Maximum Likelihood Probability Measures over Sets and Applications to Data-Driven Optimization

Juan S. Borrero[1] and Denis Sauré[2]

[1]School of Industrial Engineering and Management, Oklahoma State University
[2]Department of Industrial Engineering, University of Chile

May 15, 2023

Abstract

Motivated by data-driven approaches to sequential decision-making under uncertainty, we study maximum likelihood estimation of a distribution over a general measurable space when, unlike traditional setups, realizations of the underlying uncertainty are not directly observable but instead are known to lie within observable sets. While extant work studied the special cases when the observed sets corresponded to intervals in $\mathbb{R}^n$ for $n = 1, 2$, our work provides, to the best of our knowledge, a first rigorous treatment of the more general estimation problem. Our results show that maximum likelihood estimates concentrate on a collection of maximal intersections (CMI) sets, and can be found by solving a convex optimization problem whose size is linear in the size of the CMI. After studying the efficient computation of the CMI and the maximum likelihood estimate, we characterize convergence properties of the maximum likelihood estimate and apply our results to construct ambiguity sets and develop compact formulations for Distributionally Robust and Greedy and Optimistic Optimization. Our results show how non-parametric maximum likelihood estimation can be incorporated effectively into data-driven optimization problems, resulting in tractable formulations that are tested numerically.

1 Introduction

Motivation. We are motivated by settings of sequential decision-making under uncertainty, which typically model uncertainty as arising from the realization of a sequence of independent and identically distributed (iid) random elements $\{\boldsymbol{c}^s : s \in \mathbb{Z}_+\}$, where $\boldsymbol{c}^s$ represents the uncertainty affecting state dynamics during period $s \in \mathbb{Z}_+$. Traditionally, the literature has assumed that the underlying (common) distribution μ^0 of $\boldsymbol{c}^s$ is known to the decision-maker (DM). However, in the last decades, settings where such a distribution is initially unknown have attracted considerable attention [4, 12, 38, 50]. In such work, the DM typically receives some periodic feedback on $\boldsymbol{c}^s$ which allows her to refine her knowledge of μ^0 and thus improve the decision-making process. In particular, in settings where $\boldsymbol{c}^s$ is observed upon its realization, maximum likelihood estimation (MLE) [26, 28, 57] arises as a possible method for parametric/non-parametric inference of μ^0.

Departing from the models above, in this work we study the estimation of μ^0 when $\boldsymbol{c}^s$ is not observed directly but is known to lie within a set C^s. Consider, for example, settings of sequential interdiction where at each period $s \in \mathbb{Z}_+$ the DM implements an action x^s aimed at maximizing (in expectation) some function that depends, among other things, on an adversarial (random) response y^s, which in turn aims at minimizing some function $g(\cdot)$, so that

$$y^s(\omega) \in \arg\min \{g(y, x, \boldsymbol{c}^s(\omega)) : \ y \in Y(x)\}.$$

Depending on the application of interest, the feedback observed in period s might consist, for example, only on the response y^s, which informs rather indirectly on the realization of $\boldsymbol{c}^s$. In such a case, upon observing y^s the DM infers that $\boldsymbol{c}^s \in C^s := \{\boldsymbol{c} : y^s \in \arg\min \{g(y, x^s, \boldsymbol{c})\} : y \in Y(x^s)\}$, and thus any estimate of the distribution μ^0 must be constructed solely based on the information contained in the sequence $\{C^s : s \in \mathbb{Z}_+\}$. Settings of sequential interdiction under epistemic uncertainty have been studied recently [18–20, 61], however, they consider deterministic feedback, ignoring the relevant stochastic alternative.

A case related to the setting above comes from the inverse optimization literature [2]. Consider a DM that observes a directed network $G = (N, A)$ and seeks to estimate the distribution μ^0 of the cost vector during period s, $\boldsymbol{c}^s := (c^s_a : a \in A)$, where c^s_a is the unitary cost of moving flow in arc $a \in A$ during period $s \in \mathbb{Z}_+$. We assume it is known that the vectors $\{\boldsymbol{c}^s : s \in \mathbb{Z}_+\}$ are iid. On each period s, a user (different from the DM) observes $\boldsymbol{c}^s$ and selects the shortest path between some fixed pair of nodes assuming that the costs are given by $\boldsymbol{c}^s$. Supposing that the DM only observes some subset of the shortest path y^s traversed by the user, then the feedback obtained by the DM in period $s \in \mathbb{Z}_+$ informs that $\boldsymbol{c}^s \in C^s := \{c \in \mathbb{R}^{|A|}_+ :$ there exist a shortest path of G that contains $y^s\}$. (If more information is available, for instance, the cost of the shortest path, the sets C^s can be modified accordingly.) In this setting, the DM must estimate μ^0 using only the information contained in the sequence $\{C^s : s \in \mathbb{Z}_+\}$.

Beyond the estimation problem, decision-making under uncertainty when the underlying distribution is not known is often tackled using the robust optimization [7–9, 11, 23, 27, 59] or optimism in the face of uncertainty [5, 6, 18, 21] paradigms. A common input to these models is an *uncertainty* or *ambiguity set* upon which the DM looks for either a worst-case or best-case realization, for any given decision. In this regard, important questions that arise in these models are how to compute such ambiguity sets in a manner that is consistent with MLE, and how to solve the resulting formulations.

Objective and assumptions. Departing from the traditional setups, we study settings where the DM looks for a non-parametric estimator of μ^0 when the sequence $\{\boldsymbol{c}^s : s \in \mathbb{Z}_+\}$ is not observed but instead a sequence $\{C^s, s \in \mathbb{Z}_+\}$ is, where it is known that $\boldsymbol{c}^s \in C^s$, $s \in \mathbb{Z}_+$. When the $\boldsymbol{c}^s$s are not observed directly, as in our motivating examples, the standard MLE method cannot be (directly) employed to estimate μ^0. However, the techniques behind MLE can be extended in order to handle

this case. The overall goals of this paper are thus to (*i*) formally extend the method in order to estimate an underlying distribution where only sets containing uncertainty realizations are observed on each period; (*ii*) analyze the convergence properties of the resulting estimate; and (*iii*) study how to incorporate the resulting estimates into data-driven optimization approaches. We make no specific assumptions about μ^0, nor the space where the elements $\boldsymbol{c}^s$, $s \in \mathbb{Z}_+$, take their values, nor the forms of the sets C^s (beyond their measurability), $s \in \mathbb{Z}_+$, and use a non-parametric approach. The problem discussed in this paper is not entirely new: parametric MLE with censored data is commonplace in economics, and specific settings of non-parametric MLE with censored data give rise to specific uncertainty sets in $\mathbb{R}$ and $\mathbb{R}^2$ [31, 32, 43, 54]. However, to the best of our knowledge, there is no general and rigorous treatment of this non-parametric estimation problem in general probability spaces nor of its use in the context of data-driven optimization.

Results and Contribution. Our first contribution amounts to showing that there is a maximum likelihood probability measure (MLPM) over the σ-algebra generated by the class of sets $\mathcal{C}^t := \{C^s \colon s \in [t]\}$, where $[t] := \{1, 2, \ldots, t\}$ for each $t \in \mathbb{Z}_+$, and that this measure is concentrated in the *collection of maximal intersections* (CMI) generated by the class $\mathcal{C}^t$. While said collection might be of exponential size in the worst-case, we illustrate its efficient computation in practical instances. Related to this, we show that the MLPM can be computed by solving a convex optimization problem whose size is linear in the size of the CMI, and use the local optimality guarantees of such problem to propose a column generation procedure that avoids computing all elements of the CMI, whose efficiency we test numerically. We also show how MLPMs can be extended to larger σ-algebras that contain $\mathcal{C}^t$, using a reference measure.

A second contribution pertains to the analysis of the convergence of the MLPM. In this regard, we first show that in general the convergence of MLPMs to μ^0 cannot be guaranteed under various standard convergence modes. We then introduce a convergence notion in terms of the Wasserstein distance [34] between distributions, and extend existing measure concentration results for this case, under additional assumptions.

Finally, on a more practical side, a third contribution amounts to showing how the theoretical results for MLPMs can be used to construct ambiguity sets for data-driven optimization problems, where the DM looks for distributions that maximize the likelihood of the observed sequence of sets. Leveraging recent results for the case with complete information [27, 29], we design particular Wasserstein 'balls', formulate a Distributionally Robust Optimization (DRO) problem that provides out-of-sample guarantees (and obtain a convex reformulation of the DRO problem that extends similar formulations in the literature) and a Greedy and Optimistic Optimization (GOO) problem used in learning approaches to bilevel optimization [19]. We test the practical performance of the resulting formulations, as well as the complexity of computing the CMI and MLPM in a series of numerical experiments.

Organization of the paper. The rest of the paper is organized as follows. In Section 2 we review

the literature and articulate the novelty of our work. In Section 3 and 4 we study the theoretical and computational aspects of the estimation problem, respectively. In Section 5 we study convergence issues, and in Section 6 we review applications of our results to data-driven optimization. In Section 7 we measure the performance of the proposed algorithms through numerical experiments. Finally, Section 8 presents our conclusions. The proofs that are not in the main body of the paper are relegated to Appendix A.

Notation and Assumptions. Throughout the manuscript, we consider an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and assume that the C^ss belong to some measurable space $(\Psi, \mathcal{G})$, i.e. $C^s \in \mathcal{G}$, $s \in \mathbb{Z}_+$. Typically, $(\Psi, \mathcal{G}) = (\mathbb{R}^n, \mathbb{B}^n)$ where $\mathbb{B}^n$ denotes the Borel σ-algebra of $\mathbb{R}^n$, i.e., the σ-algebra generated by the usual open sets in $\mathbb{R}^n$ (in which case, $\boldsymbol{c}^s$ is a random vector, $s \in \mathbb{Z}_+$). We define the support of a vector $x \in \mathbb{R}^n$ as $\operatorname{supp}(x) := \{j \leq n : x_j \neq 0\}$. Given a measurable space $(\Psi, \mathcal{G})$ we write $\mu \in (\Psi, \mathcal{G})$ to indicate that $(\Psi, \mathcal{G}, \mu)$ is proper probability space. We say a probability measure μ^* maximizes the likelihood of (observing) the class $\mathcal{C}^t$ if and only if μ^* is a solution of the optimization problem

$$\sup\left\{L(\mu, t) \colon \mu \in \big(\Psi, \mathcal{G}\big)\right\}, \quad \text{where } L(\mu, t) = \prod_{s=1}^{t} \mu(C^s), \qquad \mu \in \big(\Psi, \mathcal{G}\big). \tag{1}$$

When it is well-defined, we say that the measure μ^* is an MLPM over $\big(\Psi, \mathcal{G}\big)$. For a given probability measure μ, $\mathbb{E}^\mu$ denotes the expectation operation with respect to such a measure. In addition, if $(\Psi, \mathcal{G})$ and $(\Psi', \mathcal{G}')$ are such that $\Psi \subseteq \Psi'$ and $\mathcal{G} \subseteq \mathcal{G}'$, then for any measure $\mu \in (\Psi', \mathcal{G}')$, we let $\mu_{|\mathcal{G}}$ denote the restriction of μ to $\mathcal{G}$.

For a given pair of collections of sets $\mathcal{B}$ and $\mathcal{D}$, we let $\mathcal{B} \setminus \mathcal{D}$ denote the collection of sets of the form $B \setminus D$ where $B \in \mathcal{B}$ and $D \in \mathcal{D}$. Also, we let $\sigma(\mathcal{B})$ denote the σ-algebra generated by $\mathcal{B}$, and $\mathcal{A}(\mathcal{B})$ denote the class of *atoms* of $\mathcal{B}$, which are sets with the property that they cannot be 'divided' further in terms of other sets of the class (see [48] p. 24), i.e.

$$\mathcal{A}(\mathcal{B}) = \big\{B \in \mathcal{B} \colon B \neq \emptyset \text{ and if } B' \subseteq B, \text{ with } B' \in \mathcal{B}, \text{ then } B' = B\big\}.$$

2 Literature review

MLE is a standard estimation method in statistics [28]. The basic setup involves a sequence of iid observations from an unknown probability distribution. A *parametric* approach to MLE assumes that the functional form of the underlying distribution is known but its parameters are not. In this case, MLE uses the observations to estimate parameters across a family of parametric probability distributions; under mild assumptions, MLE is known to *recover* the underlying true parameters asymptotically, i.e., the maximum likelihood (ML) estimate is consistent [26, 57]. During the last decades, these known properties of MLE have been extended to progressively more general settings, see [35, 42, 51, 58] and the references therein.

A non-parametric approach to MLE involves explicitly estimating the probability density function of the underlying distribution, instead of just parameters. In this regard, while there exists fairly general approaches, such as [58], there are also other computationally-driven methods such as kernels [41, 49], MLE with penalties [22], the method of sieves [30], stochastic programming [25], among others, see [1] and references therein. Notably, in this setting, convergence to the true distribution, i.e., consistency, is not always attained across all methods. In this more general non-parametric setting, when specific parameters or densities are not of interest, the ML estimate of the underlying distribution is the empirical distribution associated with the sample. In this case, the Glivenko-Cantelli Theorem and its generalizations (see e.g., [37] p. 20, [55], and [52] p. 828) show that the cumulative distribution function (cdf) induced by the empirical distribution converges uniformly and almost surely (a.s.) to the true cdf.

The work above assumes that the realizations of the random elements are observed with precision. In cases where there is *incomplete information* about the observations, non-parametric MLE can still be carried out. In [24, 36] the authors propose a model for incomplete information that can be seen as assuming 'two-layers' of variability: one due to the random vectors $\boldsymbol{c}^s$, $s \in \mathbb{Z}_+$, and a second one due to another sequence of random vectors that depend on the respective realization of each $\boldsymbol{c}^s$, whose output provides the data that is observed. In this model it is assumed that the conditional information of the observed data given the unobserved data is known. This model is fairly general and captures many situations arising in practice. For parametric cases, the Expectation-Maximization (EM) algorithm [24] can be employed to get the ML estimate of the parameters. The non-parametric case is studied in [36], which derives several properties of the non-parametric ML estimator, particularly *self-consistency*; it also derives an algorithm, based on the EM algorithm, in order to compute the estimate. The model studied in [24, 36], however, cannot capture settings where the observed data are sets rather than vectors, as assumed in our work. Moreover, even if the sets can be reconstructed from observed vectors, the generality of the model obscures many interesting specific properties that can be derived for our setting and becomes cumbersome to use.

An alternative model for incomplete information, which we adopt here, assumes that the vectors are not observed directly. Rather, a non-empty set C^s containing the realization of $\boldsymbol{c}^s$, $s \in \mathbb{Z}+$, is observed. This model was, to the best of our knowledge, studied first in [43], where $\boldsymbol{c}^s \in \mathbb{R}$ and the C^ss are intervals. There, the authors argue that the non-parametric estimate of the cdf should be concentrated in certain intersecting points of the intervals. In [54] these findings are expanded and the authors derive an iterative algorithm to compute the ML estimate. Multivariate extensions centered in specific applications with censored data are discussed in [32], which explicitly discusses the idea of maximal intersections in multiple dimensions and derives an explicit formula for the ML estimate under certain specific assumptions when $\boldsymbol{c}^s \in \mathbb{R}^2$. In [31] the authors study again the interval model, use the Karush-Kuhn-Tucker (KKT) conditions of a nonlinear optimization problem to compute the ML estimate, and prove the consistency of the estimator under certain assumptions. Consistency results for another specific model in $\mathbb{R}^2$ are also studied in [60], while [13] uses the KKT conditions to compute ML estimate in specific applications. In [33] the authors study the

interval setting and show that the maximal intersections can be computed by finding the maximal cliques in the intersection graph generated by the intervals. More specific models in $\mathbb{R}^2$ involving ML estimation for censored failure data are studied in [45, 46] and references therein.

Although the work discussed above studies specific instances of the problem under consideration, to the best of our knowledge there is no existing literature studying this problem in general measurable (or Euclidian) spaces, which is the objective of our work. Moreover, we are not aware of studies of these ideas beyond purely estimation problems; our work goes further and explores their use in data-driven optimization approaches.

3 Properties and construction of MLPM

Throughout this section, we consider $t \in \mathbb{Z}_+$ fixed. After introducing background material, we define the collection of maximal intersections associated with $\mathcal{C}^t$, and show that MLPMs must necessarily concentrate on such a collection. These results are then leveraged then to compute an MLPM.

Preliminaries. Let $(\Psi, \mathcal{G})$ be a measurable space and consider a sequence of iid random elements $\boldsymbol{c}^s \colon \Omega \to \Psi$, $s \in \mathbb{Z}_+$, following an unknown *probability distribution* μ^0, (by distribution we mean that μ^0 is defined by $\mu^0(G) = \mathbb{P}[\boldsymbol{c}^s \in G]$ for any $s \in \mathbb{Z}_+$ and $G \in \mathcal{G}$, see [48] p. 137). Recall that $\mathcal{C}^t = \{C^s \colon s \in [t]\}$ corresponds to the sets observed by the DM until t; we define the measurable space $\left(\Psi^t, \mathcal{G}^t\right)$, as follows

$$\Psi^t := \bigcup_{s \in [t]} C^s \qquad \text{and} \qquad \mathcal{G}^t := \sigma(\mathcal{C}^t).$$

Note that $\{\left(\Psi^t, \mathcal{G}^t\right), t \in \mathbb{Z}_+\}$ is an increasing sequence of measurable spaces (i.e., $\Psi^s \subseteq \Psi^t$ and $\mathcal{G}^s \subseteq \mathcal{G}^t$ for any $s \leq t$) and that all of them are contained in the original one, i.e., $\Psi^t \subseteq \Psi$ and $\mathcal{G}^t \subseteq \mathcal{G}$ for any $t \in \mathbb{Z}_+$.

We consider a DM who is interested in estimating the distribution μ^0 at time t based solely on the observation of the collection $\mathcal{C}^t$, and the knowledge that $\boldsymbol{c}^s \in C^s$ for any $s \in \mathbb{Z}_+$. (Note that the DM does not observe the $\boldsymbol{c}^s$s directly.) In particular, we assume that the DM seeks to estimate μ^0 by finding a distribution that maximizes the likelihood of observing $\mathcal{C}^t$ over $\left(\Psi^t, \mathcal{G}^t\right)$. An example of this setting in the context of network interdiction is presented next.

Example 1. Consider an interdictor that observes a smuggler who, at each period $s \in \mathbb{Z}_+$, traverses a path between nodes 1 and n of a network $G = (N, A)$, where N denotes the set of nodes, and A the set of arcs. Let $\boldsymbol{c}^s \in \mathbb{R}_+^{|A|}$ denote the cost vector of the network at time $s \in \mathbb{Z}_+$, and assume that the sequence $\{\boldsymbol{c}^s \colon s \in \mathbb{Z}_+\}$ is iid and follows a (absolutely continuous) distribution μ^0. At time $s \in \mathbb{Z}_+$, the interdictor blocks travel through a set of arcs $B^s \subseteq A$, and then the smuggler observes the vector $\boldsymbol{c}^s$ and moves through a shortest path P^s between nodes 1 and n in the interdicted network $G^s \equiv (A, N \setminus B^s)$. The interdictor does not observe $\boldsymbol{c}^s$ directly, but rather observes the path P^s used by the smuggler, for $s \in [t]$, and uses this information to estimate the distribution

μ^0. In this case, one has that

$$C^s = \Big\{ \boldsymbol{c} \in \mathbb{R}_+^{|A|} : P^s \in \arg\min \Big\{ \sum_{a \in P} c_a : P \text{ is a } 1-n \text{ path in } G^s \Big\} \Big\}, \quad s \in \mathbb{Z}_+. \qquad \blacksquare$$

Remark 1 shows that finding an MLPM is straightforward when the sets in $\mathcal{C}^t$ are disjoint, or when their (overall) intersection is non-empty.

Remark 1. *If the C^ss are mutually disjoint, then any MLPM μ^* is such that $\mu^*(C^s) = 1/t$ for all $s \in [t]$, so that $L(\mu^*, t) = (1/t)^t$. In particular, the empirical distribution is an MLPM. At the other extreme, if $\bigcap_{s\in[t]} C^s =: C \neq \emptyset$ then any MLPM μ^* is such that $\mu^*(C) = 1$, so that $L(\mu^*, t) = 1$.*

In general, finding MLPM is not as direct as in Remark 1. In order to characterize their support, next we introduce the concept of *maximal intersections*.

3.1 Measures over the collection of maximal intersections (CMI)

For $S \subseteq [t]$, define $I(S) := \bigcap_{s\in S} C^s$. Hereafter we refer to these sets as *C-intersections*. Let m^t be the size of the largest set associated with a non-empty C-intersection, i.e.

$$m^t := \max \{ |S| : S \subseteq [t],\, I(S) \neq \emptyset \}.$$

Note that, by definition, $I(S) = \emptyset$ for any S such that $|S| > m^t$. For $k \leq m^t$, define (recursively)

$$\mathcal{I}_k^t := \big\{ S \subseteq [t] \colon |S| = k,\, I(S) \neq \emptyset,\, S \not\subset S',\, S' \in \mathcal{I}_l^t,\, l > k \big\}, \quad \mathcal{I}_{m+1}^t = \emptyset.$$

The collection $\mathcal{I}_k^t$ consists of all subsets of periods of size k whose corresponding C-intersections are non-empty and that are not contained in any other subset of time periods that give a non-empty C-intersection. We define the *collection of maximal intersections* (CMI) as follows

$$\mathcal{M}^t := \big\{ I(S) :\, S \in \mathcal{I}_k^t,\, k \leq m^t \big\}. \tag{2}$$

The 'maximal' denomination refers to the fact that if $I(S) \in \mathcal{M}^t$, then there is no larger set S' (i.e. such that $S \subseteq S'$) such that $I(S') \in \mathcal{M}^t$. Note that $\mathcal{M}^t$ can be exponential in t in the worst case (we study the computation of $\mathcal{M}^t$ in Section 4.) However, we show later that in practice not all elements in $\mathcal{M}^t$ may need to be computed to find an MLPM.

The significance of the CMI for MLPMs is, to a large extent, due to the following property:

Lemma 1. *If $I(S) \in \mathcal{M}^t$ for $S \subseteq [t]$, then $I(S)$ is an atom of $\mathcal{G}^t$.*

Specifically, Lemma 1 allows us to construct well-defined measures over $\big(\Psi^t, \mathcal{G}^t\big)$ using as ingredients a vector of non-negative weights w and the CMI, as shown next in Lemma 2. Moreover, in the next section, we use Lemmas 1 and 2 to prove that the MLPMs are concentrated on the CMI, and thus are equivalent up to the measure assigned to the CMI.

Lemma 2. *Let $w = (w_M \colon M \in \mathcal{M}^t)$ be such that $\sum_{M\in\mathcal{M}^t} w_M = 1$ and $w_M \geq 0$, $M \in \mathcal{M}^t$. Then,*

$$P^{\boldsymbol{w}}(G) := \sum_{M\subseteq G,\, M\in\mathcal{M}^t} w_M, \qquad G \in \mathcal{G}^t \tag{3}$$

is a well-defined probability measure on $(\Psi^t, \mathcal{G}^t)$.

We close this section by noting that that the measures constructed in Lemma 2 are concentrated on $\mathcal{M}^t$ because $P^{\boldsymbol{w}}(\mathcal{M}^t) \coloneqq P^{\boldsymbol{w}}\Big(\bigcup_{M\in\mathcal{M}^t} M\Big) = 1$.

3.2 MLPMs are concentrated on the CMI.

For $S \subseteq [t]$ define $X(S)$ as the set of elements in $I(S)$ that do not belong to any other element of $\mathcal{C}^t$, i.e.

$$X(S) := I(S) \setminus \bigcup_{\ell\notin S} I(S\cup\{\ell\}) \quad S \subseteq [t],$$

where null unions are defined as the empty set. We prove the concentration result by showing that if there is a measure $\mu(\cdot)$ that is not concentrated on $\mathcal{M}^t$, then it necessarily assigns a positive measure to some non-empty set $X(S) \notin \mathcal{M}^t$. In such a case, one can construct an alternative measure with a higher likelihood by simply reassigning such a measure to $I(S) \setminus X(S)$.

Theorem 1. *Let μ be a probability measure over $\left(\Psi^t, \mathcal{G}^t\right)$ such that $\mu(\cup_{M\in\mathcal{M}^t} M) < 1$, then there exist another probability measure $\hat{\mu}$ over $\left(\Psi^t, \mathcal{G}^t\right)$ such that $L(\hat{\mu}, t) > L(\mu, t)$.*

Proof. We assume without loss of generality that $L(\mu, t) > 0$ (because, by Lemma (2), one can always construct a measure μ with $L(\mu, t) > 0$ by assigning equal probability to all elements in $\mathcal{M}^t$). Note that $X(S) = I(S)$ for all $S \subseteq [t]$ such that $I(S) \in \mathcal{M}^t$, thus if $\mu(\cup_{M\in\mathcal{M}^t} M) < 1$, then it is necessarily the case that $\mu(X(S')) > 0$ for some $S' \subseteq [t]$ for which $I(S') \notin \mathcal{M}^t$ and $X(S') \neq \emptyset$. We consider two distinct cases

Case 1: $\mu(X(S')) < \mu(I(S'))$. Define the (measurable) function f as

$$f(x) = \begin{cases} 1, & \text{if } x \notin I(S') \\ 0, & \text{if } x \in X(S') \\ \frac{\mu(I(S'))}{\mu(I(S'))-\mu(X(S'))}, & \text{if } x \in I(S') \setminus X(S'), \end{cases}$$

and define $\hat{\mu}(G) \equiv \int_G f \, d\mu$ for each $G \in \mathcal{G}^t$. In other words, $\hat{\mu}$ is given by

$$\hat{\mu}(G) = \mu(G \setminus I(S')) + \frac{\mu(I(S'))}{\mu(I(S')) - \mu(X(S'))}\mu(G \cap (I(S') \setminus X(S'))), \quad G \in \mathcal{G}^t.$$

Note that $\hat{\mu}(\cdot)$ reassigns the measure assigned to $X(S')$ uniformly to $I(S') \setminus X(S')$. Observe that

$\mu'(\cdot)$ is well-defined because f is measurable and $\int_{\Psi^t} f\, d\mu = 1$. A key observation about $\hat{\mu}(\cdot)$ is

$$\hat{\mu}(C^s) = \mu(C^s),\ s \in [t] \text{ s.t. } C^s \cap I(S') = \emptyset, \quad \text{and} \quad \hat{\mu}(C^s) \geq \mu(C^s),\ s \in [t] \text{ s.t. } C^s \cap I(S') \neq \emptyset,$$

with strict inequality for at least one $s \in [t]$. This last statement follows because $\mu(I(S')) > \mu(X(S'))$ implies that $\mu(I(S' \cup \{s'\})) > 0$ for some $s' \notin S'$ such that $C^{s'} \cap X(S') = \emptyset$, thus $\mu(C^{s'} \cap (I(S') \setminus X(S'))) > 0$ and therefore

$$\begin{aligned}
\hat{\mu}(C^{s'}) &= \mu(C^{s'} \setminus I(S')) + \frac{\mu(I(S'))}{\mu(I(S')) - \mu(X(S'))} \mu(C^{s'} \cap (I(S') \setminus X(S'))) \\
&> \mu(C^{s'} \setminus I(S')) + \mu(C^{s'} \cap (I(S') \setminus X(S'))) \\
&= \mu(C^{s'} \setminus I(S')) + \mu(C^{s'} \cap I(S')) = \mu(C^{s'}).
\end{aligned}$$

Using the above, we have that

$$L(\mu, t) = \prod_{s=1}^{t} \mu(C^s) < \prod_{s=1}^{t} \hat{\mu}(C^s) = L(\hat{\mu}, t),$$

where in the above we have used the fact that $L(\mu, t) > 0$ implies that $\mu(C^s) > 0$ for $s \in [t]$.

Case 2: $\mu(X(S')) = \mu(I(S'))$. In this case, we have that $\mu(I(S') \setminus X(S')) = 0$, and thus there exists $M \in \mathcal{M}^t$ such that $M \subset I(S') \setminus X(S')$ and $\mu(M) = 0$. Consider an alternative measure $\hat{\mu}(\cdot)$ that reassigns the measure assigned to $X(S')$ to M, i.e.

$$\hat{\mu}(G) = \mu(G \setminus I(S')) + \mu(X(S')) \mathbf{1}\{M \subseteq G\}, \quad G \in \mathcal{G}^t.$$

It is readily verified that $\hat{\mu}$ satisfies the axioms of being a probability measure on $(\Psi^t, \mathcal{G}^t)$. In addition $\hat{\mu}(C^s) = \mu(C^s)$ for $s \in [t]$ such that $C^s \cap I(S') = \emptyset$. If $C^s \cap I(S') \neq \emptyset$, then it must be the case that $M \subseteq C^s$ and thus $\hat{\mu}(C^s) \geq \mu(C^s)$. Particularly, let $M = I(\hat{S})$ and note that $S' \subset \hat{S}$. Thus, there exist at least one $s' \in \hat{S} \setminus S'$, such that $C^{s'} \cap X(S') = \emptyset$ and thus $\mu(C^{s'} \cap I(S')) = 0$. This implies that $\hat{\mu}(C^{s'}) > \mu(C^{s'})$, and the result follows. □

Theorem 1 provides a necessary condition for a measure to be an MLPM over $\left(\Psi^t, \mathcal{G}^t\right)$: it needs to be concentrated on the CMI. Although this condition is not sufficient to guarantee that a measure is an MLPM, it can be exploited to streamline the search for an MLPM, as we show next.

3.3 Finding the weights of the MLPM

We begin reviewing an auxiliary result, which provides a simple formula to compute $L(\mu, t)$ for any measure μ that is concentrated in the CMI. The formula is valid for any measurable space containing $\left(\Psi^t, \mathcal{G}^t\right)$.

Lemma 3. *Let $(\Psi', \mathcal{G}', \mu)$ be a measure space such that $\Psi^t \subseteq \Psi'$ and $\mathcal{G}^t \subseteq \mathcal{G}'$. If $\mu(\cup_{M \in \mathcal{M}^t} M) = 1$*

then

$$\mu(C^s) = \sum_{M \in \mathcal{M}^t, M \subseteq C^s} \mu(M), \qquad s \in [t].$$

From the previous section, Theorem 1 implies that in order to find an MLPM, it is sufficient to search across those measures that concentrate on the CMI. Based on (3) and Lemma 3, the following result formulates the problem of finding an MLPM.

Theorem 2. *Consider the following optimization problem over the variables* $(w_M \colon M \in \mathcal{M}^t)$,

$$\max \sum_{s \in [t]} \ln\Big(\sum_{M \subseteq C^s,\, M \in \mathcal{M}^t} w_M \Big) \tag{4a}$$

$$s.t \sum_{M \in \mathcal{M}^t} w_M = 1 \tag{4b}$$

$$w_M \geq 0, \ \forall M \in \mathcal{M}^t. \tag{4c}$$

Then, an optimal solution of (4) *exists. Let* $\hat{\boldsymbol{w}}^t = (\hat{w}^t_M \colon M \in \mathcal{M}^t)$ *denote one such an optimal solution, define* $\hat{\mu}^t \equiv P^{\hat{\boldsymbol{w}}^t}$ *as in* (3). *Then* $\hat{\mu}^t$ *is an MLPM over* $\big(\Psi^t, \mathcal{G}^t\big)$.

Proof. The first part follows because the objective function in (4) is concave with respect to $\boldsymbol{w} = (w_M \colon M \in \mathcal{M}^t)$, $w \geq 0$, therefore Problem (4) is a concave maximization problem over a compact convex set, which is (polynomially) solvable (up to a precision factor) in the number of elements of $\mathcal{M}^t$ (for example, by using gradient descent). The remaining proof follows from Theorem 1 and Lemma 3, and from the fact that Problem (4) is equivalent to

$$\begin{aligned} \max \prod_{s \in [t]} \Big(\sum_{M \subseteq C^s, M \in \mathcal{M}^t} w_M \Big) \\ \text{s.t} \sum_{M \in \mathcal{M}^t} w_M = 1 \\ w_M \geq 0, \quad M \in \mathcal{M}^t. \end{aligned}$$

This observation concludes the proof. □

Formulation (4) considers the maximization of a concave function over the simplex in $\mathcal{M}^t$. Thus, while a closed-form solution is not available in general, the KKT conditions can be used to characterize its optimal solution. In particular, the optimal solution of formulation (4) can be found by finding a feasible solution to the following equations:

$$\sum_{s \in [t]} \frac{1_{\{M \subseteq C^s\}}}{\sum_{M' \in \mathcal{M}^t, M' \subseteq C^s} w_{M'}} + \lambda_M + \lambda = 0, \qquad M \in \mathcal{M}^t \tag{5a}$$

$$\sum_{M \in \mathcal{M}^t} w_M = 1 \tag{5b}$$

$$\lambda_M \, w_M = 0, \qquad M \in \mathcal{M}^t \tag{5c}$$

$$w_M, \lambda_M \geq 0, \qquad M \in \mathcal{M}^t. \tag{5d}$$

Remark 2. *Theorem 2 implies that a measure that solves problem* (1) *always exist for* $(\Psi, \mathcal{G}) = (\Psi^t, \mathcal{G}^t)$. *Thus, in this case, we can replace* sup *by* max *in* (1).

In Section 4, we propose a column generation scheme that attempts to solve for an MLPM without having to generate all components in the CMI: the potential efficiency of the procedure rests in the following result; here, we define $\text{supp}(\boldsymbol{w}) = \{M \in \mathcal{M}^t \colon w_M > 0\}$.

Proposition 1. *There always exists a solution* $\boldsymbol{w}$ *to* (4) *such that* $|supp(\boldsymbol{w})| \leq t + 1$.

3.4 Extension of MLPMs to larger σ-algebras.

Next, we show that the results shown so far for $\mathcal{G}^t$ remain valid over any σ-algebra $\mathcal{G}'$ that contains $\mathcal{G}^t$, as long as one can construct measures on $\mathcal{G}'$ such that an MLPM over $(\Psi^t, \mathcal{G}^t)$ is *absolutely continuous* with respect to the restriction of the measure on $\mathcal{G}^t$ (recall that given two probability measures μ and λ over a measurable space $(\Psi, \mathcal{G})$, λ is absolutely continuous with respect to μ, written $\lambda << \mu$, if and only if $\mu(A) = 0$ implies $\lambda(A) = 0$, $A \in \mathcal{G}$, see [48] pg. 333). This result shows, for instance, that an MLPM can be computed when the random vectors $\boldsymbol{c}^s$, $s \in [t]$, are Borel-measurable (which is a common underlying assumption in many probabilistic models in practice). Hereafter $(\Psi', \mathcal{G}')$ denotes a measurable space such that $\Psi^t \subseteq \Psi'$ and $\mathcal{G}^t \subseteq \mathcal{G}'$ (in particular, Ψ' and $\mathcal{G}'$ can be the original Ψ and $\mathcal{G}$, respectively); $\hat{\boldsymbol{w}}^t = (\hat{w}^t_M \colon M \in \mathcal{M}^t)$ denotes an optimal solution of (4); and $\hat{\mu}^t \equiv P^{\hat{\boldsymbol{w}}^t}$ is defined over $\left(\Psi^t, \mathcal{G}^t\right)$.

Lemma 4. *Suppose that* $\mu \in (\Psi', \mathcal{G}')$ *is such that* $\hat{\mu}^t << \mu_{|\mathcal{G}^t}$. *Define* $\mu' \in (\Psi', \mathcal{G}')$ *as*

$$\mu'(G) = \sum_{M \in \mathcal{M}^t, \, \hat{w}^t_M > 0} \frac{1}{\mu(M)} \hat{w}^t_M \, \mu(G \cap M), \qquad G \in \mathcal{G}'. \tag{6}$$

Then μ' *is a probability measure over* $\left(\Psi', \mathcal{G}'\right)$. *Moreover,* $\mu'(M) = \hat{\mu}^t(M)$ *for all* $M \in \mathcal{M}^t$.

The Lemma above implies that a MLPM can be extended to any σ-algebra $\mathcal{G}'$ containing $\mathcal{G}^t$ as long as one can find a measure μ such that $\hat{\mu}^t$ is absolutely continuous with respect to $\mu_{\mathcal{G}^t}$. For simplicity, hereafter we assume $(\mathcal{G}', \mu)$ satisfies this condition. Whereas Lemma 4 shows that the MLPM can be extended to $\mathcal{G}'$, it does not necessarily imply that the extension in (6) is an MLPM over $(\Psi', \mathcal{G}')$. Next, Theorem 3 gives necessary conditions for a measure to be an MLPM over $\left(\Psi', \mathcal{G}'\right)$. This result together with Lemma 4 imply that μ' constructed in (6) is a MLPM over $\left(\Psi', \mathcal{G}'\right)$.

Theorem 3. *Let* $\hat{\mu}^t$ *be a MLPM over* $\left(\Psi^t, \mathcal{G}^t\right)$. *If* $\mu \in \left(\Psi', \mathcal{G}'\right)$ *is a MLPM over* $\left(\Psi', \mathcal{G}'\right)$, *then* $\mu(\cup_{M \in \mathcal{M}^t} M) = 1$, $L(\mu, t) = L(\hat{\mu}^t, t)$, *and* $\left\{\mu(M) \colon M \in \mathcal{M}^t\right\}$ *is an optimal solution of* (4).

Proof. Consider $\mu_{|\mathcal{G}^t} \in \left(\Psi^t, \mathcal{G}^t\right)$; because $C^s \in \mathcal{G}^t$ we have that $\mu_{|\mathcal{G}^t}(C^s) = \mu(C^s)$, for all $s \in [t]$, thus $L(\mu, t) = L(\mu_{|\mathcal{G}^t}, t)$; from the optimality of $\hat{\mu}^t$ we conclude that $L(\mu, t) \leq L(\hat{\mu}^t, t)$. On the

other hand, because μ maximizes $L(\cdot, t)$ across measures in $\left(\Psi', \mathcal{G}'\right)$, and $\hat{\mu}^t \in (\Psi', \mathcal{G}')$ it must be that $L(\mu, t) \geq L(\hat{\mu}^t, t)$, and we conclude that $L(\mu, t) = L(\hat{\mu}^t, t)$. This also implies that μ is an MLPM over $\mathcal{G}^t$ and thus, per Theorem 1 it must be that $\mu(\cup_{M \in \mathcal{M}^t} M) = \mu_{|\mathcal{G}^t}(\cup_{M \in \mathcal{M}^t} M) = 1$. Finally, Theorem 2 implies that μ is an optimal solution of (4). □

Remark 3. *Note that in contrast to $\left(\Psi^t, \mathcal{G}^t\right)$, there might be multiple MLPMs over $\left(\Psi', \mathcal{G}'\right)$ even if problem* (4) *has a unique solution. However, note that such multiple solutions are equivalent when restricted to $\mathcal{G}^t$, in which case they coincide with the solution to* (4)*.*

4 Computation of the CMI and MLPMs

From previous sections, we know that finding a MLPM amounts to solving a convex optimization problem that receives the CMI as input. Thus, we first focus on the problem of finding the CMI. Because the size of the CMI can be exponential in t, later in this section we develop a column generation procedure for computing the MLPM that attempts to bypass computing all the sets in the CMI.

4.1 Exhaustive search for the CMI

We find the elements in the CMI recursively, starting with the elements with the highest cardinality, and then using those to find the ones with lower cardinality. For a collection of indices $\mathcal{S} \subseteq 2^{[t]}$, define

$$\mathcal{V}(\mathcal{S}) := \arg\max \left\{ |S| : S \subseteq [t], I(S) \neq \emptyset,\ S \not\subseteq S',\ S' \in \mathcal{S} \right\}. \tag{7}$$

Algorithm 1 below provides pseudo-code for iteratively finding the CMI. Starting from $\mathcal{S} = \emptyset$, at each iteration, the algorithm finds the largest set $S \subseteq [t]$ that is not contained in any element of $\mathcal{S}$, and such that $I(S) \neq \emptyset$; this set is then added into $\mathcal{S}$. The algorithm stops when one can no longer find such a set. Note that the resulting $\mathcal{S}$ is the collection of time indices associated with the CMI. This is,

$$\mathcal{M}^t = \{ I(S) : S \in \mathcal{S} \}.$$

Indeed, by construction, $S \in \mathcal{S}$ is such that it is not a proper subset of any other element of $\mathcal{S}$, $I(S)$ is non-empty, and S can not be augmented without resulting in an empty intersection. Conversely, if $S \notin \mathcal{S}$, then it must be the case that either $I(S) = \emptyset$ or that $S \subseteq S'$ for some $S' \in \mathcal{S}$.

The ability to find the CMI using Algorithm 1 depends on (i) the number of elements in $\mathcal{M}^t$ and (ii) the efficiency of solving (7), which are both problem dependent. Regarding (i), the 'maximal' condition of the elements in $\mathcal{S}$ implies that not all subsets of $[t]$ can be simultaneously in $\mathcal{S}$. Indeed, suppose that t is even; Sperner's Theorem [3] implies that $\left|\mathcal{M}^t\right| \leq \binom{t}{t/2}$. Unfortunately, such a bound is tight (consider, for example, all subsets of size $t/2$), thus in the worst case there can be $O(2^t/t)$ elements in $\mathcal{M}^t$. In practice, $\mathcal{M}^t$ might be significantly smaller, as shown in Section 7. Regarding (ii), the complexity of problem (7) depends on the application at hand, and most

Algorithm 1: Computing the CMI

Input: The collection $\mathcal{C}^t$
Output: The CMI $\mathcal{M}^t$
1: Set $\mathcal{S} = \emptyset$
2: **while** $\mathcal{V}(\mathcal{S})$ is feasible **do**
3: Set $\mathcal{S} = \mathcal{S} \cup \mathcal{V}(\mathcal{S})$
4: **end while**
5: **return** $\mathcal{M}^t = \{I(S) \colon S \in \mathcal{S}\}$

importantly, on the nature of the sets in $\mathcal{C}^t$. In order to illustrate this point, we consider two examples.

Example 2 (**Rectangle Feedback**)**.** Suppose C^s are rectangles in $\mathbb{R}^n_+$, i.e. $C^s = C^s_1 \times \ldots \times C^s_n$, where $C^s_j = [l^s_j, u^s_j]$ with $u^s_j \geq l^s_j \geq 0$, $j \leq n$, $s \in [t]$. Then, one has that

$$\begin{aligned} \mathcal{V}(\mathcal{S}) := \arg\max & \sum_{s\in[t]} z^s \\ \text{s.t.} \quad u^{s'}_j + K\,(1 - z^{s'}) &\geq l^s_j\, z^s, \qquad j \in [n],\, s,\, s' \in [t] \\ \sum_{s\in[t]\setminus S} z^s &\geq 1, \qquad S \in \mathcal{S} \\ z_s \in \{0,1\}\,,\ s \in [t]. & \end{aligned}$$

Here, $z^s = 1$ represents that s belongs to $\mathcal{V}(\mathcal{S})$, and $z^s = 0$ otherwise, and K is a 'big-M' constant (which can be readily computed from the u^s_j and l^s_j). The first set of constraints checks that the intersection of the rectangles is not empty by checking that, in each dimension, the largest lower endpoint of the selected intervals is smaller than the smaller upper endpoint of the same intervals; the second set of constraints imposes that a new index is chosen outside the set S, so as to avoid choosing a proper set of S, for each set $S \in \mathcal{S}$. ■

Example 3 (**Computing the CMI in the shortest-path interdiction example**)**.** Consider the shortest-path interdiction of Example 1, and recall that for a given $s \in [t]$, B^s and P^s denote the set of blocked arcs and the path traversed by the evader, respectively, at time s. We formulate (7) by maximizing the number of periods on which one can find a common cost vector $\boldsymbol{c}$ which explains the feedback observed on such periods (i.e. $I(S) \neq \emptyset$). Because it is assumed that μ^0 is absolutely continuous, we impose that for each pair of periods $s, s' \in S$, the costs associated with paths P^s and $P^{s'}$ (under cost vector $\boldsymbol{c}$) are not equal, unless said paths coincide: this avoids situations where $I(S) \neq \emptyset$ but $I(S)$ has null Lebesgue measure. With this, we have that

$$\mathcal{V}(\mathcal{S}) := \max \sum_{s\in[t]} z^s \tag{8a}$$

$$\text{s.t.} \quad \rho^s_n - \rho^s_1 = \boldsymbol{c}^\top y^s, \qquad s \in [t] \tag{8b}$$

$$\rho^s_j - \rho^s_i \leq c_{i,j} + 1 - z^s, \qquad (i,j) \in A \setminus B^s,\, s \in [t] \tag{8c}$$

$$\left|\boldsymbol{c}^\top y^s - \boldsymbol{c}^\top y^{s'}\right| \geq \epsilon + z^s + z^{s'} - 2 \qquad s, s' \in [t] : y^s \neq y^{s'} \tag{8d}$$

$$\sum_{s \in [t] \setminus S} z^s \geq 1, \qquad S \in \mathcal{S} \tag{8e}$$

$$\sum_{a \in A} c_a = 1 \tag{8f}$$

$$\tag{8g}$$

Here, y^s stands for a vectored representation of path P^s, for all $s \in [t]$, and $\epsilon > 0$ stands for a minimum gap. The formulation above checks that $I(S) \neq \emptyset$ by finding a cost vector $\boldsymbol{c}$ that explains the path taken by the evader in the periods included in S via linear programming (LP) duality (constraints (8b)-(8c)), using ρ^s to denote the variables in the dual of the shortest path formulation in period $s \in [t]$. In addition, constraint (8d) imposes a minimum gap in costs (according to $\boldsymbol{c}$) among different paths selected by the evader in the periods in S; while non-linear in principle (as it imposes a lower bound on an absolute value), it can be linearized by introducing auxiliary binary variables. Like in the case of rectangular feedback, constraints (8e) ensure the set S is not a proper set of other set in $\mathcal{S}$. Note that, due to the fact that the evader's path choice is invariant to scaling of the cost vector, we impose (w.l.o.g.) that the vector $\boldsymbol{c}$ belongs to the simplex (constraint (8f)).

■

Note that the formulations presented above are (Mixed) Integer Programs, and as such can be tackled using state-of-the-art solvers. We close this section by noting that in the special case of rectangular feedback and $n = 1$, the CMI can be computed by finding maximal cliques on the intersection graph induced by $\mathcal{C}^t$, see [33]. This approach can be generalized to other types of sets as long as $\mathcal{C}^t$ satisfies the *Helly property*. In general, however, $\mathcal{C}^t$ might not satisfy this property, which means that in general the CMI cannot be found by means of cliques in the intersection graph, see Remark 4.

Remark 4. *The collection of sets $\mathcal{C}^t$ is said to satisfy the* Helly property *if and only if any subcollection of $\mathcal{C}^t$ with non-empty pairwise intersections has a non-empty intersection (see [16] p. 82). For example, intervals in $\mathbb{R}$ and balls in either the ℓ^1-norm or ℓ^∞-norm in $\mathbb{R}^2$ satisfy the Helly property. If $\mathcal{C}^t$ satisfies the Helly property, then the subsets in $\left\{\mathcal{I}_k^t; k \leq m\right\}$ are the maximal cliques over the* intersection graph *induced by $\mathcal{C}^t$ (the set of vertices of the intersection graph is $[t]$ and there is an edge between s and u, $s, u \in [t]$, if and only if $C^s \cap C^u \neq \emptyset$). We note that if $\mathcal{C}^t$ does not satisfy the Helly property then the CMI cannot be computed with maximal cliques, see Figure 1.*

4.2 Computation of a MLPM via column generation

In order to find an MLPM one might find the CMI using Algorithm 1 and then solve formulation (4), which has as many variables as there are elements in the CMI. Thus, one would expect the practical complexity of finding an MLPM to be strongly correlated with the size of the CMI. Considering this, here we explore the possibility of finding an MLPM without knowing all the elements of the CMI. A key observation in this regard is that (5) provides a check for local optimality, and thus can be used to design a column generation scheme, which we do next.

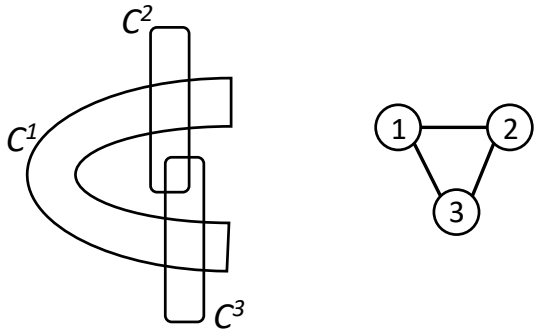

Figure 1: Consider the sequence of three sets of $\mathbb{R}^2$ in the left. The intersection graph is in the right. Clearly the maximal clique is the graph itself, but $I(\{1,2,3\}) = \emptyset$.

Suppose that we have access to a collection $\mathcal{S} \subseteq 2^{[t]}$ such that (i) for each $s \in [t]$ there exists a set $S \in \mathcal{S}$ such that $s \in S$; and (ii) such that $I(S) \in \mathcal{M}^t$ for all $S \in \mathcal{S}$. Note that one can construct such a set so that $|\mathcal{S}| \leq t$ (see, for example, Algorithm 3 below). Consider the formulation

$$\max \sum_{s\in[t]} \ln\Big(\sum_{S\in\mathcal{S},\, s\in S} w_S \Big) \tag{9a}$$

$$\text{s.t} \sum_{S\in\mathcal{S}} w_S = 1 \tag{9b}$$

$$w_S \geq 0, \quad S \in \mathcal{S}, \tag{9c}$$

let $\tilde{\boldsymbol{w}} = \{\tilde{w}_S;\, S \in \mathcal{S}\}$ denote its optimal solution (condition (i) above ensures such a solution exists), and define

$$q_s := (\sum_{S\in\mathcal{S}, s\in S} \tilde{w}_S)^{-1}, \quad s \in [t].$$

Note that $q_s > 0$ for all $s \in [t]$. Using this definition, select (arbitrarily) $\hat{S} \in \mathcal{S}$ such that $\tilde{w}_{\hat{S}} > 0$, and define $\tilde{\lambda}$ as follows

$$\tilde{\lambda} := \sum_{s\in[t]} q_s \, 1\{s \in \hat{S}\}.$$

From (5), note that $\tilde{\lambda}$ does not depend on the choice of $\hat{S}$. We use this definition to check the optimality of $\tilde{w}$ (modulo proper augmentation) for formulation (4). Define

$$\tilde{\lambda}^*(\mathcal{S}) := \max \left\{ \sum_{s\in S} q_s :\, S \subseteq [t],\, I(S) \neq \emptyset,\, S \not\subseteq S',\, S' \in \mathcal{S} \right\}, \tag{10}$$

and let $\tilde{V}(\mathcal{S})$ denote an optimal solution. We consider two cases, depending on the value of $\tilde{\lambda}^*(\mathcal{S})$ relative to $\tilde{\lambda}$.

Case 1: $\tilde{\lambda} \geq \tilde{\lambda}^*(\mathcal{S})$**.** In this case, we can augment $\tilde{\boldsymbol{w}}$ and construct a solution (5), thus checking

the optimality of the augmented solution. Specifically, for each $M \in \mathcal{M}^t$ define

$$w_M := \begin{cases} \tilde{w}_S & \text{if } S \in \mathcal{S} \text{ and } I(S) = M \\ 0 & \sim \end{cases}, \quad \lambda_M = \tilde{\lambda} - \sum_{s \in [t]: M \cap C^s \neq \emptyset} q_s. \tag{11}$$

Because $\tilde{\lambda} \geq \tilde{\lambda}^*(\mathcal{S})$ we have that $\lambda_M \geq 0$ for all $M \in \mathcal{M}^t$ and thus $(\tilde{\lambda}, \{\lambda_M\}, \{w_M\})$ satisfy (5), which in turn implies that $\mu = P^{\boldsymbol{w}}(\cdot)$ is an MLPM.

Case 2: $\tilde{\lambda} < \tilde{\lambda}^*(\mathcal{S})$. Note that $M^* \equiv I(\tilde{V}(\mathcal{S})) \in \mathcal{M}^t$ (this follows from the definition of (11), because $q_s > 0$ for all $s \in [t]$). However, because $\tilde{\lambda} < \tilde{\lambda}^*(\mathcal{S})$, the construction above is such that $\lambda_{M^*} < 0$, thus (5) does not hold.

The above suggests the following column generation procedure: starting from an initial set $\mathcal{S}$, we solve (10) and add $\tilde{V}(\mathcal{S})$ to $\mathcal{S}$ until we have that $\tilde{\lambda} \geq \tilde{\lambda}^*(\mathcal{S})$, at which point we have found an MLPM. This procedure is depicted in an algorithmic form in Algorithm 2.

Algorithm 2: Column Generation for MLPM

Input: The collection $\mathcal{C}^t$, a collection $\mathcal{S}$ such $\cup_{S \in \mathcal{S}} S = [t]$
Output: An MLPM μ

```
Set EXIT = false
while EXIT do
    Solve (9), find w̃, compute {q_s : s ∈ [t]} and λ̃
    Solve (10) and find Ṽ(S) and λ̃*(S)
    if λ̃ ≥ λ̃*(S) then
        Compute w as in (11) and set EXIT=true
    else
        Set S = S ∪ Ṽ(S)
    end if
end while
return μ = P^w
```

Note that the complexity of Algorithm 2 stems from solving formulations (9) and (10). In this regard, the later formulation is solved quite efficiently in practice (see the results in Section 7), and the former formulation is equivalent to (7) - except for a modification in the coefficients in the objective function which prioritizes finding sets that 'cover' periods that have been assigned with a lower likelihood, according to $\tilde{\boldsymbol{w}}$. Note that, in the worst case, all elements in the CMI are generated, in which case it is necessarily the case that $\tilde{\lambda} = \tilde{\lambda}^*(\mathcal{S})$, and correctness of the algorithm follows. In practice, as in most column-generation procedures, one would expect that an optimal solution is found after a relatively small number of iterations; this is supported in part by the evidence generated in our numerical experiments, where we observe that MLPM concentrates on a small number of elements in the CMI.

Finding an initial set $\mathcal{S}$. Note that Algorithm 1 can be modified to find an initial feasible set within at most t iterations: for example, one possibility is to restrict the search for new elements

of the CMI that 'cover' time periods not included by the incumbent collection. Note that this is possible because all $s \in [t]$ must be included in at least one $M \in \mathcal{M}^t$ (otherwise $\mu(C^s) = 0$ for all MLPM μ and $L(\mu, t) = 0$, which contradicts the fact that μ is an MLPM). Algorithm 3 describes such a procedure.

Algorithm 3: Finding an initial solution $\mathcal{S}$.

Input: The collection $\mathcal{C}^t$
Output: The initial set $\mathcal{S}$
1: Set $\mathcal{S} = \emptyset$
2: **while** $\cup_{S \in \mathcal{S}} S \subset [t]$ **do**
3: Set $\mathcal{V} := \arg\max \{|S| : S \subseteq [t], I(S) \neq \emptyset, S \setminus \cup_{S' \in \mathcal{S}} S' \neq \emptyset\}$
4: **end while**
5: **return** $\mathcal{S}$

5 Convergence properties of MLPMs

In this section, we consider the case $\left(\Psi, \mathcal{G}\right) = (\mathbb{R}^n, \mathbb{B}^n)$, equipped with the usual topology induced by the Euclidean distance. We are interested in analyzing the behavior of MLPMs over $\left(\mathbb{R}^n, \mathbb{B}^n\right)$ as t grows to infinity. It is known that when $C^s = \{\boldsymbol{c}^s\}$ for all $s \in \mathbb{Z}_+$, the MLPM is given by the empirical distribution. In this case, the cumulative distribution function (cdf) induced by the empirical distribution converges uniformly and a.s. to the cdf induced by the original measure μ^0 (this is a consequence of Glivenko-Cantelli Theorem and its generalizations, see Theorem 1 on Chapter 26 of [52]). Other results can be derived under other convergence notions. For instance, the empirical distribution converges to the true distribution in total variation a.s. [51] and the Wasserstein distance between μ^0 and the empirical distribution is also shown to go to zero in probability, under some conditions, as t grows [29].

For the setting of our study, under certain specific assumptions and convergence modes, it has been shown that MLPMs converge to μ^0. For instance, [31] shows that the MLPM converges to μ^0 in the topology of weak convergence assuming that $n = 1$ and that the C^s, $s \in \mathbb{Z}_+$, are specific types of intervals. A similar result for $n = 2$ is discussed in [60]. In Section 5.1 we show that, in general, convergence results such as the ones listed above do not hold, even in the case when there is a unique solution to (4). In fact, proving statistical guarantees that hold for general classes of problems is a complex task as there are many factors that might play a role in proving convergence. For instance, convergence in settings with sets that have the same geometry might depend on subtle differences in their arrangement in the space, see Example 5. These observations emphasize that convergence analyses should be carried out on a case-by-case basis.

Whereas statistical guarantees might not be available for arbitrary cases, in Section 5.2 we leverage recent results on Wasserstein distances to provide a framework that can serve to derive guarantees for particular cases. More precisely, we define a notion of convergence in terms of the Wasserstein distance between μ^0 and the set of MLPMs and show that a sufficient condition for

this distance to go to zero is that the Wasserstein distance between the empirical distribution and the set of MLPMs goes to zero. We then discuss three approaches that can be used to bound such distance. In Section 5.3, one of such approaches is used to provide statistical guarantees in a setting where μ^0 is a discrete distribution.

5.1 General statistical guarantees: counterexamples

Next, consider the following two examples that show some of the issues with providing general convergence guarantees. The first one presents a setting based on the interdiction problem in Example 1 where standard notions of convergence do not hold but where there is convergence under the notion we introduce in this section. The second example shows that having convergence goes beyond the geometry of the sets in $\mathcal{C}^t$; that is, even with the same sets there might be changes in convergence depending on how the sets are arranged in the space.

Example 4. Consider a network whose arc cost vector at each period $s \in [t]$ is drawn at random from a measure μ^0. Assume that on each period a DM only observes the cost of a shortest-path between a fixed pair of nodes on the network, and tries to infer the probability μ^0 from these observations. Specifically, consider the network depicted in Figure 2, the source-sink node pair (1,4), and assume that the DM knows that the cost of each arc belongs to [0,100]. In addition,

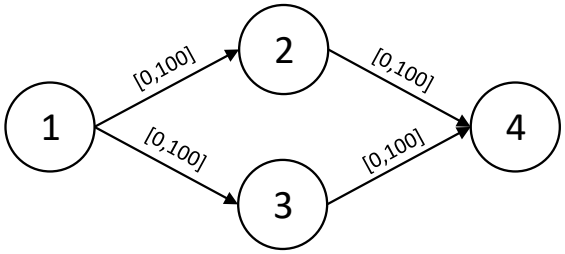

Figure 2: Network used in Example 4. The interval [0,100] above each arc means that it is known that the cost of the arc is a number between 0 and 100.

suppose that the measure μ^0 is degenerate and assigns the value 25 to the cost of each arc with probability 1. Then, the length of the shortest path between 1 and 4 is 50 for each period. Moreover, in each period we have that

$$C^s \equiv C := \{\boldsymbol{c}\colon \min\{c_{12} + c_{24}, c_{13} + c_{34}\} = 50\}, \qquad s \in \mathbb{Z}_+.$$

For each t, consider the measure $\hat{\mu}^t$ that assigns probability 1 to the point $\hat{\boldsymbol{c}} = (25, 25, 100, 100)$. One can show that $\hat{\mu}^t$ maximizes the likelihood (note that $\mathcal{M}^t = \{C\}$, $\hat{\boldsymbol{c}} \in C$, and $L(\hat{\mu}^t, t) = 1$). However, the cdf induced by $\hat{\mu}^t$ is not that induced by μ^0. Moreover, the total variation between μ^0 and $\hat{\mu}^t$ is one, which implies that $\hat{\mu}^t \not\to \mu^0$ almost surely for the total variation metric, the Wasserstein metric, or the KL divergence (the last two follow from the dual representations of the Wasserstein metric [56] and by Pinsker's inequality [44]). Similarly, $\hat{\mu}^t$ does not converge weakly to μ^0 for any given $\omega \in \Omega$ as it would require the integrals of any bounded continuous function to

eventually coincide (see, e.g., the Portmanteau theorem [48] p. 264).

Whereas in this setting we do not have convergence based on standard notions, it is readily seen from results to be introduced later in this section that the Wasserstein distance between μ^0 and the set of MLPMs is zero. ■

Example 5. Consider $n = 1$, assume that μ^0 is a discrete distribution in $\{1, 2, \ldots, J\}$ for some $J \geq 2$; and that the sets C^s are closed intervals of length $2/3$. Consider two settings: in the first, each c^s is the leftmost endpoint of C^s, whereas, in the second, c^s is the leftmost endpoint if s is odd and is the rightmost endpoint if s is even. In the first setting, it can be shown that the Wasserstein distance between μ^0 and the set of MLPMs is zero, whereas in the second there is no such convergence as it can be shown that, eventually, the elements of the CMI do not contain the support of μ^0. ■

5.2 Convergence using Wasserstein distances

Fix t and let $\hat{\boldsymbol{w}}^t$ be a solution of (4). Consider the set of MLPMs

$$\mathcal{D}^t := \big\{\mu \in \big(\mathbb{R}^n, \mathbb{B}^n\big) \colon \mu(M) = \hat{w}^t_M,\ M \in \mathcal{M}^t\big\},$$

and the (possibly unobservable) empirical distribution

$$\widetilde{\mu}^t(B, \omega) := \frac{1}{t} \sum_{s \in [t]} 1_{\{\boldsymbol{c}^s(\omega) \in B\}} \qquad B \in \mathbb{B}^n, \omega \in \Omega.$$

(In what follows we remove the ω dependency of $\widetilde{\mu}^t$; any statement or equation where ω does not appear is implicitly assumed to hold a.s. in $\omega \in \Omega$.) For any $\mu, \mu' \in \big(\mathbb{R}^n, \mathbb{B}^n\big)$ let $d(\mu, \mu')$ denote the 1-th Wasserstein distance between μ and μ' [34] and, abusing some notation, let denote by $d(\mathcal{D}^t, \mu)$ the Wasserstein distance between $\mu \in \big(\mathbb{R}^n, \mathbb{B}^n\big)$ and the set of MLPMs, i.e. $d(\mathcal{D}^t, \mu) := \inf\{d(\mu', \mu) \colon \mu' \in \mathcal{D}^t\}$. Define $\delta^t \equiv d(\mathcal{D}^t, \widetilde{\mu}^t)$: for a given $\epsilon \geq 0$ consider the following ambiguity set, which represents a 'ball' of radius $\epsilon + \delta^t$ around $\mathcal{D}^t$

$$\mathcal{U}^t_\epsilon := \big\{\mu \in \big(\mathbb{R}^n, \mathbb{B}^n\big) \colon d(\mathcal{D}^t, \mu) \leq \epsilon + \delta^t\big\}.$$

We have the following measure concentration result, which bounds the probability that μ^0 is far away from $\mathcal{U}^t_\epsilon$.

Proposition 2. *Assume that there exist $a > 1$ such that $\mathbb{E}^{\mu^0}[\exp(||\boldsymbol{c}||^a)] < \infty$ and let $\epsilon > 0$ be given. Then,*

$$\mathbb{P}\big[d(\mathcal{D}^t, \mu^0) \geq \epsilon + \delta^t\big] \leq \begin{cases} \kappa_1 \exp\big\{-\kappa_2 t \epsilon^{\max\{n,2\}}\big\}, & \text{if } \epsilon \leq 1 \\ \kappa_1 \exp\{-\kappa_2 t \epsilon^a\}, & \text{if } \epsilon > 1 \end{cases}$$

for any $t \geq 1$, $n \neq 2$[1], where κ_1 and κ_2 are positive constants that depend on n, a, and $\mathbb{E}^{\mu^0}[\exp(||\boldsymbol{c}||^a)]$.

[1]While the case of $n = 2$ admits a similar bound, we omit it here so as to avoid introducing additional notation:

From Proposition 2, we conclude that the Wasserstein distance between $\mathcal{D}^t$ and μ^0 converges to zero in probability as long as δ^t converges to zero. Motivated by this, hereafter we say that there is *convergence* (from $\mathcal{D}^t$ to μ^0) if $\delta^t \to 0$ as $t \to \infty$. Whereas in general such convergence cannot be assured (see the second setting in Example 5), next we discuss three approaches that can be used in general settings to upper-bound δ^t and potentially prove convergence.

Bounds using the definition of Wasserstein distance. Consider the following Lemma whose proof follows directly from the definition of δ^t and from the fact that a distribution that assigns weight $\hat{w}^t_M$ to $\tilde{\boldsymbol{c}}_M$, $M \in \mathcal{M}^t$, is an element of $\mathcal{D}^t$. This result shows, for instance, that there is convergence in Example 4 and in the first setting of Example 5.

Lemma 5. *Let $N(\boldsymbol{c})$ be the (random) number of times that vector $\boldsymbol{c}$ has been observed up until period t. Then*

$$\delta^t \leq \inf\Big\{\sum_{s\in[t]}\sum_{M\in\mathcal{M}^t}\lambda^s_M||\boldsymbol{c}^s-\tilde{\boldsymbol{c}}_M||\colon \sum_{s\in[t]}\lambda^s_M=\hat{w}^t_M,\ \sum_{M\in\mathcal{M}^t}\lambda^s_M=\frac{N(\boldsymbol{c}^s)}{t},\\ \lambda^s_M\geq 0,\ \tilde{\boldsymbol{c}}_M\in M,\ M\in\mathcal{M}^t,\ s\in[t]\Big\}. \tag{12}$$

Besides providing a theoretical upper bound for δ^t, Equation (12) also provides a fast method to compute such bounds. This method first finds elements $\boldsymbol{c}_M \in M$ for each $M \in \mathcal{M}^t$; then fixes $\widetilde{\boldsymbol{c}}_M = \boldsymbol{c}_M$ for all $M \in \mathcal{M}^t$; and then solves the resulting LP problem over the λs. This method is employed in the numerical experiments of Section 7 to quickly compute upper bounds of δ^t. (Clearly, the best possible such bound is found by determining the best possible $\boldsymbol{c}_M$ for each $M \in \mathcal{M}^t$; that is, by solving (12) over λ and all c_Ms. however, the resulting optimization problem is non-convex and therefore does not scale.)

Bounds in terms of frequencies. The next result upper-bounds the probabilities that MLPMs assign to the CMI in terms of the number of sets that define each element of $\mathcal{M}^t$ (see equation (2)).

Proposition 3. *Let $\mu \in \mathcal{D}^t$ be an MLPM. Then,*

$$\mu(M) \leq \frac{1}{m^t}\,\left|\{s\in[t]: M\cap C^s\neq\emptyset\}\right|, \quad M\in\mathcal{M}^t.$$

The bound in Proposition 3 implies that the sets in $\mathcal{M}^t$ that do not happen infinitely often, have probability zero under an MLPM as $t \to \infty$. This result might explain the convergent behavior observed in some of our numerical experiments in Section 7.

Convergence using KKT conditions. A final approach to proving convergence is to show that there exist MLPMs that can be made arbitrarily close, as t grows, to a solution of the KKT conditions in (5). We illustrate this approach next under the assumption that μ^0 is a discrete distribution.

see [29] for details

5.3 Convergence for discrete distributions

Throughout this section we assume that μ^0 has a finite support, i.e. $\mathbb{P}\{\boldsymbol{c}^s \in \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_J\}\} = 1$, where $\boldsymbol{d}_j \in \mathbb{R}^n$ for all $j \in [J]$ and $J < \infty$. We also assume, w.l.o.g., that $\mu^0(\{\boldsymbol{d}_j\}) > 0$ for all $j \in [J]$. (Note we do not make any assumption about whether the DM knows the values of $\boldsymbol{d}_j$, $j \in [J]$, or not). For $j \in [J]$ define

$$U_j := \bigcap_{s \in [t], \boldsymbol{d}_j \in C^s} C^s, \tag{13}$$

thus U_j is the intersection of all sets in the sequence that contain $\boldsymbol{d}_j$. The next proposition gives sufficient conditions on the collection $\mathcal{C}^t$ for the empirical distribution to be an MLPM.

Proposition 4. *Suppose that* $\{U_j : j \in [J]\} \subseteq \mathcal{M}^t$ *and that each* C^s *contains exactly one* $\boldsymbol{d}_j$, $s \in [t]$, $j \in [J]$. *Then* $\widetilde{\mu}^t \in \mathcal{D}^t$ *and thus* $\delta^t = 0$.

In general, the conditions of Proposition 4 are too strict. However, if we assume that the size of the elements in $\mathcal{C}^t$ decrease over time, we can obtain a similar result without imposing these conditions to hold, as shown next.

Proposition 5. *Suppose that* $\lim_{t\to\infty} ||C^t|| = 0$ *a.s., where* $||C^t|| := \sup\{||\boldsymbol{c} - \boldsymbol{c}'|| \colon \boldsymbol{c}, \boldsymbol{c}' \in C^t\}$. *Then,* $\lim_{t\to\infty} \delta^t = 0$ *a.s.*

We remark that the proof of Proposition 5 does not necessarily need that the size of the elements in $\mathcal{C}^t$ go to zero as t grows. Even if the sizes of the sets do not go to zero, the proof holds true as the sets in $\mathcal{C}^t$ remain sufficiently small to guarantee that there exists a $t_0 \geq 0$ such that $U_j \in \mathcal{M}^t$ for all $t \geq t_0$ and $j \in [J]$, and for which each C^s, $s \geq t_0$, only contains one $\boldsymbol{d}_j$, $j \in [J]$.

6 Data-driven stochastic optimization and MLPMs

In this section, we consider two opposite applications of MLPM to data-driven optimization, where MLPMs are used to define uncertainty sets, first in the context of robust optimization (where a DM edges against worst-case realizations of uncertainty), and second in the context of sequential decision-making under uncertainty (where a DM follows a principle of optimism in the face of uncertainty). In both settings, we assume that the DM seeks to optimize a function $f(x, \boldsymbol{c})$, which is measurable and lower semi-continuous for each $x \in X$, where $X \subseteq \mathbb{R}^l$ is known and denotes the set of possible values of the decision variables. The function depends of the random vector $\boldsymbol{c}$, thus $(\Psi, \mathcal{G}) = (\mathbb{R}^n, \mathbb{B}^n)$, and the DM does not know the distribution μ^0 and considers all distributions in an uncertainty set simultaneously, finding worst/best-case realizations, depending on the application.

There are several methods to construct ambiguity sets; see for example [23, 47, 59]. Following existing approaches [27], we assume that the ambiguity set is given by a Wasserstein 'ball' around

the set of MLPMs. This is, we consider the ambiguity set

$$\mathcal{U}_\epsilon^t := \big\{\mu \in \big(\mathbb{R}^n, \mathbb{B}^n\big) \colon d(\mathcal{D}^t, \mu) \leq \epsilon + \delta^t\big\},$$

for some $\epsilon > 0$ given.[2] The choice of this ambiguity set is justified by the following out-of-sample guarantee, which is a direct consequence of Proposition 2, and an adaptation of Theorem 3.5 of [27] to this setting (and thus is stated without proof).

Proposition 6. *Suppose that the assumptions of Proposition 2 hold, and let $f \colon \mathbb{R}^l \times \mathbb{R}^n \to \mathbb{R}$ be measurable and $x \colon \Omega \to \mathbb{R}^l$ be a random vector ($n \neq 2$). Then, for any $\beta \in (0,1)$*

$$\mathbb{P}\Big[\Big\{\omega \in \Omega \colon \mathrm{E}^{\mu^0}[f(x(\omega), \boldsymbol{c})] \leq \sup\big\{\mathrm{E}^{\mu}[f(x(\omega), \boldsymbol{c})] \colon \mu \in \mathcal{U}_{\epsilon_t(\beta)}^t(\omega)\big\}\Big\}\Big] \geq 1 - \beta, \tag{14}$$

where $\epsilon_t(\beta)$ is given by

$$\epsilon_t(\beta) = \begin{cases} \left(\frac{\log(\kappa_1 \beta^{-1})}{\kappa_2 t}\right)^{1/\max\{n,2\}} & \text{if } t \geq \log(\kappa_1\beta^{-1})/\kappa_2 \\ \left(\frac{\log(\kappa_1 \beta^{-1})}{\kappa_2 t}\right)^{1/a} & \sim, \end{cases}$$

where κ_1 and κ_2 are the same constants as in Proposition 2.

The out-of-sample guarantee in Proposition 6 states that for any decision x that depends on the observed data up the time t, the unknown expectation $E^{\mu^0}[f(x, \boldsymbol{c})]$ can be upper-bounded with high probability by the best/worst-case expectation of $f(x, \boldsymbol{c})$ over all the distributions in $\mathcal{U}_{\epsilon_t(\beta)}^t$. Observe that the sup in (14) is interpreted as a worst-case whenever f is a loss function (i.e., the distribution in $\mathcal{U}_{\epsilon_t(\beta)}^t$ that gives the largest expected loss); whereas the sup is interpreted as a best-case whenever f is a revenue function (i.e., the distribution in $\mathcal{U}_{\epsilon_t(\beta)}^t$ that gives the highest expected revenue).

6.1 Distributionally Robust Optimization

Following the distributionally robust optimization (DRO) paradigm, we assume that $f(x, \boldsymbol{c})$ is a loss function $\ell(x, \boldsymbol{c})$ and that the DM optimizes under the assumption that μ^0 takes a worst-case realization within the ambiguity set $\mathcal{U}_\epsilon^t$ for some $\epsilon > 0$. That is, the DM solves the DRO problem

$$\text{DRO: } z^t := \inf\Big\{\sup\Big\{E^{\mu}[\ell(x, \boldsymbol{c})] \colon \mu \in \mathcal{U}_\epsilon^t\Big\} \colon x \in X\Big\}. \tag{15}$$

From Proposition 6 we have that, with high probability, the value of the DRO problem upper-bounds the value of the original (unknown) optimization problem $\inf_{x \in X} E^{\mu^0}[\ell(x, \boldsymbol{c})]$. Within the context of DRO, the bound in Proposition 6 is more conservative than the bound in Theorem

[2] Implicitly, we consider the space $\big(\mathbb{R}^n, \mathbb{B}^n\big)$ equipped with the topology induced by the Wasserstein metric. In the resulting space, referred to as the Wasserstein space [40], we can define continuity and compactness notions in terms of the Wasserstein distance.

3.5 of [27] because the radius of the Wasserstein ball is larger by δ^t and because the Wasserstein ball is constructed around a set rather than a single distribution. Such over-conservativeness can be interpreted as the 'price' to pay (in this approach) for not having complete information about the realizations of the random vectors $\boldsymbol{c}^s$, $s \in \mathbb{Z}_+$. Note, however, that Proposition 6 generalizes Theorem 3.5 of [27] in the sense that if $C^s = \{\boldsymbol{c}^s\}$ and thus $\mathcal{D}^s = \{\widehat{\mu}^s\}$ for all $s \in \mathbb{Z}_+$, then both upper-bounds coincide.

While DRO is in principle a bilevel problem, we show that, under certain conditions, it can be formulated as a single-level convex optimization problem. For a given $x \in X$, $\lambda \geq 0$, $\mu \in \big(\mathbb{R}^n, \mathbb{B}^n\big)$, and $\boldsymbol{c} \in \mathbb{R}^n$, define

$$\bar{\ell}(x, \lambda, \boldsymbol{c}) := \sup_{y \in \mathbb{R}^n} \big\{\ell(x, y) - \lambda ||\boldsymbol{c} - y||\big\} \quad \text{and} \quad \bar{E}(x, \lambda, \mu) := \mathrm{E}^{\mu}[\bar{\ell}(x, \lambda, \boldsymbol{c})],$$

where in the definition of $\bar{E}(x, \lambda, \mu)$ the expectation is taken with respect to the random vector $\boldsymbol{c}$, which is distributed according to μ. We note that $\bar{\ell}(x, \lambda, \cdot)$ is measurable from any x and λ, see [14].

Theorem 4. *Suppose that each $M \in \mathcal{M}^t$ is a compact set of $\mathbb{R}^n$, that $\ell(x, \cdot)$ is lsc for each $x \in X$, and let $\epsilon > 0$ be given. Then for each $x \in X$ we have that*

$$\sup\big\{\mathbb{E}^{\mu}[\ell(x, \boldsymbol{c})] \colon \mu \in \mathcal{U}_{\epsilon}^t\big\} = \inf\Big\{\lambda(\epsilon + \delta^t) + \sum_{M \in \mathcal{M}^t} \hat{w}_M^t \sup\{\bar{\ell}(x, \lambda, \boldsymbol{c}) \colon \boldsymbol{c} \in M\} \colon \lambda \geq 0\Big\}. \tag{16}$$

Proof. For a given $\mu' \in \mathcal{D}^t$, let $H_{\mu'}$ denote the worst-case expectation of $\ell(x, \boldsymbol{c})$ across all probability measures that are at a Wasserstein distance of at most $\epsilon + \delta^t$ of μ', i.e.

$$H_{\mu'} := \sup\big\{\mathbb{E}^{\mu}[\ell(x, \boldsymbol{c})] \colon \mu \in \big(\mathbb{R}^n, \mathbb{B}^n\big),\, d(\mu, \mu') \leq \epsilon + \delta^t\big\}.$$

Then, by Theorem 1 of [14],

$$H_{\mu'} = \inf\Big\{\lambda(\epsilon + \delta^t) + \bar{E}(x, \lambda, \mu') \colon \lambda \geq 0\Big\}.$$

On the other hand, note that

$$\begin{aligned}\sup\big\{\mathbb{E}^{\mu}[\ell(x, \boldsymbol{c})] \colon \mu \in \mathcal{U}_{\epsilon}^t\big\} &= \sup\{H_{\mu'} \colon \mu' \in \mathcal{D}^t\} \\ &= \sup\Big\{\inf\Big\{\lambda(\epsilon + \delta^t) + \bar{E}(x, \lambda, \mu')] \colon \lambda \geq 0\Big\} \colon \mu' \in \mathcal{D}^t\Big\}.\end{aligned}$$

Now, for any $\lambda \geq 0$ and $\mu' \in \big(\mathbb{R}^n, \mathbb{B}^n\big)$, define $\tilde{H}(\lambda, \mu') := \lambda(\epsilon + \delta^t) + \bar{E}(x, \lambda, \mu')$. Lemmas 6 and 7, which can be found in the appendix, imply that $\tilde{H}$ is lsc and convex with respect to λ (for each μ') and upper-semicontinuous and concave with respect to μ' (for each λ). Moreover, $\mathcal{D}^t$ is a compact set from Lemma 7 (the fact that each M, $M \in \mathcal{M}^t$ is compact, implies that $\bigcup_{M \in \mathcal{M}^t} M$ is compact).

Therefore, the minimax Theorem (Corollary 3.3 of [53]) implies that

$$\sup\{\mathbb{E}^{\mu}[\ell(x,\boldsymbol{c})]\colon \mu \in \mathcal{U}_{\epsilon}^{t}\} = \inf\Big\{\sup\Big\{\lambda(\epsilon+\delta^{t}) + \bar{E}(x,\lambda,\mu')\colon \mu' \in \mathcal{D}^{t}\Big\}\colon \lambda \geq 0\Big\}.$$

The desired result then follows from Proposition 8 (see appendix) as $\bar{\ell}$ is lsc over $\boldsymbol{c}$ for each $x \in X$ and $\lambda \geq 0$ and each $M \in \mathcal{M}^{t}$ is compact. □

Theorem 4 implies that the DRO problem (15) can be formulated as

$$\inf\ \lambda(\epsilon+\delta^{t}) + \sum_{M\in\mathcal{M}^{t}} \hat{w}_{M}^{t}\, y_{M} \tag{17a}$$

$$\text{s.t. } y_{M} \geq \ell(x,\boldsymbol{c}) - \lambda \inf\{||\boldsymbol{c}-\boldsymbol{c}'||\colon \boldsymbol{c}' \in M\}, \qquad \boldsymbol{c}\in\mathbb{R}^{n}, M\in\mathcal{M}^{t} \tag{17b}$$

$$\lambda \geq 0,\ x \in X,\ y_{M} \in \mathbb{R},\ M \in \mathcal{M}^{t}. \tag{17c}$$

Problem (17) is a semi-infinite convex optimization problem as long as $\ell(x,\boldsymbol{c})$ is convex in x, which can be further reformulated as a finite convex optimization problem by following a similar procedure to the one in [27]. Alternatively, formulation (17) can be solved by a decomposition delayed constraint generation algorithm [15]. These types of approaches are typically faster in practice than solving the finite reformulation directly (see for example [10, 39] in the context of convex robust optimization) and are suitable in more general settings where the X and/or the CMI are non-convex (e.g., mixed-integer), see [17]. Finally, note that if $C^{s} = \{\boldsymbol{c}^{s}\}$, and thus $\mathcal{D}^{s} = \{\widetilde{\mu}^{s}\}$ for all $s \in [t]$, then formulation (17) becomes Formulation (11) of [27].

A compact reformulation of DRO for certain piece-wise linear loss functions. Here, we consider the special case when

$$\ell(x,\boldsymbol{c}) \equiv \min\{\boldsymbol{c}^{\top}y\colon y \in Y(x)\}, \tag{18}$$

with $Y(x) \subseteq \mathbb{R}^{n}$ compact and non-empty for any $x \in X$. That is, $\ell(x,\boldsymbol{c})$ is the value of an optimization problem with a linear objective function, where the cost vector of the objective is $\boldsymbol{c}$ and the variables in x potentially modify the feasible region of the problem. (Note that the form in (18) generalizes linear functions because if $Y(x) = x$ for all x then $\ell(x,\boldsymbol{c}) = \boldsymbol{c}^{\top}x$.)

We assume that the elements of $\mathcal{M}^{t}$ and X consist of non-negative vectors, and that $||\cdot||$ stands for the ℓ^{2}-norm in $\mathbb{R}^{n}$. Under these assumptions, we derive a simpler reformulation of problem (17) that is more amenable to standard optimization solvers. First, consider the following auxiliary result, for which we define $C(x) \equiv \{\boldsymbol{c}' \in \mathbb{R}^{n}\colon \ell(x,\boldsymbol{c}') \geq 0,\ ||\boldsymbol{c}'|| = 1\}$.

Proposition 7. *Let $x \in X \subseteq \mathbb{R}^{n}_{+}$ and $\boldsymbol{c} \in \mathbb{R}^{n}_{+}$ be given. Suppose that $\ell(x,\boldsymbol{c})$ is defined by (18), that the elements of $Y(x)$ and M are non-negative for all $x \in X$ and $M \in \mathcal{M}^{t}$, and that distances are measured using the metric induced by the ℓ^{2}-norm. If $\lambda > \sup\{\ell(x,\boldsymbol{c}')\colon \boldsymbol{c}' \in C(x)\}$, then $\bar{\ell}(x,\lambda,c) = \ell(x,\boldsymbol{c})$.*

We use the result above to reformulate the DRO.

Theorem 5. *Suppose that $\ell(x, \boldsymbol{c})$ is defined by (18), that the elements of $Y(x)$ and M are non-negative for all $x \in X$ and $M \in \mathcal{M}^t$, and that distances are measured using the metric induced by the ℓ^2-norm. Then,*

$$\sup\big\{\mathrm{E}^{\mu}[\ell(x, \boldsymbol{c})] \colon \mu \in \mathcal{U}_{\epsilon}^{t}\big\} = \sup\big\{\ell(x, \boldsymbol{c}) \colon ||\boldsymbol{c}|| \leq 1\big\}(\epsilon + \delta^t) + \sum_{M \in \mathcal{M}^t} \hat{w}_M^t \sup\{\ell(x, \boldsymbol{c}) \colon \boldsymbol{c} \in M\}.$$

Proof. From Lemma 8 (in the appendix) and Proposition 7 we see that the optimization problem in (16) is unbounded if $\lambda < \sup\{\ell(x, \boldsymbol{c}') \colon \boldsymbol{c}' \in C(x)\}$, while it is bounded if $\lambda \geq \sup\{\ell(x, \boldsymbol{c}') \colon \boldsymbol{c}' \in C(x)\}$ and in this case $\sup\{\bar{\ell}(x, \lambda, \boldsymbol{c}) \colon \boldsymbol{c} \in M\} = \sup\{\ell(x, \boldsymbol{c}) \colon \boldsymbol{c} \in M\}$ for each $M \in \mathcal{M}^t$. Consequently, Theorem 5 follows after noting that $\sup\{\ell(x, \boldsymbol{c}') \colon \boldsymbol{c}' \in C(x)\} = \sup\{\ell(x, \boldsymbol{c}') \colon ||\boldsymbol{c}'|| \leq 1\}$ because of the non-negativity assumptions on $Y(x)$ and M, $M \in \mathcal{M}^t$ (i.e., the sup cannot be attained a $\boldsymbol{c}'$ such that $\ell(x, \boldsymbol{c}') < 0$ because there exists $\boldsymbol{c}'$ with $||\boldsymbol{c}'|| = 1$ such that $\ell(x, \boldsymbol{c}') \geq 0$). □

Theorem 5 is used in Section 7 to formulate data-driven problems involving worst-case distributions. Indeed, a direct application of this result to formulation (15) results in

$$\text{DRO} : z^t = \inf\Big\{\sup\big\{\ell(x, \boldsymbol{c}) \colon ||\boldsymbol{c}|| \leq 1\big\}(\epsilon + \delta^t) + \sum_{M \in \mathcal{M}^t} \hat{w}_M^t \sup\{\ell(x, \boldsymbol{c}) \colon \boldsymbol{c} \in M\} \colon x \in X\Big\}.$$

An advantage of this formulation relative to the previous one there is no constraint tying cost vector selections across the CMI, as in (17b). Thus, for a fixed $x \in X$, evaluation of the objective function amounts to maximizing the loss function across the elements of the CMI, individually. This feature can be used to design algorithmic approaches to solving the DRO.

Remark 5. *Theorem 5 can be employed the reformulate the 'full information' special case when $C^s = \{\boldsymbol{c}^s\}$ for all $s \in [t]$ [27]. In this case the DRO reduces to*

$$z^t = \inf\Big\{\sup\big\{\ell(x, \boldsymbol{c}) \colon ||\boldsymbol{c}|| \leq 1\big\}(\epsilon + \delta^t) + \frac{1}{t}\sum_{s \in [t]} \ell(x, \boldsymbol{c}^s) \colon x \in X\Big\}.$$

6.2 Greedy and optimistic solutions for shortest-path interdiction

Following the *optimism in the face of uncertainty* principle, we assume that the DM optimizes under the assumption that μ^0 takes a best-case realization within the ambiguity set $\mathcal{U}_{\epsilon}^{t}$ for some $\epsilon > 0$. Such a principle has been applied recently in the context of interdiction problems with incomplete information [18–20, 61]. In the optimistic case, the function $f(x, \boldsymbol{c})$ is given by a revenue function $r(x, \boldsymbol{c})$. Following (18) suppose that

$$r(x, \boldsymbol{c}) \equiv \min\Big\{\boldsymbol{c}^{\top} y \colon\ y \in Y(x)\Big\},$$

with $Y(x) \subseteq \mathbb{R}^N$ compact and non-empty for any $x \in X$. In this approach, the DM solves the greedy and optimistic optimization (GOO) problem

$$\begin{aligned}\text{GOO}: \quad w^t &\equiv \sup\Big\{\sup\Big\{E^\mu[r(x,\boldsymbol{c})]\colon \mu \in \mathcal{U}_\epsilon^t\Big\} : x \in X\Big\} \\ &= \sup\Big\{\sup\big\{r(x,\boldsymbol{c})\colon ||\boldsymbol{c}|| \le 1\big\}(\epsilon+\delta^t) + \sum_{M\in\mathcal{M}^t} \hat{w}_M^t \sup\{r(x,\boldsymbol{c})\colon \boldsymbol{c}\in M\}\colon x\in X\Big\},\end{aligned}$$

where the last equation comes from applying Theorem 5 with $r(\cdot)$ instead of $\ell(\cdot)$. Note that this accommodates the network interdiction setting in Example 1 when the DM is interested in maximizing the length of the path selected by an evader. Indeed, in such a setting $r(x,\boldsymbol{c})$ corresponds to the cost of the shortest 1-n path in the interdicted network, x are the interdicted arcs, and $\boldsymbol{c}$ the cost vector, i.e.

$$r(x,\boldsymbol{c}) = \min\Big\{\boldsymbol{c}^\top y \,:\, \boldsymbol{B}\, y = \boldsymbol{b},\, y_a \ge 0,\, y_a + x_a \le 1,\, a\in A\Big\},$$

where $\boldsymbol{B}$ denotes the node-arc adjacency matrix of G and $\boldsymbol{b}$ is such that $\boldsymbol{b}_1 = -1$, $\boldsymbol{b}_n = 1$, and $\boldsymbol{b}_i = 0$ otherwise, and $x := (x_a : a \in A)$ is encoded so that $x_a = 1$ if the arc a is interdicted, and $x_a = 0$ otherwise. Note that this function has the form in (18). We consider $X := \Big\{x \in \{0,1\}^{|A|} : \sum_{a\in A} x_a \le \Lambda\Big\}$, where $\Lambda \in \mathbb{Z}_+$ represents a budget parameter.

Let $\mathcal{S}$ denote the collection of sets defining the elements of CMI with positive probability in the MLPM (which is obtained a sub-product of Algorithm 2), and $\{w_S : S \in \mathcal{S}\}$ as the solution to formulation (9). Additionally, define $\mathcal{S}_0 := \mathcal{S} \cup \{0\}$ and $w_0 := (\epsilon + \delta^t)$. We can use LP duality to formulate GOO as follows.

$$\begin{aligned}\text{GOO}: \max \sum_{S\in\mathcal{S}_0} w_S\,\left(\rho_n^S - \rho_1^S\right) & \\ \text{s.t.} \qquad \rho_j^S - \rho_i^S &\le c_{i,j}^S + 1 - x_{i,j}, && (i,j)\in A,\, S\in\mathcal{S}_0 \\ \hat{\rho}_n^{s,S} - \hat{\rho}_1^{s,S} &= (\boldsymbol{c}^S)^\top y^s, && s\in S,\, S\in\mathcal{S} \\ \hat{\rho}_j^{s,S} - \hat{\rho}_i^{s,S} &\le c_{i,j}^S, && (i,j)\in A\setminus B^s,\, s\in S,\, S\in\mathcal{S} \\ \|c^S\| &= 1 && S\in\mathcal{S}_0 \\ \sum_{a\in A} x_a &\le \Lambda & & \\ \rho,\, \hat{\rho}, c \ge 0 \qquad & x_a \in \{0,1\}. & &\end{aligned}$$

Here, y^s stands for a vectored representation of path P^s and B^s for the set of arcs blocked during period s, for $s \in [t]$. Note that variable $\boldsymbol{c}^S \in I(S)$ is a cost vector that explains the evader responses during the periods in S, for $S \in \mathcal{S}_0$ (here, we understand that $I(\emptyset) = \mathbb{R}_+^{|A|}$.) Note that GOO corresponds to the problem of a DM that faces $|S|$ evaders simultaneously, each of whom responds to a different cost vector (chosen by the DM), and whose responses are weighted differently in the DM's objective function.

7 Numerical experiments

We present numerical experiments to illustrate the convergence of the MLPMs to μ^0; how the conservativeness of the proposed DRO evolves over time; how effective is the proposed column generating approach; and how the size of the CMI evolves over time. In our experiments we generate the sequence $\mathcal{C}^t$ based on the rectangular feedback example (Example 2) and the shortest-path interdiction example (Example 1).

7.1 Convergence of MLPM to the empirical distribution and DRO

Robust Linear Assignment. In this section, we illustrate the convergence of MLPMs to the empirical distribution, as a function of t, and study how the value of the DRO in Section 6.1 converges to the optimization problem based on the expected value over the empirical distribution. For this, we consider a DM who solves a linear assignment problem; more specifically, the loss function $\ell(x, \boldsymbol{c})$ is given by

$$\ell(x, \boldsymbol{c}) = \sum_{i \in [n_1]} \sum_{j \in [n_2]} c_{ij}\, x_{ij}$$

for given $n_1, n_2 \geq 1$, and the set of feasible decisions is

$$X = \Big\{ x \in [0,1]^{n_1 \times n_2} \colon \sum_{j \in [n_2]} x_{ij} = 1,\ i \in [n_1],\ \sum_{i \in [n_1]} x_{ij} = 1,\ j \in [n_2] \Big\}. \tag{19}$$

We assume that the actual unknown distribution μ^0 is given by the discrete model introduced in Section 5.3 (i.e. we assume that $\boldsymbol{c}^s \in \{\boldsymbol{d}_j : j \in [J]\}$ a.s.), and that the elements in $\mathcal{C}^t$ are given in the form of rectangles in $\mathbb{R}^{n_1 \times n_2}$, as in Example 2. Specifically, for $s \in \mathbb{Z}_+$, we generate C^s as follows: first, we sample $\boldsymbol{c}^s$ (independently) at random from $\{\boldsymbol{d}_j : j \in [J]\}$; then, for each $(i,j) \in [n_1] \times [n_2]$ we sample $U^s_{i,j} \sim U[0,1]$ (independently across time and components), and set

$$l^s_{i,j} := c^s_{i,j} - U^s_{i,j}\, \Delta^s, \quad u^s_{i,j} := c^s_{i,j} + \Delta^s (1 - U^s_{i,j}),$$

and $C^s = \Big\{ \boldsymbol{c} \in \mathbb{R}^{n_1 \times n_2} : c_{i,j} \in [l^s_{i,j}, u^s_{i,j}], i \in [n_1], j \in [n_2] \Big\}$ for a given side length parameter Δ^s.

Instance generation. We consider $n_1 = n_2 = 3$, $J = 20$, and generate each $\boldsymbol{d}_j$ by sampling each of its components (independently) from a $U[0,1]$ distribution. We also consider a side length parameter $\Delta^s = 0.5/s^\gamma$, and generate instances using three alternative choices for γ: 0, 0.2 and 0.5. Note that in the last two cases, the sizes of the C^s go to zero as s grows.

For each $t \in \{10, 20, \ldots, 100\}$ we use $\mathcal{C}^t$ to find the CMI, and then to solve for the MLPM, which in turn we use to formulate and solve DRO; for this, we considered $\epsilon = 0.01$ and, because δ^t is hard to compute, we use the bound arising from solving the LP formulation in Lemma 5 in its place.[3]

Results and analysis. We consider 30 instances generated according to the procedure described

[3]Note that this is equivalent to using the correct distance δ^t but increasing ϵ by $U^t_w - \delta^t$, where U^t_w is the upper bound in Lemma 5.

above and report mean values across said instances. All instances were run in a machine with a 3.2Ghz 8-Core Intel Xeon W processor, with 32Gb RAM.

Figure 3 reports the evolution – as a function of t – of the mean across instances of four performance measures: (i) the upper bound U_w^t in Lemma 5, (ii) the size $\left|\mathcal{M}^t\right|$ of the CMI, (iii) the number of elements $nz(\hat{\boldsymbol{w}}^t)$ in the CMI that receive a positive value in the MLPM[4]; and (iv) the value z^t of the solution of the DRO problem.

We observe that independent on the rectangle sizes, the (average) value of U_w^t, and thus of δ^t, goes to zero as t increases, and the convergence rate to zero is faster as a function of the rate at which the length size parameter decreases. When $\Delta^s = 0.50/s^{0.2}$ or $\Delta^s = 0.50/s^{0.5}$, such behavior can be explained by Proposition 5, as the size of the elements in $\mathcal{C}^t$ decrease. For the case where Δ^s does not depend on s no such result applies; however, this convergent behavior might be explained in part by Proposition 3. That is, most of the elements of $\mathcal{M}^t$ are intersections of a small number of elements of $\mathcal{C}^t$, whereas the elements of $\mathcal{M}^t$ that contain the $\{\boldsymbol{d}_j, j \in [J]\}$ involve the intersection of a number of sets that grows infinitely often over time.

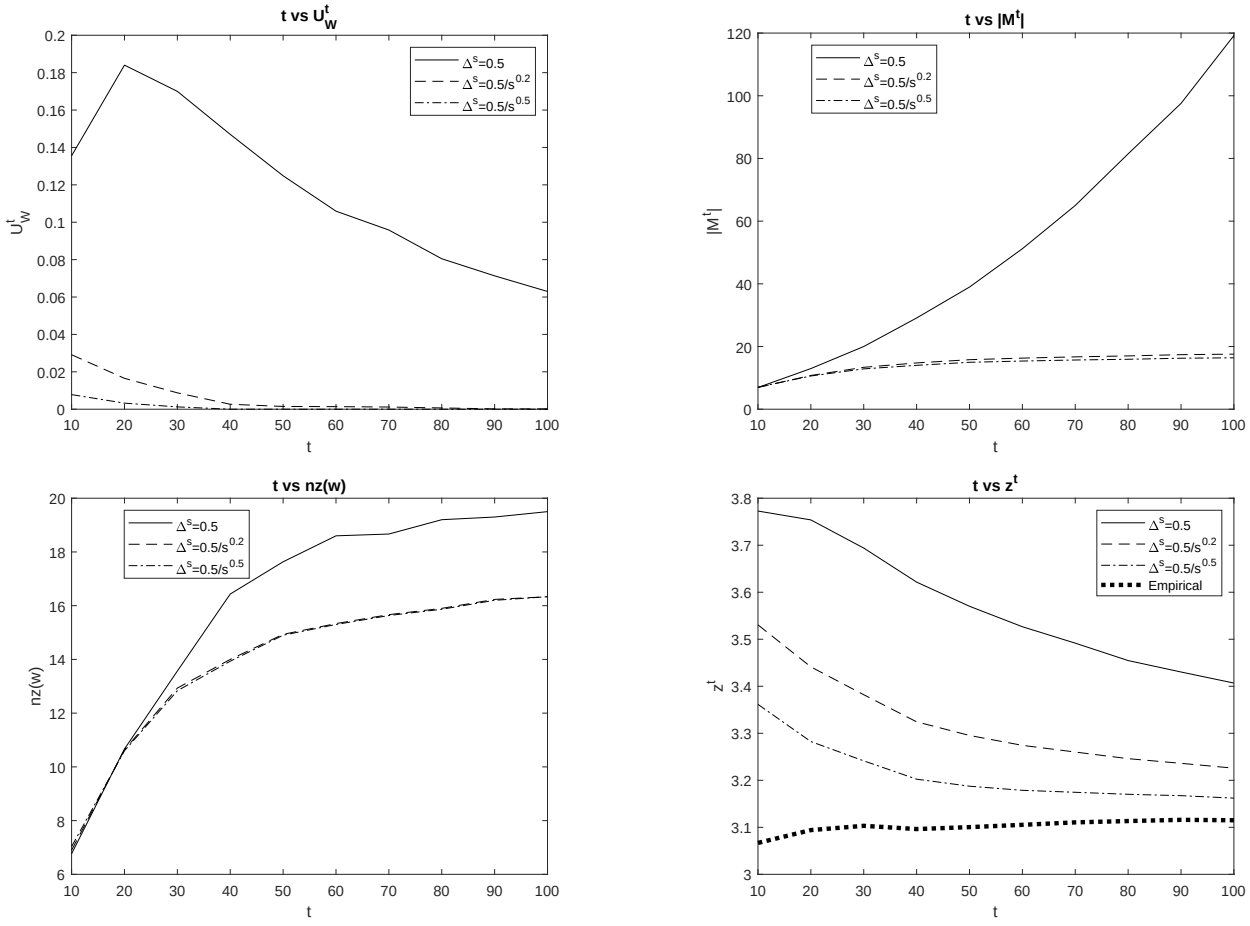

Figure 3: Behavior of U_W^t, $|M^t|$, $nz(\hat{\boldsymbol{w}}^t)$, and z^t as t grows

[4]Because of numerical precision, we report $nz(\hat{\boldsymbol{w}}^t) \approx \left|M \in \mathcal{M}^t : \hat{w}_M^t \geq 10^{-5}\right|$.

Figure 3 also shows that the size of the CMIs, $|\mathcal{M}^t|$, grow linearly and sublinearly with time thus, at least in this model, we do not observe an exponential growth (in time) of the CMI. Importantly, the number of non-zeros of the MLPM weight vector $\hat{\boldsymbol{w}}^t$ grows even slower across all cases, and there are less than 20 nonzeros across all cases at any given time, which is remarkable as the original distribution's range consists of 20 elements. Finally, Figure 3 also shows how the over-conservatism of the DRO approach is reduced as more information is available. Indeed, as t grows, the value of the DROs decreases and converges to the value of the expectation optimization problem that uses the empirical distribution to compute the expectation.

7.2 Computation of CMI via Column Generation and GOO

Shortest-path Interdiction. In this section we illustrate the efficient computation of a MLPM via the column generation procedure in Section 4.2. In particular, we compare the size of the partial set of CMI used for computing MLPM via column generation versus the full size of the CMI. The results of this section are based on a DM that observes the sets in $\mathcal{C}^t$ as in the shortest-path interdiction setting of Example 1. In addition, given the $\mathcal{C}^t$, we consider that the DM periodically solves the GOO problem as described in Section 6.2.

Instance Generation. We consider a layered graph topology with 3 layers and 3 nodes per layer: in each layer (except for the last one) each node has an arc directed toward each node on the next layer; node 1 is connected to each node in the first layer, and each node in the last layer is connected to node n. Thus, in this instance, we have that $|N| = 11$ and $|A| = 24$. On each period s, we generate $\boldsymbol{c}^s$ by drawing each component c_a^s from a $U[0, 1]$ distribution, $a \in A$, and select a set $B^s \subseteq A$ of size Λ by sampling arcs at random from A, $\Lambda = 1$ times without replacement.

For each $t \in \{10, 20, \ldots, 90, 100\}$ we use $\mathcal{C}^t$ to find the CMI via Algorithm 1, and solve formulation (9) to compute the MLPM. Additionally, we also use Algorithm 2 to compute the MLPM and keep track of the partial subset of the CMI found. In both cases, we use formulation (8) to solve for $\mathcal{V}$ imposing a minimum optimality gap of 10^{-3}. In the case of Algorithm 2, we modify the objective function in formulation (8), as described in Section 4.2, and use Algorithm 3 to find an initial set $\mathcal{S}$. We use the MLPM to formulate and solve GOO; for this we consider $\epsilon = 0.01$ and, because δ^t is hard to compute, we use the bound arising from solving the LP formulation in Lemma 5 in its place.

Results and Analysis. We consider 30 instances generated according to the procedure described above and report mean values across said instances. All instances were run in a machine with a 3.2Ghz 8-Core Intel Xeon W processor, with 32Gb RAM.

In Figure 4 (panel on the left) we observe the evolution in time of the mean number of elements of the CMI as found by Algorithm 1 (noted by CMI), the mean partial number of CMI elements as found by Algorithm 2 upon termination (noted by Partial CMI), and the mean number of elements in the CMI that are assigned a positive weight in the MLPM (noted by Optimal CMI). We observe that whereas there is an exponential increase in the size of the CMI, the number of

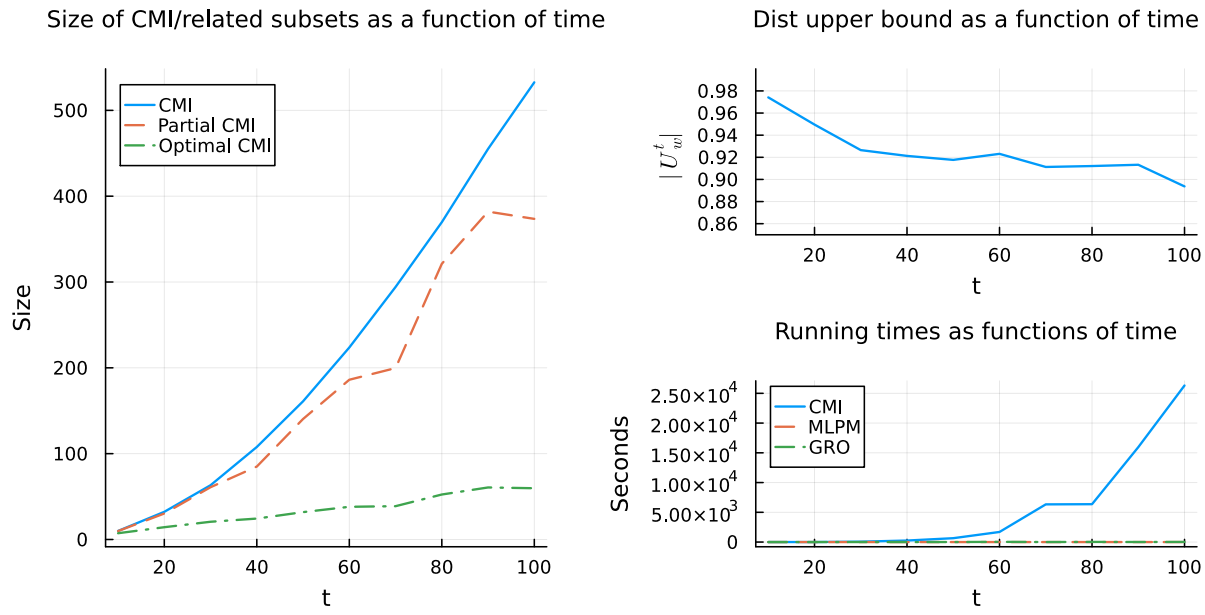

Figure 4: Behavior of set sizes, bounds and running times as t grows.

Optimal CMI sets is linear in t, as suggested by Proposition 1; this fact suggests that there is room from improving the column generation procedure, which we observe produces less elements than the full CMI collection, nonetheless, the increase on its own size also seems exponential in t.

The right panel in Figure 4 (bottom) shows the total running time to compute the CMI by applying Algorithm 1, the MLPM, and solving the GRO; we observe that running times for computing the MLPM and solving the GRO are negligible compared to that required to compute their input, the CMI. In this regard, we observe that (i) the bottleneck in computing the CMI is the number of sets that need to be found; and (ii) running time for solving formulation (8) remains quite constant, although a small increase is observed towards the largest values of t. Finally, the right panel in Figure 4 (top) shows the upper bound in Lemma 5, which we observe decreases modestly with time.

8 Conclusions

In this work, we have studied the problem of non-parametric MLE estimation when, unlike the traditional setting, random elements drawn from an unknown distribution are not directly observable but instead are known to lie within observable sets. We provide a formal treatment of the estimation problem, first identifying structural properties of the ML estimate, namely that its range lies in a collection of maximal intersections, and then providing a characterization in the form of KKT conditions. We show that while convergence results available in traditional settings are not available in general, it still occurs when adapting the notions of convergence to this new setting. More importantly from a decision-making perspective, our work shows how to compute uncertainty sets, in the context of distributionally robust optimization and greedy and optimistic optimization,

in a manner that is consistent with MLE. In future research, we expect to enhance the proposed column generation algorithm for it to be closer to the number of non-zeros observed in the optimal solutions, and to analyze convergence properties in more depth for interdiction problems under various feedback modes and assumptions on μ^0.

Acknowledgments

The research of the first author has been supported by NSF (Grant CMMI 2145553) and AFOSR (Grant FA9550-22-1-0236). The research of the second author has been supported by the grants ANID PIA/APOYO AFB220003 and Fondecyt 1211407.

References

[1] R. Agarwal, Z. Chen, and S. V. Sarma. A novel nonparametric maximum likelihood estimator for probability density functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7):1294–1308, 2016.

[2] R. K. Ahuja and J. B. Orlin. Inverse optimization. *Operations Research*, 49(5):771–783, 2001.

[3] I. Anderson. *Combinatorics of finite sets*. Oxford University Press, 1989.

[4] V. F. Araman and R. Caldentey. Dynamic pricing for nonperishable products with demand learning. *Oper. Res.*, 57(5):1169–1188, 2009.

[5] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

[6] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

[7] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton University Press, 2009.

[8] A. Ben-Tal, A. Goryashko, E. Guslitzer, and A. Nemirovski. Adjustable robust solutions of uncertain linear programs. *Mathematical programming*, 99(2):351–376, 2004.

[9] A. Ben-Tal and A. Nemirovski. Robust solutions of uncertain linear programs. *Operations research letters*, 25(1):1–13, 1999.

[10] D. Bertsimas, I. Dunning, and M. Lubin. Reformulation versus cutting-planes for robust optimization. *Computational Management Science*, 13(2):195–217, 2016.

[11] D. Bertsimas and M. Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004.

[12] O. Besbes and A. Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Oper. Res.*, 57(6):1407–1420, 2009.

[13] R. A. Betensky and D. M. Finkelstein. A non-parametric maximum likelihood estimator for bivariate interval censored data. *Statistics in Medicine*, 18(22):3089–3100, 1999.

[14] J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[15] J. W. Blankenship and J. E. Falk. Infinitely constrained optimization problems. *Journal of Optimization Theory and Applications*, 19(2):261–281, 1976.

[16] B. Bollobás and B. Béla. *Combinatorics: set systems, hypergraphs, families of vectors, and combinatorial probability*. Cambridge University Press, 1986.

[17] J. S. Borrero and L. Lozano. Modeling defender-attacker problems as robust linear programs with mixed-integer uncertainty sets. *INFORMS Journal on Computing*, 2021.

[18] J. S. Borrero, O. A. Prokopyev, and D. Sauré. Sequential shortest path interdiction with incomplete information. *Decision Analysis*, 13(1):68–98, 2016.

[19] J. S. Borrero, O. A. Prokopyev, and D. Sauré. Sequential interdiction with incomplete information and learning. *Operations Research*, 67(1):72–89, 2019.

[20] J. S. Borrero, O. A. Prokopyev, and D. Sauré. Learning in sequential bilevel linear programming. *INFORMS Journal on Optimization*, 4(2):174–199, 2022.

[21] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *CoRR*, abs/1204.5721, 2012.

[22] G. F. de Montricher, R. A. Tapia, and J. R. Thompson. Nonparametric maximum likelihood estimation of probability densities by penalty function methods. *The Annals of Statistics*, pages 1329–1348, 1975.

[23] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[25] M. X. Dong and R. J. Wets. Estimating density functions: a constrained maximum likelihood approach. *International journal of computer mathematics*, 12(4):549–595, 2000.

[26] J. L. Doob. Probability and statistics. *Transactions of the American Mathematical Society*, 36(4):759–775, 1934.

[27] P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

[28] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368, 1922.

[29] N. Fournier and A. Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

[30] S. Geman and C.-R. Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *The annals of Statistics*, pages 401–414, 1982.

[31] R. Gentleman and C. J. Geyer. Maximum likelihood for interval censored data: Consistency and computation. *Biometrika*, 81(3):618–623, 1994.

[32] J. A. Hanley and M. N. Parnes. Nonparametric estimation of a multivariate distribution in the presence of censoring. *Biometrics*, pages 129–139, 1983.

[33] M. G. Hudgens. On nonparametric maximum likelihood estimation with interval censoring and left truncation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):573–587, 2005.

[34] L. V. Kantorovich. Mathematical methods of organizing and planning production. *Management Science*, 6(4):366–422, 1960.

[35] J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906, 1956.

[36] N. Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811, 1978.

[37] M. Loeve. *Probability theory, 2nd Ed.* Van Nostrand, Princeton, NJ, 1960.

[38] S. Modaresi, D. Sauré, and J. P. Vielma. Learning in combinatorial optimization: What and how to explore. *Operations Research*, 68(5):1585–1604, 2020.

[39] A. Mutapcic and S. Boyd. Cutting-set methods for robust convex optimization with pessimizing oracles. *Optimization Methods & Software*, 24(3):381–406, 2009.

[40] V. M. Panaretos and Y. Zemel. *An invitation to statistics in Wasserstein space*. Springer Nature, 2020.

[41] E. Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

[42] M. D. Perlman. On the strong consistency of approximate maximum likelihood estimators. Technical report, University of Minnesota, 1969.

[43] R. Peto. Experimental survival curves for interval-censored data. *Journal of the Royal Statistical Society: series C (Applied Statistics)*, 22(1):86–91, 1973.

[44] M. S. Pinsker. *Information and information stability of random variables and processes*. Holden-Day, 1964.

[45] R. Prentice. Self-consistent nonparametric maximum likelihood estimator of the bivariate survivor function. *Biometrika*, 101(3):505–518, 2014.

[46] R. L. Prentice and S. Zhao. Nonparametric estimation of the multivariate survivor function: the multivariate kaplan–meier estimator. *Lifetime data analysis*, 24(1):3–27, 2018.

[47] H. Rahimian and S. Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

[48] S. Resnick. *A probability path*. Springer, 2019.

[49] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, pages 832–837, 1956.

[50] P. Rusmevichientong and H. Topaloglu. Robust assortment optimization in revenue management under the multinomial logit choice model. *Oper. Res.*, 60(4):865–882, 2012.

[51] N. Sagara. Nonparametric maximum-likelihood estimation of probability measures: existence and consistency. *Journal of statistical planning and inference*, 133(2):249–271, 2005.

[52] G. R. Shorack and J. A. Wellner. *Empirical processes with applications to statistics*. Wiley, New York, NY, 1986.

[53] M. Sion. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

[54] B. W. Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 38(3):290–295, 1976.

[55] V. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264, 1971.

[56] C. Villani. *Topics in Optimal Transportation*. Number 58. American Mathematical Soc., 2003.

[57] A. Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601, 1949.

[58] J.-L. Wang. Strong consistency of approximate maximum likelihood estimators with applications in nonparametrics. *The Annals of Statistics*, pages 932–946, 1985.

[59] W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

[60] G. Y. Wong and Q. Yu. Generalized mle of a joint distribution function with multivariate interval-censored data. *Journal of Multivariate Analysis*, 69(2):155–166, 1999.

[61] J. Yang, J. S. Borrero, O. A. Prokopyev, and D. Sauré. Sequential shortest path interdiction with incomplete information and limited feedback. *Decision Analysis*, 18(3):218–244, 2021.

A Proofs of Selected and Auxiliary Results

Proof of Lemma 1. First, note that because $\mathcal{C}^t$ is a finite collection, there is an explicit representation of $\mathcal{G}^t$, i.e.

$$\mathcal{G}^t := \left\{G : G = \cup_{j\leq k} G_j, \quad G_j = I(S_j) \cap I^c(S'_j),\, S_j,\, S'_j \subseteq [t],\, S_j \cap S'_j = \emptyset,\, k \in \mathbb{Z}_+\right\},$$

where $I^c(S) := \bigcap_{s\in S} \Phi^t \setminus C^s$, $s \subseteq [t]$, and the $\{G_j\}$ are mutually disjoint and finite.

Now, suppose that the result does not hold, and thus that there is $G \in \mathcal{G}^t$ such that $G \subset I(S)$ and $G \neq \emptyset$. This implies that $I(S) \cap G \neq \emptyset$. We can write (w.l.o.g.) $G = \cup_{j\leq k} G_j$, with $G_j = I(S_j) \cap I^c(S'_j)$ disjoint. This implies that

$$I(S) \cap G = \sum_{j\leq k} I(S) \cap I(S_j) \cap I^c(S'_j) = \sum_{j\leq k} I(S \cup S_j) \setminus I(S'_j).$$

(Here the sum of sets stands for disjoint union.) However, note that $I(S\cup S_j) = \emptyset$ unless $S_j \subseteq S$, by the maximality of S. Because $I(S)\cap G \neq \emptyset$ we need only to consider $j \leq k$ such that $I(S\cup S_j) = I(S)$. Suppose $j \leq k$ is such that $I(S) \setminus I(S'_j) \neq \emptyset$: because of the maximality of S, it must be that $S \cap S'_j = \emptyset$, in which case $I(S) \setminus I(S'_j) = I(S)$. Summarizing, $I(S) \cap G \neq \emptyset$ implies that

$$\begin{aligned} I(S) \cap G &= \sum_{j:S_j\subseteq S\,\cap\, S\cap S'_j=\emptyset} I(S) \setminus I(S'_j) + \sum_{j:S_j\subseteq S\,\cap\, S\cap S'_j\neq\emptyset} I(S) \setminus I(S'_j) + \sum_{j:S_j\not\subseteq S} I(S \cap S_j) \setminus I(S'_j) \\ &= \sum_{j:S_j\subseteq S\,\cap\, S\cap S'_j=\emptyset} I(S) \quad = \quad \begin{cases} I(S) & \exists\, j \leq k,\, S_j \subseteq S,\, S'_j \cap S = \emptyset \\ \emptyset & \sim . \end{cases} \end{aligned}$$

Note that, if $I(S) \cap G \neq \emptyset$ implies that $I(S) \cap G = I(S)$, implying that $G = I(S)$, contradicting our assumption that $G \subset I(S)$. This observation proves the result. ∎

Proof of Lemma 2. The fact that $P^{\boldsymbol{w}}$ is non-negative and at most one is immediate. Consider $G_1, G_2 \in \mathcal{G}^t$, disjoint. Because $G_1 \cap G_2 = \emptyset$ and the fact that elements of the CMI are atoms, one has that $\mathcal{M}^t \cap (G_1 \cup G_2) = \mathcal{M}^t \cap G_1 + \mathcal{M}^t \cap G_2$, (here sum stands for disjoint union), thus

$$\begin{aligned} P^{\boldsymbol{w}}(G_1 \cup G_2) &= \sum_{M\subseteq G_1\cup G_2 : M\in\mathcal{M}^t} w_M \\ &= \sum_{M\subseteq G_1 : M\in\mathcal{M}^t} w_M + \sum_{M\subseteq G_1 : M\in\mathcal{M}^t} w_M \quad = \quad P^{\boldsymbol{w}}(G_1) + P^{\boldsymbol{w}}(G_2). \end{aligned}$$

We conclude that $P^{\boldsymbol{w}}$ is additive, and thus σ-additive, because $\mathcal{G}^t$ is finite. ∎

Proof of Lemma 3. Note that the elements in $\mathcal{M}^t$ are disjoint, and define $\bar{M} := \cup_{M\in\mathcal{M}^t} M$. We

have that

$$\begin{aligned}\mu(C^s) &= \sum_{M\in\mathcal{M}^t} \mu(C^s\cap M) + \mu(C^s\cap \bar{M}^c)\\ &= \sum_{M\in\mathcal{M}^t:\, M\subseteq C^s} \mu(C^s\cap M) \quad = \sum_{M\in\mathcal{M}^t:\, M\subseteq C^s} \mu(M) \quad = \sum_{M\in\mathcal{M}^t:\, M\subseteq C^s} w_M\end{aligned}$$

where the first equation follows as $\mu(C^s\cap\bar{M}^c)\le\mu(\bar{M}^c)=0$ and the elements in $\mathcal{M}^t$ are mutually disjoint, the second because the elements of $\mathcal{M}^t$ are either completely contained in C^s or disjoint with C^s, and the last one because $C^s\cap M = M$ if $M\subseteq C^s$. ∎

Proof of Proposition 1. For any feasible solution $\boldsymbol{w}$ to (4) define $q(\boldsymbol{w}) := (q_s(\boldsymbol{w}) : s\in[t])$, where

$$q_s(\boldsymbol{w}) := \sum_{M\in\mathcal{M}^t: M\subseteq C^s} w_M, \quad s\in[t].$$

With this, the KKT conditions (5) can be written as

$$\begin{aligned}\sum_{s\in[t]:M\subseteq C^s} \frac{1}{q_s(\boldsymbol{w})} + \lambda_M + \lambda &= 0, \quad M\in\mathcal{M}^t\\ \lambda_M\, w_M &= 0, \quad M\in\mathcal{M}^t\\ \sum_{M\in\mathcal{M}^t} w_M = 1 \qquad & w_M,\, \lambda_M \ge 0, \quad M\in\mathcal{M}^t.\end{aligned}$$

Let $(\boldsymbol{w}', \lambda, (\lambda_M : M\in\mathcal{M}^t))$ be a solution to the KKT conditions above, and suppose that $|\mathrm{supp}(\boldsymbol{w}')| > t+1$ (otherwise, the result holds true). A key observation is that any non-negative vector $\boldsymbol{w}$ such that $\sum_{M\in\mathcal{M}^t} w_M = 1$, $q(\boldsymbol{w}) = q(\boldsymbol{w}')$ and $w'_M = 0 \Rightarrow w_M = 0$, $M\in\mathcal{M}^t$ is such that $(\boldsymbol{w}, \lambda, (\lambda_M,\, M\in\mathcal{M}^t))$ also solves the KKT conditions. Thus, the result follows if we find such a vector $\boldsymbol{w}$ with the property that $|\mathrm{supp}(\boldsymbol{w})|\le t+1$.

Let $\mathrm{supp}(\boldsymbol{w}')\equiv\{M_1,\ldots,M_J\}$, where $M_j\in\mathcal{M}^t$ for some finite $J > t+1$. For $j\le J$ define $z^j := \Big(z^j_s : s\in[t]\Big)\in\mathbb{R}^t$, where $z^j_s := \mathbf{1}\{M_j\subseteq C^s\}$, $s\in[t]$, and define the matrix $Z := \big(z^1\cdots z^J\big)$. Consider the (non-negative) polyhedron $\mathbf{P}$ defined by

$$\mathbf{P} = \big\{\boldsymbol{w}\in\mathbb{R}^J_+ \colon Z\boldsymbol{w} = q(\boldsymbol{w}'), \mathbf{1}^\top\boldsymbol{w} = 1\big\} \tag{A-2}$$

and note that $\boldsymbol{w}'\in\mathbf{P}$. Define $r := \mathrm{rank}(\binom{Z}{\mathbf{1}^\top}) \le t+1$: from the theory of linear programming, we know there exists a basic feasible solution $\boldsymbol{w}\in\mathbf{P}$ such that $|\mathrm{supp}(\boldsymbol{w})|\le r\le t+1$. This concludes the proof. ∎

Proof of Lemma 4. The fact that μ' is non-negative and one over Ψ' is immediate. Consider σ-additivity. Let $\{G_i \colon i\in\mathbb{Z}_+\}$ a sequence of pairwise disjoint elements of $\mathcal{G}'$. We need to prove

that

$$\mu'(\cup_i G_i) = \sum_i \mu'(G_i).$$

This will follow from the finiteness of $\mathcal{M}^t$. Indeed, note that by absolute continuity $\mu(M) > 0$ for all $M \in \mathcal{M}^t$ such that $\hat{w}^t_M > 0$. Thus, we have that

$$\begin{aligned}
\mu'(\cup_i G_i) &= \sum_{M \in \mathcal{M}^t:\, \hat{w}^t_M > 0} \frac{1}{\mu(M)} \hat{w}^t_M \, \mu(\cup_i G_i \cap M) \\
&= \sum_{M \in \mathcal{M}^t:\, \hat{w}^t_M > 0} \sum_i \frac{1}{\mu(M)} \hat{w}^t_M \mu(G_i \cap M) \\
&= \sum_i \sum_{M \in \mathcal{M}^t:\, \hat{w}^t_M > 0} \frac{1}{\mu(M)} \hat{w}^t_M \mu(G_i \cap M) \quad = \quad \sum_i \mu'(G_i),
\end{aligned}$$

where the third equality follows from monotone convergence. Finally, the fact that μ' and $\hat{\mu}^t$ coincide over $\mathcal{M}^t$ follows from the definition. ■

Proof of Proposition 2. For the sake of clarity, let $\mathcal{D}^t(\omega)$ denote the realization of $\mathcal{D}^t$ under $\omega \in \Omega$, and let $\mu(\omega)$ be the element of $\mathcal{D}^t(\omega)$ that attains the inf in (or that is arbitrarily close to) $d(\mathcal{D}^t(\omega), \widetilde{\mu}^t)$. Note that

$$\begin{aligned}
\big\{\omega \in \Omega \colon d(\mathcal{D}^t(\omega), \mu^0) \geq & \epsilon + \delta^t(\omega)\big\} \subseteq \big\{\omega \in \Omega^t \colon d(\mu(\omega), \mu^0) \geq \epsilon + \delta^t(\omega)\big\} \\
&\subseteq \big\{\omega \in \Omega \colon d(\mu(\omega), \widetilde{\mu}^t(\omega)) + d(\widetilde{\mu}^t(\omega), \mu^0) \geq \epsilon + \delta^t(\omega)\big\} \\
&\subseteq \big\{\omega \in \Omega \colon d(\widetilde{\mu}^t(\omega), \mu^0) \geq \epsilon\big\},
\end{aligned}$$

where the first equation follows because $d(\mu(\omega), \mu^0) \geq d(\mathcal{D}^t(\omega), \mu^0)$, the second from the triangle inequality, i.e., $d(\mu(\omega), \widetilde{\mu}(\omega)) + d(\widetilde{\mu}(\omega), \mu^0) \geq d(\mu(\omega), \mu^0)$, and the last one from the definition of $\delta^t(\omega)$ and μ. Therefore, we conclude that

$$P^t\big[d(\mathcal{D}^t, \mu^0) \geq \epsilon + \delta^t\big] \leq P^t\big[d(\widetilde{\mu}^t, \mu^0) \geq \epsilon\big],$$

and the result follows from Theorem 2 of [29]. ■

Proof of Proposition 3. Let $\hat{\boldsymbol{w}}^t$ be a solution of (4) and recall that $\mu(M) = \hat{w}^t_M$ for any $\mu \in \mathcal{D}^t$ and $M \in \mathcal{M}^t$. Let $S \subseteq [t]$ be such that $I(S) \in \mathcal{M}^t$ and $|S| = m^t$. Suppose that $w_M > 0$, as otherwise the proposition holds trivially. From the KKT conditions (5) and Lemma 3 we have that

$$|s \in [t] : M \cap C^s \neq \emptyset| \, \frac{1}{\mu(M)} \geq \sum_{s \in [t]: M \cap C^s \neq \emptyset} \frac{1}{\mu(C^s)} \geq \sum_{s \in S} \frac{1}{\mu(C^s)} \geq m, \tag{A-4}$$

where the first inequality comes from noting that $\mu(M) \leq \mu(C^s)$ for all $s \in [t]$ such that $M \cap C^s \neq \emptyset$ (M is an atom), the second from (5), and the third from the fact that $\mu(C^s) \leq 1$ for all $s \in [t]$. The result follows from rearranging the terms above. ■

Proof of Proposition 4. Note that $U_j \subseteq C^s$ if and only if $\boldsymbol{d}_j \in C^s$, and that the assumption that C^s contains exactly one $\boldsymbol{d}_j$ implies that $\boldsymbol{d}_j \in C^s$ if and only if $\boldsymbol{c}^s = \boldsymbol{d}_j$. Therefore, we can conclude that $\{U_j \subseteq C^s\} = \{\boldsymbol{c}^s = \boldsymbol{d}_j\}$. Define

$$w_{U_j} := \frac{1}{t} \left| s \in [t] : \boldsymbol{c}^s = \boldsymbol{d}_j \right|, \quad j \in [J],$$

and $w := \big(w_M : M \in \mathcal{M}^t\big)$. Note that w is non-negative, and adds up to one. In addition, for $j \in [J]$ and $s \in [t]$ such that $\boldsymbol{c}^s = \boldsymbol{d}_j$ we have that

$$\sum_{M \in \mathcal{M}^t \colon M \cap C^s \neq \emptyset} w_M = w_{U_j}.$$

This implies that, for any $j \in [J]$,

$$\sum_{s \in [t]} \frac{1_{\{U_j \subseteq C^s\}}}{\sum_{M \in \mathcal{M}^t \colon M \cap C^s \neq \emptyset} w_M} = \sum_{s \in [t] : \boldsymbol{c}^s = \boldsymbol{d}_j} \frac{1}{w_{U_j}} = t.$$

Thus, we have that w, $\lambda = t$, and $\lambda_M = 0$ for all $M \in \mathcal{M}^t$ fulfill the KKT conditions (5), and thus we conclude that P^w is an MLPM. The result follows from noting that $\tilde{\mu}^t|_{\mathcal{G}^t}$ coincides with P^w, and thus is in $\mathcal{D}^t$. ∎

Proof of Proposition 5. Since $\lim_{t\to\infty} ||C^t|| = 0$ then there exists a $t_0 \geq 0$ such that $U_j \in \mathcal{M}^t$ for all $t \geq t_0$ and $j \in [J]$, and for which each C^s, $s \geq t_0$, only contains one $\boldsymbol{d}_j$, $j \in [J]$. For $t \geq t_0$ define

$$w_M^t := \begin{cases} \frac{1}{t} \left| s \in [t] : \boldsymbol{c}^s = \boldsymbol{d}_j \right| & \text{if } M = U_j \text{ for some } j \in [J], \\ 0 & \sim . \end{cases}$$

For $j \in [J]$ and $s \in [t]$, if $U_j \subseteq C^s$ then for $t \geq t_0$ one has that

$$\sum_{M \in \mathcal{M}^t \colon M \cap C^s \neq \emptyset} w_M^t \geq w_{U_j}.$$

Using the above, for a given $j \in [J]$, we have that

$$\begin{aligned} \sum_{s \in [t]} \frac{1_{\{U_j \subseteq C^s\}}}{\sum_{M \in \mathcal{M}^t \colon M \cap C^s \neq \emptyset} w_M^t} &\leq \frac{1}{w_{U_j}^t} \sum_{s \in [t]} 1_{\{U_j \subseteq C^s\}} && \text{(A-5)} \\ &\leq \frac{t}{\left| s \in [t] : \boldsymbol{c}^s = \boldsymbol{d}_j \right|} \sum_{s \in [t]} 1_{\{U_j \subseteq C^s\}} && \text{(A-6)} \\ &\leq \frac{t}{\left| s \in [t] : \boldsymbol{c}^s = \boldsymbol{d}_j \right|} \Big(t_0 + \left| s \geq t_0 \colon \boldsymbol{c}^s = \boldsymbol{d}_j \right| \Big) && \text{(A-7)} \\ &\leq t \Big(1 + \frac{t_0}{\left| s \in [t] \colon \boldsymbol{c}^s = \boldsymbol{d}_j \right|} \Big), && \text{(A-8)} \end{aligned}$$

where the second to last inequality follows because if $s \geq t_0$ then $\boldsymbol{c}^s = \boldsymbol{d}_j$ if and only if $U_j \subseteq C^s$. On

the other hand, let $s \geq t_0$ and $j \in U_j$. Then $U_j \subseteq C^s$ if and only if $\boldsymbol{c}^s = \boldsymbol{d}_j$; similarly, $M \cap C^s \neq \emptyset$ if and only if $M = U_j$. Therefore,

$$\sum_{s\in[t]} \frac{1_{\{U_j\subseteq C^s\}}}{\sum_{M\in\mathcal{M}^t\colon M\cap C^s\neq\emptyset} w_M} \geq \sum_{s\in[t],s\geq t_0} \frac{1_{\{U_j\subseteq C^s\}}}{\sum_{M\in\mathcal{M}^t\colon M\cap C^s\neq\emptyset} w^t_M} \tag{A-9}$$

$$\geq \frac{1}{w^t_{U_j}} \sum_{s\in[t],s\geq t_0} 1_{\{\boldsymbol{c}^s=\boldsymbol{d}_j\}} \tag{A-10}$$

$$\geq \frac{t}{|s\in[t]\colon \boldsymbol{c}^s=\boldsymbol{d}_j|} |s\geq t_0\colon \boldsymbol{c}^s=\boldsymbol{d}_j| \tag{A-11}$$

$$\geq t\Big(1-\frac{t_0}{|s\in[t]\colon \boldsymbol{c}^s=\boldsymbol{d}_j|}\Big). \tag{A-12}$$

We conclude that for any $j \in [J]$, if $t \geq t_0$, then

$$\left(1-\frac{t_0}{|s\in[t]:\boldsymbol{c}^s=\boldsymbol{d}_j|}\right) \leq \frac{1}{t}\sum_{s\in[t]} \frac{1_{\{U_j\subseteq C^s\}}}{\sum_{M\in\mathcal{M}^t\colon M\cap C^s\neq\emptyset} w^t_M} \leq \Big(1+\frac{t_0}{|s\in[t]\colon \boldsymbol{c}^s=\boldsymbol{d}_j|}\Big).$$

Because $\mu^0(\boldsymbol{d}_j) > 0$ for all $j \in [J]$, the converse Borel-Cantelli Lemma implies that

$$\lim_{t\to\infty} \frac{t_0}{|s\in[t]:\boldsymbol{c}^s=\boldsymbol{d}_j|} = 0, \quad \text{a.s.} \quad j\in[J].$$

We can conclude that

$$\lim_{t\to\infty} \frac{1}{t}\sum_{s\in[t]} \frac{1_{\{U_j\subseteq C^s\}}}{\sum_{M\in\mathcal{M}^t\colon M\cap C^s\neq\emptyset} w_M} = 1, \quad \text{a.s.} \quad j\in[J].$$

The above results imply that defining $w = w^t$, $\lambda = -t$ and $\lambda_M = 0$, $M \in \mathcal{M}^t$ can be arbitrarily close to a solution of (5) as t grows. These observations give the desired result. ∎

Proposition 8. *Let $\phi\colon \mathbb{R}^n \to \mathbb{R}$ be a measurable function over $\big(\mathbb{R}^n, \mathbb{B}^n\big)$ such that $\sup\{\phi(\boldsymbol{c})\colon \boldsymbol{c} \in M\} < \infty$ for any M bounded, and such that the supreme is attained by an element of M, for all $M \in \mathcal{M}^t$. Then,*

$$\sup\Big\{\mathrm{E}^\mu[\phi(\boldsymbol{c})]\colon \mu\in\mathcal{D}^t\Big\} = \sum_{M\in\mathcal{M}^t} \hat{w}^t_M \sup\{\phi(\boldsymbol{c})\colon \boldsymbol{c}\in M\}.$$

Proof of Proposition 8. Note that for any $\mu \in \mathcal{D}^t$ it holds that:

$$\mathbb{E}^\mu[\phi(\boldsymbol{c})] = \sum_{M\in\mathcal{M}^t} \int_M \phi(\boldsymbol{c})\, d\mu(\boldsymbol{c}) \leq \sum_{M\in\mathcal{M}^t} \hat{w}^t_M \sup\{\phi(\boldsymbol{c})\colon \boldsymbol{c}\in M\}. \tag{A-13}$$

Let $\boldsymbol{c}^M \in \arg\max\{\phi(\boldsymbol{c})\colon \boldsymbol{c}\in M\}$ and let μ^* be defined by $\mu^*(\{\boldsymbol{c}^M\}) = \hat{w}^t_M$ for all $M \in \mathcal{M}^t$. Then clearly $\mu^* \in \mathcal{D}^t$. Moreover, μ^* attains the upper-bound in the right-hand side of Equation (A-13), and the result follows. ∎

Lemma 6. *Let $x \in X$ and $\mu \in \mathcal{P}\big(\mathbb{R}^n, \mathbb{B}^n\big)$ be given. Then*

1. *$\bar{\ell}(x, \lambda, \boldsymbol{c})$ is lower semi-continuous (lsc), convex, and non-increasing with respect to λ, for any $\boldsymbol{c} \in \mathbb{R}^n$.*
2. *$\bar{E}(x, \lambda, \mu)$ is lsc and convex with respect to λ.*

Proof of Lemma 6. Regarding the first part of the lemma, convexity follows from the definition after using the fact that the supremum over a sum is less than or equal to the sum of the suprema. On the other hand, for a fixed y it is clear that $\ell(x, y) - \lambda||\boldsymbol{c} - y||$ is continuous over λ. Lower semi-continuity follows as the supremum over continuous functions is lsc. The fact that $\bar{\ell}$ is non-decreasing follows from

$$\bar{\ell}(x, \lambda, \boldsymbol{c}) - \bar{\ell}(x, \lambda', \boldsymbol{c}) \leq \sup_{y \in \mathbb{R}^n} \{(\lambda' - \lambda)||\boldsymbol{c} - y||\}.$$

If $\lambda' < \lambda$ then the above supremum is zero (attained when $y = \boldsymbol{c}$), and therefore $\bar{\ell}(x, \lambda, \boldsymbol{c}) \leq \bar{\ell}(x, \lambda', \boldsymbol{c})$, as desired. With regard to the second part of the lemma, the convexity of $\bar{E}$ follows directly from the convexity of $\bar{\ell}$. In order to prove that $\bar{E}$ is lsc in λ, let λ_n, $n \geq 1$, be an increasing sequence of numbers such that $\lim_{n\to\infty} \lambda_n = \lambda$. We have that

$$\lim_{n\to\infty} \bar{E}(x, \lambda_n, \mu) = \lim_{n\to\infty} \mathbb{E}^{\mu}[\bar{\ell}(x, \lambda_n, \boldsymbol{c})] = \mathbb{E}^{\mu}[\lim_{n\to\infty} \bar{\ell}(x, \lambda_n, \boldsymbol{c})] = \mathbb{E}^{\mu}[\bar{\ell}(x, \lambda, \boldsymbol{c})] = \bar{E}(x, \lambda, \mu),$$

where the second equation follows from monotone convergence, as $\bar{\ell}(x, \lambda_n, \boldsymbol{c})$ is a non-increasing sequence in n, and the third equation follows from the fact that $\bar{\ell}$ is lsc in λ. The result follows because the increasing sequence λ_n is arbitrary. ■

Lemma 7. *Let $x \in X$ and $\lambda \geq 0$ be given. Then*

1. *$\mathcal{D}^t$ is convex. Moreover, if $\bigcup_{M \in \mathcal{M}^t} M$ is compact in $\mathbb{R}^n$, then $\mathcal{D}^t$ is compact in $\left(\mathbb{R}^n, \mathbb{B}^n\right)$.*
2. *$\bar{E}(x, \lambda, \mu)$ is concave and continuous in μ.*

Proof of Lemma 7. Regarding the first part of the lemma, we have that convexity is immediate from the definition. Compactness follows because $\mathcal{D}^t$ is a *tight* set of measures, by the assumption that $\bigcup_{M \in \mathcal{M}^t} A$ is compact, and by repeating the arguments of Proposition 2.2.3 and Corollary 2.2.5 of [40]. With regard to the second part of the lemma, the concavity of $\bar{E}(x, \lambda, \mu)$ follows as $\bar{E}(x, \lambda, \alpha\mu + (1-\alpha)\nu) = \alpha\bar{E}(x, \lambda, \mu) + (1-\alpha)G(x, \lambda, \nu)$. For continuity, note that if $\boldsymbol{c}, \boldsymbol{c}' \in \mathbb{R}^n$ then

$$\bar{\ell}(x, \lambda, \boldsymbol{c}) - \bar{\ell}(x, \lambda, \boldsymbol{c}') \leq \sup_{y \in \mathbb{R}^n} \{\lambda(\|\boldsymbol{c} - y\| - \|\boldsymbol{c}' - y\|)\} \leq \lambda(|\boldsymbol{c} - \boldsymbol{c}'\|),$$

and therefore $u(x, \lambda, \boldsymbol{c}) \coloneqq \bar{\ell}(x, \lambda, \boldsymbol{c})/\lambda$ is Lipschitz continuous with constant 1. Consequently, by the dual representation of the Wasserstein distance (see Theorem 1.14 in [56]),

$$\int u(x, \lambda, \boldsymbol{c})(d\mu(\boldsymbol{c}) - d\mu'(\boldsymbol{c})) \leq d(\mu, \mu')$$

for any $\mu, \mu' \in \mathcal{P}\big(\mathbb{R}^n, \mathbb{B}^n\big)$. In other words,

$$\bar{E}(x, \lambda, \mu) - \bar{E}(x, \lambda, \mu') \leq \lambda d(\mu, \mu'),$$

which implies that $\bar{E}(x, \lambda, \mu)$ is continuous over μ, as desired. ∎

Proof of Proposition 7. Because the suprema in Theorem 4 are taken over elements of M, we assume hereafter that $\boldsymbol{c} \in M$ for some $M \in \mathcal{M}^t$. Observe that the sup in the definition of $\bar{\ell}$ can discard elements such that $\ell(x, \boldsymbol{c}) < 0$. Indeed, $\boldsymbol{c}' = \boldsymbol{c}$ is a feasible solution that has a non-negative value in the objective of the sup (by the non-negativity assumptions on $Y(x)$ and M), whereas any $\boldsymbol{c}'$ with $\ell(x, \boldsymbol{c}') < 0$ gives a strictly negative value in the objective. We can further note that, because $\ell(x, \alpha\, \boldsymbol{c}) = \alpha\, \ell(x,\, \boldsymbol{c})$ for $\alpha \in \mathbb{R}_+$, we have that

$$\bar{\ell}(x, \lambda, \boldsymbol{c}) = \sup\{\alpha\, \ell(x, \boldsymbol{c}') - \lambda ||\boldsymbol{c} - \alpha\, \boldsymbol{c}'||\colon ||\boldsymbol{c}'|| = 1,\, \ell(x, \boldsymbol{c}') \geq 0,\, \boldsymbol{c}' \in \mathbb{R}^n,\, \alpha \in \mathbb{R}_+\}. \tag{A-14}$$

On the other hand, for any $\boldsymbol{c}' \in \mathbb{R}^n$ with $\ell(x, \boldsymbol{c}') \geq 0$, define $\boldsymbol{c}'' = \boldsymbol{c}'/||\boldsymbol{c}'||$ and $\alpha = ||\boldsymbol{c}'||$. Then $(\boldsymbol{c}'', \alpha)$ is feasible in (A-14) and attains the same objective function as $\boldsymbol{c}'$. Using these facts, we can write the rhs of (A-14) as a nested optimization problem:

$$\bar{\ell}(x, \lambda, \boldsymbol{c}) = \sup\Big\{\sup\big\{\alpha\, \ell(x, \boldsymbol{c}') - \lambda ||\boldsymbol{c} - \alpha\, \boldsymbol{c}'||\colon \alpha \in \mathbb{R}_+\big\}\colon ||\boldsymbol{c}'|| = 1,\, \ell(x, \boldsymbol{c}') \geq 0,\, \boldsymbol{c}' \in \mathbb{R}^n\Big\}. \tag{A-15}$$

Fix $x \in X$, $\boldsymbol{c} \in \mathbb{R}^n$, $\lambda > 0$, and $\boldsymbol{c}' \in C(x) := \{\boldsymbol{c}'' \in \mathbb{R}^n\colon \ell(x, \boldsymbol{c}'') \geq 0,\, ||\boldsymbol{c}''|| = 1\}$, and define

$$f(\alpha) = \alpha\, \ell(x, \boldsymbol{c}') - \lambda ||\boldsymbol{c} - \alpha\, \boldsymbol{c}'||, \qquad s \in \mathbb{R}. \tag{A-16}$$

Note that $f(\alpha)$ is a concave and differentiable function in α. Lemma 8 shows that if there exist $\boldsymbol{c}' \in C(x)$ such that $\lambda < \ell(x, \boldsymbol{c}')$ then $\bar{\ell}(x, \lambda, \boldsymbol{c}) = \infty$.

Suppose that there $x \in X$ is such that there is no $\boldsymbol{c} \in C(x)$ such that $\lambda < \ell(x, \boldsymbol{c})$ and let $\alpha^*(\boldsymbol{c}')$ be such that it attains the inner supremum in (A-15) for $\boldsymbol{c}'$ given (Lemma 8 provides a characterization.) From Lemma 8 and the concavity of f, if $\alpha^*(\boldsymbol{c}') \leq 0$ then it must be the case that $\sup\{f(\alpha)\colon \alpha \in \mathbb{R}_+\} = f(0) = -\lambda ||c|| < 0$. On the other hand, if $\alpha^*(\boldsymbol{c}') > 0$, then

$$\sup\{f(\alpha)\colon \alpha \in \mathbb{R}_+\} = \ell(x, \boldsymbol{c})(\boldsymbol{c}^\top \boldsymbol{c}') - \sqrt{(||\boldsymbol{c}||^2 - (\boldsymbol{c}^\top \boldsymbol{c}')^2)(\lambda^2 - \ell(x, \boldsymbol{c}')^2)}.$$

Note that setting $\boldsymbol{c}' = \boldsymbol{c}/||\boldsymbol{c}||$ (which belongs to $C(x)$) implies that $\sup\{f(\alpha)\colon \alpha \geq 0\} \geq 0$, thus $\boldsymbol{c}' \in C(x)$ such that $\alpha^* < 0$ cannot attain the optimal in $\bar{\ell}$. Therefore, we have that

$$\bar{\ell}(x, \lambda, \boldsymbol{c}) = \sup\Big\{\ell(x, \boldsymbol{c}')(\boldsymbol{c}^\top \boldsymbol{c}') - \sqrt{(||\boldsymbol{c}||^2 - (\boldsymbol{c}^\top \boldsymbol{c}')^2)(\lambda^2 - \ell(x, \boldsymbol{c}')^2)}\colon \boldsymbol{c}' \in C(x), \alpha^*(\boldsymbol{c}') > 0\Big\}. \tag{A-17}$$

We claim that the sup in (A-17) is attained at $\hat{\boldsymbol{c}}' = \boldsymbol{c}/||\boldsymbol{c}||$. Indeed, $\hat{\boldsymbol{c}}' \in C(x)$ and $\alpha^*(\hat{\boldsymbol{c}}') > 0$. Moreover, $\hat{\boldsymbol{c}}'$ attains the maximum in both terms of the objective in (A-17); particularly, for the

first term $\ell(x,\hat{\boldsymbol{c}}')\hat{(}\boldsymbol{c}^\top\boldsymbol{c}') = \ell(x,\boldsymbol{c})$. Whereas for the second term, the optimality is clear, for the first, assume that $\boldsymbol{c}' \in C(x)$. Then, for any $y \in Y(x)$, $\ell(x,\boldsymbol{c}') \leq \boldsymbol{c}'^\top y$ and

$$\ell(x,\boldsymbol{c}')(\boldsymbol{c}'^\top\boldsymbol{c}) \leq \boldsymbol{c}'^\top y\boldsymbol{c}'^\top\boldsymbol{c} \leq \boldsymbol{c}^\top y.$$

The last inequality can be proven by using the necessary KKT conditions on the quadratic optimization problem $\max\{\boldsymbol{c}'^\top y\,\boldsymbol{c}^\top\boldsymbol{c}' \colon ||\boldsymbol{c}'|| \leq 1, \boldsymbol{c}' \in \mathbb{R}^n\}$ and using the fact that $y\boldsymbol{c}^\top$ is a rank one matrix whose only non-zero eigenvalue is $\boldsymbol{c}^\top t$. Because $\ell(x,\boldsymbol{c}')(\boldsymbol{c}^\top\boldsymbol{c}') \leq \boldsymbol{c}^\top y$ for any $y \in Y(x)$ and any $\boldsymbol{c}' \in C(x)$, it must be the case that $\ell(x,\boldsymbol{c}')(\boldsymbol{c}^\top\boldsymbol{c}') \leq \min\{\boldsymbol{c}^\top y \colon y \in Y(x)\} = \ell(x,c)$ for any $\boldsymbol{c}' \in C(x)$.

Note that if $\lambda = \sup\{\ell(x,\boldsymbol{c}') \colon \boldsymbol{c}' \in C(x)\}$, then the stationary point does not exist: in such a case, from Lemma 8 and the concavity and continuity of f, it is readily seen that $\sup\{f(\alpha) \colon \alpha \in \mathbb{R}_+\} = \ell(x,\boldsymbol{c}')(\boldsymbol{c}^\top\boldsymbol{c}')$. Therefore, in such a case $\bar{\ell}(x,\lambda,\boldsymbol{c}) = \sup\{\ell(x,\boldsymbol{c}')(\boldsymbol{c}^\top\boldsymbol{c}') \colon \boldsymbol{c}' \in C(x)\}$ and, from above, we know that this sup is precisely $\ell(x,\boldsymbol{c})$. Therefore, if $\lambda = \sup\{\ell(x,\boldsymbol{c}') \colon \boldsymbol{c}' \in C(x)\}$ then $\bar{\ell}(x,\lambda,\boldsymbol{c}) = \ell(x,\boldsymbol{c})$. ∎

Lemma 8. *Let f be defined by* (A-16)*. Then*

$$\lim_{\alpha\to\infty} f(\alpha) = \begin{cases} \infty, & \text{if } \lambda < \ell(x,\boldsymbol{c}') \\ -\infty, & \text{if } \lambda > \ell(x,\boldsymbol{c}'). \end{cases}$$

In addition, if $\lambda > \ell(x,\boldsymbol{c}')$, then a stationary point of f exists and it is given by

$$\alpha^* := \boldsymbol{c}^\top\boldsymbol{c}' + \sqrt{\frac{||\boldsymbol{c}||^2 - (\boldsymbol{c}^\top\boldsymbol{c}')^2}{\lambda^2 - \ell(x,\boldsymbol{c}')^2}}. \tag{A-18}$$

Moreover,

$$f(\alpha^*) = \ell(x,\boldsymbol{c}')(\boldsymbol{c}^\top\boldsymbol{c}') - \sqrt{(||\boldsymbol{c}||^2 - (\boldsymbol{c}^\top\boldsymbol{c}')^2)(\lambda^2 - \ell(x,\boldsymbol{c}')^2)}.$$

Proof of Lemma 8. Define $y_c := y^\top\boldsymbol{c}$ and observe that

$$\begin{aligned}
\lim_{\alpha\to\infty} f(\alpha) &= \lim_{\alpha\to\infty} \frac{(\alpha\ell(x,y) - \lambda||\boldsymbol{c}-\alpha y||)(\alpha\ell(x,y) + \lambda||\boldsymbol{c}-\alpha y||)}{\alpha\ell(x,y) + \lambda||\boldsymbol{c}-\alpha y||} \\
&= \lim_{\alpha\to\infty} \frac{\alpha^2\ell(x,y)^2 - \lambda^2(||\boldsymbol{c}||^2 - 2\alpha y_c + \alpha^2)}{\alpha\ell(x,y) + \lambda\sqrt{||\boldsymbol{c}||^2 - 2\alpha y_c + \alpha^2)}} \\
&= \lim_{\alpha\to\infty} \frac{\alpha\ell(x,y)^2 - \lambda^2(||\boldsymbol{c}||^2/\alpha - 2y_c + \alpha)}{\ell(x,y) + \lambda\sqrt{||\boldsymbol{c}||^2/\alpha^2 - 2y_c/\alpha + 1)}} \\
&= \lim_{\alpha\to\infty} \frac{\alpha(\ell(x,y)^2 - \lambda^2) - \lambda^2(||\boldsymbol{c}||^2/\alpha - 2y_c)}{\ell(x,y) + \lambda\sqrt{||\boldsymbol{c}||^2/\alpha^2 - 2y_c/\alpha + 1)}}.
\end{aligned}$$

Note that the last limit goes to ∞ if $\lambda < \ell(x,y)$; it goes to $-\infty$ if $\lambda > \ell(x,y)$. Note that if $\lambda = \ell(x,y)$ then the limit is $\ell(x,y)y_c$. Suppose now that $\lambda > \ell(x,\boldsymbol{c}')$: the expression for α^* follows

after deriving and setting the derivative equal to zero. The second part follows after replacing α^* in the equation for $f(\alpha)$.

Note that the numerator in (A-18) is always non-negative: one can check that $||\boldsymbol{c}|| \geq \boldsymbol{c}^\top \boldsymbol{c}'$ for any $\boldsymbol{c}' \in C(x)$. Also, note that the stationary point can be negative because in general $\boldsymbol{c}^\top \boldsymbol{c}'$ can be negative. ■