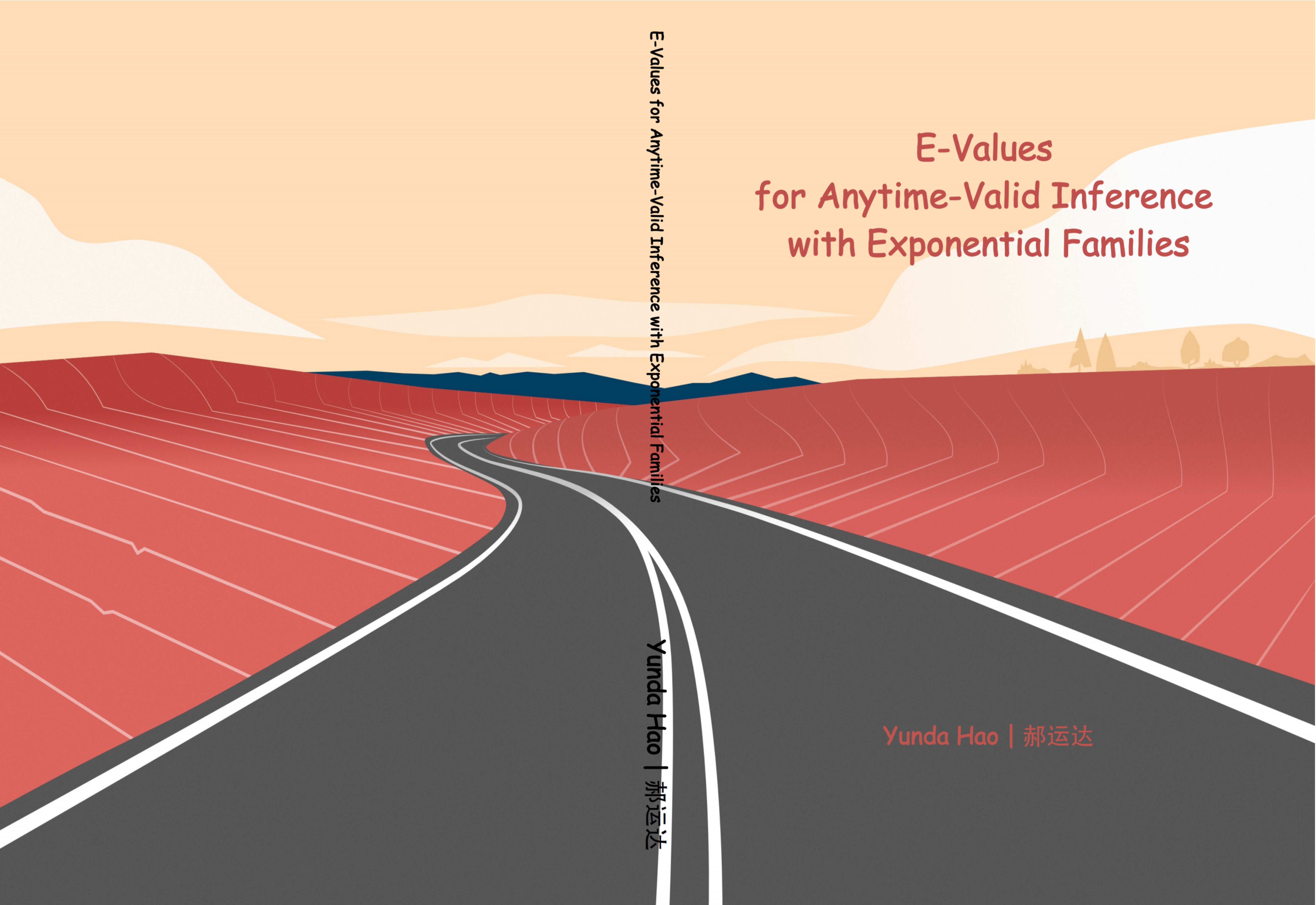


E-Values for Anytime-Valid Inference with Exponential Families

Yunda Hao | 郝运达

E-Values for Anytime-Valid Inference with Exponential Families

Yunda Hao | 郝运达

E-Values for Anytime-Valid Inference with Exponential Families

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op dinsdag 18 februari 2025
klokke 14:30 uur

door

Yunda Hao
geboren te Cangzhou, Hebei Province, China
in 1994

Promotores:
Prof.dr. P.D. Grünwald (Universiteit Leiden and Centrum Wiskunde & Informatica)
Dr. A. Ly (Centrum Wiskunde & Informatica)

Promotiecommissie:
Prof.dr. T. Dickhaus (Bremen University)
Dr. Z. Ren (University of Pennsylvania)
Dr. N.W. Koning (Erasmus University Rotterdam)
Prof.dr.ir. G.L.A. Derks
Prof.dr. M. Fiocco

The research in this dissertation was funded by the China Scholarship Council (Grant No. 202006280045). The research was performed at Centrum Wiskunde & Informatica (CWI) in Amsterdam, the national research institute for mathematics and computer science in the Netherlands.

ISBN: 978-90-6196-429-2

The cover is from https://pixabay.com/.

Universiteit Leiden
The Netherlands

CWI
Centrum Wiskunde & Informatica

Contents

Chapter 1

Introduction

In today's world, data surrounds us like air. For instance, devices like the Apple Watch continuously collect data about our daily routines, exercise habits, and more. However, this vast amount of data must be organized and analyzed before meaningful conclusions can be drawn. Solving practical problems with data is at the core of the science called *statistics*. Over the years, statistics has demonstrated its value across various research fields, such as bioinformatics and sociological research. In the era of big data, where new data constantly accumulates, we need statistical methods that can handle these continuous data flows and enable real-time decision-making.

This dissertation focuses on an *anytime-valid method* (called the *e-value method*), a powerful approach designed to tackle hypothesis testing problems within streaming data contexts. Throughout this work, we develop numerous mathematical results concerning the theory of the e-value method for hypothesis testing. In this introductory chapter, we introduce the key topics covered in the dissertation.

First, Section 1.1 introduces the hypothesis testing problem and discusses how it is addressed by classical methods. We then illustrate the problems that arise when these traditional methods are used sequentially as the data come in. In Section 1.2, we present the core concept of the e-value method and explain why it works "safely" for real-time decision-making. Section 1.3 covers preliminary knowledge that is frequently referenced in later chapters but not introduced in detail in those chapters. Section 1.4 provides an introduction to the exponential families, which are a central focus of this dissertation. Lastly, in Section 1.5, we offer an outline of each chapter of the dissertation.

1.1 Hypothesis testing

Hypothesis testing is a common practice in everyday life. For example, during winter, I often feel discomfort in my stomach after having lunch at the CWI canteen. Since I rarely ate cold food in China during winter, I suspect that the cold vegetable salad might be causing this discomfort. To *test* my suspicion, I eat the cold salad on some random days over the course of a month and monitor how my stomach feels each day. If my stomach consistently feels bad after eating the cold salad but feels fine on other days, this would provide strong evidence that my suspicion is correct. Of course, one could argue that it might just be coincidence, but such coincidence would have an extremely low probability if my suspicion is wrong.

To explain hypothesis testing further, suppose we have collected n observations, denoted $X_{(1)}, X_{(2)}, \ldots$. We are interested in whether these observations are consistent with one of two hypotheses: the *null hypothesis* ($\mathcal{H}_0$) or the *alternative hypothesis* ($\mathcal{H}_1$). In the example above, $\mathcal{H}_0$ might be “Eating cold vegetable salad does NOT cause stomach discomfort”, while $\mathcal{H}_1$ would be “Eating cold vegetable salad does cause stomach discomfort”. In general, $\mathcal{H}_0$ represents a status quo assumption or a standard model that the data might conform to, while $\mathcal{H}_1$ represents a departure from $\mathcal{H}_0$.

$\mathcal{H}_0$ and $\mathcal{H}_1$ are usually formalized as probability distributions. For example, the data $X_{(1)}, X_{(2)}, \ldots$ are independent and identically distributed (i.i.d.). Each $X_{(i)}$ is of the form $(Y_{(i)}, G_{(i)})$ with $Y_{(i)} \in \{\textsc{Feel good}, \textsc{Feel bad}\}$ and $G_{(i)} \in \{\textsc{Eat salad}, \textsc{Not eat salad}\}$. Then $\mathcal{H}_0$ is the set of conditional distributions with

$$\begin{aligned}&\Pr(Y_{(i)} = \textsc{Feel good} | G_{(i)} = \textsc{Eat salad})\\ =&\Pr(Y_{(i)} = \textsc{Feel good} | G_{(i)} = \textsc{Not eat salad});\end{aligned}$$

and $\mathcal{H}_1$ is the set of conditional distributions, in which the first probability is smaller than the second one.

There are many approaches to hypothesis testing, roughly categorized as *frequentist* or *Bayesian* methods, as explained in detail by Royall [74]. The frequentist approach is further divided into *Fisherian* and *Neyman-Pearson tests.* Fisherian testing focuses on measuring the evidence against $\mathcal{H}_0$ using the p-value (defined later); the smaller the p-value, the stronger the evidence against $\mathcal{H}_0$. In this framework, there is no explicit $\mathcal{H}_1$, as the test focuses solely on $\mathcal{H}_0$. On the other hand, the Neyman-Pearson approach explicitly compares two hypotheses, $\mathcal{H}_0$ and $\mathcal{H}_1$, with the goal of choosing one over the other. We will omit Bayesian methods here because we do not use them so much in the thesis.

In some chapters in the thesis, we use $\mathcal{H}_0$ to denote the statement defining the null hypothesis (Chapter 2 and Chapter 3), and $\mathcal{P}$ to denote the corresponding set of distributions. Similarly, we use $\mathcal{H}_1$ to denote the statement defining the alternative, and $\mathcal{Q}$ to denote the distributions. In other chapters including this introduction, we use $\mathcal{H}_0$ and $\mathcal{H}_1$ to directly denote both the set of distributions and their defining statement.

The general goal in all chapters of this thesis is to choose between a null hypothesis $\mathcal{H}_0$ and an alternative $\mathcal{H}_1$ based on the observations $x_{(1)}, x_{(2)}, \ldots$. Both $\mathcal{H}_1$ and $\mathcal{H}_0$ can be composite, meaning they may consist of multiple possible distributions rather than a single, fixed one. For example, consider testing whether a coin is fair. The null hypothesis $\mathcal{H}_0$ asserts that the coin is fair, meaning the probability of getting “heads” in a coin toss is exactly 0.5. In this case, the null model space is simple, corresponding to a Bernoulli distribution with parameter 0.5. $\mathcal{H}_0$ is referred to as a simple hypothesis. On the other hand, the alternative hypothesis $\mathcal{H}_1$ claims the coin is not fair. This means the probability of getting “heads” could be any value except 0.5, specifically any real number $\theta \in [0, 0.5) \cup (0.5, 1]$. Since this allows for a range of different probabilities, $\mathcal{H}_1$ is referred to as a composite hypothesis.

This is one of the most common settings today. While we will not cover all classical hypothesis testing methods, we will introduce one of the most well-known: the p-value method.

p-value A *p-value* is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis $\mathcal{H}_0$ is true. Mathematically, a strict p-value is a random variable p such that for any given significance level $\alpha \in [0, 1]$ and for all $P_0 \in \mathcal{H}_0$,

$$P_0(p \leq \alpha) = \alpha.$$

In practice, for a set of observations $x^{(n)} := x_{(1)}, \ldots, x_{(n)}$, we *reject* $\mathcal{H}_0$ if the p-value computed from these observations is less than or equal to α, which is commonly set at 0.05. Otherwise, we fail to reject $\mathcal{H}_0$.

To better understand this, let us look at a couple of classical examples.

Example 1. [Correct coin toss test] This test examines whether a coin is fair. The same example is discussed in Pérez-Ortiz's PhD dissertation [69]. We toss the coin n times, and from these observations $x^{(n)} = x_{(1)}, \ldots, x_{(n)}$, we record each time whether "heads" or "tails" appears. The empirical mean is defined as:

$$t(x^{(n)}) = \frac{1}{n}\#\{\text{heads in } x^{(n)}\}.$$

The number of heads follows a binomial distribution, and by the Central Limit Theorem (CLT), if $\mathcal{H}_0$ (the coin is fair) is true, $t(X^{(n)})$ approximately complies with a Gaussian distribution.

Suppose we have an observation $x^{(n)}$ and get $t(x^{(n)}) = 1/2 + t^*$. Then the event "If we would replicate the experiment, we get a result $t(X^{(n)})$ that is at least as extreme as the result actually observed" is given by $t(X^{(n)}) \in (-\infty, 1/2 - t^*] \cup [1/2 + t^*, \infty)$. Note that now $X^{(n)}$ refers to "new" data, whereas $x^{(n)}$ refers to the actually observed data—see the red intervals in Figure 1.1. Then the p-value corresponding to statistic $t(x^{(n)})$ is defined as the probability that $t(X^{(n)})$ falls within the red intervals. For each value of t^*, there is a corresponding p-value. If $t^* = 0.98/\sqrt{n}$, then the p-value is 0.05.

The p-value method would reject $\mathcal{H}_0$ if $t(X^{(n)})$ falls inside these red intervals, resulting in a *Type-I error* (i.e., false rejection of $\mathcal{H}_0$) guarantee with a probability of 0.05. That is

$$\text{Type-I error} := \Pr(\textsc{reject } \mathcal{H}_0 | \mathcal{H}_0 \textsc{ is true}),$$

which equals 0.05 in this example. This demonstrates the validity of the p-value method in this example, as it controls the Type-I error at a significance level 0.05.

Example 2. [Incorrect coin toss test] If we extend the previous example to an online streaming case, the p-value method violates the Type-I error control. Here, we keep tossing the coin, collecting new observations, and computing $t(x^{(n)})$ at each step

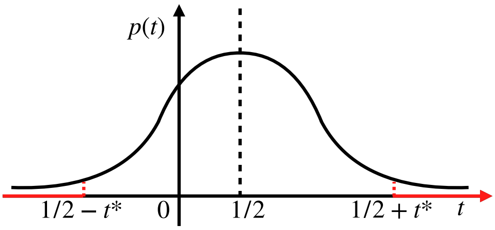


Figure 1.1: If the p-value $p = 0.05$, then $t^* = 0.98/\sqrt{n}$.

until $t(x^{(n)})$ eventually falls outside the interval $[1/2 - 0.98/\sqrt{n}, 1/2 + 0.98/\sqrt{n}]$. Due to the randomness in sampling, extreme values will occur eventually, leading us to reject $\mathcal{H}_0$ even if $\mathcal{H}_0$ is true. In this scenario, the Type-I error becomes 1, showing that the p-value method is no longer valid. This will be illustrated in Figure 1.2.

We will demonstrate that the e-value method remains valid in the scenario described in the above examples, to be discussed in Section 1.2. Furthermore, to illustrate why the Type-I error exceeds α in Example 2, we provide a clearer explanation in the next example.

Example 3. **[Multi-stage tests]** Consider the following multi-stage experiment. Initially, researchers collect a dataset $X^{(n)}$ and compute the p-value p_1. They reject $\mathcal{H}_0$ if $p_1 < 0.05$, and accept $\mathcal{H}_0$ if $p_1 \geq 0.1$. However, if $0.05 \leq p_1 < 0.1$, they deem the evidence inconclusive but promising. Therefore, they collect a new sample $X'^{(m)}$ and compute a new p-value p_2 based on joining datasets $X^{(n)}$ and $X'^{(m)}$. They reject $\mathcal{H}_0$ if $p_2 \leq 0.05$, otherwise they accept $\mathcal{H}_0$.

Let us represent the event where $0.05 < p_1 < 0.1$ as G. The total Type-I error exceeds 0.05 because:

$$\text{Type-I error} = 0.05 + P_0(G) \cdot P_0(p_2 < 0.05 \mid G) > 0.05.$$

This shows that the p-value method fails to control the Type-I error in the multi-stage testing, which already happens when there are just 2 stages. In practice, there are often multiple stages, which further increases the overall Type-I error.

In the next subsection, we will show that the e-value method consistently succeeds in the scenarios presented in these examples. More generally, the test process using the e-value method can be halted at any time without requiring a predefined stopping rule, offering greater flexibility compared to traditional methods.

1.2 Anytime-valid tests: e-value, e-process

In Section 1.1, we discussed a flaw of the p-value in sequential testing. The issue of sampling until a significant result is obtained has actually been debated by statisticians since at least the 1940s. Feller in 1940 [34] observed this issue in studies of extra-sensory perception, and Anscombe in 1954 [4] famously called it "sampling to a foregone conclusion." Robbins in 1952 [73] also pointed out this problem. However, it was not until 2019 that a fully general framework emerged to address it. In that year, four papers from different research groups were published on arXiv, laying the foundation for what would soon become known as the concept of the *e-value* [42, 95, 91, 76]. We will now introduce this general framework.

e-variable, e-value Consider a batch of data (a random vector) X, which can be collected sequentially or all at once. We define a nonnegative statistic $S(X)$, that is a function of the observed data. Let $\mathcal{H}_0 = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ be the set of distributions for X defined within their parameter space $\boldsymbol{\Theta}$. If the statistic $S(X)$ satisfies the condition:

$$\text{for all } P \in \mathcal{H}_0 : \qquad \mathbb{E}_P\left[S(X)\right] \leq 1, \tag{1.2.1}$$

then we refer to $S(X)$ as an *e-variable relative to* $\mathcal{H}_0$. The value of $S(X)$ computed from the observed data is called the *e-value*.

Similar to how we test with p-values, we first set a significance level $\alpha \in [0, 1]$ before conducting an e-value-based test. Given a sample X, we construct an e-variable $S(X)$ using the entire sample at once, and we reject $\mathcal{H}_0$ if and only if $S(X) \geq 1/\alpha$. This is analogous to the traditional p-value approach. This approach ensures a Type-I error guarantee, as can be explained by Markov's inequality:

$$\text{Type-I error} = \Pr(S(X) \geq 1/\alpha | \mathcal{H}_0 \text{ IS TRUE}) \leq \frac{\sup\limits_{P \in \mathcal{H}_0} \mathbb{E}_P[S(X)]}{1/\alpha} \leq \alpha.$$

However, data is sometimes collected sequentially, requiring us to make decisions at any time using the data collected so far. A specific example of this is provided in Example 2 and Example 3. In contrast to p-value methods, the e-value method can also be applied for testing with a data stream while maintaining a Type-I error guarantee. We divide this situation into two types: *optional continuation* and *optional stopping*.

Optional continuation Consider a stream of data batches $X_{(1)}, X_{(2)}, \ldots$. We assume the $X_{(i)}$ are independent. Let $S(X_{(i)})$ be an e-variable based on $X_{(i)}$. Then, for any positive integer $N \in \mathbb{N}^+$, we define the product $S^{(N)} := \prod\limits_{i=1}^{N} S(X_{(i)})$, which remains an e-variable. In other words,

$$\text{for all } P \in \mathcal{H}_0 : \qquad \mathbb{E}_P\left[S^{(N)}\right] \leq 1, \tag{1.2.2}$$

which can be shown as follows: since $S(X_{(1)}), \ldots, S(X_{(N)})$ are independent, it follows that for all $P \in \mathcal{H}_0$,

$$\mathbb{E}_P\left[S^{(N)}\right] = \prod_{i=1}^{N} \mathbb{E}_P\left[S(X_{(i)})\right] \leq 1.$$

Then, if we reject $\mathcal{H}_0$ when $S^{(N)} \geq 1/\alpha$, the Type I error is bounded by level α. This is because, by Markov's inequality,

$$\text{Type-I error} = \Pr\left(S^{(N)} \geq 1/\alpha \middle| \mathcal{H}_0 \text{ IS TRUE}\right) \leq \frac{\sup_{P \in \mathcal{H}_0} \mathbb{E}_P\left[S^{(N)}\right]}{1/\alpha} \leq \alpha. \tag{1.2.3}$$

This inequality holds for every fixed N. However, as shown by Grünwald et al [42], (1.2.3) still holds if N is any data dependent stopping time. A stopping time is a time determined by a rule that, at each step N, decides—based on the data observed so far, $X_1, X_2, \ldots, X_N$—whether to stop or continue. For example: "stop as soon as $S^{(N)} \geq 1/\alpha$", or "$N = 10$", or "stop if $X_{(N)}$ contains 0". For a formal definition of stopping time, we refer to Ramdas et al. [71].

In more detail, Grünwald et al [42] show that $S^{(1)}, S^{(2)}, \ldots$ is a *test supermartingale* [77], which is a nonnegative supermartingale with $\mathbb{E}_P[S^{(1)}] \leq 1$. This implies (1.2.3) and also implies that Ville's inequality [90] holds, then further implies, for all $P \in \mathcal{H}_0$, we have

$$\text{Type-I error} \leq \Pr\left(\sup_{N:N>0} S^{(N)} \geq 1/\alpha \middle| \mathcal{H}_0 \text{ IS TRUE}\right) \overset{(*)}{\leq} \frac{\mathbb{E}_P[S^{(1)}]}{1/\alpha} \leq \alpha, \tag{1.2.4}$$

where $(*)$ follows from Ville's inequality. We say that the e-value method preserves Type-I error guarantees under *optional continuation.*

Optional stopping Now suppose each $X_{(i)}$ is a single (just one) data point and we have an e-variable $S(X_{(i)})$ for each point $X_{(i)}$. But now by (1.2.4), we can do optional stopping - stop at any point we like, and still preserve Type-I error guarantees.

e-process So far, we have assumed that the data stream $X_{(1)}, X_{(2)}, \ldots$ is composed of independent observations, which also led to the independence of $S(X_{(1)}), S(X_{(2)}), \ldots$ by construction. However, in practice, they can be dependent. In this thesis, the $X_{(i)}$ in the data streams considered will always be independent, but we may use different e-variables $S^{(i)}$ that could be dependent. We directly define a stochastic process $S^{(1)}, S^{(2)}, \ldots$, with $S^{(i)}$ a nonnegative statistic of $X_{(1)}, \ldots, X_{(i)}$, satisfying

$$\text{for all } P \in \mathcal{H}_0 : \qquad \sup_{N \in \mathcal{T}} \mathbb{E}_P\left[S^{(N)}\right] \leq 1, \tag{1.2.5}$$

where $\mathcal{T}$ is the set of all valid stopping times. We call such a process an *e-process.*

Ramdas et al. [72] show that if a stochastic process $S^{(1)}, S^{(2)}, \ldots$ is an e-process, then for all $P \in \mathcal{H}_0$, there is a test martingale $M_P^{(1)}, M_P^{(2)}, \ldots$ such that $S^{(i)} \leq M_P^{(i)}$ holds for every $i > 0$. Therefore, an e-process ensures that the Type-I error remains bounded, which is derived by Ville's inequality again, as in (1.2.4).

Let us run a simple simulation comparing an e-process with the traditional p-value method, as illustrated in Example 2. The results of this simulation are shown in Figure 1.2. In this example, the p-value method fails to maintain the Type-I error

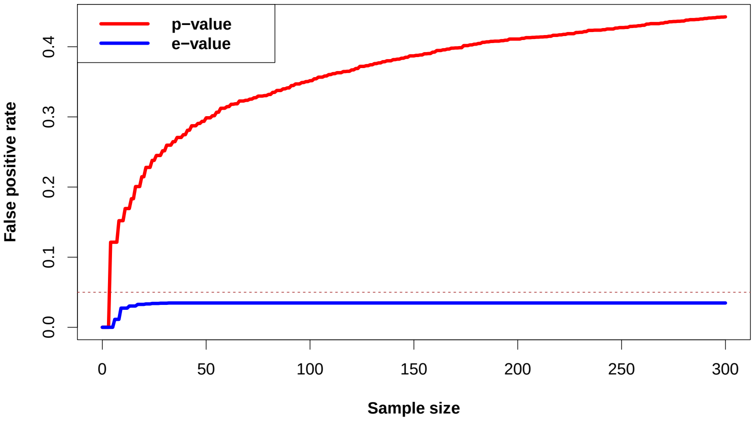


Figure 1.2: A toy simulation for Example 2 is conducted. Similar simulations are also presented by Ly et al. [65] and Turner et al. [88]. In this setup, samples are drawn from a Bernoulli(1/2) distribution. We define $S_{(i)} := \frac{p_{0.9}(X_{(i)})}{p_{0.5}(X_{(i)})}$, where $p_{0.9}(X_{(i)})$ denotes the probability mass function (pmf) of a Bernoulli(0.9) distribution. $S_{(i)}$ is an e-variable because it can easily be checked that it complies with (1.2.1). We obtain a new observation at each time step, then calculate the p-value $p(t(X^{(t)}))$ and e-value $S^{(t)} := \prod_{i=1}^{t} S_{(i)}$ at time t using the first t samples observed. This process continues until all samples are examined. If, at any time $t \in \{1, 2, \ldots, 300\}$, $p(t(X^{(t)})) \leq 0.05$, we reject $\mathcal{H}_0$; similarly, we reject $\mathcal{H}_0$ based on the e-value if there exists a time $t' \in \{1, 2, \ldots, 300\}$ such that $S^{(t')} \geq 20$. This simulation is repeated 3000 times, and we compute the rejection rates for both the p-value and e-value. This simulation shows that the Type-I error of sequentially testing $S^{(t)} \geq \frac{1}{0.05}$ remains forever below 0.05, whereas the Type-I error of sequentially using the $p(t(X^{(t)})) \leq 0.05$ violates the level 0.05.

guarantee, while the e-process method successfully controls the Type-I error rate as expected. This demonstrates the robustness of the e-process approach in contrast to the p-value method, which can be prone to inflation of Type-I error in certain scenarios.

In this thesis, we rely on various pieces of preliminary knowledge, though they may not be explicitly mentioned in each corresponding chapter. Therefore, we have included essential preliminary knowledge in the following subsections.

1.3 Preliminary knowledge: RIPr and e-power

RIPr Suppose we have a distribution Q and a set of distributions $\mathcal{P}$. The goal is to find the distribution in $\mathcal{P}$ that is 'closest' to Q. One common way to measure the divergence between two distributions is by using Kullback-Leibler (KL) Divergence, denoted as $D(Q\|P)$, which is defined as:

$$D(Q\|P) := \mathbb{E}_{X\sim Q}\left[\log\frac{q(X)}{p(X)}\right],$$

where $q(X)$ and $p(X)$ are the probability densities of Q and P, respectively.

We define $\mathcal{W}(\mathcal{P})$ to be the Choquet convex hull of $\mathcal{P}$. This means that $\mathcal{W}(\mathcal{P})$ is a convex set, and distribution $P_W \in \mathcal{W}(\mathcal{P})$, if and only of there is a proper prior W on $\mathcal{P}$ such that:

$$p_W(X) = \int p(X)dW(p).$$

It is clear that $\mathcal{P} \subseteq \mathcal{W}(\mathcal{P})$ because, for any $P \in \mathcal{P}$, we can place all the prior mass on P. We define P^*, the *Reverse Information Projection* (RIPr) of Q onto $\mathcal{P}$ [60], [27], as the distribution in $\mathcal{W}(\mathcal{P})$ that minimizes the KL-divergence from Q to $\mathcal{W}(\mathcal{P})$. This means that, if the minimum in $\mathcal{W}(\mathcal{P})$ can be attained, P^* is the distribution that is 'closest' to Q in the KL-divergence sense:

$$P^* = \arg\min_{P\in\mathcal{W}(\mathcal{P})} D(Q\|P). \tag{1.3.1}$$

In our simplified introductory statement here, the RIPr is undefined if there is no distribution in $\mathcal{W}(\mathcal{P})$ that minimizes the KL divergence. However, the RIPr can be defined for such cases as well - see [42], [43] and [58].

e-power In traditional hypothesis testing with a p-value, we use the term 'power' to describe the probability that a test rejects the null hypothesis $\mathcal{H}_0$ when the alternative hypothesis $\mathcal{H}_1$ is true. Similarly, in the context of e-values, we aim to measure a test's effectiveness through a concept called 'e-power'. The *e-power* of an e-variable $S = S(X)$ based on a data X and alternative $\mathcal{H}_1 = \{Q\}$ is defined as the expected logarithm of

S under the alternative distribution Q:

$$\mathbb{E}_{X\sim Q}[\log S(X)] := \int \log S(x)\, dQ(x),$$

as introduced by [42] and [92].

Let $\mathcal{H}_1 = \{Q\}$ and $\mathcal{H}_0 = \mathcal{P}$. Grünwald et al. [42] prove that $S_{\text{RIP}} = \frac{q(X)}{p^*(X)}$ is an e-variable, where P^* is specified as in (1.3.1). They also demonstrate that among all e-variables for $\mathcal{H}_0$, the RIPr e-variable S_{RIP} yields the highest e-power relative to the alternative $\mathcal{H}_1$ and the null $\mathcal{H}_0$.

In p-value testing, rejecting $\mathcal{H}_0$ requires a small p-value, but in the e-value framework, $\mathcal{H}_0$ is rejected when $S > 1/\alpha$. Therefore, if $\mathcal{H}_1$ is true, we want S to be as large as possible. The e-power captures this by measuring the strength of S in providing evidence against $\mathcal{H}_0$.

1.4 Preliminary knowledge: Exponential family

The probability density function (pdf) of the exponential distribution is well-known and can be expressed as:

$$p_\lambda(x) = \lambda e^{\lambda x}, x \in [0, \infty), \lambda \in (0, \infty), \tag{1.4.1}$$

where x represents the data, and λ is the rate parameter. Interestingly, many other families, such as the Gaussian, Poisson and Beta distributions, can be written in a similar form to the exponential distribution. These types of distributions are part of a broader class known as *exponential families*. We will now explain this concept in more detail.

Definition of exponential family Consider a set of probability distributions $\mathcal{P} \in \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ for data U, where $\boldsymbol{\Theta}$ is the parameter space of $\mathcal{P}$. If there exists a re-parametrization $\mathcal{P} = \{P_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \mathtt{B}\}$ with $\mathtt{B} \subset \mathbb{R}^d$ for some $d \in \mathbb{N}^+$, and a random vector X that is a function of U and the probability density (or mass) functions can be written as:

$$p_{\boldsymbol{\beta}}(U) = \frac{1}{Z(\boldsymbol{\beta})} \exp\left(\boldsymbol{\beta}^\top X\right), \tag{1.4.2}$$

where $Z(\boldsymbol{\beta})$ is a normalizing factor, then $\mathcal{P}$ is called an exponential family. We call X the *sufficient statistic* for $\boldsymbol{\beta}$, which can be verified easily using the *Fisher–Neyman factorization theorem*. $\boldsymbol{\beta}$ is the *natural* or *canonical parameter* of the distribution. (Note that in Chapter 4, we use λ for the canonical parameter, whereas in Chapter 5, we use θ.) When the functions X and $\boldsymbol{\beta} := \boldsymbol{\beta}(\boldsymbol{\theta})$ are fixed, they define a specific exponential family. Some people prefer to write the exponential family form as:

$$p_{\boldsymbol{\beta}}(U) = \frac{1}{Z(\boldsymbol{\beta})} \exp\left(\boldsymbol{\beta}^\top X\right) h(U),$$

where $h(U)$ is called the *carrier function*. However, this is essentially the same as the previous form since it can be rewritten as:

$$p_{\boldsymbol{\beta}}(U) = \frac{1}{Z(\boldsymbol{\beta})} \exp\left(\boldsymbol{\beta}'^{\top} X'\right),$$

where $\boldsymbol{\beta}' = (\boldsymbol{\beta}^{\top}, 1)^{\top}$ and $X' = (X^{\top}, \log h(U))^{\top}$.

This re-parametrization highlights how various familiar distributions can be unified under the exponential family framework. We explain it using the Gaussian example.

Gaussians are an exponential family For simplicity, we only show the one-dimensional case here. In this example, the standard parameterization would be $\boldsymbol{\theta} = (\mu, \sigma^2)$ and $\boldsymbol{\Theta} = \mathbb{R} \times \mathbb{R}^{+}$. The pdf of a Gaussian can be written as

$$\begin{aligned} p_{\mu,\sigma^2}(u) =& \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(u-\mu)^2\right) \\ =& \exp\left(-\frac{1}{2\sigma^2}u^2 + \frac{\mu}{\sigma^2}u - \frac{1}{2\sigma^2}\mu^2 - \log\sigma - \frac{1}{2}\log(2\pi)\right) \\ =& \exp\left(\boldsymbol{\beta}^{\top}x - \log Z(\boldsymbol{\beta})\right), \end{aligned}$$

where the last equality holds if we let $\boldsymbol{\beta} = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})^{\top}$, $x = (u, u^2)^{\top}$ and $\log Z(\boldsymbol{\beta}) = \log \int \exp\left(\boldsymbol{\beta}^{\top}x\right) dx = \frac{1}{2\sigma^2}\mu^2 + \log\sigma + \frac{1}{2}\log(2\pi)$.

In this section, we introduced the basic concept of the exponential family. However, throughout this dissertation, more advanced knowledge about the exponential family is frequently used, though not explicitly explained in the corresponding chapters. We have placed these more technical details in Appendix 1.A.

1.5 Outline

We provide a brief overview of the content of each chapter in the following subsections. Each chapter reports our research on various e-variables for exponential family null hypothesis and/or alternative hypothesis, addressing different problems. Additionally, Chapter 5 introduces a novel concentration inequality for multivariate exponential families.

Chapter 2: Conditions for the existence of simple e-variable

In many academic fields, such as social sciences, biology, and physics, researchers often aim to infer the underlying distribution of data. To do this, they may assume a specific model structure. For example, they might hypothesize that the data follow a binomial distribution, then estimate the parameters that best fit the data. However, the data may come from a different set of distributions, such as a set of negative binomial distribution. In mathematics, a correct assumption is referred to as a well-specified model assumption. Therefore, determining whether a model is well-specified is crucial for accurately learning the structure of the data. This involves testing whether the observed data are actually distributed according to an element of the given set of distributions.

Chapter 2 addresses this task using e-variables for exponential families. Specifically, we may want to test if a certain parameter in an exponential family is zero, or not—this includes linear regression testing as a particular case. In this chapter, we study the GRO e-variable for this task, for a single outcome U. As was announced in Section 1.3, the GRO e-variable is closely connected with the RIPr. The simplest example of GRO e-variables is the likelihood ratio between simple alternative and simple null hypotheses. However, for composite hypotheses, the situation becomes more complex. Nevertheless, GRO e-variables in the form of a likelihood ratio involving a single, specific element of the composite null hypothesis can sometimes still be found. We refer to such GRO e-variables as 'simple' e-variables. As we will demonstrate, their existence is closely linked to properties of the aforementioned RIPr.

When simple e-variables exist, they can be easily computed and are known to be optimal in terms of e-power [53, 42]. In the context of repeated experiments with a fixed stopping rule for data collection, and a simple alternative, using a simple e-variable will, asymptotically, provide the strongest evidence against the null hypothesis compared to other e-variables. Therefore, it is important to determine when simple e-variables exist in specific contexts. Chapter 2 offers a set of equivalent conditions under which simple e-variables exist for exponential family null hypotheses.

Chapter 3: General exponential family test

Chapter 3 continues from Chapter 2. In this chapter, we explore the scenario where a condition 'opposite' to the previous conditions applies, which we refer to as the *anti-simple case*, meaning that simple e-variables do not exist. For both cases—whether simple e-variables exist or not—we analyze common types of e-variables and e-processes related to composite exponential family nulls, but now for sequences of outcomes rather than a single one: we examine and compare their *e-power* [94] for i.i.d. data $U_{(1)}, U_{(2)}, \ldots$. Recall that e-power plays a pivotal role, as it is maximized by the optimal e-variable (i.e. GRO e-variable) across all e-variables defined on $U^{(n)}$. As we announced in Section 1.3, it can be determined using the *reverse information projection (RIPr)*. We denote this optimal e-variable as S_{RIP}. Additionally, we consider a sequentialized version of the RIPr e-variable, $S_{\text{SEQ-RIP}}$, which is optimal at the individual outcome level but not necessarily over the entire sample. We also investigate a *conditional* e-variable, S_{COND}, based on conditioning on the sufficient statistic, along with a well-known version of the *universal inference* e-variable, S_{UI} [95].

Instantiating such e-variables requires specifying an alternative hypothesis. We begin by considering a simple alternative $\mathcal{Q} = \{Q\}$. Our results demonstrate that the RIPr prior W that achieves the minimum in (3.1.2) is approximately Gaussian with variance $O(1/n)$ in an asymptotic sense, and exactly if $\mathcal{H}_0$ is a Gaussian location family and Q is also Gaussian. To our knowledge, this is the first time that insights into a nondegenerate RIPr prior have been obtained for the case of a parametric, non-convex null.

This result is made possible by our key theoretical insight: the conditional e-variable S_{COND} can be analyzed using a local central limit theorem with explicit bounds on the error terms [16]. Consequently, we derive not only explicit $o(1)$ bounds on its e-power but also establish that S_{COND} is closely related to S_{RIP} (in the Gaussian anti-simple case, they even coincide). We extend these results to other types of e-variables, not only under the 'true' alternative Q but also in the *misspecified case*, where the data are sampled i.i.d. from a distribution $R \neq Q$.

We employ two standard methods to design e-variables for composite alternatives $\mathcal{Q}$: the sequential plug-in method [38] and the method of mixtures [71]. We observe that, when using the method of mixtures and equipping the alternative with a prior W_1, under regularity conditions, the RIPr prior W in (3.1.2) is, approximately, the *same* prior W_1, regardless of whether we are in the simple case or not. Summarizing some of our main findings for the composite case, we derive the following relationships. Under appropriate (though mild) regularity conditions on $\mathcal{H}_0$ and $\mathcal{H}_1$, for all $Q \in \mathcal{H}_1$, (in the chapter we use $\mathcal{P}$ and $\mathcal{Q}$ because it fits better with other notations.) we have:

$$\begin{aligned}
\mathbb{E}_Q[\log S^{(n)}_{\text{RIP}}/S^{(n)}_{\text{COND}}] &= o(1).\\
\mathbb{E}_Q[\log S^{(n)}_{\text{COND}}/S^{(n)}_{\text{UI}}] &= \frac{d}{2}\log n + O(1).\\
\mathbb{E}_Q[\log S^{(n)}_{\text{SEQ-RIP}}/S^{(n)}_{\text{UI}}] &= \frac{d_{qp}}{2}\log n + O(1)\\
&\quad \text{with } 0 < d_{qp} < d, \text{ in the strict simple case.}\\
\mathbb{E}_Q[\log S^{(n)}_{\text{SEQ-RIP}}/S^{(n)}_{\text{COND}}] &\leq -n\epsilon\\
&\quad \text{for some } \epsilon > 0, \text{ all large } n, \text{ in the strict anti-simple case.}
\end{aligned}$$

Here, d represents the dimensionality of the exponential family, and d_{qp} is a measure of 'effective dimension', whose exact value depends on Q. Of course, we provide precise definitions of "strict simple, anti-simple" in Chapter 3.

Chapter 4: k-sample tests with exponential families

A k-sample test is the general version of a two-sample test. It involves analyzing data from k independent random samples, each drawn from a possibly different population. For example when $k = 2$, in studying the effectiveness of a new treatment (such as a new blood pressure medication), patients are divided into two groups: a treatment group and a placebo group. Researchers track the number of recoveries in each group after a set treatment period. If the treatment is effective, a higher recovery rate is expected in the treatment group compared to the placebo group. The statistical test is used to determine if the observed difference in recovery rates between the two groups is significant, or if it could have occurred by chance. Two-sample tests are designed to model such situations. Mathematically, the objective is to determine whether the observed difference between the two populations is statistically significant, meaning whether the difference is likely due to chance or represents a true difference between the populations.

Chapter 4 centers on k-sample tests. Some of the results presented in the chapter are special cases of those from Chapter 3. However, it remains valuable to include them in this chapter because we provide full details that were not covered in Chapter 3. We develop four (pseudo-) e-variables for k-sample tests in exponential families: $S_{\text{RIP}}, S_{\text{COND}}, S_{\text{MIX}}$, and S_{PSEUDO}. $S_{\text{RIP}}, S_{\text{COND}}, S_{\text{MIX}}$ are real e-variables, while S_{PSEUDO} is an e-variable only when it coincides with S_{RIP}, which happens whenever the latter is computationally straightforward; in other instances, it is not a true e-variable but remains useful for our theoretical analysis. Suppose the (shortest) ℓ_2-distance between the k-dimensional parameter of the alternative and the null parameter space is denoted by δ. Our results show that, for any two of the aforementioned e-variables S and S', the difference in e-power is given by $\mathbb{E}[\log S - \log S'] = O(\delta^4)$.

Chapter 5: GROW e-variables and a novel concentration inequality

The link between optimal rejection regions for anytime-valid tests at a fixed level α and optimal anytime-valid concentration inequalities is well-documented [47]. Chapter 5 explores a variation of this connection, focusing on a simple multivariate null hypothesis and a range of composite alternatives. We examine both absolute and relative *GROW* ('growth-rate optimal in the worst-case') e-variables as introduced by Grünwald et al. [42]. Further, we illustrate how these e-variables connect to a concentration inequality, which we refer to as the *Csiszár-Sanov-Chernoff* (CSC) inequality.

To start, we analyze the GROW e-variable S_{GROW} within this framework, considering cases where $\mathcal{H}_1$ is either the set $\mathcal{P}_1$ of all distributions with means in a specified *convex* set $\mathtt{M}_1$, the set $\mathcal{E}_1$ of all distributions in the exponential family generated by P_0 with means in $\mathtt{M}_1$, or any $\mathcal{H}_1$ for which $\mathcal{E}_1 \subset \mathcal{H}_1 \subset \mathcal{P}_1$. Remarkably, the GROW e-variables coincide across all such $\mathcal{H}_1$. We derive this result by applying the well-known Csiszár-Topsøe Pythagorean theorem for relative entropy, which leads us to the fundamental CSC concentration inequality. This section's focus is primarily on rephrasing established findings, familiar to the information-theoretic community but perhaps less so to those working with e-values.

Chapter 5 then introduces a novel approach, examining cases where the *complement* of $\mathtt{M}_1$ forms a connected, bounded set containing P_0 — a scenario more commonly encountered in practical applications and more aligned with the multivariate central limit theorem (CLT). This configuration, which we call the *surrounding* $\mathcal{H}_1$ case because P_0 is "surrounded" by $\mathcal{H}_1$, has rarely been considered in the derivation of CSC bounds, with an exception being the variation studied by Kaufmann and Koolen [51].

We extend the previous S_{GROW} e-variable to this surrounding $\mathcal{H}_1$ case in two ways. The first approach is a straightforward *absolute* extension of the GROW e-variable to the multivariate case, still denoted as S_{GROW}. Or we can determine a *relatively* optimal GROW e-variable S_{REL} that is as close as possible to the largest S_{GROW} among all e-variables S_{GROW} that can be defined on convex subsets of $\mathcal{H}_1$, where we define relative optimality in a minimax-regret sense. We characterize S_{GROW} for the univariate case ($d = 1$) while leaving the multidimensional case ($d > 1$) as an open problem, and we fully characterize S_{REL} for general dimensions. We then show that S_{REL} leads again to a CSC bound — and this CSC bound is new.

Appendix 1.A More exponential family preliminaries

Mean parameterization Since X is the sufficient statistic for parameter, it is often more useful to directly study X rather than U, as we often do in this dissertation. From equation (1.4.2), we know that X has the same dimension as the canonical parameter $\boldsymbol{\beta}$. We already know that when the functions $X(U)$ and $\boldsymbol{\beta}(\boldsymbol{\theta})$ are fixed, we may define a specific class of exponential family distributions, denoted by $\mathcal{P}$ (e.g., the Gaussian family). A natural approach is to represent $P_{\boldsymbol{\theta}}$ using the expectation of X, defined as $\boldsymbol{\mu}(\boldsymbol{\theta}) := \mathbb{E}_{U\sim P_{\boldsymbol{\theta}}}[X]$, which is called the *mean parameterization*, and $\boldsymbol{\mu}(\boldsymbol{\theta})$ is the *mean-value parameter*. It can be shown that for $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$, ($\boldsymbol{\Theta}$ is the standard parameter space of the distribution family $\mathcal{P}$), if $P_{\boldsymbol{\theta}} \neq P'_{\boldsymbol{\theta}}$, then $\boldsymbol{\mu}(\boldsymbol{\theta}) \neq \boldsymbol{\mu}(\boldsymbol{\theta}')$, ensuring a distinct probability model for each parameter.

Canonical parameterization For simplicity, we use the notation $\boldsymbol{\beta} := \boldsymbol{\beta}(\boldsymbol{\theta}), \boldsymbol{\mu} := \boldsymbol{\mu}(\boldsymbol{\theta})$ going forward. Since $Z(\boldsymbol{\beta})$ is the normalizing factor, we have:

$$Z(\boldsymbol{\beta}) = \int \exp\left(\boldsymbol{\beta}^\top x\right) dx,$$

where the integral becomes a sum in the discrete case. Then taking the first derivative of $\log Z(\boldsymbol{\beta})$ w.r.t. $\boldsymbol{\beta}$ gives us the mean of X under $P_{\boldsymbol{\beta}}$:

$$\frac{d \log Z(\boldsymbol{\beta})}{d\boldsymbol{\beta}} = \frac{\int x \exp\left(\boldsymbol{\beta}^\top x\right) dx}{\int \exp\left(\boldsymbol{\beta}^\top x\right) dx} = \int x \cdot \frac{1}{Z(\boldsymbol{\beta})} \exp\left(\boldsymbol{\beta}^\top x\right) dx = \mathbb{E}_{P_{\boldsymbol{\beta}}}[X] := \boldsymbol{\mu}(\boldsymbol{\beta}). \tag{1.A.1}$$

We call $\boldsymbol{\mu}(\boldsymbol{\beta})$ the mean-value parameter corresponding to $\boldsymbol{\beta}$. We continue to take the second derivative of $\log Z(\boldsymbol{\beta})$ w.r.t. $\boldsymbol{\beta}$. This gives us the covariance matrix of X under $P_{\boldsymbol{\beta}}$:

$$\Sigma_P = \frac{d^2 \log Z(\boldsymbol{\beta})}{d\boldsymbol{\beta}^2} = \frac{d\boldsymbol{\mu}(\boldsymbol{\beta})}{d\boldsymbol{\beta}}.$$

Since Σ_P is positive definite, the transformation from $\boldsymbol{\beta}$ to $\boldsymbol{\mu}$ is one-to-one, ensuring that $\mathcal{P}$ can be uniquely represented using the canonical parameter $\boldsymbol{\beta}$.

Empirical mean as MLE In the mean parameter space, the *maximum likelihood estimator (MLE)* for the sufficient statistic X generated from a data set $\mathcal{U}$ has an important property, frequently used in the following chapters. Consider a set of i.i.d. data points $x_{(1)}, x_{(2)}, \ldots, x_{(n)} := x^{(n)}$. The likelihood function is given by:

$$\mathcal{L}(\boldsymbol{\beta} \mid x^{(n)}) = \frac{1}{Z(\boldsymbol{\beta})^n} \exp\left(\boldsymbol{\beta}^\top \sum_{i=1}^{n} x_{(i)}\right).$$

To find the MLE, we take the derivative of the log-likelihood and set it to zero:

$$\frac{d\log\mathcal{L}(\boldsymbol{\beta}\mid x^{(n)})}{d\boldsymbol{\beta}} = -n\frac{d\log Z(\boldsymbol{\beta})}{d\boldsymbol{\beta}} + \sum_{i=1}^{n} x_{(i)} \overset{(a)}{=} -n\boldsymbol{\mu} + \sum_{i=1}^{n} x_{(i)} = 0. \tag{1.A.2}$$

where (a) follows from (1.A.1). If $\frac{1}{n}\sum_{i=1}^{n} x_{(i)}$ is in the interior of the the mean-value parameter space, then (1.A.2) has solution. This shows that in such cases, the MLE for the mean parameter is $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} x_{(i)}$, i.e., the empirical mean.

Robustness properties These properties are shown in detail in [35, Chapter 19]. We briefly introduce some of them that are used frequently in the following chapters. Let $\mathcal{P}$ be a regular exponential family, take $X = U$ and let $\mathtt{M}$ be its mean parameter space (See [13] or [19] for the definition of regular). Consider $P_{\boldsymbol{\mu}} \in \mathcal{P}$ such that $\mathbb{E}_{P_{\boldsymbol{\mu}}}[X] = \boldsymbol{\mu} \in \mathtt{M}$. Let Q be an arbitrary distribution with $\mathbb{E}_Q[X] = \boldsymbol{\mu}^* \in \mathtt{M}$. For all $P_{\boldsymbol{\mu}} \in \mathcal{P}$, we have:

$$\mathbb{E}_Q\left[\log\frac{p_{\boldsymbol{\mu}^*}(X)}{p_{\boldsymbol{\mu}}(X)}\right] = \mathbb{E}_{P_{\boldsymbol{\mu}^*}}\left[\log\frac{p_{\boldsymbol{\mu}^*}(X)}{p_{\boldsymbol{\mu}}(X)}\right] := D(P_{\boldsymbol{\mu}^*}||P_{\boldsymbol{\mu}}),$$

where $D(P_{\boldsymbol{\mu}^*}||P_{\boldsymbol{\mu}})$ is the KL-divergence between $P_{\boldsymbol{\mu}^*}$ and $P_{\boldsymbol{\mu}}$. This statement holds in the canonical parameter space ($\mathtt{B}$) of $\mathcal{P}$ as well. For all $P_{\boldsymbol{\beta}} \in \mathcal{P}$ with $\boldsymbol{\beta} \in \mathtt{B}$, we have:

$$\mathbb{E}_Q\left[\log\frac{p_{\boldsymbol{\beta}^*}(X)}{p_{\boldsymbol{\beta}}(X)}\right] = \mathbb{E}_{P_{\boldsymbol{\beta}^*}}\left[\log\frac{p_{\boldsymbol{\beta}^*}(X)}{p_{\boldsymbol{\beta}}(X)}\right] := D(P_{\boldsymbol{\beta}^*}||P_{\boldsymbol{\beta}}),$$

which is equivalent to the statement in $\mathtt{M}$ because $P_{\boldsymbol{\beta}}$ and $P_{\boldsymbol{\mu}}$ with $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta})$ represent the same distribution.

Moreover, $\mathbb{E}_{P_{\boldsymbol{\beta}^*}}[-\log p_{\boldsymbol{\beta}}(X)]$ is a strictly convex function with respect to $\boldsymbol{\beta}$, achieving its unique minimum at $\boldsymbol{\beta}^*$. Since $P_{\boldsymbol{\beta}}$ and $P_{\boldsymbol{\mu}}$ represent the same distribution, $\mathbb{E}_{P_{\boldsymbol{\mu}^*}}[-\log p_{\boldsymbol{\mu}}(X)]$ is also a function of $\boldsymbol{\mu}$, achieving its unique minimum at $\boldsymbol{\mu}^*$.

Chapter 2

Optimal E-Values for Exponential Families: the Simple Case

Abstract

We provide a general condition under which e-variables in the form of a simple-vs.-simple likelihood ratio exist when the null hypothesis is a composite, multivariate exponential family. Such ‘simple’ e-variables are easy to compute and expected-log-optimal with respect to any stopping time. Simple e-variables were previously only known to exist in quite specific settings, but we offer a unifying theorem on their existence for testing exponential families. We start with a simple alternative Q and a regular exponential family null. Together these induce a second exponential family $\mathcal{Q}$ containing Q, with the same sufficient statistic as the null. Our theorem shows that simple e-variables exist whenever the covariance matrices of $\mathcal{Q}$ and the null are in a certain relation. A prime example in which this relation holds is testing whether a parameter in a linear regression is 0. Other examples include some k-sample tests, Gaussian location- and scale tests, and tests for more general classes of natural exponential families. While in all these examples, the implicit composite alternative is also an exponential family, in general this is not required.

2.1 Introduction

The work of Andrew Barron has been enormously influential in the development of *e-variables*, an alternative to the p-value that is suitable for designing hypothesis

[0] This chapter is based on Peter Grünwald, Tyron Lardy, Yunda Hao, Shaul K Bar-Lev, and Martijn Jong. Optimal E-Values for Exponential Families: the Simple Case. arXiv preprint arXiv:2404.19465, 2024. It has been submitted as a contribution to the *Festschrift for Andrew Barron on the Occasion of his 65th Birthday.*

tests and confidence intervals with a flexible design, i.e. when sample sizes are not pre-specified, or when the decision to conduct new experiments may depend on past data — [71] provides a recent overview of this exciting new area of statistical research. While Andrew himself has never published on e-variables, his seminal work on *Reverse Information Projection (RIPr)* [61] (see also the work by his Ph.D. students [60] and [18]) is the cornerstone of the math underlying [42], which connects optimal e-variables to a RIPR onto the *null* hypothesis. On the other hand, there is Andrew's fundamental work on *universal coding and modeling* (there are far too many papers to cite here — the earliest one may be [22] while the most recent one is [48]). This work is directly connected to finding optimal e-variables for the *alternative* hypothesis, as explained by [42, 71]. Originally, much of it was done in the context of the Minimum Description Length *(MDL)* Principle [14]. For one of us (Grünwald) MDL was the major research topic until around 2010 [35], while e-values have become his main topic since 2015 — as such the influence of Barron's work on Grünwald's work can hardly be overstated, and he would like to express his debt and gratitude.

In this paper we bring e-variables and in particular the RIPr together with another one of Andrew Barron's central research interests: exponential families [15, 81]. An important task is to test whether these models are well-specified, that is, whether observed data are indeed distributed by an element of an exponential family; or more specifically whether a specific parameter in an exponential fmaily is 0 or not — the latter including linear regression testing as a special case. Many classic tests are well-suited for this purpose [3, 63, 79]. However, the vast majority of these methods are based on p-values, and thus designed for fixed sample size experiments. Here, we are instead interested in hypothesis tests that are based on e-values [42], which is the value taken by an e-variable. The most straightforward example of e-variables are likelihood ratios between simple alternatives and simple null hypotheses. E-variables for composite hypotheses, and in particular 'good' e-variables, are generally more complicated. However, e-variables in the form of a likelihood ratio with a single, special element of the null representing the full, composite null sometimes still exist. We refer to such e-variables as 'simple' e-variables. As we shall see below, their existence is intimately tied to properties of the aforementioned *RIPr*, connecting our work strongly to Barron's.

Simple e-variables, if they exist, can easily be computed, and are known to be optimal in an expected-log-optimality sense [53, 42]. That is, if we combine evidence from a repeated experiment where data is collected using a fixed stopping rule, then using the simple e-variable will asymptotically result in the most evidence against the null, among all e-variables; details can be found in Section 2.1.4. As such, it is desirable to find out whether or not simple e-variables exist in specific settings. The main result of this paper, Theorem 1, provides a set of equivalent conditions under which simple e-variables exist for exponential family nulls.

2.1.1 Main Result and Overview

Here we briefly describe Theorem 1, assuming prior knowledge on e-variables and exponential families, and we provide an overview of the rest of the paper — all relevant definitions and explanations are given in Section 2.1.2–2.1.4. We fix a regular multivariate exponential family null $\mathcal{P}$ for data U with some sufficient statistic vector $X = t(U)$ and a distribution Q for U, outside of $\mathcal{P}$, and with density q. As our most important regularity condition, we assume that Q has a moment generating function and that there exists $P_{\boldsymbol{\mu}^*} \in \mathcal{P}$ with the same mean of X, say $\boldsymbol{\mu}^*$, as Q. It is known that $P_{\boldsymbol{\mu}^*}$ is the *Reverse Information Projection (RIPr)* of Q onto $\mathcal{P}$ [60], that is, it achieves $\min_{P \in \mathcal{P}} D(Q\|P)$. Denoting the density of $P_{\boldsymbol{\mu}^*}$ by $p_{\boldsymbol{\mu}^*}$, it follows by Theorem 1 of [42] that $q(U)/p_{\boldsymbol{\mu}^*}(U)$ would be an e-variable in case $\inf_{P \in \text{CONV}(\mathcal{P})} D(Q\|P) = \min_{P \in \mathcal{P}} D(Q\|P)$. Our theorem establishes a sufficient condition for when this is actually the case. It is based on constructing a second exponential family $\mathcal{Q}$ with densities proportional to $\exp(\boldsymbol{\beta}^T t(U))q(U)$ for varying $\boldsymbol{\beta}$: $\mathcal{Q}$ contains Q and has the same sufficient statistic as $\mathcal{P}$. In some cases, but not all, $\mathcal{Q}$ may be thought of as the composite alternative we are interested in. Letting $\Sigma_p(\boldsymbol{\mu})$ and $\Sigma_q(\boldsymbol{\mu})$ denote the covariance matrices of the $P_{\boldsymbol{\mu}} \in \mathcal{P}$ and $Q_{\boldsymbol{\mu}} \in \mathcal{Q}$ with mean $\boldsymbol{\mu}$, Theorem 1 below implies the following: under a further regularity condition on the parameter spaces of $\mathcal{P}$ and $\mathcal{Q}$, simple e-variables exist whenever $\Sigma_p(\boldsymbol{\mu}) - \Sigma_q(\boldsymbol{\mu})$ is positive semidefinite for all $\boldsymbol{\mu}$ in the mean-value parameter space of $\mathcal{Q}$ (additionally, three equivalent conditions will be given). If this happens, then we may further conclude that for *every* element $Q_{\boldsymbol{\mu}'}$ of the constructed $\mathcal{Q}$, the likelihood ratio $q_{\boldsymbol{\mu}'}(U)/p_{\boldsymbol{\mu}'}(U)$ is an e-variable, where $P_{\boldsymbol{\mu}}$ is the element of $\mathcal{P}$ to which $Q_{\boldsymbol{\mu}}$ is projected. An example pair $(Q, \mathcal{P})$ to which the theorem applies is when, under Q, $U \sim N(m, s^2)$ for fixed m, s^2 and $\mathcal{P} = \{N(0, \sigma^2) : \sigma^2 > 0\}$ is the univariate (scale) family of normal distributions. This situation is illustrated in Figure 2.1 and is treated in detail in Section 2.4.3, and extended to linear regression testing – arguably our most important application – in Section 2.4.4.

We stress that, while our approach starts with a simple alternative Q, the results are still applicable if one is interested in a composite alternative $\mathcal{H}_1$. To this end, take any $Q \in \mathcal{H}_1$ and use our main result to determine whether a simple e-variable with respect to Q exists. If one exists for every Q, an e-variable for the full alternative can easily be constructed either by the method of mixtures or the prequential (sequential plug-in learning) method [71].

Things conceptually simplify in this composite alternative case if $\mathcal{H}_1$ can be parameterized as $\mathcal{H}_1 = \{\mathcal{Q}^{(\theta)} : \theta \in \Theta\}$ in such a way that for each $Q \in \mathcal{H}_1$, the associated family $\mathcal{Q}$ constructed from $\mathcal{P}$ and Q is equal to $\mathcal{Q}^{(\theta)}$ for some θ. As is suggested by Figure 2.1, this happens, for example, in the Gaussian scale example of Section 2.4.3, if we consider as alternative $\mathcal{H}_1$ the full Gaussian family. We can start with any $Q = N(m, s^2)$ and generate $\mathcal{Q}$ which then coincides with some $\mathcal{Q}^{(\theta)}$, corresponding to a specific sloped line in the figure. Together, all these sloped lines span $\mathcal{H}_1$. In fact, it turns out that a natural choice of $\mathcal{H}_1$ that partitions into $\mathcal{Q}^{(\theta)}$ is possible in *all* our examples, and that this $\mathcal{H}_1$ is itself an exponential family in all these examples. Nevertheless, we stress that in general our method does not in any way require $\mathcal{H}_1$ to be an exponential family — only $\mathcal{P}$ is required to be so.

A specific interpretation of the result is obtained when restricting to the 1-dimensional case. The best squared-error predictor of X sampled according to Q has Q-expected squared error prediction equal to $\text{VAR}_Q[X] = \mathbf{E}_Q(X - \boldsymbol{\mu}^*)^2$. If X is really sampled from Q but we think it comes from $P_{\boldsymbol{\mu}^*}$ and want to make the best $P_{\boldsymbol{\mu}^*}$-expected

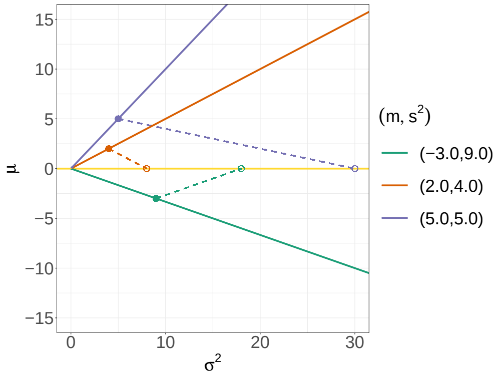


Figure 2.1: The family $\mathcal{Q}$ for various (m, s^2). The coordinate grid represents the parameters of the full Gaussian family, the horizontal line shows the parameter space of $\mathcal{P}$, the sloped lines show the parameters of the distributions in $\mathcal{Q}$, and the dashed lines show the projection of (m, s^2) onto the parameter space of $\mathcal{P}$. For example, we may start out with Q expressing $U \sim N(m, s^2)$ with $m = -3.0, s^2 = 9.0$, represented as the green dot on the green line. Its RIPr onto $\mathcal{P}$ is the green point on the yellow line. The corresponding family $\mathcal{Q}$, constructed in terms of Q and $\mathcal{P}$, is depicted by the green solid line. The theorem implies that the likelihood ratio between any point on the green line and its RIPr onto the yellow line is an e-variable; similarly for the red and blue lines.

squared error predictions, we would predict with the mean, which is still $\boldsymbol{\mu}^*$, but we assess our squared error as $\text{VAR}_{P_{\boldsymbol{\mu}^*}}[X]$ whereas the real expected squared error is still $\text{VAR}_Q[X]$. Thus, in the 1-dimensional case, in the situation that our result does *not* hold, there is a mismatch between Q and its projection $P_{\boldsymbol{\mu}^*}$ in the sense that the closest approximation we can provide to Q promises a better squared-error prediction than can be obtained with Q itself. Our result says that if the mismatch does not occur, then we cannot get closer to Q by convexifying $\mathcal{P}$.

The proof of Theorem 1 is based on convex duality properties of exponential families. In the remainder of this introductory section, we fix notation and definitions of exponential families and e-variables. In Section 2.2 we show how, based on the constructed family $\mathcal{Q}$, one can often easily construct *local* e-variables, i.e. e-variables with the null restricted to a subset of $\mathcal{P}$. Then, in Section 2.3 we present our main theorem, extending the insight to global e-variables. Section 2.4 provides several examples. This includes cases for which simple e-variables were already established, such as certain k-sample tests [88, 44] or — in an unpublished master's thesis — the linear regression model [30], as well as cases for which it was previously unknown whether simple e-variables exist, such as for a broad class of natural exponential families. Theorem 1 can thus be seen as a unification and generalization of known results on the existence of simple e-variables, leading to deeper understanding of why they sometimes exist. Section 2.5 provides the proof for Theorem 1. Finally, Section 2.6 provides a concluding discussion and points out potential future directions.

2.1.2 Formal Setting

We study general hypothesis testing problems in which the null hypothesis $\mathcal{P}$ is a regular (and hence full) d-dimensional exponential family. Here and in the sequel, we will freely use standard properties of exponential families without explicitly referring to their definitions and proofs, for which we refer to e.g. [13, 19, 33]. Each member of $\mathcal{P}$ is a distribution for a random element U, that takes values in some set $\mathcal{U}$, with a density relative to some given underlying measure ν on $\mathcal{U}$. The sufficient statistic vector is denoted by $X = (X_1, \ldots, X_d)$ with $X_j = t_j(U)$ for given measurable functions $t_1, \ldots, t_d$. We furthermore define $\mathtt{M}_p$ to be the mean-value parameter space of $\mathcal{P}$, i.e. the set of all $\boldsymbol{\mu}$ such that $\mathbb{E}_P[X] = \boldsymbol{\mu}$ for some $P \in \mathcal{P}$. For any $\boldsymbol{\mu} \in \mathtt{M}_p$, we denote by $P_{\boldsymbol{\mu}}$ the unique element of $\mathcal{P}$ with $\mathbb{E}_{P_{\boldsymbol{\mu}}}[X] = \boldsymbol{\mu}$, so that $\mathcal{P} = \{P_{\boldsymbol{\mu}} : \boldsymbol{\mu} \in \mathtt{M}_p\}$. As usual, this parameterization of $\mathcal{P}$ is referred to as its mean-value parameterization. Furthermore, we use Σ_p to denote the variance function of $\mathcal{P}$. That is, for all $\boldsymbol{\mu} \in \mathtt{M}_p$, $\Sigma_p(\boldsymbol{\mu})$ is the covariance matrix corresponding to $P_{\boldsymbol{\mu}}$.

Since $\mathcal{P}$ is an exponential family, the density of any member of $\mathcal{P}$ can be written, for each fixed $\boldsymbol{\mu}^* \in \mathtt{M}_p$, as

$$p_{\boldsymbol{\beta};\boldsymbol{\mu}^*}(u) = \frac{1}{Z_p(\boldsymbol{\beta};\boldsymbol{\mu}^*)} \exp\left(\sum_{j=1}^{d} \beta_j t_j(u)\right) \cdot p_{\boldsymbol{\mu}^*}(u), \tag{2.1.1}$$

where $Z(\boldsymbol{\beta};\boldsymbol{\mu}^*) = \int \exp(\sum \beta_j t_j(u)) p_{\boldsymbol{\mu}^*}(u) d\nu$, and $\boldsymbol{\beta} \in \mathbb{R}^d$ such that $Z_p(\boldsymbol{\beta};\boldsymbol{\mu}^*) < \infty$. Therefore, $\mathcal{P}$ can also be parameterized as $\mathcal{P} = \{P_{\boldsymbol{\beta};\boldsymbol{\mu}^*} : \boldsymbol{\beta} \in \mathtt{B}_{p,\boldsymbol{\mu}^*}\}$, where $\mathtt{B}_{p;\boldsymbol{\mu}^*} \subset \mathbb{R}^d$ denotes the canonical parameter space with respect to $\boldsymbol{\mu}^*$, i.e. the set of all $\boldsymbol{\beta}$ for which $Z_p(\boldsymbol{\beta};\boldsymbol{\mu}^*) < \infty$. We use $\boldsymbol{\beta}_p(\boldsymbol{\mu}';\boldsymbol{\mu}^*)$ to denote the $\boldsymbol{\beta} \in \mathtt{B}_{p,\boldsymbol{\mu}^*}$ such that $\mathbf{E}_{P_{\boldsymbol{\beta};\boldsymbol{\mu}^*}}[X] = \boldsymbol{\mu}'$ and set $\boldsymbol{\mu}_p(\cdot;\boldsymbol{\mu}^*) = \boldsymbol{\beta}_p^{-1}(\cdot;\boldsymbol{\mu}^*)$ to be its inverse. $\boldsymbol{\beta}_p(\cdot;\boldsymbol{\mu}^*)$ maps mean-value parameters to corresponding canonical parameters and $\boldsymbol{\mu}_p(\cdot;\boldsymbol{\mu}^*)$ vice versa. Note that $p_{\boldsymbol{\mu}^*} = p_{\mathbf{0},\boldsymbol{\mu}^*}$, and that we can see from the notation (one versus two subscripts) whether a density is given in the mean- or canonical representation, respectively.

The reason for explicitly denoting the mean $\boldsymbol{\mu}^*$ of the carrier density, which is unconventional, is that it will be convenient to simultaneously work with different canonical parameterizations, i.e. with respect to a different element of $\mathtt{M}_p$, below. These are all linearly related to one another in the sense that for each $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathtt{M}_p$, there is a fixed vector $\boldsymbol{\gamma}$ such that for all $\boldsymbol{\beta} \in \mathtt{B}_{p,\boldsymbol{\mu}_1}$ it holds that $p_{\boldsymbol{\beta};\boldsymbol{\mu}_1} = p_{\boldsymbol{\beta}+\boldsymbol{\gamma};\boldsymbol{\mu}_2}$. This can be seen by taking $\boldsymbol{\gamma} = -\boldsymbol{\beta}_p(\boldsymbol{\mu}_2; \boldsymbol{\mu}_1)$, since one then has

$$\begin{aligned} p_{\boldsymbol{\beta};\boldsymbol{\mu}_1}(u) &= \frac{1}{Z_p(\boldsymbol{\beta}; \boldsymbol{\mu}_1)} \exp\left(\sum_{j=1}^{d} (\beta_j + \gamma_j) t_j(u)\right) \exp\left(\sum_{j=1}^{d} -\gamma_j t_j(u)\right) p_{\boldsymbol{\mu}_1}(u) \\ &= \frac{Z_p(-\boldsymbol{\gamma}; \boldsymbol{\mu}_1)}{Z_p(\boldsymbol{\beta}; \boldsymbol{\mu}_1)} \exp\left(\sum_{j=1}^{d} (\beta_j + \gamma_j) t_j(u)\right) p_{-\boldsymbol{\gamma};\boldsymbol{\mu}_1}(u) \qquad (2.1.2) \\ &= \frac{1}{Z_p(\boldsymbol{\beta} + \boldsymbol{\gamma}; \boldsymbol{\mu}_2)} \exp\left(\sum_{j=1}^{d} (\beta_j + \gamma_j) t_j(u)\right) p_{\boldsymbol{\mu}_2}(u) = p_{\boldsymbol{\beta}+\boldsymbol{\gamma};\boldsymbol{\mu}_2}(u). \end{aligned}$$

2.1.3 The Composite Alternative Generated by A Simple One

We are mostly concerned with testing the null hypothesis $\mathcal{P}$ against simple alternative hypotheses of the form $\{Q\}$ for some distribution Q on $\mathcal{U}$. In particular, we will consider distributions Q that admit a moment generating function and that have a density q relative to the underlying measure ν. While the former is a strong condition, it holds in many cases of interest. For our analysis, it will be beneficial to define a second exponential family $\mathcal{Q}$ for U with distributions $Q_{\boldsymbol{\beta};\boldsymbol{\mu}^*}$ and corresponding densities

$$q_{\boldsymbol{\beta};\boldsymbol{\mu}^*}(u) = \frac{1}{Z_q(\boldsymbol{\beta}; \boldsymbol{\mu}^*)} \cdot \exp\left(\sum_{j=1}^{d} \beta_j t_j(u)\right) \cdot q(u), \qquad (2.1.3)$$

where $\boldsymbol{\mu}^*$ is the mean of X under Q, and $Z_q(\boldsymbol{\beta}; \boldsymbol{\mu}^*)$ is the normalizing constant. The notational conventions that we use for $\mathcal{Q}$ will be completely analogous to that for $\mathcal{P}$, e.g. $\boldsymbol{\beta}_q(\cdot, \mu^*)$, $\boldsymbol{\mu}_q(\cdot, \mu^*)$, Σ_q, etc. Since Q is assumed to have a moment generating function, the canonical domain $\mathtt{B}_{q,\boldsymbol{\mu}^*}$ is nonempty and contains a neighborhood of $\mathbf{0}$. Similarly, the mean-value space $\mathtt{M}_q$ is also nonempty and contains a neighborhood of $\boldsymbol{\mu}^*$. We further have the following: if we take any other $Q' \in \mathcal{Q}$, say $Q' = Q_{\boldsymbol{\mu}'}$ for $\boldsymbol{\mu}' \in \mathtt{M}_q$, then the 'constructed' family around Q', i.e. $\{q_{\boldsymbol{\beta};\boldsymbol{\mu}'} : \boldsymbol{\beta} \in \mathtt{B}_{q;\boldsymbol{\mu}'}\}$ coincides with $\mathcal{Q}$ (as was the case for $\mathcal{P}$, by (2.1.2)).

We may think of the null $\mathcal{P}$ and the generated family $\mathcal{Q}$ as two different exponential families that share the same sufficient statistic. Moreover, as we shall see below, there are many examples where their mean-value spaces are equal, that is, $\mathtt{M}_q = \mathtt{M}_p$. In this case $\mathcal{P}$ and $\mathcal{Q}$ are "matching" pairs: they share the same sufficient statistic as well as the same set of means for this statistic.

2.1.4 E-variables

We use e-variables to gather evidence against the null hypothesis $\mathcal{P}$. An e-variable is a non-negative statistic with expected value bounded by one under the null, i.e. a non-negative statistic $S(U)$ such that $\mathbb{E}_P[S(U)] \leq 1$ for all $P \in \mathcal{P}$. We give only a brief introduction to e-variables here and refer to e.g. [42, 71] for detailed discussions. The realization of an e-variable on observed data will be referred to as an e-value, though the two terms are often used interchangeably. Large e-values give evidence against the null hypothesis, since by Markov's inequality we have that $Q(S(U) \geq \frac{1}{\alpha}) \leq \alpha$ for any e-variable $S(U)$ and $Q \in \mathcal{P}$. The focus here is on a static setting, where e-variables are computed for a single block of data (i.e. one observation of U). However, the main application of e-variables is in anytime-valid settings, where data arrives sequentially and one wants a type-I error guarantee uniformly over time. Indeed, it is well-known that the product of sequentially computed e-variables again gives an e-variable, even if the definition of each subsequent e-variable depends on past e-values, which leads to an easy extension of the methods described here to such anytime-valid settings [71, 42].

Since large e-values give evidence against the null, we look for e-variables that are, on average, 'as large as possible' under the alternative hypothesis. In particular, we study growth-rate optimal (GRO) e-variables, an optimality criterion embraced implicitly or explicitly by most of the e-community [71]. [42] define the GRO e-variable for single outcome U, relative to a simple alternative $\{Q\}$, to be the e-variable S that, among all e-variables, maximizes the growth-rate $\mathbb{E}_{U \sim Q}[\log S(U)]$ (also known as e-power [94, 100]). In a celebrated result, Grünwald et al. [42] (see also [57, 58]) show that the GRO e-variable is given by:

$$\frac{q(U)}{p_{\leftsquigarrow q}(U)}, \tag{2.1.4}$$

where $p_{\leftsquigarrow q}$ denotes the reverse information projection of Q on the convex hull of the null $\mathcal{P}$. The reverse information projection of Q on $\text{CONV}(\mathcal{P})$ is defined as the distribution that uniquely achieves $\inf_{P \in \text{CONV}(\mathcal{P})} D(Q \| P)$, which is known to exist whenever the latter is finite [60, 57]. Here, $D(Q\|P)$ denotes the Kullback-Leibler (KL) divergence between Q and P, both defined as distributions for U. In this article, all reverse information projection will be on $\text{CONV}(\mathcal{P})$, so we will not explicitly mention the domain of projection everywhere (i.e. referring to it simply as 'the reverse information projection of Q'). The growth rate achieved by the GRO e-variable is given by

$$\mathbb{E}_Q \left[\log \frac{q(U)}{p_{\leftsquigarrow q}(U)} \right] = D(Q \| P_{\leftsquigarrow q}) = \inf_{P \in \text{CONV}(\mathcal{P})} D(Q \| P). \tag{2.1.5}$$

However, due to the fact that, with the exception of the Bernoulli and multinomial models, exponential families are not convex sets of distributions, finding the reverse information projection can be quite challenging [56, 44]. In this paper we provide a simple and easily verifiable condition under which

$$\inf_{P \in \text{CONV}(\mathcal{P})} D(Q \| P) = \min_{P \in \mathcal{P}} D(Q \| P), \tag{2.1.6}$$

that is, the infimum is achieved by an element of $\mathcal{P}$, so that the problem greatly simplifies.

In that case, the GRO e-variable simply takes on the form of a likelihood ratio between Q and a particular member of $\mathcal{P}$, i.e.

$$\frac{q(U)}{p(U)}, \tag{2.1.7}$$

which we will refer to as a simple e-variable relative to Q. We will frequently use the fact (following from Corollary 1 of [42, Theorem 1]) that there can be at most one simple e-variable with respect to any fixed alternative, i.e. of the form (2.1.7). This is captured by the following proposition.

Proposition 1. *Fix a probability measure Q on U. If there exists a simple e-variable relative to Q, then it must be the GRO e-variable for testing $\mathcal{P}$ against alternative $\{Q\}$.*

A big advantage of simple e-variables—besides their simplicity—is that their optimality extends beyond the static setting. That is, suppose we were to observe independent copies $U_1, U_2, \ldots$ of the data and assume that a simple e-variable of the form (2.1.7) exists. As alluded to before, we can measure the total evidence as $\prod_{i=1}^n q(U_i)/p(U_i)$, which defines an e-variable for any fixed $n \in \mathbb{N}$. Instead of thinking of this as multiplication of individual e-variables, one can think of it as a likelihood ratio of $U_1, \ldots, U_n$. Proposition 1 then implies that $\prod_{i=1}^n q(U_i)/p(U_i)$ is the GRO e-variable for testing $\mathcal{P}$ against $\{Q\}$ based on n data points. This statement shows that for any fixed sample size n, the best e-variable (in the GRO sense of 2.1.5) is the simple likelihood ratio. Moreover, for applications where the sample size is not fixed beforehand, Koolen et al. [53, Theorem 12] show that a more flexible statement is also true: if τ is any stopping time that is adapted to the data filtration, then $q(U^\tau)/p(U^\tau)$ is also a maximizer of $\mathbb{E}[\ln S_\tau]$ over all processes $S = (S_n)_{n\in\mathbb{N}}$ with $\mathbb{E}[S_\tau] \leq 1$. While we will not explicitly consider this type of sequential optimality in the following, it is one of the main motivating factors behind this work.

We assume throughout this paper that, for any considered alternative Q, there exists a $\boldsymbol{\mu}^* \in \mathtt{M}_p$ such that $\mathbb{E}_{X\sim Q}[X] = \boldsymbol{\mu}^*$. By a standard property of exponential families, the KL divergence from Q to $\mathcal{P}$ is then minimized by the element of $\mathcal{P}$ with the same mean as Q. If (2.1.6) holds, then $P_{\boldsymbol{\mu}^*}$ must therefore be the reverse information projection of Q. It follows that, if a simple e-variable with respect to Q exists, then it is given by $q(U)/p_{\boldsymbol{\mu}^*}(U)$.

2.2 Existence of Simple Local E-Variables

Here we will show how the family $\mathcal{Q}$ is related to the question of whether $q(U)/p_{\boldsymbol{\mu}^*}(U)$ is a *local* GRO e-variable around $\boldsymbol{\mu}^*$. We say that a nonnegative statistic $S(U)$ is a local e-variable around $\boldsymbol{\mu}^*$ if there exists a connected open subset $\mathtt{B}'_{\boldsymbol{\mu}^*}$ of $\mathtt{B}_{p;\boldsymbol{\mu}^*} \cap \mathtt{B}_{q;\boldsymbol{\mu}^*}$ containing $\mathbf{0}$ such that S is an e-variable relative to $\mathcal{P}' = \{P_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \mathtt{B}'_{\boldsymbol{\mu}^*}\}$, i.e. $\sup_{\boldsymbol{\beta}\in\mathtt{B}'_{\boldsymbol{\mu}^*}} \mathbb{E}_{P_{\boldsymbol{\beta};\boldsymbol{\mu}^*}}[S] \leq 1$. If S also maximizes $\mathbb{E}_Q[\ln S(U)]$ among all e-variables relative to $\mathcal{P}'$, then we say that S is a local GRO e-variable with respect to Q. A local (GRO) e-variable may not be an e-variable relative to the full null hypothesis $\mathcal{P}$, but it is a an e-variable relative to some smaller null hypothesis, restricted to all distributions in the null with mean in a neighborhood of $\boldsymbol{\mu}^*$. Investigating when local e-variables exist provides the basic insight on top of which the subsequent, much stronger Theorem 1 about 'global' e-variables is built. As stated in Section 2.1.3, we may view $\mathcal{P}$ and $\mathcal{Q}$ as two families with the same sufficient statistic, only differing in their carrier, which for $\mathcal{P}$ is $p_{\boldsymbol{\mu}^*} = p_{\mathbf{0};\boldsymbol{\mu}^*}$ and for $\mathcal{Q}$ is $q_{\mathbf{0};\boldsymbol{\mu}^*} = q = q_{\boldsymbol{\mu}^*}$: we can and will denote the original Q also by $Q_{\boldsymbol{\mu}^*}$.

Define the function $f(\cdot;\boldsymbol{\mu}^*) : \mathtt{B}_{p;\boldsymbol{\mu}^*} \cap \mathtt{B}_{q;\boldsymbol{\mu}^*} \to \mathbb{R}$ as

$$f(\boldsymbol{\beta};\boldsymbol{\mu}^*) := \log \mathbb{E}_{P_{\boldsymbol{\beta};\boldsymbol{\mu}^*}}\left[\frac{q_{\boldsymbol{\mu}^*}(U)}{p_{\boldsymbol{\mu}^*}(U)}\right] = \log Z_q(\boldsymbol{\beta};\boldsymbol{\mu}^*) - \log Z_p(\boldsymbol{\beta};\boldsymbol{\mu}^*), \tag{2.2.1}$$

where the equality comes from the fact that we can rewrite the density in the numerator as $q_{\boldsymbol{\mu}^*}(U) = Z_q(\boldsymbol{\beta};\mu^*)\exp(\sum_{j=1}^d \beta_j t_j(u))^{-1} q_{\boldsymbol{\beta};\boldsymbol{\mu}^*}(U)$ and similar for the density in the denominator. It should be clear that the function $f(\cdot;\boldsymbol{\mu}^*)$ is highly related to the question we are interested in. Indeed, $q_{\boldsymbol{\mu}^*}(U)/p_{\boldsymbol{\mu}^*}(U)$ is a local e-variable relative to $\mathcal{P}' = \{P_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \mathtt{B}'_{\boldsymbol{\mu}^*}\}$ if and only if $\sup_{\boldsymbol{\beta}\in\mathtt{B}'_{\boldsymbol{\mu}^*}} f(\boldsymbol{\beta};\boldsymbol{\mu}^*) \leq 0$. Equivalently, since $f(\mathbf{0};\boldsymbol{\mu}^*) = \mathbf{0}$, we have that $q_{\boldsymbol{\mu}^*}/p_{\boldsymbol{\mu}^*}$ is a local e-variable around $\boldsymbol{\mu}^*$ if and only if there is a local maximum at $\mathbf{0}$. To investigate when this happens, a standard result on exponential families gives the following:

$$\nabla f(\boldsymbol{\beta};\boldsymbol{\mu}^*) = \mathbb{E}_{Q_{\boldsymbol{\beta};\boldsymbol{\mu}^*}}[X] - \mathbb{E}_{P_{\boldsymbol{\beta};\boldsymbol{\mu}^*}}[X] \tag{2.2.2}$$

In particular, it follows that $\nabla f(\mathbf{0};\boldsymbol{\mu}^*) = \boldsymbol{\mu}^* - \boldsymbol{\mu}^* = \mathbf{0}$. Thus, $q_{\boldsymbol{\mu}^*}/p_{\boldsymbol{\mu}^*}$ is a local e-variable around $\boldsymbol{\mu}^*$ if and only if the $d \times d$ Hessian matrix of second partial derivatives of $f(\cdot;\boldsymbol{\mu}^*)$, is negative semidefinite in $\mathbf{0}$. By (2.2.1)-(2.2.2) and using a convex duality property of exponential families, this is equivalent to

$$I_p(\mathbf{0};\boldsymbol{\mu}^*) - I_q(\mathbf{0};\boldsymbol{\mu}^*) = \Sigma_p(\boldsymbol{\mu}^*) - \Sigma_q(\boldsymbol{\mu}^*) \text{ is positive semidefinite,}$$

where I_p and I_q denote the Fisher information matrix in terms of the canonical parameter spaces of $\mathcal{P}$ and $\mathcal{Q}$, respectively. We have thus proven our first result:

Proposition 2. *$q_{\boldsymbol{\mu}^*}(U)/p_{\boldsymbol{\mu}^*}(U)$ is a local e-variable around $\boldsymbol{\mu}^*$ (and therefore, by Proposition 1, a GRO local e-variable) if and only if $\Sigma_p(\boldsymbol{\mu}^*) - \Sigma_q(\boldsymbol{\mu}^*)$ is positive semidefinite.*

The surprising result that follows below essentially adds to this that, if for every $\boldsymbol{\mu}^* \in \mathtt{M}_q$, $q_{\boldsymbol{\mu}^*}/p_{\boldsymbol{\mu}^*}$ is a local e-variable, then also for every $\boldsymbol{\mu}^*$, we have that $q_{\boldsymbol{\mu}^*}/p_{\boldsymbol{\mu}^*}$ is a full, global e-variable!

2.3 Existence of Simple Global E-Variables (Main Result)

The theorem below gives eight equivalent characterizations of when a global GRO e-variable exists. Not all characterizations are equally intuitive and informative: the simplest ones are Part 1 and 3. To appreciate the more complicated characterizations as well, it is useful to first recall some convex duality properties concerning derivatives of KL divergences with regular exponential families [see e.g. [35], Section 18.4.3]:

$$\boldsymbol{\beta}_p(\boldsymbol{\mu}; \boldsymbol{\mu}^*) = \nabla_{\boldsymbol{\mu}} D(P_{\boldsymbol{\mu}} \| P_{\boldsymbol{\mu}^*}), \tag{2.3.1}$$

$$\left(\Sigma_p^{-1}(\boldsymbol{\mu})\right)_{ij} = \frac{d^2}{d\boldsymbol{\mu}_i d\boldsymbol{\mu}_j} D(P_{\boldsymbol{\mu}} \| P_{\boldsymbol{\mu}^*}), \tag{2.3.2}$$

and analogous for $\mathcal{Q}$. That is, the gradient of the KL divergence in its first argument at $\boldsymbol{\mu}$ is given by the canonical parameter vector corresponding to $\boldsymbol{\mu}$, and the Hessian is given by the Fisher information, i.e. the inverse covariance matrix.

Theorem 1. *Let $\mathcal{P}$ be a regular exponential family with mean-value parameter space $\mathtt{M}_p$. Fix a distribution Q for U with $\mathbb{E}_Q[X] = \boldsymbol{\mu}^*$ for some $\boldsymbol{\mu}^* \in \mathtt{M}_p \subseteq \mathbb{R}^d$ and consider the corresponding $\mathcal{Q}$ as defined above. Suppose that $\mathtt{M}_q$ is convex, $\mathtt{M}_q \subseteq \mathtt{M}_p$, and $\mathtt{B}_{p;\boldsymbol{\mu}} \subseteq \mathtt{B}_{q;\boldsymbol{\mu}}$ for all $\boldsymbol{\mu} \in \mathtt{M}_q$. Then the following statements are equivalent:*

1. *$\Sigma_p(\boldsymbol{\mu}) - \Sigma_q(\boldsymbol{\mu})$ is positive semidefinite for all $\boldsymbol{\mu} \in \mathtt{M}_q$.*
2. *$\left(\boldsymbol{\beta}_p(\boldsymbol{\mu}; \boldsymbol{\mu}') - \boldsymbol{\beta}_q(\boldsymbol{\mu}; \boldsymbol{\mu}')\right)^T \cdot (\boldsymbol{\mu} - \boldsymbol{\mu}') \leq 0$ for all $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathtt{M}_q$.*
3. *$D(Q_{\boldsymbol{\mu}} \| Q_{\boldsymbol{\mu}'}) \geq D(P_{\boldsymbol{\mu}} \| P_{\boldsymbol{\mu}'})$ for all $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathtt{M}_q$.*
4. *$\log Z_p(\boldsymbol{\beta}; \boldsymbol{\mu}) \geq \log Z_q(\boldsymbol{\beta}; \boldsymbol{\mu})$ for all $\boldsymbol{\mu} \in \mathtt{M}_q, \boldsymbol{\beta} \in \mathtt{B}_{p;\boldsymbol{\mu}}$.*
5. *$q_{\boldsymbol{\mu}}(U)/p_{\boldsymbol{\mu}}(U)$ is a global e-variable for all $\boldsymbol{\mu} \in \mathtt{M}_q$.*
6. *$q_{\boldsymbol{\mu}}(U)/p_{\boldsymbol{\mu}}(U)$ is the global GRO e-variable w.r.t. $Q_{\boldsymbol{\mu}}$ for all $\boldsymbol{\mu} \in \mathtt{M}_q$.*
7. *$q_{\boldsymbol{\mu}}(U)/p_{\boldsymbol{\mu}}(U)$ is a local e-variable for all $\boldsymbol{\mu} \in \mathtt{M}_q$.*
8. *$q_{\boldsymbol{\mu}}(U)/p_{\boldsymbol{\mu}}(U)$ is a local GRO e-variable w.r.t. $Q_{\boldsymbol{\mu}}$ for all $\boldsymbol{\mu} \in \mathtt{M}_q$.*

Note that the canonical parameter space of a full exponential family is always convex, but the mean-value space need not be [33]. Still, in all examples we consider below, the constructed family $\mathcal{Q}$ will in fact be a *regular* exponential family, and then the convexity requirement must hold.

In the one-dimensional case, the first statement simplifies to $\sigma_p^2(\mu) \geq \sigma_q^2(\mu)$ for all $\mu \in \mathtt{M}_q$. Similarly, the second statement reduces to $\boldsymbol{\beta}_q(\mu; \mu') \geq \boldsymbol{\beta}_p(\mu; \mu')$ for all $\mu \in \mathtt{M}_q$ such that $\mu > \mu'$ and $\boldsymbol{\beta}_q(\mu; \mu') \leq \boldsymbol{\beta}_p(\mu; \mu')$ for all $\mu \in \mathtt{M}_q$ such that $\mu < \mu'$ for all $\mu' \in \mathtt{M}_q$.

Using standard properties of Loewner ordering, it can be established that $\Sigma_p(\boldsymbol{\mu}) - \Sigma_q(\boldsymbol{\mu})$ is positive semidefinite if and only if $\Sigma_q^{-1}(\boldsymbol{\mu}) - \Sigma_p^{-1}(\boldsymbol{\mu})$ is [see e.g. [2]. Therefore, recalling (2.3.1) and (2.3.2), statement 1 in Theorem 1 can be thought of as a condition on the second derivative of $D(P_{\boldsymbol{\mu}} \| P_{\boldsymbol{\mu}^*}) - D(Q_{\boldsymbol{\mu}} \| Q_{\boldsymbol{\mu}^*})$, whereas statement 2 refers to its first derivative, and statement 3 to the difference in KL divergence itself. It is somewhat surprising that signs of differences between the second derivatives and separately signs of differences between the first derivatives are sufficient to determine signs of difference between a function itself.

2.3.1 Simplifying Situations

In some special situations, the conditions needed to apply Theorem 1 may be significantly simplified. We now identify two such situations, embodied by Proposition 3 and Corollary 1, that will be useful for our examples below.

First, we note that it is sometimes easy to check that either $\mathtt{M}_p = \mathtt{M}_q$ or $\mathtt{B}_{p;\boldsymbol{\mu}^*} = \mathtt{B}_{q;\boldsymbol{\mu}^*}$. The following proposition shows that, in the 1-dimensional setting, this is already sufficient to apply the theorem (we do not know whether an analogous result holds in higher dimensions):

Proposition 3. *Let $\mathcal{P}$ be a 1-dimensional regular exponential family with mean-value parameter space $\mathtt{M}_p \subseteq \mathbb{R}$. Fix a distribution Q for U with $\mathbb{E}_Q[X] = \mu^*$ for some $\mu^* \in \mathtt{M}_p$ and consider the corresponding $\mathcal{Q}$ as defined above. Suppose that for all $\mu \in \mathtt{M}_q$, we have $\sigma^2_p(\mu) \geq \sigma^2_q(\mu)$, i.e. the first condition of Theorem 1 holds. Then:*

1. *If $\mathtt{M}_q = \mathtt{M}_p$ then for all $\mu' \in \mathtt{M}_q$, $\mathtt{B}_{p;\mu'} \subseteq \mathtt{B}_{q;\mu'}$, i.e., Theorem 1 is applicable.*
2. *If for some $\mu \in \mathtt{M}_q$, we have that $\mathtt{B}_{p;\mu} = \mathtt{B}_{q;\mu}$ then $\mathtt{M}_q \subseteq \mathtt{M}_p$. Hence if for all $\mu \in \mathtt{M}_q$, we have that $\mathtt{B}_{p;\mu} = \mathtt{B}_{q;\mu}$, then Theorem 1 is applicable.*

The proof is simple and we only sketch it here: for part 1, draw the graphs of $\beta_p(\mu;\mu')$ and $\beta_q(\mu;\mu')$ as functions of $\mu \in \mathtt{M}_q$, noting that both functions must take the value 0 at the point $\mu = \mu'$. Using that $1/\sigma^2_p(\mu)$ is the derivative of $\beta_p(\mu;\mu')$ and similarly for $\sigma^2_q(\mu)$, the function $\beta_q(\mu;\mu')$ must lie above $\beta_p(\mu;\mu')$ for $\mu > \mu'$, and below for $\mu < \mu'$. Therefore the co-domain of β_q must include that of β_p. The second part goes similarly, essentially by flipping the just-mentioned graph of two functions by 90 degrees.

Second, we note that in practice we often have a composite alternative $\mathcal{H}_1$ in mind such that the union of the set of families $\mathcal{Q}$ that can be constructed from $\mathcal{P}$ and any $Q \in \mathcal{H}_1$ in fact coincides with $\mathcal{H}_1$. This is the case in the examples of Section 2.4.1, 2.4.3, 2.4.3 and 2.4.4. The following immediate corollary of Theorem 1 simplifies the analysis in such cases (although we will only explicitly need to invoke it in Section 2.4.4). While in that section, $\mathcal{H}_1$ will itself be an exponential family, we stress that in general, this need not be the case: to apply the corollary it is sufficient for $\mathcal{H}_1$ to be a *union* of exponential families.

Corollary 1. *Let $\mathcal{P}$ be a d-dimensional regular exponential family as before with mean-value parameter space $\mathtt{M}_p$, and let $\mathcal{H}_1 = \bigcup_{\theta \in \Theta} \mathcal{Q}^{(\theta)}$ where each $\mathcal{Q}^{(\theta)}$ is a d-dimensional regular exponential family with the same sufficient statistic as $\mathcal{P}$ and with mean-value parameter space $\mathtt{M}^{(\theta)}_q$ and canonical parameter spaces $\mathtt{B}^{(\theta)}_{q;\boldsymbol{\mu}}$ for $\boldsymbol{\mu} \in \mathtt{M}^{(\theta)}_q$. Suppose that, for each $\theta \in \Theta$, for each $Q \in \mathcal{Q}^{(\theta)}$, the corresponding set $\mathcal{Q}$ as constructed above in terms of $\mathcal{P}$ and Q, happens to be equal to $\mathcal{Q}^{(\theta)}$ and satisfies the pre-condition of Theorem 1, i.e. $\mathtt{M}^{(\theta)}_q$ is convex, $\mathtt{M}^{(\theta)}_q \subseteq \mathtt{M}_p$, and $\mathtt{B}_{p;\boldsymbol{\mu}} \subseteq \mathtt{B}^{(\theta)}_{q;\boldsymbol{\mu}}$ for all $\boldsymbol{\mu} \in \mathtt{M}^{(\theta)}_q$. Then we have, with $Q^{(\theta)}_{\boldsymbol{\mu}}$ (density $q^{(\theta)}_{\boldsymbol{\mu}}$) denoting the element of $\mathcal{Q}^{(\theta)}$ with mean $\boldsymbol{\mu}$. for all $\theta \in \Theta$:*

$$\text{For all } \boldsymbol{\mu} \in \mathtt{M}^{(\theta)}_q\text{: } \frac{q^{(\theta)}_{\boldsymbol{\mu}}(U)}{p_{\boldsymbol{\mu}}(U)} \text{ is the global GRO e-variable w.r.t. } Q^{(\theta)}_{\boldsymbol{\mu}} \Leftrightarrow$$
$$\text{For all } \boldsymbol{\mu} \in \mathtt{M}^{(\theta)}_q\text{: } \Sigma_p(\boldsymbol{\mu}) - \Sigma^{(\theta)}_q(\boldsymbol{\mu}) \text{ is positive semidefinite.}$$

Here $\Sigma_q^{(\theta)}(\boldsymbol{\mu})$ denotes the $d \times d$ covariance matrix of the element of $\mathcal{Q}^{(\theta)}$ with mean-value parameter vector $\boldsymbol{\mu}$.

2.4 Examples

In this section we discuss a variety of settings to which Theorem 1 can be applied. In some cases, this gives new insights into whether simple e-variables exist, and in others it simply gives a reinterpretation of existing results. The examples are broadly divided in terms of the curvature of the function $f(\cdot;\boldsymbol{\mu}^*)$, as defined in (2.2.1). Instances where $f(\cdot;\boldsymbol{\mu}^*)$ is constant will be referred to as having 'zero curvature', those with a constant second derivative as having 'constant curvature', and 'nonconstant curvature' otherwise.

2.4.1 Zero Curvature: Gaussian and Poisson k-sample tests

[44] provide GRO e-values for k-sample tests with regular exponential families. In their setting, data arrives in $k \in \mathbb{N}$ groups, or samples, and they test the hypothesis that all of the data points are distributed according to the same element of some exponential family. That is, let $U = (Y_1, \dots, Y_k)$ for $Y_i \in \mathcal{Y}$, so that $\mathcal{U} = \mathcal{Y}^k$ for some measurable space $\mathcal{Y}$. Furthermore, fix a one-dimensional regular exponential family on $\mathcal{Y}$, given in its mean-value parameterization as $\mathcal{P}_{\text{START}} = \{P_\mu : \mu \in \mathtt{M}_{\text{START}}\}$ with sufficient statistic $t_{\text{START}}(Y)$. The composite null hypothesis $\mathcal{P}$ considered in the k-sample test expresses that $Y_1, \dots, Y_k \overset{\text{i.i.d.}}{\sim} P_\mu$ for some $\mu \in \mathtt{M}_{\text{START}}$. On the other hand, the simple alternative Q that Hao et al. [44] consider is characterized by $\boldsymbol{\mu} = (\mu_1, \dots, \mu_k) \in \mathtt{M}_{\text{START}}^k$, and expresses that the $Y_1, \dots, Y_k$ are independent with $Y_i \sim P_{\mu_i}$ for $i = 1 \dots k$. They show that, for the case that $\mathcal{P}_{\text{START}}$ is either the Gaussian location family or the Poisson family,

$$S(U) := \prod_{i=1}^{k} \frac{p_{\mu_i}(Y_i)}{p_{\bar{\mu}}(Y_i)}, \text{ with } \bar{\mu} = \frac{1}{k}\sum_{i=1}^{k} \mu_i,$$

is a simple e-value relative to Q, and that its expectation is constant as the null varies. That is, for any $\mu' \in \mathtt{M}_{\text{START}}$, it holds that

$$\mathbb{E}_{U \sim P_{\mu'} \times \dots \times P_{\mu'}}[S(U)] = 1. \tag{2.4.1}$$

This finding can now be re-interpreted as an instance of Theorem 1, as we will show in detail for the Poisson family; the analysis for the Gaussian location family is completely analogous. In the Poisson case, $t_{\text{START}}(Y) = Y$, so that $\mathcal{P}$ defines an exponential family on $\mathcal{U}$ with sufficient statistic $X = \sum_{i=1}^k Y_i$ and mean-value space $\mathtt{M}_p = \mathbb{R}^+$. The latter follows because the sum of Poisson data is itself Poisson distributed with mean equal to the sum of means of the original data. Under the alternative, the mean of the sufficient statistic is given by $\mu^* := \mathbb{E}_Q[\sum_{i=1}^k Y_i] = \sum_{i=1}^k \mu_i$, so that the elements of the auxiliary exponential family $\mathcal{Q}$ as in (2.1.3) can be written as

$$q_{\beta;\mu^*}(Y_1,\ldots,Y_k) = \frac{1}{Z_q(\beta;\mu^*)} \cdot \exp\left(\beta \sum_{i=1}^{k} Y_i\right) \cdot q(Y_1,\ldots,Y_k). \tag{2.4.2}$$

Note in particular that $\mathcal{Q}$ is, by construction, a one-dimensional exponential family with sufficient statistic $\sum_{i=1}^{k} Y_i$, which does not equal (yet may be viewed as a subset of) the full k-dimensional exponential family from which Q was originally chosen. The normalizing constant $Z_q(\beta;\mu^*)$ is equal to the moment generating function of X under Q, which is given by

$$Z_q(\beta;\mu^*) = \mathbb{E}_Q\left[\exp\left(\beta \sum_{i=1}^{k} Y_i\right)\right] = \exp\left(\mu^*(e^\beta - 1)\right).$$

It follows that

$$\mathbb{E}_{Q_{\beta;\mu^*}}\left[\sum_{i=1}^{k} Y_i\right] = \frac{\mathrm{d}}{\mathrm{d}\beta} \log Z_q(\beta;\mu^*) = \mu^* e^\beta,$$

which shows that mean-value space of the alternative is again given by $\mathtt{M}_q = \mathbb{R}^+$. Therefore, via Proposition 3, the assumptions of Theorem 1 are satisfied. The element of $\mathcal{P}$ with mean μ^* is given by $P_{\bar{\mu}} \times \cdots \times P_{\bar{\mu}}$, so that

$$\frac{q_{\mu^*}(U)}{p_{\mu^*}(U)} = \prod_{i=1}^{k} \frac{p_{\mu_i}(Y_i)}{p_{\bar{\mu}}(Y_i)}.$$

Under P_{μ^*}, the sufficient statistic $\sum_{i=1}^{k} Y_i$ has the same distribution as under Q_{μ^*}, so that $Z_p(\beta;\mu^*) = Z_q(\beta;\mu^*)$. Consequently, $f(\cdot;\boldsymbol{\mu}^*)$ as in (2.2.1) is zero, so that its second derivative is zero, and condition 1 of Theorem 1 is verified. It follows that, $q_{\mu^*}(U)/p_{\mu^*}(U)$ is the global GRO e-variable with respect to Q_{μ^*}.

2.4.2 Constant Curvature: Multivariate Gaussian Location

Suppose that $\mathcal{P}$ is the multivariate Gaussian location family with some given nondegenerate covariance matrix Σ_p and let Q be any Gaussian distribution with nondegenerate covariance matrix Σ_q. Note that in this case we have that $X = U$, i.e. the sufficient statistic is simply given by the original data. The family $\mathcal{Q}$, generated from Q and $\mathcal{P}$ as in (2.1.3), is the full Gaussian location family with fixed covariance matrix Σ_q. For both $\mathcal{P}$ and $\mathcal{Q}$, the mean-value and canonical spaces are all equal to $\mathbb{R}^d$, so that Theorem 1 applies to the pair $\mathcal{P}$ and $\mathcal{Q}$. Furthermore, the covariance functions are constant, since $\Sigma_p(\boldsymbol{\mu}) = \Sigma_p$ and $\Sigma_q(\boldsymbol{\mu}) = \Sigma_q$ for all $\boldsymbol{\mu} \in \mathbb{R}^d$. It follows that, if $\Sigma_p - \Sigma_q$ is positive semidefinite, then $\Sigma_p(\boldsymbol{\mu}) - \Sigma_q(\boldsymbol{\mu})$ is positive semidefinite for all $\boldsymbol{\mu} \in \mathbb{R}^d$. In that case, Theorem 1 shows that the simple likelihood ratio $q_{\boldsymbol{\mu}}/p_{\boldsymbol{\mu}}$ is the GRO e-value w.r.t. $Q_{\boldsymbol{\mu}}$ for every $\boldsymbol{\mu} \in \mathbb{R}^d$. The growth rate is given by

$$\mathbb{E}_Q \left[\log \frac{q_{\boldsymbol{\mu}}(U)}{p_{\boldsymbol{\mu}}(U)} \right] = D_{\text{GAUSS}}(\Sigma_q \Sigma_p^{-1}),$$

where $D_{\text{GAUSS}}(B) := \frac{1}{2}\left(-\log \det(B) - (d - \operatorname{tr}(B))\right)$, i.e. the standard formula for the KL divergence between two multivariate Gaussians with the same mean.

In the case that $\Sigma_p - \Sigma_q$ is negative semidefinite, the simple likelihood ratio does not give an e-value; the GRO e-value for this case can also be derived however and will be reported on in future work.

2.4.3 Nonconstant Curvature: Univariate Examples

We now discuss three examples with nonconstant curvature. In the first two, Theorem 1 can be used to show the existence of simple e-variables. All three are univariate in nature; in the separate Section 2.4.4 we provide the example of linear regression, which has nonconstant curvature but is multivariate.

More k-Sample Tests

Consider again the k-sample test setting of Section 2.4.1. Besides the Gaussian and Poisson case, [44] identify one more model that gives rise to a k-sample test with a simple e-value: the case that $\mathcal{P}_{\text{START}}$ is the Bernoulli model. The difference with the Gaussian location- and Poisson family is that the involved e-value does not have constant expectation 1 here. Nevertheless, this result for the Bernoulli model can also be cast in terms of Theorem 1 using a different argument.

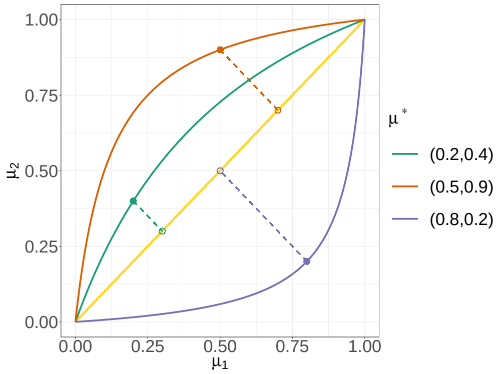


Figure 2.2: The family $\mathcal{Q}$ for various $\boldsymbol{\mu}^*$. The coordinate grid represents the parameters of the full 2-sample Bernoulli family, the straight line shows the parameter space of $\mathcal{P}$, the curved lines show the parameters of the distributions in $\mathcal{Q}$, and the dashed lines show the projection of $\boldsymbol{\mu}^*$ onto the parameter space of $\mathcal{P}$.

Again, $\mathcal{P}$ is an exponential family on $\mathcal{U}$ that states that the k samples are i.i.d. Bernoulli, which has sufficient statistic $X = \sum_{i=1}^{k} Y_i$. Its mean-value space is given by

$\mathtt{M}_p = (0, k)$, since the sum of k i.i.d. bernoulli random variables with parameter μ has a binomial distribution with parameters (k, μ). Under the alternative Q, the k samples are independently Bernoulli distributed with means given by $\boldsymbol{\mu} \in (0,1)^k$, in which case the sum has mean $\mu^* = \sum_{i=1}^k \mu_i$. When constructing the family $\mathcal{Q}$ as in (2.1.3), it can be verified that Q_{β,μ^*} is the product of Bernoulli distributions with means

$$\left(\frac{e^\beta \mu_1}{1-\mu_1+e^\beta\mu_1}, \dots, \frac{e^\beta \mu_k}{1-\mu_k+e^\beta\mu_k}\right). \tag{2.4.3}$$

This family of distributions is illustrated in Figure 2.2 for different choices of $\boldsymbol{\mu}^*$. Seen as a function of β, all entries in (2.4.3) behave as sigmoid functions, so that the sum takes values in $(0, k)$. It follows that the mean-value space of $\mathcal{Q}$ is given by $\mathtt{M}_q = (0, k)$, which equals $\mathtt{M}_p$ — ana also, the canonical spaces are all equal to $\mathbb{R}$. Furthermore, the normalizing constant $Z_q(\beta; \mu^*)$ of $\mathcal{Q}$ must be given by

$$Z_q(\beta; \mu^*) = \prod_{i=1}^k (1 - \mu_i + \mu_i e^\beta).$$

We will now verify that item 4 of Theorem 1 is satisfied by doing a similar construction for arbitrary $\mu \in (0, k)$. The element in $\mathcal{P}$ with mean μ corresponds to Bernoulli parameter μ/k, so that we have

$$Z_p(\beta; \mu) = \mathbb{E}_{P_{\mu^*}}\left[\exp\left(\beta \sum_{i=1}^k Y_i\right)\right] = \left(1 - \frac{\mu}{k} + \frac{\mu}{k}e^\beta\right)^k.$$

Furthermore, there is a corresponding $\boldsymbol{\mu}' \in (0,1)^k$ such that $\sum_{i=1}^k \mu'_i = \mu$ and $\boldsymbol{\mu}'$ can be written as (2.4.3) for a specific β. Repeating the reasoning above gives

$$Z_q(\beta; \mu) = \prod_{i=1}^k (1 - \mu'_i + \mu'_i e^\beta).$$

By concavity of the logarithm, it holds that

$$\log Z_p(\beta; \mu) = k \log\left(1 - \frac{\mu}{k} + \frac{\mu}{k}e^\beta\right) \geq \sum_{i=1}^k \log(1 - \mu'_i + \mu'_i e^\beta) = \log Z_q(\beta; \mu).$$

We can therefore conclude that $q(U)/p_{\mu^*}(U)$ is the GRO e-variable with respect to Q.

[44] investigate several other exponential families for k-sample testing, such as exponential distributions, Gaussian scale, and beta, but none of these give rise to a simple e-value. Parts 1-4 of Theorem 1 provide some insight into what separates these families from the Gaussian location, Poisson, and Bernoulli.

Gaussian Scale Family

Another setting in which Theorem 1 applies is where $\mathcal{P}$ equals the Gaussian scale family with fixed mean, which we take to be 0 without loss of generality. That is, $\mathcal{P} = \{P_{\sigma^2} : \sigma^2 \in \mathtt{M}_p\}$ where P_{σ^2} is the normal with mean 0 and variance σ^2, i.e.

$$p_{\sigma^2}(U) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{1}{2\sigma^2}U^2}. \tag{2.4.4}$$

We will substantially extend this null hypothesis, and hence this example, in Section 2.4.4. For now, note that $\mathcal{P}$ is an exponential family with sufficient statistic $X = U^2$, mean-value parameter σ^2 and mean-value space given by $\mathtt{M}_p = \mathbb{R}^+$. The canonical parameterization of the null relative to any mean-value $\sigma^2 \in \mathtt{M}_p$ is given by

$$p_{\beta;\sigma^2}(U) = \frac{1}{Z_p(\beta;\sigma^2)} \cdot e^{\beta U^2} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-U^2/(2\sigma^2)} \tag{2.4.5}$$

with canonical parameter space $\mathtt{B}_{p;\sigma^2} = (-\infty, 1/(2\sigma^2))$.

As alternative, we take Q to be a Gaussian distribution with some fixed mean $m \neq 0$ and variance s^2. We use m and s^2 instead of μ and σ^2 here to avoid confusion with the mean-value parameters of $\mathcal{P}$. The expected value of X under Q is given by $\sigma^{*2} := \mathbb{E}_Q[X] = s^2 + m^2$. The family $\mathcal{Q} = \{Q_\beta : \beta \in \mathtt{B}_{q;\sigma^{*2}}\}$ as defined by (2.1.3) therefore becomes:

$$q_{\beta;\sigma^{*2}}(U) = \frac{1}{Z_q(\beta;\sigma^{*2})} \cdot e^{\beta U^2} \cdot \frac{1}{\sqrt{2\pi}s} \cdot e^{-c(U-m)^2}, \tag{2.4.6}$$

where $c = 1/(2s^2)$, with $\mathtt{B}_{q;\sigma^{*2}} = (-\infty, c)$. Comparing (2.4.5) and the above confirms that $\mathcal{Q}$ is an exponential family that has the same sufficient statistic, namely U^2, as $\mathcal{P}$, but different carrier.

The normalizing constant Z_q can be computed using (for example) the moment generating function of the noncentral chi-squared.

$$Z_q(\beta;\sigma^{*2}) = \mathbb{E}_Q\left[e^{\beta U^2}\right] = \mathbb{E}_Q\left[e^{\beta s^2 (\frac{U}{s})^2}\right] = (1-2\beta s^2)^{-1/2} \exp\left(\frac{m^2\beta}{1-2\beta s^2}\right),$$

where we use that $(U/s)^2$ has noncentral chi-squared distribution with one degree of freedom and noncentrality parameter m^2/s^2. Plugging this back in (2.4.6) shows that $q_{\beta,\sigma^{*2}}$ is a normal density with mean $cm/(c-\beta)$ and variance $1/(2(c-\beta)) = s^2/(1-2\beta s^2)$. This gives

$$\mathbb{E}_{Q_{\beta;\sigma^{*2}}}[U^2] = \frac{2c^2m^2 - (\beta - c)}{2(\beta - c)^2} \tag{2.4.7}$$

The mean-value parameter space of $\mathcal{Q}$ is thus given by $\mathtt{M}_q = \{\mathbb{E}_{Q_{\beta;\sigma^{*2}}}[U^2], \beta < c\} = \mathbb{R}^+$ which is equal to $\mathtt{M}_p$. Thus, this constructed family does not equal the natural choice of composite alternative that Q was also chosen from, i.e. the (two-dimensional) set of all

Gaussians with arbitrary variance mean unequal to zero. However, it does correspond to a specific one-dimensional subset thereof, as was illustrated in Figure 2.1 in the introduction.

Since $\mathtt{M}_q = \mathtt{M}_p$, we get, via Proposition 3 that a simple e-variable w.r.t. Q exists if, for all $\sigma^2 > 0$, we have that $\text{VAR}_{P_{\sigma^2}}[U^2] \geq \text{VAR}_{Q_{\sigma^2}}[U^2]$. We now show this to be the case. We have

$$\text{VAR}_{P_{\sigma^2}}[U^2] = 2\sigma^4 = 2(\mathbf{E}_{P_{\sigma^2}}[U^2])^2 = 2(\mathbf{E}_{Q_{\sigma^2}}[U^2])^2.$$

It is therefore sufficient to check whether, for all $\sigma^2 > 0$, it holds that $\text{VAR}_{Q_{\sigma^2}}[U^2] \leq 2(\mathbb{E}_{Q_{\sigma^2}}[U^2])^2$. We can either verify this using existing results by noting that, no matter how m and s^2 were chosen, U^2 has a noncentral χ^2-distribution under each Q_{σ^2}, for which it is known that the inequality holds. We can also easily verify it explicitly now that we have already found an expression for $Z_q(\beta; \sigma^{*2})$: since there is no more mention of the null hypothesis, it is equivalent to check whether for each $\beta \in \mathtt{B}_{q;\sigma^{*2}}$ we have

$$\text{VAR}_{Q_{\beta;\sigma^{*2}}}[U^2] \leq 2\left(\mathbb{E}_{Q_{\beta,\sigma^{*2}}}[U^2]\right)^2.$$

To this end, the variance function in terms of β can be computed as

$$\text{VAR}_{Q_{\beta;\sigma^{*2}}}[U^2] = \frac{d^2}{d\beta^2} \log Z_q(\beta; \sigma^{*2}) = -\frac{4c^2m^2 - (\beta - c)}{2(\beta - c)^3}. \tag{2.4.8}$$

Comparing this to (2.4.7) shows that the condition above indeed holds, from which we can conclude that $q(U)/p_{\sigma^{*2}}(U)$ is an e-value.

Finally, note that even though the mean-value parameter spaces of $\mathcal{P}$ and $\mathcal{Q}$ are equal, the canonical spaces are not: $\mathtt{B}_{p;\sigma^{*2}}$ is a proper subset of $\mathtt{B}_{q;\sigma^{*2}}$. More generally, for any $\sigma'^2 > 0$ different from the σ^{*2} we started with, the canonical spaces $\mathtt{B}_{p;\sigma'^2}$ and $\mathtt{B}_{q;\sigma'^2}$ both change but remain unequal. Still, Proposition 3 ensures that we will have $\mathtt{B}_{p;\sigma'^2} \subset \mathtt{B}_{q;\sigma'^2}$.

NEFS and their Variance Functions

In this section, we consider the setting where $\mathcal{P}$ is a one-dimensional natural exponential family (NEF) and Q is also an element of an NEF. This setting is particularly suited for the analysis above, because the constructed family $\mathcal{Q}$ can be seen to equal the NEF that Q was chosen from. We therefore do not differentiate between the simple or composite alternative in this section. Furthermore, NEFs are fully characterized by the pair $(\sigma^2(\mu), \mathtt{M})$, where $\mathtt{M}$ is the mean-value parameter space and $\sigma^2(\mu)$ is the variance function as defined before. A wide variety of NEFS and their corresponding variance functions have been studied in the literature [see e.g. 66, 50, 9] and this can be used in conjunction with Theorem 1 to quickly check on a case-by-case basis whether any given pair of NEFs provides a simple e-variable.

For example, let $\mathcal{P} = \{P_{\lambda,r} : \lambda \in \mathbb{R}^+\}$ be the set of Gamma distributions for U with varying scale parameter λ and fixed shape parameter $r > 0$. The sufficient statistic is given by $X = U$ and its mean under $P_{\lambda,r}$ equals $r\lambda$, so the mean-value parameter space is $\mathtt{M}_p = \mathbb{R}^+$. The variance function is given by $\sigma_p^2(\mu) = \mu^2/r$. If we set Q to $P_{\lambda^*,r'}$ for specific $\lambda^*, r' \in \mathbb{R}^+$, then $\mathcal{Q}$ is the set of Gamma distributions with fixed shape parameter r'.

Similarly, let $\mathcal{P}$ be the set of negative binomial distributions with fixed number of successes $n \in \mathbb{N}$ and let Q be any Poisson distribution, so that $\mathcal{Q}$ equals the Poisson family. The variance functions are given by $\sigma_p^2(\mu) = \mu^2/n + \mu$ and $\sigma_q^2(\mu) = \mu$, respectively. It is trivially true that $\sigma_p^2(\mu) \geq \sigma_q^2(\mu)$ for all μ, so Theorem 1 reveals that a simple e-variables exists with respect to any element of the Poisson family. More generally, we may look at the Awad-Bar-Lev-Makov (ABM) class of NEFs [10, 5, 11] that are characterized by mean-value parameter space $\mathtt{M} = \mathbb{R}^+$ and variance function

$$\sigma_s^2(\mu) = \mu \left(1 + \frac{\mu}{s}\right)^r, \; s > 0, \; r = 0, 1, 2, \ldots$$

This class was proposed as part of a general framework for zero-inflated, over-dispersed alternatives to the Poisson model (which would arise for $r = 0$). The case $r = 1$ recovers the negative binomial distribution and $r = 2$ is called the generalized Poisson or Abel distribution. As was the case for the negative binomial distribution, it follows from Theorem 1 that simple e-variables exist for testing any of the ABM NEFs against the Poisson model.

Much more generally, consider the Tweedie-Bar-Lev-Enis class [7] of NEFs that have mean-value space $\mathtt{M} = \mathbb{R}^+$ and power variance functions

$$\sigma^2(\mu) = a\mu^\gamma, \; a > 0, \; \mu > 0, \; \gamma \geq 1.$$

We require $\gamma \geq 1$ because there are no families of this form with $\gamma \in (0, 1)$ and while there are families in this class with $\gamma < 0$, they are not regular and therefore beyond the scope of this paper. The cases $\gamma = 1$ (Poisson) and $\gamma = 2$ (Gamma families, with a depending on the shape parameter) were already encountered above. If we test between two of such families, say $\mathcal{P}$ with $\sigma_p^2(\mu) = a_p\mu^{\gamma_p}$ and $\mathcal{Q}$ with $\sigma_q^2(\mu) = a_q\mu^{\gamma_q}$ that share the same underlying sample space, there do not exist simple e-variables in general. Indeed, we have that $\sigma_p^2(\mu) \geq \sigma_q^2(\mu)$ if and only if $\mu^{\gamma_p - \gamma_q} \geq a_q/a_p$, which, for certain combinations of parameters, does not hold for all $\mu \in \mathtt{M}$. Since this condition might hold for some μ but not for others, this suggests that there may be cases where we find local e-variables that are not global.

Let us investigate this for $(a_p, \gamma_p) = (1, 2)$ and $(a_q, \gamma_q) = (1/2, 3)$, which corresponds to the family of exponential distributions and the family of inverse Gaussian distributions with shape parameter $\lambda := a_q^{-1} = 2$ respectively. In this case, it holds that $\sigma_p^2(\mu) \geq \sigma_q^2(\mu) \Leftrightarrow \mu \leq a_q^{-1}$. It follows from the analysis in Section 2.2 that $q_\mu(U)/p_\mu(U)$ is a local e-variable for $\mu \leq a_q^{-1}$. However, since the condition does not hold for all μ we cannot use Proposition 3 (or, equivalently, because, as we will see, the preconditions for Theorem 1 do not hold), this need not necessarily also be a global e-variable. In fact, the expected value under $\mu' \in \mathtt{M}$ is given by

$$\mathbb{E}_{U \sim P_{\mu'}} \left[\frac{q_\mu(U)}{p_\mu(U)}\right] = \int_0^\infty \frac{1}{\mu'} \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left(-\frac{\lambda(x - \mu)^2}{2\mu^2 x} + \frac{x}{\mu} - \frac{x}{\mu'}\right) dx, \tag{2.4.9}$$

which diverges for $\mu' \geq (1/\mu - \lambda/(2\mu^2))^{-1}$. The latter is vacuous for $\mu \leq \lambda/2$, which means that for such μ we might still get a global e-variable. For $\mu \in (\lambda/2, \lambda)$, this shows that we will get a local e-variable that is not a global e-variable. These different regimes are illustrated in Figure 2.3. For $\mu > 1$, the lines stop when the integral in (2.4.9) starts diverging. To see how the potential divergence (for large enough μ', in the regime $1 < \mu < 2$) plays out in terms of the function f in (2.2.1), consider for example $\mu = 3/2$. Then, as is immediate from the definition of exponential distributions and the inverse Gaussian density with $\lambda = 2$ we have $q_{\beta;\mu}(x) \propto \exp((\beta - 4/9)x)h(x)$ with h the probability density on $\mathbb{R}^+$ given by $h(x) = \sqrt{1/(\pi x^3)}\exp(-1/x)$, whereas $p_{\beta;\mu} \propto \exp((\beta - 2/3)x)$. We see that $\mathsf{B}_{p;\mu} = (-\infty, 6/9)$ whereas $\mathsf{B}_{q;\mu} = (-\infty, 4/9)$. Thus, as $\beta \uparrow 4/9$, we get that $\log Z_p(\beta)$ converges to a finite constant whereas $\log Z_q(\beta) \uparrow \infty$, so that $f(\beta, \mu) \to \infty$, with f the function in (2.2.1), as it should.

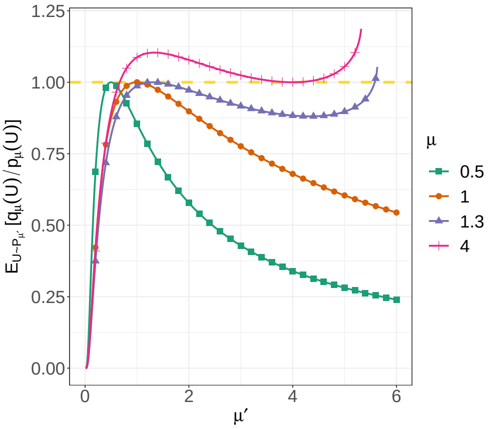


Figure 2.3: The expected value of $q_\mu(U)/p_\mu(U)$ under the null $P_{\mu'}$ for varying μ'.

2.4.4 The Linear Model

We now show that Theorem 1 allows us to conclude that simple e-variables exist for the linear model, i.e. standard linear regression with Gaussian noise, where the null hypothesis $\mathcal{P}$ is a subset of the alternative $\mathcal{H}_1$ obtained by setting the regression parameter of a control random variable to 0, as soon as we allow the variance in $\mathcal{P}$ to be a free parameter. This was shown directly, without associating a specific family $\mathcal{Q}$ to $\mathcal{P}$, in an unpublished master thesis [30]. De Jong's treatment involved a lot of hard-to-interpret calculus, much of it discovered by trial-and-error. The advantage of the present treatment is that Theorem 1 clearly guides the reasoning and suggests what formulas to verify. The setting is really a vast extension of that of Section 2.4.3, which is (essentially) retrieved if below we set $d = 0$. Interestingly, e-variables for linear models were already derived by [70] and [64], based on right-Haar priors. The current approach provides a different type of e-variable which has the advantage that it does not require the variance under the alternative to be equipped with a right-Haar prior:

while for convenience we give the treatment below for $\mathcal{H}_1$ with the variance σ^2 being left a free parameter, we can freely apply the results to any $\mathcal{H}_1' \subset \mathcal{H}_1$, in particular with $\mathcal{H}_1'$ restricted to densities with a fixed variance. The price to pay is that the e-variables derived below, while growth-optimal for the fixed $Q \in \mathcal{H}_1$ relative to which they are defined, will in general not be GROW (*worst-case* growth optimal, see [42]) in the worst-case over all distributions in $\mathcal{H}_1$ when σ^2 varies within $\mathcal{H}_1$.

Assume then that data arrives as a block of outcomes together with given covariate vectors, i.e. $U = ((Y_1, \boldsymbol{x}_1), \ldots, (Y_n, \boldsymbol{x}_n))$ with $Y_i \in \mathbb{R}$ and $\boldsymbol{x}_i = (x_{i,0}, x_{i,1}, \ldots, x_{i,d})^T \in \mathbb{R}^{d+1}$. Define the conditional normal distributions $G_{\sigma,\boldsymbol{\gamma}}$ with corresponding densities

$$g_{\sigma,\boldsymbol{\gamma}}(Y^n) := g_{\sigma,\boldsymbol{\gamma}}(Y^n \mid \boldsymbol{x}^n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot e^{-\frac{1}{2\sigma^2}\sum(Y_i - \nu_i)^2} \tag{2.4.10}$$

with $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_d)^T \in \mathbb{R}^{d+1}$ and

$$\nu_i := \boldsymbol{\gamma}^T \boldsymbol{x}_i. \tag{2.4.11}$$

Here and in the sequel, sums without explicitly denoted ranges are invariably taken to be over $i = 1..n$ and we omit the conditional $\boldsymbol{x}^n$ from the notation, since they are fixed throughout the following analysis.

We focus on the most common case in which one of the covariates, $x_{i,0}$, has a special status and we want to test whether the corresponding coefficient γ_0 is equal to 0. We thus want to design an e-variable for testing any simple alternative Q taken from the full alternative hypothesis $\mathcal{H}_1$ vs. the null $\mathcal{P}$, where $\mathcal{H}_1$ and $\mathcal{P}$ are respectively given by:

$$\mathcal{H}_1 = \{G_{\sigma,\boldsymbol{\gamma}} : \boldsymbol{\gamma} \in \mathbb{R}^{d+1}, \gamma_0 \neq 0, \sigma > 0\} \quad ; \quad \mathcal{P} = \{G_{\sigma,\boldsymbol{\gamma}} : \boldsymbol{\gamma} \in \mathbb{R}^{d+1}, \gamma_0 = 0, \sigma > 0\}. \tag{2.4.12}$$

We make the standard assumption that $n \geq d + 1$ and that the matrix $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ has maximal (i.e. $d + 1$) rank. Now define the transformed parameters $\lambda := -1/(2\sigma^2)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T$ with, for $j = 1..d$, $\beta_j := \gamma_j/\sigma^2$ and $\theta := \gamma_0/\sigma^2$ and set $t_j(Y^n) = \sum Y_i x_{i,j}$. Rewriting the likelihood (2.4.10) in terms of this new parameterization and the t_j, denoting the resulting densities by $f^{(\theta)}_{\lambda,\boldsymbol{\beta}}$, we see that

$$f^{(\theta)}_{\lambda,\boldsymbol{\beta}}(y^n) = g_{\sigma,\boldsymbol{\gamma}}(y^n) = \exp\left(\lambda \sum y_i^2 + \theta t_0(y^n) + \sum_{j=1}^{d} \beta_j t_j(y^n)\right) \cdot h_1(y^n) h_2(\sigma, \boldsymbol{\gamma}) \tag{2.4.13}$$

for some function h_1 not depending on the parameters and h_2 not depending on the data y^n. Let, for $\theta \in \mathbb{R}$, $\mathcal{Q}^{(\theta)}$ be the set of distributions $F^{(\theta)}_{\lambda,\boldsymbol{\beta}}$ with densities $f^{(\theta)}_{\lambda,\boldsymbol{\beta}}$. We see that for each $\theta \in \mathbb{R}$, $\mathcal{Q}^{(\theta)}$ is a $(d+1)$-dimensional exponential family with sufficient statistic vector

$$\left(\sum Y_i^2, t_1(Y^n), \ldots, t_d(Y^n)\right). \tag{2.4.14}$$

and mean-value parameter space $\mathtt{M}_q^{(\theta)} = (0, \infty) \times \mathbb{R}^d$. The original parameter vector corresponding to $(\lambda, \boldsymbol{\beta})$ is $(\sigma^2, \boldsymbol{\gamma})$ with $\sigma^2 := -1/(2\lambda)$ and $\boldsymbol{\gamma} = (\sigma^2\theta, \sigma^2\beta_1, \ldots, \sigma^2\beta_d)$ and the corresponding mean-value parameter vector is

$$\boldsymbol{\mu} := \left(n\sigma^2 + \sum \nu_i^2, \sum x_{i,1}\nu_i, \ldots, \sum x_{i,d}\nu_i\right)^T. \tag{2.4.15}$$

with ν_i as in (2.4.11). Observe that $\mathcal{H}_1 = \bigcup_{\theta \in \mathbb{R}\setminus\{0\}} \mathcal{Q}^{(\theta)}$ and $\mathcal{P} = \mathcal{Q}^{(0)}$. For expository convenience, we slightly deviated from our previous notation here by having a canonical parameter space vector of the form $(\lambda, \boldsymbol{\beta})$ rather than $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$; thus $\boldsymbol{\beta}$ is d-dimensional but $\boldsymbol{\mu}$ still represents a full $(d+1)$-dimensional mean-value parameter.

Having established that $\mathcal{Q}^{(\theta)}$ and $\mathcal{P}$ are, indeed, exponential families, we will now show that Theorem 1 in the form of Corollary 1 is applicable to them. Thus, fix arbitrary $Q^\circ \in \mathcal{H}_1$. We must have that $Q^\circ \in \mathcal{Q}^{(\theta^\circ)}$ for some θ° and the density of Q° can be written as $f^{(\theta^\circ)}_{\lambda^\circ, \boldsymbol{\beta}^\circ}$ or equivalently as $g_{\sigma^\circ, \boldsymbol{\gamma}^\circ}$ with σ°, $\boldsymbol{\gamma}^\circ$ and ν° related to $\theta^\circ, \lambda^\circ$ and $\boldsymbol{\beta}^\circ$ in the same way as before, in particular $\nu_i^\circ = \boldsymbol{\gamma}^{\circ T}\boldsymbol{x}_i$ (we can now see how this example extends Section 2.4.3: using the notation from that example, i.e. m the mean of U and s^2 its variance under Q, we set $n = 1$, $d = 0$, $\boldsymbol{x}_1 = 1$, $\nu_1^\circ = \gamma_0^\circ = m$ and $\sigma^{\circ 2} = s^2$).

Simple differentiation gives that the element in $\mathcal{P}$ that minimizes KL divergence, i.e. achieves $\min_{P \in \mathcal{P}} D(Q\|P)$, is given by $P = G_{\boldsymbol{\gamma}^*, \sigma^*}$ with parameters σ^{*2} and $\boldsymbol{\gamma}^* = (0, \gamma_1^*, \ldots, \gamma_d^*)$ where $\boldsymbol{\gamma}^*$ is a Euclidean projection and σ^{*2} is related to this projection via

$$\sigma^{*2} = \min_{(\gamma_1, \ldots, \gamma_d) \in \mathbb{R}^d} \frac{1}{n}\mathbf{E}_Q\left[\sum (Y_i - \sum_{j=1}^d \gamma_j \boldsymbol{x}_{i,j})^2\right]$$

This link to Euclidean projection implies, upon setting $\nu_i^* := \boldsymbol{\gamma}^{*T}\boldsymbol{x}_i$ the following easily derivable consequences:

$$\begin{aligned} &\text{for all } j \in \{1, \ldots, d\}\text{:} \sum \nu_i^\circ x_{i,j} = \sum \nu_i^* x_{i,j} \\ &\sigma^{*2} = \sigma^{\circ 2} + \frac{1}{n}\sum (\nu_i^* - \nu_i^\circ)^2 = \sigma^{\circ 2} + \frac{1}{n}\left(\sum \nu_i^{*2} - \sum \nu_i^{\circ 2}\right), \end{aligned} \tag{2.4.16}$$

where we note that (2.4.16) may be seen as versions of the standard *normal equations* in linear regression analysis. Again we define $\lambda^*, \boldsymbol{\beta}^*, \boldsymbol{\mu}^*$ correspondingly as above, in particular $\boldsymbol{\mu}^*$ is given in terms of σ^* and ν^* via (2.4.15) .

We now simply follow the steps needed to apply Theorem 1 in the form of Corollary 1. First, we reparameterize $\mathcal{P}$ in terms of the specific canonical parameterization in which $(\lambda, \boldsymbol{\beta}) = 0$ must correspond to $G_{\sigma^*, \boldsymbol{\gamma}^*}$. We obtain:

$$p_{\lambda, \boldsymbol{\beta}; \boldsymbol{\mu}^*}(y^n) = \frac{1}{Z_p(\lambda, \boldsymbol{\beta}; \boldsymbol{\mu}^*)} \cdot \exp\left(\lambda \sum y_i^2 + \sum_{j=1}^d \beta_j t_j(y^n)\right) f^{(0)}_{\lambda^*, \boldsymbol{\beta}^*}(y^n), \tag{2.4.17}$$

with $Z_p(\lambda, \boldsymbol{\beta}; \boldsymbol{\mu}^*)$ the normalizing constant, defined for all $(\lambda, \boldsymbol{\beta}) \in \mathtt{B}_{p;\boldsymbol{\mu}^*}$ where

$$\mathtt{B}_{p;\boldsymbol{\mu}^*} = \{(\lambda, \boldsymbol{\beta}) : Z_p(\lambda, \boldsymbol{\beta}; \boldsymbol{\mu}^*) < \infty\} = (-\infty, -\lambda^*) \times \mathbb{R}^d.$$

We see that this family coincides with $\mathcal{P}$. Similarly, relative to our fixed $(\theta^\circ, \lambda^\circ, \boldsymbol{\beta}^\circ)$ we define the family with densities

$$q^{(\theta^\circ)}_{\lambda,\boldsymbol{\beta}}(y^n; \boldsymbol{\mu}^*) = \frac{1}{Z^{(\theta^\circ)}_q(\lambda, \boldsymbol{\beta}; \boldsymbol{\mu}^*)} \cdot \exp\left(\lambda \sum y_i^2 + \sum_{j=1}^{d} \beta_j t_j(y^n)\right) f^{(\theta^\circ)}_{\lambda^\circ,\boldsymbol{\beta}^\circ}(y^n), \quad (2.4.18)$$

with normalizing constant $Z^{(\theta^\circ)}_q(\lambda, \boldsymbol{\beta}; \boldsymbol{\mu}^*)$. We see that this family coincides with $\mathcal{Q}^{(\theta^\circ)}$ and has canonical parameter space $\mathtt{B}^{(\theta^\circ)}_{q;\boldsymbol{\mu}^*} = (-\infty, -\lambda^\circ) \times \mathbb{R}^d$.

To apply Corollary 1, we need to verify that for each choice of θ°, we have (i) $\mathtt{M}^{(\theta^\circ)}_q \subseteq \mathtt{M}_p$, and (ii) for each $\boldsymbol{\mu} \in \mathtt{M}^{(\theta^\circ)}_q$, we have that $\mathtt{B}_{p;\boldsymbol{\mu}} \subseteq \mathtt{B}^{(\theta^\circ)}_{q;\boldsymbol{\mu}}$. We already verified $\mathtt{M}^{(\theta^\circ)}_q = \mathtt{M}_p$, implying (i), further above. As to (ii), note that the inclusion holds for $\boldsymbol{\mu} = \boldsymbol{\mu}^*$ since, using (2.4.16), $-\lambda^* = (1/2\sigma^{*2}) \leq (1/2\sigma^{\circ 2}) = -\lambda^\circ$. We next note that for each $\theta^\circ \in \mathbb{R}$, there is a 1-to-1 correspondence between the choice $(\lambda^\circ, \boldsymbol{\beta}^\circ) \in (-\infty, 0) \times \mathbb{R}^d$ used to determine $f^{(\theta^\circ)}_{\lambda^\circ,\boldsymbol{\beta}^\circ}$ and the resulting $\boldsymbol{\mu}^* \in \mathtt{M}_p$. We thus see that we can obtain the desired inclusion for arbitrarily chosen $\boldsymbol{\mu} \in \mathtt{M}_p$ by picking $(\lambda^\circ, \boldsymbol{\beta}^\circ)$ such that $\boldsymbol{\mu}^*$ becomes equal to this $\boldsymbol{\mu}$. This shows that (ii) holds for all $\boldsymbol{\mu} \in \mathtt{M}_p = \mathtt{M}^{(\theta^\circ)}_q$. Corollary 1 now gives the following: for all $\gamma^\circ \in \mathbb{R}^d$ with $\gamma^\circ_0 \neq 0$, all $\sigma^\circ > 0$, we have that $g_{\sigma^\circ,\gamma^\circ}(Y^n)/g_{\sigma^*,\gamma^*}(Y^n) = f^{(\theta^\circ)}_{\lambda^\circ,\boldsymbol{\beta}^\circ}(Y^n)/f^{(0)}_{\lambda^*,\boldsymbol{\beta}^*}(Y^n)$ is the GRO e-variable relative to $G_{\sigma^\circ,\gamma^\circ}$ if $\Sigma_p(\boldsymbol{\mu}) - \Sigma^{(\theta^\circ)}_q(\boldsymbol{\mu})$ is positive semidefinite for all $\boldsymbol{\mu} \in \mathbb{R}^d$. But this condition is readily established to hold: we do so in Appendix 2.A.

2.5 Proof of Theorem 1

To get some intuition first, we note that the distributions $P_{\boldsymbol{\beta}}$ and $Q_{\boldsymbol{\beta}}$ indexed by the $\boldsymbol{\beta}$ in the definition of $f(\boldsymbol{\beta}; \boldsymbol{\mu}^*)$, i.e. (2.2.1), are difficult to compare in the sense that they do not necessarily have any properties in common. In particular, $P_{\boldsymbol{\beta}}$ generally does not achieve $\min_{P \in \mathcal{P}} D(Q_{\boldsymbol{\beta}} \| P)$, so that $P_{\boldsymbol{\beta}}$ and $Q_{\boldsymbol{\beta}}$ do not have the same mean. This suggests to replace $f(\boldsymbol{\beta}; \boldsymbol{\mu}^*)$ by a function $g(\boldsymbol{\mu}; \boldsymbol{\mu}^*)$ on the mean-value parameter space and also to re-express $f(\boldsymbol{\beta}; \boldsymbol{\mu}^*) \leq 0$, the condition for being an e-variable, by a condition on g — and this is what we do in the proof of Theorem 1: inside the proof below we establish, using well-known convex duality properties of exponential families, that this can be done with function and condition, respectively, given by:

$$g(\boldsymbol{\mu}; \boldsymbol{\mu}^*) = D(P_{\boldsymbol{\mu}} \| P_{\boldsymbol{\mu}^*}) - D(Q_{\boldsymbol{\mu}} \| Q_{\boldsymbol{\mu}^*}), \quad (2.5.1)$$

$$g(\boldsymbol{\mu}; \boldsymbol{\mu}^*) \leq 0. \quad (2.5.2)$$

This condition on g corresponds to item 3 in Theorem 1. The key insight for showing the suitability of g is the following well-known convex-duality fact about exponential families: for all $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathtt{M}_p$, all $\boldsymbol{\beta} \in \mathtt{B}_{p;\boldsymbol{\mu}^*}$, we have:

$$-\log Z_p(\boldsymbol{\beta}; \boldsymbol{\mu}') = D(P_{\boldsymbol{\mu}_p(\boldsymbol{\beta};\boldsymbol{\mu}')} \| P_{\boldsymbol{\mu}'}) - \boldsymbol{\beta}^T \boldsymbol{\mu}_p(\boldsymbol{\beta}; \boldsymbol{\mu}') \leq D(P_{\boldsymbol{\mu}} \| P_{\boldsymbol{\mu}'}) - \boldsymbol{\beta}^T \boldsymbol{\mu}. \tag{2.5.3}$$

This can be derived as follows:

$$\begin{aligned} D(P_{\boldsymbol{\mu}_p(\boldsymbol{\beta};\boldsymbol{\mu}')} \| P_{\boldsymbol{\mu}'}) - D(P_{\boldsymbol{\mu}} \| P_{\boldsymbol{\mu}'}) &= \log \frac{Z_p(\beta_p(\boldsymbol{\mu}; \boldsymbol{\mu}'))}{Z_p(\boldsymbol{\beta}; \boldsymbol{\mu}')} + \boldsymbol{\beta}^T \boldsymbol{\mu}_p(\boldsymbol{\beta}; \boldsymbol{\mu}') - \boldsymbol{\beta}_p(\boldsymbol{\mu}; \boldsymbol{\mu}')\boldsymbol{\mu} \\ &= \log \frac{Z_p(\beta_p(\boldsymbol{\mu}; \boldsymbol{\mu}'))}{Z_p(\boldsymbol{\beta}; \boldsymbol{\mu}')} + \boldsymbol{\beta}^T (\boldsymbol{\mu}_p(\boldsymbol{\beta}; \boldsymbol{\mu}') - \boldsymbol{\mu}) - (\boldsymbol{\beta}_p(\boldsymbol{\mu}; \boldsymbol{\mu}') - \boldsymbol{\beta})^T \boldsymbol{\mu} \\ &= \boldsymbol{\beta}^T (\boldsymbol{\mu}_p(\boldsymbol{\beta}; \boldsymbol{\mu}') - \boldsymbol{\mu}) - D(P_{\boldsymbol{\mu}} \| P_{\boldsymbol{\mu}_p(\boldsymbol{\beta};\boldsymbol{\mu}')}) \\ &\leq \boldsymbol{\beta}^T (\boldsymbol{\mu}_p(\boldsymbol{\beta}; \boldsymbol{\mu}') - \boldsymbol{\mu}). \end{aligned}$$

We now prove the chain of implications in the theorem.

(1) ⇒ (2) Let $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathtt{M}_q$ and denote $\boldsymbol{\mu}(\alpha) := (1-\alpha)\boldsymbol{\mu}' + \alpha\boldsymbol{\mu}$. By assumption of convexity, we have that $\boldsymbol{\mu}(\alpha) \in \mathtt{M}_q$ for all $\alpha \in [0,1]$. Furthermore, define $h(\alpha) = (\boldsymbol{\beta}_p(\boldsymbol{\mu}(\alpha); \boldsymbol{\mu}') - \boldsymbol{\beta}_q(\boldsymbol{\mu}(\alpha); \boldsymbol{\mu}'))^T (\boldsymbol{\mu}(\alpha) - \boldsymbol{\mu}')$, so that $h(0) = 0$ and $h(1) = (\boldsymbol{\beta}_p(\boldsymbol{\mu}; \boldsymbol{\mu}') - \boldsymbol{\beta}_q(\boldsymbol{\mu}; \boldsymbol{\mu}'))^T (\boldsymbol{\mu} - \boldsymbol{\mu}')$. The derivative of h is given by

$$\begin{aligned} \frac{d}{d\alpha} h(\alpha) = &\left(\frac{d}{d\alpha} \boldsymbol{\beta}_p(\boldsymbol{\mu}(\alpha); \boldsymbol{\mu}') - \boldsymbol{\beta}_q(\boldsymbol{\mu}(\alpha); \boldsymbol{\mu}') \right)^T (\boldsymbol{\mu}(\alpha) - \boldsymbol{\mu}') \\ &+ (\boldsymbol{\beta}_p(\boldsymbol{\mu}(\alpha); \boldsymbol{\mu}') - \boldsymbol{\beta}_q(\boldsymbol{\mu}(\alpha); \boldsymbol{\mu}'))^T \frac{d}{d\alpha} \boldsymbol{\mu}(\alpha). \end{aligned}$$

The chain rule gives

$$\frac{d}{d\alpha} \boldsymbol{\beta}_p(\boldsymbol{\mu}(\alpha); \boldsymbol{\mu}') = \Sigma_p^{-1}(\boldsymbol{\mu}(\alpha))^T (\boldsymbol{\mu} - \boldsymbol{\mu}'),$$

where we use (2.3.1) and (2.3.2) together with the fact that the Jacobian of the gradient of a function equals the transpose of its Hessian. The derivative of $\boldsymbol{\beta}_q(\boldsymbol{\mu}(\alpha); \boldsymbol{\mu}')$ can be found with the same argument, so we see

$$\begin{aligned} \frac{d}{d\alpha} h(\alpha) &= \left((\Sigma_p^{-1}(\boldsymbol{\mu}(\alpha)) - \Sigma_q^{-1}(\boldsymbol{\mu}(\alpha)))^T (\boldsymbol{\mu} - \boldsymbol{\mu}') \right)^T (\boldsymbol{\mu}(\alpha) - \boldsymbol{\mu}') \\ &\quad + (\boldsymbol{\beta}_p(\boldsymbol{\mu}(\alpha); \boldsymbol{\mu}') - \boldsymbol{\beta}_q(\boldsymbol{\mu}(\alpha); \boldsymbol{\mu}'))^T (\boldsymbol{\mu} - \boldsymbol{\mu}') \\ &= \frac{1}{\alpha} (\boldsymbol{\mu}(\alpha) - \boldsymbol{\mu}')^T (\Sigma_p^{-1}(\boldsymbol{\mu}(\alpha)) - \Sigma_q^{-1}(\boldsymbol{\mu}(\alpha)))(\boldsymbol{\mu}(\alpha) - \boldsymbol{\mu}') \\ &\quad + (\boldsymbol{\beta}_p(\boldsymbol{\mu}(\alpha); \boldsymbol{\mu}') - \boldsymbol{\beta}_q(\boldsymbol{\mu}(\alpha); \boldsymbol{\mu}'))^T (\boldsymbol{\mu} - \boldsymbol{\mu}') \\ &= \frac{1}{\alpha} (\boldsymbol{\mu}(\alpha) - \boldsymbol{\mu}')^T (\Sigma_p^{-1}(\boldsymbol{\mu}(\alpha)) - \Sigma_q^{-1}(\boldsymbol{\mu}(\alpha)))(\boldsymbol{\mu}(\alpha) - \boldsymbol{\mu}') + \frac{1}{\alpha} h(\alpha). \end{aligned} \tag{2.5.4}$$

If $\Sigma_p(\boldsymbol{\mu}) - \Sigma_q(\boldsymbol{\mu})$ is positive semidefinite for all $\boldsymbol{\mu}$, then $\Sigma_p^{-1}(\boldsymbol{\mu}) - \Sigma_q^{-1}(\boldsymbol{\mu})$ is negative semidefinite (as discussed below the statement of Theorem 1). In this case, the first term in (2.5.4) is negative and, since $h(0) = 0$, the second term is also negative on $[0, 1]$. It follows that h is decreasing when $\Sigma_p(\boldsymbol{\mu}) - \Sigma_q(\boldsymbol{\mu})$ is positive semidefinite, so that $(\boldsymbol{\beta}_p(\boldsymbol{\mu}; \boldsymbol{\mu}') - \boldsymbol{\beta}_q(\boldsymbol{\mu}; \boldsymbol{\mu}'))^T(\boldsymbol{\mu} - \boldsymbol{\mu}')) \leq 0$, as was to be shown.

(2) ⇒ (3) We use a similar argument as was used to prove the previous implication, so let $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathtt{M}_q$ and denote $\boldsymbol{\mu}(\alpha) = (1-\alpha)\boldsymbol{\mu}' + \alpha\boldsymbol{\mu}$ as before. Define $h(\alpha) := g(\boldsymbol{\mu}(\alpha); \boldsymbol{\mu}')$. Using the chain rule of differentiation together with (2.3.1), we see that the derivative of h is given by

$$\begin{aligned}\frac{d}{d\alpha}h(\alpha) &= (\boldsymbol{\beta}_p(\boldsymbol{\mu}(\alpha); \boldsymbol{\mu}') - \boldsymbol{\beta}_q(\boldsymbol{\mu}(\alpha); \boldsymbol{\mu}'))^T(\boldsymbol{\mu} - \boldsymbol{\mu}') \\ &= \frac{1}{\alpha}(\boldsymbol{\beta}_p(\boldsymbol{\mu}(\alpha); \boldsymbol{\mu}') - \boldsymbol{\beta}_q(\boldsymbol{\mu}(\alpha); \boldsymbol{\mu}'))^T(\boldsymbol{\mu}(\alpha) - \boldsymbol{\mu}').\end{aligned}$$

If item (2) holds, then we have that $\frac{d}{d\alpha}h(\alpha) \leq 0$. Since $h(0) = 0$ and $h(1) = D(P_{\boldsymbol{\mu}}\|P_{\boldsymbol{\mu}'}) - D(Q_{\boldsymbol{\mu}}\|Q_{\boldsymbol{\mu}'})$, we see that item (2) implies that

$$D(P_{\boldsymbol{\mu}}\|P_{\boldsymbol{\mu}'}) - D(Q_{\boldsymbol{\mu}}\|Q_{\boldsymbol{\mu}'}) \leq 0,$$

as was to be shown.

(3) ⇒ (4) Assume that $D(P_{\boldsymbol{\mu}}\|P_{\boldsymbol{\mu}'}) - D(Q_{\boldsymbol{\mu}}\|Q_{\boldsymbol{\mu}'}) \leq 0$ for all $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathtt{M}_q$. Together with (2.5.3) this gives, for all $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathtt{M}_q$, all $\boldsymbol{\beta} \in \mathtt{B}_{p;\boldsymbol{\mu}'}$:

$$D(P_{\boldsymbol{\mu}_p(\boldsymbol{\beta};\boldsymbol{\mu}')}\|P_{\boldsymbol{\mu}'}) - \boldsymbol{\beta}^T\boldsymbol{\mu}_p(\boldsymbol{\beta}; \boldsymbol{\mu}') \leq D(P_{\boldsymbol{\mu}}\|P_{\boldsymbol{\mu}'}) - \boldsymbol{\beta}^T\boldsymbol{\mu} \leq D(Q_{\boldsymbol{\mu}}\|Q_{\boldsymbol{\mu}'}) - \boldsymbol{\beta}^T\boldsymbol{\mu}. \tag{2.5.5}$$

Applying this with $\boldsymbol{\mu} = \boldsymbol{\mu}_q(\boldsymbol{\beta}; \boldsymbol{\mu}')$ and re-arranging gives

$$-D(P_{\boldsymbol{\mu}_p(\boldsymbol{\beta};\boldsymbol{\mu}')}\|P_{\boldsymbol{\mu}'}) + \boldsymbol{\beta}^T\boldsymbol{\mu}_p(\boldsymbol{\beta}; \boldsymbol{\mu}') \geq -D(Q_{\boldsymbol{\mu}_q(\boldsymbol{\beta};\boldsymbol{\mu}')}\|Q_{\boldsymbol{\mu}'}) + \boldsymbol{\beta}^T\boldsymbol{\mu}_q(\boldsymbol{\beta}; \boldsymbol{\mu}'), \tag{2.5.6}$$

which, by the equality in key fact (2.5.3) is equivalent to $\log Z_p(\boldsymbol{\beta}; \boldsymbol{\mu}') \geq \log Z_q(\boldsymbol{\beta}; \boldsymbol{\mu}')$, which is what we had to prove.

Remaining Implications (4) ⇒ (5) now follows by the equality in (2.2.1) and the definition of an e-variable. (5) ⇒ (6) follows from proposition 1, (6) ⇒ (7) follows because a global e-variable is automatically also a local one, and (7) ⇒ (8) again follows from Proposition 1. Finally, (8) ⇒ (1) has already been established as Proposition 2. □

2.6 Conclusion and Future Work

We have provided a theorem that, under regularity pre-conditions, provides a general sufficient condition under which there exists a simple e-variable for testing a simple alternative versus a composite regular exponential family null. The characterization was given in terms of several equivalent conditions, the most direct being perhaps the condition '$\Sigma_p(\boldsymbol{\mu}) - \Sigma_q(\boldsymbol{\mu})$ is positive semidefinite for all $\boldsymbol{\mu} \in \mathtt{M}_q$'. A direct follow-up question is: can we construct GRO or close-to-GRO e-variables, in case either the regularity pre-conditions or the positive definiteness condition do *not* hold? The example of Section 2.4.3, and in particular Figure 2.3, indicated that in that case, many things can happen: under some $\boldsymbol{\mu} \in \mathtt{M}_q$ (green curve), $q_{\boldsymbol{\mu}}/p_{\boldsymbol{\mu}}$ still gives a global simple e-variable; for other $\boldsymbol{\mu}$ (blue), it gives a local but not global e-variable; for yet other $\boldsymbol{\mu}$ (pink), it does not give an e-variable at all.

Nevertheless, it turns out that if the pre-regularity conditions hold and the 'opposite' of the positive semidefinite condition holds, i.e. if $\Sigma_p(\boldsymbol{\mu}) - \Sigma_q(\boldsymbol{\mu})$ is *negative* semidefinite for all $\boldsymbol{\mu} \in \mathtt{M}_q$, then there is again sufficient structure to analyze the problem. The GRO e-variable will now be based on a mixture of elements of the null, but the specific mixture will depend on the sample size: we now need to look at i.i.d. repetitions of U rather than a single outcome U. We will provide such an analysis in future work.

Another interesting avenue for future work is to extend the analysis to *curved* exponential families [33]. While we do not have any general results in this direction yet, the analysis by [62] suggests that this may be possible. Liang [62] considers a variation of the Cochran-Mantel- Haenszel test, in which the null hypothesis expresses that the population-weighted *average* effect size over a given set of strata is equal to, or bounded by, some δ. This can be rephrased in terms of a curved exponential family null, for which [62] shows that a local e-variable exists by considering the second derivative of the function $f(\boldsymbol{\beta}; \boldsymbol{\mu}^*)$ as in (2.2.1), just like in the present paper but with $\boldsymbol{\beta}$ representing a particular suitable parameterization rather than the canonical parameterization of an exponential family. The local e-variable is then shown to be a global e-variable by a technique different from the construction of $\mathcal{Q}$ we use here. Still, the overall derivation is sufficiently similar to suggest that it can be unified with the reasoning underlying Theorem 1. Finally, the analysis of the linear model in Section 2.4.4 suggests that the results may perhaps be extended to say something about existence of *generalized* linear models without assuming a *model-X* condition [39] — a situation about which currently next to nothing is known.

Appendix 2.A Details for Section 2.4.4

We need to establish that $\Sigma_p(\boldsymbol{\mu}) - \Sigma_q^{(\theta^\circ)}(\boldsymbol{\mu}) = \Sigma_q^{(0)}(\boldsymbol{\mu}) - \Sigma_q^{(\theta^\circ)}(\boldsymbol{\mu})$ is positive semidefinite for all $\boldsymbol{\mu} \in \mathbb{R}^d$.

Thus, take any $\boldsymbol{\mu}^* \in \mathbb{R}^d$. By (2.4.15), we have that $q_{\boldsymbol{\mu}^*}^{(\theta^\circ)} = f_{\lambda^\circ, \boldsymbol{\beta}^\circ}^{(\theta^\circ)}$ and $p_{\boldsymbol{\mu}^*} = q_{\boldsymbol{\mu}^*}^{(0)} = f_{\lambda^*, \boldsymbol{\beta}^*}^{(0)}$ for some $\lambda^\circ, \boldsymbol{\beta}^\circ$ and $\lambda^*, \boldsymbol{\beta}^*$ that are related to each other via the normal equations (2.4.16). Based on the sufficient statistics (2.4.14), we can thus write, for $\theta \in \{0, \theta^\circ\}$, that

$$\Sigma_q^{(\theta)}(\boldsymbol{\mu}^*) = \begin{pmatrix} A^{(\theta)} & B^{(\theta)} \\ (B^{(\theta)})^T & C^{(\theta)} \end{pmatrix}$$

where $A^{(\theta^\circ)}$ is the variance of $\sum Y_i^2$ according to distribution $F_{\lambda^\circ, \boldsymbol{\beta}^\circ}^{(\theta^\circ)}$ and $C^{(\theta^\circ)}$ is the

$d \times d$ covariance matrix of the $t_j(Y^n)$ according to this distribution and

$$B^{(\theta^\circ)} = \left(\operatorname{cov}\left(\sum Y_i^2, t_1(Y^n)\right), \ldots, \operatorname{cov}\left(\sum Y_i^2, t_d(Y^n)\right)\right)$$

where the covariances are again under this distribution. Similarly, $A^{(0)}$ is the variance of $\sum Y_i^2$ according to distribution $F^{(0)}_{\lambda^*,\boldsymbol{\beta}^*}$ and $B^{(0)}$, $C^{(0)}$ are defined accordingly.

Positive semidefiniteness of $\Sigma_q^{(0)}(\boldsymbol{\mu}^*) - \Sigma_q^{(\theta^\circ)}(\boldsymbol{\mu}^*)$ is easily seen to be implied[1] if we can show that $C^{(0)} - C^{(\theta^\circ)}$ is positive definite and that

$$(A^{(0)} - A^{(\theta^\circ)}) - (B^{(0)} - B^{(\theta^\circ)})^T (C^{(0)} - C^{(\theta^\circ)})^{-1} (B^{(0)} - B^{(\theta^\circ)}) \geq 0. \qquad (2.A.1)$$

To show that $C^{(0)} - C^{(\theta^\circ)}$ is positive definite, note that $C^{(\theta^\circ)}$ (as is readily established, for example, by twice differentiating $\log Z_q^{(\theta^\circ)}(\lambda, \boldsymbol{\beta}; \boldsymbol{\mu}^*)$ at $\lambda = 0, \boldsymbol{\beta} = 0$) is simply the standard covariance matrix in linear regression scaled by $1/\sigma^{\circ 2}$, i.e. $C^{(\theta^\circ)} = \sigma^{\circ 2} \sum \mathbf{x}_i \mathbf{x}_i^T$ which by the maximal rank assumption is positive definite. Similarly $C^{(0)} = \sigma^{*2} \sum \mathbf{x}_i \mathbf{x}_i^T$ so that, since by assumption $\theta^\circ \neq 0$ and using the normal equations (2.4.16), we have that $C^{(0)} - C^{(\theta)} = cC^{(\theta)}$ for $c = \sigma^{*2} - \sigma^{\circ 2} > 0$ is also positive definite.

It only remains to show (2.A.1). As again easily established (for example, by twice differentiating $\log Z_q^{(\theta^\circ)}(\lambda, \boldsymbol{\beta}; \boldsymbol{\mu}^*)$ at $\lambda = 0, \boldsymbol{\beta} = 0$), we have that $A^{(\theta^\circ)} = 2\sigma^{\circ 2}\left(2(\sum \nu_i^{\circ 2}) + n\sigma^{\circ 2}\right)$ and similarly we find $A^{(0)} = 2\sigma^{*2}\left(2(\sum \nu_i^{*2}) + n\sigma^{*2}\right)$ and $B_j^{(\theta^\circ)} = -2\sigma^{\circ 2}\left(\sum \nu_i^\circ x_{i,j}\right)$ and similarly $B_j^{(0)} = -2\sigma^{*2}\left(\sum \nu_i^* x_{i,j}\right)$. By the normal equations (2.4.16) we find that $B_j^{(0)} - B_j^{(\theta^\circ)} = -2(\sigma^{*2} - \sigma^{\circ 2}) \sum \nu_i^* x_{i,j}$. After some matrix multiplications (where we may use the cyclic property of the trace of a matrix product) we get that (2.A.1) is equivalent to

$$(A^{(0)} - A^{(\theta)^\circ}) - 4(\sigma^{*2} - \sigma^{\circ 2}) \sum \nu_i^{*2} \geq 0.$$

But this is easily verified: it is equivalent to

$$2\sigma^{*2}\left(2\left(\sum \nu_i^{*2}\right) + n\sigma^{*2} - 2\left(\sum \nu_i^{*2}\right)\right) - 2\sigma^{\circ 2}\left(2\left(\sum \nu_i^{\circ 2}\right) + n\sigma^{\circ 2} - 2\left(\sum \nu_i^{*2}\right)\right) \geq 0$$

which in turn is equivalent to

$$2n\sigma^{*4} - 2n\sigma^{\circ 4} + 4(\sum \nu_i^{*2} - \sum \nu_i^{\circ 2})\sigma^{\circ 2} \geq 0$$

which by the normal equations is equivalent to

$$\sigma^{*4} - \sigma^{\circ 4} + 2(\sigma^{*2} - \sigma^{\circ 2})\sigma^{\circ 2} \geq 0$$

but this must be the case since by the normal equations, $\sigma^{*2} > \sigma^{\circ 2}$.

[1] For an explicit derivation see `https://math.stackexchange.com/questions/2280671/definiteness-of-a-general-partitioned-matrix-mathbf-m-left-beginmatrix-bf`.

Chapter 3

E-Values for Exponential Families: the General Case

Abstract

We analyze common types of e-variables and e-processes for composite exponential family nulls: the optimal e-variable based on the reverse information projection (RIPr), the conditional (COND) e-variable, and the universal inference (UI) and sequentialized RIPr e-processes. We characterize the RIPr prior for simple and Bayes-mixture based alternatives, either precisely (for Gaussian nulls and alternatives) or in an approximate sense (general exponential families). We provide conditions under which the RIPr e-variable is (again exactly vs. asymptotically) equal to the COND e-variable. Based on these and other interrelations which we establish, we determine the e-power of the four e-statistics as a function of sample size, exactly for Gaussian and up to $o(1)$ in general. For d-dimensional null and alternative, the e-power of UI tends to be smaller by a term of $(d/2) \log n + O(1)$ than that of the COND e-variable, which is the clear winner.

3.1 Introduction

Interest in e-values — a term coined only in 2019 — has exploded in recent years. Key publications include [95, 42, 91, 76]; see the introduction [71] for many more references. E-values are the values taken on by e-variables and e-processes. E-variables allow for effortless null hypothesis testing under optional continuation — combining data from different studies when the decision to perform the study may depend in unknown ways on past data. E-processes additionally allow for optional stopping within a study.

In this paper we consider various types of e-variables and -processes for multivariate exponential family null hypotheses, given in their mean-value parameter space $\mathtt{M}_p$ as

[0] This chapter is based on Yunda Hao, and Peter Grünwald. E-Values for Exponential Families: the General Case. arXiv preprint arXiv:2409.11134, 2024, under submission.

$\mathcal{P} = \{P_{\boldsymbol{\mu}} : \boldsymbol{\mu} \in \mathtt{M}_p\}$. We investigate and compare their *e-power* [94] for i.i.d. data $U_{(1)}, U_{(2)}, \ldots$. Recall that an e-variable for sample size n is any nonnegative statistic $S^{(n)} = S^{(n)}(U^{(n)})$ of the data $U^{(n)} = (U_{(1)}, \ldots, U_{(n)})$ such that, under all $P \in \mathcal{P}$, we have $\mathbb{E}_P[S^{(n)}] \leq 1$. For a simple alternative $\mathcal{Q} = \{Q\}$, expressing $U_{(1)}, U_{(2)}, \ldots \sim$ i.i.d. Q, the e-power of e-variable $S^{(n)}$ is given by

$$\mathbb{E}_Q[\log S^{(n)}]. \tag{3.1.1}$$

A central role is played by the optimal e-variable (known variously as GRO (growth rate optimal) or *numéraire*) that maximizes e-power over all e-variables that can be defined on $U^{(n)}$. A (perhaps *the*) central result in e-value theory [42, 57, 58] says that it can be calculated based on the *reverse information projection (RIPr)*. We shall denote this optimal e-variable as S_{RIP}. We also consider a sequentialized version of the RIPr e-variable, $S_{\text{SEQ-RIP}}$ that is 'locally' (outcome-wise) but not 'globally' (sample-wise) optimal [88, 83, 44]. We further look at e-variable S_{COND} that is based on *conditioning* on the sufficient statistic and a popular version of the *universal inference* e-variable S_{UI} [95]. Instantiating such e-variables requires specifying an alternative. We first consider a simple alternative $\mathcal{Q} = \{Q\}$ as in (3.1.1). In this case, under mild further conditions (see Section 3.1.2), the optimal e-variable for a sample of size n can be written as

$$S_{\text{RIP}}^{(n)} = \frac{q(U^{(n)})}{\int p_{\boldsymbol{\mu}}(U^{(n)}) dW(\boldsymbol{\mu})}, \tag{3.1.2}$$

where q is the density of Q, $p_{\boldsymbol{\mu}}$ the density of $P_{\boldsymbol{\mu}}$ and W a prior on the parameter space $\mathtt{M}_p$ of $\mathcal{P}$. In a 'companion' paper [40], we showed that under some conditions on the covariance matrices corresponding to Q and the $P_{\boldsymbol{\mu}} \in \mathcal{P}$, the RIPr is achieved by a degenerate prior, putting all mass on a single $P_{\boldsymbol{\mu}} \in \mathcal{P}$. We refer to this as the *simple* case, since in this case, a simple-vs.-simple likelihood ratio provides the optimal e-variable, even though the null is composite. Companion paper [40] (which has 'simple' in the title), only considers (a) the GRO e-variable for (b) the simple case at (c) $n = 1$; the present paper (with 'general' in the title) considers (a) various types of e-variables, in (b) general settings with (c) n varying.

Main Results A major finding of *this* paper is that, if the 'opposite' condition holds (we refer to this as the *anti-simple case*) then the RIPr prior W in (3.1.2) is a Gaussian with variance $O(1/n)$, in an approximate asymptotic sense in general, and even exactly if $\mathcal{P}$ is a Gaussian location family and Q is also Gaussian. This is the first time, as far as we know, that insight into a nondegenerate RIPr prior is obtained for the case of a parametric, non-convex null.

This finding is enabled by our main *theoretical* insight: the conditional e-variable S_{COND} can be analyzed via a local central limit theorem with explicit bounds on the error terms [16]. As a result, we not only get explicit $o(1)$ bounds on its e-power, but we also find that it is closely related to S_{RIP} (in the Gaussian anti-simple case they even coincide), leading to the result above and to explicit expressions for e-power (3.1.1). We also develop these for our other types of e-variables, not just under the 'true' alternative Q, but also in the *misspecified case* when the data are sampled i.i.d. from a distribution $R \neq Q$. In Section 3.2, Theorem 2 provides exact expressions for the case that $\mathcal{P}$ is a multivariate Gaussian location family and Q a Gaussian with a different covariance matrix. In Section 3.3, Theorem 4 we consider general exponential family nulls and Q that satisfy a regularity condition. Here our results hold up to $o(1)$. We also extend our results to a specific composite alternative, namely, a second exponential family with the same sufficient statistic as the null. Many practical parametric testing problems are of this form, as recalled in Example 4 and 6 below.

There are two standard methods to design e-variables for composite $\mathcal{Q}$: the sequential plug-in method and the method of mixtures [71]. We consider both, resulting in an (again exact) extension of Theorem 2 to Theorem 3 for composite null and alternative Gaussians (Section 3.2), and an extension of Theorem 4 to Theorem 5 (Section 3.3) for general exponential family nulls and alternatives, with results holding (mostly) up to $o(1)$. As another central contribution, we find that, when using the method of mixtures, equipping the alternative with a prior W_1, then, under regularity conditions, the RIPr prior W in (3.1.2) is, in an approximate sense, given by the *same* prior W_1, irrespective of whether we are in the simple case or not.

Additional theoretical insights regarding these results include that, in the Gaussian case of Theorem 3, the e-variable S_{COND} coincides with the e-variable obtained by equipping $\mathcal{P}$ and $\mathcal{Q}$ with the improper right Haar prior, that is suggested by Pérez-Ortiz et al. [70] (Section 3.2.2, (3.2.25)); and that the conditions needed for well-behavedness of the plug-in method and universal inference are 'dual' to each other: compare Condition 3, Section 3.3.3 with Condition 2, Section 3.3.2.

Summarizing (and for simplicity leaving some $O(1)$ terms unspecified) some of our main findings for the — practically more relevant — composite case, we obtain the following relations. These are obtained from Corollary 3 (Section 3.2.2) and Corollary 5 (Section 3.3.3). Here we used particular, convenient versions of the plug-in and mixture method to deal with composite $\mathcal{Q}$, with precise definitions (including the definition of 'strict', 'simple', 'anti-simple' and d_{qp}) in Section 3.2 and 3.3. We find that, under appropriate (yet mild) regularity conditions on $\mathcal{P}$ and $\mathcal{Q}$, that for all $Q \in \mathcal{Q}$,

$$\mathbb{E}_Q\left[\log \frac{S^{(n)}_{\text{RIP}}}{S^{(n)}_{\text{COND}}}\right] = o(1). \tag{3.1.3}$$

$$\mathbb{E}_Q\left[\log \frac{S^{(n)}_{\text{COND}}}{S^{(n)}_{\text{UI}}}\right] = \frac{d}{2}\log n + O(1). \tag{3.1.4}$$

$$\mathbb{E}_Q\left[\log \frac{S^{(n)}_{\text{SEQ-RIP}}}{S^{(n)}_{\text{UI}}}\right] = \frac{d_{qp}}{2}\log n + O(1) \text{ with } 0 < d_{qp} < d, \text{ in the strict simple case.} \tag{3.1.5}$$

$$\mathbb{E}_Q\left[\log \frac{S^{(n)}_{\text{SEQ-RIP}}}{S^{(n)}_{\text{COND}}}\right] \leq -n\epsilon \text{ for some } \epsilon > 0, \text{ all large } n, \text{ in the strict anti-simple case.} \tag{3.1.6}$$

where d is the dimensionality of the exponential family and d_{qp} is a notion of 'effective dimension' whose exact size depends on Q. Note in particular that $S_{\text{SEQ-RIP}}$ is not competitive in the anti-simple case. The resemblance of (3.1.3) to the ubiquitous BIC model selection criterion is no coincidence: both are derived via a Laplace approximation of a Bayesian marginal likelihood. The occurrence of $d_{qp} < d$ arises due to the use of plug-in methods — it also appears in the results on *prequential* model selection by [38, 54], whose techniques for analyzing log-loss of sequential plug-in estimators we employ and (vastly) generalize. Given that, in contrast to S_{RIP}, we always know how to calculate S_{COND}, it appears that S_{COND} is in some sense the clear winner — a fact that came as a surprise to us, especially since its definition requires neither a prior on the alternative (as in the method of mixtures) nor an estimator (as in the plug-in method): in a sense, it's not "learning"! However, S_{COND} does carry one big disadvantage: in contrast to S_{UI} and $S_{\text{SEQ-RIP}}$, usually (i.e. for most exponentially family nulls), S_{COND} does not define an e-process. We explain this rule of thumb, discuss exceptions and derive conditions under which S_{RIP} does not define an e-process in Section 3.6. Finally, we emphasize that while some of our results on e-power are asymptotic, the four types of e-variables we employ, ad hence the Type-I error guarantees they lead to, are invariably nonasymptotic, i.e. valid at each sample size.

Contents In the remainder of this introductory section, we first (Section 3.1.1) provide preliminaries on exponential families. We then (Section 3.1.2) define the various types of e-variables we consider. In Section 3.2, we provide our theorems for multivariate Gaussians, in Section 3.3, we provide the analogous results for general exponential family nulls. Section 3.4 provides the proofs of the Gaussian results, Section 3.5 provides the proofs of the general exponential family results — these are given in terms of various lemmas that are interesting in their own right and whose (relatively basic but very tedious) proofs are delegated to appendices. Finally, Section 3.6 provides the implications for *e-process-ness* and discusses potential future work.

3.1.1 Preliminaries on Exponential Families, Notation and KL Divergence

In all our results, the null hypothesis $\mathcal{P}$ is a regular d-dimensional exponential family defined on underlying random element U taking values in some set $\mathcal{U}$ and sufficient statistic vector $X = (X_1, \ldots, X_d)^\top$. We can write $X_j = t_j(U)$ for given functions $t_1, \ldots, t_d$. For the Gaussian case, Section 3.2, we can take $X = U$, but in general (see e.g. Example 4), distinguishing between X and U is crucial. Here and in the sequel we will freely use standard properties of exponential families without explicitly referring to their proofs, for which we refer to, e.g. [19, 13, 33]. We parameterize $\mathcal{P}$ in terms of the mean-value parameter space $\mathtt{M}_p$, so that we can write

$$\mathcal{P} = \{P_{\boldsymbol{\mu}} : \boldsymbol{\mu} \in \mathtt{M}_p\}$$

with $\mathtt{M}_p \subset \mathbb{R}^d$ and $\mathbb{E}_{P_{\boldsymbol{\mu}}}[X] = \boldsymbol{\mu}$. We denote the $d \times d$ covariance matrix of X under $P_{\boldsymbol{\mu}}$ as $\Sigma_p(\boldsymbol{\mu})$. Recall that $\Sigma_p(\boldsymbol{\mu})$ is continuous and positive definite for all $\boldsymbol{\mu} \in \mathtt{M}_p$, and, since the family is regular, $\mathtt{M}_p$ is a convex open set.

All elements of such a $\mathcal{P}$ have densities relative to some underlying measure ν. We fix such a ν and denote the density of U under $P_{\boldsymbol{\mu}}$ as $p_{\boldsymbol{\mu}}$.

As to canonical parameterizations of $\mathcal{P}$, which we will only use in Section 3.3.1, we note that we can take any $\boldsymbol{\mu} \in \mathtt{M}_p$ and define

$$p^{\text{CAN}}_{\boldsymbol{\beta}}(U) = \frac{1}{Z_p(\boldsymbol{\beta})} \cdot e^{\sum_{j=1}^{d} \beta_j t_j(U)} \cdot p_{\boldsymbol{\mu}}(U), \tag{3.1.7}$$

where $Z_p(\boldsymbol{\beta})$ is the normalizing constant and we define the canonical parameter space $\mathtt{B}_p = \{\boldsymbol{\beta} : Z_p(\boldsymbol{\beta}) < \infty\}$. As is well known, the set of distributions $\{P^{\text{CAN}}_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \mathtt{B}_p\}$ where $P^{\text{CAN}}_{\boldsymbol{\beta}}$ has density $p^{\text{CAN}}_{\boldsymbol{\beta}}$, coincides with $\mathcal{P}$.

$\mathcal{P}$ is extended to multiple outcomes by the i.i.d. assumption. It thus becomes a set of distributions for random process $U_{(1)}, U_{(2)}, \ldots$, with for $i \geq 1$, $U_{(i)}$ an i.i.d. copy of U, and $X_{(i)} = (X_{(i),1}, \ldots, X_{(i),d})$ with $X_{(i),j} = t_j(U_{(i)})$. We abbreviate $X^{(n)} = (X_{(1)}, \ldots, X_{(n)})$ and similarly $U^{(n)} = (U_{(1)}, \ldots, U_{(n)})$ and write $P_{\boldsymbol{\mu}}(U^{(n)})$ (density $p_{\boldsymbol{\mu}}(U^{(n)})$) for the marginal distribution of $U^{(n)}$ under $P_{\boldsymbol{\mu}}$. We use the notation $\mathcal{P}$ to refer both to the set of distributions for a single outcome U and for the random processes $(U_{(i)})_{i \in \mathbb{N}}$, as in each instance it will be clear what is meant. When referring to the set of marginal distributions for the first n outcomes, we use the notation $\mathcal{P}(U^{(n)}) := \{P_{\boldsymbol{\mu}}(U^{(n)}) : \boldsymbol{\mu} \in \mathtt{M}_p\}$.

Since we assume $\mathcal{P}$ to be regular, the maximum likelihood estimator (MLE) in the mean-value parameterization, $\hat{\boldsymbol{\mu}}_{|n}$, based on data $U^{(n)}$ exists, is unique and equal to $n^{-1} \sum_{i=1}^{n} X_{(i)}$ whenever the latter quantity lies in the set $\mathtt{M}_p$. With slight abuse of notation, we shall extend the definition of $\hat{\boldsymbol{\mu}}_{|n}$ and simply set it to be equal to $n^{-1} \sum_{i=1}^{n} X_{(i)}$ even if the latter quantity is not contained in $\mathtt{M}_p$; this can happen, for example, with the Bernoulli distribution if all $X_{(i)}$ are equal to 1, or all are equal to 0. We then set $p_{\hat{\boldsymbol{\mu}}_{|n}}(U^{(n)}) := \sup_{\boldsymbol{\mu} \in \mathtt{M}_p} p_{\boldsymbol{\mu}}(U^{(n)})$, which will make all quantities appearing in our results well-defined.

The *alternative* hypothesis $\mathcal{Q}$ will either (Section 3.1.2, Theorem 2 and 4) be a singleton $\mathcal{Q} = \{Q\}$ or (Theorem 3 and 5) will itself be another regular exponential family, defined on $\mathcal{U}$, with the same sufficient statistic X as $\mathcal{P}$. In the latter case, $\mathcal{Q}$ is invariably extended to sequences of outcomes by the i.i.d. assumption. In both cases we assume that all elements of $\mathcal{Q}$ have a density relative to ν. For the case that $\mathcal{Q}$ is an exponential family, we extend all notation in the obvious way: $\mathtt{M}_q$ denotes the mean-value parameter space, $\mathtt{B}_q$ is a canonical parameter space, $\Sigma_q(\boldsymbol{\mu})$ is the $d \times d$ covariance matrix corresponding to $Q_{\boldsymbol{\mu}}$, and so on.

The *KL (Kullback-Leibler) divergence* $D(\cdot \| \cdot)$ plays a central role in our analysis. For R, P distributions for i.i.d. random process $U_{(1)}, U_{(2)}, \ldots$ as above, we write $D(R(V) \| P(V))$ to denote the KL divergence between the Q- and P-marginal distributions for random vector V, respectively; for example, $D(R(U^{(n)}) \| P(U^{(n)}))$. Whenever we write $D(R \| P)$, this is meant to abbreviate $D(R(U) \| P(U))$.

3.1.2 Preliminaries on E-Variables: The E-Zoo

An *e-variable* $S^{(n)}$ for sample size n is a statistic of data $U^{(n)}$ (i.e. a random variable that can be written as a measurable function of $U^{(n)}$) that is (a) nonnegative, and (b), that satisfies, for all $P \in \mathcal{P}$, $\mathbb{E}_{U^{(n)} \sim P}[S^{(n)}] \leq 1$.

To define an e-process (a concept only analyzed in Section 3.6), suppose that $\mathcal{U} = \mathbb{R}^d$ for some $d > 0$ and let $\sigma(U^{(n)})$ be the σ-algebra generated by $U_{(1)}, \ldots, U_{(n)}$. We call $(\sigma(U^{(n)}))_{n \in \mathbb{N}}$ the *filtration induced by the data* or 'data filtration' for short. An *e-process relative to the data filtration* is a random process $(S^{(n)})_{n \in \mathbb{N}}$ defined relative to the data filtration such that for each stopping time τ (defined again relative to the data filtration), $S^{(\tau)}$ is an e-variable. We refer to [71] for more background on e-processes.

For parametric testing problems such as ours, there exist several standard ways to define e-variables and e-processes. These invariably take as their starting point a single distribution Q for random process $U_{(1)}, U_{(2)}, \ldots$. In basic cases the alternative $\mathcal{Q}$ is simple, and then we may think of Q as representing the alternative, i.e. $\mathcal{Q} = \{Q\}$. If $\mathcal{Q}$ is not simple, then Q is usually taken to be a Bayes marginal distribution or 'plug-in' distribution chosen so as to represent $\mathcal{Q}$ [71]. In both cases, the data will not be i.i.d. under Q even if they are under all elements of $\mathcal{Q}$: therefore, for now we let Q be an arbitrary distribution for random process $U_{(1)}, U_{(2)}$ such that its conditional densities $q(x_{(1)}), q(x_{(2)} \mid x_{(1)}), \ldots$ relative to the chosen background measure are well-defined.

[42] define the GRO (growth-rate optimal) e-variable for a sample of size n, relative to Q, to be the e-variable $S^{(n)}$ that, among all e-variables that can be written as a function of data $U^{(n)}$, maximizes *growth-rate* (3.1.1), also known as *e-power* [94], $\mathbb{E}_{U^{(n)} \sim Q}[\log S^{(n)}]$, and show that it is given by:

$$S^{(n)}_{Q,\text{RIP}} = \frac{q(U^{(n)})}{p_{\leftsquigarrow q(U^{(n)})}(U^{(n)})}, \tag{3.1.8}$$

where $p_{\leftsquigarrow q(U^{(n)})}$ is the *reverse information projection (RIPr)* of $Q(U^{(n)})$ on the set $\text{CONV}(\mathcal{P}(U^{(n)}))$ where CONV denotes the convex hull. More precisely, [42, Theorem 1] implies that the RIPr exists, and (3.1.8) gives the growth-optimal e-variable, whenever $D(Q \| P_{\boldsymbol{\mu}}) < \infty$ for all $\boldsymbol{\mu} \in \mathtt{M}_p$ (this result has later been generalized [57, 58] but for us the initial, simple version suffices). The RIPr often finds itself in the *Choquet convex hull* of $\mathcal{P}$ and in that case, it can be written as a Bayes marginal distribution: there then is a prior W on $\mathtt{M}_p$ such that $p_{\leftsquigarrow q(U^{(n)})}(U^{(n)}) = p_W(U^{(n)})$. Here we defined

$$p_W(U^{(n)}) = \int p_{\boldsymbol{\mu}}(U^{(n)}) dW(\boldsymbol{\mu})$$

to be the Bayes marginal distribution based on W. Whenever below we write 'optimal', we mean 'optimal in terms of e-power'. While in some nice cases (the 'simple' case of Section 3.3.1 below) calculating this optimal $S^{(n)}_{Q,\text{RIP}}$ is easy, in general, it can be quite difficult to compute. In that case we may prefer any of the following alternatives.

First, by conditioning on the sufficient statistic, we obtain an alternative, potentially sub-optimal e-variable that is always well-defined:

$$S^{(n)}_{Q,\text{COND}} = \frac{q(U^{(n)} \mid Z)}{p(U^{(n)} \mid Z)}, \tag{3.1.9}$$

where $Z = \sum_{i=1}^{n}(X_{(i),1}, \ldots, X_{(i),d})^\top$ and p is the density of the conditional distribution of $U^{(n)}$ given Z, which is identical for all $P \in \mathcal{P}$ because Z is a sufficient statistic and the family is regular. To see that $S^{(n)}_{Q,\text{COND}}$ indeed is an e-variable, note that for all $P \in \mathcal{P}$,

$$\begin{aligned}\mathbb{E}_P[S^{(n)}_{Q,\text{COND}}] &= \mathbb{E}_{Z\sim P}[\mathbb{E}_{U^{(n)}\sim P}[S^{(n)}_{Q,\text{COND}}|Z]] = \mathbb{E}_{Z\sim P}\left[\int p(u^{(n)} \mid Z) \cdot \frac{q(u^{(n)} \mid Z)}{p(u^{(n)} \mid Z)} \mathrm{d}\nu(u^{(n)})\right] \\ &= \mathbb{E}_{Z\sim P}[1] = 1.\end{aligned}$$

Although there are important exceptions (Section 3.6), in general neither the sequences $(S^{(n)}_{Q,\text{RIP}})_{n\in\mathbb{N}}$ nor $(S^{(n)}_{Q,\text{COND}})_{n\in\mathbb{N}}$ define an e-process. Here are two standard ways to define such an e-process for testing $\mathcal{Q} = \{Q\}$ vs. $\mathcal{P}$:

$$S^{(n)}_{Q,\text{SEQ-RIP}} = \prod_{i=1}^{n} \frac{q(U_{(i)} \mid U^{(i-1)})}{p_{\leftsquigarrow q(U_{(i)}|U^{(i-1)})}(U_{(i)})} \quad ; \quad S^{(n)}_{Q,\text{UI}} = \frac{q(U^{(n)})}{\sup_{\boldsymbol{\mu}\in\mathtt{M}_p} p_{\boldsymbol{\mu}}(U^{(n)})} = \frac{q(U^{(n)})}{p_{\hat{\boldsymbol{\mu}}_{|n}}(U^{(n)})}, \tag{3.1.10}$$

where $S^{(n)}_{Q,\text{SEQ-RIP}}$ is well-defined whenever all involved conditional densities are well-defined and the final equality holds whenever $p_{\hat{\boldsymbol{\mu}}_{|n}}$ is well-defined. Here $S^{(n)}_{Q,\text{SEQ-RIP}}$ is the sequential or 'local' (in time) version of (3.1.8): for each $U^{(i-1)} \in \mathcal{U}^{i-1}$, $p_{\leftsquigarrow q(U_{(i)}|u^{(i-1)})}$ is the reverse information projection of $Q(U_{(i)} \mid U^{(i-1)} = u^{(i-1)})$ on the set $\text{CONV}(\mathcal{P}(U_{(i)}))$, i.e. the convex hull of $\mathcal{P}$ restricted to a single outcome. Finally, $S_{Q,\text{UI}}$ is one possible (and in fact quite standard) instantiation of the *universal inference* method pioneered by [95].

All e-variables and processes mentioned so far are important in practice and have been studied theoretically for some specific null hypotheses $\mathcal{P}$ [71]. In particular, [70] employs RIPr e-variables; sequential RIPr e-variables are implicitly employed ubiquitously in nonparametric settings [71], and, in parametric settings, by [88, 83, 44]. UI appears in [95, 32, 85]. $S_{Q,\text{COND}}$ has been implicitly used a lot as a likelihood ratio within the likelihoodist paradigm, reducing a composite null to a simple one by conditioning on a sufficient statistic, e.g. for contingency tables; see e.g. [74, Chapter 7]. $S_{Q,\text{COND}}$ is in general only available for exponentially family nulls, all other e-variables and -processes mentioned here can be applied more generally.

By the growth optimality of $S^{(n)}_{Q,\text{RIP}}$ we have, for general Q,

$$\mathbb{E}_Q[\log S^{(n)}_{Q,\text{RIP}}] \geq \max\left\{\mathbb{E}_Q[\log S^{(n)}_{Q,\text{SEQ-RIP}}], \mathbb{E}_Q[\log S^{(n)}_{Q,\text{UI}}], \mathbb{E}_Q[\log S^{(n)}_{Q,\text{COND}}]\right\}. \tag{3.1.11}$$

Our aim in this paper is to quantify such relations in more — essentially full —

detail.

3.2 The Gaussian Location Family

In the present section we consider the special case that under the null hypothesis, $X = U$ has a multivariate Gaussian distribution with fixed covariance matrix Σ_p and the alternative is also Gaussian or a set of Gaussians, with some other fixed covariance matrix Σ_q.

3.2.1 $\mathcal{Q}$ simple, $\mathcal{P}$ multivariate Gaussian location

Let $\mathcal{P}$ be the Gaussian location family for $X = U = (X_1, \dots, X_d)$ with nondegenerate $d \times d$ covariance matrix Σ_p. Fix a particular mean vector $\boldsymbol{\mu}^* \in \mathtt{M}_p = \mathbb{R}^d$. We let $\mathcal{Q} = \{Q\}$ with Q a Gaussian for X with mean $\boldsymbol{\mu}^*$ but with nondegenerate covariance matrix $\Sigma_q \neq \Sigma_p$. For invertible $d \times d$ matrix B, we let

$$D_{\text{GAUSS}}(B) := \frac{1}{2} \left(- \log \det(B) - (d - \text{TR}(B)) \right), \tag{3.2.1}$$

where $\text{TR}(B)$ is the trace of B and $\det(B)$ is the determinant of B. The subscript derives from the fact that $D_{\text{GAUSS}}(\Sigma_q \Sigma_p^{-1})$ is the KL divergence between two d-dimensional Gaussians that share the same (arbitrary) mean vector and have covariances Σ_q and Σ_p respectively — which also tells us that for general positive definite and symmetric Σ_q, Σ_p, $D_{\text{GAUSS}}(\Sigma_q \Sigma_p^{-1}) \geq 0$ with equality iff $\Sigma_q = \Sigma_p$. The following characterization, derived from standard properties of determinant and trace, will prove useful below: letting $\lambda_1, \dots, \lambda_d$ be the eigenvalues of $\Sigma_p^{-1/2} \Sigma_q \Sigma_p^{-1/2}$, we have

$$D_{\text{GAUSS}}(\Sigma_q \Sigma_p^{-1}) = D_{\text{GAUSS}}(\Sigma_p^{-1/2} \Sigma_q \Sigma_p^{-1/2}) = \frac{1}{2} \left(\sum_{j=1}^{d} (- \log \lambda_j - (1 - \lambda_j)) \right). \tag{3.2.2}$$

It will be most helpful for stating our results to introduce the following nonstandard notation: for positive definite $d \times d$ matrices Σ_a, Σ_b, we define $d_{ab} := \text{TR}(\Sigma_a \Sigma_b^{-1})$. In particular,

$$d_{qq} = d_{pp} = d, 0 < d_{qp} \text{ and if } \Sigma_q - \Sigma_p \text{ is negative semidefinite, then } d_{qp} \leq d, \tag{3.2.3}$$

the latter inequality becoming strict if $\Sigma_q - \Sigma_p$ is negative definite. To derive the inequality, note that if $\Sigma_q - \Sigma_p$ is negative definite, then $\Sigma_p^{-1/2} \Sigma_q \Sigma_p^{-1/2} - I$ is negative definite, and then all eigenvalues λ_j above are smaller than 1, so $d_{qp} = \sum_{j=1}^{d} \lambda_j$ is smaller than d.

We now present our characterization of e-values and their growth rate for this setting. It turns out that the proof technique remains usable if we replace the sampling distribution Q by any other sampling distribution R with the same mean $\boldsymbol{\mu}^*$, so we will state and prove our results for such general R. We shall refer to the case $R \neq Q$

as the *misspecified case*, in contrast to the *well-specified* case $R = Q$. Allowing $R \neq Q$ mainly adds strength to Theorem 3 for composite alternatives further on; in Theorem 2 directly below, if a statistician employs S_Q it implies she has knowledge of $\boldsymbol{\mu}^*$, so the added generality of allowing general R with the same $\boldsymbol{\mu}^*$ is somewhat limited.

For compactness in notation, we further introduce (as was already done by [55]) the generalized KL divergence between Q and P under sampling distribution R (also assumed to have a density relative to ν) as

$$D_R(Q\|P) := \mathbf{E}_{U\sim R}\left[\log \frac{q(U)}{p(U)}\right] \overset{(a)}{=} D(R\|P) - D(R\|Q),$$

where (a) holds whenever either $D(R\|P)$ or $D(R\|Q)$ is finite. Yet $D_R(Q\|P)$ is still well-defined in some cases in which both $D(R\|P)$ or $D(R\|Q)$ are ∞ [57].

As to the conditional e-variable (3.1.9), in this context it becomes

$$S^{(n)}_{\text{COND}} := \frac{q(U^{(n)} \mid Z)}{p(U^{(n)} \mid Z)}, \tag{3.2.4}$$

where we remark that, by sufficiency, the conditional distributions of $U^{(n)}$ given $Z = \sum_{i=1}^n X_{(i)}$ are equal to each other for all $Q \in \mathcal{Q}$ (i.e. $q_{\boldsymbol{\mu}_1}(U^{(n)} \mid Z) = q_{\boldsymbol{\mu}_2}(U^{(n)} \mid Z)$ for all $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathtt{M}_q$) and similarly for $P \in \mathcal{P}$, hence (3.2.4) is well-defined. Finally, we define:

$$D_{\Sigma_r}(\Sigma_q\|\Sigma_p) := D_{\text{GAUSS}}(\Sigma_r\Sigma_p^{-1}) - D_{\text{GAUSS}}(\Sigma_r\Sigma_q^{-1}) = -\frac{1}{2}\log\det(\Sigma_q\Sigma_p^{-1}) + \frac{d_{rp} - d_{rq}}{2} \tag{3.2.5}$$

as a generalization (in the sense that $D_{\Sigma_q}(\Sigma_q\|\Sigma_p) = D_{\text{GAUSS}}(\Sigma_q\Sigma_p^{-1})$), of (3.2.1). Then, for any distribution R on X with mean $\boldsymbol{\mu}^*$ and covariance Σ_r, we have (as shown in Section 3.4 as part of the proof of the theorem below):

$$D_R(Q_{\boldsymbol{\mu}^*}\|P_{\boldsymbol{\mu}^*}) = D_{\Sigma_r}(\Sigma_q\|\Sigma_p). \tag{3.2.6}$$

In (3.2.8) below and later in (3.2.19) we use both notations D_R and D_{Σ_r} simultaneously—which, as will be seen, will greatly ease comparison to the exponential family version of our results.

Theorem 2. *Let $\mathcal{P}$ and $\mathcal{Q}$ be as above and let R be a distribution on $X = U$ with the same mean $\mathbb{E}_R[X] = \boldsymbol{\mu}^*$ as Q and with covariance matrix Σ_r. Let $U_{(1)}, U_{(2)}, \ldots$ be i.i.d. $\sim R$. Then $D_R(Q\|P_{\boldsymbol{\mu}^*})$ is finite and we have:*

1. **(UI)** *Let $S^{(n)}_{Q,\text{UI}} = \frac{q(U^{(n)})}{p_{\hat{\boldsymbol{\mu}}|n}(U^{(n)})}$ be defined as in (3.1.10). We have:*

$$\mathbb{E}_R[\log S^{(n)}_{Q,\text{UI}}] = nD_R(Q\|P_{\boldsymbol{\mu}^*}) - \frac{d_{rp}}{2}. \tag{3.2.7}$$

2. **(COND)** *Let* $S^{(n)}_{Q,\text{COND}} = \frac{q(U^{(n)}|Z)}{p(U^{(n)}|Z)}$ *be defined as in (3.1.9). We have:*

$$\mathbb{E}_R[\log S^{(n)}_{Q,\text{COND}}] = (n-1)D_R(Q\|P_{\boldsymbol{\mu}^*}) = n \cdot D_R(Q\|P_{\boldsymbol{\mu}^*}) - D_{\Sigma_r}(\Sigma_q\|\Sigma_p). \tag{3.2.8}$$

3. **(seq-RIPr/RIPr Simple Case; seq-RIPr Anti-Simple Case))** *Let* $S^{(n)}_{Q,\text{RIP}}, S^{(n)}_{Q,\text{SEQ-RIP}}$ *be defined as in (3.1.8) and (3.1.10). If* $\Sigma_q - \Sigma_p$ *is negative semidefinite (the 'simple' case) , then we have, for all n, that*

$$S^{(n)}_{Q,\text{RIP}} = S^{(n)}_{Q,\text{SEQ-RIP}} = \frac{q(U^{(n)})}{p_{\boldsymbol{\mu}^*}(U^{(n)})} \quad \textit{so that} \quad \mathbb{E}_R[\log S^{(n)}_{Q,\text{RIP}}] = nD_R(Q\|P_{\boldsymbol{\mu}^*}). \tag{3.2.9}$$

If $\Sigma_q - \Sigma_p$ *is positive semidefinite ('anti-simple case'), then* $S^{(n)}_{Q,\text{SEQ-RIP}} = 1$ *is trivial.*

4. **(RIPr, Anti-Simple Case)** *Let* $S^{(n)}_{Q,\text{RIP}}, S^{(n)}_{Q,\text{COND}}$ *and* $S^{(n)}_{Q,\text{SEQ-RIP}}$ *be defined as in (3.1.8), (3.1.9) and (3.1.10). If* $\Sigma_q - \Sigma_p$ *is positive semidefinite, then we have*

$$S^{(n)}_{Q,\text{RIP}} = \frac{q(U^{(n)})}{p_{W_0}(U^{(n)})} = S^{(n)}_{Q,\text{COND}}. \tag{3.2.10}$$

where W_0 *is a Gaussian prior with mean* $\boldsymbol{\mu}^*$ *and covariance matrix* $\Pi_0 := (\Sigma_q - \Sigma_p)/n$. $P_{W_0} = P_{\rightsquigarrow q(U^{(n)})}$ *is then the RIPr of* Q *onto* $\text{CONV}(\mathcal{P}(U^{(n)}))$.

Some remarks are in order.

Simple vs. Anti-Simple We see that, if $\Sigma_q - \Sigma_p$ is negative semidefinite, then the RIPr is simply an element of $\mathcal{P}$ and the growth-optimal e-variable is of the same form as it would be if $\mathcal{P}$ were simple. Following [40] we call this the *simple* case and indeed Part 3 of the theorem is really a direct corollary of the main result of that paper. We will formalize the notion of 'simplicity' for general exponential families in Section 3.3.

Comparing E-Power In the *strictly* simple case that $\Sigma_q - \Sigma_p$ is negative definite, we have, already mentioned below (3.2.3), that all λ_j in (3.2.2) are smaller than 1. Similarly, we are in the strict anti-simple case iff all these λ_j are greater than 1. This eigen-characterization leads to the following corollary about the e-power for the strict anti-simple and simple cases of UI vs. the other e-variables. In the strict anti-simple case, we have, by Part 4 above, $S^{(n)}_{Q,\text{COND}} = S^{(n)}_{Q,\text{RIP}}$, so, using (3.2.8) and (3.2.2) and the fact that in this strict anti-simple case, all eigenvalues λ_j are larger than 1, we find (3.2.11) below. Using (3.2.3) for the strict simple case, we find (3.2.12):

Corollary 2. **[e-power and growth optimality]**

$$\begin{aligned}\mathbb{E}_Q\left[\log S^{(n)}_{Q,\text{RIP}}/S^{(n)}_{Q,\text{UI}}\right] =& \frac{d_{qp}}{2} - D_{\Sigma_q}(\Sigma_q \| \Sigma_p) \qquad (3.2.11)\\ =& \frac{d}{2} + \frac{1}{2}\sum_{j=1}^{d} \log \lambda_j > \frac{d}{2}, \text{ in the strict anti-simple case;}\\ \mathbb{E}_Q\left[\log S^{(n)}_{Q,\text{RIP}}/S^{(n)}_{Q,\text{UI}}\right] =& \frac{d_{qp}}{2} < \frac{d}{2}, \text{ in the strict simple case.} \qquad (3.2.12)\end{aligned}$$

Even though $S_{Q,\text{UI}}$ therefore always has less e-power than $S_{Q,\text{RIP}}$, the difference (in contrast to the composite $\mathcal{Q}$ case in Section 3.2.2 below) does not keep growing with n. The conditional e-variable $S_{Q,\text{COND}}$ is identical to $S_{Q,\text{RIP}}$ in the anti-simple case but in the simple case it is hard to compare to UI; in general neither one outperforms the other. Again, we will see that with composite $\mathcal{Q}$, the situation changes.

3.2.2 $\mathcal{P}$, $\mathcal{Q}$ both multivariate Gaussian Location

We now consider the case that $\mathcal{P}$ is a d-dimensional Gaussian location family as before, but now $\mathcal{Q}$ is composite: it is itself the full d-dimensional Gaussian location family with nondegenerate covariance matrix $\Sigma_q \neq \Sigma_p$. The mean-value parameter spaces are $\mathtt{M}_q = \mathtt{M}_p = \mathbb{R}^d$.

We will establish explicit formulae and corresponding expected logarithmic growth of the various e-variables defined in Section 3.1.2. Since $\mathcal{Q}$ is composite, we need a method to estimate the distribution in $\mathcal{Q}$ given a data sequence $X_{(1)}, \ldots, X_{(n)}$, and this will determine the Q used in the definitions in that section. As stated in the introduction, we use two standard methods for this. The first method, usually called the plug-in method, is to use at each i a regularized ML estimator based on the past and defining Q as the product of predictive distributions. We use the variation studied by [38] who, following [29], call this the *prequential ML method*, setting, for each i,

$$Q^*(U_{(i)} \mid U^{(i-1)}) := Q_{\breve{\boldsymbol{\mu}}_{|i-1}}(U_{(i)}), \text{ where } \breve{\boldsymbol{\mu}}_{|n} := \frac{x_0 n_0 + \sum_{i=1}^{n} x_{(i)}}{n + n_0} \qquad (3.2.13)$$

for some fixed constant $n_0 > 0$ (not necessarily an integer) and $x_0 \in \mathtt{M}_p$. We take $n_0 > 0$ to ensure that $\breve{\boldsymbol{\mu}}_{|0}$ is well-defined and, when later applied to exponential families rather than Gaussians, to make sure that the relevant KL divergences remain finite.

Whenever we use an e-variable of the form $S_{Q^*,\cdot}$ with Q^* defined by (3.2.13), we abbreviate this to $S_{\breve{\boldsymbol{\mu}},\cdot}$. We use this method for e-variables of UI and sequential RIPr type. In particular, we set

$$S^{(n)}_{\breve{\boldsymbol{\mu}},\text{UI}} = \frac{\prod_{i=1}^{n} q_{\breve{\boldsymbol{\mu}}_{|i-1}}(U_{(i)})}{p_{\hat{\boldsymbol{\mu}}_{|n}}(U^{(n)})}, \quad S^{(n)}_{\breve{\boldsymbol{\mu}},\text{SEQ-RIP}} = \prod_{i=1}^{n} \frac{q_{\breve{\boldsymbol{\mu}}_{|i-1}}(U_{(i)})}{p_{\leftsquigarrow q_{\breve{\boldsymbol{\mu}}_{|i-1}}(U_{(i)})}(U_{(i)})}, \qquad (3.2.14)$$

where $p_{\leftsquigarrow q_{\breve{\boldsymbol{\mu}}_{|i-1}}(U_{(i)})}(U_{(i)})$ is the RIPr of $Q_{\breve{\boldsymbol{\mu}}_{|i-1}}(U_{(i)})$ onto $\textsc{conv}(\mathcal{P}(U_{(i)}))$ as in (3.1.10).

The second method to learn the alternative as data comes in, known as Robbins' method of mixtures [71], is to set Q^* to a Bayes marginal distribution, $Q^*(U^{(n)}) := Q_{W_1}(U^{(n)})$ where $q_{W_1}(U^{(n)}) = \int q_{\boldsymbol{\mu}}(U^{(n)})dW_1(\boldsymbol{\mu})$ for some prior W_1 on $\mathtt{M}_q$. In Theorem 3 below we uniquely consider Gaussian priors, i.e. priors of the form $W_1 = N(\boldsymbol{\mu}_1, \Pi_1)$ where we implicitly fix the dimension to d, i.e. $\boldsymbol{\mu}_1 \in \mathbb{R}^d$ and Π_1 a $d \times d$ nondegenerate covariance matrix. We use this method for the UI e-variables and for the growth-optimal e-variables, i.e. of RIPr type, in the anti-simple case. In particular, we set

$$S^{(n)}_{W_1,\text{UI}} = \frac{q_{W_1}(U^{(n)})}{p_{\hat{\boldsymbol{\mu}}_{|n}}(U^{(n)})}, \quad S^{(n)}_{W_1,\text{RIP}} = \frac{q_{W_1}(U^{(n)})}{p_{\leftsquigarrow q_{W_1}(U^{(n)})}(U^{(n)})} \tag{3.2.15}$$

with $p_{\leftsquigarrow q_{W_1}(U^{(n)})}$ the RIPr of $Q_{W_1}(U^{(n)})$ onto $\textsc{conv}(\mathcal{P}(U^{(n)}))$ as in (3.1.8).

Theorem 3. *Let $\mathcal{P}, \mathcal{Q}$ be as above, and let R be a distribution on $X(=U)$ with mean $\mathbb{E}_R[X] = \boldsymbol{\mu}^*$ and with covariance matrix Σ_r. Let $U_{(1)}, U_{(2)}, \ldots$ be i.i.d. $\sim R$. Then $D_R(Q_{\boldsymbol{\mu}^*} \| P_{\boldsymbol{\mu}^*})$ is finite and we have:*

1. **(UI)** *Let $S^{(n)}_{\breve{\boldsymbol{\mu}},\text{UI}}$ be as in (3.2.14) above. We have:*

$$\mathbb{E}_{U^{(n)} \sim R}[\log S^{(n)}_{\breve{\boldsymbol{\mu}},\text{UI}}] = n \cdot D_R(Q_{\boldsymbol{\mu}^*} \| P_{\boldsymbol{\mu}^*}) - \frac{d_{rq}}{2} \log n + O_{\breve{\boldsymbol{\mu}},\text{UI}}(1), \tag{3.2.16}$$

where, with $O_{\text{A}}, O_{\text{B}}$ given by (3.2.24) below,

$$O_{\breve{\boldsymbol{\mu}},\text{UI}}(1) = -\frac{d_{rp}}{2} - \frac{1}{2} \cdot O_{\text{A}}(\|x_0 - \boldsymbol{\mu}^*\|_2^2) - \frac{d_{rq}}{2} \cdot O_{\text{B}}(1).$$

Similarly, let $S^{(n)}_{W_1,\text{UI}}$ be as in (3.2.15) above. We have:

$$\mathbb{E}_{U^{(n)} \sim R}[\log S^{(n)}_{W_1,\text{UI}}] = n \cdot D_R(Q_{\boldsymbol{\mu}^*} \| P_{\boldsymbol{\mu}^*}) - \frac{d}{2} \log \frac{n}{2\pi} + O_{W_1,\text{UI}}(1), \tag{3.2.17}$$

where a precise expression for $O_{W_1,\text{UI}}$ is given in (3.2.24) below; up to $o(1)$, it is:

$$O_{W_1,\text{UI}}(1) = \frac{d_{rq} - d_{rp}}{2} + \frac{1}{2} \log \det \Sigma_q + \log w_1(\boldsymbol{\mu}^*) + o(1). \tag{3.2.18}$$

2. **(COND)** *Let $S^{(n)}_{\text{COND}}$ be as in (3.2.4) above. We have:*

$$\mathbb{E}_R[\log S^{(n)}_{\text{COND}}] = (n-1) \cdot D_{\Sigma_r}(\Sigma_q \| \Sigma_p) = n \cdot D_R(Q_{\boldsymbol{\mu}^*} \| P_{\boldsymbol{\mu}^*}) - D_{\Sigma_r}(\Sigma_q \| \Sigma_p). \tag{3.2.19}$$

3. **(seq-RIPr, Simple and Anti-Simple Case)** *Suppose that $\Sigma_q - \Sigma_p$ is negative semidefinite. Let $S^{(n)}_{\breve{\boldsymbol{\mu}},\text{SEQ-RIP}}$ be as in (3.2.14) above. We have: $P_{\leftsquigarrow q_{\breve{\boldsymbol{\mu}}_{|i-1}}(U_{(i)})}(U_{(i)}) =$*

$P_{\breve{\boldsymbol{\mu}}|i-1}(U_i)$ and

$$\mathbb{E}_R[\log S^{(n)}_{\breve{\boldsymbol{\mu}},\text{SEQ-RIP}}] = n \cdot D_R(Q_{\boldsymbol{\mu}^*} \| P_{\boldsymbol{\mu}^*}) + \frac{d_{rp} - d_{rq}}{2} \log n + O_{\text{SEQ-RIP}}(1), \tag{3.2.20}$$

where $O_{\text{SEQ-RIP}}(1) = -O_{\text{A}}(\|x_0 - \boldsymbol{\mu}^*\|_2^2) - \frac{d_{rp}-d_{rq}}{2} \cdot O_{\text{B}}(1)$. If $\Sigma_q - \Sigma_p$ is positive semidefinite, then $S^{(n)}_{\breve{\boldsymbol{\mu}},\text{SEQ-RIP}} = 1$ is trivial.

4. **(RIPr, Anti-Simple Case)** *Suppose that $\Sigma_q - \Sigma_p$ is positive semidefinite. Let $S^{(n)}_{W_1,\text{RIP}}$ be as in (3.2.15) so that $S^{(n)}_{W_1,\text{RIP}}$ is the GRO e-variable. Then*

$$p_{\leftsquigarrow q_{W_1}(U^{(n)})}(U^{(n)}) = p_{W_0}(U^{(n)}) \text{ where } W_0 = \mathcal{N}(\boldsymbol{\mu}_1, \Pi_1 + (\Sigma_q - \Sigma_p)/n), \tag{3.2.21}$$

and we have

$$S^{(n)}_{W_1,\text{RIP}} = \frac{q_{W_1}(U^{(n)})}{p_{W_0}(U^{(n)})} = S^{(n)}_{\text{COND}}. \tag{3.2.22}$$

A few remarks are in order.

Plug-In vs. Bayesian Q^* We studied the sequential RIPr (Part 3 of the theorem) only in combination with a plug-in Q^* and the full, 'global' RIPr (Part 4) only in combination with a Bayesian Q^*, but this was done for mathematical convenience only: one could in principle (though the analysis would be substantially more complicated in both cases) also study the RIPr for plug-in Q^* or the sequential RIPr for Bayesian Q^*.

Comparing E-Power of the Various Methods To determine e-power, we focus on the well-specified case again with $R = Q_{\boldsymbol{\mu}^*}$ so that $\Sigma_r = \Sigma_q$. Theorem 3 then implies:

Corollary 3. [e-power and growth optimality] *Under the conditions of Theorem 3, $\mathbb{E}_Q[\log S^{(n)}_{\text{COND}}] > n\epsilon$ for some $\epsilon > 0$ and all $n > 1$, and we have,*

$$\mathbb{E}_{Q_{\boldsymbol{\mu}^*}}[\log S^{(n)}_{W_1,\text{UI}} / S^{(n)}_{\breve{\boldsymbol{\mu}},\text{UI}}] = O(1).$$

$$\mathbb{E}_{Q_{\boldsymbol{\mu}^*}}[\log S^{(n)}_{\text{COND}} / S^{(n)}_{\breve{\boldsymbol{\mu}},\text{UI}}] = \frac{d}{2} \log n + O(1).$$

$$\mathbb{E}_{Q_{\boldsymbol{\mu}^*}}[\log S^{(n)}_{\breve{\boldsymbol{\mu}},\text{SEQ-RIP}} / S^{(n)}_{\breve{\boldsymbol{\mu}},\text{UI}}] = \frac{d_{qp}}{2} \log n + O(1) \textit{ with } d_{qp} \leq d, \text{ in simple case.} \tag{3.2.23}$$

$$\mathbb{E}_{Q_{\boldsymbol{\mu}^*}}[\log S^{(n)}_{\breve{\boldsymbol{\mu}},\text{SEQ-RIP}}] = 0 \text{ in anti-simple case.}$$

$$\mathbb{E}_{Q_{\boldsymbol{\mu}^*}}[\log S^{(n)}_{W_1,\text{RIP}} / S^{(n)}_{\text{COND}}] = 0 \text{ in anti-simple case,}$$

where d_{qp} is as in, and we used the inequality in, (3.2.3). Note that the inequality in (3.2.23) becomes strict in the strict simple case. Importantly, we only know how to explicitly calculate $S_{W_1,\text{RIP}}$ in the anti-simple case (then it is equal to S_{COND}) and we never need to specify any prior W_1 or $\breve{\boldsymbol{\mu}}$ when calculating S_{COND}.

One may now check that this is consistent with the results (3.1.4)-(3.1.6) provided informally in the introduction, where S_{UI} may stand for both $S_{W_1,\text{UI}}$ and $S_{\breve{\boldsymbol{\mu}},\text{UI}}$ and S_{RIP} stands for $S_{W_1,\text{RIP}}$. ((3.1.3) will be implied by Corollary 5 later on).

The $O(1)$ Terms Precise expressions for O_{A}, O_{B}, $O_{W_1,\text{UI}}$ are as follows.

$$O_{\text{A}}(\|x_0 - \boldsymbol{\mu}^*\|_2^2) = \sum_{i=0}^{n-1} \frac{1}{(1+i/n_0)^2} (\boldsymbol{\mu}^* - x_0)^\top \Sigma_q^{-1} (\boldsymbol{\mu}^* - x_0). \tag{3.2.24}$$

$$O_{\text{B}}(1) = \left(\sum_{i=1}^{n-1} \frac{i}{(n_0+i)^2} - \log n \right).$$

$$O_{W_1,\text{UI}}(1) = -\frac{d_{rp}}{2} - D_{\Sigma_r}(\Sigma_q \| \Sigma_q + n\Pi_1) + \frac{d}{2} \log \frac{n}{2\pi} - \frac{1}{2} O_{\text{C}}(\|\boldsymbol{\mu}^* - \boldsymbol{\mu}_1\|_2^2).$$

$$O_{\text{C}}(\|\boldsymbol{\mu}^* - \boldsymbol{\mu}_1\|_2^2) = (\boldsymbol{\mu}^* - \boldsymbol{\mu}_1)^\top (\Pi_1 + \Sigma_q/n)^{-1} (\boldsymbol{\mu}^* - \boldsymbol{\mu}_1).$$

Note that the O_{A} and O_{C} terms measure alignment between prior belief (x_0 or $\boldsymbol{\mu}_1$) and true $\boldsymbol{\mu}^*$, and become 0 if the belief is correct. O_{B} is small: for all n, we have $O_{\text{B}}(1) \leq \gamma + 1/(2n)$ (where $\gamma = 0.577\ldots$ is Euler's constant).

Relation to Optimal E-Variables for Group Invariant Testing Perez et al. [70] showed that if $\mathcal{Q}$ and $\mathcal{P}$ are both location families then the Bayes factor obtained by equipping both models with the (improper) right Haar prior gives an e-variable; it is even the GROW (worst-case growth optimal) e-variable for the two models. The right Haar prior for a location family is just the Lebesgue measure on $\boldsymbol{\mu}$. We may apply this result to our Gaussian location families. Using the fact [45] that the resulting e-variable satisfies $S^{(n)}_{\text{HAAR}}(U^{(1)}) = 1$ and

$$S^{(n)}_{\text{HAAR}}(U^{(n)}) = q_{W_1|X_{(1)}}(X_{(2)}, \ldots, X_{(n)}) / p_{W_0|X_{(1)}}(X_{(2)}, \ldots, X_{(n)}),$$

where $W_1|X_{(1)} = N(X_{(1)}, \Sigma_q)$ and $W_0|X_{(1)} = N(X_{(1)}, \Sigma_p)$ are the formal Bayes posteriors (based on the right Haar prior), after observing $X_{(1)}$, we may employ the same techniques as used in proving Theorem 3 to prove (see Appendix 3.A for details) that

$$\mathbb{E}_R[\log S^{(n)}_{\text{HAAR}}] = (n-1) \cdot D_{\Sigma_r}(\Sigma_q \| \Sigma_p), \tag{3.2.25}$$

and thus has the same e-power as the conditional e-variable. This e-power must then also be *worst-case* growth optimal (GROW) in the sense of [42], and by the results in that paper, we must have that $S^{(n)}_{\text{HAAR}} = S^{(n)}_{\text{COND}}$ are in fact *equal*.[1]

[1] Given the existence of the right Haar-based e-variable, the reader may wonder why we bother with all these other e-variables at all — the reason is that we really view them as idealized versions that should guide our analysis for general exponential families, for which Haar-based e-variables simply do not exist.

3.3 Multivariate exponential family

We extend the results of the previous section to general exponential families. We first introduce an important condition under which the form of S_{RIP} simplifies. Then, in Section 3.3.2, we present Theorem 4, the analogue to Theorem 2 (simple alternative) and then, in Section 3.3.3, Theorem 5, the analogue to Theorem 3.

3.3.1 The Simple and Anti-Simple Cases of [40]

In some special situations which we shall collectively refer to as *The Simple Case* (precise definition below), $S^{(1)}_{Q,\text{RIP}}$ (notation as in Section 3.1.2) reduces to a *simple* likelihood ratio between Q and a single element of $\mathcal{P}$. This is the upshot of the central result of [40], which we now summarize. We refer to [40] for more extensive discussion.

Fix a regular exponential family null hypothesis $\mathcal{P}$ for outcome U with sufficient statistic X. Let Q be a distribution on U. Suppose X has a moment generating function under Q and has density q relative to ν. We may then *generate* a second exponential family $\mathcal{Q}^{\text{GEN}}$ with carrier density q and the same sufficient statistic X as $\mathcal{P}$. The elements of this family are defined like (3.1.7) but with $p_{\boldsymbol{\mu}}$ replaced by $q_{\boldsymbol{\mu}}$ and Z_p by Z_q. We will adopt precisely the same notational conventions for $\mathcal{Q}^{\text{GEN}}$ as we did for $\mathcal{Q}$, in particular $\mathtt{M}_q^{\text{GEN}}$ is its mean-value parameter space and $\Sigma_q^{\text{GEN}}(\boldsymbol{\mu})$ is its $d \times d$ covariance matrix as a function of $\boldsymbol{\mu}$.

To prepare for our definitions, we note that there exist a variety of canonical parameterizations of an exponential family: by replacing the carrier density in (3.1.7) by the density of any element of the family, say with mean $\boldsymbol{\mu}^*$, we obtain a new canonical parameterization in which $p_0^{\text{CAN}} = p_{\boldsymbol{\mu}^*}$ has mean $\boldsymbol{\mu}^*$. For general regular exponential families on U that share the same sufficient statistic, we say that canonical parameterizations of $\mathcal{P}$ and $\mathcal{Q}$ *match* if $P_{\mathbf{0}}^{\text{CAN}}$ and $Q_{\mathbf{0}}^{\text{CAN}}$ are both well-defined and have the same mean, i.e. for some $\boldsymbol{\mu} \in \mathtt{M}_q \cap \mathtt{M}_p$, we have $\mathbb{E}_{P_{\mathbf{0}}^{\text{CAN}}}[X] = \mathbb{E}_{Q_{\mathbf{0}}^{\text{CAN}}}[X] = \boldsymbol{\mu}$.

Definition 1. Fix an exponential family $\mathcal{P}$ as above.

1. Let $\mathcal{Q}$ be an exponential family for U. We say that $\mathcal{P}$ and $\mathcal{Q}$ are *matching pairs* if they are both regular and they have the same sufficient statistic X and we have that (a) $\mathtt{M}_q \subseteq \mathtt{M}_p$ and, (b), for every matching canonical parameterization of $\mathcal{P}$ and $\mathcal{Q}$, we have that $\mathtt{B}_p \subseteq \mathtt{B}_q$.

2. Let $\mathcal{Q}$ be a collection of distributions on U so that each $Q \in \mathcal{Q}$, together with sufficient statistic X, generates the same exponential family $\mathcal{Q}^{\text{GEN}}$. (i) We say that we are *in the simple case* if $\mathcal{Q}^{\text{GEN}}$ is a matching pair with $\mathcal{P}$ and, for all $\boldsymbol{\mu} \in \mathtt{M}_q^{\text{GEN}}$, $\Sigma_q^{\text{GEN}}(\boldsymbol{\mu}) - \Sigma_p(\boldsymbol{\mu})$ is negative semidefinite. We say that we are in the *strictly* simple case if we are in the simple case and, for all $\boldsymbol{\mu} \in \mathtt{M}_q^{\text{GEN}}$, $\Sigma_q^{\text{GEN}}(\boldsymbol{\mu}) - \Sigma_p(\boldsymbol{\mu})$ is negative definite. (ii) We are in the $\boldsymbol{\mu}$-anti-simple case if $\boldsymbol{\mu} \in \mathtt{M}_q^{\text{GEN}} \cap \mathtt{M}_p$ and $\Sigma_q^{\text{GEN}}(\boldsymbol{\mu}) - \Sigma_p(\boldsymbol{\mu})$ is positive semidefinite ($\mathcal{P}$ and $\mathcal{Q}^{\text{GEN}}$ are not required to be matching pairs). We are in the strict $\boldsymbol{\mu}$-anti-simple case if $\boldsymbol{\mu} \in \mathtt{M}_q^{\text{GEN}}$ and $\Sigma_q^{\text{GEN}}(\boldsymbol{\mu}) - \Sigma_p(\boldsymbol{\mu})$ is positive definite.

In the following subsection we will be concerned with simple alternative $\mathcal{Q} = \{Q\}$, and then $\mathcal{Q}^{\text{GEN}}$ is well-defined as soon as X has a moment generating function under Q. Then, in Section 3.3.3, we consider $\mathcal{Q}$ that are themselves exponential families with sufficient statistic X. As is well-known, each member of a regular exponential family generates that same exponential family, so in that case we simply have $\mathcal{Q} = \mathcal{Q}^{\text{GEN}}$. In the 'companion paper' [40], we proved the following result:

Theorem [Corollary of Theorem 1 of [40]] Consider a testing problem with null $\mathcal{P}$ and alternative $\mathcal{Q}$ and suppose we are in the simple case. Then for every $\boldsymbol{\mu}^* \in \mathtt{M}_q$, the RIPr of $Q^{(n)}_{\boldsymbol{\mu}^*}$ is given by $P^{(n)}_{\boldsymbol{\mu}^*}$ so that the RIPr e-variable is given by

$$S^{(n)}_{Q_{\boldsymbol{\mu}^*},\text{RIP}} = \frac{q_{\boldsymbol{\mu}^*}(U^{(n)})}{p_{\boldsymbol{\mu}^*}(U^{(n)})}.$$

[40] gives many examples of $\mathcal{P}$ and $\mathcal{Q}$ that are 'matching pairs' and that fall under the 'simple case'. These include, for example, Bernoulli k-sample tests, the k-sample tests of Example 4 below, and also linear regression problems with Gaussian noise (Example 6). A prime example was already used implicitly in the previous section: if $\mathcal{P}$ and $\mathcal{Q}$ are both sets of multivariate Gaussians, we have $\mathtt{M}_p = \mathtt{M}_q = \mathtt{B}_p = \mathtt{B}_q = \mathbb{R}^d$ so we have matching pairs. Since in this case, the covariance matrices Σ_p and Σ_q do not depend on $\boldsymbol{\mu}$, we are in the simple case as soon as $\Sigma_q - \Sigma_p$ is negative semidefinite. This was exploited in the proofs of Theorem 2 and 3.

3.3.2 $\mathcal{Q}$ simple, $\mathcal{P}$ multivariate exponential family

We now assume $\mathcal{P}$ to be a regular exponential family and consider simple alternative $\mathcal{Q} = \{Q\}$ with Q a distribution on U with density q, mean $\mathbb{E}_Q[X] = \boldsymbol{\mu}^*$ and covariance matrix Σ_q. Below we state Theorem 4 which extends Theorem 2 to this setting. The results for UI, COND, SEQ-RIP and RIP will very closely follow those of that previous theorem, but, because of the added generality, regularity conditions are needed for some of them. We first discuss the most involved one, the **COND** condition:

We say that a d-dimensional random vector X is *of lattice form* if there exist real numbers $b_1, \ldots, b_d$ and $h_1, \ldots, h_d$ such that, for all $j \in \{1, \ldots, d\}$, all $u \in \mathcal{U}$, $X_j(u) \in \{b_j + sh_j : s \in \mathbb{Z}\}$. Obviously, most random variables with finite or countable support that are commonly encountered are of lattice form. Being either of this form or having a continuous density (with respect to Lebesgue measure) is the standard condition for the (multivariate) *local central limit theorem* [16] to hold, which is instrumental in the proof of (3.3.3) below. Concretely, we require the following:

Condition 1. *(**COND**) X has a moment generating function under Q and R in Theorem 4 below. Moreover, either X has a bounded continuous density (denoted $q^{[x]}$, $r^{[x]}$, $p^{[x]}_{\boldsymbol{\mu}}$ respectively), with respect to Lebesgue measure under Q, R and all $\boldsymbol{\mu} \in \mathtt{M}_p$, or X is of lattice form (and then has probability mass functions $q^{[x]}$, $r^{[x]}$, $p^{[x]}_{\boldsymbol{\mu}}$ respectively). The support $\mathcal{X} \subseteq \mathbb{R}^d$ ($\mathcal{X}$ is countable in the lattice case) under Q, R and all $P \in \mathcal{P}$*

coincides. Moreover, there is $a > 0$ such that for all $s > 0$,

$$\sup \left\{ \log \frac{p^{[x]}_{\boldsymbol{\mu}^*}(x)}{q^{[x]}(x)} : x \in \mathcal{X}, \|x - \boldsymbol{\mu}^*\|_2 \leq s \right\} = O(s^a). \tag{3.3.1}$$

Since all our results are only relevant for the case that $D_R(Q\|P_{\boldsymbol{\mu}^*}) < \infty$ anyway, requirement (3.3.1) merely says that the likelihood ratios cannot be super-exponentially far apart, so in that sense it is quite weak; yet it requires R to have exponentially small tails. From inspecting the proof it can be seen that if X has just three moments under R, the result (3.3.3) still holds if q and $p_{\boldsymbol{\mu}^*}$ are only polynomially apart; we have not bothered to formalize this. The requirement that X has a moment generating function under Q (rather than R) is essential for the proof though.

We further need a condition to prove a lower bound of the e-power of S_{UI}:

Condition 2. **(UI$^{\geq}$)** *R as in Theorem 4 below is such that for some $0 < \gamma < 1/2$, for all n:*

$$\mathbb{E}_R \left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n} - \boldsymbol{\mu}^*\|_2 \geq n^{-\gamma}} \cdot n \cdot D(P_{\hat{\boldsymbol{\mu}}_{|n}} \| P_{\boldsymbol{\mu}^*}) \right] = o(1).$$

This condition refers to $D(P_{\hat{\boldsymbol{\mu}}_{|n}} \| P_{\boldsymbol{\mu}^*})$ also for cases in which $\hat{\boldsymbol{\mu}}_{|n} \not\in \mathtt{M}_p$; in Appendix 3.B.1 we extend the definition to that case. We suspect the condition is quite weak: it is readily verified for one-dimensional models such as e.g. Poisson, negative binomial, exponential and so on (we illustrate that it holds for the Poisson distribution, when X has ≥ 3 moments under R), in Appendix 3.D.1), but, unlike for the 'dual' Condition 3 that we will discuss in the next subsection, we have not been able to come up with a general easy-to-verify 'n-free' condition that implies Condition 2 for multivariate families.

In this section, the covariance matrices in $\mathcal{P}$, $\mathcal{Q}$ and R are dependent on $\boldsymbol{\mu}$. We thus extend our notation for the trace (3.2.3) to $d_{ab}(\boldsymbol{\mu}) := \text{TR}(\Sigma_a(\boldsymbol{\mu})\Sigma_b^{-1}(\boldsymbol{\mu}))$. We also use (and will keep using in Theorem 5 and the proofs) the nonstandard notation $f(n) \leq O(1)$, to mean that there is a positive constant $M > 0$ such that, for all $n \in \mathbb{N}$, we have $f(n) \leq M$; $f(n) \leq o(1)$ means that there is a sequence of positive numbers $M_1, M_2, \ldots$ tending to 0 such that for all $n \in \mathbb{N}$, we have $f(n) \leq M_n$.

Theorem 4. *Let $\mathcal{P}$ be as above, and let Q be a distribution on U with density q and mean $\mathbb{E}_Q[X] = \boldsymbol{\mu}^*$ where $\boldsymbol{\mu}^* \in \mathtt{M}_p$ and with covariance matrix Σ_q. Let R be a distribution on X with the same mean $\mathbb{E}_R[X] = \boldsymbol{\mu}^*$ and with covariance matrix Σ_r, such that the first 3 moments of X exist under R and Q and $D_R(Q\|P_{\boldsymbol{\mu}^*})$ is well-defined and finite. Then, letting $U_{(1)}, U_{(2)}, \ldots$ be i.i.d. $\sim R$, we have:*

1. **(UI)** *Let $S^{(n)}_{Q,\text{UI}} = \frac{q(U^{(n)})}{p_{\hat{\boldsymbol{\mu}}_{|n}}(U^{(n)})}$ be as in (3.1.10). We have:*

$$\mathbb{E}_{U^{(n)} \sim R}[\log S^{(n)}_{Q,\text{UI}}] \leq n D_R(Q\|P_{\boldsymbol{\mu}^*}) - \frac{d_{rp}(\boldsymbol{\mu}^*)}{2} + o(1). \tag{3.3.2}$$

*If, moreover, R is such that Condition 2 ("**UI$^{\geq}$**") holds, then (3.3.2) holds with $\leq$ replaced by $=$.*

2. **(COND)** *Let* $S^{(n)}_{Q,\text{COND}} = \frac{q(U^{(n)}|Z)}{p(U^{(n)}|Z)}$ *be defined as in (3.1.9) and* $D_{\Sigma_r}(\cdot\|\cdot)$ *as in (3.2.5). Suppose Condition 1 (“***COND***”) holds. We have:*

$$\mathbb{E}_{U^{(n)}\sim R}[\log S^{(n)}_{Q,\text{COND}}] \geq nD_R(Q\|P_{\boldsymbol{\mu}^*}) - D_{\Sigma_r}(\Sigma_q\|\Sigma_p(\boldsymbol{\mu}^*)) + o(1). \tag{3.3.3}$$

*Moreover, if Condition 1 (“***COND***”) also holds with the role of* q *and* $p_{\boldsymbol{\mu}^*}$ *interchanged, (3.3.3) holds with equality.*

3. **(seq-RIPr/RIPr Simple Case, seq-RIPr Anti-Simple Case)** *Let* $S^{(n)}_{Q,\text{RIP}}, S^{(n)}_{Q,\text{SEQ-RIP}}$ *be defined as in (3.1.8) and (3.1.10). Suppose that, with alternative* $\mathcal{Q} = \{Q\}$*, we are in the simple case of Definition 1. Then we have, for all* n*, that*

$$S^{(n)}_{Q,\text{RIP}} = S^{(n)}_{Q,\text{SEQ-RIP}} = \frac{q(U^{(n)})}{p_{\boldsymbol{\mu}^*}(U^{(n)})} \quad ; \quad \mathbb{E}_{U^{(n)}\sim R}[\log S^{(n)}_{Q,\text{RIP}}] = nD_R(Q\|P_{\boldsymbol{\mu}^*}). \tag{3.3.4}$$

If $\Sigma_q - \Sigma_p(\boldsymbol{\mu}^*)$ *is positive definite (‘strict anti-simple case’) then for some* $\epsilon > 0$ *we have:*

$$\mathbb{E}_Q[\log S^{(n)}_{Q,\text{SEQ-RIP}}] = n(D(Q||P_{\boldsymbol{\mu}^*}) - \epsilon). \tag{3.3.5}$$

4. **(RIPr, Anti-Simple Case)** *Let* $S^{(n)}_{Q,\text{RIP}}, S^{(n)}_{Q,\text{COND}}$ *be defined as in (3.1.8) and (3.1.9) and* $D_{\Sigma_q}(\cdot\|\cdot)$ *as in (3.2.5). If* $\Sigma_q - \Sigma_p(\boldsymbol{\mu}^*)$ *is* positive *semidefinite, then, letting* $W_{0,(n)} = N(\boldsymbol{\mu}^*, (\Sigma_q - \Sigma_p)/n)$ *be a Gaussian distribution on* $\mathtt{M}_p$*, we have*

$$\mathbb{E}_Q[\log S^{(n)}_{Q,\text{RIP}}] \leq \mathbb{E}_Q\left[\log\frac{q(U^{(n)})}{p_{W_{0,(n)}}(U^{(n)})}\right] \leq nD(Q||P_{\boldsymbol{\mu}^*}) - D_{\Sigma_q}(\Sigma_q\Sigma_p^{-1}(\boldsymbol{\mu}^*)) + o(1). \tag{3.3.6}$$

*Moreover, if Condition 1 (‘***COND***’) holds, then (3.3.6) holds with equality up to* $o(1)$*, i.e.* $\mathbb{E}_Q[\log S^{(n)}_{Q,\text{RIP}}] = \mathbb{E}_Q[\log S^{(n)}_{Q,\text{COND}}] + o(1)$*.*

Note the very close resemblance between each part of this result and the corresponding part in Theorem 2. We briefly remark on some specifics.

First, for the **UI** part: the case in which it holds with equality is related to the celebrated Wilks phenomenon: if we let $Q \in \mathcal{P}$ and set $R = Q$, the theorem is still valid and the problem becomes ‘well-specified’. The $D_R(Q\|P_{\boldsymbol{\mu}^*})$ term then becomes 0 and we are left with $d/2$, the expectation of a χ^2-random variable with $d/2$ degrees of freedom, in accordance with Wilks’ result [54].

For the anti-simple case, we only consider the well-specified setting with $R = Q$ because the statement for general R is rather involved.

Theorem 4 shows that in the anti-simple case, the RIPr-prior that achieves GRO against Q at sample size n can be approximated, in the e-power sense, by a Gaussian with variance of order $O(1/n)$, but it does not tell us if the approximation is good enough to get something close to a real e-value if we use this prior in the denominator.[2] Note that even though we do not know if the Gaussian prior gives us an e-variable, we can still find another e-variable that provably is close to GRO – the theorem shows that we can simply use $S_{Q,\text{COND}}$.

Performing a similar analysis as in Corollary 2, we find:

Corollary 4. **[e-power and growth optimality]** *The relations of Corollary 2 still hold in the setting of Theorem 4, with $d_{qp} = d_{qp}(\boldsymbol{\mu}^*)$ now dependent on $\boldsymbol{\mu}^*$.*

Theorem 4 provides an asymptotic analysis, but we note that the Gaussian case of Theorem 2 is not the only case in which asymptotics are not required. For example, another case which does not require asymptotics for all e-variables except UI occurs if there is $P \in \mathcal{P}$ whose marginal distribution $P(X)$ for X coincides with $Q(X)$. In case $\mathcal{P}$ and $\mathcal{Q}$ are matching pairs we are then simultaneously in the (nonstrict) simple and anti-simple cases. We then have

$$D_R(Q\|P_{\boldsymbol{\mu}^*}) = \mathbb{E}_{X\sim R}\left[\log \frac{q(X)}{p_{\boldsymbol{\mu}^*}(X)}\right] + \mathbb{E}_{U\sim R}\left[\log \frac{q(U|X)}{p_{\boldsymbol{\mu}^*}(U|X)}\right] = \mathbb{E}_{U\sim R}\left[\log \frac{q(U|X)}{p_{\boldsymbol{\mu}^*}(U|X)}\right].$$

Proposition 4. *Suppose that $Q(X) = P_{\boldsymbol{\mu}^*}(X)$ for some $\boldsymbol{\mu}^* \in \mathtt{M}_p$. Then for all n, $S^{(n)}_{Q,\text{RIP}} = S^{(n)}_{Q,\text{SEQ-RIP}} = S^{(n)}_{Q,\text{COND}} = q(U^{(n)} \mid X^{(n)})/p(U^{(n)} \mid X^{(n)})$ so that*

$$\mathbb{E}_{U^{(n)}\sim R}[\log S^{(n)}_{Q,\text{RIP}}] = \mathbb{E}_{U^{(n)}\sim R}[S^{(n)}_{Q,\text{SEQ-RIP}}] = \mathbb{E}_{U^{(n)}\sim R}[\log S^{(n)}_{\text{COND}}] = D_R(Q\|P_{\boldsymbol{\mu}^*}).$$

Proof of Proposition 4 The results for $S_{Q,\text{RIP}}$ and $S_{Q,\text{SEQ-RIP}}$ follow from Theorem 4, the Simple Case, since $\Sigma_q = \Sigma_p(\boldsymbol{\mu}^*)$. This only leaves the proof for $S_{Q,\text{COND}}$, which follows by inspecting the proof for Part 2 ('**COND**') of Theorem 4 and noting that the result directly follows from (3.5.4).

Example 4. **[Two-Sample Tests]** Suppose that $U = (Y_a, Y_b)$, $X = Y_a + Y_b$, and $\mathcal{P} = \{P_{\mu^*} : \mu^* \in \mathtt{M}_p\}$ where P_{μ^*} expresses that Y_a, Y_b are independently $\sim N(\mu^*, \sigma^2)$ for some fixed σ^2, whereas $\mathcal{Q} = \{Q\}$ with, under Q, $Y_a \sim N(\mu_a, \sigma^2)$ and $Y_b \sim N(\mu_b, \sigma^2)$ independently, for some arbitrary fixed μ_a, μ_b. Note that $\boldsymbol{\mu}^* = \mathbb{E}_Q[X] = \mu_a + \mu_b$. Then, as shown by [44], $\mathcal{P}$ is an exponential family and we are in the setting of Proposition 4. They also show that the same holds if $\mathtt{M}_p = \mathbb{R}^+$ and P_μ represents that Y_a, Y_b are i.i.d according to a Poisson distribution with mean μ, whereas under Q, Y_a and Y_b are independently Poisson with means μ_a and μ_b, respectively. These examples, special cases of corresponding k-sample tests for $k = 2$, are discussed in detail in [44] and [40]. They also look at several other exponential families $\mathcal{P}$ to take the role of Gaussian and Poisson above. In all these other cases, the premise of Proposition 4 does not apply and asymptotics kick in. With the Bernoulli model, we are in the simple case; with the exponential, beta and Gaussian scale families, we are not [40, 44].

[2] Lardy et al.'s [57]'s Theorem 6 determines whether priors converging to the RIPr produce likelihood ratios converging to an e-variable, but its preconditions are not easily verified in the present case.

3.3.3 $\mathcal{P}$, $\mathcal{Q}$ both multivariate exponential families

In the section, we provide similar results as above but with composite $\mathcal{Q}$. That is,

$$\mathcal{Q} = \{Q_{\boldsymbol{\mu}} : \boldsymbol{\mu} \in \mathtt{M}_q\} \text{ and } \mathcal{P} = \{P_{\boldsymbol{\mu}} : \boldsymbol{\mu} \in \mathtt{M}_p\},$$

where $\mathcal{Q}$ and $\mathcal{P}$ are different exponential families for U with the same sufficient statistics, and we invariably assume $\mathtt{M}_q \subseteq \mathtt{M}_p$. The results involving the prequential plug-in method now require the following condition, referring to the sampling distribution R in the statement of Theorem 4 which has $\mathbb{E}_R[X] = \boldsymbol{\mu}^* \in \mathtt{M}_p$.

Condition 3. (plug-in) *There (a) exists an odd integer $m \geq 3$ such that the first m moments of X exist under R, and (b), for some $0 < \gamma < 1/2$,*

$$\mathbb{E}_R \left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n} - \boldsymbol{\mu}^*\|_2 \geq n^{-\gamma}} \cdot n \cdot D(Q_{\boldsymbol{\mu}^*} \| Q_{\breve{\boldsymbol{\mu}}_{|n}}) \right] = o(1). \tag{3.3.7}$$

As we show in Lemma 2, (3.B.8), in Appendix 3.B.1, a sufficient condition for (3.3.7) is that (a) holds, and (c) there also exist real $s > 1$ and $A > 0$ such that $\{\boldsymbol{\mu} : \|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|_2 < A\} \subset \mathtt{M}_q$ and for all $x_0 \in \mathtt{M}_q, 0 < \alpha < 1$,

$$\sup_{\boldsymbol{\mu} \in \mathtt{M}_q : \|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|_2 \geq A} \frac{D(Q_{\boldsymbol{\mu}^*} \| Q_{(1-\alpha)\boldsymbol{\mu} + \alpha x_0})}{\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|_2^{m-s}} < \infty. \tag{3.3.8}$$

As is well-known, *locally* both $D(Q_{\boldsymbol{\mu}} \| Q_{\boldsymbol{\mu}^*})$ and $D(Q_{\boldsymbol{\mu}^*} \| Q_{\boldsymbol{\mu}})$ are equal, up to lower order terms, to the quadratic form $(\boldsymbol{\mu}^* - \boldsymbol{\mu})^\top \Sigma_q^{-1} (\boldsymbol{\mu}^* - \boldsymbol{\mu}) \asymp \|\boldsymbol{\mu}^* - \boldsymbol{\mu}\|_2^2$. The $\mathbf{UI}^{\geq}$ Condition 2 implies that, up to $o(1)$, the KL divergence $D(Q_{\boldsymbol{\mu}} \| Q_{\boldsymbol{\mu}^*})$, for fixed $\boldsymbol{\mu}^*$ and varying $\boldsymbol{\mu}$, is determined by this local quadratic behaviour. The **plug-in** Condition 3 implies the same for $D(Q_{\boldsymbol{\mu}^*} \| Q_{\boldsymbol{\mu}})$. As can be seen, as $\boldsymbol{\mu}$ moves farther away from $\boldsymbol{\mu}^*$, this allows for a faster-than-quadratic increase of the KL divergences, but it should be at most polynomial in the ℓ_2-distance, and the rate of increase should be 'matched' by sufficiently high moments of R existing. The plug-in condition appears to be quite weak and is easily verified with appropriate values for m (details omitted) for 1-dimensional exponential families such as the Bernoulli, Poisson, geometric, binomial and negative binomial model. Below, and in the appendix, we verify it for some standard 2-dimensional models. Nevertheless, it does not *always* hold: for the 1-dimensional regular family $\mathcal{Q}$ with $X = U \in \mathtt{M}_q = \mathbb{R}$ generated by the *Landau distribution* [8], we find that $D(Q_{\boldsymbol{\mu}^*} \| Q_{\boldsymbol{\mu}^\circ})$ grows exponentially in $\boldsymbol{\mu}^\circ$ as $\boldsymbol{\mu}^\circ \to \infty$, thereby violating the condition. The appearance of the 'anchor' x_0 derives from the existence of a similar regularization term in the definition of $\breve{\boldsymbol{\mu}}$. Note that the condition would in some cases not hold without it; for example, in the Bernoulli model, $\sup_{\boldsymbol{\mu} \in (0,1)} D(P_{\boldsymbol{\mu}^*} \| P_{\boldsymbol{\mu}}) = D(P_{\boldsymbol{\mu}^*} \| P_1) = \infty$.

Example 5. [Gaussian Location-Scale] Let $\mathcal{Q}$ be the 2-dimensional family of Gaussian distributions $N(\mu, \sigma^2)$. In terms of the mean-value parameterization, we get $\boldsymbol{\mu} = (\mu_1, \mu_2)$ with $\mu_1 = \mu$ and $\mu_2 = \mu^2 + \sigma^2$ and $\mathtt{M}_q = \{(\mu_1, \mu_2) : \mu_1 \in \mathbb{R}, \mu_2 > \mu_1^2\}$ so that $Q_{\boldsymbol{\mu}}$ with $\boldsymbol{\mu} = (\mu_1, \mu_2)$ represents $N(\mu_1, \mu_2 - \mu_1^2)$. A standard calculation gives, with $\boldsymbol{\mu}^* = (\mu_1^*, \mu_1^{*2} + \sigma^{*2})$, and $\boldsymbol{\mu}^\circ = (\mu_1^\circ, \mu_2^\circ)$, that

$$D(Q_{\boldsymbol{\mu}^*} \| Q_{\boldsymbol{\mu}^\circ}) = \frac{1}{2} \left(\log \frac{\mu_2^\circ - \mu_1^{\circ 2}}{\sigma^{*2}} + \frac{\sigma^{*2} + (\mu_1^* - \mu_1^\circ)^2}{\mu_2^\circ - \mu_1^{\circ 2}} - 1 \right). \tag{3.3.9}$$

Now let $\boldsymbol{\mu}^\circ = (1-\alpha)\boldsymbol{\mu} + \alpha x_0$. In Appendix 3.D.2, we show (this involves some work) that for every $0 < \alpha < 1$, every $x_0 = (x_{0,1}, x_{0,2}) \in \mathtt{M}_q$, we have

$$\inf_{\boldsymbol{\mu} \in \mathtt{M}_q} (\mu_2^\circ - \mu_1^{\circ 2}) = \alpha(x_{0,2} - x_{0,1}^2). \tag{3.3.10}$$

with $\boldsymbol{\mu}^\circ = (\mu_1^\circ, \mu_2^\circ)$ a function of $\boldsymbol{\mu}$ and $x_{0,2} > x_{0,1}^2$ (so (3.3.10) > 0) because $x_0 \in \mathtt{M}_q$. This implies that the second term in (3.3.9) is $O(\sigma^{*2} + (\mu_1^* - \mu_1^\circ)^2)$. As we also show in the appendix, this then easily implies that for every $\boldsymbol{\mu}^* \in \mathtt{M}_q$, every $A > 0$, we have (3.3.8). with $m - s = 2$, verifying Condition 3 as soon as R has 5 or more moments.

With similar arguments, one can show (3.3.8) holds with $m - s = 2$ for $\mathcal{Q}$ the family of Gamma distributions; details are in Appendix 3.D.1.

Theorem 5. *Let $\mathcal{P}$ and $\mathcal{Q}$ be two regular exponential families for U with the same sufficient statistic such that $\mathcal{P} \cap \mathcal{Q} = \emptyset$ and $\mathtt{M}_q \subseteq \mathtt{M}_p$, as above. Consider a distribution R as above with mean $\mathbb{E}_R[X] = \boldsymbol{\mu}^* \in \mathtt{M}_q \subseteq \mathtt{M}_p$ and with covariance matrix Σ_r, such that the first 3 moments of X exist under R and $D_R(Q\|P)$ is finite for all $Q \in \mathcal{Q}, P \in \mathcal{P}$. Let $U_{(1)}, U_{(2)}, \ldots$ be i.i.d. $\sim R$. We have:*

1. **(UI)** *Let $S^{(n)}_{\breve{\boldsymbol{\mu}},\text{UI}}$ be as in (3.2.14). Suppose Condition 3 (“**plug-in**”) holds. We have:*

$$\mathbb{E}_R[\log S^{(n)}_{\breve{\boldsymbol{\mu}},\text{UI}}] \leq nD_R(Q_{\boldsymbol{\mu}^*} \| P_{\boldsymbol{\mu}^*}) - \frac{d_{rq}(\boldsymbol{\mu}^*)}{2} \log n + O(1). \tag{3.3.11}$$

Moreover, suppose that R has density r, that W_1 is a prior on $\mathtt{M}_q$ with continuous and strictly positive density in a neighborhood of $\boldsymbol{\mu}^*$, and such that for some $\epsilon > 0$, we have
$\mathbb{E}_R |\log p_{\boldsymbol{\mu}^*}(U)/r(U)|^{1+\epsilon} < \infty$. Then:

$$\mathbb{E}_R[\log S^{(n)}_{W_1,\text{UI}}] \leq nD_R(Q_{\boldsymbol{\mu}^*} \| P_{\boldsymbol{\mu}^*}) - \frac{d}{2} \log \frac{n}{2\pi} + O_{W_1,\text{UI}}(1), \tag{3.3.12}$$

*where $O_{W_1,\text{UI}}(1)$ was given up to $o(1)$ in (3.2.18). If moreover Condition 2 (“**UI**$^{\geq}$”) holds for R then (3.3.11) and (3.3.12) hold with equality.*

2. **(COND)** *Let $S^{(n)}_{\text{COND}}$ be as in (3.2.4) and $D_{\Sigma_r}(\cdot\|\cdot)$ be as in (3.2.5). Suppose Condition 1 (“**COND**”) holds for R, $Q = Q_{\boldsymbol{\mu}^*}$ and $P_{\boldsymbol{\mu}^*}$. We have:*

$$\mathbb{E}_R[\log S^{(n)}_{\text{COND}}] \geq nD_R(Q_{\boldsymbol{\mu}^*} \| P_{\boldsymbol{\mu}^*}) - D_{\Sigma_r}(\Sigma_q(\boldsymbol{\mu}^*) \| \Sigma_p(\boldsymbol{\mu}^*)) + o(1). \tag{3.3.13}$$

Moreover, if Condition 1 holds with the role of q and $p_{\boldsymbol{\mu}^*}$ interchanged, (3.3.13) holds with equality.

3. **(seq-RIPr, Simple and Anti-Simple Case)** *Suppose Condition 3 (“**plug-in**”) holds. Let $S^{(n)}_{\breve{\boldsymbol{\mu}},\text{SEQ-RIP}}$ be as in (3.2.14). If we are in the* simple case *of Definition 1, then we have: $P_{\leftsquigarrow q_{\breve{\boldsymbol{\mu}}|i-1}(U_{(i)})}(U_{(i)}) = P_{\breve{\boldsymbol{\mu}}|i-1}(U_i)$ and*

$$\mathbb{E}_R[\log S^{(n)}_{\breve{\boldsymbol{\mu}},\text{SEQ-RIP}}] = nD_R(Q_{\boldsymbol{\mu}^*}\|P_{\boldsymbol{\mu}^*}) + \frac{d_{rp}(\boldsymbol{\mu}^*) - d_{rq}(\boldsymbol{\mu}^*)}{2}\log n + O_{\text{SEQ-RIP}}(1) + o(1), \tag{3.3.14}$$

where $O_{\text{SEQ-RIP}}(1)$ is as below (3.2.20). On the other hand, if $\Sigma_q(\boldsymbol{\mu}^*) - \Sigma_p(\boldsymbol{\mu}^*)$ is positive definite (‘strict anti-simple case’) then for some $\epsilon > 0$, for all n,

$$\mathbb{E}_Q[\log S^{(n)}_{\breve{\boldsymbol{\mu}},\text{SEQ-RIP}}] \leq n(D(Q_{\boldsymbol{\mu}^*}\|P_{\boldsymbol{\mu}^*}) - \epsilon). \tag{3.3.15}$$

4. **(RIPr, General Case)** *Let $S^{(n)}_{W_1,\text{RIP}}$ be defined as in (3.1.8) with W_1 a prior with support contained in $\mathtt{M}_p \cap \mathtt{M}_q$ and with positive continuous density in a neighborhood of $\boldsymbol{\mu}^*$, and let $D_{\Sigma_r}(\cdot\|\cdot)$ be as in (3.2.5). We have*

$$\mathbb{E}_Q[\log S^{(n)}_{W_1,\text{RIP}}] \leq \mathbb{E}_Q\left[\log \frac{q_{W_1}(U^{(n)})}{p_{W_1}(U^{(n)})}\right] = nD(Q||P_{\boldsymbol{\mu}^*}) - D_{\text{GAUSS}}(\Sigma_q(\boldsymbol{\mu}^*)\Sigma_p^{-1}(\boldsymbol{\mu}^*)) + o(1), \tag{3.3.16}$$

*and if additionally Condition 1 (‘**COND**’) holds, then the inequality becomes an equality.*

Note again the very close similarity to the corresponding Theorem 3 for the Gaussian case — with the exception of Part 4, where allowing for asymptotics ($o(1)$) leads to a much more general result: we can take general continuous, rather than Gaussian, priors, and the result holds not just in the anti-simple case but in general.

The theorem implies the following corollary, which, like Corollary 3, commensurates with (3.1.3)-(3.1.6) in the introduction.

Corollary 5. [e-power and growth optimality] *Suppose Condition 1 (‘**COND**’) and 3 (**plug-in**) holds for $R = Q = Q_{\boldsymbol{\mu}^*}$, $\mathcal{P}$ and $\mathcal{Q}$. Then for some $\epsilon > 0$, it holds $\mathbb{E}_Q[\log S^{(n)}_{\text{COND}}] \geq n\epsilon - o(1)$, and*

$$\mathbb{E}_{Q_{\boldsymbol{\mu}^*}}[\log S^{(n)}_{\text{COND}}/S^{(n)}_{\breve{\boldsymbol{\mu}},\text{UI}}] \overset{(a)}{\geq} \frac{d}{2}\log n + O(1).$$

$$\mathbb{E}_Q[\log S^{(n)}_{\text{RIP}}/S^{(n)}_{\text{COND}}] = o(1).$$

$$\mathbb{E}_{Q_{\boldsymbol{\mu}^*}}[\log S^{(n)}_{\breve{\boldsymbol{\mu}},\text{SEQ-RIP}}/S^{(n)}_{\breve{\boldsymbol{\mu}},\text{UI}}] \geq \frac{d_{qp}(\boldsymbol{\mu}^*)}{2}\log n + O(1) \textit{ with } d_{qp} \leq d, \text{ in the simple case.}$$

$$\mathbb{E}_Q[\log S^{(n)}_{\text{SEQ-RIP}}/S^{(n)}_{\text{COND}}] \leq -n\epsilon + o(1) \text{ for some } \epsilon > 0, \text{ in the strict anti-simple case.}$$

where (a) becomes an equality if Condition 2 holds, and $d_{qp}(\boldsymbol{\mu}^*)$ is as in (3.2.3), with $d_{qp}(\boldsymbol{\mu}^*) \leq d$ in the simple case (with strict inequality in the strictly simple case) following in the same way as in (3.2.3).

Example 6. We consider two practically relevant settings in which Theorem 5 applies. A particularly simple one is given by the two-sample tests of Example 4. In the setting of that example, we may start with a simple alternative Q under which $Y_a \sim P_{\mu_a}, Y_b \sim P_{\mu_b}$ independently, and use it to generate a full alternative $\mathcal{Q} := \mathcal{Q}^{\text{GEN}}$ with the same sufficient statistic as $Y_a + Y_b$, as in Section 3.3.1. Such an alternative is often a natural choice, corresponding to testing a particular notion of *effect size*. For example, in Bernoulli two-sample tests, as shown by [40, Section 4.3], the generated alternative $\mathcal{Q}^{\text{GEN}}$ will correspond to the set of all (μ'_a, μ'_b) that have the same *log odds ratio* as the original Q, i.e. with $\delta^* = \log \mu_a(1-\mu_b)/((1-\mu_a)\mu_b)$, we have that $\mathcal{Q}^{\text{GEN}} = \{Q_{\mu'_a,\mu'_b} : \log \mu'_a(1-\mu'_b)/((1-\mu'_a)\mu'_b) = \delta^*\}$, whereas $\mathcal{P}$ is the set of all distributions with $\mu_a = \mu_b$, i.e. log odds ratio of 0. Thhe log odds ratio has long been a standard notion of effect size in two-sample tests, and our approach implicitly adopts it. Of course, in practice we would often want to test the larger alternative hypothesis that $\delta \geq \delta^*$. Clearly all four e-values we develop in this paper can be used in that setting as well, since the validity of e-values is independent of the actually chosen alternative; we suspect that they still have close to optimal e-power in a worst-case sense (the GROW sense of [42]), and we plan to investigate whether this is really so in future work. Another example is provided by the quintessential statistics problem: linear regression. We assume $X = \beta^\top Z + \gamma T + \epsilon$ for some covariate vector (Z, T) and 0-mean Gaussian noise ϵ. We want to test whether $\gamma = 0$ or $\gamma = \gamma^*$ for 'effect size' γ^*. [40, Section 4.4] derives that this standard problem can be recast as an instance of a test between two exponential families with the same sufficient statistic. Hence Theorem 5, Part 3 on the sequentialized RIPr is applicable here, and gives us bounds on e-power.

3.4 Proofs for Section 3.2

In the first subsection we gradually build up a series of remarkable properties of KL divergences between multivariate Gaussians. Once we have established all of these, proofs of the various parts of Theorem 2 and 3, established in the next subsections, will become immediate.

3.4.1 Preparation for both Proofs

Let P, Q, R and R_{G} be distributions on $X = U \in \mathbb{R}^d$. Suppose that, under R, $\mathbb{E}_R[X] = \boldsymbol{\mu}^*$ and the nondegenerate covariance matrix is Σ_r, that R_{G} is Gaussian with the same mean and covariance, and suppose that Q and P are Gaussian with $\mathbb{E}_Q[X] = \mathbb{E}_P[X] = \boldsymbol{\mu}^*$ and nondegenerate covariance matrices Σ_q and Σ_p respectively. In Section 3.2.2 we introduced the notation $D_{\Sigma_r}(\Sigma_q \| \Sigma_p)$. The following equation implies that $D_{\Sigma_r}(\Sigma_q \| \Sigma_p) = D_R(Q \| P)$, as claimed in (3.2.6); even more generally, if P is replaced by P_0 with mean $\boldsymbol{\mu}_0$ and still covariance matrix Σ_p, we obtain a generalization of the well-known formula for the KL divergence between two multivariate normals that do not necessarily have the same mean:

$$
\begin{aligned}
D_R(Q\|P_0) &= \mathbf{E}_R\left[\log\frac{q(X)}{p_0(X)}\right] = \mathbf{E}_{R_{\mathrm{G}}}\left[\log\frac{q(X)}{p_0(X)}\right] \\
&= D_{\text{GAUSS}}(\Sigma_r\Sigma_p^{-1}) - D_{\text{GAUSS}}(\Sigma_r\Sigma_q^{-1}) + \frac{1}{2}(\boldsymbol{\mu}^* - \boldsymbol{\mu}_0)^\top\Sigma_p^{-1}(\boldsymbol{\mu}^* - \boldsymbol{\mu}_0) \\
&= D_{\Sigma_r}(\Sigma_q\|\Sigma_p) + \frac{1}{2}(\boldsymbol{\mu}^* - \boldsymbol{\mu}_0)^\top\Sigma_p^{-1}(\boldsymbol{\mu}^* - \boldsymbol{\mu}_0),
\end{aligned}
\tag{3.4.1}
$$

with D_{GAUSS} as in (3.2.1) and D_{Σ_r} as in (3.2.5). Here the first equality is definition, the second follows by writing out the definitions, and in the third we used the standard (easily derived) formula for the KL divergence between two multivariate Gaussians.

Now we let $\mathcal{P}$ and $\mathcal{Q}$ be two d-dimensional Gaussian location families for i.i.d. data $U, U_{(1)}, U_{(2)}, \ldots$ as in Section 3.2.1, with nondegenerate covariance matrices Σ_p and Σ_q and notations as in that section. Let R be a distribution as above, with $\mathbb{E}_R[X] = \boldsymbol{\mu}^*$, extended to $U_{(1)}, U_{(2)}, \ldots$ by independence. We have, by a standard property of Gaussian distributions, with $Z = n^{-1/2}\sum_{i=1}^n X_i = n^{1/2}\cdot\hat{\boldsymbol{\mu}}_{|n}$, that

$$
Q_{\boldsymbol{\mu}^*}(Z) = Q_{\boldsymbol{\mu}^*}(X); P_{\boldsymbol{\mu}^*}(Z) = P_{\boldsymbol{\mu}^*}(X) \text{ so } D_R(Q_{\boldsymbol{\mu}^*}(Z)\|P_{\boldsymbol{\mu}^*}(Z)) = D_R(Q_{\boldsymbol{\mu}^*}\|P_{\boldsymbol{\mu}^*}).
$$

Also, letting W_0 and W_1 be arbitrary prior distributions on $\mathtt{M}_p = \mathtt{M}_q = \mathbb{R}^d$ (which may be degenerate, i.e. put all their mass on a single point), by sufficiency we have

$$
S_{\text{COND}} = \frac{q_{W_1}(U^{(n)}|Z)}{p_{W_0}(U^{(n)}|Z)} = \frac{q_{\boldsymbol{\mu}^*}(U^{(n)}|Z)}{p_{\boldsymbol{\mu}^*}(U^{(n)}|Z)}. \tag{3.4.2}
$$

Also, abbreviating $Q(U^{(n)})$ to $Q^{(n)}$ and similarly for $P^{(n)}$,

$$
nD_R(Q_{\boldsymbol{\mu}^*}\|P_{\boldsymbol{\mu}^*}) = D_R(Q_{\boldsymbol{\mu}^*}^{(n)}\|P_{\boldsymbol{\mu}^*}^{(n)}) = D_R(Q_{\boldsymbol{\mu}^*}^{(n)}|Z) \parallel P_{\boldsymbol{\mu}^*}^{(n)}|Z) + D_R(Q_{\boldsymbol{\mu}^*}(Z)\|P_{\boldsymbol{\mu}^*}(Z)). \tag{3.4.3}
$$

Combining (3.4.2) and (3.4.3) gives

$$
D_R(Q_{W_1}^{(n)}|Z \parallel P_{W_0}^{(n)}|Z) = D_R(Q_{\boldsymbol{\mu}^*}^{(n)}|Z \parallel P_{\boldsymbol{\mu}^*}^{(n)}|Z) = (n-1)D_R(Q_{\boldsymbol{\mu}^*}\|P_{\boldsymbol{\mu}^*}). \tag{3.4.4}
$$

A standard result (see for example [17, Chapter 2.3.]) gives that the marginal distribution of $\hat{\boldsymbol{\mu}}_{|n} \in \mathbb{R}^d$ under P_W, with $W = N(\boldsymbol{\mu}^*, \Pi)$ is given by:

$$
\hat{\boldsymbol{\mu}}_{|n} \sim N(\boldsymbol{\mu}^*, \Pi + \Sigma_p/n). \tag{3.4.5}
$$

Therefore, if W_1 puts mass 1 on $\{\boldsymbol{\mu}^*\}$ and we have $W_0 = N(\boldsymbol{\mu}^*, \Pi_0)$ with Π_0 such that

$$
\Pi_0 + \Sigma_p/n = \Sigma_q/n. \tag{3.4.6}
$$

then we get by (3.4.5) that $\hat{\boldsymbol{\mu}}_{|n}$ and therefore Z has the same distribution under q_{W_1}

as under p_{W_0} and then by (3.4.2),

$$\frac{q_{W_1}(U^{(n)})}{p_{W_0}(U^{(n)})} = \frac{q_{W_1}(U^{(n)} \mid Z) q^{\circ}_{W_1}(Z)}{p_{W_0}(U^{(n)} \mid Z) p^{\circ}_{W_0}(Z)} = S^{(n)}_{Q,\text{COND}}. \tag{3.4.7}$$

where $q^{\circ}_{W_1}$ denotes the density of Z under Q_{W_1}, and similarly for $p^{\circ}_{W_0}$, and we have $q^{\circ}_{W_0} = p^{\circ}_{W_0}$.

Now assume that $W_1 = N(\boldsymbol{\mu}^*, \Pi_1)$ rather than degenerate, with Π_1 some nondegenerate covariance matrix. We can again determine the corresponding W_0 such that (3.4.7) holds: if we can set $W_0 = N(\boldsymbol{\mu}^*, \Pi_0)$ with Π_0 such that

$$\Pi_0 + \Sigma_p/n = \Pi_1 + \Sigma_q/n, \tag{3.4.8}$$

then we again have (3.4.7).

For both choices of W_1, if a corresponding prior satisfying (3.4.6) or, respectively, (3.4.8) exists, the Bayes factor $q_{W_1}(U^{(n)})/p_{W_0}(U^{(n)})$ is equal to S_{COND} and hence becomes an e-variable. From Theorem 1 (specifically Corollary 1) of [42] we know that for every W_1 on $\mathtt{M}_q$, there can be at most one corresponding prior W_0 such that this Bayes factor is an e-variable, and if this prior exists, this e-variable is the GRO e-variable relative to Q. It follows that, if W_1 is such that a W_0 satisfying (3.4.6) or (3.4.8) exists, then S_{COND} is the GRO e-variable that maximizes e-power.

We need one more proposition before giving the actual proof of Theorem 2 and 3. It provides variations of standard results, whose proofs we delegate to Appendix 3.A — note the remarkable symmetry between (3.4.12) and (3.4.10) if $\breve{\boldsymbol{\mu}}_{|n}$ is chosen equal to $\hat{\boldsymbol{\mu}}_{|n}$.

Proposition 5. *Let $\breve{\mu}$ be as before with $n_0 \geq 0$, and let $W_1 = N(\boldsymbol{\mu}_1, \Pi_1)$ be a multivariate normal on $\mathbb{R}^d$. We have, with O_{A}–O_{C} as in (3.2.24):*

$$\mathbb{E}_{U,U^{(n)} \sim R}\left[\log \frac{q_{\boldsymbol{\mu}^*}(U)}{q_{\breve{\boldsymbol{\mu}}_{|n}}(U)}\right] = \mathbb{E}_{U^{(n)} \sim R}\left[D(Q_{\boldsymbol{\mu}^*} \| Q_{\breve{\boldsymbol{\mu}}_{|n}})\right] =$$
$$\frac{1}{2}\left(\frac{1}{(1+(n/n_0))^2}(\boldsymbol{\mu}^* - x_0)^\top \Sigma_q^{-1}(\boldsymbol{\mu}^* - x_0) + \left(\frac{n}{(n_0+n)^2}\right) d_{rq}\right) = \frac{1}{n}\frac{d_{rq}}{2} + O\left(\frac{1}{n^2}\right), \tag{3.4.9}$$

where the second equality holds if $n_0 > 0$. If $n_0 = 0$ (i.e. $\breve{\boldsymbol{\mu}}_{|n}$ is chosen equal to $\hat{\boldsymbol{\mu}}_{|n}$) and $n > 1$ we get:

$$\mathbb{E}_{U,U^{(n)} \sim R}\left[n \cdot \log\left(q_{\boldsymbol{\mu}^*}(U)/q_{\hat{\boldsymbol{\mu}}_{|n}}(U)\right)\right] = \mathbb{E}_{U^{(n)} \sim R}\left[n \cdot D(Q_{\boldsymbol{\mu}^*} \| Q_{\hat{\boldsymbol{\mu}}_{|n}})\right] = \frac{d_{rq}}{2}. \tag{3.4.10}$$

We further have:

$$\mathbb{E}_{U^{(n)}\sim R}\left[\log \frac{q_{\boldsymbol{\mu}^*}(U^{(n)})}{\prod_{i=1}^n q_{\breve{\boldsymbol{\mu}}_{|i-1}}(U_{(i)})}\right] = \sum_{i=1}^n \mathbb{E}_{U^{(i-1)}\sim R}\left[D(Q_{\boldsymbol{\mu}^*}\|Q_{\breve{\boldsymbol{\mu}}_{|i-1}})\right] = \tag{3.4.11}$$

$$\frac{d_{rq}}{2}\log n + \frac{1}{2}O_{\text{A}}(\|x_0 - \boldsymbol{\mu}^*\|_2^2) + O_{\text{B}}(1)\cdot\frac{d_{rq}}{2}.$$

$$\mathbb{E}_{U^{(n)}\sim R}\left[\log \frac{q_{\hat{\boldsymbol{\mu}}_{|n}}(U^{(n)})}{q_{\boldsymbol{\mu}^*}(U^{(n)})}\right] = \mathbb{E}_{U^{(n)}\sim R}\left[n\cdot D(Q_{\hat{\boldsymbol{\mu}}_{|n}}\|Q_{\boldsymbol{\mu}^*})\right] = \frac{d_{rq}}{2}. \tag{3.4.12}$$

$$\mathbb{E}_{U^{(n)}\sim R}\left[\log \frac{q_{\boldsymbol{\mu}^*}(U^{(n)})}{q_{W_1}(U^{(n)})}\right] = D_{\Sigma_r}(\Sigma_q\|(n\Pi_1+\Sigma_q)) + \frac{1}{2}O_{\text{C}}(\|\boldsymbol{\mu}^* - \boldsymbol{\mu}_1\|_2^2) = \tag{3.4.13}$$

$$\frac{d}{2}\log\frac{n}{2\pi} - \frac{1}{2}\log\det\Sigma_q - \log w_1(\boldsymbol{\mu}^*) - \frac{d_{rq}}{2} + o(1).$$

3.4.2 Proof of Theorem 2

Finiteness of $D_R(Q\|P)$ is immediate from evaluating the definition.

UI (3.2.7) follows almost immediately from Proposition 5, (3.4.12) applied with model $\mathcal{P}$ in the role of $\mathcal{Q}$, i.e. all q's replaced by p's (evidently it is still valid then), and using that $\mathbb{E}_R[\log(q_{\boldsymbol{\mu}^*}(U^{(n)})/p_{\boldsymbol{\mu}^*}(U^{(n)}))] = nD_R(Q_{\boldsymbol{\mu}^*}\|P_{\boldsymbol{\mu}^*})$.

COND (3.2.8) is a direct consequence of (3.4.2) and (3.4.4).

seq-RIPr Assume first that $\Sigma_q - \Sigma_p$ is negative semidefinite. According to Theorem 1 in [40], this implies that the RIPr of $Q(U)$ (i.e. Q restricted to a single outcome) is equal to $P_{\boldsymbol{\mu}^*}(U)$, i.e. the single element of $\mathcal{P}$ with the same mean as Q. This implies that $S_{Q,\text{SEQ-RIP}}$ is as in (3.2.9). Thus, $q(U^{(n)})/p_{\boldsymbol{\mu}^*}(U^{(n)})$ is an e-variable. The corollary of Theorem 1 of [42] ('there can be only one e-variable with Q in the numerator and an element P' of $\mathcal{P}$ in the denominator, and if it exists then P' is the RIPr so q/p is growth-rate optimal') implies that $P_{\boldsymbol{\mu}^*}(U^{(n)})$ is the RIPr of $Q(U^{(n)})$, and hence $S^{(n)}_{Q,\text{SEQ-RIP}} = S^{(n)}_{Q,\text{RIP}}$.

Now assume $\Sigma_q - \Sigma_p$ is positive semidefinite. $S^{(n)}_{Q,\text{SEQ-RIP}} = 1$ then follows because in this case, using the result for the Bayes marginal distribution (3.4.5) applied with $n = 1$, we find that there exists a prior W_0 such that $p_{W_0}(X) = q(X)$, namely, $W_0 = N(\boldsymbol{\mu}^*, (\Sigma_q - \Sigma_p))$. It follows that the RIPr of Q onto $\mathcal{P}$ for a single outcome U must be given by this very prior, i.e. $P_{\leftsquigarrow q(U)}(U) = P_{W_0}(U)$, and then the corresponding ratio is 1.

RIPR, Anti-Simple Case If we set $\Pi_0 = \frac{\Sigma_q - \Sigma_p}{n}$ then by positive semidefinitess of $\Sigma_q - \Sigma_p$, the prior $W_0 = N(\boldsymbol{\mu}^*, \Pi_0)$ is well-defined and (3.4.6) holds and we can conclude (3.4.7) with prior W_1 that puts all mass on $\boldsymbol{\mu}^*$ (i.e. on Q). The result (3.2.10) and the fact that P_{W_1} is the RIPr then follow by the reasoning underneath (3.4.8).

3.4.3 Proof of Theorem 3

Finiteness of $D_R(Q\|P)$ is immediate from evaluating the definition.

UI (3.2.16) can be obtained from simple algebra from Proposition 5, combined with (3.4.12) and (3.4.11), the former applied with model $\mathcal{P}$ in the role of $\mathcal{Q}$, i.e. all q's replaced by p's (evidently it is still valid then), and using that $\mathbb{E}_R[\log(q_{\boldsymbol{\mu}^*}(U^{(n)})/p_{\boldsymbol{\mu}^*}(U^{(n)}))] = nD_R(Q_{\boldsymbol{\mu}^*}\|P_{\boldsymbol{\mu}^*})$. (3.2.17) can be obtained similarly from (3.4.12) and (3.4.13).

COND (3.2.19) is a direct consequence of (3.4.4) and (3.4.2).

seq-RIPr We 'sequentialize' the reasoning of Theorem 2, Part 3: first consider the case that $\Sigma_q - \Sigma_p$ is negative semidefinite. According to Theorem 1 in [40], this implies that the RIPr of $Q_{\breve{\boldsymbol{\mu}}_{|i-1}}(U_{(i)})$ onto $\textsc{conv}(\mathcal{P}(U_{(i)}))$ is an element of $\mathcal{P}$. More specifically, the theorem says it is $P_{\breve{\boldsymbol{\mu}}_{|i-1}}(U_{(i)})$, i.e. the element of $\mathcal{P}$ with the same mean. This proves the first part of the result. (3.2.20) then follows by applying (3.4.11) in Proposition 5 twice, first directly, with elements of $\mathcal{Q}$, and then again with $\mathcal{P}$ in the role of $\mathcal{Q}$, i.e. replacing all q's by p's, and then piecing together both equations by simple algebra.

Now consider the positive semidefinite case. The fact that $S^{(n)}_{\breve{\boldsymbol{\mu}},\text{SEQ-RIP}} = 1$ now follows because, at each i, using the result for the Bayes marginal distribution (3.4.5) applied with $n = 1$, we find that with the prior $W_0 = N(\breve{\boldsymbol{\mu}}_{|i-1}, \Sigma_q - \Sigma_p)$, we get $q_{\breve{\boldsymbol{\mu}}}(U_{(i)}) = p_{W_0}(U_{(i)})$. It follows that the sequential RIPr $P_{\leftsquigarrow q_{\breve{\boldsymbol{\mu}}_{|i-1}}(U_{(i)})}(U_{(i)})$ must be given by this very prior, i.e. $P_{\leftsquigarrow q_{\breve{\boldsymbol{\mu}}_{|i-1}}(U_{(i)})}(U_{(i)}) = P_{W_0}(U_{(i)})$, and then the corresponding ratio is 1. Since this holds for each i, the result follows.

RIPr, Anti-Simple Case All results follow analogously to the RIPr Anti-Simple Case of Theorem 2. In particular, (3.2.21) and (3.2.22) follow by the same reasoning, but now the prior W_1 is nondegenerate and we can conclude (3.4.7) from the fact that (3.4.8) (instead of (3.4.6)) holds.

3.5 Proofs for Section 3.3

3.5.1 Preparation for both Proofs

We start with Proposition 6, a direct asymptotic analogue of Proposition 5, for general exponential families. It will play an analogous role in the proofs. The one (important!) difference to Proposition 5 is that now W_1 is not required to be Gaussian. Note again the remarkable symmetry, now between (3.5.1) and (3.5.2).

Proposition 6. *Let $\mathcal{Q}$ be a regular exponential family as in Section 3.3. Extend the definition of $D(P_{\hat{\boldsymbol{\mu}}_{|n}}\|P_{\boldsymbol{\mu}^*})$ as in (3.B.2), Appendix 3.B.1, so that it is defined on the whole co-domain of $\hat{\boldsymbol{\mu}}_{|n}$. Further, suppose that the first m moments of X under R exist, with $m \geq 3$ odd, and $\mathbb{E}_R[X] = \boldsymbol{\mu}^* \in \mathtt{M}_q$. Let W_1 be a fixed distribution on $\mathtt{M}_q$*

*with density w that is strictly positive and continuous in a neighborhood of $\boldsymbol{\mu}^*$. If Condition 3 ('**plug-in**') holds, then we have:*

$$\mathbb{E}_{U^{(n)}\sim R}\left[D(Q_{\boldsymbol{\mu}^*}\|Q_{\breve{\boldsymbol{\mu}}_{|n}})\right] = \frac{1}{n}\frac{d_{rq}}{2} + O\left(n^{-5/4}\right) \text{ so that}$$

$$\mathbb{E}_{U,U^{(n)}\sim R}\left[n\cdot\log\left(q_{\boldsymbol{\mu}^*}(U)/q_{\breve{\boldsymbol{\mu}}_{|n}}(U)\right)\right] = \mathbb{E}_{U^{(n)}\sim R}\left[n\cdot D(Q_{\boldsymbol{\mu}^*}\|Q_{\breve{\boldsymbol{\mu}}_{|n}})\right] = \frac{d_{rq}}{2} + O(n^{-1/4}) \tag{3.5.1}$$

$$\text{and } \mathbb{E}_{U^{(n)}\sim R}\left[\log\frac{q_{\boldsymbol{\mu}^*}(U^{(n)})}{\prod_{i=1}^n q_{\breve{\boldsymbol{\mu}}_{|i-1}}(U_{(i)})}\right] = \sum_{i=1}^n \mathbb{E}_{U^{(i-1)}\sim R}\left[D(Q_{\boldsymbol{\mu}^*}\|Q_{\breve{\boldsymbol{\mu}}_{|i-1}})\right] = \frac{d_{rq}}{2}\log n + O(1).$$

We further have:

$$\mathbb{E}_{U^{(n)}\sim R}\left[\log\frac{q_{\boldsymbol{\mu}^*}(U^{(n)})}{q_{\hat{\boldsymbol{\mu}}_{|n}}(U^{(n)})}\right] = -\mathbb{E}_{U^{(n)}\sim R}\left[n\cdot D(Q_{\hat{\boldsymbol{\mu}}_{|n}}\|Q_{\boldsymbol{\mu}^*})\right] \leq -\frac{d_{rq}}{2} + o(1), \tag{3.5.2}$$

*where the inequality becomes an equality if further Condition 2 ('**UI**$^{\geq}$') holds. And finally, if R has density r and for some $\epsilon > 0$, $\mathbb{E}_R|\log q_{\boldsymbol{\mu}^*}(U)/r(U)|^{1+\epsilon} < \infty$, then*

$$\mathbb{E}_{U^{(n)}\sim R}\left[\log\frac{q_{\boldsymbol{\mu}^*}(U^{(n)})}{q_{W_1}(U^{(n)})}\right] = \frac{d}{2}\log\frac{n}{2\pi} - \frac{1}{2}\log\det\Sigma_q(\boldsymbol{\mu}^*) - \log w_1(\boldsymbol{\mu}^*) - \frac{d_{rq}(\boldsymbol{\mu}^*)}{2} + o(1). \tag{3.5.3}$$

Proof. The proposition arises as an immediate corollary of Lemma 2, which provides one-sided versions of the statements above. It is stated in Appendix 3.B.1 and proved in Appendix 3.B.3. To prove the corollary, take $\alpha = 1/4$ in Lemma 2 and simply piece together the right equations in each case. □

3.5.2 Proof of Theorem 4

UI This is proved in exactly the same way as the **UI** proof for the Gaussian case, Theorem 2, using Proposition 6 instead of Proposition 5.

COND With $Z := n^{1/2}\hat{\boldsymbol{\mu}}_{|n}$ defined as in the proof of Theorem 2 in the Gaussian case, We can write

$$S^{(n)}_{Q,\text{COND}} = \frac{q(U^{(n)} \mid Z)}{p(U^{(n)} \mid Z)} = \frac{q(U^{(n)})}{p_{\boldsymbol{\mu}^*}(U^{(n)})}\cdot\frac{p^\circ_{\boldsymbol{\mu}^*}(Z)}{q^\circ(Z)},$$

where p° and q° are the densities, under P and Q respectively, of $Z = \sqrt{n}\hat{\boldsymbol{\mu}}_{|n}$. Taking logarithms and expectation this gives

$$\mathbb{E}_R[\log S^{(n)}_{Q,\text{COND}}] = nD_R(Q\|P_{\boldsymbol{\mu}^*}) + \mathbb{E}_R\left[\log\frac{p^\circ_{\boldsymbol{\mu}^*}(Z)}{q^\circ(Z)}\right]. \tag{3.5.4}$$

The difference to the Gaussian case is that $p^{\circ}_{\boldsymbol{\mu}^*}$ and q° are now not Gaussian densities, but rather the densities or mass functions of two distributions which, by the Central Limit theorem (CLT), both converge to a normal distribution. This already suggests that the desired result (3.3.3) might hold, but the convergence implied by the standard CLT is too weak: we need convergence of the densities (not just the distributions) to the density of a normal. This is provided by the multivariate local central limit theorem. Even the standard version of this theorem is not enough: we need explicit bounds on the errors made in the approximation. These are provided by the classic work of Bhattacharya and Rao [16]. The actual proof, bounding the final term in (3.5.4) and thereby giving (3.3.3), follows by Lemma 5 provided in Appendix 3.B, which crucially uses the results from [16] and combines them with exponential concentration bounds.

seq-RIPr The simple case goes precisely as in the Gaussian case of Theorem 2.

It remains to prove (3.3.5) in the strict anti-simple case. Since $\Sigma_q - \Sigma_p(\boldsymbol{\mu}^*)$ is now assumed positive definite, it follows from [40, Proposition 2] that $q(U^{(n)})/p_{\boldsymbol{\mu}^*}(U^{(n)})$ is not an e-variable. Yet $P_{\boldsymbol{\mu}^*}$ achieves $\min_{\boldsymbol{\mu}\in \mathtt{M}_p} D(Q\|P_{\boldsymbol{\mu}})$. For such a combination, Theorem 1 in [42] gives that the infimum over all distributions $P \in \textsc{conv}(\mathcal{P}(U))$ of $D\left(Q(U)\|P(U)\right)$ lies outside $\mathcal{P}(U)$, which implies that

$$\mathbb{E}_Q\left[\log \frac{q(U)}{p_{\leftsquigarrow q(U)}(U)}\right] = \mathbb{E}_Q\left[\log \frac{q(U)}{p_{\boldsymbol{\mu}^*}(U)}\right] - \epsilon \tag{3.5.5}$$

for some $\epsilon > 0$. The result then follows by the definition (3.1.10).

RIPr, Anti-Simple Case (3.3.6) is based on the following:

Lemma 1. *In the setting of Theorem 4, the anti-simple case (i.e. $\Sigma_q - \Sigma_p(\boldsymbol{\mu}^*)$ is positive semidefinite), we have, with $W_{0,(n)}$ as in the theorem statement:*

$$D(Q(U^{(n)})\|P_{W_{0,(n)}}(U^{(n)}))] \leq nD(Q\|P_{\boldsymbol{\mu}^*}) - D_{\text{GAUSS}}(\Sigma_q\Sigma_p^{-1}) + o(1).$$

We place the proof in Appendix 3.B.5. Now note that

$$\begin{aligned}\mathbb{E}_Q\left[\log S^{(n)}_{Q,\text{COND}}\right] \leq \mathbb{E}_Q\left[\log S^{(n)}_{Q,\text{RIP}}\right] = \inf_W D\left(Q(U^{(n)})\middle|\middle|P_W(U^{(n)})\right) \\ \leq D\left(Q(U^{(n)})\middle|\middle|P_{W_{0,(n)}}(U^{(n)})\right),\end{aligned}$$

where the first inequality follows from (3.1.11), the infimum is over all priors on $\mathtt{M}_p$ and the equality follows from Theorem 1 in [42] (see also underneath (3.1.8)). The second inequality is immediate. Together with the equality it implies (3.3.6). The first equality together with (3.3.3) (use $R = Q$) implies that, if Condition 1 ('**COND**') holds, then (3.3.6) holds with equality.

3.5.3 Proof of Theorem 5

Part 1, **UI** is proved in exactly the same way as the **UI** proof for the Gaussian case, using Proposition 6 instead of Proposition 5. Part 2, **COND**, follows in exactly the same way as the corresponding result in Theorem 4. Part 3, **seq-RIPr**, the simple case, follows again in exactly the same way as the corresponding proof for the Gaussian case, **seq-RIPr** in Theorem 3, with all applications of Proposition 5 replaced by corresponding applications of Proposition 6.

seq-RIPr, Anti-Simple Case (3.3.15) follows by extending the proof of (3.3.5) with straightforward continuity arguments; details are in Appendix 3.B.6.

RIPr To prove (3.3.16), we equip $\mathtt{M}_p$ with the *same* prior W_1 as $\mathtt{M}_q$ (with density set to 0 on the set $\mathtt{M}_p \setminus \mathtt{M}_q$). We then use the familiar Laplace approximation of the Bayesian marginal likelihood given by (3.5.3), both for $\mathcal{Q}$ and $\mathcal{P}$. Thus, we apply these bounds twice, both times with $R = Q$ and then first with q_{W_1} and $q_{\boldsymbol{\mu}^*}$ and then with p_{W_1} and $p_{\boldsymbol{\mu}^*}$ respectively; the moment regularity condition automatically holds, because it reduces to the expectation under Q of an expression that is a quadratic form of X, and Q is a member of an exponential family with sufficient statistic X, implying that such expectations are finite. This gives that

$$\begin{aligned}\mathbb{E}_Q\left[\log \frac{q_{W_1}(U^{(n)})}{p_{W_1}(U^{(n)})}\right] &= nD(Q_{\boldsymbol{\mu}^*}\|P_{\boldsymbol{\mu}^*}) + \mathbb{E}_Q\left[\log \frac{q_{W_1}(U^{(n)})}{q_{\boldsymbol{\mu}^*}(U^{(n)})} + \log \frac{p_{\boldsymbol{\mu}^*}(U^{(n)})}{p_{W_1}(U^{(n)})}\right] \\ &= nD(Q_{\boldsymbol{\mu}^*}\|P_{\boldsymbol{\mu}^*}) + \frac{1}{2}\log \frac{\det \Sigma_q(\boldsymbol{\mu}^*)}{\det \Sigma_p(\boldsymbol{\mu}^*)} + \frac{d}{2} - \frac{d_{qp}(\boldsymbol{\mu}^*)}{2} + o(1),\end{aligned}$$

where in the second equality we used (3.5.3). The result follows by recognizing that the latter three terms together are equal to $-D_{\text{GAUSS}}(\Sigma_q\Sigma_p^{-1})$.

3.6 Implications, Conclusions and Future Work

3.6.1 Implications for E-Process-Ness

It is of interest to determine when a sequence of e-variables $(S^{(n)})_{n\geq 1}$ constitutes an e-process as defined in Section 3.1.2, since in that case we can use it in optional stopping scenarios. While $S^{(n)}_{\text{SEQ-RIP}}$ and $S^{(n)}_{\text{UI}}$ provide test martingales and hence e-processes by construction, *e-processness* is not so clear for S_{COND} and S_{RIP}. We present a final little result that enables us to determine some scenarios in which they do not provide e-processes.

Proposition 7. *Suppose $\mathcal{P} = \{P_\theta : \theta \in \Theta_0\}$ is a set of distributions (not necessarily an exponential family) for U, extended to n outcomes by independence, where P_θ has density p_θ, and Q is a distribution for random process $(U^{(n)})_{n\geq 0}$ with densities $(q(U^{(n)}))_{n\geq 0}$. Let $W_1, W_2, \ldots$ be a sequence of distributions on Θ such that for every n,* *we have that $S^{(n)} := q(U^{(n)})/p_{W_n}(U^{(n)})$ is an e-variable for samples of size n. Suppose that for some n,*

$$Q(P_{W_n}(U^{(n)}) \neq P_{W_{n+1}}(U^{(n)})) > 0. \tag{3.6.1}$$

Then the process $(S^{(n)})_{n \geq 0}$ is not an e-process relative to the data filtration.

Note that (3.6.1) implies that, but is stronger than, $W_n \neq W_{n+1}$. The proposition implies that if the process of RIPr e-variables $(S^{(n)}_{Q,\text{RIP}})_{n=1,2,\ldots}$ takes on the form $(P_{W_n}(U^{(n)}))_{n=1,2,\ldots}$ for some sequence $W_1, W_2, \ldots$ as above, then the RIPr does not yield an e-process relative to the data filtration. In particular, this will be the case in the Gaussian anti-simple cases of Theorem 2 and Theorem 3. In these cases, S_{RIP} is equal to S_{COND}, implying that S_{COND} also does not provide an e-process. Although we have no formal proof, the proof for the anti-simple case for general exponential family nulls with simple alternative (anti-simple case in Theorem 4) suggests that the same holds for S_{COND} and S_{RIP} in the *strict* anti-simple setting, since then the RIPr prior W_0 can be approximated by a Gaussian prior with variance of order $\asymp 1/n$, i.e. again changing over time.

However, an interesting phenomenon happens in the case of composite alternative $\mathcal{Q}$, both in the Gaussian and in the general case. First, in the anti-simple setting in the general case (Theorem 5, Part 4, anti-simple case) the RIPr prior W_0 is approximated by taking a prior *identical* to the prior W_1 that was put on $\mathcal{Q}$, and this prior does *not* change with the sample size. Similarly, in the Gaussian case with composite alternative, the exact RIPr prior changes with n, but only very minimally so. This suggests that somehow, with the conditional and RIPr e-variables we 'almost' obtain an e-process, leading perhaps to 'approximate' handling of optional stopping. Investigating this further and formalizing 'approximate optional stopping' is a main avenue for further research.

To strengthen this conjecture further, consider the sequence of e-variables $(S^{(n)}_{\text{HAAR}})_{n \geq 0}$ based on the right-Haar (uniform) prior on $\mathtt{M}_p$ and $\mathtt{M}_q$, as in Section 3.2.2. It is known that $(S^{(n)}_{\text{HAAR}})_{n \geq 0}$ defines an e-process, but relative to a *coarser* filtration generated by $(V^{(n)})_{n \geq 0}$ with $V^{(n)} = (X_{(2)} - X_{(1)}, \ldots, X_{(n)} - X_{(1)})$, which still allows optional stopping to all practical intents and purposes [45, 70]. Theorem 3 implies, by the equality of S_{HAAR} and S_{COND}, that, at least in the anti-simple case, when $\text{RIP} = \text{COND}$, since the prior W_0 in (3.2.21) depends on n, by Proposition 7, then S_{HAAR} does *not* define an e-process for the original, data-based filtration; but the fact that it *is* an e-process in a coarser filtration suggests that perhaps something similar holds ("being an approximate e-process in a coarser filtration"), for other conditional e-variables as well; the insights on asymptotic anytime-validity by [96] may be of use here.

Finally, we note that there do exist exponential family settings in which S_{COND} provides an e-process relative to the data filtration. This happens in the settings of Proposition 4, i.e. in Example 4, in which we are in the nonstrict simple and anti-simple settings at the same time.

3.6.2 Additional Future Work

Essentially all of our results except those for S_{UI} require the existence of a moment generating function for X under the alternative(s), implying that the distribution of X has exponentially small tails under the alternative. This is a strong regularity condition. Investigating whether this condition can be relaxed is an interesting goal for future work. This strong condition aside, we provided a general analysis of e-power for the most prominent e-variables that have been proposed for parametric family nulls: $S_{\text{RIP}}, S_{\text{UI}}, S_{\text{COND}}$ and $S_{\text{SEQ-RIP}}$. However, there exist at least two additional useful e-variables that we did not investigate. First, we mention the *sequential conditional e-variable*, which, with notation in spirit of Section 3.1.2, can be defined as:

$$S^{(n)}_{Q,\text{SEQ-COND}} = \prod_{i=1}^{n} \frac{q(U_{(i)} \mid U^{(i-1)}, X_{(i)})}{p(U_{(i)} \mid X_{(i)})}, \tag{3.6.2}$$

where q and p refer to the density of the conditional distribution of U given X, with $p(u|x)$ again identical for all $P \in \mathcal{P}$. $S_{Q,\text{SEQ-COND}}$ is the sequentialized version of S_{COND}: it applies conditioning for each outcome to reduce the null to a point null, for which the likelihood ratio q/p is the natural e-variable to use. $S_{Q,\text{SEQ-COND}}$ has been used in classical sequential testing for the contingency table setting [12] but can also be fruitfully used for k-sample tests with general exponential families, as suggested and analyzed by [44]. Second, we note that, for any alternative $\mathcal{Q}$ as in this paper and any prior W_1 on $\mathtt{M}_q$, the random variables

$$\mathbb{E}_{\boldsymbol{\mu}\sim W_1}\left[S^{(n)}_{Q_{\boldsymbol{\mu}},\text{RIP}}\right] \text{ and } \mathbb{E}_{\boldsymbol{\mu}\sim W_1}\left[S^{(n)}_{Q_{\boldsymbol{\mu}},\text{SEQ-RIP}}\right], \tag{3.6.3}$$

being weighted averages of e-variables, are themselves an e-variable (in the simple case they will both be equal to each other, as implied by Theorem 4, Part 3). It can be seen that these e-variable in general do *not* coincide with either $S^{(n)}_{W_1,\text{RIP}}$ or $S^{(n)}_{W_1,\text{SEQ-RIP}}$ since the order of averaging and taking projections has been reversed; in contrast, for UI we do have $\mathbb{E}_{\boldsymbol{\mu}\sim W_1}\left[S^{(n)}_{Q_{\boldsymbol{\mu}},\text{UI}}\right] = S^{(n)}_{W_1,\text{UI}}$. It might be interesting to further study the e-variables (3.6.2) and (3.6.3) in the exponential family setting of this paper.

3.7 Acknowledgements

Peter Grünwald is also affiliated with the Mathematical Institute of Leiden University. We had substantial, generous and much appreciated help from Michael Lindon (Netflix) with an earlier version of Theorem 1, Part 3. Shaul K. Bar-Lev and Gérard Letac pointed us towards the Landau family (see discussion underneath Condition 3). We would also like to thank Tyron Lardy for various discussions. This work is partly supported by the China Scholarship Council State Scholarship Fund Nr.202006280045.

Appendix 3.A Proofs underlying Theorem 2 and 3: the Gaussian Case

Proof of Proposition 5 We first prove (3.4.12). The first equality follows by simple algebra (it is also a consequence of the robustness property of exponential families, see Lemma 2 later on). We first establish that the first expression is equal to the third. Filling in the densities we find:

$$\begin{aligned}\mathbb{E}_R\left[\log q_{\hat{\boldsymbol{\mu}}_{|n}}(U^{(n)})/q_{\boldsymbol{\mu}^*}(U^{(n)})\right] =& \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}_R\left[(X_{(i)}-\boldsymbol{\mu}^*)^\top\Sigma_q^{-1}(X_{(i)}-\boldsymbol{\mu}^*)\right] \\ &- \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}_R\left[(X_{(i)}-\hat{\boldsymbol{\mu}}_{|n})^\top\Sigma_q^{-1}(X_{(i)}-\hat{\boldsymbol{\mu}}_{|n})\right].\end{aligned}$$

We have $\mathbb{E}_R[X_{(i)}-\boldsymbol{\mu}^*]=0$ and $\textsc{cov}_R(X_{(i)}-\boldsymbol{\mu}^*, X_{(i)}-\boldsymbol{\mu}^*)=\Sigma_r$. By the property of expectation of a quadratic form, we get that for the i-th term in the first sum above that it is equal to $\textsc{tr}(\Sigma_r\Sigma_q^{-1})$. It is easy to show $\mathbb{E}_R[X_{(i)}-\hat{\boldsymbol{\mu}}_{|n}]=0$ and $\textsc{cov}_R(X_{(i)}-\hat{\boldsymbol{\mu}}_{|n}, X_{(i)}-\hat{\boldsymbol{\mu}}_{|n})=\left(1-\frac{1}{n}\right)\Sigma_r$. Again by the property of expectation of a quadratic form, we thus get for the i-th term in the second sum:

$$\mathbb{E}_{X^{(n)}\sim R}\left[(X_{(i)}-\hat{\boldsymbol{\mu}}_{|n})^\top\Sigma_q^{-1}(X_{(i)}-\hat{\boldsymbol{\mu}}_{|n})\right]=\textsc{tr}\left(\left(1-\frac{1}{n}\right)\Sigma_r\Sigma_q^{-1}\right),$$

so that the right-hand side of (3.4.12) becomes $\frac{1}{2}\textsc{tr}(\Sigma_r\Sigma_q^{-1})$.

We proceed to prove (3.4.9). The first equality is straightforward. Filling in the densities, we find that it is equal to

$$\frac{1}{2}\mathbb{E}_{X,X^{(n)}\sim R}\left[(X-\breve{\boldsymbol{\mu}}_{|n})^\top\Sigma_q^{-1}(X-\breve{\boldsymbol{\mu}}_{|n})\right]-\frac{1}{2}\mathbb{E}_R\left[(X-\boldsymbol{\mu}^*)^\top\Sigma_q^{-1}(X-\boldsymbol{\mu}^*)\right].$$

We already showed the second term is equal to $(1/2)\textsc{tr}(\Sigma_r\Sigma_q^{-1})$. As to the first term, it is easy to show $\mathbb{E}_R[X-\breve{\boldsymbol{\mu}}_{|n}]=\frac{n_0(\boldsymbol{\mu}^*-x_0)}{n_0+n}$ and $\textsc{cov}_R(X-\breve{\boldsymbol{\mu}}_{|n}, X-\breve{\boldsymbol{\mu}}_{|n})=\left(1+\frac{n}{(n_0+n)^2}\right)\Sigma_r$. By the property of expectation of a quadratic form, we have

$$\begin{aligned}\mathbb{E}_{X\sim R}\left[(X-\breve{\boldsymbol{\mu}}_{|n})^\top\Sigma_q^{-1}(X-\breve{\boldsymbol{\mu}}_{|m})\right]=&\frac{n_0^2}{(n_0+n)^2}(\boldsymbol{\mu}^*-x_0)^\top\Sigma_q^{-1}(\boldsymbol{\mu}^*-x_0)\\ &+\textsc{tr}\left(\left(1+\frac{n}{(n_0+n)^2}\right)\Sigma_r\Sigma_q^{-1}\right),\end{aligned}$$

and from this the second equality in (3.4.9) follows. The third equality is immediate, and so is (3.4.10). (3.4.11) follows by repeatedly applying, for each i, the first equality in (3.4.9) and then the second equality in (3.4.9), both with i in the role of n.

We now prove (3.4.13). If $W_1 = N(\boldsymbol{\mu}_1, \Pi_1)$ is a normal prior, we find

$$D_R(Q_{\boldsymbol{\mu}^*}(Z)\|Q_{W_1}(Z)) = D_{\Sigma_r}(\Sigma_q\|(n\Pi_1+\Sigma_q)) + \frac{1}{2}(\boldsymbol{\mu}^*-\boldsymbol{\mu}_1)^\top(\Pi_1+\Sigma_q/n)^{-1}(\boldsymbol{\mu}^*-\boldsymbol{\mu}_1) \tag{3.A.1}$$

by combining (3.4.5) and (3.4.1) above (setting P in (3.4.1) to $Q_W(Z)$) . This gives the first equality in (3.4.13). In combination with (3.4.4) and (3.4.10), which we already proved, and using that

$$-\log w_1(\boldsymbol{\mu}^*) = \frac{1}{2}(\boldsymbol{\mu}^*-\boldsymbol{\mu}_1)^\top\Pi_1^{-1}(\boldsymbol{\mu}^*-\boldsymbol{\mu}_1) + \frac{d}{2}\log 2\pi + \frac{1}{2}\log\det\Pi_1,$$

we get the second equality, by plugging it into (3.A.1) directly; the $o(1)$ derives from $\Pi_1 + \frac{\Sigma_q}{n} \to \Pi_1$ as n increases.

Proof of (3.2.25) We have:

$$\begin{aligned}
&\mathbb{E}_{U^{(n)}\sim R}[\log S^{(n)}_{\text{HAAR}}] = (n-1)D_{\Sigma_r}(\Sigma_q\|\Sigma_p) + \\
&\mathbb{E}_{X_{(1)}\sim R}\left[\mathbb{E}_{X_{(2)},\ldots,X_{(n)}}\left[\log\frac{q_{W_1|X_{(1)}}(X_{(2)},\ldots,X_{(n)})}{q_{\boldsymbol{\mu}^*}(X_{(2)},\ldots,X_{(n)})} - \log\frac{p_{W_0|X_{(1)}}(X_{(2)},\ldots,X_{(n)})}{p_{\boldsymbol{\mu}^*}(X_{(2)},\ldots,X_{(n)})}\right]\right] \\
&= \mathbb{E}_R\left[-D_{\Sigma_r}(\Sigma_q\|((n-1)\Sigma_q+\Sigma_q)) - \frac{1}{2}(\boldsymbol{\mu}^*-X_{(1)})^\top(\Sigma_q+\Sigma_q/(n-1))^{-1}(\boldsymbol{\mu}^*-X_{(1)})\right] + \\
&\mathbb{E}_R\left[D_{\Sigma_r}(\Sigma_p\|((n-1)\Sigma_p+\Sigma_p)) + \frac{1}{2}(\boldsymbol{\mu}^*-X_{(1)})^\top(\Sigma_p+\Sigma_p/(n-1))^{-1}(\boldsymbol{\mu}^*-X_{(1)})\right] + \\
&(n-1)D_{\Sigma_r}(\Sigma_q\|\Sigma_p) \qquad\qquad (3.A.2) \\
&= -D_{\Sigma_r}(\Sigma_q\|n\Sigma_q) - \frac{1}{2}\text{TR}(\Sigma_r(\Sigma_q+\Sigma_q/(n-1))^{-1}) \\
&+ D_{\Sigma_r}(\Sigma_p\|n\Sigma_p) + \frac{1}{2}\text{TR}(\Sigma_r(\Sigma_p+\Sigma_p/(n-1))^{-1}) + (n-1)D_{\Sigma_r}(\Sigma_q\|\Sigma_p) \\
&= -\frac{1}{2}\text{TR}(\Sigma_r((n\Sigma_q)^{-1}-\Sigma_q^{-1}) - \frac{1}{2}\text{TR}(\Sigma_r(\Sigma_q\cdot n/(n-1))^{-1}) \\
&+ \frac{1}{2}\text{TR}(\Sigma_r((n\Sigma_p)^{-1}-\Sigma_p^{-1}) + \frac{1}{2}\text{TR}(\Sigma_r(\Sigma_p\cdot n/(n-1))^{-1}) + (n-1)D_{\Sigma_r}(\Sigma_q\|\Sigma_p) \\
&= (n-1)D_{\Sigma_r}(\Sigma_q\|\Sigma_p),
\end{aligned}$$

where for (3.A.2) we used (3.4.13) in Proposition 5, and in the final equation we used definition (3.2.5).

Appendix 3.B Proofs underlying Theorem 4 and 5: the general case

3.B.1 Definitions for and Statement of Lemma 2

Condition 2 and Lemma 2 refer to the KL divergence $D(P_{\hat{\boldsymbol{\mu}}_{|n}} \| P_{\boldsymbol{\mu}^*})$, which is undefined if $\hat{\boldsymbol{\mu}}_{|n} \notin \mathtt{M}_p$ (which may happen even for the simple Bernoulli model if $\hat{\boldsymbol{\mu}}_{|n} \in \{0,1\}$). It is highly convenient to extend its definition to such cases, and this can be done in a straightforward manner. We first set $\bar{\mathtt{M}}_p$ to be the union of the mean-value parameter space $\mathtt{M}_p$ of $\mathcal{P}$ and the set of values that $\hat{\boldsymbol{\mu}}_{|n} = n^{-1} \sum X_{(i)}$ can take. For all $\hat{\boldsymbol{\mu}}_{|n} \in \bar{\mathtt{M}}_p \setminus \mathtt{M}_p$, we first set

$$\frac{p_{\hat{\boldsymbol{\mu}}_{|n}}(U^{(n)})}{p_{\boldsymbol{\mu}^*}(U^{(n)})} := \sup_{\boldsymbol{\mu} \in \mathtt{M}_p} \frac{p_{\boldsymbol{\mu}}(U^{(n)})}{p_{\boldsymbol{\mu}^*}(U^{(n)})}, \tag{3.B.1}$$

and we note that then, in terms of the canonical parameterization, by (3.1.7), with $\boldsymbol{\beta}^*$ the canonical parameter corresponding to $\boldsymbol{\mu}^*$ and $X_{(i)} = (t_1(U_{(i)}), \ldots, t_d(U_{(i)}))^\top$,

$$\log \frac{p_{\hat{\boldsymbol{\mu}}_{|n}}(U^{(n)})}{p_{\boldsymbol{\mu}^*}(U^{(n)})} = \sup_{\boldsymbol{\beta} \in \mathtt{B}_p} (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top \sum_{i=1}^n X_{(i)} - \log(Z_p(\boldsymbol{\beta})/Z_p(\boldsymbol{\beta}^*)),$$

which can be written as a function of $\hat{\boldsymbol{\mu}}_{|n}$. Therefore, for $\hat{\boldsymbol{\mu}}_{|n} \in \bar{\mathtt{M}}_p \setminus \mathtt{M}_p$, we can unambiguously set

$$D(P_{\hat{\boldsymbol{\mu}}_{|n}} \| P_{\boldsymbol{\mu}^*}) := \frac{1}{n} \log \frac{p_{\hat{\boldsymbol{\mu}}_{|n}}(U^{(n)})}{p_{\boldsymbol{\mu}^*}(U^{(n)})}, \tag{3.B.2}$$

since the expression on the right only depends on $U^{(n)}$ through $\hat{\boldsymbol{\mu}}_{|n}$. The rationale underlying definition (3.B.2) is that (3.B.2) holds any way as long as $\hat{\boldsymbol{\mu}}_{|n} \in \mathtt{M}_p$. This is a well-known result, with straightforward proof, sometimes referred to as the *KL robustness property for exponential families* [35, Chapter 19]. We have thus extended it to hold for all $\hat{\boldsymbol{\mu}}_{|n} \in \bar{\mathtt{M}}_p$ by definition (3.B.1), and thereby already proved the first statement in Lemma 2 below.

We only give the remaining results in the lemma under existence of an odd number of moments m under R, since, for even m', if X has m' moments then it also has $m = m' - 1$ moments and due to certain cancellations, the result one obtains with m' in the proof would not be stronger than what one obtains for m.

Lemma 2. *First, extend the definition of $D(P_{\hat{\boldsymbol{\mu}}_{|n}} \| P_{\boldsymbol{\mu}^*})$ as above. Then the robustness property for exponential families holds for all $u^{(n)} \in \mathcal{U}^n$:*

$$nD(P_{\hat{\boldsymbol{\mu}}_{|n}} \| P_{\boldsymbol{\mu}^*}) = \log \frac{p_{\hat{\boldsymbol{\mu}}_{|n}}(u^{(n)})}{p_{\boldsymbol{\mu}^*}(u^{(n)})}. \tag{3.B.3}$$

Suppose that the first m moments of X under R exist, with $m \geq 3$ odd, and $\mathbb{E}_R[X] = \boldsymbol{\mu}^*$.

Fix $0 < \alpha < 1/2$. Let Σ_p be a general $d \times d$ positive definite matrix. We have:

$$\mathbb{E}_R\left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n}-\boldsymbol{\mu}^*\|_2\geq n^{\alpha-1/2}}\cdot n\cdot\frac{1}{2}(\hat{\boldsymbol{\mu}}_{|n}-\boldsymbol{\mu}^*)^\top\Sigma_p^{-1}(\hat{\boldsymbol{\mu}}_{|n}-\boldsymbol{\mu}^*)\right] = O\left(n^{-(m-2)\alpha-1/2}\right), \tag{3.B.4}$$

$$\mathbb{E}_R\left[n\cdot\frac{1}{2}(\hat{\boldsymbol{\mu}}_{|n}-\boldsymbol{\mu}^*)^\top\Sigma_p^{-1}(\boldsymbol{\mu}^*)(\hat{\boldsymbol{\mu}}_{|n}-\boldsymbol{\mu}^*)\right] = \frac{\textsc{tr}(\Sigma_r\Sigma_p^{-1}(\boldsymbol{\mu}^*))}{2}. \tag{3.B.5}$$

Next, let $\mathcal{P}$ be a regular exponential family with $\boldsymbol{\mu}^* \in \mathtt{M}_p$. Fix $c > 0$ and let $\mathtt{M}' \subset \mathtt{M}_p$ be an open neighborhood of $\boldsymbol{\mu}^*$. Let $(W_{0,(n)})_{n\in\mathbb{N}}$ be any sequence of distributions on $\mathtt{M}_p$ with continuous densities $w_{0,(n)}$ such that for every n, $w_{0,(n)}(\boldsymbol{\mu}) > c$ for all $\boldsymbol{\mu} \in \mathtt{M}'$. We have:

$$\mathbb{E}_R\left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n}-\boldsymbol{\mu}^*\|_2\geq n^{\alpha-1/2}}\cdot\log\frac{p_{\boldsymbol{\mu}^*}(U^{(n)})}{p_{W_{0,(n)}}(U^{(n)})}\right] \leq O(n^{-(m-2)\alpha-1/2}), \tag{3.B.6}$$

and, if $W_{0,(n)}$ is as above and for some $\epsilon > 0$, $\mathbb{E}_R|\log p_{\boldsymbol{\mu}^*}(U)/r(U)|^{1+\epsilon} < \infty$, then

$$\mathbb{E}_R\left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n}-\boldsymbol{\mu}^*\|_2\geq n^{\alpha-1/2}}\cdot\log\frac{p_{W_{0,(n)}}(U^{(n)})}{p_{\boldsymbol{\mu}^*}(U^{(n)})}\right] \leq O(n^{-\alpha m\cdot(1\wedge\epsilon)}). \tag{3.B.7}$$

*Now let $\breve{\boldsymbol{\mu}}_{|n}$ be defined as in (3.2.13), with $n_0 > 0$. If (c) in Condition 3 ('**plug-in**') holds (with P. in the role of Q.), then*

$$\mathbb{E}_R\left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n}-\boldsymbol{\mu}^*\|_2\geq n^{\alpha-1/2}}\cdot n\cdot D(P_{\boldsymbol{\mu}^*}\|P_{\breve{\boldsymbol{\mu}}_{|n}})\right] = O(n^{-m\alpha+1/2}). \tag{3.B.8}$$

With

$$f(n) := \mathbb{E}_R\left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n}-\boldsymbol{\mu}^*\|_2< n^{\alpha-1/2}}\cdot n\cdot\frac{1}{2}(\hat{\boldsymbol{\mu}}_{|n}-\boldsymbol{\mu}^*)^\top\Sigma_p^{-1}(\boldsymbol{\mu}^*)(\hat{\boldsymbol{\mu}}_{|n}-\boldsymbol{\mu}^*)\right],$$

we have, with W now a fixed distribution on $\mathtt{M}_p$ with density w that is positive and continuous in a neighborhood of $\boldsymbol{\mu}^*$,

$$\mathbb{E}_R\left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n}-\boldsymbol{\mu}^*\|_2< n^{\alpha-1/2}}\cdot n\cdot D(P_{\hat{\boldsymbol{\mu}}_{|n}}\|P_{\boldsymbol{\mu}^*})\right] = f(n) + O\left(n^{\alpha-1/2}\right), \tag{3.B.9}$$

$$\mathbb{E}_R\left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n}-\boldsymbol{\mu}^*\|_2< n^{\alpha-1/2}}\cdot n\cdot D(P_{\boldsymbol{\mu}^*}\|P_{\breve{\boldsymbol{\mu}}_{|n}})\right] = f(n) + O\left(n^{\alpha-1/2}\right), \tag{3.B.10}$$

$$\begin{aligned}&\mathbb{E}_R\left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n}-\boldsymbol{\mu}^*\|_2< n^{\alpha-1/2}}\cdot\log\frac{p_{\hat{\boldsymbol{\mu}}_{|n}}(U^{(n)})}{p_W(U^{(n)})}\right] \\ =&\frac{d}{2}\log\frac{n}{2\pi} - \log w(\boldsymbol{\mu}^*) - \frac{1}{2}\log\det\Sigma_p(\boldsymbol{\mu}^*) + o(1).\end{aligned} \tag{3.B.11}$$

3.B.2 Preparatory Results for Lemma 2 and Lemma 1

Throughout the proofs for Lemma 2 and Lemma 1 and Theorem 4, Part 3 (Appendix 3.B.4), we will use the following additional notations and abbreviations:

$$\boldsymbol{\mu}^* = \mathbf{0}, \Sigma_p := \Sigma_p(\mathbf{0}), \Sigma_q := \Sigma_q(\mathbf{0}), K := (\Sigma_q - \Sigma_p)^{-1}, J(\boldsymbol{\mu}) := \Sigma_p(\boldsymbol{\mu})^{-1}, J := J(\mathbf{0}) = \Sigma_p^{-1}. \tag{3.B.12}$$

Note that the first equation, $\boldsymbol{\mu}^* = \mathbf{0}$ — which will allow substantially shortening equations — is really a definition rather than a notation; but we can enforce this without loss of generality — since $\boldsymbol{\mu}^*$ is fixed throughout all proofs, we can get the same result as for $\mathbf{0}$ with arbitrary $\boldsymbol{\mu}^*$ simply by adding $\boldsymbol{\mu}^*$ as a constant to each outcome X and transforming the resulting equations. Or, alternatively, one may modify all equations by putting in '$-\boldsymbol{\mu}^*$' at the appropriate places and note that nothing in the derivation changes.

We will make repeated use of the following basic results whose proof can be found in, for example, [38].

Lemma 3. *Fix $n_0 \geq 0$ and $x_0 \in \mathbb{R}$. Let $X, X_1, \ldots$ be (scalar-valued) i.i.d. random variables, define $\breve{\mu}_{|n} := (n_0 \cdot x_0 + \sum_{i=1}^{n} X_i)/(n + n_0)$. Suppose the first $m \in \mathbb{N}$ moments of X exist and let $\delta > 0$. Then $\Pr(|\breve{\mu}_n - \mathbb{E}[X]| \geq \delta) = O(n^{-\lceil \frac{m}{2} \rceil} \delta^{-m})$. As a consequence, via the union bound, for fixed d, our d-dimensional regularized MLE $\breve{\boldsymbol{\mu}}_{|n}$ also satisfies $\Pr(\|\breve{\boldsymbol{\mu}}_{|n} - \boldsymbol{\mu}^*\|_2 \geq \delta) = O(n^{-\lceil \frac{m}{2} \rceil} \delta^{-m})$.*

Lemma 4. *Consider the setting of Lemma 3. If the first m moments of (scalar) X exist, then $\mathbb{E}\left[\left(\breve{\mu}_{|n} - \mathbb{E}[X]\right)^m\right] = O(n^{-\lceil \frac{m}{2} \rceil})$.*

We also need the following two propositions:

Proposition 8. *Let $0 < \alpha < 1/2$ and fix $c > 0$. Then there is a constant $C > 0$ such that for all large n, for all $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_d)$, $\boldsymbol{\mu}' = (\boldsymbol{\mu}'_1, \ldots, \boldsymbol{\mu}'_d)$ with $(\|\boldsymbol{\mu}\|_2 \vee \|\boldsymbol{\mu}'\|_2) < n^{\alpha - 1/2}$, we have*

$$\left| D(P_{\boldsymbol{\mu}'} \| P_{\boldsymbol{\mu}}) - \frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}')^\top J (\boldsymbol{\mu} - \boldsymbol{\mu}') \right| \leq C \sum_{i,j,k \in [d]} |\mu_i - \mu'_i| \cdot |\mu_j - \mu'_j| \cdot |\mu_k - \mu'_k| \leq C n^{3\alpha - 3/2}.$$

Proof. Note that for all n large enough, with $\mathcal{A}_n := \{\boldsymbol{\mu} : \|\boldsymbol{\mu}\|_2 < n^{\alpha - 1/2}\}$ and $\mathcal{A}'_n := \{\boldsymbol{\mu} : \|\boldsymbol{\mu}\|_2 < n^{\alpha - 1/2}\}$, we have that $\mathcal{A}_n \cup \mathcal{A}'_n \in \mathtt{M}_p$. A third order Taylor-expansion of $D(P_{\mathbf{0}} \| P_{\boldsymbol{\mu}})$ in terms of $\boldsymbol{\mu}$ then gives that:

$$D(P_{\boldsymbol{\mu}'} \| P_{\boldsymbol{\mu}}) = \frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}')^\top J (\boldsymbol{\mu} - \boldsymbol{\mu}') + O\left(\sum_{i,j,k \in [d]} |C_{i,j,k}| |\mu_i - \mu'_i| \cdot |\mu_j - \mu'_j| \cdot |\mu_k - \mu'_k| \right),$$

where $C_{i,j,k}$ are the corresponding third-order derivatives. But since, for large n, $\mathcal{A}_n$ is a closed set in the interior of $\mathtt{M}_p$, these derivatives are bounded; the result follows. □

Proposition 9. *We have:*

$$\sup_{U^n : \|\hat{\boldsymbol{\mu}}_{|n}\|_2 < n^{\alpha-1/2}} \left| D(P_{\mathbf{0}} \| P_{\hat{\boldsymbol{\mu}}_{|n}}) - D(P_{\mathbf{0}} \| P_{\breve{\boldsymbol{\mu}}_{|n}}) \right| = O(n^{\alpha-3/2}).$$

Proof. Define $f(\boldsymbol{\mu}) := D(P_{\mathbf{0}} \| P_{\boldsymbol{\mu}})$ and $g(\boldsymbol{\mu}) := \nabla f(\boldsymbol{\mu})$. A first-order Taylor approximation gives that

$$D(P_{\mathbf{0}} \| P_{\breve{\boldsymbol{\mu}}_{|n}}) = D(P_{\mathbf{0}} \| P_{\hat{\boldsymbol{\mu}}_{|n}}) + (\breve{\boldsymbol{\mu}}_{|n} - \hat{\boldsymbol{\mu}}_{|n})^\top g(\boldsymbol{\mu}') = D(P_{\mathbf{0}} \| P_{\hat{\boldsymbol{\mu}}_{|n}}) + \frac{n_0}{n+n_0} (x_0 - \hat{\boldsymbol{\mu}}_{|n})^\top \cdot g(\boldsymbol{\mu}'), \tag{3.B.13}$$

where $\boldsymbol{\mu}'$ is some point with $\|\boldsymbol{\mu}'\|_2 < n^{\alpha-1/2}$, and the second equation is straightforward rewriting.

We can write $g(\boldsymbol{\mu}) = (g_1(\boldsymbol{\mu}), \ldots, g_d(\boldsymbol{\mu}))^\top$ with $g_j(\boldsymbol{\mu}) = (\partial / \partial \boldsymbol{\mu}_j) f(\boldsymbol{\mu})$. Taylor approximating $g_j(\boldsymbol{\mu})$ itself gives for each $j = 1..d$:

$$g_j(\boldsymbol{\mu}') = g_j(\mathbf{0}) + \left(\frac{\partial}{\partial \boldsymbol{\mu}_1} g_j(\boldsymbol{\mu}^\circ_{[j]}), \ \ldots \ , \frac{\partial}{\partial \boldsymbol{\mu}_d} g_j(\boldsymbol{\mu}^\circ_{[j]}) \right)^\top \boldsymbol{\mu}'$$

for d points $\boldsymbol{\mu}^\circ_{[j]} = (\boldsymbol{\mu}^\circ_{[j],1}, \ldots, \boldsymbol{\mu}^\circ_{[j],d})^\top$, where, for $j = \{1, \ldots, d\}$, we have $\|\boldsymbol{\mu}^\circ_{[j]}\|_2 < n^{1/2-\alpha}$. But since $\frac{\partial}{\partial \boldsymbol{\mu}_j} g_j(\boldsymbol{\mu}) = \frac{\partial^2}{\partial \boldsymbol{\mu}_j^2} f(\boldsymbol{\mu})$ and for all large n, the set $\{\boldsymbol{\mu} : \|\boldsymbol{\mu}\|_2 < n^{1/2-\alpha}\} \in \mathtt{M}_p$, these d second-derivatives are uniformly bounded. Also, $g(\mathbf{0}) = 0$ since the KL divergence is minimized at $\mathbf{0}$ and, for exponential families, is a smooth function in the mean-value parameters. It follows that the final term in (3.B.13) is $O(n^{-1} n^{\alpha-1/2})$, and the result follows. □

3.B.3 Proof of Lemma 2

Throughout the proof, we adopt the notations of (3.B.12). (3.B.3) was already proven. (3.B.11) is a well-known result, see for example [35, Chapter 8]. (3.B.5) is also known; basically, the expectation on the right can be computed exactly like we did in the proof of Proposition 5, yielding the desired result.

Proof of (3.B.4) In a variation of the proof of Markov's inequality, we can write the left-hand side, for fixed $\gamma > 0$, as

$$\begin{aligned}
&\mathbb{E}_R \left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n}\|_2 \geq n^{\alpha-1/2}} \cdot \frac{(\hat{\boldsymbol{\mu}}_{|n}^\top \Sigma_p^{-1} \hat{\boldsymbol{\mu}}_{|n})^{\gamma+1}}{(\hat{\boldsymbol{\mu}}_{|n}^\top \Sigma_p^{-1} \hat{\boldsymbol{\mu}}_{|n})^{\gamma}} \right] \leq C \cdot \mathbb{E}_R \left[\frac{(\hat{\boldsymbol{\mu}}_{|n}^\top \Sigma_p^{-1} \hat{\boldsymbol{\mu}}_{|n})^{\gamma+1}}{n^{\gamma(2\alpha-1)}} \right] \\
&= C \cdot \mathbb{E}_R \left[\frac{(\hat{\boldsymbol{\mu}}_{|n}^\top \Sigma_p^{-1} \hat{\boldsymbol{\mu}}_{|n})^{m/2}}{n^{m\alpha - m/2 - 2\alpha + 1}} \right] = O\left(n^{-m\alpha + m/2 + 2\alpha - 1} \cdot n^{-(m+1)/2} \right) \\
&= O\left(n^{-m\alpha + 2\alpha - 3/2} \right) = O\left(n^{-(m-2)\alpha - 3/2} \right)
\end{aligned}$$

for some constant $C > 0$, where we used that Σ_p^{-1} is positive definite and fixed, independent of $\hat{\boldsymbol{\mu}}_{|n}$. In the second line we set $\gamma = m/2 - 1$, using that the m-th moment of the expectation exists, and then again the fact that Σ_p^{-1} is positive definite and Lemma 4, using that m is odd. (3.B.4) follows by multiplying left- and right-hand side by n.

Proof of (3.B.6) Assume without loss of generality that n is large enough so that $\mathcal{A}_n := \{\boldsymbol{\mu} : \|\boldsymbol{\mu}\|_2 \leq n^{-1/2}\}$ is contained in $\mathtt{M}_p$ and for some constant $c > 0$, $w_{(0),n}(\boldsymbol{\mu}) \geq c$ for all $\boldsymbol{\mu} \in \mathcal{A}_n$. Then

$$-\log \frac{\int p_{\boldsymbol{\mu}}(U^{(n)}) w_{0,(n)}(\boldsymbol{\mu}) d\boldsymbol{\mu}}{p_{\mathbf{0}}(U^{(n)})} \leq -\log \frac{\int \mathbf{1}_{\mathcal{A}_n} \cdot \left(\min_{\boldsymbol{\mu}' \in \mathcal{A}_n} p_{\boldsymbol{\mu}'}(U^{(n)})\right) w_{0,(n)}(\boldsymbol{\mu}) d\boldsymbol{\mu}}{p_{\mathbf{0}}(U^{(n)})} \tag{3.B.14}$$

$$= \left(\max_{\boldsymbol{\mu} \in \mathcal{A}_n} \ -\log \frac{p_{\boldsymbol{\mu}}(U^{(n)})}{p_{\mathbf{0}}(U^{(n)})}\right) - \log W_{0,(n)}(\mathcal{A}_n) = n \cdot \max_{\boldsymbol{\mu} \in \mathcal{A}_n} f(\hat{\boldsymbol{\mu}}_{|n}, \boldsymbol{\mu}) + O\left(\frac{d}{2} \log n\right),$$

where

$$f(\boldsymbol{\mu}^\circ, \boldsymbol{\mu}) = D(P_{\boldsymbol{\mu}^\circ} \| P_{\boldsymbol{\mu}}) - D(P_{\boldsymbol{\mu}^\circ} \| P_{\mathbf{0}}),$$

and we used the robustness property of exponential families (3.B.3) in the final step.

We now move to the canonical parameterization to further analyze the function f. Let $\boldsymbol{\beta}_{\boldsymbol{\mu}}$ be the function mapping mean-value parameter vector $\boldsymbol{\mu}$ to the corresponding canonical parameter vector. For general $\boldsymbol{\mu}^\circ \in \mathtt{M}_p$, we can write

$$\begin{aligned} &\nabla_{\boldsymbol{\mu}} f(\boldsymbol{\mu}^\circ; \boldsymbol{\mu}) = \nabla_{\boldsymbol{\mu}} \left(D(P_{\boldsymbol{\mu}^\circ} \| P^{\text{CAN}}_{\boldsymbol{\beta}_{\boldsymbol{\mu}}}) - D(P_{\boldsymbol{\mu}^\circ} \| P_{\mathbf{0}})\right) = \\ &\nabla_{\boldsymbol{\mu}} \left(-(\boldsymbol{\beta}_{\boldsymbol{\mu}} - \boldsymbol{\beta}_{\mathbf{0}})^\top \boldsymbol{\mu}^\circ + \log Z(\boldsymbol{\beta}_{\boldsymbol{\mu}})\right) = J(\boldsymbol{\mu}) \cdot (\boldsymbol{\mu} - \boldsymbol{\mu}^\circ), \end{aligned}$$

which can be found by the fact that $J(\boldsymbol{\mu})$ is the matrix of partial derivatives of $\boldsymbol{\beta}_{\boldsymbol{\mu}}$ as a function of $\boldsymbol{\mu}$, and that $\boldsymbol{\mu}$ is the gradient of $\log Z(\boldsymbol{\beta})$, and using the chain rule of differentiation. A first order Taylor approximation of $f(\hat{\boldsymbol{\mu}}_{|n}, \boldsymbol{\mu})$ around $\boldsymbol{\mu} = 0$, and then bounding its value on the set $\{\boldsymbol{\mu} : ||\boldsymbol{\mu}||_2 \leq n^{-1/2}\}$, now gives:

$$\begin{aligned} n \cdot f(\hat{\boldsymbol{\mu}}_{|n}, \boldsymbol{\mu}) &= -\, n\boldsymbol{\mu}^\top J(\boldsymbol{\mu}') \cdot (\boldsymbol{\mu}' - \hat{\boldsymbol{\mu}}_{|n}) \\ &= -\, n\boldsymbol{\mu}^\top J(\boldsymbol{\mu}') c\boldsymbol{\mu} + n\boldsymbol{\mu}^\top J(\boldsymbol{\mu}') \cdot \hat{\boldsymbol{\mu}}_{|n} \leq n\boldsymbol{\mu}^\top J(\boldsymbol{\mu}') \cdot \hat{\boldsymbol{\mu}}_{|n} \\ &= O(n\boldsymbol{\mu}^\top \cdot \hat{\boldsymbol{\mu}}_{|n}) = O(n \cdot \|\boldsymbol{\mu}\|_2 \cdot \|\hat{\boldsymbol{\mu}}_{|n}\|_2) = O(\sqrt{n} \cdot \|\hat{\boldsymbol{\mu}}_{|n}\|_2), \end{aligned}$$

for $\boldsymbol{\mu}' = c\boldsymbol{\mu}$, for some $c \in [0, 1]$. Here we used that the maximum eigenvalue of $J(\boldsymbol{\mu}')$ is bounded away from 0 since $\boldsymbol{\mu}$ is inside a compact set in the interior of the parameter space, and we used Cauchy-Schwartz.

This gives that (3.B.14) is $O(\sqrt{n} \cdot \|\hat{\boldsymbol{\mu}}_{|n}\|_2 + (d/2) \log n)$. On the set with $\|\hat{\boldsymbol{\mu}}_{|n}\|_2 \geq n^{\alpha - 1/2}$, we can weaken this to give that (3.B.14) is $O(n \cdot \|\hat{\boldsymbol{\mu}}_{|n}\|_2^2)$. We can now bound this further using (3.B.4) from Lemma 2, and the result follows.

Proof of (3.B.7) Let $\mathcal{A}_n$ denote the event that $\|\hat{\boldsymbol{\mu}}_{|n} - \boldsymbol{\mu}^*\|_2 \geq n^{\alpha - 1/2}$. We have, with $W = W_{0,(n)}$:

$$\mathbb{E}_R\left[\mathbf{1}_{\mathcal{A}_n} \cdot \log \frac{p_W(U^{(n)})}{p_{\boldsymbol{\mu}^*}(U^{(n)})}\right] = \mathbb{E}_R\left[\mathbf{1}_{\mathcal{A}_n} \cdot \log \frac{p_W(U^{(n)})}{r(U^{(n)})}\right] + \mathbb{E}_R\left[\mathbf{1}_{\mathcal{A}_n} \cdot \log \frac{r(U^{(n)})}{p_{\boldsymbol{\mu}^*}(U^{(n)})}\right]. \tag{3.B.15}$$

By Lemma 3 we have $R(\mathcal{A}_n) = O(n^{(1/2-\alpha)m} \cdot n^{-(m+1)/2} = O(n^{-\alpha m - 1/2})$, which we use to further bound the first term by

$$\begin{aligned}&R(\mathcal{A}_n) \cdot \mathbb{E}_R\left[\mathbf{1}_{\mathcal{A}_n} \cdot \log \frac{p_W(U^{(n)} \mid \mathcal{A}_n)}{r(U^{(n)} \mid \mathcal{A}_n)} \mid \mathcal{A}_n\right] + R(\mathcal{A}_n) \log \frac{P_W(\mathcal{A}_n)}{R(\mathcal{A}_n)} \leq \\ &R(\mathcal{A}_n) \log(1/R(\mathcal{A}_n)) = O(n^{-\alpha m}),\end{aligned}$$

where the penultimate inequality follows by Jensen and the fact that $P_W(\mathcal{A}_n) \leq 1$. The second term in (3.B.15) can be bounded by Hölder's inequality, for arbitrary $\epsilon > 0$, as

$$(R(\mathcal{A}_n))^{\epsilon/(1+\epsilon)} \cdot \left(\mathbb{E}_R\left(\log \frac{r(U^{(n)})}{p_{\boldsymbol{\mu}^*}(U^{(n)})}\right)^{1+\epsilon}\right)^{1/(1+\epsilon)} = O(n^{(-\alpha m - 1/2)\epsilon}),$$

and it can be seen that both terms in (3.B.15) go to 0 with increasing n. The result follows.

Proof of (3.B.8) Fix $A > 0$ as in Condition 3 ('**plug-in**') with P in the role of Q. Define $\mathcal{E}_j = \{\boldsymbol{\mu} \in \bar{\mathtt{M}}_p : \|\boldsymbol{\mu}\|_2 \in [A + j - 1, A + j]\}$. With $\alpha = n_0/(n + n_0)$, we have:

$$\begin{aligned}&\mathbb{E}_R\left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n} - \boldsymbol{\mu}^*\|_2 \geq A} \cdot D(P_{\boldsymbol{\mu}^*} \| P_{\breve{\boldsymbol{\mu}}_{|n}})\right] \leq \sum_{j \in \mathbb{N}} R(\hat{\boldsymbol{\mu}}_{|n} \in \mathcal{E}_j) \max_{\hat{\boldsymbol{\mu}}_{|n} \in \mathcal{E}_j} D(P_{\boldsymbol{\mu}^*} \| P_{(1-\alpha)\hat{\boldsymbol{\mu}}_{|n} + \alpha x_0}) \\ &\overset{\text{(a)}}{=} \sum_{j \in \mathbb{N}} R(\hat{\boldsymbol{\mu}}_{|n} \in \mathcal{E}_j) O\left((A + j)^{m-s}\right) \leq \sum_{j \in \mathbb{N}} R\left(\|\hat{\boldsymbol{\mu}}_{|n}\|_2 \geq A + j - 1\right) \cdot O\left((A + j)^{m-s}\right) \\ &\overset{\text{(b)}}{=} \sum_{j \in \mathbb{N}} O\left(n^{-\lceil m/2 \rceil} \cdot (A + j - 1)^{-m}\right) \cdot (A + j)^{m-s} = O(n^{-\lceil m/2 \rceil}) \cdot \sum_{j \in \mathbb{N}} j^{-s} \overset{\text{(c)}}{=} O(n^{-m/2 - 1/2}),\end{aligned}$$

where (a) follows from Part (c) of Condition 3, (b) follows from Lemma 3, and (c) follows from the assumption $s > 1$ in Part (c) of Condition 3, and we use that m is odd. It follows that

$$\mathbb{E}_R\left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n} - \boldsymbol{\mu}^*\|_2 \geq A} \cdot D(P_{\boldsymbol{\mu}^*} \| P_{\breve{\boldsymbol{\mu}}_{|n}})\right] = O(n^{-m/2 - 1/2}). \tag{3.B.16}$$

We now show how (3.B.16) implies the result. A second order Taylor approximation gives:

$$\begin{aligned}&\mathbb{E}_R\left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n}\|_2 \geq n^{\alpha-1/2}} \cdot nD(P_{\boldsymbol{\mu}^*}\|P_{\breve{\boldsymbol{\mu}}_{|n}})\right]\\ \leq &\mathbb{E}_R\left[\mathbf{1}_{n^{\alpha-1/2}\leq\|\hat{\boldsymbol{\mu}}_{|n}\|_2\leq A}\left(\frac{1}{2}n\sup_{\boldsymbol{\mu}:\|\boldsymbol{\mu}\|_2\leq A}\left(\left(\frac{n\hat{\boldsymbol{\mu}}_{|n}+n_0x_0}{n+n_0}\right)^{\top}J(\boldsymbol{\mu})\left(\frac{n\mu+n_0x_0}{n+n_0}\right)\right)\right)\right]+\\ &\mathbb{E}_R\left[\mathbf{1}_{\|\boldsymbol{\mu}\|_2>A}\cdot nD(P_{\boldsymbol{\mu}^*}\|P_{\breve{\boldsymbol{\mu}}_{|n}})\right].\end{aligned}$$

The second term is, from the above, $O(n^{1/2-m/2})$. The first term can be bounded further, using that the supremum over $\boldsymbol{\mu}$ is on a bounded set in the interior of the parameter space, so that the eigenvalues of $J(\mu)$ are all bounded on this set, and using again Lemma 3, by

$$\begin{aligned}O\left((n\cdot A^2R(n^{\alpha-1/2}\leq\|\hat{\boldsymbol{\mu}}_{|n}\|_2\leq A)\right) &= O\left(n\cdot R(\|\hat{\boldsymbol{\mu}}_{|n}\|_2\geq n^{\alpha-1/2})\right)\\ &= O(n\cdot n^{-\lceil m/2\rceil}n^{-m(\alpha-1/2)}) = O(n^{-m\alpha+1/2}).\end{aligned}$$

The result follows.

Proof of (3.B.9) and (3.B.10) We first prove (3.B.9). Combining Proposition 9 with the first inequality in Proposition 8 (in which we bound $|\hat{\boldsymbol{\mu}}_i|$ by $n^{\alpha-1/2}$), we find that, absorbing d-factors into the $O(\cdot)$-notation:

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n}\|_2<n^{\alpha-1/2}}\cdot D(P_{\mathbf{0}}\|P_{\breve{\boldsymbol{\mu}}_{|n}})\right]=\mathbb{E}\left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n}\|_2<n^{\alpha-1/2}}\cdot D(P_{\mathbf{0}}\|P_{\hat{\boldsymbol{\mu}}_{|n}})+O\left(n^{\alpha-3/2}\right)\right]=\\ &\mathbb{E}\left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n}\|_2<n^{\alpha-1/2}}\cdot\left(\frac{1}{2}\boldsymbol{\mu}^{\top}J\boldsymbol{\mu}+O\left(n^{\alpha-1/2}\sum_{j,k\in[d]}|\hat{\boldsymbol{\mu}}_j||\hat{\boldsymbol{\mu}}_k|\right)\right)+O\left(n^{\alpha-3/2}\right)\right]=\\ &\mathbb{E}\left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n}\|_2<n^{\alpha-1/2}}\cdot\frac{1}{2}\boldsymbol{\mu}^{\top}J\boldsymbol{\mu}+\left(O\left(n^{\alpha-1/2}\max_{k\in[d]}|\hat{\boldsymbol{\mu}}_k|^2\right)\right)+O\left(n^{\alpha-3/2}\right)\right]=\\ &\mathbb{E}\left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n}\|_2<n^{\alpha-1/2}}\cdot\frac{1}{2}\boldsymbol{\mu}^{\top}J\boldsymbol{\mu}+O\left(n^{\alpha-3/2}\right)\right],\end{aligned}$$

where in the final equality we used Lemma 4. This proves (3.B.9). The proof of (3.B.10) is analogous (now only using Proposition 8, with the components of $D(\cdot\|\cdot)$ interchanged, and without the need to 'match' $\breve{\boldsymbol{\mu}}_{|n}$ with $\hat{\boldsymbol{\mu}}_{|n}$); we omit further details.

3.B.4 Proof for COND Part

Throughout the proof, we adopt the notations of (3.B.12).

We write $p^{\circ}_{\boldsymbol{\mu}^*}$ and q° as the densities of $Z = n^{1/2}\hat{\boldsymbol{\mu}}_{|n}$ under $P_{\boldsymbol{\mu}^*}$ and $Q = Q_{\boldsymbol{\mu}^*}$, respectively.

Lemma 5. *Under the regularity Condition 1 ('***COND***'), we have:*

$$\mathbb{E}_R \left[\log \frac{p^\circ_{\boldsymbol{\mu}^*}(\sqrt{n}\hat{\boldsymbol{\mu}}_{|n})}{q^\circ(\sqrt{n}\hat{\boldsymbol{\mu}}_{|n})} \right] \geq -D_{\Sigma_r}(\Sigma_q(\boldsymbol{\mu}^*) \| \Sigma_p(\boldsymbol{\mu}^*)) + o(1). \tag{3.B.17}$$

Proof. For the continuous case (X has Lebesgue density), the proof is based on the following immediate corollary of Theorem 19.2 of Bhattacharya and Rao [16] (page 192): consider an i.i.d. sequence of random vectors $X, X_{(1)}, X_{(2)}, \ldots$, with $X \sim \Psi$ with $\mathbb{E}_\Psi[X] = \mathbf{0}$ where Ψ has bounded continuous (Lebesgue) density and a moment generating function (in particular, all moments of each component of X exist). Then we have for ψ the density of $n^{-1/2}(\sum_{i=1}^n X_{(i)})$ that, for all integers $t \geq 1$, uniformly for all $y \in \mathbb{R}^d$,

$$1 + \sum_{j=1}^t n^{-j/2} f_j(y) - o\left(\frac{n^{-t/2}}{\phi_\Sigma(y)}\right) \leq \frac{\psi(y)}{\phi_\Sigma(y)} \leq 1 + \sum_{j=1}^t n^{-j/2} f_j(y) + o\left(\frac{n^{-t/2}}{\phi_\Sigma(y)}\right), \tag{3.B.18}$$

where ϕ_Σ is the density of a normal distribution with mean 0 and covariance matrix Σ and the $f_j : \mathbb{R}^d \to \mathbb{R}$ are specific $3j$-degree degree polynomials in the components of y (to see how this follows using Theorem 19.2. of [16, page 192] note that what we call t is, in their notation, $s-2$, note that their $P_0(-\phi(0,V) : \{\chi_v\}) = \phi_{0,V}$ is the density of a normal distribution with mean $\mathbf{0}$ and covariance matrix V, and their $P_1(-\phi(0,V) : \{\chi_v\})(y) = f(x)\phi_{0,V}$ [16, page 55–56]); their precondition on the characteristic function is equivalent to our condition of bounded Lebesgue density by their Theorem 19.1 and their precondition on moments being finite holds automatically because we assume Q has a moment generating function.

Below we only give the proof for the continuous case, based on (3.B.18); the proof for the discrete (lattice) case goes in exactly the same way, but now we use Theorem 22.1 of [16], which is their analogue of Theorem 19.2 for the discrete, lattice case; we omit details and continue with the continuous case.

We again adopt the notations (3.B.12), i.e. we set $\boldsymbol{\mu}^* = \mathbf{0}$ and we write $\Sigma_p := \Sigma_p(\boldsymbol{\mu}^*) = \Sigma_p(\mathbf{0})$. For fixed $A > 0$, we define

$$\begin{aligned} \mathcal{A}_n &= \left\{ x^{(n)} : \|\hat{\boldsymbol{\mu}}_{|n}\|_2 \leq \sqrt{\frac{A \log n}{n}} \right\} \\ \mathcal{B}_n &= \{x^{(n)} : \max\{\|X_{(i)}\|_2 : i \in [n]\} \leq n\} \\ \bar{\mathcal{B}}_{n,j} &= \{x^{(n)} : n + j - 1 < \max\{\|X_{(i)}\| : i \in [n]\} \leq n + j\}, \end{aligned}$$

and $\bar{\mathcal{A}}_n, \bar{\mathcal{B}}_n$ their respective complements. The expectation in (3.B.17) can be rewritten as

$$\mathbb{E}_R \left[\mathbf{1}_{\mathcal{A}_n} \log \frac{p^\circ_{\boldsymbol{\mu}^*}(\sqrt{n}\hat{\boldsymbol{\mu}}_{|n})}{q^\circ(\sqrt{n}\hat{\boldsymbol{\mu}}_{|n})} \right] + \mathbb{E}_R \left[\mathbf{1}_{\bar{\mathcal{A}}_n \cap \mathcal{B}_n} \log \frac{p^\circ_{\boldsymbol{\mu}^*}(\sqrt{n}\hat{\boldsymbol{\mu}}_{|n})}{q^\circ(\sqrt{n}\hat{\boldsymbol{\mu}}_{|n})} \right] + \mathbb{E}_R \left[\mathbf{1}_{\bar{\mathcal{A}}_n \cap \bar{\mathcal{B}}_n} \log \frac{p^\circ_{\boldsymbol{\mu}^*}(\sqrt{n}\hat{\boldsymbol{\mu}}_{|n})}{q^\circ(\sqrt{n}\hat{\boldsymbol{\mu}}_{|n})} \right]. \tag{3.B.19}$$

We now use the local central limit theorem with expansion as in (3.B.18) above to analyze (minus) the first term. Since Σ_p is positive definite, we find

$$1/\phi_{\Sigma_p}(y) = O\left(\exp\left(\frac{1}{2}\|y\|_2^2\lambda_{\max,p}\right)\right),$$

where $\lambda_{\max,p}$ is the maximum eigenvalue of Σ_p^{-1}. Setting $y = n^{-1/2}\sum_{i=1}^n x_{(i)}$ and plugging in $\|y\| \leq \sqrt{A\log n}$ which will hold on the set $\mathcal{A}_n$ (note scaling takes care of $1/\sqrt{n}$) we find that $\frac{n^{-t/2}}{\phi_{\Sigma_p}(y)} = o(1)$ if $t \geq A\lambda_{\max,p}$. The same derivation holds with Σ_q instead of Σ_p, so we get

$$\frac{n^{-t/2}}{\phi_{\Sigma_p}(y)} = o(1), \frac{n^{-t/2}}{\phi_{\Sigma_q}(y)} = o(1),$$

if $t \geq A \cdot \max\{\lambda_{\max,p}, \lambda_{\max,q}\}$. We then find that the remainder terms $n^{-t/2}f_j(y)$ are all $O((\log n)^{3j}/n)^{j/2}) = o(1)$ on all y corresponding to $x^{(n)}$ in $\mathcal{A}_n$. We can thus set ψ to p° and use the left inequality in (3.B.18) with this t and then set ψ to q° and using the right inequality with the same t to get

$$\begin{aligned}\mathbb{E}_R\left[\mathbf{1}_{\mathcal{A}_n}\log\frac{q^\circ(\sqrt{n}\hat{\boldsymbol{\mu}}_{|n})}{p^\circ_{\boldsymbol{\mu}^*}(\sqrt{n}\hat{\boldsymbol{\mu}}_{|n})}\right] \leq& \mathbb{E}_R\left[\mathbf{1}_{\mathcal{A}_n}\log\frac{\phi_{\Sigma_q}(\sqrt{n}\hat{\boldsymbol{\mu}}_{|n})(1+o(1))}{\phi_{\Sigma_p}(\sqrt{n}\hat{\boldsymbol{\mu}}_{|n})(1-o(1))}\right]\\ =& D_{\Sigma_r}(\Sigma_q(\boldsymbol{\mu}^*)\|\Sigma_p(\boldsymbol{\mu}^*)) + o(1), \end{aligned} \tag{3.B.20}$$

where the final equality follows because $Q(\mathcal{A}_n) \to 1$, which follows because we assume that Q has a moment generating function. This deals with the first term in (3.B.19).

Now consider the second term in (3.B.19). Since we assume R has a moment generating function, we can with some work (details omitted) employ the Cramèr-Chernoff method to get that for all $B > 0$, if we take A sufficiently large, we get:

$$R\left(\|\hat{\boldsymbol{\mu}}_{|n}\|_2 \geq \sqrt{\frac{A\log n}{n}}\right) = O(n^{-B}). \tag{3.B.21}$$

Now fix $B = a + 2$, with a the exponent in Condition 1, and choose A large enough so that (3.B.21) holds and then t large enough so that (3.B.20) also holds. For the second term we then get,

$$\mathbb{E}_R\left[\mathbf{1}_{\bar{\mathcal{A}}_n\cap\mathcal{B}_n}\log\frac{p^\circ_{\boldsymbol{\mu}^*}(\sqrt{n}\hat{\boldsymbol{\mu}}_{|n})}{q^\circ(\sqrt{n}\hat{\boldsymbol{\mu}}_{|n})}\right] = O(n^{-a-2}) \cdot \sup_{x^{(n)}\in\bar{\mathcal{A}}_n\cap\mathcal{B}_n}\log\frac{p_{\boldsymbol{\mu}^*}(x^{(n)})}{q(x^{(n)})} = \tag{3.B.22}$$

$$\begin{aligned}=&O(n^{-a-2}) \cdot \sup_{x^{(n)}\in\bar{\mathcal{A}}_n\cap\mathcal{B}_n}\sum_{i=1}^n\log\frac{p_{\boldsymbol{\mu}^*}(x_{(i)})}{q(x_{(i)})} \leq O(n^{-a-2}) \cdot n \cdot \sup_{x^{(n)}\in\mathcal{B}_n}\max_{i\in[n]}\log\frac{p_{\boldsymbol{\mu}^*}(x_{(i)})}{q(x_{(i)})}\\ =&O(n^{-1}),\end{aligned}$$

where we in the penultimate equality we used $x^{(n)} \in \mathcal{B}_n$ and Condition 1. The first equality follows because with $\mathcal{G}_\epsilon(v^{(n)}) = \{x^{(n)} : |\sqrt{n}(\hat{\boldsymbol{\mu}}_{|v^{(n)}} - \hat{\boldsymbol{\mu}}_{|x^{(n)}})| \leq \epsilon\}$, we have:

$$\sup_{v^{(n)} \in \bar{\mathcal{A}}_n \cap \mathcal{B}_n} \frac{p^\circ_{\boldsymbol{\mu}^*}(\sqrt{n}\hat{\boldsymbol{\mu}}_{|v^{(n)}})}{q^\circ(\sqrt{n}\hat{\boldsymbol{\mu}}_{|v^{(n)}})} = \sup_{v^{(n)} \in \bar{\mathcal{A}}_n \cap \mathcal{B}_n} \lim_{\epsilon \downarrow 0} \frac{\epsilon^{-1} \int_{\mathcal{G}_\epsilon(v^{(n)})} p_{\boldsymbol{\mu}^*}(x^{(n)}) dx^{(n)}}{\epsilon^{-1} \int_{\mathcal{G}_\epsilon(v^{(n)})} q(x^{(n)}) dx^{(n)}} \tag{3.B.23}$$

$$= \sup_{v^{(n)} \in \bar{\mathcal{A}}_n \cap \mathcal{B}_n} \lim_{\epsilon \downarrow 0} \frac{\int_{\mathcal{G}_\epsilon(v^{(n)})} p_{\boldsymbol{\mu}^*}(x^{(n)}) dx^{(n)}}{\int_{\mathcal{G}_\epsilon(v^{(n)})} q(x^{(n)}) dx^{(n)}} = \sup_{v^{(n)} \in \bar{\mathcal{A}}_n \cap \mathcal{B}_n} \frac{p_{\boldsymbol{\mu}^*}(v^{(n)})}{q(v^{(n)})},$$

where the final inequality follows by the assumed continuity of $p_{\boldsymbol{\mu}^*}$ and q.

(3.B.22) gives the second term. For the third term we get, with again a the exponent in Condition 1 and using, in the first inequality below, analogously to (3.B.23):

$$\mathbb{E}_R\left[\mathbf{1}_{\bar{\mathcal{A}}_n \cap \bar{\mathcal{B}}_n} \log \frac{p^\circ_{\boldsymbol{\mu}^*}(\sqrt{n}\hat{\boldsymbol{\mu}}_{|n})}{q^\circ(\sqrt{n}\hat{\boldsymbol{\mu}}_{|n})}\right] \leq \sum_{j \in \mathbb{N}} R(\bar{\mathcal{A}}_n, \bar{\mathcal{B}}_{n,j}) \cdot n \cdot (n+j)^a$$

$$\leq \sum_{j \in \mathbb{N}} R(\exists i \in [n] : |X_i| \geq n+j-1) \cdot n \cdot (n+j)^a \leq n^2 \sum R(|X| \geq n+j-1)(n+j)^a$$

$$\leq 2n^2 \min_k \mathbb{E}_R[|X|^k] \sum_{j \in \mathbb{N}} (n+j-1)^{-k}(n+j)^a$$

$$\leq 2 \min_k \mathbb{E}_R[|X|^k] \sum_{j \in \mathbb{N}} (n+j-1)^{-k}(n+j)^{2+a} = O\left(n^{-1}\right)$$

where the penultimate equality follows by the fact that we can employ Markov's inequality with X^k, for all k, and the final equality is obtained by setting $k = a+4$. This proves (3.B.17) and hence (3.3.3). The statement right below (3.3.3) follows by the fact that the role of $p_{\boldsymbol{\mu}^*}$ and q becomes completely symmetric if Condition 1 ('**COND**') holds with q and $p_{\boldsymbol{\mu}^*}$ interchanged. □

3.B.5 Proof of Lemma 1, underlying Anti-Simple case part of Theorem 4

Proof. We use the notations as summarized in (3.B.12). Fix $0 < \alpha < 1/6$. Without loss of generality, let $\boldsymbol{\mu}^* = 0$. The integral below is over $\mathtt{M}_p$. Since n remains fixed throughout the proof, we abbreviate $W_{0,(n)}$ to W and $w_{0,(n)}$ to w. Let $S^{(n)}_{Q,\text{RIP}'} = \frac{q(U^{(n)})}{p_W(U^{(n)})}$. It is easy to get that

$$\mathbb{E}_{U^{(n)} \sim Q}[\log S^{(n)}_{Q,\text{RIP}'}] = nD(Q\|P_{\boldsymbol{\mu}^*}) + \mathbb{E}_{U^{(n)} \sim Q}\left[\log \frac{p_{\boldsymbol{\mu}^*}(U^{(n)})}{\int w(\boldsymbol{\mu}) p_{\boldsymbol{\mu}}(U^{(n)}) d\boldsymbol{\mu}}\right],$$

so we just need to focus on the second term. By (3.B.6) in Lemma 2, we can write it as:

$$\mathbb{E}_Q\left[\log\frac{p_{\boldsymbol{\mu}^*}(U^{(n)})}{\int w(\boldsymbol{\mu})p_{\boldsymbol{\mu}}(U^{(n)})d\boldsymbol{\mu}}\right]=\mathbb{E}_Q\left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n}\|_2<n^{\alpha-1/2}}\cdot\log\frac{p_{\boldsymbol{\mu}^*}(U^{(n)})}{\int w_n(\boldsymbol{\mu})p_{\boldsymbol{\mu}}(U^{(n)})d\boldsymbol{\mu}}\right]+o(1)= \tag{3.B.24}$$

$$\mathbb{E}_Q\left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n}\|_2<n^{\alpha-1/2}}\cdot\log\frac{p_{\boldsymbol{\mu}^*}(U^{(n)})}{p_{\hat{\boldsymbol{\mu}}_{|n}}(U^{(n)})}\right]+\mathbb{E}_Q\left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n}\|_2<n^{\alpha-1/2}}\cdot\log\frac{p_{\hat{\boldsymbol{\mu}}_{|n}}(U^{(n)})}{\int w(\boldsymbol{\mu})p_{\boldsymbol{\mu}}(U^{(n)})d\boldsymbol{\mu}}\right]+o(1)=$$
$$-\frac{1}{2}\text{TR}\left(\Sigma_q\Sigma_p^{-1}\right)+\mathbb{E}_Q\left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n}\|_2<n^{\alpha-1/2}}\cdot\log\frac{p_{\hat{\boldsymbol{\mu}}_{|n}}(U^{(n)})}{\int w(\boldsymbol{\mu})p_{\boldsymbol{\mu}}(U^{(n)})d\boldsymbol{\mu}}\right]+o(1),$$

where the final equality is obtained by Lemma 2, (3.B.9). So we just need to upper bound the expectation in (3.B.24). For this, we fix any β with $\alpha<\beta<1/6$ and note that, uniformly for all $U^{(n)}\in\mathcal{U}^n$ with $\|\hat{\boldsymbol{\mu}}\|_2\leq n^{\beta-1/2}$, we have:

$$\begin{aligned}
&-\log\int w(\boldsymbol{\mu})\frac{p_{\boldsymbol{\mu}}(U^{(n)})}{p_{\hat{\boldsymbol{\mu}}_{|n}}(U^{(n)})}d\boldsymbol{\mu}=-\log\int w(\boldsymbol{\mu})\exp\left(-nD(P_{\hat{\boldsymbol{\mu}}_{|n}}\|P_{\boldsymbol{\mu}})\right)d\boldsymbol{\mu}\leq\\
&-\log\int_{\boldsymbol{\mu}:\|\boldsymbol{\mu}\|_2\leq n^{\beta-1/2}}w(\boldsymbol{\mu})\exp\left(-nD(P_{\hat{\boldsymbol{\mu}}_{|n}}\|P_{\boldsymbol{\mu}})\right)d\boldsymbol{\mu}\leq\\
&-\log\int_{\boldsymbol{\mu}:\|\boldsymbol{\mu}\|_2\leq n^{\beta-1/2}}w(\boldsymbol{\mu})\exp\left(-\frac{n}{2}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_{|n})^\top J(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_{|n})+O\left(n\cdot n^{3\beta-3/2}\right)\right)d\boldsymbol{\mu}\leq\\
&-\log\int_{\boldsymbol{\mu}:\|\boldsymbol{\mu}\|_2\leq n^{\beta-1/2}}w(\boldsymbol{\mu})\exp\left(-\frac{n}{2}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_{|n})^\top J(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_{|n})\right)d\boldsymbol{\mu}+o(1)\leq\\
&\frac{d}{2}\log\frac{2\pi}{n}-\frac{1}{2}\log\det K\\
&-\log\int_{\boldsymbol{\mu}:\|\boldsymbol{\mu}\|_2\leq n^{\beta-1/2}}\exp\left(-\frac{n}{2}\left(\boldsymbol{\mu}^\top K\boldsymbol{\mu}+(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_{|n})^\top J(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_{|n})\right)\right)d\boldsymbol{\mu}+o(1),
\end{aligned} \tag{3.B.25}$$

where we first used the robustness property of exponential families (Lemma 2, (3.B.3)) and then for the second inequality, we used the second inequality in Proposition 8 (with the α in that proposition set to $\alpha\vee\beta=\beta$). The third inequality uses that β was set $<1/6$, and in the final line we used the definition of w as the density of a Gaussian with mean $\mathbf{0}$ and covariance K^{-1}/n. By a little computation, we can rewrite the expression in the exponent in the integral in (3.B.25) as

$$\boldsymbol{\mu}^\top K\boldsymbol{\mu}+(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_{|n})^\top J(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_{|n})=(\boldsymbol{\mu}-\boldsymbol{m})^\top(K+J)(\boldsymbol{\mu}-\boldsymbol{m})+C,$$

where $\boldsymbol{m} = (K+J)^{-1}J\hat{\boldsymbol{\mu}}_{|n}$ and $C = \hat{\boldsymbol{\mu}}_{|n}^\top J\hat{\boldsymbol{\mu}}_{|n} - \boldsymbol{m}^\top(K+J)\boldsymbol{m}$. which can be simplified to $C = \hat{\boldsymbol{\mu}}_{|n}^\top \Sigma_q^{-1}\hat{\boldsymbol{\mu}}_{|n}$ by first noting $C = \hat{\boldsymbol{\mu}}_{|n}^\top(J - J(K+J)^{-1}J)\hat{\boldsymbol{\mu}}_{|n}$ and then

$$J - J(K+J)^{-1}J = J\left(I - (K+J)^{-1}J\right) = J\left(I - (I + J^{-1}K)^{-1}\right) =$$
$$J\left(I - (I - J^{-1}(I+KJ^{-1})^{-1}K)\right) = J\left(J^{-1}(I+KJ^{-1})^{-1}K\right) = (K^{-1}+J^{-1})^{-1} = \Sigma_q^{-1},$$

where the third equality follows by the reduced Woodbury matrix identity (see e.g. wikipedia). It follows that (3.B.25) can be further rewritten as

$$\frac{d}{2}\log\frac{2\pi}{n} - \frac{1}{2}\log\det K + \frac{n}{2}\cdot\hat{\boldsymbol{\mu}}_{|n}^\top\Sigma_q^{-1}\hat{\boldsymbol{\mu}}_{|n} \tag{3.B.26}$$
$$- \log\int_{\boldsymbol{\mu}:\|\boldsymbol{\mu}\|_2\le n^{\beta-1/2}} \exp\left(-\frac{n}{2}\left((\boldsymbol{\mu}-\boldsymbol{m})^\top(K+J)(\boldsymbol{\mu}-\boldsymbol{m})\right)\right)d\boldsymbol{\mu} + o(1),$$

where $\boldsymbol{m} = (K+J)^{-1}J\hat{\boldsymbol{\mu}}_{|n}$ does not depend on $\boldsymbol{\mu}$ and has norm $\|\boldsymbol{m}\|_2 = O(\|\hat{\boldsymbol{\mu}}_{|n}\|_2) = O(n^{\alpha-1/2})$ (recall that we are only evaluating the integral for values of $\hat{\boldsymbol{\mu}}_{|n}$ with $\|\hat{\boldsymbol{\mu}}_{|n}\|_2 < n^{\alpha-1/2}$). Now note (a) the distance between $\boldsymbol{m}$ and the boundary of the set over which we integrate, $\{\boldsymbol{\mu} : \|\boldsymbol{\mu}\|_2 \le n^{\beta-1/2}\}$ is therefore of order $n^{\beta-1/2} - n^{\alpha-1/2} = n^{-1/2}g(n)$ for a function $g(n)$ with $\lim_{n\to\infty} g(n) = \infty$ (recall we chose β larger than α) ; and (b) since K and J are inverses of positive definite matrices, they are themselves positive definite and so is $K+J$. Therefore the integral in (3.B.26) converges, with increasing n, to a Gaussian integral with covariance $(K+J)/n$, so we can rewrite (3.B.26) as:

$$\frac{d}{2}\log\frac{2\pi}{n} - \frac{1}{2}\log\det K + \frac{n}{2}\cdot\hat{\boldsymbol{\mu}}_{|n}^\top\Sigma_q^{-1}\hat{\boldsymbol{\mu}}_{|n} - \frac{d}{2}\log\frac{2\pi}{n} + \frac{1}{2}\log\det(K+J) + o(1)$$
$$=\frac{1}{2}\log\frac{\det(K+J)}{\det K} + \frac{n}{2}\cdot\hat{\boldsymbol{\mu}}_{|n}^\top\Sigma_q^{-1}\hat{\boldsymbol{\mu}}_{|n} + o(1),$$

and then plugging this back into (3.B.24) we see that (3.B.24) can be bounded as

$$-\frac{1}{2}\text{TR}\left(\Sigma_q\Sigma_p^{-1}\right) + \frac{1}{2}\log\frac{\det(K+J)}{\det K} + \frac{n}{2}\cdot\mathbb{E}_Q\left[\mathbf{1}_{\|\hat{\boldsymbol{\mu}}_{|n}\|_2<n^{\alpha-1/2}}\cdot\hat{\boldsymbol{\mu}}_{|n}^\top\Sigma_q^{-1}\hat{\boldsymbol{\mu}}_{|n}\right] + o(1),$$

so that , by (3.B.4) and (3.B.5) (recall we set $\boldsymbol{\mu}^* = \mathbf{0}$), (3.B.24) can be further bounded as

$$\frac{1}{2}\log\frac{\det(K+J)}{\det K} + \frac{d}{2} - \frac{\text{TR}\left(\Sigma_q\Sigma_p^{-1}\right)}{2} + o(1) \tag{3.B.27}$$
$$=\frac{1}{2}\log\det(\Sigma_p^{-1}\Sigma_q) + \frac{d}{2} - \frac{\text{TR}\left(\Sigma_q\Sigma_p^{-1}\right)}{2} + o(1) = -D_{\text{GAUSS}}(\Sigma_q\Sigma_p^{-1}) + o(1).$$

The penultimate equality holds by direct computation, using the definitions of K and J. But (3.B.27) implies the result. □

3.B.6 Proof of (3.3.15) in Theorem 5

Let $\mathcal{A}_\gamma := \{\boldsymbol{\mu}' \in \mathtt{M}_p \cap \mathtt{M}_q : \|\boldsymbol{\mu}' - \boldsymbol{\mu}^*\|_2 \leq \gamma\}$. Define $S^{(1)}_{Q_{\boldsymbol{\mu}'}} := \frac{q_{\boldsymbol{\mu}'}(U)}{p_{\leftrightsquigarrow q_{\boldsymbol{\mu}'}(U)}(U)}$ and note that $S^{(1)}_{Q_{\boldsymbol{\mu}'}}$ is an e-variable for all $\boldsymbol{\mu}' \in \mathcal{A}_\gamma$, and in particular $S^{(1)}_{Q_{\boldsymbol{\mu}^*}}$ is the GRO (optimal) e-variable relative to $Q_{\boldsymbol{\mu}^*}$. Therefore, we must have, for some $\epsilon > 0$, that

$$\mathbb{E}_{Q_{\boldsymbol{\mu}^*}}\left[\log \frac{q_{\boldsymbol{\mu}'}(U)}{p_{\leftrightsquigarrow q_{\boldsymbol{\mu}'}(U)}(U)}\right] = \mathbb{E}_{Q_{\boldsymbol{\mu}^*}}\left[\log S^{(1)}_{Q_{\boldsymbol{\mu}'}}\right] \leq \mathbb{E}_{Q_{\boldsymbol{\mu}^*}}\left[\log S^{(1)}_{Q_{\boldsymbol{\mu}^*}}\right] = \mathbb{E}_{Q_{\boldsymbol{\mu}^*}}\left[\log \frac{q_{\boldsymbol{\mu}^*}(U)}{p_{\boldsymbol{\mu}^*}(U)}\right] - \epsilon,$$

where the final equality follows because, by the same reasoning as in the proof of Theorem 4, Part 3, we have (3.5.5), with $Q_{\boldsymbol{\mu}^*}$ in the role of Q. But, taking expectations over $U_{(1)}, U_{(2)}, \ldots$ i.i.d. $\sim Q_{\boldsymbol{\mu}^*}$, from the above we immediately get, using Fubini's theorem,

$$\mathbb{E}_{U^{(n)} \sim Q_{\boldsymbol{\mu}^*}}\left[\log \prod_{i=1}^{n} \frac{q_{\breve{\boldsymbol{\mu}}_{|i-1}}(\bar{U}_{(i)})}{p_{\leftrightsquigarrow q_{\breve{\boldsymbol{\mu}}_{|i-1}}}(\bar{U}_{(i)})}\right] = \sum_{i=1}^{n} \mathbb{E}_{U^{(i-1)} \sim Q_{\boldsymbol{\mu}^*}} \mathbb{E}_{U \sim Q_{\boldsymbol{\mu}^*}}\left[\log \frac{q_{\breve{\boldsymbol{\mu}}_{|i-1}}(U)}{p_{\leftrightsquigarrow q_{\breve{\boldsymbol{\mu}}_{|i-1}}(U)}(U)}\right]$$
$$\leq n(D(Q_{\boldsymbol{\mu}^*} \| P_{\boldsymbol{\mu}^*}) - \epsilon),$$

and the result is proved.

Appendix 3.C Proofs for Section 3.6

Proof of Proposition 7 If $S^{(n)}_{n \geq 0}$ were an e-process relative to null hypothesis $\mathcal{P}$, then, as is immediate from the definition of e-process, it must also be an e-process relative to the simple null hypothesis $\{P_{W_i}\}$ for any $i \geq 0$. Thus, if $\mathbb{E}_{U^{(n)} \sim P_{W_n}}[q(U^{(n)})/p_{W_{n+1}}(U^{(n)})] > 1$, then $(S^{(n)})_{n \geq 0}$ cannot form an e-process and the result is proved. Therefore, we may assume that $\mathbb{E}_{U^{(n)} \sim P_{W_n}}[q(U^{(n)})/p_{W_{n+1}}(U^{(n)})] \leq 1$, or equivalently, $\mathbb{E}_{U^{(n)} \sim Q}\left[\frac{p_{W_n}(U^{(n)})}{p_{W_{n+1}}(U^{(n)})}\right] \leq 1$. Using Jensen's inequality, which by our assumption (3.6.1) is strict, now gives

$$\mathbb{E}_{U^{(n)} \sim Q}\left[\frac{p_{W_{n+1}}(U^{(n)})}{p_{W_n}(U^{(n)})}\right] > 1 / \left(\mathbb{E}_{U^{(n)} \sim Q}\left[\frac{p_{W_n}(U^{(n)})}{p_{W_{n+1}}(U^{(n)})}\right]\right) \geq 1,$$

or equivalently, $\mathbb{E}_{U^{(n)} \sim P_{W_{n+1}}}[q(U^{(n)})/p_{W_n}(U^{(n)})] > 1$, and the result again follows.

Appendix 3.D Additional Details for Section 3.3: checking UI and plug-in Regularity Conditions for Example Families

3.D.1 Condition UI for the Poisson Model

Example 7. **[Poisson]** Let $\mathcal{P}$ be the Poisson family, given in its standard parameterization as $p_\mu(x) = P_\mu(X = k) = \frac{\mu^k \cdot e^{-\mu}}{k!}$ with parameter $\mu > 0$, and suppose X has m moments under R. Standard and straightforward computations show that this is an exponential family with mean-value parameter $\mu = \mathbb{E}_{P_\mu}[X]$ and $D(P_{\hat{\mu}_{|n}} \| P_{\mu^*}) = \hat{\mu}_{|n} \log \frac{\hat{\mu}_{|n}}{\mu^*} + \mu^* - \hat{\mu}_{|n}$. Fix $0 < \gamma < 1/2$ and define $\mathcal{E}_j = \{\mu \in (0, \infty) : |\mu - \mu^*| \in [n^{-\gamma} + j - 1, n^{-\gamma} + j]\}$. We have

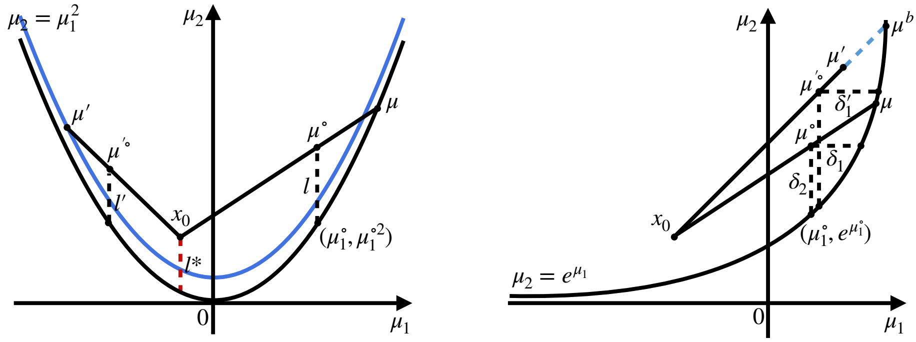
$$
\begin{aligned}
&\mathbb{E}_R \left[\mathbf{1}_{\|\hat{\mu}-\mu^*\|_2 \geq n^{-\gamma}} \cdot D(P_{\hat{\mu}} \| P_{\mu^*})\right] \leq \sum_{j \in \mathbb{N}^+} R(\hat{\mu}_{|n} \in \mathcal{E}_j) \max_{\hat{\mu}_{|n} \in \mathcal{E}_j} D(P_{\hat{\mu}_{|n}} \| P_{\mu^*}) \\
&= \sum_{j \in \mathbb{N}^+} R(\hat{\mu}_{|n} \in \mathcal{E}_j) \max_{\hat{\mu}_{|n} \in \mathcal{E}_j} \left(\hat{\mu}_{|n} \log \frac{\hat{\mu}_{|n}}{\mu^*} + \mu^* - \hat{\mu}_{|n}\right) \\
&\leq R\left(|\hat{\mu}_{|n} - \mu^*| \geq n^{-\gamma}\right) \cdot ((-\log \mu^*) \cdot O(1) + \mu^* - O(1)) \\
&\quad + \sum_{j \in \mathbb{N}^+} R\left(|\hat{\mu}_{|n} - \mu^*| \geq n^{-\gamma} + j\right) \cdot O\left((n^{-\gamma} + j + 1) \log(n^{-\gamma} + j + 1)\right) \\
&\overset{\text{(a)}}{=} O\left(n^{-\lceil m/2 \rceil} \cdot n^{\gamma m - \gamma}\right) + \sum_{j \in \mathbb{N}^+} O\left(n^{-\lceil m/2 \rceil} \cdot \left(n^{-\gamma} + j\right)^{-m}\right) \cdot O\left(j \log j\right) \\
&= O\left(n^{-\lceil m/2 \rceil} \cdot n^{\gamma m - \gamma}\right) + O(n^{-\lceil m/2 \rceil}) \cdot \sum_{j \in \mathbb{N}^+} j^{1-m} \log j := f(n, m, \gamma),
\end{aligned}
$$

where (a) follows from Lemma 3 in Appendix 3.B.2. Plugging in any $m \geq 3$ and $\gamma = 1/3$ we find that $f(n, m, \gamma) = O(n^{-4/3}) = o(n^{-1})$, which proves the result.

3.D.2 Proof for Example 5: plug-in condition for the full Gaussian

Proof of (3.3.10) We first consider the case of a $\boldsymbol{\mu}$ located on the boundary $\partial \mathtt{M}_q$ of $\mathtt{M}_q$ (actually this cannot happen because $\mathtt{M}_q$ is open, but it is still useful for proving the result as it constitutes a limiting worst-case). In this case, for fixed $0 < \alpha < 1$ and x_0, we have:

$$
\mu_2^\circ - \mu_1^{\circ 2} = (1 - \alpha)\mu_2 + \alpha x_{0,2} - \mu_1^{\circ 2} = (1 - \alpha)\mu_1^2 + \alpha x_{0,2} - ((1 - \alpha)\mu_1 + \alpha x_{0,1})^2. \tag{3.D.1}
$$



(a) Gaussian mean parameter space

(b) Gamma mean parameter space

Figure 3.1: In the left figure, the area above the black curve $\mu_2 = \mu_1^2$ is the Gaussian mean parameter space $\mathtt{M}_q$. The blue curve is obtained by vertically shifting the black one upwards such that $\boldsymbol{\mu}'$ is located on it. (3.3.10) expresses that αl^* is not larger than any other l for every $\boldsymbol{\mu}$, l' for every $\boldsymbol{\mu}'$ with $x_0 \in \mathtt{M}_q$ and $\alpha \in (0,1)$.

Taking the first derivative of $\mu_2^\circ - \mu_1^{\circ 2}$: $\frac{d}{d\mu_1}(\mu_2^\circ - \mu_1^{\circ 2}) = 2(1-\alpha)\alpha(\mu_1 - x_{0,1})$, we find

$$\arg\min_{\boldsymbol{\mu}\in\partial\mathtt{M}_q} (\mu_2^\circ - \mu_1^{\circ 2}) = (x_{0,1}, x_{0,1}^2). \tag{3.D.2}$$

(note again $\boldsymbol{\mu}^\circ$ is a function of $\boldsymbol{\mu}$). Then (3.3.10) holds with the infimum over $\partial\mathtt{M}_q$ instead of $\mathtt{M}_q$, as seen by plugging (3.D.2) into (3.D.1). The situation is depicted in Figure 3.1(a). l denotes $\mu_2^\circ - (\mu_1^\circ)^2$, l^* denotes $x_{0,2} - x_{0,1}^2$, and the statement (3.3.10) says that $\alpha l^* \leq l$.

Now consider the case of a $\boldsymbol{\mu}' \in \mathtt{M}_q$ not located on the boundary, and the corresponding $\boldsymbol{\mu}'^\circ$. In terms of the figure, we now need to prove $\alpha l^* \leq l'$, where l' denotes $\mu_2'^\circ - (\mu_1'^\circ)^2$. Let δ be the vertical translation distance between the black curve and the blue curve in the figure, i.e. the blue curve is $y = x^2 + \delta$. Extending the boundary-$\boldsymbol{\mu}$ analysis to this case, we find $l' - \delta \geq \alpha(l^* - \delta)$. This implies $l' > \alpha l^*$ (the boundary case above is really the worst-case). Note that the same reasoning still applies when x_0 is below the blue curve — then l^* is a negative real number and also l might be negative in some cases. (3.3.10) follows.

Proof that (3.3.8) holds For $\boldsymbol{\mu} \in \mathtt{S}_q \subset \{\boldsymbol{\mu} \in \mathtt{M}_q : \|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|_2 \geq A\}$ where $\mathtt{S}_q$ is compact, we have $\max_{\boldsymbol{\mu}\in\mathtt{S}_q} \frac{D(Q_{\boldsymbol{\mu}^*}\|Q_{\boldsymbol{\mu}^\circ})}{\|\boldsymbol{\mu}-\boldsymbol{\mu}^*\|_2^2} < \infty$ because $\frac{D(Q_{\boldsymbol{\mu}^*}\|Q_{\boldsymbol{\mu}^\circ})}{\|\boldsymbol{\mu}-\boldsymbol{\mu}^*\|_2^2}$ is continuous w.r.t. $\boldsymbol{\mu}$. Thus, we just need to check points that tend to the boundary of $\mathtt{M}_q$. For this, let $(\boldsymbol{\mu}_{[m]})_{m\in\mathbb{N}}$ be a sequence in $\{\boldsymbol{\mu} \in \mathtt{M}_q : \|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|_2 \geq A\}$ such that the limit in (3.D.3) below exists. We know by (3.3.10) that

$$\lim_{m\to\infty} \frac{D(Q_{\boldsymbol{\mu}^*}\|Q_{\boldsymbol{\mu}^\circ_{[m]}})}{\|\boldsymbol{\mu}_{[m]}-\boldsymbol{\mu}^*\|_2^2} = O\left(\frac{\log((\mu_2^\circ)_{[m]}-(\mu_1^{\circ 2})_{[m]})}{\|\boldsymbol{\mu}_{[m]}-\boldsymbol{\mu}^*\|_2^2}\right) + O\left(\frac{-2\mu_1^*(\mu_1^\circ)_{[m]}+(\mu_1^{\circ 2})_{[m]}}{\|\boldsymbol{\mu}_{[m]}-\boldsymbol{\mu}^*\|_2^2}\right). \tag{3.D.3}$$

We need to show that for all such sequences with $\boldsymbol{\mu}_{[m]}$ tending to the boundary of $\mathtt{M}_q$, the above limit is finite. We first note that by (3.3.10), we cannot have that $\mu_2^\circ - \mu_1^{\circ 2} \to 0$, because $\mu_2^\circ - \mu_1^{\circ 2} \geq \alpha(x_{0,2} - x_{0,1}^2)$. Thus, we only need to consider the sequences with either $\mu_2^\circ - \mu_1^{\circ 2} \to \infty$ or $|\mu_1^\circ| \to \infty$. Using that $\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|_2^2 = (\mu_1 - \mu_1^*)^2 + (\mu_1^2 + \sigma_1^2 - \mu_1^{*2} - \sigma_1^{*2})^2$, it is easily checked that (3.D.3) is finite on all such sequences.

3.D.3 Plug-in Regularity Conditions for Gamma Model

Example 8. [Gamma] The argument follows analogous stages as the one for the Gaussian case, Example 5. Let $\mathcal{Q}$ be the $\textsc{Gamma}(\alpha, \beta)$ family with densities $q(x;\alpha,\beta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha}$, for $x > 0$ and $\alpha, \beta > 0$. In terms of the mean-value parameterization, we can parameterize this family by $\boldsymbol{\mu} = (\mu_1, \mu_2)$ with $\mu_1 = \psi(\alpha) + \log\beta$ and $\mu_2 = \alpha\beta$, where $\psi(\alpha)$ is the digamma function. We first claim

$$\mathtt{M}_q = \{(\mu_1, \mu_2) : \mu_1 \in \mathbb{R}, \mu_2 > e^{\mu_1}\}. \tag{3.D.4}$$

To see this, rewrite $\boldsymbol{\mu} = (\psi(\alpha) - \log\alpha + \log(\alpha\beta), \alpha\beta)$. For any fixed $c = \alpha\beta$, we have that $\psi(\alpha) - \log\alpha$ is increasing and $\psi(\alpha) - \log\alpha \in (-\infty, 0), \alpha \in (0, \infty)$. Then the mean parameter space $\mathtt{M}_q$ is located on the upper left part of $f(\alpha\beta) = e^{\alpha\beta}$, i.e. $\mu_2 = e^{\mu_1}$, which proves (3.D.4); see Figure 3.1(b). Next, it is well-known that

$$D(Q_{\boldsymbol{\mu}^*}\|Q_{\boldsymbol{\mu}^\circ}) = \alpha^\circ \log\frac{\beta^\circ}{\beta^*} - \log\frac{\Gamma(\alpha^*)}{\Gamma(\alpha^\circ)} + (\alpha^* - \alpha^\circ)\psi(\alpha^*) - \left(\frac{1}{\beta^*} - \frac{1}{\beta^\circ}\right)\alpha^*\beta^*. \tag{3.D.5}$$

with $(\alpha^\circ, \beta^\circ)$ the parameters in the standard parameterization corresponding to $(\mu_1^\circ, \mu_2^\circ)$.

Now let $\boldsymbol{\mu}^\circ = (1-k)\boldsymbol{\mu} + kx_0$. Further below we show that for every $0 < k < 1$, every $x_0 = (x_{0,1}, x_{0,2}) \in \mathtt{M}_q$, we have

$$\mu_1^\circ - \log\mu_2^\circ = \psi(\alpha^\circ) - \log\alpha^\circ \text{ and } 0 < \alpha^\circ \leq \alpha_{x_0}, \tag{3.D.6}$$

where α_{x_0} is a constant depending on x_0. And we also show

$$\mu_2^\circ - e^{\mu_1^\circ} = (\alpha^\circ - e^{\psi(\alpha^\circ)})\beta^\circ \text{ and } \beta^\circ \geq \frac{k(x_{0,2} - e^{x_{0,1}})}{\alpha_{x_0} - e^{\psi(\alpha_{x_0})}}. \tag{3.D.7}$$

This provides constraints on the values that $D(Q_{\boldsymbol{\mu}^*}\|Q_{\boldsymbol{\mu}^\circ})$ in (3.D.5) can take. As we also show below, it implies that for every $\boldsymbol{\mu}^* \in \mathtt{M}_q$, every $A > 0$,

$$\sup_{\boldsymbol{\mu} \in \mathtt{M}_q : \|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|_2 > A} \frac{D(Q_{\boldsymbol{\mu}^*} \| Q_{\boldsymbol{\mu}^\circ})}{\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|_2^2} < \infty, \tag{3.D.8}$$

verifying Condition 3 as soon as R has 5 or more moments.

Detailed Proofs for Example 8 To prove (3.D.6) and (3.D.7), we first need to prove the following formulas (recall $\boldsymbol{\mu}^\circ$ is a function of $\boldsymbol{\mu}$):

$$\sup_{\boldsymbol{\mu} \in \mathtt{M}_q} (\mu_1^\circ - \log \mu_2^\circ) = k(x_{0,1} - \log x_{0,2}), \tag{3.D.9}$$

$$\inf_{\boldsymbol{\mu} \in \mathtt{M}_q} (\mu_2^\circ - e^{\mu_1^\circ}) = k(x_{0,2} - e^{x_{0,1}}). \tag{3.D.10}$$

Proof of (3.D.9) Reasoning analogously to the Gaussian case, Example 5, we first consider the case that $\boldsymbol{\mu}$ is located on the boundary $\partial \mathtt{M}_q$ of $\mathtt{M}_q$.

In this limiting case, for every fixed $0 < k < 1$, we get

$$\delta_1 := \mu_1^\circ - \log \mu_2^\circ = (1-k) \log \mu_2 + k \cdot x_{0,1} - \log\left((1-k)\mu_2 + k \cdot x_{0,2}\right).$$

Take the first derivative of δ_1 w.r.t. μ_2, we get

$$\arg\max_{\boldsymbol{\mu} \in \partial \mathtt{M}_q} (\mu_1^\circ - \log \mu_2^\circ) = (\log x_{0,2}, x_{0,2}).$$

This implies (3.D.9) holds (with $\mathtt{M}_q$ replaced by $\partial \mathtt{M}_q$), by plugging $(\log x_{0,2}, x_{0,2})$ into the above formula. In the case of a $\boldsymbol{\mu}'$ not located on the boundary (an instance is shown with corresponding $\boldsymbol{\mu}'^\circ$ in Figure 3.1(b)), we may consider the line connecting x_0 and $\boldsymbol{\mu}'$ in the figure; it intersects the boundary of $\mathtt{M}_q$ at some point $\boldsymbol{\mu}^b$. Letting $\boldsymbol{\mu}'^\circ = (1-k')\boldsymbol{\mu}^b + k' x_0$, we have $k' > k$ since $\boldsymbol{\mu}'^\circ = (1-k)\boldsymbol{\mu}' + k \cdot x_0$. Using the above worst-case result, we get, in terms of Figure 3.1 (b),

$$\delta_1' \leq k'(x_{0,1} - \log x_{0,2}) < k(x_{0,1} - \log x_{0,2}),$$

where the last inequality holds since $x_{0,1} - \log x_{0,2} < 0$. (3.D.9) now follows.

Proof of (3.D.10) We still consider the worst-case, $\boldsymbol{\mu}$ on the boundary, and the case with $\boldsymbol{\mu}'$ in the interior of $\mathtt{M}_q$, as above. In the boundary case, for every fixed $0 < k < 1$, we have:

$$\delta_2 := \mu_2^\circ - e^{\mu_1^\circ} = (1-k)e^{\mu_1} + k \cdot x_{0,2} - \exp\left((1-k)\mu_1 + k \cdot x_{0,1}\right).$$

Take the first derivative of δ_2 w.r.t. μ_1, we get $\arg\min_{\boldsymbol{\mu} \in \partial \mathtt{M}_q} (\mu_2^\circ - e^{\mu_1^\circ}) = (x_{0,1}, e^{x_{0,1}})$. Then this implies (3.D.10) holds (with $\mathtt{M}_q$ replaced by $\partial \mathtt{M}_q$) by plugging $(x_{0,1}, e^{x_{0,1}})$ into the above formula. The case with $\boldsymbol{\mu}' \in \mathtt{M}_q$, not on the boundary, can now be proved in the same way as (3.D.9); we omit further details.

Proof of (3.D.6) and (3.D.7) $\psi(\alpha^\circ) - \log \alpha^\circ$ is increasing and $\psi(\alpha^\circ) - \log \alpha^\circ = \mu_1^\circ - \log \mu_2^\circ \leq k(x_{0,1} - \log x_{0,2})$ from (3.D.9), which implies that α° is bounded by some α_{x_0} depending on x_0. Further, $(\alpha^\circ - e^{\psi(\alpha^\circ)})\beta^\circ = \mu_2^\circ - e^{\mu_1^\circ} \geq k(x_{0,2} - e^{x_{0,1}})$ from (3.D.10), $\alpha^\circ - e^{\psi(\alpha^\circ)}$ is increasing and $0 < \alpha^\circ \leq \alpha_{x_0}$, which implies $\beta^\circ \geq \frac{k(x_{0,2}-e^{x_{0,1}})}{\alpha_{x_0}-e^{\psi(\alpha_{x_0})}}$.

Proof of (3.D.8) We have $||\boldsymbol{\mu}-\boldsymbol{\mu}^*||_2^2 = (\psi(\alpha)+\log\beta-\psi(\alpha^*)-\log\beta^*)^2+(\alpha\beta-\alpha^*\beta^*)^2$. Reasoning analogously to the Gaussian case (3.D.3), we need to show that for any sequence $(\boldsymbol{\mu}_{[m]})_{m\in\mathbb{N}}$ with all $\boldsymbol{\mu}_{[m]} \in \{\boldsymbol{\mu} \in \mathtt{M}_q : \|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|_2 \geq A\}$, that tends to the boundary of $\mathtt{M}_q$, we have

$$\lim_{m\to\infty} \frac{D(Q_{\boldsymbol{\mu}^*} \| Q_{(\boldsymbol{\mu}^\circ)_{[m]}})}{\|\boldsymbol{\mu}_{[m]} - \boldsymbol{\mu}^*\|_2^2} < \infty.$$

We already know $0 < \alpha^\circ \leq \alpha_{x_0}$ and $\beta^\circ \geq \frac{k(x_{0,2}-e^{x_{0,1}})}{\alpha_{x_0}-e^{\psi(\alpha_{x_0})}}$, so the above limit satisfies, for some constant C,

$$\lim_{m\to\infty} \frac{D(Q_{\boldsymbol{\mu}^*} \| Q_{(\boldsymbol{\mu}^\circ)_{[m]}})}{\|\boldsymbol{\mu}_{[m]} - \boldsymbol{\mu}^*\|_2^2} = O\left(\frac{\alpha^\circ_{[m]}\left(\log \beta^\circ_{[m]} + C\right) + \log\Gamma(\alpha^\circ_{[m]})}{(\psi(\alpha_{[m]}) - \log\alpha_{[m]} + \log\beta_{[m]})^2 + (\alpha_{[m]}\beta_{[m]})^2} \right) \tag{3.D.11}$$

$$= O\left(\frac{\alpha^\circ_{[m]} \log \beta^\circ_{[m]}}{(\psi(\alpha_{[m]}) + \log\beta_{[m]})^2} \right) + O\left(\frac{\log\Gamma(\alpha^\circ_{[m]})}{(\psi(\alpha_{[m]}) + \log\beta_{[m]})^2} \right),$$

where we also plugged in (3.D.5) and the definition of $\boldsymbol{\mu}$ in terms of α, β. Any sequence tending to the boundary has $\mu_2^\circ - e^{\mu_1^\circ} \to 0$ or $\mu_2^\circ \to \infty$ or $\mu_1^\circ \to \infty$, i.e. $\alpha^\circ \to 0$ or $\alpha^\circ \to \infty$ or $\beta^\circ \to 0$ or $\beta^\circ \to \infty$. Again using that α° is bounded above and β° is bounded below, we just need to check the cases (a) $\beta^\circ \to \infty, \alpha^\circ \to \alpha^*$ with $\alpha^* \notin \{0, \infty\}$, (b) $\beta^\circ \to \beta^*$, with $\beta^* \notin \{0,\infty\}$, $\alpha^\circ \to 0$, and (c) $\beta^\circ \to \infty, \alpha^\circ \to 0$.

In Case (a), $\beta^\circ \to \infty$, and then $\mu_1^\circ = \psi(\alpha^\circ) + \log\beta^\circ \to \infty$, so $\psi(\alpha)+\log\beta = \mu_1 \to \infty$ since $\mu_1^\circ = (1-k)\mu_1 + k \cdot x_{0,1}$ and then the first term in (3.D.11) is $o(1)$ and the second $O(1)$. In Case (b), the first term in (3.D.11) is $o(1)$ and we can evaluate the second term by L'Hôpital's rule:

$$\lim_{\alpha^\circ\to 0} \frac{\log\Gamma(\alpha^\circ)}{\psi(\alpha^\circ)^2} = \lim_{\alpha^\circ\to 0} \frac{\psi(\alpha^\circ)}{2\psi(\alpha^\circ)\cdot\psi'(\alpha^\circ)} = \lim_{\alpha^\circ\to 0} \frac{1}{2\psi'(\alpha^\circ)} < \infty,$$

since $\psi'(\alpha^\circ) = \sum_{k=0}^{\infty} \frac{1}{(\alpha^\circ+k)^2}$. It remains to show Case (c); but this follows by combining Case (a) and Case (b) above.

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 $\mathcal{H}_0(\text{IID})$: 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, $\mathtt{M}$ 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 $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k) \in \mathtt{M}^k$ 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 $\mathtt{M}^k$, so that all results stated for $\boldsymbol{\mu}$ hold for any element of $\mathtt{M}^k$. 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 $\mu_0 \in \mathtt{M}$. 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 $\mu \in \mathtt{M}'$ for some sub-interval $\mathtt{M}' \subset \mathtt{M}$. 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. $S_{\text{GRO}(\boldsymbol{\mu}^1)}\mathcal{P}$ is defined as

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

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

Lemma 7. *For every set of distributions $\mathcal{P}$ on $\mathcal{X}$, $S_{\text{GRO}(\boldsymbol{\mu}^1)}\mathcal{P}$ 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 $S_{\text{GRO}(\boldsymbol{\mu}^1)}\mathcal{P}$ 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 $S_{\text{GRO}(\boldsymbol{\mu}^1)}\mathcal{P}$ 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 $S_{\text{RIPR}}, S_{\text{COND}}, S_{\text{MIX}}$ 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 $\text{CONV}(\mathcal{M})$ 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 $S_{\text{RIPR}} = S_{\text{PSEUDO}}$, see below). For all other $\mathcal{M}$, i.e. whenever the infimum is not achieved at all or is in $\text{CONV}(\mathcal{M}) \setminus \mathcal{M}$, 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:

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

Next, with W a distribution on $\mathtt{M}$, 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 $P_W \in \textsc{conv}(\mathcal{M})$. 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. $D(P_{\mu_1,\ldots,\mu_k} \| P'_{W_0}) = \inf_{P \in \textsc{conv}(\mathcal{M})} D(P_{\mu_1,\ldots,\mu_k} \| P) + \epsilon'$ for some $0 \leq \epsilon' < \epsilon$. 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 $\sup_{\mu_0 \in \mathtt{M}} \mathbb{E}_{P_{\mu_0,\ldots,\mu_0}}[S] = 1 + \delta$ 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 $\mu^*_0 \in \mathtt{M}$. Clearly, we must have $\mu^*_0 := \arg\min_{\mu_0 \in \mathtt{M}} D(P_{\boldsymbol{\mu}} \| P_{\langle \mu_0 \rangle})$. 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 $\mathtt{M}' \subset \mathtt{M}$ with μ^*_0 in the interior of $\mathtt{M}'$ so that S_{PSEUDO} is an e-variable for $\mathcal{H}_0(\mathcal{M}')$, where $\mathcal{M}' = \{P_\mu : \mu \in \mathtt{M}'\}$.*

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

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

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 $S_{\text{COND}} = 1$ 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

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

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 $\mathcal{M}$,

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

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*

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

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

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 $S_{\text{RIPR}} \neq S_{\text{PSEUDO}}$ (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 $\mathcal{H}_0(\text{IID})$) 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

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

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*

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

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 $\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log S_{\text{COND}} - \log S_{\text{MIX}}\right]$ can be written as*

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

4.4 Growth Rate Comparison for Specific Exponential Families

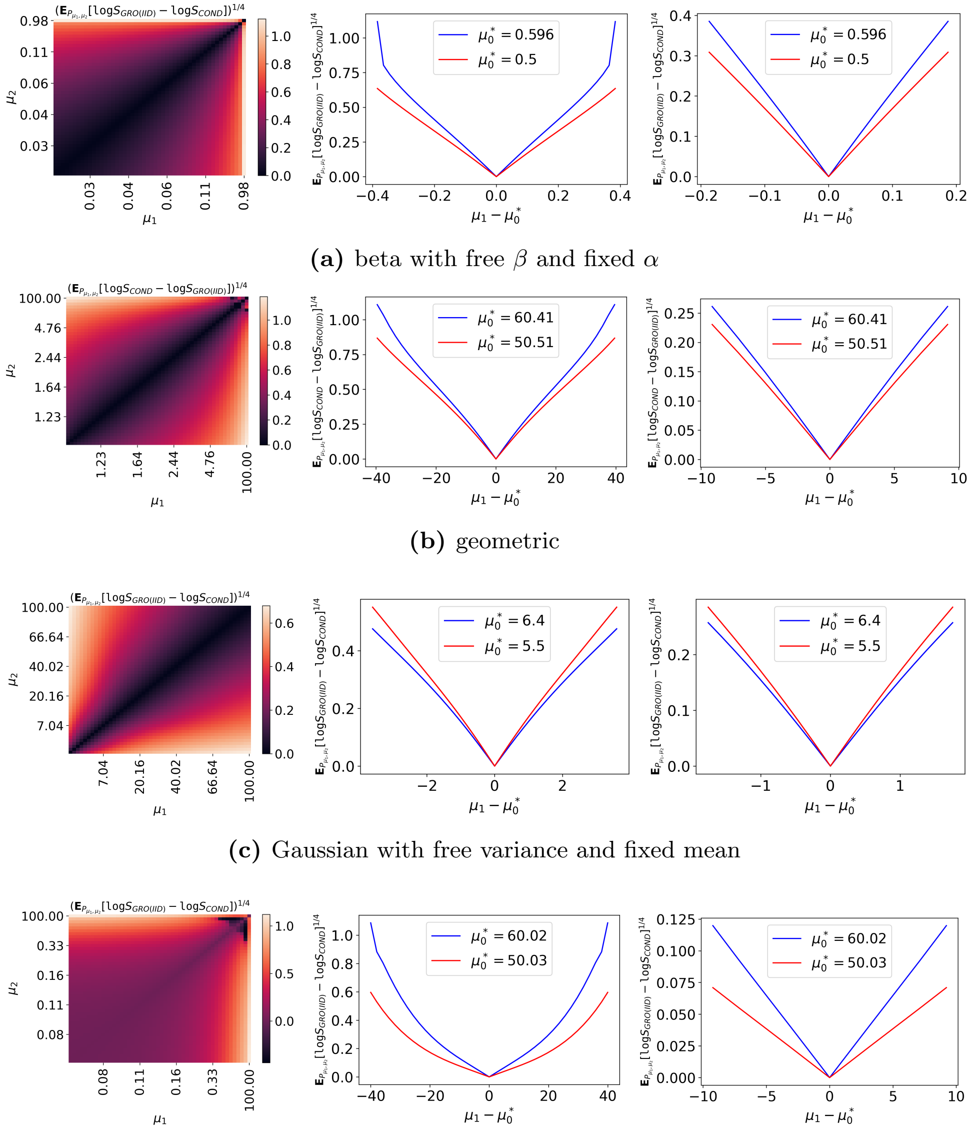
We will now establish more precise relations between the four (pseudo-) e-variables in k-sample tests for several standard exponential families, namely those listed in Table 4.1 and a few related ones, as listed at the end of this section. For each family $\mathcal{M}$ under consideration, we give proofs for which different e-variables are the same, i.e. $S = S'$, where $S, S' \in \{S_{\text{RIPR}}, S_{\text{COND}}, S_{\text{MIX}}, S_{\text{PSEUDO}}\}$. Whenever we can prove that $S_{\text{RIPR}} \neq S$ for another e-variable $S \in \{S_{\text{COND}}, S_{\text{MIX}}\}$, we can infer that $S_{\text{RIPR}} \succ S$ 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 $S_{\text{MIX}} \succ S_{\text{COND}}$ 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 $S, S' \in \{S_{\text{RIPR}}, S_{\text{PSEUDO}}, S_{\text{MIX}}, S_{\text{COND}}\}$. 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 $\mu_1, \ldots, \mu_k \in \mathtt{M}$, 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 Pareto(v, α) 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 $\delta =\mid \mu_1 - \mu_2 \mid /\sqrt{2} =\mid \mu_1 - \mu_0^* \mid \cdot\sqrt{2}$, 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 $S_{\text{RIPR}} \neq S_{\text{PSEUDO}}$ 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 $\text{CONV}(\mathcal{P})$ 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 $P_{(m)} \in \text{CONV}(\mathcal{P})$ 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 $D(Q\|P_{(m)}) \to \inf_{P \in \text{CONV}(\mathcal{P})} D(Q\|P)$. 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 $S_{\text{RIPR}} = S_{\text{PSEUDO}}$ 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})	$\sup_{\mu_0 \in \mathtt{M}} \mathbb{E}_{X_1, X_2 \sim P_{\mu_0, \mu_0}}[S_{\text{APPR}}]$
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	$(\frac{10}{3}, \frac{5}{4})$	0.51	(0.62, 0.31)	1.0047
geometric	$(\frac{10}{3}, \frac{5}{4})$	0.47	(1.84, 2.97)	1.0008
Gaussian with free variance and fixed mean	$(\frac{10}{3}, \frac{5}{4})$	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.

Acknowledgements We thank Rosanne Turner and Wouter Koolen for various highly useful discussions.

Declarations: financial support, lack of conflicting interests Partial financial support was received from China Scholarship Council State Scholarship Fund Nr.202006280045. The authors have no competing interests to declare that are relevant to the content of this article.

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_{RIP_R} 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. □

4.B.2 Proof of Proposition 11

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

$$
\begin{aligned}
g(\mu_0) &= \mathbb{E}_{p_{\langle \mu_0 \rangle}}\left[\prod_{i=1}^{k} \frac{p_{\mu_i}(X_i)}{p_{\mu_0^*}(X_i)}\right] = \prod_{i=1}^{k} \mathbb{E}_{Y \sim p_{\mu_0}}\left[\frac{p_{\mu_i}(Y)}{p_{\mu_0^*}(Y)}\right] \\
&= \prod_{i=1}^{k} \int \exp\left(\lambda(\mu_0)y - A\left(\lambda(\mu_0)\right)\right) \cdot \frac{\exp\left(\lambda(\mu_i)y - A\left(\lambda(\mu_i)\right)\right)}{\exp\left(\lambda(\mu_0^*)y - A\left(\lambda(\mu_0^*)\right)\right)} d\rho(y) \\
&= \prod_{i=1}^{k} \int \exp\left(\left(\lambda(\mu_i) + \lambda(\mu_0) - \lambda(\mu_0^*)\right) y - A\left(\lambda(\mu_i)\right) - A\left(\lambda(\mu_0)\right) + A\left(\lambda(\mu_0^*)\right)\right) d\rho(y) \\
&= \prod_{i=1}^{k} \exp\left(A\left(\lambda(\mu_0^*)\right) - A\left(\lambda(\mu_i)\right) - A\left(\lambda(\mu_0)\right)\right) \exp\left(B(\mu_i)\right) \\
&\qquad \cdot \int \exp\left(\left(\lambda(\mu_i) + \lambda(\mu_0) - \lambda(\mu_0^*)\right) y - B(\mu_i)\right) d\rho(y) \\
&= \prod_{i=1}^{k} \exp\left(A\left(\lambda(\mu_0^*)\right) - A\left(\lambda(\mu_i)\right) - A\left(\lambda(\mu_0)\right)\right) \exp\left(B(\mu_i)\right) \cdot 1 \\
&= \exp\left(kA\left(\lambda(\mu_0^*)\right) - \sum_{i=1}^{k} A\left(\lambda(\mu_i)\right) - kA\left(\lambda(\mu_0)\right) + \sum_{i=1}^{k} B(\mu_i)\right). \qquad (4.B.1)
\end{aligned}
$$

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

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

and

$$
\begin{aligned}
\frac{d^2}{d\mu_0^2} g(\mu_0) = &\left(\frac{d}{d\mu_0} g(\mu_0)\right) \cdot \frac{d}{d\mu_0}\left(\sum_{i=1}^{k} B(\mu_i) - kA\left(\lambda(\mu_0)\right)\right) \\
&+ g(\mu_0) \cdot \frac{d^2}{d\mu_0^2}\left(\sum_{i=1}^{k} B(\mu_i) - kA\left(\lambda(\mu_0)\right)\right) \\
=&g(\mu_0)\left(\sum_{i=1}^{k}(\mu_i + \mu_0 - \mu_0^*) - k\mu_0\right)^2 \qquad (4.B.3) \\
&+ g(\mu_0)\left(\sum_{i=1}^{k} \text{VAR}_{P_{\mu_i + \mu_0 - \mu_0^*}}[X] - k\text{VAR}_{P_{\mu_0}}[X]\right) \\
=&g(\mu_0)\left(\sum_{i=1}^{k} \text{VAR}_{P_{\mu_i + \mu_0 - \mu_0^*}}[X] - k\text{VAR}_{P_{\mu_0}}[X]\right) = g(\mu_0) \cdot f(\mu_0).
\end{aligned}
$$

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 $\mathtt{M}^* \subset \mathtt{M}$ with μ_0^* in the interior of $\mathtt{M}^*$ on which (4.B.1) is strictly convex. Then there must exist a point $\mu_0' \in \mathtt{M}^*$ 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 $\mathtt{M}^* \subset \mathtt{M}$ with μ_0^* in the interior of $\mathtt{M}^*$, 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:

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

We also have

$$\begin{aligned}&\frac{1}{k}\mathbb{E}_{X_1 \sim P_{\mu_0}}\left[\frac{p_{\mu_1}(X_1)}{p_0^*(X_1)}\right] + \cdots + \frac{1}{k}\mathbb{E}_{X_k \sim P_{\mu_0}}\left[\frac{p_{\mu_k}(X_k)}{p_0^*(X_k)}\right] \\ =&\frac{1}{k}\mathbb{E}_{X \sim P_{\mu_0}}\left[\frac{p_{\mu_1}(X)}{\frac{1}{k}\sum_{i=1}^{k} p_{\mu_i}(X)} + \cdots + \frac{p_{\mu_k}(X)}{\frac{1}{k}\sum_{i=1}^{k} p_{\mu_i}(X)}\right] = 1.\end{aligned} \tag{4.B.5}$$

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:

$$\begin{aligned}
\frac{1}{k}\sum_{i=1}^{k} r_i = \frac{k-1}{k}\cdot\frac{\sum_{i=1}^{k-1} r_i}{k-1} + \frac{r_k}{k} &\geq \left(\frac{\sum_{i=1}^{k-1} r_i}{k-1}\right)^{\frac{k-1}{k}} r_k^{\frac{1}{k}} \\
&= \left(\frac{k-2}{k-1}\cdot\frac{\sum_{i=1}^{k-2} r_i}{k-2} + \frac{r_{k-1}}{k-1}\right)^{\frac{k-1}{k}} r_k^{\frac{1}{k}} \\
&\geq \left(\frac{\sum_{i=1}^{k-2} r_i}{k-2}\right)^{\frac{k-2}{k}} r_{k-1}^{\frac{1}{k}} r_k^{\frac{1}{k}} \\
&\vdots \\
&\geq \left(\frac{r_1 + r_2}{2}\right)^{\frac{2}{k}} \prod_{i=3}^{k} r_i^{\frac{1}{k}} \geq \prod_{i=1}^{k} r_i^{\frac{1}{k}} \qquad (4.B.6)
\end{aligned}$$

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. □

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:

$$
\begin{aligned}
p_{\boldsymbol{\mu}}\left(x^{k-1}\middle|Z=z\right) &= \frac{p_{\boldsymbol{\mu}}\left(x^{k}\right)}{p_{\boldsymbol{\mu};[Z]}\left(z\right)} \\
&= \frac{\exp\left(\sum_{i=1}^{k}\left(\lambda(\mu_i)x_i - A(\lambda(\mu_i))\right)\right)}{\int_{y^{k-1}\in\mathcal{C}(z)}\exp\left(\sum_{i=1}^{k-1}\left(\lambda(\mu_i)y_i - A(\lambda(\mu_i)) + \lambda(\mu_k)X_k(y^{k-1},z)\right) - A(\lambda(\mu_k))\right)\right) d\rho(y^{k-1})} \\
&= \frac{\exp\left(\lambda(\mu_k)z + \sum_{i=1}^{k-1}(\lambda(\mu_i)-\lambda(\mu_k))x_i\right)}{\int_{y^{k-1}\in\mathcal{C}(z)}\exp\left(\lambda(\mu_k)z + \sum_{i=1}^{k-1}(\lambda(\mu_i)-\lambda(\mu_k))y_i\right) d\rho(y^{k-1})} \\
&= \frac{\exp\left(\sum_{i=1}^{k-1}(\lambda(\mu_i)-\lambda(\mu_k))x_i\right)}{\int_{y^{k-1}\in\mathcal{C}(z)}\exp\left(\sum_{i=1}^{k-1}(\lambda(\mu_i)-\lambda(\mu_k))y_i\right) d\rho(y^{k-1})}.
\end{aligned}
$$

□

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 $f_y(\delta) = \sum_{i=1}^{k} p_{\mu_i}(y)$ defined as in the theorem statement,

$$\begin{aligned}
F(\delta) :=& \mathbb{E}_{P_{\langle\mu_0\rangle+\boldsymbol{\alpha}\delta}} \left[\log S_{\text{PSEUDO}} - \log S_{\text{MIX}}\right] \\
=& \mathbb{E}_{P_{\boldsymbol{\mu}}} \left[\log \prod_{j=1}^{k} \left(\frac{1}{k}\sum_{i=1}^{k} p_{\mu_i}(X_j)\right) - \log p_{\langle\mu_0\rangle}(X^k)\right] \\
=& \mathbb{E}_{P_{\boldsymbol{\mu}}} \left[\sum_{j=1}^{k} \log f_{X_j}(\delta) - \sum_{j=1}^{k} \log p_{\mu_0}(X_j)\right] - k\log k \\
\overset{(a)}{=}& \sum_{j=1}^{k} \mathbb{E}_{X\sim P_{\mu_j}} \left[\log f_X(\delta) - \log p_{\mu_0}(X)\right] - k\log k \\
\overset{(b)}{=}& \overbrace{\int_{y\in\mathcal{X}} f_y(\delta)\log f_y(\delta)d\rho(y)}^{F_1(\delta)} + \overbrace{\left(-\int_{y\in\mathcal{X}} f_y(\delta)\log p_{\mu_0}(y)d\rho(y)\right)}^{F_2(\delta)} - k\log k, \quad (4.C.1)
\end{aligned}$$

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

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

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 $F_2(\delta)$.

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

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

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

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

$$\begin{aligned} F''_1(\delta) &= \int f''_y(\delta) d\rho(y) + \int \left(f''_y(\delta) \log f_y(\delta) + \frac{\left(f'_y(\delta)\right)^2}{f_y(\delta)} \right) d\rho(y) \\ &= \int \left(f''_y(\delta) \log f_y(\delta) + \frac{\left(f'_y(\delta)\right)^2}{f_y(\delta)} \right) d\rho(y) \\ F''_1(0) &= \int \left(f''_y(0) \log f_y(0) + \frac{\left(f'_y(0)\right)^2}{f_y(0)} \right) d\rho(y); \\ &= \int f''_y(0) \log p_{\mu_0}(y) d\rho(y) + \int_{y \in \mathcal{X}} \left(f''_y(0) \log k\right) d\rho(y) \qquad (4.C.5) \\ &= \int \left(f''_y(0) \log p_{\mu_0}(y)\right) d\rho(y), \end{aligned}$$

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

$$\begin{aligned} F''_2(\delta) =& - \int f''_y(\delta) \log p_{\mu_0}(y) d\rho(y) \ ; \ F''_2(0) = - \int f''_y(0) \log p_{\mu_0}(y) d\rho(y), \text{ so} \\ F''(0) =& F''_1(0) + F''_2(0) = 0. \end{aligned} \tag{4.C.6}$$

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

$$\begin{aligned}
F_1^{(3)}(\delta) &= \int \left(f_y^{(3)}(\delta) \log f_y(\delta) + \frac{f_y''(\delta) f_y'(\delta)}{f_y(\delta)} + \frac{2 f_y''(\delta) f_y'(\delta) f_y(\delta) - (f_y'(\delta))^3}{(f_y(\delta))^2} \right) d\rho(y) \\
F_1^{(3)}(0) &= \int f_y^{(3)}(0) \log f_y(0) d\rho(y) \\
&= \int f_y^{(3)}(0) \log p_{\mu_0}(y) d\rho(y) + \int f_y^{(3)}(0) \log k d\rho(y) \qquad (4.C.7) \\
&= \int f_y^{(3)}(0) \log p_{\mu_0}(y) d\rho(y) \\
F_2^{(3)}(\delta) &= - \int f_y^{(3)}(\delta) \log p_{\mu_0}(y) d\rho(y) \\
F_2^{(3)}(0) &= - \int f_y^{(3)}(0) \log p_{\mu_0}(y) d\rho(y), \text{ so } F^{(3)}(0) = F_1^{(3)}(0) + F_2^{(3)}(0) = 0,
\end{aligned}$$

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:

$$\begin{aligned}
F_1^{(4)}(\delta) = &\int \left(f_y^{(4)}(\delta) \log f_y(\delta) + \frac{f_y^{(3)}(\delta) f_y'(\delta)}{f_y(\delta)} \right) d\rho(y) \\
&+ \int 3 \cdot \frac{\left(f_y^{(3)}(\delta) f_y'(\delta) + (f_y''(\delta))^2 \right) f_y(\delta) - f_y''(\delta) \left(f_y'(\delta) \right)^2}{\left(f_y(\delta) \right)^2} d\rho(y) \\
&- \int \frac{3 \left(f_y(\delta) f_y'(\delta) \right)^2 \cdot f_y''(\delta) - 2 \left(f_y'(\delta) \right)^4 \cdot f_y(\delta)}{\left(f_y(\delta) \right)^4} d\rho(y) \; ; \qquad (4.C.8)
\end{aligned}$$

$$\begin{aligned}
F_1^{(4)}(0) &= \int \left(f_y^{(4)}(0) \log f_y(0) + \frac{3 \left(f_y''(0) \right)^2}{f_y(0)} \right) d\rho(y) \\
&= \int f_y^{(4)}(0) \log p_{\mu_0}(y) d\rho(y) + \log k \int_{y \in \mathcal{X}} f_y^{(4)}(0) d\rho(y) + \int_{y \in \mathcal{X}} \frac{3 \left(f_y''(0) \right)^2}{f_y(0)} d\rho(y) \\
&= \int f_y^{(4)}(0) \log p_{\mu_0}(y) d\rho(y) + \int_{y \in \mathcal{X}} \frac{3 \left(f_y''(0) \right)^2}{f_y(0)} d\rho(y),
\end{aligned}$$

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:

$$\begin{aligned}\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log S_{\text{PSEUDO}} - \log S_{\text{MIX}}\right] =& \frac{1}{4!}F^{(4)}(0)\delta^4 + o(\delta^4)\\ =& \frac{1}{8}\int_{y\in\mathcal{X}}\frac{\left(f_y''(0)\right)^2}{f_y(0)}d\rho(y)\cdot\delta^4 + o\left(\delta^4\right),\end{aligned}$$

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

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:

$$\begin{aligned}f(\delta) :=& \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log S_{\text{PSEUDO}} - \log S_{\text{COND}}\right]\\ =& \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log\frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle\mu_0\rangle}\left(X^k\right)} - \log\frac{p_{\boldsymbol{\mu}}\left(X^{k-1}\mid Z\right)}{p_{\langle\mu_0\rangle}\left(X^{k-1}\mid Z\right)}\right]\\ =& \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log\frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle\mu_0\rangle}\left(X^k\right)} - \log\frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle\mu_0\rangle}\left(X^k\right)} + \log\frac{\int_{\mathcal{C}(z)}p_{\boldsymbol{\mu}}\left(x^k\right)d\rho(x^{k-1})}{\int_{\mathcal{C}(z)}p_{\langle\mu_0\rangle}\left(x^k\right)d\rho(x^{k-1})}\right]\\ =& D\left(P_{\langle\mu_0\rangle+\boldsymbol{\alpha}\delta;[Z]}\|P_{\langle\mu_0\rangle;[Z]}\right).\end{aligned}$$

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.

$$\begin{aligned} f'(\delta) &= \int g'_z(\delta) \log \frac{g_z(\delta)}{g_z(0)} d\rho_{[Z]}(z) + \int g'_z(\delta) d\rho_{[Z]}(z) = \int g'_z(\delta) \log \frac{g_z(\delta)}{g_z(0)} d\rho_{[Z]}(z). \\ f''(\delta) &= \int g''_z(\delta) \log \frac{g_z(\delta)}{g_z(0)} d\rho_{[Z]}(z) + \int \frac{(g'_z(\delta))^2}{g_z(\delta)} d\rho_{[Z]}(z), \end{aligned}$$

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 $\mu_j = \mu_0 + \alpha_j \delta$,

$$\begin{aligned} g'_z(\delta) &= \int_{\mathcal{C}(z)} \frac{d}{d\delta} p_{\langle \mu_0 \rangle + \boldsymbol{\alpha}\delta}(x^k) d\rho(x^{k-1}) \\ &= \int_{\mathcal{C}(z)} \sum_{j=1}^{k} \prod_{i \in \{1,\ldots,k\} \setminus j} p_{\mu_i}(x_i) \frac{dp_{\mu_j}(x_j)}{d\delta} d\rho(x^{k-1}) \\ &= \int_{\mathcal{C}(z)} \sum_{j=1}^{k} p_{\mu_1,\ldots,\mu_{j-1},\mu_{j+1},\ldots,\mu_k}(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k) \frac{dp_{\mu_j}(x_j)}{d\mu_j} \frac{d\mu_j}{d\delta} d\rho(x^{k-1}) \\ &= \int_{\mathcal{C}(z)} \sum_{j=1}^{k} p_{\boldsymbol{\mu}}(x^k) \frac{d \log p_{\mu_j}(x_j)}{d\mu_j} \alpha_j d\rho(x^{k-1}) \\ &= \int_{\mathcal{C}(z)} \sum_{j=1}^{k} p_{\boldsymbol{\mu}}(x^k) \left(I(\mu_j) x_j - \mu_j I(\mu_j) \right) \alpha_j d\rho(x^{k-1}) \end{aligned}$$

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

$$\begin{aligned} g'_z(0) &= \int_{\mathcal{C}(z)} \sum_{j=1}^{k} p_{\langle \mu_0 \rangle}(x^k) \left(I(\mu_0) x_j - \mu_0 I(\mu_0) \right) \alpha_j d\rho(x^{k-1}) \\ &= \int_{\mathcal{C}(z)} p_{\langle \mu_0 \rangle}(x^k) I(\mu_0) \sum_{j=1}^{k} x_j \alpha_j d\rho(x^{k-1}) \qquad (4.C.10) \\ &= I(\mu_0) \cdot \int_{\mathcal{C}(z)} p_{\langle \mu_0 \rangle}(x^k) \sum_{j=1}^{k} x_j \alpha_j d\rho(x^{k-1}) \qquad (4.C.11) \end{aligned}$$

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 $k g_z'(0) = 0$ 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 $f^{(3)}(\delta)$:

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

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:

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

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:

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

which is what we had to prove. □

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

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

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

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

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). □

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).

□

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

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

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

Straightforward calculation gives

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

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 $\frac{1}{\beta_0^*} = \frac{1}{k}\sum\limits_{i=1}^{k} \frac{1}{\beta_i}$. We obtain, together with (4.D.4) and (4.D.5), that

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

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. □

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

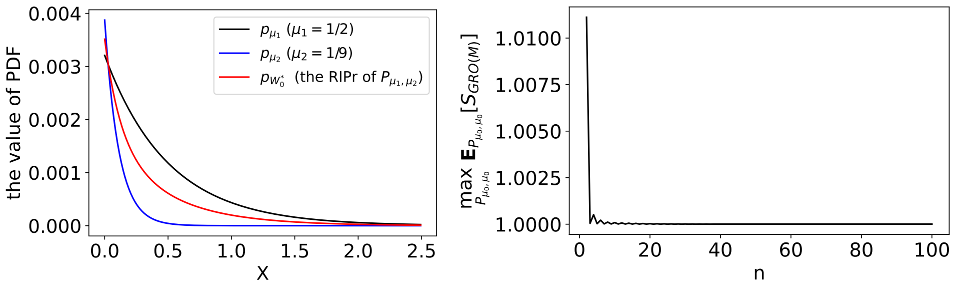


Figure 4.2: Exponential distribution. On the right, n represents number of iterations with Li's algorithm, starting at iteration 2

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

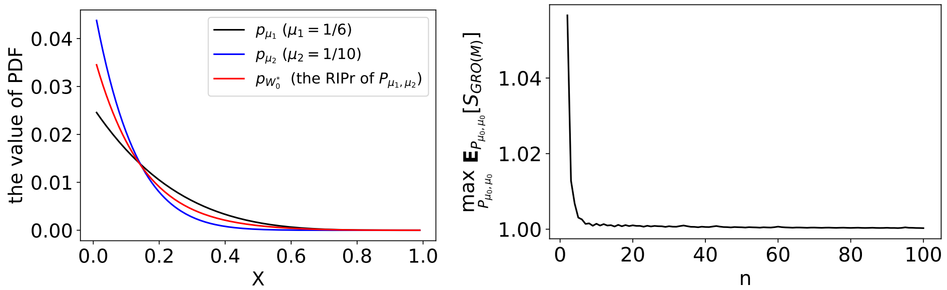


Figure 4.3: beta with free β and fixed α. On the right, n represents number of iterations with Li's algorithm, starting at iteration 2

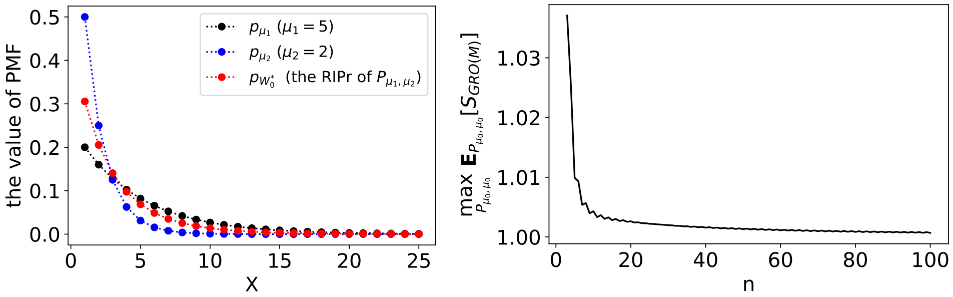


Figure 4.4: geometric distribution. On the right, n represents number of iterations with Li's algorithm, starting at iteration 3

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.

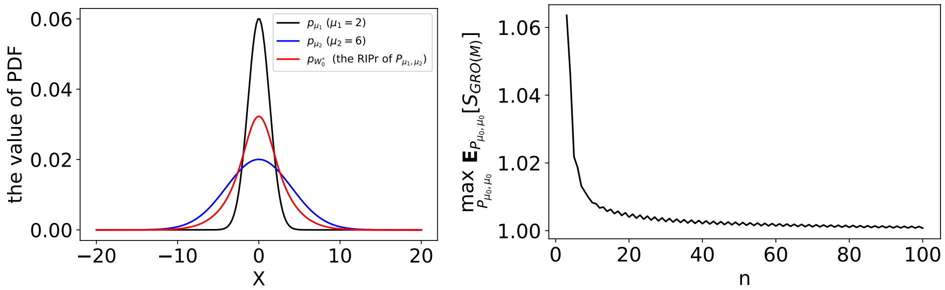


Figure 4.5: Gaussian with free variance and fixed mean. On the right, n represents number of iterations with Li's algorithm, starting at iteration 3

Distributions	(μ_1, μ_2)	α	(μ_{01}, μ_{02})	$\sup_{\mu_0 \in \mathtt{M}} \mathbb{E}_{X_1,X_2 \sim P_{\mu_0,\mu_0}}[S]$
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:

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

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

$$\begin{aligned}
\log \sum_{j=1}^{k} g_y(\lambda(\mu_j(\delta))) &\leq \log(\max_j g_y(\lambda(\mu_j(\delta)))k) \\
&= \max_j \log(g_y(\lambda(\mu_j(\delta)))) + \log k \\
&\leq \max_j \log(\max\{g_y(\lambda(\mu_j(\delta))), 1\}) + \log k \\
&\leq \sum_j \log(\max\{g_y(\lambda(\mu_j(\delta))), 1\}) + \log k \\
&\leq \sum_j |\lambda(\mu_j(\delta))y - \log A(\lambda(\mu_j(\delta)))| + \log k.
\end{aligned}$$

and

$$\begin{aligned}
\log \sum_{j=1}^{k} g_y(\lambda(\mu_j(\delta))) &= \log \frac{1}{k} \sum_{j=1}^{k} g_y(\lambda(\mu_j(\delta))) + \log k \\
&\geq \frac{1}{k} \sum_{j=1}^{k} \log g_y(\lambda(\mu_j(\delta))) + \log k \\
&= \frac{1}{k} \sum_{j=1}^{k} \lambda(\mu_j(\delta))y - A(\lambda(\mu_j(\delta))) + \log k.
\end{aligned}$$

Putting these together, we see that

$$\begin{aligned}
&|\log f_y(\delta)| \leq \\
&\max\left\{ \sum_j |\lambda(\mu_j(\delta))y - A(\lambda(\mu_j(\delta)))| + \log k, \left| \frac{1}{k} \sum_{j=1}^{k} (\lambda(\mu_j(\delta))y - A(\lambda(\mu_j(\delta)))) + \log k \right| \right\} \\
&\leq \sum_j |\lambda(\mu_j(\delta))y - A(\lambda(\mu_j(\delta)))| + \log k, \qquad (4.E.2)
\end{aligned}$$

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 $\mathtt{M}$ (see [13, Chapter 9, Theorem 9.1 and Eq. (2)]). Since $\mathtt{M}$ is open and for all $j = 1..k$, $\mu_j(0) = \mu_0 \in \mathtt{M}$, 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 $\mu \in \mathtt{M}$, 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,

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

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

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

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 $\mathtt{M}$ (see [13, Chapter 9, Theorem 9.1 and Eq. (2)]). Since this space is open and for all $j = 1..k$, $\mu_j(0) = \mu_0 \in \mathtt{M}$, 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

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

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

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

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

$$\begin{aligned}&\int \left(\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f_y(\delta) \log f_y(\delta) - f_y(\delta) \log g_y(\lambda(\mu_0))\right)^2 d\rho(y) = \\ &\int \left(f_y^{(s)}(\log f_y(\delta) - \log g_y(\lambda(\mu_0))) + f_y(\delta) \sum_u c_u \cdot b_y^{(s_2)}(\delta) \cdot \ldots \cdot b_y^{(s_u)}(\delta)\right)^2 d\rho(y) \\ &= \int (f_y^{(s)}(\log f_y(\delta) - \log g_y(\lambda(\mu_0))))^2 + \left(f_y(\delta) \sum_u c_u \cdot b_y^{(s_1)}(\delta) \cdot \ldots \cdot b_y^{(s_u)}(\delta)\right)^2 \\ &\qquad + f_y(\delta) f_y^{(s)}(\log f_y(\delta) - \log g_y(\lambda(\mu_0))) \sum_u c_u \cdot b_y^{(s_1)}(\delta) \cdot \ldots \cdot b_y^{(s_u)}(\delta) d\rho(y).\end{aligned}$$

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

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

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

Chapter 5

Growth-Optimal E-Variables and an extension to the multivariate Csiszár-Sanov-Chernoff Theorem

Abstract

We consider growth-optimal e-variables with maximal e-power, both in an absolute and relative sense, for simple null hypotheses for a d-dimensional random vector, and multivariate composite alternatives represented as a set of d-dimensional means $\mathtt{M}_1$. These include, among others, the set of all distributions with mean in $\mathtt{M}_1$, and the exponential family generated by the null restricted to means in $\mathtt{M}_1$. We show how these optimal e-variables are related to Csiszár-Sanov-Chernoff bounds, first for the case that $\mathtt{M}_1$ is convex (these results are not new; we merely reformulate them) and then for the case that $\mathtt{M}_1$ 'surrounds' the null hypothesis (these results are new).

5.1 Introduction

E-variables present a compelling alternative to traditional P-values, particularly in hypothesis tests involving optional stopping and continuation [42, 91, 76, 71, 46, 36]. As is well-known, there is a close connection between optimal rejection regions of

[0] This chapter is based on Peter Grünwald, Yunda Hao, and Akshay Balsubramani. Growth-Optimal E-Variables and an extension to the multivariate Csiszár-Sanov-Chernoff Theorem. arXiv preprint arXiv:2412.17554, 2024.

anytime-valid tests at fixed level α and optimal anytime-valid concentration inequalities [47]. In this paper we consider a variation of this connection in the context of a simple multivariate null and several types of composite alternatives. We study absolute and relative *GROW* ('growth-rate optimal in the worst-case') e-variables as introduced by [42], and we show how such e-variables are related to a concentration inequality which we call *Csiszár-Sanov-Chernoff* (CSC from now on). The 1-dimensional version of this inequality is well-known as a straightforward application of the Cramér-Chernoff method and is sometimes called the *generic* Chernoff bound. The multivariate version was apparently first derived by [26] as a (significant) strengthening of Sanov's classical theorem; we review this history in Section 5.2.1 beneath Theorem 10. Given all this we decided to name the bound 'Csiszár-Sanov-Chernoff'.

Formally, we consider a d-dimensional random vector $Y = (Y_1, \ldots, Y_d)$ supported on some (possibly finite or countable) subset $\mathcal{Y}$ of $\mathbb{R}^d$. Whenever we speak of 'a distribution for Y' we mean a distribution on $\mathcal{Y}$ equipped with its standard Borel σ-algebra. Then $\text{CONV}(\mathcal{Y})$, the convex hull of $\mathcal{Y}$, is the set of means for such distributions; we invariably assume that the zero-vector $(0, \ldots, 0)^\top$ (which we abbreviate to 0 whenever convenient) is contained in $\mathcal{Y}$. We then let the null hypothesis P_0 be a distribution for Y with mean equal to the zero vector: $\mathbb{E}_{P_0}[Y] = 0$. We fix a background measure ρ and assume that Y has a density p_0 relative to ρ under P_0 (we may in fact take ρ equal to P_0 itself without restricting the generality of our results). We assume that we are given a set of means $\mathtt{M}_1 \subset \text{CONV}(\mathcal{Y})$ and an alternative (i.e. a set of distributions on $\mathcal{Y}$) $\mathcal{H}_1$ that is *compatible* with the given $\mathtt{M}_1$, in the sense that, for all $\mu \in \text{CONV}(\mathcal{Y})$:

$$\mu \in \mathtt{M}_1 \Leftrightarrow \text{ there exists } P \in \mathcal{H}_1 \text{ with } \mathbb{E}_P[Y] = \mu. \tag{5.1.1}$$

We consider various such $\mathcal{H}_1$. In general, $\mathcal{H}_1$ is allowed to contain distributions that do not have densities, but whenever a $P_1 \in \mathcal{H}_1$ does have a density, it is denoted by small letters, i.e. p_1. We further invariably assume that P_0 and $\mathcal{H}_1$ are *separated* in terms of the mean. That is, $\inf_{\mu \in \mathtt{M}_1} \|\mu\|_2 > 0$, and finally, that Y admits a moment generating function under P_0. This is a strong assumption, but it is the only strong assumption we impose.

In Section 5.2 we consider the GROW (growth-rate-optimal in the worst-case over $\mathcal{H}_1$) e-variable S_{GROW} for this scenario, assuming either that $\mathcal{H}_1$ is the set $\mathcal{P}_1$ of *all* P with mean in some given *convex* set $\mathtt{M}_1$, or that $\mathcal{H}_1$ is the set $\mathcal{E}_1$ of all elements of the exponential family generated by P_0 with means in $\mathtt{M}_1$, or any $\mathcal{H}_1$ with $\mathcal{E}_1 \subset \mathcal{H}_1 \subset \mathcal{P}_1 1$ — it turns out that the GROW e-variables coincide for all such $\mathcal{H}_1$. We show how this result can be derived using the celebrated Csiszár-Topsøe Pythagorean theorem for relative entropy and how it leads to the basic CSC concentration inequality. We do not claim novelty for this section, which mostly contains re-formulations of results that are well-known in the information-theoretic (though perhaps not in the e-value) community. The real novelty comes in subsequent sections:

In Section 5.3 we move to the case that the *complement* of $\mathtt{M}_1$ is a connected, bounded set containing P_0 — a case that is more likely to arise in practical applications, is more closely related to the setting of the multivariate CLT, yet has, as far as we know, not been considered before when deriving CSC bounds, with the exception of

[51] who consider a variation of this setting (we return to their results in the final Section 5.6). We denote this as the *surrounding* $\mathcal{H}_1$ case, since P_0 is 'surrounded' by $\mathcal{H}_1$. We can extend the previous S_{GROW} e-variable to this case in two ways. We may either look at the straightforward *absolute* extension of the GROW e-variable to the multivariate case, which we still denote by S_{GROW}; or we can determine a *relatively* optimal GROW e-variable S_{REL} that is as close as possible to the largest S_{GROW} among all e-variables S_{GROW} that can be defined on convex subsets of $\mathcal{H}_1$, where, in this paper, as in [49], we define relative optimality in a minimax-regret sense. We characterize S_{GROW} for the case that $d = 1$ (leaving the complicated case $d > 1$ as an open question), and we characterize S_{REL} for general dimension d. We then show that S_{REL} leads again to a CSC bound, Theorem 12 — and this CSC bound is new.

The CSC bound arrived at in Theorem 12 contains a minimax regret term MMREG, which may be hard to verify in practice. In typical applications, we will have $Y = n^{-1}\sum_{i=1}^{n} X_i$ with X_i i.i.d., for some fixed sample size n. Then, if the exponential family generated by P_0 is regular (as it will be in most cases), we know that Y is equal to $\hat{\mu}_{|X_1,\ldots,X_n}$, the maximum likelihood estimator for the generated family, given in its mean-value parameterization. We can then think of the CSC bound as a concentration inequality that bounds the probability of the MLE falling in some set. Based on this instantiation of Y, we provide, in Section 5.4, based on earlier work by [23, 80], asymptotic expressions of the minimax regret term MMREG_n as a function of n, and show that, under regularity conditions on the boundary of the set $\mathtt{M}_1$, it increases as

$$\frac{d-1}{2}\log n + O(1),$$

It is no coincidence that this term is equal to the BIC/MDL model complexity for $d-1$-dimensional statistical family: it turns out that the boundary of $\mathtt{M}_1$ is the relevant quantity here, and it defines a $(d-1)$-dimensional exponential family embedded within the d-dimensional family generated by P_0. We show how this result gives us an asymptotic expression for the absolute GROW e-variable S_{GROW} after all, provided that the complement of $\mathtt{M}_1$ is a Kullback-Leibler ball around P_0.

This paper is still a work in progress. In the final section we provide additional discussion of the results, a comparison to the multivariate Central Limit Theorem, and we indicate the future work we would like to add to our current results.

5.1.1 Background on GROW e-variables

Since it will help provide the right context, in this — and only in this — subsection we allow composite null hypotheses $\mathcal{H}_0$. Each $P \in \mathcal{H}_0$ is then a distribution for Y.

Definition 7. A nonnegative statistic $S = s(Y)$ is called an *e*-variable relative to $\mathcal{H}_0$ if

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

Let $\mathcal{S}_0$ be the set of all e-variables that can be defined relative to $\mathcal{H}_0$ and such that $\mathbb{E}_P[\log S]$ is well-defined as an extended real number for all $P \in \mathcal{H}_1$, where we adopt the convention that $\log \infty = \infty$ and $\log 0 = \infty$. 'Well-defined' means that we may have $\mathbb{E}_P[(\log S) \vee 0] = \infty$ or $\mathbb{E}_P[(\log S) \wedge 0] = -\infty$ but not both. [42] defines the *worst-case optimal expected capital growth rate* (GROW) as

$$\text{GROW} := \sup_{S : S \in \mathcal{S}_0} \inf_{P \in \mathcal{H}_1} \mathbb{E}_P[\log S], \tag{5.1.2}$$

where $\mathbb{E}_P[\log S]$ is the so-called *growth rate* of S under $P \in \mathcal{H}_1$. The GROW E-variable, denoted as S_{GROW}, if it exists, is the e-variable achieving the supremum above. We refer to [42, 71] for extensive discussion on why this is, in a particular sense, the *optimal* e-variable that can be defined for the given testing problem. As a special case of their main result, Grünwald et al. [42, Theorem 2] show the following:

Theorem 9. [42, Theorem 2, Special Case] *Suppose that (a) $D(P_1 \| P_0) < \infty$ for all $P_1 \in \mathcal{H}_1, P_0 \in \mathcal{H}_0$ and (b)*

$$\min_{P_1 \in \text{CONV}(\mathcal{H}_1)} \min_{P_0 \in \text{CONV}(\mathcal{H}_0)} D(P_1 \| P_0) \tag{5.1.3}$$

is achieved by some P_1^*, P_0, then we have

$$\sup_{S \in \mathcal{S}_0} \inf_{P \in \mathcal{H}_1} \mathbb{E}_{Y \sim P}[\log S] = D(P_1^* \| P_0) = \text{GROW} = \inf_{P \in \mathcal{H}_1} \mathbb{E}_{Y \sim P}\left[\log \frac{p_1^*(Y)}{p_0(Y)}\right], \tag{5.1.4}$$

and S_{GROW}, achieving (5.1.2), is therefore given by $S_{\text{GROW}} = \dfrac{p_1^*(Y)}{p_0(Y)}$. (5.1.5)

Here p_1^* is the density of P_1^*, which exists by the finite KL assumption.

5.1.2 Simple $\mathcal{H}_0$ and the Pythagorean Property

Most recent work in e-variable theory has concentrated on the case of composite $\mathcal{H}_0$ and simple $\mathcal{H}_1$ [71]. Throughout this paper we consider the reverse case, simple $\mathcal{H}_0 = \{P_0\}$ and composite $\mathcal{H}_1$. Now the problem clearly simplifies and in fact, a lot more has been known about this special case since the 1970s, albeit expressed in the different language of data-compression: in a landmark paper, Topsøe [84, Theorem 9] proved a minimax result for relative entropy which (essentially) implies (5.1.4) for the simple $\mathcal{H}_0$ case. In fact, his result even implies that a distribution P_1^* such that (5.1.4) and (5.1.5) hold exists under much weaker conditions, in particular condition (b) above is not needed: P_1^* exists even if the minimum in $\min_{P \in \text{CONV}(\mathcal{H}_1)} D(P_1 \| P_0)$ is not achieved. The key result that Topsøe used to prove his version of (5.1.4) and (5.1.5) is Theorem 8 of his paper, a version of the *Pythagorean theorem* for KL divergence originally due to Csiszár [25, 24, 27] . We will re-state this result and explicitly use it to re-derive a version of (5.1.4) and (5.1.5) that is slightly stronger than Topsøe's and better suited to our needs (Topsøe's Theorem 9 still assumes condition (a); our derivation weakens it).

The Pythagorean theorem expresses that in the following sense, the KL divergence behaves like a squared Euclidean distance: for arbitrary P_0 and $\mathcal{H}_1$ as above, we have as long as $\mathcal{H}_1$ is *convex* and $\inf_{P_1 \in \mathcal{H}_1} D(P_1 \| P_0) < \infty$, that there exists a probability distribution P_1^*, called the *information projection of* P_0 *on* $\mathcal{H}_1$, that satisfies:

$$\text{for all } P \in \mathcal{H}_1\text{: } D(P\|P_0) \geq D(P\|P_1^*) + D(P_1^*\|P_0) \tag{5.1.6}$$

for every $Q_1, Q_2, \ldots \in \mathcal{H}_1$ with $\lim_{j\to\infty} D(Q_j\|P_0) = \inf_{P\in\mathcal{H}_1} D(P_1\|P_0)$, we have :

$$\lim_{j\to\infty} D(Q_j\|P_1^*) = 0. \tag{5.1.7}$$

$$D(P_1^*\|P_0) \leq \inf_{P\in\mathcal{H}_1} D(P\|P_0). \tag{5.1.8}$$

In standard cases, the final inequality will hold with equality; in particular we have equality if $\min_{P_1\in\mathcal{H}_1} D(P_1\|P_0)$ is achieved.

We call (5.1.6) the *Pythagorean property.* Note that it is *implied* by convexity of $\mathcal{H}_1$ and finiteness of $\inf D(P_1\|P_0)$, but it may sometimes hold even if $\mathcal{H}_1$ is not convex.

We now show, slightly generalizing Topsøe's result, how the Pythagorean property (5.1.6) implies a version of [42]'s theorem for simple $\mathcal{H}_0$ (in fact, in the reformulation as a minimax theorem for data compression, the Pythagorean property is in fact *equivalent* to the minimax statement but we will not need that fact here; see [37, Section 8] for an extended treatment of this equivalence).

Proposition 13. [Pythagoras $\Rightarrow$ $S_{\mathbf{GROW}} = p_1^*/p_0$] *Suppose that* $\mathcal{H}_0 = \{P_0\}$ *and let* $\mathcal{H}_1$ *be any set of distributions (not necessarily convex!) for* Y *such that* $\inf_{P\in\mathcal{H}_1} D(P\|P_0) < \infty$*, and suppose that* P_1^* *is such that (5.1.6)–(5.1.8) holds, so that it has a density* p_1^**. Further assume that, with 'well-defined' defined as above (5.1.2),*

$$\mathbb{E}_P[\log p_1^*(Y)/p_0(Y)] \text{ is well-defined for all } P \in \mathcal{H}_1, \tag{5.1.9}$$

so that $p_1^*(Y)/p_0(Y) \in \mathcal{S}_0$*. Then we have:*

$$\sup_{S\in\mathcal{S}_0} \inf_{P\in\mathcal{H}_1} \mathbb{E}_{Y\sim P}[\log S] = D(P_1^*\|P_0) = \inf_{P\in\mathcal{H}_1} \mathbb{E}_{Y\sim P}\left[\log \frac{p_1^*(Y)}{p_0(Y)}\right] \tag{5.1.10}$$

so that $S_{\text{GROW}} = p_1^*(Y)/p_0(Y)$.

Proof. Since we deal with a simple null, it holds that (as shown by [42]) any $S \in \mathcal{S}_0$ must be of the form $q(Y)/p_0(Y)$ for some sub-probability density q relative to the measure ρ, and any such ratio defines an e-variable: the notions are equivalent. Here 'sub-probability' means that $\int q(y)d\rho(y) \leq 1$ is allowed to be smaller than 1. We thus have, with $\sup_q$ denoting the supremum of all sub-probability density functions q for Y,

$$\begin{aligned}\sup_{S\in\mathcal{S}_0} \inf_{P\in\mathcal{H}_1} \mathbb{E}_{Y\sim P}[\log S] &= \sup_q \inf_{P\in\mathcal{H}_1} \mathbb{E}_{Y\sim P}\left[\log \frac{q(Y)}{p_0(Y)}\right] \\ &\leq \sup_q \mathbb{E}_{Y\sim P_1^*}\left[\log \frac{q(Y)}{p_0(Y)}\right] = D(P_1^*\|P_0).\end{aligned} \tag{5.1.11}$$

At the same time, the Pythagorean inequality (5.1.6) gives, by simple re-arranging of the logarithmic terms, that for all $P \in \mathcal{H}_1$ with $D(P\|P_0) < \infty$:

$$\mathbb{E}_{Y\sim P}\left[\log \frac{p_1^*(Y)}{p_0(Y)}\right] \geq \mathbb{E}_{Y\sim P_1^*}\left[\log \frac{p_1^*(Y)}{p_0(Y)}\right] = D(P_1^*\|P_0), \tag{5.1.12}$$

whereas if $D(P\|P_0) = \infty$ then by assumption (5.1.8), the right-hand side of (5.1.12) is finite. (5.1.6) then implies $D(P\|P_1) = \infty$, and. because by assumption (5.1.9), the left-hand side of (5.1.12) is well-defined, we can again re-arrange (5.1.6) to give (5.1.12). Thus, we have shown that for all $P \in \mathcal{H}_1$, (5.1.12) holds. But then

$$\sup_{S\in\mathcal{S}_0} \inf_{P\in\mathcal{H}_1} \mathbb{E}_{Y\sim P}[\log S] \geq \inf_{P\in\mathcal{H}_1} \mathbb{E}_{Y\sim P}\left[\log \frac{p_1^*(Y)}{p_0(Y)}\right] \geq D(P_1^*\|P_0). \tag{5.1.13}$$

Together, (5.1.11) and (5.1.13) imply the result. □

While Condition (5.1.9) may look complicated, it is immediately verified to hold if $D(P_1\|P_0) < \infty$ for all $P_1 \in \mathcal{H}_1$ but also under Condition ALT-$\mathcal{H}_1$ presented in the next section, which allows for $\mathcal{H}_1$ to even contain distributions P_1 with $P_1 \not\ll P_0$ (see Example 11 for an instance of this).

5.1.3 The Rôle of Exponential Families

We shall from now on tacitly assume that the convex support of P_0 is d-dimensional (see [19, Chapter 1] for the precise definition of 'convex support'). This is without loss of generality: if Y takes values in $\mathbb{R}^d$ yet the convex support does not have dimension d, it must have dimension $d' < d$, and then we can replace Y by d'-dimensional Y' that is an affine function of Y and work with Y' instead. Combined with our earlier assumption that Y has a moment generating function under P_0, it follows [19, Chapter 1] that Y and P_0 jointly *generate* a d-dimensional natural exponential family $\mathcal{E} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$: a set of distributions for Y with parameter space $\Theta \subseteq \mathbb{R}^d$. Each distribution P_θ has density p_θ relative to ρ, given by:

$$p_\theta(Y) = \frac{1}{Z(\theta)} \exp\left(\theta^\top Y\right) \cdot p_0(Y), \tag{5.1.14}$$

where $Z(\theta)$ is the normalizing factor and p_0 is the density of the *generating distribution* P_0 and $\Theta = \{\theta : Z(\theta) < \infty\}$. From now on, we freely use standard properties, terminology and definitions concerning exponential families (such as 'carrier density' and so on), that can be found, in, for example, [13, 33, 19]. We will only mention these works, and then specific sections therein, when we refer to results that are otherwise hard to find.

Parameterization (5.1.14) is called the *canonical* or *natural* parameterization. As is well-known, exponential families can be re-parameterized in terms of the mean of Y. Thus, there is a 1-to-1 mapping $\mu : \Theta \to \mathtt{M}$, mapping each $\theta \in \Theta$ to $\mu(\theta) := \mathbb{E}_{P_\theta}[Y]$, with $\mathtt{M}$ being the *mean-value parameter space*. We let $\theta(\mu)$ be the inverse of this mapping and let $\bar{P}_\mu := P_{\theta(\mu)}$ with density $\bar{p}_\mu(Y) := p_{\theta(\mu)}(Y)$. Then we can equivalently write our exponential family as

$$\mathcal{E} = \{\bar{P}_\mu : \mu \in \mathtt{M}\}. \tag{5.1.15}$$

Without loss of generality, we assume that Y is defined such that $0 \in \mathtt{M}$ and the natural parameterization is such that $\theta(0) = 0, \mu(0) = 0$. Clearly the null hypothesis $\mathcal{H}_0 = \{P_0\}$ is given by the element of $\mathcal{E}$ corresponding to the parameter vector $\theta = 0$. The mean-value parameterization will be the most 'natural' one (no pun intended) to use to define the alternative. As said in the introduction, we restrict ourselves to cases in which Y has a moment generating function under P_0. Then $\mathtt{M}$ contains an open set around 0 and the exponential family (5.1.15) exists. We define the alternative $\mathcal{H}_1$ in terms of a given set of means $\mathtt{M}_1 \subset \textsc{conv}(\mathcal{Y})$, invariably satisfying:

Condition ALT-$\mathtt{M}_1$: (a) $\mathtt{M}_1$ is closed, and $\inf_{\mu \in \mathtt{M}_1} \|\mu\|_2 > 0$; and, (b), for all $\mu \in \mathtt{M}_1$, there is a $\mu' \in \mathtt{M} \cap \mathtt{M}_1$ that lies on the straight line connecting 0 and μ.

For the actual alternative hypothesis we then invariably further assume:

Condition ALT-$\mathcal{H}_1$: (a) $\mathcal{H}_1$ and $\mathtt{M}_1$ are *compatible* in the sense of (5.1.1), and (b) $\mathcal{E}_1 := \{\bar{P}_\mu : \mu \in \mathtt{M}_1 \cap \mathtt{M}\} \subset \mathcal{H}_1$ and, (c), for all $P \in \mathcal{H}_1$, $\mathbf{E}_{Y \sim P}[Y]$ is well-defined.

To appreciate these conditions, consider first the case that $\mathtt{M} = \textsc{conv}(\mathcal{Y})$, i.e. the mean-value parameter space of family $\mathcal{E}$ contains every possible mean. Then Condition ALT-$\mathtt{M}_1$ (a) says that $\mathtt{M}_1$ is separated from 0 and that it contains its boundary (note that it does not need to be bounded: for example, in the case that $\mathcal{E}$ is the 1-dimensional normal location family, having $\mathtt{M}_1 = [1, \infty)$ is perfectly fine). Condition ALT-$\mathtt{M}_1$ (b) holds automatically if $\mathtt{M} = \textsc{conv}(\mathcal{Y})$ (e.g. in the Gaussian location case); the example below illustrates the case that $\mathtt{M}$ is a strict subset of $\textsc{conv}(\mathcal{Y})$. Condition ALT-$\mathcal{H}_1$ (b) simply says that for every mean in $\mathtt{M}_1$, $\mathcal{H}_1$ contains the element of the exponential family $\mathcal{E}$ with that mean.

Example 10. As a very simple example, suppose that $Z_1, Z_2, \ldots$ are i.i.d. Bernoulli(p) for some $0 < p < 1$, and $X_i = 1/p$ if $Z_i = 1$ whereas $X_i = -1/(1-p)$ if $Z_i = 0$. Let $Y = n^{-1} \sum_{i=1}^n X_i$. Then $\textsc{conv}(\mathcal{Y}) = [-1/(1-p), 1/p]$ and according to P_0, Y is a linear transform of a $\textsc{bin}(n, p)$ random variable with $\mathbb{E}_{P_0}[Y] = 0$. Then Condition ALT-$\mathtt{M}_1$ (a) expresses that $\mathtt{M}_1$ must not contain a neighborhood of the mean of P_0 (i.e., $\mu = 0$), and (b) that it must not be restricted to singletons at the boundary (i.e. $\mu = 1/p$ or $\mu = -1/(1-p)$). However, for example, $\mathtt{M}_1 = [1/p - \epsilon, 1/p]$ for any $0 < \epsilon < 1/p$ satisfies Condition ALT-$\mathtt{M}_1$. We see that the condition only very minimally restricts the set of $\mathtt{M}_1$ that are allowed. It becomes more restrictive for the (very seldomly encountered!) case that the generated exponential family $\mathcal{E}$ is *irregular* [13]. By construction of $\mathcal{E}$, this is equivalent to it being *not steep*. For example, let Y be 1-dimensional and let P_0 have density $p_0(y) = \mathbf{1}_{y>1} \cdot (2/y^3)$ relative to Lebesgue measure. Then we get $\mathcal{E} = \{P_\theta \mid \theta \leq 0\}$ with $p_\theta(y) \propto \exp(\theta y) p_0(y)$. Then $\mathtt{M} = (1, 2]$ yet $\mathcal{Y} = (1, \infty)$ (from the fact that $\mathtt{M}$ is not open we immediately see that $\mathcal{E}$ is not regular). Condition ALT-$\mathtt{M}_1$ now requires that $\mathtt{M}_1$ contains a $\mu < 2$, even though $P_0(Y \geq b) > 0$ for any $b \geq 2$.

5.2 Convex $\mathtt{M}_1$

The Connection between Exponential Families and the Pythagorean Theorem Although exponential families are usually employed as families that are reasonable in their own right, as is well-known [25, 37] they can also be arrived at as characterizing the information projection P_1^* in the Pythagorean property above for certain $\mathcal{H}_1$. We will heavily use this characterization below. Variations of the following result (see Figure 5.1 for illustration) are well-known:

Proposition 14. [GROW e-variable is $\bar{p}_{\mu^*}/p_0$, with $\bar{P}_{\mu^*}$ an element of exponential family $\mathcal{E}$] *Let $\mathcal{H}_0 = \{P_0\}$ and $\mathtt{M}_1$ be such that Condition ALT-$\mathtt{M}_1$ holds. Furthermore let $\mathtt{M}_1$ be convex. Then there exists $\mu^* \in \mathtt{M} \setminus \{0\}$ uniquely achieving $\inf_{\mu \in \mathtt{M}_1} D(\bar{P}_\mu \| P_0)$, and we have:*

$$\min_{\mu \in \mathtt{M}_1 \cap \mathtt{M}} D(\bar{P}_\mu \| P_0) = \min_{\mu \in \mathtt{M}_1} D(\bar{P}_\mu \| P_0) = D(\bar{P}_{\mu^*} \| P_0) = {\theta^*}^\top \mu^* - \log Z(\theta^*) \tag{5.2.1}$$

with $\theta^* := \theta(\mu^*) \in \Theta$. Furthermore let $\mathcal{H}_1$ be convex and such that Condition ALT-$\mathcal{H}_1$ holds. Then the minimum in (5.2.1) further satisfies:

$$\inf_{P \in \mathcal{H}_1} D(P \| P_0) = D(\bar{P}_{\mu^*} \| P_0) \tag{5.2.2}$$

the minimum KL on the left being achieved uniquely by $\bar{P}_{\mu^*}$. As a consequence, $S_{\text{GROW}} = \bar{p}_{\mu^*}(Y)/\bar{p}_0(Y) \in \mathcal{S}_0$ and $\text{GROW} = D(\bar{P}_{\mu^*} \| P_0)$.

Proof. The KL divergence $D(\bar{P}_\mu \| P_0)$ is continuous in μ, has its overall minimum over $\textsc{conv}(\mathcal{Y})$ in the point $\mu = 0$ and is strictly convex. This implies (i) that by Condition ALT-$\mathtt{M}_1$(a), $\min_{\mu \in \mathtt{M}_1} D(\bar{P}_{\alpha\mu} \| P_0)$ is uniquely achieved for some μ^* on the boundary of $\mathtt{M}_1$. By Condition ALT-$\mathtt{M}_1$(b), the boundary of $\mathtt{M}_1$ is included in $\mathtt{M}$, so $\mu^* \in \mathtt{M}_1 \cap \mathtt{M}$. This yields the first two equations in (5.2.1). Writing out the densities in $D(\bar{P}_{\mu^*} \| \bar{P}_0)$ then gives the rightmost equality.

It remains to prove (5.2.2). Condition ALT-$\mathcal{H}_1$ implies that $\mathcal{H}_1$ contains a $\bar{P}_\mu \in \mathcal{E}$, hence $\mathtt{M}_1$ contains a $\mu \in \mathtt{M}$, and since $D(\bar{P}_\mu \| P_0) < \infty$ for all $\mu \in \mathtt{M}$, we have $\inf_{P \in \mathcal{H}_1} D(P \| P_0) < \infty$. Therefore, it suffices to show (5.2.2) with the infimum taken over $\{P \in \mathcal{H}_1 : D(P \| P_0) < \infty\}$. In particular all P in this set have a density p. Thus, fix any P in the set $\{P \in \mathcal{H}_1 : D(P \| P_0) < \infty\}$ and let $\mu = \mathbb{E}_P[Y]$. We first consider the case that $\mu \in \mathtt{M}$, so that $\bar{P}_\mu \in \mathcal{E}$ (in particular then also $\mu \in \mathtt{M} \cap \mathtt{M}_1$ and $\bar{P}_\mu = P_\theta$ with $\theta = \theta(\mu)$; note though that we may have $P \neq \bar{P}_\mu$). Straightforward rewriting and linearity of expectation gives

$$\begin{aligned} & D(P \| P_0) = \mathbb{E}_{Y \sim P}\left[\log \frac{p(Y)}{p_0(Y)}\right] \geq \mathbb{E}_P\left[\log \frac{p_\theta(Y)}{p_0(Y)}\right] = \mathbb{E}_P\left[\log \frac{1}{Z(\theta)} \cdot e^{\theta^\top Y}\right] = \\ & \theta^\top \mu - \log Z(\theta) = D(P_\theta \| P_0) = D(\bar{P}_\mu \| P_0) \geq \min_{\mu \in \mathtt{M}_1 \cap \mathtt{M}} D(\bar{P}_\mu \| P_0), \end{aligned} \tag{5.2.3}$$

the final inequality following because $\mu \in \mathtt{M} \cap \mathtt{M}_1$. Together with (5.2.1) this shows (5.2.2) for the case that $\mu \in \mathtt{M}$. It remains to consider the case that $\mu \not\in \mathtt{M}$. In that case, Condition ALT-$\mathtt{M}_1$ and ALT-$\mathcal{H}_1$ imply that there exists a $\mu' \in \mathtt{M} \cap \mathtt{M}_1$ and $\bar{P}_{\mu'} \in \mathcal{H}_1 \cap \mathcal{E}$ such that $\mu' = \alpha\mu$ for some $0 < \alpha < 1$. Retracing the steps of (5.2.3) with $\theta' = \theta(\mu')$ in the place of μ, we find

$$D(P\|P_0) \geq \theta'^{\top}\mu - \log Z(\theta') = f(1) \tag{5.2.4}$$

where, for $\gamma \in [0,1]$, we set $f(\gamma) = \mathbb{E}_{P_{\gamma\mu}}\left[\log \frac{\bar{p}_{\mu'}(Y)}{p_0(Y)}\right]$. Since $f(0)$ is minus a KL divergence, $f(0) < 0$. Also, $f(\alpha\mu) = f(\mu') > 0$, since $f(\mu')$ is a KL divergence. Since $f(\gamma)$ is linear in γ, it follows that $f(\gamma)$ is strictly increasing so $f(1) > f(\mu')$ and then (5.2.4) gives that $D(P\|P_0) \geq D(\bar{P}_\mu\|P_0)$ which again implies the result.

It remains to prove that $S_{\text{GROW}} := S$ with $S = \bar{p}_{\mu^*}(Y)/p_0(Y)$. For this, first note that for all $P \in \mathcal{H}_1$, we have $P(p_0(Y) > 0) = 1$ by definition (namely, $\mathcal{Y}$ is the support of Y under P_0), from which it follows that $P(S > 0) = 1$, so that $\log S = \theta^{*\top}Y - \log Z(\theta^*)$ whence $\mathbf{E}_{Y\sim P}[\log S]$ is well-defined; it follows that $S \in \mathcal{S}_0$. By Topsøe's result and the assumed convexity of $\mathcal{H}_1$ and finiteness of $D(\bar{P}_{\mu^*}\|P_0)$, we may now apply Proposition 13, and the result follows. □

Example 11. Since $\mathcal{E}$ is an exponential family, we know that all elements $P \in \mathcal{E}$ have the same support as $P_0 \in \mathcal{E}$, and, by definition of P_0, this support is equal to $\mathcal{Y}$. This implies that, even if $P \in \mathcal{H}_1$ puts positive mass on an outcome $y \in \mathcal{Y}$ that has mass 0 under P_0, then (because y must be in P_0's support), well-definedness (5.1.9) may still hold. For example, consider the case that $Y = \mathbb{R}$ and $P(\{0\}) = 1/2$ and $P \mid Y \neq 0 = N(0,1)$ is a standard normal, and $\mathcal{E}$ is the normal location family so that $\bar{P}_\mu = N(\mu, 1)$. We get $\mathbb{E}_P[\log \bar{p}_\mu(Y)/p_0(Y)] = (1/2)D(P_0\|\bar{P}_\mu) + \mu^2/2$, i.e. it is well-defined. On the other hand, if we were to allow P_0 defined on a sample space $\mathcal{Y}$ with Y's support under P_0 a strict subset of $\mathcal{Y}$, and we would take $P \in \mathcal{H}_1$ that put positive mass on an outcome that is not in the support of P_0, then $\bar{p}_\mu(Y)/p_0(Y)$ would evaluate to $0/0$ with positive P-probability, and S_{GROW} of the form above would be undefined. We avoid such issues by requiring $\mathcal{Y}$ to coincide with its support under P_0. We suspect that using the ideas of [58], we can even obtain well-defined growth expressions for this case, but will leave this for future work.

5.2.1 CSC (Chernoff-Sanov-Csiszár) for convex $\mathtt{M}_1$

Note that the only role $\mathcal{H}_1$ plays in the theorem below is to make S_{GROW} well-defined; the bounds further do not depend on the specific choice of $\mathcal{H}_1$ as long as $S_{\text{GROW}} = \bar{p}_{\mu^*}(Y)/p_0(Y)$.

Theorem 10. *Suppose P_0 and $\mathtt{M}_1$ are such that $\mathtt{M}_1$ is convex and Condition ALT-$\mathtt{M}_1$ holds so that, by Proposition 14, there exists $\mu^* \in \mathtt{M}$ minimizing $D(\bar{P}_\mu\|P_0)$ over $\mathtt{M}_1$ with $\bar{P}_{\mu^*} \in \mathcal{E}$. Let $\mathcal{H}_1$ be any set of distributions such that Condition ALT-$\mathcal{H}_1$ holds, so that, by Proposition 14, $S_{\text{GROW}}(Y) := \bar{p}_{\mu^*}(Y)/p_0(Y)$. Define*

$$\underline{D} := \inf_{\mu \in \mathtt{M}_1} D(\bar{P}_\mu\|P_0). \tag{5.2.5}$$

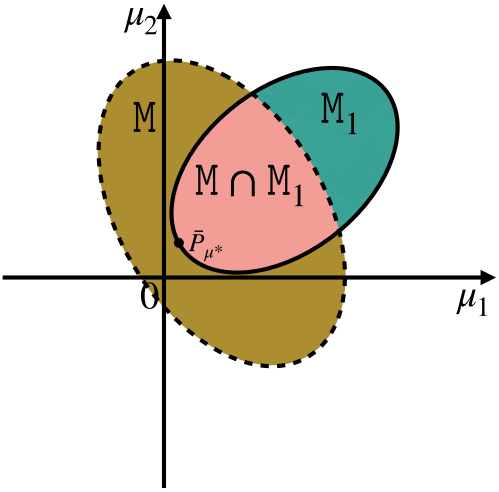


Figure 5.1: Convex $\mathtt{M}_1$. $\mathtt{M}_1$ is the mean parameter set that $\mathcal{H}_1$ is compatible with, and $\mathtt{M}$ is the mean parameter space of the exponential family generated from P_0. The setting of this figure satisfies Condition ALT-$\mathtt{M}_1$.

We have:

$$y \in \mathtt{M}_1 \Rightarrow S_{\text{GROW}}(y) \geq e^{\underline{D}} \tag{5.2.6}$$

so that

$$\mathbb{E}_{Y\sim P_0}\left[\mathbf{1}_{Y\in\mathtt{M}_1} \cdot e^{\underline{D}}\right] \leq \mathbb{E}_{Y\sim P_0}\left[\mathbf{1}_{Y\in\mathtt{M}_1} \cdot S_{\text{GROW}}\right] \leq \mathbb{E}_{Y\sim P_0}\left[S_{\text{GROW}}\right] \leq 1.$$

As a consequence, we have:

$$P_0(Y \in \mathtt{M}_1) = \mathbb{E}_{Y\sim P_0}\left[\mathbf{1}_{Y\in\mathtt{M}_1} \cdot e^{\underline{D}}\right] \cdot e^{-\underline{D}} \leq e^{-\underline{D}}, \tag{5.2.7}$$

and we also have, for the one-dimensional case with $Y \in \mathbb{R}$, for any $D > 0$ for which there exists $\mu^* \in \mathtt{M}$ such that $D(\bar{P}_{\mu^*}\|P_0) = D$, with $s = \text{sgn}(\mu^*)$,

$$P_0\left(\frac{\sup_{\mu\in\mathtt{M}} \bar{p}_\mu(Y)}{p_0(Y)} \geq e^D, \text{SGN}(Y) = s\right) = P_0\left(\frac{\bar{p}_{\mu^*}(Y)}{p_0(Y)} \geq e^D\right) \leq e^{-D}. \tag{5.2.8}$$

(5.2.7) is the bound first developed by [26], who presented it as an extension of part of Sanov's theorem. In the one-dimensional case, it can also be seen as the 'generic' Chernoff bound. This bound is usually formulated as

$$\text{“}P_0(Y \geq \mu^*) \leq \inf_{\theta>0} \mathbb{E}_{P_0}[e^{\theta^\top Y}]e^{-\theta^\top \mu^*}.\text{”}$$

Evaluating the infimum shows that the right-hand side is equal to $\exp(-\underline{D})$ with $\underline{D} = D(\bar{P}_{\mu^*}\|P_0) = \inf_{\mu\geq\mu^*} D(\bar{P}_\mu\|P_0)$, making the bounds equivalent; hence our choice of the *CSC*-terminology (this is the *generic* Chernoff bound; when authors speak of *the* Chernoff bound they usually refer to a specific instance of it, with $Y = \sum_{i=1}^{n} X_i$ and X_i binary). Links between Sanov's theorem, Csiszár's extension thereof and Chernoff have been explored before; see e.g. Van Erven [89].

We also note that, while versions of (5.2.8) have been known for a long time, it is sometimes considered surprising, because a direct, more naive application of Markov's inequality would give, with $S' = \frac{\sup_{\mu \in \mathtt{M}} \bar{p}_{\mu}(Y)}{p_0(Y)}$,

$$P_0(S' \geq e^D) \leq e^{-D} \cdot \mathbb{E}_{P_0}\left[S'\right] = e^{-D} \cdot \int p_0(y) \cdot \frac{\sup_{\mu \in \mathtt{M}} \bar{p}_{\mu}(y)}{p_0(y)} d\rho(y) \gg e^{-D},$$

which can be considerably weaker, since S' is not an e-variable. The result (5.2.8) shows that we *can* establish an underlying e-variable, and it is given by $\bar{p}_{\mu^*}/p_0$.

Even though Theorem 10 is not new, we give its proof in full since its ingredients will be reused later on:

Proof. **[of Theorem 10]** Let $\mathtt{M}_1$ and $\mathcal{H}_1$ be as in the theorem, so that ALT conditions hold. Note that we may choose $\mathcal{H}_1$ to be convex. By the final part of Proposition 14 we then know that $S_{\text{GROW}} = \frac{\bar{p}_{\mu^*}(Y)}{p_0(Y)}$ and that for all $P \in \mathcal{H}_1$, with $\mu = \mathbb{E}_P[Y]$ and $\underline{D} = D(\bar{P}_{\mu^*} \| P_0)$, we have

$$\mu \in \mathtt{M}_1 \Rightarrow \mathbb{E}_P\left[\log \frac{\bar{p}_{\mu^*}(Y)}{p_0(Y)}\right] \geq \underline{D}, \tag{5.2.9}$$

where the expectation is well-defined. Now note that the right-hand side can be rewritten, with $\theta^* = \theta(\mu^*)$, as

$$\theta^{*\top}\mu - \log Z(\theta^*) \geq \underline{D}.$$

Thus (5.2.9) can be rewritten as:

$$y \in \mathtt{M}_1 \Rightarrow \theta^{*\top} y - \log Z(\theta^*) \geq \underline{D},$$

or, again equivalently,

$$y \in \mathtt{M}_1 \Rightarrow \log \bar{p}_{\mu^*}(Y)/\bar{p}_0(Y) \geq \underline{D}.$$

The result (5.2.6) follows after exponentiating. The subsequent inequality (5.2.7) now readily follows as well.

For (5.2.8), we only consider the case $\mu^* > 0$, the case $\mu^* < 0$ being completely analogous. We have

$$\begin{aligned} &P_0\left(\log \sup_{\mu \in \mathtt{M}} \frac{\bar{p}_{\mu}(Y)}{p_0(Y)} \geq D, Y \geq \mu^*\right) = P_0\left(Y \geq \mu^*\right) = \\ &P_0\left(\frac{\bar{p}_{\mu^*}(Y)}{p_0(Y)} \geq e^D, Y \geq \mu^*\right) \leq P_0(S_{\text{GROW}} \geq D) \leq e^{-D}, \end{aligned} \tag{5.2.10}$$

which follows because, by the *robustness property of exponential families* [35, Chapter 19](but also easily verified directly by considering the natural parameterization), $\log p_{\mu^*}(y)/p_0(y) = D(\bar{P}_{\mu^*}\|P_0)$ if $y = \mu^*$, and $\log p_{\mu^*}(y)/p_0(y)$ is increasing in y for $\mu^* > 0$, which implies that the events inside the left and the right probability are identical. On the other hand,

$$P_0\left(\log \sup_{\mu\in\mathtt{M}} \frac{\bar{p}_\mu(Y)}{p_0(Y)} \geq D, 0 \leq Y < \mu^*)\right) = P_0\left(\log \frac{\bar{p}_Y(Y)}{p_0(Y)} \geq D, 0 \leq Y < \mu^*)\right) = \\ P_0(D(\bar{P}_Y\|P_0) \geq D(\bar{P}_{\mu^*}\|P_0), 0 \leq Y < \mu^*) = 0, \tag{5.2.11}$$

where the first equality follows because the μ maximizing the likelihood $\bar{p}_\mu(Y)$ is uniquely given by Y if $Y \in \mathtt{M}$, and the second is again the robustness property of exponential families, and the third follows because KL divergence $D(\bar{P}_\mu\|P_0)$ is strictly increasing in μ if $\mu > 0$.

Together, (5.2.10) and (5.2.11) imply the result. □

5.3 Surrounding $\mathtt{M}_1$

We now consider tests and concentration bounds in the often more relevant setting of 'surrounding' $\mathtt{M}_1$. Formally, we call $\mathtt{M}_1$ *surrounding* if its complement, $\mathtt{M}_1^{\mathrm{C}} := \textsc{conv}(\mathcal{Y})\backslash\mathtt{M}_1$ is an open, bounded, connected set containing 0, and contained in $\mathbb{R}^d$. We will call surrounding $\mathtt{M}_1$ *nice* if (a) $\mathtt{M}_1^{\mathrm{C}}$ is contained in the interior of the mean-value space $\mathtt{M}$ of exponential family $\mathcal{E}$ generated by P_0 and also (b) $\mathtt{M}_1^{\mathrm{C}}$ is 0-*star-shaped*, which means that for any straight line going through 0, its intersection with $\mathtt{M}_1^{\mathrm{C}}$ is an interval, so that it crosses the boundary $\mathrm{BD}(\mathtt{M}_1^{\mathrm{C}})$ only once. Note in particular that any convex $\mathtt{M}_1^{\mathrm{C}}$ is automatically star-shaped; see Figure 5.2 and 5.3 for two examples of star-shaped $\mathtt{M}_1^{\mathrm{C}}$. Note also that any *nice* $\mathtt{M}_1$ automatically satisfies Condition ALT-$\mathtt{M}_1$.

We now develop optimal e-variables for surrounding and regular $\mathtt{M}_1$. The GROW criterion is still meaningful in this setting, and we discuss it in Section 5.3.1 below. Yet alternative, *relative* growth criteria are sometimes more meaningful in hypothesis testing [71, 42] and one of these, minimax regret, more directly leads to corresponding CSC-type bounds. We consider these in Section 5.3.2 and 5.3.3.

5.3.1 GROW for $d = 1$

We may again consider the GROW criterion for general $\mathcal{H}_1$ that could be any set of distributions compatible with a given nice surrounding $\mathtt{M}_1$ and $d \geq 1$, but this turns out to be surprisingly complicated in general. We only managed to find a simple characterization of S_{GROW} for the case $d = 1$, $\mathtt{M}_1 \subset \mathtt{M}$ and $\mathcal{H}_1 = \mathcal{E}_1 = \{\bar{P}_\mu : \mu \in \mathtt{M}_1\}$; that is, we are now testing P_0, a member of 1-dimensional exponential family $\mathcal{E}$, against a subset $\mathcal{E}_1$ of $\mathcal{E}$ that is bounded away from P_0.

In the sequel we denote by $\bar{p}_W(X^n) := \int \bar{p}_\mu(X^n)dW(\mu)$ the Bayes marginal density corresponding to prior measure W.

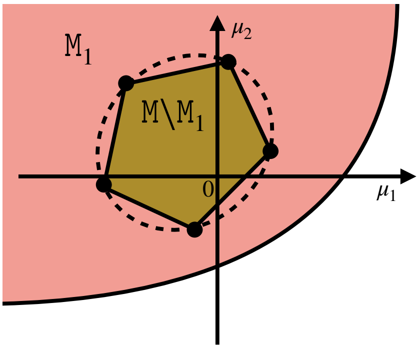


Figure 5.2: Surrounding, nice, $\mathtt{M}_1$ with a finite nice partition into convex sets. This figure is obtained by taking P_0 a Gamma distribution on X and defining $Y = (Y_1, Y_2) = (\log X, X - c)$ for a constant $c > 1$. Then $\mathcal{E}$ is a translated Gamma family with sufficient statistic Y and mean-value space $\mathtt{M} = \{(y_1, y_2) : y_1 \in \mathbb{R}, y_2 = e^{y_1} - c\}$ (unlike in Figure 5.1, we have $\mathtt{M}_1 \subset \mathtt{M}$ here).

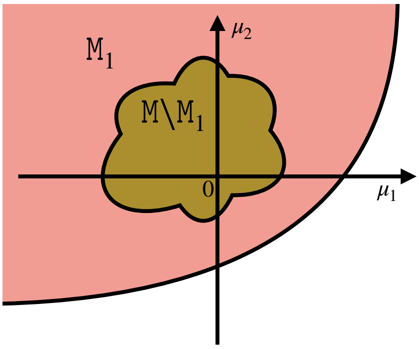


Figure 5.3: Surrounding, nice $\mathtt{M}_1$ that cannot be partitioned into a finite number of convex sets (again we show the translated Gamma family).

Theorem 11. *Let P_0 be a distribution for 1-dimensional $Y \subset \mathbb{R}$, and suppose that $\mathtt{M}_1$ is nice, i.e. $\mathtt{M}_1^{\mathrm{C}} = (\mu_1^-, \mu_1^+)$ is an open interval containing 0 and contained in the mean-value parameter space $\mathtt{M}$ for the 1-dimensional exponential family generated by P_0. Then, among all distributions W on the boundary $\mathrm{BD}(\mathtt{M}_1^{\mathrm{C}}) = \{\mu_1^-, \mu_1^+\}$, the minimum $D(\bar{P}_W || P_0)$ is achieved by a distribution W^* that satisfies*

$$D(\bar{P}_{W^*} || \bar{P}_{\mu_0}) = \mathbb{E}_{\bar{P}_{\mu_1^-}} \left[\log \frac{\bar{p}_{W^*}(Y)}{p_0(Y)} \right] = \mathbb{E}_{\bar{P}_{\mu_1^+}} \left[\log \frac{\bar{p}_{W^*}(Y)}{p_0(Y)} \right]. \tag{5.3.1}$$

The GROW e-variable relative to $\mathcal{E}_1 = \{\bar{P}_\mu : \mu \in \mathtt{M}_1 \cap \mathtt{M}\}$, denoted S_{GROW}, and the GROW e-variable relative to $\mathcal{E}_1^{\mathrm{BD}} := \{\bar{P}_\mu : \mu \in \mathrm{BD}(\mathtt{M}_1^{\mathrm{C}})\}$, denoted $S_{\mathrm{GROW}}^{\mathrm{BD}}$, are both *given by:*

$$S_{\mathrm{GROW}} = S_{\mathrm{GROW}}^{\mathrm{BD}} = \frac{\bar{p}_{W^*}(Y)}{p_0(Y)}, \tag{5.3.2}$$

i.e., the support of the prior W^* on $\mathtt{M}_1$ minimizing the KL divergence $D(\bar{P}_{W^*} \| P_0)$ is fully concentrated on the boundary of $\mathtt{M}_1^{\mathrm{C}}$.

Proof. **(of Theorem 11),** Define, for $\mu \in \mathtt{M}$ and $w^* \in [0, 1]$,

$$f(\mu, w^*) = \mathbb{E}_{\bar{P}_\mu} \left[\log \frac{(1 - w^*)\bar{p}_{\mu_1^-}(Y) + w^* \bar{p}_{\mu_1^+}(Y)}{p_{\mu_0}(Y)} \right] \tag{5.3.3}$$

and note that, for each μ, it holds that $f(\mu, w^*)$ is continuous in w^*. Now consider

$f(\mu_1^-, w^*) = -D(\bar{P}_{\mu_1^-} \| \bar{P}_{W^*}) + D(\bar{P}_{\mu_1^-} \| P_0)$. Minus the first term, $D(\bar{P}_{\mu_1^-} \| \bar{P}_{W^*})$, is 0 at $w^* = 0$ and continuously monotone increasing in w^*, since KL divergence is nonnegative and strictly convex in its second argument. Therefore $f(\mu_1^-, w^*)$ is itself continuously monotone decreasing in w^* with $f(\mu_1^-, 0) = D(\bar{P}_{\mu_1^-} \| P_0) > 0$. Also, $f(\mu_1^-, 1) = -D(\bar{P}_{\mu_1^-} \| \bar{P}_{\mu_1^+}) + D(\bar{P}_{\mu_1^-} \| P_0) < 0$, since KL divergence $D(\bar{P}_{\mu_1^-} \| \bar{P}_{\mu'})$ is strictly increasing in μ', for $\mu' > \mu_1^-$.

Analogously $f(\mu_1^+, w^*)$ is continuously monotone increasing in w^*, $f(\mu_1^+, 1) = D(\bar{P}_{\mu_1^+} \| P_0) > 0$ and $f(\mu_1^+, 0) < 0$.

This shows that there exists $0 < w^\circ < 1$ such that $f(\mu_1^-, w^\circ) = f(\mu_1^+, w^\circ)$. This implies that there exists a W^* (with $W^*(\{\mu_1^+\}) = w^\circ$) such that the rightmost equality in (5.3.1) holds. But this rightmost equality implies that for all $w' \in [0,1]$,

$$(1 - w')\mathbb{E}_{\bar{P}_{\mu_1^-}} \left[\log \frac{\bar{p}_{W^*}(Y)}{p_0(Y)} \right] + w'\mathbb{E}_{\bar{P}_{\mu_1^+}} \left[\log \frac{\bar{p}_{W^*}(Y)}{p_0(Y)} \right] = \mathbb{E}_{\bar{P}_{\mu_1^-}} \left[\log \frac{\bar{p}_{W^*}(Y)}{p_0(Y)} \right].$$

Plugging in $w' = w^\circ$, we get the left equality in (5.3.1). Now (5.3.1) in turn gives that

$$\min_{\mu \in \mathrm{BD}(\mathtt{M}_1^c)} \mathbb{E}_{\bar{P}_\mu} \left[\log \frac{\bar{p}_{W^*}(Y)}{p_0(Y)} \right] = D(\bar{P}_{W^*} || \bar{P}_{\mu_0}), \tag{5.3.4}$$

whereas for any probability density q for $\mathcal{Y}$,

$$\min_{\mu \in \mathrm{BD}(\mathtt{M}_1^c)} \mathbb{E}_{\bar{P}_\mu} \left[\log \frac{q(Y)}{p_0(Y)} \right] \leq \mathbb{E}_{\bar{P}_{W^*}} \left[\log \frac{q(Y)}{p_0(Y)} \right] \leq \mathbb{E}_{\bar{P}_{W^*}} \left[\log \frac{\bar{p}_{W^*}(Y)}{p_0(Y)} \right] = D(\bar{P}_{W^*} || \bar{P}_{\mu_0}), \tag{5.3.5}$$

which together with (5.3.4) shows that $S_{\mathrm{GROW}}^{\mathrm{BD}} = \frac{\bar{p}_{W^*}(Y)}{p_0(Y)}$, since every well-defined e-variable can be written as $q(Y)/p_0(Y)$ for some probability density q. Also, similarly to (5.3.5), we have

$$\inf_{\mu \in \mathtt{M}_1} \mathbb{E}_{\bar{P}_\mu} \left[\log \frac{q(Y)}{p_0(Y)} \right] \leq \mathbb{E}_{\bar{P}_{W^*}} \left[\log \frac{q(Y)}{p_0(Y)} \right] \leq D(\bar{P}_{W^*} || \bar{P}_{\mu_0}),$$

so if we could show

$$\inf_{\mu \in \mathtt{M}_1} \mathbb{E}_{\bar{P}_\mu} \left[\log \frac{\bar{p}_{W^*}(Y)}{p_0(Y)} \right] = D(\bar{P}_{W^*} || \bar{P}_{\mu_0}), \tag{5.3.6}$$

then the above two statements would together also imply that $S_{\mathrm{GROW}} = \frac{\bar{p}_{W^*}(Y)}{p_0(Y)}$, and we would be done. But (5.3.6) follows by (5.3.1) together with the following lemma, which thus completes the proof:

Lemma 9. *$f(\mu, w^\circ)$ is increasing on $\{\mu \in \mathtt{M}_1 : \mu \geq \mu^+\}$ and decreasing on $\{\mu \in \mathtt{M}_1 : \mu \leq \mu^-\}$.*

□

Lemma 9 is proved in Section 5.5. It follows from the fact that exponential families represent *variation reducing kernels* [20], a notion in theoretical statistics that seems to have been largely forgotten, and that we recall in Section 5.5. In that section we also explain why this result, even for $d = 1$, is difficult to prove, which also explains why proving anything nonasymptotic for the case $d > 1$ is currently beyond our reach.

5.3.2 Alternative Optimality Criteria: minimax redundancy and regret

S_{GROW} is difficult to characterize when $d > 1$ and $\mathcal{H}_1$ is surrounding; while it seems intuitive that, at least under some additional regularity conditions, it is still given by a likelihood ratio with a Bayes mixture concentrated on the boundary of $\mathtt{M}_1$, we did not manage to prove this (we indicate what the difficulty is towards the end of Section 5.5); we can say more though for the special case that $\mathtt{M}_1^{\mathrm{C}}$ is a KL ball and $Y = n^{-1} \sum_{i=1}^{n} X_i$ as $n \to \infty$; see Section 5.4.

However, in e-variable practice we often deal with $\mathcal{H}_1$ that can be partitioned into a family of subsets $\{\mathcal{H}_{1,r} : r \in \mathcal{R}\}$ such that, ideally, we would like to use the GROW e-variable relative to the $\mathcal{R}$ that actually contains the alternative. This was called the *relative GROW* criterion by [42] and it was used by, for example, [88]. We thus face a collection of e-variables $\mathcal{S} = \{S_{r,\text{GROW}} : r \in \mathcal{R}\}$ where each $S_{r,\text{GROW}}$ is GROW for $\mathcal{H}_{1,r}$. If an oracle told us beforehand "if the data come from $P \in \mathcal{H}_1$, then in fact $P \in \mathcal{H}_{1,r}$" then we would want to use $S_{r,\text{GROW}}$. Not having access to such an oracle, we want to use an e-variable that loses the least e-power compared to $S_{r,\text{GROW}}$ for the 'true' $r := r(P)$ (i.e. such that $P \in \mathcal{H}_{1,r}$), in the worst-case over all $r \in \mathcal{R}$. This is akin to what in the information theory literature is called a *minimax redundancy* approach [23, 35]. This approach is still hard to analyze in general but is amenable to the asymptotic analysis we provide in the next section. In a minor variation of this idea, used in an e-value context before by [67, 49], we may consider the e-variable that loses the least e-power compared $S_{\breve{r},\text{GROW}}$ for the $\breve{r}$ that is optimal with hindsight for the data at hand, i.e. achieving $\max_{r \in \mathcal{R}} S_{r,\text{GROW}}$; thus $\breve{r}$ can be thought of as a maximum likelihood estimator. This is akin to what information theorists call *minimax individual-sequence regret* [35]. This final approach *can* be analyzed nonasymptotically and leads to an analogue of the CSC theorem. We now formalizeboth the redundancy and regret approaches.

Thus, suppose that $\mathtt{M}_1$ is surrounding and *nice* as defined in the beginning of Section 5.3, and let $\{\mathtt{M}_{1,r} : r \in \mathcal{R}\}$ be a partition of $\mathtt{M}_1$ (see Figure 5.2). We will restrict attention to *nice* partitions, i.e. partitions such that

$$\text{for all } r \in \mathcal{R},\ \mathtt{M}_{1,r} \text{ is a closed convex subset of } \mathtt{M}_1 \text{ with } \mathtt{M}_{1,r} \cap \mathrm{BD}(\mathtt{M}_1^{\mathrm{C}}) \neq \emptyset \tag{5.3.7}$$

Let $\mathcal{H}_{1,r} := \{P \in \mathcal{H}_1 : \mathbf{E}_{Y \sim P}[Y] \in \mathtt{M}_{1,r}\}$.

By strict convexity of $D(\bar{P}_\mu \| P_0)$ in μ and the nice-ness condition, we have that $\min_{\mu \in \mathtt{M}_{1,r} \cap \mathtt{M}} D(\bar{P}_\mu \| P_0)$ exists and is achieved uniquely for a point $\mu^*(r)$ on the boundary $\mathrm{BD}(\mathtt{M}_{1,r}) \cap \mathrm{BD}(\mathtt{M}_1^{\mathrm{C}})$. Every $r \in \mathcal{R}$ is mapped to a point $f(r) \in \mathrm{BD}(\mathtt{M}_1^{\mathrm{C}})$ in this way, and the mapping $f : \mathcal{R} \to \mathrm{BD}(\mathtt{M}_1^{\mathrm{C}})$ is injective since $\mathtt{M}_{1,r} \cap \mathtt{M}_{1,r'} = \emptyset$ for every $r \neq r'$

with $r, r' \in \mathcal{R}$. Therefore we will simply *identify* $\mathcal{R}$ with a subset of $\mathrm{BD}(\mathtt{M}_1^{\mathrm{C}})$, such that $f(\mu) = \mu$ for all $\mu \in \mathcal{R}$. For $P \in \mathcal{H}_1$ with $\mathbb{E}_P[Y] = \mu$ (so $\mu \in \mathtt{M}_1$), we will now define $r(P)$ to be the $r \in \mathcal{R}$ such that $P \in \mathcal{H}_{1,r}$, i.e. such that $\mu \in \mathtt{M}_{1,r}$. Note we can think of $r(P)$ either as an index of sub-hypothesis $\mathcal{H}_{1,r}$ or as a special boundary point of the space of mean-values $\mathtt{M}_{1,r}$.

If we were to test $\mathcal{H}_0$ vs. $\mathcal{H}_{1,r}$ for given r, then we would still like to use the GROW e-variable $S_{r,\mathrm{GROW}} = \bar{p}_r/p_0$. In reality we do not know r, but we aim for an e-value that loses as little evidence as possible compared to $S_{r,\mathrm{GROW}}$, in the worst-case over all r. Formally, we seek to find e-variable $S = q(Y)/p_0(Y)$, where q achieves

$$\sup_q \inf_{P \in \mathcal{H}_1} \mathbf{E}_{Y \sim P}\left[\log \frac{q(Y)}{p_0(Y)} - \log \frac{\bar{p}_{r(P)}(Y)}{p_0(Y)}\right] = \sup_q \inf_{P \in \mathcal{H}_1} \mathbf{E}_{Y \sim P}\left[\log \frac{q(Y)}{\bar{p}_{r(P)}(y)}\right]$$
$$= -\textsc{mmred}(\mathcal{H}_1) \text{ where } \textsc{mmred}(\mathcal{H}_1) := \inf_q \sup_{P \in \mathcal{H}_1} \left(D(P\|Q) - D(P\|\bar{P}_{r(P)})\right). \quad (5.3.8)$$

where the supremum is over all probability densities on Y and $r(P)$ is again the unique $r \in \mathcal{R} \subset \mathtt{M}_1$ such that $P \in \mathcal{H}_{1,r}$. $\textsc{mmred}(\mathcal{H})$ is easily shown to be nonnegative for any $\mathcal{H}$, and both equations in (5.3.8) are immediate. From the rightmost expression, information theorists will recognize the q as minimizing the maximum *redundancy* [24, 23, 82]: the worst-case additional mean number of bits needed to encode the data by an encoder who only knows that $P \in \mathcal{H}_1$ compared to an encoder with the additional knowledge that $P \in \mathcal{H}_{1,r}$.

As said, it is easier to analyze a slight variation of this approach which makes at least as much sense: rather than comparing ourselves to the inherently unknowable $r(P)$, we may consider the actually observed data $Y = y$ and compare ourselves to (i.e. try to obtain as much evidence against P_0 as possible compared to) the $\breve{r}(y)$ we would like to have used with hindsight, after seeing and in light of data y; and rather than optimizing an expectation under an imagined distribution which we will never fully identify anyway, we will optimize in the worst-case over all data. The setup works for general functions $\breve{r}$, indicating what $r \in \mathcal{R}$ we would have liked to use with hindsight; further below we discuss intuitive choices. We note that $\breve{r}$ maps data $\mathcal{Y}$ to a point in $\mathcal{R} \subset \mathrm{BD}(\mathtt{M}_1^{\mathrm{C}}) \subset \mathtt{M}_1$, so we can think of $\breve{r}$ as an *estimator* of parameter μ; however, the estimator is restricted to a small subset of $\mathtt{M}$, namely the set $\mathcal{R}$.

Thus, we now seek to find e-variable $S_{\mathrm{REL}} = q(Y)/p_0(Y)$, where q now achieves

$$\sup_q \inf_{y \in \mathcal{Y}} \left(\log \frac{q(y)}{p_0(y)} - \log \frac{\bar{p}_{\breve{r}(y))}(y)}{p_0(y)}\right) = \sup_q \inf_{y \in \mathcal{Y}} \left(\log \frac{q(y)}{\bar{p}_{\breve{r}(y)}(y)}\right)$$
$$= -\textsc{mmreg}(\breve{r}) \text{ where } \textsc{mmreg}(\breve{r}) := \inf_q \sup_{y \in \mathcal{Y}} \left(-\log \frac{q(y)}{\bar{p}_{\breve{r}(y)}(y)}\right), \quad (5.3.9)$$

where again the supremum is over all probability densities that can be defined on Y, the quantity $\textsc{mmreg}(\breve{r})$ is easily seen to be nonnegative regardless of how $\breve{r}$ is defined, and both equalities in (5.3.9) are immediate. From the rightmost expression, information theorists will recognize the optimizing q as the q minimizing *individual sequence regret*: it minimizes the 'regret' in terms of the number of bits needed to encode the data, in the worst-case over all sequences, compared to somebody who has seen $\breve{r}(y)$ in advance; the word 'regret' is also meaningful in our setting — the aim is to minimize regret in the sense of loosing as little evidence as possible compared to the largest attainable e-value (evidence) with hindsight. In information theory, neither the minimax redundancy nor the minimax individual sequence regret is considered inherently superior or more natural, and (as we shall also see in our context, in the next section), both quantities often behave similarly.

Indeed, (5.3.9) being a variation of a standard problem within information theory and sequential prediction with the logarithmic loss, it is well-known [14, 21, 35, 41] that the solution for q is uniquely achieved by the following variation of the *Shtarkov distribution*, a notion going back to [78]:

$$q_{\text{SHTARKOV},\breve{r}}(y) = \frac{\bar{p}_{\breve{r}(y)}(y)}{\int_y \bar{p}_{\breve{r}(y))}(y)\nu(dy)} \text{ so } S_{\text{REL}} = \frac{q_{\text{SHTARKOV}}(Y)}{p_0(Y)}. \tag{5.3.10}$$

We then get that $\left(\log \frac{q(y)}{\bar{p}_{\breve{r}(y)}(y)}\ \right) = \log \int_y \bar{p}_{\breve{r}(y)}(y)\rho(dy) := \text{MMREG}(\breve{r})$ independently of y, where $\text{MMREG}(\breve{r})$ is the *maximin regret* (called 'minimax regret' originally, since in data compression the rightmost expression, without the minus sign, is the relevant one).

The most straightforward choice is to take $\breve{r}(y) := \hat{r}(y)$ the maximum likelihood estimator within $\mathcal{R}$, achieving

$$\max_{r \in \mathcal{R}} \bar{p}_r(y) \tag{5.3.11}$$

for the given y, since then q_{SHTARKOV} has minimal overhead compared to the $S_{\text{GROW},r}$ that is largest with hindsight, i.e. that provides the most evidence with hindsight — thus providing additional justification for the terminology 'regret' in this special case, which was also Shtarkov's [78] original focus. If $\breve{r}$ is set to $\hat{r}$, then Q_{SHTARKOV} is also known as the *normalized maximum likelihood (NML)* distribution relative to the set $\mathcal{R} \subset \text{BD}(\mathtt{M}_1^{\text{C}})$ (not relative to the full exponential family $\mathcal{E}$!).

In the ensuing results we will mostly be interested in the case that $\mathcal{R}$ is either finite (as in Figure 5.2), or that it is in 1-to-1 correspondence with $\text{BD}(\mathtt{M}_1^{\text{C}})$ (as in Figure 5.3). In both cases, the maximum in (5.3.11) is achieved, although it may be achieved for more than one r; in that case, we set $\hat{r}$ to be the largest r achieving (5.3.11) in lexicographical ordering, making $\hat{r}$ well-defined in all cases.

While the upcoming analogue of the CSC theorem will mention the MLE $\hat{r}$, it turns out that in the proof, and in the detailed theorem statement, we also need to refer to an estimator $\breve{r}$ that may differ from the MLE. The reason is that, intriguingly, in general we may have that for some $r \in \mathcal{R}$ and $y \in \mathtt{M}_{1,r}$ we may have that $\hat{r}(y) \neq r$, which complicates the picture. To this end, we formulated the minimax regret approach for general $\breve{r} : \text{CONV}(\mathcal{Y}) \to \mathcal{R}$, and not just the MLE, as was done earlier e.g. by [41].

Example 12. [Gaussian Example] We use the simple 2-dimensional Gaussian case to gain intuition. Thus, we let $\mathcal{Y} = \text{CONV}(\mathcal{Y}) = \mathbb{R}^2$, and P_0 a 2-dimensional Gaussian distribution on Y, with mean 0 (i.e., $(0,0)$) and Σ a positive definite 2×2 covariance matrix. Then $\mathtt{M} = \mathbb{R}^2$ and $\mathcal{E}$ is a 2-dimensional Gaussian location family. For now, take

Σ to be the identity matrix. Then $D(P_0\|P_\mu) = (1/2)\|\mu\|_2^2$ is simply the squared norm of μ, facilitating the reasoning. A simple case of a convex $\mathtt{M}_1$ is the translated half-space $\mathtt{M}_1 = \{(y_1, y_2) : y_1 \geq a\}$ for constant $a > 0$. The point $\mu \in \mathtt{M}_1$ minimizing KL divergence to P_0 must then clearly be $(a, 0)$. Therefore, if $\mathcal{H}_1$ is any set of distributions with means in $\mathtt{M}_1$ and containing $\mathcal{E}_1^{\text{BD}} = \{\bar{P}_\mu : \mu \in \text{BD}(\mathtt{M}_1^{\text{c}})\}$, we have by Proposition 14 that $S_{\text{GROW}} = \bar{p}_{(a,0)}(Y)/\bar{p}_{(0,0)}(Y)$. We see that even if we have convex $\mathcal{H}_1$, the minimax individual sequence regret approach leads to a different e-variable, if we carve up $\mathcal{H}_1$ into $\{\mathcal{H}_{1,r} : r \in \mathcal{R}\}$ for $\mathcal{R}$ with more than one element. For example, we can take $\mathcal{R} = \{(a, \mu_2) : \mu_2 \in \mathbb{R}\}$ be a vertical line and let $\mathtt{M}_{1,(a,\mu_2)}$ be the subset of $\mathtt{M}_1$ on the line connecting $(0, 0)$ with (a, μ_2). Then, with $\hat{r}$ the MLE as in (5.3.11), we get that $\hat{r}(y)$ is the point where the line $\mathcal{R}$ intersects with the line connecting $(0, 0)$ and y. Since now $\hat{r}(y)$, and hence the sub-hypothesis $\mathcal{H}_{1,\hat{r}(y)}$ we want to be almost GROW against, changes with y, we get a solution S_{REL} in (5.3.10) that differs from S_{GROW}.

In the case of convex $\mathcal{H}_1$, whether to use the absolute or relative GROW e-variable may depend on the situation (see [42] for a motivation of when absolute or relative is more appropriate). In the case of nonconvex $\mathcal{H}_1$ with $d > 1$, we simply do not know how to characterize the absolute GROW and we have to resort to the relative GROW.

5.3.3 CSC Theorem for surrounding $\mathcal{H}_1$

For given estimator $\breve{r}$ and probability density q on Y, define $\text{REG}(q, \breve{r})$ ('regret') as

$$\text{REG}(q, \breve{r}, y) := \log\left(\frac{\bar{p}_{\breve{r}(y)}(y)}{q(y)}\right) \;;\; \text{MREG}(q, \breve{r}) := \sup_{y : y \in \mathtt{M}_1} \log\left(\frac{\bar{p}_{\breve{r}(y)}(y)}{q(y)}\right). \qquad (5.3.12)$$

Whereas above we discussed an MLE $\breve{r}$, i.e. $\breve{r} = \hat{r}$, we now require it to be self-consistent, i.e. we set it to be any function of y such that for all $y \in \mathtt{M}_1$, we have $y \in \mathtt{M}_{1,\breve{r}(y)}$. The value of $r(y)$ for $y \in \text{CONV}(\mathcal{Y}) \setminus \mathtt{M}_1$ will not affect the result below.

Theorem 12. **[CSC Theorem for surrounding $\mathcal{H}_1$]** *Suppose that $\mathtt{M}_1$ is nice and let $\{\mathtt{M}_{1,r} : r \in \mathcal{R}\}$ be any nice partition of $\mathtt{M}_1$ as in (5.3.7), with $\breve{r}$ any self-consistent estimator as above. Let q be an arbitrary probability density function. Then:*

$$\mathbb{E}_{P_0}\left[\mathbf{1}_{Y \in \mathtt{M}_1} \cdot e^{D(\bar{P}_{\breve{r}(Y))}\|P_0) - \text{REG}(q, \breve{r}, Y)}\right] \leq 1.$$

so that in particular, with $\underline{D} := \inf_{\mu \in \text{BD}(\mathtt{M}_1^{\text{c}})} D(P_\mu\|P_0)$,

$$P_0\left(Y \in \mathtt{M}_1\right) \leq e^{\text{MREG}(q,\breve{r}) - \underline{D}} \overset{(*)}{=} e^{\text{MMREG}(\breve{r}) - \underline{D}} \overset{(**)}{\leq} e^{\text{MMREG}(\hat{r}) - \underline{D}}, \qquad (5.3.13)$$

where $(*)$ holds if we take $q = q_{\text{SHTARKOV},\breve{r}}$, and $(**)$ holds if the MLE estimator $\hat{r}$ is well-defined.

Proof. Folllowing precisely analogous steps as in the proof of (5.2.6) as based on (5.2.9) within the proof of Theorem 10, we obtain, for all $y \in \textsc{conv}(\mathcal{Y})$,

$$y \in \mathtt{M}_1 \Rightarrow \frac{\bar{p}_{\breve{r}(y)}(y)}{p_0(y)} \geq e^{D(\bar{P}_{\breve{r}(y)} \| P_0)}. \tag{5.3.14}$$

Then (5.3.14) gives, using definition (5.3.12):

$$\mathbb{E}_{P_0}\left[\mathbf{1}_{Y \in \mathtt{M}_1} \cdot e^{D(\bar{P}_{\breve{r}(Y)} \| P_0) - \text{REG}(q, \breve{r}, Y)}\right] \leq \mathbb{E}_{P_0}\left[\mathbf{1}_{Y \in \mathtt{M}_1} \cdot \frac{\bar{p}_{\breve{r}(Y)}(Y)}{p_0(Y)} \cdot e^{-\text{REG}(q, \breve{r}, Y)}\right] \leq$$
$$\mathbb{E}_{P_0}\left[\mathbf{1}_{Y \in \mathtt{M}_1} \cdot \frac{q(Y)}{p_0(Y)} \cdot e^{\text{REG}(q, \breve{r}, Y)} \cdot e^{-\text{REG}(q, \breve{r}, Y)}\right] \leq \mathbb{E}_{P_0}\left[\frac{q(Y)}{p_0(Y)}\right] = 1,$$

which proves the first statement in the theorem. The second statement is then immediate. □

Analyzing the CSC Result — Different Partitions $\{\mathtt{M}_{1,r} : r \in \mathcal{R}\}$ The next question is how to cleverly partition any given nonconvex $\mathtt{M}_1$ so as to get a good bound when applying Theorem 12. We first note that, for any given $\mathtt{M}_1$, the final bound (5.3.13) does not worsen if we enlarge $\mathtt{M}_1$, as long as $\underline{D} = \inf_{\mu \in \text{BD}(\mathtt{M}_1^{\text{C}})} D(P_\mu \| P_0)$ stays the same. Thus, we still get the same bound if we shrink the complement $\mathtt{M}_1^{\text{C}}$ to the $\underline{D}$-KL ball $\{\mu : D(\bar{P}_\mu \| P_0) < \underline{D}\}$, without making the bound looser. We will therefore, from now on, simply assume that we are in the situation in which $\mathtt{M}_1^{\text{C}}$ is the $\underline{D}$-KL ball. We know that such a KL ball is convex [19] — with such a convex $\mathtt{M}_1^{\text{C}}$ we are thus in the 'dual case', as it were, to the case of convex $\mathtt{M}_1$ which we discussed in Section 5.2.

There are now basically two approaches to apply the CSC Theorem 12 that suggest themselves. In the first approach, we first determine a larger $\mathtt{M}'_1$ (hence smaller $\mathtt{M}_1^{\prime\text{C}}$) that contains $\mathtt{M}_1$, such that $\mathtt{M}'_1$ can be partitioned into a finite number $|\mathcal{R}'|$ of convex subsets, $\{\mathtt{M}'_{1,r} : r \in \mathcal{R}\}$, and then we apply Theorem 12 to $\mathtt{M}'_1$ and $\mathcal{R}'$. We could, for example, take $\mathtt{M}_1^{\prime\text{C}}$ be a convex polytope with a finite number of corners, all touching $\text{BD}(\mathtt{M}_1^{\prime\text{C}})$. Such a situation is depicted in Figure 5.2, if we interpret the dashed curve the boundary of a KL ball and the $\mathtt{M}_1^{\text{C}} = \mathtt{M} \setminus \mathtt{M}_1$ in the Figure as the polytope $\mathtt{M}_1^{\prime\text{C}}$. In the second approach, we set $\mathcal{R} = \text{BD}(\mathtt{M}_1^{\text{C}})$, making it a manifold in $\mathbb{R}^d$, and set

$$\mathtt{M}_{1,r} := \{\mu \in \mathtt{M}_1 : \alpha\mu = r \text{ for some } \alpha > 0\},$$

i.e. the set of points in $\mathtt{M}_1$ on the ray starting at 0 and going through r. Then we have for all $y \in \text{BD}(\mathtt{M}_1^{\text{C}})$ that $r(y) = y$. We may think of this second approach as a limiting case of the first one, when we let the number of corners of the polytope go to infinity. In the next section we show that, if we apply the CSC Theorem 12, in our main case of interest, with $Y = n^{-1} \sum_{i=1}^n X_i$, and $\mathtt{M}_1^{\text{C}}$ a KL Ball, then for large n, the second, 'continuous' approach always leads to better bounds than the first.

5.4 Asymptotic expression of growth rate and regret

While the exact sizes of $\text{MMRED}(\mathcal{H}_1)$ and $\text{MMREG}(\breve{r})$ are hard to determine, for the case of nice $\mathtt{M}_1^{\mathrm{C}}$ and our central case of interest, with $Y = n^{-1} \sum X_i$, we can use existing results to obtain relatively sharp asymptotic (in n) approximations of $\text{MMRED}(\mathcal{H}_1)$ and $\text{MMREG}(\breve{r})$'s upper bound $\text{MMREG}(\hat{r})$. We now derive these approximations and show how they, in turn, lead to an approximation to GROW if moreover $\mathtt{M}_1^{\mathrm{C}}$ is a KL ball, which as explained above, is also the case of central interest.

Thus, we now assume that $Y := n^{-1} \sum X_i$, with $X, X_1, X_2, \ldots$ i.i.d. $\sim P'_0$, where P'_0 is a distribution on X, inducing distribution $P_0 \equiv P_0^{[Y]}$ for Y. Like P_0, we have that P'_0 also generates an exponential family, now with sufficient statistic X and densities

$$p'_\theta(x) \propto e^{\theta^\top x} p'_0(x) \tag{5.4.1}$$

extended to n outcomes by assuming independence. Now $\mathbb{E}_{P'_\theta}[Y] = n^{-1}\mathbb{E}_{P'_\theta}[\sum_{i=1}^n X_i] = \mathbb{E}_{P'_\theta}[X]$ so that the mapping $\mu(\theta)$ from canonical to mean-value parameter is the same for both the family (5.4.1) and the original family $\{P_\theta(Y) : \theta \in \Theta\}$, and the likelihood ratio of any member P'_θ of the family (5.4.1) to P'_0 on n outcomes is given by, with $\mu = \mu(\theta)$,

$$\prod_{i=1}^n \frac{\bar{p}'_\mu(X_i)}{p'_0(X_i)} = \frac{p'_\theta(X^n)}{p'_0(X^n)} = \frac{p_\theta(Y)}{p_0(Y)} = \frac{\bar{p}_\mu(Y)}{p_0(Y)},$$

which in turn implies that $D(\bar{P}_\mu \| P_0) = nD(\bar{P}'_\mu \| P'_0)$, where $D(\bar{P}'_\mu \| P'_0)$ is an expression that does not change with n. This means that if we keep $\mathtt{M}_1$ constant, the $\underline{D} = \inf_{\mu \in \mathtt{M}_1} D(\bar{P}_\mu \| P_0)$ in (5.3.13) increases linearly in n. To avoid confusion here, it is useful to make explicit the dependency of MMREG and $\underline{D}$ on n in Theorem 12, by writing it as MMREG_n and $\underline{D}_n$: we can then restate the bound (5.3.13) in the theorem as

$$P_0(Y \in \mathtt{M}_1) \leq e^{\text{MMREG}_n(\hat{r}) - \underline{D}_n} = e^{\text{MMREG}_n(\hat{r}) - n\underline{D}_1}. \tag{5.4.2}$$

Thus, as n increases, if we keep $\mathtt{M}_1$ fixed, then the quantity $\underline{D}_n$ in the bound increases linearly in n. On the other hand, we will now show that, for sufficiently smooth boundaries of $\mathtt{M}_1$, we have that $\text{MMREG}_n(\hat{r})$ only increases logarithmically in n, making the strength of bound (5.3.13) still grow exponentially in n. The result is a direct corollary of a result by [82].

Theorem 13. [Corollary of Theorem 2 of [82]; see also [80]] *Consider the setting of surrounding* $\mathtt{M}_1$ *and suppose that* $\mathtt{M}_1$ *is* nice, *as defined in the beginning of Section 5.3, and that there exists a bijective function* $\phi : \mathcal{U} \to \text{BD}(\mathtt{M}_1^{\mathrm{C}})$ *so that* $\mathcal{U}$ *is a subset of* $\mathbb{R}^{d-1}$ *with open interior,* ϕ *has at least four derivatives and these are bounded on* $\mathcal{U}$*. Then* $\hat{r}$ *is well-defined and there exists* $C > 0$ *such that for all* n*,*

$$\text{MMREG}_n(\hat{r}) \leq \frac{d-1}{2} \log n + C.$$

We also have a bound on the minimax regret for the case that $\mathcal{H}_1$ contains $\mathcal{E}_1^{\mathrm{BD}}$, the subset of exponential family $\mathcal{E}$ restricted to the boundary $\mathrm{BD}(\mathtt{M}_1^{\mathrm{C}})$:

Theorem 14. **[Corollary of Theorem 1 of [23]]** *Consider the setting and conditions of the previous theorem. Let* $\mathcal{E}_1^{\mathrm{BD}} := \{\bar{P}_\mu : \mu \in \mathrm{BD}(\mathtt{M}_1^{\mathrm{C}}) \cap \mathtt{M}\}$. *Then there exists* $C' < 0$ *such that for all* n*,*

$$\mathrm{MMRED}_n(\mathcal{E}_1^{\mathrm{BD}}) = \inf_q \sup_{\mu \in \mathrm{BD}(\mathtt{M}_1^{\mathrm{C}})} D(\bar{P}_\mu \| Q) \geq \frac{d-1}{2} \log n + C'.$$

> To see how this result follows from Clarke and Barron's, note that for this we need to verify the regularity Conditions 1–3 in Section 2 of their paper. the assumption of niceness implies that $\mathrm{BD}(\mathtt{M}_1^{\mathrm{C}})$ is contained in a compact subset $\mathtt{M}'$ of the interior of $\mathtt{M}$. Then also the corresponding natural parameters are contained in a compact subset Θ' of the interior of Θ. Condition 1 of their paper is immediately verified for parameters in the natural parameterization Θ. Since the functions $\theta : \mathtt{M} \to \Theta$ and $\phi : \mathcal{U} \to \mathtt{M}$ and their first and second derivatives are themselves bounded on $\mathcal{U}$ and $\mathtt{M}'$, Condition 1 of their paper is verified in terms of the relevant parameterization $\bar{P}_{\phi(u)}$ as well. Moreover, for all $k \in \mathbb{N}$, all partial derivatives of form $\partial^k / (\partial u_{j_1} \dots \partial u_{j_k}) \int \bar{p}_{\phi(u)}(y) d\rho(y)$, with $j_1, \dots, j_k \in \{1, \dots, d-1\}$, can be calculated by exchanging differentiation and integration (this follows from [19, Theorem 2.2]. Since we already established [23]'s Condition 1, this implies that their Condition 2 also holds, as they explain underneath their Condition 2. Their Condition 3 is immediate.

Now, for any $\mathcal{H}_1$ that contains $\mathcal{E}_1^{\mathrm{BD}}$, Theorem 14 implies that

$$\mathrm{MMRED}_n(\mathcal{H}_1) \geq \mathrm{MMRED}_n(\mathcal{E}_1^{\mathrm{BD}}) \geq \frac{d-1}{2} \log n + C', \tag{5.4.3}$$

and it is also immediate, by definition of $\hat{r}$, that

$$\mathrm{MMREG}_n(\hat{r}) \geq \mathrm{MMRED}_n(\mathcal{H}_1). \tag{5.4.4}$$

Together with (5.4.3) and (5.4.4), Theorem 13 above now gives that, under the assumption of both theorems,

$$\mathrm{MMRED}_n(\mathcal{H}_1) = \mathrm{MMREG}_n(\hat{r}) + O(1) = \frac{d-1}{2} \log n + O(1). \tag{5.4.5}$$

How to Partition $\mathtt{M}_1$, Continued We now restrict to the case that $\mathtt{M}_1^{\mathrm{C}}$ is a $\underline{D}_1$-KL ball. At the end of previous section we explained why this is the major case of interest. As above, we want to keep $\mathtt{M}_1^{\mathrm{C}}$ fixed as n increases, i.e. the set of parameters that stays in $\mathtt{M}_1$ does not change with n. This means that, when viewed as a KL ball of distributions on X, the radius of the ball remains constant with n, but when viewed as a KL ball of distributions on Y, the radius of the ball does need to scale linearly with

n, i.e. we set:

$$\mathtt{M}_1^{\mathrm{C}} = \{\mu : D(\bar{P}'_\mu \| P'_0) < \underline{D}_1\} = \{\mu : D(\bar{P}_\mu \| P_0) < n\underline{D}_1\}. \tag{5.4.6}$$

We now return to the two approaches to applying the CSC Theorem 12 which we discussed at the end of the previous section: one based on a finite $\mathcal{R}$ defining a polytope, one based on a 'continuous' $\mathcal{R} = \mathrm{BD}(\mathtt{M}_1^{\mathrm{C}})$. It turns out that if $\mathtt{M}_1$ is defined in terms of a KL ball (5.4.6), then, for large n, it is always better to take the continuous $\mathcal{R}$ approach. To see this, suppose that, in the polytope approach, we take a polytope $\mathtt{M}_1^{'\mathrm{C}}$ with k corners (e.g., $k = 5$ in Figure 5.2); let $\check{r}_k$ be the corresponding estimator and $\hat{r}_k$ be the corresponding MLE in $\mathcal{R}$ and $\underline{D}'_{1,k} < \underline{D}_1$ (Figure 5.2) indicates why the inequality holds) be the minimum KL divergence we then obtain in (5.3.13) when replacing $\mathtt{M}_1^{\mathrm{C}}$ by $\mathtt{M}_1^{'\mathrm{C}}$, applied for $n = 1$; similarly we let $\check{r}_\infty$ be the corresponding estimator for the second, continuous approach and $\hat{r}_\infty$ be the corresponding MLE in $\mathcal{R} = \mathrm{BD}(\mathtt{M}_1^{\mathrm{C}})$ and $\underline{D}'_{1,\infty} = \underline{D}_1$ be the corresponding minimum KL divergence appearing in the bound. Then the rightmost bounds in (5.3.13) will, respectively, look like

$$e^{\mathrm{MMREG}_n(\check{r}_k) - n\underline{D}'_{1,k}} \leq e^{\mathrm{MMREG}_n(\hat{r}_k) - n\underline{D}'_{1,k}} \text{ vs. } e^{\mathrm{MMREG}_n(\check{r}_\infty) - n\underline{D}_1} \leq e^{\mathrm{MMREG}_n(\hat{r}_\infty) - n\underline{D}_1}.$$

Since $\mathrm{MMREG}_n(\check{r}_k)$ is always nonnegative, no matter the definition of $\check{r}$ and the value of k and n, whereas $\mathrm{MMREG}_n(\hat{r}_\infty)$ is logarithmic, and $\underline{D}'_{1,k} < \underline{D}_1$ for all finite k, the continuous approach based on $k = \infty$ provides bounds that are eventually exponentially better in n compared to the bound based on any finite k.

An Asymptotic GROW Result In the KL ball setting, we can also say something about the asymptotic size of the worst-case optimal growth rate:

Proposition 15. *In the KL ball setting above, let $\mathcal{H}_1$ be any set of distributions that contains $\mathcal{E}_1^{\mathrm{BD}}$ and with means contained in $\mathtt{M}_1$, i.e. $\{\mathbb{E}_P[Y] : P \in \mathcal{H}_1\} \subset \mathtt{M}_1$. Then the growth rate* GROW_n *at sample size n is given by*

$$\mathrm{GROW}_n = n\underline{D} - \frac{d-1}{2}\log n + O(1).$$

Proof. We have, with the supremum being taken over all probability densities q for Y,

$$\sup_q \inf_{P \in \mathcal{H}_1} \mathbb{E}_P\left[\log \frac{q(Y)}{p_0(Y)}\right] \leq \sup_q \inf_{\mu \in \mathrm{BD}(\mathtt{M}_1^{\mathrm{C}})} \mathbb{E}_{\bar{P}_\mu}\left[\log \frac{q(Y)}{p_0(Y)}\right] =$$
$$\sup_q \inf_{\mu \in \mathrm{BD}(\mathtt{M}_1^{\mathrm{C}})} \left(\mathbb{E}_{\bar{P}_\mu}\left[\log \frac{q(Y)}{p_0(Y)} - \log \frac{\bar{p}_\mu(Y)}{p_0(Y)}\right] + \mathbb{E}_{\bar{P}_\mu}\left[\log \frac{\bar{p}_\mu(Y)}{p_0(Y)}\right] \right) =$$
$$\sup_q \inf_{\mu \in \mathrm{BD}(\mathtt{M}_1^{\mathrm{C}})} \mathbb{E}_{\bar{P}_\mu}\left[\log \frac{q(Y)}{p_0(Y)} - \log \frac{\bar{p}_\mu(Y)}{p_0(Y)}\right] + n\underline{D}_1 = -\frac{d-1}{2}\log n + O(1) + n\underline{D}_1.$$

where the last equation follows from (5.4.5). On the other hand, with $q_{\mathrm{SHTARKOV},\hat{r}}$ defined as in (5.3.10), we also have:

$$\sup_q \inf_{P\in\mathcal{H}_1} \mathbb{E}_P \left[\log \frac{q(Y)}{p_0(Y)}\right] \geq$$

$$\inf_{P\in\mathcal{H}_1} \mathbb{E}_P \left[\log \frac{q_{\text{SHTARKOV},\hat{r}}(Y)}{p_0(Y)}\right] = \inf_{\mu\in\mathtt{M}_1} \inf_{P\in\mathcal{H}_1:\mathbb{E}_P[Y]=\mu} \mathbb{E}_P \left[\log \frac{\bar{p}_{\hat{r}(Y)}(Y)}{p_0(Y)}\right] - \text{MMREG}_n(\hat{r}) \geq$$

$$\inf_{\mu\in\mathtt{M}_1} \inf_{P\in\mathcal{H}_1:\mathbb{E}_P[Y]=\mu} \mathbb{E}_P \left[\log \frac{\bar{p}_\mu(Y)}{p_0(Y)}\right] - \text{MMREG}_n(\hat{r}) \overset{*}{=}$$

$$\inf_{\mu\in\mathtt{M}_1} \mathbb{E}_{\bar{P}_\mu} \left[\log \frac{\bar{p}_\mu(Y)}{p_0(Y)}\right] - \text{MMREG}_n(\hat{r}) = \inf_{\mu\in\mathtt{M}_1} D(\bar{P}_\mu \| P_0) - \text{MMREG}_n(\hat{r}) =$$

$$n\underline{D}_1 - \frac{d-1}{2}\log n + O(1),$$

where $(*)$ follows by rewriting the quantity inside the logarithm in terms of the mean-value parameterization and evaluating the expectation. Combining the two displays above, the result follows. □

5.5 Proof of Lemma 9 and further discussion of Theorem 11

To prove Lemma 9, we first provide some background on *variation diminishing transformations* [20].

Definition 8. Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be a finite subset of $\mathbb{R}$ with $x_1 < x_2 < \ldots < x_n$ and let $g : \mathcal{X} \to \mathbb{R}$ be a function, so that $(g(x_1), \ldots, g(x_n)) \in \mathbb{R}^n$. We let $S^-(g)$ denote the number of sign changes of sequence $g(x_1), \ldots, g(x_n)$ where we ignore zeros; if g is identically 0 then we set $S^-(g)$ to 0.

Example 13. If $g'(x_1, x_2, x_3) = (-1, 0, 1)$, $g''(x_1, x_2, x_3) = (-2, 1, 4)$ and $g'''(x_1, x_2, x_4) = (-2, 0, 0, 3)$, then $S^-(g') = S^-(g'') = S^-(g''') = 1$.

Definition 9. Now consider arbitrary $\mathcal{X} \subset \mathbb{R}$ and let $g : \mathcal{X} \to \mathbb{R}$. For finite $\mathcal{V} \subset \mathcal{X}$, say $\mathcal{V} = \{x_1, \ldots, x_n\}$ with $x_1 < \ldots < x_n$, we let $g_{\mathcal{V}} = \{g(x_1), \ldots, g(x_n)\}$. We let $S^-(g)$ be the supremum of $S^-(g_{\mathcal{V}})$ over all finite subsets $\mathcal{V}$ of $\mathcal{X}$.

Intuitively, $S^-(g)$ is the number of times that the function $g(x)$ changes sign as $x \in \mathcal{X} \subset \mathbb{R}$ increases.

Lemma 10. [**20**, **Example 3.1, Proposition 3.1**] *Let P_0 and Y be as above (5.1.14), where $Y = Y_1$ is 1-dimensional, so $\mathcal{Y} \subseteq \mathbb{R}$, and consider the 1-dimensional exponential family generated by P_0 as in (5.1.14). In the terminology of [20], this family is* $\text{SVR}_n(\mathbb{R}, \Theta)$ *and hence* $\text{SVR}_n(\mathcal{Y}, \Theta)$ *for all n. Rather than giving the precise definition of SVR ('strict variation reducing'), we just state the implication of this fact that we need: for any function $g : \mathcal{Y} \to \mathbb{R}$ with $\int |g| d\rho > 0$ and $\gamma : \Theta \to \mathbb{R}$ with $\gamma(\theta) := \int p_\theta(y) g(y) \rho(dy)$, we have: $S^-(\gamma) \leq S^-(g)$.*

In words, for any function g as above, the number of sign changes of $\mathbf{E}_{P_\theta}[g(Y)]$ as we vary θ is bounded by the number of sign changes of g itself on $y \in \mathcal{Y} \subset \mathbb{R}$. Since, in one-dimensional full exponential families, $\mu(\theta)$ is a continuous, strictly increasing function of θ, this also implies that, for any function g, the expectation $\gamma(\mu) := \mathbf{E}_{\bar{P}_\mu}[g(Y)]$ also satisfies $S^-(\gamma) \leq S^-(g)$.

Now for any constant $c \in \mathbb{R}$, any $w^\circ \in [0,1]$, we set $g_c(y) = c + \log \frac{(1-w^\circ)\bar{p}_{\mu_1^-}(y) + w^\circ \bar{p}_{\mu_1^+}(y)}{p_0(y)}$. A little calculation of the derivatives shows that $g_c(y)$ is strictly convex on $\textsc{conv}(\mathcal{Y})$ and not monotonic. Therefore, $g_c(y)$ has exactly one minimum point y and is strictly monotonic on both sides of y. Thus, $g_c(y)$ as a function of $y \in \textsc{conv}(\mathcal{Y})$ changes sign twice; $g_c(y)$'s domain being restricted to $\mathcal{Y}$ (which is not the same as $\textsc{conv}(\mathcal{Y})$ in the discrete case), it changes sign at most twice. Thus, for all $c \in \mathbb{R}$, we have $S^-(g_c) \leq 2$. Lemma 10 thus implies that $S^-(\gamma_c(\mu)) \leq 2$ with $\gamma_c(\mu) := f(\mu, w^\circ) + c$ where $f(\mu, w)$ is defined as in (5.3.3). Therefore, we know that $g_0(\mu) = f(\mu, w^\circ)$ as a function of μ can at most have one minimum point achieved at some μ^* and if it has such a minimum point, it must be strictly monotonic on both sides of μ^*.

Now $f(0, w^\circ) = \mathbf{E}_{P_0}[g_0(Y)] = -D(\bar{P}_0 \| P_{W_1^\circ}) < 0$; but by (5.3.1), which we already showed, $f(\mu_1^+, w^\circ) = f(\mu_0^-, w^\circ) > 0$. It follows that a μ^* as mentioned above must exist, and that it lies in between μ_0^- and μ_0^+; the result follows.

Why the case $d > 1$ is complicated We only managed to prove a general GROW result for surrounding $\mathcal{H}_1$ for $d = 1$. To give the reader an idea where the difficulties lie, we first discuss the case $d = 1$ a little more. One may wonder why even there, we had to resort, via Lemma 9, to the pretty sophisticated theory of variation diminishing transformations. It would seem much simpler to directly calculate the derivative $(d/d\mu)f(\mu, w^\circ)$ and show that, for appropriate choice of w°, the derivative is 0 at some μ^* within the interval (μ_1^-, μ_1^+), and negative to the left and positive to the right of μ^*; this would lead to the same conclusion as just stated. Yet the derivative is given by

$$\frac{d}{d\mu} f(\mu, w^\circ) = \sigma_\mu^2 \cdot \Big(\mathbf{E}_{\bar{P}_\mu}[Y \cdot g_0(Y)] - \mathbf{E}_{\bar{P}_\mu}[Y] \cdot \mathbf{E}_{\bar{P}_\mu}[g_0(Y)] \Big), \tag{5.5.1}$$

where $\sigma_\mu^2 = \mathbf{E}_{\bar{P}_\mu}[Y^2] - (\mathbf{E}_{\bar{P}_\mu}[Y])^2$ is the variance of $\bar{P}_\mu$. While (5.5.1) looks 'clean', it is not easy to analyze — for example, it is not a priori clear whether the derivative can be 0 in only one point. Taking further derivatives does not help either in this respect; for example, the second derivative is not necessarily always-positive.

Another 'straightforward' route to show the result via differentiation might be the following. We fix any prior with finite support, with positive probability $(1-\alpha)w(\mu_1^+) > 0$ on $\{\mu_1^+\}$ and $(1-\alpha)w(\mu_1^-) > 0$ on $\{\mu_1^-\}$ and, for $j = 1, \ldots, k$, prior αw_j on $\mu'_j \in \mathtt{M}_1 \backslash \mathrm{BD}(\mathtt{M}_1^{\mathrm{C}})$, i.e. μ'_j is not on the boundary of $\mathtt{M}_1$, so that $\sum_{j=1}^k w_j = w(\mu_1^+) + w(\mu_1^-) = 1$, $0 \leq \alpha \leq 1$. We let

$$q_\alpha(Y) := \sum_{j=1}^k w_j \bar{p}_{\mu_j}(Y) + w(\mu_1^+)\bar{p}_{\mu_1^+}(Y) + w(\mu_1^-)\bar{p}_{\mu_1^-}(Y).$$

Then, if we could show that the KL divergence

$$\mathbf{E}_{Q_\alpha}\left[\log \frac{q_\alpha(Y)}{p_0(Y)}\right] \tag{5.5.2}$$

were minimized by setting $\alpha = 0$, it would follow, by applying Theorem 9, that S_{GROW} is given by $p^*(Y)/p_0(Y)$ where $p^*(Y)$ must be of the form $w(\mu_1^-)\bar{p}_{\mu_1^-} + w(\mu_1^+)\bar{p}_{\mu_1^+}$.

Yet, if we try to show this by differentiating (5.5.2) with respect to α, we end up with a similarly hard-to-analyze expression as (5.5.1), and it is again not clear how to proceed.

These difficulties with showing the result in a straightforward way, by differentiation, only get exacerbated if $d > 1$. So, instead, we might try to extend the above lemma based on variation diminishing transformations to the case $d > 1$. But, literally quoting [19, Chapter 2], 'results concerning sign changes for multidimensional families appear very weak by comparison to their univariate cousins', and indeed we have not found any existing result in the literature that allows us to extend the above lemma to $d > 1$.

5.6 Discussion, Conclusion, Future Work

We have shown how GROW e-variables relative to alternative $\mathcal{H}_1$ defined in terms of a set of means $\mathtt{M}_1$ relate to a CSC probability bound on an event defined by the same $\mathtt{M}_1$. We first considered the case of convex $\mathtt{M}_1$; here our work consisted mostly of reformulating and re-interpreting existing results. We then considered nonconvex, surrounding $\mathtt{M}_1$. We showed how GROW and the individual-sequence-regret type of *relative* GROW again relate to a version of the CSC theorem, and we established some additional results for the case that $\mathtt{M}_1^{\mathrm{C}}$ is a fixed-radius KL ball for sample size 1, whereas we let the actual sample size n grow. As far as we are aware, our CSC bounds for surrounding $\mathtt{M}_1$ that are KL balls are optimal for this setting. It is of some interest though to consider the alternative setting in which, at sample size n, we consider a KL ball that has a fixed, or very slowly growing, radius when considering distributions on $Y = n^{-1}\sum_{i=1}^n X_i$ rather than on X. Thus, instead of (5.4.6) we now set, at sample size n,

$$\mathtt{M}_1^{\mathrm{C}} = \{\mu : D(\bar{P}_\mu \| P_0) < \underline{D}_n\} = \{\mu : D(\bar{P}'_\mu \| P'_0) < \underline{D}_n / n\}, \tag{5.6.1}$$

where either $\underline{D}_n = \underline{D}$ is constant, or very slowly growing in n. First consider the case that it is constant. Then, in terms of a single outcome, the corresponding ball in Euclidean (parameter space) shrinks at rate $1/\sqrt{n}$, the familiar scaling when we consider classical parametric testing. Since the boundary $\mathrm{BD}(\mathtt{M}_1^{\mathrm{C}})$ now changes with n, the asymptotics (5.4.5) we established above are not valid anymore. Therefore, while the CSC Theorem 12 is still valid, it may be hard to evaluate the bound (5.3.13) that it provides.

Now, for the setting (5.6.1), we may also heuristically apply the multivariate Central Limit Theorem (CLT): a second order Taylor approximation of $D(\bar{P}_\mu \| P_0)$ in a neighborhood of $\mu = 0$ gives that, up to leading order, $D(\bar{P}_\mu \| P_0) = \mu^\top J(0)\mu$, with $J(\mu)$ the Fisher information matrix of $\mathcal{E}$ in terms of the mean-value parameterization, which is equal to the inverse of the covariance matrix. The multivariate CLT then immediately gives that, as $n \to \infty$, we have that $P_0(Y \in \mathtt{M}_1^{\mathrm{C}}) \to A$, where A is the probability that a normally distributed random vector V, i.e. $V \sim N(0, I)$, with I the $(d-1) \times (d-1)$ identity matrix, falls in a Euclidean ball of radius $\sqrt{\underline{D}}$. This implies that the bound (5.3.13) would only remain relevant if for all large n, its right-hand side evaluates to a constant smaller than 1. We currently do not know if this is the case; it is an interesting question for future work.

Now let us consider the scaling (5.6.1) for the case that $\underline{D}_n$ is growing at the very slow rate $a(\log(b + c \log n)$ for suitable a, b and c. [51] give an *anytime-valid bound* for this case, in which the right-hand side is also a nontrivial constant (i.e. < 1), for all large enough n. Again, we do not know if we can replicate such bounds with our analyses — it is left for future work to determine this, and to further analyze the relation between anytime-valid bounds and the bounds we derived here, which are related to e-values and hence indirectly related to anytime-validity, but are not anytime-valid themselves.

Chapter 6

Discussion

In this dissertation, we presented several key mathematical results related to e-values and e-processes within exponential families. Specifically, we demonstrated the theoretical foundations for anytime-valid testing with well-specified exponential family models and developed four types of e-variables for k-sample tests, comparing their respective e-powers. We also analyzed the GROW e-variable under a composite $\mathcal{H}_1$ and a simple $\mathcal{H}_0$, introducing a novel multivariate concentration inequality. In this discussion, we summarize the main contributions of the dissertation and highlight potential directions for future research.

6.1 Well-specified model tests

We can interpret Chapters 2 and 3, as providing a comprehensive examination of testing whether an exponential family model is well-specified. This involves determining if the data fit the hypothesized exponential family distribution under a specific null hypothesis $\mathcal{H}_0$.

Grünwald et al. [42] showed that the growth-rate-optimal (GRO) e-variable can be obtained via a specific Bayes factor between $\mathcal{H}_1$ and $\mathcal{H}_0$. When both $\mathcal{H}_0$ and $\mathcal{H}_1$ are simple hypotheses, the Bayes factor reduces to a likelihood ratio, often referred to as a "simple e-variable." However, when $\mathcal{H}_0$ is composite, the GRO e-variable does not always correspond to a simple e-variable, though it does in some cases.

In Chapter 2, we presented a theorem, under certain regularity conditions, providing a general sufficient condition for the existence of a simple e-variable when testing a simple alternative against a composite regular exponential family null. This condition can be expressed as "$\Sigma_p(\boldsymbol{\mu}) - \Sigma_q(\boldsymbol{\mu})$ is positive semidefinite for all $\boldsymbol{\mu} \in \mathtt{M}_q$." We also explored the possibility of constructing GRO or close-to-GRO e-variables when this condition does not hold. We found that for some $\boldsymbol{\mu}$ within a specific parameter range, $q_{\boldsymbol{\mu}}/p_{\boldsymbol{\mu}}$ still provides a global simple e-variable; for others, it provides a local but not global e-variable, and for some, it does not provide an e-variable at all. An interesting direction for future work involves extending these results to *curved exponential families* [33].

While we have no general results in this area yet, the work of Liang [62] suggests that this may be feasible. Liang's variation of the Cochran-Mantel-Haenszel test involves a null hypothesis that can be reframed in terms of a curved exponential family, where a local e-variable exists by considering the second derivative of a specific function. This local e-variable is shown to be global by a different method than what we used in our construction, suggesting the potential for unifying these approaches.

In Chapter 3, we investigated the 'opposite' scenario, where "$\Sigma_p(\boldsymbol{\mu}) - \Sigma_q(\boldsymbol{\mu})$ is negative semidefinite for all $\boldsymbol{\mu} \in \mathtt{M}_q$". In addition, we mainly studied various types of e-variables and e-processes ($S_{\text{RIP}}, S_{\text{UI}}, S_{\text{COND}}, S_{\text{SEQ-RIP}}$) for multivariate exponential family null hypotheses and compared their e-power for i.i.d. data. In this context, we observed that in certain scenarios, the e-power of the "conditional" e-variable S_{COND} is asymptotically equal to the e-power of GRO e-variables in the "opposite" scenario. Additionally, we discovered an interesting phenomenon when considering composite alternative hypotheses, both in the Gaussian and general case, particularly regarding the relationship between conditional and RIPr e-variables, suggesting a near "approximate optional stopping" result.

Future work We also highlighted two e-variables that have not been extensively analyzed: the *sequential conditional e-variable* and a certain *weighted average of e-variables*. The former is a sequentialized version of the conditional e-variable, used in classical sequential testing and applicable to k-sample tests with exponential families [44], which we study in the next chapter. The latter is a weighted average of RIPr e-variables across different priors on the alternative, which, though an e-variable, behaves differently from the e-variables we focused on in this thesis.

Finally, future research could focus on relaxing the assumption that the distribution of the sufficient statistics X must have exponentially small tails under the alternative hypothesis. This regularity condition underpins most of our results, but its relaxation could broaden the applicability of e-variables in exponential family settings.

6.2 k-sample tests

In Chapter 4, we introduced and analyzed four types of e-variables for testing whether k groups of data are distributed according to the same element of an exponential family. These e-variables include the GRO e-variable (S_{RIP}), a conditional e-variable (S_{COND}), a mixture e-variable (S_{MIX}), and a pseudo-e-variable (S_{PSEUDO}).

Our analysis focused on comparing the growth rates of these e-variables under a simple alternative where each of the k groups has a distinct, but fixed, distribution within the same exponential family. We demonstrated that for any pair of e-variables $S, S' \in \{S_{\text{RIP}}, S_{\text{COND}}, S_{\text{MIX}}, S_{\text{PSEUDO}}\}$, the difference in their expected log-growth rates is $O(\delta^4)$, where δ represents the ℓ_2 distance between the alternative distribution's parameters and the null parameter space. This result indicates that when the effect size is small, the performance of all four e-variables is remarkably similar. For more substantial effect sizes, S_{RIP} has the highest growth rate by definition, making it the most powerful e-variable. However, calculating S_{RIP} requires determining the reverse information projection of the alternative distribution onto the null, which is computationally challenging. We provided theoretical results showing that for certain exponential families, one of the following equalities holds: $S_{\text{PSEUDO}} = S_{\text{RIP}}$, $S_{\text{COND}} = S_{\text{RIP}}$, or $S_{\text{MIX}} = S_{\text{RIP}}$. These cases significantly reduce the computational complexity of identifying the most effective e-variable. In instances where such equalities do not hold, algorithms can approximate the reverse information projection, and we verified numerically that these approximations lead to near-optimal values for S_{RIP}. Despite this, the choice of using S_{COND} or S_{MIX} might still be preferable due to their computational efficiency. Our simulations revealed that the optimal choice between S_{COND} and S_{MIX} depends on the specific exponential family under consideration, and in some cases, no clear ordering between them emerges.

These results provide practical insights into the trade-offs between different e-variables in terms of their theoretical properties and computational demands, guiding the selection of appropriate e-variables in real-world applications.

6.3 GROW e-variables and concentration inequality

Chapter 5 demonstrated how GROW e-variables, relative to an alternative hypothesis $\mathcal{H}_1$ defined by a set of means $\mathtt{M}_1$, connect to a *Csiszár-Sanov-Chernoff* (CSC) probability bound on events determined by the same set $\mathtt{M}_1$. Initially, we focused on cases where $\mathtt{M}_1$ is convex, largely involving reformulations and reinterpretations of known results. Subsequently, we developed results for nonconvex, surrounding $\mathtt{M}_1$, showing that both GROW and a form of *relative* GROW, based on individual-sequence regret, relate to a modified CSC theorem. For cases where $\mathtt{M}_1^{\mathrm{C}}$ is defined as a fixed-radius KL ball for sample size 1, we also derived results that hold as the actual sample size n increases.

To our knowledge, the CSC bounds we derived for surrounding $\mathtt{M}_1$ characterized by KL balls are the best available for this context. It is, however, interesting to consider a different approach: for sample size n, using a KL ball with a radius, in terms of the Euclidean distance in parameter space, that is decreasing as $O(\frac{1}{n})$ or $O(\frac{f(n)}{n})$ with $f(n)$ with very slowly increasing. Firstly, we consider the case that the KL ball is decreasing as $O(\frac{1}{n})$. Since the boundary $\mathrm{BD}(\mathtt{M}_1^{\mathrm{C}})$ now varies with n, our asymptotic results for the CSC bound derived earlier in Chapter 5 no longer apply in the same form. While the CSC theorem itself remains valid, evaluating the bound may present additional challenges.

Future work Now, let us explore the case that $f(n) = a \log(b + c \log n)$ for some suitable constants a, b, and c. Kaufmann and Koolen [51] provide an *anytime-valid bound* for this setting, where the bound's right-hand side also stabilizes to a nontrivial constant (i.e., less than 1) for all sufficiently large n.

It remains an open question whether our approach could yield similar bounds; addressing this is left as a potential direction for future work. Additionally, further analysis is needed to understand the relationship between anytime-valid bounds and those derived here. Although our bounds are related to e-values and thus indirectly connected to anytime-validity, they are not anytime-valid themselves.

Bibliography

[1] Reuben Adams. Safe hypothesis tests for the 2× 2 contingency table. Master's thesis, Delft University of Technology, 2020.

[2] Akshay Agrawal. Lecture notes on Loewner order, 2018.

[3] Theodore W Anderson and Donald A Darling. A test of goodness of fit. *Journal of the American statistical association*, 49(268):765–769, 1954.

[4] Francis J. Anscombe. Fixed-sample size analysis of sequential observations. *Biometrics*, 10(1):89–100, 1954.

[5] Yaser Awad, Shaul K Bar-Lev, and Udi Makov. A new class of counting distributions embedded in the Lee–Carter model for mortality projections: A Bayesian approach. *Risks*, 10(6):111, 2022.

[6] Akshay Balsubramani and Aaditya Ramdas. Sequential nonparametric testing with the law of the iterated logarithm. *Uncertainty in Artificial Intelligence*, 2016.

[7] Shaul K Bar-Lev. Independent, though identical results: the class of Tweedie on power variance functions and the class of Bar-Lev and Enis on reproducible natural exponential families. *International Journal of Statistics and Probability*, 9(1):30–35, 2020.

[8] Shaul K Bar-Lev. The exponential dispersion model generated by the Landau distribution —A comprehensive review and further developments. *Mathematics*, 11(20):4343, 2023.

[9] Shaul K Bar-Lev, Gérard Letac, and Ad Ridder. A delineation of new classes of exponential dispersion models supported on the set of nonnegative integers. *Annals of the Institute of Statistical Mathematics*, 2024.

[10] Shaul K Bar-Lev and Ad Ridder. New exponential dispersion models for count data – the ABM and LM classes. *ESAIM: Probability and Statistics*, 25:31–52, 2021.

[11] Shaul K Bar-Lev and Ad Ridder. Exponential dispersion models for overdispersed zero-inflated count data. *Communications in Statistics-Simulation and Computation*, 52(7):3286–3304, 2023.

[12] George A. Barnard. Review of abraham wald's *Sequential Analysis. Journal of the American Statistical Association*, 42(240):658–665, 1947.

[13] Ole E. Barndorff-Nielsen. *Information and exponential families: in statistical theory.* John Wiley & Sons, 2014.

[14] Andrew Barron, Jorma Rissanen, and Bin Yu. The minimum description length principle in coding and modeling. *IEEE transactions on information theory*, 44(6):2743–2760, 1998.

[15] Andrew R Barron and Chyong-Hwa Sheu. Approximation of density functions by sequences of exponential families. *The Annals of Statistics*, 19(3):1347–1369, 1991.

[16] Rabi N. Bhattacharya and R Ranga Rao. *Normal approximation and asymptotic expansions.* SIAM, 2010.

[17] Christopher Bishop. *Pattern recognition and machine learning.* Springer, 2006.

[18] William David Brinda. *Adaptive Estimation with Gaussian Radial Basis Mixtures.* PhD thesis, Yale University, 2018.

[19] Lawrence D. Brown. *Fundamentals of Statistical Exponential Families, with Applications in Statistical Decision Theory.* Institute of Mathematical Statistics, Hayward, CA, 1986.

[20] Lawrence D. Brown, Iain M. Johnstone, and K Brenda MacGibbon. Variation diminishing transformations: a direct approach to total positivity and its statistical applications. *Journal of the American Statistical Association*, 76(376):824–832, 1981.

[21] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games.* Cambridge university press, 2006.

[22] Bertrand S Clarke and Andrew R Barron. Information-theoretic asymptotics of bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471, 1990.

[23] Bertrand S. Clarke and Andrew R. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical planning and Inference*, 41(1):37–60, 1994.

[24] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley-Interscience, New York, 1991.

[25] Imre Csiszár. *I*-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3(1):146–158, 1975.

[26] Imre Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *Annals of Probability*, 12(3):768–793, 1984.

[27] Imre Csiszár and Frantisek Matus. Information projections revisited. *IEEE Transactions on Information Theory*, 49(6):1474–1490, 2003.

[28] Donald A. Darling and Herbert Robbins. Confidence Sequences for Mean, Variance, and Median. *Proceedings of the National Academy of Sciences*, 58(1):66–68, 1967.

[29] A Philip Dawid. Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–290, 1984.

[30] Martijn de Jong. Tests of significance for linear regression using E-values. Master's thesis, Leiden University, 2021.

[31] Boyan Duan, Aaditya Ramdas, and Larry Wasserman. Interactive rank testing by betting. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 201–235, 11–13 Apr 2022.

[32] Robin Dunn, Aditya Gangrade, Larry Wasserman, and Aaditya Ramdas. Universal inference meets random projections: a scalable test for log-concavity. *arXiv:2111.09254*, 2021.

[33] Bradley Efron. *Exponential families in theory and practice.* Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2022.

[34] Willy K Feller. Statistical aspects of ESP. *The Journal of Parapsychology*, 4(2):271, 1940.

[35] Peter Grünwald. *The minimum description length principle.* MIT press, 2007.

[36] Peter Grünwald. The E-posterior. *Philosophical Transactions of the Royal Society, Series A*, 2023.

[37] Peter Grünwald and A Philip Dawid. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, 32(4):1367–1433, 2004.

[38] Peter Grünwald and Steven De Rooij. Asymptotic log-loss of prequential maximum likelihood codes. In *International Conference on Computational Learning Theory*, pages 652–667. Springer, 2005.

[39] Peter Grünwald, Alexander Henzi, and Tyron Lardy. Anytime-valid tests of conditional independence under model-x. *Journal of the American Statistical Association*, 119(546):1554–1565, 2024.

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

[41] Peter Grünwald and Nishant Mehta. A tight excess risk bound via a unified PAC-Bayesian-Rademacher-Shtarkov-MDL complexity. In *Proceedings of the Thirtieth Conference on Algorithmic Learning Theory (ALT) 2019*, 2019.

[42] Peter Grünwald, Rianne De Heide, and Wouter Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae011, 2024.

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

[44] 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.

[45] Allard Hendriksen, Rianne de Heide, and Peter Grünwald. Optional Stopping with Bayes Factors: A Categorization and Extension of Folklore Results, with an Application to Invariant Situations. *Bayesian Analysis*, 16(3):961 – 989, 2021.

[46] Alexander Henzi and Johanna F. Ziegel. Valid sequential inference on probability forecast performance. *Biometrika*, 2022.

[47] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

[48] Jun ichi Takeuchi and Andrew R. Barron. Asymptotic minimax regret for Bayes mixtures. *arXiv preprint arxiv:2406.17929*, 2024.

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

[50] Bent Jørgensen. *The Theory of Exponential Dispersion Models*, volume 76 of *Monographs on Statistics and Probability.* Chapman and Hall, London, 1997.

[51] Emilie Kaufmann and Wouter M Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021.

[52] John L. Kelly. A new interpretation of information rate. *Bell System Technical Journal*, 35:pp. 917–26, 1956.

[53] Wouter Koolen and Peter Grünwald. Log-optimal anytime-valid e-values. *International Journal of Approximate Reasoning*, 2021. Festschrift for G. Shafer's 75th Birthday.

[54] Wojciech Kotłowski and Peter Grünwald. Maximum likelihood vs. sequential normalized maximum likelihood in on-line density estimation. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 457–476, 2011.

[55] Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[56] Tyron Lardy. E-values for hypothesis testing with covariates. Master's thesis, Leiden University, 2021.

[57] Tyron Lardy, Peter Grünwald, and Peter Harremoës. Reverse information projections and optimal e-statistics. *IEEE Transactions on Information Theory*, 2024. To Appear.

[58] Martin Larsson, Aaditya Ramdas, and Johannes Ruf. The numeraire e-variable and reverse information projection. *arXiv preprint arXiv:2402.18810*, 2024.

[59] Alix Lhéritier and Frédéric Cazals. A sequential non-parametric multivariate two-sample test. *IEEE Transactions on Information Theory*, 64(5):3361–3370, 2018.

[60] Qiang Jonathan Li. *Estimation of mixture models*. Yale University, 1999.

[61] Qiang Jonathan Li and Andrew R. Barron. Mixture density estimation. In *Advances in Neural Information Processing Systems*, volume 12, pages 279–285, 2000.

[62] Haoyi Liang. Stratified safe sequential testing for mean effect. Master's thesis, University of Amsterdam, 2023.

[63] Hubert W Lilliefors. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association*, 62(318):399–402, 1967.

[64] Michael Lindon, Dae Woong Ham, Martin Tingley, and Iavor Bojinov. *Anytime-valid Inference in Linear Models and Regression-adjusted Inference*. Harvard Business School, 2024.

[65] Alexander Ly, Udo Boehm, Peter Grünwald, Aaditya Ramdas, and Don van Ravenzwaaij. Safe anytime-valid inference: Practical maximally flexible sampling designs for experiments based on e-values, 2024.

[66] Carl N Morris. Natural exponential families with quadratic variance functions. *The Annals of Statistics*, pages 65–80, 1982.

[67] Francesco Orabona and Kwang-Sung Jun. Tight concentrations and confidence sequences from the regret of universal portfolio. *arXiv:2110.14099*, 2021.

[68] Teodora Pandeva, Tim Bakker, Christian A. Naesseth, and Patrick Forré. E-valuating classifier two-sample tests. *Transactions on Machine Learning Research*, 2024.

[69] Muriel Felipe Pérez-Ortiz. *Are we there yet?: Advances in anytime-valid methods for hypothesis testing and prediction.* PhD thesis, Leiden University, 2023.

[70] Muriel Felipe Pérez-Ortiz, Tyron Lardy, Rianne de Heide, and Peter D. Grünwald. E-statistics, group invariance and anytime-valid testing. *The Annals of Statistics*, 52(4):1410 – 1432, 2024.

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

[72] Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv:2009.03167*, 2020.

[73] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

[74] Richard Royall. *Statistical evidence: a likelihood paradigm.* Chapman and Hall, 1997.

[75] Shalev Shaer, Gal Maman, and Yaniv Romano. Model-Free Sequential Testing for Conditional Independence via Testing by Betting. In *International Conference on Artificial Intelligence and Statistics*, 2023.

[76] Glenn Shafer. Testing by betting: a strategy for statistical and scientific communication (with discussion and response). *Journal of the Royal Statistical Society A*, 184(2):407–478, 2021.

[77] Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, 26(1):84–101, 2011.

[78] Yu. M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):3–17, 1987.

[79] Michael A Stephens. EDF statistics for goodness of fit and some comparisons. *Journal of the American statistical Association*, 69(347):730–737, 1974.

[80] Jun-ichi Takeuchi and Andrew R. Barron. Asymptotically minimax regret for exponential and curved exponential families, 1997. Manuscript underlying abstract at the 1998 International Symposium on Information Theory (14 pages). Available at `http://www.stat.yale.edu/ arb4/publications.html`.

[81] Jun-ichi Takeuchi and Andrew R. Barron. Asymptotically minimax regret for exponential families. In *Proceedings SITA '97*, pages 665–668, 1997.

[82] Jun-ichi Takeuchi and Andrew R. Barron. Asymptotically minimax regret by Bayes mixtures for non-exponential families. In *2013 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2013.

[83] Judith Ter Schure, Muriel Felipe Pérez-Ortiz, Alexander Ly, and Peter Grünwald. The anytime-valid logrank test: Error control under continuous monitoring with unlimited horizon. *The New England Journal of Statistics in Data Science*, 2(2):190–214, 2024.

[84] Flemming Topsøe. Information-theoretical optimization techniques. *Kybernetika*, 15(1):8–27, 1979.

[85] Timmy Tse and Anthony C Davison. A note on universal inference. *Stat*, 11(1):e501, 2022.

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

[87] Rosanne Turner and Peter Grünwald. Safe sequential testing and effect estimation in stratified count data. In *Annual AI and Statistics Conference*, 2023.

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

[89] Tim van Erven. Blog post on large deviations: Cramér vs. Sanov, 2012.

[90] Jean Ville. *Etude critique de la notion de collectif.* Gauthier-Villars Paris, 1939.

[91] Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

[92] Vladimir Vovk and Ruodu Wang. Nonparametric e-tests of symmetry. *The New England Journal of Statistics in Data Science*, 2(2):261–270, 2024.

[93] Abraham Wald. *Sequential Analysis.* John Wiley & Sons, Inc., New York; Chapman & Hall, Ltd., London, 1947.

[94] Qiuqi Wang, Ruodu Wang, and Johanna Ziegel. E-backtesting. *arXiv preprint arXiv:2209.00991*, 2022.

[95] Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.

[96] Ian Waudby-Smith, David Arbour, Ritwik Sinha, Edward H Kennedy, and Aaditya Ramdas. Time-uniform central limit theory and asymptotic confidence sequences. *arXiv:2103.06476*, 2021.

[97] Ulla-Britt Wennerholm, Sissel Saltvedt, Anna Wessberg, Mårten Alkmark, Christina Bergh, Sophia Brismar Wendel, Helena Fadl, Maria Jonsson, Lars Ladfors, Verena Sengpiel, et al. Induction of labour at 41 weeks versus expectant management and induction of labour at 42 weeks (Swedish post-term induction study, swepis): multicentre, open label, randomised, superiority trial. *British Medical Journal*, 367, 2019.

[98] David Williams. *Probability with martingales.* Cambridge university press, 1991.

[99] William Henry Young. On classes of summable functions and their fourier series. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 87(594):225–229, 1912.

[100] Zhenyuan Zhang, Aaditya Ramdas, and Ruodu Wang. On the existence of powerful p-values and e-values for composite hypotheses. *arXiv preprint arXiv:2304.16539*, 2023.

List of publications

This dissertation is based on the following papers.

- Chapter 2 is based on
  Peter Grünwald, Tyron Lardy, Yunda Hao, Shaul K Bar-Lev, and Martijn Jong. Optimal E-Values for Exponential Families: the Simple Case. arXiv preprint arXiv:2404.19465, 2024. It has been submitted as a contribution to the *Festschrift for Andrew Barron on the Occasion of his 65th Birthday*.

- Chapter 3 is based on
  Yunda Hao, and Peter Grünwald. E-Values for Exponential Families: the General Case. arXiv preprint arXiv:2409.11134, 2024, under submission.

- Chapter 4 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.

- Chapter 5 is based on
  Peter Grünwald, Yunda Hao, and Akshay Balsubramani. Growth-Optimal E-Variables and an extension to the multivariate Csiszár-Sanov-Chernoff Theorem. arXiv preprint arXiv:2412.17554, 2024.

The following paper was completed during the PhD period but is not included in this dissertation.

- Chengli Tan, Jiangshe Zhang, Junmin Liu, Yicheng Wang, and Yunda Hao. Stabilizing Sharpness-aware Minimization Through A Simple Renormalization Strategy. arXiv preprint arXiv:2401.07250, 2024. Journal of Machine Learning Research, accept after minor revision.

Samenvatting

Dit proefschrift richt zich voornamelijk op statistisch hypothesetoetsen, een onderzoeksgebied met brede relevantie in verschillende academische disciplines en industrieën. Een voorbeeld dat laat zien hoe belangrijk dit is het ontwikkelen van nieuwe geneesmiddelen. Stel dat onderzoekers een nieuw medicijn aandragen dat bedoeld is om de bloeddruk te verlagen. De nulhypothese zou kunnen stellen dat het nieuwe medicijn geen effect heeft op de bloeddruk, terwijl de alternatieve hypothese suggereert dat het medicijn de bloeddruk wèl verlaagt.

De onderzoekers voeren een klinische studie uit waarbij een groep het nieuwe medicijn krijgt en een andere groep een placebo. Stel dat na het verzamelen van de gegevens een verlaging van de bloeddruk in de eerste groep wordt waargenomen. Met een hypothesetoets kan vervolgens (statistisch) uitgesloten worden dat de waargenomen verlaging op toeval berust. Als de toets aangeeft dat de bevinding statistisch significant is, dan concluderen de onderzoekers dat het medicijn effectief is. Zonder hypothesetoetsing zouden ze niet in staat zijn om rigoureus te beoordelen of het medicijn echt werkt of dat de waarnemingen slechts het gevolg zijn van willekeurige variaties. Door een kader te bieden om fouten (zoals fout-positieve resultaten) te beheersen, zorgt hypothesetoetsing ervoor dat de meeste medicijnen die goedgekeurd worden ook daadwerkelijk effectief zijn, wat cruciaal is voor de volksgezondheid en de vooruitgang van de medische wetenschap.

De meeste klassieke methoden voor hypothesetoetsing vereisen echter dat onderzoekers – voordat de toets wordt uitgevoerd – een (vaste) steekproefgrootte bepalen. Nadat de steekproef van de vooraf vastgestelde grootte is verzameld, wordt de toets uitgevoerd en worden conclusies getrokken over het al dan niet verwerpen van de nulhypothese.

Deze vaste steekproefgroottebenadering kent verschillende beperkingen:

1. **Vastgestelde steekproefgrootte**: Onderzoekers moeten van tevoren beslissen

hoeveel gegevenspunten ze verzamelen, wat kan leiden tot studies waarbij het onderscheidend vermogen van de toets of te laag, of te hoog is.

2. **Geen tussentijdse analyse**: In klassieke hypothesetoetsing mogen onderzoekers doorgaans niet naar de gegevens kijken terwijl deze binnenkomen (om bevooroordeelde beslissingen te voorkomen) en moeten zij wachten tot de volledige steekproef beschikbaar is.

3. **Inflexibiliteit**: Als onverwachte resultaten optreden of als de steekproefgrootte na gegevensverzameling onvoldoende blijkt te zijn, kunnen onderzoekers de steekproefgrootte niet eenvoudig aanpassen zonder het risico te lopen de kans op een fout-positief resultaat substantieel te vergroten.

Ondanks deze beperkingen vormt de vaste steekproefgrootte-benadering al vele decennia de basis van statistische toetsing, en wordt deze nog steeds zeer breed toegepast. Moderne methoden, zoals zogenoemde altijd-valide toetsen (bijvoorbeeld op basis van e-waarden en e-processen), bieden echter meer flexibiliteit. Ze stellen onderzoekers in staat om het bewijs, gekwantificeerd in termen van de e-waarde, continu te evalueren terwijl gegevens worden verzameld, zonder een steekproefgrootte vooraf te specificeren. Voor e-processen hoeven zelfs de regels voor het stoppen van het experiment niet vooraf te worden bepaald.

Dit proefschrift richt zich voornamelijk op e-waarden en e-processen in de context van exponentiële families.

Hoofdstuk 2 behandelt het probleem van het bepalen of een steekproef verdeeld is volgens een distributie van een specifiek model binnen de exponentiële familie, in wezen het toetsen of een model correct is gespecificeerd. We willen bijvoorbeeld beoordelen of een steekproef een Gaussische verdeling volgt. In dit geval omvat de nulhypothese de volledige set Gaussische verdelingen, waardoor de nulhypothese composiet is. Het doel is de nulhypothese te verwerpen als er overtuigend bewijs is dat de steekproef niet verdeeld is volgens een Gaussische verdeling. We richten ons op de GRO (Growth-Rate Optimal) e-variabele, die doorgaans overeenkomt met een specifieke Bayes-factor en het hoogst onderscheidende e-vermogen heeft (d.w.z. het vermogen om alternatieven te detecteren). Het vinden van de a priori verdelingen voor de GRO e-variabele kan echter computationeel intensief zijn. Dit hoofdstuk toont aan dat in bepaalde situaties, de zogenoemde “eenvoudige gevallen”, de GRO e-variabele vereenvoudigt tot een aannemelijkheidsverhouding en biedt verschillende equivalente voorwaarden waaronder een dergelijke aannemelijkheidsverhouding bestaat voor exponentiële familie-nulhypothesen. In dit hoofdstuk worden ook (GRO) e-processen afgeleid welke gebruikt kunnen worden om een altijd valide toets te construeren.

Hoofdstuk 3 breidt het werk van Hoofdstuk 2 uit door meer algemene theoretische resultaten te bieden voor verschillende e-variabelen in het kader van modelspecificatietoetsen, zowel voor eenvoudige als composiete hypothesen. Het toont aan dat in het "anti-eenvoudige geval" (het tegenovergestelde van het eenvoudige geval), de conditionele e-waarde asymptotisch het hoogst onderscheidende e-vermogen bereikt. Dit is bijzonder waardevol omdat de GRO e-variabele moeilijk te berekenen is in dergelijke gevallen, terwijl de conditionele e-variabele computationeel werkbaar is. Daarnaast introduceert en vergelijkt dit hoofdstuk verschillende soorten e-waarden, waaronder de GRO e-variabele, de conditionele e-variabele, de UI (Universal Inference) e-variabele en sequentiële e-variabelen, met een uitgebreide analyse van hun onderscheidende e-vermogen. Sommige van deze e-waarden vormen de basis van een e-proces, dat dan weer leidt tot een altijd valide toets.

Hoofdstuk 4 richt zich op het construeren van e-variabelen die gebruikt kunnen worden om te toetsen of k steekproeven allen hetzelfde verdeeld zijn volgens een verdeling van de exponentiële familie. Bijvoorbeeld, in het geval van twee-steekproeventoets ($k = 2$) kunnen deze methoden worden toegepast om te beoordelen of een nieuwe medische behandeling effectief is voor een bepaalde ziekte door de uitkomsten tussen de behandelings- en controlegroep te vergelijken. We introduceren vier soorten e-variabelen voor de k-steekproeventoets: de GRO e-variabele, een conditionele e-variabele, een gewogen middeling van e-variabelen en een pseudo-e-variabele. Deze e-variabelen worden vergeleken op basis van hun groeisnelheden onder alternatieve hypothesen, waarbij elke groep een andere, maar vaste, verdeling heeft uit dezelfde exponentiële familie. Het proefschrift biedt theoretische resultaten die aantonen dat bij kleine effectgroottes de e-variabelen vergelijkbaar gedrag vertonen. Ook worden gevallen geïdentificeerd waarin een e-variabele vereenvoudigt tot de GRO e-variabele, wat de computationele complexiteit vermindert. In complexere situaties worden algoritmen voorgesteld om de zogenoemde omgekeerde informatieprojectie te benaderen.

Hoofdstuk 5 richt zich op *growth-rate optimal in the worst-case* (GROW) e-variabelen. We analyseren de toepassing van GROW e-variabelen binnen een hypothesetoetsingskader voor multivariate verdelingen. In dit kader heeft de nulhypotheseverdeling P_0 een verwachting van nul, en worden verschillende alternatieve hypothesen $\mathcal{H}_1$ gedefinieerd door middel van verschillende gebieden voor de verwachting.

We laten zien dat de GROW e-variabele verbonden is met een nieuwe concentratieongelijkheid die we de *Csiszár-Sanov-Chernoff* (CSC) ongelijkheid noemen. Deze CSC-ongelijkheid breidt eerder werk uit naar multivariate gevallen met een convexe of begrensde alternatieve parameterregio rond 0. Een dergelijke ongelijkheid is waarschijnlijk waardevol in praktische toepassingen, zoals *online* (sequentieel) leren en vooral op het gebied van *bandit*-algoritmen.

Samenvattend: dit proefschrift breidt de theorie van e-variabelen binnen exponentiële families uit door verschillende e-variabelen te construeren en te analyseren, en biedt praktische inzichten voor situaties waarin traditionele p-waarde-gebaseerde toetsing tekortschiet.

Summary

This dissertation primarily focuses on statistical hypothesis testing, a critical area of study with widespread importance across various academic disciplines and industries. An example that highlights the importance of hypothesis testing comes from the field of medicine, specifically in drug development. Suppose researchers are testing a new drug designed to lower blood pressure. The null hypothesis might state that the new drug has no effect on blood pressure, while the alternative hypothesis suggests that the drug does reduce blood pressure. The researchers conduct a clinical trial, giving one group the new drug and another group a placebo. After collecting the data, the method of *statistical hypothesis testing* allows them to analyze whether the observed reduction in blood pressure is significant or could have occurred by chance. If the test provides strong evidence against the null hypothesis, the researchers can confidently conclude that the drug is effective. Without statistical hypothesis testing, they cannot rigorously assess whether the drug truly works or if the results are just random variations. By providing a framework to control for errors (like false positives), hypothesis testing ensures that the drug is only approved if there is strong statistical evidence of its efficacy, which is critical for public safety and the advancement of medical science.

However, most classical hypothesis testing methods require researchers to collect a fixed sample size in advance before conducting the test. Once the data from the predetermined sample size is collected, the test is performed, and conclusions are drawn about whether to reject or fail to reject the null hypothesis.

The fixed-sample approach has several limitations:

1. **Pre-determined sample size**: Researchers must decide in advance how many data points to collect, which may lead to underpowered or overpowered studies if the chosen sample size is not optimal.

2. **No intermediate analysis**: In classical hypothesis testing, researchers are usually not allowed to look at the data as it comes in (to prevent biased decisions) and

must wait until the full dataset is available.

3. **Inflexibility**: If unexpected results occur or if the sample size turns out to be inadequate after data collection, researchers cannot easily adjust the sample size without risking inflating the Type I error (false positive rate).

Despite these limitations, the fixed-sample approach has been the foundation of statistical testing for many decades and remains widely used. However, newer methods, such as anytime-valid tests (e.g., e-value- and e-process-based methods), offer more flexibility, allowing researchers to evaluate the evidence continuously as data is collected, without needing a pre-specified sample size. With e-values, one does not even need to determine the rules for stopping the experiment before it starts.

This dissertation primarily explores e-values and e-processes within the context of exponential families.

Chapter 2 addresses the problem of determining whether a sample conforms to a specific exponential family model, essentially testing whether a model is well-specified. For example, we may want to assess whether a sample follows a Gaussian distribution. In this case, the null hypothesis includes the entire set of Gaussian distributions, making it composite. The objective is to reject the null hypothesis if the sample deviates from a Gaussian distribution. A key focus is on the GRO (Growth-Rate Optimal) e-variable, which typically corresponds to a specific Bayes factor and has the highest e-power (i.e., the ability to detect alternatives) among all e-variables. However, finding the prior for the GRO e-variable can be computationally expensive. This chapter demonstrates that in certain scenarios, termed "simple cases," the GRO e-variable simplifies to a simple-vs.-simple likelihood ratio, and it provides several equivalent conditions under which such a likelihood ratio exists for exponential family null hypotheses.

Chapter 3 extends the work of Chapter 2 by offering more general theoretical results for several e-variables in the context of testing model specification, covering both simple and composite hypotheses. It is shown that in the "anti-simple case" (the opposite of the simple case), the conditional e-value achieves asymptotically highest e-power. This is particularly valuable because, while the GRO e-variable is hard to compute in such cases, the conditional e-variable is computationally more straightforward. Additionally, this chapter introduces and compares various types of e-values, including the GRO e-variable, the conditional e-variable, the UI (Universal Inference) e-variable, and sequential e-variables, with a detailed analysis of their e-power. Some of these e-values also lead to the development of anytime-valid tests, known as e-processes.

Chapter 4 focuses on developing methods for testing whether k groups of samples are distributed according to the same element of an exponential family, using e-values.

For instance, in the case of two-sample tests ($k = 2$), these methods can be applied to assess whether a new medical treatment is effective for a particular disease by comparing the distribution of outcomes between the treatment and control groups. We introduce four types of e-variables for the k-sample test: the GRO e-variable, a conditional e-variable, a mixture e-variable, and a pseudo-e-variable. These e-variables are compared in terms of their growth rates under alternative hypotheses, where each group has a different, but fixed, distribution from the same exponential family. The paper provides theoretical results showing that, under small effect sizes, the e-variables behave similarly. It also identifies cases where one e-variable simplifies to the GRO e-variable, reducing computational complexity. In more complex settings, algorithms for estimating the reverse information projection are suggested.

Chapter 5 focuses on *growth-rate optimal in the worst-case* (GROW) e-variables. We analyze the application of GROW e-variables within a hypothesis testing framework for multivariate distributions. In this framework, the null hypothesis distribution P_0 has a mean of zero, and various alternative hypotheses $\mathcal{H}_1$ are defined by different sets of means. Interestingly, we show that the GROW e-variable connects to a new concentration inequality we call the *Csiszár-Sanov-Chernoff* (CSC) bound. This CSC inequality extends earlier work to handle multivariate cases with either a convex or bounded alternative parameter region around 0. Such an inequality is likely to be valuable in practical applications, such as online learning, and especially in the field of *bandit algorithms.*

In summary, this dissertation expands e-variable theory within exponential families by developing and analyzing several e-variables, providing practical insights for situations where traditional p-value-based testing may fall short.

总　结

本论文主要研究统计假设检验，这是一个在多个学术领域和行业中具有广泛重要性的关键研究领域。一个能够突出假设检验重要性的例子来自医学领域，特别是在药物开发中。假设研究人员正在测试一种新型降血压药物。我们设定原假设是这种新药对血压没有影响，而备择假设则表明这种药物确实能降低血压。

为验证上述假设，研究人员进行了一项临床试验，其中一组人服用新药，另一组人服用安慰剂。数据收集完成后，假设检验方法可以帮助研究人员分析观察到的血压降低是否具有统计显著性，或者是否纯属偶然。如果检验结果提供了拒绝原假设的强有力证据，研究人员就可以有信心地得出结论：该药物是有效的。假设检验通过提供一个控制错误（例如假阳性）的方法框架，确保只有存在强有力统计证据表明药物有效的情况下才会批准该药物，这对于公共安全和医学科学的发展至关重要。

然而，大多数经典的假设检验方法通常要求研究人员在检验前预先确定一个固定的样本量。一旦收集了预先确定样本量的数据，就会执行检验，并得出是否拒绝原假设的结论。

固定样本量的方法存在以下几个限制：

1. **预先确定样本量**：研究人员必须提前决定收集多少数据点，如果所选样本量不够理想，可能会导致研究的统计功效不足或过高。

2. **无法进行中间分析**：在经典的假设检验中，研究人员通常不被允许在数据收集中途查看数据（以避免决策偏差），必须等到整个数据集可用后才能进行分析。

3. **缺乏灵活性**：如果出现意外结果，或者数据收集后样本量被证明不足，研究人员无法轻松调整样本量，否则可能增加一类错误（假阳性率）的风险。

尽管存在这些限制，固定样本量的方法仍然是统计检验的基础，已被广泛应用了数十年。然而，更新的方法（如序贯分析和任意时间有效检验，例如e-value 方法和e-process）提供了更大的灵活性，使研究人员能够在数据收集中连续评估证据，而无需预先指定样本量。e-value 甚至不需要在实验开始前确定停止检验的规则。

本论文主要探讨了在指数族分布下的e-value 和e-process 方法。

第 2章研究如何判断一个样本是否符合特定的指数族模型，本质上是检验模型是否被正确地指定。例如，我们可能希望评估一个样本是否符合高斯分布。在这种情况下，原假设包括所有高斯分布的集合，使其成为复合假设。目标是当样本偏离高斯分布时拒绝原假设。本章的重点是GRO（增长率最优）e-variable，其通常对应于特定的贝叶斯因子，并具有最高的e-power（即当备择假设为真时选择备择假设的能力）。然而，找到GRO e-variable的先验分布可能在计算上非常耗费资源。本章展示了在某些情况下（被称为“简单情况”），GRO e-variable简化为似然比，并提供了多种等价条件，用以描述在指数族原假设下，何时这种似然比存在。本章还创造了一种e-process 方法，证明了GRO e-variable是一种任意有效的检验方法。

第 3章在第 2章的基础上进一步扩展，为多个e-variables 在模型正确性检验中的应用提供了更一般的理论结果，涵盖了简单和复合假设。本章表明，在“非简单情况”（简单情况的对立情况）中，条件e-value 可以实现渐近意义上的最高e-power。这一点尤为重要，因为在这些情况下，虽然GRO e-variable 难以计算，但是条件e-variable 在计算上非常简便。此外，本章介绍并比较了多种类型的e-variables，包括GRO e-variable、条件e-variable、UI（一致推断）e-variable和序贯e-variable，并对其e-power 进行了详细分析，并且给出渐近表达式。我们还证明了其中一些e-variables是任意有效检验方法，即e-process。

第 4章聚焦于多组样本检验，使用e-value方法检验k 组样本是否按照指数族中的相同元素分布。例如，在两样本检验($k = 2$) 时，运用e-value方法比较治疗组和对照组的结果分布，从而来评估一种新的医疗治疗对某种疾病是否有效。本章提出了四种k 样本检验的e-variables：GRO e-variable、条件e-variable、混合e-variable和伪e-variable。在每组样本服从相同指数族中的不同但固定的分布时，我们对这些e-variables的e-power进行了比较。本章提供的理论结果表明，在不同的分布之间的差异较小时，这些e-variables表现类似。同时，在某些特定分布下，某种e-variable等价为GRO e-variable，从而降低计算复杂性。在更复杂的情况下，我们提供了一种算法用来估计GRO e-variable。

第 5章重点研究了最差情况下增长率最优（GROW）e-variable。我们分析了GROW e-variable在多元分布假设检验框架中的应用。在此框架中，原假设分布P_0 的均值为零，而各种备择假设$\mathcal{H}_1$ 则由不同的均值集合定义。值得注意的是，我们发现GROW e-variable与一种新的集中不等式相关联，我们将其称为*Csiszár-Sanov-Chernoff (CSC)*不等式。该CSC 不等式将先前的研究扩展到处理多元情况，并支持围绕0 的凸或有界的备择参数区域。这种不等式在实际应用中可能具有重要价值，例如在线学习领域，尤其是Bandit 算法。

总结：本论文通过创立和分析多种e-variables扩展了指数族中的e-value和e-process理论，并为传统p值检验可能失效的情境提供了可行的新一套方法论。

Acknowledgements

Completing this Ph.D. has been a truly transformative experience, and I am deeply grateful for the support and encouragement I received throughout this journey. This dissertation would not have been possible without the guidance, encouragement, and assistance of many individuals and institutions.

First and foremost, I am extremely grateful to my supervisors Peter Grünwald and Alexander Ly, whose expertise, support, and insight guided me through every stage of this work. I am deeply grateful to Peter for his patient guidance, which has significantly expanded my knowledge, especially in the areas of statistics and information theory. I also appreciate his support in helping me improve my English and his kind care during the Covid-19 pandemic, which was a lonely two-year period after I had arrived in the Netherlands. I am thankful to Alexander for consistently offering insightful scientific discussions and valuable suggestions for my scientific writing.

I would also like to thank my committee members, Prof.dr. Thorsten Dickhaus, Dr. Zhimei Ren, Dr. Nick Koning, Prof.dr.ir. Gianne Derks and Prof.dr. Marta Fiocco for sharing their time, support, and expertise. Their insightful suggestions and critical perspectives helped me refine my work.

My sincere appreciation also extends to my collaborators. I am especially grateful to Wouter Koolen, who introduced me to the field of optimization and taught me to approach scientific research from multiple perspectives. I would also like to thank Tyron Lardy, Long Long, Reuben Adams, Chengli Tan, and many other remarkable researchers for sharing insightful scientific discussions. I could not have embarked on my PhD journey without the financial support from the China Scholarship Council.

I had a wonderful four years at CWI, and I'm deeply grateful to everyone in the Machine Learning group who made this journey so memorable. My heartfelt thanks to Muriel Pérez, Udo Böhm, Tao Sun, Bojian Yin, Tom Sterkenburg, Sanne van den Berg, Hongye Chen, Lucy Zhang, and Aditya Gilra for our post-lunch ping-pong sessions—they were a highlight of my days. I also want to thank CWI for providing such a comfortable office space and a supportive work environment. Special thanks to Nada Mitrovic and Duda Tepsic for their invaluable support in my research. And, of course, gratitude to the canteen staff for always serving a variety of food.

I'm equally grateful for the amazing friends I met along the way. Big thanks to Xinyu Hu, Gao Peng, Sirui Shen, Haohui Zhang, Hao Qiu, Zekun Ji, Wenjun Huang, Haitian Wang, Kaixiang Zhang, Xinyuan Wang, Tianyuan Wang, and many other wonderful people. I cherished every moment we spent playing sports, sharing dinners, and laughing together during our post-work conversations. Finally, I would like to thank my landlord, Flora Zhang, as well as my (former) flatmates—Zekai Jin, Zejing Zhu, Yiming Xu, Hikaru Takano, and Haoqing Xu in Amsterdam—for their wonderful company and the lively conversations that brought joy to my evenings.

I am deeply grateful to Prof. Limin Li for supervising me during my master's program. She introduced me to academic research, and I learned so much under her guidance. I also appreciate her warm hospitality every time I visited Xi'an. My heartfelt thanks also go to Prof. Xu Guo for inviting me to give a talk at Beijing Normal University.

I extend my gratitude to all my old friends from before my PhD journey—your companionship during that chapter of my life meant so much to me.

Lastly, and most importantly, I thank my parents and grandparents for their unwavering support. I am also profoundly thankful to my girlfriend, Qianxin Yi, who has always been by my side, offering love and encouragement, even from afar in China. Without their support and love, completing this dissertation would not have been possible.

Yunda Hao
Amsterdam, November 2024

Curriculum Vitae

The author of this dissertation was born in 1994 in Cangzhou, Hebei Province, China. After completing three years of study at Cangzhou No.2 Senior High School (2010–2013), he pursued a bachelor's degree in mathematics at Hebei University in Baoding, China (2013–2017). Following that, he earned a master's degree (Outstanding Graduate) in applied statistics at Xi'an Jiaotong University under the guidance of Prof. Dr. Limin Li in Xi'an, China (2017–2020). The author then moved to the Netherlands, where he completed his PhD at Centrum Wiskunde & Informatica (CWI) in Amsterdam under the supervision of Prof. Dr. Peter Grünwald and Dr. Alexander Ly (2020–2024).