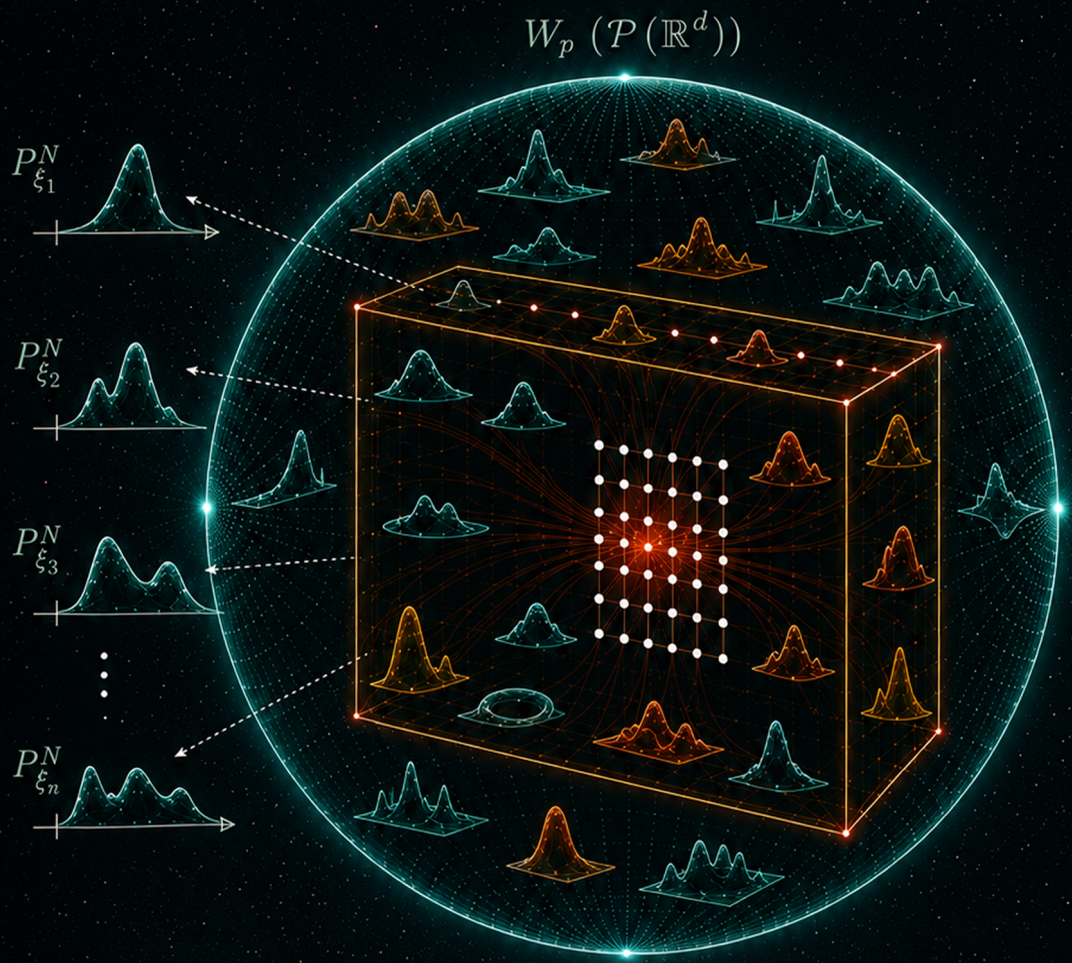
EXPLOITING STRUCTURE IN DISTRIBUTIONALLY ROBUST OPTIMIZATION

PhD thesis



Lotfi M. Chaouach

EXPLOITING STRUCTURE IN DISTRIBUTIONALLY ROBUST OPTIMIZATION

EXPLOITING STRUCTURE IN DISTRIBUTIONALLY ROBUST OPTIMIZATION

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, Prof. dr. ir. H. Bijl,
chair of the Board for Doctorates
to be defended publicly on
Monday, 29 June 2026, at 12.30

by

Lotfi-Mustapha CHAOUACH

Master of Science in Systems and Control,
National Polytechnic School, Algeria
born in Kouba, Algiers, Algeria

This Dissertation has been approved by the (co)promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. ir. T. A. E. Oomen,	Delft University of Technology/ Eindhoven University of Technology, the Netherlands, *promotor*
Dr. D. Boskos,	Delft University of Technology, the Netherlands, *copromotor*
Independent members:	
Prof. dr. ir. JW. van Wingerden,	Delft University of Technology
Prof. dr. R. R. Negenborn,	Delft University of Technology
Prof. dr. P. Patrinos,	Katholieke Universiteit Leuven, Belgium
Dr. P. Mohajerin Esfahani,	University of Toronto, Canada
Dr. A. K. Cherukuri,	University of Groningen

Keywords: Optimal transport ambiguity sets, distributional robustness, stochastic optimization, uncertain systems

Printed by: Proefschrift specialist

Cover by: Lotfi M. Chaouach

ISBN 000-00-0000-000-0

An electronic copy of this dissertation is available at
`https://repository.tudelft.nl/`.

If we knew what it was we were doing, it would not be called research, would it?

Albert Einstein

All mathematics is divided into three parts: cryptography (paid for by CIA, KGB and the like), hydrodynamics (supported by manufacturers of atomic submarines), and celestial mechanics (financed by military and other institutions dealing with missiles, such as NASA).

Vladimir I. Arnold

Science sans conscience n'est que ruine de l'âme.

François Rabelais

CONTENTS

SUMMARY

Uncertainty is unavoidable in decision-making for many real-world problems. Technological developments give rise to systems with increasing complexity; as a result, the sources and effects of uncertainty are magnified. At the same time, these developments are driven by ever-rising standards for performance and safety, which require precise parameter selection and highly accurate design models. Uncertainty stands as a tough obstacle towards effective decision-making and makes parameter selection and model design challenging. These challenges arise across a wide range of domains, including power dispatch, healthcare, finance, aerospace, and biomedical engineering. Therefore, it becomes essential to develop approaches that address the adverse effects of uncertainty in decision-making in order to meet the ambitious performance and reliability standards in these areas.

Addressing uncertainty in decision-making necessitates its modeling. The stochastic approach to model uncertainty allows for the whole range of possible outcomes to be considered along with their probability of occurring. This enables a more subtle understanding of risks by relating their adversity to their likelihood, which permits the deduction of less conservative decisions. In this framework, the modeler resorts to probability theory and uses probability distributions in order to model the uncertainty. This uncertainty modeling paradigm in decision-making yields stochastic optimization problems.

The main challenge in stochastic optimization is to inform decisions with statistical performance guarantees based on a given probability distribution. In practice, however, identifying the appropriate distribution describing a random process may be challenging. Usually, modelers exploit data-driven models like histograms or Gaussian distributions with estimated parameters. However, such inferences do not always describe the uncertainty with the necessary accuracy. Hence, uncertainty does not only arise from the randomness of the different possible scenarios, but also from the choice of the probability distribution. In other words, a deeper layer of uncertainty emerges, as we are confronted with uncertain uncertainty, which we address in this thesis.

This thesis addresses a new way of modeling uncertainty through the framework of distributionally robust optimization, by employing appropriate prior knowledge on the true distribution. The distributionally robust paradigm considers an ambiguity set of plausible distributions that could likely have generated the data, and informs the decision based on the most adverse distribution in this set. Specifically, the core objective of this thesis is to encode prior knowledge about the structure of the true distribution into ambiguity sets in order to include more relevant candidate models. Thus, it yields reliable decisions with improved performance and efficiency.

This dissertation consists of three main parts, each representing an interrelated cornerstone of the approach that is developed:

In the first part, we introduce structured ambiguity sets, which enable us to exclude

irrelevant probabilistic models of the unknown uncertainty. To achieve this, we need to restrict the class of considered uncertainties and specialize to the case where they consist of multiple independent components. The introduced ambiguity sets are structured accordingly to the independent components of the uncertainty. This significantly increases the relevance of the plausible models and improves the performance of their respective decisions.

The second part of this dissertation focuses on developing methods to solve DRO problems associated with the introduced ambiguity sets. We first provide fundamental properties of these new ambiguity sets. We exploit these properties to derive tractable reformulations of their optimization problems and obtain distributionally robust decisions. At last, we present a procedure to relax the potential numerical complexity of these reformulations while preserving the statistical guarantees characterizing the approach.

In the final part of this dissertation, we exploit the tools developed in the two previous parts to design a model predictive controller for linear systems affected by additive stochastic disturbances with an unknown probability distribution. The stochastic model predictive control algorithm is based on a tube approach and enforces chance constraints through a suitable tightening of the admissible state and control input sets. The geometry of the tube is determined offline by solving multiple distributionally robust optimization problems with the same underlying uncertainty. This spares the online part of the optimal control problem from the potential computational burden related to DRO, and leads to a practically implementable model predictive controller.

SAMENVATTING

Onzekerheid is onvermijdelijk bij het nemen van beslissingen voor veel problemen in de praktijk. Technologische ontwikkelingen leiden tot systemen met een steeds hogere complexiteit; als gevolg daarvan nemen zowel de bronnen als de effecten van onzekerheid toe. Tegelijkertijd worden deze ontwikkelingen gedreven door steeds hogere eisen aan prestaties en veiligheid, die een nauwkeurige parameterkeuze en zeer precieze ontwerpmethoden vereisen. Onzekerheid vormt daardoor een belangrijke belemmering voor effectieve besluitvorming en maakt parameterafstemming en modelontwerp bijzonder uitdagend. Deze uitdagingen doen zich voor in uiteenlopende domeinen, waaronder energievoorziening, gezondheidszorg, financiën, lucht- en ruimtevaart en biomedische techniek. Het is daarom essentieel om methoden te ontwikkelen die de nadelige effecten van onzekerheid in besluitvorming aanpakken, om zo te voldoen aan de ambitieuze eisen op het gebied van prestaties en betrouwbaarheid in deze sectoren.

Het omgaan met onzekerheid in besluitvorming vereist allereerst een geschikt model. De stochastische benadering voor het modelleren van onzekerheid maakt het mogelijk om het volledige spectrum van mogelijke uitkomsten te beschouwen, samen met hun kans van optreden. Dit biedt een genuanceerder inzicht in risico's door hun ernst te relateren aan hun waarschijnlijkheid, wat leidt tot minder conservatieve beslissingen. Binnen dit kader maakt de modelleur gebruik van kansrekening en worden waarschijnlijkheidsverdelingen ingezet om onzekerheid te beschrijven. Deze manier van onzekerheidsmodellering resulteert in stochastische optimalisatieproblemen.

De belangrijkste uitdaging binnen stochastische optimalisatie is het nemen van beslissingen met statistische prestatiegaranties op basis van een gegeven kansverdeling. In de praktijk is het echter vaak moeilijk om de juiste verdeling te identificeren die een willekeurig proces adequaat beschrijft. Doorgaans wordt gebruikgemaakt van datagedreven modellen, zoals histogrammen of Gaussische verdelingen met geschatte parameters. Dergelijke benaderingen beschrijven de onzekerheid echter niet altijd met de vereiste nauwkeurigheid. Daardoor ontstaat onzekerheid niet alleen door de willekeurigheid van mogelijke scenario's, maar ook door de keuze van de kansverdeling zelf. Met andere woorden: er ontstaat een diepere laag van onzekerheid, onzekerheid over de onzekerheid, die in dit proefschrift centraal staat.

Dit proefschrift introduceert een nieuwe manier om onzekerheid te modelleren binnen het raamwerk van distributierobuuste optimalisatie, waarbij gebruik wordt gemaakt van geschikte voorkennis over de ware verdeling. Het distributierobuuste paradigma beschouwt een ambiguïteitsverzameling van plausibele verdelingen die de waargenomen data zouden kunnen hebben gegenereerd, en baseert beslissingen op de meest ongunstige verdeling binnen deze verzameling. Het centrale doel van dit proefschrift is het coderen van structurele voorkennis over de ware verdeling in deze ambiguïteitsverzamelingen, zodat relevantere kandidaatmodellen worden meegenomen. Dit resulteert in

betrouwbare beslissingen met verbeterde prestaties en efficiëntie.

Dit proefschrift bestaat uit drie hoofddelen, die elk een onderling samenhangende pijler vormen van de ontwikkelde aanpak:

In het eerste deel introduceren we gestructureerde ambiguïteitsverzamelingen, waarmee irrelevante probabilistische modellen van de onbekende onzekerheid kunnen worden uitgesloten. Hiervoor beperken we ons tot onzekerheden die bestaan uit meerdere onafhankelijke componenten. De voorgestelde ambiguïteitsverzamelingen zijn dienovereenkomstig gestructureerd volgens deze onafhankelijke componenten. Dit verhoogt de relevantie van de plausibele modellen aanzienlijk en verbetert de kwaliteit van de resulterende beslissingen.

Het tweede deel van dit proefschrift richt zich op het ontwikkelen van methoden om distributierobuuste optimalisatieproblemen op te lossen die samenhangen met de geïntroduceerde ambiguïteitsverzamelingen. We presenteren eerst de fundamentele eigenschappen van deze nieuwe verzamelingen. Vervolgens maken we gebruik van deze eigenschappen om hanteerbare herformuleringen van de bijbehorende optimalisatieproblemen af te leiden en distributierobuuste beslissingen te verkrijgen. Tot slot stellen we een procedure voor om de potentiële numerieke complexiteit van deze herformuleringen te verminderen, terwijl de statistische garanties van de aanpak behouden blijven.

In het laatste deel van dit proefschrift passen we de in de eerste twee delen ontwikkelde methoden toe op het ontwerp van een model predictive controller voor lineaire systemen die worden beïnvloed door additieve stochastische verstoringen met een onbekende kansverdeling. Het stochastische model predictive control-algoritme is gebaseerd op een tube-benadering en handhaaft kansbeperkingen door een geschikte aanscherping van de toelaatbare toestands- en regelingsruimten. De geometrie van de tube wordt offline bepaald door het oplossen van meerdere distributierobuuste optimalisatieproblemen met dezelfde onderliggende onzekerheid. Hierdoor wordt het online deel van het optimale regelprobleem gevrijwaard van de mogelijke rekenlast die gepaard gaat met distributierobuuste optimalisatie, wat leidt tot een praktisch implementeerbare model predictive controller.

RÉSUMÉ

L'incertitude lors de la prise de décision est inévitable dans de nombreux problèmes du monde réel. Les avancées technologiques favorisent l'émergence de systèmes à la complexité croissante; ce qui, en conséquence, amplifie les sources et les effets de l'incertitude. Parallèlement, les avancées technologiques s'inscrivent dans un cadre de normes de plus en plus exigeantes en matière de performance et de sécurité, imposant une sélection rigoureuse des paramètres, ainsi que le recours à des modèles de conception de grande précision. L'incertitude constitue donc un obstacle majeur à une prise de décision efficace, rendant la sélection des paramètres ainsi que la modélisation particulièrement délicates. Ces défis apparaissent dans de nombreux domaines, notamment la répartition énergétique, la santé, la finance, l'aéronautique et l'ingénierie biomédicale. Il devient donc essentiel de développer des approches aptes à contrecarrer les effets néfastes de l'incertitude dans la prise de décision afin de pouvoir satisfaire les exigences de performance et de fiabilité propres aux secteurs précédemment évoqués.

Pour pouvoir remédier à l'incertitude dans la prise de décision, il est necessaire de pouvoir la modéliser convenablement. L'approche stochastique permet de considérer l'ensemble des issues possibles ainsi que leurs probabilités d'occurrence. Elle offre ainsi une compréhension plus fine des risques en associant leur gravité à leur vraisemblance, ce qui permet l'élaboration de décisions moins conservatrices. Dans ce cadre, le modélisateur fait appel à la théorie des probabilités et emploie des distributions de probabilité afin de décrire l'incertitude. Ce paradigme de modélisation de l'incertitude conduit naturellement à des problèmes d'optimisation stochastique.

Le principal défi de l'optimisation stochastique consiste à déduire des décisions avec des garanties de performance statistique fondées sur une distribution de probabilité donnée. En pratique, cependant, identifier la distribution qui décrit fidèlement un processus aléatoire est souvent difficile. Les modélisateurs s'appuient généralement sur des modèles construits à partir de données, tels que des histogrammes ou des distributions gaussiennes dont les paramètres sont inferrés. Toutefois, ces inférences ne décrivent pas toujours l'incertitude avec la précision requise. Ainsi, l'incertitude ne provient plus uniquement du caractère aléatoire des différents scénarios possibles, mais également du choix même de la distribution de probabilité. En d'autres termes, une couche plus profonde d'incertitude se déssine: une incertitude autour de l'incertitude elle même, que ce travail de thèse vise à traiter.

Cette thèse propose une nouvelle manière de modéliser l'incertitude à travers le cadre de l'optimisation distributionnellement robuste, tout en exploitant des connaissances préalables relatives à la distribution réelle. Le paradigme distributionnellement robuste considère un ensemble d'ambiguïté de distributions plausibles susceptibles d'avoir généré les données observées, et fonde la décision sur la distribution la plus défavorable au sein de cet ensemble. Plus précisément, l'objectif central de cette thèse est d'encoder

les connaissances préalables sur la structure de la distribution réelle dans les ensembles d'ambiguïté, afin d'y inclure des modèles candidats plus pertinents. Cette approche permet d'obtenir des décisions fiables, aboutissant à de meilleures performances et à une efficacité accrue.

Cette thèse comporte trois parties principales, chacune représentant un pilier interconnecté de l'approche développée:

Dans la première partie, nous introduisons des ensembles d'ambiguïté structurés, qui permettent d'exclure les modèles probabilistes impertinents de l'incertitude. Pour ce faire, nous restreignons la classe des incertitudes considérées au cas où celles-ci sont constituées de plusieurs composantes indépendantes. Les ensembles d'ambiguïté proposés sont alors structurés en fonction de ces composantes indépendantes. Cette structuration accroît significativement la pertinence des modèles inferrés et améliore la qualité des décisions associées.

La deuxième partie de cette thèse est consacrée au développement de méthodes permettant la résolution des problèmes d'optimisation distributionnellement robuste associés aux ensembles d'ambiguïté introduits. Nous mettons d'abord en évidence certaines propriétés fondamentales de ces nouveaux ensembles. Nous exploitons ensuite ces propriétés pour dériver des reformulations perméttant de calculer les solutions des problèmes d'optimisation correspondants. Enfin, nous proposons une procédure visant à atténuer la complexité numérique potentielle de ces reformulations, tout en préservant les garanties statistiques caractérisant l'approche.

Dans la dernière partie de cette thèse, nous exploitons les outils développés dans les deux premières parties afin de concevoir un contrôleur prédictif pour les systèmes linéaires soumis à des perturbations stochastiques additives dont la distribution est inconnue. Cet algorithme de commande prédictive stochastique repose sur une approche en tube et impose des contraintes probabilistes au moyen d'un resserrement approprié des ensembles admissibles d'états et de commandes. La géométrie du tube est déterminée hors ligne en résolvant plusieurs problèmes d'optimisation distributionnellement robuste partageant la même incertitude sous-jacente. Cette approche allège le problème de commande optimale de la complexité numérique associée à l'optimisation distributionnellement robuste, et permet d'aboutir à un contrôleur prédictif implémentable en pratique.

1

INTRODUCTION

In the first chapter of this thesis, we discuss the ubiquity of uncertainty in the real world and its adverse effect on decision-making. We introduce probability distributions, which are used to provide mathematical descriptions of uncertainty and examine their role in addressing decision-making problems in the presence of uncertainty. We emphasize the limitations of solely considering a single distribution to model uncertain phenomena in practice and consider distributionally robust optimization as a framework to guarantee reliability and performance when confronted with uncertain uncertainty. Finally, we introduce the specific problems and objectives that motivate the work in this thesis, and outline the structure of the thesis at the end of the chapter.

1

1.1. UNCERTAINTY: A CORE CHALLENGE AT THE HEART OF CONTROL AND MODERN TECHNOLOGIES

TECHNOLOGICAL innovation has impacted multiple aspects of modern society. Airplanes have revolutionized transportation and global connectivity, while precision manufacturing has established new standards of accuracy, enabling the production of increasingly complex and reliable systems. Advances in autonomous navigation are allowing safer and more efficient mobility, and wind turbines enhance our capacity to harness renewable energy. Together, these contributions demonstrate the significant progress our society is making.

Figure 1.1.: Instances of technological systems revolutionizing our modern society.

At the core of these innovations are systems and control, which form a discipline that analyzes the properties of dynamical systems, and develops algorithms in order to make them operate autonomously and with high performance. To achieve these goals, state-of-the-art approaches require accurate mathematical descriptions of the dynamics of the system. These approaches are strongly linked with decision-making, as they may seek the best model to accurately describe the dynamics of a given system or the best control actions to ensure its stability, safety, and performance.

In practice, mathematical descriptions of dynamical systems are idealized representations of reality that often fail to faithfully describe the exact behavior of a system. Facing uncertainty in this context is a common occurrence, which raises practical concerns regarding robustness, safety, and performance of control systems. Addressing this challenge is a key objective of many instances from systems and control, such as system identification [1], filtering [2, 3], state estimation [4–7], optimal control [8], and many others. Modelers often assume that the uncertainty is stochastic and has a specific probability distribution, which is usually uniform or Gaussian. This is the case, for instance, in LQG [9, 10], Kalman filtering [11], or stochastic model predictive control [12]. However, such assumptions naturally yield the following questions:

- How certain are we about the distribution of the uncertainty in practice (see

Figure 1.2)?

- How is it possible to address situations involving uncertainty over the true distribution?

Figure 1.2.: The model traditionally assumed for the result of a dice throw is the uniform distribution over the six possible outcomes. Yet, in case the thrown dice is similar to the one illustrated in the figure, its exact probability distribution is totally unclear. Actually, all dice have, at least, tiny imperfections impacting the likelihood of their outcome. Thus, we can never know the true distribution describing the results of a dice throw.

In this thesis, we consider decision-making problems in the presence of uncertainty with an unknown probability distribution. To model it, we infer from data a set of plausible probability distributions and account for the most adverse model in that range for decision-making. In order to reduce excessive conservativeness, we account for prior knowledge over the true distribution by structuring the class of considered candidate models. In the following sections, we provide further insights into decision-making problems and the difficulties arising when uncertainty is involved. We introduce the mathematical tools used to model uncertainty, along with some data-driven methods for inferring unknown probability distributions. Finally, we focus on the research objectives driving the core of this work.

1.2. UNCERTAINTY IN DECISION-MAKING

Uncertainty is the state characterizing situations with imperfect or unknown information, or the inability to exactly predict a process outcome. It is omnipresent in many aspects of our daily life and poses significant challenges in multiple areas with vital impacts on our societies. Among these areas, we have technical fields involving systems with an increasing complexity that contributes to magnifying the sources and effects of uncertainty. At the same time, expectations over performance, efficiency, and reliability of these systems are increasingly elevated. The presence of uncertainty represents a serious challenge in achieving such standards rendering decision-making under uncertainty a longstanding research problem.

In addition to systems and control, dealing with the adverse effects of uncertainty is also central in various other domains, such as healthcare [13], energy systems [14, 15], power dispatch [16] and management [17], portfolio management [18], economics [19], supply chain [20, 21], and meteorology [22].

The central problem in all the instances that were cited so far is often alluded to as *decision-making*, e.g., model parameter selection or control gain selection, *in the presence of uncertainty*. In mathematics, we also refer to decision-making problems as optimization problems. These problems consist of finding the best value for a set of variables, i.e., the optimal value, minimizing a given loss function. A mathematical optimization problem typically takes the form

$$\min_{x \in \mathscr{X}} \quad f(x) \tag{1.1.1a}$$

$$\text{subject to} \quad g_i(x) \leq 0 \qquad i \in [I], \tag{1.1.1b}$$

(see [23, Section 1.1]). Here, $\mathscr{X}$ is the decision space where the decision variable x lives. The function f is the loss function to be minimized. It can represent, for example, the cost of an operation, the prediction error of a model, or the time, energy, or fuel needed to conduct a specific task. This mathematical optimization problem is subject to a set of $I \in \mathbb{N}$ constraints determined by the constraint functions g_i in (1.1.1b). These constraints can express, for instance, the maximum cash flow available to manage a portfolio, the safe domain of operation of a system, or the maximum allowable actuation of a control system. Namely, the constraints (1.1.1b) delineate a subset of admissible values of $\mathscr{X}$ to which the optimal decision must belong.

1.3. DECISION-MAKING WITH UNCERTAIN OBJECTIVES OR CONSTRAINTS

When facing uncertainty, the loss and/or constraint functions may depend on an uncertain parameter ξ. Hence, instead of fixed functions, we deal with families $f(x,\xi)$ and $g_i(x,\xi)$, $i \in [I]$, of uncertain loss and constraint functions, respectively, which are parametrized by the uncertain parameter ξ. Thus, every value of ξ yields a different loss and different constraint functions, and it is not clear which of these possibilities to account for when solving the optimization problem, as illustrated in Figure 1.3a. Exploiting these uncertain functions in a decision-making problem requires having at disposal a description of the uncertain parameter ξ and the way it impacts the cost and constraints.

1.3.1. ROBUST OPTIMIZATION

Two main approaches may be considered in order to model uncertainty. The first one consists of addressing its adverse effect by considering the worst-case realization of the parameter ξ. This is known as the robust approach, and it leads to robust optimization problems [24–26]. In this framework, the decision is made against the scenarios yielding the worst-case loss

$$\min_{x \in \mathscr{X}} \quad \max_{\xi \in \Xi} f(x,\xi), \tag{1.2.1a}$$

depicted in Figure 1.3a, where the set Ξ spans the whole range of possible values for ξ. Solving the optimization problem (1.2.1a) provides decisions with robust loss

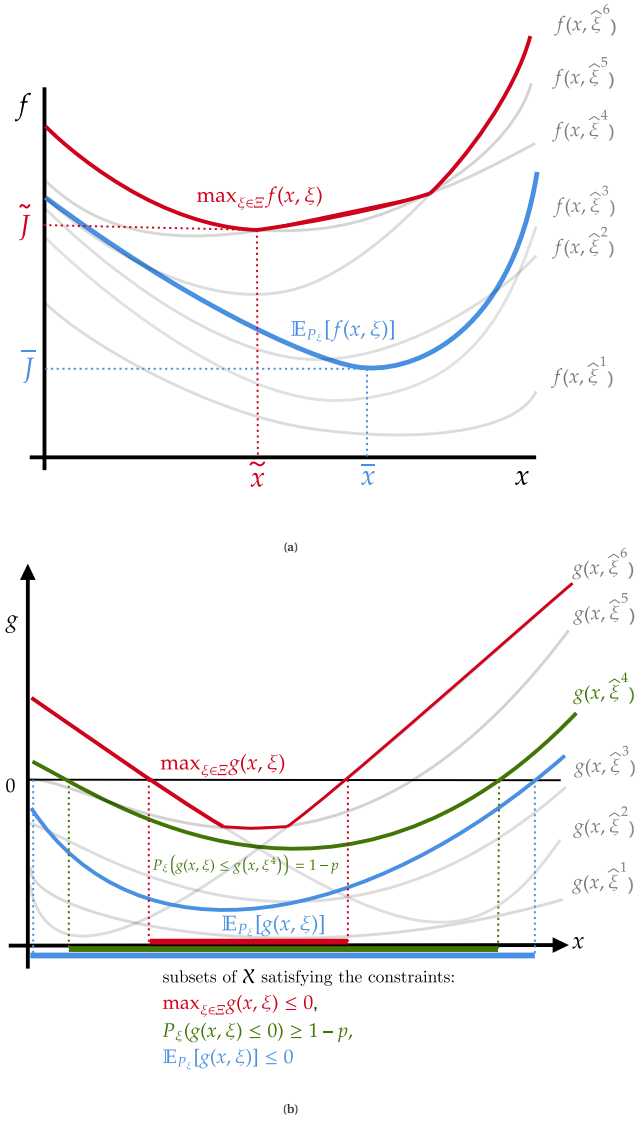


Figure 1.3.: (a) illustrates an uncertain loss function $f(x,\xi)$ where the uncertain parameter ξ can take 6 distinct values. We depict the worst-case loss in red and the expected loss in blue, together with their associated optimizers $\tilde{x}$ and $\bar{x}$. $\tilde{J}$ represents the solution of (1.2.1a) while $\bar{J}$ represents the solution of (1.2.1a).
(b) illustrates the uncertain constraint function $g(x,\xi)$, worst-case and expected values are depicted in red and blue, respectively. The green curve $g(x,\xi^4)$ represents the threshold ensuring constraint satisfaction with the desired probability $1-p$. The set of decisions x satisfying each constraint is depicted with the color associated with its respective constraint. We can observe that the worst-case approach is more restrictive than the chance constraints approach.

certificates. Namely, referring to the optimizer of (1.2.1a) by $\tilde{x}$ and denoting its associated cost by $\tilde{J}$, then $f(\tilde{x},\xi) \leq \tilde{J}$ holds for any realization of ξ. This reasoning can also be used to deduce decisions guaranteeing constraint satisfaction for any $\xi \in \Xi$, when constraints are affected by the uncertain parameter, i.e., by imposing the robust constraints

$$\max_{\xi\in\Xi} g_i(x,\xi) \leq 0 \qquad\qquad i \in [I], \tag{1.2.1b}$$

depicted in Figure 1.3b.

1.3.2. STOCHASTIC OPTIMIZATION

Solely considering worst-case realizations of the uncertainty may have the drawback of yielding overly conservative decisions, which may lead to poor performance, as illustrated in Figure 1.3. In particular, robust decisions may not always be desirable, as the worst-case scenario may be very unlikely to occur. In this case, the conservativeness can be addressed by favoring more plausible realizations, which leads to solutions over larger feasible sets and with lower average costs. This is known as the stochastic approach, and it yields stochastic optimization problems [27].

Modeling stochastic uncertainty hinges on assigning to each range of uncertain parameter values the probability that their outcome will lie in that range. This is mathematically formulated in probability theory [28], which analyzes the concept of random experiments. Therein, stochastic uncertainty is modeled through random variables, whose values follow a certain probability distribution (see Figure 1.4).

In stochastic optimization, uncertain loss or constraint functions, i.e., with random variables in their arguments, are considered. Knowing the probability distribution P_ξ of the random variable ξ, we may, for instance, determine the decision that yields the best cost on average by solving the stochastic program

$$\min_{x\in\mathscr{X}} \ \mathbb{E}_{P_\xi}[f(x,\xi)] \tag{1.2.1a}$$

(see Figure 1.3a). Similarly, we may also treat stochastic constraints involving uncertain functions $g_i(x,\xi)$ (see Figure 1.3b). In this case, we may want to enforce constraint satisfaction with a specified probability threshold $1-p$, by imposing the chance constraints

$$P_\xi\big(g_i(x,\xi) \leq 0\big) \geq 1-p \qquad\qquad i \in [I], \tag{1.2.2b}$$

[27, Chapter 4], [29], or to guarantee constraint satisfaction on average, by imposing

$$\mathbb{E}_{P_\xi}[g_i(x,\xi)] \leq 0 \qquad\qquad i \in [I], \tag{1.2.2b'}$$

[30, 31] (cf. Figure 1.3b). Risk-averse optimization can also be considered by adopting appropriate risk measures, such as the variance [32, 33], semi-deviations [34, 35], or the conditional value at risk [36]. A detailed discussion about risk-averse optimization is provided in [27, Chapter 6].

As an illustrative stochastic optimization problem, consider a simple power dispatch problem, in which a random daily power demand δ must be met. Ideally,

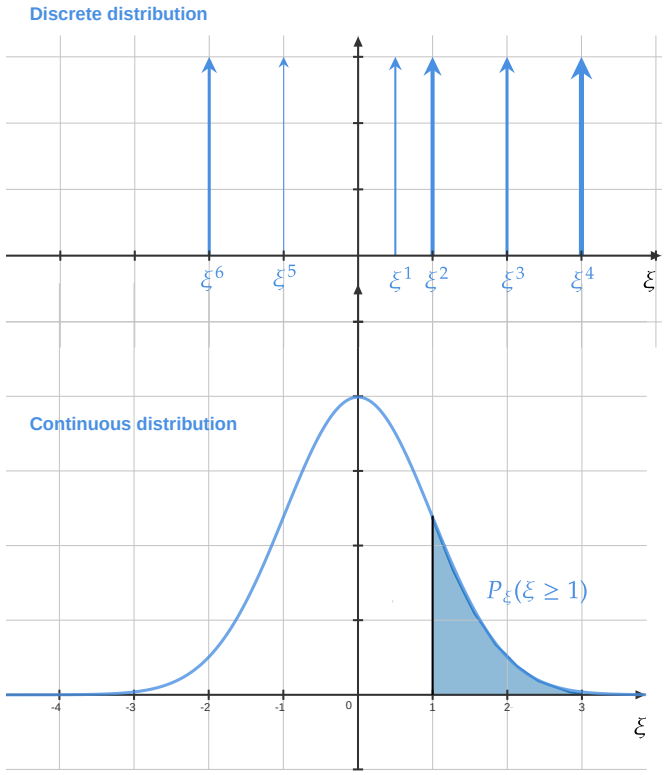


Figure 1.4.: The figure shows two examples of probability distributions. On top, a discrete distribution consisting of 6 atoms is depicted. The width of the arrows represents the probability mass of each outcome. The bottom depicts the well-known Gaussian distribution. Continuous distributions represent probability densities, where the area below their curves quantifies the probability that an event happens in the corresponding region of the ξ axis. For example, in this figure, the numerical value of the blue area quantifies the probability that $\xi \geq 1$.

this is achieved using renewable power ρ, which itself is uncertain due to its dependence on the weather conditions. Since both ρ and δ are random variables, renewable power alone cannot always suffice to meet the demand. To ensure demand coverage with a high confidence $1-p$, we introduce the decision variable x, representing the minimum additional power purchased from the market. By grouping the uncertain quantities into the random vector $\xi \equiv (\xi_1, \xi_2) := (\rho, \delta)$ with probability distribution P_ξ, the problem can be formulated as the following stochastic program

$$\begin{aligned} &\min_{x \geq 0} \quad x \\ &\text{subject to} \quad P_\xi(\xi_2 - \xi_1 - x \leq 0) \geq 1 - p. \end{aligned}$$

Another instance where stochastic optimization plays a fundamental role in is stochastic model predictive control [12]. There, we consider discrete-time stochastic

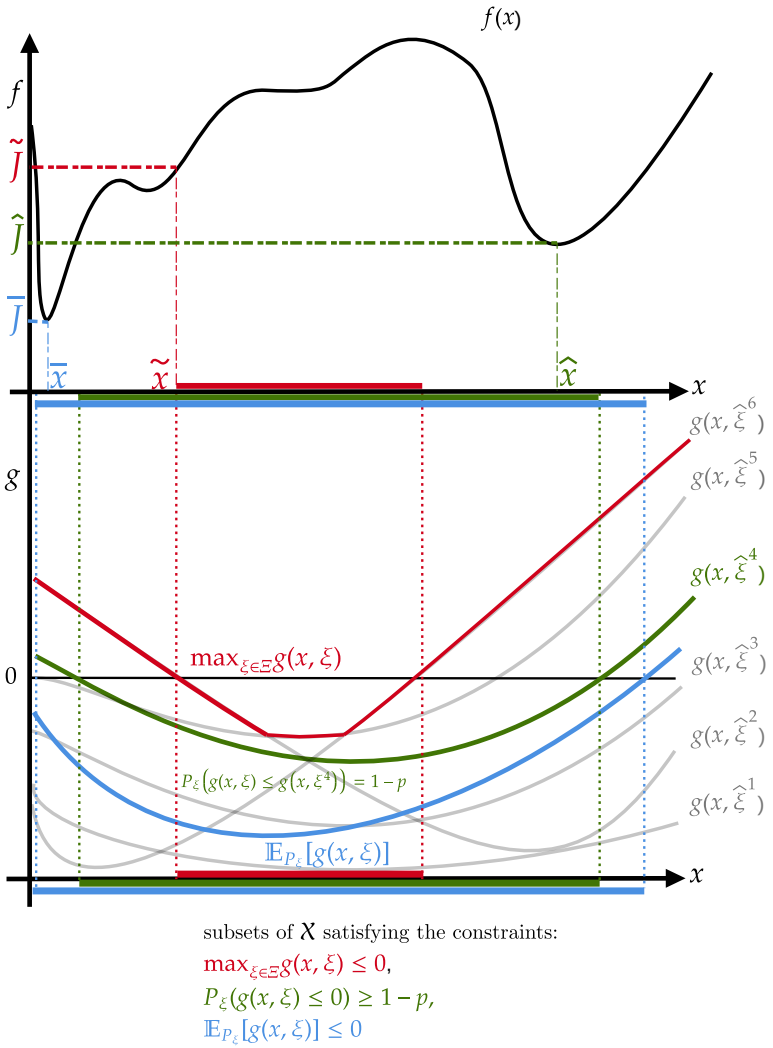


Figure 1.5.: The figure depicts the feasible domains for each considered type of constraints when the constraint functions depend on an uncertain parameter ξ. We can see that the worst-case approach, in red, leads to conservative solutions, while accounting for the stochastic nature of the uncertainty, through average constraints, in blue, or probabilistic constraints, in green, typically induces larger feasible domains and yields decisions with improved performance (lower cost) on average, or with desired probability.

systems with dynamics

$$x^+ = f(x,u,w), \tag{Σ}$$

where x is the state, u the control input and w is a stochastic disturbance with probability distribution P_w. Our typical goal is to steer the state x of (Σ) towards some reference $r \in \mathscr{X}$ while enforcing constraints over the state and control input at

every time step $t \geq 0$. This yields the recursive stochastic optimization problem

$$\begin{aligned} \min_{u_k \in \mathcal{U},\, k \in [T-1]_0} \quad & \mathbb{E}_{P_w}\left[\sum_{k=0}^{T-1} J_k(x_k - r, u_k, w_k)\right] \\ \text{subject to} \quad & P_\xi(x_k \in \mathcal{X}) \geq 1 - p_x \qquad\qquad k \in [T]_0 \\ & x_0 = x(t), \end{aligned}$$

where $J_k : \mathcal{X} \times \mathcal{U} \times \mathcal{W} \to \mathbb{R}^+$ is a stage cost function, $x(t)$ represents the state of the system measured at time t and x_{k+1} is recursively defined by

$$x_{k+1} = f(x_k, u_k, w_k) \qquad\qquad k \in [T-1]_0.$$

In general, the approach that is selected to address uncertainty in the constraints, namely, the introduction of robust constraints, expected value constraints, or chance constraints, has a direct impact on the size of the feasible set of the decision variable x. This directly impacts the conservativeness of the optimal solution as depicted in Figure 1.5. In addition, these approaches (robust, expected value and chance constraints) also have a direct effect on the tractability of their associated optimization problem, as discussed in [37], or [25, Chapter 4]. Having demonstrated that uncertainty in decision-making can be effectively addressed through appropriate mathematical models, we discuss next data-driven approaches for inferring probability distributions of uncertain quantities.

1.4. DATA-DRIVEN APPROACHES TO UNCERTAINTY MODELING AND DECISION-MAKING

In order to provide reliable decisions for stochastic optimization problems, it is essential to know the probability distribution of the uncertainty. In practice, however, such descriptions are not always available and capturing the distribution of the uncertainty necessitates large amounts of accurate data. To this end, we discuss data-driven methods to infer probability distributions of random variables and their limitations when exploited in stochastic programming. We also discuss a data-driven approach that solves stochastic optimization problems with rigorous guarantees, while sidestepping the requirement to infer a model of the uncertainty distribution.

1.4.1. DATA-DRIVEN DISTRIBUTION ESTIMATION

Data-driven distribution estimation is the field that aims at determining probabilistic models for random variables by exploiting their realizations [38, 39]. This may also be referred to as distribution fitting, e.g. [40, 41], or density estimation, e.g. [42–44], and can be addressed using various methods. These methods can be mainly classified into three distinct families, which are parametric, non-parametric, and semi-parametric methods.

The first family consists of the *parametric methods*. These methods allow for the inference of parametric distributions. More specifically, given a set of uncertainty

realizations, we seek the best candidate within a finite-dimensional parameterization of the distribution class that could have generated the data at disposal. The parametrization of the distribution class usually reflects the shape or type of the data-generating distribution. Data are exploited in order to estimate the optimal value of the parameters. Among this family, we have the maximum likelihood estimation (MLE) approach, where parameter values maximizing the likelihood of the observed data are determined (see e.g. [45–47]). The MLE is an asymptotically consistent and efficient estimator that converges in probability to the true parameters of the probability distribution, under mild conditions on the model identifiability, the parameter space, or on the model structure (see [48] or [49, Section 5.5]). It also achieves finite-sample error bounds for sufficiently regular densities (e.g., [50] studies this case for logistic regression). Another fundamental parametric method is Bayesian inference, where the true distribution parameters are treated as random variables. The prior distribution over the parameters is updated using observed data in order to obtain a posterior distribution, which is an improved estimate of the uncertainty (see [51, 52]).

The second family are the *non-parametric methods*, where no specific parametrization of the data-generating distribution is required. These methods seek to infer the best candidate model from an infinite-dimensional distribution class. This has, for instance, the advantage of not requiring assumptions about the shape of the underlying distribution. Among this family, we have data-driven histogram estimation, which is a simple density estimation method where historical realizations of the random variable are discretized into bins (intervals), counting the frequency of observations in each one of them (see [53, 54]). This estimator is consistent but sensitive to bin choice, which affects the quality of the estimated distribution, (see [53, 55]). Alternatively, historical realizations of the random variable can be used to infer the empirical distribution estimator, which is formed by considering the uniform distribution over these realizations (see [56, Chapter 2]). This estimator is unbiased and converges uniformly to the true distribution almost surely (see [56, Glivenko-Cantelli Theorem]). In addition, non-asymptotic results are also available for this estimator (see [57–59]). Another popular non-parametric distribution inference method is kernel density estimation. In this approach, the density is estimated by summing the contributions of smooth kernels placed at the samples (see [44, 60]).

The third family comprises methods combining both parametric and non-parametric features, and are thus referred to as *semi-parametric methods* (see [61, section 5.1.4]). It includes semi-parametric mixture models, where data are modeled as a weighted sum of multiple distributions [62, 63], or Copulas, which are used to describe the dependency structure between random variables, separating the marginal distributions from their joint dependence (see [64]).

The methods we have discussed in this part allow to infer the probability distribution of the random variable based on available realizations of the uncertainty. Then the deduced probabilistic model can be used in (1.2.1a), (1.2.2b), or (1.2.2b') as a data-driven proxy for the true distribution P_ξ. For instance, the empirical estimator yields the sample average approximation (SAA) of the cost (see [65]). The

SAA cost has an asymptotically normal distribution. Its properties are discussed in detail in [27, Chapter 5].

Despite the benefits of these approaches, they are also characterized by limitations that can be critical in practice. To obtain accurate models of the true distribution, inference methods require a large amount of ideal samples, which are not always guaranteed. This may negatively impact the validity of the inferred model (see e.g., [66] for the MLE and its implications in stochastic optimization). For instance, the SAA is an over-promising approach that yields optimistic decisions, i.e., the SAA cost is on average a downward bias of the true cost (see [27, Proposition 5.6, page 163]).

1.4.2. THE SCENARIO APPROACH: A DISTRIBUTION-FREE METHOD

The scenario approach is a data-driven method that directly handles uncertainty in decision-making problems without inferring the underlying distribution. This method provides decisions based on a finite number of realizations of the uncertainty or scenarios. In its simplest formulation, this approach optimizes against the worst-case available scenario. For instance, given the uncertain function $f(x,\xi)$ and the N scenarios $\xi^1,\dots,\xi^N$, we may formulate the scenario program

$$\min_{x\in\mathscr{X}} \max_{l=1,\dots,N} f(x,\xi^l) \tag{1.3.1a}$$

(cf. [67] and Figure 1.6). The same approach can be employed to address constraints involving uncertain functions through the scenario program

$$\begin{aligned} &\min_{x\in\mathscr{X}} && f(x) \\ &\text{subject to} && \max_{l=1,\dots,N} g_i(x,\xi^l) \le 0 \qquad i\in[I],\ l\in[N]. \end{aligned} \tag{1.3.1b}$$

The scenario approach can be very effective in order to address stochastic uncertainty with distribution-free guarantees. For a sufficient number of samples, it can yield decisions with a prescribed confidence level without explicitly relying on the uncertainty distribution (see [67, Theorem 1]). Namely, it can either provide an optimizer with a certificate that holds with the desired confidence, or it allows to determine regions ensuring constraint satisfaction with prescribed probability (see [68]). Besides this, it is capable of specifying tight feasibility regions for chance-constraints problems [69], that can even be exact [70].

In order to deduce reliable decisions and determine safety domains, the scenario approach requires a minimum number of samples to guarantee some desired confidence level. Again, we are not always guaranteed to have access to enough samples in practice. In addition, the samples may be corrupted, which can annihilate the probabilistic guarantees of the approach, rendering it impractical.

1.5. THE DISTRIBUTIONALLY ROBUST FRAMEWORK

In this thesis, we are particularly interested in making reliable decisions in the presence of uncertainty, based on a fixed number of data. When the number

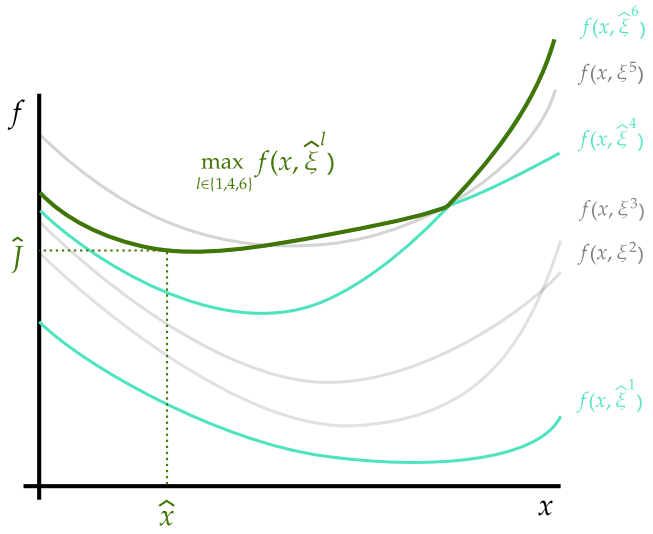


Figure 1.6.: The figure illustrates an uncertain loss function $f(x,\xi)$, where the uncertain parameter ξ can take 6 distinct values. The graphs in turquoise depict the functions resulting when the parameter ξ takes the values $\{\hat{\xi}_1,\hat{\xi}_4,\hat{\xi}_6\}$, which are known from sampling the random variable ξ. The curve in green depicts the function that is minimized in the scenario program (1.3.1a), for the scenarios $\{\hat{\xi}_1,\hat{\xi}_4,\hat{\xi}_6\}$, while $\hat{x}$ represents the optimizer and $\hat{J}$ its associated cost.

of available samples is small, it is no longer guaranteed that the inferred model represents the true distribution well. This situation induces in decision-making problems an extra layer of uncertainty, which is uncertainty over the distribution. To address this issue, modelers may exploit distributionally robust optimization (DRO), which is a decision-making framework under uncertainty where the underlying probability distribution is either unknown or has unknown parameters.

Instead of assuming a single probability distribution, the DRO paradigm considers a family of plausible distributions called ambiguity set. Then, decisions are made against the worst-case distribution inside the ambiguity set at hand, thus hedging against distributional mismatches. Namely, given an ambiguity set $\mathscr{P}$ of plausible distributions, we are interested in solving the distributionally robust counterpart

$$\min_{x\in\mathscr{X}} \sup_{P\in\mathscr{P}} \mathbb{E}_P[f(x,\xi)] \tag{1.4.1a}$$

of the stochastic program (1.2.1a), or in enforcing the constraint

$$\inf_{P\in\mathscr{P}} \big(P\left(g(x,\xi)\leq 0\right)\big) \geq 1-p, \tag{1.4.1b}$$

which is the distributionally robust counterpart of (1.2.2b). Various methods can be considered to build ambiguity sets of probability distributions. For instance, they can be constructed based on moments, i.e., by considering all distributions satisfying some moments constraints [71–74]. Another popular approach consists of

grouping all the distributions within some discrepancy gap from a baseline model. This distributional discrepancy can be measured based on statistical divergences, like the Kullback-Leibler divergence [75], which is a particular case of ϕ-divergences (see [76]). It is also possible to quantify the discrepancy between distributions by exploiting metrics like the total variation [77], or the Wasserstein distance [78]. We are particularly interested in the Wasserstein distance, due to several convenient properties of Wasserstein ambiguity sets.

1.5.1. THE WASSERSTEIN DISTANCE

The Wasserstein distance is a measure of discrepancy between probability distributions. For any two probability distributions P, Q on $\mathbb{R}^n$ with finite pth moment, their Wasserstein distance of order $p \in [1,\infty)$ is defined as

$$W_p(Q,P) := \left(\inf_{\pi \in \mathscr{C}(Q,P)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|\zeta - \xi\|^p d\pi(\zeta,\xi) \right)^{1/p},$$

where $\mathscr{C}(Q,P)$ is the set of transport plans between P and Q, i.e., distributions on $\mathbb{R}^n \times \mathbb{R}^n$ with marginals Q and P, respectively.

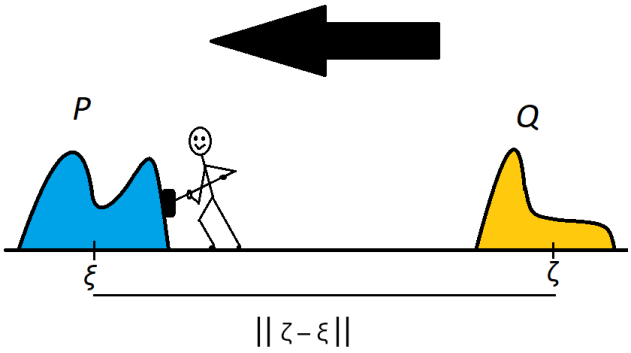


Figure 1.7.: The Wasserstein distance between P and Q is the minimum cost to transport the blue pile of sand into the yellow one.

Intuitively, if we picture the distributions Q and P as piles of sand and consider $\|\zeta - \xi\|^p$ as the cost needed to transfer one unit of sand from location ζ to location ξ, $W_p^p(Q,P)$ represent the minimum cost for transporting one pile into the other (cf. Figure 1.7). The Wasserstein distance naturally captures the geometry of the underlying uncertainty space [78, page 99], since its definition is based on the distance function of the space.

1.5.2. WASSERSTEIN AMBIGUITY SETS

Given a reference distribution Q (that may be inferred from data), we can form the Wasserstein ambiguity ball

$$\mathscr{B}_p(Q,\varepsilon) := \left\{ P \in \mathscr{P}_p(\Xi) : \; W_p(Q,P) \leq \varepsilon \right\}, \tag{1.5.1}$$

which is centered at Q and consists of all the distributions P that are ε close to Q in the Wasserstein metric, as depicted in Figure 1.8.

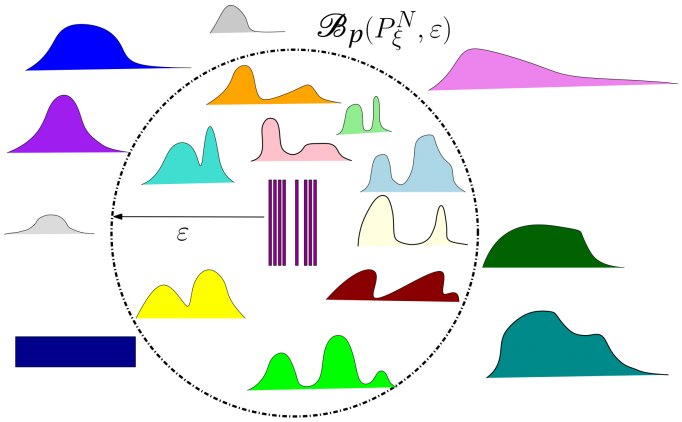


Figure 1.8.: A pictorial representation of a Wasserstein ambiguity ball centered at the empirical distribution $P_\xi^N := \frac{1}{N}\sum_{i=1}^N \xi^i$. Plausible models for the data-generating distribution are grouped up to an ε distance from the empirical model, excluding irrelevant models.

Wasserstein ambiguity balls enjoy several benefits. They are fully nonparametric and can simultaneously capture all types of distributions, e.g., with multiple modes and of any regularity, irrespectively of the nature of the distribution in their center. This facilitates the construction of ambiguity balls for data-driven problems with a limited number of uncertainty samples. Specifically, when centered at the empirical distribution of the samples, these ambiguity balls are accompanied by rigorous finite-sample guarantees of containing the true distribution with prescribed confidence [59]. Moreover, they ensure that the size of the ambiguity set converges to zero as the number of samples goes to infinity. This family of ambiguity sets also leads to tractable DRO problems [79].

These convenient properties have led to an increasing interest in exploiting Wasserstein DRO for decision problems involving uncertainty with imprecise probability distributions. Therefore, Wasserstein DRO has been widely adopted in a wide range of applications. Specific instances include machine learning, where it can be exploited to hedge against adversarial attacks [80], supply chain design, to hedge against disruption scenarios and uncertain customer demands [81], and unit commitment in power systems, to manage risk from uncertain wind power forecasted errors [82]. Other applications include motion planning in dynamical environments [83], and model predictive control [84].

1.5.3. THE CURSE OF DIMENSIONALITY

Concentration of measure results can be used to tune the radius of the Wasserstein ball (1.5.1) so that it contains the data-generating distribution with prescribed confidence. To be more specific, consider the random variable $\xi \in \mathbb{R}^d$ and assume we have access to N i.i.d. samples $\xi^1, \ldots, \xi^N$. Then, exploiting prior assumptions

about the class where the unknown distribution belongs, it is possible to bound the Wasserstein distance between the empirical distribution $P_\xi^N := \frac{1}{N}\sum_{i=1}^N \delta_{\xi^i}$ and the data-generating distribution P_ξ with high confidence. Namely, for any confidence level $1-\beta$, we can select the ambiguity radius $\varepsilon(N,\beta)$ so that

$$\mathbb{P}(W_p(P_\xi^N, P_\xi) \le \varepsilon) \ge 1-\beta.$$

The radius of the ambiguity ball typically takes the form

$$\varepsilon(N,\beta) = K\frac{1}{N^{1/\max\{d,2p\}}}, \tag{1.5.2}$$

[59], where d is the dimension of the random vector ξ, p is the selected order of the Wasserstein distance, and K is a constant that depends on p and the class where P_ξ belongs (e.g., the size of its support). The expression (1.5.2) suggests that exploiting more samples for high-dimensional random vectors does not guarantee bringing the empirical distribution P_ξ^N significantly closer to the data-generating distribution P_ξ. As a consequence, the size of the ambiguity ball $\mathscr{B}_p(P_\xi^N,\varepsilon)$ that is necessary to contain P_ξ remains large and cannot be significantly reduced by exploiting extra samples. This curse of dimensionality characterizing Wasserstein ambiguity balls negatively impacts DRO as it may lead to overly conservative decisions that yield poor performance.

In order to address the curse of dimensionality, recent work no longer requires the ambiguity set (1.5.1) to contain the true distribution with prescribed confidence and directly informs its radius from the optimization problem. This results in a decay rate of the order $1/\sqrt{N}$ for the radius that is independent of the dimension d of the random vector, while still ensuring reliable decisions [85–87]. As we shall shortly argue, despite these efforts, the curse of dimensionality still persists when solving multiple DRO problems, of the form (1.4.1a), or when enforcing multiple constraints of the form (1.4.1b), with the same underlying uncertainty.

To clarify this, assume that we inform the radius of the ambiguity set by a (single) optimization problem. By setting $\varepsilon = O(1/\sqrt{N})$, we can ensure that the distributionally robust expected cost of (1.4.1a) associated with the ambiguity ball (1.5.1) is larger than its stochastic counterpart (1.2.1a) with high confidence (see e.g., [87]). When dealing with T different optimization problems, the only straightforward way to derive similar statistical guarantees for the solutions of T different problems simultaneously is to exploit union bounds to tune the radius ε_k for each problem. In particular, assuming we pick the radii ε_k, $k \in [T]$, such that all the T desired inequalities hold individually with confidence $1-\beta$. Then, it follows from the union bound argument that the only clear lower bound for the probability that the T distributionally robust solutions upper-bound their stochastic counterpart is $1-T\beta$. Therefore, for large T, this approach leads to conservative guarantees that essentially cannot alleviate the curse of dimensionality. On the other hand, if the ambiguity set contains the true distribution with confidence $1-\beta$, the desired performance guarantee will also hold for all optimization problems simultaneously with the same confidence.

There are several instances where addressing multiple optimization problems is apparent. These include model predictive control [84, 88], controller synthesis for reach avoid specifications [89], multi-objective planning [90], distributed systems [91], or general dynamic programming [92]. Therefore, our goal is to address the curse of dimensionality when multiple optimization problems with the same uncertainty are involved.

1.6. RESEARCH SCOPE AND OBJECTIVES

In this thesis, we want to address the curse of dimensionality characterizing Wasserstein DRO when dealing with multiple DRO problems with the same underlying uncertainty. Since it is no longer possible in this case to directly inform the ambiguity radius by the optimization problem, we consider ambiguity sets that are accompanied by the requirement of containing the data-generating distribution with high confidence. Still, in general this requirement confronts us again with the excessively slow contraction rate (1.5.2).

The main idea to overcome this problem is to exploit prior information about the structure of the true distribution to restrict the class of plausible uncertainty models, and thus, infer tighter ambiguity sets. More specifically, we consider the case where the random variable consists of multiple independent components, which implies that the underlying probability distribution is necessarily a product measure. This assumption is commonly employed in applications, such as networked systems, where random inputs at different network locations are essentially independent [93], the deployment of multi-robot systems, where each agent is subjected to independent disturbances [94], or in stochastic model predictive control, where stochastic disturbances are often assumed to be independent through time [12]. To capture this prior information, we want to build alternative ambiguity sets, which we name structured ambiguity sets. These sets account for the particular structure of the true distribution by only considering distributions that are appropriately close to some product reference distribution. To this end, we identify the following objectives, which drive the development of this thesis.

- O1 Equip data-driven ambiguity sets with the independence structure of the data-generating distribution.
- O2 Establish statistical guarantees for these structured ambiguity sets to contain the true distribution with high confidence and determine how their shrinkage rate improves with the number of samples.
- O3 Derive duality results for DRO problems over structured ambiguity sets to convert infinite-dimensional optimization problems over probability distributions into finite-dimensional programs.
- O4 Derive tractable reformulations for distributionally robust optimization problems associated with the developed structured ambiguity sets.
- O5 Identify and address computational limitations of these structured ambiguity sets.

O6 Exploit structured ambiguity sets to robustify data-driven decision problems in systems and control.

1.7. THESIS ORGANIZATION

The thesis is composed of three main chapters, and their contributions are outlined below.

1.7.1. CHAPTER 2: STRUCTURED AMBIGUITY SETS FOR DISTRIBUTIONALLY ROBUST OPTIMIZATION

In the second chapter of the thesis, we address the excessively slow contraction rate of Wasserstein ambiguity balls with respect to the number of collected samples for high-dimensional random variables, by proposing alternative ambiguity sets. The contributions of this chapter are listed below.

- **Development of structured ambiguity sets**

  We develop two new classes of ambiguity sets that incorporate the independence structure of the uncertainty random variable. This structure is enforced within the ambiguity sets by considering multiple optimal transport constraints, which achieves the objective O1. The first class, called Wasserstein hyperrectangles, consists only of product distributions, which makes them non-convex. The second class, called multi-transport hyperrectangles, is a convex relaxation of the former which enjoys similar size-reduction properties as data-driven Wasserstein hyperrectangles.

- **Statistical guarantees of structured ambiguity sets**

  We show that the introduced structured ambiguity sets can be tuned to contain the true distribution with high confidence when their reference distribution is built by i.i.d. samples of the true distribution. This contribution establishes the first part of objective O2.

- **Improved contraction rate**

  We address the second part of O2 by demonstrating that structuring optimal transport ambiguity sets according to the independent components of the random variable accelerates their contraction rate with respect to the number of samples (as depicted in Figure 1.9).

- **Comparison with traditional monolithic ambiguity balls**

  In order to assess the size reduction of the newly introduced hyperrectangles compared to their traditional monolithic counterpart, we consider the smallest ambiguity ball enclosing each hyperrectangle and compare its radius with that of the traditional Wasserstein ball containing the data-generating distribution at the same confidence level, thereby fulfilling objective O2. This way, we establish that hyperrectangles contract at a significantly faster rate compared to their monolithic counterparts. In numerical simulations, we compare

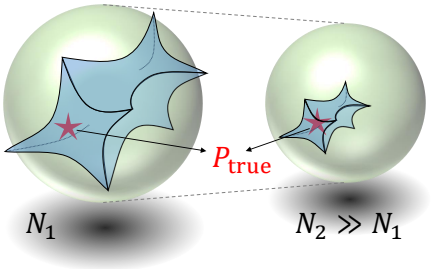


Figure 1.9.: High-dimensional hyperrectangles shrink much faster with the number of samples compared to Wasserstein ambiguity balls.

the decisions informed by each ambiguity set when calibrated to guarantee robustness with a specified confidence level.

- **Dual reformulations**

  We provide dual reformulations for both Wasserstein and multi-transport hyperrectangles and prove that dual optimizers are attained. Due to the non-convexity of Wasserstein hyperrectangles, they only admit dual reformulations for a narrow class of objective functions, whereas multi-transport hyperrectangles admit dual reformulations for a much broader class of objective functions. This achieves objective O3.

This chapter is based on the preprint:
L. M. Chaouach, T. Oomen, and D. Boskos. *Structured ambiguity sets for distributionally robust optimization.* Provisionally accepted for publication in *European Journal of Control*. 2023. arXiv: 2310.20657 [math.OC].

1.7.2. CHAPTER 3: TRACTABLE REFORMULATIONS FOR DRO PROBLEMS OVER STRUCTURED OPTIMAL TRANSPORT AMBIGUITY SETS

The third chapter of this dissertation focuses on developing tractable reformulations for distributionally robust optimization problems associated with multi-transport hyperrectangles. The contributions of this chapter are:

- **Fundamental properties of hyperrectangles**

  We establish fundamental properties characterizing multi-transport hyperrectangles in terms of weak compactness and continuity of expected losses over the distributions in the ambiguity set. We also provide conditions that guarantee finiteness of worst-case expectations over the multi-transport hyperrectangles.

- **Tractable reformulations for DRO problems over multi-transport hyperrectangles**

  Exploiting our duality results, we derive tractable reformulations of distributionally robust losses for various classes of DRO problems associated with multi-transport hyperrectangles, thereby fulfilling objective O4.

- **Development of clustering schemes to address potential numerical complexity**

  A key feature that contributes to the reduced size of the hyperrectangles while maintaining high confidence levels of containing the true distribution arises from shifting their center from the classical empirical distribution to the so-called product empirical distribution. Nevertheless, the product empirical distribution may consist of a tremendous number of atoms. This turns the manifestation of the curse of dimensionality from the excessively slow contraction rate of ambiguity sets into the intractable numerical complexity of the derived reformulations. To address the numerical complexity of the reformulations, which grows polynomially with the number of atoms of the reference distribution of the hyperrectangle, we employ clustering schemes. These schemes allow the construction of reference distributions that yield reformulations of a manageable complexity. This contribution achieves Objective O5.

This chapter is based on:
L. M. Chaouach, T. Oomen, and D. Boskos. *Tractable reformulations of DRO problems over structured optimal transport ambiguity sets.* Accepted for publication in *Transactions on Automatic Control*. 2025. arXiv: 2504.06966 [math.OC],
and partly on:
L. M. Chaouach, T. Oomen, and D. Boskos. "Comparing Structured Ambiguity Sets for Stochastic Optimization: Application to Uncertainty Quantification". In: *IEEE Int. Conf. on Decision and Control*. 2023, pp. 8274–8279.

1.7.3. CHAPTER 4: DISTRIBUTIONALLY ROBUST MODEL PREDICTIVE CONTROL WITH HORIZON ADAPTIVE AMBIGUITY SETS

In the fourth chapter of the dissertation, we fulfill Objective O6 by exploiting the tools developed in Chapters 2 and 3 to design a model predictive controller for linear systems subject to additive stochastic disturbances with an unknown probability distribution while accounting for the independence structure of the uncertainty across the prediction horizon. The contributions of this chapter are:

- **Formulation of a distributionally robust stochastic model predictive controller**

  We design a model predictive control algorithm for linear systems subject to additive stochastic disturbances with an unknown probability distribution. We enforce chance constraints by computing a suitable tightening of the admissible sets offline. In order to do so, we replace classical chance constraints by distributionally robust CVaR constrained programs, which we reformulate into convex programs that are, in turn, solved offline to determine the suitable tightening of the admissible sets across the prediction horizon. This relieves the receding-horizon optimal control problem from the computational burden typically associated with DRO. The proposed approach yields a stochastic MPC algorithm with the same complexity as nominal MPC, thereby facilitating practical implementations of the algorithm.

- **Ensuring recursive feasibility of the proposed MPC algorithm**

  We provide conditions ensuring recursive feasibility of the associated receding horizon problem, which is a fundamental requirement for any model predictive controller.

2

STRUCTURED AMBIGUITY SETS FOR DISTRIBUTIONALLY ROBUST OPTIMIZATION

Distributionally robust optimization (DRO) incorporates robustness against uncertainty in the specification of probabilistic models. This chapter focuses on mitigating the curse of dimensionality in data-driven DRO problems with optimal transport ambiguity sets. By exploiting independence across lower-dimensional components of the uncertainty, we construct structured ambiguity sets that exhibit a faster shrinkage as the number of collected samples increases. This narrows down the plausible models of the data-generating distribution and mitigates the conservativeness that the decisions of DRO problems over such ambiguity sets may face. We establish statistical guarantees for these structured ambiguity sets and provide dual reformulations of their associated DRO problems for a wide range of objective functions. The benefits of the approach are demonstrated in a numerical example.

This chapter is based on L. M. Chaouach, T. Oomen, and D. Boskos. *Structured ambiguity sets for distributionally robust optimization.* Provisionally accepted for publication in *European Journal of Control.* 2023. arXiv: 2310.20657 [math.OC].

2.1. INTRODUCTION

UNCERTAINTY in decision-making is abundant across engineering and science. Events with unpredictable outcomes add an additional layer of complexity to the decision-making process in view of the need to strike a balance between performance and risk aversion. To this end, stochastic approaches quantify the uncertainty using a probabilistic model that characterizes the range and frequency of possible outcomes [98]. In stochastic optimization, the uncertain parameters are typically assumed to follow a known distribution [99]. This, in turn, guarantees that the solution of the optimization problem enjoys some desired statistical properties. However, in practical scenarios, the true probability distribution is often uncertain and it is hard to infer it from data with sufficient accuracy. Being uncertain about the uncertainty itself can generate unreliable decisions, which may in turn lead to undesirable risks and failures of complex engineered systems. This makes addressing distributional uncertainty a problem of high importance.

Distributionally robust optimization (DRO) makes decisions in the face of uncertainty without resorting to a single probability distribution. Instead, it robustifies stochastic optimization problems by considering an ambiguity set of plausible models for the unknown distribution of the uncertainty [100]. This way, DRO hedges against model misspecification due to insufficient or corrupted data, which is the typical situation in real-life systems across engineering, finance, machine learning, medicine, and social sciences. There is, therefore, an increasing interest in exploiting DRO for stochastic decision problems, which are widespread in operations research [37], statistical learning [101, 102], and control [103, 104]. Toward applications of DRO in control, [105] develops a distributionally robust LQR framework. Data-driven formulations of Wasserstein distributionally robust stochastic control are found in [92, 106] and [107], while [108] provides a Kalman filtering design that accounts for distributional uncertainty. The problem of propagating optimal transport ambiguity sets is considered in [109–111], which take into account multiple data assimilation nonidealities. Further applications of DRO include economic dispatch in power systems [112], congestion avoidance in traffic control [113], and motion planning in dynamic environments [83].

There are multiple choices of ambiguity sets. In data-driven cases, these choices affect both the statistical properties and the tractability of their associated DRO problems. Typical ambiguity sets are constructed using statistical divergences [76, 114], moment constraints [71, 115], total variation metrics [77], and optimal transport discrepancies [116], such as the Wasserstein distance [117]. Among the favorable properties of Wasserstein ambiguity sets is that they are accompanied by finite sample guarantees of containing the unknown distribution in data-driven scenarios [59]. This, for example, cannot be achieved using divergence-based ambiguity sets when they are centered at the empirical distribution of that data [79, 118]. Wasserstein ambiguity sets are also fully nonparametric and can simultaneously capture all types of distributions (e.g., with multiple modes and of any regularity). Moreover, tractable reformulations of their associated DRO problems [79, 119, 120] are also available. In particular, for a given confidence level, the size of these ambiguity sets decreases with respect to the number of collected samples [59].

Nevertheless, this decay rate suffers from the curse of dimensionality as it becomes excessively slow with the number of samples for high-dimensional data [59, 121, 122]. To ameliorate this drawback, a recent line of work informs the ambiguity set by the specific optimization problem, rendering the ambiguity-size decay rate independent of the dimension of the uncertainty [85, 123–126]. There is also DRO literature, which considers optimal transport ambiguity sets that take into account structural properties of the unknown distribution, like heterogeneity or information about its marginals. To this end, [127] builds Wasserstein ambiguity balls using a Mahalanobis distance that allocates a higher transport cost to directions with a larger impact on the expected loss, while [128] considers a state-dependent variant of this distance. A distributionally robust decision framework for ambiguity sets of multivariate distributions with known marginals is provided in [118], which encodes dependency variations through the Wasserstein distance, while [129] establishes optimal transport duality for ambiguity sets that are defined through Fréchet classes and allow variations of their marginals.

Although important steps have been taken to develop adequate DRO approaches to address complex data-driven problems, the curse of dimensionality with respect to the dimension of the uncertainty still persists in important classes of problems. In particular, it is no longer straightforward how to inform the ambiguity set by the optimization problem when we seek to solve multiple optimization problems with the same underlying uncertainty. There are several such instances in systems and control, including model predictive control (MPC) [106, 130, 131], controller synthesis for reach-avoid specifications [132], multi-objective stochastic control [133], and general dynamic programming [92], which entail solving multiple optimization problems under the same uncertainty. We seek to address the curse of dimensionality that characterizes classical Wasserstein ambiguity sets for such problems by directly constructing alternative ambiguity sets that contain the unknown distribution with prescribed probability. To this end, we build new classes of optimal transport ambiguity sets, which shrink at favorable rates with the number of samples while containing the true distribution with a fixed confidence. Obtaining these probabilistic guarantees necessitates further assumptions regarding the class to which the distribution belongs. In this chapter, we assume independence between lower-dimensional components of the random variable and construct ambiguity sets with distributions that share similar structural properties. We will call them *structured ambiguity sets* and provide dual reformulations of their corresponding DRO problems. The independence assumption we make in the chapter is common in control and system identification problems. For example, the works [106], [105], [134], and [107] are instances where multiple independent random variables are involved.

Our first contribution is the introduction of two classes of structured ambiguity sets, which we call Wasserstein hyperrectangles and optimal-transport hyperrectangles, respectively. The former are designed to contain only product distributions while the latter contain distributions that simultaneously respect multiple optimal transport constraints. Our second contribution is to show that both ambiguity sets shrink faster than traditional Wasserstein balls in data-driven scenarios while containing the true

distribution with the same confidence level. This is established under independence of lower-dimensional components of the random variable and breaks the curse of dimensionality when these components are of sufficiently small dimension. Our third contribution is the derivation of dual reformulations of DRO problems associated with these ambiguity sets. This is enabled by directly working with the space of couplings to define the ambiguity sets, which are typically introduced in a distance-based reasoning. Due to the convexity of multi-transport hyperrectangles, which is in principle not shared by Wasserstein hyperrectangles, their DRO problems admit dual reformulations for a much broader class of objective functions. This general duality result does not assume any structure about the uncertainty and can be applied to any DRO problem where ambiguity sets are formed using multiple optimal transport discrepancies around any baseline model.

This chapter is organized as follows. In Section 2.2, we introduce mathematical preliminaries and notation. We formulate the problem in Section 2.3 and introduce two classes of structured ambiguity sets in Section 2.4. In Section 2.5, we provide probabilistic guarantees for these ambiguity sets and we present dual reformulations for their associated DRO problems in Section 2.6. In Section 2.7, we illustrate the results of the chapter in a simulation example.

2.2. PRELIMINARIES AND NOTATION

Throughout this chapter, we use the following notation. We denote by $\|\cdot\|_p$ the pth norm in $\mathbb{R}^d$ with $p \in [1,\infty]$. We denote by $\mathbb{R}_{\geq 0}$ and $\mathbb{R}_{>0}$ the positive and strictly positive real numbers, respectively, and define $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$. For $N \in \mathbb{N}\backslash\{0\}$, we denote $[N] := \{1,\ldots,N\}$. The diameter of $S \subset \mathbb{R}^d$ is $\operatorname{diam}(S) := \sup\{\|x-y\|_\infty : x, y \in S\}$. We denote by $C(\Xi)$ the class of continuous real-valued functions on a topological space Ξ, and by $C_{\text{const},2}(\Xi \times \Xi)$ the functions $\gamma \in C(\Xi \times \Xi)$ with $\gamma(\zeta,\xi) = \gamma(\zeta,\xi')$ for all $\zeta,\xi,\xi' \in \Xi$. Given the set $\Xi = \Xi_1 \times \cdots \times \Xi_n$ and $k \in [n]$, we define the projection $\operatorname{pr}_k : \Xi \to \Xi_k$ as $\operatorname{pr}_k(\xi) := \xi_k$, for all $\xi = (\xi_1,\ldots,\xi_n) \in \Xi$, and define analogously $\operatorname{pr}_{k,l}$ when projecting to two components indexed by $k,l \in [n]$. Given a normed linear space X and its topological dual X^*, the conjugate of a function $h : X \to \mathbb{R} \cup \{+\infty\}$ is defined by $h^*(x^*) := \sup_{x \in X}\{\langle x^*, x\rangle - h(x)\}$. For a vector space X and a convex cone $K \subset X$, we denote by $\succeq_K$ the order with respect to K, given by $x \succeq_K y$ iff $x - y \in K$ and will omit the dependence on K when it is clear from the context. For example, the order $\geq$ in $\mathbb{R}^d$ with respect to the positive cone $\mathbb{R}^d_{\geq 0} := \{(x_1,\ldots,x_d) \in \mathbb{R}^d : x_k \geq 0 \text{ for all } k \in [d]\}$ implies that $x \geq y$ iff $x_k \geq y_k$ for all $k \in [d]$.

Probability theory: Let Ξ be a Polish space, namely, a complete and separable metric space. We denote by ρ the metric on Ξ, by $\mathscr{B}(\Xi)$ its Borel σ-algebra, and by $\mathscr{P}(\Xi)$ the space of probability measures on $(\Xi, \mathscr{B}(\Xi))$. The Dirac distribution centered at $\xi \in \Xi$ is denoted by δ_ξ. The indicator function $\mathbb{1}_\Theta$ of $\Theta \subset \Xi$ is $\mathbb{1}_\Theta(\xi) := 1$ if $\xi \in \Theta$ and 0 otherwise. Given the measurable spaces $(\Omega, \mathscr{F})$ and $(\Omega', \mathscr{F}')$, a measurable map $\Psi : (\Omega, \mathscr{F}) \to (\Omega', \mathscr{F}')$ assigns to each (signed) measure μ in $(\Omega, \mathscr{F})$ the pushforward measure $\Psi_\# \mu$ in $(\Omega', \mathscr{F}')$ defined by $\Psi_\#\mu(B) := \mu(\Psi^{-1}(B))$ for all $B \in \mathscr{F}'$. We denote by $P \otimes Q$ the product measure of P and Q. For any $P \in \mathscr{P}(\Xi)$, its support is the

closed set $\operatorname{supp}(P) := \{x \in \Xi : P(U) > 0 \text{ for each neighborhood } U \text{ of } x\}$. Given a function $X : \Omega \to \Xi$ with the σ-algebra $\mathscr{B}(\Xi)$, we denote by $\sigma(X)$ the σ-algebra generated by X on Ω. The universal σ algebra on Ξ is defined as $\mathscr{U}(\Xi) := \cap_{P \in \mathscr{P}(\Xi)} \mathscr{B}_P(\Xi)$ (cf. [135, Definition 7.18]), where $\mathscr{B}_P(\Xi)$ refers to the completion of the σ-algebra $\mathscr{B}(\Xi)$ with respect to the measure P (cf. [136, Remark 1.70]) and satisfies $\mathscr{B}(\Xi) \subset \mathscr{B}_P(\Xi)$. We denote by $\mathfrak{m}_{\mathscr{U}}(\Xi; \mathbb{R} \cup \{+\infty\})$ the space of measurable functions from $(\Xi, \mathscr{U}(\Xi))$ to $\mathbb{R} \cup \{+\infty\}$ with its Borel σ-algebra. For any $p \geq 1$, we denote by $\mathscr{P}_p(\Xi)$ the set of probability measures in $\mathscr{P}(\Xi)$ with finite pth moment. Given $P, Q \in \mathscr{P}_p(\Xi)$, their pth Wasserstein distance is

$$W_p(Q,P) := \inf_{\pi \in \mathscr{C}(Q,P)} \left\{ \int_{\Xi \times \Xi} \rho(\zeta,\xi)^p \, d\pi(\zeta,\xi) \right\}^{\frac{1}{p}}$$

(cf. [117]). Each $\pi \in \mathscr{C}(Q,P)$ is a transport plan, i.e., a distribution on $\Xi \times \Xi$ with marginals $P = \mathrm{pr}_{2\#}\pi$ and $Q = \mathrm{pr}_{1\#}\pi$, respectively. The Wasserstein distance between P and Q is defined through the optimal cost to transfer the mass of one distribution to the other when the cost to transfer a unit of mass between two locations ζ and ξ in Ξ is $\rho(\zeta,\xi)^p$. By Katorovich duality (cf. [117, Theorem 1.3]), the optimal transportation cost $W_p^p(Q,P)$ is equal to the value of its dual optimization problem

$$K_p(Q,P) = \sup_{\substack{(\psi,\phi) \in L^1(Q) \times L^1(P) \\ \phi(\xi) - \psi(\zeta) \leq \rho(\zeta,\xi)^p}} \left\{ \int_\Xi \phi(\xi) dP(\xi) - \int_\Xi \psi(\zeta) dQ(\zeta) \right\}.$$

2.3. PROBLEM FORMULATION

In this section, we introduce data-driven stochastic optimization problems and their distributionally robust formulations that hedge against model uncertainty. Consider the stochastic optimization problem

$$\inf_{x \in \mathscr{X}} \mathbb{E}_{P_\xi}\big[f(x,\xi)\big], \tag{2.3.1}$$

where f is the objective function, $x \in \mathscr{X}$ is the decision variable, and ξ is a random variable, which takes values in a Polish space Ξ and has distribution P_ξ.

A typical situation that fits into (2.3.1) is when the distribution P_ξ is unknown and there is only access to a finite number of i.i.d. samples $\xi^1, \ldots, \xi^N$ of ξ. The usual approach to approximate the solution of (2.3.1) in this case is to replace P_ξ by the empirical distribution $P_\xi^N := \frac{1}{N}\sum_{i=1}^N \delta_{\xi^i}$. This is known as the Sample Average Approximation (SAA) of (2.3.1) and it converges to the solution of the original problem in the asymptotic limit [99].

2.3.1. DISTRIBUTIONALLY ROBUST OPTIMIZATION

When the available data are limited, the empirical distribution P_ξ^N may exhibit significant deviations from the true distribution P_ξ, which can in turn have a considerable impact on the discrepancy between the SAA and the original

optimization problem. To address this issue, uncertainty in the distribution is incorporated into (2.3.1) under the robust formulation

$$\inf_{x\in X} \sup_{P\in\mathscr{P}^N} \mathbb{E}_P\big[f(x,\xi)\big]. \tag{2.3.2}$$

In this distributionally robust optimization (DRO) problem, $\mathscr{P}^N$ is an ambiguity set of distributions that is inferred from the samples and contains plausible models of the true distribution.

A well-established approach to construct data-driven ambiguity sets is to group all distributions that are ε-close to the empirical distribution P_ξ^N in the pth Wasserstein metric for some $p \geq 1$ and $\varepsilon > 0$. In this case, $\mathscr{P}^N$ in (2.3.2) is the ball

$$\mathscr{B}_p(P_\xi^N, \varepsilon) := \{P \in \mathscr{P}_p(\Xi) : W_p(P_\xi^N, P) \leq \varepsilon\}$$

with center P_ξ^N and radius ε. Among the benefits of this choice are that Wasserstein distances capture well the underlying geometry of the space Ξ where the uncertainty lives (cf. [117, Page 99]). This yields correspondingly higher penalties to distributional variations that are farther apart in the domain and may induce larger discrepancies on the optimization problem. In addition, Wasserstein balls lead to tractable DRO problems [79] and are accompanied by rigorous finite-sample guarantees of containing the true distribution with prescribed confidence [59]. These require no parametric assumptions for the underlying distribution and ensure further that the size of the ambiguity set converges to zero as the number of samples goes to infinity. As a result, the value of (2.3.2) provides an upper bound for the expected cost (2.3.1) with prescribed confidence and approaches it asymptotically as the number of samples grows.

2.3.2. STRUCTURED AMBIGUITY SETS

The size of the ambiguity set $\mathscr{P}^N$ has a direct effect on the solution of (2.3.2) since ambiguity balls of larger sizes may lead to conservative upper bounds for (2.3.1). This can happen because an ambiguity ball that is sufficiently large to contain the true distribution with a prescribed probability may also contain several irrelevant distributions. To address this issue, we consider some prior knowledge about the uncertainty, which can facilitate the construction of ambiguity sets whose elements are more appropriate models of the unknown distribution. We make the following assumption regarding the class of the random variable.

Assumption 1. (Independent random variable components). (i) The random variable ξ takes values in the Polish space $\Xi = \Xi_1 \times \cdots \times \Xi_n$, with Ξ and Ξ_k, $k \in [n]$ equipped with the metrics ρ and ρ_k, $k \in [n]$, respectively. (ii) The components $\xi_1, \ldots, \xi_n$ of ξ are independent.

Applications involving high-dimensional uncertainty with independent components include multi-agent systems, where, for example, the noise affecting measurements or communication among agents is typically assumed independent across agents [137–139]. Analogously, in the deployment of multi-robot teams agents for tasks such

as simultaneous localization and mapping, disturbances in the dynamics and measurement noise are commonly assumed independent across the robots [140]. Further examples from optimal control and scheduling where this uncertainty structure is commonly encountered include stochastic model predictive control [12], data-enabled predictive control [107], distributionally robust control of stochastic systems [105, 141, 142], and multistage scheduling of energy systems [143, 144].

Problem formulation. Under Assumption 1, we seek to introduce *structure* in data-driven ambiguity sets so that they contain the true distribution with high probability while excluding implausible distributions and enabling the formulation of tractable DRO problems.

To this end, note that due to Assumption 1, the distribution of ξ is the product measure

$$P_\xi = P_{\xi_1} \otimes \cdots \otimes P_{\xi_n}, \tag{2.3.3}$$

with P_{ξ_k}, $k \in [n]$ denoting the distributions of its components. Thus, instead of looking for plausible models of P_ξ in an ambiguity ball, we can consider ambiguity sets whose distributions are only product measures, or at least sufficiently close to product measures. Such a set should contain a restricted class of distributions, and therefore, yield less conservative solutions for (2.3.2) under the same confidence.

2.3.3. AMBIGUITY RADIUS

By tuning the radius of the ambiguity ball, it is possible to guarantee that it contains the true distribution with prescribed probability. These guarantees hinge on concentration of measure results, which leverage prior assumptions about the class where the unknown distribution belongs, to bound the Wasserstein distance between the true and the empirical distribution. Such assumptions are the size of the distribution's support (e.g., [59, Proposition 10], [122]), its tail decay rate (e.g., [59, Theorem 2, cases (1) and (2)]), or bounds on its moments (e.g., [59, Theorem 2, case (3)], [145]). Based on these results, for any confidence $1-\beta$ and number N of i.i.d. samples, we can select the ambiguity radius $\varepsilon(N,\beta)$ so that

$$\mathbb{P}(P_\xi \in \mathscr{B}_p(P_\xi^N, \varepsilon)) \geq 1-\beta. \tag{2.3.4}$$

The radius can typically be determined by a bound of the form

$$\varepsilon(N,\beta) \leq K \frac{1}{N^{1/\max\{d,2p\}}}, \tag{2.3.5}$$

where d is the dimension of the random vector ξ. Therefore, for high-dimensional random variables, the decrease of the radius with respect to the number of samples becomes excessively slow. As a result, the exploitation of more data does not guarantee any significant improvement of the closeness between the true distribution and its empirical approximation, and hence, also of the size of the ambiguity ball. Recent work addresses this limitation by informing the ambiguity set directly from the objective function of the optimization problem [85, 123, 125]. This yields a decay rate of the order $1/\sqrt{N}$ for the radius of the Wasserstein ambiguity ball regardless of the dimension of the random variable. However, when solving multiple DRO

problems of the form (2.3.2) under the same uncertainty, the straightforward way to obtain such statistical guarantees for all problems simultaneously is to exploit union bounds based on the guarantees of the individual problems. This may lead to conservative guarantees that essentially cannot resolve the curse of dimensionality. Examples where this issue may arise include distributionally robust model predictive control, where several distributionally robust probabilistic constraints need to be enforced simultaneously [106]. Analogous considerations appear when solving stochastic reachability problems using abstraction methods [146]. There, one needs to compute upper and lower bounds for the transition probabilities among all states of the abstraction, which need to hold simultaneously with high confidence. In these situations, resorting to ambiguity sets that contain the true distribution with a high confidence is necessary to retain rigorous probabilistic guarantees. To this end, we seek to *exploit the independence Assumption 1 for the components of ξ and determine an ambiguity set that contains the true distribution with high probability and does not suffer from the curse of dimensionality with respect to d.*

2.4. AMBIGUITY HYPERRECTANGLES

In this section, we introduce two classes of structured ambiguity sets and provide some of their key statistical properties for data-driven problems. The starting point to construct these ambiguity sets are the lower-dimensional components of the random variable $\xi = (\xi_1, \ldots, \xi_n)$. Using N i.i.d. samples $\xi^1, \ldots, \xi^N$, we first build a lower-dimensional ambiguity ball $\mathscr{B}_p(P^N_{\xi_k}, \varepsilon_k)$ for each component of ξ, where $P^N_{\xi_k} := \frac{1}{N}\sum_{i=1}^N \delta_{\xi^i_k}$ denotes its corresponding empirical distribution. From these balls, we construct the *Wasserstein hyperrectangle*

$$\mathscr{H}_p(\boldsymbol{P}^N_\xi, \boldsymbol{\varepsilon}) := \{P'_{\xi_1} \otimes \cdots \otimes P'_{\xi_n} : \; P'_{\xi_k} \in \mathscr{B}_p(P^N_{\xi_k}, \varepsilon_k) \text{ for all } k \in [n]\} \tag{2.4.1a}$$

$$\boldsymbol{P}^N_\xi := P^N_{\xi_1} \otimes \cdots \otimes P^N_{\xi_n}, \quad \boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n), \tag{2.4.1b}$$

by taking the product measures across the individual distributions from the balls. We refer to the nominal model $\boldsymbol{P}^N_\xi$ around which the ambiguity set is built as the *product empirical distribution.*

Next, we establish probabilistic guarantees, which ensure that the Wasserstein hyperrectangles contain the distribution of ξ with prescribed confidence. Later, we exploit these guarantees to alleviate the curse of dimensionality regarding the shrinkage of Wasserstein balls. The following result establishes the guarantees that a Wasserstein hyperrectangle inherits from its lower-dimensional constituent ambiguity balls when the components of ξ are independent.

Theorem 2.4.1. *(Probabilistic guarantees for Wasserstein hyperrectangles). Assume that the random variable ξ satisfies Assumption 1 and that $P_\xi \in \mathscr{P}_p(\Xi)$. Given i.i.d. samples $\xi^1, \ldots, \xi^N$ of ξ, let $P^N_{\xi_1}, \ldots, P^N_{\xi_n}$ be the empirical distributions of the individual components. Assume also that each Wasserstein ball $\mathscr{B}_p(P^N_{\xi_k}, \varepsilon_k)$ contains P_{ξ_k} with confidence $1-\beta_k$. Then the Wasserstein hyperrectangle $\mathscr{H}_p(\boldsymbol{P}^N_\xi, \boldsymbol{\varepsilon})$ contains P_ξ with confidence $\prod_{k=1}^n 1-\beta_k$.*

To prove Theorem 2.4.1 we use the following lemma, whose proof is given in Appendix 2.A.1.

Lemma 2.4.2. *(Independent Wasserstein distances across empirical distributions). Assume that the random variable ξ satisfies Assumption 1 and that $P_\xi \in \mathscr{P}_p(\Xi)$. Given i.i.d. samples $\xi^1,\ldots,\xi^N$ of ξ, let $P^N_{\xi_1},\ldots,P^N_{\xi_n}$ be the empirical distributions of its components. Then for any $\varepsilon_1,\ldots,\varepsilon_n \geq 0$ the events $\{W_p(P^N_{\xi_k},P_{\xi_k}) \leq \varepsilon_k\}$, $k \in [n]$ are independent.*

Proof of Theorem 2.4.1. By Assumption 1, P_ξ is expressed as the product distribution in (2.3.3). Thus, we get from the definition of the Wasserstein hyperrectangle in (2.4.1) that

$$\mathbb{P}\big(P_\xi \in \mathscr{H}_p(\boldsymbol{P}^N_\xi,\boldsymbol{\varepsilon})\big) = \mathbb{P}\big(P_{\xi_k} \in \mathscr{B}_p(P^N_{\xi_k},\varepsilon_k) \text{ for all } k \in [n]\big). \tag{2.4.2}$$

Also, by the definition of a Wasserstein ball,

$$P_{\xi_k} \in \mathscr{B}_p(P^N_{\xi_k},\varepsilon_k) \Longleftrightarrow W_p(P^N_{\xi_k},P_{\xi_k}) \leq \varepsilon_k. \tag{2.4.3}$$

Since the components of ξ are independent, we get from the independence result of Lemma 2.4.2, (2.4.2), and (2.4.3) that

$$\mathbb{P}\big(P_\xi \in \mathscr{H}_p(\boldsymbol{P}^N_\xi,\boldsymbol{\varepsilon})\big) = \prod_{k=1}^n \mathbb{P}\big(P_{\xi_k} \in \mathscr{B}_p(P^N_{\xi_k},\varepsilon_k)\big). \tag{2.4.4}$$

Recalling that the kth Wasserstein ball contains P_{ξ_k} with confidence $1-\beta_k$ for each $k \in [n]$, we get from (2.4.4) that the hyperrectangle contains P_ξ with confidence $\prod_{k=1}^n 1-\beta_k$, which concludes the proof. □

Remark 2.4.3. (Boldface notation). We use boldface notation throughout the thesis to signify elements, which in contrast to the typical DRO literature, admit a vectorized or product representation. These include the vector Wasserstein radii $\boldsymbol{\varepsilon}$, the product empirical distribution $\boldsymbol{P}^N_\xi$ —to distinguish it from the standard empirical distribution P^N_ξ—, and vectors of dual variables $\boldsymbol{\lambda}$ that are introduced later in dual DRO reformulations.

Note that we can directly generalize the notion of a Wasserstein hyperrectangle to the case where the nominal distribution is a general product distribution $Q = Q_1 \otimes \cdots \otimes Q_n$ on the Polish space $\Xi = \Xi_1 \times \cdots \times \Xi_n$ with $Q_k \in \mathscr{P}(\Xi_k)$ for each $k \in [n]$, instead of the product empirical distribution $\boldsymbol{P}^N_\xi$. Again, the Wasserstein hyperrectangle $\mathscr{H}_p(Q,\boldsymbol{\varepsilon})$ comprises of all product distributions whose kth lower-dimensional marginal has Wasserstein distance at most ε_k from the corresponding marginal of Q. Since Wasserstein hyperrectangles contain only product distributions, they are non-convex. This restricts the class of cost functions for which the DRO problem (2.3.2) with $\mathscr{P}^N \equiv \mathscr{H}_p(Q,\boldsymbol{\varepsilon})$ admits tractable reformulations. To overcome this obstacle, we build a convex ambiguity set, which shrinks at the same favorable rate as the Wasserstein hyperrectangle with respect to the number of samples. The

distributions of this ambiguity set are defined through couplings with a nominal distribution, which need to respect a set of transport cost constraints.

In particular, consider a general reference distribution $Q \in \mathscr{P}(\Xi)$ and let

$$\Pi(Q) := \{\pi \in \mathscr{P}(\Xi \times \Xi) : \mathrm{pr}_{1\#}\pi = Q\}.$$

2

Consider also the lower semicontinuous cost functions $c_k : \Xi \times \Xi \to \mathbb{R}_{\geq 0}$, $k \in [n]$ with $c_k(\zeta,\zeta) = 0$ for all $\zeta \in \Xi$, the transport budget vector $\boldsymbol{\epsilon} = (\epsilon_1,\ldots,\epsilon_n)$ with positive entries, and let

$$\Pi(Q,\boldsymbol{\epsilon}) \equiv \Pi(Q,\boldsymbol{\epsilon};c_1,\ldots,c_n) := \Big\{\pi \in \Pi(Q) : \int_{\Xi\times\Xi} c_k(\zeta,\xi)d\pi(\zeta,\xi) \leq \epsilon_k \text{ for all } k \in [n]\Big\}. \tag{2.4.5}$$

Due to the fact that each $c_k(\zeta,\zeta) \equiv 0$, $\Pi(Q,\boldsymbol{\epsilon})$ is always nonempty. We define the *multi-transport hyperrectangle*

$$\mathscr{T}(Q,\boldsymbol{\epsilon}) := \mathrm{pr}_{2\#}\Pi(Q,\boldsymbol{\epsilon}), \tag{2.4.6}$$

which is convex and depends on the chosen cost functions $c_1,\ldots,c_n$. When Assumption 1(i) is satisfied and the costs are $c_k(\zeta,\xi) := \rho_k(\zeta_k,\xi_k)^p$ for some $p \geq 1$, we denote

$$\mathscr{T}_p(Q,\boldsymbol{\epsilon}) := \mathscr{T}(Q,\boldsymbol{\epsilon}^p), \tag{2.4.7}$$

where $\boldsymbol{\epsilon}^p := (\epsilon_1^p,\ldots,\epsilon_n^p)$. The next result delineates the relation between Wasserstein hyperrectangles and multi-transport hyperrectangles of the form (2.4.7) that are built around product distributions.

Proposition 2.4.4. *(Wasserstein hyperrectangle containment). Consider a Polish space Ξ as in Assumption 1(i) and a product distribution $Q = Q_1 \otimes \cdots \otimes Q_n \in \mathscr{P}_p(\Xi)$ with $Q_k \in \mathscr{P}_p(\Xi_k)$ for each $k \in [n]$. Then $\mathscr{H}_p(Q,\boldsymbol{\epsilon}) \subset \mathscr{T}_p(Q,\boldsymbol{\epsilon})$. In addition, for any product distribution $P \in \mathscr{T}_p(Q,\boldsymbol{\epsilon})$, also $P \in \mathscr{H}_p(Q,\boldsymbol{\epsilon})$.*

Proof. Let $P \in \mathscr{H}_p(Q,\boldsymbol{\epsilon})$. Then $P = P_1 \otimes \cdots \otimes P_n$ and $P_k \in \mathscr{B}_p(Q_k,\epsilon_k)$ for all $k \in [n]$, which implies that $W_p(Q_k,P_k) \leq \epsilon_k$. Thus, for each $k \in [n]$, there exists an optimal transport plan π_k for the Wasserstein distance between Q_k and P_k (cf. [117, Theorem 4.1]) with

$$\int_{\Xi_k\times\Xi_k} \rho_k(\zeta_k,\xi_k)^p d\pi_k(\zeta_k,\xi_k) \leq \epsilon_k^p. \tag{2.4.8}$$

Next, define

$$\pi := \bigotimes_{k=1}^n \pi_k \quad \text{and} \quad \widetilde{\pi} := T_\#\pi, \tag{2.4.9}$$

where $T : \prod_{k=1}^n \Xi_k \times \Xi_k \to \prod_{k=1}^n \Xi_k \times \prod_{k=1}^n \Xi_k$ is the linear map $T(\zeta_1,\xi_1,\ldots,\zeta_n,\xi_n) := (\zeta_1,\ldots,\zeta_n,\xi_1,\ldots,\xi_n)$. Then $\widetilde{\pi}$ is a transport plan between Q and P since

$$\widetilde{\pi}(A_1 \times \cdots \times A_n \times \Xi) = \widetilde{\pi}(A_1 \times \cdots \times A_n \times \Xi_1 \times \cdots \times \Xi_n)$$

$$\overset{(a)}{=} \pi(A_1 \times \Xi_1 \times \cdots \times A_n \times \Xi_n)$$
$$\overset{(b)}{=} \prod_{k=1}^{n} \pi_k(A_k \times \Xi_k) \overset{(c)}{=} \prod_{k=1}^{n} Q_k(A_k)$$
$$= Q_1 \otimes \cdots \otimes Q_n(A_1 \times \cdots \times A_n) = Q(A_1 \times \cdots \times A_n)$$

for any $A_k \in \mathscr{B}(\Xi_k)$, $k \in [n]$. Here, we used (2.4.9) in (a) and (b), and the fact that each π_k is a transport plan in (c). Analogously, P is also a marginal of π. In addition, we get from (2.4.8) and Fubini's theorem (cf. [147, Page 233]) that

$$\int_{\Xi_k \times \Xi_k} \rho_k(\zeta_k, \xi_k)^p \, d\widetilde{\pi}(\zeta, \xi) \leq \varepsilon_k^p$$

for each $k \in [n]$, which by (2.4.5)-(2.4.7) implies that also $P \in \mathscr{T}_p(Q, \boldsymbol{\varepsilon})$ and concludes the proof of the first claim.

For the proof of the second claim consider a product distribution $P \in \mathscr{T}_p(Q, \boldsymbol{\varepsilon})$. Then there exists a transport plan π between Q and P so that (2.4.5) holds with $c_k(\zeta, \xi) \equiv \rho_k(\zeta_k, \xi_k)^p$ and $\epsilon_k \equiv \varepsilon_k^p$. Next, let $\pi_k := \mathrm{pr}_{k,n+k\#}\pi$ (with π viewed as a distribution on $\prod_{k=1}^n \Xi_k \times \prod_{k=1}^n \Xi_k$). It follows that π_k has marginals Q_k and P_k, respectively, and that it satisfies (2.4.8). As a result, $P_k \in \mathscr{B}_p(Q_k, \varepsilon_k)$ for all $k \in [n]$ and we conclude that also $P \in \mathscr{H}_p(Q, \boldsymbol{\varepsilon})$. □

The next result follows directly from Proposition 2.4.4 and provides conditions under which a multi-transport hyperrectangle contains the true distribution with prescribed confidence.

Corollary 2.4.5. *(Probabilistic guarantees for multi-transport hyperrectangles). Assume that the random variable ξ satisfies Assumption 1 and that $P_\xi \in \mathscr{P}_p(\Xi)$. Given i.i.d. samples $\xi^1, \ldots, \xi^N$ of ξ, let $P_{\xi_1}^N, \ldots, P_{\xi_n}^N$ be the empirical distributions of its components. Assume also that each Wasserstein ball $\mathscr{B}_p(P_{\xi_k}^N, \varepsilon_k)$ contains P_{ξ_k} with confidence $1-\beta_k$. Then the multi-transport hyperrectangle $\mathscr{T}_p(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})$ contains P_ξ with confidence $\prod_{k=1}^n 1-\beta_k$.*

We conclude this section with a result that compares the size of multi-transport hyperrectangles with that of monolithic balls. Specifically, we determine the radius that a Wasserstein ball should have in order to contain a multi-transport hyperrectangle when its reference distribution is also the center of that ball. For this, we also need to relate the metric ρ on the product space Ξ with the metrics ρ_k on the components Ξ_k.

Proposition 2.4.6. *(Size of enclosing Wasserstein ball). Let $Q \in \mathscr{P}_p(\Xi)$ and assume that the metric on Ξ is*

$$\rho(\zeta, \xi) := \Big(\sum_{k=1}^{n} \rho_k(\zeta_k, \xi_k)^q \Big)^{\frac{1}{q}}, \quad \zeta, \xi \in \Xi, \tag{2.4.10}$$

for some $q \geq 1$. Then the multi-transport hyperrectangle $\mathscr{T}_p(Q, \boldsymbol{\varepsilon})$ satisfies

$$\mathscr{T}_p(Q, \boldsymbol{\varepsilon}) \subset \mathscr{B}_p(Q, \varepsilon),$$

where $\varepsilon = n^{\max\{0,1/q-1/p\}}\big(\sum_{k=1}^{n}\varepsilon_k^p\big)^{\frac{1}{p}}$. *If in addition* $p=q$ *and* $Q = Q_1\otimes\cdots\otimes Q_n$ *is a product distribution, then there exists a product distribution* $P\in\mathcal{T}_p(Q,\boldsymbol{\varepsilon})$ *with* $W_p(Q,P)=\varepsilon$.

The proof is given in Appendix 2.A.1. When Q is a product distribution and (2.4.10) holds, we get from the first parts of Propositions 2.4.4 and 2.4.6 the inclusions

$$\mathcal{H}_p(Q,\boldsymbol{\varepsilon})\subset\mathcal{T}_p(Q,\boldsymbol{\varepsilon})\subset\mathcal{B}_p(Q,\varepsilon),$$

where the radius ε of the ball is given in Proposition 2.4.6. From the second part of the same propositions, it follows that when the exponents of the product metric and the Wasserstein distance coincide, there is at least one common point in both ambiguity hyperrectangles that lies on the boundary of their enclosing ball (cf. Figure 2.1).

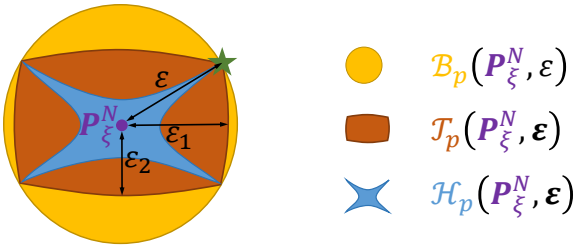


Figure 2.1.: The figure shows the Wasserstein hyperrectangle $\mathcal{H}_p(P_\xi^N,\boldsymbol{\varepsilon})$, the multi-transport hyperrectangle $\mathcal{T}_p(P_\xi^N,\boldsymbol{\varepsilon})$, and their enclosing ball $\mathcal{B}_p(P_\xi^N,\varepsilon)$ around the product empirical distribution P_ξ^N for a random variable with two independent components. The star denotes a common distribution of both hyperrectangles that lies on the boundary of the ball.

2.5. AMBIGUITY HYPERRECTANGLE SIZE BASED ON THE NUMBER OF SAMPLES

In this section, we compare Wasserstein and multi-transport hyperrectangles with Wasserstein balls in terms of the size reduction that they exhibit with the number of samples. For this comparison, we assume that both sets are constructed using the same samples and the same confidence level. Since most of the concentration of measure results for this purpose are formulated for distributions supported on Euclidean spaces (cf. [59, 110]), we focus on the case where Ξ is a bounded subset of $\mathbb{R}^d$ with the distance induced by the norm $\|\cdot\|_q$. The next result presents bounds for the Wasserstein distance between the true and the empirical distribution, which we exploit to tune the size of $\mathcal{H}_p(P_\xi^N,\boldsymbol{\varepsilon})$ and $\mathcal{T}_p(P_\xi^N,\boldsymbol{\varepsilon})$ so that they contain the true distribution with a desired confidence.

Proposition 2.5.1. *(Ambiguity radius [148, Proposition 24]). Assume that the probability distribution* P_ξ *is supported on* $\Xi\subset\mathbb{R}^d$ *with* $\rho_\Xi := \operatorname{diam}(\Xi)<\infty$. *Assume also that* $d\geq 2p+1$ *and let* $\xi^1,\ldots,\xi^N$ *be i.i.d. samples of* ξ. *Then the ambiguity radius*

$$\widehat{\varepsilon}(N,\beta,\rho_\Xi,p,q,d) := \rho_\Xi\widehat{C}(\beta,p,q,d)\frac{1}{N^{1/d}}, \tag{2.5.1}$$

where

$$\widehat{C}(\beta,p,q,d) := d^{1/q}2^{1/2p}(C(d,p)+(\ln\beta^{-1})^{1/2p})$$
$$C(d,p) := 2^{(d-2)/2p}\Big(\frac{1}{2^{1/2}-1}+\frac{1}{2^{1/2}-2^{1/2-p}}\Big)^{1/p},$$

and $1-\beta$ *is a desired confidence level, guarantees that*

$$\mathbb{P}(P_\xi \in \mathscr{B}_p(P_\xi^N,\varepsilon)) \geq 1-\beta.$$

Using this ambiguity radius and the guarantees of Theorem 2.4.1 we determine a ball around the product empirical distribution that contains both the Wasserstein hyperrectangle and the multi-transport hyperrectangle with prescribed probability. The proof of this result is given in Appendix 2.A.2.

Proposition 2.5.2. *(Size reduction of ambiguity hyperrectangles). Assume that the random variable* ξ *is supported on the compact set* $\Xi \equiv \Xi_1 \times \cdots \times \Xi_n \subset \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_n} \equiv \mathbb{R}^d$ *with* $d_k \geq 2p+1$ *for each* $k \in [n]$ *and satisfies Assumption 1 with the metric induced by* $\|\cdot\|_q$ *in each space. For any confidence* $1-\beta$*, let*

$$\beta_k := \beta\frac{d_k}{d}, \quad \varepsilon_k := \widehat{\varepsilon}(N,\beta_k,\rho_\Xi,p,q,d_k),$$

with $\widehat{\varepsilon}$ *as in* (2.5.1)*, and consider the ambiguity sets* $\mathscr{H}_p(\boldsymbol{P}_\xi^N,\boldsymbol{\varepsilon})$ *and* $\mathscr{T}_p(\boldsymbol{P}_\xi^N,\boldsymbol{\varepsilon})$*. Then both sets contain* P_ξ *with confidence* $1-\beta$ *and*

$$\mathscr{H}_p(\boldsymbol{P}_\xi^N,\boldsymbol{\varepsilon}) \subset \mathscr{T}_p(\boldsymbol{P}_\xi^N,\boldsymbol{\varepsilon}) \subset \mathscr{B}_p(\boldsymbol{P}_\xi^N,\varepsilon), \tag{2.5.2}$$

where

$$\varepsilon = cn^{1/p+\max\{0,1/q-1/p\}}\rho_\Xi\widehat{C}(\beta,p,q,d)\frac{1}{N^{1/d_{\max}}}, \tag{2.5.3}$$

$c := (\sqrt{2q+1}+1)/\big(2e^{(\sqrt{2q+1}+1)^2/8}\big)$*,* $d_{\max} := \max_{k\in[n]} d_k$*, and* $\widehat{C}$ *is defined in Proposition 2.5.1.*

Remark 2.5.3. (Component dimensionality bound). The assumption $d_k \geq 2p+1$ in Proposition 2.5.2 is made to facilitate providing the explicit multiplicative constant in (2.5.1), which is used to quantify the size of the enclosing ball in (2.5.3). Analogous formulas can be obtained by leveraging the explicit bounds on the expected Wasserstein distance between the true and the empirical distribution derived in the recent work [149] for compact subsets of $\mathbb{R}^d$. By combining these results with bounds on the concentration of the Wasserstein distance between the true and empirical distributions around its expected value, one can obtain, as in [148], explicit radii for any d_k and p. However, these expressions are more convoluted and we do no provide them here.

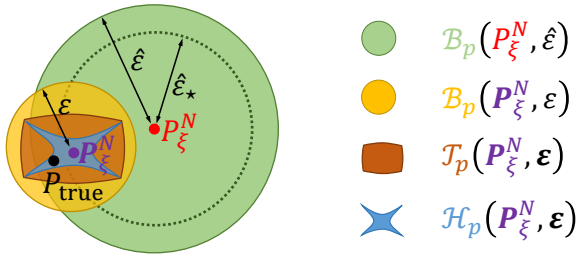


Figure 2.2.: The figure shows the hyperrectangles $\mathscr{H}_p(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})$ (in blue), $\mathscr{T}_p(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})$ (in dark red), and their enclosing ball $\mathscr{B}_p(\boldsymbol{P}_\xi^N, \varepsilon)$ with ε given in (2.5.3) (in yellow), which are all centered at the product empirical distribution $\boldsymbol{P}_\xi^N$, as well as the monolithic Wasserstein ball $\mathscr{B}_p(P_\xi^N, \varepsilon)$ (in green) around the empirical distribution P_ξ^N. All sets contain the true distribution, which is always outside the dashed ball around P_ξ^N with radius equal to the lower bound $\widehat{\varepsilon}_\star$ in (2.5.4).

Under the assumptions of Proposition 2.5.2, we can compare the size of both hyperrectangles and a monolithic ball that contains P_ξ with the same confidence. If we use the bounds of Proposition 2.5.1, then the radius ε of the Wasserstein ball that encloses the hyperrectangles is guaranteed to be strictly smaller than the radius $\widehat{\varepsilon}$ of the monolithic ball when

$$N \geq \left(cn^{1/p+\max\{0,1/q-1/p\}}\right)^{1/d_{\max}-1/d}$$

and decreases much faster for larger N, where $N^{-1/d_{\max}} \ll N^{-1/d}$ (cf. Figure 2.2). The center $\boldsymbol{P}_\xi^N$ of the ball enclosing the hyperrectangles is different from the center P_ξ^N of the monolithic ball, since P_ξ^N is the empirical distribution $\frac{1}{N}\sum_{i=1}^N \delta_{\xi^i}$, whereas $\boldsymbol{P}_\xi^N$ is the product empirical distribution

$$P_{\xi_1}^N \otimes \cdots \otimes P_{\xi_n}^N = \frac{1}{N^n} \sum_{(i_1,\ldots,i_n)\in[N]^n} \delta_{(\xi_1^{i_1},\ldots,\xi_n^{i_n})}.$$

(cf. Figures 2.2, 2.3). This also implies that under Assumption 1, an ambiguity ball that is centered at the product empirical distribution $\boldsymbol{P}_\xi^N$ will contain the true distribution with significantly higher probability compared to when it is centered at the empirical distribution P_ξ^N.

The favorable decay rate of the ambiguity rectangles is further justified by the fact that the corresponding radius of monolithic Wasserstein balls can in principle not be improved, besides potentially a constant factor that is independent of the samples. Indeed, for any distribution $P_\xi \in \mathscr{P}_p(\Xi)$ for which $\operatorname{supp}(P_\xi)$ has a non-empty interior in $\mathbb{R}^d$,

$$\widehat{\varepsilon}_\star := \left(\frac{d}{d+p}\right)^{1/p} C_\star^{-1/d} \frac{1}{N^{1/d}} \leq W_p(P_\xi^N, P_\xi) \tag{2.5.4}$$

always holds for some $C_\star > 0$, since the lower bound $\widehat{\varepsilon}_\star$ in (2.5.4) holds for any discrete distribution in place of P_ξ^N that is supported on N points (cf. [150,

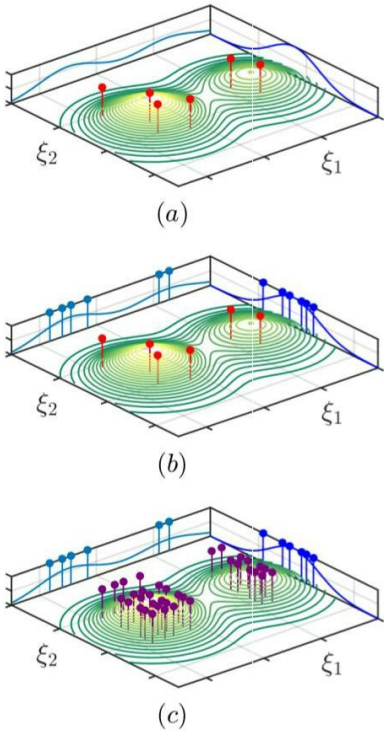


Figure 2.3.: The figure illustrates an example of the reference distributions P_ξ^N and $\boldsymbol{P}_\xi^N$. (a) shows a contour plot of the product distribution $P_\xi = P_{\xi_1} \otimes P_{\xi_2}$, its marginals P_{ξ_1} and P_{ξ_2}, and the empirical distribution P_ξ^N in red, which is formed by taking six samples from the true distribution P_ξ. (b) shows the marginals $P_{\xi_1}^N$ and $P_{\xi_2}^N$ of the empirical distribution P_ξ^N in light-blue and blue, respectively. Finally, (c) illustrates the product empirical distribution $\boldsymbol{P}_\xi^N = P_{\xi_1}^N \otimes P_{\xi_2}^N$ in purple, which is formed by taking the product of the marginal empirical distributions. The product empirical distribution $\boldsymbol{P}_\xi^N$ is clearly an improved approximation of the true distribution compared to the empirical distribution P_ξ^N.

Proposition 4.2]). Namely, to contain the distribution P_ξ with nonzero probability, the monolithic ambiguity ball centered at P_ξ^N needs to have a radius at least $\widehat{\varepsilon}_\star$, as shown with the dashed circle in Figure 2.2, which shrinks at the same rate $N^{-1/d}$ as the radius $\widehat{\varepsilon}$ in (2.5.1). From Propositions 2.5.1 and 2.5.2 it also follows that uncertainty components with heterogenous supports have proportionally heterogeneous ambiguity radii. Thus, the size of the ambiguity hyperrectangles is naturally tuned based on prior information about the components of their uncertainty.

Remark 2.5.4. (Tightness of hyperrectangle bounds). When the lower-dimensional components of ξ have the same dimension, i.e., $d_k = d/n \equiv d_{\max}$ for all $k \in [n]$, where $d_{\max}$ is given in Proposition 2.5.2, the radius ε of the Wasserstein ball enclosing the ambiguity hyperrectangles decays optimally. Here optimality is interpreted in the sense that ε has at least the same decay rate as the Wasserstein distance between

the true distribution and any discrete distribution with the same number of points as the product empirical distribution (a faster decay would otherwise imply that the enclosing ball will eventually contain the true distribution with zero probability). Indeed, assuming again that $\mathrm{supp}(P_\xi)$ has a non-empty interior, we get from [150, Proposition 4.2] that the bound

$$\left(\frac{d}{d+p}\right)^{1/p} C_\star^{-1/d} \frac{1}{N^{n/d}} \leq W_p(Q, P_\xi)$$

always holds for any discrete distribution Q that is supported on N^n points. Then the conclusion follows from (2.5.3) and the fact that $N^{1/d} = N^{n/d}$.

Remark 2.5.5. (Product empirical distribution pushforward for independent sums). Without loss of generality, consider the case where the uncertainty enters the optimization problem as the sum $\zeta \equiv g(\xi_1, \xi_2) := \xi_1 + \xi_2$ of two independent random variables ξ_1 and ξ_2. Then the pushforward of the produced empirical distribution $\boldsymbol{P}_\xi^N$ for $\xi = (\xi_1, \xi_2)$ through g, which provides an estimator for the distribution of ζ, is the convolution of $P_{\xi_1}^N$ and $P_{\xi_2}^N$, namely

$$g_\# \boldsymbol{P}_\xi^N = P_{\xi_1}^N \star P_{\xi_2}^N,$$

where the convolution $\mu \star \nu$ of two distributions μ and ν on $\mathbb{R}^d$ is defined by $\mu \star \nu(B) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \mathbb{1}_B(x+y) \mu \otimes \nu(dx, dy)$ for any $B \in \mathscr{B}(\mathbb{R}^d)$. This choice has also been favored for the density estimation of sums of independent random variables in the statistics literature (see e.g., [151, Example (b), Page 100]).

2.6. DRO REFORMULATIONS OVER AMBIGUITY HYPERRECTANGLES

In this section, we provide dual reformulations for the DRO problem (2.3.2) when the ambiguity set $\mathscr{P}^N$ is the Wasserstein hyperrectangle in (2.4.1a), or the multi-transport hyperrectangle in (2.4.6). Namely, we provide equivalent forms of the problem, which avoid the maximization over the space of probability distributions and are a stepping stone to obtain tractable optimization algorithms. We are therefore interested in reformulating the inner maximization problem $\sup_{P \in \mathscr{P}^N} \mathbb{E}_P[h(\xi)]$, where we fix the decision variable x in (2.3.2) and denote $h(\xi) := f(x, \xi)$ to facilitate notation. We next provide dual reformulations of these problems, first for Wasserstein hyperrectangles and then for multi-transport hyperrectangles. The former reformulations are applicable to a narrower class of objective functions because Wasserstein hyperrectangles are non-convex. Nevertheless, these reformulations provide sharper results since Wasserstein hyperrectangles are typically strictly contained inside multi-transport hyperrectangles.

2.6.1. DUAL REFORMULATIONS OVER WASSERSTEIN HYPERRECTANGLES

Here we provide dual reformulations of the DRO problem (2.3.2) when the set Ξ where the distribution is supported has the product-structure of Assumption 1(i) and

the ambiguity set $\mathscr{P}^N$ is the Wasserstein hyperrectangle $\mathscr{H}_p(Q,\boldsymbol{\varepsilon})$ for some product distribution $Q = Q_1 \times \cdots \times Q_n$. Thus, we are interested to determine the dual of the inner problem

$$\sup_{P\in\mathscr{H}_p(Q,\boldsymbol{\varepsilon})} \mathbb{E}_P[h(\xi)]. \tag{2.6.1}$$

Due to the non-convexity of $\mathscr{H}_p(Q,\boldsymbol{\varepsilon})$, we restrict the class of objective functions to obtain tractable dual reformulations. In particular, we assume that h can be written as the sum or product of functions that depend only on the individual components of ξ and are integrable with respect to the corresponding marginals of the reference distribution.

Assumption 2. (Sum/product decomposition). (i) The objective function h can be expressed as the sum of upper semicontinuous functions or the product of nonnegative upper semicontinuous functions that depend only on the respective components of the random variable. Namely,

$$h(\xi) = \sum_{k=1}^{n} h_k(\xi_k) \tag{2.6.2a}$$

$$\text{or} \quad h(\xi) = \prod_{k=1}^{n} h_k(\xi_k), \qquad h_k(\xi_k) \geq 0. \tag{2.6.2b}$$

(ii) Each function h_k is integrable with respect to Q_k.

We will use the following strong duality result for the maximization over Wasserstein balls.

Proposition 2.6.1. *(DRO dual over Wasserstein balls [120, Theorem 1]). Given a Polish space Ξ, consider the Wasserstein ball $\mathscr{B}_p(Q,\varepsilon)$ with $Q \in \mathscr{P}_p(\Xi)$ and the upper semicontinuous function $h \in L^1(Q)$. Then*

$$\sup_{P\in\mathscr{B}_p(Q,\varepsilon)} \mathbb{E}_P[h(\xi)] = \inf_{\lambda\geq 0} \int_{\Xi} \sup_{\xi\in\Xi} \{h(\xi) + \lambda(\varepsilon^p - \rho(\zeta,\xi)^p)\} dQ(\zeta).$$

The following result establishes strong duality for DRO problems with Wasserstein hyperrectangles when the objective function satisfies Assumption 2.

Proposition 2.6.2. *(DRO dual over Wasserstein hyperrectangles). Let the objective function h of (2.6.1) satisfy Assumption 2. Then (2.6.1) admits the duals*

$$\inf_{\boldsymbol{\lambda}\geq 0} \sum_{k=1}^{n} \int_{\Xi_k} \sup_{\xi_k\in\Xi_k} \{h_k(\xi_k) + \lambda_k(\varepsilon_k^p - \rho_k(\zeta_k,\xi_k)^p)\} dQ_k(\zeta_k) \tag{2.6.3a}$$

$$\inf_{\boldsymbol{\lambda}\geq 0} \prod_{k=1}^{n} \int_{\Xi_k} \sup_{\xi_k\in\Xi_k} \{h_k(\xi_k) + \lambda_k(\varepsilon_k^p - \rho_k(\zeta_k,\xi_k)^p)\} dQ_k(\zeta_k), \tag{2.6.3b}$$

corresponding to Assumptions (2.6.2a) and (2.6.2b), respectively, where $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_n)$.

Proof. The proof of both parts relies on Proposition 2.6.1 and basic properties of product distributions. For completeness, we provide the relevant details. The derivation of (2.6.3a) follows from the fact that under (2.6.2a),

$$\begin{aligned}
\sup_{P\in\mathscr{H}_p(Q,\boldsymbol{\varepsilon})} \mathbb{E}_P\big[h(\xi)\big] &\overset{(a_\Sigma)}{=} \sup_{P\in\mathscr{H}_p(Q,\boldsymbol{\varepsilon})} \mathbb{E}_P\Big[\sum_{k=1}^n h_k(\xi_k)\Big] \\
&\overset{(b_\Sigma)}{=} \sup_{P_k\in\mathscr{B}_p(Q_k,\varepsilon_k),k\in[n]} \mathbb{E}_{P_1\otimes\cdots\otimes P_n}\Big[\sum_{k=1}^n h_k(\xi_k)\Big] \\
&\overset{(c_\Sigma)}{=} \sup_{P_k\in\mathscr{B}_p(Q_k,\varepsilon_k),k\in[n]} \sum_{k=1}^n \mathbb{E}_{P_1\otimes\cdots\otimes P_n}\big[h_k(\xi_k)\big] \\
&\overset{(d_\Sigma)}{=} \sup_{P_k\in\mathscr{B}_p(Q_k,\varepsilon_k),k\in[n]} \sum_{k=1}^n \mathbb{E}_{P_k}\big[h_k(\xi_k)\big] \\
&\overset{(e_\Sigma)}{=} \sum_{k=1}^n \inf_{\lambda_k\geq 0} \int_{\Xi_k} \sup_{\xi\in\Xi_k} \{h_k(\xi_k)+\lambda_k(\varepsilon_k^p-\rho_k(\zeta_k,\xi_k)^p)\}dQ_k(\zeta_k) \\
&\overset{(f_\Sigma)}{=} \inf_{\boldsymbol{\lambda}\geq 0} \sum_{k=1}^n \int_{\Xi_k} \sup_{\xi\in\Xi_k} \{h_k(\xi_k)+\lambda_k(\varepsilon_k^p-\rho_k(\zeta_k,\xi_k)^p)\}dQ_k(\zeta_k).
\end{aligned}$$

Here, (a_Σ) is a consequence of Assumption 2 and (b_Σ) follows from the definition of the Wasserstein hyperrectangle in (2.4.1a). Linearity of the expectation yields (c_Σ) and (d_Σ) follows by exploiting Fubini's theorem (cf. [136, Theorem 14.19]). To derive (e_Σ), we used Proposition 2.6.1 and (f_Σ) follows from the fact that $\sum_k \inf_{\lambda_k\geq 0}\psi_k(\lambda_k) = \inf_{\boldsymbol{\lambda}\geq 0}\sum_k \psi_k(\lambda_k)$ for any functions ψ_k. Thus, (2.6.3a) holds.

In a similar manner, under (2.6.2b), we have

$$\begin{aligned}
\sup_{P\in\mathscr{H}_p(Q,\boldsymbol{\varepsilon})} \mathbb{E}_P\big[h(\xi)\big] &\overset{(a_\Pi)}{=} \sup_{P_k\in\mathscr{B}_p(Q_k,\varepsilon_k),k\in[n]} \mathbb{E}_{P_1}\Big[\cdots\mathbb{E}_{P_n}\Big[\prod_{k=1}^n h_k(\xi_k)\Big]\cdots\Big] \\
&\overset{(b_\Pi)}{=} \sup_{P_k\in\mathscr{B}_p(Q_k,\varepsilon_k),k\in[n]} \prod_{k=1}^n \mathbb{E}_{P_k}\big[h_k(\xi_k)\big] \\
&\overset{(c_\Pi)}{=} \prod_{k=1}^n \inf_{\lambda_k\geq 0} \int_{\Xi_k} \sup_{\xi\in\Xi_k} \{h_k(\xi_k)+\lambda_k(\varepsilon_k^p-\rho_k(\zeta_k,\xi_k)^p)\}dQ_k(\zeta_k) \\
&\overset{(d_\Pi)}{=} \inf_{\boldsymbol{\lambda}\geq 0} \prod_{k=1}^n \int_{\Xi_k} \sup_{\xi\in\Xi_k} \{h_k(\xi_k)+\lambda_k(\varepsilon_k^p-\rho_k(\zeta_k,\xi_k)^p)\}dQ_k(\zeta_k),
\end{aligned}$$

namely, (2.6.3b) holds. In these derivations, (a_Π) follows from Assumption 2, (2.4.1a), and Fubini's theorem (cf. [136, Theorem 14.19]), and (b_Π) from linearity of the expectation. Furthermore, (c_Π) follows from Proposition 2.6.1 and (d_Π) from the fact that $\prod_k \inf_{\lambda_k\geq 0}\psi_k(\lambda_k) = \inf_{\boldsymbol{\lambda}\geq 0}\prod_k \psi_k(\lambda_k)$ for any nonnegative functions ψ_k. The proof is now complete. □

The following corollary provides the dual reformulation of Proposition 2.6.1 for the case when the center of the Wasserstein hyperrectangle is the product empirical distribution.

Corollary 2.6.3. *(Dual of data-driven Wasserstein hyperrectangles). Let h satisfy Assumption 2(i). Then* (2.6.1) *with $Q \equiv \boldsymbol{P}_\xi^N$ admits the corresponding duals*

$$\inf_{\boldsymbol{\lambda}\geq 0} \sum_{k=1}^{n} \frac{1}{N} \sum_{i=1}^{N} \sup_{\xi_k \in \Xi_k} \{h_k(\xi_k) + \lambda_k(\varepsilon_k^p - \rho_k(\xi_k^i, \xi_k)^p)\}$$
$$\inf_{\boldsymbol{\lambda}\geq 0} \prod_{k=1}^{n} \frac{1}{N} \sum_{i=1}^{N} \sup_{\xi_k \in \Xi_k} \{h_k(\xi_k) + \lambda_k(\varepsilon_k^p - \rho_k(\xi_k^i, \xi_k)^p)\}.$$

2.6.2. DUAL REFORMULATIONS OVER MULTI-TRANSPORT HYPERRECTANGLES

Here, we provide the dual of the inner maximization problem in (2.3.2) when the ambiguity set $\mathscr{P}^N$ is the multi-transport hyperrectangle (2.4.6). Namely, we reformulate the problem

$$\sup_{P \in \mathscr{T}(Q,\boldsymbol{\epsilon})} \mathbb{E}_P[h(\xi)]. \tag{2.6.4}$$

To obtain the dual of (2.6.4), we depart from the necessity of Section 2.6.1 to have a Polish space with a product structure and assume that the uncertainty ξ belongs to a general Polish space Ξ. Our analysis generalizes the duality approach in [120], which obtains dual reformulations of DRO problems where distributional ambiguity is captured through a single optimal transport constraint. As in [120], we make the following assumption for h.

Assumption 3. (Objective function class). The objective function $h : \Xi \to \mathbb{R}$ is upper semicontinuous and $h \in L^1(Q)$.

We also assume the following regarding the cost functions in (2.4.5).

Assumption 4. (Transport costs). (i) For each $k \in [n]$ there exists a nondecreasing sequence $c_{k,m} : \Xi \times \Xi \to \mathbb{R}_{\geq 0}$, $m \in \mathbb{N}$, of *continuous* transport costs with $c_{k,m}(\zeta,\zeta) = 0$ for all ζ and $c_{k,m}(\zeta,\xi) \nearrow c_k(\zeta,\xi) \in \mathbb{R}_{\geq 0}$.
(ii) There exists a compact set $\Xi_{\text{cmp}} \subset \Xi$ such that for each m, $c_{k,m}$, $k \in [n]$ are linearly independent in $C(\Xi_{\text{cmp}} \times \Xi_{\text{cmp}})$ and $\text{span}\{c_{1,m}, \ldots, c_{n,m}\} \cap C_{2,\text{const}}(\Xi_{\text{cmp}} \times \Xi_{\text{cmp}}) = \{0\}$.

Assumption 4 is directly satisfied when Ξ has the product structure of Assumption 1 and the considered cost functions are powers of the distances between the components of the random variable, i.e., when $c_k(\zeta,\xi) = \rho_k(\zeta_k,\xi_k)^p$. We next provide some preparatory definitions and sketch the intuition behind the strong dual to problem (2.6.4), which is given later in this section.

Let $\mathscr{I}(\pi) := \int_{\Xi\times\Xi} h(\xi) d\pi(\zeta,\xi)$ and consider the set $\mathscr{P}^h(\Xi)$ of distributions ν on Ξ for which the integral of h is well defined and takes values in $\bar{\mathbb{R}}$, namely, for which either $\int_\Xi h_+(\xi) d\nu(\xi) \in \mathbb{R}$ or $\int_\Xi h_-(\xi) d\nu(\xi) \in \mathbb{R}$, were $h_+ := \max\{h, 0\}$ and $h_- := \min\{h, 0\}$. Denoting further $\Pi^h(Q) := \{\pi \in \Pi(Q) : \text{pr}_{2\#}\pi \in \mathscr{P}^h(\Xi)\}$ and analogously $\Pi^h(Q,\boldsymbol{\epsilon})$, and

recalling the definition of $\mathscr{T}_p(Q,\boldsymbol{\epsilon})$, allows us to rigorously (re)define the DRO problem (2.6.4) as

$$\sup_{P\in\mathscr{T}_p(Q,\boldsymbol{\epsilon})\cap\mathscr{P}^h(\Xi)}\int_\Xi h(\xi)dP(\xi)=\sup_{\pi\in\Pi^h(Q,\boldsymbol{\epsilon})}\int_{\Xi\times\Xi}h(\xi)d\pi(\zeta,\xi)=\sup_{\pi\in\Pi^h(Q,\boldsymbol{\epsilon})}\mathscr{I}(\pi)=:\mathscr{I}^\star. \quad (2.6.5)$$

Due to (2.4.5), this is a linear optimization problem in the space of finite signed measures on $\Xi\times\Xi$. When restricted further over the convex set of probability measures

$$\Pi_{\text{fin},\boldsymbol{c}}(Q):=\Big\{\pi\in\Pi^h(Q):\int_{\Xi\times\Xi}c_k(\zeta,\xi)d\pi(\zeta,\xi)<+\infty\text{ for all }k\in[n]\Big\},$$

over which the integrals of the costs are real-valued, and taking into account the inequality constraints (2.4.5), which always imply $\Pi^h(Q,\boldsymbol{\epsilon})\subset\Pi_{\text{fin},\boldsymbol{c}}(Q)$, its Lagrangian is given by

$$\begin{aligned}\mathscr{L}(\pi,\boldsymbol{\lambda}):&=\int_{\Xi\times\Xi}h(\xi)d\pi(\zeta,\xi)+\sum_{k=1}^n\lambda_k\Big(\epsilon_k-\int_{\Xi\times\Xi}c_k(\zeta,\xi)d\pi(\zeta,\xi)\Big)\\&=\langle\boldsymbol{\lambda},\boldsymbol{\epsilon}\rangle+\int_{\Xi\times\Xi}(h(\xi)-\langle\boldsymbol{\lambda},\boldsymbol{c}(\zeta,\xi)\rangle)d\pi(\zeta,\xi),\end{aligned}\quad (2.6.6)$$

where $\boldsymbol{\lambda}=(\lambda_1,\ldots,\lambda_n)\in\mathbb{R}^n_{\geq 0}$, $\boldsymbol{\epsilon}=(\epsilon_1,\ldots,\epsilon_n)$, and $\boldsymbol{c}(\zeta,\xi)=(c_1(\zeta,\xi),\ldots,c_n(\zeta,\xi))$. From the definition of $\mathscr{L}(\pi,\boldsymbol{\lambda})$, we have

$$\begin{aligned}\mathscr{I}^\star&=\sup_{\pi\in\Pi_{\text{fin},\boldsymbol{c}}(Q)}\inf_{\boldsymbol{\lambda}\geq 0}\mathscr{L}(\pi,\boldsymbol{\lambda})\\&=\sup_{\pi\in\Pi_{\text{fin},\boldsymbol{c}}(Q)}\inf_{\boldsymbol{\lambda}\geq 0}\Big\{\langle\boldsymbol{\lambda},\boldsymbol{\epsilon}\rangle+\int_{\Xi\times\Xi}(h(\xi)-\langle\boldsymbol{\lambda},\boldsymbol{c}(\zeta,\xi)\rangle)d\pi(\zeta,\xi)\Big\}\end{aligned}\quad (2.6.7)$$

and we get from the min–max inequality that

$$\mathscr{I}^\star\leq\inf_{\boldsymbol{\lambda}\geq 0}\Big\{\langle\boldsymbol{\lambda},\boldsymbol{\epsilon}\rangle+\sup_{\pi\in\Pi_{\text{fin},\boldsymbol{c}}(Q)}\int_{\Xi\times\Xi}(h(\xi)-\langle\boldsymbol{\lambda},\boldsymbol{c}(\zeta,\xi)\rangle)d\pi(\zeta,\xi)\Big\}. \quad (2.6.8)$$

To provide some intuition behind the dual problem to (2.6.4), assume for the moment that Ξ is compact and that h and the cost functions $c_1,\ldots,c_n$ are continuous. In this case we have that $\Pi_{\text{fin},\boldsymbol{c}}(Q)=\Pi(Q)$ and the maximization problem

$$\sup_{\pi\in\Pi(Q)}\int_{\Xi\times\Xi}(h(\xi)-\langle\boldsymbol{\lambda},\boldsymbol{c}(\zeta,\xi)\rangle)d\pi(\zeta,\xi)$$

is a linear program that can be written in the abstract form

$$\sup\langle e,\pi\rangle_1$$

The same argument can be used to resolve potential ambiguities in the reformulations of Section 2.6.1 when there are distributions in the ambiguity set that may lead to integrals of the form $+\infty-\infty$. An alternative way to address this issue is to define $+\infty-\infty=+\infty$ as in [120]. Endnote 2 in [120] also clarifies why such ambiguities do not affect the interpretation of the optimization problem.

$$\text{s.t.}\ \mathscr{A}\pi = b$$
$$\pi \succeq 0.$$

Here $\mathscr{A} \equiv \mathrm{pr}_{1\#} : \mathscr{M}(\Xi \times \Xi) \to \mathscr{M}(\Xi)$, $b \equiv Q \in \mathscr{M}(\Xi)$, $e \equiv h \circ \mathrm{pr}_2 - \sum_{k=1}^n \lambda_k c_k \in C(\Xi \times \Xi)$, $\langle \cdot, \cdot \rangle_1$ denotes the duality between $C(\Xi \times \Xi)$ and $\mathscr{M}(\Xi \times \Xi)$, and the order $\succeq$ is taken with respect to the cone of positive measures on $\Xi \times \Xi$. By linear programming duality (cf. [152]), its dual problem is given by

$$\inf \langle \varphi, b \rangle_2$$
$$\text{s.t.}\ \mathscr{A}^* \varphi \succeq e.$$

Here $\mathscr{A}^* \equiv \mathscr{K}_{\mathrm{pr}_1} : C(\Xi) \to C(\Xi \times \Xi)$ is the adjoint of $\mathscr{A}$, namely the composition, a.k.a. Koopman operator (cf. [153, Chapter 4.3]), with $\mathscr{K}_{\mathrm{pr}_1}(\varphi) := \varphi \circ \mathrm{pr}_1$, $\langle \cdot, \cdot \rangle_2$ denotes the duality between $C(\Xi)$ and $\mathscr{M}(\Xi)$, and the order $\succeq$ is taken with respect to the cone of positive continuous functions on Ξ. This in turn is an abstract representation of the dual problem

$$\inf \Big\{ \int_\Xi \varphi(\zeta) dQ(\zeta) : \varphi \in C(\Xi) \text{ and } \varphi(\zeta) \geq h(\xi) - \langle \boldsymbol{\lambda}, \boldsymbol{c}(\zeta, \xi) \rangle \text{ for all } \zeta, \xi \in \Xi \Big\}.$$

Based on these considerations, we introduce the dual of (2.6.5) in the general case, where Ξ does not need to be compact and h, c_k are not necessarily continuous. To this end, we denote

$$\Lambda \equiv \Lambda(h; c_1, \ldots, c_n) := \Big\{ (\boldsymbol{\lambda}, \varphi) : \boldsymbol{\lambda} \geq 0, \varphi \in \mathfrak{m}_{\mathscr{U}}(\Xi; \mathbb{R} \cup \{+\infty\}) \text{ and } \varphi \circ \mathrm{pr}_1 \geq h \circ \mathrm{pr}_2 - \sum_{k=1}^n \lambda_k c_k \Big\}, \tag{2.6.9a}$$

$$\mathscr{J}(\boldsymbol{\lambda}, \varphi) := \langle \boldsymbol{\lambda}, \boldsymbol{\epsilon} \rangle + \int_\Xi \varphi(\zeta) dQ(\zeta) \tag{2.6.9b}$$

and consider in analogy to [120] the dual problem

$$\begin{aligned} \mathscr{J}_\star &:= \inf_{(\boldsymbol{\lambda}, \varphi) \in \Lambda} \mathscr{J}(\boldsymbol{\lambda}, \varphi) \\ &= \inf_{\boldsymbol{\lambda} \geq 0} \Big\{ \langle \boldsymbol{\lambda}, \boldsymbol{\epsilon} \rangle + \inf \Big\{ \int_\Xi \varphi(\zeta) dQ(\zeta) : \varphi \in \mathfrak{m}_{\mathscr{U}}(\Xi; \mathbb{R} \cup \{+\infty\}) \text{ and } \\ &\qquad \varphi(\zeta) \geq h(\xi) - \langle \boldsymbol{\lambda}, \boldsymbol{c}(\zeta, \xi) \rangle \text{ for all } \zeta, \xi \in \Xi \Big\} \Big\}. \end{aligned} \tag{2.6.10}$$

Then it follows from (2.6.8) that

$$\mathscr{I}^\star \leq \mathscr{J}_\star. \tag{2.6.11}$$

From Assumptions 3 and 4(i), it follows that for any function $\varphi \in \mathfrak{m}_{\mathscr{U}}(\Xi; \mathbb{R} \cup \{+\infty\})$ the integral $\int_\Xi \varphi(\zeta) dQ(\zeta)$ is well defined. This ensures that the integral of the function $\varphi_{\boldsymbol{\lambda}}(\zeta) := \sup_{\xi \in \Xi} \{h(\xi) - \langle \boldsymbol{\lambda}, \boldsymbol{c}(\zeta, \xi) \rangle\}$ in the attainable dual pair in Theorem 2.6.4 is also well defined, since $\varphi_{\boldsymbol{\lambda}}$ is universally measurable (cf. [120, Page 16] for the justification of this fact). Measurability of the integrands in the dual reformulations of Section 2.6.1 is guaranteed in the same way.

The establishment of strong duality between the primal optimization problem and its dual hinges on showing that the reverse inequality also holds. Its proof is given in Appendix 2.B and it is based on appropriate modifications of the technical approach developed in [120].

2

Theorem 2.6.4. *(DRO dual over multi-transport hyperrectangles). Consider the problem* (2.6.4) *and let* h *and* $c_1, \ldots, c_n$ *satisfy Assumptions 3 and 4, respectively. Then*

$$\mathscr{I}^\star = \mathscr{J}_\star = \inf_{\boldsymbol{\lambda} \geq 0} \left\{ \langle \boldsymbol{\lambda}, \boldsymbol{\epsilon} \rangle + \int_\Xi \sup_{\xi \in \Xi} \{ h(\xi) - \langle \boldsymbol{\lambda}, \boldsymbol{c}(\zeta, \xi) \rangle \} dQ(\zeta) \right\} \tag{2.6.12}$$

and there exist $(\boldsymbol{\lambda}, \varphi_{\boldsymbol{\lambda}}) \in \Lambda$ *with* $\varphi_{\boldsymbol{\lambda}}(\zeta) := \sup_{\xi \in \Xi} \{ h(\xi) - \langle \boldsymbol{\lambda}, \boldsymbol{c}(\zeta, \xi) \rangle \}$, $\zeta \in \Xi$, *for which the infimum in* (2.6.10) *is attained.*

Using this result, we obtain the following explicit reformulation of the DRO problem when Ξ has a product structure and the ambiguity set is a data-driven multi-transport hyperrectangle.

Corollary 2.6.5. *(Dual of data-driven multi-transport hyperrectangles). If* Ξ *satisfies Assumption 1(i), then the optimal value of* (2.6.4) *with* $\mathscr{T}(Q, \boldsymbol{\epsilon}) \equiv \mathscr{T}_p(\boldsymbol{P}_{\hat{\xi}}^N, \boldsymbol{\epsilon})$ *is equal to*

$$\mathscr{I}^\star = \inf_{\boldsymbol{\lambda} \geq 0} \left\{ \langle \boldsymbol{\lambda}, \boldsymbol{\epsilon} \rangle + \frac{1}{N^n} \sum_{(i_1, \ldots, i_n) \in [N]^n} \sup_{\xi \in \Xi} \left\{ h(\xi) - \sum_{k=1}^n \lambda_k \rho_k(\xi_k^{i_k}, \xi_k)^p \right\} \right\}. \tag{2.6.13}$$

Remark 2.6.6. (Tractable reformulations and numerical complexity induced by the dual reformulation (2.6.13)). The dual expression (2.6.13) enables the derivation of tractable reformulations for certain classes of DRO problems over multi-transport hyperrectangles [96]. Although the complexity of these reformulations may grow exponentially with the number of independent components—due to the size of the product empirical distribution (2.4.1b)—this issue can be mitigated through clustering. This yields ambiguity sets of lower complexity that maintain their desirable probabilistic guarantees while still shrinking faster compared to their monolithic counterparts [96].

Remark 2.6.7. (Strict improvement of Wasserstein hyperrectangle optimal value). Despite the fact that multi-transport hyperrectangles admit dual reformulations over a much broader class of objective functions compared to Wasserstein hyperrectangles, the latter can exhibit a strict improvement of their optimal values compared to the former. This is justified by the containment result of Proposition 2.4.4 and is illustrated in the following toy example.

Consider the product reference distribution

$$\begin{aligned} Q = Q_1 \otimes Q_2 &:= (p_1 \delta_0 + (1 - p_1)\delta_1) \otimes (p_2 \delta_0 + (1 - p_2)\delta_1) \\ &= p_1 p_2 \delta_{(0,0)} + (1 - p_1) p_2 \delta_{(1,0)} + p_1 (1 - p_2) \delta_{(0,1)} + (1 - p_1)(1 - p_2) \delta_{(1,1)} \end{aligned}$$

on $\mathbb{R}^2$, the objective function

$$h(\xi) := \mathbb{1}_{\{(0,0)\}}(\xi) \equiv \mathbb{1}_{\{0\}}(\xi_1) \mathbb{1}_{\{0\}}(\xi_2),$$

and ambiguity radii $\varepsilon_1 \le 1-p_1$ and $\varepsilon_2 \le 1-p_2$. For each of the ambiguity sets $\mathscr{H}_p(Q,\boldsymbol{\varepsilon})$ and $\mathscr{T}_p(Q,\boldsymbol{\varepsilon})$, the distribution that maximizes the value of h is the one obtained when the largest possible amount of mass is transferred from the reference distribution to the point $(0,0)$.

For the Wasserstein hyperrectangle, this distribution is obtained through the transport plans π_1 and π_2, which move the largest possible amount of mass from 1 to 0 using the transport budgets ε_1 and ε_2, respectively. Identifying these transport plans with their restrictions to $\{0,1\}^2 \subset \mathbb{R}^2$ where they are supported, we obtain their matrix representations

$$\pi_1 \equiv \underbrace{\begin{bmatrix} p_1 & \varepsilon_1 \\ 0 & 1-p_1-\varepsilon_1 \end{bmatrix}}_{Q_1} \Big\} P_1 \quad \text{and} \quad \pi_2 \equiv \underbrace{\begin{bmatrix} p_2 & \varepsilon_2 \\ 0 & 1-p_2-\varepsilon_2 \end{bmatrix}}_{Q_2} \Big\} P_2.$$

The column sums of these transport plan matrices correspond to the reference distributions Q_1 and Q_2 and their row sums to the other marginals $P_1 := (p_1+\varepsilon_1)\delta_0 + (1-p_1-\varepsilon_1)\delta_1$ and $P_2 := (p_2+\varepsilon_2)\delta_0 + (1-p_2-\varepsilon_2)\delta_1$ of the transport plans. The distribution from $\mathscr{H}_p(Q,\boldsymbol{\varepsilon})$ that maximizes h is $P := P_1 \otimes P_2$, which is also the second marginal of the transport plan $\pi = T_\#(\pi_1 \otimes \pi_2)$ (its first marginal is Q), where $T(\zeta_1,\xi_1,\zeta_2,\xi_2) := (\zeta_1,\zeta_2,\xi_1,\xi_2)$ (see proof of Proposition 2.4.4). As above, we identify the transport plan π with its restriction to $\{(0,0),(1,0),(0,1),(1,1)\}^2 \subset \mathbb{R}^4$ where it is supported. Using the lexicographical ordering

$$(0,0) \quad (1,0) \quad (0,1) \quad (1,1) \quad \mapsto \quad 1 \quad 2 \quad 3 \quad 4$$

and the product expression $\pi(\zeta,\xi) = \pi_1(\zeta_1,\xi_1)\pi_2(\zeta_2,\xi_2)$ of π, we get its matrix representation

$$\pi \equiv \underbrace{\begin{bmatrix} p_1p_2 & \varepsilon_1p_2 & p_1\varepsilon_2 & \varepsilon_1\varepsilon_2 \\ 0 & (1-p_1-\varepsilon_1)p_2 & 0 & (1-p_1-\varepsilon_1)\varepsilon_2 \\ 0 & 0 & p_1(1-p_2-\varepsilon_2) & \varepsilon_1(1-p_2-\varepsilon_2) \\ 0 & 0 & 0 & \prod_{k=1}^2(1-p_k-\varepsilon_k) \end{bmatrix}}_{Q} \Big\} P$$

For the multi-transport hyperrectangle, it is not hard to check that the distribution that maximizes the value of h is obtained through the transport plan π', which uses the transport budgets ε_1 and ε_2 to move the largest possible amounts of mass from $(1,0)$ to $(0,0)$ and from $(0,1)$ to $(0,0)$, respectively. Thus, its corresponding matrix representation is

$$\pi' \equiv \underbrace{\begin{bmatrix} p_1p_2 & \varepsilon_1 & \varepsilon_2 & 0 \\ 0 & (1-p_1)p_2-\varepsilon_1 & 0 & 0 \\ 0 & 0 & p_1(1-p_2)-\varepsilon_2 & 0 \\ 0 & 0 & 0 & \prod_{k=1}^2(1-p_k) \end{bmatrix}}_{Q} \Big\} P'$$

Assuming without loss of generality that $p_1 < 1$ and taking into account that

$$\varepsilon_1 \le 1-p_1 \quad \text{and} \quad \varepsilon_2 \le 1-p_2,$$

it follows that the mass transported to $(0,0)$ with the Wasserstein hyperrectangle is strictly less than that with the multi-transport hyperrectangle, namely,

$$\begin{aligned}\varepsilon_1 p_2 + p_1\varepsilon_2 + \varepsilon_1\varepsilon_2 &= \varepsilon_1 p_2 + \varepsilon_2(p_1 + \varepsilon_1)\\ &\leq \varepsilon_1 p_2 + \varepsilon_2(p_1 + 1 - p_1)\\ &= \varepsilon_1 p_2 + \varepsilon_2\\ &< \varepsilon_1 + \varepsilon_2.\end{aligned}$$

Therefore, we get that

$$\mathbb{E}_P[h(\xi)] = p_1 p_2 + \varepsilon_1 p_2 + p_1\varepsilon_2 + \varepsilon_1\varepsilon_2 < p_1 p_2 + \varepsilon_1 + \varepsilon_2 = \mathbb{E}_{P'}[h(\xi)],$$

i.e., that the optimal value over the Wasserstein hyperrectangle is strictly below that of the multi-transport hyperrectangle (cf. Figure 2.4).

Since the distribution from $\mathcal{T}_p(Q,\boldsymbol{\varepsilon})$ with the largest amount of mass at $(0,0)$ has more mass at that point than any distribution from $\mathcal{H}_p(Q,\boldsymbol{\varepsilon})$, it follows that this distribution cannot belong to the convex hull of $\mathcal{H}_p(Q,\boldsymbol{\varepsilon})$. This is also why we depict the multi-transport hyperrectangle as a curved rectangle, which is strictly convex, instead of drawing a straight rectangle that would look like the convex hull of the Wasserstein hyperrectangle.

The fact that DRO problems over Wasserstein hyperrectangles can exhibit improved performance compared to their formulations over multi-transport hyperrectangles is also established in simulation for uncertainty quantification problems in the next chapter. This motivates considering the former ambiguity sets for the narrower class of problems where the duality results of Proposition 2.6.2 are applicable.

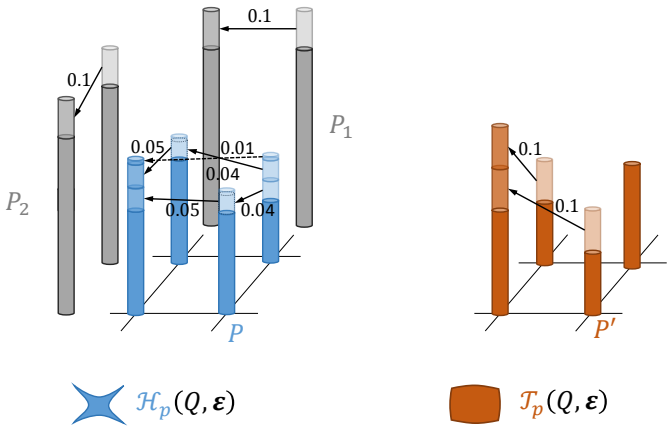


Figure 2.4.: The figure illustrates the optimal transport plans corresponding to both ambiguity sets for the concrete values $p_1 = p_2 = 0.5$ and $\varepsilon_1 = \varepsilon_1 = 0.1$. Clearly, the mass transferred to $(0,0)$ to form the distribution P that maximizes h in the case of the Wasserstein hyperrectangle is considerably smaller than that used to form P' in the case of the multi-transport hyperrectangle. This happens because P needs to retain a product distribution structure, resulting in a redundant effort to transport mass from $(1,1)$ to all points $(0,1)$, $(1,0)$, and $(0,0)$, which is not required to form P'.

2.7. SIMULATION EXAMPLE

In this section, we apply our duality results for multi-transport hyperrectangles to solve a distributionally robust power dispatch problem and compare the performance of our ambiguity sets to traditional Wasserstein balls. Our goal is to cover an excess daily power demand $D_t + \xi_1$ (it may also be negative), consisting of a deterministic time-varying term D_t, $t \in [T]$, which is known, and a stochastic term ξ_1. To this end, we use power ξ_2 from renewable energy resources, which is random, and x from the grid. Each day, we are informed about the deterministic part of the demand for the next day and optimize the scheduled power x from the grid for the case where the overall power demand on the next day cannot be met by renewables. Minimizing the expected demand-supply discrepancy yields the sequence of stochastic optimization problems

$$\min_{x \geq 0} \mathbb{E}_{P_\xi}[\mathbb{1}_\Theta(\xi)\,|D_t + \xi_1 - \xi_2 - x_t|] \qquad t \in [T], \tag{2.7.1}$$

where P_ξ is the distribution of $\xi = (\xi_1, \xi_2)$ with support Ξ and $\Theta := \{\xi \in \Xi : D_t + \xi_1 - \xi_2 \geq 0\}$. We also assume that ξ_1 and ξ_2 are independent and P_ξ is unknown. Using only i.i.d. historical data $\xi^1, \ldots, \xi^N$ from it, we reformulate (2.7.1) as the sequence of DRO problems

$$\min_{x \geq 0} \sup_{P \in \mathscr{P}^N} \mathbb{E}_{P_\xi}[\mathbb{1}_\Theta(\xi)\,|D_t + \xi_1 - \xi_2 - x_t|] \quad t \in [T], \tag{2.7.2}$$

where $\mathscr{P}^N$ denotes either a multi-transport hyperrectangle $\mathscr{T}_1(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})$ or a Wasserstein ball $\mathscr{B}_1(P_\xi^N, \varepsilon)$. Using Corollary 2.6.5 and similar arguments as in [79, proof of Corollary 5.1] we reformulate this problem with $\mathscr{P}^N \equiv \mathscr{T}_1(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})$ as the linear program (the detailed derivations are given in 2.C)

$$\begin{aligned}
&\inf_{\boldsymbol{\lambda} \geq 0, \boldsymbol{s}, \boldsymbol{\gamma}, x} \langle \boldsymbol{\lambda}, \boldsymbol{\varepsilon} \rangle + \frac{1}{N^2} \sum_{\boldsymbol{i} \in [N]^2} s_{\boldsymbol{i}} \\
\text{s.t. } &\langle \gamma_1^{\boldsymbol{i}}, d - C\xi^{\boldsymbol{i}} \rangle + D_t - x + \langle [1 \;\; -1]^\top, \xi^{\boldsymbol{i}} \rangle \leq s_{\boldsymbol{i}} \\
&\langle \gamma_2^{\boldsymbol{i}}, d - C\xi^{\boldsymbol{i}} \rangle - D_t + x + \langle [-1 \;\; 1]^\top, \xi^{\boldsymbol{i}} \rangle \leq s_{\boldsymbol{i}} \\
&\left| \mathrm{pr}_k(C^\top \gamma_1^{\boldsymbol{i}} - [1 \;\; -1]^\top) \right| \leq \lambda_k \\
&\left| \mathrm{pr}_k(C^\top \gamma_2^{\boldsymbol{i}} - [-1 \;\; 1]^\top) \right| \leq \lambda_k \\
&\gamma_j^l \geq 0 \qquad\qquad \boldsymbol{i} \in [N]^2,\ j \in [2], k \in [2],
\end{aligned} \tag{2.7.3}$$

where $\boldsymbol{\varepsilon} := (\varepsilon_1, \varepsilon_2)$, $\boldsymbol{i} = (i_1, i_2)$ $\boldsymbol{\gamma} := \{(\gamma_1^{i_1}, \gamma_2^{i_2})\}_{\boldsymbol{i} \in [N]^2}$, and $\boldsymbol{s} := \{s_{\boldsymbol{i}}\}_{\boldsymbol{i} \in [N]^2} \in \mathbb{R}^{N^2}$, with $\{s_{\boldsymbol{i}}\}_{\boldsymbol{i} \in [N]^2}$ viewed as an element of $\mathbb{R}^{N^2}$ for some ordering of $[N]^2$. The reformulation when $\mathscr{P}^N \equiv \mathscr{B}_1(P_\xi^N, \varepsilon)$ is analogous and we omit it due to space constraints.

For the simulations, we select the probability distributions $P_{\xi_1} = 2/3\delta_{-5} + 1/3\mathscr{U}_{[5,10]}$ for the stochastic part of the extra demand and $P_{\xi_2} := 1/3\mathscr{U}_{[0,5]} + 2/3\mathscr{U}_{[15,20]}$ for the renewable power supply, where $\mathscr{U}$ denotes the uniform distribution on the designated set. Both ambiguity sets are built using $N = 10$ i.i.d. samples from $P_\xi = P_{\xi_1} \otimes P_{\xi_2}$. We set $T := 3$ and select the deterministic demand sequence $3, 0, -3$.

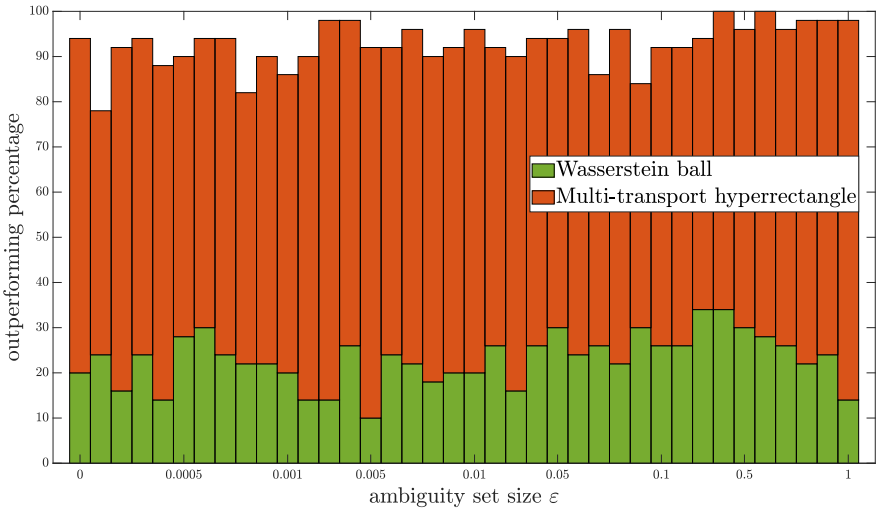


Figure 2.5.: The bar chart presents the percentage of times each ambiguity set exhibits superior out-of-sample performance across 50 trials, for the demand $D_1 = 3$ and multiple values of ε. Both the multi-transport hyperrectangle and the monolithic ball are generated by the same set of $N = 10$ samples. The radius ε of the monolithic ball is equal to that of the smallest ball enclosing the rectangle, represented by the yellow ball in Figure 2.1. Despite their comparable size, the multi-transport hyperrectangle outperforms its monolithic counterpart for all values of ε.

In the first set of simulations, we compare the out-of-sample performance of the multi-transport hyperrectangle and the monolithic ball $\mathscr{B}_1(P_\xi^N, \varepsilon)$ with the same radius as the smallest ball enclosing the rectangle. We perform this for multiple values of ε. For each ε, we assess the average performance over 50 random centers of ambiguity sets. We solve the DRO problems over both sets to determine their optimal decisions and compare their corresponding expected costs, i.e., their out-of-sample performance. The results, all for the same demand $D_1 = 3$ at $t = 1$, are shown in Figure 2.5. The bar chart shows the percentage of times the out-of-sample performance of each ambiguity set is optimal for every value of ε. Clearly, $\mathscr{T}_1(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})$ outperforms $\mathscr{B}_1(P_\xi^N, \varepsilon)$ for all ε, validating that it is better informed by the probabilistic model.

In our next experiment, we determine the smallest size ε for each ambiguity set that validates the certificate

$$\inf_{x \geq 0} \mathbb{E}_{P_\xi}[f_t(x, \xi)] \leq \inf_{x \geq 0} \sup_{P \in \mathscr{P}^N} \mathbb{E}_P[f_t(x, \xi)]) \tag{2.7.4}$$

with a prescribed confidence $1 - \beta$. To this end, we initialize ε at zero, and solve the DRO problems (2.7.1) for 100 realizations of the empirical distribution. Out of these, we count how many times the certificate (2.7.4) is met, and we increase ε until the relative frequency of (2.7.4) being true meets our desired confidence level. Figure 2.6 shows the minimum value of ε required to achieve a confidence level of 80% in (2.7.4) for each optimization problem individually, as well as for all problems

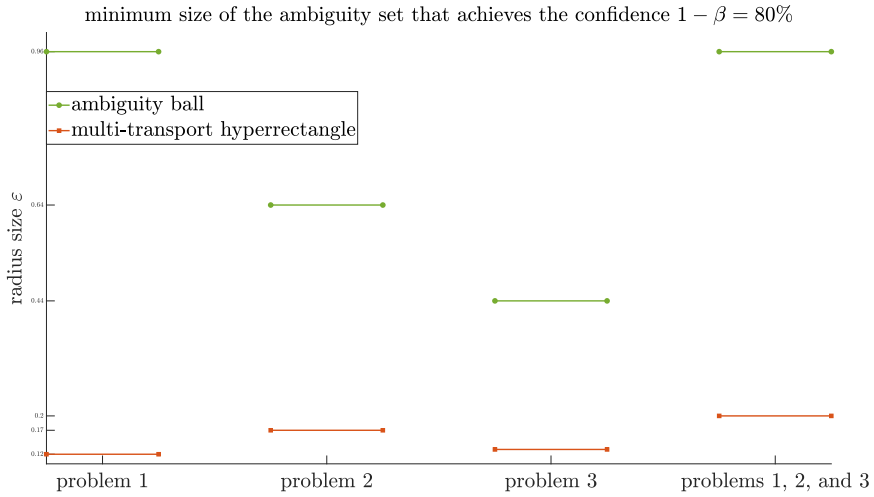


Figure 2.6.: The figure illustrates the minimum size ε of the monolithic ball and the smallest enclosing ball of the hyperrectangle to ensure that (2.7.4) is met with confidence 80% when $N = 10$ and $t \in [3]$. The size of the hyperrectangle to achieve the required confidence is significantly smaller. The figure also shows the smallest ε to ensure that all three problems simultaneously meet (2.7.4) with the same confidence, which is consistently larger than the ε required for the individual problems.

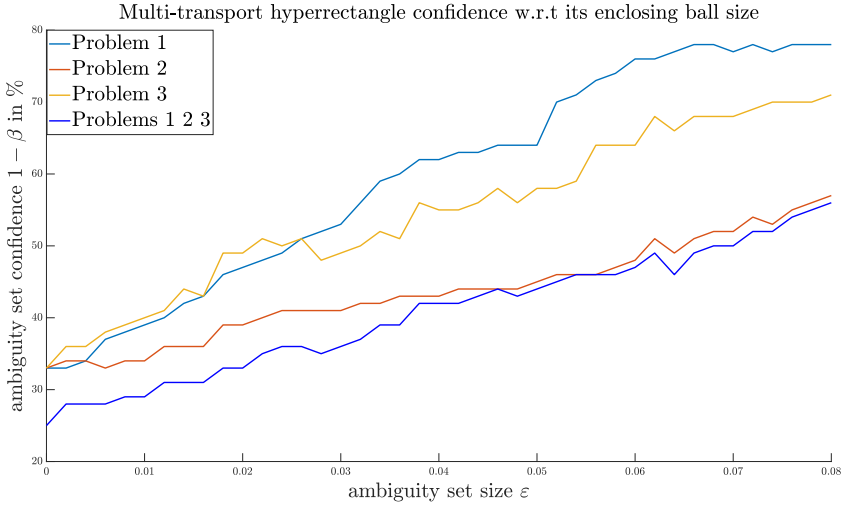


Figure 2.7.: The figure illustrates how the frequency of satisfying (2.7.4) evolves with the radius ε of the smallest ball enclosing the hyperrectangle. It is evident that simply selecting the maximum of the radii that ensure the desired confidence level for each optimization problem is insufficient to ensure the same for all problems simultaneously.

simultaneously. Clearly, $\mathscr{T}_1(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})$ requires a significantly smaller ε compared to

$\mathscr{B}_1(P_\xi^N, \varepsilon)$ to meet the desired confidence level.

Finally, we show the evolution of the confidence level with which (2.7.4) is met for each of the three DRO problems in (2.7.2) with $t \in [3]$, as a function of ε. We also compare it to the confidence level required for all three problems to meet the certificate simultaneously. The results for the hyperrectangle are shown in Figure 2.7. As expected, the confidence level with which each optimization problem meets the certificate increases with the size ε of the ambiguity set. In addition, for each ε, the confidence level of each problem meeting the certificates is consistently higher than the confidence of all three certificates to hold simultaneously. This highlights how multiple optimization problems affect the size of the ambiguity when it is required that the DRO cost upper bounds the out-of-sample cost with a desired confidence. The same observation is illustrated in Figure 2.6, which shows that the minimum ε needed to achieve an 80% confidence level across all optimization problems simultaneously is larger than the maximum of the smallest ε required to attain the same level for each problem individually. This verifies that solving multiple DRO problems under the same uncertainty requires larger ambiguity sets to ensure that their solutions remain consistently reliable for all problems simultaneously.

2.8. CONCLUSION

In this chapter, we introduced two classes of structured ambiguity sets, termed Wasserstein hyperrectangles and multi-transport hyperrectangles. In data-driven scenarios where the components of the uncertainty are statistically independent, both ambiguity sets can be tuned to contain the true distribution with prescribed confidence while exhibiting considerably faster shrinkage with the number of samples compared to monolithic ambiguity balls. We established strong duality results for DRO problems over both ambiguity sets and clarified the tradeoff between the scope of the problems that can be effectively solved for each set and the potential conservativeness of their solutions. Our numerical results certify how structured ambiguity sets can capture the uncertainty in a more effective manner than monolithic ambiguity balls and improve the task of distributionally robust decision-making. Future work includes extending the statistical guarantees for structured dependencies across the components of the uncertainty.

Appendix

2.A. PROOF FROM SECTIONS 2.4 AND 2.5

2.A.1. PROOFS FROM SECTION 2.4

The following lemma is used to prove Lemma 2.4.2.

Lemma 2.A.1. *(Independent σ-algebras [136, Theorem 2.26]). Let K be an arbitrary set and $I_k, k \in K$, arbitrary mutually disjoint index sets. Define $I = \cup_{k\in K} I_k$. If the family $\{X_i\}_{i\in I}$ is independent, then the family of σ-algebras $\{\sigma(X_j, j \in I_k)\}_{k\in K}$ is independent.*

Proof of Lemma 2.4.2. Consider the index sets $I_k = \{(k,1),\dots,(k,N)\}$ for $k \in K := [n]$ and let $I := \cup_{k\in K} I_k$. Denote by $\mathscr{F}_{I_k}$, $k \in K$, the σ-algebra generated by $\{\xi_k^i\}_{(k,i)\in I_k}$, where ξ_k^i denotes the kth component of the ith sample ξ^i. Then by Assumption 1, the fact that $\xi^1,\dots,\xi^N$ are i.i.d., and Lemma 2.A.1, the σ-algebras $\mathscr{F}_{I_k}$ are independent. Next, since $W_p(\mu_X^N,\mu_Y^N) \le (\frac{1}{N}\sum_{i=1}^N \rho(X^i,Y^i)^p)^{\frac{1}{p}}$ for any discrete distributions $\mu_X^N = \frac{1}{N}\sum_{i=1}^N \delta_{X^i}$ and $\mu_Y^N = \frac{1}{N}\sum_{i=1}^N \delta_{Y^i}$ on the Polish space Ξ (cf. [109, proof of Lemma A.2]), we deduce that each mapping $\Xi_k^N \ni (\xi_k^1,\dots,\xi_k^N) \mapsto W_p(P^N_{\xi_k}, P_{\xi_k}) \in \mathbb{R}$ is continuous, and hence, also measurable. Thus, $\sigma(W_p(P^N_{\xi_k}, P_{\xi_k})) \subset \mathscr{F}_{I_k}$ for each $k \in [n]$ and since the σ-algebras $\mathscr{F}_{I_k}$ are independent, the events $\{W_p(P^N_{\xi_k}, P_{\xi_k}) \le \varepsilon_k\}$, $k \in [n]$ are also independent. □

For the proof of Proposition 2.4.6 we will use the following auxiliary results, which relate the Wasserstein distance of two distributions in a product space with their transport cost discrepancy across the components of the product.

Proposition 2.A.2. *(Wasserstein distance of transport distributions). Consider the distributions $P,Q \in \mathscr{P}_p(\Xi)$, where $\Xi = \Xi_1 \times \cdots \times \Xi_n$ is endowed with the metric $\rho := \left(\sum_{k=1}^n \rho_k^q\right)^{1/q}$ for some $q \ge 1$ and ρ_k is the metric on Ξ_k. Assume also that there exists a transport plan $\pi \in \mathscr{C}(Q,P)$ with*

$$\int_{\Xi\times\Xi} \rho_k(\zeta_k,\xi_k)^p \, d\pi(\zeta,\xi) \le \varepsilon_k^p, \quad k \in [n], \tag{2.A.1}$$

for certain $\varepsilon_k > 0$. Then

$$W_p^p(Q,P) \le n^{\max\{0,p/q-1\}} \sum_{k=1}^n \varepsilon_k^p. \tag{2.A.2}$$

Proof. From the definition of the Wasserstein distance and the inequality $\left(\sum_{k=1}^n a_k\right)^\gamma \le n^{\max\{0,\gamma-1\}} \sum_{k=1}^n a_k^\gamma$, which holds for all $\gamma \ge 0$ and $a_k \ge 0$, we deduce that

$$W_p^p(Q,P) \le \int_{\Xi\times\Xi} \rho(\zeta,\xi)^p \, d\pi(\zeta,\xi) \le \int_{\Xi\times\Xi} n^{\max\{0,p/q-1\}} \sum_{k=1}^n \rho_k(\zeta_k,\xi_k)^p \, d\pi(\zeta,\xi)$$

$$\leq n^{\max\{0,p/q-1\}} \sum_{k=1}^{n} \varepsilon_k^p,$$

where we used (2.A.1) in the last inequality. This establishes (2.A.2). □

Proposition 2.A.3. *Consider the product distributions $Q = Q_1 \otimes \cdots \otimes Q_n$ and $P = P_1 \otimes \cdots \otimes P_n$ on $\Xi = \Xi_1 \times \cdots \times \Xi_n$ endowed with the metric $\rho := \big(\sum_{k=1}^n \rho_k^p\big)^{1/p}$, where $p \geq 1$ and ρ_k is the metric on Ξ_k. Then*

$$W_p^p(Q,P) = \sum_{k=1}^{n} W_p^p(Q_k,P_k). \tag{2.A.3}$$

Proof. By Kantorovich duality for the transport costs $W_p^p(Q_k,P_k)$ (cf. Section 2.2), we get that

$$\begin{aligned}
\sum_{k=1}^{n} W_p^p(Q_k,P_k) &= \sum_{k=1}^{n} \sup_{\substack{(\psi_k,\phi_k)\in L^1(Q_k)\times L^1(P_k)\\ \phi_k(\xi_k)-\psi_k(\zeta_k)\leq \rho_k(\zeta_k,\xi_k)^p}} \Big\{\int_{\Xi_k} \phi_k(\xi_k)dP_k(\xi_k) - \int_{\Xi_k} \psi_k(\zeta_k)dQ_k(\zeta_k)\Big\} \\
&= \sup_{\substack{(\psi_k,\phi_k)\in L^1(Q_k)\times L^1(P_k)\\ \phi_k(\xi_k)-\psi_k(\zeta_k)\leq \rho_k(\zeta_k,\xi_k)^p,\, k\in[n]}} \sum_{k=1}^{n} \Big\{\int_{\Xi_k} \phi_k(\xi_k)dP_k(\xi_k) - \int_{\Xi_k} \psi_k(\zeta_k)dQ_k(\zeta_k)\Big\} \\
&= \sup_{\substack{(\psi,\phi)\in L^1(Q)\times L^1(P)\\ \psi=\sum_{k=1}^n \psi_k\circ \mathrm{pr}_k, \phi=\sum_{k=1}^n \phi_k\circ \mathrm{pr}_k\\ (\psi_k,\phi_k)\in L^1(Q_k)\times L^1(P_k)\\ \phi_k(\xi_k)-\psi_k(\zeta_k)\leq \rho_k(\zeta_k,\xi_k)^p,\, k\in[n]}} \Big\{\int_{\Xi} \phi(\xi)dP(\xi) - \int_{\Xi} \psi(\zeta)dQ(\zeta)\Big\} \\
&\leq \sup_{\substack{(\psi,\phi)\in L^1(Q)\times L^1(P)\\ \phi(\xi)-\psi(\zeta)\leq \sum_{k=1}^n \rho_k(\zeta_k,\xi_k)^p}} \Big\{\int_{\Xi} \phi(\xi)dP(\xi) - \int_{\Xi} \psi(\zeta)dQ(\zeta)\Big\}.
\end{aligned} \tag{2.A.4}$$

Here, the second equality follows from the fact that the constraints on ψ_k and ϕ_k, $k \in [n]$ are decoupled, and the last equality from the fact that whenever $\phi_k \in L^1(P_k)$ for all $k \in [n]$ and $\phi = \sum_{k=1}^n \phi_k \circ \mathrm{pr}_k$, then $\phi \in L^1(P)$ (analogously for ψ_k, ψ) and

$$\int_{\Xi} \phi(\xi)dP(\xi) = \int_{\Xi} \sum_{k=1}^{n} \phi_k \circ \mathrm{pr}_k(\xi)dP(\xi) = \sum_{k=1}^{n} \int_{\Xi_k} \phi_k(\xi_k)dP_k(\xi_k)$$

(and analogously for $\int_{\Xi} \psi(\zeta)dQ(\zeta)$). Since $\sum_{k=1}^n \rho_k(\zeta_k,\xi_k)^p = \rho(\zeta,\xi)^p$, we get from (2.A.4) that

$$\begin{aligned}
\sum_{k=1}^{n} W_p^p(Q_k,P_k) &\leq \sup_{\substack{(\psi,\phi)\in L^1(Q)\times L^1(P)\\ \phi(\xi)-\psi(\zeta)\leq \rho(\zeta,\xi)^p}} \Big\{\int_{\Xi} \phi(\xi)dP(\xi) - \int_{\Xi} \psi(\zeta)dQ(\zeta)\Big\} \\
&= W_p^p(Q,P).
\end{aligned}$$

Conversely, following the exact same steps as in the first part of the proof of Proposition 2.4.4 and using again the fact that $\sum_{k=1}^n \rho_k(\zeta_k,\xi_k)^p = \rho(\zeta,\xi)^p$, it follows that also $W_p^p(Q,P) \leq \sum_{k=1}^n W_p^p(Q_k,P_k)$. This establishes (2.A.3) and concludes the proof. □

Proof of Proposition 2.4.6. From the definition of the multi-transport hyperrectangle and Proposition 2.A.2, we have that

$$W_p^p(Q,P) \le n^{\max\{0,p/q-1\}} \sum_{k=1}^n \varepsilon_k^p$$

for all $P \in \mathscr{T}_p(Q,\boldsymbol{\varepsilon})$. Therefore, $\mathscr{T}_p(Q,\boldsymbol{\varepsilon}) \subset \mathscr{B}_p(Q,\varepsilon)$. To prove the second part of the statement, assume that Q is the product measure $Q_1 \otimes \cdots \otimes Q_n$ and consider probability distributions P_k, $k \in [n]$ such that $W_p^p(Q_k,P_k) = \varepsilon_k^p$. Then it follows from the construction of the Wasserstein hyperrectangle (2.4.1a) (with $\boldsymbol{P}_\xi^N \equiv Q$) that $P = P_1 \otimes \cdots \otimes P_n \in \mathscr{H}_p(Q,\boldsymbol{\varepsilon})$ and we get from Proposition 2.4.4 that also $P \in \mathscr{T}_p(Q,\boldsymbol{\varepsilon})$. Further, since $p = q$, we obtain from Proposition 2.A.3 that $W_p^p(Q,P) = \sum_{k=1}^n \varepsilon_k^p = \varepsilon$, which concludes the proof. □

2.A.2. PROOFS FROM SECTION 2.5

Proof of Proposition 2.5.2. For each component of the Wasserstein hyperrectangle, we consider the confidence level $1-\beta_k$ with β_k as given in the statement. Then we get from Corollary 2.4.5 that $\mathscr{T}_p(\boldsymbol{P}_\xi^N,\boldsymbol{\varepsilon})$ contains P_ξ with confidence

$$\prod_{k=1}^n (1-\beta_k) \ge 1 - \sum_{k=1}^n \beta_k = 1 - \sum_{k=1}^n \beta \frac{d_k}{d} = 1-\beta.$$

Denoting $r_k := d/d_k$, we get from the definition of $\widehat{C}$ that

$$\begin{aligned}
\frac{\widehat{C}(\beta,d)}{\widehat{C}(\beta_k,d_k)} &= r_k^{1/q} \frac{C(d,p) + (\ln \beta^{-1})^{1/2p}}{C(d_k,p) + (\ln \beta_k^{-1})^{1/2p}} \\
&\ge r_k^{1/q} \frac{C(d,p) + (\ln \beta^{-1})^{1/2p}}{C(d,p) + (\ln \beta^{-1})^{1/2p} + (\ln r_k)^{1/2p}} \\
&\ge r_k^{1/q} \frac{1}{1 + (\ln r_k)^{1/2p}} \ge \frac{r_k^{1/q}}{1 + (\ln r_k)^{1/2}}.
\end{aligned}$$

For these derivations, we took into account that $C(d,p)$ is increasing with respect to d in the first inequality, that $(\xi+\zeta)^a \le \xi^a + \zeta^a$ for any $x,y \ge 0$ and $a \in [0,1]$ in the second inequality, and that $C(d,p) \ge 1$ and $\ln \beta^{-1} \ge 0$ in the third inequality. Taking the derivative of the function $h(\xi) = \frac{\xi^{1/q}}{1+(\ln \xi)^{1/2}}$ for $\xi \ge 1$, we get

$$\begin{aligned}
\operatorname{sign}(\dot{h}(\xi)) &= \operatorname{sign}\Big(\frac{1}{q}\xi^{1/q-1}(1+(\ln\xi)^{1/2}) - \frac{\xi^{1/q}}{2(\ln\xi)^{1/2}\xi}\Big) \\
&= \operatorname{sign}\Big(\frac{1}{q}(\ln\xi + (\ln\xi)^{1/2}) - \frac{1}{2}\Big) \\
&= \operatorname{sign}\Big(\zeta^2 + \zeta - \frac{q}{2}\Big),
\end{aligned}$$

where $\zeta = (\ln\xi)^{1/2} \ge 0$. It can be checked that $h(\xi)$ attains its minimum when $\xi = e^{\zeta^2} = e^{(\sqrt{2q+1}-1)^2/4}$. Thus

$$\frac{\widehat{C}(\beta,d)}{\widehat{C}(\beta_k,d_k)} \ge \frac{1}{c} \tag{2.A.5}$$

with c as given in the statement.

Now pick any $P \in \mathscr{T}_p(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})$. From the definition of $\mathscr{T}_p(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})$, $\int_{\Xi\times\Xi} \|\xi_k - \zeta_k\|_q^p d\pi(\xi,\zeta) \le \varepsilon_k^p$ for each $k \in [n]$. Thus, we get from (2.A.5) and Proposition 2.A.2 with each ρ_k and ρ induced by the $\|\cdot\|_q$ norm that

$$W_p^p(\boldsymbol{P}_\xi^N, P) \le n^{\max\{0,p/q-1\}} \sum_{k=1}^n \varepsilon_k^p \le n^{\max\{0,p/q-1\}} \sum_{k=1}^n c^p \rho_\Xi^p \widehat{C}(\beta,d)^p N^{-p/\overline{d}}.$$

Hence, $\mathscr{T}_p(Q,\boldsymbol{\varepsilon}) \subset \mathscr{B}_p(Q,\varepsilon)$. Since also $\mathscr{H}_p(Q,\boldsymbol{\varepsilon}) \subset \mathscr{T}_p(Q,\boldsymbol{\varepsilon})$ by Proposition 2.4.4, this concludes the proof. □

2.B. STRONG DUALITY

In this section, we prove the strong duality result of Theorem 2.6.4 over multi-transport hyperrectangles. As in [120], we prove the result progressively, starting with the case where the uncertainty space is compact and the transport costs are continuous. This has the most significant differences compared to the approach in [120] and is presented in full detail here. The extension to general costs and noncompact spaces follows very closely the analysis in [120] and is therefore only sketched in 2.B.3. Since the duality result for compact spaces also guarantees the existence of a primal optimal transport plan, which is thereafter utilized to prove duality in the most general case, we state it as a separate result.

Proposition 2.B.1. *(Duality for compact spaces). Assume that Ξ is compact, h is upper semicontinuous, and the transport costs c_k, $k \in [n]$ satisfy Assumption 4. Then $\mathscr{I}^\star = \mathscr{J}_\star$ and there exists a primal optimizer $\pi^\star \in \Pi(Q,\boldsymbol{\varepsilon})$ with $\mathscr{I}(\pi^\star) = \mathscr{I}^\star$.*

2.B.1. COMPACT UNCERTAINTY SPACE Ξ AND CONTINUOUS COSTS $c_1,\dots,c_n$ SATISFYING ASSUMPTION 4 WITH $c_{k,m} \equiv c_k$ AND $\Xi_{\mathrm{cmp}} \equiv \Xi$

Let $X = C(\Xi\times\Xi)$ and $X^* = \mathscr{M}(\Xi\times\Xi)$ be the dual pair of Banach spaces of continuous functions and finite signed measures on $\Xi\times\Xi$, equipped with the supremum and total variation norms, respectively. Next, define

$$C := \Big\{ g \in X : g = \varphi\circ \mathrm{pr}_1 + \sum_{k=1}^n \lambda_k c_k, \text{for some } \varphi \in C(\Xi) \text{ and } \lambda_k \ge 0, k \in [n] \Big\} \tag{2.B.1a}$$

$$D := \{ g \in X : g \ge h \circ \mathrm{pr}_2 \}. \tag{2.B.1b}$$

Namely, C comprises all functions $g \in X$ that have the form $g(\zeta,\xi) = \varphi(\zeta) + \langle \boldsymbol{\lambda}, \boldsymbol{c}(\zeta,\xi)\rangle$ for all ξ, ζ, where $\varphi \in C(\Xi)$ and $\boldsymbol{\lambda} \ge 0$, and D of all $g \in X$ with $g(\zeta,\xi) \ge h(\xi)$ for all ζ,ξ. Then we have the following result.

Lemma 2.B.2. *(Properties of C and D). (i) The sets C and D are nonempty and convex.*
(ii) For each $g \in C$ there is a unique $(\boldsymbol{\lambda},\varphi) \in \mathbb{R}^n_{\ge 0} \times C(\Xi)$ such that $g = \varphi\circ\mathrm{pr}_1 + \sum_{k=1}^n \lambda_k c_k$.

Proof. To show (i), note that since h is upper semicontinuous, D is always nonempty, while convexity of the sets C and D follows directly from their definitions. To show (ii), it suffices by the definition of C to prove that if $g = \varphi \circ \mathrm{pr}_1 + \sum_{k=1}^n \lambda_k c_k = \varphi' \circ \mathrm{pr}_1 + \sum_{k=1}^n \lambda'_k c_k$ for some $\boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \mathbb{R}^n_{\geq 0}$ and $\varphi, \varphi' \in C(\Xi)$, then necessarily $\boldsymbol{\lambda} = \boldsymbol{\lambda}'$ and $\varphi = \varphi'$. Indeed, by Assumption 4(ii) with $c_{k,m} \equiv c_k$ and $\Xi_{\mathrm{cmp}} \equiv \Xi$, $\mathrm{span}\{c_1, \dots, c_n\} \cap C_{2,\mathrm{const}}(\Xi \times \Xi) = \{0\}$, which implies that $(\varphi - \varphi') \circ \mathrm{pr}_1 = 0$, and $\sum_{k=1}^n (\lambda_k - \lambda'_k) c_k = 0$. Hence, $\varphi = \varphi'$ and since $c_1, \dots, c_n$ are linearly independent by Assumption 4(ii), we get from $\sum_{k=1}^n (\lambda_k - \lambda'_k) c_k = 0$ that also $\boldsymbol{\lambda} = \boldsymbol{\lambda}'$. □

Next, define the functionals $\Phi, \Gamma : X \to \bar{\mathbb{R}}$ with

$$\Phi(g) := \begin{cases} \langle \boldsymbol{\lambda}, \boldsymbol{\epsilon} \rangle + \int_\Xi \varphi(\zeta) dQ(\zeta), & \text{if } g \in C \\ +\infty, & \text{otherwise,} \end{cases} \tag{2.B.2a}$$

$$\Gamma(g) := \begin{cases} 0, & \text{if } g \in D \\ +\infty, & \text{otherwise.} \end{cases} \tag{2.B.2b}$$

By Lemma 2.B.2, both functionals Φ and Γ are well defined, convex, and have domains C and D, respectively. To prove Proposition 2.6.4, we make use of the following lemma, which determines the conjugate functionals of Φ and Γ and their respective domains.

Lemma 2.B.3. *(Conjugates of the functionals Φ, Γ and their domains). Consider the functionals Φ, Γ defined in (2.B.2). Then their conjugate functionals Φ^*, Γ^* and their respective domains C^*, D^* are given by*

$$\Phi^*(\pi) := \begin{cases} 0, & \text{if } \pi \in C^* \\ +\infty, & \text{otherwise,} \end{cases} \tag{2.B.3a}$$

$$C^* := \{\pi \in \mathscr{M}(\Xi \times \Xi) : \langle c_k, \pi \rangle \leq \epsilon_k \text{for all } k \in [n] \text{ and } \mathrm{pr}_{1\#}\pi = Q\} \tag{2.B.3b}$$

and

$$\Gamma^*(\pi) := \begin{cases} \int_{\Xi \times \Xi} h(\xi) d\pi(\zeta, \xi), & \text{if } \pi \in D^* \\ +\infty, & \text{otherwise,} \end{cases} \tag{2.B.4a}$$

$$D^* := \{\pi \in \mathscr{M}(\Xi \times \Xi) : \langle h \circ \mathrm{pr}_2, \pi \rangle < +\infty \text{ and } \pi \preceq 0\}, \tag{2.B.4b}$$

where $\langle \cdot, \cdot \rangle$ denotes the duality between $C(\Xi \times \Xi)$ and $\mathscr{M}(\Xi \times \Xi)$ and the order $\succeq$ is considered with the respect to the positive cone in $\mathscr{M}(\Xi \times \Xi)$.

Proof. The conjugate functionals Φ^* and Γ^* are equivalently defined as

$$\Phi^*(\pi) := \sup_{g \in C} \left\{ \int_{\Xi \times \Xi} g(\zeta, \xi) d\pi(\zeta, \xi) - \Phi(g) \right\}$$

$$\Gamma^*(\pi) := \sup_{g \in D} \int_{\Xi \times \Xi} g(\zeta, \xi) d\pi(\zeta, \xi),$$

and their domains are the subsets of $\mathscr{M}(\Xi\times\Xi)$ for which their values are finite. To determine Φ^* and C^*, we get from Lemma 2.B.2(ii) that for every $\pi\in\mathscr{M}(\Xi\times\Xi)$,

$$\begin{aligned}
&\sup_{g\in C}\Big\{\int_{\Xi\times\Xi} g(\zeta,\xi)d\pi(\zeta,\xi)-\Phi(g)\Big\}\\
&=\sup_{(\boldsymbol{\lambda},\varphi)\in\mathbb{R}^n_{\geq 0}\times C(\Xi)}\Big\{\int_{\Xi\times\Xi}(\varphi(\zeta)+\langle\boldsymbol{\lambda},\boldsymbol{c}(\zeta,\xi)\rangle)d\pi(\zeta,\xi)-\Big(\langle\boldsymbol{\lambda},\boldsymbol{\epsilon}\rangle+\int_{\Xi}\varphi(\zeta)dQ(\zeta)\Big)\Big\}\\
&=\sup_{(\boldsymbol{\lambda},\varphi)\in\mathbb{R}^n_{\geq 0}\times C(\Xi)}\Big\{\sum_{k=1}^n\lambda_k\Big(\int_{\Xi\times\Xi}c_k(\zeta,\xi)d\pi(\zeta,\xi)-\epsilon_k\Big)+\int_{\Xi}\varphi(\zeta)d(\mathrm{pr}_{1\#}\pi-Q)(\zeta)\Big\},\\
&=\begin{cases}0, & \text{if}\int c_k(\zeta,\xi)d\pi(\zeta,\xi)\leq\epsilon_k \text{ for all } k\in[n] \text{ and } \mathrm{pr}_{1\#}\pi=Q\\ +\infty, & \text{otherwise,}\end{cases}
\end{aligned}$$

which establishes (2.B.3a) and (2.B.3b).

To determine Γ^* and D^*, we get by the exact same arguments as in the respective part of the proof of [120, Proposition 1] that

$$\sup_{g\in D}\int_{\Xi\times\Xi}g(\zeta,\xi)d\pi(\zeta,\xi)=\begin{cases}\int_{\Xi\times\Xi}h(\xi)d\pi(\zeta,\xi), & \text{if } \pi\in\mathscr{M}(\Xi\times\Xi) \text{ is non-positive}\\ +\infty, & \text{otherwise,}\end{cases}$$

which implies (2.B.4a) and (2.B.4b). □

We also make use of the Fenchel-Rockafellar duality theorem, which we state below. This form of the theorem is more convenient to verify in our setting compared to the more general form invoked in [120], where one needs to verify conditions about the relative interior of the involved functionals that are harder to show in our case.

Theorem 2.B.4. *(Fenchel-Rockafellar duality [154, Theorem 1.12]). Let X be a normed vector space, X^* its topological dual, and $\Phi,\Gamma: X\to\mathbb{R}\cup\{+\infty\}$ two convex functionals with domains C and D, repectively. Assume further that there is some $x_0\in C\cap D$ so that Γ is continuous at x_0. Then*

$$\inf_{x\in X}\{\Phi(x)+\Gamma(x)\}=\max_{x^*\in X^*}\{-\Phi^*(x^*)-\Gamma^*(-x^*)\},$$

where Φ^* and Γ^* are the conjugates of Φ and Γ.

We now give the proof of Proposition 2.B.1 for suitable continuous costs.

Proof of Proposition 2.B.1. (For continuous costs $c_1,\dots,c_n$ that satisfy Assumption 4(ii)). From the definition of the functionals Φ and Γ in (2.B.2) and their respective domains C and D in (2.B.1), we have that

$$\inf_{g\in X}\{\Phi(g)+\Gamma(g)\}=\inf_{g\in C\cap D}\{\Phi(g)+\Gamma(g)\}=\inf\{\mathscr{J}(\boldsymbol{\lambda},\varphi):(\boldsymbol{\lambda},\varphi)\in\Lambda \text{ and } \varphi\in C(\Xi)\}.$$

with $\mathscr{J}$ and Λ as given in (2.6.9). By Lemma 2.B.3 and (2.4.5), their conjugate functionals Φ^*, Γ^* and their domains C^*, D^* satisfy

$$-\Phi^*(\pi)-\Gamma^*(-\pi)=\int_{\Xi\times\Xi}h(\xi)d\pi(\zeta,\xi)$$

and $C^* \cap -D^* = \Pi(Q, \boldsymbol{\epsilon})$. Thus, we get from (2.6.5) that

$$\sup_{\pi \in X^*} \{-\Phi^*(\pi) - \Gamma^*(-\pi)\} = \sup_{\pi \in C^* \cap -D^*} \{-\Phi^*(\pi) - \Gamma^*(-\pi)\} = \mathscr{I}^\star.$$

Next, by exploiting upper semicontinuity of h, there exists an element $g_0 \in C \cap D$ where Γ is continuous. For example, we may take $g_0(\zeta, \xi) := \sup_{\xi \in \Xi} h(\xi) + 1$, which implies that $g \in D$ for all g in a neighborhood of g_0 in X, and thus, that $\Gamma(g) = 0$, which establishes continuity at g_0. Consequently, we deduce from Theorem 2.B.4 that

$$\inf_{g \in C \cap D} \{\Phi(g) + \Gamma(g)\} = \max_{\pi \in X^*} \{-\Phi^*(\pi) - \Gamma^*(-\pi)\},$$

where the max on the right is attained for some $\pi^* \in \mathscr{M}(\Xi \times \Xi)$. We claim that $\pi^* \in C^* \cap -D^*$. Otherwise, if $\pi^* \in \mathscr{M}(\Xi \times \Xi) \setminus (C^* \cap -D^*)$, we would have that $\mathscr{I}^\star = -\Phi^*(\pi^*) - \Gamma^*(-\pi^*) = -\infty$. But this is a contradiction because h is integrable with respect to Q and $\Pi(Q, \boldsymbol{\epsilon})$ is nonempty, since $c_k(\zeta, \zeta) \equiv 0$ for all k. We therefore get that $\inf\{\mathscr{J}(\boldsymbol{\lambda}, \varphi) : (\boldsymbol{\lambda}, \varphi) \in \Lambda \text{ and } \varphi \in C(\Xi)\} = \max_{\pi \in \Pi(Q, \boldsymbol{\epsilon})} \mathscr{I}(\pi) = \mathscr{I}^\star$ and since $C(\Xi) \subset \mathfrak{m}_{\mathscr{U}}(\Xi; \mathbb{R} \cup \{+\infty\})$, it follows from (2.6.10) that

$$\mathscr{J}_\star \leq \inf\{\mathscr{J}(\boldsymbol{\lambda}, \varphi) : (\boldsymbol{\lambda}, \varphi) \in \Lambda \text{ and } \varphi \in C(\Xi)\} = \mathscr{I}^\star.$$

Combined with (2.6.11), this concludes the proof. □

2.B.2. COMPACT UNCERTAINTY SPACE Ξ AND GENERAL COSTS $c_1, \ldots, c_n$ SATISFYING ASSUMPTION 4(II)

In this section, we clarify how the machinery of the previous section can be used to establish the duality result of Proposition 2.B.1 in the general case.

Proof of Proposition 2.B.1 (sketch). The proof consists of minor modifications of the arguments in [120, proof of Proposition 2]. Notice first that due to Assumption 4, we automatically get that for each m, $c_{k,m}$, $k \in [n]$ are linearly independent in $C(\Xi \times \Xi)$ and $\operatorname{span}\{c_{1,m}, \ldots, c_{n,m}\} \cap C_{2,\text{const}}(\Xi \times \Xi) = \{0\}$. Thus it follows from the validity of the proposition for continuous costs that there exists a sequence $\{\pi_m^\star\}$ of primal optimizers for the problems

$$\mathscr{I}_m^\star := \sup_{\pi \in \Pi(Q, \boldsymbol{\epsilon}; c_{1,m}, \ldots, c_{n,m})} \mathscr{I}(\pi) = \mathscr{I}(\pi_m^\star),$$

whose corresponding ambiguity sets are defined through the costs $c_{k,m}$, $k \in [n]$. By the duality result of the same proposition, we have that $\mathscr{I}_m^\star = \mathscr{J}_{m,\star}$, where

$$\mathscr{J}_{m,\star} := \inf_{(\boldsymbol{\lambda}, \phi) \in \Lambda(h; c_{1,m}, \ldots, c_{n,m})} \mathscr{J}(\boldsymbol{\lambda}, \phi)$$

are the optimal values of the corresponding dual problems. By tightness of the sequence $\{\pi_m^\star\}$, a subsequence $\{\pi_{m_\ell}^\star\}$ converges weakly to a probability measure $\pi^\star \in \mathscr{P}(\Xi \times \Xi)$. Then the remaining proof hinges on showing that (i)

$\pi^\star \in \Pi(Q, \boldsymbol{\epsilon}; c_1, \dots, c_n)$ and (ii) that $\mathscr{I}(\pi^\star) \geq \mathscr{J}_\star$, which by weak duality establishes that $\mathscr{I}^\star = \mathscr{J}_\star$ and that $\pi^\star$ is a primal optimizer. The establishment of (i) is based on the exact same arguments as the ones in [120, proof of Proposition 2] to verify that $\int_{\Xi\times\Xi} c_k(\zeta,\xi)d\pi^\star(\zeta,\xi) \leq \epsilon_k$ for all $k \in [n]$ and that $\mathrm{pr}_{1\#}\pi^\star = Q$. The establishment of (ii) also follows the same arguments as the ones in [120, proof of Proposition 2]. It exploits that $\Lambda(h; c_{1,m}, \dots, c_{n,m}) \subset \Lambda(h; c_1, \dots, c_n)$, which holds by Assumption 4, to get that $\mathscr{J}_{m,\star} \geq \mathscr{J}_\star$ and show that $\mathscr{I}(\pi^\star) \geq \limsup_\ell \mathscr{J}_{m_\ell,\star} \geq \mathscr{J}_\star$. □

2.B.3. DUALITY FOR NON-COMPACT SPACES AND GENERAL COSTS $c_1, \dots, c_n$ SATISFYING ASSUMPTION 4(II)

Here, we sketch how the results of the previous sections can be used to establish Theorem 2.6.4. To this end, denote for each distribution $\pi \in \mathscr{P}(\Xi \times \Xi)$

$$\Xi_\pi := \mathrm{supp}(\mathrm{pr}_{1\#}\pi) \cup \mathrm{supp}(\mathrm{pr}_{2\#}\pi)$$

and for each closed set $K \subset \Xi$

$$\Lambda(K \times K) := \Big\{(\boldsymbol{\lambda}, \varphi) : \boldsymbol{\lambda} \geq 0, \varphi \in \mathfrak{m}_{\mathscr{U}}(K; \mathbb{R} \cup \{+\infty\}), \text{ and } \varphi \circ \mathrm{pr}_1(\zeta,\xi) \geq h \circ \mathrm{pr}_2(\zeta,\xi) - \sum_{k=1}^n \lambda_k c_k(\zeta,\xi) \text{for all} \, \zeta, \xi \in K\Big\}.$$

Let also

$$\Pi_{\mathrm{fin},\boldsymbol{c},h}(Q) := \Big\{\pi \in \Pi_{\mathrm{fin},\boldsymbol{c}}(Q) : \int_{\Xi\times\Xi} h(\xi)d\pi(\zeta,\xi) \in \mathbb{R}\Big\},$$

with $\Pi_{\mathrm{fin},\boldsymbol{c}}(Q)$ as defined in Section 2.6.2. For each transport plan $\pi \in \Pi_{\mathrm{fin},\boldsymbol{c},h}(Q)$, the set $\Xi_\pi \times \Xi_\pi$, which contains the support of π, can be exhausted through a sequence of compact sets over which the integrals of the objective function h and the costs c_k are uniformly bounded. This makes it possible to use the result of the previous section and obtain bounds for the values of the dual problem over non-compact subsets of the space Ξ. In particular, we have the following auxiliary result, which will be used for the proof of the main theorem.

Proposition 2.B.5. *(Dual value bounds). Let h and the cost functions $c_1, \dots, c_n$ satisfy Assumptions 3 and 4, respectively. Then for any $\pi \in \Pi_{\mathrm{fin},\boldsymbol{c},h}(Q)$, it holds that*

$$\inf_{(\boldsymbol{\lambda},\varphi)\in\Lambda(\Xi_\pi\times\Xi_\pi)} \mathscr{J}(\boldsymbol{\lambda},\varphi) \leq \mathscr{I}^\star.$$

Proof (sketch). The proof consists again of minor modifications of the proof of [120, Proposition 3]. The first step is to pick an increasing sequence of compact subsets $\Xi_m \times \Xi_m$ of $\Xi_\pi \times \Xi_\pi$ with

$$p_m := \pi(\Xi_m \times \Xi_m) \geq 1 - \frac{1}{m}$$
$$\int_{(\Xi_m\times\Xi_m)^c} c_k(\zeta,\xi)d\pi(\zeta,\xi) \leq \frac{\epsilon_k}{m} \text{ for all } k \in [m]$$

$$\int_{(\Xi_m\times\Xi_m)^c} |h(\xi)|d\pi(\zeta,\xi) \le \frac{1}{m},$$

where $(\Xi_m\times\Xi_m)^c := \Xi_\pi\times\Xi_\pi \setminus (\Xi_m\times\Xi_m)$. For each m, we denote by π_m the normalized restriction of π to $\Xi_m\times\Xi_m$, Q_m its corresponding first marginal, and $\boldsymbol{\epsilon}^m := (\epsilon_1^m,\ldots,\epsilon_n^m)$, with $\epsilon_k^m := \epsilon_k(1-\frac{1}{m})$. From Proposition 2.B.1 applied to the restriction of the DRO problem over each space Ξ_m with $\Pi(Q_m,\boldsymbol{\epsilon}^m)$ as the associated ambiguity set, there is a zero duality gap between the values of the corresponding primal and dual problems, and there exists a primal feasible transport plan $\pi_m^\star$. Namely,

$$\int_{\Xi\times\Xi} h(\xi)d\pi_m^\star(\zeta,\xi) = \mathscr{I}_m^\star = \mathscr{J}_{m,\star}.$$

Gluing the p_m-weighted version of each optimal transport plan $\pi_m^\star$ with the restriction of π on the corresponding residual set $(\Xi_m\times\Xi_m)^c$, one can deduce by the exact same arguments as in [120, proof of Proposition 3] that

$$\mathscr{I}^\star \ge p_m\mathscr{J}_{m,\star} - \frac{1}{m}. \tag{2.B.5}$$

By selecting ε-optimal vectors $\boldsymbol{\lambda}^m = (\lambda_1^m,\ldots,\lambda_n^m)$ of dual parameters for each dual optimal value $\mathscr{J}_{m,\star}$, it follows in analogy to [120, proof of Proposition 3] that

$$\langle\boldsymbol{\lambda}^m,\boldsymbol{\epsilon}^m\rangle + \int_{\Xi_m}\sup_{\xi\in\Xi_m}\Big\{h(\xi)-\sum_{k=1}^n\lambda_k^m c_k(\zeta,\xi)\Big\}dQ_m(\zeta) \le \mathscr{J}_{m,\star}+\varepsilon,$$

which together with (2.B.5) implies that

$$\begin{aligned}&\limsup_{m\to\infty}\Big\{p_m\langle\boldsymbol{\lambda}^m,\boldsymbol{\epsilon}^m\rangle + \int_{\Xi_\pi\times\Xi_\pi}\sup_{\xi\in\Xi_m}\Big\{h(\xi)-\sum_{k=1}^n\lambda_k^m c_k(\zeta,\xi)\Big\}\mathbb{1}_{\Xi_m\times\Xi_m}(\zeta,\xi')d\pi(\zeta,\xi')\Big\}\\ &\le \mathscr{I}^\star+\varepsilon.\end{aligned}$$

One can then show as in [120, proof of Proposition 3] that the sequences $\{\lambda_k^m\}_{m\in\mathbb{N}}$, $k\in[n]$ are bounded. Thus, there exists a subsequence $\{\boldsymbol{\lambda}^{m_\ell}\}_{\ell\in\mathbb{N}}$ converging to some $\boldsymbol{\lambda}^\star\ge 0$ and it can be checked along the lines of [120, proof of Lemma B.7] that

$$\liminf_{\ell\to\infty}\sup_{\xi\in\Xi_{m_\ell}}\Big\{h(\xi)-\sum_{k=1}^n\lambda_k^{m_\ell}c_k(\zeta,\xi)\Big\} \ge \sup_{\xi\in\cup_{\ell=1}^\infty\Xi_{m_\ell}}\Big\{h(\xi)-\sum_{k=1}^n\lambda_k^{m_\ell}c_k(\zeta,\xi)\Big\}$$

for all $\zeta\in\Xi_\pi$. Using the same arguments as in [120, proof of Proposition 3], this implies that

$$\begin{aligned}\mathscr{I}^\star+\varepsilon &\ge \liminf_{m\to\infty}\Big\{p_m\langle\boldsymbol{\lambda}^m,\boldsymbol{\epsilon}^m\rangle + \int_{\Xi_\pi\times\Xi_\pi}\sup_{\xi\in\Xi_m}\Big\{h(\xi)-\sum_{k=1}^n\lambda_k^m c_k(\zeta,\xi)\Big\}\mathbb{1}_{\Xi_m\times\Xi_m}(\zeta,\xi')d\pi(\zeta,\xi')\Big\}\\ &\ge \langle\boldsymbol{\lambda}^\star,\boldsymbol{\epsilon}\rangle + \int_{\Xi_\pi\times\Xi_\pi}\sup_{\xi\in\cup_{\ell=1}^\infty\Xi_{m_\ell}}\Big\{h(\xi)-\sum_{k=1}^n\lambda_k^\star c_k(\zeta,\xi)\Big\}d\pi(\zeta,\xi')\end{aligned}$$

$$= \langle \boldsymbol{\lambda}^\star, \boldsymbol{\epsilon} \rangle + \int_{\Xi_\pi} \sup_{\xi \in \Xi_\pi} \Big\{ h(\xi) - \sum_{k=1}^{n} \lambda_k^\star c_k(\zeta, \xi) \Big\} dQ(\zeta).$$

Since ε is arbitrary, selecting the pair $(\boldsymbol{\lambda}^\star, \varphi) \in \Lambda(\Xi_\pi \times \Xi_\pi)$ with $\varphi(\zeta) := \sup_{\xi \in \Xi_\pi} \{h(\xi) - \sum_{k=1}^{n} \lambda_k^\star c_k(\zeta, \xi)\}$ establishes the result. □

We need one last result whose proof we omit as it is identical to that of [120, Proposition 4].

Proposition 2.B.6. *(Integration/majorization interchange). If the objective function h and the cost functions $c_1, \ldots, c_n$ satisfy Assumptions 3 and 4, respectively, then*

$$\sup_{\pi \in \Pi_{\text{fin},\boldsymbol{c},h}(Q)} \int_{\Xi \times \Xi} (h(\xi) - \langle \boldsymbol{\lambda}, \boldsymbol{c}(\zeta, \xi) \rangle) d\pi(\zeta, \xi) = \int_{\Xi} \sup_{\xi \in \Xi} \{h(\xi) - \langle \boldsymbol{\lambda}, \boldsymbol{c}(\zeta, \xi) \rangle\} dQ(\zeta).$$

Now, we can proceed to sketch the proof of strong duality for general Polish spaces.

Proof of Theorem 2.6.4 (sketch). The proof relies on showing that $\mathscr{I}^\star \geq \mathscr{J}_\star$ and follows the steps of [120, proof of Theroem 1]. When $\mathscr{I}^\star = +\infty$, then the result follows from the fact that $\mathscr{I}^\star \leq \mathscr{J}_\star$. When $\mathscr{I}^\star < +\infty$, Proposition 2.B.5 implies that for each $\pi \in \Pi_{\text{fin},\boldsymbol{c},h}(Q)$

$$\begin{aligned} \mathscr{I}^\star &\geq \inf_{(\boldsymbol{\lambda}, \varphi) \in \Lambda(\Xi_\pi \times \Xi_\pi)} \Big\{ \langle \boldsymbol{\lambda}, \boldsymbol{\epsilon} \rangle + \int_{\Xi_\pi} \varphi(\zeta) dQ(\zeta) \Big\} \\ &\geq \inf_{\boldsymbol{\lambda} \geq 0} \Big\{ \langle \boldsymbol{\lambda}, \boldsymbol{\epsilon} \rangle + \int_{\Xi} \sup_{\xi \in \Xi_\pi} \{h(\xi) - \langle \boldsymbol{\lambda}, \boldsymbol{c}(\zeta, \xi) \rangle\} dQ(\zeta) \Big\}. \end{aligned}$$

Next, denote

$$T(\boldsymbol{\lambda}, \pi) := \langle \boldsymbol{\lambda}, \boldsymbol{\epsilon} \rangle + \int_{\Xi} \sup_{\xi \in \Xi_\pi} \{h(\xi) - \langle \boldsymbol{\lambda}, \boldsymbol{c}(\zeta, \xi) \rangle\} dQ(\zeta)$$

and $\lambda_{\max} := \max_{k=1,\ldots,n} \frac{\mathscr{I}^\star - \int_\Xi h(\zeta) dQ(\zeta)}{\epsilon_k}$. In the definition of the function $T(\boldsymbol{\lambda}, \pi)$, the transport plan π determines the set Ξ_π over which the supremum of $h(\xi) - \langle \boldsymbol{\lambda}, \boldsymbol{c}(\zeta, \xi) \rangle$ is evaluated. Then it follows by the same arguments as in [120, proof of Theroem 1(a)] that

$$\mathscr{I}^\star \geq \inf_{\boldsymbol{\lambda} \in [0, \lambda_{\max}]^n} T(\boldsymbol{\lambda}, \pi) \tag{2.B.6}$$

and that $T(\boldsymbol{\lambda}, \pi)$ is lower semicontinuous and convex with respect to $\boldsymbol{\lambda}$ and concave with respect to π. Thus, it follows from Fan's minimax theorem [155, Theorem 2] that

$$\sup_{\pi \in \Pi_{\text{fin},\boldsymbol{c},h}(Q)} \inf_{\boldsymbol{\lambda} \in [0, \lambda_{\max}]^n} T(\boldsymbol{\lambda}, \pi) = \inf_{\boldsymbol{\lambda} \in [0, \lambda_{\max}]^n} \sup_{\pi \in \Pi_{\text{fin},\boldsymbol{c},h}(Q)} T(\boldsymbol{\lambda}, \pi),$$

and we get from (2.B.6) that

$$\mathscr{I}^\star \geq \inf_{\boldsymbol{\lambda} \in [0, \lambda_{\max}]^n} \Big\{ \langle \boldsymbol{\lambda}, \boldsymbol{\epsilon} \rangle + \sup_{\pi \in \Pi_{\text{fin},\boldsymbol{c},h}(Q)} \int_{\Xi} \sup_{\xi \in \Xi_\pi} \{h(\xi) - \langle \boldsymbol{\lambda}, \boldsymbol{c}(\zeta, \xi) \rangle\} dQ(\zeta) \Big\}$$

$$\geq \inf_{\boldsymbol{\lambda}\in[0,\lambda_{\max}]^n}\Big\{\langle\boldsymbol{\lambda},\boldsymbol{\varepsilon}\rangle+\sup_{\pi\in\Pi_{\mathrm{fin},\boldsymbol{c},h}(Q)}\int_{\Xi}(h(\xi)-\langle\boldsymbol{\lambda},\boldsymbol{c}(\zeta,\xi)\rangle)dQ(\zeta)\Big\}$$
$$= \inf_{\boldsymbol{\lambda}\in[0,\lambda_{\max}]^n}\Big\{\langle\boldsymbol{\lambda},\boldsymbol{\varepsilon}\rangle+\int_{\Xi}\sup_{\xi\in\Xi}\{h(\xi)-\langle\boldsymbol{\lambda},\boldsymbol{c}(\zeta,\xi)\rangle\}dQ(\zeta)\Big\},$$

where the last equality holds due to Proposition 2.B.6. Since $(\boldsymbol{\lambda},\varphi_{\boldsymbol{\lambda}})\in\Lambda$ for all $\boldsymbol{\lambda}\in[0,\lambda_{\max}]^n$, where $\varphi_{\boldsymbol{\lambda}}(\zeta):=\sup_{\xi\in\Xi}\{h(\xi)-\langle\boldsymbol{\lambda},\boldsymbol{c}(\zeta,\xi)\rangle\}$, it follows from (2.6.10) that strong duality holds.

To show that a dual optimizer of the form $(\boldsymbol{\lambda},\varphi_{\boldsymbol{\lambda}})$ exists, let $g(\boldsymbol{\lambda}):=\langle\boldsymbol{\lambda},\boldsymbol{\varepsilon}\rangle+\int_{\Xi}\varphi_{\boldsymbol{\lambda}}(\zeta)dQ(\zeta)$. Then as in [120, proof of Theroem 1(b)], it follows that g is lower semicontinous and that $g(\boldsymbol{\lambda})\geq\langle\boldsymbol{\lambda},\boldsymbol{\varepsilon}\rangle+\int_{\Xi}h(\zeta)dQ(\zeta)$, which implies that g is radially unbounded since $\lim_{\|\boldsymbol{\lambda}\|\to+\infty}g(\boldsymbol{\lambda})=+\infty$ for $\boldsymbol{\lambda}\geq 0$. Hence, the level sets of g are compact and its infimum is always attained. Since $\mathscr{J}(\boldsymbol{\lambda},\varphi)\geq\mathscr{J}(\boldsymbol{\lambda},\varphi_{\boldsymbol{\lambda}})$ for all $(\boldsymbol{\lambda},\varphi)\in\Lambda$, the infimum of the dual problem is also attained by a pair $(\boldsymbol{\lambda}^\star,\varphi_{\boldsymbol{\lambda}^\star})\in\Lambda$ and the proof is complete. □

2.C. DRO REFORMULATIONS OF THE SIMULATION EXAMPLE

Here, we derive the dual value reformulations of the DRO problem (2.7.2). From Theorem 2.6.4, we get the equivalent dual reformulation

$$\inf_{\boldsymbol{\lambda}\geq 0}\langle\boldsymbol{\lambda},\boldsymbol{\varepsilon}\rangle+\frac{1}{N^2}\sum_{l\in[N^2]}\sup_{\xi\in\Xi}\Big\{\mathbb{1}_\Theta(\xi)\,|D_t+\xi_1-\xi_2-x_t|-\sum_{k=1}^{2}\lambda_k|\xi_k-\xi_k^l|\Big\}$$
$$=\inf_{\boldsymbol{\lambda}\geq 0}\langle\boldsymbol{\lambda},\boldsymbol{\varepsilon}\rangle+\frac{1}{N^2}\sum_{l\in[N^2]}\max\Big\{0,\sup_{\xi\in\Theta}\Big\{|D_t+\xi_1-\xi_2-x_t|-\sum_{k=1}^{2}\lambda_k|\xi_k-\xi_k^l|\Big\}\Big\}.$$

Introducing epigraphical variables $s\geq 0$, we can rewrite the dual problem as

$$\begin{aligned}&\inf_{\boldsymbol{\lambda}\geq 0}\langle\boldsymbol{\lambda},\boldsymbol{\varepsilon}\rangle+\frac{1}{N^2}\sum_{l\in[N^2]}s_l\\ &\text{s.t. }\sup_{\xi\in\Theta}|D_t+\xi_1-\xi_2-x_t|-\sum_{k=1}^{2}\lambda_k|\xi_k-\xi_k^l|\leq s_l,\qquad\forall l\in[N^2].\end{aligned}\tag{2.C.1}$$

Taking into account that

$$|D_t+\xi_1-\xi_2-x_t|=\max\{D_t+\xi_1-\xi_2-x_t,-(D_t+\xi_1-\xi_2-x_t)\}$$

and denoting each of the affine functions in this max by $h_j(x,\xi)$, $j\in\{1,2\}$ we rewrite the constraints in (2.C.1) as

$$\begin{aligned}&\sup_{\xi\in\Theta}\{h_j(x,\xi)-\sum_{k=1}^{2}\lambda_k|\xi-\xi_k^l|\}\ \sup_{\xi\in\Theta}\{h_j(x,\xi)-\max_{|\mathrm{pr}_k z_j^l|\leq\lambda_k}\langle z_j^l,\xi^l-\xi\rangle\}\\ &=\min_{|\mathrm{pr}_k z_j^l|\leq\lambda_k}\sup_{\xi\in\Theta}\{h_j(x,\xi)-\langle z_j^l,\xi^l-\xi\rangle\}\leq s_l,\qquad\forall l\in[N^2],\ \forall j\in[2],\end{aligned}$$

where the supmin interchange follows from the minimax theorem [156]. These constraints are equivalently written as

$$\begin{aligned}\sup_{\xi\in\Theta}\{D_t+\xi_1-\xi_2-x-\langle z_1^l,\xi^l-\xi\rangle\}&\le s_l\\ \sup_{\xi\in\Theta}\{-D_t-\xi_1+\xi_2+x-\langle z_2^l,\xi^l-\xi\rangle\}&\le s_l\\ \left|\mathrm{pr}_k z_j^l\right|&\le\lambda_k,\ j\in[2],\ k\in[2].\end{aligned}$$

Equivalently we have

$$\begin{aligned}D_t-x+\sup_{\xi\in\Theta}\{+\xi_1-\xi_2+\langle z_1^l,\xi\rangle\}-\langle z_1^l,\xi^l\rangle&\le s_l\\ -D_t+x+\sup_{\xi\in\Theta}\{-\xi_1+\xi_2+\langle z_2^l,\xi\rangle\}-\langle z_2^l,\xi^l\rangle&\le s_l\\ \left|\mathrm{pr}_k z_j^l\right|&\le\lambda_k,\ j\in[2],\ k\in[2].\end{aligned}$$

and using properties of the Legendre transform to rewrite the terms with the supremum, we get

$$\begin{aligned}D_t-x+\inf_{C^\top\gamma_1^l=v_1^l,\,z_1^l=v_1^l-[1\ -1]^\top}\langle\gamma_1^l,d\rangle-\langle z_1^l,\xi^l\rangle&\le s_l\\ -D_t+x+\inf_{C^\top\gamma_2^l=v_2^l,\,z_2^l=v_2^l-[-1\ 1]^\top}\langle\gamma_2^l,d\rangle-\langle z_2^l,\xi^l\rangle&\le s_l\end{aligned}$$

where we used the compact notation $\Theta\equiv\{C\xi\le d\}$. As a result, we get

$$\begin{aligned}\langle\gamma_1^l,d-C\xi^l\rangle+D_t-x+\langle[1\ -1]^\top,\xi^l\rangle&\le s_l\\ \langle\gamma_2^l,d-C\xi^l\rangle-D_t+x+\langle[-1\quad 1]^\top,\xi^l\rangle&\le s_l\end{aligned}$$

and we can rewrite the DRO problem as the linear program

$$\begin{aligned}\inf_{\boldsymbol{\lambda}\ge 0,\boldsymbol{s},\boldsymbol{\gamma},x}\ &\langle\boldsymbol{\lambda},\boldsymbol{\varepsilon}\rangle+\frac{1}{N^2}\sum_{l\in[N^2]}s_l\\ \text{s.t. }&\langle\gamma_1^l,d-C\xi^l\rangle+D_t-x+\langle[1\ -1]^\top,\xi^l\rangle\le s_l\\ &\langle\gamma_2^l,d-C\xi^l\rangle-D_t+x+\langle[-1\quad 1]^\top,\xi^l\rangle\le s_l\\ &\left|\mathrm{pr}_k(C^\top\gamma_1^l-[1\ -1]^\top)\right|\le\lambda_k\\ &\left|\mathrm{pr}_k(C^\top\gamma_2^l-[-1\quad 1]^\top)\right|\le\lambda_k\\ &\gamma_j^l\ge 0k\in[2],\ j\in[2],\ l\in[N^2].\end{aligned}$$

3

TRACTABLE REFORMULATIONS OF DRO PROBLEMS OVER STRUCTURED OPTIMAL TRANSPORT AMBIGUITY SETS

Structuring ambiguity sets in Wasserstein-based distributionally robust optimization (DRO) can improve their statistical properties when the uncertainty consists of multiple independent components. The aim of this chapter is to solve stochastic optimization problems with unknown uncertainty when we only have access to a finite set of samples from it. Exploiting strong duality of DRO problems over structured ambiguity sets, we derive tractable reformulations for certain classes of DRO and uncertainty quantification problems. We also derive tractable reformulations for distributionally robust chance-constrained problems. As the complexity of the reformulations may grow exponentially with the number of independent uncertainty components, we employ clustering strategies to obtain informative estimators, which yield problems of manageable complexity. We demonstrate the effectiveness of the theoretical results in a numerical simulation example.

This chapter is mainly based on L. M. Chaouach, T. Oomen, and D. Boskos. *Tractable reformulations of DRO problems over structured optimal transport ambiguity sets*. Accepted for publication in *Transactions on Automatic Control*. 2025. arXiv: 2504.06966 [math.OC], while the section annotated with ⋆ is based on L. M. Chaouach, T. Oomen, and D. Boskos. "Comparing Structured Ambiguity Sets for Stochastic Optimization: Application to Uncertainty Quantification". In: *IEEE Int. Conf. on Decision and Control*. 2023, pp. 8274–8279.

3.1. INTRODUCTION

DECISIONS under uncertainty are widespread in engineering, and the uncertainty often has specific sources and structure. A typical way to characterize the uncertainty is through probabilistic models. In addition to the range of an uncertain outcome, these models also capture its expected frequency in each region of its range. Stochastic optimization revolves around making decisions in the face of probabilistic uncertainty [99]. This yields an effective paradigm for system design and operation when the probabilistic model accurately reflects the nature of the uncertainty. However, such probabilistic models are often not available in practice and need to be inferred from a limited amount of data. This may in turn lead to modeling imperfections that can negatively impact design requirements such as efficiency, safety, and reliability. It is therefore essential to seek guarantees for the performance of stochastic optimization problems under uncertainty about their underlying probabilistic model.

For optimization problems with uncertain constraints that need to be met with prescribed probability, also referred to as chance-constrained problems, the scenario approach yields decisions that respect the constraints with distribution-free guarantees [67]. The nominal form of this data-driven method robustifies the optimization problem against the worst-case scenario among i.i.d. realizations of the uncertainty. This way, the scenario approach determines tight feasibility regions for chance-constrained problems [69], which can even be exact [70]. Multiple domains have successfully employed the scenario approach, including robust control [157], model predictive control [158], and interval prediction [159]. However, its guarantees regarding the satisfaction of probabilistically tight constraints with high confidence, rely on large amounts of samples, which are not always available. In addition, the realizations of the uncertainty may be corrupted by noise, and then it is no longer clear how to retain the probabilistic guarantees of the scenario theory. This motivates the consideration of distributionally robust approaches for data-driven optimization problems. Despite not being fully distribution free, such approaches can provide probabilistic guarantees for any number of samples for general classes of unknown probability distributions.

Distributionally robust optimization (DRO) aims at hedging against distributional ambiguity, which is typical in real-life stochastic uncertainty. To this end, it considers a range of probabilistic models that could likely characterize the uncertainty and determines an optimal solution that is robust against all the distributions in that range. This paradigm has proven to be useful in several applications that require effective handling of data or parameter uncertainty, including machine learning [85], portfolio optimization [160], power dispatch [161], and scheduling [162]. DRO has also been employed to solve problems in stochastic control, such as distributionally robust linear quadratic regulator problems [92, 105], model predictive control algorithms [103, 130], and more general formulations in distributionally robust dynamic programming [163].

A widely used approach is to group plausible probability distributions inside a ball in the Wasserstein metric, which is usually centered at the empirical distribution of the data. This choice enables the designer to tune the radius of the

ambiguity ball so that it contains the true probability distribution with prescribed probability [59] and yields DRO problems that admit tractable reformulations [79, 119, 120]. Decisions of DRO problems over Wasserstein ambiguity balls are also asymptotically consistent [59]. Namely, as the number of samples goes to infinity, the ambiguity radius can be tuned so that the optimal value and decision of the DRO problem converge to the corresponding quantities of the ideal stochastic optimization problem. In the finite sample regime, Wasserstein ambiguity balls that maintain probabilistic guarantees of containing the true distribution suffer from the curse of dimensionality. In particular, their contraction rate with respect to the number of samples scales poorly with the dimension of the uncertainty [59, 95, 164], as seen in Chapter 2.

A recent line of research aims at rendering the decay rate of the ambiguity ball radius independent of the uncertainty dimension by informing it from the associated optimization problem [85, 123–126], thereby suppressing the curse of dimensionality. Despite these results, the curse of dimensionality still persists when solving multiple distributionally robust optimization problems with the same underlying uncertainty. Some instances of this situation include model predictive control (MPC) [106, 130, 131], strategy synthesis for Markov decision process [132], and in general dynamic programming problems [92]. In these cases, the reliability of the DRO solution can be assured if the associated ambiguity set contains the data generating distribution with high confidence. Yet, achieving this with traditional Wasserstein ambiguity balls comes with the compromise of their slow shrinkage rate with the number of samples for high-dimensional uncertainty.

For DRO problems where the ambiguity set contains the true distribution with high confidence, we introduced in the previous chapter a new class of ambiguity sets that can ameliorate the curse of dimensionality. This is achievable when the uncertainty consists of several independent components. The corresponding ambiguity sets, termed multi-transport hyperrectangles (MTHs), are constructed by imposing multiple optimal transport constraints, one for each component of the uncertainty. These hyperrectangles shrink at much faster rates with the number of samples compared to their monolithic counterparts while maintaining the same probabilistic guarantees of containing the data-generating distribution. This in turn, can facilitate reducing the conservativeness of their associated DRO problems compared to Wasserstein balls. DRO problems over MTHs are also accompanied by finite-dimensional duality results, see Theorem 2.6.4 of Chapter 2. However, there are yet no systematic reformulations of their dual problems that can be solved by tractable algorithms.

Our goal in this chapter is to go beyond the duality results in Chapter 2, both in terms of the key theoretical properties shared by MTHs and their practical exploitation for deriving tractable reformulations of several classes of DRO problems. To this end, our first contribution is the establishment of fundamental properties shared by MTHs, including weak compactness and conditions under which their associated DRO cost is finite. These results have standalone theoretical value and are therefore presented in the general setting of probability distributions on Polish spaces. Our second set of contributions exploits DRO duality over hyperrectangles

to derive tractable reformulations of distributionally robust average-cost problems, including those with piecewise affine and quadratic costs. We further leverage the results for piecewise affine costs to solve distributionally robust uncertainty quantification problems over general polyhedral sets. These results generalize existing work in the literature, both in terms of the ambiguity sets considered and the sets over which uncertainty is quantified, which we allow to be nonconvex.

Our third contribution is the tractable reformulation of distributionally robust chance-constrained problems over MTHs. To obtain this reformulation, we generalize a stochastic minimax theorem, which allows us to handle constraints whose uncertainty-dependent arguments may become unbounded. This result is presented in the greatest possible generality and constitutes one of our main theoretical contributions. Our last contribution revolves around how to cluster the reference distribution (center) of the MTHs to reduce the complexity of their DRO reformulations when the uncertainty consists of many independent components. We also elaborate on tradeoffs between the complexity reduction and the potential inflation of the hyperrectangles to retain their probabilistic guarantees.

The chapter is organized as follows. Section 3.2 introduces necessary preliminaries and notation. Section 3.3 formulates the problem and Section 3.4 presents basic properties of MTHs. In Section 3.5, we provide tractable reformulations for DRO problems over multi-transport ambiguity hyperrectangles. In Section 3.6, we specialize these reformulations for distributionally robust uncertainty quantification problems. Section 3.7 introduces corresponding reformulations for distributionally robust chance-constrained problems. In Section 3.8, we address the computational complexity of the reformulations. We illustrate the results in a simulation example in Section 3.9.

3.2. NOTATION AND MATHEMATICAL PRELIMINARIES

General notation: Throughout this chapter, we use the following notation. We denote by $\|\cdot\|$ an arbitrary norm on $\mathbb{R}^d$ and by $\|\cdot\|_p$ the pth norm for $p \in [1,\infty]$. The dual of a norm $\|\cdot\|$ is defined through the inner product in $\mathbb{R}^d$ and is denoted by $\|\cdot\|_*$. In particular, $\|z\|_* := \sup_{\|\xi\|\leq 1}\langle z,\xi\rangle$. We denote by $\mathbb{N}_{>0}$ the positive integers, by $\mathbb{R}_{\geq 0}$ and $\mathbb{R}_{>0}$ the positive and strictly positive real numbers, respectively, and define $\bar{\mathbb{R}} := \mathbb{R}\cup\{-\infty,+\infty\}$. For $N \in \mathbb{N}_{>0}$, we denote $[N] := \{1,\ldots,N\}$. We denote by $\mathbb{S}_n$ the set of $n\times n$ real-valued symmetric matrices. We use the notation $x_+ := \max\{x,0\}$ for a real number x. Given a nonempty product set $\Xi = \Xi_1 \times \ldots \times \Xi_n$ and distinct $k,l \in [n]$, we define the projection operators $\mathrm{pr}_k : \Xi \to \Xi_k$ and $\mathrm{pr}_{kl} : \Xi \to \Xi_k \times \Xi_l$ as $\mathrm{pr}_k(\xi) := \xi_k$ and $\mathrm{pr}_{kl}(\xi) := (\xi_k,\xi_l)$, respectively, for all $\xi = (\xi_1,\ldots,\xi_n) \in \Xi$. Let Ξ be formed by the product of d identical copies of some nonempty set Θ, i.e., $\Xi = \Theta^d$ for some $d \in \mathbb{N}$ with $d = d_1 + \cdots + d_n$, and denote $\boldsymbol{d} := (d_1,\ldots,d_n) \in \mathbb{N}^n$. We define $\mathrm{pr}_k^{\boldsymbol{d}} : \Xi \to \Xi_k := \Theta^{d_k}$ by $\mathrm{pr}_k^{\boldsymbol{d}}(\theta_1,\ldots,\theta_d) := (\theta_{\ell+1},\ldots,\theta_{\ell+d_k})$ with $\ell := d_1 + \cdots + d_{k-1}$ for $k > 1$ and $\ell := 0$ for $k = 1$. The conjugate of the function $h : \mathbb{R}^d \to \bar{\mathbb{R}}$ is defined by $h^*(z) := \sup_{\xi\in\mathbb{R}^d}\{\langle z,\xi\rangle - h(\xi)\}$. Given a set $\Theta \subset \mathbb{R}^d$, its characteristic function is defined by $\chi_\Theta(\xi) := 0$ if $\xi \in \Theta$ and $\chi_\Theta(\xi) := +\infty$, otherwise, and its support function by $\sigma_\Theta(\xi) := \sup_{\xi\in\Theta}\langle z,\xi\rangle$. The support function of a set is equal to the conjugate of

its characteristic function. Given $x, y \in \mathbb{R}^d$, we denote $x \succeq y$ if all components of $x - y$ are positive. Vectors will be interpreted as column vectors in linear algebra operations unless indicated by a transpose.

Probability theory: Let (Ξ, ρ) be a Polish space, i.e., a complete and separable space Ξ with metric ρ. We denote by $\mathscr{P}(\Xi)$ the space of probability distributions on Ξ with its Borel σ-algebra. For any $p \geq 1$, $\mathscr{P}_p(\Xi)$ denotes the set of distributions in $\mathscr{P}(\Xi)$ with finite pth moment. The class $\mathscr{G}_r^{\text{up}}$ consists of all real-valued functions h on Ξ that satisfy the growth bound

$$h(\xi) \leq C(1 + \rho(\zeta, \xi)^r) \text{ for all } \xi \in \Xi, \tag{3.2.1}$$

for some $C > 0$ and $\zeta \in \Xi$. We also denote by $\mathscr{G}_r$ the class of functions h with both $-h, h \in \mathscr{G}_r^{\text{up}}$. We refer to the set of discrete distributions supported on $K \in \mathbb{N}$ atoms in Ξ as $\mathscr{P}^K(\Xi)$. Given the measurable spaces $(\Omega, \mathscr{F})$ and $(\Omega', \mathscr{F}')$, a measurable map $\Psi : (\Omega, \mathscr{F}) \to (\Omega', \mathscr{F}')$ assigns to each measure μ in $(\Omega, \mathscr{F})$ the pushforward measure $\Psi_{\#}\mu$ in $(\Omega', \mathscr{F}')$ defined by $\Psi_{\#}\mu(B) := \mu(\Psi^{-1}(B))$ for all $B \in \mathscr{F}'$. We denote by $P \otimes Q$ the product measure of P and Q. The Dirac distribution centered at $\xi \in \Xi$ is denoted by δ_ξ. The indicator function $\mathbb{1}_\Theta$ of a set $\Theta \subset \Xi$ is $\mathbb{1}_\Theta(\xi) := 1$ if $\xi \in \Theta$ and 0 otherwise. For any $p \geq 1$, the pth Wasserstein distance between two probability distributions $P, Q \in \mathscr{P}_p(\Xi)$ is defined as

$$W_p(Q, P) := \left(\inf_{\pi \in \mathscr{C}(Q,P)} \left\{ \int_{\Xi \times \Xi} \rho(\zeta, \xi)^p \mathrm{d}\pi(\zeta, \xi) \right\} \right)^{1/p}$$

Each $\pi \in \mathscr{C}(Q, P)$ is a transport plan, a.k.a. coupling between P and Q, i.e., a distribution on $\Xi \times \Xi$ with marginals $P = \mathrm{pr}_{2\#}(\pi)$ and $Q = \mathrm{pr}_{1\#}(\pi)$, respectively.

3.3. MOTIVATION AND PROBLEM FORMULATION

Stochastic optimization is decision-making in the presence of probabilistic uncertainty [99]. It includes optimization problems of the form

$$\inf_{x \in \mathscr{X}} \mathbb{E}_{P_\xi}[f(x, \xi)], \tag{3.3.1}$$

where f is the objective function, $x \in \mathscr{X}$ is the decision variable, and $\xi \in \Xi$ is a random variable with distribution P_ξ. This problem seeks decisions that are optimal on average.

Another class of stochastic decision-making problems involves constraints that are not deterministic, but are required to be satisfied with a prescribed probability. These are chance-constrained problems of the form

$$\begin{aligned} &\inf_{x \in \mathscr{X}} g(x) \\ &\text{s.t. } P_\xi(F(x, \xi) \leq 0) \geq 1 - \alpha, \end{aligned} \tag{3.3.2}$$

where $1 - \alpha$ designates a probabilistic threshold with which we require the constraint to hold.

A challenge for both problems is that the distribution P_ξ is often unknown and there is only access to a finite number of i.i.d. samples $\xi^1,\ldots,\xi^N$ of the uncertainty ξ. Using these samples, it is possible to approximate P_ξ by a data-driven estimator $\widehat{P}_\xi$, such as the empirical distribution $\widehat{P}_\xi \equiv P_\xi^N := \frac{1}{N}\sum_{i=1}^N \delta_{\xi^i}$ of the samples. Choosing the empirical distribution as a proxy for P_ξ yields the so-called Sample Average Approximation (SAA) of (3.3.1), which provides reliable decisions for large amounts of samples. However, when the number of samples is limited, the SAA may become a poor surrogate of the original optimization problem since the empirical distribution may deviate significantly from the true distribution. To hedge against this uncertainty about the distribution, we consider the distributionally robust formulation

$$\inf_{x\in\mathscr{X}} \sup_{P\in\mathscr{P}^N} \mathbb{E}_P\big[f(x,\xi)\big] \tag{3.3.3}$$

of the average-cost problem (3.3.1). In this distributionally robust optimization (DRO) problem, $\mathscr{P}^N$ is an ambiguity set of distributions that contains plausible models for the true distribution and can be inferred from the collected samples.

In an analogous way, taking into account that

$$P_\xi(F(x,\xi)\le 0)\ge 1-\alpha \iff P_\xi(F(x,\xi)>0)\le\alpha,$$

the chance-constrained problem (3.3.2) is robustified against distributional uncertainty through the formulation

$$\begin{array}{ll}\inf\limits_{x\in\mathscr{X}} & g(x)\\ \text{s.t.} & \sup\limits_{P\in\mathscr{P}^N} P_\xi(F(x,\xi)>0) \equiv \sup\limits_{P\in\mathscr{P}^N} \mathbb{E}_P[\mathbb{1}_{\{\xi\in\Xi:\,F(x,\xi)>0\}}(\xi)] \le \alpha,\end{array} \tag{3.3.4}$$

which hedges against uncertainty over the true distribution P_ξ.

A popular approach to construct data-driven ambiguity sets is to group all distributions that are ε-close to the empirical distribution P_ξ^N in the Wasserstein metric. This yields the Wasserstein ambiguity ball

$$\mathscr{B}_p(P_\xi^N,\varepsilon) := \{P\in\mathscr{P}_p(\Xi) : W_p(P_\xi^N,P)\le\varepsilon\},$$

with center P_ξ^N and radius ε. Higher values of the exponent $p\ge 1$ yield larger weights to distribution dissimilarities that are farther apart. This is aligned with the fact that Wasserstein distances penalize horizontal variations of the reference distribution and can effectively capture their impact on the value of the optimization problem. Wasserstein ambiguity balls also yield DRO problems with tractable reformulations [79, 119, 120] and enjoy finite-sample guarantees of containing the data-generating distribution [59]. As a result, the optimal cost of (3.3.3) provides an upper bound for the optimal cost of (3.3.1) with prescribed confidence [79, Theorem 3.5]. The same conclusion also holds for the robust version (3.3.4) of the chance-constrained problem (3.3.2). In particular, if $\mathscr{P}^N$ contains the true distribution with high confidence, then the feasible set of x in (3.3.4) is contained in the feasible set of (3.3.2) with the same confidence. Since the value of a minimization problem over a feasible set is less than or equal to its value over any (not necessarily strict) subset of that set, the optimal cost of (3.3.4) provides an upper bound on the cost of (3.3.2) with high confidence.

3.3.1. STRUCTURED OPTIMAL TRANSPORT AMBIGUITY SETS

Structured ambiguity sets seek to exclude distributions that significantly deviate from the true distribution to reduce the gap between the original stochastic optimization problem and its distributionally robust approximation. To obtain such ambiguity set for data-driven problems, we consider the case where the uncertainty $\xi \equiv (\xi_1, \ldots, \xi_n) \in \Xi \equiv \Xi_1 \times \ldots \times \Xi_n$ consists of n independent components ξ_k, $k \in [n]$. This implies that P_ξ is necessarily a product measure, i.e., $P_\xi = P_{\xi_1} \otimes \ldots \otimes P_{\xi_n}$. The structured ambiguity set $\mathscr{P}^N$ is then constructed so that its elements are closer to a product reference measure than in a monolithic ambiguity ball. In particular, given a vector $\boldsymbol{\varepsilon} := (\varepsilon_1, \ldots, \varepsilon_n) \geq 0$ of transport budgets and a reference distribution Q, we introduce, similarly to the previous chapter, the multi-transport hyperrectangle (MTH)

$$\mathscr{T}_p(Q, \boldsymbol{\varepsilon}) := \Big\{ \mathrm{pr}_{2\#}\pi : \pi \in \mathscr{P}(\Xi \times \Xi),\ \mathrm{pr}_{1\#}\pi = Q, \text{ and } \int_{\Xi\times\Xi} \rho_k(\zeta_k, \xi_k)^p \mathrm{d}\pi(\zeta, \xi) \leq \varepsilon_k^p \text{ for all } k \in [n] \Big\}. \tag{3.3.5}$$

This ambiguity set consists of all distributions that respect a set of mass transport constraints, one for each component of the uncertainty. For data-driven problems where we have N i.i.d. samples of ξ, we select the product empirical distribution

$$Q \equiv \boldsymbol{P}_\xi^N := P_{\xi_1}^N \otimes \ldots \otimes P_{\xi_n}^N \tag{3.3.6}$$

as a reference distribution. Here $P_{\xi_k}^N := \frac{1}{N}\sum_{i=1}^N \delta_{\xi_k^i}$, $k \in [n]$ are the empirical distributions of the components of ξ. The MTH $\mathscr{T}_p(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})$ shrinks at a favorable rate with the number of samples compared to Wasserstein ambiguity balls, under the premise that it is designed to contain the true distribution with prescribed probability. When $\Xi \subset \mathbb{R}^d$, this ameliorates the curse of dimensionality that Wasserstein balls face with respect to d [59], especially when the number n of independent components is large, see Proposition 2.5.2 of Chapter 2.

The independence property across the components of the uncertainty arises in several problems of interest, such as the control of linear stochastic systems [141, 142], model predictive control formulations [106], multi-agent systems [137–139], and data-driven multistage scheduling of energy systems [143, 144]. As we further elaborate in Section VIII, the results of this approach can still be exploited when dependencies between components are present and quantified in terms of the Wasserstein distance between the joint distribution and the product of its marginals, by appropriately inflating the ambiguity sets. Our ambiguity set design can also be generalized to data-driven cases where a different number of realizations is available for each independent component ξ_k of ξ, as these realizations may be collected separately or asynchronously. Then each marginal empirical distribution is constructed using a different number of samples.

Remark 3.3.1. (Connection to multistage DRO). Product ambiguity sets have been employed in multistage DRO (see [142–144, 165–167]). In this setting, uncertainty is typically assumed to be stage-wise independent, which leads naturally to ambiguity

sets constructed as products of marginal ambiguity sets, typically Wasserstein balls. However, such ambiguity sets generally lack convexity. As a result, it is no longer clear when strong duality holds for the associated maximization problems, and exact solutions to the resulting DRO formulations typically rely on affine policy parameterizations or on specific structural properties of the underlying optimization problem.

Products of Wasserstein balls were in fact introduced in our work [164], where they were termed Wasserstein hyperrectangles. In the previous chapter, we showed that MTHs constitute tight convex over-approximations of Wasserstein hyperrectangles. In particular, the two sets share exactly the same product distributions, Proposition 2.4.4, as well as the same extreme distributions, i.e., the distributions furthest from the reference model, when the order of the Wasserstein metric matches that of the underlying norm-induced distance (i.e., $p = q$; see Proposition 2.4.6).

Most notably, MTHs are convex and therefore admit strong dual reformulations for a very broad class of objective functions (see Theorem 2.6.4). In light of these considerations, there is prospect in exploiting MTHs to obtain exact (dual) reformulations for multistage DRO problems under arbitrary finite-dimensional policy parameterizations.

3.3.2. TRACTABILITY

Both the DRO problem (3.3.3) and the robustified chance-constrained problem (3.3.4) typically involve solving infinite-dimensional average-cost maximization problems over a space of probability distributions, namely

$$\sup_{P\in\mathscr{P}^N} \mathbb{E}_P\big[f(x,\xi)\big] \quad \text{and} \quad \sup_{P\in\mathscr{P}^N} \mathbb{E}_P[\mathbb{1}_{\{\xi\in\Xi:\, F(x,\xi)>0\}}(\xi)],$$

respectively. Thus, their numerical solution necessitates their reformulation as finite-dimensional optimization problems. It is further desirable to identify problem classes for which these reformations can be solved by efficient algorithms. The first objective, i.e., the derivation of dual finite-dimensional reformulations of such DRO problems over MTHs $\mathscr{P}^N \equiv \mathscr{T}_p(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})$ has been established in Chapter 2. Our goal in this chapter is also to address the second objective for DRO problems over MTHs, namely, to develop further key properties of MTHs and combine them with the duality results from the previous chapter to derive tractable reformulations of DRO problems for multiple classes of interest.

Building on reformulations for Wasserstein DRO and chance-constrained problems for suitable classes of costs and constraints, such as piecewise-affine or quadratic [79, 101, 168], *we seek to derive tractable DRO reformulations over MTHs.* To this end, *we also aim to establish fundamental properties of MTHs, such as weak compactness,* which play a key role in obtaining certain reformulations. Our final goal is to address the potential complexity of the derived reformulations, which depends on the number of atoms comprising the product empirical distribution $\boldsymbol{P}_\xi^N$, i.e., the reference distribution of the hyperrectangle. Since the number of these atoms grows rapidly with the number n of the uncertainty components, we seek to determine

clustering strategies that guarantee a manageable complexity for DRO reformulations while preserving the probabilistic guarantees characterizing MTHs.

3.4. MULTI-TRANSPORT HYPERRECTANGLE (MTH) DRO: FUNDAMENTAL PROPERTIES

In this section, we establish several fundamental properties of MTHs. These include general weak compactness and duality results, which are used in later sections to obtain tractable reformulations for concrete classes of DRO and chance-constrained problems over MTHs.

To introduce the relevant notions, let (Ξ, ρ) be a Polish space. The weak topology on $\mathscr{P}(\Xi)$ is defined as the weakest topology that makes all mappings $P \mapsto \mathbb{E}_P[h]$ continuous for every bounded and continuous function h on Ξ. A subset $\mathscr{A} \subset \mathscr{P}(\Xi)$ is called *weakly compact* if it is compact in the weak topology. Since $\mathscr{P}(\Xi)$ is metrizable when Ξ is Polish [169, Proposition 7.20], the notions of weak compactness, continuity, and semicontinuity coincide with their weak *sequential* counterparts. We next assume the following product structure for the Polish space Ξ.

Assumption 5. (Metric space class). Let $\Xi := \Xi_1 \times \cdots \times \Xi_n$ and consider the Polish spaces (Ξ, ρ) and (Ξ_k, ρ_k), $k \in [n]$. We assume that:
(i) The spaces (Ξ_k, ρ_k), $k \in [n]$ are proper, i.e., their closed and bounded subsets are compact.
(ii) The metric ρ on Ξ is equivalent to the product metric $(\sum_{k=1}^n \rho_k^q)^{1/q}$ for some $q \in [1, +\infty]$.

Under this assumption, we establish weak compactness of MTHs.

Theorem 3.4.1. *(Weak compactness). Let Assumption 5 hold. Then, for any reference measure $Q \in \mathscr{P}_p(\Xi)$ the MTH $\mathscr{T}_p(Q, \boldsymbol{\varepsilon})$ defined in (3.3.5) is weakly compact.*

This result is used later to establish dual reformulations of distributionally robust chance-constrained problems. The proof is given in the Appendix. For the rest of the section, we focus on the inner maximization problem of (3.3.3) associated with (3.3.5), which is carried out over an infinite-dimensional space of probability distributions. Namely, we consider the problem

$$\mathscr{I} := \sup_{P \in \mathscr{T}_p(Q, \boldsymbol{\varepsilon})} \mathbb{E}_P[h(\xi)], \tag{3.4.1}$$

where we fix the decision variable x in (3.3.3) and denote $h(\xi) := f(x, \xi)$ for notational ease. From a modeling perspective, we assume that the cost in (3.4.1) is finite-valued for both the reference and the true distribution. It is therefore meaningful to determine conditions under which this finitness property is also retained for the worst-case distribution from the MTH. These are delineated in the next result.

Proposition 3.4.2. *(Finiteness of the optimal value). Let Assumption 5(ii) hold, assume that the reference distribution Q has a finite pth moment, and let $\boldsymbol{\varepsilon} := (\varepsilon_1, \ldots, \varepsilon_n)$ with $\varepsilon_k > 0$, for all $k \in [n]$. Then the worst-case expectation (3.4.1) is finite if and only if $h \in \mathscr{G}_p^{\mathrm{up}}$.*

Proof. The proof relies on finding two Wasserstein balls, one that encloses the MTH and one that is enclosed by it, and exploiting validity of the desired result for these balls [170, Theorem 2]. To this end, note that by Assumption 5(ii) there is some $c > 0$ such that $c\rho_k \leq \rho$ for every $k \in [n]$ and let $\varepsilon_\star := c\min\{\varepsilon_k,\, k \in [n]\}$. For any $P \in \mathscr{B}_p(Q, \varepsilon_\star)$, we can pick by [117, Theorem 4.1] a transport plan $\pi \in \mathscr{C}(Q, P)$ such that

$$\int_{\Xi\times\Xi} \rho(\zeta,\xi)^p \pi(\zeta,\xi) \leq \varepsilon_\star^p.$$

Since $c\rho_k \leq \rho$, we get from the definition of $\varepsilon_\star$ that

$$\int_{\Xi\times\Xi} \rho_k(\zeta_k,\xi_k)^p d\pi(\zeta,\xi) \leq \int_{\Xi\times\Xi} \rho(\zeta,\xi)^p / c^p \, d\pi(\zeta,\xi) \leq \varepsilon_\star^p / c^p = \varepsilon_k^p$$

for all $k \in [n]$, which yields $\mathscr{B}_p(Q, \varepsilon_\star) \subset \mathscr{T}_p(Q, \boldsymbol{\varepsilon})$. From Proposition 2.4.6 of Chapter 2 and Assumption 5(ii), we can also select a sufficiently large ball $\mathscr{B}_p(Q, \varepsilon^\star)$ that contains $\mathscr{T}_p(Q, \boldsymbol{\varepsilon})$. Thus,

$$\sup_{P\in\mathscr{B}_p(Q,\varepsilon_\star)} \mathbb{E}_P[h(\xi)] \leq \sup_{P\in\mathscr{T}_p(Q,\boldsymbol{\varepsilon})} \mathbb{E}_P[h(\xi)] \leq \sup_{P\in\mathscr{B}_p(Q,\varepsilon^\star)} \mathbb{E}_P[h(\xi)]$$

and it follows from [170, Theorem 2] that the optimal value of (3.4.1) is $+\infty$ if the conditions in the statement are not met and finite otherwise. This completes the proof. □

In the next result, we strengthen the conditions of Proposition 3.4.2 to establish that the supremum in (3.4.1) can indeed be attained. These conditions also yield weak continuity of the expected value in (3.4.1) with respect to the distributions in the MTH.

Theorem 3.4.3. *(Existence of optimal solution & continuity over the distribution). Let Assumption 5 hold and assume the reference distribution Q has a finite pth moment. Then:*
(i) If $h \in \mathscr{G}_r^{\mathrm{up}}$ for some $r \in [0, p)$ and h is upper semicontinous, the supremum in (3.4.1) is attained.
(ii) If $h \in \mathscr{G}_r$ for some $r \in [0, p)$ and h is continous, the map $\Psi : \mathscr{P}_p(\Xi) \to \mathbb{R}$ with $\Psi(P) := \mathbb{E}_P[h(\xi)]$ is weakly continuous on $\mathscr{T}_p(Q, \boldsymbol{\varepsilon})$.

We provide the proof in the Appendix of this chapter. From (3.3.5), one can directly verify that MTHs are convex as they are defined through a finite number of linear constraints in the space of couplings on $\Xi \times \Xi$. This enables the derivation of dual reformulations for DRO problems over MTHs. These equivalent formulations avoid the maximization in the infinite-dimensional space of probability distributions and provide the stepping stone to obtain tractable DRO problems. Next, recalling Theorem 2.6.4 of Chapter 2, we can evaluate (3.4.1), when the reference distribution is discrete, i.e., $Q = \sum_{l=1}^M \vartheta_l \delta_{\xi^l}$, by solving the following dual optimization problem

$$\inf_{\boldsymbol{\lambda}\geq 0} \Big\{ \langle \boldsymbol{\lambda}, \boldsymbol{\varepsilon} \rangle + \sum_{l=1}^M \vartheta_l \sup_{\xi\in\Xi} \Big\{ h(\xi) - \sum_{k=1}^n \lambda_k \rho_k(\xi_k^l, \xi_k)^p \Big\} \Big\}.$$

The expression of the dual problem over discrete distributions, such as the empirical or the product empirical distribution, hints at the fact that tractable reformulations hinge on explicit ways to evaluate the suprema therein with respect to ξ. This will indeed be the case for all the specific problem classes that we tackle in the later sections.

3.5. TRACTABLE REFORMULATIONS FOR DRO PROBLEMS ASSOCIATED WITH MTHS

In this section, we consider MTHs (3.3.5) that are centered at a discrete distribution Q and exploit Theorem 2.6.4 of the previous chapter to derive tractable reformulations of the DRO problem (3.4.1). We make the following assumption regarding the reference distribution Q.

Assumption 6. (Reference distribution). The reference distribution Q is discrete and consists of M atoms $\xi^1,\ldots,\xi^M$ with masses $\vartheta_1,\ldots,\vartheta_M$, respectively.

Here, we mainly build on the convex reformulations from [79], which considers ambiguity balls using the 1-Wasserstein distance in $\mathbb{R}^d$ equipped with some norm $\|\cdot\|$. We therefore focus on $\mathbb{R}^d$-valued random variables $\xi = (\xi_1,\ldots,\xi_n) \in \Xi \subset \mathbb{R}^d \equiv \mathbb{R}^{d_1} \times \ldots \times \mathbb{R}^{d_n}$. To transform the optimization problem (3.4.1) into a tractable convex program, we also need to make certain assumptions about the domain of the uncertainty Ξ and the objective function h. As in [79], we assume that the objective function can be expressed as the point-wise maximum of a finite family of concave functions.

Assumption 7. (Convex decomposition). The set $\Xi \subset \mathbb{R}^d$ is convex and closed. Further, the objective function $h : \mathbb{R}^d \to \mathbb{R}$ is real-valued and has the form $h(\xi) = \max_{j\in[m]} h_j(\xi)$, for some $m \in \mathbb{N}_{>0}$, where the functions $-h_j : \mathbb{R}^d \to \bar{\mathbb{R}}$, $j \in [m]$ are convex, proper, lower semi-continuous, and not identically $+\infty$ on Ξ (although some of the functions h_j may attain the value $-\infty$ individually, we assume their collective maximum to be always finite).

In the next result, we exploit Assumptions 6 and 7 to derive a finite-dimensional convex reformulation for (3.4.1). The proofs of this section follow the same reasoning as the corresponding proofs in [79]. The main difference, which does not introduce any additional difficulty in carrying out the proofs, is that they rely on the duality result of Theorem 2.6.4 of Chapter 2 over MTHs rather than on standard duality theory for Wasserstein balls. For completeness, the proofs of the section are provided in Appendix 3.A.

Proposition 3.5.1. *(Finite convex program). Consider the optimization problem* (3.4.1) *with* $p = 1$ *and some* $\varepsilon \geq 0$. *If* Q *and* h *satisfy Assumptions 6 and 7, respectively, then the worst-case expectation* (3.4.1) *can be evaluated by solving the finite convex*

program

$$\begin{aligned}\inf_{\substack{\boldsymbol{\lambda},s_l,z_{lj},v_{lj}\\ l\in[M],\,j\in[m]}}\ & \langle \boldsymbol{\lambda},\boldsymbol{\varepsilon}\rangle + \sum_{l=1}^{M}\vartheta_l s_l \\ \text{s.t.}\ & [-h_j]^*(z_{lj}-v_{lj}) + \sigma_\Xi(v_{lj}) - \langle z_{lj},\xi^l\rangle \le s_l \\ & \|\mathrm{pr}_k^{\boldsymbol{d}}(z_{lj})\|_* \le \lambda_k \\ & l\in[M],\ j\in[m],\ k\in[n].\end{aligned} \tag{3.5.1}$$

3

Here $\boldsymbol{\lambda} := (\lambda_1,\ldots,\lambda_n)$, $\boldsymbol{d} := (d_1,\ldots,d_n)$, and σ_Ξ is the support function of Ξ.

Next, we specialize this result to piecewise affine objective functions. These form a rather general function class, as they can approximate any nonlinear function of bounded variation with arbitrary accuracy [171].

Proposition 3.5.2. *(Piecewise affine objective functions). Assume $p=1$ and suppose that the domain of the uncertainty set is the polyhedral set $\Xi := \{\xi\in\mathbb{R}^d : C\xi \preceq q\}$, where $C\in\mathbb{R}^{r\times d}$, $q\in\mathbb{R}^r$, and that $h_j(\xi) := \langle \alpha_j,\xi\rangle + b_j$, $j\in[m]$.*
(i) If $h(\xi) = \max_{j\in[m]} h_j(\xi)$, then (3.4.1) can be evaluated by solving

$$\begin{aligned}\inf_{\substack{\boldsymbol{\lambda},s_l,\gamma_{lj}\\ l\in[M],\,j\in[m]}}\ & \langle \boldsymbol{\lambda},\boldsymbol{\varepsilon}\rangle + \sum_{l=1}^{M}\vartheta_l s_l \\ \text{s.t.}\ & b_j + \langle \alpha_j,\xi^l\rangle + \langle \gamma_{lj}, q - C\xi^l\rangle \le s_l \\ & \left\|\mathrm{pr}_k^{\boldsymbol{d}}(C^\top\gamma_{lj}-\alpha_j)\right\|_* \le \lambda_k \\ & \gamma_{lj}\succeq 0,\ l\in[M],\ j\in[m],\ k\in[n].\end{aligned} \tag{3.5.2}$$

(ii) If $h(\xi) = \min_{j\in[m]} h_j(\xi)$, then (3.4.1) can be evaluated by solving

$$\begin{aligned}\inf_{\substack{\boldsymbol{\lambda},s_l,\gamma_l,\theta_l\\ l\in[M]}}\ & \langle \boldsymbol{\lambda},\boldsymbol{\varepsilon}\rangle + \sum_{l=1}^{M}\vartheta_l s_l \\ \text{s.t.}\ & \langle \theta_l, b + A\xi^l\rangle + \langle \gamma_l, q - C\xi^l\rangle \le s_l \\ & \left\|\mathrm{pr}_k^{\boldsymbol{d}}(C^\top\gamma_l - A^\top\theta_l)\right\|_* \le \lambda_k \\ & \langle \theta_l,\mathbf{1}\rangle = 1 \\ & \gamma_l\succeq 0,\ \theta_l\succeq 0,\ l\in[M],\ k\in[n].\end{aligned} \tag{3.5.3}$$

In these reformulations, $A\in\mathbb{R}^{m\times d}$ is the matrix with rows $\alpha_j^\top$, b is the column vector with entries b_j, $\mathbf{1}$ is a vector of ones, $\boldsymbol{\lambda} := (\lambda_1,\ldots,\lambda_n)$, and $\boldsymbol{d} := (d_1,\ldots,d_n)$.

Based on [101] we also consider quadratic objective functions and derive tractable reformulation when $p=2$, $\Xi=\mathbb{R}^d$ and $\|\cdot\|$ is the Euclidean norm.

Proposition 3.5.3. *(Indefinite quadratic objective functions). Assume $\Xi=\mathbb{R}^d$ is equipped with the Euclidean norm, and consider the indefinite quadratic loss function*

$h(\xi) = \xi^\top \mathcal{Q}\xi + 2q^\top \xi$ with $\mathcal{Q} \in \mathbb{S}_d$ and $q \in \mathbb{R}^d$. Then, under Assumption 6, the worst-case expectation (3.4.1) with $p = 2$, can be evaluated by solving the tractable semi-definite program

$$\begin{array}{ll} \inf\limits_{\substack{\boldsymbol{\lambda}\geq 0, s_l \geq 0 \\ l\in[M]}} & \langle \boldsymbol{\lambda}, \boldsymbol{\epsilon}\rangle + \sum\limits_{l=1}^{M} \vartheta_l s_l \\ \text{s.t.} & \begin{bmatrix} \operatorname{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda}) - \mathcal{Q} & q + \operatorname{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda})\xi^l \\ q^\top + {\xi^l}^\top \operatorname{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda}) & s_l + {\xi^l}^\top \operatorname{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda})\xi^l \end{bmatrix} \succeq 0 \\ & l \in [M] \end{array} \tag{3.5.4}$$

where $\boldsymbol{\epsilon} := (\epsilon_1^2, \ldots, \epsilon_n^2)$, $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_n)$, $\boldsymbol{d} := (d_1, \ldots, d_n)$, and $\operatorname{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda})$ refers to the $d \times d$ diagonal matrix that takes the value λ_k from the $\ell_{k-1}+1$th to the ℓ_kth entry of its diagonal, where $\ell_k := d_1 + \cdots + d_k$ for $k \geq 1$ and $\ell_0 := 0$.

3.6. UNCERTAINTY QUANTIFICATION USING STRUCTURED AMBIGUITY SETS

In this section, we develop tractable reformulations to solve uncertainty quantification problems over MTHs. Such problems can be of important practical interest in applications where we seek to assess whether a physical or engineered system is safe or not. To this end, we determine a set of constraints that must be satisfied by the uncertain state of the system with a specified probability so that the system can be qualified as safe or unsafe.

We focus on scenarios where the distribution of the state is unknown and we can only exploit samples to infer worst-case probabilities of being either safe or unsafe with high confidence. To achieve this, we exploit the reformulations from [79, Corollary 5.3] that allow determining worst- and best-case probabilities for a random variable ξ to belong to a polyhedral safe set $\mathbb{A}$ or not and extend it to the more general case where the safe set is the union of polyhedral sets. Specifically, we solve the problem

$$\sup_{P \in \mathcal{T}_1(Q,\boldsymbol{\epsilon})} P(\xi \in \mathbb{A}), \tag{3.6.1}$$

where $\mathbb{A} := \cup_{j=1}^m \mathbb{A}_j$ and $\mathbb{A}_j$ are polyhedral sets for all $j \in [m]$.

Theorem 3.6.1. *(Uncertainty quantification for unions of polyhedral sets). Let the support of P_ξ be the nonempty polyhedral set $\Xi := \{\xi \in \mathbb{R}^d : C\xi \preceq f\}$ and assume that each set $\mathbb{A}_j := \{\xi \in \mathbb{R}^d : A_j\xi \preceq b_j\}$, $j \in [m]$ has nonempty intersection with Ξ. Then, under Assumption 6, the probability (3.6.1) can be evaluated by solving the convex*

program

$$\begin{aligned}\inf_{\substack{\boldsymbol{\lambda},s_l,\gamma_{lj},\theta_{lj}\\ l\in[M],j\in[m]}}\ & \langle\boldsymbol{\lambda},\boldsymbol{\varepsilon}\rangle+\sum_{l=1}^{M}\vartheta_l s_l\\ \text{s.t.}\ & 1+\langle\theta_{lj},b_j-A_j\xi^l\rangle+\langle\gamma_{lj},f-C\xi^l\rangle\le s_l\\ & \|\mathrm{pr}_k^{\boldsymbol{d}}(A_j^\top\theta_{lj}+C^\top\gamma_{lj})\|_*\le\lambda_k\\ & \gamma_{lj}\ge 0,\ \theta_{lj}\ge 0,\ s_l\ge 0\\ & l\in[M],\ j\in[m],\ k\in[n],\end{aligned}\tag{3.6.2}$$

where $\boldsymbol{\lambda}:=(\lambda_1,\ldots,\lambda_n)$ *and* $\boldsymbol{d}:=(d_1,\ldots,d_n)$.

Proof. To prove the theorem, we exploit the fact that

$$\sup_{P\in\mathcal{T}_1(Q,\boldsymbol{\varepsilon})}P(\xi\in\mathbb{A})=\sup_{P\in\mathcal{T}_1(Q,\boldsymbol{\varepsilon})}\mathbb{E}_P[\mathbb{1}_{\mathbb{A}}].$$

and use Proposition 3.5.1 with appropriate loss functions h_j to recover the indicator function $\mathbb{1}_{\mathbb{A}}$. To this end, let $h_j:=1-\chi_{\mathbb{A}_j}$ for each $j\in[m]$, set $h_{m+1}:=0$, and note that

$$\mathbb{1}_{\mathbb{A}}(\xi)=\mathbb{1}_{\mathbb{A}_1\cup\cdots\cup\mathbb{A}_m}(\xi)=\max\{h_1(\xi),\ldots,h_{m+1}(\xi)\}=\max\{1-\chi_{\mathbb{A}_1}(\xi),\ldots,1-\chi_{\mathbb{A}_m}(\xi),0\}$$

for all $\xi\in\mathbb{R}^d$. The conjugates of the first m functions inside the max are

$$\begin{aligned}[-h_j]^*(z)&=\sup_{\xi\in\mathbb{A}_j}\langle z,\xi\rangle+1\\ &=\begin{cases}\sup\limits_{\xi}\ \langle z,\xi\rangle+1\\ \text{s.t. } A_j\xi\preceq b_j\end{cases}=\begin{cases}\inf\limits_{\theta\ge 0}\ \langle\theta,b_j\rangle+1\\ \text{s.t. } A_j^\top\theta=z.\end{cases}\end{aligned}\tag{3.6.3}$$

Here the last equality follows from the assumption that $\mathbb{A}_j$ is nonempty and linear programming duality. We also express the support function of Ξ as

$$\sigma_\Xi(v)=\begin{cases}\sup\limits_{\xi}\ \langle v,\xi\rangle\\ \text{s.t. } C\xi\preceq f\end{cases}=\begin{cases}\inf\limits_{\gamma\ge 0}\ \langle\gamma,f\rangle\\ \text{s.t. } C^\top\gamma=v,\end{cases}\tag{3.6.4}$$

where the last equality follows from linear programming duality.

Substituting the expressions in (3.6.3) and (3.6.4) into the first set of constrainst in (3.5.1), we get for each $l\in[M]$ and $j\in[m]$ the constraints

$$[-h_j]^*(z_{lj}-v_{lj})+\sigma_\Xi(v_{lj})-\langle z_{lj},\xi^l\rangle\equiv 1+\langle\theta_{lj},b_j\rangle+\langle\gamma_{lj},f\rangle-\langle z_{lj},\xi^l\rangle\le s_l$$

and $C^\top\gamma_{lj}=v_{lj}$, $A_j^\top\theta_{lj}=z_{lj}-v_{lj}$, $\gamma_{lj}\ge 0$, $\theta_{lj}\ge 0$, which are equivalent to $z_{lj}=A_j^\top\theta_{lj}+C^\top\gamma_{lj}$, $\gamma_{lj}\ge 0$, $\theta_{lj}\ge 0$. Substituting also the last expression for z_{lj} in the above inequality constraint and the second set of constraints in (3.5.1), yields for each $l\in[M]$, $k\in[n]$, and $j\in[m]$ the equivalent set of constraints

$$1+\langle\theta_{lj},b_j-A_j\xi^l\rangle+\langle\gamma_{lj},f-C\xi^l\rangle\le s_l$$

$$\left\| \mathrm{pr}_k^{\boldsymbol{d}}(A_j^\top \theta_{lj} + C^\top \gamma_{lj}) \right\|_* \le \lambda_k,\ \gamma_{lj} \succeq 0, \theta_{lj} \ge 0.$$

Similarly, we find that $[-h_{m+1}]^*(z) = 0$ if $z = 0$ and $+\infty$ otherwise. Exploiting (3.6.4) and using similar steps as above, for each l and k, the corresponding constraints in (3.5.1) become

$$\begin{aligned} \inf_{\gamma_{l,m+1} \succeq 0} \langle \gamma_{l,m+1}, f - C\xi^l \rangle &\le s_l \\ \left\| \mathrm{pr}_k^{\boldsymbol{d}}(C^\top \gamma_{l,m+1}) \right\|_* &\le \lambda_k, \end{aligned}$$

Since $f - C\xi^l \succeq 0$, the expression on the left-hand side of the first constraint attains its infimum for $\gamma_{l,m+1} = 0$ while automatically satisfying the second constraint. Hence, we get the equivalent constraints $s_l \ge 0$ for all $l \in [M]$. This establishes the desired result. □

Note that besides generalizing [79, Corollary 5.3(ii)] from Wasserstein ambiguity balls to MTHs, Theorem 3.6.1 also generalizes the evaluation of worst/best case probability from polyhedral sets to the union of polyhedral sets.

Instead of assessing how safe the set $\mathbb{A}$ can be, we may also want to know with high confidence how likely it is for ξ to lie outside $\mathbb{A}$. This yields the uncertainty quantification problem

$$\sup_{P \in \mathcal{T}_1(Q,\boldsymbol{\varepsilon})} P(\xi \notin \mathbb{A}), \tag{3.6.5}$$

where $\mathbb{A} := \cup_{j=1}^m \mathbb{A}_j$ and each $\mathbb{A}_j$ is a polyhedral set. The next result evaluates this worst-case probability.

Corollary 3.6.2. *(Uncertainty quantification for complements of unions of polyhedral sets). Assume that P_ξ is supported on the nonempty polyhedral set $\Xi := \{\xi \in \mathbb{R}^d : C\xi \preceq f\}$ and let $\mathbb{A}_j := \{\xi \in \mathbb{R}^d : \langle a_j^l, \xi \rangle < b_j^l \text{ for all } l \in [\alpha_j]\}$, $j \in [m]$. Then, under Assumption 6, the value of the program*

$$\begin{aligned} \inf_{\substack{\boldsymbol{\lambda}, s_l, \gamma_{lq}, \theta_{lq} \\ l \in [M], q \in Q}} \ & \langle \boldsymbol{\lambda}, \boldsymbol{\varepsilon} \rangle + \sum_{l=1}^M \vartheta_l s_l \\ \text{s.t.} \ & 1 - \langle \theta_{lq}, b_q - A_q \xi^l \rangle + \langle \gamma_{lq}, f - C\xi^l \rangle \le s_l \\ & \left\| \mathrm{pr}_k^{\boldsymbol{d}} \left(C^\top \gamma_{lq} - A_q^\top \theta_{lq} \right) \right\|_* \le \lambda_k \\ & \gamma_{lq} \succeq 0,\ \theta_{lq} \ge 0,\ s_l \ge 0 \\ & l \in [M],\ q \in Q,\ k \in [n] \end{aligned}$$

is equal to the probability (3.6.5). Here $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_n)$, $\boldsymbol{d} := (d_1, \ldots, d_n)$ and for any $q := (q_1, \ldots, q_m) \in \prod_{j=1}^m [\alpha_j]$, $A_q \in \mathbb{R}^{m \times d}$ is the matrix formed by concatenating the row vectors $(a_j^{q_j})^\top$ and $b_q := (b_1^{q_1}, \ldots, b_m^{q_m})$. The set Q comprises all indices $q \in \prod_{j=1}^m [\alpha_j]$ for which the sets $\{\xi \in \mathbb{R}^d : A_q \xi \succeq b_q\}$ have nonempty intersection with Ξ.

Proof. Since each polyhedral set $\mathbb{A}_j$ is the intersection of the half-spaces $\mathscr{A}_j^{q_j} := \{\xi \in \mathbb{R}^d : \langle a_j^{q_j}, \xi \rangle < b_j^{q_j}\}$, we have that $\mathbb{A}_j = \cap_{q_j=1}^{\alpha_j} \mathscr{A}_j^{q_j}$. Taking also into account that

$$\sup_{P \in \mathscr{T}_1(Q,\boldsymbol{\varepsilon})} P(\xi \notin \mathbb{A}) = \sup_{P \in \mathscr{T}_1(Q,\boldsymbol{\varepsilon})} P\left(\xi \in \mathbb{A}^c\right)$$

and that $\mathbb{A} = \cup_{j=1}^m \mathbb{A}_j$, we get

$$\begin{aligned}\mathbb{A}^c &= \cap_{j=1}^m \mathbb{A}_j{}^c = \cap_{j=1}^m (\cap_{q_j=1}^{\alpha_j} \mathscr{A}_j^{q_j})^c = \cap_{j=1}^m \cup_{q_j=1}^{\alpha_j} (\mathscr{A}_j^{q_j})^c \\ &= \cup_{(q_1,\ldots,q_m) \in [\alpha_1] \times \cdots \times [\alpha_m]} \cap_{j=1}^m (\mathscr{A}_j^{q_j})^c,\end{aligned}$$

which follows from De Morgan's law and the distributivity between unions and intersections. Denoting $\mathbb{B}_q := \cap_{j=1}^m (\mathscr{A}_j^{q_j})^c$, we deduce that $\mathbb{A}^c$ is the union of the sets $\mathbb{B}_q$ where $q := (q_1, \ldots, q_m) \in [\alpha_1] \times \cdots \times [\alpha_m]$. These sets are equivalently expressed as

$$\begin{aligned}\mathbb{B}_q &= \{\xi \in \mathbb{R}^d : \langle a_j^{q_j}, \xi \rangle \geq b_j^{q_j} \text{ for all } j \in [m]\} \\ &= \{\xi \in \mathbb{R}^d : A_q \xi \geq b_q\} = \{\xi \in \mathbb{R}^d : -A_q \xi \leq -b_q\}. \end{aligned} \tag{3.6.6}$$

Keeping only the indexes q for which $\mathbb{B}_q$ has a nonempty intersection with Ξ, and invoking Theorem 3.6.1, we obtain the worst-case probability (3.6.5) by replacing A_j and b_j in the first and second set of constraints in (3.6.2) with $-A_q$ and $-b_q$ from (3.6.6), respectively. □

3.7. DISTRIBUTIONALLY ROBUST CHANCE-CONSTRAINED PROBLEMS

In this section, we generalize tractable reformulations for chance-constrained problems over Wasserstein ambiguity balls to MTHs. We consider optimization problems that may have multiple chance constraints (see e.g., [172, 173]). Such problems are typical in multistage decision making [99, Chapter 3], including stochastic model predictive control [12, 174, 175]. Introducing several chance constraints provides the freedom to assign higher tolerance levels to design requirements that are softer than others. For data-driven problems, verifying these constraints individually with high confidence may require accurate knowledge of how the probability mass is distributed across different regions of the uncertainty domain. Hedging against such distribution imperfections can be addressed by considering an ambiguity set that contains the data-generating distribution with high confidence. This further motivates our consideration of MTHs, since they share refined probabilistic guarantees of containing the true distribution. On the other hand, the alternative of grouping all individual chance constraints into a single probabilistic constraint may result in overly conservative decisions or even infeasibility, especially if they are violated for disjoint events.

We build on the approach in [176], which provides reformulations of problems with a single distributionally robust chance constraint over Wasserstein ambiguity balls. Here, we are interested in solving problems of the form

$$\begin{aligned} &\inf_{x\in\mathscr{X}} \langle c, x\rangle \\ &\text{s.t. } P_\xi(F_i(x,\xi)\le 0)\ge 1-\alpha_i \quad i\in[I], \end{aligned} \tag{3.7.1}$$

where $\mathscr{X}\subset\mathbb{R}^\ell$, $\xi\in\Xi\subset\mathbb{R}^d\equiv\mathbb{R}^{d_1}\times\cdots\times\mathbb{R}^{d_n}$ is a random variable, $c\in\mathbb{R}^\ell$, $I\in\mathbb{N}$, and each $F_i:\mathscr{X}\times\Xi\to\mathbb{R}$, $i\in[I]$ designates a chance constraint that needs to be fulfilled with probability at least $1-\alpha_i\in(0,1)$. We further assume that Ξ is closed and $\mathscr{X}$ is closed and convex. This optimization problem may be non-convex, even when the functions F_i are convex. To address this issue, in analogy to [176], we approximate the chance constraints using the conditional value at risk of $F(x,\xi)$. The *conditional value at risk* $\mathrm{CVaR}^P_{1-\alpha}(\gamma)$ of a random variable γ at level α is

$$\mathrm{CVaR}^P_{1-\alpha}(\gamma) := \inf_{\tau\in\mathbb{R}}\{\alpha^{-1}\mathbb{E}_P[(\gamma+\tau)_+]-\tau\},$$

(see [177, Theorem 2]). Exploiting the fact that

$$\mathrm{CVaR}^{P_\xi}_{1-\alpha}(F(x,\xi))\le 0 \implies P_\xi(F(x,\xi)\le 0)\ge 1-\alpha,$$

we approximate (3.7.1) by the CVaR-constrained problem

$$\begin{aligned} &\inf_{x\in\mathscr{X}} \langle c, x\rangle \\ &\text{s.t. } \inf_{\tau\in\mathbb{R}}\mathbb{E}_{P_\xi}[(F_i(x,\xi)+\tau)_+-\tau\alpha_i]\le 0 \quad i\in[I]. \end{aligned} \tag{3.7.2}$$

Our goal is to provide tractable reformulations for a distributionally robust version of this problem. We will build on the reformulations of the previous section to solve it when the distribution P_ξ belongs to the MTH $\mathscr{T}_p(Q,\boldsymbol{\varepsilon})$. We are interested in solving the distributionally robust chance-constrained problem

$$\begin{aligned} &\inf_{x\in\mathscr{X}} \langle c, x\rangle \\ &\text{s.t. } \sup_{P\in\mathscr{T}_p(Q,\boldsymbol{\varepsilon})}\ \inf_{\tau\in\mathbb{R}}\mathbb{E}_P[(F_i(x,\xi)+\tau)_+-\tau\alpha_i]\le 0 \quad i\in[I], \end{aligned} \tag{3.7.3}$$

which represents a robustified version of (3.7.2) against discrepancies in the reference distribution. Tractable reformulations for this distributionally robust chance-constrained problem rely on obtaining a tractable expression for each CVaR constraint. We make the following assumption regarding the functions F_i, $i\in[I]$.

Assumption 8. (Constraint function class). The functions $F_i:\mathscr{X}\times\Xi\to\mathbb{R}$ satisfy the following properties:
(i) For each $i\in[I]$ and $\xi\in\Xi$, the function $x\to F_i(x,\xi)$ is convex;
(ii) There exists $r\in[0,p)$ such that for each $i\in[I]$ and $x\in\mathscr{X}$, the function $\xi\to F_i(x,\xi)$ is continuous and belongs to $\mathscr{G}_r$.

Suppressing for the moment the subscript i of the functions in the constraints and their satisfaction probabilities, we seek a tractable characterization for the feasible set

$$\Big\{x \in \mathscr{X} : \sup_{P \in \mathscr{T}(Q,\boldsymbol{\varepsilon})} \inf_{\tau \in \mathbb{R}} \mathbb{E}_P[(F(x,\xi)+\tau)_+ - \tau\alpha] \leq 0\Big\}. \tag{3.7.4}$$

of a single chance-constraint where the function F satisfies Assumption 8.

Determining tractable reformulations for (3.7.4) hinges on interchanging the order between the sup and inf of the distributionally robust constraint. This possibility is guaranteed by the following result, which is based on the fact that MTHs are weakly compact. This result extends [168, Lemma I.V.2] to the case where the ambiguity set is the MTH (3.3.5). It also entails the consideration of a broader class of constraints, since in [168] the mappings $\xi \mapsto F(x,\xi)$ are assumed to be bounded. Here, instead, they only need to respect the growth condition of Assumption 8(ii).

Lemma 3.7.1. *(Min-max equality for the* CVaR *over MTHs). Suppose that Assumption 8 holds. Then:*
(i) For all $x \in \mathscr{X}$,

$$\sup_{P \in \mathscr{T}_p(Q,\boldsymbol{\varepsilon})} \inf_{\tau \in \mathbb{R}} \mathbb{E}_P[(F(x,\xi)+\tau)_+ - \tau\alpha] = \inf_{\tau \in \mathbb{R}} \sup_{P \in \mathscr{T}_p(Q,\boldsymbol{\varepsilon})} \mathbb{E}_P[(F(x,\xi)+\tau)_+ - \tau\alpha]; \tag{3.7.5}$$

(ii) The infimum on the right-hand-side of (3.7.5) with respect to τ is attained.

The proofs of this lemma and the other results of this section are given in the Appendix. Next, we exploit DRO duality for MTHs to derive a tractable characterization of the feasible set (3.7.4). The following result characterizes this set through a finite number of convex constraints.

Proposition 3.7.2. *(Equivalent characterization of the feasible set (3.7.4)). Under Assumptions 6 and 8, (3.7.4) consists of all $x \in \mathscr{X}$ for which the convex constraints*

$$\begin{aligned} &\langle \boldsymbol{\lambda}, \boldsymbol{\varepsilon} \rangle + \sum_{l=1}^{M} \vartheta_l s_l \leq \tau\alpha \\ &\sup_{\xi \in \Xi} \Big\{ F(x,\xi) + \tau - \sum_{k=1}^{n} \lambda_k \|\xi_k - \xi_k^l\|^p \Big\} \leq s_l \quad l \in [M] \end{aligned}$$

are met for some $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_n) \geq 0$ and $s_l \geq 0$, where $l \in [M]$.

In the next proposition, we consider the case where the function f is piecewise affine with respect to ξ. Therefore, we derive a general reformulation when the function is expressed as the maximum of affine functions over a polyhedral set.

Proposition 3.7.3. *(Equivalent characterization of (3.7.4) for piecewise affine constraints). Let Assumption 6 hold and consider the compact set $\Xi := \{\xi \in \mathbb{R}^d : C\xi \preceq h\}$, where $C \in \mathbb{R}^{q \times d}$ and $h \in \mathbb{R}^q$. Assume also that $F(x,\xi) := \max_{j \in [m]} \langle x, A_j\xi \rangle + b_j(x)$, where*

$A_j \in \mathbb{R}^{\ell\times d}$ and each $b_j : \mathbb{R}^n \to \mathbb{R}$ is convex. Then, (3.7.4) *consists of all $x \in \mathscr{X}$ for which the constraints*

$$\begin{aligned}
&\langle \boldsymbol{\lambda}, \boldsymbol{\varepsilon}\rangle + \sum_{l=1}^{M} \vartheta_l s_l \le \tau\alpha \\
&b_j(x) + \tau + \langle A_j^\top x - C^\top \eta_{lj}, \xi^l\rangle + \langle \eta_{lj}, h\rangle \le s_l \\
&\|\mathrm{pr}_k^{\boldsymbol{d}}(A_j^\top x - C^\top \eta_{lj})\|_* \le \lambda_k \\
&l \in [M],\ j \in [m],\ k \in [n]
\end{aligned}$$

are met for some $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_n) \ge 0$ and $s_l \ge 0$, $\eta_{lj} \ge 0$, $l \in [M]$, $j \in [m]$, with $\boldsymbol{d} := (d_1, \ldots, d_n)$.

Applying Propositions 3.7.2 and 3.7.3 to the distributionally robust CVaR constraint of (3.7.3), we obtain the following two corollaries. These respectively yield finite-dimensional reformulations for convex distributionally robust chance-constrained problems and tractable reformulations for problems with piecewise affine constraints.

Corollary 3.7.4. *(Epigraphical reformulation of CVaR-constrained problem* (3.7.3)*). Under Assumption 8, problem* (3.7.3) *is equivalent to the convex program*

$$\begin{aligned}
\inf_{\substack{x\in\mathscr{X},\tau_i,\boldsymbol{\lambda}_i,s_{li}\\ l\in[M],\, i\in[I]}}\quad & \langle c, x\rangle \\
\text{s.t.}\quad & \langle \boldsymbol{\lambda}_i, \boldsymbol{\varepsilon}\rangle + \sum_{l=1}^{M} \vartheta_l s_{li} \le \tau_i \alpha_i \\
& \sup_{\xi\in\Xi}\Big\{ f_i(x,\xi) + \tau_i - \sum_{k=1}^{n} \lambda_{ki}\|\xi_k - \xi_k^l\|^p \Big\} \le s_{li} \\
& s_{li} \ge 0,\ l \in [M],\ i \in [I].
\end{aligned}$$

Corollary 3.7.5. *(Tractable reformulation of* (3.7.3) *with piecewise affine constraints). Assume further that $\Xi := \{\xi \in \mathbb{R}^d : C\xi \le h\}$ is compact, where $C \in \mathbb{R}^{q\times d}$ and $h \in \mathbb{R}^q$, and $F_i(x,\xi) := \max_{j\in[m]} \langle x, A_{ji}\xi\rangle + b_{ji}(x)$, where $A_{ji} \in \mathbb{R}^{\ell\times d}$ and each $b_{ji} : \mathbb{R}^\ell \to \mathbb{R}$ is convex. Then* (3.7.3) *is equivalent to the convex optimization problem*

$$\begin{aligned}
\inf_{\substack{x\in\mathscr{X},\tau_i,\boldsymbol{\lambda}_i,s_{li},\eta_{lji}\\ l\in[M],\, j\in[m],\, i\in[I]}}\quad & \langle c, x\rangle \\
\text{s.t.}\quad & \langle \boldsymbol{\lambda}_i, \boldsymbol{\varepsilon}\rangle + \sum_{l=1}^{M} \vartheta_l s_{li} \le \tau_i \alpha \\
& b_{ji}(x) + \tau_i + \langle A_{ji}^\top x - C^\top \eta_{lji}, \xi^l\rangle + \langle \eta_{lji}, h\rangle \le s_{li} \\
& \Big\|\mathrm{pr}_k^{\boldsymbol{d}}(A_{ji}^\top x - C^\top \eta_{lji})\Big\|_* \le \lambda_{ik} \\
& s_{li} \ge 0 \quad \eta_{lji} \ge 0 \\
& l \in [M],\ j \in [m],\ i \in [I],\ k \in [n],
\end{aligned}$$

where $\boldsymbol{\lambda}_i := (\lambda_{i1}, \ldots, \lambda_{in})$ and $\boldsymbol{d} := (d_1, \ldots, d_n)$.

3.8. CLUSTERING THE PRODUCT EMPIRICAL DISTRIBUTION

In the previous sections, we established tractable reformulations for various classes of DRO problems over MTHs that are centered at an atomic reference distribution $\widehat{P}_\xi$. Recalling our motivation from data-driven problems where the uncertainty ξ consists of several independent components, in this section we turn our attention to the case where $\widehat{P}_\xi$ has a product atomic structure like the product empirical distribution $\boldsymbol{P}_\xi^N$ in (3.3.6) (c.f. Figure 3.1). The motivation for choosing it as a reference distribution comes from the fact that the corresponding MTH enjoys favorable probabilistic guarantees of containing the true underlying distribution, as discussed in Chapter 2. Nevertheless, these come with the cost of a higher complexity for the reformulations derived in the previous sections, which may increase exponentially with the number of independent lower-dimensional components n. For example, given a random vector with $n=5$ lower dimensional components and 100 collected samples, we would get a reference distribution comprising $M = N^n = 10^{10}$ atoms. Each atom corresponds to multiple variables and constraints in the associated reformulation, which grow linearly with M and are the primary factors determining its computational complexity. Therefore, directly using $\boldsymbol{P}_\xi^N$ for large n renders the results derived so far computationally intractable.

To address this issue, we cluster the product empirical distribution $\boldsymbol{P}_\xi^N$ into an atomic distribution which can yield DRO problems of an acceptable complexity. We also investigate how much it is required to enlarge the vector of transport budgets to retain the statistical guarantees of the new hyperrectangle containing the true distribution. Throughout the section, we will consider random variables ξ supported on a subset Ξ of $\mathbb{R}^d \equiv \mathbb{R}^{d_1} \times \ldots \times \mathbb{R}^{d_n}$ and we fix an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^d$.

Next, we present strategies to construct manageable reference distributions by clustering the atoms of $\boldsymbol{P}_\xi^N$. The clustered distributions enjoy the benefits of having a lower complexity than $\boldsymbol{P}_\xi^N$ while being typically closer to the true distribution compared to empirical models that only rely on i.i.d. samples of ξ. Clustering methods to reduce the complexity of data-driven DRO problems have also been considered in [178, 179]. Our key distinctive feature here is to adapt the MTH discrepancy structure to the clustered distributions so that the probabilistic guarantees associated with $\mathcal{T}_p(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})$ are maintained for the shifted hyperrectangle.

To achieve this objective, we consider the following approaches:

- **Direct clustering:** The first approach is to directly cluster the points on which $\boldsymbol{P}_\xi^N$ is supported into a discrete distribution $\widehat{P}_\xi$, which has a smaller number of atoms.
- **Component-wise clustering:** The second option is to obtain clustered versions $\widehat{P}_{\xi_k}$ of the marginal empirical distributions $P_{\xi_k}^N$, $k \in [n]$, and deduce a lower-complexity reference model from their product $\widehat{P}_{\xi_1} \otimes \cdots \otimes \widehat{P}_{\xi_n}$.
- **Multi-component clustering:** This option provides a middle ground, where one can form ℓ groups of consecutive components of ξ, decompose $\boldsymbol{P}_\xi^N$ into products with marginals $\boldsymbol{P}^N_{(\xi_1,\ldots,\xi_{m_1})}, \ldots, \boldsymbol{P}^N_{(\xi_{m_\ell},\ldots,\xi_n)}$, cluster them, and then take the product of these clusters.

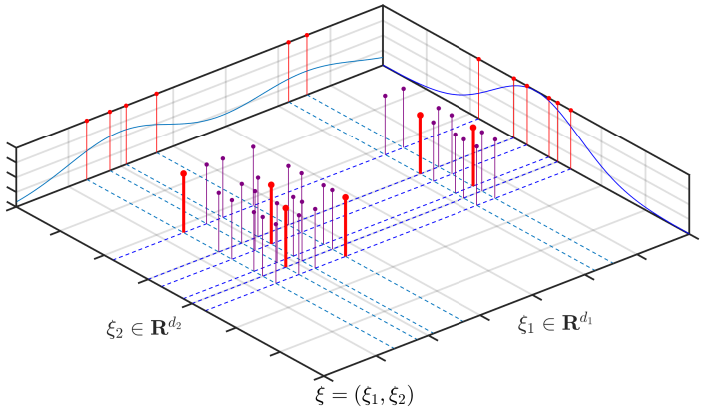


Figure 3.1.: The figure shows the empirical distribution P_ξ^N of $N = 6$ samples, represented by the thick red impulses, and the product empirical distribution $\boldsymbol{P}_\xi^N$, represented by the thin purple impulses. The product empirical distribution is the product of the marginal empirical distributions $P_{\xi_1}^N$ and $P_{\xi_2}^N$ of the independent components of ξ, which are depicted by the thin red impulses.

The derivation of clusters that are optimal in terms of their Wasserstein discrepancy from the reference distribution $\boldsymbol{P}_\xi^N$ can be performed using Lloyd's algorithm [180], [181, Chapter 9], which is guaranteed to converge to local minima. This is convenient since the complexity of each gradient step of the algorithm is linear in the number of points to be clustered. The clustering process can be further adapted by adjusting the complexity of lower-dimensional clusters and the choice to cluster in a monolithic or a component-wise fashion.

By modifying the reference distribution of the hyperrectangles, the probabilistic guarantees characterizing $\mathscr{T}_p(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})$ (cf. Proposition 2.5.2 in the previous chapter) do not necessarily hold anymore. Yet, we show how they can be recovered by appropriately adjusting the transport budgets ε_k in each direction. For simplicity, we only consider the two extreme cases where we either directly cluster all the samples into a monolithic distribution or cluster all marginal empirical distributions and form their associated product.

Proposition 3.8.1. *(Containment guarantees via hyperrectangle inflation). Assume that $P_\xi \in \mathscr{T}_p(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})$ and consider the reference distribution $\widehat{P}_\xi$.*

(i) If there is a transport plan $\hat{\pi} \in \mathscr{C}(\widehat{P}_\xi, \boldsymbol{P}_\xi^N)$ such that $\int_{\Xi\times\Xi} \|\eta_k - \zeta_k\|^p d\hat{\pi}(\eta,\zeta) \le \epsilon_k^p$ holds with some $\epsilon_k \ge 0$, for all $k \in [n]$, then

$$P_\xi \in \mathscr{T}_p(\widehat{P}_\xi, \boldsymbol{\varepsilon} + \boldsymbol{\epsilon}) \tag{3.8.1}$$

where $\boldsymbol{\epsilon} := (\epsilon_1, \ldots, \epsilon_n)$.

(ii) If additionally $\widehat{P}_\xi$ has the product structure $\widehat{P}_{\xi_1} \otimes \cdots \otimes \widehat{P}_{\xi_n}$, then (3.8.1) also holds with $\boldsymbol{\epsilon} := (\epsilon_1, \ldots, \epsilon_n)$ and $\epsilon_k := W_p(P_{\xi_k}^N, \widehat{P}_{\xi_k})$, for all $k \in [n]$.

Proof. Since $P_\xi \in \mathscr{T}_p(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})$, there is by (3.3.5) a transport plan π between $\boldsymbol{P}_\xi^N$ and P_ξ so that

$$\int_{\Xi\times\Xi} \|\zeta_k - \xi_k\|^p \mathrm{d}\pi(\zeta,\xi) \leq \varepsilon_k^p \tag{3.8.2}$$

for all $k \in [n]$. To prove part (i), let $\widehat{\pi}$ be a transport plan as in the statement. Since $\widehat{\pi} \in \mathscr{C}(\widehat{P}_\xi, \boldsymbol{P}_\xi^N)$, and thus π and $\widehat{\pi}$ share $\boldsymbol{P}_\xi^N$ as their first and second marginal, respectively, it follows from the Gluing Lemma [117] that there exists a distribution $\widetilde{\pi}$ on Ξ^3 such that $\mathrm{pr}_{12\#}(\widetilde{\pi}) = \widehat{\pi}$ and $\mathrm{pr}_{23\#}(\widetilde{\pi}) = \pi$. Then if we consider the transport plan $\pi' := \mathrm{pr}_{13\#}\widetilde{\pi}$, it follows that

$$\begin{aligned}
&\Big(\int_{\Xi\times\Xi} \|\eta_k - \xi_k\|^p d\pi'(\eta,\xi)\Big)^{\frac{1}{p}} \\
&\quad= \Big(\int_{\Xi^3} \|\eta_k - \xi_k\|^p d\widetilde{\pi}(\eta,\zeta,\xi)\Big)^{\frac{1}{p}} \\
&\quad\leq \Big(\int_{\Xi^3} (\|\eta_k - \zeta_k\| + \|\zeta_k - \xi_k\|)^p d\widetilde{\pi}(\eta,\zeta,\xi)\Big)^{\frac{1}{p}} \\
&\quad\leq \Big(\int_{\Xi^3} \|\eta_k - \zeta_k\|^p d\widetilde{\pi}(\eta,\zeta,\xi)\Big)^{\frac{1}{p}} + \Big(\int_{\Xi^3} \|\zeta_k - \xi_k\|^p d\widetilde{\pi}(\eta,\zeta,\xi)\Big)^{\frac{1}{p}} \\
&\quad= \Big(\int_{\Xi\times\Xi} \|\eta_k - \zeta_k\|^p d\widehat{\pi}(\eta,\zeta)\Big)^{\frac{1}{p}} + \Big(\int_{\Xi\times\Xi} \|\zeta_k - \xi_k\|^p d\pi(\zeta,\xi)\Big)^{\frac{1}{p}} \leq \varepsilon_k + \epsilon_k,
\end{aligned}$$

for all $k \in [n]$. Here, we used the triangle inequality in the first inequality, Minkowski's inequality in the second one, and Fubini's theorem in the last equality. By the definition of the MTH (3.3.5), this establishes part (i).

To show part (ii), since $W_p(P_{\xi_k}^N, \widehat{P}_{\xi_k}) = \epsilon_k$, there exists by [117, Theorem 4.1] an optimal transport plan $\widehat{\pi}_k$ that minimizes the transport cost between $P_{\xi_k}^N$ and $\widehat{P}_{\xi_k}$ and therefore

$$\int_{\Xi\times\Xi} \|\xi_k - \zeta_k\|^p d\widehat{\pi}_k(\xi_k,\zeta_k) = \epsilon_k^p$$

for each $k \in [n]$. Then if we define

$$\widehat{\pi} := T_\# \bigotimes_{k=1}^n \widehat{\pi}_k, \tag{3.8.3}$$

where $T : \prod_{k=1}^n \Xi_k \times \Xi_k \to \prod_{k=1}^n \Xi_k \times \prod_{k=1}^n \Xi_k$ is the linear map $T(\zeta_1,\xi_1,\ldots,\zeta_n,\xi_n) := (\zeta_1,\ldots,\zeta_n,\xi_1,\ldots,\xi_n)$, one can readily check as in proof of Proposition 2.4.4 of Chapter 2 that it is a transport plan between $\boldsymbol{P}_\xi^N$ and $\widehat{P}_\xi$. Using again the Gluing Lemma between $\widehat{\pi}$ and π in (3.8.2), it follows exactly as in the proof of part (i) that

$$\begin{aligned}
&\Big(\int_{\Xi\times\Xi} \|\eta_k - \xi_k\|^p d\pi'(\eta,\xi)\Big)^{\frac{1}{p}} \\
&\qquad\leq \Big(\int_{\Xi\times\Xi} \|\xi_k - \eta_k\|^p d\pi(\xi,\eta)\Big)^{\frac{1}{p}} + \Big(\int_{\Xi\times\Xi} \|\eta_k - \zeta_k\|^p d\widehat{\pi}(\eta,\zeta)\Big)^{\frac{1}{p}} \leq \varepsilon_k + \epsilon_k,
\end{aligned}$$

where π' is again a transport plan between $\widehat{P}_\xi$ and P_ξ. Thus, we deduce from (3.3.5) that (3.8.1) also holds with $\boldsymbol{\epsilon}$ as in the statement of part (ii), which concludes the proof. □

3.8.1. STATISTICAL GUARANTEES

Proposition 3.8.1 clarifies how much we need to inflate the hyperrectangle to retain the probabilistic guarantees of containing the data-generating distribution. In addition to clustering the product empirical distribution into $\widehat{P}_\xi$, Lloyd's algorithm also yields a transport plan that associates each atom of $\boldsymbol{P}_\xi^N$ to its cluster, enabling the direct computation of $\boldsymbol{\epsilon}$ in case (i). Analogously, in case (ii), the Wasserstein distances ϵ_k between $P_{\xi_k}^N$ and $\widehat{P}_{\xi_k}$ can be easily computed by solving a linear program [182, section 3.1], or, when $\widehat{P}_{\xi_k}$ is not specified beforehand, directly deduced using Lloyd's algorithm. Since the motivation to introduce MTHs for data-driven problems is their faster shrinkage compared to Wasserstein balls, in Appendix 3.B, we also establish the degree to which such desirable shrinkage rates are retained for clustered reference distributions.

We briefly elaborate on how the same result can be utilized to construct ambiguity sets with statistical guarantees of containing the data-generating distribution when the discrepancy between the true distribution and the product of its marginals is quantified using optimal transport. In particular, invoking recent work on optimal-transport-based measures of dependence [183], we may assume that P_ξ is ϵ-far from being independent for some $\epsilon > 0$, in the sense that

$$W_p(P_\xi, P_{\xi_1} \otimes \cdots \otimes P_{\xi_n}) \leq \epsilon,$$

where $P_{\xi_1}, \ldots, P_{\xi_n}$ denote the marginals of P_ξ. Suppose we have access to N independent samples $\xi^1, \ldots, \xi^N$ of ξ. Following a similar reasoning as in Chapter 2, we can first determine radii ε_k, $k \in [n]$ such that $W_p(P_{\xi_k}^N, P_{\xi_k}) \leq \varepsilon_k$ with probability $1-\beta_k$ for each $k \in [n]$. Invoking a union bound argument, we can assert that

$$\mathbb{P}(P_{\xi_1} \otimes \cdots \otimes P_{\xi_n} \in \mathcal{T}_p(P_{\xi_1}^N \otimes \cdots \otimes P_{\xi_n}^N, \boldsymbol{\varepsilon})) \geq 1 - \beta,$$

where $\beta = \beta_1 + \cdots + \beta_n$ (note that $1 - \beta \approx \prod_{k=1}^n (1-\beta_k)$ for small β_k, yielding guarantees analogous to Corollary 2.4.5 of Chapter 2. Combining this bound with the assumed distance from independence and Proposition 3.8.1(i), we conclude that $P_\xi \in \mathcal{T}_p(P_{\xi_1}^N \otimes \cdots \otimes P_{\xi_n}^N, \boldsymbol{\varepsilon} + \boldsymbol{\epsilon})$ with probability at least $1-\beta$. This yields the desired statistical guarantee for the dependent case.

3.9. SIMULATION EXAMPLES

In this section, we provide two real-world examples that show the superiority of structured ambiguity sets centered at product distributions or clustered product distributions over monolithic ambiguity balls, validating the theoretical results. In particular, we compare the performance of traditional Wasserstein balls and hyperrectangles, when the latter are centered either at the product empirical distribution or its clustered version.

3.9.1. OPTIMAL POWER DISPATCH

In this section, we provide a real-world example that shows the superiority of MTHs centered at clustered product distribution over monolithic ambiguity balls,

validating the theoretical results. In particular, we compare the performance of traditional Wasserstein balls and MTHs, when the latter are centered either at the product empirical distribution, or its clustered version. Moreover, we also conduct a comparison with a MTH centered at a resampled version of the product empirical distribution in order to demonstrate that clustering is more effective than the resampling strategy.

PROBLEM FORMULATION

We consider a power dispatch problem, where a daily power demand $d+\xi_2$ needs to be covered. The demand consists of a nominal term d, and a stochastic fluctuation ξ_2 representing the gap between the nominal and actual demand. To meet the demand, we exploit the power ξ_1 from a renewable energy resource, which is random, as it depends on the daily weather conditions. Since both ξ_1 and ξ_2 are random, there is no guarantee that the daily power demand $d+\xi_2$ can be met the renewable resource ξ_1. For this reason, we want to ensure that the demand is covered with probability at least $1-\alpha$ by ordering the minimum extra amount x of power from the market (cf. Figure 3.2).

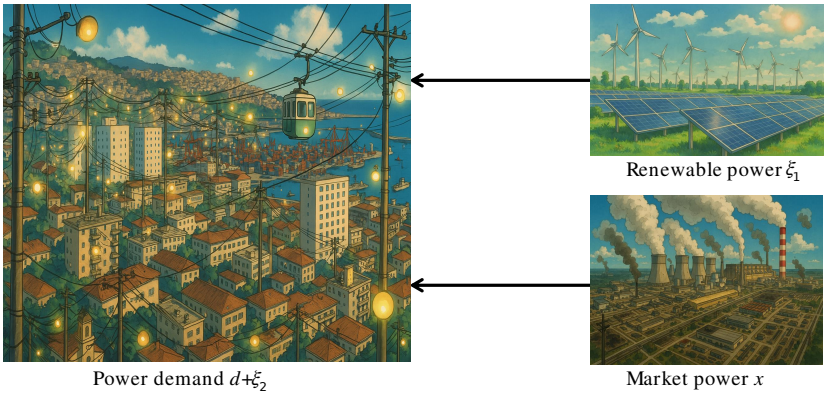


Figure 3.2.: Illustration of the considered power dispatch problem. The goal is to cover the power demand $d+\xi_2$ with probability at least $1-\alpha$, using the renewable power ξ_1 and the smaller possible amount x of power from the market.

This yields the chance-constrained problem

$$\begin{aligned} &\min_{x\geq 0} x \\ &\text{s.t. } \mathrm{CVaR}^{\mathbb{P}}_{1-\alpha}(d+\xi_2-\xi_1-x)\leq 0, \end{aligned} \tag{3.9.1}$$

where the CVaR constraint ensures the satisfaction of the chance constraint $\mathbb{P}(d+\xi_2-\xi_1-x\leq 0)\geq 1-\alpha$ as in Section 3.7. We assume that the distributions of ξ_2 and ξ_1 are unknown and we only have access to N i.i.d. samples $\xi^1,\ldots,\xi^N$ of $\xi:=(\xi_1,\xi_2)$. Thus, instead of (3.9.1) we solve the distributionally robust problem

$$\begin{aligned} &\min_{x\geq 0} x \\ &\text{s.t. } \sup_{P\in\mathscr{P}^N} \mathrm{CVaR}^P_{1-\alpha}(d+\xi_2-\xi_1-x)\leq 0, \end{aligned} \tag{3.9.2}$$

for three distinct data-driven ambiguity sets $\mathscr{P}^N$. In particular, we consider the cases where $\mathscr{P}^N$ is the Wasserstein ball $\mathscr{B}\equiv\mathscr{B}_1(P^N_\xi,\varepsilon)$ with $P^N_\xi := \frac{1}{N}\sum_{i=1}^N \delta_{\xi^i}$, the MTH $\mathscr{T}\equiv\mathscr{T}_1(\boldsymbol{P}^N_\xi,\boldsymbol{\varepsilon})$ with $\boldsymbol{P}^N_\xi = P^N_{\xi_1}\otimes P^N_{\xi_2}$, the MTH $\mathscr{T}_{\mathrm{cl}}\equiv\mathscr{T}_1(\widehat{P}_\xi,\boldsymbol{\varepsilon})$, where $\widehat{P}_\xi$ is a product distribution whose marginals are obtained by clustering each marginal of $\boldsymbol{P}^N_\xi$, and $\mathscr{T}_{\mathrm{resampl}}\equiv\mathscr{T}_1(\widehat{P}_{\mathrm{resampl}},\boldsymbol{\varepsilon})$, where $\widehat{P}_{\mathrm{resampl}}$ is is discrete distribution obtained by randomly sampling $\boldsymbol{P}^N_\xi$.

Let $\Xi := \{\xi\in\mathbb{R}^2 : C\xi\leq h\}$ be the support of P_ξ. We also assume $\mathbb{R}^2$ equipped with the 1-norm and denote $c := (-1,1)$. Using Corollary 3.7.5, we reformulate (3.9.2) as the linear program

$$\begin{aligned} \min_{\substack{x\geq 0,\boldsymbol{\lambda},\tau\\ s_l,\eta_l,l\in[M]}}\ & x \\ \text{s.t. }\ & \langle\boldsymbol{\lambda},\boldsymbol{\varepsilon}\rangle + \sum_{l=1}^M \vartheta_l s_l \leq \tau\alpha \\ & d - x + \tau + \langle c,\xi^l\rangle + \langle\eta_l, h - C\xi^l\rangle \leq s_l \\ & \left\|\mathrm{pr}_k^{(1,1)}(c - C^\top\eta_l)\right\|_\infty \leq \lambda_k \\ & s_l\geq 0,\ \eta_l\geq 0,\ l\in[M],\ k\in[2], \end{aligned}$$

where $\boldsymbol{\lambda} := (\lambda_1,\lambda_2)$. The solution of (3.9.2) is random due to the samples we use to construct the ambiguity set in each case. Depending on the ambiguity set we denote the solution by $\widehat{x}_{\mathscr{B}}$, $\widehat{x}_{\mathscr{T}}$, $\widehat{x}_{\mathscr{T}_{\mathrm{cl}}}$, and $\widehat{x}_{\mathscr{T}_{\mathrm{resampl}}}$ respectively.

SIMULATION RESULTS

For the simulations, we choose the distribution $P_\xi := P_{\xi_1}\otimes P_{\xi_2}$ with

$$\begin{aligned} P_{\xi_1} &:= 0.4\mathscr{U}\big([11,\,16]\big) + 0.6\mathscr{U}\big([24,\,27]\big) \\ P_{\xi_2} &:= 0.6\mathscr{U}\big([3,\,6]\big) + 0.4\mathscr{U}\big([10,\,11]\big), \end{aligned}$$

where $\mathscr{U}$ denotes the uniform distribution on the designated set. We also select the parameters $d = 4.5$ and $\alpha = 0.2$ and exploit $N = 20$ samples to build the empirical and product empirical distributions. We cluster the latter using $M = 9\times 8 = 72$ atoms, by clustering $P^N_{\xi_1}$ and $P^N_{\xi_2}$ into 9 and 8 atoms, respectively. Aditionally, a discrete distribution with an equal number of atoms to the clustered distribution is built using 72 samples drawn from $\boldsymbol{P}^N_\xi$. Next, we tune the size of the four ambiguity sets so that

$$\sup_{P\in\mathscr{P}^N} \mathrm{CVaR}^P_{1-\alpha}(d+\xi_2-\xi_1-x)\leq 0$$

with confidence at least 0.9 for each. To this end, we generate a large number of realizations of $N = 20$ data samples and construct the respective centers P_ξ^N, $\boldsymbol{P}_\xi^N$, $\widehat{P}_\xi$, and $\widehat{P}_{\text{resampl}}$. We solve the optimization problem over different ε values for the radius of the corresponding Wasserstein ball and the smallest ball enclosing the MTHs, and determine the constraint satisfaction frequencies of the associated decisions for the true distribution (cf. Figure 3.3). Next, we determine for each ambiguity set the smallest value of ε for which the CVaR constraint is met with the desired 0.9 confidence margin and solve all four instances of the problem for a large number of data realizations. Figure 3.4 shows the cumulative distribution of these solutions.

From Figure 3.3, it is clear that for each ε, the decisions of the rectangles ensure constraint satisfaction for a significantly higher confidence level than the monolithic ball. In particular, to achieve the desired level of 0.9 confidence, the monolithic ambiguity ball needs to have a radius $\varepsilon_{\min}(\mathscr{B}) = 0.65$ that is larger than $\varepsilon_{\min}(\mathscr{T}_{\text{resampl}}) = 0.625$, and much larger than the radii $\varepsilon_{\min}(\mathscr{T}) = 0.4875$ and $\varepsilon_{\min}(\mathscr{T}_{\text{cl}}) = 0.5$ of the balls enclosing the two rectangles. We also note that $\mathscr{T}_{\text{cl}}$ has almost similar confidence levels as $\mathscr{T}$ despite the smaller complexity of its clustered center (72 instead of 400 atoms), and a much larger confidence than $\mathscr{T}_{\text{resampl}}$ while having the exact same complexity as the latter. Furthermore, It only needs to be slightly larger than $\mathscr{T}$ to achieve the same guarantees.

In Figure 3.4, we plot the cumulative distributions of the solutions $\widehat{x}_{\mathscr{B}}$, $\widehat{x}_{\mathscr{T}_{\text{cl}}}$, $\widehat{x}_{\mathscr{T}}$, $\widehat{x}_{\mathscr{T}_{\text{resampl}}}$ and their expectations using colored vertical lines. First, we notice that the distribution mass accumulates faster for the solutions resulting from DRO problems associated with MTHs compared to the solution resulting from the DRO problem associated with the monolithic ambiguity ball. This means that the DRO problems associated with MTHs have solutions that are significantly less conservative while still guaranteeing the satisfaction of the CVaR constraint with the desired confidence level. Similarly to our previous observation, we notice that clustering leads to improved performance compared to resampling in our simulation example, thereby suggesting that it is an effective strategy to mitigate the complexity of the product empirical measure.

3.9.2. UNCERTAINTY QUANTIFICATION: COOPERATIVE SEARCH AND RESCUE MISSION ⋆

In this section, we solve an uncertainty quantification problem to illustrate the properties of multi-transport and Wasserstein hyperrectangles, where the latter are introduced in Chapter 2, and compare their performances. We consider two drones that must reach a region within a specific deadline to perform a search-and-rescue task. The probability that these drones succeed in reaching the region before the deadline determines whether a fallback plan for the mission has to be used or not.

The maximum velocity and the distance from the region are unknown and assumed to be random and independent across the drones. Only historical information about the distance and maximum velocity of the drones from previous deployments can be used. The goal is to build an ambiguity set from these data and determine a lower probability bound for two different scenarios. The first is that at least one drone

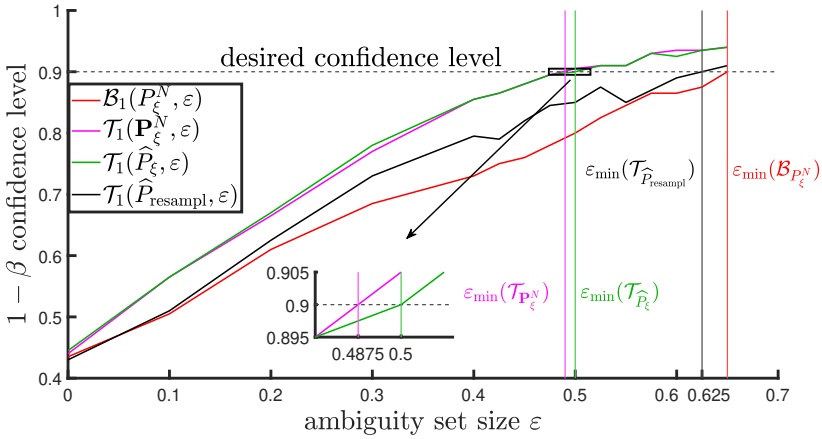


Figure 3.3.: Satisfaction probability of the CVaR constraint for each ambiguity set with respect to its radius ε for the parameters $d = 4.5$, $\alpha = 0.2$, $N = 20$ and $M = 72$. For each hyperrectangle, ε represents the radius of the smallest ball enclosing it. The minimum radii ensuring the desired 0.9 confidence level for each ambiguity set are indicated by the vertical lines. For every ε, MTHs satisfy the chance-constraint with confidence levels that are always greater than that of the corresponding ambiguity ball, which turns out to be much more conservative. We also observe that despite the clear complexity reduction of clustered distribution compared to the product empirical distribution, the confidence levels of their associated MTHs are very close for every ε. In addition, we can also clearly notice that the resampled distribution, on the other hand, yields decisions with a much smaller confidence compared to its clustered counterpart, despite the fact that both are supported on the same number of atoms.

reaches the region and the other that both drones reach the region in due time. Namely, we want to determine the worst-case probability for each event to happen among all the distributions in the inferred ambiguity sets.

Let τ denote the deadline, while v_k and r_k denote the maximum velocity and distance of each drone from the target, see Figure 3.5. Each drone k reaches the region iff

$$r_k - \tau v_k < 0 \iff a_k \xi < 0,\ k = 1,2,$$

where $a_1 = \begin{bmatrix} 1 & -\tau & 0 & 0 \end{bmatrix}$, $a_2 = \begin{bmatrix} 0 & 0 & 1 & -\tau \end{bmatrix}$, and $\xi = (r_1, v_1, r_2, v_2)^\top$ represents the random variable of our problem. Denoting $\mathscr{R}^k$ the event that drone k reaches the target within the deadline, we get

$$\mathscr{R}^k = \{\xi \in \mathbb{R}^4 : a_k \xi < 0\},\ k = 1,2.$$

Since the event that at least one drone reaches the region before the deadline is described by the set $\mathscr{E}_1 := \mathscr{R}^1 \cup \mathscr{R}^2$, we aim to determine the worst-case probability

$$\min_{P \in \mathscr{P}^N} P[\xi \in \mathscr{E}_1] = 1 - \max_{P \in \mathscr{P}^N} P[\xi \notin \mathscr{E}_1]. \tag{3.9.3}$$

where $\mathscr{P}^N$ is an ambiguity set that we infer from N i.i.d. samples of ξ. Since $\max_{P \in \mathscr{P}^N} P[\xi \notin \mathscr{E}_1] = \max_{P \in \mathscr{P}^N} \mathbb{E}[\mathbb{1}_{(\mathscr{E}_1)^c}(\xi)]$, the objective function $\mathbb{1}_{(\mathscr{E}_1)^c}(\xi)$ does not

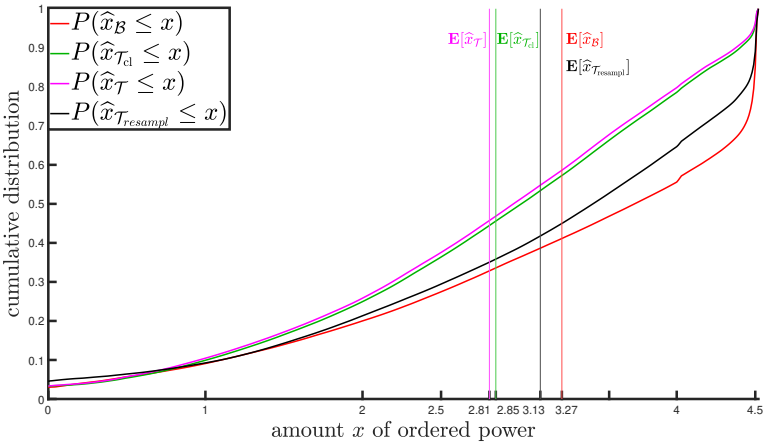


Figure 3.4.: Cumulative distributions of $\hat{x}_{\mathcal{B}}$ in red, $\hat{x}_{\mathcal{T}_{\text{cl}}}$ in green, $\hat{x}_{\mathcal{T}}$ in magenta, and $\hat{x}_{\mathcal{T}_{\text{resampl}}}$ in black for the smallest size of each ambiguity set that ensures the satisfaction of the CVaR constraint with a confidence level of 0.9 for the parameters $d = 4.5$, $\alpha = 0.2$, $N = 20$ and $M = 72$. The mean values of the distributionally robust solutions $\hat{x}_{\mathcal{B}}$ in red, $\hat{x}_{\mathcal{T}_{\text{cl}}}$ in green, $\hat{x}_{\mathcal{T}}$ in magenta, and $x_{\mathcal{T}_{\text{resampl}}}$ in black are depicted with vertical lines. The obtained DRO values with the MTHs are significantly smaller than the ones with the traditional ambiguity balls. Further, the average solution when the MTH is centered at the clustered distribution is very close to the one where the MTH is centered at the product empirical distribution, despite the considerable difference in the number of atoms of the two centers. At the same time, the average solution when the MTH is centered at the resampled distribution is significantly less efficient than the solutions obtained its clustered counterpart, despite the fact that they are supported at the same number of atoms.

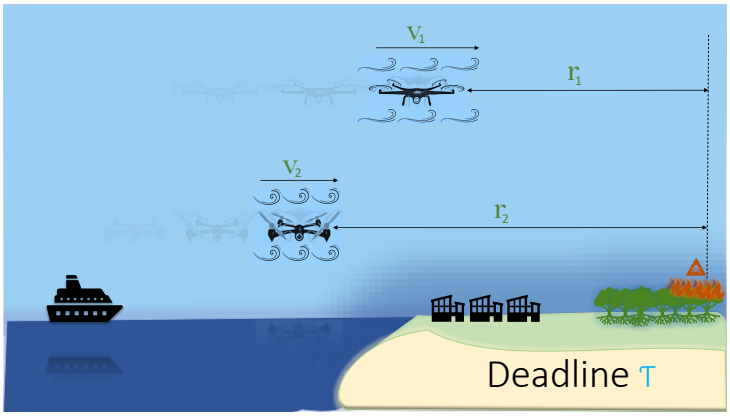


Figure 3.5.: A pictorial representation of the considered uncertainty quantification problem. The drones 1 and 2 must reach, at every emergency occurrence, the random location of the emergency, within the specific time deadline τ. Each drone must cross, at every occurrence, a random distance $r_{1,2}$ with a random velocity $v_{1,2}$.

have the decoupled structure of Proposition 2.6.2 and we cannot evaluate (3.9.3) using Wasserstein hyperrectanles. Nevertheless, since $\mathscr{E}_1$ is the union of two open polytopes, we can exploit Corollary 3.6.2 to evaluate the worst-case probability (3.9.3)

over a multi-transport hyperrectangle, which we also compare to a Wasserstein ball.

Analogously, the event that both drones reach the region in due time is described by the set $\mathscr{E}_2 := \mathscr{R}^1 \cap \mathscr{R}^2$. Then, we aim to determine the worst-case probability

$$\begin{aligned} \min_{P\in\mathscr{P}^N} P[\xi \in \mathscr{E}_2] &= 1 - \max_{P\in\mathscr{P}^N} P[\xi \notin \mathscr{E}_2], \\ &= 1 - \max_{P\in\mathscr{P}^N} P[\xi \in \mathscr{E}_2{}^c], \end{aligned} \tag{3.9.4}$$

where $(\mathscr{E}_2)^c = (\mathscr{R}^1)^c \cup (\mathscr{R}^2)^c$. The set $(\mathscr{E}_2)^c$ is the union of two closed polytopes. Thus, we can evaluate (3.9.4) with $\mathscr{P}^N = \mathscr{B}_1(P_\xi^N, \varepsilon)$ or $\mathscr{P}^N = \mathscr{T}_1(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})$ by exploiting program (3.6.2) in Theorem 3.6.1. In the case where $\mathscr{P}^N = \mathscr{H}_1(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})$, similarly to [164, equation (17)] we have

$$\min_{P\in\mathscr{H}_1(\boldsymbol{P}_\xi^N,\boldsymbol{\varepsilon})} P[\xi \in \mathscr{E}_2] = \prod_{k=1}^{2} \Big(1 - \max_{P_k\in\mathscr{B}_1(P_{\xi_k}^N,\varepsilon_k)} P_k(\mathscr{R}^k)^c)\Big). \tag{3.9.5}$$

Consequently, solving the quantification problem using the Wasserstein hyperrectangle becomes tractable, since it can get reformulated as (3.9.5) which is composed of two robust uncertainty quantification problems over closed polytopic sets using Wasserstein balls.

For the simulations, the initial distances (in km) of the drones 1-2 follow the distributions $0.5\mathscr{U}[6,10] + 0.5\mathscr{U}[10.1,11.1]$ and $0.95\mathscr{U}[9,10] + 0.05\mathscr{U}[10.1,11.1]$, respectively, where $\mathscr{U}$ denotes the uniform distribution. All velocities (in m/sec) follow the distribution $\mathscr{U}[50,50.5]$ and the deadline is set to $\tau = 200$sec. The exact supports of the distributions are assumed known yet not the distributions themselves. Using this information, and to avoid the conservativeness that might be posed by the confidence bounds of Proposition 2.5.1, we select a radius for the monolithic ball and the relative size of the hyperrectangles such that the radius ε_k of each hyperrectangle component satisfies $\varepsilon_k \leq c\frac{\rho_k}{\rho} N^{-\frac{1}{2}+\frac{1}{4}}\varepsilon$. (see [164] for more details).

Figure 3.7 shows the solution of problem (3.9.3) across 30 realizations of the simulation that leverage 18 samples each. The multi-transport rectangle exhibits superior performance compared to the Wasserstein ball since the worst-case values are above the probability threshold (set at 0.8) in 90% of the realizations in the former case compared to 16.67% in the latter. The true probability of at least one of the drones reaching the target equals 0.975, which implies that the bounds from the multi-transport hyperrectangle are much closer to the true probability in relative terms compared to those of the ball. Figure 3.6 shows the solution of (3.9.4) across 30 realizations of the simulation that leverage 18 samples each. This figure allows us to compare the worst-case probabilities across all three ambiguity sets. As expected, the results from both hyperrectangles outperform those from the ambiguity ball in every case. In addition, although the multi-transport hyperrectangle is more conservative than the Wasserstein one, it leads to an improvement of a similar scale, thus supporting that in general, it exhibits a convenient compromise between tractability and conservativeness-reduction.

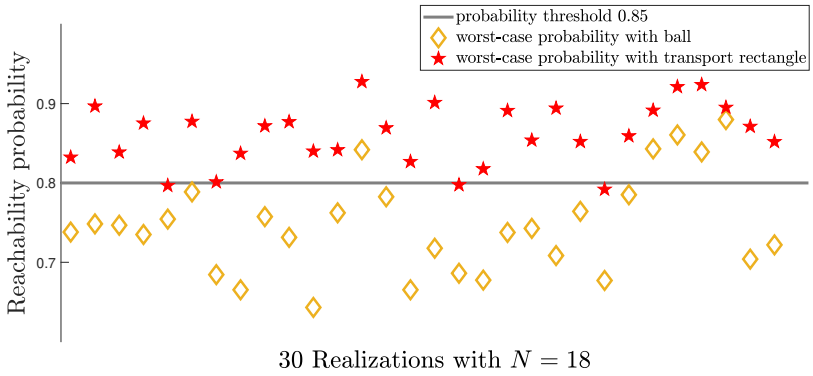


Figure 3.6.: The figure shows the worst-case probabilities of at least one drone reaching the target across 30 realizations. The results obtained by using the monolithic ball are depicted by the diamonds while those obtained from the multi-transport hyperrectangle are depicted by the red stars. In all cases, the ambiguity sets are built using 18 samples. The results obtained by the hyperrectangle show a clear decrease of conservativenness. Moreover, they are most often above the minimum probability threshold while it is rarely the case for those from the monolithic ball.

3.10. CONCLUSION

In this chapter, we derived tractable reformulations for several classes of distributionally robust stochastic optimization problems where distributional uncertainty is captured through ambiguity sets formed by multiple optimal transport constraints. The sets, called multi-transport hyperrectangles (MTHs) share probabilistic guarantees of containing the unknown distribution, which scales conveniently with the number of samples when the uncertainty consists of independent low-dimensional components. For the reformulations, we also established basic properties of MTHs, including weak compactness and conditions ensuring finiteness of their associated DRO values.

We derived reformulations for objective functions whose dependence on the

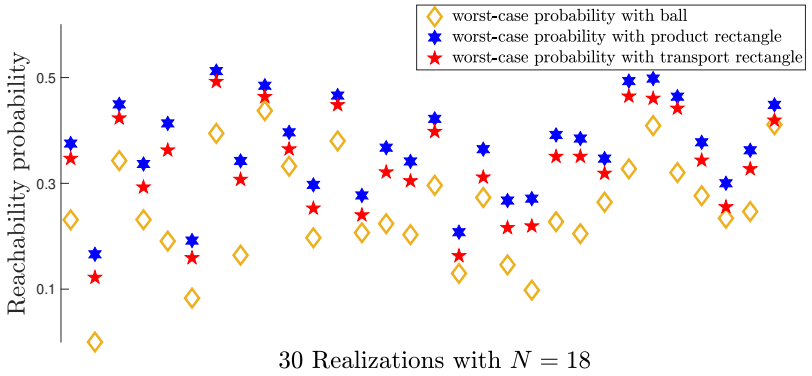


Figure 3.7.: The figure shows the lower probability bound that the target is reached by both drones across 30 realizations. The results obtained by using the monolithic ball are depicted by the diamonds and the results with the product measure hyperrectangle by the blue stars while the results with the transport rectangle are depicted by the red stars. In all cases, the ambiguity sets are built using 18 samples. The results obtained by both hyperrectangles outperform those obtained by the monolithic ball.

uncertainty is quadratic or expressed as the maximum of finitely many concave functions, including piecewise affine ones. Using the results for the piecewise affine class, we solved uncertainty quantification problems over unions of polyhedral sets. We also derived tractable reformulations for distributionally robust chance-constrained problems over MTHs. To mitigate the complexity of the reformulations for data-driven problems, we clustered the reference distribution of the MTHs. The numerical simulations illustrate the benefits of MTHs compared to traditional Wasserstein balls when the uncertainty consists of independent components. Clustering the center of the MTH yielded results of comparable performance under a noticeable complexity reduction.

Future work will exploit the tools developed in this chapter to solve distributionally robust control problems with independent uncertainty components and test them on real-world data sets. These include stochastic model predictive control algorithms where uncertainty is independent across stages and the control of multi-agent systems where agents are subject to independent disturbances.

Appendix

3.A. TECHNICAL PROOFS

3.A.1. PROOFS FROM SECTION 3.4

In this part, we provide the proofs of some statements of Section 3.4. To prove Theorem 3.4.1, we also introduce some further preliminaries and results from probability theory. A collection $\mathscr{S} \subset \mathscr{P}(\Xi)$ of probability distributions is called tight if, for any $\varepsilon > 0$, there exists a compact subset $B \subset \Xi$ such that $P(\Xi \setminus B) \leq \varepsilon$, for all $P \in \mathscr{S}$. The following theorem of Prokhorov provides a link between weak compactness and tightness.

Theorem 3.A.1. *(Prokhorov's theorem). A collection $\mathscr{S} \subset \mathscr{P}(\Xi)$ of probability measures is tight if and only if the weak closure of $\mathscr{S}$ is weakly compact in $\mathscr{P}(\Xi)$.*

We will exploit the following result from [117] that establishes the lower semicontinuity of the transport-cost functional. We adopt it here for positive transport costs.

Lemma 3.A.2. *(Lower semicontinuity of the cost functional [117, Lemma 4.3]). Let Ξ_1 and Ξ_2 be two Polish spaces and $c : \Xi_1 \times \Xi_2 \to [0, +\infty]$ be a lower semicontinuous cost function. Then $F : \pi \to \int_{\Xi_1 \times \Xi_2} c d\pi$ is lower semicontinuous on $\mathscr{P}(\Xi_1 \times \Xi_2)$, equipped with the topology of weak convergence.*

We also need the next result, which certifies the tightness of the set of transport plans between tight sets of probability distributions.

Lemma 3.A.3. *(Tightness of couplings [117, Lemma 4.4]). Let Ξ_1 and Ξ_2 be two Polish spaces and let $\mathscr{R}_1$ and $\mathscr{R}_2$ be tight subsets of $\mathscr{P}(\Xi_1)$ and $\mathscr{P}(\Xi_2)$, respectively. Then the set $\mathscr{C}(\mathscr{R}_1, \mathscr{R}_2)$ of all couplings with marginals in $\mathscr{R}_1$ and $\mathscr{R}_2$, respectively, is itself tight in $\mathscr{P}(\Xi_1 \times \Xi_2)$.*

The next result ensures that properness is maintained on the product of proper spaces under equivalence of any product metric with the considered metric on the product space.

Lemma 3.A.4. *(Product of proper metric spaces). Consider the proper spaces (Ξ_k, ρ_k), $k \in [n]$, and let Assumption 5 hold. Then, the metric space (Ξ, ρ) is also proper.*

Proof. It suffices to show that the ball $B_\rho(\zeta, r) := \{\xi \in \Xi : \rho(\zeta, \xi) \leq r\}$ is compact for each $\zeta \in \Xi$ and $r > 0$. To this end, we exploit Assumption 5(ii), which asserts that there exists $C > 0$ such that $\rho_k(\zeta_k, \xi_k) \leq C\rho(\zeta, \xi)$ for all $\zeta, \xi \in \Xi$ and $k \in [n]$. This in turn implies that

$$B_\rho(\zeta, r) \subset B_{\rho_1}(\zeta_1, Cr) \times \ldots \times B_{\rho_n}(\zeta_n, Cr).$$

Since the sets $B_{\rho_k}(\zeta_k, Cr)$, $k \in [n]$ are compact by properness of each (Ξ_k, ρ_k), their cartesian product is also compact, which establishes the result. □

Proposition 3.A.5. *(Weak compactness of the set of transport plans). Let Assumption 5 hold and consider the set of transport plans*

$$\Pi_p(Q,\boldsymbol{\varepsilon}) := \Big\{\pi \in \mathscr{P}(\Xi\times\Xi) : \mathrm{pr}_{1\#}\pi = Q \text{ and } \int_{\Xi\times\Xi} \rho_k(\zeta_k,\xi_k)^p d\pi(\zeta,\xi) \le \varepsilon_k \text{ for all } k\in[n]\Big\}.$$

where $\boldsymbol{\varepsilon} := (\varepsilon_1,\ldots,\varepsilon_n) \ge 0$ *and* Q *has finite* p*th moment. Then* $\Pi_p(Q,\boldsymbol{\varepsilon})$ *is weakly compact.*

Proof. From Proposition 2.4.6 in Chapter 2 and Assumption 5(ii) we can enclose the MTH $\mathscr{T}_p(Q,\boldsymbol{\varepsilon})$ inside a Wassertein ball $\mathscr{B}_p(Q,\varepsilon)$ of a sufficiently large radius ε. Since (Ξ,ρ) is proper by Lemma 3.A.4 and the reference distribution Q has a finite pth moment, we get by [170, Theorem 1] that this enclosing Wasserstein ball is also weakly compact. From the reverse implication of Prokhorov's theorem (Theorem 3.A.1) this ball is tight, which implies that $\mathscr{T}_p(Q,\boldsymbol{\varepsilon})$ is also tight as it is contained in $\mathscr{B}_p(Q,\varepsilon)$. Recalling that $\Pi_p(Q,\boldsymbol{\varepsilon})$ is the set of transport plans between $\{Q\}$ and $\mathscr{T}_p(Q,\boldsymbol{\varepsilon})$, by virtue of Lemma 3.A.3, $\Pi_p(Q,\boldsymbol{\varepsilon})$ is also tight. Thus, to show that $\Pi_p(Q,\boldsymbol{\varepsilon})$ is weakly compact, it suffices from the direct implication of Prokhorov's theorem to show that it is weakly closed.

Since each ρ_k, $k\in[n]$ is a metric, its is nonnegative. Thus, we get from Lemma 3.A.2 with $\Xi_1\equiv\Xi_2\equiv\Xi$ that the sublevel sets $F_k^{-1}((-\infty,\varepsilon_k])$ of the functionals $F_k : \pi \to \int_{\Xi\times\Xi}\rho_k^p d\pi$, $k\in[n]$ are weakly closed. Since each of those is equal to the set

$$\Pi_{p,k}(Q,\varepsilon_k) := \Big\{\pi\in\mathscr{P}(\Xi\times\Xi) : \mathrm{pr}_{1\#}\pi = Q \text{ and } \int_{\Xi\times\Xi}\rho_k(\zeta_k,\xi_k)^p d\pi(\zeta,\xi)\le\varepsilon_k\Big\},$$

and taking into account that $\Pi_p(Q,\boldsymbol{\varepsilon}) = \cap_{k=1}^n \Pi_{p,k}(Q,\varepsilon_k)$, it follows that $\Pi_p(Q,\boldsymbol{\varepsilon})$ is weakly closed as well, which establishes the result. □

Now we have gathered all the necessary ingredients to prove weak compactness of $\mathscr{T}_p(Q,\boldsymbol{\varepsilon})$.

Proof of Theorem 3.4.1. First, recall that from Proposition 2.4.6 and Assumption 5(ii) we can enclose the MTH $\mathscr{T}_p(Q,\boldsymbol{\varepsilon})$ inside a Wassertein ball $\mathscr{B}_p(Q,\varepsilon)$ of a sufficiently large radius ε. Since (Ξ,ρ) is proper by Lemma 3.A.4 and the reference distribution Q has a finite pth moment, we get by [170, Theorem 1] that this enclosing ball is weakly compact, which implies tightness of $\mathscr{T}_p(Q,\boldsymbol{\varepsilon})$. Hence, in order to show the weak compactness of $\mathscr{T}_p(Q,\boldsymbol{\varepsilon})$, it is enough to show that $\mathscr{T}_p(Q,\boldsymbol{\varepsilon})$ is closed. To this end let $\{P_j\}_{j\in\mathbb{N}}\subset\mathscr{T}_p(Q,\boldsymbol{\varepsilon})$ be a sequence converging to some $P\in\mathscr{P}(\Xi)$. By the definition of $\mathscr{T}_p(Q,\boldsymbol{\varepsilon})$, this sequence consists of the second marginals of a sequence $\{\pi_j\}_{j\in\mathbb{N}}\subset\Pi_p(Q,\boldsymbol{\varepsilon})$, i.e., $P_j = \mathrm{pr}_{2\#}\pi_j$ for each j. Since $\Pi_p(Q,\boldsymbol{\varepsilon})$ is weakly compact by Proposition 3.A.5, we can extract a subsequence of $\{\pi_j\}_{j\in\mathbb{N}}$ that converges weakly to some $\pi\in\Pi_p(Q,\boldsymbol{\varepsilon})$ and denote $P' := \mathrm{pr}_{2\#}\pi$. With a slight abuse of notation, we also index this sequence by j. Therefore, for any continuous and bounded function g on Ξ, we have

$$\lim_{j\to\infty}\int_{\Xi\times\Xi} g(\xi)dP_j(\xi) = \lim_{j\to\infty}\int_{\Xi\times\Xi} g(\xi)d\mathrm{pr}_{2\#}\pi_j(\zeta,\xi) = \lim_{j\to\infty}\int_{\Xi\times\Xi} g\circ\mathrm{pr}_2(\zeta,\xi)d\pi_j(\zeta,\xi)$$

$$= \int_{\Xi\times\Xi} g\circ \mathrm{pr}_2(\zeta,\xi)\,d\pi(\zeta,\xi) = \int_{\Xi\times\Xi} g(\xi)\mathrm{pr}_{2\#}d\pi(\zeta,\xi) = \int_{\Xi\times\Xi} g(\xi)dP'(\xi).$$

By uniqueness of the weak limit, $P = P' \equiv \mathrm{pr}_{2\#}\pi$, and since $\pi \in \Pi_p(Q,\boldsymbol{\varepsilon})$, it follows by the definition of $\mathcal{T}_p(Q,\boldsymbol{\varepsilon})$ that P belongs to $\mathcal{T}_p(Q,\boldsymbol{\varepsilon})$, which establishes that the set is closed and concludes the proof. □

Proof of Theorem 3.4.3. The proof of part (i) is analogous to the proof of [170, Theorem 3]. From Assumption 5(ii) and Proposition 2.4.6 we can enclose $\mathcal{T}_p(Q,\boldsymbol{\varepsilon})$ inside a sufficiently large Wasserstein ball. Thus, it follows from [170, Lemma 1] that the MTH $\mathcal{T}_p(Q,\boldsymbol{\varepsilon})$ has a uniformly bounded pth moment. Together with our assumptions on h, this implies along the lines of the proof of [170, Theorem 3] that the map $P \to \mathbb{E}_P[h]$ is upper semicontinuous on $\mathcal{T}_p(Q,\boldsymbol{\varepsilon})$. Combining this with weak compactness of $\mathcal{T}_p(Q,\boldsymbol{\varepsilon})$ we establish that h attains its supremum.

To prove part (ii), let $C>0$ and $\zeta\in\Xi$ such that (3.2.1) holds for both h and $-h$, consider the ball $B_\rho(\zeta,R) := \{\xi\in\Xi\colon \rho(\zeta,\xi)\le R\}$ for some $R>0$, and let $\{P_k\}_{k\in\mathbb{N}}$ be a sequence in $\mathcal{T}_p(Q,\boldsymbol{\varepsilon})$ converging weakly to some $P_\star$. To prove the weak continuity of Ψ, we show that $\lim_{k\to\infty}\Psi(P_k) = \Psi(P_\star)$. Namely, we will show that for any $\epsilon>0$ there exists $k_0\in\mathbb{N}$ such that

$$|\Psi(P_k) - \Psi(P_\star)| < \epsilon \quad \text{for all } k \ge k_0. \tag{3.A.1}$$

To this end, we introduce a truncated version of h. We fix $R>0$ and define

$$h_R(\xi) := \min\{h^\star(\xi), C(1+R^r)\}.$$

where $h^\star(\xi) := \max\{h(\xi), -C(1+R^r)\}$. We then have that

$$|h(\xi) - h_R(\xi)| \le \begin{cases} C\rho(\zeta,\xi)^r, & \text{if } \rho(\zeta,\xi)\ge R \\ 0, & \text{otherwise.}\end{cases}$$

This bound implies that there exists $b>0$ such that

$$\begin{aligned}\Big|\int_\Xi h(\xi)dP(\xi) - \int_\Xi h_R(\xi)dP(\xi)\Big| &\le \int_\Xi |h(\xi)-h_R(\xi)|\,dP(\xi)\\ &\le C\int_{\Xi\setminus B_\rho(\zeta,R)} \rho(\zeta,\xi)^r dP(\xi) \le Cb/R^{(p-r)}\end{aligned} \tag{3.A.2}$$

for all $P\in\mathcal{T}_p(Q,\boldsymbol{\varepsilon})$. Indeed, in analogy to the proof of [170, Theorem 3], this follows from the fact that $\rho(\zeta,\xi)^r = \rho(\zeta,\xi)^p/\rho(\zeta,\xi)^{r-p} \le \rho(\zeta,\xi)^p/R^{r-p}$ over the domain of integration and [170, Lemma 1]. Then, we obtain from the triangle inequality that

$$|\Psi(P_k) - \Psi(P_\star)| = \Big|\int_\Xi h(\xi)dP_k(\xi) - \int_\Xi h(\xi)dP_\star(\xi)\Big|$$

If $\rho(\zeta,\xi)<R$, then $|h(\xi)|\le C(1+R^r)$ and thus $h(\xi)-h_R(\xi)=0$. If $\rho(\zeta,\xi)\ge R$, then either $h(\xi)\ge C(1+R^r)$, which implies $0\le h(\xi)-h_R(\xi)\le C\rho(\zeta,\xi)^r$ by (3.2.1), or $h(\xi)\le -C(1+R^r)$, which again by (3.2.1) implies $-C\rho(\zeta,\xi)^r \le h(\xi)-h_R(\xi)\le 0$, or $|h(\xi)|\le C(1+R^r)$, which implies $h(\xi)-h_R(\xi)=0$.

This lemma can be applied to $\mathcal{T}_p(Q,\boldsymbol{\varepsilon})$ since it is possible to enclose it inside a sufficiently large Wasserstein ball that is centered at Q Proposition 2.4.6.

$$\leq \left| \int_{\Xi} h_R(\xi) d(P_k - P_\star)(\xi) \right| + \left| \int_{\Xi} (h(\xi) - h_R(\xi)) dP_k(\xi) \right| + \left| \int_{\Xi} (h(\xi) - h_R(\xi)) dP_\star(\xi) \right|$$

for all k. The last two terms can be rendered arbitrarily small by (3.A.2) for a large enough R. Since h is continuous, which implies that h_R is as well, and P_k converges weakly to $P_\star$, we can also make the first term as small as desired for large enough k. This establishes (3.A.1) and concludes the proof. □

3.A.2. PROOFS FROM SECTION 3.5

Here we gather all proofs from Section 3.5.

Proof of Proposition 3.5.1. Since h is real-valued by Assumption 7, it is also bounded on the atoms of Q. Thus, it satisfies the conditions of Theorem 2.6.4 of the previous chapter, which yields

$$\begin{aligned} \sup_{P_\xi \in \mathcal{T}_1(Q,\boldsymbol{\varepsilon})} \mathbb{E}_{P_\xi}[h(\xi)] &= \inf_{\boldsymbol{\lambda} \geq 0} \left\{ \langle \boldsymbol{\lambda}, \boldsymbol{\varepsilon} \rangle + \sum_{l \in [M]} \vartheta_l \sup_{\xi \in \Xi} \left\{ h(\xi) - \sum_{k=1}^{n} \lambda_k \| \xi_k^l - \xi_k \| \right\} \right\} \\ &= \inf_{\boldsymbol{\lambda} \geq 0} \left\{ \langle \boldsymbol{\lambda}, \boldsymbol{\varepsilon} \rangle + \sum_{l \in [M]} \vartheta_l \max_{j \in [m]} \sup_{\xi \in \Xi} \left\{ h_j(\xi) - \sum_{k=1}^{n} \lambda_k \| \xi_k^l - \xi_k \| \right\} \right\}, \end{aligned}$$

where we used the fact that $h(\xi) = \max_{j \in [m]} h_j(\xi)$ in the last equality. Introducing epigraphical variables s_l, $l \in [M]$, the problem is equivalent to

$$\begin{aligned} &\inf_{\boldsymbol{\lambda}, s_l} \langle \boldsymbol{\lambda}, \boldsymbol{\varepsilon} \rangle + \sum_{l=1}^{M} \vartheta_l s_l \\ &\text{s.t.} \sup_{\xi \in \Xi} \left\{ h_j(\xi) - \sum_{k=1}^{n} \lambda_k \| \xi_k^l - \xi_k \| \right\} \leq s_l \quad l \in [M],\ j \in [m] \\ &\quad \lambda_k \geq 0 \qquad\qquad\qquad\qquad\qquad\quad k \in [n]. \end{aligned} \tag{3.A.3}$$

By adopting the arguments from [79, cf. (12d) & (12e)] to the vectorial with respect to $\boldsymbol{\lambda}$ form of the constraints in (3.A.3), and considering for each $l \in [M]$ and $j \in [m]$ the dual variable $z_{lj} \in \mathbb{R}^d$, we get that

$$\begin{aligned} &\sup_{\xi \in \Xi} \left\{ h_j(\xi) - \sum_{k=1}^{n} \max_{\| \mathrm{pr}_k^{\boldsymbol{d}}(z_{lj}) \|_* \leq \lambda_k} \langle \mathrm{pr}_k^{\boldsymbol{d}}(z_{lj}), \xi_k - \xi_k^l \rangle \right\} \\ &\quad = \sup_{\xi \in \Xi} \min_{\| \mathrm{pr}_k^{\boldsymbol{d}}(z_{lj}) \|_* \leq \lambda_k, k \in [n]} \left\{ h_j(\xi) - \sum_{k=1}^{n} \langle \mathrm{pr}_k^{\boldsymbol{d}}(z_{lj}), \xi_k - \xi_k^l \rangle \right\} \\ &\quad = \min_{\| \mathrm{pr}_k^{\boldsymbol{d}}(z_{lj}) \|_* \leq \lambda_k, k \in [n]} \sup_{\xi \in \Xi} \left\{ h_j(\xi) - \sum_{k=1}^{n} \langle \mathrm{pr}_k^{\boldsymbol{d}}(z_{lj}), \xi_k - \xi_k^l \rangle \right\}. \end{aligned}$$

The last equality follows by compactness of $\{z_{lj} : \|\mathrm{pr}_k^{\boldsymbol{d}}(z_{lj})\|_* \leq \lambda_k, k \in [n]\}$ and the minimax theorem [156, Proposition 5.5.4].Then each of the first set of constraints in (3.A.3) is equivalent to

$$\begin{cases} \sup\limits_{\xi\in\Xi}\{h_j(\xi) - \langle z_{lj}, \xi - \xi^l\rangle\} \leq s_l \\ \|\mathrm{pr}_k^{\boldsymbol{d}}(z_{lj})\|_* \leq \lambda_k, \quad k \in [n], \end{cases} \iff \begin{cases} \sup\limits_{\xi\in\Xi}\{h_j(\xi) + \langle z_{lj}, \xi\rangle\} - \langle z_{lj}, \xi^l\rangle \leq s_l \\ \|\mathrm{pr}_k^{\boldsymbol{d}}(z_{lj})\|_* \leq \lambda_k, \quad k \in [n]. \end{cases} \tag{3.A.4}$$

where we substituted z_{lj} by $-z_{lj}$ in the right-hand side of the equivalence. In addition, from Assumption 7, which establishes that $-h_j(\xi)$ is proper, convex, and lower semi-continuous, we obtain by using the exact same arguments as in [79] that

$$\sup_{\xi\in\Xi}\{h_j(\xi) + \langle z_{lj}, \xi\rangle\} = \mathrm{cl}\big(\inf_{v_{lj}}\{\sigma_\Xi(v_{lj}) + [-h_j]^*(z_{lj} - v_{lj})\}\big), \tag{3.A.5}$$

where $\mathrm{cl}(f)(x) := \liminf_{z\to x} f(z)$ (cf. [184, Page 14]). Therefore, by taking further into account (3.A.3)–(3.A.5), we obtain the reformulation in the statement of the proposition. □

Proof of Proposition 3.5.2. The proof relies on Proposition 3.5.1 and the exact same reasoning as in the proof of [79, Corollary 5.1]. □

Proof of Proposition 3.5.3. By Theorem 2.6.4 of Chapter 2 and Assumption 6, (3.4.1) is equal to the value of the dual problem

$$\begin{aligned} &\inf_{\boldsymbol{\lambda}\geq 0}\langle \boldsymbol{\lambda}, \boldsymbol{\epsilon}\rangle + \sum_{l=1}^M \vartheta_l \sup_{\xi\in\Xi}\Big\{\xi^\top \mathscr{Q}\xi + 2q^\top\xi - \sum_{k=1}^n \lambda_k \|\xi_k^l - \xi_k\|^2\Big\} \\ &\quad = \inf_{\boldsymbol{\lambda}\geq 0}\langle \boldsymbol{\lambda}, \boldsymbol{\epsilon}\rangle + \sum_{l=1}^M \vartheta_l \sup_{\xi\in\Xi}\{\xi^\top \mathscr{Q}\xi + 2q^\top\xi - (\xi^l - \xi)^\top \mathrm{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda})(\xi^l - \xi)\} \\ &\quad = \inf_{\boldsymbol{\lambda}\geq 0}\langle \boldsymbol{\lambda}, \boldsymbol{\epsilon}\rangle + \sum_{l=1}^M \vartheta_l \sup_{\xi\in\Xi}\{\xi^\top (\mathscr{Q} - \mathrm{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda}))\xi + 2(q + \mathrm{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda})\xi^l)^\top\xi - {\xi^l}^\top \mathrm{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda})\xi^l\}. \end{aligned}$$

Since $\Xi \equiv \mathbb{R}^d$, we have that

$$\sup_{\xi\in\mathbb{R}^d}\{\xi^\top (\mathscr{Q} - \mathrm{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda}))\xi + 2(q + \mathrm{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda})\xi^l)^\top\xi - {\xi^l}^\top \mathrm{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda})\xi^l\} = +\infty$$

if either $\mathscr{Q} - \mathrm{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda}) \not\preceq 0$ or $q + \mathrm{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda})\xi^l \notin \mathrm{range}(\mathscr{Q} - \mathrm{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda}))$. Taking this into account and introducing epigraphical variables, we can rewrite the dual problem as

$$\begin{aligned} &\inf_{\boldsymbol{\lambda}\geq 0}\ \langle \boldsymbol{\lambda}, \boldsymbol{\epsilon}\rangle + \sum_{l=1}^M \vartheta_l s_l \\ &\text{s.t. } \sup_{\xi\in\Xi}\{\xi^\top (\mathscr{Q} - \mathrm{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda}))\xi + 2(q + \mathrm{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda})\xi^l)^\top\xi - {\xi^l}^\top \mathrm{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda})\xi^l\} \leq s_l \qquad l \in [M] \\ &\qquad\qquad q + \mathrm{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda})\xi^l \in \mathrm{range}(\mathscr{Q} - \mathrm{diag}^{\boldsymbol{d}}) \qquad l \in [M] \\ &\qquad\qquad \mathscr{Q} - \mathrm{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda}) \preceq 0. \end{aligned}$$

Since $\mathcal{Q} - \operatorname{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda}) \preceq 0$ and $q + \operatorname{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda})\xi^l \in \operatorname{range}(\mathcal{Q} - \operatorname{diag}^{\boldsymbol{d}})$ for each $l \in [M]$, the quadratic program of the corresponding epigraphical constraint admits an explicit solution [23, page 653], which yields the equivalent constraint

$$(q + \operatorname{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda})\xi^l)^\top (\operatorname{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda}) - \mathcal{Q})^\dagger (q + \operatorname{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda})\xi^l) - {\xi^l}^\top \operatorname{diag}^{\boldsymbol{d}}(\boldsymbol{\lambda})\xi^l - s_l \leq 0.$$

Combining this with the characterization of positive semidefiniteness through the Schur complement [23, page 651], we obtain the desired result. □

3.A.3. PROOFS FROM SECTION 3.7

Here we provide the proofs from Section 3.7. To this end, we generalize a saddle point result from [185, Theorem 2.1], which enables the interchange of min and max operations for convex expected-value problems over general ambiguity sets of probability distributions. We first introduce some necessary preliminaries. Let $\mathscr{A}$ be a nonempty and convex set of probability distributions on the measurable space $(\Xi, \mathscr{F})$. We also consider a closed convex subset S of $\mathbb{R}^n$, a convex neighborhood V of S, a function $\varphi : \mathbb{R}^n \times \Xi \to \mathbb{R}$, and make the following assumption.

Assumption 9. (Saddle function & domain). With $\mathscr{A}$, S, V, φ as above we assume that:
(A1) For all $x \in V$ and $P \in \mathscr{A}$ the function $\xi \mapsto \varphi(x, \xi)$ is $\mathscr{F}$-measurable and P-integrable, i.e.,

$$g(x, P) := \mathbb{E}_P[\varphi(x, \xi)] \in \mathbb{R}.$$

(A2) For all $\xi \in \Xi$ the function $x \mapsto \varphi(x, \xi)$ is convex on V.
(A3) For all $x \in V$ the max function

$$f(x) := \sup_{P \in \mathscr{A}} \mathbb{E}_P[\varphi(x, \xi)] \equiv \sup_{P \in \mathscr{A}} g(x, P)$$

is finite valued, i.e., $f(x) < +\infty$.

Consider also a topology $\mathscr{T}$ on $\mathscr{A}$ and assume that:
(B1) The set $\mathscr{A}$ is sequentially compact.
(B2) For every $x \in V$ the function $P \mapsto g(x, P)$ is continuous.

The following theorem summarizes the saddle-point property obtained by Propositions 2.1 and 2.2 in [185] for our case where $\mathscr{A}$ is convex.

Theorem 3.A.6. *(Existence of saddle-point). Let Assumptions 9(A1)-(A3) hold and $\bar{x}$ be an optimal solution of the problem*

$$\inf_{x \in S} \sup_{P \in \mathscr{A}} \mathbb{E}_\mu[\varphi(x, \xi)] \equiv \inf_{x \in S} \sup_{P \in \mathscr{A}} g(x, P) \equiv \inf_{x \in S} f(x). \tag{3.A.6}$$

Suppose further that

$$\partial f(\bar{x}) = \cup_{P \in \mathscr{A}^\star(\bar{x})} \partial g_P(\bar{x}), \tag{3.A.7}$$

where $\mathscr{A}^\star(\bar{x}) := \operatorname{argmax}_{P\in\mathscr{A}} g(\bar{x},P)$. *Then there exists* $\bar{P}\in\mathscr{A}^\star(\bar{x})$ *such that* $(\bar{x},\bar{P})$ *is a saddle point of* (3.A.6)*, namely*

$$\inf_{x\in S}\sup_{P\in\mathscr{A}} g(\bar{x},\bar{P}) = \sup_{P\in\mathscr{A}}\inf_{x\in S} g(\bar{x},\bar{P}). \tag{3.A.8}$$

Using this result enables us to generalize [185, Theorem 2.1] in terms of the growth of the objective function for minimax DRO problems over MTHs.

Theorem 3.A.7. *(Stochastic min-max equality). Let Assumption 9 hold, i.e., all (A1)-(A3) and (B1), (B2) hold, and assume that* $\bar{x}$ *is an optimal solution of* (3.A.6)*. Then* (3.A.6) *also admits a saddle point* $(\bar{x},\bar{P})$ *and the minimax equality* (3.A.8) *holds.*

Proof. Following [185], we first note that by [186, Theorem 3, Page 201], it holds that

$$\partial f(\bar{x}) = \operatorname{cl}\big(\cup_{P\in\mathscr{A}^\star(\bar{x})}\partial g_P(\bar{x})\big).$$

Therefore, to establish (3.A.7) and conclude the proof, it suffices to show that $\cup_{P\in\mathscr{A}^\star(\bar{x})}\partial g_P(\bar{x})$ is closed. To this end, let $z_k\to z\in\mathbb{R}^n$ with each $z_k\in\partial g_{P_k}(\bar{x})$ for some $P_k\in\mathscr{A}^\star(\bar{x})$. Since $\{P_k\}\subset\mathscr{A}$ and $\mathscr{A}$ is sequentially compact by (B1), a subsequence converges to some $P_\star\in\mathscr{A}$. With a slight abuse of notation, we also index this sequence by k. Taking further into account that $P\mapsto g_P(\bar{x})$ is continuous by (B2), we have that $g_{P_k}(\bar{x})\to g_{P_\star}(\bar{x})$. By the fact that each $P_k\in\mathscr{A}^\star(\bar{x})$ and the definition of $\mathscr{A}^\star$, we get that also $P_\star\in\mathscr{A}^\star(\bar{x})$. Thus, it suffices to show that $z\in\partial g_{P_\star}(\bar{x})$. Since $z_k\in\partial g_{P_k}(\bar{x})$, we have for each $x\in V$ that

$$g_{P_k}(x)-g_{P_k}(\bar{x})\geq\langle z_k,x-\bar{x}\rangle \Longrightarrow g_{P_\star}(x)-g_{P_\star}(\bar{x})\geq\langle z,x-\bar{x}\rangle,$$

where we exploited again continuity of g with respect to P. The proof is now complete. □

We next formulate a result that enables us to apply this min-max interchange to DRO problems over MTHs.

Corollary 3.A.8. *(Stochastic min-max equality for MTHs). Consider the Polish space* Ξ *with its Borel* σ*-algebra, let* $\mathscr{A}\equiv\mathscr{T}_p(Q,\boldsymbol{\varepsilon})$*,* $\mathscr{T}$ *be the weak topology on* $\mathscr{A}$*, and assume that there exists* $r\in[0,p)$*, such that for each* $x\in V$*, the function* $\xi\mapsto\varphi(x,\xi)$ *is continuous and belongs to the class* $\zeta\in\mathscr{G}_r$*. Let also Assumptions 9(A1)-(A3) hold and* $\bar{x}$ *be an optimal solution of* (3.A.6)*. Then* (3.A.6) *admits a saddle point* $(\bar{x},\bar{P})$ *and the minimax equality* (3.A.8) *holds.*

Proof. From Theorem 3.A.7 it suffices to establish (B1) and (B2), namely, that $\mathscr{A}$ is weakly (sequentially) compact and that each function $P\mapsto g(x,P)$ is continuous. These properties follow directly from Theorems 3.4.1 and 3.4.3(ii), respectively. □

Using this min-max interchange property, we can prove the results of Section 3.7 that yield tractable reformulations of distributionally robust chance-constrained problems.

Proof of Lemma 3.7.1. The proof follows by using Corollary 3.A.8, weak compactness of $\mathscr{T}_p(\boldsymbol{P}_\xi^N,\boldsymbol{\varepsilon})$, and the exact same arguments as in the proof of [168, Lemma IV.2.]. □

Proof of Proposition 3.7.2. The proof is analogous to the proof of [168, Proposition IV.3]. From Lemma 3.7.1(i), Assumption 8(ii), and the strong duality result of Theorem 2.6.4 of the previous chapter, we get that

$$\sup_{P\in\mathscr{T}_p(Q,\boldsymbol{\varepsilon})} \inf_{\tau\in\mathbb{R}} \left\{\mathbb{E}_P[(f(x,\xi)+\tau)_+] - \tau\alpha\right\}$$
$$= \inf_{\tau\in\mathbb{R}} \inf_{\boldsymbol{\lambda}\geq 0} \left\{\langle \boldsymbol{\lambda},\boldsymbol{\varepsilon}\rangle - \tau\alpha + \sum_{l\in[M]} \vartheta_l \sup_{\xi\in\Xi} \left\{(f(x,\xi)+\tau)_+ - \sum_{k=1}^n \lambda_k \|\xi_k^l - \xi_k\|^p\right\}\right\}.$$

The inner infimum with respect to $\boldsymbol{\lambda}\geq 0$ is attained by Theorem 2.6.4 of Chapter 2 and the infimum with respect to τ is also attained because of Lemma 3.7.1(ii). Thus, introducing epigraphical variables $s_l\in\mathbb{R}$, $l\in[M]$, the feasible set (3.7.4) comprises the points $x\in\mathscr{X}$ for which

$$\langle \boldsymbol{\lambda},\boldsymbol{\varepsilon}\rangle + \sum_{l=1}^M \vartheta_l s_l \leq \tau\alpha$$
$$\sup_{\xi\in\Xi} \left\{(f(x,\xi)+\tau)_+ - \sum_{k=1}^n \lambda_k \|\xi_k - \xi_k^l\|^p\right\} \leq s_l \quad l\in[M]$$

for some $\boldsymbol{\lambda}\geq 0$ and $s_l\in\mathbb{R}$, where $l\in[M]$. Using the same arguments as in the final part of the proof of [168, Proposition IV.3], this constraint is equivalent to

$$\left(\sup_{\xi\in\Xi} \left\{(f(x,\xi)+\tau)_+ - \sum_{k=1}^n \lambda_k \|\xi_k - \xi_k^l\|^p\right\}\right)_+ \leq s_l \quad l\in[M],$$

which yields the desired result. □

Proof of Proposition 3.7.3. Since $f(x,\xi)$ is affine in ξ and convex in x, and Ξ is a compact polyhedral set, it satisfies Assumption 8. Therefore, the feasible set of (3.7.3) can be determined by Proposition 3.7.2. Using the same arguments as in the proof of [168, Proposition V.1], the constraints involving the auxiliary variables s_l are equivalently written

$$s_l \geq \left(b_j(x) + \tau + \sup_{\xi\in\Xi} \left\{\langle x, A_j\xi\rangle - \sum_{k=1}^n \lambda_k \|\xi_k - \xi_k^l\|\right\}\right)_+ \quad l\in[M],\ j\in[m]. \tag{3.A.9}$$

By adjusting the derivations in [168, equation (23)] to these constraints and considering for each $l\in[M]$ and $j\in[m]$ the dual variable $z_{lj}\in\mathbb{R}^d$, we get

$$\sup_{\xi\in\Xi} \left\{\langle x, A_j\xi\rangle - \sum_{k=1}^n \lambda_k \|\xi_k - \xi_k^l\|\right\} \overset{(a)}{=} \sup_{\xi\in\Xi} \left\{\langle x, A_j\xi\rangle - \sum_{k=1}^n \sup_{\|\mathrm{pr}_k^{\boldsymbol{d}}(z_{lj})\|_*\leq\lambda_k} \langle \mathrm{pr}_k^{\boldsymbol{d}}(z_{lj}), \xi_k - \xi_k^l\rangle\right\},$$
$$\overset{(b)}{=} \inf_{\|\mathrm{pr}_k^{\boldsymbol{d}}(z_{lj})\|_*\leq\lambda_k,\ k\in[n]} \left\{\langle z_{lj}, \xi^l\rangle + \sup_{\xi\in\Xi}\{\langle A_j^\top x - z_{lj}, \xi\rangle\}\right\},$$
$$\overset{(c)}{=} \inf_{\|\mathrm{pr}_k^{\boldsymbol{d}}(z_{lj})\|_*\leq\lambda_k,\ k\in[n]} \left\{\langle z_{lj}, \xi^l\rangle + \inf_{\substack{z_{lj}=A_j^\top x - C^\top \eta_{lj} \\ \eta_{lj}\geq 0}} \langle \eta_{lj}, h\rangle\right\},$$

$$= \inf_{\substack{\|\mathrm{pr}_k^{\boldsymbol{d}}(z_{lj})\|_* \le \lambda_k,\, k\in[n] \\ \eta_{lj}\ge 0,\, z_{lj} = A_j^\top x - C^\top \eta_{lj}}} \big\{\langle z_{lj}, \xi^l\rangle + \langle \eta_{lj}, h\rangle\big\},$$

$$= \inf_{\substack{\|\mathrm{pr}_k^{\boldsymbol{d}}(A_j^\top x - C^\top \eta_{lj})\|_* \le \lambda_k \\ k\in[n],\, \eta_{lj}\ge 0}} \big\{\langle A_j^\top x - C^\top \eta_{lj}, \xi^l\rangle + \langle \eta_{lj}, h\rangle\big\},$$

Here (a) follows from the definition of the dual norm and (c) from linear programming duality for $\sup_{\xi\in\Xi}\langle A_j^\top x - z_{lj}, \xi\rangle$. Since the supremum in (b) is taken over the compact set Ξ, it is also attained. Thus, by linear programming duality, the inner infimum in (c) is also attained for some $\eta_{lj} \ge 0$. From these derivations, we can equivalently rewrite for each $j \in [m]$ and $l \in [M]$ the constraint (3.A.9) as

$$s_l \ge \Big(b_j(x) + \tau + \inf_{\substack{\eta_{lj}\ge 0 \\ \|\mathrm{pr}_k^{\boldsymbol{d}}(A_j^\top x - C^\top \eta_{lj})\|_* \le \lambda_k}} \big\{\langle A_j^\top x - C^\top \eta_{lj}, \xi^l\rangle + \langle \eta_{lj}, h\rangle\big\}\Big)_+.$$

Recalling that the infimum in this expression is attained for some $\eta_{lj} \ge 0$, the constraint is satisfied if and only if there exist $s_l \ge 0$ and $\eta_{lj} \ge 0$ such that

$$\begin{aligned} b_j(x) + \tau + \langle A_j^\top x - C^\top \eta_{lj}, \xi^l\rangle + \langle \eta_{lj}, h\rangle \le s_l \qquad & l \in [M],\ j \in [m], \\ \Big\|\mathrm{pr}_k^{\boldsymbol{d}}(A_j^\top x - C^\top \eta_{lj})\Big\|_* \le \lambda_k \qquad & l \in [M],\ j \in [m],\ k \in [n]. \end{aligned}$$

This yields the reformulation given in the statement and concludes the proof. □

3.B. SHRINKAGE OF MTHS WITH CLUSTERED REFERENCE DISTRIBUTION

Here, we quantify how the clustering process affects the size of data-driven ambiguity sets under the requirement that they contain the true distribution with prescribed probability. In particular, we seek to determine how the rate by which they shrink with the number of samples can be associated with the number of clusters when these are also allowed to depend on the sample size. This is motivated by the fact that MTHs shrink at favorable rates compared to Wasserstein balls and we want to establish to which degree such rates are retained under the clustering. To this end, we use concentration-of-measure results that yield confidence bounds on the Wasserstein distance between compactly supported distributions and their empirical approximations.

Proposition 3.B.1. *(Distance between true and empirical distribution [110, Proposition 4.2]). Assume P_ξ is supported on $\Xi \subset \mathbb{R}^d$ with $\rho_\Xi := 1/2\mathrm{diam}(\Xi) < \infty$ and that $d > 2p$. Consider the empirical distribution P_ξ^N inferred from N i.i.d. samples of P_ξ. Then*

$$\mathbb{P}\big(W_p(P^K, \boldsymbol{P}_\xi^N) \le \rho_\Xi c_1(d,p)(c_2(d,p) + (\ln \beta^{-1})^{1/2p})N^{-1/d}\big) \ge 1 - \beta,$$

where c_1, c_2 are positive constants and c_1 also depends on the norm on $\mathbb{R}^d$.

Exploiting this result, we determine an upper bound for the Wasserstein distance between the clustered distribution $\widehat{P}_\xi$ and the product empirical distribution $\boldsymbol{P}_\xi^N$.

Proposition 3.B.2. *(Wasserstein distance from optimal quantizer). Assume $\widehat{P}_\xi$ is an optimal quantizer of $\boldsymbol{P}_\xi^N$ supported at K atoms with*

$$\log_N K = q \tag{3.B.1}$$

for some $q > 1$ and let $d > 2p$. Then, we have

$$W_p(\widehat{P}_\xi, \boldsymbol{P}_\xi^N) \leq C(\rho_\Xi, d, p) N^{-q/d},$$

with $C(\rho_\Xi, d, p) := \rho_\Xi c_1(d,p) c_2(d,p)$ and ρ_Ξ, c_1, c_2 as given in Proposition 3.B.1.

Proof. First, fix a realization of the product empirical distribution $\boldsymbol{P}_\xi^N \in \mathscr{P}^{N^n}(\Xi)$ and let $P^K \in \mathscr{P}^K(\Xi)$ be supported on K atoms of $\boldsymbol{P}_\xi^N$. Then from Proposition 3.B.1, the probabilistic argument suggests that we can select for any $\beta \in (0,1)$ a distribution P^K such that

$$\begin{aligned} W_p(P^K, \boldsymbol{P}_\xi^N) &\leq \rho_\Xi c_1(d,p)(c_2(d,p) + (\ln \beta^{-1})^{1/2p}) K^{-1/d} \\ &= \rho_\Xi c_1(d,p)(c_2(d,p) + (\ln \beta^{-1})^{1/2p}) N^{-q/d}, \end{aligned}$$

where we used (3.B.1) to obtain the second line. Thus, we can consider a sequence of probabilities $\beta_i \nearrow 1$ and discrete distributions P_i^K with

$$W_p^p(P_i^K, \boldsymbol{P}_\xi^N) \leq \varepsilon_i := \rho_\Xi c_1(d,p)(c_2(d,p) + (\ln \beta_i^{-1})^{1/2p}) N^{-q/d}.$$

Identifying each discrete distributions P_i^K with a finite sequence of points in $\Xi^K \subset \mathbb{R}^{Kd}$, we deduce by compactness of Ξ^K that a subsequence of P_i^K converges to some $P_\star^K$ with

$$W_p(P_\star^K, \boldsymbol{P}_\xi^N) \leq \lim_{i \to \infty} \varepsilon_i = C(\rho_\Xi, d, p) N^{-q/d},$$

where C is given in the statement. Since $\widehat{P}_\xi$ is an optimal quantizer of $\boldsymbol{P}_\xi^N$ consisting of K atoms, which implies that $W_p(\widehat{P}_\xi, \boldsymbol{P}_\xi^N) \leq W_p(P_\star^K, \boldsymbol{P}_\xi^N)$, we obtain the desired result. □

Remark 3.B.3. (Quantization quality of the clustered distribution). The optimal quantization problem is non-convex and its solution using Lloyd-type methods only guarantees convergence to some local optimum. However, this often yields satisfactory, i.e., near-optimal approximations of the global optimum [187–189]. As a result, a bound close to the one derived in Proposition 3.B.2 is likely to hold in practice.

The following variant of Proposition 3.B.1 establishes that MTHs built from random variables with multiple independent components that contain the true distribution with prescribed confidence shrink at favorable rates with the number of samples compared to Wasserstein balls.

Proposition 3.B.4. *(Ambiguity hyperrectangle size dependence on data [164, Proposition 5.2].). Let $P_\xi := P_{\xi_1} \otimes \ldots \otimes P_{\xi_n}$, where Ξ_k is a compact subset of $\mathbb{R}^{d_k}$ and $d_k > 2p$ for each $k \in [n]$. Then one can select the radii $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ so that*

$$\mathbb{P}(P_\xi \in \mathscr{T}_p(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})) \geq 1 - \beta \quad \text{and} \quad \mathscr{T}_p(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon}) \subset \mathscr{B}_p(\boldsymbol{P}_\xi^N, \varepsilon),$$

where

$$\varepsilon = \widehat{C}(\rho_\Xi, \beta, n, d, p) N^{-1/d_{\max}}, \tag{3.B.2}$$

$d_{\max} := \max_{k \in [n]} d_k$, and $\widehat{C}$ also depends on the norm in $\mathbb{R}^d$.

Combining Propositions 3.8.1, 3.B.2, together with Proposition 3.B.4, we can determine how a shifted MTH that is centered at a quantizer of the product empirical distribution shrinks with the number of samples.

Corollary 3.B.5. *(Size of inflated MTHs). With P_ξ as in Proposition 3.B.4, let $\widehat{P}_\xi$ be an optimal quantizer of $\boldsymbol{P}_\xi^N$ supported at K atoms and assume that (3.B.1) holds for some $q > 1$. Consider also an optimal transport plan π for the W_p distance between $\widehat{P}_\xi$ and $\boldsymbol{P}_\xi^N$. Then with $\boldsymbol{\varepsilon} := (\varepsilon_1, \ldots, \varepsilon_n)$ as in Proposition 3.B.4 and $\boldsymbol{\epsilon} := (\epsilon_1, \ldots, \epsilon_n)$ where $\epsilon_k^p := \int_{\Xi \times \Xi} \|\eta_k - \zeta_k\|^p d\pi(\eta, \zeta)$ for each $k \in [n]$, we have*

$$\mathbb{P}\big(P_\xi \in \mathscr{T}_p(\widehat{P}_\xi, \boldsymbol{\varepsilon} + \boldsymbol{\epsilon})\big) \geq 1 - \beta \quad \text{and} \quad \mathscr{T}_p(\widehat{P}_\xi, \boldsymbol{\varepsilon} + \boldsymbol{\epsilon}) \subset \mathscr{B}_p(\widehat{P}_\xi, \varepsilon^\star), \tag{3.B.3}$$

where

$$\varepsilon^\star := C(\rho_\Xi, d, p) N^{-q/d} + \widehat{C}(\rho_\Xi, \beta, n, d, p) N^{-1/d_{\max}}$$

and C and $\widehat{C}$, $d_{\max}$ are given in Proposition 3.B.2 and 3.B.4, respectively.

Proof. From Proposition 3.B.4 we select $\boldsymbol{\varepsilon}$ so that $\mathbb{P}\big(\mathscr{T}_p(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon})\big) \geq 1 - \beta$. Since also $\mathscr{T}_p(\boldsymbol{P}_\xi^N, \boldsymbol{\varepsilon}) \subset \mathscr{T}_p(\widehat{P}_\xi, \boldsymbol{\varepsilon} + \boldsymbol{\epsilon})$ by Proposition 3.8.1(i), we deduce the inequality in (3.B.3). The inclusion in (3.B.3) follows directly by Propositions 3.B.2 and 3.B.4, the definition of $\varepsilon^\star$, and the triangle inequality

$$W_p(\widehat{P}_\xi, P_\xi) \leq W_p(\widehat{P}_\xi, \boldsymbol{P}_\xi^N) + W_p(\boldsymbol{P}_\xi^N, P_\xi)$$

for the Wasserstein distance. □

This corollary suggests that choosing the number $K \equiv K(N)$ of clusters, or equivalently the exponent q so that $1/d < q/d \leq 1/d_{\max}$, provides a tunable tradeoff between the size of the MTH and the complexity of its reference with fixed probabilistic guarantees. We also note that by clustering the product empirical distribution, we automatically obtain a (suboptimal) transport plan that can be used to compute explicit upper bounds for $\epsilon_1, \ldots, \epsilon_n$.

4

DISTRIBUTIONALLY ROBUST MODEL PREDICTIVE CONTROL USING HORIZON-ADAPTIVE AMBIGUITY SETS

Controlled systems often involve disturbances that can be modeled as a stochastic process, often with imprecise knowledge of the distribution. The aim of this chapter is to develop a distributionally robust model predictive control algorithm for discrete-time linear systems with stochastic disturbances that have an unknown distribution. In the presented method, we infer, from a limited number of samples, structured ambiguity sets that capture the uncertain distribution with high confidence. Exploiting reformulations from distributionally robust optimization, these ambiguity sets yield convex receding horizon control problems that are computationally tractable and recursively feasible. To control the complexity of the problem, the parameterization of each ambiguity set is tailored to the horizon stage. This is achieved by clustering the points that are used to generate the reference distributions around which the ambiguity sets are built.

4.1. INTRODUCTION

MODEL predictive control (MPC) is a well-established control design that optimizes system commands over a receding horizon and is particularly adept at handling constraints on both controls and states [190–192]. Its effectiveness hinges on employing a predictive model to ensure closed-loop stability while satisfying the constraints in a computationally tractable manner. A key challenge in MPC is to account for the typical inadequacy of idealized models to precisely describe real-world systems. In that case, discrepancies between model predictions and the actual system behavior may lead to degraded control performance, instability, or constraint violations.

Two main categories of MPC formulations systematically account for uncertainties [12]. The first is robust model predictive control, which considers the worst-case realization of the uncertainty [193]. Despite its effectiveness in ensuring robust stability and constraint satisfaction, this approach may need to account for highly unlikely scenarios, which can lead to conservative control actions that limit closed-loop performance.

The other category is stochastic MPC [12] (SMPC), which incorporates probabilistic descriptions of the uncertainty and enforces chance constraints on the state of the system [194, 195]. The computational efficiency of SMPC problems typically hinges on the construction of stochastic reachability tubes [196], and properties like recursive feasibility of the scheme can be ensured via indirect feedback [174]. There are also SMPC algorithms that control the average number of constraint violations [197] and formulations that ensure constraint satisfaction under correlated disturbances [198]. Many SMPC formulations often rely on an exact characterization of the underlying probability distribution of the uncertainty [195, 197]. In other cases, they require knowledge of specific distribution parameters, such as the mean and variance [174, 194] or a correlation bound [198]. In practice, however, such information is not always available.

When the distribution of the uncertainty is unknown, it can be inferred from a finite number of available samples. Nevertheless, the resulting model may still deviate considerably from the true distribution. To hedge against this uncertainty about the uncertainty model, distributionally robust formulations reinforce stochastic optimization problems by optimizing against the worst-case distributions from an ambiguity set of plausible probabilistic models [100]. Such ambiguity sets are usually constructed by constraining the moments of the distribution, e.g. [199], or considering all the distributions within some distance from a reference model. Convenient distances for this purpose, like the Wasserstein metric, are devised using tools from optimal transport, e.g. [79, 119, 120].

Among Distributionally Robust MPC (DRMPC) formulations, [130] learns moment-based ambiguity sets online, [200] updates its feasible set as more data are gathered along iterations, and [201] formulates SMPC problems over Wasserstein ambiguity sets of zero-mean distributions. Further approaches that rely on the propagation of ambiguity sets include [106], which focuses on linear systems, and [131], which also treats the nonlinear case. Data-driven distributionally robust MPC formulations usually rely on the assumption that the uncertainty is inferred by

i.i.d. samples. Optimal transport ambiguity sets for this purpose use the samples to form independent trajectories of the noise that are further exploited to build high-dimensional empirical models of the noise distribution along the horizon. Such monolithic ambiguity sets suffer from the curse of dimensionality when it comes to the number of independent realizations that are required to accurately capture the unknown distribution [59]. Despite the progress in formulating Wasserstein-based distributionally robust MPC algorithms, the curse of dimensionality inherent to DRO, which is exacerbated by the length of the MPC prediction horizons, is not solved yet. The aim of this chapter is to address this problem using structured ambiguity sets, developed in the previous chapters, which exploit independence across the components of high-dimensional uncertainty.

In this chapter, we design a distributionally robust model predictive control algorithm for linear systems affected by additive i.i.d. stochastic disturbances with unknown distribution. Unlike existing data-driven approaches, we quantify the disturbance propagation by informing the ambiguity set at each instance of the prediction horizon by the whole set of data. To construct the ambiguity set, we impose a product structure on its reference distribution and use a separate optimal transport constraint for each component of the uncertainty instead of considering a monolithic discrepancy budget, which can be arbitrarily allocated to the components. The ambiguity set is accompanied by statistical guarantees of containing the disturbance trajectory along the prediction horizon, which extend to all time instances simultaneously. Exploiting tractable reformulations for structured ambiguity sets [96], we determine distributionally robust tightenings for the feasible sets of chance constraints that are imposed on the state of the system. These tightenings are reliant on solving convex programs offline and do not impose any computational burden on the online part of the MPC algorithm. Since the complexity of these convex programs increases considerably with the length of the prediction horizon, we provide an equivalent formulation of the DRO problem that utilizes a telescopic sequence of structured ambiguity sets. The sets in this sequence capture the uncertainty over successive initial fragments of the horizon and have progressively increasing complexity. We systematically reduce this complexity by clustering the reference distribution of the ambiguity sets in a time-adaptive manner with respect to the horizon instance. We also establish that the proposed MPC scheme is recursively feasible. The complexity of our online MPC algorithm is equivalent to that of nominal MPC, thereby rendering it suitable for practical implementations.

This chapter is organized as follows. Section 4.2 introduces the preliminaries and notation used throughout the chapter. Section 4.3 formulates the control problem that we seek to solve. Section 4.4 introduces the model predictive control scheme that we adopt to achieve our control objectives. In Section 4.5, we introduce data-driven structured ambiguity sets and discuss their properties. We derive tractable reformulations of their associated chance constraints in Section 4.6. Section 4.7 addresses the computational complexity of the resulting optimization problems and Section 4.8 the recursive feasibility of the proposed MPC scheme. We provide numerical simulations in Section 4.9.

4.2. PRELIMINARIES AND NOTATION

Throughout this chapter, we use the following notation. Given a norm $\|\cdot\|$ in $\mathbb{R}^d$ we denote by $\|\cdot\|_*$ its dual norm. For a positive semi-definite matrix $M \in \mathbb{R}^{n\times n}$ and any $x \in \mathbb{R}^n$, we denote $\|x\|_M := \langle x, Mx\rangle$. Given $N \in \mathbb{N}$ we denote $[N] := \{1,\dots,N\}$, and $[N]_0 := [N] \cup \{0\}$. For any vector and $h \in \mathbb{R}^n$ and matrix $H \in \mathbb{R}^{n\times m}$ with $m, n \in \mathbb{N}$, h_j and H_j denote the j-th component of h and row of H, respectively. We denote by $\succeq$ the partial order on $\mathbb{R}^d$ with $x \succeq y$ iff $x_k \geq y_k$ for all $k \in [d]$. We use the notation $x_+ := \max\{x, 0\}$ for any real number x. Given a set Ξ and $k \in [m]$, we define the projection operators $\mathrm{pr}_k : \Xi^m \to \Xi$ and $\mathrm{pr}_{[k]} : \Xi^m \to \Xi^k$ as $\mathrm{pr}_k(\boldsymbol{\xi}_m) := \xi_k$ and $\mathrm{pr}_{[k]}(\boldsymbol{\xi}_m) = (\xi_1,\dots,\xi_k)$, respectively, for all $\boldsymbol{\xi}_T \equiv (\xi_1,\dots,\xi_T) \in \Xi^m$. We also use the compact notation $\boldsymbol{\xi}_{[k]} := \mathrm{pr}_{[k]}(\boldsymbol{\xi}_m)$. The Minkowski sum and Pontryagin difference of two sets $X, Y \subset \mathbb{R}^n$, are defined as $X \oplus Y := \{x + y : x \in X,\ y \in Y\}$ and $X \ominus Y := \{x \in X : x + y \in X \text{ for all } y \in Y\}$, respectively. The Kronecker product of two matrices A and B is denoted by $A \otimes B$.

Given two measurable spaces $(\Omega, \mathscr{F})$, $(\Omega', \mathscr{F}')$, the measurable map $\Psi : (\Omega, \mathscr{F}) \to (\Omega', \mathscr{F}')$, assigns to each measure μ in $(\Omega, \mathscr{F})$ the pushforward measure $\Psi_\# \mu$ in $(\Omega', \mathscr{F}')$ with $\Psi_\#\mu(B) := \mu(\Psi^{-1}(B))$ for all $B \in \mathscr{F}'$. Given $\Xi \subset \mathbb{R}^d$, the 1-Wasserstein distance of two distributions $P, Q \in \mathscr{P}_1(\Xi)$ (i.e., with finite 1st moment) is

$$W(Q,P) := \inf_{\pi \in \mathscr{C}(Q,P)} \left\{ \int_{\Xi\times\Xi} \|\zeta - \xi\| \, d\pi(\zeta,\xi) \right\}$$

(cf. [117]). Each $\pi \in \mathscr{C}(Q,P)$ is a transport plan, i.e., a distribution on $\Xi \times \Xi$ with marginals $P = \mathrm{pr}_{2\#}\pi$ and $Q = \mathrm{pr}_{1\#}\pi$, respectively. We denote by $P \otimes Q$ the product distribution of P and Q and by $P^{\otimes t}$ the t-fold product of P with itself for some $t \in \mathbb{N}$. Given a distribution $\mu \in \mathscr{P}(\Xi)$ we define the identity coupling $\pi_{\mathrm{Id}}(\mu)$ on $\Xi \times \Xi$ as

$$\pi_{\mathrm{Id}}(\mu) := T_{\mathrm{diag}\#}\mu, \quad \text{where} \quad T_{\mathrm{diag}}(\xi) := (\xi, \xi).$$

(cf. [117]). The Dirac distribution centered at $\xi \in \Xi$ is denoted by δ_ξ. The CVaR of a random variable X with distribution P at level α is given by

$$\mathrm{CVaR}^P_{1-\alpha}(X) := \inf_\tau \{\alpha^{-1}\mathbb{E}_P[(X+\tau)_+] - \tau\}.$$

4.3. PROBLEM FORMULATION

Consider the discrete-time LTI system

$$x_{t+1} = Ax_t + Bu_t + w_t, \tag{4.3.1}$$

where $x, w \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, the pair (A, B) is stabilizable, and the initial condition x_0 is known. The stochastic noise sequence $\{w_t\}_{t\in\mathbb{N}}$ is independent and identically distributed according to some *unknown* distribution P_w, for which we make the following assumption.

Assumption 10. (Noise distribution class). The distribution P_w is supported on the compact polytope $\mathscr{W} := \{w \in \mathbb{R}^n : Fw \preceq f\}$, $f \in \mathbb{R}^{n_w}$, which contains the origin.

Under full-state measurements, which is a standing assumption in this chapter, we can use the dynamical model (4.3.1) of the system to retrieve past realizations of the noise (under suitable assumptions, this is also possible when we only have partial measurements [202]). In particular, we consider access to N i.i.d. samples $\widehat{w}^1,\dots,\widehat{w}^N$ of w that are obtained by ideal full-state measurements and denote $\widehat{\boldsymbol{w}} := (\widehat{w}^1,\dots,\widehat{w}^N)$. We also assume that these samples have been collected *prior* to the initialization of the system and are independent of the noise process $\{w_t\}_{t\in\mathbb{N}}$. As a result, the data-driven part of our algorithm design is separated (independent) from how the noise process affects the system. Thus, *we assume throughout that all uncertain elements, i.e, the samples $\widehat{w}^1,\dots,\widehat{w}^N$ and the process $\{w_t\}_{t\in\mathbb{N}}$ belong to the same probability space and denote by $\mathbb{P}$ its underlying probability measure.*

Given the state x_t of the system at time t, we model its predicted states by

$$x_{k+1|t} = Ax_{k|t} + Bu_{k|t} + w_{k+t} \quad k\in\mathbb{N}_0$$

where $x_{0|t} = x(t)$. We also consider a (static) measurable control policy $u_t \equiv \pi(x_t)$. This policy fixes a probability measure on the system and its predicted states by setting $u_{k|t} \equiv \pi(x_{k|t})$. The system is also subject to chance constraints on the predicted states and hard constraints on the input. Namely, at every time t we consider the constraints

$$\mathbb{P}(\langle H_j^\top, x_{k|t}\rangle \le h_j | x_t) \ge 1-p_j \quad k\in\mathbb{N},\ j\in[J], \tag{4.3.2}$$

$$u_t \in \mathscr{U} \tag{4.3.3}$$

where $h\in\mathbb{R}^J$ and $\mathscr{U}$ is a polytopic set.

To control system (4.3.1) we solve a quadratic optimal control problem under the above constraints on its state and input. Since optimal control laws for such problems typically lack closed-form parameterizations, we follow the MPC paradigm, which approximates infinite-horizon formulations by receding horizon optimization problems. While there is extensive literature on stochastic MPC under a known noise distribution P_w [12], the consideration of distributional uncertainty entails several open questions. Here, we are interested in obtaining a distributionally robust formulation of the MPC problem for system (4.3.1) by building from the data $\widehat{w}^1,\dots,\widehat{w}^N$ a suitable ambiguity set that contains the true distribution with high confidence.

We want to provide a tractable approximation of the feasible set of x_t, that shares rigorous probabilistic guarantees without being conservative compared to the feasible set (4.3.2) when P_w is known. To achieve this, we seek a reliable approximation of the joint distribution of the noise trajectory along the prediction horizon. This motivates the key problems that we aim to address, which are *i) how to exploit distributionally robust optimization to derive tractable characterizations of the feasible set* (4.3.2) *under distributional uncertainty, and, ii) how to optimally exploit the available data and parameterize ambiguity sets with probabilistic guarantees of containing the distribution of the noise trajectory.*

4.4. DISTRIBUTIONALLY ROBUST MPC

Here we start laying down the building blocks of our solution to the distributionally robust MPC problem. We first provide a concrete formulation of the optimal control problem that we seek to solve in a receding horizon fashion.

4.4.1. CONTROL POLICY AND CLOSED-LOOP DYNAMICS

To get a tractable optimization problem, we need to fix a suitable control policy. Let z_t refer to the nominal part of the state of (4.3.1), which describes its evolution in the absence of disturbances, and denote the error between the true and nominal state by $e_t := x_t - z_t$. We consider a control policy that consists of a nominal term v_t and a correction term that is linear in the propagation error e_t. Namely, we set

$$u_t = v_t + Ke_t \tag{4.4.1}$$

for a stabilizing feedback controller K for (4.3.1), i.e., such that $A+BK$ is Schur. Now, fixing the prediction horizon length T, and parameterizing further the nominal control v_t to depend affinely on Kz_t, we obtain from (4.3.1) and (4.4.1) the equivalent prediction models for the closed-loop dynamics at every time $t \in \mathbb{N}$ and for every $k \in [T-1]_0$

$$z_{k+1|t} = Az_{k|t} + Bv_{k|t} \tag{4.4.2a}$$
$$v_{k|t} = c_{k|t} + Kz_{k|t} \tag{4.4.2b}$$
$$e_{k+1|t} = (A+BK)e_{k|t} + w_{k+t}. \tag{4.4.2c}$$

Here $z_{k|l}$ is set equal to x_k and therefore $e_{k|0} = 0$ for each time step k. Next, considering the $n \times Tn$ matrices

$$\boldsymbol{G}_k := \begin{bmatrix} (A+BK)^{k-1} \cdots A+BK & I_n & 0 \dots 0 \end{bmatrix}, \tag{4.4.3}$$

for $k \in [T-1]_0$, the error equation is equivalently written

$$e_{k|t} = \boldsymbol{G}_k \boldsymbol{\xi}_{T|t}, \tag{4.4.4}$$

where $\boldsymbol{\xi}_{T|t}$ represents the disturbance trajectory throughout the prediction horizon, i.e.,

$$\boldsymbol{\xi}_{T|t} := (w_t, \dots, w_{t+T-1}) \quad \text{and} \quad \boldsymbol{\xi}_T := \boldsymbol{\xi}_{T|0}. \tag{4.4.5}$$

The model predictive control algorithm that we want to design, should compute the nominal control term v_t at every time instant t to maximize the performance of the closed-loop system (4.4.2) while guaranteeing the satisfaction of the constraints (4.3.2) and (4.3.3) at every time instant $t > 0$.

4.4.2. CONSTRAINT TIGHTENING

Here we determine tightenings of the chance-constraints (4.3.2) that hedge against the lack of knowledge of the noise distribution. This hinges on exploiting the expression (4.4.4) for the propagation error to derive successive tightenings of the constraints. To this end, we provide an equivalent characterization of the feasible set of (4.3.2) in terms of the nominal state and the propagation error.

Proposition 4.4.1. *(Equivalent nominal constraints [195, Proposition 1]). Under Assumption 10, system (4.4.2) satisfies the chance constraints (4.3.2) for $k \in [T]$ and $j \in [J]$ if and only if the nominal system (4.4.2a) satisfies the constraint*

$$z_{k|t} \in \mathcal{Z}_k := \{z \in \mathbb{R}^n : Hz \preceq \tilde{h}_k\}, \qquad k \in [T], \tag{4.4.6}$$

where each component $\tilde{h}_{k,j}$ of $\tilde{h}_k$ is equal to the optimal value of the problem

$$\begin{aligned} &\max_{\gamma \in \mathbb{R}} \gamma \\ &\text{s.t. } P_{\boldsymbol{\xi}_T}(\gamma \le h_j - \langle \boldsymbol{G}_k^\top H_j^\top, \boldsymbol{\xi}_T \rangle) \ge 1 - p_j. \end{aligned} \tag{4.4.7}$$

Note that since the error equation is initiated at zero at each time t and the process noise is i.i.d., the random element $\boldsymbol{\xi}_T$ in (4.4.7) that represents the noise over the horizon is independent of the time instant t.

Remark 4.4.2. (Advantage of single chance constraints). In this work, we consider single chance constraints on every component of the state. These enable to enforce different violation levels for different states, e.g., we may be interested in imposing harder constraints for the position of a system and softer ones for its velocity. They also have the advantage of not leading to an increased number of tightened constraints (see e.g., [195, Section IV.C] for more details).

From Proposition 4.4.1, we can replace the chance constraints (4.3.2) by equivalent linear constraints on the nominal state z that we seek to satisfy along the prediction horizon. These constraints are expressed through the optimization problem (4.4.7), which is typically nonconvex and therefore difficult to solve. To address this issue, we tighten the chance-constraint (4.4.7), i.e., we shrink its feasible set, using its conditional value at risk approximation. This is commonly employed in chance-constrained optimization [96, 106, 176] and yields a tractable convex problem. Namely, we consider for each $j \in [J]$ and $k \in [T]$ the CVaR-constrained problem

$$\begin{aligned} &\max_{\gamma \in \mathbb{R}} \gamma \\ &\text{s.t. } \mathrm{CVaR}_{1-p_j}^{P_{\boldsymbol{\xi}_T}}(\langle \boldsymbol{G}_k^\top H_j^\top, \boldsymbol{\xi}_T \rangle + \gamma - h_j) \le 0, \end{aligned} \tag{4.4.8}$$

which yields a threshold γ that also satisfies the chance constraint (4.4.7).

Since the distribution of w and therefore also $P_{\boldsymbol{\xi}_T}$ is unknown, we cannot directly solve this chance-constrained problem. To remedy this, we follow the distributionally robust paradigm. Namely, we consider an ambiguity set $\widehat{\mathscr{P}}_{\boldsymbol{\xi}_T}$ of distributions that contains $P_{\boldsymbol{\xi}_T}$ with high confidence and solve the distributionally robust problem

$$\begin{aligned} &\max_{\gamma \in \mathbb{R}} \gamma \\ &\text{s.t. } \sup_{P \in \widehat{\mathscr{P}}_{\boldsymbol{\xi}_T}} \mathrm{CVaR}_{1-p_j}^{P}(\langle \boldsymbol{G}_k^\top H_j^\top, \boldsymbol{\xi}_T \rangle + \gamma - h_j) \le 0 \end{aligned} \tag{4.4.9}$$

in place of (4.4.8). We are particularly interested in choosing a data-driven ambiguity set that is inferred from the available i.i.d. realizations of the process noise.

Ideally, such an ambiguity set should yield tractable optimization problems and simultaneous out-of-sample guarantees for all constraints while being minimally conservative. Namely, given a confidence level $1-\beta$ with $\beta \in (0,1)$, our goal is to obtain the out-of-sample guarantee

$$\mathbb{P}(\widehat{\gamma}^{\star}_{k,j} \leq \gamma^{\star}_{k,j} \text{ for all } j \in [J] \text{ and } k \in [T]) \geq 1-\beta, \tag{4.4.10}$$

where $\widehat{\gamma}^{\star}_{k,j}$ and $\gamma^{\star}_{k,j}$ denote the solutions of (4.4.9) and (4.4.8), respectively.

Remark 4.4.3. (Probabilistic input constraints). The same constraint tightening approach can also be used to impose distributionally robust chance constraints on the input u (cf. [195]). Namely, when u is determined through the control policy (4.4.1), the constraints $P_{\boldsymbol{\xi}_T}(G_j u \leq g_j) \geq 1-q_j,\ j \in [J']$ are tightened by requiring that $G_j v \leq \widetilde{g}_{k,j}$ for each j, where $g_{k,j}$ is equal to the optimal value of

$$\begin{aligned} &\max_{\gamma \in \mathbb{R}} \gamma \\ &\text{s.t.} \sup_{P \in \widehat{\mathscr{P}}_{\boldsymbol{\xi}_T}} \operatorname{CVaR}^{P}_{1-q_j}(\langle \boldsymbol{G}_k^{\top} K^{\top} G_j^{\top}, \boldsymbol{\xi}_T \rangle + \gamma - g_j) \leq 0. \end{aligned}$$

4.4.3. STOCHASTIC MPC

We want to design a stochastic model predictive controller that measures at every time instant $t \in \mathbb{N}$ the state x_t and updates its control action by solving the receding horizon optimal control problem

$$\begin{aligned} \min_{\boldsymbol{v},\boldsymbol{c}} \ & \textstyle\sum_{k=0}^{T} \|z_{k|t}\|_Q^2 + \|v_{k|t}\|_R^2 && && (4.4.11)\\ \text{s.t. } & z_{k+1|t} = A z_{k|t} + B v_{k|t} && k \in [T-1]_0 \\ & v_{k|t} = c_{k|t} + K z_{k|t} && k \in [T-1]_0 \\ & z_{k|t} \in \mathscr{Z}_k(\widehat{\mathscr{P}}_{\xi_T}) && k \in [T] \\ & v_{k|t} \in \mathscr{U} \ominus K\mathscr{E}_k && k \in [T-1]_0 \\ & z_{0|t} = x_t. \end{aligned}$$

Here $T \in \mathbb{N}$ is the prediction-horizon length, Q is a positive semi-definite symmetric matrix, R is a positive definite matrix, $\mathscr{Z}_k(\widehat{\mathscr{P}}_{\boldsymbol{\xi}_T})$ is a nominal state-constraint set that is determined by the data-driven ambiguity set $\widehat{\mathscr{P}}_{\boldsymbol{\xi}_T}$, and $\mathscr{E}_k$ is obtained by the recursion

$$\mathscr{E}_{k+1} = (A+BK)\mathscr{E}_k \oplus \mathscr{W}, \qquad \mathscr{E}_0 = \{0\}.$$

The robust constraints on the nominal input are introduced to ensure the satisfaction of (4.3.3) for all $w \in \mathscr{W}$.

By selecting an ambiguity set that contains the true distribution of $\boldsymbol{\xi}_T$ with prescribed confidence $1-\beta$, we also ensure the simultaneous satisfaction of all chance constraints across time with the same confidence. In particular, selecting $\mathscr{Z}_k(\widehat{\mathscr{P}}_{\boldsymbol{\xi}_T})$ according to (4.4.6), where each threshold $\widetilde{h}_{k,j}$ is given by (4.4.9), we

ensure that (4.4.10) is satisfied for all times. As this guarantee only depends on $\boldsymbol{\xi}_T$, conditioned on the event that the ambiguity set contains the true distribution, all chance constraints will be met across time as long as (4.4.11) is feasible. Once the optimization problem is solved, yielding an optimal pair $(\boldsymbol{v}^\star, \boldsymbol{c}^\star)$, the controller provides the input $u_t = c^\star_{0|t} + Kx_t = v^\star_{0|t} + Ke_t$ to the system and repeats the same steps. This way, we obtain a static control policy $u_t \equiv \pi(x_t, \widehat{\boldsymbol{w}})$ that depends on the initially collected samples $\widehat{\boldsymbol{w}}$, and is defined over all $x_t \in \mathbb{R}^n$ for which the optimization problem (4.4.11) is feasible.

4.5. DATA-DRIVEN STRUCTURED AMBIGUITY SETS

To determine the tightenings of the nominal feasible sets $\mathscr{Z}_k(\widehat{\mathscr{P}}_{\boldsymbol{\xi}_T})$, we need to solve TJ DRO problems over a data-driven ambiguity set. This ambiguity set is built from the collected data $\widehat{\boldsymbol{w}}$ of the noise and captures plausible models of the noise trajectory distribution $P_{\boldsymbol{\xi}_T}$ along the prediction horizon. We will consider optimal transport ambiguity sets, which are accompanied by statistical guarantees of containing the true distribution of the data.

To present the key ideas, it is convenient to consider random vectors $\boldsymbol{\xi}_k := (w_0, \ldots, w_{k-1})$ with an arbitrary number of i.i.d. copies of w, instead of focusing on the case when k is equal to the horizon length T. As before, we also consider N i.i.d. samples from w that we compactly also denote by the vector $\widehat{\boldsymbol{w}} := (\widehat{w}^1, \ldots, \widehat{w}^N)$. To build a data-driven ambiguity set, we construct a reference distribution $\widehat{P} \equiv \widehat{P}(\widehat{\boldsymbol{w}})$ from the samples and group all distributions that lie within a certain optimal transport discrepancy from this reference. We will present two options for both the reference distribution and the discrepancy measure. Choosing the first option for both yields traditional Wasserstein balls, whereas choosing the second option for both yields structured ambiguity sets with enhanced statistical guarantees.

To present the two options regarding the reference distribution, we assume without loss of generality that the number N of samples from w is an integer multiple of k, i.e., $N = mk$ for some $m \in \mathbb{N}$. Then, partitioning the samples into m consecutive clusters $(\widehat{w}^1, \ldots, \widehat{w}^k), \ldots, (\widehat{w}^{(m-1)k+1}, \ldots, \widehat{w}^{mk})$, we obtain m independent copies $\{\widehat{\boldsymbol{\xi}}^i_k\}_{i\in[m]}$ of $\boldsymbol{\xi}_k$ and build the empirical distribution

$$\widehat{P}(\widehat{\boldsymbol{w}}) \equiv P^m_{\boldsymbol{\xi}_k} := \frac{1}{m}\sum_{i=1}^m \delta_{\widehat{\boldsymbol{\xi}}^i_k}$$

as a reference model for $P_{\boldsymbol{\xi}_k}$. The other altrenative is to first approximate P_w by the empirical distribution

$$P^N_w := \frac{1}{N}\sum_{i=1}^N \delta_{\widehat{w}^i}$$

and then take the product measure

$$\widehat{P}(\widehat{\boldsymbol{w}}) \equiv (P^N_w)^{\otimes T} = \frac{1}{N^k}\bigotimes_{\kappa=1}^k \sum_{i=1}^N \delta_{\widehat{w}^i}. \tag{4.5.1}$$

as a reference distribution for $P_{\boldsymbol{\xi}_k} \equiv P_w^{\otimes k}$. We will later also consider variants of (4.5.1), which are again discrete and supported on a smaller number of points.

The first of the two options to group the distributions of the ambiguity set is to consider the Wasserstein ball

$$\mathscr{B}(\widehat{P},\varepsilon) := \{P \in \mathscr{P}_1(\mathbb{R}^{kn}) : W(\widehat{P},P) \leq \varepsilon\},$$

which has radius ε an is centered at the reference distribution. When $\widehat{P} = P_{\boldsymbol{\xi}_k}^m$, this yields the traditional Wasserstein ambiguity sets that are used in data-driven problems.

The other option to build a data-driven ambiguity set around the reference distribution is to consider the *multi-transport hyperrectangle* of vector radius $\boldsymbol{\varepsilon}_k := (\varepsilon_1,\ldots,\varepsilon_k)$ that has been introduced in Chapter 2 and is given by

$$\begin{aligned}\mathscr{T}(\widehat{P},\boldsymbol{\varepsilon}_k) := \big\{&\mathrm{pr}_{2\#}\pi : \pi \in \mathscr{P}(\mathscr{W}^k \times \mathscr{W}^k), \mathrm{pr}_{1\#}\pi = \widehat{P}, \text{ and} \\ &\int_{\mathscr{W}^k\times\mathscr{W}^k} \|\zeta_\kappa - \xi_\kappa\|\, d\pi(\zeta,\xi) \leq \varepsilon_\kappa, \text{for all } \kappa \in [k]\big\}. \end{aligned} \tag{4.5.2}$$

Unlike the Wasserstein ball, the multi-transport hyperrectangle encodes distributions that are close to $\widehat{P}$ by introducing multiple optimal transport constraints; namely, a separate constraint along each component of the uncertainty. As a result, it provides a more refined representation of the uncertainty since the monolithic discrepancy budget of a Wasserstein ball can be arbitrarily allocated to the components. This characteristic of the ball is juxtaposed in our case with the fact that all marginal distributions of $\boldsymbol{\xi}_T$ are identical and should be equally far away from a product reference model like $(P_w^N)^{\otimes k}$.

The following result certifies how probabilistic guarantees of $\mathscr{T}((P_w^N)^{\otimes k},\boldsymbol{\varepsilon}_k)$ containing the distribution of the whole noise trajectory $\boldsymbol{\xi}_k$ are directly inherited by the guarantee that the noise distribution at a single time instant lies in the Wasserstein ball $\mathscr{B}(P_w^N,\varepsilon)$ when $\boldsymbol{\varepsilon}_k \equiv (\varepsilon,\ldots,\varepsilon)$.

Proposition 4.5.1. *(Probabilistic guarantees for multi-transport hyperrectangle). Consider the reference distribution $(P_w^N)^{\otimes k}$ given by (4.5.1), the vector of radii $\boldsymbol{\varepsilon}_k \equiv (\varepsilon,\ldots,\varepsilon)$, and assume $\mathscr{B}(P_w^N,\varepsilon)$ contains P_w with confidence $1-\beta$, $\beta \in (0,1)$. Then the multi-transport hyperrectangle (4.5.2) contains $P_w^{\otimes k}$ with confidence $1-\beta$ and*

$$\mathscr{T}((P_w^N)^{\otimes k},\boldsymbol{\varepsilon}_k) \subset \mathscr{B}((P_w^N)^{\otimes k}, k\varepsilon). \tag{4.5.3}$$

Proof. To prove the first part of the statement, we introduce as in Chapter 2 the ambiguity set

$$\mathscr{H}((P_w^N)^{\otimes k},\boldsymbol{\varepsilon}_T) := \{P'_{w_1}\otimes\cdots\otimes P'_{w_k} \in \mathscr{P}(\mathscr{W}^k) : \quad P'_{w_\kappa} \in \mathscr{B}(P_w^N,\varepsilon_\kappa) \text{ for all } \kappa \in [k]\}.$$

Since $P_w^{\otimes k}$ is a product measure, it follows from Proposition 2.4.4 of Chapter 2 that

$$\begin{aligned}\mathbb{P}\big(P_w^{\otimes k} \in \mathscr{T}((P_w^N)^{\otimes k},\boldsymbol{\varepsilon}_k)\big) &= \mathbb{P}\big(P_w^{\otimes k} \in \mathscr{H}((P_w^N)^{\otimes k},\boldsymbol{\varepsilon}_k)\big) \\ &= \mathbb{P}(P_w \in \mathscr{B}(P_w^N,\varepsilon))\end{aligned}$$

$$\geq 1-\beta.$$

The proof of the second part is a direct consequence of Proposition 2.4.6 of Chapter 2. □

This proposition establishes that (4.5.2) does not suffer from the curse of dimensionality that traditionally characterizes Wasserstein balls for containing the data-generating distribution of high-dimensional random variables [59, 95, 164], as explained in Chapter 2. This is justified by the containment (4.5.3), which establishes that the bound on the Wasserstein distance between the true distribution $P_w^{\otimes k}$ of the trajectory $\boldsymbol{\xi}_k$ and the reference measure $(P_w^N)^{\otimes k}$ is analogous (up to a constant at most k) to the same bound for the distance between the pointwise distribution P_w of the uncertainty and its empirical distribution P_w^N. In particular, this happens with prescribed probability, and the ratio of these bounds does not depend on the number of samples.

Remark 4.5.2. (Probabilistic guarantees for distributionally robust chance constraints). Given any number N of samples, we can choose $\boldsymbol{\varepsilon}_T = (\varepsilon, \dots, \varepsilon)$ according to Proposition 2.5.2 of Chapter 2, to ensure that $\mathscr{T}((P_w^N)^{\otimes T}, \boldsymbol{\varepsilon}_T))$ as given by (4.5.2) with the reference distribution $(P_w^N)^{\otimes T}$ given by (4.5.1) contains $P_w^{\otimes T}$ with prescribed confidence. This in turn implies that all chance constraints in (4.4.9) are met with $\widehat{\mathscr{P}}_{\boldsymbol{\xi}_T} \equiv \mathscr{T}((P_w^N)^{\otimes T}, \boldsymbol{\varepsilon}_T))$ and therefore (4.4.10) also holds for all times.

4.6. REFORMULATION OF CVaR-CONSTRAINED PROBLEMS OVER MULTI-TRANSPORT HYPERRECTANGLES

Here we exploit duality results from distributionally robust optimization (DRO) to determine tractable reformulations of the distributionally robust CVaR-constrained problem (4.4.9). We provide tractable reformulations of the problem

$$\begin{aligned} &\max_{\gamma \in \mathbb{R}} \gamma \\ &\text{s.t.} \sup_{P \in \mathscr{T}(\widehat{P}, \boldsymbol{\varepsilon})} \mathrm{CVaR}_{1-p_j}^{P}(\langle \boldsymbol{G}_k^\top H_j^\top, \boldsymbol{\xi}_T \rangle + \gamma - h_j) \leq 0, \end{aligned} \tag{4.6.1}$$

where $\widehat{P}$ is a discrete distribution that is supported on M atoms and represents a desirable reference model for $P_{\boldsymbol{\xi}_T}$. To this end, we leverage reformulations of DRO problems over multi-transport hyperrectangles. The next result provides reformulations of distributionally robust CVaR-constrained problems with piecewise affine constraints.

Proposition 4.6.1. *(Reformulation of piecewise affine* CVaR*-constrained problems Corollary 3.7.5 of Chapter 3). Consider a random variable ξ and assume its distribution is supported on the compact polytope $\Xi := \{\xi \in \mathbb{R}^n : C\xi \leq d\}$. Then the DRO problem*

$$\inf_{\gamma \in \Gamma} \langle g, \gamma \rangle \tag{4.6.2}$$

$$\text{s.t.} \sup_{P\in\mathcal{T}(\widehat{P},\boldsymbol{\varepsilon})} \mathrm{CVaR}^{P}_{1-p}(\langle a,\xi\rangle + b(\gamma)) \le 0,$$

where $\Gamma \subset \mathbb{R}^{\ell}$ *is convex and closed,* $b:\mathbb{R}^{\ell}\to\mathbb{R}$ *is convex, and* $\widehat{P}$ *is the discrete distribution* $\sum_{l=1}^{M}\theta^{l}\delta_{\xi^{l}}$ *with mass* θ^{l} *at each point* ξ^{l}*, admits the convex reformulation*

$$\begin{aligned}
\inf_{\substack{\gamma\in\Gamma,\boldsymbol{\lambda}\ge 0,\tau\in\mathbb{R}\\ s^{l}\ge 0,\eta^{l}\ge 0,l\in[N]}} \quad & \langle g,\gamma\rangle \\
\text{s.t.} \quad & \langle \boldsymbol{\lambda},\boldsymbol{\varepsilon}\rangle - \tau p + \sum_{l=1}^{M}\theta^{l}s^{l} \le 0 \\
& b(\gamma) + \tau + \langle a - C^{\top}\eta^{l},\xi^{l}\rangle + \langle \eta^{l},d\rangle \le s^{l} \\
& \|\mathrm{pr}_{q}(a - C^{\top}\eta^{l})\|_{*} \le \lambda_{q}, \quad q\in[T], l\in[M].
\end{aligned}$$

Using this result, we obtain the following tractable reformulation of the distributionally robust CVaR constrained problem (4.6.1).

Corollary 4.6.2. *(Tractable reformulation of* (4.6.1)*). Under Assumption 10, the solution of the distributionally robust* CVaR *constrained problem* (4.6.1) *can be evaluated by solving the convex program*

$$\begin{aligned}
\max_{\substack{\gamma\in\mathbb{R},\boldsymbol{\lambda}\ge 0,\tau\in\mathbb{R}\\ s^{l}\ge 0,\eta^{l}\ge 0,l\in[n]}} \quad & \gamma \\
\text{s.t.} \quad & \langle \boldsymbol{\lambda},\boldsymbol{\varepsilon}\rangle + \sum_{l=1}^{M} s^{l}\theta^{l} \le \tau p_{j} \\
& \gamma - h_{j} + \tau + \langle \boldsymbol{G}_{k}^{\top}H_{j}^{\top} - F_{T}^{\top}\eta^{l},\boldsymbol{\xi}_{T}^{l}\rangle + \langle \eta^{l},f_{T}\rangle \le s^{l} \\
& \|\mathrm{pr}_{q}(\boldsymbol{G}_{k}^{\top}H_{j}^{\top} - F_{T}^{\top}\eta^{l})\|_{*} \le \lambda_{q}, \\
& l\in[M], q\in[T],
\end{aligned}$$

where $f_{T} := (f,\dots,f)\in\mathbb{R}^{kn}$*,* $F_{T} := I_{T}\otimes F$*, and* $\boldsymbol{G}_{k}$ *is given in* (4.4.3).

Proof. The result is a direct consequence of Proposition 4.6.1 applied with $a \equiv \boldsymbol{G}_{k}^{\top}H_{j}^{\top}$, $p \equiv p_{j}$, and $b(\gamma)\equiv\gamma - h_{j}$. In particular, convexity of the set $\Xi \equiv \{\xi\in\mathbb{R}^{Tn} : F_{T}\xi \le f_{T}\}$ where $\boldsymbol{\xi}_{T}$ is supported, with F_{T} and f_{T} as given in the statement, follows from (4.4.5) and Assumption (10). □

This reformulation yields a tractable optimization to determine a nominal constraint set that guarantees the satisfaction of the chance constraints (4.3.2). A key advantage of the overall approach is that this optimization problem can be solved offline and does not add any computational burden to the online part of the MPC implementation. This renders our approach suitable in practice, as its online computational complexity is analogous to that of a nominal model predictive controller.

4.7. COMPUTATIONAL COMPLEXITY REDUCTION

A drawback of solely considering the product empirical distribution $(P_w^N)^{\otimes T}$ in (4.5.1) as a baseline model to infer the ambiguity set (4.5.2), is that the complexity of the reformulation in Section 4.6 grows exponentially with the length of the prediction horizon. Namely, the number of variables and constraints in the reformulation of Corollary 4.6.2 is proportional to the number of atoms of the discrete distribution $\widehat{P} \equiv (P_w^N)^{\otimes T}$. As this number grows exponentially with T, the reformulation become intractable for large horizons, even when solved offline.

To address this issue, we consider two complementary adjustments of the distributionally robust chance constraints. The first is merely an equivalent representation of the ambiguity sets appearing in CVaR constraints (4.4.9). I hinges on the observation that for the first k prediction steps, only the k first n-dimensional components of $\boldsymbol{\xi}_T$ are involved in the associated CVaR constraint. Namely, we only need to consider the k first marginals of the distributions in $\mathscr{T}((P_w^N)^{\otimes T}, \boldsymbol{\varepsilon}_T)$. The next result establishes that these marginals are precisely the distributions of the lower-dimensional hyperrectangle $\mathscr{T}((P_w^N)^{\otimes k}, \boldsymbol{\varepsilon}_k)$.

Proposition 4.7.1. *(Prefix projection of multi-transport hyperrectangle). Consider the multi-transport hyperrectangle $\mathscr{T}((P_w^N)^{\otimes T}, \boldsymbol{\varepsilon}_T)$, with $\boldsymbol{\varepsilon}_T := (\varepsilon_1, \dots, \varepsilon_T)$. Then*

$$\mathrm{pr}_{[k]\#}\mathscr{T}((P_w^N)^{\otimes T}, \boldsymbol{\varepsilon}_T) = \mathscr{T}((P_w^N)^{\otimes k}, \boldsymbol{\varepsilon}_k)$$

for every $k \in [T]$, where $\boldsymbol{\varepsilon}_k := (\varepsilon_1, \dots, \varepsilon_k)$.

Proof. We proceed by double inclusion. To this end, let $P \in \mathscr{T}((P_w^N)^{\otimes T}, \boldsymbol{\varepsilon}_T)$. Then there exists a transport plan $\pi \in \mathscr{P}(\mathscr{W}^T \times \mathscr{W}^T)$ such that the transport constraints in (4.5.2) are satisfied and $P = \mathrm{pr}_{2\#}\pi$. Denoting by $\widetilde{\mathrm{pr}}_{[k]} : \mathscr{W}^T \times \mathscr{W}^T \to \mathscr{W}^k \times \mathscr{W}^k$ the projection operator

$$\widetilde{\mathrm{pr}}_{[k]}(\boldsymbol{\zeta}_T, \boldsymbol{\xi}_T) := (\mathrm{pr}_{[k]}(\boldsymbol{\zeta}_T), \mathrm{pr}_{[k]}(\boldsymbol{\xi}_T)),$$

we get that

$$\mathrm{pr}_{2\#}\widetilde{\mathrm{pr}}_{[k]\#}\pi = \mathrm{pr}_{[k]\#}\mathrm{pr}_{2\#}\pi = \mathrm{pr}_{[k]\#}P \tag{4.7.1a}$$

$$\mathrm{pr}_{1\#}\widetilde{\mathrm{pr}}_{[k]\#}\pi = \mathrm{pr}_{[k]\#}\mathrm{pr}_{1\#}\pi = (P_w^N)^{\otimes k}, \tag{4.7.1b}$$

where pr_1 and pr_2 are applied with $\Xi \equiv \mathscr{W}^k$ and $\Xi \equiv \mathscr{W}^T$ in the left- and right-hand-side of the respective equalities. Denoting also $\varphi_q((\boldsymbol{\zeta}_k, \boldsymbol{\xi}_k)) := \|\zeta_q - \xi_q\|$, $q \in [k]$, we have

$$\begin{aligned}
&\int_{\mathscr{W}^k \times \mathscr{W}^k} \|\zeta_q - \xi_q\| \, d(\widetilde{\mathrm{pr}}_{[k]\#}\pi)(\boldsymbol{\zeta}_k, \boldsymbol{\xi}_k) \\
&\quad \equiv \int_{\mathscr{W}^k \times \mathscr{W}^k} \varphi_q((\boldsymbol{\zeta}_k, \boldsymbol{\xi}_k)) d(\widetilde{\mathrm{pr}}_{[k]\#}\pi)(\boldsymbol{\zeta}_k, \boldsymbol{\xi}_k) \\
&\quad = \int_{\mathscr{W}^T \times \mathscr{W}^T} \varphi_q \circ \mathrm{pr}_2((\boldsymbol{\zeta}_k, \boldsymbol{\xi}_k)) d\pi(\boldsymbol{\zeta}_T, \boldsymbol{\xi}_T) \\
&\quad = \int_{\mathscr{W}^T \times \mathscr{W}^T} \|\zeta_q - \xi_q\| d\pi(\boldsymbol{\zeta}_T, \boldsymbol{\xi}_T) \leq \varepsilon_k
\end{aligned}$$

for all $q \in [k]$. Together with (4.7.1), this yields the desired inclusion $\mathrm{pr}_{[k]\#}P \in \mathscr{T}((P_w^N)^{\otimes k}, \boldsymbol{\varepsilon}_k)$.

Conversely, let $P' \in \mathscr{T}((P_w^N)^{\otimes k}, \boldsymbol{\varepsilon}_k)$ and consider a coupling $\pi' \in \mathscr{P}(\mathscr{W}^k \times \mathscr{W}^k)$ such that the transport constraints in (4.5.2) are satisfied and $P' = \mathrm{pr}_{2\#}\pi'$. Let also $\pi_{\mathrm{Id}}(P_w^{N\otimes(T-k)}) \in \mathscr{P}(\mathscr{W}^{T-k}, \mathscr{W}^{T-k})$ be the identity coupling associated to the distribution $P_w^{N\otimes(T-k)}$ (as defined in the preliminaries section). Then we have that $\pi' = \widetilde{\mathrm{pr}}_{[k]\#}\pi$ and it is not hard to check that $\mathrm{pr}_{2\#}\pi \in \mathscr{T}((P_w^N)^{\otimes T}, \boldsymbol{\varepsilon}_T)$ and

$$P' = \mathrm{pr}_{2\#}\widetilde{\mathrm{pr}}_{[k]\#}\pi = \mathrm{pr}_{[k]\#}\mathrm{pr}_{2\#}\pi.$$

Namely we have $P' \in \mathrm{pr}_{[k]\#}\mathscr{T}((P_w^N)^{\otimes T}, \boldsymbol{\varepsilon}_T)$, which completes the proof. □

4

The next corollary utilizes the projection of the hyperrectangle over the horizon to its initial fragments, to obtain equivalent robust CVaR expressions over the projected ambiguity sets in the desired chance constraints.

Corollary 4.7.2. *(Telescopic representation of* CVaR *expressions). With ε_T and ε_k as in Proposition 4.7.1, we have*

$$\begin{aligned}&\sup_{P\in\mathscr{T}((P_w^N)^{\otimes T},\boldsymbol{\varepsilon}_T)} \mathrm{CVaR}^P_{1-p_j}(\langle \boldsymbol{G}_k^\top H_j^\top, \boldsymbol{\xi}_T\rangle + \gamma - h_j)\\ &= \sup_{P\in\mathscr{T}((P_w^N)^{\otimes k},\boldsymbol{\varepsilon}_k)} \mathrm{CVaR}^P_{1-p_j}(\langle \widetilde{\boldsymbol{G}}_k^\top H_j^\top, \boldsymbol{\xi}_k\rangle + \gamma - h_j)\end{aligned} \tag{4.7.2}$$

for all $k \in [T]$, where

$$\widetilde{\boldsymbol{G}}_k := \begin{bmatrix}(A+BK)^{k-1}\cdots A+BK & I_n\end{bmatrix}.$$

Proof. We have

$$\begin{aligned}&\sup_{P\in\mathscr{T}((P_w^N)^{\otimes T},\boldsymbol{\varepsilon}_T)} \mathrm{CVaR}^P_{1-p_j}(\langle \boldsymbol{G}_k^\top H_j^\top, \boldsymbol{\xi}_T\rangle + \gamma - h_j)\\ &= \sup_{P\in\mathscr{T}((P_w^N)^{\otimes T},\boldsymbol{\varepsilon}_T)} \mathrm{CVaR}^P_{1-p_j}(\langle \widetilde{\boldsymbol{G}}_k^\top H_j^\top, \mathrm{pr}_{[k]}\boldsymbol{\xi}_T\rangle + \gamma - h_j),\end{aligned}$$

which in turn is equal to the right-hand side of (4.7.2) by Proposition 4.7.1. □

Remark 4.7.3. (Prefix projection of ambiguity ball). Taking into account that Wasserstein ambiguity balls are multi-transport hyperrectangles with a single optimal transport constraint, we also get that

$$\mathrm{pr}_{[k]\#}\mathscr{B}(P^N_{\boldsymbol{\xi}_T}, \varepsilon) = \mathscr{B}(\mathrm{pr}_{[k]\#}P_{P^N_{\boldsymbol{\xi}_T}}, \varepsilon).$$

According to Corollary 4.7.2, we can equivalently replace the CVaR expression in (4.6.1) with that of (4.7.2). This shows that we can drastically reduce the computational complexity of the chance constraints at the initial steps of the prediction horizon. The approach hinges on using a different hyperrectangle for each step, which yields the same feasible set as before and retains the desirable probabilistic guarantees.

While the consideration of separate hyperrectangles along the steps of the prediction horizon resolves the complexity of the initial steps, the complexity associated with the optimization problems of the final steps persists. To address this issue, we cluster the reference models $(P_w^N)^{\otimes k}$ for larger $k \in [T]$. These clustered distributions share the benefits of having a lower complexity than $(P_w^N)^{\otimes k}$ and being typically closer to the true distribution $P_w^{\otimes k}$ than empirical models that only rely on independent trajectories of $\boldsymbol{\xi}_k$ (see e.g., [106]).

We consider multiple clustering strategies in accordance with the work in Section 3.8 of Chapter 3. The first consists of directly clustering the points on which $(P_w^N)^{\otimes k}$ is supported into a discrete distribution $\widehat{P}$ with a smaller number of points. The other strategies rely on a two-step approach that exploits the structure of the reference distribution across the prediction horizon. The most direct approach along these lines is to obtain clustered versions $\widehat{P}_1, \dots, \widehat{P}_k$ of the empirical distribution P_w^N of the noise and form their product $\widehat{P}_1 \otimes \cdots \otimes \widehat{P}_k$ as a lower-complexity reference model. There is also a middle ground, where one can decompose $(P_w^N)^{\otimes k}$ into lower-dimensional products $P_w^{N \otimes m_1}, \dots, P_w^{N \otimes m_\ell}$ with $m_1 + \cdots + m_\ell = k$, cluster them, and then take the product of these clusters. A key motivation to consider two-step clustering strategies for large k is that the mere enumeration of the atoms in $(P_w^N)^{\otimes k}$ may be prohibitive to directly cluster the distribution.

To derive clusters with an optimal Wasserstein discrepancy from $(P_w^N)^{\otimes k}$, we use Lloyd's algorithm, which is guaranteed to converge to local minima and whose gradient step complexity is linear in the number of points to be clustered. The clustering process is adapted to the prediction step k by adjusting the complexity of lower-dimensional clusters and the choice to cluster in a monolithic or component-wise fashion with the step of the horizon. For instance, it is favorable to consider lower-dimensional clusters with a larger number of points during the initial time steps and reduce this number as time progresses to keep the size of their product under control. Such a choice can be further refined for larger k by exploiting coarser clusterings for the initial components of $(P_w^N)^{\otimes k}$ and finer clusterings for its final components. This is justified by the fact that closed-loop stability of the error dynamics suppresses the effect of initial disturbances on later time steps of the horizon.

By adjusting the reference distribution of the hyperrectangles, the probabilistic guarantees of Proposition 4.5.1 do not necessarily hold anymore. Yet, we show in Proposition 3.8.1 of Chapter 3 that these probabilistic guarantees can be recovered by appropriately adjusting the transport budget of $\mathcal{T}((P_w^N)^{\otimes k}, \boldsymbol{\varepsilon}_k)$ in each direction. Indeed, this result quantifies the extra transport budget that is required to retain the probabilistic guarantees when shifting the multi-transport hyperrectangle center from the product empirical distribution $(P_w^N)^{\otimes k}$ to the clustered distribution $\widehat{P}$. Following Proposition 3.8.1, we can opt for multiple clustering strategies, either by directly clustering the atoms of the product empirical distribution $(P_w^N)^{\otimes k}$, or by clustering its marginals P_w^N and then considering the product of the clustered marginals. Using the projected hyperrectangles reduces the conservatism of the clustering for the first prediction steps. Namely, when the number of clusters is fixed, $(P_w^N)^{\otimes k}$ is typically closer to its clustered version in the Wasserstein metric

for smaller k, see Corollary 3.B.5 of Chapter 3. At the same time, stability of the closed-loop error dynamics enables the designer to perform a heavy clustering of the initial components of $(P_w^N)^{\otimes k}$ while keeping its complexity under control. As a result, it is possible in practice to exploit ambiguity sets that avoid the conservativeness associated with the curse of dimensionality and still have a manageable complexity.

4.8. RECURSIVE FEASIBILITY

In this section, we discuss strategies that guarantee the recursive feasibility of the MPC algorithm. A related subtlety in stochastic MPC is linked to the fact that the probability of violating the constraints at the kth predicted step of time t given x_t differs from the same violation probability at the $k-1$th predicted step of time $t+1$ given x_{t+1}. In particular (cf. [195]), we have

$$Hz_{k|t} \preceq \widetilde{h}_k \not\Rightarrow Hz_{k-1|t+1} \preceq \widetilde{h}_{k-1},$$

which implies that the tightening of Proposition 4.4.1 is not enough to ensure recursive feasibility. This happens because there may be realizations of x_{t+1} for which feasibility of the optimization problem is no longer guaranteed.

To address this issue, we impose extra constraints on the nominal state trajectory across the prediction horizon. We follow the approach in [195], which guarantees recursive feasibility by imposing extra constraints on the first prediction step of the MPC formulation. This approach has the advantage of being less conservative than other tightenings used for the same purpose, such as the one in [203]. To this end, we first introduce the set of feasible states

$$\begin{aligned}\mathscr{Z}_{\text{feasible}} := \{z_0 \in \mathbb{R}^n : \text{there exist } v_0, \dots, v_{T-1} \in \mathbb{R}^m \text{ s.t.} \\ z_{k+1} = Az_k + Bv_k \text{ for all } k \in [T-1]_0, \\ z_k \in \mathscr{Z}_k \text{ for all } k \in [T], \\ \text{and } v \in \mathscr{U} \ominus K\mathscr{E}_k \text{ for all } k \in [T-1]_0\},\end{aligned}$$

where the nominal constraint sets $\mathscr{Z}_k$ are given by (4.4.6) with the thresholds $\widetilde{h}$ as determined by (4.4.9). The computation of $\mathscr{Z}_{\text{feasible}}$ can be performed using the approaches proposed in [204]. We also consider its associated maximum robust controlled invariant subset $\mathscr{Z}_{\text{rc}}$, which satisfies

$$\begin{aligned}z \in \mathscr{Z}_{\text{rc}} \subset \mathscr{Z}_{\text{feasible}} \Longrightarrow \\ Az + Bu + \mathscr{W} \subset \mathscr{Z}_{\text{rc}} \text{ for some } u \in \mathscr{U}\end{aligned}$$

and is maximal with respect to this property. Then it follows by the same arguments as in [195, Proposition 4] that the tightened MPC formulation

$$\begin{aligned}&\min_{\boldsymbol{v},\boldsymbol{c}} \textstyle\sum_{k=0}^{T-1} \|z_{k|t}\|_Q^2 + \|v_{k|t}\|_R^2 && && (4.8.1)\\ &\text{s.t. } z_{k+1|t} = Az_{k|t} + Bv_{k|t} && k \in [T-1]_0 \\ &\quad v_{k|t} = c_{k|t} + Kz_{k|t} && k \in [T-1]_0\end{aligned}$$

$$
\begin{aligned}
&z_{k|t} \in Z_k && k \in [T-1] \\
&v_{k|t} \in \mathscr{U} \ominus K\mathscr{E}_k && k \in [T-1]_0 \\
&z_{1|t} \in \mathscr{Z}_{\mathrm{rc}} \ominus \mathscr{W} \\
&z_{0|t} = x_t
\end{aligned}
$$

is recursively feasible.

4.9. SIMULATION EXAMPLE

In this section, we exploit the distributionally robust CVaR reformulation together with the clustering scheme to evaluate the constraint sets and feasible set of the model predictive controller (4.4.11) when (4.3.1) is the double integrator system with

$$A := \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad B := \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}.$$

We consider the following admissible sets for the state and the control input

$$
\begin{aligned}
\mathscr{X} &:= \{x \in \mathbb{R}^2 : (0,-1) \leq x \leq (5,3)\} \\
&\equiv \{x \in \mathbb{R}^2 : Hx \leq h\}, \\
\mathscr{U} &:= \{u \in \mathbb{R} : -3.1 \leq u \leq 3.1\}.
\end{aligned} \tag{4.9.1}
$$

We also enforce the chance constraints

$$\mathbb{P}(x \in \mathscr{X} | x_t) \geq 1 - p_j,\ j \in [J],\ t \in \mathbb{N},$$

while robust constraints are imposed on the control input u, i.e., $u \in \mathscr{U}$ for all $w \in \mathscr{W}$. For the design of the MPC controller, we select the cost matrices $Q = I_2$ and $R = 1$ and select the feedback gain K by solving the LQR problem with weights $Q_{\mathrm{LQR}} = I_2$ and $R_{\mathrm{LQR}} = 1$.

We consider the product distribution $P_w = P_{w_1} \otimes P_{w_2}$ for the noise, with $P_{w_1} := \mathscr{U}(-10^{-5}, 10^{-5})$ and $P_{w_2} := 0.7\mathscr{U}(0, 0.005) + 0.3\mathscr{U}(2, 2.5)$, where $\mathscr{U}(a,b)$ denotes the uniform distribution on the closed interval $[a,b]$. We also assume the distribution of w is not known and we only have access to N i.i.d. samples of it. Using these samples, we build multi-transport hyperrectangles for the disturbance trajectory over the horizon and compute the corresponding tightenings of the desired chance constrains. We compare the results with the case where the ambiguity set is a Wasserstein ball centered at the empirical distribution of the disturbance trajectory, as for instance in [106].

For the simulations, we set the probability of satisfaction to 0.95 for each constraint, the prediction horizon length to $T = 8$, and consider $N = 56$ i.i.d. samples of w. Exploiting those samples, we construct the empirical distribution $P^7_{\boldsymbol{\xi}_8} = \frac{1}{7}\sum_{i=1}^{7} \delta_{\boldsymbol{\xi}_8^i}$ of $\boldsymbol{\xi}_8 := (w_0, \ldots, w_7)$ using 7 independent disturbance trajectories and the empirical distribution $P^{56}_w = \frac{1}{56}\sum_{i=1}^{56} \delta_{w^i}$ of w. Based on those two reference distributions, we construct the ambiguity sets $\mathscr{B}(P^7_{\boldsymbol{\xi}_8}, \varepsilon)$ and $\mathscr{T}(\widehat{P}^{56|k}_w, \boldsymbol{\varepsilon})$, $k \in [8]$, where each $\widehat{P}^{56|k}_w$ is a

clustered approximation of the product distribution $(P_w^{56})^{\otimes k}$. To this end, we cluster P_w^{56} into the discrete distributions $\{\widehat{P}_q\}_{q\in\{1,3,5,7\}}$ with 1, 3, 5, and 7 atoms, respectively and build the reference distributions

$$\begin{aligned}
\widehat{P}_w^{56|k} &:= \widehat{P}_7^{\otimes k}, \quad k \in [3]\\
\widehat{P}_w^{56|4} &:= \widehat{P}_3 \otimes \widehat{P}_7^{\otimes 3}\\
\widehat{P}_w^{56|k} &:= \widehat{P}_1^{\otimes (k-4)} \otimes \widehat{P}_3 \otimes \widehat{P}_7^{\otimes 3}, \quad k \in [5:7]\\
\widehat{P}_w^{56|8} &:= \widehat{P}_1^{\otimes 4} \otimes \widehat{P}_5 \otimes \widehat{P}_7^{\otimes 3},
\end{aligned}$$

whose clustering is adapted along the prediction horizon.

Our goal is to determine the minimum size of each type of ambiguity set, so that at every instant of the prediction horizon, the corresponding CVaR constraints are satisfied with prescribed confidence. Thereafter, we also seek to compare the associated constraint tightenings. To this end, we compute the tightened constraints $\tilde{h}_k^{\text{ball}}$ and $\tilde{h}_k^{\text{rect}}$, $k \in [8]$ for the Wasserstein ball using Corollary 4.6.2 with $\mathscr{T}(\widehat{P}, \boldsymbol{\varepsilon}) \equiv \mathscr{B}(P_{\boldsymbol{\xi}_8}^7, \varepsilon)$, and using Corollaries 4.6.2 and 4.7.2 with $\mathrm{pr}_{k\#}\mathscr{T}(\widehat{P}, \boldsymbol{\varepsilon}) \equiv \mathscr{T}(\widehat{P}_w^{56|k}, \boldsymbol{\varepsilon}_k)$ for the corresponding multi-transport hyperrectangles. We solve the CVaR-constrained problems for increasing values ε of the Wasserstein ball and the smallest radius of the ball enclosing the hyperrectangles. For each ε, we compute the relative frequency with which the CVaR constraints are met at each individual step of the horizon, using 100 realizations of $N = 56$ samples of w. We increase ε until this frequency reaches the desired confidence level, which we set to 0.8. For this smallest ε that guarantees the desired constraint satisfaction, we compute the average tightening over the 100 reference distributions for every ambiguity set and determine the resulting feasible state sets of the MPC.

Although our choice of ambiguity set size is not so well aligned with our motivation to have simultaneous probabilistic guarantees across the horizon, it provides a simple empirical tuning based on the constraints at each time step. In particular, since the clustering of the hyperrectangle centers yields enclosing balls of different size across the horizon, it is less clear how to select their individual radii simultaneously to obtain a collective constraint satisfaction guarantee. To have a fair practical comparison between structured ambiguity sets and balls, we tune the radius of $\mathscr{B}(P_{\boldsymbol{\xi}_8}^7, \varepsilon)$ at each time step in the same data-driven way. Nevertheless, in this case the tuning can also be performed unambiguously by handling all constraints across the horizon simultaneously.

We depict the selected size ε of each ambiguity set across the prediction horizon in Figure 4.1. For each $k \in [8]$, the radius of the smallest ball enclosing the set $\mathscr{T}(\widehat{P}_w^{56|k}, \boldsymbol{\varepsilon}_k)$ is significantly smaller than that of the corresponding ball $\mathscr{B}(P_{\boldsymbol{\xi}_8}^7, \varepsilon)$ to achieve the desired 80% confidence level of CVaR constraint satisfaction. In addition, while the size of the monolithic balls remains at the same levels, the size of the hyperrectangles increases noticeably throughout the prediction horizon. This happens because the increasing number of lower-dimensional components of $\boldsymbol{\xi}_k$ with respect to the prediction step yields an exponential growth of the atoms in the corresponding product empirical distributions. Since there is a limit on the number

of clusters to approximate these models, the accuracy with which they approximate the true distribution of the noise sequence deteriorates father in the prediction horizon.

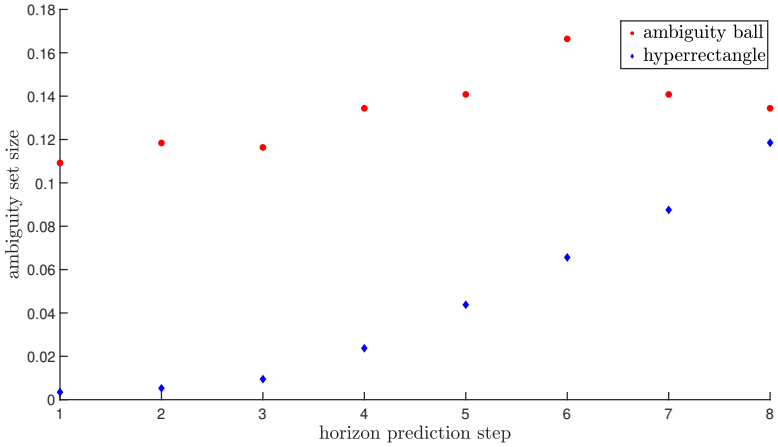


Figure 4.1.: The figure depicts the minimum size necessary for every ambiguity set in order to induce tightenings of the constraints that satisfy the desired CVaR level with a confidence of 80%. The results show that on every prediction step hyperrectangles need a smaller size to ensure the desired confidence level compared to their traditional counterpart.

Figure 4.2 shows the average bounds on the components of the nominal state z across the prediction horizon. These bounds are computed by solving the corresponding CVaR constrained problems for the ambiguity ball and the multi-transport hyperrectangle we determined for each prediction step k. We also show the same bounds that yield the desired CVaR constraints for the true distribution. Clearly, the bounds for the ambiguity ball are much more conservative compared to the ones of the corresponding hyperrectangle for all $k \in [2:8]$. Figure 4.3 depicts the corresponding feasible set of the initial nominal sate that is introduced in Section 4.8. We observe that the hyperrectangle yields an accurate approximation of the feasible set (in blue), which is close to the one obtained with the true distribution (in yellow) and much larger than the one obtained with the ball (in red). These results confirm that accounting for the structure of the uncertainty enhanced the performance of distributionally robust tightening. By appropriately clustering the reference models of the associated ambiguity sets, we obtained reformulations of a manageable computational complexity that yielded considerably improved constraint sets.

4.10. CONCLUSION

In this chapter, we introduce a recursively feasible model predictive control algorithm that can enforce probabilistic constraints, with high confidence, while inducing an online complexity similar to that of a nominal MPC algorithm. Nominal

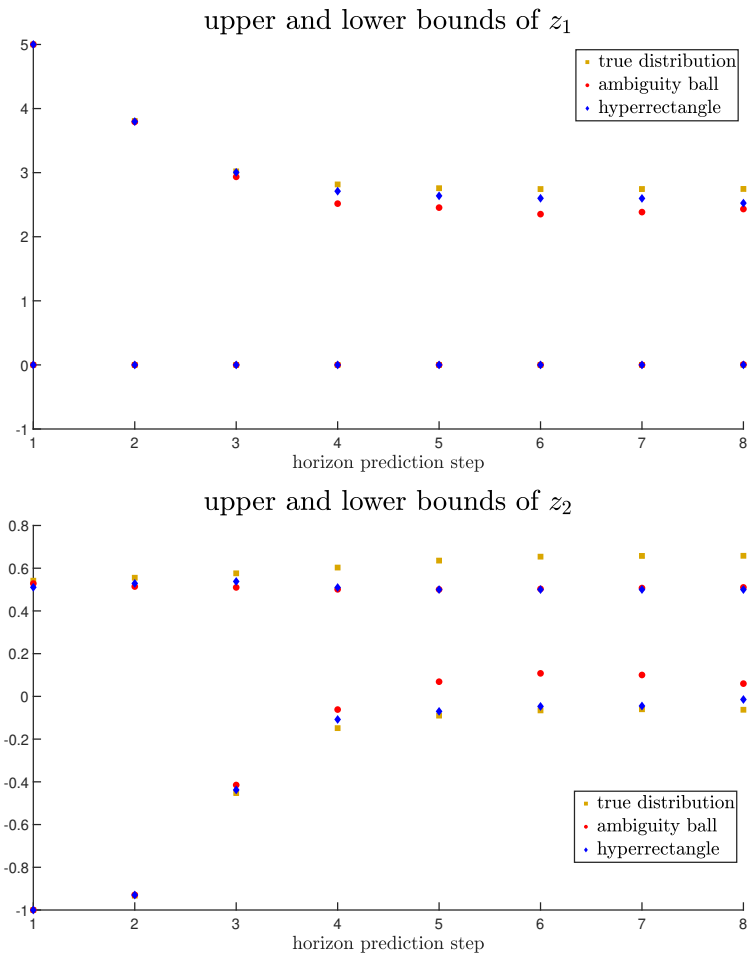


Figure 4.2.: The figure depicts the average bounds on the nominal state z across the prediction horizon. These are deduced by computing the average solution over 100 tries of the distributionally robust CVaR problems with the ambiguity sets $\mathscr{B}(P^7_{\boldsymbol{\xi}_8}, \varepsilon)$ and $\mathscr{T}(\widehat{P}^{56|k}_w, \boldsymbol{\varepsilon})$, respectively, for $k \in [8]$. The size ε and $\boldsymbol{\varepsilon}$ of each ambiguity set is set to the values depicted in Fig. 4.1 for every prediction step $k \in [8]$.

constraints are determined by solving, offline, a linear program equivalent to a distributionally robust CVaR constrained optimization problem. An adaptive clustering scheme is employed to address the potential numerical complexity of the tractable reformulations. These results yield a computationally efficient tool for data-driven predictive control under distributional uncertainty with potential applications to, e.g., autonomous driving and robotics. Future work includes the determination of the characterizations ensuring closed-loop system stability and the quantification of an asymptotic average performance bound.

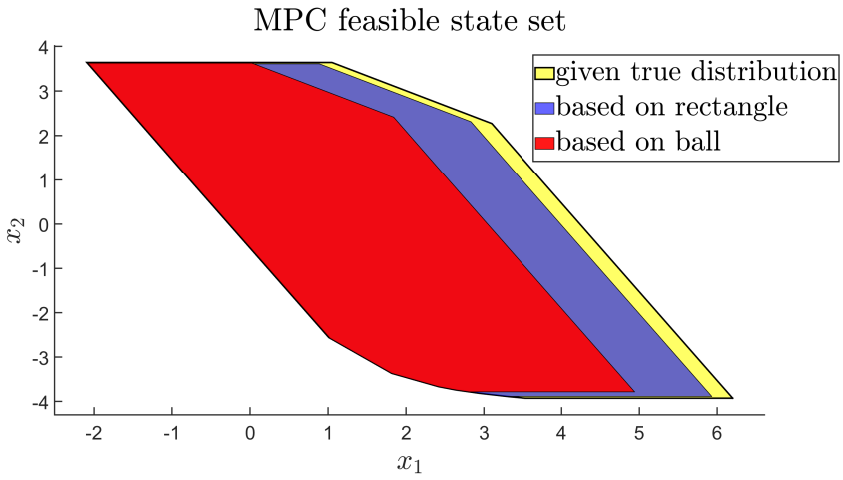


Figure 4.3.: The figure depicts the feasible sets for the model predictive controller in the three cases: (yellow) when the true distribution is known, (blue) using the multi-transport hyperrectangle with the proposed clustering scheme, and (red) using the monolithic ambiguity ball. The three feasible state sets are deduced from the constraint sets in Fig. 4.2.

5

CONCLUSION

In this thesis, we address the curse of dimensionality characterizing Wasserstein ambiguity sets by exploiting the independence structure of random variables, thereby contributing to the field of distributionally robust optimization. In this final chapter, we provide a summary of the contributions of this work and explore possible directions for future research.

THIS thesis contributes to the field of distributionally robust optimization by addressing the curse of dimensionality inherent in data-driven Wasserstein ambiguity sets. The contributions of the thesis enable the modeling of probability distributions based on a limited number of samples, allowing for reliable decision-making in the presence of uncertainty. We address the conservativeness of Wasserstein ambiguity balls by restricting our attention to the case where multiple independent random variables are involved. The main idea behind the results is to integrate the underlying independence between the random variable components into the structure of optimal transport ambiguity sets. In addition to enjoying a faster contraction rate with respect to the number of samples, these ambiguity sets can improve decisions, especially when multiple optimization problems are involved. We next discuss how the objectives set at the beginning of the thesis have been achieved.

5.1. DISCUSSION OF THE RESEARCH OBJECTIVES ACHIEVED

O1 • **Equipping data-driven ambiguity sets with the independence structure of the data-generating distribution**

In order to capture product distributions, we encoded their product structure into optimal transport ambiguity sets by considering multiple transport constraints. Each transport constraint informs one marginal of the distribution. This leads to two distinct classes of structured ambiguity sets: Wasserstein hyperrectangles and multi-transport hyperrectangles. The first class only contains product distributions. This is achieved by considering a distinct transport plan for each component of the uncertainty. Multi-transport hyperrectangles, on the other hand, are encoded by a set of transport constraints that we impose on the same transport plan. This results in a convex overapproximation of Wasserstein hyperrectangles that does not contain many more distributions.

O2 • **Establishing statistical guarantees that these structured ambiguity sets contain the true distribution with high confidence, and determining how their shrinkage rate improves with the number of samples**

Exploiting the independence between the components of the random variable, we showed that Wasserstein hyperrectangles contain the data-generating distribution iff their lower-dimensional Wasserstein ambiguity balls contain its marginals. We then establish that multi-transport hyperrectangles contain the true distribution with the same confidence level, since they form an overapproximation of Wasserstein hyperrectangles. This overapproximation is tight in the sense that both sets share exactly the same product distributions.

We also showed that these sets shrink at a much faster rate with respect to the number of collected samples compared to their traditional counterpart. This shrinkage rate increases with the number of independent components of the random variable. The geometry of the structured

ambiguity sets captures the heterogeneity between the independent components of the uncertainty, enabling a more effective allocation of the transport budget across the transport constraints to capture each marginal distribution. In addition, despite being an overapproximation of Wasserstein hyperrectangles, multi-transport hyperrectangles enjoy the same size reduction properties characterizing the former structured ambiguity set.

O3 • **Derivation of dual reformulations of DRO problems associated with structured ambiguity sets**

We derived dual formulations of DRO problems associated with structured ambiguity sets. These enabled the conversion of infinite-dimensional optimization problems into finite-dimensional ones, which is a stepping stone to also obtain tractable reformulations of these problems. The dual reformulations derived for Wasserstein hyperrectangles only hold for a very narrow class of loss functions, due to their non-convexity. On the other hand, the class of loss functions for which the dual reformulations hold for multi-transport hyperrectangles is much larger.

O4 • **Derivation of tractable reformulations for DRO problems associated with multi-transport hyperrectangles**

Exploiting the derived dual reformulations, we provided tractable reformulations that enable the effective computation of worst-case expectations, the solution of uncertainty quantification problems, and the derivation of convex relaxations for distributionally robust chance-constrained problems.

O5 • **Identification of computational limitations inherent to multi-transport hyperrectangles and their mitigation**

When the uncertainty admits independent components, multi-transport hyperrectangles effectively mitigate the curse of dimensionality that affects Wasserstein data-driven ambiguity sets. To achieve this, multi-transport hyperrectangles need to be centered at a product empirical distribution, whose number of atoms grows exponentially with the number of independent components. This poses a significant limitation, since the resulting reformulations involve a number of variables and constraints proportional to the number of atoms of the distribution from which the hyperrectangle is inferred. To overcome this issue, we developed clustering schemes based on K-means clustering that approximate the product empirical distribution by a distribution supported on a much smaller number of atoms, and we adapted the size of the hyperrectangle in order to maintain its probabilistic confidence of containing the data-generating distribution. A theoretical analysis is also presented to determine conditions ensuring that the hyperrectangle centered at the clustered distribution still enjoys advantageous size reduction properties compared to traditional Wasserstein ambiguity balls.

O6 • **Exploiting multi-transport hyperrectangles to robustify data-driven decision problems in systems and control**

We formulated a distributionally robust model predictive controller that enforces distributionally robust CVaR constraints for linear systems affected by additive stochastic disturbances with an unknown distribution. We based the controller on a tube approach and enforced the constraints through suitable tightenings of the admissible state and control input sets. These tightenings are determined by solving multiple distributionally robust optimization problems. We exploited multi-transport hyperrectangles, which are structured accordingly to the independence between the realizations of the uncertainty to capture the disturbance trajectory distribution with reduced conservativeness. This enabled the determination of tighter polytopic regions that contain the propagation error of the state with desired probability and prescribed confidence. We also ensure the recursive feasibility of the proposed MPC scheme by introducing an extra constraint over the first-step-ahead state of the prediction horizon.

5

5.2. FUTURE RESEARCH DIRECTIONS

The thesis presents contributions to decision-making in the presence of multiple independent uncertainties by incorporating independence into the structure of optimal transport ambiguity sets. The results achieved can be exploited in diverse fields, such as statistics, optimization, control, and dynamic programming. In this section, we highlight some of the research gaps and open questions that could represent potential directions for future work, building on the work achieved in this thesis.

- **Determination of transport-information conditions characterize family of distributions with informative clustered product estimator**

  To reduce the complexity characterizing product empirical distributions, we resort to clustering schemes to deduce simpler baseline probabilistic models. However, clustering the product empirical distribution generally yields a less accurate estimate of the true distribution. This is reflected by an increased Wasserstein distance between the new reference model and the true distribution, necessitating an additional transport budget to retain statistical guarantees of containing the true distribution. In this work, we identify a condition based on a minimum number of clusters that ensures a more favorable decay rate of the bounds for hyperrectangles in order to contain the true distribution with high confidence. This poses the challenge of identifying families of distributions whose product empirical estimators are still good approximations of the true distribution after being clustered. A possible approach to characterize such families of distributions would be to exploit transport information inequalities.

- **Application to stochastic dynamic programming**

An instance involving multiple optimization problems is dynamic programming. Since optimal transport hyperrectangles are designed to capture the true distribution with high confidence while mitigating the curse of dimensionality, they may serve as a valuable tool for distributionally robust dynamic programming problems with structured uncertainty. A notable application is distributionally robust LQG control.

- **Enrichment of the results related to model predictive control**

  The model predictive control algorithm developed in this thesis is recursively feasible; however, its stability has not yet been established. Therefore, we aim to prove the stochastic stability of the closed-loop system with the model predictive controller proposed in Chapter 4. In addition, we seek to derive distributionally robust bounds on the closed-loop system performance and derive direct tractable reformulations of the chance constraints without resorting to convex relaxations, i.e., without relying on the CVaR.

- **Derivation of statistical guarantees for uncertainty components with a known structural dependency**

  The statistical guarantees of hyperrectangles containing the data-generating distribution hold only when the uncertainty has independent components. A promising line of research would be to extend this framework to random variables that exhibit a known dependency structure among their components. This would allow addressing the curse of dimensionality of Wasserstein ambiguity sets that contain the data-generating distribution with prescribed probability across a broader range of applications.

BIBLIOGRAPHY

[1] R. Pasquier and I. F. C. Smith. “Robust system identification and model predictions in the presence of systematic uncertainty”. In: *Advanced Engineering Informatics* 29.4 (2015), pp. 1096–1109.

[2] P. Shi. “Filtering on sampled-data systems with parametric uncertainty”. In: *IEEE Transactions on Automatic Control* 43.7 (1998), pp. 1022–1027.

[3] J. C. Geromel. “Optimal linear filtering under parameter uncertainty”. In: *IEEE Transactions on Signal processing* 47.1 (1999), pp. 168–175.

[4] A. Bargiela, W. Pedrycz, and M. Tanaka. “A study of uncertain state estimation”. In: *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 33.3 (2003), pp. 288–301.

[5] D. Bertsekas and I. Rhodes. “Recursive state estimation for a set-membership description of uncertainty”. In: *IEEE Transactions on Automatic Control* 16.2 (1971), pp. 117–128.

[6] I. R. Petersen and A. V. Savkin. *Robust Kalman filtering for signals and systems with large uncertainties*. Springer Science & Business Media, 1999.

[7] L. Xie and Y. C. Soh. “Robust Kalman filtering for uncertain systems”. In: *Systems & Control Letters* 22.2 (1994), pp. 123–129.

[8] M. Quincampoix and V. M. Veliov. “Optimal control of uncertain systems with incomplete information for the disturbances”. In: *SIAM journal on control and optimization* 43.4 (2004), pp. 1373–1399.

[9] A. Barkefors, M. Sternad, and L. J. Brännmark. “Design and analysis of linear quadratic Gaussian feedforward controllers for active noise control”. In: *IEEE/ACM transactions on audio, speech, and language processing* 22.12 (2014), pp. 1777–1791.

[10] S. B. Choi, S. R. Hong, K. G. Sung, and J. W. Sohn. “Optimal control of structural vibrations using a mixed-mode magnetorheological fluid mount”. In: *International Journal of Mechanical Sciences* 50.3 (2008), pp. 559–568.

[11] P. J. Hargrave. “A tutorial introduction to Kalman filtering”. In: *IEE Colloquium on Kalman Filters: Introduction, Applications and Future Developments*. 1989.

[12] A. Mesbah. “Stochastic Model Predictive Control: An Overview and Perspectives for Future Research”. In: *IEEE Control Systems Magazine* 36.6 (2016), pp. 30–44.

[13] N. Mackintosh and N. Armstrong. “Understanding and managing uncertainty in health care: revisiting and advancing sociological contributions”. In: *Sociology of Health & Illness* 42 (2020), pp. 1–20.

[14] A. Soroudi and T. Amraee. “Decision making under uncertainty in energy systems: State of the art”. In: *Renewable and Sustainable Energy Reviews* 28 (2013), pp. 376–384.

[15] G. Mavromatidis, K. Orehounig, and J. Carmeliet. “Uncertainty and global sensitivity analysis for the optimal design of distributed energy systems”. In: *Applied Energy* 214 (2018), pp. 219–238.

[16] S. M. Mohseni-Bonab and A. Rabiee. “Optimal reactive power dispatch: a review, and a new stochastic voltage stability constrained multi-objective model at the presence of uncertain wind power generation”. In: *IET Generation, Transmission & Distribution* 11.4 (2017), pp. 815–829.

[17] H. Jung and M. Pedram. “Dynamic power management under uncertain information”. In: *2007 Design, Automation & Test in Europe Conference & Exhibition*. IEEE. 2007, pp. 1–6.

[18] K. Park, H. G. Jung, T. S. Eom, and S. W. Lee. “Uncertainty-aware portfolio management with risk-sensitive multiagent network”. In: *IEEE Transactions on Neural Networks and Learning Systems* 35.1 (2022), pp. 362–375.

[19] L. Davidson. “Uncertainty in economics”. In: *Uncertainty, International Money, Employment and Theory: Volume 3: The Collected Writings of Paul Davidson*. Springer, 1999, pp. 30–37.

[20] E. Simangunsong, L. C. Hendry, and M. Stevenson. “Supply-chain uncertainty: a review and theoretical foundation for future research”. In: *International journal of production research* 50.16 (2012), pp. 4493–4523.

[21] A. Gupta and C. D. Maranas. “Managing demand uncertainty in supply chain planning”. In: *Computers & chemical engineering* 27.8-9 (2003), pp. 1219–1227.

[22] R. C. Gilliam, C. Hogrefe, J. M. Godowitch, S. Napelenok, R. Mathur, and S. T. Rao. “Impact of inherent meteorology uncertainty on air quality model predictions”. In: *Journal of Geophysical Research: Atmospheres* 120.23 (2015), pp. 12–259.

[23] S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[24] A. Ben-Tal and A. Nemirovski. “Robust convex optimization”. In: *Mathematics of operations research* 23.4 (1998), pp. 769–805.

[25] A. Ben-Tal, A. Nemirovski, and L. E. Ghaoui. “Robust optimization”. In: (2009).

[26] D. Bertsimas and M. Sim. “The price of robustness”. In: *Operations research* 52.1 (2004), pp. 35–53.

[27] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

[28] P. Billingsley. *Probability and measure*. John Wiley & Sons, 2017.

[29] S. Ahmed and A. Shapiro. “Solving chance-constrained stochastic programs via sampling and integer programming”. In: *State-of-the-art decision-making tools in the information-intensive age*. Informs, 2008, pp. 261–269.

[30] W. Wang and S. Ahmed. “Sample average approximation of expected value constrained stochastic programs”. In: *Operations Research Letters* 36.5 (2008), pp. 515–519.

[31] H. Rahimian and B. Pagnoncelli. “Contextual Stochastic Programs with Expected-Value Constraints”. In: *Optimization Online* (2024).

[32] D. Kuhn. “Convergent bounds for stochastic programs with expected value constraints”. In: *Journal of optimization theory and applications* 141.3 (2009), pp. 597–618.

[33] P. Li, H. Arellano-Garcia, and G. Wozny. “Chance constrained programming approach to process optimization under uncertainty”. In: *Computers & chemical engineering* 32.1-2 (2008), pp. 25–45.

[34] W. Ogryczak and A. Ruszczyński. “From stochastic dominance to mean-risk models: Semideviations as risk measures”. In: *European journal of operational research* 116.1 (1999), pp. 33–50.

[35] D. S. Kalogerias and W. B. Powell. “Recursive optimization of convex risk measures: Mean-semideviation models”. In: *arXiv preprint arXiv:1804.00636* (2018).

[36] R. T. Rockafellar and S. Uryasev. “Conditional value-at-risk for general loss distributions”. In: *Journal of banking & finance* 26.7 (2002), pp. 1443–1471.

[37] D. Bertsimas, D. B. Brown, and C. Caramanis. “Theory and applications of robust optimization”. In: *SIAM Review* 53.3 (2011), pp. 464–501.

[38] G. G. Roussas. *An introduction to probability and statistical inference*. Elsevier, 2003.

[39] R. Bartoszynski and M. Niewiadomska-Bugaj. *Probability and statistical inference*. John Wiley & Sons, 2008.

[40] B. Ghojogh, A. Ghojogh, M. Crowley, and F. Karray. “Fitting a mixture distribution to data: tutorial”. In: *arXiv preprint arXiv:1901.06708* (2019).

[41] D. Cousineau, S. Brown, and A. Heathcote. “Fitting distributions using maximum likelihood: Methods and packages”. In: *Behavior Research Methods, Instruments, & Computers* 36.4 (2004), pp. 742–756.

[42] W. Wertz and B. Schneider. “Statistical density estimation: a bibliography”. In: *International Statistical Review/Revue Internationale de Statistique* (1979), pp. 155–175.

[43] L. Devroye. *A course in density estimation*. Birkhauser Boston Inc., 1987.

[44] B. W. Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

[45] I. J. Myung. “Tutorial on maximum likelihood estimation”. In: *Journal of mathematical Psychology* 47.1 (2003), pp. 90–100.

[46] T. W. Anderson and I. Olkin. “Maximum-likelihood estimation of the parameters of a multivariate normal distribution”. In: *Linear algebra and its applications* 70 (1985), pp. 147–171.

[47] S. R. Eliason. *Maximum likelihood estimation: Logic and practice*. Sage Publications, 1993.

[48] H. White. "Maximum likelihood estimation of misspecified models". In: *Econometrica: Journal of the econometric society* (1982), pp. 1–25.

[49] A. W. V. der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000.

[50] H. Chardon, M. Lerasle, and J. Mourtada. "Finite-sample performance of the maximum likelihood estimator in logistic regression". In: *arXiv preprint arXiv:2411.02137* (2024).

[51] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[52] J. K. Ghosh and R. V. Ramamoorthi. *Bayesian nonparametrics*. Springer, 2003.

[53] D. W. Scott. "On optimal and data-based histograms". In: *Biometrika* 66.3 (1979), pp. 605–610.

[54] D. Freedman and P. Diaconis. "On the histogram as a density estimator: L 2 theory". In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 57.4 (1981), pp. 453–476.

[55] G. Lugosi and A. Nobel. "Consistency of data-driven histogram methods for density estimation and classification". In: *The Annals of Statistics* 24.2 (1996), pp. 687–706.

[56] A. W. V. D. Vaart and J. A. Wellner. "Weak convergence". In: *Weak convergence and empirical processes: with applications to statistics*. Springer, 1996, pp. 16–28.

[57] P. Massart. "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality". In: *The annals of Probability* (1990), pp. 1269–1283.

[58] M. Naaman. "On the tight constant in the multivariate Dvoretzky–Kiefer–Wolfowitz inequality". In: *Statistics & Probability Letters* 173 (2021), p. 109088.

[59] N. Fournier and A. Guillin. "On the rate of convergence in Wasserstein distance of the empirical measure". In: *Probability theory and related fields* 162.3 (2015), pp. 707–738.

[60] E. Parzen. "On estimation of a probability density function and mode". In: *The annals of mathematical statistics* 33.3 (1962), pp. 1065–1076.

[61] J. Shao. *Mathematical statistics, 2nd Edition*. Springer, 2003.

[62] B. G. Lindsay and M. L. Lesperance. "A review of semiparametric mixture models". In: *Journal of statistical planning and inference* 47.1-2 (1995), pp. 29–39.

[63] S. Xiang, W. Yao, and G. Yang. "An overview of semiparametric extensions of finite mixture models". In: (2019).

[64] R. B. Nelsen. *An Introduction to Copulas*. Springer, 2006.

[65] S. Kim, R. Pasupathy, and S. G. Henderson. “A guide to sample average approximation”. In: *Handbook of simulation optimization* (2014), pp. 207–243.

[66] Z. Psaradakis and M. Sola. “Finite-sample properties of the maximum likelihood estimator in autoregressive models with Markov switching”. In: *Journal of Econometrics* 86.2 (1998), pp. 369–386.

[67] M. C. Campi and S. Garatti. *Introduction to the scenario approach*. SIAM, 2018.

[68] M. C. Campi, S. Garatti, and M. Prandini. “The scenario approach for systems and control design”. In: *Annual Reviews in Control* 33.2 (2009), pp. 149–157.

[69] M. C. Campi and S. Garatti. “A sampling-and-discarding approach to chance-constrained optimization: feasibility and optimality”. In: *Journal of optimization theory and applications* 148.2 (2011), pp. 257–280.

[70] M. C. Campi and S. Garatti. “The exact feasibility of randomized solutions of uncertain convex programs”. In: *SIAM Journal on Optimization* 19.3 (2008), pp. 1211–1230.

[71] E. Delage and Y. Ye. “Distributionally robust optimization under moment uncertainty with application to data-driven problems”. In: *Operations Research* 58.3 (2010), pp. 595–612.

[72] Y. Zhang, R. Jiang, and S. Shen. “Ambiguous chance-constrained binary programs under mean-covariance information”. In: *SIAM Journal on Optimization* 28.4 (2018), pp. 2922–2944.

[73] M. R. Wagner. “Stochastic 0–1 linear programming under limited distributional information”. In: *Operations Research Letters* 36.2 (2008), pp. 150–156.

[74] X. Yu and S. Shen. “Multistage distributionally robust mixed-integer programming with decision-dependent moment-based ambiguity sets”. In: *Mathematical Programming* 196.1 (2022), pp. 1025–1064.

[75] Z. Hu and L. J. Hong. “Kullback-Leibler divergence constrained distributionally robust optimization”. In: *Available at Optimization Online* 1.2 (2013), p. 9.

[76] R. Jiang and Y. Guan. “Data-driven chance constrained stochastic program”. In: *Mathematical Programming* 158.1-2 (2016), pp. 291–327.

[77] I. Tzortzis, C. D. Charalambous, and T. Charalambous. “Dynamic programming subject to total variation distance ambiguity”. In: *SIAM Journal on Control and Optimization* 53.4 (2015), pp. 2040–2075.

[78] C. Villani. “The Wasserstein distances”. In: *Optimal transport: old and new*. Springer, 2009, pp. 93–111.

[79] P. M. Esfahani and D. Kuhn. “Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations”. In: *Mathematical Programming* 171.1 (2018), pp. 115–166.

[80] A. Sinha, H. Namkoong, R. Volpi, and J. Duchi. “Certifying some distributional robustness with principled adversarial training”. In: *arXiv preprint arXiv:1710.10571* (2017).

[81] C. Ge, L. Zhang, and Z. Yuan. “Distributionally robust optimization for the closed-loop supply chain design under uncertainty”. In: *AIChE Journal* 68.12 (2022), e17909.

[82] R. Zhu, H. Wei, and X. Bai. “Wasserstein metric based distributionally robust approximate framework for unit commitment”. In: *IEEE Transactions on Power Systems* 34.4 (2019), pp. 2991–3001.

[83] A. Hakobyan and I. Yang. “Wasserstein distributionally robust motion control for collision avoidance using conditional Value-at-Risk”. In: *IEEE Transactions on Robotics* (2021).

[84] Z. Zhong, E. A. del Rio-Chanona, and P. Petsagkourakis. *Data-driven distributionally robust MPC using the Wasserstein metric*. 2021.

[85] S. Shafieezadeh-Abadeh, D. Kuhn, and P. M. Esfahani. “Regularization via mass transportation”. In: *Journal of Machine Learning Research* 20.103 (2019), pp. 1–68.

[86] J. Blanchet, Y. Kang, and K. Murthy. “Robust Wasserstein profile inference and applications to machine learning”. In: *Journal of Applied Probability* 56.3 (2019), pp. 830–857.

[87] R. Gao. “Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality”. In: *Operations Research* 71.6 (2023), pp. 2291–2306.

[88] Q. Li, Y. Shi, Y. Jiang, Y. Shi, H. Wang, and H. V. Poor. “A distributionally robust model predictive control for static and dynamic uncertainties in smart grids”. In: *IEEE Transactions on Smart Grid* 15.5 (2024), pp. 4890–4902.

[89] Y. Chen, Y. Li, S. Li, and X. Yin. “Distributionally Robust Control Synthesis for Stochastic Systems with Safety and Reach-Avoid Specifications”. In: *arXiv preprint arXiv:2501.03137* (2025).

[90] Y. Li, K. Li, R. Fan, J. Chen, and Y. Zhao. “Multi-objective planning of distribution network based on distributionally robust model predictive control”. In: *Frontiers in Energy Research* 12 (2024), p. 1478040.

[91] S. Yüksel and T. Basar. “Stochastic networked control systems”. In: *AMC* 10 (2013), p. 12.

[92] I. Yang. “Wasserstein distributionally robust stochastic control: A data-driven approach”. In: *IEEE Transactions on Automatic Control* 66.8 (2021), pp. 3863–3870.

[93] C. Ma, T. Li, and J. Zhang. “Consensus control for leader-following multi-agent systems with measurement noises”. In: *Journal of Systems Science and Complexity* 23.1 (2010), pp. 35–49.

[94] V. A. Battagello and C. H. C. Ribeiro. “Analysis of the effects of failure and noise in the distributed connectivity maintenance of a multi-robot system”. In: *2014 IEEE 15th International Symposium on Computational Intelligence and Informatics (CINTI)*. IEEE. 2014, pp. 427–432.

[95] L. M. Chaouach, T. Oomen, and D. Boskos. *Structured ambiguity sets for distributionally robust optimization*. Provisionally accepted for publication in *European Journal of Control*. 2023. arXiv: 2310.20657 [math.OC].

[96] L. M. Chaouach, T. Oomen, and D. Boskos. *Tractable reformulations of DRO problems over structured optimal transport ambiguity sets*. Accepted for publication in *Transactions on Automatic Control*. 2025. arXiv: 2504.06966 [math.OC].

[97] L. M. Chaouach, T. Oomen, and D. Boskos. "Comparing Structured Ambiguity Sets for Stochastic Optimization: Application to Uncertainty Quantification". In: *IEEE Int. Conf. on Decision and Control*. 2023, pp. 8274–8279.

[98] K. Marti. *Stochastic Optimization Methods: Applications in Engineering and Operations Research*. 3rd. Springer Berlin, Heidelberg, 2015.

[99] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. Vol. 16. SIAM, 2014.

[100] W. Wiesemann, D. Kuhn, and M. Sim. "Distributionally Robust Convex Optimization". In: *Operations Research* 62 (Dec. 2014), pp. 1358–1376.

[101] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. "Wasserstein distributionally robust optimization: Theory and applications in machine learning". In: *Operations research & management science in the age of analytics*. Informs, 2019, pp. 130–166.

[102] R. Chen and I. C. Paschalidis. "A Robust Learning Approach for Regression Models Based on Distributionally Robust Optimization". In: *Journal of Machine Learning Research* 19.13 (2018), pp. 1–48.

[103] B. P. G. Van Parys, D. Kuhn, P. J. Goulart, and M. Morari. "Distributionally robust control of constrained stochastic systems". In: *IEEE Transactions on Automatic Control* 61.2 (2015), pp. 430–442.

[104] B. Li, Y. Tan, A. Wuo, and G. Duan. "A Distributionally Robust Optimization Based Method for Stochastic Model Predictive Control". In: *IEEE Transactions on Automatic Control* 67.11 (2022), pp. 5762–5776.

[105] I. Tzortzis, C. D. Charalambous, and C. N. Hadjicostis. "A Distributionally Robust LQR for Systems with Multiple Uncertain Players". In: *IEEE Int. Conf. on Decision and Control*. 2021, pp. 3972–3977.

[106] L. Aolaritei, M. Fochesato, J. Lygeros, and F. Dörfler. "Wasserstein Tube MPC with Exact Uncertainty Propagation". In: *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE. 2023, pp. 2036–2041.

[107] J. Coulson, J. Lygeros, and F. Dörfler. "Distributionally Robust Chance Constrained Data-Enabled Predictive Control". In: *IEEE Transactions on Automatic Control* 67.7 (2022), pp. 3289–3304.

[108] S. Shafieezadeh-Abadeh, V. A. Nguyen, D. Kuhn, and P. M. Esfahani. "Wasserstein Distributionally Robust Kalman Filtering". In: *Advances in Neural Information Processing Systems*. 2018, pp. 8474–8483.

[109] D. Boskos, J. Cortés, and S. Martinez. “Data-driven ambiguity sets with probabilistic guarantees for dynamic processes”. In: *IEEE Transactions on Automatic Control* 66.7 (2021), pp. 2991–3006.

[110] D. Boskos, J. Cortés, and S. Martínez. “High-confidence data-driven ambiguity sets for time-varying linear systems”. In: *IEEE Transactions on Automatic Control* 69.2 (2024), pp. 797–812.

[111] L. Aolaritei, N. Lanzetti, H. Chen, and F. Dörfler. “Distributional Uncertainty Propagation via Optimal Transport”. In: *IEEE Transactions on Automatic Control* (2025).

[112] B. K. Poolla, A. R. Hota, S. Bolognani, D. S. Callaway, and A. Cherukuri. “Wasserstein distributionally robust look-ahead economic dispatch”. In: *IEEE Transactions on Power Systems* 36.3 (2020), pp. 2010–2022.

[113] D. Li, D. Fooladivanda, and S. Martínez. “Data-driven variable speed limit design for highways via distributionally robust optimization”. In: *European Control Conference*. 2019, pp. 1055–1061.

[114] G. C. Calafiore and L. E. Ghaoui. “On distributionally robust chance-constrained linear programs”. In: *Journal of Optimization Theory & Applications* 130.1 (2006), pp. 1–22.

[115] I. Popescu. “Robust mean-covariance solutions for stochastic optimization”. In: *Operations Research* 55.1 (2007), pp. 98–112.

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

[117] C. Villani. *Optimal transport: old and new*. Vol. 338. Springer, 2008.

[118] R. Gao and A. J. Kleywegt. “Data-Driven Robust Optimization with Known Marginal Distributions”. In: 2017.

[119] R. Gao and A. J. Kleywegt. “Distributionally Robust Stochastic Optimization with Wasserstein Distance”. In: *Mathematics of Operations Research* 48.2 (2023), pp. 603–655.

[120] J. Blanchet and K. Murthy. “Quantifying Distributional Model Risk via Optimal Transport”. In: *Mathematics of Operations Research* 44.2 (2019), pp. 565–600.

[121] S. Dereich, M. Scheutzow, and R. Schottstedt. “Constructive quantization: Approximation by empirical measures”. In: *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques* 49.4 (2013), pp. 1183–1203.

[122] J. Weed and F. Bach. “Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance”. In: *Bernoulli* 25.4A (2019), pp. 2620–2648.

[123] J. Blanchet, Y. Kang, and K. Murthy. “Robust Wasserstein profile inference and applications to machine learning”. In: *Journal of Applied Probability* 56.3 (2019), pp. 830–857.

[124] J. Blanchet, K. Murthy, and N. Si. “Confidence Regions in Wasserstein Distributionally Robust Estimation”. In: *Biometrika* 109 (2 2021), pp. 295–315.

[125] R. Gao. “Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality”. In: *Operations Research* 0.0 (2022), null.

[126] N. Si, J. Blanchet, S. Ghosh, and M. Squillante. “Quantifying the Empirical Wasserstein Distance to a Set of Measures: Beating the Curse of Dimensionality”. In: *Advances in Neural Information Processing Systems*. 2020, pp. 21260–21270.

[127] J. Blanchet, Y. Kang, K. Murthy, and F. Zhang. “Data-Driven Optimal Transport Cost Selection For Distributionally Robust Optimization”. In: *2019 Winter Simulation Conference (WSC)*. 2019, pp. 3740–3751.

[128] J. Blanchet, K. Murthy, and F. Zhang. “Optimal Transport-Based Distributionally Robust Optimization: Structural Properties and Iterative Schemes”. In: *Mathematics of Operations Research* 47.2 (2022), pp. 1500–1529.

[129] D. Bartl, M. Kupper, T. Lux, A. Papapantoleon, and S. Eckstein. “Marginal and Dependence Uncertainty: Bounds, Optimal Transport, and Sharpness”. In: *SIAM Journal on Control and Optimization* 60.1 (2022), pp. 410–434.

[130] P. Coppens and P. Patrinos. “Data-Driven Distributionally Robust MPC for Constrained Stochastic Systems”. In: *IEEE Control Systems Letters* 6 (2022), pp. 1274–1279.

[131] F. Wu, M. E. Villanueva, and B. Houska. “Ambiguity tube MPC”. In: *Automatica* 146 (2022), p. 110648.

[132] I. Gracia, D. Boskos, L. Laurenti, and M. Mazo Jr. “Distributionally robust strategy synthesis for switched stochastic systems”. In: *Proceedings of the 26th ACM International Conference on Hybrid Systems: Computation and Control*. 2023, pp. 1–10.

[133] A. A. Malikopoulos, V. Maroulas, and J. Xiong. “A multiobjective optimization framework for stochastic control of complex systems”. In: *2015 American Control Conference (ACC)*. 2015, pp. 4263–4268.

[134] V. Shestak, J. Smith, H. J. Siegel, and A. A. Maciejewski. “A stochastic approach to measuring the robustness of resource allocations in distributed systems”. In: *2006 International Conference on Parallel Processing (ICPP’06)*. IEEE. 2006, pp. 459–470.

[135] D. Bertsekas and S. E. Shreve. *Stochastic optimal control: the discrete-time case*. Athena Scientific, 1996.

[136] A. Klenke. *Probability theory: a comprehensive course*. Springer, 2013.

[137] M. Huang and J. H. Manton. “Coordination and consensus of networked agents with noisy measurements: Stochastic algorithms and asymptotic behavior”. In: *SIAM Journal on Control and Optimization* 48.1 (2009), pp. 134–161.

[138] T. Li and J. Zhang. “Mean square average-consensus under measurement noises and fixed topologies: Necessary and sufficient conditions”. In: *Automatica* 45.8 (2009), pp. 1929–1936.

[139] L. Cheng, Z. G. Hou, and M. Tan. “A mean square consensus protocol for linear multi-agent systems with communication noises and fixed topologies”. In: *IEEE Transactions on Automatic Control* 59.1 (2013), pp. 261–267.

[140] N. Atanasov, J. L. Ny, K. Daniilidis, and G. J. Pappas. “Decentralized active information acquisition: Theory and application to multi-robot SLAM”. In: *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2015, pp. 4775–4782.

[141] L. Aolaritei, N. Lanzetti, and F. Dörfler. “Capture, propagate, and control distributional uncertainty”. In: *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE. 2023, pp. 3081–3086.

[142] B. Taskesen, D. Iancu, Ç. Koçyiğit, and D. Kuhn. “Distributionally robust linear quadratic control”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 18613–18632.

[143] C. Ning, A. Ma, X. Ma, L. Li, G. Pan, W. Gu, W. Du, Z. Dong, and M. Shahidehpour. “Multi-Stage Distributionally Robust Scheduling with Structured Mixture Ambiguity for Hydrogen-Based Integrated Energy Systems: Finite-Sample Guarantees and Equivalent Reformulations”. In: *IEEE Transactions on Power Systems* (2025).

[144] C. Ning, A. Ma, and Z. Dong. “Data-driven multi-stage distributionally robust scheduling for coupled electricity-hydrogen-refinery systems”. In: *Applied Energy* 401 (2025), p. 126620.

[145] J. Dedecker and F. Merlevède. “Behavior of the empirical Wasserstein distance in R^d under moment conditions”. In: *Electronic Journal of Probability* 24 (2019).

[146] I. Gracia, D. Boskos, L. Laurenti, and M. Lahijanian. “Data-driven strategy synthesis for stochastic systems with unknown nonlinear disturbances”. In: *6th Annual Learning for Dynamics & Control Conference*. 2024, pp. 1633–1645.

[147] P. Billingsley. *Probability and measure*. John Wiley, 2008.

[148] D. Boskos, J. Cortés, and S. Martínez. “High-Confidence Data-Driven Ambiguity Sets for Time-Varying Linear Systems”. In: *arXiv e-prints* (2021), arXiv–2102.

[149] N. Fournier. “Convergence of the empirical measure in expected wasserstein distance: non-asymptotic explicit bounds in R^d”. In: *ESAIM: Probability and Statistics* 27 (2023), pp. 749–775.

[150] B. Kloeckner. “Approximation by finitely supported measures”. In: *ESAIM: Control, Optimisation and Calculus of Variations* 18 (2012), pp. 343–359.

[151] K. Politis and S. M. Pitts. “Nonparametric Estimation in Renewal Theory II: Solutions of Renewal-Type Equations”. In: *The Annals of Statistics* 28.1 (2000), pp. 88–115.

[152] A. Barvinok. *A course in convexity*. Vol. 54. American Mathematical Society, 2002.

[153] T. Eisner, B. Farkas, M. Haase, and R. Nagel. *Operator theoretic aspects of ergodic theory*. Springer, 2015.

[154] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, 2010.

[155] K. Fan. "Minimax Theorems*". In: *Proceedings of the National Academy of Sciences* 39.1 (1953), pp. 42–47.

[156] D. P. Bertsekas. *Convex Optimization Theory*. Belmont. Athena Scientific, 2009.

[157] S. Garatti and M. C. Campi. "Modulating robustness in control design: Principles and algorithms". In: *IEEE Control Systems Magazine* 33.2 (2013), pp. 36–51.

[158] G. C. Calafiore and L. Fagiano. "Robust model predictive control via scenario optimization". In: *IEEE Transactions on Automatic Control* 58.1 (2012), pp. 219–224.

[159] S. Garatti, M. C. Campi, and A. Care. "On a class of interval predictor models with universal reliability". In: *Automatica* 110 (2019), p. 108542.

[160] D. Bertsimas, V. Gupta, and N. Kallus. "Data-driven robust optimization". In: *Mathematical Programming* 167 (2018), pp. 235–292.

[161] J. Liu, Y. Chen, C. Duan, J. Lin, and J. Lyu. "Distributionally Robust Optimal Reactive Power Dispatch with Wasserstein Distance in Active Distribution Network". In: *Journal of Modern Power Systems and Clean Energy* 8.3 (2020), pp. 426–436.

[162] R. Jiang, M. Ryu, and G. Xu. "Data-Driven Distributionally Robust Appointment Scheduling over Wasserstein Balls". In: *arXiv preprint arXiv:1907.03219* (2019).

[163] A. Nilim and L. E. Ghaoui. "Robust control of Markov decision processes with uncertain transition matrices". In: *Operations Research* 53.5 (2005), pp. 780–798.

[164] L. M. Chaouach, D. Boskos, and T. Oomen. "Uncertain uncertainty in data-driven stochastic optimization: towards structured ambiguity sets". In: *IEEE Int. Conf. on Decision and Control*. 2022, pp. 4776–4781.

[165] R. Gao, R. Arora, and Y. Huang. "Data-driven multistage distributionally robust linear optimization with nested distance". In: *arXiv preprint arXiv:2407.16346* (2024).

[166] N. Lanzetti, A. Terpin, and F. Dörfler. "Optimality of linear policies for distributionally robust linear quadratic Gaussian regulator with stationary distributions". In: *arXiv preprint arXiv:2410.22826* (2024).

[167] R. D. R. D. McAllister and P. M. Esfahani. "Distributionally robust model predictive control: Closed-loop guarantees and scalable algorithms". In: *IEEE Transactions on Automatic Control* (2024).

[168] A. R. Hota, A. Cherukuri, and J. Lygeros. *Data-Driven Chance Constrained Optimization under Wasserstein Ambiguity Sets*. 2018.

[169] D. Bertsekas and S. E. Shreve. *Stochastic optimal control: the discrete-time case*. Athena Scientific, 1978.

[170] M. C. Yue, D. Kuhn, and W. Wiesemann. "On linear optimization over Wasserstein balls". In: *Mathematical Programming* 195.1-2 (June 2021), pp. 1107–1122.

[171] J. Kristensen and F. Rindler. "Piecewise affine approximations for functions of bounded variation". In: *Numerische Mathematik* 132 (2016), pp. 329–346.

[172] G. Schildbach, L. Fagiano, and M. Morari. "Randomized solutions to convex programs with multiple chance constraints". In: *SIAM Journal on Control and Optimization* 23.4 (2013), pp. 2479–2501.

[173] J. Sengupta. "Stochastic linear programming with chance constraints". In: *International Economic Review* 11.1 (1970), pp. 101–116.

[174] L. Hewing, K. P. Wabersich, and M. N. Zeilinger. "Recursively feasible stochastic model predictive control using indirect feedback". In: *Automatica* 119 (2020), p. 109095. ISSN: 0005-1098.

[175] M. Fiacchini, M. Mammarella, and F. Dabbene. *Measured-state conditioned recursive feasibility for stochastic model predictive control*. 2024. arXiv: 2406.13522.

[176] A. Hota, A. Cherukuri, and J. Lygeros. "Data-driven chance constrained optimization under wasserstein ambiguity sets". English. In: *Proceedings of the American Control Conference, ACC 2019*. Institute of Electrical and Electronics Engineers Inc., July 2019, pp. 1501–1506.

[177] R. T. Rockafellar, S. Uryasev, *et al.* "Optimization of conditional value-at-risk". In: *Journal of risk* 2 (2000), pp. 21–42.

[178] D. Li and S. Martínez. "Data Assimilation and Online Optimization With Performance Guarantees". In: *IEEE Transactions on Automatic Control* 66.5 (2021), pp. 2115–2129.

[179] F. Fabiani and P. J. Goulart. "The optimal transport paradigm enables data compression in data-driven robust control". In: *2021 American Control Conference (ACC)*. 2021, pp. 2412–2417.

[180] S. Lloyd. "Least squares quantization in PCM". In: *IEEE transactions on information theory* 28.2 (1982), pp. 129–137.

[181] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006.

[182] G. Peyre and M. Cuturi. "Computational Optimal Transport". In: *Foundations and Trends in Machine Learning* 11.5-6 (2019), pp. 355–607.

[183] M. Catalano and H. Lavenant. "Measures of Dependence based on Wasserstein distances". In: *arXiv preprint arXiv:2510.06034* (2025).

[184] R. T. Rockafellar and R. J. B. Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften. Springer, 2010.

[185] A. Shapiro and A. Kleywegt. “Minimax analysis of stochastic problems”. In: *Optimization Methods and Software* 17.3 (2002), pp. 523–542.

[186] A. D. Ioffe and V. M. Tihomirov. *Theory of extremal problems: Theory of extremal problems*. North-Holland Publishing Company.

[187] R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy. “The effectiveness of Lloyd-type methods for the k-means problem”. In: *Journal of the ACM (JACM)* 59.6 (2013), pp. 1–22.

[188] D. Arthur and S. Vassilvitskii. “k-means++: The advantages of careful seeding”. In: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA ’07. New Orleans, Louisiana: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[189] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii. “Scalable k-means++”. In: *Proc. VLDB Endow.* 5.7 (Mar. 2012), pp. 622–633.

[190] D. Mayne, J. Rawlings, C. Rao, and P. Scokaert. “Constrained model predictive control: Stability and optimality”. In: *Automatica* 36.6 (2000), pp. 789–814.

[191] B. Kouvaritakis and M. Cannon. “Model predictive control”. In: *Switzerland: Springer International Publishing* 38 (2016), pp. 13–56.

[192] J. B. Rawlings, D. Q. Mayne, M. Diehl, *et al. Model predictive control: theory, computation, and design*. Vol. 2. Nob Hill Publishing Madison, WI, 2017.

[193] D. Q. Mayne and W. Langson. “Robustifying model predictive control of constrained linear systems”. In: *Electronics letters* 37.23 (2001), pp. 1422–1423.

[194] M. Farina, L. Giulioni, L. Magni, and R. Scattolini. “A probabilistic approach to Model Predictive Control”. In: *52nd IEEE Conference on Decision and Control*. 2013, pp. 7734–7739.

[195] M. Lorenzen, F. Dabbene, R. Tempo, and F. Allgöwer. “Constraint-tightening and stability in stochastic model predictive control”. In: *IEEE Transactions on Automatic Control* 62.7 (2016), pp. 3165–3177.

[196] L. Hewing and M. N. Zeilinger. “Stochastic Model Predictive Control for Linear Systems Using Probabilistic Reachable Sets”. In: *2018 IEEE Conference on Decision and Control (CDC)*. 2018, pp. 5182–5188.

[197] M. Korda, R. Gondhalekar, F. Oldewurtel, and C. N. Jones. “Stochastic MPC framework for controlling the average constraint violation”. In: *IEEE Transactions on Automatic Control* 59.7 (2014), pp. 1706–1721.

[198] L. M. Chaouach, M. Fiacchini, and T. Alamo. “Stochastic model predictive control for linear systems affected by correlated disturbances”. In: *ROCOND 2022 - 10th IFAC Symposium on Robust Control Design*. Kyoto, Japan, 2022.

[199] J. Nie, L. Yang, S. Zhong, and G. Zhou. “Distributionally robust optimization with moment ambiguity sets”. In: *Journal of Scientific Computing* 94.1 (2023), p. 12.

[200] A. Zolanvari and A. Cherukuri. “Data-driven distributionally robust iterative risk-constrained model predictive control”. In: *2022 European Control Conference (ECC)*. 2022, pp. 1578–1583.

[201] C. Mark and S. Liu. “Stochastic MPC with Distributionally Robust Chance Constraints”. In: *IFAC-PapersOnLine* 53.2 (2020), pp. 7136–7141.

[202] S. Gillijns and B. De Moor. “Unbiased minimum-variance input and state estimation for linear discrete-time systems”. In: *Automatica* 43.1 (2007), pp. 111–116. ISSN: 0005-1098.

[203] M. Korda, R. Gondhalekar, J. Cigler, and F. Oldewurtel. “Strongly feasible stochastic model predictive control”. In: *2011 50th IEEE Conference on Decision and Control and European Control Conference*. IEEE. 2011, pp. 1245–1251.

[204] P. Gutman, M. Cwikel, *et al.* “An algorithm to find maximal state constraint sets for discrete-time linear dynamical systems with bounded controls and states”. In: *IEEE Transactions on Automatic Control* 32.3 (1987), pp. 251–254.

[205] L. M. Chaouach, T. Oomen, and D. Boskos. *Distributionally robust model predictive control using horizon-adaptive ambiguity sets*. Submitted for journal publication. 2026.

[206] L. M. Chaouach, D. Boskos, and T. Oomen. “Reformulations for data-driven stochastic optimization problems with structured ambiguity sets”. In: *42nd Benelux Meeting on Systems and Control 2023*. 2023.

[207] L. M. Chaouach, D. Boskos, and T. Oomen. “Tightening ambiguity set characterizations for data-driven distributionally robust optimization”. In: *41st Benelux Meeting on Systems and Control 2022*. 2022.

Acknowledgements

If I have seen further it is by standing on the shoulders of Giants.

Isaac Newton

First and foremost, I wish to express my sincere gratitude to my supervisor Dimitris, for everything he has taught me. I am deeply grateful for his exceptional guidance, mentorship, and support throughout this doctoral journey. His insightful feedback, intellectual rigor, great intuition, and unwavering belief in this research have been instrumental in shaping the quality of this work. His consistent encouragement to think critically and independently cultivated an environment where curiosity was valued, and work was genuinely enjoyable. Beyond his technical guidance, Dimitris also instilled in me a deep passion for research and mathematics, which has been pivotal to my growth as a researcher. I am especially thankful for his patience during challenging periods, as well as his generosity and enthusiasm in discussing ideas both within and beyond the scope of this thesis.

I would also like to express my gratitude to Tom, who has shown a genuine investment in this work. His thoughtful guidance was instrumental in bringing this research to its current level of quality. With his advanced expertise and complementary perspective, he consistently pushed me to think outside the box and to adopt a more critical stance toward my own work. He also played a key role in helping me realize that the quality of scientific presentation is just as important as the quality of the results themselves. I am additionally thankful for his patience and continuous encouragement, especially when the path forward seemed uncertain.

Next, I want to thank my fantastic friends and research group fellows, Ilias and Tea. It was an honor and a great pleasure to interact with such brilliant people. I am deeply grateful to Ilias for the enriching discussions and brainstorming sessions; I had a lot of fun collaborating with you. Thank you for the lively discussions, the shared frustrations, the celebrations of small victories, and the unspoken understanding of what pursuing a PhD means.

The members of DCSC have been both colleagues and friends throughout this journey. Therefore, a huge thanks to all of them for creating a fantastic environment that I enjoyed being part of. In particular, I would like to thank Alex, Rayyan, Sasan, Darya, Ivo, Leila, Claudia, Léonore, Silvia, Coen, Pradyumna, Steven, David, Anil, Luca, Edoardo, Jonas, Filippo, Rogier, Francesco, Alessandro, Amin, Frederik, Athina, Gianpietro, Jesse, Robin, Mahshad, Giorgos, Ashkan, Thomas, Salim, Afra, Zoé, Arghya, Eva, Alexandra, Tolga, Changrui, Suad, Fritz, Sreeshma, Maarteen, Sam, Marcus, Wolfram, Yun, Haoyu, Matteo, Shahzeb, Tim, Maria de Neves de Fonseca,

Maria Bartzioka, Reza Rahimi Baghbadorani, Reza Riahi Samani, Mohammad Boveiri, and Mohammad Shokri. I am also deeply appreciative of the professors and senior scientific staff. A distinctive thanks to Kim, Manon, Jan-Willem, Riccardo, Manuel, Peyman, Gabriel, Carlo, Azita, Koty, Luca, Sebastiaan, Nitin, Shengling, Marta, and Amin. I would like to extend a special thank you to Max and Mohamed for the insightful, engaging, and sometimes thought-provoking discussions we shared. Those exchanges were truly intellectually enriching, and they pushed me to, sometimes, reconsider several key aspects of my own thinking. I deeply appreciate the time and genuine interest they brought to every conversation.

I would like to acknowledge the administrative staff of DCSC whose contributions, though often behind the scenes, have been essential to this research. In particular, thank you, Francy, Helen, Anna, Bo, Sandra, Renate, Erica, and Marieke for managing the bureaucratic aspects and institutional requirements.

I am deeply grateful to all my friends, who have provided essential balance, perspective, and joy in my life. Omid and Liwia deserve special mention. Their warmth and generosity have meant more than words can express. They are not just friends, but represent the family I have in the Netherlands. Adel's friendship has been equally cherished. Our shared fascination with mathematics, history, and geopolitics has sparked countless conversations that have pushed my thinking forward. Beyond our intellectual kinship, his unwavering belief in constant growth and evolution has always been a great source of inspiration. I am grateful to Dalil, Islem, Oussama, and Malek for their genuine esteem, thoughtful consideration, and for continually challenging me to develop my perspectives. Iskander, Yanis, and Yacine have been sources of good humor throughout, always present, always willing to help, and always ready to lighten the mood. I am also thankful to Rachika, Abdelmalek, Yacine, and Camélia for their refined artistic sensibility, exceptional musical taste, and effortless coolness. Finally, I extend my heartfelt gratitude to all of my friends: Samy, Walid, Rania, Chakib, Lotfi, Kader, Aissam, Samed, Mohamed, Youssef, Lyes, Fadhlo, Haidar, Moumouh, Younes, Kamil, Yanis, Riad, Nadim, Amine, Nassila, and Namiz for the friendship, laughter, and countless moments of connection.

Je souhaite à présent exprimer ma profonde reconnaissance à l'ensemble des enseignants ayant contribué à ma formation scolaire et académique. À l'École Nationale Polytechnique, j'ai eu la chance d'étudier auprès d'enseignants rigoureux et passionnés qui m'ont transmis une solide base et qui m'ont inspiré à poursuivre une carrière dans les domaines scientifique et technique. En particulier, j'adresse mes remerciements distingués à Madame Laleg, ainsi qu'à Messieurs Tadjine, Kebli et Amokrane. Je tiens aussi à honorer la mémoire de Monsieur Abdelouel, qui nous a malheureusement quittés dans de tragiques circonstances. Pendant ma formation de master à l'Université Grenoble Alpes, j'ai bénéficié de l'enseignement de professeurs d'exception dont l'expertise, la rigueur intellectuelle et l'engagement envers leurs étudiants m'ont profondément marqué. Ces fondations solides et cette formation rigoureuse ont été essentielles à ma progression en tant que chercheur et ont jeté les bases de ce doctorat. Je serai éternellement reconnaissant pour les enseignements que j'ai reçus. En particulier, j'aimerais exprimer ma profonde gratitude à Messieurs Witrant et Fiacchini. J'aimerais aussi témoigner ma reconnaissance envers les

enseignants Bendjoudi, Sayeh, Maaradji, Mekhfi, Maayouf, Aksa et Benlala.

Pour finir, à ma famille,

Je souhaite dédier ce travail à mes parents Ali et Nadéra, dont l'amour, la bienveillance et les sacrifices ont constitué le fondement de tout ce que j'ai pu accomplir. Leur soutien indéfectible m'a accompagné dans chacun des défis que j'ai rencontrés au cours de mon existence. À Tarik et Sabrina, dont l'exemple, les conseils et la présence ont profondément influencé la personne que je suis devenue aujourd'hui. À Amina et Arslane, avec qui j'ai grandi et partagé les souvenirs et les joies qui ont façonné notre lien et notre complicité au fil des années. À mes oncles Amor et Houari, mes tantes Hafedha et Fouzia, ainsi qu'à mes cousins Samy, Ghouti, Zak, Sonia, Kamel, Redha, Zoubir, Mohamed, Bilal, Abdelkader, Ahmed, Karima, Malika, Lila, Amina, Meriem, Bachira, Saadia, mes neveux Élias, Tania, Mario, Ulysse, et Swan, ainsi que les petits Wassim et Célia pour leur affection, leur présence et tous les moments partagés qui donnent à la famille sa véritable richesse. Enfin, j'ai une pensée particulière pour mes grands-parents Zitouni, el Alia, Khadidja, Mustapha et Fatiha ainsi que pour mes oncles Tahar, Yagoub, Abdelkader, Zitouni et Mohamed et tantes Kharfia et Safia disparus. Leur souvenir ou leur mémoire continuent de m'accompagner aujourd'hui. À travers cette thèse, c'est une part de chacun de vous que je porte avec moi.

Chaleureusement,
Lotfi M. Chaouach

About the author

Born on December 13, 1996 in Algeria, Lotfi completed his joint Eng. and M.Sc. degrees in Automatic Control from the National Polytechnic School (Algiers) in 2020. His initial research focused on observer design for time-scale systems, conducted under the supervision of Prof. M. Tadjine and Prof. T. M. Laleg-Kirati.

Seeking to deepen his expertise in the mathematics of control, Lotfi pursued the Control Systems Theory track of the Master in Systems, Control and Information Technologies at Université Grenoble Alpes, graduating in 2021, at the top of his cohort. His master's thesis, completed at GIPSA-lab under the supervision of Dr. M. Fiacchini, addressed the characterization of feasibility and convergence conditions for stochastic model predictive control for LTI systems affected by correlated disturbances.

Since October 2021, Lotfi has been pursuing his Ph.D. at the Delft Center for Systems and Control, Delft University of Technology, under the guidance of Prof. dr. ir. T. Oomen and Dr. D. Boskos. His doctoral research addresses the curse of dimensionality inherent in Wasserstein-based distributionally robust optimization and develops improved distributionally robust model predictive control formulations. Lotfi's broader research interests span data-driven uncertainty quantification, system identification, robust and stochastic optimization, optimal control, and machine learning.

List of Publications

1. Journal preprints

3. L. M. Chaouach, T. Oomen, and D. Boskos. *Distributionally robust model predictive control using horizon-adaptive ambiguity sets*. Submitted for journal publication. 2026

2. L. M. Chaouach, T. Oomen, and D. Boskos. *Tractable reformulations of DRO problems over structured optimal transport ambiguity sets*. Accepted for publication in *Transactions on Automatic Control*. 2025. arXiv: 2504.06966 [math.OC]

1. L. M. Chaouach, T. Oomen, and D. Boskos. *Structured ambiguity sets for distributionally robust optimization*. Provisionally accepted for publication in *European Journal of Control*. 2023. arXiv: 2310.20657 [math.OC]

2. Peer-reviewed articles in conference proceedings

2. L. M. Chaouach, T. Oomen, and D. Boskos. "Comparing Structured Ambiguity Sets for Stochastic Optimization: Application to Uncertainty Quantification". In: *IEEE Int. Conf. on Decision and Control*. 2023, pp. 8274–8279

1. L. M. Chaouach, D. Boskos, and T. Oomen. "Uncertain uncertainty in data-driven stochastic optimization: towards structured ambiguity sets". In: *IEEE Int. Conf. on Decision and Control*. 2022, pp. 4776–4781

3. Abstracts in conference proceedings

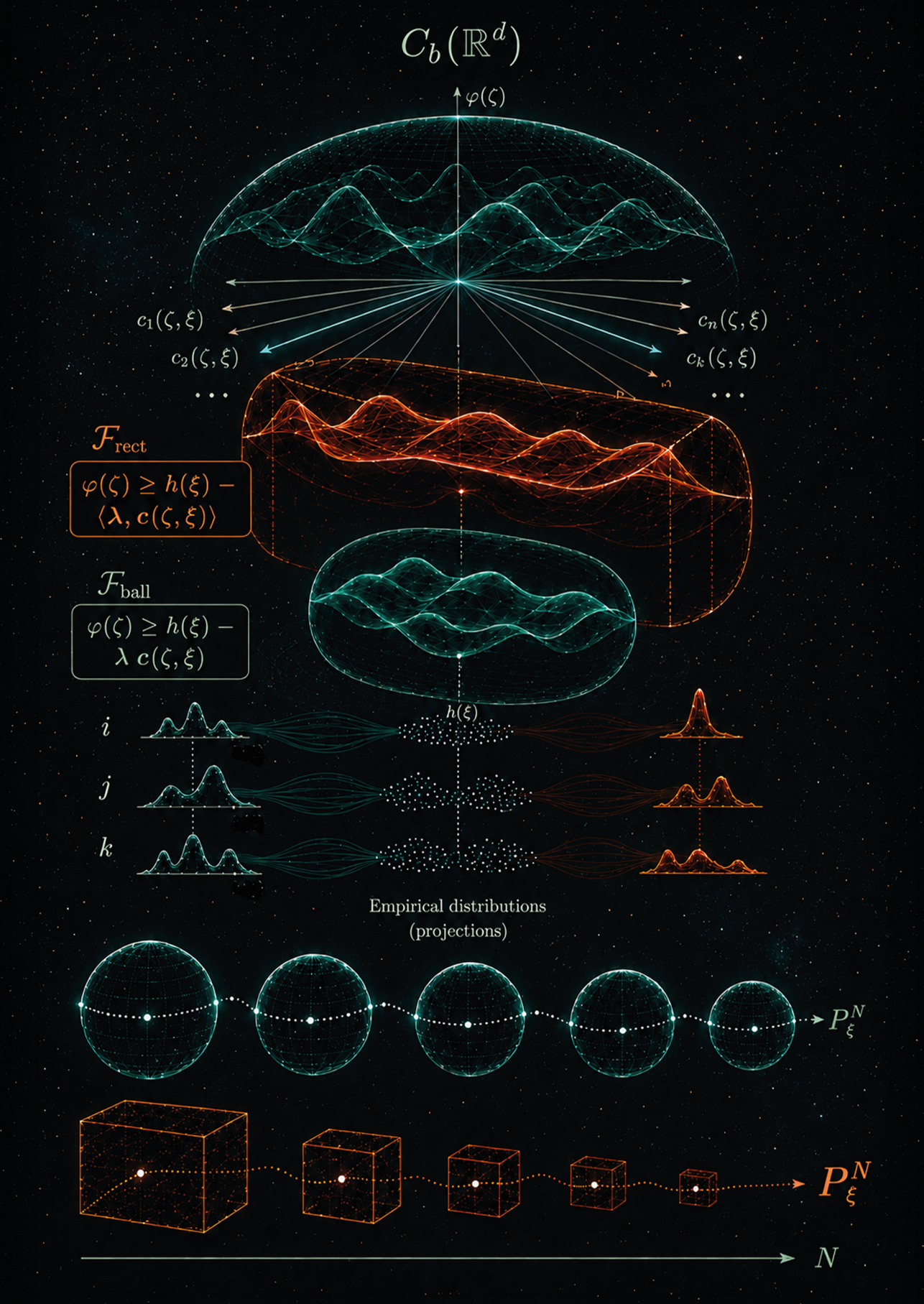
2. L. M. Chaouach, D. Boskos, and T. Oomen. "Reformulations for data-driven stochastic optimization problems with structured ambiguity sets". In: *42nd Benelux Meeting on Systems and Control 2023*. 2023

1. L. M. Chaouach, D. Boskos, and T. Oomen. "Tightening ambiguity set characterizations for data-driven distributionally robust optimization". In: *41st Benelux Meeting on Systems and Control 2022*. 2022

Unrelated peer-reviewed articles in conference proceedings

1. L. M. Chaouach, M. Fiacchini, and T. Alamo. "Stochastic model predictive control for linear systems affected by correlated disturbances". In: *ROCOND*

2022 - 10th IFAC Symposium on Robust Control Design. Kyoto, Japan, 2022


$C_b(\mathbb{R}^d)$
$\varphi(\zeta)$
$c_1(\zeta,\xi)$
$c_2(\zeta,\xi)$
$c_n(\zeta,\xi)$
$c_k(\zeta,\xi)$
$\mathcal{F}_{\text{rect}}$
$\varphi(\zeta) \geq h(\xi) - \langle \boldsymbol{\lambda}, \mathbf{c}(\zeta,\xi) \rangle$
$\mathcal{F}_{\text{ball}}$
$\varphi(\zeta) \geq h(\xi) - \lambda\, c(\zeta,\xi)$
$h(\xi)$
i
j
k
Empirical distributions
(projections)
P_ξ^N
P_ξ^N
N