Testing maximum entropy models with *e*-values

Francesca Giuffrida [1,2,*] Diego Garlaschelli [1,2] and Peter Grünwald [3,4]
[1] *IMT School for Advanced Studies, Lucca, Italy*
[2] *Lorentz Institute for Theoretical Physics (LION), Leiden University, Leiden, The Netherlands*
[3] *Centrum Wiskunde & Informatica, Amsterdam, The Netherlands*
[4] *Mathematical Institute, Leiden University, Leiden, The Netherlands*

(Received 3 October 2025; accepted 16 April 2026; published 15 May 2026)

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, in both 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 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.

DOI: 10.1103/xhf5-1l7p

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 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 *optimale*-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

*Contact author: francesca.giuffrida@imtlucca.it

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 (nonstationary) 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 subcase of 2×2 tables (to reappear as Examples A–C in our paper), 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 Sec. II we introduce *e*-variables and growth-rate optimality. In Sec. III we address the problem of finding optimal *e*-variables for testing between two maximum entropy models, either microcanonical (Sec. III A) or canonical (Sec. III B), that differ in their sufficient statistics. We introduce a method to construct optimal *e*-variables. 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 Sec. 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 (Sec. IV A) and then generalize to $2 \times k$ (Sec. IV B). In both cases, we provide fully worked-out examples, allowing for exact calculations of the corresponding *e*-variables and a detailed comparison between microcanonical and canonical constructions. We show that, in all scenarios considered, 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:

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

The realized value of E evaluated on data $\mathbf{x}$ is called an *e-value*. Unlike *p*-values, *largere*-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]:

(1) *Type I error control.* The condition $\mathbb{E}_0[E] \leqslant 1$ 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 $0 \leqslant \alpha \leqslant 1$, by Markov's inequality, we have

$$P_0(E \geqslant 1/\alpha) \leqslant \alpha, \tag{4}$$

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.

(2) *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.

(3) *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.

We emphasize that *post-hoc* error control and Type I error control under optional continuation *cannot* be achieved with *p*-value-based and other classical hypothesis testing methods—in particular, classical meta-analyses in the medical and the social sciences come without any error guarantees. Regarding the latter, [26] contains an *e*-value-based reanalysis of two of 28 classic psychological findings considered in the *ManyLabs2 Project* [27]. The goal of this project was to assess the replicability of the original results. Each of the 28 psychological studies was repeated in many labs across the world, each repetition (called *replication attempt*) involving newly acquired, expensive experimental data. [26] found that, for the two findings considered, if *e*-values had been used, correct and statistically valid conclusions could have been drawn based

on a much smaller number of replication attempts than had actually been done, saving enormous amounts of time and money. For an overview of other practical applications of e-values (there are many, besides meta-analysis), we refer to [4,5].

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 *optimale*-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

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

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 $\boldsymbol{\Theta}_j \subset \mathbb{R}^d$ ($d > 0$), assuming they are continuous and strictly positive for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_j$. 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* [28–30] 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* [31], i.e., the ratio between two pieces of Bayesian evidence:

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

The fact that the solution to (6) looks like (7) is not at all obvious—it is the central result of two breakthrough papers, [7,12]. 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

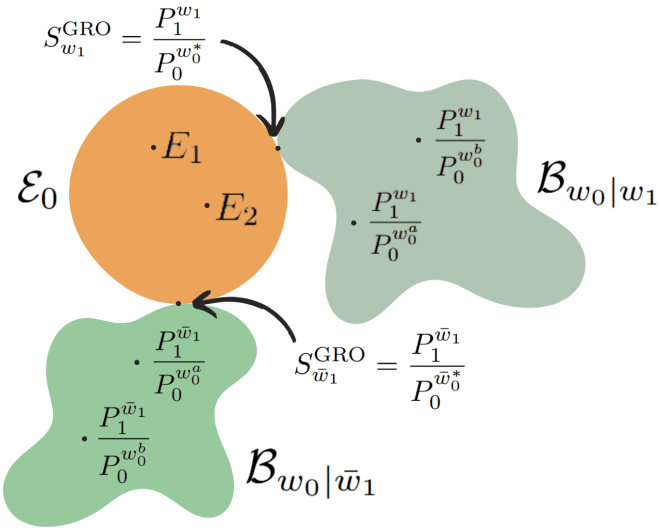

FIG. 1. The GRO e-variable S^{GRO} is the unique intersection between the set $\mathcal{B}_{w_0|w_1}$ of all Bayes factors for a given $P_1^{w_1}$ and varying $P_0^{w_0}$, and the set $\mathcal{E}_0$ of all e-variables relative to model $\mathcal{M}_0$. At the same time, it is the unique e-variable maximizing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

hypothesis. Specifically, w_0^* is the solution to the following optimization problem:

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

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*:

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

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 Fig. 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

[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.

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 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

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

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 the following optimality principle: among all *e*-variables, one would seek the *minimax optimale*-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* [32–34]. 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 MDL 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 [32,35], the code length to describe $\mathbf{x}$, using a code that compresses optimally in expectation under Q, is (up to rounding) $-\log Q(\mathbf{x})$ nats (log denoting natural logarithm here); thus, $-\log \bar{P}_j(\mathbf{x})$ is the code length implied by $\bar{P}_j$.

1. 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 $\mathcal{M}_j$ [32].

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

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

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 to define 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 [32]). 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 nonempty, 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 sublinearly in m.

A standard result [32] states that for every INECCSI subset $\boldsymbol{\Theta}'_j$ and regular prior w_j, there exists $C > 0$ such that for all m:

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

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

(1) The minimum achievable worst-case redundancy grows as $(d_j/2)\log m$;

(2) 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.

2. *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 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 $\mathcal{M}_0$:

$$-\log S(\mathbf{x}) = -\log \bar{P}_1(\mathbf{x}) + \big[\log P_0^{w_0}(\mathbf{x})\big]. \quad (20)$$

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 (20) 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 (Open Problem No. 9, p. 413 [32]). 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 [33]. 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 $S_{\mathrm{cand}} = \bar{P}_1(\mathbf{x})/P_0^{w_0^*}(\mathbf{x})$ in definition 15, as $\mathrm{REG}_1(\boldsymbol{\Theta}'_1;\bar{P}_1)$. Then, for any distribution $\bar{P}_1$, according to definitions (15) and (12) (see Sec. S1 of the Supplemental Material [36]) it holds:

$$\mathrm{REG}_1(\boldsymbol{\Theta}'_1;\bar{P}_1) \leqslant \mathrm{RED}_1(\boldsymbol{\Theta}'_1;\bar{P}_1). \quad (21)$$

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 $\bar{P}_1$ to the Bayesian mixture $P_1^{w_1}$ (with a regular prior), which yields small regret of order $(d_1/2)\log m + O(1)$. This choice is also common in the literature, as Bayesian mixtures are widely adopted in the construction of e-variables [4]. In Sec. S7 [36], we show the same results for another universal distribution, the Normalized Maximum Likelihood $\bar{P}_1^{\mathrm{NML}}$, which yields regret of the same order, and has already been used (implicitly) in [37] as an e-variable numerator.

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

$$\Omega(\mathbf{c}) := \sum_{\mathbf{x}\,:\,\mathbf{c}(\mathbf{x})=\mathbf{c}} 1. \tag{22}$$

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

$$P_{\text{mic}}(\mathbf{x};\mathbf{c}) = \begin{cases} \frac{1}{\Omega(\mathbf{c})}, & \text{if } \mathbf{c}(\mathbf{x}) = \mathbf{c}; \\ 0, & \text{else.} \end{cases} \tag{23}$$

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:

$$P_{\text{can}}(\mathbf{x};\boldsymbol{\theta}) = \frac{e^{-\boldsymbol{\theta}\cdot\mathbf{c}(\mathbf{x})}}{Z(\boldsymbol{\theta})}, \tag{24}$$

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. (23), 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. (24).

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:

$$\boldsymbol{\mu}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{c}(\mathbf{x})]. \tag{25}$$

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

(1) The canonical parameter space is the full space $\boldsymbol{\Theta}_{\text{can}} = \mathbb{R}^d$;

(2) 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" expected constraint values.[2] In the rest of the paper, we will make use of this bijection and employ whichever parametrization 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. 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

$$S^{\text{GRO}}_{\text{mic}} = \frac{P^{W_1}_{\text{mic},1}(\mathbf{x})}{P^{W_0^*}_{\text{mic},0}(\mathbf{x})}, \tag{26}$$

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

$$W_0^* = \arg\min_{W\in\mathcal{W}_{\mathbf{c}_0}} D_{\text{KL}}\big(P^{W_1}_{\text{mic},1} \| P^{W_0}_{\text{mic},0}\big), \tag{27}$$

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 is in Sec. S8 [36]; here we report only the main results) and find the optimal prior distribution on the null:

$$W_0^*(\mathbf{c}_0) = \sum_{\mathbf{x}\,:\,\mathbf{c}_0(\mathbf{x})=\mathbf{c}_0} P^{W_1}_{\text{mic},1}(\mathbf{x}), \tag{28}$$

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:

$$\text{there exists a function } f : \mathcal{C}_1 \to \mathcal{C}_0 \text{ s.t. } \mathbf{c}_0(\mathbf{x}) = f(\mathbf{c}_1(\mathbf{x})). \tag{29}$$

Under Condition A, one can write

$$W_0^*(\mathbf{c}_0) = \sum_{\mathbf{c}_1\,:\,f(\mathbf{c}_1)=\mathbf{c}_0} W_1(\mathbf{c}_1), \tag{30}$$

[2]It follows from the general theory of exponential families with finite support (Theorem 9.2 [38]), under a technical condition known as *steepness*, which holds when $\mathcal{C}$ is finite.

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

$$S_{\text{mic}}^{\text{GRO}}(\mathbf{x}) = \frac{\Omega_0(\mathbf{c}_0(\mathbf{x}))}{\Omega_1(\mathbf{c}_1(\mathbf{x}))} \frac{W_1(\mathbf{c}_1(\mathbf{x}))}{W_0^*(\mathbf{c}_0(\mathbf{x}))}. \tag{31}$$

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

$$\mathbb{E}_0\big[S_{\text{mic}}^{\text{GRO}}\big] = 1 \quad \forall P_{\text{mic},0} \in \mathcal{M}_{\text{mic},0}. \tag{32}$$

This direct proof, as well as the section's other detailed calculations and proofs, can be found in Sec. S2 [36]. 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 $\mathbf{x}^a = (x_1^a, \dots, x_{n^a}^a)$ and $\mathbf{x}^b = (x_1^b, \dots, x_{n^b}^b)$, with n^a and n^b the respective group sizes. The total sample size is $n = n_a + n_b$. We denote by $n_1^a = \sum_{i=1}^{n_a} x_i^a$ and $n_1^b = \sum_{i=1}^{n_b} x_i^b$ the total number of ones in $\mathbf{x}^a$ and $\mathbf{x}^b$, and by $n_1 = n_1^a + n_1^b$ the total number of ones 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 ones in each group, $\mathbf{c}_1 = (n_1^a, n_1^b)$, and the null sufficient statistic equal to the total number of ones, $c_0 = n_1$. In the microcanonical formulation, these quantities are treated as fixed in the respective models. To find the microcanonical GRO e-variable, we apply formula (31), where

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

(2) $\Omega_1(n_1^a, n_1^b) = \binom{n^a}{n_1^a}\binom{n^b}{n_1^b}$ is the number of permutations of $\mathbf{x}$ preserving the total number of ones 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:

$$W_1\big(n_1^a, n_1^b\big) = \mathcal{U}_a\big(n_1^a\big)\,\mathcal{U}_b\big(n_1^b\big) = \frac{1}{n^a+1}\,\frac{1}{n^b+1}. \tag{33}$$

In this case Condition A (29) 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:

$$W_0^*(n_1) = \begin{cases} \frac{n_1+1}{(n^a+1)(n^b+1)}, & \text{if } 0 \leqslant n_1 \leqslant \min(n^a, n^b), \\ \frac{\min(n^a,n^b)+1}{(n^a+1)(n^b+1)}, & \text{if } \min(n^a,n^b) < n_1 \leqslant \max(n^a,n^b), \\ \frac{n^a+n^b+1-n_1}{(n^a+1)(n^b+1)}, & \text{if } \max(n^a,n^b) < n_1 \leqslant n^a+n^b, \\ 0, & \text{if otherwise,} \end{cases} \tag{34}$$

as shown in Fig. 2. The microcanonical GRO-optimal e-value is finally obtained by substituting each term in Eq. (31).

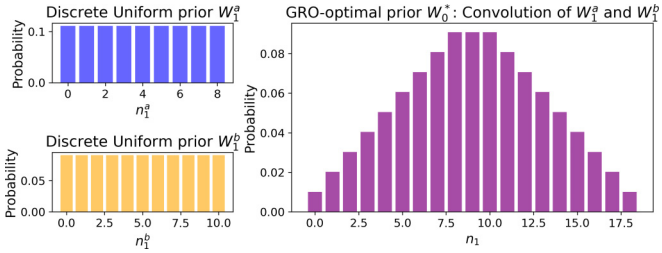

FIG. 2. In the microcanonical Example A, when the prior on the alternative sufficient statistics n_1^a and n_1^b are uniform distributions (on the left), the resulting GRO-optimal prior on the null sufficient statistic n_1 is the convolution of the two uniform distributions, which results in a triangular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

$$S_{\text{can}}^{\text{GRO}} = \frac{\bar{P}_{\text{can},1}(\mathbf{x})}{P_{\text{can},0}^{w_0^*}(\mathbf{x})}, \tag{35}$$

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

$$w_0^* = \arg\min_{w \in \mathcal{W}_{\mu_0}} D_{\text{KL}}\big(\bar{P}_1 \| P_0^{w_0}\big). \tag{36}$$

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*. As we will see, the first serves 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 Sec. S3 [36]:

(1) 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

$$\bar{P}_{\text{can},1} = P_{\text{mic},1}^{W_{\text{can},1}} \tag{37}$$

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

$$W_{\text{can},j}(\mathbf{c}_j) = \sum_{\mathbf{x}\,:\,\mathbf{c}_j(\mathbf{x}) = \mathbf{c}_j} \bar{P}_{\text{can},j}(\mathbf{x}), \tag{38}$$

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

(2) 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:

$$\begin{aligned} &\mathbb{E}_P[\,E\,] \leqslant 1 \;\; \forall P \in \mathcal{M}_{\text{mic}} \\ &\qquad \Rightarrow \mathbb{E}_P[\,E\,] \leqslant 1 \;\; \forall P \in \mathcal{M}_{\text{can}}. \end{aligned} \tag{39}$$

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 $S^{\text{GRO}}_{\text{mic}}$ for the microcanonical test based on the corresponding $P^{W_{\text{can},1}}_{\text{mic},1} = \bar{P}_{\text{can},1}$. Thus, (26) becomes

$$S^{\text{GRO}}_{\text{mic}} = \frac{\bar{P}_{\text{can},1}(\mathbf{x})}{P^{W_0^*}_{\text{mic},0}(\mathbf{x})} = \bar{P}_{\text{can},1}\frac{\Omega_0(\mathbf{c}_0(\mathbf{x}))}{W_0^*(\mathbf{c}_0(\mathbf{x}))}, \tag{40}$$

where, readapting (28),

$$W_0^*(\mathbf{c}_0) = \sum_{\mathbf{x}\,:\,\mathbf{c}_0(\mathbf{x})=\mathbf{c}_0} \bar{P}_{\text{can},1}(\mathbf{x}). \tag{41}$$

Given the second fact (39), 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:

$$\mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S^{\text{GRO}}_{\text{mic}}\right] \leqslant \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S^{\text{GRO}}_{\text{can}}\right]. \tag{42}$$

The pseudo-approximation is further built over the microcanonical one. The prior W_0^* (used to build $S^{\text{GRO}}_{\text{mic}}$), 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 Sec. S5 [36]. 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

$$S_{\text{pseudo}} = \frac{\bar{P}_{\text{can},1}(\mathbf{x})}{P^{w_{\text{pseudo},0}}_{\text{can},0}(\mathbf{x})} \tag{43}$$

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

$$D_{\text{KL}}\left(\bar{P}_{\text{can},1}\middle\|P^{w_0^*}_{\text{can},0}\right) \leqslant D_{\text{KL}}\left(\bar{P}_{\text{can},1}\middle\|P^{w_{\text{pseudo},0}}_{\text{can},0}\right) \tag{44}$$

or, equivalently,

$$\mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S^{\text{GRO}}_{\text{can}}\right] \leqslant \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S_{\text{pseudo}}\right]. \tag{45}$$

Consequently, one has

$$\begin{aligned}\mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S^{\text{GRO}}_{\text{mic}}\right] &\leqslant \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S^{\text{GRO}}_{\text{can}}\right] \\ &\leqslant \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S_{\text{pseudo}}\right],\end{aligned} \tag{46}$$

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 (38). Then we compute $S^{\text{GRO}}_{\text{mic}}$ according to the formulas expressed in the previous section [Eqs. (28) and (31)].

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

$$r = \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S_{\text{pseudo}}\right] - \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log S^{\text{GRO}}_{\text{mic}}\right] \geqslant 0 \tag{47}$$

where, for future reference, it is useful to note that, using definitions (43) and (40) and sufficiency, we can rewrite

$$\begin{aligned} r &= \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log P^{W_0^*}_{\text{mic},0}(\mathbf{x}) - \log P^{w_{\text{pseudo},0}}_{\text{can},0}(\mathbf{x})\right] \\ &= \mathbb{E}_{\bar{P}_{\text{can},1}}\left[\log W_0^*(\mathbf{c}_0(\boldsymbol{x})) - \log W_{\text{pseudo},0}(\mathbf{c}_0(\mathbf{x}))\right],\end{aligned} \tag{48}$$

where $W_{\text{pseudo},0}(\mathbf{c}_0(\mathbf{x}))$ is defined as in (38).

The evaluation above is under $\bar{P}_{\text{can},1}$ expectation; this makes sense if we use a Bayesian universal distribution $\bar{P}_{\text{can},1} = P^{w_1}_{\text{can},1}$ 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 (see Sec. S7 [36]), 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 Eq. (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 $\boldsymbol{\mu}_1 = \boldsymbol{\mu}(\boldsymbol{\theta}_1)$:

$$\begin{aligned} r'(\boldsymbol{\mu}_1) &= \mathbb{E}_{\boldsymbol{\mu}_1}\left[\log S_{\text{pseudo}} - \log S^{\text{GRO}}_{\text{mic}}\right] \\ &= \mathbb{E}_{\boldsymbol{\mu}_1}\left[\log W_0^*(\mathbf{c}_0(\boldsymbol{x})) - \log W_{\text{pseudo},0}(\mathbf{c}_0(\mathbf{x}))\right],\end{aligned} \tag{49}$$

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':

$$r' := \max_{\boldsymbol{\mu}_1 \in \mathtt{M}_1} r'(\boldsymbol{\mu}_1). \tag{50}$$

It can be easily argued that

$$r \geqslant 0 \Rightarrow r' \geqslant 0. \tag{51}$$

In case r (or r') is small, we know that our easily computable microcanonical e-variable $S^{\text{GRO}}_{\text{mic}}$ is close to optimal according to the GRO criterion for the canonical problem, and hence can be used instead of the canonical $S^{\text{GRO}}_{\text{can}}$. 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 ones 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:

$$P^{w_1}_{\text{can},1} = \int_0^1 p_a^{n_1^a}(1-p_a)^{n^a-n_1^a} dp_a \int_0^1 p_b^{n_1^b}(1-p_b)^{n^b-n_1^b} dp_b$$
$$= \binom{n^a}{n_1^a}^{-1} \frac{1}{n^a+1} \binom{n^b}{n_1^b}^{-1} \frac{1}{n^b+1}. \quad (52)$$

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 $P^{w_1}_{\text{can},1} = P^{W_{\text{can},1}}_{\text{mic},1}$. By inspecting Eq. (52), 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 $S^{\text{GRO}}_{\text{mic}}$ 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 $w^1_{\text{pseudo},0}$ in Fig. 3 to be distinct from $w^2_{\text{pseudo},0}$ of the following Example C). A schematic representation of how $S^{\text{GRO}}_{\text{mic}}$ and S^{pseudo} are built is shown in Fig. 3.

Once $S^{\text{GRO}}_{\text{mic}}$ and S^{pseudo} are computed, we can compute the gap between their e-power, r, which contains the optimal e-power: the smaller r, the better the microcanonical approximation works. In our simple example, r is already very small for small sample sizes and converges very rapidly to 0 as the latter increases, as shown in Fig. 4).

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 Eq. (47) converges very fast to zero 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

$$P_{\text{can}}(\mathbf{x};\boldsymbol{\theta}) = \frac{e^{-\boldsymbol{\theta}\cdot\mathbf{c}(\mathbf{x})}}{Z(\boldsymbol{\theta})}, \quad (53)$$

and induces the mean-value mapping

$$\boldsymbol{\mu}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{c}(\mathbf{x})], \quad (54)$$

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

$$\mathbf{s}^{(m)}(\mathbf{y}^{(m)}) = \sum_{j=1}^{m} \mathbf{c}(\mathbf{x}_j). \quad (55)$$

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

$$Q^w(\boldsymbol{\mu} \in \mathtt{M}') := \int_{\mathtt{M}'} w(\boldsymbol{\mu})\, d\boldsymbol{\mu}. \quad (56)$$

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 $\mathtt{M} \subset \mathbb{R}^d$. Then, for any INECCSI (Definition 1) subset $\mathtt{M}'$ of $\mathtt{M}$, we have

$$\left| P^{w^{(m)}}_{\text{can}} \left\{ \frac{\mathbf{s}^{(m)}(\mathbf{y}^{(m)})}{m} \in \mathtt{M}' \right\} - Q^w\{\boldsymbol{\mu} \in \mathtt{M}'\} \right| = O\left(\frac{\log m}{m}\right). \quad (57)$$

In words: under the Bayesian marginal likelihood $P^{w^{(m)}}_{\text{can}}$, 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 (55) 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:

(1) $\mathbf{s}_1^{(m)} = (n_1^{a\,(m)}, n_1^{b\,(m)})$: number of ones in each group under the alternative,

(2) $\mathrm{s}_0^{(m)} = \frac{1}{2}(n_1^{a\,(m)} + n_1^{b\,(m)})$: 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: $P^{w_1(m)}_{\text{can},1}$ induces a probability $W^{(m)}_{\text{can},1}$ 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 (Fig. 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, $V_1^{(m)}$

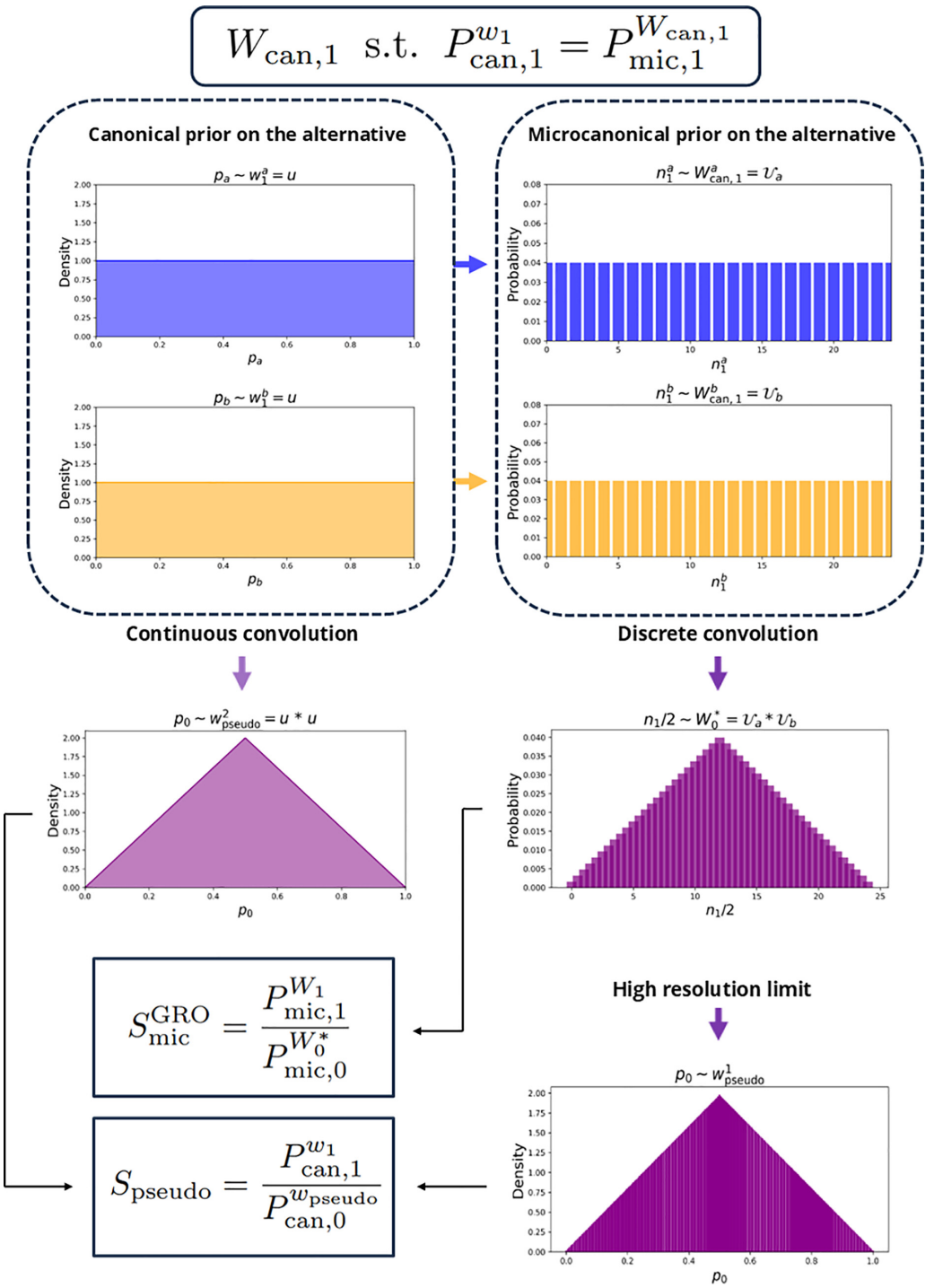

FIG. 3. Procedures to compute the microcanonical approximation $S^{\text{GRO}}_{\text{mic}}$ and the pseudo approximation S_{pseudo} for testing between two binary data streams, under uniform priors (as in Examples A, B, and C). Starting from two independent continuous uniform priors on the alternative (top left), we construct discrete microcanonical priors (top right) satisfying $P^{w_1}_{\text{can},1} = P^{W_{\text{can},1}}_{\text{mic},1}$. In the specific case of continuous uniform priors, these discrete priors are also uniform. The optimal discrete prior W_0^*, used in $S^{\text{GRO}}_{\text{mic}}$, 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 ($w^1_{\text{pseudo},0}$), or by directly convolving the original continuous priors ($w^2_{\text{pseudo},0}$).

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 1 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 $w^1_{\text{pseudo},0}$. This distribution can be

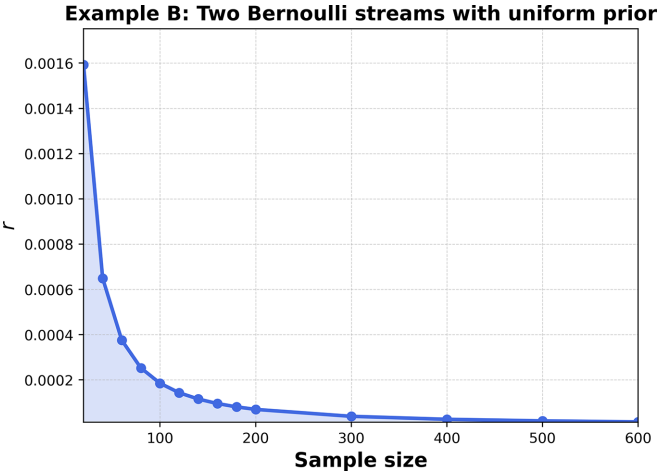

FIG. 4. Convergence of the interval width r for a canonical test between two streams of binary data (as in Example B), for $n^a = n^b = m$, as the sample size $n = 2m$ grows.

approximated by considering a very large m and "smoothening" the corresponding $V_0^{*(m)}$ to $w^1_{\text{pseudo},0}$. This limiting procedure is precisely what we referred to earlier as the *high-resolution limit*. Once $w^1_{\text{pseudo},0}$ is obtained, $P_{\text{can}}^{w^1_{\text{pseudo},0}(m)}$ induces a probability distribution $W^{1(m)}_{\text{pseudo},0}$ on the null sufficient statistics. As above, we can define

$$V^{1(m)}_{\text{pseudo},0}\left(\frac{\mathbf{s}_0^{(m)}}{m}\right) := W^{1(m)}_{\text{pseudo},0}\left(\mathbf{s}_0^{(m)}\right). \tag{58}$$

Now we invoke Theorem 1 again: it indicates that $V^{1(m)}_{\text{pseudo},0}$ converges to $w^1_{\text{pseudo},0}$. Thus, one may expect $V_0^{*(m)}$ and $V^{1(m)}_{\text{pseudo},0}$, and consequently $W_0^{*(m)}$ and $W^{1(m)}_{\text{pseudo},0}$, to be close and r to be small, according to Eq. (48). 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 $w^2_{\text{pseudo},0}$ 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 $V^{2(m)}_{\text{pseudo},0}$ on $\mathbf{s}_0^{(m)}/m$ converges to a discretized version of $w^2_{\text{pseudo},0}$. At the same time, $V_0^{*(m)}$ is constructed by first discretizing the prior w_1 and then computing the discrete convolution of its two components. In the high-resolution limit $m \to \infty$, this procedure yields the discretized version of the *continuous* convolution of w_1. Thus, in the large-m limit, $w^1_{\text{pseudo},0}$ and $w^2_{\text{pseudo},0}$ 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 (Fig. 3).

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 Fig. 3); 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 Fig. 3) 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 (57) is too weak to formally imply $r \to 0$ (it concerns probabilities of sets, whereas (48) involves expectations of log densities). Nevertheless, it provides strong heuristic evidence, and we do observe convergence numerically (see Fig. 4). 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, 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.

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., [39,40] for connections between network modeling and contingency tables, and the discussion in Sec. 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 (37) and then applying formula (40). 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 $\mathbf{x}_a = (x_1^a, \ldots, x_{n^a}^a)$ and $\mathbf{x}_b = (x_1^b, \ldots, x_{n^b}^b)$, 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:

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

$$\mathbf{c}_1 = \left(n_1^a, n_1^b\right). \tag{59}$$

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

$$\mathrm{c}_0 = n_1. \tag{60}$$

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 Sec. 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.

1. 2 × 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

$$P_{\mathrm{mic},\,0}(\mathbf{x}; n_1) = \begin{cases} \frac{1}{\Omega_0(n_1)}, & \text{if } n_1(\mathbf{x}) = n_1, \\ 0, & \text{else,} \end{cases} \tag{61}$$

where

$$\Omega_0(n_1) = \binom{n}{n_1} \tag{62}$$

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

$$P_{\mathrm{mic},\,1}(\mathbf{x}; n_1^a, n_1^b) = \begin{cases} \frac{1}{\Omega_1(n_1^a, n_1^b)}, & \text{if } \left(n_1^a(\mathbf{x}), n_1^b(\mathbf{x})\right) \\ & \quad = \left(n_1^a, n_1^b\right), \\ 0, & \text{else,} \end{cases} \tag{63}$$

where

$$\Omega_1\left(n_1^a, n_1^b\right) = \binom{n^a}{n_1^a}\binom{n^b}{n_1^b} \tag{64}$$

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

For any given prior W_1 on the alternative sufficient statistics, $S_{\mathrm{mic}}^{\mathrm{GRO}}$ is found exactly by computing W_0^* and applying (31). In this case Condition A (29) is satisfied, as the null sufficient statistics can be written as a function of the alternative one:

$$n_1 = n_1^a + n_1^b. \tag{65}$$

Thus, following (30), 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:

$$W_1\left(n_1^a, n_1^b\right) = W_1^a\left(n_1^a\right) \cdot W_1^b\left(n_1^b\right), \tag{66}$$

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

$$W_0^* = W_1^a * W_1^b, \tag{67}$$

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

Notice that an example application of this formula for the case of two independent uniform priors has already been carried out in Example A of Sec. III A.

2. 2 × 2 canonical test

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

$$P_{\mathrm{can}}(\mathbf{x}; \theta_0) = \frac{e^{-\theta_0 \cdot n_1(\mathbf{x})}}{(1 + e^{-\theta_0})^n}, \tag{68}$$

which can be rewritten in the mean-value parametrization:

$$P_{\mathrm{can}}(\mathbf{x}; p_0) = p_0^{n_1(\mathbf{x})}(1 - p_0)^{n - n_1(\mathbf{x})} \tag{69}$$

upon defining

$$p_0 = \frac{e^{-\theta_0}}{1+e^{-\theta_0}}. \tag{70}$$

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:

$$P_{\text{can}}(\mathbf{x};\theta_a,\theta_b) = \frac{e^{-\theta_a\cdot n_1^a(\mathbf{x})-\theta_b\cdot n_1^b(\mathbf{x})}}{(1+e^{-\theta_a})^{n^a}(1+e^{-\theta_b})^{n^b}}, \tag{71}$$

or, equivalently,

$$\begin{aligned} P_{\text{can}}(\mathbf{x};p_a,p_b) = {} & p_a^{n_1^a(\mathbf{x})}(1-p_a)^{n^a-n_1^a(\mathbf{x})} \\ & \times p_b^{n_1^b(\mathbf{x})}(1-p_b)^{n^b-n_1^b(\mathbf{x})} \end{aligned} \tag{72}$$

upon defining the mean-value parameters:

$$p_a = \frac{e^{-\theta_a}}{1+e^{-\theta_a}} \quad \text{and} \quad p_b = \frac{e^{-\theta_b}}{1+e^{-\theta_b}}. \tag{73}$$

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 Sec. III B to different choices of $\bar{P}_{\text{can},1}$.

Example 1: Independent beta priors. The beta probability distribution reads:

$$\text{Beta}(y;\alpha,\beta) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)}, \quad \text{for } y\in(0,1), \tag{74}$$

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

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt. \tag{75}$$

The beta prior represents a popular choice because it is flexible enough to encompass several cases of interest (Table S1 [36]). 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

$$\bar{P}_{\text{can},1}(\mathbf{x}) = \frac{B(\bar{\alpha}^a,\bar{\beta}^a)}{B(\alpha^a,\beta^a)}\frac{B(\bar{\alpha}^b,\bar{\beta}^b)}{B(\alpha^b,\beta^b)}, \tag{76}$$

where

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

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

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

with

$$W_{\text{can},1}^a(n_1^a) = \Omega_1^a(n_1^a)\,\frac{B(\bar{\alpha}^a,\bar{\beta}^a)}{B(\alpha^a,\beta^a)} \tag{78}$$

and

$$W_{\text{can},1}^b(n_1^b) = \Omega_1^b(n_1^b)\,\frac{B(\bar{\alpha}^b,\bar{\beta}^b)}{B(\alpha^b,\beta^b)}. \tag{79}$$

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 $W_{\text{can},1}^a(n_1^a)$ and $W_{\text{can},1}^b(n_1^b)$. 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 (34). When no closed form is available, the convolution can be computed numerically.

In Fig. S1 [36] we show W_1^a, W_1^b, W_0^*, $w_{\text{pseudo},0}^1$, and $w_{\text{pseudo},0}^2$ 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, $w_{\text{pseudo},0}^1$ and $w_{\text{pseudo},0}^2$ are almost indistinguishable. This is a consequence of Theorem 1: the distribution of the mean value parameter ($w_{\text{pseudo},0}^2$) and that of the sufficient statistic ($w_{\text{pseudo},0}^1$) 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 $\gamma > 0$.

3. 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 (36). We find that the e-power of the microcanonical approximation converges to that of the canonical GRO e-variable as the total size grows (Fig. S2 [36]). 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 $S_{\text{can}}^{\text{GRO}}$.

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 Fig. S3 [36], we show the plot of r as defined earlier to evaluate the effectiveness

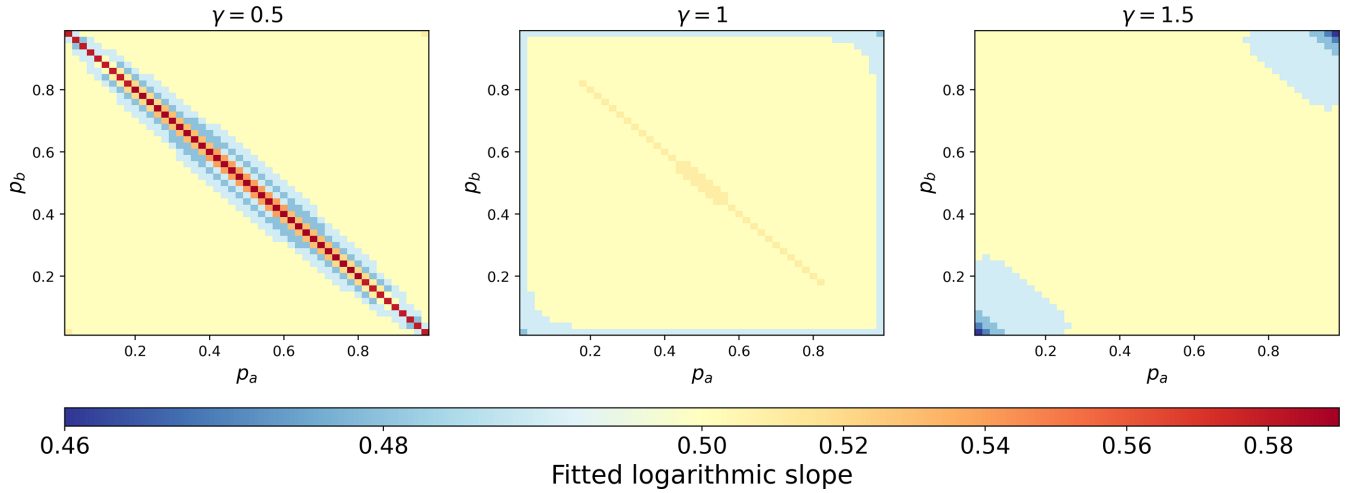

FIG. 5. Fitted slope of the logarithmic growth $a \log m + b$ of the microcanonical approximation regret (14) in the 2×2 case ($n^a = n^b = m$), shown for different combinations of the alternative parameters (p_a, p_b). The expected asymptotic slope is 0.5 (yellow). Three alternative beta priors are 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.

of the microcanonical approximation in different scenarios. The results confirm those of Fig. S2 [36], as in all cases considered, r converges to 0. In conclusion, we argue that the microcanonical approximation is a perfect candidate in this case.

4. Results on regret

Let's again consider the m-dimensional i.i.d. extension of our models. In Sec. S6 [36], we show that, if the error $r'(\boldsymbol{\mu}_1)$ in (49) vanishes as the sample size m increases, both the canonical growth-optimal e-variable $S^{\mathrm{GRO}}_{\mathrm{can}}$ and its microcanonical approximation $S^{\mathrm{GRO}}_{\mathrm{mic}}$ satisfy

$$\mathrm{REG}_1(\boldsymbol{\mu}_1, \cdot) = \frac{d_1 - d_0}{2} \log m + O(1). \tag{80}$$

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

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

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 Fig. 5, we observe a clear dichotomy:

(1) For $\gamma < 1$, the convolution of the beta priors w_1^a and w_1^b is nondifferentiable at $p_0 = 1/2$, as shown in Fig. S1 [36] (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 $S^{\mathrm{GRO}}_{\mathrm{can}}$ and $S^{\mathrm{GRO}}_{\mathrm{mic}}$, and (80) no longer holds (see Fig. S4 [36]). Indeed, in Fig. 5, we see that even on small INECCSI sets, the regret grows like $a \log m + b$ for some $a > 1/2$.

(2) For $\gamma = 1$, convergence is moderate. Although r' decays quickly (Fig. S4 [36]), the experimental values of Fig. 5 show 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 ($m \leqslant 1800$).

(3) 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 Fig. 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, the Jeffreys prior [32] 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 [41]]. However, in our setting, the 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 $\mathbf{x}_i = (x_1^i, \ldots, x_{n^i}^i)$ for $i = 1, \ldots, k$, 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 goal is to test whether all groups share the same probability of observing $x = 1$. This problem again translates into a hypothesis testing problem, where

(1) Under the alternative hypothesis, the groups are allowed to have different probabilities, meaning the number of ones is constrained separately in each group:

$$\mathbf{c}_1 = \left(n_1^1, n_1^2, \ldots, n_1^k\right). \tag{81}$$

(2) Under the null hypothesis, all groups share the same probability, meaning only the total number of ones is constrained:

$$c_0 = n_1. \tag{82}$$

Rejection of the null, therefore, indicates a lack of homogeneity across groups, or equivalently, that at least two groups differ in their outcome distribution.

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 Sec. 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.

1. $2 \times k$ microcanonical test

While the null model stays the same [Eq. (61)], the alternative model is simply the extension of (63) from 2 to k groups:

$$P_{\text{mic},1}(\mathbf{x}; \{n_1^i\}) = \begin{cases} \frac{1}{\Omega_1(\{n_1^i\})}, & \text{if } \left(n_1^1(\mathbf{x}), \ldots, n_1^k(\mathbf{x})\right) \\ & \quad = \left(n_1^1, \ldots, n_1^k\right), \\ 0, & \text{else}, \end{cases} \tag{83}$$

where

$$\Omega_1\left(\{n_1^i\}\right) = \prod_{i=1}^{k} \binom{n^i}{n_1^i}. \tag{84}$$

As in the 2×2 case, Condition A (29) is satisfied:

$$n_1 = \sum_{i=1}^{k} n_1^i, \tag{85}$$

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,

$$W_1\left(\{n_1^i\}\right) = \prod_{i=1}^{k} W_1^i\left(n_1^i\right), \tag{86}$$

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

$$W_0^* = W_1^1 * \cdots * W_1^k. \tag{87}$$

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 $k \gg 1$:

$$W_0^*(n_1) \approx \frac{1}{N(\mu_k, \sigma_k)} \exp\left(-\frac{(n_1 - \mu_k)^2}{2\sigma_k^2}\right), \tag{88}$$

where N is the normalization constant and

$$\mu_k = \sum_{i=1}^{k} \mathbb{E}_{W_1^i}\left[n_1^i\right],$$

$$\sigma_k^2 = \sum_{i=1}^{k} \mathrm{Var}_{W_1^i}\left(n_1^i\right).$$

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

$$W_1\left(\{n_1^i\}\right) = \prod_{i=1}^{k} \mathcal{U}_i(n_1^i) = \prod_{i=1}^{k} \frac{1}{n^i + 1}, \tag{89}$$

the GRO null prior is again the convolution of all the individual priors, i.e., the convolution of k discrete uniform distributions, which reads [42]

$$W_0^*(n_1) = \sum_{S \subseteq \{1, \ldots k\}} (-1)^{|S|} \binom{n_1 + k - 1 - \sum_{j \in S}(n^j - n)}{n_1 - 1} \times \left[\prod_{i=1}^{k} n^i + 1\right]^{-1}, \tag{90}$$

where the sum runs over all possible subsets of $\{1, \ldots, k\}$ and $|\, S \,|$ 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

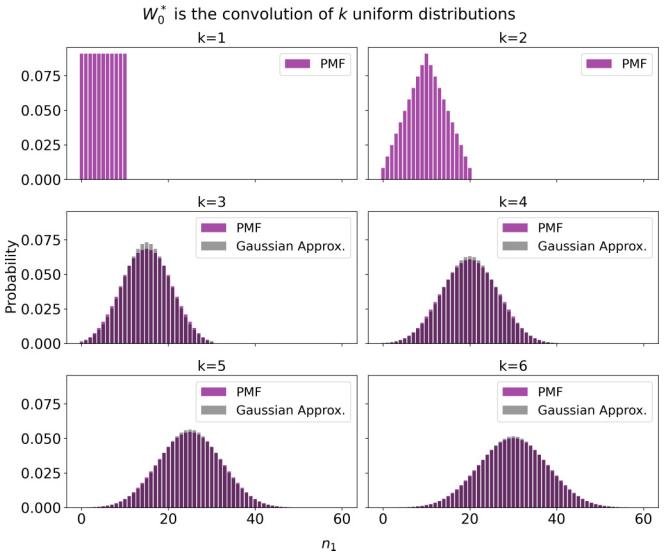

FIG. 6. The microcanonical GRO-optimal prior on the null W_0^* for testing $2 \times k$ tables is obtained as the convolution of the k independent priors on the alternative, which are discrete uniform priors in the case shown in this picture. Each convolution, for $k > 2$, is superposed to its discrete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

reads:

$$W_0^*(n_1) = \sum_{j=1}^{\lfloor n_1/(m+1)\rfloor} (-1)^j \binom{n}{j} \binom{n_1 - j(m+1) + k - 1}{k-1} \times \left[\prod_{i=1}^{k} n^i + 1\right]^{-1}, \tag{91}$$

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

2. $2 \times k$ canonical test

The null canonical model is the same as in the 2×2 case, Eq. (68), 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. (71) and (72) to the case of k groups, by constraining the expected values of n_1^i separately for each group i, leading to the expression

$$P_{\text{can}}(\mathbf{x}; \boldsymbol{\theta}) = \frac{e^{-\sum_{i=1}^{k} \theta_i n_1^i(\mathbf{x})}}{\prod_{i=1}^{k} (1 + e^{-\theta_i})^{n^i}}, \tag{92}$$

or, equivalently, in the mean-value parametrization:

$$P_{\text{can}}(\mathbf{x}; \boldsymbol{p}) = \prod_{i=1}^{k} p_i^{n_1^i(\mathbf{x})} (1 - p_i)^{n^i - n_1^i(\mathbf{x})}, \tag{93}$$

where we define the group-specific probabilities as

$$p_i = \frac{e^{-\theta_i}}{1 + e^{-\theta_i}}, \quad \text{for each } i \in \{1, \ldots, k\}. \tag{94}$$

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 Sec. III B to different choices of $\bar{P}_{\text{can},1}$.

Example 3: Independent beta priors. Here we extend Example 1 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:

$$\bar{P}_{\text{can},1}(\mathbf{x}) = \prod_{i=1}^{k} \frac{B(\bar{\alpha}^i, \bar{\beta}^i)}{B(\alpha^i, \beta^i)}, \tag{95}$$

where

$$\bar{\alpha}^i = n_1^i + \alpha^i$$
$$\bar{\beta}^i = n^i - n_1^i + \beta^i \quad \text{for each } i \in \{1, \ldots, k\}.$$

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:

$$W_{\text{can},1}(\{n_1^i\}) = \prod_{i=1}^{k} W_{\text{can},1}^i(n_1^i) \tag{96}$$

with

$$W_{\text{can},1}^i(n_1^i) = \Omega_1^i(n_1^i) \cdot \frac{B(\bar{\alpha}^i, \bar{\beta}^i)}{B(\alpha^i, \beta^i)} \quad \text{for each } i \in \{1, \ldots, k\}. \tag{97}$$

$W_0^*(n_1)$ 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 (90). 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 (88). Analogously, a continuous Gaussian approximation can be used to approximate w_{pseudo}. In Fig. S5 [36], we show $W_{\text{can},1}^i$, W_0^*, and w_{pseudo} for $k = 10$.

3. 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 $i \in \{1, \ldots, k\}$. In this scenario we have that $n = m\,k$. 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 $w_{\text{pseudo},\,0}^1$, according to the procedure described in Example C, which is easily extended to the case of k groups. Our experiments (Fig. S6 [36]) show the following:

(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 $S_{\text{mic}}^{\text{GRO}}$ 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 [14,16–18,43].

For instance, when applied to networks, maximum entropy models in their canonical formulations are known as *exponential random graph models* [18,44,45]. Examples of commonly used maximum entropy network models are the ErdősRényi 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, Sec. II A establishes a link between e-values and the Minimum Description Length principle—a framework increasingly used in recent years for network inference, model selection [46–49].

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 binary network ensembles [16,18]. Indeed, a binary network is represented by a binary adjacency matrix, which is a (structured) collection of ones and zeros, corresponding to the presence or absence of a link between two nodes.

In particular, the null model considered here, in both its canonical and microcanonical formulation, corresponds to the well-known *Erdős-Rényi* (ER) model, where the sufficient statistic is the total number of links, equal to (half, if the network is undirected) the total number of ones 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 intragroup links. For instance, in models of networks with community structure, intragroup link probabilities are larger than intergroup ones. The sufficient statistics are the number of links in each block. Testing an SBM against an ER model corresponds exactly to testing whether the connection probabilities are identical across all blocks (i.e., communities are absent), and this SBM vs ER problem reduces to our canonical or microcanonical $2 \times k$ contingency table test.

In the *Partial Configuration Model* (PCM) for bipartite networks [50], 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 ones 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 nonstationarity 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 [49,51]. 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, 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 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.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available. The data are available from the authors upon reasonable request.

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

[2] D. J. Benjamin, J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, *et al.*, Redefine statistical significance, Nat. Hum. Behav. **2**, 6 (2017).

[3] B. B. McShane, D. Gal, A. Gelman, C. Robert, and J. L. Tackett, Abandon statistical significance, Am. Stat. **73**, 235 (2019).

[4] A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer, Game-theoretic statistics and safe anytime-valid inference, Stat. Sci. **38**, 576 (2023).

[5] A. Ramdas and R. Wang, Hypothesis testing with e-values, Found. Trends Stat. **1**, 1 2025.

[6] Y. Zhang, S. Glancy, and E. Knill, Asymptotically optimal data analysis for rejecting local realism, Phys. Rev. A **84**, 062118 (2011).

[7] P. Grünwald, R. de Heide, and W. Koolen, Safe testing, J. R. Stat. Soc. B **86**, 1091 (2024).

[8] L. Wasserman, A. Ramdas, and S. Balakrishnan, Universal inference, Proc. Natl. Acad. Sci. USA **117**, 16880 (2020).

[9] V. Vovk and R. Wang, E-values: Calibration, combination and applications, Ann. Stat. **49**, 1736 (2021).

[10] G. Shafer, Testing by betting: A strategy for statistical and scientific communication, J. R. Stat. Soc. Ser. A **184**, 407 (2021).

[11] T. Lardy, P. Grünwald, and P. Harremoës, Reverse information projections and optimal e-statistics, IEEE Trans. Inf. Theory **70**, 7616 (2024).

[12] M. Larsson, A. Ramdas, and J. Ruf, The numeraire e-variable and reverse information projection, Ann. Stat. **53**, 1015 (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, Cambridge, 2010).

[14] E. T. Jaynes, Information theory and statistical mechanics, Phys. Rev. **106**, 620 (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, Phys. Rep. **757**, 1 (2018).

[18] G. Cimini, T. Squartini, F. Saracco, D. Garlaschelli, A. Gabrielli, and G. Caldarelli, The statistical physics of real-world networks, Nat. Rev. Phys. **1**, 58 (2019).

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

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

[21] R. Turner and P. Grünwald, Anytime-valid confidence intervals for contingency tables and beyond, Stat. Probab. Lett. **198**, 109835 (2023).

[22] R. Turner, A. Ly, and P. Grünwald, Generic e-variables for exact sequential k-sample tests that allow for optional stopping, J. Stat. Plan. Inference **230**, 106116 (2024).

[23] Y. Hao and P. Grünwald, e-values for exponential families: The general case, arXiv:2409.11134.

[24] P. Grünwald, T. Lardy, Y. Hao, S. K. Bar Lev, M. de Jong, Optimal e-values for exponential families: The simple case, arXiv:2404.19465.

[25] P. D. Grünwald, Beyond Neyman–Pearson: E-values enable hypothesis testing with a data-driven alpha, Proc. Natl. Acad. Sci. USA **121**, e2302098121 (2024).

[26] A. Ly, U. Boehm, P. Grünwald, A. Ramdas, and D. van Ravenzwaaij, A tutorial on safe anytime-valid inference: Practical maximally flexible sampling designs for experiments based on e-values (2025), https://doi.org/10.31234/osf.io/h5vae_v3.

[27] R. A Klein, M. Vianello, F. Hasselman, B. G. Adams, R. B. Adams Jr, S. Alper, M. Aveyard, J. R. Axt, M. T. Babalola, Š. Bahník, *et al.*, Many Labs 2: Investigating variation in replicability across sample and setting, Adv. Methods Prac. Psychol. Sci. **1**, 443 (2018).

[28] Z. Zhang, A. Ramdas, and R. Wang, On the existence of powerful p-values and e-values for composite hypotheses, Ann. Stat. **52**, 2241 (2024).

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

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

[31] R. D. Morey, J.-W. Romeijn, and J. N. Rouder, The philosophy of Bayes factors and the quantification of statistical evidence, J. Math. Psychol. **72**, 6 (2016).

[32] P. Grünwald, *The Minimum Description Length Principle* (MIT press, Cambridge, MA, 2007).

[33] P. Grünwald and T. Roos, Minimum description length revisited, Int. J. Math. Ind. **11**, 1930001 (2019).
[34] K. Yamanishi, *Learning with the Minimum Description Length Principle* (Springer Nature, Singapore, 2023).
[35] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, 3rd ed. (Springer, New York, 2008).
[36] See Supplemental Material at http://link.aps.org/supplemental/10.1103/xhf5-1l7p for additional methods, derivations, and figures, which includes additional Refs. [52–54].
[37] K. Jang, K.-S. Jun, I. Kuzborskij, and F. Orabona, Tighter PAC-Bayes bounds through coin-betting, in *Proceedings of the 36th Conference on Learning Theory* (PMLR, 2023), Vol. 195, pp. 2240–2264
[38] O. E. Barndorff-Nielsen, *Information and Exponential Families in Statistical Theory* (Wiley, Chichester, UK, 1978).
[39] H. Alves, P. Brito, and P. Campos, Community detection in interval-weighted networks, Data Min. Knowl. Discovery **38**, 653 (2024).
[40] M. Jerdee, A. Kirkley, and M. E. J. Newman, Mutual information and the encoding of contingency tables, Phys. Rev. E **110**, 064306 (2024).
[41] B. S. Clarke and A. R. Barron, Jeffreys' prior is asymptotically least favorable under entropy risk, J. Stat. Plann. Inference **41**, 37 (1994).
[42] M. Earnest, Extended stars-and-bars problem (where the upper limit of the variable is bounded), https://math.stackexchange.com/q/3182858.
[43] A. Golan, G. Judge, and D. Miller, *Maximum Entropy Econometrics: Robust Estimation with Limited Data* (John Wiley, Chichester, UK, 1996).
[44] P. W. Holland and S. Leinhardt, An exponential family of probability distributions for directed graphs, J. Am. Stat. Assoc. **76**, 33 (1981).
[45] D. R. Hunter and M. S. Handcock, Inference in curved exponential family models for networks, J. Comput. Graphical Stat. **15**, 565 (2006).
[46] T. P. Peixoto, Nonparametric Bayesian inference of the microcanonical stochastic block model, Phys. Rev. E **95**, 012317 (2017).
[47] T. P. Peixoto, Network reconstruction via the minimum description length principle, Phys. Rev. X **15**, 011065 (2025).
[48] 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).
[49] F. Giuffrida, T. Squartini, P. Grünwald, and D. Garlaschelli, Description length of canonical and microcanonical models, Phys. Rev. Res. **7**, 043057 (2025).
[50] Q. Zhang and D. Garlaschelli, Strong ensemble nonequivalence in systems with local constraints, New J. Phys. **24**, 043011 (2022).
[51] T. Squartini, J. de Mol, F. den Hollander, and D. Garlaschelli, Breaking of ensemble equivalence in networks, Phys. Rev. Lett. **115**, 268701 (2015).
[52] I. Csiszár, Sanov property, generalized I-projection and a conditional limit theorem, Ann. Prob, **12**, 768 (1984).
[53] P. Grünwald, Y. Hao, and A. Balsubramani, Growth-optimal e-variables and an extension to the multivariate Csiszár-Sanov-Chernoff theorem, arXiv:2412.17554.
[54] P. P. A. Staniczenko, M. J. Smith, and S. Allesina, Selecting food web models using normalized maximum likelihood, Methods Ecol. Evol. **5**, 551 (2014).