Universiteit Leiden
The Netherlands

Optimal test statistics for anytime-valid hypothesis tests

Lardy, T.D.

3 | On the Optimality of E-statistics

In the previous chapter, we defined the log-optimal, or GRO, e-statistic in the context of anytime-valid testing. In this chapter, we shift the focus to the properties of the GRO e-statistic, independent of that specific context. In particular, we discuss the form of the GRO e-statistic, as well as limitations of Definition 2.6.

To this end, it has recently been shown that, for testing simple alternatives against composite null hypotheses, there is a one-to-one correspondence between the GRO e-statistic and the reverse information projection (RIPr). The latter is an object that arises in information theory and is defined as the measure in the null that is closest to the alternative in information divergence. However, the RIPr as well as the GRO e-statistic are not uniquely defined when the infimum information divergence between the null and alternative hypothesis is infinite. We show that in such scenarios, under some assumptions, there still exists a measure in the null that is closest to the alternative in a specific sense. Whenever the information divergence is finite, this measure coincides with the usual RIPr. It therefore gives a natural extension of the RIPr to certain cases where the latter was previously not defined. This extended notion of the RIPr is shown to lead to optimal e-statistics in a sense that is a novel, but natural, extension of the GRO criterion. We also give conditions under which the (extension of the) RIPr is a strict sub-probability measure, as well as conditions under which an approximation of the RIPr leads to approximate e-statistics. For this case we provide tight relations between the corresponding approximation rates.

Throughout this chapter, we assume that the null hypothesis is convex, which is not true for most models used in practice. However, from the perspective of testing with e-statistics, it does not matter whether one considers a given model or its convex hull, because taking the convex hull does not change the set of all e-statistics.

3.1 Introduction

We write $D(\nu\|\lambda)$ for the information divergence (Kullback-Leibler divergence, (Kullback and Leibler, 1951; Csiszár, 1963; Liese and Vajda, 1987)) between two finite measures ν and λ given by

$$D(\nu\|\lambda) = \begin{cases} \int_\Omega \ln\left(\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}\right) \mathrm{d}\nu - (\nu(\Omega) - \lambda(\Omega)), & \text{if } \nu \ll \lambda; \\ \infty, & \text{else.} \end{cases}$$

For probability measures the interpretation of $D(\nu\|\lambda)$ is that it measures how much we gain by coding according to ν rather than coding according to λ if data are distributed according to ν. Many problems in probability theory and statistics, such as conditioning and maximum likelihood estimation, can be cast as minimization in either or both arguments of the information divergence. In particular, this is the case within the recently established and now flourishing theory of hypothesis testing based on e-statistics that allows for optional continuation of experiments (see Section 3.2.3)(Grünwald et al., 2024; Ramdas et al., 2023; Vovk and Wang, 2021; Shafer, 2021; Henzi and Ziegel, 2022). That is, a duality has been established between optimal e-statistics for testing a simple alternative P against a composite null hypothesis $\mathcal{C}$ and reverse information projections (Grünwald et al., 2024). Here, the reverse information projection (RIPr) of P on $\mathcal{C}$ is — if it exists — a unique measure $\hat{Q}$ such that every sequence $(Q_n)_{n\in\mathbb{N}}$ in $\mathcal{C}$ with $D(P\|Q_n) \to \inf_{Q\in\mathcal{C}} D(P\|Q)$ converges to $\hat{Q}$ in a particular norm (Li, 1999; Csiszár and Matúš, 2003). Li (1999) showed that whenever $\mathcal{C}$ is convex and $D(P\|\mathcal{C}) := \inf_{Q\in\mathcal{C}} D(P\|Q) < \infty$, the RIPr $\hat{Q}$ exists and the likelihood ratio between P and $\hat{Q}$ is an e-statistic (this result is restated as Theorem 3.1 below). Grünwald et al. (2024) showed (restated as Theorem 3.3 below) that it is even the optimal e-statistic for testing P against $\mathcal{C}$. However, it is clear that the RIPr cannot be defined in this way if the information divergence between P and $\mathcal{C}$ is infinite, i.e. $D(P\|\mathcal{C}) = \infty$. This leaves a void in the theory of optimality of e-statistics. In this chapter we remedy this by realizing that even if all measures in $\mathcal{C}$ are infinitely worse than P at describing data distributed according to P itself, there can still be a measure that performs best relative to the elements of $\mathcal{C}$. To find such a measure, we consider the *description gain* (Topsøe, 2007) given by

$$D(P\|Q \rightsquigarrow Q') = \int_\Omega \ln\left(\frac{\mathrm{d}Q'}{\mathrm{d}Q}\right) \mathrm{d}P - (Q'(\Omega) - Q(\Omega)) \tag{3.1}$$

whenever this integral is well-defined. If the quantities involved are finite then the description gain reduces to

$$D(P\|Q \rightsquigarrow Q') = D(P\|Q) - D(P\|Q'). \tag{3.2}$$

In analogy to the interpretation of information divergence for coding, the description gain measures how much we gain by coding according to Q' rather than Q if data are distributed according to P. Furthermore denote

$$D(P\|Q \rightsquigarrow \mathcal{C}) := \sup_{Q' \in \mathcal{C}} D(P\|Q \rightsquigarrow Q'),$$

where undefined values are counted as $-\infty$ when taking the supremum. If there exists at least one $Q^* \in \mathcal{C}$ such that $P \ll Q^*$, then $D(P\|Q \rightsquigarrow \mathcal{C})$ is a well-defined number in $[0, \infty]$ for any $Q \in \mathcal{C}$. This quantity should be seen as the maximum description gain one can get by switching from Q to any other measure in $\mathcal{C}$. Intuitively, if there is a best descriptor in $\mathcal{C}$, nothing can be gained by switching away from it. Indeed, in Proposition 3.6 we show that $\inf_{Q \in \mathcal{C}} D(P\|Q \rightsquigarrow \mathcal{C})$ is finite if and only if it is equal to zero.

3.1.1 Contents and Overview

Below, in Section 3.2, we start by giving an overview of existing results on both the reverse information projection and e-statistics, which we define and briefly motivate, and the growth-rate optimality (GRO) criterion, a natural replacement of statistical power within the context of e-value based hypothesis testing. Section 3.3 states Theorem 3.5, our first central result. It shows that — under very mild conditions — there exists a unique measure $\hat{Q}$ such that every sequence $(Q_n)_{n \in \mathbb{N}}$ in $\mathcal{C}$ with

$$D(P\|Q_n \rightsquigarrow \mathcal{C}) \to 0$$

converges to $\hat{Q}$ in a specific metric which we define. Thus, Theorem 3.5 may be viewed as a generalization of Li's result stated below as Theorem 3.1. We refer to $\hat{Q}$ as the RIPr, as it coincides with the original notion of the RIPr whenever the information divergence is finite. The remainder of Section 3.3 provides further discussion of this result, as well as an example showing that our extended notion of the RIPr can be well-defined whereas the RIPr was previously undefined. In the specific case that all initial measures are probability measures, both Li's original result and ours leave open

the possibility that $\hat{Q}$ may be a strict sub-probability measure, integrating to less than 1. In Sub-Section 3.3.1 we give a further example showing that this can indeed be the case, and we provide, via Theorem 3.9, a condition under which $\hat{Q}$ is guaranteed to be a standard (integrating to 1) probability measure. Sub-Section 3.3.2 then extends the greedy algorithm of Li and Barron (1999) and Brinda (2018) for approximating the RIPr in settings where $D(P\|\mathcal{C}) < \infty$ to settings where the information divergence might be infinite.

In Section 3.4 we turn to e-statistics. It contains our second central result, Theorem 3.13, which shows that whenever our extended notion of the RIPr $\hat{Q}$ exists, the likelihood ratio of P and $\hat{Q}$ is an optimal e-statistic according to the criterion of Definition 3.11, which can be seen as a strict generalization of GRO, the standard optimality criterion for e-statistics. As such, this result may be viewed as a generalization of a result of Grünwald et al. (2024) stated below as Theorem 3.3. After illustrating the result by an example, Sub-Section 3.4.2 provides another technical result, Theorem 3.16, which relates approximations in terms of information gain, to approximations in terms of e-statisticity: conditions are given under which a sequence $Q_1, Q_2, \ldots$ converging to $\hat{Q}$ in terms of information gain at a certain rate also satisfies that the likelihood ratio between P and $Q_1, Q_2, \ldots$ converges to an e-statistic, and tight bounds on the corresponding rates are given. After a discussion of related work, the chapter ends with a summary and ideas for future work in Section 3.5. All proofs are delegated to Appendix A.1. Appendix A.2 provides a general method for constructing RIPrs that are strict sub-probability measures. Finally, Appendix A.3 provides a discussion on the assumption of convexity that we will make throughout.

3.2 Background

3.2.1 Preliminaries

We work with a measurable space $(\Omega, \mathcal{F})$ and, unless specified otherwise, all measures will be defined on this space. Throughout, P will denote a finite measure and $\mathcal{C}$ a set of finite measures, such that P and all $Q \in \mathcal{C}$ have densities w.r.t. a common σ-finite measure μ. These densities will be denoted with lowercase, i.e. p and q respectively. We will assume throughout that $\mathcal{C}$ is convex, i.e. closed under finite mixtures. In Section 3.4.1 and in more detail in Appendix A.3 we discuss how our results would be affected if we were to adopt stronger notions of convexity like σ-convexity (closed under countable mixtures), or Choquet-convexity (closed under arbitrary mixtures).

Furthermore, we assume that there exists at least one $Q^* \in \mathcal{C}$ such that $P \ll Q^*$. This assumption is needed to ensure that $D(P\|Q \rightsquigarrow \mathcal{C})$ is a well-defined number in $[0, \infty]$ for any $Q \in \mathcal{C}$.

3.2.2 The Reverse Information Projection

As mentioned briefly above, the reverse information projection is the result of minimizing the information divergence between P and $\mathcal{C}$. If $\mathcal{C}$ is an exponential family, this problem is well understood (Csiszár and Matúš, 2003), but we focus here on the case that $\mathcal{C}$ is a general convex set. In this setting, the following theorem establishes existence and uniqueness of a limiting object for any sequence $(Q_n)_{n\in\mathbb{N}}$ in $\mathcal{C}$ such that $D(P\|Q_n) \to D(P\|\mathcal{C})$ whenever the latter is finite. This limit (i.e. $\hat{Q}$ in the following) is called the reverse information projection of P on $\mathcal{C}$.

Theorem 3.1 (Li (1999))**.** *If P and all $Q \in \mathcal{C}$ are probability measures s.t. $D(P\|\mathcal{C}) < \infty$, then there exists a unique (potentially sub-) probability measure $\hat{Q}$ such that:*

1. *We have that $\ln q_n \to \ln \hat{q}$ in $L_1(P)$ for all sequences $(Q_n)_{n\in\mathbb{N}}$ in $\mathcal{C}$ such that $\lim_{n\to\infty} D(P\|Q_n) = D(P\|\mathcal{C})$.*
2. *$\int_\Omega \ln \frac{\mathrm{d}P}{\mathrm{d}\hat{Q}} \,\mathrm{d}P = D(P\|\mathcal{C})$,*
3. *$\int_\Omega \frac{\mathrm{d}P}{\mathrm{d}\hat{Q}} \,\mathrm{d}Q \leq 1$ for all $Q \in \mathcal{C}$.*

3.2.3 E-Statistics and Growth Rate Optimality

The *e*-value has recently emerged as a popular alternative to the *p*-value for hypothesis testing (Ramdas et al., 2023; Henzi and Ziegel, 2022). Unlike the *p*-value, it is eminently suited for testing under optional continuation — and more generally, when the rule for stopping or continuing to analyze an additional batch of data is not under control of the data analyst, and may even be unknown or unknowable. It can be thought of as a measure of statistical evidence that is intimately linked with numerous ideas, such as likelihood ratios, test martingales (Ville, 1939) and tests of randomness (Levin, 1976). Formally, an *e*-value is defined as the value taken by an *e*-statistic, which is defined as a random variable $E : \Omega \to [0, \infty]$ that satisfies $\int_\Omega E \,\mathrm{d}Q \leq 1$ for all $Q \in \mathcal{C}$ (Vovk and Wang, 2021). The set of all *e*-statistics is denoted as $\mathcal{E}_\mathcal{C}$. Large *e*-values constitute evidence against $\mathcal{C}$ as null hypothesis, so that the null can be rejected when the computed *e*-value exceeds a certain threshold. For example, the test that rejects the null hypothesis when $E \geq 1/\alpha$ has a type-I error guarantee of α by

a simple application of Markov's inequality: $Q(E \geq 1/\alpha) \leq \alpha \int_\Omega E \, \mathrm{d}Q \leq \alpha$. For all further details, as well as an extensive introduction to the concept, and how it relates to optional stopping and continuation, we refer to Grünwald et al. (2024) and the overview paper by Ramdas et al. (2023).

In general, the set $\mathcal{E}_\mathcal{C}$ of e-statistics is quite large, and the above does not tell us *which* e-statistic to pick. This question was studied by Grünwald et al. (2024) and a log-optimality criterion coined GRO (*Growth-Rate Optimality*) was introduced for the case that the interest is in gaining as much evidence as possible relative to an alternative hypothesis given by a single probability measure P. GRO is a natural replacement of statistical power, which cannot meaningfully be used in an optional stopping/continuation context. This criterion can be traced back to the information-theoretic Kelly betting criterion by Kelly (1956) and is further discussed at length by Shafer (2021); Ramdas et al. (2023); Grünwald et al. (2024), to which we refer for more discussion.

Definition 3.2. If it exists, an e-statistic $\hat{E} \in \mathcal{E}_\mathcal{C}$ is Growth-Rate Optimal (GRO) if it achieves

$$\int_\Omega \ln \hat{E} \, \mathrm{d}P = \sup_{E \in \mathcal{E}_\mathcal{C}} \int_\Omega \ln E \, \mathrm{d}P.$$

The following theorem establishes a duality between GRO e-statistics and reverse information projections. For a limited set of testing problems, it states that GRO e-statistics exist and are uniquely given by likelihood ratios.

Theorem 3.3 (Grünwald et al. (2024), Theorem 1)**.** *If P and all $Q \in \mathcal{C}$ are probability measures such that $D(P\|\mathcal{C}) < \infty$, $p(\omega) > 0$ for all $\omega \in \Omega$, and $\hat{Q}$ is the RIPr of P on $\mathcal{C}$, then $\hat{E} = \frac{\mathrm{d}P}{\mathrm{d}\hat{Q}}$ is GRO with rate equal to $D(P\|\mathcal{C})$, i.e.*

$$\sup_{E \in \mathcal{E}_\mathcal{C}} \int_\Omega \ln E \, \mathrm{d}P = \int_\Omega \ln \hat{E} \, \mathrm{d}P = D(P\|\mathcal{C}).$$

Furthermore, for any GRO e-statistic $\tilde{E}$, we have that $\tilde{E} = \hat{E}$ holds P-almost surely.

3.3 The Reverse Information Projection

In this section, we state a result analogous to Theorem 3.1 in a more general setting. Rather than convergence of the logarithm of densities in $L_1(P)$, we consider convergence with respect to a different metric on the set of measurable positive functions,

i.e. $\mathcal{M}(\Omega, \mathbb{R}_{>0}) = \{f : \Omega \to \mathbb{R}_{>0} : f \text{ measurable}\}$. For $f, f' \in \mathcal{M}(\Omega, \mathbb{R}_{>0})$ we define

$$m_P^2(f, f') := \frac{1}{2} \int_\Omega \ln\left(\frac{\overline{f}}{f}\right) + \ln\left(\frac{\overline{f}}{f'}\right) \mathrm{d}P, \tag{3.3}$$

where $\overline{f} := {}^{(f+f')}/_2$. This is a divergence that can be thought of as the averaged Bregman divergence associated with the convex function $\gamma(x) = x - 1 - \ln(x)$. In particular, this means that for $Q, Q' \in \mathcal{C}$ such that $P \ll Q$ and $P \ll Q'$, we have that

$$m_P^2(q, q') = \frac{1}{2} D(P \| Q \rightsquigarrow \bar{Q}) + \frac{1}{2} D(P \| Q' \rightsquigarrow \bar{Q}). \tag{3.4}$$

Chen et al. (2008) study averaged Bregman divergences in detail for general γ, and they show that the function

$$m_\gamma^2(x, y) = \frac{1}{2}\gamma(x) + \frac{1}{2}\gamma(y) - \gamma\left(\frac{x+y}{2}\right)$$

is the square of a metric if and only if $\ln(\gamma''(x))'' \geq 0$. In our case, $\ln(\gamma''(x))'' = 2x^{-2}$, so this result holds. This can be used together with an application of the Minkowski inequality to show that the triangle inequality holds for the square root of the divergence (3.3), i.e. m_P, on $\mathcal{M}(\Omega, \mathbb{R}_{>0})$. It should also be clear that for $f, g \in \mathcal{M}(\Omega, \mathbb{R}_{>0})$ if $f = g$ everywhere, then $m_P(f, g) = 0$. Conversely $m_P(f, g) = 0$ only implies that $P(f \neq g) = 0$. This prevents us from calling m_P a metric on $\mathcal{M}(\Omega, \mathbb{R}_{>0})$, and we therefore define, analogous to $\mathcal{L}^p$ and L^p spaces, $M(\Omega, \mathbb{R}_{>0})$ as the set of equivalence classes of $\mathcal{M}(\Omega, \mathbb{R}_{>0})$ under the relation '$\sim$' given by $f \sim g \Leftrightarrow P(f \neq g) = 0$. By the discussion above, m_P properly defines a metric on $M(\Omega, \mathbb{R}_{>0})$. In the following we will often ignore this technicality and simply act as if m_P defines a metric on $\mathcal{M}(\Omega, \mathbb{R}_{>0})$, since we are not interested in what happens on null sets of P.

Considering convergence with respect to m_P will be useful for our analyses in the following. In particular, we will exploit on numerous occasions that m_P can be interpreted as a symmetrized version of the description gain, as described in Equation (3.4). However, other than mathematical convenience, there is no fundamental difference between considering convergence with respect to m_P and convergence of the logarithms in $L_1(P)$, as considered in Theorem 3.1. Indeed, Lemma A.2 in Appendix A.1 shows that the two types of convergence are equivalent. It is also this result from which the following proposition follows.

Proposition 3.4. *The metric space* $(M(\Omega, \mathbb{R}_{>0}), m_P)$ *is complete.*

Everything is now in place to state the main result.

Theorem 3.5. *If* $\inf_{Q\in\mathcal{C}} D(P\|Q \rightsquigarrow \mathcal{C}) < \infty$*, then there exists a measure* $\hat{Q}$ *that satisfies the following for every sequence* $(Q_n)_{n\in\mathbb{N}}$ *in* $\mathcal{C}$ *such that* $D(P\|Q_n \rightsquigarrow \mathcal{C}) \to \inf_{Q\in\mathcal{C}} D(P\|Q \rightsquigarrow \mathcal{C})$ *as* $n \to \infty$*:*

1. $q_n \to \hat{q}$ *in* m_P*.*
2. *If* P' *is a measure such that* $|\inf_{Q\in\mathcal{C}} D(P\|Q \rightsquigarrow P')| < \infty$*, then*

$$\int_\Omega \ln \frac{\mathrm{d}P'}{\mathrm{d}\hat{Q}} \,\mathrm{d}P = \lim_{n\to\infty} \int_\Omega \ln \frac{\mathrm{d}P'}{\mathrm{d}Q_n} \,\mathrm{d}P.$$

3. *For any* $Q \in \mathcal{C}$*,*

$$\int_\Omega \frac{\mathrm{d}P}{\mathrm{d}\hat{Q}} \,\mathrm{d}Q \leq P(\Omega) + Q(\Omega) - \liminf_{n\to\infty} Q_n(\Omega).$$

Theorem 3.1 is a special case of Theorem 3.5 when P and all $Q \in \mathcal{C}$ are probability measures such that $D(P\|\mathcal{C}) < \infty$. This follows because Equation (3.2) implies that minimizing $D(P\|Q \rightsquigarrow Q')$ over Q is equivalent to minimizing $D(P\|Q)$ and because convergence of the densities in m_P is equivalent to convergence of the logarithms in $L_1(P)$ by Lemma A.2 in Appendix A.1.1. We therefore refer to $\hat{Q}$ as the *reverse information projection* of P on $\mathcal{C}$, thereby extending the definition of the latter (we refrain from the term 'generalized RIPr', because it has already been used for the RIPr whenever it is not attained by an element of $\mathcal{C}$ (Csiszár and Matúš, 2003) or when the log score is replaced by another loss function (Grünwald and Mehta, 2020)). However, the density of the measure $\hat{Q}$ is only unique as an element of $M(\Omega, \mathbb{R}_{>0})$, since convergence of the densities holds in m_P. This causes no ambiguity here, so that we simply refer to it as 'the' RIPr.

Note that Theorem 3.5 implies that if there exists a $Q \in \mathcal{C}$ with $D(P\|Q \rightsquigarrow \mathcal{C}) = 0$, then Q is the RIPr of P on $\mathcal{C}$. This matches with the intuition that the maximum gain we can get from switching away from the 'best' code in $\mathcal{C}$ should be equal to zero. The following result establishes this more formally.

Proposition 3.6. *The following conditions are equivalent:*

1. *There exists a measure* P' *such that* $D(P\|P' \rightsquigarrow \mathcal{C})$ *is finite.*
2. *There exists a measure* Q *in* $\mathcal{C}$ *such that* $D(P\|Q \rightsquigarrow \mathcal{C})$ *is finite.*

3. *There exists a sequence of measures* $Q_n \in \mathcal{C}$ *such that* $D(P\|Q_n \rightsquigarrow \mathcal{C}) \to 0$ *for* $n \to \infty$.

Consequently, whenever $\inf_{Q\in\mathcal{C}} D(P\|Q \rightsquigarrow \mathcal{C}) < \infty$*, it must actually be equal to zero.*

To show that the reverse information projection exists, it is therefore enough to prove that one of these equivalent conditions holds. Which condition is easiest to check will depend on the specific setting, as exemplified by the following propositions.

Proposition 3.7. *Suppose that* $\mathcal{C}$ *is the convex hull of finitely many distributions, that is,* $\mathcal{C} = \operatorname{conv}(\{Q_1, \ldots, Q_n\})$*, then for any probability measure* P *with* $P \ll Q_i$ *for at least one* i*, it holds that* $D(P\|\frac{1}{n}\sum Q_i \rightsquigarrow \mathcal{C}) < \infty$.

Example 3.1. Let $\mathcal{C}$ be a singleton whose single element Q is given by the standard Gaussian and let P be the standard Cauchy distribution. Since the Cauchy distribution is exponentially heavier-tailed than the Gaussian, we have that $D(P\|\mathcal{C}) = \infty$. However, since both distributions have full support, it follows that

$$D(P\|Q \rightsquigarrow \mathcal{C}) = D(P\|Q \rightsquigarrow Q) = 0.$$

By Theorem 3.5 (1), Q is therefore the reverse information projection of P on $\mathcal{C}$.

This example can be extended to composite $\mathcal{C}$ by considering all mixtures of the Gaussian distributions $\mathcal{N}(-1, 1)$ and $\mathcal{N}(1, 1)$ with mean ± 1 and variance 1. Proposition 3.7 guarantees the existence of a reverse information projection although the information divergence is still infinite because a Cauchy distribution is more heavy tailed than any finite mixture of Gaussian distributions. Symmetry implies that the reverse information projection must be equal to the uniform mixture of $\mathcal{N}(-1, 1)$ and $\mathcal{N}(1, 1)$, which coincides with the result one would intuitively expect.

Proposition 3.8. *Assume that* $\mathcal{C}$ *is a convex set of probability measures that has finite minimax regret and with normalized maximum likelihood distribution* $Q^* \in \mathcal{C}$*. Then for any probability measure* P *that is absolutely continuous with respect to* Q^**, it holds that* $D(P\|Q^* \rightsquigarrow \mathcal{C}) < \infty$.

For an extensive discussion on minimax regret in the present coding context, as well as the normalized maximum likelihood distribution (also known as *Shtarkov* distribution), see e.g. Grünwald and Harremoës (2009); van Erven and Harremoës (2014). In short, the minimax regret is defined as $\inf_{Q\in\mathcal{C}} \sup_{Q'\in\mathcal{C},\omega\in\Omega} \ln {}^{q'(\omega)}/_{q(\omega)}$. This quantity is known to be finite if and only if the normalized maximum likelihood distribution Q^*, defined as $q^*(\omega) = {}^{\sup_{Q\in\mathcal{C}} q(\omega)}/\left(\int_\Omega \sup_{Q\in\mathcal{C}} q \,\mathrm{d}\mu\right)$, is well-defined. One-dimensional

exponential families with finite minimax regret have been classified by Grünwald and Harremoës (2009).

3.3.1 Strict Sub-Probability Measure

We return now to the familiar setting where P is a probability measure and $\mathcal{C}$ a convex set of probability measures. It is easy to verify that the RIPr $\hat{Q}$ of P on $\mathcal{C}$ is then a sub-probability measure. This follows because we know that there exists a sequence $(Q_n)_{n\in\mathbb{N}}$ in $\mathcal{C}$ such that q_n converges point-wise P-a.s. to $\hat{q}$ and Fatou's Lemma tells us

$$\int_\Omega \hat{q}\,\mathrm{d}\mu = \int_\Omega \liminf_{n\to\infty} q_n\,\mathrm{d}\mu \leq \liminf_{n\to\infty} \int_\Omega q_n\,\mathrm{d}\mu = 1. \tag{3.5}$$

It is not clear a priori whether this can ever be a strict inequality. For example, if the sample space is finite, the set of probability measures is compact, so the limit of any sequence of probability measures (i.e. the reverse information projection) will also be a probability measure. The following example illustrates that this is not always the case for infinite sample spaces, and it can in fact already go wrong for a countable sample space with $D(P\|\mathcal{C}) < \infty$.

Example 3.2. Let $\Omega = \mathbb{N}$ and $\mathcal{F} = 2^{\mathbb{N}}$. Furthermore, let P denote the probability measure δ_1 concentrated in the point $i = 1$ and $\mathcal{C}$ the set of distributions Q satisfying

$$\sum_{i=1}^{\infty} \frac{1}{i} q(i) = \frac{1}{2}.$$

This set is defined by a linear constraint, so that $\mathcal{C}$ is convex, and for any $Q \in \mathcal{C}$, we have

$$q(1) + \sum_{i=2}^{\infty} \frac{1}{i} q(i) = \sum_{i=1}^{\infty} \frac{1}{i} q(i) = \frac{1}{2},$$

implying that $q(1) \leq 1/2$. It follows that $D(P\|Q) = -\ln(q(1)) \geq \ln(2)$. The sequence $Q_n = \frac{n-2}{2n-2}\delta_1 + \frac{n}{2n-2}\delta_n$ satisfies $Q_n \in \mathcal{C}$ and

$$D(P\|Q_n) = \ln \frac{2n-2}{n-2} \to \ln(2).$$

Consequently, it must hold that $D(P\|\mathcal{C}) = \ln(2)$. The sequence Q_n converges to the strict sub-probability measure $(1/2)\delta_1$, which must therefore be the RIPr of P on $\mathcal{C}$.

A more general example, which can be seen as a template to create such situa-

tions, is given in Appendix A.2. The common theme is that $\mathcal{C}$ is defined using only constraints of the form $\sum_i f_1(i)q(i) = c$, where f_1 is some positive function such that $\lim_{n\to\infty} f_1(n) = 0$. Since $\mathcal{C}$ only contains probability measures, there is the additional constraint that $\sum_i f_0(i)q(i) = 1$, where f_0 denotes the constant function $f_0 \equiv 1$. This function f_0 dominates all other constraints f_1 in the sense that $\lim_{i\to\infty} {}^{f_1(i)}/_{f_0(i)} = 0$, but is itself not dominated by any of the constraints in the same manner. It turns out that this is the precise condition that dictates whether or not a constraint has to be respected when taking point-wise limits of elements in $\mathcal{C}$. Indeed, as shown in the theorem below, any constraint on $\mathcal{C}$ that is dominated by another constraint in the sense described above cannot be violated by taking point-wise limits. Therefore, if we add a restriction to $\mathcal{C}$ that dominates the constant function 1, i.e. that is defined by some function f_1 with $\lim_{n\to\infty} f_1(n) = \infty$, then the RIPr cannot be a strict sub-probability measure.

Theorem 3.9. *Take $\Omega = \mathbb{N}, \mathcal{F} = 2^{\mathbb{N}}$, and let $\mathcal{C}$ be a convex set of probability measures. Suppose that for $f_0, f_1 : \mathbb{N} \to \mathbb{R}_{>0}$, we have that $\sum_i f_0(i)q(i) \leq \lambda_0$ and $\sum_i f_1(i)q(i) = \lambda_1$ for all $Q \in \mathcal{C}$. If Q_n denotes a sequence of measures in $\mathcal{C}$ that converges point-wise to some distribution Q^*, and f_0 dominates f_1 in the sense that*

$$\lim_{i\to\infty} \frac{f_1(i)}{f_0(i)} = 0, \tag{3.6}$$

then

$$\sum_i f_1(i) \cdot q^*(i) = \lambda_1. \tag{3.7}$$

3.3.2 Greedy Approximation

So far, we have discussed the existence and properties of the RIPr of P on $\mathcal{C}$. However, there will be many situations where it is infeasible to compute this exact projection, as it requires solving a complex minimization problem. For example, if $\mathcal{C}$ is given by the convex hull of some parameterized family of distributions, the reverse information projection might be an arbitrary mixture of elements of this family, and the minimization problem need not be convex in the parameters of the family. To this end, Li and Barron (1999) propose an iterative greedy algorithm for the case that $\mathcal{C}$ is given by the σ-convex hull (all countable mixtures, see Appendix A.3) of a parameterized family of distributions, i.e. $\mathcal{C} = \sigma\text{-conv}(\{Q_\theta : \theta \in \Theta\})$, and $D(P\|\mathcal{C}) < \infty$. The algorithm starts by setting $Q_1 := Q_{\theta_1}$, where θ_1 minimizes $D(P\|Q_{\theta_1})$, and then

iteratively defining $Q_k := (1-\alpha_k)Q_{k-1} + \alpha_k Q_{\theta_k}$, where $\alpha_k = 2/(k+1)$[1] and θ_k is chosen to minimize $D(P\|Q_k)$. It is shown that, if $\sup_{x,\theta_1,\theta_2} \log q_{\theta_1}(x)/q_{\theta_2}(x)$ is bounded, then $D(P\|Q_k)$ converges to $D(P\|\mathcal{C})$ at rate $1/k$. Later, Brinda (2018) showed that the condition that the likelihood ratio has to be uniformly bounded in x can be relaxed to the condition that (3.8) below is finite. In both of these previous works, it is simply assumed that a minimizer in each step exists, though it need not necessarily be unique. We will do likewise in the following, where we give an adaptation of the algorithm that works when the KL divergence is infinite.

Algorithm 1 Greedy Approximation of the RIPr

Fix $Q^* \in \mathcal{C}$ s.t. $|\inf_{\theta\in\Theta} \int_\Omega \log q^*/q_\theta \,\mathrm{d}P| < \infty$
Let $Q_1 = Q_{\theta_1}$, where $\theta_1 = \arg\min_{\theta'\in\Theta} D(P\|Q_{\theta'} \rightsquigarrow Q^*)$
for $k = 2, 3, \ldots$ **do**
    Choose $\alpha_k = \frac{2}{k+1}$ and $\theta_k = \arg\min_{\theta'\in\Theta} D(P\|(1-\alpha_k)Q_{k-1} + \alpha_k Q_{\theta'} \rightsquigarrow Q^*)$
    Let $Q_k = (1-\alpha_k)Q_{\theta_{k-1}} + \alpha_k Q_{\theta_k}$
end for

Proposition 3.10. *Suppose that* $\inf_{Q\in\mathcal{C}} D(P\|Q \rightsquigarrow \mathcal{C}) < \infty$*, let* $(Q_k)_{k\in\mathbb{N}}$ *be the output of Algorithm 1, and let* Q *be any measure in* $\mathcal{C}$*, so that* $q = \sum_{\theta\in\Theta'} q_\theta \cdot w_Q(\theta)$ *for some probability mass function* w_Q *on a countable* $\Theta' \subset \Theta$*. If* $D(P\|Q' \rightsquigarrow Q'')$ *is finite for all* $Q', Q'' \in \mathcal{C}$*, then it holds that*

$$D(P\|Q_k \rightsquigarrow Q) \leq \frac{b_Q^{(k)}(P)}{k},$$

where $b_Q^{(k)}(P)$ *is given by*

$$\int_\Omega \left(1 + \sup_{\theta^*\in\{\theta_i\}_{i=1}^k} \log \frac{\sup_{\theta\in\Theta} q_\theta}{q_{\theta^*}}\right) \frac{\sum_{\theta\in\Theta'} q_\theta^2 \cdot w_Q(\theta)}{q^2} \,\mathrm{d}P \leq \sup_{Q\in\mathcal{C}} \int_\Omega \left(1 + \sup_{\theta^*,\theta\in\Theta} \log \frac{q_\theta}{q_{\theta^*}}\right) \frac{\sum_{\theta\in\Theta'} q_\theta^2 \cdot w_Q(\theta)}{q^2} \,\mathrm{d}P. \tag{3.8}$$

It follows that if $b_Q^{(k)}$ is uniformly bounded over all $Q \in \mathcal{C}$, in particular if (3.8) is finite, then $D(P\|Q_k \rightsquigarrow \mathcal{C})$ converges to zero, i.e. Q_k converges to the RIPr of P on $\mathcal{C}$, at rate $1/k$. The former holds under the strong, but often imposed assumption that the

[1] Li actually proposes to either minimize over α_k or use $\alpha_2 = 2/3$ and $\alpha_k = 2/k$ for $k > 2$; the formulation given here is a slight simplification by Brinda (2018).

likelihood ratios in $\mathcal{C}$ are uniformly bounded; for example when $\mathcal{C}$ is given by the σ-convex hull of Gaussian densities restricted to a cube (Li, 1999, Example 1). However, (3.8) might also be finite under weaker assumptions. For example, consider the set of Gaussian mixtures as in Example 3.1, that is, $\mathcal{C} = \{w \cdot \mathcal{N}(-1,1) + (1-w) \cdot \mathcal{N}(1,1) : w \in [0,1]\}$. It can be seen that $b_Q(P) < \infty$ for all $Q \in \mathcal{C}$ whenever P has a finite first moment. Moreover, if the latter holds, then $b_Q(P)$ is uniformly bounded over all $Q \in \mathcal{C}'$ where $\mathcal{C}' = \{w \cdot \mathcal{N}(-1,1) + (1-w) \cdot \mathcal{N}(1,1) : w \in [c, 1-c]\}$ for some $c \in (0, 1/2)$.

Whereas Proposition 3.10 is a satisfying theoretical result, we must concede that Algorithm 1 might not be the fastest to implement in practice. This arises from the fact that the objective $D(P\|(1-\alpha_k)Q_{k-1} + \alpha_k Q_{\theta'} \rightsquigarrow Q^*)$ need not be convex in θ'. One might therefore have to resort to an exhaustive search over a discretization of the parameter space. On top of that, there is no guarantee that the information gain is easily computable. As an alternative for the case that Θ is finite and $D(P\|\mathcal{C}) < \infty$, one might use the iterative algorithm proposed by Csiszár and Tusnády (1984, Theorem 5). A big advantage of the latter is that their recursive update step has an explicit formula, which makes each iteration considerably faster. The downside is that, whereas convergence in terms of KL is guaranteed, it is unclear at what rate this happens in general. Furthermore, proving convergence of their algorithm in the setting where $D(P\|\mathcal{C}) = \infty$ seems far from a straightforward exercise.

3.3.3 Discussion

The results in this section might be regarded as a generalization of large parts of Chapters 3 and 4 in Li's Ph.D. thesis (Li, 1999) and in fact the tools in this section were initially developed to clear up some ambiguity around the proof of Theorem 3.1, Part 1 as provided by Li. That is, Li states that for all sequences $(Q_n)_{n\in\mathbb{N}}$ in $\mathcal{C}$ such that $\lim_{n\to\infty} D(P\|Q_n) = D(P\|\mathcal{C})$ it holds that $\ln q_n \to \ln \hat{q}$ in $L_1(P)$. However, the proof thereof refers to his Lemma 4.3, which only shows existence of one such a sequence. Then, in Lemma 4.4, Li also shows that if $\hat{Q}$ is such that $\log q_n \to \log \hat{q}$ in $L_1(P)$ for some sequence $(Q_n)_{n\in\mathbb{N}}$ that achieves $\lim_{n\to\infty} D(P\|Q_n) = D(P\|\mathcal{C})$, then it must hold that $D(P\|\hat{Q}) = D(P\|\mathcal{C})$. However, it is a priori not clear whether every sequence $(Q_n)_{n\in\mathbb{N}}$ that achieves $\lim_{n\to\infty} D(P\|Q_n) = D(P\|\mathcal{C})$ has such a limit. Moreover, it is never shown that, if it exists, this limit must be the same for every such sequence. Note that it is not at all our intention here to criticize Li's fundamental and ground-breaking work. Li's is one of those rare theses that have had a major impact outside of their own research area: being a thesis on information-theory, it served

as the central tool and inspiration for papers on fast convergence rates in machine learning theory (van Erven et al., 2015; Grünwald and Mehta, 2020), and also for Grünwald et al. (2024), which led to a breakthrough in (e-based) hypothesis testing. Our aim is merely to indicate that Theorem 3.5 ties up some loose ends in Li's original, pioneering results.

3.4 Optimal E-Statistics

In this section, we assume that P and all $Q \in \mathcal{C}$ are probability measures, and we are interested in the hypothesis test with P as alternative and $\mathcal{C}$ as null. To this end, Theorem 3.5 shows that — whenever it exists — the likelihood ratio of P and its RIPr is an e-statistic. A natural question is whether the optimality of the RIPr in terms of describing data distributed according to P carries over to some sort of optimality of the e-statistic, as is true for the GRO criterion in the case that $D(P\|\mathcal{C}) < \infty$. It turns out that this is true in terms of an intuitive extension of the GRO criterion. Completely analogously to the coding story, we simply have to change from absolute to pairwise comparisons.

Definition 3.11. For e-statistics $E, E' \in \mathcal{E}_{\mathcal{C}}$, we say that E is *stronger* than E' if the following integral is well-defined and nonnegative, possibly infinite:

$$\int_{\Omega} \ln \left(\frac{E}{E'} \right) \mathrm{d}P, \tag{3.9}$$

where we adhere to the conventions $\ln({}^{0}\!/_{c}) = -\infty$ and $\ln({}^{c}\!/_{0}) = \infty$ for all $c \in \mathbb{R}_{>0}$. Furthermore, an e-statistic $E^* \in \mathcal{E}_{\mathcal{C}}$ is a *strongest* e-statistic if it is stronger than any other e-statistic $E \in \mathcal{E}_{\mathcal{C}}$.

The notion of optimality in Definition 3.11 comes down to the simple idea that if one e-statistic E is stronger than another e-statistic E', then *repeatedly* testing based on E eventually becomes more powerful than repeatedly testing based on E' in the sense that there is a higher probability of rejecting a false null-hypothesis. Let us explain in more detail what we mean by this. Suppose that we conduct the same experiment N times independently to test the veracity of the hypothesis $\mathcal{C}$, resulting in outcomes $\omega_1, \ldots, \omega_N$. For any given e-statistic $E \in \mathcal{E}_{\mathcal{C}}$, we have that $\prod_{i=1}^{N} E(\omega_i)$ is still an e-statistic, not just for fixed N but even if N is a random (i.e. data-dependent) stopping time. So, as indicated before, it can be used to test $\mathcal{C}$ with type-I error guarantees. Yet, for two e-statistics $E, E' \in \mathcal{E}_{\mathcal{C}}$, the law of large numbers states that

if P is true, it will almost surely hold that

$$\frac{\prod_{i=1}^{n} E(\omega_i)}{\prod_{i=1}^{n} E'(\omega_i)} = \exp\left(n \int_\Omega \ln\left(\frac{E}{E'}\right) \mathrm{d}P + o(n)\right).$$

It follows that if the integral $\int_\Omega \ln\left({}^{E}/_{E'}\right) \mathrm{d}P$ is positive then with high probability E will, for large enough n, give more evidence against $\mathcal{C}$ than E' if the alternative is true, i.e. a test based on E will asymptotically have more power than a test based on E'.

Since we assume throughout that there exists a $Q^* \in \mathcal{C}$ such that $P \ll Q^*$, it follows that for any e-statistic E we must have $P(E = \infty) = 0$, which simplifies any subsequent analyses greatly.

Proposition 3.12. *Assume that $\mathcal{C}$ is a set of probability measures and that P is a probability measure. If there is an $E' \in \mathcal{E}_\mathcal{C}$ such that $\sup_{E \in \mathcal{E}_\mathcal{C}} \int_\Omega \ln\left({}^{E}/_{E'}\right) \mathrm{d}P < \infty$, then a strongest e-statistics exists. Furthermore, if E_1 and E_2 are both strongest e-statistics then $E_1 = E_2$ holds P-a.s.*

The strongest e-statistic in Definition 3.11 can be seen as a generalization of the GRO e-statistic, because if $\int_\Omega \ln E \, \mathrm{d}P$ and $\int_\Omega \ln E' \, \mathrm{d}P$ are both finite, (3.9) can be written as the difference between the two logarithms, that is, $\int_\Omega \ln\left({}^{E}/_{E'}\right) \mathrm{d}P = \int_\Omega \ln E \, \mathrm{d}P - \int_\Omega \ln E' \, \mathrm{d}P$. In this case, finding the strongest e-statistic therefore corresponds to maximizing $\int_\Omega \ln E \, \mathrm{d}P$ over all e-statistics, thus recovering the original GRO criterion. As an extension of that case, we prove that whenever the RIPr exists, it always leads to the strongest e-statistic.

Theorem 3.13. *Suppose that both P and all $Q \in \mathcal{C}$ are probability measures and that $\inf_{Q \in \mathcal{C}} D(P \| Q \rightsquigarrow \mathcal{C}) < \infty$. If $\hat{Q}$ denotes the RIPr of P on $\mathcal{C}$, then $\hat{E} = {}^{\mathrm{d}P}/_{\mathrm{d}\hat{Q}}$ is the strongest e-statistic.*

The likelihood ratio between P and its RIPr is in fact the only e-statistic in the form of a likelihood ratio with P in the numerator, as the following proposition shows. Though the statement is more general, the proof is completely analogous to part of the proof of Lemma 4.1 by Li (1999).

Proposition 3.14. *Suppose that $\mathcal{C}$ is a set of probability measures and that P is a probability measure. If there exists a measure $Q^* \in \mathcal{C}$ such that ${}^{\mathrm{d}P}/_{\mathrm{d}Q^*} \in \mathcal{E}_\mathcal{C}$, then $D(P \| Q^* \rightsquigarrow \mathcal{C}) = 0$, i.e. Q^* is the RIPr of P on $\mathcal{C}$.*

We now return to Example 3.1, where the GRO criterion is not able to distinguish between e-variables, but we are able to do so with Definition 3.11 and Theorem 3.13.

Example 3.1 (continued)**.** In the case that P is the standard Cauchy and $\mathcal{C} = \{Q\}$, where Q is the standard Gaussian, it is straightforward to see that the likelihood ratio between P and Q is an e-statistic, i.e.

$$\int_\Omega \frac{\mathrm{d}P}{\mathrm{d}Q}\,\mathrm{d}Q = \int_\Omega \mathrm{d}P = 1.$$

However, for the growth rate it holds that

$$\int_\Omega \ln\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\,\mathrm{d}P = D(P\|Q) = \infty.$$

The same argument can be used to show that for any $0 < c \leq 1$, we have an e-statistic given by $c\,\mathrm{d}P/\mathrm{d}Q$, which still has infinite growth rate. The GRO criterion in Definition 3.2 is not able to tell which of these e-statistics is preferable. However, since Q is the RIPr of P on $\mathcal{C}$, it follows from Theorem 3.13 that $\mathrm{d}P/\mathrm{d}Q$ is the strongest e-statistic, and in particular it is stronger than $c\,\mathrm{d}P/\mathrm{d}Q$ for all $0 < c < 1$.

3.4.1 Convexity

In the discussion above, the null hypothesis $\mathcal{C}$ is assumed to be convex, which does not hold for many of the null hypotheses commonly employed in statistics, such as the set of all Gaussian distributions with varying mean and/or variance. However, it follows from the Fubini-Tonelli theorem that the set of e-statistics on $\mathcal{C}$ equals the set of e-statistics on the convex hull of $\mathcal{C}$. The same is true if the convex hull is replaced by the σ-convex hull where countable mixtures are allowed or by the Choquet-convex hull where arbitrary mixtures are allowed (see Appendix A.3 for precise definitions). Therefore, if there exists a strongest e-statistic for testing the alternative P against any of these notions of the convex hull, then that is also the strongest e-statistic for testing P against the original null hypothesis, regardless of whether that was convex. It follows from Theorem 3.13 that, to find the strongest e-statistic, it suffices to find the RIPr of P on any of the notions of the convex hull. In particular, if RIPrs exists on more than one of these, they must coincide; on the other hand, none of the three RIPrs may exist, and our results also do not rule out the possibility that the RIPr exists on just one or two of the three convex hulls. To witness, in Appendix A.3 we give an example (Example A.1) in which the RIPr of P on the σ-convex hull exists, whereas the RIPr of P on the convex hull does not. At the same time, there are constraints: Theorem A.8 in Appendix A.3 implies that if the RIPr on the convex hull of $\mathcal{C}$ exists, then the RIPr on the σ-convex hull of $\mathcal{C}$ also exists (and then they must be equal).

Things become much more clear-cut if the RIPr $\hat{Q}$ of P on a certain notion of the convex hull exists *and is an element of that set*. In that case, $\hat{Q}$ is also the RIPr of P on any stronger notion of the convex hull. Indeed, the different levels of convex hulls are nested, and their corresponding sets of e-statistics coincide, so this follows directly from Proposition 3.14:

Corollary 3.15. *Let $\mathcal{C}$ denote a set of probability measures (not necessarily convex) and let P denote a probability measure. If the RIPr of P on the convex hull of $\mathcal{C}$ exists and is given by $\hat{Q} \in \mathrm{conv}(\mathcal{C})$, then $\hat{Q}$ is also, (a) the RIPr of P on the σ-convex and, (b), on the Choquet-convex hull of $\mathcal{C}$. Similarly, if $\hat{Q} \in \sigma\text{-conv}(\mathcal{C})$ is the RIPr of P on $\sigma\text{-conv}(\mathcal{C})$, then (c) $\hat{Q}$ is also the RIPr of P on the Choquet-convex hull of $\mathcal{C}$.*

Further details regarding convexity are presented in Appendix A.3. In particular, Theorem A.8 in the latter gives an analogous result to Corollary 3.15, Part (a), for the case that P and $\mathcal{C}$ are not restricted to be probability measures, and the RIPr is not assumed to be attained in the set.

3.4.2 Approximation

In Section 3.3.2, we discussed an algorithm that provides an approximation of the RIPr for scenarios where it is not possible to explicitly compute the latter. However, the convergence guarantee given by Proposition 3.10 is in terms of the information gain. That is, if Q_k is the approximation of the projection after k iterations, then under suitable conditions it holds that $D(P\|Q_k \rightsquigarrow \mathcal{C}) \to 0$. This is not enough if we want to use such an approximation for hypothesis testing: we need that p/q_k gets closer and closer to being an e-statistic. The following theorem gives a condition under which this is true.

Theorem 3.16. *Assume $\inf_{Q\in\mathcal{C}} D(P\|Q \rightsquigarrow \mathcal{C}) < \infty$, fix $Q, Q' \in \mathcal{C}$, set $\delta := D(P\|Q \rightsquigarrow \mathcal{C})$ and suppose that there exists $\beta \in (0, \infty]$ such that $\|q'/q\|_{1+\beta} < \infty$. If $\beta \leq 1$ or $D(P\|Q' \rightsquigarrow \mathcal{C}) \leq K\delta$, then it holds that*

$$\int_\Omega \frac{p}{q}\,\mathrm{d}Q' = \int_\Omega \frac{q'}{q}\,\mathrm{d}P = 1 + O\left(C_\beta \cdot \delta^{\frac{\beta}{1+\beta}}\right) \text{ as } \delta \to 0, \tag{3.10}$$

where $C_\beta = \|q'/q\|_{1+\beta}$ if $\beta \leq 1$ and $C_\beta = K^{\frac{\beta-1}{2(1+\beta)}} \|q'/q\|_{1+\beta}$ otherwise.

Here, we use $\|f\|_p$ for $p \in (0, \infty]$ to denote the $\mathcal{L}^p(\Omega, P)$ norm of a function $f \in \mathcal{M}(\Omega, \mathbb{R}_{>0})$, i.e. $\left(\int_\Omega (f)^p \,\mathrm{d}P\right)^{1/p}$. Explicit values for the constants in (3.10) can be

found in the proof in the appendix. In particular, Theorem 3.16 implies the following: if there are $C, \delta_0 > 0$ such that $\|q'/q\|_2 \leq C$ for all $Q' \in \mathcal{C}$ and all $Q \in \mathcal{C}$ with $D(P\|Q \rightsquigarrow \mathcal{C}) \leq \delta_0$, then any sequence $Q_1, Q_2, \ldots$ with $D(P\|Q_k \rightsquigarrow \mathcal{C}) \to 0$ will have $\sup_{q' \in \mathcal{C}} \int_\Omega q'/q_k \, \mathrm{d}P = 1 + O(\delta_k^{1/2})$, where $\delta_k = D(P\|Q_k \rightsquigarrow \mathcal{C})$. This gives an easy to check condition for the convergence of p/q_k to an e-statistic. This square-root rate of convergence cannot be improved in general without an extra assumption, even if all likelihood ratios are bounded, i.e. $\|q'/q\|_\infty < \infty$. This can be seen by taking P and Q to be Bernoulli distributions with parameter $1/2$ and $1/2 + \epsilon$ respectively, $\mathcal{C}$ the set of Bernoulli distributions with parameters in $[1/4, 3/4]$ and Q' Bernoulli $1/4$. Then $\delta = D(P\|Q \rightsquigarrow \mathcal{C}) = 2\epsilon^2(1 + o(1))$ yet $\int_\Omega q'/q \, \mathrm{d}P = 1 + 4\epsilon(1 + o(1))$. But if likelihood ratios are bounded and we additionally consider Q' in a 'neighborhood' of Q (i.e. $D(P\|Q' \rightsquigarrow \mathcal{C}) \leq K\delta$), then a linear rate is possible as shown in Theorem 3.16 by letting β tend to infinity; the rate then interpolates between $\delta^{1/2}$ and δ depending on the largest β for which the $(1 + \beta)$-th moment exists. Furthermore the following example shows that in general bounds on the integrated likelihood ratios are necessary for the convergence to hold at all.

Example 3.3. Let $\mathcal{Q}$ represent the family of geometric distributions on $\Omega = \mathbb{N}_0$ and let $\mathcal{C} = \mathrm{conv}(\mathcal{Q})$. The elements of $\mathcal{Q}$ are denoted by Q_θ with density $q_\theta(n) = \theta^n(1-\theta)$, where $\theta \in [0, 1)$ denotes the probability of failure. For simplicity, assume that $P \in \mathcal{Q}$ so that the reverse information projection of P on $\mathcal{C}$ is equal to P. Take for example $P = Q_{1/2}$, then for any $\theta, \theta' \in [0, 1)$

$$\begin{aligned} \int_\Omega \frac{q_{\theta'}}{q_\theta} \, \mathrm{d}P &= \sum_{n=0}^{\infty} \left(\frac{1}{2}\frac{\theta'}{\theta}\right)^n \frac{1}{2}\frac{1-\theta'}{1-\theta} \\ &= \begin{cases} \frac{1}{1-\frac{\theta'}{2\theta}} \cdot \frac{1}{2}\frac{1-\theta'}{1-\theta}, & \text{if } \theta' < 2\theta; \\ \infty, & \text{otherwise;} \end{cases} \end{aligned} \tag{3.11}$$

whereas

$$\begin{aligned} D(P\|Q_\theta) &= \sum_{n=0}^{\infty} \left(\frac{1}{2}\right)^{n+1} (-n\log(2\theta) - \log 2(1-\theta)) \\ &= \log\frac{1/2}{1-\theta} + \log\frac{1/2}{\theta}, \end{aligned}$$

Now consider a sequence $1/3 < \theta_1 < \theta_2 < \theta_3 \ldots$ that converges to $1/2$. Then by the

above,

$$D(P\|Q_{\theta_i}) \to 0 = D(P\|\mathcal{C}).$$

We also see that for all i and all $\theta' \in [2\theta_i, 1)$, we have

$$\int_\Omega \frac{q_{\theta'}}{q_{\theta_i}} \mathrm{d}P = \infty,$$

i.e. for all i we have $\sup_{\theta' \in [0,1)} \int_\Omega {}^{q_{\theta'}}/_{q_{\theta_i}} \mathrm{d}P = \infty$.

3.4.3 Related Work

The results on the existence of optimal e-statistics displayed in this section bear similarities with work concurrently done by Zhang et al. (2024). In particular, they show that if $\mathcal{C}$ is a convex polytope, then there exists an e-statistic in the form of a likelihood ratio between two unspecified measures. Since a convex polytope contains the uniform mixture of its vertices, which can be shown to have finite information gain, this also follows from our Proposition 3.7. However, the techniques used to prove their results appear to be of a completely different nature than the ones used in this chapter, as they rely mostly on classical results in convex geometry together with results on optimal transport (and with these techniques, they provide various other results incomparable to ours).

In the case of compact alternative they furthermore discuss a property which they refer to as nontrivial e-power. That is, if the alternative is a convex polytope $\mathcal{A}$, then at least one of their e-statistics in the form of a likelihood ratio satisfies $\inf_{P\in\mathcal{A}} \int_\Omega \ln E \, \mathrm{d}P > 0$. We now show that the existence of such an e-statistic also follows from our results. In fact, if $\mathcal{A}$ is any convex set (not just a polytope) such that $\inf_{P\in\mathcal{A}} D(P\|\mathcal{C}) < \infty$, then (as Zhang et al. (2024) point out) a similar result is already implied by Grünwald et al. (2024) as long as the infimum is achieved on the left. Indeed, they show that the likelihood ratio of the distribution that achieves the infimum and its RIPr is an e-statistic that has nontrivial e-power. This leaves the case that $\inf_{P\in\mathcal{A}} D(P\|\mathcal{C}) = \infty$. Indeed, the current work implies that also in this case, an e-statistic with nontrivial (in fact, infinite) e-power exists, as long as $\mathcal{A}$ is a convex polytope. That is, if we use P^* to denote the uniform mixture of the vertices of $\mathcal{A}$, then for any vertex $P \in \mathcal{C}$, we have that

$$\int_\Omega \ln \frac{\mathrm{d}P^*}{\mathrm{d}\hat{Q}^*} \, \mathrm{d}P \geq \int_\Omega \ln \frac{\frac{1}{n}\mathrm{d}P}{\mathrm{d}\hat{Q}^*} \, \mathrm{d}P \geq D(P\|\mathcal{C}) - \ln(n),$$

where $\hat{Q}^*$ denotes the RIPr of P^*. It follows that $\inf_{P \in \mathcal{A}} \int_\Omega \ln \frac{\mathrm{d}P^*}{\mathrm{d}\hat{Q}^*} \, \mathrm{d}P = \infty$, so that the e-statistic given by the likelihood ratio of P^* to its RIPr has "nontrivial e-power". However, more work is needed to determine whether such constructions are in any way optimal and whether the restriction that $\mathcal{A}$ is a convex polytope can be relaxed.

Second, after the first version of this manuscript was made available online, a follow-up paper appeared by Larsson et al. (2024). They show that, under no conditions on P and $\mathcal{C}$ whatsoever, there exists an e-statistic E^* that is the strongest e-statistic in the sense of Definition 3.11. This e-statistic, which they call 'the numeraire', gives rise to a measure Q^* such that ${}^{\mathrm{d}Q^*}/_{\mathrm{d}P} = {}^{1}/_{E^*}$. Whenever the conditions of Theorem 3.5 hold, Q^* coincides with the reverse information projection of P on $\mathcal{C}$, so that it provides (in their words) "[...] a natural definition of the RIPr in the absence of any assumptions on $\mathcal{C}$ or P." We refer to their work (Larsson et al., 2024) for all further details.

3.5 Summary and Future Work

We have shown that, under very mild conditions, there exists a measure that achieves the minimax description gain over a convex set of measures $\mathcal{C}$ relative to a measure P. Whenever the information divergence between P and $\mathcal{C}$ is finite, this measure coincides with the reverse information projection of P on $\mathcal{C}$. As such, it provides a natural extension of the reverse information projection to cases where the the minimax description gain is finite, while the information divergence is infinite. In the context of hypothesis testing, this extended notion of the RIPr can be used to define an e-statistic for testing the simple alternative P against the composite null $\mathcal{C}$. This e-statistic is optimal in a sense that is a natural, but novel extension of the previously known GRO optimality criterion for e-statistics. We have shown an example where GRO is unable to differentiate between e-statistics, whereas our novel criterion can, so that it is a strict extension. Additionally, we discussed an algorithm that can be used to approximate the reverse information projection in scenarios where it is not explicitly computable and show under what circumstances this also leads to an approximation of the optimal e-statistic.

The results presented thus far suggest various avenues for further research of which we discuss two. First, Theorem 3.5 is formulated for general measures so one may ask for an interpretation of the RIPr in the case that P and $\mathcal{C}$ are not probability measures. If Ω is finite and λ is a measure on Ω, then we may define a probability measure $Po(\lambda)$ as the product measure $Po(\lambda) = \bigotimes_{\omega \in \Omega} Po\left(\lambda(\omega)\right)$, where $Po\left(\lambda(\omega)\right)$

denotes the Poisson distribution with mean $\lambda(\omega)$. With this definition we get

$$D(P\|Q \rightsquigarrow Q') = D(Po(P)\|Po(Q) \rightsquigarrow Po(Q')).$$

Furthermore, it can be shown that if the RIPr $\hat{Q}$ of P on $\mathcal{C}$ exists and is an element of $\mathcal{C}$, then $Po(\hat{Q})$ is also the RIPr of $Po(P)$ on the convex hull of $\mathcal{C}' := \{Po(Q)|Q \in \mathcal{C}\}$. Consequently, ${}^{Po(P)}/_{Po(\hat{Q})}$ can be thought of as an e-statistic for $\mathcal{C}'$. More work is needed to determine whether this interpretation has any applications and if it can be generalized to arbitrary Ω.

Second, even if $D(P\|\mathcal{C}) = \infty$, the Rényi divergence $D_\alpha(P\|Q)$ (see e.g. van Erven and Harremoës (2014)) may be a well-defined nonnegative real number for $\alpha \in (0, 1)$ and $Q \in \mathcal{C}$. These Rényi divergences are jointly convex in P and Q (van Erven and Harremoës, 2014) and for each $0 < \alpha < 1$ one may define a reverse Rényi projection $\hat{Q}_\alpha$ of P on $\mathcal{C}$ (Kumar and Sason, 2016). Larsson et al. (2024) show that one may use this projection to define an e-statistic that is optimal for a polynomial rather than a logarithmic utility function — the theory is developed in completely analogous fashion to the logarithmic/standard Kullback-Leibler information case. We conjecture that the projections $\hat{Q}_\alpha$ will converge to the RIPr for α tending to 1, which might lead to further applications.