Analysis and Improvement of Entropy Estimators in NIST SP 800-90B for Non-IID Entropy Sources

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

Speaker: Shuangyi Zhu

Outline

- Background
  - Entropy source and min-entropy
  - NIST SP 800-90B
- Contents and contributions
  - Problem of statistic-based estimators
  - A new estimator specific to Markov process
  - Some experiments
- Conclusion

Background

- Random numbers are essential to cryptographic applications (e.g. key generation, initialization vectors, nonce and so on) .

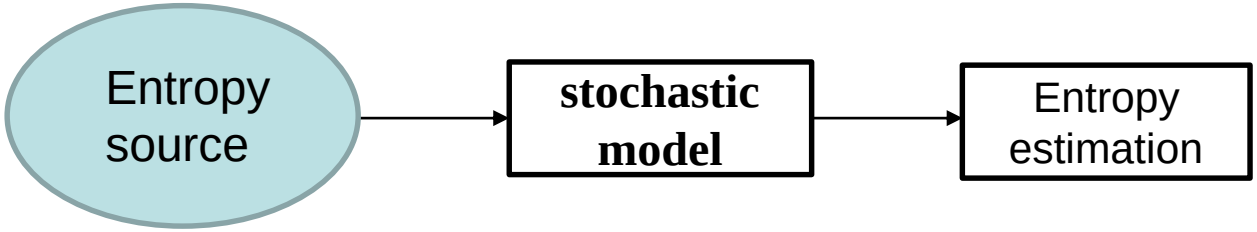
- An entropy source generates true random numbers on the basis of some physical phenomena.

- Entropy is employed to assess and quantify the quantity of randomness.

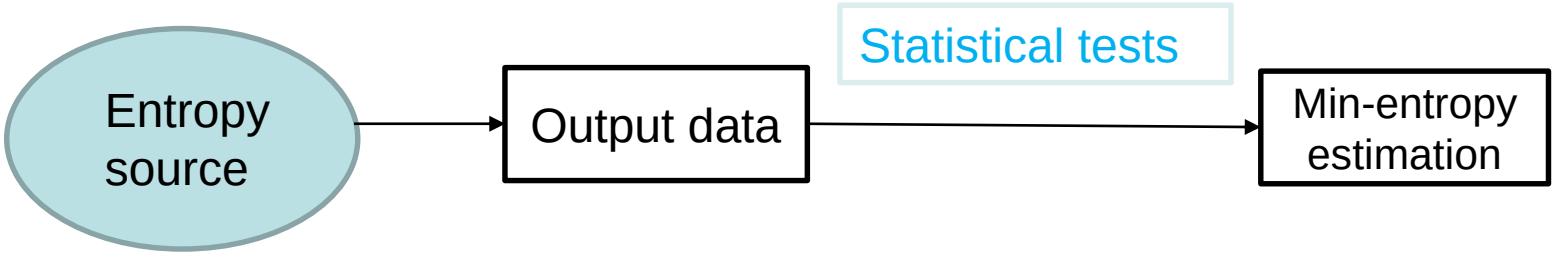
- There are many types of entropy, while the min-entropy is used in the 90B. It corresponds to the difficulty of guessing the most likely output of the entropy source.

How to estimate the entropy of an entropy source

- In AIS 31, developed by German BSI (Federal Office for Information Security)



- In NIST SP 800-90B,

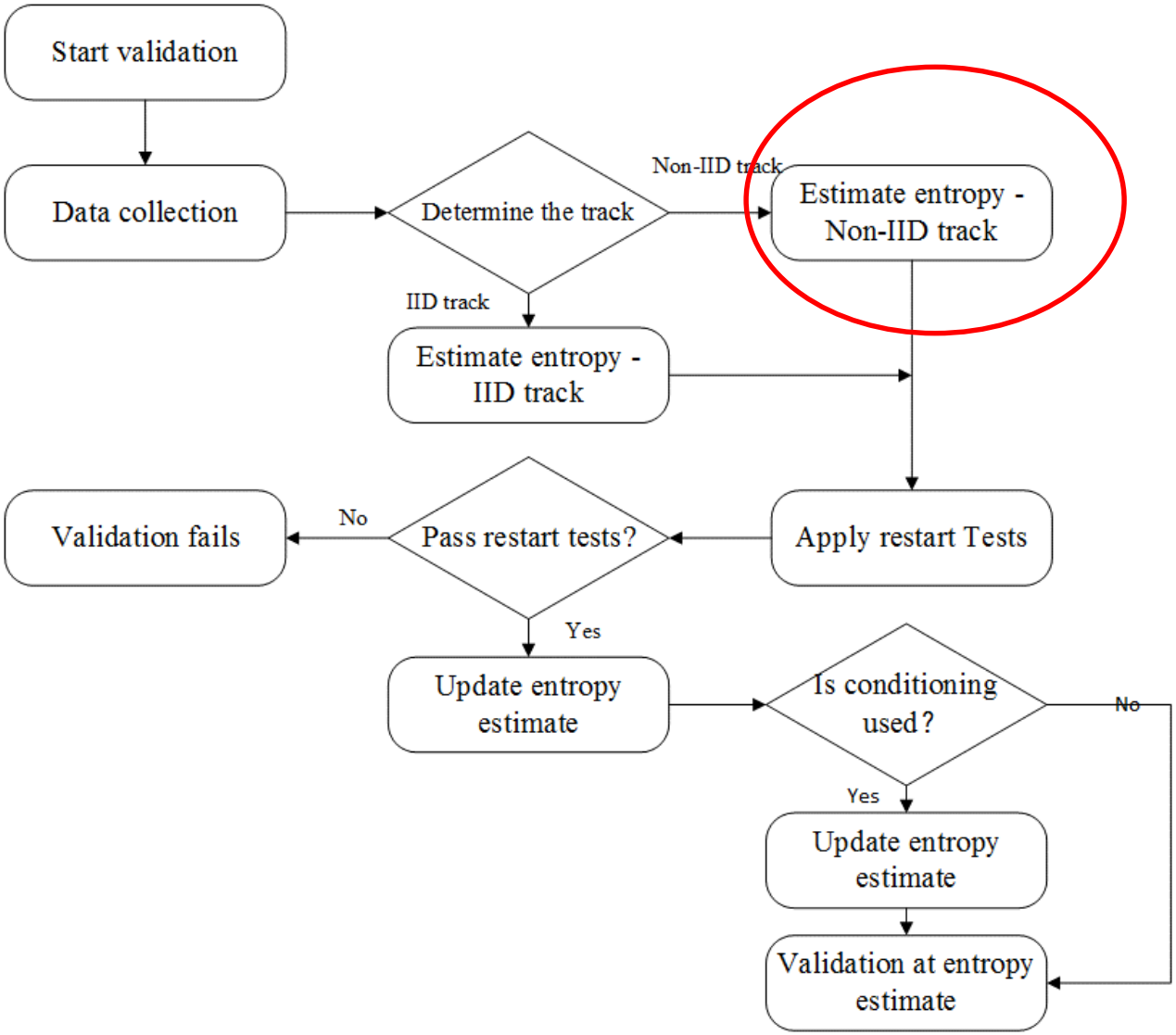


Brief Introduction of SP 800-90B

Version of 90B	Date
2nd draft	January 2016
Current version	January 2018

- When we studied the 90B, there is only the 2nd draft version of 90B, while the standard version is published in January 2018.
- The problem we have found in the old version is still existing in the current version.

SP 800-90B Test procedure



Brief Introduction of SP 800-90B

- In IID track, the entropy is easy to estimate. Only one simple estimator is used to estimate the min-entropy of the entropy source.
- In non-IID track, 10 estimators are employed to estimate the min-entropy. Finally the minimum one among ten estimates is selected as the min-entropy of the entropy source.

Basic type	Statistic based type	Prediction based type
Most Common Value Estimate	Collision Estimate	MultiMCW Prediction Estimate
t-Tuple Estimate	Compression Estimate	Lag Prediction Estimate
LRS Estimate		MultiMMC Prediction Estimate
Markov Estimate		LZ78Y Prediction Estimate

Our work on SP 800-90B

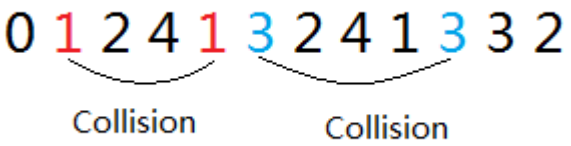
- We find that the SP 800-90B often underestimates the entropy source in the Non-IID track.
- We indicate the impact of underestimation.
- We point out the reason of underestimation:
  - The Collision Estimate and Compression Estimate are flawed.
- Experiments results support our conclusion.

Impact of underestimation

- For one estimator, underestimation is less harmful than overestimation. But for the SP 800-90B test suite, the underestimation of an estimator is more harmful than the overestimation.
- In 90B, each estimator calculates its own estimation independently. The minimum estimation of all is selected as the final result.
- Therefore, most overestimated values are not reflected in the final result. However, if an estimator provides a significant underestimate, the final result will be assigned to this underestimated value no matter how correct other estimators are.

The problem of Collision Estimate

- Collision Estimator:
  - If any value of the sequence repeats, we call there exists a collision. This estimator calculates the mean number of samples to the first collision as the statistic. It provides an min-entropy estimation on the basis of this statistic.
  - For example,



  - The Collision statistic is (4+5)/2=4.5

The problem of Collision Estimate

- For non-IID track, this estimator calculates the min-entropy on the basis of a statistic calculated from non-IID data, however, the formula g used to calculate the min-entropy is derived from the hypothesis that the test data are IID.
- That is to say, g is unsuitable for non-IID data.

The problem of Collision Estimate

- For example, if the tested data obey a first-order Markov process of {0, 1} whose transfer matrix is $\begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix}$.
  - When p > 0.5，that means the data have positive correlations, the Collision Estimator will underestimate the min-entropy of these data.
  - When p < 0.5， it will overestimate the min-entropy.
- In SP 800-90B, the Collision statistic $\bar{X}$ obeys following equation:

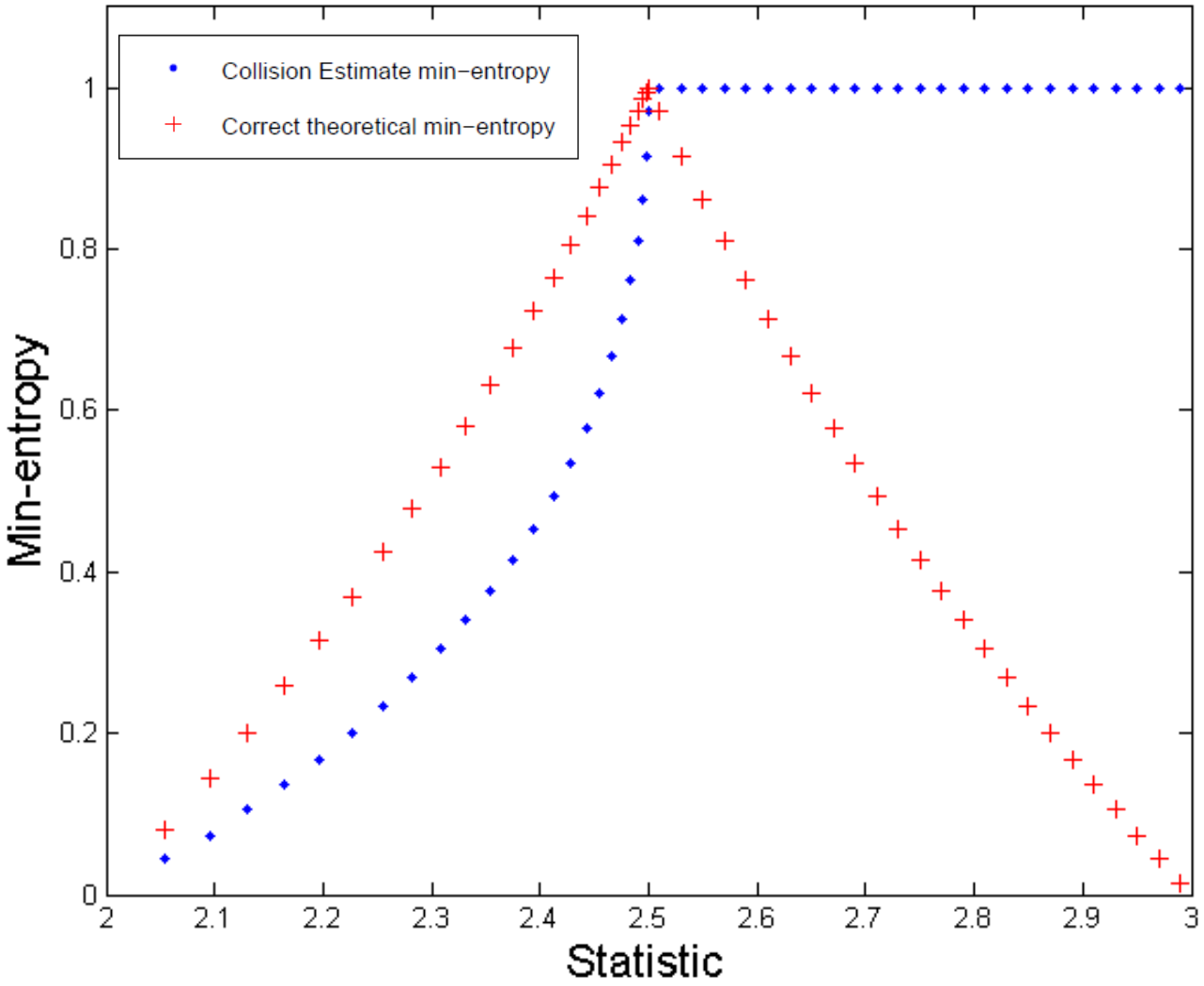
$$\bar{X} = pq^{-2}\left(1+\frac{1}{2}\left(p^{-1}-q^{-1}\right)\right)F\left(q\right)-pq^{-1}\frac{1}{2}\left(p^{-1}-q^{-1}\right)$$

Where $q=1-p, p\geq q, F(1/z)=\Gamma(3,z)z^{-3}e^{z}$.

- But in fact, for this simple Markov process, the true relationship between $\bar{X}$ and p is: $\bar{X}=3-p$

. .

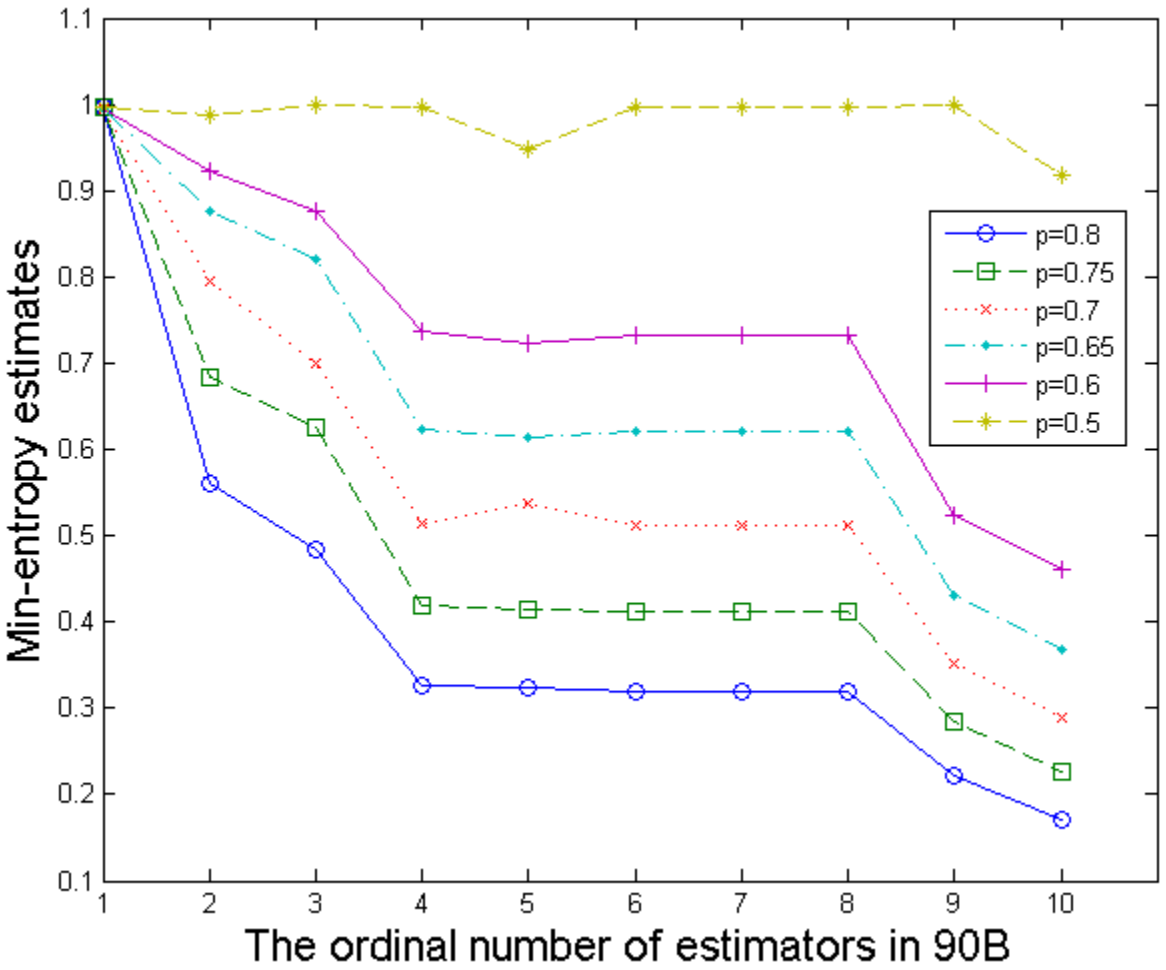
The problem of Collision Estimate



Experiments results

- We test an entrop
  data size of each gr

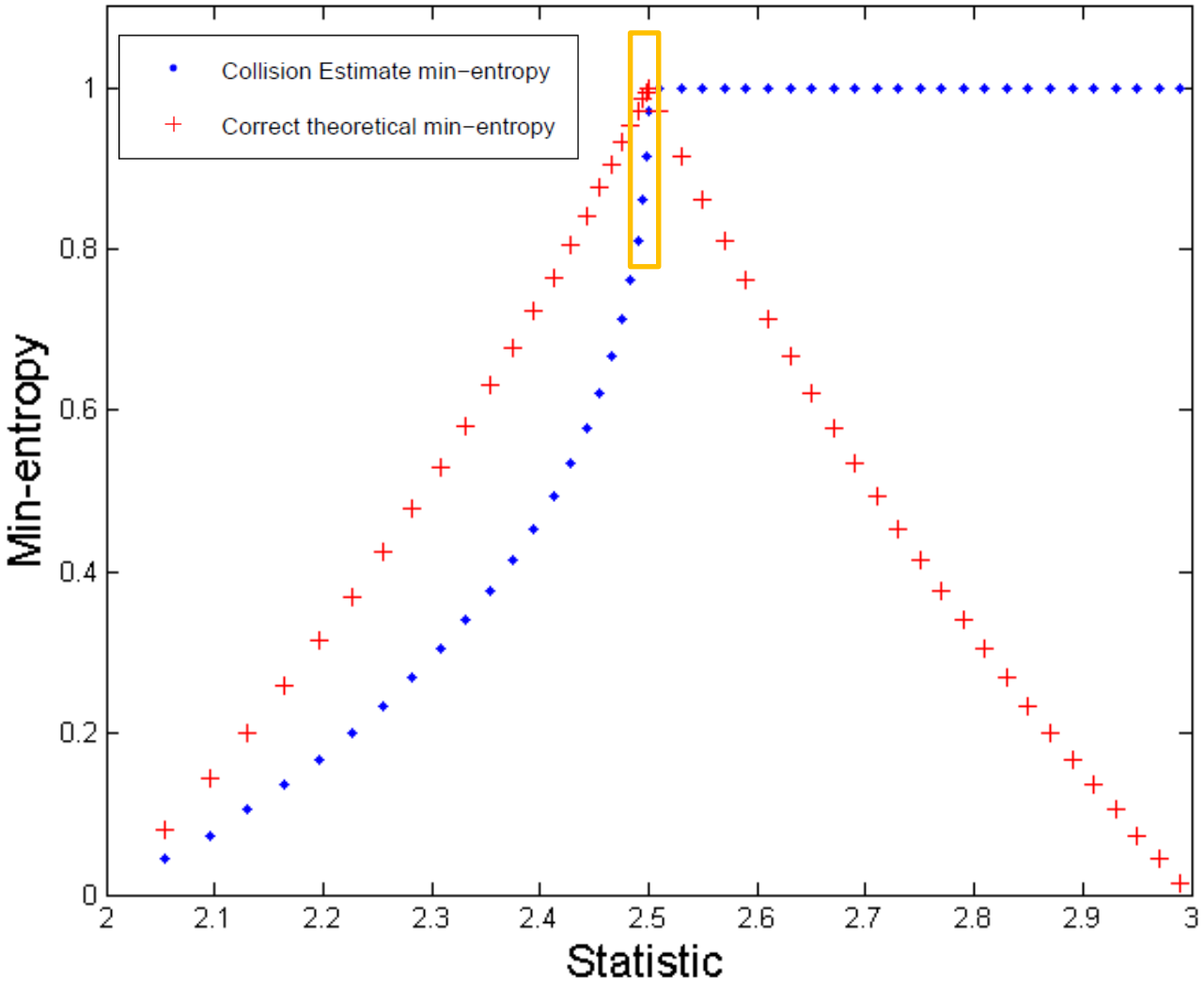
Estimator
1. MostCommon
2. LRS
3. MultiMCW
4. Markov
5. t-Tuple
6. Lag
7. MultiMMC
8. LZ78Y
9. **Compression**
10. **Collision**
Theoretical min-entropy



The problem of Collision Estimate

- For IID track, the Collision Estimate is highly sensitive to the statistic when the entropy sources are (almost) perfectly random.
- In addition, it takes the 99% confidence interval's lower bound of $\bar{X}$ to calculate the min-entropy, which means a value slightly smaller than $\bar{X}$ is calculated to estimate the min-entropy. And this causes a significant reduction of the min-entropy estimation.

The problem of Collision Estimate



The problem of Collision Estimate

- We choose several PRNGs in NIST SP 800-22 to generate random numbers and test them using 90B estimate suite.

Estimator	BBS	LCG	MODEXPG	MSG
MostCommon	0.9877	0.9893	0.9882	0.9832
LRS	0.9984	0.9996	0.9951	0.9803
MultiMCW	0.9873	0.9918	0.9834	0.9910
Markov	0.9927	0.9941	0.9853	0.9856
t-Tuple	0.9227	0.9359	0.9319	0.9430
Lag	0.9881	0.9947	0.9913	0.9931
MultiMMC	0.9910	0.9926	0.9883	0.9912
LZ78Y	0.9945	0.9910	0.9931	0.9924
Compression	1	1	1	1
Collision	**0.8250**	**0.8527**	**0.8630**	**0.8742**

The problem of Compression Estimate

- The Compression Estimate has the same problem with the Collision Estimate.
- In addition, in the second draft of SP 800-90B, the standard deviation in the Compression Estimate is inaccurate. It accounts for the abnormal results of Compression Estimate.

Compression	1	1	1	1
Collision	0.8250	0.8527	0.8630	0.8742

- In the current SP 800-90B, the standard deviation in the Compression Estimate is corrected. However the underestimation problem still exists.

A new estimator specific to Markov process

- The SP 800-90B only gives the formula of min-entropy for independent random variable X that takes values from the set $A = \{x_1, x_2, ..., x_k\}$ with probability $\Pr\{X = x_i\} = p_i$ for i =1,…,k :

$$H = \min_{1 \le i \le k} (-\log_2 p_i),$$

$$= -\log_2 \max_{1 \le i \le k} p_i.$$

- Therefore every estimator in SP 800-90B is trying to calculate something like $\max_{1 \le i \le k} p_i$ to get the min-entropy.

A new estimator specific to Markov process

- We propose an estimator directly according to the min-entropy of the high order Markov process, that is a very common model for non-IID entropy sources in real world.
- If we have known some outputs of an entropy source, we focus on the most-likely probability of the value of the subsequent output. The min-entropy is calculated using conditional probabilities:

$$H_D = -\log_2 \left(\sum_{\xi \in A^d} \Pr\{s_d = \xi\} \max_{1 \le i \le k} p_{\xi,i} \right)$$

- Instead of calculating something like $\max_{1 \le i \le k} p_i$, we directly estimate these conditional probabilities ($\Pr\{s_d = \xi\}$ and $\max_{1 \le i \le k} p_{\xi,i}$). So our estimator is very suitable for high order Markov process.

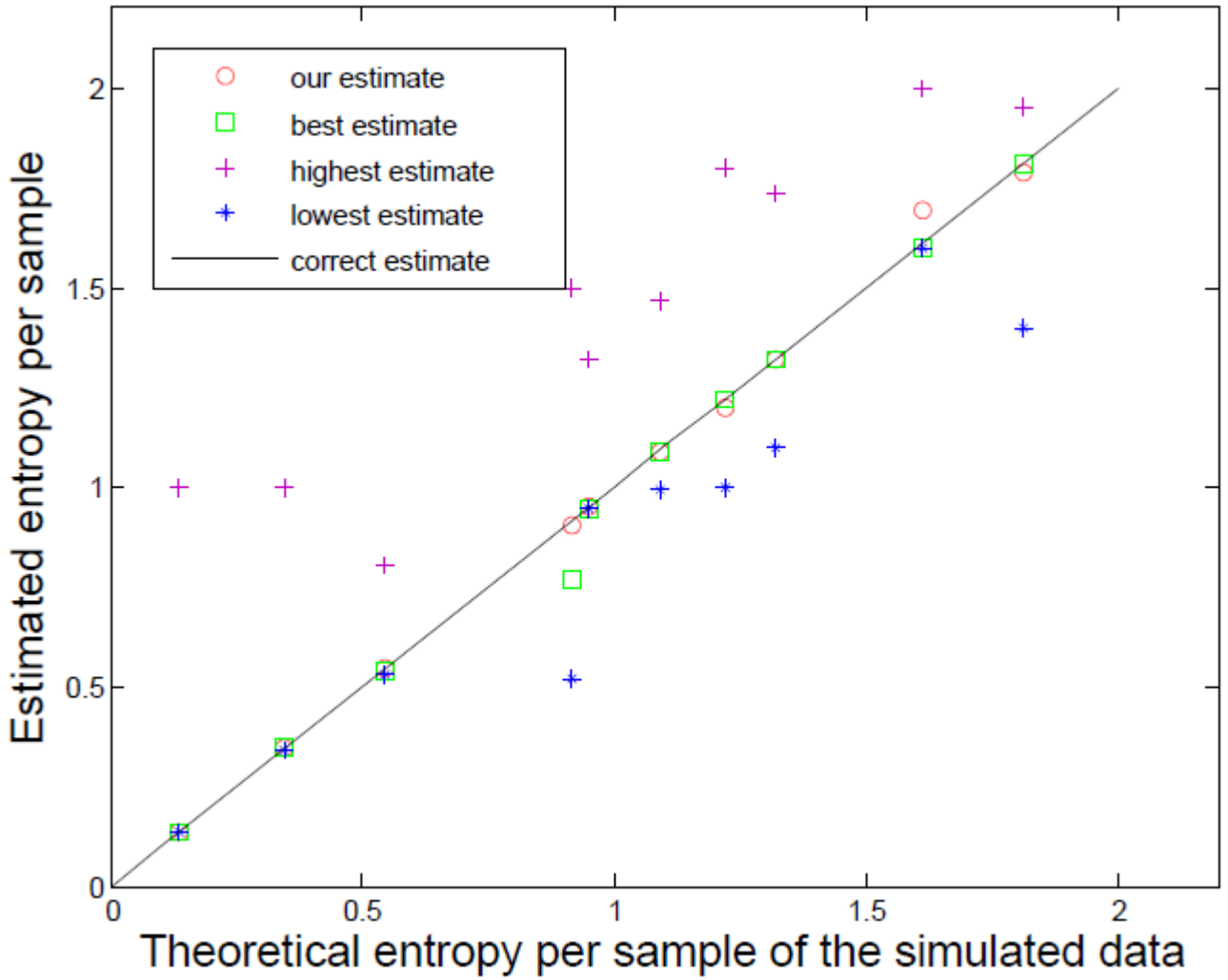
Comparison with 90B estimators

- The Markov Estimate in 90B is specific to the Markov process. However it is only suitable for the first-order Markov model due to the heavy computation complexity. Our estimator has lower computation complexity.

- The prediction based estimators in 90B are newly proposed. Our estimator is a frequency based estimator, but has comparable performance with the prediction based estimators.

Comparison with 90B estimators

- We ...del and Os...



Conclusion

- Accurate entropy estimation is critical for the evaluation of RNG security. We prove that the Collision Estimate and the Compression Estimate could provide significant underestimations for non-IID data.
- we propose a new estimator to calculate the min-entropy of non-IID sources, on the basis of conditional probability.
- Experiments on non-IID sources show that our estimator provides close estimates to the prediction based estimators in 90B.

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

Thanks !

T: 86 010-88889999 F: 86 010-88886666
E:daiyongming@zhongguokeyueyuan.com
地址：北京市海淀区闵庄路甲89号 100195