Information Theory (2021/22)
Homework 1: Estimating Entropies

Tommaso Bergamasco (ID: 2052409)

Contents

1 Introduction

In this report I'm going to describe a procedure to estimate, starting from a vector of realizations of a generic unknown random variable, some entropies and other related quantities.

In Section 2 I am going to consider a single vector of realizations x and I will try to estimate:

- Alphabet A_x and pmd p_x of x.
- Shannon Entropy $H(x)$.
- Collision Entropy $H_{coll}(x)$.
- Guessing Entropy $H_{min}(x)$.

In Section 4 I am going to extend the analysis to the "multivariate" case for which I will try to estimate the following quantities regarding 2 vectors of realizations x and y:

- Alphabets A_x and A_y and pmds p_x and p_y of x and y.
- Joint Entropy $H(x, y)$.
- Conditional Entropy $H(x|y)$.
- Relative (Kullback-Leibler Divergence) Entropy $D(p_x||p_y)$.
- Mutual Information $I(x; y)$.

Each one of these procedures are then tested by comparing the estimated entropies to their "true" values for some selected well known random distributions.

The results and the plots have been obtained by using `Python 3.8.8` and in particular its libraries `NumPy` and `Matplotlib`. In the Homework's folder the code I used to obtain such results can be found in the `Jupyter Notebook`.

2 Estimating the Empirical Entropies of x

2.1 Alphabet and pmd of x

Given the arbitrarily long vector of realizations x, I used the function `np.unique()` to extract all the values occurring at least once in x (Alphabet A_x) and to count (for each one of them) its absolute frequence. I then divided such values for the length of x obtaining the *relative frequencies* and hence an empirical estimate of the pmd, namely:

```
A_x, pmd = np.unique(x, return_counts=True)
p_x = np.divide(pmd, x.shape[0])
```

where:

$$A_x = [a_1, a_2, \dots, a_M] \tag{1}$$
$$p_x = [p_x(a_1), p_x(a_2), \dots, p_x(a_M)]. \tag{2}$$

2.2 Shannon Entropy $H(x)$

Using the pmd of x from Equation 2, we can easily compute $H(x)$ by definition:

$$H(x) = \sum_{a \in A_x} p_x(a) \underbrace{\log_2 \frac{1}{p_x(a)}}_{i_x(a)}. \tag{3}$$

(Note that we don't need any information about the Alphabet A_x, the vector p_x is sufficient).

2.3 Collision Entropy $H_{coll}(x)$

Again using only the pmd vector p_x in Equation 2 we first obtain the expression for the *Collision Probability* and then for the *Collision Entropy* itself:

$$P_{coll}(x) = \sum_{a \in A_x} p_x^2(a) \tag{4}$$

$$H_{coll}(x) = \log_{1/2} P_{coll}(x). \tag{5}$$

2.4 Guessing Entropy $H_{min}(x)$

Once again using only Equation 2 we obtain:

$$H_{min}(x) = \min_{a \in A_x} i_x(a) = \log_{1/2} P_{max}(x). \tag{6}$$

where P_{max} is the largest value of the vector p_x.

3 Tests on the Estimates for x

The tests are performed by generating random vectors of realizations from a selection of well known random distributions such as:

- Uniform Distribution over $A_x = \{1, \ldots, M\}$.
- Binary Distribution with $p_x(0) = q$.
- Geometric Distribution with $p_x(k) = (1 - \lambda)\lambda^k$.

For each one of these distributions I compare my estimate with their "true" theoretical value and I also plot its relative precision $\varepsilon = \frac{|\hat{H}(x) - H(x)|}{H(x)}$.

3.1 Tests on $x \sim \mathcal{U}[1, M]$

3.1.1 Uniform Theoretical Values

For a Theoretical random variable $x \sim \mathcal{U}[1, M]$ we note that the alphabet cardinality is

$$|A_x| = (M - 1) + 1 = M.$$

Therefore all the quantities of interest can be found (using the same equations I used in 2) noticing that:

$$p_x(a) = \frac{1}{M}, \quad \forall a \in A_x.$$

In particular we will have:

$$H(x) = \log_2 M = H_{min}(x) = H_{coll}(x) \tag{7}$$

3.1.2 Varying M

In Figure 1 I compare the estimates and the true entropies when varying the alphabet cardinality M in the range $M \in [2, 100]$ with a fixed length of the empirical vector x of $L = 10000$.

In Figure 2 I plot the relative precision defined as $\varepsilon = \frac{|\hat{H}(x) - H(x)|}{H(x)}$ for each one of the 3 entropies.

Note the following:

- The 3 entropies have the same theoretical value.
- ε tends to grow with the value of M.

This is quite intuitive since a larger M implies a larger entropy (uncertainty) which makes our estimation harder.

e.g. if $L = 100$ and $M = 2$ then we will expect that $\hat{p}_x(1) \simeq \hat{p}_x(2)$. If $L = 100$ *but* also $M = 100$ we wouldn't even be able (with high probability) to correctly identify the alphabet, leading to not neglegible estimation errors.

- It holds that: $\varepsilon[H_{min}(x)] >> \varepsilon[H_{coll}(x)], \varepsilon[H(x)]$ (where $\varepsilon[a]$ is the relative estimation error regarding the quantity a).

In facts note that both the Shannon and Collision entropy depend on some sort of averages:

$$H(x) = \mathbb{E}[i_x(x)] = \mathbb{E}[\log_{1/2} p_x(x)] \tag{8}$$

$$H_{coll}(x) = \log_{1/2} \left(\sum_{A_x} p_x(a) p_x(a) \right) = \log_{1/2} \mathbb{E}[p_x(x)]. \tag{9}$$

Which makes them more robust to noise, while the Guessing Entropy depends only on a single estimated probability $P_{max}(x)$ and therefore is more exposed to estimated errors.

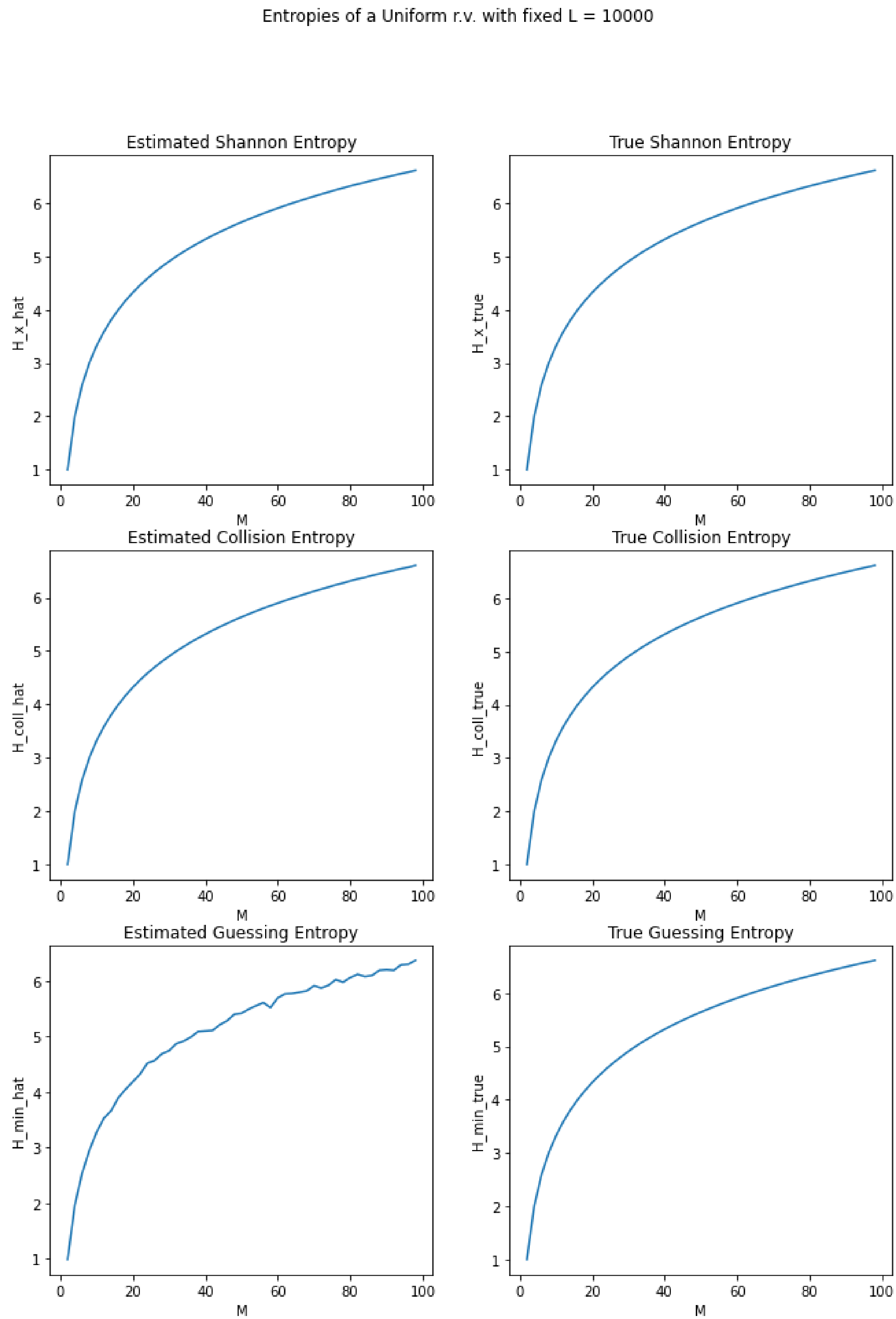

Figure 1: Estimates (left) vs True (right) Entropies comparison for a Uniform r.v. and a varying M

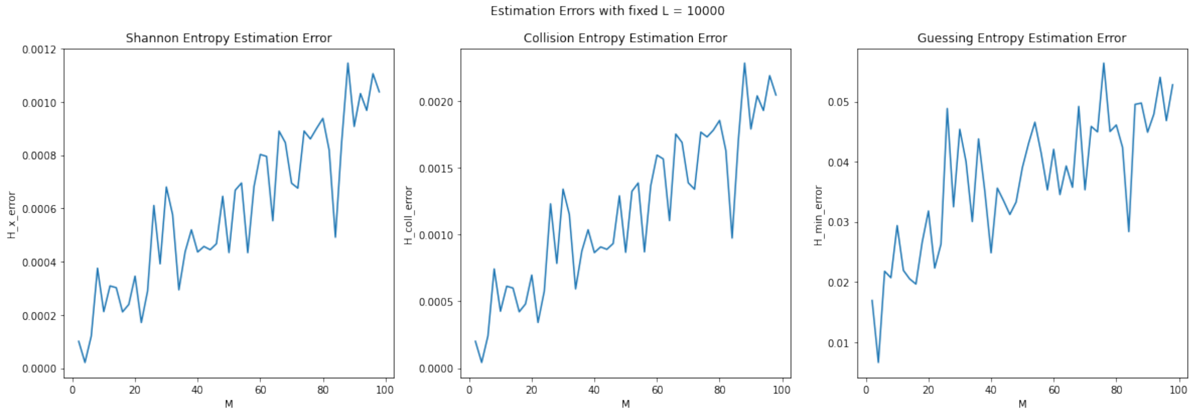

Figure 2: Relative Precision of the estimates for a uniform r.v. and a varying M

3.1.3 Varying L

I now compare in Figure 3 the estimates and the true entropies with $M = 3$ fixed and the length of the vector of realizations x ranging in the interval $L \in [30, 10000]$ with step 400.

In Figure 4 the relative estimation errors are plotted.

Note that:

- The error decreases when increasing L (this is trivial).
- While the $H(x)$ and $H_{coll}(x)$ estimates are almost equally precise, $H_{min}(x)$ keeps oscillating conspicuously even for $L >>$. This confirms what have been said in the paragraph above.

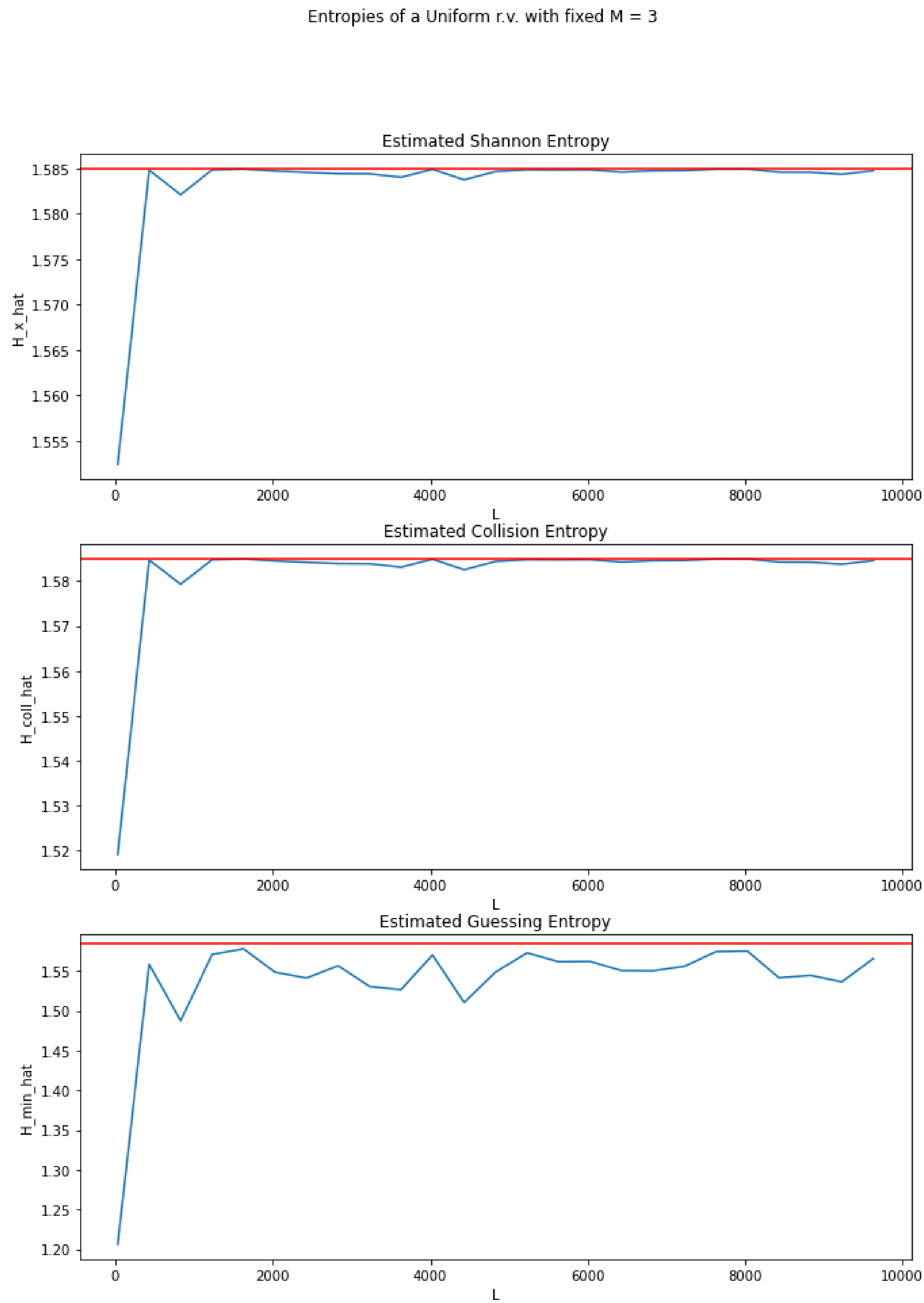

Figure 3: Estimates (blue) vs True (red) entropies comparison for a Uniform r.v., with fixed M=3 and a varying L

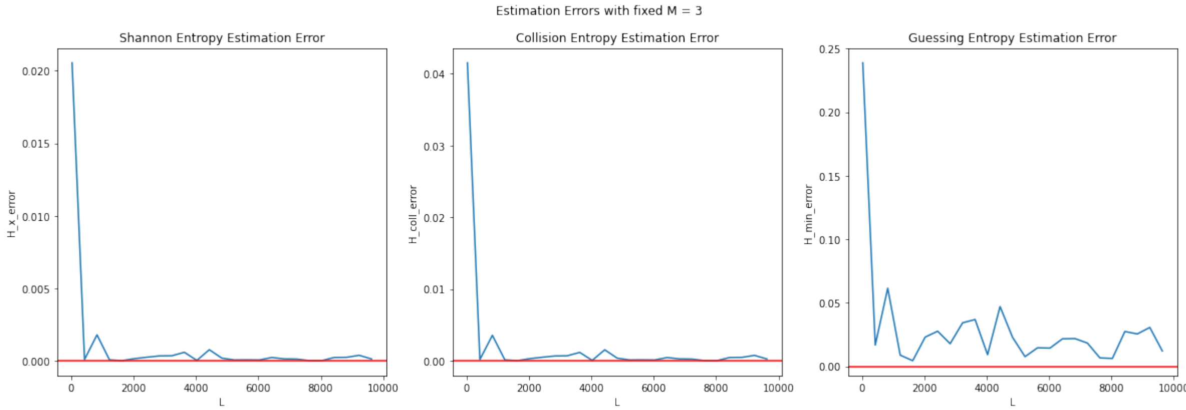

Figure 4: Relative precision of the estimates for a uniform r.v., with fixed M=3 and a varying L

3.2 Tests on $x \sim \mathcal{B}(1-q)$

3.2.1 Binary Theoretical Values

Letting $p_x(0) = q$ and $p_x(1) = 1 - q = p$, we have:

$$H(x) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p} \tag{10}$$

$$H_{coll}(x) = \log_2 \frac{1}{2p^2 - 2p + 1} \tag{11}$$

$$H_{min}(x) = \min\{i_x(0), i_x(1)\} \tag{12}$$

3.2.2 Varying q

Similarly to what I did for the uniform distribution, in Figure 5 and 6 are plotted the same quantities for a Binary r.v. with "failure probability" $q \in [0, 1]$.

Note that:

- All the entropies are (and must be) symmetric w.r.t. the line $x = 0.5$ where the peak for all the entropies occur.
- $H(x)$ and $H_{coll}(x)$ are almost equally precise.
- $H(x) \geq H_{coll}(x)$.

In facts, by recalling the Equations 8 and 9, by noticing that x is **not a.s. constant** $\forall q \neq 0, 1$ and that $\log_{1/2}$ is strictly convex, we can state using the Jensen inequality that:

$$H(x) = \mathbb{E}[\log_{1/2} p_x(x)] \geq \log_{1/2} \mathbb{E}[p_x(x)] = H_{coll}(x) \tag{13}$$

The equality holds for $q = 0, 1$ (for which x is a.s. constant) and for $q = 0.5$ (for which x is a Uniform r.v. with $M = 2$ and we already showed that $H(x) = H_{coll}(x)$ in Equation 7).

- $H_{coll}(x) \geq H_{min}(x)$

In facts

$$H_{coll}(x) = \log_{1/2} \mathbb{E}[p_x(x)] \geq \log_{1/2} P_{max}(x) = H_{coll}(x) \tag{14}$$

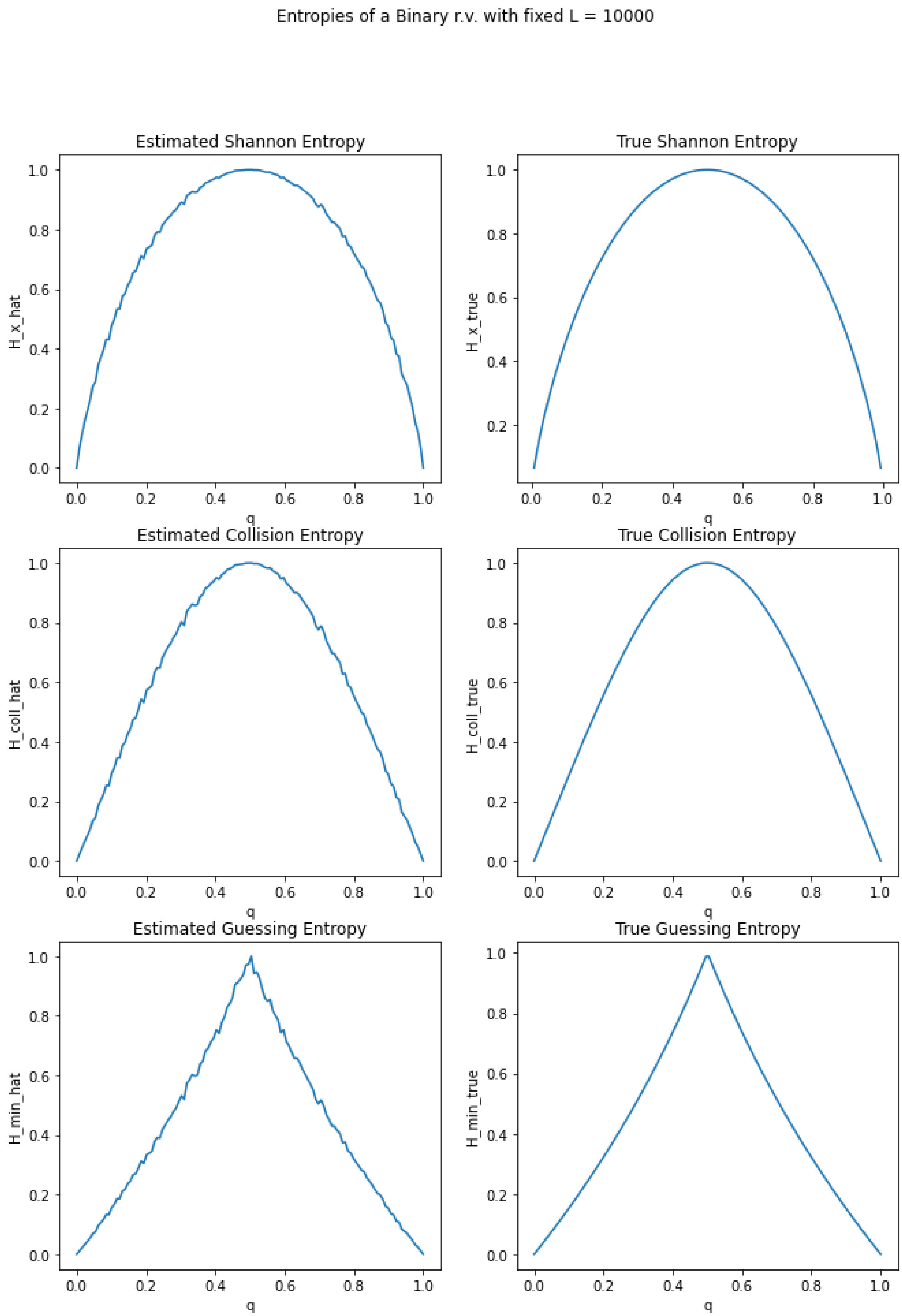

Figure 5: Estimates (left) vs True (right) Entropies comparison for a Binary r.v. and a varying q.

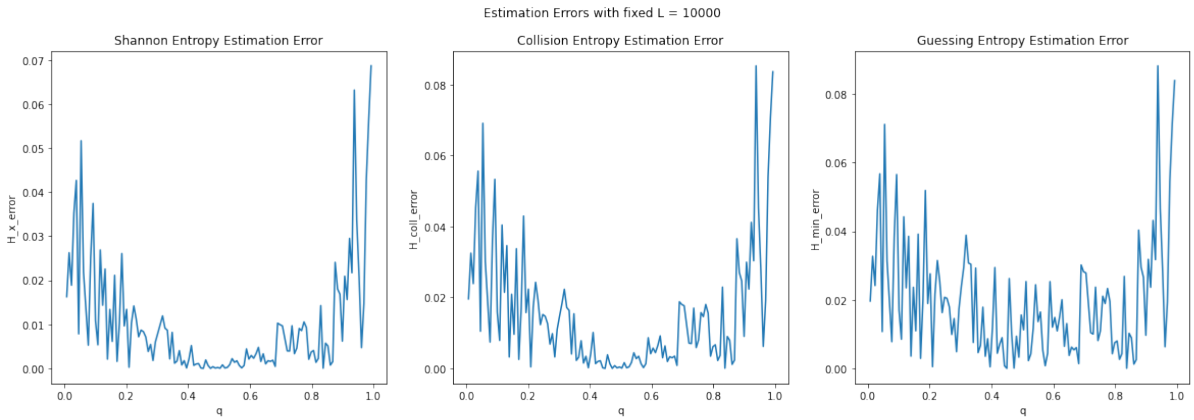

Figure 6: Relative Precision of the estimates for a Binary r.v. and a varying q.

The equality holds when $\mathbb{E}[p_x(x)] = P_{max}(x) \Leftrightarrow q = 0, 1, 0.5$, namely when x is a.s. constant or it's a Uniform r.v. with $M = 2$ (see Equation 7).

- The largest errors take place when q is "small" or "large".

This is due to the fact that is more difficult to estimate "small" probabilities when we are bound to non-infinite-length empircal vectors.
e.g. if $L = 10$ and $q = 0.1$ then even 2 failures in the vector of realizations of length L would lead to estimate $\hat{q} = 2q$ (if $q = 0.5$ the estimates are more robust).

- Once again $H_{min}(x)$ is the less robust quantity since, as in the uniform case, it doesn't benefit from any sort of averaging. We can see this well when $q \simeq 0.5$ where instead $H(x)$ and $H_{coll}(x)$ perform significantly better.

3.2.3 Varying L

I now consider in Figure 7 and 8 the entropies and errors for a Binary r.v. and the length of the empirical vector L ranging in $L \in [30, 10000]$. I choosed as failure probability $q = 0.5$ in order to highlight the difference between $H(x), H_{coll}(x)$ and $H_{min}(x)$.

- Note once again how $H(x)$ and $H_{coll}(x)$ behave similarly, while $H_{min}(x)$ is the less precise even for $L >>$.

3.3 Tests on $x \sim \mathcal{G}(\lambda)$

3.3.1 Geometric Theoretical Values

Let λ be the "failure probability", then the probability to "fail" exactly k times is: $p_x(k) = (1-\lambda)\lambda^k$. Let $p = 1 - \lambda$, then with some easy computations, by using some infinite summation properties, we obtain:

$$H(x) = p_x(0)i_x(0) + p_x(1)i_x(1) + \ldots = \sum_{k=0}^{\infty}(1-p)^k p \log_2 \frac{1}{(1-p)^k p} = \frac{p-1}{p}\log_2(1-p) - \log_2 p \quad (15)$$

$$H_{coll}(x) = \log_2 \frac{1}{P_{coll}(x)} = \log_2 \frac{1}{\sum_{k=0}^{\infty}(1-p)^{2k}p^2} = \log_2 \frac{-p^2+2p}{p^2} \quad (16)$$

$$H_{min}(x) = \min\left\{\log_2 \frac{1}{p}, \log_2 \frac{1}{(1-p)p}, \log_2 \frac{1}{(1-p)^2 p}, \ldots\right\} = \log_2 \frac{1}{p} \quad (17)$$

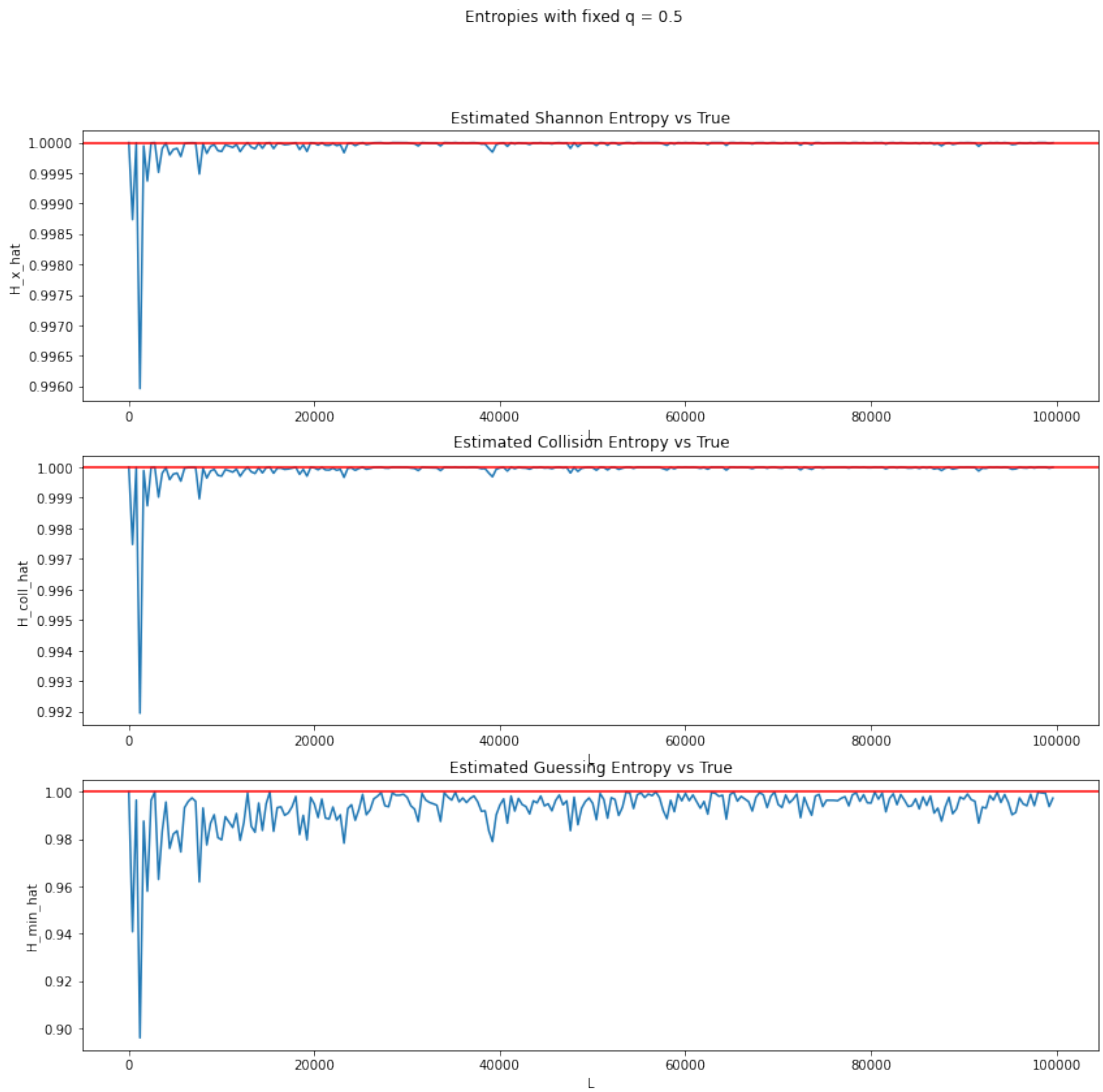

Figure 7: Estimates (blue) vs True (red) entropies comparison for a Binary r.v., with fixed q=0.5 and a varying L.

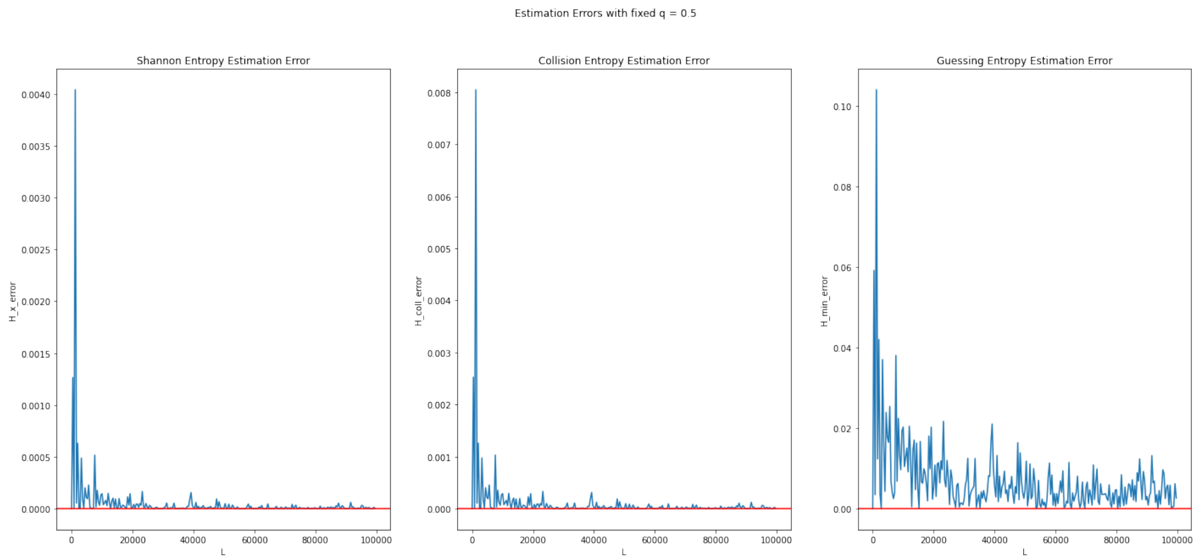

Figure 8: Relative precision of the estimates for a Binary r.v., with fixed q=0.5 and a varying L.

3.3.2 Varying λ

See Figure 9 and 10 to see the entropies and relative precision for a geometric r.v. with failure probability λ ranging in $\lambda \in [0, 1]$ and fixed length of the empirical vector $L = 10000$.

Note that:

- It holds that:

$$\lim_{\lambda \to 1} H(x) = \lim_{p \to 0} H(x) = \lim_{p \to 0} H_{coll}(x) = \lim_{p \to 0} H_{min}(x) = +\infty$$

This is due to the fact that when $\lambda \to 1$, all the symbols of the infinite alphabet of the geometric r.v. become (on average) more likely, increasing the uncertainty.

- It still holds that: $H(x) \geq H_{coll}(x) \geq H_{min}(x)$.

This follows from equations 13 and 14, and the equality holds only when $\lambda = 0$ when there is no uncertainty and all the entropies are 0, or when $\lambda \to 1$ when all the 3 entropies are infinite.

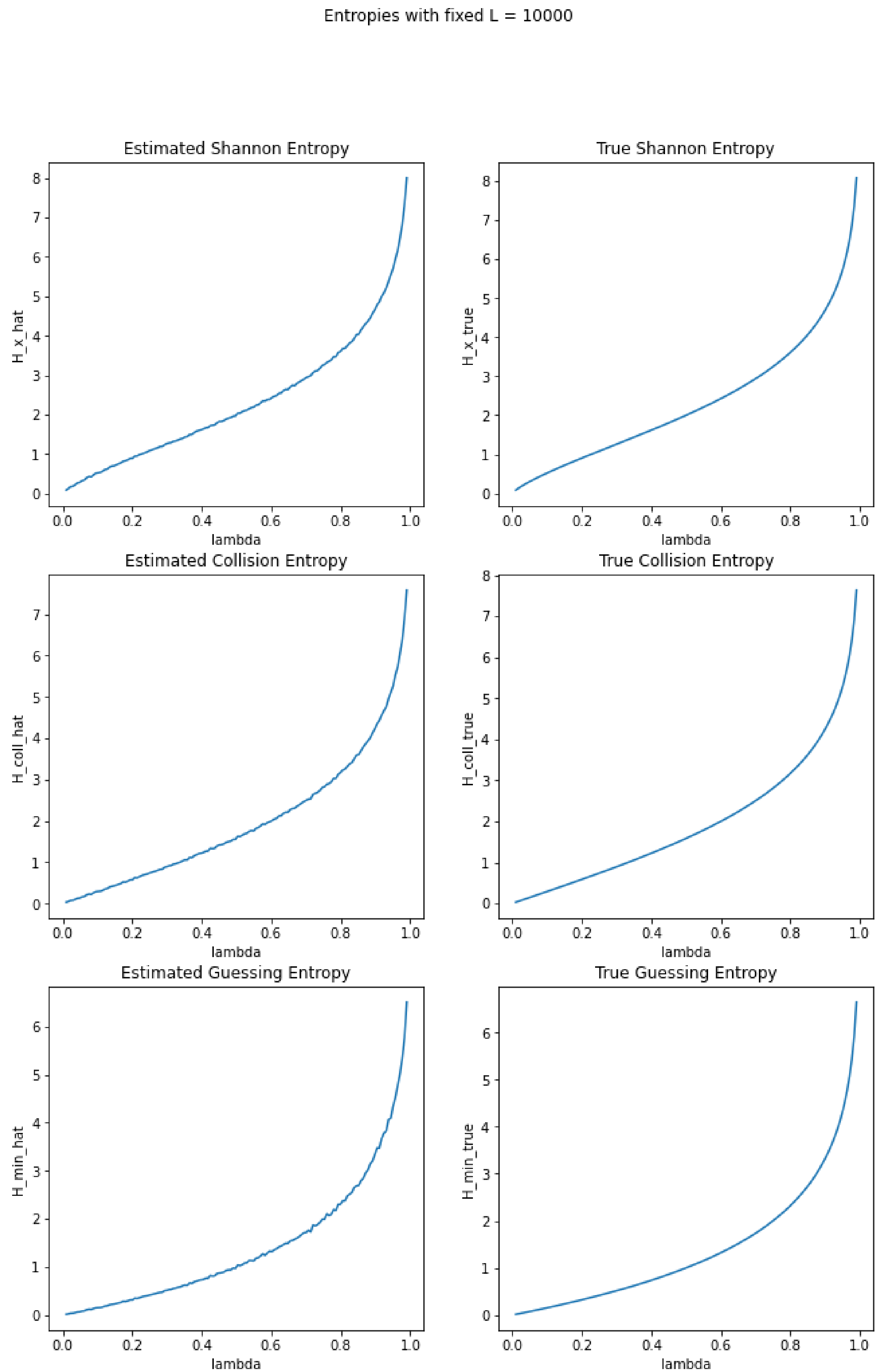

Figure 9: Estimates (left) vs True (right) Entropies comparison for a Geometric r.v. and a varying q.

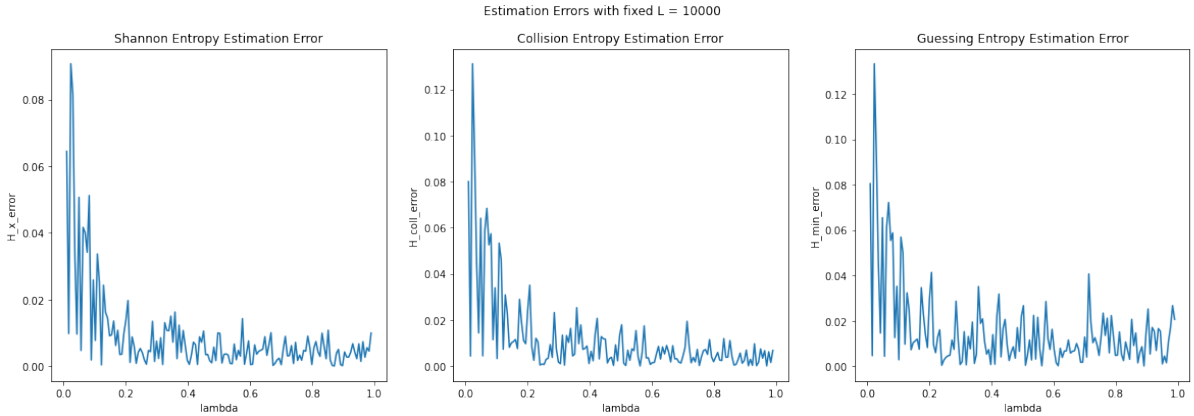

Figure 10: Relative Precision of the estimates for a Binary r.v. and a varying q.

3.3.3 Varying L

In figure 11 and 12 are plotted the entropies and relative precisions for a geometric r.v. with fixed failure probability $\lambda = 0.5$ and L ranging in $[30, 100000]$.

Note once again that:

- $H_{min}(x)$ is the harder to estimate (pay attention to the plots' scale).
- In order to obtain a relative precision similar to the Uniform and Binary case we had to increase the range of L from $[0, 10^4]$ to $[0, 10^5]$.

This is due to the fact that the geometric r.v has infinite alphabet. If we lower the failure probability λ (virtually "restricting" the alphabet) we could accept even a more restricted range for L.

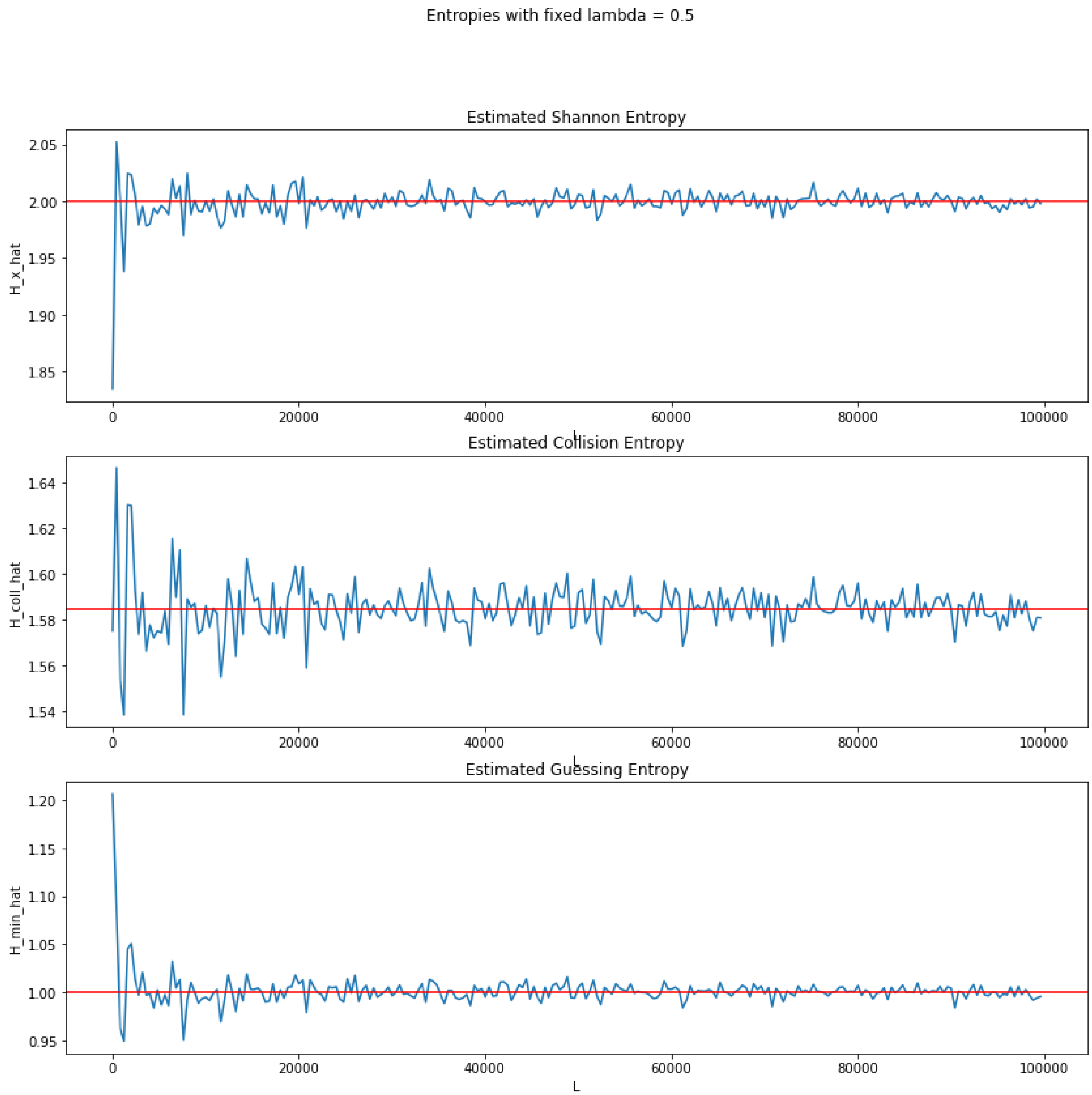

Figure 11: Estimates (blue) vs True (red) entropies comparison for a Geometric r.v., with fixed $\lambda = 0.5$ and a varying L.

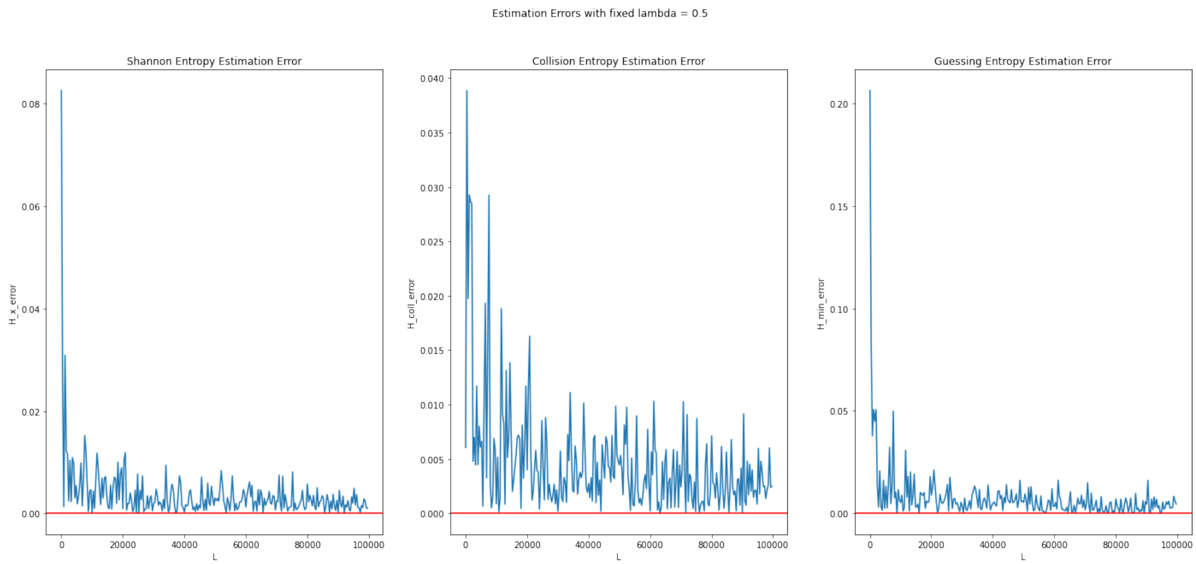

Figure 12: Relative precision of the estimates for a Geometric r.v., with fixed $\lambda = 0.5$ and a varying L.

4 Estimating Joint Quantities for x, y

I now describe the strategy I used to estimate some joint quantities of 2 empirical vectors x and y (See Section 1 for the list).

4.1 Alphabets and pmds

A_x, A_y, p_x, p_y are trivially obtained by using the same approach defined in 2.1.

The joint pmd $p_{xy}(a, b)$ cannot be found performing an outer product between p_x and p_y since we cannot assume x,y to be independent. The problem is solved by looking at the single realizations of x and y at positions $0, 1, 2, \ldots$ and by counting how many times the realizations a, b $\forall a \in A_x$, $\forall b \in A_y$ occur at the same time.

Note therefore that p_{xy} must be stored in matrix of $|A_x|$ rows and $|A_y|$ columns.

$$p_{xy} = \begin{bmatrix} p_{00} & p_{01} & p_{02} & \cdots & p_{0|A_y|} \\ p_{10} & p_{11} & \cdots & & \vdots \\ \vdots & \vdots & \ddots & & \\ p_{|A_x|0} & \cdots & & & p_{|A_x||A_y|} \end{bmatrix} \tag{18}$$

4.2 Joint Entropy $H(x, y)$

By definition, using only the joint pmd found above we obtain:

$$H(x, y) = \sum_{a \in A_x, b \in A_y} p_{xy}(a, b) \log_{1/2} p_{xy}(a, b). \tag{19}$$

4.3 Conditional Entropy $H(x|y)$

By using the joint entropy found in equation 19 and extracting the Shannon entropy of y with the same approach of paragraph 2.2 we can find the conditional entropy as:

$$H(x|y) = H(x, y) - H(y) \tag{20}$$

4.4 Relative Entropy $D(p_x||p_y)$

To estimate the relative entropy, I first check that $A_x \subset A_y$ (if not print that $D(p_x||p_y) = +\infty$), and then I simply use the definition, using the univariate pmds found in 4.1, namely:

$$D(p_x||p_y) = \sum_{a \in A_x} p_x(a) \log_2 \frac{p_x(a)}{p_y(a)} \tag{21}$$

4.5 Mutual Information $I(x;y)$

By using the Shannon entropies of x and y, and the joint entropy found above, the mutual information is given by:

$$I(x;y) = H(x) + H(y) - H(x,y) \tag{22}$$

5 Tests on the Estimates for x and y

The tests are performed by generating 2 empirical vectors x and y for each one of the following cases:

- $x \sim \mathcal{U}[1, M]$ and $y = x + z$, where $z \sim \mathcal{U}[-1, 1]$.
- $x \sim \mathcal{G}(\lambda_x)$ and $y \sim \mathcal{G}(\lambda_y)$.

5.1 Tests on Multivariate Uniform

5.1.1 Multivariate Uniform Theoretical Values

The "true" values for the entropies has been computed using the following equations.

Let's first notice that the matrix p_{xy} has M rows all equal to 0, except for 3 values (for each row) for which $p_{xy} = \frac{1}{3M}$, therefore:

$$H(x,y) = 3 \times M \times \left[\frac{1}{3M} \log_2 3M \right] = \log_2 3M. \tag{23}$$

The conditional entropy is obtained by definition by previously computed quantities as:

$$H(x|y) = H(x,y) - H(y) \tag{24}$$

The relative entropy is obtained by computing $p_x(1), \dots, p_x(M)$ and $p_y(1), \dots, p_y(M)$, as:

$$D(p_x||p_y) = 2 \times \left[\frac{1}{M} \log_2 \frac{3}{2} \right] + (M-2) \times \left[\frac{1}{M} \log_2 1 \right] = \frac{2}{M} \log_2 \frac{3}{2}. \tag{25}$$

Finally for the mutual information I used the equation:

$$I(x;y) = H(x) + H(y) - H(x,y). \tag{26}$$

For which we already have all the useful quantitites.

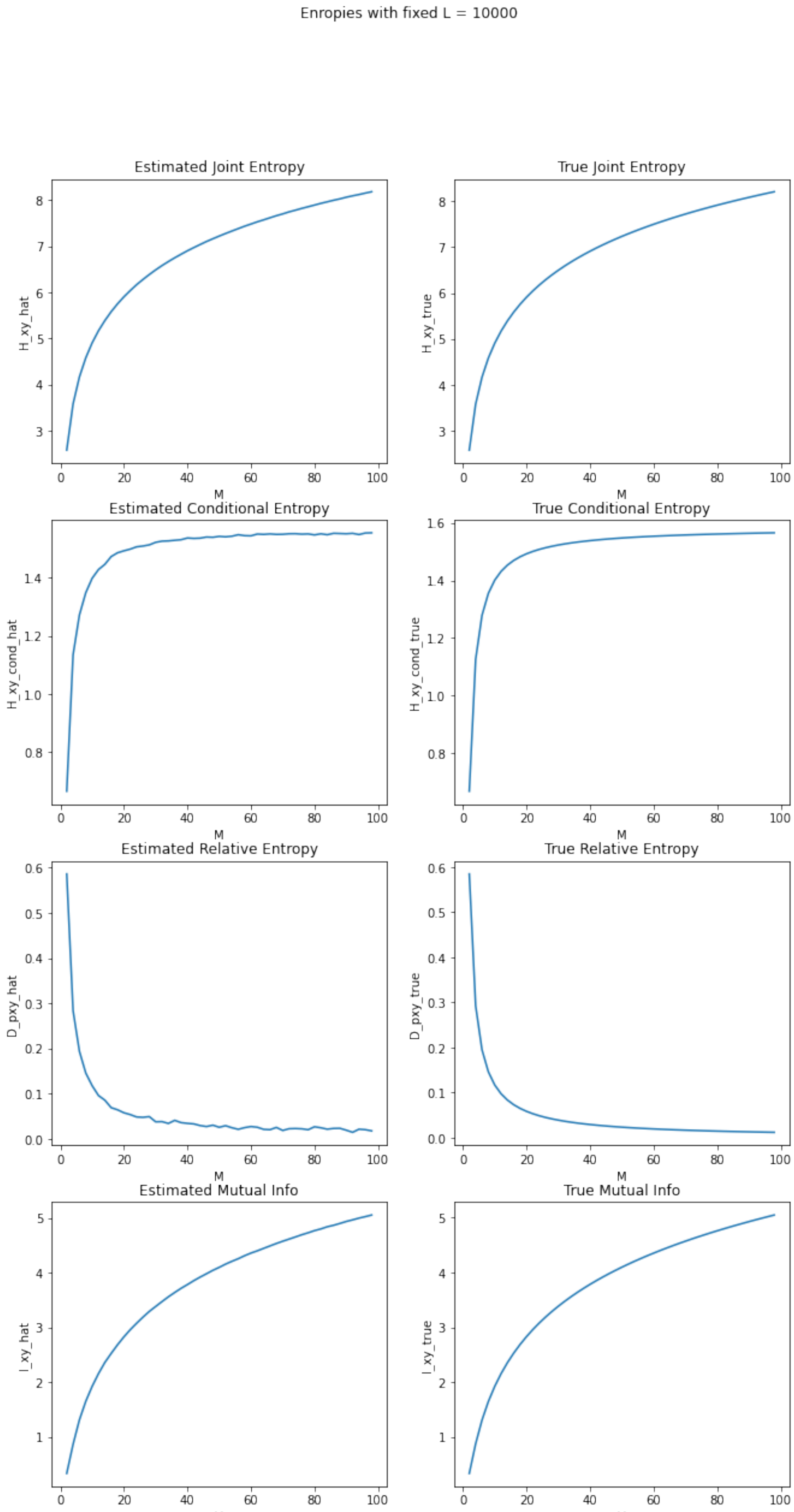

Figure 13: Estimates (left) vs True (right) Entropies comparison for x and y with a varying $|A_x| = M$.

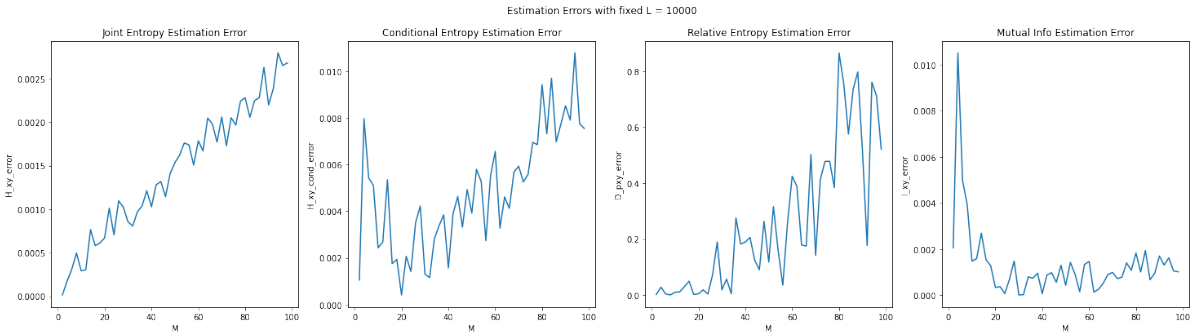

Figure 14: Relative Precision of the estimates for x and y with a varying $|A_x| = M$.

5.1.2 Varying M

In Figure 13 and 14 are plotted the entropies and the errors regarding x and y defined above as: $x \sim \mathcal{U}[1, M]$ and $y = x + z$, where $z \sim \mathcal{U}[-1, 1]$ with a varying $M \in [2, 100]$ and a fixed $L = 10000$.

Note that:

- The joint entropy $H(x, y)$ has the same shape of a uniform r.v.'s Shannon entropy.

  In facts we can interpret it as:

$$\lim_{M\to\infty} H(x, y) = \lim_{M\to\infty} \underbrace{H(x|y)}_{\substack{\text{Entropy of a Uniform} \\ \text{with } |A|=3}} + \underbrace{H(y)}_{\text{Almost a uniform}} = \log_2 3 + \log_2 M \tag{27}$$

- $D(p_x||p_y) \to 0$ as $M \to \infty$.

  In facts when $M \to \infty$ the symbols 0 and M+1 of the alphabet of y start to become less and less "important" and y becomes more and more similar (on average) to the distribution x.

- $I(x; y) \to \infty$ as $M \to \infty$.

  In facts:

$$I(x; y) = H(x) + H(y) - H(x, y) = H(x) + H(y) - H(x|y) - H(y) = H(x) - H(x|y)$$

  and hence (by exploiting what we observed above for $H(x, y)$)

$$\lim_{M\to\infty} I(x; y) = \lim_{M\to\infty} \log_2 M - \log_2 3 = +\infty$$

- It's harder to correctly estimate the mutual information when M is small. (e.g. if $M = 2$ and $|A_y| = 4 = 2M$ it's difficult to detect any correlation between x and y).
- More similarly to the univariate uniform, the 3 entropies (joint, conditional, relative) are harder to estimate when M grows.

5.1.3 Varying L

In Figure 15 and 16 are plotted the entropies and relative precisions for x and y as described above, with fixed $M = 2$ and L ranging in $[30, 10000]$.

- Note that the best estimate is the one regarding the joint entropy $H(x, y)$.

This was expected since the conditional entropy and the mutual information are obtained starting from other estimated quantitites while the relative entropy (containing the term p_x/p_y in its equation) is more prone to error propagation (particularly when M is small).

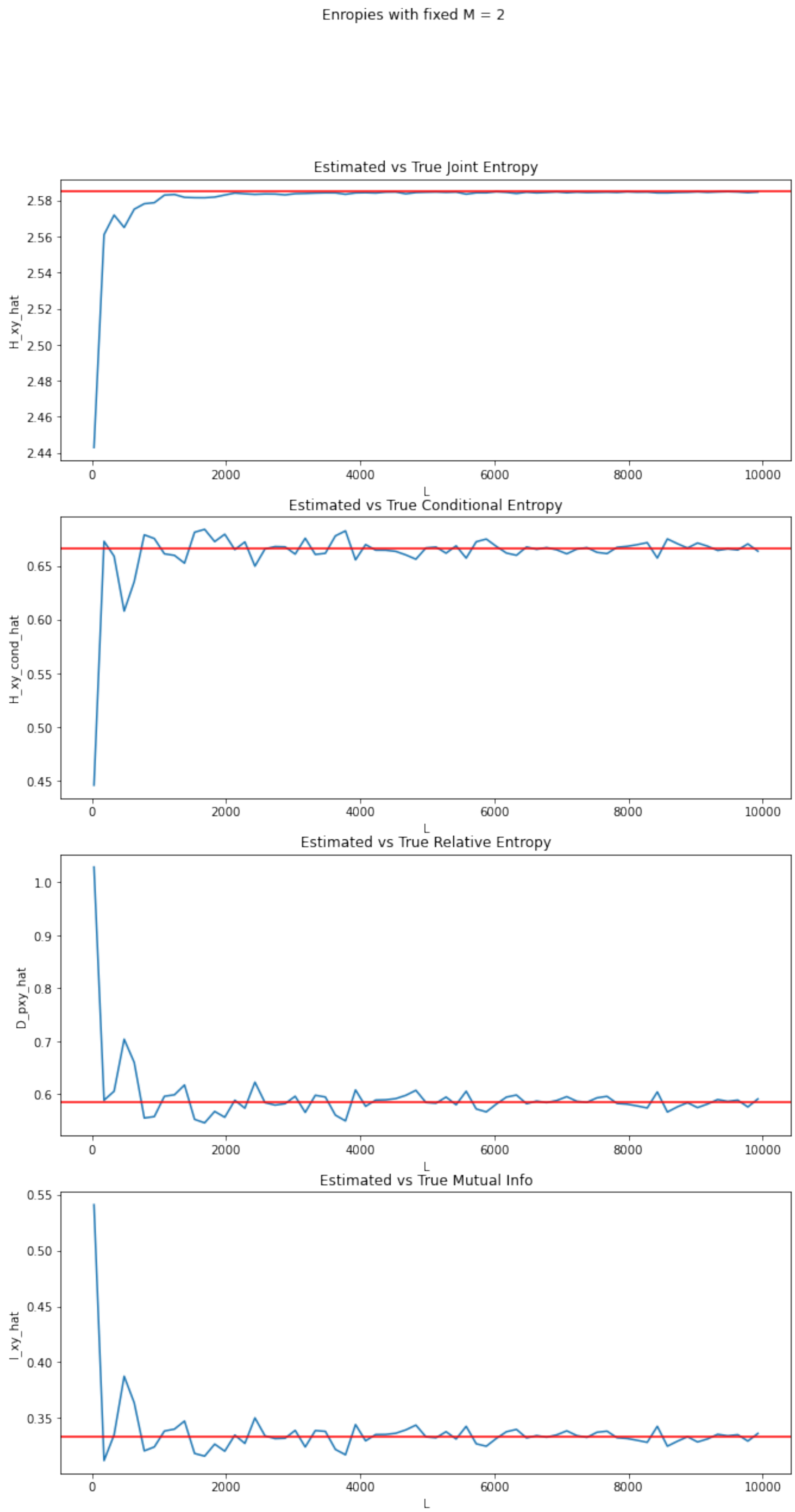

Figure 15: Estimates (blue) vs True (red) entropies comparison for x (uniform) and y, with fixed M = 2 and a varying L.

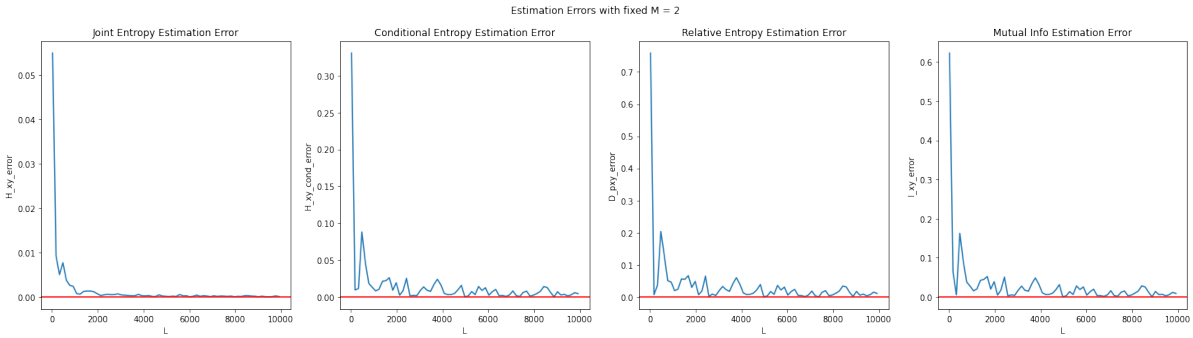

Figure 16: Relative precision of the estimates for x (uniform) and y, with fixed M = 2 and a varying L.

5.2 Tests on Multivariate Geometric

5.2.1 Multivariate Geometric Theoretical Values

We now study the joint quantities regarding x, y s.t.:

$$p_x(k) = (1-\lambda_x)\lambda_x^k$$
$$p_y(k) = (1-\lambda_y)\lambda_y^k.$$

First of all we notice that x and y are built to be indepent. This makes some computations a lot easier, in facts:

$$H(x,y) = H(x) + H(y) \tag{28}$$
$$H(x|y) = H(x) \tag{29}$$
$$D(p_x||p_y) = \sum_{k=0}^{\infty}(1-\lambda_x)\lambda_x^k \log_2 \frac{(1-\lambda_x)\lambda_x^k}{(1-\lambda_y)\lambda_y^k} \tag{30}$$
$$= (1-\lambda_x)\left[\frac{1}{1-\lambda_x}\log_2\frac{1-\lambda_x}{1-\lambda_y} + \frac{\lambda_x}{(1-\lambda_x)^2}\log_2\lambda_x - \frac{\lambda_x}{(1-\lambda_y)^2}\log_2\lambda_y\right] \tag{31}$$
$$I(x;y) = 0 \tag{32}$$

5.2.2 Varying λ_x, λ_y

In Figure 17 and 18 are plotted the entropies and the errors regarding x and y defined above. For simplicity I considered a fixed $\lambda_y = 0.8$ and a varying $\lambda_x \in [0.01, 0.99]$, while $L = 10000$ fixed.

Note that:

- Both $H(x,y)$ and $H(x|y)$ diverge as $\lambda_x \to 1$ (since we expressed them in relation to the single geometric entropies $H(x), H(y)$ which diverge).
- $D(p_x||p_y)$ is very precise for $\lambda_x < 0.8$ (see Figure 18). For $\lambda_x > 0.8$ the true $D(p_x||p_y)$ goes to ∞ while our estimate does not exist since it (wrongly) detect that $A_y \subset A_x$ (theoretically both alphabets should be of infinite cardinality).
- For high λ_x our procedure wrongly detect a peak of mutual information (estimated as $I(x;y) = H(x) + H(y) - H(x,y)$) probably due to an underestimation of the joint entropy as we can see in the first row of plots of Figure 17 as $\lambda_x \to 1$. We should increase L in order to get better estimations of $p(k)$ even for bigger values of k and be able to decrease this peak (see Figure 20).

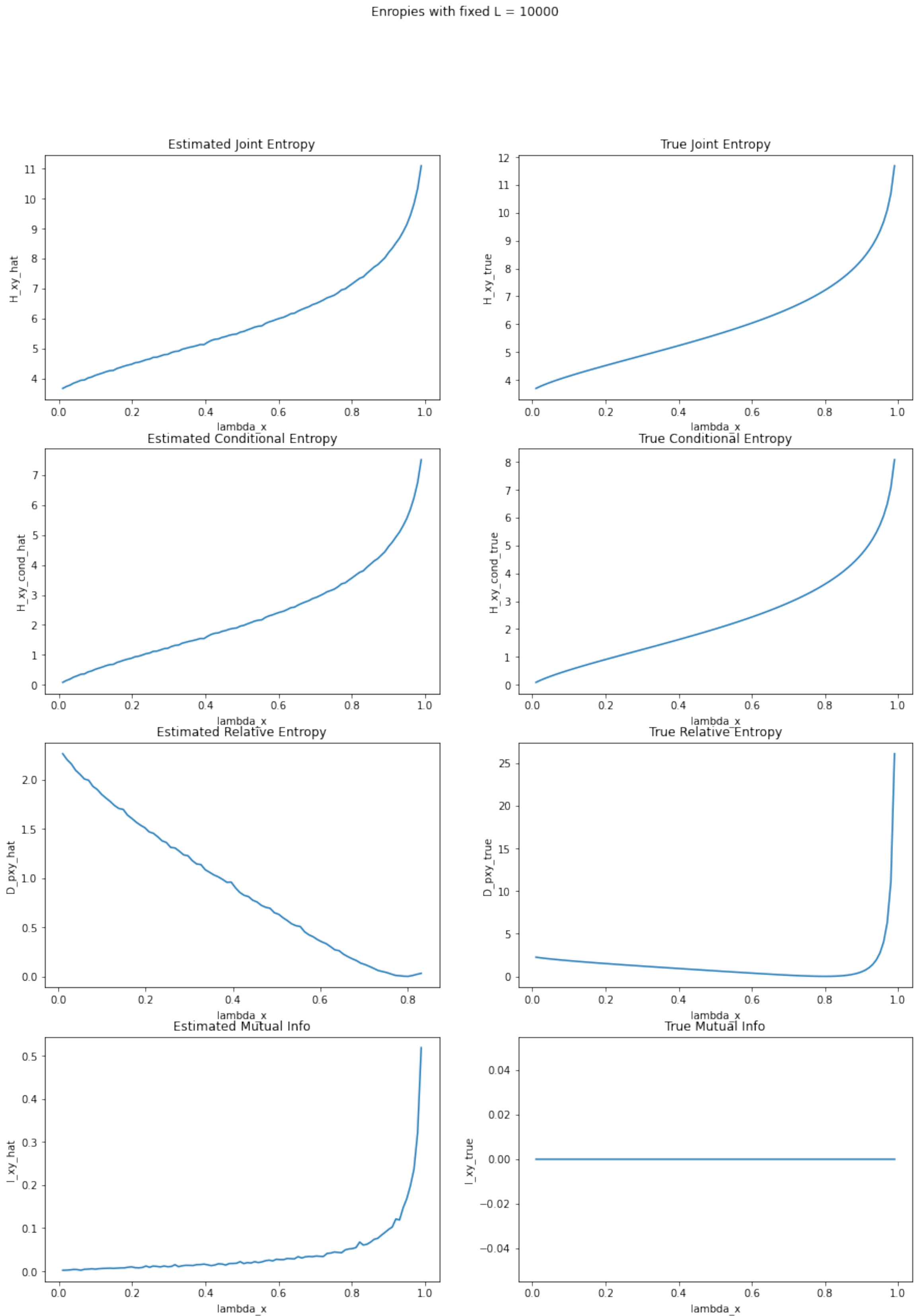

Figure 17: Estimates (left) vs True (right) Entropies comparison for x and y with a varying λ_x, fixed $\lambda_y = 0.8$, fixed $L = 10000$.

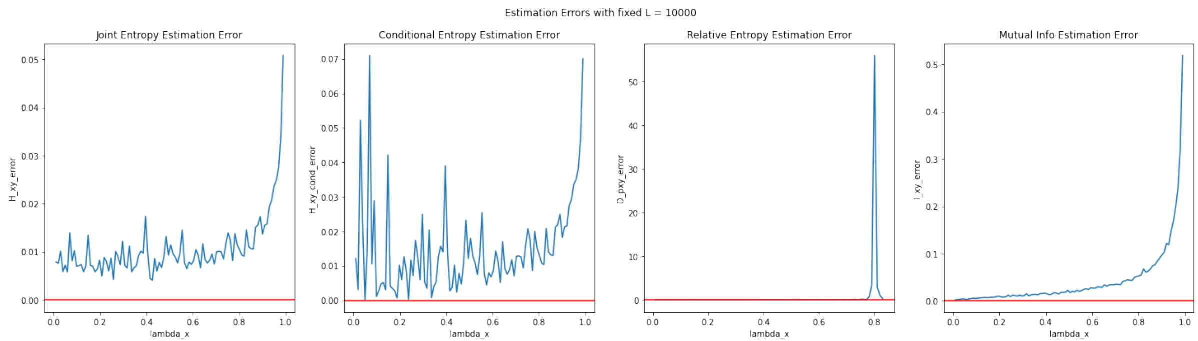

Figure 18: Relative Precision of the estimates for x and y with a varying λ_x, fixed $\lambda_y = 0.8$, fixed $L = 10000$.

5.2.3 Varying L

In Figure 19 and 20 are plotted the entropies and relative precisions for x and y as described above, with fixed $\lambda_x = 0.3$, $\lambda_y = 0.8$ and L ranging in $[30, 10000]$.

Note that:

- $\varepsilon[H(x,y) = H(x) + H(y)] < \varepsilon[H(x|y) = H(x)]$.

This seems counterintuitive but is due to the fact that is more difficult to estimate the entropy of a geometric r.v. with a small λ such as x (See Figure 12). Since we are studying the **relative** precision of $\hat{H}(x,y)$, the estimation error is "absorbed" by the larger value of $H(y)$.

- The mutual information is the one quantity which suffer the most of a small value of L (since it is based on the computation of 2 Shannon entropies and a joint one leading to a non-neglegible error propagation), but converge fast to 0 as $L \uparrow$.

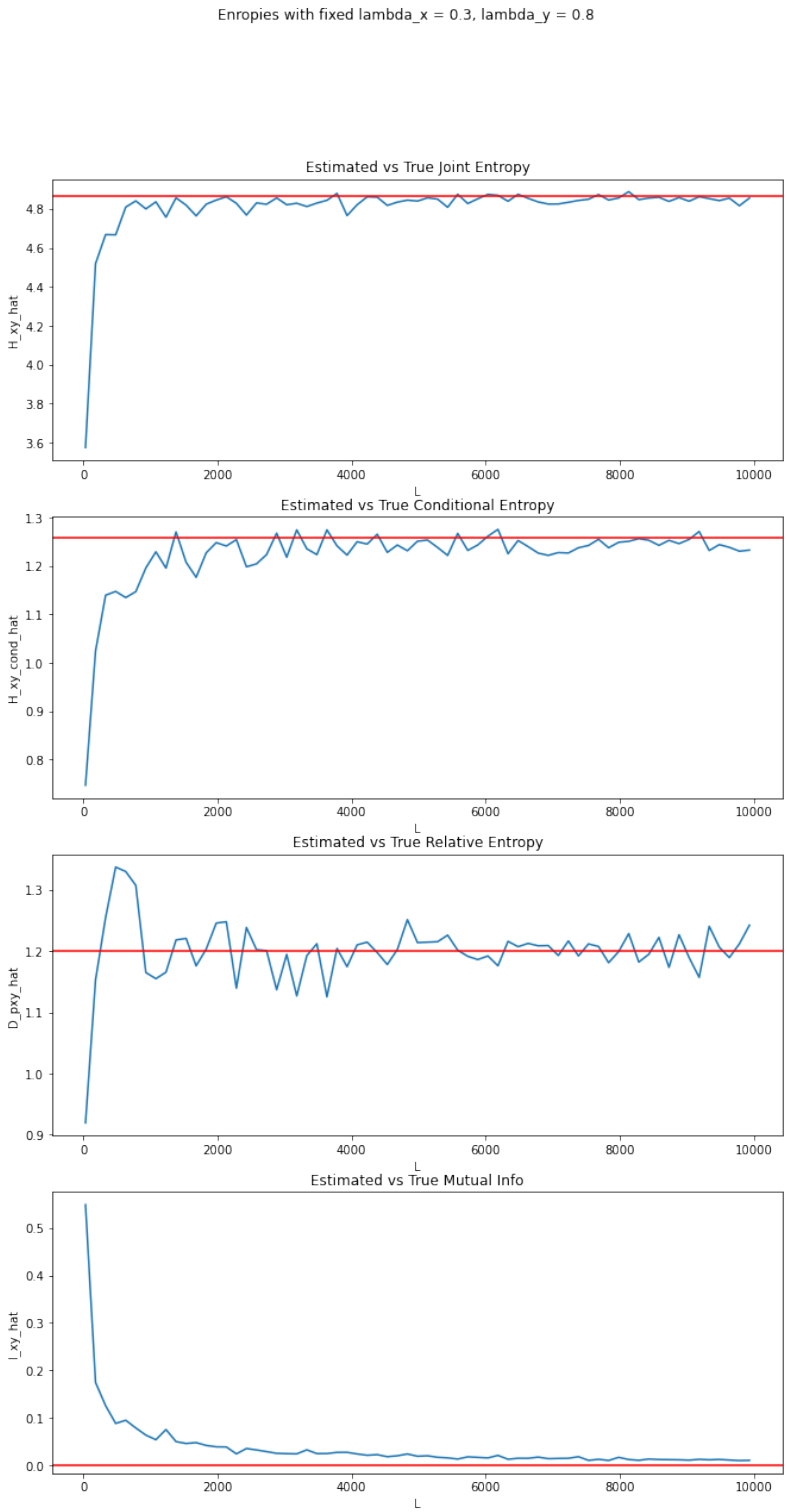

Figure 19: Estimates (blue) vs True (red) entropies comparison for x and y, with fixed $\lambda_x = 0.3$, $\lambda_y = 0.8$ and a varying L.

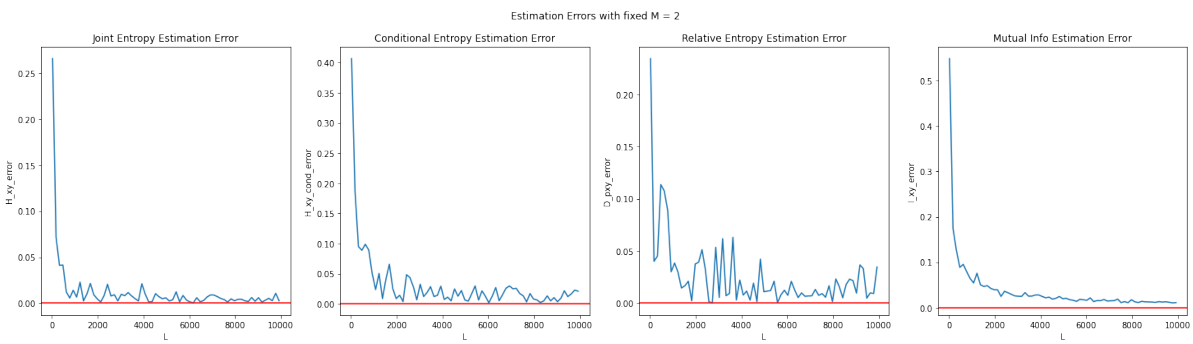

Figure 20: Relative precision of the estimates for x (uniform) and y, with fixed $\lambda_x = 0.3$, $\lambda_y = 0.8$ and a varying L.