----

1. This sketch summarizes the key results of this work.
2. The full paper is attached at the end.
3. This work was conducted under the guidance of Prof. Lek-Heng Lim.

Contents

1 Motivation.

- Let μ, ν be probability measures on measurable space (Ω, Σ).
- The total variation (TV) distance and the Kolmogorov–Smirnov (KS) distance can both be written in the form of
$$\sup_{A \in \mathcal{A}} |\mu(A) - \nu(A)|.$$
where $\mathcal{A} \subset \Sigma$ is a collection of sets.
- We want to explore the deeper intrinsic connections between them.
- Total variation distance is f-divergence, and hence has an integral representation
$$\mathsf{TV}(\mu, \nu) = \sup_{A \in \Sigma} |\mu(A) - \nu(A)| = \int \frac{1}{2} \cdot \left| \frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1 \right| \mathrm{d}\nu. \tag{1}$$
- A natural question is: *can we have a similar expression like equation (1) for KS distance?*

2 Preliminary.

Metric Projection.

- Let (X, d) be a metric space, $M \subseteq X$ is nonempty, then the *metric projection* onto M, denoted by $\mathrm{proj}_M : X \to 2^M$ is a set-value function:
$$\mathrm{proj}_M(x) := \left\{ y \in M : d(x, y) = \arg\min_{z \in M} d(x, z) \right\}.$$
- In other words, $\mathrm{proj}_M(x)$ is the points in M that have the same nearest distance to x.
- When $\mathrm{proj}_M(x)$ is single-valued for any $x \in X$, the set M is called a *Chebyshev set*.
- If X is a Hilbert space, M is a convex and closed subset of X, M is a Chebyshev set, and proj_M is continuous.

Class of Rays.

- Define the class of open rays as $\mathcal{R}_{)} = \{(-\infty, a) : a \in \mathbb{R}\} \cup \{\varnothing, \mathbb{R}\}$
- Define the class of closed rays as $\mathcal{R}_{]} = \{(-\infty, a] : a \in \mathbb{R}\} \cup \{\varnothing, \mathbb{R}\}$.
- Define the class of rays as $\mathcal{R} = \mathcal{R}_{]} \cup \mathcal{R}_{)}$.
- KS distance is defined as: $\mathsf{KS}(\mu, \nu) = \sup_{A \in \mathcal{R}_{]}} (\mu(A) - \nu(A))$.

- We can prove that

$$\sup_{A\in\mathcal{R}_]}(\mu(A)-\nu(A)) = \sup_{A\in\mathcal{R}_)}(\mu(A)-\nu(A)) = \sup_{A\in\mathcal{R}}(\mu(A)-\nu(A))$$

  thus, we will analyze only the third one for simplicity.
- Please first ignore the difference between absolute value and parentheses in the definition.
- For total variation distance, there is no difference:

$$\sup_{A\in\Sigma}(\mu(A)-\nu(A)) = \sup_{A\in\Sigma}|\mu(A)-\nu(A)|$$

  because Σ is a σ-algebra and for any $A\in\Sigma, A^c\in\Sigma$ as well.
- For KS distance,

$$\sup_{A\in\mathcal{R}}|\mu(A)-\nu(A)| = \max\left\{\sup_{A\in\mathcal{R}}(\mu(A)-\nu(A)), \sup_{A\in\mathcal{R}}(\nu(A)-\mu(A))\right\}$$

$\mathcal{R}$-measurable.

- Define a function f is $\mathcal{R}$-measurable if for any $r\in\mathbb{R}$, the preimage $(f>r)=f^{-1}(r,\infty)\in\mathcal{R}$.
- Let $\mathcal{M}_{\mathcal{R}}$ be the collection of *$\mathcal{R}$-measurable function* in $L_2(\nu)$:

$$\mathcal{M}_{\mathcal{R}} = \{f\in L_2(\nu) : (f>r)\in\mathcal{R}, r\in\mathbb{R}\}.$$

- We proved that $\mathcal{M}_{\mathcal{R}}$ is a closed convex cone in $L_2(\nu)$.
- Thus $\mathcal{M}_{\mathcal{R}}$ is a Chebyshev set, and the metric projection operator $\mathrm{proj}_{\mathcal{M}_{\mathcal{R}}}$ is single-valued and continuous.

3 Main Results

Key lemma.

- Although this result served as a lemma in the paper, we consider it insightful and significant and thus state it as a theorem.
- This result states that for any $f\in L_2^+$, its metric projection onto $\mathcal{M}_{\mathcal{R}}$ has the same integral as f over any $E\in\mathcal{E}\subset\mathcal{R}$, as shown in Figure 1.

Theorem 1. *Let $f\in L_2^+$, let $\mathcal{E}=\{(\mathrm{proj}_{\mathcal{M}_{\mathcal{R}}}f>r) : r\geq 0\}\cup\{\mathbb{R}\}$, then for any $E\in\mathcal{E}$*

$$\int_E \mathrm{proj}_{\mathcal{M}_{\mathcal{R}}}f\,\mathrm{d}\nu = \int_E f\,\mathrm{d}\nu. \tag{2}$$

and for any $A\in\mathcal{R}$,

$$\int_A \mathrm{proj}_{\mathcal{M}_{\mathcal{R}}}f\,\mathrm{d}\nu \geq \int_A f\,\mathrm{d}\nu. \tag{3}$$

- We think this extends the Radon-Nikodym (RN) theorem.
- Let $\mu\ll\nu$, the RN theorem implies for any sub σ-algebra $\mathcal{G}\subset\Sigma$, there exists a $\mathcal{G}$-measurable function $\rho_{\mathcal{G}}$ such that $\mu(A)=\int_A\rho_{\mathcal{G}}\,\mathrm{d}\nu$ for any $A\in\mathcal{G}$.
- For σ-algebras $\mathcal{G}\subset\Sigma$, the blue part is a stronger condition; the red part is a weaker condition.

- A question is, *"what if $\mathcal{G}$ is not a σ-algebra?"*
- This theorem implies, for $\mathcal{R} \subset \Sigma$, there exists a $\mathcal{R}$-measurable function $\rho_{\mathcal{R}}$ and a collection $\mathcal{E} \subset \mathcal{R}$, such that

$$\mu(A) = \int_A \rho_{\mathcal{R}} \, \mathrm{d}\nu, \quad \text{for any } A \in \mathcal{E}.$$

- where $\rho_{\mathcal{R}} = \mathrm{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}$.

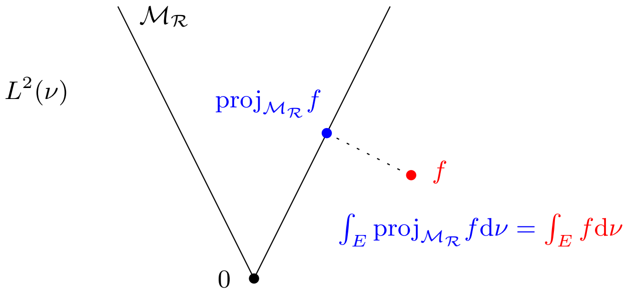

$\mathcal{M}_{\mathcal{R}}$
$L^2(\nu)$
$\mathrm{proj}_{\mathcal{M}_{\mathcal{R}}} f$
f
$\int_E \mathrm{proj}_{\mathcal{M}_{\mathcal{R}}} f \mathrm{d}\nu = \int_E f \mathrm{d}\nu$
0

Figure 1: metric projection.

Main theorem.

- Back to our original question: Is there an integral representation of KS distance similar to that of TV distance:

$$\mathsf{TV}(\mu, \nu) = \sup_{A \in \Sigma} |\mu(A) - \nu(A)| = \int \frac{1}{2} \cdot \left| \frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1 \right| \mathrm{d}\nu.$$

- The answer is Yes!

Theorem 2. *Let μ, ν be probability measures, then*

$$\mathsf{KS}(\mu, \nu) = \sup_{A \in \mathcal{R}} (\mu(A) - \nu(A)) = \int \frac{1}{2} \cdot \left| \mathrm{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1 \right| \mathrm{d}\nu.$$

- Observe that when replacing $\mathcal{R}$ with the σ-algebra Σ back, it recovers to TV distance.
- This led us to introduce a concept of $(\mathcal{R}, f)$-divergence and extend the properties of KS distance (e.g., GC theorem) and f-divergence (e.g., joint range theorem) to this new conception.

$(\mathcal{R}, f)$**-divergence and its properties.**

Let $\mathcal{F} = \{f$ be a continuous convex function such that $f(1) = 0\}$.

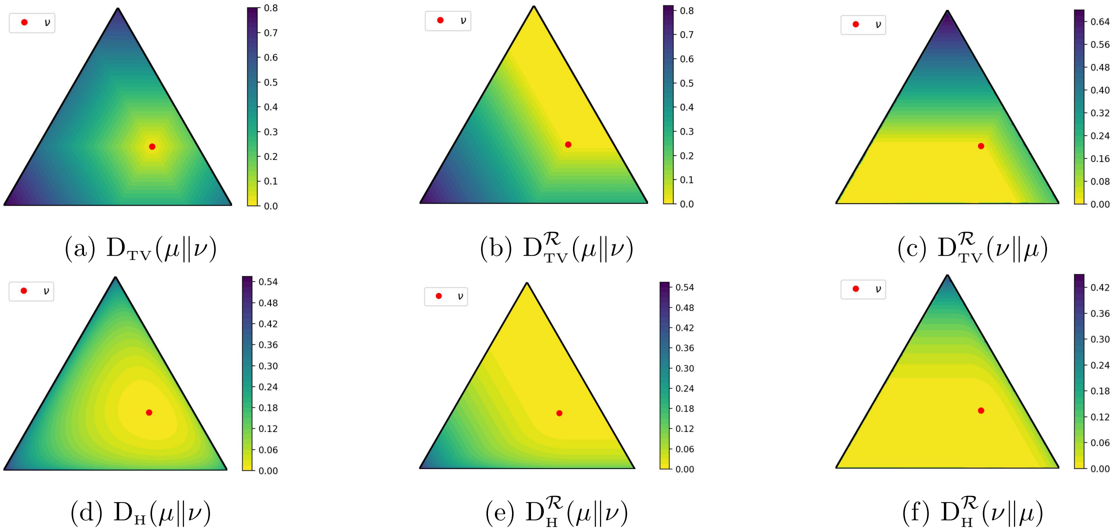
Definition 1 ($(\mathcal{R}, f)$-divergence)**.** *For any $f \in \mathcal{F}$, define the $(\mathcal{R}, f)$-divergence as*

$$\mathrm{D}_f^{\mathcal{R}}(\mu \| \nu) = \int f \left(\mathrm{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \right) \mathrm{d}\nu,$$

- When $\mathcal{R}$ is a σ-algebra, the $(\mathcal{R}, f)$-divergence reduces to standard f-divergence.
- We developed some basic properties for $(\mathcal{R}, f)$-divergence.

Proposition 1 (Basic Properties)**.** *Let μ, ν be probability measures on measurable space (Ω, Σ), then the $(\mathcal{R}, f)$-divergence has*

1. *Linearity:* $\mathrm{D}^{\mathcal{R}}_{\sum_{i=1}^n a_i f_i}(\mu \| \nu) = \sum_{i=1}^n a_i \, \mathrm{D}^{\mathcal{R}}_{f_i}(\mu \| \nu), a_i \geq 0$*;*
2. *Non-negativity:* $\mathrm{D}^{\mathcal{R}}_f(\mu \| \nu) \geq 0$*;*
3. *Let* $g(x) = f(x) + c(x-1)$*, then* $\mathrm{D}^{\mathcal{R}}_f(\mu \| \nu) = \mathrm{D}^{\mathcal{R}}_g(\mu \| \nu)$*;*
4. $\mathrm{D}^{\mathcal{R}}_f(\mu \| \nu) \leq \mathrm{D}_f(\mu \| \nu)$*;*
5. *if* $\mathrm{D}^{\mathcal{R}}_f(\mu \| \nu) = \mathrm{D}^{\mathcal{R}}_f(\nu \| \mu) = 0$*, then* $\mu = \nu$*.*



(a) $\mathrm{D}_{\mathrm{TV}}(\mu \| \nu)$

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

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

(d) $\mathrm{D}_{\mathrm{H}}(\mu \| \nu)$

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

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

Figure 2: 40 Level curves of total variation D_{TV}, $\mathcal{R}$-partial variation $\mathrm{D}^{\mathcal{R}}_{\mathrm{TV}}$, Hellinger distance D_{H} and $\mathcal{R}$-Hellinger distance $\mathrm{D}^{\mathcal{R}}_{\mathrm{H}}$ for fixed $\nu = [0.2, 0.5, 0.3]$ as μ ranges over the simplex of distributions on a three-element set.

General Glivenko–Cantelli theorem

Theorem 3. *Let ν is a probability measure, ν_n is the empirical measures, then*

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

almost surely.

General joint range theorem

we can define the joint range w.r.t. $(\mathcal{R}, f)$ and $(\mathcal{R}, g)$-divergence as

$$\mathcal{I}^{\mathcal{R}} = \left\{ (D^{\mathcal{R}}_f(\mu \| \nu), D^{\mathcal{R}}_g(\mu \| \nu)) : \mu, \nu \begin{array}{l} \text{are probability measures on} \\ \text{some measurable space} \end{array} \right\};$$

$$\mathcal{I}^{\mathcal{R}}_k = \left\{ (D^{\mathcal{R}}_f(\mu \| \nu), D^{\mathcal{R}}_g(\mu \| \nu)) : \mu, \nu \text{ are probability measures on } [k] \right\}.$$

Theorem 4.

$$\mathcal{I}^{\mathcal{R}} = \mathrm{co}(\mathcal{I}_2^{\mathcal{R}}) = \mathcal{I}_4^{\mathcal{R}}.$$

where co *denotes the convex hull with a natural extension of convex operations to* $[0, \infty]^2$.

- An important corollary is: any inequality between f-divergence holds for $(\mathcal{R}, f)$-divergence as well.
- For example, Pinsker's inequality

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

- TV and (Squared) Hellinger

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

etc.

A UNIFICATION OF THE KOLMOGOROV–SMIRNOV DISTANCE AND f-DIVERGENCE

HAOMING WANG AND LEK-HENG LIM

Abstract. In this work, we have established an integral representation of the Kolmogorov–Smirnov (KS) distance, analogous to that of the total variation (TV) distance. This led us to define a generalized f-divergence associated with a subcollection of the ground σ-algebra, and we have developed several properties of this generalized f-divergence. Additionally, within our general framework, we have incorporated the Harremoës-Vajda Joint Range Theorem and the Glivenko–Cantelli Theorem, often referred to as the Fundamental Theorem of Statistics.

1. Introduction

1.1. Glivenko–Cantelli theorem.

Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space, and X is a random variable on Ω, let $\nu = \mathbb{P}_{\#X}$ be the distribution of X, thus ν is a probability measure on real Borel space $(\mathbb{R}, \mathcal{B})$. Let $(\mathbb{R}, \mathcal{B}, \nu)$ be probability space defined as above, let $X_1, \dots, X_n \sim \nu$ i.i.d. Then define the empirical distribution as a random measure

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

for any $A \in \mathcal{B}$. Fix $w \in \Omega$, the empirical distribution satisfies the probability axioms and hence is a probability measure on space $(\mathbb{R}, \mathcal{B})$; fix $A \in \mathcal{B}$, the empirical distribution is a random variable on Ω. For any $A \in \mathcal{B}$, the strong law of large number shows $|\nu_n(A) - \nu(A)| \to 0$ almost surely. However, the following example shows such convergence is not uniform with respect to $A \in \Sigma$.

Example 1.1 ([1]). Suppose ν is the uniform measure on $([0,1], \mathcal{B}_{[0,1]}), X_1, \dots, X_n \sim \nu$ i.i.d., and ν_n is the corresponding empirical distribution. Then for each $n \in \mathbb{N}$, and any $w \in \Omega$, there exists set $C_{n,w} = \{X_1(w), \dots, X_n(w)\} \subseteq [0,1]$ such that $\nu_n(C_{n,w}) = 1$ but $\nu(C_{n,w}) = 0$, thus $\sup_{A \in \mathcal{B}_{[0,1]}} |\nu_n(A) - \nu(A)| \equiv 1$ for any n.

Definition 1.2 (Class of Rays)**.** Define the class of open rays as $\mathcal{R}_{)} = \{(-\infty, a) : a \in \mathbb{R}\} \cup \{\varnothing, \mathbb{R}\}$, the class of closed rays as $\mathcal{R}_{]} = \{(-\infty, a] : a \in \mathbb{R}\} \cup \{\varnothing, \mathbb{R}\}$, and the class of rays $\mathcal{R} = \mathcal{R}_{]} \cup \mathcal{R}_{)}$.

Proposition 1.3. *Let μ, ν be Borel probability measures, then*

$$\sup_{A \in \mathcal{R}_{]}} (\mu(A) - \nu(A)) = \sup_{A \in \mathcal{R}_{)}} (\mu(A) - \nu(A)).$$

Proof. By the continuity of probability measure, for any $t \in \mathbb{R}$, set $\epsilon \downarrow 0$, we have $\mu(-\infty, t + \epsilon) \downarrow \mu(-\infty, t]$ and $\nu(-\infty, t + \epsilon) \downarrow \nu(-\infty, t]$. Let

$$A = \{\mu(-\infty, t] - \nu(-\infty, t] : t \in \mathbb{R}\} \subseteq \mathbb{R}$$
$$B = \{\mu(-\infty, t) - \nu(-\infty, t) : t \in \mathbb{R}\} \subseteq \mathbb{R}$$

Then for any $x \in A$, there exists a sequence $x_n \in B$ such that $x_n \to x$, which implies $\sup A \le \sup B$. The other direction can be proved by same way, thus $\sup A = \sup B$. □

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, joint range, Glivenko–Cantelli theorem.

The Kolmogorov–Smirnov Distance between probability measures μ, ν is defined as $\sup_{A \in \mathcal{R}_]} |\mu(A) - \nu(A)|$. However, for simplicity, we will ignore the absolute value and just consider the quaitity $\sup_{A \in \mathcal{R}_]} (\mu(A) - \nu(A))$. The proposition 1.3 implies $\sup_{A \in \mathcal{R}_]} (\mu(A) - \nu(A)) = \sup_{A \in \mathcal{R}} (\mu(A) - \nu(A))$. Thus in our following analysis, the Kolmogorov–Smirnov Distance refers to $\sup_{A \in \mathcal{R}} (\mu(A) - \nu(A))$ without further instruction.

Theorem 1.4 (Glivenko–Cantelli)**.** *Let ν be the target probability measure on measurable space $(\mathbb{R}, \mathcal{B})$, and ν_n be the empirical measures, then*

$$\sup_{A \in \mathcal{R}} (\nu_n(A) - \nu(A)) \to 0, \quad \textit{almost surely.} \tag{1}$$

Glivenko–Cantelli Theorem is also called the Fundamental Theorem of Statistics. The theorem is a cornerstone of empirical process theory, which deals with the asymptotic behavior of stochastic processes formed by empirical measures [2]. It is also crucial in non-parametric statistics and forms the basis for statistical consistency in many estimation problems [2]. In machine learning, the Glivenko–Cantelli theorem is used to show the consistency of the ERM principle [3] and helps derive bounds on generalization error by ensuring the difference between empirical and true distributions diminishes with larger samples, aiding in over-fitting control [4]. The Glivenko–Cantelli theorem also plays a role in the Kolmogorov–Smirnov test. The theorem ensures that the test statistic has desirable asymptotic properties [5]. In statistical resampling methods like the bootstrap, the Glivenko–Cantelli theorem guarantees that the empirical distribution derived from resampled data will approximate the true distribution well as the number of samples grows [6].

1.2. **Motivation.**

Definition 1.5 (Total Variation)**.** Let μ, ν be probability measures on $(\mathbb{R}, \mathcal{B})$, the total variation distance between them is defined by

$$\mathrm{D}_{\mathrm{TV}}(\mu \| \nu) = \sup_{A \in \Sigma} (\mu(A) - \nu(A)) .$$

Let $\mathcal{F} := \{f$ be a continuous convex function such that $f(1) = 0\}$. Let $(\mathbb{R}, \mathcal{B})$ be a measurable space and μ, ν be probability measures on $(\mathbb{R}, \mathcal{B})$. Assume that $\nu \ll \mu$, then the Radon–Nikodym derivate $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ exists. Then, we can define the f-divergence as

$$\mathrm{D}_f(\mu \| \nu) = \int f\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\nu.$$

For different selection of f, we can define Kullback–Liebler [7], Le Cam [8], Jensen–Shannon [9] , Jeffreys [10], Chernoff [11], Pearson χ^2 [12], Hellinger squared [13], exponential [14], and alpha–beta [15] -divergences, and so on. And in particular, let $f(t) = |t-1|/2$, we can obtain the total variation distance [16] by

$$\mathrm{D}_{\mathrm{TV}}(\mu \| \nu) = \sup_{A \in \Sigma} (\mu(A) - \nu(A)) = \int \frac{1}{2} \left| \frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1 \right| \mathrm{d}\nu. \tag{2}$$

The set A in equation (2) ranges over the σ-algebra Σ and hence we call D_{TV} the *Total Variation*. Heuristically, if we constrain the set A to range over the class $\mathcal{R} \subseteq \Sigma$, we may call the distance defined by $\mathrm{D}_{PV}(\mu \| \nu) := \sup_{A \in \mathcal{R}} (\mu(A) - \nu(A))$ the *Partial Variation*. As we can see, the total variation is a special case of f-divergence when $f(t) = |t-1|/2$, then a natural question is, can we express such "Partial variation" as a special case of f-divergence type of statistical distance? In words,

> *(Q1) Can we have a similar expression like equation (2) when we replace the σ-algebra Σ with the class of rays $\mathcal{R}$?*

If the answer is yes, we can extend such partial variation to other f-divergence, and we will call these the $(\mathcal{R}, f)$-divergence. And write the $(\mathcal{R}, f)$-divergence distance between probability measures μ, ν as $\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu)$. Furthermore, we can generalize the Glivenko–Cantelli theorem to $(\mathcal{R}, f)$-divergence for all $f \in \mathcal{F}$. Hence, although $\mathrm{D}_f(\nu_n\|\nu)$ does not converge to zero almost surely, we can define a f-divergence type of statistical distance–$(\mathcal{R}, f)$-divergence that guarantees $\mathrm{D}_f^{\mathcal{R}}(\nu_n\|\nu)$ converges to zero almost surely.

To answer the question Q1, we need a concept *Radon–Nikodym Property*, The Radon–Nikodym theorem is, for σ-finite measure μ, ν on measurable space $(\mathbb{R}, \mathcal{B})$ s.t. $\nu \ll \mu$ then there exist a Σ-measurable function ρ_Σ such that $\mu(A) = \int_A \rho_\Sigma \, \mathrm{d}\nu$, for any $A \in \Sigma$. And Radon–Nikodym Property is that, if $\mathcal{G} \subseteq \Sigma$ is a sub σ-algebra of Σ, then there exist a $\mathcal{G}$-measurable function $\rho_\mathcal{G}$ such that

$$\mu(A) = \int_A \rho_\mathcal{G} \, \mathrm{d}\nu, \quad \text{for any } A \in \mathcal{G}.$$

then a nature question is, what if $\mathcal{G}$ is not a σ-algebra? for instance $\mathcal{G} = \mathcal{R}$ in our case:

> *(Q2) Is there a $\mathcal{R}$-measurable function $\rho_\mathcal{R}$ and a sub-collection $\mathcal{E} \subseteq \mathcal{R}$ such that $\mu(A) = \int_A \rho_\mathcal{R} \, \mathrm{d}\nu$ for any $A \in \mathcal{E}$?*

We will first answer the question Q2 in Theorem 2.13, and then solve the question Q1 based on it in theorem 2.21 in section 2. Based on these results, we can define the $(\mathcal{R}, f)$-divergence and develop good properties for $(\mathcal{R}, f)$-divergence in section 3; and finally generalize the Glivenko–Cantelli theorem in section 4.

2. $(\mathcal{R}, f)$ - DIVERGENCE

An important tool we will use is metric projection. Let (X, d) be a metric space, $M \subseteq X$ is nonempty, then the *metric projection* onto M, denoted by $\mathrm{proj}_M : X \to 2^M$ is a set-value function:

$$\mathrm{proj}_M(x) \coloneqq \left\{ y \in M : d(x, y) = \operatorname*{arg\,min}_{z \in M} d(x, z) \right\}.$$

In other words, $\mathrm{proj}_M(x)$ is the points in M that have the same nearest distance to x. When $\mathrm{proj}_M(x)$ is single-valued for any $x \in X$, the set M is called a *Chebyshev set*. If X is a Hilbert space, M is a convex and closed subset of X, M is a Chebyshev set, and proj_M is continuous [17].

Let $(\mathbb{R}, \mathcal{B}, \nu)$ be a probability space, and $L_2(\mathbb{R}, \mathcal{B}, \nu) =: L_2(\nu)$ is the space of measurable, square-integrable functions, in which two functions that differ on a null set are equivalent. Thus, when we say a function f is non-increasing, we ignore the case which does not hold on a null set. $L_2(\nu)$ is a Hilbert space, let $L_2^+(\nu) \coloneqq \{f \in L_2(\nu) : f \geq 0\}$. We will abbreviate them as L_2 and L_2^+ if ν is given in context. We first develop some properties that will be used. Let ν is a Borel probability measure on $(\mathbb{R}, \mathcal{B})$, and $\mathcal{M}_\mathcal{R}$ is the collection of *$\mathcal{R}$-measurable function* in $L_2(\nu)$:

$$\mathcal{M}_\mathcal{R} = \{f \in L_2(\nu) : (f > r) \in \mathcal{R}, r \in \mathbb{R}\}.$$

Let $f \in L_2$ and (f_n) be a sequence of measurable functions. We say f_n converge to f ν-almost surely (a.s.), and denoted as $f_n \xrightarrow{a.s.} f$, if f_n converges to f point-wise except for a ν-null set. We say f_n converge to f in ν-probability, denoted as $f_n \xrightarrow{\nu} f$, if $\lim_{n\to\infty} \nu(|f_n - f| > \epsilon) = 0$ for any $\epsilon > 0$. We say f_n converge to f in L_2-norm, denoted as $f_n \xrightarrow{L_2} f$, if $\|f_n - f\|_{L_2} = (\int (f_n - f)^2 \, \mathrm{d}\nu)^{1/2} \to 0$.

Lemma 2.1 (Modes of convergence, [18])**.** *Let ν be a Borel probability measure, then*
(1) $f_n \xrightarrow{L_2} f$ implies $f_n \xrightarrow{\nu} f$;
(2) $f_n \xrightarrow{\nu} f$ implies there is a sub-sequence f_{n_k} such that $f_{n_k} \xrightarrow{a.s.} f$;
(3) $f_n \xrightarrow{a.s.} f$, and there exists $g \in L_2(\nu)$ dominates f_n, then $f_n \xrightarrow{L_2} f$.

Lemma 2.2. *If $f_n, f \in L_p$ and $f_n \xrightarrow{L_p} f$, $1 \leq p < \infty$, then $\int f \, \mathrm{d}\nu = \lim_{n\to\infty} \int f_n \, \mathrm{d}\nu$.*

Proof.

$$\left|\int f_n - f \,\mathrm{d}\nu\right| \leq \int |f_n - f| \,\mathrm{d}\nu \leq \left(\int (f_n - f)^p \,\mathrm{d}\nu\right)^{1/p} = \|f_n - f\|_{L_p} \to 0.$$

□

Proposition 2.3. *$\mathcal{M}_{\mathcal{R}}$ is the collection of non-increasing functions in $L_2(\nu)$.*

Proof. If f is a non-increasing function, for any $r \in \mathbb{R}$, let $a = \sup\{x : f(x) > r\}$, then for any $x \in \{f > r\}, x \leq \sup\{f > r\} = a$, thus $\{f > r\} \subseteq (-\infty, a]$. If there exists $x < a = \sup\{f > r\}$ such that $f(x) \leq r$, then there exists $y > x$ such that $f(y) > r \geq f(x)$, which is a contradiction since f is non-increasing. Thus $(\infty, a) \subseteq \{f > r\} \subseteq (-\infty, a]$, and then $\{f > r\} \in \mathcal{R}, f \in \mathcal{M}_{\mathcal{R}}$.

For the other direction, if $f \in \mathcal{M}_{\mathcal{R}}$. Assume that there exists $x < y$ such that $f(x) < f(y)$. There exists a such that $(-\infty, a) \subseteq \{f > f(x)\} \subseteq (-\infty, a]$. Since $f(y) > f(x)$, then $y \in \{f > f(x)\}$, thus $x < y < a$, which implies $x \in \{f > f(x)\}$ i.e. $f(x) > f(x)$, which is contradictory. Thus, f is non-increasing. □

Proposition 2.4. *$\mathcal{M}_{\mathcal{R}}$ is a closed convex cone in $L_2(\nu)$.*

Proof. It suffices to show that the non-negative linear combination and the L_2-norm limit of non-increasing functions are still non-increasing. (1) Let f, g are non-increasing functions, then for $a \leq b$, $f(a) \leq f(b)$, $g(a) \leq g(b)$. Let $\lambda, \eta \geq 0$, then $\lambda f(a) + \eta g(a) \leq \lambda f(b) + \eta g(b)$, thus $\lambda f + \eta g$ is non-increasing. (2) Let (f_n) be a sequence of non-increasing functions, and $\|f_n - f\|_{L_2} \to 0$, then by Lemma 2.1, $f_n \to f$ in ν-measure, and hence there exists a sub-sequence (f_{n_k}) that converges to f ν-almost surely. In words, there exists a null set A, such that $\lim_{k\to\infty} f_{n_k}(x) = f(x)$ for $x \in A^c$, where f_{n_k} is non-increasing, thus f is non-increasing (a.s.). □

Thus, $\mathcal{M}_{\mathcal{R}}$ is a Chebyshev set, and we would be interested in the metric projection on it. Given $f \in L_2(\nu)$, to study $\operatorname{proj}_{\mathcal{M}_{\mathcal{R}}} f$, it is equivalent to study the nearest non-increasing function to f in L_2-norm.

Remark 2.5. The reason we do not choose $\mathcal{R}_]$ is that $\mathcal{M}_{\mathcal{R}_]}$ is not a closed set, and hence we can not discuss the metric projection on it properly. Firstly, any function $f \in \mathcal{M}_{\mathcal{R}_]}$ (except for the constant function) can not be continuous because the preimage of the open set $f^{-1}(r, \infty) = \{f > r\} = (-\infty, a]$ is not open. Let $f_n = -\sum_{k\in\mathbb{Z}} \frac{k}{n} \cdot \mathbb{1}_{(\frac{k}{n}, \frac{k+1}{n}]}$, then $\mathcal{M}_{\mathcal{R}_]} \ni f_n \xrightarrow{L_2} f = -x \notin \mathcal{M}_{\mathcal{R}_]}$.

Class $\{A_1, \ldots, A_n\}$ is called an *order partition* of $\mathbb{R}$ if (1) $A_i \cap A_j = \varnothing$ for any $i \neq j$, (2) $\bigcup_{i=1}^n A_i = \mathbb{R}$ and (3) $\sup A_i \leq \inf A_j$ for each $i \leq j$. A function $f : \mathbb{R} \to \mathbb{R}$ is called an *order simple function* if it is a simple function on an order partition $\{A_1, \ldots, A_n\}$. A well-known result is that for any $f \in L_2^+$, there exists a sequence of non-negative simple function g_n increasingly approximate f. We aim to develop a similar result for order simple function in Corollary 2.7.

Proposition 2.6. *Let ν be a probability measure on $(\mathbb{R}, \mathcal{B})$, $f \in L_2^+$ is bounded and continuous, then there exists a sequence of non-negative order simple function g_n such that $g_n \uparrow f$.*

Proof. Let order partition to be $\mathcal{P}_n = \left\{[\frac{nk}{2^n}, \frac{n(k+1)}{2^n}) : k = -2^n, -2^n + 1, \ldots, 2^n - 1\right\} \cup \{(-\infty, -n), [n, \infty)\}$. Then $\mathcal{P}_n$ is an order partition, then $g_n = \sum_{A\in\mathcal{P}_n} \inf_{x\in A} f(x) \cdot \mathbb{1}_A$ is order simple function. Then for any $\epsilon > 0$, $x \in \mathbb{R}$, there is a $n \in \mathbb{N}$ and $A_{(n)} \in \mathcal{P}_n$ such that $x \in A_{(n)} \subseteq B_\epsilon(x)$, and hence

$$\inf_{y\in B_\epsilon(x)} f(y) \leq \inf_{y\in A_{(n)}} f(y) = g_n(x) \leq f(x),$$

where $\inf_{y\in B_\epsilon(x)} f(y) \uparrow f(x)$ as $\epsilon \downarrow 0$ since f is continuous. Thus $g_n(x) \to f(x)$ as $n \to \infty$. Sequence g_n is increasing since for any $n \leq m$, then $A \subseteq B$ for any $A \in \mathcal{P}_m$, $B \in \mathcal{P}_n$, hence $g_n(x) \leq g_m(x)$ for any $x \in \mathbb{R}$. □

Thus, for any continuous function $f \in L_2^+$, there exists a sequence of order simple function g_n such that converges to f point-wise and is dominated by f, then by Lemma 2.1, $g_n \xrightarrow{L_2} f$. Since the space $C_c(\mathbb{R})$ of continuous functions with compact support is dense in $L_2(\mathbb{R})$ [19], then for any $\epsilon > 0$, there exists a continuous functions with compact support φ such that $\|f - \varphi\|_{L_2} < \epsilon/2$. Since φ is continuous and has compact support, then it is bounded and hence in L_2, and there exists an order simple function g such that $\|\varphi - g\|_{L_2} < \epsilon/2$, and hence

$$\|f - g\|_{L_2} \leq \|f - \varphi\|_{L_2} + \|\varphi - g\|_{L_2} < \epsilon.$$

In words, for any $f \in L_2^+$, there exists a sequence of order simple function g_n such that $g_n \xrightarrow{L_2} f$. Furthermore, by Lemma 2.1, there exists a sub-sequence $\pi(n)$ such that $h_n \coloneqq g_{\pi(n)} \xrightarrow{a.s.} f$. Note that for each n, g_n is bounded by a bounded continuous function, thus $\|g_n\|_{L_\infty} < \infty$. However, there is no guarantee such that $\sup_n \|g_n\|_{L_\infty} < \infty$.

Corollary 2.7. *Let $f \in L_2^+$, then there exists sequence of bounded order simple functions (g_n) such that $g_n \xrightarrow{L_2} f$, and $g_n \xrightarrow{a.s.} f$. If f is also continuous a.s. then there exists sequence of bounded order simple functions (g_n) such that $g_n \uparrow f$ a.s.*

Since we will only discuss functions in L_2^+, i.e. space of equivalence classes of non-negative integrable functions, assume that N is a null set such that $h_n \to f$ point-wise in $\mathbb{R}\backslash N$, then for each n, let $h_n' = h_n \mathbb{1}_{\mathbb{R}\backslash N} + f \mathbb{1}_N$, which is equivalent with h_n, and $h_n \to f$ point-wise. Thus, we will omit "almost surely" in the following analysis.

Proposition 2.8. *Let ν be a probability measure on $(\mathbb{R}, \mathcal{B})$, $g = \sum_{i=1}^n \alpha_i \mathbb{1}_{A_i} \in L_2^+$ is an order simple function, then $\mathrm{proj}_{\mathcal{M}_\mathcal{R}} g$ is also an order simple function on order partition $\{A_1, \ldots, A_n\}$.*

Proof. Assume that $\mathrm{proj}_{\mathcal{M}_\mathcal{R}} g$ is not constant on A_j for some $j \in [n]$. Then define

$$\varphi = \begin{cases} \mathrm{proj}_{\mathcal{M}_\mathcal{R}} g, & w \in A_j^c \\ c, & w \in A_j \end{cases}$$

where $c = \mathrm{proj}_{\mathcal{M}_\mathcal{R}} g(w^*)$ and $w^* = \arg\min_{w \in \bar{A}_j} \left|\mathrm{proj}_{\mathcal{M}_\mathcal{R}} g(w) - g\right|$, then $\varphi \in \mathcal{M}_\mathcal{R}$ and

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

Which leads to a contradiction. □

Proposition 2.9 (Monotonicity)**.** *Let ν be a Borel probability measure, for $f_1, f_2 \in L_2^+(\nu)$,*

$$f_1 \leq f_2 \Rightarrow \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_1 \leq \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_2$$

Proof. We first show it is true for order simple functions $f_1 = \sum_{i=1}^n a_i \mathbb{1}_{A_i}$, $f_2 = \sum_{j=1}^m b_j \mathbb{1}_{B_j}$, where $(A_1, \ldots, A_n)$, $(B_1, \ldots, B_m)$ are order partitions. Assume that $f_1 \leq f_2$, and $\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_1 > \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_2$ on $E = A_i \cap B_j$ for some $i \in [n]$, $j \in [m]$. First, functions $\varphi = \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_1 \cdot \mathbb{1}_{E^c} + \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_2 \cdot \mathbb{1}_E$ and $\psi = \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_2 \cdot \mathbb{1}_{E^c} + \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_1 \cdot \mathbb{1}_E$ are both non-increasing functions, and hence $\mathcal{R}$-measurable.

There are six cases on E:

(1) $f_2 \geq f_1 \geq \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_1 \geq \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_2$, (2) $f_2 \geq \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_1 \geq f_1 \geq \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_2$

(3) $f_2 \geq \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_1 \geq \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_2 \geq f_1$, (4) $\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_1 \geq f_2 \geq f_1 \geq \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_2$

(5) $\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_1 \geq f_2 \geq \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_2 \geq f_1$, (6) $\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_1 \geq \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_2 \geq f_2 \geq f_1$

Cases (1), (2) and (3) imply $\|\psi - f_2\|_{L_2} \leq \left\|\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_2 - f_2\right\|_{L_2}$; Cases (5) and (6) imply $\|\varphi - f_1\|_{L_2} \leq \left\|\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_1 - f_1\right\|_{L_2}$; and case (4) implies either $\|\psi - f_2\|_{L_2} \leq \left\|\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_2 - f_2\right\|_{L_2}$ or $\|\varphi - f_1\|_{L_2} \leq \left\|\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f_1 - f_1\right\|_{L_2}$. Thus, all of these cases lead to contradiction.

In general case, if $f_1, f_2 \in L_2^+$, then $f_0 \coloneqq f_2 - f_1 \in L_2^+$. There exists sequence of non-negative order simple functions $(g_{0,n})$, $(g_{1,n})$ such that $g_{0,n} \xrightarrow{L_2} f_0$ and $g_{1,n} \xrightarrow{L_2} f_1$, hence $g_{2,n} \coloneqq g_{0,n} + g_{1,n}$ is non-negative order simple function such that $g_{2,n} \xrightarrow{L_2} f_2$ and $g_{2,n} \geq g_{1,n}$ for each n. Then by the continuity of operator $\operatorname{proj}_{\mathcal{M}_\mathcal{R}}$, we have

$$\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f_1 = \lim_{n\to\infty} \operatorname{proj}_{\mathcal{M}_\mathcal{R}} g_{1,n} \leq \lim_{n\to\infty} \operatorname{proj}_{\mathcal{M}_\mathcal{R}} g_{2,n} = \operatorname{proj}_{\mathcal{M}_\mathcal{R}} f_2.$$

□

Remark 2.10. For any $f \in L_2^+$, let g_n be the sequence of order simple function such that $g_n \to f$ a.s. and in L_2 norm; by the continuity of operator $\operatorname{proj}_{\mathcal{M}_\mathcal{R}}$, $\operatorname{proj}_{\mathcal{M}_\mathcal{R}} g_n \xrightarrow{L_2} \operatorname{proj}_{\mathcal{M}_\mathcal{R}} f$. Then there exists a sub-sequence $\pi(n)$ such that both $g_{\pi(n)} \to f$ a.s. and in L_2 norm, and $\operatorname{proj}_{\mathcal{M}_\mathcal{R}} g_n \to \operatorname{proj}_{\mathcal{M}_\mathcal{R}} f$ a.s. and in L_2 norm.

Corollary 2.11. *Let ν be a Borel probability measure, $f_n, f \in L_2^+(\nu)$, and $f_n \uparrow f$, then* $\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f_n \uparrow \operatorname{proj}_{\mathcal{M}_\mathcal{R}} f$.

Proposition 2.12. *Let ν be a Borel probability measure, for any $A \in \mathcal{R}$,* $\operatorname{proj}_{\mathcal{M}_\mathcal{R}}(f \cdot \mathbb{1}_A) = \operatorname{proj}_{\mathcal{M}_\mathcal{R}}(f \cdot \mathbb{1}_A) \cdot \mathbb{1}_A$. *Hence For any $A \in \mathcal{R}$*

$$\operatorname{proj}_{\mathcal{M}_\mathcal{R}}(f \cdot \mathbb{1}_A) \leq (\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f) \cdot \mathbb{1}_A.$$

Proof.

$$\begin{aligned}\int (f \cdot \mathbb{1}_A - &\operatorname{proj}_{\mathcal{M}_\mathcal{R}}(f \cdot \mathbb{1}_A))^2 \,\mathrm{d}\nu \\ &= \int_A \big(f \cdot \mathbb{1}_A - \operatorname{proj}_{\mathcal{M}_\mathcal{R}}(f \cdot \mathbb{1}_A)\big)^2 \,\mathrm{d}\nu + \int_{A^c} \big(f \cdot \mathbb{1}_A - \operatorname{proj}_{\mathcal{M}_\mathcal{R}}(f \cdot \mathbb{1}_A)\big)^2 \,\mathrm{d}\nu \\ &= \int_A \big(f \cdot \mathbb{1}_A - \operatorname{proj}_{\mathcal{M}_\mathcal{R}}(f \cdot \mathbb{1}_A)\big)^2 \,\mathrm{d}\nu + \int_{A^c} \big(\operatorname{proj}_{\mathcal{M}_\mathcal{R}}(f \cdot \mathbb{1}_A)\big)^2 \,\mathrm{d}\nu \\ &\geq \int_A \big(f \cdot \mathbb{1}_A - \operatorname{proj}_{\mathcal{M}_\mathcal{R}}(f \cdot \mathbb{1}_A)\big)^2 \,\mathrm{d}\nu \\ &= \int \big(f \cdot \mathbb{1}_A - \operatorname{proj}_{\mathcal{M}_\mathcal{R}}(f \cdot \mathbb{1}_A) \cdot \mathbb{1}_A\big)^2 \,\mathrm{d}\nu\end{aligned}$$

since the projection is unique, then $\operatorname{proj}_{\mathcal{M}_\mathcal{R}}(f \cdot \mathbb{1}_A) = \operatorname{proj}_{\mathcal{M}_\mathcal{R}}(f \cdot \mathbb{1}_A) \cdot \mathbb{1}_A$. Since $f \cdot \mathbb{1}_A \leq f$, then

$$\operatorname{proj}_{\mathcal{M}_\mathcal{R}}(f \cdot \mathbb{1}_A) = \operatorname{proj}_{\mathcal{M}_\mathcal{R}}(f \cdot \mathbb{1}_A) \cdot \mathbb{1}_A \leq (\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f) \cdot \mathbb{1}_A.$$

□

Theorem 2.13 (Answer to Question (Q2))**.** *Let ν be a Borel probability measure, $f \in L_2^+$, let $\mathcal{E} = \big\{(\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f > r) : r \geq 0\big\} \cup \{\mathbb{R}\}$, then for any $E \in \mathcal{E}$*

$$\int_E \operatorname{proj}_{\mathcal{M}_\mathcal{R}} f \,\mathrm{d}\nu = \int_E f \,\mathrm{d}\nu. \tag{3}$$

Proof. We prove this theorem in four steps:
(1) it is true if f is an order simple function;
(2) it is true for $E = \big\{\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f > r\big\}$ if $\nu(\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f = r) = 0$;
(3) it is true for any $E = \big\{\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f > r\big\}$;
(4) it is true for $E = \mathbb{R}$.

Step (1). We first show that it is true for the order simple function g. Let $g = \sum_{i=1}^n \alpha_i \mathbb{1}_{A_i}$, $\mathrm{proj}_{\mathcal{M}_\mathcal{R}} g = \sum_{i=1}^n \beta_i \mathbb{1}_{A_i}$, and $E = \left\{\mathrm{proj}_{\mathcal{M}_\mathcal{R}} g > r\right\} = \bigcup_{i=1}^k A_i \in \mathcal{E}$ such that $\beta_k > r$ and $\beta_{k+1} \le r$.

Let $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)^\top$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_n)^\top$, and $\boldsymbol{w} = (w_1, \dots, w_n)^\top$ where $w_i = \nu(A_i)$ and $W = \mathrm{diag}(\boldsymbol{w})$. Then $\boldsymbol{\beta}$ is the solution of the following quadratic programming problem:

$$\begin{aligned} \min \quad & (\boldsymbol{\alpha} - \boldsymbol{\beta})^\top W (\boldsymbol{\alpha} - \boldsymbol{\beta}) \\ \text{s.t.} \quad & \beta_1 \ge \beta_2 \ge \cdots \ge \beta_n \ge 0 \end{aligned}$$

Let

$$\Pi = \begin{bmatrix} 1 & -1 & & \\ & 1 & -1 & \\ & & \ddots & -1 \\ & & & 1 \end{bmatrix}, \quad \text{and hence} \quad \Pi^{-1} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ & 1 & \dots & 1 \\ & & \ddots & \vdots \\ & & & 1 \end{bmatrix}$$

let $\boldsymbol{\gamma} = \Pi \boldsymbol{\beta}$, then the equivalent optimization problem is

$$\begin{aligned} \min \quad & \boldsymbol{\gamma}^\top \Pi^{-\top} W \Pi^{-1} \boldsymbol{\gamma} - 2\boldsymbol{\alpha}^\top \Pi^{-1} W \boldsymbol{\gamma} \\ \text{s.t.} \quad & -\boldsymbol{\gamma} \preceq 0 \end{aligned}$$

then by KKT condition (stationarity) we have $\boldsymbol{\lambda} = 2\Pi^{-\top} W (\Pi^{-1}\boldsymbol{\gamma} - \boldsymbol{\alpha}) = 2\Pi^{-\top} W (\boldsymbol{\beta} - \boldsymbol{\alpha})$, that is

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix} = 2 \cdot \begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ \vdots & \vdots & \ddots & \\ 1 & 1 & \dots & 1 \end{bmatrix} \cdot \begin{bmatrix} w_1 & & & \\ & w_2 & & \\ & & \ddots & \\ & & & w_n \end{bmatrix} \cdot \begin{bmatrix} \beta_1 - \alpha_1 \\ \beta_2 - \alpha_2 \\ \vdots \\ \beta_n - \alpha_n \end{bmatrix} \tag{4}$$

Since $\beta_k = (\Pi^{-1}\boldsymbol{\gamma})_k = \gamma_k + \gamma_{k+1} + \cdots + \gamma_n > r$ and $\beta_{k+1} = \gamma_{k+1} + \gamma_{k+2} + \cdots + \gamma_n \le r$, we have that $\gamma_k > 0$, then by KKT condition (complementary slackness) we have $\lambda_k = 2(\Pi^{-\top} W(\boldsymbol{\beta} - \boldsymbol{\alpha}))_k = 2\sum_{i=1}^k w_i(\beta_i - \alpha_i) = 0$. Thus

$$\int (\mathrm{proj}_{\mathcal{M}_\mathcal{R}} g) \cdot \mathbb{1}_E \,\mathrm{d}\nu = \sum_{i=1}^k \beta_i w_i = \sum_{i=1}^k \alpha_i w_i = \int g \cdot \mathbb{1}_E \,\mathrm{d}\nu.$$

Step (2). For general $f \in L_2^+$, let g_n be the sequence of non-negative order simple functions such that $g_n \to f$ a.s. and in L_2 norm; and $\mathrm{proj}_{\mathcal{M}_\mathcal{R}} g_n \to \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f$ a.s. and in L_2 norm. Let $E = \left\{\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f > r\right\}$ and $E_n = \left\{\mathrm{proj}_{\mathcal{M}_\mathcal{R}} g_n > r\right\}$, assume $\nu(\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f = r) = 0$. Then

$$\begin{aligned} \|\mathbb{1}_E - \mathbb{1}_{E_n}\|_{L_2}^2 &= \int (\mathbb{1}_E - \mathbb{1}_{E_n})^2 \,\mathrm{d}\nu \\ &= \nu(E) + \nu(E_n) - 2\nu(E \cap E_n) = \nu(E \triangle E_n) \\ &= \nu(E \cap E_n^c) + \nu(E_n \cap E^c). \end{aligned}$$

Let $N = \left\{\mathrm{proj}_{\mathcal{M}_\mathcal{R}} g_n \not\to \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f\right\}$, thus N is a null set. $E \cap E_n^c = \left\{\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f > r\right\} \cap \left\{\mathrm{proj}_{\mathcal{M}_\mathcal{R}} g_n \le r\right\}$, thus if $x \in E \cap E_n^c$ infinitely often, then $\liminf_{n\to\infty} \mathrm{proj}_{\mathcal{M}_\mathcal{R}} g_n(x) \le r < \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f(x)$, that is $x \in N$ and hence $\limsup_{n\to\infty} E \cap E_n^c \subseteq N$. On the other hand,

$$E_n \cap E^c = \underbrace{\left[\left\{\mathrm{proj}_{\mathcal{M}_\mathcal{R}} g_n > r\right\} \cap \left\{\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f < r\right\}\right]}_{:=F_n} \cup \underbrace{\left[\left\{\mathrm{proj}_{\mathcal{M}_\mathcal{R}} g_n > r\right\} \cap \left\{\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f = r\right\}\right]}_{:=G_n}.$$

Where G_n is a null set for each n. If $x \in F_n$ infinitely often, then $\limsup_{n\to\infty} \operatorname{proj}_{\mathcal{M}_\mathcal{R}} g_n(x) \geq r > \operatorname{proj}_{\mathcal{M}_\mathcal{R}} f(x)$, thus $x \in N$ as well, hence $\limsup_{n\to\infty} F_n \subseteq N$. Thus

$$\begin{aligned}\limsup_{n\to\infty} \|\mathbb{1}_E - \mathbb{1}_{E_n}\|_{L_2}^2 &\leq \limsup_{n\to\infty} \Big[\nu(E \cap E_n^c) + \nu(F_n) + \nu(G_n)\Big] \\ &\leq \limsup_{n\to\infty} \nu(E \cap E_n^c) + \limsup_{n\to\infty} \nu(F_n) \\ &\leq \nu\left(\limsup_{n\to\infty} E \cap E_n^c\right) + \nu\left(\limsup_{n\to\infty} F_n\right) = 0\end{aligned}$$

Thus $\mathbb{1}_{E_n} \xrightarrow{L_2} \mathbb{1}_E$, then

$$\begin{align*}\|f\mathbb{1}_E - g_n\mathbb{1}_{E_n}\|_{L_1} &= \|f\mathbb{1}_E - f\mathbb{1}_{E_n} + f\mathbb{1}_{E_n} - g_n\mathbb{1}_{E_n}\|_{L_1} \\ &\leq \|f\mathbb{1}_E - f\mathbb{1}_{E_n}\|_{L_1} + \|f\mathbb{1}_{E_n} - g_n\mathbb{1}_{E_n}\|_{L_1} \\ &= \|f \cdot (\mathbb{1}_E - \mathbb{1}_{E_n})\|_{L_1} + \|(f - g_n) \cdot \mathbb{1}_{E_n}\|_{L_1} \\ &\leq \underbrace{\|f\|_{L_2}}_{<\infty} \cdot \underbrace{\|\mathbb{1}_E - \mathbb{1}_{E_n}\|_{L_2}}_{\to 0} + \underbrace{\|f - g_n\|_{L_2}}_{\to 0} \cdot \underbrace{\|\mathbb{1}_{E_n}\|_{L_2}}_{\leq 1} \tag{5} \\ &\to 0.\end{align*}$$

where inequality (5) is by Hölder's inequality. By same way, $\mathbb{1}_{E_n} \operatorname{proj}_{\mathcal{M}_\mathcal{R}} g_n \xrightarrow{L_1} \mathbb{1}_E \operatorname{proj}_{\mathcal{M}_\mathcal{R}} f$ as well, and hence equation (6) and (7) can be obtained by Lemma 2.2:

$$\begin{align*}\int f \cdot \mathbb{1}_E \, \mathrm{d}\nu &= \lim_{n\to\infty} \int g_n \cdot \mathbb{1}_{E_n} \, \mathrm{d}\nu \tag{6} \\ &= \lim_{n\to\infty} \int (\operatorname{proj}_{\mathcal{M}_\mathcal{R}} g_n) \cdot \mathbb{1}_{E_n} \, \mathrm{d}\nu \\ &= \int (\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f) \cdot \mathbb{1}_E \, \mathrm{d}\nu. \tag{7}\end{align*}$$

Step (3). Let $x = \inf\big\{\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f = r\big\}$. without loss of generality, assume $x > -\infty$. Then any $y < x$, $\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f(y) \geq r$, hence $\inf_{y\in(-\infty,x)} \operatorname{proj}_{\mathcal{M}_\mathcal{R}} f(y) \geq r$.
(i) If $\inf_{y\in(-\infty,x)} \operatorname{proj}_{\mathcal{M}_\mathcal{R}} f(y) > r$: Then there exists $\epsilon > 0$ such that $\inf_{y\in(-\infty,x)} \operatorname{proj}_{\mathcal{M}_\mathcal{R}} f(y) > r+\epsilon$, thus $E_r = \big\{\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f > r\big\} = \big\{\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f > r + \epsilon\big\} =: E_{r+\epsilon}$; and also, $\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f \neq r + \epsilon$ a.s. hence $\nu(\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f = r + \epsilon) = 0$. Thus

$$\int_{E_r} \operatorname{proj}_{\mathcal{M}_\mathcal{R}} f \, \mathrm{d}\nu = \int_{E_{r+\epsilon}} \operatorname{proj}_{\mathcal{M}_\mathcal{R}} f \, \mathrm{d}\nu = \int_{E_{r+\epsilon}} f \, \mathrm{d}\nu = \int_{E_r} f \, \mathrm{d}\nu.$$

(ii) If $\inf_{y\in(-\infty,x)} \operatorname{proj}_{\mathcal{M}_\mathcal{R}} f(y) = r$: Then for any $k \in \mathbb{N}$, there exists $x_k \in (-\infty, x)$ such that $r < \operatorname{proj}_{\mathcal{M}_\mathcal{R}} f(x_k) < r + \frac{1}{k}$. Since interval $(r, \operatorname{proj}_{\mathcal{M}_\mathcal{R}} f(x_k)]$ is uncountable, then there exists a $r_k \in (r, \operatorname{proj}_{\mathcal{M}_\mathcal{R}} f(x_k)]$ such that $r < r_k < r + \frac{1}{k}$ and $\nu(\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f = r_k) = 0$. Let $E_k = (\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f > r_k)$, then $E_k \uparrow$ and $E_k \subseteq E_r$, thus $\bigcup_k E_k \subseteq E_r$. On the other hand, if $y \in E_r$, i.e. $\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f(y) > r$, then there exists r_k such that $r_k < \operatorname{proj}_{\mathcal{M}_\mathcal{R}} f(y)$, thus $y \in E_k$, thus $E_k \uparrow E_r$. Since for any $A \in \mathcal{B}$,

$$\zeta(A) := \int_A \operatorname{proj}_{\mathcal{M}_\mathcal{R}} f \, \mathrm{d}\nu, \quad \xi(A) := \int_A f \, \mathrm{d}\nu$$

define finite measures respectively, then by the continuity from below, we have

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

Step (4). We now show that $\int \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f \, \mathrm{d}\nu = \int f \, \mathrm{d}\nu$. We first show it is true for an order simple function $g = \sum_{i=1}^n \alpha_i \mathbb{1}_{A_i}$. Without loss of generality, let $\beta_k > 0$ and $\beta_{k+1} = \cdots = \beta_n = 0$. Then, by equation (4) and the result in Step (1), we have

$$\begin{aligned}\lambda_n &= 2\sum_{i=1}^n w_i(\beta_i - \alpha_i) = \underbrace{2\sum_{i=1}^k w_i(\beta_i - \alpha_i)}_{=-\lambda_k=0} +2\sum_{j=k+1}^n w_i(\beta_i - \alpha_i) \\ &= -2(w_{k+1}\alpha_{k+1} + w_{k+2}\alpha_{k+2} + \cdots + w_n\alpha_n)\end{aligned}$$

By KKT condition (dual feasibility), we have $\lambda_n \geq 0$. Since $w_i, \alpha_i \geq 0$ for each $i = k+1, \dots, n$, then we have $\lambda_n = -2(w_{k+1}\alpha_{k+1} + w_{k+2}\alpha_{k+2} + \cdots + w_n\alpha_n) = 0$. Then from equation (4), we have $\lambda_n = 2\sum_{i=1}^n w_i(\beta_i - \alpha_i) = 0$, thus

$$\int \mathrm{proj}_{\mathcal{M}_\mathcal{R}} g \, \mathrm{d}\nu = \sum_{i=1}^n \beta_i w_i = \sum_{i=1}^n \alpha_i w_i = \int g \, \mathrm{d}\nu.$$

For general $f \in L_2^+$, let non-negative order simple functions $g_n \xrightarrow{L_2} f$, then by Lemma 2.1,

$$\int f \, \mathrm{d}\nu = \lim_{n\to\infty} \int g_n \, \mathrm{d}\nu = \lim_{n\to\infty} \int (\mathrm{proj}_{\mathcal{M}_\mathcal{R}} g_n) \, \mathrm{d}\nu = \int (\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f) \, \mathrm{d}\nu.$$

□

Lemma 2.14 ([18])**.** *Let μ, ν be finite measures on (Ω, Σ) such that $\mu(\Omega) = \nu(\Omega)$, $\mathcal{A}$ is a π-system that generates Σ, if $\mu(A) = \nu(A)$ for any $A \in \mathcal{A}$, then $\mu = \nu$.*

Lemma 2.15 ([20])**.** *If $f : X \to Y$ be any function, $\mathcal{E} \subseteq \mathcal{P}(Y)$, then*

$$\sigma(f^{-1}(\mathcal{E})) = f^{-1}(\sigma(\mathcal{E})).$$

Corollary 2.16. *Let ν be a Borel probability measure, $f \in L_2^+$, let $\mathcal{E} = \{(\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f > r) : r \geq 0\}$, then for any $E \in \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f^{-1}(\mathcal{B})$*

$$\int_E \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f \, \mathrm{d}\nu = \int_E f \, \mathrm{d}\nu.$$

Proof. Now, we have shown that the equation (3) is true for any $E \in \mathcal{E}$. Observe that the Left-hand side and the right-hand side of the equation (3) define a finite measure η, τ respectively such that $\eta(\mathbb{R}) = \tau(\mathbb{R})$ by step (4). Since the class $\mathcal{E} = \{(\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f > r) : r \geq 0\}$ is a π-system. Then by Lemma 2.14 and Lemma 2.15, we have the equation (3) is true for any $E \in \sigma(\mathcal{E}) = \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f^{-1}(\mathcal{B})$. □

We call Theorem 2.13 the *$\mathcal{R}$-measurable Radon–Nikodym Property* because it answers the question (Q2) and hence extends the Radon–Nikodym property. The theorem 2.13 implies that there exist a $\mathcal{R}$-measurable function $\rho_\mathcal{R} = \mathrm{proj}_{\mathcal{M}_\mathcal{R}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ such that

$$\mu(A) = \int_A \rho_\mathcal{R} \, \mathrm{d}\nu, \quad \text{for any } A \in \rho_\mathcal{R}^{-1}(\mathcal{B}).$$

Let $\mathcal{E} = \{(\rho_\mathcal{R} > r), r \geq 0\} \subseteq \mathcal{R}$, then $\rho_\mathcal{R}^{-1}(\mathcal{B})$ is the σ-algebra generated by a sub-collection $\mathcal{E}$ of $\mathcal{R}$.

Remark 2.17. However, in general, we cannot obtain the Radon–Nikodym Property when we replace $\mathcal{R}$ in Theorem 2.13 with a σ-algebra $\mathcal{G}$. Because assume μ, ν be probability measures on Σ, $\rho_\mathcal{G}^{-1}(\mathcal{B}) = (\mathrm{proj}_{\mathcal{M}_\mathcal{G}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu})^{-1}(\mathcal{B}) = \frac{\mathrm{d}\mu|_\mathcal{G}}{\mathrm{d}\nu|_\mathcal{G}}^{-1}(\mathcal{B}) \subsetneq \mathcal{G}$ in general. We will give an example to state it.

Example 2.18. Let $\Omega = \{a, b, c, d\}$ and $\Sigma = \mathcal{P}(\Omega)$, and define probability measures μ, ν on (Ω, Σ) as

$$\mu(a) = \mu(d) = \frac{1}{2}, \mu(c) = \mu(d) = 0$$
$$\nu(a) = \nu(b) = \nu(c) = \mu(d) = \frac{1}{4}$$

then the Radon–Nikodym derivate is $\frac{\mathrm{d}\mu}{\mathrm{d}\nu} = \sum_{v\in\Omega} \frac{\mu(v)}{\nu(v)} \mathbb{1}_{\{v\}} = 2\mathbb{1}_{\{a,d\}}$ and hence $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}^{-1}(\mathcal{B}) = \sigma\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}^{-1}(\mathcal{Q})\right) = \sigma(\{a, d\}) = \{\varnothing, \Omega, \{a, d\}, \{b, c\}\} \subsetneq \Sigma = \mathcal{P}(\Omega)$.

Corollary 2.19. *Let ν be a Borel probability measure, $f \in L_2^+$, for any $A \in \mathcal{R}$,*

$$\int_A \operatorname{proj}_{\mathcal{M}_\mathcal{R}} f \,\mathrm{d}\nu \geq \int_A f \,\mathrm{d}\nu.$$

Proof. By Proposition 2.12 and Theorem 2.13,

$$\begin{aligned}\int_A \operatorname{proj}_{\mathcal{M}_\mathcal{R}} f \,\mathrm{d}\nu &= \int (\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f) \cdot \mathbb{1}_A \,\mathrm{d}\nu \\ &\geq \int \operatorname{proj}_{\mathcal{M}_\mathcal{R}} (f \cdot \mathbb{1}_A) \,\mathrm{d}\nu \\ &= \int_A f \,\mathrm{d}\nu\end{aligned}$$

□

Corollary 2.20. *Let ν be a Borel probability measure, $f \in L_2^+$, and $\mathcal{E} := \{(\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f > r) : r \geq 0\} \subseteq \mathcal{R}$, then for any $E \in \mathcal{E}$*

$$(\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f) \cdot \mathbb{1}_E = \operatorname{proj}_{\mathcal{M}_\mathcal{R}} (f \cdot \mathbb{1}_E).$$

Proof. By Theorem 2.13,

$$\int (\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f) \cdot \mathbb{1}_E \,\mathrm{d}\nu = \int f \cdot \mathbb{1}_E \,\mathrm{d}\nu = \int \operatorname{proj}_{\mathcal{M}_\mathcal{R}} (f \cdot \mathbb{1}_E) \,\mathrm{d}\nu$$

since $(\operatorname{proj}_{\mathcal{M}_\mathcal{R}} f) \cdot \mathbb{1}_E \geq \operatorname{proj}_{\mathcal{M}_\mathcal{R}} (f \cdot \mathbb{1}_E)$, which implies they are identical. □

Theorem 2.21 (Answer to Question (1))**.** *Let μ, ν be Borel probability measures, then*

$$\int \frac{1}{2} \left| \operatorname{proj}_{\mathcal{M}_\mathcal{R}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1 \right| \mathrm{d}\nu = \sup_{A\in\mathcal{R}} (\mu(A) - \nu(A)).$$

Proof. Let $E = \{\operatorname{proj}_{\mathcal{M}_\mathcal{R}} \mathrm{d}\mu/\mathrm{d}\nu > 1\}$, then

$$\begin{aligned}\int \frac{1}{2} \left| \operatorname{proj}_{\mathcal{M}_\mathcal{R}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1 \right| \mathrm{d}\nu &= \frac{1}{2} \int_E \operatorname{proj}_{\mathcal{M}_\mathcal{R}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1 \,\mathrm{d}\nu + \frac{1}{2} \int_{E^c} 1 - \operatorname{proj}_{\mathcal{M}_\mathcal{R}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \,\mathrm{d}\nu \\ &= \frac{1}{2} \left[2 \int_E \operatorname{proj}_{\mathcal{M}_\mathcal{R}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \,\mathrm{d}\nu - 1 \right] - \frac{1}{2} \left[2 \int_E \mathrm{d}\nu - 1 \right] \qquad (8) \\ &= \int_E \operatorname{proj}_{\mathcal{M}_\mathcal{R}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \,\mathrm{d}\nu - \nu(E) = \mu(E) - \nu(E) \qquad (9) \\ &\leq \sup_{A\in\mathcal{R}} (\mu(A) - \nu(A)). \qquad (10)\end{aligned}$$

The equation (8) and (9) is because $\int_A \operatorname{proj}_{\mathcal{M}_\mathcal{R}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \,\mathrm{d}\nu = \int_A \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \nu = \mu(A)$ for any $A \in \operatorname{proj}_{\mathcal{M}_\mathcal{R}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}^{-1}(\mathcal{B})$ by Theorem 2.21. And the equation (10) is because $E = \left\{ \operatorname{proj}_{\mathcal{M}_\mathcal{R}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} > 1 \right\} \in \mathcal{R}$ since $\operatorname{proj}_{\mathcal{M}_\mathcal{R}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \in$

$\mathcal{M}_{\mathcal{R}}$. To show the other direction, by Corollary 2.19, for any $A \in \mathcal{R}$ we have that

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

Thus $\int_A \operatorname{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\,\mathrm{d}\nu \geq \mu(A) - \nu(A)$ for any $A \in \mathcal{R}$, and by equation (9), we have

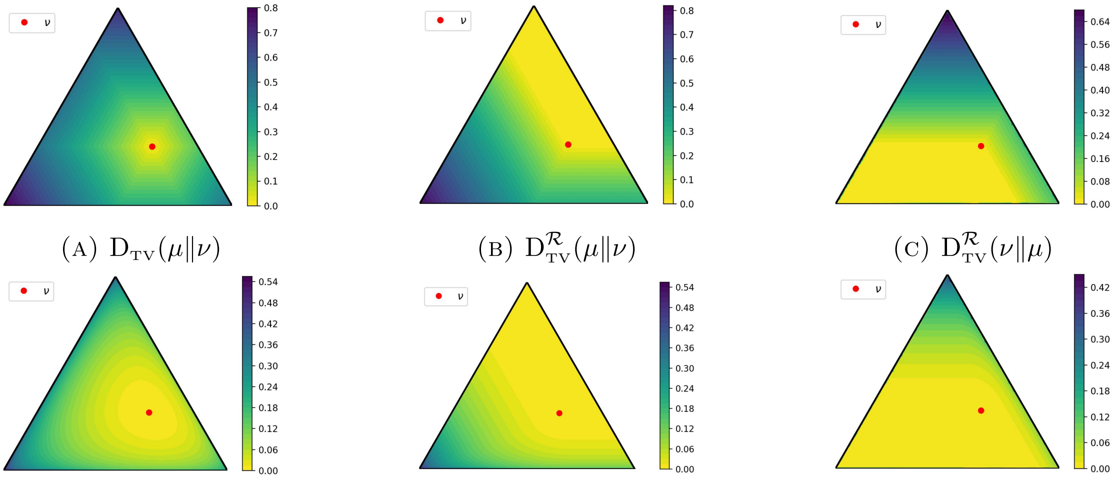
$$\begin{aligned} \int \frac{1}{2}\left|\operatorname{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\right|\mathrm{d}\nu &= \int_E \operatorname{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\,\mathrm{d}\nu \\ &= \sup_{A\in\mathcal{R}} \left[\int_A \operatorname{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\,\mathrm{d}\nu - \nu(A)\right] \\ &\geq \sup_{A\in\mathcal{R}} (\mu(A) - \nu(A)). \end{aligned}$$

□

The Theorem 2.21 shows that the "partial variation" is indeed a special case of f-divergence type of statistical distance $\int f\left(\operatorname{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\nu$ when $f(t) = |t-1|/2$. Hence, we can define the $(\mathcal{R}, f)$-divergence.

Definition 2.22 ($(\mathcal{R}, f)$-divergence)**.** Let μ, ν be probability measures on measurable space (Ω, Σ), $\mathcal{R} = \{(-\infty, a] : a \in \mathbb{R}\}$, then the $(\mathcal{R}, f)$-divergence is defined as

$$\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu) \coloneqq \int f\left(\operatorname{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\nu,$$



(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 $\mathrm{D}_{\mathrm{TV}}, \mathrm{D}_{\mathrm{TV}}^{\mathcal{R}}$, D_{H} and $\mathrm{D}_{\mathrm{H}}^{\mathcal{R}}$ for fixed $\nu = [0.2, 0.5, 0.3]$ as μ ranges over the simplex of distributions on a three-element set.

Proposition 2.23 (Basic Properties)**.** *Let μ, ν be probability measures on measurable space (Ω, Σ), then the $(\mathcal{R}, f)$-divergence has*
(1) Linearity: $\mathrm{D}^{\mathcal{R}}_{\sum_{i=1}^n a_i f_i}(\mu\|\nu) = \sum_{i=1}^n a_i \,\mathrm{D}^{\mathcal{R}}_{f_i}(\mu\|\nu), a_i \geq 0$;
(2) Non-negativity: $\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu) \geq 0$;
(3) Let $g(x) = f(x) + c(x-1)$, then $\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu) = \mathrm{D}_g^{\mathcal{R}}(\mu\|\nu)$;
(4) $\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu) \leq \mathrm{D}_f(\mu\|\nu)$;
(5) if $\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu) = \mathrm{D}_f^{\mathcal{R}}(\nu\|\mu) = 0$, then $\mu = \nu$.

Proof. 1) $\mathrm{D}^{\mathcal{R}}_{\sum_{i=1}^n a_i f_i}\mu = \int \sum_{i=1}^n a_i f_i(\mathrm{proj}_{\mathcal{M}_\mathcal{R}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu})\,\mathrm{d}\nu = \sum_{i=1}^n a_i \int f_i(\mathrm{proj}_{\mathcal{M}_\mathcal{R}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu})\,\mathrm{d}\nu = \sum_{i=1}^n a_i\,\mathrm{D}^{\mathcal{R}}_{f_i}(\mu\|\nu)$.

2): since f is convex, then by Jensen inequality and Theorem 2.13, we have

$$\int f\left(\mathrm{proj}_{\mathcal{M}_\mathcal{R}}\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\nu \geq f\left(\int \mathrm{proj}_{\mathcal{M}_\mathcal{R}}\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\,\mathrm{d}\nu\right) = f\left(\int \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\,\mathrm{d}\nu\right) = f(1) = 0.$$

3):

$$\begin{aligned}\mathrm{D}^{\mathcal{R}}_g(\mu\|\nu) &= \int f\left(\mathrm{proj}_{\mathcal{M}_\mathcal{R}}\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) + c\left(\mathrm{proj}_{\mathcal{M}_\mathcal{R}}\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\right) d\nu \\ &= \int f\left(\mathrm{proj}_{\mathcal{M}_\mathcal{R}}\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\nu + c\cdot\underbrace{\int \mathrm{proj}_{\mathcal{M}_\mathcal{R}}\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1d\nu}_{=0} = \mathrm{D}^{\mathcal{R}}_f(\mu\|\nu).\end{aligned}$$

We postpone the proof for (4) and (5) into section 3. □

Thus the $(\mathcal{R}, TV)$-divergence satisfies $\mathrm{D}^{\mathcal{R}}_{\mathrm{TV}}(\mu\|\nu) = \int \frac{1}{2}\left|\mathrm{proj}_{\mathcal{M}_\mathcal{R}}\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\right|\mathrm{d}\nu = \sup_{A\in\mathcal{R}}(\mu(A) - \nu(A))$. Note that it is not symmetric, and the symmetrized $(\mathcal{R}, TV)$-divergence has $\check{\mathrm{D}}^{\mathcal{R}}_{\mathrm{TV}}(\mu\|\nu) = \sup_{A\in\mathcal{R}}|\mu(A) - \nu(A)|$. Glivenko–Cantelli theorem states the empirical distribution converges to the target distribution almost surely in $(\mathcal{R}, TV)$-divergence, and we will generalize this result to all families of $(\mathcal{R}, f)$-divergence in next section.

3. Inequality between $(\mathcal{R}, f)$-divergence

Lemma 3.1. *Let ν be a Borel probability measure, and $f, \psi \in L_2^+$ be continuous a.s., if ψ is non-decreasing, then*

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

Proof. Let $f = \sum_{i=1}^n \alpha_i \mathbb{1}_{A_i}$, and $\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f = \sum_{i=1}^n \beta_i \mathbb{1}_{A_i}$, then $\psi\circ\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f = \sum_{i=1}^n \psi(\beta_i)\mathbb{1}_{A_i}$. Assume that $\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f$ take $m, (m \leq n)$ values $\tau_1 > \tau_2 > \cdots > \tau_m$, and we can write it as $\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f = \sum_{i=1}^m \tau_i \mathbb{1}_{\{\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f = \tau_i\}}$. For each $k \in [m]$, define $E_k = \bigcup_{i=1}^k \{\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f = \tau_i\} \in \{\{\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f > r\} : r \geq 0\}$. Let $\zeta_m = \psi(\tau_m)$, and for each $j = 1, 2, \ldots, m-1$, let

$$\zeta_{m-j} = \psi(\tau_{m-j}) - \zeta_{m-j+1} - \zeta_{m-j+2} - \cdots - \zeta_m,$$

then we can rewrite $\psi\circ\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f$ as $\sum_{i=1}^m \zeta_i \mathbb{1}_{E_i}$, then

$$\begin{aligned}\int(\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f)\cdot(\psi\circ\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f)\,\mathrm{d}\nu &= \int(\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f)\cdot\sum_{i=1}^m \zeta_i\mathbb{1}_{E_i}\,\mathrm{d}\nu \\ &= \sum_{i=1}^m \zeta_i\int_{E_i}\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f\,\mathrm{d}\nu = \sum_{i=1}^m \zeta_i \int_{E_i} f\,\mathrm{d}\nu \\ &= \int f\cdot\sum_{i=1}^m \zeta_i\mathbb{1}_{E_i}\,\mathrm{d}\nu = \int f\cdot(\psi\circ\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f)\,\mathrm{d}\nu.\end{aligned}$$

For continuous a.s. $f \in L_2^+$, let sequence of order simple function $g_n \uparrow f$ both a.s. and in L_2 norm, then $\mathrm{proj}_{\mathcal{M}_\mathcal{R}} g_n \uparrow \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f$ in L_2 norm. Thus there exists a sub-sequence $\pi(n)$ such that $\mathrm{proj}_{\mathcal{M}_\mathcal{R}} g_{\pi(n)} \uparrow \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f$ a.s., since ψ is non-decreasing, then $\psi\circ\mathrm{proj}_{\mathcal{M}_\mathcal{R}} g_{\pi(n)} \uparrow \psi\circ\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f$ a.s.

Then, by the monotone convergence theorem,

$$\int (\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f)\cdot(\psi \circ \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f)\,\mathrm{d}\nu = \lim_{n\to\infty} \int (\mathrm{proj}_{\mathcal{M}_\mathcal{R}} g_{\pi(n)}) \cdot (\psi \circ \mathrm{proj}_{\mathcal{M}_\mathcal{R}} g_{\pi(n)})\,\mathrm{d}\nu$$

$$= \lim_{n\to\infty} \int g_{\pi(n)} \cdot (\psi \circ \mathrm{proj}_{\mathcal{M}_\mathcal{R}} g_{\pi(n)})\,\mathrm{d}\nu = \int f \cdot (\psi \circ \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f)\,\mathrm{d}\nu.$$

□

Proposition 3.2. *Let ν be a Borel probability measure, $\varphi \in L_2$ is a convex, bounded from below, continuous and differentiable a.s. function, $f \in L_2^+$ be continuous a.s., then*

$$\int \varphi \circ \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f\,\mathrm{d}\nu \le \int \varphi \circ f\,\mathrm{d}\nu.$$

Proof. Since φ is convex, then for any x, $\varphi(\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f(x)) - \varphi(f(x)) \le \varphi'(\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f(x))\cdot(\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f(x) - f(x))$. Thus By Lemma 3.1,

$$\int \varphi \circ \mathrm{proj}_{\mathcal{M}_\mathcal{R}} f - \varphi \circ f\,\mathrm{d}\nu \le \int \varphi'(\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f)\cdot(\mathrm{proj}_{\mathcal{M}_\mathcal{R}} f - f)\,\mathrm{d}\nu = 0.$$

□

For f-divergence, the f is bounded below and differentiable (a.s.) convex function, thus

$$\mathrm{D}_f^\mathcal{R}(\nu\|\mu) = \int f\left(\mathrm{proj}_{\mathcal{M}_\mathcal{R}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\nu \le \int f\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\nu = \mathrm{D}_f(\nu\|\mu).$$

Definition 3.3 (Joint Range, [21, 22])**.** Let $f, g \in \mathcal{F}$, the joint range (w.r.t f, g) is a subset of $[0,\infty]^2$ as

$$\mathcal{I} \coloneqq \{(\mathrm{D}_f(\mu\|\nu), \mathrm{D}_g(\mu\|\nu)) : \mu, \nu \text{ are probability measures on some measurable space}\};$$

and the joint range overall k-ary distribution is defined as

$$\mathcal{I}_k \coloneqq \{(\mathrm{D}_f(\mu\|\nu), \mathrm{D}_g(\mu\|\nu)) : \mu, \nu \text{ are probability measures on } [k]\}.$$

Theorem 3.4 ([21, 22])**.**

$$\mathcal{I} = \mathrm{co}(\mathcal{I}_2) = \mathcal{I}_4,$$

Where co *denotes the convex hull with a natural extension of convex operations to* $[0,\infty]^2$.

Similarly, we can define the joint range w.r.t. $(\mathcal{R}, f)$ and $(\mathcal{R}, g)$-divergence as

$$\mathcal{I}^\mathcal{R} \coloneqq \left\{(D_f^\mathcal{R}(\mu\|\nu), D_g^\mathcal{R}(\mu\|\nu)) : \mu, \nu \text{ are probability measures on some measurable space}\right\};$$

$$\mathcal{I}_k^\mathcal{R} \coloneqq \left\{(D_f^\mathcal{R}(\mu\|\nu), D_g^\mathcal{R}(\mu\|\nu)) : \mu, \nu \text{ are probability measures on } [k]\right\}.$$

Theorem 3.4 states that, for probability measures μ, ν on and measurable space, there exist 4-ary probability measures μ_1, ν_1, such that $\mathrm{D}_f(\mu\|\nu) = \mathrm{D}_f(\mu_1\|\nu_1)$. And there exists utmost 3 pairs of 2-ary probability measures (μ_1, ν_1), (μ_2, ν_2), (μ_3, ν_3) [23], such that $\mathrm{D}_f(\mu\|\nu)$ is the convex combination of $\mathrm{D}_f(\mu_1\|\nu_1)$, $\mathrm{D}_f(\mu_2\|\nu_2)$, and $\mathrm{D}_f(\mu_3\|\nu_3)$. And we can extend such results for $(\mathcal{R}, f)$-divergence.

Theorem 3.5. *(1) $\mathcal{I}_k^\mathcal{R} = \mathcal{I}_k$ for each $k \in \mathbb{N}$ and $\mathcal{I}^\mathcal{R} = \mathcal{I}$, furthermore (2)*

$$\mathcal{I}^\mathcal{R} = \mathrm{co}(\mathcal{I}_2^\mathcal{R}) = \mathcal{I}_4^\mathcal{R}.$$

We separate the proof of theorem 3.5 into the following propositions.

Proposition 3.6. *Let μ, ν probability measures on (Ω, Σ), then there exists a probability measure μ^* on (Ω, Σ) such that for any $f \in \mathcal{F}$,*

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

Proof. We want to show there exists probability measures μ^*, ν^*, such that for any $f \in \mathcal{F}$, $\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu) = \int f\left(\operatorname{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\nu = \int f(\frac{\mathrm{d}\mu^*}{\mathrm{d}\nu}) \mathrm{d}\nu = \mathrm{D}_f(\mu^*\|\tau)$. It is suffices to show there exists probability measures ξ such that $\operatorname{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} = \frac{\mathrm{d}\mu^*}{\mathrm{d}\nu}$. Define a measure μ^* as, for any $B \in \mathcal{B}$,

$$\mu^*(B) =: \int_B \operatorname{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \,\mathrm{d}\nu.$$

By Theorem 2.13, we have $\mu^*(\Omega) = \int \operatorname{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \mathrm{d}\nu = \int \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \mathrm{d}\nu = 1$, thus μ^* is a probability, and $\frac{\mathrm{d}\mu^*}{\mathrm{d}\nu} = \operatorname{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}$. □

Proposition 3.6 implies $\mathcal{I}_k^{\mathcal{R}} \subseteq \mathcal{I}_k$ for each $k \in \mathbb{N}$, and $\mathcal{I}^{\mathcal{R}} \subseteq \mathcal{I}$. The following theorem proves the reverse direction.

Proposition 3.7. *For any $k \in \mathbb{N}$, let μ, ν be be k-ary probability measures, then there exists be k-ary probability measures μ_*, ν_* such that for any $f \in \mathcal{F}$,*

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

Proof. Let $\mu = (\mu_1, \mu_2, \dots, \mu_k)$, $\nu = (\nu_1, \nu_2, \dots, \nu_k)$, and $\alpha_i = \mu_i/\nu_i$, then $\mathrm{D}_f(\mu\|\nu) = \sum_{i=1}^k f(\alpha_i) \cdot \nu_i$, and $\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu) = \sum_{i=1}^k f(\beta_i) \cdot \nu_i$, where $(\beta_1, \dots, \beta_k)$ is the nearest vector to $(\alpha_1, \dots, \alpha_k)$ in ℓ_2 norm. If $\alpha_1 \geq \dots, \alpha_k$, then $\beta_j = \alpha_j$ for each $j \in [k]$, and thus $\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu) = \mathrm{D}_f(\mu\|\nu)$. Let $\alpha_{(j)}$ be the j-th maximum value in $\alpha_1, \dots, \alpha_k$, and $\{\pi(i) : i \in [k]\}$ be a rearrangement of $[k]$ such that $\alpha_{\pi(j)} = \alpha_{(j)}$ for each $j \in [k]$. Let $(\mu_*)_i = \mu_{\pi(i)}$, $(\nu_*)_i = \nu_{\pi(i)}$ for each $i \in [k]$, then μ_*, ν_* define probability measures. Since $(\alpha_*)_i = (\mu_*)_i/(\nu_*)_i = \mu_{\pi(i)}/\nu_{\pi(i)} = \alpha_{(i)}$, then $(\alpha_*)_1 \geq (\alpha_*)_2 \geq \dots \geq (\alpha_*)_k$, then

$$\begin{aligned}\mathrm{D}_f(\mu\|\nu) &= \sum_{i=1}^k f(\alpha_i) \cdot \nu_i = \sum_{i=1}^k f(\alpha_{\pi(i)}) \cdot \nu_{\pi(i)} \\ &= \mathrm{D}_f(\mu_*\|\nu_*) = \mathrm{D}_f^{\mathcal{R}}(\mu_*\|\nu_*).\end{aligned}$$

□

Combine Proposition 3.6, 3.7, we have $\mathcal{I}_k^{\mathcal{R}} = \mathcal{I}_k$ for each $k \in \mathbb{N}$. Thus by Theorem 3.4, we have $\operatorname{co}(\mathcal{I}_2^{\mathcal{R}}) = \operatorname{co}(\mathcal{I}_2) = \mathcal{I}_4 = \mathcal{I}_4^{\mathcal{R}}$, and $\mathcal{I}^{\mathcal{R}} \subseteq \mathcal{I} = \mathcal{I}_4 = \mathcal{I}_4^{\mathcal{R}} \subseteq \mathcal{I}^{\mathcal{R}}$. Thus $\mathcal{I}^{\mathcal{R}} = \mathcal{I}_4^{\mathcal{R}} = \operatorname{co}(\mathcal{I}_2^{\mathcal{R}})$. An important conclusion is that we preserve all the inequality from f-divergence to $(\mathcal{R}, f)$-divergence.

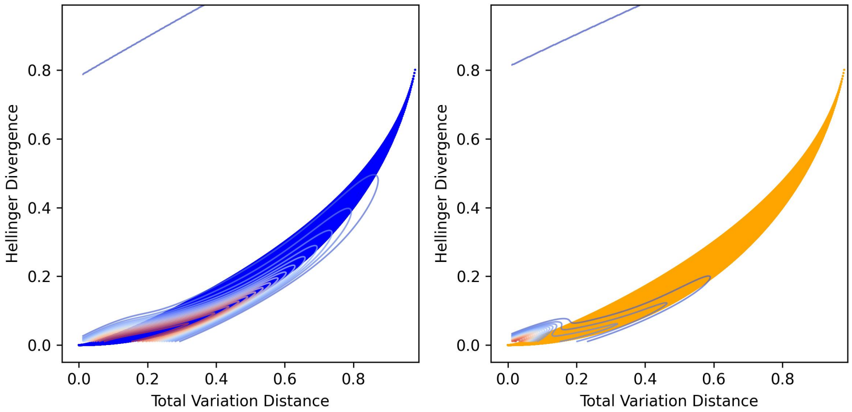


FIGURE 2. The joint range over 2-ary distributions for Total variation and Hellinger divergence, with the contour of distribution of 2-ary distributions.

Corollary 3.8. *Let $f, g \in \mathcal{F}$, and ψ, φ are some functions such that the inequality*

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

holds for all probability measures μ, ν, if and only if

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

holds for all probability measures μ, ν.

Proof. Assume that $\psi(\mathrm{D}_f(\mu\|\nu)) \le \varphi(\mathrm{D}_g(\mu\|\nu))$ for any μ, ν, then $\psi(\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu)) = \psi(\mathrm{D}_f(\mu^*\|\nu)) \le \varphi(\mathrm{D}_g(\mu^*\|\nu)) = \varphi(\mathrm{D}_g^{\mathcal{R}}(\mu\|\nu))$. Reversely, if $\psi(\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu)) \le \varphi(D_g^{\mathcal{R}}(\mu\|\nu))$ for any μ, ν, then $\psi(\mathrm{D}_f(\mu\|\nu)) = \psi(\mathrm{D}_f^{\mathcal{R}}(\mu_*\|\nu_*)) \le \varphi(\mathrm{D}_g^{\mathcal{R}}(\mu_*\|\nu_*)) = \varphi(\mathrm{D}_g(\mu\|\nu))$. □

The following proposition gives a lower bound for any f-divergence in terms of total variation.

Proposition 3.9 ([24])**.** *Let μ, ν be probability measures on (Ω, Σ), for any $f \in \mathcal{F}$, let $\psi(x) = f(1+x) + f(1-x)$, then ψ is strictly increasing and*

$$\psi(\mathrm{D}_{\mathrm{TV}}(\mu\|\nu)) \le \mathrm{D}_f(\mu\|\nu).$$

Then by Corollary 3.8, we can extends it to $(\mathcal{R}, f)$-divergence, for any $f \in \mathcal{F}$, $\psi(x) = f(1+x) + f(1-x)$, then

$$\psi(D_{TV}^{\mathcal{R}}(\mu\|\nu)) = f\left(1 + D_{TV}^{\mathcal{R}}(\mu\|\nu)\right) + f\left(1 - D_{TV}^{\mathcal{R}}(\mu\|\nu)\right) \le D_f^{\mathcal{R}}(\mu\|\nu).$$

and $D_f^{\mathcal{R}}(\mu\|\nu) = 0$ implies $\psi(D_{TV}^{\mathcal{R}}(\mu\|\nu)) = 0$. And since ψ is strictly increasing, then $D_{TV}^{\mathcal{R}}(\mu\|\nu) = 0$. Since $\mathcal{R}$ is a π-system generating Σ, and μ, ν are probability measures, then by Lemma 2.14,

$$\mathrm{D}_{\mathrm{TV}}^{\mathcal{R}}(\mu\|\nu) \vee \mathrm{D}_{\mathrm{TV}}^{\mathcal{R}}(\nu\|\mu) = \sup_{A \in \mathcal{R}} |\mu(A) - \nu(A)| = 0$$

implies $\mu = \nu$. Then we have the following Corollary.

Corollary 3.10. *Let μ, ν be probability measures on (Ω, Σ), for any $f \in \mathcal{F}$, $\mu = \nu$ if and only if $\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu) = \mathrm{D}_f^{\mathcal{R}}(\nu\|\mu) = 0$.*

There are many inequalities related to f-divergence, and by Corollary 3.8, they are also held for $(\mathcal{R}, f)$-divergence.

- TV and (Squared) Hellinger [21]

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

  and [25]

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

- KL and TV

  Pinsker's inequality [16]:

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

  stronger than Pinsker's inequality [26] :

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

  Reverse Pinsker[27]:

$$\mathrm{D}_{\mathrm{KL}}(\mu\|\nu) \le \log\left(1 + \frac{2}{\nu_{\min}}\mathrm{D}_{\mathrm{TV}}(\mu\|\nu)^2\right) \le \frac{2\log e}{\nu_{\min}}\mathrm{D}_{\mathrm{TV}}(\mu\|\nu)^2, \quad \nu_{\min} = \min_x \nu(x)$$

- TV and χ^2 [21]

$$\mathrm{D}_{\chi^2}(\mu\|\nu) \geq f(\mathrm{D}_{\mathrm{TV}}(\mu\|\nu)) \geq 4\,\mathrm{D}^2_{\mathrm{TV}}(\mu\|\nu), \quad f(t) = \begin{cases} 4t^2. & t \leq \frac{1}{2} \\ \frac{t}{1-t} & t \geq \frac{1}{2}. \end{cases}$$

  and

$$\mathrm{D}_{\mathrm{TV}}(\mu\|\nu) \leq \frac{1}{2}\sqrt{\mathrm{D}_{\chi^2}(\mu,\nu)}$$

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

- KL and Hellinger [21]

$$\mathrm{D}_{\mathrm{KL}}(\mu\|\nu) \geq 2\log\frac{2}{2-\mathrm{D}^2_{\mathrm{H}}(\mu\|\nu)} \geq \log e \cdot \mathrm{D}^2_{\mathrm{H}}(\mu\|\nu)$$

  the reverse direction

$$\mathrm{D}_{\mathrm{KL}}(\mu\|\nu) \leq \frac{\log\left(\frac{1}{\nu_{\min}}-1\right)}{1-2\nu_{\min}} \cdot \left(1-\left(1-\mathrm{D}^2_{\mathrm{H}}(\mu\|\nu)\right)^2\right).$$

- KL and χ^2 [21]

$$0 \leq \mathrm{D}_{\mathrm{KL}}(\mu\|\nu) \leq \log\left(1+\mathrm{D}_{\chi^2}(\mu\|\nu)\right) \leq \log e \cdot \mathrm{D}_{\chi^2}(\mu\|\nu)$$

- Le Cam and Hellinger [8]

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

- Le Cam and Jensen-Shannon [28]

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

4. General Glivenko–Cantelli theorem

The example 1.1 implies $\limsup_{n\to\infty}\mathrm{D}_{\mathrm{TV}}(\nu_n\|\nu) > 0$. For any $f \in \mathcal{F}$, since $\psi(x) = f(1+x) + f(1-x)$ is strictly increasing and continuous, and $\psi(0) = 0$ [24], then f-divergence distance between the empirical distribution and the target does not converge to 0 almost surely in general, since

$$0 < \psi\left(\limsup_{n\to\infty}\mathrm{D}_{\mathrm{TV}}(\nu_n\|\nu)\right) = \limsup_{n\to\infty}\psi\left(\mathrm{D}_{\mathrm{TV}}(\nu_n\|\nu)\right) \leq \mathrm{D}_f(\nu_n\|\nu).$$

In this section, however, we prove that $(\mathcal{R}, f)$-divergence between the empirical distribution and the target converges to 0 almost surely, i.e. for any $f \in \mathcal{F}$, $\lim_{n\to\infty}\mathrm{D}^{\mathcal{R}}_f(\nu_n\|\nu) = 0$ a.s., under some mild assumptions. Firstly, since $\dim_{VC}(\mathcal{R}) < \infty$ then by Vapnik–Chervonenkis Theorem

$$\lim_{n\to\infty}\mathrm{D}^{\mathcal{R}}_{\mathrm{TV}}(\nu_n\|\nu) = \lim_{n\to\infty}\sup_{A\in\mathcal{R}}(\nu_n(A)-\nu(A)) = 0, \quad a.s. \tag{11}$$

Assume that $\frac{\mathrm{d}\nu_n}{\mathrm{d}\nu}$, $n \in \mathbb{N}$ exist and are uniformly bounded. Under such assumption, we can see that the key reason why $\sup_{A\in\Sigma}(\nu_n(A)-\nu(A)) \not\to 0$ but $\sup_{A\in\mathcal{R}}(\nu_n(A)-\nu(A)) \to 0$ is that the sequence of Radon–Nikodym derivatives do not converge to 1, but the sequence of the projection of the derivatives converges to 1, i.e., $\frac{\mathrm{d}\nu_n}{\mathrm{d}\nu} \not\to 1$ but $\mathrm{proj}_{\mathcal{M}_{\mathcal{R}}}\frac{\mathrm{d}\nu_n}{\mathrm{d}\nu} \to 1$. To see this, assume that $\lim_{n\to\infty}\frac{\mathrm{d}\nu_n}{\mathrm{d}\nu} = 1$, then

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

Then $\lim_{n\to\infty} \mathrm{D}_{\mathrm{TV}}(\nu_n\|\nu) = 0$, which leads to a contradiction, and hence the statement $\lim_{n\to\infty} \frac{\mathrm{d}\nu_n}{\mathrm{d}\nu} = 1$ is false. Since $(\frac{\mathrm{d}\nu_n}{\mathrm{d}\nu})_n$ are uniformly bounded, then there exists M such that $\frac{\mathrm{d}\nu_n}{\mathrm{d}\nu} \leq M$ for each n, then by Proposition 2.9, $\mathrm{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\nu_n}{\mathrm{d}\nu} \leq \mathrm{proj}_{\mathcal{M}_{\mathcal{R}}} M = M$ for each n, i.e. the sequence $(\mathrm{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\nu_n}{\mathrm{d}\nu})_n$ are uniformly bounded as well. By equation (11) and Bounded Convergence Theorem, we have that

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

since the integrand is non-negative, then $\left|\lim_{n\to\infty} \mathrm{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\nu_n}{\mathrm{d}\nu} - 1\right| = 0$ and hence $\lim_{n\to\infty} \mathrm{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\nu_n}{\mathrm{d}\nu} = 1$. By this result, we can generalize the Glivenko–Cantelli theorem to others $(\mathcal{R}, f)$-divergence.

Proposition 4.1. *Let ν is a probability measure on measurable space (Ω, Σ), ν_n is the empirical measures, $\mathcal{R}$ is the class of rays, then for any $f \in \mathcal{F}$,*

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

Proof. Apply the result $\lim_{n\to\infty} \mathrm{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\nu_n}{\mathrm{d}\nu} = 1$, and the Reverse Fatou's lemma, we have

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

thus $\lim_{n\to\infty} \int f\left(\mathrm{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\nu_n}{\mathrm{d}\nu}\right) \mathrm{d}\nu = \lim_{n\to\infty} D_f^{\mathcal{R}}(\nu_n\|\nu) = 0$. □

Furthermore, assume that the sequence $\frac{\mathrm{d}\nu}{\mathrm{d}\nu_n}$, $n \in \mathbb{N}$ exists and uniformly bounded.

Proposition 4.2. *Let ν is a probability measure on measurable space (Ω, Σ), ν_n is the empirical measures, $\mathcal{R}$ is the class of rays, then*

$$\lim_{n\to\infty} \mathrm{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\nu}{\mathrm{d}\nu_n} = 1$$

Proof. Let $x_n := \frac{1}{2}\left|\mathrm{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\nu}{\mathrm{d}\nu_n} - 1\right| \in L^2(\nu_n)$, then by Glivenko–Cantelli theorem, we have

$$\lim_{n\to\infty} \int x_n \,\mathrm{d}\nu_n = \lim_{n\to\infty} \nu_n x_n = 0$$

assume that $\lim_{n\to\infty} x_n \neq 0$, then there is a $\epsilon > 0$ and a sub-sequence (x_{n_k}) s.t. $\|x_{n_k}\| = \int x_{n_k}\,\mathrm{d}\nu > \epsilon$. By assumption, $(\frac{\mathrm{d}\nu}{\mathrm{d}\nu_n})_n$ are uniformly bounded, then there exists M s.t. $\frac{\mathrm{d}\nu}{\mathrm{d}\nu_n} < M$ for any n. Then

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

which leads to a contradiction, thus $\lim_{n\to\infty} x_n = 0$, and hence $\lim_{n\to\infty} \mathrm{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\nu}{\mathrm{d}\nu_n} = 1$. □

Proposition 4.3. *Let ν is a probability measure on measurable space (Ω, Σ), ν_n is the empirical measures, $\mathcal{R}$ is the class of rays, then*

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

Proof. Let $y_n \coloneqq \left| f\left(\operatorname{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\nu}{\mathrm{d}\nu_n}\right)\right|$, then $\lim_{n\to\infty} y_n = \left| f\left(\lim_{n\to\infty} \operatorname{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\nu}{\mathrm{d}\nu_n}\right)\right| = 0$. Then

$$\left| \int f\left(\operatorname{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\nu}{\mathrm{d}\nu_n}\right) \mathrm{d}\nu_n \right| \leq \int \left| f\left(\operatorname{proj}_{\mathcal{M}_{\mathcal{R}}} \frac{\mathrm{d}\nu}{\mathrm{d}\nu_n}\right)\right| \mathrm{d}\nu_n = \nu_n y_n.$$

Assume that $\nu_n y_n \not\to 0$, then there is a $\epsilon > 0$ and a sub-sequence (n_k) s.t.

$$0 < \epsilon < \|\nu_{n_k} y_{n_k}\| \leq \|\nu_{n_k}\| \cdot \|y_{n_k}\|.$$

Since for measurable function φ, we have $\lim_{n\to\infty} \nu_n \varphi = \nu\varphi$, then $\sup_n \|\nu_n y\| < \infty$ for any $y \in L^1(\nu)$. By Uniformly Bounded Principle, we have that $\sup_n \|\nu_n\| < \infty$, and hence

$$0 < \epsilon \leq \|\nu_{n_k}\| \cdot \|y_{n_k}\| \leq \sup_n \|\nu_n\| \cdot \|y_{n_k}\| \to 0.$$

which leads to a contradiction, thus $\lim_{n\to\infty} \nu_n y_n = 0$ and hence $\lim_{n\to\infty} \mathrm{D}_f^{\mathcal{R}}(\nu\|\nu_n) = 0$. □

We have shown that under mild assumptions, both $\mathrm{D}_f^{\mathcal{R}}(\nu_n\|\nu)$ and $\mathrm{D}_f^{\mathcal{R}}(\nu\|\nu_n)$ converge to 0 almost surely. Thus by corollary 3.10, $\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu) + \mathrm{D}_f^{\mathcal{R}}(\nu\|\mu)$ or $\max\{\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu), \mathrm{D}_f^{\mathcal{R}}(\nu\|\mu)\}$ defines a distance that converges to 0 almost surely between the target distribution and empirical distributions.

5. Discussion

5.1. **Choquet Integral.**

In this subsection, we state our initial idea to solve the question (Q1) using the Choquet integral and show why it does not work. Lebesgue integral defines integral w.r.t measure, and Choquet integral extends it; it defines integral for a more general set function, i.e., capacity. Let Ω be the ground set, $\mathcal{S} \subseteq 2^{\Omega}$ and $\varnothing, \Omega \in \mathcal{S}$. $\nu : \mathcal{S} \to \mathbb{R}$ is a *capacity* on $\mathcal{S}$ if (1) $\nu(\varnothing) = 0$ and (2) for $A, B \in \mathcal{S}$, $A \subseteq B$, then $\nu(A) \leq \nu(B)$. function $f : \Omega \to \mathbb{R}$ is $\mathcal{S}$-*measurable* if $(f > t) \in \mathcal{S}$ for any $t \in \mathbb{R}$. And let $B^+(\mathcal{S})$ be the set of all bounded $\mathcal{S}$-measurable non-negative functions. Let $f \in B^+(\mathcal{S})$, The decumulative (distribution) function of f w.r.t. ν is $G_{\nu,f}(t) = \nu(f > t)$, $t \in \mathbb{R}$. The *Choquet integral* of f w.r.t. ν is defined by

$$(C)\int f \,\mathrm{d}\nu = \int_0^{\infty} G_{\nu,f}(t)\,\mathrm{d}t.$$

where the right-hand-side integral is the Riemann integral. If capacity ν is a measure, then the Choquet integral coincides with the Lebesgue integral [29]. An nature idea to define $(\mathcal{R}, f)$-divergence is letting $\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu) \coloneqq \mathrm{D}_f(\mu_{\mathcal{R}}\|\nu_{\mathcal{R}})$, where $\mu_{\mathcal{R}}, \nu_{\mathcal{R}}$ are the restrictions of μ, ν onto $\mathcal{R}$. The idea came from the fact: it recovers "partial variation" as we want when one replace $\mathcal{R}$ with a sub σ-algebra $\mathcal{G} \subseteq \Sigma$, because if μ, ν are probability measures on Σ, then $\mu_{\mathcal{G}}, \nu_{\mathcal{G}}$ are probability measures on $\mathcal{G}$, and hence

$$\mathrm{D}_f^{\mathcal{G}}(\mu\|\nu) \coloneqq \mathrm{D}_f(\mu_{\mathcal{G}}\|\nu_{\mathcal{G}}) = \sup_{A\in\mathcal{G}} \left(\mu(A) - \nu(A)\right).$$

For the case of class of rays $\mathcal{R}$, it is easy to check $\mu_{\mathcal{R}}, \nu_{\mathcal{R}}$ are capacities, and hence we tried to define $\mathrm{D}_f(\mu_{\mathcal{R}}\|\nu_{\mathcal{R}})$ using Choquet integral, base on the results in [30, 31, 32, 33, 34]. However, the inevitable obstacle is there is no guarantee such that the derivate $\frac{\mathrm{d}\mu_{\mathcal{R}}}{\mathrm{d}\nu_{\mathcal{R}}}$ exists and $f(\frac{\mathrm{d}\mu_{\mathcal{R}}}{\mathrm{d}\nu_{\mathcal{R}}}), f \in \mathcal{F}$ is $\mathcal{R}$-measurable (i.e., non-increasing) for the class of rays $\mathcal{R}$, instead of σ-algebra or even algebra. Thus the stuff $\mathrm{D}_f(\mu_{\mathcal{R}}\|\nu_{\mathcal{R}})$ can not be well-defined. Our another attempt is to define $\hat{\mu}_{\mathcal{R}}(A) = \mu(A)$ if $A \in \mathcal{R}$, and $\hat{\mu}_{\mathcal{R}}(A) = 0$ if $A \in \Sigma\backslash\mathcal{R}$. Then $\hat{\mu}_{\mathcal{R}}$ is a set function with domain Σ. Let $\mathrm{D}_f^{\mathcal{R}}(\mu\|\nu) \coloneqq$

$\mathrm{D}_f(\hat{\mu}_{\mathcal{R}}\|\hat{\nu}_{\mathcal{R}})$. This idea came from another heuristic fact:

$$\begin{aligned}\mathrm{D}^{\mathcal{R}}_{\mathrm{TV}}(\mu\|\nu) = \mathrm{D}_{\mathrm{TV}}(\hat{\mu}_{\mathcal{R}}\|\hat{\nu}_{\mathcal{R}}) &= \sup_{A\in\Sigma}|\hat{\mu}_{\mathcal{R}}(A)-\hat{\nu}_{\mathcal{R}}(A)| \\ &= \max\left\{\sup_{A\in\mathcal{R}}|\hat{\mu}_{\mathcal{R}}(A)-\hat{\nu}_{\mathcal{R}}(A)|,\ \sup_{B\in\Sigma\setminus\mathcal{R}}\underbrace{|\hat{\mu}_{\mathcal{R}}(B)-\hat{\nu}_{\mathcal{R}}(B)|}_{\equiv 0}\right\} \\ &= \sup_{A\in\mathcal{R}}|\mu(A)-\nu(A)|,\end{aligned}$$

which is the "partial variation" we want. However, another obstacle is that set functions $\hat{\mu}_{\mathcal{R}}, \hat{\nu}_{\mathcal{R}}$ do not satisfy the monotonicity, i.e., they are not capacities. Thus, $\mathrm{D}_f(\hat{\mu}_{\mathcal{R}}\|\hat{\nu}_{\mathcal{R}})$ can not be well-defined either by Choquet integral.

5.2. **Characteristic for Variational Class.** In the following subsections, we propose two open questions left by this paper. Let (Ω,Σ) be a measurable space. We would call a class $\mathcal{C}\subseteq\Sigma$ is a *variational class*, if for any probability measures μ,ν on (Ω,Σ), one has $\mathcal{M}_{\mathcal{C}} := \{f\in L_2(\nu) : (f > r)\in\mathcal{C}, r\in\mathbb{R}\}$ is a Chebyshev set in $L_2(\nu)$ and

$$\int \frac{1}{2}\left|\mathrm{proj}_{\mathcal{M}_{\mathcal{C}}}\frac{\mathrm{d}\mu}{\mathrm{d}\nu}-1\right|\mathrm{d}\nu = \sup_{A\in\mathcal{C}}(\mu(A)-\nu(A)) \tag{12}$$

If $\mathcal{C}$ is a Chebyshev set, then we can define the $(\mathcal{C},f)$-divergence as $\mathrm{D}^{\mathcal{C}}_f(\mu\|\nu) = \int f(\mathrm{proj}_{\mathcal{M}_{\mathcal{C}}}\frac{\mathrm{d}\mu}{\mathrm{d}\nu})\,\mathrm{d}\nu$. We require the equation (12) is because we want the $(\mathcal{C},f)$-divergence has the property $\mathrm{D}^{\mathcal{C}}_{\mathrm{TV}}(\mu\|\nu) = \sup_{A\in\mathcal{C}}(\mu(A)-\nu(A))$, in words, we want the "$\mathcal{C}$-partial variation distance" is a special case of $(\mathcal{C},f)$-divergence when $f(t)=|1-t|/2$, just like the total variation. We have shown that the class $\mathcal{R}$ is a variational class. A nature question is to develop a characteristic for the variational class and to extend the $(\mathcal{R},f)$-divergence to for all others variational class $\mathcal{C}$. If $\mathcal{C}$ satisfies (1) $\{\varnothing,\Omega\}\subseteq\mathcal{C}$, and (2) it is closed under countable union and intersection, then $\mathcal{M}_{\mathcal{C}}$ is a closed convex cone in $L_2(\nu)$, and hence it is a Chebyshev set. It is a convex cone since for $f,g\in\mathcal{M}_{\mathcal{C}}$, and $\lambda,\eta\geq 0$, we have $\lambda f, \eta g\in\mathcal{M}_{\mathcal{C}}$, and

$$(f+g>c) = \bigcup_{q\in\mathbb{Q}}(f>c-q)\cap(g>q)\in\mathcal{C}.$$

And it is closed because, if $f_n\in\mathcal{M}$ such that $\|f_n-f\|_{L_2}\to 0$, then $f_n\to f$ in ν-measure, and hence there exists a sub-sequence $(f_{\pi(n)})$ such that $f_{\pi(n)}\to f$ point-wise (ν-a.s.), then we have

$$(f>c) = \bigcup_{k\geq 1}\bigcap_{m\geq 1}\bigcup_{n\geq m}\left(f_{\pi(n)}>c+\frac{1}{k}\right)\in\mathcal{C}.$$

The σ-algebra Σ is a variational class, and when $\mathcal{C}=\Sigma$, then $\mathcal{M}_\Sigma = L_2$ and $\mathrm{proj}_{\mathcal{M}_\Sigma}f = f$ for $f\in L_2$, hence we can recover the vanilla f-divergence:

$$\mathrm{D}^{\Sigma}_{\mathrm{TV}}(\mu\|\nu) = \int\frac{1}{2}\left|\mathrm{proj}_{\mathcal{M}_\Sigma}\frac{\mathrm{d}\mu}{\mathrm{d}\nu}-1\right|\mathrm{d}\nu = \int\frac{1}{2}\left|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}-1\right|\mathrm{d}\nu = \sup_{A\in\Sigma}|\mu(A)-\nu(A)|.$$

Any sub σ-algebra $\mathcal{G}\subseteq\Sigma$ is also a variational class; we will give examples to show this.

Example 5.1. Let μ,ν be probability measures on (Ω,Σ), and $\mathcal{G}\subseteq\Sigma$ be a sub σ-algebra, it is clear that $\mathcal{G}$ is a variational class. Let $\mu|_{\mathcal{G}}$ is the restriction of μ on $\mathcal{G}$, then $\mu|_{\mathcal{G}}$ is still a probability measure [20], and furthermore, if $g\geq 0$ is $\mathcal{G}$-measurable, i.e. $(g>c)\in\mathcal{G}$ for any $c\in\mathbb{R}$, then for any $A\in\mathcal{G}$

$$\int_A g\,\mathrm{d}\mu|_{\mathcal{G}} = \int_A g\,\mathrm{d}\mu$$

beyond this, note that if $\mu \ll \nu$, then $\mu|_{\mathcal{G}} \ll \nu|_{\mathcal{G}}$, and hence the Radon–Nikodym derivate $\frac{\mathrm{d}\mu|_{\mathcal{G}}}{\mathrm{d}\nu|_{\mathcal{G}}}$ exists. For any $A \in \mathcal{G}$

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

that is, $\frac{\mathrm{d}\mu|_{\mathcal{G}}}{\mathrm{d}\nu|_{\mathcal{G}}}$ is indeed the conditional expectation: $\frac{\mathrm{d}\mu|_{\mathcal{G}}}{\mathrm{d}\nu|_{\mathcal{G}}} = \mathbb{E}_\mu\left[\frac{\mathrm{d}\mu}{\mathrm{d}\nu} \mid \mathcal{G}\right]$, and hence $\frac{\mathrm{d}\mu|_{\mathcal{G}}}{\mathrm{d}\nu|_{\mathcal{G}}} = \operatorname{proj}_{\mathcal{M}_{\mathcal{G}}} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ [35]. Then we have

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

5.3. **Characteristic for Glivenko–Cantelli Class.** The second open question is that for what kind of variational class $\mathcal{C}$, we have $\lim_{n\to\infty} \operatorname{proj}_{\mathcal{M}_{\mathcal{C}}} \frac{\mathrm{d}\nu_n}{\mathrm{d}\nu} = 1$ almost surely, we would call such a class a *Glivenko–Cantelli Class.* The Vapnik–Chervonenkis theorem [36] implies, if variational class $\mathcal{C}$ has finite VC dimension, then $\lim_{n\to\infty} \operatorname{proj}_{\mathcal{F}_{\mathcal{C}}} \frac{\mathrm{d}\nu_n}{\mathrm{d}\nu} = 1$, and hence it is a Glivenko–Cantelli class. We will give an example of a Glivenko–Cantelli class with an infinite VC dimension.

Example 5.2. If $\mathcal{E} = (E_n)_{n\in\mathbb{N}}$ is a countable partition of Ω, then any set in $\mathcal{G} = \sigma(\mathcal{E})$ is countable union of sets in $\mathcal{E}$ [37]. And hence any $\mathcal{G}$-measurable function g must be a simple function like $g = \sum_{n=1}^{\infty} \alpha_n \mathbb{1}_{E_n}$. Let $\mu|_{\mathcal{G}}, \nu|_{\mathcal{G}}$ be the restrictions of probability measures μ, ν, and write $\frac{\mathrm{d}\mu|_{\mathcal{G}}}{\mathrm{d}\nu|_{\mathcal{G}}} = \sum_{n=1}^{\infty} \alpha_n \mathbb{1}_{E_n}$, then for any m, we have

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

thus we have the closed form of the Radon–Nikodym derivate $\frac{\mathrm{d}\mu|_{\mathcal{G}}}{\mathrm{d}\nu|_{\mathcal{G}}} = \sum_{n=1}^{\infty} \frac{\mu|_{\mathcal{G}}(E_n)}{\nu|_{\mathcal{G}}(E_n)} \mathbb{1}_{E_n}$. Thus

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

then assume the sequence $\left(\frac{\nu_n|_{\mathcal{G}}(E_k)}{\nu|_{\mathcal{G}}(E_k)}\right)_k$ are uniformly bounded, we have

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

References

[1] R. M. Shortt. Empirical measures. *The American Mathematical Monthly*, 91(6):358–360, 1984.

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

[3] Vladimir N. Vapnik. *Statistical learning theory.* Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, Inc., New York, 1998. A Wiley-Interscience Publication.

[4] Olivier Bousquet and André Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2(3):499–526, 2002.

[5] Galen R. Shorack and Jon 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. Reprint of the 1986 original [MR0838963].

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

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

[8] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer-Verlag, New York, 1986.

[9] Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Trans. Inform. Theory*, 37(1):145–151, 1991.

[10] Harold Jeffreys. *Theory of probability*. Oxford Classic Texts in the Physical Sciences. The Clarendon Press, Oxford University Press, New York, 1998. Reprint of the 1983 edition.

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

[12] Karl 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*, pages 11–28. Springer New York, New York, NY, 1992.

[13] E. Hellinger. Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *J. Reine Angew. Math.*, 136:210–271, 1909.

[14] Stanislav Molchanov. Book review: Geometric modeling in probability and statistics. *Bulletin of the American Mathematical Society*, 55:1, 05 2017.

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

[16] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[17] Charles L. Lawson. The approximation of functions, vol. ii (john r. rice). *SIAM Review*, 14(1):187–188, 1972.

[18] Donald L. Cohn. *Measure theory*. Birkhäuser Advanced Texts: Basler Lehrbücher. [Birkhäuser Advanced Texts: Basel Textbooks]. Birkhäuser/Springer, New York, second edition, 2013.

[19] John K. Hunter and Bruno Nachtergaele. *Applied analysis*. World Scientific Publishing Co., Inc., River Edge, NJ, 2001.

[20] René L. Schilling. *Measures, integrals and martingales*. Cambridge University Press, Cambridge, second edition, 2017.

[21] Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2024.

[22] Peter Harremoës and Igor Vajda. On pairs of f-divergences and their joint range. *IEEE Trans. Inform. Theory*, 57(6):3230–3235, 2011.

[23] László Lovász and Michael D. Plummer. *Matching theory*. AMS Chelsea Publishing, Providence, RI, 2009. Corrected reprint of the 1986 original [MR0859549].

[24] Jochen Bröcker. A lower bound on arbitrary f–divergences in terms of the total variation. 2009.

[25] Gustavo L. Gilardoni. On Pinsker's and Vajda's type inequalities for Csiszár's f-divergences. *IEEE Trans. Inform. Theory*, 56(11):5377–5386, 2010.

[26] Igor Vajda. Note on discrimination information and variation. *IEEE Trans. Inform. Theory*, IT-16:771–773, 1970.

[27] Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Trans. Inform. Theory*, 62(11):5973–6006, 2016.

[28] Flemming Topsø e. Some inequalities for information divergence and related measures of discrimination. *IEEE Trans. Inform. Theory*, 46(4):1602–1609, 2000.

[29] Dieter 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.

[30] Michio Sugeno. A note on derivatives of functions with respect to fuzzy measures. *Fuzzy Sets and Systems*, 222:1–17, 2013.

[31] Vicen¸c Torra, Yasuo Narukawa, and Michio Sugeno. On the f-divergence for non-additive measures. *Fuzzy Sets and Systems*, 292:364–379, 2016.

[32] John Harding, Massimo Marinacci, Nhu T. Nguyen, and Tonghui Wang. Local Radon-Nikodym derivatives of set functions. *Internat. J. Uncertain. Fuzziness Knowledge-Based Systems*, 5(3):379–394, 1997.

[33] Siegfried Graf. A Radon-Nikodým theorem for capacities. *J. Reine Angew. Math.*, 320:192–214, 1980.

[34] Yasuo Narukawa. Distances defined by choquet integral. In *2007 IEEE International Fuzzy Systems Conference*, pages 1–6, 2007.

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

[36] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, Cham, 2015. Reprint of Theor. Probability Appl. **16** (1971), 264–280.

[37] Erhan ¸Cı nlar. *Probability and stochastics*, volume 261 of *Graduate Texts in Mathematics*. Springer, New York, 2011.

Computational and Applied Mathematics Initiative, University of Chicago, Chicago, IL 60637-1514.
Email address: `haomingwang@uchicago.edu, lekheng@uchicago.edu`