Testing maximum entropy models with e -values

Testing maximum entropy models with e-values

Francesca Giuffrida,[1, 2, 3, *] Diego Garlaschelli,[1, 2] and Peter Grünwald[4, 5]

[1] *IMT School for Advanced Studies, Lucca (Italy)*
[2] *Lorentz Institute for Theoretical Physics (LION), Leiden University, Leiden (The Netherlands)*
[3] *Dipartimento di Fisica e Chimica, University of Palermo, Palermo (Italy)*
[4] *Centrum Wiskunde & Informatica, Amsterdam (The Netherlands)*
[5] *Mathematical Institute, Leiden University, Leiden (The Netherlands)*

E-values have recently emerged as a robust and flexible alternative to p-values for hypothesis testing, especially under optional continuation, i.e., when additional data from further experiments are collected. In this work, we define optimal e-values for testing between maximum entropy models, both in the microcanonical (hard constraints) and canonical (soft constraints) settings. We show that, when testing between two hypotheses that are both microcanonical, the so-called growth-rate optimal e-variable admits an exact analytical expression, which also serves as a valid e-variable in the canonical case. For canonical tests, where exact solutions are typically unavailable, we introduce a microcanonical approximation and verify its excellent performance via both theoretical arguments and numerical simulations. We then consider constrained binary models, focusing on $2 \times k$ contingency tables — an essential framework in statistics and a natural representation for various models of complex systems. Our microcanonical optimal e-variable performs well in both settings, constituting a new tool that remains effective even in the challenging case when the number k of groups grows with the sample size, as in models with growing features used for the analysis of real-world heterogeneous networks and time-series.

I. INTRODUCTION

In recent years, scientific interest in complex data modeling has surged, due to the increasing availability of both global-scale structured data and computational power. At the same time, rising concerns about the misuse of p-values and significance testing [1–3] underscore the need for reliable statistical methods to extract knowledge from data. As a robust and flexible alternative to p-values for hypothesis testing, *e-values* [4, 5] have recently gained considerable attention. Having been independently (re)discovered several times in different contexts (including by physicists [6] — see [4] for early history) over the past decades, interest suddenly exploded in 2019 when the first versions of several breakthrough papers [7–10] appeared on arXiv.

An *e-variable* is simply a nonnegative random variable whose expected value under the null hypothesis is at most one. The value it takes on the given sample is called the e-value. This simple definition yields several desirable properties: e-values provide rigorous control of the Type I error, retain it under optional continuation (i.e., when data from additional experiments become available), and can be interpreted as a measure of evidence against the null hypothesis. However, not all e-variables are equally useful as test statistics. To address this, a notion of *optimality* is introduced. An *optimal* e-variable is one that grows quickly under the alternative hypothesis, accumulating strong evidence against the null when the latter is false. In this paper, we focus specifically on *growth-rate optimal* (GRO) e-variables [7]. The results in [7], later extended in [11, 12], provide a general theoretical framework for constructing GRO e-variables in broad testing scenarios.

The aim of this work is to develop optimal e-variables for hypothesis testing between maximum entropy models (MEMs). These models are derived by considering each possible realization $\mathbf{x} \in \mathcal{X}$ of the data (where $\mathcal{X}$ is the set of allowed realizations) and looking for the probability distribution $P(\mathbf{x})$ that maximizes Shannon entropy

$$\mathcal{S}[P] = -\sum_{\mathbf{x}\in\mathcal{X}} P(\mathbf{x}) \log P(\mathbf{x}) \tag{1}$$

under a set of constraints, typically defined through a vector of observables $\mathbf{c}(\mathbf{x})$ over the data. This approach, due to Gibbs [13] and Jaynes [14], outputs ensembles of data reproducing the constrained quantities and randomizing everything else maximally.

Two main formulations of MEMs exist, depending on how the constraints are enforced. If the constraints are imposed as exact values, i.e., $\mathbf{c}(\mathbf{x}) = \mathbf{c}^*$ on each realizable $\mathbf{x}$, one obtains a *microcanonical model*, where only configurations satisfying the constraints are assigned nonzero probability. If, instead, the constraints are satisfied only on average, i.e., $\mathbb{E}_P[\mathbf{c}(\mathbf{x})] = \mathbf{c}^*$, one obtains a *canonical model*, where fluctuations are allowed and the probability distribution has exponential form. In statistical terminology, by varying $\mathbf{c}^*$ one obtains an *exponential family with discrete outcome space and uniform carrier* [15]. In both cases, the probability of $\mathbf{x}$ is entirely determined by the value of $\mathbf{c}(\mathbf{x})$, which plays the role of sufficient statistic.

MEMs are commonly used to model complex systems that give rise to structured data, e.g., in network science [16–18] and time-series analysis [19, 20], where they capture structural properties such as (heterogeneous) node degrees in networks or empirical trends in (non-

* francesca.giuffrida@imtlucca.it

stationary) temporal data, respectively. However, while statistical tests for exponential family models are well established, testing procedures specifically tailored to maximum entropy models remain far less developed, especially when applied in the microcanonical setting. In fact, e-variables for testing between general MEMs have so far not been developed at all: the only related works we are aware of are [21, 22] and [23, 24]. The former concentrates on the very specific sub-case of 2×2 tables (to re-appear as Example A–C in our paper later on), but uses e-variables which are designed for purely sequential purposes, and are therefore not optimal in the sense we define below, neither in the canonical nor in the microcanonical setting. The latter works, [23, 24], studied e-variables for testing between two exponential families with the same sufficient statistic but different carriers. By contrast, testing between MEMs amounts to testing exponential families with different sufficient statistics but the same (uniform) carrier. This is exactly the aim of this work.

This paper is organized as follows. In section II, we introduce e-variables and growth-rate optimality. In section III, we address the problem of finding optimal e-variables for testing between two maximum entropy models, either microcanonical (III A) or canonical (III B), that differ in their sufficient statistics. We introduce a method to construct optimal e-variables in both Bayesian and non-Bayesian settings. We show that the microcanonical GRO e-variable is also a valid canonical e-variable, and that in some cases, it asymptotically coincides with the optimal one. This is particularly relevant: while canonical models are way more commonly used in the literature, calculating the optimal canonical e-variable is usually analytically impossible and computationally infeasible. Here, we provide a method to explicitly compute the optimal microcanonical e-variable and to further verify how well it approximates the optimal canonical e-variable. In section IV, we explicitly apply these results to contingency tables, underlying connections with important problems in network science. We first analyze the case of 2×2 contingency tables (IV A) and then generalize to $2 \times k$ (IV B). We show that, in these cases and for both Bayesian and non-Bayesian examples, the GRO microcanonical e-variable is not only a valid canonical e-variable but also an excellent approximation of the optimal canonical one.

II. INTRODUCTION TO E-VARIABLES

Consider the typical hypothesis testing scenario, where the goal is to test a *null hypothesis* $\mathcal{M}_0$ against an *alternative hypothesis* $\mathcal{M}_1$. Both $\mathcal{M}_0$ and $\mathcal{M}_1$ are assumed to be parametric statistical models, i.e., families of distributions sharing the same functional form:

$$\mathcal{M}_j = \{P_j(\mathbf{x};\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_j}, \quad j \in \{0,1\}, \tag{2}$$

where $\boldsymbol{\theta}$ represents the vector of model parameters and $\boldsymbol{\Theta}_j$ the corresponding parameter space for model $\mathcal{M}_j$.

An *e-variable* E is a non-negative random variable that satisfies the following condition under all distributions in the null hypothesis:

$$\mathbb{E}_0[\,E\,] \leq 1 \quad \forall P_0 \in \mathcal{M}_0. \tag{3}$$

The realized value of E evaluated on data $\mathbf{x}$ is called an *e-value*. Unlike p-values, *larger* e-values indicate stronger evidence against the null. This follows directly from their defining property that, under the null, their expectation is bounded by one. This simple yet powerful definition has several important implications [4, 5]:

- **Type I error control**: The condition $\mathbb{E}_0[\,E\,] \leq 1$ ensures that a test based on e-values controls the Type I error, that is, the probability of rejecting the null hypothesis when it is actually true. Given a significance level $0 \leq \alpha \leq 1$, by Markov's inequality, we have

$$P_0(E \geq 1/\alpha) \leq \alpha, \tag{4}$$

  for all $P_0 \in \mathcal{M}_0$. This guarantees that the probability of wrongly rejecting the null hypothesis does not exceed the significance level α, regardless of the true parameter value within the null model.

- **Post-hoc error control:** E-values allow a variation of valid Type-I error control even when the significance level is chosen *after* observing the data [25]. Specifically, if e is the observed e-value, then rejecting the null hypothesis at level $1/e$ preserves a Type I risk bound despite this level being data-dependent. This contrasts with traditional p-values, which only guarantee valid inference when the significance level is fixed in advance.

- **Optional continuation:** E-values support valid testing under *optional continuation*, making them well-suited for sequential analyses and meta-analyses across independent studies. If $e_{(1)}, e_{(2)}, \ldots$ are e-values computed on independent data batches (e.g., studies), their product remains a valid e-value — even if the decision to analyze further batches, to perform tests on them, or to incorporate specific prior knowledge into the e-values is guided by the outcomes of earlier batches. In this way, Type I error control is preserved, enabling flexible and robust hypothesis testing across repeated or cumulative experimental settings.

While all random variables satisfying condition (3) qualify as e-variables, not all of them are informative. For instance, the constant random variable $E(\mathbf{x}) \equiv 1$ satisfies the definition but provides no information. To address this, a notion of *optimal* e-variables was introduced in [7]. In particular, the authors define the *Growth Rate Optimal* (GRO) e-variable as the unique solution to a specific

optimization problem based on a growth criterion, which we present below.

As a first step toward understanding GRO e-variables, we introduce the concept of *Bayesian evidence* (also known as the *Bayesian marginal likelihood*) of model j with prior density w_j, defined as:

$$P_j^{w_j}(\mathbf{x}) = \int_{\mathbf{\Theta}_j} P_j(\mathbf{x};\boldsymbol{\theta}) w_j(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{5}$$

This quantity reflects the overall support the data provide for model j, by averaging the likelihood over the prior; it is widely used in Bayesian model selection, where models with higher evidence are preferred. We shall mostly work with prior densities w_j defined on convex parameter spaces $\mathbf{\Theta}_j \subset \mathbb{R}^d$ $(d > 0)$, assuming they are continuous and strictly positive for all $\boldsymbol{\theta} \in \mathbf{\Theta}_j$. We refer to such priors as *regular priors*.

Given a fixed prior w_1 (regular or not) on the alternative hypothesis, the GRO e-variable S^{GRO} is the unique solution to the following optimization problem:

$$S^{\mathrm{GRO}} = \arg\max_{E \in \mathcal{E}_0} \mathbb{E}_{P_1^{w_1}}[\log E], \tag{6}$$

where $\mathcal{E}_0$ denotes the set of all e-variables relative to the null model $\mathcal{M}_0$, i.e., the set of all random variables satisfying (3).

This optimization can be interpreted as a growth criterion: while the expected value of any e-variable is bounded under the null, a well-designed e-variable should grow rapidly assuming the alternative is true, when the prior w_1 is correctly specified. The use of the logarithmic growth in this criterion is motivated and discussed in more detail in [7]. The quantity $\mathbb{E}_{P_1^{w_1}}[\log E]$, known as the *e-power* [26–28] of E, has become a standard measure for evaluating the performance of an e-variable.

In the most common case, a GRO e-variable solving the aforementioned optimization problem takes the form of a *Bayes factor* [29], i.e., the ratio between two Bayesian evidences:

$$S(\mathbf{x}) = \frac{P_1^{w_1}(\mathbf{x})}{P_0^{w_0}(\mathbf{x})}. \tag{7}$$

Equation (7) represents the Bayes factor comparing models $\mathcal{M}_1$ and $\mathcal{M}_0$. It measures the relative support that the data provide for one model over the other. However, not all Bayes factors qualify as e-variables; to ensure the e-variable property (3), while w_1 may be chosen freely, a specific prior w_0^*, depending on w_1, must then be chosen for the null hypothesis. Specifically, w_0^* is the solution to the following optimization problem:

$$w_0^* = \arg\min_{w \in \mathcal{W}_{\boldsymbol{\theta}_0}} D_{\mathrm{KL}}(P_1^{w_1} \| P_0^{w_0}) \tag{8}$$

where $\mathcal{W}_{\boldsymbol{\theta}_0}$ is the space of all priors on $\boldsymbol{\theta}_0$, and $D_{\mathrm{KL}}(P_1^{w_1} \| P_0^{w_0})$ denotes the *Kullback-Leibler divergence*:

$$\begin{aligned} D_{\mathrm{KL}}(P_1^{w_1} \| P_0^{w_0}) &= \sum_{\mathbf{x} \in \mathcal{X}} P_1^{w_1}(\mathbf{x}) \log \frac{P_1^{w_1}(\mathbf{x})}{P_0^{w_0}(\mathbf{x})} \\ &= \mathbb{E}_{P_1^{w_1}}[\log S(\mathbf{x})]. \end{aligned} \tag{9}$$

Theorem 1 in [7] proves that given w_1, among all Bayes factors of the form (7), the random variable

$$S^{\mathrm{GRO}}(\mathbf{x}) = \frac{P_1^{w_1}(\mathbf{x})}{P_0^{w_0^*}(\mathbf{x})}, \tag{10}$$

is the *only* e-variable, assuming that a w_0^* achieving the minimum in (8) exists [1].

To sum up, the GRO e-variable is the unique solution of two different optimization problems, defined on two different sets: for a given $P_1^{w_1}$ and null model $\mathcal{M}_0$, S^{GRO} is the only e-variable among Bayes factors of the form 7; at the same time it is the only e-variable maximizing the e-power (see Figure 1).

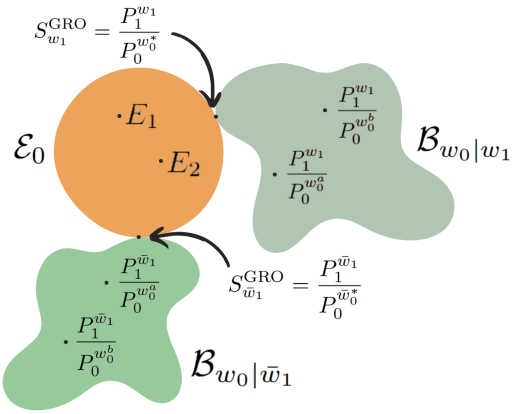

FIG. 1. The GRO e-variable S^{GRO} is the unique intersection between the set $\mathcal{B}_{w_0|w_1}$ of all Bayes factors for a given $P_1^{w_1}$ and varying $P_0^{w_0}$, and the set $\mathcal{E}_0$ of all e-variables relative to model $\mathcal{M}_0$. At the same time, it is the unique e-variable maximizing the e-power relative to $P_1^{w_1}$. This schematic representation considers two possible alternative priors, w_1 and $\bar{w}_1$.

The reader may have noticed a seeming asymmetry: while e-values are defined in a frequentist sense — requiring Type I error control for *all* $P_0 \in \mathcal{M}_0$ — our optimality criterion for GRO relies on a prior w_1 over the alternative model $\mathcal{M}_1$, and thus relies on a Bayesian formulation.

It would be conceptually appealing to define an optimality criterion that, like the e-variable condition, provides performance guarantees over *all* $P_1 \in \mathcal{M}_1$ rather than *on average according to a prior* w_1. As it turns out, this is indeed possible by drawing on ideas from information theory.

To move in that direction, we first note that the result of [7] is not limited to Bayes factors. It applies to more

[1] As shown in [7], multiple distinct minimizers w_0^* may exist, but they yield the same $P_0^{w_0^*}$. Even when a minimizer does not exist, $P_0^{w_0^*}$ can be defined as a limit along a minimizing sequence w_j, ensuring that (10) remains a valid e-variable.

general e-variables of the form

$$S(\mathbf{x}) = \frac{\bar{P}_1(\mathbf{x})}{P_0^{w_0^*}(\mathbf{x})}, \tag{11}$$

where $\bar{P}_1$ is any probability distribution over the data space. For such constructions to be useful, however, the choice of $\bar{P}_1$ must be guided by an appropriate extension of the GRO criterion.

This leads us to the concept of *regret*, also referred to as *relative growth* (*regrow*) in [7], which we adopt here using more common terminology. Regret quantifies the power loss incurred when using a candidate e-variable instead of the ideal one, which is designed for the true data-generating distribution.

To define it precisely, suppose that the data are generated according to a fixed but unknown distribution $P_1(\mathbf{x}; \boldsymbol{\theta}_1) \in \mathcal{M}_1$. If we knew $\boldsymbol{\theta}_1$, we could construct the GRO e-variable $S^{\mathrm{GRO}(\boldsymbol{\theta}_1)}$ optimal for that specific alternative:

$$S^{\mathrm{GRO}(\boldsymbol{\theta}_1)} = \frac{P_1(\mathbf{x}; \boldsymbol{\theta}_1)}{P_0^{\tilde{w}_0^*}(\mathbf{x})} \tag{12}$$

where

$$\tilde{w}_0^* = \arg \min_{w_0 \in \mathcal{W}_{\boldsymbol{\theta}_0}} \mathbb{E}_{\boldsymbol{\theta}_1}\left[\log \frac{P_1(\mathbf{x}; \boldsymbol{\theta}_1)}{P_0^{w_0}}\right] \tag{13}$$

and $\mathbb{E}_{\boldsymbol{\theta}_j}$ denotes the expected value under $P_j(\cdot\,, \boldsymbol{\theta}_j)$. Here, the alternative hypothesis reduces to a *singleton* — a statistical model containing only one distribution — and in some cases, such as the $2 \times k$ contingency tables considered later in this paper, $S^{\mathrm{GRO}(\boldsymbol{\theta}_1)}$ can be computed exactly. The regret of a candidate e-variable S_{cand} is then given by:

$$\mathrm{REG}_1(\boldsymbol{\theta}_1; S_{\mathrm{cand}}) := \mathbb{E}_{\boldsymbol{\theta}_1}\left[\log S^{\mathrm{GRO}(\boldsymbol{\theta}_1)} - \log S_{\mathrm{cand}}\right], \tag{14}$$

which quantifies the expected loss in log-growth due to not knowing the true parameter.

Since $\boldsymbol{\theta}_1$ is unknown, a natural robustness criterion is to consider the *worst-case regret* across the entire alternative:

$$\mathrm{REG}_1(\boldsymbol{\Theta}_1; S_{\mathrm{cand}}) := \max_{\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1} \mathrm{REG}_1(\boldsymbol{\theta}_1; S_{\mathrm{cand}}). \tag{15}$$

This leads to a new optimality principle: among all e-variables, one would seek the *minimax optimal* e-variable — i.e., the e-variable that minimizes $\mathrm{REG}_1(\boldsymbol{\Theta}_1; S_{\mathrm{cand}})$ over all valid choices of S_{cand}. However, computing this minimax-optimal e-variable is generally infeasible in practice, as we currently lack efficient algorithms for solving the corresponding optimization problem. Nevertheless, when the models $\mathcal{M}_0$ and $\mathcal{M}_1$ exhibit sufficient regularity — as is the case for Maximum Entropy models, discussed in the next section — GRO e-variables constructed from (11) with appropriately chosen $\bar{P}_1$ can closely approximate the minimax optimal solution. In particular, one can consider e-variables of the form (11), where the numerator $\bar{P}_1$ is set to a *universal distribution* relative to the alternative model $\mathcal{M}_1$. Universal distributions, which include Bayesian mixtures $P_1^{w_1}$ as special cases, arise naturally in the theory of the *Minimum Description Length (MDL) Principle* [30–32]. Such choices of $\bar{P}_1$ lead to e-variables that, while not exactly minimax-optimal, are typically close to optimal in terms of regret minimization, and therefore provide a practical and principled strategy for robust hypothesis testing. To clarify this connection, we take a brief detour to explain how e-value-based methods relate to the MDL Principle and its central concept, the universal distribution.

A. GRO e-values and description lengths

The Minimum Description Length Principle provides a general framework for model selection: from a set of candidate models, it chooses the one that yields the shortest encoding of the observed data. In this approach, each model is represented by a single probability distribution, and models are compared via their *description length*. The preferred model is the one with the smallest description length.

When comparing two models $\mathcal{M}_0$ and $\mathcal{M}_1$, the difference in description lengths is

$$\begin{aligned}\Delta\mathrm{DL}(\mathbf{x}) &= \mathrm{DL}_1(\mathbf{x}) - \mathrm{DL}_0(\mathbf{x}) \\ &= -\log \bar{P}_1(\mathbf{x}) + \log \bar{P}_0(\mathbf{x}),\end{aligned} \tag{16}$$

where $\bar{P}_1$ and $\bar{P}_0$ are the representative distributions for $\mathcal{M}_1$ and $\mathcal{M}_0$. By Kraft's inequality [30, 33], the code length to describe $\mathbf{x}$, using a code that compresses optimally in expectation under Q, is (up to rounding) $-\log Q(\mathbf{x})$ bits; thus, $-\log \bar{P}_j(\mathbf{x})$ is the code length implied by $\bar{P}_j$.

Universal distributions and worst-case redundancy

The key point is how to determine a single probability distribution $\bar{P}_j$ representing $\mathcal{M}_j$: it should perform well regardless of which specific distribution within $\mathcal{M}_j$ generated the data. In other words, if a distribution $P \in \mathcal{M}_j$ achieves a short expected code length $\mathbb{E}_P[-\log P(\mathbf{x})]$, then $\bar{P}_j$ should yield a similarly short one. Such $\bar{P}_j$ are called *universal distributions* for the model $\mathcal{M}_j$ [30].

To be more precise, we can define the *redundancy* of $\bar{P}_j$ relative to a parameter $\boldsymbol{\theta}_j$ as

$$\mathrm{RED}_j(\boldsymbol{\theta}_j; \bar{P}_j) := \mathbb{E}_{\boldsymbol{\theta}_j}\big[-\log \bar{P}_j(\mathbf{x}) + \log P_j(\mathbf{x}; \boldsymbol{\theta}_j)\big] \tag{17}$$

This quantity measures the expected extra bits needed when using $\bar{P}_j$ instead of the expected optimal code for $P_j(\cdot; \boldsymbol{\theta}_j)$. The latter is not available in practice, since the true $\boldsymbol{\theta}_j$ is typically unknown. Thus, it is useful defining the *worst-case redundancy*:

$$\mathrm{RED}_j(\boldsymbol{\Theta}_j; \bar{P}_j) := \max_{\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j} \mathrm{RED}_j(\boldsymbol{\theta}_j; \bar{P}_j). \tag{18}$$

A distribution is universal if this quantity is small.

Ideally, one would like to find a $\bar{P}_j$ that minimizes the worst-case redundancy, but this is generally infeasible. However, for a d_j-dimensional parametric model under standard regularity conditions (satisfied by the Maximum Entropy models considered later), the Bayesian choice $\bar{P}_j = P_j^{w_j}$ with a regular prior w_j attains near-optimal performance: its redundancy is within a constant of the optimal value as the sample size grows. This is formalized in the following result.

Definition 1 (INECCSI sets [30]) *Let*

$$\mathcal{M}_j = \{P_j(\mathbf{x};\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_j}, \quad j \in \{0,1\}.$$

A subset $\boldsymbol{\Theta}'_j \subset \boldsymbol{\Theta}_j$ is an INECCSI subset *if its interior is a non-empty, convex, compact subset of the interior of $\boldsymbol{\Theta}_j$.*

INECCSI subsets exclude boundary effects and ensure regular asymptotics. For instance, in the Bernoulli model with $\Theta = [0,1]$, any $[\epsilon, 1-\epsilon]$ with $0 < \epsilon < 1/2$ is INECCSI.

Let $\mathcal{M}_j^{(m)}$ be the i.i.d. extension of $\mathcal{M}_j$ to m observations, i.e., a model over $\mathbf{y}^m = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ where each $\mathbf{x}_i \in \mathcal{X}$ is independently sampled from $P_j(\cdot;\boldsymbol{\theta})$. Let $P_j^{(m)}$ be the i.i.d. extension of P_j and $\bar{P}_j^{(m)}$ be a distribution on $\mathbf{y}^m$. Let $\mathrm{RED}^m(\boldsymbol{\Theta}_j; \bar{P}_j^{(m)})$ be the worst-case redundancy attained by $\bar{P}_j^{(m)}$. Since $\mathbb{E}_{\boldsymbol{\theta}_j}[-\log P_j^{(m)}(\mathbf{y}^m;\boldsymbol{\theta}_j)]$ grows linearly in m, universality of $\bar{P}_j^{(m)}$ requires RED^m to grow sub-linearly in m.

A standard result [30] states that for every INECCSI subset $\boldsymbol{\Theta}'_j$ and regular prior w_j, there exists $C > 0$ such that for all m:

$$\begin{aligned}\frac{d_j}{2}\log m - C \;&\le\; \inf_{\bar{P}_j^{(m)}} \mathrm{RED}^m(\boldsymbol{\Theta}'_j; \bar{P}_j^{(m)}) \\ &\le\; \mathrm{RED}^m(\boldsymbol{\Theta}'_j; P_j^{w_j(m)}) \;\le\; \frac{d_j}{2}\log m + C, \quad (19)\end{aligned}$$

where the infimum is over all distributions on $\mathcal{X}^{(m)}$. The key implications of these results are:

- the minimum achievable worst-case redundancy grows as $(d_j/2)\log m$;
- Bayesian marginal likelihoods are universal distributions, and their redundancy exceeds the minimum attainable by at most a constant — in this sense, they are asymptotically almost optimal.

In a variation of the definition of universality, we may search for $\bar{P}_j$ such that $-\log \bar{P}_j(\mathbf{x}) - \log P_{\hat{\boldsymbol{\theta}}_j}(\mathbf{x})$ is small, in the worst-case over all possible data realizations $\mathbf{x}$; here $\hat{\boldsymbol{\theta}}_j(\mathbf{x})$ is the maximum likelihood estimator of $\boldsymbol{\theta}_j$ relative to $\mathbf{x}$. This leads to comparing $\bar{P}_j(\mathbf{x})$ to the best-fitting model *a posteriori*, that is, the model that, with hindsight, would give the shortest code for the observed data.

Within the MDL literature, this more stringent criterion is often viewed as the ideal one. The distribution that achieves this is called the Normalized Maximum Likelihood (NML) distribution [30–32]. It assigns probabilities by maximizing the likelihood of the observed data while normalizing over all possible datasets of the same size:

$$\bar{P}_j^{\mathrm{NML}}(\mathbf{x}) = \frac{P_j(\mathbf{x}; \hat{\boldsymbol{\theta}}_j(\mathbf{x}))}{\sum_{\mathbf{y}\in\mathcal{X}} P_j(\mathbf{y}; \hat{\boldsymbol{\theta}}_j(\mathbf{y}))}. \quad (20)$$

Importantly, the NML distribution does not require any prior, and achieves universality both in worst-case data and in worst-case expected regret. In fact, for the MEM models we introduce below, inequality (19) also holds when $P_j^{w_j(m)}$ is replaced by $\bar{P}_j^{\mathrm{NML}(m)}$.

Universal distributions guarantee low-regret e-variables

We can now formalize the connection between e-values and MDL: we show that using a universal distribution $\bar{P}_1$ as the numerator in the e-variable construction (11) leads to small regret. This provides a principled justification for the use of Bayesian mixtures and NML in e-value methods.

To see this, notice that (minus) the log-ratio of any variable of the form

$$S(\mathbf{x}) = \frac{\bar{P}_1(\mathbf{x})}{P_0^{w_0}(\mathbf{x})}$$

induces a difference in description lengths between models $\mathcal{M}_1$ and $\mathcal{M}_0$:

$$-\log S(\mathbf{x}) = -\log \bar{P}_1(\mathbf{x}) - [\log P_0^{w_0}(\mathbf{x})]. \quad (21)$$

This mirrors expression (16), but with one crucial difference: in order for S to be an e-variable, the denominator $P_0^{w_0}$ cannot be chosen freely, as it must be the prior w_0^* that ensures that S qualifies as a GRO e-variable (i.e., satisfies the e-variable condition).

This formulation, however, brings a clear interpretative advantage: the description length difference expressed in (21) now has a direct statistical interpretation. Indeed, if S is an e-variable, the corresponding code-length difference can be mapped to a statistical significance measure, since Type I error control is guaranteed. This grounds the MDL code-length difference in a frequentist hypothesis testing framework. In particular, smaller values of $-\log S(\mathbf{x})$ correspond to larger e-values and hence stronger evidence against the null model $\mathcal{M}_0$. This observation addresses a longstanding issue in MDL: although it provides a principled model comparison method, it lacks explicit statistical guarantees such as Type I error control [30, Open Problem No.

9, page 413]. Restricting attention to code-length differences that admit an e-value interpretation not only provides such guarantees but also makes it possible to assign a well-defined evidential value to differences in description lengths. This can be seen as the natural solution to the problem — at least for the two-model comparison case [31]. Extending this insight to multiple models remains an important open challenge.

The connection between e-values and MDL becomes even more compelling when considering the regret of e-variables. Indeed, we now show that the worst-case regret of an e-variable using numerator $\bar{P}_1$ is never larger than the worst-case redundancy of $\bar{P}_1$. Let us restrict attention to an INECCSI subset $\boldsymbol{\Theta}_1' \subset \boldsymbol{\Theta}_1$. Moreover, for clarity, let's denote the regret relative to the GRO e-variable associated to $\bar{P}_1$, i.e. the regret obtained by putting $S_{\text{cand}} = \bar{P}_1(\mathbf{x})/P_0^{w_0^*}(\mathbf{x})$ in definition 15, as $\text{REG}_1(\boldsymbol{\Theta}_1'; \bar{P}_1)$. Then, for any distribution $\bar{P}_1$, according to definitions (15) and (12) (see section S1) it holds:

$$\text{REG}_1(\boldsymbol{\Theta}_1'; \bar{P}_1) \leq \text{RED}_1(\boldsymbol{\Theta}_1'; \bar{P}_1). \tag{22}$$

This shows that, in the worst-case over $\boldsymbol{\theta}_1$, the regret of an e-variable built with numerator $\bar{P}_1$ is upper bounded by the redundancy of $\bar{P}_1$. Consequently, choosing a universal distribution $\bar{P}_1$, which by definition provides small redundancy, guarantees small regret.

This motivates our choices in the next sections: we will construct e-variables by setting $\bar{P}_1$ either to the Bayesian mixture $P_1^{w_1}$ (with a regular prior), or to the Normalized Maximum Likelihood $\bar{P}_1^{\text{NML}}$. Both yield small regret of order $(d_1/2)\log m + O(1)$. These choices are also common in the literature. The NML distribution was used (implicitly) in [34] as an e-variable numerator, and Bayesian mixtures are widely adopted in the construction of e-variables [4].

III. APPLICATION TO MAXIMUM ENTROPY MODELS

Here, we provide explicit formulas for hypothesis tests that involve either microcanonical or canonical maximum entropy models. We focus on the case of discrete data. For a given choice of sufficient statistics $\mathbf{c}(\mathbf{x})$, we denote by $\mathcal{C}$ the discrete set of values of $\mathbf{c}(\mathbf{x})$ that are realizable by at least one $\mathbf{x} \in \mathcal{X}$; for mathematical convenience, we assume that $\mathcal{C}$ is a (finite or countable) subset of $\mathbb{R}^d$ for some $d \in \mathbb{N}$. Moreover, for any given value $\mathbf{c} \in \mathcal{C}$, let $\Omega(\mathbf{c})$ represent the number of configurations satisfying the constraint $\mathbf{c}(\mathbf{x}) = \mathbf{c}$, formally defined as:

$$\Omega(\mathbf{c}) := \sum_{\mathbf{x}\,:\,\mathbf{c}(\mathbf{x})=\mathbf{c}} 1. \tag{23}$$

Entropy maximization, when the hard constraints $\mathbf{c}(\mathbf{x}) = \mathbf{c}$ are enforced on each realizable configuration $\mathbf{x}$, yields a *microcanonical* model whose functional form is a uniform distribution over data satisfying the constraints:

$$P_{\text{mic}}(\mathbf{x}; \mathbf{c}) = \begin{cases} \frac{1}{\Omega(\mathbf{c})}, & \text{if } \mathbf{c}(\mathbf{x}) = \mathbf{c}; \\ 0, & \text{else.} \end{cases} \tag{24}$$

The parameters of a microcanonical model correspond to the sufficient statistics themselves, with values in the discrete parameter space $\boldsymbol{\Theta}_{\text{mic}} = \mathcal{C}$.

When soft constraints $\mathbb{E}[\mathbf{c}(\mathbf{x})] = \mathbf{c}$ are enforced (that is, the value $\mathbf{c}$ of the sufficient statistic is to be met only as an ensemble average), the maximization of the entropy returns a *canonical* model where the resulting functional form of the probability distribution is, instead, exponential, with positive probability for all possible data:

$$P_{\text{can}}(\mathbf{x}; \boldsymbol{\theta}) = \frac{e^{-\boldsymbol{\theta}\cdot\mathbf{c}(\mathbf{x})}}{Z(\boldsymbol{\theta})} \tag{25}$$

where $Z(\boldsymbol{\theta}) \equiv \sum_{\mathbf{x}\in\mathcal{X}} e^{-\boldsymbol{\theta}\cdot\mathbf{c}(\mathbf{x})}$ is a normalization term known as *partition function*. Canonical models coincide with what is called *exponential families with uniform carrier function* in the statistics literature [15], and the formula above is generally referred to as canonical parametrization, where the parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\text{can}}$ may be viewed as the Lagrange multipliers resulting from the entropy maximization. For each value $\mathbf{c}$ defining the microcanonical model in Eq. (24), there is a corresponding value $\boldsymbol{\theta}$ such that $\mathbb{E}_{\boldsymbol{\theta}}[\mathbf{c}(\mathbf{x})] = \mathbf{c}$ under the canonical distribution in Eq. (25).

The above 'duality' between canonical and microcanonical models implies that, alternatively, canonical models can also be parameterized using the expected value of the sufficient statistics. Given parameters $\boldsymbol{\theta}$, define the mean value vector:

$$\boldsymbol{\mu}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{c}(\mathbf{x})]. \tag{26}$$

This defines a smooth, one-to-one mapping between the canonical parameter space $\boldsymbol{\Theta}_{\text{can}}$ and the set of realizable mean values, which we denote by $\mathtt{M}$. In exponential family theory, $\boldsymbol{\mu}$ is known as the *mean value parameter*. For future reference, we refer to $P_{\boldsymbol{\mu}} = P_{\text{can}}(\cdot; \boldsymbol{\theta}(\boldsymbol{\mu}))$ as the canonical distribution defined in its mean value parametrization, where $\boldsymbol{\theta}(\boldsymbol{\mu})$ is the mapping from mean-value parameters to corresponding canonical parameters, i.e. the inverse of $\boldsymbol{\mu}(\boldsymbol{\theta})$.

A well-known result in this setting is that, if the set of possible constraint values $\mathcal{C}$ is finite, then:

- the canonical parameter space is the full space $\boldsymbol{\Theta}_{\text{can}} = \mathbb{R}^d$;
- the corresponding space of mean values $\mathtt{M}$ coincides with the interior of the convex hull of $\mathcal{C}$.

This result ensures that the mapping $\boldsymbol{\theta} \mapsto \boldsymbol{\mu}$ is not only bijective but also covers all "physically meaningful" ex-

pected constraint values[2]. In the rest of the paper, we will make use of this bijection and employ whichever parameterization is most convenient. In particular, when dealing with Bayesian marginal likelihoods and their priors, we will typically work in the mean-value space, bearing in mind that all results can be equivalently expressed in the canonical parameter space via the mapping $\boldsymbol{\theta} \mapsto \boldsymbol{\mu}(\boldsymbol{\theta})$.

In what follows, we define GRO e-variables for tests where both the null and the alternative hypotheses are two microcanonical or two canonical MEMs that differ in the choice of constraints. Our main theoretical results are presented in a general form, but to guide the reader through the derivations, we will use a running example throughout (Examples A, B, and C): a simple 2×2 contingency table, representing two groups of binary data.

A. Microcanonical test

Consider a test where the null $\mathcal{M}_{\text{mic},0}$ is a microcanonical model with sufficient statistics $\mathbf{c}_0$ taking values in set $\mathcal{C}_0$ and the alternative $\mathcal{M}_{\text{mic},1}$ is a microcanonical model with sufficient statistics $\mathbf{c}_1$ taking values in set $\mathcal{C}_1 \neq \mathcal{C}_0$. The parameters of microcanonical models are discrete and correspond to their sufficient statistics. As shown in [36], the NML microcanonical distribution is equivalent to a Bayesian distribution with a uniform prior over the sufficient statistics. Therefore, in this section, we restrict our analysis to the case of microcanonical Bayesian universal distributions for the alternative hypothesis, denoted by $P^{W_1}_{\text{mic},1}$, where W_1 is a probability mass function defined on $\mathcal{C}_1$. Thus, the microcanonical GRO e-variable reads

$$S^{\text{GRO}}_{\text{mic}} = \frac{P^{W_1}_{\text{mic},1}(\mathbf{x})}{P^{W^*_0}_{\text{mic},0}(\mathbf{x})} \tag{27}$$

and it solves the discrete version of the GRO optimization problem:

$$W^*_0 = \arg \min_{W \in \mathcal{W}_{\mathbf{c}_0}} D_{\text{KL}}(P^{W_1}_{\text{mic},1} \| P^{W_0}_{\text{mic},0}) \tag{28}$$

where instead of prior densities w_0, we need to consider prior probability mass functions W_0, and $\mathcal{W}_{\mathcal{C}_0}$ is the set of all such distributions on the parameter space $\mathcal{C}_0$. We solve the microcanonical GRO optimization problem explicitly and exactly (full derivation in SM; here we report only the main results) and find the optimal prior distribution on the null:

$$W^*_0(\mathbf{c}_0) = \sum_{\mathbf{x}\,:\,\mathbf{c}_0(\mathbf{x})=\mathbf{c}_0} P^{W_1}_{\text{mic},1}(\mathbf{x}), \tag{29}$$

[2] It follows from the general theory of exponential families with finite support [35, Theorem 9.2], under a technical condition known as *steepness*, which holds when $\mathcal{C}$ is finite.

i.e., $W^*_0(\mathbf{c}_0)$ is the marginal distribution of the null sufficient statistic $\mathbf{c}_0(\mathbf{x})$ induced by $P^{W_1}_{\text{mic},1}$. In the special case where the alternative sufficient statistics completely determine the value of the null, we say that *Condition A* holds:

Condition A: (30)
there exists a function $f : \mathcal{C}_1 \to \mathcal{C}_0$ s.t. $\mathbf{c}_0(\mathbf{x}) = f(\mathbf{c}_1(\mathbf{x}))$.

Under Condition A, one can write:

$$W^*_0(\mathbf{c}_0) = \sum_{\mathbf{c}_1\,:\,f(\mathbf{c}_1)=\mathbf{c}_0} W_1(\mathbf{c}_1), \tag{31}$$

i.e., the GRO-optimal prior on the null is the distribution induced on the null sufficient statistics by the alternative prior, or equivalently, the marginal distribution of $\mathbf{c}_0$ induced by W_1.

Once that W^*_0 is computed, the microcanonical GRO-optimal e-variable can always be expressed as

$$S^{\text{GRO}}_{\text{mic}}(\mathbf{x}) = \frac{\Omega_0(\mathbf{c}_0(\mathbf{x}))}{\Omega_1(\mathbf{c}_1(\mathbf{x}))} \frac{W_1(\mathbf{c}_1(\mathbf{x}))}{W^*_0(\mathbf{c}_0(\mathbf{x}))}. \tag{32}$$

Finally, although the fact that $S^{\text{GRO}}_{\text{mic}}$ is an e-variable follows from a general theorem (Theorem 1 in [7], as mentioned above Equation (10)), we additionally provide a further, direct proof showing that its expected value under the null is exactly one:

$$\mathbb{E}_0[S^{\text{GRO}}_{\text{mic}}] = 1 \quad \forall P_{\text{mic},0} \in \mathcal{M}_{\text{mic},0}. \tag{33}$$

This direct proof, as well as the section's other detailed calculations and proofs, can be found in section S2. For clarity, we now provide a first example application.

Example A

Let us consider the dataset $\mathbf{x} = (\mathbf{x}^a, \mathbf{x}^b)$ consisting of two groups of binary data, represented as $\mathbf{x}^a = (x^a_1, ..., x^a_{n^a})$ and $\mathbf{x}^b = (x^b_1, ..., x^b_{n^b})$, with n^a and n^b the respective group sizes. The total sample size is $n = n_a + n_b$. We denote by $n^a_1 = \sum_{i=1}^{n_a} x^a_i$ and $n^b_1 = \sum_{i=1}^{n_b} x^b_i$ the total number of 1s in $\mathbf{x}^a$ and $\mathbf{x}^b$, and by $n_1 = n^a_1 + n^b_1$ the total number of 1s in $\mathbf{x}$. The aim is to build a microcanonical test to check whether the probability of observing $x = 1$ changes according to the different groups. To do so, we set the alternative sufficient statistics equal to the number of 1s in each group, $\mathbf{c}_1 = (n^a_1, n^b_1)$, and the null sufficient statistic equal to the total number of 1s, $c_0 = n_1$. In the microcanonical formulation, these quantities are treated as fixed in the respective models. To find the microcanonical GRO e-variable, we apply formula (32), where:

- $\Omega_0(n_1) = \binom{n}{n_1}$ is the number of permutations of $\mathbf{x}$ preserving the total number of 1s;

- $\Omega_1(n_1^a, n_1^b) = \binom{n^a}{n_1^a}\binom{n^b}{n_1^b}$ is the number of permutations of $\mathbf{x}$ preserving the total number of 1s in each group.

For the sake of this example, we put independent, discrete uniform priors on the alternative parameters n_1^a and n_1^b:

$$W_1(n_1^a, n_1^b) = \mathcal{U}_a(n_1^a)\,\mathcal{U}_b(n_1^b) = \frac{1}{n^a+1}\,\frac{1}{n^b+1}. \quad (34)$$

In this case, Condition A (30) holds, as the null sufficient statistics can be written as a function of the alternative one: $n_1 = n_1^a + n_1^b$. Thus, the optimal prior on the null W_0^* is the distribution of n_1 induced by W_1. In this case, that is simply the convolution of $\mathcal{U}_a$ and $\mathcal{U}_b$, which is a triangular discrete function:

$$W_0^*(n_1) = \begin{cases} \frac{n_1+1}{(n^a+1)(n^b+1)}, & \text{if } 0 \le n_1 \le \min(n^a, n^b), \\ \frac{\min(n^a,n^b)+1}{(n^a+1)(n^b+1)}, & \text{if } \min(n^a,n^b) < n_1 \le \max(n^a,n^b), \\ \frac{n^a+n^b+1-n_1}{(n^a+1)(n^b+1)}, & \text{if } \max(n^a,n^b) < n_1 \le n^a+n^b, \\ 0, & \text{if otherwise.} \end{cases} \quad (35)$$

as shown in Figure 2.

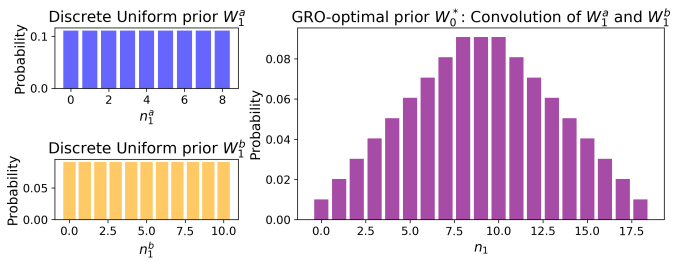

FIG. 2. In the microcanonical Example A, when the prior on the alternative sufficient statistics n_1^a and n_1^b are uniform distributions (on the left), the resulting GRO-optimal prior on the null sufficient statistic n_1 is the convolution of the two uniform distributions, which results in a triangular distribution (on the right). In this example, $n^a = 8$ and $n^b = 10$.

B. Canonical test

Consider a test where the null $\mathcal{M}_{\text{can},0}$ is a canonical model with sufficient statistics $\mathbf{c}_0$ taking values in set $\mathcal{C}_0$, and the alternative $\mathcal{M}_{\text{can},1}$ is a canonical model with sufficient statistics $\mathbf{c}_1$ taking values in set $\mathcal{C}_1 \neq \mathcal{C}_0$. The goal is to find the canonical GRO e-variable:

$$S_{\text{can}}^{\text{GRO}} = \frac{\bar{P}_{\text{can},1}(\mathbf{x})}{P_{\text{can},0}^{w_0^*}(\mathbf{x})} \quad (36)$$

where w_0^* is a prior density on the mean-value parameter space $\mathtt{M}_0$ and solves the optimization problem

$$w_0^* = \arg\min_{w \in \mathcal{W}_{\boldsymbol{\mu}_0}} D_{\text{KL}}(\bar{P}_1 \| P_0^{w_0}). \quad (37)$$

No exact general solution is currently available for this problem. While in some cases it can be solved analytically or numerically, in the majority of cases, there is neither a known analytic solution nor a feasible numerical approach. Here, we propose two candidate approximations, the *microcanonical approximation* and the *pseudo approximation*. The first will serve as an actual approximation, and the second as a tool to assess whether the former approximation is good.

The definition of the microcanonical approximation is based on two facts, proven in section S3:

- A canonical universal distribution $\bar{P}_{\text{can},1}$ with sufficient statistics $\mathbf{c}_1$ can always be expressed as a microcanonical Bayesian marginal likelihood, i.e., there always exists a prior probability mass function $W_{\text{can},1}(\mathbf{c})$ such that

$$\bar{P}_{\text{can},1} = P_{\text{mic},1}^{W_{\text{can},1}} \quad (38)$$

with $W_{\text{can},1}$ obtained by setting (for $j = 1$):

$$W_{\text{can},j}(\mathbf{c}_j) = \sum_{\mathbf{x}\,:\,\mathbf{c}_j(\mathbf{x}) = \mathbf{c}_j} \bar{P}_{\text{can},j}(\mathbf{x}), \quad (39)$$

i.e., $W_{\text{can},j}(\mathbf{c}_j)$ is equal to the distribution of $\mathbf{c}_j(\mathbf{x})$ induced by $\bar{P}_{\text{can},j}(\mathbf{x})$.

- Given the canonical and microcanonical models $\mathcal{M}_{\text{can}}$ and $\mathcal{M}_{\text{mic}}$ built upon the same sufficient statistic $\mathbf{c}(\mathbf{x})$, a microcanonical e-variable E is always a canonical e-variable:

$$\mathbb{E}_P[E] \le 1 \;\; \forall P \in \mathcal{M}_{\text{mic}} \quad \Rightarrow \quad \mathbb{E}_P[E] \le 1 \;\; \forall P \in \mathcal{M}_{\text{can}}. \quad (40)$$

Following the first fact, given $\bar{P}_{\text{can},1}$ and using the results of the previous section, we can build the approximating microcanonical GRO e-variable $S_{\text{mic}}^{\text{GRO}}$ for the microcanonical test based on the corresponding $P_{\text{mic},1}^{W_{\text{can},1}} = \bar{P}_{\text{can},1}$. Thus (27) becomes

$$S_{\text{mic}}^{\text{GRO}} = \frac{\bar{P}_{\text{can},1}(\mathbf{x})}{P_{\text{mic},0}^{W_0^*}(\mathbf{x})} \quad (41)$$

where, readapting (29)

$$W_0^*(\mathbf{c}_0) = \sum_{\mathbf{x}\,:\,\mathbf{c}_0(\mathbf{x}) = \mathbf{c}_0} \bar{P}_{\text{can},1}(\mathbf{x}). \quad (42)$$

Given the second fact (40), the resulting microcanonical GRO e-variable is a valid canonical e-variable, i.e., it is an e-variable for the test between two canonical models, even if, for this test, it is not the GRO-optimal one. As such, from (6), it will have a smaller e-power than the canonical GRO one unless the two coincide:

$$\mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S_{\text{mic}}^{\text{GRO}}\right] \le \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S_{\text{can}}^{\text{GRO}}\right] \quad (43)$$

The pseudo approximation is further built over the microcanonical one. The prior W_0^* (used to build $S_{\text{mic}}^{\text{GRO}}$), defined on $\mathcal{C}_0$, is transformed into a smooth density $w_{\text{pseudo},0}$ over the corresponding (continuous) mean-value parameter space $\mathtt{M}_0$. This is obtained through a high resolution limit, by computing $W_0^*(\mathbf{c}_0)$ for a much higher dimension and by properly rescaling and normalizing it such that it is interpreted as a Riemann approximation of a continuous density on $\boldsymbol{\mu}_0$. A practical example of this procedure, which might seem abstract at this stage, is given in Examples B and C. Moreover, a pseudo-code is provided in section S5. Given that, in general, $w_{\text{pseudo},0}$ is different from the GRO-optimal prior w_0^*, which in most cases remains unknown, the resulting variable

$$S_{\text{pseudo}} = \frac{\bar{P}_{\text{can},1}(\mathbf{x})}{P_{\text{can},0}^{w_{\text{pseudo},0}}(\mathbf{x})} \tag{44}$$

is not an e-variable, unless $w_{\text{pseudo},0} = w_0^*$. Indeed, from Theorem 1 of [7], $S_{\text{can}}^{\text{GRO}}$ is the only e-variable of that form. Moreover, from (8), it holds:

$$D_{\text{KL}}(\bar{P}_{\text{can},1} \| P_{\text{can},0}^{w_0^*}) \leq D_{\text{KL}}(\bar{P}_{\text{can},1} \| P_{\text{can},0}^{w_{\text{pseudo},0}}) \tag{45}$$

or, equivalently:

$$\mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S_{\text{can}}^{\text{GRO}}\right] \leq \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S_{\text{pseudo}}\right] \tag{46}$$

Consequently, one has

$$\begin{aligned} \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S_{\text{mic}}^{\text{GRO}}\right] &\leq \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S_{\text{can}}^{\text{GRO}}\right] \\ &\leq \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S_{\text{pseudo}}\right], \end{aligned} \tag{47}$$

i.e., the two approximations provide an upper and a lower bound for the e-power of the canonical GRO e-variable. In summary, when the canonical GRO e-variable is not available, we can follow a two-step procedure:

1. We build the corresponding microcanonical approximation, knowing that it is a valid candidate e-variable. To build it, we first transform the canonical universal distribution into a microcanonical one, by finding $W_{\text{can},1}$ as in (39). Then, we compute $S_{\text{mic}}^{\text{GRO}}$ according to the formulas expressed in the previous section (Equations (29) and (32)).

2. The goodness of the microcanonical approximation can be evaluated by looking at the width of the interval

$$r = \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S_{\text{pseudo}}\right] - \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S_{\text{mic}}^{\text{GRO}}\right] \geq 0 \tag{48}$$

where, for future reference, it is useful to note that, using definitions (44) and (41) and sufficiency, we can rewrite

$$\begin{aligned} r &= \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log P_{\text{mic},0}^{W_0^*}(\mathbf{x}) - \log P_{\text{can},0}^{w_{\text{pseudo},0}}(\mathbf{x})\right] \\ &= \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log W_0^*(\mathbf{c}_0(\boldsymbol{x})) - \log W_{\text{pseudo},0}(\mathbf{c}_0(\mathbf{x}))\right]. \end{aligned} \tag{49}$$

where $W_{\text{pseudo},0}(\mathbf{c}_0(\mathbf{x}))$ is defined as in (39).

The evaluation above is under $\bar{P}_{\text{can},1}$-expectation; this makes sense if we use a Bayesian universal distribution $\bar{P}_{\text{can},1} = P_{\text{can},1}^{w_1}$ and the prior w_1 is a reasonable expression of our uncertainty. If we are not so sure about our priors, or if $\bar{P}_{\text{can},1}$ is non-Bayesian, we may be interested in a more stringent, worst-case measure for evaluating the performance of the microcanonical approximation. In analogy with the worst-case REG defined in 15, we define an alternative version of r, denoted by r', which can be defined both relatively to a single parameter $\boldsymbol{\theta}_1$ (equivalently and more conveniently relative to $\boldsymbol{\mu}_1 = \boldsymbol{\mu}(\boldsymbol{\theta}_1)$):

$$\begin{aligned} r'(\boldsymbol{\mu}_1) &= \mathbb{E}_{\boldsymbol{\mu}_1}\left[\log S_{\text{pseudo}} - \log S_{\text{mic}}^{\text{GRO}}\right] \\ &= \mathbb{E}_{\boldsymbol{\mu}_1}\left[\log W_0^*(\mathbf{c}_0(\boldsymbol{x})) - \log W_{\text{pseudo},0}(\mathbf{c}_0(\mathbf{x}))\right], \end{aligned} \tag{50}$$

where $\mathbb{E}_{\boldsymbol{\mu}}$ denotes the expected value under $P_{\boldsymbol{\mu}}$, and in its worst-case version, which for clarity will be simply denoted by r':

$$r' := \max_{\boldsymbol{\mu}_1 \in \mathtt{M}_1} r'(\boldsymbol{\mu}_1). \tag{51}$$

It can be easily argued that

$$r \geq 0 \Rightarrow r' \geq 0. \tag{52}$$

In case r (or r') is small, we know that our easily computable microcanonical e-variable $S_{\text{mic}}^{\text{GRO}}$ is close to optimal according to the GRO criterion for the canonical problem, and hence can be used instead of the canonical $S_{\text{can}}^{\text{GRO}}$. In the following example, which is a continuation of Example A, we show a practical case where this turns out to be true.

Example B (continued from Example A)

We consider the same setting as in Example A, but in this case, we are interested in constructing a canonical test. In a canonical formulation, the observed number of 1s is fixed only in expectation. As a result, the null model is a collection of n i.i.d. Bernoulli variables, where the parameter is the probability $p_0 \in [0,1]$ of observing $x = 1$, which is the same regardless of the group. The alternative model, instead, assumes that data in the two groups are independent Bernoulli variables, where the parameters are the probabilities $(p_a, p_b) \in [0,1]^2$ of observing $x = 1$, which depend on the group. The aim of the tests is to assess whether p_a and p_b are the same or whether they are different. Again, for the sake of this example, we put independent, continuous uniform priors on the alternative parameters p_a and p_b, $w_1(p_a, p_b) = u(p_a)u(p_b)$ where $u(p) = 1$ if $p \in [0,1]$ and $u(p) = 0$ else. In this simple case, the Bayesian marginal likelihood can be computed analytically, and it reads:

$$\begin{aligned} P_{\text{can},1}^{w_1} &= \int_0^1 p_a^{n_1^a}(1-p_a)^{n^a - n_1^a} dp_a \int_0^1 p_b^{n_1^b}(1-p_b)^{n^b - n_1^b} dp_b \\ &= \binom{n^a}{n_1^a}^{-1} \frac{1}{n^a+1} \binom{n^b}{n_1^b}^{-1} \frac{1}{n^b+1} \end{aligned} \tag{53}$$

Following the procedure described in this section, we first build the microcanonical approximation. To do so, we need to compute the probability $W_{\text{can},1}$ induced by $P_1^{w_1}$ on the alternative sufficient statistics, such that $P_{\text{can},1}^{w_1} = P_{\text{mic},1}^{W_{\text{can},1}}$. By inspecting Eq. (53), it is easy to observe that $W_{\text{can},1}$ is the uniform distribution: $W_{\text{can},1} = (n_a+1)^{-1}(n_b+1)^{-1} = \mathcal{U}_a(n_1^a)\,\mathcal{U}_b(n_1^b)$. Thus, we can compute the microcanonical approximation $S_{\text{mic}}^{\text{GRO}}$ by using the results of Example A. As a second step, we check whether this microcanonical e-variable is a good approximation by studying the behavior of the interval width r as the total size n increases. To evaluate S_{pseudo}, we compute the prior $w_{\text{pseudo},0}$ as described above (denoted by $w_{\text{pseudo},0}^1$ in Figure 4 to be distinct from $w_{\text{pseudo},0}^2$ of the following Example C). A schematic representation of how $S_{\text{mic}}^{\text{GRO}}$ and S^{pseudo} are built is shown in Figure 4.

Once $S_{\text{mic}}^{\text{GRO}}$ and S^{pseudo} are computed, we can compute r and show that the microcanonical approximation works very well in our simple example (see Figure 3).

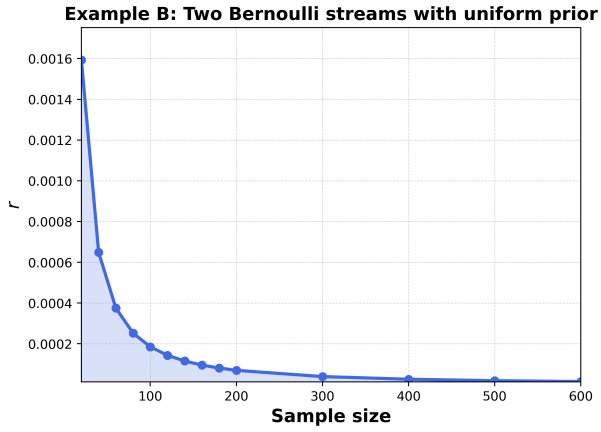

FIG. 3. Convergence of the interval width r for a canonical test between two streams of binary data (as in Example B), for $n^a = n^b = m$, as the sample size $n = 2m$ grows.

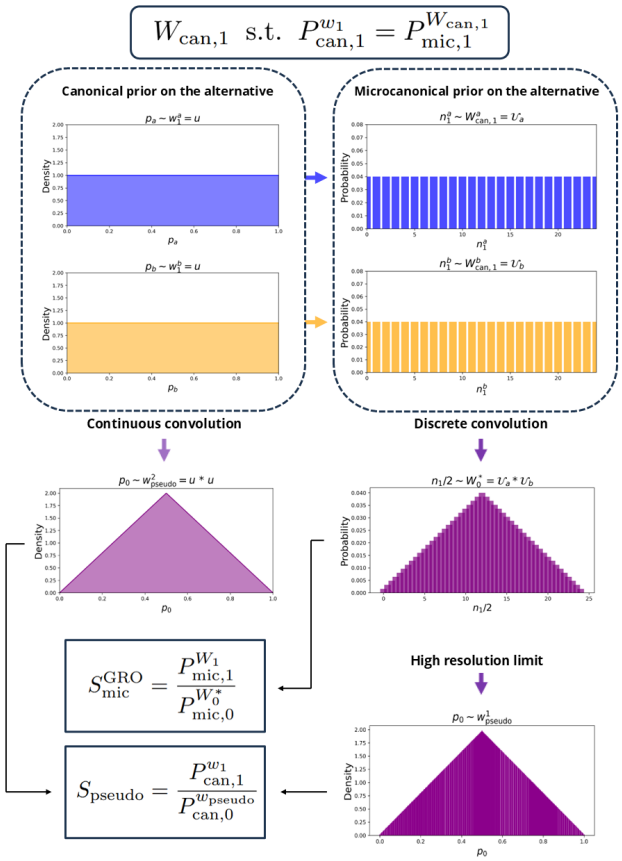

FIG. 4. Procedures to compute the microcanonical approximation $S_{\text{mic}}^{\text{GRO}}$ and the pseudo approximation S_{pseudo} for testing between two binary data streams, under uniform priors (as in Examples A, B, and C). Starting from two independent continuous uniform priors on the alternative (top left), we construct discrete microcanonical priors (top right) satisfying $P_{\text{can},1}^{w_1} = P_{\text{mic},1}^{W_{\text{can},1}}$. The optimal discrete prior W_0^*, used in $S_{\text{mic}}^{\text{GRO}}$, is obtained by convolving the alternative priors. The continuous prior $w_{\text{pseudo},0}$ for S_{pseudo} is derived either from W_0^* through a high resolution limit ($w_{\text{pseudo},0}^1$), or by directly convolving the original continuous priors ($w_{\text{pseudo},0}^2$).

C. Asymptotic justification for the microcanonical approximation

We now provide a theoretical result that explains why the microcanonical approximation tends to perform very well in practice, even when the canonical GRO e-variable is not available. Specifically, it suggests that in many cases the gap r defined in Equation (48) converges very fast to 0 as the sample size increases.

First, let $\mathcal{M}$ be a canonical maximum entropy model with sufficient statistic $\mathbf{c}(\mathbf{x})$ taking values in a finite set $\mathcal{C} \subset \mathbb{R}^d$. The canonical distribution has an exponential form

$$P_{\text{can}}(\mathbf{x};\boldsymbol{\theta}) = \frac{e^{-\boldsymbol{\theta}\cdot\mathbf{c}(\mathbf{x})}}{Z(\boldsymbol{\theta})}, \tag{54}$$

and induces the mean-value mapping

$$\boldsymbol{\mu}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{c}(\mathbf{x})], \tag{55}$$

with $\boldsymbol{\mu}$ taking values in the mean-value parameter space $\mathtt{M}$.

Next, consider the i.i.d. extension $\mathcal{M}^{(m)}$ in which $\mathbf{y}^{(m)} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ are m i.i.d. samples from $P_{\text{can}}(\cdot;\boldsymbol{\theta})$. The sufficient statistic of $\mathbf{y}^{(m)}$ is the sum

$$\mathbf{s}^{(m)}(\mathbf{y}^{(m)}) = \sum_{j=1}^{m} \mathbf{c}(\mathbf{x}_j). \tag{56}$$

Let w be a prior density over $\mathtt{M}$, and let Q^w be the induced probability for any measurable $\mathtt{M}' \subseteq \mathtt{M}$:

$$Q^w(\boldsymbol{\mu} \in \mathtt{M}') := \int_{\mathtt{M}'} w(\boldsymbol{\mu})\, d\boldsymbol{\mu}, \tag{57}$$

The following theorem shows that the normalized sufficient statistic converges in distribution to Q^w. Convergence is quantified in terms of probabilities of subsets, with error decaying at rate $O(\log m/m)$.

Theorem 1 *Let w be any regular prior density on the mean value parameter space $\mathtt{M} \subset \mathbb{R}^{\mathtt{d}}$. Then, for any INECCSI (Definition 1) subset $\mathtt{M}'$ of $\mathtt{M}$, we have*

$$\left| P_{\mathrm{can}}^{w^{(m)}} \left\{ \frac{\mathbf{s}^{(m)}(\mathbf{y}^{(m)})}{m} \in \mathtt{M}' \right\} - Q^w \left\{ \boldsymbol{\mu} \in \mathtt{M}' \right\} \right| = O\left(\frac{\log m}{m} \right). \quad (58)$$

In words: under the Bayesian marginal likelihood $P_{\mathrm{can}}^{w^{(m)}}$, the normalized sufficient statistic $\mathbf{s}^{(m)}/m$ becomes increasingly close to being distributed according to the prior over mean-value parameters.

Assume we can extend both canonical models $\mathcal{M}_0$ and $\mathcal{M}_1$ to i.i.d. models $\mathcal{M}_0^{(m)}$ and $\mathcal{M}_1^{(m)}$ as above, where (56) holds for both $\mathbf{c} = \mathbf{c}_0$ (sum denoted by $\mathbf{s}_0$) and $\mathbf{c} = \mathbf{c}_1$ (sum denoted by $\mathbf{s}_1$). In such settings, Theorem 1 supports the claim that the gap r between the microcanonical and pseudo approximations vanishes for large m. This is best explained in terms of our running example.

Example C (continued from Examples A and B)

Suppose $n^a = n^b = m$, so that data can be grouped into m i.i.d. *blocks*, each consisting of one binary outcome from group a and one from b. For each m the sufficient statistics are:

- $\mathbf{s}_1^{(m)} = (n_1^{a\,(m)}, n_1^{b\,(m)})$: number of ones in each group under the alternative,
- $\mathrm{s}_0^{(m)} = \frac{1}{2}\big(n_1^{a\,(m)} + n_1^{b\,(m)}\big)$: average number of ones across both groups under the null. The division by 2 is required to ensure that, for a single outcome, $\mathtt{M}_0 = [0,1]$, and can be interpreted, intuitively, as a set of probabilities.

After normalization by m, $\mathbf{s}_1^{(m)}/m \in \mathtt{M}_1 = [0,1]^2$ and $\mathrm{s}_0^{(m)}/m \in \mathtt{M}_0 = [0,1]$. Notice that every discrete distribution $W_j^{(m)}$ on the sufficient statistics of $\mathcal{M}_j^{(m)}$ induces a discrete distribution $V_j^{(m)}$ on the normalized sufficient statistics: $P_{\mathrm{can},1}^{w_1(m)}$ induces a probability $W_{\mathrm{can},1}^{(m)}$ on the alternative sufficient statistics $\mathbf{s}_1$, and a corresponding one, denoted here by $V_1^{(m)}$, on the normalized alternative sufficient statistics $\mathbf{s}_1/m$. Similarly, the microcanonical optimal prior $W_0^{*(m)}$ on the null sufficient statistic s_0 induces a distribution V_0^* on s_0/m.

In Example B, we used a prior w_1 under which p_a and p_b were independently and uniformly distributed, i.e., $w_1(p_a, p_b) = w_1^a(p_a) \cdot w_1^b(p_b)$ with $w_1^a = w_1^b = u$. The independent uniform prior has a remarkable property: $V_1^{(m)}$ coincides *exactly* with the product of two independent discrete uniforms, each defined on $\{0, 1/m, \ldots, 1\}$, corresponding to the components of $\mathbf{s}_1^{(m)}/m$. Consequently, the distribution $V_0^{*(m)}$ is *exactly* equal to a triangular discrete distribution, which is the convolution of these two discrete uniforms (Figure 2). Theorem 1 indicates that something analogous, but now in an asymptotic sense, will happen for every regular prior $w_1(p_a, p_b) = w_1^a(p_a) w_1^b(p_b)$, as long as p_a and p_b are still independent under w_1. More in detail, even if w_1^a and/or w_1^b are not uniform, $V_1^{(m)}$ will converge to a distribution on $\mathtt{M}_1$ that is a discretized version of w_1, and $V_0^{*(m)}$ will still be the exact convolution of the two components of $V_1^{(m)}$, which are the (approximate) discretized versions of the components of w_1. To illustrate, in Example 3 below, w_1^a and w_1^b will be taken to be of general beta form rather than restricted to uniform, and then we will see Theorem 1 in action, the correspondence becoming asymptotic rather than precise at each m. Still, $V_0^{*(m)}$ will converge, as m grows, to a continuous, strictly positive density on $\mathtt{M}_0$, denoted by $w^1_{\mathrm{pseudo},0}$. This distribution can be approximated by considering a very large m and "smoothening" the corresponding $V_0^{*(m)}$ to $w^1_{\mathrm{pseudo},0}$. This limiting procedure is precisely what we referred to earlier as the *high resolution limit*. Once $w^1_{\mathrm{pseudo},0}$ is obtained, $P_{\mathrm{can}}^{w^1_{\mathrm{pseudo},0}(m)}$ induces a probability distribution $W^{1(m)}_{\mathrm{pseudo},0}$ on the null sufficient statistics. As above, we can define

$$V^{1(m)}_{\mathrm{pseudo},0}\left(\frac{\mathbf{s}_0^{(m)}}{m} \right) := W^{1(m)}_{\mathrm{pseudo},0}\big(\mathbf{s}_0^{(m)}\big). \quad (59)$$

Now we invoke Theorem 1 again: it indicates that $V^{1(m)}_{\mathrm{pseudo},0}$ converges to $w^1_{\mathrm{pseudo},0}$. Thus, one may expect $V_0^{*(m)}$ and $V^{1(m)}_{\mathrm{pseudo},0}$, and consequently $W_0^{*(m)}$ and $W^{1(m)}_{\mathrm{pseudo},0}$, to be close and r to be small, according to Eq. (49). The microcanonical e-variable becomes thus an excellent approximation of the canonical one — which is what we set out to argue.

In the current example, we can go further: let $w^2_{\mathrm{pseudo},0}$ be the continuous convolution of the independent priors w_1^a and w_1^b. Applying Theorem 1 to $\mathcal{M}_0$ with this density shows that the induced distribution $V^{2(m)}_{\mathrm{pseudo},0}$ on $\mathbf{s}_0^{(m)}/m$ converges to a discretized version of $w^2_{\mathrm{pseudo},0}$. At the same time, $V_0^{*(m)}$, being the convolution of a discretized w_1, converges to the discretized convolution of w_1. Thus, for large m, $w^1_{\mathrm{pseudo},0}$ and $w^2_{\mathrm{pseudo},0}$ become indistinguishable. In practice, one can compute S_{pseudo} either by directly convolving the continuous components of w_1, or by taking the discrete convolution $W_0^{*(m)}$ and then its high-resolution limit: both approaches yield the same result (Figure 4).

How precise and general is this? In the reasoning above, we invoked Theorem 1 several times to go back and forth between prior distributions on mean-value parameters and marginal distributions on sufficient statistics. Specifically: (a) at the level of $\mathcal{M}_1$ (blue and yellow arrows in Figure 4); and (b) at the level of $\mathcal{M}_0$, for relating $W_0^{*(m)}$ to $P_{\text{can}}^{w^1_{\text{pseudo},0}}$ (b1, bottom right arrow in Figure 4) and $P_{\text{can}}^{w^2_{\text{pseudo},0}}$ (b2, bottom left arrow).

In step (a), the theorem is not really needed when w_1^a and w_1^b are uniform (as in the figure). Nevertheless, as long as the priors remain independent and regular, Theorem 1 suggests that step (a) holds even if they are not uniform. More generally, moving from the binary 2-group case to a general MEM $\mathcal{M}_1$, Theorem 1 still suggests that step (a) is valid whenever w_1 factorizes into independent regular priors, making the mean-value parameters independent. We write "suggest" rather than "prove" because the convergence in (58) is too weak to formally imply $r \to 0$ (it concerns probabilities of sets, whereas (49) involves expectations of log densities). Nevertheless, it provides strong heuristic evidence, and we do observe convergence numerically (see Figure 3). All reasoning based on Theorem 1 should thus be understood as heuristic rather than fully formal.

Turning now to step (b) for general $\mathcal{M}_0$ and $\mathcal{M}_1$: as long as $\mathbf{s}_0^{(m)}$ is a linear function of $\mathbf{s}_1^{(m)}$, the use of Theorem 1 in steps (b1) and (b2) remains heuristically justified, provided w_1 factorizes into regular independent priors as above. This linearity condition holds in all our examples (e.g. in Example C, $\mathbf{s}_0^{(m)}$ is the average of the components of $\mathbf{s}_1^{(m)}$). It guarantees that the limiting density $w^1_{\text{pseudo},0}$ exists, and Theorem 1 then suggests that it coincides with $w^2_{\text{pseudo},0}$.

In the more general case where $\mathbf{s}_0^{(m)}$ is a function (not necessarily linear) of $\mathbf{s}_1^{(m)}$ — i.e. Condition A holds — then it may still be true that $V_0^{*(m)}$ converges to a high-resolution limiting density $w^1_{\text{pseudo},0}$. In that case, Theorem 1 still suggests that step (b1) remains valid, so that r becomes small with growing sample size, making the microcanonical approximation effective. However, in such settings, it is less clear whether the approach based on $w^2_{\text{pseudo},0}$ still makes sense.

We stress that this asymptotic justification does not rely on $\bar{P}_1$ being a Bayesian mixture with prior w_1. The construction leading to $w^1_{\text{pseudo},0}$ applies to any universal distribution $\bar{P}_1$ on the alternative (including the NML), since $\bar{P}_1$ always induces a discrete distribution on the alternative sufficient statistic; from this, one can derive $V_0^{*(m)}$ and then obtain $w^1_{\text{pseudo},0}$ via the high-resolution limit. By contrast, the alternative route based on $w^2_{\text{pseudo},0}$ explicitly requires a factorized regular prior on the alternative, and is therefore not directly available in the non-Bayesian case.

IV. APPLICATION TO CONTINGENCY TABLES AND RELATED MODELS

Contingency tables are a fundamental tool in statistical analysis for examining the relationship between categorical variables. Given a dataset where observations are classified according to categorical factors, a contingency table provides a structured way to summarize the frequencies of different category combinations.

Formally, a contingency table is an $l \times k$ matrix where each entry represents the count of occurrences for a particular combination of row and column categories. Such tables are widely used in fields where categorical data naturally arise, such as biostatistics, social sciences, and market analysis.

In network science, this approach plays a crucial role in link analysis, where the presence or absence of an edge ($x = 1$ or $x = 0$) in a network is studied across different subsets of nodes. For instance, in community detection, one may ask whether the probability of forming a link differs within and between predefined groups of nodes. This idea is closely related to the Stochastic Block Model (SBM), a generative model in which nodes are assigned to latent groups, and connection probabilities are determined by group memberships. Contingency tables provide a natural way to summarize and test the differences in connection probabilities across groups, helping to assess whether observed patterns deviate from a null model where edges are formed independently of group structure. See e.g., [37, 38] for connections between network modeling, and contingency tables and the discussion in IV C of this paper.

In this work, we focus on binary categorical data, which corresponds to $l = 2$ in the general $l \times k$ contingency tables setting. We first apply our results to the simple case of two groups, i.e., 2×2 contingency tables. We consider microcanonical and canonical tests. For canonical tests, our main focus will be that of finding the microcanonical approximation in practical cases; this translates into finding the induced prior on the alternative (38) and then applying formula (41). We will finally verify the approximation validity by evaluating the interval width r, and show results on the regret. Later, we extend these results to the more general case of $2 \times k$ contingency tables.

A. 2×2 contingency tables

A 2×2 contingency table is a fundamental tool to assess whether the distribution of a binary outcome differs between two groups. Given a dataset where each observation consists of a binary variable $x \in \{0, 1\}$ and a categorical label indicating group membership, the data can be summarized in the following 2×2 table:

	Group A	Group B	Total
$x=1$	n_1^a	n_1^b	n_1
$x=0$	n_0^a	n_0^b	n_0
Total	n^a	n^b	n

The dataset $\mathbf{x}$ consists of two groups, represented as $\mathbf{x}_a = (x_1^a, ..., x_{n^a}^a)$ and $\mathbf{x}_b = (x_1^b, ..., x_{n^b}^b)$, where n^a and n^b are the respective group sizes. The table reports the number of ones (n_1^a and n_1^b) and zeros (n_0^a and n_0^b) in each group, along with their totals, n_1 and n_0. The key question is whether the probability of observing $x = 1$ differs between the two groups. This problem translates into a hypothesis testing problem, where:

- In the alternative hypothesis, the two groups are distinct, meaning the number of ones is constrained separately in each group:

$$\mathbf{c}_1 = (n_1^a, n_1^b). \tag{60}$$

- In the null hypothesis, the groups are indistinguishable, so only the total number of ones is constrained:

$$\mathrm{c}_0 = n_1. \tag{61}$$

These constraints define the sufficient statistics under each hypothesis and form the basis for the microcanonical and canonical tests discussed next. The reader may have noticed that this is exactly the setting of Examples A, B, and C in section III. Nevertheless, for the sake of clarity, in this section all quantities will be defined again and in more detail, at the cost of repeating ourselves.

$\mathbf{2 \times 2}$ microcanonical test

In the microcanonical formulation, we enforce hard constraints on the observed counts, treating them as fixed quantities. The null model with sufficient statistics n_1 reads

$$P_{\text{mic, }0}(\mathbf{x}; n_1) = \begin{cases} \frac{1}{\Omega_0(n_1)}, & \text{if } n_1(\mathbf{x}) = n_1; \\ 0, & \text{else;} \end{cases} \tag{62}$$

where

$$\Omega_0(n_1) = \binom{n}{n_1} \tag{63}$$

is the number of permutations of $\mathbf{x}$ preserving the total number of 1s. The alternative model with sufficient statistics (n_1^a, n_1^b) reads

$$P_{\text{mic, }1}(\mathbf{x}; n_1^a, n_1^b) = \begin{cases} \frac{1}{\Omega_1(n_1^a, n_1^b)}, & \text{if } (n_1^a(\mathbf{x}), n_1^b(\mathbf{x})) \\ & \quad = (n_1^a, n_1^b); \\ 0, & \text{else,} \end{cases} \tag{64}$$

where

$$\Omega_1(n_1^a, n_1^b) = \binom{n^a}{n_1^a}\binom{n^b}{n_1^b} \tag{65}$$

is the number of permutations of $\mathbf{x}$ preserving the total number of 1s in each group.

For any given prior W_1 on the alternative sufficient statistics, $S_{\text{mic}}^{\text{GRO}}$ is found exactly by computing W_0^* and applying (32). In this case, Condition A (30) is satisfied, as the null sufficient statistics can be written as a function of the alternative one:

$$n_1 = n_1^a + n_1^b. \tag{66}$$

Thus, following (31), the optimal prior on the null is the distribution of n_0 induced by $W_1(n_1^a, n_1^b)$. If n_1^a and n_1^b are independently distributed:

$$W_1(n_1^a, n_1^b) = W_1^a(n_1^a) \cdot W_1^b(n_1^b), \tag{67}$$

then W_0^* is simply the convolution of W_1^a and W_1^b:

$$W_0^* = W_1^a * W_1^b, \tag{68}$$

where $f * g$ represents the convolution between functions f and g.

Example 1: Microcanonical test with NML In the microcanonical case, resorting to the Normalized Maximum Likelihood approach is completely equivalent to putting a uniform prior on both parameters of the alternative model, as shown in [36]. Consequently, this case is reduced to Example A in III A, and is not considered further.

$\mathbf{2 \times 2}$ canonical test

The null canonical model obtained by constraining the average number of 1s, i.e., the expected value of n_1, is represented by the exponential distribution

$$P_{\text{can}}(\mathbf{x}; \theta_0) = \frac{e^{-\theta_0 \cdot n_1(\mathbf{x})}}{(1 + e^{-\theta_0})^n}, \tag{69}$$

which can be rewritten in the mean-value parametrization:

$$P_{\text{can}}(\mathbf{x}; p_0) = p_0^{n_1(\mathbf{x})}(1 - p_0)^{n - n_1(\mathbf{x})} \tag{70}$$

upon defining

$$p_0 = \frac{e^{-\theta_0}}{1 + e^{-\theta_0}}. \tag{71}$$

The null model is the distribution of a collection of n i.i.d. Bernoulli variables, where the occurrence of $x = 1$ has the same probability p_0 regardless of the group.

The alternative model, obtained by constraining the expected values of n_1^a and n_1^b, reads:

$$P_{\rm can}(\mathbf{x};\theta_a,\theta_b) = \frac{e^{-\theta_a\cdot n_1^a(\mathbf{x})-\theta_b\cdot n_1^b(\mathbf{x})}}{(1+e^{-\theta_a})^{n^a}(1+e^{-\theta_b})^{n^b}}, \tag{72}$$

or, equivalently,

$$\begin{aligned} P_{\rm can}(\mathbf{x};p_a,p_b) = {} & p_a^{n_1^a(\mathbf{x})}(1-p_a)^{n^a-n_1^a(\mathbf{x})} \\ & \times p_b^{n_1^b(\mathbf{x})}(1-p_b)^{n^b-n_1^b(\mathbf{x})} \end{aligned} \tag{73}$$

upon defining the mean-value parameters:

$$p_a = \frac{e^{-\theta_a}}{1+e^{-\theta_a}} \quad \text{and} \quad p_b = \frac{e^{-\theta_b}}{1+e^{-\theta_b}}. \tag{74}$$

The alternative model assumes that data in group A and group B are independent Bernoulli variables, where the probability of $x = 1$ is different according to the group. In this scenario, the aim of the test is to assess whether p_a and p_b are the same or whether they are different. In what follows, we explicitly apply the procedure described in section III B to different choices of $\bar{P}_{\rm can,1}$.

Example 2: Canonical test with NML. First, we focus on the test between two canonical models with an NML approach, i.e., $\bar{P}_{\rm can,1} = P_{\rm can,1}^{\rm NML}$. For a model with two independent Bernoulli distributions, the exact expression of the NML distributions reads [36, 39]:

$$P_{\rm can,1}^{\rm NML}(\mathbf{x}) = P_{\rm can,1}^{\rm a,\ NML}(\mathbf{x})\cdot P_{\rm can,1}^{\rm b,\ NML}(\mathbf{x}) \tag{75}$$

with

$$P_{\rm can,1}^{\rm i,\ NML}(\mathbf{x}) = \frac{\left(\frac{n_1^i(\mathbf{x})}{n^i}\right)^{n_1^i(\mathbf{x})}\left(1-\frac{n_1^i(\mathbf{x})}{n^i}\right)^{n^i-n_1^i(\mathbf{x})}}{\frac{e^{n^i}\Gamma(n^i,n^i)}{(n^i)^{n^i-1}}+1} \tag{76}$$

for $i \in \{a,b\}$. In the formula above, $\Gamma(s,t)$ is the upper incomplete gamma function. According to the procedure described in III B, a good candidate e-variable in this case is the microcanonical approximation, i.e., the GRO e-variable of the corresponding microcanonical test. To build it, we need the distribution of the alternative sufficient statistics (n_1^a, n_1^b) induced by $P_{\rm can,1}^{\rm i,\ NML}$, which is

$$W_{\rm can,1}(n_1^a,n_1^b) = W_{\rm can,1}^a(n_1^a)\cdot W_{\rm can,1}^b(n_1^b) \tag{77}$$

with

$$W_{\rm can,1}^a(n_1^a) = \Omega_1^a(n_1^a)\cdot\frac{\left(\frac{n_1^a(\mathbf{x})}{n^a}\right)^{n_1^a(\mathbf{x})}\left(1-\frac{n_1^a(\mathbf{x})}{n^a}\right)^{n^a-n_1^a(\mathbf{x})}}{\frac{e^{n^a}\Gamma(n^a,n^a)}{(n^a)^{n^a-1}}+1} \tag{78}$$

and

$$W_{\rm can,1}^b(n_1^b) = \Omega_1^b(n_1^b)\cdot\frac{\left(\frac{n_1^b(\mathbf{x})}{n^b}\right)^{n_1^b(\mathbf{x})}\left(1-\frac{n_1^b(\mathbf{x})}{n^b}\right)^{n^b-n_1^b(\mathbf{x})}}{\frac{e^{n^b}\Gamma(n^b,n^b)}{(n^b)^{n^b-1}}+1}. \tag{79}$$

Given that n_1^a and n_1^b are independently distributed, $W_0^*(n_1)$ is the convolution of $W_{\rm can,1}^a(n_1^a)$ and $W_{\rm can,1}^b(n_1^b)$, which can be computed numerically.

Example 3: Independent beta priors. The beta probability distribution reads:

$$\text{Beta}(y;\alpha,\beta) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)}, \quad \text{for } y\in(0,1), \tag{80}$$

where $B(\alpha,\beta)$ is the beta function, defined as

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt. \tag{81}$$

The beta prior represents a popular choice because it is flexible enough to encompass several cases of interest (Table S1 of the Supplementary Material). Here, we put two independent beta priors w_1^a and w_1^b with parameters, respectively, (α_a,β_a) and (α_b,β_b), on p_a and p_b. The Bayesian marginal likelihood resulting from this choice can be written explicitly as

$$\bar{P}_{\rm can,1}(\mathbf{x}) = \frac{B(\bar{\alpha}^a,\bar{\beta}^a)}{B(\alpha^a,\beta^a)}\cdot\frac{B(\bar{\alpha}^b,\bar{\beta}^b)}{B(\alpha^b,\beta^b)} \tag{82}$$

where

$$\begin{aligned} \bar{\alpha}^a &= n_1^a+\alpha^a, & \bar{\beta}^a &= n^a-n_1^a+\beta^a, \\ \bar{\alpha}^b &= n_1^b+\alpha^b, & \bar{\beta}^b &= n^b-n_1^b+\beta^b. \end{aligned}$$

As in the previous example, to obtain the microcanonical approximation for this problem, we look for the probability mass function induced by $\bar{P}_{\rm can,1}$ on the alternative sufficient statistics, which reads:

$$W_{\rm can,1}(n_1^a,n_1^b) = W_{\rm can,1}^a(n_1^a)\cdot W_{\rm can,1}^b(n_1^b) \tag{83}$$

with

$$W_{\rm can,1}^a(n_1^a) = \Omega_1^a(n_1^a)\cdot\frac{B(\bar{\alpha}^a,\bar{\beta}^a)}{B(\alpha^a,\beta^a)} \tag{84}$$

and

$$W_{\rm can,1}^b(n_1^b) = \Omega_1^b(n_1^b)\cdot\frac{B(\bar{\alpha}^b,\bar{\beta}^b)}{B(\alpha^b,\beta^b)}. \tag{85}$$

With this choice, W_1^a and W_1^b are *beta-binomial distributions*. Given that n_1^a and n_1^b are independently distributed, as we expected because we put independent priors on p_a and p_b, $W_0^*(n_1)$ is the convolution of $W_{\rm can,1}^a(n_1^a)$ and $W_{\rm can,1}^b(n_1^b)$. Whether this expression can

be written in closed form depends on the specific values of the beta parameters chosen. For example, if all beta parameters are equal to 1, W_0^* reduces to the convolution between two discrete uniform distributions (35). When no closed form is available, the convolution can be computed numerically.

In Figure S1 we show W_1^a, W_1^b, W_0^*, $w^1_{\text{pseudo},0}$ and $w^2_{\text{pseudo},0}$ for different choices of the beta parameters. As expected, in all cases where w_1^a and w_1^b are well defined in the whole parameter space, $w^1_{\text{pseudo},0}$ and $w^2_{\text{pseudo},0}$ are almost indistinguishable. This is a consequence of Theorem 1: the distribution of the mean value parameter ($w^2_{\text{pseudo},0}$) and that of the sufficient statistic ($w^1_{\text{pseudo},0}$) resemble each other when the sample size is big (high resolution limit).

In the next section, we show results obtained by numerical simulations for what concerns the optimality of the microcanonical approximation and the regret, measured in the examples reported in this section. For simplicity, when necessary, we will assume that the two groups have the same sample size, i.e., $n^a = n^b = m$, and that the independent beta priors on the alternative, denoted by w_1^a and w_1^b, have all parameters equal to a certain value $\gamma > 0$.

Evaluating the microcanonical approximation

In order to evaluate the goodness of the microcanonical approximation, we employ two approaches: a direct comparison and a comparison through r.

In the first case, we directly compare the e-power of the microcanonical approximation to the GRO-optimal canonical one, where the latter is computed by numerically solving the optimization problem (37). We find that the e-power of the microcanonical approximation converges to that of the canonical GRO e-variable as the total size grows (Figure S2). The e-power of the pseudo approximation converges as well, even though the convergence is slower compared to that of the microcanonical one. From these plots, we can already conclude that the microcanonical one works as a good approximation of $S^{\text{GRO}}_{\text{can}}$.

Notice that, if n_1^a and n_1^b are both big enough, using $P^{\text{NML}}_{\text{can},1}$ is asymptotically equivalent to using a Bayesian universal distribution with a *Jeffreys prior* [30] on the alternative parameters, which in our case is equivalent to a beta prior with parameters all equal to $\gamma = 0.5$. More precisely, let us, for simplicity, set $n^a = n^b = m$. Then, for any INECCSI subset $\mathtt{M}'_1 \subset \mathtt{M}_1$, as $m \to \infty$, with w_1 equal to the density of Jeffreys prior,

$$\sup_{\boldsymbol{\mu}_1 \in \mathtt{M}_1} \mathbb{E}_{\boldsymbol{\mu}_1}\left[-\log P^{w_1}_{\text{can},1}(\mathbf{x}^m) + \log P^{\text{NML}}_{\text{can},1}\right] = o(1), \quad (86)$$

where $o(1)$ denotes a quantity that goes to 0 as $m \to +\infty$. Nevertheless, a difference persists at the boundaries (outside $\mathtt{M}'_1$), where Jeffreys prior diverges and the NML induced priors do not. This difference becomes even more important when convoluting the independent Jeffreys priors to compute W_0^*. This explains why the first and third pictures in Figure S2 are quite different. For this reason, in all simulations, we implement the exact NML formula instead of its Jeffreys approximation.

The numerical approach to directly compare the e-power is feasible in a few simple cases and only for relatively small sample sizes. Conversely, the value of r can be easily evaluated, even for very large system sizes. In Figure S3, we show the plot of r as defined earlier to evaluate the effectiveness of the microcanonical approximation in different scenarios. The results confirm those of Figure S2, as in all cases considered, r converges to 0. In conclusion, we argue that the microcanonical approximation is a perfect candidate in this case.

Results on regret

Let's again consider the m-dimensional i.i.d. extension of our models. In section S6 of the Supplementary Materials, we show that, if the error $r'(\boldsymbol{\mu}_1)$ in (50) vanishes as the sample size m increases, both the canonical growth-optimal e-variable $S^{\text{GRO}}_{\text{can}}$ and its microcanonical approximation $S^{\text{GRO}}_{\text{mic}}$ satisfy:

$$\text{REG}_1(\boldsymbol{\mu}_1, \cdot) = \frac{d_1 - d_0}{2} \cdot \log m + O(1). \quad (87)$$

In the 2×2 case, where $d_1 = 2$ and $d_0 = 1$, this becomes:

$$\text{REG}_1(\boldsymbol{\mu}_1; \cdot) = \frac{1}{2} \log m + O(1).$$

This result holds uniformly over all $\boldsymbol{\mu}_1 \in \mathtt{M}'_1$, provided that $\mathtt{M}'_1$ is an INECCSI set (i.e., excluding regions near the boundary of the parameter space). However, it does not extend to the full parameter space $\mathtt{M}$, where the asymptotic form (19) may fail to hold even in well-specified cases.

Our experiments confirm these insights. We evaluated worst-case regret in the 2×2 setting for different values of the beta prior parameter γ. Notice that in what follows we apply our reasoning to the mean value parameter spaces, $(p_a, p_b) \in \mathtt{M}_1 = [0,1]^2$ and $p_0 \in \mathtt{M}_0 = [0,1]$, and that we consider INECCSI sets with respect to $\mathtt{M}_1$. From experimental results, collected in Figure 5, we observe a clear dichotomy:

- For $\gamma < 1$, the convolution of the beta priors w_1^a and w_1^b is non-differentiable at $p_0 = 1/2$, as shown in Figure S1 (e.g., for $\gamma = 0.5$). Consequently, the convergence of $V_0^{*(m)}$ to a density over the mean-value space $\mathtt{M}_0$ (as discussed under Theorem 1) may be very slow or fail altogether. In this case, S^{pseudo} becomes incomparable to $S^{\text{GRO}}_{\text{can}}$ and $S^{\text{GRO}}_{\text{mic}}$, and (87) no longer holds (see Figure S4). Indeed, in Figure 5, we see that even on small INECCSI sets, the regret grows like $a \log m + b$ for some $a > 1/2$.

- For $\gamma = 1$, convergence is moderate. Although r' decays quickly (Figure S4), the experimental values of Figure 5 shows areas (e.g., the yellow counter-diagonal) where regret exceeds the expected rate. These may still belong to an INECCSI set, but convergence has not yet been reached at the sample sizes considered ($m \leq 1800$).

- For $\gamma > 1$, the convolution is differentiable, and convergence of $V_0^{*(m)}$ is fast. The asymptotic behavior $(1/2)\log m + O(1)$ is observed on INECCSI sets (see again Figure 5)

These findings imply that, from a minimax perspective, using priors with $\gamma < 1$ is generally suboptimal. Such priors fail to achieve the expected regret rate of $(1/2)\log m + O(1)$ even when the true parameters lie well inside the parameter space.

This has implications for default prior choices. In both the Bayesian and MDL literatures, Jeffreys prior [30] is often recommended as a default when no prior knowledge is available, and is justified in the MDL framework because it achieves asymptotically minimax optimal redundancy (i.e., the middle inequality in (19) becomes an equality [40]). However, in our setting, Jeffreys prior corresponds to $\gamma = 1/2$, which, despite its MDL-optimality, is *not* optimal with respect to worst-case regret under e-values.

B. $2 \times k$ contingency tables

A $2 \times k$ contingency table is a natural extension of the 2×2 case, allowing us to assess whether the distribution of a binary outcome differs across multiple (k) groups. Given a dataset where each observation consists of a binary variable $x \in \{0, 1\}$ and a categorical label indicating group membership (among k different groups), the data can be summarized in the following $2 \times k$ table:

	Group 1	Group 2	$\cdots$	Group k	Total
$x = 1$	n_1^1	n_1^2	$\cdots$	n_1^k	n_1
$x = 0$	n_0^1	n_0^2	$\cdots$	n_0^k	n_0
Total	n^1	n^2	$\cdots$	n^k	n

The dataset consists of k groups, represented as $\mathbf{x}_i = (x_1^i, ..., x_{n^i}^i)$ for $i = 1, ..., k$, where n^i denotes the size of group i. The table reports the number of ones (n_1^i) and zeros (n_0^i) in each group, along with their respective totals, n_1 and n_0.

The key question remains whether the probability of observing $x = 1$ differs between groups. This problem again translates into a hypothesis testing problem, where:

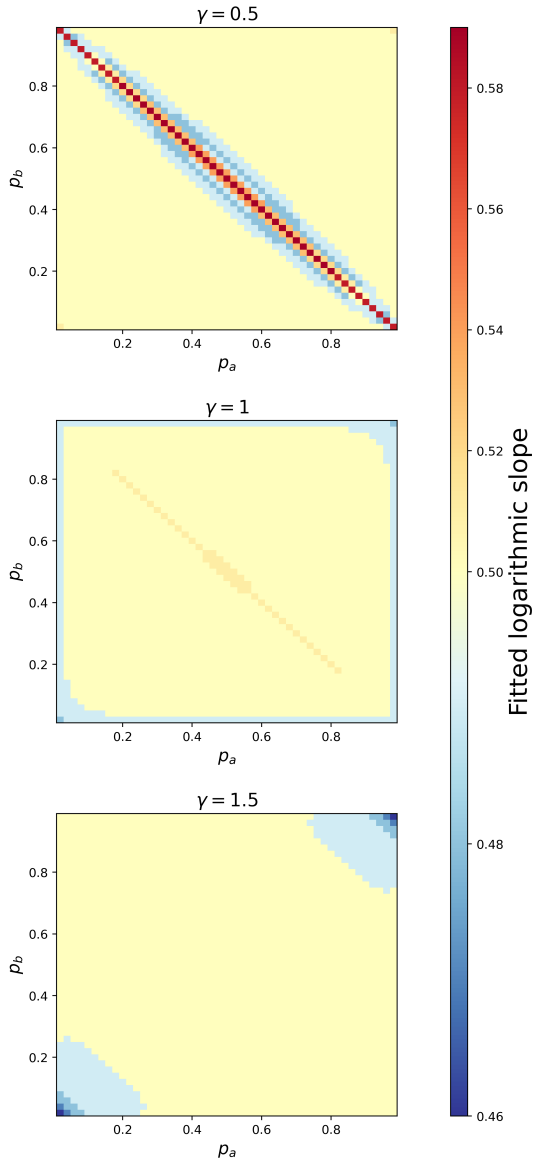

FIG. 5. Fitted slope of the logarithmic growth $a \log m + b$ of the microcanonical approximation regret 14 in the 2×2 case ($n^a = n^b = m$), shown for different combinations of the alternative parameters (p_a, p_b). The expected asymptotic slope is 0.5 (yellow). Three alternative beta priors are considered: $\alpha = \beta = 0.5$, $\alpha = \beta = 1$ and $\alpha = \beta = 1.5$. The sample sizes used for fitting are $m \in \{600, 800, 1000, 1200, 1400, 1600, 1800\}$. p_a and p_b vary in the interval $[0.02, 0.98]$, with a grid step of 0.02. Values at the boundaries are excluded to improve the readability of the plots.

- Under the alternative hypothesis, the groups are distinct, meaning the number of ones is constrained separately in each group:

$$\mathbf{c}_1 = (n_1^1, n_1^2, \ldots, n_1^k). \tag{88}$$

- Under the null hypothesis, the groups are indistinguishable, meaning only the total number of ones

is constrained:

$$\mathrm{c}_0 = n_1. \tag{89}$$

These constraints define the sufficient statistics under each hypothesis of the microcanonical and canonical tests discussed next. As the examples will illustrate, most of the results in this section naturally extend from the 2×2 case. The key distinction is that, in the latter case, the only relevant asymptotic behavior is as the total sample size n grows large. In contrast, in the present setting, both n and k can grow large, with different scenarios arising depending on the application (see IV C). In all cases, the asymptotic behavior of e-variables plays a crucial role, particularly in the canonical test, where only asymptotic approximations are available, and we need to assess whether the microcanonical approximation can be used.

$2 \times$ k microcanonical test

While the null model stays the same (Eq. (62), the alternative model is simply the extension of (64) from 2 to k groups:

$$P_{\text{mic, }1}(\mathbf{x}; \{n_1^i\}) = \begin{cases} \frac{1}{\Omega_1(\{n_1^i\}),} & \text{if } (n_1^1(\mathbf{x}), ..., n_1^k(\mathbf{x})) \\ & \quad = (n_1^1, ..., n_1^k), \\ 0, & \text{else,} \end{cases} \tag{90}$$

where

$$\Omega_1(\{n_1^i\}) = \prod_{i=1}^{k} \binom{n^i}{n_1^i}. \tag{91}$$

As in the 2×2 case, Condition A (30) is satisfied:

$$n_1 = \sum_{i=1}^{k} n_1^i \tag{92}$$

and the optimal prior on the null is the marginal distribution of n_0 induced by $W_1(\{n_1^i\})$. If all n_1^i are independently distributed:

$$W_1(\{n_1^i\}) = \prod_{i=1}^{k} W_1^i(n_1^i) \tag{93}$$

then W_0^* is simply the convolution of the individual alternative priors:

$$W_0^* = W_1^1 * ... * W_1^k. \tag{94}$$

Interestingly, when the number of groups k is large, and the priors are regular enough, a Central Limit Theorem holds; thus, W_0^* is well approximated by a *discrete Gaussian distribution*, i.e., if $k \gg 1$:

$$W_0^*(n_1) \approx \frac{1}{N(\mu_k, \sigma_k)} \exp\left(-\frac{(n_1 - \mu_k)^2}{2\sigma_k^2}\right) \tag{95}$$

where N is the normalization constant and

$$\mu_k = \sum_{i=1}^{k} \mathbb{E}_{W_1^i}[n_1^i]$$
$$\sigma_k^2 = \sum_{i=1}^{k} \mathrm{Var}_{W_1^i}(n_1^i).$$

This result is particularly convenient: when k is big enough, the only effect of the choice of priors on the alternative, as long as they are independent and regular enough, is in determining the average and the variance of the optimal (approximated) Gaussian prior on the null.

Example 4: Independent uniform priors. Here, we extend Example 1, i.e., Example A, to the case of k groups. When a uniform discrete prior $\mathcal{U}$ is put on each parameter of the alternative:

$$W_1(\{n_1^i\}) = \prod_{i=1}^{k} \mathcal{U}_i(n_1^i) = \prod_{i=1}^{k} \frac{1}{n^i + 1}, \tag{96}$$

the GRO null prior is again the convolution of all the individual priors, i.e., the convolution of k discrete uniform distributions, which reads [41]:

$$W_0^*(n_1) = \\ = \sum_{S \subseteq \{1, \dots k\}} (-1)^{|S|} \binom{n_1 + k - 1 - \sum_{j \in S}(n^j - n)}{n_1 - 1} \\ \times \left[\prod_{i=1}^{k} n^i + 1\right]^{-1}, \tag{97}$$

where the sum runs over all possible subsets of $\{1, \dots k\}$ and $\mid S \mid$ is the number of elements of set S. In the formula, the first factor stands for the number of ways in which a set of k non-negative numbers ($\{n_1^i\}$) can be chosen uniformly such that their sum is equal to n_1, with the constraint that for each i, n_1^i must be smaller than or equal to n^i. The second factor represents a normalization constant. If all n^i are equal to a certain value m, the formula simplifies and reads:

$$W_0^*(n_1) = \sum_{j=1}^{\lfloor n_1/(m+1) \rfloor} (-1)^j \binom{n}{j} \binom{n_1 - j(m+1) + k - 1}{k - 1} \\ \times \left[\prod_{i=1}^{k} n^i + 1\right]^{-1}. \tag{98}$$

where $\lfloor x \rfloor$ is the floor function of x. This is the formula used to generate Figure 6, where we show W_0^*, along with its Gaussian approximation, for increasing values of k.

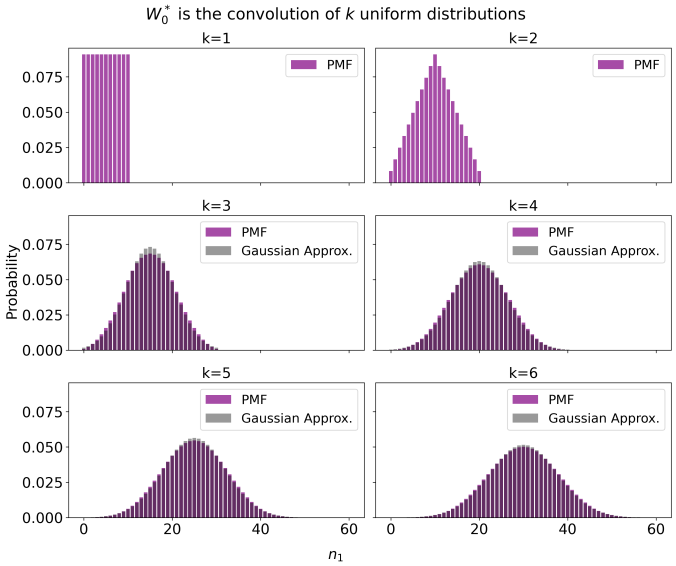

FIG. 6. The microcanonical GRO-optimal prior on the null W_0^* for testing $2 \times k$ tables is obtained as the convolution of the k independent priors on the alternative, which are discrete uniform priors in the case shown in this picture. Each convolution, for $k > 2$, is superposed to its discrete Gaussian approximation.

$2 \times$ k canonical test

The null canonical model is the same as in the 2×2 case, Eq. (69), i.e., a collection of n i.i.d. Bernoulli trials, where the probability of observing $x = 1$ is the same across all groups. The alternative model extends Eq. (72) and (73) to the case of k groups, by constraining the expected values of n_1^i separately for each group i, leading to the expression:

$$P_{\text{can}}(\mathbf{x};\boldsymbol{\theta}) = \frac{e^{-\sum_{i=1}^k \theta_i n_1^i(\mathbf{x})}}{\prod_{i=1}^k (1+e^{-\theta_i})^{n^i}}, \tag{99}$$

or, equivalently, in the mean-value parametrization:

$$P_{\text{can}}(\mathbf{x};\boldsymbol{p}) = \prod_{i=1}^k p_i^{n_1^i(\mathbf{x})}(1-p_i)^{n^i - n_1^i(\mathbf{x})}, \tag{100}$$

where we define the group-specific probabilities as:

$$p_i = \frac{e^{-\theta_i}}{1+e^{-\theta_i}}, \quad \text{for each } i \in \{1,\dots,k\}. \tag{101}$$

In this formulation, the alternative model assumes that data in each group are independent Bernoulli variables, where the probability of $x = 1$ depends on the group. The goal of the hypothesis test is to determine whether these probabilities are equal across all groups ($p_1 = p_2 = \cdots = p_k$) or whether they differ, indicating that the probability of observing $x = 1$ is group-dependent.

In the following sections, we apply the procedure described in section III B to different choices of $\bar{P}_{\text{can},1}$.

Example 5: Canonical test with NML. Extending results from Example 2, for a model with k independent Bernoulli distributions, the NML distribution reads [36]:

$$P_{\text{can},1}^{\text{NML}}(\mathbf{x}) = \prod_{i=1}^k P_{\text{can},1}^{\text{i, NML}}(\mathbf{x}) \tag{102}$$

where $P_{\text{can},1}^{\text{i, NML}}$ is that of Eq. (76). The microcanonical approximation is obtained by defining

$$W_{\text{can},1}(\{n_1^i\}) = \prod_{i=1}^k W_{\text{can},1}^i(n_1^i) \tag{103}$$

with

$$W_{\text{can},1}^i(n_1^i) = \Omega_1^i(n_1^i) \cdot \frac{\left(\frac{n_1^i(\mathbf{x})}{n^i}\right)^{n_1^i(\mathbf{x})} \left(1-\frac{n_1^i(\mathbf{x})}{n^i}\right)^{n^i - n_1^i(\mathbf{x})}}{\frac{e^{n^i}\Gamma(n^i,n^i)}{(n^i)^{n^i-1}} + 1}. \tag{104}$$

W_0^* is then the convolution of all $W_{\text{can},1}^i(n_1^i)$, which again can be computed numerically or by resorting to the Gaussian approximation (95).

Example 6: Independent beta priors. Here, we extend Example 3 to the case of k groups. We assume that each p_i is independently distributed according to a beta prior with parameters (α^i, β^i). The Bayesian marginal likelihood reads:

$$\bar{P}_{\text{can},1}(\mathbf{x}) = \prod_{i=1}^k \frac{B(\bar{\alpha}^i, \bar{\beta}^i)}{B(\alpha^i, \beta^i)} \tag{105}$$

where

$$\begin{aligned} \bar{\alpha}^i &= n_1^i + \alpha^i \\ \bar{\beta}^i &= n^i - n_1^i + \beta^i \quad \text{for each } i \in \{1,\dots,k\}. \end{aligned}$$

To derive the microcanonical approximation, we compute the probability mass function induced by $\bar{P}_{\text{can},1}(\mathbf{x})$ on $\{n_1^i\}$, which reads:

$$W_{\text{can},1}(\{n_1^i\}) = \prod_{i=1}^k W_{\text{can},1}^i(n_1^i) \tag{106}$$

with

$$W_{\text{can},1}^i(n_1^i) = \Omega_1^i(n_1^i) \cdot \frac{B(\bar{\alpha}^i, \bar{\beta}^i)}{B(\alpha^i, \beta^i)} \quad \text{for each } i \in \{1,\dots,k\}. \tag{107}$$

$W_0^*(n_1)$ is, then, the convolution of all $W_{\text{can},1}^i(n_1^i)$. If all beta parameters are equal to 1, W_0^* reduces to the convolution between k discrete uniform distributions (97). When the beta parameters are such that no closed form

is available, the convolution must be computed numerically. Alternatively, if k is big enough, one can resort to the discrete Gaussian approximation (95). Analogously, a continuous Gaussian approximation can be used to approximate w_{pseudo}. Figure (S5), we show $W^i_{\text{can},1}$, W^*_0, and w_{pseudo} for $k = 10$.

Evaluating the microcanonical approximation

To assess the effectiveness of the microcanonical approximation, we study the behavior of the interval width r in different cases. To simplify the problem, all beta priors considered in our results have parameters equal to the same number, γ, and all groups share the same size, i.e., $n_i = m$ for all $i \in \{1, \dots, k\}$. In this scenario, we have that $n = m \cdot k$. We consider three cases: m increases and k is fixed; n is fixed, and m and k change accordingly; finally, m and k grow together according to a certain law. We evaluate S_{pseudo}, and consequently r, by using $w^1_{\text{pseudo, }0}$, according to the procedure described in Example C, which is easily extended to the case of k groups. Our experiments (Figure S6) show that:

1. r converges quickly to 0 for fixed k as the m increases (or, equivalently, the total sample size increases);
2. r grows slowly for n fixed and k getting bigger;
3. r converges quickly to 0 whenever k and m grow together according to different power laws.

The only case where r does not converge to 0 corresponds to a decreasing m as $O(1/k)$. Our conclusion is that our microcanonical approximation $S^{\text{GRO}}_{\text{mic}}$ is an optimal candidate as long as m, i.e., the data size of each group, is big enough.

C. Connection to models of networks and time series

maximum entropy models are widely used to construct null models of complex systems that preserve specific structural or temporal features, while remaining otherwise random [14, 16–18, 42].

For instance, when applied to networks, maximum entropy models in their canonical formulations are known as *exponential random graph models* [18, 43, 44]. Examples of commonly used maximum entropy network models are the Erdős–Rényi model, Configuration Models, and Stochastic Block Models [16, 18]. The framework presented here is fully general and can be applied to build and compute e-values when testing between general maximum entropy network models with different sufficient statistics, in both their canonical and microcanonical formulations. Moreover, section II A establishes a link between e-values and the Minimum Description Length principle — a framework increasingly used in recent years for network inference and model selection [36, 45–47].

In particular, the hypothesis tests for contingency tables developed here have a direct correspondence with hypothesis tests between common network models. This mapping arises because the sufficient statistics in our contingency tables capture the same structural constraints as those imposed in standard network ensembles [16, 18]. Indeed, a binary network is represented by a binary adjacency matrix, which is a (structured) collection of 1s and 0s, corresponding to the presence or absence of a link between two nodes.

The null model considered here, in both its canonical and microcanonical formulation, corresponds to the well-known *Erdős–Rényi* model (ER), where the sufficient statistic is the total number of links, equal to (half, if the network is undirected) the total number of 1s observed in the adjacency matrix.

In the *Stochastic Block Model* (SBM), nodes are partitioned into groups and the adjacency matrix of a network is structured in k blocks, corresponding to the presence of inter- and intra-group links. For instance, in models of networks with community structure, intra-group link probabilities are larger than inter-group ones. The sufficient statistics are the number of links in each block. Testing an SBM against an ER model corresponds exactly to testing whether connection probabilities are identical across all blocks (i.e., communities are absent), and this SBM vs ER problem reduces to our canonical or microcanonical contingency table $2 \times k$ test.

In the *Partial Configuration Model* (PCM) for bipartite networks [48], the degree of each node in one layer is constrained, while connections to the other layer are otherwise random. The (bi-)adjacency matrix is a $k \times m$ rectangular binary matrix, and the sufficient statistics are the number of links connected to each node in the constrained layer, i.e., the number of 1s in each row. Testing a PCM against a bipartite ER model corresponds to testing whether all nodes in the constrained layer have the same connection probability (and therefore the same expected degree), i.e., testing for homogeneity of node properties in the graph. This again maps to a $2 \times k$ contingency table, where each constrained node represents a "group" and each group size equals the number m of nodes in the unconstrained layer.

Besides network models, a final connection worth mentioning is the one between binary contingency tables and multivariate time series data describing, e.g., a system of units being active (1) or inactive (0) at discrete time steps (such as spiking neurons data). The PCM can, in this case, represent a model enforcing, for each time step, a different activation probability of the various units. Therefore, testing the PCM against a bipartite ER model corresponds in this case to testing non-stationarity vs stationarity of the observed process over time.

We therefore conclude that our microcanonical e-variable for contingency tables can be directly applied to a wide range of problems, both exactly in the

microcanonical case and as an approximation for the canonical case. Moreover, our results on the behavior of r show that the microcanonical approximation works very well in both scenarios, as long as the size of each group is large enough. This circumstance is particularly convenient when studying models of large complex systems with a growing number of heterogeneous features, such as PCMs where the number of nodes in both layers can diverge in the "thermodynamic limit" of infinitely large graphs, SBMs used to model networks with a growing number of communities, and models of high-dimensional multivariate (nonstationary) time series. As we mentioned, the growing number of features (and parameters) in these models is generally needed to replicate the heterogeneous properties of real-world networks and time series more closely. At the same time, it makes the study of these models more challenging because of the breakdown of various useful approximations valid for a finite number of parameters —and even of the asymptotic equivalence between canonical and microcanonical versions of the resulting ensembles [36, 49]. Despite these complications, the results derived here nicely apply to those regimes.

V. CONCLUSION

In this work, we have developed a general framework for constructing optimal e-values for hypothesis testing between maximum entropy models with different constraints, in both microcanonical and canonical formulations. Our main theoretical contribution is the exact derivation of the microcanonical GRO e-variable and its use as a valid approximation to the canonical GRO e-variable when the latter is intractable. We provided analytical and numerical evidence that this approximation becomes asymptotically exact in many relevant regimes.

We illustrated our results through applications to 2×2 contingency tables, in both Bayesian and non-Bayesian (NML) settings, showing numerically that the microcanonical approximation provides a good proxy for the canonical solution, confirming our theoretical results. We then extended the analysis to general $2 \times k$ tables, where numerical results suggest that the microcanonical approximation works and remains asymptotically optimal for different interplays between k and the group sizes, as long as the latter are sufficiently large. Interestingly, when k becomes large, the microcanonical e-variable is itself well approximated by choosing a discrete Gaussian prior on the null. We highlighted that this framework can be naturally translated into network-science terms, where many important models can be derived as maximum entropy models.

A central role in our construction is played by universal distributions. These are the same distributions that underlie the Minimum Description Length (MDL) principle, where they achieve minimax redundancy. Our results show that such universal distributions (including Bayesian and NML ones) can be conveniently used to build GRO e-variables as well, thus providing a direct and convenient connection between description lengths and e-variables. A possible direction to explore in future work is to extend this connection beyond pairwise model comparisons and investigate how GRO e-variables and MDL can be combined to design tests involving multiple models at once.

[1] J. P. A. Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

[2] D. J. Benjamin et al. Redefine statistical significance. *Nature Human Behaviour*, 2(1):6–10, 2017.

[3] B. B. McShane, D. Gal, A. Gelman, C. Robert, and J. L. Tackett. Abandon statistical significance. *The American Statistician*, 73(sup1):235–245, 2019.

[4] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statist. Sci.*, 38(4):576–601, 2023.

[5] Aaditya Ramdas and Ruodu Wang. Hypothesis testing with e-values. *Foundations and Trends in Statistics*, 2025. To Appear.

[6] Yanbao Zhang, Scott Glancy, and Emanuel Knill. Asymptotically optimal data analysis for rejecting local realism. *Physical Review A*, 84(6):062118, 2011.

[7] P. Grünwald, R. de Heide, and W. Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128, 2024.

[8] L. Wasserman, A. Ramdas, and S. Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.

[9] V. Vovk and R. Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3), 2021.

[10] G. Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(2):407–431, 2021.

[11] Tyron Lardy, Peter Grünwald, and Peter Harremoës. Reverse information projections and optimal e-statistics. *IEEE Transactions on Information Theory*, 70(11):7616–7631, 2024.

[12] Martin Larsson, Aaditya Ramdas, and Johannes Ruf. The numeraire e-variable and reverse information projection. *Annals of Statistics*, 2025.

[13] J. Gibbs. *Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundation of Thermodynamics*. Cambridge Library Collection - Mathematics. Cambridge University Press, 2010.

[14] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, 1957.

[15] L.D. Brown. *Fundamentals of Statistical Exponential Families, with Applications in Statistical Decision Theory*. Institute of Mathematical Statistics, Hayward, CA,

1986.

[16] T. Squartini and D. Garlaschelli. *Maximum-Entropy Networks: Pattern Detection, Network Reconstruction and Graph Combinatorics*. Springer Cham, 2017.

[17] T. Squartini, G. Caldarelli, G. Cimini, A. Gabrielli, and D. Garlaschelli. Reconstruction methods for networks: The case of economic and financial systems. *Physics Reports*, 757, 2018.

[18] G. Cimini, T. Squartini, F. Saracco, D. Garlaschelli, A. Gabrielli, and G. Caldarelli. The statistical physics of real-world networks. *Nature Reviews Physics*, 1, 2019.

[19] R. Marcaccioli and G. Livan. Correspondence between temporal correlations in time series, inverse problems, and the spherical model. *Physical Review E*, 102(1):012112, 2020.

[20] R. Marcaccioli and G. Livan. Maximum entropy approach to multivariate time series randomization. *Scientific Reports*, 10(1):10656, 2020.

[21] Rosanne Turner and Peter Grünwald. Anytime-valid confidence intervals for contingency tables and beyond. *Statistics and Probability Letters*, 2023.

[22] Rosanne Turner, Alexander Ly, and Peter Grünwald. Generic e-variables for exact sequential k-sample tests that allow for optional stopping. *Statistical Planning and Inference*, 230:106116, 2024.

[23] Yunda Hao and Peter Grünwald. E-values for exponential families: the general case. *arXiv: 2409.11134*, 2024.

[24] Peter Grünwald, Tyron Lardy, Yunda Hao, Shaul K. Bar Lev, and Martijn de Jong. Optimal e-values for exponential families: the simple case. *arXiv: 2404.19465*, 2024.

[25] Peter D. Grünwald. Beyond neyman–pearson: E-values enable hypothesis testing with a data-driven alpha. *Proceedings of the National Academy of Sciences*, 121(39):e2302098121, 2024.

[26] Z. Zhang, A. Ramdas, and R. Wang. On the existence of powerful p-values and e-values for composite hypotheses. *arXiv: 2305.16539*, 2024.

[27] Q. Wang, R. Wang, and J. Ziegel. E-backtesting. *arXiv: 2209.00991*, 2024.

[28] V. Vovk and R. Wang. Efficiency of nonparametric e-tests. *arXiv: 2208.08925*, 2024.

[29] Richard D Morey, Jan-Willem Romeijn, and Jeffrey N Rouder. The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, 72(6-18):36, 2016.

[30] P. Grünwald. *The minimum description length principle*. MIT press, 2007.

[31] P. Grünwald and T. Roos. Minimum description length revisited. *International journal of mathematics for industry*, 11(01):1930001, 2019.

[32] K. Yamanishi. *Learning with the Minimum Description Length Principle*. Springer Nature Singapore, 2023.

[33] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, New York, 3rd edition, 2008.

[34] Kyoungseok Jang, Kwang-Sung Jun, Ilja Kuzborskij, and Francesco Orabona. Tighter pac-bayes bounds through coin-betting. In *Proceedings COLT 2023*, 2023.

[35] O.E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley, Chichester, UK, 1978.

[36] F. Giuffrida, T. Squartini, P. Grünwald, and D. Garlaschelli. Description length of canonical and microcanonical models. *Phys. Rev. Res.*, 2025.

[37] Hélder Alves, Paula Brito, and Pedro Campos. Community detection in interval-weighted networks. *Data Mining and Knowledge Discovery*, 38(2):653–698, 2024.

[38] Max Jerdee, Alec Kirkley, and Mark E. J. Newman. Mutual information and the encoding of contingency tables. *Physical review. E*, 110 6-1:064306, 2024.

[39] P. P. A. Staniczenko, M.J. Smith, and S. Allesina. Selecting food web models using normalized maximum likelihood. *Methods in Ecology and Evolution*, 5(6):551–562, 2014.

[40] B.S. Clarke and A.R. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41:37–60, 1994.

[41] M. Earnest (user). Extended stars-and-bars problem(where the upper limit of the variable is bounded). Mathematics Stack Exchange. URL:https://math.stackexchange.com/q/3182858 (version: 2019-04-14).

[42] A Golan, G Judge, and D Miller. *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. John Wiley), Chichester, UK, 1996.

[43] P. W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50, 1981.

[44] D. R. Hunter and M. S. Handcock. Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15(3):565–583, 2006.

[45] T. P. Peixoto. Nonparametric bayesian inference of the microcanonical stochastic block model. *Phys. Rev. E*, 95:012317, 2017.

[46] Tiago P. Peixoto. Network reconstruction via the minimum description length principle. *Phys. Rev. X*, 15:011065, Mar 2025.

[47] T. Vallès-Català, T. P. Peixoto, M. Sales-Pardo, and R. Guimerà. Consistencies and inconsistencies between model selection and link prediction in networks. *Phys. Rev. E*, 97:062316, 2018.

[48] Qi Zhang and Diego Garlaschelli. Strong ensemble nonequivalence in systems with local constraints. *New Journal of Physics*, 24(4):043011, 2022.

[49] Tiziano Squartini, Joey de Mol, Frank den Hollander, and Diego Garlaschelli. Breaking of ensemble equivalence in networks. *Physical review letters*, 115(26):268701, 2015.

[50] I. Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *The Annals of Probability*, pages 768–793, 1984.

[51] Peter Grünwald, Yunda Hao, and Akshay Balsubramani. Growth-optimal e-variables and an extension to the multivariate Csiszár-Sanov-Chernoff theorem. *arXiv: 2412.17554*, 2024.

Supplementary Materials

S1. REDUNDANCY AND REGRET

Here we show that, given the null and alternative models $\mathcal{M}_0 = \{P_0(\mathbf{x};\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0}$ and $\mathcal{M}_1 = \{P_1(\mathbf{x};\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_1}$, the regret (15) of the GRO e-variable (11), i.e.,

$$S^{\mathrm{GRO}}(\mathbf{x}) = \frac{\bar{P}_1(\mathbf{x})}{P_0^{w^*}(\mathbf{x})} \tag{S1}$$

for a given distribution $\bar{P}_1$, is bounded by the redundancy (17) of $\bar{P}_1$. Indeed, for any INECCSI (Def. 1) subset $\boldsymbol{\Theta}'$ of $\boldsymbol{\Theta}$:

$$\begin{aligned}
\mathrm{REG}(\boldsymbol{\Theta}'_{\mathbf{1}};\bar{P}_1) &= \max_{\boldsymbol{\theta}_1\in\boldsymbol{\Theta}'_1} \mathbb{E}_{\boldsymbol{\theta}_1}\left[\log S^{\mathrm{GRO}(\boldsymbol{\theta}_1)} - \log\frac{\bar{P}_1(\mathbf{x})}{P_0^{w_0^*}(\mathbf{x})}\right] \\
&= \max_{\boldsymbol{\theta}_1\in\boldsymbol{\Theta}'_1} \min_{w'_0\in\mathcal{W}_{\boldsymbol{\theta}_0}} \mathbb{E}_{\boldsymbol{\theta}_1}\left[\log\frac{P_1(\mathbf{x};\boldsymbol{\theta}_1)}{P_0^{w'_0}(\mathbf{x})} - \log\frac{\bar{P}_1(\mathbf{x})}{P_0^{w_0^*}(\mathbf{x})}\right] \\
&= \max_{\boldsymbol{\theta}_1\in\boldsymbol{\Theta}'_1}\left(\min_{w'_0\in\mathcal{W}_{\boldsymbol{\theta}_0}} \mathbb{E}_{\boldsymbol{\theta}_1}\left[\log\frac{P_0^{w_0^*}(\mathbf{x})}{P_0^{w'_0}(\mathbf{x})}\right] + \mathrm{RED}_1(\boldsymbol{\theta}_1;\bar{P}_1)\right) \\
&\le \mathrm{RED}_1(\boldsymbol{\Theta}'_1;\bar{P}_1).
\end{aligned} \tag{S2}$$

S2. MICROCANONICAL TEST

In this section, the subscript "mic" is omitted for the sake of clarity, as all the models considered are microcanonical models.

A. Exact solution of the optimization problem

We consider a test between a microcanonical alternative $\mathcal{M}_1$ and a microcanonical null $\mathcal{M}_0$. Given a universal microcanonical distribution $\bar{P}_1$ (either NML or Bayesian) on the alternative, the GRO-optimal microcanonical e-variable

$$S^{\mathrm{GRO}} = \frac{\bar{P}_1}{P_0^{W_0^*}} \tag{S3}$$

is found by solving the optimization problem

$$\begin{aligned}
W_0^* &= \arg\min_{W\in\mathcal{W}_{\mathbf{c}_0}} D_{\mathrm{KL}}(\bar{P}_1\|P_0^W) \\
&= \arg\min_{W\in\mathcal{W}_{\mathbf{c}_0}} \sum_{\mathbf{x}\in\mathcal{X}} \bar{P}_1(\mathbf{x})\log\frac{\bar{P}_1(\mathbf{x})}{\bar{P}_0^W(\mathbf{x})} \\
&= \arg\max_{W\in\mathcal{W}_{\mathbf{c}_0}} \sum_{\mathbf{x}\in\mathcal{X}} \bar{P}_1(\mathbf{x})\log\bar{P}_0^W(\mathbf{x}).
\end{aligned} \tag{S4}$$

For a microcanonical model with sufficient statistics $\mathbf{c}_i$, the Bayesian marginal likelihood reads [36]

$$P_i^{W_i}(\mathbf{x}) = \sum_{\mathbf{c}_i\in\mathcal{C}_i} P_i(\mathbf{x};\mathbf{c}_i)W_i(\mathbf{c}_i) = P_i(\mathbf{x};\mathbf{c}_i(\mathbf{x}))W_i(\mathbf{c}_i(\mathbf{x})) = \frac{W_i(\mathbf{c}_i(\mathbf{x}))}{\Omega_i(\mathbf{c}_i(\mathbf{x}))} \tag{S5}$$

where the latter equality is due to the definition of the microcanonical model, which assigns a positive probability only if $\mathbf{c}_i(\mathbf{x}) = \mathbf{c}_i$. By putting this result in (S4), one gets

$$W_0^* = \arg\max_{W\in\mathcal{W}_{\mathbf{c}_0}} \sum_{\mathbf{x}\in\mathcal{X}} \bar{P}_1(\mathbf{x})[\log W_0(\mathbf{c}_0(\mathbf{x})) - \log \Omega_0(\mathbf{c}_0(\mathbf{x}))] \tag{S6}$$
$$= \arg\max_{W\in\mathcal{W}_{\mathbf{c}_0}} \sum_{\mathbf{x}\in\mathcal{X}} \bar{P}_1(\mathbf{x}) \log W_0(\mathbf{c}_0(\mathbf{x})).$$

The latter expression can be written as a sum over the values of $\mathbf{c}_0$:

$$W_0^* = \arg\max_{W\in\mathcal{W}_{\mathbf{c}_0}} \sum_{\mathbf{c}_0\in\mathcal{C}_0} \left[\sum_{\mathbf{x}\,:\,\mathbf{c}(\mathbf{x})=\mathbf{c}_0} \bar{P}_1(\mathbf{x})\right] \log W(\mathbf{c}_0). \tag{S7}$$

According to Gibbs inequality, the distribution maximizing the quantity above is

$$W_0^*(\mathbf{c}_0) = \sum_{\mathbf{x}\,:\,\mathbf{c}_0(\mathbf{x})=\mathbf{c}_0} \bar{P}_1(\mathbf{x}) \tag{S8}$$

i.e., the marginal distribution of the null sufficient statistic $\mathbf{c}_0(\mathbf{x})$ induced by $\bar{P}_1$, hereby denoted, for simplicity, by $\bar{P}_1^{\mathbf{c}_0}$.

In the special case where the alternative sufficient statistics completely determine the value of the null one, we can write:

Condition A: (S9)
there exists a function $f : \mathcal{C}_1 \to \mathcal{C}_0$ s.t.$\mathbf{c}_0(\mathbf{x}) = f(\mathbf{c_1}(\mathbf{x}))$,

and thus

$$W_0^*(\mathbf{c}_0) = \sum_{\mathbf{x}\,:\,\mathbf{c}_0(\mathbf{x})=\mathbf{c}_0} \bar{P}_1(\mathbf{x}) \tag{S10}$$
$$= \sum_{\mathbf{x}\,:\,\mathbf{c}_0(\mathbf{x})=\mathbf{c}_0} \frac{W_1(\mathbf{c}_1(\mathbf{x}))}{\Omega_1(\mathbf{c}_1(\mathbf{x}))}$$
$$= \sum_{\mathbf{c}_1\,:\,f(\mathbf{c_1})=\mathbf{c}_0} \Omega_1(\mathbf{c}_1)\frac{W_1(\mathbf{c}_1)}{\Omega_1(\mathbf{c}_1)}$$
$$= \sum_{\mathbf{c}_1\,:\,f(\mathbf{c_1})=\mathbf{c}_0} W_1(\mathbf{c}_1).$$

In other words, the GRO-optimal prior on the null is the distribution induced on the null sufficient statistics by the alternative prior, or, equivalently, the marginal distribution of $\mathbf{c}_0$ induced by W_1, hereby denoted by $W_1^{\mathbf{c}_0}$.

Once that W_0^* is computed, given that all universal microcanonical distributions considered here are, in fact, Bayesian marginal likelihoods, the microcanonical GRO-optimal e-variable can be expressed as

$$S^{\mathrm{GRO}}(\mathbf{x}) = \frac{P_1^{W_1}(\mathbf{x})}{P_0^{W_0^*}(\mathbf{x})} \tag{S11}$$
$$= \frac{P_1(\mathbf{x};\mathbf{c}_1(\mathbf{x}))}{P_0(\mathbf{x};\mathbf{c}_0(\mathbf{x}))}\frac{W_1(\mathbf{c}_1(\mathbf{x}))}{W_0^*(\mathbf{c}_0(\mathbf{x}))}$$
$$= \frac{\Omega_0(\mathbf{c}_0(\mathbf{x}))}{\Omega_1(\mathbf{c}_1(\mathbf{x}))}\frac{W_1(\mathbf{c}_1(\mathbf{x}))}{W_0^*(\mathbf{c}_0(\mathbf{x}))}.$$

B. The average under the null of the GRO e-variable is always unitary

Here, we show that $\mathbb{E}_0[\,S^{\mathrm{GRO}}\,] = 1 \quad \forall P_0 \in \mathcal{M}_0$.

We start by picking a generic distribution inside the null model:

$$P_0(\mathbf{x};\bar{\mathbf{c}}_0) = \begin{cases} \dfrac{1}{\Omega_0(\bar{\mathbf{c}}_0)} & \text{if } \mathbf{c}_0(\mathbf{x}) = \bar{\mathbf{c}}_0 \\ 0 & \text{else.} \end{cases} \tag{S12}$$

Then, we compute the average under $P_0(\mathbf{x};\bar{\mathbf{c}}_0)$ of S^{GRO}:

$$\mathbb{E}_0[S^{\mathrm{GRO}}] = \mathbb{E}_0\left[\frac{\bar{P}_1}{P_0^{W_0^*}}\right] = \sum_{\mathbf{x}\in\mathcal{X}} P_0(\mathbf{x};\bar{\mathbf{c}}_0)\frac{\bar{P}_1(\mathbf{x})}{P_0^{W_0^*}(\mathbf{x})} \tag{S13}$$

$$= \frac{1}{\Omega_0(\bar{\mathbf{c}}_0)} \sum_{\mathbf{x}\,:\,\mathbf{c}_0(\mathbf{x})=\bar{\mathbf{c}}_0} \frac{\bar{P}_1(\mathbf{x})}{P_0^{W_0^*}(\mathbf{x})} \tag{S14}$$

In what follows, we use the explicit expression of the Bayesian marginal likelihood (S5), i.e., $P_0^{W_0}(\mathbf{x}) = \frac{W_0(\mathbf{c}_0(\mathbf{x}))}{\Omega_0(\mathbf{c}_0(\mathbf{x}))}$, and that of the GRO optimal prior (S8)

$$\mathbb{E}_0[\,S^{\mathrm{GRO}}\,] = \frac{1}{\Omega_0(\bar{\mathbf{c}}_0)} \sum_{\mathbf{x}\,:\,\mathbf{c}_0(\mathbf{x})=\bar{\mathbf{c}}_0} \frac{\Omega_0(\mathbf{c}_0(\mathbf{x}))\bar{P}_1(\mathbf{x})}{W_0^*(\mathbf{c}_0(\mathbf{x}))} \tag{S15}$$

$$= \frac{1}{\Omega_0(\bar{\mathbf{c}}_0)}\frac{\Omega_0(\bar{\mathbf{c}}_0)}{W_0^*(\bar{\mathbf{c}}_0)} \sum_{\mathbf{x}\,:\,\mathbf{c}_0(\mathbf{x})=\bar{\mathbf{c}}_0} \bar{P}_1(\mathbf{x}) = 1.$$

S3. MICROCANONICAL APPROXIMATION

A. Every microcanonical e-variable is a canonical e-variable

Given a sufficient statistics $\mathbf{c}(\mathbf{x})$, the microcanonical probability distribution can be expressed as a conditional canonical one:

$$P_{\mathrm{mic}}(\mathbf{x};\mathbf{c}) = P_{\mathrm{can}}(\mathbf{x};\boldsymbol{\theta} \mid \mathbf{c}). \tag{S16}$$

where the probability of $\mathbf{x}$ is conditioned on a certain value of $\mathbf{c}(\mathbf{x})$. According to the law of total expectation, for every random variable $S(\mathbf{x})$ defined on $\mathcal{X}$:

$$\mathbb{E}_{\mathrm{can}}[\,S\,] = \mathbb{E}_{\mathrm{can},}\,[\,\mathbb{E}_{\mathrm{can}}[\,S \mid \mathbf{c}\,]\,] = \mathbb{E}_{\mathrm{can}}\,[\,\mathbb{E}_{\mathrm{mic}}[\,S\,]\,]\,. \tag{S17}$$

It follows that if E_{mic} is a microcanonical e-variable, it is also a canonical one:

$$\mathbb{E}_{\mathrm{mic}}[\,E_{\mathrm{mic}}\,] \le 1 \quad \forall P_{\mathrm{mic}} \in \mathcal{M}_{\mathrm{mic}} \quad \Rightarrow \quad \mathbb{E}_{\mathrm{can}}[\,E_{\mathrm{mic}}\,] \le 1 \quad \forall P_{\mathrm{can}} \in \mathcal{M}_{\mathrm{can}} \tag{S18}$$

Moreover, as proven in the last section, it holds:

$$\mathbb{E}_{\mathrm{mic}}[\,S_{\mathrm{mic}}^{\mathrm{GRO}}\,] = 1. \tag{S19}$$

Thus, from (S17), it follows that

$$\mathbb{E}_{\mathrm{can}}[\,S_{\mathrm{mic}}^{\mathrm{GRO}}\,] = 1. \tag{S20}$$

B. A canonical universal distribution can always be expressed as microcanonical Bayesian marginal likelihood

In what follows, we show that:

1. Given a canonical universal distribution $\bar{P}_{\mathrm{can}}$ relative to a sufficient statistic $\mathbf{c}$, we can always define a prior distribution $W_{\mathrm{can}}(\mathbf{c})$ on the sufficient statistic such that

$$\bar{P}_{\mathrm{can}} = \bar{P}_{\mathrm{mic}}^{W_{\mathrm{can}}}. \tag{S21}$$

2. The opposite is not true: for some $\bar{P}_{\mathrm{mic}}^{W}$, there is no choice of prior density $w(\boldsymbol{\theta})$ such that $\bar{P}_{\mathrm{mic}}^{W} = \bar{P}_{\mathrm{can}}^{w}$.

Proof 1

By construction, a canonical universal distribution $\bar{P}_{\text{can}}(\mathbf{x})$ relative to the sufficient statistics $\mathbf{c}$ always assigns the same probability mass to configurations sharing the same value of the sufficient statistic, just as the corresponding $P_{\text{can}}(\mathbf{x};\boldsymbol{\theta})$ does. Consequently, the following holds:

$$\bar{P}_{\text{can}}(\mathbf{x}\,|\,\mathbf{c}) = P_{\text{mic}}(\mathbf{x};\mathbf{c}), \tag{S22}$$

i.e., when conditioned on the sufficient statistic, the universal canonical distribution reduces to a uniform distribution over the number of configurations corresponding to the observed value, which is the microcanonical distribution. Moreover, we can always write

$$\bar{P}_{\text{can}}(\mathbf{x}) = \bar{P}_{\text{can}}(\mathbf{x}\mid\mathbf{c}(\mathbf{x}))\bar{P}^{\mathbf{c}}_{\text{can}}(\mathbf{c}(\mathbf{x})) = P_{\text{mic}}(\mathbf{x};\mathbf{c}(\mathbf{x}))\bar{P}^{\mathbf{c}}_{\text{can}}(\mathbf{c}(\mathbf{x})) \tag{S23}$$

where $\bar{P}^{\mathbf{c}}_{\text{can}}(\mathbf{c})$ is the distribution of $\mathbf{c}$ induced by $\bar{P}_{\text{can}}$. The proof follows by comparing the expression above with the general expression of the microcanonical Bayesian marginal likelihood (S5) and by setting $W_{\text{can}}(\mathbf{c}) = \bar{P}^{\mathbf{c}}_{\text{can}}(\mathbf{c})$.

Proof 2

The canonical Bayesian marginal likelihood $\bar{P}^{w}_{can}$

$$P^{w}_{\text{can}}(\mathbf{x}) = \int_{\boldsymbol{\Theta}} P_{\text{can}}(\mathbf{x};\boldsymbol{\theta})w(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{S24}$$

is the weighted sum of positive functions; indeed, $P_{\text{can}}(\mathbf{x};\boldsymbol{\theta})$ is an exponential function that assigns positive probability to all $\mathbf{x}\in\mathcal{X}$. As such, for all proper choices of the prior density $w(\boldsymbol{\theta})$, $\bar{P}^{w}_{\text{can}}$ is strictly positive everywhere:

$$\bar{P}^{w}_{can}(x) > 0 \quad \forall x \in \mathcal{X}. \tag{S25}$$

Consider a microcanonical prior $W(\mathbf{c})$ such that $W(\mathbf{c}^*) = 0$ for a certain value $\mathbf{c}^*$ of the sufficient statistics. Consequently,

$$\bar{P}^{W}_{\text{mic}}(x^*) = 0 \quad \forall\, x : \mathbf{c}(\mathbf{x}) = \mathbf{c}^*. \tag{S26}$$

Thus, there is no choice of canonical prior $w(\boldsymbol{\theta})$ s.t. $\bar{P}^{W}_{\text{mic}} = \bar{P}^{w}_{\text{can}}$.

S4. PROOF OF THEOREM 1

In the proof, we freely use well-known properties of exponential families as described by, e.g., [35]. Set $B := \mathtt{M}'$. Fix a constant a and consider the sets, for $m = 1, 2, \ldots$:

$$B^{+}_{m} = \left\{\boldsymbol{\mu} : \inf_{\boldsymbol{\mu}'\in B} \|\boldsymbol{\mu}-\boldsymbol{\mu}'\|_2^2 \le \frac{a\log m}{m}\right\}, B^{-}_{m} = \left\{\boldsymbol{\mu}\in B : \inf_{\boldsymbol{\mu}'\notin B} \|\boldsymbol{\mu}-\boldsymbol{\mu}'\|_2^2 \ge \frac{a\log m}{m}\right\}.$$

B^{+}_{m} is a superset of B, including a small region (whose volume tends to 0 with sample size) just outside B's boundary; similarly B^{-}_{m} excludes a small region just inside B's boundary. To shorten notation we write $\hat{\boldsymbol{\mu}} := \mathbf{s}(\mathbf{y}^{(m)})/m$ and $P^{w} := P^{w\,(m)}_{\text{can}}$ and, for any measurable subset $B'\subset\mathtt{M}$, $Q^{w}(B') := Q^{w}(\boldsymbol{\mu}\in B')$, and $P_{\boldsymbol{\mu}} := P_{\text{can}}(\cdot;\boldsymbol{\theta}(\boldsymbol{\mu}))$ where $\boldsymbol{\theta}(\boldsymbol{\mu})$ is the mapping from mean-value parameters to corresponding canonical parameters, i.e. the inverse of the (1-to-1) mapping $\boldsymbol{\mu}(\boldsymbol{\theta})$ defined in the main text. We have:

$$\begin{aligned}
P^{w}(\hat{\boldsymbol{\mu}}\in B) &= \int_{\boldsymbol{\mu}\in\mathtt{M}} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}}\in B)w(\boldsymbol{\mu})d\boldsymbol{\mu} \ge \int_{\boldsymbol{\mu}\in B^{-}_{m}} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}}\in B)w(\boldsymbol{\mu})d\boldsymbol{\mu} \\
&\ge Q^{w}(B^{-}_{m}) \inf_{\boldsymbol{\mu}\in B^{-}_{m}} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}}\in B) \ge Q^{w}(B^{-}_{m}) \inf_{\boldsymbol{\mu}\in B^{-}_{m}} (1 - P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}}\notin B)) \\
&\ge Q^{w}(B^{-}_{m}) - \sup_{\boldsymbol{\mu}\in B^{-}_{m}} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}}\notin B)Q^{w}(B) + O\left(\frac{\log m}{m}\right) - \sup_{\boldsymbol{\mu}\in B^{-}_{m}} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}}\notin B).
\end{aligned} \tag{S27}$$

and also

$$P^w(\hat{\boldsymbol{\mu}} \in B) = \int_{\boldsymbol{\mu} \in B_m^+} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in B) w(\boldsymbol{\mu}) d\boldsymbol{\mu} + \int_{\boldsymbol{\mu} \notin B_m^+} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in B) w(\boldsymbol{\mu}) d\boldsymbol{\mu}$$
$$\leq Q^w(B_m^+) + \sup_{\boldsymbol{\mu} \in \mathtt{M}, \boldsymbol{\mu} \notin B_m^+} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in B) = Q^w(B) + O\left(\frac{\log m}{m}\right) + \sup_{\boldsymbol{\mu} \in \mathtt{M} \setminus B_m^+} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in B). \tag{S28}$$

We will now further bound the supremum terms in the above two formulas, showing that they are both of order $O(m^{-a})$. The result then follows by plugging in $a = 1$.

a. Supremum in (S28) To bound the supremum in (S28), note first that B is a convex set. We can therefore use Csiszár's [50] multivariate generalization of Chernoff's concentration inequality (see [51] for general discussion) to get that

$$\sup_{\boldsymbol{\mu} \in \mathtt{M} \setminus B_m^+} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in B) \leq \sup_{\boldsymbol{\mu} \in \mathtt{M} \setminus B_m^+, \boldsymbol{\mu}' \in B} e^{-m D_{\mathrm{KL}}(P_{\boldsymbol{\mu}'} \| P_{\boldsymbol{\mu}})} = e^{-m \inf_{\boldsymbol{\mu} \in \mathtt{M} \setminus B_m^+, \boldsymbol{\mu}' \in B} D_{\mathrm{KL}}(P_{\boldsymbol{\mu}'} \| P_{\boldsymbol{\mu}})} \leq e^{-m \inf_{\boldsymbol{\mu} \in \partial B_m^+, \boldsymbol{\mu}' \in B} D_{\mathrm{KL}}(P_{\boldsymbol{\mu}'} \| P_{\boldsymbol{\mu}})}, \tag{S29}$$

where $D_{\mathrm{KL}}(P_{\boldsymbol{\mu}'} \| P_{\boldsymbol{\mu}})$ is the Kullback-Leibler divergence between $P_{\boldsymbol{\mu}'}$ and $P_{\boldsymbol{\mu}}$ defined at a single outcome, and in the final inequality we used the standard fact that the KL divergence between members of an exponential family is strictly convex in its first argument.

Now since B is an INECCSI subset of $\mathtt{M}$, clearly there exists another INECCSI subset of $\mathtt{M}$, say $\bar{B}$, and a finite m_0 such that $B_m^+ \subset \bar{B}$ for all $m \geq m_0$. Since $\bar{B}$ is INECCSI, there exist c, C with $0 < c < C < \infty$ such that all eigenvalues of the Fisher information matrix $I(\boldsymbol{\mu})$ in the mean-value parameterization are in between c and C for all $\boldsymbol{\mu} \in \bar{B}$. This means that the KL divergence between any $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \bar{B}$ satisfies

$$(1/2)c^k \frac{a \log m}{m} \leq \frac{1}{2} c^k \|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_2^2 \leq D_{\mathrm{KL}}(P_{\boldsymbol{\mu}'} \| P_{\boldsymbol{\mu}}) \leq \frac{1}{2} C^k \|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_2^2 \leq (1/2) C^k \frac{a \log m}{m}. \tag{S30}$$

The result now follows by plugging in the lower bound on the KL divergence implied by the above into (S29) and then setting $a = 1$ and plugging further into (S28).

b. Supremum in (S27) To bound the supremum in (S27), fix any $\boldsymbol{\mu} \in B_m^-$ and let $R_{m,\boldsymbol{\mu}}$ be a hyper-rectangle centered at $\boldsymbol{\mu}$ that is a subset of B and that has side-length $2\epsilon_m$. By construction there is $c' > 0$ such that, for all m, we can take $\epsilon_m = c' \cdot \sqrt{(a \log m / m)}$. Noting that we can write $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_d)$, with d the dimensionality of both the canonical space $\boldsymbol{\Theta}$ and the mean-value space $\mathtt{M}$, we set $H^{\geq}_{\boldsymbol{\mu},j,\epsilon} := \{\boldsymbol{\mu}' \in \mathtt{M} : \boldsymbol{\mu}'_j \geq \boldsymbol{\mu}_j + \epsilon\}$ and $H^{\leq}_{\boldsymbol{\mu},j,\epsilon} := \{\boldsymbol{\mu}' \in \mathtt{M} : \boldsymbol{\mu}'_j \leq \boldsymbol{\mu}_j - \epsilon\}$. We have:

$$\sup_{\boldsymbol{\mu} \in B_m^-} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \notin B) \leq \sup_{\boldsymbol{\mu} \in B_m^-} P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \notin R_{m,\boldsymbol{\mu}}) \leq \sum_{j=1}^{d} \left(P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in H^{\leq}_{\boldsymbol{\mu},j,\epsilon_m}) + P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in H^{\geq}_{\boldsymbol{\mu},j,\epsilon_m}) \right)$$
$$\leq \sum_{j=1}^{d} \left(e^{-m \inf_{\boldsymbol{\mu}' \in H^{\leq}_{\boldsymbol{\mu},j,\epsilon_m}} D_{\mathrm{KL}}(P_{\boldsymbol{\mu}'} \| P_{\boldsymbol{\mu}}) \|)} + e^{-m \inf_{\boldsymbol{\mu}' \in H^{\geq}_{\boldsymbol{\mu},j,\epsilon_m}} D_{\mathrm{KL}}(P_{\boldsymbol{\mu}'} \| P_{\boldsymbol{\mu}}) \|)} \right) = O\left(m^{-a}\right). \tag{S31}$$

Here the second inequality is the union bound and the third is once again Csiszár's [50] multivariate generalization of Chernoff's concentration inequality (in (S29), we bounded $P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in B)$ where B was a convex set, allowing us to use Csiszár's result directly; but in (S31), we need to bound $P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \notin B) = P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in \mathtt{M} \setminus B)$; since $\mathtt{M} \setminus B$ is not a convex set, yet convexity is required by Csiszŕ's result, we now first need to cover it by $2d$ convex sets $H^{\cdot}_{\boldsymbol{\mu},\cdot,\epsilon_m}$, for each of which we then use Csiszár's result). The final inequality follows by using (S30) again.

S5. PSEUDO APPROXIMATION THROUGH THE HIGH RESOLUTION LIMIT

Below, we provide pseudocode to compute the density $w^1_{\text{pseudo},0}$ for a canonical test on 2×2 contingency tables, under the alternative hypothesis with independent beta priors on the mean-value parameters. Notice that for $\alpha_a = \beta_a = \alpha_b = \beta_b = 1$, we retrieve the uniform prior of Example B.

The procedure starts from the induced distribution on the sufficient statistics under the alternative, denoted $W^a_{\text{can},1}$ and $W^b_{\text{can},1}$ in the main text, which follow a beta-binomial form. To approximate the continuous limit, we increase the resolution by multiplying the original group sizes by a scaling factor (Steps 1–2). The higher the scaling factor, the better the approximation — at the cost of greater computational effort.

The two high-resolution distributions are then convolved to obtain an approximation of the GRO-optimal prior on the null, W_0^* (Step 3). Since we want a density over $p_0 \in [0, 1]$, we define a pseudo-continuous support for p_0 accordingly (Step 4). Finally, the convolved distribution is normalized to form a proper density over this support (Step 5).

```
Input:
    n_a, n_b               // group sizes
    scaling_factor         // resolution multiplier
    alpha_a, beta_a        // beta-binomial parameters for group a
    alpha_b, beta_b        // beta-binomial parameters for group b

Step 1: Define high-resolution support
    x_high_a = 0 to scaling_factor * n_a
    x_high_b = 0 to scaling_factor * n_b

Step 2: Compute high-resolution beta-binomial PMFs
    For each i in x_high_a:
        p_high_a[i] = BetaBinomialPMF(i, scaling_factor * n_a, alpha_a, beta_a)
    For each i in x_high_b:
        p_high_b[i] = BetaBinomialPMF(i, scaling_factor * n_b, alpha_b, beta_b)

Step 3: Convolve the two distributions
    conv_high = Convolve(p_high_a, p_high_b)

Step 4: Define pseudo-continuous support over [0, 1]
    p_0_fine = [0, 1, ..., len(conv_high)-1] / (scaling_factor * (n_a + n_b))

Step 5: Normalize the convolved distribution
    step = p_0_fine[1] - p_0_fine[0]
    conv_high = conv_high / (sum(conv_high) * step)

Output:
    conv_high   // approximate density over [0, 1]
    p_fine      // corresponding support
```

S6. BOUND ON REGRET

Here we provide a bound for the regret of a maximum entropy model (15).

Let $\mathcal{M}_0$ be a maximum entropy model. We begin by recalling an extension of the classical $(d/2)\log m$ asymptotic redundancy result (see Equation (19)) that remains valid even when the model $\mathcal{M}_0$ is misspecified — i.e., it does not contain the true distribution P^*. This generalization appears in [30], and is formalized in [23, Proposition 3].

Let $\mathbf{x}^m$ be an i.i.d. sample from a distribution P^* that may lie outside $\mathcal{M}_0$. Suppose that there exists a distribution $P_{\tilde{\boldsymbol{\theta}}_0} \in \mathcal{M}_0$ minimizing the Kullback-Leibler divergence to P^*:

$$P_{\tilde{\boldsymbol{\theta}}_0} = \arg\min_{\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0} D_{\mathrm{KL}}(P^* \| P_{\boldsymbol{\theta}_0}).$$

Then, for any regular prior density w_0, we have:

$$\begin{aligned}\mathrm{RED}_0(P^*; P_0^{w_0}) &:= \mathbb{E}_{P^*}\left[-\log P_0^{w_0}(\mathbf{x}^m) + \log P_{\tilde{\boldsymbol{\theta}}_0}(\mathbf{x}^m)\right] \\ &= \frac{d_0}{2}\log m + O(1). \quad \text{(S32)}\end{aligned}$$

In the well-specified case where $P^* \in \mathcal{M}_0$, it holds that $P^* = P_{\tilde{\boldsymbol{\theta}}_0}$, and $\mathrm{RED}_0(P^*; P_1^{w_0})$ coincides with our earlier definition $\mathrm{RED}_0(\tilde{\boldsymbol{\theta}}_0, P_1^{w_0})$ (up to notation), thus recovering the classical result (19).

We now apply this general result in the setting where $P^* = P_{\boldsymbol{\theta}_1} \in \mathcal{M}_1$, in order to bound the regret $\mathrm{REG}(\boldsymbol{\theta}_1; P_1^{w_1})$ of the Bayesian marginal $P_1^{w_1}$ for a regular prior w_1. Consider the pseudo-e-variable S_{pseudo}, used as a proxy for either $S_{\mathrm{mic}}^{\mathrm{GRO}}$ or $S_{\mathrm{can}}^{\mathrm{GRO}}$. Using the previous result to refine equation (22), we obtain:

$$\begin{aligned}\mathrm{REG}(\boldsymbol{\theta}_1; S_{\mathrm{pseudo}}) &= \mathbb{E}_{\boldsymbol{\theta}_1}\left[\log S^{\mathrm{GRO}(\boldsymbol{\theta}_1)} - \log \frac{P_1^{w_1}(\mathbf{x})}{P_0^{w_{\mathrm{pseudo},0}}(\mathbf{x})}\right] \\ &= \mathbb{E}_{\boldsymbol{\theta}_1}\left[\log \frac{P_0^{w_{\mathrm{pseudo},0}}(\mathbf{x})}{P_0^{\tilde{w}_0'}(\mathbf{x})}\right] + \mathrm{RED}_1(\boldsymbol{\theta}_1; P_1^{w_1}) \\ &= \mathbb{E}_{\boldsymbol{\theta}_1}\left[\log \frac{P_0^{w_{\mathrm{pseudo},0}}(\mathbf{x})}{P_{\tilde{\boldsymbol{\theta}}_0}(\mathbf{x})}\right] + \mathrm{RED}_1(\boldsymbol{\theta}_1; P_1^{w_1}) \\ &= \mathrm{RED}_1(\boldsymbol{\theta}_1; P_1^{w_1}) - \mathrm{RED}_0(P_{\boldsymbol{\theta}_1}; P_0^{w_{\mathrm{pseudo},\,0}}) + O(1) \\ &\overset{(a)}{=} \frac{d_1 - d_0}{2} \cdot \log m + O(1). \quad \text{(S33)}\end{aligned}$$

The first two equalities are direct, and the third follows from [23, Theorem 3 (Parts 3,4)]. Equality (a) holds provided that $w_{\mathrm{pseudo},0}$ is regular, in particular, that it has a continuous density.

S7. SUPPLEMENTARY TABLES AND FIGURES

Case	Parameter Condition	Behavior
Uniform	$\alpha = \beta = 1$	Flat distribution
Jeffreys Prior	$\alpha = \beta = 0.5$	U-shaped
Bimodal (U-shaped)	$\alpha, \beta < 1$	Peaks at 0 and 1
Left-skewed	$\alpha > 1, \beta < 1$	Peak near 1
Right-skewed	$\alpha < 1, \beta > 1$	Peak near 0
Bell-shaped	$\alpha, \beta > 1$	Normal-like
Highly concentrated	$\alpha = \beta \gg 1$	Sharp peak
Degenerate (Dirac Delta)	$\alpha, \beta \to \infty$	Point mass at $\frac{\alpha}{\alpha+\beta}$

TABLE S1. Behaviors of the beta distribution for different parameter values.

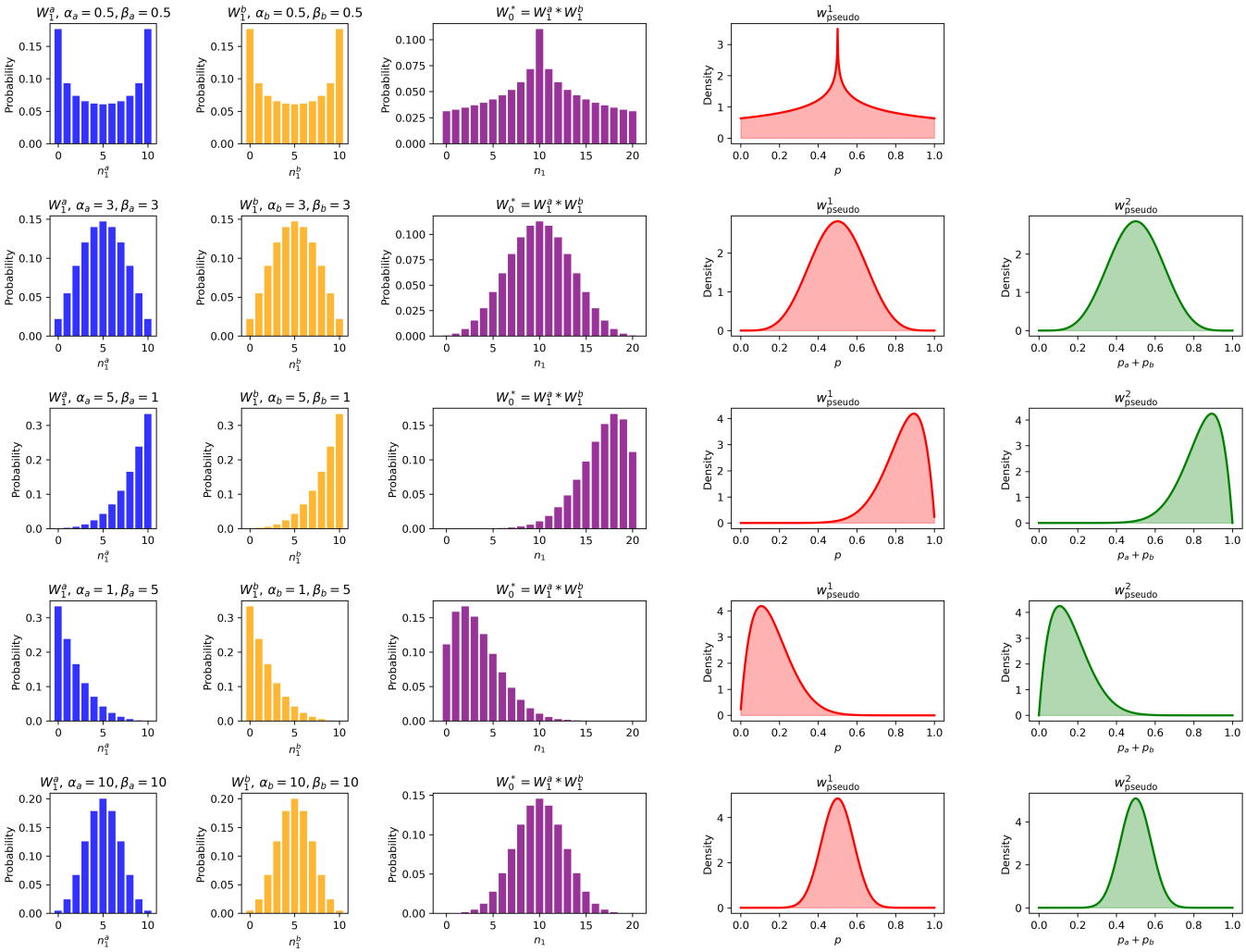

FIG. S1. Construction of the microcanonical optimal prior W_0^* (third column) and the pseudo-prior approximations $w^1_{\text{pseudo},0}$ (fourth column) and $w^2_{\text{pseudo},0}$ (fifth column), for 2×2 contingency tables with independent beta priors on the alternative. Starting from the induced independent beta-binomial distribution on the alternative sufficient statistics (first and second columns), the microcanonical GRO-optimal prior W_0^* is obtained as their convolution. The pseudo prior approximation $w^1_{\text{pseudo},0}$ is obtained starting from W_0^* through a high-resolution limit. The pseudo prior approximation $w^2_{\text{pseudo},0}$ is obtained by directly convoluting the original continuous beta priors, when well defined on the whole parameter space (i.e., for $\gamma \geq 1$).

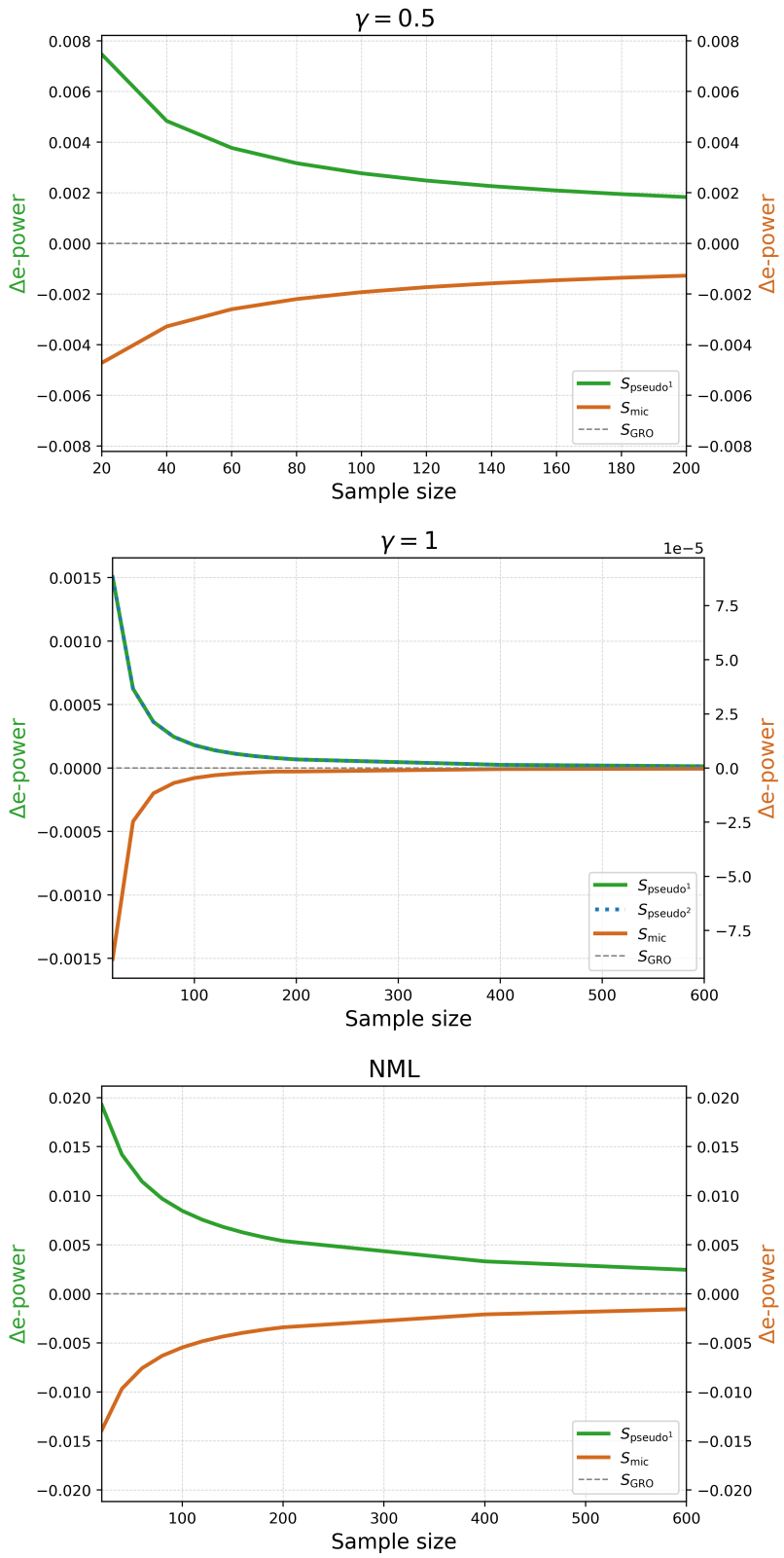

FIG. S2. E-power difference between the canonical GRO e-variable (computed numerically), its microcanonical approximation (orange curve), and the pseudo approximation (green curve), across sample sizes and different choices on the alternative (NML and beta with all parameters equal to γ). The microcanonical and pseudo approximations provide a lower and upper bound for the canonical GRO e-power, converging to it as the sample size grows. Results are shown for the 2×2 contingency tables canonical test with $n^a = n^b = m$ and sample size equal to $2m$.

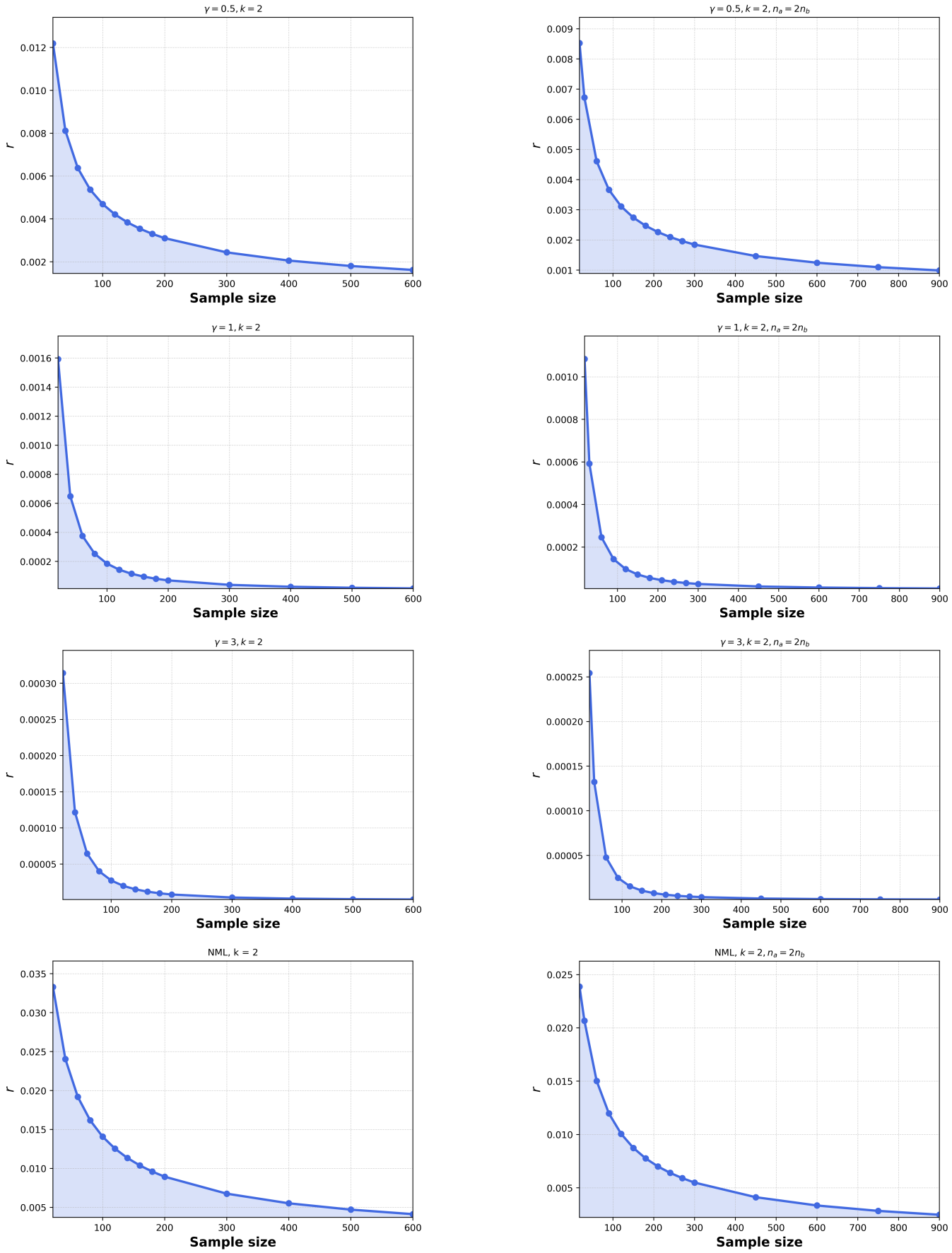

FIG. S3. Convergence of r in 2×2 contingency tables, for $n^a = n^b$ (left column) and $n^a = 2n^b$ (right column), as the sample size $n = n^a + n^b$ grows. Results are shown for different choices of $\bar{P}_{\text{can},1}$: beta independent priors with all parameters equal to $\gamma = 0.5$ (first row), 1 (second row), 3 (third row), and NML (fourth row).

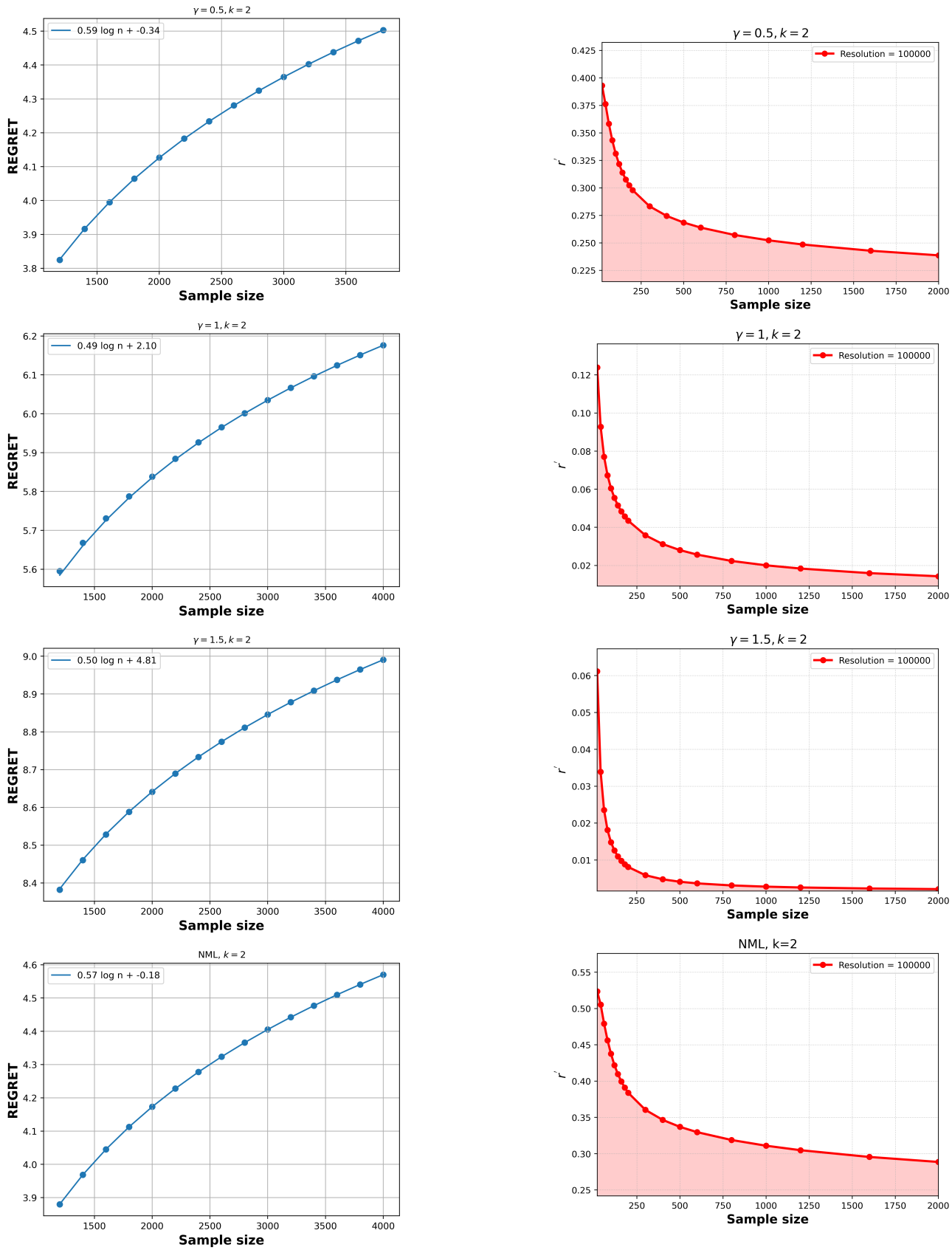

FIG. S4. Worst case regret (left) and convergence of r' (right), for 2×2 tables with $n^a = n^b = m$ and sample size equal to $n = 2m$. Different choices of the alternative are considered: Bayesian with identical independent beta priors $B(\gamma, \gamma)$ with $\gamma = 0.5$, 1, 1.5, and NML. r' is computed by considering $w^1_{\text{pseudo},0}$, obtained through a high resolution limit with resolution scale equal to 100000.

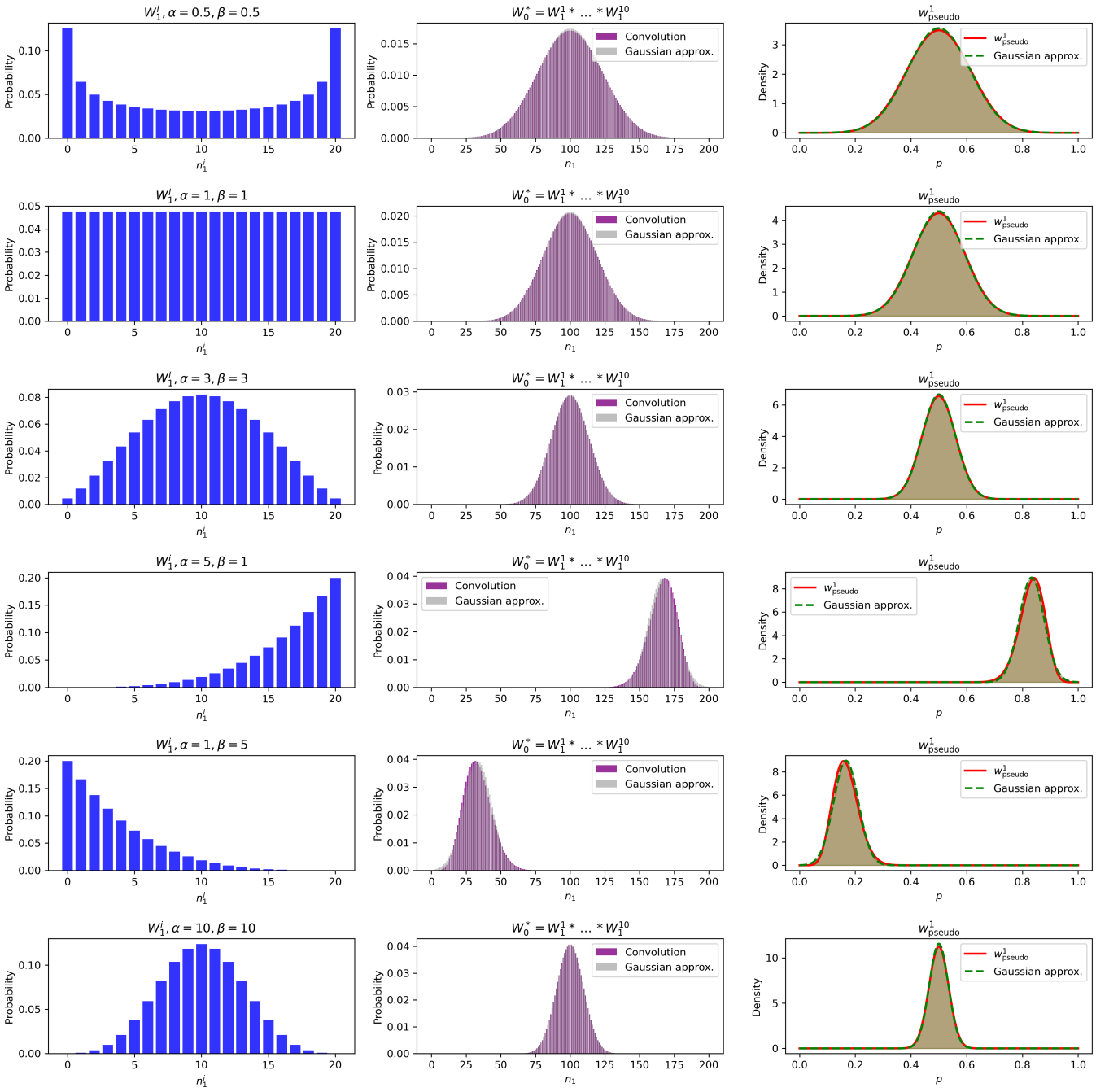

FIG. S5. The first column shows the discrete distribution of each component of the alternative sufficient statistics, here denoted simply by W_1^i, induced by independent identical beta priors on the alternative, for different prior parameters. The GRO-optimal microcanonical prior on the null W_0^* (second column) for the $2 \times k$ test is obtained by convolving these discrete distributions k times. The pseudo prior density w^1_{pseudo} (third column) is instead obtained by directly convolving k times the corresponding beta priors. Both W_0^* and w^1_{pseudo} are shown together with their Gaussian approximations (discrete for W_0^* and continuous for w^1_{pseudo}). In this example, $k = 10$.

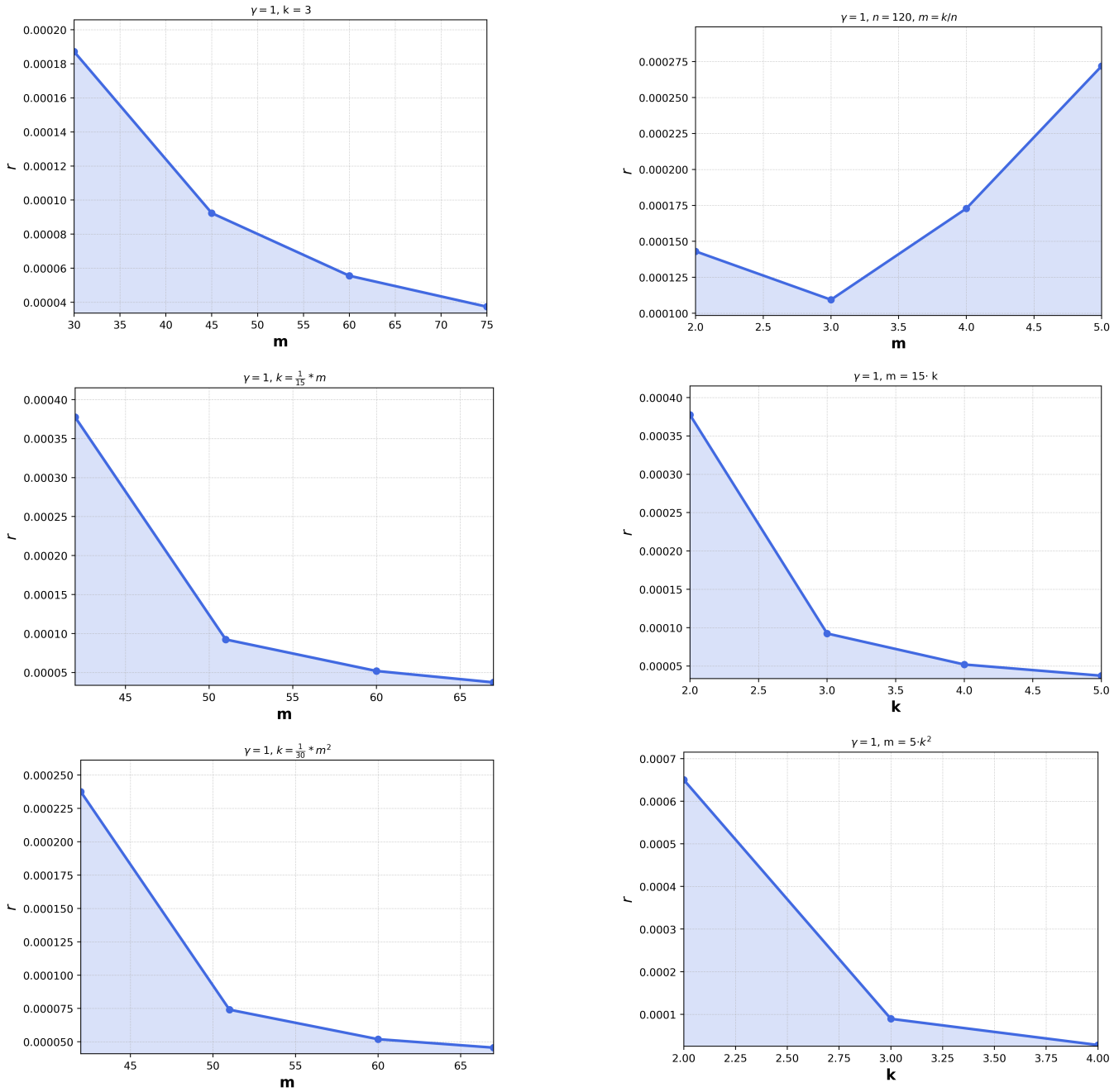

FIG. S6. Convergence to 0 of the interval width r in the case of $2 \times k$ contingency tables, where all groups have same size m, for different interplays between the number of groups k and the size of each group m. Results are shown for identical independent beta priors on the alternative, with all parameters equal to $\gamma = 1$.