SSL Weekly Presentation

ANM Zahid Milkan
200041202

Amit Bin Tariqul
200041214

Sahab al Chowdhury
200041255

Supervisors

Dr. Kamrul Hasan
Professor

Dr. Hasan Mahmud
Professor

Syed Rifat Ryan
Lecturer

Systems and Software Lab (SSL)
Dept. of Computer Science & Engineering
Islamic University of Technology

February 26, 2025

Introduction

- The EXP watermarking algorithm embeds signals in text generation.
- Detection relies on statistical properties of token selection.
- Key idea: transformed random numbers follow an exponential distribution.

Why Use an Exponential Distribution?

- Random numbers from $U(0,1)$ are transformed as:

$$X_i = -\log(1 - r_i)$$

- Ensures $X_i \sim \text{Exponential}(1)$.
- Sum of transformed values follows a Gamma distribution:

$$S = \sum_{i=1}^{k} X_i \sim \Gamma(k, 1)$$

- Enables detection using a statistical test.

Transformation to Exponential(1)

Given: $R \sim U(0,1)$

- Transformation: $X = -\log(1-R)$
- Compute CDF:

$$F_X(x) = P(X \leq x) = P(-\log(1-R) \leq x)$$
$$= P(R \leq 1 - e^{-x}) = 1 - e^{-x}, \quad x \geq 0.$$

- Differentiate to get PDF:

$$f_X(x) = \frac{d}{dx}(1 - e^{-x}) = e^{-x}, \quad x \geq 0.$$

- This matches the PDF of Exponential(1), confirming the transformation.

Sum of Exponential Distributions is Gamma

Given: $X_1, X_2, \ldots, X_k \sim \text{Exponential}(\lambda)$ independently.

- Define the sum: $Y = X_1 + X_2 + \cdots + X_k$.
- Moment-Generating Function (MGF):

$$M_X(t) = \mathbb{E}[e^{tX}] = \int_0^\infty e^{tx} \lambda e^{-\lambda x} dx = \lambda \int_0^\infty e^{(t-\lambda)x} dx.$$

- Evaluating:

$$M_X(t) = \frac{\lambda}{\lambda - t}, \quad t < \lambda; \quad M_Y(t) = \left(\frac{\lambda}{\lambda - t}\right)^k.$$

- This matches the MGF of Gamma(k, λ):

$$f_Y(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}, \quad y \geq 0.$$

- Conclusion: $Y \sim \text{Gamma}(k, \lambda)$.

Why Use $u^{(1/\text{probs})}$ in Sampling?

- Token selection formula:

$$\text{argmax}\left(u^{(1/p)}\right)$$

- Ensures higher probability tokens are exponentially favored.
- Prevents low-probability tokens from dominating.
- Embeds a statistical pattern that can be detected later.

Detection Process

- Compute transformed values:

$$X_i = -\log(1 - r_i)$$

- Compute total score:

$$S = \sum_{i=1}^{k} X_i$$

- Compare against $\Gamma(k, 1)$ distribution.
- Compute p-value:

$$p\text{-value} = P(S_{\text{null}} > S_{\text{observed}})$$

- If p-value $<$ threshold, watermark detected.

Conclusion

- EXP watermarking modifies token probabilities in a detectable way.
- Detection relies on transformed random numbers following a Gamma distribution.
- Low p-values indicate watermark presence.