Statistical physics and information theory for systems with local constraints

Zhang, Q.

Chapter 5

Information theory with coupled sources under ensemble nonequivalence

Abstract

Information theory is build to describe information transmission and storage in different systems. Restricted by the initial setting of the electronic communication system, the traditional information theory is based on a fundamental assumption that the signal generation by those information sources has an identical probability distribution and is independent of time. However, recent research on nervous systems and social networks shows that the information flows in those systems are generated by the numerous interacting units, and the signal generation in those systems is under temporal dependencies. It means the classical information theory based on the i.i.d assumption cannot deal with the coupled sources with temporal and spatial dependencies in non-artificial communication systems.

Motivated by the recent works on systems with local constraints in statistical physics, a generalization of information theory with coupled sources is built to find the limits of information transmission and storage based on the descriptions of information sequences with statistical ensembles under local constraints in this work. We find that the microcanonical ensemble description or the *Boltzmann entropy* is closer to the real limit of information storage than the canonical ensemble description with soft constraints or *Shannon entropy*. We also find that the classical information theory is a particular case of the canonical ensemble description when the dependencies are homogeneous. Moreover, the effectivity of classical information theory only holds when the microcanonical ensemble description and the canonical ensemble description of the signal generation are under the ensemble equivalence. Our result also shows that the finite temporal dependences of units in the information source are not enough to break the ensemble equivalence. The ensemble nonequivalence is formed by the extensive spatial interactions among all the units [1].

[1] This chapter is based on the coming paper:
Qi Zhang, Diego Garlaschelli, "Information theory with coupled source under ensemble nonequivalence" (2021)

5.1 Introduction

The classical information theory established in 1948 by Shannon is build to estimate the information-theoretical bounds of the information storage and the information transmission in communication systems [33]. This theory is based on an essential assumption that the information source works independently with an identical probability distribution (i.i.d) in the process of signal generation. According to this assumption, Shannon found that almost all the information generated by the i.i.d. information source $\mathbf{x}$ is carried by a set of equiprobable sequences $\{x_1, x_2, \cdots, x_n\}$, which is named as typical set $T_\epsilon^{(n)}$. Therefore, the smallest space needs to store the information generated by the information source is equal to the logarithm of the cardinality of the typical set $\ln |T_\epsilon^{(n)}|$. This limit of information storage will converge to the $n \times s(\mathbf{x})$ as a function of Shannon entropy $s(\mathbf{x})$ of the information sources when $(n \to \infty)$ the length of sequences goes to infinite [29]. This typicality is the asymptotic equipartition property (AEP). It also can be found in statistical physics, which shows that the equilibrium behaviour of a system in the thermodynamic limit is determined by typical microstates [8].

The signal generation under the i.i.d. assumption as a stationary process ignores the possible temporal and spatial dependencies between the units in information sources. However, in natural systems, especially when there are multivariates in the information source, these dependencies generally exist. For example, in the vertebrate retina, the activity of neurons is determined by pairwise correlations among neurons, and the limited energy can be used for each neuron simultaneously [69]. The pairwise correlations are spatial interactions among all the neurons. The finite energy that can be used by each neuron in the whole process of signal generation is the temporal constraint. The spatial and temporal dependencies also exist in the changes of cars' flow in the urban traffic networks [70] and fluctuations of the stock market [3]. These heterogeneous dependencies among the units in the information source make it impossible to find the limit of information storage by the classical AEP [71, 30]. Also, it may affect the symbol rate used to reliable transport the information through different channels. Thus, we need a new theory to describe the signal generation with spatial and temporal dependencies and find the information-theoretical bounds.

According to the classical information theory, the information generated by the source is carried by information sequences, which are used to record the behaviour of units in information sources. So even when the signal generation of the multivariate source is under spatial and temporal dependencies, the information generated by it is still carried by the multivariate information sequences $\{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_n\}$. Thus, to find the information-theoretical bounds of those non-stationary signal generations, we should focus on the information sequences, not the information sources.

However, those sequences with heterogeneous dependencies and the increased length are impossible to be described by random variables with finite outcomes, but those macroscopic properties are analogue with the that in states of thermodynamic systems. Both of them have numerous interacting units, and the numbers of units go to infinite in the thermodynamic limit. It means statistical ensembles in physics can be used to describe the multivariate and heterogeneous dependent information sequences, to find its information-theoretical bounds [1, 27].

In statistical physics, systems with different macroscopic properties need to be described by different ensembles. The microcanonical ensemble is used to describe systems with fixed total energy E^*. The canonical ensemble is used to describe systems with fixed temperature $\beta = 1/kT$ [1, 49, 27]. The two ensembles will conjugate with each other by setting the parameter β in the canonical ensemble equal to β^*, which makes the average value of the total energy in the canonical ensemble equal to the fixed total energy in the microcanonical ensemble ($\langle E \rangle = E^*$) [15].

Normally, in the thermodynamic limit, the two conjugate ensemble descriptions are believed to be equivalent. The microcanonical ensemble can be replaced by the canonical ensemble, which is mathematically easy to calculate. This one phenomenon is called ensemble equivalence (EE) [8]. However, in the past decades, the breakdown of EE also has been observed in various physical systems [39, 25, 21]. Especially when there are extensive local constraints in the system, the EE breaks in the whole parameter space of this system [5, 27]. Therefore, when statistical ensembles are used to describe the information sequences with different macroscopic properties, their information-theoretical bounds are affected by the possible appearance of the ensemble nonequivalence (EN).

In this work, a matrix ensemble with local constraints is used to describe the information sequences that are generated in the signal generation with heterogeneous dependencies. These heterogeneous dependencies are quantified by the total correlation (multi-information) among units in sequences [72]. We find that the classical information theory is a particular case of the canonical ensemble description with soft constraints. We also prove that the effectivity of the classical AEP in information theory is based on the EE of the signal generation. Most importantly, we find that the EN in the non-stationary process is led by the variable spatial interactions among units in the source, not the finite temporal dependence.

5.2 Ensemble described information sequences

In traditional statistical physics, statistical ensembles are under global constraints such as the fixed total energy and fixed temperature. The interactions among all units are homogeneous. However, this assumption breaks when the statistical ensembles are used to describe the information sequences generated by the information source with heterogeneous dependent units since the interactions among those units are homogeneous and time variant. It means the description of the information sequences needs the statistical ensemble with local constraints [27].

The information sequences are generated by the information source with heterogeneous spatial interactions and temporal dependencies. If the information source has m finite units, then the information generated by it in n times sampling should be recorded by the $m \times n$ matrix $\mathbf{X}$. Each unit x_{ji} represents the state of unit j

in information source at time i, and matrix $\mathbf{X}$ represents a particular state of the information sequences.

According to the definition of matrix should above, the spatial dependence among m units in the information source at different time can be modelled by the *column local constraints* $\vec{c} = [c_1, c_2, \cdots, c_i, \cdots, c_n]$, where c_i is the sum of all the units in column i of matrix $\mathbf{X}$ as $c_i = \sum_{j=1}^{m} x_{ij}$. The temporal dependence of each variable in the information sources is modelled by the *row local constraints* $\vec{r} = [r_1, r_2, \cdots, r_j, \cdots, r_m]$, where each $r_j = \sum_{i=1}^{n} x_{ji}$ is the sum of all the element in the jth row of matrix $\mathbf{X}$. In nervous systems, r_j represents the total energy theta can be used by the neuron X_j in the whole signal generation, c_i represents the total energy that can be used by m interacting units in time i.

The relationship between local constraints in the matrix ensemble and the dependencies in ensemble sequences is shown in FIG.5.1.

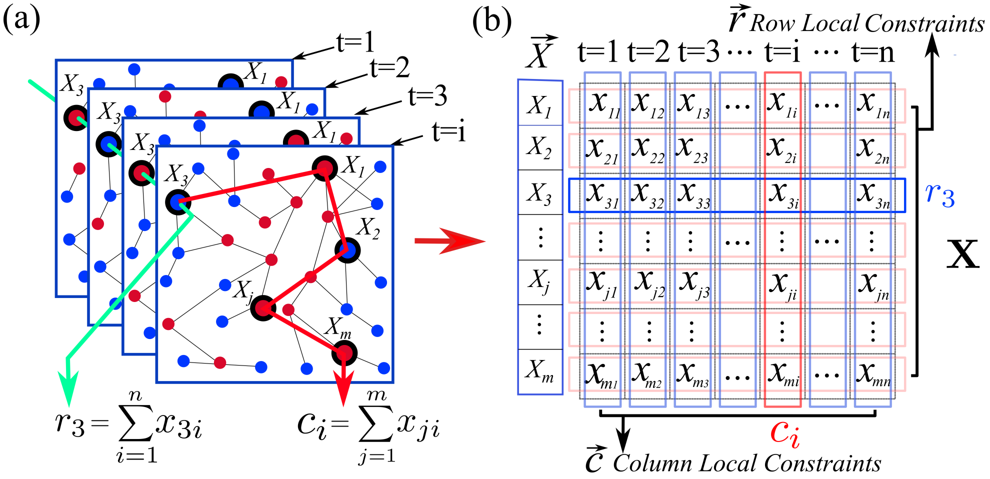
(a)
t=1
t=2
t=3
t=i
X_1
X_2
X_3
X_j
X_m
$r_3 = \sum_{i=1}^{n} x_{3i}$
$c_i = \sum_{j=1}^{m} x_{ji}$
(b)
$\vec{r}$ Row Local Constraints
$\vec{X}$
t=1 t=2 t=3 ⋯ t=i ⋯ t=n
X_1 X_2 X_3 ⋮ X_j ⋮ X_m
x_{11} x_{12} x_{13} ⋯ x_{1i} ⋯ x_{1n}
x_{21} x_{22} x_{23} x_{2i} x_{2n}
x_{31} x_{32} x_{33} x_{3i} x_{3n}
x_{j1} x_{j2} x_{j3} x_{ji} x_{jn}
x_{m1} x_{m2} x_{m3} ⋯ x_{mi} ⋯ x_{mn}
r_3
$\mathbf{X}$
c_i
$\vec{c}$ Column Local Constraints

Figure 5.1. The status of m units in different times show in (a). The red and blue colour of each node represents it is active or non-active. The selected nodes (the big nodes with the black circle as margin) are used as units in the information source. The green line across different layers represents temporal dependence. The red line among nodes in each layer represents the spatial interaction among them. The localization of the two dependencies in the matrix shows in (b). The spatial interactions among all the units in the information source at time i are quantified by the column local constraints c_i. The temporal dependence of unit j is represented by the row local constraints r_j.

When the dependence is homogeneous, the information sequences should be modelled by the matrix ensemble with global constraints as $t^* = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ji}$, which is the sum of all elements in the sequence [27].

The statistical ensemble is a probability distribution of all possible states in a specific thermodynamic system. The macroscopic property of constraints determines the value of probability for each state. For example, when the constraints are hard, states of it are equiprobable. When the constraints are soft, the probability of each state is different. The two different kinds of constraints represent two different ways to describe the signal generation under dependencies. When the constraints are hard, we need the microcanonical ensemble. Otherwise, when the constraints are soft, we need the canonical ensemble [15].

5.2.1 Canonical ensemble description

When heterogeneous dependencies in sequences are *soft*, the constraints of each microscopic configuration are different. The average value of the constraints for all the sequences is equal to the hard constraints in the conjugate microcanonical ensemble. Then the sequences will be described by the canonical ensemble.

The probability of each state in the canonical ensemble is a parameter solution under the realization of the average constraints and the maximization of Shannon entropy

$$P_{\text{can}}(\mathbf{X}|\vec{\beta}) = e^{-H(\mathbf{X},\vec{\beta})}/Z(\vec{\beta}). \tag{5.1}$$

Vector $\vec{\beta}$ is an extension of the maximum likelihood parameter β base on extensive local constraints in it [27].

The partition function $Z(\vec{\beta})$ is a normalization constant, which collect the exponential function of all the possible configurations of the sequences $\mathbf{X}$ as $Z(\vec{\beta}) = \sum_{\mathbf{X}\in\mathcal{X}} e^{-H(\mathbf{X},\vec{\beta})}$. The symbol $H(\mathbf{X},\vec{\beta})$ represents the *Hamiltonian* of a sequence $\mathbf{X}$. It is a liner combination of constraint and maximum likelihood parameter, $H(\mathbf{X},\vec{\beta}) = \vec{C}(\mathbf{X}) \cdot \vec{\beta}$ [15].

The microcanonical ensemble and canonical ensemble are conjugate with each other by setting the parameter $\vec{\beta} = \vec{\beta}^*$, to make the average value of constraints $\langle\vec{C}(\mathbf{X})\rangle$ in the canonical ensemble is equal to hard constraints $\vec{C}^*$ in the microcanonical ensemble as

$$\langle\vec{C}(\mathbf{X})\rangle = \sum_{\mathbf{X}\in\mathcal{X}} P_{\text{can}}(\mathbf{X}|\vec{\beta}^*)\vec{C}(\mathbf{X}) = \vec{C}^*, \tag{5.2}$$

where $\mathcal{X}$ represents the collection of all the possible information sequences.

As constraints $\vec{C}(\mathbf{X})$ is the sum of elements in each column and row, the probability of a sequence in the conjugate canonical ensemble is equal to the product of the probability of each unit in each sequence as

$$P_{\text{can}}(\mathbf{X}|\vec{\beta}^*) = \prod_{i=1}^{n}\prod_{j=1}^{m} \frac{e^{-x_{ij}\beta^*_{ij}}}{\sum_{x_{ij}\in\S_{ij}} e^{-x_{ij}\beta^*_{ij}}}. \tag{5.3}$$

The value of $P_{\text{can}}(\mathbf{X}|\vec{\beta}^*)$ is governed by the parameter $\vec{\beta}^*$ and the value of each units x_{ij} in sequence $\mathbf{X}$. The symbol $\S_{ij}$ represents the collection of all the possible values of x_{ij}.

Then, according to the classical information theory, it is easy to find that the smallest space to store the information carried by canonical ensemble sequences is equal to the Shannon entropy of it as

$$S_{\text{can}} = \sum_{\mathbf{X}\in\mathcal{X}} P_{\text{can}}(\mathbf{X}|\vec{\beta}^*) \ln P_{\text{can}}(\mathbf{X}|\vec{\beta}^*). \tag{5.4}$$

This result coincides with the consequence in Shannon's information theory that the smallest space needs to store the information is determined by its uncertainty. However, the quantification of the uncertainty is affected by the appearance of heterogeneous interactions.

The Eq.5.3 shows that each unit in the sequence $\mathbf{X}$ under soft constraints is independent. Probability of each unit to gets value x_{ij} is governed by the localized parameter β^*_{ij}. Thus, the canonical ensemble is equal to the production of the marginal probability of each unit as

$$P_{\text{can}}(\mathbf{X}|\vec{\beta}^*) = \prod_{i=1}^{n}\prod_{j=1}^{m} P(x_{ji}). \tag{5.5}$$

Comparing with the microcanonical ensemble that realization all the constraints exactly, the canonical one is more like the localization of the dependencies on each unit in the information sequence.

When the two local constraints are working simultaneously, sequences are under coupled local constraints. The *Hamiltonian* is equal to $H = \sum_{i=1}^{n} \alpha^*_i c_i + \sum_{j=1^m} \beta^*_j r_j$. Then, we can get the probability distribution and the Shannon entropy of this canonical ensemble description. Details of the calculation are shown in the Appendix 5.C.

When only column-local constraints work on the signal generation or the signal generation is only constrained by soft spatial interactions among all the units in information sources, the canonical ensemble description of this signal generation still can be described by the coupled local constraints, but with the row local constraints equal to each other as $r^*_j = r^*$ and $\beta^*_j = \beta^*$. Then the *Hamiltonian* should equal to $H = \sum_{j=1}^{m}\sum_{i=1}^{n}(\beta^* + \alpha^*_i)x_{ji}$. The partition function and the Shannon entropy of this canonical ensemble can be found in the Appendix 5.B. It also can be described by the one-sided local constraints matrix in [27].

When the signal generation is only limited by the soft temporal dependence or when the units in the multivariate information source are independent, the signal generation also can be described by the coupled local constraints canonical ensembles with the unit in column local constraints equal to each other as $c^*_i = c^*$. The corresponding maximum likelihood parameter is also equal to each other as $\alpha^*_i = \alpha^*$ The *Hamiltonian* of this canonical ensemble description is $H = \sum_{j=1}^{m}\sum_{i=1}^{n}(\beta^*_j + \alpha^*)x_{ji}$. The Shannon entropy can be found in the Appendix 5.A.

5.2.2 Microcanonical ensemble description

In the microcanonical ensemble description, constraints of each state have the same value $\vec{C}^*$. The column and row local constraints are fixed exactly in each matrix.

Thus, in the signal generation, the possible state of each unit in the information source will be limited by the spatial dependencies and temporal interactions exactly. For example, the activity of neurons in the nervous system is decided by the energy that the neuron can use in signal generation. The fixed column local constraints $\vec{c}^*$ means the total energy that can be used by all neurons each time is finite. The fixed row local constraints $\vec{r}^*$ mean the total energy that can be used by each neuron in the whole process of signal generation is finite. Then the more energy cost by other units in the information source, the less energy will be left for the specific one to have different states. For each unit, the more energy cost in the past, the less will be left for the future.

The *hard* constraints of information sequences require the probability of each state with constraints $\vec{C}^*$ equal to each other as

$$P_{\text{mic}}(\mathbf{X}|\vec{C}^*) = 1/\Omega_{\vec{C}^*}, \tag{5.6}$$

where $\Omega_{\vec{C}^*} = |\mathcal{X}_{\text{mic}}|$ represents the total number of sequences with constraint $\vec{C}^*$ in the microcanonical ensemble.

According to the AEP, we can find that all sequences in the microcanonical ensemble belong to the typical set T_{mic} of it. Space needs to store the information carried by the typical set is equal to the *Boltzmann entropy* of information sequence $\mathbf{X}$ [10]

$$S_{\text{mic}} = \ln |T_{\text{mic}}| = \ln \Omega_{\text{mic}}. \tag{5.7}$$

The value of it is determined by the total number of configurations of information sequence with hard constraints $\vec{C}^*$.

The $\Omega_{\vec{C}^*}$ is hard to calculate analytical, especially when information sequences are under coupled constraints. However, according to the mechanism in [37], we can estimate the value of it by the covariance matrix of constraints in the canonical ensemble.

In the two conjugate ensembles, the total number of states in the microcanonical ensemble is equal to the number of states in the canonical ensemble with constraints equal to $\vec{C}^*$ [37]. Thus, the number of states in the microcanonical ensemble can be calculated from the canonical probability distribution with Dirac delta function (δ function) as

$$\begin{aligned} \Omega_{\vec{C}^*} &= \sum_{\mathbf{X}\in\mathcal{X}} \int_{-\vec{\pi}}^{\vec{\pi}} \frac{\mathrm{d}\vec{\psi}}{(2\pi)^K} e^{i\vec{\psi}[\vec{C}^* - \vec{C}(\mathbf{X})]} \\ &= \int_{-\vec{\pi}}^{\vec{\pi}} \frac{\mathrm{d}\vec{\psi}}{(2\pi)^K} P_{\text{can}}^{-1}(\mathbf{X}^*|\vec{\theta}^* + i\vec{\psi}). \end{aligned} \tag{5.8}$$

The integral is difficult to calculate when it is under coupled constraints [37]. But it is still possible to use saddle point technology to approach the value of $\Omega_{\vec{C}^*}$ as

$$\Omega_{\vec{C}^*} = \frac{e^{S_{\text{can}}}}{\sqrt{\det(2\pi\boldsymbol{\Sigma}^*)}} \prod_{k=1}^{K} [1 + O(1/\lambda_k^*)], \tag{5.9}$$

where $\mathbf{\Sigma}^*$ is the covariance matrix of constraints in the canonical ensemble whose entries are defined as

$$\begin{aligned}\Sigma^*_{ij} &\equiv \frac{\partial^2 \ln Z(\vec{\theta})}{\partial\theta_i \partial\theta_j}\, |_{\vec{\theta}=\vec{\theta}^*} \\ &= \mathrm{Cov}[C_i, C_j]_{\vec{\theta}^*} \\ &= \langle C_i C_j \rangle_{\vec{\theta}^*} - \langle C_i \rangle_{\vec{\theta}^*} \langle C_j \rangle_{\vec{\theta}^*},\end{aligned} \tag{5.10}$$

The $\{\lambda^*_k\}$ is the eigenvalue of covariance matrix $\mathbf{\Sigma}^*$ [37]. K is the number of constraints in the matrix $\mathbf{X}$

Then we can have the *Boltzmann* entropy of the microcanonical ensemble $S_{\mathrm{mic}} = \ln \Omega_{\vec{C}^*}$ is equal to the Shannon entropy of the conjugate canonical ensemble minus the correction part based on the covariance matrix of constraints in the canonical ensemble as

$$S_{\mathrm{mic}} = S_{\mathrm{can}} - \ln \sqrt{\det(2\pi \mathbf{\Sigma}^*)} + \sum_{k=1}^{K} \ln[1 + O(1/\lambda^*_k)]. \tag{5.11}$$

The correction part $\sum_{k=1}^{K} \ln[1 + O(1/\lambda^*_k)$ is negligible when the eigenvalue value of the covariance matrix λ^*_k is big enough. Thus, the space to store the information carried by the microcanonical ensemble sequences is smaller than the canonical one. Because the hard constraint in the microcanonical ensemble strictly modelled the influence of heterogeneous dependencies in information sequences. Therefore, the microcanonical ensemble description is closer to the natural process of signal generation under heterogeneous dependence than the canonical ensemble one, which is the maximum entropy approximation.

5.3 Total correlations of information sequences

The nonnegligible difference between the microcanonical and canonical ensemble description of the information sequences shows that the two different descriptions of the heterogeneous dependencies in the information sequences contains different information about the signal generation with temporal and spatial dependencies. However, we do not know, if this difference is related with the correaltions among units in the information sequences. Therefore, it is important to check if the nonnegligible ensemble difference is a manifestation of the correaltions of the units in the information sequences.

The matrix ensemble $\mathbf{X}$ gives a possible model for us to quantify the dependence among all units in information sequences as the total correlations $\mathbb{C}$ [72]

$$\mathbb{C} = \sum_{\mathbf{X} \in \mathcal{X}} P(\mathbf{X}) \ln \frac{P(\mathbf{X})}{\prod_{i=1}^{n} \prod_{j=1}^{m} P(x_{ji})}. \tag{5.12}$$

The symbol $\mathcal{X}$ represents the collection of all possible configurations of sequences $\mathbf{X}$ under heterogeneous dependencies.

However, with heterogeneous dependencies, both the actual probability $P(\mathbf{X})$ of the sequence $\mathbf{X}$ and the production of the marginal probability of each unit $\prod_{i=1}^{n}\prod_{j=1}^{m}P(x_{ij})$ are difficult to calculate. Therefore, the two different ensemble descriptions show above that are based on the maximum entropy principle proposed by Jaynes [15] give a way to approach the probability $P(\mathbf{X})$ and the production of marginal probability $\prod_{i=1}^{n}\prod_{j=1}^{m}P(x_{ij})$ from the biased information we know.

The probability of the matrix $\mathbf{X}$ with constraints $\vec{C}(\mathbf{X})$ to appear in the signal generation is decided by the number of configurations $\Omega_{\vec{C}(\mathbf{X})}$ in it. The microcanonical ensemble description strictly satisfies the requirement of signal generation. Thus, when the constraints are fixed, the probability $P(\mathbf{X})$ in the definition of total correlations can be replaced by the probability of states in the microcanonical ensemble description as

$$P(\mathbf{X}) = P_{\text{mic}}(\mathbf{X}|\vec{C}^{*}). \tag{5.13}$$

The production of each unit's marginal probability in Eq.5.12 is based on the assumption that the probability of each unit $P(x_{ji})$ can be calculated independently. However, when there is heterogeneous dependence, we can only use the canonical ensemble to approach the production of marginal probabilities, as the canonical ensemble has localized the dependencies of all the units in the sequence by the parameter β_{ji}. Thus, we can have the production of the marginal probability of the matrix $\mathbf{X}$ as

$$\prod_{i=1}^{n}\prod_{j=1}^{m}P(x_{ji}) = P_{\text{can}}(\mathbf{X}|\vec{\beta}^{*}). \tag{5.14}$$

It is determined by the parameter $\vec{\beta}^{*}$ and the corresponding constraints $\vec{C}(\mathbf{X})$ simultaneously.

Therefore, the total correlation in sequences with constraints $\vec{C}^{*}$ (the hard constraints in microcanonical ensemble, and the average value of constraints in canonical ensemble) can be approached by the relative entropy $S(P_{\text{mic}}||P_{\text{can}})$ between the two ensembles as

$$\mathbb{C} = S(P_{\text{mic}}||P_{\text{can}}) = \sum_{\mathbf{X}\in\mathcal{X}} P_{\text{mic}}(\mathbf{X}|\vec{C}^{*}) \ln \frac{P_{\text{mic}}(\mathbf{X}|\vec{C}^{*})}{P_{\text{can}}(\mathbf{X}|\vec{\beta}^{*})} \tag{5.15}$$

The probability of states in microcanonical ensemble with constraints not equal to $\vec{C}^{*}$ is 0, so the total correlation also can be calculated as

$$\mathbb{C} = \ln P_{\text{mic}}(\mathbf{X}^{*}) - \ln P_{\text{can}}(\mathbf{X}^{*}), \tag{5.16}$$

where $\mathbf{X}^{*}$ represents sequence with constraints $\vec{C}^{*}$ [27].

The Shanon entopy of the canonical ensemble is equal to $S_{\text{can}} = \ln P_{\text{can}}(\mathbf{X}^{*})$. The Boltzmann entropy of the microcanonical ensembl equal to $S_{\text{mic}} = \ln P_{\text{mic}}(\mathbf{X}^{*})$.Thus, the total correaltions $\mathbb{C}$ is the difference between the shannon entropy and Boltzmann entropy of the sequences with heterogeneous dependencies.

As the relative entropy density is the indicator of the measure ensemble nonequivalence [8], the total correlation $\mathbb{C}$ of the matrices $\mathbf{X}$ also has a close relationship with the EN in it.

Because the microcanonical entropy can be obtained by the covariance matrix of constraints in the canonical ensemble, we can find that the total correlation also can be calcualted by the covariance matrix of constraints in the canonical ensemble as

$$\mathbb{C} = \ln \sqrt{\det(2\pi\mathbf{\Sigma}^*)} - \ln \prod_{k=1}^{K} [1 + O(1/\lambda_k^*)]. \tag{5.17}$$

This result shows that the extra information needs to describe states in the canonical ensemble is determined by the relative fluctuation between the constraints in the canonical ensemble and the conjugate microcanonical ensemble. The bigger the fluctuation, the more difference between the two ensembles. The more information we need to store under the canonical ensemble description.

5.4 Classical information theory with ensemble description

As we already mentioned before, the ensemble description is an extension of the classical information theory with heterogeneous dependencies. Thus, in this part, we will show that the classical information theory is a particular case of the ensemble description of information sequences.

First, we will introduce the typical description of signal generating in classical information theory.

In classical information theory, there is a basic assumption that each unit in the sequence is independent. Moreover, the probability distribution of each unit to get different values is the same (independent-identical-distribution). For example, in the binary information source x, the probability of the sequence $A = [a_1, a_2, \cdots, a_i, \cdots, a_n]$ with t units equal to 1 is

$$P(A|t) = p^t(1-p)^{n-t}, \tag{5.18}$$

where p is the probability of the information source x to have the value of 1, and t is the total number of units with value 1 in the sequence.

When the length n of the sequence A goes to infinite, and the probability p of each unit to get value 1 fixed, we can find the average number of units in the sequence A to has value 1 is equal to $\langle t \rangle = t^* = n \times p$. It is a manifestation of the large number law.

The information that is generated by the information source x is carried by information sequences A in the typical set T_ϵ. This typical set can be detected by the AEP as

$$\frac{1}{n} \ln P(A|t) = \frac{t}{n} \ln p + \frac{n-t}{n} \ln(1-p). \tag{5.19}$$

When the length $n \to \infty$, the value of $\frac{t}{n}$ in the rescaled proability $\frac{1}{n} \ln P(A|t)$ will equal to $\frac{t^*}{n} = p$. Thus, the limit of the rescaled probability is equal to

$$\lim_{n\to\infty} \frac{1}{n} \ln P(A|p) = p \ln p + (1-p)\ln(1-p), \tag{5.20}$$

which is the minus of the Shannon entropy of the information source x. The Shannon entropy $s(x)$ of the information source is equal to

$$s(x) = -\frac{t^*}{n} \ln \frac{t^*}{n} - \frac{n-t^*}{n} \ln \frac{n-t^*}{n}. \tag{5.21}$$

Therefore, the typical set T_ϵ is the collection of all the sequences that satisfy the following condition

$$T_\epsilon = \{A | e^{-n(s(x)-\epsilon)} \leq P(A) \leq e^{-n(s(x)-\epsilon)}\}. \tag{5.22}$$

The space to store sequences in the typical set is equal to $\ln |T_{\epsilon=0}| = n \times s(x)$. The result shows one of the main results in Shannon's information theory that the space needs to store the information generated by the information source is decided by the uncertainty of the information source, which is the Shannon entropy of the information source x [33].

Next we will prove that the classical information theory is a particular case of the canonical ensemble described information sequences with coupled constraints when the parameter β^*_{ij} is equal to each other as β^*, and $m = 1$. It is also can be described by the canonical ensemble under global constraints $t = \sum_{i=1}^n a_i$.

When use the canonical ensemble undee global constraints to describe the information sequence in the classical information theory, the *Hamiltonian* of classical case is equal to $H(A) = t \cdot \beta^*$. The partition function is equal to $Z(\beta^*) = (1 + e^{-\beta^*})^n$. Probability of the sequence with global constraint equal to t is

$$P_{\text{can}}(A|(\beta^*, t)) = \frac{e^{-t\beta^*}}{(1 + e^{-\beta^*})^n}. \tag{5.23}$$

The soft constraints require the average value of global constraints in the canonical ensemble equals to t^* as

$$\langle t \rangle_{\text{can}} = \sum_{A \in \mathcal{X}_{\text{can}}} t(A) P_{\text{can}}(A|\beta^*, t) = t^*. \tag{5.24}$$

Thus, the probability of unit a_i to have the value of 1 in the canonical ensemble is equal to

$$p = \frac{e^{-\beta^*}}{1 + e^{-\beta^*}} = \frac{t^*}{n}. \tag{5.25}$$

The Shannon entropy of the information source under canonical ensemble description is equal to

$$s_{\text{can}}(x) = -[\frac{t^*}{n} \ln \frac{t^*}{n} + \frac{n-t^*}{n} \ln \frac{n-t^*}{n}]. \tag{5.26}$$

Hence, the typical set of sequences in the classical information theory can be obtained from the AEP as

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\ln P_{\text{can}}(A|\beta^*,t) &= \frac{1}{n}\sum_{i=1}^{n}\ln p(a_i)\\ &\to E\ln p(x)\\ &= -s_{\text{can}}(x),\end{aligned} \tag{5.27}$$

The sequences in the typical set should satisfy the following condition

$$T_{\text{can}} = \{A|e^{-ns_{\text{can}}(x)-\epsilon} \leq P(A) \leq e^{-ns_{\text{can}}(x)+\epsilon}\}. \tag{5.28}$$

According to the relationship show in Eq.5.23, we find sequences that belong to the typical set also can be represented by the Shannon entropy of the canonical ensemble. As the sum of n rescaled entropy $s_{\text{can}}(x)$ is equal to

$$\begin{aligned}n\times s_{\text{can}}(x) &= \frac{t^{*t^*}(n-t^*)^{n-t^*}}{n^n}\\ &= \ln P_{\text{can}}(A|(\beta^*,t^*))\\ &= S_{\text{can}}\end{aligned} \tag{5.29}$$

Thus, the space to store the information carried by those sequences is equal to

$$\ln|T_\epsilon^{\text{can}}| = S_{\text{can}} = n\times s(x). \tag{5.30}$$

The result of the canonical ensemble description is equivalent to the classical description of the information sources. It proves that when the interactions in the information sequences are homogeneous, the classical description in information theory is a special case of the canonical ensemble description.

Then we need to check what happens when we use the microcanonical ensemble to describe the classical signal generating of the binary information source x. The microcanonical ensemble description is different from the canonical one. The probability of these sequences with t^* units have a value 1 can be obtained by the microcanonical ensemble description as

$$P_{\text{mic}}(A|t^*) = 1/\binom{n}{t^*}. \tag{5.31}$$

All the sequences in the microcanonical ensemble belong to the typical set T_{mic} of it, so the smallest space needs to store the information carried by those sequences $\ln|T_{\text{mic}}|$ is equal to the Boltzmann entropy of the microcanonical ensemble

$$\ln|T_{\text{mic}}| = \ln\binom{n}{t^*} = \ln\Omega_{\text{mic}} = S_{\text{mic}}, \tag{5.32}$$

where Ω_{mic} is the total number of sequences under this hard constraints t^*. We can find that the canonical ensemble description showed above is equivalent to the classical information theory when the i.i.d. assumption hold.

The difference between the space to store the information carried by the sequences in the two ensembles is equal to

$$\ln|T_{\mathrm{can}}| - \ln|T_{\mathrm{mic}}| = S_{\mathrm{can}} - S_{\mathrm{mic}}. \tag{5.33}$$

It is the total correlations $\mathbb{C}$ in the sequence A

$$\mathbb{C} = \sum_{A\in\mathcal{A}} P_{\mathrm{mic}}(A|t^*) \ln \frac{P_{\mathrm{mic}}(A|t^*)}{P_{\mathrm{can}}(A|\theta^*)}. \tag{5.34}$$

When it is rescaled n,

$$\frac{1}{n}\mathbb{C} = \frac{1}{n}[\ln \frac{n^n}{t^{*t^*}(n-t^*)^{n-t^*}} - \ln \binom{n}{t^*}], \tag{5.35}$$

It is equal to the relative entropy density between the two ensembles.

According to the Stirling approximation, the limit of the rescaled difference is equal to 0

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{C} = \lim_{n\to\infty} \frac{1}{n}[\frac{1}{2}\ln 2\pi t^*(1-\frac{t^*}{n})] = 0. \tag{5.36}$$

It means the canonical ensemble will converge to the microcanonical one in the thermodynamic limit. It also shows that the limit of the information storage is the same in the two ensemble descriptions.

In statistical physics, this is the measure-level ensemble equivlaence [8]. The logarithm difference $\ln|T_{\mathrm{can}}| - \ln|T_{\mathrm{mic}}|$ is the relative entropy

$$\ln|T_{\mathrm{can}}| - \ln|T_{\mathrm{mic}}| = S_{\mathrm{can}} - S_{\mathrm{mic}} = S(P_{\mathrm{mic}}||P_{\mathrm{can}}), \tag{5.37}$$

which grows like $o(n)$.

The microcanonical ensemble description has realized the constraints in the signal generation exactly. The classical information theory under this case is a particular example of the canonical ensemble. It means the effectiveness of the classical information theory is based on the EE between the microcanonical ensemble and the canonical ensemble with the global constraints t^*.

5.5 Extensive number of constraints and ensemble nonequivalence

As already strictly prov proved in Chapter 3, that the ensemble nonequivalence is generally exist in the systems with extensive local constraints. When we use K to represents the numbers of constraints in a systems. In the matrix under global constraints, there is only one constraint, $K = 1$. When the $m \times n$ matrix is under row local constraits, there is m constraints in it, $K = m$. If the matrix is under column local constraints, then there is $K = n$ constraints in it. The matrix under coupled local constraints has $K = m + n$ constraints. The extension of constraints in the matrix $\mathbf{X}$ is a localization of dependencies in information sequences. According to the canonical ensemble description, the units in the matrix under global constraints have a homogeneous interactions. When the matrix are under row local constraints, the units can be divided into m parts and each part have the same interactions. Obviously, when the $m \times n$ matrix under coupled constraints, the interactions have be localized to each unit.

The classical information theory is a special case of the global constrained canonical ensemble description, when the matrix $\mathbf{X}$ only has one row $m = 1$ and n columns, $t = \sum_{i=1}^{n} x_{1i}$ [27]. It is under EE, and the limit of information storage in the two ensemble descriptions is equivalent.

The single generation by the m independent variables with different probability distribution in the network information theory is an m extension of the global constrained classical information theory. We can use the matrix with the finite row local constraints $\vec{r}$ to describe the sequences generated by the i.i.d multivariate information source in the network information theory [27, 30]. There are $K = m$ constraints in it. It is also under EE. We can find the description of it in Appendix 5.A.

If we only focus on the spatial dependences, there will be n constraints in the sequence, $K = n$. The sequences generated by this non-stationary process can be represented by the states of the matrices with local column constraint $\vec{c}$. According to Stirling's approximation, the limit of rescaled total correlations is bigger than 0 as

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} [\frac{1}{2} \ln 2\pi c_i^* (1 - \frac{c_i^*}{m})] > 0. \tag{5.38}$$

Sequences with this spatial dependence are under EN. Details of proof are in Appendix 5.B.

The coupled local constraints are the two kinds of dependences work simultaneously, $\vec{C}(\mathbf{X}) = [\vec{c}, \vec{r}]$. There are $K = m + n$ constraints in it. The probability of the microcanonical ensemble is difficult to calculate, but we can still get the conjugate canonical probability as

$$P_{\text{can}}(\mathbf{X}|\vec{\theta}^*) = \prod_{i=1}^{n} \prod_{j=1}^{m} \frac{e^{-(\alpha_i^* + \beta_j^*) x_{ji}}}{e^{-(\alpha_i^* + \beta_j^*)} + 1}. \tag{5.39}$$

According to the results in [27], those sequences are also under EN.

The constraints' extension has two paths: the first one is from global to row-local constraints, then coupled local constraints, the second one is from global to column-local constraints, then to the coupled local constraints. As we already know, both of the two paths will break the EE, but the difference in the two paths will show which kind of dependence subleading the breaking of EE.

In order to check how is the total correlation will change when the constraint is extended in the system, we set a series of models with homogeneous dependencies. The homogeneous spatial interaction means the c_i in the column local constraint is equal to each other as c^*. The homogeneous temporal dependencies need each element in the row local constraints equal to each other as $r_j = r^*$. And the global constraint is equal to $t^* = n \times c^* = m \times r^*$. All the three special cases can be described by coupled constraints matrix ensemble with different definitions of *Hamilotonian*. Thus, we have a series of models where the spatial interactions and temporal dependencies are homogeneous but still under extensive local constraints.

The signal generation by multivariate information source under the same temporal dependencies should have the same value with row local constraint $r_j^* = r^*$. The canonical entropy of this matrix ensemble is equal to

$$S_{\text{can}}^{(K=m)} = m \times \ln \frac{n^n}{r^{*r^*}(n-r^*)^{n-r^*}}. \tag{5.40}$$

As we already mentioned before, it is a m times linear extension of the single independent identical signal generating in the classical information theory, and it is under EE.

In information sequences with homogeneous spatial interactions in the information sources, the canonical entropy should equal to

$$S_{\text{can}}^{K=n} = n \times \ln \frac{n^n}{r^{*r^*}(n-r^*)^{n-r^*}}. \tag{5.41}$$

It is equivalent to the signal generated by the multivariate information source with an identical probability distribution, but variables in the information source are not independent. It has nonneglected dependencies among them.

If the homogeneous spatial interactions and temporal dependencies work simultaneously, information sequences with these coupled constraints have the same column and row local constraint as $c_i^* = c^*$, $r_j^* = r^*$. The canonical entropy of this matrix ensemble should equal to

$$S_{\text{can}}^{(K=m+n)} = n \times \ln \frac{m^m}{c^{*c^*}(m-c^*)^{m-c^*}} \tag{5.42}$$

when we focus on the spatial interactions.

The canonical entropy is equal to

$$S_{\text{can}}^{(K=m+n)} = m \times \ln \frac{n^n}{r^{*r^*}(n-r^*)^{n-r^*}} \tag{5.43}$$

when the calculation is focused on the temporal dependencies. It is also equivalent to the matrix ensemble with one side column local constraints,

Canonical entropies show above are all equivalent to the one with global constraints t^*, when the r^* is replaced by $r^* = t^*/m$, or the c^* is replaced by $c^* = t^*/n$. It shows that the canonical ensemble description of information Sequences under four different local constraints that are implied by the homogeneous dependencies are equivalent to each other,

$$S_{\text{can}}^{(K=1)} = S_{\text{can}}^{(K=m)} = S_{\text{can}}^{(K=n)} = S_{\text{can}}^{(K=m+n)}. \tag{5.44}$$

Thus, under soft constraints, the four homogeneous signal generations have the same information-theoretical bounds. The space to store the information generated by it is the same.

However, the Boltzmann entropy of information sequences with the four different dependences is different. Under the global constraint, it is equal to

$$S_{\text{mic}}^{(K=1)} = \ln \binom{mn}{t^*}. \tag{5.45}$$

Under row-local constraints, the Boltzmann entropy is equal to

$$S_{\text{mic}}^{(K=m)} = m \times \ln \binom{n}{r^*}. \tag{5.46}$$

When it is under column-local constraints, the Boltzmann entropy is equal to

$$S_{\text{mic}}^{(K=n)} = n \times \ln \binom{m}{c^*}. \tag{5.47}$$

When the two dependence is working Simultaneously, the Boltzmann entropy of the hard constrained sequences is difficult to calculate, but we can still get the approximation by the Eq (5.11), and the value of it is smaller than $S_{\text{mic}}^{(K=n)}$. Thus, the relationship between the four Boltzmann entropy is

$$S_{\text{mic}}^{(K=1)} > S_{\text{mic}}^{(K=m)} > S_{\text{mic}}^{(K=n)} > S_{\text{mic}}^{(K=m+n)}. \tag{5.48}$$

The relationship between the breaking of ensemble equivalence and the extension of constraints is shown in Fig.5.2.

These results show above proved that the temporal dependence of each independent variable in information sources is not enough to break the EE between the hard and soft constraint's description of information sequences. This EE allows classical information theory can be applied in the independent multivariate information sources system to estimate the limit of information storage [30].

5.6 Information transmission with coupled sources

In classical information theory, the maximum speed of reliable information transmission through a channel is the channel capacity. It is the other vital bounds of the information theory, and it is equal to the mutual information between the information source $\mathbf{x}$ and the received signal $\mathbf{y}$ [33].

To transport the information carried by matrix $\mathbf{X}$ with local constraints, we still need to code it by codes G with length L. In the receiver, we will receive a matrix $\mathbf{Y}$. The information we will transport through this channel is equal to $L \times R$, and it is decided by the uncertainty from the channel $\mathbf{H}(\mathbf{X}|\mathbf{Y})$ and the information carried

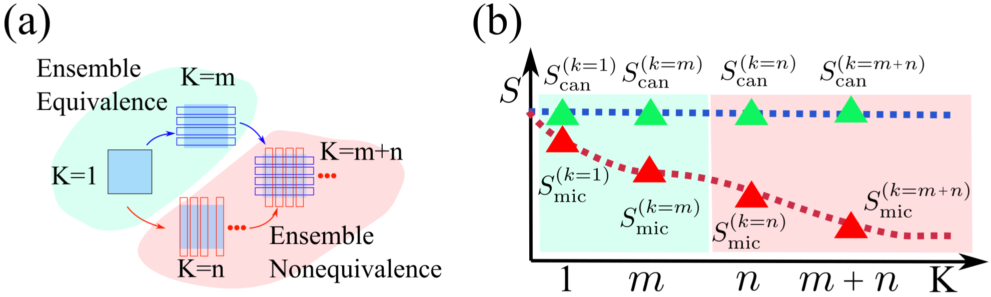

Figure 5.2. The relationship between the constraints' extension and the breaking of EE is shown in (a). The extension of constraints in matrix described systems have two paths. First, the constraints is extend from global constraint to row local constraints then to coupled constraints. The other path is the extendtion from global constraints to column local constraints then coupled constraints. Systems are still under EE when the constraints extend from global to row local constraints ($K = 1 \to K = m$), but when the constraints extend to the coupled local constraints from row local constraints $K = m \to K = m + n$, the EE is broken. On the other path, the extension of global constraint to column local constraints already breaks the EE. Obviously, the extension from column local constraints to coupled local constraints is already under EN. This result shows that the extension of global constraints to finite numbers local constraints is not enough to break the EE. Only when the extension of local constraints has the same order as the increasing of system's size, or even faster, EE will break by the extension of constraints. On the other hand, the relationship between the canonical entropy and microcanonical entropy of the systems with extensive constraints under homogeneous dependencies in (b) shows that the non-vanished fluctuation of the constraints in the canonical ensemble does not lead to the breaking of EE. The EN is formed by the reduction of possible configurations of sequences in the microcanonical ensemble. Because the canonical ensemble with soft constraints has the same Shannon entropy when it is under homogeneous local and global constraints. But the Boltzmann entropy of the microcanonical ensemble with hard constraints is declined.

by the sequences $\mathbf{X}$ sent to the receiver $\mathbf{I}(\mathbf{X};\mathbf{Y})$. Thus, we will have the relationship as

$$L \times R = \mathbf{H}(\mathbf{X}|\mathbf{Y}) + \mathbf{I}(\mathbf{X};\mathbf{Y}) \geq S(\mathbf{X}). \tag{5.49}$$

In the zero-error channel, the matrix we received $\mathbf{Y}$ is the matrix we sent $\mathbf{X}$, the value of $\mathbf{H}(\mathbf{X}|\mathbf{Y}) = 0$, the mutual information $\mathbf{I}(\mathbf{X};\mathbf{Y}) = S(\mathbf{X})$, then the effective information we need transport is equal to the Shannon entropy of the sequence $S(\mathbf{X})$. When the message we received is a random matrix, the mutual information $\mathbf{I}(\mathbf{X};\mathbf{Y}) = 0$, but the channel uncertainty is equal to $\mathbf{H}(\mathbf{X}|\mathbf{Y}) = S(\mathbf{X})$.

Both in the two cases, the smallest information needs to transport are all decided by the entropy of the sequences $S(\mathbf{X})$. Thus, the symbol rate $R = S(\mathbf{G})/L$ or the channel capacity is also affected by the hard or soft constraints in this signal generation process. When using the microcanonical to describe the sequences, the Boltzmann entropy S_{mic} is smaller than the Shannon entropy in the conjugate canonical ensemble S_{can}. Thus, the symbol rate R_{mic} needs for the microcanonical ensemble is smaller than the symbol rate R_{can},

$$R_{\text{mic}} \leq R_{\text{can}}. \tag{5.50}$$

It means the microcanonical ensemble with local constraints is reliable than the canonical ensemble one when we use the same channel to transport the information carried by the sequences ensemble. Because it needs less information carried for each symbol in the code, the system will have more redundancy.

5.7 Conclusions

The information is not only the bits that flow in the electronic communication systems. It also generally exists in the natural systems, e.g., the activity of neurons in the nervous system and the fluctuation of the financial system. The birth of classical information theory ignores the interactions and dependencies among the information sources, as it is limited by the initial structure of artificial communication systems. However, this imperfection has been amplified when the classical information theory is attempted to describe the signal generation in natural systems with heterogeneous dependencies and interactions.

The existence of heterogeneous dependencies in the signal generation has broken the i.i.d assumption in the classical information theory, which is the cornerstone of the finding that the uncertainty of information sources decides the limit of the information storage. Thus, finding the information-theoretical bounds (i.e., the limit of information storage and channel capacity) of the natural signal generation should focus on the information sequences, not the interacting and temporal dependent information source.

The information sequence with heterogeneous interactions has an extensive length. It is analogous to the state in the thermodynamic system. That is why statistical ensembles can be used to describe the information sequences in this work.

In ensemble descriptions of the information sequence, the heterogeneous dependencies imply local constraints in the statistical ensemble. The microcanonical ensemble description requires all the local constraints fixed in the information sequences generated by the heterogeneous interacting information sources. It is closer to the real process of signal generation in natural systems. On the contrary, the canonical ensemble description is a maximum entropy approximation of the signal generation. The total correlation is equal to the relative entropy between the microcanonical and canonical ensemble, which is the indicator of the measure level EN. It means the EN appearance in the heterogeneous interacted signal generation also has a connection with the degree of dependences among the units in it.

The sequences described by the microcanonical ensemble with hard local constraints need less space to store the information carried by them compared with the canonical one. In the information transmission, the microcanonical ensemble also needs a lower symbol rate to transport the information than the canonical ensemble. Both results show that the microcanonical ensemble description is better than the canonical one from the limit of information storage. However, the probability distribution of the microcanonical ensemble is mathematically difficult to calculate than the canonical ensemble. Thus, when we want to build a communication system with a multi-coupled source, we still need to consider the trade-off between the cost of calculation to maintain the hard constraints in the microcanonical ensemble and the waste of space and channel capacity to use the canonical ensemble with soft constraints.

We also find that the classical information theory under the i.i.d. assumption is a special case of the canonical ensemble description when the column-local constraints have the same value or when the sequences have homogeneous spatial dependence. The effectiveness of the classical information theory is based on the EE between the canonical ensemble and the natural signal generation.

This non-stationary process also gives a chance to learn how is the extension of constraints causes EN in a system. From the canonical ensemble description, the global constraints and two different local constraints are special cases of the coupled constraints matrix ensemble. The extension of constraints in it has two paths. The first one is from global constraints to row local constraints then coupled constraints. The second one is from global constraints to column local constraints, then coupled constraints. The matrix under global constraints and finite row local constraints is in the EE. However, this equivalence will break when there are column-local constraints and coupled constraints. The column-local constraints are used to describe the spatial interaction among the units in the information source. Thus, the EN in sequences is caused by the spatial interactions among the units in the signal generation, not the temporal dependence of finite independent variables in the information source. Even the row local constraints also have influences on all the units in sequences.

The same Shannon entropy and the decreased Boltzmann entropy in the process of constraints' extention illustrate that the breaking of EE is caused by the reduction of the possible configurations in the microcanonical ensemble when there are homogeneous dependencies. This mechanism is different from the traditional one, which shows that the EN is caused by the non-vanished fluctuation of constraints in the canonical ensemble. This finding extended our understanding of the EN in statistical physics.

Appendix 5.A Row local constraints and multivariate independence source

There is m rows in the matrices $\mathbf{X}$, which means there are m units in the information source. The signal generation of the multivariate independent information source is under ensemble equivalence, as there is a finite number of local constraints in it, and there is no phase transition in the information sequence [27]. This signal generation still can be described by the classical information theory. Then according to the AEP, we can find the limit of information storage.

The sequence generated by the information source $\vec{x} = [b_1, b_2, \cdots, b_j, \cdots, b_m]$ with m independent variables is an $m \times n$ matrix B. As the m units in the information source are independent with each other, the sequence B can be divided into m row vectors $\mathbf{B} = \{\vec{R}_1; \vec{R}_2; \cdots; \vec{R}_m\}$. Each row vectors $\vec{R}_j$ of the matrix B has n elements in it. According to the classical information theory, the m i.i.d. random variables may have different probability to get different value, so the probability of sequence $\mathbf{B}$ to appear in the signal generation is equal to

$$P(\mathbf{B}) = \prod_{j=1}^{m} P(\vec{R}_j) = \prod_{j=1}^{m} p_j^{r_j}(1-p_j)^{n-r_j}. \tag{5.51}$$

Here we still focus on the binary information sequence. Thus, p_j is the probability of each unit in row j to have value 1. The r_j is the number of units in row j to have a value of 1, and it will affect the process of signal generating when different constraints model it.

The value of p_j can be obtained by the average value of total units with value 1 in each row as $p_j = r_j^*/n$. When the jth variable is represented by b_j, then the AEP will be generalized as

$$\begin{aligned} \frac{1}{n}\ln P(\mathbf{B}) &= \sum_{j=1}^{m}[\frac{r_j^*}{n}\ln\frac{r_i^*}{n} + \frac{n-r_j^*}{n}\ln\frac{n-r_j^*}{n}] \\ &\to -\sum_{j=1}^{m} s(b_j) \end{aligned} \tag{5.52}$$

Sequences belonging to the typical set of this system with multivariate independence information source should satisfy the following condition

$$T_\epsilon = \{\mathbf{B}|e^{-n\sum_{j=1}^{m} s(b_j)-\epsilon} \leq P(\mathbf{B}) \leq e^{-n\sum_{j=1}^{m} s(b_j)+\epsilon}\}. \tag{5.53}$$

The space to store the information carried by it is equal to

$$\ln|T_\epsilon| = n \times \sum_{j=1}^{m} s(b_j). \tag{5.54}$$

This result shows that the uncertainty of the information source still decides the limit of information storage. Even there are m independent variables in it.

Next, we will introduce the ensemble description of the information sequence generated by the multivariate independent information sources.

Canonical ensemble description- The information sequence $\mathbf{B}$ generated by the independent variables with different probability distribution can be modeled by the matrix $\mathbf{X}$ with the row local constraints $\vec{r}(\mathbf{X}) = [r_1^*, r_2^*, \cdots, r_m^*]$, where $r_j^* = \sum_{i=1}^n x_{ji}$. The maximum likelihood parameter $\vec{\beta} = [\beta_1^*, \beta_2^*, \cdots, \beta_m^*]$ has m elements. The *Hamiltonian* is still the linear combination of the constraints and parameters $H(\mathbf{X}) = \sum_{j=1}^m \beta_j^* r_j^*$. In the binary case, the partition function of this matrix ensemble is

$$Z(\vec{\beta}) = \prod_{j=1}^{m} (1 + e^{-\beta_j^*})^n. \tag{5.55}$$

Then we can have the canonical probability of the state $\mathbf{X}$ with row local constraints $\vec{r}(\mathbf{X})$ as

$$P_{\text{can}}(\mathbf{X}) = \prod_{j=1}^{m} \frac{e^{-\beta_j^* r_j^*}}{(1 + e^{-\beta_j^*})^n}. \tag{5.56}$$

The parameter β_j^* is decided by the corresponding average value of row local constraints $\langle r_j \rangle$.

Space to store the information carried by the information sequences $\mathbf{B}$ still can be quantified by the AEP. The rescaled logarithm of the probability is equal to

$$\frac{1}{n} \ln P_{\text{can}}(\mathbf{X}) = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n} \ln \frac{e^{-\beta_j^* x_{ji}}}{1 + e^{-\beta_j^*}}. \tag{5.57}$$

When the length n goes to infinite, the probability of the units in jth row to have value 1 is equal to average value of x_{ji} as

$$\langle x_{ji} \rangle = \frac{e^{-\beta_j^*}}{1 + e^{-\beta_j^*}} = \frac{r_j^*}{n}. \tag{5.58}$$

The sum of the logarithms of the probability for the $m \times n$ units will equal the sum of the m variables Shannon entropy

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \ln P_{\text{can}}(\mathbf{X}) &= \lim_{n\to\infty} \sum_{j=1}^{m} [\frac{r_j^*}{n} \ln \frac{r_j^*}{n} + \frac{n - r_j^*}{n} \ln \frac{n - r_j^*}{n}] \\ &\to \sum_{j=1}^{m} s(b_j). \end{aligned} \tag{5.59}$$

Then space to store the information carried by the sequences in the typical set $T_{\text{can}}^{\epsilon=0}$, which satisfy the following condition

$$T_{\text{can}}^{\epsilon=0} = \{\mathbf{X} | P_{\text{can}}(\mathbf{X}) = e^{n \sum_{j=1}^{m} s(b_j)}\} \tag{5.60}$$

The logarithm of the number of sequence in the typical set also equal to $\ln|T_{\text{can}}| = n\sum_{j=1}^{m} s(b_j)$, which is the Shannon entropy of the sequences $S_{\text{can}}(\mathbf{X})$

$$\ln|T_{\text{can}}^{\epsilon=0}| = S_{\text{can}}(\mathbf{X}) = \sum_{j=1}^{m}\ln[\frac{n^n}{{r_j^*}^{r_j^*}(n-r_j^*)^{n-r_j^*}}]. \tag{5.61}$$

It is equal to the $\ln|T_\epsilon| = n \times \sum_{j=1}^{m} s(b_j)$ in the classical information theory. This result shows that the method in the classical information theory is a particular case of the canonical ensemble description.

Microcanonical ensemble description- The microcanonical ensemble under the hard constraints needs the total number of units with the value 1 in each row of all the sequences $\mathbf{X}$ are fixed the same as $\vec{r^*} = [r_1^*, r_2^*, \cdots, r_j^*, \cdots, r_m^*]$. The probability of each sequence decided by the total number of configurations of this local row constrained sequences with constraints $\vec{r^*}$ as

$$P_{\text{mic}}(\mathbf{X}) = 1/\prod_{j=1}^{m}\binom{n}{r_j^*}. \tag{5.62}$$

All the sequences in the microcanonical ensemble belong to the typical set of it, the space to store the information carried by it is equal to the *Boltzmann entropy* of the sequences as

$$\ln|T_{\text{mic}}^{(\vec{r^*})}| = \sum_{j=1}^{m}\ln\binom{n}{r_j^*} = \ln\Omega_{\vec{r^*}}. \tag{5.63}$$

The rescaled difference between the logarithm of the two ensemble's probability is close to the rescaled total correaltions $\frac{1}{n}\mathbb{C}$ as

$$\frac{1}{n}\mathbb{C} = \frac{1}{n}\sum_{j=1}^{m}[\ln\frac{n^n}{{r_j^*}^{r_j^*}(n-r_j^*)^{n-r_j^*}} - \ln\binom{n}{r_j^*}] \tag{5.64}$$

The asymptotic behaviour of $\frac{1}{n}\mathbb{C}$ decided by m, in this work, $m = o(n)$, thus the limit of $\frac{1}{n}\mathbb{C}$ is equal to 0 as

$$\lim_{n\to\infty}\frac{1}{n}\mathbb{C} = \lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{m}[\frac{1}{2}2\pi r_j^*(1-\frac{r_j^*}{n})] = 0. \tag{5.65}$$

The signal generalization by the classical independent multivariate information sources is under ensemble equivalence. That is why we can still use the AEP in the estimation of information-theoretical bounds.

Appendix 5.B Column local constraints and non-stationary process

In the binary matrix $\mathbf{X}$, when the constraints are the time-variation total energy can be used by all the m units in different times, the process recorded by the matrices ensemble $\mathcal{X}$ is under column local constraints $\vec{c^*}(\mathbf{X}) = [c_1^*, c_2^*, \cdots c_n^*]$. At time i, the constraint is the sum of all the units in column i of matrice $\mathbf{X}$ as $c_i^* = \sum_{j=1}^m x_{ji}$. The probability of the variable getting different values in the signal generation will change with time, so applying the classical information theory in this process is impossible, but we can still use the ensemble descriptions.

Microcanonical ensemble description- In the microcanonical ensemble description of sequences, the probability of each state is based on the total number of configurations in it as

$$P_{\text{mic}}(\mathbf{X}|\vec{c^*}) = 1/\prod_{i=1}^{n}\binom{m}{c_i^*}. \tag{5.66}$$

Obviously, all the sequences with hard constraints still belong to the typical set of it, so the space to store the information carried by it is equal to the *Boltzmann entropy* of it as

$$\ln|T_{\text{mic}}^{(\vec{c}^*)}| = -\ln P_{\text{mic}}(\mathbf{X}|\vec{c^*}) = \sum_{i=1}^{n}\ln\binom{m}{c_i^*}. \tag{5.67}$$

This is the smallest space need to store the information generated under the constraints $\vec{c}^*$.

Canonical ensemble description- We need the canonical ensemble to describe the sequences $\mathbf{X}$ with soft constraints. The correspond parameter $\vec{\alpha}^* = [\alpha_1^*, \cdots, \alpha_n^*]$ has n elements in it. The *Hamiltonian* of the binary matrix still is a linear combination of the parameter and constraints, $H = \sum_{i=1}^n c_i \alpha_i^*$. The partition function is equal to

$$Z(\vec{\alpha}^*) = \prod_{i=1}^{n}(e^{-\alpha_i^*}+1)^m, \tag{5.68}$$

and the probability of each sequence in the system is equal to

$$P_{\text{can}}(\mathbf{X}|\vec{\alpha}^*) = \prod_{i=1}^{n}\frac{e^{-\alpha_i^* c_i}}{(e^{-\alpha_i^*}+1)^m}. \tag{5.69}$$

The average value of x_{ji} equal to

$$\langle x_{ji}\rangle = \frac{e^{-\alpha_i^*}}{e^{-\alpha_i^*}+1} = \frac{c_i^*}{m}. \tag{5.70}$$

This condition imply the relationship between the parameter α_i^* and the corresponding constraints c_i^* as

$$e^{-\alpha_i^*} = \frac{c_i^*}{m-c_i^*}. \tag{5.71}$$

The space to store the information carried by it can not be approached by the AEP, but we can find the Shannon entropy of the whole sequences as

$$\begin{aligned} S_{\text{can}} &= \sum_{i=1}^{n} \alpha_i^* c_i^* + \ln Z(\vec{\alpha}^*) \\ &= \sum_{i=1}^{n} [-c_i^* \ln \frac{c_i^*}{m-c_i^*} + m \ln \frac{m}{m-c_i^*}] \\ &= \sum_{i=1}^{n} [\ln \frac{m^m}{c_i^{*c_i^*}(m-c_i^*)^{m-c_i^*}}]. \end{aligned} \tag{5.72}$$

Under this constraint, it is difficult to find the typical set of the sequences by the classical information theory, but we can still find the difference of the rescaled logarithm of the probability between microcanonical and canonical ensemble is equal to the rescaled total correlations in the sequence as $\frac{1}{n} \ln \frac{P_{\text{mic}}(D)}{P_{\text{can}}(D)} = \frac{1}{n}\mathbb{C}$, which is equal to

$$\frac{1}{n}\mathbb{C} = \frac{1}{n} \sum_{i=1}^{n} [\ln \frac{m^m}{c_i^{*c_i^*}(m-c_i^*)^{m-c_i^*}} - \ln \binom{m}{c_i^*}]. \tag{5.73}$$

As we already know the total correlations is equal to the relative entropy, thus the difference between the rescaled logarithm of the probability is equal to the relative entropy density as $\frac{1}{n}\mathbb{C} = \frac{1}{n}S(P_{\text{mic}}||P_{\text{can}})$. When the value of n goes to infinite, the limit

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{C} = O(n) > 0, \tag{5.74}$$

so this signal generation is under ensemble nonequivalence.

The total correlation between the microcanonical ensemble and canonical ensemble of the matrix $\mathbf{X}$ with column local constraints $\vec{c}^*$ grows like $O(n)$ in the thermodynamic limit. It means the information carried by the typical set of the two ensemble descriptions is different under the spatial interactions. This difference also manifested in the appearance of measure level ensemble nonequivalence in it.

Appendix 5.C Coupled local constraints and multi-coupled process

In the former two subsections, we have introduced the ensemble description of the information generating with independent spatial or temporal dependencies. They are modelled by the matrix ensemble with local column or row constraints. In this subsection, we will study what will happens when the two constraints work simultaneously $\vec{C}(\mathbf{X}) = [\vec{c}, \vec{r}]$.

Canonical ensemble description- As we know, under the coupled local constraints, the maximum likelihood parameter will be $\vec{\theta} = [\vec{\alpha}, \vec{\beta}]$. The constraint $\vec{c}$ is the local column constraints. The $\vec{r}$ is the local row constraints. The corresponding parameters $\vec{\theta}$ also comes from the maximum likelihood parameter of row and column local constraints. The *Hamiltonian* is still the linear combination of constraints and parameters as

$$H = \sum_{i=1}^{n} \sum_{j=1}^{m} (\alpha_i + \beta_j) x_{ji}. \tag{5.75}$$

We still focus on the binary matrix, so the partition function of these sequences with coupled local constraints is

$$Z(\vec{\theta}) = \prod_{i=1}^{n} \prod_{j=1}^{m} [e^{-(\alpha_i + \beta_j)} + 1]. \tag{5.76}$$

Probability of the sequence $\mathbf{X}$ to appears in the signal generating process is equal to

$$P_{\text{can}}(\mathbf{X}) = \prod_{i=1}^{n} \prod_{j=1}^{m} \frac{e^{-(\alpha_i^* + \beta_j^*) x_{ji}}}{e^{-(\alpha_i^* + \beta_j^*)} + 1}. \tag{5.77}$$

The probability of each unit in matrix $\mathbf{X}$ is decided by the value of x_{ji} and the corresponding parameter α_i and β_j. The smallest space needs to store the information carried by them is still equal to the Shannon entropy of it as

$$S_{\text{can}} = \sum_{i=1}^{n} \sum_{j=1}^{m} (\alpha_i + \beta_j) \langle x_{ji} \rangle + \ln Z(\vec{\theta}). \tag{5.78}$$

The average value of x_{ji} is difficult to get exactly, but we can still find the exact value of it under the special setting of the constraints.

Appendix 5.D Homogeneous dependencies under different constraints

As a special case of heterogeneous dependencies, homogeneous dependencies gives a chance for us to check what will happens when the canonical ensemble descriptions are equivalent under different constraints, but the microcanonical ensemble descriptions are different. Obviously, the global constraints and one-sided local constraints (both the column and row local constraints) are all the special cases of the coupled constraints under homogeneous dependencies. Thus, in this part, we will introduce the coupled constrained ensemble descriptions of the signal generation with homogeneous dependencies under different constraints.

5.D.1 Global constraint t^*

The canonical ensemble description of the information sequence under global constraints is a special case of the coupled local constraints when there is only one constraint, and the corresponding maximum likelihood parameter is equal to $\alpha^* + \beta^*$.

The *Hamiltonian* of it is equal to

$$H = \sum_{i=1}^{n}\sum_{j=1}^{m}(\alpha^* + \beta^*)x_{ji}. \tag{5.79}$$

The partition function will be

$$Z(\theta^*) = \prod_{i=1}^{n}\prod_{j=1}^{m}[e^{-(\alpha^*+\beta^*)} + 1] = [e^{-(\alpha^*+\beta^*)} + 1]^{mn}. \tag{5.80}$$

The probability of the states under this special case is equal to

$$P_{\text{can}}(\mathbf{X}) = \frac{e^{-\sum_{j=1}^{m}\sum_{i=1}^{n}(\alpha^*+\beta^*)x_{ji}}}{[e^{-(\alpha^*+\beta^*)} + 1]^{mn}}. \tag{5.81}$$

When the sum of all elements in sequence is equal to t, the probability of $\mathbf{X}$ to appears in the signal generation is equal to that in the canonical ensemble with global constraints t^* as

$$P_{\text{can}}(\mathbf{X}) = \frac{e^{-(\alpha^*+\beta^*)t}}{[e^{-(\alpha^*+\beta^*)} + 1]^{mn}}. \tag{5.82}$$

When the average value of the total number of units in the information sequence with value 1 $\langle t \rangle = t^*$ as

$$\langle t \rangle = \sum_{\mathbf{X}\in\mathcal{X}} t(\mathbf{X})P_{\text{can}}(\mathbf{X}) = t^*. \tag{5.83}$$

We can have the parameter $\alpha^* + \beta^*$ is equal to

$$e^{-(\alpha^*+\beta^*)} = \frac{t^*}{mn - t^*}. \tag{5.84}$$

Then we can find the Shannon entropy of the information sequence with the number of constraints equal to 1 is equal to $\ln P_{\text{can}}(\mathbf{X}|t^*)$ as

$$S_{\text{can}}^{(K=1)} = \ln \frac{mn^{mn}}{t^{*t^*}(mn - t^*)^{mn-t^*}}. \tag{5.85}$$

Then microcanonical ensemble description of the information sequence with global constraints t^* have Ω_{t^*} states in it. Thus, the probability of each state in it is equal to

$$P_{\text{mic}}(\mathbf{X}|t^*) = \frac{1}{\Omega_{t^*}} = 1/\binom{mn}{t^*}. \tag{5.86}$$

Therefore, the *Boltzmann* entropy of the microcanonical ensemble with global constraints is equal to

$$S_{\text{mic}}^{(K=1)} = \ln \binom{mn}{t^*} \tag{5.87}$$

According to Stirling's approximation and the results in [27], the difference between the Shannon entropy of canonical and the microcanonical ensemble is equal to

$$S_{\text{can}}^{(K=1)} - S_{\text{mic}}^{(K=1)} = \frac{1}{2}\ln\left[2\pi t^*\left(1-\frac{t^*}{\pm mn}\right)\right][1+o(1)]. \tag{5.88}$$

Thus, it is under ensemble equivalence.

5.D.2 Row local constraints

When we use $r^* = t^*/m$ as the row local constraints, there are m constraints in the information sequence. It also can be model by the matrix with coupled constraints, when the elements in the column local constraints is equal to each other as $c_i^* = c^*$ but the elements in the row local constraints is $\vec{r}^*$. We will have another special case of the coupled constrained ensemble description, which is that the interactions among all the units in the information sources are identified in the whole process of signal generating, but the temporal dependence of each unit is different. Then the corresponding parameter will change as $\alpha_i^* = \alpha^*$. Thus, the *Hamiltonian* of this special case will be

$$H = \sum_{i=1}^{n}\sum_{j=1}^{m}(\alpha^* + \beta_j^*)x_{ji}. \tag{5.89}$$

The partition function of this coupled local constraints sequences will be

$$Z(\vec{\theta}^*) = \prod_{i=1}^{n}\prod_{j=1}^{m}[e^{-(\alpha^*+\beta_j^*)} + 1]. \tag{5.90}$$

We can get the probability of states in the sequences under this special constraints as

$$P_{\text{cam}}(\mathbf{X}|\vec{\theta}^*) = \prod_{i=1}^{n}\prod_{j=1}^{m}\frac{e^{-(\alpha^*+\beta_j^*)x_{ji}}}{e^{-(\alpha^*+\beta_j^*)} + 1}. \tag{5.91}$$

Because each element in the column local constraints are equal to each other as $c_i^* = c^*$, the average value of x_{ji} in this sequence is equal to

$$\langle x_{ji}\rangle = \frac{e^{-(\alpha^*+\beta_j^*)}}{e^{-(\alpha^*+\beta_j^*)} + 1} = \frac{r_j^*}{n}. \tag{5.92}$$

It implies the relationship between $\alpha^* + \beta_j^*$ and r_j^* as

$$e^{-(\alpha^*+\beta_j^*)} = \frac{r_j^*}{n - r_j^*}. \tag{5.93}$$

The Shannon entropy of this sequences under this local constraints is equal to

$$\begin{aligned}
S_{\text{can}}^{(K=m)} &= \ln P_{\text{can}}(\mathbf{X}^*|\vec{\theta}^*) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}[(\alpha^* + \beta_j^*)\langle x_{ji}\rangle] + \ln Z(\vec{\theta}^*) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}[-\frac{r_j^*}{n}\ln\frac{r_j^*}{n-r_j^*}] + \ln Z(\vec{\theta}^*) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}[-\frac{r_j^*}{n}\ln\frac{r_j^*}{n-r_j^*} + \ln\frac{n}{n-r_j^*}] \\
&= \sum_{j=1}^{m}[-r_j^*\ln\frac{r_j^*}{n-r_j^*} + \ln\frac{n^n}{(n-r_j^*)^n}] \\
&= \sum_{j=1}^{m}[\ln\frac{n^n}{r_j^{*r_j^*}(n-r_j^*)^{n-r_j^*}}],
\end{aligned} \tag{5.94}$$

It also equals the Shannon entropy of the sequences under one-sided row local constraints.

The microcanonical ensemble description of this special case is equal to the one where there are only row local constraints in the information sequence. Thus, we can have the *Boltzmann* entropy of this information source as

$$S_{\text{mic}}^{(K=m)} = \sum_{j=1}^{m}\ln\binom{n}{r_j^*} \tag{5.95}$$

It is also under ensemble equivalence [27].

5.D.3 Column local constraints

When all the elements in the row constraints is equal to each other as $r_j^* = r^*$, the column local constraints still remain as $\vec{c}^*$, the corresponding maximum likelihood parameter will be $\vec{\theta}^* = [\vec{\alpha}^*, \vec{\beta}^*]$, but all the elements in $\vec{\beta}^*$ is equal to each other as $\beta_j^* = \beta^*$. Then the *Hamiltonian* of this coupled constrained canonical ensemble will be

$$H = \sum_{i=1}^{n}\sum_{j=1}^{m}(\alpha_i^* + \beta^*)x_{ji}. \tag{5.96}$$

The partition function need to consider all the possible configurations as

$$Z(\vec{\theta}^*) = \prod_{i=1}^{n}\prod_{j=1}^{m}[e^{-(\alpha_i^*+\beta^*)} + 1]. \tag{5.97}$$

Probability of states under this constraint is equal to

$$P_{\text{cam}}(\mathbf{X}|\vec{\theta}^*) = \prod_{i=1}^{n}\prod_{j=1}^{m} \frac{e^{-(\alpha_i^*+\beta^*)x_{ji}}}{e^{-(\alpha_i^*+\beta^*)}+1}. \tag{5.98}$$

The average value of each element $\langle x_{ij}\rangle$ is

$$\langle x_{ij}\rangle = \frac{e^{-(\alpha_i^*+\beta^*)}}{e^{-(\alpha_i^*+\beta^*)}+1}. \tag{5.99}$$

Each element in the row local constraints is equal to each other as $r_j^* = r^*$, so the average value of element x_{ji} should also equal to the $\frac{c_i^*}{m}$. Thus, we will have the relationship follows

$$\frac{e^{-(\alpha_i^*+\beta^*)}}{e^{-(\alpha_i^*+\beta^*)}+1} = \frac{c_i^*}{m}, e^{-(\alpha_i^*+\beta^*)} = \frac{c_i^*}{m-c_i^*}. \tag{5.100}$$

The smallest space to store the information carried by the canonical ensemble described sequences is equal to the Shannon entropy of it as

$$\begin{aligned}
S_{\text{can}}^{(K=n)} =& \ln P_{\text{cam}}(\mathbf{X}^*|\vec{\theta}^*)\\
=&\vec{C}^*\cdot\vec{\theta}^* + \ln Z(\vec{\theta}^*)\\
=&\sum_{i=1}^{n}\sum_{j=1}^{m}[(\alpha_i^*+\beta^*)\frac{e^{-(\alpha_i^*+\beta^*)}}{e^{-(\alpha_i^*+\beta^*)}+1}\\
&+\ln[e^{-(\alpha_i^*+\beta^*)}+1]]\\
=&\sum_{i=1}^{n}\sum_{j=1}^{m}[-\frac{c_i^*}{m}\ln\frac{c_i^*}{m-c_i^*}+\ln\frac{m}{m-c_i^*}]\\
=&\sum_{i=1}^{n}[-\ln\frac{c_i^{*c_i^*}}{(m-c_i^*)^{c_i^*}}+\ln\frac{m^m}{(m-c_i^*)^m}]\\
=&\sum_{i=1}^{n}\ln\frac{m^m}{c_i^{*c_i^*}(m-c_i^*)^{m-c_i^*}},
\end{aligned} \tag{5.101}$$

It is the same as the Shannon entropy of the sequences under soft constraints when there is only column local constraints $\vec{c}^*$.

The microcanonical ensemble description is also equal to the one when there only has column-local constraints. Therefore, the entropy of this microcanonical ensemble description is equal to

$$S_{\text{mic}}^{(K=n)} = \sum_{i=1}^{n}\ln\binom{m}{c_i^*}. \tag{5.102}$$

According to the results in section 5.B, this coupled constrianted information sequence is under ensemble nonequivlaence.