ELEN090-2 Information And Coding Theory
Project 1: Information measures

Sébastien Laurent (s201561) - Duy Vu Dinh (s2401627)

1 Implementation

1.1 Question 1

Mathematical formula

The entropy $H(\mathcal{X})$ of a random variable $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$ from its probability distribution $P_{\mathcal{X}} = (p_1, p_2, \ldots, p_n)$ is defined as:

$$H(\mathcal{X}) \triangleq -\sum_{i=1}^{n} p_i \log_2 p_i \tag{1}$$

where:

- $p_i \in [0, 1]$ represents the probability of each possible event X_i, the i-th possible element of X.
- The logarithm $log_2 p_i$ is taken in base 2, ensuring that entropy is measured in bits (or Shannon bits).
- Since $p_i \in [0, 1]$, $log_2 p_i \leq 0$. Therefore, the negative sign ensures that the entropy is non-negative. In other words, $H(\mathcal{X}) \geq 0$.
- If some events have zero probability, i.e., $p_i = 0$, we can calculate the entropy by excluding them from the sum, which means assuming $p_i \log_2 p_i \equiv 0$, since $\lim_{p \to 0} p \log_2 p = 0$.

Explanation of implementation

The function `entropy(Px)` operates in the following manner:

- `where=(Px > 0)`: Ensures zero-probability events are ignored, implementing the mathematical limit convention.
- `-np.sum(Px * np.log2(Px))`: Computes the entropy. First, each probability p_i is multiplied by its log value $\log_2 p_i$ for all i, then we sum over all terms and negate the sum to ensure we get the non-negative entropy value.

Intuitive interpretation

Intuitively, the entropy measures the average information provided by the random variable. In other words, it measures the amount of uncertainty in a random variable.

- High entropy: The distribution is more random and the variable is highly unpredictable (e.g. uniform distribution). If $p_1 = p_2 = \ldots = p_n$, i.e. $p_i = \frac{1}{n}$, then the entropy $H(\mathcal{X})$ is maximum. This means that if all outcomes are equally likely, entropy is maximized.
- Low entropy: The outcome is more predictable. If $p_i = 1$ (one event has the probability of 1), then the entropy $H(\mathcal{X}) = 0$, i.e. there is no uncertainty.

1.2 Question 2

Mathematical formula

The joint entropy $H(\mathcal{X}, \mathcal{Y})$ of two discrete random variables $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$ and $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_m\}$ with joint probability distribution $P_{\mathcal{X},\mathcal{Y}}$ is defined as:

$$H(\mathcal{X}, \mathcal{Y}) \triangleq -\sum_{i=1}^{n} \sum_{j=1}^{m} P(X_i, Y_j) \log_2 P(X_i, Y_j) \tag{2}$$

where:

- $P(X_i, Y_j)$ is the probability of the joint event $\mathcal{X} = X_i$, $\mathcal{Y} = Y_j$.

$$P(X_i, Y_j) = P(X_i)P(Y_j|X_i) = P(Y_j)P(X_i|Y_j) = P(X_i \cap Y_j)$$

Explanation of implementation

- `where=(Pxy > 0)`: Ensures zero-probability events are ignored, implementing the mathematical limit convention.
- `-np.sum(Pxy * np.log2(Pxy))` computes the joint entropy. First, each probability p_{ij} is multiplied by its log value $\log_2 p_{ij}$, then we sum over all terms and negate the sum to ensure we get the non-negative joint entropy value.

Comparison between entropy and joint entropy

- The joint entropy simply amounts to setting $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ and applying the definition of entropy (eq. 1) to this new random variable, whose possible values are the combinations of the values of $\mathcal{X}$ and $\mathcal{Y}$.
- Entropy $H(\mathcal{X})$ measures the uncertainty of a single random variable, while joint entropy $H(\mathcal{X}, \mathcal{Y})$ measures the combined uncertainty of two random variables.

$$H(\mathcal{X}, \mathcal{Y}) = H(\mathcal{X}) + H(\mathcal{Y}|\mathcal{X}) = H(\mathcal{Y}) + H(\mathcal{X}|\mathcal{Y}) \tag{3}$$

- $H(\mathcal{X}, \mathcal{Y}) \leq H(\mathcal{X}) + H(\mathcal{Y})$ with equality iff $\mathcal{X} \perp \mathcal{Y}$.
- $H(\mathcal{X}, \mathcal{Y}) \geq \max\{H(\mathcal{X}), H(\mathcal{Y})\}$

Question 3

Mathematical formula

The conditional entropy $H(\mathcal{X}|\mathcal{Y})$ of a discrete random variable $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$ given another discrete random variable $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_m\}$, with joint probability distribution $P_{\mathcal{X},\mathcal{Y}}$, is defined as:

$$H(\mathcal{X}|\mathcal{Y}) \triangleq -\sum_{i=1}^{n}\sum_{j=1}^{m} P(X_i, Y_j) \log_2 P(X_i|Y_j) \tag{4}$$

where:

- $P(X_i, Y_j)$ is the joint probability of X_i and Y_j.
- $P(X_i|Y_j) = \frac{P(X_i,Y_j)}{P(Y_j)}$ is the conditional probability of X_i given Y_J.
- $P(Y_j)$ is the marginal probability of $\mathcal{Y}$.
- The logarithm is taken in base 2 to measure entropy in bits.

Explanation of implementation

The implementation follows Eq. (5):

- `joint_entropy(Pxy)`: Computes the joint entropy $H(\mathcal{X}, \mathcal{Y})$, introduced in question 2.
- `np.sum(Pxy, axis=0)`: Computes the marginal probability distribution $P(Y_j)$ by summing over all X_i.
- `entropy(np.sum(Pxy, axis=0))`: Computes the entropy $H(\mathcal{Y})$, introduced in question 1.

Equivalent way of computing that quantity

An equivalent formula for conditional entropy is:

$$H(\mathcal{X}|\mathcal{Y}) = H(\mathcal{X}, \mathcal{Y}) - H(\mathcal{Y}) \tag{5}$$

where:

- $H(\mathcal{X}, \mathcal{Y})$ is the joint entropy of $\mathcal{X}$ and $\mathcal{Y}$.
- $H(\mathcal{Y})$ is the entropy of $\mathcal{Y}$.

Question 4

Mathematical formula

The mutual information $I(\mathcal{X};\mathcal{Y})$ between two discrete random variables $\mathcal{X}$ and $\mathcal{Y}$ with joint probability distribution $P_{\mathcal{X},\mathcal{Y}}$ is defined as:

$$\mathcal{I}(\mathcal{X};\mathcal{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{m} P(X_i, Y_j) \log_2 \frac{P(X_i, Y_j)}{P(X_i)P(Y_j)} \tag{6}$$

where:

- $P(X_i, Y_j)$ is the joint probability distribution of $\mathcal{X}$ and $\mathcal{Y}$.
- $P(X_i) = \sum_j P(X_i, Y_j)$ is the marginal probability of $\mathcal{X}$.
- $P(Y_j) = \sum_i P(X_i, Y_j)$ is the marginal probability of $\mathcal{Y}$.

Explanation of implementation

The implementation follows Eq. (7):

- `np.sum(Pxy, axis=1)`: Computes the marginal probability distribution $P(\mathcal{X})$ by summing over Y.
- `entropy(np.sum(Pxy, axis=1))`: Computes the entropy $H(\mathcal{X})$, introduced in question 1.
- `conditional_entropy(Pxy)`: Computes $H(\mathcal{X}|\mathcal{Y})$, introduced in question 3.

Interpretation of mutual information

Mutual information $I(\mathcal{X};\mathcal{Y})$ measures the reduction in uncertainty about $\mathcal{X}$ given that we know $\mathcal{Y}$. It quantifies the dependency between $\mathcal{X}$ and $\mathcal{Y}$. In other words, Higher values of $I(\mathcal{X};\mathcal{Y})$ indicate stronger relationships (or higher correlation) between X and Y:

$$I(\mathcal{X};\mathcal{Y}) = H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Y}) = H(\mathcal{Y}) - H(\mathcal{Y}|\mathcal{X}) \tag{7}$$

- $I(\mathcal{X};\mathcal{Y}) \geq 0$ with equality iff $\mathcal{X} \perp \mathcal{Y}$.
- $I(\mathcal{X};\mathcal{Y}) \leq \min\{H(\mathcal{X}), H(\mathcal{Y})\}$ with equality of $I(\mathcal{X};\mathcal{Y}) = H(\mathcal{X})$ iff $\mathcal{Y} = g(\mathcal{X})$.

Question 5

Mathematical formula

The conditional joint entropy $H(\mathcal{X}, \mathcal{Y}|\mathcal{Z})$ quantifies the remaining uncertainty of $(\mathcal{X}, \mathcal{Y})$ given that $\mathcal{Z}$ is known. It is defined as:

$$\begin{aligned}
H(\mathcal{X}, \mathcal{Y}|\mathcal{Z}) &= -\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{l} P(X_i, Y_j, Z_k) \log_2 P(X_i, Y_j|Z_k) && (8)\\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{l} P(X_i, Y_j, Z_k) \log_2 \frac{P(X_i, Y_j, Z_k)}{P(Z_k)} && (9)\\
&= H(\mathcal{X}, \mathcal{Y}, \mathcal{Z}) - H(\mathcal{Z}) && (10)\\
&= H(\mathcal{X}|\mathcal{Z}) + H(\mathcal{Y}|\mathcal{X}, \mathcal{Z}) && (11)\\
&= H(\mathcal{Y}|\mathcal{Z}) + H(\mathcal{X}|\mathcal{Y}, \mathcal{Z}) && (12)
\end{aligned}$$

where:

- $P(X_i, Y_j, Z_k)$ is the joint probability of $(\mathcal{X}, \mathcal{Y}, \mathcal{Z})$.
- $P(X_i, Y_j|Z_k) = \frac{P(X_i, Y_j, Z_k)}{P(Z_k)}$ is the conditional probability.

The conditional mutual information $\mathcal{I}(\mathcal{X}; \mathcal{Y}|\mathcal{Z})$ measures the information shared between $\mathcal{X}$ and $\mathcal{Y}$ when $\mathcal{Z}$ is known:

$$\begin{aligned}
I(\mathcal{X}; \mathcal{Y}|\mathcal{Z}) &\triangleq H(\mathcal{X}|\mathcal{Z}) - H(\mathcal{X}|\mathcal{Y}, \mathcal{Z}) && (13)\\
&= H(\mathcal{X}|\mathcal{Z}) + H(\mathcal{Y}|\mathcal{Z}) - H(\mathcal{X}, \mathcal{Y}|\mathcal{Z}) && (14)
\end{aligned}$$

Alternatively, the conditional mutual information can be expressed as:

$$\begin{aligned}
I(\mathcal{X}; \mathcal{Y}|\mathcal{Z}) &= \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{l} P(X_i, Y_j, Z_k) \log_2 \frac{P(X_i, Y_j|Z_k)}{P(X_i|Z_k)P(Y_j|Z_k)} && (15)\\
&= \sum_{k=1}^{l} P(Z_k) I(\mathcal{X}; \mathcal{Y}|Z_k) && (16)
\end{aligned}$$

where

$$I(\mathcal{X}; \mathcal{Y}|Z_k) = \sum_{i=1}^{n}\sum_{j=1}^{m} P(X_i, Y_j|Z_k) \log_2 \frac{P(X_i, Y_j|Z_k)}{P(X_i|Z_k)P(Y_j|Z_k)} \qquad (17)$$

Explanation of implementation

Function: `cond_joint_entropy(Pxyz)`

- `np.sum(Pxyz, axis=(0,1), keepdims=True)` computes $P(Z_k)$ (marginal probability of $\mathcal{Z}$).
- `Pxyz / np.sum(Pxyz, axis=(0,1), keepdims=True)` calculates the conditional probability $P(X_i, Y_j|Z_k)$.
- `-np.sum(Pxyz * np.log2(...))` computes the sum while ignoring zero probabilities.

Function: `cond_mutual_information(Pxyz)`

- `conditional_entropy(np.sum(Pxyz, axis=1))` computes $H(\mathcal{X}|\mathcal{Z})$.
- `conditional_entropy(np.sum(Pxyz, axis=0))` computes $H(\mathcal{Y}|\mathcal{Z})$.
- `cond_joint_entropy(Pxyz)` computes $H(\mathcal{X}, \mathcal{Y}|\mathcal{Z})$.

2 Predicting the result of the Information and Coding Theory exam

2.1 Question 6

Theoretical justification

Eq. (1) provide the formulation for the entropy of an random variable $\mathcal{X}$.

Entropy $H(\mathcal{X})$ is maximized when all possible values of $\mathcal{X}$ are equally likely, meaning the probability distribution is uniform.

For a discrete random variable $\mathcal{X}$ with n possible values, Entropy $H(\mathcal{X})$ is maximized when all possible values of $\mathcal{X}$ are equally likely, meaning the probability distribution is uniform, i.e.:

$$P(\mathcal{X} = X_i) = \frac{1}{n}, \quad \forall i \in \{1, 2, ..., n\}$$

and:

$$\begin{aligned} H_{max}(\mathcal{X}) &= -\sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} \\ &= -\sum_{i=1}^{n} \frac{1}{n}(-\log_2 n) \\ &= \log_2 n \end{aligned}$$

Therefore, when a random variable follows a perfectly uniform distribution, its entropy reaches its maximum value of $\log_2(n)$. In other words, the entropy of each variable is bounded by its cardinality. Conversely, if entropy is much lower than $\log_2 n$, then one or more categories dominate, making the variable less uncertain.

Table 1 computes the maximum entropies corresponding to the cardinalities.

Table 1: Maximum entropies for different cardinalities

Cardinality ($\|\mathcal{X}\| = n$)	Max. entropy ($\log_2 n$) [bit]
2	1.000
3	1.585
4	2.000
5	2.322

Analysis

The entropy values, cardinalities, and probability distributions for the variables in the dataset is provided in Table 2.

Table 2: Entropy, cardinality, and probability distribution of each variable

Variable, $\mathcal{X}$	Entropy $H(\mathcal{X})$ [bit]	Cardinality $\|\mathcal{X}\|$	Probability distribution, $P_{\mathcal{X}}$
Weather the week before the exam	0.980	2	[0.582, 0.418]
Exam result	0.991	2	[0.555, 0.445]
Master	1.481	3	[0.503, 0.301, 0.196]
Grade for the probability class	1.489	3	[0.495, 0.305, 0.200]
Interest in the course	1.514	3	[0.411, 0.397, 0.192]
Evalens score of the course	1.516	3	[0.464, 0.323, 0.213]
Date	1.584	3	[0.347, 0.332, 0.321]
Location	1.585	3	[0.334, 0.333, 0.333]
Project grade	1.494	4	[0.518, 0.256, 0.224, 0.002]
Time spent studying	1.622	5	[0.589, 0.251, 0.065, 0.057, 0.038]
Time spent on project	1.978	5	[0.477, 0.205, 0.135, 0.128, 0.055]

Both *Weather the week before the exam* ($H \approx 0.980$ bit) and *Exam result* ($H \approx 0.991$ bit) are close to their maximum entropy, 1 bits. This indicates that both variables have an almost uniform distribution, suggesting that the weather conditions and the exam results were almost evenly split over the dataset, but the distribution of *Exam result* is closer to uniform distribution compared to the other.

Similarly, *Date* and *Location* have high entropy values (1.584 and 1.585 bit), nearly reaching their max entropy of 1.585 bit. This means they were uniformly distributed across the three options. On the other hand, *Evalens score of the course*, *Interest in the course*, *Grade for the probability class*, and *Master* have lower entropy values than the maximum entropy in the descending way. This shows that the random variables are skewed and in each variable, there is at least one category that has a higher probability than the rest. It can be clearly seen that in their probability distributions, there exists a category that has a probability of more than 40%.

Furthermore, the entropy of *Project grade* is lower than maximum entropy, meaning some project grades were more frequent than others.

For variables with five possible values, the maximum entropy is $\log_2(5) \approx 2.322$ bit. However, the observed entropy for *Time spent studying* ($H \approx 1.622$ bit) and *Time spent on project* ($H \approx 1.978$ bit) is notably lower, indicating that some categories occur more frequently than others, deviating from a uniform distribution.

Question 7

The conditional entropies of *Exam result* given each of the other variables are shown in Table 3.

Table 3: Conditional entropy of *Exam result* given each of the other variables

Variable, $\mathcal{Y}$	Conditional entropy $H(\texttt{Exam result}\vert\mathcal{Y})$ [bit]
Time spent studying	0.865
Interest in the course	0.910
Time spent on project	0.914
Project grade	0.918
Evalens score of the course	0.920
Date	0.971
Grade for the probability class	0.981
Weather the week before the exam	0.990
Master	0.991
Location	0.991

(a) When the conditioning variable is *Interest in the course*

In this section, the variables of *Exam result* and of *Interest in the course* are denoted as $\mathcal{X}$ and $\mathcal{Y}$.

The conditional entropy of *Exam result* given *Interest in the course* is $H(\mathcal{X}|\mathcal{Y}) \approx 0.910$ bit, which is lower than the individual entropy of *Exam result*, $H(\mathcal{X}) \approx 0.991$ bit. In addition, this conditional entropy is also relatively low compared to other variables.

This suggests that students' interest in the course has a meaningful impact on their exam performance. Intuitively, students who are interested in the course are more likely to succeed, while those who are not interested might have a higher probability of failure. However, this conditional entropy is not extremely low, meaning that while *Interest in the course* is informative, it does not fully determine exam results, i.e. other factors also play a role.

(b) When the conditioning variable is *Master*

In this section, the variables of *Exam result* and of *Master* are denoted as $\mathcal{X}$ and $\mathcal{Y}$.

In contrast, the conditional entropy of *Exam result* given *Master* approximates the individual entropy of *Exam result*, $H(\mathcal{X}|\mathcal{Y}) \approx 0.991$ bit $\approx H(\mathcal{X})$. This means that *Master* is independent of *Exam result*, i.e. a student's field of study does not provide significant information about their likelihood of success or failure.

Question 8

The mutual information between *Location* and *Evalens score of the course* is 0.0002 bit, which is very close to zero. This indicates that knowing the *Location* of exam provides almost no information about the *Evalens score of the course*, i.e. exam location does not significantly influence students' evaluations of the course.

The mutual information between *Time spent on project* and *Project grade* is 0.685 bit. This value is significantly higher, indicating a stronger relationship between the two variables. However, the relationship is not perfectly deterministic, some students might spend a lot of time but still receive low grades, while others might do well with less effort.

Question 9

The mutual information of *Exam result* when knowing each of the other variables is described in Table 4.

Table 4: Mutual information of each of the other variables when knowing *Exam result*

Variable $\mathcal{Y}$	Mutual information $I(\texttt{Exam result}; \mathcal{Y})$ [bit]
Time spent studying	0.126
Interest in the course	0.081
Time spent on project	0.077
Project grade	0.073
Evalens score of the course	0.071
Date	0.021
Grade for the probability class	0.010
Weather the week before the exam	0.001
Master	0.0001
Location	0.0001

The highest mutual information value with "Exam result" is *Time spent studying*, 0.126 bit. This indicates that knowing a student's `Time spent studying` reduces the uncertainty about their `Exam result` more than any other variable in the dataset. Therefore, if the student can access only one variable, he should choose *Time spent studying* to maximize his ability to predict exam outcomes.

Besides, conditional entropy could also be used as an alternative measure. Conditional entropy $H(\texttt{Exam result} \mid \mathcal{Y})$ quantifies the remaining uncertainty in the exam result after observing a specific variable $\mathcal{Y}$. A lower conditional entropy means that the variable provides more certainty about the outcome of the exam. Since mutual information and conditional entropy are related according to Eq. (7), using conditional entropy should lead to the same variable choice (the one that minimizes $H(\texttt{Exam result} \mid \mathcal{Y})$). So, with the lowest conditional entropy, *Time spent studying* is also the top choice in terms of the conditional entropy.

Therefore, there is no difference between the two approaches, maximizing mutual information and minimizing conditional entropy.

Question 10

The conditional mutual information of *Exam result* and each of the other variables given *Interest in the course* is described in Table 5.

Table 5: Conditional entropy and mutual information of *Exam result* and each other variable given *Interest in the course*

Variable, $\mathcal{Y}$	Conditional mutual information [bit] $I(\texttt{Exam result}; \mathcal{Y} \mid \texttt{Interest in the course})$
Time spent studying	0.045
Evalens score of the course	0.036
Date	0.030
Grade for the probability class	0.011
Project grade	0.009
Weather the week before the exam	0.002
Location	0.002
Time spent on project	0.001
Master	0.001

- Even after accounting for *Interest in the course*, *Time spent studying* remains the most informative variable, with the conditional mutual information of $I = 0.045$ bit. This confirms that the ranking of the best variable does not change when conditioning on Interest in the course. *Time spent studying* consistently provides the most predictive information about *Exam result*, both with and without conditioning on *Interest in the course.*
- The conditional mutual information of *Time spent studying* is lower than its original mutual information (0.045 bit vs. 0.126 bit), meaning that *Interest in the course* already explains some of the variation in *Exam result.*
- Other variables, such as *Time spent on project* and *Project grade*, have very low conditional mutual information values, meaning they provide very little additional predictive power once *Interest in the course* is known.

Final Conclusion

Even when considering *Interest in the course* as known, Time spent studying remains the best predictor of Exam result. This shows that while a student's interest in the course plays a role in their performance, their actual study effort is still the most important factor in determining their exam success.

3 Playing with information theory-based strategy

Let us define:

- The discrete random variables: $\mathrm{S}_i = \{1, \dots, \mathcal{C}\}\ \forall i \in \{1, \dots, \mathcal{S}\}$. The random variable S_i represents the value of the i^{th} slot.
- The discrete random variable: $\mathrm{M} = (\mathrm{S}_1, \dots, \mathrm{S}_{\mathcal{S}})$. This random variable represents the state of the entire Mastermind game.
- The set of feedback i-j-C, which is the set of feedback, where the color of the i^{th} slot is correctly guessed and is equal to j. Similarly, we define the set of feedback i-j-I, when the colors are incorrectly guessed. A set of feedback can be viewed as an event, as it reduces the sample space in a manner similar to how an event does. We will differentiate a set of feedback F from its corresponding event (F).
- An operator $\#(\cdot)$ that take n events and return the number of Mastermind configurations that are compatible with these n events, i.e. for n events $\mathrm{A}_1, \dots, \mathrm{A}_n$, $\#(\mathrm{A}_1, \dots, \mathrm{A}_n)$ returns the number of Mastermind configurations that are compatible $(\mathrm{P}(\mathrm{M} = i|\mathrm{A}_1, \dots, \mathrm{A}_n) > 0)$ with $\mathrm{A}_1, \dots, \mathrm{A}_n$. We denote by $\#(/)$ the total number of Mastermind configurations, which is the total number of values that M can take according to its definition. We assume that all Mastermind configurations are equiprobable.

Question 11

First, let us derive an important formula that we will use in the following questions. For events $\mathrm{A}_1, \dots, \mathrm{A}_n$ and events $\mathrm{B}_1, \dots, \mathrm{B}_m$, we use the definition of the conditional probability

$$\mathrm{P}(\mathrm{A}_1, \dots, \mathrm{A}_n|\mathrm{B}_1, \dots, \mathrm{B}_m) = \frac{\mathrm{P}(\mathrm{A}_1, \dots, \mathrm{A}_n, \mathrm{B}_1, \dots, \mathrm{B}_m)}{\mathrm{P}(\mathrm{B}_1, \dots, \mathrm{B}_m)}$$

We know that all Mastermind configurations are equiprobable, so

$$\mathrm{P}(\mathrm{A}_1, \dots, \mathrm{A}_n, \mathrm{B}_1, \dots, \mathrm{B}_m) = \frac{\#(\mathrm{A}_1, \dots, \mathrm{A}_n, \mathrm{B}_1, \dots, \mathrm{B}_m)}{\#(/)}$$

and

$$\mathrm{P}(\mathrm{B}_1, \dots, \mathrm{B}_m) = \frac{\#(\mathrm{B}_1, \dots, \mathrm{B}_m)}{\#(/)}$$

Meaning that

$$\mathrm{P}(\mathrm{A}_1, \dots, \mathrm{A}_n|\mathrm{B}_1, \dots, \mathrm{B}_m) = \frac{\#(\mathrm{A}_1, \dots, \mathrm{A}_n, \mathrm{B}_1, \dots, \mathrm{B}_m)}{\#(\mathrm{B}_1, \dots, \mathrm{B}_m)}$$

To compute the entropy of a discrete variable $\mathcal{X}$ that has n possible values, we use the following formula

$$\mathrm{H}(\mathcal{X}) = -\sum_{i=1}^{n} \mathrm{P}(\mathcal{X} = i) \log_2 \mathrm{P}(\mathcal{X} = i)$$

Let us apply this formula to the S_i. We suppose that the distribution of the S_i is uniform, because there is no reason why a color should appear more than another one, so $P(\mathrm{S}_i = j) = 1/5 \; \forall j \in \{1, 2, 3, 4, 5\}$.

$$\begin{aligned}
\mathrm{H}(\mathrm{S}_i) &= -\sum_{j=1}^{5} \mathrm{P}(\mathrm{S}_i = j) \log_2 \mathrm{P}(\mathrm{S}_i = j) \\
&= -\sum_{j=1}^{5} (1/5) \log_2(1/5) \\
&= \log_2(5) \\
&= 2.32 \text{ Shannon bits}
\end{aligned}$$

To compute the entropy of M, we use the fact that M has $\#(/)$ possible values, the fact that all Mastermind configurations are equiprobable and the formula of the entropy

$$\begin{aligned}
\mathrm{H}(\mathrm{M}) &= -\sum_{i=1}^{\#(/)} \mathrm{P}(\mathrm{M} = i) \log_2 \mathrm{P}(\mathrm{M} = i) \\
&= -\sum_{i=1}^{\#(/)} \frac{\#(\mathrm{M} = i)}{\#(/)} \log_2 \frac{\#(\mathrm{M} = i)}{\#(/)}
\end{aligned}$$

By definition of M, we know that $\#(\mathrm{M} = i) = 1$, and because there are five slots, which can take five equiprobable values independently of each other, we can compute that $\#(/) = 5^5$, so

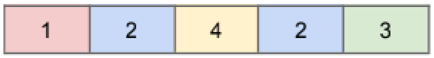
$$\begin{aligned}
\mathrm{H}(\mathrm{M}) &= -\sum_{i=1}^{\#(/)} \frac{\#(\mathrm{M} = i)}{\#(/)} \log_2 \frac{\#(\mathrm{M} = i)}{\#(/)} \\
&= -\sum_{i=1}^{5^5} \frac{1}{5^5} \log_2 \left(\frac{1}{5^5}\right) \\
&= \log_2 \left(5^5\right) \\
&= 5 \log_2(5) \\
&= 11.61 \text{ Shannon bits}
\end{aligned}$$

We observe that

$$\mathrm{H}(\mathrm{M}) = \mathrm{H}(\mathrm{S}_1, \mathrm{S}_2, \mathrm{S}_3, \mathrm{S}_4, \mathrm{S}_5) = \mathrm{H}(\mathrm{S}_1) + \mathrm{H}(\mathrm{S}_2) + \mathrm{H}(\mathrm{S}_3) + \mathrm{H}(\mathrm{S}_4) + \mathrm{H}(\mathrm{S}_5)$$

and we know that this equality only holds if and only if the S_i are independent, so we can conclude that the S_i are independent.

Question 12



1 correctly placed peg at the first position

Feedback 1-1-C is given. Its corresponding event is (1-1-C). The entropy of S_1 has changed, because the probability distribution of the color of the first slot has changed. In fact, the probability of one of the color is equal to 1 and equal to 0 for all the other colors. Using the formula of the entropy and basic understanding of what the entropy is (no uncertainty ⇒ no entropy), we know that

$$\mathrm{H}(\mathrm{S}_1|(\text{1-1-C})) = 0 \text{ Shannon bit}$$

As for the other slots, since we received no feedback about them, we know they can not be blue, yellow, or green. Therefore, slots 2, 3, 4, and 5 can only be red or brown, with equal probability, as there is no reason to favor one color over the other. Also, the entropy of slots 2, 3, 4, and 5 are equal, because their situations are identical. We compute their entropy using the formula $\forall i \in \{2, 3, 4, 5\}$

$$\begin{aligned}
\mathrm{H}(\mathrm{S}_i|(\text{1-1-C})) &= -\sum_{j=1}^{5} \mathrm{P}(\mathrm{S}_i = j|(\text{1-1-C})) \log_2 \mathrm{P}(\mathrm{S}_i = j|(\text{1-1-C})) \\
&= -(\mathrm{P}(\mathrm{S}_i = 1|(\text{1-1-C})) \log_2 \mathrm{P}(\mathrm{S}_i = 1|(\text{1-1-C})) + \mathrm{P}(\mathrm{S}_i = 5|(\text{1-1-C})) \log_2 \mathrm{P}(\mathrm{S}_i = 5|(\text{1-1-C}))) \\
&= -2\frac{1}{2} \log_2 \left(\frac{1}{2}\right) \\
&= \log_2(2) \\
&= 1 \text{ Shannon bit}
\end{aligned}$$

To compute $\mathrm{H}(\mathrm{M}|(\text{1-1-C}))$, we need to compute $\#(\mathrm{M} = i, (\text{1-1-C}))$ and $\#((\text{1-1-C}))$. Firstly, there are five slots, and to be compatible with 1-1-C, the first slot of the configuration must be equal to 1, while the four other slots can take two equiprobable values independently of each other, so

$$\#((\text{1-1-C})) = 1 \cdot 2^4$$

Let us define $\mathrm{M}_{\text{1-1-C}}$ the set of all values of M that are compatible with 1-1-C. By definition, the cardinality of $\mathrm{M}_{\text{1-1-C}} = \#((\text{1-1-C}))$. By definition of M, we deduce that

$$\#(\mathrm{M} = i, (\text{1-1-C})) = \begin{cases} 1 & \text{if } i \in \mathrm{M}_{\text{1-1-C}} \\ 0 & \text{if } i \notin \mathrm{M}_{\text{1-1-C}} \end{cases}$$

Let us compute the entropy of $\mathrm{M}|(\text{1-1-C})$

$$
\begin{aligned}
\mathrm{H}(\mathrm{M}|(\text{1-1-C})) &= -\sum_{i=1}^{\#(/)} \mathrm{P}(\mathrm{M}=i|(\text{1-1-C}))\log_2 \mathrm{P}(\mathrm{M}=i|(\text{1-1-C})) \\
&= -\sum_{i\in \mathrm{M}_{\text{1-1-C}}} \mathrm{P}(\mathrm{M}=i|(\text{1-1-C}))\log_2 \mathrm{P}(\mathrm{M}=i|(\text{1-1-C})) \\
&\quad -\sum_{i\notin \mathrm{M}_{\text{1-1-C}}} \mathrm{P}(\mathrm{M}=i|(\text{1-1-C}))\log_2 \mathrm{P}(\mathrm{M}=i|(\text{1-1-C})) \\
&= -\sum_{i\in \mathrm{M}_{\text{1-1-C}}} \frac{\#(\mathrm{M}=i,(\text{1-1-C}))}{\#((\text{1-1-C}))}\log_2 \frac{\#(\mathrm{M}=i,(\text{1-1-C}))}{\#((\text{1-1-C}))} \\
&\quad -\sum_{i\notin \mathrm{M}_{\text{1-1-C}}} \frac{\#(\mathrm{M}=i,(\text{1-1-C}))}{\#((\text{1-1-C}))}\log_2 \frac{\#(\mathrm{M}=i,(\text{1-1-C}))}{\#((\text{1-1-C}))} \\
&= -\sum_{i\in \mathrm{M}_{\text{1-1-C}}} \frac{1}{\#((\text{1-1-C}))}\log_2 \frac{1}{\#((\text{1-1-C}))}
\end{aligned}
$$

Now, we use the value of the $\#((\text{1-1-C}))$ and the cardinality of $\mathrm{M}_{\text{1-1-C}}$

$$
\begin{aligned}
\mathrm{H}(\mathrm{M}|(\text{1-1-C})) &= -\sum_{i\in \mathrm{M}_{\text{1-1-C}}} \frac{1}{\#((\text{1-1-C}))}\log_2 \frac{1}{\#((\text{1-1-C}))} \\
&= -\#((\text{1-1-C}))\frac{1}{\#((\text{1-1-C}))}\log_2 \frac{1}{\#((\text{1-1-C}))} \\
&= \log_2 \#((\text{1-1-C})) \\
&= \log_2\left(2^4\right) \\
&= 4\log_2(2) \\
&= 4 \text{ Shannon bits}
\end{aligned}
$$

Again, we have the following equality

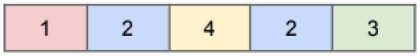
$$
\begin{aligned}
\mathrm{H}(\mathrm{M}|(\text{1-1-C})) &= \mathrm{H}(\mathrm{S}_1,\mathrm{S}_2,\mathrm{S}_3,\mathrm{S}_4,\mathrm{S}_5|(\text{1-1-C})) \\
&= \mathrm{H}(\mathrm{S}_1|(\text{1-1-C})) + \mathrm{H}(\mathrm{S}_2|(\text{1-1-C})) + \mathrm{H}(\mathrm{S}_3|(\text{1-1-C})) + \mathrm{H}(\mathrm{S}_4|(\text{1-1-C})) + \mathrm{H}(\mathrm{S}_5|(\text{1-1-C}))
\end{aligned}
$$

so we can conclude that the S_i remain independent after the guess.

To compute the quantity of information that the guess brought, we just compute the difference between the entropy of the game before and after the feedback

$$
\mathrm{H}(\mathrm{M}) - \mathrm{H}(\mathrm{M}|(\text{1-1-C})) = 5\log_2(5) - 4 = 7.61 \text{ Shannon bits}
$$

Question 13



1 correctly placed peg at the first position and 1 incorrectly placed peg at the second position

Feedback 1-1-C and 2-2-I are occurring. The event corresponding to the conjunction of feedback 1-1-C and feedback 2-2-I is (1-1-C, 2-2-I). The probability distribution of S_1 does not change from the question 12, so its entropy does

$$H(S_1|(1\text{-}1\text{-}C, 2\text{-}2\text{-}I)) = H(S_1|(1\text{-}1\text{-}C)) = 0 \text{ Shannon bit}$$

First, notice that we received no feedback for slots three and five, so there is no yellow or green in the code. Also, there can only be one blue in the code; Otherwise, slot four would also be incorrect (4-2-I).

The probability distribution of S_2 and S_4 change. Indeed, they can only be red or brown, with equal probability, as there is no reason to favor one color over the other. Also, the entropy of slots two and four are equal, because their situations are identical. We compute their entropy using the formula $\forall i \in \{2, 4\}$

$$\begin{aligned}
H(S_i|(1\text{-}1\text{-}C, 2\text{-}2\text{-}I)) =& -\sum_{j=1}^{5} P(S_i = j|(1\text{-}1\text{-}C, 2\text{-}2\text{-}I)) \log_2 P(S_i = j|(1\text{-}1\text{-}C, 2\text{-}2\text{-}I)) \\
=& -(P(S_i = 1|(1\text{-}1\text{-}C, 2\text{-}2\text{-}I)) \log_2 P(S_i = 1|(1\text{-}1\text{-}C, 2\text{-}2\text{-}I))+ \\
& \quad P(S_i = 5|(1\text{-}1\text{-}C, 2\text{-}2\text{-}I)) \log_2 P(S_i = 5|(1\text{-}1\text{-}C, 2\text{-}2\text{-}I))) \\
=& -2\frac{1}{2}\log_2\left(\frac{1}{2}\right) \\
=& \log_2(2) \\
=& 1 \text{ Shannon bit}
\end{aligned}$$

Now, we compute $H(S_3|(1\text{-}1\text{-}C, 2\text{-}2\text{-}I))$ and $H(S_5|(1\text{-}1\text{-}C, 2\text{-}2\text{-}I))$. Since the situations of S_3 and S_5 are symmetrical (We have the exact same informations about slots three and five), we know that

$$H(S_3|(1\text{-}1\text{-}C, 2\text{-}2\text{-}I)) = H(S_5|(1\text{-}1\text{-}C, 2\text{-}2\text{-}I))$$

To compute $H(S_3|(1\text{-}1\text{-}C, 2\text{-}2\text{-}I))$, we need to compute

$$P(S_3 = i|(1\text{-}1\text{-}C, 2\text{-}2\text{-}I)) \ \forall i \in \{1, 2, 3, 4, 5\}$$

We know that they can not be yellow or green, so

$$P(S_3 = 3|(1\text{-}1\text{-}C, 2\text{-}2\text{-}I)) = P(S_3 = 4|(1\text{-}1\text{-}C, 2\text{-}2\text{-}I)) = 0$$

Again, we can not distinguish color brown and red in this case, so

$$P(S_3 = 1|(1\text{-}1\text{-}C, 2\text{-}2\text{-}I)) = P(S_3 = 5|(1\text{-}1\text{-}C, 2\text{-}2\text{-}I)) = \frac{1}{2}P(S_3 \neq 2|(1\text{-}1\text{-}C, 2\text{-}2\text{-}I))$$

Let us compute $P(S_3 = 2|(1\text{-}1\text{-}C, 2\text{-}2\text{-}I))$

$$P(S_3 = 2|(1\text{-}1\text{-}C, 2\text{-}2\text{-}I)) = \frac{\#(S_3 = 2, (1\text{-}1\text{-}C, 2\text{-}2\text{-}I))}{\#((1\text{-}1\text{-}C, 2\text{-}2\text{-}I))}$$

Let us compute $\#((\text{1-1-C}, \text{2-2-I}))$. 1-1-C and 2-2-I imply that S_1 is red, that S_2 and S_4 are either red, or brown, and that $(S_3 = 2) \oplus (S_5 = 2)$. There is only one possibility for the value of S_1, two equiprobable values for S_2 and S_4 and the number of pairs (S_3, S_5) that contain exactly one color blue is equal to 4, because one slot will be blue and the other will be either red or brown (blue-red, blue-brown, red-blue, brown-blue). We conclude that

$$\#((\text{1-1-C}, \text{2-2-I})) = 1 \cdot 2 \cdot 2 \cdot 4$$

We compute $\#(S_3 = 2, \text{1-1-C}, \text{2-2-I})$ in the same way. The only difference is that the number of pairs (S_3, S_5) that contain exactly one blue slot is equal to 2, because S_3 takes the value 2 and S_5 will be either red or brown (blue-red, blue-brown).

$$\#(S_3 = 2, (\text{1-1-C}, \text{2-2-I})) = 1 \cdot 2 \cdot 2 \cdot 2$$

and so

$$P(S_3 = 2|(\text{1-1-C}, \text{2-2-I})) = \frac{\#(S_3 = 2, (\text{1-1-C}, \text{2-2-I}))}{\#((\text{1-1-C}, \text{2-2-I}))} = \frac{1 \cdot 2 \cdot 2 \cdot 2}{1 \cdot 2 \cdot 2 \cdot 4} = \frac{1}{2}$$

Because

$$P(S_3 \neq 2|(\text{1-1-C}, \text{2-2-I})) + P(S_3 = 2|(\text{1-1-C}, \text{2-2-I})) = 1$$

we know that

$$P(S_3 \neq 2|(\text{1-1-C}, \text{2-2-I})) = 1 - P(S_3 = 2|(\text{1-1-C}, \text{2-2-I})) = 1 - \frac{1}{2} = \frac{1}{2}$$

and that

$$P(S_3 = 1|(\text{1-1-C}, \text{2-2-I})) = P(S_3 = 5|(\text{1-1-C}, \text{2-2-I})) = \frac{1}{2}P(S_3 \neq 2|(\text{1-1-C}, \text{2-2-I})) = \frac{1}{4}$$

Now, we just use the entropy formula

$$\begin{aligned}
H(S_3(|\text{1-1-C}, \text{2-2-I})) &= -\sum_{i \in \{1,2,3,4,5\}} P(S_3 = i|(\text{1-1-C}, \text{2-2-I})) \log_2 P(S_3 = i|(\text{1-1-C}, \text{2-2-I})) \\
&= -\sum_{i \in \{1,2,5\}} P(S_3 = i|(\text{1-1-C}, \text{2-2-I})) \log_2 P(S_3 = i|(\text{1-1-C}, \text{2-2-I})) \\
&= -\left(\frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) \\
&= -\frac{1}{2}\left(\log_2\left(\frac{1}{4}\right) + \log_2\left(\frac{1}{2}\right)\right) \\
&= \frac{1}{2}(\log_2(4) + \log_2(2)) \\
&= \frac{1}{2}(2 + 1) \\
&= 1.5 \text{ Shannon bit}
\end{aligned}$$

To compute $\mathrm{H}(\mathrm{M}|(\text{1-1-C}, \text{2-2-I}))$, we need the values of $\#(\mathrm{M} = i, (\text{1-1-C}, \text{2-2-I}))$ and $\#((\text{1-1-C}, \text{2-2-I}))$. We already know

$$\#((\text{1-1-C}, \text{2-2-I})) = 1 \cdot 2 \cdot 2 \cdot 4 = 16$$

Let us define $\mathrm{M}_{\text{1-1-C,2-2-I}}$ the set of all values of M that are compatible with 1-1-C and 2-2-I. By definition, the cardinality of $\mathrm{M}_{\text{1-1-C,2-2-I}} = \#((\text{1-1-C}, \text{2-2-I}))$. By definition of M, we deduce that

$$\#(\mathrm{M} = i, (\text{1-1-C}, \text{2-2-I})) = \begin{cases} 1 & \text{if } i \in \mathrm{M}_{\text{1-1-C,2-2-I}} \\ 0 & \text{if } i \notin \mathrm{M}_{\text{1-1-C,2-2-I}} \end{cases}$$

Let us compute the entropy of $\mathrm{M}|(\text{1-1-C}, \text{2-2-I})$. We proceed in the exact same way as in question 12.

$$\begin{aligned} \mathrm{H}(\mathrm{M}|(\text{1-1-C}, \text{2-2-I})) &= -\sum_{i=1}^{\#(/)} \mathrm{P}(\mathrm{M} = i|(\text{1-1-C}, \text{2-2-I})) \log_2 \mathrm{P}(\mathrm{M} = i|(\text{1-1-C}, \text{2-2-I})) \\ &\dots \\ &= \log_2 \#((\text{1-1-C}, \text{2-2-I})) \\ &= \log_2(16) \\ &= 4 \text{ Shannon bits} \end{aligned}$$

Let us compute the sum of the entropy of the slots

$$\begin{aligned} &\mathrm{H}(\mathrm{S}_1|(\text{1-1-C}, \text{2-2-I})) + \mathrm{H}(\mathrm{S}_2|(\text{1-1-C}, \text{2-2-I})) + \mathrm{H}(\mathrm{S}_3|(\text{1-1-C}, \text{2-2-I})) \\ &\quad + \mathrm{H}(\mathrm{S}_4|(\text{1-1-C}, \text{2-2-I})) + \mathrm{H}(\mathrm{S}_5|(\text{1-1-C}, \text{2-2-I})) = 0 + 1 + 1.5 + 1 + 1.5 = 5 \text{ Shannon bits} \end{aligned}$$

As we can see, the sum of the entropy of the five slots is not equal to the entropy of the entire game, so we can deduce that the five slots are not independent anymore, due to the conjunction of 1-1-C and 2-2-I. Also, we can see that the entropy of the game is smaller than the sum of the entropy of the five slots. This is expected, because for any n random variables $\mathcal{X}_1, \dots, \mathcal{X}_n$

$$\mathrm{H}(\mathcal{X}_1, \dots, \mathcal{X}_n) \leq \sum_{i=1}^{n} \mathrm{H}(\mathcal{X}_i)$$

with the equality holding when the n variables are independent. Let us compute the quantity of information that the guess brought

$$\mathrm{H}(\mathrm{M}) - \mathrm{H}(\mathrm{M}|(\text{1-1-C}, \text{2-2-I})) = 5\log_2(5) - 4 = 7.61 \text{ Shannon bits}$$

As we can see, the information gain in question 13 is equal to the one in question 12. It might seem unexpected that we gained the same quantity of information in both cases because the feedback from question 13 appears to provide more information (the feedback contains two statements, instead of one), and we might even be tempted to write

$$\mathrm{H}(\mathrm{M}|\text{1-1-C}, \text{2-2-I}) \leq \mathrm{H}(\mathrm{M}|\text{1-1-C})$$

with the equality holding when $\mathrm{M} \perp$ 2-2-I|1-1-C. However, it make no sense, because 1-1-C and 2-2-I are feedback and not events. In consequence, the set of configurations compatible with 1-1-C and 2-2-I is not a subset of the set of configurations compatible with 1-1-C (e.g. red-red-blue-red-red is compatible with 1-1-C and 2-2-I, but not with only 1-1-C), while it should be the case if 1-1-C and 2-2-I were truly events. Because the gain in entropy is equal in both questions, we can conclude that we have as much uncertainty about the code with both feedback.

Question 14

Obviously, the maximum entropy can only be reach when we have not made any guess, because guesses can only reduce entropy by providing information.

In this case, we have $\mathcal{S}$ slots that can take $\mathcal{C}$ values independently of each other, so $\#(/) = \mathcal{C}^{\mathcal{S}}$. Now, we define

$$p_i = \mathrm{P}(\mathrm{M} = i)$$

$$\begin{aligned}
\mathrm{H}(\mathrm{M}) &= -\sum_{i=1}^{\#(/)} \mathrm{P}(\mathrm{M}=i)\log_2 \mathrm{P}(\mathrm{M}=i) \\
&= -\sum_{i=1}^{\#(/)} p_i \log_2 p_i \\
&= \mathrm{H}_{\#(/)}\left(p_1, \ldots, p_{\#(/)}\right)
\end{aligned}$$

We know a theorem about the entropy function

$$\mathrm{H}_n(p_1, \ldots, p_n) \leq \log_2(n), \text{ with equality } \Leftrightarrow \forall i : p_i = \frac{1}{n}$$

We apply it in our case to prove that

$$\mathrm{H}(\mathrm{M}) = \mathrm{H}_{\#(/)}\left(p_1, \ldots, p_{\#(/)}\right) \leq \log_2 \#(/) = \log_2 \mathcal{C}^{\mathcal{S}} = \mathcal{S}\log_2 \mathcal{C}$$

So if all Mastermind configurations are equiprobable, as we supposed in the previous questions, the entropy is maximal and is equal to $\mathcal{S}\log_2\mathcal{C}$.

As expected, both the number of colors and the number of slots make the entropy of the game increases. Something interesting to point out is the fact that an increase of the number of slots has a greater impact on the maximum entropy than an increase of the number of colors: the maximum entropy increases linearly with the number of slots and logarithmically with the number of colors.

Question 15

We would like to solve the general setting, so we must take into account the previous feedback. To do so, we totally ignore code which are not compatible with previous feedback. We denote $\mathrm{M_{PF}}$, the set of all configurations which are compatible with the previous feedback. $|\mathrm{M_{PF}}|$ denotes the cardinality of $\mathrm{M_{PF}}$.

Given a code and a guess, each slot can be assigned to one of three values: Correct (C), Incorrect (I), or Nothing (N). When the opponent communicates a feedback, he communicates a $\mathcal{S}$-tuple $\{\mathrm{C},\mathrm{I},\mathrm{N}\}^{\mathcal{S}}$. However, given a guess, multiple of these $\mathcal{S}$-tuple can be equivalent. For example, at question 13, $(\mathrm{C},\mathrm{I},\mathrm{N},\mathrm{N},\mathrm{N})$ and $(\mathrm{C},\mathrm{N},\mathrm{N},\mathrm{I},\mathrm{N})$ are equivalent. We define $\mathrm{F}(\mathrm{m}, g)$, the set of equivalent $\mathcal{S}$-tuples that can be given to a guess g, when the code is m. We define $\mathrm{F(g)} = \{\mathrm{F(m,g)}|\mathrm{m} \in \mathrm{M_{PF}}\}$. An information-based approach would be to make the guess that minimizes the uncertainty about the code given the feedback.

$$\arg\min_{\mathrm{g}} \mathrm{H(M|F(M,g))}$$

$\mathrm{F(M,g)}$ is a random variable, since it is a function of the random variable M. For two random variables $\mathcal{X}$ and $\mathcal{Y}$, we know that (equivalent of bayes formula for entropy)

$$\mathrm{H}(\mathcal{X}|\mathcal{Y}) = \mathrm{H}(\mathcal{Y}|\mathcal{X}) + \mathrm{H}(\mathcal{X}) - \mathrm{H}(\mathcal{Y})$$

We apply this formula to our case

$$\arg\min_{\mathrm{g}} \mathrm{H(M|F(M,g))} = \arg\min_{\mathrm{g}} \left(\mathrm{H(F(M,g)|M)} + \mathrm{H(M)} - \mathrm{H(F(M,g))}\right)$$

If we know M, then $\mathrm{F(M,g)}$ is known, so $\mathrm{H(F(M,g)|M)} = 0$, and because H(M) does not depend on g, we can ignore it.

$$\arg\min_{\mathrm{g}} \mathrm{H(M|F(M,g))} = \arg\min_{\mathrm{g}} -\mathrm{H(F(M,g))} = \arg\max_{\mathrm{g}} \mathrm{H(F(M,g))}$$

We obtain a simple rule: if we want to maximize our information gain on the code, we must choose a guess that maximize the expected quantity of information in the feedback. Using the formula of the entropy

$$\mathrm{H(F(M,g))} = -\sum_{f \in \mathrm{F(g)}} \mathrm{P}(\mathrm{F(M,g)} = f) \log_2 \mathrm{P}(\mathrm{F(M,g)} = f)$$

And we can compute

$$\mathrm{P}(\mathrm{F(M,g)} = f) = \frac{\#(\mathrm{F(M,g)} = f)}{|\mathrm{M_{PF}}|}$$

We can compute $\mathrm{H(F(M,g))}$ for all possible guess and pick the one that lead to the greatest value. $\mathrm{F(m,g)}$ can be easily (not efficiently) computed for each m by verifying which feedback is compatible with m and g. Then, $\mathrm{F(g)}$ can be easily obtained from the $\mathrm{F(m,g)}$. Finally, $\#(\mathrm{F(M,g)} = f)$ can be computed by iterating on the set of configurations $\mathrm{M_{PF}}$ and using the $\mathrm{F(m,g)}$ previously computed. This naive approach is inefficient, but it is sound and complete.