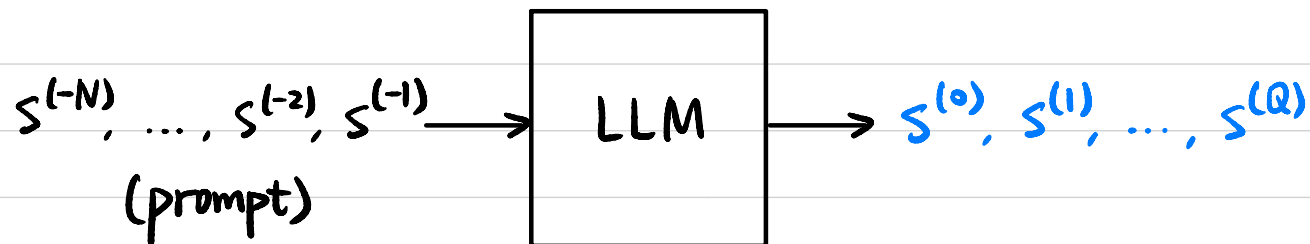
- **Low False Negatives / Robustness** (to local perturbations):
AI-generated content (even with small changes) will be detected
- **Low False Positives**:
Non-AI-generated content won't be flagged
- **Indistinguishability / Undetectability / Quality**:
Watermarking doesn't hurt quality of AI-generated content.
- **Unforgeability**:
Watermarked content cannot be generated without knowledge of sk.

High-Level Idea



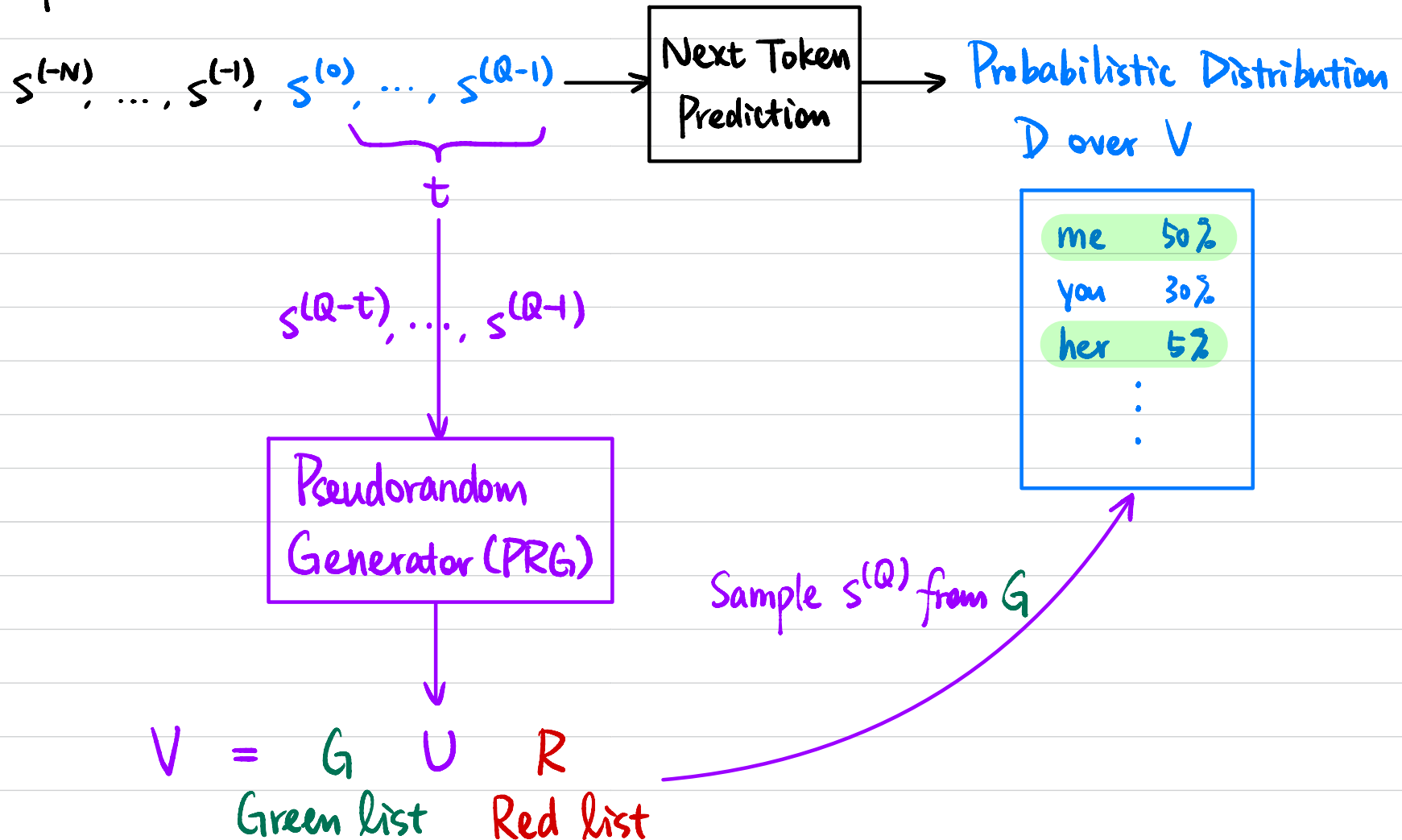
Language Model Basics

Vocabulary V : words / word fragments, "tokens"



me	50%
you	30%
her	5%
⋮	
⋮	

Example: Green-Red Watermark



How to detect watermark?

Properties?

Any issues?

Other Variants

- "Soft" watermark: for low-entropy text (role of entropy)

Green list more likely to be sampled

- Watermark strength vs. text quality

$$|G| = \gamma \cdot |V| \quad |R| = (1 - \gamma) \cdot |V|$$

- Watermark strength vs. number of tokens

- Private watermark: use sk to watermark/detect

Pseudorandom Function (PRF) $F_k(s^{(Q-t)}, \dots, s^{(Q-1)})$

Attacks ?

Attacks

- Emoji attack: write an essay but inserting an emoji after every word.

- Translation / Paraphrasing attacks

↓
Impossibility of "strong" watermarking

Ultimately: use watermark at a "semantic" level?

THANK YOU 😊