Aalto University
School of Science

Online Entropy Estimator

Andrea Röck

Aalto University, School of Science
Department of Information and Computer Science

Crypto research seminar, April 18, 2011

Outline

Introduction

Estimator for Unknown Entropy Sources

New Estimator

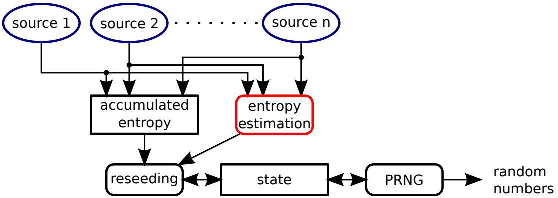
Empirical Results

Conclusion

Introduction

PRNG with Entropy Input

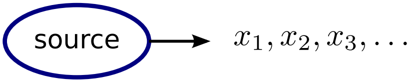
- **Pseudo Random Number Generator with entropy sources**



- **PRNG is deterministic**: A specific state will always produce the same output
- Uncertainty come from **entropy sources**
- When **"enough uncertainty is collected"** reseed the state
  - **How do we know when to reseed?**

Entropy - Information Theoretical Model

- **Entropy Source**



- Outputs element from **sample space** $\mathcal{X}$
- **Concrete output** at **time** t: $x_t \in \mathcal{X}$
- Source represents sequence of identical and independent distributed **random variables** $X_1, X_2, \ldots$

$$\Pr(X_t = \eta) = p_\eta \text{ for all } t,\ p = (p_\eta)_{\eta \in \mathcal{X}}$$

- **Sequence** of **random variables** $\mathbf{X}_{[t_1,t_2]} = X_{t_1}, X_{t_1+1}, \ldots, X_{t_2}$
- **Sequence** of **concrete outputs** $\mathbf{x}_{[t_1,t_2]} = x_{t_1}, x_{t_1+1}, \ldots, x_{t_2}$
- **Empirical distribution** of $\mathbf{x}_{[1,n]}$: $\hat{p}_\eta = \frac{\#\{t : x_t = \eta\}}{n}$

Entropy - Definitions

- **Shannon entropy**: Measure of average number of binary questions before guessing the output
$$H(p) = -\sum_{\eta \in \mathcal{X}} p_\eta \log_2 p_\eta$$
- **Rényi entropy**: Measure of correlation probability
$$H_\alpha(p) = \frac{1}{1-\alpha} \log_2 \left(\sum_{\eta \in \mathcal{X}} p_\eta^\alpha \right)$$
- **Min entropy**: Lower bound for all entropy measures
$$H_\infty(p) = -\log_2 \left(\max_{\eta \in \mathcal{X}} p_\eta \right)$$
- **Relation** for $\alpha > 1$:
$$H_\infty(p) \leq H_\alpha(p) \leq H(p) \leq \log_2 |\mathcal{X}|$$

Estimator for Unknown Entropy Sources

Requirements

- Estimator $\hat{H}$ should:
  - Work with **unknown sources**
  - Be **pessimistic** $E(\hat{H}) \leq H(p)$
  - Be **efficient**
  - Given an estimate for **each output**
    - Want $\hat{H}$ such that $\frac{1}{n-r} \sum_{t=r+1}^{n} \hat{H}(\mathbf{X}_{[t-r,t]}) \xrightarrow{n \to \infty} H(p)$
  - Work with **any source**

Known Estimators (1)

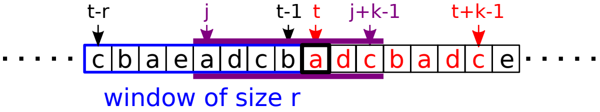
- **Plug-in** or **maximum-likelihood** estimator: use $H(\hat{p})$
  - Estimate only for **whole data set**
  - Need counter for **each** $\eta \in \mathcal{X}$
  - Can be applied on a data window but still not efficient

- **Compression** based on **frequency counting**:
  e.g. Huffman coding, arithmetic coding
  - **Not efficient** (Huffman coding need tree of size $|\mathcal{X}|$)

Known Estimators (2)

- **Compression** based on **match length** [Lempel Ziv]

$$L_t^r(\mathbf{x}) = 1 + \max_{t-r \leq j \leq t-1} \{k : \mathbf{x}_{[t,t+k-1]} = \mathbf{x}_{[j,j+k-1]}\}$$



- **Classical estimator**:

$$\frac{L_t^r(\mathbf{X})}{\log_2 r} \to \frac{1}{H} \text{ a.s. } (r \to \infty)$$

  - Estimate only for **whole data set**
  - Need **future values**, no upper bound for $L_t^r(\mathbf{x})$

Known Estimators (3)

- **LZ estimator** with intermediate values, $\hat{H}^r_{\mathrm{LZ}}(\mathbf{x}_{[t,n]}) = \frac{\log_2 r}{L^r_t(\mathbf{x})}$

$$\frac{1}{n-r} \sum_{t=r+1}^{n} \hat{H}^r_{\mathrm{LZ}}(\mathbf{x}_{[t,n]}) \to H \text{ a.s. } (n, r \to \infty)$$

  - Gives an estimate for **each** $t \geq r$
  - Need **future values**, no upper bound for $L^r_t(\mathbf{x})$

Known Estimators (4)

- Based on **transition frequencies** [Bucci Luzzi 2005]
- **Count transitions** from 0 to 1 or from 1 to 0
- **Expected number of transition** in $n+1$ bits: $n2p_0(1-p_0)$
- Use: $-log_2 y \geq \frac{1}{\ln(2)}(1-y)$ for $0 < y < 1$
- **Entropy**:

$$-p_0 \log_2 p_0 - (1-p_0)\log_2(1-p_0) \geq \frac{1}{\ln(2)} 2p_0(1-p_0)$$

  - Only **binary sources**

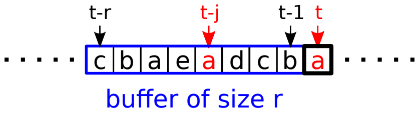
- **Our idea**: Extend to **non-binary case**

New Estimator

Idea

- **Number of comparisons** before finding **last occurrence** of **current element**:

$$\ell_t^r = \begin{cases} r & \text{if } x_t \neq x_{t-j}, 1 \leq j \leq r, \\ \min\{0 \leq j \leq r-1 : x_{t-1-t} = x_t\} & \text{otherwise.} \end{cases}$$



- **Estimator:** $\hat{H}^r_{\mathrm{pv}}(\mathbf{x}_{[t-r,t]}) = \frac{1}{\ln(2)} \sum_{j=1}^{\ell_t^r} \frac{1}{j}$

Theory

- **Expected value**: $\mathrm{E}\left(\hat{H}^r_{\mathrm{pv}}(\mathbf{X}_{[t-r,t]})\right) = \frac{1}{\ln(2)} \sum_{\eta\in\mathcal{X}} p_\eta \sum_{i=1}^{r} \frac{(1-p_\eta)^i}{i}$
- Using results on **(r+1)-dependent random variables**:

$$\frac{1}{n-r}\sum_{t=r+1}^{n} \hat{H}^r_{\mathrm{pv}}(\mathbf{x}_{[t,n]}) \to \mathrm{E}\left(\hat{H}^r_{\mathrm{pv}}(\mathbf{X}_{[t-r,t]})\right) \text{ a.s. } (n\to\infty)$$

- Taylor series of logarithm ($0 < x < 1$):

$$\ln\frac{1}{x} = \sum_{i=1}^{\infty}\frac{(1-x)^i}{i} \geq \sum_{i=1}^{r}\frac{(1-x)^i}{i}$$

- **Lower bound for entropy**:

$$\mathrm{E}\left(\hat{H}^r_{\mathrm{pv}}(\mathbf{X}_{[t-r,t]})\right) \leq \sum_{\eta\in\mathcal{X}} p_\eta \log_2\frac{1}{p_\eta} = H(p)$$

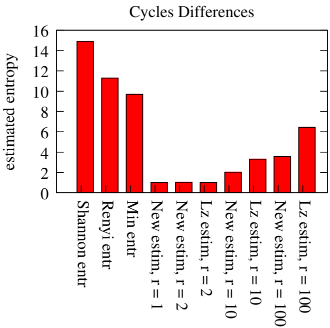
Empirical Results

Test Data

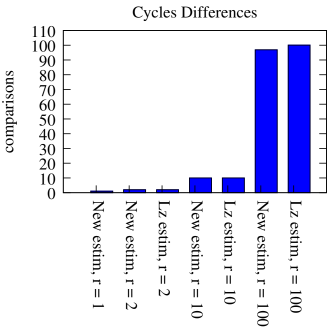
- Input data for **Linux random number generator**
  - Cooperation with Lacharme, Strubel, Videau
  - Time between interrupts from **user input (mouse, keyboard)**
  - Time measured in **jiffies** and **cycles**
- Data from **iPhone GPS** device
  - Cooperation with Lauradoux, Ponge
  - Measurement of altitude, longitude, latitude, acceleration and compass (heading)
  - **Indoor** (less movement) and **outdoor** (more movement) measurements
- Compare to **LZ estimator**

Linux RNG - Cycle Count Difference

Entropy



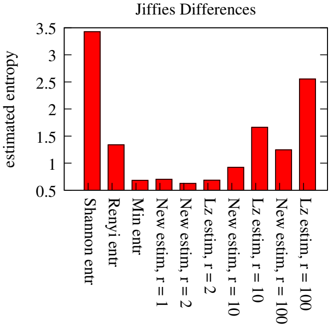
Number of comparisons



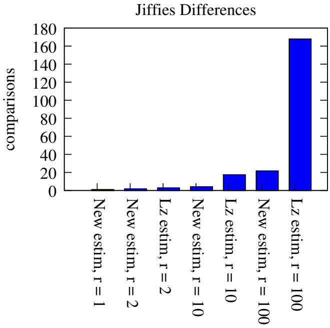
- **LZ estimator** has **complexity as good as our** estimator

Linux RNG - Jiffies Count Difference

Entropy



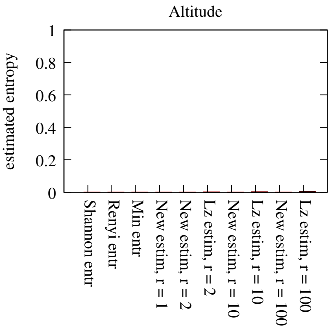
Number of comparisons



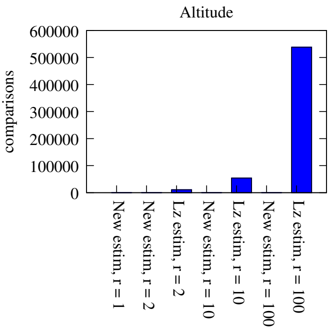
- **Better complexity** than LZ estimator

iPhone - Altitude - Indoor

Entropy



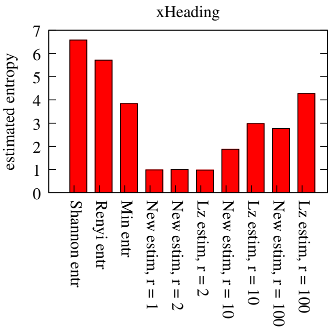
Number of comparisons



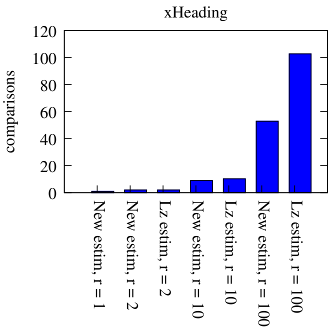
- **Extreme case for LZ estimator** if source has (almost) **no entropy**

iPhone - xHeading - Outdoor

Entropy



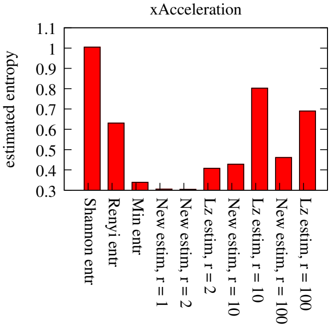
Number of comparisons



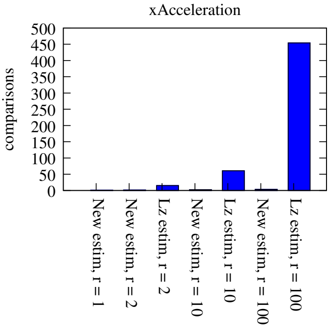
- Estimated entropy of new estimator **not far from LZ estimate**

iPhone - xAcceleration - Indoor

Entropy



Number of comparisons



- **Better complexity** than LZ estimator

Conclusion

Conclusion

- **Entropy estimator** for **unknown** entropy sources with **changing behavior**
- **At most r comparisons**
- For **independent sequences** the **expected value** of the estimator is **lower bound** for entropy
- Gives estimate for **each output value**

- **Only Shannon** entropy, not Rényi or Min entropy
- Proof **not** for **correlated sources**
- Estimates **lower bound**, exact value only for $r \to \infty$