Learning under constraints: inference, selection and validation with maximum entropy models

Giuffrida, F.

CHAPTER 2

Description length of maximum entropy models

This chapter is based on:
F. Giuffrida, T. Squartini, P. Grünwald, and D. Garlaschelli. Description length of canonical and microcanonical Models. *Phys. Rev. Research 7, 043057* (2025).

Abstract

The (non)equivalence of canonical and microcanonical ensembles is a fundamental question in statistical physics, concerning whether the use of soft and hard constraints in the maximum-entropy construction leads to the same description of a system. Despite the fact that maximum-entropy models are also commonly used in statistical inference, pattern detection, and hypothesis testing, a complete understanding of the effects of ensemble nonequivalence on statistical modeling is still missing. Here, we study this problem from a rigorous model selection perspective by comparing canonical and microcanonical models via the minimum description length principle, which yields a trade-off between likelihood, measuring model accuracy, and complexity, measuring model flexibility and its potential to overfit data. We compute the normalized maximum likelihood (NML) of both formulations and find that (1) microcanonical models always achieve higher likelihood but are always more complex; (2) the optimal model choice depends on the empirical values of the constraints—the canonical model performs best when its fit to the observed data exceeds its uniform average fit across all realizations; (3) in the thermodynamic limit, the difference in description length per node vanishes when ensemble equivalence holds but persists otherwise, showing that nonequivalence implies extensive differences between large canonical and microcanonical models. Finally, we compare the NML approach to Bayesian methods, showing that (4) the choice of priors, practically irrelevant in equivalent models, becomes crucial when an extensive number of constraints are enforced, possibly leading to very different outcomes.

2.1 Introduction

Entropy maximization provides a principled approach to inference, resulting in maximally random models under specified constraints. These models have found applications across a wide range of scientific disciplines, such as network science [31, 6, 36] and time-series analysis [37, 38].

Two distinct formulations of maximum-entropy models are possible depending on how constraints are enforced. Canonical models are defined by requiring the value of the constraints to be satisfied *on average*. Microcanonical models are defined by requiring the value of the constraints to be satisfied *exactly*. These two formulations do not necessarily lead to the same description of a given system, i.e., they are, in general, *non-equivalent*.

The concept of ensemble equivalence, introduced by Gibbs [2], plays an important role in statistical mechanics. Traditional studies have shown that canonical and microcanonical ensembles tend to become equivalent in the thermodynamic limit for systems with short-range interactions. Violations of this asymptotic equivalence, known as ensemble non-equivalence, have historically been linked to systems with long-range interactions or other forms of non-additivity, such as mean-field spin models or self-gravitating systems [39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 7].

More recently, it has become clear that ensemble non-equivalence can also arise due to an additional mechanism: the presence of an *extensive number of constraints* [12, 13, 14, 15]. This has been demonstrated, for example, in network models with local constraints, such as the Configuration Model. In these cases – unlike the classical examples from statistical physics – non-equivalence is not necessarily associated with long-range interactions or phase transitions, and is not confined to specific parameter regimes, but is present across the entire parameter space. Moreover, this phenomenon can have practical implications for the expected macroscopic properties of the system under study (see [50] for an example concerning bipartite networks).

Whenever the two approaches are equivalent, they identify the same set of typical configurations. In these cases, preferring one formulation over the other becomes a matter of convenience (e.g., computational efficiency). Whenever the two approaches are non-equivalent, the choice depends on the specific characteristics of the system under study and the nature of the constraints involved. For example, in the context of choosing a null model to be compared with real data, for e.g., pattern detection, it has been argued that, when non-equivalence holds, the ensemble choice should be based on a theoretical guiding principle [31, 14, 51, 15, 52]. In this case, if the data are expected to be noisy, meaning that the measured values of the constraints generally do not match the true (unobservable) ones, one should prefer the canonical version of the null model because it allows for fluctuations in the constraint quantities, while the microcanonical one would paradoxically assign zero probability to the true configuration of the system. Conversely, if, by hypothesis, the measured values of the constraints are not affected by error or noise, the microcanonical model should be preferred.

When there is no prior expectation available about the presence or absence of noise in the empirical values of the constraints, the decision between hard and soft

constraints should be based on posterior evidence, i.e., which of the two null models best describes the data. This immediately leads to a problem of model selection between canonical and microcanonical ensembles.

Information theory provides a powerful framework for model selection by offering criteria that capture the trade-off between a model's "goodness of fit," quantified by its log-likelihood, and its "complexity." In general, model complexity measures the flexibility of a model in fitting different datasets. In other words, it measures how easily the model can overfit the data. A common way to quantify model complexity is by counting the number k of the model's free parameters, which, in the case of maximum-entropy models, corresponds to the number of constraints. This is the case of the two historically most popular model selection criteria: the *Akaike Information Criterion* (AIC) [10] and the *Bayesian Information Criterion* (BIC) [11], according to which one should pick the model minimizing

$$\mathrm{AIC} = -\ln \mathcal{L} + k \tag{2.1}$$

and

$$\mathrm{BIC} = -\ln \mathcal{L} + \frac{k}{2} \ln V, \tag{2.2}$$

respectively, where $\mathcal{L}$ is the model likelihood and V is the sample size. Canonical and microcanonical models defined by the same set of constraints share the same number of parameters. Therefore, according to both AIC and BIC, their complexity is the same, and a comparison based on these criteria would reduce to a mere comparison between their log-likelihood functions. This would naively lead to the conclusion that the microcanonical model is always the best-scoring one since it has the highest likelihood. However, neither AIC nor BIC can be computed for microcanonical models since they are derived under regularity assumptions that are not even remotely met by the latter. Moreover, both criteria are asymptotic results derived under the assumption of a finite number of parameters as the system size approaches infinity [9]. This assumption is not met by models whose number of parameters scales with the size of the system.

To overcome these issues, here we consider the Minimum Description Length (MDL) principle [16, 17, 18], a family of information-theoretic approaches to model selection seeking the model that provides the shortest description of data. Specifically, we focus on the approach based on the Normalized Maximum Likelihood (NML), or Shtarkov, distribution [53], with the aim of investigating the trade-off between the difference of log-likelihoods and that of complexities in light of ensemble non-equivalence.

A particularly useful definition of non-equivalence, naturally connected to information theory, is the *measure-level* one, which involves the Kullback-Leibler (KL) divergence between the microcanonical and canonical probability distributions. Under mild conditions, this approach is equivalent to other definitions commonly used in statistical physics [7]. According to this definition, (non-)equivalence is characterized by a (non-)vanishing relative entropy density as the system size approaches infinity. Interestingly, the KL divergence between canonical and microcanonical distributions coin-

cides with the difference between the corresponding log-likelihood functions [12, 13]. Thus, measure-level (non-)equivalence can be inspected by simply considering the asymptotic behavior of this log-likelihood difference [12], providing a natural connection to statistical inference.

While the impact of non-equivalence on the difference between log-likelihoods is relatively well understood [12, 14], its effects on the difference between complexities remain largely unexplored. To encompass a broad spectrum of applications, our study focuses on maximum-entropy models suitable for studying systems represented by binary rectangular matrices (e.g., bipartite networks, time series). More precisely, we aim at determining i) whether the influence of non-equivalence persists even when the complexity term is accounted for, and ii) whether non-equivalent canonical and microcanonical models should, in fact, be regarded as statistically distinct models. We conclude that the answer to both questions is yes, raising the question of whether non-equivalence can be defined at the model selection level. Furthermore, we show that in the presence of extensively many constraints, the choice of prior in a Bayesian framework can critically affect the resulting description length, underscoring the importance of this choice in non-equivalent settings.

The rest of the paper is organized as follows. Section 2.2 introduces MDL for discrete data and the Normalized Maximum Likelihood (NML) approach to define description lengths. In Section 2.3, we provide a general expression for the difference between the description lengths of canonical and microcanonical models as a function of the Kullback-Leibler divergence between the two distributions. In Section 2.4, we explicitly derive and compare the description lengths of both the canonical and the microcanonical formulations of models defined by one global constraint (i.e., the sum of the entries of a matrix) as well as one-sided local constraints (i.e., the row-specific sums of a matrix). While the first constraint leads to ensemble equivalence, the second set of constraints leads to ensemble non-equivalence. Finally, Section 2.5 reviews the relationship between the NML-based and Bayesian approaches when non-equivalence holds.

2.2 The Minimum Description Length principle

The MDL principle embodies the idea that the best model to describe a given dataset is the one providing the shortest description. We now introduce the basic formalism by focusing on the discrete data $\boldsymbol{x} \in \mathcal{X}$, where $\mathcal{X}$ represents the space of all possible samples of fixed size V. A parametric statistical model $\mathcal{M}$ consists of a family of probability distributions sharing the same functional form, i.e.,

$$\mathcal{M} = \{P_{\mathcal{M}}(\boldsymbol{x}; \boldsymbol{\theta})\}_{\boldsymbol{\theta} \in \Theta} \tag{2.3}$$

where $\boldsymbol{\theta}$ is the vector of parameters whose values lie in the set Θ.

In order to apply the MDL principle, one needs to compute the description length of the data $\boldsymbol{x}$ provided by a model $\mathcal{M}$. By Kraft's inequality [16, 54], the optimal number of natural digits (nats) needed to describe the outcome $\boldsymbol{x}$ of a probability distribution $P(\boldsymbol{x})$ is given by $-\ln P(\boldsymbol{x})$. Since, however, $\mathcal{M}$ encompasses many distributions corresponding to different parameter values, a natural choice is that of

considering the description length provided by the distribution minimizing the quantity $-\ln P_{\mathcal{M}}(\boldsymbol{x};\boldsymbol{\theta})$, reading

$$\mathcal{L}_{\mathcal{M}}(\boldsymbol{x}) \equiv P_{\mathcal{M}}(\boldsymbol{x};\hat{\boldsymbol{\theta}}(\boldsymbol{x})), \tag{2.4}$$

with $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ being the maximum-likelihood (ML) estimator of $\boldsymbol{\theta}$, i.e.

$$\hat{\boldsymbol{\theta}}(\boldsymbol{x}) = \arg\max_{\boldsymbol{\theta}\in\Theta} P_{\mathcal{M}}(\boldsymbol{x};\boldsymbol{\theta}). \tag{2.5}$$

However, the distribution $\mathcal{L}_{\mathcal{M}}(\boldsymbol{x})$ is affected by (at least) two problems having the same origin: the parameters appearing in its definition are evaluated for each individual data in $\mathcal{X}$. As a consequence, i) $\mathcal{L}_{\mathcal{M}}(\boldsymbol{x})$ cannot be a probability distribution with support in $\mathcal{X}$ as it is not properly normalized, and ii) adopting $\mathcal{L}_{\mathcal{M}}(\boldsymbol{x})$ may lead to severe overfitting since it is defined *a posteriori* relative to the observations.

To solve these problems, one can build a so-called *universal distribution* relative to $\mathcal{M}$, i.e., a unique, representative distribution $\bar{P}_{\mathcal{M}}(\boldsymbol{x})$ of the statistical model $\mathcal{M}$. Once equipped with a universal distribution, the description length of $\boldsymbol{x}$ through is $\mathcal{M}$ is defined as

$$\mathrm{DL}^{\bar{P}}_{\mathcal{M}}(\boldsymbol{x}) \equiv -\ln \bar{P}_{\mathcal{M}}(\boldsymbol{x}). \tag{2.6}$$

Universal distributions (which don't necessarily belong to $\mathcal{M}$) can be defined in several ways. In what follows, we will consider the MDL recipe based upon the NML universal distribution (for a more formal introduction to universal coding, we redirect the interested reader to [16]).

2.2.1 The Normalized Maximum Likelihood universal distribution

The NML distribution corresponds to the distribution attaining the minimum pointwise *redundancy* (or *regret*) in the worst-case scenario [16], i.e.,

$$\bar{P}^{\mathrm{NML}}_{\mathcal{M}} \equiv \arg\min_{\bar{P}} \max_{\boldsymbol{x}\in\mathcal{X}} \left\{ -\ln \bar{P}(\boldsymbol{x}) - [-\ln P_{\mathcal{M}}(\boldsymbol{x};\hat{\boldsymbol{\theta}}(\boldsymbol{x}))] \right\}. \tag{2.7}$$

In words, it is the distribution that *a priori* requires the minimum number of nats when compared to the shortest description length *a posteriori*, that is $-\ln P_{\mathcal{M}}(\boldsymbol{x};\hat{\boldsymbol{\theta}}(\boldsymbol{x}))$. The minimum is found in the worst-case scenario, i.e., for the data $\boldsymbol{x}$ for which this difference is maximal. The unique solution to this *minimax* problem reads

$$\bar{P}^{\mathrm{NML}}_{\mathcal{M}}(\boldsymbol{x}) = \frac{P_{\mathcal{M}}(\boldsymbol{x};\hat{\boldsymbol{\theta}}(\boldsymbol{x}))}{\sum_{\boldsymbol{y}\in\mathcal{X}} P_{\mathcal{M}}(\boldsymbol{y};\hat{\boldsymbol{\theta}}(\boldsymbol{y}))}, \tag{2.8}$$

which is a normalized maximum likelihood. Following the universal coding approach, the description length of $\boldsymbol{x}$ through model $\mathcal{M}$ is computed as

$$\begin{aligned} \mathrm{DL}^{\mathrm{NML}}_{\mathcal{M}}(\boldsymbol{x}) &\equiv -\ln \bar{P}^{\mathrm{NML}}_{\mathcal{M}}(\boldsymbol{x}) \\ &= -\ln P_{\mathcal{M}}(\boldsymbol{x};\hat{\boldsymbol{\theta}}(\boldsymbol{x})) + \ln \sum_{\boldsymbol{y}\in\mathcal{X}} P_{\mathcal{M}}(\boldsymbol{y};\hat{\boldsymbol{\theta}}(\boldsymbol{y})). \end{aligned} \tag{2.9}$$

While the first term is (minus) a model maximum log-likelihood, the second term

$$\mathrm{COMP}_{\mathcal{M}} \equiv \ln \sum_{\boldsymbol{y}\in\mathcal{X}} P_{\mathcal{M}}(\boldsymbol{y};\hat{\boldsymbol{\theta}}(\boldsymbol{y})) = \ln \sum_{\boldsymbol{y}\in\mathcal{X}} \mathcal{L}_{\mathcal{M}}(\boldsymbol{y}) \tag{2.10}$$

is named *parametric complexity* of $\mathcal{M}$. Notice that $\mathcal{X}$ represents the set of all possible data sets of fixed size V, i.e., all the observable samples of size V. Thus, $\mathrm{COMP}_{\mathcal{M}}$ depends on both the model and $\mathcal{X}$. Finally, we can rewrite equation (2.9) as

$$\mathrm{DL}^{\mathrm{NML}}_{\mathcal{M}}(\boldsymbol{x}) = -\ln \mathcal{L}_{\mathcal{M}}(\boldsymbol{x}) + \mathrm{COMP}_{\mathcal{M}}. \tag{2.11}$$

Similarly to other criteria rooted in information theory, this description length consists of two terms: the log-likelihood term, evaluated on the data and favoring models with a high goodness-of-fit, and the complexity term, depending only on the chosen model and penalizing the ones that are too complex.

Once we have evaluated the description length of our data according to a basket of competing models, the MDL principle prescribes selecting the one providing the minimum DL.

2.3 Description length of canonical and microcanonical models

Although computing the complexity term within the NML-based description length can be challenging, models defined by a sufficient statistic admit a simpler expression for this term. This section introduces an important class of models of this kind, the maximum-entropy models (MEMs), both in their canonical and microcanonical formulations.

Specifically, we will focus on discrete data that can be represented as a binary $n \times m$ matrix $\mathbf{G}$, i.e.

$$\mathbf{G} = \{g_{ij}\}_{\substack{i=\{1\ldots n\}\\ j=\{1\ldots m\}}} \tag{2.12}$$

and denote by $\mathcal{G}$ the set of all such matrices that, when endowed with a probability distribution $P(\mathbf{G})$, constitute a proper ensemble of matrices. The matrix representation is often used to describe systems composed of n elements to be modeled (corresponding to the number of rows), each characterized by m state variables or degrees of freedom (corresponding to the number of columns). The size of the data set is fixed and supposed to represent the number of independent entries of the matrix, i.e., nm.

In the most common scenario, we can access a vector of quantities $\mathbf{c}^* = \mathbf{c}(\mathbf{G}^*)$, evaluated on the only available matrix $\mathbf{G}^*$, and representing the sufficient statistic of the model. Maximum-entropy models are, then, obtained by looking for the probability distribution that maximizes Shannon entropy

$$\mathcal{S}[P] = -\sum_{\mathbf{G}\in\mathcal{G}} P(\mathbf{G}) \ln P(\mathbf{G}) \tag{2.13}$$

while constraining the sufficient statistics. If the model is required to reproduce the value of the constraints exactly, i.e., $\mathbf{c}(\mathbf{G}) = \mathbf{c}^*$, one obtains a *microcanonical* model.

If, instead, the model is required to reproduce the value of the constraints on average, i.e., $\langle \mathbf{c}(\mathbf{G}) \rangle_P = \mathbf{c}^*$ with $\langle \cdot \rangle_P$ being the ensemble average induced by distribution P, one obtains the corresponding *canonical* formulation. This approach, introduced by Gibbs [2] and reprised by Jaynes [1], leads to ensembles of matrices that reproduce (either exactly or on average) the constrained quantities while randomizing everything else.

In what follows, we will adopt the NML-based approach to compute the description lengths of microcanonical and canonical MEMs, which is optimal (in a minimax sense) when assuming no prior knowledge on the parameters. Thus, the description length is taken as the one defined in equation (2.11). For models admitting a sufficient statistic, the complexity term takes the convenient form

$$\mathrm{COMP} = \ln \sum_{\mathbf{c} \in \mathcal{C}} \Omega(\mathbf{c}) \mathcal{L}(\mathbf{c}), \tag{2.14}$$

where $\mathcal{L}(\mathbf{c})$ denotes the value of the maximum likelihood in correspondence with a sufficient statistic reading $\mathbf{c}$ and $\Omega(\mathbf{c})$ denotes the number of configurations with $\mathbf{c}(\mathbf{G}) = \mathbf{c}$, i.e.,

$$\Omega(\mathbf{c}) \equiv \sum_{\mathbf{G} \in \mathcal{G}} \delta_{\mathbf{c}(\mathbf{G}), \mathbf{c}} \tag{2.15}$$

with $\delta_{i,j}$ being the Kronecker delta. As this set may be empty for some values of $\mathbf{c}$, we will solely focus on the *graphical values*, i.e., the set $\mathcal{C} = \{\mathbf{c} : \Omega(\mathbf{c}) \neq \emptyset\}$ of values that are verified by at least one configuration, and denote it with $\mathcal{N}_{\mathcal{C}} = |\mathcal{C}|$ its cardinality. In other words, the sum in equation (2.14) runs over the graphical values of the sufficient statistics rather than over all possible data.

2.3.1 Description length of microcanonical models

When considering microcanonical models, entropy maximization yields the following functional form

$$P_{\mathrm{mic}}(\mathbf{G}; \mathbf{c}) = \begin{cases} \frac{1}{\Omega(\mathbf{c})} & \text{if } \mathbf{c}(\mathbf{G}) = \mathbf{c} \\ 0 & \text{else,} \end{cases} \tag{2.16}$$

where $\mathbf{c}$ is the vector of parameters whose values lie in the discrete set $\mathcal{C}$ - for microcanonical models, in fact, the vector of parameters corresponds to the sufficient statistics itself - and $P_{\mathrm{mic}}(\mathbf{G}; \mathbf{c})$ is a uniform probability whose mass differs from zero only in correspondence of those configurations that verify the condition $\mathbf{c}(\mathbf{G}) = \mathbf{c}$. The derivation of this probability functional form as the solution of entropy maximization under hard constraints is given in more detail in Section 2.A of the Appendix.

We will now evaluate the description length of a microcanonical model. Its maximum log-likelihood simply reads

$$\ln \mathcal{L}_{\mathrm{mic}}(\mathbf{c}) = \ln \frac{1}{\Omega(\mathbf{c})} = -\ln \Omega(\mathbf{c}). \tag{2.17}$$

While combining this equation with equation (2.14), one obtains the following expression for the related complexity:

$$\mathrm{COMP}_{\mathrm{mic}} = \ln \sum_{\mathbf{c} \in \mathcal{C}} \Omega(\mathbf{c}) \cdot \frac{1}{\Omega(\mathbf{c})} = \ln \sum_{\mathbf{c} \in \mathcal{C}} 1 = \ln \mathcal{N}_{\mathcal{C}}. \tag{2.18}$$

Hence, the description length of a microcanonical maximum-entropy model is

$$\mathrm{DL}_{\mathrm{mic}}^{\mathrm{NML}}(\mathbf{G}^*) = \ln \Omega(\mathbf{c}^*) + \ln \mathcal{N}_{\mathcal{C}}. \tag{2.19}$$

2.3.2 Description length of canonical models

When considering canonical models, entropy maximization yields the following well-known exponential form

$$P_{\mathrm{can}}(\mathbf{G};\boldsymbol{\theta}) = \frac{e^{-\boldsymbol{\theta}\cdot\mathbf{c}(\mathbf{G})}}{Z(\boldsymbol{\theta})} \tag{2.20}$$

with $Z(\boldsymbol{\theta}) = \sum_{\mathbf{G}\in\mathcal{G}} e^{-\boldsymbol{\theta}\cdot\mathbf{c}(\mathbf{G})}$ being the *partition function*. In the context of network theory, the maximum-entropy canonical models are also known as *Exponential Random Graph Models* [4, 5, 6]. In this case, the vector of parameters $\boldsymbol{\theta}$ corresponds to the vector of Lagrange multipliers needed to carry out a constrained entropy maximization. The ML estimators $\hat{\boldsymbol{\theta}}(\mathbf{G})$ are found by solving the system of equations

$$\langle \mathbf{c} \rangle_{\hat{\boldsymbol{\theta}}(\mathbf{G})} = \mathbf{c}(\mathbf{G}) \tag{2.21}$$

with $\langle\cdot\rangle_{\boldsymbol{\theta}}$ representing the ensemble average induced by $P_{\mathrm{can}}(\mathbf{G};\boldsymbol{\theta})$.

The maximum log-likelihood of a canonical model is linked to the microcanonical one through the Kullback-Leibler divergence:

$$S(P_{\mathrm{mic}}||P_{\mathrm{can}})|_{\mathbf{c},\boldsymbol{\theta}} = \sum_{\mathbf{G}\in\mathcal{G}} P_{\mathrm{mic}}(\mathbf{G};\mathbf{c}) \ln \frac{P_{\mathrm{mic}}(\mathbf{G};\mathbf{c})}{P_{\mathrm{can}}(\mathbf{G};\boldsymbol{\theta})}. \tag{2.22}$$

In fact, it can be shown [13, 12] that

$$\begin{aligned} D_{\mathrm{KL}}(\mathbf{c}^*) &\equiv S(P_{\mathrm{mic}}||P_{\mathrm{can}})|_{\mathbf{c}^*,\boldsymbol{\theta}^*} \\ &= \ln \frac{P_{\mathrm{mic}}(\mathbf{G}^*;\mathbf{c}^*)}{P_{\mathrm{can}}(\mathbf{G}^*;\boldsymbol{\theta}^*)} \\ &= \ln \mathcal{L}_{\mathrm{mic}}(\mathbf{c}^*) - \ln \mathcal{L}_{\mathrm{can}}(\mathbf{c}^*) \end{aligned} \tag{2.23}$$

where $\boldsymbol{\theta}^* = \hat{\boldsymbol{\theta}}(\mathbf{c}^*)$. Hence,

$$\mathcal{L}_{\mathrm{can}}(\mathbf{c}) = \mathcal{L}_{\mathrm{mic}}(\mathbf{c}) \cdot e^{-D_{\mathrm{KL}}(\mathbf{c})} = \frac{1}{\Omega(\mathbf{c})} \cdot e^{-D_{\mathrm{KL}}(\mathbf{c})} \tag{2.24}$$

for any value of the sufficient statistic, and

$$\begin{aligned} \mathrm{COMP}_{\mathrm{can}}^{\mathrm{NML}} &= \ln \sum_{\mathbf{c}\in\mathcal{C}} \Omega(\mathbf{c}) \cdot \frac{1}{\Omega(\mathbf{c})} \cdot e^{-D_{\mathrm{KL}}(\mathbf{c})} \\ &= \ln \sum_{\mathbf{c}\in\mathcal{C}} e^{-D_{\mathrm{KL}}(\mathbf{c})}. \end{aligned} \tag{2.25}$$

Consequently, the description length of a canonical model can be expressed as

$$\begin{aligned} \mathrm{DL}_{\mathrm{can}}^{\mathrm{NML}}(\mathbf{G}^*) &= -\ln \mathcal{L}_{\mathrm{can}}(\mathbf{c}^*) + \ln \sum_{\mathbf{c}\in\mathcal{C}} e^{-D_{\mathrm{KL}}(\mathbf{c})} \\ &= \ln \Omega(\mathbf{c}^*) + D_{\mathrm{KL}}(\mathbf{c}^*) + \ln \sum_{\mathbf{c}\in\mathcal{C}} e^{-D_{\mathrm{KL}}(\mathbf{c})}. \end{aligned} \tag{2.26}$$

2.3.3 Comparing canonical and microcanonical models

We can now compare the description length of canonical and microcanonical models. Given that D_{KL} is guaranteed to be non-negative, we have

$$\begin{aligned}\Delta \ln \mathcal{L}(\mathbf{G}^*) &\equiv \ln \mathcal{L}_{\mathrm{can}}(\mathbf{G}^*) - \ln \mathcal{L}_{\mathrm{mic}}(\mathbf{G}^*) \\ &= -D_{\mathrm{KL}}(\mathbf{c}^*) \leqslant 0 \end{aligned} \tag{2.27}$$

i.e., the canonical likelihood is always smaller than or equal to the microcanonical one. For the same reason, it is straightforward to see that the canonical complexity is always smaller than, or equal to, the microcanonical one:

$$\begin{aligned}\Delta\mathrm{COMP} &\equiv \mathrm{COMP}_{\mathrm{can}} - \mathrm{COMP}_{\mathrm{mic}} \\ &= \ln \frac{\sum_{\mathbf{c}\in\mathcal{C}} e^{-D_{KL}(\mathbf{c})}}{\mathcal{N}_{\mathcal{C}}} \leqslant 0. \end{aligned} \tag{2.28}$$

Thus, microcanonical models achieve a higher goodness-of-fit but, at the same time, are more complex. The interplay between the two differences determines the difference between the canonical and the microcanonical description length:

$$\begin{aligned}\Delta\mathrm{DL}(\mathbf{G}^*) &\equiv \mathrm{DL}^{\mathrm{NML}}_{\mathrm{can}}(\mathbf{G}^*) - \mathrm{DL}^{\mathrm{NML}}_{\mathrm{mic}}(\mathbf{G}^*) \\ &= -\Delta \ln \mathcal{L}(\mathbf{G}^*) + \Delta\mathrm{COMP} \\ &= D_{\mathrm{KL}}(\mathbf{c}^*) + \ln \frac{\sum_{\mathbf{c}\in\mathcal{C}} e^{-D_{\mathrm{KL}}(\mathbf{c})}}{\mathcal{N}_{\mathcal{C}}}. \end{aligned} \tag{2.29}$$

The following reasoning provides a different way of looking at the expression above. The canonical distribution $P_{\mathrm{can}}(\mathbf{G};\boldsymbol{\theta})$ induces the following distribution on the values of the sufficient statistic

$$Q_{\mathrm{can}}(\mathbf{c};\boldsymbol{\theta}) = \Omega(\mathbf{c}) P_{\mathrm{can}}(\mathbf{c};\boldsymbol{\theta}) \tag{2.30}$$

where $P_{\mathrm{can}}(\mathbf{c};\boldsymbol{\theta})$ is the canonical probability of a matrix $\mathbf{G}$, such that $\mathbf{c}(\mathbf{G}) = \mathbf{c}$, for a given value of the vector of parameters $\boldsymbol{\theta}$. The maximum likelihood assigned to $\mathbf{c}$ by the correspondent model $\{Q_{\mathrm{can}}(\mathbf{c},\boldsymbol{\theta})\}_{\theta\in\Theta}$ is, thus,

$$\begin{aligned} Q_{\mathrm{can}}(\mathbf{c};\hat{\boldsymbol{\theta}}(\mathbf{c})) &= \Omega(\mathbf{c}) P_{\mathrm{can}}(\mathbf{c};\hat{\boldsymbol{\theta}}(\boldsymbol{c})) \\ &= \frac{\mathcal{L}_{\mathrm{can}}(\mathbf{c})}{\mathcal{L}_{\mathrm{mic}}(\mathbf{c})} = e^{-D_{\mathrm{KL}}(\mathbf{c})} \end{aligned} \tag{2.31}$$

and

$$\Delta\mathrm{DL}(\mathbf{G}^*) = -\ln Q_{\mathrm{can}}(\mathbf{c}^*;\boldsymbol{\theta}^*) + \ln \frac{\sum_{\mathbf{c}\in\mathcal{C}} Q_{\mathrm{can}}(\mathbf{c};\hat{\boldsymbol{\theta}}(\mathbf{c}))}{\mathcal{N}_{\mathcal{C}}}; \tag{2.32}$$

upon defining

$$\bar{Q}_{\mathrm{can}} \equiv \frac{\sum_{\mathbf{c}\in\mathcal{C}} Q_{\mathrm{can}}(\mathbf{c},\hat{\boldsymbol{\theta}}(\mathbf{c}))}{\mathcal{N}_{\mathcal{C}}} \tag{2.33}$$

as the uniform average of $Q_{\mathrm{can}}(\mathbf{c};\hat{\boldsymbol{\theta}}(\mathbf{c}))$ over all possible values of $\mathbf{c}\in\mathcal{C}$, we find that the difference between description lengths can be expressed as

$$\Delta\mathrm{DL}(\mathbf{G}^*) = \ln \frac{\bar{Q}_{\mathrm{can}}}{Q^*_{\mathrm{can}}}, \tag{2.34}$$

where $Q^*_{\text{can}} \equiv Q_{\text{can}}(\mathbf{c}^*, \boldsymbol{\theta}^*)$. This expression depends only on the canonical model, providing a new interpretation of the comparison between the canonical and the microcanonical description length: the canonical model provides a shorter description length whenever $\Delta\text{DL}(\mathbf{G}^*) < 0$, i.e., whenever $Q^*_{\text{can}} > \bar{Q}_{\text{can}}$. In other words, the canonical model is the best-scoring model whenever the maximum likelihood it assigns to the observed value $\mathbf{c}^*$ is higher than the maximum likelihood averaged over all possible values of the sufficient statistic (or, more intuitively, if it performs 'better than on average' on the observed sufficient statistic). Remarkably, this is in open contrast with a comparison based only on the likelihoods, according to which the microcanonical formulation should always be preferred, independently of the observed data.

Expression (2.29) shows that the entity of the difference between the two description lengths crucially depends on the value of the KL divergence. As said in the introduction, the canonical and the microcanonical ensembles are (non-)equivalent if the relative D_{KL} is (non-)vanishing as the system size approaches infinity. When studying matrices, identifying the system size with the number of entries, i.e., nm, may be inappropriate, as adding (only) one entry to a matrix without altering its structure is, in fact, impossible. In these cases, it may be much more reasonable to identify the system size with the number n of items: increasing the system size from n to $n+1$ would, thus, correspond to adding a node to a network or a time-point to a time series.

In this scenario, the canonical and the microcanonical ensembles are equivalent in the measure sense [7] if

$$\lim_{n\to\infty} \frac{D_{\text{KL}}(\mathbf{c}^*)}{n} = \lim_{n\to\infty} \frac{|\Delta \ln \mathcal{L}(\mathbf{c}^*)|}{n} = 0 \tag{2.35}$$

or, equivalently, if

$$D_{\text{KL}}(\mathbf{c}^*) = |\Delta \ln \mathcal{L}(\mathbf{c}^*)| = o(n). \tag{2.36}$$

where $o(x)$ stands for a quantity that goes to 0 when divided by x as $n \to +\infty$.

This definition of non-equivalence relies on comparing the likelihood functions of the two models. We aim to extend it in order to take into account the models' complexities. As a first step, we consider the order of ΔDL. Notice that

$$D_{\text{KL}}(\mathbf{c}) = o(n) \quad \forall \mathbf{c} \in \mathcal{C} \quad \Rightarrow \quad \Delta\text{DL}(\mathbf{c}) = o(n) \quad \forall \mathbf{c} \in \mathcal{C} \tag{2.37}$$

as it follows directly from (2.29). Thus, if the two ensembles are equivalent, the difference in description lengths between the corresponding models will be $o(n)$. However, a similar conclusion cannot be immediately derived when non-equivalence holds and D_{KL} grows at least linearly with n. Indeed, the likelihood and complexity terms in (2.29) are of the same order and have opposite signs; therefore, evaluating the order of ΔDL is a non-trivial task in this case. The following section addresses this problem by explicitly evaluating ΔDL in two notable examples of maximum-entropy models.

2.4 Applications to maximum-entropy models

The simplest constraint one may imagine to enforce is the sum of the elements of a matrix, i.e.

$$l(\mathbf{G}) \equiv \sum_{i=1}^{n}\sum_{j=1}^{m} g_{ij}. \tag{2.38}$$

In the case of bipartite binary networks, this quantity corresponds to the total number of links and induces the well-known Erdös-Rényi model. While this model is one of the most popular ones, it is also widely recognized to fall short in capturing the topological heterogeneity that characterizes most real-world systems. An alternative approach to introduce heterogeneity is to constrain the sum of elements of each row, i.e.

$$r_i(\mathbf{G}) \equiv \sum_{j=1}^{m} g_{ij} \quad i = 1 \dots n. \tag{2.39}$$

In the case of bipartite binary networks, the sequence $r_i(\mathbf{G})$, $i = 1 \dots n$ coincides with the degree sequence of the nodes belonging to just one layer and induces the canonical MEM known as Bipartite Partial Configuration Model [55]. As we will employ asymptotic results to compare the canonical and the microcanonical description lengths, we need to specify the scaling of the constraints $\mathbf{c}^*$ as $n \to \infty$. In what follows, we will focus on *dense matrices*, i.e., we will assume that $g^*_{ij} = \Theta(1)\ \forall (i,j)$, where $\Theta(x)$ indicates a quantity of the same order of x as $n \to \infty$. This choice leads to $r^*_i = \Theta(m)$ and $l^* = \Theta(nm)$, the 'actual' order of these quantities depending on the growth rate of m. We will consider two cases: (i) m finite in the limit, i.e., $m = \Theta(1)$ and (ii) m growing linearly with n, i.e., $m = \Theta(n)$. As a consequence of our assumption, the matrix density

$$p^* = \frac{l^*}{nm} \tag{2.40}$$

and the row densities

$$p^*_i = \frac{r^*_i}{m} \quad i = 1 \dots n \tag{2.41}$$

will remain finite in the asymptotic limit, i.e., $p^* = \Theta(1)$ and $p^*_i = \Theta(1)$, $\forall\, i$.

2.4.1 Enforcing one global constraint

Since we are considering binary matrices, constraining the total number of 1s in a microcanonical fashion leads to a number of configurations equal to $\Omega(l) = \binom{nm}{l}$. The corresponding microcanonical distribution, thus, becomes

$$P_{\text{mic}}(\mathbf{G}; l) = \begin{cases} \frac{1}{\binom{nm}{l}} & \text{if } l(\mathbf{G}) = l \\ 0 & \text{else,} \end{cases} \tag{2.42}$$

inducing the maximum log-likelihood

$$\ln \mathcal{L}_{\text{mic}}(l) = -\ln \binom{nm}{l}. \tag{2.43}$$

The complexity equals the logarithm of the number of graphical values, i.e.

$$\mathrm{COMP}_{\mathrm{mic}}^{\mathrm{NML}} = \sum_{l=0}^{nm} 1 = \ln[nm+1]. \tag{2.44}$$

The microcanonical description length, thus, reads

$$\mathrm{DL}_{\mathrm{mic}}^{\mathrm{NML}}(\mathbf{G}^*) = \ln \binom{nm}{l^*} + \ln[nm+1]. \tag{2.45}$$

The canonical probability can be computed explicitly, as this model represents one of the few canonical MEMs whose partition function can be computed analytically. Specifically, it reads $Z(\theta) = (1+e^{\theta})^{nm}$ and induces the expression

$$P_{\mathrm{can}}(\mathbf{G};\theta) = \frac{e^{-\theta l(\mathbf{G})}}{(1+e^{-\theta})^{nm}}. \tag{2.46}$$

Upon defining

$$p \equiv \frac{e^{-\theta}}{1+e^{-\theta}}, \tag{2.47}$$

we can re-write the canonical probability via the so-called *mean-value parametrization* leading to the expression

$$P_{\mathrm{can}}(\mathbf{G};p) = p^{l(\mathbf{G})}(1-p)^{nm-l(\mathbf{G})} \tag{2.48}$$

i.e., the distribution of a collection of nm i.i.d. Bernoulli variables, indicating that each matrix element is either 1, with probability p, or 0, with probability $1-p$. The ML estimator of p is given by the matrix density $\hat{p}(\mathbf{G}) = l(\mathbf{G})/nm$ and the maximum log-likelihood, obtained by inserting $\hat{p}$ into (2.48), reads

$$\ln \mathcal{L}_{\mathrm{can}}(l) = -nm \cdot h\left(\frac{l}{nm}\right) \tag{2.49}$$

where $h(p) = -p\ln(p) - (1-p)\ln(1-p)$ is the entropy of a Bernoulli distribution with a parameter equal to p. An exact expression for the canonical complexity has been derived in [56] and reads

$$\mathrm{COMP}_{\mathrm{can}}^{\mathrm{NML}} = \ln\left[\frac{e^{nm}\Gamma(nm,nm)}{nm^{nm-1}} + 1\right] \tag{2.50}$$

where $\Gamma(s,x)$ is the upper incomplete gamma function. In case s is a positive integer (as it happens to be our case), it can be expressed as

$$\Gamma(s,x) = e^{-x}(s-1)!\sum_{k=0}^{s-1}\frac{x^k}{k!}. \tag{2.51}$$

In conclusion, the canonical description length reads

$$\mathrm{DL}_{\mathrm{can}}^{\mathrm{NML}}(\mathbf{G}^*) = nm \cdot h\left(\frac{l^*}{nm}\right) + \ln\left[\frac{e^{nm}\Gamma(nm,nm)}{nm^{nm-1}} + 1\right]. \tag{2.52}$$

Notably, the details of the matrix structure play no role as $\mathrm{DL}_{\mathrm{can}}^{\mathrm{NML}}$ solely depend on global quantities, namely the number of 1s and the sample size - in our case, nm.

So far, all the reported results are exact. Nevertheless, as we are interested in the order of ΔDL when $n \to \infty$, we consider the following asymptotic results, derived from the exact ones (see section 2.B.1 of the Appendix):

$$\Delta \ln \mathcal{L}(\mathbf{G}^*) = -\frac{1}{2} \ln[nm] - \frac{1}{2} \ln[2\pi p^*(1-p^*)] + o(1), \tag{2.53}$$

$$\Delta\mathrm{COMP} = -\frac{1}{2} \ln[nm] + \frac{1}{2} \ln\left[\frac{\pi}{2}\right] + o(1) \tag{2.54}$$

and, upon combining them, the (asymptotic) difference between description lengths becomes

$$\Delta\mathrm{DL}(\mathbf{G}^*) = \frac{1}{2} \ln[\pi^2 p^*(1-p^*)] + o(1). \tag{2.55}$$

The expressions above provide two significant insights. First, the order of the difference between log-likelihood functions is $o(n)$, regardless of the growth rate of m, a result implying that condition (2.36) is met and ensemble equivalence holds. Second, the leading terms of expressions (2.53) and (2.54) are identical, thus canceling out when compared. Consequently, the (asymptotic) difference between description lengths depends on the size of the system only through the matrix density p^*, assumed to be finite in the dense regime. Thus, such a difference is finite in the same regime, in agreement with (2.37).

2.4.2 Enforcing n local constraints

Let us now consider canonical and microcanonical models induced by an extensive number of parameters, each one corresponding to the sum of the elements of a different row. With this choice, the matrix can be viewed as a collection of n independent strips of size $1 \times m$. Since we impose a single constraint per row, all the prior results hold row-wise, with m replacing nm and r_i replacing l. Specifically, the microcanonical probability distribution reads

$$P_{\mathrm{mic}}(\mathbf{G};\mathbf{r}) = \begin{cases} \prod_{i=1}^{n} \frac{1}{\binom{m}{r_i}} & \text{if } \mathbf{r}(\mathbf{G}) = \mathbf{r} \\ 0 & \text{else,} \end{cases} \tag{2.56}$$

and the related maximum log-likelihood is the sum of the individual ones, i.e.,

$$\ln \mathcal{L}_{\mathrm{mic}}(\mathbf{G}^*) = -\sum_{i=1}^{n} \ln \binom{m}{r_i^*}. \tag{2.57}$$

Analogously, the microcanonical complexity reads

$$\mathrm{COMP}_{\mathrm{mic}} = n \ln[m+1], \tag{2.58}$$

inducing the expression

$$\mathrm{DL}_{\mathrm{mic}}^{\mathrm{NML}}(\mathbf{G}^*) = \sum_{i=1}^{n} \ln \binom{m}{r_i^*} + n \ln[m+1]. \tag{2.59}$$

Similarly, the canonical probability is the product of row-specific canonical probabilities

$$P_{\text{can}}(\mathbf{G};\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{e^{-\theta_i r_i(\mathbf{G})}}{(1+e^{-\theta_i})^m}, \tag{2.60}$$

and can be rewritten as

$$P_{\text{can}}(\mathbf{G};\mathbf{p}) = \prod_{i=1}^{n} p_i^{r_i(\mathbf{G})}(1-p_i)^{m-r_i(\mathbf{G})} \tag{2.61}$$

upon considering the mean-value parametrization

$$p_i \equiv \frac{e^{-\theta_i}}{(1+e^{-\theta_i})^m} \quad i = 1 \dots n. \tag{2.62}$$

Since for each row the ML estimator of the parameter p_i is given by the row density $\hat{p}_i(\mathbf{G}) = \frac{r_i(\mathbf{G})}{m}$, the maximum log-likelihood reads

$$\ln \mathcal{L}_{\text{can}}(\mathbf{r}) = -m \sum_{i=1}^{n} h\left(\frac{r_i}{m}\right) \tag{2.63}$$

and the complexity reads

$$\text{COMP}_{\text{can}} = n \ln\left[\frac{e^m \Gamma(m,m)}{m^{m-1}} + 1\right]. \tag{2.64}$$

In conclusion, the canonical description length reads

$$\text{DL}_{\text{can}}(\mathbf{G}^*) = m \sum_{i=1}^{n} h\left(\frac{r_i}{m}\right) + n \ln\left[\frac{e^m \Gamma(m,m)}{m^{m-1}} + 1\right]. \tag{2.65}$$

In the hypothesis of m growing linearly with n, i.e., $m = \Theta(n)$, we consider the following asymptotic expressions (for a derivation, see section 2.B.2 of the Appendix)

$$\begin{aligned} \Delta \ln \mathcal{L}(\mathbf{G}^*) = &-\frac{n}{2}\ln[m] - \frac{1}{2}\sum_{i=1}^{n} \ln[2\pi p_i^*(1-p_i^*)] \\ &+ \frac{n}{12m} - \frac{1}{12m}\sum_{i=1}^{n} \frac{1}{p_i^*(1-p_i^*)} + o(1), \end{aligned} \tag{2.66}$$

$$\begin{aligned} \Delta\text{COMP} = &-\frac{n}{2}\ln[m] + \frac{n}{2}\ln\left[\frac{\pi}{2}\right] \\ &+ a \cdot \frac{n}{\sqrt{m}} + b \cdot \frac{n}{m} + o(1) \end{aligned} \tag{2.67}$$

with

$$\begin{aligned} a &= \frac{2}{3}\sqrt{\frac{2}{\pi}} \simeq 0.53, \\ b &= -\frac{11}{12} - \frac{4}{9\pi} \simeq -1.06. \end{aligned}$$

By combining these expressions, we obtain the asymptotic expression

$$\Delta \mathrm{DL}(\mathbf{G}^*) = \frac{1}{2}\sum_{i=1}^{n} \ln[\pi^2 p_i^*(1-p_i^*)] + a \cdot \frac{n}{\sqrt{m}} + c \cdot \frac{n}{m} + \frac{1}{12m}\sum_{i=1}^{n} \frac{1}{p_i^*(1-p_i^*)} + o(1)$$

where $c = -1 - \frac{4}{9\pi} \simeq -1.14$.

As for the case of one global constraint, we focus on two aspects of these results. First, the difference between log-likelihood functions is either $\Theta(n)$ or $\Theta(n \ln n)$ depending on the order of m, i.e., $m = \Theta(1)$ in the first case and $m = \Theta(n)$ in the second case. In any case, according to (2.36), the ensemble equivalence is broken. In particular, if $m = \Theta(n)$, the leading terms of expressions (2.66) and (2.67) are both of order $\Theta(n \ln n)$, thus canceling out when compared, precisely as in the equivalent case. Nonetheless, the remaining terms are still of order $\Theta(n)$. Hence, the (asymptotic) difference between description lengths increases linearly with the system size, regardless of the growth rate of m.

2.4.3 Towards model-level equivalence

At this point, it is important to emphasize that comparing canonical and microcanonical *ensembles* is not the same as comparing the corresponding *models*. The former involves comparing two probability distributions, each obtained for specific parameter values, whereas the latter involves comparing two sets of distributions (i.e., two models), with each set sharing a common functional form within itself. While in the context of the measure-level definition of ensemble equivalence, the comparison between ensembles can be reduced to comparing their likelihood terms, the comparison between models must take into account both their likelihoods and complexities.

In this section, we have shown two examples where the leading-order terms of the differences in likelihood and complexity are identical. This confirms that the complexities play as crucial a role in model comparison as the likelihoods. Furthermore, we validated relation (2.37) in the case of a single global constraint: when ensemble equivalence holds, the difference in description lengths between the canonical and microcanonical ensembles turns out to be $o(n)$. At the same time, our result that ΔDL grows with n in the case of n local constraints shows that the signature of broken ensemble equivalence remains visible when model complexities are included in the comparison.

These findings lead to the following question: can we generalize the definition of ensemble equivalence at the model level? To do so, we need a quantity that plays a role analogous to the relative entropy in comparing ensembles, but that applies to comparing models.

ΔDL might seem like a good candidate, but it is not suitable for this purpose: while it does take into account the model complexities, it still depends on the observed values. To be more general, we propose a definition involving the universal NML distributions representing the two models. Inspired by the measure level definition of

non-equivalence [7] of equation (2.35), we consider the following KL divergence

$$\begin{aligned} D_{\mathrm{KL}}^{\mathrm{NML}} &\equiv S(P_{\mathrm{mic}}^{\mathrm{NML}} || P_{\mathrm{can}}^{\mathrm{NML}}) \\ &= \sum_{\mathbf{G}\in\mathcal{G}} \bar{P}_{\mathrm{mic}}^{\mathrm{NML}}(\mathbf{G}) \ln \frac{\bar{P}_{\mathrm{mic}}^{\mathrm{NML}}(\mathbf{G})}{\bar{P}_{\mathrm{can}}^{\mathrm{NML}}(\mathbf{G})} \\ &= \sum_{\mathbf{G}\in\mathcal{G}} \bar{P}_{\mathrm{mic}}^{\mathrm{NML}}(\mathbf{G}) \Delta\mathrm{DL}(\mathbf{G}) \\ &= \sum_{\mathbf{c}\in\mathcal{C}} \Omega(\mathbf{c}) \frac{1}{\Omega(\mathbf{c})\mathcal{N}_{\mathcal{C}}} \Delta\mathrm{DL}(\mathbf{c}) \\ &= \overline{\Delta\mathrm{DL}(\mathbf{c})}, \end{aligned} \tag{2.68}$$

where $\overline{\Delta\mathrm{DL}(\mathbf{c})}$ represents the uniform average of $\Delta\mathrm{DL}(\mathbf{c})$ over all possible values of $\mathbf{c} \in \mathcal{C}$. Following the standard definition and assuming that the more relevant scale to consider is still the system size n, we say that the canonical and microcanonical models are equivalent *on the model level* if

$$\lim_{n\to\infty} \frac{D_{\mathrm{KL}}^{\mathrm{NML}}}{n} = \lim_{n\to\infty} \frac{\overline{\Delta\mathrm{DL}(\mathbf{c})}}{n} = 0. \tag{2.69}$$

Notice that, according to (2.37)

$$D_{\mathrm{KL}}(\mathbf{c}) = o(n) \quad \forall \mathbf{c} \in \mathcal{C} \quad \Rightarrow \quad D_{\mathrm{KL}}^{\mathrm{NML}} = o(n). \tag{2.70}$$

i.e., whenever ensemble equivalence holds, the corresponding microcanonical and canonical models are equivalent on the model level. The reverse implication, verified here in the specific case of one-sided local constraints, will be investigated in future work.

2.5 NML-based VS Bayesian approach to model selection

In the previous sections, we implicitly assumed that no prior knowledge about the parameters was available. Under this assumption, the NML-based universal distribution is the best choice in a precise minimax sense, as shown in section 2.2. Here, we consider the Bayesian approach to model selection [16, 17] by introducing the Bayesian universal distribution

$$\bar{P}_{\mathcal{M}}^{\mathrm{Bayes}}(\boldsymbol{x}) = \int_{\Theta} P_{\mathcal{M}}(\boldsymbol{x};\boldsymbol{\theta}) w(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{2.71}$$

(see [57, 58] for applications in the context of network models). The main difference with respect to the NML description length lies in the presence of a prior on the parameters $w(\boldsymbol{\theta})$ to be supplied by the user. This implementation, which is formally equivalent to the *Bayes factor method* [59], provides a description length for $\boldsymbol{x}$ through model $\mathcal{M}$ that is defined as

$$\mathrm{DL}_{\mathcal{M}}^{\mathrm{Bayes}}(\boldsymbol{x}) \equiv -\ln \bar{P}_{\mathcal{M}}^{\mathrm{Bayes}}(\boldsymbol{x}). \tag{2.72}$$

In some cases, the NML-based approach can be retrieved in a Bayesian context once the Bayesian distribution is equipped with a prior such that the two description lengths are, at least asymptotically, the same. We will refer to these priors as *NML-optimal priors*. In the microcanonical case, the uniform prior is an NML-optimal prior. Indeed, when the parameter space is a discrete set, the integral in (2.72) is replaced by a sum, and the microcanonical Bayesian description length reads

$$\begin{aligned} \mathrm{DL}_{\mathrm{mic}}^{\mathrm{Bayes}}(\mathbf{G}^*) &= -\ln \mathcal{L}_{\mathrm{mic}}(\mathbf{G}^*) - \ln w(\mathbf{c}^*) \\ &= \ln \Omega(\mathbf{c}^*) - \ln w(\mathbf{c}^*), \end{aligned} \tag{2.73}$$

which coincides with the NML-based description length provided by equation (2.19), upon identifying $w(\mathbf{c})$ with a uniform prior over $\mathcal{C}$:

$$w^{\mathrm{Uniform}}(\mathbf{c}) = \frac{1}{\mathcal{N}_{\mathcal{C}}} \quad \forall\, \mathbf{c} \in \mathcal{C}. \tag{2.74}$$

In the canonical case, the situation is quite different. First, consider V i.i.d. outcomes $\boldsymbol{x} = (x_1, \dots x_V)$ from a general k-dimensional exponential model. When $\boldsymbol{x}$ is such that the ML estimators of $\boldsymbol{\theta}$ are bounded away from the boundaries of Θ as the sample size goes to infinity, and under some regularity conditions on $\mathcal{M}$ which hold in our setting, the following asymptotic results hold true:

$$\mathrm{DL}_{\mathcal{M}}^{\mathrm{NML}}(\boldsymbol{x}) = -\ln \mathcal{L}_{\mathcal{M}}(\boldsymbol{x}) + \frac{k}{2} \ln \frac{V}{2\pi} + \ln \int_{\Theta} \sqrt{\det \mathbf{I}(\boldsymbol{\theta})} d\boldsymbol{\theta} + o(1), \tag{2.75}$$

$$\mathrm{DL}_{\mathcal{M}}^{\mathrm{Bayes}}(\boldsymbol{x}) = -\ln \mathcal{L}_{\mathcal{M}}(\boldsymbol{x}) + \frac{k}{2} \ln \frac{V}{2\pi} - \ln w(\hat{\boldsymbol{\theta}}(\boldsymbol{x})) + \ln \sqrt{\det \mathbf{I}(\hat{\boldsymbol{\theta}}(\boldsymbol{x}))} + o(1), \tag{2.76}$$

where $\det \mathbf{I}(\theta)$ is the determinant of the $k \times k$ *normalized Fisher information matrix*.

If we equip the Bayesian universal distribution with the Jeffreys prior [60]

$$J(\boldsymbol{\theta}) = \frac{\sqrt{\det \mathbf{I}(\boldsymbol{\theta})}}{\int_{\Theta} \sqrt{\det \mathbf{I}(\boldsymbol{\theta})} d\boldsymbol{\theta}} \tag{2.77}$$

by posing $w(\boldsymbol{\theta}) \equiv J(\boldsymbol{\theta})$, the two asymptotic expressions coincide (for a formal derivation of these results, we redirect the interested reader to [53] and [16]). Thus, whenever the asymptotic formulas (2.75) and (2.76) hold true, the NML-based description length is asymptotically equivalent to the one provided by the Bayesian distribution equipped with Jeffreys prior (hereby, Bayes-Jeffreys description), which is NML-optimal.

This is verified in the case of one global constraint, where the resulting canonical model turns out to be an exponential i.i.d. model. The asymptotic formulas do not provide the rate of convergence between the two description lengths, which can be derived by directly comparing the exact expressions of the NML description length $\mathrm{DL}_{\mathrm{can}}^{\mathrm{NML}}$ (2.65) and the Bayes-Jeffreys description length

$$\mathrm{DL}_{\mathrm{can}}^{\text{Bayes-Jeffreys}}(\mathbf{G}^*) = \ln[\pi(nm)!] - \ln\left[\Gamma(nm - l^* + 1/2)\Gamma(l^* + 1/2)\right] \tag{2.78}$$

(see section 2.C.1 of the Appendix for a derivation). In the limit $n \to \infty$ one finds

$$\mathrm{DL}_{\mathrm{can}}^{\mathrm{NML}}(\mathbf{G}^*) - \mathrm{DL}_{\mathrm{can}}^{\text{Bayes-Jeffreys}}(\mathbf{G}^*) = \frac{2}{3}\sqrt{\frac{2}{\pi nm}} + \Theta\left(\frac{1}{nm}\right). \tag{2.79}$$

The same comparison can be carried out in the case of n local constraints by applying the results above to each row. In this case, the Bayes-Jeffreys description length reads

$$\mathrm{DL}_{\mathrm{can}}^{\text{Bayes-Jeffreys}}(\mathbf{G}^*) = n\ln[\pi m!] - \sum_{i=1}^{n} \ln\left[\Gamma(m - r_i^* + 1/2)\Gamma(r_i^* + 1/2)\right] \tag{2.80}$$

and in the limit $n \to \infty$ one finds

$$\mathrm{DL}_{\mathrm{can}}^{\mathrm{NML}}(\mathbf{G}^*) - \mathrm{DL}_{\mathrm{can}}^{\text{Bayes-Jeffreys}}(\mathbf{G}^*) = \frac{2}{3}\frac{\sqrt{2}n}{\sqrt{\pi m}} + \Theta\left(\frac{n}{m}\right). \tag{2.81}$$

As $n \to \infty$, the last difference diverges both in case m stays finite, i.e., $m = \Theta(1)$, and in case m diverges, upon assuming that the rate at which m diverges does not exceed that of n. Consequently, in this case, it is no longer true that the NML approach can be asymptotically retrieved from the Bayesian one equipped with Jeffreys prior. Although the existence of a modified Jeffreys prior extending the equivalence of the two approaches to the case of n local constraints can indeed be hypothesized, for the time being, we merely suggest not to assume this to be true when ensemble non-equivalence holds.

In the remainder of this section, we show that the Bayesian description length becomes more sensitive to the choice of prior whenever local constraints are enforced. Let us start by considering the relationship

$$\mathrm{DL}_{\mathrm{can}}^{\mathrm{NML}} = \mathrm{DL}_{\mathrm{mic}}^{\text{Bayes-Rissanen}} \tag{2.82}$$

stating that the NML-based description length of a canonical model coincides with the Bayesian description length of its microcanonical variant, equipped with the so-called *canonical prior*, introduced by Rissanen in [61] (see the definition in section 2.C.2 of the Appendix). Moreover,

$$\mathrm{DL}_{\mathrm{mic}}^{\mathrm{NML}} = \mathrm{DL}_{\mathrm{can}}^{\text{Bayes-Uniform}} \tag{2.83}$$

i.e., the NML-based description length of the microcanonical model coincides with the Bayesian description length of its canonical variant equipped with a uniform prior on the mean-value parameters.

The two identities, derived in 2.C.2, show that there exists a prior guaranteeing that the Bayesian description length of the canonical (microcanonical) model coincides with the NML-based description length of the microcanonical (canonical) model. This might lead to the conclusion that, in the framework of model selection, canonical and microcanonical models coincide even when non-equivalence holds if the right priors are chosen. Nevertheless, these priors are not NML-optimal. While this non-optimality is negligible in the case of one global constraint, it could play a crucial role when

local constraints are enforced. In other words, in this case, different choices of prior can result in very different Bayesian description lengths for microcanonical as for canonical models.

The n local constraints case examined in the previous sections provides an instructive example. In this case, we showed that the difference between the canonical and microcanonical NML description lengths is $\Theta(n)$. Combining this result with (2.82), we can conclude that when it comes to the microcanonical model, choosing between the uniform prior (which is NML-optimal) and the canonical prior can amount to a difference in description length of $\Theta(n)$ nats. Similarly, when it comes to the canonical model, it can be easily shown that, based on (2.81, 2.83) when $m = \Theta(n)$, choosing between the uniform and the Jeffreys prior can lead to a difference in description length of $\Theta(\sqrt{n})$ nats. In contrast, the same differences are $\Theta(1)$ when one global constraint is involved.

This simple example shows that when non-equivalence holds, the selection of different priors can result in increasingly larger differences in DL when the system size grows. In other words, fine-tuning the prior can result in a significant gain - or, conversely, a loss - in terms of compression, particularly for large-scale systems.

2.6 Conclusion

This study explored the implications of ensemble non-equivalence in the context of MDL-based model selection. We specifically compared the NML-based description lengths of canonical and microcanonical maximum-entropy models. The NML approach enables a comparison of model description lengths derived from the same underlying principle without requiring prior assumptions about the parameters.

We found that, although microcanonical models fit the data better than their canonical counterparts, they are always more complex. Determining the best-scoring model - i.e., the one that minimizes the description length - is not straightforward, as it depends on the observed values of the sufficient statistics. Specifically, we found that the canonical model performs best when its fit to the observed data exceeds its average fit across all possible values of the sufficient statistics.

Additionally, we explicitly compared the description lengths of canonical and microcanonical models for dense $n \times m$ binary matrices in two cases of interest: one with a global constraint, such as the sum of the matrix entries, and another with n local constraints, such as the sum of the entries in each row. In the former case, the model is homogeneous, where the probability of an entry being 1 is the same for all entries, while in the latter, the model becomes heterogeneous, with the probability of an entry being 1 depending on the specific row. In this second scenario, the canonical and microcanonical ensembles are no longer equivalent.

By comparing the description lengths of the canonical and microcanonical formulations of these models, we found that the difference remains finite when a global constraint is enforced but grows linearly with system size when local constraints are imposed. Our findings highlight the impact of choosing hard or soft constraints on data compression, leading us to conclude that canonical and microcanonical models should be treated as distinct from a model selection perspective in the presence

of ensemble non-equivalence. Furthermore, they suggest a new definition of model equivalence based on MDL, which requires further investigation.

In addition, we investigated the relationship between NML-based and Bayesian description lengths. We defined NML-optimal priors as those that, when applied in a Bayesian framework, result in description lengths that asymptotically match the NML-based approach. While the Jeffreys prior is well known to be NML-optimal for canonical models under a single global constraint, we show that this optimality does not hold when local constraints are imposed. Indeed, in such cases, the difference between the corresponding description lengths increases with system size. This observation serves as a caution to practitioners, suggesting that the two approaches (NML and Jeffreys priors) should not be treated as interchangeable in the presence of ensemble non-equivalence. It also raises the question of whether an NML-optimal prior exists when local constraints are enforced.

Lastly, we showed that different priors can result in largely different description lengths when local constraints are involved. This finding underscores the importance for practitioners of carefully selecting priors, since the choice of prior could greatly impact the accuracy and effectiveness of model selection.

Appendix 2.A Derivation of the microcanonical distribution from entropy maximization

It is well known that canonical maximum-entropy models can be derived by maximizing the Shannon entropy subject to soft constraints [1]. Here, we show that microcanonical models can also be seen as the solution to an entropy maximization problem under hard constraints, as a direct consequence of the Shannon-Khinchin axioms that uniquely identify Shannon entropy [62, 3].

Let $\mathcal{G}$ denote the set of all configurations, and let $\mathcal{G}_{\text{valid}} \subset \mathcal{G}$ denote the subset of configurations that exactly satisfy the hard constraints. Consider the problem of maximizing the Shannon entropy

$$S[P] = -\sum_{\mathbf{G}\in\mathcal{G}} P(\mathbf{G}) \log P(\mathbf{G}) \tag{2.84}$$

over probability distributions $P(\mathbf{G})$ supported on $\mathcal{G}$, under the requirement that $P(\mathbf{G}) = 0$ for all $\mathbf{G} \notin \mathcal{G}_{\text{valid}}$. That is, configurations violating the constraints are assigned zero probability.

The third Shannon-Khinchin axiom states that $S[P]$ is expansible, i.e., it does not change if an outcome with zero probability is added. This implies that, in this case, maximizing entropy over the full space $\mathcal{G}$ with support restricted to $\mathcal{G}_{\text{valid}}$ is equivalent to maximizing entropy over $\mathcal{G}_{\text{valid}}$ only. Then, according to the second Shannon-Khinchin axiom, which states that entropy is maximized by the uniform distribution in the absence of further constraints, the solution is the uniform distribution over $\mathcal{G}_{\text{valid}}$:

$$P_{\text{mic}}(\boldsymbol{x}) = \begin{cases} \frac{1}{|\mathcal{G}_{\text{valid}}|} & \text{if } \mathbf{G} \in \mathcal{G}_{\text{valid}}, \\ 0 & \text{else.} \end{cases} \tag{2.85}$$

With this derivation, canonical and microcanonical ensembles emerge as distinct solutions to the same entropy maximization principle, differing only in how the constraints are imposed.

Appendix 2.B Asymptotic expansions

Here we derive the asymptotic expansions used to obtain the asymptotic results in section 2.4.

2.B.1 One global constraint

We start by deriving an asymptotic expansion for the microcanonical log-likelihood:

$$\begin{aligned}\ln \mathcal{L}_{\text{mic}}(l) &= -\ln \binom{nm}{l} \\ &= -\ln(nm)! + \ln(nm-l)! + \ln l!. \end{aligned} \tag{2.86}$$

In the dense regime, l is of order $\Theta(n)$ and Stirling's approximation

$$\ln x! = x \ln x - x + \frac{1}{2}\ln(2\pi x) + \Theta(1/x) \tag{2.87}$$

can be used to evaluate all the factorials above, yielding

$$\ln \mathcal{L}_{\text{mic}}(l) = -nm \ln(nm) + (nm-l)\ln(nm-l) + l \ln l - \frac{1}{2}\ln\left(\frac{nm}{2\pi l(nm-l)}\right) + o(1), \tag{2.88}$$

which can be further expressed as a function of the density $p = l/nm$:

$$\ln \mathcal{L}_{\text{mic}}(p) = -nm \cdot h(p) + \frac{1}{2}\ln(nm) + \frac{1}{2}\ln(2\pi p(1-p)) + o(1). \tag{2.89}$$

By combining this expression with the canonical log-likelihood

$$\ln \mathcal{L}_{\text{can}}(p) = -nm \cdot h(p), \tag{2.90}$$

we obtain the asymptotic log-likelihood difference of equation (2.53).

Similarly, we combine the asymptotic expansion of the microcanonical complexity

$$\begin{aligned}\text{COMP}_{\text{mic}} &= \ln(nm+1) = \ln(nm) + \ln\left(1 + \frac{1}{nm}\right) \\ &= \ln(nm) + o(1), \end{aligned} \tag{2.91}$$

which follows from $\ln\left(1+\frac{1}{x}\right) = \Theta(1/x)$, and the following asymptotic expansion of the canonical complexity

$$\text{COMP}_{\text{can}} = \frac{1}{2}\ln(nm) + \frac{1}{2}\ln\left(\frac{\pi}{2}\right) + o(1). \tag{2.92}$$

The latter is easily derived by asymptotically expanding the following approximation

$$\mathrm{COMP}_{\mathrm{can}}^{\mathrm{NML}} \simeq \ln\left[\sqrt{\frac{\pi nm}{2}} + \frac{2}{3} + \frac{\sqrt{2\pi}}{24\sqrt{nm}} - \frac{4}{135nm} + \frac{\sqrt{2\pi}}{576\sqrt{(nm)^3}} + \frac{8}{2835(nm)^2}\right], \tag{2.93}$$

provided in [56], where the authors show it to be very precise already for small values of the total number of entries nm. Finally, equation (2.54) is obtained by comparing the two asymptotic complexities. Notice that the asymptotic formula (2.92) for the canonical complexity represents a well-known result, as it can be derived directly by applying the asymptotic formula (2.75) to the case of i.i.d. Bernoulli random variables.

2.B.2 One-sided local constraints

Here, we consider the regime in which $m = \Theta(n)$. Similarly to the previous case, we need to expand the microcanonical likelihood

$$\begin{aligned}\ln \mathcal{L}_{\mathrm{mic}}(\mathbf{r}) &= -\sum_{i=1}^{n} \ln \binom{m}{r_i} \\ &= -n\ln m! + \sum_{i=1}^{n} \ln(m-r_i)! + \ln r_i!. \end{aligned} \tag{2.94}$$

In the dense regime, all r_i's are of order $\Theta(n)$, and we could again apply Stirling's formula to the factorials above. Nevertheless, because of the factor n ahead of everything, Stirling's formula is not enough, and we turn to Stirling's series [63]

$$\ln x! = x\ln x - x + \frac{1}{2}\ln(2\pi x) + \frac{1}{12x} + \Theta(1/x^3). \tag{2.95}$$

The resulting asymptotic expression can be expressed as a function of the row densities $p_i = r_i/m$:

$$\begin{aligned}\ln \mathcal{L}_{\mathrm{mic}}(\mathbf{r}) = -m\sum_{i=1}^{n} h(p_i) + \frac{n}{2}\ln m + \frac{1}{2}\sum_{i=1}^{n}[\ln(2\pi p_i(1-p_i))] \\ - \frac{n}{12m} + \frac{1}{12m}\sum_{i=1}^{n}\left[\frac{1}{p_i(1-p_i)}\right] + o(1). \end{aligned} \tag{2.96}$$

By combining this expression with the canonical log-likelihood

$$\ln \mathcal{L}_{\mathrm{can}}(\mathbf{p}) = -m\sum_{i=1}^{n} h(p_i), \tag{2.97}$$

one obtains the asymptotic log-likelihood difference of equation (2.66).

Similarly, the asymptotic expansion of the microcanonical complexity

$$\begin{aligned}\mathrm{COMP}_{\mathrm{mic}} &= n\ln(m+1) = \ln m + n\ln\left(1+\frac{1}{m}\right) \\ &= n\ln m + \frac{n}{m} + o(1), \end{aligned} \tag{2.98}$$

which follows from $\ln\left(1+\frac{1}{x}\right)=\frac{1}{x}+\Theta(1/x^2)$, is compared to the following asymptotic expansion of the canonical complexity, based on (2.93)

$$\mathrm{COMP}_{\mathrm{can}}=\frac{n}{2}\ln m+\frac{n}{2}\ln\frac{\pi}{2}+\frac{2}{3}\sqrt{\frac{2}{\pi}}\cdot\frac{n}{\sqrt{m}}+\left(\frac{1}{12}-\frac{4}{9\pi}\right)\cdot\frac{n}{m}+o(1) \tag{2.99}$$

to express (2.64). The asymptotic expression (2.67) is obtained as the difference between the two asymptotic complexities.

Appendix 2.C NML and Bayes

2.C.1 Bayesian-Jeffreys description lengths

In what follows, we compute the Bayesian description length $\mathrm{DL}_{\mathrm{can}}^{\mathrm{B,\ Jeffreys}}$ of the canonical model obtained by constraining the sum l over the matrix. As already stated, this model is equivalent to modeling nm i.i.d. Bernoulli variables, and this result can be found in Example 8.3 of [16]. In the paper, we extend this result to the case of n one-sided local constraints.

First, we compute Jeffreys prior. For a Bernoulli variable, the Fisher information for the parameter p is $\mathbf{I}(p)=p^{-1}(1-p)^{-1}$ and $\int_0^1\sqrt{\mathbf{I}(p)}=\pi$. Thus, the Jeffreys prior reads

$$J(p)=\frac{1}{\pi\sqrt{p(1-p)}} \tag{2.100}$$

and the Bayesian-Jeffreys DL is

$$\begin{aligned}\mathrm{DL}_{\mathrm{can}}^{\text{Bayes-Jeffreys}}(\mathbf{G}^*)&=-\ln\int_0^1 P_{\mathrm{can}}(\mathbf{G}^*;p)J(p)dp\\&=-\ln\frac{1}{\pi}\int_0^1 p^{l^*-\frac{1}{2}}(1-p)^{nm-l^*-\frac{1}{2}}dp.\end{aligned} \tag{2.101}$$

The integral above is the Beta function

$$B(x,y)=\int_0^1 t^{x-1}(1-t)^{y-1}dt=\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \tag{2.102}$$

computed for $x=l^*+\frac{1}{2}$ and $y=nm-l^*+\frac{1}{2}$, where $\Gamma(x)$ is the Gamma function

$$\Gamma(x)=\int_0^{+\infty}t^{x-1}e^{-t}dt. \tag{2.103}$$

Thus

$$\mathrm{DL}_{\mathrm{can}}^{\text{Bayes-Jeffreys}}(\mathbf{G}^*)=-\ln\frac{\Gamma\left(l^*+\frac{1}{2}\right)\Gamma\left(nm-l^*+\frac{1}{2}\right)}{\pi(nm)!}, \tag{2.104}$$

which corresponds to Equation (2.78). By employing the Stirling series, this expression can be expanded asymptotically as

$$\mathrm{DL}_{\mathrm{can}}^{\text{Bayes-Jeffreys}}(\mathbf{G}^*)=nm\cdot h\left(\frac{l^*}{nm}\right)+\frac{1}{2}\ln\frac{\pi nm}{2}+\Theta\left(1/nm\right). \tag{2.105}$$

The last expression is compared with the asymptotic expansion of the canonical NML description length $\mathrm{DL}_{\mathrm{can}}^{\mathrm{NML}}$, obtained as the difference between the asymptotic expansion of (2.93) and the canonical log-likelihood:

$$\mathrm{DL}_{\mathrm{can}}^{\mathrm{NML}}(\mathbf{G}^*) = nm \cdot h\left(\frac{l^*}{nm}\right) + \frac{1}{2}\ln\frac{\pi nm}{2} + \frac{2}{3}\sqrt{\frac{2}{\pi nm}} + o(1). \tag{2.106}$$

Finally, Equation (2.79) is obtained as the difference between (2.106) and (2.105)

2.C.2 Proving identities

In what follows, we prove identities (2.82) and (2.83) of section 2.5.

Proving Identity (2.82). We begin with identity (2.82), namely:

$$\mathrm{DL}_{\mathrm{can}}^{\mathrm{NML}} = \mathrm{DL}_{\mathrm{mic}}^{\text{Bayes-Rissanen}}, \tag{2.107}$$

where the canonical prior $\hat{W}$ inducing the Bayes-Rissanen description length is expressed as

$$\begin{aligned}\hat{W}(\mathbf{c}) &= \frac{\sum_{\mathbf{G}:\mathbf{c}(\mathbf{G})=\mathbf{c}} P_{\mathrm{can}}(\mathbf{G};\hat{\boldsymbol{\theta}}(\mathbf{G})}{\sum_{\mathbf{G}} P_{\mathrm{can}}(\mathbf{G};\hat{\boldsymbol{\theta}}(\mathbf{G}))} \\ &= \frac{\Omega(\mathbf{c})\mathcal{L}_{\mathrm{can}}(\mathbf{c})}{\sum_{\mathbf{G}} \mathcal{L}_{\mathrm{can}}(\mathbf{G})}.\end{aligned} \tag{2.108}$$

We recognize the canonical complexity in the denominator of $\hat{W}$. Thus, putting this expression in the Bayesian microcanonical description length, we get

$$\begin{aligned}\mathrm{DL}_{\mathrm{mic}}^{\text{Bayes-Rissanen}}(\mathbf{G}^*) &= \ln\Omega(\mathbf{c}^*) - \ln\hat{W}(\mathbf{c}^*) \\ &= -\ln\mathcal{L}_{\mathrm{can}}(\mathbf{c}) + \mathrm{COMP}_{\mathrm{can}} \\ &= \mathrm{DL}_{\mathrm{can}}^{\mathrm{NML}}(\mathbf{G}^*),\end{aligned} \tag{2.109}$$

which proves the identity.

Proving Identity (2.83). We will now prove identity (2.83), namely:

$$\mathrm{DL}_{\mathrm{mic}}^{\mathrm{NML}} = \mathrm{DL}_{\mathrm{can}}^{\text{Bayes-Uniform}}, \tag{2.110}$$

for the specific maximum-entropy models considered in this work, starting from the case of one global constraint. The uniform prior on the mean-value parameter p reads

$$w^{\mathrm{Uniform}}(p) = 1 \quad \text{for } p \in [0,1]. \tag{2.111}$$

Putting this prior in the Bayesian description length yields:

$$\begin{aligned}
\mathrm{DL}_{\mathrm{can}}^{\text{Bayes-Uniform}}(\mathbf{G}^*) &= -\ln \int_0^1 P_{\mathrm{can}}(\mathbf{G}^*; p) dp \\
&= -\ln \int_0^1 p^{l^*} (1-p)^{nm-l^*} dp \\
&= \ln \binom{nm}{l^*} + \ln(nm+1) \\
&= \mathrm{DL}_{\mathrm{mic}}^{\mathrm{NML}}(\mathbf{G}^*),
\end{aligned} \tag{2.112}$$

where the integral is computed by integrating by parts l^* times. This proves the identity for the case of one global constraint.

The identity holds as well for the case of n one-sided local constraints, with the uniform prior reading

$$w^{\mathrm{Uniform}}(\mathbf{p}) = 1 \quad \text{for } \mathbf{p} \in [0,1]^n. \tag{2.113}$$

Indeed, we have that

$$\begin{aligned}
\mathrm{DL}_{\mathrm{can}}^{\text{Bayes-Uniform}}(\mathbf{G}^*) &= -\ln \int_0^1 P_{\mathrm{can}}(\mathbf{G}^*; p) dp \\
&= -\sum_{i=1}^{n} \ln \int_0^1 p_i^{r_i^*} (1-p)^{m-r_i^*} dp_i \\
&= \sum_{i=1}^{n} \ln \binom{m}{r_i^*} + n \ln(m+1) \\
&= \mathrm{DL}_{\mathrm{mic}}^{\mathrm{NML}}(\mathbf{G}^*),
\end{aligned} \tag{2.114}$$

which proves the identity for the case of n one-sided local constraints.