Supplementary material for 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)*

S1. REDUNDANCY AND REGRET

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

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

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

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

S2. MICROCANONICAL TEST

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

A. Exact solution of the optimization problem

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

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

is found by solving the optimization problem

* francesca.giuffrida@imtlucca.it

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

For a microcanonical model with sufficient statistics $\mathbf{c}_i$, the Bayesian marginal likelihood reads [1]

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

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

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

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

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

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

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

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

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

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

and thus

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

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

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

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

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

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

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

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

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

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

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

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

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

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

S3. MICROCANONICAL APPROXIMATION

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

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

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

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

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

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

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

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

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

Thus, from (S17), it follows that

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

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

In what follows, we show that:

1. Given a canonical universal distribution $\bar{P}_{\text{can}}$ relative to a sufficient statistic $\mathbf{c}$, we can always define a prior distribution $W_{\text{can}}(\mathbf{c})$ on the sufficient statistic such that
$$\bar{P}_{\text{can}} = \bar{P}_{\text{mic}}^{W_{\text{can}}}. \tag{S21}$$
2. The opposite is not true: for some $\bar{P}_{\text{mic}}^{W}$, there is no choice of prior density $w(\boldsymbol{\theta})$ such that $\bar{P}_{\text{mic}}^{W} = \bar{P}_{\text{can}}^{w}$.

Proof 1

By construction, a canonical universal distribution $\bar{P}_{\text{can}}(\mathbf{x})$ relative to the sufficient statistics $\mathbf{c}$ always assigns the same probability mass to configurations sharing the same value of the sufficient statistic, just as the corresponding $P_{\text{can}}(\mathbf{x};\boldsymbol{\theta})$ does. Consequently, the following holds:
$$\bar{P}_{\text{can}}(\mathbf{x} \,|\, \mathbf{c}) = P_{\text{mic}}(\mathbf{x};\mathbf{c}), \tag{S22}$$
i.e., when conditioned on the sufficient statistic, the universal canonical distribution reduces to a uniform distribution over the number of configurations corresponding to the observed value, which is the microcanonical distribution. Moreover, we can always write
$$\bar{P}_{\text{can}}(\mathbf{x}) = \bar{P}_{\text{can}}(\mathbf{x} \mid \mathbf{c}(\mathbf{x}))\bar{P}_{\text{can}}^{\mathbf{c}}(\mathbf{c}(\mathbf{x})) = P_{\text{mic}}(\mathbf{x};\mathbf{c}(\mathbf{x}))\bar{P}_{\text{can}}^{\mathbf{c}}(\mathbf{c}(\mathbf{x})) \tag{S23}$$
where $\bar{P}_{\text{can}}^{\mathbf{c}}(\mathbf{c})$ is the distribution of $\mathbf{c}$ induced by $\bar{P}_{\text{can}}$. The proof follows by comparing the expression above with the general expression of the microcanonical Bayesian marginal likelihood (S5) and by setting $W_{\text{can}}(\mathbf{c}) = \bar{P}_{\text{can}}^{\mathbf{c}}(\mathbf{c})$.

Proof 2

The canonical Bayesian marginal likelihood $\bar{P}_{can}^{w}$
$$P_{\text{can}}^{w}(\mathbf{x}) = \int_{\boldsymbol{\Theta}} P_{\text{can}}(\mathbf{x};\boldsymbol{\theta})w(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{S24}$$
is the weighted sum of positive functions; indeed, $P_{\text{can}}(\mathbf{x};\boldsymbol{\theta})$ is an exponential function that assigns positive probability to all $\mathbf{x} \in \mathcal{X}$. As such, for all proper choices of the prior density $w(\boldsymbol{\theta})$, $\bar{P}_{\text{can}}^{w}$ is strictly positive everywhere:
$$\bar{P}_{can}^{w}(x) > 0 \quad \forall x \in \mathcal{X}. \tag{S25}$$
Consider a microcanonical prior $W(\mathbf{c})$ such that $W(\mathbf{c}^*) = 0$ for a certain value $\mathbf{c}^*$ of the sufficient statistics. Consequently,
$$\bar{P}_{\text{mic}}^{W}(x^*) = 0 \quad \forall\, x : \mathbf{c}(\mathbf{x}) = \mathbf{c}^*. \tag{S26}$$
Thus, there is no choice of canonical prior $w(\boldsymbol{\theta})$ s.t. $\bar{P}_{\text{mic}}^{W} = \bar{P}_{\text{can}}^{w}$.

S4. PROOF OF THEOREM 1

In the proof, we freely use well-known properties of exponential families as described by, e.g., [2]. Set $B := \mathtt{M}'$. Fix a constant a and consider the sets, for $m = 1, 2, \ldots$:
$$B_m^+ = \left\{ \boldsymbol{\mu} : \inf_{\boldsymbol{\mu}' \in B} \|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_2^2 \leq \frac{a \log m}{m} \right\}, B_m^- = \left\{ \boldsymbol{\mu} \in B : \inf_{\boldsymbol{\mu}' \notin B} \|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_2^2 \geq \frac{a \log m}{m} \right\}.$$
B_m^+ is a superset of B, including a small region (whose volume tends to 0 with sample size) just outside B's boundary; similarly B_m^- excludes a small region just inside B's boundary. To shorten notation we write $\hat{\boldsymbol{\mu}} := \mathbf{s}(\mathbf{y}^{(m)})/m$ and

$P^w := P_{\text{can}}^{w(m)}$ and, for any measurable subset $B' \subset \mathtt{M}$, $Q^w(B') := Q^w(\boldsymbol{\mu} \in B')$, and $P_{\boldsymbol{\mu}} := P_{\text{can}}(\cdot; \boldsymbol{\theta}(\boldsymbol{\mu}))$ where $\boldsymbol{\theta}(\boldsymbol{\mu})$ is the mapping from mean-value parameters to corresponding canonical parameters, i.e. the inverse of the (1-to-1) mapping $\boldsymbol{\mu}(\boldsymbol{\theta})$ defined in the main text. We have:

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

and also

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

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

a. Supremum in (S28) To bound the supremum in (S28), note first that B is a convex set. We can therefore use Csiszár's [3] multivariate generalization of Chernoff's concentration inequality (see [4] for general discussion) to get that

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

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

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

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

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

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

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

Here the second inequality is the union bound and the third is once again Csiszár's [3] multivariate generalization of Chernoff's concentration inequality (in (S29), we bounded $P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in B)$ where B was a convex set, allowing us to use Csiszár's result directly; but in (S31), we need to bound $P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \notin B) = P_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}} \in \mathtt{M} \setminus B)$; since $\mathtt{M} \setminus B$ is not a convex set, yet convexity is required by Csiszŕ's result, we now first need to cover it by $2d$ convex sets $H_{\boldsymbol{\mu},\cdot,\epsilon_m}^{\cdot}$, for each of which we then use Csiszár's result). The final inequality follows by using (S30) again.

S5. PSEUDO APPROXIMATION THROUGH THE HIGH RESOLUTION LIMIT

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

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

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

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

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

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

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

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

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

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

S6. BOUND ON REGRET

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

Let $\mathcal{M}_0$ be a maximum entropy model. We begin by recalling an extension of the classical $(d/2)\log m$ asymptotic redundancy result (see Equation 19 of the main text) that remains valid even when the model $\mathcal{M}_0$ is misspecified — i.e., it does not contain the true distribution P^*. This generalization appears in [5], and is formalized in [6, Proposition 3].

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

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

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

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

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

We now apply this general result in the setting where $P^* = P_{\boldsymbol{\theta}_1} \in \mathcal{M}_1$, in order to bound the regret $\mathrm{REG}(\boldsymbol{\theta}_1; P_1^{w_1})$ of the Bayesian marginal $P_1^{w_1}$ for a regular prior w_1. Consider the pseudo-e-variable S_{pseudo}, used as a proxy for either $S_{\mathrm{mic}}^{\mathrm{GRO}}$ or $S_{\mathrm{can}}^{\mathrm{GRO}}$. Using the previous result to refine equation Eq. 21 of the main text, we obtain:

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

The first two equalities are direct, and the third follows from [6, Theorem 3 (Parts 3,4)]. Equality (a) holds provided that $w_{\mathrm{pseudo},0}$ is regular, in particular, that it has a continuous density.

S7. TESTING WITH THE NML UNIVERSAL DISTRIBUTION

In this section, we propose an alternative to the Bayesian approach, based on the *Normalized Maximum Likelihood distribution*. In a variation of the definition of universality given in the main text, we may search for $\bar{P}_j$ such that $-\log \bar{P}_j(\mathbf{x}) + \log P_{\hat{\boldsymbol{\theta}}_j}(\mathbf{x})$ is small, in the worst-case over all possible data realizations $\mathbf{x}$; here $\hat{\boldsymbol{\theta}}_j(\mathbf{x})$ is the maximum likelihood estimator of $\boldsymbol{\theta}_j$ relative to $\mathbf{x}$. This leads to comparing $\bar{P}_j(\mathbf{x})$ to the best-fitting model *a posteriori*, that is, the model that, with hindsight, would give the shortest code for the observed data.

Within the MDL literature, this more stringent criterion is often viewed as the ideal one. The distribution achieving this is called the Normalized Maximum Likelihood (NML) distribution [5, 7, 8]. For each observed dataset $\mathbf{x}$, this distribution assigns its maximum likelihood value, normalized over all possible datasets in $\mathcal{X}$:

$$\bar{P}_j^{\mathrm{NML}}(\mathbf{x}) = \frac{P_j(\mathbf{x}; \hat{\boldsymbol{\theta}}_j(\mathbf{x}))}{\sum_{\mathbf{y} \in \mathcal{X}} P_j(\mathbf{y}; \hat{\boldsymbol{\theta}}_j(\mathbf{y}))}. \tag{S34}$$

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

In what follows, we explicitly compute NML-based GRO e-variables for contingency tables, introduced in the main text, starting from the 2×2 case and generalizing to the $2 \times k$ case. In the microcanonical case, resorting to

the Normalized Maximum Likelihood approach is equivalent to putting a uniform prior on both parameters of the alternative model, as shown in [1]. Consequently, this case is reduced to Example A of the main text and is not considered further. Instead, we focus on computing the microcanonical approximation of the canonical GRO e-variable.

Canonical test with NML. We first consider the case of a 2×2 contingency table, where we test between two canonical models through an NML approach, i.e., $\bar{P}_{\text{can},1} = P^{\text{NML}}_{\text{can},1}$. For a model with two independent Bernoulli distributions, the exact expression of the NML distribution reads [1, 9]:

$$P^{\text{NML}}_{\text{can},1}(\mathbf{x}) = P^{\text{a, NML}}_{\text{can},1}(\mathbf{x}) \cdot P^{\text{b, NML}}_{\text{can},1}(\mathbf{x}) \tag{S35}$$

with

$$P^{\text{i, NML}}_{\text{can},1}(\mathbf{x}) = \frac{\left(\frac{n^i_1(\mathbf{x})}{n^i}\right)^{n^i_1(\mathbf{x})} \left(1 - \frac{n^i_1(\mathbf{x})}{n^i}\right)^{n^i - n^i_1(\mathbf{x})}}{\frac{e^{n^i}\Gamma(n^i, n^i)}{(n^i)^{n^i - 1}} + 1} \tag{S36}$$

for $i \in \{a, b\}$. In the formula above, $\Gamma(s, t)$ is the upper incomplete gamma function. According to the procedure described in the main text, a natural candidate e-variable is given by the microcanonical approximation i.e., the GRO e-variable of the corresponding microcanonical test. To build it, we first need the distribution of the alternative sufficient statistics (n^a_1, n^b_1) induced by $P^{\text{i, NML}}_{\text{can},1}$, which is

$$W_{\text{can},1}(n^a_1, n^b_1) = W^a_{\text{can},1}(n^a_1) \cdot W^b_{\text{can},1}(n^b_1) \tag{S37}$$

with

$$W^a_{\text{can},1}(n^a_1) = \Omega^a_1(n^a_1) \cdot \frac{\left(\frac{n^a_1(\mathbf{x})}{n^a}\right)^{n^a_1(\mathbf{x})} \left(1 - \frac{n^a_1(\mathbf{x})}{n^a}\right)^{n^a - n^a_1(\mathbf{x})}}{\frac{e^{n^a}\Gamma(n^a, n^a)}{(n^a)^{n^a - 1}} + 1} \tag{S38}$$

and

$$W^b_{\text{can},1}(n^b_1) = \Omega^b_1(n^b_1) \cdot \frac{\left(\frac{n^b_1(\mathbf{x})}{n^b}\right)^{n^b_1(\mathbf{x})} \left(1 - \frac{n^b_1(\mathbf{x})}{n^b}\right)^{n^b - n^b_1(\mathbf{x})}}{\frac{e^{n^b}\Gamma(n^b, n^b)}{(n^b)^{n^b - 1}} + 1}. \tag{S39}$$

Given that n^a_1 and n^b_1 are independently distributed, $W^*_0(n_1)$ is the convolution of $W^a_{\text{can},1}(n^a_1)$ and $W^b_{\text{can},1}(n^b_1)$, which can be computed numerically. Once that $W^*_0(n_1)$ is computed, the microcanonical approximation is evaluated through Eq. 40 of the main text. The accuracy of the microcanonical approximation is assessed by directly inspecting the difference in e-power (Figure S2) and through the interval width r (bottom row of Figure S3), which is shown to be already close to 0 for small sample sizes and to converge towards 0 as the sample size increases.

Notice that, if n^a_1 and n^b_1 are both large enough, using $P^{\text{NML}}_{\text{can},1}$ is asymptotically equivalent to using a Bayesian universal distribution with a *Jeffreys prior* [5] on the alternative parameters. The Jeffreys prior, defined as

$$w_J(\boldsymbol{\theta}) \propto \sqrt{\det I(\boldsymbol{\theta})}, \tag{S40}$$

where $I(\boldsymbol{\theta})$ is the Fisher information matrix, is invariant under any smooth reparametrization of $\boldsymbol{\theta}$ and is therefore commonly used as a canonical non-informative prior. For a Bernoulli model, such as ours, the Jeffreys prior is equal to a beta prior with parameters all equal to $\gamma = 0.5$:

$$w_J(p_a, p_b) = w^a_J(p_a) w^b_J(p_b) = \frac{1}{\pi^2} \frac{1}{\sqrt{p_a(1-p_a)}} \frac{1}{\sqrt{p_b(1-p_b)}}. \tag{S41}$$

Asymptotically, and in an expected-regret sense, the NML distribution coincides with a Bayesian universal distribution based on Jeffreys prior when the true parameter lies in the interior of $\mathtt{M}_1 = [0, 1]^2$.

More explicitly, let us set $n^a = n^b = m$. Then, for any INECCSI subset $\mathtt{M}'_1 \subset \mathtt{M}_1$, as $m \to \infty$:

$$\sup_{\boldsymbol{\mu}_1 \in \mathtt{M}_1} \mathbb{E}_{\boldsymbol{\mu}_1} \left[-\log P^{w_J}_{\text{can},1}(\mathbf{x}^m) + \log P^{\text{NML}}_{\text{can},1}(\mathbf{x}^m) \right] = o(1), \tag{S42}$$

where $o(1)$ denotes a quantity that goes to 0 as $m \to +\infty$. At the boundaries of the parameter space, however, the two constructions differ: Jeffreys prior, which assigns densities that diverge to infinity at the boundaries, leads to substantially different data probabilities than the NML-induced distribution. This difference becomes even more important when convoluting the independent Jeffreys priors to compute W_0^*. This explains why the first and third pictures in Figure S2 are quite different. For this reason, in all simulations, we implement the exact NML formula instead of its Jeffreys approximation.

We now extend the solution to $2 \times k$ contingency tables. For a model with k independent Bernoulli distributions, the NML distribution reads [1]:

$$P_{\text{can},1}^{\text{NML}}(\mathbf{x}) = \prod_{i=1}^{k} P_{\text{can},1}^{\text{i, NML}}(\mathbf{x}) \tag{S43}$$

where $P_{\text{can},1}^{\text{i, NML}}$ is given by Eq. (S36). The microcanonical approximation is obtained by defining

$$W_{\text{can},1}(\{n_1^i\}) = \prod_{i=1}^{k} W_{\text{can},1}^i(n_1^i) \tag{S44}$$

with

$$W_{\text{can},1}^i(n_1^i) = \Omega_1^i(n_1^i) \cdot \frac{\left(\frac{n_1^i(\mathbf{x})}{n^i}\right)^{n_1^i(\mathbf{x})} \left(1 - \frac{n_1^i(\mathbf{x})}{n^i}\right)^{n^i - n_1^i(\mathbf{x})}}{\frac{e^{n^i}\Gamma(n^i, n^i)}{(n^i)^{n^i - 1}} + 1}. \tag{S45}$$

W_0^* is then the convolution of all $W_{\text{can},1}^i(n_1^i)$, which can be computed numerically or by resorting to the Gaussian approximation described in the main text. Once that $W_0^*(n_1)$ is computed, the microcanonical approximation is evaluated as described in the main text.

S8. SUPPLEMENTARY TABLES AND FIGURES

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

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

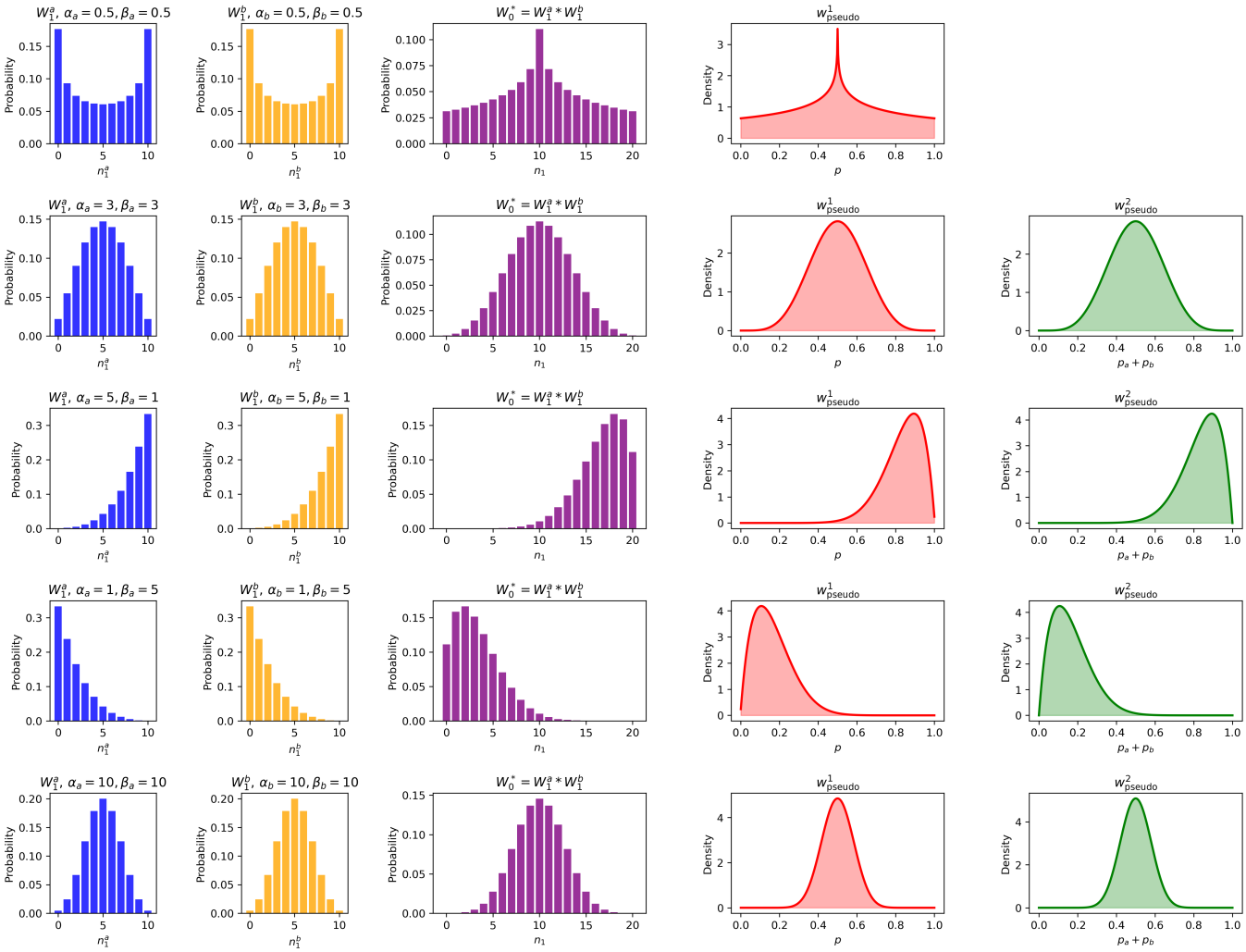

FIG. S1. Construction of the microcanonical optimal prior W_0^* (third column) and the pseudo-prior approximations $w^1_{\text{pseudo},0}$ (fourth column) and $w^2_{\text{pseudo},0}$ (fifth column), for 2×2 contingency tables with independent beta priors on the alternative. Starting from the induced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

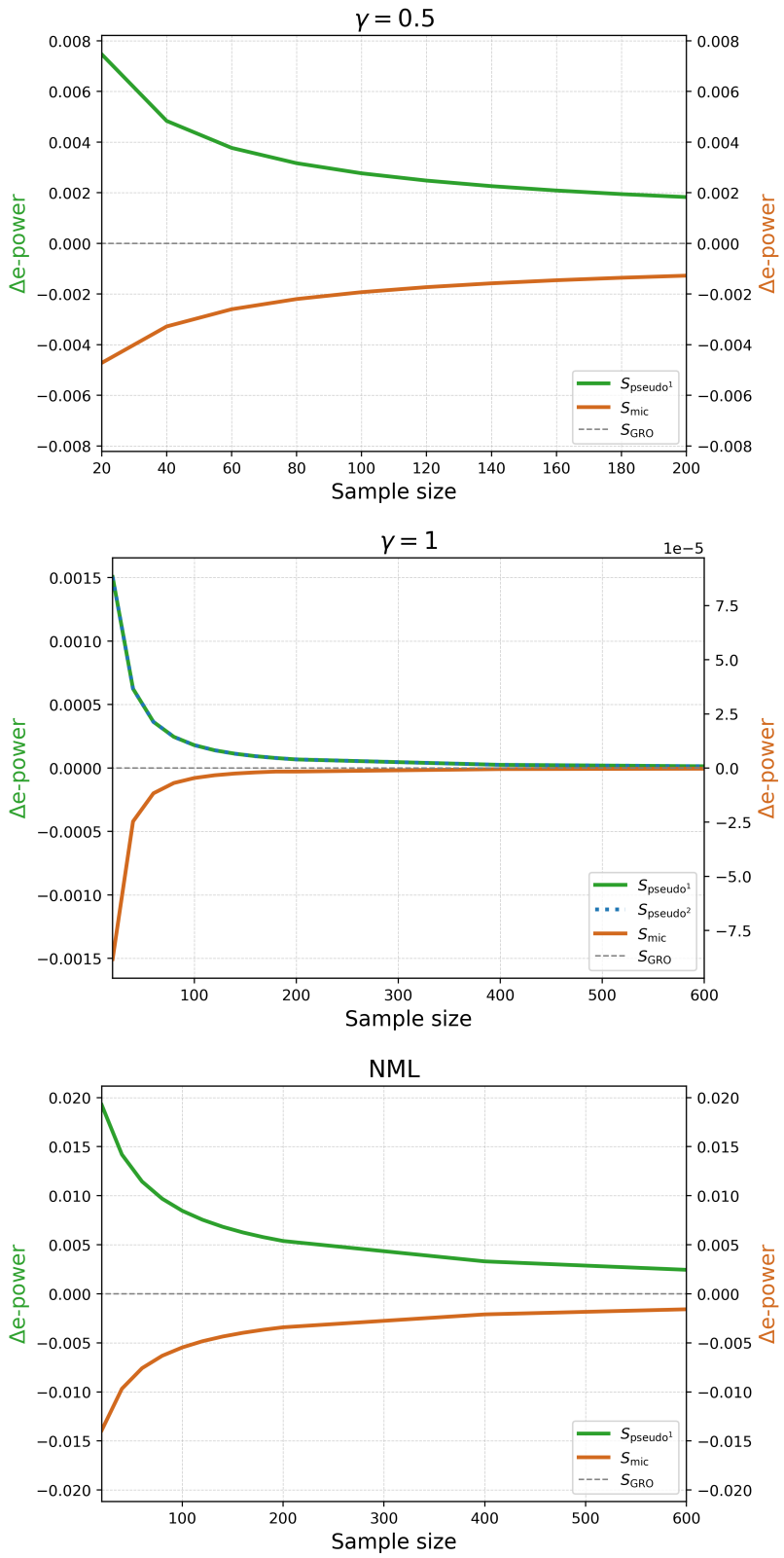

FIG. S2. E-power difference between the canonical GRO e-variable (computed numerically), its microcanonical approximation (orange curve), and the pseudo approximation (green curve), across sample sizes and different choices on the alternative (NML and beta with all parameters equal to γ). The microcanonical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

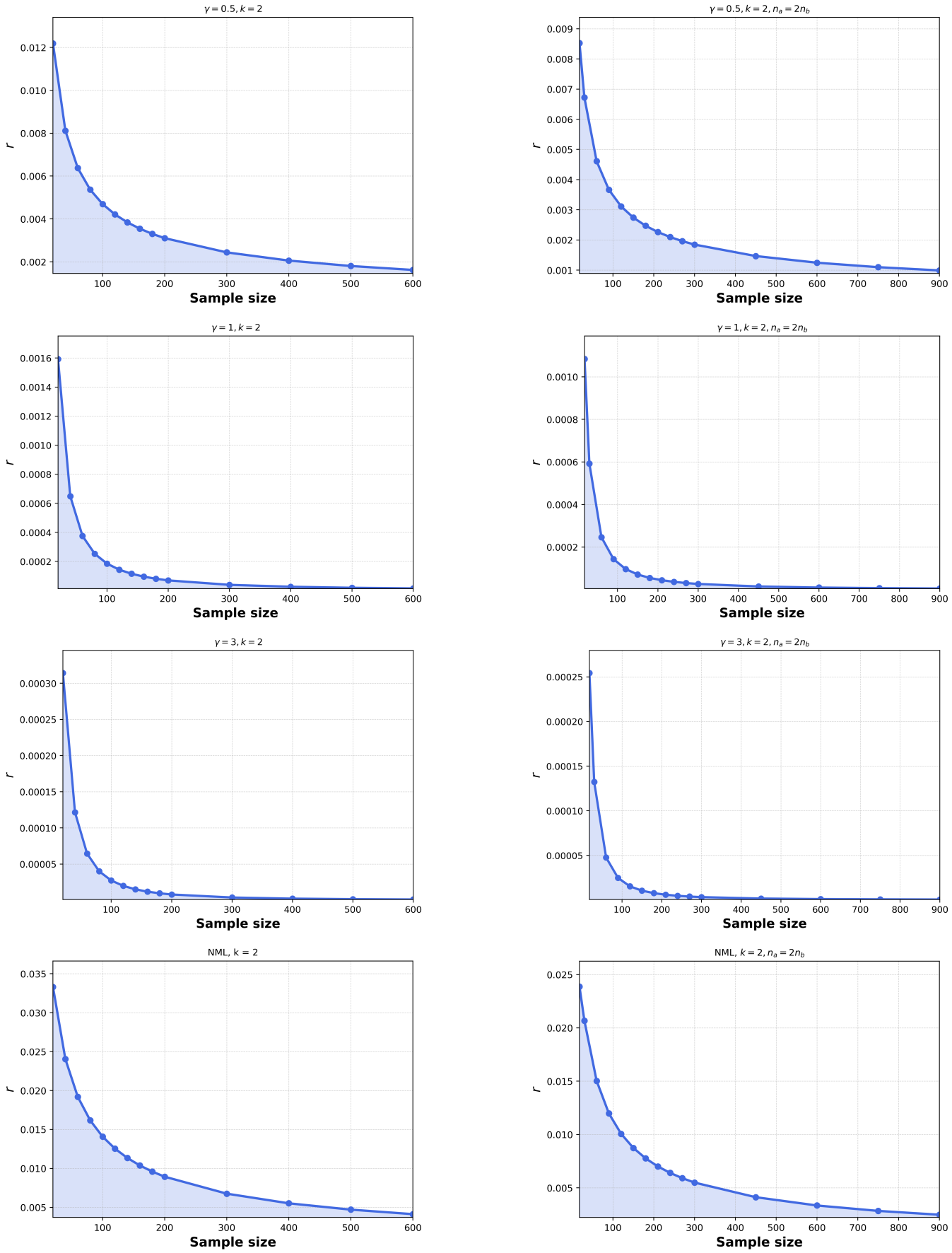

FIG. S3. Convergence of r in 2×2 contingency tables, for $n^a = n^b$ (left column) and $n^a = 2n^b$ (right column), as the sample size $n = n^a + n^b$ grows. Results are shown for different choices of $\bar{P}_{\text{can},1}$: beta independent priors with all parameters equal to $\gamma = 0.5$ (first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

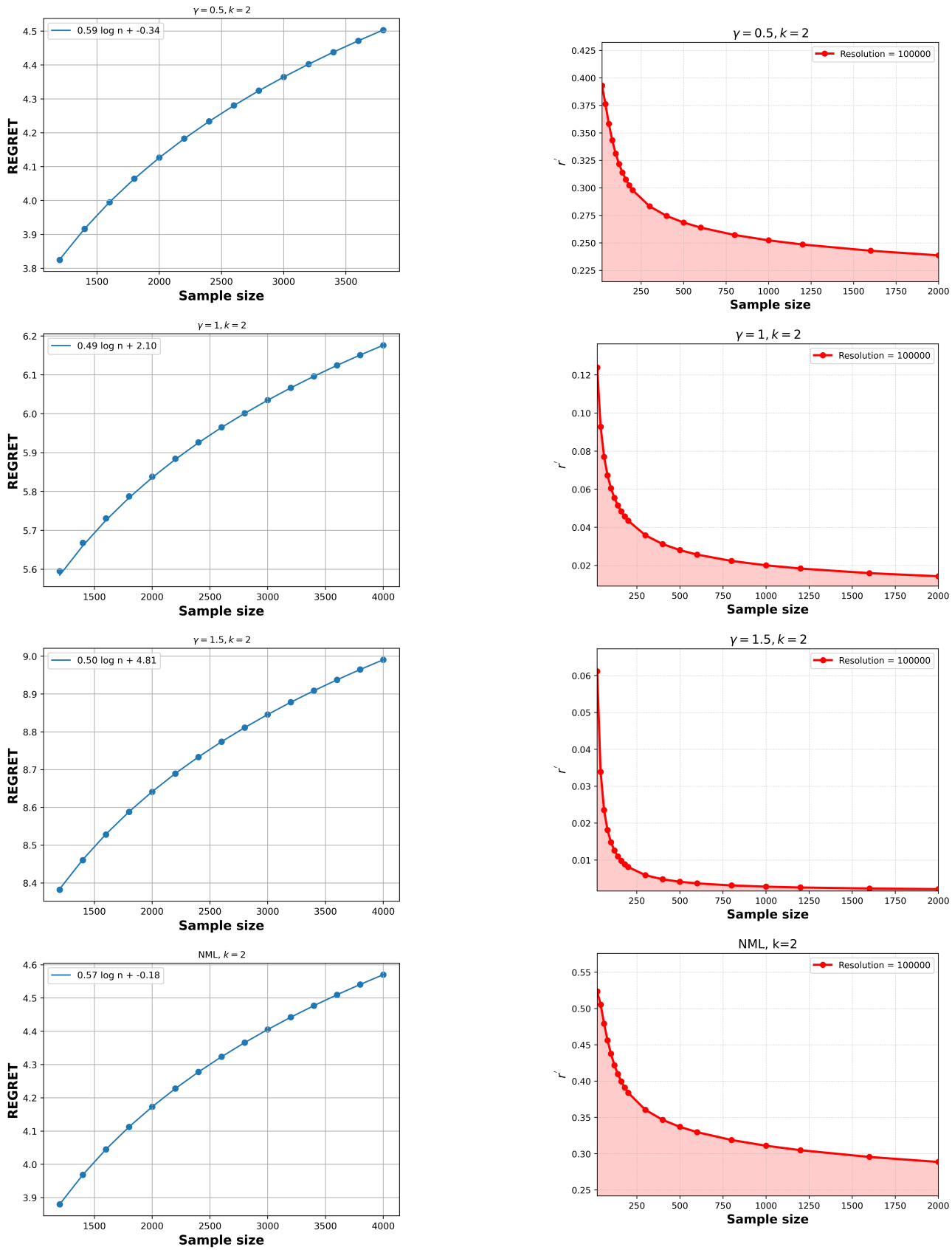

FIG. S4. Worst case regret (left) and convergence of r' (right), for 2×2 tables with $n^a = n^b = m$ and sample size equal to $n = 2m$. Different choices of the alternative are considered: Bayesian with identical independent beta priors $B(\gamma, \gamma)$ with $\gamma = 0.5$, 1, 1.5, and NML. r' is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

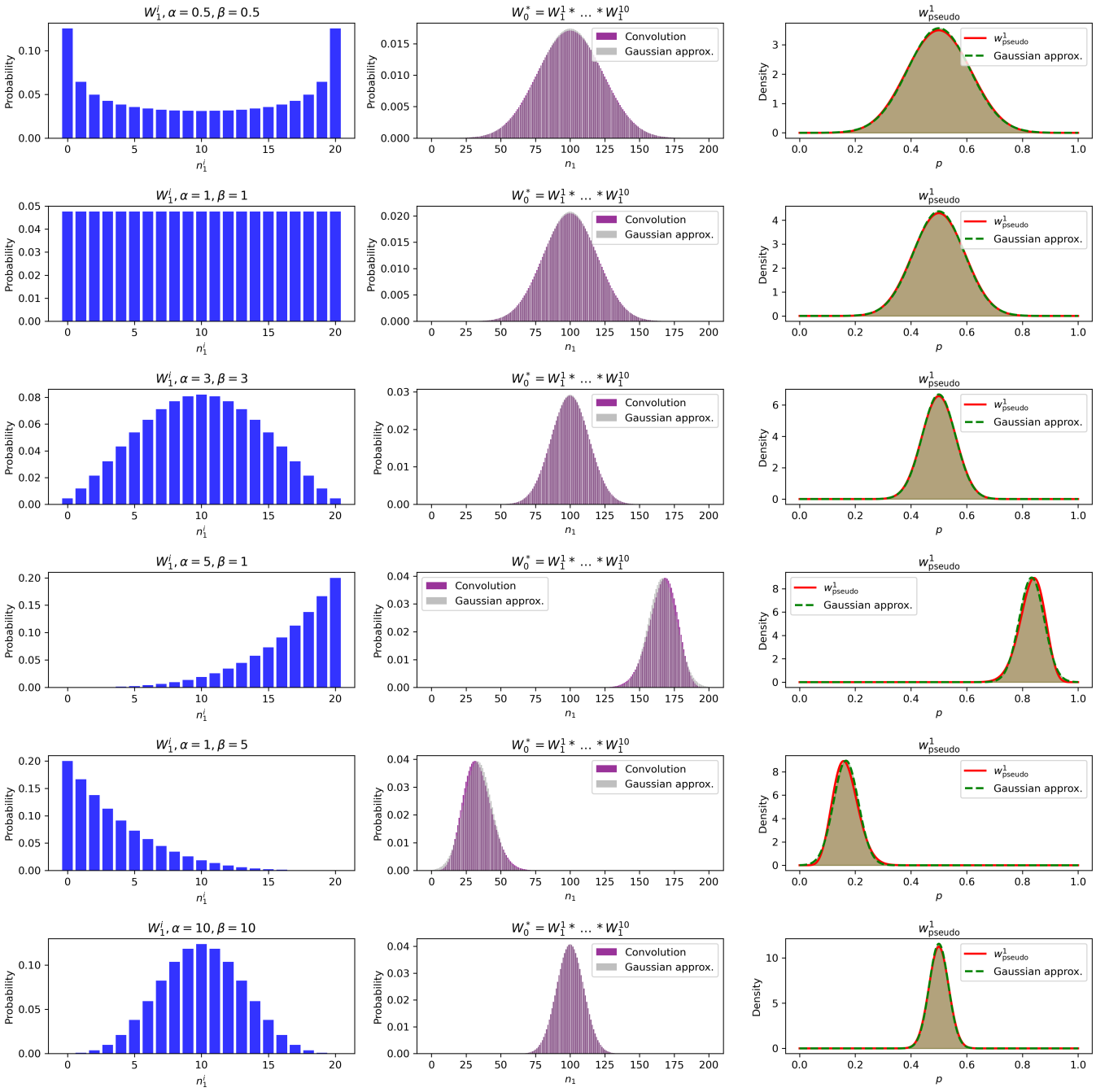

FIG. S5. The first column shows the discrete distribution of each component of the alternative sufficient statistics, here denoted simply by W_1^i, induced by independent identical beta priors on the alternative, for different prior parameters. The GRO-optimal microcanonical prior on the null W_0^* (second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

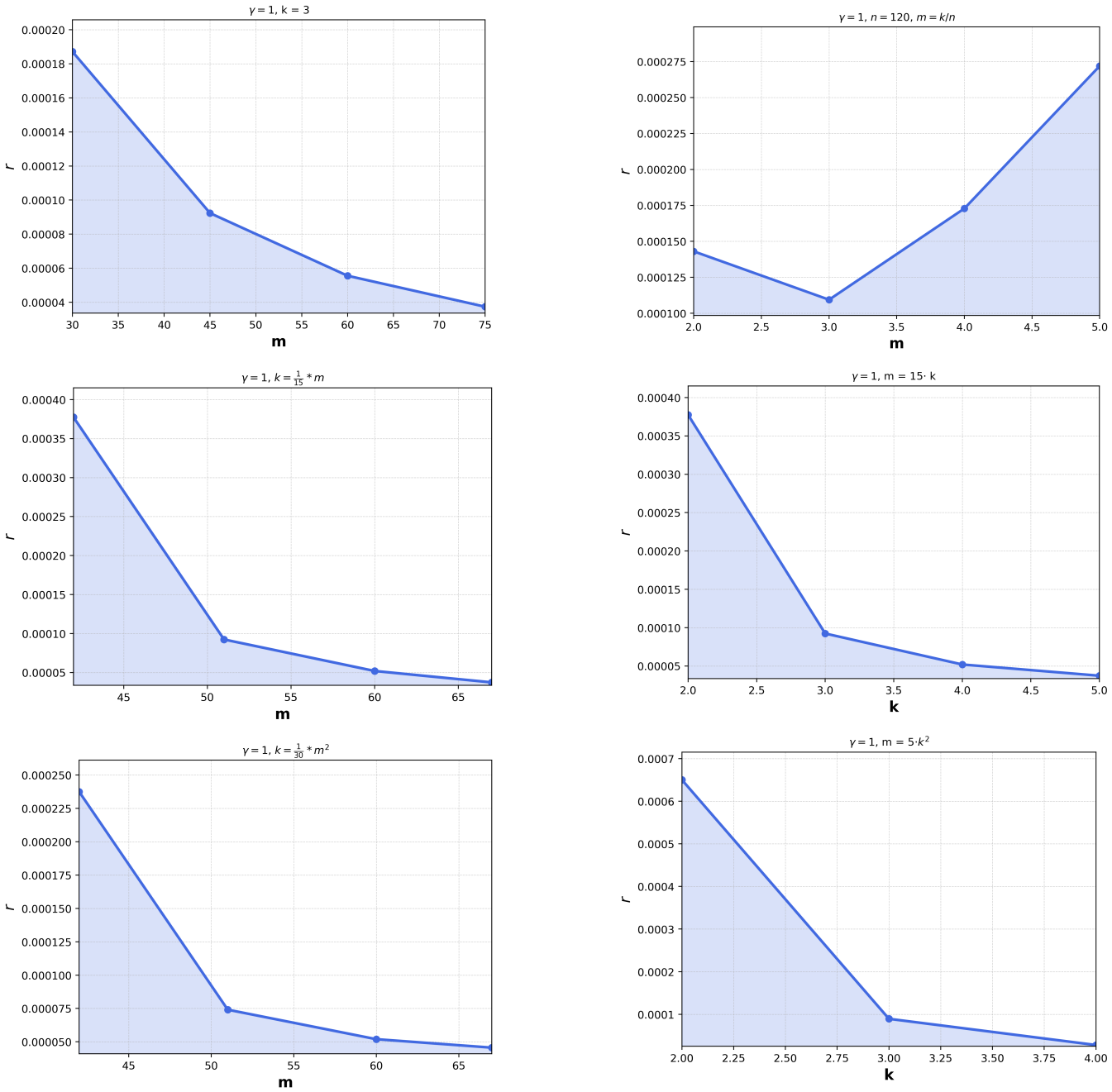

FIG. S6. Convergence to 0 of the interval width r in the case of $2 \times k$ contingency tables, where all groups have same size m, for different interplays between the number of groups k and the size of each group m. Results are shown for identical independent beta priors on the alternative, with all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

[1] F. Giuffrida, T. Squartini, P. Grünwald, and D. Garlaschelli, Description length of canonical and microcanonical models, Phys. Rev. Res. 10.1103/1sd2-rxmy (2025).

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

[3] I. Csiszár, Sanov property, generalized I-projection and a conditional limit theorem, The Annals of Probability , 768 (1984).

[4] P. Grünwald, Y. Hao, and A. Balsubramani, Growth-optimal e-variables and an extension to the multivariate Csiszár-Sanov-Chernoff theorem, arXiv: 2412.17554 (2024), arXiv:2412.17554 [cs.IT].

[5] P. Grünwald, *The minimum description length principle* (MIT press, 2007).

[6] Y. Hao and P. Grünwald, E-values for exponential families: the general case, arXiv: 2409.11134 (2024), arXiv:2409.11134 [stat.ME].

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

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

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