Optimal transport for ϵ-contaminated credal sets*

Michele Caprio[1,2]

[1]Department of Computer Science, University of Manchester, United Kingdom
[2]Manchester Centre for AI Fundamentals (MCAIF), United Kingdom

Abstract

We present generalized versions of Monge's and Kantorovich's optimal transport problems with the probabilities being transported replaced by lower probabilities. We show that, when the lower probabilities are the lower envelopes of ϵ-contaminated sets, then our version of Monge's, and a restricted version of our Kantorovich's problems, coincide with their respective classical versions. We also give sufficient conditions for the existence of our version of Kantorovich's optimal plan, and for the two problems to be equivalent. As a byproduct, we show that for ϵ-contaminations the lower probability versions of Monge's and Kantorovich's optimal transport problems need not coincide. The applications of our results to Machine Learning and Artificial Intelligence are also discussed.

Keywords. optimal transport, credal sets, ϵ-contamination, lower probabilities

1. Introduction

The concept of stochasticity is pervasive in modern-day artificial intelligence (AI) and machine learning (ML), allowing to capture the lack of determinism that underpins virtually all interesting applications, ranging from the medical domain [15, 64] to trajectory prediction of ballistic missiles [37].

Two objects that are often of interest are a random quantity ξ_1, distributed according to a probability measure P_1, $\xi_1 \sim P_1$, and a transformation of ξ_1 via a function T, that we write $\xi_2 = T(\xi_1)$, which in turn is distributed according to a probability measure P_2, $T(\xi_1) = \xi_2 \sim P_2$. Simple – but important – examples of these instances are height and body mass index (BMI) of a population, and a mother's income and her children's (future) income.

From a classical measure-theoretic argument [60], we can obtain P_2 as the pushforward measure of P_1 via T, $P_2(\xi_2 \in B) = P_1(T(\xi_1) \in B) = P_1(\xi_1 \in T^{-1}(B))$, where B is an arbitrary subset of the space that ξ_2 take values on. Then, we can write $P_2 \equiv T_{\#}P_1 := P_1 \circ T^{-1}$, so that P_2 is indeed the pushforward measure $T_{\#}P_1$ of P_1 via T.

An interesting question we may ask ourselves at this point is whether we can turn the problem around. Given P_1 and P_2, there are many functions T that push P_1 to P_2. Is there an "optimal" one, that is, one that makes the transformation from P_1 to P_2 as efficient (i.e. less expensive) as possible? This question, which lays at the heart of the field of Optimal Transport (OT), is similar to one that Napoleonic engineers were asked by Napoleon himself. They were tasked to find the cheapest way of transporting iron ore from the mines to the factories [69].

To find such an optimal T, in the late 1700s Gaspard Monge suggested the following optimization problem,

$$\arg\inf \left\{ \int c(\xi_1, T(\xi_1)) P_1(\mathrm{d}\xi_1) \,:\, T_{\#}P_1 = P_2 \right\}.$$

It seeks to find the function T that makes transporting the probability mass encoded in P_1 to that encoded in its pushforward via T, $P_2 = T_{\#}P_1$, as "cheap" as possible. The latter is gauged by considering a cost function c that gives us the cost of moving one unit of probability mass from ξ_1 to $\xi_2 = T(\xi_1)$. In other words, c gives us a measure of the efficiency of "moving probability bits" from P_1 to $P_2 = T_{\#}P_1$.

Alas, an optimal solution T to this problem may not exist [66, Section 1.2]. Fortunately, though, Leonid Kantorovich came up with an equivalent formulation of the problem that, under mild conditions, is guaranteed to be well-posed. His expression is the following

$$\arg\inf \left\{ \int c(\xi_1, \xi_2) \mathrm{d}\alpha(\xi_1, \xi_2) \,:\, \alpha \in \Gamma(P_1, P_2) \right\},$$

where $\Gamma(P_1, P_2)$ is the set of all joint probability measures whose marginals are P_1 and P_2. Instead of looking for the most efficient transportation map T from ξ_1 to ξ_2, it seeks the "cheapest" *transportation plan* α between the distributions P_1 and P_2. The relationship between the optimal transportation plan α and the theory of copulas [52][1] was studied e.g. in Chi et al. [21] and Liu, Wang, Wang, and Zhuang [44].

*To the memory of Sayan Mukherjee

[1]Recall that a *copula* is a multivariate cumulative distribution function, for which the marginal probability distribution of each variable is Uniform on the interval $[0, 1]$.

Another notable Leonid, Wasserstein, used the tools developed by Kantorovich and other optimal transport theory scholars to study a class of probability metrics that bears his name: for $p \geq 1$,

$$W_p(P_1, P_2) := \left[\inf_{\alpha \in \Gamma(P_1, P_2)} \int d(\xi_1, \xi_2)^p \mathrm{d}\alpha(\xi_1, \xi_2) \right]^{\frac{1}{p}}$$

is the p-Wasserstein metric (on the probability space P_1 and P_2 are defined on), where cost function $c(\xi_1, \xi_2)$ is the p-th power of some metric d on the state space where ξ_1 and ξ_2 are defined on, e.g. some norm $\|\xi_1 - \xi_2\|$. In a sense, the Wasserstein metric allows to endow the probability space with a metric derived from the distance defined on the underlying state space. The concept of Wasserstein distance is ubiquitous in AI and ML, spanning fields such as data-driven control [43] and uncertainty quantification [61].

Contributions. In this paper, we ask ourselves:

Question 1: What do Monge's and Kantorovich's problems look like, when instead of transporting probability measures, we transport *lower probabilities*?

Lower probabilities are the imprecise counterpart of classical probabilities that allow to describe the ambiguity faced by the scholar around the true data generating processes [7, 68, 70]. We give the first (to the best of our knowledge) definitions of Monge's and Kantorovich's problems for lower probabilities, and then we focus our attention on sets of probabilities $\mathcal{M}(\underline{P})$ (called *cores* of a coherent lower probability, a special type of *credal sets*) that are completely characterized by their lower envelope $\underline{P}$ (that is a lower probability).[2] This means that the whole set $\mathcal{M}(\underline{P})$ can be reconstructed by simply looking at lower probability $\underline{P} = \inf_{P \in \mathcal{M}(\underline{P})} P$. A pictorial representation of our endeavor is given in Figure 1.

It is necessary to mention that we are not the first to extend the study of OT beyond classical probability theory. Lorenzini, Petturiti, and Vantaggi [47] and Nguyen [53] do so for belief functions and random sets, and Gal and Niculescu [29], Rachev and Olkin [56], and Torra [67] do so for capacities and non-additive measures.

Question 2: Is there a class of credal sets completely characterized by their lower probabilities (LPs), for which the LP versions of Monge's and Kantorovich's problems coincide with their classical counterparts?

We show that, for the class of ϵ-contaminated credal sets (which we introduce in (3)), the answer to Question 2 is positive.[3] This is an important result, as it promises to be fraught with fruitful consequences for many possible applications. We also give sufficient conditions (i) for the existence of the lower probability version of Kantorovich's optimal plan, and (ii) for the two problems formulations to coincide. A byproduct of the latter is that, in general, the lower probability versions of Monge's and Kantorovich's problems need not coincide.

Motivation and related work. Besides being interested in these results for their own mathematical beauty, our motivations to study them stem from the field of Imprecise Probabilistic Machine Learning (IPML) [12, 13, 16, 17, 26, 27, 36, 48, 73]. Credal Machine Learning (CML), a subfield of IPML, devotes itself to developing ML theory and methods working with credal sets. Our findings in this paper can be immediately applied to CML in at least three different contexts.

First, they can be used to define new uncertainty measures enjoying the axiomatic desiderata in Abellán and Klir [1], Jiroušek and Shenoy [38], and Sale, Caprio, and Hüllermeier [62] based on Hausdorff-type distances between sets of probabilities, which would extend the results in Sale, Bengs, Caprio, and Hüllermeier [61] to credal sets. To be more specific, the Hausdorff distance based on the Wasserstein metric between two credal sets could be expressed as the Wasserstein distance between their lower probabilities. The latter is intimately related to the Kantorovich's OT problem that we define in this paper.

A second immediate application is robust Hypotheses Testing (HT) [2, 6, 19, 32, 35, 45, 51], where a test statistic based on the optimal transport cost between lower probabilities characterizing credal sets could be used to test whether the true data generating process – that produces the data accruing to the phenomenon of interest – belongs to either of the credal sets. In HT notation, $H_0 : P^{\text{true}}$ belongs to $\mathcal{M}(\underline{P}_1)$, versus $H_1 : P^{\text{true}}$ belongs to $\mathcal{M}(\underline{P}_2)$. Let us give a hand-wavy example. Acharya, Daskalakis, and Kamath [2] construct a procedure to test $H_0 : P^{\text{true}} \in \mathcal{M}(\underline{P}_1)$ versus $H_1 : P^{\text{true}} \notin \mathcal{M}(\underline{P}_1)$. To carry out the test, they first split the available sample size into two parts, and use the first part to "project" P^{true} onto $\mathcal{M}(\underline{P}_1)$; denote by $\hat{P}^{\text{true}}$ such a projection. Then, with the second part, they test $H_0' : P^{\text{true}} \in B_\eta(\hat{P}^{\text{true}})$, where $B_\eta(\hat{P}^{\text{true}})$ is a ball (with respect to some metric, e.g. the Wasserstein one) of radius η centered at $\hat{P}^{\text{true}}$, versus $H_1' : P^{\text{true}} \notin B_\eta(\hat{P}^{\text{true}})$, and they show that this procedure corresponds to the original test. In the spirit of Ramdas, Trillos, and Cuturi [59], we conjecture that – if the goal of the researcher is to verify which credal set between $\mathcal{M}(\underline{P}_1)$ and $\mathcal{M}(\underline{P}_2)$ contains P^{true} – this procedure could be extended to one that includes the notion of Wasserstein metric between $\underline{P}_1$ and $\underline{P}_2$ to take into account how far $\mathcal{M}(\underline{P}_1)$ and $\mathcal{M}(\underline{P}_2)$ are from each other, thus making our contributions relevant.

Finally, our findings can be seminal in starting a new field of inquiry that generalizes ergodic transport theory

[2] In general, credal sets are in one-to-one correspondence with lower prevision functionals [68].

[3] What we mean is that solving the LP version of Monge's and Kantorovich's problems is equivalent to solving their classical versions. The solutions, then, will trivially coincide.

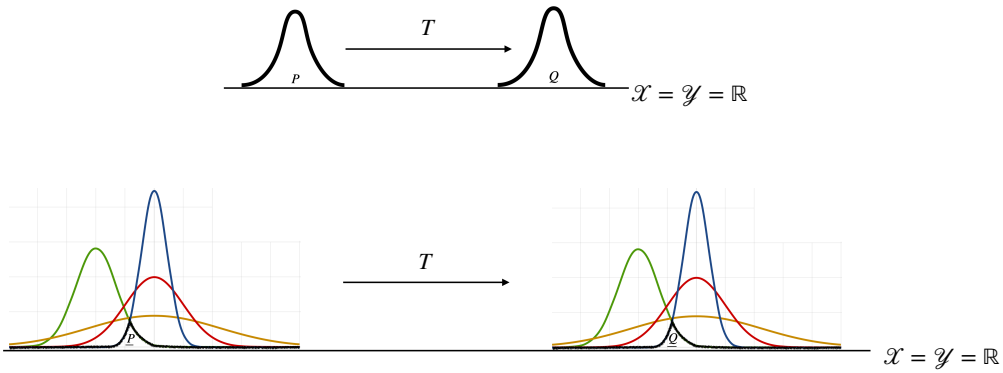

Figure 1. *Top: the optimal transport map T between two bell-shaped distributions P and Q on $\mathbb{R}$. Bottom: the optimal transport map T between lower probabilities $\underline{P}$ and $\underline{Q}$ (both depicted as black brushstrokes) that completely characterize credal sets $\mathcal{P}$ and $\mathcal{Q}$ of probabilities on $\mathbb{R}$. The colored distributions are elements of the respective credal sets.*

[20, 39, 46] to the credal setting. This may also be related to the field of computer vision (and especially the theory of convolutional autoencoders [72]). To see how, we refer the interested reader to Caprio [11], where the ergodicity of (a version of) imprecise Markov processes is related to the behavior of the outputs of a convolutional autoencoder, as the inputs are perturbed.

Structure of the paper. The paper is arranged as follows. Section 2 gives the necessary background on credal sets. Our results pertaining the lower probability version of Monge's and Kantorovich's problems are presented in Section 3. Section 4 concludes our work.

2. Background

In this section, we introduce the concepts of the core and of the Choquet integral. The reader who is familiar with these notions can skip to Section 3.

The main tool we work with in this paper is a particular type of a credal set (convex and weak*-closed set of probabilities [42]),[4] that is, what economists and operations researchers call the *core* (of an exact capacity) [13, 18, 50].

Given a capacity of interest – in this paper, it will always be a lower probability $\underline{P}$, i.e. a set function on the σ-algebra of interest, mapping in $[0,1]$, which is the lower envelope of a weak*-compact set [18, Section 2.1.(viii)] – on a generic measurable space $(\mathcal{X},\mathcal{F})$, the core is defined as

$$\mathcal{M}^{\text{fa}}(\underline{P}) := \{P \in \Delta^{\text{fa}}_{\mathcal{X}} : P(A) \geq \underline{P}(A), \forall A \in \mathcal{F}\}, \quad (1)$$

where $\Delta^{\text{fa}}_{\mathcal{X}}$ denotes the set of finitely additive probabilities on $(\mathcal{X},\mathcal{F})$.

We focus on cores for two main reasons. First, in general we have that the convex hull of a finite set of finitely additive probabilities on $\mathcal{X}$ is a *proper subset* of the core of the lower probability associated with that set, see e.g. [3, Example 1] and [4, Examples 6, 7, 8]. That is, given $\{P_k\}_{k=1}^K \subset \Delta^{\text{fa}}_{\mathcal{X}}$, $K < \infty$, we have that $\text{CH}(\{P_k\}_{k=1}^K) \subset \mathcal{M}^{\text{fa}}(\underline{P})$, where CH denotes the convex hull operator, and $\underline{P}(A) = \inf_{P \in \text{CH}(\{P_k\}_{k=1}^K)} P(A)$, for all $A \in \mathcal{F}$. Hence, focusing on the core gives us more generality.

Second, the core is *uniquely identified* by its lower probability [33]. To see this, notice that by knowing $\underline{P}$, we can reconstruct the set by simply considering all finitely additive probability measures on $\mathcal{X}$ that set-wise dominate $\underline{P}$.

Before proceeding to the main results of this paper, we need to introduce the concepts of pushforward lower probability (PLP) and of Choquet integral.

Definition 2.1 (Pushforward Lower Probability, PLP)**.** Given two measurable spaces $(\mathcal{X},\mathcal{F})$ and $(\mathcal{Y},\mathcal{G})$, a measurable mapping $T : \mathcal{X} \to \mathcal{Y}$, and a (coherent à la Walley [70, Section 2.5]) lower probability $\underline{P} = \inf_{P \in \mathcal{M}^{\text{fa}}(\underline{P})} P$, the pushforward of $\underline{P}$ is the (set) function $T_{\#}\underline{P} : \mathcal{G} \to [0,1]$ such that

$$T_{\#}\underline{P}(B) = \underline{P}(T^{-1}(B)), \quad \forall B \in \mathcal{G}.$$

Lemma 2.1 (PLPs are well-defined)**.** *The pushforward lower probability $T_{\#}\underline{P}$ in Definition 2.1 is a well-defined lower probability.*

Lemma 2.1 – whose proof relates PLPs to the mathematics of ambiguity [49], and is postponed to Appendix C – entails that $T_{\#}\underline{P}$ is a coherent lower probability on $\mathcal{G}$, hence a superadditive version of a pushforward probability measure [70, Section 1.6.4]. We now introduce Choquet integrals [22], [68, Section C.2].

Definition 2.2 (Choquet Integral)**.** Let $(\mathcal{Z},\mathcal{H})$ be a generic measurable space, and $\underline{P}$ be a generic lower probability on $\mathcal{Z}$. For each real-valued function f on $\mathcal{Z}$, we

[4] The weak* convergence is the setwise convergence on the fixed σ-algebra.

associate the extended real number

$$f \mapsto \int_{\mathcal{Z}} f \mathrm{d}\underline{P} := \int_0^\infty \underline{P}^\star(\{f^+ \geq t\})\mathrm{d}t - \int_0^\infty \left[\underline{P}^\star(\mathcal{Z}) - \underline{P}^\star(\{f^- \geq t\}^c)\right]\mathrm{d}t \tag{2}$$

called the *Choquet integral* of f with respect to $\underline{P}$, provided that the difference on the right-hand side is well defined. There, $f^+ := 0 \vee f$, and $f^- := -(0 \wedge f)$. Also, $\underline{P}^\star(A) := \sup_{B \subseteq A} \underline{P}(B)$, for all $A \in \mathcal{H}$, is the inner lower probability [70, Chapter 3.1],[5] and the integrals are (improper) Riemannian integrals.[6]

When (2) is well defined, we say that the Choquet integral $\int_{\mathcal{Z}} f \mathrm{d}\underline{P}$ of f with respect to $\underline{P}$ *exists*. We now report Troffaes and Cooman [68, Proposition C.3], which gives an alternative expression of $\int_{\mathcal{Z}} f \mathrm{d}\underline{P}$, and a sufficient condition for its existence.

Proposition 2.1 (Characterizing the Choquet Integral)**.** *Using the same notation as Definition 2, suppose the Choquet integral $\int_{\mathcal{Z}} f \mathrm{d}\underline{P}$ of f with respect to $\underline{P}$ exists. Then,*

$$\int_{\mathcal{Z}} f \mathrm{d}\underline{P} = \int_0^\infty \underline{P}^\star(\{f \geq t\})\mathrm{d}t + \int_{-\infty}^0 \left[\underline{P}^\star(\{f \geq t\}) - \underline{P}^\star(\mathcal{Z})\right]\mathrm{d}t.$$

In addition, if f is bounded or Borel measurable, then it is Choquet integrable *with respect to $\underline{P}$, that is, its Choquet integral $\int_{\mathcal{Z}} f \mathrm{d}\underline{P}$ exists.*

Corollary 2.1 (A Simplification of the Choquet Integral)**.** *Using the same notation as Definition 2, if f is positive and measurable, then $\int_{\mathcal{Z}} f \mathrm{d}\underline{P} = \int_0^\infty \underline{P}(\{f \geq t\})\mathrm{d}t$. If f is also bounded, then the weak inequality can be substituted by a strict one.*

Proof. The first part of the statement comes from Proposition 2.1 and Marinacci and Montrucchio [49, Equation (11)]. The second part is a consequence of Marinacci and Montrucchio [49, Proposition 17]. A similar result to Corollary 2.1 was proven by Grabisch [34]. □

3. Main results

In this section, we answer the two questions that we put forth in the Introduction. First, we give the general definitions of Monge's and Kantorovich's problems for transporting lower probabilities $\underline{P}$ and $\underline{Q}$. Then, we notice how for a special type of cores $\mathcal{M}(\underline{P})$ and $\mathcal{M}(\underline{Q})$, namely those associated with contaminated (countably additive) probabilities P and Q, such problems are equivalent to the classical ones, where the transport happens between P and Q. This is because $\mathcal{M}(\underline{P})$ and $\mathcal{M}(\underline{Q})$ are completely characterized by the lower probabilities $\underline{P} = (1-\epsilon)P$ and $\underline{Q} = (1-\epsilon)Q$. We show formally how the intuition that the $(1-\epsilon)$ scaling factor does not play a role when solving Monge's and Kantorovich's problems is correct. As a result, the lower probabilities and the classical versions of such problems coincide.

We begin by remarking two notational choices and an assumption that we make in the rest of the paper. We will put $\mathcal{M}^{\mathrm{fa}}(\underline{P}) \equiv \mathcal{M}(\underline{P})$, and we will call $\Delta_{\mathcal{X}}^{\mathrm{ca}}$ the space of countably additive probabilities on $\mathcal{X}$. We will also assume that the cost function c is (Borel) measurable, so that (i) inner lower probability $\underline{P}^\star$ and "classical" lower probability $\underline{P}$ coincide, and (ii) the Choquet integrals that we consider exist.

The special cores that we consider are the so-called *ϵ-contaminated credal sets*. That is, given a countably additive probability measure P on $\mathcal{X}$, $P \in \Delta_{\mathcal{X}}^{\mathrm{ca}}$, we consider the set

$$\mathcal{P}_\epsilon = \{\Pi \in \Delta_{\mathcal{X}}^{\mathrm{fa}} : \Pi(A) = (1-\epsilon)P(A) + \epsilon R(A), \forall R \in \Delta_{\mathcal{X}}^{\mathrm{fa}}, \forall A \in \mathcal{F}\}, \tag{3}$$

where ϵ is a parameter in $[0,1]$.

Lemma 3.1 (Properties of ϵ-Contaminated Credal Sets)**.** *Let $\mathcal{P}_\epsilon$ be an ϵ-contaminated credal set as in (3). Then, $\underline{P}'(A)$ is given by*

$$\inf_{\Pi \in \mathcal{P}_\epsilon} \Pi(A) = \begin{cases} (1-\epsilon)P(A), & \text{for all } A \in \mathcal{F} \setminus \{\mathcal{X}\} \\ 1, & \text{for } A = \mathcal{X} \end{cases} \tag{4}$$

and

$$\mathcal{P}_\epsilon = \mathcal{M}(\underline{P}') = \{\Pi \in \Delta_{\mathcal{X}}^{\mathrm{fa}} : \Pi(A) \geq \underline{P}'(A), \forall A \in \mathcal{F}\}. \tag{5}$$

Proof. Both these properties were proven in Wasserman and Kadane [71, Example 3] and in Walley [70, Section 2.9.2]. □

A remark is in order. The elements of $\mathcal{P}_\epsilon$ must be finitely additive probabilities, and not merely countably additive, because if that were not the case, then $\mathcal{P}_\epsilon$ would not be weak*-compact, and so it would not be a well-defined core. Let us give an example, borrowed from Walley [70]. Consider an ϵ-contamination model on the Naturals $\mathbb{N}$, with $\epsilon = 1$ (this is the vacuous lower probability that assigns lower probability 0 to every natural number). The sequence $(\delta_n)_{n \in \mathbb{N}}$ of Dirac measures δ_n assigning mass 1 to $n \in \mathbb{N}$ is a sequence of countably additive probability measures that belongs to $\mathcal{P}_\epsilon$. But this sequence has no weak* converging subsequence to a countably additive probability measure. If it did, it

[5] We need to work with $\underline{P}^\star$ because f may not be measurable. When it is, $\underline{P}^\star = \underline{P}$.

[6] For a primer on Riemannian integrals, see Troffaes and Cooman [68, Section C.1].

would have to assign probability 0 to all of the Naturals $n \in \mathbb{N}$. Hence, $\mathcal{P}_\epsilon$ cannot be weak*-compact in this case.

This is a technicality which does not affect the interpretation of our results, for two main reasons. First, all countably additive probabilities are also finitely additive, that is, $\Delta_{\mathcal{X}}^{\text{ca}} \subset \Delta_{\mathcal{X}}^{\text{fa}}$. Second, we consider contaminations of a countably additive probability $P \in \Delta_{\mathcal{X}}^{\text{ca}}$.[7] This is because we want to relate the lower probability versions of Monge's and Kantorovich's OT problems (that we introduce later in this Section) with the classical ones, that are formulated for countably additive probabilities.

In the remainder of the paper, we will work with the (incoherent, according to Walley [70, Section 2.5]) lower probability $\underline{P}$ such that

$$\underline{P}(A) = (1-\epsilon)P(A), \quad \text{for all } A \in \mathcal{F}, \tag{6}$$

in place of $\underline{P}'$. This is an unnormalized countably additive measure, which is Radon measure if the underlying space is separable. The reason we work with $\underline{P}$ in (6) is twofold: calculations are easier to carry out, and also the following lemma holds. The interested reader can find a further discussion on this choice in Appendix A.

Lemma 3.2 (A More Convenient Core)**.** *Pick any countably additive probability measure P on $\mathcal{X}$, any $\epsilon \in [0,1]$, and consider the two lower probabilities $\underline{P}'$ and $\underline{P}$ in (4) and (6), respectively. Let $\mathcal{M}(\underline{P}) = \{\Pi \in \Delta_{\mathcal{X}}^{\text{fa}} : \Pi(A) \geq \underline{P}(A), \forall A \in \mathcal{F}\}$. Then, $\mathcal{M}(\underline{P}') = \mathcal{M}(\underline{P})$.*

Proof. We begin by noting that $\underline{P}'$ is a well-defined lower probability by Cerreia-Vioglio, Maccheroni, and Marinacci [18, Section 2.1.(viii)]. Now, pick any $\Pi \in \Delta_{\mathcal{X}}^{\text{fa}}$. We have that $\Pi(A) \geq \underline{P}(A)$ if and only if $\Pi(A) \geq \underline{P}'(A)$, for all $A \in \mathcal{F} \setminus \{\mathcal{X}\}$. In addition, $\Pi(\mathcal{X}) = \underline{P}'(\mathcal{X}) = 1 > 1-\epsilon = \underline{P}(\mathcal{X})$. In turn, this shows that $\mathcal{M}(\underline{P}') = \mathcal{M}(\underline{P})$, concluding the proof. □

The intuition behind Lemma 3.2 is that $\underline{P}'$ and $\underline{P}$ only disagree (by ϵ much) on the value to assign to $\mathcal{X}$. But any finitely additive probability measure Π assigns probability 1 to the whole state space $\mathcal{X}$. So, to determine whether Π belongs to $\mathcal{M}(\underline{P}') = \mathcal{M}(\underline{P})$, it is enough to check if Π set-wise dominates $\underline{P}'$ and $\underline{P}$ on the events in $\mathcal{F} \setminus \{\mathcal{X}\}$. An immediate consequence of this argument is that we can write $\mathcal{M}(\underline{P}') = \mathcal{M}(\underline{P}) = \{\Pi \in \Delta_{\mathcal{X}}^{\text{fa}} : \Pi(A) \geq \underline{P}(A) = \underline{P}'(A), \forall A \in \mathcal{F} \setminus \{\mathcal{X}\}\}$.

Now, let $\mathcal{Q}_\epsilon \subset \Delta_{\mathcal{Y}}^{\text{fa}}$ be a credal set defined similarly to $\mathcal{P}_\epsilon$, and consider its "associated" lower probability $\underline{Q} = (1-\epsilon)Q$, $Q \in \Delta_{\mathcal{Y}}^{\text{ca}}$. We are interested in the Optimal Transport (OT) map between $\mathcal{M}(\underline{P})$ and $\mathcal{M}(\underline{Q})$. Because these sets are completely characterized by $\underline{P}$ and $\underline{Q}$, respectively (as we have seen in Section 2), we focus our attention on such lower probabilities.

[7] Of course, in general we may have $P \in \Delta_{\mathcal{X}}^{\text{fa}}$.

3.1. Lower probability Monge's (LPM) problem.

We begin our endeavor of finding the OT map by writing a version of Monge's optimal transport problem involving $\underline{P}$ and $\underline{Q}$. It is the following.

Definition 3.1 (Lower Probability Monge's OT Problem, LPM)**.** Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be a Borel measurable (cost) function. Given lower probabilities $\underline{P}$ and $\underline{Q}$ on $\mathcal{X}$ and $\mathcal{Y}$, respectively, we want to find the (measurable) optimal transport map $T : \mathcal{X} \to \mathcal{Y}$ that solves the following optimization problem

$$\arg\inf \left\{ \int_{\mathcal{X}} c(x, T(x)) \underline{P}(\mathrm{d}x) : T_{\#}\underline{P} = \underline{Q} \right\}. \tag{7}$$

We assume c to be Borel measurable and T to be measurable to ensure that $c(x, T(x))$ is Choquet integrable. Let us also notice that we work with the Choquet integral $\int_{\mathcal{X}} c(x, T(x)) \underline{P}(\mathrm{d}x)$ because, in the case of ϵ-contaminations, it corresponds to the lower expectation $\inf_{P \in \mathcal{M}(\underline{P})} \int_{\mathcal{X}} c(x, T(x)) P(\mathrm{d}x)$; this is because $\underline{P}$ is 2-monotone. In general, though, we have that $\inf_{P \in \mathcal{M}(\underline{P})} \int_{\mathcal{X}} c(x, T(x)) P(\mathrm{d}x) \leq \int_{\mathcal{X}} c(x, T(x)) \underline{P}(\mathrm{d}x)$; the Choquet integral (with respect to a lower probability) is an upper bound for the worst-case (i.e. lower) expectation. In the future, we will study how the solution to LPM changes if we consider more general cores (i.e. not necessarily ϵ-contaminations), and if we work with lower expectations in place of Choquet integrals.

Notice that the OT map T need not exist,[8] so we will need to verify its existence in every application of interest. We now show that, for ϵ-contaminated credal sets $\mathcal{P}_\epsilon$ and $\mathcal{Q}_\epsilon$, LPM is equivalent to the classical Monge's problem of finding the OT map between the contaminated probabilities P and Q. Throughout the rest of the paper, we perpetrate an abuse of terminology and refer to $\underline{P}$ and $\underline{Q}$ as "lower envelopes", even though they are not normalized at 1.

Theorem 3.1 (LPM Coincides with Classical Monge for ϵ-Contaminated Credal Sets)**.** *Suppose $\mathcal{X}$ and $\mathcal{Y}$ are separable, so that the elements of $\Delta_{\mathcal{X}}^{\text{ca}}$ and $\Delta_{\mathcal{Y}}^{\text{ca}}$ are Radon measures. If $\underline{P}$ and $\underline{Q}$ are the lower envelopes of the ϵ-contaminations $\mathcal{P}_\epsilon \subseteq \Delta_{\mathcal{X}}^{\text{fa}}$ and $\mathcal{Q}_\epsilon \subseteq \Delta_{\mathcal{Y}}^{\text{fa}}$ of $P \in \Delta_{\mathcal{X}}^{\text{ca}}$ and $Q \in \Delta_{\mathcal{Y}}^{\text{ca}}$, respectively, then the LPM of Definition 3.1 is equivalent to the classical Monge's OT Problem involving P and Q.*

Proof. By Proposition 2.1, Corollary 2.1, and Lemma 3.2, we have that

$$\begin{aligned} &\int_{\mathcal{X}} c(x, T(x)) \underline{P}(\mathrm{d}x) \\ &\quad = \int_0^\infty \underline{P}(\{x \in \mathcal{X} : c(x, T(x)) \geq t\})\, \mathrm{d}t \end{aligned}$$

[8] In the classical (precise) case, an optimal transport map T does not exist when P is a Dirac measure, but Q is not.

$$= \int_0^\infty (1-\epsilon)P\left(\{x \in \mathcal{X} : c(x, T(x)) \geq t\}\right) \mathrm{d}t \quad (8)$$
$$= (1-\epsilon) \int_0^\infty P\left(\{x \in \mathcal{X} : c(x, T(x)) \geq t\}\right) \mathrm{d}t$$
$$= (1-\epsilon) \int_{\mathcal{X}} c(x, T(x)) P(\mathrm{d}x).$$

In addition,

$$T_{\#}\underline{P}(B) = \underline{P}(T^{-1}(B)) = (1-\epsilon)P(T^{-1}(B)), \quad (9)$$

for all $B \in \mathcal{G}$, and

$$\underline{Q}(B) = (1-\epsilon)Q(B), \quad \forall B \in \mathcal{G}. \quad (10)$$

Hence, by (9) and (10), the constraint in (7) becomes

$$T_{\#}\underline{P} = \underline{Q} \iff (1-\epsilon)P \circ T^{-1} = (1-\epsilon)Q$$
$$\iff P \circ T^{-1} = Q \iff T_{\#}P = Q.$$

In turn, we can rewrite the LPM in equation (7) as

$$\arg\inf \left\{ \int_{\mathcal{X}} c(x, T(x)) \underline{P}(\mathrm{d}x) : T_{\#}\underline{P} = \underline{Q} \right\}$$
$$= \arg\inf \left\{ (1-\epsilon) \int_{\mathcal{X}} c(x, T(x)) P(\mathrm{d}x) : T_{\#}P = Q \right\}$$
$$= \arg\inf \left\{ \int_{\mathcal{X}} c(x, T(x)) P(\mathrm{d}x) : T_{\#}P = Q \right\}, \quad (11)$$

which is the classical Monge's OT problem. The last equality holds because c is a positive functional. Notice also that since $\underline{P}$ and $\underline{Q}$ are both countably additive measures normed to $(1-\epsilon)$, the Choquet integrals in this proof coincide with classical Lebesgue-Stieltjes integrals. □

In Theorem 3.1 we assume separability of $\mathcal{X}$ and $\mathcal{Y}$ to further relate the LPM and the classical Monge's OT problems. Indeed, in the latter, separability ensures the existence of Borel measurable selections, which is crucial for defining transport maps. Notice also that for Theorem 3.1 to hold we do not need to implicitly assume that the contaminating parameter ϵ is the same for both $\mathcal{P}_\epsilon$ and $\mathcal{Q}_\epsilon$. That is, we could consider $\mathcal{P}_\epsilon$ and $\mathcal{Q}_{\epsilon'}$, $\epsilon' \neq \epsilon$. This because, for the equivalences below (10) to work, it must be that $(1-\epsilon)/(1-\epsilon') = 1$, and so $\epsilon' = \epsilon$ must hold.

We now give an example, formulated as a corollary, in which Theorem 3.1 proves useful.

Corollary 3.1 (OT Map for LPM when $\mathcal{X} = \mathcal{Y} = \mathbb{R}$)**.** *Let $\mathcal{P}_\epsilon$ and $\mathcal{Q}_\epsilon$ denote the ϵ-contaminations of countably additive probability measures P and Q on $\mathcal{X} = \mathcal{Y} = \mathbb{R}$. Choose cost function c such that $c(x, y) = h(x-y)$, where h is a strictly convex, positive, Borel measurable functional. If P and Q have finite p-th moment, $p \in [1, \infty)$, and P has no atom, then the unique solution to LPM is $T = F_Q^{-1} \circ F_P$, where F_P and F_Q are the cdf's of P and Q, respectively.*

Proof. Rachev and Rüschendorf [57] show that, given our assumptions on P and Q, an optimal transport map T that attains the infimum in (11) exists, is unique, and is given by $T = F_Q^{-1} \circ F_P$. By Theorem 3.1, then, we know that the same OT map attains the infimum in (7). This concludes the proof. □

3.2. Lower probability Kantorovich's (LPK) problem. Adopting the Kantorovich formulation of the OT problem would strengthen our result, since – as we shall see in Corollary 3.2 – a suitable choice of the cost function c will ensure us that the OT map T exists. In addition, since most existing OT results are expressed as a solution to the classical Kantorovich OT problem, we would be able to immediately use them in the context of ϵ-contaminated credal sets.

The main difficulty coming from studying Kantorovich's version is that its extension to lower probabilities is not as immediate as the one in Definition 3.1. To see this, notice that a lower probability version of Kantorovich's OT problem is the following.

Definition 3.2 (Lower Probability Kantorovich's OT Problem, LPK)**.** Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be a Borel measurable (cost) function. Given lower probabilities $\underline{P}$ and $\underline{Q}$ on $\mathcal{X}$ and $\mathcal{Y}$, respectively, we want to find the joint lower probability $\underline{\alpha}$ (also called the *lower optimal transport plan*) on $\mathcal{X} \times \mathcal{Y}$ that solves the following optimization problem

$$\arg\inf \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \mathrm{d}\underline{\alpha}(x, y) : \underline{\alpha} \in \Gamma(\underline{P}, \underline{Q}) \right\}, \quad (12)$$

where $\Gamma(\underline{P}, \underline{Q})$ is the collection of all joint lower probabilities on $\mathcal{X} \times \mathcal{Y}$ whose marginals on $\mathcal{X}$ and $\mathcal{Y}$ are $\underline{P}$ and $\underline{Q}$, respectively.

Considerations on the choice of a Borel measurable cost function c and of working with a Choquet integral similar to those pointed out below Definition 3.1 hold also for Definition 3.2. In Imprecise Probability theory [7, 68, 70] there is not a unique way to perform conditioning [12, 14, 33], so we need to be extra careful when defining $\Gamma(\underline{P}, \underline{Q})$ in (12). In this work, we consider the joint lower probabilities resulting from *geometric conditioning*, and write $\Gamma(\underline{P}, \underline{Q}) \equiv \Gamma^{\text{geom}}(\underline{P}, \underline{Q})$. In that case, the conditional lower probability resulting from a joint probability $\underline{G}$ is derived as

$$\underline{G}(A \mid B) = \frac{\underline{G}(A, B)}{\underline{G}_{\mathcal{Y}}(B)}, \quad \forall A \in \mathcal{F}, \forall B \text{ s.t. } \underline{G}_{\mathcal{Y}}(B) > 0, \quad (13)$$

where $\underline{G}_{\mathcal{Y}} \equiv \underline{Q}$ is the marginal lower probability of $\underline{G}$ on $\mathcal{Y}$, and similarly for $\underline{G}_{\mathcal{X}} \equiv \underline{P}$. The importance of the choice of geometric conditioning is further discussed

in Appendix B. Let us mention here that – as pointed out by Gong and Meng [33] – the agent that chooses the geometric rule as a mechanism to update their belief is a pessimist. In fact, the geometric rule endorses a stringent interpretation of what counts as evidence for both the query (A) and conditioning (B) events, by admitting only evidence that supports its constituents into the lower conditional probability.

3.3. Restricted lower probability Kantorovich's (RLPK) problem. If the marginal lower probabilities correspond to the lower envelopes of ϵ-contaminated credal sets, then using joint lower probabilities that can be decomposed as in (13) entails that the elements of $\Gamma(\underline{P}, \underline{Q})$ are such that, for all $A \in \mathcal{F}$ and all $B \in \mathcal{G}$ such that $\underline{Q}(B) > 0$,

$$\underline{G}(A \mid B) = \frac{\underline{G}(A,B)}{\underline{Q}(B)} = \frac{\underline{G}(A,B)}{(1-\epsilon)Q(B)}$$

and similarly, $\underline{G}(B \mid A) = \frac{\underline{G}(B,A)}{\underline{P}(A)} = \frac{\underline{G}(A,B)}{(1-\epsilon)P(A)}$.

We can then consider a restricted version of LPK.

Definition 3.3 (Restricted Lower Probability Kantorovich's OT Problem, RLPK)**.** Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be a Borel measurable (cost) function. Given lower probabilities $\underline{P}$ and $\underline{Q}$ on $\mathcal{X}$ and $\mathcal{Y}$, respectively, we want to find the joint lower probability $\underline{\alpha}$ on $\mathcal{X}\times\mathcal{Y}$ that that solves the following optimization problem

$$\arg\inf \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) \mathrm{d}\underline{\alpha}(x,y) : \underline{\alpha} \in \Gamma_R(\underline{P}, \underline{Q}) \right\}, \quad (14)$$

where $\Gamma_R(\underline{P}, \underline{Q}) \subset \Gamma(\underline{P}, \underline{Q})$ is the collection of all joint lower probabilities (i) that can be written as an ϵ-contamination of countably additive joint probability measures $G \in \Delta^{\mathrm{ca}}_{\mathcal{X}\times\mathcal{Y}}$, and (ii) whose marginals on $\mathcal{X}$ and $\mathcal{Y}$ are $\underline{P}$ and $\underline{Q}$, respectively.

Definition 3.3 entails that, if $\underline{P}$ and $\underline{Q}$ are the lower envelopes of the ϵ-contaminations $\mathcal{P}_\epsilon$ and $\mathcal{Q}_\epsilon$, respectively, then an element $\underline{G} \in \Gamma_R(\underline{P}, \underline{Q})$ is such that

$$\underline{G}(A \mid B) = \frac{\underline{G}(A,B)}{\underline{Q}(B)} = \frac{(1-\epsilon)G(A,B)}{(1-\epsilon)Q(B)} = \frac{G(A,B)}{Q(B)},$$

and similarly,

$$\underline{G}(B \mid A) = \frac{\underline{G}(A,B)}{\underline{P}(A)} = \frac{(1-\epsilon)G(A,B)}{(1-\epsilon)P(A)} = \frac{G(A,B)}{P(A)}.$$

Notice how working with $\Gamma_R(\underline{P}, \underline{Q})$ is reminiscent of the covariate shift condition in the Machine Learning literature [58, Section 3.3.4.4]. That is, a situation in which there is ambiguity on the marginal distribution of the input features of the model (but not on the conditional distribution of the output, given the input), which may be different (i.e. may have changed) from the one that the model has "seen" during training and validation.

We now show that, for ϵ-contaminated credal sets $\mathcal{P}_\epsilon$ and $\mathcal{Q}_\epsilon$, RLPK is equivalent to the classical Kantorovich's OT problem. The result need not hold if either the unrestricted LPK or a different type of conditioning are considered. We will expand on this in Remark 3.1.

Theorem 3.2 (RLPK Coincides with Classical Kantorovich for ϵ-Contaminated Credal Sets)**.** *Suppose $\mathcal{X}$ and $\mathcal{Y}$ are separable, so that the elements of $\Delta^{\mathrm{ca}}_{\mathcal{X}}$ and $\Delta^{\mathrm{ca}}_{\mathcal{Y}}$ are Radon measures. If $\underline{P}$ and $\underline{Q}$ are the lower envelopes of the ϵ-contaminations $\mathcal{P}_\epsilon \subseteq \Delta^{\mathrm{fa}}_{\mathcal{X}}$ and $\mathcal{Q}_\epsilon \subseteq \Delta^{\mathrm{fa}}_{\mathcal{Y}}$ of $P \in \Delta^{\mathrm{ca}}_{\mathcal{X}}$ and $Q \in \Delta^{\mathrm{ca}}_{\mathcal{Y}}$, respectively, then the RLPK of Definition 3.3 is equivalent to the classical Kantorovich's OT Problem involving P and Q.*

Proof. Pick any element $\underline{G}$ of $\Gamma_R(\underline{P}, \underline{Q})$. We have that

$$\begin{aligned}
\underline{G}(A,B) &= \underline{G}(A \mid B)\underline{Q}(B) = \frac{G(A,B)}{Q(B)}(1-\epsilon)Q(B) \quad (15)\\
&= \underline{G}(B \mid A)\underline{P}(A) = \frac{G(A,B)}{P(A)}(1-\epsilon)P(A) \quad (16)\\
&= (1-\epsilon)G(A,B).
\end{aligned}$$

So we can write $\Gamma_R(\underline{P}, \underline{Q}) = (1-\epsilon)\Gamma(P,Q) = \{(1-\epsilon)G : G \in \Gamma(P,Q)\}$, where set $\Gamma(P,Q)$ is the the collection of all (countably additive) probability measures on $\mathcal{X}\times\mathcal{Y}$ whose marginals on $\mathcal{X}$ and $\mathcal{Y}$ are P and Q, respectively. This shows that $\Gamma_R(\underline{P}, \underline{Q})$ is nonempty if and only if $\Gamma(P,Q) \neq \emptyset$. In addition, it is easy to see that $\Gamma_R(\underline{P}, \underline{Q})$ inherits the convexity and the weak*-compactness from $\Gamma(P,Q)$.[9] In turn,

$$\begin{aligned}
&\arg\inf \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) \mathrm{d}\underline{\alpha}(x,y) : \underline{\alpha} \in \Gamma_R(\underline{P}, \underline{Q}) \right\} =\\
&\arg\inf \left\{ (1-\epsilon) \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) \mathrm{d}\alpha(x,y) : \alpha \in \Gamma(P,Q) \right\} \quad (17)\\
&= \arg\inf \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) \mathrm{d}\alpha(x,y) : \alpha \in \Gamma(P,Q) \right\}, \quad (18)
\end{aligned}$$

where (17) comes from Proposition 2.1 and our definition of $\Gamma_R(\underline{P}, \underline{Q})$, and the last equality comes from c being positive. The fact that (18) is the classical Kantorovich's OT Problem [40] concludes our proof. Notice also that since $\underline{P}$ and $\underline{Q}$ are both countably additive measures normed to $(1-\epsilon)$, the Choquet integrals in this proof coincide with classical Lebesgue-Stieltjes integrals. □

[9] It is easy to see that $\Gamma(P,Q)$ is convex. In addition, its weak*-compactness comes from Prokhorov's theorem, following our separability assumptions and the well-known fact that $\Gamma(P,Q)$ is weak*-closed.

Notice that for Theorem 3.2 too we do not need to implicitly assume that the contaminating parameter ϵ is the same for both $\mathcal{P}_\epsilon$ and $\mathcal{Q}_\epsilon$. That is, we could consider $\mathcal{P}_\epsilon$ and $\mathcal{Q}_{\epsilon'}$, $\epsilon' \neq \epsilon$. This because, by (15), we have that $\underline{G}(A,B) = (1-\epsilon')G(A,B)$, and, by (16), that $\underline{G}(A,B) = (1-\epsilon)G(A,B)$. But they must be equal to each other, and so $\epsilon = \epsilon'$ must hold.

We now give sufficient conditions for the minimizer of (14) to exist, in the context of ϵ-contaminated credal sets. First, we need to introduce the concept of tightness of $\Gamma_R(\underline{P},\underline{Q})$.

Definition 3.4 (Tightness of $\Gamma_R(\underline{P},\underline{Q})$)**.** Let $(\mathcal{X}\times\mathcal{Y},\tau)$ be a Hausdorff space. Let $\Sigma_{\mathcal{X}\times\mathcal{Y}}$ be a σ-algebra on $\mathcal{X}\times\mathcal{Y}$ that *contains* τ. That is, every τ-open subset of $\mathcal{X}\times\mathcal{Y}$ is measurable, and $\Sigma_{\mathcal{X}\times\mathcal{Y}}$ is at least as fine as the Borel σ-algebra on $\mathcal{X}\times\mathcal{Y}$. We say that $\Gamma_R(\underline{P},\underline{Q})$ is *tight* if, for all $\delta\in(0,1]$, there exists a τ-compact set $K_\delta\in\Sigma_{\mathcal{X}\times\mathcal{Y}}$ such that, for all $\underline{\alpha}\in\Gamma_R(\underline{P},\underline{Q})$, we have that $\underline{\alpha}(K_\delta) > 1-\delta$.

Lemma 3.3 (Necessary and Sufficient Condition for $\Gamma_R(\underline{P},\underline{Q})$ to be Tight)**.** *Let $\mathcal{X},\mathcal{Y}$ be metric spaces, and $\epsilon\in[0,1)$. Then, set $\Gamma_R(\underline{P},\underline{Q})$ is tight if and only if set $\Gamma(P,Q)$ is tight.*

Proof. Suppose $\Gamma_R(\underline{P},\underline{Q})$ is tight. Given our assumption that $\mathcal{X}$ and $\mathcal{Y}$ are metric spaces, this implies that $\mathcal{X}$ and $\mathcal{Y}$ are separable. Pick any $\delta\in(0,1]$ and any $\underline{\alpha}\in\Gamma_R(\underline{P},\underline{Q})$. Then, by Definition 3.4, we have that $\underline{\alpha}(K_\delta) > 1-\delta$. By the proof of Theorem 3.2, we know that there exists $\alpha\in\Gamma(P,Q)$ such that $\underline{\alpha}(A) = (1-\epsilon)\alpha(A)$, for all $A\in\Sigma_{\mathcal{X}\times\mathcal{Y}}$. In turn, this implies that $(1-\epsilon)\alpha(K_\delta) > 1-\delta \iff \alpha(K_\delta) > \frac{1-\delta}{1-\epsilon}$. Now let $\frac{1-\delta}{1-\epsilon} =: 1-\gamma$, and put $K_\delta \equiv K_\gamma$. We obtain $\alpha(K_\gamma) > 1-\gamma$. But δ and $\underline{\alpha}$ were chosen arbitrarily, which allows us to conclude that $\Gamma(P,Q)$ is tight.

Suppose instead that $\Gamma(P,Q)$ is tight. As before, given our assumption that $\mathcal{X}$ and $\mathcal{Y}$ are metric spaces, this implies that $\mathcal{X}$ and $\mathcal{Y}$ are separable. Pick any $\delta\in(0,1]$, and any $\alpha\in\Gamma(P,Q)$. Then, by Definition 3.4, we have that $\alpha(K_\delta) > 1-\delta$. This holds if and only if $(1-\epsilon)\alpha(K_\delta) = \underline{\alpha}(K_\delta) > (1-\epsilon)(1-\delta)$, where $\epsilon\in[0,1)$ is the same parameter of Definition 3.3. Now let $(1-\epsilon)(1-\delta) =: 1-\gamma$, and put $K_\delta \equiv K_\gamma$. We obtain $\underline{\alpha}(K_\gamma) > 1-\gamma$. But δ and α were chosen arbitrarily, which allows us to conclude that $\Gamma_R(\underline{P},\underline{Q})$ is tight. □

We are ready for our result.

Corollary 3.2 (Existence of OT Plan)**.** *Let $\mathcal{X},\mathcal{Y}$ be metric spaces, and $\epsilon\in[0,1)$. If $\Gamma_R(\underline{P},\underline{Q})$ is tight, and if cost function c in (14) is also lower semicontinuous, then a minimizer for (14) exists.*

Proof. Let $\Gamma_R(\underline{P},\underline{Q})$ be tight. Given our assumption that $\mathcal{X}$ and $\mathcal{Y}$ are metric spaces, this implies that $\mathcal{X}$ and $\mathcal{Y}$ are separable. Ambrosio, Gigli, and Savaré [5] show that if $\Gamma(P,Q)$ is tight, and if c is lower semicontinuous, then there is a minimizer for the classical Kantorovich's OT problem. By Lemma 3.3, we know that – for any $\epsilon\in[0,1)$ – if $\Gamma_R(\underline{P},\underline{Q})$ is tight, then so is $\Gamma(P,Q)$. The proof follows by the equivalence established in Theorem 3.2. □

The tightness condition in Corollary 3.2 is satisfied e.g. when $\mathcal{X}$ and $\mathcal{Y}$ are both Polish spaces.[10] This is an immediate consequence of Thorpe [66, Proposition 1.5].

3.4. Equivalence between LPM and RLPK problems. We now inspect when do RLPK and LPM coincide, in the context of ϵ-contaminated credal sets.

Theorem 3.3 (RLPK is Equivalent to LPM)**.** *Suppose $\mathcal{X}$ and $\mathcal{Y}$ are separable, so that the elements of $\Delta_{\mathcal{X}}^{\mathrm{ca}}$ and $\Delta_{\mathcal{Y}}^{\mathrm{ca}}$ are Radon measures. Let $\underline{P}$ and $\underline{Q}$ be the lower envelopes of the ϵ-contaminations $\mathcal{P}_\epsilon\subseteq\Delta_{\mathcal{X}}^{\mathrm{fa}}$ and $\mathcal{Q}_\epsilon\subseteq\Delta_{\mathcal{Y}}^{\mathrm{fa}}$ of $P\in\Delta_{\mathcal{X}}^{\mathrm{ca}}$ and $Q\in\Delta_{\mathcal{Y}}^{\mathrm{ca}}$, respectively. When the minimizer $\underline{\alpha}$ of RLPK is such that $\mathrm{d}\underline{\alpha}(x,y) = \underline{P}(\mathrm{d}x)\delta_{y=T(x)}$, then T is an optimal transport map and RLPK is equivalent to LPM.*

Proof. Given the way we defined $\underline{P}$ and $\underline{Q}$, we have that $\mathrm{d}\underline{\alpha}(x,y) = \underline{P}(\mathrm{d}x)\delta_{y=T(x)} \iff (1-\epsilon)\mathrm{d}\alpha(x,y) = (1-\epsilon)P(\mathrm{d}x)\delta_{y=T(x)} \iff \mathrm{d}\alpha(x,y) = P(\mathrm{d}x)\delta_{y=T(x)}$. Thorpe [66, Section 1.2] shows that when the minimizer α of Kantorovich's classical OT problem is such that $\mathrm{d}\alpha(x,y) = P(\mathrm{d}x)\delta_{y=T(x)}$,[11] then T is Monge's OT map, and Monge's and Kantorovich's problems are equivalent. The proof, then, follows by Theorems 3.1 and 3.2. □

We now give an example, formulated as a corollary, in which Theorem 3.3 proves useful.

Corollary 3.3 (Multivariate Normal Case)**.** *Let $\underline{P}$ and $\underline{Q}$ be the lower envelopes of the ϵ-contaminations $\mathcal{P}_\epsilon,\mathcal{Q}_\epsilon\subseteq\Delta_{\mathbb{R}^d}^{\mathrm{fa}}$ of $P=\mathcal{N}_d(0,\Sigma_P)$ and $Q=\mathcal{N}_d(0,\Sigma_Q)$, two (countably additive) multivariate Normals on $\mathcal{X}=\mathcal{Y}=\mathbb{R}^d$, respectively. Select $c(x,y)=|y-Ax|^2/2$, where $A\in\mathbb{R}^{d\times d}$ is invertible. Then, the optimal map that solves LPM is $x\mapsto T(x)$,*

$$T(x) = (A^\top)^{-1}\Sigma_P^{-1/2}\left(\Sigma_P^{1/2}A^\top\Sigma_Q A\Sigma_P^{1/2}\right)^{1/2}\Sigma_P^{-1/2}x \tag{19}$$

and the optimal plan that solves RLPK is $\mathrm{d}\underline{\alpha}(x,y) = \underline{P}(\mathrm{d}x)\delta_{y=T(x)}$.

Proof. Notice that $\mathbb{R}^d$ is separable. Galichon [30] shows that if $P=\mathcal{N}_d(0,\Sigma_P)$, $Q=\mathcal{N}_d(0,\Sigma_Q)$, and $c(x,y)=|y-Ax|^2/2$, then Monge's (classical) OT map is the one in (19), and also that $\mathrm{d}\alpha(x,y) = P(\mathrm{d}x)\delta_{y=T(x)}$, so that Monge's and Kantorovich's (classical) problems are equivalent.

[10] Separable, completely metrizable topological spaces

[11] Conditions sufficient for such a condition can be found in Thorpe [66, Chapter 4].

Now, by Theorem 3.1, we have that if $\underline{P}$ and $\underline{Q}$ are lower envelopes of the ϵ-contaminations $\mathcal{P}_\epsilon, \mathcal{Q}_\epsilon \subseteq \Delta^{\text{fa}}_{\mathbb{R}^d}$ of $P = \mathcal{N}_d(0, \Sigma_P)$ and $Q = \mathcal{N}_d(0, \Sigma_Q)$, respectively, then LPM coincides with the classical Monge's OT Problem involving P and Q. In turn, this implies that the map in (19) is also the OT map between $\underline{P}$ and $\underline{Q}$. In addition, by Theorem 3.2, we know that RLPK coincides with the classical Kantorovich's OT Problem involving P and Q. This implies that the optimal lower coupling $\underline{\alpha}$ is given by $\mathrm{d}\underline{\alpha}(x, y) = \underline{P}(\mathrm{d}x)\delta_{y=T(x)}$. □

Remark 3.1 (On the Equivalence of Monge and Kantorovich)**.** A consequence of Theorem 3.3 is that, for ϵ-contaminated credal sets, LPM and LPK with $\Gamma(\underline{P}, \underline{Q}) \equiv \Gamma^{\text{geom}}(\underline{P}, \underline{Q})$ **need not** coincide. To see this, notice that Theorem 3.2 only holds for the **restricted** LPK (RLPK) in Definition 3.3. Had we not specified that the joint lower probabilities $\underline{G} \in \Gamma_R(\underline{P}, \underline{Q})$ are ϵ-contaminations of countably additive joint probabilities $G \in \Delta^{\text{ca}}_{\mathcal{X}\times\mathcal{Y}}$, then Theorem 3.2 may not have held. Similarly, had we considered generalized Bayes' conditioning, or other conditioning mechanisms for lower probabilities [14], Theorem 3.2 may not have held as well.

Whether this is a phenomenon pertaining only to ϵ-contaminated credal sets, or a more general one, will be the subject of future studies.

4. Conclusion

The conclusion that we can derive from this work is that Questions 1 and 2 in the Introduction have a positive answer. We can formulate a version of Monge's and Kantorovich's problems for lower probabilities. In addition, we can indeed find one class of credal sets completely characterized by their lower probability (the class of ϵ-contaminations) for which the optimal transport map and plan coincide with the classical cases. We also inspected when our versions of the two problems coincide, and find out that this need not hold in general.

With this work, we begin to explore the exciting venue of optimal transport between lower probabilities completely characterizing credal sets. In the future, we plan to further our study of optimal transport between ϵ-contaminations by deriving a Brenier-type theorem [10] and a Kantorovich-Rubinstein-type duality result [30] (which could potentially have a significant impact in economics [23–25, 41]). We also intend to explore the machine learning applications (especially concerning out-of-distribution detection) of our findings, and of distributionally robust optimization [28]. Finally, we will extend our focus to other types of credal sets that are not necessarily completely characterized by lower probabilities, such as finitely generated credal sets (the convex hull of finitely many distributions), to signed measures, and to second-order distributions, that is, distributions over distributions.

A. On the choice of $\underline{P}$

In this section, we discuss our choice of the core $\mathcal{M}(\underline{P})$ in Lemma 3.2. We work with the (core of the incoherent, according to Walley [70, Section 2.5]) lower probability $\underline{P}$ because it makes it easier to derive our desired results, e.g. the proof of Theorem 3.1.

Had we worked with $\underline{P}'$ instead, we would have had (by Corollary 2.1) $\int_{\mathcal{X}} c(x, T(x))\underline{P}'(\mathrm{d}x) = \int_0^\infty \underline{P}'(\{x \in \mathcal{X} : c(x, T(x)) \geq t\})\,\mathrm{d}t$. It is not immediate to show that the latter is equal to $(1-\epsilon)\int_{\mathcal{X}} c(x, T(x))P(\mathrm{d}x)$. To see this, notice that there might be some value $\bar{t} \in \mathbb{R}_+$ for which all $x \in \mathcal{X}$ are such that $c(x, T(x)) \geq \bar{t}$. In that case, $\underline{P}'(\{x \in \mathcal{X} : c(x, T(x)) \geq \bar{t}\}) = \underline{P}'(\mathcal{X}) = 1$, and so the "trick" that we used in (8) does not work anymore.

To achieve the desired result easily while working with $\underline{P}'$, we would have had to require that the cost function c is bounded. Indeed, suppose that the latter holds, and call $\underline{c} := \inf_{x\in\mathcal{X}} c(x, T(x))$ and $\overline{c} := \sup_{x\in\mathcal{X}} c(x, T(x))$. Then, we can use Troffaes and Cooman [68, Theorem C.3.(ii).(C.7)] to get

$$\begin{aligned}
&\int_{\mathcal{X}} c(x, T(x))\underline{P}'(\mathrm{d}x) \\
&\quad = \underline{P}'(\mathcal{X}) \cdot \underline{c} + \int_{\underline{c}}^{\overline{c}} \underline{P}'(\{x \in \mathcal{X} : c(x, T(x)) > t\})\,\mathrm{d}t \\
&\quad = \underline{c} + (1-\epsilon)\int_{\underline{c}}^{\overline{c}} P(\{x \in \mathcal{X} : c(x, T(x)) > t\})\,\mathrm{d}t \\
&\quad = (1-\epsilon)\int_{\mathcal{X}} c(x, T(x))P(\mathrm{d}x).
\end{aligned}$$

As we can see, the desired result becomes either harder to prove (if we only ask for c to be Borel measurable), or it needs an extra assumption (a bounded cost function c). A similar argument holds also for the Kantorovich's results.

Since the main goal of the paper is to transport lower probabilities *that completely characterize credal sets*, and since by Lemma 3.2 we know that $\mathcal{M}(\underline{P}) = \mathcal{M}(\underline{P}')$, we opted for using the incoherent lower probability $\underline{P}$ instead of the coherent one $\underline{P}'$.

We conclude with a remark. We acknowledge that working with the incoherent lower probability $\underline{P}$ makes it harder to use the techniques that we employ in this work, for models that are more complex than the ϵ-contaminations that we study. How to overcome this shortcoming will be the object of future work.

B. On the difference between conditioning methods

Let us illustrate the difference that the choice of conditioning rule makes when working with imprecise probabilities. Suppose that, instead of considering geometric,

we choose generalized Bayes' conditioning [70, Section 6.4]. That is, for a generic credal set $\mathcal{P}$, for all $A \in \mathcal{F}$ and all $B \in \mathcal{G}$ such that $G(B) > 0$,

$$\underline{G}^{\text{GBC}}(A \mid B) := \inf_{P \in \mathcal{P}} \left[\frac{G(A,B)}{G(B)} \right].$$

By Walley [70, Theorem 6.4.6], we have that

$$\inf_{P \in \mathcal{P}} \left[\frac{G(A,B)}{G(B)} \right] = \frac{\underline{G}(A,B)}{\overline{G}(B)},$$

so

$$\begin{aligned} \underline{G}^{\text{GBC}}(A \mid B) &= \frac{\underline{G}(A,B)}{\overline{G}(B)} \\ &\leq \frac{\underline{G}(A,B)}{\underline{G}(B)} =: \underline{G}^{\text{geom}}(A \mid B), \end{aligned}$$

since $\overline{G}(B) \geq \underline{G}(B)$, for all $B \in \mathcal{G}$. This was also proven in Gong and Meng [33, Lemma 4.3]. This inequality still holds true even in a simple model like ϵ-contaminated credal sets that we consider in the present work. To see this, notice that – in the same notation as Lemma 3.1 – by Wasserman and Kadane [71, Example 3] in an ϵ-contamination model $\mathcal{P}_\epsilon$ we have that $\overline{P}'(A) = (1-\epsilon)P(A) + \epsilon$, for all $A \in \mathcal{F} \setminus \{\emptyset\}$, and $\overline{P}'(\emptyset) = 0$. Similarly to what we did in the main body of the paper, we can focus instead the incoherent upper probability $\overline{P}(A) = (1-\epsilon)P(A) + \epsilon$, for all $A \in \mathcal{F}$. Then, by Walley [70, Section 6.6.2], we have that

$$\underline{G}^{\text{GBC}}(A \mid B) = \frac{(1-\epsilon)G(A,B)}{(1-\epsilon)Q(B) + \epsilon}$$

and, similarly, that

$$\underline{G}^{\text{GBC}}(B \mid A) = \frac{(1-\epsilon)G(A,B)}{(1-\epsilon)P(A) + \epsilon}.$$

In turn, this implies that the elements of $\Gamma_R^{\text{GBC}}(\underline{P}, \underline{Q})$ are such that $\underline{G}(A,B) = \underline{G}^{\text{GBC}}(A \mid B)\overline{Q}(B) = \underline{G}^{\text{GBC}}(B \mid A)\overline{P}(A)$, for all $A \in \mathcal{F}$ and all $B \in \mathcal{G}$ (having positive upper probability). They are different than the elements of $\Gamma_R^{\text{geom}}(\underline{P}, \underline{Q})$ that we introduced in Definition 3.3.

C. Proof of Lemma 2.1

We begin by noting that $T_\# \underline{P}$ is a real-valued set function, and that $T_\# \underline{P}(\emptyset) = \underline{P}(T^{-1}(\emptyset)) = \underline{P}(\emptyset) = 0$.[12] In turn, $T_\# \underline{P}$ is what Marinacci and Montrucchio [49, Section 2.1] call a *game*. In addition, since $\underline{P}(T^{-1}(B)) \in [0,1]$, for all $B \in \mathcal{G}$, we have that the co-domain of $T_\# \underline{P}$ is $[0,1]$, and so $T_\# \underline{P}$ is what Marinacci and Montrucchio [49, Section 2.1.2] call a *bounded game*. We can then consider the core of such a bounded game Marinacci and Montrucchio [49, Section 2.2], which is a slight generalization of the core introduced earlier. It is defined as $\mathcal{M}^{\text{bc}}(T_\# \underline{P}) := \{\mu \in \text{bc}(\mathcal{G}) : \mu(B) \geq T_\# \underline{P}(B), \forall B \in \mathcal{G}, \text{ and } \mu(\mathcal{Y}) = T_\# \underline{P}(\mathcal{Y})\}$, where $\text{bc}(\mathcal{G})$ is the vector spaces of all bounded charges (signed, finitely additive measures) on $\mathcal{G}$.

Notice that $\mathcal{M}^{\text{bc}}(T_\# \underline{P})$ is nonempty because it contains the pushforward measures of the elements of $\mathcal{M}^{\text{fa}}(\underline{P})$ in (1) through map T. In formulas, $P \in \mathcal{M}^{\text{fa}}(\underline{P}) \implies T_\# P \in \mathcal{M}^{\text{bc}}(T_\# \underline{P})$. To see that this is the case, notice that a finitely additive probability is a special case of a bounded charge, and that $\mathcal{M}^{\text{fa}}(\underline{P})$ contains all the finitely additive probabilities on $\mathcal{X}$ that set-wise dominate $\underline{P}$. More formally, pick any $\tilde{B} \in \mathcal{G}$ and any $P \in \mathcal{M}^{\text{fa}}(\underline{P})$. Let $\tilde{A} = T^{-1}(\tilde{B})$. Then, by (1), $T_\# P(\tilde{B}) = P(T^{-1}(\tilde{B})) = P(\tilde{A}) \geq \underline{P}(\tilde{A}) = \underline{P}(T^{-1}(\tilde{B})) = T_\# \underline{P}(\tilde{B})$. But then $T_\# P \in \mathcal{M}^{\text{bc}}(T_\# \underline{P})$, which shows that $\mathcal{M}^{\text{bc}}(T_\# \underline{P}) \neq \emptyset$. In turn, by Marinacci and Montrucchio [49, Proposition 3], we have that $\mathcal{M}^{\text{bc}}(T_\# \underline{P})$ is weak*-compact.

Now, notice that $\mathcal{M}^{\text{fa}}(T_\# \underline{P}) := \{Q \in \Delta_{\mathcal{Y}}^{\text{fa}} : Q(B) \geq T_\# \underline{P}(B), \forall B \in \mathcal{G}\}$ is a proper subset of $\mathcal{M}^{\text{bc}}(T_\# \underline{P})$, since $\mathcal{M}^{\text{fa}}(T_\# \underline{P})$ only considers the *finitely additive* probabilities (and not all the bounded charges) that set-wise dominate $T_\# \underline{P}$.

Let now $(Q_\alpha)_{\alpha \in \mathcal{I}}$ be a net in $\mathcal{M}^{\text{fa}}(T_\# \underline{P})$ that weak*-converges to $Q \in \Delta_{\mathcal{Y}}^{\text{fa}}$. Here $\mathcal{I}$ is a generic index set. This means that $Q_\alpha(B) \to Q(B)$, for all $B \in \mathcal{G}$. That is, pick any $B \in \mathcal{G}$; then,

$$\forall \epsilon > 0, \exists \tilde{\alpha}_\epsilon : \forall Q_\alpha \succeq Q_{\tilde{\alpha}_\epsilon}, |Q_\alpha(B) - Q(B)| < \epsilon. \quad (20)$$

Equation (20) implies that, for all $Q_\alpha \succeq Q_{\tilde{\alpha}_\epsilon}$, we have that $Q(B) > Q_\alpha(B) - \epsilon \geq T_\# \underline{P}(B) - \epsilon$. Letting $\epsilon \to 0$, this implies that $Q(B) \geq T_\# \underline{P}(B)$. But B was chosen arbitrarily in $\mathcal{G}$, and so $Q \in \mathcal{M}^{\text{fa}}(T_\# \underline{P})$. Hence, $\mathcal{M}^{\text{fa}}(T_\# \underline{P})$ is weak* sequentially closed, and therefore it is weak* closed. Being a weak* closed subset of a weak* compact space, we can conclude that $\mathcal{M}^{\text{fa}}(T_\# \underline{P})$ is weak* compact itself.

By Cerreia-Vioglio, Maccheroni, and Marinacci [18, Section 2.1.(viii)], we know that a (set) function $\nu : \mathcal{G} \to [0,1]$ is a lower probability if and only if there exists a weak*-compact set $\mathcal{M} \subseteq \Delta_{\mathcal{Y}}^{\text{fa}}$ such that $\nu(B) = \min_{Q \in \mathcal{M}} Q(B)$, for all $B \in \mathcal{G}$. Letting $\mathcal{M} \equiv \mathcal{M}^{\text{fa}}(T_\# \underline{P})$ and $\nu \equiv T_\# \underline{P}$, we obtain that $T_\# \underline{P}(B) \leq \min_{Q \in \mathcal{M}^{\text{fa}}(T_\# \underline{P})} Q(B)$, for all $B \in \mathcal{G}$. Suppose now for the sake of contradiction that there is $\tilde{B} \in \mathcal{G}$ such that $T_\# \underline{P}(\tilde{B}) < \min_{Q \in \mathcal{M}^{\text{fa}}(T_\# \underline{P})} Q(\tilde{B})$. This would imply that $\mathcal{M}^{\text{fa}}(T_\# \underline{P}) \supsetneq \mathcal{M}^{\text{fa}}(T_\# \underline{P})$, a contradiction. Hence, we can conclude that $T_\# \underline{P}(B) = \min_{Q \in \mathcal{M}^{\text{fa}}(T_\# \underline{P})} Q(B)$, for all $B \in \mathcal{G}$, thus completing the proof. □

[12] We always (implicitly) assume that $T^{-1}(\emptyset) = \emptyset$.

Additional author information

Acknowledgements. We wish to express our gratitude to Yusuf Sale for insightful discussions, particularly on Lemma 2.1, Corollary 3.3, and Remark 3.1. We are also grateful to Krikamol Muandet and Siu Lun (Alan) Chau for suggesting the relationship between RLPK and the covariate shift condition in Machine Learning, and to four anonymous reviewers for their insightful comments.

We would also like to point out how this work was inspired by Daniel Kuhn's seminar at the Society for Imprecise Probabilities on December 12, 2023. There, he presented some of his and his colleagues' remarkable works on the relationship between optimal transport and distributionally robust optimization, a field studying decision problems under uncertainty framed as zero-sum games against Nature [8, 9, 28, 31, 54, 55, 63, 65]. This immediately led us to think about the possibility of studying optimal transport between lower probabilities characterizing credal sets.

Finally, as highlighted on the title page, we would like to dedicate the present work to the memory of Sayan Mukherjee. Without his guidance, we would never have had the strength – and perhaps the hubris – to start studying imprecise probability theory.

References

[1] Joaquín Abellán and George J. Klir. "Additivity of uncertainty measures on credal sets". In: *International Journal of General Systems* 34.6 (2005), pp. 691–713.

[2] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. "Optimal Testing for Properties of Distributions". In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett. Vol. 28. Curran Associates, Inc., 2015. URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/1f36c15d6a3d18d52e8d493bc8187cb9-Paper.pdf.

[3] Massimiliano Amarante and Fabio Maccheroni. "When an Event Makes a Difference". In: *Theory and Decision* 60 (2006), pp. 119–126.

[4] Massimiliano Amarante, Fabio Maccheroni, Massimo Marinacci, and Luigi Montrucchio. "Cores of non-atomic market games". In: *International Journal of Game Theory* 34 (2006), pp. 399–424.

[5] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2005.

[6] Thomas Augustin. "Neyman-Pearson testing under interval probability by globally least favorable pairs: Reviewing Huber-Strassen theory and extending it to general interval probability". In: *Journal of Statistical Planning and Inference* 105.1 (2002). Imprecise Probability Models and their Applications, pp. 149–173. DOI: 10.1016/S0378-3758(01)00208-7.

[7] Thomas Augustin, Frank P. A. Coolen, Gert de Cooman, and Matthias C. M. Troffaes, eds. *Introduction to Imprecise Probabilities*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2014.

[8] Jose Blanchet and Yang Kang. "Sample Out-of-Sample Inference Based on Wasserstein Distance". In: *Operations Research* 69 (2021), pp. 985–1013.

[9] Jose Blanchet, Yang Kang, and Karthyek Murthy. "Robust Wasserstein profile inference and applications to machine learning". In: *Journal of Applied Probability* 56.3 (2019), pp. 830–857. DOI: 10.1017/jpr.2019.49.

[10] Yann Brenier. "Polar factorization and monotone rearrangement of vector-valued functions". In: *Communications on Pure and Applied Mathematics* 44.4 (1991), pp. 375–417.

[11] Michele Caprio. "Imprecise Markov Semigroups and their Ergodicity ". 2024. arXiv: 2405.00081.

[12] Michele Caprio and Ruobin Gong. "Dynamic precise and imprecise probability kinematics". In: *Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications*. Ed. by Enrique Miranda, Ignacio Montes, Erik Quaeghebeur, and Barbara Vantaggi. Vol. 215. Proceedings of Machine Learning Research. 2023, pp. 72–83.

[13] Michele Caprio and Sayan Mukherjee. "Ergodic theorems for dynamic imprecise probability kinematics". In: *International Journal of Approximate Reasoning* 152 (2023), pp. 325–343.

[14] Michele Caprio and Teddy Seidenfeld. "Constriction for sets of probabilities". In: *Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications*. Ed. by Enrique Miranda, Ignacio Montes, Erik Quaeghebeur, and Barbara Vantaggi. Vol. 215. Proceedings of Machine Learning Research. 2023, pp. 84–95.

[15] Michele Caprio, David Stutz, Shuo Li, and Arnaud Doucet. "Conformalized Credal Regions for Classification with Ambiguous Ground Truth". In: *Transactions on Machine Learning Research* (2025). URL: https://openreview.net/forum?id=L7sQ8CW2FY.

[16] Michele Caprio, Maryam Sultana, Eleni Elia, and Fabio Cuzzolin. “Credal Learning Theory”. In: *Proceedings of NeurIPS 2024* (2024).

[17] Michele Caprio and Souradeep Dutta and Kuk Jin Jang and Vivian Lin and Radoslav Ivanov and Oleg Sokolsky and Insup Lee. “Credal Bayesian Deep Learning”. In: *Transactions on Machine Learning Research* (2024). URL: https://openreview.net/forum?id=4NHF9AC5ui.

[18] Simone Cerreia-Vioglio, Fabio Maccheroni, and Massimo Marinacci. “Ergodic theorems for lower probabilities”. In: *Proceedings of the American Mathematical Society* 144 (2015), pp. 3381–3396.

[19] Siu Lun Chau, Antonin Schrab, Arthur Gretton, Dino Sejdinovic, and Krikamol Muandet. “Credal Two-Sample Tests of Epistemic Ignorance”. In: *The 28th International Conference on Artificial Intelligence and Statistics*. 2025.

[20] Yifan Chen and Wuchen Li. “Optimal transport natural gradient for statistical manifolds with continuous sample space”. In: *Information Geometry* 3 (2020), pp. 1–32.

[21] Jinjin Chi et al. “Approximate continuous optimal transport with copulas”. In: *International Journal of Intelligent Systems* 37.8 (2022), pp. 5354–5380. DOI: 10.1002/int.22795.

[22] Gustave Choquet. “Theory of capacities”. In: *Annales de l’Institut Fourier* 5 (1954), pp. 131–295.

[23] Roberto Corrao. *Mediation Markets: The Case of Soft Information*. 2023. URL: https://economics.mit.edu/sites/default/files/inline-files/Roberto_Corrao_JMP1_15.pdf.

[24] Roberto Corrao and Yifan Dai. *The Bounds of Mediated Communication*. 2023. arXiv: 2303.06244 [econ.TH].

[25] Roberto Corrao, Drew Fudenberg, and David Levine. *Risk, surprise, randomization, and adversarial forecasters*. 2023. URL: https://economics.mit.edu/sites/default/files/inline-files/Risk_and_surprise%20-%202023-07-21_online.pdf.

[26] Thierry Denœux and Lalla Meriem Zouhal. “Handling possibilistic labels in pattern classification using evidential reasoning”. In: *Fuzzy Sets and Systems* 122.3 (2001), pp. 409–424.

[27] S. Destercke, D. Dubois, and E. Chojnacki. “Unifying practical uncertainty representations – I: Generalized p-boxes”. In: *International Journal of Approximate Reasoning* 49.3 (2008), pp. 649–663.

[28] Peyman Mohajerin Esfahani and Daniel Kuhn. “Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations”. In: *Mathematical Programming* 171 (2018), pp. 115–166.

[29] Sorin G. Gal and Constantin P. Niculescu. “Kantorovich’s Mass Transport Problem For Capacities”. In: *Proceedings of the Romanian Academy Series A* 20 (2019), pp. 337–345.

[30] Alfred Galichon. *Optimal Transport Methods in Economics*. Princeton University Press, 2016.

[31] Rui Gao. “Finite-Sample Guarantees for Wasserstein Distributionally Robust Optimization: Breaking the Curse of Dimensionality”. In: *Operations Research* 71 (2022), pp. 2291–2306.

[32] Rui Gao, Liyan Xie, Yao Xie, and Huan Xu. “Robust Hypothesis Testing Using Wasserstein Uncertainty Sets”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/a08e32d2f9a8b78894d964ec7fd4172e-Paper.pdf.

[33] Ruobin Gong and Xiao-Li Meng. “Judicious judgment meets unsettling updating: dilation, sure loss, and Simpson’s paradox”. In: *Statistical Science* 36.2 (2021), pp. 169–190.

[34] Michel Grabisch. *Set Functions, Games and Capacities in Decision Making*. Springer, 2016. DOI: 10.1007/978-3-319-28702-3.

[35] Peter J. Huber and Volker Strassen. “Minimax Tests and the Neyman-Pearson Lemma for Capacities”. In: *The Annals of Statistics* 1.2 (1973), pp. 251–263. DOI: 10.1214/aos/1176342363.

[36] Eyke Hüllermeier and Willem Waegeman. “Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods”. In: *Machine Learning* 3.110 (2021), pp. 457–506.

[37] Ruiping Ji, Yan Liang, Linfeng Xu, and Zhenwei Wei. “Trajectory prediction of ballistic missiles using Gaussian process error model”. In: *Chinese Journal of Aeronautics* 35.1 (2022), pp. 458–469. DOI: 10.1016/j.cja.2021.05.011.

[38] Radim Jiroušek and Prakash P. Shenoy. “A new definition of entropy of belief functions in the Dempster–Shafer theory”. In: *International Journal of Approximate Reasoning* 92 (2018), pp. 49–65.

[39] Rabiul Hasan Kabir and Kooktae Lee. “Receding-Horizon Ergodic Exploration Planning using Optimal Transport Theory”. In: *2020 American Control Conference (ACC)*. 2020, pp. 1447–1452. DOI: 10.23919/ACC45564.2020.9147930.

[40] Lev Kantorovich. “On the Translocation of Masses”. In: *Comptes Rendus de l’Académie des Sciences de l’URSS* 37 (1942), pp. 199–201.

[41] Anton Kolotilin, Roberto Corrao, and Alexander Wolitzky. *Persuasion and Matching: Optimal Productive Transport*. 2023. arXiv: 2311.02889 [econ.TH].

[42] Isaac Levi. *The Enterprise of Knowledge*. London, UK : MIT Press, 1980.

[43] Vivian Lin et al. “DC4L: Distribution shift recovery via data-driven control for deep learning models”. In: *Proceedings of the 6th Annual Learning for Dynamics and Control Conference*. Ed. by Alessandro Abate, Mark Cannon, Kostas Margellos, and Antonis Papachristodoulou. Vol. 242. Proceedings of Machine Learning Research. 2024, pp. 1526–1538.

[44] Haiyan Liu, Bin Wang, Ruodu Wang, and Sheng Chao Zhuang. *Distorted optimal transport*. 2023. arXiv: 2308.11238 [math.OC].

[45] Xing Liu and François-Xavier Briol. “On the Robustness of Kernel Goodness-of-Fit Tests”. 2024. arXiv: 2408.05854.

[46] Artur O. Lopes and Jairo K. Mengue. “Duality Theorems in Ergodic Transport”. In: *Journal of Statistical Physics* 149 (2012), pp. 921–942.

[47] Silvia Lorenzini, Davide Petturiti, and Barbara Vantaggi. *Optimal Transport in Dempster-Shafer Theory and Choquet-Wasserstein Pseudo-Distances*. 2024. URL: https://ipmu2024.inesc-id.pt/files/paper_1160.pdf.

[48] Pengyuan Lu, Michele Caprio, Eric Eaton, and Insup Lee. “IBCL: Zero-shot Model Generation for Task Trade-offs in Continual Learning”. 2024. arXiv: 2305.14782.

[49] Massimo Marinacci and Luigi Montrucchio. “Introduction to the mathematics of ambiguity”. In: *Uncertainty in economic theory: a collection of essays in honor of David Schmeidler’s 65th birthday*. Ed. by Itzhak Gilboa. London : Routledge, 2004.

[50] Enrique Miranda and Ignacio Montes. “Centroids of the core of exact capacities: a comparative study”. In: *Annals of Operations Research* 321 (2023), pp. 409–449.

[51] Thomas Mortier, Viktor Bengs, Eyke Hüllermeier, Stijn Luca, and Willem Waegeman. “On the Calibration of Probabilistic Classifier Sets”. In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. Ed. by Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent. Vol. 206. Proceedings of Machine Learning Research. 2023, pp. 8857–8870.

[52] Roger B. Nelsen. *An Introduction to Copulas*. 2nd. Springer Series in Statistics. Springer, New York, 2006.

[53] Hung T. Nguyen. “A prelude to statistics arising from optimal transport theory”. In: *Asian Journal of Economics and Banking* 7.2 (July 2023), pp. 166–179. URL: https://ideas.repec.org/a/eme/ajebpp/ajeb-05-2023-0038.html.

[54] Georg Pflug and David Wozabal. “Ambiguity in portfolio selection”. In: *Quantitative Finance* 7.4 (2007), pp. 435–442.

[55] Georg Ch. Pflug and Alois Pichler. *Multistage Stochastic Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2014.

[56] S. T. Rachev and I. Olkin. “Mass transportation problems with capacity constraints”. In: *Journal of Applied Probability* 36.2 (1999), pp. 433–445. DOI: 10.1239/jap/1032374463.

[57] Svetlozar T. Rachev and Ludger Rüschendorf. *Mass Transportation Problems – Volume 1: Theory*. Probability and Its Applications. Springer New York, NY, 1998.

[58] Jenni Raitoharju. “Chapter 3 - Convolutional neural networks”. In: *Deep Learning for Robot Perception and Cognition*. Ed. by Alexandros Iosifidis and Anastasios Tefas. Academic Press, 2022, pp. 35–69.

[59] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. “On Wasserstein Two-Sample Testing and Related Families of Nonparametric Tests”. In: *Entropy* 19.2 (2017). DOI: 10.3390/e19020047.

[60] Walter Rudin. *Real and complex analysis*. 3rd. New york : McGraw-Hill, 1987.

[61] Yusuf Sale, Viktor Bengs, Michele Caprio, and Eyke Hüllermeier. “Second-Order Uncertainty Quantification: A Distance-Based Approach”. In: *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Ruslan Salakhutdinov et al. Vol. 235. Proceedings of Machine Learning Research. 2024, pp. 43060–43076.

[62] Yusuf Sale, Michele Caprio, and Eyke Hüllermeier. "Is the volume of a credal set a good measure for epistemic uncertainty?" In: *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*. Ed. by Robin J. Evans and Ilya Shpitser. Vol. 216. Proceedings of Machine Learning Research. 2023, pp. 1795–1804.

[63] Soroosh Shafieezadeh-Abadeh, Liviu Aolaritei, Florian Dörfler, and Daniel Kuhn. "New Perspectives on Regularization and Computation in Optimal Transport-Based Distributionally Robust Optimization". 2023. arXiv: 2303.03900.

[64] David Stutz and Abhijit Guha Roy and Tatiana Matejovicova and Patricia Strachan and Ali Taylan Cemgil and Arnaud Doucet. "Conformal prediction under ambiguous ground truth". In: *Transactions on Machine Learning Research* (2023). URL: https://openreview.net/forum?id=CAd6V2qXxc.

[65] Bahar Taşkesen, Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Karthik Natarajan. "Discrete Optimal Transport with Independent Marginals is #P-Hard". In: *SIAM Journal on Optimization* 33.2 (2023), pp. 589–614.

[66] Matthew Thorpe. *Introduction to Optimal Transport*. F2.08, Centre for Mathematical Sciences: University of Cambridge, 2018.

[67] Vicenç Torra. "The transport problem for non-additive measures". In: *European Journal of Operational Research* 311.2 (2023), pp. 679–689. DOI: 10.1016/j.ejor.2023.03.016.

[68] Matthias C.M. Troffaes and Gert de Cooman. *Lower Previsions*. Chichester, United Kingdom : John Wiley and Sons, 2014.

[69] Cedric Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics 58. American Mathematical Society, Providence, Rhode Island, 2003.

[70] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Vol. 42. Monographs on Statistics and Applied Probability. London : Chapman and Hall, 1991.

[71] Larry A. Wasserman and Joseph B. Kadane. "Bayes' Theorem for Choquet Capacities". In: *The Annals of Statistics* 18.3 (1990), pp. 1328–1339.

[72] Qien Yu, Chen Li, Ye Zhu, and Takio Kurita. "Convolutional autoencoder based on latent subspace projection for anomaly detection". In: *Methods* 214 (2023), pp. 48–59.

[73] Marco Zaffalon. "The naive credal classifier". In: *Journal of Statistical Planning and Inference* 105.1 (2002), pp. 5–21.