Information Theory
Lab 1: Familiarising with Entropy

A. Giorlandino

04/2022

1 Empirical Estimates Of Different Definitions of Entropy

Our goal is to show the validity of the law of large numbers by investigating how the empirical estimate of entropy converges to its theoretical value as the number of samples, from which we empirically estimate the probability mass function, increases. This procedure had been carried out for three different distributions and Shannon entropy, collision entropy and guessing entropy had been investigated.
The results that follow show the mean behaviour with the relative uncertainty evaluated by running the procedure for 100 independent cycles.

Uniformly Distributed Variables

Given N samples drawn from the distribution given by eq.(1) whose alphabet has cardinality M, the theoretical values of the three entropies under investigation coincide with one of nominal entropy, as depicted in eq.(2)

$$p_x(x) = \mathcal{U}(1, M) = \frac{\chi(x)_{[1,M]\cap\mathbb{N}}}{M} \tag{1}$$

$$H(x) = H_2(x) = H_{min}(x) = \log_2(M) \tag{2}$$

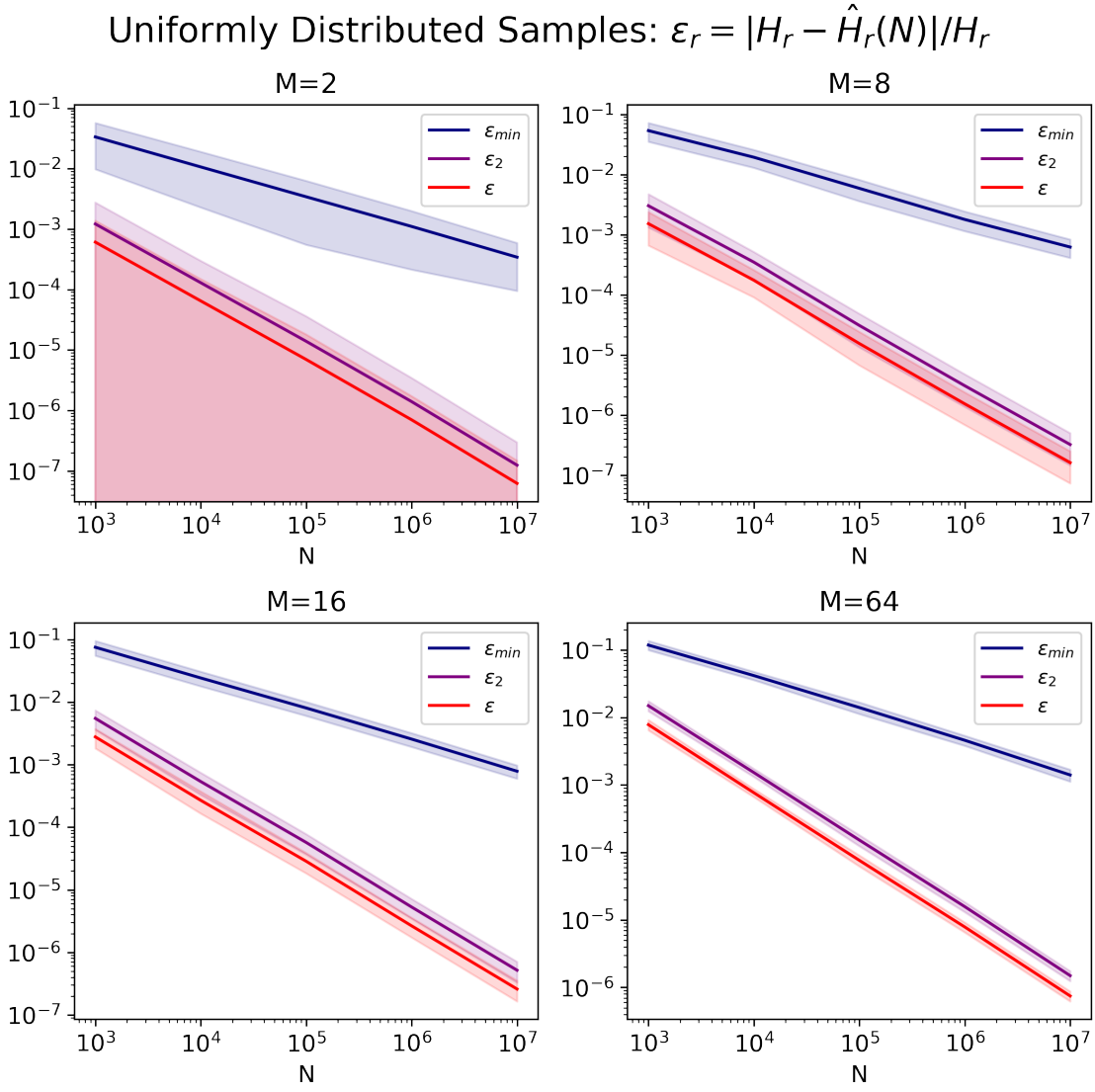

Figure 1: Relative error of the estimates of guessing Entropy, collision Entropy and Shannon Entropy versus the number of samples that are drawn from a uniform distribution of various sizes M.

While the relative error committed in the evaluation of collision entropy and Shannon entropy seem to have the same power law; the relative error committed in estimating the guessing entropy still decays exponentially, but with a slower rate. It is interesting to see that the fluctuation of the relative error decreases as the cardinality of the alphabet increases.

Bernoulli Distributed Variables

For a Bernoulli distributed variable the probability mass function is:

$$f(k) = q\delta(k) + (1-q)\delta(k-1) \qquad q \in [0,1] \tag{3}$$

Hence, k has binary alphabet and q represents the probability of drawing a zero; the consequent entropies are:

$$\begin{aligned} H(k) &= -q\log_2(q) - (1-q)\log_2(1-q) \\ H_2(k) &= -\log_2\left(q^2 + (1-q)^2\right) \\ H_{min}(k) &= min\{-\log_2(q), -\log_2(1-q)\} \end{aligned}$$

In fig.(2), the relative error as a function of the number of the drawn samples is reported for different values of q. Except for the special case where $q = 1 - q = 0.5$, the relative error has the same behavior for the three entropies.

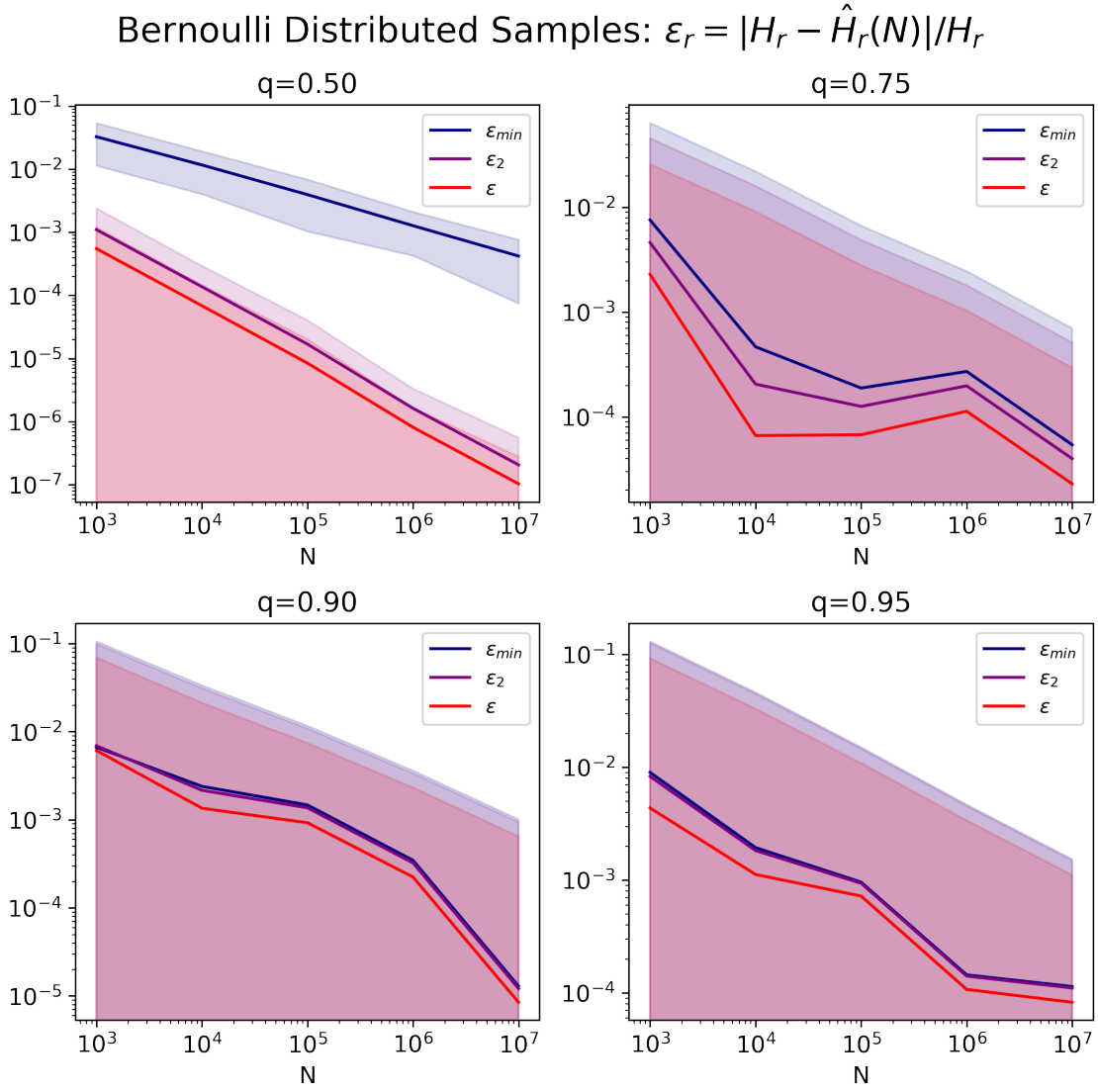

Figure 2: Relative error of the estimates of Guessing Entropy, Collision Entropy and Shannon Entropy versus the number of samples that are drawn from a Bernoulli distribution for various values of q.

Geometrically Distributed Variables

Given a Bernoulli process with probability of success λ, the geometric distribution gives the probability that the first occurrence of success requires k independent trials and it can be written as:

$$p(k) = (1-\lambda)^{k-1}\lambda \qquad k \in \mathbb{Z}^+ \tag{4}$$

Finding the associated probabilities is easy by exploiting the properties of geometric series; the results are:

$$H(k) = -\log_2(\lambda) - \frac{(1-\lambda)}{\lambda}\log_2(1-\lambda)$$
$$H_2(k) = -\log_2(\lambda/(2-\lambda))$$
$$H_{min}(k) = -\log_2(\lambda)$$

The simulations are reported in fig.(3). In this case the fluctuations are more visible than the previous cases, but the upper bound still decays exponentially. This effect could be toned down by increasing the number of cycles used in determining the mean and the variance of relative error.

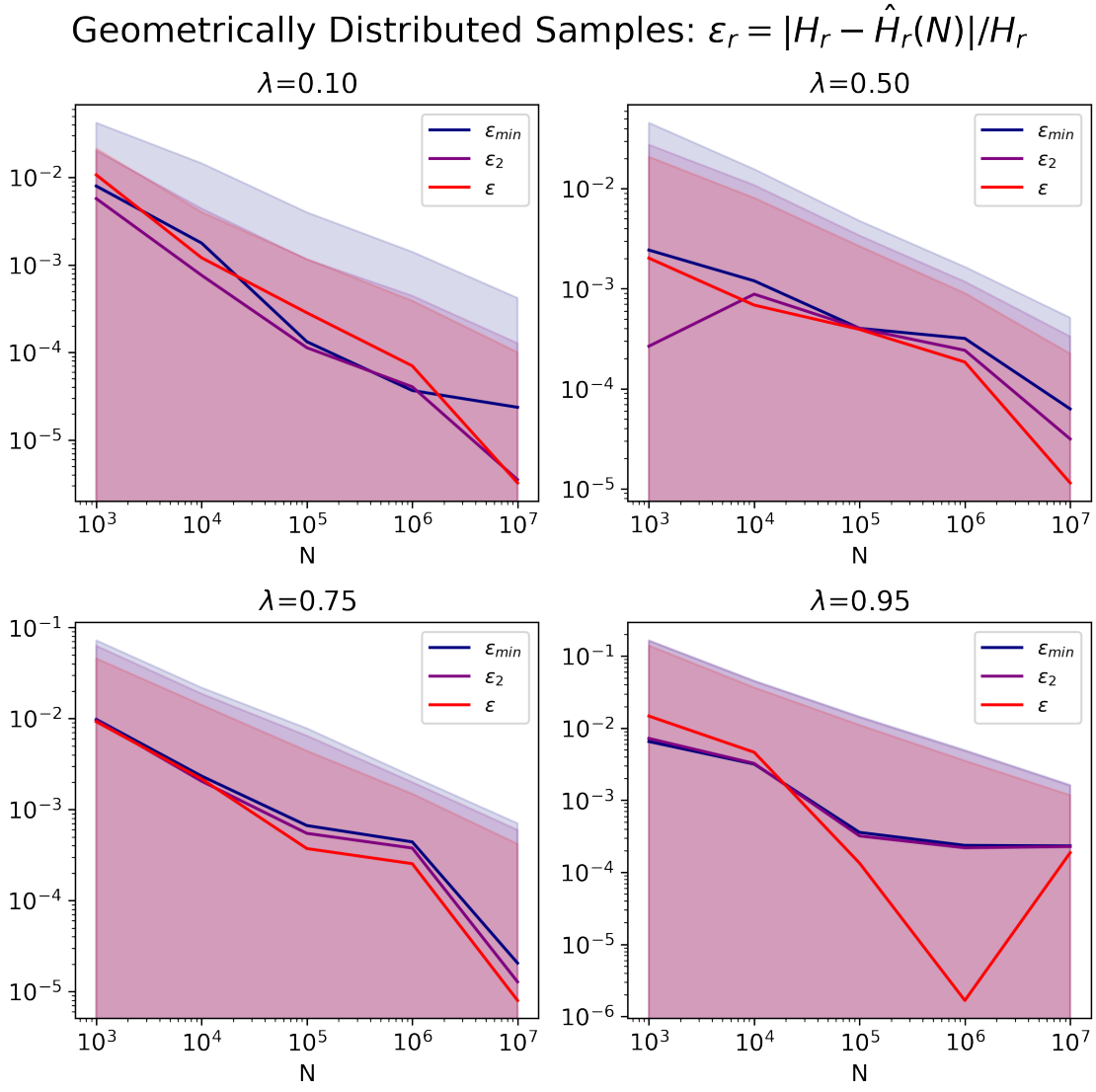

Figure 3: Relative error of the estimates of Guessing Entropy, Collision Entropy and Shannon Entropy versus the number of samples that are drawn from a Geometric distribution for various values of λ.

2 Joint Empirical Entropy and Relative Quantities

Similarly as we proceeded for single sequences, we can estimate the joint probability and the the conditional probabilities of two random variables by exploiting the law of large numbers when two arbitrarily long lists of samples for two discrete random variables x and y are given.
Once we have estimated empirically the aforementioned probabilities, it is possible to evaluate of the corresponding Shannon entropies. While samples within the same sequence are $i.i.d$; samples belonging to different sequences might not even be independent. For this reason, we shall focus both of the two cases.

Sequences Of Correlated Random Variables

Given the sequences X and Y of the same length N, all the realization within the same sequence are independent from each other i.e. $p(x_i, x_j) = p(x_i)p(x_j), \forall i \neq j$ and the same $\forall y_i \neq y_j$. Also $p(x_i, y_j) = p(x_i)p(y_j) \quad \forall i \neq j$.
The correlation might be within x_i and y_i. For example we focus on the following case:

$$x_i \sim \mathcal{U}(0, M) \qquad z_i \sim \mathcal{U}(-1, 1) \qquad y_i = x_i + z_i \tag{5}$$

While x_i and z_i are independent, clearly y_i is correlated to x_i. The resulting distribution for y is given by eq.(6).

$$p_y = \begin{cases} 1/3M & y = 0, M+1 \\ 2/3M & y = 1, M \\ 1/M & y \in [2, M-1] \end{cases} \tag{6}$$

From which we can easily find $H(y)$ by applying Shannon's definition. Given how y is built, it is sensitive to conclude that the expression of $p(y|x)$ is:

$$p(y|x) = \begin{cases} 1/3 & y \in [x-1, x+1] \\ 0 & else \end{cases} \qquad \Longrightarrow \qquad H(y|x) = \log_2(3) \tag{7}$$

The other important quantities follow using the relations:

$$H(x,y) = H(y|x) + H(x) \qquad H(x|y) = H(x,y) - H(y) \qquad I(x;y) = H(x) + H(y) - H(x,y) \tag{8}$$

For sufficiently large sequence $\mathcal{A}_x \subset \mathcal{A}_y$, so it is possible to evaluate the Kullback–Leibler divergence from its definition.

We generate the sequences X and Y varying their length we calculate each time the joint and conditional frequencies. By doing this for 100 cycles we are able to estimate the mean behaviour and the uncertainty for a given configuration of N.
When comparing the empirical results to the expected theoretical values for different choices of M and varying the length N of the sequences, the results follow in figs.(4- 5). Note that the power law that determines the decay of the relative errors for the joint and conditional probability is visibly independent from M: in log-log scale the slope doesn't seem to depend on M. The same cannot be concluded for the relative error committed on the mutual information and on the Kullback-Leibler divergence as the statistical fluctuations make this effect not clear.

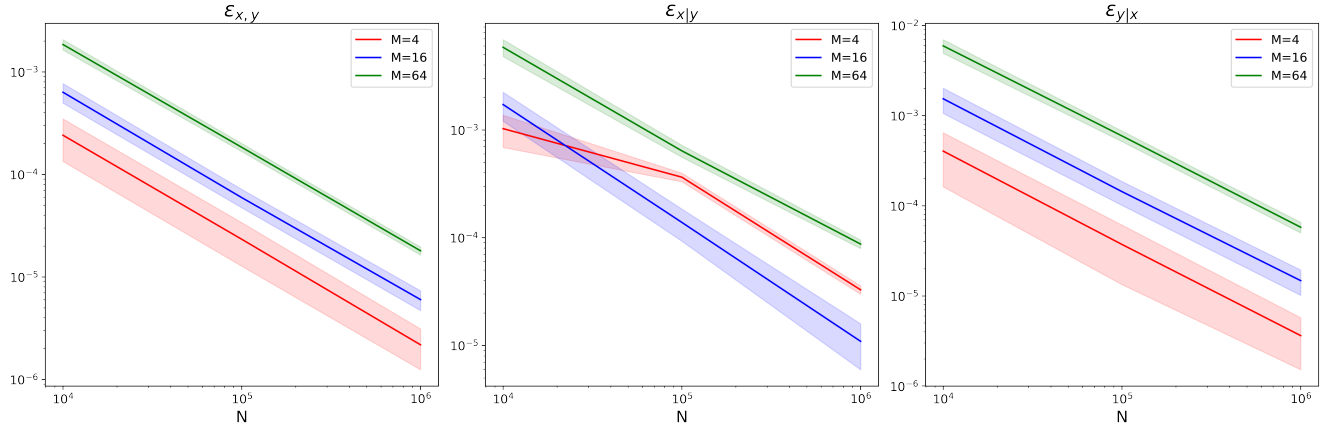

Figure 4: Relative error of the estimates of joint entropy and the respective conditional entropies for x and y versus the length of the sequences.

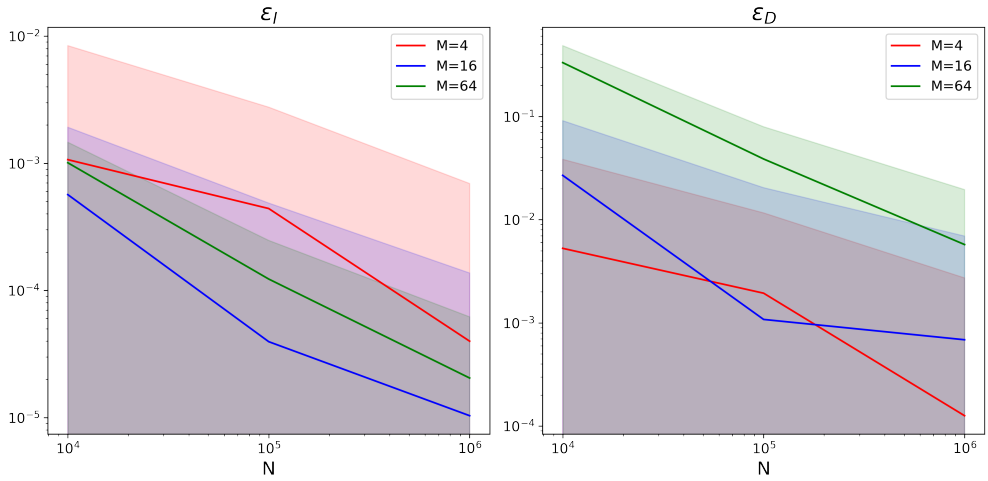

Figure 5: Relative error committed in the evaluation of the mutual information and the Kullback-Leibler divergence of two correlated random variables versus the number of samples used in the empirical estimation of the quantities in consideration.

Sequences Of Independent Random Variables

Contrary to what has just been discusses, let's now treat the case of two independent random variables. We generate two independent sequences X and Y of i.i.d realizations, where $x_i \sim p_{\lambda_x}(x_i)$ and $y_i \sim p_{\lambda_y}(y_i)$ and $p_\lambda(k)$ is the geometric distribution discussed previously in eq.(4). As the two variables are independent the expressions of the joint and conditional entropies are trivial.
Moreover, the theoretical mutual information is zero; hence it's not possible to show the relative error. For this reason the empirical mutual information is directly reported for nine different combinations of λ_x and λ_y in fig.(6). Note how it exponentially tends to zero (its theoretical value) as N gets larger and larger. This behaviour doesn't depend on the particular combination (λ_x, λ_y).

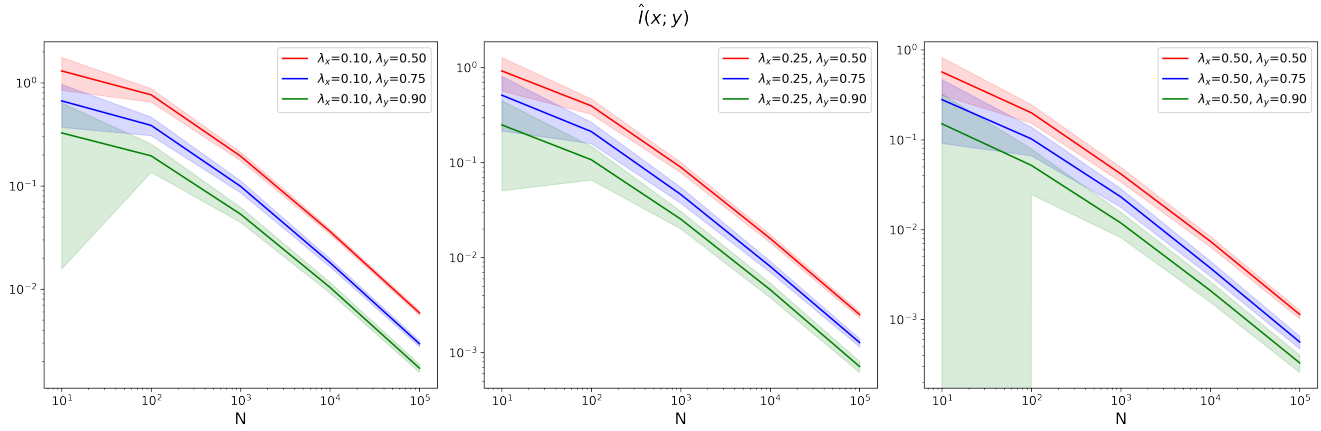

Figure 6: Empirical mutual information of two independent geometrically distributed random variables (with parameters λ_x and λ_y) versus the number of samples used.

In this case the alphabets of the two random variables coincide with $\mathbb{N}$, which clearly has infinite cardinality. For this reason the empirical estimation of the Kullback-Leibler divergence is not always well-defined *i.e.* generating finite sequences it can happen that $\hat{\mathcal{A}}_x \not\subset \hat{\mathcal{A}}_y$ and $\hat{\mathcal{A}}_y \not\subset \hat{\mathcal{A}}_x$, which leads to an infinite KL divergence. But, theoretically we expect the contrary of this.
If $\lambda_x = \lambda_y$ then the KL divergence is zero; more generally:

$$D(p_{\lambda_x}||p_{\lambda_y}) = \log_2\left(\frac{\lambda_x}{\lambda_y}\right) + \frac{1-\lambda_x}{\lambda_x}\log_2\left(\frac{1-\lambda_x}{1-\lambda_y}\right) \tag{9}$$

and the analogous for $D(p_{\lambda_y}||p_{\lambda_x})$. In the simulation the most pathological case, as could be expected, was $\lambda_x = \lambda_y = 0.5$: the theoretical value for the KL divergence is zero but only $\sim 64\%$ of the generated couple of

sequences was such that the estimated alphabets were one the subset of the other.

On the other hand the estimates of the joint and conditional probabilities agreed with their theoretical values as the relative error exponentially goes to zero, independently of the choice of λ_x and λ_y, as visible in fig.(7).

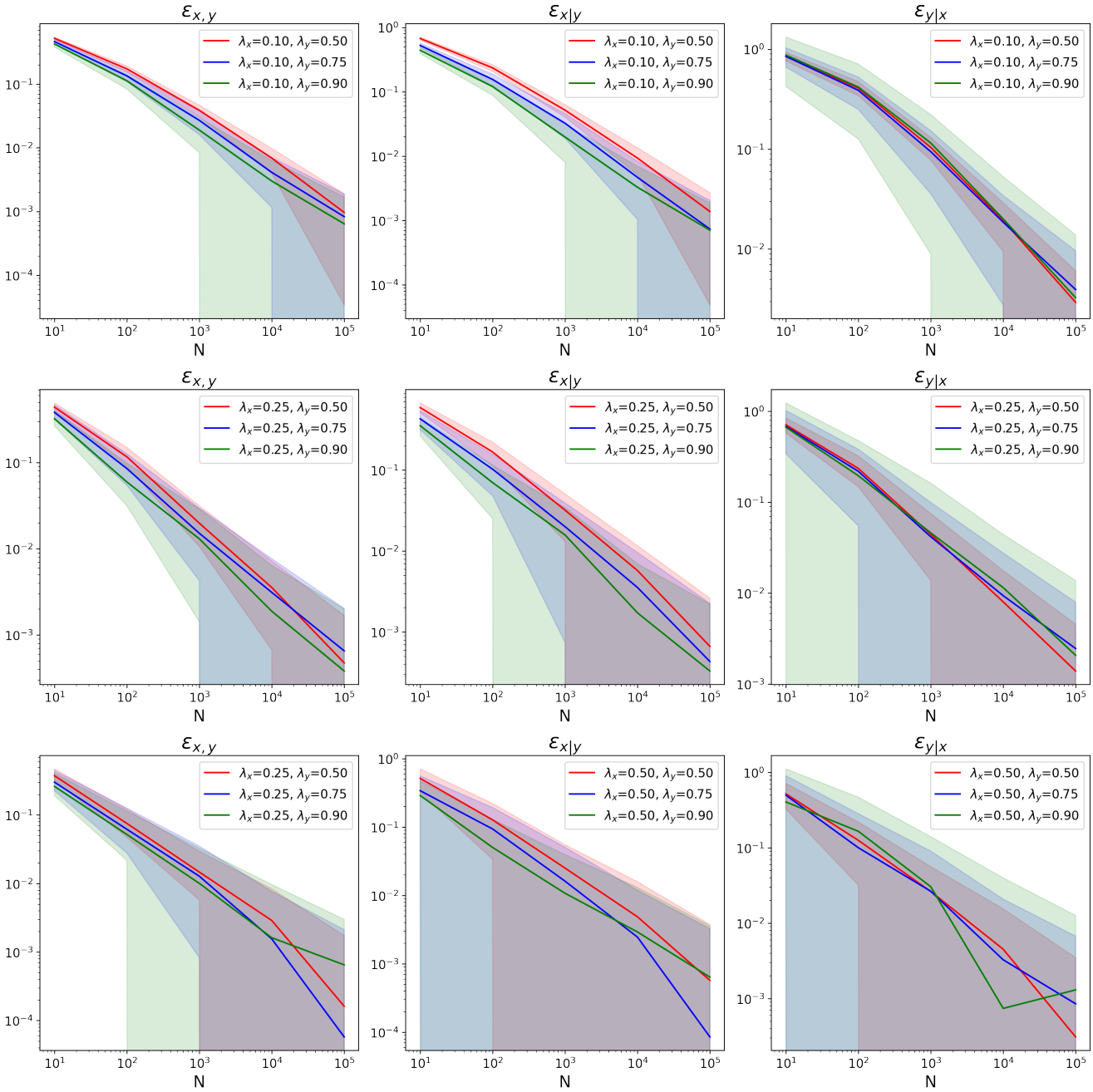

Figure 7: Relative error of the estimates of joint entropy and the respective conditional entropies for x and y independent geometrically distributed variables versus the length of the sequences used in the empirical estimation.

3 Code Source

The code of the all the simulations is visible on GitHub at `https://github.com/aGiorlandino/Information-Theory-Labs/blob/main/lab1/lab1.ipynb`