# GLIVENKO–CANTELLI FOR $f$-DIVERGENCE

HAOMING WANG AND LEK-HENG LIM

Abstract. We extend the celebrated Glivenko–Cantelli theorem, sometimes called the fundamental theorem of statistics, from its standard setting of total variation distance to all $f$-divergences. A key obstacle in this endeavor is to define $f$-divergence on a subcollection of a $\sigma$-algebra that forms a $\pi$-system but not a $\sigma$-subalgebra. This is a side contribution of our work. We will show that this notion of $f$-divergence on the $\pi$-system of rays preserves nearly all known properties of standard $f$-divergence, yields a novel integral representation of the Kolmogorov–Smirnov distance, and has a Glivenko–Cantelli theorem. We will also discuss the prospects of a Vapnik–Chervonenkis theory for $f$-divergence.

## 1. Introduction

The Glivenko–Cantelli theorem [22, 8] is a cornerstone of empirical process theory. It is likely the best known statement regarding the asymptotic behavior of stochastic processes formed by empirical measures [16]. It is also crucial in nonparametric statistics and forms the basis for statistical consistency in many estimation problems [16]. In the Kolmogorov–Smirnov test, the theorem ensures that the test statistic has desirable asymptotic properties [46]. In statistical resampling methods like the bootstrap, it guarantees that the empirical distribution derived from resampled data will approximate the true distribution as the number of samples grows [18]. Because of its many central roles, the Glivenko–Cantelli theorem is often called the fundamental theorem of statistics [49, 27, 14, 44].

In machine learning, the Glivenko–Cantelli theorem can be extended to the Vapnik–Chervonenkis theorem [55], also known as the fundamental theorem of learning theory [45]. It is used to show the consistency of the principle of empirical risk minimization (ERM) [54] and to derive bounds on generalization error by ensuring the difference between empirical and true distributions diminishes with larger samples, aiding in over-fitting control [6].

Given its pivotal role in both statistics and machine learning, it is surprising that the Glivenko–Cantelli theorem is somewhat limited in scope—it only works with the *total variation norm* restricted to a Glivenko–Cantelli class. Modern AI applications, on the other hand, have ushered in a plethora of alternative distances between two probability distributions $\mu$ and $\nu$, most of them easier to compute than the total variation norm. The most prominent of these are the $f$-divergences

$$\mathrm{D}_f(\mu\|\nu) = \int_\Omega f\Big(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\Big)\,\mathrm{d}\nu = \int_\Omega f\big(g(x)\big)\,\mathrm{d}\nu(x), \tag{1}$$

where $g$ is the Radon–Nikodym derivative with $\mathrm{d}\mu(x) = g(x)\,\mathrm{d}\nu(x)$. The total variation norm is itself an $f$-divergence with $f(t) = |t-1|/2$ but many other $f$-divergence with yet other $f$'s have played a prominent role in important AI applications recently. In [31], the Kullback–Leibler (KL) divergence is used to regularize the posterior distribution in the Variational Autoencoder (VAE), aligning it with a chosen prior. In [23], Generative Adversarial Networks (GANs) are trained by minimizing the Jensen–Shannon (JS) divergence between the model and real data distributions. In [33], Rényi $\alpha$-divergences provide a smooth interpolation from the evidence lower bound to the log

2020 *Mathematics Subject Classification.* 28A33, 28A50, 46E27, 60E05, 60E15, 60F17, 94A17.
*Key words and phrases.* probability measures, Kolmogorov–Smirnov distance, total variation distance, $f$-divergence, Glivenko–Cantelli theorem.

marginal likelihood that is controlled by the parameter $\alpha$. The goal of our article is to generalize the Glivenko–Cantelli theorem to any $f$-divergence, restricted to a Glivenko–Cantelli class defined with respect to this $f$-divergence.

At this point, one may wonder about Wasserstein metrics, another class of distances between probability distributions that rivals $f$-divergences in importance and popularity. But as was shown in [5, 35], there is no equivalent of Glivenko–Cantelli theorem for $p$-Wasserstein metric. Such a result will always be trivial as the "Glivenko–Cantelli class" with respect to any $p$-Wasserstein metric is just the entire Borel $\sigma$-algebra: see [5] for $p \in [1, \infty)$ and [35] for $p = \infty$. On the other hand, the main contribution of our article shows that there is a natural, nontrivial generalization of the Glivenko–Cantelli theorem to $f$-divergences of which the special case $f(t) = |t - 1|/2$ gives us the classical Glivenko–Cantelli theorem.

We remind the reader of this classical version, taking the opportunity to set notations, define basic notions, and establish some facts for later use. Let $(\Omega, \Sigma, P)$ be a probability space and $X : \Omega \to \mathbb{R}$ be a random variable. Let $\nu$ be the distribution of $X$, i.e., the pushforward measure of $P$ by $X$. Thus $\nu$ is a probability measure on the real Borel space $(\mathbb{R}, \mathcal{B})$. For $X_1, \dots, X_n \sim \nu$ independent and identically distributed random variables on the probability space $(\mathbb{R}, \mathcal{B}, \nu)$, the *empirical distribution* is the random measure

$$\nu_n(A, \omega) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i(\omega) \in A\}},$$

defined for any $A \in \mathcal{B}$ and $\omega \in \Omega$. For a fixed $\omega \in \Omega$, $\nu_n(\cdot, \omega)$ is a probability measure on $(\mathbb{R}, \mathcal{B})$; and for a fixed $A \in \mathcal{B}$, $\nu_n(A) \coloneqq \nu_n(A, \cdot)$ is a random variable on $\Omega$.

As is well known, the strong law of large number states that as $n \to \infty$,

SLLN: for each fixed $A$, $\nu_n(A) - \nu(A) \to 0$ almost surely.

The Glivenko–Cantelli theorem may be viewed as an attempt to replace the qualifier "for each fixed $A$" by "for all $A \in \mathcal{C}$" where $\mathcal{C}$ is an appropriately chosen class of subsets of $\Omega$ called a Glivenko–Cantelli class [46], while maintaining the property that as $n \to \infty$,

GC: $$\sup_{A \in \mathcal{C}} \big(\nu_n(A) - \nu(A)\big) \to 0 \text{ almost surely.}$$

The following example shows that one cannot simply take $\mathcal{C} = \mathcal{B}$, the Borel $\sigma$-algebra.

**Example 1.1.** Let $\nu$ be the standard Gaussian distribution. Let $X_1, \dots, X_n \sim \nu$ be independent and identically distributed random variables with realizations $x_1, \dots, x_n$. If $A = \{x_1, \dots, x_n\} \in \mathcal{B}$, then $\nu_n(A) = 1$ but $\nu(A) = 0$. Thus $\sup_{A \in \mathcal{B}} \big(\nu_n(A) - \nu(A)\big) = 1$ for all $n$ and $\mathcal{B}$ is not a Glivenko–Cantelli class.

One might think that $\mathcal{B}$ is too big but even if we choose a much smaller $\mathcal{C}$, it could still fail to be a Glivenko–Cantelli class, as the following example from [47] shows.

**Example 1.2** (Shortt)**.** Let $\nu$ be the uniform measure on $[0, 1]$ and $\mathcal{C}$ be the collection of all finite subset of $[0, 1]$. Let $X_1, \dots, X_n \sim \nu$ be independent and identically distributed random variables with realizations $x_1, \dots, x_n$. If $A = \{x_1, \dots, x_n\} \in \mathcal{C}$, then $\nu_n(C) = 1$ but $\nu(C) = 0$. Thus $\sup_{A \in \mathcal{C}} \big(\nu_n(A) - \nu(A)\big) = 1$ for all $n$ and $\mathcal{C}$ is not a Glivenko–Cantelli class.

So the existence of (nontrivial) Glivenko–Cantelli classes is not obvious. The Glivenko–Cantelli theorem shows that they indeed exists, with the class of rays the original [22, 8] and best known example.

**Definition 1.3** (Rays)**.** The class of *open rays* and the class of *closed rays* are

$$\mathcal{R}_{\mathrm{o}} = \{(-\infty, a) : a \in \mathbb{R}\} \cup \{\varnothing, \mathbb{R}\}, \quad \mathcal{R}_{\mathrm{c}} = \{(-\infty, a] : a \in \mathbb{R}\} \cup \{\varnothing, \mathbb{R}\}$$

respectively. Their union is the class of *rays*, denoted

$$\mathcal{R} = \mathcal{R}_{\mathrm{c}} \cup \mathcal{R}_{\mathrm{o}}.$$

These are all $\pi$-systems [17, p. 202] but not $\sigma$-algebras. In this article, we restrict ourselves to "left" rays, i.e., where the left-end point of the interval is always $-\infty$. In topology, $\mathcal{R}_{\mathrm{o}}$ is called the left-ray topology or left-order topology.

Glivenko [22] showed that $\mathcal{R}_{\mathrm{c}}$ is a Glivenko–Cantelli class whereas Cantelli [8] showed that $\mathcal{R}_{\mathrm{o}}$ is a Glivenko–Cantelli class but we will soon see that it makes no difference whether we use $\mathcal{R}_{\mathrm{c}}$, $\mathcal{R}_{\mathrm{o}}$, or $\mathcal{R}$; all three statements are equivalent. Obviously, the name "Glivenko–Cantelli class" came about much later [46], earlier statements of the result were more of the following form:

**Theorem 1.4** (Glivenko–Cantelli)**.** *Let $\nu$ be a Borel probability measure and $\nu_n$ be the corresponding empirical measure, $n \in \mathbb{N}$. Then, almost surely,*

$$\sup_{A \in \mathcal{R}} \big(\nu_n(A) - \nu(A)\big) \to 0. \tag{2}$$

There is no substantive difference if we instead use $\mathcal{R}_{\mathrm{o}}$ or $\mathcal{R}_{\mathrm{c}}$ in place of $\mathcal{R}$ in Theorem 1.4 as

$$\sup_{A \in \mathcal{R}_{\mathrm{c}}} \big(\mu(A) - \nu(A)\big) = \sup_{A \in \mathcal{R}_{\mathrm{o}}} \big(\mu(A) - \nu(A)\big) = \sup_{A \in \mathcal{R}} \big(\mu(A) - \nu(A)\big) \tag{3}$$

for any probability measures $\mu$ and $\nu$ on $(\mathbb{R}, \mathcal{B})$, a fact that follows easily from the continuity of probability measures. The common value in (3) defines a distance between probability measures called the Kolmogorov–Smirnov distance, although it is usually defined in terms of $\mathcal{R}_{\mathrm{c}}$ [30].

The *total variation distance* is defined by

$$\mathrm{D}_{\mathrm{TV}}(\mu \| \nu) = \sup_{A \in \mathcal{B}} \big(\mu(A) - \nu(A)\big). \tag{4}$$

So the Kolmogorov–Smirnov distance is just the total variation distance "restricted" to $\mathcal{R} \subseteq \mathcal{B}$, i.e., a partial variation over a smaller class of sets than the full Borel $\sigma$-algebra. As we have alluded to earlier, the total variation is an example of $f$-divergence, an extensive class of distance between probability measures. Note that whether we take absolute values or not in (4) makes no difference to the value because $A \in \mathcal{B}$ iff $A^{\mathsf{c}} \in \mathcal{B}$. Throughout this article, an integral sign $\int$ unadorned with upper and lower limits is taken to mean $\int_{\mathbb{R}}$.

**Definition 1.5** ($f$-divergence)**.** Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous convex function with $f(1) = 0$. Let $\mu$ and $\nu$ be probability measures on $(\mathbb{R}, \mathcal{B})$ be such that $\mu \ll \nu$, i.e., $\mu$ is absolutely continuous with respect to $\nu$, so that we may speak of Radon–Nikodym derivative $\mathrm{d}\mu/\mathrm{d}\nu$. Then the $f$-divergence [41, 13, 37, 1] between $\mu$ and $\nu$ is given by

$$\mathrm{D}_f(\mu \| \nu) = \int f\Big(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\Big)\, \mathrm{d}\nu. \tag{5}$$

Different choices of $f$ yield various well-known divergences, including Kullback–Leibler [3], Le Cam [32], Jensen–Shannon [34], Jeffreys [29], Chernoff [10], Pearson $\chi^2$ [39], Hellinger squared [26], exponential [36], and alpha–beta [19] divergences, and so on. For $f(t) = |t-1|/2$, we get [52]:

$$\int \frac{1}{2}\Big|\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\Big|\, \mathrm{d}\nu = \sup_{A \in \mathcal{B}} \big(\mu(A) - \nu(A)\big),$$

the total variation distance. To obtain a Glivenko–Cantelli theorem for an arbitrary $f$-divergence, we first need an analog of Kolmogorov–Smirnov distance corresponding to an $f$-divergence.

**Question 1.** Is there a notion of $f$-divergence over $\mathcal{R}$ for general $f$ that (a) reduces to the Kolmogorov–Smirnov distance in (3) when $f(t) = |t-1|/2$; and (b) reduces to the standard $f$-divergence in (5) when $\mathcal{R}$ is replaced by $\mathcal{B}$?

The answer to Question 1 is yes, provided by what we will call an $f$-divergence over $\mathcal{R}$ and denoted $\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu)$. We will establish the existence of $f$-divergence over $\mathcal{R}$ in Section 3 and develop some of its properties in Section 4, showing in particular that this new notion preserves some of the best-known relations between different $f$-divergences. In Section 5, we will prove the result claimed in the title of our article, namely, $\mathrm{D}_f^{\mathcal{R}}(\nu_n\|\nu)$ converges to zero almost surely. Note that when $f(t) = |t-1|/2$, this reduces to the Glivenko–Cantelli theorem in Theorem 1.4.

As we will see, it takes some effort to establish the notion of $f$-divergence over $\mathcal{R}$. For the total variation distance, getting from (4) to (3) is a matter of simply replacing $\mathcal{B}$ by $\mathcal{R}$. But for general $f$-divergence, $\mathcal{B}$ does not even appear directly in (5) and it is no longer a matter of simply replacing $\mathcal{B}$ by $\mathcal{R}$. Indeed we know of no straightforward answer to Question 1.

In particular, we emphasize that the groundbreaking work of Vapnik and Chervonenkis [55] sheds no light on whether the Glivenko–Cantelli theorem remains true when total variation distance is replaced by an $f$-divergence. They showed that if the VC-dimension of a class $\mathcal{C} \subseteq \Sigma$, i.e., the maximum number of points that $\mathcal{C}$ can shatter, is finite, then $\mathcal{C}$ is a Glivenko–Cantelli class, i.e., $\sup_{A\in\mathcal{C}}\big(\nu_n(A) - \nu(A)\big) \to 0$ almost surely. More generally, existing Vapnik–Chervonenkis theory is still cast entirely in terms of the total variation distance [54]. Indeed, our work in this article suggests that a "Vapnik–Chervonenkis theory for $f$-divergence" is likely within reach. We will establish some initial results in Section 6.

Lastly, readers familiar with Choquet integrals [11, 15] may think that it provides a plausible answer to Question 1, as we did initially; we will explain why it does not in Section 7.

## 2. Notations and terminologies

2.1. **Functions.** We reserve the letter $f$ for $f$-divergence in this article. Other functions will be denoted $g$, $h$, $\varphi$, $\psi$. Depending on context, the notation $X_k \uparrow X$ could mean an increasing sequence $X_1 \le X_2 \le \cdots \le X_k \in \mathbb{R}$ converging to a real value $X \in \mathbb{R}$ or a nested sequence of sets $X_1 \subseteq X_2 \subseteq \cdots \subseteq X_k \subseteq \mathbb{R}$ converging to a set $X \subseteq \mathbb{R}$. Similarly for $X_k \downarrow X$.

We write $\|\cdot\|_p$ for the $L^p$-norm for all $p \in [1,\infty]$ except $p = 2$. The $L^2$-norm will just be denoted as $\|\cdot\|$. We write $L^2(\nu) \coloneqq L^2(\mathbb{R}, \mathcal{B}, \nu)$ for the space of measurable functions with finite $L^2$-norm and

$$L^2_+(\nu) \coloneqq \{g \in L^2(\nu) : g \ge 0\}$$

for the cone of functions nonnegative almost everywhere. As usual, by "function" we mean an equivalence class of functions that differ at most on a measure-zero set and conditions like "$h \ge g$" will always be in the almost everywhere sense without specification. On the other hand, we will specify the measure every time since we often have to deal with multiple measures.

Composition of functions will always be denoted by $\circ$ and pointwise product of real-valued functions will be denoted by $\cdot$ if necessary for emphasis or otherwise left unmarked.

Let $\varnothing \ne M \subseteq L^2(\nu)$ be a closed convex subset. The *metric projection* onto $M$ is the operator $\mathrm{proj}_M : L^2(\nu) \to L^2(\nu)$ that takes $g \in L^2(\nu)$ to the closest point $g_* \in M$, i.e., $\|g - g_*\| = \min_{h\in M}\|h - g\|$. The existence and uniqueness of $g_*$ is guaranteed by the conditions on $M$ and so $\mathrm{proj}_M$ is well-defined. It is also well-known that $\mathrm{proj}_M$ is continuous [2, p. 52].

2.2. **Sequences.** We will remind the readers of the three notions of convergence used in our article: For a sequence of measurable functions $g_n$ converging to $g$ in $L^2(\nu)$. The convergence is said to be (i) $\nu$-almost surely if $g_n$ converges to $g$ pointwise except on a $\nu$-null set; (ii) in $\nu$-measure if $\lim_{n\to\infty} \nu(|g_n - g| > \varepsilon) = 0$ for all $\varepsilon > 0$; (iii) in $L^2$-norm if $\|g_n - g\| \to 0$. The following lemma, adapted from [12, pp. 80–83], summarizes the relations between them.

**Lemma 2.1.** *Let $\nu$ be a Borel probability measure.*

(a) *If $g_n \to g$ in $L^2$-norm, then $g_n \to g$ in $\nu$-measure.*
(b) *If $g_n \to g$ in $\nu$-measure, then there is a subsequence $(g_{n_k})$ such that $g_{n_k} \to g$ almost surely.*

(c) *If $g_n \to g$ almost surely and there exists $g \in L^2(\nu)$ that dominates $g_n$, then $g_n \to g$ in $L^2$-norm.*

We will also need an observation that follows from $|\int g_n - g \,\mathrm{d}\nu| \le \int |g_n - g| \,\mathrm{d}\nu \le \left(\int (g_n - g)^p \,\mathrm{d}\nu\right)^{1/p}$.

**Lemma 2.2.** *Let $g_n \in L^p(\nu)$. If $g_n \to g$ in $L^p$-norm for some $1 \le p < \infty$, then*

$$\int g \,\mathrm{d}\nu = \lim_{n\to\infty} \int g_n \,\mathrm{d}\nu.$$

2.3. **Measures.** We write $2^\Omega$ for the power set of $\Omega$. A collection of subsets $\mathcal{E} \subseteq 2^\Omega$ is called a *$\pi$-system* if it is closed under finite intersections [17, p. 202]. The following results from [12, p. 38] and [43, p. 58] are reproduced here for easy reference.

**Lemma 2.3.** *Let $\mu, \nu$ be finite measures on $(\Omega, \Sigma)$ such that $\mu(\Omega) = \nu(\Omega)$. If $\mathcal{A}$ is a $\pi$-system that generates $\Sigma$ and $\mu(A) = \nu(A)$ for any $A \in \mathcal{A}$, then $\mu = \nu$.*

**Lemma 2.4.** *Let $g : X \to Y$ be any function and $\mathcal{E} \subseteq 2^Y$. Then $\sigma(g^{-1}(\mathcal{E})) = g^{-1}(\sigma(\mathcal{E}))$.*

Whenever we write $\mathrm{d}\mu/\mathrm{d}\nu$, we implicitly assume that $\mu \ll \nu$.

## 3. $f$-DIVERGENCE OVER RAYS

The goal of this section is to resolve Question 1 in the affirmative. The answer is an analogue of $f$-divergence with respect to the class of rays $\mathcal{R}$ that we will call the $f$-divergence over $\mathcal{R}$. This appears only at the end of this section in Definition 3.12. It will take the groundwork developed over the course of the whole section before we can show that the notion is indeed well-defined. There are two milestones to watch for: We will see in Theorem 3.15 that the construction of such a divergence would work for any class $\mathcal{C} \subseteq \mathcal{B}$ that satisfies a certain "Radon–Nikodym property." We will see in Corollary 3.11, which follows from Theorem 3.9, that $\mathcal{C} = \mathcal{R}$ has this property.

We begin by establishing some basic properties related to $\mathcal{R}$. Throughout this section, $\nu$ denotes a Borel probability measure on $(\mathbb{R}, \mathcal{B})$.

**Proposition 3.1.** *The set*

$$G(\mathcal{R}) \coloneqq \{g \in L^2(\nu) : \{g > r\} \in \mathcal{R} \text{ for all } r \in \mathbb{R}\} \tag{6}$$

*is a closed convex cone in $L^2(\nu)$ comprising all nonincreasing functions in $L^2(\nu)$.*

*Proof.* We first show that $G(\mathcal{R})$ is exactly the set of nonincreasing $L^2$-functions. Let $g$ be such a function and let $r \in \mathbb{R}$. Set $a = \sup\{x : g(x) > r\}$. Any $x \in \{g > r\}$ has $x \le a$ so $\{g > r\} \subseteq (-\infty, a]$. If there is an $x < a$ with $g(x) \le r$, then there must be some $y > x$ with $g(y) > r \ge g(x)$. This contradicts the assumption that $g$ is nonincreasing. Hence $(\infty, a) \subseteq \{g > r\} \subseteq (-\infty, a]$. It follows that $\{g > r\} \in \mathcal{R}$ and so $g \in G(\mathcal{R})$. Conversely, let $g \in G(\mathcal{R})$. Suppose we have $x < y$ with $g(x) < g(y)$. Then there is some $a$ with $(-\infty, a) \subseteq \{g > g(x)\} \subseteq (-\infty, a]$. Since $g(y) > g(x)$, it follows that $y \in \{g > g(x)\}$, and so $x < y < a$. Consequently, $x \in \{g > g(x)\}$, i.e., $g(x) > g(x)$, a contradiction. Hence $g$ must be nonincreasing.

$G(\mathcal{R})$ is a closed convex cone because nonnegative linear combinations or $L^2$-limits of nonincreasing functions remain nonincreasing. The former is routine. The latter follows from Lemma 2.1: A sequence $(g_n)$ of nonincreasing functions with $\|g_n - g\| \to 0$ must also converge in $\nu$-measure. So there is a subsequence $(g_{n_k})$ that converges to $g$ almost surely. So there is a null set $A$ where $\lim_{k\to\infty} g_{n_k}(x) = g(x)$ for all $x \notin A$. Since each $g_{n_k}$ is nonincreasing, $g$ is almost surely nonincreasing. □

An immediate consequence of Proposition 3.1 is that $G(\mathcal{R})$ has a well-defined metric projection. For a given $g \in L^2(\nu)$, finding $\mathrm{proj}_{G(\mathcal{R})}\, g$ is equivalent to finding the nearest nonincreasing function to $g$ in $L^2$-norm.

**Definition 3.2.** A collection $\{A_1, \dots, A_n\}$ is called an *ordered partition* of $\mathbb{R}$ if it satisfies the following conditions:

(i) disjoint: $A_i \cap A_j = \varnothing$ for any $i \neq j$;
(ii) cover: $A_1 \cup \cdots \cup A_n = \mathbb{R}$;
(iii) ordered: $\sup A_i \leq \inf A_j$ for each $i \leq j$.

An *ordered simple function* $h : \mathbb{R} \to \mathbb{R}$ is a simple function defined on an ordered partition $\{A_1, \dots, A_n\}$.

It is well-known that any $g \in L^2_+(\nu)$ can be approximated to arbitrary accuracy by an increasing sequence of nonnegative simple functions $(h_n)$. We will show in Corollary 3.4 that $h_n$ can be chosen to be ordered simple functions.

**Proposition 3.3.** *If $g \in L^2_+(\nu)$ is almost surely continuous, then there exists a monotone increasing sequence of nonnegative ordered simple functions $(h_n)$ that converges pointwise to $g$ almost surely.*

*Proof.* Define the ordered partition

$$\mathcal{P}_n := \left\{ \left[ \frac{nk}{2^n}, \frac{n(k+1)}{2^n} \right) : k = -2^n, -2^n + 1, \dots, 2^n - 1 \right\} \cup \left\{ (-\infty, -n), [n, \infty) \right\}$$

and the ordered simple function $h_n = \sum_{A \in \mathcal{P}_n} \inf_{x \in A} g(x) \mathbb{1}_A$. Since $g$ is almost surely continuous, we can find a null set $S$ such that for any $\varepsilon > 0$ and $x \in \mathbb{R} \backslash S$, there exist $n \in \mathbb{N}$ and $A \in \mathcal{P}_n$ with $x \in A \subseteq B_\varepsilon(x)$. Consequently,

$$\inf_{y \in B_\varepsilon(x)} g(y) \leq \inf_{y \in A} g(y) = h_n(x) \leq g(x),$$

where $\inf_{y \in B_\varepsilon(x)} g(y) \uparrow g(x)$ as $\varepsilon \downarrow 0$ since $g$ is continuous on $\mathbb{R} \backslash S$. Therefore, $h_n(x) \to g(x)$ almost surely as $n \to \infty$. For any $n \leq m$, we have $A \subseteq B$ for any $A \in \mathcal{P}_m$ and $B \in \mathcal{P}_n$; so $h_n(x) \leq h_m(x)$ for all $x \in \mathbb{R}$. □

In Proposition 3.3, $h_n$ is dominated by $g$; so we also have $h_n \to g$ in $L^2$-norm by Lemma 2.1. The almost sure continuity in Proposition 3.3 may sometimes be overly restrictive. We derive an alternative version without this requirement.

**Corollary 3.4.** *Let $g \in L^2_+(\nu)$. Then there exists a sequence of bounded ordered simple functions $(h_n)$ that converges to $g$ in $L^2$-norm and almost surely.*

*Proof.* Given that the space of continuous functions with compact support $C_c(\mathbb{R})$ is dense in $L^2(\nu)$ [28, p. 354], there is a continuous function $\varphi$ with compact support that approximates $g$ to accuracy $\|g - \varphi\| < \varepsilon/2$ for any $\varepsilon > 0$. On the other hand, the continuous function $\varphi$ is in $L^2(\nu)$; so by Proposition 3.3 there is an ordered simple function $h$ that approximates it to accuracy $\|\varphi - h\| < \varepsilon/2$. Hence $\|h - g\| < \varepsilon$. This shows that for any $g \in L^2_+(\nu)$, there exists a sequence of ordered simple function $h_n$ such that $h_n \to g$ in $L^2$-norm. By passing through a subsequence if necessary, Lemma 2.1 ensures $h_n \to g$ almost surely as well. □

We next see that the metric projection of an ordered simple function onto $G(\mathcal{R})$ remains ordered simple.

**Proposition 3.5** (Projection)**.** *Let $h = \sum_{i=1}^n \alpha_i \mathbb{1}_{A_i} \in L^2_+(\nu)$ be an ordered simple function with $\{A_1, \dots, A_n\}$ an ordered partition of $\mathbb{R}$. Then $\operatorname{proj}_{G(\mathcal{R})} h$ is also an ordered simple function on $\{A_1, \dots, A_n\}$.*

*Proof.* Suppose that $\operatorname{proj}_{G(\mathcal{R})} h$ is not constant on $A_j$ for some $j = 1, \dots, n$. Define

$$\varphi = \begin{cases} \operatorname{proj}_{G(\mathcal{R})} h & \text{on } A_j^{\mathsf{c}}, \\ c & \text{on } A_j, \end{cases}$$

where $c = \operatorname{proj}_{G(\mathcal{R})} h(\omega_*)$ and $\omega_* = \operatorname{argmin}_{\omega \in A_j} |h(\omega) - \operatorname{proj}_{G(\mathcal{R})} h(\omega)|$. Then $\varphi \in G(\mathcal{R})$ and

$$\int (h - \operatorname{proj}_{G(\mathcal{R})} h)^2 \, \mathrm{d}\nu \geq \int (h - \varphi)^2 \, \mathrm{d}\nu,$$

a contradiction. □

A consequence of Proposition 3.5 is that metric projection onto $G(\mathcal{R})$ preserves ordering.

**Proposition 3.6** (Monotonicity)**.** *Let $g_1, g_2 \in L^2_+(\nu)$. If $g_1 \leq g_2$, then*

$$\operatorname{proj}_{G(\mathcal{R})} g_1 \leq \operatorname{proj}_{G(\mathcal{R})} g_2.$$

*Proof.* We first show that the statement holds for ordered simple functions. Let $h_1 = \sum_{i=1}^n a_i \mathbb{1}_{A_i}$ and $h_2 = \sum_{j=1}^m b_j \mathbb{1}_{B_j}$ where $(A_1, \ldots, A_n)$ and $(B_1, \ldots, B_m)$ are ordered partitions. Suppose $h_1 \leq h_2$ but that $\operatorname{proj}_{G(\mathcal{R})} h_1 > \operatorname{proj}_{G(\mathcal{R})} h_2$ on $E = A_i \cap B_j$ for some $i$ and $j$. Define

$$\varphi = (\operatorname{proj}_{G(\mathcal{R})} h_1) \cdot \mathbb{1}_{E^c} + (\operatorname{proj}_{G(\mathcal{R})} h_2) \cdot \mathbb{1}_E,$$
$$\psi = (\operatorname{proj}_{G(\mathcal{R})} h_2) \cdot \mathbb{1}_{E^c} + (\operatorname{proj}_{G(\mathcal{R})} h_1) \cdot \mathbb{1}_E.$$

Both $\varphi$ and $\psi$ are clearly nonincreasing and thus $\mathcal{R}$-measurable. There are six possibilities for the values of these functions on $E$:

(a) $h_2 \geq h_1 \geq \operatorname{proj}_{G(\mathcal{R})} h_1 \geq \operatorname{proj}_{G(\mathcal{R})} h_2$,
(b) $h_2 \geq \operatorname{proj}_{G(\mathcal{R})} h_1 \geq h_1 \geq \operatorname{proj}_{G(\mathcal{R})} h_2$,
(c) $h_2 \geq \operatorname{proj}_{G(\mathcal{R})} h_1 \geq \operatorname{proj}_{G(\mathcal{R})} h_2 \geq h_1$,
(d) $\operatorname{proj}_{G(\mathcal{R})} h_1 \geq h_2 \geq h_1 \geq \operatorname{proj}_{G(\mathcal{R})} h_2$,
(e) $\operatorname{proj}_{G(\mathcal{R})} h_1 \geq h_2 \geq \operatorname{proj}_{G(\mathcal{R})} h_2 \geq h_1$,
(f) $\operatorname{proj}_{G(\mathcal{R})} h_1 \geq \operatorname{proj}_{G(\mathcal{R})} h_2 \geq h_2 \geq h_1$.

For Cases (a), (b), and (c),

$$\|\psi - h_2\| \leq \|\operatorname{proj}_{G(\mathcal{R})} h_2 - h_2\|;$$

for Cases (e) and (f),

$$\|\varphi - h_1\| \leq \|\operatorname{proj}_{G(\mathcal{R})} h_1 - h_1\|;$$

and for Case (d), either of the following

$$\|\psi - h_2\| \leq \|\operatorname{proj}_{G(\mathcal{R})} h_2 - h_2\|, \quad \|\varphi - h_1\| \leq \|\operatorname{proj}_{G(\mathcal{R})} h_1 - h_1\|$$

must hold; all leading to contradictions.

For the general case with $g_1, g_2 \in L^2_+(\nu)$, we set $g_0 \coloneqq g_2 - g_1 \in L^2_+(\nu)$. By Corollary 3.4, we have two sequences of nonnegative ordered simple functions $(h_{0,n})$ and $(h_{1,n})$ such that $h_{0,n} \to g_0$ and $h_{1,n} \to g_1$ in $L^2$-norm. Set $h_{2,n} \coloneqq h_{0,n} + h_{1,n}$, which is nonnegative, ordered simple, and satisfies $h_{2,n} \to g_2$ in $L^2$-norm and $h_{2,n} \geq h_{1,n}$ for each $n$. By the continuity of the metric projection operator $\operatorname{proj}_{G(\mathcal{R})}$,

$$\operatorname{proj}_{G(\mathcal{R})} g_1 = \lim_{n\to\infty} \operatorname{proj}_{G(\mathcal{R})} h_{1,n} \leq \lim_{n\to\infty} \operatorname{proj}_{G(\mathcal{R})} h_{2,n} = \operatorname{proj}_{G(\mathcal{R})} g_2,$$

as required. □

A slight refinement of the last step of the proof above yields the following.

**Corollary 3.7** (Sequential monotonicity)**.** *Let $g \in L^2_+(\nu)$. If $(h_n)$ is a monotone increasing sequence in $L^2_+(\nu)$ converging to $g$, then $(\operatorname{proj}_{G(\mathcal{R})} h_n)$ is a monotone increasing sequence in $L^2_+(\nu)$ converging to $\operatorname{proj}_{G(\mathcal{R})} g$. Convergence is both in the sense of $L^2$-norm and almost surely.*

*Proof.* Let $h_n$ be a sequence of ordered simple function such that $h_n \to g$ almost surely and in $L^2$-norm. Continuity of $\operatorname{proj}_{G(\mathcal{R})}$ yields

$$\operatorname{proj}_{G(\mathcal{R})} h_n \to \operatorname{proj}_{G(\mathcal{R})} g$$

in $L^2$-norm. By passing to a subsequence if necessary, we may assume that this convergence also holds in the almost sure sense. □

**Proposition 3.8.** *Let $A \in \mathcal{R}$. Then*

$$\operatorname{proj}_{G(\mathcal{R})}(g\mathbb{1}_A) = \operatorname{proj}_{G(\mathcal{R})}(g\mathbb{1}_A) \cdot \mathbb{1}_A \le (\operatorname{proj}_{G(\mathcal{R})} g) \cdot \mathbb{1}_A.$$

*Proof.* We have

$$\begin{aligned}
&\int (g\mathbb{1}_A - \operatorname{proj}_{G(\mathcal{R})}(g\mathbb{1}_A))^2 \,\mathrm{d}\nu \\
&\quad = \int_A (g\mathbb{1}_A - \operatorname{proj}_{G(\mathcal{R})}(g\mathbb{1}_A))^2 \,\mathrm{d}\nu + \int_{A^{\mathrm{c}}} (g\mathbb{1}_A - \operatorname{proj}_{G(\mathcal{R})}(g\mathbb{1}_A))^2 \,\mathrm{d}\nu \\
&\quad = \int_A (g\mathbb{1}_A - \operatorname{proj}_{G(\mathcal{R})}(g\mathbb{1}_A))^2 \,\mathrm{d}\nu + \int_{A^{\mathrm{c}}} (\operatorname{proj}_{G(\mathcal{R})}(g\mathbb{1}_A))^2 \,\mathrm{d}\nu \\
&\quad \ge \int_A (g\mathbb{1}_A - \operatorname{proj}_{G(\mathcal{R})}(g\mathbb{1}_A))^2 \,\mathrm{d}\nu \\
&\quad = \int (g\mathbb{1}_A - \operatorname{proj}_{G(\mathcal{R})}(g\mathbb{1}_A) \cdot \mathbb{1}_A)^2 \,\mathrm{d}\nu.
\end{aligned}$$

Since the projection is unique, we have $\operatorname{proj}_{G(\mathcal{R})}(g\mathbb{1}_A) = \operatorname{proj}_{G(\mathcal{R})}(g\mathbb{1}_A) \cdot \mathbb{1}_A$. As $g\mathbb{1}_A \le g$, the monotonicity in Proposition 3.6 yields

$$\operatorname{proj}_{G(\mathcal{R})}(g\mathbb{1}_A) = \operatorname{proj}_{G(\mathcal{R})}(g\mathbb{1}_A) \cdot \mathbb{1}_A \le (\operatorname{proj}_{G(\mathcal{R})} g) \cdot \mathbb{1}_A. \qquad \square$$

Let $\mathcal{C} \subseteq \mathcal{B}$ be a $\sigma$-subalgebra. The Radon–Nikodym theorem states that for $\sigma$-finite measures $\mu$ and $\nu$ on the measurable space $(\mathbb{R}, \mathcal{B})$ with $\mu \ll \nu$, there exists a $\mathcal{C}$-measurable function $\rho_{\mathcal{C}}$ such that $\mu(A) = \int_A \rho_{\mathcal{C}} \,\mathrm{d}\nu$ for any $A \in \mathcal{C}$.

What happens if $\mathcal{C}$ is not a $\sigma$-subalgebra? For us, the most important case to consider is when $\mathcal{C} = \mathcal{R}$, which is not a $\sigma$-subalgebra. We will show in Corollary 3.11 that $\mathcal{R}$ satisfies a kind of Radon–Nikodym property: There is a $\mathcal{R}$-measurable function $\rho_{\mathcal{R}}$ and a subcollection $\mathcal{E} \subseteq \mathcal{R}$ such that $\mu(A) = \int_A \rho_{\mathcal{R}} \,\mathrm{d}\nu$ for any $A \in \mathcal{E}$. The bulk of the work is in establishing the following technical result.

**Theorem 3.9.** *Let $g \in L^2_+(\nu)$. Define the $\pi$-system $\mathcal{E}(g) \subseteq \mathcal{R}$ by*

$$\mathcal{E}(g) \coloneqq \big\{\{\operatorname{proj}_{G(\mathcal{R})} g > r\} : r \ge 0\big\} \cup \{\mathbb{R}\}. \tag{7}$$

*Then for any $E \in \mathcal{E}(g)$,*

$$\int_E \operatorname{proj}_{G(\mathcal{R})} g \,\mathrm{d}\nu = \int_E g \,\mathrm{d}\nu. \tag{8}$$

*Proof.* We establish the result in four steps, showing that

(i) it holds when $g$ is an ordered simple function $h$;
(ii) it holds for $E = \{\operatorname{proj}_{G(\mathcal{R})} g > r\}$ under the assumption that $\nu(\{\operatorname{proj}_{G(\mathcal{R})} g = r\}) = 0$;
(iii) it extends to all $E = \{\operatorname{proj}_{G(\mathcal{R})} g > r\}$;
(iv) it holds for $E = \mathbb{R}$.

Step (i): First we assume that we have an ordered simple function

$$h = \sum_{i=1}^n \alpha_i \mathbb{1}_{A_i}, \quad \operatorname{proj}_{G(\mathcal{R})} h = \sum_{i=1}^n \beta_i \mathbb{1}_{A_i}.$$

So $E = \{\operatorname{proj}_{G(\mathcal{R})} h > r\}$ can be written as $E = A_1 \cup \cdots \cup A_k \in \mathcal{E}(h)$ where $\beta_k > r$ and $\beta_{k+1} \le r$. Write $\alpha = (\alpha_1, \ldots, \alpha_n)$, $\beta = (\beta_1, \ldots, \beta_n)$, $\omega = (\omega_1, \ldots, \omega_n)$ where $\omega_i = \nu(A_i)$, and let $W = \operatorname{diag}(\omega)$. Then $\beta$ is the solution to the following quadratic programming problem:

$$\begin{aligned}
\text{minimize} \quad & (\alpha - \beta)^{\mathrm{T}} W (\alpha - \beta) \\
\text{subject to} \quad & \beta_1 \ge \cdots \ge \beta_n \ge 0.
\end{aligned}$$

Observe that for

$$\Pi = \begin{bmatrix} 1 & -1 & & \\ & 1 & -1 & \\ & & \ddots & -1 \\ & & & 1 \end{bmatrix}, \qquad \Pi^{-1} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ & 1 & \dots & 1 \\ & & \ddots & \vdots \\ & & & 1 \end{bmatrix}.$$

Let $\gamma = \Pi\beta$. Then the quadratic programming problem above is equivalent to

$$\begin{aligned} \text{minimize} \quad & \gamma^{\mathsf{T}}\Pi^{-\mathsf{T}}W\Pi^{-1}\gamma - 2\alpha^{\mathsf{T}}\Pi^{-1}W\gamma \\ \text{subject to} \quad & -\gamma \preceq 0. \end{aligned}$$

Standard KKT conditions yield

$$\lambda = 2\Pi^{-\mathsf{T}}W(\Pi^{-1}\gamma - \alpha) = 2\Pi^{-\mathsf{T}}W(\beta - \alpha)$$

or

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix} = 2 \begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ \vdots & \vdots & \ddots & \\ 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} \omega_1 & & & \\ & \omega_2 & & \\ & & \ddots & \\ & & & \omega_n \end{bmatrix} \begin{bmatrix} \beta_1 - \alpha_1 \\ \beta_2 - \alpha_2 \\ \vdots \\ \beta_n - \alpha_n \end{bmatrix}. \tag{9}$$

Since $\beta_k = (\Pi^{-1}\gamma)_k = \gamma_k + \gamma_{k+1} + \dots + \gamma_n > r$ and $\beta_{k+1} = \gamma_{k+1} + \gamma_{k+2} + \dots + \gamma_n \le r$, we must have $\gamma_k > 0$. By the KKT complementary slackness condition,

$$\lambda_k = 2(\Pi^{-\mathsf{T}}W(\beta - \alpha))_k = 2\sum_{i=1}^{k} \omega_i(\beta_i - \alpha_i) = 0. \tag{10}$$

Hence we have

$$\int (\operatorname{proj}_{G(\mathcal{R})} h) \cdot \mathbb{1}_E \,\mathrm{d}\nu = \sum_{i=1}^{k} \beta_i\omega_i = \sum_{i=1}^{k} \alpha_i\omega_i = \int h\mathbb{1}_E \,\mathrm{d}\nu.$$

Step (ii): Next we drop the ordered simple assumption and just require that $g \in L^2_+(\nu)$ be a general function with

$$\nu(\{\operatorname{proj}_{G(\mathcal{R})} g = r\}) = 0. \tag{11}$$

Let $(h_n)$ be a sequence of nonnegative ordered simple functions with $h_n \to g$ almost surely and in $L^2$-norm. So $\operatorname{proj}_{G(\mathcal{R})} h_n \to \operatorname{proj}_{G(\mathcal{R})} g$ in $L^2$-norm and thus in measure. By passing through a subsequence if necessary, we may assume that $\operatorname{proj}_{G(\mathcal{R})} h_n \to \operatorname{proj}_{G(\mathcal{R})} g$ almost surely and in $L^2$-norm. Set $E = \{\operatorname{proj}_{G(\mathcal{R})} g > r\}$ and $E_n = \{\operatorname{proj}_{G(\mathcal{R})} h_n > r\}$, bearing in mind that we assume (11). Then

$$\begin{aligned} \|\mathbb{1}_E - \mathbb{1}_{E_n}\|^2 &= \int (\mathbb{1}_E - \mathbb{1}_{E_n})^2 \,\mathrm{d}\nu \\ &= \nu(E) + \nu(E_n) - 2\nu(E \cap E_n) = \nu(E \triangle E_n) = \nu(E \cap E_n^{\mathsf{c}}) + \nu(E_n \cap E^{\mathsf{c}}). \end{aligned}$$

Consider the null set $N = \{\operatorname{proj}_{G(\mathcal{R})} h_n \not\to \operatorname{proj}_{G(\mathcal{R})} g\}$. If

$$x \in E \cap E_n^{\mathsf{c}} = \{\operatorname{proj}_{G(\mathcal{R})} g > r\} \cap \{\operatorname{proj}_{G(\mathcal{R})} h_n \le r\}$$

infinitely often, then

$$\liminf_{n\to\infty} \operatorname{proj}_{G(\mathcal{R})} h_n(x) \le r < \operatorname{proj}_{G(\mathcal{R})} g(x),$$

and so $x \in N$. Consequently, $\limsup_{n\to\infty} E \cap E_n^{\mathsf{c}} \subseteq N$. On the other hand, we may decompose

$$\begin{aligned} E_n \cap E^{\mathsf{c}} &= \big[\{\operatorname{proj}_{G(\mathcal{R})} h_n > r\} \cap \{\operatorname{proj}_{G(\mathcal{R})} g < r\}\big] \cup \big[\{\operatorname{proj}_{G(\mathcal{R})} h_n > r\} \cap \{\operatorname{proj}_{G(\mathcal{R})} g = r\}\big] \\ &=: F_n \cup G_n. \end{aligned}$$

Since $G_n$ is a null set for each $n$, if $x \in F_n$ infinitely often, then

$$\limsup_{n\to\infty} \operatorname{proj}_{G(\mathcal{R})} h_n(x) \geq r > \operatorname{proj}_{G(\mathcal{R})} g(x),$$

and so $x \in N$ as well. Consequently, $\limsup_{n\to\infty} F_n \subseteq N$. We obtain

$$\begin{aligned}\limsup_{n\to\infty} \|\mathbb{1}_E - \mathbb{1}_{E_n}\|^2 &\leq \limsup_{n\to\infty} \big[\nu(E \cap E_n^{\mathsf{c}}) + \nu(F_n) + \nu(G_n)\big] \\ &\leq \limsup_{n\to\infty} \nu(E \cap E_n^{\mathsf{c}}) + \limsup_{n\to\infty} \nu(F_n) \\ &\leq \nu\Big(\limsup_{n\to\infty} E \cap E_n^{\mathsf{c}}\Big) + \nu\Big(\limsup_{n\to\infty} F_n\Big) = 0,\end{aligned}$$

and by Hölder,

$$\begin{aligned}\|h\mathbb{1}_E - h_n\mathbb{1}_{E_n}\|_1 &= \|h\mathbb{1}_E - h\mathbb{1}_{E_n} + h\mathbb{1}_{E_n} - h_n\mathbb{1}_{E_n}\|_1 \\ &\leq \|h \cdot (\mathbb{1}_E - \mathbb{1}_{E_n})\|_1 + \|(h - h_n) \cdot \mathbb{1}_{E_n}\|_1 \\ &\leq \|h\|\|\mathbb{1}_E - \mathbb{1}_{E_n}\| + \|h - h_n\|\|\mathbb{1}_{E_n}\| \to 0.\end{aligned}$$

The same argument also yields

$$(\operatorname{proj}_{G(\mathcal{R})} h_n) \cdot \mathbb{1}_{E_n} \to (\operatorname{proj}_{G(\mathcal{R})} g) \cdot \mathbb{1}_E$$

in $L^1$-norm. By Lemma 2.2, we get

$$\int g\mathbb{1}_E \,\mathrm{d}\nu = \lim_{n\to\infty} \int h_n\mathbb{1}_{E_n} \,\mathrm{d}\nu = \lim_{n\to\infty} \int (\operatorname{proj}_{G(\mathcal{R})} h_n) \cdot \mathbb{1}_{E_n} \,\mathrm{d}\nu = \int (\operatorname{proj}_{G(\mathcal{R})} g) \cdot \mathbb{1}_E \,\mathrm{d}\nu$$

as required.

Step (iii): We next omit the condition (11). Let $x = \inf\{\operatorname{proj}_{G(\mathcal{R})} g = r\}$ and assume $x > -\infty$ for nontriviality. For any $y < x$, we have $\operatorname{proj}_{G(\mathcal{R})} g(y) \geq r$ and so

$$\inf_{y\in(-\infty,x)} \operatorname{proj}_{G(\mathcal{R})} g(y) \geq r. \tag{12}$$

If we have strict inequality in (12), then there exists $\varepsilon > 0$ such that $\inf_{y\in(-\infty,x)} \operatorname{proj}_{G(\mathcal{R})} g(y) > r+\varepsilon$. Thus

$$E_r \coloneqq \{\operatorname{proj}_{G(\mathcal{R})} g > r\} = \{\operatorname{proj}_{G(\mathcal{R})} g > r + \varepsilon\} =: E_{r+\varepsilon}.$$

Since $\operatorname{proj}_{G(\mathcal{R})} g \neq r + \varepsilon$ almost surely, $\nu(\{\operatorname{proj}_{G(\mathcal{R})} g = r + \varepsilon\}) = 0$, and

$$\int_{E_r} \operatorname{proj}_{G(\mathcal{R})} g \,\mathrm{d}\nu = \int_{E_{r+\varepsilon}} \operatorname{proj}_{G(\mathcal{R})} g \,\mathrm{d}\nu = \int_{E_{r+\varepsilon}} g \,\mathrm{d}\nu = \int_{E_r} g \,\mathrm{d}\nu.$$

If we have equality in (12), then for any $k \in \mathbb{N}$, we may pick $x_k \in (-\infty, x)$ such that

$$r < \operatorname{proj}_{G(\mathcal{R})} g(x_k) < r + \frac{1}{k}$$

and $r_k \in (r, \operatorname{proj}_{G(\mathcal{R})} g(x_k)]$ such that

$$r < r_k < r + \frac{1}{k} \quad \text{and} \quad \nu(\operatorname{proj}_{G(\mathcal{R})} g = r_k) = 0.$$

Hence $E_k \coloneqq \{\operatorname{proj}_{G(\mathcal{R})} g > r_k\}$, $k \in \mathbb{N}$, is an increasing sequence of nested sets with $E_k \subseteq E_r$ and so $\bigcup_{k=1}^{\infty} E_k \subseteq E_r$. Conversely, if $y \in E_r$, i.e., $\operatorname{proj}_{G(\mathcal{R})} g(y) > r$, then there exists $r_k$ such that $r_k < \operatorname{proj}_{G(\mathcal{R})} g(y)$; so $y \in E_k$; and so $E_k \uparrow E_r$. Define the finite measures

$$\zeta(A) \coloneqq \int_A \operatorname{proj}_{G(\mathcal{R})} g \,\mathrm{d}\nu, \qquad \xi(A) \coloneqq \int_A g \,\mathrm{d}\nu$$

for any $A \in \mathcal{B}$. By the continuity of measure from below,

$$\zeta(E_r) = \lim_{k\to\infty} \zeta(E_k) = \lim_{k\to\infty} \xi(E_k) = \xi(E_r)$$

as required.

STEP (iv): It remains to treat the case $E = \mathbb{R}$. We will show that $\int \operatorname{proj}_{G(\mathcal{R})} g \,\mathrm{d}\nu = \int g \,\mathrm{d}\nu$ by first establishing it for an ordered simple function $h = \sum_{i=1}^n \alpha_i \mathbb{1}_{A_i}$. We may assume that $\beta_k > 0$ and $\beta_{k+1} = \beta_{k+2} = \cdots = \beta_n = 0$ without loss of generality. By (9) and Step (i),

$$\begin{aligned}\lambda_n &= 2\sum_{i=1}^n \omega_i(\beta_i - \alpha_i) = 2\sum_{i=1}^k \omega_i(\beta_i - \alpha_i) + 2\sum_{j=k+1}^n \omega_i(\beta_i - \alpha_i)\\ &= -\lambda_k - 2(\omega_{k+1}\alpha_{k+1} + \omega_{k+2}\alpha_{k+2} + \cdots + \omega_n\alpha_n).\end{aligned}$$

By (10), $\lambda_k = 0$. By KKT dual feasibility, $\lambda_n \geq 0$. Since we also have $\omega_i, \alpha_i \geq 0$, $i = k+1, k+2, \ldots, n$,

$$\lambda_n = -2(\omega_{k+1}\alpha_{k+1} + \omega_{k+2}\alpha_{k+2} + \cdots + \omega_n\alpha_n) = 0.$$

It follows from (9) that $0 = \lambda_n = 2\sum_{i=1}^n \omega_i(\beta_i - \alpha_i)$ and so

$$\int \operatorname{proj}_{G(\mathcal{R})} h \,\mathrm{d}\nu = \sum_{i=1}^n \beta_i w_i = \sum_{i=1}^n \alpha_i w_i = \int h \,\mathrm{d}\nu.$$

For a general function $g \in L^2_+(\nu)$, pick a sequence of nonnegative ordered simple functions $h_n \to g$ in $L^2$-norm and apply Lemma 2.1 to get

$$\int g \,\mathrm{d}\nu = \lim_{n\to\infty} \int h_n \,\mathrm{d}\nu = \lim_{n\to\infty} \int \operatorname{proj}_{G(\mathcal{R})} h_n \,\mathrm{d}\nu = \int \operatorname{proj}_{G(\mathcal{R})} g \,\mathrm{d}\nu. \qquad \square$$

We state an immediate corollary of Theorem 3.9.

**Corollary 3.10.** *Let $g \in L^2_+(\nu)$. If $\operatorname{proj}_{G(\mathcal{R})} g(E) \in \mathcal{B}$, then*

$$\int_E \operatorname{proj}_{G(\mathcal{R})} g \,\mathrm{d}\nu = \int_E g \,\mathrm{d}\nu.$$

*Proof.* Theorem 3.9 says that (8) holds for any $E \in \mathcal{E}(g)$. The left and right sides of (8) define finite measures $\zeta$ and $\xi$ respectively, and $\zeta(\mathbb{R}) = \xi(\mathbb{R})$ as in Step (iii) of the proof of Theorem 3.9. Since $\mathcal{E}(g)$ is a $\pi$-system, $\sigma(\mathcal{E}(g)) = (\operatorname{proj}_{G(\mathcal{R})} g)^{-1}(\mathcal{B})$ by Lemmas 2.3 and 2.4. Hence (8) holds for all $E \in (\operatorname{proj}_{G(\mathcal{R})} g)^{-1}(\mathcal{B})$. $\square$

We now arrive at the Radon–Nikodym property for $\mathcal{R}$ mentioned earlier.

**Corollary 3.11** (Radon–Nikodym over rays)**.** *Let $\mu, \nu$ be Borel probability measures on $(\mathbb{R}, \mathcal{B})$. Then the $\mathcal{R}$-measurable function*

$$\rho_{\mathcal{R}} = \operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu},$$

*and the $\pi$-system $\mathcal{E}(\mathrm{d}\mu/\mathrm{d}\nu) = \big\{\{\rho_{\mathcal{R}} > r\} : r \geq 0\big\} \cup \{\mathbb{R}\} \subseteq \mathcal{R}$ satisfy*

$$\mu(E) = \int_E \rho_{\mathcal{R}} \,\mathrm{d}\nu$$

*for any $E \in \mathcal{E}(\mathrm{d}\mu/\mathrm{d}\nu)$.*

*Proof.* For any $E \in \mathcal{E}(\mathrm{d}\mu/\mathrm{d}\nu)$, it follows from Theorem 3.9 that

$$\mu(E) = \int_E \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \,\mathrm{d}\nu = \int_E \operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \,\mathrm{d}\nu = \int_E \rho_{\mathcal{R}} \,\mathrm{d}\nu. \qquad \square$$

We will show that the following definition of $f$-divergence over $\mathcal{R}$ provides an affirmative answer to Question 1: Theorem 3.15 shows that (a) when $f$ is chosen to be $f(t) = |t-1|/2$, we recover the Kolmogorov–Smirnov distance in (3); and Definition 3.12 shows that by replacing $\mathcal{R}$ by $\mathcal{B}$, we recover the standard $f$-divergence in (5). A reminder of our convention in Section 2.3: we assume

that $\mu \ll \nu$ whenever we write $\mathrm{d}\mu/\mathrm{d}\nu$. This is not an additional imposition; it is also a requirement for standard $f$-divergence in Definition 1.5.

**Definition 3.12** ($f$-divergence over rays)**.** Let $\mu$ and $\nu$ be Borel probability measures and $\mathcal{R}$ be the class of rays. Then the $f$-divergence over $\mathcal{R}$ is defined as

$$\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu) \coloneqq \int f\Big(\mathrm{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\Big)\, \mathrm{d}\nu. \tag{13}$$

Henceforth we will use this terminology in conjunction with standard nomenclatures. For example when we speak of total variation or Hellinger distance over $\mathcal{R}$ we mean the $f$-divergence over $\mathcal{R}$ with $f(t) = |t-1|/2$ or $f(t) = (\sqrt{t}-1)^2$ respectively. Obviously, if we replace $\mathcal{R}$ in (13) by the Borel $\sigma$-algebra $\mathcal{B}$, then $\mathrm{D}_f^{\mathcal{B}}(\mu\|\nu) = \mathrm{D}_f(\mu\|\nu)$ and we recover the standard $f$-divergence in Definition 1.5. So we may view the standard definition as "$f$-divergence over $\mathcal{B}$."

We will provide abundant evidence in Section 4 that our $f$-divergence over $\mathcal{R}$ in Definition 3.12 retains almost all known properties of the standard $f$-divergence. But our most rudimentary justification is Theorem 3.15, i.e., the total variation distance on $\mathcal{R}$ is exactly the Kolmogorov–Smirnov distance. To get there we need two corollaries.

**Corollary 3.13.** *Let $g \in L^2_+(\nu)$. Then for any $A \in \mathcal{R}$,*

$$\int_A \mathrm{proj}_{G(\mathcal{R})}\, g\, \mathrm{d}\nu \geq \int_A g\, \mathrm{d}\nu.$$

*Proof.* By Proposition 3.8 and Theorem 3.9, we have

$$\int_A \mathrm{proj}_{G(\mathcal{R})}\, g\, \mathrm{d}\nu = \int (\mathrm{proj}_{G(\mathcal{R})}\, g) \cdot \mathbb{1}_A\, \mathrm{d}\nu \geq \int \mathrm{proj}_{G(\mathcal{R})}(g\mathbb{1}_A)\, \mathrm{d}\nu = \int_A g\, \mathrm{d}\nu. \qquad \square$$

The $\pi$-system $\mathcal{E}(g)$ as defined in (7) has the following property.

**Corollary 3.14.** *Let $g \in L^2_+(\nu)$. Then for any $E \in \mathcal{E}(g)$,*

$$(\mathrm{proj}_{G(\mathcal{R})}\, g) \cdot \mathbb{1}_E = \mathrm{proj}_{G(\mathcal{R})}(g\mathbb{1}_E).$$

*Proof.* By Theorem 3.9,

$$\int (\mathrm{proj}_{G(\mathcal{R})}\, g) \cdot \mathbb{1}_E\, \mathrm{d}\nu = \int g\mathbb{1}_E\, \mathrm{d}\nu = \int \mathrm{proj}_{G(\mathcal{R})}(g\mathbb{1}_E)\, \mathrm{d}\nu.$$

Since $(\mathrm{proj}_{G(\mathcal{R})}\, g) \cdot \mathbb{1}_E \geq \mathrm{proj}_{G(\mathcal{R})}(g\mathbb{1}_E)$, we must in fact have equality. $\square$

We now arrive at the main result of this section. The key to Theorem 3.15 is Theorem 3.9. If we have an analogue of Theorem 3.9 for a class of sets $\mathcal{C}$, then we have an "$f$-divergence on $\mathcal{C}$" for which an analogue of Theorem 3.15 with $\mathcal{C}$ in place of $\mathcal{R}$ holds.

**Theorem 3.15** (Kolmogorov–Smirnov as special case)**.** *Let $\mu, \nu$ be Borel probability measures on $(\mathbb{R}, \mathcal{B})$. Then*

$$\int \frac{1}{2}\Big|\mathrm{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\Big|\, \mathrm{d}\nu = \sup_{A \in \mathcal{R}} \big(\mu(A) - \nu(A)\big). \tag{14}$$

*Proof.* Let $E = \{\mathrm{proj}_{G(\mathcal{R})}\, \mathrm{d}\mu/\mathrm{d}\nu > 1\}$. By Theorem 3.9 and Corollary 3.11,

$$\int_A \mathrm{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\, \mathrm{d}\nu = \int_A \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\, \mathrm{d}\nu = \mu(A)$$

for any $A \in \mathcal{E}(\mathrm{d}\mu/\mathrm{d}\nu)$. So

$$\begin{aligned}\int \frac{1}{2}\Big|\operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\Big| \,\mathrm{d}\nu &= \frac{1}{2}\int_E \operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1 \,\mathrm{d}\nu + \frac{1}{2}\int_{E^{\mathrm{c}}} 1 - \operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \,\mathrm{d}\nu \\ &= \frac{1}{2}\Big[2\int_E \operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \,\mathrm{d}\nu - 1\Big] - \frac{1}{2}\Big[2\int_E \mathrm{d}\nu - 1\Big] \\ &= \int_E \operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \,\mathrm{d}\nu - \nu(E) \\ &= \mu(E) - \nu(E) \le \sup_{A\in\mathcal{R}} \big(\mu(A) - \nu(A)\big),\end{aligned} \tag{15}$$

noting that $E = \{\operatorname{proj}_{G(\mathcal{R})} \mathrm{d}\mu/\mathrm{d}\nu > 1\} \in \mathcal{R}$ as $\operatorname{proj}_{G(\mathcal{R})} \mathrm{d}\mu/\mathrm{d}\nu \in G(\mathcal{R})$. For the reverse inequality, by Corollary 3.13, we have

$$\int_A \Big(\operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\Big)\mathrm{d}\nu \ge \mu(A) - \nu(A),$$

for any $A \in \mathcal{R}$. By (15),

$$\begin{aligned}\int \frac{1}{2}\Big|\operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\Big| \,\mathrm{d}\nu &= \int_E \Big(\operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\Big)\mathrm{d}\nu \\ &= \sup_{A\in\mathcal{R}} \Big(\int_A \operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \,\mathrm{d}\nu - \nu(A)\Big) \ge \sup_{A\in\mathcal{R}} \big(\mu(A) - \nu(A)\big)\end{aligned}$$

as required. □

## 4. Properties of the $f$-divergences over rays

Theorem 3.15 serves as an indication that our notion of $f$-divergence over rays in Definition 3.12 is the right one because it restricts to the Kolmogorov–Smirnov distance for a special choice of $f$. In this section, we will see that just about every property that we know holds for $f$-divergences also holds for $f$-divergences over $\mathcal{R}$. Henceforth we will denote the convex cone of functions convex on $[0,\infty)$ and both vanishing and strictly convex at 1 by

$$\mathcal{F} \coloneqq \{f : [0,\infty) \to \mathbb{R} : f \text{ is convex, } f \text{ is strictly convex at 1, and } f(1) = 0\}. \tag{16}$$

The strict convexity at 1 is as in [40, p. 120]: For all $x, y \in [0,\infty)$ and $t \in (0,1)$ such that $tx + (1-t)y = 1$, we have $tf(x) + (1-t)f(y) > 0$. This condition guarantees that $\mathrm{D}_f(\mu\|\nu) = 0$ if and only if $\mu = \nu$. As a precaution, in situations like the Kullback–Leibler divergence where $f(x) = x\log x$, the value $f(0)$ may only be defined in a limiting sense.

Again $\mu, \nu$ will denote Borel probability measures on $(\mathbb{R}, \mathcal{B})$ throughout this section, with the implicit assumption that $\mu \ll \nu$ whenever we speak of their Radon–Nikodym derivative $\mathrm{d}\mu/\mathrm{d}\nu$.

**Theorem 4.1.** *Let $f, g \in \mathcal{F}$. Then the $f$-divergence over $\mathcal{R}$ has the following properties:*

(i) *Linearity:* $\mathrm{D}^{\mathcal{R}}_{\alpha f+\beta g}(\mu\|\nu) = \alpha\,\mathrm{D}^{\mathcal{R}}_f(\mu\|\nu) + \beta\,\mathrm{D}^{\mathcal{R}}_g(\mu\|\nu)$ *for any* $\alpha, \beta \ge 0$.
(ii) *Nonnegativity:* $\mathrm{D}^{\mathcal{R}}_f(\mu\|\nu) \ge 0$.
(iii) *Affine invariance: If* $g(x) = f(x) + c(x-1)$, $c \in \mathbb{R}$, *then* $\mathrm{D}^{\mathcal{R}}_f(\mu\|\nu) = \mathrm{D}^{\mathcal{R}}_g(\mu\|\nu)$.
(iv) *Boundedness:* $\mathrm{D}^{\mathcal{R}}_f(\mu\|\nu) \le \mathrm{D}_f(\mu\|\nu)$.
(v) *Identity: If* $\mathrm{D}^{\mathcal{R}}_f(\mu\|\nu) = 0 = \mathrm{D}^{\mathcal{R}}_f(\nu\|\mu)$, *then* $\mu = \nu$.

*Proof.* Item (i) is routine from definition. For item (ii), since $f$ is convex, applying Jensen's inequality and Theorem 3.9 gives

$$\int f\Big(\operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\Big)\,\mathrm{d}\nu \ge f\Big(\int \operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\,\mathrm{d}\nu\Big) = f\Big(\int \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\,\mathrm{d}\nu\Big) = f(1) = 0.$$

Item (iii) follows from

$$\begin{aligned}\mathrm{D}_g^{\mathcal{R}}(\mu\|\nu) &= \int f\Big(\mathrm{proj}_{G(\mathcal{R})}\,\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\Big) + c\Big(\mathrm{proj}_{G(\mathcal{R})}\,\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\Big)\,\mathrm{d}\nu \\ &= \int f\Big(\mathrm{proj}_{G(\mathcal{R})}\,\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\Big)\,\mathrm{d}\nu + c\int\Big(\mathrm{proj}_{G(\mathcal{R})}\,\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\Big)\mathrm{d}\nu = \mathrm{D}_f^{\mathcal{R}}(\mu\|\nu),\end{aligned}$$

noting that the last integral vanishes. The proofs for items (iv) and (v) are considerably more involved and will be deferred to Proposition 4.3 and Corollary 4.7 respectively. □

We caution the reader that special properties of a specific $f$-divergence may sometimes be lost for its counterpart over $\mathcal{R}$. For example, it is well-known that while $f$-divergence is generally not symmetric, i.e., $\mathrm{D}_f(\mu\|\nu) \neq \mathrm{D}_f(\nu\|\mu)$, the total variation distance is. This symmetry is lost for the total variation distance over $\mathcal{R}$ as

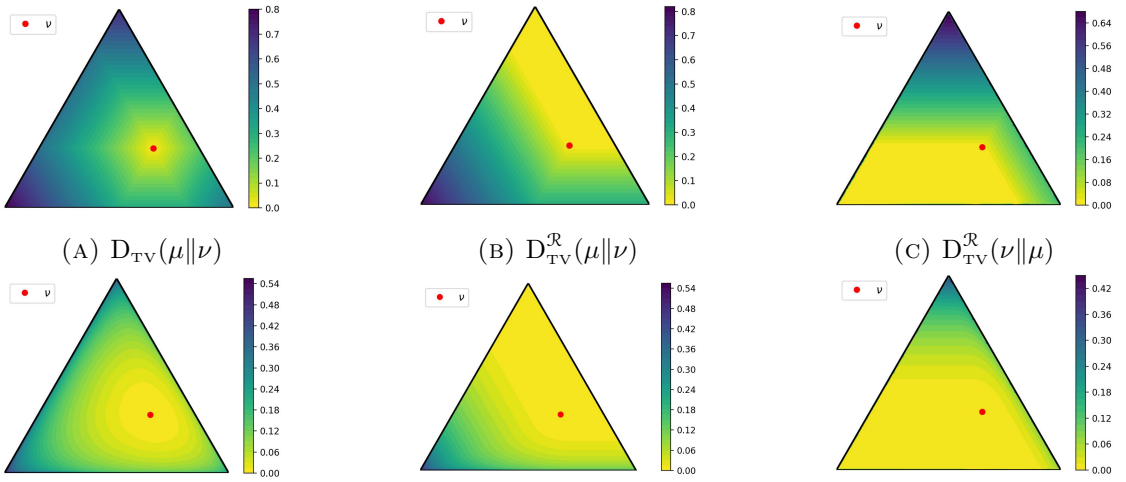
$$\mathrm{D}_{\mathrm{TV}}^{\mathcal{R}}(\mu\|\nu) = \int \frac{1}{2}\Big|\mathrm{proj}_{G(\mathcal{R})}\,\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\Big|\,\mathrm{d}\nu = \sup_{A\in\mathcal{R}}\big(\mu(A) - \nu(A)\big);$$

although one may recover it by symmetrizing:

$$\widehat{\mathrm{D}}_{\mathrm{TV}}^{\mathcal{R}}(\mu\|\nu) \coloneqq \max\big\{\mathrm{D}_{\mathrm{TV}}^{\mathcal{R}}(\mu\|\nu), \mathrm{D}_{\mathrm{TV}}^{\mathcal{R}}(\nu\|\mu)\big\} = \sup_{A\in\mathcal{R}}|\mu(A) - \nu(A)|. \tag{17}$$

Strictly speaking, it is this symmetrized version $\widehat{\mathrm{D}}_{\mathrm{TV}}^{\mathcal{R}}$ that should be called the Kolmogorov–Smirnov *distance* [30] although we have been loosely using the term for the *divergence* $\mathrm{D}_{\mathrm{TV}}^{\mathcal{R}}$ as well.

This is also the reason why we need both $\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu) = 0$ and $\mathrm{D}_f^{\mathcal{R}}(\nu\|\mu) = 0$ in Theorem 4.1(v). It would not be true otherwise. We illustrate this in Figure 1 with the total variation and Hellinger distances over $\mathcal{R}$. One can see that $\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu) = 0$ or $\mathrm{D}_f^{\mathcal{R}}(\nu\|\mu) = 0$ does not necessarily imply $\mu = \nu$ but that $\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu) = \mathrm{D}_f^{\mathcal{R}}(\nu\|\mu) = 0$ does.




(A) $\mathrm{D}_{\mathrm{TV}}(\mu\|\nu)$ (B) $\mathrm{D}_{\mathrm{TV}}^{\mathcal{R}}(\mu\|\nu)$ (C) $\mathrm{D}_{\mathrm{TV}}^{\mathcal{R}}(\nu\|\mu)$

(D) $\mathrm{D}_{\mathrm{H}}(\mu\|\nu)$ (E) $\mathrm{D}_{\mathrm{H}}^{\mathcal{R}}(\mu\|\nu)$ (F) $\mathrm{D}_{\mathrm{H}}^{\mathcal{R}}(\nu\|\mu)$

FIGURE 1. 40 Level curves of total variation $\mathrm{D}_{\mathrm{TV}}$, total variation over rays $\mathrm{D}_{\mathrm{TV}}^{\mathcal{R}}$, Hellinger distance $\mathrm{D}_{\mathrm{H}}$ and Hellinger distance over rays $\mathrm{D}_{\mathrm{H}}^{\mathcal{R}}$ for fixed $\nu = [0.2, 0.5, 0.3]$ as $\mu$ ranges over the simplex of distributions on a three-element set.

Evidently, the projection operator $\mathrm{proj}_{G(\mathcal{R})}$ plays a key role in our definition of $f$-divergence over $\mathcal{R}$. The next two results show how it interacts with nondecreasing functions and with convex functions.

**Lemma 4.2.** *Let $f : \mathbb{R} \to \mathbb{R}$ be nondecreasing. If $g \in L^2_+(\nu)$ is almost surely continuous, then*

$$\int (\operatorname{proj}_{G(\mathcal{R})} g) \cdot (f \circ \operatorname{proj}_{G(\mathcal{R})} g) \, \mathrm{d}\nu = \int g \cdot (f \circ \operatorname{proj}_{G(\mathcal{R})} g) \, \mathrm{d}\nu.$$

*Proof.* Let $h = \sum_{i=1}^n \alpha_i \mathbb{1}_{A_i}$ and $\operatorname{proj}_{G(\mathcal{R})} h = \sum_{i=1}^n \beta_i \mathbb{1}_{A_i}$. Suppose $\operatorname{proj}_{G(\mathcal{R})} h$ takes $m$ distinct values $\tau_1 > \cdots > \tau_m$ where $m \le n$. For each $k = 1, \dots, m$, set

$$E_k \coloneqq \bigcup_{i=1}^k \{\operatorname{proj}_{G(\mathcal{R})} h = \tau_i\} \in \mathcal{E}(h).$$

Let $\zeta_m \coloneqq f(\tau_m)$ and $\zeta_{m-j} \coloneqq f(\tau_{m-j}) - \zeta_{m-j+1} - \zeta_{m-j+2} - \cdots - \zeta_m$, $j = 1, \dots, m-1$. Then

$$\begin{aligned}
\int (\operatorname{proj}_{G(\mathcal{R})} h) \cdot (f \circ \operatorname{proj}_{G(\mathcal{R})} h) \, \mathrm{d}\nu &= \int (\operatorname{proj}_{G(\mathcal{R})} h) \cdot \Big( \sum_{i=1}^n f(\beta_i) \mathbb{1}_{A_i} \Big) \, \mathrm{d}\nu \\
&= \int (\operatorname{proj}_{G(\mathcal{R})} h) \cdot \sum_{i=1}^m \zeta_i \mathbb{1}_{E_i} \, \mathrm{d}\nu \\
&= \sum_{i=1}^m \zeta_i \int_{E_i} \operatorname{proj}_{G(\mathcal{R})} h \, \mathrm{d}\nu = \sum_{i=1}^m \zeta_i \int_{E_i} h \, \mathrm{d}\nu \\
&= \int h \cdot \sum_{i=1}^m \zeta_i \mathbb{1}_{E_i} \, \mathrm{d}\nu = \int h \cdot (f \circ \operatorname{proj}_{G(\mathcal{R})} h) \, \mathrm{d}\nu.
\end{aligned}$$

Now let $h_n$ be a nondecreasing sequence of nonnegative ordered simple function that converges to $g$ almost surely and in $L^2$-norm. By passing through a subsequence if necessary, $\operatorname{proj}_{G(\mathcal{R})} h_n$ is a nonincreasing sequence that converges to $\operatorname{proj}_{G(\mathcal{R})} g$ almost surely. Since $f$ is nondecreasing,

$$f \circ \operatorname{proj}_{G(\mathcal{R})} h_n \uparrow f \circ \operatorname{proj}_{G(\mathcal{R})} g$$

almost surely. By the monotone convergence theorem,

$$\begin{aligned}
\int (\operatorname{proj}_{G(\mathcal{R})} g) \cdot (f \circ \operatorname{proj}_{G(\mathcal{R})} g) \, \mathrm{d}\nu &= \lim_{n\to\infty} \int (\operatorname{proj}_{G(\mathcal{R})} h_n) \cdot (f \circ \operatorname{proj}_{G(\mathcal{R})} h_n) \, \mathrm{d}\nu \\
&= \lim_{n\to\infty} \int h_n \cdot (f \circ \operatorname{proj}_{G(\mathcal{R})} h_n) \, \mathrm{d}\nu = \int g \cdot (f \circ \operatorname{proj}_{G(\mathcal{R})} g) \, \mathrm{d}\nu. \qquad \square
\end{aligned}$$

**Proposition 4.3.** *Let $f : \mathbb{R} \to \mathbb{R}$ be convex. If $g \in L^2_+(\nu)$ is almost surely continuous, then*

$$\int f \circ \operatorname{proj}_{G(\mathcal{R})} g \, \mathrm{d}\nu \le \int f \circ g \, \mathrm{d}\nu.$$

*Proof.* By convexity,

$$f\big(\operatorname{proj}_{G(\mathcal{R})} g(x)\big) - f\big(g(x)\big) \le f'\big(\operatorname{proj}_{G(\mathcal{R})} g(x)\big) \cdot \big(\operatorname{proj}_{G(\mathcal{R})} g(x) - g(x)\big)$$

almost surely. Integrating with respect to $\nu$ and applying Lemma 4.2 to $f'$, we get

$$\int f \circ \operatorname{proj}_{G(\mathcal{R})} g - f \circ g \, \mathrm{d}\nu \le \int f'(\operatorname{proj}_{G(\mathcal{R})} g) \cdot (\operatorname{proj}_{G(\mathcal{R})} g - g) \, \mathrm{d}\nu = 0. \qquad \square$$

Theorem 4.1(iv) follows from Proposition 4.3 since any $f \in \mathcal{F}$ is convex and so

$$\mathrm{D}_f^{\mathcal{R}}(\nu \| \mu) = \int f\Big(\operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\Big) \, \mathrm{d}\nu \le \int f\Big(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\Big) \, \mathrm{d}\nu = \mathrm{D}_f(\nu \| \mu).$$

It remains to establish Theorem 4.1(v), whose proof will take four intermediate results: Propositions 4.4 and 4.5, Corollaries 4.6 and 4.7, each of independent interest. We begin by showing how we may convert $f$-divergence over rays to standard $f$-divergence.

**Proposition 4.4.** *For any Borel probability measures $\mu$ and $\nu$, there exists a Borel probability measure $\zeta$ such that for any $f \in \mathcal{F}$,*

$$\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu) = \mathrm{D}_f(\zeta\|\nu).$$

*Proof.* We define $\zeta$ by

$$\zeta(B) \coloneqq \int_B \operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\, \mathrm{d}\nu$$

for any $B \in \mathcal{B}$. By Corollary 3.11, $\zeta$ is a probability measure absolutely continuous with respect to $\nu$. Hence

$$\frac{\mathrm{d}\zeta}{\mathrm{d}\nu} = \operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}$$

and so

$$\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu) = \int f\Big(\operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\Big)\, \mathrm{d}\nu = \int f\Big(\frac{\mathrm{d}\zeta}{\mathrm{d}\nu}\Big)\, \mathrm{d}\nu = \mathrm{D}_f(\zeta\|\nu). \qquad \square$$

Next we do the reverse: converting standard $f$-divergence to $f$-divergence over rays.

**Proposition 4.5.** *For any Borel probability measures $\mu$ and $\nu$, there exist Borel probability measures $\eta, \tau$ such that for any $f \in \mathcal{F}$,*

$$\mathrm{D}_f(\mu\|\nu) = \mathrm{D}_f^{\mathcal{R}}(\eta\|\tau).$$

*Proof.* The distribution function[1] $m : [0,\infty) \to [0,1]$ of $\mathrm{d}\mu/\mathrm{d}\nu$ is given by

$$m(\alpha) \coloneqq \nu\Big(\Big\{x \in \Omega : \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x) > \alpha\Big\}\Big)$$

for any $\alpha \in [0,\infty)$. Since $m$ may not be injective, we define its generalized inverse $\rho : [0,1] \to [0,\infty)$,

$$\rho(u) \coloneqq \inf\{\alpha \in [0,\infty) : m(\alpha) \le u\}$$

for any $u \in [0,1]$. Thus $\rho$ may be viewed a decreasing rearrangement of $\mathrm{d}\mu/\mathrm{d}\nu$ on $[0,1]$ satisfying

$$\nu\Big(\Big\{x \in \Omega : \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x) > \alpha\Big\}\Big) = \lambda\big(\{u \in [0,1] : \rho(u) > \alpha\}\big),$$

with $\lambda$ the Lebesgue measure on $[0,1]$ [20, p. 199]. Let $T : [0,1] \to \mathbb{R}$ be a strictly increasing measure-preserving transformation. Define the probability measures

$$\tau(A) \coloneqq \lambda(T^{-1}(A)), \qquad \eta(A) \coloneqq \int_{T^{-1}(A)} \rho\, \mathrm{d}\lambda$$

for any $A \in \mathcal{B}$. Since $\rho$ is nonincreasing and $T^{-1}$ is increasing, the Radon–Nikodym derivative

$$\frac{\mathrm{d}\eta}{\mathrm{d}\tau} = \rho \circ T^{-1}$$

is also nonincreasing on $\mathbb{R}$, so $\mathrm{d}\eta/\mathrm{d}\tau = \operatorname{proj}_{G(\mathcal{R})} \mathrm{d}\eta/\mathrm{d}\tau$. Hence, for any $f \in \mathcal{F}$,

$$\mathrm{D}_f(\mu\|\nu) = \int f\Big(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\Big)\, \mathrm{d}\nu = \int_{[0,1]} f(\rho)\, \mathrm{d}\lambda = \int f(\rho \circ T^{-1})\, \mathrm{d}\tau = \int f\Big(\operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\eta}{\mathrm{d}\tau}\Big)\, \mathrm{d}\tau = \mathrm{D}_f^{\mathcal{R}}(\eta\|\tau)$$

as required. $\square$

Collectively, Propositions 4.4 and 4.5 lead to the following conclusion: An inequality relation between two divergences hold if and only if that same relation hold for their counterparts over rays.

[1] In this article, we use the *decumulative* distribution function instead of the more common cumulative distribution function. They are equivalent — we could have used the latter but it introduces more notational clutter. Another advantage is consistency with our discussion of Choquet integrals in Section 7.

**Corollary 4.6** (Preservation of inequality relations)**.** *Let $f, g \in \mathcal{F}$ and $\psi, \varphi : [0,\infty) \to \mathbb{R}$. Then*

$$\psi(\mathrm{D}_f(\mu\|\nu)) \le \varphi(\mathrm{D}_g(\mu\|\nu))$$

*holds for all Borel probability measures $\mu$ and $\nu$ if and only if*

$$\psi(\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu)) \le \varphi(\mathrm{D}_g^{\mathcal{R}}(\mu\|\nu))$$

*holds for all Borel probability measures $\mu$ and $\nu$.*

*Proof.* Suppose $\psi(\mathrm{D}_f(\mu\|\nu)) \le \varphi(\mathrm{D}_g(\mu\|\nu))$ for all $\mu$ and $\nu$. By Proposition 4.4, there exists $\zeta$ such that

$$\psi(\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu)) = \psi(\mathrm{D}_f(\zeta\|\nu)) \le \varphi(\mathrm{D}_g(\zeta\|\nu)) = \varphi(\mathrm{D}_g^{\mathcal{R}}(\mu\|\nu)).$$

Conversely, suppose $\psi(\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu)) \le \varphi(\mathrm{D}_g^{\mathcal{R}}(\mu\|\nu))$ for all $\mu$ and $\nu$. By Proposition 4.5, there exist $\eta$ and $\tau$ such that

$$\psi(\mathrm{D}_f(\mu\|\nu)) = \psi(\mathrm{D}_f^{\mathcal{R}}(\eta\|\tau)) \le \varphi(\mathrm{D}_g^{\mathcal{R}}(\eta\|\tau)) = \varphi(\mathrm{D}_g(\mu\|\nu)).$$ □

As most readers familiar with $f$-divergences would suspect, the total variation distance plays a somewhat special role among all $f$-divergences. One striking property [7] is that it serves as a kind of universal lower bound for all $f$-divergence, in the sense that

$$\bar{f}\big(\mathrm{D}_{\mathrm{TV}}(\mu\|\nu)\big) \le \mathrm{D}_f(\mu\|\nu),$$

where $\bar{f}(x) \coloneqq f(1+x) + f(1-x)$ is the symmetrized function of $f$. Note that $\mathrm{D}_{\mathrm{TV}}(\mu\|\nu)$ takes value in $[0,1]$, so the $f(1-x)$ term is always evaluated within the domain of $f \in \mathcal{F}$. A consequence of our definition of $\mathcal{F}$ in (16) is that $\bar{f}$ is increasing and $\bar{f}(x) > 0$ when $x > 0$. So by Corollary 4.6, the same result holds verbatim for divergences over rays.

**Corollary 4.7** (Total variation over rays as universal lower bound)**.** *Let $f \in \mathcal{F}$ and set $\bar{f}(x) \coloneqq f(1+x) + f(1-x)$ for $x \in [0,1]$. Then*

$$\bar{f}\big(\mathrm{D}_{\mathrm{TV}}^{\mathcal{R}}(\mu\|\nu)\big) \le \mathrm{D}_f^{\mathcal{R}}(\mu\|\nu).$$

Theorem 4.1(v) now follows directly from Corollary 4.7: If $\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu) = 0$, then $\bar{f}\big(\mathrm{D}_{\mathrm{TV}}^{\mathcal{R}}(\mu\|\nu)\big) = 0$. So $\mathrm{D}_{\mathrm{TV}}^{\mathcal{R}}(\mu\|\nu) = 0$ as $\bar{f}$ is strictly increasing and $\bar{f}(0) = 0$. Since we have also assumed that $\mathrm{D}_f^{\mathcal{R}}(\nu\|\mu) = 0$, we get $\mathrm{D}_{\mathrm{TV}}^{\mathcal{R}}(\nu\|\mu) = 0$, and thus $\widehat{\mathrm{D}}_{\mathrm{TV}}^{\mathcal{R}}(\nu\|\mu) = 0$. By (17), we deduce that $\mu = \nu$ over $\mathcal{R}$. But $\mathcal{R}$ is a $\pi$-system that generates $\mathcal{B}$. So Lemma 2.3 assures that we have $\mu = \nu$.

We end this section with another demonstration of the utility of Corollary 4.6: It allows us to extend essentially all known relations between divergences to their counterparts over rays.

**Theorem 4.8** (Relations between $f$-divergences over rays)**.** *The following inequalities hold verbatim when each $\mathrm{D}_f$ is replaced by $\mathrm{D}_f^{\mathcal{R}}$.*

(a) *Total variation and Hellinger:*

$$\frac{1}{2}\,\mathrm{D}_{\mathrm{H}}^2(\mu\|\nu) \le \mathrm{D}_{\mathrm{TV}}(\mu\|\nu) \le \mathrm{D}_{\mathrm{H}}(\mu\|\nu)\sqrt{1 - \frac{\mathrm{D}_{\mathrm{H}}^2(\mu\|\nu)}{4}} \le 1,$$

$$\mathrm{D}_{\mathrm{TV}}(\mu\|\nu) \le \sqrt{-2\log\Big(1 - \frac{\mathrm{D}_{\mathrm{H}}^2(\mu\|\nu)}{2}\Big)}.$$

(b) *Kullback–Leibler and total variation:*

$$\mathrm{D}_{\mathrm{TV}}^2(\mu\|\nu) \le \frac{1}{2}\,\mathrm{D}_{\mathrm{KL}}(\mu\|\nu),$$

$$\log\Big(\frac{1+\mathrm{D}_{\mathrm{TV}}(\mu\|\nu)}{1-\mathrm{D}_{\mathrm{TV}}(\mu\|\nu)}\Big) - \frac{2\,\mathrm{D}_{\mathrm{TV}}(\mu\|\nu)}{1+\mathrm{D}_{\mathrm{TV}}(\mu\|\nu)} \le \mathrm{D}_{\mathrm{KL}}(\mu\|\nu),$$

$$\mathrm{D}_{\mathrm{KL}}(\mu\|\nu) \le \log\Big(1 + \frac{1}{2\nu_{\min}}\,\mathrm{D}_{\mathrm{TV}}(\mu\|\nu)^2\Big) \le \frac{1}{2\nu_{\min}}\,\mathrm{D}_{\mathrm{TV}}(\mu\|\nu)^2,$$

*where* $\nu_{\min} \coloneqq \min_x \nu(x)$. *These are known respectively as Pinsker, strong Pinsker, and reverse Pinsker inequality.*

(c) *Total variation and* $\chi^2$*:*

$$\mathrm{D}_{\mathrm{TV}}^2(\mu\|\nu) \le \frac{1}{4}\,\mathrm{D}_{\chi^2}(\mu,\nu), \qquad \mathrm{D}_{\mathrm{TV}}(\mu\|\nu) \le \max\left\{\frac{1}{2}, \frac{\mathrm{D}_{\chi^2}(\mu\|\nu)}{1+\mathrm{D}_{\chi^2}(\mu,\nu)}\right\},$$
$$4\,\mathrm{D}_{\mathrm{TV}}^2(\mu\|\nu) \le \varphi(\mathrm{D}_{\mathrm{TV}}(\mu\|\nu)) \le \mathrm{D}_{\chi^2}(\mu\|\nu),$$

*where*

$$\varphi(t) = \begin{cases} 4t^2 & t \le 1/2, \\ t/(1-t) & t \ge 1/2. \end{cases}$$

(d) *Kullback–Leibler and Hellinger:*

$$\mathrm{D}_{\mathrm{H}}^2(\mu\|\nu) \le 2\log\frac{2}{2-\mathrm{D}_{\mathrm{H}}^2(\mu\|\nu)} \le \mathrm{D}_{\mathrm{KL}}(\mu\|\nu),$$
$$\mathrm{D}_{\mathrm{KL}}(\mu\|\nu) \le \frac{\log\big(1/\nu_{\min}-1\big)}{1-2\nu_{\min}}\big(1-(1-\mathrm{D}_{\mathrm{H}}^2(\mu\|\nu))^2\big).$$

(e) *Kullback–Leibler and* $\chi^2$*:*

$$\mathrm{D}_{\mathrm{KL}}(\mu\|\nu) \le \log\big(1+\mathrm{D}_{\chi^2}(\mu\|\nu)\big) \le \mathrm{D}_{\chi^2}(\mu\|\nu).$$

(f) *Le Cam and Hellinger:*

$$\frac{1}{2}\,\mathrm{D}_{\mathrm{H}}^2(\mu\|\nu) \le \mathrm{D}_{\mathrm{LC}}(\mu\|\nu) \le \mathrm{D}_{\mathrm{H}}^2(\mu\|\nu).$$

(g) *Le Cam and Jensen–Shannon:*

$$\mathrm{D}_{\mathrm{LC}}(\mu\|\nu) \le \mathrm{D}_{\mathrm{JS}}(\mu\|\nu) \le 2\log 2\cdot\mathrm{D}_{\mathrm{LC}}(\mu\|\nu).$$

*Proof.* These are well-known inequalities between the respective $f$-divergences: the two inequalities in (a) may be found in [40, p. 124] and [21, Theorem 12] respectively; the three inequalities in (b) in [52, p. 88], [53], and [42, Theorem 28] respectively; those in (c), (d), and (e) may be found in [40, p. 133]; (f) in [32, p. 48]; (g) in [50, Theorem 3.2]. It follows from Corollary 4.6 that these relations also hold for the corresponding $f$-divergences over $\mathcal{R}$. □

Ultimately, a definition is justified by the richness of results that can be proved with it. We view the results in this section as evidence that the $f$-divergence over rays as defined in Definition 3.12 is the "right" one. In this regard, the next section will supply another major piece of evidence.

## 5. Glivenko–Cantelli theorem for $f$-divergences over rays

Put in our context, the original Glivenko–Cantelli theorem in Theorem 1.4 states that empirical distributions converge to the target distribution almost surely in total variation distance over $\mathcal{R}$, as defined in Definition 3.12. Here we generalize this result to all $f$-divergences over $\mathcal{R}$.

As in the case of total variation distance, this convergence does not always hold if we replace $\mathcal{R}$ with $\mathcal{B}$ or some other subsets of $\mathcal{B}$. Examples 1.1 and 1.2, where $\limsup_{n\to\infty}\mathrm{D}_{\mathrm{TV}}(\nu_n\|\nu) > 0$, can be readily extended to any $f \in \mathcal{F}$ by way of Corollary 4.7:

$$0 < \bar{f}\big(\limsup_{n\to\infty}\mathrm{D}_{\mathrm{TV}}(\nu_n\|\nu)\big) = \limsup_{n\to\infty}\bar{f}\big(\mathrm{D}_{\mathrm{TV}}(\nu_n\|\nu)\big) \le \mathrm{D}_f(\nu_n\|\nu),$$

i.e., the $f$-divergence between the empirical distribution and the target does not in general converge to 0 almost surely. Our main result is that it holds for $f$-divergence over rays: For any $f \in \mathcal{F}$, Theorems 5.2 and 5.4 collectively show that almost surely

$$\lim_{n\to\infty}\mathrm{D}_f^{\mathcal{R}}(\nu_n\|\nu) = 0 = \lim_{n\to\infty}\mathrm{D}_f^{\mathcal{R}}(\nu\|\nu_n). \tag{18}$$

Note that an $f$-divergence is generally not symmetric in its arguments so we need to establish both equalities.

We will examine more carefully why a sequence of empirical measures can converge on $\mathcal{R}$ but fails to converge on $\mathcal{B}$. It turns out that a necessary condition is that its corresponding sequence of Radon–Nikodym derivatives does not converge to 1, but the sequence of their projections onto $G(\mathcal{R})$ does.

**Proposition 5.1.** *Let $\nu$ be a Borel probability measure and $\nu_n$ be the corresponding empirical measure, $n \in \mathbb{N}$. Suppose the Radon–Nikodym derivative $\mathrm{d}\nu_n/\mathrm{d}\nu$ exists for all $n \in \mathbb{N}$ and are uniformly bounded. If*

$$\sup_{A\in\mathcal{B}} \big(\nu_n(A) - \nu(A)\big) \not\to 0 \quad \textit{and} \quad \sup_{A\in\mathcal{R}} \big(\nu_n(A) - \nu(A)\big) \to 0,$$

*then*

$$\frac{\mathrm{d}\nu_n}{\mathrm{d}\nu} \not\to 1 \quad \textit{and} \quad \operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\nu_n}{\mathrm{d}\nu} \to 1,$$

*where convergence are all in an almost sure sense.*

*Proof.* Suppose $\lim_{n\to\infty} \mathrm{d}\nu_n/\mathrm{d}\nu = 1$. Then

$$0 = \int \frac{1}{2}\left|\lim_{n\to\infty} \frac{\mathrm{d}\nu_n}{\mathrm{d}\nu} - 1\right| \mathrm{d}\nu \geq \limsup_{n\to\infty} \int \frac{1}{2}\left|\frac{\mathrm{d}\nu_n}{\mathrm{d}\nu} - 1\right| \mathrm{d}\nu = \limsup_{n\to\infty} \mathrm{D}_{\mathrm{TV}}(\nu_n\|\nu) \geq 0.$$

So $\lim_{n\to\infty} \mathrm{D}_{\mathrm{TV}}(\nu_n\|\nu) = 0$, a contradiction, and thus $\mathrm{d}\nu_n/\mathrm{d}\nu \not\to 1$. By assumption, $\mathrm{d}\nu_n/\mathrm{d}\nu \leq c$ for all $n \in \mathbb{N}$ and for some constant $c > 0$. It then follows from Proposition 3.6 that

$$\operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\nu_n}{\mathrm{d}\nu} \leq \operatorname{proj}_{G(\mathcal{R})} c = c$$

for all $n$, i.e., the sequence $(\operatorname{proj}_{G(\mathcal{R})} \mathrm{d}\nu_n/\mathrm{d}\nu)_n$ is also uniformly bounded. Since $\dim_{\mathrm{VC}}(\mathcal{R}) < \infty$, it follows from the Vapnik–Chervonenkis theorem (see p. 4 for a statement) that

$$\lim_{n\to\infty} \mathrm{D}^{\mathcal{R}}_{\mathrm{TV}}(\nu_n\|\nu) = \lim_{n\to\infty} \sup_{A\in\mathcal{R}} \big(\nu_n(A) - \nu(A)\big) = 0$$

almost surely; taken together with the bounded convergence theorem,

$$\begin{aligned} 0 &= \lim_{n\to\infty} \sup_{A\in\mathcal{R}} \big(\nu_n(A) - \nu(A)\big) = \lim_{n\to\infty} \int \frac{1}{2}\left|\operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\nu_n}{\mathrm{d}\nu} - 1\right| \mathrm{d}\nu \\ &= \int \frac{1}{2}\left|\lim_{n\to\infty} \operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\nu_n}{\mathrm{d}\nu} - 1\right| \mathrm{d}\nu \geq 0. \end{aligned}$$

Hence the last integrand must converge to zero, or, equivalently,

$$\lim_{n\to\infty} \operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\nu_n}{\mathrm{d}\nu} = 1. \tag{19}$$

□

With the observation in Proposition 5.1, we obtain the left half of (18).

**Theorem 5.2** (Glivenko–Cantelli theorem for $f$-divergence I)**.** *Let $\mathcal{R}$ be the class of rays, $f \in \mathcal{F}$ as in (16), $\nu$ a Borel probability measure, and $\nu_n$ the corresponding empirical measure, $n \in \mathbb{N}$. Then almost surely*

$$\lim_{n\to\infty} \mathrm{D}^{\mathcal{R}}_f(\nu_n\|\nu) = 0.$$

*Proof.* By (19) and the reverse Fatou lemma,

$$\begin{aligned} 0 &= \int \left|f\Big(\lim_{n\to\infty} \operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\nu_n}{\mathrm{d}\nu}\Big)\right| \mathrm{d}\nu = \int \lim_{n\to\infty} \left|f\Big(\operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\nu_n}{\mathrm{d}\nu}\Big)\right| \mathrm{d}\nu \\ &\geq \limsup_{n\to\infty} \int \left|f\Big(\operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\nu_n}{\mathrm{d}\nu}\Big)\right| \mathrm{d}\nu \geq \limsup_{n\to\infty} \left|\int f\Big(\operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\nu_n}{\mathrm{d}\nu}\Big) \mathrm{d}\nu\right| \geq 0. \end{aligned}$$

Hence

$$\lim_{n\to\infty} \int f\Big(\operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\nu_n}{\mathrm{d}\nu}\Big)\,\mathrm{d}\nu = \lim_{n\to\infty} D_f^{\mathcal{R}}(\nu_n\|\nu) = 0. \qquad \square$$

The following "reciprocal" of (19) will lead us to the right half of (18).

**Proposition 5.3.** *Let $\nu$ be a Borel probability measure and $\nu_n$ be the corresponding empirical measure, $n \in \mathbb{N}$. Suppose the Radon–Nikodym derivative $\mathrm{d}\nu_n/\mathrm{d}\nu$ exists for all $n \in \mathbb{N}$ and are uniformly bounded. Then almost surely*

$$\lim_{n\to\infty} \operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\nu}{\mathrm{d}\nu_n} = 1.$$

*Proof.* Let $x_n \coloneqq |\operatorname{proj}_{G(\mathcal{R})}(\mathrm{d}\nu/\mathrm{d}\nu_n) - 1|/2$. Since $\dim_{\mathrm{VC}}(\mathcal{R}) < \infty$, it follows from the Vapnik–Chervonenkis theorem that

$$\sup_{A\in\mathcal{R}} \big(\nu(A) - \nu_n(A)\big) = \lim_{n\to\infty} \int x_n \,\mathrm{d}\nu_n = 0.$$

By assumption, $\mathrm{d}\nu/\mathrm{d}\nu_n < c$ for all $n \in \mathbb{N}$ and for some constant $c > 0$. If $\lim_{n\to\infty} x_n \neq 0$, then there exists $\varepsilon > 0$ and subsequences $(x_{n_k})$ and $(\nu_{n_k})$ such that

$$\varepsilon < \int x_{n_k}\,\mathrm{d}\nu = \int x_{n_k} \frac{\mathrm{d}\nu}{\mathrm{d}\nu_{n_k}}\,\mathrm{d}\nu_{n_k} \leq c \int x_{n_k}\,\mathrm{d}\nu_{n_k} \to 0,$$

a contradiction, and thus $\lim_{n\to\infty} x_n = 0$. $\square$

**Theorem 5.4** (Glivenko–Cantelli theorem for $f$-divergence II)**.** *Let $\mathcal{R}$ be the class of rays, $f \in \mathcal{F}$ as in (16), $\nu$ a Borel probability measure, and $\nu_n$ the corresponding empirical measure, $n \in \mathbb{N}$. Then almost surely*

$$\lim_{n\to\infty} \mathrm{D}_f^{\mathcal{R}}(\nu\|\nu_n) = 0.$$

*Proof.* Let $y_n \coloneqq |f(\operatorname{proj}_{G(\mathcal{R})} \mathrm{d}\nu/\mathrm{d}\nu_n)|$. Since $f$ is continuous at 1,

$$\lim_{n\to\infty} y_n = \Big|f\Big(\lim_{n\to\infty} \operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\nu}{\mathrm{d}\nu_n}\Big)\Big| = 0.$$

So

$$\Big|\int f\Big(\operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\nu}{\mathrm{d}\nu_n}\Big)\,\mathrm{d}\nu_n\Big| \leq \int \Big|f\Big(\operatorname{proj}_{G(\mathcal{R})} \frac{\mathrm{d}\nu}{\mathrm{d}\nu_n}\Big)\Big|\,\mathrm{d}\nu_n = \nu_n y_n.$$

If $\nu_n y_n \not\to 0$, then there exists $\varepsilon > 0$ and a subsequence $(y_{n_k})$ such that

$$\varepsilon < |\nu_{n_k} y_{n_k}| \leq \|y_{n_k}\| \to 0,$$

a contradiction, and thus $\lim_{n\to\infty} \nu_n y_n = 0$, giving us the required result. $\square$

The following is an immediate consequence of Theorems 5.2 and 5.4.

**Corollary 5.5** (Glivenko–Cantelli theorem for $f$-divergence III)**.** *Let $\mathcal{R}$ be the class of rays, $f \in \mathcal{F}$ as in (16), $\nu$ a Borel probability measure, and $\nu_n$ the corresponding empirical measure, $n \in \mathbb{N}$. Then almost surely*

$$\lim_{n\to\infty} \max\big\{\mathrm{D}_f^{\mathcal{R}}(\nu\|\nu_n), \mathrm{D}_f^{\mathcal{R}}(\nu_n\|\nu)\big\} = 0.$$

One advantage afforded by the version in Corollary 5.5 is that the expression

$$\widehat{\mathrm{D}}_f^{\mathcal{R}}(\mu\|\nu) \coloneqq \max\big\{\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu), \mathrm{D}_f^{\mathcal{R}}(\nu\|\mu)\big\}$$

defines a distance (and not just a divergence) between probability measures. For $f(t) = |t-1|/2$, this is exactly the Kolmogorov–Smirnov distance in (17). We may of course also replace the expression with an $L^p$ variant for any $p \in [1, \infty)$.

## 6. Vapnik–Chervonenkis theory for $f$-divergence

A natural follow-up question from the last section is: Given that we have a Glivenko–Cantelli theorem for $f$-divergence, is a Vapnik–Chervonenkis theory for $f$-divergence within reach? A first goal of such a theory would be to extend the work in this article from the class of rays $\mathcal{R}$ to more general classes $\mathcal{C} \subseteq \mathcal{B}$. Surprisingly, we found that the original Vapnik–Chervonenkis theorem [55] provides a key.

Replace $\mathcal{R}$ by a class $\mathcal{C} \subseteq \mathcal{B}$ in the definition of $G(\mathcal{R})$ in (6). As Section 3 shows, if $G(\mathcal{C})$ satisfies the conclusions of Proposition 3.1 and Theorem 3.15 with $\mathcal{C}$ in place of $\mathcal{R}$, then Definition 3.12 with $\mathcal{C}$ in place of $\mathcal{R}$ gives a well-defined $f$-divergence over $\mathcal{C}$. We formalize this with a definition.

**Definition 6.1** (Pre-Glivenko–Cantelli class)**.** We say that $\mathcal{C} \subseteq \mathcal{B}$ is a pre-Glivenko–Cantelli class if for any Borel probability measure $\nu$,

$$G(\mathcal{C}) \coloneqq \big\{ g \in L^2(\nu) : \{g > r\} \in \mathcal{C} \text{ for all } r \in \mathbb{R} \big\}$$

is a closed convex cone in $L^2(\nu)$; and for any pair of Borel probability measures $\mu$ and $\nu$,

$$\sup_{A \in \mathcal{C}} \big(\mu(A) - \nu(A)\big) = \int \frac{1}{2} \Big| \operatorname{proj}_{G(\mathcal{C})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1 \Big| \, \mathrm{d}\nu.$$

We may speak freely of $\mathrm{D}_f^{\mathcal{C}}$ whenever we have a pre-Glivenko–Cantelli class $\mathcal{C}$ and an $f \in \mathcal{F}$. We show that the first requirement in Definition 6.1 follows as long as $\mathcal{C}$ is closed under union and intersection, noting that both $\mathcal{R}$ and any $\sigma$-subalgebra of $\mathcal{B}$ will have this property.

**Lemma 6.2.** *If $\mathcal{C} \subseteq \mathcal{B}$ is closed under countable unions and intersections, then $G(\mathcal{C})$ is a closed convex cone in $L^2(\nu)$.*

*Proof.* The sum of two functions in $G(\mathcal{C})$ remains in it as

$$\{g_1 + g_2 > c\} = \bigcup_{q \in \mathbb{Q}} \{g_1 > c - q\} \cap \{g_2 > q\} \in \mathcal{C}.$$

From this it is routine to verify that $G(\mathcal{C})$ is a convex cone. If $g_n \in G(\mathcal{C})$ is a sequence with $\|g_n - g\| \to 0$, then $g_n \to g$ in $\nu$-measure. By passing through a subsequence if necessary, we may assume that $g_n \to g$ almost surely. It then follows from the definition of pointwise convergence that

$$\{g > c\} = \bigcup_{k \geq 1} \bigcap_{m \geq 1} \bigcup_{n \geq m} \Big\{ g_n > c + \frac{1}{k} \Big\} \in \mathcal{C}$$

and so $g \in G(\mathcal{C})$. Hence $G(\mathcal{C})$ is closed. □

The second requirement in Definition 6.1 is usually harder to establish. For $\mathcal{B}$ it is immediate but for $\mathcal{R}$ it took nearly the whole of Section 3. We will extend the former to include all $\sigma$-subalgebras of $\mathcal{B}$.

**Lemma 6.3.** *Any $\sigma$-subalgebra $\mathcal{C} \subseteq \mathcal{B}$ is a pre-Glivenko–Cantelli class.*

*Proof.* If $\mathcal{C}$ is a $\sigma$-subalgebra, then the restrictions of Borel probability measures $\mu$ and $\nu$ to $\mathcal{C}$, $\mu|_{\mathcal{C}}$ and $\nu|_{\mathcal{C}}$, remain probability measures [43, p. 30]. Let $g$ be an almost surely nonnegative $\mathcal{C}$-measurable function, i.e., $\{g > c\} \in \mathcal{C}$ for all $c \in \mathbb{R}$. Then

$$\int_A g \, \mathrm{d}\nu|_{\mathcal{C}} = \int_A g \, \mathrm{d}\nu$$

for any $A \in \mathcal{C}$, i.e., the value of the integral is preserved. If $\mu \ll \nu$, then $\mu|_{\mathcal{C}} \ll \nu|_{\mathcal{C}}$, ensuring that the Radon–Nikodym derivative $\mathrm{d}\mu|_{\mathcal{C}}/\mathrm{d}\nu|_{\mathcal{C}}$ exists and that

$$\int_A \frac{\mathrm{d}\mu|_{\mathcal{C}}}{\mathrm{d}\nu|_{\mathcal{C}}} \, \mathrm{d}\nu = \int_A \frac{\mathrm{d}\mu|_{\mathcal{C}}}{\mathrm{d}\nu|_{\mathcal{C}}} \, \mathrm{d}\nu|_{\mathcal{C}} = \mu|_{\mathcal{C}}(A) = \mu(A) = \int_A \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \, \mathrm{d}\nu$$

for any $A \in \mathcal{C}$. So $\mathrm{d}\mu|_{\mathcal{C}}/\mathrm{d}\nu|_{\mathcal{C}}$ is indeed the conditional expectation and

$$\frac{\mathrm{d}\mu|_{\mathcal{C}}}{\mathrm{d}\nu|_{\mathcal{C}}} = \mathbb{E}_{\mu}\Big[\frac{\mathrm{d}\mu}{\mathrm{d}\nu} \;\Big|\; \mathcal{C}\Big] = \operatorname{proj}_{G(\mathcal{C})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \tag{20}$$

by [4, p. 90]. Hence

$$\begin{aligned} \mathrm{D}_{\mathrm{TV}}^{\mathcal{C}}(\mu\|\nu) &= \int \frac{1}{2}\Big|\operatorname{proj}_{G(\mathcal{C})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\Big| \,\mathrm{d}\nu = \int \frac{1}{2}\Big|\frac{\mathrm{d}\mu|_{\mathcal{C}}}{\mathrm{d}\nu|_{\mathcal{C}}} - 1\Big| \,\mathrm{d}\nu \\ &= \int \frac{1}{2}\Big|\frac{\mathrm{d}\mu|_{\mathcal{C}}}{\mathrm{d}\nu|_{\mathcal{C}}} - 1\Big| \,\mathrm{d}\nu|_{\mathcal{C}} = \sup_{A\in\mathcal{C}} \big(\mu(A) - \nu(A)\big). \end{aligned}$$

□

So Definition 6.1 is not vacuous: The class of rays $\mathcal{R}$ and any $\sigma$-subalgebra of $\mathcal{B}$ give two examples of pre-Glivenko–Cantelli classes. The reason for introducing this notion is that any finite VC-dimensional pre-Glivenko–Cantelli class gives us a Vapnik–Chervonenkis theorem for $f$-divergence.

**Theorem 6.4** (Vapnik–Chervonenkis theorem for $f$-divergence)**.** *Let $f \in \mathcal{F}$ and $\mathcal{C}$ be a pre-Glivenko–Cantelli class. If $\dim_{\mathrm{VC}}(\mathcal{C}) < \infty$, then almost surely*

$$\lim_{n\to\infty} \mathrm{D}_f^{\mathcal{C}}(\nu_n\|\nu) = 0 = \lim_{n\to\infty} \mathrm{D}_f^{\mathcal{C}}(\nu\|\nu_n). \tag{21}$$

*Proof.* All that we really need to observe is that the whole of Section 5 does not depend on the fact that $\mathcal{R}$ is the class of rays — every result therein holds verbatim with $\mathcal{C}$ in place of $\mathcal{R}$ as long as $\mathrm{D}_f^{\mathcal{C}}$ is well-defined, i.e., as long as $\mathcal{C}$ is a pre-Glivenko–Cantelli class. The original Vapnik–Chervonenkis theorem [55], as applied in Propositions 5.1 and 5.3, shows that a pre-Glivenko–Cantelli class $\mathcal{C}$ with finite VC-dimension must satisfy (21) almost surely. □

With Theorem 6.4 we are led to the following definition in analogy with standard Glivenko–Cantelli class (i.e., with respect to total variation distance).

**Definition 6.5** (Glivenko–Cantelli class for $f$-divergence)**.** Let $f \in \mathcal{F}$ and $\mathcal{C} \subseteq \mathcal{B}$ be a pre-Glivenko–Cantelli class for $f$. We say that $\mathcal{C}$ is a Glivenko–Cantelli class for $f$ if (21) holds almost surely.

Note that we need to restrict ourselves to pre-Glivenko–Cantelli classes so that $\mathrm{D}_f^{\mathcal{C}}$ is well-defined. It is nevertheless possible for such a Glivenko–Cantelli class to have infinite VC-dimension, so Theorem 6.4 merely provides a sufficient criterion.

**Example 6.6** (Infinite VC-dimensional Glivenko–Cantelli class for $f$-divergence)**.** Let $\mathcal{E} = (E_n)_{n\in\mathbb{N}}$ be a countable partition of $\mathbb{R}$ with measurable $E_n$, $n \in \mathbb{N}$. Let $\mathcal{C} = \sigma(\mathcal{E}) \subseteq \mathcal{B}$ be the $\sigma$-algebra generated by $\mathcal{E}$. So any set in $\mathcal{C}$ can be expressed as a countable union of sets in $\mathcal{E}$ [9, p. 4]. Thus any $\mathcal{C}$-measurable function must be a simple function $h = \sum_{n=1}^{\infty} \alpha_n \mathbb{1}_{E_n}$. Any pair of Borel probability measures $\mu$ and $\nu$ restrict to probability measures $\mu|_{\mathcal{C}}$ and $\nu|_{\mathcal{C}}$ with Radon–Nikodym derivative

$$\frac{\mathrm{d}\mu|_{\mathcal{C}}}{\mathrm{d}\nu|_{\mathcal{C}}} = \sum_{n=1}^{\infty} \alpha_n \mathbb{1}_{E_n}.$$

For any $m \in \mathbb{N}$,

$$\mu|_{\mathcal{C}}(E_m) = \int_{E_m} \frac{\mathrm{d}\mu|_{\mathcal{C}}}{\mathrm{d}\nu|_{\mathcal{C}}} \,\mathrm{d}\nu|_{\mathcal{C}} = \int \mathbb{1}_{E_m} \sum_{n=1}^{\infty} \alpha_n \mathbb{1}_{E_n} \,\mathrm{d}\nu|_{\mathcal{C}} = \alpha_m \cdot \nu|_{\mathcal{C}}(E_m).$$

By (20),

$$\operatorname{proj}_{G(\mathcal{C})} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} = \frac{\mathrm{d}\mu|_{\mathcal{C}}}{\mathrm{d}\nu|_{\mathcal{C}}} = \sum_{n=1}^{\infty} \mathbb{1}_{E_n} \frac{\mu|_{\mathcal{C}}(E_n)}{\nu|_{\mathcal{C}}(E_n)},$$

and so the $f$-divergence over $\mathcal{C}$ is given by

$$\mathrm{D}_f^{\mathcal{C}}(\nu_n\|\nu) = \mathrm{D}_f(\nu_n|_{\mathcal{C}}\|\nu|_{\mathcal{C}}) = \sum_{i=1}^{\infty} f\Big(\frac{\nu_n|_{\mathcal{C}}(E_i)}{\nu|_{\mathcal{C}}(E_i)}\Big)\nu|_{\mathcal{C}}(E_i).$$

If the sequence $\big(\nu_n|_{\mathcal{C}}(E_k)/\nu|_{\mathcal{C}}(E_k)\big)_{k\in\mathbb{N}}$ is uniformly bounded, then

$$\begin{aligned}\limsup_{n\to\infty} \mathrm{D}_f^{\mathcal{C}}(\nu_n\|\nu) &= \limsup_{n\to\infty} \sum_{i=1}^{\infty} f\Big(\frac{\nu_n|_{\mathcal{C}}(E_i)}{\nu|_{\mathcal{C}}(E_i)}\Big)\nu|_{\mathcal{C}}(E_i)\\ &\le \sum_{i=1}^{\infty} \limsup_{n\to\infty} f\Big(\frac{\nu_n|_{\mathcal{C}}(E_i)}{\nu|_{\mathcal{C}}(E_i)}\Big)\nu|_{\mathcal{C}}(E_i) = \sum_{i=1}^{\infty} f\Big(\lim_{n\to\infty} \frac{\nu_n|_{\mathcal{C}}(E_i)}{\nu|_{\mathcal{C}}(E_i)}\Big)\nu|_{\mathcal{C}}(E_i) = 0.\end{aligned}$$

So $\lim_{n\to\infty} \mathrm{D}_f^{\mathcal{C}}(\nu_n\|\nu) = 0$ almost surely. Likewise $\lim_{n\to\infty} \mathrm{D}_f^{\mathcal{C}}(\nu\|\nu_n) = 0$ almost surely. Hence $\mathcal{C}$ satisfies (21) almost surely even though $\dim_{\mathrm{VC}}(\mathcal{C}) = \infty$.

## 7. Choquet integral

To address Question 1, our initial thought was to employ the Choquet integral [11, 15], promising because it allows one to work with an arbitrary class $\mathcal{C} \subseteq 2^{\Omega}$ with essentially no condition imposed (unlike $\pi$-system or $\sigma$-algebra). This is perfect for us as it seems that we could have taken $\mathcal{C} = \mathcal{R}$ but it did not work. We think it might be instructive to record the difficulties encountered as we remain optimistic that it is not a dead end.

The Lebesgue integral defines integration with respect to a measure; Choquet integral extends this to so-called "capacities." Let $\Omega$ be a set and $\mathcal{C} \subseteq 2^{\Omega}$ be a collection of subsets with $\varnothing, \Omega \in \mathcal{C}$. A function $\nu : \mathcal{C} \to \mathbb{R}$ is called a *capacity* on $\mathcal{C}$ if it satisfies

(a) normalization: $\nu(\varnothing) = 0$;
(b) monotonicity: $\nu(A) \le \nu(B)$ whenever $A \subseteq B$, $A, B \in \mathcal{C}$.

A function $g : \Omega \to \mathbb{R}$ is $\mathcal{C}$-measurable if $\{g > t\} \in \mathcal{C}$ for all $t \in \mathbb{R}$. Let $B_+(\mathcal{C})$ be the set of all bounded, nonnegative, $\mathcal{C}$-measurable functions. The distribution of $g \in B_+(\mathcal{C})$ is the function $m_{\nu,g} : \mathbb{R} \to [0,\infty)$, $t \mapsto \nu(\{g > t\})$. The Choquet integral of the function $g$ with respect to the capacity $\nu$ is then

$$\text{⨍} g \,\mathrm{d}\nu := \int_0^{\infty} m_{\nu,g}(t)\,\mathrm{d}t,$$

where the right-hand side is a Riemann integral. Choquet coincides with Lebesgue when $\nu$ is a measure [15, p. 62].

For the class of rays $\mathcal{R}$, it is easy to see that $\nu_{\mathcal{R}}$, the restriction of $\nu$ to $\mathcal{R}$, is a capacity. So we could simply define $f$-divergence over $\mathcal{R}$ as

$$\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu) := \mathrm{D}_f(\mu_{\mathcal{R}}\|\nu_{\mathcal{R}}) = \text{⨍} f\Big(\frac{\mathrm{d}\mu_{\mathcal{R}}}{\mathrm{d}\nu_{\mathcal{R}}}\Big)\,\mathrm{d}\nu_{\mathcal{R}}.$$

Two immediate problems are that the Radon–Nikodym derivative $\mathrm{d}\mu_{\mathcal{R}}/\mathrm{d}\nu_{\mathcal{R}}$ cannot be easily defined and the function $f(\mathrm{d}\mu_{\mathcal{R}}/\mathrm{d}\nu_{\mathcal{R}})$ may not be $\mathcal{R}$-measurable, i.e., nonincreasing. The first problem has been discussed extensively [24, 25, 38, 48, 51] but all assumed that $\mathcal{C}$ is a $\sigma$-algebra and thus inapplicable to $\mathcal{R}$. One might think that a way around these problems is to extend the domain of $\mu_{\mathcal{R}}$ and $\nu_{\mathcal{R}}$ back to the whole of $\mathcal{B}$, i.e.,

$$\widehat{\nu}_{\mathcal{R}}(A) := \begin{cases} \nu(A) & \text{if } A \in \mathcal{R},\\ 0 & \text{if } A \in \mathcal{B} \setminus \mathcal{R},\end{cases}$$

and likewise for $\widehat{\mu}_{\mathcal{R}}$. But now $\widehat{\mu}_{\mathcal{R}}$ and $\widehat{\nu}_{\mathcal{R}}$ fail monotonicity and are not even capacities.

We are slightly disappointed that the discussions in [24, 25, 38, 48, 51] are all based on the assumption that $\mathcal{C}$ is a $\sigma$-algebra (if so, then why not just use standard Lebesgue integral?) but interested readers may have better luck where we and these authors failed:

**Open Problem 1.** Is it possible to define an $f$-divergence using Choquet integral for a reasonably general class $\mathcal{C} \subseteq \mathcal{B}$ so that Theorems 4.1 and 4.8 hold true?

By "reasonably general" we expect $\mathcal{C}$ to cover at least the requirements of Lemma 6.2, or perhaps be a $\pi$-system or a $\lambda$-system, certainly not just a $\sigma$-algebra. A positive answer to this open problem should provide an alternative to Definition 3.12 and could potentially be useful for developing the Vapnik–Chervonenkis theory for $f$-divergence in Section 6.

## 8. Conclusion

As we have noted in Section 1, the total variation distance plays an outsize role in traditional theoretical statistics but modern AI applications often relies on $f$-divergences other than the total variation distance. The work in this article would hopefully shed light on how one may extend classical results in statistics to other $f$-divergences.

## References


[1] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):131–142, 1966.

[2] A. R. Alimov and I. G. Tsar'kov. *Geometric approximation theory*. Springer Monographs in Mathematics. Springer, Cham, 2021.

[3] C. M. Bishop. *Pattern recognition and machine learning*. Information Science and Statistics. Springer, New York, 2006.

[4] A. Bobrowski. *Functional analysis for probability and stochastic processes*. Cambridge University Press, Cambridge, 2005.

[5] F. Bolley, A. Guillin, and C. Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probab. Theory Related Fields*, 137(3–4):541–593, 2007.

[6] O. Bousquet and A. Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2(3):499–526, 2002.

[7] J. Bröcker. A lower bound on arbitrary $f$-divergences in terms of the total variation, 2009. Preprint, arXiv:0903.1765.

[8] F. P. Cantelli. Sulla determinazione empirica delle leggi di probabilita. *Giorn. Ist. Ital. Attuari*, 4:421–424, 1933.

[9] E. Çinlar. *Probability and stochastics*, volume 261 of *Graduate Texts in Mathematics*. Springer, New York, 2011.

[10] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statistics*, 23:493–507, 1952.

[11] G. Choquet. Theory of capacities. *Ann. Inst. Fourier (Grenoble)*, 5:131–295, 1953/54.

[12] D. L. Cohn. *Measure theory*. Birkhäuser Advanced Texts: Basel Textbooks. Birkhäuser/Springer, New York, second edition, 2013.

[13] I. Csiszar. Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. *Magyer Tud. Akad. Mat. Kutato Int. Koezl.*, 8:85–108, 1964.

[14] H. Dehling. Almost sure convergence of random variables. In M. Lovric, editor, *International Encyclopedia of Statistical Science*, pages 32–35. Springer, Berlin Heidelberg, 2011.

[15] D. Denneberg. *Non-additive measure and integral*, volume 27 of *Theory and Decision Library. Series B: Mathematical and Statistical Methods*. Kluwer Academic Publishers Group, Dordrecht, 1994.

[16] R. M. Dudley. *Uniform central limit theorems*, volume 142 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, New York, second edition, 2014.

[17] E. B. Dynkin. *Markov processes. Vol. II*, volume Band 122 of *Die Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin-Göttingen-Heidelberg; Academic Press, Inc., Publishers, New York, 1965. Translated with the authorization and assistance of the author by J. Fabius, V. Greenberg, A. Maitra, G. Majone.

[18] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York, 1993.

[19] S. Eguchi. A differential geometric approach to statistical inference on the basis of contrast functionals. *Hiroshima Math. J.*, 15(2):341–391, 1985.

[20] G. Folland. *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 1999.

[21] G. L. Gilardoni. On Pinsker's and Vajda's type inequalities for Csiszár's $f$-divergences. *IEEE Trans. Inform. Theory*, 56(11):5377–5386, 2010.
[22] V. Glivenko. Sulla determinazione empirica delle leggi di probabilita. *Gion. Ist. Ital. Attauri.*, 4:92–99, 1933.
[23] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.
[24] S. Graf. A Radon–Nikodým theorem for capacities. *J. Reine Angew. Math.*, 320:192–214, 1980.
[25] J. Harding, M. Marinacci, N. T. Nguyen, and T. Wang. Local Radon–Nikodým derivatives of set functions. *Internat. J. Uncertain. Fuzziness Knowledge-Based Systems*, 5(3):379–394, 1997.
[26] E. Hellinger. Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *J. Reine Angew. Math.*, 136:210–271, 1909.
[27] N. Henze. *Asymptotic Stochastics: An Introduction with a View Towards Statistics*. Mathematics Study Resources. Springer, Berlin Heidelberg, 2024.
[28] J. K. Hunter and B. Nachtergaele. *Applied analysis*. World Scientific Publishing Co., Inc., River Edge, NJ, 2001.
[29] H. Jeffreys. *Theory of probability*. Oxford Classic Texts in the Physical Sciences. The Clarendon Press, Oxford University Press, New York, 1998.
[30] M. Kelbert. Survey of distances between the most popular distributions. *Analytics*, 2(1):225–245, 2023.
[31] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations (ICLR)*, 2014.
[32] L. Le Cam. *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer, New York, 1986.
[33] Y. Li and R. E. Turner. Rényi divergence variational inference. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 1081–1089. Curran Associates, Inc., 2016.
[34] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Trans. Inform. Theory*, 37(1):145–151, 1991.
[35] A. Liu, J.-G. Liu, and Y. Lu. On the rate of convergence of empirical measure in $\infty$-Wasserstein distance for unbounded density function. *Quart. Appl. Math.*, 77(4):811–829, 2019.
[36] S. A. Molchanov. *Geometric modeling in probability and statistics* [book review of MR3308142]. *Bull. Amer. Math. Soc. (N.S.)*, 55(1):109–111, 2018.
[37] T. Morimoto. Markov processes and the h-theorem. *Journal of the Physical Society of Japan*, 18(3):328–331, 1963.
[38] Y. Narukawa. Distances defined by Choquet integral. In *2007 IEEE International Fuzzy Systems Conference*, 2007.
[39] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. In S. Kotz and N. L. Johnson, editors, *Breakthroughs in Statistics: Methodology and Distribution*, pages 11–28. Springer, New York, 1992.
[40] Y. Polyanskiy and Y. Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2024.
[41] A. Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pages 547–562. University of California Press, 1961.
[42] I. Sason and S. Verdú. $f$-divergence inequalities. *IEEE Trans. Inform. Theory*, 62(11):5973–6006, 2016.
[43] R. L. Schilling. *Measures, integrals and martingales*. Cambridge University Press, Cambridge, second edition, 2017.
[44] C. Schumacher, F. Schwarzenberger, and I. Veselić. A Glivenko–Cantelli theorem for almost additive functions on lattices. *Stochastic Process. Appl.*, 127(1):179–208, 2017.
[45] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.
[46] G. R. Shorack and J. A. Wellner. *Empirical processes with applications to statistics*, volume 59 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2009.
[47] R. M. Shortt. Empirical measures. *Amer. Math. Monthly*, 91(6):358–360, 1984.
[48] M. Sugeno. A note on derivatives of functions with respect to fuzzy measures. *Fuzzy Sets and Systems*, 222:1–17, 2013.
[49] J. C. Taylor. *An introduction to measure and probability*. Springer, New York, 1997.
[50] F. Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE Trans. Inform. Theory*, 46(4):1602–1609, 2000.
[51] V. Torra, Y. Narukawa, and M. Sugeno. On the $f$-divergence for non-additive measures. *Fuzzy Sets and Systems*, 292:364–379, 2016.

[52] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.
[53] I. Vajda. Note on discrimination information and variation. *IEEE Trans. Inform. Theory*, IT-16:771–773, 1970.
[54] V. N. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, Inc., New York, 1998.
[55] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theor. Probability Appl.*, 16(2):264–280, 1971.

Computational and Applied Mathematics Initiative, Department of Statistics, University of Chicago, Chicago, IL 60637
*Email address*: haomingwang@uchicago.edu

Computational and Applied Mathematics Initiative, Department of Statistics, University of Chicago, Chicago, IL 60637
*Email address*: lekheng@uchicago.edu