

# Universal Reverse Information Projections and Optimal E-statistics

Tyron Lardy, Peter Grünwald and Peter Harremoës, *Member, IEEE*

## Abstract

Information projections have found important applications in probability theory, statistics, and related areas. In the field of hypothesis testing in particular, the reverse information projection (RIPr) has recently been shown to lead to so-called growth-rate optimal (GRO)  $e$ -statistics for testing simple alternatives against composite null hypotheses. However, the RIPr as well as the GRO criterion are undefined whenever the infimum information divergence between the null and alternative is infinite. We show that in such scenarios there often still exists an element in the alternative that is ‘closest’ to the null: the universal reverse information projection. The universal reverse information projection and its non-universal counterpart coincide whenever information divergence is finite. Furthermore, the universal 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 universal RIPr is a strict sub-probability distribution, as well as conditions under which an approximation of the universal RIPr leads to approximate  $e$ -statistics. For this case we provide tight relations between the corresponding approximation rates.

## Index Terms

Reverse Information Projections, Description Gain, Hypothesis Testing, E-variables.

## I. INTRODUCTION

WE write  $D(\nu\|\lambda)$  for the information divergence (Kullback-Leibler divergence, [1]–[3]) between two finite measures  $\nu$  and  $\lambda$  given by

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

For probability measures the interpretation of  $D(\nu\|\lambda)$  is that it measures how much we gain by coding according to  $\nu$  rather than coding according to  $\lambda$  if data are distributed according to  $\nu$ . 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 II-C) [4]–[8]. 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 [4]. 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  $D(P\|Q_n) \rightarrow \inf_{Q \in \mathcal{C}} D(P\|Q)$  converges to  $\hat{Q}$  in a particular norm [9], [10]. Li [9] 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 1 below). Grünwald et al. [4] showed (restated as Theorem 2 below) that it is even the optimal  $e$ -statistic for testing  $P$  against  $\mathcal{C}$ . However, it is clear that the RIPr does not exist 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 work 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* [11] given by

$$D(P\|Q \rightsquigarrow Q') = \int_{\Omega} \ln\left(\frac{dQ'}{dQ}\right) dP - (Q'(\Omega) - Q(\Omega)) \quad (1)$$

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

$$D(P\|Q \rightsquigarrow Q') = D(P\|Q) - D(P\|Q'). \quad (2)$$

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

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

T. Lardy and P. Grünwald are affiliated with CWI, Amsterdam, and Leiden University, The Netherlands. P. Harremoës is at Copenhagen Business College, Copenhagen, Denmark. CWI is the national research institute for mathematics and computer science in the Netherlands.

A five-page abstract of this paper, containing a subset of the theorems but no proofs, was presented at ISIT 2023, Taipei.

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

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 2 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.

### A. Contents and Overview

Below, in Section II, 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 III states Theorem 3 about the universal RIPr, 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

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

converges to  $\hat{Q}$  in a specific metric which we define. Thus, Theorem 3 may be viewed as a generalization of Li's result stated below as Theorem 1. We call  $\hat{Q}$  the universal RIPr, as it coincides with the RIPr whenever the information divergence is finite. The remainder of Section III provides further discussion of this result, as well as an example showing that the universal RIPr can be well-defined whereas the standard RIPr is not. 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 III-A we give a further example showing that this can indeed be the case, and we provide, via Theorem 4, a condition under which  $\hat{Q}$  is guaranteed to be a standard (integrating to 1) probability measure. Sub-Section III-B then extends the greedy algorithm of [12] and [13] for approximating the RIPr to the universal RIPr.

In Section IV we turn to  $e$ -statistics. It contains our second central result, Theorem 5, which shows that whenever the universal RIPr  $\hat{Q}$  exists, the likelihood ratio of  $P$  and  $\hat{Q}$  is an optimal  $e$ -statistic according to the criterion of Definition 2, 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 Grünwald et al. [4]'s result, stated below as Theorem 2. After illustrating the result by example, Sub-Section IV-A provides another technical result, Theorem 6, which relates approximations in terms of information gain, to approximations in terms of  $e$ -statisticity: conditions are given under which a sequence  $Q_1, Q_2, \dots$  converging to the universal RIPr  $\hat{Q}$  in terms of information gain at a certain rate also satisfies that the likelihood ratio between  $P$  and  $Q_1, Q_2, \dots$  converges to an  $e$ -statistic, and tight bounds on the corresponding rates are given. After a discussion of related work, the paper ends with a summary and ideas for future work in Section V. All proofs are delegated to Appendix A. Appendix B provides a general method for constructing RIPrs that are strict sub-probability measures.

## II. BACKGROUND

### A. 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  $\sigma$ -finite measure  $\mu$ . These densities will be denoted with lowercase, i.e.  $p$  and  $q$  respectively. We will assume throughout that  $\mathcal{C}$  is  $\sigma$ -convex, i.e. closed under countable mixtures, though we will refer to this simply as 'convex'. Furthermore, we assume that there exists at least one  $Q^* \in \mathcal{C}$  such that  $P \ll Q^*$ . On the one hand, this ensures that  $D(P\|Q \rightsquigarrow \mathcal{C})$  is a well-defined number in  $[0, \infty]$  for any  $Q \in \mathcal{C}$ . On the other hand, it aligns with our philosophy when we turn to hypothesis testing, in which case  $P$  and all  $Q \in \mathcal{C}$  will be probability distributions and serve as the alternative and null hypothesis respectively. We will consider  $P$  mostly as a tool to gather evidence against  $\mathcal{C}$ , so that it does not make sense to consider the case in which  $P$  puts mass on events that cannot occur according to the null, as the null hypothesis can be discredited in such scenarios regardless of how much mass  $P$  puts on the event.

### B. 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 [10], 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  $D(P\|Q_n) \rightarrow D(P\|\mathcal{C})$  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 1** (Li [9], Definition 4.2 and Theorem 4.3). *If  $P$  and all  $Q \in \mathcal{C}$  are probability distributions such that  $D(P\|\mathcal{C}) < \infty$ , then there exists a unique (potentially sub-) probability distribution  $\hat{Q}$  such that:*

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

### C. $e$ -statistics and Growth Rate Optimality

The  $e$ -value has recently emerged as a popular alternative to the  $p$ -value for hypothesis testing [5], [8], [14]. 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 [15] and tests of randomness [16]. Formally, an  $e$ -value is defined as the value taken by an  $e$ -statistic, which is defined as a random variable  $E : \Omega \rightarrow [0, \infty]$  that satisfies  $\int_{\Omega} E dQ \leq 1$  for all  $Q \in \mathcal{C}$  [6]. 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  $\alpha$  by a simple application of Markov's inequality:  $Q(E \geq 1/\alpha) \leq \alpha \int_{\Omega} E dQ \leq \alpha$ . 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 [4] and the overview paper [5].

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 in [4] 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 distribution  $P$ . 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 in [17] and is further discussed at length by [4], [5], [7], to which we refer for more discussion.

**Definition 1.** If it exists, an  $e$ -statistic  $\hat{E} \in \mathcal{E}_{\mathcal{C}}$  is Growth-Rate Optimal (GRO) if it achieves

$$\int_{\Omega} \ln \hat{E} dP = \sup_{E \in \mathcal{E}_{\mathcal{C}}} \int_{\Omega} \ln E dP.$$

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 2** (Grünwald et al. [4], Theorem 1). *If  $P$  and all  $Q \in \mathcal{C}$  are probability distributions 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  $\hat{E} = \frac{dP}{d\hat{Q}}$  is GRO with rate equal to  $D(P||\mathcal{C})$ , i.e.*

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

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

### III. THE UNIVERSAL REVERSE INFORMATION PROJECTION

In this section, we state a result analogous to Theorem 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 metric on the set of measurable positive functions, i.e.  $\mathcal{M}(\Omega, \mathbb{R}_{>0}) = \{f : \Omega \rightarrow \mathbb{R}_{>0} : f \text{ measurable}\}$ . For  $f, f' \in \mathcal{M}(\Omega, \mathbb{R}_{>0})$  we define

$$m_P^2(f, f') := \frac{1}{2} \int_{\Omega} \ln \left( \frac{\bar{f}}{f} \right) + \ln \left( \frac{\bar{f}}{f'} \right) dP, \quad (3)$$

where  $\bar{f} := (f+f')/2$ . 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 [18], such divergences are studied in detail for general  $\gamma$ . In particular, 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), 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  $P$ .

**Proposition 1.** *The metric space  $(M(\Omega, \mathbb{R}_{>0}), m_P)$  is complete.*

Everything is now in place to state the main result.

**Theorem 3.** *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  $\mathcal{C}$  s.t.  $D(P \| Q_n \rightsquigarrow \mathcal{C}) \rightarrow \inf_{Q \in \mathcal{C}} D(P \| Q \rightsquigarrow \mathcal{C})$  as  $n \rightarrow \infty$ :*

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

$$\int_{\Omega} \ln \frac{dP'}{d\hat{Q}} dP = \lim_{n \rightarrow \infty} \int_{\Omega} \ln \frac{dP'}{dQ_n} dP.$$

- 3) For any  $Q \in \mathcal{C}$ ,

$$\int_{\Omega} \frac{dP}{dQ} dQ \leq P(\Omega) + Q(\Omega) - \liminf_{n \rightarrow \infty} Q_n(\Omega).$$

Theorem 1 is a special case of Theorem 3 when  $P$  and all  $Q \in \mathcal{C}$  are probability distributions such that  $D(P \| \mathcal{C}) < \infty$ . This follows because Equation (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  $m_P$  implies convergence of the logarithms in  $L_1(P)$  by Lemma 2 in Appendix A-A. The measure  $\hat{Q}$  as in Theorem 3 therefore extends the notion of the reverse information projection of  $P$  on  $\mathcal{C}$ . We call  $\hat{Q}$  the universal reverse information projection of  $P$  on  $\mathcal{C}$  ('generalized' has already been used for the RIPr whenever it is not attained by an element of  $\mathcal{C}$  [10] or when the log score is replaced by another loss function [19]). 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  $m_P$ . In the current work this causes no ambiguity, so that we simply refer to it as 'the' universal RIPr.

Note that Theorem 3 implies that if there exists a  $Q \in \mathcal{C}$  with  $D(P \| Q \rightsquigarrow \mathcal{C}) = 0$ , then  $Q$  is the universal 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, i.e. whenever  $\inf_{Q \in \mathcal{C}} D(P \| Q \rightsquigarrow \mathcal{C}) < \infty$ , it must actually be equal to zero.

**Proposition 2.** *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  $D(P \| Q_n \rightsquigarrow \mathcal{C}) \rightarrow 0$  for  $n \rightarrow \infty$ .*

To show that the universal 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.** *If  $\mathcal{C}$  is the convex hull of finitely many distributions, i.e.  $\mathcal{C} = \text{conv}(\{Q_1, \dots, Q_n\})$ , 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 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 tails of the Cauchy distribution are exponentially heavier-tailed than those of the Gaussian, we have that  $D(P \| \mathcal{C}) = \infty$ . However, since both distributions have full support, it follows that

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

By Theorem 3 (1),  $Q$  is therefore the universal 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  $\pm 1$  and variance 1. Proposition 3 guarantees the existence of a universal reversed 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 universal reversed 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 4.** *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 a definition of minimax regret in the present coding context, as well as the normalized maximum likelihood distribution (also known as *Shtarkov* distribution) that, if the minimax regret is finite, achieves it, see e.g. [20] or [21]. One-dimensional exponential families with finite minimax regret have been classified in [20].

#### A. Strict sub-probability measure

We return now to the familiar setting where  $P$  is a probability distribution and  $\mathcal{C}$  a convex set of probability distributions. 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

$$\int_{\Omega} \hat{q} d\mu = \int_{\Omega} \liminf_{n \rightarrow \infty} q_n d\mu \leq \liminf_{n \rightarrow \infty} \int_{\Omega} q_n d\mu = 1. \quad (4)$$

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 2.** Let  $\Omega = \mathbb{N}$  and  $\mathcal{F} = 2^{\mathbb{N}}$ . Furthermore, let  $P$  denote the probability measure  $\delta_1$  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

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

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 (universal) RIPr of  $P$  on  $\mathcal{C}$ .

A more general example, which can be seen as a template to create such situations, is given in Appendix B. The common theme is that  $\mathcal{C}$  is defined using only constraints of the form  $\sum g(i)q(i) = c$ , where  $g$  is some positive function such that  $\lim_{n \rightarrow \infty} g(n) = 0$ . Since  $\mathcal{C}$  only contains probability measures, there is the additional constraint that  $\sum q(i)f(i) = 1$ , where  $f$  denotes the constant function  $f \equiv 1$ . This function  $f$  dominates all other constraints  $g$  in the sense that  $\lim_{i \rightarrow \infty} g^{(i)}/f(i) = 0$ , but is itself not dominated by any of the constraints in the same manner. As shown in the theorem below, any constraint on  $\mathcal{C}$  that is dominated by another in this way cannot be violated by taking the point-wise limit of elements of  $\mathcal{C}$ . Therefore, if we add a restriction to  $\mathcal{C}$  that dominates the constant function 1, i.e. that is defined by some function  $g$  with  $\lim_{n \rightarrow \infty} g(n) = \infty$ , then the RIPr cannot be a strict sub-probability measure.

**Theorem 4.** Take  $\Omega = \mathbb{N}$ ,  $\mathcal{F} = 2^{\mathbb{N}}$ , and let  $\mathcal{C}$  be a convex set of probability distributions. Suppose that for  $f_0, f_1 : \mathbb{N} \rightarrow \mathbb{R}_{>0}$ , 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

$$\lim_{i \rightarrow \infty} \frac{f_1(i)}{f_0(i)} = 0, \quad (5)$$

then

$$\sum_i f_1(i) \cdot q^*(i) = \lambda_1. \quad (6)$$

## B. Greedy Approximation

So far, we have discussed the existence and properties of the universal 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 universal 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 [12] propose an iterative greedy algorithm for the case that  $\mathcal{C}$  is given by the  $\sigma$ -convex hull of a parameterized family of distributions, i.e.  $\sigma\text{-conv}(\{Q_\theta : \theta \in \Theta\})$ . The algorithm starts by setting  $Q_1 := Q_{\theta_1}$ , where  $\theta_1$  minimizes  $D(P\|Q_{\theta_1})$ , and then iteratively defining  $Q_k := (1 - \alpha_k)Q_{k-1} + \alpha_k Q_{\theta_k}$ , where  $\alpha_k = 2/(k+1)^*$  and  $\theta_k$  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 [13] showed that the condition that the likelihood ratio has to be uniformly bounded in  $x$  can be relaxed to the condition that (7) 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.

\*Li actually proposes to either minimize over  $\alpha_k$  or use  $\alpha_2 = 2/3$  and  $\alpha_k = 2/k$  for  $k > 2$ ; the formulation given here is a slight simplification by Brinda [13].

---

**Algorithm 1** Greedy Approximation of the Universal RIPr
 

---

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

**Proposition 5.** 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

$$\begin{aligned} & \int_{\Omega} \left( 1 + \sup_{\theta^* \in \{\theta_i\}_{i=1}^k} \log \frac{\sup_{\theta \in \Theta} q_{\theta}}{q_{\theta^*}} \right) \frac{\sum_{\theta \in \Theta'} q_{\theta}^2 \cdot w_Q(\theta)}{q^2} dP \leq \\ & \sup_{Q \in \mathcal{C}} \int_{\Omega} \left( 1 + \sup_{\theta^*, \theta \in \Theta} \log \frac{q_{\theta}}{q_{\theta^*}} \right) \frac{\sum_{\theta \in \Theta'} q_{\theta}^2 \cdot w_Q(\theta)}{q^2} dP. \end{aligned} \quad (7)$$

It follows that if  $b_Q^{(k)}$  is uniformly bounded over all  $Q \in \mathcal{C}$ , in particular if (7) is finite, then  $D(P||Q_k \rightsquigarrow \mathcal{C})$  converges to zero, i.e.  $Q_k$  converges to the universal RIPr of  $P$  on  $\mathcal{C}$ , at rate  $1/k$ . While this 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  $\theta'$ . 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  $\Theta$  is finite and  $D(P||\mathcal{C}) < \infty$ , one might use the iterative algorithm proposed by Csiszár and Tusnády [22, 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, while 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.

### C. 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 [9] and in fact the tools in this section were initially developed to clear up some ambiguity around the proof of Theorem 1, Part 1 as provided by Li. That is, in [9] it is stated that for all sequences  $(Q_n)_{n \in \mathbb{N}}$  in  $\mathcal{C}$  such that  $\lim_{n \rightarrow \infty} D(P||Q_n) = D(P||\mathcal{C})$  it holds that  $\ln q_n \rightarrow \ln \hat{q}$  in  $L_1(P)$ . However, the proof thereof refers to [9, Lemma 4.3], which only shows existence of one such a sequence. Then [9, Lemma 4.4] also shows that if  $\hat{Q}$  is such that  $\log q_n \rightarrow \log \hat{q}$  in  $L_1(P)$  for some sequence  $(Q_n)_{n \in \mathbb{N}}$  that achieves  $\lim_{n \rightarrow \infty} D(P||Q_n) = D(P||\mathcal{C})$ , 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  $\lim_{n \rightarrow \infty} D(P||Q_n) = D(P||\mathcal{C})$  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 [19], [23], and also for [4], which led to a breakthrough in ( $e$ -based) hypothesis testing. Our aim is merely to indicate that Theorem 3 ties up some loose ends in Li's original, pioneering results.

## IV. OPTIMAL $E$ -STATISTICS

In this section, we assume that  $P$  and all  $Q \in \mathcal{C}$  are probability distributions, and we are interested in the hypothesis test with  $P$  as alternative and  $\mathcal{C}$  as null. To this end, Theorem 3 shows that — if it exists — the likelihood ratio of  $P$  and its universal RIPr is an  $e$ -statistic. A natural question is whether the optimality of the universal RIPr in terms of describing data distributed according to  $P$  carries over to some sort of optimality of the  $e$ -statistic, similar as 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 2.** 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 non-negative, possibly infinite:

$$\int_{\Omega} \ln \left( \frac{E}{E'} \right) dP, \quad (8)$$

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 the *strongest*  $e$ -statistic if it is stronger than any other  $e$ -statistic  $E \in \mathcal{E}_{\mathcal{C}}$ .

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. The optimality criterion in Definition 2 can be seen as a generalization of GRO, because if  $\int_{\Omega} \ln E \, dP$  and  $\int_{\Omega} \ln E' \, dP$  are both finite, (8) can be written as the difference between the two logarithms, thus recovering the original GRO criterion. Analogously to that case, we prove that whenever the universal RIPr exists, it leads to an optimal  $e$ -statistic.

**Theorem 5.** *Suppose that both  $P$  and all  $Q \in \mathcal{C}$  are probability distributions such that  $\inf_{Q \in \mathcal{C}} D(P||Q \rightsquigarrow \mathcal{C}) < \infty$ . If  $\hat{Q}$  denotes the universal RIPr of  $P$  on  $\mathcal{C}$ , then  $\hat{E} = {}^{dP/d\hat{Q}}$  is the strongest  $e$ -statistic. Furthermore, for any other strongest  $e$ -statistic  $\tilde{E}$  we must have that  $\tilde{E} = \hat{E}$  holds  $P$ -a.s.*

The likelihood ratio between  $P$  and its universal 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 in [9].

**Proposition 6.** *Suppose that  $Q^* \in \mathcal{C}$  such that  ${}^{dP/dQ^*} \in \mathcal{E}_{\mathcal{C}}$ , then  $D(P||Q^* \rightsquigarrow \mathcal{C}) = 0$ , i.e.  $Q^*$  is the universal RIPr of  $P$  on  $\mathcal{C}$ .*

The notion of optimality in Definition 2 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  $\omega_1, \dots, \omega_N$ . 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

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

It follows that if the integral  $\int_{\Omega} \ln\left(\frac{E}{E'}\right) dP$  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  $E'$ . We now return to Example 1, where the GRO criterion is not able to distinguish between  $e$ -variables, but we are able to do so with Definition 2 and Theorem 5.

**Example 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.

$$\int_{\Omega} \frac{dP}{dQ} dQ = \int_{\Omega} dP = 1.$$

However, for the growth rate it holds that

$$\int_{\Omega} \ln\left(\frac{dP}{dQ}\right) dP = D(P||Q) = \infty.$$

The same argument can be used to show that for any  $0 < c \leq 1$ , we have an  $e$ -statistic given by  $cdP/dQ$ , which still has infinite growth rate. The GRO criterion in Definition 1 is not able to tell which of these  $e$ -statistics is preferable. However, since  $Q$  is the universal RIPr of  $P$  on  $\mathcal{C}$ , it follows from Theorem 5 that  ${}^{dP/dQ}$  is the strongest  $e$ -statistic, and in particular stronger than  $cdP/dQ$  for all  $0 < c < 1$ .

### A. Approximation

In Section III-B, we discussed an algorithm that provides an approximation of the universal RIPr for scenarios where it is not possible to explicitly compute the latter. However, the convergence guarantee given by Proposition 5 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  $D(P||Q_k \rightsquigarrow \mathcal{C}) \rightarrow 0$ . 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. For  $p \in (0, \infty]$  we use  $\|f\|_p$  to denote the  $\mathcal{L}^p(\Omega, P)$  norm of a function  $f \in \mathcal{M}(\Omega, \mathbb{R}_{>0})$ , i.e.  $(\int_{\Omega} (f)^p dP)^{1/p}$ .

**Theorem 6.** *Assume that  $\inf_{Q \in \mathcal{C}} D(P||Q \rightsquigarrow \mathcal{C}) < \infty$ , 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*

$$\int_{\Omega} \frac{p}{q} dQ' = \int_{\Omega} \frac{q'}{q} dP = 1 + O\left(C_{\beta} \cdot \delta^{\frac{\beta}{1+\beta}}\right) \text{ as } \delta \rightarrow 0, \quad (9)$$

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.

Explicit values for the constants in (9) can be found in the proof in the appendix. In particular, Theorem 6 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  $D(P\|Q) \leq \delta_0$ , then any sequence  $Q_1, Q_2, \dots$  with  $D(P\|Q_k) \rightarrow 0$  will have  $\sup_{q' \in \mathcal{C}} \int_{\Omega} q'/q_k dP = 1 + O(\delta_k^{1/2})$ , where  $\delta_k = D(P\|Q_k)$ . 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  $\delta = D(P\|Q) = 2\epsilon^2(1 + o(1))$  yet  $\int_{\Omega} q'/q dP = 1 + 4\epsilon(1 + o(1))$ . But if likelihood ratios are bounded and we additionally consider  $Q'$  in a ‘neighborhood’ of  $Q$  (i.e.  $D(P\|Q') \leq K\delta$ ), then a linear rate is possible as shown in Theorem 6 by letting  $\beta$  tend to infinity; the rate then interpolates between  $\delta^{1/2}$  and  $\delta$  depending on the largest  $\beta$  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.** Let  $\mathcal{Q}$  represent the family of geometric distributions on  $\Omega = \mathbb{N}_0$  and let  $\mathcal{C} = \text{conv}(\mathcal{Q})$ . The elements of  $\mathcal{Q}$  are denoted by  $Q_\theta$  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  $\theta, \theta' \in [0, 1)$

$$\int_{\Omega} \frac{q_{\theta'}}{q_{\theta}} dP = \sum_{n=0}^{\infty} \left(\frac{1-\theta'}{2-\theta}\right)^n \frac{1-\theta'}{2-\theta} = \begin{cases} \frac{1-\theta'}{1-\frac{\theta'}{2}} \cdot \frac{1}{2} \frac{1-\theta'}{1-\theta}, & \text{if } \theta' < 2\theta; \\ \infty, & \text{otherwise;} \end{cases} \quad (10)$$

whereas

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

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

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

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

$$\int_{\Omega} \frac{q_{\theta'}}{q_{\theta_i}} dP = \infty,$$

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

## B. 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. [24]. 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. However, the techniques used to prove their results appear to be of a completely different nature than the ones used in this paper, 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, and which can be seen as an analogue to the classical property of ‘unbiasedness’ for tests based on  $p$ -values. 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  $\inf_{P \in \mathcal{A}} \int_{\Omega} \ln E dP > 0$ . 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  $\inf_{P \in \mathcal{A}} D(P\|C) < \infty$ , then (as [24] point out) a similar result is already implied by Grünwald et al. [4] 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  $\inf_{P \in \mathcal{A}} D(P\|C) = \infty$ . 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

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

where  $\hat{Q}^*$  denotes the universal RIPr of  $P^*$ . It follows that  $\inf_{P \in \mathcal{A}} \int_{\Omega} \ln \frac{dP^*}{d\hat{Q}^*} dP = \infty$ , so that the  $e$ -statistic given by the likelihood ratio of  $P^*$  to its universal 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.



## V. 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  $P$ . This measure coincides with the reverse information projection whenever the information divergence between  $P$  and  $\mathcal{C}$  is finite, so we refer to it as the universal reverse information projection. In the context of hypothesis testing, the universal 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, while our novel criterion can, so that it is a strict extension. Additionally, we discussed an algorithm that can be used to approximate the universal 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 is formulated for general measures so one may ask for an interpretation of the universal RIPr in the case that  $P$  and  $\mathcal{C}$  are not probability measures. If  $\Omega$  is finite and  $\lambda$  is a measure on  $\Omega$ , then we may define a probability measure  $Po(\lambda)$  as the product measure  $Po(\lambda) = \otimes_{\omega \in \Omega} Po(\lambda(\omega))$ , where  $Po(\lambda(\omega))$  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 universal RIPr  $\hat{Q}$  of  $P$  on  $\mathcal{C}$  exists and is an element of  $\mathcal{C}$ , then  $Po(\hat{Q})$  is also the universal 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  $\Omega$ .

Second, even if  $D(P\|\mathcal{C}) = \infty$ , the Rényi divergence  $D_\alpha(P\|Q)$  (see e.g. [21]) may be a well-defined non-negative real number for  $\alpha \in (0, 1)$  and  $Q \in \mathcal{C}$ . These Rényi divergences are jointly convex in  $P$  and  $Q$  [21] and for each  $0 < \alpha < 1$  one may define a reversed Rényi projection  $\hat{Q}_\alpha$  of  $P$  on  $\mathcal{C}$  [25]. If it exists, it can be shown that this distribution will satisfy

$$\int_{\Omega} \left( \frac{dP}{d\hat{Q}_\alpha} \right)^\alpha dQ \leq 1$$

for all  $Q \in \mathcal{C}$ , i.e.  $(dP/d\hat{Q}_\alpha)^\alpha$  is an  $e$ -statistic. We conjecture that the projections  $\hat{Q}_\alpha$  will converge to the universal RIPr for  $\alpha$  tending to 1, which might lead to further applications.

## REFERENCES

- [1] S. Kullback and R. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.
- [2] I. Csiszár, "Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der ergodizität von Markoffschen Ketten," *Publ. Math. Inst. Hungar. Acad.*, vol. 8, pp. 95–108, 1963.
- [3] F. Liese and I. Vajda, *Convex Statistical Distances*. Leipzig: Teubner, 1987.
- [4] P. Grünwald, R. de Heide, and W. Koolen, "Safe testing," *J. Roy. Stat. Soc.*, 2023, to appear. [Online]. Available: <https://arxiv.org/abs/1906.07801>
- [5] A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer, "Game-theoretic statistics and safe anytime-valid inference," 2022. [Online]. Available: <https://arxiv.org/abs/2210.01948>
- [6] V. Vovk and R. Wang, "E-values: Calibration, combination and applications," *Ann. Stat.*, vol. 49, no. 3, pp. 1736–1754, 2021.
- [7] G. Shafer *et al.*, "Testing by betting: A strategy for statistical and scientific communication," *J. Roy. Stat. Soc.: Series A (Stat. in Soc.)*, vol. 184, no. 2, pp. 407–431, 2021.
- [8] A. Henzi and J. F. Ziegel, "Valid sequential inference on probability forecast performance," *Biometrika*, vol. 109, no. 3, pp. 647–663, 2022.
- [9] J. Li, "Estimation of mixture models," Ph.D. dissertation, Yale University, New Haven, CT, 1999.
- [10] I. Csiszár and F. Matúš, "Information projections revisited," *IEEE Trans. Inform. Theory*, vol. 49, no. 6, pp. 1474–1490, 2003.
- [11] F. Topsøe, "Information theory at the service of science," in *Entropy, Search, Complexity*, ser. Bolyai Society Mathematical Studies, I. Csiszár, G. O. H. Katona, and G. Tardos, Eds. János Bolyai Mathematical Society and Springer-Verlag, 2007, vol. 16, pp. 179–207. [Online]. Available: <http://www.math.ku.dk/~topsoe/aspects.pdf>
- [12] J. Li and A. Barron, "Mixture density estimation," *Advances in neural information processing systems*, vol. 12, 1999.
- [13] W. D. Brinda, "Adaptive estimation with Gaussian radial basis mixtures," Ph.D. dissertation, Yale University, 2018.
- [14] P. Grünwald, A. Henzi, and T. Lardy, "Anytime valid tests of conditional independence under model-X," *J. Am. Stat. Assoc.*, pp. 1–12, 2023. [Online]. Available: <https://doi.org/10.1080/01621459.2023.2205607>
- [15] J. Ville, "Étude critique de la notion de collectif," *Bull. Amer. Math. Soc.*, vol. 45, no. 11, p. 824, 1939.
- [16] L. A. Levin, "Uniform tests of randomness," in *Doklady Akademii Nauk*, vol. 227, no. 1. Russian Academy of Sciences, 1976, pp. 33–35.
- [17] J. L. Kelly, "A new interpretation of information rate," *IRE Trans. Inf. Theory*, vol. 2, pp. 185–189, 1956.
- [18] P. Chen, Y. Chen, and M. Rao, "Metrics defined by Bregman divergences," *Commun. Math. Sci.*, vol. 6, no. 4, pp. 915–926, 2008. [Online]. Available: <http://projecteuclid.org/euclid.cms/1229619676>
- [19] P. D. Grünwald and N. A. Mehta, "Fast rates for general unbounded loss functions: from ERM to generalized Bayes," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 2040–2119, 2020.
- [20] P. Grünwald and P. Harremoës, "Finiteness of redundancy, regret, Shtarkov sums, and Jeffreys integrals in exponential families," in *Proceedings for the International Symposium for Information Theory, Seoul, 2009*. IEEE, June 2009, pp. 714–718. [Online]. Available: <http://www.harremoes.dk/Peter/submit2009f.pdf>
- [21] T. van Erven and P. Harremoës, "Rényi divergence and Kullback-Leibler divergence," *IEEE Trans. Inform. Theory*, vol. 60, no. 7, pp. 3797–3820, 2014. [Online]. Available: <http://arxiv.org/pdf/1206.2459v1.pdf>
- [22] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," *Statistics and Decisions*, vol. Supplementary Issue 1, pp. 205–237, 1984.
- [23] T. van Erven, P. D. Grünwald, N. A. Mehta, M. D. Reid, and R. C. Williamson, "Fast rates in statistical and online learning," *Journal of Machine Learning Research*, vol. 16, pp. 1793–1861, 2015.
- [24] Z. Zhang, A. Ramdas, and R. Wang, "When do exact and powerful p-values and e-values exist?" 2023, arXiv preprint arXiv:2305.16539. [Online]. Available: <https://arxiv.org/abs/2305.16539>
- [25] M. A. Kumar and I. Sason, "Projection theorems for the Rényi divergence on  $\alpha$ -convex sets," *IEEE Trans. Inform. Theory*, vol. 62, no. 9, pp. 4924–4935, 2016.

APPENDIX A  
PROOFS

A. Proofs for Section III

Before giving the intended results, we note that we introduced  $m_P$  as the averaged Bregman divergence associated with  $\gamma(x) = x - 1 - \ln(x)$ . For the proof, it will be useful to also define the Bregman divergence associated with  $\gamma(x) = x - 1 - \ln(x)$  itself, which is the so-called Itakura-Saito divergence. For  $f, g \in \mathcal{M}(\Omega, \mathbb{R}_{>0})$ , it is given by

$$IS_P(f, g) = \int_{\Omega} \left( \frac{f}{g} - 1 - \ln \frac{f}{g} \right) dP.$$

By definition, it holds that

$$m_P^2(f, g) = \frac{1}{2} IS \left( f, \frac{f+g}{2} \right) + \frac{1}{2} IS \left( g, \frac{f+g}{2} \right).$$

Furthermore, for  $Q \in \mathcal{C}$ , we have  $IS_P(q, p) = D(P||Q)$ . We now state some auxiliary results before giving the proofs for Section III.

**Lemma 1.** For  $x, y \in \mathbb{R}_{>0}$ , we have

$$|\ln(x) - \ln(y)| = g(m_{\gamma}^2(x, y)),$$

where  $g$  denotes the function

$$g(t) = 2t + 2 \ln \left( 1 + (1 - \exp(-2t))^{1/2} \right).$$

The function  $g$  is concave and satisfies  $g(t) \geq 2t$ .

*Proof.* Let  $m = \frac{x+y}{2}$ . Our goal is to determine the function  $g$  function such that

$$|\ln(x) - \ln(y)| = g(m_{\gamma}^2(x, y)).$$

We first rewrite the right-hand side

$$\begin{aligned} g(m_{\gamma}^2(x, y)) &= g \left( \ln(m) - \frac{1}{2} \ln(x) - \frac{1}{2} \ln(y) \right) \\ &= g \left( \frac{1}{2} \ln \left( \frac{m^2}{x \cdot y} \right) \right) \\ &= g \left( \frac{1}{2} \ln \left( \frac{\left( \frac{m}{y} \right)^2}{\frac{x}{y}} \right) \right) \\ &= g \left( \frac{1}{2} \ln \left( \frac{\left( \frac{1+\frac{x}{y}}{2} \right)^2}{\frac{x}{y}} \right) \right). \end{aligned}$$

Plugging this back in and replacing  $\frac{x}{y}$  by  $w$  leads to

$$|\ln(w)| = g \left( \frac{1}{2} \ln \left( \frac{\left( \frac{1+w}{2} \right)^2}{w} \right) \right)$$

Then we solve the equation

$$\frac{1}{2} \ln \left( \frac{\left( \frac{1+w}{2} \right)^2}{w} \right) = t,$$

which gives

$$\begin{aligned} w &= 2 \exp(2t) - 1 + 2 \cdot (\exp(4t) - \exp(2t))^{1/2} \\ g(t) &= \ln \left( 2 \exp(2t) - 1 + 2 \cdot (\exp(4t) - \exp(2t))^{1/2} \right) \\ &= 2t + \ln \left( 2 - \exp(-2t) + 2 \cdot (1 - \exp(-2t))^{1/2} \right) \\ &= 2t + 2 \ln \left( 1 + (1 - \exp(-2t))^{1/2} \right). \end{aligned}$$

The derivatives of  $g$  are

$$g'(t) = 2 + 2 \frac{(1 - \exp(-2t))^{-1/2} \exp(-2t)}{1 + (1 - \exp(-2t))^{1/2}} = \frac{2}{(1 - \exp(-2t))^{1/2}}$$

$$g''(t) = \frac{-\exp(-t/2)}{2^{1/2}(\sinh t)^{3/2}}.$$

We see that  $g''(t) < 0$  and conclude that  $g$  is concave. Finally, we have

$$g(t) = 2t + 2 \ln\left(1 + (1 - \exp(-2t))^{1/2}\right) \geq 2t,$$

because  $1 - \exp(-2t) \geq 0$ . □

**Lemma 2.** Let  $(f_n)_{n \in \mathbb{N}}$  be a sequence of elements of  $\mathcal{M}(\Omega, \mathbb{R}_{>0})$ , then

$$\limsup_{m, n \rightarrow \infty} m_P(f_m, f_n) = 0 \Rightarrow \limsup_{m, n \rightarrow \infty} \int_{\Omega} \left| \ln \left( \frac{f_m}{f_n} \right) \right| dP = 0.$$

*Proof.* By Lemma 1, we have for  $m, n \in \mathbb{N}$ ,

$$\begin{aligned} \int_{\Omega} \left| \ln \left( \frac{f_n}{f_m} \right) \right| dP &= \int_{\Omega} g(m_{\gamma}^2(f_n, f_m)) dP \\ &\leq g\left(\int_{\Omega} m_{\gamma}^2(f_n, f_m) dP\right) \\ &= g(m_P^2(f_n, f_m)). \end{aligned}$$

The result follows by continuity of  $g$ . □

**Lemma 3.** For  $Q_1, Q_2 \in \mathcal{C}$  such that  $P \ll Q_i$  for  $i \in \{1, 2\}$ , we have

$$m_P^2(q_1, q_2) \leq \frac{D(P \| Q_1 \rightsquigarrow \mathcal{C}) + D(P \| Q_2 \rightsquigarrow \mathcal{C})}{2}.$$

*Proof.* Let  $\bar{Q}$  denote the midpoint between  $Q_1$  and  $Q_2$ . Then we have

$$\begin{aligned} \frac{D(P \| Q_1 \rightsquigarrow \mathcal{C}) + D(P \| Q_2 \rightsquigarrow \mathcal{C})}{2} &= \frac{\sup_{Q \in \mathcal{C}} D(P \| Q_1 \rightsquigarrow Q) + \sup_{Q \in \mathcal{C}} D(P \| Q_2 \rightsquigarrow Q)}{2} \\ &\geq \frac{D(P \| Q_1 \rightsquigarrow \bar{Q}) + D(P \| Q_2 \rightsquigarrow \bar{Q})}{2} = m_P^2(q_1, q_2). \end{aligned}$$

□

*Proof of Proposition 1.* This follows as a direct corollary of Lemma 2. □

We now deviate slightly from the order of the results in Section III and first state the proof of Proposition 2, so that we can use its results in the proof of Theorem 3.

*Proof of Proposition 2.* The implications (3)  $\rightarrow$  (2)  $\rightarrow$  (1) are obvious, so we show here only the implication (1)  $\rightarrow$  (3). Assume that  $P'$  is a measure such that  $-\infty < D(P \| P' \rightsquigarrow \mathcal{C}) < \infty$ . Then there exists a sequence of measures  $Q_n \in \mathcal{C}$  such that

$$D(P \| P' \rightsquigarrow Q_n) \rightarrow D(P \| P' \rightsquigarrow \mathcal{C})$$

for  $n \rightarrow \infty$ . Without loss of generality we may assume that  $-\infty < D(P \| P' \rightsquigarrow Q_n) < \infty$  for all  $n$ . The result follows because

$$D(P \| P' \rightsquigarrow \mathcal{C}) = D(P \| P' \rightsquigarrow Q_n) + D(P \| Q_n \rightsquigarrow \mathcal{C})$$

and all involved quantities are finite. □

*Proof of Theorem 3 (1).* Let  $(Q_n)_{n \in \mathbb{N}}$  denote a sequence in  $\mathcal{C}$  such that

$$\lim_{n \rightarrow \infty} D(P \| Q_n \rightsquigarrow \mathcal{C}) = \inf_{Q \in \mathcal{C}} D(P \| Q \rightsquigarrow \mathcal{C}) = 0,$$

where the last equality follows from Proposition 2. Without loss of generality, we may assume that  $D(P \| Q_n \rightsquigarrow \mathcal{C}) < \infty$  for all  $n$ , so that  $P \ll Q_n$  for all  $n$ . It then follows from Lemma 3 that for  $m, n \in \mathbb{N}$  we have

$$m_P^2(q_m, q_n) \leq \frac{D(P \| Q_m \rightsquigarrow \mathcal{C}) + D(P \| Q_n \rightsquigarrow \mathcal{C})}{2}.$$

It follows that  $(q_n)_{n \in \mathbb{N}}$  is a Cauchy sequence with respect to  $m_P$ , so that  $(q_n)_{n \in \mathbb{N}}$  converges to some function  $\hat{q}$  in  $m_P$ . The latter follows from the completeness of  $(\mathcal{M}(\Omega, (0, \infty)), m_P)$ , i.e. Proposition 1.

Furthermore, suppose that  $(Q'_n)_{n \in \mathcal{C}}$  is another sequence in  $\mathcal{C}$  such that

$$\lim_{n \rightarrow \infty} D(P \| Q'_n \rightsquigarrow \mathcal{C}) = 0.$$

Then, by the same reasoning as before,  $Q_1, Q'_1, Q_2, Q'_2, Q_3, Q'_3, \dots$  is also a Cauchy sequence that converges and since a Cauchy sequence can only converge to a single element this implies the desired uniqueness.  $\square$

*Proof of Theorem 3 (2).* The equality

$$\int_{\Omega} \ln \frac{p'}{\hat{q}} dP = \lim_{n \rightarrow \infty} \int_{\Omega} \ln \frac{p'}{q_n} dP$$

follows from Theorem 3 (1) together with the fact that convergence of  $q_n$  in  $m_P$  implies convergence of the logarithms in  $L_1(P)$ .  $\square$

*Proof of Theorem 3 (3).* Let  $(Q_n)_{n \in \mathcal{C}}$  denote a sequence in  $\mathcal{C}$  such that

$$\lim_{n \rightarrow \infty} D(P \| Q_n \rightsquigarrow \mathcal{C}) = 0.$$

Without loss of generality, we may assume that  $D(P \| Q_n \rightsquigarrow \mathcal{C}) < \infty$  for all  $n$  and that  $q_n$  converges to  $\hat{q}$   $P$ -almost surely. The latter is valid, because convergence in  $m_P$  implies convergence of the logarithms in  $L_1(P)$  by Lemma 2, which gives the existence of an almost surely converging sub-sequence.

Let  $\tilde{Q} = (1-t)Q_1 + tQ$  for fixed  $Q \in \mathcal{C}$  and fixed  $0 < t < 1$ . Let  $Q_{n,s}$  denote the convex combination  $Q_{n,s} = (1-s_n)Q_n + s_n\tilde{Q}$  and  $s_n \in [0, 1]$ . By Theorem 3 (1), we know that there exists some  $\tilde{Q}$  such that  $q_n \rightarrow \hat{q}$  in  $m_P$ .

Since  $Q_{n,s} \in \mathcal{C}$  by convexity, we have that  $D(P \| Q_{n,s} \rightsquigarrow \mathcal{C}) \leq D(P \| Q_n \rightsquigarrow \mathcal{C})$ . We also have

$$\begin{aligned} D(P \| Q_{n,s} \rightsquigarrow \mathcal{C}) &= s_n D(P \| Q_n \rightsquigarrow \tilde{Q}) + s_n IS_P(\tilde{q}, q_{n,s}) + (1-s_n) IS_P(q_n, q_{n,s}) \\ &\geq s_n D(P \| Q_n \rightsquigarrow \tilde{Q}) + s_n IS_P(\tilde{q}, q_{n,s}). \end{aligned}$$

Hence

$$s_n D(P \| Q_n \rightsquigarrow \tilde{Q}) + s_n IS_P(\tilde{q}, q_{n,s}) \leq D(P \| Q_n \rightsquigarrow \mathcal{C}).$$

Division by  $s_n$  gives

$$D(P \| Q_n \rightsquigarrow \tilde{Q}) + IS_P(\tilde{q}, q_{n,s}) \leq \frac{D(P \| Q_n \rightsquigarrow \mathcal{C})}{s_n}.$$

Choosing  $s_n = D(P \| Q_n \rightsquigarrow \mathcal{C})^{1/2}$ , this gives

$$D(P \| Q_n \rightsquigarrow \tilde{Q}) + IS_P(\tilde{q}, q_{n,s}) \leq s_n^{1/2}.$$

Then we get

$$\begin{aligned} IS_P(\tilde{q}, q_{n,s}) &\leq D(P \| \tilde{Q} \rightsquigarrow Q_n) + s_n^{1/2}. \\ \int_{\Omega} \left( \frac{\tilde{q}}{q_{n,s}} + \ln \frac{q_{n,s}}{q_n} \right) dP &\leq P(\Omega) + \tilde{Q}(\Omega) - Q_n(\Omega) + s_n^{1/2}. \end{aligned}$$

Writing  $q_n$  as  $\frac{q_{n,s} - s_n \tilde{q}}{1-s_n}$ , we see

$$\begin{aligned} \ln \frac{q_{n,s}}{q_n} &= \ln \frac{q_{n,s}}{\frac{q_{n,s} - s_n \tilde{q}}{1-s_n}} \\ &= \ln(1-s_n) - \ln \frac{q_{n,s} - s_n \tilde{q}}{q_{n,s}} \\ &= \ln(1-s_n) - \ln \left( 1 - s_n \frac{\tilde{q}}{q_{n,s}} \right) \\ &\geq \ln(1-s_n) + s_n \frac{\tilde{q}}{q_{n,s}}. \end{aligned}$$

Hence

$$\ln(1-s_n) + (1+s_n) \int_{\Omega} \frac{\tilde{q}}{q_{n,s}} dP \leq P(\Omega) + \tilde{Q}(\Omega) - Q_n(\Omega) + s_n^{1/2}.$$

As  $\lim_{n \rightarrow \infty} s_n = 0$ , taking the limit inferior as  $n \rightarrow \infty$  on both sides gives

$$\liminf_{n \rightarrow \infty} \int_{\Omega} \frac{\tilde{q}}{q_{n,s}} dP \leq P(\Omega) + \tilde{Q}(\Omega) - \liminf_{n \rightarrow \infty} Q_n(\Omega).$$

An application of Fatou's lemma gives

$$\int_{\Omega} \frac{dP}{d\tilde{Q}} d\tilde{Q} \leq P(\Omega) + \tilde{Q}(\Omega) - \liminf_{n \rightarrow \infty} Q_n(\Omega).$$

Since  $\tilde{Q} = (1-t)Q_1 + tQ$  we get the inequality

$$\begin{aligned} \int_{\Omega} \frac{dP}{d\tilde{Q}} d((1-t)Q_1 + tQ) &\leq P(\Omega) + (1-t)Q_1(\Omega) + tQ(\Omega) - \liminf_{n \rightarrow \infty} Q_n(\Omega), \\ (1-t) \int_{\Omega} \frac{dP}{d\tilde{Q}} dQ_1 + t \int_{\Omega} \frac{dP}{d\tilde{Q}} dQ &\leq P(\Omega) + (1-t)Q_1(\Omega) + tQ(\Omega) - \liminf_{n \rightarrow \infty} Q_n(\Omega). \end{aligned}$$

Finally we let  $t$  tend to one and obtain the desired result.  $\square$

*Proof of Proposition 3.* Let  $Q \in \mathcal{C}$  arbitrarily. Then there exists a sequence  $(w_i)_{i=1}^n$  in  $[0, 1]$  with  $\sum_i w_i = 1$  such that  $Q = \sum_{i=1}^n w_i Q_i$ . It follows that

$$\begin{aligned} D\left(P \parallel \frac{1}{n} \sum_i Q_i \rightsquigarrow Q\right) &= \int_{\Omega} \ln \frac{\sum_i w_i Q_i}{\frac{1}{n} \sum_i Q_i} dP \\ &\leq \int_{\Omega} \ln \frac{\max_i w_i \sum_i Q_i}{\frac{1}{n} \sum_i Q_i} dP \\ &= \ln(n) + \ln(\max_i w_i) \leq \ln(n). \end{aligned}$$

The proposition follows by taking the supremum over  $Q$  on both sides.  $\square$

*Proof of Proposition 4.* Since  $Q^*$  is the normalized maximum likelihood distribution we have  $\sup_Q \sup_{\omega} \ln \frac{dQ}{dQ^*} < \infty$ . In particular

$$\begin{aligned} \sup_{Q \in \mathcal{C}} D(P \parallel Q^* \rightsquigarrow Q) &= \sup_{Q \in \mathcal{C}} \int_{\Omega} \ln \frac{dQ}{dQ^*} dP \\ &\leq \sup_{Q \in \mathcal{C}} \sup_{\omega} \ln \frac{dQ}{dQ^*}(\omega) < \infty. \end{aligned}$$

$\square$

*Proof of Proposition 5.* We can write

$$D(P \parallel Q_{\theta} \rightsquigarrow Q^*) = D(P \parallel Q_{\theta} \rightsquigarrow Q) + D(P \parallel Q \rightsquigarrow Q^*).$$

By assumption all terms are finite so that minimising  $D(P \parallel Q_{\theta} \rightsquigarrow Q^*)$  over  $\theta$  must be equivalent to minimising  $D(P \parallel Q_{\theta} \rightsquigarrow Q)$  over  $\theta$ . The same argument holds for step 5 in Algorithm 1. The result then follows from [13, Theorem 3.0.13]. While the algorithm described there works by choosing  $\theta_k$  to minimise  $\int_{\Omega} \log((1-\alpha_k)q_{\theta_{k-1}} + \alpha_k q_{\theta}) dP$ , the proof relies on [9, Lemma 5.9], which indeed uses minimisation of  $D(P \parallel (1-\alpha_k)Q_{\theta_{k-1}} + \alpha_k Q_{\theta} \rightsquigarrow Q)$  as described here.  $\square$

*Proof of Theorem 4.* For any  $a \in \mathbb{R}$  we have

$$f_0(i) + a \cdot f_1(i) = f_0(i) \cdot \left(1 + a \cdot \frac{f_1(i)}{f_0(i)}\right). \quad (11)$$

Since  $\frac{f_1(i)}{f_0(i)} \rightarrow 0$  for  $i \rightarrow \infty$  we have that  $f_0(i) + a \cdot f_1(i) \geq 0$  for  $i$  sufficiently large. Therefore, we can apply Fatou's lemma to the function and obtain

$$\begin{aligned} \sum f_0(i) \cdot q^*(i) + a \cdot \sum f_1(i) \cdot q^*(i) &= \sum (f_0(i) + a \cdot f_1(i)) \cdot q^*(i) \\ &= \sum \liminf_{n \rightarrow \infty} (f_0(i) + a \cdot f_1(i)) \cdot q_n(i) \\ &\leq \liminf_{n \rightarrow \infty} \sum_i (f_0(i) + a \cdot f_1(i)) \cdot q_n(i) \\ &= \liminf_{n \rightarrow \infty} \left( \sum_i f_0(i) \cdot q_n(i) + a \cdot \sum_i f_1(i) \cdot q_n(i) \right) \\ &= \liminf_{n \rightarrow \infty} (\lambda_0 + a \cdot \lambda_1) = \lambda_0 + a \cdot \lambda_1. \end{aligned}$$

Hence

$$a \cdot \left( \sum f_1(i) \cdot q^*(i) - \lambda_1 \right) \leq \lambda_0 - \sum f_0(i) \cdot q^*(i). \quad (12)$$

This inequality should hold for all  $a \in \mathbb{R}$ , which is only possible if

$$\begin{aligned}\sum f_1(i) \cdot q^*(i) - \lambda_1 &= 0. \\ \sum f_1(i) \cdot q^*(i) &= \lambda_1.\end{aligned}$$

□

### B. Proofs for Section IV

*Proof of Theorem 5.* Firstly, since  $\hat{E} > 0$  holds  $P$ -almost surely, we have that  $\hat{E}$  is stronger than any  $E' \in \mathcal{E}_C$  with  $P(E' = 0) > 0$ .

Secondly, let  $E \in \mathcal{E}_C$  be an  $e$ -statistic for which  $E > 0$  holds  $P$ -almost surely. Furthermore, let  $Q_n$  be a sequence of measures in  $\mathcal{C}$  such that  $D(P\|Q_n \rightsquigarrow \mathcal{C}) \rightarrow 0$ . We can define a sequence of sub-probability measures  $R_n$  by  $R_n(F) = \int_F E dQ_n$ , which satisfies  $dR_n/dQ_n = E$ . We see

$$\begin{aligned}\int_{\Omega} \ln\left(\frac{\hat{E}}{E}\right) dP &= \int_{\Omega} \ln\left(\frac{dQ_n}{d\hat{Q}}\right) dP + D(P\|R_n) + (P(\Omega) - R_n(\Omega)) \\ &\geq \int_{\Omega} \ln\left(\frac{dQ_n}{d\hat{Q}}\right) dP.\end{aligned}$$

The last expression goes to zero as  $n \rightarrow \infty$ , so we see that  $\hat{E}$  is stronger than  $E$ . □

*Proof of Proposition 6.* Using the fact that  $\ln(x) \leq x - 1$  for  $x > 0$ , we see

$$D(P\|Q^* \rightsquigarrow Q) = \int_{\Omega} \ln \frac{dQ}{dQ^*} dQ \leq \int_{\Omega} \left(\frac{dQ}{dQ^*} - 1\right) dP \leq 0,$$

where the last inequality follows from the fact that  $dP/dQ^*$  is an  $e$ -statistic. □

*Proof of Theorem 6.* Without loss of generality, assume that  $\int_{\Omega} q'/q dP = 1 + \epsilon$  for some  $\epsilon > 0$ . For the sake of brevity, we write  $c_{\beta} := \|q'/q\|_{(1+\beta)/(1-\beta)}$ . We now define a function  $g : [0, 1] \rightarrow \mathbb{R}_{\geq 0}$  as

$$g(\alpha) := D(P\|(1-\alpha)Q + \alpha Q' \rightsquigarrow \mathcal{C}).$$

Notice that  $g(0) = \delta$  and  $g(\alpha) \geq 0$ , since  $(1-\alpha)Q + \alpha Q' \in \mathcal{C}$ . This function and its derivatives will guide the rest of the proofs, and we now list some properties that we will need:

$$g'(\alpha) := \frac{d}{d\alpha} g(\alpha) = \int_{\Omega} \frac{q - q'}{(1-\alpha)q + \alpha q'} dP, \text{ so} \quad (13)$$

$$g'(0) = \int_{\Omega} \left(1 - \frac{q'}{q}\right) dP = -\epsilon, \quad (14)$$

$$g''(\alpha) := \frac{d^2}{d\alpha^2} g(\alpha) = \int_{\Omega} \left(\frac{q' - q}{(1-\alpha)q + \alpha q'}\right)^2 dP, \quad (15)$$

$$\text{so } g''(0) = \int_{\Omega} \left(1 - \frac{q'}{q}\right)^2 dP = 1 - 2(1 + \epsilon) + c_1 \text{ and}$$

$$0 \leq g''(\alpha) \leq \frac{1}{(1-\alpha)^2} g''(0). \quad (16)$$

We now prove (9). We start with the case  $\beta = 1$  and will use the result for  $\beta = 1$  to prove the case for  $\beta < 1$ . The proof for the case  $\beta > 1$  comes later; it requires a completely different proof.

$\beta = 1$ . The general idea is simple: at  $\alpha = 0$  the function  $g(\alpha)$  is equal to  $\delta$  and has derivative  $-\epsilon$ . Its second derivative is positive and bounded by constant times  $g''(0) \leq c_1$  for all  $\alpha \leq 1/2$ . Thus, if  $\epsilon$  is larger than a certain threshold,  $g(\alpha)$  will become negative at some  $\alpha \leq 1/2$ , but this is not possible since  $g$  is a description gain and we would arrive at a contradiction. The details to follow simply amount to calculating the threshold as a function of  $\delta$ .

By Taylor's theorem, we have for any  $\alpha \in [0, 1/2]$  that

$$\begin{aligned}g(\alpha) &= g(0) + g'(0)\alpha + \max_{0 \leq \alpha^{\circ} \leq \alpha} \frac{g''(\alpha^{\circ})}{2} \alpha^2 \\ &\leq g(0) + g'(0)\alpha + 2g''(0)\alpha^2 \\ &\leq \delta - \epsilon\alpha + 2\alpha^2 c_1,\end{aligned}$$

where we use the properties derived above. This final expression has a minimum in  $\alpha^* = \min\{\epsilon/4c_1, 1/2\}$ . By non-negativity of  $g$ , we know that  $\delta - \epsilon\alpha^* + 2\alpha^{*2}c_1 \geq 0$ . This gives  $\epsilon \leq (8c_1\delta)^{1/2}$  in the case that  $\alpha^* = \epsilon/4c_1 < 1/2$ , and  $\epsilon \leq 2\delta + c_1$  otherwise. In the latter case, it holds that  $c_1 < \epsilon/2$ , so the bound can be loosened slightly to find the simplification  $\epsilon \leq 4\delta$ . This concludes the proof for  $\beta = 1$ , which we now use to prove the case that  $\beta < 1$ .

$\beta < 1$ . For any  $a > 0$ , it holds that

$$\int_{\Omega} \frac{q'}{q} dP = \int_{\Omega} \frac{q'}{q} \mathbb{1}\{q'/q \leq a\} dP + \int_{\Omega} \frac{q'}{q} \mathbb{1}\{q'/q > a\} dP. \quad (17)$$

We write  $q'' := q' \mathbb{1}\{q'/q \leq a\}$  and we will bound the first term on the right-hand side of (17) using the proof above with  $Q'$  replaced by  $Q''$ . Since  $Q''$  is not necessarily an element of  $\mathcal{C}$ , we need to verify non-negativity, which follows because for each  $\alpha \in (0, 1)$ , we have that  $D(P\|(1-\alpha)Q + \alpha Q'' \rightsquigarrow \mathcal{C}) \geq D(P\|(1-\alpha)Q + \alpha Q' \rightsquigarrow \mathcal{C}) \geq 0$ . Furthermore, it holds that

$$\begin{aligned} \left\| \frac{q''}{q} \right\|_2^2 &= \int_{\Omega} \left( \frac{q''}{q} \right)^2 dP \\ &= \int_{\Omega} \left( \frac{q''}{q} \right)^{1+\beta} \left( \frac{q''}{q} \right)^{1-\beta} dP \\ &\leq a^{1-\beta} c_{\beta} \end{aligned}$$

The results above therefore give

$$\int_{\Omega} \frac{q''}{q} dP \leq 1 + \max\{(8a^{1-\beta} c_{\beta} \delta)^{1/2}, 2\delta\}.$$

For the second term on the right-hand side of (17), we use a Markov-type bound, i.e.

$$\begin{aligned} \int_{\Omega} \frac{q'}{q} \mathbb{1}\{q'/q > a\} dP &\leq \int_{\Omega} \frac{q'}{q} \left( \frac{q'/q}{a} \right)^{\beta} \mathbb{1}\{q'/q > a\} dP \\ &\leq a^{-\beta} c_{\beta}. \end{aligned}$$

Putting this together gives

$$\int_{\Omega} \frac{q'}{q} dP \leq 1 + \max\{(8a^{1-\beta} c_{\beta} \delta)^{1/2}, 4\delta\} + a^{-\beta} c_{\beta}.$$

Since this holds for any  $a$ , we now pick it to minimize this bound. To this end, consider

$$\frac{d}{da} (8a^{1-\beta} c_{\beta} \delta)^{1/2} + a^{-\beta} c_{\beta} = \frac{(1-\beta)(8c_{\beta} \delta)^{1/2}}{2} a^{-(1+\beta)/2} - \beta a^{-(1+\beta)} c_{\beta}.$$

Setting this to zero, we find

$$a^* = \left( \frac{\beta c_{\beta}^{1/2}}{(1-\beta)(2\delta)^{1/2}} \right)^{2/(1+\beta)}.$$

The proof is concluded by noting that

$$\begin{aligned} (8a^{*1-\beta} c_{\beta} \delta)^{1/2} &= \left( 8 \left( \frac{\beta c_{\beta}^{1/2}}{(1-\beta)(2\delta)^{1/2}} \right)^{2(1-\beta)/(1+\beta)} c_{\beta} \delta \right)^{1/2} \\ &= 2c_{\beta}^{1/(\beta+1)} (2\delta)^{\beta/(\beta+1)} \left( \frac{\beta}{1-\beta} \right)^{(1-\beta)/(1+\beta)} \end{aligned}$$

and

$$\begin{aligned} a^{*-\beta} c_{\beta} &= \left( \frac{\beta c_{\beta}^{1/2}}{(1-\beta)(2\delta)^{1/2}} \right)^{-2\beta/(1+\beta)} c_{\beta} \\ &= c_{\beta}^{1/(\beta+1)} \left( \frac{\beta}{1-\beta} \right)^{-2\beta/(1+\beta)} (2\delta)^{\beta/(1+\beta)}. \end{aligned}$$

$\beta > 1$ . We now prove the result for  $\beta \in (1, \infty)$ ; the proof for  $\beta = \infty$  follows by a minor modification of (19). If  $\epsilon \leq 0$  there is nothing to prove, so without loss of generality we can write  $\epsilon = \gamma\delta$  for some  $\gamma > 0$ ; we will bound  $\gamma$ . Whereas the previous proof exploited the fact that the second derivative  $g''(\alpha)$  was bounded above in terms of  $\delta$  and hence ‘not too large’, the proof below uses the condition that  $c_{\beta}$  is finite to show first, (a), that  $g''(\alpha)$  can also be bounded *below* in terms of  $(\gamma, \delta)$ . Therefore, if  $\epsilon$  exceeds a certain threshold, as  $\alpha$  moves away from the  $\alpha^*$  at which  $g(\alpha)$  achieves its minimum in the direction



of the furthest boundary point (i.e. if  $\alpha^* < 1/2$ , we consider  $\alpha \uparrow 1$ , if  $\alpha^* \geq 1/2$  we consider  $\alpha \downarrow 0$ ),  $g(\alpha)$  will become larger than  $K\delta$  or  $\delta$  respectively, and we arrive at a contradiction. (b) below gives the detailed calculation of this threshold.

*Proof of (a).* Fix some  $0 \leq \tilde{\alpha} < 1$  (we will derive a bound for any such  $\tilde{\alpha}$  and later optimize for  $\tilde{\alpha}$ ; for a sub-optimal yet easier derivation take  $\tilde{\alpha} = 1/2$ ). By Taylor's theorem, we have  $0 \leq g(\tilde{\alpha}) = \delta - \tilde{\alpha}\epsilon + (1/2)\tilde{\alpha}^2 g''(\alpha^\circ)$  for some  $0 \leq \alpha^\circ \leq \tilde{\alpha}$ . Plugging in  $\epsilon = \gamma\delta$  we find that

$$g''(\alpha^\circ) \geq \frac{2}{\tilde{\alpha}^2}(\tilde{\alpha}\gamma - 1)\delta.$$

This gives a lower bound on  $g''(\alpha^\circ)$  for *some*  $\alpha^\circ$  in terms of  $(\gamma, \delta)$ . We now turn this into a weaker lower bound on *all*  $\alpha$ . First, using (16) and then  $\alpha^\circ \leq \tilde{\alpha}$  and then the above lower bound, we find

$$\begin{aligned} g''(0) &\geq \max_{\alpha \in [0, \tilde{\alpha}]} (1 - \alpha)^2 g''(\alpha) \geq (1 - \alpha^\circ)^2 g''(\alpha^\circ) \\ &\geq (1 - \tilde{\alpha})^2 g''(\alpha^\circ) \geq 2f_{\tilde{\alpha}}(\gamma, \delta), \end{aligned} \quad (18)$$

where  $f_{\tilde{\alpha}}(\gamma, \delta) := ((1 - \tilde{\alpha})/\tilde{\alpha})^2(\tilde{\alpha}\gamma - 1)\delta$  is a function that is linear in  $\gamma$  and  $\delta$ . We have now lower bounded  $g''(0)$  in terms of  $\gamma, \delta$ . We next show that, under our condition that  $c_\beta < \infty$ , this implies a (weaker) lower bound on  $g''(\alpha)$  for all  $\alpha$ . For this, fix any  $C > 1$ . We have for all  $0 < \alpha \leq 1$ :

$$\begin{aligned} g''(\alpha) &\geq \int_{\Omega} \mathbf{1}_{q' \leq Cq} \cdot \left( \frac{q' - q}{(1 - \alpha)q + \alpha q'} \right)^2 dP \\ &\geq \int_{\Omega} \mathbf{1}_{q' \leq Cq} \cdot \left( \frac{q' - q}{(1 - \alpha)q + \alpha Cq} \right)^2 dP \\ &= \int_{\Omega} \mathbf{1}_{q' \leq Cq} \cdot \left( \frac{q' - q}{q} \right)^2 dP \cdot \frac{1}{(1 + \alpha(C - 1))^2} \\ &\geq \frac{1}{(1 + (C - 1))^2} \left( g''(0) - \int_{\Omega} \mathbf{1}_{q' > Cq} \cdot \left( \frac{q'}{q} - 1 \right)^2 dP \right) \\ &\geq \frac{1}{C^2} (2f_{\tilde{\alpha}}(\gamma, \delta) - C^{1-\beta} c_\beta), \end{aligned} \quad (19)$$

where in the fourth line we used the definition of  $g''(0)$ , and in the fifth line we used (18) and a Markov-type bound on the integral, i.e. we used that  $\int_{\Omega} \mathbf{1}_{q' > Cq} \cdot (q'/q - 1)^2 dP$  is bounded by

$$\int_{\Omega} \mathbf{1}_{q' > Cq} \cdot \left( \frac{q'}{q} \right)^2 dP \leq \int_{\Omega} \left( \frac{q'}{C} \right)^{\beta-1} \cdot \left( \frac{q'}{q} \right)^2 dP = C^{1-\beta} c_\beta.$$

By differentiation we can determine the  $C$  that maximizes the bound (19). This gives  $C^{1-\beta} = f_{\tilde{\alpha}}(\gamma, \delta)/(4/c_\beta(1+\beta))$ . and with this choice of  $C$ , (19) becomes

$$g''(\alpha) \geq f_{\tilde{\alpha}}(\gamma, \delta)^{(\beta+1)/(\beta-1)} c_\beta^{2/(1-\beta)} h(\beta) \quad (20)$$

where  $h(\beta) = (4/(1 + \beta))^{2/(\beta-1)} \cdot (2(\beta - 1))/(1 + \beta)$ . We are now ready to continue to:

*Proof of (b).* Let  $\alpha^* \in [0, 1]$  be the point at which  $g(\alpha)$  achieves its minimum. If  $\alpha^* \leq 1/2$ , a second-order Taylor approximation of  $g(1)$  around  $\alpha^*$  gives that

$$\begin{aligned} K\delta &\geq g(1) \geq (1/2)(1 - \alpha^*)^2 \min_{\alpha \in [\alpha^*, 1]} g''(\alpha) \\ &\geq (1/8)f_{\tilde{\alpha}}(\gamma, \delta)^{(\beta+1)/(\beta-1)} c_\beta^{2/(1-\beta)} h(\beta), \end{aligned}$$

so that after some manipulations

$$f_{\tilde{\alpha}}(\gamma, \delta)^{(1+\beta)/(\beta-1)} \leq 8K' c_\beta^{2/(\beta-1)} \cdot h(\beta)^{-1} \delta. \quad (21)$$

with  $K' = K$ . If  $\alpha^* > 1/2$ , we perform a completely analogous second-order Taylor approximation of  $g(0)$  around  $\alpha^*$ , which will then give (21) again but with  $K'$  replaced by 1. We thus always have (21) with  $K' = \max\{K, 1\}$ . Unpacking  $f_{\tilde{\alpha}}$  in (21) and rearranging gives:

$$\gamma \leq \frac{\tilde{\alpha}}{(1 - \tilde{\alpha})^2} \cdot V + \frac{1}{\tilde{\alpha}}$$

with  $V = c_\beta^{2/(1+\beta)} \cdot (8K'/h(\beta))^{(\beta-1)/(1+\beta)} \delta^{-2/(1+\beta)}$ . We now pick the  $\tilde{\alpha}$  that makes both terms on the right equal, so that the right-hand side becomes equal to  $2/\tilde{\alpha}$ . This is the solution to the equation  $(\tilde{\alpha}/(1 - \tilde{\alpha}))^2 V = 1$  which must clearly be obtained for some  $0 < \tilde{\alpha} < 1$ , so this  $\tilde{\alpha}$  satisfies our assumptions. Basic calculation gives

$$\gamma \leq \frac{2}{\tilde{\alpha}} = 2 \cdot \left( V^{1/2} + 1 \right)$$

and unpacking  $V$  we obtain

$$\epsilon = \gamma\delta \leq c^* \cdot \delta^{\frac{\beta}{1+\beta}} + 2\delta.$$

where

$$c^* = c_\beta^{1/(1+\beta)} \cdot (8K'/h(\beta))^{\frac{\beta-1}{2(1+\beta)}}.$$

Unpacking  $h(\beta)$  gives the desired result.  $\square$

## APPENDIX B RIPR STRICT SUB-PROBABILITY MEASURE

In this section, we discuss a general way to construct a measure  $P$  and convex set of distributions  $\mathcal{C}$  such that the reverse information projection of  $P$  on  $\mathcal{C}$  is a strict sub-probability measure. For simplicity, we take  $\Omega = \mathbb{N}$  and  $\mathcal{F} = 2^{\mathbb{N}}$ , though the idea should easily translate to more general settings.

**Proposition 7.** *Let  $g : \mathbb{N} \rightarrow \mathbb{R}_{>0}$  be a function, and let  $\mathcal{C}$  denote the set of measures  $\{Q : \sum_i g(i)q(i) \leq \nu\}$  for some  $\nu > 0$ . Then for any  $P$  that is not in  $\mathcal{C}$  we have that  $E(i) = g(i)/\nu$  is the optimal  $e$ -statistic.*

*Proof.* The extreme points in  $\mathcal{C}$  are the measure with total mass 0 and measures of the form  $\frac{\nu}{g(i)}\delta_i$ , i.e. measures concentrated in single points. An  $e$ -statistic  $E$  must satisfy

$$\sum_j E(j) \frac{\nu}{g(i)} \delta_i(j) \leq 1$$

or, equivalently,  $E(i) \frac{\nu}{g(i)} \leq 1$ . Hence  $E \leq g/\nu$  so the optimal  $e$ -statistic is  $g/\nu$ .  $\square$

Let  $g : \mathbb{N} \rightarrow \mathbb{R}_{>0}$  be any function that satisfies

$$\lim_{n \rightarrow \infty} g(n) = 0.$$

Furthermore, let  $P$  denote a probability measure on the natural numbers such that

$$\sum_i \frac{p(i)}{g(i)} = c$$

for some  $c \in \mathbb{R}_{>0}$ . For  $\nu \in (0, 1/c)$  and let  $\mathcal{C}_\nu$  denote the set of measures  $\{Q : \sum_i g(i)q(i) \leq \nu\}$ . Note that we do not yet require all measures in  $\mathcal{C}_\nu$  to be probability measures so that the set  $\mathcal{C}_\nu$  is compact. It follows that there exists a unique element of  $\mathcal{C}$  that minimizes  $\sum_i p(i) \ln(p(i)/q(i))$ .

The optimal  $e$ -statistic is  $E_\nu = g/\nu$ , and we may define the measure  $\hat{Q}_\nu$  by

$$\hat{q}_\nu(i) = \frac{p(i)}{E_\nu(i)} = \nu p(i)/g(i),$$

and we can check that  $\hat{Q}_\nu \in \mathcal{C}$ . Hence  $\hat{Q}_\nu$  minimizes  $\sum_i p(i) \ln(p(i)/q(i))$ .

This is a strict sub-probability measure:

$$\begin{aligned} \sum_i \hat{q}_\nu(i) &= \nu \sum_i \frac{p(i)}{g(i)} \\ &= \nu c \\ &< 1, \end{aligned}$$

where we use that  $\nu < 1/c$ .

The next step is to prove that the information projection does not change if we restrict to the set of probability measures in  $\mathcal{C}_{\nu^*}$ , which we denote by  $\tilde{\mathcal{C}}_{\nu^*}$ . To this end, note first that for  $\nu < \nu^*$ , we have that  $\sum g(i)q_\nu(i) < \nu^*$ , so that for all  $\nu < \nu^*$  there exists  $n_\nu \in \mathbb{N}$  such that the probability measure defined by

$$q_\nu(i) + \left(1 - \sum_j q_\nu(j)\right) \delta_{n_\nu}(i)$$

is an element of  $\tilde{\mathcal{C}}_{\nu^*}$ . Hence

$$\begin{aligned} D(P\|\tilde{\mathcal{C}}) &\leq D\left(P\left\|Q_\lambda + \left(1 - \sum_j q_\nu(i)\right)\delta_{n_\nu}\right.\right) \\ &= \sum_{i \in \mathbb{N}} p(i) \ln\left(\frac{p(i)}{Q_\nu(i) + \left(1 - \sum_{j \in \mathbb{N}} q_\nu(j)\right)\delta_{n_\nu}(i)}\right) \\ &= -p(n_\nu) \ln\left(\frac{p(n_\nu)}{q_\nu(n_\nu)}\right) + p(n_\nu) \ln\left(\frac{p(n_\nu)}{q_\nu(n_\nu) + 1 - \sum_{j \in \mathbb{N}} q_\nu(j)}\right) + \sum_{i=1}^{\infty} p(i) \ln\left(\frac{p(i)}{q_\nu(i)}\right). \end{aligned}$$

The first term can be written as

$$\begin{aligned} p(n_\nu) \ln\left(\frac{p(n_\nu)}{q_\nu(n_\nu)}\right) &= q_\nu(n_\nu) \frac{p(n_\nu)}{q_\nu(n_\nu)} \ln\left(\frac{p(n_\nu)}{q_\nu(n_\nu)}\right) \\ &= q_\nu(n_\nu) \frac{g(n_\nu)}{\nu} \ln\left(\frac{g(n_\nu)}{\nu}\right) \end{aligned}$$

Then notice that for  $\nu \rightarrow \nu^*$ , we must have that  $n_\nu \rightarrow \infty$ . Using that  $c \ln(c) \rightarrow 0$  for  $c \rightarrow 0$  we see the first term tends to 0 for  $\nu \rightarrow \nu^*$ . Similarly, the second term can be written as

$$\begin{aligned} p(n_\nu) \ln\left(\frac{p(n_\nu)}{q_\nu(n_\nu) + 1 - \sum_{j \in \mathbb{N}} q_\nu(j)}\right) &= \\ &\left(q_\nu(n_\nu) + 1 - \sum_{j \in \mathbb{N}} q_\nu(j)\right) \frac{p(n_\nu)}{q_\nu(n_\nu) + 1 - \sum_{j \in \mathbb{N}} q_\nu(j)} \times \ln\left(\frac{p(n_\nu)}{q_\nu(n_\nu) + 1 - \sum_{j \in \mathbb{N}} q_\nu(j)}\right). \end{aligned}$$

We also have

$$\frac{p(n_\nu)}{q_\nu(n_\nu) + 1 - \sum_{j \in \mathbb{N}} q_\nu(j)} \rightarrow 0$$

for  $\nu \rightarrow \nu^*$  and using that  $c \ln(c) \rightarrow 0$  for  $c \rightarrow 0$  we get the second term tends to 0 for  $\nu \rightarrow \nu^*$ . Therefore we see

$$\begin{aligned} D(P\|\tilde{\mathcal{C}}) &\leq \lim_{\nu \rightarrow \nu^*} D\left(P\left\|Q_\nu + \left(1 - \sum_i q_\nu(i)\right)\delta_{n_\nu}\right.\right) \\ &\leq \sum_i p(i) \ln\left(\frac{p(i)}{q_{\nu^*}(i)}\right) \\ &= \inf_{Q \in \mathcal{C}} \sum_i p(i) \ln\left(\frac{p(i)}{q(i)}\right). \end{aligned}$$

The inequality trivially also holds the other way around, so we find that

$$D(P\|\tilde{\mathcal{C}}) = \inf_{Q \in \mathcal{C}} \sum_i p(i) \ln\left(\frac{p(i)}{q(i)}\right).$$

It follows that  $Q_{\nu^*}$  is a strict sub-probability measure, and at the same time it is the reverse information projection of  $P$  onto  $\tilde{\mathcal{C}}_{\nu^*}$ .