E-values for anytime-valid inference with exponential families
Hao, Y.

Chapter 4

E-values for k-Sample Tests With Exponential Families

Abstract

We develop and compare e-variables for testing whether k samples of data are drawn from the same distribution, the alternative being that they come from different elements of an exponential family. We consider the GRO (growth-rate optimal) e-variables for (1) a 'small' null inside the same exponential family, and (2) a 'large' nonparametric null, as well as (3) an e-variable arrived at by conditioning on the sum of the sufficient statistics. (2) and (3) are efficiently computable, and extend ideas from [88] and [93] respectively from Bernoulli to general exponential families. We provide theoretical and simulation-based comparisons of these e-variables in terms of their logarithmic growth rate, and find that for small effects all four e-variables behave surprisingly similarly; for the Gaussian location and Poisson families, e-variables (1) and (3) coincide; for Bernoulli, (1) and (2) coincide; but in general, whether (2) or (3) grows faster under the alternative is family-dependent. We furthermore discuss algorithms for numerically approximating (1).

4.1 Introduction

E-variables (and the value they take, the *e-value*) provide an alternative to p-values that is inherently more suitable for testing under optional stopping and continuation, and that lies at the basis of *anytime-valid* confidence intervals that can be monitored continuously [42, 91, 76, 71, 46, 36]. While they have their roots in the work on anytime-valid testing by H. Robbins and students (e.g. [28]), they have begun to be investigated in detail for composite null hypotheses only very recently. E-variables can be associated with a natural notion of optimality, called GRO (growth-rate optimality),

---

[0] This chapter is based on Yunda Hao, Peter Grünwald, Tyron Lardy, Long Long, and Reuben Adams. E-values for k-Sample Tests with Exponential Families. Sankhya A, 86(1):596–636, 2024.

introduced and studied in detail by [42]. GRO may be viewed as an analogue of the uniformly most powerful test in an optional stopping context. In this paper, we develop GRO and near-GRO e-variables for a classical statistical problem: parametric k-sample tests. Pioneering work in this direction appears already in [93]: as we explain in Example 9, his SPRT for a sequential test of two proportions can be re-interpreted in terms of e-values for Bernoulli streams. Wald's e-values are not optimal in the GRO sense — GRO versions were derived only very recently by [88, 86], but again only for Bernoulli streams. Here we develop e-variables for the case that the alternative is associated with an arbitrary but fixed exponential family, $\mathcal{M}$, with data in each of the k groups sequentially sampled from a different distribution in that family. We mostly consider tests against the null hypothesis, denoted by $\mathcal{H}_0(\mathcal{M})$ that states that outcomes in all groups are i.i.d. by a single member of $\mathcal{M}$. We develop the GRO e-variable S_{RIPR} for this null hypothesis, but it is not efficiently computable in general. Therefore, we introduce two more tractable e-variables: S_{MIX} and S_{COND}. The former is defined as the GRO e-variable, for the much larger null hypothesis that the k groups are i.i.d. from an arbitrary distribution, denoted by $\mathcal{H}_0(\mathrm{IID})$: since an e-variable relative to a null hypothesis $\mathcal{H}_0$ is automatically an e-variable relative to any null that is a subset of $\mathcal{H}_0$, S_{MIX} is automatically also an e-variable relative to $\mathcal{H}_0(\mathcal{M})$. Whenever below we refer to 'the null', we mean the smaller $\mathcal{H}_0(\mathcal{M})$. The use of S_{MIX} rather than S_{RIPR} for this null, for which it is not GRO, is justifiable by ease of computation and robustness against misspecification of the model $\mathcal{M}$. However, exactly this robustness might also cause it to be too conservative when $\mathcal{M}$ is well-specified. The third e-variable we consider, S_{COND}, does not have any GRO status, but is specifically tailored to $\mathcal{H}_0(\mathcal{M})$, so that it might still be better than S_{MIX} in practice. Finally, we introduce a pseudo-e-variable S_{PSEUDO}, which coincides with S_{RIPR} whenever the latter is easy to compute; in other cases it is not a real e-variable, but it is still highly useful for our theoretical analysis.

Results Besides defining S_{RIPR}, S_{MIX} and S_{COND} and proving that they achieve what they purport to, we analyze their behaviour both theoretically and by simulations. Our main theoretical results, Theorem 7 and 8 reveal some surprising facts: for any exponential family, the four types of (pseudo-) e-variables achieve almost the same growth rate under the alternative, hence are almost equally good, whenever the 'distance' between null and alternative is sufficiently small. That is, suppose that the (shortest) ℓ_2-distance between the k dimensional parameter of the alternative and the parameter space of the null is given by δ. Then for any two of the aforementioned e-variables S, S', we have $\mathbb{E}[\log S - \log S'] = O(\delta^4)$, where the expectation is taken under the alternative. Here, $\mathbb{E}[\log S]$ can be interpreted as the growth rate of S, as explained in Section 4.1.1.

While S_{MIX} and S_{COND} are efficiently computable for the families we consider, this is generally not the case for S_{RIPR}, since to compute it we need to have access to the *reverse information projection* (RIPr; [60, 42]) of a fixed simple alternative to the set $\mathcal{H}_0(\mathcal{M})$. In general, this is a convex combination of elements of $\mathcal{H}_0(\mathcal{M})$, which can only be found by numerical means. Interestingly, we find that for three families, Gaussian with fixed variance, Bernoulli and Poisson, the RIPr is attained at a single point (i.e. a mixture putting all its mass on that point) that can be efficiently computed.

Furthermore, in these cases S_{RIPR} coincides with one of the other e-variables (S_{MIX} for Bernoulli, S_{COND} for Gaussian and Poisson). For other exponential families, for $k = 2$, we approximate the RIPr and hence S_{RIPR} using both an algorithm proposed by Li [60] and a brute-force approach. We find that we can already get an extremely good approximation of the RIPr with a mixture of just *two* components. This leads us to conjecture that perhaps the deviation from the RIPr is just due to numerical imprecision and that the actual RIPr really can be expressed with just two components. The theoretical interest of such a development notwithstanding, we advise to use S_{COND} or S_{MIX} rather than S_{RIPR} for practical purposes whenever more than one component is needed for the RIPr, as their growth rates are not much worse, and they are much easier to compute. If furthermore robustness against misspecification of the null is required, then S_{MIX} is the most sensible choice.

Method: Restriction to Single Blocks and Simple Alternatives   The main interest of e-variables is in analyzing sequential, anytime-valid settings: the data arrives in k streams corresponding to k groups, and we may want to stop or continue sampling at will (optional stopping); for example, we only stop when the data looks sufficiently good; or we stop unexpectedly, because we run out of money to collect new data. Nevertheless, in this paper we focus on what happens in a single *block*, i.e. a vector $X^k = (X_1, \ldots, X_k)$, where each X_j denotes a single outcome in the j-th stream. By now, there are a variety of papers (see e.g. [42, 71, 88]) that explain how e-variables defined for such a single block can be combined by multiplication to yield e-processes (in our context, coinciding with *nonnegative supermartingales*) that can be used for testing the null with optional stopping if blocks arrive sequentially — that is, one observes one outcome of each sample at a time. Briefly, one multiplies the e-variables and at any time one intends to stop, one rejects the null if the product of e-values observed so-far exceeds $1/\alpha$ for pre-specified significance level α. This gives an *anytime-valid* test at level α: irrespective of the stopping rule employed, the Type-I error is guaranteed to be below α. Similarly, one can extend the method to design *anytime-valid confidence intervals* by inverting such tests, as described in detail by [71]. This is done for the 2-sample test with Bernoulli data by [86]; their inversion methods are extendable to the general exponential family case we discuss here. Thus, we refer to the aforementioned papers for further details and restrict ourselves here to the 1-block case. Also, [88, 87] describe how one can adapt an e-process for data arriving in blocks to general streams in which the k streams do not produce data points at the same rate; we briefly extend their explanation to the present setting in Appendix 4.A. Finally, we mainly restrict to the case of a simple alternative, i.e. a single member of the exponential family under consideration. While this may seem like a huge restriction, extension from simple to composite alternatives (e.g. the full family under consideration) is straightforward using the *method of mixtures* (i.e. Bayesian learning of the alternative over time) and/or the plug-in method. We again refer to [42, 71] for detailed explanations, and [88] for an explanation in the 2-sample Bernoulli case, and restrict here to the simple alternative case: all the 'real' difficulty lies in dealing with composite null hypotheses, and that, we do explicitly and exhaustively in this paper.

Related Work and Practical Relevance   As indicated, this paper is a direct (but far-reaching) extension of the papers [88, 86] on 2-sample testing for Bernoulli streams as well as Wald's [93] sequential two-sample test for proportions to streams coming from an exponential family. There are also *nonparametric* sequential [59] and anytime-valid 2-sample tests [6, 68] that tackle a somewhat different problem. They work under much weaker assumptions on the alternative (in some versions the samples could be arbitrary high-dimensional objects such as pictures and the like). The price to pay is that they will need a much larger sample size before a difference can be detected. Indeed, while our main interest is theoretical (how do different e-variables compare? in what sense are they optimal?), in settings where data are expensive, such as randomized clinical trials, the methods we describe here can be practically very useful: they are exact (existing methods are often based on chi-squared tests, which do not give exact Type-I error guarantees at small sample size), they allow for optional stopping, and they need small amounts of data due to the strong parametric assumptions for the alternative. As a simple illustration of the practical importance of these properties, we refer to the recent SWEPIS study [97] which was stopped early for harm. As demonstrated by [88], if an anytime-valid two-sample test had been used in that study, substantially stronger conclusions could have been drawn.

    We also mention that k-sample tests can be viewed as independence tests (is the outcome independent of the group it belongs to?) and as such this paper is also related to recent papers on e-values and anytime-valid tests for conditional independence testing [39, 75, 31]. Yet, the setting studied in those papers is quite different in that they assume the covariates (i.e. indicator of which of the k groups the data belongs to) to be i.i.d.

Contents   In the remainder of this introduction, we fix the general framework and notation and we briefly recall how e-variables are used in an anytime-valid/optional stopping setting. In Section 4.2 we describe our four (pseudo-) e-variables in detail, and we provide preliminary results that characterize their behaviour in terms of growth rate. In Section 4.3 we provide our main theoretical results which show that, for all regular exponential families, the expected growth of the four types of e-variables is of surprisingly small order δ^4 if the parameters of the alternative are at ℓ_2-distance δ to the parameter space of the null. In Section 4.4 we give more detailed comparisons for a large number of standard exponential families (Gaussian, Bernoulli, Poisson, exponential, geometric, beta), including simulations that show what happens if δ gets larger. Section 4.5 provides some additional simulations about the RIPr. All proofs, and some additional simulations, are in the appendix.

4.1.1   Formal Setting

Consider a regular one-dimensional exponential family $\mathcal{M} = \{P_\mu : \mu \in \mathtt{M}\}$ given in its mean-value parameterization (see e.g. [13] for more on definitions and for all the proofs of all standard results about exponential families that are to follow). Each member of the family is a distribution for some random variable U, taking values in some set $\mathcal{U}$, with density $p_{\mu;[U]}$ relative to some underlying measure $\rho_{[U]}$ which, without loss of

generality, can be taken to be a probability measure. For regular exponential families, $\mathbb{M}$ is an open interval in $\mathbb{R}$ and $p_{\mu;[U]}$ can be written as:

$$p_{\mu;[U]}(U) = \exp\left(\lambda(\mu) \cdot t(U) - A(\lambda(\mu))\right), \tag{4.1.1}$$

where $\lambda(\mu)$ maps mean-value μ to canonical parameter β. We then have $\mu = \mathbb{E}_{P_\mu}[t(U)]$, where $t(U)$ is a measurable function of U and $A(\beta)$ is the log-normalizing factor. The measure $\rho_{[U]}$ induces a corresponding (marginal) measure $\rho := \rho_{[X]}$ on the *sufficient statistic* $X := t(U)$, and similarly the density (4.1.1) induces a corresponding density $p_\mu := p_{\mu;[X]}$ on X, i.e. we have

$$p_\mu(X) := p_{\mu;[X]}(X) = \exp\left(\lambda(\mu) \cdot X - A(\lambda(\mu))\right). \tag{4.1.2}$$

All e-variables that we will define can be written in terms of the induced measure and density of the sufficient statistic of X; in other words, we can without loss of generality act as if our family is *natural*. Therefore, from now on we simply assume that we observe data in terms of their sufficient statistics X rather than the potentially more fine-grained U, and will be silent about U; for simplicity we thus abbreviate $p_{\mu;[X]}$ to p_μ and $\rho_{[X]}$ to ρ. Note that exponential families are more usually defined with a carrier function $h(X)$ and ρ set to Lebesgue or counting measure; we cover this case by absorbing h into ρ, which we do not require to be Lebesgue or counting.

The data comes in as a block $X^k = (X_1, \ldots, X_k) \in \mathcal{X}^k$, where $\mathcal{X}$ is the support of ρ. To calculate our e-values we only need to know $X^k \in \mathcal{X}^k$, and under the alternative hypothesis, all X_j, $j = 1 \ldots k$ are distributed according to some element P_{μ_j} of $\mathcal{M}$. In our main results we take the alternative hypothesis to be *simple*, i.e. we assume that $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k) \in \mathbb{M}^k$ is fixed in advance. The alternative hypothesis is thus given by

$$\text{simple } \mathcal{H}_1 : X_1 \sim P_{\mu_1}, X_2 \sim P_{\mu_2}, \ldots, X_k \sim P_{\mu_k} \text{ independent.}$$

Note that we will keep $\boldsymbol{\mu}$ fixed throughout the rest of this section and Section 4.2. This is without loss of generality as $\boldsymbol{\mu}$ is defined as an arbitrary element of $\mathbb{M}^k$, so that all results stated for $\boldsymbol{\mu}$ hold for any element of $\mathbb{M}^k$. The extension to composite alternatives by means of the method of mixtures or the plug-in method is straightforward, and done in a manner that has become standard for e-value based testing [71].

Our null hypothesis is directly taken to be composite, for as regards the null, the composite case is inherently very different from the simple case [71, 42]. It expresses that the X^k are identically distributed. We shall consider various variants of this null hypothesis, all composite: let $\mathcal{P}$ be a set of distributions on $\mathcal{X}$, then the null hypothesis *relative to $\mathcal{P}$*, denoted $\mathcal{H}_0(\mathcal{P})$, is defined as

$$\text{composite } \mathcal{H}_0(\mathcal{P}) : X_1 \sim P, X_2 \sim P, \ldots, X_k \sim P \text{ i.i.d. for some } P \in \mathcal{P}.$$

Our most important instantiation for the null hypothesis will be $\mathcal{H}_0 = \mathcal{H}_0(\mathcal{M})$ for the same exponential family $\mathcal{M}$ from which the alternative was taken; then $\mathcal{H}_0(\mathcal{M})$ is a one-dimensional parametric family expressing that the X_i are i.i.d. from P_{μ_0} for $\mu_0 \in \mathbb{M}$. Still, we will also consider $\mathcal{H}_0 = \mathcal{H}_0(\mathcal{P})$ where $\mathcal{P}$ is the much larger set of *all*

distributions on $\mathcal{X}$. Then the null simply expresses that the X^k are i.i.d.; we shall abbreviate this null to $\mathcal{H}_0(\text{IID})$. Finally we sometimes consider $\mathcal{H}_0 = \mathcal{H}_0(\mathcal{M}')$ where $\mathcal{M}' \subset \mathcal{M}$ is a subset of $P_\mu \in \mathcal{M}$ with $\mu \in \texttt{M}'$ for some sub-interval $\texttt{M}' \subset \texttt{M}$. The statistics that we use to gain evidence against these null hypotheses are e-variables.

Definition 2. We call any nonnegative random variable S on a sample space Ω (which in this paper will always be $\Omega = \mathcal{X}^k$) an *e-variable relative to* $\mathcal{H}_0$ if it satisfies

$$\text{for all } P \in \mathcal{H}_0 : \qquad \mathbb{E}_P[S] \leq 1. \tag{4.1.3}$$

4.1.2 The GRO E-variable for General $\mathcal{H}_0$

In general, there exist many e-variables for testing any of the null hypotheses introduced above. Each e-variable S can in turn be associated with a growth rate, defined by $\mathbb{E}_{P_\mu}[\log S]$. Roughly, this can be interpreted as the (asymptotic) exponential growth rate one would achieve by using S in consecutive independent experiments and multiplying the outcomes if the (simple) alternative was true (see e.g. [42, Section 2.1] or [52]). The Growth Rate Optimal (GRO) e-variable is then the e-variable with the greatest growth rate among all e-variables. The central result (Theorem 1) of [42] states that, under very weak conditions, GRO e-variables take the form of likelihood ratios between the alternative and the *reverse information projection* [60] of the alternative onto the null. We instantiate their Theorem 1 to our setting by providing Lemma 6 and 7, both special cases of their Theorem 1. Before stating these, we need to introduce some more notation and definitions. For $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$ we use the following notation:

$$p_{\boldsymbol{\mu}}(X^k) := \prod_{i=1}^{k} p_{\mu_i}(X_i).$$

Whenever in this text we refer to KL divergence $D(Q\|R)$, we refer to measures Q and R on $\mathcal{X}^k$. Here Q is required to be a probability measure, while R is allowed to be a sub-probability measure, as in [42]. A *sub-* probability measure R on $\mathcal{X}^k$ is a measure that integrates to 1 or less, i.e $\int_{x \in \mathcal{X}} dR(x) \leq 1$.

The following lemma follows as a very special case of Theorem 1 (simplest version) of [42], when instantiated to our k-sample testing set-up:

Lemma 6. *Let $\mathcal{P}$ be a set of probability distributions on $\mathcal{X}^k$ and let $\text{CONV}(\mathcal{P})$ be its convex hull. Then there exists a sub-probability measure P_0^* with density p_0^* such that*

$$D(P_{\boldsymbol{\mu}}\|P_0^*) = \inf_{P \in \text{CONV}(\mathcal{P})} D(P_{\boldsymbol{\mu}}\|P). \tag{4.1.4}$$

P_0^* *is called the* reverse information projection (RIPr) *of $P_{\boldsymbol{\mu}}$ onto* $\text{CONV}(\mathcal{P})$.

Clearly, if $P_0^* \in \text{CONV}(\mathcal{P})$ (the minimum is achieved) then P_0^* is a probability measure, i.e. integrates to exactly one. We show that this happens for certain specific exponential families in Section 4.4. However, in general we can neither expect the minimum to be achieved, nor the RIPr to integrate to one. Lemma 7 below, again a

special case of [42, Theorem 1], shows that the RIPr characterizes the GRO e-variable, and explains the use of the term GRO in the definition below.

Definition 3. $S_{\mathrm{GRO}(\boldsymbol{\mu}^1)}\mathcal{P}$ is defined as

$$S_{\mathrm{GRO}(\boldsymbol{\mu}^1)}\mathcal{P} := \frac{p_{\boldsymbol{\mu}}(X^k)}{p_0^*(X^k)} \tag{4.1.5}$$

where p_0^* is the density of the RIPr of $P_{\boldsymbol{\mu}}$ onto $\mathrm{CONV}(\mathcal{P})$.

Lemma 7. *For every set of distributions $\mathcal{P}$ on $\mathcal{X}$, $S_{\mathrm{GRO}(\boldsymbol{\mu}^1)}\mathcal{P}$ is an e-variable for $\mathcal{H}_0(\mathcal{P})$. Moreover, it is the GRO (Growth-Rate-Optimal) e-variable for $\mathcal{H}_0(\mathcal{P})$, i.e. it essentially uniquely achieves*

$$\sup_S \mathbb{E}_{P_{\boldsymbol{\mu}}}[\log S]$$

where the supremum ranges over all e-variables for $\mathcal{H}_0(\mathcal{P})$.

Here, essential uniqueness means that any other GRO e-variable must be equal to $S_{\mathrm{GRO}(\boldsymbol{\mu}^1)}\mathcal{P}$ with probability 1 under $P_{\boldsymbol{\mu}}$. This in turn implies that the measure P_0^* is in fact unique, as members of regular exponential families must have full support. Thus, once we have fixed our alternative and defined our null as $\mathcal{H}_0(\mathcal{P})$ for some set of distributions $\mathcal{P}$ on $\mathcal{X}$, the optimal (in the GRO sense) e-variable to use is the $S_{\mathrm{GRO}(\boldsymbol{\mu}^1)}\mathcal{P}$ e-variable as defined above.

4.2 The Four Types of E-variables

In this section, we define our four types of e-variables; the definitions can be instantiated to any underlying 1-parameter exponential family. More precisely, we define three 'real' e-variables $S_{\mathrm{RIPR}}, S_{\mathrm{COND}}, S_{\mathrm{MIX}}$ and one 'pseudo-e-variable' S_{PSEUDO}, a variation of S_{RIPR} which for some exponential families is an e-variable, and for others is not.

4.2.1 The GRO E-variable for $\mathcal{H}_0(\mathcal{M})$ and the pseudo e-variable

We now consider the GRO e-variable for our main null of interest, $\mathcal{H}_0(\mathcal{M})$. In practice, for some exponential families $\mathcal{M}$, the infimum over $\mathrm{CONV}(\mathcal{M})$ in (4.1.4) is actually achieved for some $P_{\mu_0^*} \in \mathcal{M}$. In this *easy* case we can determine S_{RIPR} analytically (this happens if $S_{\mathrm{RIPR}} = S_{\mathrm{PSEUDO}}$, see below). For all other $\mathcal{M}$, i.e. whenever the infimum is not achieved at all or is in $\mathrm{CONV}(\mathcal{M}) \setminus \mathcal{M}$, we do not know if S_{RIPR} can be determined analytically. In this *hard* case will numerically approximate it by S'_{RIPR} as defined below. First, for a fixed parameter $\mu_0 \in \mathtt{M}$ we define the vector $\langle \mu_0 \rangle$ as the vector indicating the distribution on $\mathcal{X}^k$ with all parameters equal to μ_0:

$$\langle \mu_0 \rangle := (\mu_0, \dots, \mu_0) \in \mathtt{M}^k \tag{4.2.1}$$

Next, with W a distribution on M, we define

$$p_W := \int p_{\langle \mu_0 \rangle}(X^k) dW(\mu_0) \tag{4.2.2}$$

to be the Bayesian marginal density obtained by marginalizing over distributions in $\mathcal{H}_0(\mathcal{M})$ according to W. Clearly, if W has finite support then the corresponding distribution P_W has $P_W \in \text{CONV}(\mathcal{M})$. We now set

$$S'_{\text{RIPR}} := \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{W'_0}(X^k)}$$

where W'_0 is chosen so that $p_{W'_0}$ is within a small ϵ of achieving the minimum in (4.1.4), i.e. $D(P_{\mu_1,\ldots,\mu_k} \| P'_{W_0}) = \inf_{P \in \text{CONV}(\mathcal{M})} D(P_{\mu_1,\ldots,\mu_k} \| P) + \epsilon'$ for some $0 \le \epsilon' < \epsilon$. Then, by Corollary 2 of Grünwald et al. [42], S'_{RIPR} will *not* be an e-variable unless $\epsilon' = 0$, but in each case (i.e. for each choice of $\mathcal{M}$) we verify numerically that $\sup_{\mu_0 \in \text{M}} \mathbb{E}_{P_{\mu_0,\ldots,\mu_0}}[S] = 1 + \delta$ for negligibly small δ, i.e. δ goes to 0 quickly as ϵ' goes to 0. We return to the details of the calculations in Section 4.5.

We now consider the 'easy' case in which $P_0^* = P_{\langle \mu_0^* \rangle}$ for some $\mu_0^* \in \text{M}$. Clearly, we must have $\mu_0^* := \arg\min_{\mu_0 \in \text{M}} D(P_{\boldsymbol{\mu}} \| P_{\langle \mu_0 \rangle})$. An easy calculation shows that then

$$\mu_0^* = \frac{1}{k} \sum_{i=1}^{k} \mu_i. \tag{4.2.3}$$

Definition 4. S_{PSEUDO} is defined as

$$S_{\text{PSEUDO}} := \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle \mu_0^* \rangle}(X^k)}.$$

S_{PSEUDO} is not always a real e-variable relative to $\mathcal{H}_0(\mathcal{M})$, which explains the name 'pseudo'. Still, it will be very useful as an auxiliary tool in Theorem 7 and derivations. Note that, if it is an e-variable then we know that it is equal to S_{RIPR}:

Proposition 10. S_{PSEUDO} *is an e-variable for* $\mathcal{M}$ *iff* $S_{\text{PSEUDO}} = S_{\text{RIPR}}$.

The proposition above does not give any easily verifiable condition to check whether S_{PSEUDO} is an e-variable or not. The following proposition does provide a condition which is sometimes easy to check (and which we will heavily employ below). With μ_0^* as in (4.2.3), define

$$f(\mu_0) := \sum_{i=1}^{k} \text{VAR}_{P_{\mu_i + \mu_0 - \mu_0^*}}[X] - k\text{VAR}_{P_{\mu_0}}[X].$$

Proposition 11. *If* $f(\mu_0^*) > 0$, *then* S_{PSEUDO} *is not an e-variable. If* $f(\mu_0^*) < 0$, *then there exists an interval* $\text{M}' \subset \text{M}$ *with* μ_0^* *in the interior of* M' *so that* S_{PSEUDO} *is an e-variable for* $\mathcal{H}_0(\mathcal{M}')$, *where* $\mathcal{M}' = \{P_\mu : \mu \in \text{M}'\}$.

4.2.2 The GRO E-variable for $\mathcal{H}_0(\text{IID})$

Recall that we defined $\mathcal{H}_0(\text{IID})$ as the set of distributions under which X_j, $j = 1, \ldots k$, are i.i.d. from some arbitrary distribution on $\mathcal{X}$. By the defining property of e-variables, i.e. expected value bounded by one under the null (4.1.3), it should be clear that any e-variable for $\mathcal{H}_0(\text{IID})$ is also an e-variable for $\mathcal{H}_0(\mathcal{M})$, since $\mathcal{H}_0(\mathcal{M}) \subset \mathcal{H}_0(\text{IID})$. In particular, we can also use the GRO e-variable for $\mathcal{H}_0(\text{IID})$ in our setting with exponential families. It turns out that this e-variable, which we will denote as S_{MIX}, has a simple form that is generically easy to compute. We now show this:

Theorem 6. *The minimum KL divergence $\inf_{P \in \text{CONV}(\mathcal{H}_0(\text{IID}))} D(P_{\boldsymbol{\mu}} \| P)$ as in Lemma 6 is achieved by the distribution P_0^* on $\mathcal{X}^k$ with density*

$$p_0^*(x^k) = \prod_{j=1}^{k} \frac{1}{k} \sum_{i=1}^{k} p_{\mu_i}(x_j).$$

Hence, S_{MIX}, as defined below, is the GRO e-variable for $\mathcal{H}_0(\text{IID})$.

Definition 5. S_{MIX} is defined as

$$S_{\text{MIX}} := \frac{p_{\boldsymbol{\mu}}(X^k)}{\prod_{j=1}^{k} \left(\frac{1}{k} \sum_{i=1}^{k} p_{\mu_i}(X_j) \right)}.$$

The proof of Theorem 6 extends an argument of [88] for the 2-sample Bernoulli case to the general k-sample case. The argument used in the proof does not actually require the alternative to equal the product distribution of k independent elements of an exponential family — it could be given by the product of k arbitrary distributions. However, we state the result only for the former case, as that is the setting we are interested in here.

4.2.3 The Conditional E-variable S_{cond}

So far, we have defined e-variables as likelihood ratios between $P_{\boldsymbol{\mu}}$ and cleverly chosen elements of either $\mathcal{H}_0(\mathcal{M})$ or $\mathcal{H}_0(\text{IID})$. We now do things differently by not considering the full original data $X_1, \ldots X_k$, but instead conditioning on the sum of the sufficient statistics, i.e. $Z = \sum_{i=1}^{k} X_i$. It turns out that doing so actually collapses $\mathcal{H}_0(\mathcal{M})$ to a single distribution, so that the null becomes simple. That is, the distribution of $X^k \mid Z$ is the same under all elements of $\mathcal{H}_0(\mathcal{M})$, as we will prove in Proposition 12. This means that instead of using a likelihood ratio of the original data, we can use a likelihood ratio conditional on Z, which 'automatically' gives an e-variable.

Definition 6. Setting Z to be the random variable $Z := \sum_{i=1}^{k} X_i$, S_{cond} is defined as

$$S_{\text{COND}} := \frac{p_{\boldsymbol{\mu}}\left(X^{k-1} \mid Z\right)}{p_{\langle \mu_0 \rangle}\left(X^{k-1} \mid Z\right)},$$

with $\mu_0 \in \mathtt{M}$ and (X) the sufficient statistic as in (4.1.2).

Proposition 12. *For all $\boldsymbol{\mu}' = (\mu_1', \ldots, \mu_k') \in \mathtt{M}^k$, we have that $p_{\boldsymbol{\mu}'}(x^{k-1} \mid Z = z)$ depends on $\boldsymbol{\mu}'$ only through $\lambda_j := \lambda(\mu_j') - \lambda(\mu_k')$, $j = 1, \ldots k-1$, i.e. it can be written as a function of $(\lambda_1, \ldots, \lambda_{k-1})$. As a special case, for all $\mu_0, \mu_0' \in \mathtt{M}$, it holds that $p_{\langle \mu_0 \rangle}(x^k \mid Z) = p_{\langle \mu_0' \rangle}(x^k \mid Z)$. As a direct consequence, S_{COND} is an e-variable for $\mathcal{H}_0(\mathcal{M})$,*

Example 9. [The Bernoulli Model] If $\mathcal{M}$ is the Bernoulli model and $k = 2$, then the conditional e-variable reduces to a ratio between the conditional probability of $(X_1, X_2) \in \{0,1\}^2$ given their sum $Z \in \{0,1,2\}$. Clearly, for all $\mu_1', \mu_2' \in \mathtt{M} = (0,1)$, we have $p_{\mu_1', \mu_2'}((0,0) \mid Z = 0) = p_{\mu_1', \mu_2'}((1,1) \mid Z = 2) = 1$, so $S_{\mathrm{COND}} = 1$ whenever $Z = 0$ or $Z = 2$, irrespective of the alternative: data with the same outcome in both groups is effectively ignored. A non-sequential version of S_{COND} for the Bernoulli model was analyzed earlier in great detail by [1].

Furthermore, for any $c \in \mathbb{R}$, we have that $\mathtt{M}_c := \{(\mu_1', \mu_2') : \lambda(\mu_1) - \lambda(\mu_2) = c\}$ is the line of distributions within $\mathtt{M}^2$ with the same odds ratio $\log(\mu_1(1-\mu_2)/((1-\mu_1)\mu_2)) = c$. The sequential probability ratio test of two proportions from [93] was based on fixing a c for the alternative (viewing it as a notion of 'effect size') and analyzing sequences of paired data $X_{(1)}, X_{(2)}, \ldots$ with $X_{(i)} = (X_{i,1}, X_{i,2})$ by the product of conditional probabilities

$$\frac{p_c(X_{(i)} \mid Z_{(i)})}{p_0(X_{(i)} \mid Z_{(i)})} = S_{\mathrm{COND}}(X_i),$$

thus effectively using S_{COND} (here, we abuse notation slightly, writing $p_c(x \mid z)$ when we mean $p_{\mu_1', \mu_2'}(x \mid z)$ for any $\mu_1', \mu_2' \in \mathtt{M}_c$). It is, however, important to note that this product was not used for an anytime-valid test but rather for a classical sequential test with a fixed stopping rule especially designed to optimize power.

4.3   Growth Rate Comparison of Our E-variables

Above we provided several recipes for constructing e-variables $S = S^{\boldsymbol{\mu}}$ whose definition implicitly depended on the chosen alternative $\boldsymbol{\mu}$. To compare these, we define, for any non-negative random variables $S_1^{\boldsymbol{\mu}}$ and $S_2^{\boldsymbol{\mu}}$, $S_1^{\boldsymbol{\mu}} \succeq S_2^{\boldsymbol{\mu}}$ to mean that for all $\boldsymbol{\mu} \in \mathtt{M}^k$, it holds that $\mathbb{E}_{P_{\boldsymbol{\mu}}}[\log S_1^{\boldsymbol{\mu}}] \geq \mathbb{E}_{P_{\boldsymbol{\mu}}}[\log S_2^{\boldsymbol{\mu}}]$. We write $S_1^{\boldsymbol{\mu}} \succ S_2^{\boldsymbol{\mu}}$ if $S_1^{\boldsymbol{\mu}} \succeq S_2$ and there exists $\boldsymbol{\mu} \in \mathtt{M}^k$ for which equality does not hold. From now on we suppress the dependency on $\boldsymbol{\mu}$ again, i.e. we write S instead of $S^{\boldsymbol{\mu}}$. We trivially have, for every underlying exponential family $\mathcal{M}$,

$$S_{\mathrm{PSEUDO}} \succeq S_{\mathrm{RIPR}} \succeq S_{\mathrm{MIX}} \text{ and } S_{\mathrm{RIPR}} \succeq S_{\mathrm{COND}}. \tag{4.3.1}$$

We proceed with Theorem 7 and 8 below (proofs in the Appendix). These results go beyond the qualitative assessment above, by numerically bounding the difference in growth rate between S_{PSEUDO} and S_{MIX} (and, because S_{RIPR} must lie in between them, also between these two and S_{RIPR}) and S_{PSEUDO} and S_{COND} respectively. Theorem 7 and 8 are asymptotic (in terms of difference between mean-value parameters) in nature. To

give more precise statements rather than asymptotics we need to distinguish between individual exponential families; this is done in the next section.

To state the theorems, we need a notion of effect size, or discrepancy between the null and the alternative. So far, we have taken the alternative to be fixed and given by $\boldsymbol{\mu}$, but effect sizes are usually defined with the null hypothesis as starting point. To this end, note that each $P_{\langle \mu_0 \rangle} \in \mathcal{H}_0(\mathcal{M})$ corresponds to a whole set of alternatives for which $P_{\langle \mu_0 \rangle}$ is the closest point in KL within the null. This set of alternatives is parameterized by $\mathtt{M}^{(k)}(\mu_0) = \{\mu'_1, \ldots, \mu'_k \in \mathtt{M} : \frac{1}{k} \sum_{i=1}^{k} \mu'_i = \mu_0\}$, as in (4.2.3). We can re-parameterize this set as follows, using the special notation $\langle \mu_0 \rangle$ as given by (4.2.1). Let $\mathtt{A}$ be the set of unit vectors in $\mathbb{R}^k$ whose entries sum to 0, i.e. $\boldsymbol{\alpha} \in \mathtt{A}$ iff $\sqrt{\sum_{j=1}^{k} \alpha_j^2} = 1$ and $\sum_{j=1}^{k} \alpha_j = 0$. Clearly $\boldsymbol{\mu} \in \mathtt{M}^{(k)}(\mu_0)$ if and only if $\mu_1, \ldots, \mu_k \in \mathtt{M}$ and $\boldsymbol{\mu} = \langle \mu_0 \rangle + \delta \boldsymbol{\alpha}$ for some scalar $\delta \geq 0$ and $\boldsymbol{\alpha} \in \mathtt{A}$. We can think of δ as expressing the magnitude of an effect and $\boldsymbol{\alpha}$ as its direction. Note that, if $k = 2$, then there are only two directions, $\mathtt{A} = \{\boldsymbol{a}_1, \boldsymbol{a}_{-1}\}$ with $\boldsymbol{a}_1 = (1/\sqrt{2}, -1/\sqrt{2})$ and $\boldsymbol{a}_{-1} = -\boldsymbol{a}_1$, corresponding to positive and negative effects: we have $\mu_1 - \mu_2 = \sqrt{2} \cdot \delta$ if $\boldsymbol{\alpha} = \boldsymbol{a}_1$ and $\mu_1 - \mu_2 = -\sqrt{2} \cdot \delta$ if $\boldsymbol{\alpha} = \boldsymbol{a}_{-1}$, as illustrated later on in Figure 4.1. Also note that, for general k, in the theorem below, we can simply interpret δ as the Euclidean distance between $\boldsymbol{\mu}$ and $\langle \mu_0 \rangle$.

Theorem 7. *Fix some $\mu_0 \in \mathtt{M}$, some $\boldsymbol{\alpha} \in \mathtt{A}$ and let $\boldsymbol{\mu} = \langle \mu_0 \rangle + \delta \boldsymbol{\alpha}$ for $\delta \geq 0$ such that $\boldsymbol{\mu} \in \mathtt{M}^{(k)}(\mu_0)$. The difference in growth rate between S_{PSEUDO} and S_{MIX} is given by*

$$\mathbb{E}_{P_{\boldsymbol{\mu}}} [\log S_{\mathrm{PSEUDO}} - \log S_{\mathrm{MIX}}] = \frac{1}{8} \int_x \frac{(f''_x(0))^2}{f_x(0)} d\rho(x) \cdot \delta^4 + o(\delta^4) = O(\delta^4), \quad (4.3.2)$$

where $f_x(\delta) = \sum_{i=1}^{k} p_{\mu_0 + \delta \alpha_i}(x) = \sum\limits_{i=1}^{k} p_{\mu_i}(x)$ and f''_x is the second derivative of f_x, so that $f_x(0) = k p_{\mu_0}(x)$ and (with some calculation) $f''_x(0) = \frac{d^2}{d\mu^2} p_\mu(x) |_{\mu = \mu_0}$.

As is implicit in the $O(\cdot)$-notation, the expectation on the left is well-defined and finite and the integral in the middle equation is finite as well. The theorem implies that for general exponential families, S_{MIX} is surprisingly close $(O(\delta^4))$ to the optimal S_{RIPr} in the GRO sense, whenever the distance δ between $\mathcal{H}_1$ and $\mathcal{H}_0(\mathcal{M})$ is small. This means that, whenever $S_{\mathrm{RIPr}} \neq S_{\mathrm{PSEUDO}}$ (so S_{RIPr} is hard to compute and S_{PSEUDO} is not an e-variable), we might consider using S_{MIX} instead: it will be more robust (since it is an e-variable for the much larger hypothesis $\mathcal{H}_0(\mathrm{IID})$) and it will only be slightly worse in terms of growth rate.

Theorem 7 is remarkably similar to the next theorem, which involves S_{COND} rather than S_{MIX}. To state it, we first set $X_k(x^{k-1}, z) := z - \sum_{i=1}^{k-1} x_i$, and we denote the marginal distribution of $Z = \sum_{i=1}^{k} X_i$ under $P_{\boldsymbol{\mu}}$ as $P_{\boldsymbol{\mu};[Z]}$, noting that its density $p_{\boldsymbol{\mu};[Z]}$ is given by

$$p_{\boldsymbol{\mu};[Z]}(z) = \int_{\mathcal{C}(z)} p_{\boldsymbol{\mu}} (x^{k-1}, x_k) d\rho(x^{k-1}), \quad (4.3.3)$$

where ρ is extended to the product measure of ρ on $\mathcal{X}^{k-1}$ and

$$\mathcal{C}(z) := \left\{ x^{k-1} \in \mathcal{X}^{k-1} : X_i(x^{k-1}, z) \in \mathcal{X} \right\}. \tag{4.3.4}$$

Theorem 8. *Fix some $\mu_0 \in \mathtt{M}$, $\boldsymbol{\alpha} \in \mathtt{A}$ and let $\boldsymbol{\mu} = \langle \mu_0 \rangle + \delta \boldsymbol{\alpha}$ for $\delta \geq 0$ such that $\boldsymbol{\mu} \in \mathtt{M}^{(k)}(\mu_0)$. The difference in growth rate between S_{PSEUDO} and S_{COND} is given by*

$$\mathbb{E}_{P_{\boldsymbol{\mu}}} [\log S_{\mathrm{PSEUDO}} - \log S_{\mathrm{COND}}] = \frac{1}{8} \int_z \frac{(g_z''(0))^2}{g_z(0)} d\rho_{[Z]}(z) \cdot \delta^4 + o\left(\delta^4\right) = O(\delta^4), \tag{4.3.5}$$

where $g_z(\delta) := p_{\langle \mu_0 \rangle + \boldsymbol{\alpha}\delta; [Z]}(z)$ and $\rho_{[Z]}$ denotes the measure on Z induced by the product measure of ρ on $\mathcal{X}^k$; an explicit expression for $g_z''(0)$ is

$$\int_{\mathcal{C}(z)} p_{\langle \mu_0 \rangle} \left(x^k \right) \sum_{j=1}^{k} \left[I'(\mu_0)(x_j - \mu_0) - I(\mu_0) \right] d\rho(x^{k-1}),$$

where $I(\mu)$ denotes the Fisher information for μ and $I'(\mu)$ is its first derivative.

Again, the expectation on the left is well-defined and finite and the integral on the right is finite. Comparing Theorem 8 to Theorem 7, we see that $f_x(0)$, the sum of k identical densities evaluated at x, is replaced by $g_z(0)$, the density of the sum of k i.i.d. random variables evaluated at z.

Corollary 6. *With the definitions as in the two theorems above, the growth-rate difference $\mathbb{E}_{P_{\boldsymbol{\mu}}} [\log S_{\mathrm{COND}} - \log S_{\mathrm{MIX}}]$ can be written as*

$$\frac{1}{8} \left(\int_x \frac{(f_x''(0))^2}{f_x(0)} d\rho(x) - \int_z \frac{(g_z''(0))^2}{g_z(0)} d\rho_{[Z]}(z) \right) \cdot \delta^4 + o\left(\delta^4\right) = O\left(\delta^4\right). \tag{4.3.6}$$

4.4 Growth Rate Comparison for Specific Exponential Families

We will now establish more precise relations between the four (pseudo-) e-variables in k-sample tests for several standard exponential families, namely those listed in Table 4.1 and a few related ones, as listed at the end of this section. For each family $\mathcal{M}$ under consideration, we give proofs for which different e-variables are the same, i.e. $S = S'$, where $S, S' \in \{S_{\mathrm{RIPR}}, S_{\mathrm{COND}}, S_{\mathrm{MIX}}, S_{\mathrm{PSEUDO}}\}$. Whenever we can prove that $S_{\mathrm{RIPR}} \neq S$ for another e-variable $S \in \{S_{\mathrm{COND}}, S_{\mathrm{MIX}}\}$, we can infer that $S_{\mathrm{RIPR}} \succ S$ because S_{RIPR} is the GRO e-variable for $\mathcal{H}_0(\mathcal{M})$. Whenever both S_{COND} and S_{MIX} are not equal to S_{RIPR}, we will investigate via simulation whether $S_{\mathrm{MIX}} \succ S_{\mathrm{COND}}$ or vice versa — our theoretical results do not extend to this case. All simulations are carried out for the case $k = 2$ in the paper. Theorem 7 and Theorem 8 show that in the neighborhood of $\delta = 0$ ($\mu_1, \ldots, \mu_k$ all close together), the difference $\mathbb{E}_{P_{\boldsymbol{\mu}}}[\log S - \log S']$ is of order δ^4 when $S, S' \in \{S_{\mathrm{RIPR}}, S_{\mathrm{PSEUDO}}, S_{\mathrm{MIX}}, S_{\mathrm{COND}}\}$. Hence in the figures we will

show $(\mathbb{E}_{P_\mu}[\log S - \log S'])^{1/4}$, since then we expect the distances to increase linearly as we move away from the diagonal, making the figures more informative.

Our findings, proofs as well as simulations, are summarised in Table 4.1. For each exponential family, we list the rank of the (pseudo-)e-variables when compared with the order '$\succ$'. The ranks that are written in black are proven in Appendix 4.D, while the ranks in blue are merely conjectures based on our simulations as stated above. The results of the simulations on which these conjectures are based are given in Figure 4.1. Furthermore, the rank of S_{PSEUDO} is colored red whenever it is not an e-variable for that model, as shown in the Appendix. Note that whenever any of the e-variables have the same rank, they must be equal ρ-almost everywhere, by strict concavity of the logarithm together with full support of the distributions in the exponential family. For example, the results in the table reflect that for the Bernoulli family, we have shown that $S_{\text{PSEUDO}} = S_{\text{RIPR}} = S_{\text{MIX}}$ and that $S_{\text{PSEUDO}} \succ S_{\text{COND}}$. Also, for the geometric family and beta with free β and fixed α, we have proved that S_{PSEUDO} is not an e-variable, that $S_{\text{RIPR}} \neq S_{\text{MIX}}$ and that $S_{\text{RIPR}} \neq S_{\text{COND}}$, so that it follows from (4.3.1) that $S_{\text{PSEUDO}} \succ S_{\text{RIPR}}$, $S_{\text{RIPR}} \succ S_{\text{MIX}}$ and $S_{\text{RIPR}} \succ S_{\text{COND}}$. Then the findings of the simulations shown in Figure 4.1a suggest that $S_{\text{MIX}} \succ S_{\text{COND}}$ for beta with free β and fixed α and in Figure 4.1b suggest that $S_{\text{COND}} \succ S_{\text{MIX}}$ for geometric family, but these are not proven. Figure 4.1c shows that $S_{\text{MIX}} \succ S_{\text{COND}}$ for Gaussians with free variance and fixed mean. Finally, Figure 4.1d shows that for the exponential, there is no clear relation between S_{MIX} and S_{COND}. That is, S_{MIX} grows faster than S_{COND} for some $\mu_1, \ldots, \mu_k \in \mathbb{M}$, and slower for others, which is indicated by rank $(3) - (4)$ in the table.

Exponential Family	S_{PSEUDO}	S_{RIPR}	S_{MIX}	S_{COND}
Bernoulli	(1)	(1)	(1)	(2)
Gaussian with free mean and fixed variance	(1)	(1)	(2)	(1)
Poisson	(1)	(1)	(2)	(1)
beta with free β and fixed α	(1)	(2)	(3)	(4)
geometric	(1)	(2)	(4)	(3)
Gaussian with free variance and fixed mean	(1)	(2)	(3)	(4)
Exponential	(1)	(2)	(3)-(4)	(3)-(4)

Table 4.1: The ranks of the four different e-variables when compared with the relation '$\succ$'. The ranks in black are proved in Appendix 4.D, while the ranks in blue are conjectures based on the simulations in Figure 4.1. The rank of S_{PSEUDO} is denoted in red whenever it is not an e-variable, as shown in Appendix 4.D

Finally, we note that for each family listed in the table, the results must extend to any other family that becomes identical to it if we reduce it to the natural form (4.1.2). For example, the family of Pareto distributions with fixed minimum parameter v can be reduced to that of the exponential distributions: if $U \sim \text{Pareto}(v, \alpha)$, then we can do a transformation $X = t(U)$ with $t(U) = \log(U/v)$, and then $X \sim \text{Exp}(\alpha)$. Thus, the k-sample problem for U with the $\text{Pareto}(v, \alpha)$ distributions, with α as free parameter, is equivalent to the k-sample problem for X with the exponential distributions; the e-value

(a) beta with free β and fixed α

(b) geometric

(c) Gaussian with free variance and fixed mean
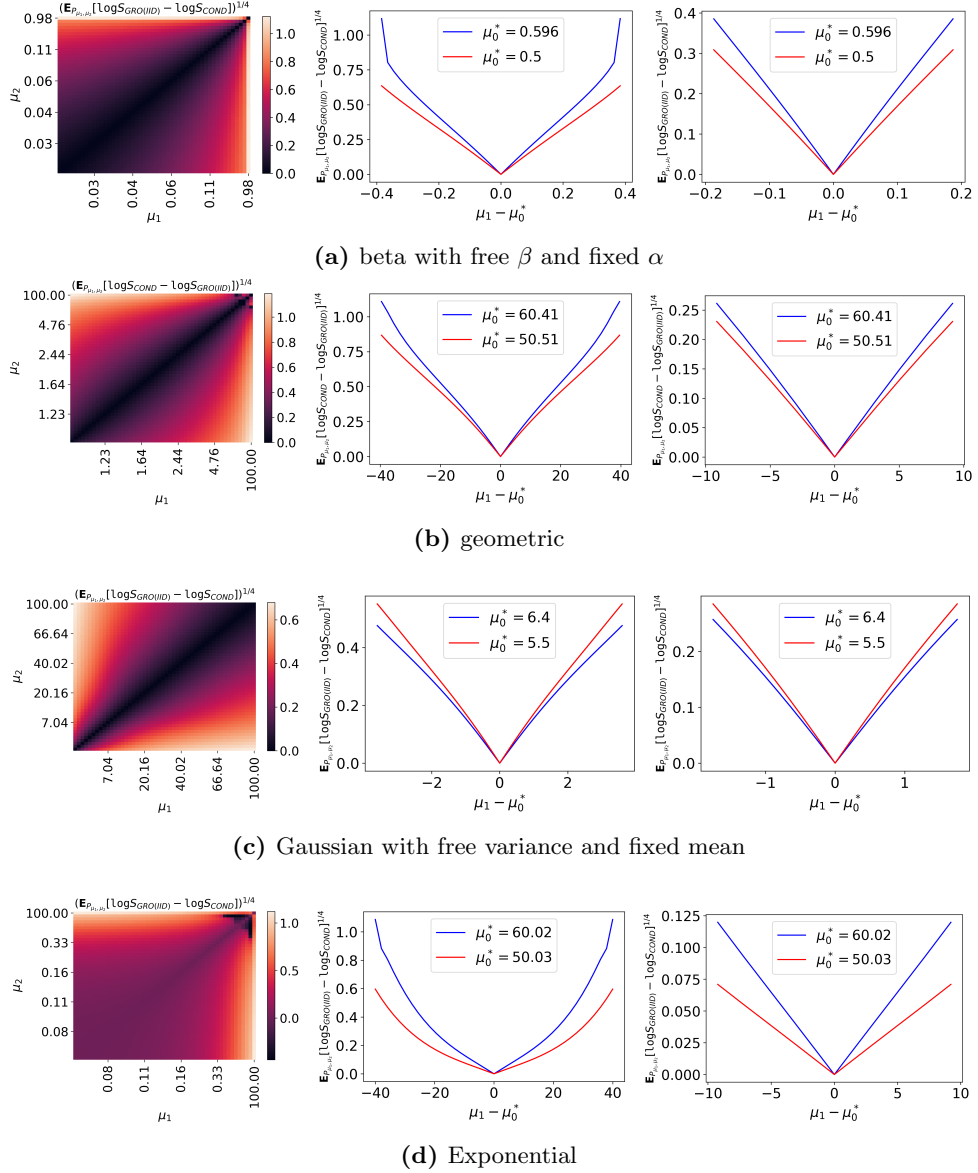
(d) Exponential

Figure 4.1: A comparison of S_{MIX} and S_{COND} for four exponential families. We evaluated the expected growth difference on a grid of 50×50 alternatives (μ_1, μ_2), equally spaced in the standard parameterization (explaining the nonlinear scaling on the depicted mean-value parameterization). On the left are the corresponding heatmaps. On the right are diagonal 'slices' of these heatmaps: the red curve corresponds to the main diagonal (top left - bottom right), the blue curve corresponds to the diagonal starting from the second tick mark (10th discretization point) top left until the second tick mark bottom right. These slices are symmetric around 0, their value only depending on $\delta = |\mu_1 - \mu_2| / \sqrt{2} = |\mu_1 - \mu_0^*| \cdot \sqrt{2}$, where $\mu_0^* = (\mu_1 + \mu_2)/2$ and δ is as in Theorem 7

S_{RIPR} obtained with a particular alternative in the Pareto setting for observation U coincides with S_{RIPR} for the corresponding alternative in the exponential setting for observation $X = t(U)$, and the same holds for S_{MIX} and S_{COND}. Therefore, the ordering for Pareto must be the same as the ordering for exponential in Table 4.1. Similarly, the e-variables for the log-normal distributions (with free mean or variance) can be reduced to the two corresponding normal distribution e-variables.

4.5   Simulations to Approximate the RIPr

Because of its growth optimality property, we may sometimes still want to use the GRO e-variable S_{RIPR}, even in cases where it is not equal to the easily calculable S_{PSEUDO}. To this end we need to approximate it numerically. The goal of this section is twofold: first, we want to illustrate that this is feasible in principle; second, we show that this raises interesting additional questions for future work. Thus, below we consider in more detail simulations to approximate S_{RIPR} for the exponential families with $S_{\mathrm{RIPR}} \neq S_{\mathrm{PSEUDO}}$ that we considered before, i.e. beta, geometric, exponential and Gaussian with free variance; for simplicity we only consider the case $k = 2$. In Appendix 4.E we provide some graphs illustrating the RIPr probability densities for particular choices of μ_1, μ_2; here, we focus on how to approximate them, taking our findings for $k = 2$ as suggestive for what happens with larger k.

4.5.1   Approximating the RIPr via Li's Algorithm

[60] provides an algorithm for approximating the RIPr of distribution Q with density q onto the convex hull $\mathrm{CONV}(\mathcal{P})$ of a set of distributions $\mathcal{P}$ (where each $P \in \mathcal{P}$ has density p) arbitrarily well in terms of KL divergence. At the m-th step, this algorithm outputs a finite mixture $P_{(m)} \in \mathrm{CONV}(\mathcal{P})$ of at most m elements of $\mathcal{P}$. For $m > 1$, these mixtures are determined by iteratively setting $P_{(m)} := \alpha P_{(m-1)} + (1 - \alpha)P'$, where $\alpha \in [0, 1]$ and $P' \in \mathcal{P}$ are chosen so as to minimize KL divergence $D(Q\|\alpha P_{(m-1)} + (1 - \alpha)P')$. Here, $P_{(1)}$ is defined as the single element of $\mathcal{P}$ that minimizes $D(Q\|P_{(1)})$. It is thus a greedy algorithm, but Li shows that, under some regularity conditions on $\mathcal{P}$, it holds that $D(Q\|P_{(m)}) \to \inf_{P \in \mathrm{CONV}(\mathcal{P})} D(Q\|P)$. That is, $P_{(m)}$ approximates the RIPr in terms of KL divergence. This suggests, but is not in itself sufficient to prove, that $\sup_{P \in \mathcal{P}} \mathbb{E}_P[q(X)/p_{(m)}(X)] \to 1$, i.e. that the likelihood ratio actually tends to an e-variable.

We numerically investigated whether this holds for our familiar setting with $k = 2$, Q is equal to $P_{\boldsymbol{\mu}}$ for some $\boldsymbol{\mu} = (\mu_1, \mu_2) \in \mathtt{M}^2$, and $\mathcal{P} = \mathcal{H}_0(\mathcal{M})$. To this end, we applied Li's algorithm to a wide variety of values (μ_1, μ_2) for the beta, exponential, geometric and Gaussian with free variance. In all these cases, after at most $m = 15$ iterations, we found that $\sup_{\mu_0 \in \mathtt{M}} \mathbb{E}_{P_{\mu_0, \mu_0}}[p_{\mu_1, \mu_2}(X_1, X_2)/q_{(m)}(X_1, X_2)]$ was bounded by 1.005: Li's algorithm convergences quite fast; see Appendix 4.E for a graphical depiction of the convergence and design choices in the simulation.

(note that, since we have proved that $S_{\mathrm{RIPR}} = S_{\mathrm{PSEUDO}}$ for Bernoulli, Poisson and Gaussian with free mean, there is no need to approximate S_{RIPR} for those families).

4.5.2 Approximating the RIPr via Brute Force

While Li's algorithm converges quite fast, it is still highly suboptimal at iteration $m = 2$, due to its being greedy. This motivated us to investigate how 'close' we can get to an e-variable by using a mixture of just two components. Thus, we set $p_A(x^k) := \alpha p_{\langle \mu_{01} \rangle}(x^k) + (1 - \alpha) p_{\langle \mu_{02} \rangle}(x^k)$ and, for various choices of $\boldsymbol{\mu} = (\mu_1, \mu_2)$, considered

$$S_{\text{APPR}} := \frac{p_{\boldsymbol{\mu}}(X^k)}{p_A(X^k)} \tag{4.5.1}$$

as an approximate e-variable, for the specific values of $\alpha \in [0, 1]$ and μ_{01}, μ_{02} that minimize

$$\sup_{\mu_0 \in \mathtt{M}} \mathbb{E}_{P_{\langle \mu_0 \rangle}}[S_{\text{APPR}}].$$

(in practice, we maximize μ_0 over a discretization of $\mathtt{M}$ with 1000 equally spaced grid points and minimize $\alpha, \mu_{01}, \mu_{02}$ over a grid with 100 equally sized grid points, with left- and right- end points of the grids over $\mathtt{M}$ determined by trial and error).

The simulation results, for $k = 2$ and particular values of μ_1, μ_2 and the exponential families for which approximation makes sense (i.e. $S_{\text{RIPR}} \neq S_{\text{PSEUDO}}$) are presented in Table 4.2. We tried, and obtained similar results, for many more parameter values; one more parameter pair for each family is given in Table 4.3 in Appendix 4.E. The term $\sup_{\mu_0 \in \mathtt{M}} \mathbb{E}_{P_{\langle \mu_0 \rangle}}[S_{\text{APPR}}]$ is remarkably close to 1 for all of these families. Corollary 2 of Grünwald et al. [42] implies that if the supremum *is* exactly 1, i.e. S_{APPR} is an e-variable, then S_{APPR} must also be the GRO e-variable relative to $P_{\boldsymbol{\mu}}$. This leads us to speculate that perhaps all the exceedance beyond 1 is due to discretization and numerical error, and the following might (or might not — we found no way of either proving or disproving the claim) be the case:

Conjecture For $k = 2$, the RIPr, i.e. the distribution achieving

$$\min_{Q \in \text{CONV}(\mathcal{H}_0(\mathcal{M}))} D(P_{\mu_1, \mu_2} \| Q)$$

can be written as a mixture of just two elements of $\mathcal{H}_0(\mathcal{M})$.

4.6 Conclusion and Future Work

In this paper, we introduced and analysed four types of e-variables for testing whether k groups of data are distributed according to the same element of an exponential family. These four e-variables include: the GRO e-variable (S_{RIPR}), a conditional e-variable (S_{COND}), a mixture e-variable (S_{MIX}), and a pseudo-e-variable (S_{PSEUDO}). We compared the growth rate of the e-variables under a simple alternative where each of the k groups has a different, but fixed, distribution in the same exponential family. We have shown that for any two of the e-variables $S, S' \in \{S_{\text{RIPR}}, S_{\text{COND}}, S_{\text{MIX}}, S_{\text{PSEUDO}}\}$, we have $\mathbb{E}[\log S - \log S'] = O(\delta^4)$ if the ℓ_2 distance between the parameters of this alternative distribution and the parameter space of the null is given by δ. This shows that when

Distributions	(μ_1, μ_2)	α	(μ_{01}, μ_{02})	$\sup\limits_{\mu_0 \in \mathbb{M}} \mathbb{E}_{X_1, X_2 \sim P_{\mu_0, \mu_0}}[S_{\text{APPR}}]$
beta	$(0.5, 0.25)$	0.22	$(0.24, 0.81)$	1.0052
Exponential	$(0.5, 0.25)$	0.56	$(0.35, 0.51)$	1.0000
Gaussian with free variance and fixed mean	$(0.5, 0.25)$	0.37	$(0.5, 0.5)$	1.0000
Exponential	$\left(\frac{10}{3}, \frac{5}{4}\right)$	0.51	$(0.62, 0.31)$	1.0047
geometric	$\left(\frac{10}{3}, \frac{5}{4}\right)$	0.47	$(1.84, 2.97)$	1.0008
Gaussian with free variance and fixed mean	$\left(\frac{10}{3}, \frac{5}{4}\right)$	0.08	$(3.64, 2.73)$	1.0002

Table 4.2: For given values of $\boldsymbol{\mu} = (\mu_1, \mu_2)$, we show α, μ_{01} and μ_{02} for the corresponding two-component mixture $\alpha p_{\mu_{01}}(X_1) p_{\mu_{01}}(X_2) + (1 - \alpha) p_{\mu_{02}}(X_1) p_{\mu_{02}}(X_2)$ arrived at by brute-force minimization of the KL divergence as in Section 4.5.2, and we show how close the corresponding likelihood ratio S_{APPR} is to being an e-variable

the effect size is small, all the e-variables behave surprisingly similar. For more general effect sizes, we know that S_{RIPR} has the highest growth rate by definition. Calculating S_{RIPR} involves computing the reverse information projection of the alternative on the null, which is generally a hard problem. However, we proved that there are exponential families for which one of the following holds $S_{\text{PSEUDO}} = S_{\text{RIPR}}$, $S_{\text{COND}} = S_{\text{RIPR}}$ or $S_{\text{MIX}} = S_{\text{RIPR}}$, which considerably simplifies the problem. If one is interested in testing an exponential family for which is not the case, there are algorithms to estimate the reverse information projection. We have numerically verified that approximations of the reverse information projection also lead to approximations of S_{RIPR}. However, the use of S_{COND} or S_{MIX} might still be preferred over S_{RIPR} due to the computational advantage. Our simulations show that depends on the specific exponential family which of them is preferable over the other, and that sometimes there is even no clear order.

Appendix 4.A   Application in Practice: k Separate I.I.D. Data Streams

In the simplest practical applications, we observe one block at a time, i.e. at time n, we have observed $\boldsymbol{X}_{(1)}, \ldots, \boldsymbol{X}_{(n)}$, where each $\boldsymbol{X}_{(i)} = (X_{i,1}, \ldots, X_{i,k})$ is a block, i.e. a

vector with one outcome for each of the k groups. This is a rather restrictive setup, but we can easily extend it to blocks of data in which each group has a different number of outcomes. For example, if data comes in blocks with m_j outcomes in group j, for $j = 1 \ldots k$, $X_{(i)} = (X_{i,1,1}, \ldots, X_{i,1,m_1}, X_{i,2,1}, \ldots, X_{i,2,m_2}, \ldots, X_{i,k,1}, \ldots, X_{i,k,m_k})$, we can re-organize this having $k' = \sum_{j=1}^{k} m_j$ groups, having 1 outcome in each group, and having an alternative in which the first m_1 entries of the outcome vector share the same mean $\mu'_1 = \ldots = \mu'_{m_1} = \mu_1$; the next m_2 entries share the same mean $\mu'_{m_1+1} = \ldots = \mu'_{m_1+m_2} = \mu_2$, and so on.

Even more generally though, we will be confronted with k separate i.i.d streams and data in each stream may arrive at a different rate. We can still handle this case by pre-determining a multiplicity $m_1, \ldots, m_k$ for each stream. As data comes in, we fill virtual 'blocks' with m_j outcomes for group j, $j = 1 \ldots k$. Once a (number of) virtual block(s) has been filled entirely, the analysis can be performed as usual, restricted to the filled blocks. That is, if for some integer B we have observed Bm_j outcomes in stream j, for all $j = 1 \ldots k$, but for some j, we have not yet observed $(B+1)m_j$ outcomes, and we decide to stop the analysis and calculate the evidence against the null, then we output the product of e-variables for the first B blocks and ignore any additional data for the time being. Importantly, if we find out, while analyzing the streams, that some streams are providing data at a much faster rate than others, we may adapt $m_1, \ldots, m_k$ dynamically: whenever a virtual block has been finished, we may decide on alternative multiplicities for the next block; see [88] for a detailed description for the case that $k = 2$.

Appendix 4.B   Proofs for Section 4.2

In the proofs we freely use, without specific mention, basic facts about derivatives of (log-) densities of exponential families. These can all be found in, for example, [13].

4.B.1   Proof of Proposition 10

Proof. Since S_{RIPR} was already shown to be an E-variable in Lemma 7, the 'if' part of the statement holds. The 'only-if' part follows directly from Corollary 2 to Theorem 1 in [42], which states that there can be at most one E-variable of the form $p_{\boldsymbol{\mu}}(X^k)/r(X^k)$ where r is a probability density for X^k. $\qquad\square$

4.B.2 Proof of Proposition 11

Proof. Define $g(\mu_0) := \mathbb{E}_{p_{\langle \mu_0 \rangle}} [S_{\text{PSEUDO}}]$ and $B(\mu_i) := A\left(\lambda(\mu_i) + \lambda(\mu_0) - \lambda(\mu_0^*)\right)$.

$$g(\mu_0) = \mathbb{E}_{p_{\langle \mu_0 \rangle}} \left[\prod_{i=1}^{k} \frac{p_{\mu_i}(X_i)}{p_{\mu_0^*}(X_i)} \right] = \prod_{i=1}^{k} \mathbb{E}_{Y \sim p_{\mu_0}} \left[\frac{p_{\mu_i}(Y)}{p_{\mu_0^*}(Y)} \right]$$

$$= \prod_{i=1}^{k} \int \exp\left(\lambda(\mu_0)y - A\left(\lambda(\mu_0)\right)\right) \cdot \frac{\exp\left(\lambda(\mu_i)y - A\left(\lambda(\mu_i)\right)\right)}{\exp\left(\lambda(\mu_0^*)y - A\left(\lambda(\mu_0^*)\right)\right)} d\rho(y)$$

$$= \prod_{i=1}^{k} \int \exp\left(\left(\lambda(\mu_i) + \lambda(\mu_0) - \lambda(\mu_0^*)\right)y - A\left(\lambda(\mu_i)\right) - A\left(\lambda(\mu_0)\right) + A\left(\lambda(\mu_0^*)\right)\right) d\rho(y)$$

$$= \prod_{i=1}^{k} \exp\left(A\left(\lambda(\mu_0^*)\right) - A\left(\lambda(\mu_i)\right) - A\left(\lambda(\mu_0)\right)\right) \exp\left(B(\mu_i)\right)$$

$$\cdot \int \exp\left(\left(\lambda(\mu_i) + \lambda(\mu_0) - \lambda(\mu_0^*)\right)y - B(\mu_i)\right) d\rho(y)$$

$$= \prod_{i=1}^{k} \exp\left(A\left(\lambda(\mu_0^*)\right) - A\left(\lambda(\mu_i)\right) - A\left(\lambda(\mu_0)\right)\right) \exp\left(B(\mu_i)\right) \cdot 1$$

$$= \exp\left(kA\left(\lambda(\mu_0^*)\right) - \sum_{i=1}^{k} A\left(\lambda(\mu_i)\right) - kA\left(\lambda(\mu_0)\right) + \sum_{i=1}^{k} B(\mu_i)\right). \tag{4.B.1}$$

Taking first and second derivatives with respect to μ_0, we find

$$\frac{d}{d\mu_0} g(\mu_0) = g(\mu_0) \cdot \frac{d}{d\mu_0} \left(\sum_{i=1}^{k} B(\mu_i) - kA\left(\lambda(\mu_0)\right)\right) \tag{4.B.2}$$

and

$$\frac{d^2}{d\mu_0^2} g(\mu_0) = \left(\frac{d}{d\mu_0} g(\mu_0)\right) \cdot \frac{d}{d\mu_0} \left(\sum_{i=1}^{k} B(\mu_i) - kA\left(\lambda(\mu_0)\right)\right)$$

$$+ g(\mu_0) \cdot \frac{d^2}{d\mu_0^2} \left(\sum_{i=1}^{k} B(\mu_i) - kA\left(\lambda(\mu_0)\right)\right)$$

$$= g(\mu_0) \left(\sum_{i=1}^{k} (\mu_i + \mu_0 - \mu_0^*) - k\mu_0\right)^2 \tag{4.B.3}$$

$$+ g(\mu_0) \left(\sum_{i=1}^{k} \text{VAR}_{P_{\mu_i + \mu_0 - \mu_0^*}}[X] - k\text{VAR}_{P_{\mu_0}}[X]\right)$$

$$= g(\mu_0) \left(\sum_{i=1}^{k} \text{VAR}_{P_{\mu_i + \mu_0 - \mu_0^*}}[X] - k\text{VAR}_{P_{\mu_0}}[X]\right) = g(\mu_0) \cdot f(\mu_0).$$

where the second equality holds by (4.B.2), $(d/d\lambda(\mu))A(\lambda(\mu)) = \mathbb{E}_{P_\mu}[X]$ and $(d^2/d\lambda(\mu)^2)A(\lambda(\mu)) = \text{VAR}_{P_\mu}[X]$. (4.B.3) is continuous with respect to μ_0. Therefore, if $f(\mu_0^*) > 0$ holds, it means that there exists an interval $\text{M}^* \subset \text{M}$ with μ_0^* in the interior of M^* on which (4.B.1) is strictly convex. Then there must exist a point $\mu_0' \in \text{M}^*$ satisfying $\mathbb{E}_{P_{\langle\mu_0'\rangle}}[S_{\text{PSEUDO}}] > \mathbb{E}_{P_{\langle\mu_0^*\rangle}}[S_{\text{PSEUDO}}] = 1$, i.e. S_{PSEUDO} is not an E-variable. Conversely, $f(\mu_0^*) < 0$ means that there exists an interval $\text{M}^* \subset \text{M}$ with μ_0^* in the interior of M^*, on which (4.B.1) is strictly concave. The result follows. □

4.B.3   Proof of Theorem 6

To prepare for the proof of Theorem 6, let us first recall Young's inequality [99]:

Lemma 8. [Young's inequality] *Let p, q be positive real numbers satisfying $\frac{1}{p} + \frac{1}{q} = 1$. Then if a, b are nonnegative real numbers, $ab \leq \frac{a^p}{p} + \frac{b^q}{q}$.*

The proof of Theorem 6 follows exactly the same argument as the one used by [88] to prove this statement in the special case that $\mathcal{M}$ is the Bernoulli model.

Proof. We first show that S_{MIX} as defined in the theorem statement is an E-variable. For this, we set $p_0^*(X) = \frac{1}{k} \sum_{i=1}^{k} p_{\mu_i}(X)$. We have:

$$\mathbb{E}_{X^k \sim P_{\langle\mu_0\rangle}}[S_{\text{MIX}}] = \mathbb{E}_{X_1 \sim P_{\mu_0}}\left[\frac{p_{\mu_1}(X_1)}{p_0^*(X_1)}\right] \cdot \ldots \cdot \mathbb{E}_{X_k \sim P_{\mu_0}}\left[\frac{p_{\mu_k}(X_k)}{p_0^*(X_k)}\right]. \qquad (4.B.4)$$

We also have

$$\frac{1}{k}\mathbb{E}_{X_1 \sim P_{\mu_0}}\left[\frac{p_{\mu_1}(X_1)}{p_0^*(X_1)}\right] + \cdots + \frac{1}{k}\mathbb{E}_{X_k \sim P_{\mu_0}}\left[\frac{p_{\mu_k}(X_k)}{p_0^*(X_k)}\right]$$

$$= \frac{1}{k}\mathbb{E}_{X \sim P_{\mu_0}}\left[\frac{p_{\mu_1}(X)}{\frac{1}{k}\sum_{i=1}^{k} p_{\mu_i}(X)} + \cdots + \frac{p_{\mu_k}(X)}{\frac{1}{k}\sum_{i=1}^{k} p_{\mu_i}(X)}\right] = 1. \qquad (4.B.5)$$

We need to show that (4.B.4) ≤ 1, for which we can use (4.B.5). Stated more simply, it

is sufficient to prove $\prod_{i=1}^{k} r_i \leq 1$ with $\frac{1}{k} \sum_{i=1}^{k} r_i \leq 1$, $r_i \in \mathbb{R}^+$. But this is easily established:

$$
\frac{1}{k} \sum_{i=1}^{k} r_i = \frac{k-1}{k} \cdot \frac{\sum_{i=1}^{k-1} r_i}{k-1} + \frac{r_k}{k} \geq \left(\frac{\sum_{i=1}^{k-1} r_i}{k-1} \right)^{\frac{k-1}{k}} r_k^{\frac{1}{k}}
$$

$$
= \left(\frac{k-2}{k-1} \cdot \frac{\sum_{i=1}^{k-2} r_i}{k-2} + \frac{r_{k-1}}{k-1} \right)^{\frac{k-1}{k}} r_k^{\frac{1}{k}}
$$

$$
\geq \left(\frac{\sum_{i=1}^{k-2} r_i}{k-2} \right)^{\frac{k-2}{k}} r_{k-1}^{\frac{1}{k}} r_k^{\frac{1}{k}}
$$

$$
\vdots
$$

$$
\geq \left(\frac{r_1 + r_2}{2} \right)^{\frac{2}{k}} \prod_{i=3}^{k} r_i^{\frac{1}{k}} \geq \prod_{i=1}^{k} r_i^{\frac{1}{k}} \qquad (4.\text{B}.6)
$$

where the first inequality holds because of Young's inequality, by setting $\frac{1}{p} := \frac{k-1}{k}, \frac{1}{q} := \frac{1}{k}, a^p := \frac{\sum_{i=1}^{k-1} r_i}{k-1}, b^q := r_k$ in Lemma 8. The other inequalities are established in the same way. It follows that $\prod_{i=1}^{k} r_i^{\frac{1}{k}} \leq 1$ and further $\prod_{i=1}^{k} r_i \leq 1$.

This shows that S_{MIX} is a e-variable. It remains to show that S_{MIX} is indeed the GRO e-variable relative to $\mathcal{H}_0(\text{IID})$; once we have shown this, it follows by Lemma 2 that it is the unique such e-variable and therefore by Lemma 1 that P_0^* achieves the minimum in Lemma 1. Since we already know that S_{MIX} is an e-variable, the fact that it is the GRO e-variable relative to $\mathcal{H}_0(\text{IID})$ follows immediately from Corollary 2 of Theorem 1 in [42], which states that there can be at most one e-variable of form $p_{\boldsymbol{\mu}}(X^k)/r(X^k)$ where r is a probability density. Since S_{MIX} is such an e-variable, Lemma 1 gives that it must be the GRO e-variable. $\qquad \square$

4.B.4   Proof of Proposition 12

Proof. The observed values of $X_1, X_2, \ldots, X_k$ are denoted as x^k $(:= x_1, \ldots, x_k)$. With $X_k(x^{k-1}, z) := z - \sum_{i=1}^{k-1} x_i$ and $\mathcal{C}(z)$ as in (4.3.4) and $p_{\boldsymbol{\mu};[Z]}(z)$ and $\rho(x^{k-1})$ as in

(4.3.3), we get:

$$p_{\boldsymbol{\mu}}\left(x^{k-1}\big|Z=z\right)=\frac{p_{\boldsymbol{\mu}}\left(x^k\right)}{p_{\boldsymbol{\mu};[Z]}\left(z\right)}$$

$$=\frac{\exp\left(\sum\limits_{i=1}^{k}\left(\lambda(\mu_i)x_i-A(\lambda(\mu_i))\right)\right)}{\int_{y^{k-1}\in\mathcal{C}(z)}\exp\left(\sum\limits_{i=1}^{k-1}\left(\lambda(\mu_i)y_i-A(\lambda(\mu_i))+\lambda(\mu_k)X_k(y^{k-1},z))-A(\lambda(\mu_k))\right)\right)d\rho(y^{k-1})}$$

$$=\frac{\exp\left(\lambda(\mu_k)z+\sum\limits_{i=1}^{k-1}\left(\lambda(\mu_i)-\lambda(\mu_k)\right)x_i\right)}{\int_{y^{k-1}\in\mathcal{C}(z)}\exp\left(\lambda(\mu_k)z+\sum\limits_{i=1}^{k-1}\left(\lambda(\mu_i)-\lambda(\mu_k)\right)y_i\right)d\rho(y^{k-1})}$$

$$=\frac{\exp\left(\sum\limits_{i=1}^{k-1}\left(\lambda(\mu_i)-\lambda(\mu_k)\right)x_i\right)}{\int_{y^{k-1}\in\mathcal{C}(z)}\exp\left(\sum\limits_{i=1}^{k-1}\left(\lambda(\mu_i)-\lambda(\mu_k)\right)y_i\right)d\rho(y^{k-1})}.$$

$\square$

Appendix 4.C   Proofs for Section 4.3

4.C.1   Proof of Theorem 7

Proof. We prove the theorem using an elaborate Taylor expansion of $F(\delta)$, defined below, around $\delta=0$. We first calculate the first four derivatives of $F(\delta)$. Thus we define and derive, with $\mu_i=\mu_0+\alpha_i\delta$ and $f_y(\delta)=\sum\limits_{i=1}^{k}p_{\mu_i}(y)$ defined as in the theorem

statement,

$$F(\delta) := \mathbb{E}_{P_{\langle\mu_0\rangle + \alpha\delta}} \left[\log S_{\text{PSEUDO}} - \log S_{\text{MIX}} \right]$$

$$= \mathbb{E}_{P_{\boldsymbol{\mu}}} \left[\log \prod_{j=1}^{k} \left(\frac{1}{k} \sum_{i=1}^{k} p_{\mu_i}(X_j) \right) - \log p_{\langle\mu_0\rangle}(X^k) \right]$$

$$= \mathbb{E}_{P_{\boldsymbol{\mu}}} \left[\sum_{j=1}^{k} \log f_{X_j}(\delta) - \sum_{j=1}^{k} \log p_{\mu_0}(X_j) \right] - k \log k$$

$$\overset{(a)}{=} \sum_{j=1}^{k} \mathbb{E}_{X \sim P_{\mu_j}} \left[\log f_X(\delta) - \log p_{\mu_0}(X) \right] - k \log k$$

$$\overset{(b)}{=} \overbrace{\int_{y \in \mathcal{X}} f_y(\delta) \log f_y(\delta) d\rho(y)}^{F_1(\delta)} + \overbrace{\left(- \int_{y \in \mathcal{X}} f_y(\delta) \log p_{\mu_0}(y) d\rho(y) \right)}^{F_2(\delta)} - k \log k, \quad (4.\text{C}.1)$$

where we define $F_1(\delta)$ to be equal to the leftmost term in (4.C.1) and $F_2(\delta)$ to be equal to the second, and (a) and (b) both hold provided that

$$\text{for all } j \in \{1, \dots, k\}: \mathbb{E}_{X_j \sim P_{\mu_j}} \left[|\log f_{X_j}(\delta) - \log p_{\mu_0}(X_j)| \right] < \infty \quad (4.\text{C}.2)$$

is finite. In the online supplementary material we verify that this condition, as well as a plethora of related finiteness-of-expectation-of-absolute-value conditions hold for all δ sufficiently close to 0. Together these not just imply (a) and (b), but also (c) that we can freely exchange integration over y and differentiation over δ for all such δ when computing the first k derivatives of $F_1(\delta)$ and $F_2(\delta)$, for any finite k and (d) that all these derivatives are finite for δ in a compact interval including 0 (since the details are straightforward but quite tedious and long-winded we deferred these to the supplementary material). Thus, using (c), we will freely differentiate under the integral sign in the remainder of the proof below, and using (d), we will be able to conclude that the final result is finite.

For each derivative, we first compute the derivative of $F_1(\delta)$ and then that of $F_2(\delta)$.

$$F_1'(\delta) = \int f_y'(\delta) d\rho(y) + \int f_y'(\delta) \log f_y(\delta) d\rho(y) = 0,$$

$$F_2'(\delta) = - \int f_y'(\delta) \log p_{\mu_0}(y) d\rho(y) = 0, \text{ so } F'(0) = F_1'(0) + F_2'(0) = 0, \quad (4.\text{C}.3)$$

where the above formulas hold since $f'_x(0) = 0$ for all $x \in \mathcal{X}$, which can be obtained by

$$f'_x(\delta^\circ) = \sum_{j=1}^{k} \frac{dp_{\mu_j}(x)}{d\mu_j} \frac{d\mu_j}{d\delta}(\delta^\circ),$$

$$f'_x(0) = \frac{dp_{\mu_0}(x)}{d\mu_0} \sum_{j=1}^{k} \frac{d\mu_j}{d\delta}(0) = \frac{dp_{\mu_0}(x)}{d\mu_0} \sum_{j=1}^{k} \alpha_j = 0, \tag{4.C.4}$$

where we used that all μ_j are equal to μ_0 at $\delta = 0$. We turn to the second derivatives:

$$F''_1(\delta) = \int f''_y(\delta) d\rho(y) + \int \left(f''_y(\delta) \log f_y(\delta) + \frac{\left(f'_y(\delta) \right)^2}{f_y(\delta)} \right) d\rho(y)$$

$$= \int \left(f''_y(\delta) \log f_y(\delta) + \frac{\left(f'_y(\delta) \right)^2}{f_y(\delta)} \right) d\rho(y)$$

$$F''_1(0) = \int \left(f''_y(0) \log f_y(0) + \frac{\left(f'_y(0) \right)^2}{f_y(0)} \right) d\rho(y);$$

$$= \int f''_y(0) \log p_{\mu_0}(y) d\rho(y) + \int_{y \in \mathcal{X}} \left(f''_y(0) \log k \right) d\rho(y) \tag{4.C.5}$$

$$= \int \left(f''_y(0) \log p_{\mu_0}(y) \right) d\rho(y),$$

where $\int f''_y(\delta) d\rho(y) = 0$ because $\int f_y(\delta) d\rho(y) = k$, in which k is a constant that does not depend on δ. Then $F''_2(\delta)$ is given by

$$F''_2(\delta) = - \int f''_y(\delta) \log p_{\mu_0}(y) d\rho(y) \; ; \; F''_2(0) = - \int f''_y(0) \log p_{\mu_0}(y) d\rho(y), \text{ so}$$

$$F''(0) = F''_1(0) + F''_2(0) = 0. \tag{4.C.6}$$

Now we compute the third derivative of $F(\delta)$, denoted as $F^{(3)}(\delta)$.

$$F_1^{(3)}(\delta) = \int \left(f_y^{(3)}(\delta) \log f_y(\delta) + \frac{f_y''(\delta) f_y'(\delta)}{f_y(\delta)} + \frac{2 f_y''(\delta) f_y'(\delta) f_y(\delta) - (f_y'(\delta))^3}{(f_y(\delta))^2} \right) d\rho(y)$$

$$F_1^{(3)}(0) = \int f_y^{(3)}(0) \log f_y(0) d\rho(y)$$

$$= \int f_y^{(3)}(0) \log p_{\mu_0}(y) d\rho(y) + \int f_y^{(3)}(0) \log k d\rho(y) \tag{4.C.7}$$

$$= \int f_y^{(3)}(0) \log p_{\mu_0}(y) d\rho(y)$$

$$F_2^{(3)}(\delta) = - \int f_y^{(3)}(\delta) \log p_{\mu_0}(y) d\rho(y)$$

$$F_2^{(3)}(0) = - \int f_y^{(3)}(0) \log p_{\mu_0}(y) d\rho(y), \text{ so } F^{(3)}(0) = F_1^{(3)}(0) + F_2^{(3)}(0) = 0,$$

which holds since $f_y'(0) = 0$ and $\int f_y(0) d\rho(y) = k$.

The fourth derivative of $F(\delta)$ can be computed as follows:

$$F_1^{(4)}(\delta) = \int \left(f_y^{(4)}(\delta) \log f_y(\delta) + \frac{f_y^{(3)}(\delta) f_y'(\delta)}{f_y(\delta)} \right) d\rho(y)$$

$$+ \int 3 \cdot \frac{\left(f_y^{(3)}(\delta) f_y'(\delta) + (f_y''(\delta))^2 \right) f_y(\delta) - f_y''(\delta) \left(f_y'(\delta) \right)^2}{\left(f_y(\delta) \right)^2} d\rho(y)$$

$$- \int \frac{3 \left(f_y(\delta) f_y'(\delta) \right)^2 \cdot f_y''(\delta) - 2 \left(f_y'(\delta) \right)^4 \cdot f_y(\delta)}{\left(f_y(\delta) \right)^4} d\rho(y) ; \tag{4.C.8}$$

$$F_1^{(4)}(0) = \int \left(f_y^{(4)}(0) \log f_y(0) + \frac{3 \left(f_y''(0) \right)^2}{f_y(0)} \right) d\rho(y)$$

$$= \int f_y^{(4)}(0) \log p_{\mu_0}(y) d\rho(y) + \log k \int_{y \in \mathcal{X}} f_y^{(4)}(0) d\rho(y) + \int_{y \in \mathcal{X}} \frac{3 \left(f_y''(0) \right)^2}{f_y(0)} d\rho(y)$$

$$= \int f_y^{(4)}(0) \log p_{\mu_0}(y) d\rho(y) + \int_{y \in \mathcal{X}} \frac{3 \left(f_y''(0) \right)^2}{f_y(0)} d\rho(y),$$

and $F_2^{(4)}(\delta)$ can be computed by

$$F_2^{(4)}(\delta) = -\int f_y^{(4)}(\delta)\log p_{\mu_0}(y)d\rho(y),\ \ F_2^{(4)}(0) = -\int f_y^{(4)}(0)\log p_{\mu_0}(y)d\rho(y),\ \text{so}$$

$$F^{(4)}(0) = F_1^{(4)}(0) + F_2^{(4)}(0) = \int \frac{3\left(f_y''(0)\right)^2}{f_y(0)}d\rho(y) > 0.$$

Based on the above derivatives, we can now do a fourth-order Taylor expansion of $F(\delta)$ around $\delta = 0$, which gives:

$$\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log S_{\text{PSEUDO}} - \log S_{\text{MIX}}\right] = \frac{1}{4!}F^{(4)}(0)\delta^4 + o(\delta^4)$$

$$= \frac{1}{8}\int_{y\in\mathcal{X}} \frac{\left(f_y''(0)\right)^2}{f_y(0)}d\rho(y)\cdot\delta^4 + o\left(\delta^4\right),$$

where $f_y(0) = \sum_{i=1}^k p_{\mu_0}(y) = kp_{\mu_0}(y)$ and $f_y''(0) = \left(\sum_{i=1}^k \alpha_i^2\right)\cdot\frac{d^2}{d\mu^2}p_{\boldsymbol{\mu}}(y)\mid_{\mu=\mu_0} = \frac{d^2}{d\mu^2}p_{\boldsymbol{\mu}}(y)\mid_{\mu=\mu_0}$. $\quad\square$

4.C.2   Proof of Theorem 8

Proof. We obtain the result using an even more involved Taylor expansion than in the previous theorem. As in that theorem, we will freely differentiate (with respect to δ) under the integral sign — that this is allowed is again verified in the online supplementary material.

Let $\boldsymbol{\mu}, \boldsymbol{\alpha}, \mathcal{C}(z), \rho(x^{k-1}), P_{\boldsymbol{\mu}}$ etc. be as in the theorem statement. We have:

$$f(\delta) := \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log S_{\text{PSEUDO}} - \log S_{\text{COND}}\right]$$

$$= \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log\frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle\mu_0\rangle}\left(X^k\right)} - \log\frac{p_{\boldsymbol{\mu}}\left(X^{k-1}\mid Z\right)}{p_{\langle\mu_0\rangle}\left(X^{k-1}\mid Z\right)}\right]$$

$$= \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log\frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle\mu_0\rangle}\left(X^k\right)} - \log\frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle\mu_0\rangle}\left(X^k\right)} + \log\frac{\int_{\mathcal{C}(z)}p_{\boldsymbol{\mu}}\left(x^k\right)d\rho(x^{k-1})}{\int_{\mathcal{C}(z)}p_{\langle\mu_0\rangle}\left(x^k\right)d\rho(x^{k-1})}\right]$$

$$= D\left(P_{\langle\mu_0\rangle+\boldsymbol{\alpha}\delta;[Z]}\|P_{\langle\mu_0\rangle;[Z]}\right).$$

We will prove the result by doing a Taylor expansion for $f(\delta)$ around $\delta = 0$. It is obvious that $f(0) = 0$ and the first derivative $f'(0) = 0$ since $f(0)$ is the minimum of $f(\delta)$ over an open set, and $f(\delta)$ is differentiable. We proceed to compute the second derivative of $f(\delta)$, using the notation $g_z(\delta) = p_{\langle\mu_0\rangle+\boldsymbol{\alpha}\delta;[Z]}(z)$ as in the theorem statement, with

g_z' and g_z'' denoting first and second derivatives.

$$f'(\delta) = \int g_z'(\delta) \log \frac{g_z(\delta)}{g_z(0)} d\rho_{[Z]}(z) + \int g_z'(\delta) d\rho_{[Z]}(z) = \int g_z'(\delta) \log \frac{g_z(\delta)}{g_z(0)} d\rho_{[Z]}(z).$$

$$f''(\delta) = \int g_z''(\delta) \log \frac{g_z(\delta)}{g_z(0)} d\rho_{[Z]}(z) + \int \frac{(g_z'(\delta))^2}{g_z(\delta)} d\rho_{[Z]}(z),$$

where in the first line, the second equality follows since the second term does not change if we interchanging differentiation and integration and the fact that $\int g_z(\delta) dz = 1$ is constant in δ. We obtain

$$f''(0) = \int \frac{(g_z'(0))^2}{g_z(0)} d\rho_{[Z]}(z), \tag{4.C.9}$$

and, with x_k set to $X_k(x^{k-1}, z)$ and recalling that $\boldsymbol{\mu} = \langle \mu_0 \rangle + \boldsymbol{\alpha} \delta$ and $\mu_j = \mu_0 + \alpha_j \delta$,

$$g_z'(\delta) = \int_{\mathcal{C}(z)} \frac{d}{d\delta} p_{\langle \mu_0 \rangle + \boldsymbol{\alpha}\delta}(x^k) d\rho(x^{k-1})$$

$$= \int_{\mathcal{C}(z)} \sum_{j=1}^k \prod_{i \in \{1,\dots,k\} \setminus j} p_{\mu_i}(x_i) \frac{dp_{\mu_j}(x_j)}{d\delta} d\rho(x^{k-1})$$

$$= \int_{\mathcal{C}(z)} \sum_{j=1}^k p_{\mu_1,\dots,\mu_{j-1},\mu_{j+1},\dots,\mu_k}(x_1,\dots,x_{j-1},x_{j+1},\dots,x_k) \frac{dp_{\mu_j}(x_j)}{d\mu_j} \frac{d\mu_j}{d\delta} d\rho(x^{k-1})$$

$$= \int_{\mathcal{C}(z)} \sum_{j=1}^k p_{\boldsymbol{\mu}}(x^k) \frac{d \log p_{\mu_j}(x_j)}{d\mu_j} \alpha_j d\rho(x^{k-1})$$

$$= \int_{\mathcal{C}(z)} \sum_{j=1}^k p_{\boldsymbol{\mu}}(x^k) \left(I(\mu_j) x_j - \mu_j I(\mu_j) \right) \alpha_j d\rho(x^{k-1})$$

where $I(\mu_j)$ is the Fisher information. The final equality follows because, with $\lambda(\mu_j)$ the canonical parameter corresponding to μ_j, we have $d\lambda(\mu_j)/d\mu_j = I(\mu_j)$ and $dA(\beta)/d\beta \mid_{\beta=\lambda(\mu_j)} = \mu_j$; see e.g. [35, Chapter 18]. Now

$$g_z'(0) = \int_{\mathcal{C}(z)} \sum_{j=1}^k p_{\langle \mu_0 \rangle}(x^k) \left(I(\mu_0) x_j - \mu_0 I(\mu_0) \right) \alpha_j d\rho(x^{k-1})$$

$$= \int_{\mathcal{C}(z)} p_{\langle \mu_0 \rangle}(x^k) I(\mu_0) \sum_{j=1}^k x_j \alpha_j d\rho(x^{k-1}) \tag{4.C.10}$$

$$= I(\mu_0) \cdot \int_{\mathcal{C}(z)} p_{\langle \mu_0 \rangle}(x^k) \sum_{j=1}^k x_j \alpha_j d\rho(x^{k-1}) \tag{4.C.11}$$

where the second equality follows from $\sum_{j=1}^{k} \alpha_j = 0$. Because X^k i.i.d. $\sim P_{\mu_0}$ under $P_{\langle\mu_0\rangle}$ and the integral in (4.C.10) is over a set of exchangeable sequences, (For understanding the statement, we can consider the simple case $k = 2$, X_1 and X_2 can be exchangeable because they are 'symmetric' for given $\mathcal{C}(z)$.) we must have that (4.C.10) remains valid if we re-order the α_j's in round-robin fashion, i.e. for all $i = 1..k$, we have, with $\alpha_{j,i} = \alpha_{(j+i-1) \mod k}$,

$$g_z'(0) = I(\mu_0) \cdot \int_{\mathcal{C}(z)} p_{\langle\mu_0\rangle}(x^k) \sum_{j=1}^{k} x_j \alpha_{j,i} d\rho(x^{k-1}).$$

Summing these k equations we get, using that $\sum_{i=1}^{k} \alpha_i = 0$, that $kg_z'(0) = 0$ so that $g_z'(0) = 0$. From (4.C.9) we now see that

$$f''(0) = 0.$$

Now we compute the third derivative of $f(\delta)$, denoted as $f^{(3)}(\delta)$:

$$f^{(3)}(\delta) = \int \left(g_z^{(3)}(\delta) \log \frac{g_z(\delta)}{g_z(0)} + \frac{g_z''(\delta)g_z'(\delta)}{g_z(\delta)} \right) d\rho_{[Z]}(z)$$
$$+ \int \left(\frac{2g_z''(\delta)g_z'(\delta)g_z(\delta) - (g_z'(\delta))^3}{(g_z(\delta))^2} \right) d\rho_{[Z]}(z)$$

So since $g_z'(0) = 0$ we must also have

$$f^{(3)}(0) = 0.$$

The fourth derivative of $f(\delta)$ is now computed as follows:

$$f^{(4)}(\delta) = \int \left(g_z^{(4)}(\delta) \log \frac{g_z(\delta)}{g_z(0)} + \frac{g_z^{(3)}(\delta) \cdot g_z'(\delta)}{g_z(\delta)} \right) d\rho_{[Z]}(z)$$
$$+ \int 3 \cdot \frac{\left(g_z^{(3)}(\delta) \cdot g_z'(\delta) + (g_z''(\delta))^2 \right) g_z(\delta) - g_z''(\delta) \cdot (g_z'(\delta))^2}{(g_z(\delta))^2} d\rho_{[Z]}(z).$$

Then

$$f^{(4)}(0) = \int \frac{3 \left(g_z''(0) \right)^2}{g_z(0)} d\rho_{[Z]}(z) > 0.$$

We now have all ingredients for a fourth-order Taylor expansion of $f(\delta)$ around $\delta = 0$,

which gives:

$$\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log S_{\text{PSEUDO}} - \log S_{\text{COND}}\right] = \frac{1}{8}\int \frac{(g_z''(0))^2}{g_z(0)}d\rho_{[Z]}(z) \cdot \delta^4 + o\left(\delta^4\right)$$

which is what we had to prove. $\qquad\square$

Appendix 4.D   Proofs for Section 4.4

In this section, we prove all the statements in Table 4.1.

4.D.1   Bernoulli Family

We prove that for $\mathcal{M}$ equal to the Bernoulli family, we have $S_{\text{PSEUDO}} = S_{\text{RIPR}} = S_{\text{MIX}} \succ S_{\text{COND}}$.

Proof. We set $\mu_0^* = \frac{1}{k}\sum\limits_{i=1}^{k}\mu_i$.

$$S_{\text{MIX}} := \frac{p_{\boldsymbol{\mu}}(X^k)}{\prod\limits_{j=1}^{k}\left(\frac{1}{k}\sum\limits_{i=1}^{k}p_{\mu_i}(X_j)\right)} = \frac{p_{\boldsymbol{\mu}}(X^k)}{\prod\limits_{j=1}^{k}\left(\frac{1}{k}\sum\limits_{i=1}^{k}\left(\mu_i^{X_j}(1-\mu_i)^{1-X_j}\right)\right)} \qquad (4.\text{D}.1)$$

$$= \frac{p_{\boldsymbol{\mu}}(X^k)}{\prod\limits_{j=1}^{k}((\mu_0^*)^{X_j}(1-\mu_0^*)^{1-X_j})}$$

$$= \frac{p_{\boldsymbol{\mu}}(X^k)}{\prod\limits_{j=1}^{k}p_{\mu_0^*}(X_j)} = S_{\text{PSEUDO}} \qquad (4.\text{D}.2)$$

where the third equality holds since $X_i \in \{0,1\}$. So S_{PSEUDO} is an E-variable and $S_{\text{PSEUDO}} = S_{\text{RIPR}}$ according to Theorem 10. Then the claim follows using (4.3.1) together with the fact that when $Z = 0$ or $Z = 2$, we have $S_{\text{COND}} = 1$, while this is not true for the other e-variables, so that $S_{\text{COND}} \neq S_{\text{RIPR}} = S_{\text{PSEUDO}} = S_{\text{MIX}}$. The result then follows from (4.3.1). $\qquad\square$

4.D.2   Poisson and Gaussian Family With Free Mean and Fixed Variance

We prove that for $\mathcal{M}$ equal to the family of Gaussian distributions with free mean and fixed variance σ^2, we have $S_{\text{PSEUDO}} = S_{\text{RIPR}} = S_{\text{COND}} \succ S_{\text{MIX}}$. The proof that the same holds for $\mathcal{M}$ equal to the family of Poisson distributions is omitted, as it is completely analogous.

Proof. Note that if we let $Z := \sum_{i=1}^{k} X_i$, then we have that $Z \sim \mathcal{N}(\sum_{i=1}^{k} \mu_i, k\sigma^2)$ if $X^k \sim P_{\boldsymbol{\mu}}$. Let μ_0^* be given by (4.2.3) relative to fixed alternative $P_{\boldsymbol{\mu}}$ as in the definition of S_{PSEUDO} underneath (4.2.3). Since $k\mu_0^* = \sum_{i=1}^{k} \mu_i$, we have that Z has the same distribution for $X^k \sim P_{\langle \mu_0^* \rangle}$. This can be used to write

$$S_{\text{COND}} = \frac{p_{\boldsymbol{\mu}}\left(X^k \mid Z\right)}{p_{\langle \mu_0^* \rangle}\left(X^k \mid Z\right)} = \frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle \mu_0^* \rangle}\left(X^k\right)} \frac{p_{\langle \mu_0^* \rangle}(Z)}{p_{\boldsymbol{\mu}}(Z)} = \frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle \mu_0^* \rangle}\left(X^k\right)} = S_{\text{PSEUDO}}.$$

Therefore, S_{PSEUDO} is also an e-variable, so we derive that $S_{\text{PSEUDO}} = S_{\text{RIPR}}$ by Theorem 10. Furthermore, we have that the denominator of S_{MIX} is given by a different distribution than $p_{\langle \mu_0^* \rangle}$, so that $S_{\text{MIX}} \neq S_{\text{RIPR}} = S_{\text{PSEUDO}} = S_{\text{COND}}$. The result then follows from (4.3.1). $\qquad\square$

4.D.3   The Families for Which S_{pseudo} Is Not an E-variable

Here, we prove that S_{PSEUDO} is not an e-variable for $\mathcal{M}$ equal to the family of beta distributions with free β and fixed α. It then follows from (4.3.1) that $S_{\text{PSEUDO}} \succ S_{\text{RIPR}}$. (4.3.1) also gives $S_{\text{RIPR}} \succeq S_{\text{MIX}}$ and $S_{\text{RIPR}} \succeq S_{\text{COND}}$. The same is true for $\mathcal{M}$ equal to the family of geometric distributions and the family of Gaussian distributions with free variance and fixed mean, as the proof that S_{PSEUDO} is not an e-variable is entirely analogous to the proof for the beta distributions given below. In all of these cases, one easily shows by simulation that in general, $S_{\text{RIPR}} \neq S_{\text{MIX}}$ and $S_{\text{RIPR}} \neq S_{\text{COND}}$, so then $S_{\text{RIPR}} \succ S_{\text{MIX}}$ and $S_{\text{RIPR}} \succ S_{\text{COND}}$ follow.

Proof. First, let $Q_{\alpha,\beta}$ represent a beta distribution in its standard parameterization, so that its density is given by

$$q_{\alpha,\beta}(u) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-1}(1-u)^{\beta-1}, \qquad \alpha, \beta > 0; u \in [0, 1].$$

To simplify the proof, we assume $\alpha = 1$ here. Then

$$q_{1,\beta}(u) = \frac{\Gamma(1+\beta)}{\Gamma(\beta)}(1-u)^{\beta-1} = \frac{1}{1-u} \exp\left(\beta \log(1-u) - \log\frac{1}{\beta}\right)$$

where the first equality holds since $\Gamma(1 + \beta) = \beta\Gamma(\beta)$. Comparing this to (4.1.1), we see that β is the canonical parameter corresponding to the family $\{Q_{1,\beta} : \beta > 0\}$, and we have

$$\lambda(\mu) = \beta, \quad t(u) = \log(1-u), \quad A(\beta) = \log\frac{1}{\beta}.$$

To prove the statement, according to Proposition 11, we just need to show, for any $\mu_1, \ldots, \mu_k$ that are not all equal to each other, that, with $X = t(U) = \log(1 - U)$ and

$\mu_0^* = \frac{1}{k} \sum_{i=1}^{k} \mu_i$ defined as in (4.2.3), we have

$$\sum_{i=1}^{k} \text{VAR}_{P_{\mu_i}}[X] - k\text{VAR}_{P_{\mu_0^*}}[X] > 0. \tag{4.D.3}$$

Straightforward calculation gives

$$\text{VAR}_{P_{\mu_i}}[X] = \text{VAR}_{Q_{1,\beta_i}}[X] = \frac{d^2}{d^2\beta_i}(\log\frac{1}{\beta_i}) = \frac{1}{\beta_i^2} \text{ in particular } \text{VAR}_{P_{\mu_0^*}}[X] = \frac{1}{(\beta_0^*)^2} \tag{4.D.4}$$

where β_i corresponds to μ_i, i.e. $\mathbb{E}_{Q_{1,\beta_i}}[(X)] = \mu_i$. We also have:

$$\mathbb{E}_{P_{\beta_0^*}}[(X)] = \mu_0^* = \frac{1}{k} \sum_{i=1}^{k} \mu_i = \frac{1}{k} \sum_{i=1}^{k} \mathbb{E}_{P_{\beta_i}}[(X)]. \tag{4.D.5}$$

While $\mathbb{E}_{P_{\beta_i}}[(X)] = \frac{d}{d\beta_i}(\log\frac{1}{\beta_i}) = -\frac{1}{\beta_i}$, therefore $\frac{1}{\beta_0^*} = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{\beta_i}$. We obtain, together with (4.D.4) and (4.D.5), that

$$\sum_{i=1}^{k} \text{VAR}_{P_{\mu_i}}[(X)] - k\text{VAR}_{P_{\mu_0^*}}[(X)] = \sum_{i=1}^{k} \frac{1}{(\beta_i)^2} - k\left(\frac{1}{k}\sum_{i=1}^{k}\frac{1}{\beta_i}\right)^2. \tag{4.D.6}$$

Jensen's inequality now gives that (4.D.6) is strictly positive, whenever at least one of the μ_i is not equal to μ_0^*, which is what we had to show. $\qquad\square$
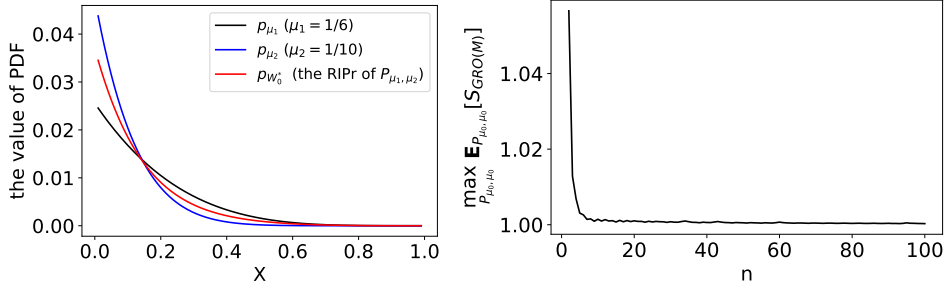
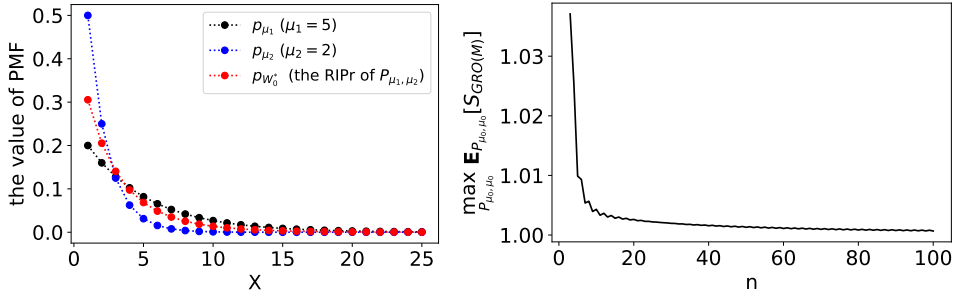# Appendix 4.E   Graphical Depiction of RIPr-Approximation and Convergence of Li's Algorithm
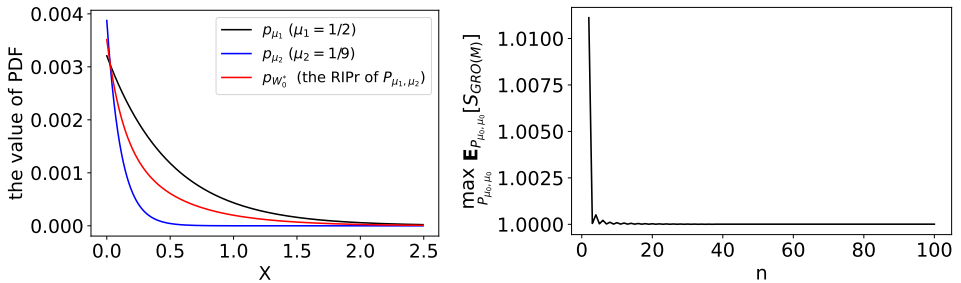
Figure 4.2: Exponential distribution. On the right, n represents number of iterations with Li's algorithm, starting at iteration 2

## 4.E. Graphical Depiction of RIPr-Approximation and Convergence of Li's Algorithm

Figure 4.3: beta with free β and fixed α. On the right, n represents number of iterations with Li's algorithm, starting at iteration 2

Figure 4.4: geometric distribution. On the right, n represents number of iterations with Li's algorithm, starting at iteration 3

We illustrate RIPr-approximation and convergence of Li's algorithm with four distributions: exponential, beta with free β and fixed α, geometric and Gaussian with free variance and fixed mean, each with one particular (randomly chosen) setting of the parameters. The pictures on the left in Figure 4.2– 4.5 give the probability density functions (for geometric distributions, discrete probability mass functions) after $n = 100$ iterations of Li's algorithm. The pictures on the right illustrate the speed of convergence of Li's algorithm. The pictures on the right do not show the first (or the first two, for geometric and Gaussian with free variance) iteration(s), since the worst-case expectation $\sup_{\mu_0 \in \mathtt{M}}[S_{\mathrm{RIPR}}]$ is invariably incomparably larger in these initial steps. We empirically find that Li's algorithm converges quite fast for computing the true S_{RIPR}. In each step of Li's algorithm, we searched for the best mixture weight α in $P_{(m)}$ over a uniformly spaced grid of 100 points in $[0, 1]$, and for the novel component $P' = P_{\mu',\mu'}$ by searching for μ' in a grid of 100 equally spaced points inside the parameter space $\mathtt{M}$ where the left- and right- endpoints of the grid were determined by trial and error. While with this ad-hoc discretization strategy we obviously cannot guarantee any formal approximation results, in practice it invariably worked well: in all cases, we found that $\max_{\mu_0 \in \mathtt{M}} \mathbb{E}_{P_{\mu_0,\mu_0}}[S_{\mathrm{RIPR}}] \leq 1.005$ after 15 iterations.
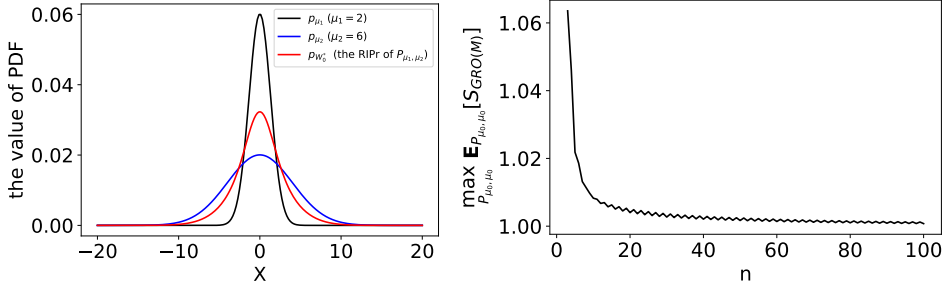
Figure 4.5: Gaussian with free variance and fixed mean. On the right, n represents number of iterations with Li's algorithm, starting at iteration 3

Distributions	(μ_1, μ_2)	α	(μ_{01}, μ_{02})	$\sup\limits_{\mu_0 \in \mathbb{M}} \mathbb{E}_{X_1, X_2 \sim P_{\mu_0, \mu_0}}[S]$
beta	$(\frac{1}{6}, \frac{1}{10})$	0.57	$(0.12, 0.16)$	1.00071
geometric	$(5, 2)$	0.39	$(2.52, 4.21)$	1.00035
Exponential	$(\frac{1}{2}, \frac{1}{9})$	0.53	$(0.13, 0.51)$	1.00083
Gaussian with free variance and fixed mean	$(2, 6)$	0.41	$(5.82, 3.36)$	1.00035

Table 4.3: Analogue of Table 4.2 for μ_1, μ_2 corresponding to the parameters used in Figures 4.2–4.5

For comparison, we show the best approximation that can be obtained by brute-force combining of just two components, for the same parameter values, in Table 4.3.

Supplementary Material

In this supplement we verify that all conditions are met for the implicit use of Fubini's theorem and differentiation under the integral sign in the proofs of Theorem 2 and 3, and that all derivatives of interest are bounded.

Theorem 2

In the paper, notation is as follows:

$$\mu_j = \mu_0 + \delta\alpha_j$$
$$\lambda(\mu_j) = \text{nat. param. } \lambda \text{ corresponding to mean } \mu = \mu_j$$
$$p_\mu(y) = e^{\lambda(\mu)y - A(\lambda(\mu))}$$
$$f_y(\delta) = \sum_{j=1}^{k} p_{\mu_i}(y).$$

4.E. Graphical Depiction of RIPr-Approximation and Convergence of Li's Algorithm

As this will simplify the notation for the derivatives, we write $g_y(\lambda) = e^{\lambda y - A(\lambda)}$, so that

$$f_y(\delta) = \sum_{j=1}^{k} g_y(\lambda(\mu_j)) \text{ and } p_{\mu_0}(y) = g_y(\lambda(\mu_0)). \tag{4.E.1}$$

To stress dependence on δ, we write $\mu_j(\delta)$ instead of μ_j in the following.

Step 1 We first establish the finiteness condition (4.C.2). We note that

$$\log \sum_{j=1}^{k} g_y(\lambda(\mu_j(\delta))) \leq \log(\max_j g_y(\lambda(\mu_j(\delta)))k)$$

$$= \max_j \log(g_y(\lambda(\mu_j(\delta)))) + \log k$$

$$\leq \max_j \log(\max\{g_y(\lambda(\mu_j(\delta))), 1\}) + \log k$$

$$\leq \sum_j \log(\max\{g_y(\lambda(\mu_j(\delta))), 1\}) + \log k$$

$$\leq \sum_j |\lambda(\mu_j(\delta))y - \log A(\lambda(\mu_j(\delta)))| + \log k.$$

and

$$\log \sum_{j=1}^{k} g_y(\lambda(\mu_j(\delta))) = \log \frac{1}{k} \sum_{j=1}^{k} g_y(\lambda(\mu_j(\delta))) + \log k$$

$$\geq \frac{1}{k} \sum_{j=1}^{k} \log g_y(\lambda(\mu_j(\delta))) + \log k$$

$$= \frac{1}{k} \sum_{j=1}^{k} \lambda(\mu_j(\delta))y - A(\lambda(\mu_j(\delta))) + \log k.$$

Putting these together, we see that

$$|\log f_y(\delta)| \leq$$

$$\max\left\{ \sum_j |\lambda(\mu_j(\delta))y - A(\lambda(\mu_j(\delta)))| + \log k, \left| \frac{1}{k} \sum_{j=1}^{k} (\lambda(\mu_j(\delta))y - A(\lambda(\mu_j(\delta)))) + \log k \right| \right\}$$

$$\leq \sum_j |\lambda(\mu_j(\delta))y - A(\lambda(\mu_j(\delta)))| + \log k, \tag{4.E.2}$$

and, more trivially,

$$|\log g_y(\lambda(\mu_0))| \leq |\lambda(\mu_0)y - A(\lambda(\mu_0)|. \tag{4.E.3}$$

We know that $\lambda(\mu_j(\delta))$ and $A(\lambda(\mu_j(\delta)))$ are smooth, hence finite functions for $\mu_j(\delta)$ in the interior of the mean-value parameter space $\mathbb{M}$ (see [13, Chapter 9, Theorem 9.1 and Eq. (2)]). Since $\mathbb{M}$ is open and for all $j = 1..k$, $\mu_j(0) = \mu_0 \in \mathbb{M}$, it follows that $|\log f(y)(\delta) - \log g_y(\lambda(\mu_0))|$ can be written as a smooth, in particular finite function of $|y|$ for all δ in a compact subset of $\mathbb{R}$ with 0 in its interior. Since $|y| \leq 1 + y^2$ has finite expectation under all P_μ with $\mu \in \mathbb{M}$, finiteness of (4.C.2) follows by (4.E.1).

Step 2   We now proceed to establish that we can differentiate with respect to δ for δ in a compact subset of $\mathbb{R}$ with 0 in its interior. The proof will make use of (4.E.2) and (4.E.3). We denote derivatives of functions f_y and g_y as

$$g_y^s(\lambda) = \frac{\mathrm{d}^s}{\mathrm{d}\lambda^s} g_y(\lambda) \quad \text{and} \quad f_y^s(\delta) = \frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f_y(\delta).$$

We will argue that, for any $s \in \mathbb{N}$, the family $\{\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f_y(\delta) \log f_y(\delta) - f_y(\delta) \log g_y(\lambda(\mu_0)) : \delta \in \Delta\}$ is uniformly integrable for any compact $\Delta \subset \mathbb{R}$, so that we are allowed to interchange differentiation and integration [see e.g. 98, Chapter A16].

Using standard results for exponential families, we have, for λ in the interior of the canonical parameter space,

$$g_y^{(1)}(\lambda) = (y - \mu(\lambda))g_y(\lambda)$$
$$g_y^{(2)}(\lambda) = -I(\lambda)g_y(\lambda) + (y - \mu(\lambda))^2 g_y(\lambda),$$

where $\mu(\lambda)$ denotes the mean-value parameter corresponding to λ and $I(\lambda)$ the corresponding Fisher information.

Continuing this using the fact that $(d^s/d\lambda^s)A(\lambda)$ is continuous for all s, gives

$$g_y^{(s)}(\lambda) = g_y(\lambda) \cdot h_{y,s}(\lambda) \text{ with } h_{y,s}(\lambda) = \sum_{t=1}^{s} h_{[t,s]}(\lambda)(y - \mu(\lambda))^t \qquad (4.\text{E}.4)$$

for some smooth functions $h_{[1,s]}, h_{[2,s]}, \ldots, h_{[s,s]}$ of λ (we do not need to know precise definitions of these functions). Similarly

$$f_y^{(1)}(\delta) = \sum_j g_y^{(1)}(\lambda_{\mu_j(\delta)}) \cdot (\lambda(\mu_j(\delta)))'$$

where $\lambda(\mu_j(\delta))' = \frac{\mathrm{d}}{\mathrm{d}\delta}\lambda(\mu_j(\delta))$. We know that $\lambda'(\mu_j(\delta))$ and further derivatives are smooth functions for $\mu_j(\delta)$ in the interior of the mean-value parameter space $\mathbb{M}$ (see [13, Chapter 9, Theorem 9.1 and Eq. (2)]). Since this space is open and for all $j = 1..k$, $\mu_j(0) = \mu_0 \in \mathbb{M}$, it follows that $\lambda'(\mu_j(\delta))$ are smooth functions of δ for δ in a compact subset of $\mathbb{R}$ with 0 in its interior. Thus, analogously to what we did above with $g^{(s)}$, we get that

$$f_y^{(s)}(\delta) = \sum_j \sum_{t=1}^{s} g_y^{(t)}(\lambda(\mu_j(\delta))) \cdot r_{t,s}(\mu_j) \qquad (4.\text{E}.5)$$

for some smooth functions $r_{t,s}$, the details of which we do not need to know. In particular this gives, with

$$b_y^{(s)} := \frac{f_y^{(s)}(\delta)}{f_y(\delta)}$$

that

$$\left| b_y^{(s)} \right| \leq \frac{\sum_j g_y(\lambda(\mu_j(\delta))) \cdot \left(\sum_{t=1}^s |h_{y,t}(\lambda(\mu_j(\delta))) \cdot r_{t,s}(\mu_j(\delta))| \right)}{\sum_j g_y(\lambda(\mu_j(\delta)))}$$

$$\leq \sum_j \sum_{t=1}^s |h_{y,t}(\lambda(\mu_j(\delta))) \cdot r_{t,s}(\mu_j(\delta))|.$$

Inspecting the proof in the main text, we informally note that all terms without logarithms in the first four derivatives of $F_0(\delta)$ and $F_1(\delta)$ can be written as products $f_y(\delta) \cdot b_y^{(s_1)}(\delta) \cdot \ldots \cdot b_y^{(s_u)}(\delta)$ for the $b_y^{(s)}$ we just bounded in terms of polynomials in $|y|$; similarly, the terms involving logarithms can be bounded in terms of such polynomials as well using (4.E.2) and (4.E.3), suggesting that all terms inside all integrals can be such bounded. This is indeed the case: formalizing the reasoning, we see that

$$\int \left(\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f_y(\delta) \log f_y(\delta) - f_y(\delta) \log g_y(\lambda(\mu_0)) \right)^2 d\rho(y) =$$

$$\int \left(f_y^{(s)}(\log f_y(\delta) - \log g_y(\lambda(\mu_0))) + f_y(\delta) \sum_u c_u \cdot b_y^{(s_2)}(\delta) \cdot \ldots \cdot b_y^{(s_u)}(\delta) \right)^2 d\rho(y)$$

$$= \int (f_y^{(s)}(\log f_y(\delta) - \log g_y(\lambda(\mu_0))))^2 + \left(f_y(\delta) \sum_u c_u \cdot b_y^{(s_1)}(\delta) \cdot \ldots \cdot b_y^{(s_u)}(\delta) \right)^2$$

$$+ f_y(\delta) f_y^{(s)}(\log f_y(\delta) - \log g_y(\lambda(\mu_0))) \sum_u c_u \cdot b_y^{(s_1)}(\delta) \cdot \ldots \cdot b_y^{(s_u)}(\delta) d\rho(y).$$

By (4.E.2) and (4.E.3) and the bound on $|b_y^{(s)}|$ given above, all the terms within the integral can be bounded by polynomials in y (or $|y|$), so the integral is given by linear functions of moments of ρ and P_μ. Therefore, using also that ρ is itself a probability measure and a member of the exponential family under consideration (equal to P_μ with $\lambda(\mu) = 0$), the integral can be uniformly bounded over δ in a compact subset of the mean-value parameter space. It follows that the family $\{ \frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f_y(\delta) \log f_y(\delta) - f_y(\delta) \log g_y(\lambda(\mu_0)) : \delta \in \Delta \}$ is uniformly integrable [see e.g. 98, Chapter 13.3], so integration and differentiation may be interchanged freely [see e.g. 98, Chapter A16]. It also follows that the quantity on the right-hand side in the theorem statement is bounded.

Theorem 3

As in the proof of Theorem 3, let $f(\delta) = \mathbb{E}_{P_\mu} \left[\log \frac{p_\mu(X^k)}{p_{\langle \mu_0 \rangle}(X^k)} - \log \frac{p_\mu(X^k|Z)}{p_{\langle \mu_0 \rangle}(X^k|Z)} \right]$.

To validate the proof in the main text we merely need to show that $f(\delta)$ is finite, and that we can interchange differentiation and expectation with respect to δ in a compact interval containing $\delta = 0$. Thus, we want to show that, for any $s \in \mathbb{N}$, we have that

$$\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f(\delta) = \mathbb{E}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\left(\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle\mu_0\rangle}(X^k)} - \log \frac{p_{\boldsymbol{\mu}}(X^k \mid Z)}{p_{\langle\mu_0\rangle}(X^k \mid Z)}\right)\right].$$

To show this, first note that both $\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle\mu_0\rangle}(X^k)}\right]$ and $\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log \frac{p_{\boldsymbol{\mu}}(X^k|Z)}{p_{\langle\mu_0\rangle}(X^k|Z)} \mid Z\right]$ are KL divergences between members of exponential families (the fact that conditioning on a sum of sufficient statistics results in a new, derived full exponential family is shown by, for example, [19]), which are finite as long as δ is in a sufficiently small interval containig 0 in its interior (since then μ is in the interior of the mean-value parameter space). This already shows that $f(\delta)$ is finite, and it also allows us to rewrite

$$f(\delta) = \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle\mu_0\rangle}(X^k)}\right] - \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log \frac{p_{\boldsymbol{\mu}}(X^k \mid Z)}{p_{\langle\mu_0\rangle}(X^k \mid Z)}\right].$$

Furthermore, [19, Theorem 2.2] in combination with Theorem 9.1. and Chapter 9, Eq.2. of [13] shows that for any full exponential family, for any finite $k > 0$, the k-th derivative of the KL divergence with respect to its first argument, given in the mean-value parameterization, exists, is finite, and can be obtained by differentiating under the integral sign, at any μ in the interior of the mean-value parameter space. We are therefore allowed to interchange expectation and differentiation for such terms separately for all δ in any compact interval containing 0. Thus, starting with the previous display, we can write

$$\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f(\delta) = \frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle\mu_0\rangle}(X^k)}\right] - \frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log \frac{p_{\boldsymbol{\mu}}(X^k \mid Z)}{p_{\langle\mu_0\rangle}(X^k \mid Z)}\right]$$

$$= \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle\mu_0\rangle}(X^k)}\right] - \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu}}(X^k \mid Z)}{p_{\langle\mu_0\rangle}(X^k \mid Z)}\right]$$

$$= \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle\mu_0\rangle}(X^k)}\right] - \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle\mu_0\rangle}(X^k)} + \log \frac{p_{\boldsymbol{\mu};[Z]}(Z)}{p_{\langle\mu_0\rangle;[Z]}(Z)}\right] =$$

$$\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle\mu_0\rangle}(X^k)}\right] - \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle\mu_0\rangle}(X^k)}\right] + \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu};[Z]}(Z)}{p_{\langle\mu_0\rangle;[Z]}(Z)}\right]$$

$$= \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu};[Z]}(Z)}{p_{\langle\mu_0\rangle;[Z]}(Z)}\right],$$

where in the last line we use that all involved terms are finite. This is what we had to show.

4.5. Graphical Depiction of RIPr-Approximation and Convergence of Li's Algorithm